СПЕЦКУРС
Логический анализ данных в распознавании
(Logical data analysis in recognition)

лектор д.ф.-м.н. Елена Всеволодовна Дюкова

Спецкурс посвящён вопросам применения аппарата дискретной математики в задачах интеллектуального анализа данных. Излагаются общие принципы, лежащие в основе логического подхода к задачам машинного обучения. Описываются методы конструирования процедур классификации по прецедентам с использованием понятий теории булевых функций и теории покрытий булевых матриц. Рассматриваются основные модели логических процедур классификации, вопросы сложности их реализации и качества решения прикладных задач.

Спецкурс для бакалавров 2-4 курсов ВМК МГУ им. М.В. Ломоносова.

По спецкурсу издано учебное пособие:
http://www.ccas.ru/frc/papers/djukova03mp.pdf

Лекция 1
Сущность дискретного подхода к задаче классификации (распознавания) по прецедентам

- В различных областях человеческой деятельности возникают задачи, в которых требуется найти решение на основе анализа большого объема накопленных знаний. Рассмотрим некоторые примеры.
- 1) ***Медицинская диагностика и прогнозирование.*** На основе анализа медицинских карт (результатов обследований) пациентов требуется выявить наиболее важные симптомы заболевания, сделать прогноз насколько эффективна методика лечения.
- 2) ***Обработка социологической информации.*** На основе анализа ответов респондентов на вопросы анкеты требуется выделить однородные группы респондентов и классифицировать вопросы по степени их важности.
- 3) ***Техническое прогнозирование.*** На основе сравнения описаний состава и свойств известных сплавов и требуется осуществить прогнозирование свойств новых сплавов с заданными составами. Такие же исследования могут проводиться для прогнозирования свойств новых химических соединений, лекарственных препаратов и т.д.

- 4) ***Геологическое прогнозирование***. На основе анализа косвенных признаков (измеряемых и наблюдаемых характеристик) известных месторождений полезных ископаемых требуется определить наличие и масштабы новых месторождений.
- 5) ***Прогнозирование возможности возврата банковских кредитов.*** На основе анализа документации о цели кредита и состоятельности получателя кредита нужно сделать прогноз относительно возможности возврата этого кредита.
- 6) ***Анализ пользовательской среды сети Интернета***, например, с целью выделения основных категорий пользователей для рекламодателя.
- Итак, рассматриваются задачи, возникающие в плохо формализованных областях. Для их решения достаточно сложно, а иногда и просто невозможно построение математических моделей в общепринятом смысле. Другой подход основан на применении методов машинного обучения и заключается данный подход в использовании эвристических информационных моделей, при построении которых на этапе формализации используются различные экспертные знания. Такие же знания используются при интерпретации полученных результатов на языке пользователя. Сам же процесс принятия решения осуществляется в результате строгого математического исследования и применения содержательно обоснованных алгоритмов.

• По существу речь идет о формализации известной способности человека узнавать, точнее классифицировать, различные предметы или явления, руководствуясь накопленным опытом. Так, каждый из нас свободно читает рукописные тексты, написанные разными людьми, легко узнает знакомых, даже если они изменили свой внешний облик. Некоторые профессии напрямую связаны с умением классифицировать. Например, врачи могут диагностировать заболевания, геологи по косвенным признакам устанавливать наличие и масштабы месторождений и т.д. Дело в том, что в каждом из нас заложен особый механизм, называемый феноменом восприятия, благодаря которому и происходит сравнение вновь наблюдаемого объекта или явления с известными ранее наблюдаемыми примерами подобных объектов или явлений. Феномен восприятия присутствует практически во всех сферах человеческой деятельности. Естественно возникает вопрос, как научить ЭВМ распознавать (классифицировать) предметы и явления? На этот вопрос отвечает наука, центральной задачей которой является задача классификации по прецедентам.

- **Что понимается под прецедентной информацией?**

• Под прецедентной (обучающей) информацией понимается совокупность примеров изучаемых объектов, в которой каждый объект представлен в виде числовой последовательности, полученной на основе измерения или наблюдения ряда его параметров или характеристик. Подлежащие измерению или наблюдению параметры и характеристики называются признаками. В самом простом случае прецеденты делятся на два класса (класс положительных и класс отрицательных примеров). В общем случае число классов может быть больше двух. Например, месторождения полезных ископаемых могут быть мелкими, средними и крупными. Требуется уметь классифицировать объекты, не вошедшие в обучающую выборку, т.е. по признаковому описанию каждого такого объекта, определять какому классу он принадлежит.

- Ниже приведена **формальная постановка задачи классификации по прецедентам.**
-
- Исследуется некоторое множество объектов $\boldsymbol{M}$. Известно, что $\boldsymbol{M}$ представимо в виде объединения $\boldsymbol{l}$ подмножеств $\boldsymbol{K_1, \dots, K_l}$, называемых классами. Объекты множества $\boldsymbol{M}$ описываются признаками $\boldsymbol{x_1, \dots, x_n}$. Имеется конечный набор $\boldsymbol{S_1, \dots, S_m}$ объектов из множества $\boldsymbol{M}$, о которых известно, каким классам они принадлежат. Это прецеденты или обучающие объекты. Пусть их описания имеют вид $\boldsymbol{S_1 = (a_{11}, \dots, a_{1n})}$, $\boldsymbol{S_2 = (a_{21}, \dots, a_{2n}), \dots, S_m = (a_{m1}, \dots, a_{mn})}$, здесь $\boldsymbol{a_{ij}}$ - значение признака $\boldsymbol{x_j}$ для объекта $\boldsymbol{S_i}$. Требуется по предъявленному набору значений признаков $\boldsymbol{(b_1, \dots, b_n)}$, описывающему некоторый объект из M, о котором, вообще говоря, не известно какому классу он принадлежит, определить этот класс.
- Фактически нужно сравнить вновь предъявленное описание с материалом обучения. Существуют разные мнения о том, как проводить подобное сравнение.

- Первоначально считалось, что данное направление является частью математической статистики. Для анализа сложных описаний с помощью статистических методов необходимо принимать на веру дополнительные предположения вероятностного характера, т.е. предъявлять достаточно сильные требования к пространствам описаний исследуемых объектов. Кроме того, для получения надежных результатов на основе статистического подхода требуются чрезвычайно большие массивы прецедентов, т.е. обучающая выборка должна быть достаточно представительной. Оказалось, что набор большого числа прецедентов требует, как правило, дорогостоящих и трудоемких работ, а в некоторых случаях вообще невозможен. Например, такая ситуация имеет место в задачах прогнозирования месторождений редких ископаемых, прогнозирования свойств твёрдых сплавов. Для решения подобных задач не было адекватных математических методов и их пришлось создавать на основе совершенно новых идей, в частности, на основе применения логических методов анализа данных.

- Логический анализ прецедентной информации использует аппарат дискретной математики, в том числе методы преобразования нормальных форм логических функций и построения покрытий булевых целочисленных матриц. Такой анализ сводится к поиску в исходных данных определенных закономерностей. Найденные закономерности или элементарные классификаторы позволяют различать объекты из разных классов и, как правило, имеют содержательное описание в терминах той прикладной области, в которой решается задача. По их наличию или наоборот отсутствию в описании распознаваемого объекта решается вопрос о его классификации. Центральными являются вопросы корректного обучения. Алгоритм классификации считается корректным, если он безошибочно классифицирует обучающие объекты.
- Логический подход нацелен на обработку данных с признаками, имеющими конечное множество допустимых значений. Допустимые значения признака обычно кодируется целыми числами. Особенно эффективен этот подход в случае целочисленной информации низкой значности.
- Поясним сказанное на примере бинарных признаков.

- Пусть каждый признак x_j , $j \in \{1, 2, .., n\}$, является бинарным, то есть может принимать значение 0 или 1. Тогда элементарный классификатор (эл.кл.) – это элементарная конъюнкция над переменными $x_1, \dots, x_n$, определённая на признаковых описаниях объектов. Если на описании некоторого объекта элементарная конъюнкция обращается в единицу, то говорят, что этот объект содержит данный эл.кл. ***Эл.кл. считается корректным для класса*** K , $K \in \{K_1, \dots, K_l\}$, если не существует двух обучающих объектов S' и S'' , $S' \in K$, $S'' \notin K$, содержащих данный эл.кл. Аналогичным образом вводится понятие эл.кл. в случае целочисленных данных.
- В классических моделях этап обучения основан на построении специальных семейств корректных эл.кл. При этом ищутся наиболее информативные эл.кл. На этапе классификации каждый найденный эл.кл. участвует в процедуре «голосования». В результате вычисляются оценки принадлежности распознаваемого объекта к классам. Корректность таких моделей логических классификаторов обеспечивается корректностью используемых эл.кл.

• Не всегда удается найти достаточное количество информативных корректных эл.кл. Подобная ситуация возникает, например, когда информация целочисленная и каждый признак имеет слишком много значений. Один из способов решения проблемы – применение логических корректоров (логических классификаторы, при построении которых используются произвольные, не обязательно корректные эл.кл.). Корректность алгоритма достигается за счёт построения корректных наборов эл.кл. ***Набор эл.кл. называется корректным для класса*** $\boldsymbol{K}$, $\boldsymbol{K} \in \{\boldsymbol{K_1}, \dots, \boldsymbol{K_l}\}$, если для любых двух обучающих объектов $\boldsymbol{S'}$ и $\boldsymbol{S''}$, $\boldsymbol{S'} \in \boldsymbol{K}$, $\boldsymbol{S''} \notin \boldsymbol{K}$, в данном наборе существует эл.кл. такой, что один из объектов $\boldsymbol{S'}$ и $\boldsymbol{S''}$ его содержит, а другой не содержит.

• Главное достоинство логических методов – это возможность классификации в случае небольшого числа прецедентов (в этом их преимущество перед статистическими методами). Не требуется также задание метрики в пространстве описаний объектов. Однако использование дискретного аппарата приводит к трудностям вычислительного характера, особенно в случае, когда для описания прецедентов используется значительное число их характеристик.

• Основное направление теоретических исследований – изучение сложности дискретных перечислительных задач, возникающих при поиске корректных эл.кл. и поиске корректных наборов эл.кл. Сложность (***труднорешаемость***) таких задач обусловлена двумя аспектами: экспоненциальным ростом числа решений при увеличении размера задачи и сложностью их нахождения (перечисления). Эффективность алгоритма оценивается сложностью одного шага (сложностью нахождения каждого нового решения). Наиболее эффективным считается алгоритм, имеющий полиномиальный от размера входа (размера задачи) шаг, при этом оценка сложности шага алгоритма даётся для худшего случая (для самого сложного варианта задачи). Такой алгоритм называется алгоритмом ***с полиномиальной задержкой.***

• Главной перечислительной задачей считается ***монотонная дуализация.*** Это задача построения сокращённой дизъюнктивной нормальной формы монотонной булевой функции, заданной конъюнктивной нормальной формой (КНФ). Кроме приведённой формулировки, монотонная дуализация имеет графовую и матричную постановки, в которых соответственно используются понятие вершинного покрытия гиперграфа и понятие покрытия булевой матрицы. Хотя эта задача поставлена ещё в 60-х годах прошлого века, полиномиальные алгоритмы (алгоритмы с полиномиальной задержкой) удалось построить лишь некоторых ограничениях, например, когда в исходной КНФ каждая дизъюнкция содержит не более двух переменных. Поэтому требования к алгоритму были ослаблены. Обозначились ***два основных направления исследований.***

- ***Первое направление***, разрабатываемое в основном за рубежом, основано на построении так называемых ***инкрементальных*** алгоритмов, когда алгоритму разрешено на каждом шаге просматривать решения, найденные на предыдущих шагах. В 1996 году М. Фридманом и Л. Хачияном построен инкрементальный алгоритм дуализации с квазиполиномиальным шагом, фактически определяемым не только размером входа задачи, но и размером её выхода. Такой алгоритм интересен исключительно для теории, поскольку в худшем случае число решений дуализации (размер выхода задачи) растёт экспоненциально с ростом размера её входа.
- ***Второе направление*** (предложено Е.В. Дюковой в 1977 году) основано на построении ***асимптотически оптимальных*** алгоритмов монотонной дуализации. В этом случае алгоритму разрешено делать лишние полиномиальные шаги при условии, что их число должно быть мало по сравнению с числом всех решений задачи. В результате удалось построить алгоритмы, эффективные не всегда, а почти всегда (для почти всех вариантов задачи). Эти алгоритмы на сегодняшний день являются лидерами по скорости счёта.

- При решении прикладных задач большой размерности для сокращения вычислительных затрат приходится отказываться от перечисления всех решений монотонной дуализации и строить приближённые алгоритмы, например, стохастические.
- Существует ряд проблем, от успешного решения которых зависит качество решения прикладных задач классификации. Для дискретного подхода основной является проблема обработки целочисленной информации высокой значности и вещественнозначной информации. Среди вопросов общего характера следует выделить вопросы, связанные с предварительным анализом обучающей выборки с целью выделения наиболее информативных признаков и обучающих объектов, типичных для своих классов.

• Тематика развивается в научной школе академика РАН Ю.И. Журавлева более 40 лет. Одной из первых работ в этом направлении была статья [3], в которой рассматривалась задача прогнозирования золотоносных месторождений. Для анализа обучающей выборки в [3] использовалось хорошо известное в дискретной математике понятие теста, которое первоначально применялось в задачах контроля управляющих систем [8]. Упомянутая выше статья Ю.И. Журавлёва с соавторами [3], а также статьи М.Н. Вайнвайга и М.М. Бонгарда [1, 2], в которых описывалась модель распознающего алгоритма под названием “Кора”, положили начало отечественным исследованиям по применению методов логического (дискретного) анализа данных в задаче классификации по прецедентам. Идея построения логических процедур распознавания с использованием корректных наборов эл.кл. предложена в [5] и получила существенное развитие в работах [4, 6]. В [4] решена важная в методологическом плане задача построения общей схемы синтеза корректных логических процедур распознавания. В [7] построены логические классификаторы для данных, в которых на множестве допустимых значений каждого признака задан конечный частичный порядок.

- ОСНОВНАЯ ЛИТЕРАТУРА

- 1. *Бонгард М. М.* Проблема узнавания // М.: Физматгиз, 1967. 321 с.
- *2. Вайнцвайг М. Н.* Алгоритм обучения распознаванию образов «Кора». // Алгоритмы обучения распознаванию образов / Под ред. В. Н. Вапник._ М.: Советское радио, 1973.
- *3. Дмитриев А.И., Журавлев Ю.И., Кренделев Ф.П.* О математических принципах классификации предметов или явлений // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1966. Вып. 7.
- *4. Дюкова Е.В., Журавлёв Ю.И., Прокофьев П.А.* Логические корректоры в задаче классификации по прецедентам // Ж. Вычислительная математика и математическая физика, 2017. Т. 57. № 11.
- *5. Дюкова Е.В., Журавлёв Ю.И., Рудаков К.В.* Об алгебраическом синтезе корректных процедур распознавания на базе элементарных алгоритмов. Ж. Вычислительная математика и математическая физика, 1996. Т. 36. № 8.
- *6. Дюкова Е.В., Журавлёв Ю.И., Сотнезов Р.М.* Построение коллектива логических корректоров на базе элементарных классификаторов // Pattern Recognition and Image Analysis. 2011. Vol. 21. №2.
- *7. Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О логическом анализе данных с частичными порядками в задаче классификации по прецедентам // Ж. Вычислительная математика и математическая физика, 2019. Т. 59. № 9.
- *8. Чегис И.А., Яблонский С.В.* Логические способы контроля электрических схем // Тр. МИАН СССР, М., 1958.