Российская академия наук
Вычислительный центр РАН ФИЦ ИУ РАН
Центр хранения и анализа больших данных МГУ
Центр компетенций НТИ по направлению
«Искусственный интеллект» МФТИ

Интеллектуализация обработки информации

13-я Международная конференция

Москва, 2020

УДК 004.85+004.89+004.93+519.2+519.25+519.7
ББК 22.1:32.973.26-018.2
И 73

Интеллектуализация обработки информации: Тезисы докладов 13-й Международной конференции, г. Москва 2020 г. — М.: Российская академия наук, 2020. — 472 с.

ISBN 978-5-907366-16-9

В сборнике представлены тезисы докладов 13-й Международной конференции «Интеллектуализация обработки информации», проводимой Российской академией наук, Вычислительным центром Федерального исследовательского центра «Информатика и управление» РАН, Центром компетенций НТИ по направлению «Искусственный интеллект» на базе Московского физико-технического института, Центром хранения и анализа больших данных на базе Московского государственного университета имени М. В. Ломоносова.

Конференция проводится регулярно, начиная с 1989 г., и является представительным научным форумом в области интеллектуального анализа данных, машинного обучения, распознавания образов, анализа изображений, обработки сигналов, дискретного анализа.

Сайт конференции `http://mmro.ru`.

ISBN 978-5-907366-16-9

UDK 004.85+004.89+004.93+519.2+519.25+519.7
BBK 22.1:32.973.26-018.2

Intelligent Data Processing: Theory and Applications: Book of abstract of the 13th International Conference, Moscow, 2020. — Moscow: Russian Academy of Sciences, 2020. — 472 p.

ISBN 978-5-907366-16-9

The volume contains the abstracts of the 13th International Conference "Intelligent Data Processing: Theory and Applications". The conference is organized by the Russian Academy of Sciences, Federal Research Center "Computer Science and Control" of RAS, Center of big data storage and analysis at the Moscow State University, and the Competence Center of the National Technological Initiative "Artificial intelligence" at the Moscow Institute of Physics and Technology.

The conference has being held biennially since 1989. It is one of the most recognizable scientific forums on data mining, machine learning, pattern recognition, image analysis, signal processing, and discrete analysis.

The conference website `http://mmro.ru/en/`.

ISBN 978-5-907366-16-9

Оргкомитет

Председатель: Журавлев Юрий Иванович, *акад. РАН*, *ФИЦ ИУ РАН*

Заместитель: Чехович Юрий Викторович, *к.ф.-м.н.*

Борисова Татьяна Игоревна
Горнов Александр Юрьевич, *д.т.н.*
Громов Андрей Николаевич
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Мотренко Анастасия Петровна, *к.ф.-м.н.*
Петров Игорь Борисович, *член-корр. РАН*
Помазкова Евгения Владимировна
Рейер Иван Александрович, *к.т.н.*
Соколов Игорь Анатольевич, *акад. РАН*
Татарчук Александр Игоревич, *к.ф.-м.н.*
Чехович Юлия Викторовна
Шананин Александр Алексеевич, *чл.-корр. РАН*

Редактор: Грабовой Андрей Валериевич

Программный комитет

Сопредседатели: Рудаков Константин Владимирович, *акад. РАН*, *ФИЦ ИУ РАН*
Зорин Денис Николаевич, *проф.*, *CIMS NYU USA*

Ученый секретарь: Стрижов Вадим Викторович, *д.ф.-м.н.*

Воронцов Константин Вячеславович, *д.ф.-м.н.*
Гимади Эдуард Хайрутдинович, *д.ф.-м.н.*
Громова Ольга Алексеевна, *д.м.н.*
Двоенко Сергей Данилович, *д.ф.-м.н.*
Краснопрошин Виктор Владимирович, *д.т.н.*
Матвеев Иван Алексеевич, *д.т.н.*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.т.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Сойфер Виктор Александрович, *акад. РАН*
Чуличков Алексей Иванович, *д.ф.-м.н.*
Хачай Михаил Юрьевич, *д.ф.-м.н.*

Organizing Committee

Chair: Yury Zhuravlev, *acad. of RAS*, *FRCCSC*

Secretary: Yury Chekhovich, *C.Sc.*

Tatiana Borisova
Alexander Gornov, *D.Sc.*
Andrey Gromov
Andrey Inyakin, *C.Sc.*
Andrey Kokoshkin, *acad. of RAS*
Anastasiya Motrenko, *C.Sc.*
Evgenia Pomazko
Ivan Reyer, *C.Sc.*
Igor Sokolov, *acad. of RAS*
Alexander Tatrshuk, *D.Sc.*
Yulia Chekhovich
Alexander Shananin, *corr. member of RAS*

Editor: Andrey Grabovoy

Program Committee

Chair: Konstantin Rudakov, *acad. of RAS*, *FRCCSC*
Denis Zorin, *professor of computer CIMS NYU USA*

Secretary: Vadim Strijov, *D.Sc.*

Konstantin Vorontsov, *D.Sc.*
Edward Gimadi, *D.Sc.*
Olga Gromova, *D.Sc.*
Sergey Dvoenko, *D.Sc.*
Alexander Kel'manov, *D.Sc.*
Viktor Krasnoproshin *D.Sc.*
Ivan Matveev *D.Sc.*
Leonid Mestetskiy, *D.Sc.*
Vadim Mottl, *D.Sc.*
Genady Osipov, *D.Sc.*
Yury Pytiev, *D.Sc.*
Vladimir Ryazanov, *D.Sc.*
Viktor Soyfer, *acad. of RAS*
Alexey Chulichkov, *D.Sc.*
Michael Khachay, *D.Sc.*

Рецензенты

Адуенко А. А.
Анциперов В. Е.
Бахтеев О. Ю.
Бунакова В. Р.
Вальков А. С.
Ветров Д. П.
Визильтер Ю. В.
Владимирова М. Р.
Володин С. Е.
Воронцов К. В.
Гасников А. В.
Генрихов И. Е.
Гнеушев А. Н.
Голиков А. И.
Гончаров А. В.
Гороховский К. Ю.
Грабовой А. В.
Двоенко С. Д.
Дударенко М. А.
Дьяконов А. Г.
Жариков И. Н.
Животовский Н. К.
Загоруйко Н. Г.
Зайцев А. А.
Ивахненко А. А.
Игнатов А. Д.
Игнатов Д. И.
Игнатьев В. Ю.
Инякин А. С.
Исаченко Р. Г.
Ишкина Ш. Х.
Карасиков М. Е.
Каркищенко А. Н.
Катруца А. М.
Копылов А. В.
Кочедыков Д. А.
Кочетов Ю. А.
Красоткина О. В.
Крымова Е. А.
Кудинов М. С.
Кузнецов М. П.
Кузнецова М. В.
Кузьмин А. А.
Кулунчаков А. С.
Кушнир О. А.
Ланге М. М.
Ломов Н. А.
Лукашевич Н. В.
Майсурадзе А. И.
Максимов Ю. В.
Матвеев И. А.
Матросов М. П.
Местецкий Л. М.
Миркин Б. Г.
Михеева А. В.
Мнухин В. Б.
Мотренко А. П.
Мурашов Д. М.
Неделько В. М.
Нейчев Р. Г.
Новик В. П.
Одиноких Г. А.
Панов А. И.
Панов М. Е.
Потапенко А. А.
Пушняков А. С.
Рейер И. А.
Рудой Г. И.
Рябенко Е. А.
Сафонов И. В.
Сенько О. В.
Середин О. С.
Сотнезов Р. М.
Стенина М. М.
Стрижов В. В.
Сулимова В. В.
Талипов К. И.
Таханов Р. С.
Торшин И. Ю.
Трёкин А. Н.
Турдаков Д. Ю.
Федоряка Д. С.
Фрей А. И.
Хачай М. Ю.
Хританков А. С.
Царьков С. В.
Черепанов Е. В.
Чичева М. А.
Чуличков А. И.
Янина А. О.

Reviewers

Aduenko A.
Antsiperov V.
Bakhteev O.
Bunakova V.
Cherepanov E.
Chicheva M.
Chulichkov A.
Dudarenko M.
Dvoenko S.
D'yakonov A.
Fedoryaka D.
Frei A.
Gasnikov A.
Genrikhov I.
Gneushev A.
Golikov A.
Goncharov A.
Gorokhovskiy K.
Grabovoy A.
Ignat'ev V.
Ignatov A.
Ignatov D.
Inyakin A.
Isachenko R.
Ishkina Sh.
Ivakhnenko A.
Karasikov M.
Karkishchenko A.
Katrutsa A.
Khachay M.
Khritankov A.
Kochedykov D.
Kochetov Yu.
Kopylov A.
Krasotkina O.
Krymova E.
Kudinov M.
Kulunchakov A.
Kushnir O.
Kuz'min A.
Kuznetsov M.
Kuznetsova M.
Lange M.
Lomov N.
Lukashevich N.
Maksimov Yu.
Matrosov M.
Matveev I.
Maysuradze A.
Mestetskiy L.
Mikheeva A.
Mirkin B.
Mnukhin V.
Motrenko A.
Murashov D.
Nedel'ko V.
Nejchev R.
Novik V.
Odinokikh G.
Panov A.
Panov M.
Potapenko A.
Pushnyakov A.
Reyer I.
Rudoy G.
Ryabenko E.
Safonov I.
Sen'ko O.
Seredin O.
Sotnezov R.
Stenina M.
Strizhov V.
Sulimova V.
Takhanov R.
Talipov K.
Torshin I.
Trekin A.
Tsar'kov S.
Turdakov D.
Val'kov A.
Vetrov D.
Vizil'ter Yu.
Vladimirova M.
Volodin S.
Vorontsov K.
Yanina A.
Zagorujko N.
Zajtsev A.
Zharikov I.
Zhivotovskiy N.

Краткое оглавление

Brief contents

О наследуемости диагностических заключений при пополнении обучающей выборки новыми эмпирическими данными

Забежайло Михаил Иванович[1]⋆ zabezhailo@yandex.ru

[1]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В интеллектуальном анализе данных (ИАД), по-видимому, подавляющее большинство подходов реализуют так называемую *интерполяционно-экстраполяционную*(термин Ю.И. Журавлева) технику порождения эмпирических зависимостей (ЭЗ) при анализе обучающей выборки и классификации (диагностике) новых прецедентов. Однако, в ряде практически значимых приложений (в медицинской и технической диагностике, борьбе с мошенничеством в финансовой сфере, информационной безопасности и др.) дополнительно необходимо обеспечить *наследуемость (устойчивость)* формируемых интерполяций при пополнении обучающей выборки новыми данными. Важная причина этому – необходимость организовать, опираясь на результаты интерполяции, эффективные меры противодействия последствиям диагностируемых эффектов. Один из путей к искомой наследуемости – формирование интерполяций с выделением *причин* (факторов влияния), «вынуждающих» возникновение диагностируемого эффекта. В этом случае необходимо порождение таких интерполяционно-экстраполяционных (ИЭ) зависимостей, которые описывали бы причинные механизмы возникновения целевого эффекта на прецедентах «обучающей» выборки, обеспечивая сохранение корректности построенного описания – *«наследуемость»* этих ЭЗ, их способность адекватно описывать анализируемую причинность – при расширении «обучающей» выборки новыми данными.

Формируя результативные подходы здесь необходимо не только указать, как именно идентифицировать информацию о, собственно, *причинности* в описаниях прецедентов, но и учитывать, что используемые ИЭ-схемы анализа данных далеко не всегда позволяют «наследовать» уже сформированную интерполяционную «модель» (систему ЭЗ, . . .) при расширении обучающей выборки новыми данными.

В [1-3 и др.] предложен вариант математической техники ИАД, удовлетворяющей названным выше условиям. Речь идет о так называемых характеристических функций (ХФ), позволяющих описывать накапливаемые в обучающей выборке примеры (прецеденты, где идентифицирован анализируемый целевой эффект) и контрпримеры (прецеденты, где целевого эффекта нет) логическими условиями $ChF(\varphi)$, которые обращаются в истину на всех фактах φ, описываемых примерами, и ложны на всех фактах φ, описываемых контрпримерами текущей базы фактов **FB** (обучающей выборки прецедентов):

$$\forall\varphi\ [(\varphi{\in}\mathbf{FB}) \supset (ChF(\mathbf{FB})|{-}\ \varphi)]$$

Математическая техника ХФ позволяет формировать открытые теории, где все накопленные на текущий момент факты в **FB** могут быть описаны как логи-

ческие следствия множества образующих такую теорию утверждений. Алгоритмика порождения ХФ основана на анализе сходства описаний прецедентов (фактов в **FB**), формализуемого как алгебраическая операция. При этом постулируется, что сходства примеров (фактов наличия целевого эффекта) позволяют выделять «причинные влияния», приводящие к появлению целевого эффекта, а описания контрпримеров обязаны не содержать таких «причинных влияний» (ведь результатом «причинного влияния» должно быть наличие соответствующего эффекта, однако, на контрпримерах из текущей **FB** его нет). Проверка *неоспариваемости* формируемых заключений (в данном случае – $ChF(\varphi)$, порождаемых из **FB**) организована в рамках аргументационной схемы: сравниваются доводы ЗА (гипотезы о «причинных влияниях», отражающие сходства примеров в **FB**) и доводы ПРОТИВ («вложимость» таких гипотез в контрпримеры из **FB**).

Начатое в [1-3] обсуждение сложностных характеристик переборных задач, возникающих при формировании ХФ (в т.ч. - экспоненциально-быстро растущей «емкости» множества $ChF(\varphi)$ порождаемых на FB характеристических функций ChF; перечислительной полноты ряда переборных задач; полиномиальной разрешимости задачи о каузальной репрезентативности текущей **FB** – непустоте соответствующего множества $ChF(\varphi$ и др.) может быть продолжено анализом *наследуемости (устойчивости)* ЭЗ при расширении текущей FB описаниями $\Delta\mathbf{FB}$ новых примеров и контрпримеров.

Рассматриваются две актуальные задачи, возникающие при переходе от текущей базы фактов $\mathbf{FB_1}$ к ее расширению $\mathbf{FB_2} = \mathbf{FB_1} \cup \Delta\mathbf{FB}$:

1. Множество ЭЗ (частично-определенных ChF из $\boldsymbol{ChF}(\mathbf{FB_1})$, до-определяемых на $\Delta\mathbf{FB}$ при переходе от $\mathbf{FB_1}$ к $\mathbf{FB_2}$), общих для $\mathbf{FB_1}$ и $\mathbf{FB_2}$, всегда не пусто?

2. Можно ли надеяться на решение задачи о выделении ЭЗ, наследуемых при переходе от $\boldsymbol{ChF}(\mathbf{FB_1})$, до-определяемых на FB при переходе от $\mathbf{FB_1}$ к $\mathbf{FB_2}$, методом «грубой силы» (полным перебором элементов соответствующего множества $\boldsymbol{ChF}(\mathbf{FB_1})$)?

Отрицательный ответ в первом случае (приводятся соответствующие контрпримеры) дополнен примерами ситуаций, когда свойство *каузальной репрезентативности* $\mathbf{FB_1}$ при переходе к $\mathbf{FB_2}$ (т.е. непустота соответствующих множеств ХФ) наследуется, однако, при этом наследуемых из $\boldsymbol{ChF}(\mathbf{FB_1})$ в $\boldsymbol{ChF}(\mathbf{FB_2})$ эмпирических зависимостей (т.е. ChF, «достраиваемых» в $\boldsymbol{ChF}(\mathbf{FB_2})$ из $\boldsymbol{ChF}(\mathbf{FB_1})$ нет). Для второй задачи также получен отрицательный ответ: приводится пример множества ХФ, экспоненциально быстро растущего при линейном росте размеров исходной базы фактов. Вместе с тем показано, что непустота множества ЭЗ, наследуемых при переходе от $\mathbf{FB_1}$ к $\mathbf{FB_2}$, проверяется полиномиально быстро.

Работа поддержана проектом «Мат. основы интелл. анализа больших данных» ЦХАБД МГУ им. М.В.Ломоносова и Фонда НТИ (Договор № 7/1251/2019 от 15.08.2019).

[1] *Забежайло М.И.* О емкости семейств характ. функций, обеспеч. корректное решение задач диагност. типа // 19 Всероссийская конференция «Мат. методы распознавания образов» (ММРО-2019), 2019. С. 305–306.

[2] *Грушо А.А., Забежайло М.И., Тимонина Е.Е.* О каузальной репрезентативности обучающих выборок прецедентов в задачах диагностического типа. // Инфор. и ее применения., 2020. Т. 14. № 1. С. 80–86.

[3] *Забежайло М.И.* Об оценках размеров семейств характеристических функций, обеспечивающих корректное решение задач диагностического типа // ЖВМ и МФ., 2020. № 12.

To the heritability of diagnostic conclusions at extension of training sample by new empirical data

Michael Zabezhailo[1]⋆ m.zabezhailo@yandex.ru

[1]Moscow, FRCCSC of the Russian Academy of Sciences

Now to design empirical dependencies (ED) characterizing collected empirical data evidently the majority of modern computer data analysis tools are implementing so-called *interpolation-extrapolation* (IE) mathematical models (the term introduced by academician Yu.I.Zhuravlev) to analyze training sample (TS) and to diagnose new cases. However, basing on experience of some important applications (e.g. in medical or technical diagnostics, fraud protection in banking\finance, data security etc.) it's quite necessary to provide *heritability* of the designed interpolations when TS is extended by new empirical data (case descriptions). This heritability is critically important to form effective control activities preventing possible negative consequences of diagnosed effect\phenomenon. To design TS-interpolations basing on reliable description of *causal reasons* enforcing the effect under diagnosis\classification may be one of the resultative solutions.

It's natural to expect that extracted causal dependencies will be heritable ("stable", preserving correct representation of causal reasons) with respect to extensions of TS by new empirical data. E.g. it'll be strange to "treat" the same "disease" (TS-extension by additional empirical case descriptions) changing the "goals" of the counteraction\"treatment" every time receiving new empirical data about diagnosed effect\phenomenon.

So, exploring mathematical models, methods and algorithms to solve the discussed problems it's necessary not only to explicate how (by what type of data analysis tools) to identify causal reasons describing effect\phenomenon under diagnostics, but in addition to take into consideration that many IE-techniques, now popular in Machine Learning and Artificial Intelligence, aren't able to provide heritability of the designed TS-extrapolation when this TS is extended by new case descriptions. No less important to be able to avoid artefacts of causality analysis (e.g. overfitting etc.).

In [1-3] there were proposed one of the possible variants of mathematical technique providing intelligent data analysis following the above formulated requirements. This is a so-called Characteristic Function (ChF) formalism designed to describe accumulated in TS-cases (combining *examples* – precedents of diagnosed effect\phenomenon identification, and *counter-examples* – precedents where the absence of diagnosed effect\phenomenon is identified) by special functions of many-valued logics. Every $ChF(\varphi)$ has truth-value "*true*" on every example φ from current training sample **FB** and truth-value "*false*" for every counter-example φ from **FB**:

$$\forall\varphi\ [(\varphi\in\mathbf{FB}) \supset (ChF(\mathbf{FB}) \vdash \varphi)]$$

The proposed mathematical ChF-technique is a tool to form *open theory*, where collected in current **FB** facts\cases may be described as logical consequences of true

formulas\statements forming this theory. To design ChFs, it's supposed that analyzed effect\phenomenon is enforced by deterministic "causality": at least some enforcing "causality" factors are presented in descriptions of TS-cases (i.e. are representable by the used knowledge representation language). There are analyzed all significant combinations of such enforcing "causality" factors (corresponding to fix points of the Galois closure formed by the similarity of analyzed cases descriptions that is formalized as a binary algebraic operation). It is postulated that similarities of examples represent "causal influences" enforcing existence of diagnosed effect\phenomenon and, in addition, counter-examples must not "contain" these "causal influence" factors. (Indeed, causal reasons must "enforce" diagnosed effect\phenomenon existence, but by definition counter-examples are cases where the absence of the goal effect is identified). ChF's (i.e. calculated EDs) *falsifiability* checking is implemented by so-called argumentation scheme: arguments PRO (i.e. hypotheses about "causal influences" represented by similarities of examples) are compared to arguments CONTRA ("embeddability" of such hypotheses in counter-examples from current **FB**). Falsified arguments are rejected.

Starting to discuss (see [1-3]) computational complexity of *ChF*-related combinatorial problems (e.g. exponential speed of growth of the number of $\boldsymbol{ChF}(\mathbf{FB})$ elements; so-called $\#P$-completeness of some related combinatorial problems; polynomial complexity of the problem of current **FB** causal representativity – the set $\boldsymbol{ChF}(\mathbf{FB})$ non-emptiness - checking, etc.), it's natural to continue analyzing formed *ChFs heritability* when initial **FB** is extended by new data $\Delta\mathbf{FB}$ (examples and counter-examples).

Two actual algorithmic problems characterizing transition from $\mathbf{FB_1}$ (i.e. current TS) to its extension $\mathbf{FB_2} = \mathbf{FB_1} \cup \Delta\mathbf{FB}$ are under investigation:

1 Let $\boldsymbol{Com\_ChF}(\mathbf{FB_1},\mathbf{FB_2})$ be the set of common (heritable) EDs for $\mathbf{FB_1}$ and $\mathbf{FB_2}$ – partially-defined characteristic functions extended from $\mathbf{FB_1}$ to $\mathbf{FB_2}$ by clarification of their values on new cases from $\Delta\mathbf{FB}$. What about $\boldsymbol{Com\_ChF}(\mathbf{FB_1},\mathbf{FB_2})$ non-emptiness? Is it always non-empty?

2 Is it possible to extract heritable (from $\mathbf{FB_1}$ to $\mathbf{FB_2}$) empirical dependencies by "brute force" method (i.e. by inspection of all "candidates" - characteristic functions from $\boldsymbol{ChF}(\mathbf{FB_1})$)?

Negative solution for the first algorithmic problem (some counter-examples are demonstrated) may be added by examples of situations where causal representativity (non-emptiness of the corresponding set of ChFs) is heritable from $\mathbf{FB_1}$ to $\mathbf{FB_2}$, but the set $\boldsymbol{Com\_ChF}(\mathbf{FB_1},\mathbf{FB_2})=\varnothing$: there are no any "common" ChFs in $\boldsymbol{ChF}(\mathbf{FB_1})$ and $\boldsymbol{ChF}(\mathbf{FB_2})$.

For the second algorithmic problem negative answer is demonstrated too. There was designed an example of **FB** where the number of elements in the corresponding set $\boldsymbol{ChF}(\mathbf{FB})$ demonstrates exponential speed of growth with respect to linear growths of the number of cases in this **FB**. However, it is proved that non-emptiness of $\boldsymbol{Com\_ChF}(\mathbf{FB_1},\mathbf{FB_2})$ - the set of heritable ChFs - may be checked effectively

(i.e. there exist algorithm of polynomial complexity to solve this combinatorial problem).

[1] *Zabezhailo M.* To the Complexity of Characteristic Function Sets Providing Correct Diagnostic Solutions // Proc. 19 All-Russian Conf. "Mathematical Methods for Pattern Recognition" (MMPR-2019), 2019. Pp. 305–306.

[2] *Grusho A., Zabezhailo M., Timonina E.* On Causal Representativeness of Training Samples of Precedents in Diagnostic Type Tasks // Informatics and Applications, 2020. Vol.14. No 1. Pp. 80–86.

[3] *Zabezhailo M.* On the Complexity of Characteristic Function Sets Providing Correct Solutions for Diagnostic Type Tasks // Computational Mathematics and Mathematical Physics, 2020. No 12.

Выбор моделей и ансамблей

Стрижов Вадим Викторович[1,2]⋆ strijov@phystech.edu
Адуенко Александр Александрович[1] aduenko@phystech.edu
Бахтеев Олег Юрьевич[1] bakhteev@phystech.edu
Исаченко Роман Владимирович[1] isa-ro@yandex.ru
Грабовой Андрей Валериевич[1] grabovoy.av@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

Задача выбора модели машинного обучения не является корректно поставленной задачей по Адамару. Нет гарантии, что адекватная модель присутствует в семействе, из которого производится выбор, выбранная модель не единственна, более того, предполагается, что модель имеет избыточную сложность, и получаемое решение неустойчиво к изменениям. Для решения задачи выбора адекватной модели, необходимо выставить семейство гипотез поражения данных: предположение о стохастической природе моделируемого объекта, и назначить семейство моделей, из которых производится выбор. Оно не должно противоречить детерминированной природе моделируемого объекта. Модель задается суперпозицией функций-примитивов. Чтобы выбрать модель, необходимо указать на структуру суперпозиции, получаемую с помощью функций-примитивов. Чтобы сделать прогноз, необходимо назначить оптимальные параметры, согласно критерию качества модели. Модель машинного обучения выбирается оптимизацией четырех переменных. Это параметры модели и их распределение, структура модели и ее распределение. Изменения данных подразумевают изменения в структуре модели. Для сложного набора данных приходится иметь дело с ансамблем выбранных моделей. Предполагается, что ансамбль объединяет разные модели, которые адекватно описывают данные. В докладе будут обсуждаться принципы выбора модели для отдельной модели и для ансамбля моделей.

Работа выполнена при поддержке РФФИ (проекты 19-07-01155, 19-07-00875) и НТИ (проект 13/1251/2018).

[1] *Грабовой А. В., Стрижов В. В.* Выбор априорного распредления для смеси экспертов // Журнал вычислительной математики и математической физики, 2021. Т. 61. №5.

Selection of models and ensembles

Vadim Strijov[1,2]⋆ strijov@phystech.edu
Alexandr Aduenko[1] aduenko@phystech.edu
Oleg Bakhteev[1] bakhteev@phystech.edu
Roman Isachenko[1] isa-ro@yandex.ru
Andrey Grabovoy[1] grabovoy.av@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

According to Jacques Hadamard, the machine learning model selection is not a well-posed problem. All tree items hold: there is no guarantee an adequate model exists, the solution is not unique, the solution is not stable. To solve the model selection problem, one has to declare a family of data generation hypotheses: assumption about stochastic nature of the modelled object and a family of models: assumptions of its deterministic nature. A model is declared by a superposition of some elementary, or primitive, functions. To get a model, one shall point to a structure of superposition and a set of primitive functions. These two form a class of selected models. A model is a parametric family of functions. To make a forecast, one has to assign optimal parameters according to a criterion of optimality and following, a criterion of model quality. Four variables make a machine learning model selected: the model parameters and their distribution, the model structure and its distribution. Changes in data implies changes in the model structure. For a complex data set one has to deal with an ensemble of selected models. An ensemble is supposed to collect different models, which fit the data holistically. This talk will discuss principles of model selection for a single model and for an ensemble.

This research was supported by RFBR (projects 19-07-01155, 19-07-00875) and NTI (project 13/1251/2018).

[1] *Grabovoy A. V. Strijov V. V.* Prior distribution selection for a mixture of experts // Computational Mathematics and Mathematical Physics, 2021. Vol. 61. No 5.

О качественном поведении компонент разложений критериев качества решающих функций

Неделько Виктор Михайлович nedelko@math.nsc.ru
Институт математики им. С. Л. Соболева

При выборе оптимальной сложности метода построения решающих функций важным инструментом является разложение критерия качества на компоненты. Наиболее известно разложение на смещение и разброс (bias-variance decomposition).

Принято считать (и на практике это чаще всего выполняется), что с ростом сложности метода компонента смещения монотонно убывает, а компонента разброса — растёт.

Проведённое исследование показывает, что в некоторых случаях такое поведение нарушается. В работе [1] доказаны утверждения о том, что с ростом сложности смещение может расти, а разброс — уменьшаться.

На рисунке приведены зависимости смещения и разброса от глубины дерева решений (параметр сложности). Использована классическая реализация метода из sklearn (DecisionTreeRegressor). В качестве модели для генерации данных взята одномерная линейная зависимость, объём выборки равен 300. На левой диаграмме приведён результат при нулевом шуме, на правой – с нормальным шумом ($\sigma = 0{,}02$).

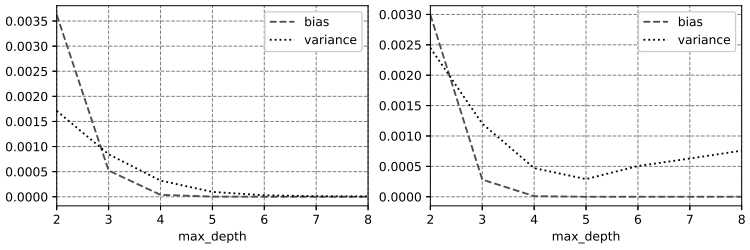


Рис. 1. Зависимости смещения и разброса от глубины дерева.

Видим, что в одном случае разброс везде уменьшается, а в другом — возрастает только с некоторого момента.

Данные факты свидетельствуют о том, что разложение на смещение и разброс, по-видимому, не может считаться вполне достаточным объяснением качественного поведения кривой обучения (зависимости качества от сложности).

Существует также другое разложение: на меру адекватности и меру устойчивости (Г. С. Лбов, Н. Г. Старцева, 1989). Компоненты данного разложения являются заведомо монотонными. Однако, данное разложение также имеет недо-

статки, в частности, есть сложность с определением меры адекватности для методов, содержащих регуляризацию по объёму выборки.

Актуальным представляется исследование вопроса, в какой мере и в каких случаях смещение может использоваться в качестве оценки меры адекватности. В рамках этого направления в работе исследуется сходство и различие упомянутых разложений.

Работа поддержана грантом РФФИ № 18-07-00600-а.

[1] *Nedel'ko V. M.* On Decompositions of Decision Function Quality Measure // Известия Иркутского государственного университета. Серия Математика, 2020. Т. 33. С. 64–79.

On the Shape of Components of Decompositions of Quality Criteria for Decision Functions

Victor Nedel'ko nedelko@math.nsc.ru
Sobolev Institute of Mathematics

When choosing the optimal complexity of the method for constructing decision functions, an important tool is to decompose the quality criterion into components. The bias-variance decomposition is the most well-known one.

It is generally assumed that with increasing complexity of the method, the bias component monotonically decreases, and the variance component — increases. The study shows that in some cases this behavior is violated. In [1] we have proved the statements that with increasing complexity, the bias can grow, and the variance can decrease.

These facts indicate that the bias-variance decomposition, apparently, can not be considered a sufficient explanation for the qualitative behavior of the learning curve (the dependence of quality on complexity).

Another approach (G.S. Lbov, N.G. Startseva, 1989) is a decomposition into a measure of adequacy and a measure of statistical stability (robustness). The idea of the approach is to decompose the prediction error into approximation error and statistical error.

It is relevant to study the question to what extent and in what cases bias can be used as an estimate of the adequacy measure. In this direction, we study the similarity and distinctions of the above-mentioned decompositions.

This research is funded by RFBR, grant 18-07-00600.

[1] *Nedel'ko V.* On Decompositions of Decision Function Quality Measure // The Bulletin of Irkutsk State University. Series Mathematics, 2020. Vol. 33. Pp. 64–79.

Исследование зависимости качества классификации от выбора частичных порядков на множествах значений признаков

Бакланова Анастасия Олеговна[1] baklanova_an@mail.ru
Дюкова Елена Всеволодовна[2] edjukova@mail.ru
Масляков Глеб Олегович[3]⋆ gleb-mas@mail.ru
[1]Москва, МГУ ВМК
[2]Москва, ВЦ ФИЦ ИУ РАН
[3]Москва, ООО «Яндекс»

Рассматривается одна из центральных задач машинного обучения — задача классификации на основе прецедентов. Для решения этой задачи успешно применяется аппарат логического анализа данных. Основным преимуществом похода является возможность получения результата при отсутствии дополнительных предположений вероятностного характера и при небольшом числе прецедентов.

Прикладные задачи не всегда могут быть описаны в рамках классической постановки логической классификации, когда отдельные значения целочисленного признака сравниваются с использованием отношения равенства. В [1] предложена схема синтеза корректных логических алгоритмов классификации при условии, что на множествах значений признаков заданы конечные частичные порядки; обобщены базовые понятия, используемые при построении классических моделей логических классификаторов, и приведены условия корректности построенных моделей общего вида.

В настоящей работе с целью улучшения качества классификации и повышения скорости обучения рассмотрена задача построения стохастических композиций классификаторов, предложенных в [1]. Следует отметить, что обучение логического классификатора над произведением частичных порядков по сути близко к обучению классического (корректного) решающего дерева. В том и другом случае происходит разбиение признакового пространства на прямоугольные области, содержащие объекты только одного класса. При этом решающее дерево строит своё разбиение «жадным» способом, а логический классификатор для построения разбиения решает сложную в вычислительном плане задачу дуализации над произведением частичных порядков. Проведено экспериментальное сравнение построенных на базе бустинга и бэгинга стохастических композиции логических классификаторов с аналогичными композициями решающих деревьев, успешно используемыми при решении практических задач, а именно, бэггингом над деревьями, случайным лесом и бустингом над деревьями.

При проведении экспериментов в исходные описания объектов были добавлены признаки с обратным отношением порядка. Такое преобразование признакового пространства гарантирует корректность рассматриваемой модели логи-

ческого классификатора, базирующейся на построении на этапе обучения «корректных элементарных классификаторов». Эксперименты на реальных задачах показали преимущество данной модели логического классификатора над бэггингом над деревьями и случайным лесом в случае небольшого числа прецедентов относительно числа признаков. В обратной ситуации лучшие результаты у случайного леса. Бустинг над деревьями ненамного превзошёл бустинг над элементарными классификаторами. Полученные результаты позволяют говорить о перспективности дальнейшего усовершенствования стохастических моделей логических классификаторов.

На практике отношение частичного порядка на множестве значений признака часто не задаётся. В рамках данного исследования с целью повышения качества классификации рассмотрена задача выбора на этапе предварительного анализа обучающей выборки «хорошего» частичного порядка и для её решения разработаны две процедуры, первая из которых основана на «независимом» выборе частичных порядков, вторая на «согласованном» выборе.

В первом случае для каждого класса K и для каждого признака x оценивается информативность каждого значения признака x, которое встречается в описаниях прецедентов из K. Считается, что информативность рассматриваемого значения тем выше, чем чаще это значение встречается в описаниях прецедентов класса K и чем реже оно встречается в описаниях прецедентов из других классов. На множестве значений признака x устанавливается линейный порядок согласно полученным оценкам. Предлагаемая процедура позволяет добиться улучшения качества классификации по сравнению с классическим подходом, однако имеет очевидный недостаток, который заключается в том, что частичный порядок на множестве значений признака подбирается независимо от выбора частичных порядков для других признаков.

В случае согласованного упорядочивания задача поиска частичного порядка сводится к оптимизации функционала эмпирического риска по множеству перестановочных матриц. Задача может быть решена методами градиентной оптимизации с применением усечённого оператора Синхорна. При этом в случае классификатора с дифференцируемой решающей функцией оптимизация осуществляется одновременно по параметрам решающей функции и по параметрам частичного порядка. В указанном случае описанный подход позволяет подбирать оптимальный с точки зрения функционала эмпирического риска частичный порядок совместно для всех признаков и, как показывают эксперименты на реальных данных, имеет преимущество перед независимым выбором частичного порядка. Однако данная процедура не может быть напрямую применена к алгоритмам классификации с недифференцируемой решающей функцией, к которым относятся и логические классификаторы. Поэтому в экспериментах использовался частичный порядок, предварительно полученный применением выбранной оптимизационной процедуры к линейной решающей функции. Результаты тестирования по качеству логической классификации двух рассмот-

ренных процедур выбора частичного порядка свидетельствуют о преимуществе независимого выбора перед согласованным выбором.

Работа частично финансирована грантом РФФИ № 19-01-00430.

[1] *Дюкова Е. В., Масляков Г. О., Прокофьев П. А.* О логическом анализе данных с частичными порядками в задаче классификации по прецедентам // Журнал вычислительной математики и математической физики, 2019. Т. 59. № 9. С. 1605–1616.

Investigation of the dependence of the supervised classification quality on the choice of partial orders on feature values sets

Anastasia Baklanova[1] baklanova_an@mail.ru
Elena Djukova[2]⋆ edjukova@mail.ru
Gleb Masliakov[3]⋆ gleb-mas@mail.ru
[1]Moscow, MSU CMC
[2]Moscow, CC FRC CSC RAS
[3]Moscow, LLC "Yandex"

We consider one of the central machine learning problems — the supervised classification problem. Methods of logical data analysis are successfully used to solve this problem. The main advantage of the approach is the possibility of obtaining a result without additional probabilistic assumptions and with a small number of precedents.

Applied problems can't always be described within the classical statement of logical classification in which integer feature values are compared using simple equality relation. In [1], a scheme for the synthesis of correct logical classification algorithms is proposed, provided that finite partial orders are defined on the sets of feature values; the basic concepts used in the construction of classical models of logical classifiers are generalized, and the conditions for the correctness of the constructed general type models are given.

In this paper, in order to improve the quality of classification and decrease the learning time, we consider the problem of constructing stochastic compositions of classifiers proposed in [1]. It should be noted that training a logical classifier over a product of partial orders is essentially close to training a classical (correct) decision tree. In both cases, the feature space is divided into rectangular areas containing objects of only one class. The difference is that the decision tree constructs its partition in a "greedy" way, and the logical classifier for constructing the partition solves the computationally complex dualization problem over the product of partial orders. An experimental comparison of stochastic compositions of logical classifiers built on the basis of boosting and bagging with similar compositions of decision trees is conducted. We mean such compositions as bagging over trees, random forest and boosting over trees. They are successfully used for solving practical problems.

During the experiments, features with the reverse orders were added to the original descriptions of objects. This transformation of the feature space guarantees the correctness of the logical classifier model, which is based on the construction of "correct elementary classifiers" at the training stage. Experiments on real tasks have shown the advantage of this model of logical classifier over bagging over trees and random forest in the case of a small number of precedents relative to the number of features. In the reverse situation, random forest has the best results. Boosting over trees is a little better than boosting over elementary classifiers. The obtained

results allow us to speak about the prospects for further improvement of stochastic models of logical classifiers.

In practice, the partial orders relation on the sets of feature values is often not specified. In this study, in order to improve the quality of classification, the problem of choosing a "good" partial order at the stage of preliminary analysis of the training set is considered. Two procedures are developed for its solution, the first of which is based on "independent" choice of partial orders, the second on "coordinated" choice.

In the first case, for each feature x the information content of each feature value that is found in the descriptions of the precedents from class K is estimated. This is done for each class K. It is considered that the informative content of the examined value is the higher in K the more often this value is found in the descriptions of precedents from class K and the less often it is found in the descriptions of precedents from other classes. Then according to the obtained estimates, the linear order is defined on the set of values of the feature x. The proposed procedure allows to improve the quality of classification in comparison with the classical approach, but it has an obvious drawback, which is that the partial order on the set of attribute values is chosen independently of the choice of partial orders for other features.

In the case of coordinated ordering, searching for partial orders is reduced to optimizing the empirical risk functional over the set of permutation matrices. The problem can be solved by methods of gradient optimization with the use of truncated Sinkhorn operator. In the case of a classifier with a differentiable decision function, optimization is performed simultaneously for the parameters of the decision function and for parameters of the partial order. In this case, the described approach allows to choice the optimal partial order from the point of view of the empirical risk functional collectively for all features, and, as experiments on real data show, it has an advantage over the independent choice of partial orders. However, this procedure can't be directly applied to classification algorithms with an undifferentiable decision function, which also include logical classifiers. Therefore, we used partial orders previously obtained in the experiments by applying the chosen optimization procedure to a linear decision function. The results of testing the quality of logical classification indicate the advantage of independent selection choice over the consistent choice.

This research is partially funded by RFBR, grant 19-01-00430.

[1] *Djukova E. V., Maslyakov G. O., Prokofyev P. A.* On the Logical Analysis of Partially Ordered Data in the Supervised Classification Problem // Computational Mathematics and Mathematical Physics, 2019. Vol. 59. No 9. Pp. 1542–1552.

Восстановление графов суперпозиций функций в задаче символьной регрессии

Нейчев Радослав Георгиев[1]⋆ neychev@phystech.edu
Шибаев Иннокентий Андреевич[1] shibaev.kesha@gmail.com
Стрижов Вадим Викторович[1 2] strijov@gmail.com
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Решается задача символьной регрессии для построения интерпретируемых моделей в задаче обучения с учителем. В рассматриваемом подходе дерево решения восстанавливается из предсказанной матрицы смежности для вычислительного графа модели. Рассматриваются несколько методов восстановления графов суперпозиций функций из их зашумленного матричного описания. Предлагаемый подход к восстановлению дерева решения основывается на $(2 - \varepsilon)$-аппроксимирующем алгоритме решения задачи PCST. Вычислительный эксперимент и сравнение различных подходов производится на синтетических данных.

Работа поддержана грантом РФФИ № 19-07-01155, № 19-07-00875 и НТИ проект № 13/1251/2018

Optimal superposition trees restoration in symbolic regression

Radoslav Neychev[1]⋆ neychev@phystech.edu
Innokenty Shibaev[1] shibaev.kesha@gmail.com
Vadim Strijov[1 2] strijov@gmail.com
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The symbolic regression problem is considered. Solutions to symbolic regression problems tend to be easier to interpret and more effective. The symbolic regression problem is solved via restoration of the superposition tree from the predicted adjacency matrix. The proposed approach is based on the search of minimum spanning tree in weighted coloured graph.The restoration procedure is sensitive to noise in the predicted matrix. This paper presents a novel approach based on Prize–Collecting Steiner Tree (PCST) algorithm. The proposed algorithm is compared with existing approaches to superposition trees restoration from noised adjacency matrix. Synthetic data is used in the experimental setup.

This research is funded by RFBR, grants 19-07-01155, 19-07-00875 and NTI project 13/1251/2018

Критерий одноклассовой классификации при наличии нетипичных объектов в обучающей выборке

Ларин Александр Олегович* ekzebox@gmail.com
Середин Олег Сергеевич oseredin@yandex.ru
Копылов Андрей Валериевич and.kopylov@gmail.com
Тула, Тульский государственный университет

Наиболее распространенным представителями граничных методов одноклассовой классификации являются метод одноклассового SVM (англ. One-class SVM), предложенный в работах Шолькопфа и Support Vector Data Description (SVDD, метод описания данных опорными векторами), предложенный Д. Тэксом и Р.Дьюином. Суть первого состоит в построении границы, обеспечивающей максимальный зазор между объектами обучающей совокупности и нулевой точкой признакового пространства. Использование метода потенциальных функций позволяет при этом строить границу заданной сложности. Моделью описания данных во втором методе является гиперсфера, представляющая собой ближайшую внешнюю границу вокруг целевого набора данных. Хотя позднее было показано, что при использовании радиальной базисной функции Гаусса (RBF) методы SVDD и One-class SVM эквивалентны, SVDD имеет интуитивно более понятную формулировку и геометрическую интерпретацию. Тем не менее, хотя оба метода используются для обнаружения аномалий в данных и нетипичных объектов, наличие таких объектов способно сильно смещать границу принятия решений. Главной проблемой описываемых выше формулировок задачи одноклассовой классификации является серьезное допущение в их математической постановке, связанное со стратегией штрафования нетипичных объектов. Исходная постановка задачи не является геометрически верной, т.к. величина штрафа допустимого выхода объектов обучающей выборки за пределы описывающей гиперсферы является несоизмеримой с расстоянием до её центра в оптимизационной задаче. Приведенные в нашей работе экспериментальные исследования показали, что это не существенно влияет на качество работы метода в случае отсутствия аномалий в обучающей выборке и позволяет сильно упростить решение исходной задачи, но в случае наличия нетипичных объектов в обучающей совокупности качество работы метода сильно ухудшается. В работе [1] предлагается новая версия критерия SVDD, позволяющая устранить геометрические несоответствия штрафов на выход объектов обучающей выборки за пределы описывающей гиперсферы и расстояние до её центра. Предложен критерий оптимизации и его эквивалентная форма. В результате удалось получить более устойчивое решающее правило по отношению к наличию аномальных объектов, чем при традиционных постановках, что подтверждается результатами модельных экспериментов. Для количественной оценки устойчивости классификаторов к наличию нетипичных объектов в обучающей выборке предложен метод, основанный на вычислении меры Жаккара между областями принятия

решения в пользу целевого класса при наличии и отсутствии нетипичных объектов.

Работа поддержана грантами РФФИ № 18-07-00942 и № 20-07-00441.

[1] *Ларин А. О., Середин О. С., Копылов А. В.* Модифицированный критерий для описания данных гиперсферой с учетом нетипичных объектов // Известия ТулГУ, Технические науки, 2020.

Criterion for one-class classification in the presence of outliers in the training set

Aleksandr Larin* ekzebox@gmail.com
Oleg Seredin oseredin@yandex.ru
Andrey Kopylov and.kopylov@gmail.com
Tula State University

A new version of one-class classification criterion robust to anomalies in the training dataset is proposed based on support vector data description (SVDD). The original formulation of the problem is not geometrically correct, since the value of the penalty for the admissible escape of the training sample objects outside the describing hypersphere is incommensurable with the distance to its center in the optimization problem and the presence of outliers can greatly affect the decision boundary. The proposed criterion is intended to eliminate this in-consistency. The equivalent form of criterion without constraints lets us use a kernel-based approach without transition to the dual form to make a flexible de-scription of the training dataset. The substitution of the non-differentiable objective function by the smooth one allows us to apply an algorithm of sequential optimizations to solve the problem. We apply the Jaccard measure for a quantitative assessment of the robustness of a decision rule to the presence of outliers. A comparative experimental study of existing one-class methods shows the superiority of the proposed criterion in anomaly detection.

This research is funded by RFBR, grants 18-07-00942 and 20-07-00441.

[1] *Larin A., Seredin O., Kopylov A.* Modified criterion for description of data with a hypersphere taking into account outliers // Izvestiya TulGU, 2020.

Снижение размерности в задаче декодирования временных рядов

Исаченко Роман Владимирович[1]⋆ roman.isachenko@phystech.edu
Стрижов Вадим Викторович[1 2] strijov@gmail.com
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Рработа посвящена задаче декодирования временных рядов в пространстве высокой размерности. Проблема заключается в избыточности исходного описания данных. Кроме того, рассматривается многомерный случай, целевая переменная является вектором. Компоненты целевой переменной являются зависимыми. Для устранения избыточной корреляции в признаковом описании объектов используются методы снижения размерности и выбора признаков.

Регрессия методом частных наименьших квадратов (PLS) используется в качестве базовой модели для снижения размерности пространства. Данная модель проецирует входные объекты и ответы в скрытое пространство и максимизирует ковариации между проекциями. Сочетание зависимостей входных объектов и ответов позволяет построить устойчивую модель.

Снижение размерности не поозволяет построить разреженную модель. Разреженность достигается путем выбора признаков. Большинство методов выбора признаков не используют зависимости в пространстве ответов. В работе предлагается новый подход к выбору признаков в случае многомерной регрессии. Для учета корреляций в матрице ответов предлагается обобщить идею алгоритма выбора признаков с помощью квадратичного программирования (QPFS). Алгоритм QPFS выбирает некоррелированные объекты, которые релеванты столбцам матрицы ответов. Предлагаемые методы накла- дывают веса на столбцы матрицы ответов. Идея состоит в том, чтобы оштрафовать коррелированные столбцы и уменьшить их влияние на выбор признаков.

Вычислительный эксперимент проводится на реальном наборе данных электрокортикограмм (ЭКОГ). Предложенные алгоритмы сравниваются по различным критериям, таким как стабильность и точность прогноза. Алгоритмы показывают результаты выше базового алгоритма. Сравнивается модель линейной регрессии с использованием QPFS алгоритма и модель регрессии частных наименьших квадратов. Наилучший результат достигается комбинацией алгоритмов QPFS и PLS.

[1] *Isachenko R. V. Strijov V V.* Quadratic Programming Feature Selection for Multicorrelated Signal Decoding with Partial Least Squares // Expert Systems with Applications, 2020.

Dimensionality reduction for time series decoding

Roman Isachenko[1]⋆ roman.isachenko@phystech.edu
Vadim Strijov[1 2] strijov@gmail.com
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The research investigates the problem of decoding for multi-dimensional time-series. The challenge to build the model is redundancy in initial data description. The features are highly correlated due to spatial nature of the data. It leads to redundant measurements and instability of the final model. To overcome this problem dimensionality reduction and feature selection methods are used.

The dimensionality reduction algorithms find the optimal combinations of the initial features and use these combinations as the model features. Partial least squares (PLS) is used as a base model. The algorithm projects the features and the targets onto the joint latent space and maximizes the covariances between projected vectors. It allows to save information about initial input and target matrices and find their relations. The dimensionality of latent space is much less than the size of initial data description. It leads to a stable linear model built on the small number of features. For this model we obtain the linear model with small latent dimension. However, the final model use the whole range of the initial features and it does not allow to remove useless features.

Feature selection is a special case of dimensionality reduction when the latent representation is a subset of initial data description. Here the model are built on the subset of the features. One of the approach to feature selection is to maximize feature relevances and minimize pairwise feature redundancy. Quadratic programmic feature selection (QPFS) uses this approach to construct the optimization problem.

The experiments were carried out in the ECoG data from the NeuroTycho project .We compared the proposed methods for multivariate feature selection with the baseline strategy and with PLS algorithm. The stability of the proposed methods were investigated. The proposed algorithms outperform the baseline algorithm with the same number of features. The combination of the feature selection procedure and the PLS algorithm gives the best performance.

[1] *Isachenko R. V. Strijov V V.* Quadratic Programming Feature Selection for Multicorrelated Signal Decoding with Partial Least Squares // Expert Systems with Applications, 2020.

Нижняя граница и избыточность вероятности ошибки классификации

Ланге Михаил Михайлович[1]⋆ lange_mm@ccas.ru
Парамонов Семен Владимирович[1] psvpobox@gmail.com
[1]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

В работе [1] введена функция «взаимная информация–вероятность ошибки» $R(\varepsilon)$ для схемы многоклассовой классификации объектов, связанных с множеством классов стохастическим каналом наблюдения. Эта функция не зависит от решающего алгоритма, монотонно убывает и дает нижнюю границу вероятности ошибки ε при средней взаимной информации R между множеством классифицируемых объектов и множеством решений об их классах. Для функции R_ε получена нижняя граница

$$R(\varepsilon) \geqslant \underline{R}(\varepsilon) = I(\mathbf{X};\Omega) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min})\ln(c-1)$$

в области значений $\varepsilon_{\min} \leqslant \varepsilon \leqslant \varepsilon_{\max}$, где $I(\mathbf{X};\Omega)$ — средняя взаимная информация между множеством объектов $\mathbf{X}$ и множеством классов $\Omega = \{\omega_i, i = 1,\ldots,c\}$, $c \geqslant 2$, $h(z) = -z\ln(z) - (1-z)\ln(1-z)$, $\underline{R}(\varepsilon_{\min}) = I(\mathbf{X};\Omega)$ и $\underline{R}(\varepsilon_{\max}) = 0$. При заданной энтропии $H(\Omega)$ множества Ω с априорными вероятностями $P(\omega_i)$, $i = 1,\ldots,c$, величина $\varepsilon_{\min}$ определяется условной энтропией $H(\Omega|\mathbf{X}) = H(\Omega) - I(\mathbf{X};\Omega)$, а $\varepsilon_{\max} = (c-1)\min_{i=1}^{c} P(\omega_i)$.

В настоящей работе демонстрируется возможность применения границы $\underline{R}(\varepsilon)$ для оценивания избыточности средней вероятности ошибки классификации по заданному набору разделяющих функций $G = \{g_j(\mathbf{x}),\ \mathbf{x} \in \mathbf{X}, j = 1,\ldots,c\}$. Такие функции дают правдоподобия решений $\omega_j \in \hat{\Omega}$, $j = 1,\ldots,c$ по предъявляемому объекту $\mathbf{x}$, причем решение считается ошибочным, когда $i \neq j$.

Набор G порождает условные распределения вероятностей решений

$$\left\{ Q(\omega_j|x) = \frac{g_j(\mathbf{x})}{\sum_{i=1}^{c} g_i(\mathbf{x})}, \quad j = 1,\ldots,c; \forall \mathbf{x} \in \mathbf{X} \right\},$$

которые дают среднюю взаимную информацию

$$I_G(\mathbf{X};\hat{\Omega}) = \sum_{\mathbf{x}\in\mathbf{X}} P(\mathbf{x}) \sum_{j=1}^{c} Q(\omega_j|\mathbf{x}) \ln \frac{Q(\omega_j|\mathbf{x})}{Q(\omega_j)}$$

и среднюю вероятность ошибки

$$E_G(\mathbf{X};\hat{\Omega}) = \sum_{\mathbf{x}\in\mathbf{X}} P(\mathbf{x}) \sum_{j=1}^{c} Q(\omega_j|\mathbf{x}) \left(1 - P(\omega_j|\mathbf{x})\right),$$

где $P(\mathbf{x}) = \sum_{i=1}^{c} P(\omega_i)P(\mathbf{x}|\omega_i)$ и $P(\mathbf{x}|\omega_i)$, $i = 1, \ldots, c$ — безусловные и условные по классам вероятности объектов, а $P(\omega_j|\mathbf{x})$, $j = 1, \ldots, c$ — апостериорные вероятности классов.

Величины $I_G = I_G(\mathbf{X}; \hat{\Omega})$ и $E_G = E_G(\mathbf{X}; \hat{\Omega})$ позволяют ввести избыточность $r_G = E_G - \varepsilon$ средней вероятности ошибки E_G относительно потенциально возможного значения ε при условии, что $\underline{R}(\varepsilon) = I_G$. Наименьшую избыточность обеспечивают байесовские разделяющие функции, когда $Q(\omega_j|\mathbf{x}) = P(\omega_j|\mathbf{x})$. В этом случае $\underline{R}(\varepsilon_{\min}) = I(\mathbf{X}; \Omega)$ и E_G наиболее близко к значению $\varepsilon_{\min}$. Байесовские разделяющие функции для блоков объектов позволяют достичь нулевой избыточности.

Предлагаемый подход допускает использование $L > 1$ слабых наборов разделяющих функций G_l, $l = 1, \ldots, L$, для которых $I_{G_l} < I(\mathbf{X}; \Omega)$. Такие наборы порождают композитный набор функций G в виде произведений

$$g_j(\mathbf{x}) = \prod_{l=1}^{L} g_{jl}^{s_l}(\mathbf{x}), \quad j = 1, \ldots, c$$

степеней слабых разделяющих функций при $s_l \geqslant 1$, $l = 1, \ldots, L$. Увеличение L должно привести к сближению характеристик I_G и E_G композитного набора с нижней границей $\underline{R}(\varepsilon_{\min})$ и, следовательно, к уменьшению избыточности r_G.

	Bayes	NTC	NTC2	CTC	
$p = 1$				$s_l = 1$	$s_l = -\ln E_{G_l}$
I_G	3.148	2.089	2.015	3.011	3.126
E_G	0.027	0.265	0.293	0.048	0.029
r_G	0.011	0.066	0.077	0.016	0.012
LBE	0.016	0.199	0.216	0.032	0.017
$p = 2$					
I_G	3.208	2.531	2.483	3.081	3.182
E_G	0.004	0.186	0.202	0.030	0.008
r_G	0.002	0.077	0.084	0.013	0.004
LBE	0.002	0.109	0.118	0.017	0.004

Выполнены эксперименты с разделяющими функциями экспоненциального вида $g_{jl}(\mathbf{x}) = \exp\left(-\nu_{jl} d^p(\mathbf{x}, \mathbf{x}_{jl})\right)$ с параметром $\nu_{jl} > 0$, где $d^p(\mathbf{x}, \mathbf{x}_{jl})$ — степень порядка $p \geqslant 1$ расстояния между объектом $\mathbf{x} \in \mathbf{X}$ и представителем $\mathbf{x}_{jl} \in \mathbf{X}_j$ j-го класса. Исследованы два слабых набора с эвристическими оценками $\hat{\nu}_{jl}$, $\hat{\mathbf{x}}_{jl}$, построенными на ближайших представителях классов NTC (Nearest Template-based Collection, $l = 1$) либо на вторых ближайших представителях классов NTC2 ($l = 2$) к предъявляемому объекту. Указанные оценки вычислены на обучающих выборках в режиме «leave-one-out». В композитных наборах CTC

(Composite Template-based Collection) использованы степенные коэффициенты $s_l = 1$ или $s_l = -\ln E_{G_l}$, $l = 1, 2$.

Разделяющие функции с параметром $p = 1$ и $p = 2$ протестированы на множестве изображений лиц от $c = 25$ персон, по 40 изображений в классах с одинаковыми априорными вероятностями. Полученные численные характеристики представлены в таблице. Значения нижней границы для вероятности ошибки указаны в строках LBE.

Полученные численные характеристики демонстрируют уменьшение избыточности на наборах CTC по сравнению с наборами NTC и NTC2, причем логарифмические коэффициенты дают характеристики, близкие к байесовским. Предложенный подход допускает использование параметрических разделяющих функций различного вида, когда априорное распределение классов и условные по классам распределения объектов неизвестны.

Работа частично поддержана грантами РФФИ № 18-07-01231 и № 18-07-01385.

[1] *Lange M., Lange A., Paramonov S.* On Data Classification Efficiency Based a Trade-off Relation between Mutual Information and Error Probability // IEEE Proceedings of the 6-th International Conference on Information Technology and Nanotechnology, ITNT-2020, 2020.

A lower bound and a redundancy of classification error probability

Mikhail Lange[1]⋆ lange_mm@ccas.ru
Semion Paramonov[1] psvpobox@gmail.com
[1]Moscow, Federal Research Center "Computer Science and Control" of RAS

In [1], for a multiclass object classification scheme with a given probabilistic observation channel, a "mutual information-error probability function" $R(\varepsilon)$ has been defined. The function $R(\varepsilon)$ is independent on any decision algorithm and due to monotonic decreasing it yields a lower bound to the error probability ε subject to a given value R of the average mutual information in a set of the objects relative to a set of the class-label decisions. For the function $R(\varepsilon)$, we have obtained the following lower bound

$$R(\varepsilon) \geqslant \underline{R}(\varepsilon) = I(\mathbf{X};\Omega) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min})\ln(c-1)$$

in a segment $\varepsilon_{\min} \leqslant \varepsilon \leqslant \varepsilon_{\max}$. Here, $I(\mathbf{X};\Omega)$ is the average mutual information between a set of the objects $\mathbf{X}$ and a set of the classes $\Omega = \{\omega_i, i = 1,\ldots,c\}$, $c \geqslant 2$; $h(z) = -z\ln(z) - (1-z)\ln(1-z)$; $\underline{R}(\varepsilon_{\min}) = I(\mathbf{X};\Omega)$ and $\underline{R}(\varepsilon_{\max}) = 0$. Given entropy $H(\Omega)$ of the set Ω with prior probabilities $P(\omega_i)$, $i = 1,\ldots,c$, the value of $\varepsilon_{\min}$ is determined by the conditional entropy $H(\Omega|\mathbf{X}) = H(\Omega) - I(\mathbf{X};\Omega)$ and $\varepsilon_{\max} = (c-1)\min_{i=1}^{c} P(\omega_i)$.

The present paper shows a possibility of using the bound $\underline{R}(\varepsilon)$ to calculate a redundancy of the average classification error probability that is yielded by a given collection of discriminant functions $G = \{g_j(\mathbf{x}),\ \mathbf{x} \in \mathbf{X}, j = 1,\ldots,c\}$. These functions give the likelihood values for the possible decisions $\omega_j \in \hat{\Omega}$, $j = 1,\ldots,c$ about a submitted object $\mathbf{x}$ and a decision is false when $i \neq j$.

The collection G produces the decision conditional probability distributions

$$\left\{ Q(\omega_j|x) = \frac{g_j(\mathbf{x})}{\sum_{i=1}^{c} g_i(\mathbf{x})}, \quad j = 1,\ldots,c; \forall \mathbf{x} \in \mathbf{X} \right\}.$$

Thus, these distributions yield the average mutual information

$$I_G(\mathbf{X};\hat{\Omega}) = \sum_{\mathbf{x}\in\mathbf{X}} P(\mathbf{x}) \sum_{j=1}^{c} Q(\omega_j|\mathbf{x}) \ln \frac{Q(\omega_j|\mathbf{x})}{Q(\omega_j)}$$

and the average error probability

$$E_G(\mathbf{X};\hat{\Omega}) = \sum_{\mathbf{x}\in\mathbf{X}} P(\mathbf{x}) \sum_{j=1}^{c} Q(\omega_j|\mathbf{x}) \left(1 - P(\omega_j|\mathbf{x})\right).$$

Here, $P(\mathbf{x}) = \sum_{i=1}^{c} P(\omega_i)P(\mathbf{x}|\omega_i)$ and $P(\mathbf{x}|\omega_i)$, $i = 1,\ldots,c$ are the unconditional and class-conditional probabilities of the objects, and $P(\omega_j|\mathbf{x})$, $j = 1,\ldots,c$ are the posterior probabilities of the classes.

The values $I_G = I_G(\mathbf{X}; \hat{\Omega})$ and $E_G = E_G(\mathbf{X}; \hat{\Omega})$ allow us to define a redundancy $r_G = E_G - \varepsilon$ of the average error probability E_G relative to the potentially possible value ε such that $\underline{R}(\varepsilon) = I_G$. For $Q(\omega_j|\mathbf{x}) = P(\omega_j|\mathbf{x})$, the Bayes discriminant functions provide the minimal redundancy. In this case, $\underline{R}(\varepsilon_{\min}) = I(\mathbf{X}; \Omega)$ and E_G is the most closed on $\varepsilon_{\min}$. Moreover, the Bayes discriminant functions for the blocks of the objects allow us to achieve the zero redundancy.

The proposed approach permits us to use $L > 1$ weak collections of the discriminant functions G_l, $l = 1, \ldots, L$ that satisfy the condition $I_{G_l} < I(\mathbf{X}; \Omega)$. So, the weak collections yield a composite collection of the functions as the products

$$g_j(\mathbf{x}) = \prod_{l=1}^{L} g_{jl}^{s_l}(\mathbf{x}), \quad j = 1, \ldots, c$$

taken with the power weights $s_l \geqslant 1$, $l = 1, \ldots, L$. An increase of L should make the composite characteristics I_G and E_G closer to the lower bound $\underline{R}(\varepsilon_{\min})$ providing a decrease of the redundancy r_G.

	Bayes	NTC	NTC2	CTC	
$p = 1$				$s_l = 1$	$s_l = -\ln E_{G_l}$
I_G	3.148	2.089	2.015	3.011	3.126
E_G	0.027	0.265	0.293	0.048	0.029
r_G	0.011	0.066	0.077	0.016	0.012
LBE	0.016	0.199	0.216	0.032	0.017
$p = 2$					
I_G	3.208	2.531	2.483	3.081	3.182
E_G	0.004	0.186	0.202	0.030	0.008
r_G	0.002	0.077	0.084	0.013	0.004
LBE	0.002	0.109	0.118	0.017	0.004

The simulated experiments with the exponential discriminant functions of the form $g_{jl}(\mathbf{x}) = \exp\left(-\nu_{jl} d^p(\mathbf{x}, \mathbf{x}_{jl})\right)$ have been performed. Here, $d^p(\mathbf{x}, \mathbf{x}_{jl})$ is a p-power distance between a submitted object $\mathbf{x} \in \mathbf{X}$ and the jth class template $\mathbf{x}_{jl} \in \mathbf{X}_j$, and $\nu_{jl} > 0$ is a parameter. Two weak collections of the functions with the heuristic estimates $\hat{\nu}_{jl}$, $\hat{\mathbf{x}}_{jl}$ have been evaluated. In these collections, the functions use the first or the second nearest templates in classes. So, these functions give NTC (Nearest Template-based Collection, $l = 1$) and NTC2 ($l = 2$), respectively. The templates and the appropriate estimates of the parameters are evaluated using the leave-one-out procedure. The composite collection CTC (Composite Template-based Collection) consists of the product functions with the power weights $s_l = 1$ or $s_l = -\ln E_{G_l}$, $l = 1, 2$.

The discriminant functions with the parameters $p = 1$ and $p = 2$ have been tested in the set of face images taken from $c = 25$ persons by 40 images in the classes of

the same prior probabilities. The numerical characteristics are given in the following table. The lower bound values of the error probability are shown in LBE lines.

The obtained numerical characteristics illustrate a decrease of the redundancy for the composite discriminant functions CTC as against the weak functions NTC and NTC2. Moreover, the logarithmic power weights yield the characteristics that are sufficiently closed to the Bayes values. The proposed approach permits us to use the different parametric discriminant functions when the class prior distribution and the object class-conditional distributions are unknown.

The research is partially supported by RFBR, grants 18-07-01231 and 18-07-01385.

[1] *Lange M., Lange A., Paramonov S.* On Data Classification Efficiency Based a Trade-off Relation between Mutual Information and Error Probability // IEEE Proceedings of the 6-th International Conference on Information Technology and Nanotechnology, ITNT-2020, 2020.

Ансамблевый кластерный анализ с использованием разнородного трансферного обучения

Бериков Владимир Борисович1[1]⋆ berikov@math.nsc.ru

[1]Новосибирск, Институт математики им. С. Л. Соболева СО РАН

В работе [1] предложен метод ансамблевого кластерного анализа, использующий трансферное обучение. Рассматривается задача кластерного анализа, в которой в дополнении к основному набору данных имеется дополнительный набор размеченных данных, в некоторм смысле "похожий" на основной. При этом наборы данных могут иметь различное признаковое описание. Метод основан на формировании мета-признаков, описывающих структурные особенности данных и их переносе с дополнительного набора на основной. В данной работе предлагается модификация алгоритма, основанная на использовании метода стохастического градиентного спуска при поиске зависимости между мета-признаками. Исследование алгоритма с помощью статистического моделирования подтвердило его эффективность. По сравнению с другими подобными методами, предложенный алгоритм позволяет анализировать наборы данных с различными признаковыми описаниями и является менее трудоемким.

Работа поддержана грантом РФФИ № 18-07-00600.

Ensemble Clustering with Heterogeneous Transfer Learning

Vladimir Berikov[1]⋆ berikov@math.nsc.ru

[1]Novosibirsk, Sobolev Institute of Mathematics SB RAS

The work [1] proposes an ensemble clustering method using transfer learning approach. We consider a clustering problem, in which in addition to data under consideration, "similar" labeled data are available. The datasets can be described with different features. The method is based on the constructing meta-features describing structural characteristics of data, and their transfer from source to target domain. In the given work, we propose a modification of the method using stochastic gradient descent to learn the dependence between meta-features. An experimental study of the method using Monte Carlo modeling has confirmed its efficiency. In comparison with other similar methods, the proposed one is able to work under arbitrary feature descriptions of source and target domains; it has smaller complexity.

This research is funded by RFBR grant 18-07-00600.

Нелинейный метод средних решающих правил с умными подвыборками для решения больших двухклассовых задач SVM-классификации

Макарова Александра Игоревна[1]⋆ aleksarova@gmail.com
Курбаков Михаил Юрьевич[1] muwsik@mail.ru
Сулимова Валентина Вячеславовна[1] vsulimova@yandex.ru
[1]Тула, Тульский государственный университет

Одним из популярных способов решения больших задач двухклассовой SVM-классификации является извлечение небольших подвыборок из полного обучающего множества для формирования более мелких и, соответственно, легче решаемых подзадач на каждом шаге некоторого итерационного процесса. Этот принцип лежит в основе большой группы методов стохастической аппроксимации, методов декомпозиции, последовательной минимальной оптимизации (SMO) и многих других.

В основе данной работы лежит нелинейный метод средних решающих правил (Kernel-based Mean Decision Rules Method, KMDR). Он заключается во взятии небольших случайных обучающих подвыборок, отдельном независимом обучении по каждой из них и последующем усреднении полученных частных решающих правил для получения окончательного. Этот метод также основан на указанном принципе, но, в отличие от многих других методов этой группы, он 1) позволяет строить нелинейные границы между классами и 2) не имеет зависимостей по данным между итерациями и, как следствие, обеспечивает возможность эффективного распараллеливания.

В данной статье предлагаются два новых подхода, совместное использование которых позволяет улучшить метод KMDR и быстрее найти лучшее (или не сильно отличающееся от лучшего) решение больших SVM задач, по сравнению с другими существующими реализациями SVM.

Первый предложенный подход - это новый (интеллектуальный) способ формирования подвыборок, который использует свойства SVM и KMDR, выбирая только те объекты, которые являются кандидатами на роль опорных, то есть такие объекты, которые находятся недалеко от разделяющей гиперплоскости и, следовательно, могут повлиять на результирующее решающее правило. А именно, предлагается формировать такие "умные"подвыборки только из опорных объектов, которые получаются в результате обучения на обычных небольших случайных подвыборках обучающей совокупности.

Высокая скорость формирования умных выборок в этом случае определяется относительно небольшим размером случайных подвыборок, SVM-обучение для которых может выполняться очень быстро. Более того, лучшее качество умных подвыборок по сравнению со случайными подвыборками в простом методе KMDR, обеспечивает более быструю сходимость метода, требует гораздо

меньшее число обучений на относительно больших выборках и, как результат, позволяет получить лучшее качество с меньшими временными затратами.

Кроме того, обучение по умным подвыборкам позволяет сократить время распознавания по сравнению с KMDR. Это объясняется тем, что KMDR имеет тенденцию к увеличению числа опорных объектов в итоговом решающем правиле с ростом числа случайных подвыборок. Фактически, число опорных объектов в KMDR соответствует размеру умной выборки. Но при обучении по умной выборке в большинстве практических случаев результирующее число опорных объектов существенно сокращается. Таким образом, поскольку этап распознавания требует сравнения новых объектов с каждым опорным объектом, применение умных выборок дает возможность сократить время распознавания.

Второй предлагаемый подход - это новая стратегия данных для ускорения произвольного доступа к большим наборам данных, хранящимся в традиционном формате libsvm. Предлагаемая стратегия основана на двух идеях: 1) осуществление предварительной разметки файла данных и 2) работа с файлом, отображенным в память процесса, вместо традиционной работы с файлом на диске.

Необходимость предварительной разметки обусловлена особенностью формата libsvm, затрудняющего произвольный доступ к объектам из-за невозможности вычисления начальной позиции объектов. Предварительная разметка предназначена для заполнения дополнительной структуры данных в памяти, которая для каждого объекта сохраняет позицию его начала в файле. Хотя для хранения разметки требуется дополнительная память, объем которой зависит от количества объектов, этот объем значительно меньше объема, необходимого для хранения всего обучающего набора.

Реализация второй идеи опирается на использование специальных функций операционной системы, позволяющих отобразить весь файл (или его часть) в область динамической памяти и вернуть указатель на эту область. В результате весь дальнейший доступ к данным осуществляется через указатели, что более удобно и быстро по сравнению с традиционной работой с дисковым файлом.

Таким образом, эта стратегия требует однократного последовательного чтения всего файла данных перед началом обучения, но ускоряет произвольный доступ к объектам.

Предлагаемый подход применим для любых методов, основанных на выборках, включая методы стохастического градиента.

Результирующий подход (Smart Sample KMDR) позволяет получить лучшее (или близкое к лучшему) решение больших двухклассовых задач SVM быстрее, чем существующие библиотеки, традиционно используемые для решения соответствующих задач. Но, тем не менее, у предложенного метода есть дополнительный резерв для повышения качества и скорости вычислений. Так, например, качество, как правило, можно повысить путем увеличения количества умных подвыборок, а время обучения можно дополнительно существенно

уменьшить, используя библиотеку liblinear (или другой быстрый инструмент) для формирования умных выборок. В свою очередь, это позволит охватить большее число объектов за меньшее время и получить более стабильное и качественное решение.

Работа выполнена при финансовой поддержке РФФИ, гранты 18-07-01087, 18-07-00942, 20-07-00055.

Исследования проведены с использованием оборудования Центра коллективного пользования сверхвысокопроизводительными вычислительными ресурсами МГУ имени М.В. Ломоносова.

[1] *Makarova A., Kurbakov M., Sulimova V.* Mean Decision Rules Method with Smart Sampling for Fast Large-Scale Binary SVM Classification // International Conference on Pattern Recognition — IEEE, 2020.

Smart Sample Kernel-based Mean Decision Rules Method for Big Binary SVM Classification Problems

Alexandra Makarova[1]⋆ aleksarova@gmail.com
Mikhail Kurbakov[1] muwsik@mail.ru
Valentina Sulimova[1] vsulimova@yandex.ru
[1]Tula, Tula State University

One of the popular ways to solve big binary SVM classification problems is to take small data samples of the entire training set to form a reduced subproblem at each step of some iteration process. This principle underlies a large group of stochastic approximation methods, decomposition methods, Sequential Minimal Optimization (SMO), and others.

This paper relies on the Kernel-based Mean Decision Rule method (KMDR) that consists in taking small random samples of the full dataset and separate training for each of them with consecutive averaging the respective particular decision rules to obtain a final one. This method is also based on sampling, but, unlike many other methods of this group, it 1) allows us to construct nonlinear boundaries between classes, and 2) has no data dependencies between iterations and, as a result, can be effectively parallelized.

This paper proposes two new approaches, which joint using allows to improve the KMDR and to obtain the best (or near the best) decision of large-scale binary SVM problems faster, compared to existing SVM solvers.

The first proposed approach is a new intellectual sampling technique, that exploits SVM and MDR properties to fast form so-called smart samples by selecting only those objects that are candidates to be the support ones, i.e. such objects that are near to a separate hyperplane and therefore can affect the resulting decision rule. Namely, it is proposed to form a "smart" sample only from support objects that are obtained as a result of training for small simple random samples.

The high speed of forming smart samples in this case is determined by the relatively small size of random samples, the SVM-training for which can be done very fast. Moreover, the better quality of smart samples, unlike random samples in simple KMDR, provides faster convergence of the method, requires much less training in relatively large samples and, as a result, makes it possible to obtain the best quality in less time.

In addition, training on smart samples allows for faster recognition times compared to KMDR. This is because KMDR tends to increase the number of support objects in the final decision rule as the number of random subsamples increases. In fact, the number of support objects in the KMDR corresponds to the smart sample size. But when training on a smart sample, in most practical cases, the resulting number of support objects is significantly reduced. Thus, since the recognition stage requires comparing new objects with each support object, the use of smart samples makes it possible to reduce the recognition time.

The second proposed approach is a new data strategy to accelerate random access to large datasets stored in the traditional libsvm format.

The proposed strategy is based on two ideas: 1) to pre-mark the data file, and 2) to work with the file mapped to the process memory instead of the traditional working with the file on the disk.

The pre-mark make random access to objects possible, which is difficult due to the libsvm format the peculiarity, because it does not allow to calculate the objects positions. The purpose of pre-mark is to store for each object its position in the file. Although additional memory is required for storage, which depends on the number of objects, this amount is significantly less than that required to store the entire training set.

The implementation of this idea relies on the use of special functions of the operating system that allows to map the entire file (or its part) into the dynamic memory area and return a pointer to this area. As a result, all further data access is performed through pointers, what is more convenient and fast compared to traditional working with a disk file.

So, this strategy requires a single sequential reading of the entire data file before starting the training but accelerates the random access to the objects.

The proposed approach is suitable for any sampling-based methods, including stochastic gradient methods.

The resulted Smart Sample KMDR approach allows to obtain the best (or near the best) decision of large-scale binary SVM problems faster compared to the existing SVM solvers. But nevertheless it has an additional reserve to increase the quality and speed of computations in contrast to reported values. So, the quality, as a rule, can be increased with increasing the number of smart samples and the training time can be additionally essentially decreased by using liblinear (or another fast tool) to form smart samples. In its turn, it will allow to cover more objects in less time and to obtain more stable and qualitative decision.

This research is funded by RFBR, grant 18-07-01087, 18-07-00942, 20-07-00055.

The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

[1] *Makarova A., Kurbakov M., Sulimova V.* Mean Decision Rules Method with Smart Sampling for Fast Large-Scale Binary SVM Classification // International Conference on Pattern Recognition — IEEE, 2020.

Прогнозирование на базе решения набора задач классификации с учителем и степеней принадлежности

Луканин Артем Александрович[1]⋆ lukanin@phystech.edu
Рязанов Владимир Васильевич[2] rvvccas@mail.ru
[1]Москва, Московский физико-технический институт
[2]Москва, ФИЦ "Информатика и управление" РАН

В данной работе решается задача восстановления зависимостей по выборкам прецедентов. Предлагается использовать степени принадлежности объектов каждому классу при распознавании в модели восстановления зависимостей «линейный корректор», основанной на решении набора задач классификации, сформированных по обучающей выборке, и последующей коррекции в пространстве значений целевого признака. В качестве классификаторов предлагаемой модели используются алгоритмы вычисления оценок с двумя различными функциями близости: метрической функцией и функцией близости для произвольных порядковых признаков. Производится сравнение работы предлагаемой модели с исходным методом и с известными методами анализа данных. Исследуется зависимость работы линейного корректора от его параметров. Предлагаемая версия линейного корректора применима к реальным прикладным задачам, что подтверждается экспериментально на примере решения задачи оценки параметров кристаллической решетки мелилитов.

Работа поддержана грантом РФФИ № 18-01-00557.

[1] *Lukanin A.*, *Ryazanov V.*, *Kiselyova N.* Prediction Based on the Solution of the Set of Classification Problems of Supervised Learning and Degrees of Membership // Pattern Recognition and Image Analysis, 2020. Vol. 30. Pp. 63–69.

Prediction based on the solution of the set of classification problems of supervised learning and degrees of membership

Artem Lukanin[1]⋆ lukanin@phystech.edu
Vladimir Ryazanov[2] rvvccas@mail.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS

The paper solves the problem of restoring dependences from samples of precedents. It is proposed to use the degrees of membership of objects to each class in the recognition process in the linear corrector model based on the solution of the set of classification problems generated using the training sample, and subsequent correction in the space of the target values. The classifiers of the proposed model are algorithms for calculating estimates with two different proximity functions: the metric function and the proximity function for arbitrary ordinal features. The work of the proposed model is compared with the original method and with the well-known data analysis methods. The dependence of the work of the linear corrector on its parameters is studied. The proposed version of the linear corrector is applicable to real applied problems, which is confirmed experimentally by the example of solving the problem of estimating the parameters of a melilite crystal lattice.

This research is funded by RFBR, grant 18-01-00557.

[1] *Lukanin A.*, *Ryazanov V.*, *Kiselyova N.* Prediction Based on the Solution of the Set of Classification Problems of Supervised Learning and Degrees of Membership // Pattern Recognition and Image Analysis, 2020. Vol. 30. Pp. 63–69.

Максимальные логические закономерности для построения решающих правил распознавания

Масич Игорь Сергеевич[1]⋆ i-masich@yandex.ru
Краева Екатерина Михайловна[1] em-kraeva@yandex.ru
[1]Красноярск, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева

Логические алгоритмы классификации или классификаторы, основанные на правилах, являются подходом к распознаванию, основанным на интеллектуальной обработке данных и отличающимся «прозрачным» классификатором, представляющим собой набор логических выражений [1, 2]. Основными задачами в рамках этого подхода являются нахождение информативных закономерностей на основе имеющихся данных и отбор наиболее полезных из них. Требуется найти такие правила (закономерности), которые выполняются для как можно большего числа наблюдений одного класса. Такие правила имеют большее подтверждение на имеющихся прецедентах и, предположительно, обладают большей обобщающей способностью. При этом желательно, чтобы закономерности не выполнялись для наблюдений других классов либо выполнялись для как можно меньшего их числа, то есть были наиболее однородными.

Существует множество критериев для оценки полезности или информативности закономерностей. Часто они сводятся к свертке двух описанных выше показателей (покрытию и однородности) [3]. Также, согласно [4], среди закономерностей, являющихся термами (конъюнкцией условий или литералов) можно выделить закономерности различных типов. Для этого на множестве закономерностей вводятся отношения частичного предпорядка – отношения простоты, избирательности и доказательности, которые основываются соответственно на множествах литералов (условий) закономерностей, объемах соответствующих булевых подкубов (множество содержащихся в подкубе точек) и множествах покрываемых наблюдений.

Закономерности, (локально) оптимальные по отношению доказательности, названы сильными. Предполагается, что закономерности с большим покрытием обладают высокой обобщающей способностью. Закономерность, (локально) оптимальная по отношению простоты, названа первичной. Такие закономерности состоят из наименее возможного числа литералов, при котором выполняется условие однородности.

Закономерности, максимальные одновременно по отношениям доказательности и простоты, называются сильными первичными. Закономерности, максимальные одновременно по отношениям доказательности и избирательности, называются сильными охватывающими.

Первичные закономерности более «объемные», их использование уменьшает число нераспознанных наблюдений. Охватывающие закономерности более избирательны, и их использование снижает ошибку распознавания. Более того,

с точки зрения интерпретируемости решающего правила, первичные закономерности являются более простыми и поэтому легко интерпретируемыми. Но охватывающие закономерности являются более «уверенными». Таким образом, если при построении классификатора обратить внимание на получение и использование закономерностей определенного типа, то можно добиться улучшения необходимых для конкретной ситуации показателей.

Но еще больше преимуществ от такого подхода можно получить при совместном использовании двух этих видов закономерностей. Предлагается выявлять и использовать логические закономерности попарно – сильную первичную и сильную охватывающую. Парные закономерности имеют одинаковое покрытие на обучающей выборке, то есть покрывают одни и те же наблюдения. Но при этом охватывающая закономерность состоит из большего числа литералов-условий.

Предлагаемый подход рассматривается для классификаторов, основанных на голосовании логических правил. Для построения классификатора решается серия задач поиска сильной α-закономерности, то есть терма, покрывающего наблюдение α класса K^+ и не покрывающего наблюдения другого класса K^-. Это задача псевдобулевой оптимизации [5]:

$$\sum_{\beta\in K^+}\prod_{\substack{i=1\\ \beta_i\neq\alpha_i}}^{n}(1-y_i)\to\max_{Y},\quad \sum_{\substack{i=1\\ \gamma_i\neq\alpha_i}}^{n} y_i\geqslant 1 \text{ для всех } \gamma\in K^-.$$

где $y_i=\begin{cases}1, & \text{если } i\text{-ый признак фиксирован в } P^\alpha,\\ 0, & \text{иначе.}\end{cases}$

Оптимальное решение данной задаче соответствует сильной закономерности, однако здесь присутствует неопределенность относительно того, является ли она первичной или охватывающей. Эти неоднозначности можно исключить путем введения дополнительных ограничений. Решение полученной задачи осуществляется с помощью алгоритма оптимизации, основанного на схеме ветвей и границ и жадного эвристического алгоритма, реализующего свойства рассматриваемого класса задач условной псевдобулевой оптимизации [6].

Работа выполнена в рамках государственного задания № FEFE-2020-0013 Минобрнауки России.

[1] *Дюкова Е. В., Журавлёв Ю. И., Прокофьев П. А.* Логические корректоры в задаче классификации по прецедентам // Журнал вычислительной математики и математической физики, 2017. Т. 57. № 11. С. 1906–1927.

[2] *Vorontsov K. V., Ivahnenko A. A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4-th Int'l Conf. on Pattern Recognition and Machine Intelligence, 2011. Pp. 66–73.

[3] *An A., Cercone N.* Rule Quality Measures for Rule Induction Systems: Description and Evaluation // Computational Intelligence, 2001. Vol. 17. No 3. Pp. 409–424.

[4] *Hammer P. L., Kogan A., Simeone B., Szedmak S.* Pareto-optimal patterns in logical analysis of data // Discrete Appl. Math., 2004. Vol. 144. Pp. 79–102.

[5] *Bonates T. O., Hammer P. L., Kogan A.* Maximum patterns in datasets // Discrete Appl. Math., 2008. Vol. 156. Pp. 846–861.

[6] *Kazakovtsev L. A., Masich I. S.* A branch-and-bound algorithm for a pseudo-boolean optimization problem with black-box functions // Facta Universitatis, Series Mathematics and Informatics, 2018. Vol. 33. No 2. Pp. 337–360.

Maximum logical patterns for constructing decision rules for recognition

Igor Masich[1]⋆ i-masich@yandex.ru
Ekaterina Kraeva[1] em-kraeva@yandex.ru
[1]Krasnoyarsk, Reshetnev Siberian State University of Science and Technology

Logical classification algorithms or rule-based classifiers are an intelligent data processing approach to recognition that features a "transparent" classifier, which is a set of logical expressions [1, 2]. The main problems within this approach are finding informative patterns based on the available data and selecting the most useful ones. It is required to find such rules (patterns) that are fulfilled for as many observations of the same class as possible. Such rules have more evidence on the existing precedents and, presumably, have more generalizing ability. In this case, it is desirable that the patterns are not fulfilled for observations of other classes, or they are fulfilled for the smallest possible number of them, that is, the patterns are the most homogeneous.

There are many criteria for assessing the usefulness or information content of patterns. They are often reduced to the convolution of the two above indicators (coverage and homogeneity) [3]. Also, according to [4], we can distinguish patterns of various types among the patterns that are terms (conjunction of conditions or literals). For this, partial preorder relations are introduced on a set of patterns: relations of simplicity, selectivity and evidence, which are based, respectively, on the sets of literals (conditions) of patterns, the volumes of the corresponding Boolean subcubes (the set of points contained in the subcube) and the sets of covered observations.

A pattern that is (locally) optimal with respect to evidence is called strong. It is assumed that patterns with large coverage have a high generalizing ability. A pattern that is (locally) optimal with respect to simplicity is called primary. Such patterns consist of the least possible number of literals for which the homogeneity condition is satisfied.

A pattern that is maximal at the same time in terms of evidence and simplicity is called strong primary. A pattern that is maximal at the same time in terms of evidence and selectivity is called strong spanned. These are the two types of patterns that are most useful. But at the same time, they have significant differences in terms of results obtained in the analysis of data and recognition.

Primary patterns are more "voluminous", their use reduces the number of unrecognized observations. Spanned patterns are more selective, and using them reduces recognition error. Moreover, in terms of interpretability of a decision rule, the primary patterns are simpler and therefore easily interpretable. But spanned patterns are more "confident". Thus, if we pay attention to obtaining and using patterns of a certain type when constructing a classifier, then we can improve the indicators necessary for a specific situation.

But even more benefits from this approach can be obtained by using these two types of patterns together. It is proposed to identify and use logical patterns in pairs: strong primary and strong spanned. Paired patterns have the same coverage on the training set, that is, they cover the same observations. But at the same time, the spanned pattern consists of a larger (if these patterns do not coincide) number of literals (conditions).

The proposed approach is considered for classifiers based on voting of logical rules. To construct a classifier, a series of problems is solved to reveal a strong α-pattern, that is, a term covering an observation α of the class K^+ and not covering observations of another class K^-. This is an optimization problem [5]:

$$\sum_{\beta\in K^+}\ \prod_{\substack{i=1\\ \beta_i\neq\alpha_i}}^{n}(1-y_i)\to\max_{Y},\quad \sum_{\substack{i=1\\ \gamma_i\neq\alpha_i}}^{n} y_i\geqslant 1 \text{ for all } \gamma\in K^-.$$

where $y_i=\begin{cases}1, & \text{if } i^{\text{th}} \text{ attribute is fixed in } P^\alpha,\\ 0, & \text{otherwise.}\end{cases}$

The optimal solution to this problem corresponds to a strong pattern, but there is uncertainty as to whether it is primary or spanned. These ambiguities can be eliminated by introducing additional constraints.

The obtained problem is solved using an optimization algorithm based on a branch-and-bound scheme and a greedy heuristic that implements the properties of the considered class of conditional pseudo-Boolean optimization problems [6].

Results were obtained in the framework of the state task FEFE-2020-0013 of the Ministry of Science and Higher Education of the Russian Federation.

[1] *Djukova E. V., Zhuravlev Y. I., Prokofjev P. A.* Logical correctors in the problem of classification by precedents // Computational Mathematics and Mathematical Physics, 2017. Vol. 57. No 11. Pp. 1866–1886.

[2] *Vorontsov K. V., Ivahnenko A. A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4-th Int'l Conf. on Pattern Recognition and Machine Intelligence, 2011. Pp. 66–73.

[3] *An A., Cercone N.* Rule Quality Measures for Rule Induction Systems: Description and Evaluation // Computational Intelligence, 2001. Vol. 17. No 3. Pp. 409–424.

[4] *Hammer P. L., Kogan A., Simeone B., Szedmak S.* Pareto-optimal patterns in logical analysis of data // Discrete Appl. Math., 2004. Vol. 144. Pp. 79–102.

[5] *Bonates T. O., Hammer P. L., Kogan A.* Maximum patterns in datasets // Discrete Appl. Math., 2008. Vol. 156. Pp. 846–861.

[6] *Kazakovtsev L. A., Masich I. S.* A branch-and-bound algorithm for a pseudo-boolean optimization problem with black-box functions // Facta Universitatis, Series Mathematics and Informatics, 2018. Vol. 33. No 2. Pp. 337–360.

Исследование методов сокращения опорной выборки при коллективном выводе с помощью нечетких логических систем

Полякова Анастасия Сергеевна[1]⋆ polyakova_nasty@mail.ru
Липинский Леонид Витальевич[1] lipinskiyl@mail.ru
Семенкин Евгений Станиславович[1] eugenesemenkin@yandex.ru
[1]Красноярск, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева

Одним из главных методов в процессах сокращения данных является метод отбора экземпляров (Instance Selection Method). Сокращение набора данных имеет две основных цели: уменьшение требований к вычислительным ресурсам, а также времени на обработку задачи обучения. Таким образом, выбор правильного подмножества примеров позволяет уменьшить размер выборки и повысить эффективность обучения. Но в отношении регрессионных наборов данных эта проблема изучена не полностью, в том числе из-за сложности этого типа набора данных.

Коллективный метод принятия решения на основе нечеткой логики показал свою эффективность при решении задач регрессии и классификации. При формировании коллектива и обучении агента может использоваться вся обучающая выборка. Однако использовать всю обучающую выборку для определения степени уверенности агента на тестовой точке – является вычислительно затратным. Формирование опорной выборки (опорного множества точек) позволит сократить эти затраты. В данной работе исследуется задача уменьшения размера опорного множества точек при коллективном принятии решений.

Под опорной выборкой подразумевается выборка, которая используется в коллективном выводе с помощью системы на нечеткой логике (т.е. опорная выборка представляет собой подвыборку обучающего множества). Нечеткий контроллер принимает решение о выборе агента для каждой точки из тестового множества. Для точки из тестового множества определяется ближайшая точка из опорной выборки. В зависимости от того, насколько эта точка близка к объекту из тестового множества и насколько успешно справляется на ней алгоритм, определяется его уверенность на данной тестовой точке. С помощью нечеткого контроллера итоговое решение принимает либо агент с наибольшей уверенностью, либо k – лучших агентов (k – параметр алгоритма) методом усреднения.

В рамках этой работы предлагается применить отбор экземпляров в опорное множество из обучающего при решении задач регрессии на основе таких методов как генетический алгоритм, алгоритм кластеризации k-средних, и случайный отбор экземпляров выборки. В большинстве существующих методах выбор экземпляров применялся к задачам классификации (дискретное прогнозирование), предлагаемый же подход используется для получения методов выбора экземпляров для задач регрессии.

Эффективность применения методов отбора примеров при формировании опорного множества в коллективном выводе на основе нечеткой логики была исследована на 3 известных наборах данных. В качестве критерия используется коэффициент корреляции согласованности. В результате проведенных исследований на тестовых задачах регрессии можно сделать выводы о том, что сокращение опорной выборки не приводит к снижению точности работы коллектива, а в некоторых случаях повышает ее.

Для формирования малых опорных выборок (количество примеров менее 500) генетический алгоритм оказывается лучше других подходов. При этом удается получить точность работы коллектива выше, чем на полной выборке. Это можно объяснить тем, что не всегда ближайшие точки к оцениваемому примеру являются показательными, например, из-за переобучения отдельных агентов или особенностей зависимости между входными параметрами и выходом.

Алгоритм k-средних на малых объемах опорной выборки может показывать хорошие результаты, однако при этом эксперименты показывают большой разброс точности решения коллектива. В первую очередь это объясняется высокой зависимостью эффективности алгоритма k-средних от генерации стартовых центров кластеров. При этом даже простой метод случайного отбора экземпляров выборки может оказаться не хуже, чем k-средних. С ростом объема опорной выборки снижается разница в эффективности представленных методов. На достаточно больших объемах (8000 и более) разница между данными методами практически отсутствует. При этом, k-средних и генетический алгоритм являются более затратными в вычислительном плане. Для таких объемов опорной выборки лучше использовать метод случайного отбора экземпляров выборки.

Вычислительные эксперименты показывают, что при эффективном выборе точек в опорное множество можно существенно снизить затраты вычислительных ресурсов при сохранении точности результата.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации в рамках базового госбюджетного финансирования по проекту № FEFE-2020-0013.

[1] *Polyakova A., Lipinskiy L., Semenkin E.* Investigation of Reference Sample Reduction Methods for Ensemble Output with Fuzzy Logic - Based Systems // 8th International Congress on Advanced Applied Informatics — IEEE, 2019. Pp. 583–586.

Investigation of Reference Sample Reduction Methods for Ensemble Output with Fuzzy Logic-Based Systems

Anastasiya Polyakova[1]⋆ polyakova_nasty@mail.ru
Leonid Lipinskiy[1] lipinskiyl@mail.ru
Eugene Semenkin[1] eugenesemenkin@yandex.ru
[1]Krasnoyarsk, Reshetnev Siberian State University of Science and Technology

One of the main methods in data reduction processes is the instance selection method. Reducing the dataset has two main objectives: reducing the requirements for computing resources, as well as the time for processing the learning task. Thus, the choice of the correct subset of examples allows reducing the sample size and increasing the efficiency of training. But with regards to regression datasets, this problem is not fully understood, including due to the complexity of this type of dataset.

The method collective decision making based on fuzzy logic has shown its effectiveness in solving regression and classification problems. During the ensemble formation and training of the agent the whole training sample can be used. However, using the entire training sample to determine the confidence level of the agent on the test point is computationally expensive. The formation of a reference sample will reduce these costs. In this paper, we study the problem of reducing the size of a reference set of points (reference sample) during collective decision making.

In the article, the reference sample refers to the sample that is used in ensemble output using of fuzzy logic system (i.e. the reference sample is a subsample of the training set). The fuzzy controller makes a decision on which agent should be used for each point from the test set. For a point from the test set, the nearest point from the reference sample is determined. Depending on the distance to the object from the test set and the successfulness of the algorithm on this object, the confidence of the algorithm on this test point is determined. With the help of fuzzy controller, the final decision is made either by the agent with the greatest confidence or by the k – best agents (k is the parameter of the algorithm) by averaging method.

As part of this work, it is proposed to apply the instance selection to select instances for the reference set from the training set when solving regression problems based on such methods as genetic algorithms (GA), the k-means clustering algorithm, and the random instance selection (RIS). In most existing methods, instance selection was applied to classification problem, and the proposed approach is used to perform instance selection for regression problem.

The effectiveness of the use of instances sample methods in forming a reference set in ensemble output using of fuzzy logic system has been explored on three known datasets. The concordance correlation coefficient is used as a criterion. As a result of the research on the test problems, it can be concluded that the reduction of the reference sample does not lead to a decrease in the accuracy of the ensemble

(etc. ensemble output with fuzzy logic-based systems) and in some cases increases it.

For the formation of small reference samples (the number of instances is less than 500), GA is better than other approaches. At the same time, it is possible to obtain the accuracy of the ensemble work higher than in the full sample. This can be explained by the fact that not always the nearest points to the estimated in-stance are indicative, for example, due to retraining of individual agents or the peculiarities of the relationship between the input parameters and the output.

The k-means algorithm can show not bad results on small volumes of the reference sample, but the experiments show a large spread of the accuracy of the team. This is primarily due to the high dependence of the k-means algorithm efficiency on the generation of cluster start centers. At the same time, even a simple RIS can be no worse than k-means. The difference in the efficiency of the presented methods decreases with the growth of the reference sample size. On sufficiently large volumes (8000 and more) the difference between these methods is practically absent. At the same time, k-means and GA are more computationally expensive. It is better to use RIS for such reference sample sizes.

Computational experiments show that effective instance selection in the reference set can significantly reduce the computational costs while maintaining the accuracy of the result.

Research is performed with the support of the Ministry of Education and Science of the Russian Federation within State Assignment project, grant #FEFE-2020-0013.

[1] *Polyakova A., Lipinskiy L., Semenkin E.* Investigation of Reference Sample Reduction Methods for Ensemble Output with Fuzzy Logic - Based Systems // 8th International Congress on Advanced Applied Informatics — IEEE, 2019. Pp. 583–586.

Аддитивная регуляризация для выбора структуры сетей глубокого обучения

Потанин Марк Станиславович[1]⋆ mark.potanin@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Рассматривается влияние способа построения функции ошибки на выбор структуры модели глубокого обучения. Требуется разработать метод, позволяющий понизить сложность модели при сохранении ее точности. Предлагается совместить аддитивную регуляризацию и прореживание сети генетическим алгоритмом для выбора оптимальной структуры. Регуляризация способствует повышению обобщающей способности модели и снижению риска переобучения.

Исследуется влияние весов регуляризаторов на сложность и точность модели. Веса регуляризаторов изменяются в ходе процедуры оптимизации структуры. Для этого составляется расписание оптимизации метапараметров аддитивной регуляризации. Предлагается совместить процесс оптимизации параметров нейросети с итерациями генетического алгоритма для оптимизации метапараетров регуляризации.

Для снижения структурной сложности нейросети использовался генетечиский алгоритм, который выступает как процедура прореживания нейросети. В вычислительном эксперименте определяются оптимальные значение метапараметров регуляризации, а так же исследуется зависимость точности, сложности и устойчивости модели от процедуры регуляризации, задаваемой метапараметрами. Результаты эксперимента показывают, что нейросеть с регуляризованными параметрами сходится к менее структурно сложному и качественно более точному решению, а так же имеет более высокую устойчивость по сравнению с моделью, которая не использует аддитивную регуляризацию.

Работа выполнена при поддержке РФФИ (проекты 19-07-01155, 19-07-00875) и НТИ (проект 13/1251/2018).

Additive regularization schedule for neural architecture search

Mark Potanin[1]⋆ mark.potanin@phystech.edu
Vadim Strijov[1] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The research paper considers the influence of the method of constructing the error function on the choice of the structure of the deep learning model. It is required to develop a method to reduce the complexity of the model while maintaining its accuracy. We propose to combine additive regularization and thinning by genetic algorithm to obtain optimal structure of neural networks. Regularization promotes increasing the generalizing ability of the model and reducing the risk of overfitting.

The research paper studies the influence of the weights of regularizes on the complexity and accuracy of the model. The metaparameters are changed in the course of the structure optimization procedure. This change is called the optimization schedule. It is proposed to combine the process of optimization of neural network parameters with iterations of a genetic algorithm to optimize regularization metaparameters.

To reduce the structural complexity of the neural network the GA-NAS was used. The GA-NAS in this case acts as a thinning procedure of a neural network. The goal of the computational experiment is to determine the optimal values of the regularization metaparameters, as well as to study the dependence of the accuracy, complexity and stability of the model on the regularization procedure specified by the metaparameters. Based on the experiment results, the regularized models work and are more accurate than the non-regularized models.

This research was supported by RFBR (projects 19-07-01155, 19-07-00875) and NTI (project 13/1251/2018).

О поиске частых элементов в небинарных данных на основе технологии CUDA

Генрихов Игорь Евгеньевич[1]⋆ ingvar1485@rambler.ru
Дюкова Елена Всеволодовна[2] edjukova@mail.ru
[1]Химки, ООО «Мобайл парк ИТ»
[2]Москва, ВЦ ФИЦ ИУ РАН

Задача поиска частых наборов в данных актуальна для многих прикладных областей. Одним из основных приложений является нахождение ассоциативных правил в базах данных (Agrawal R., Imielinski T., Swami A., 1993). Приведём классическую постановку рассматриваемой задачи для наиболее исследованного случая, а именно для случая бинарных данных.

Дано некоторое множество P, элементы которого называются атрибутами и дана база данных D, содержащая некоторые подмножества множества P, не обязательно различные. Подмножества множества P называются наборами атрибутов, а те из них, которые содержатся в D, называются транзакциями. Набор атрибутов Z называется *s-частым*, если отношение числа транзакций, содержащих Z, к числу всех транзакций не менее s. Требуется найти все s-частые наборы атрибутов.

В настоящее время существует большое число алгоритмов поиска частых наборов атрибутов в бинарных данных, среди которых одним из наиболее известных является алгоритм FP-Growth. Этот алгоритм основан на построении FP-дерева, которое фактически является структурированным представлением исходных данных.

В более общей постановке каждый атрибут имеет некоторое множество числовых значений и вместо наборов атрибутов рассматриваются наборы их значений. Как правило, поиск частых наборов значений атрибутов сводится к бинарному случаю путем задания для каждого атрибута некоторого числа (порога), позволяющего перекодировать исходные данные в бинарные.

В 2014 г. К. Эльбассиони поставил задачу поиска ассоциативных правил в данных, представленных в виде декартового произведения частичных порядков, и ввел понятие s-частого элемента для множества $P = P_1 \times \ldots \times P_n$, где $P_1, \ldots, P_n$ — конечные частично упорядоченные числовые множества. Считается, что элемент $y = (y_1, \ldots, y_n) \in P$ следует за элементом $x = (x_1, \ldots, x_n) \in P$ ($x \preceq y$), если y_i следует за x_i при $i = 1, \ldots, n$. Элементы $x, y \in P$ называются *сравнимыми*, если либо $x \preceq y$, либо $y \preceq x$. В противном случае x и y называются *несравнимыми*.

Предполагается, что каждое множество P_i имеет *наименьший элемент*, т. е. такой элемент l_i, для которого выполнено $l_i \preceq x_i$ для любого $x_i \in P_i$. Элемент $x_i \in P_i$ называется *существенным значением* элемента $x = (x_1, \ldots, x_i, \ldots, x_n) \in P$, если $x_i \neq l_i$. Предполагается также, что база данных D представлена в виде

некоторой совокупности элементов множества P (не обязательно различных) и не содержит транзакцию $l = (l_1, \ldots, l_n)$.

В случае бинарных данных $P_i = \{0, 1\}$, $i = 1, \ldots, n$, и в P_i установлен порядок $0 \preceq 1$, $0 \neq 1$. Здесь $l_i = 0$ и каждое P_i имеет всего один существенный элемент, равный 1.

Пусть $x \in P$, $x \neq l$. Число транзакций z в D таких, что $x \preceq z$ обозначается $S_D(x)$. Элемент x называется *s-частым*, если $S_D(x)/|D| \geqslant s$, иначе x называется *s-нечастым*. Заметим, что если $x \preceq y$, то $S_D(x) \geqslant S_D(y)$. Элемент x называется *максимальным s-частым* элементом, если $S_D(x) \geqslant s$ и $S_D(z) < s$ для любого z, такого что $x \preceq z$, $x \neq z$ (т.е. из условия $x \preceq z$, $x \neq z$ следует, что z — s-нечастый элемент в P).

В [1] предложена схема бинаризации множества $P = P_1 \times \ldots \times P_n$, согласно которой для каждого атрибута P_i, $i \in \{1, \ldots, n\}$, строится множество «значимых порогов» $Q_i \sqsubseteq P_i$. Числу p, $p \in P_i$, поставлен в соответствие элемент $\varphi(p) = (x_1, \ldots, x_n) \in P$, в котором $x_i = p$ и $x_j = l_i$ при $j \neq i$. Тогда число p называется *значимым порогом*, если $S_D(\varphi(p)) \geqslant s$. Предполагается, что каждое P_i имеет хотя бы один значимый порог. Набор порогов $H = \{p_1, \ldots, p_n\}$, в котором $p_i \in Q_i$ при $i = 1, \ldots, n$, порождает один из возможных вариантов перекодировки исходных данных в произведение бинарных частичных порядков P^H. Частые элементы множества P^H названы *пороговыми* частыми элементами. Поставлена задача перечисления максимальных пороговых s-частых элементов, порождаемых всеми возможными вариантами бинарной перекодировки (находить частые элементы, не являющиеся максимальными, неэффективно как по времени, так и по памяти). Для решения поставленной задачи построено полное пороговое FP-дерево (FTFP-дерево).

Конструкция FTFP-дерева является модификацией классического FP-дерева. В FTFP-дереве для каждого небинарного атрибута P_i строится полная вершина (P_i, Q_i). Пороги в Q_i сортируются по убыванию значения величины $S_D(\varphi(p))$, $p \in Q_i$. Сортировка важна для сокращения временных затрат. При спуске из полной вершины строится либо следующая полная вершина, либо корневая вершина классического бинарного FP-дерева. Бинарное FP-дерево строится тогда, когда в текущей ветви все значения небинарных атрибутов перекодированы в бинарные.

В реальных задачах может формироваться большое число наборов порогов перекодировки и для поиска максимальных пороговых частых элементов с использованием FTFP-дерева требуются существенные вычислительные ресурсы. В настоящей работе с целью ускорения процедуры поиска искомых частых элементов разработаны параллельные алгоритмы на основе технологии CUDA. Указанная технология позволяет выполнить операции, не требующие длительного времени, на центральном процессоре, а все сложные операции на графическом процессоре (GPU). Реализовано несколько схем распараллеливания: три блочных и одна одиночная. В блочной схеме множество всех «значимых» набо-

ров порогов H_D разбивается на непересекающиеся подмножества, каждое из которых подаётся на отдельный вычислительный блок GPU для синтеза максимальных пороговых частых элементов. При одиночном распараллеливании для нахождения максимальных пороговых частых элементов, порождаемых набором порогов из H_D, используется один вычислительный блок GPU. Приведены результаты тестирования построенных параллельных алгоритмов на модельных данных и на реальных задачах из репозитория UCI.

Работа частично финансирована грантом РФФИ № 19-01-00430-а.

[1] *Генрихов И. Е., Дюкова Е. В.* Поиск частых элементов произведения частичных порядков и ассоциативные правила // Сборник трудов VI Международной конференции и молодёжной школы «Информационные технологии и нанотехнологии» (ИТНТ-2020), 2020. Т. 4. С. 620–629.

About searching for frequent elements in nonbinary data based on CUDA technology

Igor Genrikhov[1]⋆ ingvar1485@rambler.ru
Elena Djukova[2] edjukova@mail.ru
[1]LLC ≪Mobile Park IT≫, Khimki, Russia
[2]CC FRC CSC RAS, Moscow, Russia

The problem of finding frequent sets in data is actual for many applications. One of the main applications is finding association rules in databases (Agrawal R., Imielinski T., Swami A., 1993). Here is the classical statement of the problem under consideration for the most studied case, namely for the case of binary data.

Let P be a set the elements of which are called attributes and D be a database containing some of certain subsets of P (not necessarily different). The subsets of P are called collections (sets) of attributes, and the subsets included in D are called transactions. The set of attributes Z is called *s-frequent* if the ratio of the number of transactions containing Z to the total number of transactions is not less than s. It is required to find all s-frequent sets of attributes.

Currently, there are a large number of algorithms for searching for frequent sets of attributes in binary data, among which one of the most well-known is the FP-Growth algorithm. This algorithm is based on the construction of an FP-tree which is actually a structured representation of the original database.

In a more general statement, each attribute can take a set of numerical values, and instead of sets of attributes, sets of their values are considered. As a rule, the search for frequent sets of values attributes is reduced to the simple binary case by specifying for each nonbinary attribute a certain number (threshold) for representing the original data by binary data.

In 2014 K. Elbassioni formulated the problem of finding association rules in data presented as a Cartesian product of partial orders, and introduced the concept of a s-frequent element for a set $P = P_1 \times \ldots \times P_n$, where $P_1, \ldots, P_n$ — finite partially ordered number sets. It is considered that the element $y = (y_1, \ldots, y_n) \in P$ followed by the element $x = (x_1, \ldots, x_n) \in P$ ($x \preceq y$) if y_i succeeds x_i for $i = 1, \ldots, n$. The elements $x, y \in P$ are said to be *comparable* if either $x \preceq y$ or $y \preceq x$. Otherwise, x and y are called *incomparable*.

We assume that each set P_i has the *least element*, i. e. an element l_i for which $l_i \preceq x_i$ for every $x_i \in P_i$. The element $x_i \in P_i$ is called a *significant value* of the element $x = (x_1, \ldots, x_i, \ldots, x_n) \in P$ if $x_i \neq l_i$. We also assume that the database D is represented by a collection of elements of the set P (not necessarily different) and does not contain the transaction $l = (l_1, \ldots, l_n)$.

In the case of binary data $P_i = \{0, 1\}$, $i = 1, \ldots, n$, and there is the order $0 \preceq 1$, $0 \neq 1$, in each P_i. Here $l_i = 0$ and each P_i has only one significant element equal to 1.

Let $x \in P$, $x \neq l$. The number of transactions z in D such that $x \preceq z$ is denoted by $S_D(x)$. The element x is said to be *s-frequent* if $S_D(x)/|D| \geqslant s$, otherwise is called *s-infrequent*. The element x is said to be the *maximum s-frequent* element if $S_D(x) \geqslant s$ and $S_D(z) < s$ for any z such that $x \preceq z$, $x \neq z$ (i. e. from the condition $x \preceq z$, $x \neq z$ it follows that z is an s-infrequent element in P).

In [1], a binarization scheme for the set $P = P_1 \times \ldots \times P_n$ is proposed, according to which a set of "significant thresholds" $Q_i \sqsubseteq P_i$ is constructed for each attribute P_i, $i \in \{1, \ldots, n\}$. The number p, $p \in P_i$, corresponds to the element $\varphi(p) = (x_1, \ldots, x_n) \in P$ in which $x_i = p$ and $x_j = l_i$ for $j \neq i$. Then the number p is called a *significant threshold* if $S_D(\varphi(p)) \geqslant s$. We assume that each P_i has at least one significant threshold. A set of thresholds $H = \{p_1, \ldots, p_n\}$, in which $p_i \in Q_i$ for $i = 1, \ldots, n$, generates one of the possible variants of encoding the original data into a product of binary partial orders P^H. Frequent elements of the set P^H are called *threshold* frequent elements. The problem is to enumerate the maximum threshold s-frequent elements generated by all possible variants of binary encoding (finding frequent elements that are not maximal is inefficient both in time and in memory). To solve this problem, a full threshold FP-tree (FTFP-tree) is constructed.

The FTFP-tree construction is a modification of the classical FP-tree construction. In the FTFP-tree, a full vertex (P_i, Q_i) is constructed for each nonbinary attribute P_i. Thresholds in Q_i are sorted in descending order of the value of $S_D(\varphi(p))$, $p \in Q_i$. Sorting is important to reduce time spent. When descending from a full vertex, either the next full vertex or the root vertex of a classical binary FP-tree is constructed. A binary FP-tree is constructed when all values of nonbinary attributes in the current branch are recoded to binary values.

In real problems, a large number of sets of encoding thresholds can be formed, and the search for maximum threshold frequent elements in the FTFP-tree requires significant computational resources. In this paper, parallel algorithms based on CUDA technology are developed to speed up the search for sought elements. This technology allows you to perform operations that do not require a long time on the central processor, and all complex operations on the graphics processor (GPU). Several parallelization schemes are implemented: three block and one single. In a block scheme, the set of all "significant" sets of thresholds H_D is divided into disjoint subsets, each of which is fed to a separate GPU computing unit for synthesizing maximum threshold frequent elements. In single parallelization, a single GPU computing unit is used for synthesizing the maximum threshold frequent elements generated by a set of thresholds from H_D. The results of testing the constructed parallel algorithms on model data and on real problems from the UCI repository are presented.

This research is partial financial supported of RFBR, grant 19-01-00430-a.

[1] *Genrikhov I. E., Djukova E. V.* Finding frequent elements for a product of partial orders and association rules // Proceedings of the VI International conference and youth school "Information technologies and nanotechnologies", 2020. Vol. 4. Pp. 620–629.

Задача обучения с экспертом для построение интерпретируемых моделей машинного обучения

Грабовой Андрей Валериевич[1]⋆ grabovoy.av@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Построение интерпретируемых моделей является одной из основных проблем в машинном обучении. Для повышения качества аппроксимации повышается сложность модели, из-за чего снижается ее интерпретируемость. Для повышения качества аппроксимации без повышения сложности модели предлагается использовать экспертную информацию о данных. Метод машинного обучения, основанный на экспертной информации, называется *обучением с экспертом*.

Данное исследование посвящено построению интерпретируемых моделей на основе экспертного априорного представления о решаемой задаче. Решается задача аппроксимации кривых второго порядка на контурном изображении. Предполагается, что кривая второго порядка описывается одной моделью. Экспертная информация позволяет отобразить точки изображения в новое пространство, где данная кривая описывается линейной моделью. При аппроксимации нескольких кривых на одном изображении строится мультимодель на основе смеси экспертов.

Решается задача аппроксимации радужки глаза. Глаз представляется как две концентрические окружности на изображении. Каждая окружность аппроксимируется одной линейной моделью. Изображение глаза является набором точек, которые требуется аппроксимировать. В вычислительном эксперименте анализируется качество аппроксимации контурного изображения при помощи предложенного метода. Проводится анализ качества аппроксимации радужки глаза.

Работа выполнена при поддержке РФФИ (проекты 19-07-01155, 19-07-00875) и НТИ (проект 13/1251/2018).

[1] *Грабовой А. В., Стрижов В. В.* Выбор априорного распредления для смеси экспертов // Журнал вычислительной математики и математической физики, 2021. Т. 61. №5.

Expert learning for interpretable model selection

Andrey Grabovoy[1]⋆ grabovoy.av@phystech.edu
Vadim Strijov[1] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The interpretable model selection is an important problem in machine learning. To improve the approximation quality one has to increase the number of model parameters. It results in a less interpretable model. The authors propose to use an expert prior information to improve the quality without increasing the complexity of the model. This information is called the expert information. The machine learning method based on an expert information is called expert learning.

This work analyzes the interpretable model selection. It solves the problem of approximating second-order curves on a given contour image. It is assumed that the second-order curve is described by one model, and the expert information maps the points of an image into a new space, where this curve is described by a linear model. A mixture of experts approximate to approximate several curves on one image.

A circle is an example of a second-order curve. The iris is represented as two concentric circles in an image. Each circle is approximated by one linear model.

The computational experiment analyses the proposed method for the contour image approximation. It uses synthetic and real data to test the proposed method. The real data is a human eye image from the iris detection problem.

This research was supported by RFBR (projects 19-07-01155, 19-07-00875) and NTI (project 13/1251/2018).

[1] *Grabovoy A. V. Strijov V. V.* Prior distribution selection for a mixture of experts // Computational Mathematics and Mathematical Physics, 2021. Vol. 61. No 5.

Анализ временных рядов с учетом критерия стационарности

Сенько Олег Валентинович[1,2] senkoov@mail.ru
Добролюбова Ольга Анатольевна[2]⋆ dbrl.olga@gmail.com>
[1]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук
[2]Москва, Московский государственный университет имени М.В.Ломоносова

Прогнозирование временных рядов является одной из самых исследуемых задач. При анализе временного ряда возникает множество вопросов: какие причинные связи присутствуют в данных, присутствует ли автокорреляция, имеет ли место наличие сезонных эффектов, тренда и т.д. Такое большое разнообразие свойств ряда требует глубокой аналитической работы и неизбежно ведет к использованию ряда упрощений. Соответственно необходимо использование критериев качества, которые позволят судить о правильности подобранной настройки для модели прогнозирования.

Традиционно для построения выводов о качестве прогноза анализируются дисперсия, среднее значение и различные однокомпонентные метрики. В общем виде задача прогнозирования состоит в выборе такого алгоритма, который обеспечивает максимальное качество прогноза относительно выбранной функции потерь. В зависимости от особенностей прогнозируемых данных исследователь отдает предпочтение той или иной метрике.

Одной из проверок качества спецификации модели является анализ стационарности остатков. Если остатки нестационарны, то модель имеет неодинаковую точность прогноза в разные периоды времени. Такая систематически разная ошибка говорит о том, что модель нуждается в корректировке. Проверка остатков в данном случае используется на уровне принятия и отклонения гипотезы о стационарности.

В работе предлагается способ улучшения прогноза путем добавления информации о стационарности остатков в функцию потерь. Способ основан на концепции коинтеграции Грейнджера, который широко используется в парном трейдинге и для выявления ложных регрессий. Использование концепции коинтеграции позволяет делать вывод о наличии статистически значимой связи между прогнозом и реальными значениями.

Новая функция потерь дополняется информацией о «степени» статистически значимой связи между прогнозом и реальными значениями. Для измерения «степени» стационарности ряда используется p-значение тестов на стационарность. То есть чем выше статистическая значимость утверждения о стационарности остатков, тем с большей уверенностью можно говорить о том, что модель настроена верно. В качестве тестов на стационарность используются тесты единичного корня Дики-Фуллера, Квятковского-Филлипса-Шмидта-Шина (КПСС) и тест Зивота и Эндрюса.

Релевантность использования сконструированной функции потерь проверялась на моделях регрессий и решающих деревьев. Выбор настроек модели прогнозирования основывался на выборе настройки с минимальным возможным p-значением теста на стационарность и минимальной среднеквадратичной ошибкой. Остатки плохо специфицированной модели могут иметь высокую «степень» стационарности, поэтому для получения хорошего прогноза необходимо контролировать и среднеквадратичную ошибку.

Проверка предложенного метода проведена на экономических рядах и сгенерированых псевдовыборках с нормальным распределением. Результаты исследования позволяют говорить о том, что использование информации о p-значении тестов на стационарность в функции потерь позволяет уменьшить ошибку прогноза в сравнении с классическим подходом к проблеме прогнозирования.

[1] *Кирилюк И. Л., Сенько О. В.* Выбор моделей оптимальной сложности методами Монте-Карло (на примере моделей производственных функций регионов Российской Федерации) // Информатика и её применения, 2020. Т. 14. № 2. С. 111–118.

Time series analysis with stationarity criterion

Oleg Senko[1,2] senkoov@mail.ru
Olga Dobroliubova[2]⋆ dbrl.olga@gmail.com
[1]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[2]Moscow, Moscow State University

Time-series forecasting is one of the most researched tasks. Analyzing a time series raises many questions: what causal relationships are present in the data, whether autocorrelation is present, seasonal effects, trends, etc. Such a wide variety of properties of a series requires deep analytical work and inevitably leads to many simplifications. Accordingly, it is necessary to use quality criteria that will make it possible to conclude the correctness of the selected setting for the forecasting model.

Traditionally, variance, mean, and various one-component metrics are analyzed to conclude forecast quality. In general, the forecasting problem consists of choosing an algorithm that provides the maximum forecast quality to the selected loss function. Depending on the features of the predicted data, the researcher gives preference to one or another metrics.

One of the quality checks of the model specification is to analyze the stationarity of the residuals. If the residuals are not stationary, then the model has unequal forecast accuracy in different periods of time. Such a systematically different error indicates that the model needs to be adjusted. In this case, checking the residuals is used at the acceptance and rejection of the stationarity hypothesis.

The work proposes improving the forecast by adding information about the residuals' stationarity to the loss function. The method is based on the concept of Granger cointegration, which is widely used in pair trading and to detect false regressions. Using the cointegration concept makes it possible to conclude a statistically significant relationship between the forecast and real values.

The new loss function is complemented by information about the "degree" of a statistically significant relationship between the forecast and real values. To measure the "degree" of stationarity is the p-value of stationarity tests is used. The higher the statistical significance of the statement about the residuals' stationarity, the more confident we can say that the model is configured correctly. As tests for stationarity, the unit root tests of Dickey-Fuller, Kwiatkowski–Phillips–Schmidt–Shin(KPSS), and the Zivot-Andrews test are used.

The relevance of using the constructed loss function was tested in regression models and decision trees. The choice of the prediction model settings was based on the setting's choice with the lowest possible stationarity test's p-value for stationarity and the lowest root means square error. The residuals of a poorly specified model can have a high "degree" of stationarity. Therefore, to obtain a good forecast, it is necessary to control the root mean square error.

The proposed method was tested on economic series and generated pseudo-samples with a normal distribution. The study results suggest that the use of information about the p-value of tests for stationarity in the loss function makes it possible to reduce the forecast error compared to the classical approach to the forecasting problem.

[1] *Kirilyuk I. and Senko O.* Selection of optimal complexity models by methods of non-parametric statistics (on the example of production function models of the regions of the Russian Federation // Informatics and applications, 2020. Vol. 14. No 2.

Выбор мультимоделей в задачах классификации и фильтрация выбросов

Адуенко Александр Александрович[1]⋆ aduenko@phystech.edu
Стрижов Вадим Викторович[1, 2] strijov@ccas.ru
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В работе рассматривается задача выбора мультимоделей при решении задач двухклассовой и многоклассовой классификации. Мультимодели позволяют учесть неоднородность зависимости целевой переменной от признаков без потери интерпретируемости одиночной модели. В работе рассматриваются многоуровневые модели и смеси моделей, построенные из обобщенно-линейных моделей. Мультимодель может содержать большое число близких моделей, что ведет к низкому качеству прогноза и отсутствию интерпретируемости. Для решения этой проблемы предлагается метод статистического сравнения моделей для прореживания мультимодели и построения адекватной мультимодели, все модели в которой попарно статистически различимы. Для статистического сравнения моделей вводится функция близости между апостериорными распределениями параметров моделей. Функция должна быть определена для пары распределений, которые могут быть заданы на разных признаковых пространствах, а также должна не различать два распределения, одно из которых является малоинформативным. Предложена функция близости для пары распределений, которая удовлетворяет этим требованиям, получены асимптотические свойства ее распределения в условиях истинности гипотезы о совпадении моделей.

Выборка может содержать шумовые объекты (выбросы), что также ведет к снижению качества прогноза. Ставится задача фильтрации выбросов. Эта задача сведена к задаче построения адекватной мультимодели. Для улучшения качества прогноза для малых классов при многоклассовой классификации предложена модификация метода статистического сравнения моделей. Вычислительный эксперимент на реальных и синтетических данных демонстрирует статистически значимое улучшение качества классификации и сокращение числа моделей в построенных мультимоделях.

[1] *Адуенко А. А., Мотренко А. П., Стрижов В. В.* Отбор объектов в кредитном скоринге с помощью ковариационной матрицы оценок параметров // Анналы исследования операций, 2020. Т. 260. № 1-2. С. 3–21.

Multimodel Selection for Classification and Outlier Filtering

Alexander Aduenko[1]⋆ aduenko@phystech.edu
Vadim Strijov[1, 2] strijov@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

In this article we consider the problem of a multimodel selection for two-class and multi-class classification problems. Multimodels allow accounting for inhomogeneity in relation between the target variable and features' values without losing interpretability inherent to a single model. In this article we consider multilevel models and mixtures of models consisting of generalized linear models. A multimodel may contain multiple similar models that leads to a worse classification quality and a loss of interpretability. A method for statistical models' comparison is introduced to prune a multimodel. The pruning leads to a construction of an adequate multimodel, i.e. the multimodel that has all the constituent models pairwise statistically distinguishable. A similarity function between the posterior distributions of models' parameters is used for model comparison. Such a similarity function must be defined for a pair of distributions in different features spaces, and must not distinguish the non-informative distribution from any other. We propose the similarity function that satisfies these conditions. The asymptotic properties of its distribution are obtained in case the models' true parameters are identical.

Outliers contained in a sample may also decrease the classification quality. The problem of outlier filtering is considered. This problem is reformulated as a problem of an adequate multimodel construction. To improve the prognosis quality for small classes in a multiclass classification, we propose a modification of the statistical models' comparison method. Computational experiments show statistically significant improvement in classification quality for real and synthetic datasets and substantial multimodel size reduction.

[1] *Aduenko A., Motrenko A., Strijov V.* Object selection in credit scoring using covariance matrix of parameters estimations // Annals of Operations Research, 2018. Vol. 260. No 1-2. Pp. 3–21.

Особенности группировки панельных данных на примере показателей, характеризующих экономическое развитие российских регионов

Кирилюк Игорь Леонидович[1]⋆ igokir@rambler.ru
Сенько Олег Валентинович[2] senkoov@mail.ru
[1]Москва, Институт экономики РАН
[2]Москва, Федеральный государственный центр "Информатика и управление" РАН

В различных экономических исследованиях возникает необходимость выявить структуру множества исследуемых объектов, выделив в них объективно существующие группы. Разбиения могут производится на основании какого-то одного информативного по мнению эксперта признака с использованием критериев, соответствующих выдвигаемой гипотезе.

Особое значение имеют ситуации, когда разбиение производится в соответствие с принципом максимального сходства объектов внутри групп с максимальными различиями между группами. При этом меры сходство (различия) между объектами вычисляются по всей совокупности признаков. Такого рода группы принято называть кластерами, а соответствующее направление интеллектуального анализа данных принято называть кластерным анализом.

Кластеризация может быть реализована различными способами, приводящими иногда к существенно различающимся результатам. Важную роль играет вопрос об объективности выявленной кластерной структуры. Поскольку кластерный анализ относится к методам многомерного статистического анализа, одним из способов доказательства объективности кластерной структуры является оценка её статистической значимости, которая может быть сформулирована в терминах проверки нулевых статистических гипотез. Нулевые гипотезы в кластерном анализе — это гипотезы, соответствующие интуитивному пониманию об «отсутствии кластерной структуры», которые могут быть по разному конкретизированы. Например, в этом качестве могут быть использованы нормально распределенные, или же равномерно распределенные случайные величины. В качестве статистик, на которых основывается проверка нулевых гипотез при верификации кластеризации, могут быть использованы известные индексы качества кластеризации, включая индексы Дана или Силуэта.

Задача обобщения методов кластерного анализа на случай, когда исследуемые объекты представлены не отдельными точками в пространствах признаков, а наборами временных рядов, составляя панельные данные, освещена в ряде обзоров. Однако, новые особенности, которые возникают при кластерном анализе сложных нестационарных временных рядов, требуют ещё существенного анализа.

В частности, для верификации результатов кластерного анализа при исследовании временных рядов в панельных данных возможный набор вариантов нулевых гипотез увеличивается из-за необходимости рассмотрения также гипо-

тез о порождении временных рядов белым шумом или процессом случайного блуждания. Такие нулевые гипотезы рассматривались авторами в работе [1] для верификации моделей Кобба-Дугласа. Для систем, где важную роль играют изменения во времени, применим как статический кластерный анализ по усреднённым во времени данным, так и динамический, описывающий возможные временные изменения кластерной структуры. Как показал практический опыт, задача о качестве и достоверности кластеризации может быть формализована с помощью введения большого набора индексов качества кластеризации, соответствующих различным аспектам интуитивного понимания того, что такое выраженность кластеризации.

Нами проведён анализ кластерной структуры данных российских регионов, характеризующих их социально-экономическое развитие с использованием разных альтернативных вариантов индексов, представленных, например, в пакете NbClust, написанном на языке R. Показаны существенные различия в поведении индексов для исследуемых данных. Оценена достоверность кластеризации с использованием нескольких альтернативных вариантов нулевых гипотез.

[1] *Kirilyuk I. Senko O.* Verification of the Returns to Scale of Production Type for the Russian Federation Regions // EPJ Web of Conferences 224, 2019. Pp. 1–6.

Peculiarities of grouping of panel data on the example of indicators characterising the economic development of Russian regions

Igor Kirilyuk[1]⋆ igokir@rambler.ru
Oleg Senko[2] senkoov@mail.ru
[1]Moscow, Institute of Economics of RAS
[2]Moscow, FRC "Informatics and Control" of RAS

In various economic studies, it becomes necessary to identify the structure of the set of objects under study, highlighting objectively existing groups in them. Partitions can be made on the basis of any one informative, in the opinion of the expert, index using criteria corresponding to the hypothesis put forward.

Of particular importance are situations when the division is made in accordance with the principle of maximum similarity of objects within groups and with maximum differences between groups. In this case, measures of similarity (differences) between objects are calculated for the entire set of features. Such groups are usually called clusters, and the corresponding direction of data mining is usually called cluster analysis.

Clustering can be implemented in a variety of ways, sometimes leading to significantly different results. An important role is played by the question of the objectivity of the identified cluster structure. Since cluster analysis refers to the methods of multivariate statistical analysis, one of the ways to prove the objectivity of the cluster structure is to assess its statistical significance, which can be formulated in terms of testing null statistical hypotheses. Null hypotheses in cluster analysis are hypotheses that correspond to the intuitive understanding of the "absence of a cluster structure", which can be specified in different ways. For example, normally distributed or uniformly distributed random variables can be used in this capacity. Known clustering quality indices, including Dunn or Silhouette indices, can be used as statistics on which the null hypothesis testing is based in clustering verification.

The problem of generalizing the methods of cluster analysis for the case when the objects under study are represented not by individual points in the feature spaces, but by sets of time series, composing panel data, is highlighted in a number of reviews. However, new features that arise in the cluster analysis of complex nonstationary time series still require significant analysis.

In particular, in order to verify the results of cluster analysis when studying time series and panel data, the possible set of variants of null hypotheses increases due to the need to consider also hypotheses about the generation of time series by white noise or a random walk process. Such null hypotheses were considered by the authors in [1] to verify the Cobb-Douglas models. For systems where changes in time play an important role, we can use both static cluster analysis based on time-averaged data, and dynamic, describing possible temporal changes in the clusters structure. As practical experience has shown, the problem of the validity of clustering can be

formalized by introducing a large set of clustering quality indices that correspond to various aspects of the intuitive understanding of what the validity of clustering is.

We have analyzed the cluster structure of data from Russian regions that characterize their socio-economic development using different alternative variants of indices, presented, for example, in the NbClust package written in the R language. Significant differences in the behavior of indices for the data under study are shown. The validity of clustering was estimated using several alternative variants of null hypotheses.

[1] *Kirilyuk I. Senko O.* Verification of the Returns to Scale of Production Type for the Russian Federation Regions // EPJ Web of Conferences 224, 2019. Pp. 1–6.

Подход к использованию содержательного контекста для построения и численной проверки гипотез о скрытых закономерностях в данных

Журавлёв Юрий Иванович[1] zhur@ccas.ru
Рязанов Владимир Васильевич[1] rvvccas@mail.ru
Сенько Олег Валентинович[1] senkoov@mail.ru
Докукин Александр Александрович[1] alex_dok@mail.ru
Виноградов Александр Петрович[1]⋆ vngrccas@mail.ru
Нелюбина Елена Андреевна[2] e.nelubina@gmail.com
Стефановский Дмитрий Владимирович[3] dstefanovskiy@gmail.com
[1]Москва, ФИЦ ИУ РАН
[2]Калининград, КГТУ
[3]Москва, РАНХиГС

В настоящее время в различных прикладных областях востребованны методы целенаправленного поиска и извлечения полезной информации из больших выборок данных, а также методы её дальнейшего анализа в различных аспектах. Мы исследуем и представляем здесь новый подход, который принадлежит к классу методов поиска скрытых закономерностей, допускающих параметрическое описание и многократно проявляющихся в числовых данных. Числовые закономерности могут наблюдаться при использовании различных типов представления и хранения первичных данных, не обязательно числовых. В данном случае основным методом перевода значимого запроса в числовую форму был подсчет повторений различных позиций в первичных записях, поэтому темой работы является построение и проверка параметрических гипотез о существовании полезных закономерностей именно такого вида. Центральным этапом для подхода в целом является построение цифровой модели потенциально присутствующих зависимостей между компонентами данных. Для точной формулировки гипотезы о наличии той или иной зависимости и её численной проверки использованы обобщенные прецеденты и аналоги преобразования Хафа в повышенных размерностях в качестве как концептуальной, так и вычислительной базы. Особое внимание при этом уделялось задействованию априорной информации и экспертного опыта в цифровых моделях, учитывались также некоторые аспекты перевода исходной информации в числа повторений (применение суперпозиции Хаф-подобных преобразований). Были проанализированы особенности взаимодействия различных факторов такого рода с целевыми параметрами анализа выборки, предложен ряд приёмов использования этих особенностей. Получены положительные результаты тестирования подхода на примере массива цифровых архивных записей большого объёма в розничной торговле, где обнаружены некоторые скрытые закономерности в поведении данных, которые могут быть полезны в аналитике и в практике ритейла. Ранее подход применялся

успешно также в других приложениях, что позволяет говорить о его перспективности и необходимости дальнейшей разработки.

Работа выполнена при частичной поддержке грантов РФФИ № 18-01-00557, 18-29-03151, 20-01-00609.

An approach to using meaningful context to construct and numerically test hypotheses about hidden regularities in data

Yury Zhouravlev[1] zhur@ccas.ru
Vladimir Ryazanov[1] rvvccas@mail.ru
Oleg Senko[1] senkoov@mail.ru
Aleksandr Dokukin[1] alex_dok@mail.ru
Aleksandr Vinogradov[1]⋆ vngrccas@mail.ru
Elena Nelyubina[2] e.nelubina@gmail.com
Dmitry Stefanovskiy[3] dstefanovskiy@gmail.com
[1]Moscow, FRC CSC RAS
[2]Kalinigrad, KSTU
[3]Moscow, RANEPA

Currently, methods of purposeful search and extraction of useful information from large samples of data, as well as methods of its further analysis in various aspects, are in demand in various application areas. We investigate and present here a new approach that belongs to the class of methods for searching for hidden regularities that allow a parametric description and are manifested many times in numerical data.

Numerical regularities can be observed using different types of representation and storage of primary data, not necessarily numerical. In this case, the main method of translating a meaningful query into a numerical form was the calculation of repetitions of various positions in primary records, so the topic of work is the construction and verification of parametric hypotheses about the existence of useful dependencies of this kind.

The central step for the approach as a whole is to build a digital model of potentially present dependencies between data components. For the exact formulation of the hypothesis about the presence of a particular dependence and its numerical verification, generalized precedents and analogues of the Hough transform in higher dimensions are used as both a conceptual and computational base. Special attention was paid to the use of a priori information and expertise in digital models, some aspects of translating the original information into repetition numbers were also taken into account (the use of superposition of Hough-type transforms). The features of the interaction of various factors of this kind with the target parameters of the sample analysis were analyzed, a set of techniques for using these features were proposed.

Positive results were obtained from testing the approach using the example of an array of digital archival records of a large volume in retail, where some hidden regularities in the behavior of data were found that can be useful in analytics and retail practice. Previously, the approach was also successfully applied in other applications, which approves its promise and the need for further development.

The work was done with partial support of RFBR grants 18-01-00557, 18-29-03151, 20-01-00609.

Спектральный ансамблевый кластерный анализ с использованием малоранговых представлений и нейросетевого автоэнкодера

Бериков Владимир Борисович[1] berikov@math.nsc.ru
[1]Новосибирск, Институт математики им. С.Л. Соболева СО РАН

В работе [1] был предложен алгоритм кластерного ансамбля, использующий комбинацию подходов, основанных на малоранговом представлении матрицы коассоциации, спектральной кластеризации и однослойном автокодирующем преобразованим. Пусть $X = \{x_1, \ldots, x_n\}$ - набор наблюдений, $x_i \in \mathcal{R}^d$. Цель анализа - разбить X на K подмножеств (кластеров), которые дают наилучшее значение некоторого критерия качества. В данной работе представлена модификация алгоритма с использованием глубокого автокодера (Deep Autoencoder, DA).

В машинном обучении широко используются различные методы извлечения признаков (например, метод главных компонент) для получения низкоразмерного представления данных. DA применяется для нелинейного преобразования данных с целью уменьшения размерности и выявления структур данных. DA - это искусственная нейронная сеть, преобразовывающая (кодирующая) входные данные X в некоторое представление X', так чтобы обратное преобразование (декодированные данные) реконструировало входные данные с максимальной точностью. Типичная архитектура автоэнкодера показана на Рис.1.

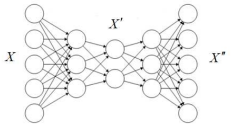


Рис. 1. Пример DA. Левый слой соответствует входу X; центральный слой представляет кодированные данные X'. Правый слой выводит декодированные данные X''.

Наилучшую структуру автокодировщика определить достаточно сложно, поэтому будем рассматривать набор архитектур; на основе каждого набора будем получать различные варианты кластеризации.

В процессе генерации кластерного ансамбля, конфигурации (с использованием разного количества слоев или нейронов в каждом слое) приводят к различным вариантам преобразованных данных. Каждый m-й вариант ($m = 1, \ldots, M$) многократно анализируется алгоритмом кластеризации, например k-средних с различным числом кластеров; при этом получаются базовые разбиения.

Ансамблевая кластеризация на основе матрицы коассоциации рассматривается как двухэтапный процесс. Сначала вычисляется усредненная матрица коассоциации $\mathbf{H}$ с элементами, указывающими на связь между x_i и x_j: $H(i,j) =$

$= \sum_{m=1}^{M} \alpha_m \sum_{l=1}^{L_m} \gamma_{m,l} h_{m,l}(i,j)$, где $\alpha_1, \ldots, \alpha_M$ - веса вариантов автоэнкодера, $\alpha_m \geqslant 0$, $\sum_m \alpha_m = 1$, $\gamma_{m,l} \geqslant 0$ - веса базовых вариантов кластеризации для m-го варианта архитектуры на l-м запуске, $\sum_l \gamma_{l,m} = 1$. Веса пропорциональны оценкам качества кластеризации.

На втором этапе формируется окончательное разбиение; элементы $\mathbf{H}$ рассматриваются как попарные расстояния (или меры сходства). В данной работе используется модификация алгоритма ансамблевой спектральной кластеризации, основанная на малоранговом представлении матрицы коассоциации. Алгоритм имеет близкое к линейному время работы и затраты памяти.

Алгоритм спектрального кластерного анализа (SC) известен своей эффективностью в обнаружении сложных структур данных. Существует несколько модификаций алгоритма SC. Рассматрим алгоритм следующего вида.

Сначала вычислим матрицу лапласиана $\mathbf{L} = \mathbf{D} - \mathbf{H}$, где $\mathbf{D} = \mathrm{diag}(D_{1,1}, \ldots, D_{nn})$, $D_{i,i} = \sum_j H(i,j)$. Затем найдем собственные вектора $v_1, ..., v_K$, которые соответствуют первым ненулевым наименьшим собственным значениям $\mathbf{L}$. Наконец, проведем кластеризацию с помощью алгоритма k-средних в пространстве признаков, определяемом собственными векторами.

Реализация SC достаточно трудоемка из-за сложности нахождения собственных векторов. Однако, как будет показано ниже, малоранговое представление матрицы $\mathbf{H}$ позволяет экономить время и память за счет использования эффективных матричных операций. Применим метод степенных итераций для вычисления собственных значений $\mathbf{L}$.

В этом методе, для любой $n \times n$ симметричной матрицы G и начального значения $v^{(0)}$, наибольшее собственное значение по абсолютному значению итеративно вычисляется как $v^{(k)} = \frac{G\, x^{(k-1)}}{||v^{(k-1)}||}, k = 1, 2, \ldots$ до достижения сходимости. Итерационный процесс сходится при некоторых условиях регулярности.

Представим усредненную матрицу коассоциации в виде $\mathbf{H} = \sum_{r=1}^{R} w_r H_r$ где H_l - матрица коассоциации, определенная для r-го разбиения, $R = M(L_1 + \cdots + L_M)$, w_r - соответствующий ему вес, $w_r = \alpha_\cdot \cdot \gamma_\cdot$. Через A_r обозначим матрицу ассоциации размерности $n \times K_r$ для r-го разбиения: $A_r(i,k) = I[c(x_i) = k]$, $i = 1, \ldots, n$, $k = 1, \ldots, K_r$, где K_r - количество кластеров в разбиении, $I[\cdot]$ - предикатная функция, $c(x_i)$ - метка кластера, назначенная x_i.

Усредненная матрица коассоциации может быть представлена в виде: $\mathbf{H} = BB^T, B = [B_1 B_2 \ldots B_r]$, где B - блочная матрица, $B_r = \sqrt{w_r} A_r$. Как правило, $m = \sum_r K_r \ll n$, что дает возможность экономить память за счет хранения $n \times m$ разреженной матрицы вместо полной $n \times n$ матрицы. Сложность умножения $\mathbf{H} \cdot x$ снижается с $O(n^2)$ до $O(nm)$.

Как легко увидеть, матрица $G = \mathbf{D} - \mathbf{H} = \mathbf{D} - BB^T$ представляется в малоранговой форме. Также известно, что можно свести задачу вычисления наименьших собственных значений симметричной положительно определенной матрицы к задаче нахождения ее наибольшего собственного значения.

Алгоритм был численно исследован с использованием биологических данных. Эксперименты показали преимущество предложенного подхода.

Работа поддержана грантом РФФИ № 19-29-01175.

[1] *Berikov V.* Autoencoder-based Low-Rank Spectral Ensemble Clustering of Biological Data // Cognitive Sciences, Genomics and Bioinformatics (CSGB) — IEEE, 2020. Pp. 43–46.

Low-Rank Spectral Ensemble Clustering Using Autoencoder Network

Vladimir Berikov[1] berikov@math.nsc.ru

[1]Novosibirsk, Sobolev Institute of Mathematics SB RAS

In the work [1], we have proposed a cluster ensemble algorithm using a combination of low-rank co-association matrix approach, spectral clustering, and shallow autoencoder transformation. Let $X = \{x_1, \ldots, x_n\}$ be a sample of observations, $x_i \in \mathcal{R}^d$. The purpose of the analysis is to partition X into K subsets (clusters) which give the best value for some criterion of clustering quality. In this work, we present a modification of the algorithm using Deep Autoencoder (DA) structure.

Different feature extraction techniques (such as Principal Components Analysis) are widely used in machine learning to get low-dimensional data representation. DA is applied for nonlinear transformation of data aimed at dimensionality reduction and revealing meaningful data representations. DA is an artificial neural network trained to transform (encode) input data X into some representation X' so that the inverse transformation (decoded data) reconstructs the input as accurately as possible. Typical autoencoder architecture is shown in Fig. 1.

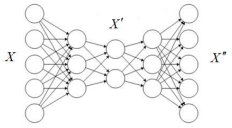


Figure 2. Example of DA. The left layer corresponds to input X; the central layer presents encoded data X'. The right layer outputs decoded data X''.

Because of the difficulty in finding the best autoencoder network structure, we generate a series of transformations and produce variants of clustering partitions based on the transformed data.

In the process of cluster ensemble generation, various configurations of the transformation procedure (using a different number of layers or neurons in each layer) yield a number of variants of the transformed data. Each mth variant ($m = 1, \ldots, M$) is an input for a clustering algorithm, k-means for example. The obtained base partitions are fed to a given consensus-finding procedure.

Co-association matrix-based ensemble clustering is considered as a two-stage process. Firstly, we calculate the averaged co-association matrix $\mathbf{H}$ with elements indicating the relationship between points x_i and x_j: $H(i,j) = \sum_{m=1}^{M} \alpha_m \sum_{l=1}^{L_m} \gamma_{m,l} h_{m,l}(i,j)$ where $\alpha_1, \ldots, \alpha_M$ are weights of the autoencoder variants, $\alpha_m \geqslant 0$, $\sum_m \alpha_m = 1$, $\gamma_{m,l} \geqslant 0$ are weights of base clustering variants for mth algorithm in its lth run, $\sum_l \gamma_{l,m} = 1$. The weights are proportional to clustering quality estimates.

The final consensus partition is obtained on the second stage. The elements of $\mathbf{H}$ are considered as pairwise distances (or similarity measures) between objects, and it is possible to use any distance- or similarity-based algorithm to partition the sample into a given number of clusters. In this work, we use ensemble spectral clustering algorithm based on the low-rank decomposition of the averaged co-association matrix, which has near-linear time and storage complexity.

Spectral clustering (SC) is known for its efficiency in discovering complex data structures. There are several modifications of SC algorithm. We consider SC in the following form.

We firstly calculate graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{H}$, where $\mathbf{D} = = \mathrm{diag}(D_{1,1}, \ldots, D_{nn})$, $D_{i,i} = \sum_j H(i,j)$. Then we find eigenvectors $v_1, ..., v_K$ which correspond to the first nonzero smallest eigenvalues of $\mathbf{L}$. Finally, we perform k-means clustering in the feature space defined by the eigenvectors.

The straightforward implementation of SC is rather time-consuming due to the eigensolver complexity. However, as it will be shown below, the low-rank form of $\mathbf{H}$ allows us to save time and memory by usage of efficient matrix operations. Namely, we apply power iteration method from numerical linear algebra to calculate eigenvalues of $\mathbf{L}$.

In this method, given any $n \times n$ symmetric matrix G and initial value $v^{(0)}$, the largest eigenvalue in absolute value is iterated as $v^{(k)} = \frac{G\, x^{(k-1)}}{||v^{(k-1)}||}, k = 1, 2, \ldots$ until convergence. It is known that the iteration process converges under some regularity assumptions.

Let us present the averaged co-association matrix in the form $\mathbf{H} = \sum\limits_{r=1}^{R} w_r H_r$ where H_l is co-association matrix defined for rth partition, $R = M(L_1 + ..., +L_M)$, w_r is its corresponding weight, $w_r = \alpha_{\cdot} \cdot \gamma_{\cdot}$. By matrix A_r of dimensionality $n \times K_r$ we denote cluster assignment matrix for rth partition: $A_r(i,k) = I[c(x_i) = = k]$, $i = 1, \ldots, n$, $k = 1, \ldots, K_r$, where K_r is the number of clusters in the partition, $I[\cdot]$ is a predicate function, $c(x_i)$ is a cluster number assigned to x_i. The averaged co-association matrix admits low-rank decomposition in the form: $\mathbf{H} = = BB^T, B = [B_1 B_2 \ldots B_r]$, where B is a block matrix, $B_r = \sqrt{w_r} A_r$. As a rule, $m = \sum_r K_r \ll n$, thus it gives one an opportunity of saving memory by storing $n \times m$ sparse matrix instead of full $n \times n$ co-association matrix. The complexity of matrix-vector multiplication $\mathbf{H} \cdot x$ is decreased from $O(n^2)$ to $O(nm)$.

It is easy to see that matrix $G = \mathbf{D} - \mathbf{H} = \mathbf{D} - BB^T$ has a low-rank form that decreases the cost of matrix-vector multiplication from quadratic to linear. It is also known that it is possible to reduce the problem of calculating the smallest eigenvalues of the symmetric PSD matrix to the problem of finding its largest eigenvalue.

The suggested algorithm has been numerically studied using biological datasets. The experiments have demonstrated the advantage of the proposed combination.

This research is funded by RFBR grant 19-29-01175.

[1] *Berikov V.* Autoencoder-based Low-Rank Spectral Ensemble Clustering of Biological Data // Cognitive Sciences, Genomics and Bioinformatics (CSGB) — IEEE, 2020. Pp. 43–46.

Поиск границ радужной оболочки при помощи сверточных нейронных сетей

Ефимов Юрий Сергеевич[1,2]⋆ yuri.efimov@phystech.edu
Соломатин Иван Андреевич[1,2] ivan.solomatin@phystech.edu
Одиноких Глеб Андреевич[2] g.odinokikh@gmail.com
[1]Долгопрудный, МФТИ
[2]Москва, Samsung R&D Institute Russia

Идентификация по радужной оболочке глаза — известная и широко используемая технология. Однако ее применение в мобильных устройствах ограничено ввиду алгоритмической сложности существующих методов, а также их малой универсальности относительно разнообразия условий регистрации. Предлагается метод выделения области радужки путем аппроксимации ее границ двумя окружностями, при этом относительная ошибка определения параметров окружностей должна составлять не более 5%. Применены свёрточные нейронные сети, оптимизированные по количеству параметров. Данный способ позволяет превзойти описанные в литературе классические (ненейросетевые) методы решения этой задачи сегментации. Работа метода проверена на базах изображений радужки из открытых источников.

Работа поддержана грантами РФФИ № 19-31-90171; 19-31-90167.

[1] *Ефимов Ю. С., Соломатин И. А., Одиноких Г. А.* Поиск границ радужной оболочки при помощи сверточных нейронных сетей // Известия РАН. Теория и системы управления, 2021. № 6. С. 89–98.

Finding the borders of the iris using convolutional neural networks

Yuriy Efimov[1,2]⋆ yuri.efimov@phystech.edu
Ivan Solomatin[1,2] ivan.solomatin@phystech.edu
Gleb Odinokikh[2] g.odinokikh@gmail.com
[1]Dolgoprudny, MIPT
[2]Moscow, Samsung R&D Institute Russia

Iris recognition is a well-known and widely used technology. Due to several reasons its application in modern smartphones is limited. The aforementioned reasons include high computational complexity of existing state of the art approaches and significant variety of eye image registration conditions. A method of iris area localization by approximating its borders with circles is proposed, which has relative circular approximation error less than 5%. The approach is based on applying convolutional neural networks (CNNs) and outperforms modern state of the art solutions, which do not use neural network backbone. The method is tested on several open source iris image databases.

This research is funded by RFBR, grants 19-31-90171; 19-31-90167.

[1] *Efimov Yu., Solomatin I., Odinokikh G.* Finding the borders of the iris using convolutional neural networks // Journal of Computer and Systems Sciences International, 2021. Vol. 59. No 6.

Оптимизация регрессионных нейросетевых моделей прямой оценки параметров объектов на изображениях модифицированными методами Adam

Денис Юрьевич Нарцев[1]⋆ dennartsev@gmail.com
Гнеушев Александр Николаевич[2] gneushev@ccas.ru
[1]Москва, Московский физико-технический институт
[2]Москва, Федеральный исследовательский центр "Информатика и управление" РАН

В области анализа изображений важной подзадачей является предварительная обработка, подготовка областей изображения объекта для устойчивого выделения признаков. На этапе предобработки так же решается задача фильтрации тех изображений, на которых уверенное распознавание не может быть достигнуто. Для решения этих задач предлагается подход прямой оценки параметров непосредственно путем решения задачи регрессии по изображению с помощью обучения нейросетевых моделей. В работе рассматривались задачи оценки степени размытия изображений глаз для фильтрации изображений низкого качества в системах идентификации личности по радужной оболочке глаза, а так же задача определения ориентации лица на изображении. Путем обучения и оценки точности предсказаний нейросетевой модели ResNet18 сравнивались различные методы оптимизации, такие как стохастический градиентный спуск с моментом (SGDM), Adam и его модификации, AdamW и RAdam. Обучающие выборки генерировались размытием гауссовским фильтром изображений глаз и поворотами изображений лица. Предложенная в работе модификация процедуры регуляризации алгоритма RAdam в рассматриваемых задачах позволила уменьшить ошибку более чем в 1.5 раза по сравнению с моделью, обученной исходным алгоритмом. Тестовая и обучающая выборка для задачи оценки были сформированы на основе баз изображений глаз BATH и CASSIA.

Работа поддержана грантом РФФИ № 19-07-01231.

[1] *Нарцев Д. Ю., Гнеушев А. Н.* Сравнение модифицированных методов обучения Adam в задачах оценки параметров регрессионных моделей по изображению // Информационные технологии, 2021.

Optimization of regression neural network models for direct estimation of object parameters in images by modified Adam methods

Denis Nartsev[1] ⋆ dennartsev@gmail.com
Alexander Gneushev[2] gneushev@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS

In the field of image analysis, an important subtask is preprocessing, preparation of image areas of an object for stable feature extraction. The task of filtering those images on which confident recognition cannot be achieved is also solving at the stage of preprocessing. To solve these problems, direct estimation of parameters directly by solving the image regression problem by training neural network models is proposed. The paper considered the tasks of evaluating the parameter of blurring of eye images for filtering low quality images in personality identification systems by the iris of the eye, as well as the task of determining the orientation of a face in an image. Various optimization methods, such as stochastic gradient descent with momentum (SGDM), Adam and its modifications, AdamW and RAdam, were compared by training and evaluating the prediction accuracy of the ResNet18 neural network model. The training samples were generated by blurring the Gaussian filter of the eye images and rotating the face images. The modification of the regularization procedure of the RAdam algorithm proposed in the work reduced the error by more than 1.5 times in comparison with the model trained by the original algorithm. The test and training sample for the assessment problem was formed on the basis of the BATH and CASSIA eye image databases.

This research is funded by RFBR, grant 19-07-01231

[1] *Nartsev D. Yu., Gneushev A. N.* Adam optimization method modifications comparison in the regression models parameters evaluation tasks // Information Technology, 2021.

Построение нейросетевого классификатора в пространстве дескрипторов преобразования Радона для эффективного детектирования пешеходов

Самсонов Никита Андреевич[1]⋆ nikita.samsonov@phystech.edu
Гнеушев Александр Николаевич[1] gneushev@ccas.ru
Матвеев Иван Алексеевич[1] matveev@ccas.ru

[1]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

Задача детектирования пешеходов на изображении решается на основе построения интегральных дескрипторов объекта и обучения на них бинарного классификатора. В работе для построения интегрального дескриптора объекта используется семейство локальных дескрипторов изображения — Гистограмм Аккумуляторного пространств Хафа (Hough accumulated histograms, HAH, HAH-AR). Локальные дескрипторы строятся с помощью оконного преобразования Хафа, полученного на основе рассмотрения модели лучевого преобразования Радона градиентного поля изображения. Преобразование Хафа может описывать контуры и пространственные особенности объектов на изображении, что позволяет учитывать локальное распределение ориентаций и положений контурных признаков. В работе предлагается использовать полносвязную нейронную сеть для увеличения разделяющей способности HAH семейства дескрипторов. Для её обучения используется нейросетевой метод опорных векторов. Эксперименты, проведенные на базах изображений пешеходов INRIA и CityPersons, показали, что использование нейросетевого классификатора HAH-AR дескрипторов позволяет уменьшить ошибки детектирования по сравнению с использованием SVM классификатора HAH и HOG дескрипторов. Результаты тестов предлагаемого метода показали близкую к сверточным нейронным сетям точность, при этом выигрывая в скорости на CPU в полтора раза.

Работа поддержана грантом РФФИ № 19-07-01231.

[1] *Самсонов Н. А. Гнеушев А. Н. Матвеев И. А.* Обучение классификатора на дескрипторах в пространстве преобразования радона // Известия российской академии наук, Теория и системы управления, 2020. С. 111–125.

Neural network classifier in the space of Radon Transform descriptors for efficient pedestrian detection

Nikita Samsonov[1]⋆ nikita.samsonov@phystech.edu
Alexander Gneushev[1] gneushev@ccas.ru
Ivan Matveev[1] matveev@ccas.ru
[1]Moscow, Federal Research Center "Computer Science and Control" of RAS

The problem of detecting pedestrians in the image is solved by constructing integral descriptors of the object and training a binary classifier on this discriptors. In this paper, family of local image descriptors — Hough accumulated histograms (HAH, HAH-AR) is used to construct the integral object descriptor. Local descriptors are constructed using the Hough window transform by applying Radon ray transform to the image gradient field. The Hough transform can describe the contours and spatial features of objects in the image, which allows to take into account the local distribution of orientations and positions of contour features. In this paper, we propose to use a fully connected neural network as binary classifier to increase the separating ability of the HAH descriptor. For its training, a neural support vector machine is used. Experiments carried out on the INRIA and CityPersons pedestrian image databases have shown that the use of the neural network classifier with HAH-AR descriptors reduces detection errors in comparison with the use of the SVM classifier together with HAH and HOG descriptors. Test results of the proposed method showed that its accuracy is comparable to the accuracy of convolutional neural networks, and it is 1.5 times faster on CPU.

This research is funded by RFBR, grant 19-07-01231.

[1] *Samsonov N, Gneushev A, Matveev I.* Training a Classifier by descriptors in the Space of the Radon Transform // Journal of Computer and Systems Sciences international, 2020. Vol. 59. No 3. Pp. 415–429.

Автоматическое проектирование интерпретируемых ансамблей на основе нечетких систем и нейронных сетей

Ахмедова Шахназ Агасувар кызы[1]⋆ shahnaz@inbox.ru
Становов Владимир Вадимович[1] vladimirstanovov@yandex.ru
Камия Юкихиро[2] kamiya@ist.aichi-pu.ac.jp
[1]Красноярск, Сибирский государственный университет науки и технологий имени академика Решетнева
[2]Накагуте, Университет префектуры Аичи

Технологии машинного обучения для решения задач классификации из различных областей приобретают все большую важность в связи с наступлением эры Интернета Вещей. Развитие современных алгоритмов для решения задач классификации привело к широкому применению нейронных сетей, нечетких систем, машин опорных векторов и т.д. в индустрии, финансовой сфере, медицине, технических и социальных науках. Однако классфикаторы на основе нечеткой логики являются одними из наиболее важных технологий машинного обучения, если необходима интерпретируемость результатов, полученных классификаторами. С другой стороны, нейронные сети обладают высокой точностью при низкой интерпретируемости, в то время как зачастую точность классификации нечеткими системами ниже.

Идея комбинирования нескольких методов анализа данных для решения задач классификации, зачастую реализуемое путем применения схемы голосования, не является новой. В настоящее время существуют различные схемы проектирования ансамблей: бэггинг, бустинг, случайные леса, стекинг и другие. Преимущество схемы голосования заключается в том, что она обычно приводит к высокой точности классификации и является более работоспособной для прикладных задач. Самой простой схемой проектирования ансамблей для решения задач классификации является схема голосования большинством. Однако стоит отметить, что наиболее эффективные алгоритмы проектирования ансамблей основаны на схеме взвешенного голосования.

В [1] авторами представлен новый способ взвешенного голосования. Данная схема использует значения функций принадлежности, которые возвращаются вместе с номером присвоенного класса в большинстве нечетких классификаторах. Упомянутые значения функций принадлежности используются как показатели уверенности принадлежности объекта к тому или иному классу, и затем голосованием на основе этих показателей принимается решение, использовать ли нечеткий классификатор или другой, являющийся дополнительным (в данном случае это нейронная сеть). Главное преимущество предложенного подхода заключается в комбинировании двух классификаторов, где первый является основным, а второй – вспомогательным. Подобная схема позволяет получать интерпретируемые и точные результаты.

Итак, выбранные нейронная сеть и нечеткая система возвращают некоторый показатель уверенности в принадлежности объекта к определенному классу. В базе правил нечеткой системы определяется правило-победитель путем сравнения значений функций принадлежности всех правил. Если значение функции принадлежности победившего правила равно 1, то данное правило полностью описывает рассматриваемый объект, если же это значение равно 0, то ни одно правило не может определить какой класс ему присвоить. Более того, если все значения функций принадлежности равны 0, то предполагается, что объект не классифицирован.

Для нейронной сети в данной схеме также используется функция softmax [1], возвращающая свои показатели уверенности, которые рассматриваются как вероятности принадлежности к некоторому классу. Если нейронная сеть не может определить класс, то все значения функции softmax равны $1/k$, где k - это число классов. Таким образом, в отличии от нечеткой системы нейронная сеть в любом случае присвоит некоторый класс объекту.

Совместное использование нечеткой системы и нейронной сети осуществляется следующим образом: после получения значений функций принадлежности нечеткой системы и присвоения класса каждому объекту из тестовой выборки отбираются 25% объектов с наиболее низкими значениями. Эти объекты классифицируются вспомогательным методом (нейронной сетью). Предложенная схема названа Confidence-Based Voting (CBV).

Разработанная схема была применена для решения двух задач классификации, связанных с автоматическим определением состояния человека или его реакции в заданный момент времени. В [1] рассмотрены задачи определения состояния человека и его реакции во время прослушивания музыки. Таким образом, 10 человек различных полов и возрастов участвовали в экспериментах. Их состояние наблюдалось с помощью бесконтактных датчиков в три этапа: во время прослушивания музыки, нравящейся и не нравящейся участнику эксперимента, а также в то время, когда музыка не прослушивалась.

Полученные данные были пред-обработаны и нормированы. В итоге были сформулированы следующие задачи классификации:

— задача "слушали данным присвоен класс "1 если участник эксперимента слушал музыку, и "0"в противном случае;
— задача "нравится данным присвоен класс "1 если участнику эксперимента нравилась музыка, и "0"в противном случае.

Результаты экспериментов представлены в Таблицах 1 и 2. В этих таблицах нечеткая система и нейронная сеть обозначены как "FS"и "NN"соответственно.

Таблица 1. Полученные значения точности классификации

Задача	**FS**	**NN**	**CBV**
Нравится	0.616	0.550	0.600
Слушали	0.613	0.750	0.688

Таблица 2. Полученные значения F-меры

Задача	**FS**		**NN**		**CBV**	
	1	0	1	0	1	0
Нравится	0.535	0.465	0.419	0.482	0.539	0.482
Слушали	0.220	0.784	0.000	0.857	0.155	0.823

Итак, для задач определения состояния человека разработанный алгоритм CBV позволяет достигнуть лучшие результаты, но это возможно только в том случае, когда вспомогательный классификатор имеет более высокую точность, чем нечеткая система. Предложенный алгоритм является обобщенной схемой объединения любых двух классификаторов. В дальнейшем будут рассмотрены другие технологии анализа данных для алгоритма CBV.

Работа поддержана Министерством науки и высшего образования Российской Федерации в рамках государственного контракта № FEFE-2020-0013.

[1] *Stanovov V., Akhmedova Sh., Kamiya Y.* Confidence-Based Voting for the Design of Interpretable Ensembles with Fuzzy Systems // Algorithms, 2020. Vol. 13. No 4. Pp. 86–92.

Automated design of interpretable ensembles based on fuzzy systems and neural networks

Shakhnaz Akhmedova[1]⋆ shahnaz@inbox.ru
Vladimir Stanovov[1] vladimirstanovov@yandex.ru
Yukihiro Kamiya[2] kamiya@ist.aichi-pu.ac.jp
[1]Krasnoyarsk, Reshetnev Siberian State University of Science and Technology
[2]Nakagute, Aichi Prefectural University

The importance of data analysis and machine learning techniques for solving classification problems from various areas is increasing due to the coming of the Internet of Things (IoT) era. To be more specific, the development of modern classification systems has led to the wide application of the neural networks, fuzzy systems, support vector machines and so on in industry, financial sphere, medicine, engineering and social sciences. However, classifiers based on the fuzzy logic are one of the most important technologies for machine learning in cases where interpretable classifiers are required. On the other hand, the neural networks have high accuracy and low interpretability at the same time, while the fuzzy systems are usually characterized by lower precision.

The idea to combine several data mining methods together to solve the classification problems is not new, and it is usually based on some implementation of voting. Nowadays there are various ensemble design schemes, which include bagging, boosting, random forest, stacking, and others. The advantage of voting scheme is that it usually leads to higher classification accuracies, and therefore it is more useful in the real-world scenarios. For classification problems the easiest way to create an ensemble is to use the voting by majority. It should be noted, that the most efficient algorithms for ensemble generation use the weighting schemes.

In [1] a novel weighting technique developed by authors is introduced. This technique uses the membership values, returned together with the class number in most fuzzy classification systems. These membership values are used as confidence levels, and the confidence-based voting decides whether the fuzzy classifier should be used, or the other, supporting classifier (neural network here). The main advantage of the proposed approach is that it concentrates on combining two classifiers, where the first one is the main, and the second is the assisting classifier. This inequality of classifier roles allows receiving more interpretable and accurate results.

Thus, the chosen neural network and the fuzzy logic method return some measure of confidence of the classifier together with the class number. The fuzzy rule bases generated by the fuzzy logic method are used to find the winner rule by comparing the membership values of every rule, so that the membership value of the winner rule changes in range $(0, 1)$. If this value equals one, then the classifier fully describes a given instance with this rule, and if the membership is zero, than among all the rules in the base, even the winner rule has low confidence about the true class number.

Furthermore, if all rules have zero confidence in their decision, then the instance is rejected, namely, considered as misclassified.

The neural network model implemented to classifiers introduced in [1] also has softmax layer, which has its own set of confidence values returned, these values are considered as probabilities. If the neural network is completely unaware about the class number, then all values of the softmax layer are equal to $1/k$, where k is the number of classes. So, the neural network does not have the rejected classification situation like fuzzy logic classifiers.

To use the fuzzy logic classifier together with other classifier, the following combination algorithm is proposed: after the fuzzy classifier returned the membership value and class number for all instances of the test set, 25% of instances with lowest membership values are chosen, and a threshold value is set. These instances are classified by the assisting method. The resulting method was called the Confidence-Based Voting (CBV).

Proposed classifiers were applied to two classification problems related to the automated detection of human condition or reaction to something at a given moment. In [1], as an example, human condition and reaction while listening and not listening to music were determined. For this purpose, firstly ten people of different genders and ages were asked to participate in experiments. Their condition was monitored by non-contact vital sensing using Doppler sensors over three stages: listening to music that a participant admitted to like and dislike in three different time periods each, and not listening to music at all in two different time periods.

Received signals were pre-processed by using and the obtained data were normalized. Thus, the following two classification problems were formulated:

— the problem "listened", where each instance was labelled as "1" if the participant listened to the music and "0" otherwise;
— the problem "liked", where each instance was labelled as "1" if the participant liked the music and "0" otherwise.

The results of the experiments are presented in Tables 1 and 2. In these tables fuzzy system and neural network are denoted as "FS" and "NN" respectively.

Table 1. Obtained values of the accuracy

Problem	**FS**	**NN**	**CBV**
Liked	0.616	0.550	0.600
Listened	0.613	0.750	0.688

Table 2. Obtained values of the F-measure

Problem	**FS**		**NN**		**CBV**	
	1	0	1	0	1	0
Liked	0.535	0.465	0.419	0.482	0.539	0.482
Listened	0.220	0.784	0.000	0.857	0.155	0.823

Thus, for the problems related to human condition detection the CBV allowed to achieve better results, but that happened only in cases when the assisting classifier

was better than the fuzzy rule base. The proposed algorithm is a general scheme for combining any two or more classifiers. Further studies may include experimenting with different basic learners and their combinations by using the CBV scheme.

This research is funded by the Ministry of Science and Higher Education of the Russian Federation within limits of state contract FEFE-2020-0013.

[1] *Stanovov V., Akhmedova Sh., Kamiya Y.* Confidence-Based Voting for the Design of Interpretable Ensembles with Fuzzy Systems // Algorithms, 2020. Vol. 13. No 4. Pp. 86–92.

Легковесная Шумоподавляющая Фильтрующая Нейронная Сеть Для Алгоритма FBP

Ямаев Андрей Викторович[1,2]⋆ rewin1996@gmail.com
Чукалина Марина Петровна[1,3,4] m.chukalina@smartengines.com
Николаев Дмитрий Валерьевна[1,3] d.p.nikolaev@smartengines.com
Шешкус Александер Владимирович[1] asheshkus@smartengines.com
Чуличков Алексей Иванович[2] achulichkov@gmail.com

[1]Москва, Smart Engines Service LLC
[2]Москва, Московский государственный университет имени М. В. Ломоносова
[3]Москва, Институт проблем передачи информации
[4]Москва, Институт кристаллографии и фотоники

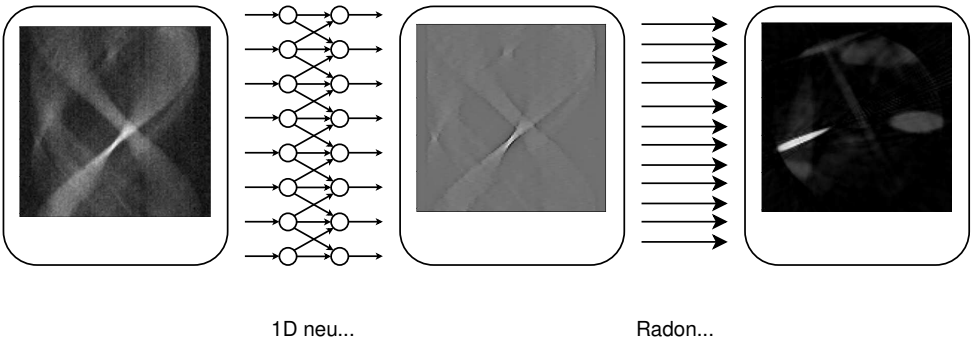


Рис. 1. Визуальное представление предложенного в работе подхода

В работе предложен подход, при котором одномерная нейронная сеть выполняет операцию фильтрации синограммы перед этапом обратного проецирования в схеме алгоритма FBP. Нейронная сеть выступает не только в качестве фильтра перед обратным проецированием, но и одновременно в качестве шумоподавляющего алгоритма. Мы перекладываем операцию фильтрации на нейронную сеть с целью уменьшения количества вычислений и в предположении, что это повысит качество реконструкции, нежели при использовании нейронной сети только с целью подавления шума на синограмме. При этом одномерность нейронной сети выбирается в целях ускорения вычисления реконструкции и минимазации потребляемой памяти, в противовес текущей тенденции использования тяжелах сетей для задач компьютерной томографии. Для обучения нейронной сети использовались генерируемые фантомы из пакета Adler. От фантомов строились синограммы с помощью преобразования Радона в параллельной схеме с количеством пикселей детектора равным 183 и 128 углами измерения. Посчитанные синограммы зашумлялись на основе квантовых свойств рентгеновского

излучения. Для сравнения реконструкции и фантома, по синограмме которого строилась реконструкция, использовалась нестандартная функция потерь, эффективность которой показали наши эксперименты,

Алгоритм	FBP	Одномерная свертка	Сверточная нейронная сеть	UNet 1D	Unet	LPDR	LPDR (SSIM loss)
Количество умножений на одну синограмму	0	$8,4 * 10^5$	$2,6 * 10^7$	$2,9 * 10^8$	$1,7 * 10^9$	$4,1 * 10^9$	$4,1 * 10^9$
Ширина рецептивного поля	183 x 128	51	50	50	50 x 50	183 x 128	183 x 128
Количество обучаемых параметров	0	$5,2 * 10^1$	$1,6 * 10^3$	$4,6 * 10^4$	$1,1 * 10^6$	$2,5 * 10^5$	$2,5 * 10^5$
Количество умножений на один обучаемый параметр	–	$16 * 10^3$	$16 * 10^3$	$6 * 10^3$	$1,5 * 10^3$	$16 * 10^3$	$16 * 10^3$
PSNR	25	24.5	31.4	33.2	34.0	**38.7**	35.5
SSIM	0.51	0.61	0.85	0.91	0.92	**0.96**	0.96
Время рекоснтрукции на CPU (ms)	44	**21**	81	90	105	1214	1214

$$Loss = -\sum_i SSIM(Phantom_i, BP(NN(NoisySin_i, W))), \quad (1)$$

где NN – функция нейронной сети, $NoisySin_i$ – зашумленная синограмма, подающаяся на вход нейронной сети и полученная из проеций с малыми временами

экспозиций, W – параметры и веса нейронной сети, BP – операция обратного проецирования, $Phantom_i$ – фантом, послуживший основой для создания синограммы $NoisySin_i$, $SSIM$ - метрика структурного подобия SSIM. Метрика SSIM была выбрана по причине того, что наши эксперименты показали значительное увеличение качества результатов работы сети после обучения по этой метрике, нежели при обучении по l^2 метрике. Архитектура предложенной нейронной сети представляет собой одномерный вариант архитектуры UNet. В работе идет сравнение нашего подхода с нейронной сетью из статьи Learned Primal Dual Reconstruction(LPDR), UNet, сверточной сетью, одномерной сверткой и алгоритмом FBP. В результате сравнения было выявлено, что наш подход может поддерживать достаточный уровень качества реконструкции (SSIM $>$ 0.9 на реконструкции к фантому) при этом являясь очень быстрой сетью(в 11-14 раз быстрее LPDR).

Работа поддержана грантами РФФИ №19-01-00790 и 18-29-26020.

Lightweight Denoising Filtering Neural Netwrok For FBP Algorithm.

Andrei Yamaev[1,2]⋆ rewin1996@gmail.com
Marina Chukalina[1,3,4] m.chukalina@smartengines.com
Dmitry Nikolaev[1,3] d.p.nikolaev@smartengines.com
Alexander Sheshkus[1] asheshkus@smartengines.com
Alexey Chulichkov[2] achulichkov@gmail.com
[1]Moscow, Smart Engines Service LLC
[2]Moscow, Moscow State University
[3]Moscow, Institute for Information Transmission Problems (Kharkevich Institute) RAS
[4]Moscow, FSRC "Crystallography and Photonics" RAS

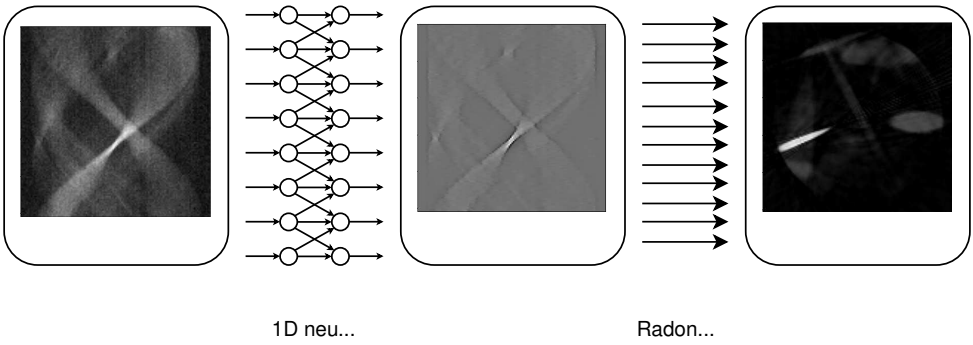


Figure 2. Visual representation of the proposed approach

The paper proposes an approach in which a one-dimensional neural network performs the sinogram filtering operation before the back-projection stage in the FBP algorithm scheme. The neural network acts not only as a filter before rear projection but also as a noise-canceling algorithm at the same time. We shift the filtering operation to a neural network in order to reduce the number of computations and on the assumption that this will improve the quality of reconstruction, rather than using a neural network only to suppress noise on the sinogram. In this case, the one-dimensionality of the neural network is chosen in order to speed up the reconstruction computation and minimize the consumed memory, as opposed to the current trend of using heavy networks for computed tomography tasks. Generated phantoms from the Adler package were used to train the neural network. Sinograms were constructed from the phantoms using the Radon transform in a parallel circuit with the number of detector pixels equal to 183 and 128 measurement angles. The calculated sinograms were noisy based on the quantum properties of X-ray radiation. To compare the reconstruction and the phantom, according to the sinogram

of which the reconstruction was constructed, a non-standard loss function was used, the effectiveness of which was shown by our experiments,

Algorithm	FBP	1d convolution	Convolution neural network	UNet 1D	Unet	LPDR	LPDR (SSIM loss)
Number of multiplications by one sinogram	0	$8,4 * 10^5$	$2,6 * 10^7$	$2,9 * 10^8$	$1,7 * 10^9$	$4,1 * 10^9$	$4,1 * 10^9$
Receptive field width	183 x 128	51	50	50	50 x 50	183 x 128	183 x 128
Training parameters amount	0	$5,2 * 10^1$	$1,6 * 10^3$	$4,6 * 10^4$	$1,1 * 10^6$	$2,5 * 10^5$	$2,5 * 10^5$
Number of multiplications by one training parameter	–	$16 * 10^3$	$16 * 10^3$	$6 * 10^3$	$1,5 * 10^3$	$16 * 10^3$	$16 * 10^3$
PSNR	25	24.5	31.4	33.2	34.0	**38.7**	35.5
SSIM	0.51	0.61	0.85	0.91	0.92	**0.96**	0.96
Work time (ms)	44	**21**	81	90	105	1214	1214

$$Loss = -\sum_i SSIM(Phantom_i, BP(NN(NoisySin_i, W))), \tag{2}$$

where NN is neural network function, $NoisySin_i$ is noisy sinogram fed to the neural network input, which obtained from projections with short exposure times, W are trainable parameters of the neural network NN, BP is Radon back-projection, $Phantom_i$ is a phantom(ground truth reconstruction) and base of created sinogram $NoisySin_i$, $SSIM$ is a structural similarity metric SSIM. The SSIM metric was chosen because our experiments showed a significant increase in the quality of neural network results after training using this metric, rather than when training using the

$L2$ metric. The architecture of the proposed neural network is a 1d representation of UNet architecture. The paper compares our approach with the neural network from the article Learned Primal-Dual Reconstruction (LPDR), UNet, convolutional network, one-dimensional convolution, and the FBP algorithm. As a result of the comparison, it was revealed that our approach can maintain a sufficient level of reconstruction quality (SSIM¿ 0.9 for reconstruction to the phantom) while being a very fast network (11-14 times faster than LPDR).

This research is funded by RFBR, grants 19-01-00790 and 18-29-26020.

Анализ и прогнозирование уровня глюкозы в крови на основе нейронных сетей и данных суточного мониторинга

Привезенцев Денис Геннадьевич[1]⋆ dgprivezencev@mail.ru
Жизняков Аркадий Львович[1] lvovich1975@mail.ru
Белякова Анна Сергеевна[1] asbelyakova@rambler.ru
[1]Муром, Муромский институт (филиал) ВлГУ

Работа направлена на разработку программного обеспечения для проведения исследований по формированию структуры и обучению многослойной нейронной сети для анализа и прогнозирования уровня глюкозы в крови. В ходе исследования был проведен анализ поставленной задачи, рассмотрены инструменты, методы и алгоритмы решения проблемы. Была исследована модель искусственной нейронной сети, разработана структура и проведено обучение.

В результате работы была разработана система исследования структуры и обучения многослойной нейронной сети для анализа и прогнозирования уровня глюкозы в крови.

Был проведен обзор инструментов реализации поставленной задачи и выбраны методы решения. Рассмотрены теория создания искусственных нейронных сетей и исходные данные, исследованы типы и алгоритмы обучения нейронной сети.

Анализ разработанной нейронной сети для прогнозирования уровня глюкозы в крови показал, что точность прогноза составляет 78%. Однако точность прогнозирования уровня глюкозы в крови в периоды без приема углеводов составляет 93%, а в периоды потребления углеводов - 56%. Из этого следует, что необходимо разработать индивидуальную модель усвоения углеводов.

Дальнейшее развитие, совершенствование и использование системы предполагает разработку методов создания индивидуальных моделей усвоения пищи и инсулина.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (Госзадание ВлГУ ГБ-1187/20)

[1] *Privezentsev D. G., Zhiznyakov A. L., Belyakova A. S., Astafiev A. A.* Formation of the structure and training of a multilayer neural network for the analysis and blood glucose level prediction // Journal of Physics: Conference Series, 2021.

Analysis and prediction of blood glucose levels based on neural networks and daily monitoring data

Denis Privezentsev[1]⋆ dgprivezencev@mail.ru
Arcady Zhiznyakov[1] lvovich1975@mail.ru
Anna Belyakova[1] asbelyakova@rambler.ru
[1]Murom, Murom institute (branch) Vladimir State University

The work aims to develop software for conducting research on the formation of the structure and training a multilayer neural network for analyzing and blood glucose level prediction. The paper analyzes the task, considers the tools, methods and algorithms for solving the problem. A model of an artificial neural network was investigated, the structure was developed and training was carried out.

As a result of the work, a system was developed for researching the formation of the structure and training of a multilayer neural network for analyzing and blood glucose level prediction.

A review of tools for the implementation of the task was carried out and methods of solution were chosen. The theory of the creation of artificial neural networks and the initial data are considered, the types and algorithms for learning the neural network are investigated.

Analysis of the developed neural network for predicting blood glucose levels showed that the prediction accuracy is 78%. However, the accuracy of predicting blood glucose levels during periods without carbohydrate intake is 93%, and during periods of carbohydrate intake is 56%. From this it follows that a new, individualized pattern of carbohydrate absorption is needed.

Further development, improvement and use of the system suggests the development of methods for creating individual models of food and insulin absorption.

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation (State task of VlSU GB-1187/20).

[1] *Privezentsev D. G., Zhiznyakov A. L., Belyakova A. S., Astafiev A. A.* Formation of the structure and training of a multilayer neural network for the analysis and blood glucose level prediction // Journal of Physics: Conference Series, 2021.

Графовые нейронные сети для несвязанных графов в химических реакцях.

Никитин Филипп Александрович[1]⋆ filipp.nikitin@phystech.edu
Стрижов Вадим Викторович[1, 2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, ВЦ им. А. А. Дородницына ФИЦ ИУ РАН

Разработана точная и интерпретируемая модель предсказания продуктов и отображения атомов химических реакций. Предложена новая архитектура нейронной сети, DRACON, для классификации вершин в несвязанных графах. Архитектура обобщает графовые сверточные нейронные сети для несвязанных графов, используется механизм внимания и построение иерархического представления исходных графов. Модель также обобщает идею векторных представлений графов использующих идею псевдо-вершин, демонстрирующую наилучшее качество для множества задач в вычислительной химии. Результаты на USPTO_STEREO показывают, что DRACON предсказывает атомы основного продукта и центры для химических реакций с высокой точностью.

DRACON имеет интерпретируемую структуру: использует знания о типе химической связи и признаки атомов в исходных молекулах. Представленные псевдо-вершины химических реакций неявно выучивают сходство химических реакций. Это может быть полезно для основанных на правилах системах, так как распознает и категоризирует реакции.

Авторы рассматривают применение DRACON к молекулярным графам в химических реакциях. Представленный подход применим к несвязанным графам в общем виде. Это расширяет применимость графовых сверточных нейронных сетей для различных задач в вычислительной химии таких как классификация атомов в молекулярных графах, классификации молекулярных графов, предсказании свойств атомов в реакциях и растворах.

Работа выполнена при поддержке РФФИ (проекты 19-07-01155, 19-07-00875) и НТИ (проект 13/1251/2018).

[1] *Nikitin F., Isayev O., and Strijov V.* DRACON: Disconnected Graph Neural Network for Atom Mapping in Chemical Reactions // Phys. Chem. Chem. Phys., The Royal Society of Chemistry., 2020

Graph neural networks for disconnected graphs in chemical reactions.

Filipp Nikitin[1]⋆ filipp.nikitin@phystech.edu
Vadim Strijov[1, 2] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

We developed an interpretable and accurate model for outcome prediction and atom mapping in chemical reactions. A novel neural network architecture, DRACON were proposed for node classification problem in disconnected graphs. It generalizes graph convolution neural network for a disconnected graph with self-attention mechanism and learning hierarchical representations of source graphs. The model also generalizes the idea of graph representation learning with pseudo-nodes, which is state-of-the-art for a variety of problems in drug discovery. The model was analyzed on the large-scale USPTO_STEREO dataset. The results demonstrate that DRACON predicts atoms of the main product and centers in a chemical reaction with high accuracy.

DRACON has an interpretable structure: it uses types of chemical bonds and characteristics of atoms in source molecules. The introduced pseudo-nodes of chemical reactions implicitly learn the similarity of chemical reactions. This feature is useful for rule-based Named Reaction Identification systems that allow the recognition and categorization of reactions from their connection tables, error identification, reaction visualization, and retrieval tasks.

This authors consider the application of DRACON to molecular graphs in chemical reactions. The approach presented can be applied to disconnected graphs in general. It expands the graph convolution neural networks for various problems in computational chemistry such as atom classification in molecular graphs, classification of molecular graphs, different prediction of atom's properties in reactions and solutions.

This research was supported by RFBR (projects 19-07-01155, 19-07-00875) and NTI (project 13/1251/2018).

[1] *Nikitin F., Isayev O., and Strijov V.* DRACON: Disconnected Graph Neural Network for Atom Mapping in Chemical Reactions // Phys. Chem. Chem. Phys., The Royal Society of Chemistry., 2020

Оценка качества работы нейронной сети для предсказания 3D модели объекта

Князь Владимир Владимирович[1,2]⋆ vl.kniaz@gosniias.ru
Мизгинов Владимир Андреевич[1] vl.mizginov@gosniias.ru
Гродзицкий Лев Владимирович[1] moshkantsev@gosniias.ru
Мошканцев Пётр Владиславович[1] lev.grodzickiy@gosniias.ru
[1]Москва, ФГУП «ГосНИИАС»
[2]Москва, Московский физико-технический институт

На сегодняшний день фотограмметрические сканеры со структурированным подсветом интенсивно используются для решения таких задач, как оптическая метрология, контроль качества на конвейере и документирование объектов культурного наследия. Активные фотограмметрические системы на основе структурированного света демонстрируют высокую точность и высокую производительность для получения нескольких трехмерных координат поверхности объекта. Несмотря на высокую конкуренцию среди производителей (более 20 компаний разрабатывают коммерчески доступные сканеры структурированного света), точность технологии структурированного света имеет ограниченное быстродействие при работе в реальном времени. Зачастую на поверхности восстановленных трёхмерных моделей возникают расхождения, особенно если текстура модели имеет серьезные изменения в яркости или обладает высокой отражающей способностью. Было предложено множество методов компенсации ошибки стереотождествления для систем со структурированным подсветом. Хотя эти методы уменьшают ошибку реконструкции, они не могут устранить расхождения, вызванные неравномерной яркостью текстуры.

Алгоритмы на основе глубоких нейронных сетей оказались эффективными для восстановления трехмерных моделей. Недавно было предложено новое поколение нейронных сетей, которое называется генеративно-состязательными сетями (GAN). Эти сети могут быть обучены решать сложные задачи обработки изображений, такие как устранение смаза, шумов на изображении; преобразование изображения одного объекта в изображение другого; колоризация черно-белых изображений; увеличение разрешения изображений без потери качества. Модель GAN состоит из двух сетей: генератора G и дискриминатора D. Две сети обучаются одновременно для решения противоположных задач. Цель дискриминатора D состоит в том, чтобы отличить «реальные» изображения из обучающего набора данных от «поддельных» поддельных изображений, генерируемых генератором G. Цель генератора G - синтез «поддельных» выборок, которые будут неотличимы от случайных изображений из обучающего набора данных.

Данная работа посвящена оценке качества работы нейронной сети SSZ (Single Shot Z-space segmentation) [**?**] для преобразования одного изображения в семантическую воксельную модель. Полученные результаты сравниваются с

результатами восстановления трёхмерной модели с помощью сканера структурированного света. Оцениваемая нейронная сеть SSZ одновременно выполняет реконструкцию трехмерной воксельной модели и семантическую сегментацию трехмерной сцены из одного изображения. Предполагается, что семантическая сегментация классов 3D-объектов будет способствовать увеличению точности восстановления трехмерной сцены глубокой нейронной сетью. С этой целью была предложена цветовая сегментация воксельной модели каждого класса в сцене. В качестве базовой архитектуры генератора используется сеть U-Net, которая содержит инвертированные редуцированные блоки и может проецировать прямой переход из 2D в 3D, используя контурные соответствия между изображением и 3D-моделью.

Работа выполнена при поддержке Российского научного фонда, грант РНФ № 19-11-11008.

3D Reconstruction Neural Network Quality Evaluation

Vladimir Kniaz[1,2]⋆ vl.kniaz@gosniias.ru
Vladimir Mizginov[1] vl.mizginov@gosniias.ru
Lev Grodzitsky[1] moshkantsev@gosniias.ru
Petr Moshkantsev[1] lev.grodzickiy@gosniias.ru
[1]Moscow, FGUP ¡¡GosNIIAS¿¿
[2]Moscow, Moscow Institute of Physics and Technology

Structured light scanners are intensively exploited in various applications such as non-destructive quality control at an assembly line, optical metrology, and cultural heritage documentation. While more than 20 companies develop commercially available structured light scanners, structured light technology accuracy has limitations for fast systems. Model surface discrepancies often present if the texture of the object has severe changes in brightness or reflective properties of its texture. The primary source of such discrepancies is errors in the stereo matching caused by complex surface texture. Many methods were proposed to compensate for the error in stereo matching for structured light systems. While these methods reduce surface distance error between reconstructed and the ground truth models, they could not eliminate the discrepancies caused by uneven texture brightness.

Algorithms based on deep neural networks have proven to be effective tools for reconstructing three-dimensional models. Recently a new generation of neural networks has been proposed that is commonly named Generative Adversarial Networks (GANs). These networks could be trained for complex image-to-image translation tasks such as object transfiguration, image super-resolution and noise reduction. A GAN model consists of two networks: a generator G and a discriminator D. Two networks are trained simultaneously for concurrent tasks. The aim of the discriminator D is to distinguish 'real' samples B from the training dataset from 'fake' samples $\hat{B}$ produced by the generator G. The objective of the generator G is the synthesis of 'fake' samples $\hat{B}$ that are indistinguishable from the random samples B from the training datasets.

This paper is focused on the evaluation of a deep neural network SSZ (Single Shot Z-space segmentation) [**?**] for translation a single image into a semantic voxel model. The results obtained are compared with the results of the reconstruction of a three-dimensional model using a structured light scanner. The evaluated neural network SSZ simultaneously performs 3D voxel model reconstruction and semantic segmentation of a 3D scene from a single image. It is proposed that the semantic segmentation of classes of 3D objects will increase the accuracy of reconstruction of a three-dimensional scene by a deep neural network. Semantic labeling of 3D object classes on 3D scene was proposed. We use a U-net-like generator with inverted residuals blocks as a starting point. Such 3D representation allows us to design direct 2D-to-3D skip connections, that leverage contour correspondences between an image and a 3D model.

The reported study was funded by the Russian Science Foundation (RSF) according to the research project N^{o} 19-11-11008.

Глубокое обучение для задач отображения улучшенного видения на ИЛС

Гродзицкий Лев Владимирович[1]⋆ moshkantsev@gosniias.ru
Данилов Сергей Юрьевич[1,2] danilov@gosniias.ru
Князь Владимир Владимирович[1,2] vl.kniaz@gosniias.ru
[1]Москва, ФГУП «ГосНИИАС»
[2]Москва, Московский физико-технический институт

Ситуационная осведомленность экипажа имеет решающее значение для безопасности полета. Индикатор на лобовом стекле (ИЛС) позволяет отображать всю необходимую полетную информацию перед пилотом на фоне закабинной обстановки. ИЛС создан для решения проблемы информационной перегрузки при пилотировании самолета.

В то время как средствами компьютерной графики цифровая модель местности и ее масштабы могут быть легко спроецированы на ИЛС, ошибки в калибровке проекционного дисплея ИЛС для правильного представления данных с оптических датчиков (камер видимого и инфракрасного диапазона) вызывают проблемы. Основная проблема возникает из-за параллакса между глазами пилота и положением камеры за бортом.

Данная статья посвящена разработке алгоритма калибровки проекции на ИЛС в режиме реального времени для согласования проекции трехмерной модели местности и ВПП на проекционном дисплее ИЛС. Алгоритм базируется на методах предложенных в предыдущих исследованиях [?]. Цель алгоритма – совмещение объектов, видимых через стекло кабины, с их проекциями на проекционном дисплее ИЛС. Для повышения точности проецирования используется дополнительный многоканальный оптический датчик, установленный на самолете. В его состав входят сенсоры видимого и инфракрасного диапазонов. Выходные данные с него представляют собой видеопоследовательность комплексированных кадров видимого и инфракрасного диапазонов.

В разработанном алгоритме сочетаются классические фотограмметрические методы с современными подходами к глубинному обучению. В частности, используется модель нейронной сети для обнаружения объектов, чтобы найти зону взлётно-посадочной полосы (ВПП) и согласовать проекцию ВПП с ее фактическим расположением. Кроме того, для устранения ошибок, вызванных параллаксом, используется повторная проекция кадра закабинного датчика на трехмерную модель рельефа.

Для оценки работоспособность алгоритма разработан симулятор среды окружения. С помощью симулятора подготовлен большой набор обучающих данных. Набор данных включает 2000 видеопоследовательностей, представляющих движение самолета во время взлета, посадки и руления. Результаты оценки качества работы алгоритма показывают, что качественно и количественно пред-

ложенный алгоритм способен точно согласовать контурны трехмерных моделей, проецируемых на ИЛС, и контуры соответствующих им реальных объектов.

Работа выполнена при поддержке Российского научного фонда, грант РНФ № 19-11-11008.

Deep Learning for Projection of the Enhanced Vision on the HUD

Lev Grodzitsky[1]⋆ moshkantsev@gosniias.ru
Sergey Danilov[1,2] danilov@gosniias.ru
Vladimir Kniaz[1,2] vl.kniaz@gosniias.ru
[1]Moscow, FGUP ¡¡GosNIIAS¿¿
[2]Moscow, Moscow Institute of Physics and Technology

Situational awareness of the crew is critical for the safety of the air flight. The Head-up display allows providing all required flight information in front of the pilot over the cockpit view visible through the cockpit's front window. This device has been created for solving the problem of informational overload during the piloting of an aircraft.

While computer graphics allows projecting flight information and digital terrain model on such display, errors in the Head-up display alignment for correct presenting of sensor data pose challenges.The main problem arises from the parallax between the pilot's eyes and the position of the camera.

This paper is focused on the development of an online calibration algorithm for the conform projection of the 3D terrain and runway models on the pilot's head-up display. We use assumptions made by Danilov et. al [**?**] as the starting point for our research. The aim of our algorithm is to align the objects visible through the cockpit glass with their projections on the Head-up display. To improve the projection accuracy, we use an additional multi-channel optical sensor installed on the aircraft. It includes visible and infrared sensors. The output data from it is a video sequence of complexed frames of the visible and infrared ranges.

The developed algorithm combines classical photogrammetric methods with modern approaches to deep learning. In particular, an object detection neural network model is used to find the runway area and match the runway projection with its actual location. In addition, the sensor image is re-projected onto the 3D terrain model to eliminate errors caused by parallax.

An environment simulator was developed to evaluate the performance of the algorithm. A large training dataset was generated using the simulator. The dataset includes 2,000 video sequence images representing the movement of an aircraft during takeoff, landing, and taxi. The results of the algorithm evaluation show that both qualitatively and quantitatively, the proposed algorithm is able to accurately match the 3D models projected on a Head-up display.

The reported study was funded by the Russian Science Foundation (RSF) according to the research project N^{o} 19-11-11008.

Опыт применения многослойных свёрточных нейронных сетей и технологий Big Data на примере искусственного медицинского интеллекта ФтизисБиоМед

Гогоберидзе Юрий Тенгизович[1] gut@vector.ru
Классен Виктор Иванович[1] kvi@vector.ru
Натензон Михаил Яковлевич[2]⋆ mnatenzon4@gmail.com
Просвиркин Илья Александрович[1] pia@vector.ru
Сафин Артем Альбертович[1] saal@vector.ru
[1]Чистополь, ООО «ФтизисБиоМед»
[2]Москва, НПО «Национальное телемедицинское агентство»

В первые два десятилетия 21-века технологии искусственного интеллекта (ИИ) сделали огромный скачок, совершив качественный переход от теоретических моделей и лабораторных конструктов вначале 2000-х до продуктов частного потребления и отраслевых сервисов в 2020-м. Распознавание образов и анализ больших данных прочно вошли во многие сферы деятельности человечества: маркетинг, управление транспортным движением, безопасность, коммуникации, строительство, дизайн, индустрия развлечений, здравоохранение.

ООО «ФтизисБиоМед» занимается развитием технологий ИИ в области диагностической медицины с 2015 года.

В рамках данной работы разработан ансамбль сверточных нейронных сетей (СНС) осуществляющих анализ медицинских изображений органов грудной клетки (рентгенограмм, флюорограмм) на предмет обнаружения патологий[1,2]. Имея исходный массив из 300000 флюорограмм, группа математиков и рентгенологов создала базу данных из размеченных снимков, в которых области верифицированных патологий выделялись контурами. Сформированная выборка включала изображения с масками 21-й патологии. На изображениях из данной базы было проведено 10000 эпох дообучения ансамбля СНС, что позволило получить приемлемый для врачей инструмент поддержки принятия решений.

С февраля 2020 года сервис, разработанный компанией, принимает участие в Эксперименте по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы.

Калибровочные тесты ИИ, проведенные ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» (ГБУЗ НПКЦ ДиТ ДЗМ), показали увеличение диагностической точности сервиса.

Площадь под ROC-кривой составила 0,9 (0,84 – 0,96) с доверительной вероятностью 0,95. По критерию максимизации прогностической ценности отрицательного результата в результате калибровочного тестирования определен порог активации нейронной сети 0,2.

Результаты тестов, представленные в Таблице 1 и на Рисунке 1. показали, что сервис подходит для поддержки клинических решений, принимаемых в рамках первичной диагностики.

Таблица 1. Результаты калибровочного тестирования сервиса

№	Наименование	Заявленное значение	Полученное значение на эталонном наборе данных	
			Метод определения оптимального порога	
			индекс Юдена (YI)	максимизации прогностической ценности отрицательного результата (max NPV)
1	Значение площади под ROC-кривой (AUC)	0,79	0,9	
2	точность	0,7397 (0,6238-0,8355)	0,86 (0,77-0,92)	0,8 (0,71-0,87)
3	чувствительность	0,913 (0,7196-0,9893)	0,8 (0,66-0,9)	0,94 (0,83-0,99)
4	специфичность	0,66 (0,5123-0,7879)	0,92 (0,81-0,98)	0,66 (0,51-0,79)
5	Удельный вес ложноотрицательных результатов		0,2	0,06
6	Удельный вес ложноположительных результатов		0,08	0,34
7	Оптимальный порог		0,78	0,2

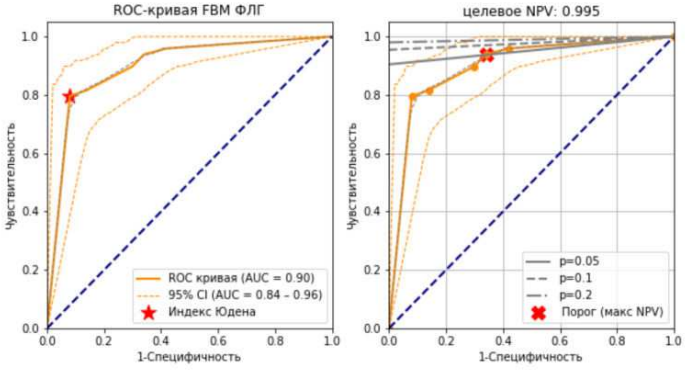


Рисунок 1. Слева ROC – кривая, построенная на основании обработки ИИ-сервисом «FBM » набора из 100 исследований для флюорографии. Справа на том же графике отмечены прямые для NPV=0.995 и для пре-тестовых вероятностей 0.05, 0.1 и 0.2

На основании, собранной в ходе эксперимента статистики, можно будет сделать выводы о готовности технологий ИИ к работе в системе здравоохранения Российской Федерации и целесообразности широкой интеграции на нынешнем этапе технического развития.

В заключении будут обозначены перспективы и дальнейшие пути развития технологий ИИ в области здравоохранения.

[1] *Гогоберидзе Ю. Т., Классен В. И., Натензон М. Я., Просвиркин И. А., Сафин А. А.* Особенности имплементации систем искусственного интеллекта в задаче анализа двухмерных радиологических изображений // Математические методы распознавания образов, 2019. С. 307–308.

[2] *Klassen V. I., Safin A. A. Maltsev A. V., Adrianov N. G., Morozov S. P., Vladzymyrskyy A. V., Ledikhova N. V., Sokolina I. A., Kulberg N. S., Gombolevsky V. A., Kuzmina E. S.,* AI-based screening of pulmonary tuberculosis: diagnostic accuracy // Journal of eHealth Technology and Application, 2018. Vol. 16. No 1. Pp. 28–32.

Experience of using multilayer convolutional neural networks and Big Data technologies on the example of artificial medical intelligence FtisisBioMed

Yuriy Gogoberidze[1] gut@vector.ru
Victor Klassen[1] kvi@vector.ru
Mikhail Natenzon[2]⋆ mnatenzon4@gmail.com
Iliya Prosvirkin[1] pia@vector.ru
Artem Safin[1] saal@vector.ru
[1]Chistopol, FtizisBioMed LLC
[2]Moscow, National Telemedicine Agency

In the first two decades of the 21st century, artificial intelligence (AI) technologies made a huge leap, making a qualitative transition from theoretical models and laboratory constructs in the early 2000s to private consumer products and industry services in 2020. Pattern recognition and big data analysis have become firmly established in many areas of human activity: marketing, traffic management, security, communications, construction, design, entertainment, healthcare.

FtisisBioMed LLC has been developing AI technologies in the field of diagnostic medicine since 2015.

Within the boundaries of this work, an ensemble of convolutional neural networks (CNN) has been developed that analyzes medical images of the chest organs (X-rays, fluorograms) for the detection of pathologies[1,2]. Having an initial array of 300,000 fluorograms, a group of mathematicians and radiologists created a database of labeled images, in which areas of verified pathologies were highlighted by contours. The formed sample included images with masks of the 21st pathology. On the images from this database, 10,000 epochs of additional training of the CNN ensemble were carried out, which made it possible to obtain a decision support tool acceptable for doctors.

Since February 2020, the service developed by the company has been taking part in the Experiment on the use of innovative technologies in the field of computer vision for the analysis of medical images and further application in the healthcare system of the city of Moscow. AI calibration tests carried out by Moscow Radiology showed an increase in the diagnostic accuracy of the service. The AUC metric value was 0.9 (0.84 - 0.96) with a confidence level of 0,95. According to the criterion of maximizing the predictive value of a negative result as a result of calibration testing, the neural network activation threshold was determined as 0.2.

The test results presented in Table 1 and Figure 1 showed that the service is suitable to support clinical decisions made in the framework of primary diagnosis.

The report is devoted to the problems of introducing AI technologies into the diagnostic practice of healthcare institutions in the Russian Federation, as well as to the practical results and analysis of the experience of using artificial intelligence systems during the Experiment.

Table 1. Service calibration test results

№	Metric	Expected value	The resulting value on the reference dataset	
			Method for determining the optimal threshold	
			Yuden Index (YI)	Maximization of Negative Predictive Value (max NPV)
1	Area under curve (AUC)	0,79	0,9	
2	accuracy	0,7397 (0,6238-0,8355)	0,86 (0,77-0,92)	0,8 (0,71-0,87)
3	sensitivity	0,913 (0,7196-0,9893)	0,8 (0,66-0,9)	0,94 (0,83-0,99)
4	specificity	0,66 (0,5123-0,7879)	0,92 (0,81-0,98)	0,66 (0,51-0,79)
5	False Negative Rate (FNR)		0,2	0,06
6	False Positive Rate (FPR)		0,08	0,34
7	Optimal threshold		0,78	0,2

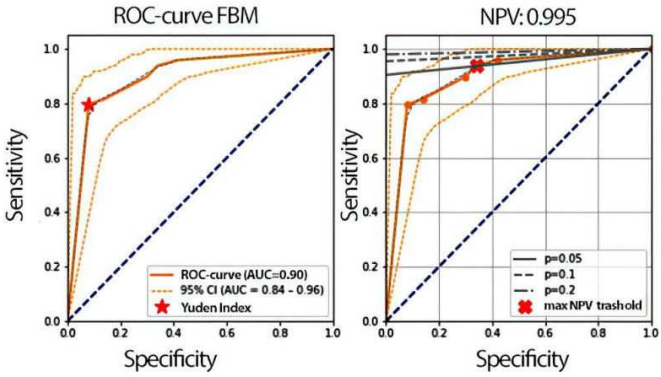


Figure 1. Left: Roc-curve, based on AI analysis of 100 images from calibrating sample. Right; same graph with lines for NPV=0,995

The report will consider the problems of both a technological nature, including the problems of compatibility and implementation of AI services in an already formed system for storing and processing diagnostic data, and problems associated with the human factor. In addition, the issues of legal and legislative regulation, as well as compliance with medical ethics standards will be discussed.

Based on the statistics collected during the experiment, it will be possible to draw conclusions about the readiness of AI technologies to work in the healthcare system of the Russian Federation and the feasibility of widespread integration at the current stage of technical development.

In conclusion, the prospects and further ways of developing AI technologies in the field of healthcare will be outlined.

[1] *Gogoberidze Y., Klassen V., Natenzon M., Prosvirkin I., Safin A.* Features of the implementation of artificial intelligence systems in the task of analysis of two-dimensional radiological images // Mathematical Methods for Pattern Recognition, 2019. Pp. 309–310.

[2] *Klassen V. I., Safin A. A. Maltsev A. V., Adrianov N. G., Morozov S. P., Vladzymyrskyy A. V., Ledikhova N. V., Sokolina I. A., Kulberg N. S., Gombolevsky V. A., Kuzmina E. S.,* AI-based screening of pulmonary tuberculosis: diagnostic accuracy // Journal of eHealth Technology and Application, 2018. Vol. 16. No 1. Pp. 28–32.

Об одном робастном подходе к поиску центров кластеров

Шибзухов Заур Мухадинович[1,2]⋆ intellimath@mail.ru

[1]Москва, Московский физико-технический институт

[2]Москва, Институт математики и информатики Московского педагогического государственного университета

Стандартная постановка задачи поиска центра $\mathbf{c}(\mathbf{X})$ конечного множества $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$ имеет вид:

$$\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \sum_{k=1}^{N} \|\mathbf{x}_k - \mathbf{c}\|^2.$$

Функция $\|\mathbf{z}\|^2$ чувствительна к существенным искажениям. Рассмотривая эквивалентную постановку задачи

$$\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \tfrac{1}{N} \sum_{k=1}^{N} \tfrac{1}{n} \sum_{i=1}^{n} (x_{ki} - c_i)^2$$

и заменяя среднее арифметическое на дифференцируемые операторы усреднения нечувствительные (малочувствительные) к выбросам, получаем следующие постановки задачи:

1) $\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \sum_{k=1}^{N} \overset{[}{i}= 1]n\mathsf{M}(x_{ki} - c_i)^2.$

2) $\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \overset{[}{k}= 1]N\mathsf{M}\, \overset{[}{i}= 1]n\mathsf{M}(x_{ki} - c_i)^2.$

3) $\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \overset{[}{k}= 1]N\mathsf{M}\|\mathbf{x}_k - \mathbf{c}\|^2.$

где M – оператор усреднения, нечувствительный к выбросам [1, 2].

Рассматриваются алгоритмы типа итеративного перевзвешивания для поиска центров кластеров, основанные на постановках 1)–3). На модельных примерах показываются их возможности по преодолению влияния выбросов в $\mathbf{X}$ на поиск центров кластеров. Работа выполнена при поддержке гранта РФФИ №18-01-00050.

[1] *Shibzukhov Z. M.* On the principle of empirical risk minimization based on averaging aggregation functions // Doklady Mathematics, 2017. Vol. 96. No. 3. Pp. 494–497.

[2] *Shibzukhov Z. M., Kazakov M. A.* Clustering based on the principle of finding centers and robust averaging functions of aggregation // Journal of Physics: Conference Series, 2019.

About one robust approach to the search for cluster centers

Zaur Shibzukhov[1,2]⋆ intellimath@mail.ru

[1]Moscow, Moscow Institute of Physics and Technology

[2]Moscow, Institute of Mathematics and Computer Science of Moscow Pedagogical State University

The standard formulation of the problem of finding the center $\mathbf{c}(\mathbf{X})$ of a finite set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$ has the form:

$$\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \sum_{k=1}^{N} \|\mathbf{x}_k - \mathbf{c}\|^2.$$

The function $\|\mathbf{z}\|^2$ is sensitive to significant distortions. Because of this, $\mathbf{c}(\mathbf{X})$ is also prone to corruption. Considering the equivalent problem statement

$$\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \tfrac{1}{N} \sum_{k=1}^{N} \tfrac{1}{n} \sum_{i=1}^{n} (x_{ki} - c_i)^2$$

and replacing the arithmetic mean by differentiable averaging operators that are insensitive to outliers, we obtain the following problem statements:

1) $\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \sum_{k=1}^{N} \mathsf{M}[i = 1] n \mathsf{M}(x_{ki} - c_i)^2.$

2) $\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \mathsf{M}[k = 1] N \mathsf{M}[i = 1] n \mathsf{M}(x_{ki} - c_i)^2.$

3) $\mathbf{c}(\mathbf{X}) = \arg\min_{\mathbf{c}} \mathsf{M}[k = 1] N \mathsf{M} \|\mathbf{x}_k - \mathbf{c}\|^2.$

where M is the outlier insensitive averaging operator [**?**, **?**].

Algorithms of the iteratively reweighting type for finding cluster centers based on statements 1)–3) are considered. Illustrative examples show their capabilities to overcome the influence of outliers in $\mathbf{X}$ on the search for cluster centers. This work was supported by the RFBR grant No 18-01-00050.

[1] *Shibzukhov Z. M.* On the principle of empirical risk minimization based on averaging aggregation functions // Doklady Mathematics, 2017. Vol. 96. No. 3. Pp. 494–497.

[2] *Shibzukhov Z. M., Kazakov M. A.* Clustering based on the principle of finding centers and robust averaging functions of aggregation // Journal of Physics: Conference Series, 2019.

Алгоритм ближайших выпуклых оболочек на основе использования линейного программирования

Немирко Анатолий Павлович[1]⋆ apn-bs@yandex.ru
Дула Хосе[2] jhdula@cba.ua.edu
[1]Санкт-Петербург, СПбГЭТУ "ЛЭТИ"
[2]Таскалуса, США, Университет Алабамы

В работе рассмотрены алгоритмы классификации, основанные на анализе ближайших к испытуемой точке выпуклых оболочек. Предложен новый способ оценки близости испытуемой точки к выпуклой оболочке класса. Он основан на решении оптимизационной задачи линейного программирования, которая приводит к измерительной процедуре приближенной оценки близости тестовой точки к выпуклой оболочке класса. Эта же процедура дает возможность определить расположение тестовой точки: внутри или вне выпуклой оболочки класса.

На основе этого способа оценки близости построен алгоритм классификации ближайшей выпуклой оболочки. Классификатор такого типа хорошо подходит для задач биометрии с большим числом классов и малыми объемами обучающих выборок по классам, например, для задач распознавания людей по лицам или отпечаткам пальцев.

Приведены результаты экспериментальных исследований на реальной задаче медицинской диагностики. Сравнение эффективности предложенного классификатора с другими типами классификаторов, в частности с классификатором LNCH [1], показало высокую эффективность предложенного классификатора, реализующего новый метод оценки близости на основе линейного программирования.

Работа поддержана грантами РФФИ № 19-29-01009 и № 18-29-02036.

[1] *Nemirko A. P.* Lightweight nearest convex hull classifier // Pattern Recogn. Image Anal., 2019. Vol. 29. No 2. Pp. 360–365

Nearest convex hulls algorithm based on linear programming — IDP-13

Anatoliy Nemirko[1]⋆ apn-bs@yandex.ru
Jose Dula[2] jhdula@cba.ua.edu
[1]Saint Petersburg, ETU "LETI"
[2]Tuscaloosa, University of Alabama

The paper considers classification algorithms based on the analysis of convex hulls closest to the test point. A new method for estimating the proximity of the test point to the convex hull of the class is proposed. It is based on solving an optimization problem of linear programming, which leads to a measuring procedure for approximate estimation of the proximity of the test point to the convex hull of the class. The same procedure makes it possible to determine the location of the test point: inside or outside the convex hull of the class.

Based on this method of proximity estimation, an algorithm for classifying the nearest convex hull is constructed. This type of classifier is well suited for biometrics tasks with a large number of classes and small volumes of training samples by class, for example, for recognizing people by face or fingerprint.

The results of experimental studies on the real problem of medical diagnostics are presented. Comparison of the efficiency of the proposed classifier with other types of classifiers, in particular with the LNCH classifier [1], showed the high efficiency of the proposed classifier, which implements a new method for evaluating proximity based on linear programming.

This research is funded by RFBR, grants No 19-29-01009 and No 18-29-02036.

[1] *Nemirko A. P.* Lightweight nearest convex hull classifier // Pattern Recogn. Image Anal., 2019. Vol. 29. No 2. Pp. 360–365

Трехэтапная вычислительная технология оптимизации атомно-молекулярных кластеров Морса сверхбольших размерностей

Сороковиков Павел Сергеевич[1]⋆ sorokovikov.p.s@gmail.com
Горнов Александр Юрьевич[1] gornov@icc.ru
[1]Иркутск, Институт динамики систем и теории управления имени В.М. Матросова СО РАН

Одной из классических проблем вычислительной химии является задача поиска низкопотенциальных атомно-молекулярных кластеров. С математической точки зрения, задача сводится к поиску глобального оптимума невыпуклых потенциальных функций – специальных моделей. Основной сложностью указанного класса задач является астрономический рост числа локальных экстремумов с увеличением количества атомов. Несмотря на это, современные оптимизационные алгоритмы, запущенные на высокопроизводительной вычислительной технике, способны находить «наилучшие из известных» решения, которые, возможно, являются глобальными оптимумами.

Для решения задач оптимизации атомно-молекулярных кластеров реализована трехэтапная вычислительная технология, которая на каждом этапе включает различные методы из базового набора. В основе первого этапа лежит стохастическая аппроксимация окрестности известной рекордной точки с помощью коллекции алгоритмов-генераторов разнохарактерных начальных приближений. На втором этапе решается задача быстрого спуска из полученных на первом этапе приближений, для решения которой применяется один из «алгоритмов-стартеров» – алгоритмов, позволяющих, в действительности, находить уже низкопотенциальные кластеры с минимальными вычислительными затратами. Реализованы четыре варианта «алгоритмов-стартеров»: модификация метода Б.Т. Поляка 1969 г., декомпозиционный градиентный метод («рейдер-метод»), модификация многомерного дихотомического метода, генетический алгоритм. Для уточняющих локальных спусков, реализуемых на третьем этапе, применяется квазиньютоновский метод L-BFGS.

В работе рассматривается задача поиска низкопотенциальных кластеров Морса сверхбольших размерностей. Целевая функция имеет следующий вид: $f(x) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} e^{\rho(1-r_{ij})} \left(e^{\rho(1-r_{ij})} - 2\right)$, где r_{ij} – расстояние между частицами i и j, N – число атомов. Выполнены системные вычислительные эксперименты по поиску низкопотенциальных состояний кластеров Морса с размерностями от 301 до 330 атомов с шагом 1 при $\rho = 3$ (см. Таблицу 1).

Сравнительный анализ результатов экспериментов не выявил резких отклонений от наблюдаемой закономерности найденных значений потенциалов, описывающей их рост в зависимости от числа атомов. Сравнения с расчетами других авторов провести, к сожалению, не удалось по причине отсутствия таковых.

Таблица 1. Наилучшие найденные значения (от 301 до 330 атомов)

N	Значение	N	Значение
301	-3742.786756500	316	-3971.398067351
302	-3756.943941215	317	-3985.575953832
303	-3771.562170128	318	-4000.902271527
304	-3786.863828202	319	-4016.001113765
305	-3803.364623270	320	-4031.192528437
306	-3818.985147947	321	-4046.184649397
307	-3831.900251012	322	-4061.074419771
308	-3848.610152404	323	-4077.757995930
309	-3863.643211031	324	-4093.396925659
310	-3879.968258938	325	-4109.568207462
311	-3894.091767219	326	-4124.693731294
312	-3908.579015736	327	-4137.002740071
313	-3926.825309287	328	-4155.545778724
314	-3939.085618390	329	-4170.595348250
315	-3955.606154874	330	-4185.823810024

Работа поддержана грантом РФФИ № 18-07-00587.

[1] *Sorokovikov P., Gornov A., Anikin A.* Computational technology for the study of atomic-molecular Morse clusters of extremely large dimensions // IOP Conference Series: Materials Science and Engineering, 2020. Vol. 734. No. 1.

Three-stage computational technology for optimization of atomic-molecular Morse clusters of extremely large dimensions

Pavel Sorokovikov[1]⋆ sorokovikov.p.s@gmail.com
Alexander Gornov[1] gornov@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of SB RAS

One of the classical problems in computational chemistry is the problem of finding low-potential atomic-molecular clusters. From a mathematical point of view, we can consider these problems as the problem of searching for a global minimum of non-convex potential functions – specific models. The main difficulty of this class of problems is the astronomical increase in the number of local extremums with increasing dimension. However, modern optimization algorithms running on high-performance computing techniques can find the "best of known" solutions that may be global optima.

Table 2. The best-found values (from 301 to 330 atoms)

N	Value	N	Value
301	-3742.786756500	316	-3971.398067351
302	-3756.943941215	317	-3985.575953832
303	-3771.562170128	318	-4000.902271527
304	-3786.863828202	319	-4016.001113765
305	-3803.364623270	320	-4031.192528437
306	-3818.985147947	321	-4046.184649397
307	-3831.900251012	322	-4061.074419771
308	-3848.610152404	323	-4077.757995930
309	-3863.643211031	324	-4093.396925659
310	-3879.968258938	325	-4109.568207462
311	-3894.091767219	326	-4124.693731294
312	-3908.579015736	327	-4137.002740071
313	-3926.825309287	328	-4155.545778724
314	-3939.085618390	329	-4170.595348250
315	-3955.606154874	330	-4185.823810024

To solve the problems of optimizing atomic-molecular clusters, a three-stage computational technology has been implemented, which includes various methods from a set of basic at each stage. The first stage is focused on the stochastic approximation of a neighborhood of a known record point and relies on a set of algorithms-generators of various initial approximations. At the second stage, the problem of fast descent of the approximations generated at the first stage is solved using one of the "starter algorithms" – algorithms that allow finding low-potential clusters with minimal com-

putational costs. Four variants of "starter algorithms" have been implemented: the modification of B.T. Polyak's method, the decomposition gradient method ("raider method"), the variant of the multivariate dichotomy method and genetic algorithm. For local descents implemented in the third stage, the quasi-Newtonian L-BFGS method is used.

The paper considers the problem of finding low-potential Morse clusters of extremely large dimensions. The objective function is as follows: $f(x) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} e^{\rho(1-r_{ij})} \left(e^{\rho(1-r_{ij})} - 2\right)$, where r_{ij} is the distance between particles i and j, N is the number of atoms. System computational experiments were performed to search for low-potential states of Morse clusters with dimensions from 301 to 330 atoms with a step of 1, $\rho = 3$ (see Table 1).

Comparative analysis of the experimental results did not reveal sharp deviations from the observed regularity of the found values of potentials, which describes their growth depending on the number of atoms. Unfortunately, it was not possible to make comparisons with the computations of other authors due to the absence of such.

This research is funded by RFBR, grant 18-07-00587.

[1] *Sorokovikov P., Gornov A., Anikin A.* Computational technology for the study of atomic-molecular Morse clusters of extremely large dimensions // IOP Conference Series: Materials Science and Engineering, 2020. Vol. 734. No. 1.

Численное решение задач минимизации потенциала Китинга с размерностями до 300 миллионов переменных

Аникин Антон Сергеевич[1]⋆ anikin@icc.ru

[1]Иркутск, Институт динамики систем и теории управления имени В.М. Матросова СО РАН

Рассматривается задача минимизации потенциала Китинга, предложенного ещё в 1966 [1] для моделирования деформаций, возникающих в квантовых точках «кремний-германий». Несмотря на весьма «почтенный» возраст эта модель до сих пор используется прикладными специалистами как в её исходном виде [2], так и в упрощённом (что, очевидно, связано с определёнными сложностями, возникающими при поиске решения) варианте [3]. Поэтому задача разработки эффективных и надёжных численных алгоритмов, способных работать с большими и сверхбольшими постановками продолжает оставаться актуальной.

В работе исследуется практическая эффективность различных алгоритмов унимодальной оптимизации, включая квазиньютоновский метод LBFGS и разные варианты метода сопряжённых градиентов. Изучено влияние различных алгоритмов одномерного (линейного) поиска на скорость сходимости и «качество работы» выбранных методов оптимизации. Выполнена параллельная (технология OpenMP) программная реализация предложенных алгоритмов. Представлены результаты численного решения для различных кристаллических структур типа «кремний-германий», содержащих до 100 миллионов атомов ($3 \cdot 10^8$ оптимизируемых переменных).

Работа поддержана грантом РФФИ № 18-07-00587.

[1] *Keating P. N.* Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure // Phys. Rev., 1966. Vol. 145. Pp. 637–645.

[2] *Yakimov A.I. Bloshkin A.A. Dvurechenskii A.V.* Calculating of energy spectrum and electronic structure of two holes in a pair of coupled Ge/Si quantum dots // Phys. Rev., 2010. Vol. 81. Pp. 1–11.

[3] *Давыдов С.Ю.* Простой модельный потенциал для описания упругих свойств однослойного графена // Физика твердого тела, 2013. Т. 55. № 4. С. 813–815.

Numerical solution of Keating potential minimization problems with dimensions up to 300 million variables

Anton Anikin[1]⋆ anikin@icc.ru

[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of SB RAS

We consider the problem of minimizing the Keating potential, which was proposed in 1966 [1] for modeling deformations that occur in "silicon-germanium" quantum dots. Despite its very "venerable" age, this model is still used by applied specialists both in its original form [2] and in a simplified version (which is obviously due to certain difficulties that arise when searching for a solution) [3]. Therefore, the task of developing efficient and reliable numerical algorithms that can work with large and huge-scale problems continues to be relevant.

The paper examines the practical effectiveness of various unimodal optimization algorithms, including the quasi-Newtonian LBFGS method and different versions of the conjugate gradient method. The influence of various line-search (one-dimensional) algorithms on the convergence rate and "convergence quality" of the selected optimization methods is studied. Parallel software implementation (OpenMP technology) of the proposed algorithms is performed. The results of a numerical solution for various "silicon-germanium" crystal structures containing up to 100 million atoms ($3 \cdot 10^8$ optimized variables) are presented.

This research is funded by RFBR, grant 18-07-00587.

[1] *Keating P. N.* Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure // Phys. Rev., 1966. Vol. 145. Pp. 637–645.

[2] *Yakimov A.I. Bloshkin A.A. Dvurechenskii A.V.* Calculating of energy spectrum and electronic structure of two holes in a pair of coupled Ge/Si quantum dots // Phys. Rev., 2010. Vol. 81. Pp. 1–11.

[3] *Davydov S.Yu.* Simple model potential for describing elastic properties of single-layer graphene // Physics of the Solid State, 2013. Vol. 55. No 4. Pp. 813–815.

Оптимизационные методы численного решения систем линейных интервальных уравнений, связанных с задачами построения линейных зависимостей при интервальной неопределенности данных

Ерохин Владимир Иванович[1]⋆ erohin_v_i@mail.ru
Кадочников Андрей Павлович[1] kado162@mail.ru
Сотников Сергей Владимирович[1] svsotnikov66@gmail.com
Маркина Мария Константиновна[1] kravshenok@mail.ru

[1]Санкт-Петербург, Военно-космическая академия имени А.Ф. Можайского

Пусть задана следующая совокупность условий

$$Ax = b, \tag{1}$$

$$\underline{A} \leqslant A \leqslant \bar{A}, \tag{2}$$

$$\underline{b} \leqslant b \leqslant \bar{b}, \tag{3}$$

где $\underline{A}, \bar{A} \in \mathbb{R}^{m\times n}$ – заданные матрицы; $\underline{b}, \bar{b} \in \mathbb{R}^m$ – заданные векторы, такие что $\underline{A} \leqslant \bar{A}$, $\underline{b} \leqslant \bar{b}$; $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ – неизвестные (подлежащие определению) матрица и векторы.

В наиболее содержательном случае $\underline{A} \neq \bar{A}$, $\underline{b} \neq \bar{b}$, условия (1)-(3) задают *интервальную систему линейных алгебраических уравнений* (ИСЛАУ) общего вида и задачу поиска её *слабого решения* (A, x, b) [1].

Эквивалентное (1)-(3) представление ИСЛАУ может быть записано с помощью *средней матрицы* $A_c = \frac{1}{2}\left(\underline{A} + \bar{A}\right)$, *матрицы радиусов* $A_r = \frac{1}{2}\left(\bar{A} - \underline{A}\right)$, *среднего вектора* $b_c = \frac{1}{2}\left(\underline{b} + \bar{b}\right)$ и *вектора радиусов* $b_r = \frac{1}{2}\left(\bar{b} - \underline{b}\right)$ [1].

ИСЛАУ представляются естестественной моделью построения линейных зависимостей по данным, обладающим интервальной неопределенностью (см., например [2, 3, 4]).

Численные методы решения ИСЛАУ представляют значительный интерес. В докладе предполагается обсудить подходы, являющиеся развитием методов, изложенных в работах [5, 6, 7, 8]. Пусть исследуемая ИСЛАУ такова, что ее «центральная» система $A_c x = b_c$ несовместна, а матрица и вектор радиусов имеют специальный вид $A_r = \mu \cdot 1_m 1_n^\top$, $b_r = \delta \cdot 1_m$, где $\mu, \delta > 0$ – некоторые (малые) константы, $1_m \in \mathbb{R}^m$ и $1_n \in \mathbb{R}^n$ – векторы, составленные из единиц.

Рассмотрим задачу

$$Z: \quad \begin{array}{c} \text{Найти } A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, x \in \mathbb{R}^n \text{ такие, что} \\ \|A_c - A\|_{\ell_\infty} \leqslant \mu, \|b_c - b\|_\infty \leqslant \delta, \\ \text{система } Ax = b \text{ совместна}, \|x\|_1 \to \min, \end{array}$$

где $\|\cdot\|_1$, $\|\cdot\|_\infty$ – векторные нормы Гёльдера с показателями 1 и ∞, $\|A = (a_{ij})\|_{\ell_\infty} = \max\limits_{i,j} |a_{ij}|$ – норма Гёльдера с показателем ∞ на множестве матриц размера $m \times n$ [9].

Задача Z, с одной стороны, является одним из возможных способов формализации проблемы поиска слабого решения описанной выше ИСЛАУ специального вида. С другой стороны, задача Z представляет собой частный случай из класса задач, являющихся обобщением регуляризованного метода наименьших квадратов А.Н. Тихонова, использующего евклидовы векторные и матричные нормы [10], на случай использования полиэдральных норм. Эволюцию математического аппарата, применяемого для решения указанного класса задач, можно проследить по работам [5, 6, 7]. В явном виде задача Z, в числе некоторых других, рассмотрена в работе [8]. Справедлива следующая

Теорема 1. *Задача Z разрешима тогда и только тогда, когда разрешима задача*

$$R : \|b_c - A_c x\|_\infty \leqslant \mu\|x\|_1 + \delta, \|x\|_1 \to \min .$$

Если $\widehat{x}$ – решение задачи R, то $\widehat{x}$, $\widehat{A}$, $\widehat{b}$,
где $\widehat{A} = A_c + \mu \frac{b_c - A_c\widehat{x}}{\mu\left\|\widehat{x}\right\|_1 + \delta} \mathrm{diag}(\mathrm{sign}(\widehat{x})) \cdot 1_n^\top$, $\widehat{b} = b_c + \delta \frac{A_c\widehat{x} - b_c}{\mu\left\|\widehat{x}\right\|_1 + \delta}$ – решение задачи Z ($\mathrm{sign}(\cdot)$ – функция поэлементного вычисления знаков элементов вектора).

Теорема 1 закладывает основу для конструирования (оптимизационных) численных методов точного или приближенного решения задачи Z, которые и предполагается обсудить в докладе.

[1] *Фидлер М., Недома Й., Рамик Я., Рон И., Циммерман К.* Задачи линейной опимизации с неточными данными // Институт компьютерных исследований, 2008. С. 288.

[2] *Панюков А. В.*, *Голодов В. А.* Программная реализация алгоритма решения системы линейных алгебраических уравнений с интервальной неопределенностью в исходных данных // Управление большими системами, 2013. № 43. С. 78–94.

[3] *Шарый С. П.* Метод максимума согласования для восстановления зависимостей по данным с интервальной неопределенностью // Известия РАН. Теория и системы управления, 2017. № 6. С. 3–19.

[4] *Шарый С. П.* Задача восстановления зависимостей по данным с интервальной неопределенностью // Заводская лаборатория. Диагностика материалов, 2020. Т. 86. № 1. С. 62–74.

[5] *Горелик В. А.*, *Ерохин В. И.*, *Печенкин Р. В.* Минимаксная матричная коррекция несовместимых систем линеиных алгебраических уравнений с блочными матрицами коэффициентов // Известия РАН. Теория и системы управления, 2006. № 5. С. 52–62.

[6] *Волков В. В.*, *Ерохин В. И.*, *Какаев В. В.*, *Онуфрей А. Ю.* Обобщения регуляризованного метода наименьших квадратов Тихонова на векторные нормы, отличные от евклидовой // Журн. вычисл. матем. и матем. физ, 2017. Т. 57. № 9. С. 1433–1443.

[7] *Ерохин В. И.*, *Волков В. В.*, *Хвостов М. Н.* Модификация метода А.Н. Тихонова решения приближенных систем линейных алгебраических уравнений с использованием норм, отличных от евклидовой // Вестник ВГУ. Серия: Физика. Математика, 2018. № 4. С. 84–101.

[8] *Erokhin V.*, *Volkov V.*, *Krasnikov A.*, *Khvostov M.* Chapter Seventeen. Solving Approximate Systems ofLincar Algebraic Equations using Polyhedral Norms // In bk.: Recent Advances of the Russian Operations Research Society, 2020. С. 280–292.

[9] *Хорн Р.*, *Джонсон Ч.* Матричный анализ. М: Мир, 1989. 656 с.

[10] *Тихонов А. Н.* О приближенных системах линейных алгебраических уравнений // Журн. вычисл. матем. и матем. физ, 1980. Т. 20. № 6. С. 1373–1383.

Optimization methods for the numerical solution of systems of linear interval equations associated with the problems of constructing linear dependencies with interval data uncertainty

Vladimir Erokhin[1]⋆ erohin_v_i@mail.ru
Andrey Kadochnikov[1] kado162@mail.ru
Sergey Sotnikov[1] svsotinkov66@gmail.com
Maria Markina[1] kravshenok@mail.ru

[1]Saint-Petersburg, A.F. Mozhaysky Military-Space Academy

Let the following set of conditions be given

$$Ax = b, \tag{1}$$

$$\underline{A} \leqslant A \leqslant \bar{A}, \tag{2}$$

$$\underline{b} \leqslant b \leqslant \bar{b}, \tag{3}$$

where $\underline{A}, \bar{A} \in \mathbb{R}^{m\times n}$ is given matrices; $\underline{b}, \bar{b} \in \mathbb{R}^m$ are given vectors such that $\underline{A} \leqslant \bar{A}$, $\underline{b} \leqslant \bar{b}$; $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ are unknown (to be determined) matrix and vectors.

In the most meaningful occasion $\underline{A} \neq \bar{A}$, $\underline{b} \neq \bar{b}$, conditions (1)-(3) define *interval system of linear algebraic equation* (ISLAE) of general form and the problem of finding its *weak solution* (A, x, b) [1].

The equivalent (1)-(3) representation of ISLAE can be written using *middle matrix* $A_c = \frac{1}{2}\left(\underline{A} + \bar{A}\right)$, *radius matrices* $A_r = \frac{1}{2}\left(\bar{A} - \underline{A}\right)$, *middle vector* $b_c = \frac{1}{2}\left(\underline{b} + \bar{b}\right)$ and *radius vector* $b_r = \frac{1}{2}\left(\bar{b} - \underline{b}\right)$ [1].

ISLAE is a natural model for constructing linear dependencies from data with interval uncertainty (see for example [2, 3, 4]).

Numerical methods for solving ISLAE are of considerable interest. The report is supposed to discuss the approaches that are the development of the methods outlined in [5, 6, 7, 8]. Let the investigated ISLAE be such that its "central" system $A_c x = b_c$ is inconsistent, and the matrix and the vector of radii have a special form $A_r = \mu \cdot 1_m 1_n^\top$, $b_r = \delta \cdot 1_m$, where $\mu, \delta > 0$ are some (small) constants, $1_m \in \mathbb{R}^m$ and $1_n \in \mathbb{R}^n$ are vectors composed of ones.

Let's consider the problem

$$Z: \quad \begin{array}{c} \text{To find } A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, x \in \mathbb{R}^n \text{ such that} \\ \|A_c - A\|_{\ell_\infty} \leqslant \mu, \|b_c - b\|_\infty \leqslant \delta, \\ \text{system } Ax = b \text{ is compatible}, \|x\|_1 \to \min, \end{array}$$

where $\|\cdot\|_1$, $\|\cdot\|_\infty$ – are Hölder vector norms with exponents 1 and ∞, $\|A = (a_{ij})\|_{\ell_\infty} = \max\limits_{i,j} |a_{ij}|$ s the Hölder norm with exponent ∞ on the set of matrices of size $m \times n$ [9].

Problem Z, on the one hand, is one of the possible ways to formalize the problem of finding a weak solution to the above-described ISLAE of a special form. On the other hand, the problem Z is a special case from the class of problems that are a generalization of the regularized least squares method of A.N. Tikhonov, using Euclidean vector and matrix norms [10], to the case of using polyhedral norms. The evolution of the mathematical apparatus used to solve the specified class of problems can be traced back to papers [5, 6, 7]. Explicitly, the problem Z, among some others, is discussed in the paper [8]. The following theorem is true.

Theorem 1. *The problem Z is solvable if and only if the problem*

$$R : \|b_c - A_c x\|_\infty \leqslant \mu\|x\|_1 + \delta, \|x\|_1 \to \min .$$

is solvable. If $\widehat{x}$ is a solution to problem R, then $\widehat{x}$, $\widehat{A}$, $\widehat{b}$, where $\widehat{A} = A_c + \mu\frac{b_c - A_c\widehat{x}}{\mu\left\|\widehat{x}\right\|_1 + \delta}\mathrm{diag}(\mathrm{sign}(\widehat{x})) \cdot 1_n^\top$, $\widehat{b} = b_c + \delta\frac{A_c\widehat{x} - b_c}{\mu\left\|\widehat{x}\right\|_1 + \delta}$ is the solution to problem Z ($\mathrm{sign}(\cdot)$ is a function of elementwise calculation of signs of vector elements).

Theorem 1 lays the foundation for constructing numerical (optimization) methods of exact or approximate solution of the problem Z, which are supposed to be discussed in the talk.

[1] *Fiedler M., Nedoma J., Ramik J., Rohn J., Zimmerman K.* Linear optimization problems with inexact data // Springer, 2006. — 224 p.

[2] *Panyukov A., Golodov V.* Software implementation of algorithm for solving a set of linear equations under interval uncertainty // UBS, 2013. No 43. Pp. 78–94.

[3] *Shary S.* Maximum compatibility method for data fitting under interval uncertainty // Journal of Computer and Systems Sciences International, 2017. Vol. 56. No 6. Pp. 897–913.

[4] *Shary S.* Data fitting problem under interval uncertainty in data // Industrial Laboratory. Diagnostics of Materials, 2020. Vol. 86. No 1. Pp. 62–74.

[5] *Gorelik V., Erokhin V., Pechenkin R.* Minimax Matrix Correction of Inconsistent Systems of Linear Algebraic Equations with Block Matrices of Coefficients // Journal of Computer and Systems Sciences International, 2006. Vol. 45. No 5. Pp. 727–737.

[6] *Volkov V., Erokhin V., Kakaev V., Onufrei A.* Generalizations of Tikhonov's Regularized Method of Least Squares to Non-Euclidean Vector Norms // Computational Mathematics and Mathematical Physics, 2017. Vol. 57. No 9. Pp. 1416–1426.

[7] *Erokhin V., Volkov V., Khvostov M.* Modification of the A.N. Tikhonov's method for solving approximate systems of linear algebraic equations for non-euclidean norms //Proceedings of Voronezh State University. Series: Physics. Mathematics, 2018. No. 4. Pp. 84–101.

[8] *Erokhin V., Volkov V., Krasnikov A., Khvostov M.* Chapter Seventeen. Solving Approximate Systems of Lincar Algebraic Equations using Polyhedral Norms // In bk.: Recent Advances of the Russian Operations Research Society, 2020. Pp. 280–292.

[9] *Horn R., Johnson Ch.* Matrix analysis // Cambridge: Cambridge University Press, 1985. Pp. 561.

[10] *Tikhonov A.* Approximate systems of linear algebraic equations // USSR Comput. Math. Math. Phys., 1980. Vol. 20. No 6. Pp. 1373–1383.

Методика стресс-тестирования программных комплексов для оптимизации нелинейных управляемых динамических систем

Зароднюк Татьяна Сергеевна[1]⋆ tz@icc.ru
Горнов Александр Юрьевич[1] gornov@icc.ru
[1]Иркутск, ИДСТУ СО РАН

Основным методом экспериментальной оценки эффективности алгоритмов оптимизации и соответствующих программных комплексов является тестирование. Традиционное сравнительное тестирование подразумевает поиск решения задачи разными методами с целью дальнейшего сравнения полученных результатов. В работе предложена альтернативная методика – стресстестирование комплексов программ для решения задач оптимального управления.

Стресс-тестирование направлено на получение информации о предельных свойствах исследуемых алгоритмов. Основой методики служит специальный набор тестовых примеров, опирающийся на использование семейства функций (Neumaier A. Rational functions with prescribed global and local minimizers), позволяющих моделировать особенности, характерные для задач глобальной оптимизации – существование большого числа локальных экстремумов, наличие экстремумов с узкой областью притяжения, близость локального и глобального экстремумов и другие. С его использованием сформирован ряд задач оптимального управления различного уровня сложности с известным множеством достижимости и глобальным экстремумом.

Разработанные тесты включены в тестовую коллекцию невыпуклых задач оптимального управления, использованную для тестирования семейства программных комплексов OPTCON, ориентированных на исследование сложных невыпуклых задач оптимального управления.

Работа поддержана грантом РФФИ No 18-07-00587.

[1] *Gornov A. Yu., Zarodnyuk T. S., Anikin A. S., Finkelstein E. A.* Extension technology and extrema selections in a stochastic multistart algorithm for optimal control problems // Journal of Global Optimization, 2020, Vol. 76, No 3. Pp. 533–543.

Stress testing technique of numerical investigating software for optimization of nonlinear controlled dynamical systems

Tatiana Zarodnyuk[1]⋆ tz@icc.ru
Alexander Gornov[1] gornov@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory SB RAS

The main method of the experimental evaluation of the effec- tiveness of the optimization algorithms and corresponding software is testing. The traditional comparative testing involves finding the solution of the problem using different methods in order to further compare the obtained results. The paper proposes an alternative technique which is the stress testing of the software for solving opti- mal control problems.

Stress testing is aimed at obtaining information about the limiting properties of the studied algorithms. The methodology is based on a special set of test examples based on using the functions family (Neumaier A. Rational functions with prescribed global and local minimizers: http://solon.cma.univie.ac.at), which allow simulating the typical features for the global optimization problems, such as, the existence of a large number of local extrema, the presence of extrema with a narrow attraction region, the proximity of the local and global extrema, and others. With its use, a number of optimal control problems of various complexity levels with known reachable set and global extremum have been constructed.

The developed tests are included in the test collection of nonlinear controlled dynamical systems, used to numerical investigating the family of OPTCON software applied for the complex optimal control problems.

This research is funded by RFBR, grant 18-07-00587.

[1] *Gornov A. Yu., Zarodnyuk T. S., Anikin A. S., Finkelstein E. A.* Extension technology and extrema selections in a stochastic multistart algorithm for optimal control problems // Journal of Global Optimization, 2020, Vol. 76, No 3. Pp. 533–543.

Линейная сходимость в гладкой выпуклой задаче min-min с сильной выпуклостью по одной из групп переменных

Гладин Егор Леонидович[1,2]⋆ gladin.el@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Сколтех

Рассматривается задача вида

$$\min_{x\in\mathbb{R}^n}\min_{y\in Q} F(x,y), \tag{1}$$

где функция $F(x,y)$ — выпуклая и L-гладкая по совокупности переменных, а также μ-сильно выпуклая по x. Такая постановка возникает, например, при поиске равновесий в транспортных сетях [1]. В машинном обучении задачи такого типа соответствуют случаю, когда регуляризация осуществляется по одной из двух групп параметров модели (сильная выпуклость по x). Особенность предлагаемого в работе решения заключается в том, что удаётся получить линейную скорость сходимости несмотря на отсутствие сильной выпуклости по одной из групп переменных.

Пусть $Q \subset \mathbb{R}^d$ — выпуклое компактное множество, $n \gg d$. Введём функцию

$$f(x) = \min_{y\in Q} F(x,y), \tag{2}$$

которая также будет L-гладкой ([1], утверждение 2.3.3). Тогда можно переписать задачу (6) следующим образом:

$$f(x) \to \min_{x\in\mathbb{R}^n} \tag{3}$$

При решении (8) некоторым итерационным методом необходимо на каждом его шаге решать вспомогательную задачу (7). Это предлагается делать с помощью метода Вайды [2], имеющего сложность $O(d\log\frac{dr}{\varepsilon})$, где r — радиус множества Q. Стоимость итерации алгоритма, помимо вычисления градиента по y, равна $O(d^\omega)$, где константа $\omega \in (2,3)$ определяется сложностью перемножения матриц. Поскольку этот алгоритм сходится за линейное время, то можно считать, что вспомогательная задача (7) решается сколь угодно точно. В силу того, что размерность y относительно небольшая, зависимость числа итераций от d не вызывает затруднений.

Внешнюю задачу (8) можно решать быстрым градиентным методом (БГМ) Нестерова. В силу L-гладкости и μ-сильной выпуклости f, этот метод обеспечивает верхнюю оценку на число итераций $O\left(\sqrt{\frac{L}{\mu}}\log\frac{R^2}{\varepsilon}\right)$, где $R := \|x^0 - x^*\|_2$ — расстояние от начального приближения до решения задачи (8).

Итоговая оценка сложности предлагаемого метода составляет $\tilde{O}(d\sqrt{L/\mu})$ вычислений $\nabla_y F$ и $\tilde{O}(\sqrt{L/\mu})$ вычислений $\nabla_x F$, где $\widetilde{O}(\cdot)$ означает $O(\cdot)$ с точностью

до логарифмического по ε^{-1} множителя в степени 1 или 2. Примечательно, что тут имеет место линейная сходимость, хотя исходная функция сильно выпуклая только по одной группе переменных — x. Цена, которую приходится за это заплатить — множитель d в оценке числа вызовов $\nabla_y F$.

Если гладкость F имеет место только по переменным x, то нужно использовать другой порядок взятия $\min\limits_x \min\limits_y$:

$$g(y) = \min_{x \in \mathbb{R}^n} F(x, y), \tag{4}$$

$$g(y) \to \min_{y \in Q} \tag{5}$$

Тогда метод Вайды будет применяться к внешней задаче (10), и оценки изменятся таким образом: $\tilde{O}(d)$ вычислений $\nabla_y F$ и $\tilde{O}(d\sqrt{L/\mu})$ вычислений $\nabla_x F$. Какой вариант лучше, зависит от сложности оракулов. Учитывая, что $\dim x \gg \dim y$, то и стоимость вычисления $\nabla_x F$ часто может превышать стоимость вычисления $\nabla_y F$. Впрочем, иногда это различие не заметно — $\ln(\sum_{k=1}^n \exp(\langle a_k, x\rangle - b_k) + + \sum_{k=1}^m \exp(\langle c_k, y\rangle - d_k))$.

Особый интерес для приложений с машинном обучении представляет случай, когда $F(x, y)$ имеет вид суммы большого числа функций (обозначим это число за m). Тогда обычный БГМ требует вычисления m градиентов на каждом шаге, что может быть трудозатратно при больших m. В такой ситуации лучше применять к (9) ускоренный градиентный метод с редукцией дисперсии [3], который также имеет оптимальную (линейную) скорость сходимости:

$$N = O\left(m \log m + \sqrt{\frac{mL}{\mu}} \log \frac{D_0}{\varepsilon}\right),$$

$$\text{где } D_0 := 2\left(f\left(x^0\right) - f\left(x^*\right)\right) + \frac{3}{2} L R^2.$$

В таком случае итогая оценка сложности составляет

$$\tilde{O}\left(d \cdot \left(m + \sqrt{\frac{mL}{\mu}}\right)\right) \text{ обращений к оракулу } \nabla_x F,$$

$$\tilde{O}\left(dm\right) \text{ обращений к оракулу } \nabla_y F.$$

Повторим, что гладкость по y для получения этих оценок не требуется.

[1] *Гасников А. В., Гасникова Е. В.* Модели равновесного распределения транспортных потоков в больших сетях: учебное пособие // Москва: МФТИ, 2020.

[2] *Vaidya P.* A new algorithm for minimizing convex functions over convex sets // Mathematical Programming, 1996. Pp. 291–341.

[3] *Lan G.* First-order and Stochastic Optimization Methods for Machine Learning // Atlanta: Springer, 2020.

Linear convergence for smooth convex min-min problem with strong convexity in one of the groups of variables

Egor Gladin[1,2]⋆ gladin.el@phystech.edu

[1]Moscow, Moscow Institute of Physics and Technology

[2]Moscow, Skoltech

Consider the problem of the form

$$\min_{x\in\mathbb{R}^n}\min_{y\in Q} F(x,y), \tag{6}$$

where function $F(x,y)$ is convex and L-smooth in (x,y), and also μ-strongly convex in x. Such problems arise, for example, in traffic assignment models [1]. In machine learning problems of this type correspond to the case when regularization is applied to one of the two groups of model's parameters (hence strong convexity in x). Advantage of the approach presented below is that it achieves linear convergence despite the fact that target function is not strongly convex in one of the variables.

Let $Q \subset \mathbb{R}^d$ be convex compact set, $n \gg d$. Consider function

$$f(x) = \min_{y\in Q} F(x,y), \tag{7}$$

which is also L-smooth (see [1], proposition 2.3.3). We can rewrite (6) as follows:

$$f(x) \to \min_{x\in\mathbb{R}^n} \tag{8}$$

When using some iterative method to solve (8), we have to find a solution of the auxiliary problem (7). We propose to use Vaidya's cutting plane method [2] for this purpose. The method has complexity $O(d\log\frac{dr}{\varepsilon})$, where r is the radius of Q. The cost of iteration (in addition to the cost of finding a gradient w.r.t. y) is $O(d^\omega)$, where $\omega \in (2,3)$ corresponds to complexity of matrix multiplication. Due to the fact that this algorithm converges linearly, we can assume that the auxiliary problem (7) can be solved with arbitrary precision. Since dimension of y is relatively small, dependence on d doesn't cause difficulties.

The outer problem (8) can be solved via Nesterov's Fast Gradient Method (FGM). As f is L-smooth and μ-strongly convex, FGM results in upper bound on number of iterations $O\left(\sqrt{\frac{L}{\mu}}\log\frac{R^2}{\varepsilon}\right)$, where $R := \|x^0 - x^*\|_2$ is distance between initial point and solution of (8).

The final complexity consists of $\tilde{O}(d\sqrt{L/\mu})$ computations of $\nabla_y F$ and $\tilde{O}(\sqrt{L/\mu})$ computations of $\nabla_x F$, where $\widetilde{O}(\cdot)$ means $O(\cdot)$ up to logarithmic in ε^{-1} factor to the power of 1 or 2. It is remarkable that we have linear convergence despite the fact that target function is strongly convex only in one of the groups of variables — x. The price we pay for this is factor d in the number of calls to $\nabla_y F$.

If F is smooth only in x, we can use the reverse order of $\min\limits_x \min\limits_y$:

$$g(y) = \min_{x \in \mathbb{R}^n} F(x, y), \tag{9}$$

$$g(y) \to \min_{y \in Q} \tag{10}$$

This way Vaidya's method is applied to the outer problem (10), and complexity estimates are as follows: $\tilde{O}(d)$ computations of $\nabla_y F$ and $\tilde{O}(d\sqrt{L/\mu})$ computations of $\nabla_x F$. Which way is the best, depends on oracles' complexities. Due to the fact that $\dim x \gg \dim y$, the cost of computing $\nabla_x F$ may often exceed the cost of computing $\nabla_y F$. However, sometimes this difference is not significant — $\ln(\sum_{k=1}^n \exp(\langle a_k, x\rangle - - b_k) + \sum_{k=1}^m \exp(\langle c_k, y\rangle - d_k))$.

The case when $F(x, y)$ is a sum of a large number of functions (we denote this number by m) is of particular interest because of applications in machine learning. Then, the usual FGM requires computing m gradients at each step, which is tedious for large m. In this case it's better to use Variance-Reduced Accelerated Gradient Descent [3], which also has optimal (linear) convergence rate:

$$N = O\left(m \log m + \sqrt{\frac{mL}{\mu}} \log \frac{D_0}{\varepsilon}\right),$$

$$\text{where } D_0 := 2\left(f\left(x^0\right) - f\left(x^*\right)\right) + \frac{3}{2} L R^2.$$

Thus, the final complexity is

$$\tilde{O}\left(d \cdot \left(m + \sqrt{\frac{mL}{\mu}}\right)\right) \text{ calls to } \nabla_x F \text{ oracle},$$

$$\tilde{O}\left(dm\right) \text{ calls to } \nabla_y F \text{ oracle}.$$

Recall that smoothness in y is not required for these complexity estimates.

[1] *Gasnikov A., Gasnikova E.* Traffic assignment models. Numerical aspects (in Russian) // Moscow: MIPT, 2020.

[2] *Vaidya P.* A new algorithm for minimizing convex functions over convex sets // Mathematical Programming, 1996. Pp. 291–341.

[3] *Lan G.* First-order and Stochastic Optimization Methods for Machine Learning // Atlanta: Springer, 2020.

Ускорение стохастических методов на примере децентрализованного SGD

Тримбач Екатерина Алексеевна[1]⋆ trimbach.ea@phystech.edu
Рогозин Александр Викторович[1] aleksandr.rogozin@phystech.edu
[1]Москва, Московский физико-технический институт

Рассмотрим задачу децентрализованной оптимизации, а именно поиска минимум суммы функций

$$x^\star = argmin_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\right]$$

где каждая f_i имеет вид $f_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i} F_i(\mathbf{x},\xi_i)$ и каждая из функций f_i хранится на отдельном узле некоторой вычислительной сети. В работе рассматривается способ ускорения децентрализованного градиентного спуска, который гарантированно находит минимум с точностью ε за число итераций

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu n\varepsilon} + \frac{\sqrt{L}\tau(\bar{\sigma}\frac{\sqrt{p}}{\sqrt{\tau}} + \bar{\zeta})}{\mu p\sqrt{\varepsilon}} + \frac{\tau L}{p\mu}\right)$$

L, μ -константы выпуклости и сильной выпуклости.
p, τ - константы графа сети. p - коэффициент сжатия, т.е так что

$$\mathbb{E}_W\|XW - \bar{X}\|_F^2 \leqslant (1-p)\|X - \bar{X}\|_F^2$$

W - матрица графа, описывающего сеть, $\bar{X} := X\cdot\frac{1}{n}11^\top$. τ - число итераций, за которые гарантированно данное сжатие.
Основным результатом работы является ускорение алгоритма децентрализованного стохастического градиентного спуска для точности ε до числа итераций

$$\tilde{O}\left(\frac{\bar{\sigma}^2}{\mu n\varepsilon} + \frac{L^{\frac{1}{4}}(\bar{\zeta}\tau + \bar{\sigma}\sqrt{p\tau})}{\mu^{\frac{3}{4}}p\sqrt{\varepsilon}} + \frac{\sqrt{L}\tau}{\sqrt{\mu}p}\right)$$

[1] *Trimbach E., Rogozin A.* Acceleration of stochastic methods on the example of decentralized SGD // https://arxiv.org/abs/2011.07585

Acceleration of stochastic methods on the example of decentralized SGD

Ekaterina Trimbach[1]⋆ trimbach.ea@phystech.edu
Alexander Rogozin[1] aleksandr.rogozin@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology

Consider the problem of decentralized optimization, namely, finding the minimum sum of functions

$$x^\star = argmin_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\right]$$

where each f_i has the form

$$f_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i} F_i\left(\mathbf{x}, \xi_i\right)$$

and each of the functions f_i is stored on a separate node of some computer network. The paper considers a way to accelerate decentralized gradient descent, which is guaranteed to find the minimum with an accuracy of ε for the number of iterations

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu n\varepsilon} + \frac{\sqrt{L}\tau(\bar{\sigma}\frac{\sqrt{p}}{\sqrt{\tau}} + \bar{\zeta})}{\mu p\sqrt{\varepsilon}} + \frac{\tau L}{p\mu}\right)$$

L, μ -constants of convexity, and strong convexity. p, τ - network graph constants. p - compression ratio, i.e. such that

$$\mathbb{E}_W\|XW - \bar{X}\|_F^2 \leqslant (1-p)\|X - \bar{X}\|_F^2$$

W is a graph matrix describing the network, $\bar{X} := X \cdot \frac{1}{n}11^\top$.
τ - number of iterations for which this compression is guaranteed.
The main result of the work is the acceleration of the decentralized stochastic gradient descent algorithm for accuracy ε to the number of iterations

$$\tilde{O}\left(\frac{\bar{\sigma}^2}{\mu n\varepsilon} + \frac{L^{\frac{1}{4}}(\bar{\zeta}\tau + \bar{\sigma}\sqrt{p\tau})}{\mu^{\frac{3}{4}} p\sqrt{\varepsilon}} + \frac{\sqrt{L}\tau}{\sqrt{\mu}p}\right)$$

[1] *Trimbach E., Rogozin A.* Acceleration of stochastic methods on the example of decentralized SGD // https://arxiv.org/abs/2011.07585

Самонастраивающийся алгоритм поиска с чередующимися окрестностями для почти оптимального решения задачи кластеризации k-средних

Казаковцев Лев Александрович * levk@bk.ru
Рожнов Иван Павлович ris2005@mail.ru
Попов Алексей Михайлович vm_popov@sibsau.ru
Товбис Елена Михайловна sibstu2006@rambler.ru

Красноярск, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева

Задача k-средних — одна из самых популярных моделей в машинном обучении без учителя, в которой минимизируется сумма квадратов расстояний (также называемая суммой квадратов ошибок, SSE — sum of squared errors) от группируемых объектов $A_1, ..., A_N \in \mathbb{R}^n$ до искомых центров кластеров (центроидов) $X_1, ..., X_k \in \mathbb{R}^n$:

$$SSE(X_1, ..., X_k) = \sum_{i=1}^{N} \min_{X \in \{X_1, ..., X_k\}} ||A_i - X||^2 \to \min_{X_1, ..., X_k \in \mathbb{R}^n}. \quad (1)$$

Здесь k должно быть известно заранее.

Простота алгоритмической реализации побуждает исследователей применять эту модель в различных инженерных и научных областях. Тем не менее доказано, что проблема NP-трудна, что делает точные алгоритмы неприменимыми для крупномасштабных задач, а самые простые и популярные алгоритмы приводят к очень высоким значениям суммы квадратов расстояний.

Процедура Ллойда, самый популярный алгоритм для задачи k-средних, довольно быстра. Тем не менее, для конкретных наборов данных, включая географические/геометрические данные, этот алгоритм приводит к решению, которое очень далеко от значения глобального минимума целевой функции (1), и режим мультистарта лишь незначительно улучшает результат. Более точные методы кластеризации k-средних намного медленнее. Тем не менее, недавние достижения в области высокопроизводительных вычислений позволяют нам обрабатывать большой объем вычислений с использованием процедуры Ллойда, встроенной в более сложные алгоритмические схемы. Таким образом, очевидна потребность в алгоритмах кластеризации, представляющие собой компромисс между временем, затрачиваемым на вычисления, и результирующим значение целевой функции (1). Тем не менее, в некоторых случаях при решении задачи (1) требуется получить результат (значение целевой функции) за ограниченное фиксированное время, такое, чтобы его было бы сложно улучшить известными методами без значительного увеличения вычислительных затрат. Такие результаты требуются, если цена ошибки высока, а также для оценки более быстрых алгоритмов в качестве эталонных решений.

Если задача должна быть решена в течение ограниченного времени с максимально возможной точностью, которую было бы трудно повысить с использованием известных методов без увеличения вычислительных затрат, алгоритмы поиска с чередующимися окрестностями (VNS - Variable Neighborhood Search), которые ведут поиск в рандомизированных окрестностях, образованных применением жадных агломеративных процедур, являются конкурентоспособными.

Агломеративные процедуры, несмотря на их относительно высокую вычислительную сложность, могут быть успешно интегрированы в более сложные схемы поиска. Их можно использовать в составе VNS-алгоритмов. Более того, такие алгоритмы представляют собой компромисс между точностью решения и временными затратами. В нашем исследовании под точностью мы понимаем исключительно способность алгоритма (решателя) получать минимальные значения целевой функции (1). Такие процедуры начинаются с недопустимого решения с избыточным числом центроидов $k + r$. На каждой итерации такие процедуры устраняют один или несколько центроидов и улучшают результат с использованием процедуры Ллойда. Использование алгоритмов VNS, которые ищут в окрестностях, сформированных путем применения жадных агломерационных процедур к известному (текущему) решению, позволяет получить хорошие результаты за фиксированное время, приемлемое для интерактивных режимов работы. Выбор таких процедур, их последовательность и параметры оставались открытым вопросом. Эффективность таких процедур показана экспериментально на некоторых тестовых и практических задачах. Различные версии VNS-алгоритмов, основанные на жадных агломеративных процедурах, существенно различаются по своим результатам, что затрудняет использование таких алгоритмов в практических задачах. Практически невозможно спрогнозировать относительную производительность конкретного алгоритма VNS на основе таких обобщенных числовых характеристик задачи, как размер выборки и количество кластеров. Причем эффективность таких процедур зависит от их параметров. Однако вид и характер этой зависимости не были изучены.

В данной работе мы систематизируем подходы к построению алгоритмов поиска в окрестностях, сформированных с помощью жадных агломеративных процедур, и определяем окрестность $GREEDY_r$, где r - ее параметр (количество избыточных центроидов в промежуточном решении). Мы исследуем влияние наиболее важного параметра таких окрестностей (параметра r) на вычислительную эффективность и предлагаем новый алгоритм (вычислительное средство) на основе VNS, реализованный на графическом процессоре (GPU), который автоматически подстраивает этот параметр. Сравнительный анализ решения задач на наборах данных, содержащих до нескольких миллионов объектов, демонстрирует преимущество нового алгоритма по сравнению с известными алгоритмами локального поиска (включая поиск в окрестностях SWAP, таких как j-Means) в течение фиксированного времени, что позволяет проводить

онлайн-вычисления, при уровне значимости 0,01 (критерий Уилкоксона-Манна-Уитни).

Работа выполнена при поддержке Министерства науки и высшего образования РФ, проект № FEFE-2020-0013.

[1] *Kazakovtsev L.* Self-Adjusting Variable Neighborhood Search Algorithm for Near-Optimal k-Means Clustering // Computation, 2020, Vol. 8. No 4.

Self-Adjusting Variable Neighborhood Search Algorithm for Near-Optimal k-Means Clustering

Lev Kazakovtsev ⋆ levk@bk.ru
Ivan Rozhnov ris2005@mail.ru
Aleksey Popov vm_popov@sibsau.ru
Elena Tovbis sibstu2006@rambler.ru
Krasnoyarsk, Reshetnev Siberian State University of Science and Technology

The k-means problem is one of the most popular models in the unsupervised machine learning that minimizes the sum of the squared distances (also called sum of squared errors, SSE) from clustered objects $A_1, ..., A_N \in \mathbb{R}^n$ to the sought cluster centers (centroids) $X_1, ..., X_k \in \mathbb{R}^n$:

$$SSE(X_1, ..., X_k) = \sum_{i=1}^{N} \min_{X \in \{X_1, ..., X_k\}} ||A_i - X||^2 \to \min_{X_1, ..., X_k \in \mathbb{R}^n}. \quad (2)$$

Here, k must be known in advance.

The simplicity of its algorithmic implementation encourages researchers to apply this model in a variety of engineering and scientific branches. Nevertheless, the problem is proven to be NP-hard which makes exact algorithms inapplicable for large scale problems, and the simplest and most popular algorithms result in very poor values of the squared distances sum.

Lloyd's procedure, the most popular k-means clustering algorithm, is rather fast. Nevertheless, for specific datasets including geographic/geometrical data, this algorithm results in a solution which is very far from the global minimum of the objective function (1), and the multi-start operation mode does not improve the result significantly. More accurate k-means clustering methods are much slower. Nevertheless, recent advances in high-performance computing enable us to work through a large amount of computation using the Lloyd's procedure embedded into more complex algorithmic schemes. Thus, the demand for clustering algorithms that compromise on the time spent for computations and the resulting objective function (1) value is apparent. Nevertheless, in some cases, when solving problem (1), it is required to obtain a result (a value of the objective function) within a limited fixed time, which would be difficult to improve on by known methods without a significant increase in computational costs. Such results are required if the cost of error is high, as well as for evaluating faster algorithms, as reference solutions.

If a problem must be solved within a limited time with the maximum accuracy, which would be difficult to improve using known methods without increasing computational costs, the variable neighborhood search (VNS) algorithms, which search in randomized neighborhoods formed by the application of greedy agglomerative procedures, are competitive.

Agglomerative procedures, despite their relatively high computational complexity, can be successfully integrated in tomore complex search schemes. They can be used as a part of the VNS algorithms. Moreover, such algorithms are a compromise between the solution accuracy and time costs. In our research, by accuracy, we mean exclusively the ability of the algorithm (solver) to obtain the minimum values of the objective function (1). Such procedures start from an infeasible solution with an excessive number of centroids $k+r$. In each iteration, such procedures eliminate one or more centroids and improve th result with the use of Lloyd's procedure. The use of VNS algorithms, that search in the neighborhoods, formed by applying greedy agglomerative procedures to a known (current) solution, enables us to obtain good results in a fixed time acceptable for interactive modes of operation. The selection of such procedures, their sequence and their parameters remained an open question. The efficiency of such procedures has been experimentally shown on some test and practical problems. Various versions of VNS algorithms based on greedy agglomerative procedures differ significantly in their results which makes such algorithms difficult to use in practical problems. It is practically impossible to forecast the relative performance of a specific VNS algorithm based on such generalized numerical features of the problem as the sample size and the number of clusters. Moreover, the efficiency of such procedures depends on their parameters. However, the type and nature of this dependence has not been studied.

In this work,we systematize approaches to the construction of search algorithms in neighborhoods, formed by the use of greedy agglomerative procedures, and define the $GREEDY_r$ neighborhood where r is its parameter (a number of excessice centroids in the intermediate solution). We investigate the influence of the most important parameter of such neighborhoods (parameter r) on the computational efficiency and propose a new VNS-based algorithm (solver), implemented on the graphics processing unit (GPU), which adjusts this parameter. Benchmarking on data sets composed of up to millions of objects demonstrates the advantage of the new algorithm in comparison with known local search algorithms (including those seaching in SWAP neighborhoods such as j-Means), within a fixed time, allowing for online computation, at significance level 0.01 (Wilcoxon-Mann-Whitney test).

This research is funded by the Ministry of Science and Higher Edication of the Russian Federation, project FEFE-2020-0013.

[1] *Kazakovtsev L.* Self-Adjusting Variable Neighborhood Search Algorithm for Near-Optimal k-Means Clustering // Computation, 2020, Vol. 8. No 4.

Интервальный подход в проблеме аппроксимации Эйлеровой константы

Карацуба Екатерина Анатольевна[1]⋆ ekar@ccas.ru
[1]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Интервальная арифметика появилась в ответ на вопрос об улучшении надёжности полученного вычислительного результата.

Получить реалистичную оценку погрешности вычисления часто не менее трудно, чем найти подходящий вычислительный алгоритм. Следствием чего является отсутствие оценок прогрешности вычислений во многих пакетах программ численных вычислений, основанных на замене точного значения одним приближённым. При этом пользователи таких программ не могут быть уверены в точности полученного результата.

Интервальный подход заключается в том, что вычисляемое значение заменяется не единственным элементом того же класса, а конечно-представимым множеством элементов, содержащим нужный элемент – вычисляемое значение. Термин "интервальный" этот подход получил благодаря тому, что интервал, задаваемый как правило парой рациональных чисел – границ интервала, является простейшим видом конечно-представимого множества, включающего простейший элемент – вещественное число. При этом исходные данные и промежуточные результаты представляются граничными значениями, над которыми и проводятся все операции. Преимуществом *интервального подхода* является автоматический учёт погрешностей и гарантированная точность вычисления. В настоящее время *интервальные методы* получили распространение во многих областях прикладной науки: от машиностроения и метрологии до экономического планирования и прогнозирования.

Интервальный подход можно применить также в области чистой математики; особенно в таком разделе этой науки как *аппроксимация* (специальных функций и классических констант).

В настоящей работе рассматривается проблема эффективного вычисления классической Эйлеровой константы γ, определяемой выражением:

$$\gamma = \lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n\right).$$

Последовательность

$$D_n = 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n;\ \ n=1,2,3,\ldots;$$

сходится к числу γ очень медленно. Для разности $D_n-\gamma$ справедливы оценки

$$\frac{1}{2(n+a)} \leqslant D_n-\gamma \leqslant \frac{1}{2(n+b)},$$

с такими доказанными разными авторами константами a и b:
$a = 1$, $b = 1$ (Тимс и Тирэлл), $a = 1$, $b = 0$ (Янг), $a = \frac{2\gamma-1}{2-2\gamma}$, $b = \frac{1}{6}$ (Альцер).

ДеТэмпл рассматривал сходимость к Эйлеровой константе следующей последовательности

$$R_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log\left(n + \frac{1}{2}\right);\ n = 1, 2, 3, \ldots;$$

и доказал, что для любого натурального n для разности $R_n - \gamma$ справедливы оценки

$$\frac{1}{24(n+1)^2} \leqslant R_n - \gamma \leqslant \frac{1}{24n^2}.$$

Обозначим

$$H(n) = n^2\,(R_n - \gamma)\,, \quad H(n) = n^2 h(n).$$

В конце 90-х гг. М. Вуоринен сформулировал гипотезу:
функция $H(n)$ *возрастает монотонно от* $H(1) = -\gamma + 1 - \log\frac{3}{2} = 0.0173\ldots$ *до* $\frac{1}{24} = 0.0416\ldots$ *при* $n \to +\infty$; $n = 1, 2, 3, \ldots$

В [1] автор доказал гипотезу Вуоринена и получил такие *интервалы* для разностей $D_n - \gamma$ и $R_n - \gamma$ при $n = 1, 2, 3, \ldots$:

$$\frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} - \frac{1}{126n^6} \geqslant D_n - \gamma \leqslant \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4},$$
$$\frac{1}{24n^2} - \frac{1}{24n^3} + \frac{1}{120n^4} - \frac{1}{126n^6} \leqslant R_n - \gamma \leqslant \frac{1}{24n^2} - \frac{1}{24n^3} + \frac{23}{960n^4}.$$

Для построенного автором на основе метода БВЕ быстрого алгоритма вычисления Эйлеровой константы, быстрая аппроксимация к константе Эйлера γ вычисляемых сумм определяется следующим интервалом:

$$-\frac{2}{(12k)!} - 2k^2e^{-k} \ \leqslant\ \gamma - 1 + \log k \sum_{r=1}^{12k+1} \frac{(-1)^{r-1}k^{r+1}}{(r-1)!(r+1)}$$
$$-\sum_{r=1}^{12k+1} \frac{(-1)^{r-1}k^{r+1}}{(r-1)!(r+1)^2} \ \leqslant\ \frac{2}{(12k)!} + 2k^2e^{-k}, \quad k \geqslant 1.$$

[1] *Karatsuba E. A.* On the computation of the Euler constant gamma // J. of Numerical Algorithms, 2000. Vol. 24. Pp. 83–97.

The interval approach in the problem of approximation of the Euler constant

Ekaterina Karatsuba[1]⋆ ekar@ccas.ru

[1]Moscow, FRCCSC of the Russian Academy of Sciences

Interval arithmetic is a response to the question of improving the reliability of the computational result.

Get a realistic estimate of the calculation error are often no less difficult than to find a suitable computational algorithm. As a result, there are no estimates of the calculational errors in many software packages for numerical computations based on replacing the exact value with an approximate one. Moreover, users of such programs cannot be sure in the accuracy of the result obtained.

In *interval approach*, the value to be calculated is replaced not by a single element of the same class, but by a finite-representable set of elements containing the desired element – the value to be calculated. The term "interval" this approach has got due to the fact that the interval, defined as a rule by a pair of rational numbers – the boundaries of the interval, is the simplest form of a finite-representable set, which includes the simplest element – a real number. In this case, the initial data and intermediate results are represented by the boundary values, over which all operations are carried out. The advantage of the *interval approach* is that errors are automatically taken into account and the accuracy of the calculation is guaranteed. At present, *interval methods* are widely used in many areas of applied science: from mechanical engineering and metrology to economic planning and forecasting.

The *interval approach* can also be applied in area of pure mathematics; especially in such a field of this science as *approximation* (special functions and classical constants).

In the present work, we consider the problem of efficient calculation of the classical Euler constant γ, defined by the formula:

$$\gamma = \lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n\right).$$

The sequence

$$D_n = 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n;\ n=1,2,3,\ldots;$$

converges to γ very slowly. For the difference

$D_n-\gamma$ the following bounds are valid

$$\frac{1}{2(n+a)} \leqslant D_n-\gamma \leqslant \frac{1}{2(n+b)},$$

with the constants a and b proved by different authors:
$a=1$, $b=1$ (Tims and Tyrell), $a=1$, $b=0$ (Young), $a=\frac{2\gamma-1}{2-2\gamma}$, $b=\frac{1}{6}$ (Alzer).

DeTemple considered the convergence to the Euler constant of the following sequence

$$R_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log\left(n + \frac{1}{2}\right);\ n = 1, 2, 3, \ldots;$$

and proved that for any natural number n for the difference $R_n - \gamma$ the following bounds are valid

$$\frac{1}{24(n+1)^2} \leqslant R_n - \gamma \leqslant \frac{1}{24n^2}.$$

Denote

$$H(n) = n^2\,(R_n - \gamma)\,, \quad H(n) = n^2 h(n).$$

In the late 90s. M. Vuorinen formulated a conjecture:
the function $H(n)$ increases monotonically from $H(1) = -\gamma + 1 - \log\frac{3}{2} = 0.0173\ldots$ to $\frac{1}{24} = 0.0416\ldots$ for $n \to +\infty$; $n = 1, 2, 3, \ldots$

In [1], the author proved Vuorinen's conjecture and obtained such *intervals* for the differences $D_n - \gamma$ and $R_n - \gamma$ for $n = 1, 2, 3, \ldots$:

$$\frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} - \frac{1}{126n^6} \geqslant D_n - \gamma \leqslant \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4},$$
$$\frac{1}{24n^2} - \frac{1}{24n^3} + \frac{1}{120n^4} - \frac{1}{126n^6} \leqslant R_n - \gamma \leqslant \frac{1}{24n^2} - \frac{1}{24n^3} + \frac{23}{960n^4}.$$

For the fast algorithm for computation of Euler's constant constructed by author on the basis of the FEE method, the fast approximation to the Euler constant γ of the corresponding sums to be calculated is determined by the following interval:

$$-\frac{2}{(12k)!} - 2k^2e^{-k} \ \leqslant\ \gamma - 1 + \log k \sum_{r=1}^{12k+1} \frac{(-1)^{r-1}k^{r+1}}{(r-1)!(r+1)}$$
$$- \sum_{r=1}^{12k+1} \frac{(-1)^{r-1}k^{r+1}}{(r-1)!(r+1)^2} \ \leqslant\ \frac{2}{(12k)!} + 2k^2e^{-k}, \quad k \geqslant 1.$$

[1] *Karatsuba E. A.* On the computation of the Euler constant gamma // J. of Numerical Algorithms, 2000. Vol. 24. Pp. 83–97.

Вероятностное моделирование процесса стохастического оценивания межкадровых геометрических деформаций изображений

Ташлинский Александр Григорьевич[1]⋆ tag@ulstu.ru
Сафина Галина Леонидовна[2] minkinag@mail.ru
[1]Ульяновск, Ульяновский государственный технический университет
[2]Москва, Национальный исследовательский Московский государственный строительный университет

Оценивание параметров геометрических деформаций последовательности изображений является одной из актуальных задач и понимания обработки изображений. Одним из подходов к решению этой задачи является стохастическое оценивание. Известны асимптотически оптимальные по скорости сходимости процедуры стохастической аппроксимации, однако точностные возможности процедур этого класса исследованы только в асимптотике. Подходы к улучшению (акселеризации) и анализу точности оценок алгоритмов стохастической аппроксимации при конечном числе итераций основаны, как правило, на учете априорной информации об оптимальном решении, которая задается финитной плотностью распределения вероятностей (ПРВ). При отсутствии такой априорной информации оптимальные алгоритмы на конечных итерациях могут приводить к оценкам очень далеким от оптимальных.

В докладе предложена методика вероятностного финитного моделирования стохастического оценивания параметров межкадровых геометрических деформаций изображений. Моделирование процесса стохастического оценивания параметров деформаций, требует учета сложного комплекса влияющих факторов. Факторами, не зависящими от параметров процедуры оценивания, являются ПРВ и корреляционные функции изображений и мешающих шумов, а также вид целевой функции качества оценивания. К характеристикам процедуры, на которые можно воздействовать, можно отнести способ расчета стохастического градиента, матрицу усиления, число итераций и начальное приближение вектора оценок параметров.

Для реализуемости процедуры моделирования целесообразно использовать минимальный набор величин, характеризующих независимые факторы достаточно для нахождения вероятностных оценок параметров деформаций как функции управляемых характеристик процедуры. В качестве таких величин применены вероятности сноса оценок (улучшения, ухудшения и не изменения оценок по отношению оптимальных значениям). Для получения расчетных выражений вероятностей сноса оценок использована нормализуемость стохастического градиента целевой функции при увеличении объема локальной выборки отсчетов изображений, по которой он находится. Получены выражения для ситуаций использования в качестве целевых функций оценивания коэффициента межкадровой корреляции и среднего квадрата межкадровой разности. Для по-

следнего на рисунке 1 приведены графики вероятности ρ_h^+ улучшения оценки параллельного сдвига h изображений от погрешности оценивания ε_h при объеме выборки 1 (кривая 1), 4 (кривая 2) и 10 (кривая 3). Там же показаны экспериментальные результаты (крестики), полученные статистическим моделированием на имитированных изображениях с аналогичными параметрами.

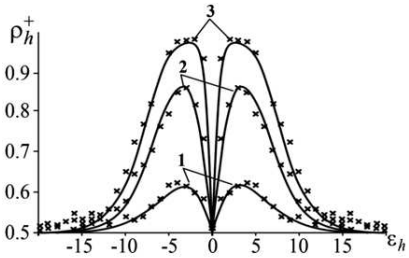


Рисунок 1

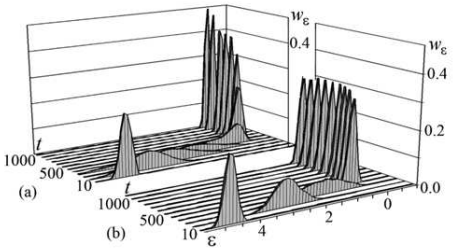


Рисунок 2

Кроме того методика предполагает дискретизацию области определения параметров деформаций. Особенностью методики является и то, что в ней модели анализируемых изображений и шумов задаются ПРВ и автокорреляционными функциями. Рассмотрено нахождение с использованием методики ПРВ оценок параметров деформаций изображений, формируемых безыдентификационной релейной процедурой за заданное конечное числе итераций. На рисунке 2 приведены примеры расчета ПРВ w_ε погрешности ε оценки сдвига изображений на итерациях релейного оценивания при изменяющемся (рисунок 2а) и постоянном (рисунок 2б) коэффициентах матрицы усиления.

Работа поддержана грантами РФФИ № 19-29-09048 и 18-41-730006.

[1] *Tashlinskii A. G., Safina G. L., Kovalenko R. O.* Probabilistic finite modeling of stochastic estimation of image inter-frame geometric deformations // Journal of Physics: Conference Series, 2019. Vol. 1368. No 3.

Probabilistic modeling of stochastic estimation process of image inter-frame geometric deformations

Tashlinskii Alexander Grigorievich[1]⋆ tag@ulstu.ru
Safina Galina Leonidovna[2] minkinag@mail.ru
[1]Ulyanovsk, Ulyanovsk State Technical University
[2]Moscow, National Research Moscow State University of Civil Engineering

Estimation of geometric deformations the parameters of an image sequence is one of the urgent problems and understanding of image processing. One of the approaches to solve this problem is stochastic estimation. Stochastic approximation procedures that are asymptotically optimal in convergence rate are known, but the accuracy capabilities of procedures of this class have been investigated only in asymptotics. Approaches to improving (accelerating) and analyzing the accuracy of estimates of stochastic approximation algorithms for a finite number of iterations are based, as a rule, on taking into account a priori information about the optimal solution, which is specified by a finite probability density (PD). In the absence of such a priori information, optimal algorithms at final iterations can lead to estimates that are very far from optimal.

In the report a method of probabilistic finite modeling of stochastic estimation of image inter-frame geometric deformations parameters is proposed. Modeling the process of stochastic estimation of deformation parameters requires taking into account a complex set of influencing factors. The factors that do not depend on the parameters of the estimation procedure are the PD and correlation functions of images and interfering noise, as well as the form of the objective function of the estimation quality. The characteristics of the procedure that can be influenced include the method for calculating the stochastic gradient, the gain matrix, the number of iterations, and the initial approximation of the vector of parameter estimates.

In order to realize the modeling procedure, it is advisable to use the minimum set of values characterizing the independent factors sufficient to find probabilistic estimates of the deformation parameters as a function of the controlled characteristics of the procedure. As such values, the probabilities of the demolition of estimates (improvement, deterioration, and no change in estimates in relation to the optimal values) are used. To obtain the calculated expressions for the probabilities of the demolition of the estimates, we used the normalizability of the stochastic gradient of the objective function with an increase in the volume of the local sample of image samples from which it is found. Expressions are obtained for two objective functions for estimating: the interframe correlation coefficient and the mean square of the interframe difference. For the last, Figure 1 shows the dependences of the probability ρ_h^+ of the parallel shift estimate improvement on the estimation error ε_h for a sample size of 1 (curve 1), 4 (curve 2) and 10 (curve 3). The same figure shows the experimental results (crosses) obtained by statistical modeling on simulated images with similar parameters.

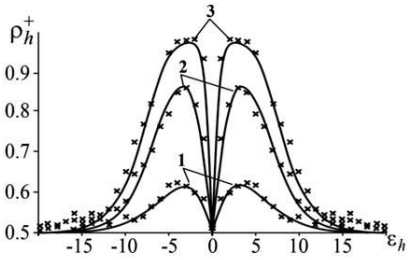


Рисунок 1

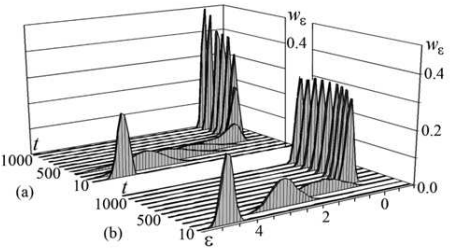


Рисунок 2

In addition, the technique assumes discretization of the domain for determining the deformation parameters. A feature of the method is that in it the models of the analyzed images and noise are set by the PD and autocorrelation functions. Finding the image deformations parameters estimates formed by a non-identification relay procedure for a given finite number of iterations with usage of the PD method, is considered. Figure 2 shows examples of calculating the PD w_ε of the image shift estimation error ε at the relay estimation iterations with changing (Figure 2a) and constant (Figure 2b) gain matrix coefficients.

When modeling the process of stochastic gradient estimation at each iteration, an adaptive limitation of the boundaries of the modeling window in the parameter space is applied. In this case, in order to preserve the accuracy of the simulation, the drift probabilities are corrected at the nodes of the discretized domain for determining the parameters near the boundaries of the simulation window. This allowed several times to reduce the computational costs while maintaining the adequacy of the model.

The proposed method of probabilistic finite modeling can be used to find the accuracy and probabilistic characteristics of stochastic algorithms for estimating interframe geometric deformations of images for a given number of iterations.

This research is funded by RFBR, grants 19-29-09048 and 18-41-730006.

[1] *Tashlinskii A. G., Safina G. L., Kovalenko R. O.* Probabilistic finite modeling of stochastic estimation of image inter-frame geometric deformations // Journal of Physics: Conference Series, 2019. Vol. 1368. No 3.

Восстановление решений уравнений типа Урысона

Белозуб Владимир Антонович[1] disstroier@mail.ru
Козлова Маргарита Геннадьевна[1] art-inf@mail.ru
Лукъяненко Владимир Андреевич[1]⋆ art-inf@yandex.ru

[1]Симферополь, ФГАУО ВО «Крымский федеральный университет им. В. И. Вернадского»

Рассматривается задача восстановления решений уравнения Урысона в предположении, что известна априорная информация о решении и дополнительная информация о решении близких уравнений. Здесь возникают две самостоятельные задачи обработки и интерпретации косвенно получаемых данных. Обработка полученных экспериментальных данных проводится с целью выделения максимума достоверной информации о реальных характеристиках восстанавливаемых функций. Наблюдаются только интегральные характеристики, т. е. учитывающие суммарный эффект от всех точек наблюдаемого объекта. Такие характеристики нечуствительны даже к большим изменениям величин, характеризующих объект, если эти изменения компенсируют друг друга. Таким образом, задача интерпретации состоит в решении обратной задачи $Az = u$. Прямая задача состоит в измерении интегральных характеристик объекта — правых частей интегральных уравнений типа Урысона 1-го рода u по заданным исходным зависимостям z, характеризующим объект.

Информация о монотонности искомого решения $z(s)$ позволяет свести задачу решения нелинейного уравнения Урысона 1-го рода к линейному уравнению типа свертки. Учет разнообразной информации уже требует разработки интеллектуализированной системы обработки данных. Данный подход приводит к различным сценариям нахождения искомых величин по результатам косвенных измерений с учетом разнообразной информации. Здесь естественно использовать решение близкого уравнения. Для обоснования восстановления решений исследуемых уравнений с помощью двух уравнений, рассматривается схема решения одного уравнения $Az = u$ на основе решения близкого ему уравнения $\widetilde{A}\widetilde{z} = \widetilde{u}$. Задача сводится к двум близким экстремальным задачам $M^\alpha[z] = \alpha||z||^2_{W_2^1} + ||Az - u||^2_{L_2}, \widetilde{M}^\alpha[\widetilde{z}] = \widetilde{\alpha}||\widetilde{z}||^2_{W_2^1} + ||\widetilde{A}\widetilde{z} - \widetilde{u}||^2_{L_2}$ или к эквивалентным уравнениям $Bz \equiv \alpha(-z'' + z) + A^*Az = A^*u \equiv g,\ \widetilde{B}\widetilde{z} \equiv \widetilde{\alpha}(-\widetilde{z}'' + \widetilde{z}) + \widetilde{A}^*\widetilde{A}\widetilde{z} = \widetilde{A}^*\widetilde{u} \equiv \widetilde{g}$ с параметрами регуляризации $\alpha, \widetilde{\alpha}$.

Если выполняется условие $||\widetilde{B}^{-1}(B - \widetilde{B})|| < 1$, тогда итерационный процесс $z_{n+1} = z_n\widetilde{B}^{-1}[Bz_n - g]$ — сходится в W_2^1. Все элементы z_n обладают свойством $z^* - z_n \subset W_2^1$. Справедлива формула для погрешности:

$$||z_n - z^*|| \leqslant \frac{||\widetilde{B}^{-1}(\widetilde{B} - B)||^n}{1 - ||\widetilde{B}^{-1}(\widetilde{B} - B)||}||\widetilde{B}^{-1}g||_{W_2^1}.$$

Разнообразие алгоритмов следует из применяемых методов регуляризации, для метода регуляризации Лаврентьева $F(z) = Az - u + \alpha(z - z_m)$ итерационная процедура построения z_{n+1} приближения $z_{n+1} = z_n + h$ приводит к алгоритму нахождения h из линейного уравнения $[\alpha I + (A'z_n)]h - F(z_n) = 0$ или

$$z_{n+1} = z_n - \gamma[\alpha I + (A'z_n)]^{-1}[Az_n - u + \alpha(z_n - z_m)].$$

В качестве близких уравнений следует рассматривать уравнения

$$\mathbb{K}h \equiv [\alpha I + (A'z_n)]h = Az_n - u + \alpha(z_n - z_m) \equiv g,$$

$$\widetilde{\mathbb{K}}\widetilde{h} \equiv [\widetilde{\alpha} I + (\widetilde{A}'\widetilde{z}_n)]\widetilde{h} = \widetilde{A}\widetilde{z}_n - \widetilde{u} + \widetilde{\alpha}(\widetilde{z}_n - \widetilde{z}_m) \equiv \widetilde{g},$$

где, обеспечив $||\widetilde{\mathbb{K}}^{-1}(\mathbb{K} - \widetilde{\mathbb{K}})|| < 1$, можем найти решение через решение близкого уравнения, более простого по своей структуре, или решение, которое уже найдено для различных уровней погрешности (прецедентная информация). Для модифицированного варианта метода Левенберга-Марквардта решения нелинейных уравнений в частном случае имеет вид

$$z_{n+1} = z_n - \gamma[\widetilde{A}'(z_n)^*\widetilde{A}'(z_n) + \alpha I]^{-1}[\widetilde{A}'(z_n)^*(Az_n - u_\delta) + \alpha(z_n - \widetilde{z}_m)].$$

Здесь $\widetilde{z}_m$ — начальное приближение для искомого решения.

Таким образом, наличие эффективно решаемых близких уравнений позволяет строить алгоритмы для исходных уравнений.

[1] *Lukyanenko V. A.* Some Tasks for Integral Equations of Urison's Type // Proceedings of the International Conference «Integral Equations – 2010», 2010. Pp. 80–84.

Reconstruction of solutions of equations of Uryson type

Vladimir Belozub[1] disstroier@mail.ru
Margarita Kozlova[1] art-inf@mail.ru
Vladimir Lukianenko[1]⋆ art-inf@yandex.ru
[1]Simferopol, V.I. Vernadsky Crimean Federal University

We consider the problem of restoring solutions to the Urysohn equation under the assumption that we know a priori information about the solution and additional information about the solution of close equations. Here there are two separate tasks of processing and interpretation of the data obtained indirectly. Processing of the obtained experimental data is carried out in order to extract the maximum reliable information about the real characteristics of the restored functions. Only integral characteristics are observed, i.e. they take into account the total effect of all points of the observed object. Such characteristics are not sensitive even to large changes in the values that characterize the object, if these changes compensate for each other. Thus, the problem of interpretation consists in solving the inverse problem $Az = u$. The direct problem is to measure the integral characteristics of an object — the right-hand sides of Uryson-type integral equations of the 1st kind u from the given initial dependencies z that characterize the object.

Information about the monotonicity of the desired solution $z(s)$ allows us to reduce the problem of solving the nonlinear Uryson equation of the 1st kind to a linear convolution equation. Accounting for a variety of information already requires the development of an intelligent data processing system. This approach leads to different scenarios for finding the desired values based on the results of indirect measurements, taking into account a variety of information. Here it is natural to use the solution of a close equation. To justify the recovery of solutions of the studied equations using two equations, we consider a scheme for solving a single equation $Az = u$ based on the solution of a close equation $\widetilde{A}\widetilde{z} = \widetilde{u}$. The problem is reduced to two close extreme problems

$$M^{\alpha}[z] = \alpha||z||^2_{W_2^1} + ||Az - u||^2_{L_2}, \widetilde{M}^{\alpha}[\widetilde{z}] = \widetilde{\alpha}||\widetilde{z}||^2_{W_2^1} + ||\widetilde{A}\widetilde{z} - \widetilde{u}||^2_{L_2}$$

or to equivalent equations $Bz \equiv \alpha(-z'' + z) + A^*Az = A^*u \equiv g$,

$$\widetilde{B}\widetilde{z} \equiv \widetilde{\alpha}(-\widetilde{z}'' + \widetilde{z}) + \widetilde{A}^*\widetilde{A}\widetilde{z} = \widetilde{A}^*\widetilde{u} \equiv \widetilde{g}$$

where $\alpha, \widetilde{\alpha}$ — the parameters of regularization.

If the condition $||\widetilde{B}^{-1}(B - \widetilde{B})|| < 1$ is met, then the iterative process $z_{n+1} = z_n\widetilde{B}^{-1}[Bz_n - g]$ — converges to W_2^1. All elements of z_n have the property $z^* - z_n \subset W_2^1$. The formula for error:

$$||z_n - z^*|| \leqslant \frac{||\widetilde{B}^{-1}(\widetilde{B} - B)||^n}{1 - ||\widetilde{B}^{-1}(\widetilde{B} - B)||}||\widetilde{B}^{-1}g||_{W_2^1}.$$

The variety of algorithms follows from the regularization methods used, for the Lavrentiev regularization method

$$F(z) = Az - u + \alpha(z - z_m)$$

iterative construction procedure z_{n+1} approximations $z_{n+1}{=}z_n{+}h$ leads to an algorithm for finding h from the linear equation $[\alpha I + (a' z_n)]h - F(z_n) = 0$ or

$$z_{n+1} = z_n - \gamma[\alpha I + (A_n)]^{-1}[Az_n - u + \alpha(z_n - z_m)].$$

The equations should be considered as close equations

$$\mathbb{K}h \equiv [\alpha I + (A' z_n)]h = Az_n - u + \alpha(z_n - z_m) \equiv g,$$

$$\widetilde{\mathbb{K}}\widetilde{h} \equiv [\widetilde{\alpha} I + (\widetilde{A}' \widetilde{z}_n)]\widetilde{h} = \widetilde{A}\widetilde{z}_n - \widetilde{u} + \widetilde{\alpha}(\widetilde{z}_n - \widetilde{z}_m) \equiv \widetilde{g},$$

where, by providing $||\widetilde{\mathbb{K}}^{-1}(\mathbb{K} - \widetilde{\mathbb{K}})|| < 1$, we can find a solution through the solution of a close equation that is simpler in its structure, or a solution that has already been found for various error levels (case information). For a modified version of the Levenberg-Marquardt method, solutions of nonlinear equations in the special case have the form

$$z_{n+1} = z_n - \gamma[\widetilde{A}'(z_n)^* \widetilde{A}'(z_n) + \alpha I]^{-1}[\widetilde{A}'(z_n)^*(Az_n - u_\delta) + \alpha(z_n - \widetilde{z}_m)],$$

where $\widetilde{z}_m$ — initial approximation for the desired solution.

Thus, the presence of effectively solved close equations allows us to build algorithms for the original equations.

[1] *Lukyanenko V. A.* Some Tasks for Integral Equations of Urison's Type // Proceedings of the International Conference ¡¡Integral Equations – 2010¿¿, 2010. Pp. 80–84.

Скелет символа как модель следа пера для распознавания по восстановленной траектории

Арсеев Сергей Петрович[1]⋆ 9413serg@gmail.com
Местецкий Леонид Моисеевич[1] mestlm@mail.ru
[1]Москва, Московский государственный университет имени М.В. Ломоносова

Распознавание рукописного текста является активно исследуемой задачей компьютерного зрения, находящей своё применение во многих практических областях, например, в автоматической обработке документов или в работе с архивами. Существует два основных подхода к решению этой задачи: offline- и online-распознавание.

Offline-распознавание представляет собой задачу распознавания написанного текста по его изображению. Эта задача чаще всего встречается на практике. Online-распознавание представляет собой задачу распознавания текста по записанному в процессе его написания следу пера.Такая задача часто является более простой в решении, чем задача offline-распознавания, но основной проблемой такого подхода является необходимость использования специального записывающего оборудования при написании текста, что существенно ограничивает применение этого подхода на практике.

Цель данной работы – попытаться свести задачу offline-распознавания рукописных символов к задаче online-распознавания посредством реконструкции возможного следа пера по изображению. Такое сведение позволит решать широкий спектр задач распознавания рукописного текста методами online-распознавания, которые часто оказываются более эффективными и не столь требовательны к объёму обучающей выборки. Основной идеей предложенного подхода является построение модели символа в виде геометрического графа с вершинами в точках, близких к положению центра пера при начертании символа, и рёбрами, соответствующими отрезкам траектории движения пера. После построения такой модели восстановленный след пера будет являться маршрутом в данном графе. Для этого используется медиальное представление или скелет символа на этом изображении, где скелетом называется множество центров всех вписанных пустых кругов фигуры.

Алгоритм сведения задачи offline-распознавания к задаче online-распознавания, предложенный в этой работе, состоит из следующих шагов: построение скелета символа и представление его в виде графа; построение по данному графу вспомогательного графа, называемого метаграфом, для облегчения разрешения неоднозначностей обхода; построение обхода метаграфа; построение обхода исходного графа по обходу метаграфа и генерализация обхода графа посредством агрегирования вершин и рёбер.

Основным критерием пригодности восстановленного возможного следа пера для распознавания символа алгоритмами online-распознавания является сравнение качества работы алгоритма распознавания на реальной траектории пера и

на восстановленной по записи символа. Критерием же возможности распознавания посредством сведения задачи offline-распознавания к online-распознаванию является сравнение качества распознавания с использованием предложенного алгоритма и распознавания традиционными методами offline-распознавания.

Для распознавания символов с использованием получившейся трассы использовался алгоритм, основанный на использовании рекуррентной нейронной сети, состоящей из 24 слоёв управляемых рекуррентных блоков с сигмоидной функцией активации. В качестве альтернативного метода offline-распознавания использовалась нейросеть VGG16, предобученная на базе ImageNet. Разбиение на обучающую и тестовую выборки проводилось также в соотношении 4:1. Эксперименты проводились на наборе данных, содержащем последовательности точек, соответствующих положениям пера. Последовательности были записаны с помощью сенсорного контроллера, сохраняющего последовательности координат при движении пера. Набор данных состоит из семи классов, для каждого класса имеется по 100 примеров. Для получения изображений текста для offline-распознавания эти символы были растеризованы. Впоследствии алгоритм восстановления возможного следа пера работал с этими растеризованными символами как с изображениями, без использования исходных данных о положении пера.

Результаты экспериментального исследования методов распознавания приведены в таблице 1. Методы обозначены следующим образом: RNN-True – рекуррентная нейронная сеть, обученная на истинных данных о положении пера; RNN-Offline – рекуррентная нейронная сеть, обученная на восстановленной траектории пера; VGG – свёрточная сеть VGG16, предобученная на ImageNet и дообученная на изображениях из набора.

Метод	Точность
RNN-True	0.93
RNN-Offline	0.87
VGG	0.60

Таблица 1. Результаты экспериментального исследования.

Видно, что предложенный метод предсказуемо отстаёт от online-распознавания по истинным данным, но при этом он значительно опережает по своим характеристикам метод offline-распознавания, несмотря на то, что сеть VGG16 предобучалась на наборе данных ImageNet, а модели online-распознавания были обучены с нуля на достаточно малой обучающей выборке. Таким образом, использование рекуррентной нейронной сети совместно с предложенным алгоритмом для online-распознавания обеспечивает возможность обучения и распознавания даже в условиях малой обучающей выборки, на которой неспособны обучиться современные свёрточные модели.

Дальнейшие направления исследования могут включать в себя совершенствование метода восстановления траектории пера, расширение коллекции данных на весь алфавит с увеличением числа почерков, эксперименты с распознаванием непрерывных последовательностей символов, а также эксперименты за зашумлённых данных.

Работа поддержана грантом РФФИ № 20-01-00664.

[1] *Арсеев С. П., Местецкий Л. М.* Распознавание рукописного текста по восстановленному следу пера с помощью медиального представления // Информационные технологии и нанотехнологии (ИТНТ-2020), 2020. С. 683–689.

Symbol skeleton as a pen trace model for recognition using reconstructed trace

Sergey Arseev[1]⋆ 9413serg@gmail.com
Leonid Mestetskiy[1] mestlm@mail.ru
[1]Moscow, Lomonosov Moscow State University

The task of handwritten text recognition is a major part of computer vision which has various practical uses such as automatic processing of documents or digitalization of handwritten text archives. There are two main approaches to that problem: offline and online recognition.

Offline recognition approach only uses the text image as an input. This approach is the most widely used in practical applications. Online recognition, on the other hand, is based on a recorded pen trace. This task can be easier than offline recognition in some cases, however the main difficulty limiting this approach is the necessity to use special hardware at the time of recording which limits the practicality of this approach.

The goal of this paper is to try and reduce the offline handwritten symbol recognition task to online recognition via reconstructing the possible pen trace. This reduction allows us to apply online recognition methods to a wider group of practical tasks potentially gaining performance and reducing the necessary training set size compared to offline methods. The main idea of the proposed approach is to model the symbol as a geometrical graph with nodes close to pen position at the time of writing and edges representing the pen movement between the nodes. After this, a walk in this graph that includes all nodes of the graph will be modelling a potential pen trace. For this purpose, a skeleton or a medial axis of a sybbol is used where a skeleton oof a polygonal figure is the locus of the centers of circles that are tangent to the polygon edge in two or more points, where all such circles are contained in the figure.

The pen trace reconstruction algorithm presented in this paper consists of the following steps: medial representation construction and its representation as a graph; construction of an auxiliary graph called a metagraph of the original graph which simplifies possible tracing ambiguities; calculation of a walk in the metagraph calculation of a walk in the initial graph using the metagraph walk and trace generalization via aggregation of short edges and closely positioned vertices.

The main quality criterion of the reconstructed pen trace for the recognition task is the comparison between offline recognition, online recognition with true pen trace and offline-to-online recognition with the reconstructed pen trace. A recurrent neural network based algorithm was used for online symbol recognition where the network consisted of 24 gated recurrent units with sigmoidal activation function. A VGG16 convolutional neural network pre-trained on ImageNet dataset was used as an offline recognition method. Images were split into training and testing sets in 4:1 proportion. The experiments were conducted on a dataset containing point

sequences corresponding to pen positions. The dataset consists of 7 classes with 100 examples for each class. To get training images for offline recognition, these traces were rasterized. The pen trace reconstruction algorithm then processed them as images without additional information about the pen position.

The results of the evaluation are shown in Table I. Recognition methods are marked as follows: RNN-True – recurrent neural network trained on the real pen trace data; RNN-Offline – same network trained on reconstructed pen trace data provided by the proposed algorithm; VGG – the VGG16 convolutional network pre-trained on ImageNet and fine-tuned on the training set.

Method	Accuracy
RNN-True	0.93
RNN-Offline	0.87
VGG	0.60

Table 2. Experimental evaluation results.

The proposed method ranks below the online recognition on true data but has a significant advantage over the offline recognition method despite the VGG16 network being pre-trainied on the ImageNet dataset and the online recognition models being trained from scratch on a very small training set. This way, a recurrent neural network used for online recognition in conjunction with the proposed algorithm allows successful learning and recognition even for the small training set insufficient to train modern convolutional models.

Further research is underway and will include improvements in the pen trace reconstruction methods for more accurate reconstruction of continuous strokes and experimental evaluation of different offline and online recognition algorithms on larger datasets containing more symbol classes as well as continuous sequences of symbols.

This research is funded by RFBR, grant 20-01-00664.

[1] *Arseev S., Mestetskiy L.* Handwritten text recognition using reconstructed pen trace with medial representation // Information Technology and Nanotechnology (ITNT-2020), 2020. Pp. 683–689.

Применение генеративных нейросетей для повышения пространственного разрешения спутниковых изображений

Мурынин Александр Борисович[1,2]⋆ amurynin@bk.ru
Матвеев Иван Алексеевич[1] matveev@ccas.ru
Игнатьев Владимир Юрьевич[3] vladimir.ignatiev.mipt@gmail.com
[1]Москва, ФИЦ ИУ РАН
[2]Москва, НИИ «АЭРОКОСМОС»
[3]Москва, Сколковский ин-т науки и технологий

Генеративно-состязательные нейронные сети применены для повышения разрешения спутниковых изображений определенного класса без привлечения дополнительных данных. Оценка качества получаемых изображений повышенного разрешения проводится отношением сигнал/шум и мерой структурного сходства. На основе известных функций потерь, используемых в генеративно-состязательных нейронных сетях, получена функция, специфичная для решаемой задачи. Обучение и тестирование ведется на примере изображений объектов железнодорожной инфраструктуры, выборка представляет около 78 км железных дорог. Особенность разработанного подхода заключается в объединении метода на основе генеративно-состязательных нейронных сетей и классических методов. Для синтеза мультиспектральных каналов могут использоваться алгоритмы слияния изображений, основанные на вероятностном анализе, вейвлет-анализе, методе главных компонент, преобразовании цветовых компонент.

Работа выполнена при поддержке Минобрнауки России (уникальный идентификатор проекта RFMEFI60719X0312)

[1] *Игнатьев В. Ю., Матвеев И. А., Мурынин А. Б., Усманова А. А., Цурков В. И.* Повышение пространственного разрешения панхроматических спутниковых изображений на основе генеративных нейросетей // Известия РАН. Теория и системы управления, 2021. № 1.

Application of generative neural networks to increase the spatial resolution of satellite images

Alexander Murynin[1], 2⋆ amurynin@bk.ru
Ivan Matveev[1] matveev@ccas.ru
Vladimir Ignatiev[3] vladimir.ignatiev.mipt@gmail.com
[1]Moscow, FRC CSC RAS
[2]Moscow, AEROCOSMOS Research Institute for Aerospace Monitoring
[3]Moscow, Skolkovo Institution of Science and Technology

Generative adversarial neural networks are used to increase the resolution of satellite images of a certain class without involving additional data. The quality of the obtained high-resolution images is assessed by the signal-to-noise ratio and the measure of structural similarity. Based on the known loss functions used in generative adversarial neural networks, a function specific to the problem being solved is obtained. Training and testing is carried out on the example of images of railway infrastructure objects, the sample represents about 78 km of railways. The peculiarity of the developed approach is to combine the method based on generative adversarial neural networks and classical methods. To synthesize multispectral channels, image fusion algorithms based on probabilistic analysis, wavelet analysis, principal component analysis, and color component transformation can be used. This research is supported by Ministry of Education and Science, project ID RFMEFI60719X0312.

[1] *Ignatiev V., Matveev I., Murynin A., Usmanova A., Tsurkov V.* Increasing the spatial resolution of panchromatic satellite images based on generative neural networks // Journal of Computer and Systems Sciences International, 2021. No 1.

Метод распознавания шрифтов на основе медиального представления

Липкина Анна Львовна[1]⋆ lipkina96@mail.ru
Местецкий Леонид Моисеевич[1] mestlm@mail.ru
[1]Москва, Московский государственный университет имени М.В. Ломоносова

В статье описывается метод распознавания шрифтов на основе медиального представления, интегрированный в систему распознавания шрифтов по цифровому изображению текста. Эта система ищет похожие шрифты, упорядоченные по схожести, на шрифт, изображенный на введенном пользователем изображении текста.

Работа системы основана на решении двух задач машинного обучения: распознавания текста на изображении и распознавания шрифта по изображению текста. Для решения первой задачи используется понятие математической модели графемы, основанной на непрерывном медиальном представлении символа. Решение задачи распознавания шрифта основано на понятии морфологической ширины фигуры, также тесно связанной с медиальным представлением. Мы предлагаем метод использования функции морфологической ширины для поиска наиболее похожих шрифтов из известной базы.

Проведенные эксперименты показывают высокую точность поиска наиболее похожих шрифтов. Для базы, состоящей из 2543 шрифтов, точность составляет 0.991 по метрике *top*@5 для правильно распознанного текста в размере шрифта 100 пикселей на изображении.

Работа поддержана грантом РФФИ № 20-01-00664.

[1] *Липкина А. Л., Местецкий Л. М.* Метод распознавания шрифтов на основе медиального представления // Труды 30-й Международной конференции по компьютерной графике и машинному зрению GraphiCon, 2020.

Medial representation based font recognition method

Anna Lipkina[1]⋆ lipkina96@mail.ru
Leonid Mestetskiy[1] mestlm@mail.ru
[1]Moscow, Lomonosov Moscow State University

In this article a method of font recognition based on the medial representation, integrated into the font recognition system based on a digital image of text is described. This system searches for similar fonts, ordered by similarity, to the font shown in the user-entered text image.

The system is based on solving two machine learning problems: text recognition on an image and font recognition on a text image. To solve the first problem, we use the concept of a mathematical model of a grapheme based on a continuous medial representation of a symbol. The solution of the font recognition problem is based on the concept of the morphological width of the figure, which is also closely related to the medial representation. We propose a method for using the morphological width function to find the most similar fonts from a known database.

The experiments show high accuracy of searching for the most similar fonts. For a database consisting of 2543 fonts, the accuracy is 0.991 according to the metric *top*@5 for correctly recognized text in the font size of 100 pixels in the image.

This research is funded by RFBR, grant 20-01-00664.

[1] *Lipkina A., Mestetskiy L.* Medial representation based font recognition method // In proceedings of the 30th international conference on computer graphics and machine vision GraphiCon, 2020.

Бигармоническое сглаживание изображений

Василенко Вера Викторовна[2] t9288487681@gmail.com
Сафронов Алексей Павлович[1] mf.cellan@gmail.com
Смыслов Александр Андреевич[1] sarhanishe@yandex.ru
Цепляев Даниил Павлович[1] xximikk1696@gmail.com
Марковский Алексей Николаевич[1]⋆ mrkvsk@yandex.ru

[1] Краснодар, Кубанский государственный университет

[2] Краснодар, Краснодарское высшее военное авиационное училище летчиков им. Героя Советского Союза А. К. Серова

Рассматривается задача сглаживания монохромных цифровых изображений. Предлагаемый подход можно отнести к диффузионным методам сглаживания. Цифровые изображения моделируются функциями из пространства $L_2(Q)$, где Q поле зрения. Рассматривается расширение классического оператора Лапласа с выделением подпространства его взаимно однозначного действия и, далее, строится обратный к нему – гармонический и бигармонический оператор сглаживания. Такие операторы есть свертка исходного (четкого) изображения и фундаментального решения за минусом проекции этой свертки на гармоническое или бигармоническое подпространство. Рассматривается разложение пространства $L_2(Q)$ в ортогональную сумму гармонического, бигармонического и новиковского подпространства. Приводится алгоритм метода базисных потенциалов выделения бигармонической составляющей цифрового изображения; метод опирается на полную систему базисных потенциалов. Рассматривается дискретный случай прямого и обратного гармонического и бигармонического оператора; вводится спектральный параметр, определяется сумма и композиция таких операторов. Рассматривается двупараметрическое семейство сглаживающих и концентрирующих преобразований, для которых спектральные параметры выступают мерой гладкости. Представлены результаты вычислительных экспериментов при различных значениях спектральных параметров. Случай гармонического сглаживания рассматривался в [1].

[1] *Василенко В. В., Сафронов А. П., Смыслов А. А., Цепляев Д. П., Марковский А. Н.* Гармоническое сглаживание цифровых изображений // Экологический Вестник Научных Центров ЧЭС, 2020. Т. 17. № 1(2). С. 8–15.

Biharmonic smoothing the images — IDP-13

Vera Vasilenko[2] t9288487681@gmail.com
Aleksey Safronov[1] mf.cellan@gmail.com
Aleksander Smyslov[1] sarhanishe@yandex.ru
Daniil Tsyplyaev[1] xximikk1696@gmail.com
Aleksey Markovskiy[1]⋆ mrkvsk@yandex.ru

[1] Krasnodar, Kuban State University
[2] Krasnodar, Krasnodar Higher Military Aviation School Of Pilots. Hero Of The Soviet Union A. K. Serov

The problem of smoothing monochrome digital images is considered. The proposed approach can be attributed to diffusion smoothing methods. Digital images are modeled by functions from the space $L_2(Q)$, where Q is the field of view. We consider an extension of the classical Laplace operator with the allocation of a subspace of its one-to-one action, and then construct the inverse of it-the harmonic and biharmonic smoothing operator. Such operators are the convolution of the original (clear) image and the fundamental solution minus the projection of this convolution on a harmonic or biharmonic subspace. We consider the decomposition of the space $L_2(Q)$ into the orthogonal sum of the harmonic, biharmonic, and Novikov subspaces. The algorithm of the method of basic potentials for selecting the biharmonic component of a digital image is presented; the method is based on a complete system of basic potentials. We consider the discrete case of forward and reverse harmonic and biharmonic operators; a spectral parameter is entered, and the sum and composition of such operators are determined. We consider a two-parameter family of smoothing and concentrating transformations for which the spectral parameters are a measure of smoothness. The results of computational experiments for various values of spectral parameters are presented. The case of harmonic smoothing was considered in [1].

[1] *Vasilenko V. V., Safronov A. P., Smyslov A. A., Tseplyaev D. P., Markovskiy A. N.* Harmonic smoothing of digital images // Environmental Bulletin of the BSEC Scientific Centers, 2020. Vol. 17. No 1(2). Pp. 8–15.

Информационная модель для метода обеспечения качества автоматической сегментации изображений

Мурашов Дмитрий Михайлович[1] d_murashov@mail.ru

[1]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

При сегментации изображений возникает проблема выбора параметров применяемых алгоритмов. Параметры назначаются, исходя из наилучшего разбиения изображения на сегменты. Критерием качества разбиения может быть визуальная оценка эксперта или какой-либо количественный показатель. Ранее, исходя из гипотезы о минимизации избыточности информации на ранних стадиях обработки сигнала в зрительной системе человека, для определения наилучшего разбиения изображения на сегменты из множества разбиений, полученных при различных значениях параметра алгоритма сегментации, предложено использовать критерий минимума меры избыточности информации. Представляемая работа посвящена исследованию указанного критерия качества сегментации цифровых изображений.

Задача обеспечения качества сегментации формулируется следующим образом. Пусть операция сегментации описывается соотношением:

$$\boldsymbol{V} = \boldsymbol{F}(\boldsymbol{U}, \boldsymbol{t}),$$

где $\boldsymbol{U}$ - входное изображение, представленное как вектор $\boldsymbol{U} \in \mathbb{R}^N$; $\boldsymbol{V}$ - сегментированное изображение, $\boldsymbol{V} \in \mathbb{R}^N$; $\boldsymbol{F}$ - вектор-функция (нелинейная в общем случае), описывающая алгоритм сегментации; $\boldsymbol{t} \in \mathbb{R}^P$ - вектор параметров. Варьируя параметры, получим множестово из J разбиений изображения $\boldsymbol{U}$ на сегменты $\mathcal{V} = \{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_j, \ldots, \boldsymbol{V}_J\}$. Требуется выбрать разбиение $\boldsymbol{V}_j$, обеспечивающее минимум критерия качества $M(\boldsymbol{U}, \boldsymbol{V}_j)$:

$$\boldsymbol{V}_{\min} = \arg\min_j \left(M(\boldsymbol{U}, \boldsymbol{V}_j)\right), j = 1, 2, \ldots, J.$$

Для применения теоретико-информационного подхода необходима вероятностная модель связи между исходным и сегментированным изображениями. Пусть исходное и сегментированное изображения являются входом и выходом стохастической информационной системы. Значения уровней серого изображений описываются дискретными случайными переменными U и V со значениями u и v. Переменная U имеет L, а V может иметь $1 \leqslant l \leqslant L$ градаций серого тона.

Операция сегментации будет представлена моделью информационного канала:

$$V = F(U + \eta, t),$$

где U - сигнал на входе, V - выход канала, F - функция преобразования, t - параметр, η - шум канала, переменные V и η независимы. Рассматривается

критерий качества сегментации изображений в виде минимума меры избыточности канала, определяемой в виде

$$R = \frac{H(V|U)}{H(V)},$$

где $H(V|U)$ - условная энтропия выхода канала V при условии, что вход равен U, $H(V)$ - энтропия выхода.

Для исследования характера изменения меры избыточности предлагается упрощенная модель совместного двумерного дискретного распределения уровней полутонов входа и выхода системы сегментации, изменяющегося при уменьшении количества сегментов (см. [1]). Предположим, что совместное распределение уровня серого тона изображений U и V может быть представлено K компонентами, соответствующими сегментам V. Для простоты предполагается, что все компоненты однотипные и состоят из L составляющих, которые соответствуют частоте появления пикселей уровня серого в области изображения U, соответствующей сегменту k изображения V, $1 \leqslant l \leqslant L$. Каждая из компонент распределения имеет пик, соответствующий доминирующему уровню серого. Пусть $P(u_l, v_k) = P$, если $l = k$. Соотношение между значениями вероятностей уровней серого l в компонентах определяется коэффициентом α, $0 < \alpha \leqslant 1$. Например, в сегменте изображения V закодированном уровнем v_1,

$$P(u_2, v_1) = P(u_3, v_1) = \ldots = P(u_L, v_1) = \alpha P(u_1, v_1) = \alpha P,$$

где $P = 1\,[(L-1)\,\alpha + 1]$; $\alpha = \alpha(k)$ зависит от количества сегментов на изображении V:

$$\alpha(k) = \frac{a}{\left(1 + e^{c(K-b)}\right)} + d,$$

где a, b, c и d – параметры, $d = 1 - a$.

Показано, что для предложенной модели совместного распределения существует минимум меры избыточности, которому соответствует наилучшее разбиение изображения. Соответствие модели реальной системе сегментации подтверждено вычислительным экспериментом на изображениях из базы Berkeley Segmentation Dataset (BSDS500). Выявлено, что для достаточно большой группы тестовых изображений результаты сегментации, полученные из условия минимума меры избыточности, обеспечивают минимум информационной меры различия (вариации информации) при сравнении с эталонными изображениями из базы BSDS500. Проведено сравнение результатов сегментации изображений из базы BSDS500 системой EDISON по критерию минимума информационной избыточности и известному критерию в виде относительной скорости убывания энтропии изображения при уменьшении количества сегментов. Сегментированное изображение, полученное из условия минимума меры избыточности, имеет большее сходство с входным изображением, чем изображение, полученное по

энтропийному критерию, и в большинстве случаев показало меньшее отличие от эталонных сегментаций.

Работа поддержана грантом РФФИ № 18-07-01385.

[1] *Murashov D. M.* An Information Model of Image Segmentation Algorithm Based on Redundancy Minimization // IEEE Proceedings, 2020.

Information model for quality assessment method applied to automatic image segmentation

Dmitry Murashov[1] d_murashov@mail.ru

[1]Moscow, FRCCSC of the Russian Academy of Sciences

When segmenting images, the problem of setting the parameters of the applied algorithms arises. For different tasks of image analysis, various quality criteria should be selected. The criterion for the quality of the partition can be a visual assessment of an expert or some quantitative measure. Earlier, based on the hypothesis of minimizing the redundancy of information at the early stages of signal processing in the human visual system, to determine the best partitioning from the set of partitions of the image into segments obtained at different values of the parameter of the segmentation algorithm, it was proposed to use the criterion of the minimum information redundancy. The presented work is devoted to the study of the specified quality criterion for the segmentation of digital images.

The segmentation operation can be described by the following expression:

$$\boldsymbol{V} = \boldsymbol{F}(\boldsymbol{U}, \boldsymbol{t}),$$

where $\boldsymbol{U}$ is an input image represented as a vector, $\boldsymbol{U} \in \mathbb{R}^N$; $\boldsymbol{V} \in \mathbb{R}^N$, is a segmented image; $\boldsymbol{F}$ is a vector-function (non-linear in general case) describing segmentation algorithm, $\boldsymbol{t} \in \mathbb{R}^P$ is a vector of parameters. Varying coordinates of parameter vector, one will obtain a set of segmented images. $\mathcal{V} = \{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_j, \ldots, \boldsymbol{V}_J\}$. It is necessary to choose an image $\boldsymbol{V}_j$ giving minimum to performance criterion $M(\boldsymbol{U}, \boldsymbol{V}_j)$:

$$\boldsymbol{V}_{\min} = \arg\min_j \left(M(\boldsymbol{U}, \boldsymbol{V}_j)\right), j = 1, 2, \ldots, J.$$

To apply information-theoretic approach, a probabilistic model of relationship between input and segmented images is needed. For simplicity, we will consider the segmentation process of grayscale images. Let the source and segmented images be the input and output of the stochastic information system. Discrete random variables U and V with values u and v describe the grayscale values of the system input and output, respectively. We assume that the image U has L, and V can have $1 \leqslant l \leqslant L$ gray levels. The segmentation operation will be represented by an information channel model:

$$V = F(U + \eta, t),$$

where U is an input signal, V is a channel output, F is a transformation function, t is a parameter, and η is a channel noise. We assume that variables V and η are independent.

As a quality criterion of image segmentation, we consider the minimum of information channel redundancy, defined as

$$R = \frac{H(V|U)}{H(V)},$$

where $H(V|U)$ is a conditional entropy of the output V under condition that the input is equal to U, $H(V)$ is an entropy of the output. To study the qualitative properties of the redundancy measure, we propose a simplified model of a joint two-dimensional discrete distribution of the gray scale values at the segmentation system input and output (see [1]). Suppose that the joint grayscale distribution of the images U and V can be represented by K components corresponding to the segments of image V. For simplicity, we assume that all components have the same size. Each component have L elements that correspond to the frequency of occurrence of pixel's gray level l in the region of image U corresponding to k-th segment of V. Each component has a peak corresponding to a dominant gray level. Let $P(u_l, v_k) = P$, if $l = k$. The relationship between the probabilities of the gray levels in the components is determined by the coefficient α, $0 < \alpha \leqslant 1$. For example, in a segment of image V encoded with level v_1,

$$P(u_2, v_1) = P(u_3, v_1) = \ldots = P(u_L, v_1) = \alpha P(u_1, v_1) = \alpha P,$$

where $P = 1\,[(L-1)\,\alpha + 1]$; $\alpha = \alpha(k)$ depends on the number of segments in the image V:

$$\alpha(k) = \frac{a}{\left(1 + e^{c(K-b)}\right)} + d,$$

where a, b, c and d are the parameters, $d = 1 - a$.

It is shown that for the proposed model of joint distribution, there exists a minimum of the redundancy measure, which corresponds to the best image partitioning. The model validity was confirmed by a computational experiment on images from the Berkeley Segmentation Dataset database. It was revealed that for a sufficiently large group of test images, the segmentation results obtained from the condition of the minimum measure of redundancy produce a minimum information measure of difference (variation of information) when compared with the ground truth images from the BSDS500 database. A comparison of the results of segmenting images from the BSDS500 database by the EDISON system was carried out using the criterion of minimum information redundancy and the well-known criterion in the form of the relative rate of decrease of the image entropy. The segmented image obtained from the condition of the minimum of the redundancy measure is more similar to the input image than the image obtained using the entropy criterion, and in most cases, it showed less dissimilarity from the ground truth segmentations.

This research is funded by RFBR, grant 18-07-01385.

[1] *Murashov D. M.* An Information Model of Image Segmentation Algorithm Based on Redundancy Minimization // IEEE Proceedings, 2020.

Определение Движения Оптической Системы на Основе Метода Согласованного Оценивания

Фурсов Владимир Алексеевич[1,2]⋆ fursov@ssau.ru
Минаев Евгений Юрьевич[1] e.minaev@gmail.com
Котов Антон Петрович[1,2] kotov@ssau.ru

[1] Самара, Самарский национальный исследовательский университет им. академика С.П. Королева
[2] Самара, Институт систем обработки изображений - филиал ФНИЦ «Кристаллография и фотоника» РАН

Одним из важных направлений исследований в области систем технического зрения является создание систем автономной навигации транспортных средств с использованием оптических средств наблюдения [1]. Задача определения относительного движения формулируется как задача построения оптического потока, создаваемого последовательностью изображений динамически меняющейся сцены (динамическим изображением). Оптический поток – это векторное поле скоростей движения точек или фрагментов изображения. Оптический поток восстанавливается путем анализа и сравнения, чаще всего путем совмещения, последовательности кадров динамического изображения.

В последние годы активно развивается подход к построению автономных систем оптической навигации на основе визуальной одометрии. В частности, используется метод визуальной одометрии, в котором объектив оптического регистрирующего устройства направлен вниз, т.е. на кадрах регистрируется последовательность изображений опорной поверхности, по которой движется транспортное средство [2].

В настоящей статье на основе этих двух идей (применения оптического потока и использования последовательности изображений опорной поверхности) мы предлагаем технологию определения движения, основанную на методе согласованного оценивания.

В классической задаче вычисления оптического потока предполагается, что на видеокадрах пиксель в точке (x, y, t) с интенсивностью $I(x, y, t)$ перемещается между двумя кадрами. При этом происходит сдвиг $(\Delta x, \Delta y, \Delta t)$ между этими кадрами. Также предполагается, что яркость постоянна, т.е. пиксель переходит к следующему кадру без изменений: $I(x, y, t) \cong I(x + \Delta x, y + \Delta y, t + \Delta t)$. Дополнительно предполагают, что смежные N пикселей в некоторой небольшой окрестности смещаются на одинаковое расстояние и решают переопределенную систему уравнений:

$$\mathbf{I}_{x,y}\boldsymbol{\Delta} = \mathbf{I}_t, \tag{1}$$

где $\boldsymbol{\Delta} = [\Delta x, \Delta y]^T$, $\mathbf{I}_t$ — матрица-столбец, размерности $N \times 1$, $\mathbf{I}_{x,y} = [\mathbf{I}_x, \mathbf{I}_y]$ — $N \times 2$ -матрица, а $\mathbf{I_x}, \mathbf{I_y}, \mathbf{I_t}$ — $N \times 1$ -матрицы, составленные в соответствии с соотношениями:

$\mathbf{I}_x = \left[\frac{\partial I_1}{\partial x}, ..., \frac{\partial I_N}{\partial x}\right]^T, \mathbf{I}_y = \left[\frac{\partial I_1}{\partial y}, ..., \frac{\partial I_N}{\partial y}\right]^T, \mathbf{I_t} = [-\partial I_1, ..., -\partial I_N]^T.$

Для решения переопределенной системы уравнений (1) мы применяем метод согласованного оценивания, который развивается авторами [3]. Кратко существо метода сводится к следующему. Из исходной системы (1) формируется L подсистем *верхнего уровня* размерности P:

$$\mathbf{I}_{x,y}(l)\,\mathbf{\Delta}(l) = \mathbf{I}_t(l) + \xi_l, \quad l = \overline{1,L}, \quad M < P < N.$$

Для кажой $\hat{l}$-ой подсистемы получаем множество оценок Θ_i для которых вводится функция согласованности

$$W(l) = \frac{2}{K(K-1)} \sum_{k,j=1,j>k}^{K} \left\| \hat{\mathbf{\Delta}}(l,k) - \hat{\mathbf{\Delta}}(l,j) \right\|,$$

являющаяся мерой взаимной близости оценок. Далее ищется $\hat{l}$-я подсистема верхнего уровня, для которой

$$W\left(\hat{l}\right): \quad W\left(\hat{l}\right) = \min_{l=\overline{1,L}} W(l). \tag{2}$$

Наилучшим считается решение, полученное на этой $\left(\hat{l}\right)$-й подсистеме верхнего уровня. Это решение может быть найдено любым из известных способов. Простейшей является МНК-оценка:

$$\hat{\mathbf{\Delta}}(l) = \left[\mathbf{I}_{x,y}^T\left(\hat{l}\right)\mathbf{I}_{x,y}\left(\hat{l}\right)\right]^{-1}\mathbf{I}_{x,y}^T\left(\hat{l}\right)\mathbf{I}_t^T\left(\hat{l}\right). \tag{3}$$

Следует заметить, что ($\hat{l}$ -я) подсистема верхнего уровня, на которой вычислена оценка (3), вообще говоря, включает уравнения переопределенной системы, которые соответствуют точкам изображения, имеющим близкие (согласованные) скорости оптического потока. Поэтому метод согласованного оценивания может рассматриваться как метод сегментации поля скоростей оптического потока.

Описанный метод согласованного оценивания применялся авторами при построении трехмерных моделей местности по разноракурсным изображениям [4]. Особенность настоящей задачи состоит в том, что априори известно, что все точки изображений опорной поверхности обязаны иметь одинаковые сдвиги. Поэтому в рамках настоящей технологии реализация метода согласованного оценивания сводится к следующему. На последовательных изображениях опорной поверхности необходимо найти множество наиболее согласованных решений, которые в соответствии с критерием (2) и являются наиболее точными оценками

искомых сдвигов. В докладе приводятся различные варианты реализации описанного выше метода согласованного оценивания, учитывающие свойство одинаковости сдвигов, а также результаты экспериментальных исследований технологии на тестовых данных, полученных путем съемки опорных поверхностей с различными текстурами.

Работа выполнена в рамках государственного задания по теме FSSS-0777-2020-0017.

[1] *Mohamed S. A. Haghbayan M. H., Westerlund T, Heikkonen J., Tenhunen H., Plosila J.* A survey on odometry for autonomous navigation systems // IEEE Access, 2019. Pp. 97466–97486.

[2] *Nourani-Vatani N., Borges P. V.* Correlation-based visual odometry for ground vehicles // Journal of Field Robotics, 2011. Pp. 742–768.

[3] *Fursov V. A., Kotov A. P., Goshin Ye. V.* Solution of overdetermined systems of equations using the conforming subsystem selection // Journal of Physics: Conference Series Procedia Engineering, 2019. Pp. 708–717.

[4] *Kotov A. P., Goshin Ye. V., Fursov V. A.* DEM generation based on RPC model using relative conforming estimate criterion // Procedia Engineering, 2017. Pp. 708–717.

Motion Detection of Optical Systems Based on the Conformed Estimation Method

Vladimir Fursov[1,2]⋆ fursov@ssau.ru
Evgeniy Minaev[1] e.minaev@gmail.com
Anton Kotov[1,2] kotov@ssau.ru
[1]Samara, Samara National Research University
[2]Samara, IPSI RAS - branch of the FSRC "Crystallography and Photonics" RAS

The development of autonomous navigation systems for vehicles using optical surveillance equipment is one of the major challenges in bulding computer vision systems [1]. The problem of determining the relative motion is formulated as the problem of constructing an optical flow created by a sequence of images of a dynamically changing scene (dynamic image). Optical flow is a vector field of velocities of motion of points or image fragments. The optical flow is restored by analysis and comparison, most often by combining, a sequence of frames of a dynamic image.

In recent years, an approach to the construction of autonomous optical navigation systems based on visual odometry has been actively developed. In particular, the visual odometry method is used, in which the lens of the optical recording device is directed downward, i.e. on the frames, a sequence of images of the reference surface is recorded on which the vehicle is moving.

In this article, based on these two ideas (using optical flow and using a sequence of reference surface images), we propose a motion detection technology based on the method of consistent estimation [2].

In the classical problem of calculating the optical flow, it is assumed that on video frames, a pixel at a point (x, y, t) with intensity $I(x, y, t)$ moves between two frames. In this case, a shift $(\Delta x, \Delta y, \Delta t)$ occurs between these frames. It is also assumed that the brightness is constant, i.e. the pixel goes to the next frame unchanged: $I(x, y, t) \cong I(x + \Delta x, y + \Delta y, t + \Delta t)$. Additionally assumed that adjacent N pixels in some small neighborhood are displaced by the same distance and solve the overdetermined system of equations:

$$\mathbf{I}_{x,y}\boldsymbol{\Delta} = \mathbf{I}_t, \tag{1}$$

where $\boldsymbol{\Delta} = [\Delta x, \Delta y]^T$, $\mathbf{I}_t$ — matrix-column, dimensions $N \times 1$, $\mathbf{I}_{x,y} = [\mathbf{I}_x, \mathbf{I}_y]$ — $N \times 2$ -matrix, and $\mathbf{I_x}, \mathbf{I_y}, \mathbf{I_t}$ — $N \times 1$ -matrices, composed in accordance with the relations:

$\mathbf{I}_x = \left[\frac{\partial I_1}{\partial x}, ..., \frac{\partial I_N}{\partial x}\right]^T, \mathbf{I}_y = \left[\frac{\partial I_1}{\partial y}, ..., \frac{\partial I_N}{\partial y}\right]^T, \mathbf{I_t} = [-\partial I_1, ..., -\partial I_N]^T$.

To solve the overdetermined system of equations (1) we use the Conformed Estimates Method (CEM), which is developed by the authors [3]. Briefly, the essence of the proposed method is as follows. From the original system (1) L subsystems of the *upper level* of dimension P are formed:

$$\mathbf{I}_{x,y}(l)\,\boldsymbol{\Delta}(l) = \mathbf{I}_t(l) + \xi_l, \quad l = \overline{1, L}, \quad M < P < N.$$

To characterize sets Θ_i a conformity function is introduced

$$W(l) = \frac{2}{K(K-1)} \sum_{k,j=1,j>k}^{K} \left\| \hat{\mathbf{\Delta}}(l,k) - \hat{\mathbf{\Delta}}(l,j) \right\|,$$

which is a closeness measure of estimates. Next, the $\hat{l}$-th subsystems of the *upper level* is searched for, for which

$$W\left(\hat{l}\right): \quad W\left(\hat{l}\right) = \min_{l=\overline{1,L}} W(l). \tag{2}$$

It is considered the best solution obtained in the $\left(\hat{l}\right)$ subsystems of the upper level. This solution can be found in any of the known ways. The simplest is a least-squares estimate:

$$\hat{\mathbf{\Delta}}(l) = \left[\mathbf{I}_{x,y}^T\left(\hat{l}\right)\mathbf{I}_{x,y}\left(\hat{l}\right)\right]^{-1}\mathbf{I}_{x,y}^T\left(\hat{l}\right)\mathbf{I}_t^T\left(\hat{l}\right). \tag{3}$$

It should be noted that the ($\hat{l}$ -th) *upper level* subsystem on which estimate (3) is calculated, generally speaking, includes the equations of the overdetermined system, which correspond to image points having close (consistent) optical flow rates. Therefore, the CEM can be considered as a method of segmentation of the optical flow velocity field.

The described method of conformed estimation was used by the authors for digital elevetaion model (DEM) generation from multi-angle satellite images [4]. The peculiarity of this problem is that it is known a priori that all points of the images of the reference surface must have the same shifts. Therefore, within the framework of this technology, the implementation of the CEM is reduced to the following. On successive images of the reference surface, it is necessary to find the set of the most conformed solutions, which, in accordance with criterion (2), are the most accurate estimates of the required shifts. The report presents various options for implementing the above-described method of conformed estimation, taking into account the property of the same shifts, as well as the results of experimental studies of the technology on test data obtained by shooting reference surfaces with different textures.

The research was carried out within the state assignment theme FSSS-0777-2020-0017.

[1] *Mohamed S. A. Haghbayan M. H., Westerlund T, Heikkonen J., Tenhunen H., Plosila J.* A survey on odometry for autonomous navigation systems // IEEE Access, 2019. Pp. 97466–97486.

[2] *Nourani-Vatani N., Borges P. V.* Correlation-based visual odometry for ground vehicles // Journal of Field Robotics, 2011. Pp. 742–768.

[3] *Fursov V. A., Kotov A. P., Goshin Ye. V.* Solution of overdetermined systems of equations using the conforming subsystem selection // Journal of Physics: Conference Series Procedia Engineering, 2019. Pp. 708–717.

[4] *Kotov A. P., Goshin Ye. V., Fursov V. A.* DEM generation based on RPC model using relative conforming estimate criterion // Procedia Engineering, 2017. Pp. 708–717.

Определение эффективности алгоритмов выделения линий на изображении

Применко Дмитрий Владимирович[1] dima-primenko777@yandex.ru
Панищев Владимир Славиевич[1] gskunk@yandex.ru
Бурцев Олег Александрович[1]⋆ olegon_web@mail.ru
[1]Курск, Юго-Западный государственный университет

Несмотря на постоянно растущий уровень развития компьютерной техники, и возможностей, которые с каждым годом увеличиваются все больше и больше, остается не решенными целый ряд практических задач, с которыми компьютер и по сей день не может справиться. К числу таких задач относится задача автоматического распознавания и интерпретации визуальной информации. Это во многом связано со сложностью формализации процесса восприятия видимых образов. Поэтому, несмотря на очевидную легкость, с которой человек решает задачу распознавания окружавших его предметов, все еще нет «универсального» математического или технологического подхода, позволяющего конструктивно разрабатывать методы, алгоритмы и автоматические устройства, эффективно осуществляющие процесс распознавания.Однако, для некоторых частных ситуаций, когда математические модели оказывается подходящими для той или иной практической задачи, удается получить приемлемые результаты.

Задача распознавания простейших геометрических объектов на двумерном изображении имеет большое значение для различных отраслей знаний и многих технических приложений. В настоящее время разработано большое количество различных алгоритмов и методов, позволяющих осуществлять распознавание объектов на графических изображениях и видеопотоках. Наиболее распространенные алгоритмы, используемые при выделении линий на изображениях, такие как начальная обработка (нормализация) изображения является алгоритм преобразования Радона и преобразование Хафа и их модификации.

Целью данной работы является проектирование и реализация программного продукта для исследования алгоритмов выделения линий на изображении.

Для достижения поставленной цели необходимо решить следующие задачи: рассмотреть и дать описание существующим алгоритмам по выделению геометрических объектов на изображении; составить математическую модель и выполнить разработку программного продукта для исследования алгоритмов выделения линий на изображении; провести эксперименты некоторых алгоритмов выделения линий на изображении.

В ходе работы была разработана программа для исследования алгоритмов выделения линий на изображениях. В качестве примера были рассмотрены: алгоритм преобразования Хафа и его модификации, а также простое преобразование Радона.

Были проведены эксперименты для определения эффективности алгоритмов выделения линий на изображении при различных количествах линий на

исходном изображении и при различном размере исходного изображения по показателям эффективности и скорости. В эксперименте использовались следующие алгоритмы: обычное преобразования Хафа, случайное преобразования Хафа, адаптивное преобразования Хафа, вероятностное преобразования Хафа, простое преобразование Радона.

Для определения эффективности алгоритмов выделения линий на изображении при различных количествах линий на исходном изображении были подготовлены три различных изображения одинакового размера (400x400) и с одинаковым фоном с количеством линий 5, 10, 15.

Для определения эффективности алгоритмов выделения линий на изображении при различных количествах линий на изображении при различном размере исходного изображения были подготовлены три различных изображения разного размера (400x400, 1000x1000, 1500x1500) и с одинаковым фоном с количеством линий (10).

При выполнении экспериментов выявлено, что наилучший средний результат по показателям скорость и эффективность показал алгоритм обычного преобразования Хафа, при различных количествах линий эффективность нахождения и выделения линий оставалась примерно на одном показателе, скорость нахождения и выделения значительно ниже, чем, к примеру у алгоритма простого преобразования Радона при той же эффективности; наилучший средний результат по показателям скорость и эффективность показал алгоритм простого преобразования Радона, при различных размерах изображений эффективность нахождения и выделения линий оставалась примерно на одном показателе, скорость нахождения и выделения примерно на одном уровне при обработке изображения 1000x1000 и 1500x1500.

[1] *Пестунов И. А˙ Рылов С. А˙* Алгоритмы спектрально-текстурной сегментации спутниковых изображений высокого пространственного разрешения // Вестник Кемеровского государственного университета, 2014, №4. С. 104–109.

Determining the effectiveness of line identification algorithms in an image

Dmitry Primenko[1] dima-primenko777@yandex.ru
Vladimir Panishchev[1] gskunk@yandex.ru
Oleg Burtsev[1] olegon_web@mail.ru
[1]Kurks, South-West State University

Despite the constantly growing level of computer hardware development, and the opportunities that are increasing more and more year after year, there are still a number of practical tasks that the computer can not cope with up to the present day. Among these tasks is the problem of automatic identification and interpretation of visual information. This is largely due to the complexity of formalizing the process of visible images perception. Therefore, despite the obvious ease with which a person solves the problem of recognizing objects around him, there is still no "universal" mathematical or technological approach that allows us to constructively develop methods, algorithms and automatic devices, which effectively implement the identification process. However, for some particular situations, when mathematical models are suitable for this or that practical problem, it is possible to obtain acceptable results.

The task of identifying the simplest geometric objects in a two-dimensional image is of great importance for various branches of knowledge and many engineering applications. At present, a large number of different algorithms and methods have been developed that allow carrying out object identification in graphic images and video streams. The most common algorithms used in identifying lines in images, such as initial processing (normalization) of an image is the Radon transform algorithm and the Hough transform algorithm and their modifications.

The goal of the given paper is to design and implement a software product for studying algorithms to identify lines in an image.

To achieve this goal, it is necessary to solve the following tasks: to consider and describe existing algorithms for identifying geometric objects in an image; to create a mathematical model and develop a software product for studying line identification algorithms in an image; to conduct experiments of some algorithms for identifying lines in an image.

In the course of the work, a program was developed to study algorithms for identifying lines in images. The following examples were considered: the Hough transform algorithm and its modifications, as well as the ordinary Radon transform.

Experiments have been conducted to determine the effectiveness of algorithms for identifying lines in an image with different numbers of lines in the original image and with different sizes of the original image in terms of efficiency and speed. The following algorithms were used in the experiment: ordinary Hough transform,random Hough transform, adaptive Hough transform, probabilistic Hough transform, and ordinary Radon transform.

To determine the effectiveness of algorithms for identifying lines in an image with different numbers of lines in the original image, three different images of the same size were prepared (400x400). They had the same background with the number of lines 5, 10, 15.

To determine the effectiveness of algorithms for identifying lines in an image with different numbers of lines at different sizes of the original image, three different images of different sizes (400x400, 1000x1000, 1500x1500) and with the same background having the number of lines (10) were prepared.

When performing experiments, it was found that the best average result in terms of speed and efficiency was shown by the ordinary Hough transform algorithm, with different numbers of lines, the efficiency of finding and identifying lines remained approximately at the same level, the speed of finding and identifying lines is significantly lower than, for example, for the ordinary Radon transform algorithm with the same efficiency; the best average result in terms of speed and efficiency was shown by the ordinary Radon transform algorithm, for different image sizes, the efficiency of finding and identifying lines remained approximately at the same level, and the speed of finding and identifying lines remained approximately at the same level when processing 1000x1000 and 1500x1500 images.

[1] *Pestunov I. Rylov S.* Algoritmi spektralno teksturnoi segmentacii sputnikovih izobrajenii visokogo prostranstvennogo razresheniya // Vestnik Kemerovskogo gosudarstvennogo universiteta, 2014, No 4. Pp. 104–109.

Метод повышения точности классификации изображений с использованием обучения с подкреплением

Елизаров Артем Александрович[1]⋆ artelizar@gmail.com
Разинков Евгений Викторович[1] evgeny@razinkov.ai
[1]Казань, Казанский (Приволжский) федеральный университет

В настоящее время стремительно развивается такое направление машинного обучения, как обучение с подкреплением. Как следствие методы обучения с подкреплением активно используются при решении различных задач компьютерного зрения, в том числе при решении задачи классификации изображений, которая на сегодняшний день является одной из важнейших задач искусственного интеллекта.

В работе предложен метод классификации изображений с использованием подходов обучения с подкреплением. Основная идея разработанного метода базируется на предположении об увеличении точности классификации за счет вырезания правильной области на картах признаков изображения в процессе классификации. В качестве базовой модели для классификации изображений используется нейронная сеть ResNet, предобученная на используемом наборе данных. Также используется дополнительная нейронная сеть, которую назовем агентом. Агент получает на вход выход группы остаточных блоков ResNet и должен предсказать, какую область нужно вырезать из полученных карт признаков изображения, чтобы в процессе классификации увеличилась уверенность сети ResNet в принадлежности объекта на исходном изображении правильному классу. Такая задача агента сводится к задаче о контекстном многоруком бандите, а обучение агента производится с помощью алгоритмов обучения с подкреплением и стратегий достижения компромисса между эксплуатацией и исследованием при выборе действий.

Проведены эксперименты на подмножестве набора данных ImageNet по выбору архитектуры агента, алгоритма обучения с подкреплением и стратегии выбора действий при обучении, а также анализ по выбору места расположения агента в архитектуре ResNet. Рассмотрены такие стратегии выбора действий, как ε-жадная, ε-softmax, ε-decay-softmax и метод UCB1, и такие алгоритмы обучения с подкреплением, как DQN, REINFORCE и A2C.

При проведении экспериментов изображения из валидационной выборки агента использовались при обучении сети ResNet, чтобы оценить эффективность работы агента для входа из выходов группы остаточных блоков ResNet из одного распределения. При таком подходе разработанная модель показала точность классификации, значительно превосходящую точность классификации базовой модели, что дает задел для будущих исследований. Также предложен подход, в котором изображения из обучающей и валидационной выборок агента не используются при обучении ResNet. Этот подход тоже демонстрирует увеличение точности классификации по сравнению с базовой моделью.

[1] *Елизаров А. А., Разинков Е. В.* Классификация изображений с использованием обучения с подкреплением // Электронные библиотеки, 2020. Т. 23. № 6. С. 23.

Image classification accuracy improvement method using reinforcement learning

Artem Elizarov[1]⋆ artelizar@gmail.com
Evgenii Razinkov[1] evgeny@razinkov.ai
[1]Kazan, Kazan Federal University

Currently, reinforcement learning as a direction of machine learning is rapidly developing. As a result, reinforcement learning methods are actively used in solving various problems of computer vision, including the problem of image classification, which today is one of the most important tasks of artificial intelligence.

In this paper is proposed a method for image classification using reinforcement learning approaches. The main idea of the developed method is based on the assumption that the classification accuracy will be increased by cutting out the correct area on the image feature maps during the classification process. As a basic model for image classification is used the ResNet neural network pre-trained on the used dataset. Also is used an additional neural network, which we will call an agent. The agent receives at the input the output of a group of residual ResNet blocks and must predict which area needs to be cut from the obtained image feature maps, so that increases the confidence of the ResNet network in the belonging of the object on the original image to the correct class during the classification process. Such an agent's task is reduced to the task of a contextual multi-armed bandit. The agent is trained using reinforcement learning algorithms and strategies for compromising exploitation and exploration in the choice of actions.

Experiments have been carried out on a subset of the ImageNet dataset to select an agent architecture, a reinforcement learning algorithm and a strategy for choosing actions during training, as well as an analysis on choosing an agent's location in the ResNet architecture. Action selection strategies such as ε-greedy, ε-softmax, ε-decay-softmax, and the UCB1 method, and reinforcement learning algorithms such as DQN, REINFORCE, and A2C are considered.

In the experiments, images from the agent's validation set were used to train the ResNet network to evaluate the agent's performance for input from the outputs of a group of residual ResNet blocks from the same distribution. With this approach, the developed model showed a classification accuracy that significantly exceeds the classification accuracy of the base model, which provides a groundwork for future research. Also is proposed an approach, in which images from the training and validation sets of the agent are not used for training ResNet. This approach also demonstrates an increase in classification accuracy compared to the base model.

[1] *Elizarov A., Razinkov E.* Image classification using reinforcement learning // Russian Digital Libraries Journal, 2020. Vol. 23. No 6. Pp. 23.

«Формула Эйлера» для морфологического анализа мозаичных изображений

Визильтер Юрий Валентинович[1]⋆ viz@gosniias.ru
Выголов Олег Вячеславович[1] o.vygolov@gosniias.ru
Желтов Сергей Юрьевич[1] zhl@gosniias.ru
Брянский Станислав Андреевич[1] sbrianskiy@gosniias.ru
[1]Москва, ФГУП «ГосНИИАС»

В рамках простейшей морфологии Пытьева [1] изображения традиционно рассматриваются как кусочно-постоянные функции вида

$$f(x,y) = \sum_{i=1}^{n} f_{Fi}\chi_{Fi}(x,y) \tag{1}$$

где n – число областей разбиения F кадра Ω площади S на связные непересекающиеся области постоянной яркости, $F = \{F_1, \ldots, F_n\}; f = (f_{F1}, \ldots, f_{Fn})^T$ – вектор значений яркости. Такие изображения называются мозаичными.

Морфологическое сравнение мозаичных форм по сложности традиционно осуществляется в терминах отношения частичного порядка «не сложнее по форме». Множество мозаичных форм образует по данному отношению алгебраическую структуру типа «решетка», в которой для любых форм F и G можно указать форму более сложную $F \bigwedge G$ и менее сложную $F \bigvee G$. Более сложные формы получаются из менее сложных разбиением, а менее сложные из более сложных – слиянием областей. В терминах множеств (классов) изображений, F «не сложнее по форме» G, если $F \subseteq G$. В терминах проекторов, F «не сложнее по форме» G, если $P_G P_F = P_F$.

В работе [2] была введено обобщенное отношение полного порядка, основанное на определении меры сложности формы, которая равна 0 для простейшей формы O (одна область на весь кадр) и монотонно увеличивается при последовательных разбиениях областей. Для мозаичного разбиения F с площадями областей $S_{Fi} = Sp_i, i = 1, .., n$ такой мерой является ОГО-сложность

$$\mu_H(F) = \sum_{i=1,..,n} p_i(1 - p_i) \tag{2}$$

связанная с метрикой оценки геометрических отличий (ОГО) [3]:

$$d_H(F,G) = \sum_{j=1,..,l} \sum_{i=1,..,n} S_{ij} d_H(F_i, G_j) \tag{3}$$

где $d_H(F_i, G_j) = S_{Fi} + S_{Gj} - 2S_{ij}$ – расстояние Хэмминга между областями разбиения F_i и G_j; S_{ij} – площадь области $F_i \cap G_j$. ОГО-метрика может интерпретироваться как метрика редактирования мозаичных форм. Другая ее интерпретация сязана с реляционным описанием мозаичных форм [4].

Определим реляционную форму мозаичного изображения как предикат бинарного отношения «пикселы не принадлежат одной области»:

$$\pi_F(x,y,u,v) = \begin{cases} 0 & \text{если } \forall i : \chi_{Fi}(x,y) = \chi_{Fi}(u,v) \\ 1 & \text{в противном случае} \end{cases} \quad (4)$$

L^1-метрика между реляционными формами π_F и π_G в точности эквивалентна ОГО-метрике между соответствующими мозаичными формами F и G:

$$\begin{aligned} d_\pi(\pi_F,\pi_G) &= \|\pi_F(x,y,u,v) - \pi_G(x,y,u,v)\| = \\ &= \sum_{i=1,..,n} \sum_{j=1,..,m} S_{ij}(s_{Fi} + S_{Gj} - 2S_{ij}) = \\ &= \sum_{i=1,..,n} \sum_{j=1,..,m} d_H(F_i, G_j) = d_H(F,G) \end{aligned} \quad (5)$$

Значит, нормированное значение L^1-нормы реляционной формы π_F:

$$\begin{aligned} \mu_H(\pi_F) &= \|\pi_F\| - \|\pi_O\| / \|\pi_I\| - \|\pi_O\| = \\ = d_\pi(\pi_F,\pi_O)/d_\pi(\pi_I,\pi_O) &= d_\pi(\pi_F,\pi_O)/d_\pi(\pi_I,\pi_O) = \\ = \sum_{i=1,..,n} S_{Fi}(1-S_{Fi})/S^2 &= \sum_{i=1,..,n} p_i(1-p_i) = \\ &= d_H(F,O)/d_H(I,O) = \mu_H(F) \end{aligned} \quad (6)$$

есть ничто иное, как нормированная ОГО-сложность $\mu_H(F)$ (14).

Рассмотрим теперь сравнение изображений по форме, которое традиционно осуществляется при помощи морфологического проектора [1]:

$$\mathrm{g}_F(x,y) = P_F g(x,y) = \sum_{i=1}^{n} \mathrm{g}_{Fi}\chi_{Fi}(x,y), \mathrm{g}_{Fi} = \frac{(\chi_{Fi}, \mathrm{g})}{\|\chi_{Fi}\|^2}, i = 1, ..., n$$

Если изображение $\mathrm{g}(x,y)$ само является мозаичным, тогда:

$$\begin{aligned} \mathrm{g}_F(x,y) &= P_F \mathrm{g}(x,y) = P_F \sum_{j=1,..,m} \mathrm{g}_{Gj}\chi_{Gj}(x,y) = \\ = \sum_{j=1,..,m} \mathrm{g}_{Gj} P_F \chi_{Gj}(x,y) &= \sum_{j=1,..,m} \mathrm{g}_{Gj}\chi_{GFj}(x,y), \chi_{GFj}(x,y) \\ &= \sum_{i=1,..,n} \chi_{GFij}\chi_{Fi}(x,y), \chi_{GFij}(x,y) = \\ &= (\chi_{Gj}(x,y), \chi_{Fi}(x,y))/\|\chi_{Gj}(x,y)\|^2 = \\ &= S_{Wij}/S_{Fi}, i = 1,..,n; j = 1,..,m \end{aligned} \quad (7)$$

Таким образом, мы получаем два альтернативных описания проекции $\mathrm{g}_F(x,y)$: на основе четкой модели $\langle \boldsymbol{\chi}_F, \mathbf{g}_F \rangle$ и нечеткой модели $\langle \boldsymbol{\chi}_{GF}, \mathbf{g}_G \rangle$. В силу этого оператор P_F может быть рассмотрен уже не как оператор в пространстве изображений, а как оператор в пространстве мозаичных форм. В такой схеме нечеткая (диффузная [5]) мозаичная модель G_F автоматически возникает как проекция мозаичной модели на другую четкую мозаичную модель:

$$\boldsymbol{\chi}_{GF} = P_F \boldsymbol{\chi}_G \iff G_F = P_F G$$

Определим морфологический коэффициент корреляции форм (МККФ):

$$K_M(G,F) = \|P_F G\| / \|G\| = \|\boldsymbol{\chi}_{GF}(x,y)\| / \|\boldsymbol{\chi}_G(x,y)\| \tag{8}$$

Подставив (19) в (20) получим явное выражение для $K_M(G,F)$:

$$K_M^2(G,F) = \|\boldsymbol{\chi}_{GF}(x,y)\|^2 / \|\boldsymbol{\chi}_G(x,y)\|^2 = \sum_{j=1,..,m} \sum_{i=1,..,n} S_{ij}^2 / (S S_{Fi}) = \sum_{j=1,..,m} \sum_{i=1,..,n} p_{ij}^2 / p_{Fi} \tag{9}$$

где $p_{ij} = S_{ij}/S, p_{Fi} = S_{Fi}/S$.

Заметим, что выражение (21) уже было ранее получено нами из совершенно иных, статистических соображений [6, 7]. Тогда было предложено называть среднеквадратичным эффективным коэффициентом морфологической корреляции (СКМК) форм F и G корень из отношения среднего квадрата нормы проекции изображения из F на форму G к среднему квадрату нормы проецируемого изображения. Формула для СКМК в предположении о взаимной независимости яркостей областей на G, оказалась в точности такой же, как полученное нами здесь выражение для МККФ (21).

Наконец, сравнивая при помощи K_M^2 форму F с простейшей формой O:

$$K_M^2(F,O) = \sum_{i=1,..,n} \frac{S_{Fi}^2}{S^2} = \sum_{i=1,..,n} p_{Fi}^2 \tag{10}$$

легко убедиться, что:

$$\mu_H(F) = \sum_{i=1,..,n} p_{Fi}(1 - p_{Fi}) = \sum_{i=1,..,n} p_{Fi} - \sum_{i=1,..,n} p_{Fi}^2 = 1 - K_M^2(F,O) \tag{11}$$

Таким образом, мера сложности $\mu_H(F)$ оказалась непосредственно связана с морфологической корреляцией $K_M^2(F,O)$. Более того, теперь можно записать:

$$\mu_H(F) = \|\pi_F - \pi_O\|_{L1} / \|\pi_I - \pi_O\|_{L1} = d_H(F,O)/d_H(I,O) = 1 - K_M^2(F,O) \tag{12}$$

что является замечательным подтверждением внутреннего единства морфологического анализа Пытьева во всех его формах – атрибутной и реляционной, метрической и операторной, геометрической и статистической, четкой (проективной) и нечеткой (диффузной). Это выражение также демонстрирует скрытую внутреннюю связь между морофлогичесикми инструментами сравнения по сходству/различию и по сложности. В некотором смысле для морфологического анализа это такое же «объединительное» равенство, как знаменитое тождество Эйлера $e^{i\pi} + 1 = 0$ для комплексного анализа, поскольку оно сводит воедино, казалось бы, совершенно разные аспекты рассматриваемого предмета. В частности, тождество (12) показывает, что ОГО-метрика действительно является естественной метрикой форм, которая напрямую связана не только с мозаичным и реляционным представлением форм, но и с морфологическим проектором и морфологическим коэффициентом корреляции.

Работа поддержана грантом РНФ № 16-11-00082.

[1] *Пытьев Ю. П., Чуличков А. И.* Методы морфологического анализа изображений // М: Физматлит, 2010. С. 336.

[2] *Визильтер Ю. В., Рубис А. Ю.* Морфологическое сравнение образов по сложности // Всероссийская конференция ММРО-16, 2013. С. 70.

[3] *Визильтер Ю. В., Рубис А. Ю.* Метрическое пространство форм изображений // 9-я международная конференция. «Интеллектуализация обработки информации», 2012. С. 406–410.

[4] *Визильтер Ю. В., Рубис А. Ю., Горбацевич В. С.* Реляционные модели формы изображений и метрики их сравнения // 9-я международная конференция. «Интеллектуализация обработки информации», 2012. С. 410–414.

[5] *Vizilter Yu V., Gorbatsevich V. S., Rubis A. Yu., Zheltov S. Yu.* Shape-Based Image Matching Using Heat Kernels and Diffusion Maps // Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., 2014. Vol. 3. Pp. 357–364.

[6] *Vizilter Yu V., Zheltov S. Yu.* Geometrical Correlation and Matching of 2D Image Shapes // ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., 2012. No 3. Pp. 191–196.

[7] *Визильтер Ю. В., Выголов О. В., Рубис А. Ю.* Экспериментальное исследование морфологических методов сравнения форм изображений в задачах комплексирования многоспектральной видеоинформации // Вестник компьютерных и информационных технологий, 2013. № 8. С. 3–9.

"Euler Identity" for Morphological Image Analysis

Yury Vizilter[1]⋆ viz@gosniias.ru
Oleg Vygolov[1] o.vygolov@gosniias.ru
Sergei Zheltov[1] zhl@gosniias.ru
Stanislav Brianskiy[1] sbrianskiy@gosniias.ru
[1]Moscow, FGUP "GosNIIAS"

In the framework of Pyt'ev morphology [1] we consider images as piecewise-constant 2D functions

$$f(x,y) = \sum_{i=1}^{n} f_{Fi}\chi_{Fi}(x,y) \tag{13}$$

where n is a number of tessellation F on frame Ω with area S into connected regions of constnt intensity, $F = \{F_1, \ldots, F_n\}; f = (f_{F1}, \ldots, f_{Fn})^T$ is an intensity value vector. Such images we call the mosaic images.

Morphological comparison of mosaic shapes by complexity is traditionally implemented in terms of a partial order relation "not more complex by shape". The set of mosaic shapes has an algebraic lattice structure: for any shapes F and G we can find the more complex shape $F \bigwedge G$ and less complex shape $F \bigvee G$. More complex shapes are obtained by region splitting, and less complex shapes are obtained by regions merging. In terms of sets (classes) of images, F is not more complex than G, if $F \subseteq G$. In terms of morphological projectors, F is not more complex than G if $P_G P_F = P_F$.

In [2] we intoduce a generalized full-order relation based on the definition of the shape complexity measure, which is equal to 0 for the simplest shape O (one region on a frame) and increases monotonouslly with successive region splitting. For mosaic tessellation F with areas of regions $S_{Fi} = Sp_i, i = 1, .., n$ we proposed to use the GDD-complexity:

$$\mu_H(F) = \sum_{i=1,..,n} p_i(1-p_i) \tag{14}$$

which is connected with Geometric Difference Distance (GDD) between shapes [3]:

$$d_H(F,G) = \sum_{j=1,..,l} \sum_{i=1,..,n} S_{ij} d_H(F_i, G_j) \tag{15}$$

where $d_H(F_i, G_j) = S_{Fi} + S_{Gj} - 2S_{ij}$ is a Hamming distance between regions F_i and G_j; S_{ij} – area of region $F_i \cap G_j$ Such GDD metrics can be interpreted as a mosaic shape editing metrics. The other interpretation of GDD is connected with relational models of mosaic shapes [4].

We define the relational shape of the mosaic image as a predicate of the binary relation "pixels do not belong to the same region":

$$\pi_F(x,y,u,v) = \begin{cases} 0 & \text{if } \forall i : \chi_{Fi}(x,y) = \chi_{Fi}(u,v) \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

L^1-distance between relational models π_F and π_G is exactely equal to GDD-distance between mosaic tesselations F and G:

$$\begin{aligned} d_\pi(\pi_F, \pi_G) = \|\pi_F(x,y,u,v) - \pi_G(x,y,u,v)\| = \\ = \sum_{i=1,..,n} \sum_{j=1,..,m} S_{ij}(s_{Fi} + S_{Gj} - 2S_{ij}) = \\ = \sum_{i=1,..,n} \sum_{j=1,..,m} d_H(F_i, G_j) = d_H(F,G) \end{aligned} \tag{17}$$

Therefore, the normalized value of L^1-norm for relational model π_F:

$$\begin{aligned} \mu_H(\pi_F) = \|\pi_F\| - \|\pi_O\|/\|\pi_I\| - \|\pi_O\| = \\ = d_\pi(\pi_F, \pi_O)/d_\pi(\pi_I, \pi_O) = d_\pi(\pi_F, \pi_O)/d_\pi(\pi_I, \pi_O) = \\ = \sum_{i=1,..,n} S_{Fi}(1 - S_{Fi})/S^2 = \sum_{i=1,..,n} p_i(1-p_i) = \\ = d_H(F,O)/d_H(I,O) = \mu_H(F) \end{aligned} \tag{18}$$

is nothing other than the normalized GDD-complexity $\mu_H(F)$ (14).

Let us now consider the image comparison by shape, which is traditionally determined using a morphological projector [1]:

$$\mathrm{g}_F(x,y) = P_F g(x,y) = \sum_{i=1}^{n} \mathrm{g}_{Fi}\chi_{Fi}(x,y), \mathrm{g}_{Fi} = \frac{(\chi_{Fi}, \mathrm{g})}{\|\chi_{Fi}\|^2}, i = 1, ..., n$$

If image $\mathrm{g}(x,y)$ is mosaic too, then:

$$\begin{aligned} \mathrm{g}_F(x,y) = P_F \mathrm{g}(x,y) = P_F \sum_{j=1,..,m} \mathrm{g}_{Gj}\chi_{Gj}(x,y) = \\ = \sum_{j=1,..,m} \mathrm{g}_{Gj} P_F \chi_{Gj}(x,y) = \sum_{j=1,..,m} \mathrm{g}_{Gj}\chi_{GFj}(x,y), \chi_{GFj}(x,y) \\ = \sum_{i=1,..,n} \chi_{GFij}\chi_{Fi}(x,y), \chi_{GFij}(x,y) = \\ = (\chi_{Gj}(x,y), \chi_{Fi}(x,y))/\|\chi_{Gj}(x,y)\|^2 = \\ = S_{Wij}/S_{Fi}, i = 1, .., n; j = 1, .., m \end{aligned} \tag{19}$$

Thus, we obtain two alternative descriptions of the projection of $\mathrm{g}_F(x,y)$: based on a usual mosaic model $\langle \boldsymbol{\chi}_F, \mathbf{g}_F \rangle$ and a fuzzy model of $\langle \boldsymbol{\chi}_{GF}, \mathbf{g}_G \rangle$. Hence, the P_F operator can be considered no longer as an operator in image space, but as an operator in the space of mosaic shapes. In such scheme, the fuzzy (diffusion [5]) model G_F automatically arises as a projection of the mosic model onto another mosaic model:

$$\boldsymbol{\chi}_{GF} = P_F \boldsymbol{\chi}_G \iff G_F = P_F G$$

Let us define the Morphological Shape Correlation Corefficient (MSCC) as

$$K_M(G,F) = \|P_F G\|/||G\| = \|\boldsymbol{\chi}_{GF}(x,y)\|/\|\langle\boldsymbol{\chi}_G(x,y)\| \tag{20}$$

Then we substitute (19) into (20) and obtain the evident formula for $K_M(G,F)$:

$$K_M^2(G,F) = \|\boldsymbol{\chi}_{GF}(x,y)\|^2/\|\boldsymbol{\chi}_G(x,y)\|^2$$
$$= \sum_{j=1,..,m}\sum_{i=1,..,n} S_{ij}^2/(SS_{Fi}) = \sum_{j=1,..,m}\sum_{i=1,..,n} p_{ij}^2/p_{Fi} \tag{21}$$

where $p_{ij} = S_{ij}/S, p_{Fi} = S_{Fi}/S$.

Note that was previously receive the expression (21) from completely different, statistical considerations [6, 7]. We determined the mean square effective morphological correlation coefficient (MSEMCC) as the root of the ration of average square of the projection norm of the image shape F to the sape G to the average square of the projected image norm. The formula for the MSEMCC based on assumption of the mutual independence of the region intensities on G, turned out to be exactly the same as the expression we obtained here for the MSCC (21).

Finally, one can use the K_M^2 to compare shape F with simplest shape O:

$$K_M^2(F,O) = \sum_{i=1,..,n} \frac{s_{Fi}^2}{S^2} = \sum_{i=1,..,n} p_{Fi}^2 \tag{22}$$

and find that

$$\mu_H(F) = \sum_{i=1,..,n} p_{Fi}(1-p_{Fi}) = \sum_{i=1,..,n} p_{Fi} - \sum_{i=1,..,n} p_{Fi}^2 = 1 - K_M^2(F,O) \tag{23}$$

Thus, we found that the complexity measure $\mu_H(F)$ is directly related to the morphological correlation $K_M^2(F,O)$. Moreover, now we can outline:

$$\mu_H(F) = \|\pi_F - \pi_O\|_{L1}/\|\pi_I - \pi_O\|_{L1} = d_H(F,O)/d_H(I,O) =$$
$$= 1 - K_M^2(F,O) \tag{24}$$

This is a remarkable confirmation of the internal unity of Pytyev's morphological analysis in all of its forms: attribute and relational, metric and operator, geometric and statistical, mosaic (projective) and fuzzy (diffusion). This expression also demonstrates the hidden intrinsic relationship between morophlogic comparison by shape similarity/difference and by shape complexity. In some sense, for morphological analysis, this "unifying" equality is analogous to the famous Euler identity $e^{i\pi} + 1 = 0$ for complex analysis, since it brings together seemingly completely different aspects of the subject under consideration. In particular, the identity (24) shows that the GDD-distance is a really natural metrics in the shape space, which is directly related both to the mosaic and relational representation of shapes, and to the Pyt'ev's morphological projector and morphological correlation coefficient.

Funded by RSF, project 16-11-00082.

[1] *Pyt'ev Yu. P., Chulichkov A. I.* Morphological methods for image analysis // Moscow: Fizmatlit Publisher, 2010. Pp. 336.

[2] *Vizilter Yu. V., Rubis A. Yu.* Morphological Image Comparison by Complexity // Mathematical Methods of Pattern Recognition, 2013. Pp. 70.

[3] *Vizilter Yu. V., Rubis A. Yu.* Metric space of image shapes // Intellegent Information Processing IIP 9, 2012. Pp. 406–409.

[4] *Vizilter Yu. V., Rubis A. Yu., Gorbatsevich V. S.* Relational Shapes and Comparison Metrics // Intellegent Information Processing IIP 9, 2012. — p. 410–414.

[5] *Vizilter Yu V., Gorbatsevich V. S., Rubis A. Yu., Zheltov S. Yu.* Shape-Based Image Matching Using Heat Kernels and Diffusion Maps // Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., 2014. Vol. 3. Pp. 357–364.

[6] *Vizilter Yu V., Zheltov S. Yu.* Geometrical Correlation and Matching of 2D Image Shapes // ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., 2012. No 3. Pp. 191–196.

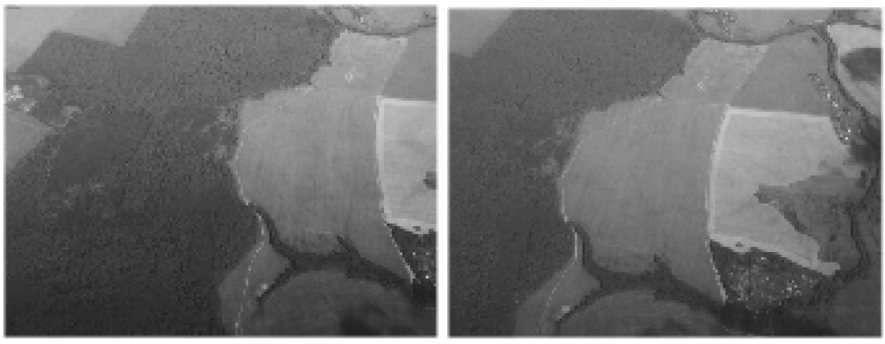
[7] *Vizilter Yu. V., Vygolov O. V., Rubis A. Yu.* Comparison of 2D Image Shape Similarity Measures // 11-th International Conference Pattern Recognition and Image Analysis: New Information Technologies, 2013. Pp. 345–348.

Применение квазиоптимальной кластеризации пикселей в задаче комплексирования разноракурсных изображений

Ханыков Игорь Георгиевич[1]⋆ igk@iias.spb.su
Ненашев Вадим Александрович[2] nenashev@guap.com
[1]Санкт-Петербург, Федеральный исследовательский центр Российской академии наук
[2]Санкт-Петербург, Государственный университет аэрокосмического приборостроения

В работе [1] предлагается методика совмещения разноракурсных локационных изображений в одно по точкам контура выделенных областей. Области выделяются путем использования метода высокоскоростной квазиоптимальной кластеризации пикселей изображения. Используемый метод кластеризации обходит проблему вычислительной сложности за счет разделения всего процесса обработки изображения на три последовательных этапа.

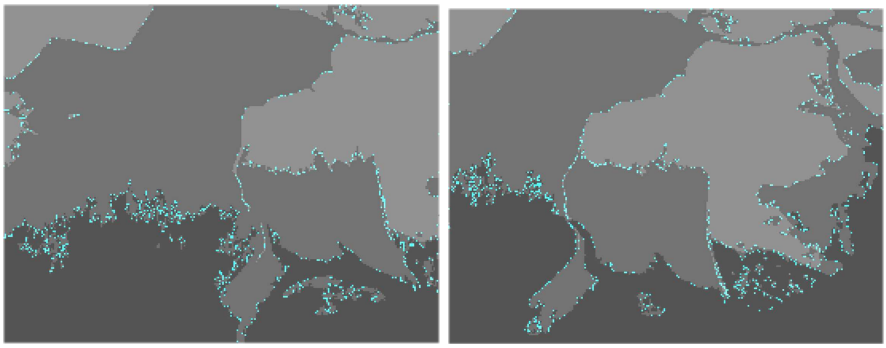
На входе алгоритма квазиоптимальной кластеризации, помимо изображения, задается точность вычислений, определяемая параметром числа суперпикселей Nсп. Параметр Nсп задает фиксированное число кластеров, для которого реструктуризуется иерархия разбиений. Диапазон значений задаваемого параметра Nсп – от 1 до N, где N – общее количество пикселей в изображении. Большему значению параметра Nсп соответствует лучшее качество реструктуризации, но и большее время обработки. На выходе алгоритм генерирует серию кусочно-постоянных разбиений. Количество выводимых разбиений задается пользователем в диапазоне от 1 до N.



(a) (b)

Рис. 1. Исходные левое и правое разноракурсные изображения: a) левое разноракурсное изображение; b) правое разноракурсное изображение

На первом этапе алгоритма квазиоптимальной кластеризации выполняется быстрое построение грубой иерархии сегментов применением либо модели Мамфорда-Шаха, либо классического метода Уорда по частям изображения.



(a) (b)

Рис. 2. Кластеризованные левое и правое разноракурсные изображения: a) левое изображение из 3 кластеров; b) правое изображение из 3 кластеров

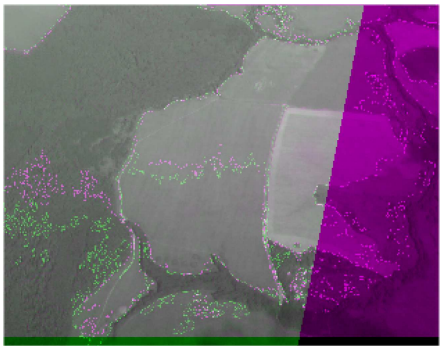


Рис. 3. Совмещенное изображение из двух исходных разноракурсных

На втором этапе выполняется промежуточное улучшение качества, формируется Nсп суперпикселей. На третьем этапе сформированные суперпиксели кластеризуются классическим методом Уорда.

Методика комплексирования разноракурсных изображений состоит из 4 шагов. На 1 шаге исходные разноракурсные изображения состыковывают в один единый снимок, который кластеризуется трехэтапным алгоритмом квазиоптимальной кластеризации пикселей при заданном числе Nсп. На выходе алгоритма получается серия разбиений с увеличивающимся числом кластеров. Совместная обработка изображений позволяет подобным образом выделить одинаковые по структуре области на разноракурсных изображениях. На 2 шаге из полученной серии кластеризованных разбиений выбирают одно, которое обратно разделяют на отдельные изображения. На выбранных и разделенных изображениях

выделяют контуры характерных областей и определяют точки схожих по структуре контуров. На 3 шаге через решение корреляционно-экстремальной задачи по точкам контуров, найденных на кластеризованных изображениях, подбирается функциональное преобразование, при котором значение функции корреляции принимает максимальное значение. Найденное функциональное преобразование применяют к исходным изображениям. На 4 шаге выполняется оценка качества сформированного совмещенного изображения. При неудовлетворительной степени субъективного восприятия переходят к предыдущим шагам методики совмещения. Либо на 2 шаге берем пару кластеризованных снимков с большим числом кластеров для последующего выделения новых характерных точек контуров и уточнения положения ранее найденных. Либо, перейдя к 1 шагу, увеличиваем значение параметра детализации Nсп. Повышая степень детализации сформированных снимков, увеличивается количество точек контуров, и как следствие уточняется вид искомого функционального преобразования. Методика уточнения функционального преобразования с повышением степени детализации изображения следует повторять до тех пор, пока качества совмещенного изображения ни будет приемлемым. Детализация осуществляется как за счет увеличения числа кластеров (цветов) в серии разбиений, так и за счет увеличения параметра точности квазиоптимальной обработки. Данная методика применима для бортовых локационных станций многопозиционных систем в целях комплексирования информации, в том числе разнородной.

На рисунках 1.a, 1.b представлены исходные левое и правое разноракурсные изображения. На рисунках 2.a, 2.b представлены левое и правое кластеризованные разноракурсные изображения, состоящие из трех кластров (цветов). На рисунке 3 представлен результат совмещения двух исходных изображений в одно.

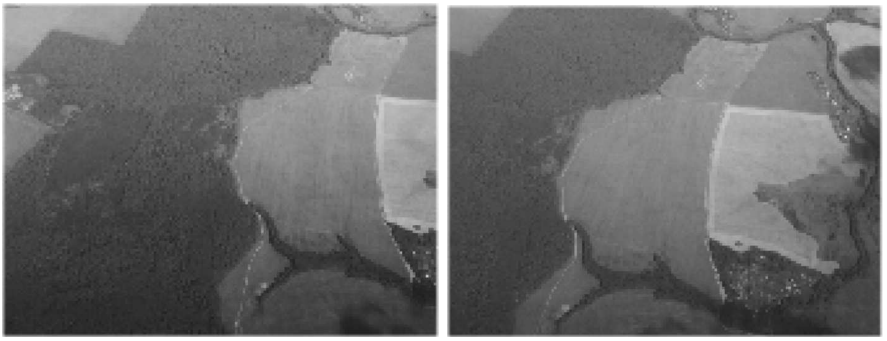
Исследование выполнено за счет гранта Российского научного фонда (проект No 19-79-00303).

[1] *Khanykov I. G.* The Application of the High-Speed Pixel Clustering Method in Combining Multi-Angle Images Obtained from Airborne Optical-Location Systems // XXIII International Conference on Wave Electronics and Infocommunication Systems, 2020. Pp. 1–7.

The application of quasi-optimal pixel clustering in the problem of combining multi-angle images

Igor Khanykov[1]⋆ igk@iias.spb.su
Vadim Nenashev[2] nenashev@guap.com
[1]Saint Petersburg, Federal Research Centre of the Russian Academy of Sciences
[2]Saint Petersburg, State University of Aerospace Instrumentation

The paper [1] proposes a technique for combining multi-angle location images into one at the points of the contour. Areas are isolated using the method of high-speed quasi-optimal clustering of image pixels. The used clustering method bypasses the problem of computational complexity by dividing the entire image processing process into three sequential stages.

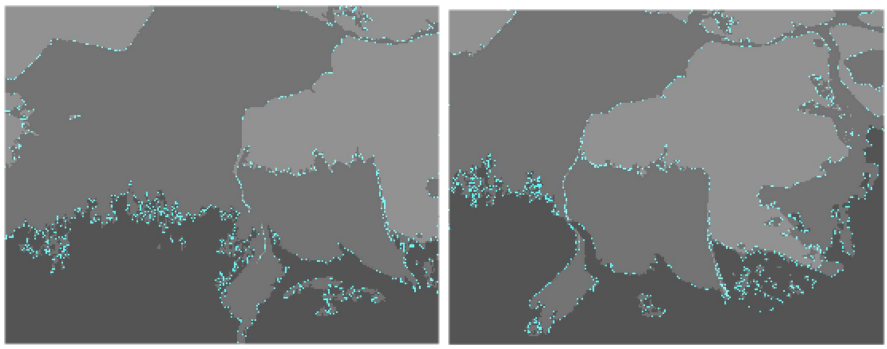


(a) (b)

Figure 4. Initial left and right multi-angle images: a) left multi–angle image; b) right multi-angle image

At the input of the quasi-optimal clustering algorithm, in addition to the image, the calculation accuracy is specified, which is determined by the parameter of the number of superpixels Nsp. The Nsp parameter specifies a fixed number of clusters for which the hierarchy of partitions is restructured. The range of values of the given parameter Nsp is from 1 to N, where N is the total number of pixels in the image. A larger value of the Nsp parameter corresponds to a better restructuring quality, but also a longer processing time. At the output, the algorithm generates a series of piecewise-constant partitions. The number of displayed splits is set by the user in the range from 1 to N.

At the first stage of the quasi-optimal clustering algorithm, a rapid construction of a rough hierarchy of segments is performed using either the Mumford-Shah model or the classical Ward's method by parts of the image. At the second stage, an intermediate quality improvement is performed, Nsp superpixels are formed. At third stage, the formed superpixels are clustered using the classical Ward's method.



(a) (b)

Figure 5. Clustered left and right multi-angle images: a) left image of 3 clusters; b) right image of 3 clusters

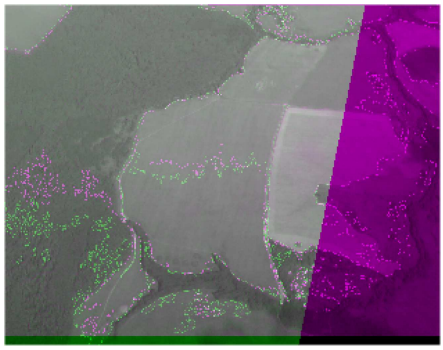


Figure 6. Combined image from two original multi-angle images

The technique of combining multi-angle images consists of 4 steps. At the 1 step, the original multi-angle images are docked into one single image, which is clustered by a three-stage quasi-optimal pixel clustering algorithm for a given number of Nsp. The output of the algorithm is a series of partitions with an increasing number of clusters. Joint processing of images allows to isolate in a similar way areas of the same structure in different multi-angle images. At the 2 step, one joint clustered partition is selected from the obtained series. Then it is divided back into separate parts. The contours of areas are distinguished and characteristic points of contours similar in structure are determined on the selected and divided images. At the 3 step, a functional transformation is selected through the solution of the correlation-extremal problem by the contours points found on the clustered images. The correlation function value must take the maximum value. The found functional

transformation is applied to the original images. At the 4 step, the quality of the generated combined image is assessed. If the degree of subjective perception takes on an unsatisfactory value, then we proceed to the previous steps of the combining technique. Or, at the 2 step, we take a couple of clustered images with a larger number of clusters for the subsequent selection of new characteristic contour points and refining the position of previously found. Or, proceeding to the 1 step, we increase the value of the detail parameter Nsp. Increasing the degree of detail of the formed images, the number of contour points rises, and as a result, the type of the desired functional transformation is refined. The technique of refining the functional transformation with increasing the degree of image detail should be repeated until the degree of quality assessment of the combined image is acceptable. Detailing is carried out both by increasing the number of clusters (colors) in a series of partitions, and by increasing the accuracy parameter of quasi-optimal processing. This technique is applicable to onboard radar stations of multi-position systems in order to integrate information, including heterogeneous information.

Figures 4.a, 4.b show initial left and right multi-angle iamges. Figures 5.a, 5.b show the left and right clustered multi-angle images, consisting of three clusters (colors). The figure 6 shows the result of combining two original images into one.

The research was carried out at the expense of a grant from the Russian Science Foundation (project No 19-79-00303).

[1] *Khanykov I. G.* The Application of the High-Speed Pixel Clustering Method in Combining Multi-Angle Images Obtained from Airborne Optical-Location Systems // XXIII International Conference on Wave Electronics and Infocommunication Systems, 2020. Pp. 1–7.

Обнаружение и сопровождение целей с БПЛА при помощи нейронных сетей

Калмыков Никита Сергеевич[1,2]⋆ kalmikov.nik@gmail.com
[1]Москва, Институт проблем управления им. В. А. Трапезникова РАН
[2]Королев, ООО «Альбатрос»

Комплексные технологии оперативного мониторинга территорий, различных объектов с помощью беспилотных летательных аппаратов, являются одними из наиболее перспективных и востребованных в ситуациях, требующих незамедлительного реагирования, но, в то же время, и весьма сложных для практической реализации. Особое значение оперативный мониторинг имеет для задач обслуживания транспортной инфраструктуры, охраны объектов, мониторинга территорий, нужд силовых ведомств и при предупреждении и ликвидации последствий чрезвычайных ситуаций, поиска людей. В таких задачах наиболее применимым является видео или тепловизионный мониторинг, из чего вырастает актуальность данной работы. Целью данной НИОКР является повышение комфорта работы оператора БПЛА через автоматизацию рабочих операций во время выполнения видео- и тепловизионного мониторинга с беспилотного летательного аппарата. Данная цель реализуется через создание следующих модулей для беспилотных летательных аппаратов:

1. гиростабилизированная трехосевая платформа бесконечного вращения с видеокамерой;
2. программное обеспечение управления полетом БПЛА;
3. специализированное ПО для определения объектов по видеопотоку, получаемому с камеры, и автоматического удержания захваченных объектов в поле зрения видеокамеры.

Таким образом, оператор БПЛА, выполняя работы по мониторингу, получает возможность указать БПЛА объект интереса и детально его изучить, не отвлекаясь на управление беспилотником, в то время как БПЛА в автоматическом режиме будет следовать за объектом интереса и держать его в поле зрения видеокамеры. Ключевые преимущества проектной технологии перед любыми другими видами мониторинга:

1. оперативность захвата цели;
2. автоматизация удержания захваченной цели;
3. площадь наблюдения;
4. автоматизация наблюдения.

Превосходством проектной технологии перед съемкой беспилотными аппаратами без модуля автоматического слежения за целью является повышение качества и надежности дистанционного мониторинга путем захвата и фиксации средств мониторинга на неподвижной или движущейся цели, что, в свою очередь, существенно повышает производительность и эффективность выполнения

работ по видеомониторингу. Также, среди преимуществ можно выделить более высокую надежность удержания цели в плоскости изображения по сравнению с беспилотными аппаратами, в которых полезная нагрузка (ПН) управляется оператором вручную. Данная работа поддержана Фондом содействия инноваций, контракт № 3577ГС2/60682 от 13.07.2020 по направлению «Разработка и испытания опытного образца автономной системы слежения за объектами по видеопотоку, получаемому с БПЛА»

Target detection and tracking using neural networks on UAV

Nikita Kalmykov* kalmikov.nik@gmail.com
Moscow, V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences
Korolyov, "Albatros" LTD

Integrated technologies for operational monitoring of territories and various objects using unmanned aerial vehicles are among the most promising and in demand in situations that require immediate response, but at the same time are very difficult to implement. Operational monitoring is of particular importance for the maintenance of transport infrastructure, security, monitoring of territories, the needs of law enforcement agencies, and in the prevention and elimination of the emergency situations, search and rescuing. In such problems most applicable are video and thermal monitoring, from which it grows relevance of this work. The purpose of this R&D is to improve the comfort of the UAV operator through the automation of work operations during video and thermal imaging monitoring from an unmanned aerial vehicle. This goal is realized through the establishment of the following modules for unmanned aerial vehicles:

1. gyro-stabilized three-axis platform of infinite rotation with a video camera;
2. UAV flight control software;
3. specialised object tracking software;

Thus, the UAV operator, carrying out monitoring work, is able to specify the UAV object of interest and study it in detail, not being distracted by the UAV control, while the UAV automatically follows the object of interest and keep it in the camera's field of view. Key advantages of design technology over any other types of monitoring:

1. efficiency of target acquisition;
2. automation of holding the captured target;
3. observation area;
4. automation of surveillance.

The superiority of the design technology over capturing without an automatic target tracking module aims to improve remote monitoring of quality and reliability by capturing and locking monitoring tools on the stationary or moving targets, which in turn significantly increases the productivity and efficiency of work. Also among the advantages we can allocate higher reliability retention purposes in the image as compared to unmanned vehicles, in which the payload is controlled manually by the operator. This work was supported by the Foundation to promote innovation, contract number No 3577GC2/60682 of 13.7.2020 on the subject "Development and testing of a prototype of autonomous tracking system for objects on a video stream from UAV"

Десять открытых проблем вероятностного тематического моделирования

Воронцов Константин Вячеславович[1]⋆ k.v.vorontsov@phystech.edu
[1]Московский Физико-Технический Институт

Вероятностная тематическая модель коллекции текстовых документов описывает каждый документ d дискретным распределением $p(t|d)$ на множестве тем t, а каждую тему t — дискретным распределением $p(w|t)$ на множестве слов w. Тематическое моделирование называют «мягкой» кластеризацией, так как каждый документ не относится «жёстко» к одному кластеру-теме, а моделируется вероятностной смесью тем: $p(w|d) = \sum_t p(w|t)p(t|d)$. В простейшей постановке задача сводится к построению неотрицательного матричного разложения и имеет бесконечное множество решений. Более сложные постановки включают различные критерии регуляризации, позволяющие учитывать дополнительные данные о документах или особенности строения текстов на естественных языках.

В результате моделирования строятся тематические векторные представления как документов $p(t|d)$, так и слов $p(t|w)$, причём не только глобальные для всей коллекции, но и локальные $p(t|d, w)$, определяемые контекстом слова w в документе d.

Данное обстоятельство потенциально ставит тематическое моделирование в один ряд с современными нейросетевыми моделями языка, которые также строят латентные векторные представления (эмбединги) любых текстовых фрагментов, при этом различая глобальные и локальные эмбединги слов.

Преимуществом тематических эмбедингов является их разреженность и интерпретируемость. Каждая координата вектора соответствует некоторой теме, которая описывается семантически связанными словами и фразами естественного языка. Образно говоря, любая тема способна сама рассказать о себе, тогда как нейросетевые эмбединги такой возможностью не обладают.

С другой стороны, тематические модели сильно уступают нейросетевым (BERT, GPT и их вариантам) по выразительным способностям, точности языкового моделирования и качеству решения трудных задач текстовой аналитики. Нейросетевые модели языка предсказывают следующее слово в предложении на уровне человека (значение перплексии около 10), тогда как перплексия тематических моделей обычно в сотни раз выше. Причина не только в том, что нейронные сети сложнее устроены, имеют больше параметров и кодируют в эмбедингах нетривиальную информацию о семантике (природу которой мы пока не понимаем и не умеем моделировать в явном виде). Тематические модели исходно предназначены лишь для верхнеуровневого понимания больших объёмов текста и ответов на общие вопросы: «какие темы представлены во всей коллекции», «какие темы затронуты в данном тексте» и «что ещё полезно знать об этих темах».

В докладе обсуждаются открытые задачи вероятностного тематического моделирования, возникающие в приложениях разведочного информационного поиска (exploratory search).

1. *Гарантирование качества тем.* На практике многие темы оказываются неинтерпретируемыми, неверно интерпретируемыми, мусорными, ошибочно разделёнными или объединёнными. Одна из естественных причин этих явлений кроется в несбалансированности тем, когда объём одних тем в коллекции превышает объём других в сотни раз. Критерий максимума правдоподобия имеет тенденцию выравнивать темы по объёму, а не по семантической однородности. Для несбалансированных коллекций необходимо модифицировать оптимизационный критерий.

2. *Разделение лексики на тематическую и общую.* Общая лексика мешает интерпретировать темы. Необходимы алгоритмы автоматического оценивания тематичности слов и словосочетаний, как в локальном контексте, так и в контексте коллекции, не требующие экспериментального подбора гиперпараметров.

3. *Моделирование тематики связного текста* с учётом порядка слов. Модели, основанные на гипотезе «мешка слов», слишком хаотично относят близко стоящие слова к разным темам. Перспективным подходом представляется моделирование внимания по аналогии с нейросетевыми моделями языка.

4. *Динамическое создание событийных тем* в текстовых потоках. Для выделения тем с коротким времени жизни необходима слаженная работа нескольких механизмов: классификации тем на событийные и перманентные, распознавания наличия новых тем в документе, определения того, что тема устарела.

5. *Обеспечение устойчивости тематических моделей.* Эксперименты показывают, что для построения полного набора существенно различных и хорошо интерпретируемых тем необходимы десятки или даже сотни запусков моделирования. Открытым остаётся вопрос, возможно ли построение полного набора тем за один запуск, и какие регуляризаторы для этого необходимы.

6. *Автоматизация подбора гиперпараметров.* При использовании нескольких регуляризаторов и модальностей их весовые коэффициенты обычно подбираются вручную, что требует многократного перестроения модели по всем данным. Необходим механизм быстрой адаптивной многокритериальной оптимизации гиперпараметров в режиме пакетной обработки коллекции.

7. *Бережное слияние нескольких коллекций* с выделением общих и сохранением уникальных тем. Определение общей тематики при объединении текстовых коллекций сильно затруднено, когда они несбалансированы по объёму.

8. *Автоматическое именование и описание тем.* Большинство приложений тематического моделирования связано с визуальным представлением тем пользователю. Для этого необходимо генерировать для любой темы краткое название, релевантные фразы и краткое связное описание (суммаризацию), оптимизированные под заданный объём визуального пространства.

9. *Создание предобученных тематических моделей* по большим текстовым коллекциям объёма не менее Википедии. Разработка методов дообучения тематических эмбедингов по новым данным, в том числе путём увеличения их размерности.

10. *Применение гиперграфовых тематических моделей*, уже имплементированных в библиотеке BigARTM с открытым кодом. Как наиболее широкое обобщение ARTM, они имеют большой потенциал в анализе разнородных транзакционных данных.

В докладе также рассматриваются новые способы повышения качества тематических моделей: статистические тесты для проверки условной независимости, регуляризаторы для быстрой векторизации текста и для несбалансированных коллекций.

Работа поддержана грантом РФФИ № 20-07-00936.

[1] *Ирхин И. А., Булатов В. Г., Воронцов К. В.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста // Компьютерные исследования и моделирование, 2020.

Ten open problems in probabilistic topic modeling

Konstantin Vorontsov[1]⋆ k.v.vorontsov@phystech.edu

[1]Moscow Institute of Physics and Technology

Probabilistic topic model of a text document collection represents each document d by a discrete distribution $p(t|d)$ over topics t, and describes each topic t with a discrete distribution $p(w|t)$ over words w. Topic model is a "soft" clustering technique, which model each document as a probabilistic mixture of topics $p(w|d) = \sum_t p(w|t)p(t|d)$ instead of "hard" assigning the document to only one cluster. In the simplest setting, learning the topic model is done by the likelihood maximization for a non-negative matrix factorization. This optimization problem is ill-posed because of it has an infinite number of solutions. More complex settings include various regularization criteria, allowing to take into account some external data about documents or some linguistic observations about natural language structures. Topic model results in probabilistic embeddings for both documents and words in a form of probability distributions $p(t|d)$ and $p(t|w)$ respectively. Moreover, the local context embeddings of words are also available $p(t|d, w)$ for each word position in the document collection.

This opportunity potentially puts topic modeling on a par with recent neural networks for language modeling, which also give embeddings for words and phrases, also distinguishing them in global and local contexts.

The advantage of topical embeddings is their sparseness and interpretability. Each vector coordinate corresponds to some topic, which is described by semantically related words and phrases. Figuratively speaking, any topic can tell about itself, whereas neural embeddings do not have such an opportunity.

On the other hand, topic models are much inferior to neural language models such as BERT, GPT and their variants in their expressive ability and accuracy in solving difficult text analysis tasks. Neural language models predict the next word in a sentence at the human level of perplexity about 10, whereas the perplexity of topic models is usually hundreds of times higher. The reason is not only that neural networks are more complex, they have more parameters and encode nontrivial semantic information (the nature of which we do not yet understand and do not know how to model in an explicit way). Topic models are originally intended only for high-level understanding of large amounts of text and answering to general questions such as "what topics are presented in the whole collection", "what topics are covered in this text" and "what else is good to know about these topics".

In this report we discuss ten open problems of probabilistic topic modeling, which arise in exploratory search and other topic modeling applications.

1. *The automatic quality assurance for topics.* In practice, topics can be uninterpreted, misinterpreted, noisy, mistakenly separated or combined. One of the natural reasons is the imbalance in topics when the size of some topics in the collection exceeds the size of others by hundreds of times. The likelihood maximization tends

to align all topics in their size rather than semantic integrity. The optimization criterion is to be modified for unbalanced collections.

2. *Separation of vocabulary into topical and common words.* Common words may obstruct the interpretation of topics. Therefore, automatic estimates of the topicality of words and phrases are to be elaborated for using in both local and global context. Automatic means that the calculation of these estimates should require no experimental selection of hyperparameters.

3. *Topic modeling of continuous text* taking into account the order of words. The bag-of-words hypothesis results in too chaotic intra-text topic dynamics. Borrowing ideas from self-attention neural language models seems to be a promising approach.

4. *New event topic detection in text streams.* The automatic detection of short-lifetime topics requires a well-coordinated interaction of several mechanisms: (i) classification of topics into event and permanent, (ii) recognition is there a new topic in the document, (iii) determining that a topic is out of date and can be removed.

5. *Ensuring the completeness of the topic set.* Experiments have shown that dozens or even hundreds of modeling runs are needed in order to find a complete set of different, well-interpreted topics. The question remains open, is it possible to build a complete set of topics in one run, and what regularizers are needed for this.

6. *Automatic hyperparameters learning.* When using multiple regularizers and modalities, the researcher usually selects their weighting factors manually. It takes a long time for multiple restarts of the model. An effective multi-criteria optimization technique is needed, in which the hyperparameters would be adaptively learned from sufficiently small batches instead of the entire collection.

7. *Gently merging heterogeneous collections* detecting which topics are common and which are unique for individual collections. Determining common topics becomes a challenging task in the case when collections to be merged are unbalanced in their size.

8. *Automatic topic labeling and topic summarization.* Most topic modeling applications has to do with visualization of topics to the user. Showing a topic title and/or a brief explanation of topic is very important in most visualization scenarios. However, little research is known about generating relevant topic summarizations within a given placement budget.

9. *Learning and sharing ready-to-use topic models*, pre-trained on large text collections of at least Wikipedia size. Development of techniques for fine tuning topical word embeddings on additional small data. including a technique for increasing the dimension. In particular, new techniques are needed to increase the dimension of embeddings both when new topics are discovered and when the functionality of the topical model is expanded.

10. *Application of hypergraph topic models* already implemented in `BigARTM` open source software. As the broadest generalization of ARTM, the hypergraph topic models have a great potential for processing big heterogeneous transactional data.

In the report, we also discuss new ways to improve the quality of topic models: statistical tests for checking conditional independence, regularizers for fast vectorization of text and for unbalanced collections.

This research is funded by RFBR, grant 20-07-00936.

[1] *Irkhin I. A., Bulatov V. G., Vorontsov K. V.* Additive Regularizarion of topic models with fast text vectorizartion // Computer research and modeling, 2020.

Поиск почти-дубликатов в рукописных текстах школьных сочинений

Бахтеев Олег Юрьевич[1,2] bakhteev@ap-team.ru
Кузнецова Рита Валерьевна[3] rita.kuznetsova@phystech.edu
Хазов Андрей Вячеславович[2] khazov@ap-team.ru
Огальцов Александр Владимирович[2,4] ogaltsov@ap-team.ru
Сафин Камиль Фанисович[1] kamil.safin@phystech.edu
Горленко Татьяна Александровна[2] gorlenko@ap-team.ru
Суворова Марина Алексеевна[2] suvorova@ap-team.ru
Ивахненко Андрей Александрович[2] ivahnenko@ap-team.ru
Чехович Юрий Викторович [2,5] chehovich@ap-team.ru
Моттль Вадим Вячеславович [5] vmottl@cass.ru

[1]Москва, Московский физико-технический институт (Государственный университет)
[2]Москва, Антиплагиат
[3]Цюрих, ETH
[4]Москва, Высшая школа экономики (Научно-исследовательский институт)
[5]Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)

Рассматривается задача поиска почти-дубликатов в коллекции школьных сочинениях. Актуальность задачи обсуловлена наличием больших библиотек школьных сочинений, которые могут использоваться школьниками в качестве источника заимствования при написании собственного сочинения. Данная проблема является важной для системы образования, что было обозначено в ряде работ [1, 2], посвященных анализу нарушений, выявляемых при написании сочинений и прохождении академических испытаний. Несмотря на массовость проблемы, на текущий момент не существует автоматических методов анализа сочинений на наличие корректных и некорректных цитирований, а также заимствований.

Задача поиска почти-дубликатов рассматривается как задача информационного поиска. Предполагается, что авторы сочинений с допущенными заимствованиями используют в качестве источника заимствования только один текст. Сочинение представляется набором изображений рукописного текста, написанного автором. Традиционные системы поиска почти-дубликатов и заимствований рассматривают в качестве объекта печатный текст [3]. Основные работы в области анализа текстов рукописных сочинений основаны на методах распознавания текста. Несмотря на успехи в области распознавания печатного текста, а также рукописного текста, написанного с использованием сенсорных устройств, применение данных методов для рассмотренной задачи затруднительно. В отличие от методов, применяемых для распознавания печатного текста, методы распознавания рукописного текста обладают достаточно низким качеством[4],

не позволяющим использовать их для поиска заимствований в тексте сочинений. Основной проблемой при обработке школьных сочинений является невозможность системы распознавания текста адаптироваться к большому числу вариантов почерка. Поскольку сочинения пишутся разными людьми, то возможность провести дообучение системы по почерку автора отсутствует, что также усложняет задачу. Другой проблемой методов поиска почти-дубликатов, основанных на распознавании текста, является обязательное наличие разметки — соответствия между регионом изображения и соответствующим ему текстом. Разметка подобного рода для рукописных текстов является трудозатратной, и встречается в открытом доступе крайне редко. Предлагаемый метод поиска почти-дубликатов не требует наличия детально размеченного текста, что позволяет применять его в большом количестве задач, связанных с извлечением информации из изображений рукописного текста.

Для решения задачи предлагается рассматривать текст, находящийся в изображении, как последовательность [5]. Предлагается метод, заключающийся в выделении слов в изображении для дальнейшего извлечения графических признаков. В качестве алгоритма извлечения слов применяется метод, основанный на выделении компонент связности [6]. Текст характеризуется последовательностью признаков, получение которых значительно проще, чем распознавание самого слова, что позволяет эффективнее работать с различными вариантами почерка. Примером таких признаков являются длина и высота слова в изображении, наличие для слова характерных лигатур. В рамках проведенного эксперимента текст характеризуется нормализованными длинами извлеченных из изображения слов. Полученные статистики являются инвариантными по отношению к почерку автора, а также могут использоваться как для рукописных, так и для машиночитаемых текстов. В качестве функции схожести полученных признаковых описаний школьных сочинений рассматривается набор методов выравнивания последовательностей и временных рядов [7, 8]. Для подтверждения работоспособности метода проводится эксперимент на выборке изображений рукописных текстов школьных сочинений.

Работа поддержана грантом РФФИ № 19-29-14100.

[1] *Ma H. J., Wan G., Lu E. Y.* Digital cheating and plagiarism in schools // Theory Into Practice, 2008. Vol. 47. No 3. Pp. 197–203.

[2] *Wrigley S.* Avoiding “de-plagiarism”: Exploring the affordances of handwriting in the essay-writing process // Active Learning in Higher Education, 2019. Vol. 20. No 2. Pp.167–179.

[3] *Журавлев Ю. И., и др.* Система распознавания интеллектуальных заимствований «Антиплагиат» // Математические методы распознавания образов, 2005. Т. 12. № 1. С. 329–332.

[4] *Puigcerver J.* Are multidimensional recurrent layers really necessary for handwritten text recognition? // 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) — IEEE, 2017. Vol. 1. Pp. 67–72.

[5] *Kuznetsov M., Motrenko A., Kuznetsova R., Strijov V.* Methods for intrinsic plagiarism detection and author diarization // Notebook for PAN at CLEF 2016.

[6] *Louloudis G. et al.* Text line and word segmentation of handwritten documents // Pattern recognition, 2009. Vol. 42. No 12. Pp. 3169–3183.

[7] *Salvador S., Chan P.* Toward accurate dynamic time warping in linear time and space // Intelligent Data Analysis, 2007. Vol. 11. No 5. Pp. 561–580.

[8] *Giorgino T., et al.* Computing and visualizing dynamic time warping alignments in R: the dtw package // Journal of statistical Software, 2009. Vol. 31. No 7. Pp.1–24.

Near-duplicate detection in handwritten school essays

Oleg Bakhteev[1,2] bakhteev@ap-team.ru
Rita Kuznetsova[3] rita.kuznetsova@phystech.edu
Andrey Khazov[2] khazov@ap-team.ru
Aleksandr Ogaltsov[2,4] ogaltsov@ap-team.ru
Kamil Safin[1] kamil.safin@phystech.edu
Tatyana Gorlenko[2] gorlenko@ap-team.ru
Marina Suvorova[2] suvorova@ap-team.ru
Andrey Ivahnenko[2] ivahnenko@ap-team.ru
Yury Chekhovich[2,5] chehovich@ap-team.ru
Vadim Mottl[5] vmottl@cass.ru

[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Antiplagiat
[3]Moscow, High School of Economics
[4]Zurich, ETH
[5] Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

The paper is devoted to the problem of near-duplicate detection in collections of school essays. Due to the large amount of online resources with available school essays the problem of text reuse detection is actual. This problem is important for the education system, which was noted in a number of works [1, 2], devoted to the analysis of cheating in writing essays and passing academic tests. Despite the importance of the problem, at the moment there are no automatic methods for text reuse detection available for the school essays analysis.

The task of near-duplicate detection is considered as an information search problem. It is assumed that authors of essays usually use only one text as a source of text reuse. The essay is represented as a sequence of scanned images of handwritten essay text. Traditional search engines for near-duplicate detection consider the text as an object to analyse [3]. Despite advances in recognition of printed text, as well as handwritten text written using touch devices, the use of these methods for current task is a challenge. Unlike the methods used for the recognition of printed text, the methods of handwritten text recognition have a rather low quality [4], which does not allow them to be used to search for the near-duplicates in the essays texts. The major problem in school essays text recognition is the inability of the text recognition system to adapt to a large variety of handwriting styles. Since the essays are written by different people, there is no opportunity to fine-tune to the specific author writing style, which also complicates the task. Another problem of near-duplicate detection methods based on text recognition is the requirements of the markup — the correspondence between the image region and the text in it. Such a markup can rarely be found in the public domain. The proposed method of near-duplicate detection does not require detailed markup text, which makes it possible to use it

in a large number of tasks related to the information extraction from the images of handwritten text.

The paper presents a method based on sequence analysis [5]. The image is segmented into words for further extraction of graphic features. For the word segmentation we use a method based on the connected components detection [6]. The text is characterized by a sequence of features, which are invariant to the author writing style. An example of such features is the length of a word in the image, the height or the presence of ligatures for a word. These features are invariant with respect to the author's handwriting style and can also be used for both handwritten and machine-readable texts. For the similar text detection we employ different methods of time series and sequence alignment [7, 8]. The computational experiment is conducted on the real dataset of the images of handwritten school essays.

This research is funded by RFBR, grant 19-29-14100.

[1] *Ma H. J., Wan G., Lu E. Y.* Digital cheating and plagiarism in schools // Theory Into Practice, 2008. Vol. 47. No 3. Pp. 197–203.

[2] *Wrigley S.* Avoiding "de-plagiarism": Exploring the affordances of handwriting in the essay-writing process // Active Learning in Higher Education, 2019. Vol. 20. No 2. Pp.167–179.

[3] *Zhuravlyov Y. et al.* The system of intellectual reuse detection Antiplagiat // Matematicheskie metody raspoznavanija obrazov, 2005. Vol. 12. No. 1. Pp. 329–332.

[4] *Puigcerver J.* Are multidimensional recurrent layers really necessary for handwritten text recognition? // 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) — IEEE, 2017. Vol. 1. Pp. 67–72.

[5] *Kuznetsov M., Motrenko A., Kuznetsova R., Strijov V.* Methods for intrinsic plagiarism detection and author diarization // Notebook for PAN at CLEF 2016.

[6] *Louloudis G. et al.* Text line and word segmentation of handwritten documents // Pattern recognition, 2009. Vol. 42. No 12. Pp. 3169–3183.

[7] *Salvador S., Chan P.* Toward accurate dynamic time warping in linear time and space // Intelligent Data Analysis, 2007. Vol. 11. No 5. Pp. 561–580.

[8] *Giorgino T., et al.* Computing and visualizing dynamic time warping alignments in R: the dtw package // Journal of statistical Software, 2009. Vol. 31. No 7. Pp.1–24.

Оценка близости смысловому эталону без поиска перифраз и иерархия тематических текстов

Михайлов Дмитрий Владимирович[1]⋆ mdv74@list.ru
Емельянов Геннадий Мартинович[1] gennady.emelyanov@novsu.ru
[1]Великий Новгород, Новгородский государственный университет

Настоящая работа посвящена проблеме численного оценивания взаимной смысловой зависимости тематических текстов относительно наиболее рациональных (эталонных) вариантов описания представляемых ими фрагментов знаний. Данная проблема актуальна при определении значимости источников информации относительно решаемых пользователем задач. Примером является поиск оптимального порядка работы с первоисточниками при формировании индивидуальной образовательной траектории обучаемого (студента). В предлагаемом решении [1] основой оценки близости текста эталону является разбиение слов каждой его фразы на классы по значению меры TF-IDF относительно текстов корпуса D, предварительно формируемого экспертом. Близость текста эталону оценивается без поиска перифраз. В роли анализируемых текстов выступают аннотации научных статей вместе с их заголовками. Для группы фраз **Ts**, первая из которых — заголовок статьи, а остальные представляют аннотацию, используются два предложенных авторами ранее варианта оценки близости эталону, в равной мере предусматривающие минимум среднеквадратического отклонения (СКО) значения близости эталону по всем фразам группы. Первый подразумевает максимальную близость эталону для заголовка. Второй предполагает максимизацию близости эталону по всем фразам анализируемого текста. Максимальный итоговый рейтинг по коллекции, из которой производится отбор, получает статья с наибольшим значением первого варианта, лежащим в одном кластере со значением второго варианта оценки для той же статьи. При этом значение первого варианта оценки для статьи с максимальным рейтингом и наибольшее значение первого варианта оценки по коллекции должны быть в одном кластере. В случае отсутствия в коллекции статьи, отвечающей данному требованию, максимальный итоговый рейтинг получает статья с наибольшим значением первого варианта оценки по анализируемой коллекции. Рекурсивным применением данного принципа исходная коллекция сортируется по убыванию итогового рейтинга статей. Смысловые образы наиболее близких эталону текстов будут определять слова с наибольшими значениями TF-IDF, которые при расположении по соседству в линейном ряду фразы с наибольшей вероятностью связаны по смыслу и образуют ключевые сочетания вместе со словами, близкими среднему значению данной меры. Для отнесения сочетания слов к ключевым используется интерпретация меры TF-IDF, раздельно учитывающая случаи совместной встречаемости слов сочетания и встречаемость без одновременного вхождения во фразу. При этом значение TF-IDF

ключевого сочетания не должно быть ниже минимума указанной меры по его отдельным словам.

Для построения иерархии документов отсортированной коллекции S_{res} используется аналогия с задачей вероятностного тематического моделирования, где иерархия тем моделирует стратегию поиска с постепенным фокусированием внимания пользователя на подтемах. Применительно к значениям TF-IDF ключевых терминов в решаемой авторами задаче это выражается в более высоких значениях TF-IDF этих терминов в родительском документе по сравнению с дочерним в формируемой иерархии.

Пусть X — упорядоченная по убыванию последовательность значений TF-IDF слов исходной фразы относительно некоторого документа $d \in D$; $H_1 \dots H_r$ — последовательность кластеров, на которые разбивается X. При этом для оценки близости фразы смысловому эталону содержательный интерес представляют слова кластеров H_1 (слова-термины исходной фразы, наиболее уникальные для d), $H_{r/2}$ (общая лексика, обеспечивающая перифразы, и термины-синонимы), а также H_r (слова-термины, преобладающие в корпусе). Введём обозначения: $\mathbf{H}_1(Ts_i)$, $\mathbf{H}_{r/2}(Ts_i)$ и $\mathbf{H}_r(Ts_i)$ — множества слов кластеров H_1, $H_{r/2}$ и H_r, соответственно, для фразы $Ts_i \in \mathbf{Ts}$ относительно документа $d \in D$, по которому получен максимум близости эталону, $\mathbf{Ts} \in S_{res}$; $\mathbf{H_1}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \mathbf{H}_1(Ts_i)$; $\mathbf{H}_{\overline{Z}}(Ts_i)$ — множество слов фразы Ts_i с ненулевыми значениями TF-IDF относительно того же документа d; $\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \big(\mathbf{H}_{\overline{Z}}(Ts_i) \setminus \mathbf{H}_1(Ts_i)\big)$. По определению, в составе S_{res} документы (группы фраз) расположены по убыванию итогового рейтинга. Пусть $\mathbf{Ts}_i$ и $\mathbf{Ts}_j$ — тексты из входящих в S_{res}, причём $i > j$, то есть рейтинг статьи, отвечающей группе фраз $\mathbf{Ts}_i$, выше, чем по $\mathbf{Ts}_j$. Выдвигается следующая гипотеза: мера, в которой текст $\mathbf{Ts}_j$ дополняется по смыслу текстом $\mathbf{Ts}_i$, соответствует величине $\big|\big(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\big) \cap \mathbf{H_1}(\mathbf{Ts}_i)\big|$.

Сама дополняемость текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$ определяется как

$$K_1\big(\mathbf{Ts}_j, \mathbf{Ts}_i\big) = \frac{\big|\big(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\big) \cap \mathbf{H_1}(\mathbf{Ts}_i)\big|}{\big|\mathbf{H_1}(\mathbf{Ts}_i)\big|}. \tag{1}$$

Пусть $\mathbf{Kw}(\mathbf{Ts}_i)$ — множество ключевых сочетаний слов, найденных для $\mathbf{Ts}_i$, $\mathbf{H_{Kw}}(\mathbf{Ts}_i)$ — множество слов в составе этих сочетаний. Введём в рассмотрение $\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i) \subset \mathbf{Kw}(\mathbf{Ts}_i)$, которое получается следующим образом: по каждой фразе $Ts_{jk} \in \mathbf{Ts}_j$ в него войдут сочетания слов множества $\mathbf{H}_{\overline{Z}}(Ts_{jk}) \setminus \mathbf{H_1}(\mathbf{Ts}_j)$, причём для каждого сочетания минимум одно слово должно принадлежать $\mathbf{H_1}(\mathbf{Ts}_i)$. При этом в целях учёта искомых сочетаний из множества в числителе формулы (2) исключаются слова в составе сочетаний из $\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)$, а к самому значению в числителе прибавляется $\big|\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)\big|$. Аналогично из множества в знаменателе исключают-

ся слова в составе $\mathbf{H_{Kw}}(\mathbf{Ts}_i)$, а само значение в знаменателе увеличивается на $\left|\mathbf{Kw}(\mathbf{Ts}_i)\right|$.

Для подтверждения наличия связи смыслового эталона текста $\mathbf{Ts}_j$ с эталоном текста $\mathbf{Ts}_i$ в дополнение к (2) вводится оценка представленности слов в кластерах $H_1, H_{r/2}$ и H_r для фраз $Ts_{jk} \in \mathbf{Ts}_j$, предполагающая максимизацию $|\mathbf{H}_\mathrm{m}(Ts_{jk})|\,/len(Ts_{jk})$ при минимуме СКО указанной величины по всем $\mathrm{m} \in \{1, \mathrm{r}/2, \mathrm{r}\}$, где $len(Ts_{jk})$ — число слов во фразе Ts_{jk}. При этом $\mathbf{H}_1(Ts_{jk})$ объединяется с множеством $\left(\mathbf{H}_{\overline{\mathrm{Z}}}(Ts_{jk}) \setminus \mathbf{H}_1(Ts_{jk})\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$, а из $\mathbf{H}_{\mathrm{r}/2}(Ts_{jk})$ и $\mathbf{H}_\mathrm{r}(Ts_{jk})$ элементы указанного множества удаляются. Если в целом по $\mathbf{Ts}_j$ значение вышеупомянутой оценки при этом не убывает, то текст $\mathbf{Ts}_i$ допустим в качестве вышестоящего для $\mathbf{Ts}_j$ в формируемой иерархии. Окончательный выбор вышестоящего текста ведётся по максимуму оценки (2).

Работа поддержана грантом РФФИ № 19-01-00006.

[1] *Михайлов Д. В., Емельянов Г. М.* Иерархизация тематических текстов на основе оценки близости смысловому эталону без перефразирования // Pattern Recognition and Image Analysis, 2020. Т. 30, № 3. С. 440–449.

Estimation for the closeness to a semantic pattern without paraphrasing, and a hierarchy of topical texts

Dmitry Mikhaylov⋆ mdv74@list.ru
Gennady Emelyanov gennady.emelyanov@novsu.ru
Russia, Veliky Novgorod, Yaroslav-the-Wise Novgorod State University

The offered work is devoted to the problem of numerical estimation of mutual semantic dependence of topical texts concerning the most rational linguistic variants (i. e. semantic patterns or sense standards) of the description of the knowledge fragments their represent. The closeness of a text to its standard is estimated herewith without revelation of periphrases. This problem is relevant when estimating the significance of information sources concerning the tasks solved by the user. As an example of practical application here can be the search of optimal order of how the trainee (in particular, a student) should work with primary sources at the creation of its educational trajectory. In the suggested solution [1], the basis of estimation of the closeness of the text to the semantic pattern is the splitting of the words of each of its phrases into classes by the TF"=IDF metric value relative to texts of a corpus D preformed by an expert. Abstracts of scientific papers together with their titles are analyzed. For a group of phrases **Ts**, first of which is the title of the article and others represent its abstract, two variants for estimation of the affinity to the sense standard are used. Both variants are previously proposed by authors and equally assumed the minimum of root"=mean"=square deviation (RMSD) for the value of affinity to the standard for all phrases of the group. The first variant assumes the maximal closeness to the standard for the article title. The second variant assumes maximizing the affinity to the standard for all phrases of the analyzed text. The maximal final rank in the collection for paper selection will be designated to the article with a greatest value of the first variant of estimation related to the same cluster with the value of the second variant for the same paper. Herewith the value of the first estimation variant for article with a maximal final rank, and a maximal value of the first estimation variant in the collection must be in the same cluster. In a case of absence of article meets this requirement, the maximal final rank will be designated to the article with a greatest value of the first variant of estimation in analyzed collection. By recursively using of given principle, the initial collection will be sorted descending the values of the final rank of articles. Here, the semantic images of the texts closest to the semantic pattern specify the words with the highest TF"=IDF values, which, when placed next to each other in the linear series of a phrase, are, most probably, semantically related and form key combinations with words whose mentioned metric is close to average. To identify a key combination of words the interpretation of TF"=IDF metrics which separately takes into account the cases of co"=occurrence of combination words and occurrence without a simultaneous presence in a phrase is used. Herewith the value of TF"=IDF metrics for

key word combination should not be less than the minimum of values of mentioned measure for its separate words.

To form a hierarchy of documents from the sorted collection S_{res} we use the analogy with the task of probabilistic topic modeling, where the topical hierarchy emulates a natural human strategy to focus on subtopics. Concerning TF"=IDF values of key terms in the problem being solved by authors it is expressed in a greater TF"=IDF values of these terms in a parent document in comparison with a child in the formed hierarchy.

Let X be a descent"=ordered sequence of TF"=IDF values for words of the initial phrase relatively to a document $d \in D$; $H_1 \dots H_r$ be the sequence of clusters as a result of splitting the initial X. Herewith the most important clusters to estimate the affinity of some phrase to the semantic pattern will be the cluster H_1 (the terms from the source phrase which are the most unique for d); the "median" cluster $H_{r/2}$ which will host general vocabulary that ensures periphrases and synonymous terms; and the cluster H_r to which the terms that prevail in the corpus have corresponded.

Let's enter the following denotations: $\mathbf{H}_1(Ts_i)$, $\mathbf{H}_{r/2}(Ts_i)$, and $\mathbf{H}_r(Ts_i)$ are sets of words of clusters H_1, $H_{r/2}$ and H_r, respectively, for the phrase $Ts_i \in \mathbf{Ts}$ relative to the document $d \in D$, concerning which the maximum of affinity to the standard has been achieved, $\mathbf{Ts} \in S_{res}$; $\mathbf{H_1}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \mathbf{H}_1(Ts_i)$; $\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_i)$ is a set of words of the phrase Ts_i with nonzero values of TF"=IDF relative to the same document d; $\mathbf{H}_{\bar{\mathbf{Z}}}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \big(\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_i) \setminus \mathbf{H}_1(Ts_i)\big)$. According to the definition, within S_{res} documents (i. e. phrase groups) are ordered descending their final ranks. Let $\mathbf{Ts}_i$ and $\mathbf{Ts}_j$ be texts from S_{res} and $i > j$, i. e., the rank of the article that matches a phrase group $\mathbf{Ts}_i$ is higher than for $\mathbf{Ts}_j$. Hypothesize the following: the measure of how the text $\mathbf{Ts}_j$ is complemented by a sense of the text $\mathbf{Ts}_i$ has corresponded to the value $\big|\big(\mathbf{H}_{\bar{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\big) \cap \mathbf{H_1}(\mathbf{Ts}_i)\big|$.

The sense complementarity of text $\mathbf{Ts}_j$ by text $\mathbf{Ts}_i$ can be defined as

$$K_1\big(\mathbf{Ts}_j, \mathbf{Ts}_i\big) = \frac{\big|\big(\mathbf{H}_{\bar{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\big) \cap \mathbf{H_1}(\mathbf{Ts}_i)\big|}{\big|\mathbf{H_1}(\mathbf{Ts}_i)\big|}. \tag{2}$$

Let $\mathbf{Kw}(\mathbf{Ts}_i)$ be a set of key combinations of words found for $\mathbf{Ts}_i$, $\mathbf{H_{Kw}}(\mathbf{Ts}_i)$ be the set of words within these combinations. Let's enter into consideration $\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i) \subset \mathbf{Kw}(\mathbf{Ts}_i)$, which is formed as follows: for each phrase $Ts_{jk} \in \mathbf{Ts}_j$ it will include a combination of words from the set $\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_{jk}) \setminus \mathbf{H_1}(\mathbf{Ts}_j)$, and for each combination, at least one word must belong to $\mathbf{H_1}(\mathbf{Ts}_i)$. Herewith to take into account the desired combinations, words of combinations from $\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)$ are excluded from the set in the numerator of formula (2), and to the numerator value, $\big|\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)\big|$ is added. Similarly, from the set in denominator words of $\mathbf{H_{Kw}}(\mathbf{Ts}_i)$ will be excluded, and the denominator value is increased by $\big|\mathbf{Kw}(\mathbf{Ts}_i)\big|$.

To confirm the relation of the sense standard for the text $\mathbf{Ts}_j$ with the sense standard of $\mathbf{Ts}_i$ in addition to formula (2) the estimation of representation of words in clusters $H_1, H_{r/2}$, and H_r for phrases $Ts_{jk} \in \mathbf{Ts}_j$ is entered. This estimation assumes the maximization of $|\mathbf{H}_{\mathrm{m}}(Ts_{jk})| / len(Ts_{jk})$ at a minimum of RMSD of the mentioned value for all $\mathrm{m} \in \{1, \mathrm{r}/2, \mathrm{r}\}$, where $len(Ts_{jk})$ is the number of words in the phrase Ts_{jk}. Herewith $\mathbf{H}_1(Ts_{jk})$ unites with the set $\left(\mathbf{H}_{\overline{2}}(Ts_{jk}) \setminus \mathbf{H}_1(Ts_{jk})\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$, and elements of this set will be removed from $\mathbf{H}_{\mathrm{r}/2}(Ts_{jk})$ and $\mathbf{H}_{\mathrm{r}}(Ts_{jk})$. If on the whole for $\mathbf{Ts}_j$ the value of the above"=mentioned estimation does not decrease here, then the text $\mathbf{Ts}_i$ can be assumed as a parent for $\mathbf{Ts}_j$ in the formed hierarchy. The criterion of the final choice of parent text is the maximum of estimation (2).

This research is funded by RFBR, grant 19-01-00006.

[1] *Mikhaylov D., Emelyanov G.* Hierarchization of topical texts based on the estimate of proximity to the semantic pattern without paraphrasing // Pattern Recognition and Image Analysis, 2020. Vol. 30. No 3. Pp. 440–449.

Генерация SPARQL-запросов для ответа на сложные вопросы с помощью BERT и BiLSTM

Евсеев Дмитрий Андреевич* dmitrij.euseew@yandex.ru
Архипов Михаил Юрьевич arkhipov@yahoo.com
Москва, Московский физико-технический институт

Вопросно-ответная система — это информационная система, которая получает вопрос на естественном языке, обрабатывает его и генерирует ответ. Вопросно-ответные системы активно используются в виртуальных ассистентах (Алиса, Amazon Alexa, и т. д.).

Вопросно-ответные системы могут использовать в качестве источников набор текстовых документов или базы знаний. Преимущество баз знаний состоит в структурированном представлении фактов по сравнению с текстом, поэтому ответы на вопросы при использовании баз знаний краткие и точные.

В данной работе описывается вопросно-ответная система для ответа на сложные вопросы по базе знаний Wikidata. В отличие от простых вопросов, для ответа на которые требуется найти один факт в базе знаний, сложные вопросы требуют извлечения более 1 триплета, а также логические или сравнительные рассуждения. Предложенная система переводит вопрос на естественном языке в запрос на языке SPARQL, выполнение которого дает ответ. В состав системы входят модели, которые определяют шаблон SPARQL-запроса, соответствующего вопросу, и затем заполняют пустые места в шаблоне сущностями, отношениями и численными значениями. Для извлечения сущностей мы использовали модель маркировки последовательностей на основе BERT. Ранжирование возможных отношений для вопроса происходит в два этапа с помощью моделей на основе BiLSTM и BERT. Предложенные модели - первое решение для датасета LC-QUAD2.0. Система способна отвечать на вопросы, требующие сравнительное или логическое рассуждение.

[1] *Evseev D. A., Arkhipov M. Yu.* SPARQL query generation for complex question answering with BERT and BiLSTM-based model // 26th International Conference on Computational Linguistics and Intellectual Technologies, 2020. Pp 270–282.

SPARQL query generation for complex question answering with BERT and BiLSTM-based model

Dmitry Evseev ⋆ dmitrij.euseew@yandex.ru
Mikhail Arkhipov arkhipov@yahoo.com
Moscow, Moscow Institute of Physics and Technology

Question answering system is an informational system which recieves a natural language question, processes it and returns the answer. Question answering systems are widely used in virtual assistants (Alisa, Amazon Alexa etc.).

Question answering systems can look for an answer in a collection of raw text documents or in knowledge bases. Knowledge bases are structured sources so they return short, precise and interpretable answers.

In this work we describe question answering system for answering of complex questions over Wikidata knowledge base. Unlike simple questions, which require extraction of single fact from the knowledge base, complex questions are based on more than one triplet and need logical or comparative reasoning. The proposed question answering system translates a natural language question into a query in SPARQL language, execution of which gives an answer. The system includes the models which define the SPARQL query template corresponding to the question and then fill the slots in the template with entities, relations and numerical values. For entity detection we used BERT-based sequence labelling model. Ranking of candidate relations is performed in two steps with BiLSTM and BERT-based models. The proposed models are the first solution for LC-QUAD2.0 dataset. The system is capable of answering complex questions which involve comparative or boolean reasoning.

[1] *Evseev D. A., Arkhipov M. Yu.* SPARQL query generation for complex question answering with BERT and BiLSTM-based model // 26th International Conference on Computational Linguistics and Intellectual Technologies, 2020. Pp 270–282.

Методы Big Math и интеграция математических знаний

Елизаров Александр Михайлович[1] amelizarov@gmail.com
Липачев Евгений Константинович[1]⋆ elipachev@gmail.com
[1]Казань, Казанский (Приволжский) федеральный университет

Термин “Big Data”, широко используемый в настоящее время в различных предметных областях, применительно к математике требует определенных уточнений: в математике все данные существенны, кроме того, в математических документах многие их части, особенно формулы, являются своеобразным кодом, требующим расшифровки и специального толкования. Далее, при решении математических задач являются существенно большими ожидания от использования ИКТ. Здесь можно провести аналогию с тем, как вычислительные машины полностью устранили ручные вычисления. Вычисления всегда требовали применения особых методов и нестандартных организационных решений, позволяющих справиться с объемом (Volume – одна из характеристик больших данных) и преодолеть барьер вычислительных возможностей отдельного человека. Если говорить о Velocity как одной из характеристик больших данных, то длительность ручных вычислений иллюстрирует пример вычисления числа Пи: В. Шенкс (William Shanks, 1873 г.) потратил 15 лет на вычисление 707 знаков этого числа (однако только 555 из них оказались верными). Помимо вычислений и подготовки документов необходимы инструменты интеллектуального поиска, в том числе, рекомендательные системы для нахождения научных статей, близких по содержанию; сервисы терминологического аннотирования; персональные информационные помощники и цифровые платформы для автоматизации издательской деятельности.

Недавно J. Carette, W.M. Farmer, M. Kohlhase и F. Rabe (arXiv:1904.10405v1 [cs.MS] 23 April 2019) предложили использовать, по аналогии с Big Data, термин Big Math для обозначения области создания методов и разработки программных систем поддержки математических исследований. Ими выделены пять основных аспектов Big Math:

• Inference (вывод утверждений путем дедукции),

• Computation (алгоритмическое преобразование представлений математических объектов в формы, более легкие для понимания),

• Tabulation (создание статических, конкретных данных, относящихся к математическим объектам и структурам, которые можно легко хранить, запрашивать и совместно использовать),

• Narration (приведение результатов в форму, которая может быть усвоена людьми),

• Organization (модульная организация математических знаний).

Основная задача математических программных систем заключается сегодня в интеграции указанных аспектов, составляющих Big Math. Система цифровых математических библиотек, создаваемая в настоящее время, призвана консоли-

дировать и сделать доступными как современные математические знания, так и математические знания, содержащиеся в документах, созданных в доцифровой период. Для достижения этой цели в рамках цифровых библиотек разрабатываются методы управления цифровой информацией, учитывающие особенности представления математического контента.

В области интеграции математических знаний наиболее значительными являются инициатива Global Digital Mathematics Library (GDML, https://doi.org/10.1007/978-3-319-62075-6_5) и проект World Digital Mathematics Library (WDML, https://arxiv.org/ftp/arxiv/papers/1404/1404.1905.pdf). Его основная задача – объединение в распределенной системе электронных коллекций всего корпуса цифровых математических документов. На интеграцию европейских математических ресурсов направлен проект The European Digital Mathematics Library (EuDML, https://initiative.eudml.org/). Этот проект рассматривается как один из этапов построения WDML.

В соответствии с основными принципами WDML в Казанском университете создается цифровая математическая библиотека Lobachevskii Digital Mathematics Library (Lobachevskii-DML, https://lobachevskii-dml.ru/). Построение этой библиотеки предполагает разработку инструментов управления математическим контентом, учитывающих не только специфику математических текстов, но и особенности обработки русскоязычных текстов. Еще одной задачей этой цифровой библиотеки является интеграция математических ресурсов Казанского университета и их включение в глобальную научную инфраструктуру, в частности, MathNet.Ru и EuDML.

В исследованиях, выполненных нашей группой (см. [1]), разработаны подходы к управлению большими коллекциями цифровых математических документов, основанные на семантических методах и согласующиеся с принципами WDML, а также относящиеся к направлениям, составляющим Big Math. Эти подходы развиваются и уже частично практически реализованы в цифровой математической библиотеке Lobachevskii-DML. Предложены методы формирования цифровых коллекций из набора документов – научных статей, монографий, докладов, представленных в различных форматах хранения. На основе анализа структуры документов и стилевых особенностей их оформления разработан алгоритм экстракции их метаданных. Создан программный инструмент разделения сборников статей на отдельные документы и формирования их семантического представления. На примере «Трудов Математического центра им. Н.И. Лобачевского», имеющих различные формат и структуру, реализован алгоритм создания цифровой коллекции и ее включения в Lobachevskii-DML.

Разработаны: алгоритмы пополнения электронных коллекций цифровой библиотеки Lobachevskii-DML и формирования метаданных документов этих коллекций в выбранных форматах; сервисы нормализации этих метаданных в соответствии с DTD-правилами и XML-схемами NISO JATS и DBLP; алгоритмы

создания обязательного и фундаментального наборов метаданных коллекций в соответствии с правилами EuDML.

Работа выполнена в рамках программы развития Регионального научно-образовательного математического центра Приволжского федерального округа, соглашение № 075-02-2020-1478/1.

[1] *Elizarov A. M. and Lipachev E. K.* Big Math Methods in Lobachevskii–DML Digital Library // Data Analytics and Management in Data Intensive Domains, 2019. Pp. 59–72.

Big Math Methods and Mathematical Knowledge Integration

Alexander Elizarov[1] amelizarov@gmail.com
Evgeny Lipachev[1]⋆ elipachev@gmail.com
[1] Kazan, Kazan (Volga Region) Federal University

The term "Big Data", which is currently widely used in various subject areas, in relation to mathematics requires certain clarifications: in mathematics, all data is essential, in addition, in mathematical documents, many of their parts, especially formulas, are a kind of code that requires decoding and special interpretation. Further, when solving mathematical problems, expectations from the use of ICT are significantly higher. An analogy can be drawn here with how computers completely eliminated manual computation. Computing has always required the use of special methods and non-standard organizational solutions to cope with volume (Volume is one of the characteristics of big data) and overcome the barrier of the computational capabilities of an individual. If we talk about Velocity as one of the characteristics of big data, then the duration of manual calculations illustrates an example of calculating the number Pi: W. Shanks (William Shanks, 1873) spent 15 years calculating 707 digits of this number (however, only 555 of them turned out to be correct). In addition to calculations and preparation of documents, intelligent search tools are needed, including recommendation systems for finding scientific articles that are similar in content; terminological annotation services; personal information assistants and digital platforms for publishing automation.

J. Carette, W.M. Farmer, M. Kohlhase and F. Rabe (arXiv: 1904.10405v1 [cs.MS] 23 April 2019) proposed to use, by analogy with the term Big Data, the term Big Math to denote the field of creating methods and developing software systems to support mathematical research. They highlighted 5 main aspects of Big Math:

- Inference (output of statements by deduction);
- Computation (algorithmic transformation of representations of mathematical objects into forms that are easier to understand);
- Tabulation (creating static, specific data related to mathematical objects and structures that can be easily stored, queried and shared);
- Narration (bringing the results into a form that people can assimilate);
- Organization (modular organization of mathematical knowledge).

The main task of mathematical software systems today is to integrate the aspects that make up Big Math. The system of digital mathematical libraries currently being created is intended to consolidate and make accessible both modern mathematical knowledge and the knowledge contained in articles and books published in the pre-digital period. To achieve this goal, in the framework of digital libraries, methods for managing digital information are developed that take into account the characteristics of the presentation of mathematical content.

In the area of integrating mathematical knowledge, the most significant are the Global Digital Mathematics Library initiative (GDML, https://doi.org/10.1007/978-3-319-62075-6_5) and the World Digital Mathematics Library project (WDML, https://arxiv. org/ftp/arxiv/papers/1404/1404.1905.pdf). Its main task is to unite the entire corpus of digital mathematical documents in a distributed system of electronic collections. The European Digital Mathematics Library project (EuDML, https://initiative.eudml.org/) aims to integrate European mathematical resources. This project is considered as one of the stages of building WDML.

In accordance with the basic principles of WDML, a digital library Lobachevskii Digital Mathematics Library (Lobachevskii-DML, https://lobachevskii-dml.ru/) is being created at the Kazan University. The construction of this library involves the development of management tools for mathematical content that take into account not only the specifics of mathematical texts, but also the peculiarities of processing Russian-language texts. Another objective of this digital library is the integration of the mathematical resources of Kazan University and their inclusion in the global scientific infrastructure, in particular, MathNet.Ru and EuDML.

In the studies carried out by our group (see [1]), approaches to managing large collections of digital mathematical documents based on semantic methods and consistent with the WDML principles, as well as related to the directions that make up Big Math, have been developed. These approaches are being developed and already partially practically implemented in the digital mathematical library Lobachevskii-DML. Methods for the formation of digital collections from a set of documents – scientific articles, monographs, reports, presented in various storage formats are proposed. Based on the analysis of the structure of documents and the style features of their design, an algorithm for extracting their metadata has been developed. A software tool has been created for dividing collections of articles into separate documents and forming their semantic representation. On the example of "Proceedings of the N.I. Lobachevskii Mathematical Center", which have different formats and structures, an algorithm for creating a digital collection and its inclusion in Lobachevskii-DML was implemented.

We have developed: algorithms for replenishing electronic collections of the digital library Lobachevskii-DML and forming metadata for documents of these collections in selected formats; services for normalizing this metadata in accordance with DTD rules and NISO JATS and DBLP XML schemas; algorithms for creating mandatory and fundamental sets of collection metadata in accordance with EuDML rules.

The work was carried out within the framework of the development program of the Regional Scientific and Educational Mathematical Center of the Volga Federal District, agreement No. 075-02-2020-1478/1.

[1] *Elizarov A. M. and Lipachev E. K.* Big Math Methods in Lobachevskii–DML Digital Library // Data Analytics and Management in Data Intensive Domains, 2019. Pp. 59–72.

Вариационное моделирование правдоподобия с триплетными ограничениями в задачах информационного поиска

Кузнецова Рита Валерьевна[1]⋆ rita.kuznetsova@phystech.edu

[1]Московский физико-технический институт

Задача построения отображения данных, пришедших из разных источников (доменов), в одно общее скрытое пространство (пространство оценок) [1], крайне актуальна для многих областей машинного обучения, таких как информационный поиск, перевод между доменами и генерация объектов [2]. В данной работе рассматриваются данные, которые являются разными модальностями одного объекта и являются объектами одинаковой структуры. Примером задачи с данными одинаковой структуры может являться машинный перевод — дана выборка соответствующих предложений на разных языках, нужно построить отображение из одного языка в другой. Также, по одной имеющейся модальности объекта можно осуществлять кросс-доменный поиск и формировать поисковую выдачу, состоящую из соответсвующих модальностей другого домена. Примерами такого поиска могут являться тематический кросс-языковой поиск [3], или поиск фотографий одной и той же местности с разных ракурсов [4].

В данной работе предлагается обучать отображение объектов доменов в одно общее пространство меньшей размерности, чем исходное. В этом пространстве, с помощью полученных скрытых представлений этих объектов, отражающих все модальности, с ними можно будет работать напрямую, решая поставленные задачи и не прибегая к промежуточным этапам, например таким как использование отдельной системы для перевода между доменами. Промежуточные этапы обладают еще одним недостатком — при сведении данных разных доменов к одному, могут теряются важные характеристики объектов доменов. Другой актуальной задачей является генерация данных, в случае работы с двумя доменами — генерация условно-похожих пар. Таким образом, предлагаемая модель должна быть *генеративной* — исследователь должен иметь возможность генерировать пары из общего пространства, например, для задач пополнения выборок. Удобным инструментом для оценки $p(\mathfrak{D})$ и обучения внутренних представлений является вариационный автокодировщик, представленный в [5], с помощью которого можно проще решить все вышеперечисленные задачи, используя вариационные методы [6, 5].

При решении различных практических задач анализа данных часто приходится сталкиваться с выборками, содержащими ошибки разметки [9]. Исследования показывают, что генеративные модели [8] довольно уязвимы для шума в данных. Очень небольшое возмущение в выборке может легко обмануть модель. На практике, собрать выборку пар высокого качества сложно и дорого с точки зрения затрат человеческого труда и времени. Из-за непрекращающегося роста данных для таких областей, как обработка естественного языка, компьютер-

ное зрение, речь и т.д. эта задача становится практически невозможной. Таким образом, предлагаемый алгоритм должен быть в состоянии использовать эти объекты для обучения.

В данной работе, для создания алгоритма, устойчивого к небольшому количеству выбросов в выборке, предлагается использовать подход из метрического обучения, а именно так называемых "триплетных ограничений" [10]. Использование относительных ограничений как информации о том, что один объект более близок к другому, чем некоторый третий (также распространено название "ложный сосед"), позволяет форсировать обучение модели в верном направлении, при этом однако не накладывая жестких правил на соответствие объектов друг другу. В данной работе идея модифицируется для случая двух доменов — предполагается, что объекты в паре, составленной из разных доменов, близки друг другу, однако на каждой итерации обучения выбирается "ложный сосед" из других объектов доменов, не входящих в пару. Таким образом, формируется триплет. Моделируя правдоподобие такой тройки объектов (объекты в паре и выбранный "ложный сосед") и используя его в основной модели правдоподобия, как компоненту штрафа, можно научить модель разносить объекты в некорректно сопоставленной паре. Если выбранный "ложный сосед" оказывается ближе к объекту в паре, чем назначенный, накладываемый на модель штраф разносит их, не позволяя выучить некорректное соответствие. Модификация триплетных ограничений для случая двух доменов позволит решить рассматриваемые в данной работе задачи. Так как модель содержит штраф, основанный на триплетных ограничениях, это позволяет снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач.

Вычислительный эксперимент проведен на различных выборках, содержащих изображения и текстовые данные. Показано, что предлагаемый метод сопоставим и превосходит предыдущие методы на этих наборах данных.

Работа выполнена при финансовой поддержке РФФИ (проект № 18-07-01441)

[1] *Рудаков К. В.* Алгебраическая теория универсальных и локальных ограничений для алгоритмов распознавания // М.: ВЦ РАН, 1992.

[2] *Suzuki M., Nakayama K. Matsuo Y.* Joint multimodal learning with deep generative models // 2016.

[3] *Wei L., Deng Z.* A Variational Autoencoding Approach for Inducing Cross-lingual Word Embeddings // Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017. Pp. 4165–4171.

[4] *Yunjey C., Min-Je C., Munyoung K., Jung-Woo H., Sunghun K., Jaegul C.* StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation // Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017. Pp. 4165–4171.

[5] *Kingma D., Welling M.* Auto-encoding variational bayes // 2013.

[6] *Jordan M., Ghahramani Z., Jaakkola, T., Saul L.* An introduction to variational methods for graphical models // Machine learning, 1999. Vol. 37. No 2. Pp. 183–233.

[7] *Fetaya E., Jacobsen J., Grathwohl W., Zemel R.* Understanding the limitations of conditional generative models // 2019.

[8] *Futami F., Sato I., Sugiyama M.* Variational inference based on robust divergences // International Conference on Artificial Intelligence and Statistics, 2018. Pp. 813–822.

[9] *Angluin D., Laird P.* Learning from noisy examples // Machine Learning, 1998. Vol. 2. No 4. Pp. 343–370.

[10] *Karaletsos T., Belongie S., Rätsch G.* Bayesian representation learning with oracle constraints // 2015.

Variational Bi-domain Triplet Modeling in Information Retrieval

Rita Kuznetsova[1]⋆ rita.kuznetsova@phystech.edu

[1]Moscow Institute of Physics and Technology

Learning distributed representations from data [1] that comes from different domains is one of the most challenging task in many machine learning problems [2] like information retrieval, translation and object generation.

Recent advances in probabilistic deep generative models allow us to specify a model as joint probability distribution over the data and latent variable consider the representations as samples from the posterior distribution on latent variables given data. In this work we consider the case when data comes from different domains, where the domains have a similar structure (e. g. texts [3], images [4]). We want to construct the common latent space for such kind of data. There we also assume that we have the correspondence between domains. We work with a paired dataset $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}, \mathbf{y})_i\}_{i=1}^N$, where $\mathbf{X}$ and $\mathbf{Y}$ are domains. But this dataset may contain errors, i.e. pairs with incorrect correspondence. In addition, the correspondence between domain objects can be many to many. Our main goals in such problem settings are:

— To learn powerful latent representations of domain data in order to work with them directly. For example, by using latent representations we can compare texts in different languages — which is an important task for cross-lingual plagiarism detection. We could do this without using an intermediate step-like machine translation. In addition, there are language pairs for which machine translation produces poor quality owing to insufficient parallel data.
— To use the learned distributed representations for domain adaptation task.
— To extend the model objective for successful work with noisy data and with data that consists a limited number of labeled examples.

Advances in variational autoencoders (VAEs) [5] that estimate the data using variational inference [6] with a few assumptions about data distribution and approximate posterior distribution. They make it possible to use latent variables as our learned representation. As pairs may be noisy [9], i.e. with irrelevant correspondence, we extend the objective function of our approach by the relative constraints or learning triplets to help our model capture the domain characteristics and similarity between domain objects. This constraint also penalizes our model in case of noisy pairs. The key idea of triplets comes from the metric learning approach: point a should be more similar to point b than to point c. We model this information about object similarity in the latent space like [10], but our main contribution is completely different: we do this in the shared latent space across domains to deal with noisy datasets or datasets with a limited number of labels. Thus we use implicit knowledge about the data and improve performance.

Thus the proposed approach builds the joint probability of domains $\mathbf{X}$, $\mathbf{Y}$ with the help of metric learning constraints. The performance of the proposed model on

different tasks such as bi-directional image generation, image-to-image translation and cross-lingual document classification (even on non-parallel data). We show that our method is comparable to existing methods performed on these datasets. It outperforms some state-of-the-art methods.

This research is funded by RFBR, grant 18-07-01441.

[1] *Rudakov K. V.* Algebraicheskaja theorija raspoznavanija obrazov // 1992.

[2] *Suzuki M., Nakayama K. Matsuo Y.* Joint multimodal learning with deep generative models // 2016.

[3] *Wei L., Deng Z.* A Variational Autoencoding Approach for Inducing Cross-lingual Word Embeddings // Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017. Pp. 4165–4171.

[4] *Yunjey C., Min-Je C., Munyoung K., Jung-Woo H., Sunghun K., Jaegul C.* StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation // Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017. Pp. 4165–4171.

[5] *Kingma D., Welling M.* Auto-encoding variational bayes // 2013.

[6] *Jordan M., Ghahramani Z., Jaakkola, T., Saul L.* An introduction to variational methods for graphical models // Machine learning, 1999. Vol. 37. No 2. Pp. 183–233.

[7] *Fetaya E., Jacobsen J., Grathwohl W., Zemel R.* Understanding the limitations of conditional generative models // 2019.

[8] *Futami F., Sato I., Sugiyama M.* Variational inference based on robust divergences // International Conference on Artificial Intelligence and Statistics, 2018. Pp. 813–822.

[9] *Angluin D., Laird P.* Learning from noisy examples // Machine Learning, 1998. Vol. 2. No 4. Pp. 343–370.

[10] *Karaletsos T., Belongie S., Rätsch G.* Bayesian representation learning with oracle constraints // 2015.

Математические и лингвистические аспекты моделирования медиадискурса

Кальян Виктор Петрович[1,2] vkalyan@mail.ru

[1]Москва, Российский Университет Дружбы Народов(РУДН)

[2]Москва, Федеральный Исследовательский Центр «Информатика и Управление» Российской Академии Наук

Опыт исследований сферы медиадискурса вплотную подводит нас к необходимости исследований влияния медиа на общественное мнение и возможности прогнозирования динамики их взаимодействия. Работа посвящена применению точных методов к изучению медиадискурса, разработке стратегии и инструментария его моделирования.

Информация о составе когерентных интенциональных групп и их отношении к происходящим событиям извлекалась из текстов дискуссий в социальных сетях, критических статей в прессе, телевизионных и радио- обзоров. Именованные сущности выявлялись с помощью метода вероятностного тематического моделирования. Отношения между ними как и морфология изучаемой ситуации устанавливались с помощью метода разметки семантической ролей (SRL).

Статика и динамика развития ситуации, взаимовлияние интенциональных групп и образование семантических доминант моделировалась с помощью методов теории игр, оптимального управления, параметрического резонанса. Этот же инструментарий использовался при прогнозировании динамики реакций аудитории, определении силы коалиций мнений и выявление истоков зарождающихся конфликтов. Результаты сопоставлялись с данными статистических опросов.

В работе предложена стратегия интеграции результатов работы исследователей в виде сети взаимо-портируемых моделей при использовании единого метаязыка в системах искусственного интеллекта. Этот подход может быть применён к моделированию ситуаций, описанию процессов как по текстовым, так и графическим, фото, видео и аудио данным.

Публикация выполнена при поддержке Программы стратегического академического лидерства РУДН.

[1] *Кальян В. П.* Modeling media discourse. Goals, strategies, tools // Ситуация, язык, речь. Модели и приложения. Полные тексты докладов конференций SLS 2018/2019, 2020.

Mathematical and linguistic aspects of media discourse modeling

Victor Kaliyan[1,2] vkalyan@mail.ru

[1]Moscow, Peoples Friendship University of Russia (RUDN University)

[2]Moscow, Federal Research Center "Computer Science and Control" of RAS.

The experience of research in the sphere of media discourse brings us very close to the need to research the influence of media on public opinion and the possibility of predicting the dynamics of their interaction. The work is devoted to the application of precise methods to the study of media discourse, the development of strategies and tools for its modeling.

Information about the composition of coherent intentional groups and their attitude to current events was extracted from the texts of discussions on social networks, critical articles in the press, television and radio reviews. Named entities were identified using probabilistic topic modeling. The relationship between them, as well as the morphology of the studied situation, were established using the semantic roles markup method (SRL).

The statics and dynamics of the situation development, the mutual influence of intentional groups and the formation of semantic dominants were modeled using the methods of game theory, optimal control, and parametric resonance. The same methods were used to predict the dynamics of audience reactions, determine the coalitions opinions strength and identify the origins of incipient conflicts. The results were compared with the data of statistical surveys.

The report proposes a strategy for integrating the results of researchers' work in the form of a inter-portable models network using a single metalanguage in artificial intelligence systems. This approach can be applied to modeling situations, describing processes using both text and graphic, photo, video and audio data.

The work shows that mathematical and linguistic methods of extracting knowledge from data, automatic understanding of texts, modeling situations are urgently needed and meet the most urgent requirements of the theory and practice of communication.

This paper has been supported by the RUDN University Strategic Academic Leadership Program.

[1] *Kaliyan V.* Modeling media discourse. Goals, strategies, tools. // Situation, Language, Speech. Models and Applications. Full texts of reports of conferences SLS 2018/2019, 2020.

Задачи систем обнаружения заимствований в применении к поиску заимствований в учебных работах средней школы

Беленькая Ольга Сергеевна[1] belenkaya@antiplagiat.ru
Суворова Марина Алексеевна[1] suvorova@antiplagiat.ru
Филиппова Ольга Анатольевна[1] filippova@antiplagiat.ru
Чехович Юрий Викторович[2] yury.chekhovich@gmail.com

[1]Москва, Антиплагиат

[2]Москва, Федеральный исследовательский центр Информатика и управление Российской академии наук (ФИЦ ИУ РАН)

Системы обнаружения заимствований (ОЗ) стали за последние 15 лет привычным инструментом преподавателей в высших учебных заведениях [1]. Системами ОЗ не менее активно пользуются при проверке научных работ [2]: статей, диссертаций, монографий и т.п. Преподаватели отмечают, что вчерашние школьники, становясь студентами, оказываются не готовы к тому, что в их письменных работах начинают контролировать количество и качество заимствований. В то же время общеобразовательная система как в России, так и за рубежом практически не использует инструментарий для поиска заимствований.

Изменения, которые в нашу жизнь принесла пандемия COVID-19, приводят к серьезному усилению роли дистанционного обучения во всех ступенях образования. Очевидно, что в дальнейшем доля дистанционного образования в структуре занятий продолжит увеличиваться. Это, в свою очередь, приводит к увеличению доли работ, которые ученики выполняют самостоятельно и результаты которых предоставляются в виде тех или иных письменных форматов: эссе, сочинений, презентаций, рефератов и т.п. То есть форматов, которые в наибольшей степени подвержены рискам использования некорректных заимствований.

При этом шаблонный перенос в школу инструментов и практик, используемых в вузах, обречен на неудачу. Во-первых, работы школьников обладают существенной спецификой, в том числе они имеют небольшой средний размер, могут содержать высокую долю заимствованного текста. Во-вторых, предъявляются более высокие требования к отсутствию орфографических ошибок. Наконец, требования, предъявляемые к школьным работам, очевидно, должны быть гораздо более мягкими, по сравнению с требованиями к студенческим работам.

Основная цель проверки работ заключается не в наказании, а в знакомстве школьников с принципами подготовки самостоятельных письменных работ, наработке опыта работы с источниками, приобретении навыков корректного указания ссылок на использованные материалы. Этой целью определяется и основной набор дополнительных к проверке на заимствования задач, которые

система ОЗ должна решать, чтобы быть пригодной к использованию в среде общего образования.

К таким задачам относится предложение корректного оформления цитат с обнаруженным заимствованным текстом. Ключевыми проблемами здесь являются поиск первоисточника работы, для чего необходимо наличие корректной и полной датировки возможных источников, определение допустимого объема цитаты, учитывающего тип задания и предполагаемый объем итогового документа. Для решения обеих задач используются методы машинного обучения. Для настройки алгоритмов по датировке источников осуществляется генерация признаков как основанных на содержании документа, так и на взаимном расположении объектов верстки. Задача определения допустимого объема цитаты может оказаться существенно более сложной в силу отсутствия подготовленных наборов данных для обучения.

Другой важной задачей является поиск подходящих источников. Инструментом для поиска может стать тематическое моделирование, с помощью которого для обрабатываемого документа осуществляется поиск подходящих по теме источников. Естественно, в силу специфики общего образования, в дополнение к тематическому поиску понадобится редуцирование поисковой выдачи, например, за счет коллаборативной фильтрации.

Еще одной необходимой функцией является контроль списывания, то есть использования похожих текстов учениками в близкое время. Для решения такой задачи могут использоваться относительно простые методы поиска нечетких дубликатов [3].

Задача автоматической проверки орфографии является, с одной стороны, исторически самой традиционной, с другой стороны, находящейся наиболее далеко от получения решений приемлемого качества. На данный момент создание алгоритмов, обеспечивающих проверку орфографии в школьных работах, является вызовом, с которым ученым в области машинного обучения и обработки естественных языков предстоит работать в течение нескольких лет [4].

С точки зрения использования средств ОЗ в общем образовании, важным вопросом оказывается определение основных пользователей этих средств. Если переносить опыт высших учебных и научных учреждений, то пользователями должны быть учителя. Однако описанная выше общеобразовательная специфика подсказывает, что ученики должны, как минимум, знакомится с результатами проверок. В идеале школьник должен получать промежуточные результаты и подсказки в процессе подготовки работы. Это позволило бы достичь понимания учениками «границы дозволенного» в относительно мягкой форме.

Очевидно, что многие из описанных решений пока не реализованы на уровне ведущих промышленных систем ОЗ. Тем не менее, все перечисленные задачи представляются пригодными для решения методами современного мирового уровня или методами, которые будут определять мировой уровень в ближайшее время.

Работа поддержана грантом РФФИ № 19-29-14130.

[1] *Никитов А. В., Орчаков О. А., Чехович Ю. В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия // Университетское управление: практика и анализ, 2021. № 5 (81). С. 61–68.

[2] *Чехович Ю. В., Беленькая О. С.* Оценка корректности заимствований в текстах научных публикаций // В сборнике: Научное издание международного уровня – 2018: редакционная политика, открытый доступ, научные коммуникации. Материалы 7-й международной научно-практической конференции, 2018. С. 158–162.

[3] *Зеленков Ю. Г., Сегалович И. В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, 2007.

[4] Про//чтение – Технологический конкурс UP Great —- URL: `https://ai.upgreat.one/about-project/` (дата обращения: 02.12.2020)

Tasks of text reuse detection systems when applied to the text reuse detection in secondary school written works

Olga Belenkaya[1] belenkaya@antiplagiat.ru
Marina Suvorova[1] suvorova@antiplagiat.ru
Olga Filippova[1] filippova@antiplagiat.ru
Yury Chekhovich[2] chehovich@antiplagiat.ru
[1]Moscow, Antiplagiat
[2]Moscow, Federal Research Center Computer Science and Management Russian Academy of Sciences (FRC CSM RAS)

Text reuse detection (TRD) systems have become a common tool for professors and instructors at higher education institutions over the past 15 years [1]. TRD systems are also actively used to analyze scientific papers, dissertations, monographs, etc. [2]. Professors note that recent pupils, becoming students, are not ready for their paperwork to be checked for quantity and quality of text reuse. At the same time educational systems both in Russia and abroad hardly use tools for text reuse detection.

The changes brought to our lives by the COVID-19 pandemic are leading to a dramatic increase of distance learning role at all levels of education. It is obvious that the proportion of distance learning in the structure of classes will continue to grow in the future. This leads to an increase in the proportion of research that pupils do by themselves and present the results in written formats: essays, presentations, abstracts, etc. And these are the formats most prone to incorrect text reuse.

Herewith a simple transfer of the tools and practices used in higher education to secondary schools will fail. Firstly, school written works possess significant differences due to several factors including small average size and high proportion of reused text. Secondly, less orthographic mistakes are allowed. Finally, the requirements applicable to school written works are obviously should be much more lenient comparing to student papers.

The main goal of checking of works is not to punish pupils, but to show them the principles of self-organized preparation of written works, to gain experience in working with sources, and to acquire the skills of making correct citations to the used materials. This goal also determines the main set of tasks, additional to reuse detection, that TRD system must perform in order to be suitable for use in the general education environment.

Such tasks include proposing correct formatting of the quotations of the detected reused text. The key problems here are to find the original source of the work, wherefore all possible sources must have correct and complete publish dates, and to determine the quotation volume that is allowed taking into account the type of the work and the estimated volume of the final document. Machine learning methods are used to solve both problems. Features are generated both based on the content of the document and on the relative position of the layout objects to fine-tune algo-

rithms for dating sources. Determination of the acceptable quotation amount may be much more difficult task due to the lack of prepared training data sets.

Search for suitable sources is another important task. Topic modeling can be an appropriate tool to find topic-related sources for the processed document. Due to the specifics of general education, reducing results of the topic search, for example, through collaborative filtering is required.

Another important feature is control of cheating when pupils are using similar texts at the same time. Relatively simple methods can be used for searching fuzzy duplicates to solve this problem [3].

The automatic spell check task, on the one hand, is historically the most traditional, but on the other hand, lacks solutions with an acceptable quality. At the moment, development of the algorithms performing spell check in school written works is a challenge that specialists in machine learning and natural language processing will work on for several years [4].

It is important to define main users of TRD systems to use these tools in general education. If we transfer the experience of higher educational and scientific institutions then teachers should be the users. However, the general educational specifics described above suggest that pupils should at least be familiar with the check results. Ideally, the pupils should receive intermediate results and tips while preparing their written work. This would allow pupils to understand "the boundaries" in a relatively mild form.

Obviously, many of the described solutions have not yet been implemented in the leading industrial TRD systems. Nevertheless, all of the above tasks seem to be solvable by the state-of-the-art methods or methods that will determine the state-of-the-art in the near future.

This research is funded by RFBR, grant 19-29-14130.

[1] *Nikitov A., Orchakov O., Chekhovich Yu.* Plagiarism in the written works of students and postgraduates: the problem and methods of counteraction // Universitetskoe upravlenie: praktika i analiz, 2012. No 5 (81). Pp. 61–68.

[2] *Chekhovich Yu., Belenkaya O.* Assessment of the text reuse correctness in scientific publications // In the digest: Scientific world-level publication – 2018: editorial policy, open access, scientific communications. Materials of the 7th International Scientific and Practical Conference, 2018. Pp. 158–162.

[3] *Zelenkov Yu., Segalovich I.* Comparative analysis o near-duplicate detection methods of Web-documents // Proceedings of the 9th All-Russian Scientific Conference ≪Digital Libraries: Advanced Methods and Technologies, Digital Collections≫ – RCDL'2007, 2007.

[4] Pro//chtenie – Technology competition UP Great —- URL: `https://ai.upgreat.one/about-project/` (request date: 02.12.2020)

Анализ рельефа кремниевых пластин методами геоморфометрии

Дедкова Анна Александровна[1]⋆ dedkova@ckp-miet.ru
Флоринский Игорь Васильевич[2] iflor@mail.ru
[1]Москва, Национальный исследовательский университет «МИЭТ»
[2]Пущино, ИМПБ РАН, филиал ИПМ им. М. В. Келдыша РАН

Анализ рельефа поверхности кремниевых пластин и мембран на пластинах является важным этапом производства микроэлектромеханических систем [1]. В частности, такой анализ позволяет измерять деформацию (изгиб) этих объектов и рассчитывать величину их механических напряжений. При этом используют цифровые модели рельефа (ЦМР) пластин и мембран с субмиллиметровым разрешением и перепадом высот микрометрового диапазона, которые получают способом оптической профилометрии.

При анализе рельефа пластин и мембран обычно ограничиваются определением прогиба. Иногда применяется аппарат классической дифференциальной геометрии: анализируются модели Гауссовой и главных кривизн, которые рассчитываются по ЦМР этих объектов. Нами применены методы геоморфометрии, в основе которой лежит теория топографической поверхности, существенно развивающая классические положения.

Анализировался рельеф пластин с диаметрами 100 и 150 мм и перепадами высот от 5 до 45 мкм. По их ЦМР были рассчитаны модели 22-х морфометрических величин, в частности, серии кривизн: горизонтальной, вертикальной, разностной, двух избыточных, минимальной, максимальной, средней, Гауссовой, и др. Полученные морфометрические модели значительно информативнее, чем исходная ЦМР: они позволяют выявить структуру, конфигурацию и величину неровностей поверхности пластины. Анализ этих сведений дает возможность провести соответствие между ними и проведенными операциями технологического маршрута для их дальнейшей корректировки.

Исследование выполнено с использованием оборудования ЦКП «МСТ и ЭКБ» МИЭТ.

[1] *Дедкова А. А., Дюжев Н. А., Гусев Е. Э., Штерн М. Ю.* Оперативная неразрушающая методика анализа прогиба мембран, расположенных на пластине // Дефектоскопия, 2020. № 5. С. 52–59.

Analysis of topography of silicon wafers by geomorphometric methods

Anna Dedkova[1]⋆ dedkova@ckp-miet.ru
Igor Florinsky[2] iflor@mail.ru
[1]Moscow, National Research University of Electronic Technology
[2]Pushchino, IMPB KIAM RAS

Analysis of topography of silicon wafers and membranes on silicon wafers is an important stage of manufacturing of microelectromechanical systems [1]. In particular, such an analysis allows one to measure deformation (bending) of these objects and to estimate a magnitude of their mechanical stresses. In such works, one can use digital elevation models (DEMs) of wafers and membranes with a submillimeter resolution and a micrometer-range elevation difference, which are produced by optical profilometry.

Analyzing wafer/membrane topography, researchers usually limit themselves to determining deflection. The apparatus of the classic differential geometry is sometimes used analyzing models of the Gaussian and principal curvatures derived from wafer/membrane DEMs. We applied methods of geomorphometry, which is based on a theory of topographic surface essentially developing the classical concepts.

We analyzed topography of wafers with diameters of 100 mm and 150 mm and elevations ranging from 5 μm to 45 μm. We derived models of 22 morphometric variables from wafer DEMs, in particular, a series of curvatures: horizontal, vertical, difference, two excessive, minimal, maximal, mean, Gaussian, etc. The derived morphometric models are much more informative than the original DEM: they allowed us to reveal a structure, configuration, and value of irregularities of the wafer surface. Analysis of these data makes it possible to find a correspondence between them and the operations performed on the technological route for their further correction.

The study was performed with equipment of the R&D Center "MEMSE" (MIET).

[1] *Dedkova A. A., Dyuzhev N. A., Gusev E. E., Shtern M. Yu.* Fast nondestructive technique for analyzing deflection of membranes located on the substrate // Russian Journal of Nondestructive Testing, 2020. Vol. 56. No. 5. Pp. 452–459.

Метод обеспечения отказоустойчивости вычислительных комплексов на основе оценки характеристик надежности

Никулин Владимир Сергеевич[1]⋆ nikulin-94@inbox.ru
Пестунов Андрей Игоревич[1] pestunov@gmail.com
[1]Новосибирск, Новосибирский государственный университет экономики и управления «НИНХ»

Существующие методы обеспечения отказоустойчивости вычислительных комплексов, основанные на регулярных проверках и репликации, не подходят из-за возникающих сложностей при эксплуатации и масштабировании оборудования. В связи с этим, уменьшение отрицательных последствий вызванных отказом оборудования и предоставление точных прогнозов с достаточным временем на восстановление остается сложной исследовательской проблемой. Это обусловливает необходимость наличия эффективного и упреждающего метода обеспечения отказоустойчивости, направленного на минимизацию эффекта отказов в системе.

В процессе эксплуатации вычислительных комплексов, отслеживание текущего состояния осуществляется системой мониторинга. Полученные данные содержат статистическую информацию об отказах оборудования в виде выборки полных наработок.

Известно, что для большинства сложных технических систем (в т.ч. вычислительных комплексов) невозможно описать возникающие в них отказы с помощью конкретного закона распределения случайных величин. В этих случаях для статистической оценки используют непараметрические методы.

Для решения задачи обеспечения отказоустойчивости вычислительных комплексов, предлагается метод на основе точечного оценивания, нахождения плотности распределения времени до отказа. Использование непараметрического метода Парзена-Розенблатта позволяет рассчитать плотность по формуле (1):

$$F(t) = \frac{1}{l\sigma} \sum_{i=1}^{l} \frac{K(t-x)}{\sigma}, \tag{1}$$

где K(x) - ядро функции, σ - параметр локальности.

Проведенный анализ на основе смоделированной функции распределения показал, что использование в качестве ядра функции Гаусса (2) позволяет более точно описать эмпирическую функцию распределения:

$$K(x) = \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{x^2}{2}} \tag{2}$$

Также отмечено, что наблюдаемая величина для которой строится плотность распределения – время, задано на области определения $[0, +\infty)$. Поскольку выбранная в качестве ядра функция Гаусса является непрерывной на $(-\infty, +\infty)$, то в окрестности точки ноль будет присутствовать смещение оценки и не

будет выполняться условие $F(0) = 0$, подразумевающее работоспособность в момент начала эксплуатации. Метод зеркального отображения ядра позволяет компенсировать данное смещение заменой симметричного ядра $\frac{K(t-x)}{\sigma}$ в (1), отраженным нормальным ядром с установленной в ноль нижней границы области определения функции распределения. В этом случае для единичной наработки функция распределения F(t) примет вид (3):

$$F(t) = [\frac{K(t-x)}{\sigma} + \frac{K(t+x)}{\sigma} - 1] \tag{3}$$

На базе полученной плотности осуществляется расчет показателей надежности вычислительных комплексов: функции распределения времени до отказа, вероятности безотказной работы (функции надежности), интенсивности отказов, средней наработки до отказа, нестационарного коэффициента готовности (функции готовности). Полученные показатели надежности позволяют обеспечивать отказоустойчивость вычислительных комплексов на заданном уровне и используются для принятия управляющих решений.

[1] *Никулин В. С.* Методика подготовки данных для интеллектуального анализа надежности вычислительных комплексов // Вестник СибГУТИ, 2020. №3. С. 26–37.

Method of maintaining fault tolerance of computing systems based on the assessment of reliability characteristics

Vladimir Nikulin[1]⋆ nikulin-94@inbox.ru
Andrey Pestunov[1] pestunov@gmail.com
[1]Novosibirsk, Novosibirsk State University of Economics and Management

The existing methods of ensuring the fault tolerance computing systems based on regular checks and replication are not suitable due to the difficulties arising in the operation and scaling of equipment. As such, mitigating the negative impacts caused by equipment failure and providing accurate predictions with sufficient recovery time remains a challenging research challenge. This necessitates an effective and predicted method of ensuring fault tolerance aimed at minimizing the effect of failures in the system.

During the operation of computing systems, the current state is monitored by a monitoring system. The data obtained contains statistical information on equipment failures in the form of a sample of complete operating time.

It is known that for the majority of complex technical systems (including computing systems) it is impossible to describe the failures occurring in them using a specific distribution law of random variables. In these cases, nonparametric methods are used for statistical evaluation.

To solve the problem of ensuring fault tolerance of computing systems, a method is proposed based on point estimation, finding the distribution density of the time to failure. Using the nonparametric Parsen-Rosenblatt method, the density can be calculated using the formula (4):

$$F(t) = \frac{1}{l\sigma} \sum_{i=1}^{l} \frac{K(t-x)}{\sigma}, \tag{4}$$

where K(x) is the kernel function, σ is the locality parameter.

The analysis performed on the basis of the simulated distribution function showed that the use of the Gaussian function (5) as the kernel allows a more accurate description of the empirical distribution function:

$$K(x) = \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{x^2}{2}} \tag{5}$$

It is also noted that the observable value for which the distribution density is constructed - time, is specified on the domain of definition $[0, +\infty)$. Since the Gaussian function chosen as the kernel is continuous on $(-\infty, +\infty)$, then in the vicinity of the point zero there will be an estimate bias and the condition $F(0) = 0$ will not satisfied, which implies operability at the start exploitation. The kernel mirroring method allows you to compensate for this bias by replacing the symmetric kernel

$\frac{K(t-x)}{\sigma}$ in (4), reflected by a normal kernel with the lower boundary of the distribution function definition set to zero. In this case, for a unit operating time, the distribution function F(t) will take the form (6):

$$F(t) = [\frac{K(t-x)}{\sigma} + \frac{K(t+x)}{\sigma} - 1] \tag{6}$$

On the basis of the obtained density, the reliability indicators of computing systems are calculated: the distribution function of the time to failure, the probability of failure-free operation (the reliability function), the failure rate, the mean time to failure, the non-stationary availability factor (the availability function). The obtained reliability indicators make it possible to ensure the fault tolerance of computing systems at a given level and are used to make control decisions.

[1] *Nikulin V.* Methods of data preparation data for intelligent analysis of the computer systems reliability // Vestnik SibGUTI, 2020. No 3. Pp. 26–37.

Мониторинг и прогнозирование в слабо-структурированных ситуациях с использованием временных рядов и когнитивного моделирования

Авдеева Зинаида Константиновна avdeeva@ipu.ru
Гребенюк Елена Алексеевна lngrebenuk12@yandex.ru
Коврига Светлана Вадимовна[1] kovriga@ipu.ru
[1]Москва, Институт проблем управления им. В.А. Трапезникова РАН

В экономической и финансовой областях методы анализа временных рядов являются востребованным прогностическим инструментом работы с большими массивами данных, отражающих закономерности поведения исследуемых процессов. Однако их потенциала недостаточно для прогнозирования развития ситуации в условиях непредсказуемости поведения изучаемого процесса, например, в случаях резкого перехода из одного состояния в другое, обусловленного событием, вызывающим скачкообразное изменение значений процесса или нарушения сезонности в процессах при переходе от стабильного состояния к кризису.

Предлагается подход к мониторингу и прогнозированию нестационарных процессов с использованием временных рядов и когнитивного моделирования ситуации. Такое сочетание направлено на учет информации, отраженной в данных (временных рядах) и информации о возможных вариантах развития ситуации на основе обработки экспертных знаний и гипотез и разнородных информационных источников, посредством анализа и моделирования на когнитивной карте ситуации (ККС) – экспертной модели причинно-следственных влияний между значимыми факторами, обеспечивающей углублённое понимание ситуации. Мы используем когнитивное моделирование для мониторинге информационного пространства с целью направленного поиска экспертно значимых событий (инфоповодов), способных оказать существенное влияние на прогнозируемый процесс.

Подход состоит в следующем.

Объект прогнозирования. Рассматриваются процессы, описываемые нестационарными временными рядами. При формировании прогнозов исследуется не только сам процесс, но и фон, на котором он развивается, иначе внешняя среда, охватывающая связанные с ним другие процессы и значимые системообразующие факторы влияния на развитие процесса.Компоненты процесса могут быть как зависимыми, так и независимыми друг от друга, причем сила связей и их направленность могут изменяться под воздействием факторов внешней среды. Например, на процесс изменения цен могут оказывать влияние другие процессы (изменения тарифов, объемов продаж, запасов, стоимости энергии и пр.). Таким образом, прогнозируемый показатель в момент времени t может зависеть: (1) от своих прошлых значений, (2) от прошлых значений других компонент процесса, (3) от измеряемых значений процессов, которые могут оказывать воздействие на его поведение, (4) от событий внешней среды, которые могут приводить либо

к резким изменениям свойств процесса, либо вызывать постепенные неявные изменения: нарушения взаимосвязей, медленные изменения трендов, увеличение или уменьшение волатильности, (5) от случайных факторов.

Долгосрочный прогноз значений процесса включает следующие шаги:

1) построение набора количественных прогнозных моделей, описывающих динамику изменения прогнозируемого параметра с использованием факторов (1), (2), (3), (5);
2) коррекцию на основе сигналов о наличии и идентификации изменений в прогнозируемом параметре и связанных с ним, полученных в результате мониторинга;
3 коррекцию на основе сигналов, получаемых в результате когнитивного моделирования внешней среды, упреждающих и корректирующих результаты, полученные с использованием количественного анализа данных.

Коррекция долгосрочного прогноза значений компонент процесса обусловлена актуализацией сигналов мониторинга и когнитивного моделирования в режиме текущих наблюдений. Для активации процедур:

1) отбора количественных моделей прогнозируемого параметра и связанных с ним компонент;
2) коррекции этих моделей, обусловленной идентификацией (в результате мониторинга) экспертно-значимых событий (инфоповодов) – потенциальных сигналов влияния на прогнозируемый показатель и/или связанные с ним компоненты,

разработан ряд критериев актуализации сигналов внешней среды по результатам когнитивного моделирования.

ККС, отражающая согласованное целостное представление о ситуации разнопрофильных экспертов и специалистов предметной области, служит онтологией, которая является основой для фильтрации информационного пространства и идентификации параметров информационного поиска при формировании базы данных (в виде временных рядов) и базы данных о событиях. Результаты когнитивного моделирования, как оценки последствий воздействия инфоповода на прогнозируемый процесс, служат обоснованием выбора событийного инфоповода для активации процедуры коррекции прогнозных моделей.

В рамках реализации подхода разработаны алгоритмы:

1) построения набора количественных прогнозных моделей, выбора актуальной модели на текущий период и формирования прогноза;
2) обнаружения «неявных» изменений тенденций, свойств и взаимосвязей между прогнозируемым параметром и связанных с ним компонент в реальном времени, замаскированных неконтролируемыми случайными возмущениями;
3) коррекции полученных прогнозов по результатам обнаруженных изменений и рекомендациям ККС.

Практическая значимость подхода подтверждена на задаче формирования закупочных цен на черный металлолом металлургическими комбинатами на вторичном рынке сырья.

[1] *Z. K. Avdeeva, E. A. Grebenyuk E, S. V. Kovriga.* Forecasting of Key Indicators of the Manufacturing System in Changing External Environment // IFAC-Papers OnLine, 2020. Vol. 53. No 3.

Monitoring and forecasting in ill-structured situations based on time series and cognitive modelling

Zinaida Avdeeva[1]⋆ zinaida.avdeeva@gmail.com
Elena Grebenyuk[1] lngrebenuk12@yandex.ru
Svetlana Kovriga[1] kovriga@ipu.ru
[1]Moscow, Trapeznikov V.A. Institute of Control Sciences of RAS

In the economic and financial fields, time series analysis methods are a popular predictive tool for working with large data sets that reflect the patterns of behaviour of the studied processes in these areas. However, their potential is not sufficient to predict the development of the situation in conditions that violate the predictability of the behaviour of the studied process, for example, in cases of a sharp transition from one state to another due to an event that causes a jump in the values of the process or a violation of seasonality in processes during the transition from a stable state to a crisis.

We propose an approach to long-term forecasting of non-stationary processes using time series and cognitive modelling of a situation. This combination allows to take into account both the information contained in the data (time series), and information about possible scenarios based on the practice expertise and hypotheses and the information from heterogeneous sources through analysis and simulation on the cognitive map of the situation (CMS). We consider the CMS as an expert model of cause-and-effect influences between significant factors, providing an in-depth understanding of the situation. We use the cognitive modelling of a situation to monitor the information space in order to search for expert-relevant events (information causes) that can have a significant impact on the predicted process.

The approach is as follows.

Forecasting object. Processes represented by non-stationary time series are considered. When forming forecasts, we study not only the process itself, but also the background against which it develops, or the external environment that covers other processes associated with it and significant system-forming factors of influence on the development of the process. The components of the process can be both dependent and independent of each other, and the strength of connections and their direction can change under the influence of environmental factors. For example, the processes of changing tariffs, sales volumes, inventory, energy costs, and so on may affect the process of changing prices. Thus, the predicted indicator at time t may depend on: (1) its past values, (2) the past values of other components of the process, (3) the measured values of processes that may affect its behaviour, (4) environmental events that may lead either to abrupt changes in the properties of the process, or cause gradual implicit changes: violations of relationships, gradual changes in trends, an increase or decrease in volatility, (5) random factors.

Long-term forecasting of the values of the process includes the following steps:

1) building a set of quantitative forecast models describing the dynamics of changes in the predicted parameter using factors (1), (2), (3), (5);
2) correction based on signals about the presence and identification of changes in the predicted parameter and related ones obtained as a result of monitoring;
3) correction based on signals obtained as a result of cognitive modelling of the external environment, anticipating and correcting the results obtained using quantitative data analysis.

Correction of the long-term forecast of the values of the process components is due to the updating of monitoring signals and cognitive modelling in the current observation mode.

We have proposed a number of criteria for updating environmental signals based on the results of the cognitive modelling to activate procedures:

1) selection of quantitative models of the predicted parameter and related components;
2) correction of these models due to identification (as a result of monitoring) of expert-significant events (information causes) that are potential signals of influence on the predicted indicator and/or related components.

The built CMS, which reflects a consistent holistic view of the situation of various specialised experts and subject area specialists, serves as an ontology that is the basis for filtering the information space and identifying information search parameters when forming a database (in the form of time series) and a database of events. The rationale for choosing an event-based information to activate the procedure for correcting forecasting models is the results of the cognitive modelling as an assessment of the consequences of the information causes on the predicted process.

As part of the implementation of the approach, we have developed algorithms:

1) building a set of predictive models based on quantitative data, selecting an up-to-date model for the current period and forming a forecast;
2) detecting "implicit" changes in trends, properties, and relationships between the predicted parameter and its associated components in real time, masked by uncontrolled random disturbances;
3 correcting the obtained forecasts based on the results of detected changes and recommendations of the CMS.

The practical significance of the approach is confirmed by the problem of forming purchase prices for ferrous scrap by metallurgical plants in the secondary raw material market.

[1] *Z. K. Avdeeva, E. A. Grebenyuk E, S. V. Kovriga.* Forecasting of Key Indicators of the Manufacturing System in Changing External Environment // IFAC-Papers OnLine, 2020. Vol. 53. No 3.

Применение многомерных моделей гауссовых смесей для анализа заказов службы такси

Андриянов Никита Андреевич[1,2]⋆ nikita-and-nov@mail.ru
Дементьев Виталий Евгеньевич[2] dve@ulntc.ru
Ташлинский Александр Григорьевич[2] tag@ulstu.ru
[1]Москва, Финансовый университет при Правительстве РФ
[2]Ульяновск, Ульяновский государственный технический университет

Работа посвящена алгоритмам кластеризации заказов службы такси по нескольким параметрам. По результатам работы службы заказа такси была подготовлена выборка из 100 заказов и сводка характеристик по ним. В характеристики заказа попали следующие параметры: время заказа (утро, день, вечер, ночь), расстояние между начальной и конечной точками маршрута, погодные условия (жарко, холодно, осадки, комфортные условия), класс обслуживания (эконом, стандарт, бизнес). По таким заказам также добавлена их итоговая стоимость. На основе полученного датасета эксперты службы заказа такси (менеджеры) выполнили бинарную классификацию предложенных заказов, на выходе которой в исходные данные был добавлен еще один параметр, названный «выгода от заказа». Эксперты ставили метки «заказ выгоден службе» и «заказ не выгоден службе». По результатам такой классификации была принята первичная попытка обучиться определять выгодность и невыгодность заказов, что впоследствии поможет найти соотношение выгодных/невыгодных заказов и принимать меры по оптимизации политики ценообразования для невыгодных заказов с учетом отобранных критериев.

Для кластеризации данных различным параметрам были сопоставлены целочисленные значения. Затем были обучены нейронные сети с обратным распространением ошибок. Также были рассмотрены алгоритмы на базе бинарных деревьев решений. Однако для реализации данных алгоритмов требуются дополнительные затраты на обучение. Чтобы избежать данных затрат, был предложен алгоритм на базе моделей гауссовых смесей. Он позволяет кластеризировать данные на основе их статистических характеристик нормального распределения (математического ожидания и дисперсия). При этом характеристики определяются для всех рассматриваемых параметров. Таким образом, формируется многомерные пространства с центрами кластеров по средним значениям параметров для разных классов. Результаты сравнения показали, что алгоритм на базе смесей гауссовых моделей не уступает алгоритмам машинного обучения на базе нейронных сетей и деревьев решений. При этом точность кластеризации по эталонам от экспертов в тестовой выборке составляет порядка 90%. Таким образом, выполнены исследования по интеллектуальному анализу данных работы службы заказа такси на базе смесей гауссовых моделей [1].

Работа поддержана грантом РФФИ № 19-29-09048.

[1] *Andriyanov N. A., Tashlinsky A. G., Dementiev V. E.* Detailed Clustering Based on Gaussian Mixture Models // Advances in Intelligent Systems and Computing, 2021. Vol. 1251. Pp. 437–448.

Application of multi-dimensional models of Gaussian mixtures models for analysis of taxi service

Nikita Andriyanov[1,2]⋆ nikita-and-nov@mail.ru
Vitaly Dementiev[2] dve@ulntc.ru
Alexandr Tashlinskii[2] tag@ulstu.ru
[1]Moscow, Financial University under the Government of the Russian Federation
[2]Ulyanovsk, Ulyanovsk State Technical University

The work is devoted to algorithms for clustering taxi service orders by several parameters. Based on the results of the taxi ordering service, a sample of 100 orders and a summary of their characteristics were prepared. The following parameters were included in the order characteristics such as order time (morning, afternoon, evening, night), distance between the starting and ending points of the route, weather conditions (hot, cold, precipitation, comfortable conditions), service class (economy, standard, business). For such orders, their total cost has also been added. On the basis of the obtained dataset, the experts of the taxi ordering service (taxi service managers) performed a binary classification of the proposed orders, at the output of which one more parameter was added to the initial data. This column in the data table was called 'benefit from the order'. Experts put the tags 'order is beneficial for service' and 'order is not beneficial for service'. Based on the results of this classification, an initial attempt was made to learn how to determine the profitability and disadvantage of orders, which will subsequently provide the estimation of the ratio of profitable / unfavorable orders and take measures to optimize the pricing policy for disadvantageous orders, taking into account the selected criteria.

To cluster the data, integer values were mapped to various parameters. Then backpropagation neural networks were trained. Algorithms based on binary decision trees were also considered. However, the implementation of these algorithms requires additional training costs. To avoid these costs, an algorithm based on Gaussian mixture models was proposed. It allows one's to cluster data based on their statistical characteristics of the normal distribution (mean and variance). In this case, the characteristics are determined for all considered parameters. Thus, multidimensional spaces are formed with cluster centers based on the average values of the parameters for different classes. The comparison results showed that the algorithm based on Gaussian mixture models is not inferior to machine learning algorithms based on neural networks and decision trees. At the same time, the accuracy of clustering according to standards from experts in the test sample is about 90%. Thus, research has been carried out on the data mining of a taxi ordering service based on mixtures of Gaussian models [1].

This research is funded by the Russian Foundation for Basic Research, grant 19-29-09048.

[1] *Andriyanov N. A., Tashlinsky A. G., Dementiev V. E.* Detailed Clustering Based on Gaussian Mixture Models // Advances in Intelligent Systems and Computing, 2021. Vol. 1251. Pp. 437–448.

Идентификация штатной работы оборудования на основе прямых геометрических методов

Некрасов Иван Васильевич[1] ivannekr@mail.ru

[1]Москва, ФГБУН Институт Проблем Управления имени В. А. Трапезникова РАН

Задача классификации состояния оборудования является базовой в системах мониторинга и диагностики. В первом приближении любая система диагностики должна как минимум уметь различать состояния объекта по бинарному признаку «нормальная работа» – «нештатное состояние». Общая идея диагностики оборудования может быть интерпретирована в терминах теории фазового пространства [1] как задача построения областей притяжения точек состояния, где каждая область соответствует определенному режиму функционирования (как штатного, так и нештатного). Построение указанных областей удобно вести на основании реальных измерений, интерпретируя каждый срез измерений как точку пространства состояний $State=(X_1,X_2,X_3, \ldots, X_N)$, принадлежащую конкретной области (пример для трехмерного случая см.рис.1).

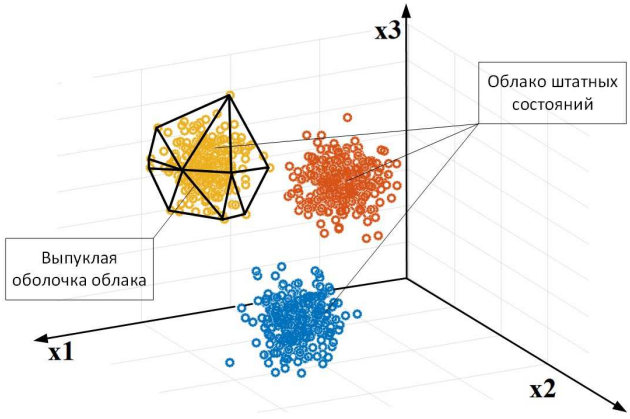


Рис. 1. Состояния оборудования как точки фазового пространства (на примере трехмерного пространства состояний)

Развитие данного подхода очевидным образом задействует методы, заимствованные из разделов математики, связанных с пространственными геометрическими построениями [2]. Основой проводимого исследования является гипотеза принадлежности близких друг другу точек состояния (например, точек работы оборудования на одном штатном режиме) ограниченным множествам [3] в фазовом пространстве. Такой подход позволяет аппроксимировать облака измеренных точек штатной работы оборудования выпуклыми оболочками [2], а задачу классификации текущего состояния на штатное и нештатное интер-

претировать как задачу определения нахождения текущей точки внутри или снаружи оболочки, соответственно.

Выделение замкнутой области фазового пространства на основе облака точек состояния является самостоятельной нетривиальной задачей. В настоящее время существуют эффективные алгоритмы построения выпуклой оболочки на плоскости [4] – например, алгоритмы Джарвиса, Грэхема, Чана – и их модификации и в трехмерном пространстве [5]. Доказанная сложность алгоритмов:

$$C(N = 2) = O[n \cdot log(n)] \tag{1}$$

где n – количество точек на плоскости, N – размерность пространства (в данном случае $N{=}2$).

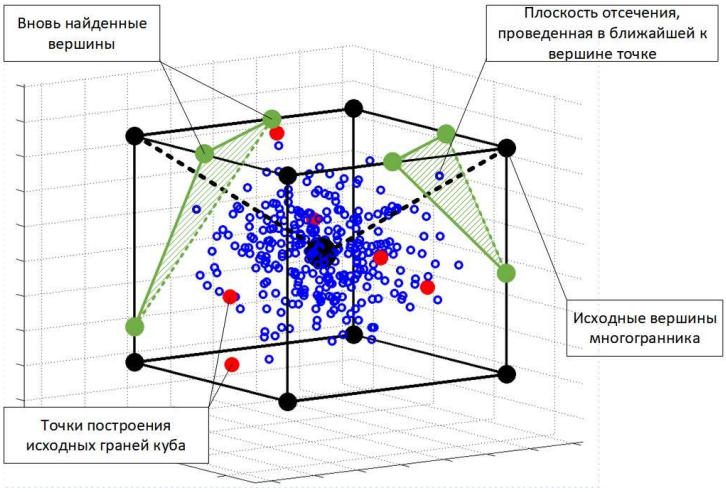


Рис. 2. Последовательная аппроксимации выпуклой оболочки многогранником (на примере 3-параметрического пространства состояний)

Однако размерность фазового пространства состояния сложных технических объектов значительно превосходит $N{=}3$ (может достигать десятков и сотен переменных), что заставляет исследователей обратиться к алгоритмам приближенного построения внешней оболочки известного облака точек. В настоящей работе рассмотрен алгоритм последовательной аппроксимации выпуклой оболочки многогранником, количество граней которого увеличивается на каждой итерации. Работа алгоритма начинается с построения максимального гиперкуба, гарантированно включающего все точки облака (в качестве примера для $N{=}3$ см.рис.2). Затем для каждой вершины осуществляется поиск наименее удаленной точки облака, через которую проводится касательная отсекающая гиперплоскость, являющаяся новой дополнительной гранью (см.рис.2). Далее

работа алгоритма аналогичным образом повторяется для всех вновь полученных вершин многогранника.

На основании полученной выпуклой оболочки можно настроить классификатор, определяющий, находится ли точка состояния, полученная на основе новых измерений, внутри или снаружи данного многогранника. Нахождение за его пределами свидетельствует о «подозрении» на нештатную работу оборудования в текущий момент. Вопросы настройки самого алгоритма классификации, а также организации работы системы диагностики в целом выходят за рамки представленного исследования и будут освещены в других работах автора.

[1] *Шаталов А. С.* Отображение процессов управления в пространстве состояний // Москва: Энергоатомиздат, 1986. С. 256.

[2] *Препарата Ф., Шеймос М.* Вычислительная геометрия: Введение: пер. с англ. // Москва: Мир, 1989. С. 478.

[3] *Корн Г., Корн Т.* Справочник по математике (для научных работников и инженеров): пер. с англ. // Москва: Наука, 1974. С. 831.

[4] *Ивановский С. А., Преображенский А. С., Симончик С. К.* Алгоритмы вычислительной геометрии. Выпуклые оболочки: простые алгоритмы. // Компьютерные инструменты в образовании, 2007. № 1. С. 4–19.

[5] *Ивановский С. А., Преображенский А. С., Симончик С. К.* Алгоритмы вычислительной геометрии. Выпуклые оболочки в трехмерном пространстве. // Компьютерные инструменты в образовании, 2007. Vol. 3. С. 4–17.

Direct Geometric Approach for Asset Normal State Identification

Ivan Nekrasov[1] ivannekr@mail.ru

[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

The problem of identification of asset state belongs to the ground items that form theoretical basis of monitoring and diagnostic systems design. As minimal requirement any diagnostic system must be able to classify the asset state against a Boolean criterion "normal functioning" – "abnormal state". A generalized approach of asset diagnostics can be formalized in terms of state-space theory [1] as a problem of rendering groups of state points where each group corresponds to a definite functioning mode. The rendering process of the named groups can be effectively conducted based on asset real state-points that are composed from the measured values of asset online parameters $State=(X_1,X_2,X_3, \ldots, X_N)$ (an example for three-dimensional case is depicted in the fig.1).

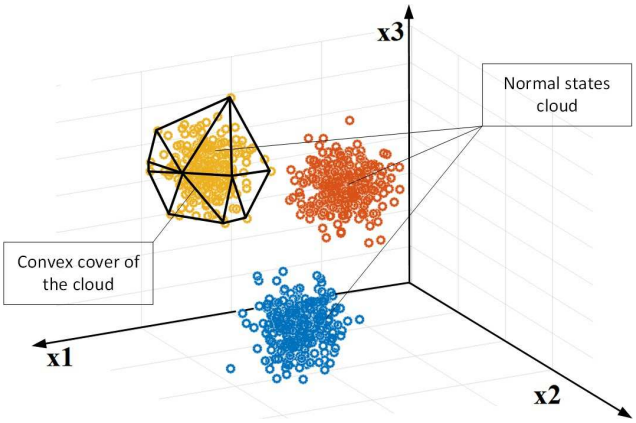


Figure 3. Asset states represented as state-space points (for an instance of three-dimensional space)

Development of the named approach obviously comes to an idea to utilize methods commonly used in geometrical analysis [2]. The conducted research grounds on the hypothesis that the state points located close enough to each other belong to a bounded set [**?**] in the multidimensional space. This assumption gives us ability to approximate the cloud of measured normal asset states with a convex cover [2] and to convert the initial "normal – abnormal state" classification problem into an exercise where we have to define whether the current measured point is located inside or outside of the defined cover.

Creating a convex cover around a point cloud is a nontrivial mathematical problem by itself. At present there exist effective algorithms to build a minimal convex

outline in a two-dimensional plane [4] – for instance "Jarvis march", "Graham scan", "Chan's algorithm" – and their adoptions for three-dimensional case [5]. The provn computational complexity of named algorithms is:

$$C(N = 2) = O[n \cdot log(n)] \tag{2}$$

where n – number of state points in the set on the plane, N – tate-space dimensionality (for this case *N=2*).

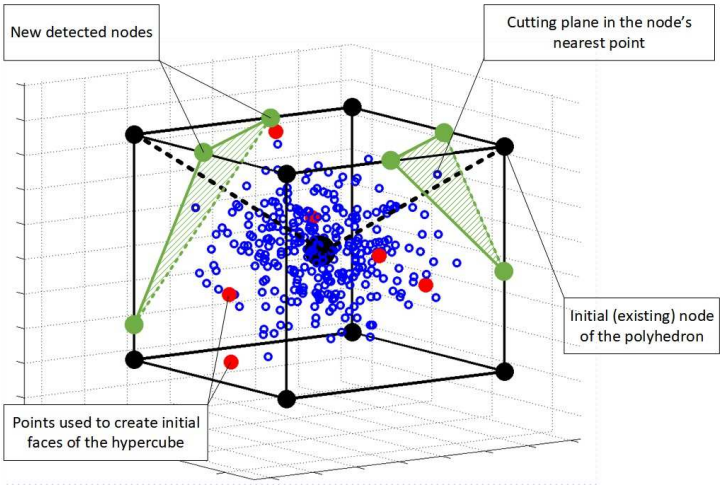


Figure 4. Consequent approximation of the convex cover with a polyhedral (for an instance of three-dimensional space)

However the dimensionality of the state-space for real technical plants is far beyond *N=3* (it can be tens or hundreds of variables). This fact forces researches to refer towards approximate algorithms of convex covers creation. This paper represents an algorithm that conducts consequent approximation of the convex cover with a polyhedral. Each iteration of the algorithm adds new face in the point that is nearest to a known node of the previously generated polyhedral. The algorithm starts with construction of a maximal hypercube that covers all points of the cloud with guarantee (see a three-dimensional example in the fig.2). After that each existing node of the hypercube is scaned for the nearest point that is used to build a new face. Orientation of the new face is defined as a tangent hyperplane in the found nearest point (see fig.2). Further algorithm run is cyclically conducted for all newly generated nodes with the same operation list.

After the target convex cover is generated we can proceed to the next step of the diagnostic system development – we should create a classifier, that defines whether the current measured state point resides inside or outside the polyhedral. Since

the initial cloud was generated based on the normal state measurements finding the state point outside the cover testifies the current state as "suspicious" to be abnormal. The questions of classifying algorithm itself and other aspects of diagnostic system creating are out of scope of this paper and will be / are placed in separate publications of the author.

[1] *Shatalov A. S.* Depicting control processes in state-space // Moskva: Energoatomizdat, 1986. Pp. 256.

[2] *Preparata F., Shamos M.* Computational geometry: an Introduction // New York: Springer-Verlag, 1985. Pp. 478 p.

[3] *Korn G., Korn T.* Mathematical handbook (for scientists and engineers) // New York: McGraw -Hill Book Company, 1968. Pp. 831.

[4] *Ivanovskij S. A., Preobrazhenskij A. S., Simonchik S. K.* Computational geometry algorithms. Convex covers: simple algorithms // Computer tools for education, 2007. No 1. P. 4–19.

[5] *Ivanovskij S. A., Preobrazhenskij A. S., Simonchik S. K.* Computational geometry algorithms. Convex covers in three-dimensional space // Computer tools for education, 2007. No 3. Pp. 4–17.

Использование системы технического зрения в системе прослеживания производства железнодорожных колес

Кульков Ярослав Юрьевич[1] y_mail@mail.ru
Жизняков Аркадий Львович[1] lvovich1975@mail.ru
Привезенцев Денис Геннадьевич[1] dgprivezencev@mail.ru
Запатрин Михаил Георгиевич[1]⋆ misha.zapatrin@yandex.ru
[1]Муром, Муромский институт (филиал) ВлГУ

На металлургических предприятиях, производящих колеса для подвижного железнодорожного транспорта часто используется система прослеживания движения продукции. Одним из элементов этой системы является система технического зрения для распознавания маркировки изделий. Задача распознавания символов на изображении имеет множество готовых алгоритмов решения [1]. При этом, имеется ряд проблем, препятствующих их внедрение в рассматриваемую систему.

Основной проблемой является технологический процесс производства колес, в котором маркируемая поверхность после штамповки полируется до зеркального блеска. Такая поверхность приводит к появлению бликов при получении изображения, которые при попадании в область маркировки существенно осложняют правильную идентификацию. В рамках предлагаемой работы рассматривается построение алгоритма этапа формирования изображения, обнаружения и выделения символов горячей и холодной маркировки, а также последующее их распознавание.

После формирования бинарного изображения часть информации, маскируемой бликами теряется безвозвратно. Поэтому первой задачей является выбор оптимального алгоритма бинаризации. Так как область нанесения маркировки на колесе известна, то предварительно для уменьшения влияния темных участков средней части колеса на работу алгоритма, на само изображение накладывается бинарная маска, удаляющая лишние части. В итоге получаем кольцевое изображение части колеса, в области которой должна находиться маркировка.

В ходе экспериментальных исследований наиболее устойчивые к бликам результаты при бинаризации показал адаптивный метод Кристиана. Так как угол ориентации колеса относительно камеры не задан, то маркировка может иметь случайное расположение в пределах заданной кольцевой области.

По бинарному изображению выполняется обнаружение области маркировки [2], после чего выделяются отдельные символы. Вычисляется их ориентация и производится нормировка их положения к вертикальному.

Выделенные символы поступают на вход нейросети, предварительно обученной на шаблонах символов горячей и холодной маркировки, используемой в процесс производства.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (Госзадание ВлГУ ГБ-1187/20)

[1] *Lee S., Yun J. P., Koo G., Kim S. W.* End-to-end recognition of slab identification numbers using a deep convolutional neural network // Knowledge-Based Systems, 2017. No 132. С. 1–10.

[2] *Болотова Ю. А., Спицын В. Г., Осина П. М.* Обзор алгоритмов детектирования текстовых областей на изображениях и видеозаписях // Компьютерная оптика, 2017. Т. 41. № 3. С. 411–452.

Use of a machine vision system for tracking the production of railway wheels

Yaroslav Kulkov[1] y_mail@mail.ru
Arcady Zhiznyakov[1] lvovich1975@mail.ru
Denis Privezentsev[1] dgprivezencev@mail.ru
Mikhail Zapatrin[1]⋆ misha.zapatrin@yandex.ru

[1]Murom, Murom institute (branch) Vladimir State University

In metallurgical plants that produce wheels for mobile rail vehicles, a product tracking system is often used. One of the elements of this system is a vision system for recognizing product markings. The problem of character recognition in an image has many ready-made algorithms for solving it [1]. At the same time, there are a number of problems that prevent their introduction into the system under consideration.

The main problem is the technological process of wheel production, in which the marked surface after stamping is polished to a mirror finish. Such a surface leads to the appearance of glare during image acquisition, which, when it enters the marking area, significantly complicates correct identification. Within the framework of the proposed work, the construction of an algorithm for the stage of image formation, detection and extraction of hot and cold marking symbols, as well as their subsequent recognition is considered.

After the formation of a binary image, part of the information masked by glare is irretrievably lost. Therefore, the first task is to choose the optimal binarization algorithm. Since the area of marking on the wheel is known, a binary mask is applied to the image itself, which removes unnecessary parts, to reduce the influence of the dark areas of the middle part of the wheel on the operation of the algorithm. As a result, we get an annular image of the part of the wheel, in the area of which the marking should be located.

In the course of experimental studies, the most resistant to glare results with binarization was shown by the adaptive method of Christian. Since the angle of orientation of the wheel relative to the camera is not specified, the marking may be randomly located within the specified annular area.

By the binary image, the marking area is detected [2], after which individual characters are selected. Their orientation is calculated and their position is normalized to vertical.

The selected symbols are fed to the input of a neural network, pre-trained on templates of hot and cold marking symbols used in the production process.

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation (State task of VlSU GB-1187/20).

[1] *Lee S., Yun J. P., Koo G., Kim S. W.* End-to-end recognition of slab identification numbers using a deep convolutional neural network // Knowledge-Based Systems, 2017. No 132. C. 1–10.

[2] *Bolotova Y. A., Spitcin V. G., Osina P. V.* Overview of algorithms for detecting text areas in images and videos // Computer optics, 2017. Vol. 41. No 3. Pp. 411–452.

Вычисление координат точки захвата плоского объекта роботом

Кульков Ярослав Юрьевич[1]⋆ y_mail@mail.ru
Садыков Султан Сидыкович[1] sadykovss@yandex.ru
Орлов Александр Дмитриевич[1] j-awtrope75@mail.ru
Баюров Сергей Вячеславович[1] bayurovsv@yandex.ru

[1]Муром, Муромский институт (филиал) ВлГУ

Нахождение координат точки захвата объекта роботом манипулятором при случайном расположении объекта в рабочей области является одной из основных при разработки автоматизированных сборочных и сортировочных комплексов [1]. Часто, из-за сложной геометрии объекта, захват в центральной точке не представляется возможным. В таких случаях оператору приходится вручную указывать координаты наиболее подходящей для захвата точки, что делает невозможным автоматизацию процесса. Для решения этой проблемы необходим алгоритм, позволяющий определять заданную для каждого типа объекта точку захвата вне зависимости от его расположения и ориентации на рабочем поле робота.

Предлагаемый подход позволяет задавать относительные координаты для каждого объекта. В ходе работы после этапа распознавания класса объекта [2] система выполняет все шаги алгоритма и выдает координаты в глобальной системе координат робота манипулятора.

Исходным является изображение объекта, полученное с камеры системы технического зрения (СТЗ), входящей в состав роботизированной системы. На данном шаге считаем, что точка начала отсчета двумерной системы координат СТЗ совпадает с таковой системы координат манипулятора $E = \{O, X, Y\}$ на рабочем поле. На этапе настройки роботизированной системы оператор задает координаты точки захвата $P_E(x, y)$.

На первом шаге выполняется бинаризация полученного изображения объекта. Используя моменты второго порядка вычисляется угол α ориентации оси объекта относительно оси X глобальной системы координат E.

Далее выполняется построение описанного прямоугольника минимальной площади для найденного объекта используя метод вращающихся калиперов, представляя объект как множество точек $P_E = \{p_1, p_n, \ldots, p_n\}$. Угловую точку полученного прямоугольника, ближайшую к левой части объекта, будем считать точкой $(0, 0)$ начала локальной системы координат $\hat{E} = \{\hat{O}, X, Y\}$ самого объекта.

Полученная система координат $\hat{E}$ поворачивается на угол α, в результате чего положение оси X будет совпадать с осью ориентации объекта в системе E.

Вычисляется преобразование координат точки P_E в координаты локальной системы $P_{\hat{E}}$. Данные координаты сохраняются в базе системы.

Для нахождения координат точки захвата в рабочей области манипулятора (глобальные координаты точки захвата), на случайно расположенном объекте в процессе работы системы, выполняются те же шаги: строится минимальный описанный прямоугольник; вычисляется угол ориентации оси объекта; строится локальная система координат объекта. В полученной системе координат $\hat{E}$ положение оси X также будет совпадать с вычисленной осью ориентации объекта. Используя хранимые в базе координаты локальной точки, путем вращения системы координат $\hat{E}$ на угол $-\alpha$, приводятся в соответствие с глобальной системой координат E робота. Данные координаты передаются в манипулятор, который используя схват или пневматическую присоску захватывает объект в заданной точке для перемещения.

Экспериментальные и вычислительные исследования показали, что предложенный алгоритм позволяет определять заданную точку захвата с погрешностью не более 1-2 пикселей, что связано ошибкой округления при вращении системы координат в дискретном пространстве изображения. Физическая линейная погрешность зависит от разрешения используемой камеры СТЗ и расстояния от нее до самого объекта.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (Госзадание ВлГУ ГБ-1187/20)

[1] *Niu L., Saarinen M., Tuokko R., Matti J.* Integration of Multi-Camera Vision System for Automatic Robotic Assembly // Procedia Manufacturing, 2019. No 37. Pp. 380-384.

[2] *Садыков С. С., Кульков Я. Ю.* Идентификация отдельных тестовых и реальных плоских объектов по безразмерным признакам контуров их двухградационных изображений // Вестник Рязанского государственного радиотехнического университета, 2016. № 56. С. 136–143.

Calculation of the flat objects gripping point coordinates by a robot

Yaroslav Kulkov[1]⋆ y_mail@mail.ru
Sultan Sadykov[1] sadykovss@yandex.ru
Alexandr Orlov[1] j-awtrope75@mail.ru
Sergey Bayurov[1] bayurovsv@yandex.ru

[1]Murom, Murom institute (branch) Vladimir State University

Finding the coordinates of the gripping point of an object by a robotic manipulator with a random object location at the robot workspace, is one of the main ones in the development of automated assembly and sorting complexes [1]. Often, due to the complex geometry of the object, it is not possible to capture at the center point. In such cases, the operator has to manually specify the coordinates of the most suitable point to capture, which makes it impossible to automate the process. To solve this problem, an algorithm is needed that allows to determine the gripping point specified for each type of object, regardless of its location and orientation on the robot's working field.

The proposed approach allows to set relative coordinates for each object. During the work, after the stage of object class recognition [2], the system performs all the steps of the algorithm and set outputs coordinates in the global coordinate system of the manipulator robot.

The initial image of the object is obtained from the camera of the technical vision system (STS), which is part of the robotic system. At this step, we assume that the origin point of the two-dimensional coordinate system of the STS coincides with that of the manipulator coordinate system $E = O, X, Y$ on the working field. At the stage of setting up the robotic system, the operator sets the coordinates of the capture point $P_E(x, y)$.

At the first step, the resulting image of the object is binarized. Using the moments of the second order, the angle α of orientation of the object axis relative to the X axis of the global coordinate system E is calculated.

Next, the described rectangle of the minimum area for the found object is constructed using the method of rotating calipers, representing the object as a set of points $P_E = p_1, p_n, \ldots, p_n$. The corner point of the resulting rectangle, which is closest to the left side of the object, will be considered the point $(0, 0)$ of the origin of the local coordinate system $\hat{E} = \hat{O}, X, Y$ of the object itself.

The resulting coordinate system $\hat{E}$ is rotated through an angle α, as a result of which the position of the X axis will coincide with the object's orientation axis in the E.

The transformation of the coordinates of the point P_E into the coordinates of the local system $P_{\hat{E}}$ is calculated. These coordinates are saved in the system base.

To find the coordinates of the grip point in the working area of the manipulator (global coordinates of the grip point), on a randomly located object in the process

of the system operation, the same steps are performed: the minimum bounded rectangle is constructed; the angle of orientation of the object axis is calculated; the local coordinate system of the object is built. In the resulting coordinate system $\hat{E}$, the position of the X axis will also coincide with the calculated orientation axis of the object. Using the coordinates of the local point stored in the base, by rotating the coordinate system $\hat{E}$ by the angle $-\alpha$, they are brought into line with the global coordinate system E of the robot. These coordinates are transferred to the manipulator, which, using a gripper or pneumatic suction cup, grabs the object at a given point for movement.

Experimental and computational studies have shown that the proposed algorithm makes it possible to determine a given capture point with an error of no more than 1-2 pixels, which is associated with a rounding error when the coordinate system is rotated in the discrete image space. The physical linear error depends on the resolution of the used STS camera and the distance from it to the object itself.

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation (State task of VlSU GB-1187/20).

[1] *Niu L., Saarinen M., Tuokko R., Matti J.* Integration of Multi-Camera Vision System for Automatic Robotic Assembly // Procedia Manufacturing, 2019. No 37. Pp. 380-384.

[2] *Sadykov S. S., Kulkov Y. Yu.* Identification of individual test and real flat objects by dimensionless features of the contours of their two-gradation images // Vestnik of Ryazan state radioengineering university, 2016. No 56. Pp. 136–143.

Разработка алгоритма позиционирования объекта по данным с активной сенсорной сети Bluetooth Low Energy маяков

Астафьев Александр Владимирович[1]⋆ Alexandr.Astafiev@mail.ru
Демидов Антон Александрович[1] AADemidov@list.ru
Кондрушин Илья Евгеньевич[1] IEkondrushin@list.ru
Макаров Михаил Вячеславович[1] nauka-murom@mail.ru

[1]Муром, Владимирский государственный университет

Навигация внутри помещений стала очень актуальной научно-технической отраслью. Несмотря на достаточно высокую точность навигации по сетям ГЛОНАСС и GPS вопрос навигации внутри помещений все еще остается нерешенным. Исходя из этого, разработка новых алгоритмов позиционирования объекта внутри помещений является актуальной научно-технической задачей. На рынке развиваются продукты для навигации внутри помещений: Google Maps, Open StreetMap и т.д. Несмотря на это, алгоритмическая поддержка навигации внутри помещений до сих пор остается в основном нетронутой. Целью исследования является разработка алгоритма позиционирования объекта внутри помещений на основе беспроводной сенсорной сети с низким энергопотреблением. Работу разработанного алгоритма можно представить следующей последовательностью этапов:

1. Получение многомерного временного ряда уровней сигналов RSSI.
2. Упорядочивание многомерного временного ряда по уникальным UUID маяков.
3. Редактирование выбросов и экстремальных значений.
4. Фильтрация полученных данных упрощенным фильтром Калмана.
5. Аппроксимация полученных временных рядов для получения расстояния.
6. Перевод уровня сигнала RSSI в расстояние с использование искусственной нейронной сети прямого распространения.
7. Комплексирование полученных на этапе 5 и 6 временных рядов.
8. Позиционирование объекта по методу мультилатерации.
9. Перевод относительных координат в абсолютные.
10. Получение относительных координат положения объекта на контролируемой территории.

Схема работы алгоритма позиционирования представлена на рисунке 1.

Экспериментальные исследования разработанного алгоритма показали, что средняя ошибка позиционирования составила 0,23 метра. Ошибка позиционирования не превышает 0.7 м в закрытом помещении размера 5х5,5 м. Исходя из этого точность позиционирования мобильного устройства с использованием предлагаемого метода в проведенном эксперименте выше на 40,9% по сравнению с аналогами.

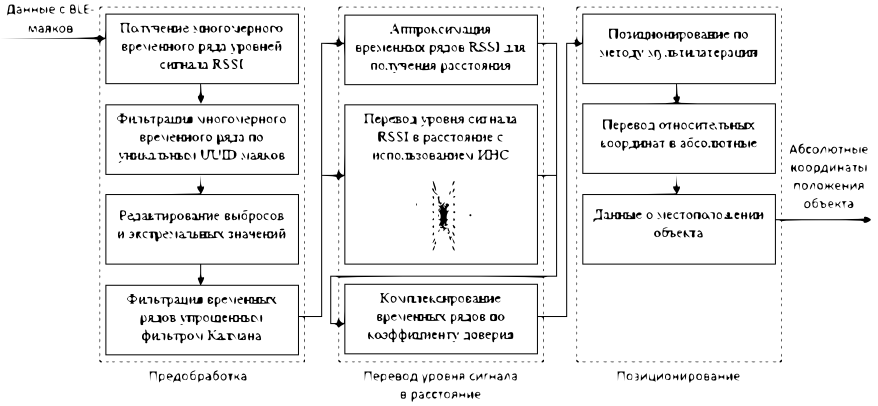


Рис. 1. Схема работы алгоритма позиционирования

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (Госзадание ВлГУ ГБ-1187/20).

[1] *Astafiev A. V.* Development of a Methodology for Positioning Small Scale Mechanization Tools at Industrial Enterprises for the Construction of Unmanned Product Movement Control Systems // 22th International Conference on Digital Signal Processing and its Applications (DSPA), 2020. Pp. 1–6.

Development of algorithm for positioning an object according to data from an active sensor network of Bluetooth Low Energy beacons

Aleksander Astafiev[1]⋆ Alexandr.Astafiev@mail.ru
Anton Demidov[1] AADemidov@list.ru
Ilya Kondrushin [1] IEkondrushin@list.ru
Michail Makarov[1] nauka-murom@mail.ru

[1]Murom, Vladimir State University

Indoor navigation has become a very hot science and technology industry. Despite the sufficiently high accuracy of navigation on the GLONASS and GPS networks, the issue of indoor navigation is still unresolved. Based on this, the development of new algorithms for positioning an object indoors is an urgent scientific and technical task. Indoor navigation products are developing in the market: Google Maps, Open StreetMap, etc. Regardless, algorithmic support for indoor navigation is still largely intact. The aim of the study is to develop an indoor object positioning algorithm based on a wireless sensor network with low power consumption. The work of the developed algorithm can be represented by the following sequence of stages:

1. Obtaining a multidimensional time series of levels of RSSI signals.
2. Ordering a multidimensional time series by unique UUIDs of beacons.
3. Editing outliers and extreme values.
4. Filtration of the received data with a simplified Kalman filter.
5. Approximation of the obtained time series to obtain the distance.
6. Converting the RSSI signal level into distance using a feed forward artificial neural network.
7. Integration of the time series obtained at stage 5 and 6.
8. Positioning of the object using the multilateration method.
9. Conversion of relative coordinates to absolute.
10. Conversion of relative coordinates to absolute.
11. Obtaining the relative coordinates of the position of the object in the controlled area.

The positioning algorithm is shown in Figure 1.

Experimental studies of the developed algorithm showed that the average positioning error was 0.23 meters. The positioning error does not exceed 0.7 m in a closed room with a size of 5x5.5 m. Based on this, the positioning accuracy of a mobile device using the proposed method in the experiment is 40.9% higher than that of analogs.

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation (State assignment of VlGU GB-1187/20)

[1] *Astafiev A. V.* Development of a Methodology for Positioning Small Scale Mechanization Tools at Industrial Enterprises for the Construction of Unmanned Product Move-

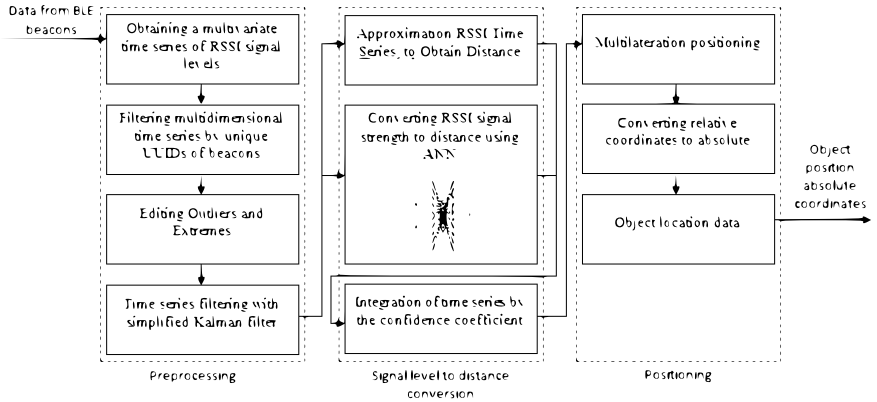


Figure 2. Positioning algorithm operation diagram

ment Control Systems // 22th International Conference on Digital Signal Processing and its Applications (DSPA), 2020. Pp. 1–6.

Об одном подходе к статистическому моделированию транспортных потоков на МКАД и управлению въездами

Старожилец Всеволод Михайлович[1]⋆ starvsevol@gmail.com
Чехович Юрий Викторович[1] chehovich@forecsys.ru
[1]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

На сегодняшний день моделирование крупных транспортных сетей представлено в работах [1, 2] в виде примеров применения существующих программных пакетов таких как SUMO (Simulation of Urban Mobility), iTETRIS ("An Integrated Wireless and Traffic Platform for Real-Time Road Traffic Management Solutions") и других. Детальное описание подхода к моделированию автомагистрали в данных пакетах зачастую отсутствует.

Моделирование транспортных потоков на автомагистрали тесно связано с задачей оптимизации светофорного управления в транспортной сети. Например в работе [3] используется обучение с подкреплением для получение оптимальной схемы управления перекрестком. В большинстве работ, посвящённых светофорному управлению дороги на перекрестке, все дороги считаются равнозначными и не ставиться задача обеспечения максимальной пропускной способности выделенной автомагистрали, как это происходит в данной статье.

Классический подход к моделированию транспортных потоков представлен микроскопическим [4] подходом, где моделируется движения каждого автомобиля в отдельности и макроскопическим [5], в котором движение АТС уподобляется жидкости или газу и моделирование сводится к решению систем нелинейных уравнений. Мы предлагаем подход основанный на движении групп АТС, что приводит к существенным упрощениям по сравнению с макроскопическим подходам. Скорость же групп транспортных средств предлагается рассчитывать с помощью фундаментальной диаграммы поток-плотность на магистрали [6]. Такой подход позволяет быстро обсчитывать достаточно большие транспортные сети, в том числе такую магистраль как МКАД, что необходимо для решения оптимизационных задач для которых проводится моделирование. Модель с использованием данного подхода разработана и представлена в статье [7].

В данной работе рассматривается задача моделирования крупной транспортной автомагистрали — Московской кольцевой автомобильной дороги (МКАД). Целью моделирования является проверка гипотезы о возможности посредством управления потоком въезжающих на магистраль автомобильно-транспортных средств (АТС) уменьшить суммарные потери времени на МКАД и увеличить пропускную способность автомагистрали.

В работе строится модель одной из сторон МКАД с крупными въездами и съездами с неё. На основе имеющихся данных с дорожных датчиков на некоторых из въездов и статистических данных ЦОДД генерируются модельные данные на въездах на автомагистраль двух типов — с утренней пиковой загрузкой и с вечерней, что соответствует большому потоку автомобилей в Москву

и из Москвы. Проводится моделирование поведения автомагистрали с различным модельными данными на въездах и сравниваются результаты с контролем на въездах и без него. Для проверки гипотезы об эффективности управления въездами рассчитываются временные потери на проезд по МКАД за день, а также число автомобилей, проехавших по автомагистрали.

Работа поддержана грантом РФФИ № 20-07-01057 А.

[1] *Yuta A., Nobuyasu I., Hajime I., Tetsuo I., Uchitane T.* Traffic simulation of Kobe-city // Proceedings of the international conference on social modeling and simulation, plus Econophysics Colloquium, 2015. Vol. 229. Pp. 255–264.

[2] *Bieker L., Krajzewicz D., Morra A., Michelacci C., Cartolano F.* Traffic simulation for all: a real world traffic scenario from the city of Bologna // Modeling Mobility with Open Data, 2015. Vol. 229. Pp. 47–60.

[3] *El-Tantawy S., Abdulhai B., Abdelgawad H.* Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto // IEEE Transactions on Intelligent Transportation Systems, 2013. Vol. 14. No 3. Pp. 1140–1150.

[4] *Гасников А. В., и др.* Введение в математическое моделирование транспортных потоков // М.: Litres, 2015. С. 89.

[5] *Whitham J. B.* Linear and nonlinear waves // Wiley, 1974. Pp. 656.

[6] *Алексеенко А. Е., Холодов Я. А., Холодов А. С., Горева А. И., Васильев М. О., Чехович Ю. В., Мишин В. Д., Старожилец В. М.* Разработка, калибровка и верификация модели движения трафика в городских условиях // Компьютерные исследования и моделирование, 2015. Т. 7. № 6. С. 1185–1203.

[7] *Староэюилец В.М, Чехович Ю.В.* Об одном подходе к статистическому моделированию транспортных потоков // Журнал Вычислительной математики и математической физики, 2021. Т. 61. № 5.

About one approach to traffic flows statistical modeling on Moscow Ring Road and enters control

Vsevolod Starozhilets[1]⋆ starvsevol@gmail.com
Yury Chekhovich[1] chehovich@forecsys.ru
[1]Moscow, Dorodnicyn Computing Centre FRC CSC RAS

At this moment, modeling of large transport networks presented in the works [1, 2] in form of examples of using existing software packages such as SUMO (Simulation of Urban Mobility), iTETRIS ("An Integrated Wireless and Traffic Platform for Real-Time Road Traffic Management Solutions") and others. A detailed description of the highway modeling approach in these packages is often missing.

Traffic flow modeling on a highway is closely related to the task of optimizing traffic light control in a transport network. For example, in the article [3] reinforcement learning is used to obtain an optimal crossroad control scheme. In most works devoted to traffic light control at crossroads all roads are considered equivalent and do not set the task of ensuring the maximum throughput of a dedicated highway, as it is in this article.

Classic approach to traffic modelling is represented by macroscopic [4], where movement of each vehicle modelled, and microscopic [5], where we suppose that traffic is like gas or liquid which lead us to solving system of nonlinear equations. We propose an approach based on modelling the movement of groups of vehicles, including dozens of cars, instead of modeling the movement of each individual car as it is done in microscopic model. Car groups speed is calculated using flow-density fundamental diagram on a highway [6]. This approach makes possible to quickly calculate sufficiently large transport networks such as Moscow Ring Road, which is necessary to solve optimization problems for which modeling is carried out.

This paper is about a problem of modeling a large transport highway — Moscow Ring Road. Main purpose of modelling is to test that it is possible to reduce total time loss for driving along Moscow Ring Road and increase the throughput of highway by controlling traffic flow on its enters.

In this work we create a model of one of sides of the Moscow Ring Road with major entrances and exits from it. For modelling we create model data for Moscow Ring Road entrances based on the available information from road traffic detectors at some of entrances and statistical data from Moscow Traffic management center. There are two types of modelling data — with morning peak load and with evening load, which corresponds to a large flow of cars to and from Moscow. Modeling the behavior of the highway with different model data at the entrances is carried out and the results are compared with and without control at the entrances. To test the hypothesis about the efficiency of entry traffic flow control, time losses for travel along the Moscow Ring Road per day are calculated, as well as the number of cars that have driven along the highway.

This research is funded by RFBR, grant 20-07-01057 A.

[1] *Yuta A., Nobuyasu I., Hajime I., Tetsuo I., Uchitane T.* Traffic simulation of Kobe-city // Proceedings of the international conference on social modeling and simulation, plus Econophysics Colloquium, 2015. Vol. 229. Pp. 255–264.

[2] *Bieker L., Krajzewicz D., Morra A., Michelacci C., Cartolano F.* Traffic simulation for all: a real world traffic scenario from the city of Bologna // Modeling Mobility with Open Data, 2015. Vol. 229. Pp. 47–60.

[3] *El-Tantawy S., Abdulhai B., Abdelgawad H.* Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto // IEEE Transactions on Intelligent Transportation Systems, 2013. Vol. 14. No 3. Pp. 1140–1150.

[4] *Gasnikov A. V. and other* Vvedenie v matematicheskoe modelirovanie transportnykh potokov // M.: Litres, 2015. Pp. 89.

[5] *Whitham J. B.* Linear and nonlinear waves // Wiley, 1974. Pp. 656.

[6] *Alekseenko A. E., Kholodov Ya. A., Kholodov A. S., Goreva A. I., Vasil'ev M. O., Chekhovich Yu. V., Mishin V. D., Starozhilecz V. M.* Razrabotka, kalibrovka i verifikacziya modeli dvizheniya trafika v gorodskikh usloviyakh // Computer Research and Modeling, 2015. T. 7. No 6. Pp. 1185–1203.

[7] *Starozhilets V.M, Chekhovich Y.V.* About one approach to traffic flows statistical modeling // Computational Mathematics and Mathematical Physics, 2021. Vol. 61. No 5.

Разработка метода ранней и дифференциальной диагностики болезни Паркинсона и эссенциального тремора с помощью анализа всплескообразной активности мышц

Сушкова Ольга Сергеевна[1]⋆ o.sushkova@mail.ru
Морозов Алексей Александрович[1] morozov@cplire.ru
Габова Александра Васильевна[2] agabova@yandex.ru
Карабанов Алексей Вячеславович[3] doctor.karabanov@mail.ru

[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, ИВНД и НФ РАН
[3]Москва, ФГБНУ «Научный центр неврологии»

Исследованы проблемы разработки метода ранней и дифференциальной диагностики болезни Паркинсона (БП) и эссенциального тремора (ЭТ) с помощью анализа всплескообразной активности мышц. При исследовании пациентов с БП сложно найти пациентов, которые ранее не принимали противопаркинсонические препараты, то есть, являются нелеченными. Для решения проблемы малых выборок было предложено искусственно увеличить выборки испытуемых при помощи бутстрепа. Для увеличения выборок были созданы различные модели испытуемых. При построении моделей пациентов возникает две проблемы: проблема неадекватности и проблема недостаточной обобщающей способности модели. Поэтому при построении моделей пациентов необходимо найти компромисс между точностью и обобщающей способностью моделей. Предложены модели, которые, по мнению авторов, являются перспективными для ранней и дифференциальной диагностики БП и ЭТ. Была получена точность распознавания БП на ранней стадии около 100%, точность дифференциальной диагностики (различения БП и ЭТ) составила более 90%.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта No. 18-37-20021.

[1] *Сушкова О. С., Морозов А. А., Габова А. В., Карабанов А. В.* Разработка метода ранней и дифференциальной диагностики болезни Паркинсона и эссенциального тремора с помощью анализа всплескообразной активности мышц // Сборник статей ИТНТ–2020, 2020. Т. 4. С. 170–178.

Development of a method for early and differential diagnosis of Parkinson's disease and essential tremor based on analysis of wave train electrical activity of muscles

Olga Sushkova[1]⋆ o.sushkova@mail.ru
Alexei Morozov[1] morozov@cplire.ru
Alexandra Gabova[2] agabova@yandex.ru
Alexei Karabanov[3] doctor.karabanov@mail.ru

[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, IHNA&NPh RAS
[3]Moscow, FSBI "Research Center of Neurology"

The problems associated with the development of a method for early and differential diagnosis of Parkinson's disease (PD) and essential tremor (ET) using the analysis of wave train muscle activity are investigated. In the study of patients with PD, it is very difficult to find untreated patients who have not previously taken antiparkinsonian drugs. To solve the problem of small samples, it was proposed to increase artificially the sample of subjects using the bootstrap method. Three artificially enlarged samples of subjects were created: a sample of patients with PD, a sample of patients with ET, and a sample of control subjects. When constructing models, two problems arise: the problem of inadequacy and the problem of insufficient generalizing ability of the model. Both of these issues must be considered when building models. The paper proposes models that, according to the authors, are promising for the early and differential diagnosis of PD and ET. Using these models, a good accuracy of PD recognition at the early stage was obtained (about 100%); the accuracy of differential diagnostics (distinguishing between PD and ET) was more than 90%.

The reported study was funded by RFBR according to the research project No. 18-37-20021.

[1] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V.* Razrabotka metodov ranney i differentsial'noy diagnostiki Parkinsona i essentsial'nogo tremora s pomoshch'yu analiza vspleskoobraznoy aktivnosti myshts // Sbornik statey ITNT–2020, 2020. V. 4. Pp. 170–178.

Расширение прикладной интеллектуальной системы диагностики качества жизни пациентов с неврологической патологией с учетом психологической безопасности

Янковская Анна Ефимовна[1] ayyankov@gmail.com
Обуховская Виктория Борисовна[1,2]⋆ diada1991@gmail.com
[1]Россия, Томск, НИ ТГУ
[2]Россия, Томск, СибГМУ

В настоящее время неврологическая патология является одной из наиболее социально-значимых проблем человечества [1]. В работе с пациентами рассматриваемой категории приоритетным направлением служит как своевременное лечение, так и выявление пациентов с начальными формами снижения качества жизни. Качество жизни является фактором, определяющим ощущение психологической безопасности, которое позволяет сохранять трудоспособность на протяжении длительного периода [2]. Актуальность своевременной диагностики качества жизни и психологической безопасности пациентов с неврологической патологией не вызывает сомнений. Необходимость расширения прикладной интеллектуальной системы (ИС) диагностики качества жизни пациентов с неврологической патологией (ДИАКАЖ) путем объединения с прикладной ИС диагностики психологической безопасности (ДИПСИБ) очевидна. Ниже излагаются проведённые нами исследования по структуризации проблемной области и построению базы данных и знаний, расширению прикладной ИС ДИАКАЖ путем объединения с прикладной ИС ДИПСИБ.

На основе проведённого нами анализа пациентов с различной неврологической патологией (болезнь Паркинсона, рассеянный склероз, остеохондроз позвоночника, последствия инсульта, головокружения и нарушения устойчивости) были выявлены параметры (признаки), определяющие качество жизни (физический компонент здоровья (4 составляющих), психический компонент здоровья (4 составляющих) и психологическую безопасность (психологическое благополучие (6 составляющих), базисные убеждения (3 составляющих), жизнестойкость (3 составляющих)) пациентов с неврологической патологией. Осуществлена структуризация данных и знаний на основе матрицы описаний $\mathbf{Q}$ объектов (пациентов) в пространстве характеристических признаков и различений $\mathbf{R}$, задающих различные механизмы (заболевания) разбиения объектов на классы эквивалентности [3]. Строки матрицы $\mathbf{R}$ сопоставлены строкам матрицы $\mathbf{Q}$, а столбцы – классификационным признакам (диагностическим решениям). Сформированы и заполнены матрицы описаний и различений изучаемых заболеваний, служащие основой для построения базы данных и знаний прикладной ИС ДИАКАЖ, расширенной путем объединения с прикладной ИС ДИПСИБ [4].

Расширение прикладной ИС ДИАКАЖ, предназначенной для своевременной и эффективной диагностики пациентов с неврологической патологией, путем объединения с ИС ДИПСИБ осуществляется на основе кратких опросников.

Математический аппарат прикладных ИС базируется на конвергенции нескольких наук и научных направлений [5]; основан на матричном способе представления данных и знаний (матрицы описаний и различений); оригинальных тестовых методах распознавания образов; выявлении различного рода закономерностей, включая отрицательные образы; альтернативные, зависимые и сигнальные признаки; отказоустойчивые безызбыточные и смешанные диагностические тесты и их весовые коэффициенты; принятии решения и их обоснования с использованием графических, включая когнитивные, средств [3, 5]. Расширение ИС ДИАКАЖ путем объединения с прикладной ИС ДИПСИБ будет осуществлено на базе интеллектуального инструментального средства (ИИС) ИМСЛОГ [6], предназначенного для выявления различного рода закономерностей, включая отказоустойчивые диагностические тесты и их весовые коэффициенты; принятия решения и их обоснования с использованием средств когнитивной графики. База данных и знаний будет создана на основе результатов исследования пациентов с неврологической патологией.

Впервые предлагается расширение прикладной интеллектуальной системы диагностики качества жизни пациентов с неврологической патологией ИС ДИАКАЖ путем объединения с прикладной ИС ДИПСИБ. Впервые проведён анализ, структуризация данных и знаний пациентов с различной неврологической патологией. Поскольку сконструированные на базе ИИС ИМСЛОГ более 30 прикладных ИС показали высокую эффективность, есть все основания, что применение расширенной прикладной ИС ДИАКАЖ позволит повысить качество диагностики.

Работа поддержана грантом РФФИ № 18-013-00937.

[1] *Neurological disorders: public health challenges.* WHO Library Cataloguing-in-Publication Data: World Health Organization, 2006. Pp. 232.

[2] *Авров М. В.* Качество жизни пациентов с хронической ишемией головного мозга // Журнал неврологии и психиатрии им. С.С. Корсакова,, 2017. Т. 4. С. 56–58.

[3] *Янковская А. Е.* Логические тесты и средства когнитивной графики // Издательский Дом: LAP LAMBERT Academic Publishing, 2011. С. 92.

[4] *Yankovskaya A. E., Obukhovskaya V. B.* Basics of creating an applied intelligent system for diagnosing the psychological safety of patients with neurological pathology // Materials of the International Conference “Scientific research of the SCO countries: synergy and integration” — Part 2: Participants’ reports in English, 2019. Pp. 184–190.

[5] *Янковская А. Е.* Анализ данных и знаний на основе конвергенции нескольких наук и научных направлений // Интеллектуализация обработки информации: 8-я международная конференция, 2010. С. 196–199.

[6] *Yankovskaya A. E.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis, 2003. Vol. 13,No 3. Pp. 650–657.

An expansion of applied intelligent system for diagnosing the quality of life of patients with neurological pathology with considering psychological safety

Anna Yankovskaya[1] ayyankov@gmail.com
Victoria Obukhovskaya[1,2]⋆ diada1991@gmail.com
[1]Russia, Tomsk, NR TSU
[2]Russia, Tomsk, SSMU

Currently, neurological pathology is one of the most socially significant problems of humanity [1]. In working with patients of this category, the priority is both timely treatment and the identification of patients with initial forms of decline in the quality of life. The quality of life is a factor that determines the feeling of psychological safety, which allows you to maintain ability to work for a long period [2]. The relevance of timely diagnosis of the quality of life and psychological safety of patients with neurological pathology is beyond doubt. The need to expand the applied intelligent system (IS) for diagnosing the quality of life of patients with neurological pathology (DIAQOL) by combining it with the applied IS for the diagnosis of psychological safety (DIPSYS) is obvious. Below, we briefly outline our studies on structuring the problem area and construction a data and knowledge base, the expansion of the applied IS DIAQOL by combining it with the applied IS DIPSYS.

Based on our analysis of patients with various neurological pathologies (Parkinson's disease, multiple sclerosis, spinal osteochondrosis, the effects of stroke, dizziness and impaired stability), we revealed the parameters (features) that determine the quality of life (the Physical Health (4 components), the Mental Health (4 components) and psychological safety (psychological well-being (6 components), basic beliefs (3 components), hardiness (3 components)) of patients with neurological disorders. The data and knowledge were structured using the matrix of description **Q** of objects (patients) in the space of features and the distinguishing matrix **R**, defining various mechanisms of partition of objects into equivalent classes [3]. The rows of the matrix **R** correspond to the rows of the matrix **Q**, and the columns are compared classification features (diagnostic decision). We have formed and filled in matrices of descriptions and distinctions of the studied diseases, serving as the basis for construction of a database and knowledge of the applied IS DIAQOL, expanded by combining with the applied IS DIPSYS [4].

Expansion of the applied IS DIAQOL, intended for the timely and effective diagnosis of patients with neurological pathology, by combining it with IS DIPSYS, is carried out on the basis of short questionnaires. The mathematical apparatus of the applied IS is based on the convergence of several sciences and scientific directions [5]; based on the matrix method of data and knowledge representation (matrices of descriptions and distinctions); original test methods for pattern recognition; revealing different kind of regularities, including negative patterns; alternative, dependent and signal features; fault-tolerant irredundant and mixed diagnostic tests and their

weight coefficients; decision making and their justification using graphic, including cognitive, tools [3, 5]. Expansion of IS DIAQOL by combining with the applied IS DIPSYS will be carried out on the base of intelligent instrumental software (IIS) IMSLOG [6], designed to revealing various of regularities, including fault-tolerant diagnostic tests and their weight coefficients; decision making and justification using cognitive graphics tools. The data and knowledge base will created using the results of research of patients with neurological pathology.

For the first time, it is proposed to expand the applied IS DIAQOL of patients with neurological pathology by combining it with the applied IS DIPSYS. For the first time, analysis, structurization of data and knowledge of patients with various neurological pathologies were carried out. Since more than 30 applied IS designed on the basis of the ISS IMSLOG have shown high efficiency, there is every reason that the use of the extended applied IS DIAQOL can improve the quality of diagnostics.

This research is funded by RFBR, grant 18-013-00937.

[1] *Neurological disorders: public health challenges.* WHO Library Cataloguing-in-Publication Data: World Health Organization, 2006. Pp. 232.

[2] *Avrov M. V.* Quality of life of patients with chronic cerebral ischemia // S.S. Korsakov Journal of Neurology and Psychiatry, 2017. Vol. 4. Pp. 56–58.

[3] *Yankovskaya A. E.* Logical tests and cognitive graphics // LAP LAMBERT Academic Publishing, 2011. Pp. 92.

[4] *Yankovskaya A. E., Obukhovskaya V. B.* Basics of creating an applied intelligent system for diagnosing the psychological safety of patients with neurological pathology // Materials of the International Conference "Scientific research of the SCO countries: synergy and integration" — Part 2: Participants' reports in English, 2019. Pp. 184–190.

[5] *Yankovskaya A. E.* Analysis of data and knowledge based on the convergence of several sciences and scientific areas // Intellectualisation of information processing: 8th international conference, 2010. Pp. 196–199.

[6] *Yankovskaya A. E.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis, 2003. Vol. 13,No 3. Pp. 650–657.

Сегментация длительных сигналов ЭЭГ на области интереса и способ дифференциации эпилептических приступов от артефактов жевания

Кершнер Иван Андреевич[1] ⋆ ivan_kershner@mail.ru
Обухов Юрий Владимирович[1] yuvobukhov@mail.ru
Синкин Михаил Владимирович[2] mvsinkin@gmail.com
[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, НИИ СП им. Н.В. Склифосовского

Для автоматического сегментирования длительных (несколько суток) сигналов ЭЭГ на области временные интервалы, значимые при диагностике посттравматической и послеоперационной эпилепсии, был разработан метод, основанный на анализе хребтов вейвлет спектров. Для сигнала ЭЭГ рассчитывалась вейвлет спектрограмма. Для точек хребта вейвлет спектрограммы анализировалась гистограмма СПМ, чтобы найти пороговое значение СПМ, при котором точки хребта разделяются на фоновую активность и на области интереса. Разработанный метод позволяет уменьшить нагрузку на медицинский персонал, сократив время, затрачиваемое на маркировку сигналов суточного мониторинга ЭЭГ, которая проводится вручную.

При нахождении отличительных характеристик эпилептической активности от артефактов жевания изучалась периодичность пиков во временных точках отсчетов хребта вейвлет спектрограммы, соответствующих эпилептической активности и артефакту жевания. Для срезов, лежащих выше частоты хребта, вычислялись спектры Фурье. По параметрам спектров артефакты жевания хорошо различимы от эпилептических приступов.

Работа выполнена в рамках государственного задания и частично поддержана Российским фондом фундаментальных исследований, проект РФФИ № 18-29-02035 мк.

[1] *Кершнер И. А., Обухов Ю. В., Синкин М. В.* Сегментация областей интереса в данных длительного мониторинга ЭЭГ послеоперационных больных эпилепсией // Физика и радиоэлектроника в медицине и экологии: Доклады 14-й международной научной конференции, 2020. С. 253–256.

Segmentation of long-term EEG signals on the area of interest and a method for differentiating epileptic seizures from chewing artifacts

Ivan Kershner[1] ⋆ ivan_kershner@mail.ru
Yury Obukhov[1] yuvobukhov@mail.ru
Mikhail Sinkin[2] mvsinkin@gmail.com
[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, N.V.Sklifosovsky Research Institute of Emergency Medicine

For automatic segmentation of long-term (several days) EEG signals on the regions of interest, which are significant in the diagnosis of post-traumatic and postoperative epilepsy, a method based on the analysis of the ridges of wavelet spectra was developed. The wavelet spectrogram with a complex Morlet basis function was calculated. For the points of the ridge of the wavelet spectrogram, the histogram of the PSD was analyzed to find the threshold value of the PSD at which the ridge points are divided into background activity and regions of interest. The developed method makes it possible to reduce the burden on medical personnel, reducing the time spent on marking the signals of daily EEG monitoring, which is currently carried out manually.

When solving the problem of identifying the distinctive characteristics of epileptic activity from chewing artifacts, the frequency of peaks at the time points of the wavelet spectrogram ridge corresponding to the peak-wave epileptic activity and chewing artefacts was studied. The sections of the wavelet spectra (vector of PSD values) at frequencies above the maximum value of the frequency of the ridge of the wavelet spectrogram were considered. The Fourier spectrum was calculated for these areas. By the parameters of Fourier spectra of the slices, it was possible to distinguish these activities.

This work was carried out within the framework of a state assignment and partially supported by the Russian Foundation for Basic Research, project RFBR grant 18-29-02035 MK.

[1] *Kershner I. A., Obukhov Yu. V., Sinkin M. V.* Segmentation of areas of interest in long-term EEG monitoring data for postoperative patients with epilepsy // Physics and Radioelectronics in medicine and ecology: Reports of the 14-th international scientific conference, 2020. Pp. 253–256.

Реконструкция функциональной структуры мозга человека по данным электроэнцефалографии

Рыкунов Станислав Дмитриевич[1]⋆ rykunov@impb.ru
Бойко Анна Ивановна[1] a.boyko@list.ru
Маслова Ольга Александровна [2] olga-a-m@mail.ru
Устинин Михаил Николаевич[1] u_m_n@mail.ru
[1]Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН
[2]Москва, ИПМ им. М.В. Келдыша РАН

Предложен новый метод анализа данных, позволяющий преобразовать многоканальные временные ряды в пространственную структуру изучаемой системы. Метод успешно использовался для изучения биологических и физических объектов с помощью измерений магнитного поля. В данной работе выполнено обобщение развитого подхода для анализа данных экспериментов, в которых измеряется электрическое поле. При помощи электроэнцефалографа с 19 каналами, расставленными по схеме 10–20, регистрировалась активность мозга человека в состоянии с закрытыми глазами. Измерялись электроэнцефалограммы в состоянии покоя, а также при произвольных движениях рук. Было выполнено преобразование Фурье полных временных рядов, что позволило получить детальные многоканальные спектры. На всех спектрах виден широкий пик альфаритма в полосе частот 9–12 Гц. Для всех спектральных компонент в этой полосе была решена обратная задача и была построена трехмерная карта активности мозга – функциональная структура источников альфа-ритма. Обратная задача решалась в приближении эквивалентного токового диполя в однослойном сферическом проводнике, без каких-либо ограничений положения источника. Совместное рассмотрение магнитно-резонансной томограммы и функциональной структуры позволяет сделать вывод о разумном согласии этой структуры с существующими представлениями об альфа-ритме человека. Также была построена трехмерная карта векторного поля доминирующих направлений источников альфа ритма. Метод может быть использован для изучения пространственного распределения активности мозга в любом спектральном диапазоне данных электроэнцефалографии.

Исследование выполнено за счет гранта Российского научного фонда (проект № 18-11-00178).

[1] *Устинин М. Н., Рыкунов С. Д., Бойко А. И., Маслова О. А.* Реконструкция функциональной структуры мозга человека по данным электроэнцефалографии // Математическая биология и биоинформатика, 2020. Т. 15 № 1. С. 106–117.

Reconstruction of the Human Brain Functional Structure Based on the Electroencephalography Data

Stanislav Rykunov[1]⋆ rykunov@impb.ru
Anna Boyko[1] a.boyko@list.ru
Olga Maslova[2] olga-a-m@mail.ru
Mikhail Ustinin[1] u_m_n@mail.ru
[1]Pushchino, IMPB RAS - Branch of KIAM RAS
[2]Moscow, Keldysh Institute of Applied Mathematics RAS

New method for the data analysis was proposed, making it possible to transform multichannel time series into the spatial structure of the system under study. The method was successfully used to investigate biological and physical objects based on the magnetic field measurements. In this paper we further develop this method to analyze the data of the experiments where the electric field is measured. The brain activity in the state of subject "eyes closed" was registered by the 19-channel electric encephalograph, using the 10-20 scheme. The electroencephalograms were obtained in resting state and with arbitrary hands motions. Detailed multichannel spectra were obtained by the Fourier transform of the whole time series. All spectral data revealed the broad alpha rhythm peak in the frequency band 9-12 Hz. For all spectral components in this band the inverse problem was solved, and the 3D map of the brain activity was calculated. The inverse problem was solved in elementary current dipole model for one-layer spherical conductor without any restrictions for the source position. The combined analysis of the magnetic resonance image and the brain functional structure leads to the conclusion that this structure generally corresponds to the modern knowledge about the alpha rhythm. The 3D map of the vector field of the dominating directions of the alpha rhythm sources was also generated. The proposed method can be used to study the spatial distribution of the brain activity in any spectral band of the electroencephalography data.

This work was supported by the Russian Science Foundation (grant 18-11-00178).

[1] *Ustinin M. N.,Rykunov S. D., Boyko A. I., Maslova O. A.* Reconstruction of the Human Brain Functional Structure Based on the Electroencephalography Mathematical Biology and Bioinformatics, 2020. Vol. 15. No 1. Pp. 106–117.

Реконструкция пространственной структуры нервной и мышечной системы тела человека по его магнитному полю

Устинин Михаил Николаевич[1]⋆ u_m_n@mail.ru
Рыкунов Станислав Дмитриевич[1] rykunov@impb.ru
Бойко Анна Ивановна[1] a.boyko@list.ru
[1]Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

Электрическая активность человеческого тела была неинвазивно реконструирована по записям магнитного поля, измеренного массивом внешних датчиков. Экспериментальные данные были получены с помощью 275-канального магнитного энцефалографа Нью-Йоркского университета. Для анализа данных использовался новый метод, основанный на преобразовании Фурье полного временного ряда и анализе паттернов магнитного поля десятков тысяч частотных компонент. Для каждой частоты восстанавливается многоканальный временной ряд, методом анализа независимых компонент выделяются когерентные составляющие, каждая из которых имеет постоянный паттерн поля. Для каждого паттерна методом полного перебора решается обратная задача в модели токового диполя. Задается сетка с пространственным разрешением от 1 до 6 мм, в зависимости от изучаемого объекта, в каждом узле сетки размещается от 8 до 12 разнонаправленных пробных диполей. Для нескольких миллионов пробных диполей вычисляются паттерны магнитного поля и выполняется их сравнение с экспериментальным паттерном. Решением обратной задачи считается диполь, порождающий паттерн, наиболее близкий к экспериментальному. Распределение спектральной мощности по координатам найденных источников образует функциональную томограмму изучаемой части человеческого тела. Метод был верифицирован на симулированных данных и на физическом фантоме, после чего использовался для восстановления функциональной структуры мозга, сердца и скелетных мышц. Также получены данные о локализации боли в мышцах спины. Полученные результаты разумно интерпретируются анатомически, что позволяет говорить о применимости метода в задачах диагностики.

Исследование выполнено за счет гранта Российского научного фонда (проект №18-11-00178).

[1] *Llinás R R., Ustinin M., Rykunov S., Walton K. D., Rabello G. M., Garcia J., Boyko A., Sychev V.* Noninvasive muscle activity imaging using magnetography // Proceedings of the National Academy of Sciences, 2020. Vol, 117. No 9. Pp. 4942–4947.

Reconstruction of the spatial structure of the human body nervous and muscular systems based on its magnetic field

Mikhail Ustinin[1]⋆ u_m_n@mail.ru
Stanislav Rykunov[1] rykunov@impb.ru
Anna Boyko[1] a.boyko@list.ru
[1]Pushchino, IMPB RAS - Branch of KIAM RAS

The electrical activity of the human body was noninvasively reconstructed from the magnetic fields registered by the array of outer sensors. The experimental data were obtained by the 275-channel magnetic encephalograph of the New York University. For the data analysis we used a new method based on the Fourier transform of the full time series and analysis of the magnetic field patterns of the tens of thousands of frequency components. For each frequency, the multichannel time series is reconstructed. The Independent Component Analysis method is used to extract coherent component, having constant pattern of the field. For each pattern, the inverse problem in the current dipole model is solved by the exhaustive search. The spatial grid is constructed with resolution from 1 to 6 mm depending on the object under study. Every node of the grid contains 8-12 directions of the trial dipoles. For several millions of trial dipoles the magnetic field patterns are calculated and compared with experimental pattern. The dipole producing the pattern, closest to experimental pattern, is supposed to be the inverse problem solution. Distribution of the spectral power between the coordinates of sources is the functional tomogram of the part of human body under study. The method was verified on simulated data and on physical phantom. Then it was used to reconstruct the functional structure of the brain, heart and skeletal muscles. Also the localization of the pain in the back muscles was performed. The results obtained can be reasonably interpreted anatomically, leading to the conclusion about good prospects of the proposed method in diagnostics.

The research was supported by the Russian Science Foundation (grant 18-11-00178).

[1] *Llinás R R., Ustinin M., Rykunov S., Walton K. D., Rabello G. M., Garcia J., Boyko A., Sychev V.* Noninvasive muscle activity imaging using magnetography // Proceedings of the National Academy of Sciences, 2020. Vol, 117. No 9. Pp. 4942–4947.

Мониторинг межканальной фазовой синхронизации ЭЭГ у пациентов с черепно-мозговой травмой до и после реабилитации

Толмачева Рената Алексеевна[1]⋆ tolmatcheva@ya.ru
Обухов Юрий Владимирович[1] yuvobukhov@mail.ru
Жаворонкова Людмила Алексеевна[2] lzhavoronkova@hotmail.com
[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, ИВНД и НФ РАН

Учитывая нарушение связей между различными областями головного мозга пациентов с черепно-мозговой травмой (ЧМТ), необходимы методы анализа электроэнцефалограмм (ЭЭГ), позволяющие определить степень связанности биопотенциалов мозга между различными каналами. Мы предложили новый метод для оценки межканальной фазовой синхронизации ЭЭГ, основанный на вычислении и сравнении фаз сигналов в точках хребтов их вейвлет-спектрограмм. Были выделены фазово-связанные пары отведений ЭЭГ, путем построения гистограмм долей разности фаз в двух отведениях ЭЭГ. Далее была построена зависимость максимального значения долей разности фаз от номера пары отведений для записи ЭЭГ без теста и во время когнитивного теста. Рассматривая разность максимальных значений долей разности фаз при когнитивном тесте и без теста, сортированную по парам отведений ЭЭГ в порядке возрастания и ее производную, целесообразно считать пары отведений с номерами большими, чем в точке резкого возрастания производной, как фазово-связанные. Сравнивая фазово-связанные пары отведений ЭЭГ у пациентов с ЧМТ до и после реабилитации с фазово-связанными парами отведений ЭЭГ у контрольных испытуемых, можно определять положительную или отрицательную динамику реабилитации.

Работа поддержана грантом РФФИ № 18-07-00609.

[1] *Толмачева Р. А., Обухов Ю. В., Жаворонкова Л. А.* Мониторинг межканальной фазовой синхронизации ЭЭГ у пациентов с черепно-мозговой травмой до и после реабилитации // Сборник статей ИТНТ-2020, 2020. Т. 4. С. 561–567.

Monitoring of inter-channel EEG phase synchronization in patients with traumatic brain injury before and after rehabilitation

Renata Tolmacheva[1]⋆ tolmatcheva@ya.ru
Yury Obukhov[1] yuvobukhov@mail.ru
Ludmila Zhavoronkova[2] lzhavoronkova@hotmail.com
[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, IHNA&NPh RAS

Taking into consideration the impaired connections between different areas of the brain of patients with traumatic brain injury (TBI), methods of analysis of electroencephalograms (EEG) are needed to determine the degree of connectivity of brain biopotentials between different channels. We proposed a new method for estimating the inter-channel phase synchronization of EEG based on calculating and comparing the phases of signals at the points of the ridges of their wavelet spectrograms. Phase-coupled pairs of EEG channels were determined by plotting histograms of the portions of the phase difference in two EEG channels. Then, the graph of the maximum values of portions for the phase difference for two EEG channels versus the number of pairs of channels for EEG record in cognitive tests, and without test was plotted. Considering the difference of the maximum values of the portions of the phase difference in the cognitive test and without the test, sorted by pairs of EEG channels in increasing order and its derivative, it is advisable to consider pairs of channels with numbers greater than at the point of the sharp increase of the derivative as phase-coupled. Comparing phase-coupled pairs of EEG channels in patients with TBI before and after rehabilitation with phase- coupled pairs of EEG channels in control subjects, it is possible to determine the positive or negative dynamics of rehabilitation.

This research is funded by RFBR, grant 18-07-00609.

[1] *Tolmacheva R. A., Obukhov Y. V., Zhavoronkova L. A.* Monitoring mezhkanalnoi fazovoi sinkhronizatsii EEG u patsientov s cherepno-mozgovoi travmoi do i posle reabilitatsii // Sbornik statey ITNT–2020, 2020. V. 4. Pp. 561–567.

Метод генерации признаковых описаний, основанный на расстояниях до эталонов, в биомедицинских исследованиях

Сенько Олег Валентинович[1,2] senkoov@mail.ru
Салманов Махир Юсиф оглы[1]⋆ sy.mahir@gmail.com
Брусов Олег Сергеевич[3] oleg.brusow@yandex.ru
Матвеев Иван Алексеевич[2] matveev@ccas.ru
Кузнецова Анна Викторовна[4] azfor@yandex.ru

[1]Москва, МГУ им. М.В.Ломоносова
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН
[3]Москва, ФБГНУ НЦПЗ РАН
[4]Москва, Институт биохимической физики им. Н.М.Эмануэля

Существующие методы машинного обучения во многом основываются на принципе компактности, предполагающие близкое расположение описаний объектов со сходными значениями целевой переменной Y. Среди различных способов реализации принципа компактности может быть выделен подход, основанный на использовании метрик заданных на множестве признаковых описаний. В качестве примера можно привести метод k-ближайших соседей и метод опорных векторов. В первом случае прогноз Y в точке $\boldsymbol{x}$ вычисляется по значениям Y на k ближайших к точке $\boldsymbol{x}$ объектов обучающей выборки. В наиболее популярном варианте метода опорных векторов прогноз значения Y вычисляется по линейной комбинации ядерных функций, каждая из которых зависит от расстояния между $\boldsymbol{x}$ и одним из опорных объектов обучающей выборки. Высокая эффективность методов машинного обучения, основанных на использовании при решении задач распознавания расстояний до эталонных объектов, подтверждена многочисленными экспериментами. Следует также отметить, что методы, основанные на оценке общей похожести рассматриваемого объекта с ранее встречавшимися случаями, часто используются при принятии решений, например, в медицине. Способы применения расстояний до эталонных объектов не могут сводится только лишь к использованию линейных комбинаций монотонно трансформированных расстояний. Альтернативным подходом является использование набора расстояний между произвольным объектом $\boldsymbol{x}$ и эталонными объектами в качестве нового векторного описания $\boldsymbol{z(x)}$, по которому далее производится прогноз значения Y с помощью алгоритма, который был обучен на выборке, состоящей из полученных таким же образом новых векторных описаний объектов исходной обучающей выборки. При этом обучение может производиться с помощью самых разных технологий, включая случайный лес и градиентный бустинг. Использование ансамблей решающих деревьев позволяет не только описывать сложные нелинейные зависимости, но и добиваться присущей ансамблевым методам более высокой устойчивости обучения.

Указанный подход был использован для решения задачи диагностики шизофрении по характеру тромбодинамики, то есть по динамике образования спон-

танных фибриновых сгустков. Необходимость использования рассматриваемой технологии при решении именно этой задачи связана с тем, что диагностику предполагается производить по кривым, показывающей временную зависимость интенсивности отражённого света на фибриновых сгустках.

Среди существующих подходов работы с такими кривыми можно выделить способ, основанный на подсчёте для каждой кривой нескольких характеризующих её параметров. Недостатком данного подхода является определённая утрата информации. Альтернативным способом является использование признаков, соответствующих отдельным временным точкам отсчёта. Каждый из таких признаков принимает значения интенсивности отражённого света соответствующей точке. Недостатками подхода являются высокая размерность при большом числе точек отсчёта и высокая коррелированность признаков. В наших исследованиях сравнивалась эффективность методов, основанного на расстояниях до эталонных объектов, и метода, основанного на использования в качестве признаков измерений на всевозможных точках отсчёта, при распознавании группу из пациентов с шизофренией и контрольной группы из испытуемых. В рамках подхода, основанного на расстояниях до эталонов исследовалась эффективность использования трёх метрик: эвклидовой метрики, косинусной метрики и метрики Минковского (p=1). Использовались два способа отбора эталонных объектов:

1. отбирались опорные вектора из всех объектов на основе метода SVC (Support Vector Classifier);
2. в качестве эталонов использовались все объекты обучающей выборки.

Метод кросс-валидации с тестированием на каждом шаге на одном объекте (Leave One Out) был использован для сравнения эффективности логистической егрессии (ЛР), случайного решающего леса (СРЛ) и градиентного бустинга (ГБ) над решающими деревьями. Оценка доверительного интервала результата классификации ROC AUC на основе бутстрапа показала, что полученные результаты стабильны. Для обеспечения наглядности результатов и возможности их интерпретации наряду с многофакторными методами распознавания использовался также метод интеллектуального анализа данных, основанный на оптимальных достоверных разбиениях признакового пространства (метод ОДР), в рамках подхода, представленного в работе [1].

Значения ROC AUC для показавшей наилучшие результаты метрики Минковского (p=1) представлены в таблице. В качестве эталонов использовались все объекты обучающей выборки и объекты, отобранных SVC:

	ЛР	СЛ	ГБ
с отбором объектов	0.737	0.778	0.706
без отбора	0.723	0.778	0.762

Проведённые исследования выявили существование выраженной внутренней структуры в данных, генерируемых с использованием расстояний до эталонных объектов. Внутренняя структура при этом характеризует также различиями между группой пациентов с шизофренией и контрольной группой.

Работа выполнена при поддержке РФФИ, проект 20-01-00609.

[1] *Доровских И.В., Сенько О.В., Чучупал В.Я., Докукин А.А., Кузнецова А.В.* Исследование возможности диагностики деменции по сигналам ЭЭГ с помощью методов машинного обучения. // Математическая биология и биоинформатика, 2019. Т. 14. № 2. С. 543–553.

The feature descriptions generating method based on distances to standards in biomedical research

Oleg Senko[1,2] senkoov@mail.ru
Mahir Salmanov[1]⋆ sy.mahir@gmail.com
Oleg Brusow[3] oleg.brusow@yandex.ru
Ivan Matveev[2] matveev@ccas.ru
Anna Kuznetsova[4] azfor@yandex.ru

[1]Moscow, Lomonosov Moscow State University
[2]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[3]Moscow, Federal State Budgetary scientific institution "Scientific Center for Mental Health" RAS
[4]Moscow, N.M. Emanuel Institute of Biochemical Physics of the Russian Academy of Sciences

The existing machine learning techniques are largely based on the principle of compactness, that assumes proximity of descriptions of objects with similar Y. An approach based on the use of metrics that are given on a set of feature descriptions can be distinguished among the various ways of implementation compactness principle. As example k-nearest neighbors and support vector machines algorithms can be given. In the first case, the prediction Y at the point $\boldsymbol{x}$ is calculated from the values of Y on the k objects of the training sample closest to the point $\boldsymbol{x}$. In the most popular version of the support vector machine, the forecast of the value of Y is calculated from a linear combination of kernel functions, each of which depends on the distance between $\boldsymbol{x}$ and one of the support objects of the training sample. The high efficiency of machine learning methods based on the use of distance to reference objects in pattern recognition problems has been confirmed by numerous experiments. It should also be noted that methods based on assessing the general similarity of the object under consideration with previously encountered cases are often used in decision-making, for example, in medicine. The methods of using distances to reference objects cannot be reduced only to the use of linear combinations of monotonically transformed distances. An alternative approach is to use a set of distances between an arbitrary object $\boldsymbol{x}$ and reference objects as a new vector description $\boldsymbol{z(x)}$, which is then used to predict thevalue of Y using an algorithm that was trained on a sample consisting of new vector descriptions of objects obtained in the same way the original training sample. At the same time, training can be performed using a variety of techniques, including random forest and gradient boosting. Using ensembles of decision trees allows not only to describe complex nonlinear dependencies, but also to achieve higher learning stability inherent in ensemble methods.

This approach was used to solve the problem of diagnostics by the characteristics of thrombodynamics, that is, by the dynamics of the formation of spontaneous fibrin clots.

Among the existing approaches to working with such curves, one can single out a method based on calculating several parameters characterizing it for each curve. The disadvantage of this approach is a certain loss of information. An alternative way is to use features corresponding to individual time points of reference. Each of these features takes on the values of the intensity of the reflected light at the corresponding point. The disadvantages of the approach are high dimensionality with a large number of reference points and high correlation of features. Our studies compared the effectiveness of methods based on distances to reference objects and a method based on the use of measurements at all possible reference points as signs when recognizing a group of patients with schizophrenia and a control group of subjects. Within the framework of the approach based on distances to standards, the efficiency of using three metrics was investigated: the Euclidean metric, the cosine metric, and the Minkowski metric (p=1). The two methods for selecting reference objects were used:

1. support vectors were selected from all objects based on the SVC method (Support Vector Classifier),
2. all objects of the training sample were used as reference objects.

The cross-validation method with testing at each step on one object (Leave One Out) was used to compare the effectiveness of logistic regression (LR), random decision forest (RDF) and gradient boosting over decision trees (GB). The estimation of the confidence interval of the classification result ROC AUC based on the bootstrap showed that the results obtained are stable. To ensure the clarity of the results and the possibility of their interpretation, along with multifactorial recognition methods, the data mining method was also used, based on the optimal valid partitioning of the feature space (OVP method), within the framework of the approach presented in [1].

The ROC AUC values for the best-performing Minkowski metric (p = 1) are presented in the table. All objects of the training sample and objects selected by SVC were used as reference objects:

	LR	RDF	GB
with objects selection	0.737	0.778	0.706
without objects selection	0.723	0.778	0.762

The conducted research revealed the existence of a pronounced internal structure in the data generated using the distances to the reference objects. The internal structure also characterizes the differences between the group of patients with schizophrenia and the control group.

This research is funded by RFBR, grant 20-01-00609

[1] *Dorovskih I.V., Senko O.V., Chuchupal V.Ya., Dokukin A.A., Kuznetsova A.V.* On Possibility of Machine Learning Application for Diagnosing Dementia by Eeg Signals. // Mathematical biology and bioinformatics, 2019. Vol. 14. No 2. Pp. 543–553

Пандемия Covid19 и методы интеллектуального анализа рисков

Ройзензон Григорий Владимирович[1,2,3]⋆ rgv@isa.ru
Соколов Александр Витальевич[4] alexander.v.sokolov@gmail.com
Черешкин Дмитрий Семенович[1] dchereshkin@yandex.ru
Комендантова Надежда Павловна[5] komendan@iiasa.ac.at
Голубков Виктор Владимирович[1] golvic@mail.ru
Бритков Владимир Борисович[1] britkov@mail.ru

[1]Москва, ИСА ФИЦ ИУ РАН
[2]Москва, Московский физико-технический институт
[3]Москва, МЭИ
[4]Москва, ИППИ РАН
[5]Лаксенбург, ИИАСА

Пандемия коронавируса (далее CV19) продемонстрировала неготовность как многих стран (включая развитые), так и различных крупных международных организаций (ООН, ВОЗ и др.) противостоять новым угрозам, а также формулировать и оценивать новые типы рисков. В работе предложен новый и усовершенствован имеющейся математический инструментарий, ориентированный на решение различных задач анализа рисков и прогнозирования. Особое внимание требует разработка новых методов оценки эффективности при введении различных социальных ограничений для успешного противостояния пандемии CV19 (самоизоляция, мониторинг пассажиропотоков, перевод на удаленную форму работы и др.). Современные реалии требуют, наряду с классическими методами, использования для решения поставленных задач технологий искусственного интеллекта (далее ИИ).

Методы анализа рисков (далее МАР) можно условно разделить на четыре большие группы. Первая группа — вероятностные (или инженерные) методы. В рамках данного подхода основные усилия направлены на сбор статистических данных о поломках, авариях и т.п. Инженерные методы ориентированы на количественный расчет вероятности поломок, отказов и других нежелательных событий. Вторая группа МАР — построение моделей на основе данных (data driven modeling). Данный подход предполагает моделирование процессов, которые определяют динамику системы, которая может попасть в различные нежелательные состояния (аварии, эпидемии и т.п.). При этом важным этапом является выбор модели соответствующей количеству и качеству исходной информации (цифровым данным и знаниям о функционировании моделируемого объекта). Для этого предлагается использовать метод сбалансированной идентификации, который позволяет достигнуть компромисса между сложностью математической модели и погрешностью описания используемого массива данных. Для использования метода необходимо: 1) определить (параметрическое) семейство моделей, пригодных для удовлетворительного воспроизведения измерений;

2) формализовать понятия сложности модели (для выбранного семейства); 3) задать меру близости траектории модели к исходным данным; 4) определить процедуру оценки погрешности моделирования (например, использовать процедуру перекрестной проверки); 5) провести поиск оптимального компромисса между сложностью модели и близостью к измерениям на основе минимизации погрешности моделирования измерений. Кроме того, авторами работы на протяжении длительного периода развивается оригинальный подход глобального моделирования, основанный на использовании сразу нескольких моделей (мультимодельный подход), применение которого для решения поставленной задачи представляется весьма перспективным. Мультимодельный подход использован при решении различных крупномасштабных практических задач (например, оценки экономических и демографических потенциалов стран). Третья группа МАР — экспертные. При применении инженерного и модельного подходов достаточно часто возникают ситуации, когда наблюдается дефицит статистических данных (или есть сомнения в их достоверности). Кроме того, при построении моделей в ряде случаев затруднительно выявить различные зависимости (слабоструктурированные задачи). В такой ситуации фактически единственным источником сведений являются эксперты. В рамках экспертного подхода хорошо себя зарекомендовали методы вербального анализа решений, ориентированные на слабоструктурированные задачи многокритериального стратегического выбора. Экспертные методы могут быть непосредственно использованы как для оценки рисков, так и для оценки эффективности от введения дополнительных ограничительных мер, что предполагает разработку специальных систем критериев, по которым можно будет судить о степени достижения поставленных задач (целей). При этом критерии могут быть условно разделены на три большие группы. Первую группу образуют критерии, позволяющие для противодействия пандемии CV19 оценивать имеющийся ресурс (количество больничных коек, приборов ИВЛ, медицинского персонала, медикаментов и т.п.). Вторая группа критериев позволяет оценивать скорость расходования и прироста ресурса во времени (возможность использования медицинских специалистов смежных специальностей, ускоренный ввод в эксплуатацию объектов медицинской сферы, возможности закупки лекарств и оборудования за рубежом и др.) необходимого для борьбы с пандемией CV19. Наконец, третья группа критериев позволяет сделать вывод о степени достижения поставленных целей (например, предполагается, что в течение определенного срока (например, года) вирусом заразиться не более 2 процентов населения, или процент падения ВВП и т.п.). Таким образом, оценки по указанным составным критериям, позволят сделать вывод насколько принятые меры, и полученные результаты, являются эффективными. Кроме того, важнейшим направлением использования экспертного подхода для анализа риска является возможность исследования вопросов безопасности критических инфраструктур в условиях пандемии CV19. Авторами работы разработаны теория и методы управления рисками нарушения

безопасности критических инфраструктур и их критически важных объектов. Четвертый подход MAP — социологический. В рамках данного подхода предполагается измерить восприятие населением и его отдельными группами того или иного риска. В рамках работы предполагается провести дополнительные исследования влияния человеческого фактора на безопасность критически важных объектов и сформулировать рекомендации по его учету при расчете и анализе рисков в условиях пандемии CV19. Работа поддержана грантами РФФИ № 19-07-00522 и № 19-010-00423.

[1] *Chereshkin D., Royzenson G., Britkov V.* Multidimensional classifier of risk analysis methods // 11th World Conference «Intelligent Systems for Industrial Automation», 2021.

Covid19 pandemic and artificial intelligence methods for risk analysis

Gregory Royzenson[1,2,3]⋆ rgv@isa.ru
Alexander Sokolov[4] alexander.v.sokolov@gmail.com
Dmitriy Chereshkin[1] dchereshkin@yandex.ru
Nadejda Komendantova[5] komendan@iiasa.ac.at
Viktor Golubkov[1] golvic@mail.ru
Vladimir Britkov[1] britkov@mail.ru

[1]Moscow, ISA FRCCSC of RAS
[2]Moscow, Moscow Institute of Physics and Technology
[3]Moscow, MPEI
[4]Moscow, IITP of RAS
[5]Laxenburg, IIASA

The coronavirus pandemic (hereinafter CV19) has demonstrated the unwillingness of both many countries (including developed) and various large international organizations (UN, WHO, etc.) to confront new threats, as well as to formulate and assess new types of risks. The paper proposes a new and improved existing mathematical toolkit, focused on solving various problems of risk analysis and forecasting. Particular attention is required to develop new methods for assessing the effectiveness of the introduction of various social constraints for a successful response to the pandemic CV19 (self-isolation, monitoring of passenger traffic, transfer to a remote form of work, etc.). Modern realities require, along with classical methods, the use of artificial intelligence technologies (hereinafter AI) for solving the assigned tasks.

Risk analysis methods (hereinafter RAM) can be roughly divided into four large directions. The first direction is probabilistic (or engineering) methods. The main efforts are focused in this direction on collecting statistical data on failures and accidents that involve leakage of harmful substances into environment. The engineering direction concentrates on quantitative calculation of the probability of failures, malfunctions, and other undesirable events. The second direction of RAM is data driven modeling. This approach involves modeling the processes that determine the dynamics of the system, which can get into various undesirable states (accidents, epidemics, etc.). In this case, an important stage is the choice of a model corresponding to the quantity and quality of the initial information (digital data and knowledge about the functioning of the modeled object). For this, it is proposed to use the balanced identification method, which allows reaching a compromise between the complexity of the mathematical model and the error in describing the data array used. To use the method, it is necessary to: 1) define a (parametric) family of models suitable for satisfactory reproduction of measurements; 2) formalize the concept of model complexity (for the selected family); 3) set a measure of the proximity of the model trajectory to the original data; 4) define a procedure for estimating the modeling error (for example, use a cross-validation procedure); 5) to search for the optimal

compromise between the complexity of the model and the proximity to measurements based on minimizing the measurement modeling error. In addition, over a long period of time, the authors of the work have been developing an original approach to global modeling based on the use of several models (multi-model approach), the use of which for solving the problem is very promising. The multi-model approach is used in solving various large-scale practical problems (for example, assessing the economic and demographic potentials of countries). The third direction of RAM is expert. When applying the engineering and model approaches, situations often arise when there is a shortage of statistical data (or there are doubts about their reliability). In addition, when constructing models, in some cases it is difficult to identify various dependencies (semi-structured problems). In such a situation, in fact, the only source of information is experts. Within the framework of the expert approach, methods of verbal decision analysis focused on semi-structured problems of multi-criteria strategic choice, have proven themselves well. Expert methods can be directly used both for assessing risks and for assessing the effectiveness of the introduction of additional restrictive measures, which implies the development of special systems of criteria by which it will be possible to judge the degree of achievement of the tasks (goals). In this case, the criteria can be conditionally divided into three large groups. The first group is formed by the criteria that make it possible to assess the available resource (the number of hospital beds, ventilators, medical personnel, medicines, etc.) to counter the CV19 pandemic. The second group of criteria makes it possible to assess the rate of expenditure and increase in the resource over time (the possibility of using medical specialists of related specialties, accelerated commissioning of medical facilities, the possibility of purchasing drugs and equipment abroad, etc.) necessary to combat the CV19 pandemic. Finally, the third group of criteria allows us to conclude about the degree of achievement of the set goals (for example, it is assumed that within a certain period (for example, a year) no more than 2 percent of the population will become infected with the virus, or the percentage of GDP decline, etc.). Thus, the assessments according to the specified complex criteria will make it possible to conclude how effective the measures taken and the results obtained are. In addition, the most important area of using the expert approach for risk analysis is the ability to study the security of critical infrastructures in the context of the CV19 pandemic. The authors of the work have developed the theory and methods of risk management of security breaches of critical infrastructures and their critical facilities. The fourth RAM direction is sociological. Within the framework of this approach, it is supposed to measure the perception of the population and its individual groups of a particular risk. As part of the work, it is planned to conduct additional studies of the influence of the human factor on the safety of critical facilities and formulate recommendations for taking it into account when calculating and analyzing risks in the context of the CV19 pandemic. This research is funded by RFBR, grants 19-07-00522 and 19-010-00423.

[1] *Chereshkin D., Royzenson G., Britkov V.* Multidimensional classifier of risk analysis methods // 11th World Conference ¡¡Intelligent Systems for Industrial Automation¿¿, 2021.

Технология сбалансированной идентификации: выбор модели динамики COVID-19 по имеющимся данным

Соколов Александр Витальевич[1]⋆ alexander.v.sokolov@gmail.com
Соколова Любовь Александровна[2] las.sokolova@gmail.com
[1]Москва, Институт проблем передачи информации РАН
[2]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук Институт Системного Анализа

Разделить сложное явление на составляющие; рассмотреть процессы, определяющие его динамику; формализовать принятые гипотезы в виде математических уравнений; подобрать соответствующий экспериментальный и статистический материал и, в итоге, построить математическую модель – типичные задачи естественно-научного исследования. В данной работе сложное био-социальное явление эпидемии COVID-19 исследуется на основе технологии сбалансированной идентификации. Это позволило рассмотреть ряд моделей, определить биологические закономерности взаимодействия вируса с человеком (общие для всех популяций) и социальные особенности управления эпидемией в рассматриваемых странах и регионах (различные в различных популяциях). В качестве исходных данных используются новые (ежесуточные) случаи заражения (new cases) – официальные статистические данные для ряда стран и регионов. Полученные оценки числа невыявленных зараженных являются оценками снизу. Привлечение дополнительной информации (число носителей антител) позволяет получить более реалистичные оценки.

Динамика развития эпидемии (в т.ч. и COVID-19) определяется как биологическими особенностями взаимодействия человеческого организма и вируса, так и социальными аспектами взаимодействия человека и общества.

Биологические особенности определяют:

1) потенциальное количество зараженных одним больным,
2) вероятность самостоятельного выздоровления (без выявления и изоляции),
3) видимость (степень манифестации) симптомов (для выявления и последующей изоляции).

Общество может управлять эпидемией тремя способами:

1) ограничивать контакты (самоизоляция и карантин для имевших контакт с заражёнными) или делать их более безопасными (дистанция, маски и т.д.),
2) выявлять и изолировать инфицированных,
3) проводить вакцинацию населения и стимулировать прием профилактических лекарств (в данном исследовании не рассматриваются).

В работе используются популяционные модели распространения вируса в популяции человека (типа «паразит-хозяин» или «хищник-жертва»). Особое внимание уделяется разделению биологических процессов взаимодействия хищника

(вируса) и жертвы (человека) и социальных механизмов противодействия общества эпидемии. В используемой (динамической) модели внутренняя структура соответствует современным представлениям о биологической природе объекта, а внешние управления отражают специфику противоэпидемических мероприятий различных стран.

Динамику эпидемии определяют невыявленные зараженные (НВЗ). Причина в том, что выявленные зараженные, в той или иной степени, изолируются обществом, а невыявленные продолжают "размножаться" – заражать других. Таким образом, рассматривается только популяция НВЗ – тех, кто не попадает в статистику по новым случаям заражения.

Скорость эпидемического процесса в значительной степени определяется заразностью носителя вируса, которая существенно зависит от длительности заражения. Так, например, больной COVID-19 становится заразным (латентный период) в среднем через 4-5 дней с момента заражения. Следовательно, для описания динамики НВЗ целесообразно разбить популяцию НВЗ на группы по времени, прошедшему с момента заражения, которое будем называть длительностью заражения (ДЗ). Модели такого типа давно известны и широко используются в демографии, экологии и эпидемиологии.

Для построения (идентификации) моделей используется только один вид статистической информации – стандартные официальные данные по количеству новых (выявленных за сутки) случаев заражения (new cases). Данные используются "как есть", без предварительной обработки. Страны с нестандартными данными (например, Китай) не рассматриваются. Число выбранных стран и регионов (всего 7) определялось наглядностью отображения результатов моделирования на одном рисунке. Используемая технология сбалансированной идентификации позволяет увеличить их количество в несколько раз. В построенной модели рассматриваются 7 популяций – населения стран/регионов: Великобритания (Gbr), Германия (Deu), Италия (Ita), Испания (Esp), Франция (Fra), Россия без Москвы и Московской области (Rus-) и города Москвы с Московской областью (Mos+). Разбиение России на две части вызвано существенно различающейся динамикой частей.

Принятая в итоге модель была выбрана из нескольких возможных как та, что наилучшим образом соответствует объему и качеству статистических данных. Для этого использовался метод сбалансированной идентификации [1],[2], который позволил количественно оценить, насколько принятый набор гипотез о функционировании объекта (био-социальной системы) соответствует доступному фактическому материалу (статистическим данным).

Работа поддержана грантом РФФИ № 20-07-00701.

[1] *Соколов А. В., Волошинов В. В.* Выбор математической модели: баланс между сложностью и близостью к измерениям // International Journal of Open Information Technologies, 2018. Т. 6. № 9. С. 33–41.

[2] *Sokolov A. V., Voloshinov V. V.* Model Selection by Balanced Identification: the Interplay of Optimization and Distributed Computing // Open Computer Science, 2020. Т. 10. С. 283–295.

Balanced Identification Technology: Choosing COVID-19 Dynamics Model for Available Data

Alexander Sokolov[1]⋆ alexander.v.sokolov@gmail.com
Lyubov Sokolova[2] las.sokolova@gmail.com
[1]Moscow, Institute for Information Transmission Problem RAS
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS Institute for Systems Analysis

Typical tasks of scientific research include breaking down a complex phenomenon into its components, considering the processes that determine its dynamics, formalizing the accepted hypotheses in mathematical equations, selecting appropriate experimental and statistical material, and ultimately, constructing a mathematical model. This paper explores a complex bio-social phenomenon (COVID-19) using a specific data processing method - balanced identification. The method combined with appropriate information technology made it possible to consider a number of models, determine the general biological laws of the virus vs. human interaction (common to all populations), and the country specific social epidemic management in the populations under consideration. As the initial data, only new cases were used. Data from different countries was taken from official sources and processed in a uniform way. The obtained estimates of the number of undetected infected are lower estimates. Further information (antibody carriers estimation) accounts for more realistic estimates.

The dynamics of epidemics (including COVID-19) is determined both by biological characteristics of the human body vs. virus interaction, and by social aspects of the interaction of man and society.

Biological characteristics determine:

1) "potential" number of people infected by one person,
2) the probability of recovery (without detection and isolation),
3) "visibility" (manifestation) of symptoms (for identification and subsequent isolation).

Society can manage an epidemic in three ways:

1) limit contacts (self-isolate and quarantine those who had contacts with infected) or make them safer (distancing, masks, etc.)
2) identify and isolate those infected,
3) vaccination and/or prophylactic of population (out of scope of the paper).

The paper uses population-based models of the virus spreading in human population (such as "host-parasite" or "predator-prey"). Particular attention is paid to the separation of the biological processes of interaction between the predator (virus) and the prey (human) and the social mechanisms of society counteracting an epidemic. In the (dynamic) model used, the internal structure corresponds to modern ideas

about the biological nature of the object, and external controls reflect the specifics of anti-epidemic measures in different countries.

The epidemic dynamics is determined by undetected infected people (UDI) – those identified are isolated (more or less carefully) by the society, and those not identified continue to "multiply", i.e. infect others. So, only the population of UDI (overlooked by daily new cases statistics) is considered.

Epidemics spreading speed is largely defined by the contagiousness of the infected, which substantially depends on the duration of infection. The typical time from infection to the moment when the host becomes contagious (latent period) is about 4-5 days for COVID-19. Therefore, to describe the dynamics of UDI population it should be divided into groups according to the time elapsed since the infection, which we will call the duration of infection (DI). Models of this type have long been known and are widely used in demography, ecology and epidemiology.

A singe type of statistical information is used to build (identify) the models – the standard official data on the number of new (detected per day) cases of infection. Data is used "as is" without pre-processing. Countries with abnormal data (for example, China) are ignored. The number of countries present (7 populations) was determined by visual restrictions of showing the simulation results in one figure. Meanwhile the technology of balanced identification allows to increase the number several times. The constructed model considers 7 populations — the populations of Great Britain (Gbr), Germany (Deu), Italy (Ita), Spain (Esp), France (Fra), Russia excluding the city of Moscow and the Moscow Region (Rus-) and the city of Moscow including the Moscow Region (Mos+). The division of Russia into two parts is caused by the significant difference in the dynamics.

The model under consideration was selected as the one that best fits the quantity and quality of the selected statistics. The balanced identification method used [1],[2] allowed us to quantify how much the accepted set of hypotheses about the functioning of an object (bio-social system) corresponds to the available factual material (statistics).

This research is funded by RFBR, grant 20-07-00701.

[1] *Sokolov A. V., Voloshinov V. V.* Choice of mathematical model: balance between complexity and proximity to measurements. // International Journal of Open Information Technologies, 2018. Vol. 6 No 9. Pp. 33–41.

[2] *Sokolov A. V., Voloshinov V. V.* Model Selection by Balanced Identification: the Interplay of Optimization and Distributed Computing // Open Computer Science, 2020. Vol. 10. Pp. 283–295.

Разработка и развитие базы данных двухспиральных мотивов белковых молекул и вычислительные сервисы для их анализа

Руднев Владимир Ремович[1]⋆ volodyrv@mail.ru
Куликова Людмила Ивановна[1,2] likulikova@mail.ru
Кайшева Анна Леонидовна[3] kaysheva1@gmail.com
Тихонов Дмитрий Анатольевич[1,2] dmitry.tikhonov@gmail.com

[1]Пущино, Институт теоретической и экспериментальной биофизики РАН
[2]Пущино, Институт математических проблем биологии РАН филиал ИПМ им. М.В. Келдыша РАН
[3]Москва, Научно–исследовательский институт биомедицинской химии имени В.Н. Ореховича

Данная работа посвящена созданию базы данных структурных мотивов белковых молекул, состоящих из двух элементов вторичной структуры, имеющих уникальные укладки полипептидной цепи в пространстве. Исследуемые мотивы представляют собой пары любого типа спиралей, соединёнными между собой различной ненулевой длины и различной конформации перетяжкой. Пространственная ориентация двух спиралей определяет тип спиральной пары: $\alpha - \alpha$ — уголок, V — структура, L — структура, $\alpha - \alpha$ — шпилька и др. Для каждой структуры рассчитаны геометрические параметры. Описываемая в работе база данных «Structural Elements Database» объединяет сервисы хранения данных и вычислительные алгоритмы для их анализа. Strucural Elements DB на сегодня содержит свыше 45000 аннотированных белковых мотивов – спиральных пар. Был создан новый интерфейс базы данных «Structural Elements DB», который дополнен возможностями графического представления данных отдельной спиральной пары, а также статистической обработки выборки отсортированных спиральных пар. Интерфейс позволяет:

- строить выборки структурных мотивов по интересующим геометрическим параметрам;
- исследовать взаимосвязь геометрии пространственных структур с аминокислотной последовательностью с помощью разработанных инструментов;
- выполнять операции поиска, сортировки, фильтрации по всем параметрам;
- получать выборки структур с заданными геометрическими характеристиками;
- проводить статистический анализ и строить гистограммы распределения различных характеристик структур в выборке;
- просматривать 3D модели двухспиральных мотивов.

Предусмотрена возможность загрузки результатов пространственного и математического анализа отдельных спиральных пар, а также выборки спиральных пар.

Работа поддержана грантом РФФИ №18-07-01031-а.

[1] *Rudnev V. R., Tikhonov D. A., Kulikova L. I., Gubin M. Yu., Efimov A. V.* Database of two-helical motifs of protein molecules and computer services for their analysis. // Journal of Bioinformatics and Genomics, 2019. Vol. 3, No 12.

Creation and development of a database of two helical motifs of protein molecules and computational services for their analysis

Vladimir Rudnev[1]⋆ volodyrv@mail.ru
Ludmila Kulikova[1,2] likulikova@mail.ru
Anna Kaysheva[3] kaysheva1@gmail.com
Dmitry Tikhonov[1,2] dmitry.tikhonov@gmail.com

[1] Pushchino, Institute of Theoretical and Experimental Biophysics of RAS
[2]Pushchino, Institute of Mathematical Problems of Biology Branch of Keldysh Institute of Applied Mathematics of RAS
[3]Moscow, V.N. Orekhovich Institute of Biomedical Chemistry

This work devoted to the development of structural motifs database of protein molecules consisting of two elements of a secondary structure that have unique spatial stacking of a polypeptide chain. The motives investigated are pairs of any type of helix, connected by a different non-zero length and different conformation of the connection. The spatial orientation of the two α — helices determines the type of helical pair: $\alpha - \alpha$ — corner, V – structure, L – structure, $\alpha - \alpha$ — hairpin et al. For each structure, geometric parameters are calculated. The Structural Elements Database combines data storage services and computational algorithms for their analysis. Structural Elements DB currently contains over 45000 annotated protein motifs – helical pairs. A new interface for the "Structural Elements DB" database was created, which was supplemented with the capabilities of graphical presentation of data for an individual helical pair, as well as statistical processing of a sample of sorted helical pairs. The interface allows:

- to build a selection of structural motives according to the geometric parameters of interest;
- to investigate the relationship of the geometry of spatial structures with the amino acid sequence using the developed tools;
- perform operations of search, sorting, filtering by all parameters;
- to receive samples of structures with specified geometric characteristics;
- conduct statistical analysis and build histograms of the distribution of various characteristics of structures in the sample;
- View 3D models of two helical motifs.

It is possible to download the results of spatial and mathematical analysis of individual helical pairs, as well as a sample of helical pairs.

This research is funded by RFBR, grant 18-07-01031-a.

[1] *Rudnev V. R., Tikhonov D. A., Kulikova L. I., Gubin M. Yu., Efimov A. V.* Database of two-helical motifs of protein molecules and computer services for their analysis. // Journal of Bioinformatics and Genomics, 2019. Vol. 3, No 12.

Подходы к мультиклассовой классификации датасета потенциалов P300

Гончаренко Владислав Владимирович[1,2] vladislav.goncharenko@phystech.edu
Григорян Рафаэль Каренович[1,3] grraph.bio@gmail.com
Самохина Алина Максимовна[1,2]⋆ alina.samokhina@phystech.edu
[1]Москва, Neiry
[2]Москва, Московский физико-технический институт
[3]Москва, МГУ

Нейрокомпьютерные интерфейсы (НКИ) изначально применялись только для пациентов с ограниченными возможностями. На сегодняшний день существует множество попыток использовать НКИ на неинвазивных электродах для здоровых людей, к примеру, в играх.

Мы публикуем авторский датасет потенциалов P300, полученных на визуальных стимулах в игре в виртуальной реальности. В работе приводятся описания структуры данных датасета, игрового процесса, участников и оборудования. Также описан препроцессинг данных и указаны оптимальные параметры для всех этапов обработки данных, найденные в результате вычислительных экспериментов.

Основная задача данной работы — задача определения объекта внимания человека по электроэнцефалограмме (ЭЭГ). По данным, после их предобработки, проводится мультиклассовая классификация для определения целевого визуального стимула. В случае работы со стимулами P300 по ‘oddball‘-парадигме, мультиклассовая классификация основывается на результатах бинарной классификации. Бинарный классификатор обучается на каждом человеке. При этом задача определения наличия P300 в отрезке ЭЭГ имеет дисбаланс классов (только один стимул из семи может быть целевым).

В связи с этим, к стандартным шагам препроцессинга ЭЭГ (децимация, фильтрация, ресэмплинг, клиппинг и нормировка) предлагается добавить аугментацию данных. Данный шаг позволяет сбалансировать классы для обучения бинарного классификатора и, соответственно, повысить качество мультиклассовой классификации. В данной работе приводится два способа аугментаций. Назовём эпохой отрезок времени, где мы предполагаем наличие стимула P300, отсчитываемый от момента предъявления визуального стимула. Тогда первый вариант аугментации данных — смещение точки отсчёта старта эпох. Второй — изменение данных ЭЭГ в эпохе с помощью алгоритма SMOTE. Использование аугментаций позволяет повысить качество мультиклассовой классификации на величину до 6%.

Также для определения потенциально неверных ответов итогового классификатора вводится понятие "уверенности"(уровень уверенности модели в ответе). Предполагается, что если распределение вероятностей стимулов не имеет явно выраженного максимума, то ответ классификатора ошибочный. Тогда по

предсказаниям бинарного классификатора можно построить критерий уверенности модели в ответе. Из различных рассмотренных эвристик наиболее удачным определением критерия уверенности является разность максимума и медианы вероятностей, полученных для всех стимулов. Если значение уверенности ниже порогового ($\sim 20\%$), то ответ классификатора не принимается средой (к примеру, в игровом процессе вместо выстрела происходит осечка). К дальнейшему рассмотрению предлагается использование отдельной модели на вероятностях бинарных классификаторов для предсказания уровня уверенности в конечном ответе.

[1] *Goncharenko V., Grigoryan R., Samokhina A.* Raccoons vs Demons: multiclass labeled P300 dataset // `https://arxiv.org/abs/2005.02251`

Approaches to multiclass classification of the P300 dataset

Vladislav Goncharenko[1,2] vladislav.goncharenko@phystech.edu
Rafael Grigoryan[1,3] grraph.bio@gmail.com
Alina Samokhina[1,2]⋆ alina.samokhina@phystech.edu
[1]Moscow, Neiry
[2]Moscow, Moscow Institute of Physics and Technology
[3]Moscow, MSU

Firstly Brain-Computer Interfaces (BCI) were developed mostly for disabled patients. Nowadays there are numerous attempts to employ BCI solutions for healthy people using noninvasive electrodes, for example, in games.

We publish dataset of P300 potentials, recorded on visual stimuli of the game in virtual reality (VR). In this work we describe the data structure of the dataset, gameplay, equipment and consolidated information on the participants. Also, we provide information on the data preprocessing and its optimal parameters found during computational experiments.

The main purpose of this paper is to identify the object of a person's attention by electroencephalogram (EEG). We do multiclass classification on preprocessed data to identify the target stimuli. In case of the oddball paradigm, multiclass classification is based on the results of a binary one. Binary classifier is trained for each person independently and the task of determining the P300 presence in EEG is unbalanced (only one stimulus out of seven can be the target one).

To solve this problem we suggest to add data augmentation to the standard preprocessing (decimation, filtering, resampling, clipping, normalization). The augmentation allows us to balance classes for the training of binary classifier and to improve the multiclass classification accuracy up to 6%. In this paper we propose two methods of augmentation. We call the EEG segment, where we expect to see P300, counted from the moment of visual stimulus appearance — epoch. So the first type of augmentation is shifting the epoch start. The second type is altering the epoch with SMOTE algorithm.

To separate wrongly classified answers we introduce the confidence level. Supposedly, if the distribution of stimuli probabilities doesn't have explicit maximum, the given answer is likely to be wrong. So we can define a criterion of model's confidence level. We looked at different heuristics and the most informative was the difference between maximum and median of stimuli probabilities. If the value of confidence is less than some threshold ($\sim 20\%$), then the answer of classifier won't be accepted by the system (e.g. in the game there will be a misfire instead of a shot). For further discussion we propose the usage of a separate model on binary classifiers' probabilities for confidence prediction.

[1] *Goncharenko V., Grigoryan R., Samokhina A.* Raccoons vs Demons: multiclass labeled P300 dataset // https://arxiv.org/abs/2005.02251

Использование однородной семантической сети для классификации результатов генетического анализа

Куликов Алексей Михайлович[1] amkulikov@gmail.com
Харламов Александр Александрович[2, 3, 4]⋆ kharlamov@analyst.ru
[1]Москва, РАН, Институт биологии развития им. Кольцова
[2]Москва, РАН, Институт высшей нервной деятельности и нейрофизиологии
[3]Москва, Московский государственный лингвистический университет
[4]Москва, Высшая школа экономики

В работе показано использование механизма сравнения семантических сетей текстов в задаче диагностики заболеваний с использованием сигнальных сетей. Выявление степени пересечения семантических сетей текстов позволяет говорить о степени их смыслового подобия. Однородная семантическая сеть как множество узлов, связанных дугами, имеет численные характеристики – частоты появления слов, а также пар слов в тексте, которые перенормируются с использованием n-граммной модели текста. Такие сети как смысовые портреты текстов могут служить для сравнения (и, следовательно, для классиффикации) текстов. Генетический квазитекст может быть представлен, в том числе, в виде сигнальной или генной сети. Сигнальные сети разных классов генетических событий могут быть использованы для классификации этих текстов. В этом случае концентрации белков, выявленные в процессе эксперимента, используются для вычисления числовых характеристик узлов сети. Приведены примеры сравнения сетей генетических квазитекстов, соответствующих норме и паталогии.

Работа поддержана грантом Фонда поддержки малого и среднего бизнеса №550ГРНТИС5/49492 от 26.09.2019.

[1] *Kulikov A. M., Kharlamov A. A.* Using a Homogeneous Semantic Network to Classify the Results of Genetic Analysis // Neuroinformatics and Semantic Representations. Theory and Applications. Collective Monography. Chapter Eleven, 2020. Pp. 244–254.

Using a homogeneous semantic network to classify the result of genetic analysis

Alexey Kulikov[1] amkulikov@gmail.com
Alexander Kharlamov[2, 3, 4]⋆ kharlamov@analyst.ru
[1]Moscow, RAS, Koltsov Institute of Developmental Biology
[2]Moscow, RAS, Institute of Higher Nervous Activity and Neurophysiology
[3]Moscow, Moscow State Linguistic University
[4]Moscow, Higher School of Economics

The paper describes the use of a mechanism for comparing semantic networks of texts in the problem of diagnosing diseases using signaling networks. The identification of the degree of intersection for semantic networks of texts makes it possible to state the degree of their semantic similarity. A homogeneous semantic network as a set of nodes connected by arcs has numerical characteristics: the frequency of occurrence of words, as well as word pairs in the text that are renormalized using the n-gram model of the text. Networks such as semantic portraits of texts can be used to compare (and, therefore, to classify) texts. A genetic quasi-text can be represented, in particular, in the form of a signaling or gene network. Signaling networks of various classes of genetic events can be used to classify these texts. In this case, the protein concentrations detected during the experiment are used to calculate the numerical characteristics of the network nodes. Examples of networks comparison for genetic quasi-texts corresponding to normal and pathological conditions are provided.

This research is funded by FASIE, grant 550GRNTIS5/49492 26.09.2019.

[1] *Kulikov A. M., Kharlamov A. A.* Using a Homogeneous Semantic Network to Classify the Results of Genetic Analysis // Neuroinformatics and Semantic Representations. Theory and Applications. Collective Monography. Chapter Eleven, 2020. Pp. 244–254.

Один тип искусственной нейронной сети на основе нейронов с временной суммацией сигналов

Харламов Александр Александрович[1, 2, 3]⋆ kharlamov@analyst.ru

[1]Москва, РАН, Институт высшей нервной дщестельности и нейрофизиологии
[2]Москва, Московский государственный лингвистический университет
[3]Москва, Высшая школа экономики

Большая часть современных нейросетевых парадигм манипулирует нейроподобными элементами с пространственной суммацией сигналов. В настоящее время искусственные нейронные сети усложнились введением в модель нейрона свойств накопления значения потенциала в синапсе, задержек при проведении сигнала по аксону – импульсные сети [1]. Интересные свойства искусственная нейронная сеть приобретает в случае использования нейроподобных элементов с временной суммацией сигналов – включением в модель нейрона задержек в дендрите – на входе в нейрон. Влияние электронекомпактности нейрона на процессы взаимодействия постсинаптических потенциалов приводит к формированию принципиально иного, чем в большинстве существующих искусственных нейронных сетей, подхода к обработке информации такими нейронами. Возникающая в результате упомянутой электронекомпактности зависимость реакции нейрона на временную структуру последовательности входных сигналов дает возможность избирательно адресовать один конкретный нейрон из множества подобных. Такая избирательная адресация приводит к принципиально иному пониманию структуры обработки информации в нейроне. Различные распределения возбуждающих и тормозных синапсов на этой модели дендрита как адреса конкретных нейронов моделируются вершинами n-мерного единичного гиперкуба в многомерном пространстве R^n. Тогда любая входная последовательность может быть представлена как последовательность сработавших нейронов – траектория в многомерном сигнальном пространстве.

[1] *Kharlamov A. A.* On a Type of Artificial Neural Network Based on Neurons with Temporal Summation of Signals //Neuroinformatics and Semantic Representations, 2020. Pp. 47–56.

On a type of artificial neural network based on neurons with temporal summation of signals

Alexander Kharlamov[1, 2, 3] kharlamov@analyst.ru

[1]Moscow, RAS, Institute of Higher Nervous Activity and Neurophysiology
[2]Moscow, Moscow State Linguistic University
[3]Moscow, Higher School of Economics

Most modern neural network paradigms manipulate neural-like elements with spatial summation of signals. At present, artificial neural networks have become complicated by the introduction of the properties of synaptic potential value accumulation into the neuron model, and delays in conducting the signal along the axon – so called impulse networks [1]. An artificial neural network acquires interesting properties when using neural-like elements with a temporal summation of signals – inclusion of dendritic delays into the neuron model, at the neuron input. The effect of electric non-compactness of the neuron on the interaction processes of postsynaptic potentials leads to the formation of a fundamentally different approach to information processing by such neurons than in most existing artificial neural networks. The dependence of the neuron reaction on the temporal structure of the sequence of input signals arising as a result of the mentioned electric non-compactness makes it possible to selectively address one particular neuron from many similar ones. Such selective addressing leads to a fundamentally different understanding of the structure of information processing in a neuron. Different distributions of excitatory and inhibitory synapses of this dendrite model as an address of specific neurons are modelled by nodes of an n-dimensional unit hypercube in a multidimensional space R^n. Then, any input sequence can be represented as a sequence of triggered neurons – a trajectory in a multidimensional signal space, provided that there will be a complete set of these addresses (neurons with corresponding distributions of excitatory and inhibitory synapses).

[1] *Kharlamov A. A.* On a Type of Artificial Neural Network Based on Neurons with Temporal Summation of Signals //Neuroinformatics and Semantic Representations, 2020. Pp. 47–56.

О новых типах медианы Кемени

Двоенко Сергей Данилович[1]⋆ sergedv@yandex.ru
Пшеничный Денис Олегович[1] denispshenichny@yandex.ru
[1]Тула, Тульский государственный университет

Согласованное ранжирование как мнение экспертной группы может быть представлено медианой Кемени в известной задаче агрегирования рангов. Такое решение в наименьшей степени отличается от других ранжирований и свободно от некоторых противоречий проблемы правила большинства. Как математический принцип, медиана Кемени дает решение в любом случае, в частности, для конфликтующих экспертов или групп. На практике обычно выявляются конкурирующие мнения, где для достижения требуемого уровня консенсуса используются специальные процедуры. Один из известных подходов заключается в присвоении весов экспертным мнениям. В данной работе сформулирована корректная математическая основа для новых типов медианы Кемени, таких как метрическая и взвешенная медианы. Задача построения медианы для линейной комбинации экспертных ранжирований исследуется на основе известного локально-оптимального алгоритма Кемени. Предложено исследовать задачу агрегирования рангов на основе новых типов медианы.

Работа поддержана грантами РФФИ № 20-07-00055, 18-07-01087, 18-07-00942.

[1] *Dvoenko S. D., Pshenichny D. O.* Rank Aggregation Based on New Types of the Kemeny's Median // Pattern Recognition and Image Analysis, 2021. Vol. 31. № 1.

On New Types of the Kemeny's Median

Sergey Dvoenko[1]⋆ sergedv@yandex.ru
Denis Pshenichny[1] denispshenichny@yandex.ru
[1]Tula, Tula State University

A coordinated ranking as an opinion of an expert group can be represented by the Kemeny's median in a well-known rank aggregation problem. Such a decision is a least different from other rankings, and is free of some contradictions of the majority rule problem. As a mathematical principle, the Kemeny's median gives a decision in any case, in particular, for conflicting experts or groups. In practice, competing opinions are usually identified, and special procedures are used to achieve the required level of consensus. One known approach consists in assigning weights to experts' opinions. In this paper, the correct mathematical basis is formulated for new types on the Kemeny's median, like metric and weighted ones. The problem to find the median for a linear combination of experts' rankings is investigated based on the well-known locally optimal Kemeny's algorithm. It is proposed to investigate the rank aggregation problem based on new types of the median.

This research is funded by RFBR, grants 20-07-00055, 18-07-01087, 18-07-00942.

[1] *Dvoenko S. D., Pshenichny D. O.* Rank Aggregation Based on New Types of the Kemeny's Median // Pattern Recognition and Image Analysis, Pleiades Publishing, 2021. Vol. 31. No. 1.

Вейвлет-модель вариаций геомагнитного поля

Мандрикова Оксана Викторовна[1] oksanam1@mail.ru
Родоманская Анастасия Игоревна[1]⋆ pantina_anastasia@mail.ru
[1]Петропавловск-Камчатский, Институт космофизических исследований и распространения радиоволн ДВО РАН

В работе предложена вейвлет-модель вариаций геомагнитного поля, описывающая его регулярные изменения и иррегулярные особенности в месте регистрации данных. Рассмотрена модификация модели для зоны авроральных широт. На примере данных магнитной станции Абиско (Швеция, координаты 68°21.7′N 18°43.4′E) описан процесс идентификации модели и показано ее применение в задаче определения периодов суббурь и оценки их интенсивности. Численная реализация модели обеспечивает возможность её применения в режиме, близком к реальному времени.

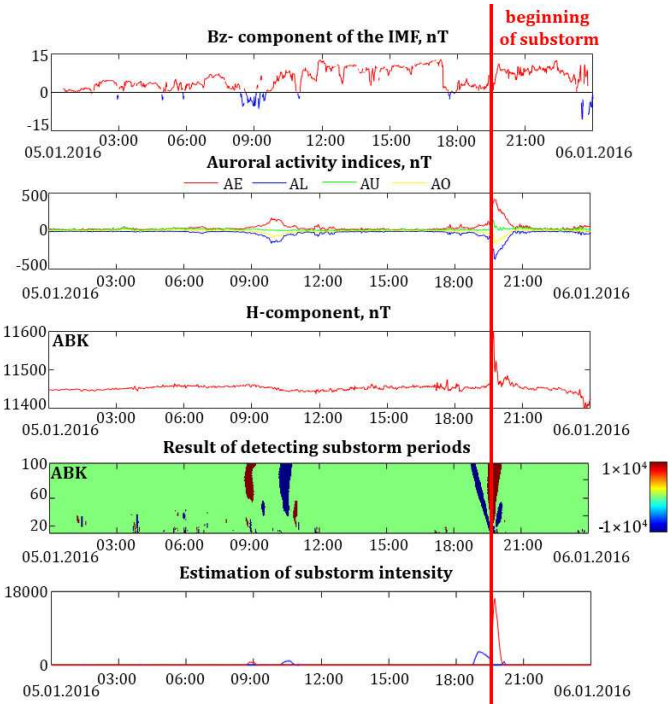


Рис. 1. Результаты применения метода в период суббури 05 января 2016 г

Результаты работы представляют интерес в задачах оценки состояния околоземного космического пространства и прогноза космической погоды.

Wavelet model of geomagnetic field variations

Oksana Mandrikova[1] oksanam1@mail.ru
Anastasia Rodomanskay[1]⋆ pantina_anastasia@mail.ru
[1]Petropavlovsk-Kamchatskiy, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS

The paper proposes a wavelet model of variations in the geomagnetic field, which describes its regular changes and irregular features at the place of data registration.A modification of the model for the auroral zone is considered. On the basis of the data from Abisko station (Sweden, coordinates 68°21.7′N 18°43.4′E), the process of model identification is described and its application is shown in the problem of determining the periods of substorms and assessing their intensity. The numerical implementation of the model provides the possibility of its application in a mode close to real time.

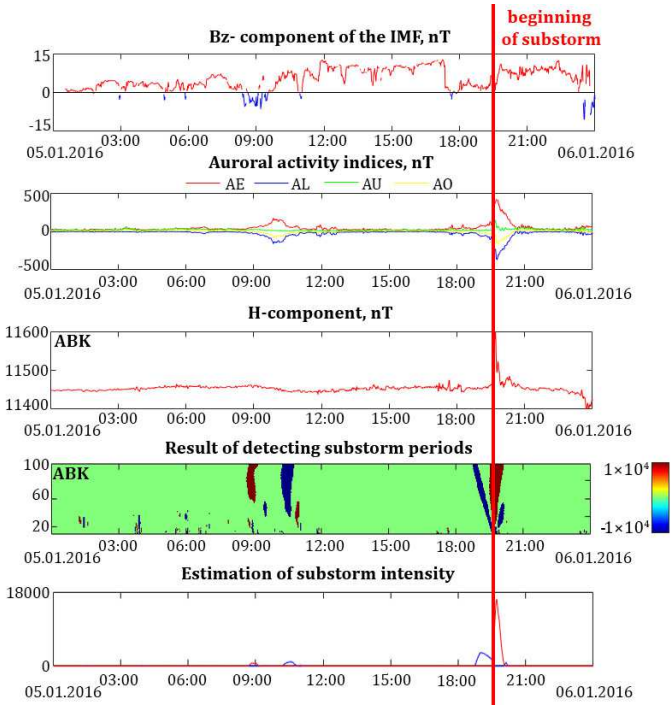


Figure 2. Results of applying the method during the substorm period on January 5, 2016

The results of the work are of interest in the problems of assessing the state of near-earth space and forecasting space weather.

Метод обнаружения аномальных эффектов в сложном сигнале

Геппенер Владимир Владимирович[1] geppener@mail.ru
Мандрикова Богдана Сергеевна[2]⋆ 555bs5@mail.ru

[1]Санкт-Петербург, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[2]Паратунка, Институт космофизических исследований и распространения радиоволн ДВО РАН

Предложен метод обнаружения и идентификации аномальных эффектов в сигнале сложной структуры, основанный на суперпозиции конструкций вейвлет-преобразования. Показано, что метод позволяет идентифицировать

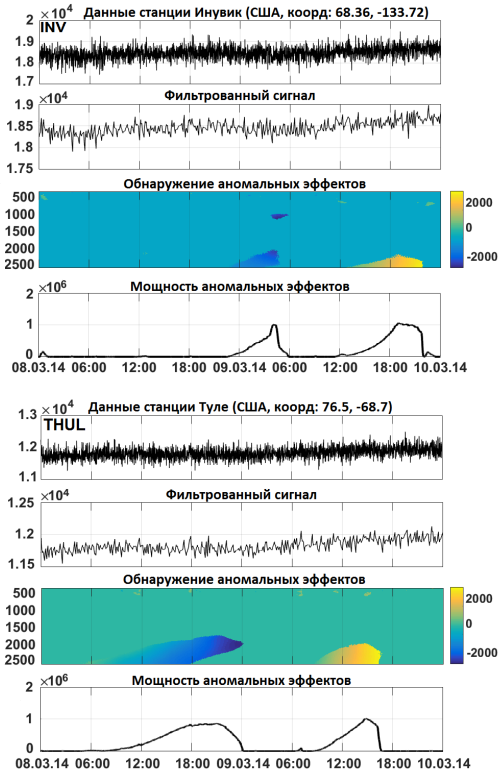


Рис. 1. Результат применения метода

аномальные эффекты разной формы и длительности. Исследуются вопросы вы-

бора аппроксимирующего вейвлет-базиса и пути их оптимизации. На примере данных мировой сети нейтронных мониторов доказана эффективность метода для задачи обнаружения спорадических (аномальных) эффектов в динамике космических лучей. Исследование космических лучей и обнаружение спорадических эффектов представляет интерес в изучении астрофизических процессов, а также в решении многих практических задач, в числе которых мониторинг и прогноз космической погоды, обеспечение радиационной безопасности космонавтов и др. Вследствие сложной нестационарной структуры данных нейтронных мониторов, высокого уровня шума и отсутствия адекватных математических моделей задача эффективного обнаружения и идентификации спорадических эффектов в космических лучах в настоящее время не решена. Предложенный в работе подход, использующий наборы словарей вейвлет-базисов, рассматривается в качестве одного из возможных решений данной проблемы. Пример применения метода представлен на рисунке 1.

Работа выполнена в рамках ГЗ по теме «Динамика физических процессов в активных зонах ближнего космоса и геосфер» (2018-2020) № АААА-А17-117080110043-4. Авторы выражают благодарность институтам, выполняющим поддержку станций нейтронных мониторов, которые использовались в работе.

[1] *Geppener V. V., Mandrikova B. S.* An automated method for detecting sporadic effects in cosmic rays // E3S Web of Conferences, 2020.

Method for detecting anomalous effects in a complex signal

Vladimir Geppener[1] geppener@mail.ru
Bogdana Mandrikova[2]⋆ 555bs5@mail.ru
[1]Saint-Petersburg , Saint Petersburg Electrotechnical University "LETI"
[2]Paratunka, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS

A method for detecting and identifying anomalous effects in a signal of a complex structure, based on the superposition of wavelet transform structures, is proposed. It is shown that the method allows one to identify anomalous effects of various forms and duration. The questions of the choice of the approximating wavelet basis and

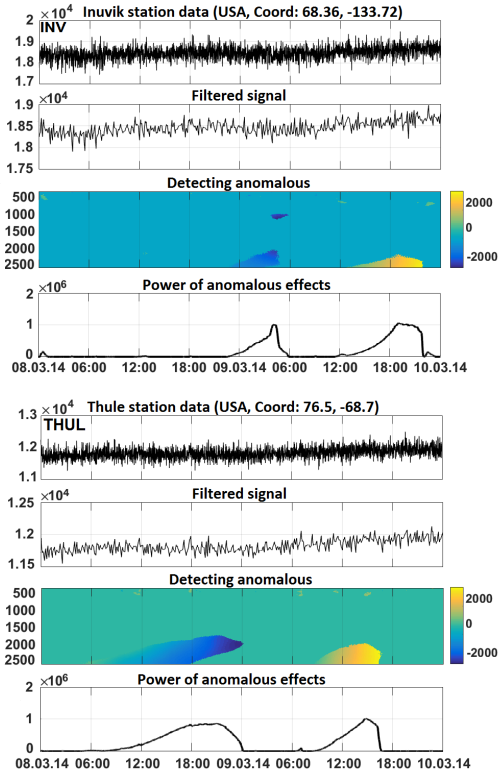


Figure 2. Method result

the ways of their optimization are investigated. Using the data of the world network of neutron monitors as an example, the effectiveness of the method is proved for the problem of detecting sporadic (anomalous) effects in the dynamics of cosmic

rays. The study of cosmic rays and the detection of sporadic effects is of interest in the study of astrophysical processes, as well as in solving many practical problems, including monitoring and forecasting space weather, ensuring the radiation safety of astronauts, etc. In the absence of adequate mathematical models, the problem of effective detection and identification of sporadic effects in cosmic rays has not yet been solved. The proposed approach, using sets of dictionaries of wavelet bases, is considered as one of the possible solutions to this problem. An example of the application of the method is presented in Figure 1.

The work was carried out in the framework of the State Task on the topic "Dynamics of physical processes in the active zones of near space and geospheres" (2018-2020) No. Registration AAAA-A17-117080110043-4. The authors are grateful to the institutions that support the neutron monitor stations that were used in the work.

[1] *Geppener V. V., Mandrikova B. S.* An automated method for detecting sporadic effects in cosmic rays // E3S Web of Conferences, 2020.

Восстановление 3D-модели объектов инфраструктуры на основе использования нейросетевых методов обработки спутниковых изображений

Кошелева Наталья Владимировна[2]⋆ antipova@phystech.edu
Гвоздев Олег Геннадьевич[1,3] gvozdev@miigaik.ru
Козуб Владимир Александрович[1] postbox-kozub@yandex.ru
Мурынин Александр Борисович[1,2] amurynin@bk.ru
Рихтер Андрей Александрович[1] urfin17@yandex.ru

[1]Москва, Научно-исследовательский институт аэрокосмического мониторинга "АЭРОКОСМОС"
[2]Москва, Федеральный исследовательский центр "Информатика и управление" РАН
[3]Москва, ФГБОУ ВО "Московский государственный университет геодезии и картографии"

Рассматривается задача восстановления трехмерной модели объектов инфраструктуры по одному спутниковому изображению без использования метаданных. Предлагаемый метод заключается в последовательной обработке изображения: выявление и классификация объектов хозяйственной инфраструктуры и собственно восстановление моделей объектов. Модели объектов восстанавливаются по растровым областям изображения, полученного в результате работы локального анализа, основанного на регрессионном анализе, методе эквивалентных фигур, линеаризации и поляризации контура. Интегральный анализ нацелен на локализацию объектов, локальный – на извлечение данных, необходимых для определения геометрии объекта и их интерпретацию.

В работе рассматриваются четыре класса областей: крыши, стены, рельсы, опорные столбы. Модель объекта складывается из контуров крыши объекта, крыш его пристроек и надстроек, ортогональных отрезков высот и превышений (при наличии стен). В работе приведены примеры трёхмерных моделей трёх зданий, восстановленных по растрам областей классов стен и крыш. В результате формируется продукт, пригодный для дальнейшей машинной или ручной обработки.

Работа выполнена при поддержке Министерства науки и высшего образования РФ (уникальный идентификатор проекта RFMEFI60719X0312).

[1] *Гвоздев О. Г., Козуб В. А., Кошелева Н. В., Мурынин А. Б., Рихтер А. А.* Построение трехмерных моделей ригидных объектов по спутниковым изображениям высокого пространственного разрешения с использованием сверточных нейронных сетей // Известия РАН. Физика атмосферы и океана, 2020. Т. 56. № 12.

Reconstruction of a 3D model of infrastructure objects based on the usage of neural network methods for processing satellite images

Natalia Kosheleva[1,2] antipova@phystech.edu
Oleg Gvozdev[1,3] gvozdev@miigaik.ru
Vladimir Kozub[1] postbox-kozub@yandex.ru
Aleksander Murynin[1,2] amurynin@bk.ru
Andrey Richter[1] urfin17@yandex.ru

[1]Moscow, AEROCOSMOS Research Institute for Aerospace Monitoring
[2]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[3]Moscow, State University of Geodesy and Cartography

The problem of reconstructing a three-dimensional model of infrastructure objects from one satellite image without using metadata is considered. The proposed method consists in sequential image processing: identification and classification of the economic infrastructure objects and the actual restoration of the objects' models. Object models are reconstructed from raster areas of the image obtained as a result of local analysis based on regression analysis, the method of equivalent figures, linearization and polarization of the contour. Integral analysis is aimed at localizing objects, while local analysis is aimed at extracting data necessary to determine the geometry of an object and interpret it.

The work considers four classes of areas: roofs, walls, rails, support pillars. The object model consists of the contours of the object roof, roofs of its extensions and superstructures, orthogonal segments of heights and elevations (if there are walls). The paper provides examples of three-dimensional models of three buildings, reconstructed from raster areas of class walls and roofs. As a result, a product is formed, suitable for further machine or manual processing.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI60719X0312).

[1] *Gvozdev O. G., Kozub V. A., Kosheleva N. V., Murynin A. B, Richter A. A.* Construction of Three-Dimensional Models of Rigid Objects from Satellite Images of High Spatial Resolution Using Convolutional Neural Networks // Izvestiya, Atmospheric and Oceanic Physics, 2020. Vol. 56. No 12.

Совместная оценка карты рассеивания и атмосферной освещенности с использованием вероятностной гамма-нормальной модели для задачи устранения тумана на изображении

Филин Андрей Игоревич[1]⋆ andrewifilin@gmail.com
Грачева Инесса Александровна[1] gia1509@mail.ru
Копылов Андрей Валериевич[1] andkopylov@gmail.com
[1]Тула, Тульский государственный университет

Распространенной проблемой в системах технического зрения, работающих в широком диапазоне погодных условий и освещения является общее ухудшение видимости объектов сцены. Удаление тумана и мелкодисперсной пыли является актуальной задачей технического зрения, поскольку, во-первых, данная задача играет значительную роль в системах улучшенного зрения, а во-вторых, в большинстве алгоритмов анализа изображений, независимо от того выполняют ли они обработку, обнаружение или распознавание объектов, на вход традиционно поступают значения интенсивности пикселей изображения. Искаженное или слабоконтрастное входное изображение напрямую влияет на точность и эффективность алгоритмов обработки и анализа. Ввиду важности этой проблемы для систем технического зрения, интерес к ней не ослабевает, несмотря на появление в последнее время достаточно большого количества эффективных методов. В то же время растущий размер обрабатываемых изображений и переход от статического анализа изображений к обработке видео предъявляют повышенные требования к вычислительной сложности алгоритмов.

В данной работе предлагается новый вычислительно эффективный метод удаления тумана на изображении, основанный на совместной оценке карты рассеивания и атмосферной освещенности, с использованием вероятностной гамма-нормальной модели [1].

Устойчивая оценка атмосферной освещенности является одной из основных сложностей в рассматриваемой задаче, возникающей при наличии локализованных источников света (фары машин, дорожные фонари). Присутствие локализованных источников света часто нарушает априорные допущения, обычно принимаемые для оценки атмосферного света, что приводит к переэкспонированию участков и потере деталей на результирующем изображении. Для выделения атмосферного света был разработан универсальный метод на основе одноклассовой классификации. Этот метод позволяет создать маску, скорректированную методами морфологической фильтрации, и исключить области, соответствующие локализованным источникам, из оценки атмосферного света. Основным недостатком этого метода является независимая классификация каждого элемента изображения и построение бинарной маски, не позволяющей учесть ореолы вокруг источников света. В данной работе предлагается усовершенствованная версия универсального метода выделения атмосферного света.

Степень принадлежности пикселя к локализованному источнику света определяется расстоянием от цвета пикселя до центра гиперсферы, построенной в процессе обучения в гильбертовом пространстве. Взаимное согласование локальных решений предлагается выполнить с помощью процедуры фильтрации со свойствами переноса структуры, основанной на вероятностной гамма-нормальной модели. Результатом предложенной процедуры будет согласованная вероятностная оценка принадлежности элементов изображения к источникам света.

Поскольку оценка карты рассеивания и атмосферного освещения фактически выполняется на основе одной и той же процедуры, а исходное изображение используется в качестве носителя информации о структуре данных, эти этапы можно объединить, оценив вероятностные отношения между элементами данных только один раз. В сочетании с линейной вычислительной сложностью процедуры оценки скрытого компонента в гамма-нормальной модели по отношению к количеству пикселей изображения это позволяет построить новый алгоритм удаления тумана на изображении с высокой скоростью вычислений.

Для проведения экспериментального исследования разработанного метода удаления тумана на изображении были использованы три группы баз данных: базы данных с естесственным туманом, базы данных с искусственным туманом и базы данных с дополнительными локализованными источниками освещения. Результаты обработки этих баз данных предлагаемым методом были сравнены с другими известными методами удаления тумана на изображении по качеству обработки, с использованием метрик PSNR, SSIM, автоматической оценки качества изображений (Neural Image Assessment, NIMA), и по времени работы алгоритмов. Результаты экспериментов показывают, что предлагаемый метод имеет сопоставимые результаты качества и меньшее время вычислений, чем другие методы удаления тумана.

Работа поддержана грантами РФФИ № 18-07-00942, № 20-07-00441.

[1] *Filin A. I., Gracheva I. A., Kopylov A. V.* Haze removal method based on joint transmission map estimation and atmospheric-light extraction // ACM International Conference Proceedings Series, 2020.

Combined transmission map estimation and atmospheric-light extraction using the probabilistic gamma-normal model for haze removal problem

Andrei Filin[1]⋆ andrewifilin@gmail.com
Inessa Gracheva[1] gia1509@mail.ru
Andrei Kopylov[1] andkopylov@gmail.com
[1]Tula, Tula State University

A common problem in outdoor technical vision systems worked in a wide range of weather and lighting conditions is a general visibility degradation of objects on scene. Haze and fine dust removal is an actual task in technical vision systems, since, firstly, this task plays a significant role in improved vision systems, and secondly, most image analysis algorithms, regardless of whether they perform processing, detection or object recognition traditionally use values of pixels intensity on input image. A distorted or low-contrast input image directly affects the accuracy and efficiency of image processing and analysis algorithms. In view of the importance of this problem for technical vision systems, interest in it does not wane, despite the appearance of a fairly large number of effective methods recently. At the same time, the growing size of processed images and the transition from static image analysis to video processing place high demands on the computational complexity of algorithms.

In this paper, a new computationally efficient method of image haze removal based on a joint transmission map estimation and atmospheric light extraction using a probabilistic gamma-normal model is proposed [1].

Stable estimation of atmospheric light is one of the main difficulties in the problem under consideration, which arises in presence of localized light sources (car headlights, lanterns). Presence of localized light sources often violates a priori assumptions commonly made for atmospheric light estimation that allows to overexposure of areas and loss of detail in resulting image. The universal method of atmospheric light extraction has been developed and based on the one-class classification. This method creates a mask, corrected by morphological filtering methods, exclude areas corresponding to localized sources from atmospheric light estimation. The main disadvantage of this method is independent classification of each image element and construction of binary mask, which take account of halos around light sources. This paper proposes the improved version of the universal method for atmospheric light extraction. Belonging degree of pixel to localized light source is determined by distance from pixel color to center of hypersphere, built in training process in Hilbert space. Mutual adjustment of local solutions is proposed to be performed using the filtering procedure with structure-transferring properties based on the probabilistic gamma-normal model. Similarly, you can find the transmission map estimation. Result of the proposed procedure will be an agreed probabilistic assessment of image elements belonging to light sources.

Since the transmission map and atmospheric light are actually evaluated using the same procedure, and original image is used as a carrier of information about the objects structure, these steps can be combined by evaluating the probabilistic relationships between data items only once. This procedure in combination with the linear computational complexity of the procedure for estimating the latent component in the gamma-normal model in relation to number of image pixels allows constructing a new algorithm for image haze removal with a high computational speed.

Three groups of databases were used to conduct the experimental research of the proposed image haze removal method: databases with natural haze, databases with artificial haze and databases with additional localized light sources. Processing results of these databases by the proposed method were compared with other well-known image haze removal methods in processing quality, using PSNR, SSIM metrics, automatic image quality assessment (Neural Image Assessment, NIMA), and computation time of algorithms. Experimental results show that the proposed method has comparable quality results and less computation time than other haze removal methods.

This research is funded by RFBR, grant 18-07-00942 and grant 20-07-00441.

[1] *Filin A. I., Gracheva I. A., Kopylov A. V.* Haze removal method based on joint transmission map estimation and atmospheric-light extraction // ACM International Conference Proceedings Series, 2020.

Возможности использования деревьев решений в задаче атрибуции публицистических текстов XIX века

Рогов Александр Александрович[1] rogov@petrsu.ru
Москин Николай Дмитриевич[1]⋆ moskin@petrsu.ru
Абрамов Роман Владимирович[2] monset008@gmail.com
Кулаков Кирилл Александрович[1] kulakov@cs.karelia.ru
[1]Петрозаводск, Петрозаводский государственный университет
[2]Санкт-Петербург, Национальный исследовательский университет ИТМО

В статье рассматривается математическая и программная поддержка задачи атрибуции (установления авторства) анонимных текстов [1]. Исследование проводилось на публицистических статях XIX века из журналов «Время» (1861-1863), «Эпоха» (1864-1865) и еженедельника «Гражданин» (1873-1874). Известно, что Ф. М. Достоевский (вместе со своим братом М. М. Достоевским) редактировал и возглавлял эти журналы, поэтому уже давно ведутся исследования на предмет принадлежности его перу данных произведений. Большое количество этих статей опубликовано анонимно, т. е. либо без подписи, либо под псевдонимами. Впрочем, это относится и к статьям, которые исследователи давно приписывали Достоевскому, более или менее основываясь на документальных данных. Текст изучают на разных уровнях: пунктуационном, орфографическом, синтаксическом, лексико-фразеологическом и стилистическом (отметим, что под «авторским стилем» обычно понимают последние три уровня). Исследование проводилось с помощью информационной системы «Смалт» («Статистические методы анализа литературных текстов»), разработанной в Петрозаводском государственном университете [2].

В настоящее время для решения задачи атрибуции используют методы машинного обучения. В статье основное внимание уделено поиску признаков атрибуции с использованием технологий «деревья решений» и «случайный лес». Как известно, одним из достоинств данных технологий является хорошая интерпретация полученных результатов, что является ключевым при их признании специалистами в области филологии. Был исследован метод, основанный на дереве решений, для классификации статей на два класса: «Ф. М. Достоевский» и «другие». В качестве признаков были взяты статистики n-грамм частей речи (последовательностей из n закодированных частей речи). Для определения частотных характеристик текстов применялась информационная система СМАЛТ, а для построения деревьев решений использовалась среда Python 3.6. Используя лишь одну последовательность, удалось достичь результата 89% точности классификации текстов. С помощью полученных результатов было проанализировано влияние выбора глубины дерева, длины n-граммы, размера текста и остальных параметров на конечную точность алгоритма. Кроме того, были выявлены наиболее информативные части речи и их комбинации для данной задачи.

Ф. М. Достоевский хорошо понимал важность влияния сильных позиций текста на читателя, поэтому мог уделять больше внимания внесению правок в начальные и в конечные абзацы текстов чужих статей. Поэтому решая вопросы, связанные с атрибуцией текстов в журналах «Время» и «Эпоха», в качестве одной из задач выделен специальный анализ данных элементов текста. Была рассмотрена совокупность статей Ф. М. Достоевского и других авторов (М. М. Достоевский, Н. Н. Страхов, А. А. Головачев, И. Н. Шилль, А. Григорьев, А. У. Порецкий, Я. П. Полонский), опубликованных в этих журналах в период 1861-1865 гг. В текстах были выделены фрагменты размером 500, 700 и 1000 слов. При этом для увеличения объема выборки использовался шаг для отсчета начала следующего фрагмента: 100, 200 слов и т. п. На основе частеречного распределения фрагментов текстов были построены деревья решений, в узлах которых находятся условия ветвления, основанные на частоте встречаемости той или иной n-граммы. Анализ сильных позиций данных текстов с помощью деревьев решений показывает возможность стилистической правки, которую вносил Ф. М. Достоевский в тексты изначальных авторов.

Лексический спектр является значимой характеристикой для решения задачи определения авторства текстов (например, его использовал Г. Хетсо при исследовании текстов Ф. М. Достоевского). Однако применение деревьев решений требует представления спектра в виде одного числа, которое адекватно отражало бы его структуру. Была рассмотрена аппроксимация лексических спектров (на уровне словаря и на уровне текста) гиперболическими и экспоненциальными кривыми, в результате чего получаются две характеристики для каждой кривой. На материале статей из дореволюционного журнала «Время» (1861-1863) показано, что коэффициенты гиперболической регрессии аппроксимируют данные гораздо лучше, чем коэффициенты экспоненциальной кривой. Построение спектров осуществлялось с помощью информационной системы СМАЛТ.

При решении задачи атрибуции текстов возникает проблема определения авторского стиля писателя, который создал меньшее количество текстов (как количественно, так и по общему объему слов) в сравнении с другими авторами из числа анализируемых. Были рассмотрены возможные варианты решения этой проблемы на примере определения стиля Аполлона Григорьева. В качестве метода построения ансамбля классификаторов в работе использовался бэггинг (bootsrtap aggregating). В результате расчетов выяснилось, что относительная частота биграммы «частица-прилагательное» больше 6,5 является отличительной особенностью публицистического стиля Аполлона Григорьева. Также было проведено исследование статьи «Стихотворения А. С. Хомякова», которое подтверждает ранее сделанный вывод о том, что нет оснований считать ее принадлежащей Аполлону Григорьеву.

Применение других методов машинного обучения (рекуррентные сети и параллельные рекурретные сети) показали результаты, сравнимые с деревьями ре-

шений. Эфективнее оказалась модель трансформера, однако она требует большого объема данных для обучения.

Работа поддержана грантом РФФИ № 18-012-90026.

[1] *Абрамов Р. В.* Применение дерева решений для атрибуции текстов // Процессы управления и устойчивость, 2020. Т. 7. № 1. С. 183–187.

[2] *Кулаков К. А., Рогов А. А., Москин Н. Д.* Программная поддержка в решении задачи атрибуции текстов // Программная инженерия, 2019. Т. 10. № 5. С. 234–240.

Possibilities of using decision trees in the problem of attribution of publicistic texts of the XIX century

Alexander Rogov[1] rogov@petrsu.ru
Nikolai Moskin[1]⋆ moskin@petrsu.ru
Roman Abramov[2] monset008@gmail.com
Kirill Kulakov[1] kulakov@cs.karelia.ru
[1]Petrozavodsk, Petrozavodsk State University
[2]Saint Petersburg, ITMO University

The article discusses the mathematical and software support for the problem of attribution (establishing authorship) of anonymous texts [1]. The research was carried out on publicistic articles of the 19th century from the magazines ≪Time≫ (1861-1863), ≪Epoch≫ (1864-1865) and the weekly ≪Citizen≫ (1873-1874). It is known that F. M. Dostoevsky (together with his brother M. M. Dostoevsky) edited and headed these magazines, so research has long been conducted on the subject of belonging to his pen of these works. A large number of these articles were published anonymously, i.e. either without a signature or under pseudonyms. However, it also applies to articles that researchers have long attributed to Dostoevsky, more or less based on documentary data. The text is studied at different levels: punctuation, orthographic, syntactic, lexical-phraseological and stylistic (note that the last three levels are usually understood as ≪author's style≫). The research was carried out using the information system ≪Smalt≫ (≪Statistical Methods for the Analysis of Literary Texts≫), developed at the Petrozavodsk State University [2].

Currently, machine learning methods are used to solve the attribution problem. The article focuses on the search for attribution features using the ≪decision trees≫ and ≪random forest≫ technologies. It is known that one of the advantages of this technologies is a good interpretation of the obtained results, that is key for their acceptance by philologists. A method based on a decision tree was investigated for classifying articles into two classes: ≪F. M. Dostoevsky≫ and ≪others≫. Statistics of n-grams of parts of speech (sequences of n encoded parts of speech) were taken as features. The SMALT information system was used to determine the frequency characteristics of texts, and the Python 3.6 environment was used to build decision trees. Using only one sequence, it was possible to achieve the result of 89% accuracy of text classification. Using the obtained results, the influence of the choice of tree depth, n-gram length, text size and other parameters on the final accuracy of the algorithm was analyzed. In addition, the most informative parts of speech and their combinations for this task were identified.

F. M. Dostoevsky well understood the importance of the influence of strong positions of the text on the reader, so he could pay more attention to making corrections in the initial and final paragraphs of the texts of other people's articles. Therefore, when solving issues related to the text attribution in the ≪Time≫ and ≪Epoch≫ magazines, a special analysis of these text elements is highlighted as one

of the tasks. The authors reviewed a set of articles by F. M. Dostoevsky and other authors (M. M. Dostoevsky, N. N. Strakhov, A. A. Golovachev, I. N. Shill, A. Grigoriev, A. U. Poretsky, Ya. P. Polonsky) published in these journals in the period 1861-1865. Fragments of 500, 700 and 1000 words were highlighted in the texts. At the same time to increase the sample size a step was used to count the beginning of the next fragment: 100, 200 words, etc. On the basis of the part-of-speech distribution of text fragments decision trees were built, at the nodes of which there are branching conditions based on the frequency of occurrence of a particular n-gram. The analysis of the strong positions of these texts using decision trees shows the possibility of stylistic corrections made by F. M. Dostoevsky into the texts of the original authors.

The lexical spectrum is a significant characteristic for solving the problem of determining the authorship of texts (for example, it was used by G. Kjetsaa in the study of texts by F. M. Dostoevsky). However, the application of decision trees requires representing the spectrum as a single number that would adequately reflect its structure. The approximation of lexical spectrum (at the dictionary level and at the text level) by hyperbolic and exponential curves was considered, as a result of which two characteristics are obtained for each curve. Based on the articles from the pre-revolutionary journal ≪Time≫ (1861-1863), it is shown that the coefficients of hyperbolic regression approximate the data much better than the coefficients of the exponential curve. The spectrums were constructed using the SMALT information system.

When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises. We consider possible solutions to this problem by the example of determining the style of Apollon Grigoriev. As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*). As a result of calculations we can assume that the relative frequency of the ≪particle-adjective≫ bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. There also was a study of the article ≪Poems by A. S. Khomyakov≫, which confirms the previously conclusion that there is no reason to consider it as belonging to Apollon Grigoriev.

Application of other machine learning methods (recurrent networks and parallel recurrent networks) has shown results comparable to decision trees. The transformer model turned out to be more effective, but it requires a large amount of data for training.

This research is funded by RFBR, grant 18-012-90026.

[1] *Abramov R. V.* Decision tree application for text attribution // Control Processes and Stability, 2020. Vol. 7. No 1. Pp. 183–187.

[2] *Kulakov K., Rogov A., Moskin N.* Software support in solving the problem of text attribution // Software engineering, 2019. Vol. 10. No 5. Pp. 234–240.

Комбинированный метод учета эпсилон-ограничений для решения задач распределения нагрузки с помощью дифференциальной эволюции

Становов Владимир Вадимович[1,2]⋆ vladimirstanovov@yandex.ru
Ахмедова Шахназ Агасувар кызы[1,2] shahnaz@inbox.ru
Семенкин Евгений Станиславович[1,2] eugenesemenkin@yandex.ru
[1]Красноярск, Сибирский институт прикладного системного анализа
[2]Красноярск, Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева

За последние десятилетия было предложено несколько эффективных методов учета ограничений в области эволюционных вычислений, и метод ε-ограничений считается современным методом как для однокритериальной, так и для многокритериальной оптимизации. Тем не менее, было предпринято мало попыток улучшить этот метод в применении к алгоритму дифференциальной эволюции. В этой работе [1] предлагается несколько новых методов учета ограничений, где уровни ε устанавливаются для каждого ограничения, при этом комбинация пригодности и нарушения ограничений используется для определения недопустимых решений. Предложенные подходы демонстрируют качество работы, превосходящее таковое в сравнении с другими подходами в смысле доли допустимых решений в многомерных пространствах поиска, а также в смысле сходимости к глобальному оптимуму. Эксперименты проведены с использованием тестовых функций Конгресса по Эволюционным Вычислениям 2017 (Congress on Evolutionary Computation, CEC), а также набора задач распределения нагрузки в электросетях.

Набор задач распределения нагрузки (Economic Load Dispatch, ELD) был изначально предложен в рамках соревнования по тестированию эволюционных алгоритмов для реальных задч оптимизации в рамках Конгресса по Эволюционным вычислениям. В рамках соревнования, эти задачи были представлены как задачи безусловной оптимизации, в то время как их формулировка содержала от 2 до 7 ограничений, которые были добавлены в функцию пригодности со статическими коэффициентами. В этой работе эти задачи рассматривались как задачи условной оптимизации, их характеристики представлены в Таблице 1.

Целью задач распределения нагрузки (ELD_6-ELD_140) было минимизировать затраты топлива на генерацию электроэнергии в течении определенного периода работы, в динамических задачах распределения (DED_5 and DED_10) цель состояла в минимизации затрат в течении 24-часового периода с изменяющимися требованиями по мощности, а для задач гидротермического планирования (HS_1 to HS_3), которые также относятся к динамическим задачам, цель состояла в минимизации генерации электроэнергии тепловыми электростанциями. Набор ограничений содержал ограничения по балансу мощности,

Таблица 1. Свойства задач распределения нагрузки

Задача	Размерность	Число ограничений
DED_5	120	4
DED_10	216	4
ELD_6	6	4
ELD_13	13	2
ELD_15	15	4
ELD_40	40	2
ELD_140	140	4
HS_1	96	6
HS_2	96	7
HS_3	96	6

ограничения генераторов, ограничения угла рампы, а также зоны запрещенного функционирования, т.е. особые наборы параметров, где функционирование запрещено в силу ограничений объекта (например, нестабильность компонентов или вибрации) для задач ELD и DED, и для задач HS ограничения включали нелинейные соотношения, такие как каскадная природу гидросистемы, задержки переноса воды, и временные связи между последующими планами, тем самым усложняя задачи.

Эксперименты по учету индивидуальных штрафов независимо друг от друга в отличие от их комбинирования в одно значения показали, что такой подход может существенно улучшить качество работы. Более того, важно, что поддержание нескольких уровней ε не приводит к существенному увеличению вычислительной сложности или требованиям к памяти, если число ограничений сохраняется в пределах разумного. Подход с использованием длины вектора, EC_L, похож по качеству работы на EC_P, так как представляет собой другой способ комбинации нескольких ограничений.

Разработка новых универсальных методов учета ограничений имеет особую важность не только для области эволюционных вычислений, как для однокритериальных, так и многокритериальных задач, но и для реальных приложений. Эта работа не только предлагает несколько методов учета ε-ограничений, среди которых EC_{IFL} оказывается самым эффективным, но также показывает некоторые важные свойства эволюционных алгоритмов в применении к задачам условной оптимизации, включая тот факт, что способ разделения популяции существенно влияет на качество работы. Разработанные подходы могут быть применены к любой другой модификации дифференциальной эволюции или эволюционному алгоритму, который способен использовать принцип превосходства ε-допустимых решений, заменяя классическую селекцию или замещения и вводя лексикографический порядок. Дальнейшие исследования в этом

направлении могут включать, но не ограничены: автоматическим масштабированием ограничений и пригодностей для определения лексикографического порядка, определением оптимального значения скорости снижения индекса θ, поддержанием двух уровней ε для каждого ограничения-равенства путем разграничения между нарушениями в каждом направлении и включением методов EC в современные многокритериальные эволюционные алгоритмы.

Работа была поддержана Министерством науки и высшего образования Российской Федерации в рамках государственного контракта № FEFE-2020-0013.

[1] *Stanovov V., Akhmedova S., Semenkin E.* Combined fitness–violation epsilon constraint handling for differential evolution // Soft Computing, 2020. Vol. 24. Pp. 7063–7079.

Combined Epsilon-Constraint Handling Method for Solving Economic Load Dispatch Problems with Differential Evolution

Vladimir Stanovov[1]⋆ vladimirstanovov@yandex.ru
Shakhnaz Akhmedova[1] shahnaz@inbox.ru
Eugene Semenkin[1] eugenesemenkin@yandex.ru
[1]Krasnoyarsk, Siberian Institute of Applied System Analysis, Reshetnev Siberian State University of Science and Technology

Over recent decades, several efficient constraint-handling methods have been proposed in the area of evolutionary computation, and the ε constraint method is considered as a state-of-the-art method for both single and multiobjective optimization. Still, very few attempts have been made to improve this method when applied to the differential evolution algorithm. This study [1] proposes several novel constraint-handling methods following similar ideas, where the ε level is defined based on the current violation in the population, individual ε levels are maintained for every constraint, and a combination of fitness and constraint violation is used for determining infeasible solutions. The proposed approaches demonstrates superior performance compared to other approaches in terms of the feasibility rate in high-dimensional search spaces, as well as convergence to global optima. The experiments are performed using the Congress on Evolutionary Computation (CEC) 2017 constrained suite benchmark functions and a set of Economic Load Dispatch problems in power grids.

The set of Economic Load Dispatch (ELD) problems was originally proposed within the Congress on Evolutionary Computation Competition on Testing Evolutionary Algorithms on Real World Optimization Problems. In the competition, these problems were presented as bound constraint problems, while their formulation contained from 2 to 7 constraints, which were included in fitness function with static coefficients. In this study these problems are considered as constraint optimization problems, having the properties described in Table 1.

The goal of static Economic Load Dispatch problems (ELD_6-ELD_140) was to minimize the fuel cost for power generation for a specific period of operation, for Dynamic Economic Dispatch (DED_5 and DED_10) problems the goal was to minimize cost during 24 hours of operation with varying power demands, and for Hydrothermal Scheduling problems (HS_1 to HS_3), which were also dynamic problems, the goal was to minimize the power generation by thermal units. The set of constraints includes power balance constraints, generator constraints, ramp rate limits and also prohibited operating zones, i.e. specific certain parameter sets where the operation is prohibited due to limitations of the units (for example, components instability or vibrations) for ELD and DED, and for HS the constraints included nonlinear relationships, including the cascaded nature of the hydraulic system, water carry delays and the time link between consecutive schedules, adding to the complexity of the problems.

Table 2. Properties of ELD problems

Problem	Dimension	Number of constraints
DED_5	120	4
DED_10	216	4
ELD_6	6	4
ELD_13	13	2
ELD_15	15	4
ELD_40	40	2
ELD_140	140	4
HS_1	96	6
HS_2	96	7
HS_3	96	6

The experiments with handling individual penalties independently rather than combining them into a single value have shown that such approach can significantly improve the performance. Moreover, it is important that maintaining several ε levels does not introduce significant computational overheads in terms of computational complexity or memory consumption, if the number of constraints is limited by a reasonable value. The vector length approach, EC_L, is similar in performance to EC_P, because it represents another way of combining several constraints.

The development of novel universal constraint handling techniques has major importance not only for the area of evolutionary computation, for both single-objective and multi-objective optimization, but also for real-world applications. This study not only proposes several ε-constraint methods, with EC_{IFL} being the most efficient, but also reveals important properties of evolutionary algorithms when applied to constrained optimization problems, such as the fact that the population division method significantly influences the performance. The developed approaches could be applied to any other modification of differential evolution, or evolutionary algorithms which is capable of utilizing the superiority of ε-feasible solutions, replacing the classical selection or replacement step, and introducing the novel lexicographic ordering procedure. Further studies in this direction may include but are not limited to: automatic scaling of constraints and fitness value to define lexicographic ordering, identifying the optimal θ index decrease rate, maintaining two ε levels for each equality constraint by distinguishing between violation in every direction, and incorporation of EC methods into modern Multi-Objective Evolutionary Algorithms.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation within limits of state contract No. FEFE-2020-0013.

[1] *Stanovov V., Akhmedova S., Semenkin E.* Combined fitness–violation epsilon constraint handling for differential evolution // Soft Computing, 2020. Vol. 24. Pp. 7063–7079.

Бионические алгоритмы для для оптимизации расписания в промышленности

Семенкина Ольга Евгеньевна[1]⋆ semenkinaolga@gmail.com
Попов Евгений Александрович[1] epopov@bmail.ru
Семенкин Евгений Станиславович[1] eugenesemenkin@yandex.ru
[1]Красноярск, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева

Оперативное планирование производства является управлением системой почти в реальном времени, поэтому время, затрачиваемое на вычисления, имеет большое значение. Эффективность принятия решений зависит от скорости вычислений, а значит и от многих факторов, таких как временной горизонт, количество единиц оборудования, количество сотрудников, разнообразие технологических процессов и так далее. Существует несколько способов преодолеть проблему комбинаторного взрыва, например, использование проблемно-ориентированных эвристик. Однако гораздо более интересный подход - применение существующих алгоритмов глобальной оптимизации. Задача составления расписания, а также другие хорошо известные комбинаторные проблемы, такие как задача коммивояжера (TSP), могут быть сформулированы как задача составления расписания проекта с условием ограничения на ресурсы (RCPSP).

Авторы предлагают учитывать все детали производственного процесса с помощью имитационной модели, а также использовать алгоритмы оптимизации для задачи верхнего уровня, что позволяет оптимизировать некоторые параметры системы. Подобная иерархическая структура задачи гарантирует, что все решения в пространстве поиска в любом случае допустимы, и, в то же время, значительно снижает размерность проблемы и количество ограничений. В этой статье мы рассматриваем задачу планирования, преобразованную в задачу иерархической оптимизации, содержащую комбинаторную задачу упорядочения (TSP) и вложенную RCPSP, которая была заменена моделью. Подобная модификация позволяет упростить задачу, облегчая применение методов оптимизации при оперативном планировании производства.

В качестве входных данных модели мы рассматриваем список операций для всех партий в определенном порядке, которые необходимо обработать. Модуль планирования выполняет операции в заданном порядке, помещая их в ближайшую свободную точку с доступными ресурсами. Такой подход гарантирует соблюдение всех ограничений на ресурсы. Модуль планирования реализует всю бизнес-логику и может построить допустимое расписание в соответствии с ограничениями производственного процесса. В этой статье мы сравниваем два подхода, а именно перестановку приоритетов операций и перестановку самих партий, что соответствующим образом формирует входные данные модели.

Бионические алгоритмы, такие как генетический алгоритм (GA), алгоритм умных капель (IWDs) и алгоритм муравьиных колоний (ACO), показывают

конкурентоспособные результаты при решении TSP, поэтому мы также использовали их в этой работе. В дополнение к ним используется эвристика Лин-Кернигана (LKH). Все бионические алгоритмы имеют множество параметров, которые необходимо выбирать, и это их большой недостаток. Кроме того, лучшие настройки под конкретную задачу могут отличаться, и их невозможно спрогнозировать заранее. При решении практических задач обычно нет возможности тратить ресурсы на определение наилучших настроек алгоритма. А исправить этот недостаток можно с помощью метода самоконфигурирования для управления параметрами «на лету».

Эффективность алгоритмов оценивалась при решении шести задач, которые были сгенерированы с использованием псевдослучайных чисел. Сравнение алгоритмов и представлений решений производилось для одного и того же количества вычислений целевой функции, а также результаты были усреднены по 50 запускам. Результаты, усредненные по всем 6 задачам, показаны на рисунке 1, где более темным цветом обозначено представление в виде перестановки приоритетов операций. Результаты самоконфигурируемых алгоритмов, а именно самоконфигурируеммого GA (ScGA) и самоконфигурируеммого ACO (ScACO), выделены отдельно синим цветом.

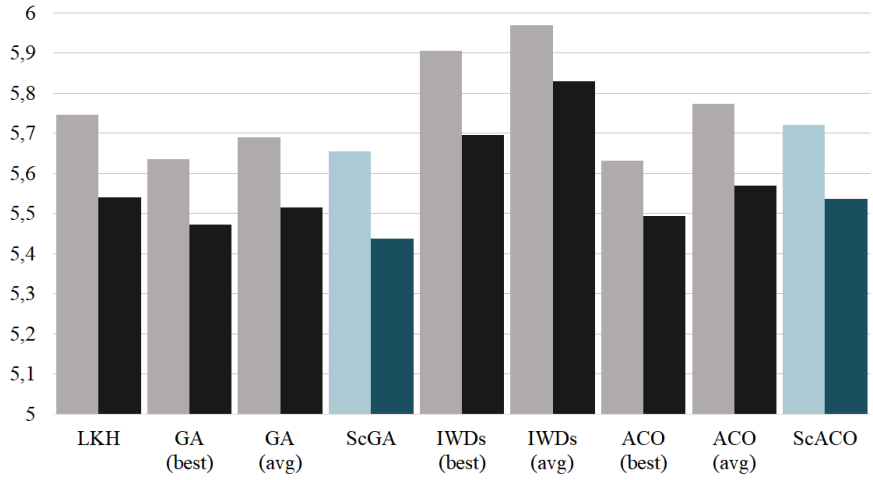


Рис. 1. Усредненное по 6 задачам сравнение алгоритмов

Как видно на рисунке, представление решения в виде порядка приоритетов операций дает лучшие результаты. Скорее всего, это связано с тем, что после выбора порядка для партии (в случае представления с порядком партий) для каждой из них расписание строится жадным образом, то есть выбирают-

ся первые доступные ресурсы. Эта стратегия может привести к прерывистому расписанию для некоторого ресурса, когда в его расписании появляется пустое пространство между двумя проставленными операциями, хотя расписание для конкретной партии будет плотным. Постановка задачи с упорядочением приоритетов значительно увеличивает размерность задачи, усложняя ее решение, но в то же время расширяет пространство поиска, не ограничивая его сцепленными цепочками операций.

Результаты исследования показывают, что метод самоконфигурирования является эффективной модификацией и имеет существенное преимущество, так как пользователю не нужно выбирать параметры алгоритма, но он по-прежнему может получать хорошие результаты. Более того, для решения сложных проблем использование разных настроек на разных этапах процесса поиска может быть хорошей стратегией.

Работа выполнена при поддержке Министерства науки и высшего образования РФ, проект № FEFE-2020-0013

[1] *Semenkina O. Popov E.* Nature-inspired algorithms for a scheduling problem inoperational planning // IOP Conf. Series: Materials Science and Engineering, 2020. Vol. 734.

Nature-inspired algorithms for scheduling optimization in industry

Olga Semenkina[1]⋆ semenkinaolga@gmail.com
Eugene Popov[1] epopov@bmail.ru
Eugene Semenkin[1] eugenesemenkin@yandex.ru
[1]Krasnoyarsk, Siberian State University of Science and Technologies

Operational planning in manufacturing systems is a kind of near real-time control where simulation runtime is an important aspect. The efficiency of short-term decision-making hangs on the speed of online simulation, which in turn depends on many factors such as time horizon, amount of equipment, number of employees, variety of technological processes, and so on. There are several ways to overcome the problem of combinatorial explosion, for example, the use of problem-oriented heuristics. However, a much more interesting approach is to find how to apply different existing global optimization algorithms to the problem. The scheduling problem as well as other well-known combinatorial problems such as the travelling salesman problem (TSP) can be formulated as resource-constrained project scheduling problems (RCPSP). The problem consists in finding a schedule with minimal makespan by assigning a machine tool, an employee, and a start time for all activities of a project.

The authors propose to consider all the details of a production process using a simulation model and also use optimization algorithms for a top-level problem that allows some system parameters to be found that are optimal in some sense. An earlier investigation shows that the hierarchical structure of the problem makes it possible to be sure that all solutions in a search space are feasible in any case and at the same time significantly reduces the problem dimension as well as the number of constraints. In this paper, we consider scheduling problem transformed into a hierarchical optimization problem containing a combinatorial ordering problem (regarded as TSP) and nested RCPSP replaced by a model with some rules. Modification of this kind makes it possible to simplify the problem in the case of an existing production model and allows the dimension and number of constraints to be reduced making the application of optimization methods in operational production planning easier.

As an input of the model, we consider a list of operations (activities) for all lots in a certain order that needs to be processed. The scheduler module takes operations in the given order and puts them at the first accessible point when resources are available. This approach guarantees that all restrictions on the resources are met. A start point of operation is selected as the nearest free point with an available resource that uses a greedy strategy. The scheduler module realized all the business logic and can construct a valid schedule according to the constraints of the production process. In this paper, we compare two approaches, namely permutation of activity priorities and permutation of lots which form the model input accordingly.

Nature-inspired algorithms such as genetic algorithms (GA), intelligent water drops algorithm (IWDs), and ant colony optimization (ACO) show competitive results on the TSP which is why we also used them in this work. In addition to them, Lin-Kernighan Heuristics (LKH) is used. All bionic algorithms have many parameters that must be chosen, and this is their great disadvantage. Besides, the best settings on a particular task may differ, and it is impossible to forecast them in advance. Real-world problems do not usually allow resources to be spent on determining the best algorithm settings. The way to fix this disadvantage is through a self-configuring method for the parameter control "on a fly".

Algorithm performance was compared by solving six tasks that were generated using pseudo-random numbers. A comparison of the algorithms as well as a comparison of the solution representations were performed on the same objective function calculation amount and were also averaged over 50 runs. The results of the experiments are shown in Figure 1, where the darker color indicates the solution representation as a permutation of activity priorities. Figure 1 shows the result for all algorithm averaged by 6 tasks. The results of the self-configuring algorithms, namely Self-Configuring GA (ScGA) and Self-Configuring ACO (ScACO), are highlighted separately in blue.

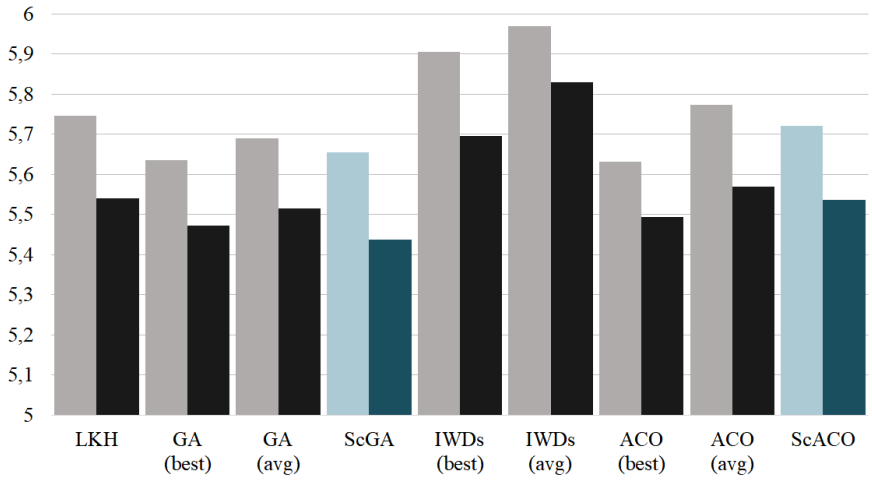


Figure 2. Algorithm comparison averaged on 6 tasks

As can be seen from the figure, the statement of the problem through the search of the activity priority order shows better results. Most likely this is because after choosing the lot order (in the case of the lot order problem) for each of them the schedule is built in a greedy manner that is, the first available resources are selected.

This strategy can lead to a discontinuous schedule for the current resource, where a blank space appears in the schedule of this resource between two already set activities, which no other activity can fit into, although the schedule of a specific lot is dense. The problem statement with priority ordering significantly increases the dimension of the problem, complicating its solution, but at the same time expands the search space without limiting it to concatenated chains of activities.

The results of the investigation show that the self-configuring method is an effective standard one modification and has a significant advantage in that the user does not need to select the algorithm settings, but is still able to receive competitive results. Moreover, for solving complex problems, using different settings at different stages of the search process can be a good strategy.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation within limits of state contract FEFE-2020-0013.

[1] *Semenkina O. Popov E.* Nature-inspired algorithms for a scheduling problem inoperational planning // IOP Conf. Series: Materials Science and Engineering, 2020. Vol. 734.

Знаниеориентированные модели маршрутизации многих коммивояжеров

Германчук Мария Сергеевна[1] m.german4uk@ya.ru
Козлова Маргарита Геннадьевна[1] art-inf@mail.ru
Лукьяненко Владимир Андреевич[1]⋆ art-inf@ya.ru
[1]Симферополь, Крымский федеральный университет им. В. И. Вернадского

Рассматриваются знаниеориентированные модели, задачи и алгоритмы построения маршрутов в сложных сетях многими агентами-коммивояжерами. Формализация приводит к моделям псевдобулевой дискретной оптимизации с ограничениями, учитывающими специфику построения маршрутов. Разработка приближенных алгоритмов выбора маршрутов в сложных сетях связана с учетом знаний о свойствах структуры сети, ее сложности, наличие ограничений, предписаний, условий достижимости, числа агентов-коммивояжеров. Показано, что решение задач маршрутизации может базироваться на применении многоагентного подхода в сочетании с кластеризацией исходной задачи и метаэвристиках. Разнообразие алгоритмов также связано с наличием априорных знаний о решении или структуре сети, прецедентным характером знаний и требованиями к точности решения. Рационально использование, как точных, так и приближенных алгоритмов и их композиций. Заметим, что задачи прикладной маршрутизации возникают в сочетании с другими известными задачами: задача о ранце, распределение ресурсов, кластеризации, максимального разреза, покрытия и т. п. Специфика задач маршрутизации в сложных сетях, в отличие от классической теории графов, связана с рядом уникальных задач: о нахождении метрических характеристик сложных сетей; поиск минимального (максимального) среднего пути в сети; коэффициентов кластеризации; изучения информационных потоков в сети; выявление критичных мест в сети; определения кластеров; выявление блоков, компонент, мостов, точек сочленения (перемычек). Многоагентные системы с роевым интеллектом используются для решения сложных задач дискретной оптимизации, которые нельзя эффективно решать классическими алгоритмами. Агентная модель для сложной сети задачи типа многих коммивояжеров становится интеллектуализированной системой, определяющей эвристические алгоритмы поиска оптимального решения реактивными агентами (следующих заложенным в них правилам). В работе применяются композиции алгоритмов: модификация генетического алгоритма, муравьиный, роевой (пчелиной колонии), имитации отжига. Предложен и реализован обобщенный алгоритм, в котором исходной сети ставится в соответствие более простая сеть (сеть облета). Алгоритм инспирирован рядом актуальных прикладных задач: задачей планирования многодневных туристических маршрутов на инфраструктурной сети достопримечательностей Крыма и задачей доставки ресурсов агентами-коммивояжерами по территории Ялты в условиях чрезвычайных ситуаций (ЧС).

В многоагентной системе (МАС) сочетаются задачи выбора решения; управления; распределения ресурсов; синтеза сети (вершин-источников ресурсов); устойчивости сети в зависимости от удаления вершины, дуги или некоторого маршрута; кластеризации сети в зависимости от изменяющихся условий; обмена информацией между агентами; потоковые задачи; задачи прокладки кратчайших путей и замкнутых маршрутов. Методология разработки алгоритма решения задач маршрутизации может быть основана на формировании по исходной сложной сети более простой (относительно реализации алгоритмов маршрутизации) по своей структуре сети.

Численный эксперимент проведен для задачи маршрутизации по карте ГИС для городской инфраструктуры. Реализованы алгоритмы кластеризации, в которых первоначально пройденные маршруты уточняются с помощью алгоритмов 2-opt, имитации отжига и других метаэвристик. Построение рациональных решений на сетях большой размерности реализуется по схеме:

1) решается задача распределения сети между коммивояжерами с помощью кластеризации;
2) решаются задачи коммивояжера на каждом кластере с помощью метаэвристик;
3) в зависимости от полученного результата уточняются границы используемых кластеров.

Дальнейшие исследования связаны с обучением агентов-коммивояжеров, их автономностью и организацией обмена прецедентной информацией между агентами (системами управления).

Knowledgeoriented routing models for many traveling salesmen

Mariia Germanchuk m.german4uk@ya.ru
Margarita Kozlova art-inf@mail.ru
Vladimir Lukianenko⋆ art-inf@ya.ru
Simferopol, V. I. Vernadsky Crimean Federal University

Knowledge-oriented models, tasks, and algorithms for constructing routes in complex networks by many traveling salesmen are considered. Formalization leads to models of pseudo-Boolean discrete optimization with restrictions that take into account the specifics of route construction. The development of approximate algorithms for selecting routes in complex networks involves taking into account knowledge about the properties of the network structure, its complexity, restrictions, requirements, reachability conditions, and the number of sales agents. It is shown that the solution of routing problems can be based on the application of a multi-agent approach in combination with clustering of the original problem and metaheuristics. The variety of algorithms is also related to the presence of a priori knowledge about the solution or network structure, the case-based nature of knowledge, and the requirements for the accuracy of the solution. Rational use of both exact and approximate algorithms and their compositions. Note that applied routing problems occur in combination with other well-known problems: the knapsack problem, resource allocation, clustering, maximum cut, coverage, and so on. The specifics of routing problems in complex networks, in contrast to classical graph theory, are associated with a number of unique problems: finding metric characteristics of complex networks; finding the minimum (maximum) average path in the network; clustering coefficients; studying information flows in the network; identifying critical places in the network; determining clusters; identifying blocks, components, bridges, and junction points (jumpers).

Multi-agent systems with swarm intelligence are used to solve complex discrete optimization problems that cannot be effectively solved by classical algorithms. The agent model for a complex network of problems like many traveling salesmen becomes an intellectualized system that defines heuristic algorithms for finding the optimal solution by reactive agents (following the rules laid down in them). The paper uses several algorithms: modification of the genetic algorithm, ant, swarm (bee colony), simulated annealing. A generalized algorithm is proposed and implemented, in which a simpler network (a flyover network) is matched to the source network. The algorithm is inspired by a number of actual applied tasks: the task of planning multi-day tourist routes on the infrastructure network of attractions in the Crimea and the task of delivering resources by traveling salesmen on the territory of Yalta in emergency situations. A multi-agent system (MAC) combines the tasks of solution selection; management; resource allocation; network synthesis (vertexes-resource sources); network stability depending on the removal of a vertex, arc, or some route; network clustering depending on changing conditions; informa-

tion exchange between agents; streaming tasks; tasks of laying shortest paths and closed routes. The methodology for developing an algorithm for solving routing problems can be based on the formation of a simpler network structure (relative to the implementation of routing algorithms) based on the original complex network.

A numerical experiment was performed for the problem of routing on a GIS map for urban infrastructure. Clustering algorithms are implemented, in which the initially traversed routes are refined using 2-opt algorithms, simulated annealing, and other metaheuristics. The construction of rational solutions on large-dimensional networks is implemented according to the scheme:

1) the problem of network distribution between salesmen is solved using clustering;
2) traveling salesman problems are solved on each cluster using metaheuristics;
3) depending on the result obtained, the boundaries of the clusters used are specified.

Further research is related to the training of sales agents, their autonomy, and the organization of the exchange of case information between agents (management systems).

Классификация асимметричных задач коммивояжера по квантилям распределения сложности индивидуальных задач

Жукова Галина Николаевна[1] gzhukova@hse.ru
Ульянов Михаил Васильевич[2,3]⋆ muljanov@mail.ru
[1]Москва, Национальный исследовательский университет «Высшая школа экономики»
[2]Москва, ИПУ РАН им. В. А. Трапезникова
[3]Москва, МГУ им. М. В. Ломоносова

В аспекте проблемы прогнозирования временных характеристик для задач большой вычислительной сложности в докладе рассматривается вариант классификации асимметричных задач коммивояжера по квантилям логнормального распределения сложности индивидуальных задач. Далее под сложностью индивидуальной задачи коммивояжера понимается число вершин поискового дерева решений, порожденных классической реализацией метода ветвей и границ, которая предложена Литлом, Мерти, Суини и Кэролом в [1]. В работе [2] на основании статистической обработки экспериментальных данных — результатов измерений сложности индивидуальных задач — было показано, что логнормальное распределение удовлетворительно аппроксимирует распределение значений сложности при фиксированной размерности задачи. Исследование проводилось для пула в 100 000 сгенерированных матриц асимметричной задачи коммивояжера для каждой размерности от 20 до 49. Генерация осуществлялась стандартным генератором псевдослучайных чисел с равномерным распределением.

Пусть $C(A)$ - сложность индивидуальной задачи, заданной матрицей A размерности n (очевидно, что сложность коррелирована с временем решения задачи). Пусть C_n - сложность, как случайная величина при фиксированной размерности n, т.е. $C(A)$ является реализацией C_n. В [2] на основе статистического анализа экспериментальных данных показано, что случайная величина, представляющая собой линейное преобразование логарифма сложности $L(\ln C_n) = = \frac{\ln C_n - b}{n}$ имеет нормальное распределение (и, следовательно C_n имеет логнормальное распределение), и параметры нормального распределения $L(\ln C_n)$ не зависят от размерности задачи в диапазоне 20-49.

Полученный результат позволяет унифицировано ввести классификацию задач коммивояжера по сложностям индивидуальных задач, при принятии гипотезы об унификации, состоящей в том, что распределение $L(\ln C_n)$ не меняет своего типа и вида зависимости параметров от размерности задачи и при дальнейшем увеличении размерности задачи. Такую классификацию мы предлагаем ввести на основании квантилей нормального распределения, опираясь на формальный подход 80/20 для распределения $L(\ln C_n)$. При этом переход в реальный диапазон квантилей 10% и 90% сложностей индивидуальных задач

осуществляется по формуле

$$q_p^{C_n} \approx e^{(an+h)q_p^{N(0,1)}+dn+f}, \quad q_p^{N(0,1)} = \Phi^{-1}(p), \tag{1}$$

где $p = 10\%, 90\%$, $\Phi^{-1}(p)$ — обратная функция к функции распределения стандартного нормального закона.

Вычисляя при фиксированном n значения 10% и 90% квантилей, получаем классификацию задач по сложности:

— простые задачи — задачи, попадающие в интервал от 0 до 10%-го квантиля нормального распределения $L(\ln C_n)$;
— средние задачи — задачи, попадающие в диапазон от квантиля 10% до квантиля 90% нормального распределения $L(\ln C_n)$
— сложные задачи — задачи, имеющие сложность более 90%-го квантиля нормального распределения $L(\ln C_n)$

Отметим, что предложенная классификация инвариантна по размерности задачи при принятии гипотезы об унификации.

Приведем численный пример для размерности 40. Квантили 10% и 90% стандартного нормального распределения равны -1.28 и 1.28 соответственно; для равномерного распределения элементов матрицы стоимостей экспериментально получены оценки параметров формулы (2): $a = 0.018$, $d = 0.77$, $f = 0.7$, $h = 0.77$. По формуле (2) при $n = 40$ получаем значения квантилей 10% и 90% сложности задачи 453 и 7216 соответственно, следовательно

— простые задачи — задачи с индивидуальной сложностью не более чем в 453 порожденных вершин поискового дерева решений;
— средние задачи — задачи с индивидуальной сложностью между 453 и 7216 порожденных вершин поискового дерева решений;
— сложные задачи — задачи с индивидуальной сложностью более чем в 7216 порожденных вершин поискового дерева решений.

Заметим, что максимально в эксперименте со 100 000 матриц наблюдалось 598893 порожденных вершин.

Работа поддержана грантом РФФИ № 20-58-S52006.

[1] *Little J., Murty K., Sweeney D., Karel C.* An algorithm for the traveling salesman problem // Operations Research, 1963. Vol. 11. Pp. 972–989.

[2] *Головешкин В. А., Жукова Г. Н., Ульянов М. В., Фомичев М. И.* Вероятностный прогноз сложности индивидуальных задач коммивояжера на основе идентификации распределения сложности по экспериментальным данным // Автоматика и телемеханика, 2018. № 7. С. 149–166.

Classification of asymmetric traveling salesman problems by quantiles of the distribution of the complexity of individual problems

Galina Zhukova[1] gzhukova@hse.ru
Mikhail Ulyanov[2,3]⋆ muljanov@mail.ru
[1]Moscow, National Research University Higher School of Economics
[2]Moscow, I.P. V. A. Trapeznikova
[3]Moscow, Moscow State University M. V. Lomonosov

In the aspect of the problem of forecasting working time characteristics for problems of high computational complexity, the report considers a variant of the classification of asymmetric traveling salesman problems by quantiles of the lognormal distribution of the complexity of individual problems. In what follows, the complexity of an individual traveling salesman problem is understood as the number of vertices of the search decision tree generated by the classical implementation of the branch and bound method proposed by Little, Murty, Sweeney, and Carol in [**?**]. In the work [2], on the basis of statistical processing of experimental data — the results of measuring the complexity of individual problems — it was shown that the lognormal distribution satisfactorily approximates the distribution of complexity values for a fixed dimension of the problem. The study was conducted for a pool of 100,000 generated matrices of the asymmetric traveling salesman problem for each dimension from 20 to 49. The generation was carried out by a standard pseudo-random number generator with a uniform distribution.

Let $C(A)$ be the complexity of an individual problem, given by a matrix A of dimension n (it is obvious that the complexity is correlated with the time of solving the problem). Let C_n be the complexity as a random variable for a fixed dimension n, i.e. $C(A)$ is an implementation of C_n. In [2], based on the statistical analysis of experimental data, it is shown that a random variable representing a linear transformation of the logarithm of complexity $L(\ln C_n) = \frac{\ln C_n - b}{n}$ has a normal distribution (and, therefore, C_n has a lognormal distribution), and the parameters of the normal distribution $L(\ln C_n)$ do not depend on the dimension of the problem in the range 20-49.

The result obtained allows us to unify the classification of traveling salesman problems according to the complexity of individual problems, when accepting the hypothesis of unification, which is that the distribution $L(\ln C_n)$ does not change its type and the form of the dependence of parameters on the dimension of the problem and with further increase in the dimension of the problem. We propose to introduce such a classification on the basis of the quantiles of the normal distribution, relying on the formal 80/20 approach for the distribution $L(\ln C_n)$. In this case, the transition to the real range of quantiles 10 % and 90 % of the complexity of individual tasks

is carried out according to the formula

$$q_p^{C_n} \approx e^{(an+h)q_p^{N(0,1)}+dn+f}, \quad q_p^{N(0,1)} = \Phi^{-1}(p), \tag{2}$$

where $p = 10\%, 90\%$, $\Phi^{-1}(p)$ is the inverse function to the distribution function of the standard normal law.

Calculating for a fixed n the values of 10 % and 90 % quantiles, we obtain a classification of problems by complexity:

— simple problems - problems that fall into the interval from 0 to 10 % - quantile of the normal distribution $L(\ln C_n)$;
— medium problems - problems that fall in the range from quantile 10 % to quantile 90 % of the normal distribution $L(\ln C_n)$
— complex problems - problems with complexity over 90 % - quantile of the normal distribution $L(\ln C_n)$

Note that the proposed classification is invariant with respect to the dimension of the problem when the unification hypothesis is accepted.

Let's give a numerical example for dimension 40. Quantiles 10 % and 90 % of the standard normal distribution are equal to -1.28 and 1.28, respectively; for the uniform distribution of the elements of the cost matrix, estimates of the parameters of the formula (2) were obtained experimentally: $a = 0.018$, $d = 0.77$, $f = 0.7$, $h = 0.77$. By formula (2) for $n = 40$ we obtain the values of the quantiles 10 % and 90 % of the complexity of the problem 453 and 7216, respectively, hence

— simple problems - problems with an individual complexity of no more than 453 generated vertices of the search decision tree;
— medium problems - problems with individual complexity between 453 and 7216 generated vertices of the search decision tree;
— complex problems - problems with an individual complexity of more than 7216 generated vertices of the search decision tree.

Note that the maximum in the experiment with 100,000 matrices was 598893 generated vertices.

This research is funded by RFBR, grant 20-58-S52006.

[1] *Little J., Murty K., Sweeney D., Karel C.* An algorithm for the traveling salesman problem // Operations Research, 1963. Vol. 11. Pp. 972–989.

[2] *Goloveshkin V. A., Zhukova G. N., Ulyanov M. V., Fomichev M. I.* Probabilistic Prediction of the Complexity of Traveling Salesman Problems Based on Approximating the Complexity Distribution from Experimental Data // Autom. Remote Control, 2018. Vol. 79. No 7. Pp. 1296–1310.

Эвристическая ребалансировка на основе приоритетов в задаче управления данными с вероятностными ограничениями

Токарева Виктория Андреевна[1] victoria.tokareva@kit.edu

[1]Карлсруэ, Технологический институт Карлсруэ

В настоящее время хранение и обработка больших объемов разнородных данных вызывает большой интерес со стороны бизнеса и исследователей. Наиболее актуальным ответом на этот вызов являются т.н. озёра данных (англ. data lakes) [1], которые позволяют наладить обработку потоков данных для отдельных проектов, внутри организаций или международных научных экспериментов [2, 3]. Однако в ситуации проектирования платформ, поддерживающих мета- и междисциплинарные исследования, а так же при проектировании платформ для свободного доступа к данным [4], требуется более высокий уровень абстракции. Обслуживание и балансировка потоков данных в системах такого типа является актуальной проблемой, требующей активных исследований со стороны научного соотщества. В данном докладе модель такой системы исследуется на примере разработки системы аггрегированного сбора данных для экспериментов астрофизики частиц [5]. Произведено исследование предметной области и составлена математическая модель происходящих в ней конкуретных процессов. Аггрегация данных в системе описана как задача смешанного цеха с вероятностными ограничениями. Поставлена задача динамической балансировки пользовательских заявок к системе с целью оптимизации временных затрат. Предлагается эвристический подход к распределению приоритетов заданий в очереди, учитывающий вероятностные зависимости и огнаничения, имеющие место быть в рассматриваемой системе.

Работа поддержана грантом Объединения немецких научно-исследовательских центров им. Гельмгольца (Helmholtz Society) № HRSF-0027.

[1] *Beheshti A., et al.* Intelligent Knowledge Lakes: The Age of Artificial Intelligence and Big Data // Web Information Systems Engineering, 2020. Pp. 24–34.

[2] *Barberis D., et al.* The ATLAS EventIndex: data flow and inclusion of other metadata // J. Phys. Conf. Ser., 2016. Vol. 762. No 1.

[3] *Ambroz, L.* Performance studies of CMS workflows using Big Data technologies // Bologna: Bologne University, 2016.

[4] *Mons B., et al.* Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. // Information Services & Use, 2017. Pp. 49–56.

[5] *Tokareva V., et al.* Data aggregation platform for experiments of astroparticle physics // Proceedings of the 4th International Workshop on Data Life Cycle in Physics,2020.

Priority-based rebalancing heuristic for a mixed shop problem with probabilistic constraints

Victoria Tokareva[1] victoria.tokareva@kit.edu

[1]Karlsruhe, Karlsruhe Institute of Technology

At the present time, storing and processing large amounts of heterogeneous data evokes strong interest among business representatives and researchers. The most relevant approach to this challenge employs data lakes [1] that allow to set up data-flow processing within individual projects, organizations or scientific experiments [2, 3]. However, thinking at a higher level of abstraction is required when designing platforms that support meta-analysis and interdisciplinary research, as well as free access to data within the open-science [4] paradigm. Data stream maintenance and balancing in systems of this type is a topical issue that requires active research from the scientific community.

In this talk, modeling of such a system is investigated via an example of a data-aggregation system for astroparticle physics experiments [5]. A study of the subject area was carried out and a mathematical model of the competitive processes taking place in it was designed. Data aggregation in the system is described in a form of a mixed-shop problem with probabilistic constraints. The problem of dynamic balancing of user requests for the system is set in order to optimize time costs. A heuristic approach to the task priority distribution in the queue is proposed, that takes into account the relevant probabilistic dependencies and constraints.

This research is funded by the Helmholtz Society, grant HRSF-0027.

[1] *Beheshti A., et al.* Intelligent Knowledge Lakes: The Age of Artificial Intelligence and Big Data // Web Information Systems Engineering, 2020. Pp. 24–34.

[2] *Barberis D., et al.* The ATLAS EventIndex: data flow and inclusion of other metadata // J. Phys. Conf. Ser., 2016. Vol. 762. No 1.

[3] *Ambroz, L.* Performance studies of CMS workflows using Big Data technologies // Bologna: Bologne University, 2016.

[4] *Mons B., et al.* Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. // Information Services & Use, 2017. Pp. 49–56.

[5] *Tokareva V., et al.* Data aggregation platform for experiments of astroparticle physics // Proceedings of the 4th International Workshop on Data Life Cycle in Physics,2020.

Модели координации задач планирования закупки сырья и выпуска конечной продукции промышленного предприятия

Некрасов Иван Васильевич[1] ivannekr@mail.ru
Правдивец Николай Александрович[1]⋆ pravdivets@ipu.ru
[1]Москва, ФГБУН Институт Проблем Управления имени В. А. Трапезникова РАН

Современные подходы к управлению предприятием требуют сквозной интеграции всех бизнес-процессов в единую информационную модель [1]. Примечательно, что сама по себе информационная составляющая в данном вопросе играет второстепенную роль – основной экономический эффект достигается за счет функциональной согласованности деятельности служб предприятия на всех уровнях [2] – в частности, за счет координации целевых показателей смежных служб, моделей их расчета и методов их достижения. С точки зрения планирования производственного процесса, указанное взаимодействие целесообразно формализовать в виде трех взаимосвязанных контуров управления, изображенных на рис. 1.

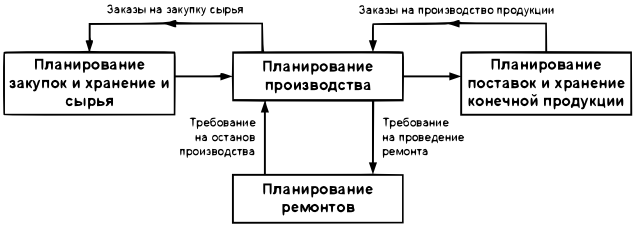


Рис. 1. Подзадачи планирования на предприятии и контуры их взаимодействия

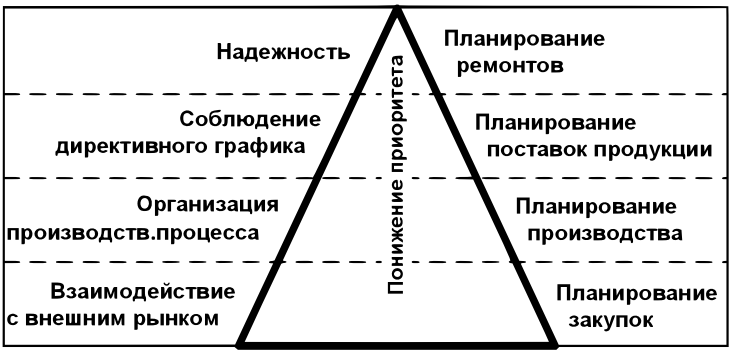


Рис. 2. Пример назначения приоритетов задач планирования (на примере опасного производства с гос. заказом)

Полная модель планирования производства должна включать ограничения по всем перечисленным подзадачам, что значительно увеличивает размерность оптимизационной задачи и усложняет алгоритм ее решения. В настоящее время функционирование систем планирования на предприятиях, как правило, разделено по направлениям [3] и ведется независимо для каждого блока рис. 1. Согласованность задач планирования обеспечивается «директивным способом» за счет жестких ограничений, накладываемых на переменные менее приоритетных задач по результатам вычисления на более критичных участках. Приоритеты задач специфичны для каждой отрасли. В частном случае, для особо опасного производства [4] с жестким производственным планом (как пример – атомная промышленность с государственным заказом), иерархия приоритетов и задач планирования имеет вид:

При рассмотрении предприятий других отраслей распределение приоритетов может кардинально отличаться. С точки зрения теории оптимизации представленный подход эквивалентен разбиению общей задачи поиска глобального оптимального плана всего предприятия на совокупность локальных подзадач оптимизации [5] деятельности отдельных участков (направлений) производства, что не гарантирует нахождение глобального оптимума.

Настоящая работа является частью проекта по разработке постановок совместного согласованного решения задач планирования смежных участков/служб предприятия. Взаимный учет целей и ограничений задач планирования приближает постановку задачи к полной модели рис. 1 и обеспечивает лучшее приближение оптимального глобального плана предприятия. Например, эффект согласованного планирования основного производства и процессов обслуживания и ремонта оборудования показан в работе [6]. В настоящей статье представлен аналогичный подход для взаимного учета планов основного производства и поступления закупленного сырья. Проанализированы методы расширения стандартных моделей планирования производственного процесса дополнительными ограничениями на основе предопределенных жестких графиков поступления сырья. В развитие подхода предложено ввести в указанные жесткие ограничения дополнительные переменные, позволяющие свободно модифицировать график закупок совместно с расписанием основного производства. Практическим эфектом предложенных решений является дополнительная экономия за счет гибкой подстройки производственного и логистического процессов предприятия под такие меняющиеся внешние факторы, как стоимость сырья, его доступность на внешнем рынке, переменное время доставки и т.п.

Работа частично поддержана грантом РФФИ № 18-07-00656 А.

[1] *ANSI/ISA–95.00.03–2005.* Enterprise-Control System Integration Part 3: Activity Models of Manufacturing Operations Management // American National Standard, 2005. Pp. 104.

[2] *Шульц Т., Некрасов И. В., Лежнин Д. В.* Обзор модели стандартной архитектуры и компонентов “Industry 4.0” // Автоматизация в промышленности, 2018. №10. С. 39–46.

[3] *Данилина М. Г.* Основы планирования на предприятии: курс лекций для студентов экономических специальностей, бакалавров и магистров по направлениям «Экономика», «Менеджмент» и «Торговое дело» // Москва: МИИТ, 2012. С. 86.

[4] *116-ФЗ* «О промышленной безопасности опасных производственных объектов» // Федеральный закон Российской Федерации.

[5] *Ковалев М. М.* Дискретная оптимизация (целочисленное программирование) // Москва: Едиториал УРСС, 2003. С. 78–98.

[6] *Nekrasov I.* Coordinated Production and Maintenance Scheduling for an Industrial Enterprise // Управление развитием крупномасштабных систем MLSD, 2019. С. 934–935.

Coordination Models for Purchasing and Production Scheduling Processes of an Industrial Enterprise

Ivan Nekrasov[1] ivannekr@mail.ru
Nikolay Pravdivets[1]⋆ pravdivets@ipu.ru
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

Modern approaches to enterprise management require end-to-end integration of all business processes into a single information model [1]. The information component in this matter plays a secondary role. The main economic effect is achieved due to the functional coherence of of the enterprise services at all levels [2], in particular through the coordination of target indicators of related services, models for their calculation and methods of achieving them. From the point of view of planning the production process, it is advisable to formalize this interaction in the form of three interconnected control loops shown in Fig. 1.

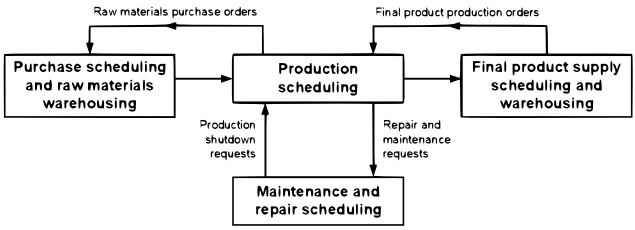


Figure 3. Subtasks of planning in the enterprise and the contours of their interaction

The complete production planning model should include constraints on all of the listed subtasks, which significantly increases the dimension of the optimization problem and complicates the algorithm for its solution. At present, the functioning of planning systems at enterprises is usually divided into directions [3] and is conducted independently for each block of Fig. 1. The coherence of planning tasks is ensured in a "directive way" with strict constraints on the variables of lower priority tasks based on the results of calculations in more critical tasks. Task priorities are specific to each industry. In a particular case, for especially hazardous production [4] with a rigid production plan (as an example, the state demand nuclear industry), the hierarchy of priorities and planning tasks are presented in fig. 4.

When considering enterprises in other industries, the priorities can be radically different. From the point of view of optimization theory, the presented approach is equivalent to splitting the general problem of finding a global optimal plan for the entire enterprise into a set of local optimization subtasks [5] of the activities of individual production areas (directions), which does not guarantee finding the global optimum.

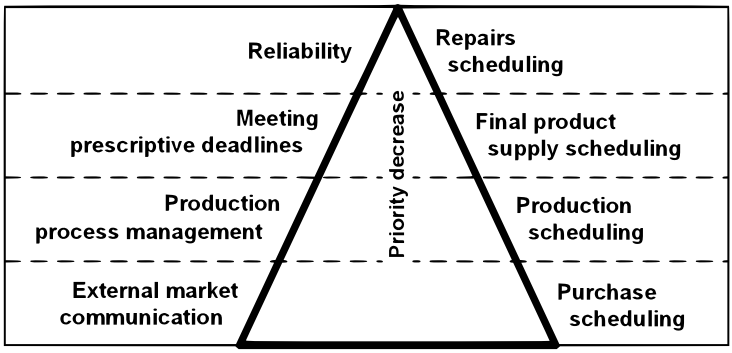


Figure 4. An example of the assignment of priorities for planning tasks (on the example of a state demand hazardous production)

This work is a part of a project for the development of a coordinated solution of planning problems for adjacent areas/services of the enterprise. Relative consideration of the goals and limitations of planning tasks brings the problem statement closer to the complete model of Fig. 1 and provides a better approximation of the optimal global enterprise plan. For example, the effect of coordinated planning of main production and equipment maintenance and repair processes is shown in the work [6]. This article presents a similar approach for the mutual accounting of plans for the main production and the receipt of purchased raw materials. Methods for extending the standard models for planning the production process with additional restrictions based on predefined rigid schedules of arriving raw materials are analyzed. In the development of the approach, it is proposed to introduce additional variables into the specified rigid restrictions, which make it possible to freely modify the procurement plan together with the schedule of the main production. The practical effect of the proposed solutions appears as additional savings due to flexible adjustment of the production and logistics processes of the enterprise according to changing external factors, such as the cost of raw materials, their availability in the external market, variable delivery times, etc.

This work was partially supported by the RFBR (project 18-07-00656 A).

[1] *ANSI/ISA–95.00.03–2005.* Enterprise-Control System Integration Part 3: Activity Models of Manufacturing Operations Management // American National Standard, 2005. Pp. 104.

[2] *Shults T., Nekrasov I. V., Lezhnin D. V.* Overview of the Model of Standard Architecture and Components “Industry 4.0” // Industrial automation, 2018. Vol. 10. Pp. 39–46.

[3] *Danilina M. G.* Fundamentals of planning in an enterprise: a course of lectures for students of economic specialties, bachelors and masters in Economics, Management and Trade // Moscow: MIIT, 2012. Pp. 86.

[4] *Federal law 116* "On industrial safety of hazardous production facilities" // Federal Law of the Russian Federation.

[5] *Kovalyov M. M.* Discrete optimization (integer programming) // Moscow: Editorial URSS, 2003. Pp. 78–98

[6] *Nekrasov I.* Coordinated Production and Maintenance Scheduling for an Industrial Enterprise // Materials of the XII International Conference MLSD, 2019. Pp. 934–935.

Задача о биназначениях в приложении к проблеме воднотранспортного обслуживания островных и городских агломераций

Федосенко Юрий Семенович[1] fds1707@mail.ru
Хандурин Дмитрий Константинович[1] kaf_isuit@vsuwt.ru
Шеянов Анатолий Владимирович[1]⋆ asheyanov@ya.ru

[1]Нижний Новгород, Волжский государственный университет водного транспорта

Имеются совокупность скоростных пассажирских судов $I = \{1, 2, \ldots, n\}$ и два множества маршрутов $P = \{p_1, p_2, \ldots, p_n\}$ и $Q = \{q_1, q_2, \ldots, q_n\}$. Каждое судно должно быть назначено для перевозок пассажиров на один из маршрутов множества P и на один из маршрутов множества Q; на каждый маршрут должно быть назначено только одно судно.

Полагаются заданными $(n \times n)$-матрицы $A = \{a_{ij}\}$ и $B = \{b_{ij}\}$ численных оценок, где a_{ij} — оценка выполнения перевозок судном i по маршруту p_j, b_{ij} — оценка выполнения тем же судном перевозок по маршруту q_j, $i = \overline{1, n}$, $j = \overline{1, n}$.

Введем следующие обозначения: $\Pi_1 = \{\pi_1(i), I\}$ — совокупность назначений судов на маршруты из множества P, $\Pi_2 = \{\pi_2(i), I\}$ — совокупность назначений судов на маршруты из множества Q. Как назначение Π_1, так и назначение Π_2 представляет собой взаимно однозначное отображение множества I в себя: равенство $\pi_1(i) = j$ означает назначение судна i на маршрут p_j; аналогично равенство $\pi_2(i) = j$ означает назначение судна i на маршрут q_j.

Биназначениями [1] именуем пары вида $\langle \pi_1(i), \pi_2(i) \rangle$; считается, что при реализации такого биназначения каждое судно i из множества I, начиная от момента времени 0, выполняет сначала перевозку пассажиров по маршруту с номером $\pi_1(i)$, после чего немедленно приступает к выполнению маршрута с номером $\pi_2(i)$.

В общем виде задача о биназначениях (ЗБН) с минимаксным критерием, обобщающая классическую задачу о назначениях [2], записывается в виде

$$\min_{\pi_1, \pi_2} \left(\max_{\alpha} [a_{\alpha\pi_1(\alpha)} + b_{\alpha\pi_1(\alpha)}] \right) \tag{1}$$

Если матрицами A и B установлены длительности выполнения перевозок по соответствующим маршрутам судами, то в результате решения задачи (1), определится биназначение, обеспечивающее минимальность общей продолжительности выполнения всего комплекса перевозок по маршрутам $\{p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_n\}$.

Пусть i — не превышающая n натуральная константа, W_1, W_2 — произвольные i-элементные подмножества I.

Через $Z(i, W_1, W_2)$ обозначим подзадачу задачи (1), в которой между судами множества I следует распределить маршруты с нижними индексами (номерами) из подмножеств W_1 и W_2; при этом каждое судно должно быть назначено

только на один маршрут из множества P с нижним индексом, входящим в подмножество W_1, и только один маршрут из множества Q с индексом, входящим в подмножество W_2. Оптимальное значение критерия задачи (1) обозначим H_{opt}.

Согласно концепции динамического программирования состояние процесса парного распределения маршрутов множеств P и Q между судами множества I на шаге i однозначно определяется тройкой (i, W_1, W_2), где i – не превышающая n натуральная константа, позволяющая выполнить назначение судов с номерами $j, k = 1, 2, \ldots, i$, и W_1, W_2 произвольные i-элементные подмножества множеств индексов маршрутов P, Q соответственно, доступные для распределения.

Оптимальное значение критерия в задаче $Z(i, W_1, W_2)$ обозначим $H(i, W_1, W_2)$. Как очевидно, $H(i, W_1, W_2)$ — функция Беллмана для задачи (1), причем

$$H(1, \{j\}, \{k\}) = a_{1j} + b_{1k};\ j, k \in N. \tag{2}$$

Для решения задачи (1) запишем следующие соотношения динамического программирования

$$\begin{gathered} H(i, W_1, W_2) = \min(\max_{\alpha,\beta}[(a_{i\alpha} + b_{i\beta}), \\ H(i-1, W_1 \setminus \{\alpha\}, W_2 \setminus \{\beta\})]); \end{gathered} \tag{3}$$

$$\begin{gathered} H_{opt}(n, I, I) = \min(\max_{\alpha,\beta}[(a_{n\alpha} + b_{n\beta}), \\ H(n-1, I \setminus \{\alpha\}, I \setminus \{\beta\})]), \end{gathered} \tag{4}$$

где (α, β) — произвольные пары индексов из множества $W_1 \times W_2$.

Выполнение реализующего по этим соотношениям алгоритма, обозначаемого далее DP, начинается с определения величин $H(1, \{j\}, \{k\})$ для всех одноэлементных множеств W_1 и W_2.

Далее последовательно в порядке возрастания параметра i для всех возможных наборов W_1 и W_2 по формуле (3) определяются значения функции Беллмана $H(i, W_1, W_2)$, $i = \overline{1, n}$.

Определяемое по соотношению (4) значение $H(n, I, I)$ представляет собой оптимальное в задаче (1) значение критерия H_{opt}.

В процессе выполнения алгоритма DP для каждой тройки (i, W_1, W_2) значений аргументов функции Беллмана следует фиксировать пару (α, β), на которой реализуется минимум правой части соотношений (3), (4). Это позволит после отыскания оптимального значения критерия H_{opt} однозначно определить соответствующее ему биназначение.

Соотношения (2)—(4) предполагают реализацию схемы прямого счета динамического программирования, без учета состояний, недостижимых из начального.

Сложность алгоритма DP определяется числом вычисляемых значений функции Беллмана и, как очевидно, определяется величиной $O(4^n)$.

Полученные в результате экспериментов данные по оценке быстродействия алгоритма DP на практически значимых значениях размерности (n =10–13) демонстрируют его практическую значимость для рассматриваемого типа прикладных задач.

Перспективы дальнейших исследований состоят в том, чтобы модифицировать предложенную модель на более широкий набор прикладных задач, в том числе требующих многокритериальные постановки.

[1] *Федосенко Ю. С., Хандурин Д. К.* Модель и алгоритмы синтеза биназначений // Системы управления и информационные технологии, 2020. Т. 4. № 82.

[2] *Votaw D. F., Orden A.* The personnel assignment problem // In Symposium on Linear Inequalities and Programming. Scientific Computation of Optimum Programs. Project SCOOP, 1952. No 10. Pp. 155–163.

Bi-assignment problem application to the problem of water transport services for island and urban agglomerations

Yuriy Fedosenko[1] fds1707@mail.ru
Dmitriy Khandurin[1] kaf_isuit@vsuwt.ru
Anatoliy Sheyanov[1]⋆ asheyanov@ya.ru
[1]Nizhny Novgorod, Volga State University of Water Transport

There are a set of high-speed passenger ships $I = \{1, 2, \ldots, n\}$ and two sets of routes $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$. Each ship must be assigned to carry passengers on one of the routes from the set P and one of the routes from the set Q; only one ship must be assigned to each one of the routes.

The $(n \times n)$-matrices $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ of numerical estimates are assumed to be given, where a_{ij} is the evaluation of the transportation by ship i along the route p_j, while b_{ij} is the same along the route q_j, $i = \overline{1,n}$, $j = \overline{1,n}$.

Let us introduce the following notation: $\Pi_1 = \{\pi_1(i), I\}$ is the set of assignments of ships to routes from P, $\Pi_2 = \{\pi_2(i), I\}$ is the set of assignments of ships to routes from Q. Both the assignment Π_1 and the assignment Π_2 are a bijection of the set I to itself. The equality $\pi_1(i) = j$ means the appointment of ship i to route p_j; similarly, the equality $\pi_2(i) = j$ means the appointment of ship i to route q_j.

We denote pairs of the form $\langle \pi_1(i), \pi_2(i) \rangle$ as bi-assignment [1]; it is supposed that when implementing such bi-assignment, each ship i from I, first starting from time 0, carries out the transportation of passengers along the route number $\pi_1(i)$, and then immediately proceeds to route number $\pi_2(i)$.

In a general form, Bi-assignment Problem (BAP) with a minimax criterion, which is is a generalization of Assignment Problem (AP) [2], is written as follows:

$$\min_{\pi_1, \pi_2} \left(\max_{\alpha} [a_{\alpha\pi_1(\alpha)} + b_{\alpha\pi_1(\alpha)}]\right) \tag{5}$$

If the transportation duration by ships along corresponding routes is determined by matrices A and B, then result of solving problem(1) will be bi-assignment, ensuring the minimum total duration of the entire complex of transportation along the routes $\{p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_n\}$.

Let i be a natural constant not exceeding n, W_1, W_2 be arbitrary i-element subsets of I.

By $Z(i, W_1, W_2)$ we denote the subproblem of the problem (1), in which among the ships of the set I one should distribute routes with lower indices (numbers) from the subsets W_1 and W_2; in this case, each ship should be assigned to only one route from the set P with a subscript included in the subset W_1, and to only one route from the set Q with the index included in the subset W_2. The optimal value of the criterion of the problem (1) is denoted by H_{opt}.

According to the concept of dynamic programming, the state of the process of the pair distribution of the routes from the sets P, Q between the ships of the set I at

step i is uniquely determined by the triple (i, W_1, W_2), where i is a natural constant not exceeding n, restricting the assignment to ships with numbers $j, k = 1, 2, \ldots, i$, and W_1, W_2 arbitrary i-element subsets of the sets of indices of the routes P, Q respectively, available for distribution.

The optimal criterion value in the problem $Z(i, W_1, W_2)$ is denoted by $H(i, W_1, W_2)$.

$H(i, W_1, W_2)$ is the Bellman function for problem (1), and

$$H(1, \{j\}, \{k\}) = a_{1j} + b_{1k};\ j, k \in N. \tag{6}$$

For solving problem (1), we write the following recurrent relations of dynamic programming

$$H(i, W_1, W_2) = \min(\max_{\alpha,\beta}[(a_{i\alpha} + b_{i\beta}), H(i-1, W_1 \setminus \{\alpha\}, W_2 \setminus \{\beta\})]); \tag{7}$$

$$H_{opt}(n, I, I) = \min(\max_{\alpha,\beta}[(a_{n\alpha} + b_{n\beta}), H(n-1, I \setminus \{\alpha\}, I \setminus \{\beta\})]), \tag{8}$$

where (α, β) are arbitrary pairs of indices from the set $W_1 \times W_2$.

The implementation of the computational algorithm along these relations, denoted below by DP, begins with the determination of the quantities $H(1, \{j\}, \{k\})$ for all singleton sets W_1 and W_2.

Next, sequentially in increasing order of parameter i for all possible sets W_1 and W_2 the values of the Bellman function $H(i, W_1, W_2)$, $i = \overline{1, n}$ are determined by formula (3).

The value $H(n, I, I)$ determined by relation (4) is the optimal criterion value H_{opt} in problem (1).

In the process of executing the DP algorithm for each triple (i, W_1, W_2) of the values of the arguments of the Bellman function, we should fix the pair (α, β), on which the minimum of the right-hand side of relations (3), (4) is realized. This will allow, after finding the optimal value of the H_{opt} criterion, to uniquely determine the corresponding bi-assignment.

Relations (2)—(4) suppose implementation of the direct calculation scheme, without taking into account the state, unattainable from the initial one.

The complexity of the DP algorithm is determined by the number of calculated values of the Bellman function and is estimated as $O(4^n)$.

The results of experiments on performance evaluating of the DP algorithm for practically significant values of the dimension (n =10–13) demonstrate its practical importance for the type of applied problems under consideration.

Prospects for further research are to modify the proposed model for a wider range of applied problems, including those requiring multi-criteria formulation.

[1] *Fedosenko Y., Khandurin D.* Model and Algorithms for Synthesis of Bi-Assignment // Journal ofInformational Tecnologies and Control, 2020. Vol. 4 No 82.

[2] *Votaw D. F., Orden A.* The personnel assignment problem // In Symposium on Linear Inequalities and Programming. Scientific Computation of Optimum Programs. Project SCOOP, 1952. No 10. Pp. 155–163.

Решение задачи минимизации времени выполнения заказа для рекурсивного конвейера

Куприянов Борис Васильевич[1]⋆ kuprianovb@mail.ru
Лазарев Александр Алексеевич[1] jobmath@mail.ru
[1]Москва, ИПУ РАН

В докладе рассматривается метод решения RCPSP [2] задачи, минимизации времени выполнения множества заказов конвейером описываемым конечным набором рекурсивных функций. RCPSP задачи, как правило, являются полиномиально или NP трудными. В данном случае решение осуществляется сведением к Задаче Удовлетворения Ограничений (ЗУО), являющейся составной частью Constraint Programming. Модель конвейера представляет собой связный ациклический ориентированный граф с единственной конечной вершиной. Вершины графа помечены номерами из множества $I = 1, 2, \cdots, n$. Множество дуг графа упорядоченные пары вида (i, j), где $i, j \in I$.

Вершины графа помечены с помощью отображения $type : I \to E$, где $E = \{\mathbf{bop}, \mathbf{op}, \mathbf{and}, \mathbf{mul}, \mathbf{red}, \mathbf{get1}, \mathbf{get2}, \mathbf{put}\}$ — конечное перечислимое множество типов вершин. С каждым типом вершины связана определенная интерпретация и соответствующая рекурсивная функция. Расписание выполнения операций строится с помощью вычисления суперпозиции рекурсивных функций $t(i, k)$, определенных на множестве $I \times K$, где $K \subset N_0$ конечное множество и вычисляющих время завершения обработки i-й операцией k-го заказа при $k \in K$. Множество рекурсивных функций и их интерпретация подробно описаны в [1]. Каждая операция i конвейера использует некоторый возобновляемый ресурс $m_i \in D_i$ и характеризуется временем выполнения p_{ij}. D_i множество ресурсов, которое используется для выполнения операции i. M — множество возобновляемых ресурсов конвейера и $D_i \subseteq M$. В каждый момент времени каждая операция может использовать только один ресурс и каждый ресурс может использоваться только одной операцией.

Исходная задача сводится к минимизации функции $t(n, k)$ для заданного k на конечном множестве возобновляемых ресурсов M.

В теории ЗУО рассматривается как четверка (V, D, R, C), где $V = \{v_1, v_2, \cdots, v_n\}$ —множествопеременных, $D = \{D_1, \cdots, D_n\}$ — множество доменов переменных, R — множество отношений различной арности над подмножествами D. $C = \{C_1, \cdots, C_m\}$ — множество ограничений, связывающих множество значений переменных из V посредством отношений из R. Решение ЗУО это присвоение значений всем переменным множества V, которые удовлетворяют всем ограничениям из C. Целью решения ЗУО может быть нахождение одного или всех решений.

Перенесем общую постановку ЗУО в рассматриваемую нами прикладную область. Задача УО может быть представлена в виде сети ограничений. Было бы естественно в качестве такой сети взять граф модели конвейера, в котором вер-

шине соответствует некоторая операция. Однако, в связи с тем, что операция i выполняется k раз при наличии k заказов, она может использовать различные ресурсы для разных заказов, т.к. с операцией i связано множество ресурсов D_i. Чтобы преодолеть эту проблему сгенерируем из исходного графа модели для заданного $\hat{k}$ развернутый граф, в котором каждой вершине соответствует пара (номер операции, номер заказа) и сохраняются отношения предшествования. Построение такого графа возможно для любой модели конвейера и не представляет сложности. Если исходный граф имеет n вершин (операций), то новый граф имеет $n' = n\hat{k}$ вершин. Перенумеруем вершины развернутого графа некоторым способом. Пусть $I' = \{1, 2, \cdots, n'\}$ множество новых номеров. Будем считать по умолчанию, что $i' \in I'$ обозначает номер вершины нового графа и существуют отображения $\psi : I' \to I$ и $\varphi : I' \to K$.

После рассмотренных преобразований четверка ЗУО будет выглядеть следующим образом $V = \{x_1, \cdots, x_{n'}\}$ — множество дискретных переменных для каждой из которых задана область определения (домен). Переменная x_i взаимнооднозначно связана с парой (i, k) такой, что $i = \psi(x_i)$, $k = \varphi(x_i)$. $D_{i'} = D_i$ — домен или множество значений переменной $x_{i'}$, связанный с соответствующей i-й вершиной исходного графа, т.е. $i' \to (i, k)$, $R \subseteq (D_1 \times D_2 \times \cdots \times D_{n'})$, C — множество ограничений.

Множество ограничений делится на унарные, бинарные, тернарные и глобальные ограничения. Первые три задают ограничения на времена завершения операций, обусловленные отношениями предшествования и используемыми ресурсами. Например:

$$tnk(i,k) = \begin{cases} p_{ij}, \quad if(x_{i'} = m_j)\&(k = 0); \\ tnk(i, k-1) + p_{ij}, \quad if(x_{i'} = x_{pk(i')})\&(x_{i'} = m_j); \\ tnk(i, k-1) - p_{ij} + p_{il}, \quad if(x_{i'} \neq x_{pk(i')})\&(x_{i'} = m_l)\&(x_{pk(i')} = m_j); \end{cases}$$

Суть глобальных ограничений вида all-different в том, что все пересекающиеся во времени интервалы выполнения операций должны использовать различные ресурсы. В данном случае ЗУО будет с дискретной переменной и конечным множеством значений. Построение расписания для конвейера вычислением суперпозиции рекурсивных функций однозначно предполагает для решения ЗУО вариант поиска с возвратами.

[1] *Лазарев А. А., Гафаров Е. Р.* Теория расписаний. Задачи и алгоритмы // Москва: Изд-во МГУ, 2011. С. 223.

[2] *Куприянов Б. В.* Оценка и оптимизация производительности рекурсивного конвейера // Автоматика и телемеханика, 2020. № 5. С. 6–25.

Solving the problem of minimizing order lead time for a recursive conveyor

Boris Kupriyanov[1]⋆ kuprianovb@mail.ru
Alexander Lazarev[1] jobmath@mail.ru
[1]Moscow, Institute of Control Sciences of the Russian Academy of Sciences

The report discusses a method for solving the RCPSP [2] problem, minimizing the execution time of a set of jobs by a conveyor described by a finite set of recursive functions. RCPSP problems are usually polynomial or NP hard. In this case, the solution is carried out by reducing to the Constraint Satisfaction Problem, which is an integral part of Constraint Programming. The conveyor model is a connected acyclic directed graph with a single finite vertex. The vertices of the graph are marked with numbers from the set $I = \{1, 2, \cdots, n\}$. The set of arcs of a graph is jobed pairs of the form (i, j), where $i, j \in I$.

The graph vertices are marked using the mapping $type : I \to E$, where $E = \{\mathbf{bop}, \mathbf{op}, \mathbf{and}, \mathbf{mul}, \mathbf{red}, \mathbf{get1}, \mathbf{get2}, \mathbf{put}\}$ is a finite enumerable set of vertex types. Each vertex type is associated with a specific interpretation and a corresponding recursive function. The schedule of operations is constructed by calculating the superposition of recursive functions $t(i, k)$ defined on the set $I \times K$, where $K \subset N_0$ is a finite set and calculating the completion time of processing i by the k-th operation of the K-th job at $k \in K$. The set of recursive functions and their interpretation are described in detail in [1]. Each I operation in the conveyor uses some renewable resource $m_i \in D_i$ and is characterised by the processing requirement p_{ij}. D_i is the set of machines that is used to perform the i operation. M — the set of renewable conveyor resources and $D_i \subseteq m$. At any given time, each operation can only use one resource, and each resource can only be used by one operation.

The original problem is reduced to minimising the function $t(n, k)$ for a given k on a finite set of renewable resources M. In CSP theory, is considered as a four (V, D, R, C), where $V = v_1, v_2, \cdots, v_n$ — set of variables, $D = D_1, \cdots, D_n$ — set of variable domains, R — set of relations of different arity over subsets of D. $C = C_1, \cdots, C_m$ — set of constraints that bind the set of variable values from V by means of relations from R. The CSP solution is to assign values to all variables in the set V that satisfy all the constraints of C. The goal of a CSP solution may be to find one or all of the solutions.

Let's transfer the General statement of the CSP to the applied area we are considering. The CSP can be represented as a network of constraints. It would be natural to take as such a network the graph of the conveyor model, in which a vertex corresponds to some operation. However, since the i operation is performed k times when there are k jobs, it can use different resources for different jobs, since the i operation has many D_i resources associated with it. To overcome this problem, we will generate an expanded graph from the source graph of the model for the given $\hat{k}$, in which each vertex corresponds to a pair (operation number, job number) and the

precedence relations are preserved. The construction of such a graph is possible for any model of the conveyor and is not difficult. If the original graph has n vertices (operations), then the new graph has $n' = n\hat{k}$ vertices. Renumber the vertices of the expanded graph in some way. Let $I' = \{1, 2, \cdots, n'\}$ be the set of new numbers. By default, we assume that $i' \in I'$ denotes the vertex number of the new graph and there are mappings $\psi : I' \to I$ and $\varphi : I' \to K$.

After the considered transformations, the four of CSP will look like $V = \{x_1, \cdots, x_{n'}\}$ — a set of discrete variables for each of which a domain is defined. The variable x_i is one-to-one related to the pair (i, k) such that $i = \psi(x_i), k = \varphi(x_i)$ and $D_{i'} = D_i$ — domain or set of values of the variable $x_{i'}$ associated with the corresponding i-th vertex of the source graph, i.e. $i' \to (i, k)$, $R \subseteq (D_1 \times D_2 \times \cdots \times D_{n'})$, C — set of restrictions.

The set of constraints is divided into unary, binary, ternary, and global constraints. The first three set limits on the completion times of operations based on the precedence relationships and resources used. For example:

$$tnk(i,k) = \begin{cases} p_{ij}, \quad if(x_{i'} = m_j)\&(k = 0); \\ tnk(i, k-1) + p_{ij}, \quad if(x_{i'} = x_{pk(i')})\&(x_{i'} = m_j); \\ tnk(i, k-1) - p_{ij} + p_{il}, \quad if(x_{i'} \neq x_{pk(i')})\&(x_{i'} = m_l)\&(x_{pk(i')} = m_j); \end{cases}$$

The essence of global restrictions of the form all-different is that all overlapping time intervals of operations must use different resources. In this case, the CSP will be with a discrete variable and a finite set of values. Building a schedule for the conveyor by calculating a superposition of recursive functions unambiguously assumes a search option with backtracks for solving the CSP.

[1] *Kupriyanov B. V.* Evaluation and optimization of recursive pipeline performance // Automation and telemechanics, 2020. No 5. Pp. 6–25.

[2] *Lazarev A. A., Gafarov E. P.* The scheduling theory. Tasks and algorithms // Moscow: Izdatelstvo MSU, 2011. Pp. 223.

Стратегии комбинирования решений трехиндексной задачи о назначениях

Афраймович Лев Григорьевич[1] levafraimovich@gmail.com
Емелин Максим Денисович[1]⋆ makcum888e@mail.ru
[1]Нижний Новгород, ННГУ

Введение

Трёхиндексная аксиальная задача о назначениях имеет широкую область применения, примеры приведены в [1]. Мы рассматриваем задачу оптимального комбинирования допустимых решений трёхиндексной аксиальной задачи о назначениях. Данный алгоритм может быть применен в качестве дополнения к известным эвристическим или приближенным алгоритмам для постобработки полученных приближенных решений задачи о назначениях.

Постановка задачи

Пусть заданы три непересекающихся множества индексов $I, J, K, |I| = |J| = |K| = n$, а также трёхиндексная матрица стоимостей и трёхиндексная матрица неизвестных, $c_{ijk}, x_{ijk}, i \in I, j \in J, k \in K$, поставлена трёхиндексная аксиальная задача о назначениях:

$$\sum_{j\in J}\sum_{k\in K} x_{ijk} = 1, i \in I, \tag{1}$$

$$\sum_{i\in I}\sum_{k\in K} x_{ijk} = 1, j \in J, \tag{2}$$

$$\sum_{i\in I}\sum_{j\in J} x_{ijk} = 1, k \in K, \tag{3}$$

$$x_{ijk} \in \{0, 1\}, i \in I, j \in J, k \in K, \tag{4}$$

$$\sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c_{ijk}x_{ijk} \to min. \tag{5}$$

и известны m допустимых решений данной задачи $x^1, x^2, ..., x^m$.

Введем множество $W(x)$ следующим образом: $W(x) = \{(i, j, k)|x_{ijk} = 1, i \in I, j \in J, k \in K\}$. Обозначим через $Z(W(x^1, x^2, ..., x^m))$ задачу (1)-(6), где $W(x^1, x^2, ..., x^m) = W(x^1) \cup W(x^2) \cup ... \cup W(x^m)$

$$x_{ijk} = 0, (i, j, k) \notin W(x^1, x^2, ..., x^m) \tag{6}$$

Подходы к решению

Для случая $m = 2$ был разработан полиномиальный алгоритм решения поставленной задачи[2]. Для случая $m > 2$ было разработано несколько эвристических подходов для решения поставленной задачи. Будем комбинировать решения, используя алгоритм оптимальной комбинации двух допустимых решений

трехиндексной аксиальной задачи о назначениях[2]. Определим последовательную комбинацию решений следующим образом: комбинируем первое решение со вторым, на каждом следующем шаге комбинируем результат с предыдущего шага со следующим невыбранным решением.

Стратегия 1. Упорядочить решения в случайном порядке, провести последовательную комбинацию решений.

Стратегия 2. Упорядочить решения в порядке возрастания критерия, провести последовательную комбинацию решений.

Стратегия 3. Упорядочить решения в порядке возрастания критерия, провести последовательную комбинацию решений. Затем k раз упорядочить решения в порядке возрастания критерия, выбрать случайно половину решений и поменять их местами в случайном порядке, провести последовательную комбинацию решений. Для $k+1$ полученного решения провести последовательную комбинацию.

Вычислительный эксперимент

Эксперимент проводился на тестах, генерируемых по той же идее, что и в [3]. Параметры c_{ijk} выбирались равновероятно из отрезка [0,300]. Для каждого теста генерировалось n^3 случаных решений, каждое из которых проходило через процедуру локальной оптимизации[4]. При проведении эксперимента параметр k из стратегии 3 был выбран равным 4. Для сравнения мы также нашли минимум из этих решений. Для серии экспериментов будем оценивать среднее отклонение от оптимума в серии. В каждой серии было 10 задач одинаковой размерности.

n	Минимум из решений	Стратегия 1	Стратегия 2	Стратегия 3
10	3,195%	2,399%	2,399%	1,478%
11	8,883%	7,413%	6,332%	4,274%
12	15,054%	14,185%	14,185%	10,923%
13	24,319%	24,319%	24,319%	22,010%
14	41,210%	34,207%	34,387%	23,276%
15	54,856%	47,248%	52,372%	46,553%
16	72,872%	69,792%	70,677%	69,473%
17	68,145%	68,145%	59,684%	51,387%
18	87,201%	81,402%	82,272%	66,769%
19	100,464%	88,211%	90,114%	81,358%

Согласно приведенным выше результатам минимум из решений отклоняется в среднем на 47,620%, Стратегия 1 на 43,732%, Стратегия 2 на 43,674%, Стратегия 3 на 37,750%.

[1] *Афраймович Л. Г.* Многоиндексные транспортные задачи с декомпозиционной структурой // Автоматика и телемеханика, 2012. № 1. С. 130–147.

[2] *Афраймович Л. Г. Емелин М. Д.* Решение задачи оптимального комбинирования двух допустимых решений трёхиндексной аксиальной задачи о назначениях // Материалы XII международного семинара «Дискретная математика и ее приложения», 2019. С. 193–196.

[3] *Balas E. Saltzman M. J.* An Algorithm for the Three-Index Assignment Problem // Operations Research, 1991. Vol. 39. No 1. Pp. 150–161.

[4] *Huang G. Lim A.* A hybrid genetic algorithm for the Three-Index Assignment Problem // European Journal of Operational Research, 2006. Vol. 172. Pp. 249–257.

Strategies for combining solutions to a three-index assignment problem

Lev Afraimovich[1] levafraimovich@gmail.com
Maksim Emelin[1]⋆ makcum888e@mail.ru
[1]Nizhny Novgorod, NNSU

Introduction

The three-index axial assignment problem has a wide range of applications, examples are given in [1]. We consider the problem of optimal combination of feasible solutions to the three-index axial assignment problem. This algorithm can be used as a supplement to the known heuristic or approximate algorithms for post-processing the obtained approximate solutions to the assignment problem.

Problem statement

Suppose that there are three disjoint sets of indexes $I, J, K, |I| = |J| = |K| = n$, as well as a three-index matrix of values and a three-index matrix of unknowns, $c_{ijk}, x_{ijk}, i \in I, j \in J, k \in K$, a three-index axial assignment problem is defined:

$$\sum_{j\in J}\sum_{k\in K} x_{ijk} = 1, i \in I, \tag{1}$$

$$\sum_{i\in I}\sum_{k\in K} x_{ijk} = 1, j \in J, \tag{2}$$

$$\sum_{i\in I}\sum_{j\in J} x_{ijk} = 1, k \in K, \tag{3}$$

$$x_{ijk} \in \{0, 1\}, i \in I, j \in J, k \in K, \tag{4}$$

$$\sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c_{ijk}x_{ijk} \to min. \tag{5}$$

and m possible solutions to this problem are known $x^1, x^2, ..., x^m$. We introduce the set W(x) as follows: $W(x) = \{(i, j, k)|x_{ijk} = 1, i \in I, j \in J, k \in K\}$. Denote by $Z(W(x^1, x^2, ..., x^m))$ problem (1)-(6), where $W(x^1, x^2, ..., x^m) = W(x^1) \cup W(x^2) \cup ... \cup W(x^m)$

$$x_{ijk} = 0, (i, j, k) \notin W(x^1, x^2, ..., x^m) \tag{6}$$

Solution approaches

For the case $m = 2$, a polynomial algorithm for solving the problem was developed[2]. For the case $m > 2$, several heuristic approaches have been developed to solve the problem. We will combine the solutions using the algorithm of optimal combination of two feasible solutions to the three-index axial assignment problem[2]. Let's define a sequential combination of solutions as follows: we combine the first solution with the second one, and at each next step we combine the result from the previous step with the next unselected solution.

Strategy 1. Arrange the solutions in a random order, run a sequential combination of solutions.

Strategy 2. Arrange the solutions in ascending order of the criterion, and perform a sequential combination of solutions.

Strategy 3. Arrange the solutions in ascending order of the criterion, and perform a sequential combination of solutions. Then arrange the solutions k times in ascending order of the criterion, select half of the solutions randomly and swap them in random order, and perform a sequential combination of solutions. For the $k+1$ solutions received, perform a sequential combination.

Computational experiment

The experiment was conducted on tests generated by the same idea as in [3]. Parametrs c_{ijk} were chosen equally from the segment [0,300]. For each test, n^3 random solutions were generated, each of which went through the local optimization procedure[4]. During the experiment, parameter k from strategy 3 was chosen to be 4. For comparison, we also found the minimum of these solutions. For a series of experiments, we will estimate the average deviation from the optimum in the series. Each series had 10 problems of the same dimension.

n	Minimum of solutions	Strategy 1	Strategy 2	Strategy 3
10	3,195%	2,399%	2,399%	1,478%
11	8,883%	7,413%	6,332%	4,274%
12	15,054%	14,185%	14,185%	10,923%
13	24,319%	24,319%	24,319%	22,010%
14	41,210%	34,207%	34,387%	23,276%
15	54,856%	47,248%	52,372%	46,553%
16	72,872%	69,792%	70,677%	69,473%
17	68,145%	68,145%	59,684%	51,387%
18	87,201%	81,402%	82,272%	66,769%
19	100,464%	88,211%	90,114%	81,358%

According to the results above, the minimum of solutions deviate on average by 47.620%, Strategy 1 by 43.732%, Strategy 2 by 43.674%, Strategy 3 by 37.750%.

[1] *Afraimovich L.* A Multi-index Transport Problems with Decomposition Structure // Autom. Remote Control, 2012 Vol. 73. No 1. Pp. 118–133.

[2] *Afraimovich L. Emelin M.* Solving the problem of optimal combination of two acceptable solutions to the three-index axial assignment problem // Materials of the XIII International seminar "Discrete Mathematics and its Application", 2019. Pp. 193–196.

[3] *Balas E. Saltzman M. J.* An Algorithm for the Three-Index Assignment Problem // Operations Research, 1991. Vol. 39. No 1. Pp. 150–161.

[4] *Huang G. Lim A.* A hybrid genetic algorithm for the Three-Index Assignment Problem // European Journal of Operational Research, 2006. Vol. 172. Pp. 249–257.

О проекте цифровой эко-системы для создания виртуального рынка цифровых двойников предприятий электротехнической промышленности

Скобелев Петр Олегович[1] petr.skobelev@gmail.com
Ларюхин Владимир Борисович[1]⋆ vladimir.larukhin@live.ru
[1]Самара, Самарский Государтсвенй Технический Университет

В условиях растущей сложности и неопределенности современной экономики в отрасли промышленности наблюдается существенное снижение прибыли и обострение конкуренции предприятий, который при этом еще часто должны сочетать производство гражданской и военной продукции. В ряде случае, например, в электротехнической промышленности ни одно из предприятий на рынке часто не способно целиком выполнить крупный заказ Газпрома или Лукойла (например, на поставку комплектных электротехнических подстанций), требующий в комплектации широкой номенклатурой покупных и производимых изделий.

Возникающая при этом новая сложная задача состоит в том, чтобы автоматизировать процесс формирования цепочек кооперации такого рода предприятий, которые бы осуществлялись в реальном времени в сам момент формирования запроса от крупного заказчика с учетом текущей загрузки, компетенций, ресурсных мощностей и ограничений каждого предприятия и возможностей их кооперации.

При этом ручной режим такого рода переговоров крайне сложный и трудоемкий – каждому предприятию потребуется держать целый штат, чтобы знать состояние и планы предприятия, проверять наличие комплектующих на складе или стоимость и срок и их заказа на стороне, планировать производственные процессы с учетом особенностей изделий, технологических процессов, станков и компетенций рабочих предприятия, а также отвечать через формирование технико-коммерческих предложений (ТКП) на каждый запрос, которых может приходить десяток в день от разных заказчиков. Однако, наиболее сложная часть этого процесса состоит в том, чтобы, анализируя получаемые ТКП, понять, кто из предприятий может войти в формируемую цепочку и в какой части производимого изделия, что будет наиболее выгодно как заказчику, так и другим участникам цепочки, и где каждый исполнитель должен пойти на компромисс и уступить, в счет той прибыли, которой получит от принятой части. Важным стимулом участия в такой кооперации могут быть принципы «солидарной экономики», которые, если цепочка сложится, и сводное ТКП будет принято, из образующейся прибыли позволят возместить уступки тем предприятиям, кто согласился на уменьшение цены или отказ от части своей поставки в угоду интересов цепочки в целом.

Этот подход может способствовать развитию новых подходов Индустрии 5.0 в части внедрения систем искусственного интеллекта (ИИ), участвующих в та-

ких запросах как со стороны заказчика, так и потенциального исполнителя, цифровизации знаний и формирования цифровых эко-систем колоний систем ИИ («систем систем»). А также принципов Общества 5.0, основанного на внедрении такого рода цифровых систем ИИ, в части модели «coopetition» (от англ. «cooperation» - кооперация и «competition» - конкуренция), т.е. динамического переключения от конкуренции к кооперации, и наоборот, в зависимости от складывающейся ситуации.

В результате предлагаемого подхода впервые будет построена цифровая экосистема умных цифровых двойников (ИСУР) различных предприятий, организованная как «система систем» с p2p взаимодействием отдельных систем на общей цифровой платформе.

Результатом внедрения является решения сложной задачи обеспечения поставок по комплексным запросам, повышение гибкости и эффективности управления ресурсами, сокращение времени в 100-100 раз на принятие решений, прозрачность и снижение зависимости от человеческого фактора, возможность масштабирования бизнеса без роста численности административного персонала.

Работа поддержана грантом РФФИ № 20-37-90052.

[1] *Rzhevski G., Skobelev P.* Managing Complexity // London-Boston: WIT Press, 2014. Pp. 156.

About the digital ecosystem project to create a virtual market for digital twins of electrical industry enterprises

Petr Skobelev[1] petr.skobelev@gmail.com
Vladimir Laryukhin[1]⋆ vladimir.larukhin@live.ru
[1]Samara, Samara State Technical University

In the context of the growing complexity and uncertainty of the modern economy, the industry of complete sets of electrical equipment has seen a significant decline in profits and increased competition between enterprises.

In some cases, none of the enterprises in the market of complete sets of electrical equipment is often able to fully fulfill a large order from Gazprom or LUKOIL (for example, for the supply of complete electrical substations), which requires a wide range of purchased and manufactured products. The new complex task that arises is to automate the process of forming chains of cooperation of such enterprises, which would be carried out in real time at the very moment of forming a request from a large customer, taking into account the current load, competencies, resource capacities and limitations of each enterprise and the possibilities of their cooperation.

In this manual mode of such negotiations is extremely complex and time-consuming – each company will need to keep a large staff of experts to know the current status and plans of the company, to check the availability of components in the warehouse or the cost and time for their order on the party, planning the production process with the features of products, processes, tools and competences of workers of the enterprise and to respond through the formation of technical and commercial proposals on each request, which can come a dozen a day from different customers.

However, the most difficult part of this process is to analyze the received proposals on customer side, to understand which of the enterprises can enter the formed chain and in which part of the manufactured product, which will be most profitable for both the customer and other participants in the chain, and where each performer should compromise and give up, at the expense of the profit that he will receive from the accepted part. An important incentive to participate in such cooperation can be the principles of "solidary economy", which, if the chain is formed and the consolidated proposal is adopted, from the resulting profits will allow to compensate concessions to those enterprises who agreed to reduce the price or refuse part of their supply in favor of the interests of the chain as a whole.

This approach can contribute to the development of new approaches in the 5.0 Industry in terms of implementing artificial intelligence (AI) systems that participate in such requests from both the customer and the potential performer, digitizing knowledge, and creating digital eco-systems for colonies of AI systems ("systems of systems"). As well as the principles of Society 5.0, based on the introduction of this kind of digital AI systems, in terms of the "coopetition" model: "cooperation" and

"competition", i.e. dynamic switching from competition to cooperation, and Vice versa, depending on the current situation.

As a result of the proposed approach, a digital ecosystem of smart digital twins of various enterprises will be built for the first time, organized as a "system of systems" with p2p interaction of individual systems on a common digital platform.

The implementation results in solving the complex task of ensuring deliveries for complex requests, increasing the flexibility and efficiency of resource management, reducing the time for decision-making by 100-100 times, transparency and reducing dependence on the human factor, and the ability to scale the business without increasing the number of administrative staff.

This research is funded by RFBR, grant 20-37-90052.

[1] *Rzhevski G., Skobelev P.* Managing Complexity // London-Boston: WIT Press, 2014. Pp. 156.

Математическое моделирование планирования подготовки космонавтов

Джуманов Ратмир Рамаевич[1] dzhumanov.r19@physics.msu.ru
Хуснуллин Наиль Фаридович[1]⋆ Nhusnullin@gmail.com
Лазарев Александр Алексеевич[1] jobmath@mail.ru

[1]Москва, МГУ им. М.В. Ломоносова, Институт проблем управления им. В. А. Трапезникова РАН

На сегодняшний день проблема автоматического составления расписаний является актуальным в специализированных центрах подготовки.Подготовкой космнонавтов занимается Центр подготовки космонавтов(ЦПК) им. Ю.А. Гагарина. Задача этих центров — планирование подготовки экипажей на тренажерах и формирования определенных навыков и умений, необходимых для выполнения задач космического полета.

В данный момент расписание составляется вручную, что является очень трудоемкой работой. Каждое изменение плана приводит к значительным трудозатратам из-за широких горизонтов планирования. Существует множество как точных, так и эвристических методов решения подобных задач . В [1] были использованы метод целочисленного линейного программирования и метод программирования в ограничениях. Последний оказался наиболее эффективным. Эксперименты были проведены как на псевдореальных, так и на реальных данных из ЦПК им. Ю.А. Гагарина.

В итоге разработанные методы позволяют за несколько минут выдать допустимое расписание для экипажа, вместо нескольких дней ручного труда. В дальнейшем планируется доработка существующих алгоритмов и математических моделей составления расписаний для нескольких экипажей одновременно. Это поможет не только сократить трудозатраты на составление расписаний, но и ускорить сам процесс планирования подготовки экипажей.

Работа выполнена при частичной поддержке гранта РФФИ № 20-58-S52006

[1] *Лазарев А. А., Бронников С. В., Герасимов А. В., Мусатова Е. Г., Петров А. С., Пономорев К. В., Харламов М. М., Хуснуллин Н. Ф., Ядренцев Д. А.* Математическое моделирование планирования подготовки космонавтов // Управление большими системами, 2016. С. 25.

Mathematical modeling of cosmonaut training planning

Ratmir Jumanov[1] dzhumanov.r19@physics.msu.ru
Nail Husnullin[1]⋆ Nhusnullin@gmail.com
Alexander Lazarev[1] jobmath@gmail.com
[1]Moscow, Lomonosov Moscow State University, Institute of Control Sciences RAS

Nowadays, automatic scheduling is relevant in specialized training centres. Cosmonauts are trained in the Yu.A. Gagarin Research & Test Cosmonaut Training Centre. The task of these centers is to plan the training of crews using simulators and the formation of certain skills and abilities necessary for performing space flight tasks.

At the moment, the schedule drawn up by hand, which is a very time consuming job. Therefore, each change to the plan leads to significant labor costs due to the wide planning horizons. There are many methods for solving such problems, both precise and heuristic. In [1] were used the integer linear programming method and the constraint programming method. The latter proved to be the most effective. The experiments were carried out both on the pseudo-real data and on real data from the Yu.A. Gagarin Research & Test Cosmonaut Training Centre.

As a result, the developed methods make it possible to issue an permissible schedule for the crew in a few minutes, instead of several days of manual labor. In the future, it is planned to refine the existing algorithms and mathematical models for scheduling for several crews at the same time. This will help not only reduce scheduling labor, but also speed up the planning process for crew training.

This research was supported by RFBR project 20-58-S52006

[1] *Lazarev A. A., Bronnikov S. V., Gerasimov A. V., Musatova E. G., Petrov A. S., Ponomorev K. V., Kharlamov M. M., Husnullin N. F., Yadrentsev D. A.* Mathematical modeling of cosmonaut training planning // Large-Scale Systems Control, 2016. Pp. 25.

Распределение комплекса работ по исполнителям

Макаровских Татьяна Анатольевна[1] Makarovskikh.T.A@susu.ru
Панюкова Александра Анатольевна[2]⋆ 3meandme@gmail.com
[1]Челябинск, Южно-Уральский государственный университет
[2]Москва, ГАПОУ Колледж предпринимательства

Пусть дан комплекс работ в виде ациклического оргрфа $G(V, A)$, где V — множество работ, A — отношение непосредственного предшествования на множестве работ: из $(v_i, v_j) \in A$ следует, что работа $v_i \in V$ должна быть выполнена до начала выполнения работы $v_j \in V$.

Пусть имеется множество U возможных исполнителей. На множестве $V \times U$ задана вектор-функция $f : V \times U \to \mathbb{Z}^{+N}$. Координатные функции $f_i(v, u)$ $i = 1, 2, \dots N$ определяют различные показатели качества выполнения работы $v \in V$ исполнителем $u \in U$.

Отношение совместимости пары соисполнителей $\{u_k, u_l\} \subset U$ при выполнении смежных работ $(v_i, v_j) \in A$ моделируются вектор функцией $g : A \times \{\{u_1, u_2\} \subseteq U\} \to \mathbb{Z}^{+M}$. Координатные функции $g_i((v_i, v_j), u_1, u_2)$ $i = 1, 2, \dots M$ определяют различные показатели качества выполнения смежных работ $(v_i, v_j) \in A$ исполнителями u_1, u_2.

Ставится задача нахождения однозначного отображения $\varphi : V \to U$, то есть назначению каждой работе $v \in V$ исполнителя $u \in U$, для которого выполнены ограничения

$$\underline{f} \leqslant f(v, \varphi(v) \leqslant \overline{f}, \quad \underline{g} \leqslant g(v, \varphi(v) \leqslant \overline{g},$$

соответствующие распределению работ по исполнителям, при котором значения функций f и g были бы допустимы (находились в допустимых интервалах). На f и g можно определить целевой функционал и оптимизировать его. Значения функций могут быть заданы в виде таблиц, хранящихся в базе данных либо рассчитываемых алгоритмически.

Поставленная задача является обобщением задачи Вебера. Разработкой этой проблематики занимались Иорданский М.А., Забудский Г.Г. [1], Панюков А.В. [2, 3], Сергеев С.И., Сигал И.Х., Стоян Ю.Г., Трубин В.А., Adolfson D., Beckmann M.J., Burcard R.E. [4], Francis R.L. [5], Koopmans T.C., Tamir A., Wesolowsky G.O. и другие [6].

Планируется использование решения данной задачи в мобильных приложениях. Отличительной особенностью этих приложений является относительно небольшой граф G и достаточно большое множество U, представляющее собой базу данных о возможных исполнителях, большинство из которых могут являться самозанятыми. Например, при заказе фотосессии помимо услуг фотографа заказчик арендует студию, пользуется услугами визажиста, а фотограф в свое время арендует определенное оборудование. Требуется на основе имеющих-

ся данных подобрать для выполнения проекта исполнителей, удовлетворяющих критериям заказчика.

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования РФ (государственное задание FENU–2020–0022).

[1] *Забудский Г. Г., Лагздин А. Ю.* Полиномиальные алгоритмы решения квадратичной задачи о назначениях на сетях // Журнал вычислительной математики и математической физики, 2010, Т. 50. № 11. С. 2052–2059.

[2] *Панюков А. В., Пельцвергер Б. Ф., Шафир А. Ю.* Оптимальное размщение точек ветвления транспортной сети на цифровой модели местности // АиТ, 1990. № 9. С. 153–162.

[3] *Панюков А. В., Шангин Р. Э.* Точный алгоритм решения дискретной задачи Вебера для k-дерева // Дискретный анализ и исследование операций, 2014. Т. 21. № 3. С. 64–75.

[4] *Burkard R., Cela E.* Quadratic and three-dimensional assignments: An annotated bibliography // Computational Optimization, 1998. Vol. 4. Pp. 373–392.

[5] *Francis R.L., McGinnis L.F., WHite J.A.* Futility Layout and Location: An Analytical Approach // Prentice Hall, Englewood Cliffs, 1991.

[6] *Шангин Р. Э.* Алгоритм точного решения дискретной задачи Вебера для простого цикла // Прикладная дискретная математика, 2013. № 4. С. 96–102.

Distribution of the Complex of Jobs by Performer

Tatiana Makarovskikh[1] makarovskikh.t.a@susu.ru
Alexandra Panyukova[2]⋆ 3meandme@gmail.com
[1]Chelyabinsk, South Ural State University
[2]Moscow, College of Entrepreneurship No. 11

Let the complex of jobs is represented by acyclic digraph $G(V, A)$, where V is the set of jobs, A is an immediate precedence relation on a set of jobs, and if $(v_i, v_j) \in A$ then job $v_i \in V$ is hold before the job $v_j \in V$ starts.

Let we have a set U of possible performers. A vector-function $f : V \times U \to \mathbb{Z}^{+N}$ is defined on $V \times U$. Coordinate functions $f_i(v, u)$ $i = 1, 2, \dots N$ define the different performance indicators $v \in V$ by $u \in U$.

The relation of co-performers $\{u_k, u_l\} \subset U$ compatibility while executing the adjacent jobs $(v_i, v_j) \in A$ may be modelled by vector-function $g : A \times \{\{u_1, u_2\} \subseteq U\} \to \mathbb{Z}^{+M}$. The coordinate functions $g_i((v_i, v_j), u_1, u_2)$ $i = 1, 2, \dots M$ define the different quality indicators of adjacent jobs $(v_i, v_j) \in A$ by performers u_1, u_2.

The task is to find injection $\varphi : V \to U$, i.e. the assignment of a performer $u \in U$ to each job $v \in V$. The following restrictions are hold for this performer:

$$\underline{f} \leqslant f(v, \varphi(v) \leqslant \overline{f}, \quad \underline{g} \leqslant g(v, \varphi(v) \leqslant \overline{g}.$$

They correspond to distribution of jobs by performers when the values of functions f and g are permissible (belong to permissible intervals). It is possible to define and optimize the objective function on f and g. The values of these functions may be obtained either from database tables or calculated automatically.

The problem posed is a generalization of Weber's problem. The development of this problem was carried out by M.A. Iordansky, G.G. Zabudsky [1], A.V. Panyukov [2, 3], S.I. Sergeev, I.Kh. Sigal, Yu.G. Stoyan, V.A. Trubin, Adolfson D., Beckmann M.J., Burcard R.E. [4], Francis R.L. [5], Koopmans T.C., Tamir A., Wesolowsky G.O. and others [6].

The implementation of this problem solution is planned for mobile apps. A distinctive feature of these applications is a relatively small graph G and a fairly large set of U, which is a database of possible performers, most of whom may be self-employed. For example, when ordering a photo session, in addition to the services of a photographer, the customer rents a studio, uses the services of a make-up artist, and the photographer at one time may rent some certain equipment. It is required, on the basis of the available data, to select for the implementation of the project a set of performers who meet the criteria of the customer.

The work was supported by the Ministry of Science and Higher Education of the Russian Federation (government order FENU-2020-0022).

[1] *Zabudsky G. G., Lagzdin A. Yu.* Polynomial Algorithms for Solving the Quadratic Assignment Problem on Networks // Journal of Computational Mathematics and Mathematical Physics, 2010. Vol. 50. No 11. Pp. 2052–2059.

[2] *Panyukov A. V., Pelzverger B. F., Shafir A. Yu.* Optimal placement of transport network branch points on a digital terrain model // Automation and Remote Control, 1990. No. 9. Pp. 153–162.

[3] *Panyukov A. V., Shangin R. E.* An exact algorithm for solving the discrete Weber problem for a k-tree // Discrete Analysis and Operation Research, 2014. Vol. 21. No 3. Pp. 64–75.

[4] *Burkard R., Cela E.* Quadratic and three-dimensional assignments: An annotated bibliography // Computational Optimization, 1998. Vol. 4. Pp. 373–392.

[5] *Francis R.L., McGinnis L.F., WHite J.A.* Futility Layout and Location: An Analytical Approach // Prentice Hall, Englewood Cliffs, 1991.

[6] *Shangin R. E.* Algorithm for the exact solution of the discrete Weber problem for a simple cycle // Applied Discrete Mathematics, 2013. No. 4. Pp. 96–102.

Методы автоматической сборки белков

Гришин Егор Максимович[1,2] grishin.em16@physics.msu.ru
[1]Москва, Институт Проблем Управления им. В. А. Трапезникова РАН
[2]Москва, Московский Государственный Университет им. М. В. Ломоносова

Основная задача протеомики – собрать белок по спектру, полученному при помощи масс-спектрометрии. Пики спектра соответствуют пептидам. Информация о пептидах, полученная по спектру называется ридами (а составляющие рида – контиги). Пептиды состоят из аминокислот, которые, в свою очередь, кодируются триплетами нуклеотидов (A, T, G, C). Характеристики всех аминокислот известны. На первом этапе необходимо обработать спектр (удалить шумы, выровнять и др.). Далее необходимо сгенерировать пептиды и определить последовательность аминокислот, которые наилучшим образом описывают спектр (пики). И, наконец, по набору полученных пептидов и дополнительной информации об исходном белке (например, его масса) необходимо восстановить исходный белок. При этом на любом этапе идентификации белка возможны пропуски, неточности и ошибки, которые необходимо учитывать и исправлять.

Качественным и количественным исследованием белков ученые занимаются уже много лет. Одним из первых методов была деградация по Эдману [1], заключающаяся в повторяющемся отщеплении меченых концевых аминокислотных остатков, с их последующей идентификацией при помощи хроматографии. Однако реактивы для этого метода достаточно дорогие, а сам метод медленный. С появлением компьютеров были разработаны новые методы исследования белков. В современных методах исследование производится на основе данных, полученных при помощи более дешевой и быстрой масс-спектрометрии. На сегодняшний день существуют два основных подхода: поиск по базам данных и подход de novo. В [2] подробно представлены многие аспекты протеомики, в частности, различные методы идентификации белков.

В общем случае алгоритмы поиска по базам данных состоит из нескольких шагов. На первом шаге экспериментально получают спектр исследуемого белка при помощи масс-спектрометрии. На втором шаге производится обработка полученного спектра. На третьем шаге происходит поиск по базе данных с различными параметрами, и отбираются потенциальные совпадающие белки. Производится обработка теоретического спектра для сравнения с экспериментальным. На последнем шаге происходит сравнение двух спектров. При этом оценивается вероятность совпадения исследуемого и теоретически полученного белков. В [3] представлен один из алгоритмов поиска по базам данных. Следует отметить, что при помощи поиска по базам данных можно идентифицировать только известные белки, в отличие от подходов de novo и Overlap Layout Consensus.

Метод Overlap Layout Consensus является предшественником de novo. В этом методе среди всех полученных из спектра ридов (последовательности амино-

кислот одинаковой длины) производится поиск пересечений, и на основе этого строится полная последовательность пептидов. Основным методом поиска пересечений является динамическое программирование. Сложность разработанных алгоритмов не ниже $O(n^2)$, что приводит к большим временам сборки белка. Однако преимуществом подхода Overlap Layout Consensus является сравнительно малое количество требуемой памяти. В [4] предложен метод типа Overlap Layout Consensus.

Наибольший интерес представляет подход de novo сборки. Его достоинствами являются линейная сложность $O(n)$ и возможность получения неизвестных пептидных последовательностей. Основной идеей данного подхода является построение графа де Брюина возможных перестановок на основе данных масс-спектрометрии. Пики спектрограммы соответствуют вершинам графа, а расстояния между пиками соответствуют ребрам. В [5] предложен алгоритм поиска Евклидова пути в подобном графе, который и является исследуемой последовательностью для сборки генома. Однако размеры построенных графов велики, и для поиска приходится использовать супер компьютеры. В [6] проведен подробный анализ использования графов для сборки белковых последовательностей, построенных по спектрам, а также представлены основные задач, решение которых необходимо для более эффективного применения подхода de novo.

Работа поддержана грантом фонда Базис № 20-2-9-12-1.

[1] *Edman P.* Method for determination of the amino acidsequence in peptides // Acta Chem, Scand., 1950. Vol. 4. Pp. 283–293.

[2] *Simon J. H., Andrew R. J.* Proteome Bioinformatics // Springer, New-York, 2010. Pp. 397.

[3] *Xu H., Freitas A. F.* A high mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data // BMC Bioinformatics, 2007. Vol. 8. Pp. 133.

[4] *Sutton G., White O., Adams M., Kerlavage A.* TIGR assembler: A new tool for assembling large shotgun sequencing projects // Genome Sci Technol, 1995. Vol. 1. Pp. 9–19.

[5] *Pavel A. P., Haixu T., Michael S. W.* An Eulerian path approach to DNA fragment assembly // Proceedings of the National Academy of Sciences of the United States of America, 2001. Vol. 98. No 17. Pp. 9748–9753.

[6] *Raffaella R., Stefano B., Murray P., Yuri P., Marco P., Gianluca D. V., Paola B.* Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era // Quantitative Biology, 2019. Vol. 7. Pp. 278–292.

Methods of automated protein assembly

Egor Grishin[1,2] grishin.em16@physics.msu.ru
[1]Moscow, V.A. Trapeznikov Institute of Control Science RAS
[2]Moscow, M.V. Lomonosov Moscow State University

The main goal of proteomics is to assemble protein from the spectrum obtained by mass spectrometry. Spectrum peaks correspond to peptides. Spectrum information on peptides is called reads (and the read components are contigs). Peptides consist of amino acids, which in turn are encoded by triplets of nucleotides (A, T, G, C). The characteristics of all amino acids are known. In the first stage, it is necessary to prepare the spectrum (remove noises, level out, etc.). Then peptides must be generated and the sequence of amino acids that represent the best way to describe the spectrum (peaks) must be determined. Finally, using the set of peptides obtained and additional information about the original protein (e.g. its mass), it is necessary to reconstruct the original protein. At any stage of protein identification, omissions, inaccuracies and errors may occur that need to be taken into account and corrected.

Scientists have been engaged in qualitative and quantitative proteins analysis for many years to date. One of the first methods was Edman degradation [1], which consisted in the repeated detachment of labeled amino acid residues, with their subsequent identification by chromatography. However, the reagents for this method are quite expensive and the method itself is slow. With the advances in computer technology, new methods for protein analysis have been developed. In modern methods, research is based on data obtained by using cheaper and faster mass spectrometry. Today, there are two main approaches: database search and the de novo approach. Many aspects of proteomics are detailed in [2], in particular the various methods of protein identification are presented.

In general, database search algorithms consist of several steps. The first step is to experimentally obtain the spectrum of the protein under study using mass spectrometry. In the second step, the obtained spectrum is processed. In the third step, a database search with different parameters is performed, and potential matching proteins are selected. The theoretical spectrum is processed for comparison with the experimental one. In the last step, the two spectra are compared. The probability that the protein under study and the theoretically obtained protein will coincide is assessed. One of the database search algorithms is presented in [3]. It should be noted that database search can only identify known proteins, as opposed to the de novo and Overlap Layout Consensus approaches.

The Overlap Layout Consensus method is the predecessor of de novo. In this method, overlaps are found among all the reads from the spectrum (amino acid sequences of the same length) and a complete sequence of peptides is built on this basis. The main method for finding intersections is dynamic programming. The complexity of the developed algorithms is no less than $O(n^2)$, which leads to long assembly times of the protein. However, the advantage of the Overlap Layout Con-

sensus approach is the relatively small amount of memory required. A method of the Overlap Layout Consensus type has been proposed in [4].

The most interesting is the de novo assembly approach. Its advantages are the linear complexity $O(n)$ and the possibility of obtaining unknown peptide sequences. The main idea of this approach is to construct a de Bruin graph of possible permutations based on mass spectrometry data. The peaks of the spectrum correspond to the vertices of the graph, and the distances between the peaks correspond to the edges. An algorithm for finding the Euclidian path in a similar graph, which is the sequence under study for assembling the genome, is proposed in [5]. However, the size of the graphs built is large, and super computers have to be used for searching. A detailed analysis of the use of graphs for assembly of protein sequences built by spectrum has been carried out in [6], and the main problems that have to be solved in order to apply the de novo approach more effectively are presented.

This research is funded by Bazis foundation, grant 20-2-9-12-1.

[1] *Edman P.* Method for determination of the amino acidsequence in peptides // Acta Chem, Scand., 1950. Vol. 4. Pp. 283–293.

[2] *Simon J. H., Andrew R. J.* Proteome Bioinformatics // Springer, New-York, 2010. Pp. 397.

[3] *Xu H., Freitas A. F.* A high mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data // BMC Bioinformatics, 2007. Vol. 8. Pp. 133.

[4] *Sutton G., White O., Adams M., Kerlavage A.* TIGR assembler: A new tool for assembling large shotgun sequencing projects // Genome Sci Technol, 1995. Vol. 1. Pp. 9–19.

[5] *Pavel A. P., Haixu T., Michael S. W.* An Eulerian path approach to DNA fragment assembly // Proceedings of the National Academy of Sciences of the United States of America, 2001. Vol. 98. No 17. Pp. 9748–9753.

[6] *Raffaella R., Stefano B., Murray P., Yuri P., Marco P., Gianluca D. V., Paola B.* Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era // Quantitative Biology, 2019. Vol. 7. Pp. 278–292.

Верхние и нижние границы параллельного партийного планирования для одной машины с учетом последовательности работ

Гафаров Евгений Рашидович[1,2]⋆ axel73@mail.ru
Долгий Александр Борисович[2] alexandre.dolgui@imt-atlantique.fr
Сомов Михаил Львович[3] somovml1999@gmail.com

[1]Москва, Институт проблем управления им. В. А. Трапезникова РАН
[2]Нант, IMT Atlantique
[3]Москва, Московский Государственный Университет им. М. В. Ломоносова

Рассматривается задача параллельного партийного планирования на одной машине с учетом ограничений на последовательность работ. Задача формулируется следующим образом. Имеется набор из $N = \{1, 2, ..., n\}$ работ. Работа $j \in N$ должна быть выполнена за время p_j без прерываний. Кроме того, отношение последовательности работ задаются ациклическим направленным графом $G = (N, V)$. Работы выполняются партиями, в партию входит подмножество работ, и каждая работа должна быть включена в какую-нибудь партию. Время выполнения одной партии работ равно максимальному времени выполнения работы из этой партий. Если между работами есть отношения в последовательности, то они не могут быть в одной партии. Цель состоит в том, что бы определить партии для всех работ $j = 1, 2, ..., n$ и выполнялись следующие условия:

— Соблюдены все отношения последовательности работ
— $C_{max} = \sum_{l=1}^{L} P_l$ - минимизирован, где P_l - время выполнения партии работ l, а L - количество партий в решении

Представлены шесть полиномиальных нижних и верхних границ для задачи, где отношение последовательности работ задается цепочками. А также их экспериментальное сравнение и относительные погрешности. Показано, что относительная погрешность некоторых простых нижних и верхних границ не ограничена константой. Для более сложных нижних и верхних границ приведен численный эксперимент. В дальнейшем планируется рассмотреть верхние и нижние границы для общего случая, когда граф отношений работ более сложный и имеет циклы.

Работа выполнена при частичной поддержке гранта РФФИ № 20-58-S52006

On lower and upper bounds for single machine parallel batch scheduling subject to chains of jobs.

Evgeny Gafarov[1,2]⋆ axel73@mail.ru
Alexandre Dolgui[2] alexandre.dolgui@imt-atlantique.fr
Somov Mikhail[3] somovml1999@gmail.com
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences
[2]Nantes, IMT Atlantique
[3]Moscow, Moscow State University

We consider single machine parallel batch scheduling problem subject to precedence constraints. The problem is formulate as follows. Given a set $N = \{1, 2, ..., n\}$ of jobs. Job $j \in N$ has to be processed for $p_j \geqslant 0$ time units without preemption. Furthermore, finish-start precedence relations $i \to j$ are defined between the jobs according to an acyclic directed graph $G = (N, V)$. The jobs are processed in batches, where a batch is a subset of jobs, and we require that the batches form a partition of the set of all jobs. The processing time of a batch is equal to the maximum processing time among the jobs in this batch. If there is a precedence relation between jobs i and j, then these jobs cannot be processed in the same batch. The objective is to determine a batch for each job $j, j = 1, 2, ..., n$, in such a way that:

— The given precedence relations are fulfilled
— $C_{max} = \sum_{l=1}^{L} P_l$ - minimized, where P_l - processing time of batch l and L – a number of batches in a solution.

Six polynomial lower and upper bounds are presented for the task, where the ratio of the job sequence is set by chains. And also their experimental comparison and relative errors are presented. It is shown, that the relative error of some simple lower and upper bounds are not bounded by a constant. For more complicated lower and upper bounds a numerical experiment is provided. In the future, it is planned to consider the upper and lower boundaries for the more general case when the graph of work relationships is more complex and has cycles.

This research was supported by RFBR project 20-58-S52006

Методы повышения эффективности энергетических систем

Галахов Семен Алексеевич[1,2] galakhov.sa16@physics.msu.ru
[1]Москва, Институт Проблем Управления им. В. А. Трапезникова РАН
[2]Москва, Московский Государственный Университет им. М. В. Ломоносова

С развитием технологий растет потребность в более эффективном их использовании. Например, наступление эпохи интернета вещей (IoT) существенно увеличит нагрузку не только на телекоммуникационную, но и на энергетическую сети. Под интернетом вещей можно понимать множество технических устройств и датчиков, объединенных друг с другом различными каналами связи и подключенных к одной большой сети. Поскольку окружающий мир находится в постоянном движении, такая концепция требует высокой автономности всех вещей, то есть каждое устройство или датчик должны "научиться" быть менее зависимым от электросети, а в идеале получать электроэнергию из окружающей среды. Однако, в настоящее время большинство "вещей" работают на аккумуляторах, что ставит перед исследователями задачу создания алгоритма, позволяющего оптимальным образом распределить энергию между заряжаемыми устройствами.

В общей формулировке задачи на пункт зарядки поступают различные "вещи", которые требуют определенное количество заряда. Работы поступают в режиме реального времени и характеристики каждой работы становятся известны только в момент ее поступления. Такая задача может быть сформулирована в рамках задачи онлайн рюкзака, являющейся NP-полной задачей в общем случае. В статье [1] авторы рассматривают задачу зарядки электромобилей на зарядной станции. Однако в рамках концепции интернета вещей, вычислительные мощности бывают настолько ограничены, что необходимо разработать не эффективный алгоритм, а алгоритм число операций которого не больше заданного числа N.

Самыми быстрыми алгоритмами являются некоторые эвристические алгоритмы. Они наименее ресурсозатратны, однако качество решения (в общем случае) довольно невысоко. Таким образом, планируется разработать наиболее простой эффективный алгоритм решения подобной задачи, который будет иметь преимущество по сравнению с последовательной зарядкой всех поступающих требований. Эти преимущества должны учитывать особенности современных аккумуляторов [2] и обеспечить удобство конечного пользователя.

Работа поддержана грантом фонда Базис № 20-2-9-32-1.

[1] *Sun B., Zeynali A., Li T., Hajiesmaili M., Wierman A., Tsang T.* Competitive Algorithms for the Online Multiple Knapsack Problem with Application to Electric Vehicle Charging // `https://arxiv.org/pdf/2010.00412.pdf`

[2] *Kim M., Baek J., Han S.* Optimal Charging Method for Effective Li-ion Battery Life Extension Based on Reinforcement Learning // `https://arxiv.org/pdf/2005.08770.pdf`

Methods for improving the efficiency of energy systems

Galakhov Semen[1,2] galakhov.sa16@physics.msu.ru
[1]Moscow, ICS RAS
[2]Moscow, MSU

With the development of technologies, needs to use them more efficiently increases. For example, the advent of the Internet of Things (IoT) era will significantly increase the load not only on telecommunications but also on the energy networks. The Internet of Things can be understood as a set of technical devices and sensors combined with each other by different communication channels and connected in one large network. Since the surrounding world is in a permanent motion, this concept requires a high autonomy of all things. That is, each device or sensor must "learn" to be less dependent on the electricity grid, and to get electricity from the environment in the best way. At present, however, most "things" use batteries, that sets the challenge for researchers to develop an algorithm that allows to distribute energy between the charged devices optimally.

In general, various "things" come to the charging station and require a certain amount of charge. The jobs comes in real time and the characteristics of each job become known only at the moment its release. Such a problem can be formulated as part of the online knapsack problem, which is an NP complete in the general case. In [1] the authors consider the problem of charging electric vehicles at the charging station. However, within the framework of the Internet of Things concept, computational power can be so limited that it is necessary to develop not an effective algorithm, but an algorithm with number of operations less than given amount N.

Several heuristic algorithms are the fastest ones. They are the least resource consuming, but the quality of their solution (in general) is rather low. Thus, it is planned to develop the simplest effective algorithm for solving such problem, which will have an advantage in comparison with sequential charging of all incoming jobs. These advantages should take into account the features of modern batteries [2] and provide convenience to the end customer.

This research is funded by RFBR, grant 20-2-9-32-1.

[1] *Sun B., Zeynali A., Li T., Hajiesmaili M., Wierman A., Tsang T.* Competitive Algorithms for the Online Multiple Knapsack Problem with Application to Electric Vehicle Charging // https://arxiv.org/pdf/2010.00412.pdf

[2] *Kim M., Baek J., Han S.* Optimal Charging Method for Effective Li-ion Battery Life Extension Based on Reinforcement Learning // https://arxiv.org/pdf/2005.08770.pdf

Расширение возможностей метрического подхода на основе теории средних и теории ошибок

Сидельников Юрий Валентинович[1]⋆ sidelnikovy@mail.ru

[1]Москва, Институт проблем управления им. В.А.Трапезникова РАН

Метрики достаточно широко используются в фундаментальных и прикладных исследованиях. Например, в теории расписания, которая является фундаментальной областью дискретной оптимизации [2]. Полагаем, что метрический подход для решения задач может получить новый импульс при использовании, не только метрик, но и их расширений, например псевдометрик. В этом случае, ослабляется первая аксиома (рефлексивность метрики), и не требуется выполнение следующего условия: чтобы из $(x, y) = 0$ следовало $x = y$. В данном случае, мы рассматриваем вариант, когда ослабляется вторая аксиома (симметричность метрики). В таком случае, можно использовать ассиметричные меры, например, для числового случая $E(x, y) = |x - y|/|y|$. Обычно такие меры называют показателями ошибки. Ряд исследований рассматривают взаимосвязи между разработками в области теорий ошибок и средних с другими областями исследования. Такими были, например, работы А.И. Орлова, а также автора данного доклада установившего взаимосвязи между различными видами числовых средних и характеристиками классов эквивалентности числовых показателей ошибок. В данном исследовании был усилен результат полученный ак. А.Н. Колмогоровым, который показал, что если при рассмотрении ассоциативной средней добавить ряд условий, включая симметричность, то она будет иметь аналитический вид [3]. В работе была получена формула для ассоциативной средней, но уже без условия симметричности [1].

[1] *Сидельников Ю. В.* Технология экспертного прогнозирования // Автореферат на соискание ученой степени д.т.н., 2002.

[2] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling, 2021. No 89. Pp. 1163–1176.

[3] *Kolmogorov, A. N.* Sur la notion de moyenne // Cl. Sci. Fis. Mat. Natur., 1930. Vol. 12. No 6. Pp. 388–391.

Expansion of the possibilities of metric approach on the basic of the theory of averages and theory of the errors

Yury Sidelnikov[1]⋆ sidelnikovy@mail.ru

[1]Moscow, ICS RAS

Metrics are widely used in fundamental and applied studies. For example, in schedule theory, which is a fundamental area of discrete optimization [2]. We believe that the metric approach to solving problems can get a new impulse when using not only metrics, but also their extensions, such as pseudometrics. In this case, the first axiom (reflexivity of the metric) is weakened, and the following condition is not required: for $(x, y) = 0$ to follow $x = y$. In this case, we consider the case when the second axiom (symmetry of the metric) is weakened. In this case, you can use asymmetric measures, for example, for the numeric case. $E(x, y) = |x - y|/|y|$. These measures are usually referred to as error indicators. A number of studies examine the relationship between developments in error and average theories and other areas of research. Such were, for example, the works of A. I. Orlov, as well as the author of this report, who established the relationship between different types of numerical averages and the characteristics of equivalence classes of numerical error indicators [1]. In this study, the result obtained by A. N. Kolmogorov was strengthened, which showed that if a number of conditions, including symmetry, are added to the associative mean, it will have an analytical form [3]. In this paper, we obtained a formula for the associational average, but without the symmetry condition [1].

[1] *Sidelnikov Y.* Technology of the expert prognostication // Author's abstract to the competition of the scientific degree "Doctor of Engineering Science", 2002.

[2] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling, 2021. No 89. Pp. 1163–1176.

[3] *Kolmogorov, A. N.* Sur la notion de moyenne // Cl. Sci. Fis. Mat. Natur., 1930. Vol. 12. No 6. Pp. 388–391.

Аппроксимация целевой функции задач теории расписаний

Барашов Егор Борисович[1,2]⋆ barashov.eb@gmail.com
Лазарев Александр Алексеевич[1] jobmath@mail.ru
Правдивец Николай Александрович[1] pravdivets@ya.ru
[1]Москва, Институт Проблем Управления им. В. А. Трапезникова РАН
[2]Москва, Московский Государственный Университет им. М. В. Ломоносова

Рассматривается линейная аппроксимация для задачи одного прибора теории расписаний: предполагается, что существует линейная относительно моментов окончания обслуживания требований целевая функция. Расписания, построенные «вручную», являются оптимальными относительно этой целевой функции. Аппроксимируются неизвестные значения весовых коэффициентов целевой функции, что сводится, как будет показано, к решению системы линейных неравенств относительно этих коэффициентов.

Исследуется задача $1|r_j|\sum w_jC_j$: имеются один прибор и множество $J = \{1, 2, ..., n\}$ из n требований, которые необходимо обслужить на приборе. Для каждого требования $j \in J$ заданы момент поступления на прибор r_j и длительность обслуживания p_j. Отношения предшествования отсутствуют (не накладывается ограничений на очередность обслуживания требований), прерывания в обслуживании требований и искусственные простои прибора запрещены. Расписание π однозначно задаётся порядком обслуживания требований $(j_1, \ldots, j_n)$. В задаче $1|r_j|\sum w_jC_j$ необходимо найти расписание π^0, минимизирующее суммарное взвешенное время завершения обслуживания требований $\sum w_jC_j$, где C_j-момент окончания обслуживания требования j, а $w_j > 0$ весовой коэффициент соответствующего времени завершения обслуживания.

Примером I задачи $1||\sum w_jC_j$ *будем называть множество значений длительностей обслуживания требований*:

$$I = \{p_1, ..., p_n\}.$$

Систему неравенств относительно весовых коэффициентов w_j:

$$\frac{w_{j_1^k}}{p_{j_1^k}} \geqslant \frac{w_{j_2^k}}{p_{j_2^k}} \geqslant \ldots \geqslant \frac{w_{j_n^k}}{p_{j_n^k}} \tag{1}$$

будем называть исходной системой неравенств задачи аппроксимации весовых коэффициентов для случая $r_1 = ... = r_n$. Исходная система (1) содержит $N(n-1)$ неравенств. Пусть $K = (1, \ldots, k, \ldots, N)$ есть множество индексов k, соответствующее заданным парам (I_k, π_k^0) примеров задачи и их оптимальных расписаний, и пусть $\widetilde{K} \subset K$ есть некоторое подмножество множества K. Далее запись вида:

$$\min_{k\in\tilde{K}} \text{ или } \max_{k\in\tilde{K}}$$

будет означать минимум (максимум) по всем возможным парам (I_k, π_k^0) таким, что $\min\limits_{k\in\tilde{K}}$.

Выберем произвольным образом пару различных требований $i, j \in (1, ..., n), i \neq = j$. Разобьем множество K на два подмножества $K_{i,j}$ и $K_{j,i}$ в зависимости от взаимного расположения требований i,j в расписании π_k^0:

$$K_{i,j} = (k \in K) : \pi_k^0 = (\ldots, i, \ldots, j, \ldots),$$

$$K_{j,i} = (k \in K) : \pi_k^0 = (\ldots, j, \ldots, i, \ldots).$$

Тогда из неравенств (1) исходной системы для соответствующих весовых коэффициентов w_i, w_j имеем:

$$\frac{w_j}{w_i} \leqslant \frac{p_j^k}{p_i^k}, k \in K_{i,j}, \tag{2a}$$

$$\frac{w_j}{w_i} \geqslant \frac{p_j^k}{p_i^k}, k \in K_{j,i}. \tag{2b}$$

Далее, пусть

$$Y(i,j) = \min_{k\in K_i,j}(\tfrac{p_j^k}{p_i^k}),$$
$$X(i,j) = \min_{k\in K_j,i}(\tfrac{p_j^k}{p_i^k}),$$

тогда система неравенств для выбранных i, j эквивалентна двойному неравенству:

$$X(i,j) \leqslant \frac{w_j}{w_i} \leqslant Y(i,j). \tag{3}$$

Алгоритм аппроксимации весовых коэффициентов целевой функции основан на решении эффективной системы неравенств:

$$\begin{cases} i, j \in (1, ..., n), i \neq j \\ K = (k), k = \overline{1, N} \\ K_i, j = (k \in K : \pi_k^0 = (...i, ..., j...)) \\ K_j, i = (k \in K : \pi_k^0 = (...j, ..., i...)) \\ X(i,j) = max_{k\in K_j,i}(\frac{p_j^k}{p_i^k}) \\ Y(i,j) = max_{k\in K_i,j}(\frac{p_j^k}{p_i^k}) \\ X(i,j) \leqslant \frac{w_j}{w_i} \leqslant Y(i,j). \end{cases}$$

Индекс l выбирается произвольным образом, для аппроксимации коэффициентов w_j необходимо:

1. построить множества K_i, j, K_j, i;
2. заполнить матрицы $X(i,j), Y(i,j)$;
3. вычислить матрицы $\widetilde{X}(i,j), \widetilde{Y}(i,j)$;
4. вычислить $w_j = \begin{cases} 1, j = l; \\ (\frac{\widetilde{X}(l,j)+\widetilde{Y}(l,j))}{2}), j \neq l, \text{где индекс l любое число.} \end{cases}$

Результатом алгоритма является набор таких весовых коэффициентов w_j, $j = \overline{1,n}$, что для каждого из N заданных примеров оптимальное расписание, найденное для аппроксимированных значений весовых коэффициентов либо полностью совпадает с его заданным оптимальным расписанием, соответствующим неизвестному истинному набору весовых коэффициентов w_j^0, либо имеет с ним одинаковое значение целевой функции.

Исследование выполнено при частичной финансовой поддержке РФФИ и Министерством по науке и технологиям Тайваня в рамках научного проекта №20-58-S52006.

[1] *Черников С.Н.* Линейные неравенства // Итоги науки и техники. Серия «Алгебра. Топология. Геометрия», 1966. С. 137–187.

Approximation of the objective function of scheduling problems

Egor Barashov[1,2]⋆ barashov.eb@gmail.com
Alexander Lazarev[1] jobmath@mail.ru
Nikolay Pravdivets[1] pravdivets@ya.ru
[1]Moscow, V. A. Trapeznikov Institute of Control Science RAS
[2]Moscow, M. V. Lomonosov Moscow State University

We consider a linear approximation for a single machine problem: it is assumed that there is a linear objective function regarding to completion times of jobs. Schedules that are constructed "manually" are optimal according to the objective function. Unknown values of the weight coefficients of the objective function are approximated, which is reduced to solving a system of linear inequalities.

The problem $1|r_j|\sum w_jC_j$ is under research: there is one machine and a set of jobs $J = \{1, 2, ..., n\}$ that have to be processed on the machine. For each job $j \in J$ release date r_j and processing time p_j are known. There are no precedence relations, and interrupts in the processing of jobs are prohibited. The schedule π is the order in which jobs are processed $(j_1, ..., j_n)$. In the problem $1|r_j|\sum w_jC_j$ we need to construct a schedule π^0 that minimizes the total weighted completion time $\sum w_jC_j$, where C_j is the job j completion time, and $w_j > 0$ is the weight coefficient of the corresponding completion time.

We will denote by instance I of the problem $1||\sum w_jC_j$ the set of values for the duration of processing jobs:

$$I = \{p_1, ..., p_n\}.$$

The following system of inequalities:

$$\frac{w_{j_1^k}}{p_{j_1^k}} \geqslant \frac{w_{j_2^k}}{p_{j_2^k}} \geqslant \ldots \geqslant \frac{w_{j_n^k}}{p_{j_n^k}} \tag{4}$$

will be called the initial system of inequalities of the weight coefficients approximation problem for the case $r_1 = \ldots = r_n$. The original system (4) contains $N(n-1)$ inequalities. Let $K = (1, \ldots, k, \ldots, N)$ be the set of indices k corresponding to the given pairs (I_k, π_k^0) of instances of the problem and their optimal schedules. Let $\tilde{K} \subset K$ will be a subset of the set K. Next, the expression:

$$\min_{k\in\tilde{K}} \text{ or } \max_{k\in\tilde{K}}$$

will means the minimum (maximum) for all possible pairs (I_k, π_k^0) such that $\min\limits_{k\in\tilde{K}}$. Let's randomly select a pair of different jobs $i, j \in (1, ..., n), i \neq j$. We divide the set K into two subsets $K_{i,j}$ and $K_{j,i}$ depending on the relative position of the jobs i, j in the schedule π_k^0:

$K_{i,j} = (k \in K) : \pi_k^0 = (..., i, ..., j, ...)$,

$K_{i,j} = (k \in K) : \pi_k^0 = (..., j, ..., i, ...)$, Then from the inequalities (4) of the original system for the corresponding weight coefficients w_i, w_j we have:

$$\frac{w_j}{w_i} \leqslant \frac{p_j^k}{p_i^k}, k \in K_{i,j} \tag{5a}$$

$$\frac{w_j}{w_i} \geqslant \frac{p_j^k}{p_i^k}, k \in K_{j,i} \tag{5b}$$

Next, let it be

$$Y(i,j) = \min_{k \in K_i, j} (\frac{p_j^k}{p_i^k})$$

$$X(i,j) = \min_{k \in K_j, i} (\frac{p_j^k}{p_i^k})$$

then the system of inequalities for the selected i, j is equivalent to the double inequality:

$$X(i,j) \leqslant \frac{w_j}{w_i} \leqslant Y(i,j). \tag{6}$$

The algorithm for weight coefficients approximation of the objective function is based on solving an effective system of inequalities:

$$\begin{cases} i, j \in (1, ..., n), i \neq j \\ K = (k), k = \overline{1, N} \\ K_i, j = (k \in K : \pi_k^0 = (...i, ..., j...)) \\ K_j, i = (k \in K : \pi_k^0 = (...j, ..., i...)) \\ X(i,j) = max_{k \in K_j, i} (\frac{p_j^k}{p_i^k}) \\ Y(i,j) = max_{k \in K_i, j} (\frac{p_j^k}{p_i^k}) \\ X(i,j) \leqslant \frac{w_j}{w_i} \leqslant Y(i,j). \end{cases}$$

The index l is chosen arbitrarily, for the approximation of the coefficients of the w_j it is necessary:

1. to construct sets $K_{i,j}, K_{j,i}$;
2. to fill in the matrices $X(i,j), Y(i,j)$;
3. to calculate the matrices $\widetilde{X}(i,j), \widetilde{Y}(i,j)$;
4. to calculate $w_j = \begin{cases} 1, j = l; \\ (\frac{\tilde{X}(l,j) + \tilde{Y}(l,j))}{2}), j \neq l, \text{where index l is any number.} \end{cases}$

The result of the algorithm is a set of such weight coefficients w_j, $j = \overline{1,n}$ that for each of the N given examples, the optimal schedule found for the approximated values of the weight coefficients either completely coincides with its given optimal schedule corresponding to the unknown true set of weight coefficients w_j^0, or has the same value of the objective function with it.

The research was partially supported by RFBR and MOST (project No 20-58-S52006).

[1] *Chernikov S. N.* Linear inequalities // Results of science and technology. Series "Algebra. Topology. Geometry", 1966. Pp. 137–187.

Метрический подход для задач железнодорожного планирования

Лазарев Александр Алексеевич jobmath@mail.ru
Лемтюжникова Дарья Владимировна* darabbt@gmail.com
Москва, Институт проблем управления им. В. А. Трапезникова РАН

Большинство задач теории расписаний NP-трудны в сильном смысле. Однако, для многих из них существуют полиномиально разрешимые случаи. В данном исследовании показан метрический подход [1] для задачи планирования однопутной железной дороги с двумя станциями. Этот подход позволяет строить решения с гарантированной точностью за полиномиальное время, используя специальные случаи исходной NP-трудной задачи.

Пример задачи планирования представляет точку в $f(n)$-мерном пространстве Ω, где n — количество работ. Если одно и то же расписание π используется в качестве решения для двух разных произвольных примеров A и B, то можно сделать оценку разности между их значениями целевой функции $V_A(\pi)$ и $V_B(\pi)$. Оценка формируется как метрика $\rho(A, B)$, определенная на $\Omega \times \Omega$:

$$|V_A(\pi) - V_B(\pi)| \leqslant \rho(A, B). \tag{1}$$

Пусть существуют два оптимальных расписания π^A и π^B для примеров A и B соответственно. Тогда можно построить оценку для выражения (2) и доказать,что оно зависит от $\rho(A, B)$. $\Delta(\rho(A, B))$ является абсолютной точностью для случая, когда π^B используется в качестве решения, например, A вместо реального оптимального решения.

$$V_A(\pi^B) - V_A(\pi^A) \leqslant \Delta(\rho(A, B)). \tag{2}$$

Метод применен к нескольким классическим NP-трудным задачам планирования. Показано изменение среднего отношения абсолютной ошибки к ее верхней границе Δ для множества тестовых примеров.

Работа выполнена при частичной поддержке гранта РФФИ № 20-58-S52006

[1] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling, 2021. No 89. Pp. 1163–1176.

Metric approach for railway planning problems

Alexandr Lazarev jobmath@mail.ru
Darya Lemtyuzhnikova⋆ darabbt@gmail.com
Moscow, V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences

Most scheduling problems are NP-hard in a strong sense. However for many of them there are polynomial solvable cases. This research shows a metric approach [1] for the problem of planning a single-track railway with two stations. This approach allows us to construct solutions with guaranteed accuracy in polynomial time using special cases of the original NP-hard problem.

An instance of a scheduling problem represents a point in the $f(n)$–dimensional space Ω where n is the number of jobs. If the same schedule π is used as a solution for two different arbitrary instances A and B, then we can estimate the difference between their values of the objective function $V_A(\pi)$ and $V_B(\pi)$. The score is formed as a metric $rho(A, B)$ defined on $\Omega \times \Omega$:

$$|V_A(\pi) - V_B(\pi)| \leqslant \rho(A, B). \tag{3}$$

Let there be two optimal schedules π^A and π^B for the instances A and B, respectively. Then we can construct an estimate for expression (2) and prove that it depends on $\rho(A, B)$. $\Delta(\rho(A, B))$ is absolute precision for the case when π^B is used as a solution, for example, A instead of the real optimal solution.

$$V_A(\pi^B) - V_A(\pi^A) \leqslant \Delta(\rho(A, B)). \tag{4}$$

The method is applied to several classical NP-hard scheduling problems. The change in the average ratio of the absolute error to its upper bound Δ is shown for a set of test cases.

This research was supported by RFBR project 20-58-S52006

[1] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling, 2021. No 89. Pp. 1163–1176.

Метрическая интерполяция в задачах теории расписаний

Лемтюжникова Дарья Владимировна[1] darabbt@gmail.com
Тюняткин Андрей Александрович[2]⋆ andtun@yandex.ru
[1]Москва, Институт проблем управления им. В. А. Трапезникова РАН
[2]Москва, Московский Государственный Университет им. М. В. Ломоносова

Подавляющее большинство задач теории расписаний NP-трудны. Для решения каждой такой задачи необходим отдельный алгоритм: аппроксимационный или полиномиальный в среднем. Производительность таких алгоритмов сильно зависит от входных данных.

Предлагается универсальный подход – *метрическая интерполяция*. Данный метод можно использовать для уменьшения погрешности при использовании эвристических алгоритмов, получения начального решения для точных методов и аппроксимации любых задач теории расписаний при любых входных данных с гарантированной абсолютной погрешностью.

Интерполируются значения целевой функции: подбираются точки интерполирования, вычислительная сложность нахождения решения в которых приемлема. Оценивается значение целевой функции исходного примера. Далее, используя метрический подход [1], строится оптимальное расписание. При наличии некоторой полиномиально разрешимой области, решение исходного примера может быть аппроксимировано, используя метрический подход.

На данный момент строятся различные эффективные метрики и проводятся численные эксперименты с ними. Планируется разработка специальных алгоритмов нахождения интерполяционных прямых.

В этом исследовании предложен метод метрической интерполяции и представлены результаты численных экспериментов на тестовых примерах.

Работа выполнена при частичной поддержке гранта РФФИ № 20-58-S52006

[1] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling, 2021. No 89. Pp. 1163–1176.

Metric interpolation in scheduling problems

Darya Lemtyuzhnikova[1] darabbt@gmail.com
Andrey Tyunyatkin[2]⋆ andtun@yandex.ru
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences
[2]Moscow, Lomonosov Moscow State University

The vast majority of scheduling problems are NP-hard. To solve each such problem, we need a separate algorithm: an approximation scheme or an algorithm with polynomial average time. The performance of such algorithms strongly depends on the input data.

A universal approach is proposed – *a metric interpolation*. This method can be used to reduce the heuristic algorithms errors, obtain an initial solution for exact methods, and approximately solve any problems in scheduling theory for any input data with guaranteed absolute error.

The values of the objective function are interpolated: those interpolating nodes are selected for which the computational complexity is acceptable. The value of the target function of the original instance is then approximated. After that, using the metric approach [1], an optimal schedule is constructed. If a polynomially solvable domain exists, the solution of the original instance can be approximated using the metric approach.

Currently, various effective metrics are being developed and numerical experiments with metrics are being conducted. It is planned to develop special algorithms for finding interpolation lines.

In this research, a metric interpolation method is proposed and the results of numerical experiments on test instances are presented.

This research was supported by RFBR project 20-58-S52006

[1] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling, 2021. No 89. Pp. 1163–1176.

Оптимальное управление системами массового обслуживания в условиях применения для описания их состояния методов структурно-классификационной и экспертно-статистической обработки

Мандель Александр Соломонович[1]⋆ almandel@yandex.ru
Лаптин Виктор Алексеевич[2] straqker@bk.ru
[1]Москва, Институт проблем управления им. В. А. Трапезникова РАН
[2]Москва, Московский государственный университет им. М. В. Ломоносова

Рассматривается управляемая система массового обслуживания (СМО), которая принадлежит к тому классу управляемых СМО, что были исследованы в работе [1]. Отмечено, что наибольшей проблемой при практическом применении этих моделей является оценка вероятностных распределений и других вероятностных характеристик, которые входят в описание этих моделей. Предлагается новая интеллектуальная процедура анализа исходных статистических данных, которая опирается на методы экспертно-классификационного и экспертно-статистического анализа.

Итак, рассматривается многолинейная СМО, в которой в качестве управления используется число включенных в работу каналов обслуживания. Решения о включении (из числа резервных каналов) или отключении рабочих каналов принимаются в дискретные периодические моменты времени. Считается, что в эти моменты времени интенсивность простейшего входящего потока претерпевает марковские скачки в соответствии с матрицей вероятностей перехода $P = \|p_{ij}\|$. При этом задано конечное множество Λ возможных значений интенсивности входящего потока.

Оптимальная стратегия переключения каналов удовлетворяет следующему уравнению дискретного динамического программирования:

$$C_1^*(\lambda_i, m) = \min_{u \geqslant \underline{u_i}} C^{(1)}(\lambda_i, m, u),$$

$$C_n^*(\lambda_i, m) = \min_{u \geqslant \underline{u_i}} (C^{(1)}(\lambda_i, m, u) + \alpha \sum_{j=1}^{k} p_{ij} C_{n-1}^*(\lambda_j, u)),$$

где $C^{(1)}(\lambda_i, m, u)$ – это функция одношаговых затрат, $\lambda_i \in \Lambda$, m – число рабочих каналов, с которого начинается очередной шаг, u – число каналов, которое следует включить (управление), α – коэффициент дисконтирования, а $C_n^*(\lambda_i, m)$ – минимально возможное значение затрат на n последних шагах периода планирования. Показано, что оптимальная стратегия переключения каналов имеет пороговый характер и указан способ вычисления соответствующих пороговых значений и что использование этих стратегий приносит значительный экономический эффект.

Главная проблема при практическом использовании модели заключается в том, что необходимо уметь оценивать матрицу вероятностей перехода $P = \|p_{ij}\|$ и значения λ_i интенсивностей входящего потока. Важно понимать, что предложенная математическая модель является лишь некоторым приближением к реальной действительности. Для того, чтобы оценивать ее параметры необходимо использовать не только статистическую информацию, но и прибегать к помощи экспертов-предметников. Иначе говоря, воспользоваться возможностями экспертно-классификационного и экспертно-статистического подходов.

В рассматриваемом случае доступная статистика может сводиться к отрезкам временных рядов, содержащих моменты поступления в СМО очередных требований, разнообразной информации о состоянии рынка и других сведениям. Как правило, разумно представлять эти временные ряды в виде последовательности чисел, характеризующих числа требований, поступавших в систему на каждом шаге.

Вся эта информация образует многомерное пространство признаков, которые подвергаются классификационному, экспертно-классификационному и экспертно-статистическому анализу. При этом число выделяемых классов должно быть согласовано с экспертами-предметниками. При выделении в процессе классификации каждого класса с присвоением ему соответствующего номера i вычисляется и эталон класса. Значение интенсивности λ, соответствующее этому, эталонному, объекту становится одним из элементов λ_i множества Λ.

Затем вычисляется начальное приближение к оценкам вероятностей перехода:

$$p_{ji}^{(1)} = \frac{\alpha_j^{(1)}}{R_{ji}^{(1)}},$$

а нормирующий множитель $\alpha_j^{(1)}$ вычисляется по формуле:

$$\alpha_j^{(1)} = \frac{1}{\sum_{i=1}^{r} \frac{1}{R_{ji}^{(1)}}},$$

где символом $R_{ji}^{(1)}$ обозначено расстояние между j-м объектом пространства признаков и i-м эталоном.

Можно также выписать соотношения для последующего уточнения оценок вероятностей p_{ij}, которые будут опубликованы в полном тексте доклада.

[1] *Mandel A., Laptin V.* Myopic Channel Switching Strategies for Stationary Mode: Threshold Calculation Algorithms // Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, 2018. Vol. 919. Pp. 410–420.

Optimal control of queuing systems (QS) when using the methods of structural-classification and expert-statistical processing to describe the QS state

Alexander Mandel[1]⋆ almandel@yandex.ru
Viktor Laptin[2] straqker@bk.ru
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences RAS
[2]Moscow, M. V. Lomonosov Moscow State University

A controlled queuing system (QS) is considered, which belongs to the class of controlled QS that was investigated in [**?**]. It is noted that the greatest problem in the practical application of these models is the assessment of probability distributions and other probabilistic characteristics that are included in the description of these models. A new intellectual procedure for the analysis of the initial statistical data is proposed, which is based on the methods of expert-classification and expert-statistical analysis.

So, we consider a multi-channel QS, in which the number of service channels included in the operation is used as a control. Decisions to turn on (from the number of backup channels) or turn off working channels are made at discrete periodic moments of time. It is believed that at these moments in time the intensity of the simplest incoming flow undergoes Markov jumps in accordance with the transition probability matrix $P = \|p_{ij}\|$. In this case, a finite set Λ of possible values of the intensity of the incoming flow is given.

The optimal channel switching strategy satisfies the following discrete dynamic programming equation:

$$C_1^*(\lambda_i, m) = \min_{u \geqslant \underline{u_i}} C^{(1)}(\lambda_i, m, u),$$

$$C_n^*(\lambda_i, m) = \min_{u \geqslant \underline{u_i}} (C^{(1)}(\lambda_i, m, u) + \alpha \sum_{j=1}^{k} p_{ij} C_{n-1}^*(\lambda_j, u)),$$

where $C^{(1)}(\lambda_i, m, u)$ is the function of one-step costs, $\lambda_i \in \Lambda$, m is the number of working channels from which the next step begins, u is the number of channels to be switched on (control) , α is the discount factor, and $C_n^*(\lambda_i, m)$ is the minimum possible cost value at the last n steps of the planning period. It is shown also that the optimal channel switching strategy has a threshold character and a method for calculating the corresponding threshold values is indicated, and that the use of these strategies brings a significant economic effect.

The main problem in the practical use of the model is that it is necessary to be able to estimate the matrix of transition probabilities $P = \|p_{ij}\|$ and the values λ_i of the intensities of the incoming flow. It is important to understand that the proposed mathematical model is only some approximation to reality. In order to assess its

parameters, it is necessary to use not only statistical information, but also resort to the help of subject experts. In other words, take advantage of the possibilities of expert-classification and expert-statistical approaches.

In the case under consideration, the available statistics can be reduced to time series segments containing the times of demand arrivals in the QS, various information about the state of the market and other information. As a rule, it is reasonable to represent these time series as a sequence of numbers characterizing the number of demands entering the system at each step.

All this information forms a multidimensional space of features, which are subject to classification, expert-classification and expert-statistical analysis. In this case, the number of allocated classes should be agreed with subject experts. When each class is allocated in the process of classification and the corresponding number i is assigned to it, the class standard is also calculated. The intensity value λ corresponding to this reference object becomes one of the elements λ_i of the set Λ.

Then the initial approximation to the transition probabilities is calculated as

$$p_{ji}^{(1)} = \frac{\alpha_j^{(1)}}{R_{ji}^{(1)}},$$

and the normalizing factor $\alpha_j^{(1)}$ is calculated by the formula:

$$\alpha_j^{(1)} = \frac{1}{\sum_{i=1}^{r} \frac{1}{R_{ji}^{(1)}}},$$

where the symbol $R_{ji}^{(1)}$ denotes the distance between the j-th object of the feature space and the i-th standard.

The relations are also given for the subsequent refinement of the estimates of the probabilities p_{ij}, which will be published in the full text of the report.

[1] *Mandel A., Laptin V.* Myopic Channel Switching Strategies for Stationary Mode: Threshold Calculation Algorithms // Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, 2018. Vol. 919. Pp. 410–420.

О потенциале кластерных схем синтеза оптимальных расписаний для моделей воднотранспортной логистики

Резников Михаил Борисович[1]⋆ mirekez@gmail.com
Федосенко Юрий Семенович[1] fds1707@mail.ru
[1]Нижний Новгород, Волжский государственный университет водного транспорта

С задачами поиска решений для оптимизационных моделей приходится постоянно сталкиваться в процессах оперативного планирования и управления воднотранспортной логистикой. В первую очередь это относится к периодам северного завоза, обеспечения нефтепродуктами социальной, хозяйственной и производственной инфраструктуры полярных регионов, транспортировки минерально-строительных материалов, добываемых на русловых месторождениях внутренних водных путей РФ.

В условиях различных технологических особенностей, ограниченности навигационных периодов и значительных объемов предъявляемых к перевозке грузов задачи повышения эффективности использования крупного парка существенно различных по техническим и экономическим параметрам грузовых судов и танкерного флота объективно требуют привлечения для решения эксплуатационных задач адекватных математических моделей и разработки быстродействующих алгоритмов синтеза проектов управляющих решений.

Все известные оптимизационные постановки в рассматриваемой прикладной области относятся к классу NP-трудных. Данное обстоятельство в ряде случаев является существенным сдерживающим фактором. Ниже в рамках канонической задачи диспетчеризации [1] рассматриваются потенциальные возможности и ограничения кластерных схем реализации алгоритма динамического программирования (ДП) для синтеза оптимальных решений.

Рассматривается n-элементный детерминированный поток Z независимых объектов z_1, z_2, …, z_n. Каждый объект z_i, $i = \overline{1,n}$ подлежит однофазному однократному обслуживанию стационарным процессором P и характеризуется целочисленными параметрами: t_i – момент поступления, τ_i – продолжительность обслуживания, a_i – величина штрафа за единицу времени пребывания в системе обслуживания ($0 \leqslant t_1 \leqslant \ldots\ t_i\ \ldots \leqslant t_n$). В начальный момент времени $t = 0$ процессор P свободен и находится в состоянии готовности к выполнению обслуживания объектов потока Z. Расписание обслуживания потока Z отождествляется с перестановкой $p = (p(1), p(2), \ldots, p(i), \ldots, p(n))$ множества индексов объектов и считается компактным, т.е. момент t'_i начала обслуживания очередного объекта $z_{p(i)}$, определяется соотношениями $t'_1 = t_{p(1)}$, $t'_i = max\{t'_{i-1} + \tau_{p(i-1)}, t_{p(i)}\}$. Момент t'_n завершения обслуживания потока Z определяется как момент завершения обслуживания объекта $z_{p(n)}$. Каноническая задача диспетчеризации заключается в построении расписания p^*, обеспечивающего минимизацию суммарного штрафа по всем объектам потока Z, т.е. $W(p) = \sum_{i=1}^{n} a_{p(i)}(t'_i - t_{p(i)}) \to min$.

Пусть t – момент дискретного времени принятия решения об обслуживании следующего объекта, S – множество ранее обслуженных объектов, $W_k^{min}(t,S)$ – минимальная величина суммарного штрафа при обслуживании потока Z процессором P, освободившимся в момент времени t после обслуживания объекта $z_{p(k)}$ и множества объектов S ($S \in Z$), k – порядковый номер итерации выбора объекта для обслуживания, $t' = \max(t,t_i)$ - время начала обслуживания следующего объекта $z_{p(k+1)}$.Тогда рекуррентные соотношения динамического программирования (ДП) имеют вид $W_k^{min}(t,S) = min_{i=\overline{1,n};z_i \notin S}\ (W_{k+1}^{min}(t' + + \tau_i, S \cup z_i) + a_i(t' - t_i))$.

Решение задачи может быть получено, например, методом обратного прохода алгоритма от частных решений $W_n^{min}([1;T],Z)$, где T - временно́е ограничение на процесс обслуживани. Каждая операция вычисления $W_k^{min}(t,S)$, $k =$ n-1, n-2, n-3, ..., 0 потребует в качестве операндов некоторые ранее сохраненные частные решения $W_{k+1}^{min}(t,S)$. В кластерных архитектурах нахождение значения $W_k^{min}(t,S)$ возможно, если имеется доступ ко всем ранее рассчитанным значениям $W_{k+1}^{min}(t,S)$. Алгоритм параллельного ДП реализуем в виде вычислений набором из M узлов всех частных решений для каждой итерации k от n до 0 и их сохранения в оперативной памяти в виде таблицы

Для равномерного распределения задач по однотипным кластерным ресурсам все состояния (t,S) системы обслуживания на фиксированной итерации k вычислительного процесса были пронумерованы. С этой целью множество S синтезировалось в виде битовой маски $B_S^k = (b_1,\ b_2,\ \ldots\ ,\ b_n)$, где каждый бит отвечает за наличие соответствующего объекта в множестве ранее обслуженных объектов. Начиная с набора из k единиц и $n-k$ нолей, производились перестановки битов для получения всевозможных битовых масок, одновременно с этим давая последовательный номер каждой. Описанный способ нумерации $N(S)$ позволяет разделить множество состояний (задач) (t,S) на равные части, используя для определения номера узла m формулу вида $m = N(S) mod M$.

Вычисленными экспериментами подтверждена возможность равномерной балансировки задач по узлам кластера, однако, одновременно с этим отмечена и высокая потребность кластерной схемы в сетевом ресурсе для обмена данными между узлами. При максимальном числе возможных значений S на итерации $k = n/2$ равном $C_n^{n/2}$ между узлами будет суммарно передано порядка $(n/2)C_n^{n/2}$ частных решений задачи, т.е. используемый объем сетевого трафика значительно превышает объем требуемой алгоритмом оперативной памяти. Соответственно, основным способом снижения продолжительности синтеза решения кластерным алгоритмом ДП является снижение требований вычислительной схемы к ресурсам обмена данными между узлами. В качестве способа понижения требований алгоритма к сетевому ресурсу сформулирован альтернативный принцип балансировки вычислительных задач по узлам кластера: $m = B_S^k mod M$.

Анализ альтернативного способа балансировки задач по кластеру показывает, что на всем диапазоне размерностей потока объектов n от 28 до 48 может

быть подобран такой размер кластера, при котором накладные расходы на сетевой обмен данными составят от 10 до 30% по сравнению с равномерным способом балансировки. В то же время с ростом размера кластера накладные расходы возрастают, и для значений $n < 50$ значительного снижения этих расходов можно добиться только для кластеров размером не более 1024 узлов. Данные выводы подтверждают полученные при моделировании алгоритма ограничения на максимальное число узлов суперкомпьютера при решении задачи.

[1] *Коган Д. И.*, *Пушкин А. М.*, *Дуничкина Н. А.*, *Федосенко Ю. С.* Задачи диспетчеризации обслуживания стационарных объектов в одномерной рабочей зоне процессора // Автоматика и телемеханика, 2016. No 4. С. 67–83.

About potential of cluster schemas for optimal schedule synthesis for models of water transport logistics

Mikhail Reznikov[1]⋆ mirekez@gmail.com
Yuriy Fedosenko[1] fds1707@mail.ru
[1]Nizhny Novgorod Volga State University of Water Transport

Processes of operational planning and water transport logistics control are usually connected to problems of solution finding to optimization models. Firstly, it is related to periods of Northern Delivery, petroleum products supply for economics and industrial infrastructure of Polar Regions, minerals and construction materials transportation, which are mined in riverbed deposits of inner waterways of Russian Federation.

In conditions of different technological specifics, navigation period limitations, significant volumes of cargos requested for transportation, problems of efficiency improvement for large vehicles and tankers, which are different by technical and economical parameters, objectively require development and usage of fast algorithms for control projects solution synthesis.

All known optimizations in the considered application domain are belong to NP-difficult class. This fact is a significant constraint in many cases. Below in the terms of canonical problem of dispatching [1] potential features and limitations of cluster systems are considered for implementation of a dynamic programming algorithm of optimal solution synthesis.

Lets consider an n-element determined flow Z of independent objects z_1, z_2, …, z_n. Each object z_i is intended for one-phase one-time service by stationary processor P and characterized by the following integer parameters: t_i – moment of readiness for service, τ_i – service duration, a_i – penalty value per time unit object being in service system ($0 \leqslant t_1 \leqslant \ldots \ t_i \ \ldots \ \leqslant t_n$). In the start moment of time $t = 0$ processor P is free and ready to start servicing of objects of the flow Z. A schedule of servicing of the objects flow Z is considered as permutation of object indexes $p = (p(1), p(2), \ldots, p(i), \ldots, p(n))$ and compact, i.e. moment t'_i of next object $z_{p(i)}$ servicing, is defined by $t'_1 = t_{p(1)}$, $t'_i = max\{t'_{i-1} + \tau_{p(i-1)}, t_{p(i)}\}$. Moment t'_n of an objects flow Z service finishing is defined as moment when object $z_{p(n)}$ servicing is finished. The canonical problem of dispatching is to synthesize schedule p^* of servicing, which provides minimization of overall penalty for all objects of the flow Z, i.e. $W(p) = \sum_{i=1}^{n} a_{p(i)}(t'_i - t_{p(i)}) \to min$.

Let t be a moment of discrete time when decision is made about next object for service, S ($S \in Z$) is a set of already serviced objects to this moment, $W_k^{min}(t, S)$ – minimal possible overall penalty for objects servicing by processor P since it finished servicing of some object $z_{p(k)}$ at moment of time t, k is serial number of iteration of decision making. Then the dynamic programming (DP) equation will be $W_k^{min}(t, S) = min_{i=\overline{1,n};z_i \notin S} \ (W_{k+1}^{min}(t' + \tau_i, S \cup z_i) + a_i(t' - t_i))$.

The problem solution can be found by using reverse passage algorithm starting at states $W_n^{min}([1;T],Z)$ where T is a maximum allowed time limit for work. Each operation of calculation of $W_k^{min}(t,S)$, $k = n$-1, n-2, n-3, ..., 0 will require some pre-calculated partial solutions $W_{k+1}^{min}(t,S)$ as operands. In cluster architectures the $W_k^{min}(t,S)$ value can be found only if there is an access to all pre-calculated values $W_{k+1}^{min}(t,S)$. DP algorithm is implemented as calculation of all partial solutions on each iteration k from n to 0 by a set of M nodes and then saving it to a table in memory.

For balanced distribution over all cluster nodes all states (t,S) of servicing system was enumerated on the iteration k. For this the S set was synthesized as a bit mask $B_S^k = (b_1, b_2, \ldots, b_n)$, where each bit is responsible for respective object in a set of already serviced objects. Starting with a combination of k ones and $(n$-$k)$ zeroes, the permutations of bits were done to form each possible bit mask, with this providing each mask with its serial number. The described enumeration method $N(S)$ allows to balance all states (tasks) (t,S) to equal subsets, using the following equation to calculate node number: $m = N(S) mod M$.

By computational experiments, the possibility of uniform balancing of tasks over cluster nodes was proved, but with this, the notable need in network intercommunication resource by the cluster schema was observed. For the maximum possible different S sets number for an iteration $k = n/2$ is $C_n^{n/2}$, there will be transmitted about $(n/2)C_n^{n/2}$ partial solutions of the problem, i.e. network traffic volume significantly exceeds overall used memory volume. Therefore the main way for reducing synthesis time for DP cluster algorithm is reducing of calculation schema requirements for intercommunication resource. As one of the ways for network resource usage reduction, the alternative method of task balancing over cluster nodes was suggested: $m = B_S^k mod M$.

Analysis of this alternative balancing schema shows that for any objects flow size n in a range from 28 to 48 there can be found such cluster size that network usage overhead can be reduced to values from 10 to 30% in comparison to uniform balancing method. With this with cluster size growth the overhead will increase and for values of problem size $n < 50$ it is possible to significantly reduce it only for clusters of size not more than 1024 nodes. This conclusions prove results which were observed in modelling of algorithm implementation for large supercomputers.

[1] *Kogan D. I.*, *Pushkin A. M.*, *Dunechkina N. A.*, *Fedosenko Yu. S.* Problems of stationary objects service dispatching in one dimension working zone of a processor // Automatics and Telemechanics, 2016. No 4. Pp. 67–83.

Жадный алгоритм решения классической NP-трудной задачи минимизации суммарного запаздывания

Саратов Анатолий Алексеевич[1] ⋆ sapfor58@gmail.com
[1]Город Тула, ООО «ИНТЕРСАП»

Рассматривается NP-трудная задача теории расписаний. Необходимо обслужить множество требований $N = \{1, 2, .., n\}$ на одном приборе. Прерывания и обслуживание более одного требования в любой момент времени запрещены. Для требования $j \in N$ заданы продолжительность обслуживания $p_j > 0,\ p_j \in Z$ и директивный срок окончания обслуживания d_j. Все требования поступают на обслуживание одновременно в момент времени t_0, с которого прибор готов начать обслуживание требований. Требуется построить расписание π^* обслуживания требований множества N, при котором достигается минимум функции $F(\pi) = \sum\limits_{j=1}^{n} \max\{0, c_j(\pi) - d_j\}$, где $c_j(\pi)$ — момент завершения обслуживания требования j при расписании π. Пусть $\pi = (j_1, .., j_n)$, тогда $c_{j_1}(\pi) = t_0 + p_{j_1}$ и $c_{j_k}(\pi) = c_{j_{k-1}} + p_{j_k}$ для $k = 2, 3, .., n$. Величина $T_j(\pi) = \max\{0, c_j(\pi) - d_j\}$ называется запаздыванием требования j при расписании π, а $F(\pi)$ — суммарным запаздыванием требований при расписании π. Если обслуживание требования i предшествует обслуживанию требования j, то для этого будем использовать запись $(i \to j)_\pi$. Предложен метод расчёта смещений частичных расписаний при перестановках требований [1]. Поскольку в оптимальном расписании π^* любое перемещение требования j, и любая парная перестановка требований i и j не может привести к уменьшению $F(\pi)$, то для всех $i, j \in \pi^*$ должно выполняться условие (1):

$$\left\{ \frac{\nabla T_j^{'}(\pi) > \sum\limits_{j+1}^{j+k} \nabla T_i^{'}(\pi),\ j>i}{\nabla T_j^{''}(\pi) < \sum\limits_{j-k}^{j-1} \nabla T_i^{''}(\pi),\ j<i} \right. \quad \text{и} \quad \left\{ \frac{\nabla T_{ij}(\pi) > \sum\limits_{i+1}^{j-1} \nabla T_v^{'}(\pi),\ p_i>p_j}{\nabla T_{ji}(\pi) < \sum\limits_{i+1}^{j-1} \nabla T_v^{''}(\pi),\ p_i<p_j} \right. , \qquad (1)$$

где $\nabla T_j^{'}(\pi), \nabla T_j^{''}(\pi)$ — изменение запаздывания требования j при смещении j в очереди на k позиций, соответственно, вправо и влево; $\sum\limits_{j+1}^{j+k} \nabla T_i^{'}(\pi)$ — уменьшение запаздывания требований $i = \{j+1, ..,\ j+k\} \in N$ при их смещении на k позиций влево;

$$\sum_{i+1}^{i+k} \nabla T_i^{'}(\pi) = \sum_{j+1}^{j+k} \left\{ \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m - d_i\right)\right] - \right.$$

$$\left. - \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m - p_j - d_i\right)\right] \right\}; \quad (2)$$

$\sum\limits_{j-k}^{j-1} \nabla T_i^{''}(\pi)$ — увеличение запаздывания требований $i = \{j-k, .., j-1\} \in N$ при их смещении на k вправо;

$$\sum_{i-k}^{i-1} \nabla T_i^{''}(\pi) = \sum_{j-k}^{j-1} \left\{ \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m + p_j - d_i\right)\right] - \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m - d_i\right)\right] \right\}; \quad (3)$$

$\nabla T_{ij}(\pi)$ — увеличение запаздывания требования i при замене j;
$\nabla T_{ji}^{'}(\pi)$ — уменьшение запаздывания требования j при замене i;
$\sum\limits_{i+1}^{j-1} \nabla T_v^{'}(\pi)$ — уменьшение запаздывания требований $v = \{i+1, .., j-1\}$ при перестановке требований i, j и $p_i > p_j$;
$\sum\limits_{i+1}^{j-1} \nabla T_v^{''}(\pi)$ — увеличение запаздывания требований $v = \{i+1, .., j-1\}$ при перестановке требований i, j и $p_i < p_j$.

Множества перемещаемых требований формул (2),(3) назовём частичными расписаниями π' и π''; $\pi' = \{j+1, ..., j+k\} \in N$, $\pi'' = \{j-k, ..., j-1\} \in N$.

Выражение (1) при $k = 1, 2, .., n-1$, является эффективным инструментом оценки оптимальности построенного расписания, ибо вычисляется за $0\left(n^2\right)$ операций.

Для построения π^* на основе формулы (1) необходим быстрый алгоритм формирования частичных расписаний π' и π'', эквивалентных по соотношению суммарных смещений $\sum\limits_{i+1}^{i+k} \nabla T_i^{'}(\pi)$, $\sum\limits_{i-k}^{i-1} \nabla T_i^{''}(\pi)$ с суммарными смещениями в расписании π^*. Такая эквивалентность может быть достигнута в частичных расписаниях π', π'' отвечающих (1) при $k = 1$. Алгоритм А построения π' имеет вид:

1. Принимаем время старта $t = t_0$.

2. Попарно переставляем требования $i \to j, j \to i$ и вычисляем сумму $\nabla T_{ij} = T_i(\pi) + T_j(\pi)$. Если при $i \to j$ суммарное смещение меньше, чем при $j \to i$, то i помечается, как приоритетное требование i^*, и продолжает участвовать в перестановках, а j исключается . Если суммарные смещения для $i \to j$ и $j \to i$ равны, то помечается требование с меньшим d_j.

3. Исключаем i^*из N и помещаем в N^*, добавляя справа.

4. Принимаем время старта $t = t_0 + p_i$.

5. Повторяем шаги 2-4, пока $N \neq 0$.

6. Корректируем π'на соответствие (1).

Обозначим как π_j расписание, состоящее из π' и следующего за ним требования j, а смещение требований π_j как $F(\pi_j)$.

Если принять условие (1) достаточным для проверки оптимальности π', то в расписании $\pi_j = (\pi', j)$ с минимальным суммарным смещением $F(\pi_j)$ требование $j \in \pi_j$ будет последним в оптимальном расписании π^*. Исходя из этого, алгоритм построения π^* имеет вид:

1. Для каждого требования — перемещаем в конец очереди. Остальные работы упорядочиваем, согласно алгоритму А. Вычисляем суммарное смещение $F(\pi)$.
2. Выбираем требование i с минимальным $F(\pi)$.
3. Исключаем i из N и помещаем в список требований N^*, добавляя слева.
4. Повторяем шаги 1, 2, 3, пока в N останутся только успевающие работы.
5. Объединяем N и N^*.
6. Конец.

[1] *Саратов А. А.* Конкурентный метод синтеза производственных расписаний // Изв. ТулГУ. Технические науки, 2014. № 3. С. 104–110.

Greedy algorithm for solving the NP-hard problem of minimizing total tardiness for a single machine

Anatoly Saratov[1] sapfor58@gmail.com
[1]Tula, "INTERSAP" Ltd.

Considered classical the NP-Hard problem of minimizing total tardiness for a single machine. The interruptions when servicing and service more one requirements any time prohibited. Length of the service $p_j > 0,\ p_j \in Z$ and deadline $d_j$of the completion of the service is given for requirement $j \in N = \{1, 2, .., n\}$. All requirements enter on service at moment of time t_0 simultaneously, with which instrument ready to begin servicing the requirements. It Is Required build scheduling of the servicing the requirements ensemble N, under which is reached minimum to functions $F(\pi) = \sum\limits_{j=1}^{n} \max\{0, c_j(\pi) - d_j\}$, where $c_j(\pi)$- a moment of the termination of the servicing the requirement j at timetable π. Let $\pi = (j_1, .., j_n)$, then $c_{j_1}(\pi) = t_0 + p_{j_1}$ and $c_{j_k}(\pi) = c_{j_{k-1}} + p_{j_k}$ for $k = 2, 3, .., n$. The value $T_j(\pi) = \max\{0, c_j(\pi) - d_j\}$ is identified tardiness requirements j at timetable π, and $F(\pi)$- total tardiness requirements. If servicing the requirement i precedes servicing the requirement j then for this shall use record $(i \to j)_\pi$. For estimation of the reductions of tardiness requirement consecutively move in the end local optimized partial timetables [1]. Since in optimum timetable any displacement requirement j, and any fresh transposition of the requirements i, j cannot bring about reduction then for all must be executed condition (1):

$$\left\{ \frac{\nabla T_j'(\pi) > \sum\limits_{j+1}^{j+k} \nabla T_i'(\pi),\ j>i}{\nabla T_j''(\pi) < \sum\limits_{j-k}^{j-1} \nabla T_i''(\pi),\ j<i} \right. \quad \text{and} \quad \left\{ \frac{\nabla T_{ij}(\pi) > \sum\limits_{i+1}^{j-1} \nabla T_v'(\pi),\ p_i>p_j}{\nabla T_{ji}(\pi) < \sum\limits_{i+1}^{j-1} \nabla T_v''(\pi),\ p_i<p_j} \right., \qquad (1)$$

where $\nabla T_j'(\pi), \nabla T_j''(\pi)$ — change of tardiness when moving j on k positions, accordingly, to the right and to the left;

$\sum\limits_{j+1}^{j+k} \nabla T_i'(\pi)$ — total reduction of tardiness $i = \{j+1, .., j+k\} \in N$ under their displacement in queue on k positions to the left;

$$\sum_{i+1}^{i+k} \nabla T_i'(\pi) = \sum_{j+1}^{j+k} \left\{ \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m - d_i\right)\right] - \right.$$

$$\left. - \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m - p_j - d_i\right)\right] \right\}; \qquad (2)$$

$\sum_{j-k}^{j-1} \nabla T_i^{''}(\pi)$ — total increase of tardiness requirements $i = \{j-k, .., j-1\} \in N$ under their displacement in queue on k positions to the right;

$$\sum_{i-k}^{i-1} \nabla T_i^{''}(\pi) = \sum_{j-k}^{j-1} \left\{ \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m + p_j - d_i\right)\right] - \right.$$

$$\left. - \max\left[0, \left(t_0 + \sum_{m=0}^{i} p_m - d_i\right)\right] \right\}; \qquad (3)$$

$\nabla T_{ij}(\pi)$ — increase of tardiness requirements i under transposition j;

$\nabla T_{ji}^{'}(\pi)$ — reduction of tardiness requirements j under transposition i;

$\sum_{i+1}^{j-1} \nabla T_v^{'}(\pi)$ — total reduction of tardiness requirements $v = \{i+1, .., j-1\}$ after permutation i and j, $p_i > p_j$;

$\sum_{i+1}^{j-1} \nabla T_v^{''}(\pi)$ — total increase of tardiness requirements $v = \{i+1, .., j-1\}$ after permutation i and j, $p_i < p_j$.

The ensemble of the movable requirements molded (2), (3) shall name the partial timetables and π' and π''; $\pi' = \{j+1, ..., j+k\} \in N$, $\pi'' = \{j-k, ..., j-1\} \in N$.

The Expression (1) under $k = 1, 2, .., n-1$, is an efficient instrument of the estimation built timetables, is since calculated in $O\left(n^2\right)$ operations. For ensuring the efficient process of the syntheses π^* on base of the formula (1) necessary quick algorithm shaping the partial timetables $\pi'$and π'', equivalent on correlation of the total offsets $\sum_{i+1}^{i+k} \nabla T_i^{'}(\pi)$, $\sum_{i-k}^{i-1} \nabla T_i^{''}(\pi)$ with total tardiness partial timetables π^*. Such equivalence can be reached in partial timetable π', π'' answering (1) under $k = 1$. Sequence of jobs $j \in \pi'$ is defined by correlations d_j,p_j, so timetables meets condition (1) under shall name local optimum. Algorithm A buildings π' are of the form of:

1. Take time of the start $t = t_0$.
2. Prototype two by two transpositions of the requirements $i \to j, j \to i$ and calculate $\nabla T_{ij} = T_i(\pi) + T_j(\pi)$. If under $i \to j$ total offset less, than under $j \to i$, that i is marked as priority requirement i^*, and continues to participate in transposition, but j is excluded. If total offsets for $i \to j$ and $j \to i$ alike, that is marked requirement with smaller d_j
3. Exclude i^* from N and place in N^* , adding list on the right.
4. Take time of the start $t = t_0 + p_i$.
5. Repeat the steps 2-4 while $N \neq 0$.
6. The Adjustment π' on correspondence to (1).

We shall mark as π_j timetable, consisting of π' and the following for its requirements j, and total offset of the requirements π_j as $F(\pi_j)$. If take the condition (1) sufficient for checking optimality π' then in timetable $\pi_j = (\pi', j)$ with minimum total offset $F(\pi_j)$ of the requirements $j \in \pi_j$ will be last in optimum timetable π^*. Coming thereof, algorithm of the building is of the form of:

1. For each requirement — move in the end queue. Rest work regularize, according to algorithm A. Calculate total tardiness $F(\pi)$.

2. Choose work i with minimum $F(\pi)$.

3. Exclude i from N and place in list of the last requirements N^*, adding list on the left.

4. Repeat the steps 1, 2, 3 while in N will remain only having time to work.

5. Unite N and N^*.

6. The end.

[1] *Saratov A. A.* Competitive method for the synthesis of production schedules // Izv. TulSU. Technical science, 2014. No 3. Pp. 104–110.

Разработка общедоступного набора видеоданных в терагерцовом диапазоне и программной платформы для экспериментов с интеллектуальным видеонаблюдением в терагерцовом диапазоне

Морозов Алексей Александрович* morozov@cplire.ru
Сушкова Ольга Сергеевна o.sushkova@mail.ru

Москва, ИРЭ им. В.А. Котельникова РАН

Разработаны общедоступный набор видеоданных в терагерцовом диапазоне и программная платформа для экспериментов с интеллектуальным видеонаблюдением в терагерцовом диапазоне. Набор видеоданных включает короткие видеоролики людей с объектами, спрятанными под одеждой. Набор данных является мультимодальным, то есть он содержит синхронизированные видео различных типов: терагерцовые, тепловизионные, RGB, 3D и в ближнем инфракрасном диапазоне. Специальная программная платформа разработана для сбора и предварительной обработки видеоданных. Программная платформа включает в себя транслятор Акторного Пролога в Java и библиотеку с открытым исходным кодом встроенных классов для сбора и обработки видеоданных. В частности, программа позволяет проецировать терагерцовые и тепловизионные видеоданные на трёхмерные облака точек с использованием трёхмерных таблиц соответствий. Описан эксперимент по обучению CNN различных архитектур на наборе терагерцовых видеоданных.

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 16-29-09626 и № 18-07-01295
(www.fullvision.ru).

[1] *Morozov A. A., Sushkova O. S.* Development of a Publicly Available Terahertz Video Dataset and a Software Platform for Experimenting with the Intelligent Terahertz Visual Surveillance // Advances in Intelligent Systems and Computing, 2020. Vol. 1255. Pp. 105–113.

Development of a Publicly Available Terahertz Video Dataset and a Software Platform for Experimenting with the Intelligent Terahertz Visual Surveillance

Alexei Morozov⋆ morozov@cplire.ru
Olga Sushkova o.sushkova@mail.ru
Moscow, Kotel'nikov IRE RAS

A publicly available terahertz video dataset and a software platform for experimenting with the terahertz intelligent video surveillance are developed. The video dataset includes short videos of people with objects hidden under the clothing. The dataset is multimodal, that is, it contains synchronized videos of various kinds: terahertz, thermal, visible, near-infrared, and 3D. A special software platform is developed for the acquisition and preprocessing of the video data. The software platform includes a translator of the Actor Prolog language to Java and an open-source library of built-in classes for data acquisition and processing. In particular, the software enables one to project terahertz/thermal video data onto three-dimensional point clouds using 3D lookup tables. An experiment with the terahertz video data analysis based on various CNN architectures is described.

This research was supported by the Russian Foundation for Basic Research, projects 16-29-09626 and 18-07-01295 (www.fullvision.ru).

[1] *Morozov A. A., Sushkova O. S.* Development of a Publicly Available Terahertz Video Dataset and a Software Platform for Experimenting with the Intelligent Terahertz Visual Surveillance // Advances in Intelligent Systems and Computing, 2020. Vol. 1255. Pp. 105–113.

Оптимизация работы системы слежения, основанной на сети камер видеонаблюдения

Чигринский Виктор Владимирович[1]⋆ chigrinskiy.viktor@phystech.edu
Матвеев Иван Алексеевич[2] matveev@ccas.ru
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Отслеживание перемещения объектов на видеопоследовательностях — важная задача компьютерного зрения, имеющая широкое практическое применение. Ключевыми моментами работы систем слежения являются обнаружение объекта и, в случае если он был обнаружен не в первый раз, его повторная идентификация. Представлена быстро и корректно работающая система слежения, использующая несколько камер и включающая в себя: детектирование и сегментацию объектов на изображении, получение дескрипторов их внешнего вида, сравнение каждого нового объекта с уже накопленными, принятие решения о повторной идентификации. Реализована базовая конфигурация системы, в которой в качестве составляющих используются лучшие на текущий момент алгоритмы детектирования и модели получения дескрипторов внешнего вида. На этой основе произведены модификации как отдельных модулей, так и всей системы в целом. Выполнен вычислительный эксперимент, количественно подтверждающий преимущество доработанной системы относительно базовой.

Работа поддержана грантом РФФИ № 19-07-01231.

[1] *Чигринский В. В., Матвеев И. А.* Оптимизация работы системы слежения, основанной на сети камер видеонаблюдения // Известия РАН. Теория и системы управления, 2020. № 4. С. 110–124.

Optimization of the multi-target multi-camera tracking system

Victor Chigrinsky[1]⋆ chigrinskiy.viktor@phystech.edu
Ivan Matveev[2] matveev@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

Tracking the motion of objects in video sequences is an important problem of computer vision that has a wide range of applications. The key points in tracking systems is the detection of an object and, if it was detected repeatedly, its reidentification. A fast correctly working tracking system that uses a number of cameras is described. The system includes detection and segmentation of objects in images, construction of their appearance descriptors, comparison of each new object with earlier collected objects, and making a decision about their reidentification. The basic system configuration is implemented in which the state-of-the art detection algorithms and models for constructing the appearance descriptors are used as the constituent parts. Based on this, the system as a whole and some of its modules are modified. A computational experiment that quantitatively confirms the advantages of the modified system over the basic system is performed.

This research is funded by RFBR, grant 19-07-01231.

[1] *Chigrinsky V., Matveev I.* Optimization of a tracking system based on a network of cameras // Journal of Computer and Systems Sciences International, 2020. Vol. 59. No 4. Pp. 583–597.

Слежение за множеством объектов на видео изображениях с помощью ре-идентификации с предфильтрацией дескрипторов по качеству

Григорьев Алексей Дмитриевич[1]⋆ grigorev.ad@phystech.edu
Гнеушев Александр Николаевич[2] gneushev@ccas.ru
[1]Москва, Московский физико-технический институт (НИУ)
[2]Москва, Федеральный исследовательский центр "Информатика и управление" РАН

Данная работа посвящена задаче слежения за множеством объектов в видео потоке изображений. Существующие решения являются либо ресурсоемкими, либо неустойчивыми к большой плотности объектов, что приводит к срывам слежения и накоплению ошибок. В работе формулируется задача сопровождения множества объектов на основе факторизации апостериорного распределения параметров объектов при допущении линейности и независимости движения. Решение сводится к фильтру Калмана и подзадаче о назначениях измерений и объектов. Для увеличения устойчивости назначений вводятся поправки к коэффициентам матрицы стоимости с помощью ре-идентификации, сопоставления дескрипторов областей изображения с историей вдоль траекторий. Для уменьшения вычислительной сложности предлагается использовать предварительный отбор дескрипторов вдоль траектории на основе показателя "качества" , который позволяет исключить из сопоставления дескрипторы с низкой информативностью и объекты, не являющиеся предметом наблюдения. Для оценки параметра "качества" использовались как известные биометрические подходы, так и показатель уверенности детектора.

Вычислительный эксперимент, проведенный на выборках MOT20-01 и MOT20-02, показал высокую вычислительную эффективность предложенных методов и увеличение устойчивости слежения. Проанализировано влияние числа n лучших отбираемых дескрипторов вдоль траектории для ре-идентификации.

Работа поддержана грантом РФФИ 19-07-01231.

[1] *Григорьев А. Д., Гнеушев А. Н.* Ре-идентификация с предфильтрацией по качеству изображений в задаче слежения за множеством объектов // Информационные технологии, 2021.

Multiple object tracking on video via re-identification with descriptor pre-filtering based on quality assessment

Alexey Grigorev[1]⋆ grigorev.ad@phystech.edu
Alexander Gneushev[2] gneushev@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS

The work solves the problem of multiple object tracking. The existing methods are either resource-intensive or not resistant to high object densities, this leads to tracking disruptions and error accumulation. We formulate the problem of multiple object tracking based on the factorization of the posterior distribution of objects' parameters under the assumption of linearity and independence of objects' motion. The final solution contains two parts: the Kalman filter and the assignments problem between measurements and objects. We introduce corrections to the coefficients of the cost matrix via re-identification comparing the descriptors of the image parts with the history along the trajectories in order to increase the stability of assignments. The preliminary selection of descriptors along the trajectory based on the parameter of "quality" is proposed to decrease the computational complexity of the algorithm. Such selection allows us to exclude objects with low information usefulness and objects unrelated to the subject of observation from re-identification. Both known biometric approaches and an alternative method based on the detector confidence are considered as methods of quality assessment.

Computational experiments were conducted on MOT20-01 and MOT20-02 datasets using detectors of various complexity and showed high computational efficiency and tracking stability of the proposed methods. The influence of the number n of the best descriptors selected along the trajectory for re-identification was analyzed.

This research is funded by RFBR, grant 19-07-01231.

[1] *Grigorev A., Gneushev A.* Re-identification with pre-filtering by image quality for multiple objects tracking // Information Technologies, 2021.

Метод размещения иммунных детекторов на основе оценки риска безопасности сетевых узлов

Сычугов Алексей Алексеевич* xru2003@list.ru
Токарев Вячеслав Леонидович tokarev22@yandex.ru
[1]Тула, Тульский государственный университет

Одним из наиболее эффективных средств своевременного обнаружения атак в компьютерных сетях являются системы обнаружения вторжений (СОВ), построенные на основе иммунных детекторов [1], позволяющие обнаруживать атаки различных классов, включая ранее неизвестные.

Однако важное значения для эффективности применения таких СОВ имеет рациональное размещение иммунных детекторов по узлам сети. Показано [2], что состав и размещение иммунных детекторов тогда позволяют достичь наибольшего эффекта, когда они контролируют узлы со сравнительно высоким риском нарушения информационной безопасности. Источник риска может быть многоступенчатой, многовариантной уязвимостью и охватывать несколько узлов, что значительно осложняет оценку риска безопасности узлов. Для преодоления этой сложности, предложено использовать статистическую формальную модель на основе Марковских цепей в сочетании с метриками анализа уязвимостей, что позволяет определить критические узлы, в которых нарушители могут быть наиболее опасны. Основываясь на получаемой с помощью модели информации, сетевой администратор может именно на этих узлах установить иммунные детекторы, что позволит существенно улучшить систему защиты.

В качестве оценок уязвимостей используются скоринговые оценки CVSS [3], которые используют три вида метрик: базовые, временные и контекстные.

Нарушители обычно проникают в компьютерные сети с помощью цепочки эксплойтов, каждый элемент которой создает основу для следующего элемента. Сочетание таких эксплойтов составляет цепочку, называемую траекторией атаки, совокупность которых образуют граф возможных траекторий (ГВТ) заканчивающийся в состоянии, где нарушитель может успешно достичь своей цели. Существует ряд алгоритмов, которые были разработаны для построения ГВТ атак [4, 5, 6]. Однако анализ сети с помощью ГВТ может оказаться нетривиальной задачей.

Для построения формальной модели доступа к узлу предлагается использовать Марковские цепи [7], отражающие реальное поведение атакующего в том смысле, что нарушитель может использовать разные траектории (последовательность узлов) до достижения цели-узла.

Предполагается: 1) выбор наилучшего промежуточного узла зависит от трех факторов, а именно: эксплойтности, характеризующей уязвимости подсистемы доступа; влияния уязвимостей на нарушения конфиденциальности, целостности и доступности, а также индивидуального навыка атакующего; 2) переходные состояния не зависят от времени; 3) может быть определена некоторая матрица

вероятностей перехода P(x, y) и начальное распределение вероятностей $R = \{ r_1, r_2, ..., r_n \}$.

Тогда, имея матрицу P(x, y), вектор начальных рисков R, используя основные свойства Марковского процесса можно определить риски узлов и риск всей сети.

Основным компонентом предлагаемой модели является ГВТ, который строится путем изучения топологии сети, служб, запущенных на каждом узле, правил, определенных на брандмауэрах, и уязвимостей, связанных с каждым узлом, на котором запущены различные службы.

Предполагается, что:

1) в рассматриваемой сети присутствует ограниченное число узлов, каждый из которых запускает различные виды услуг и там же могут существовать различные уязвимости, для которых определены CVSS - системой соответствующие баллы, которыми можно пометить ребра ГВТ, используемого для определения вероятности использования нарушителем i-й уязвимости;
2) нарушитель выберет уязвимость, которая максимизирует шансы успеха в компрометации состояния узла-цели;
3) если нарушитель, по какой-либо причине, завершает атаку, то он перейдет в исходное состояние.

Центральная составляющая предлагаемой модели - ГВТ доступа к узлу.

Нарушитель может атаковать узел, к которому имеет непосредственный доступ, и, преодолев защиту этого узла, развивает атаки, пока не достигнет цели. При этом перед ним возникает задача выбора следующего узла для атаки. Этот выбор зависит от двух параметров: Exp, характеризующий сложность преодоления защиты узла, и Imp, характеризующий уязвимости узла.

Принимая окончательное решение о переходе с одного узла на другой, атакующий опирается также и на собственные навыки и опыт. Это субъективный фактор, влияющий на выбор нарушителем очередного узла для атаки, предлагается учесть в модели как фактор смещения. В результате функцию выбора можно представить следующим выражением

$$a_{jk} = \beta \cdot Exp(\vartheta_k) + (1 - \beta) \cdot Imp(\vartheta_k) \quad (1)$$

где a_{jk} - это “выгода” от перемещения от узла j до узла k, v_k - функция уязвимости, значение которой характеризует возможность преодоления защиты k-го узла нарушителем; - коэффициент смещения, принимающий значение от 0 до 1.

Если значения a_{jk} определить для каждой пары узлов сети, то ее защищенность от атак, с точки зрения нарушителя, можно охарактеризовать матрицей смежности:

$$A = \begin{bmatrix} 0 & a_{01} & \dots & a_{0g} & \dots & a_{0n} \\ a_{10} & 0 & \dots & a_{1g} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n0} & a_{n1} & \dots & a_{ng} & \dots & 0 \end{bmatrix} \tag{2}$$

Нормализация значений a_{jk} матрицы A позволяет получить значения функции принадлежности нечеткому множеству «Узел Sk доступен для атаки из узла Sj»:

$$\mu_{jk} = \frac{a_{jk}}{\sum_i a_{ji}} \tag{3}$$

Тогда характеристика сети в матричном виде:

$$M = DA \tag{4}$$

где M - матрица переходов, определяющая возможность перехода нарушителя из одного узла в другой, D-диагональная матрица, вычисленная с помощью правила нормализации

$$d_{jk} = \begin{cases} 1/\sum_i a_{ij} & j = k \\ 0 & otherwise \end{cases} \tag{5}$$

Анализ риска основан на относительном значении ранга для каждого узла ГВТ. Начальное значение вектора риска вычисляется на основе количества узлов, присутствующих в ГВТ. Если в ГВТ существует n узлов, то можно установить все ранги узлов равные 1/n.

Величина риска rk для узла k вычисляется с помощью итерационной процедуры до получения устойчивого значения:

$$r_k(t) = \sum_j r_k(t-1)\mu_{jk} \tag{6}$$

Эта итерация сходится к устойчивому значению вектора R*, как собственному вектору матрицы M.

Метод размещения иммунных детекторов включает следующие этапы.

Этап 1. Инициализация. Каждому значению вектора рисков присваивается начальное значение 1/n.

Этап 2. Итерационная процедура коррекции значений рисков до наступления условия сходимости для каждой вершины ГВТ.

Этап 3. Размещение иммунных детекторов в узлах сети с наибольшим значением рисков.

Этап 4. Определение суммы рисков как общего показателя информационной безопасности сети.

Предложенный метод может быть использован для построения системы защиты информации автоматизированных систем.

[1] *Токарев В. Л., Сычугов А. А.* Обнаружение вредоносного программного обеспечения с использованием иммунных детекторов // Известия Тульского государственного университета. Технические науки, 2017. № 10. С. 216–230.

[2] *Tokarev V. L., Sychugov A. A.* Multi-Agent system for network attack detection // International Journal of Civil Engineering and Technology, 2018. Vol. 9. No 6.

[3] Банк данных угроз безопасности информации. Калькулятор CVSS v2. // `https://bdu.fstec.ru/calc`.

[4] *Токарев В. Л.* Распознавание стратегии противодействующей стороны по текущим наблюдениям // Доклады Томского государственного университета систем управления и радиоэлектроники, 2014. С. 184–187.

[5] *Jha S., Sheyner O., Wing J.* Two Formal Analyses of Attack Graphs // Proceedings of 15th IEEE Computer Security Foundations Workshop, 2002. Pp. 49–63.

[6] *Mehta V., Bartzis C., Zhu H., Clarke E, Wing J.* Ranking Attack Graphs. // International Workshop on Recent Advances in Intrusion Detection, 2006. Pp 127–144.

[7] *Дынкин Е. Б.* Основания теории марковских процессов // ФИЗМАТЛИТ, 2006. С. 228.

Method of immunne detectors placement based on network node security risk assessment

Alexander Sychugov* xru2003@list.ru
Alexander Anchishkin alexanderanchishkin@yandex.ru
Tula, Tula State University

One of the most effective means of timely detection of attacks in computer networks is the intrusion detection systems (IDS) based on immune detectors [1], which allow detecting attacks of various classes, including previously unknown ones.

However, the rational placement of immune detectors on network nodes is important for the effectiveness of the use of such systems. It is shown [2] that the composition and placement of immune detectors then allow achieving the greatest effect when they control nodes with a relatively high risk of information security violations. The source of the risk can be a multi-stage, multi-variant vulnerability and cover several nodes, which significantly complicates the assessment of the node security risk. To overcome this complexity, it is proposed to use a statistical formal model based on Markov chains in combination with vulnerability analysis metrics, which allows us to determine the critical nodes where violators can be most dangerous. Based on the information obtained using the model, the network administrator can install immune detectors on these nodes, which will significantly improve the protection system.

CVSS scoring is used as vulnerability estimates [3], which use three types of metrics: basic, temporal, and contextual.

Intruders usually break into computer networks using a chain of exploits, each element of which creates the basis for the next element. The combination of such exploits makes up a chain called an attack trajectory, which together form a graph of possible trajectories (GPT) ending in a state where the intruder can successfully achieve his goal. There are a number of algorithms that have been developed for constructing the GPT of intrusions [4, 5, 6]. However, network analysis using GPT can be a non-trivial task.

To build a formal model of access to a node, it is proposed to use Markov chains [7], which reflect the real behavior of the attacker in the sense that the intruder can use different trajectories (a sequence of nodes) before reaching the target node.

It is suggested that: 1) selection of the best relay node depends on three factors, namely: exploitati characterizing the vulnerability of the subsystem access; the impact of vulnerabilities to the confidentiality, integrity and availability, as well as the individual skill of the attacker; 2) the transients do not depend on time; 3) some matrix of transition probabilities P(x, y) and the initial probability distribution can be determined.

Then, having a matrix P(x, y), a vector of initial risks R, using the basic properties of the Markov process, we can determine the risks of nodes and the risk of the entire network.

The main component of the proposed model is the GPT, which is built by examining the network topology, the services running on each node, the rules defined on firewalls, and the vulnerabilities associated with each node running various services.

It is supposed that:

1) in the considered network there is a limited number of nodes, each of which runs different types of services and there can be multiple vulnerabilities for which CVSS system sets relevant points, which can mark the edges of GPT used to determine the probability of use of the i-th vulnerability by the infringer;
2) the intruder will choose a vulnerability that maximizes the chances of success in compromising the state of the target node;
3) if the intruder, for any reason, completes the attack, it will return to its original state.

The central component of the proposed model of the GPT access to the node. An intruder can attack a node that he has direct access to, and after overcoming the protection of this node, develops the attacks until they reach the target. In this case, he is faced with the task of selecting the next node to attack. This choice depends on two parameters: Exp, which characterizes the difficulty of overcoming node protection, and Imp, which characterizes node vulnerabilities.

When making the final decision to move from one node to another, the attacker also relies on his own skills and experience. This is a subjective factor that affects the violator's choice of the next node to attack, which is proposed to be taken into account in the model as a bias factor . As a result, the selection function can be represented by the following expression

$$a_{jk} = \beta \cdot Exp(\vartheta_k) + (1 - \beta) \cdot Imp(\vartheta_k) \tag{7}$$

where a_{jk} – is the "benefit" of moving from node j to node k, v_k - vulnerability function, the value of which characterizes the ability of the intruder to overcome the protection of the k-th node; - offset coefficient, which takes a value from 0 to 1.

If the a_{jk} values are determined for each pair of network nodes, then its protection from attacks, from the point of view of the intruder, can be characterized by the adjacency matrix:

$$A = \begin{bmatrix} 0 & a_{01} & \dots & a_{0g} & \dots & a_{0n} \\ a_{10} & 0 & \dots & a_{1g} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n0} & a_{n1} & \dots & a_{ng} & \dots & 0 \end{bmatrix} \tag{8}$$

Normalization of the a_{jk} values of matrix A allows us to get the values of the fuzzy set membership function " Node S_k is available for attack from node S_j »:

$$\mu_{jk} = \frac{a_{jk}}{\sum_i a_{ji}} \tag{9}$$

Then the characteristic of the network in matrix form is

$$M = DA \tag{10}$$

where M is the transition matrix that determines whether the intruder can move from one node to another, and D is the diagonal matrix calculated using the normalization rule

$$d_{jk} = \begin{cases} 1/\sum_i a_{ij} & j = k \\ 0 & otherwise \end{cases} \tag{11}$$

The risk analysis is based on the relative rank value for each GPT node. The initial value of the risk vector is calculated based on the number of nodes present in the GPT. If there are n nodes in the GPT, then we can set all node ranks equal to 1/n.

The risk value rk for node k is calculated using an iterative procedure until a stable value is obtained:

$$r_k(t) = \sum_j r_k(t-1)\mu_{jk} \tag{12}$$

This iteration converges to the stable value of the vector R* as the eigenvector of the matrix M.

The method of placing immune detectors includes the following steps.

Stage 1. Initialization. Each value of the risk vector is assigned an initial value of 1/n.

Stage 2. Iterative procedure for correcting risk values before the convergence condition occurs for each vertex of the GPT.

Stage 3. Placement of immune detectors in the network nodes with the highest risk value.

Stage 4. Determining the amount of risks as a general indicator of network information security.

The proposed method can be used to build an information security system for automated systems.

[1] *Tokarev V. L., Sychugov A. A.* Detection of malicious software using immune detectors // Proceedings of Tula State University. Technical science, 2017. No 10. Pp. 216–230.

[2] *Tokarev V. L., Sychugov A. A.* Multi-Agent system for network attack detection // International Journal of Civil Engineering and Technology, 2018. Vol. 9. No 6.

[3] Data bank of information security threats. CVSS V2 calculator. // `https://bdu.fstec.ru/calc`.

[4] *Tokarev V. L.* Recognition of the strategy of the opposing side based on current observations // Reports of Tomsk State University of Control Systems and Radioelectronics, 2014. Pp. 184–187.

[5] *Jha S., Sheyner O., Wing J.* Two Formal Analyses of Attack Graphs // Proceedings of 15th IEEE Computer Security Foundations Workshop, 2002. Pp. 49–63.

[6] *Mehta V., Bartzis C., Zhu H., Clarke E, Wing J.* Ranking Attack Graphs. // International Workshop on Recent Advances in Intrusion Detection, 2006. Pp 127–144.

[7] *Dynkin E. B.* Bases of the theory of Markov processes // FIZMATLIT, 2006. Pp. 228.

Содержание

Contents

Авторский указатель

Author index

MachineLearning.ru

http://www.machinelearning.ru/

Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Цели ресурса — сконцентрировать информацию о достижения ведущих научных школ; способствовать обмену опытом, накоплению и распространению научных знаний; предоставить площадку для виртуальных научных семинаров и обсуждений.

Научное издание

Интеллектуализация обработки информации

Тезисы докладов
13-й Международной конференции

Подписано в печать 17.12.2020
Формат 60×84 1/8
Усл.-печ. л. 20,1. Уч.-изд. л. 21,17
Тираж 50 экз

Издатель — Российская Академия Наук

Печать — УНИД РАН

Отпечатано в экспериментальной цифровой типографии РАН

Издается по распоряжению президиума РАН
и распространяется бесплатно