На правах рукописи

СОТНЕЗОВ Роман Михайлович

Исследование в области сложности алгебро-логического анализа данных и синтеза распознающих процедур

Специальность:
01.01.09 – Дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва - 2011

Работа выполнена на кафедре математических методов прогнозирования факультета вычислительной математики и кибернетики
Московского государственного университета им. М.В. Ломоносова

Научный руководитель: доктор физико-математических наук
Е.В. Дюкова

Официальные оппоненты: доктор физико-математических наук

кандидат физико-математических наук

Ведущая организация:

Защита состоится «___» _____ 2012 г. в _____ часов на заседании диссертационного совета Д002.017.02 в Учреждении Российской академии наук Вычислительный центр им. А.А. Дородницына РАН по адресу: 119991, Москва, ул. Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН

Автореферат разослан «___» _____ 2012 г.

Ученый секретарь
диссертационного совета
д.ф.-м.н., профессор В.В. Рязанов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Рассматриваются задачи, в которых требуется найти решение на основе анализа большого объема накопленных знаний. К ним относятся задачи классификации, распознавания и прогнозирования, возникающие в различных плохо формализованных областях таких, как медицинская диагностика и прогнозирование, обработка социологической информации, техническое и геологическое прогнозирование, анализ банковской деятельности и.т.д. Для решения перечисленных задач успешно применяются методы распознавания образов, в частности методы, основанные на обучении по прецедентам.

Постановка задачи распознавания по прецедентам заключается в следующем. Исследуется некоторое множество объектов M, про которое известно, что оно может быть разбито на непересекающиеся подмножества (классы) $K_1, \dots, K_l,\ l \geq 2$. Под прецедентной (обучающей) информацией понимается совокупность примеров описаний изучаемых объектов, полученная на основе измерения или наблюдения ряда характеристик этих объектов, а также те «ответы», которые должен был бы дать «идеальный» алгоритм, решая задачу классификации для заданной совокупности описаний. Подлежащие измерению или наблюдению характеристики называются признаками. Требуется уметь классифицировать объекты, не вошедшие в обучающую выборку, т.е. по признаковому описанию каждого такого объекта определять, какому классу он принадлежит. Фактически нужно сравнить вновь предъявленное описание с материалом обучения. Существуют разные мнения о том, как проводить подобное сравнение.

Развиваемый в данной работе подход к задаче распознавания по прецедентам базируется на применении аппарата дискретной математики с использованием логических и алгебро-логических методов анализа данных. Основы проблематики были заложены в работах С.В. Яблонского, Ю.И. Журавлёва, М.Н. Вайнцвайга и М.М. Бонгарда.

Использование аппарата дискретной математики для решения прикладных задач распознавания имеет целый ряд достоинств, к числу которых, прежде всего, следует отнести возможность получения результата при отсутствии сведений о функциях распределения и при наличии малых обучающих выборок. Не требуется также задание метрики в пространстве описаний объектов. В данном случае для каждого признака определяется бинарная функция близости между его значениями, позволяющая различать объекты и их подописания. Особенно эффективен рассматриваемый подход в случае дискретной (целочисленной) информации низкой значности, например, бинарной. Вещественнозначная информация часто рассматривается как целочисленная высокой значности. Поэтому актуальной является задача корректного понижения значности исходных целочисленных данных.

Важнейшими для рассматриваемого направления являются вопросы эффективного поиска конъюнктивных закономерностей в признаковых описаниях объектов, которые играют роль элементарных классификаторов. В решающем правиле используется процедура голосования по каждому из построенных элементарных классификаторов. Как правило, корректность распознающего алгоритма (способность правильно классифицировать объекты из обучающей выборки) обеспечивается корректностью каждого из порождаемых элементарных классификаторов, что является основной логического синтеза распознающих процедур.

В [1] предложена идея построения корректных процедур распознавания на базе произвольных эл.кл., то есть эл.кл. необязательно являющихся корректными. В качестве корректирующей функции рассмотрена монотонная булева функция. Предлагаемый в [1] подход сочетает алгебраические и логические методы построения корректных распознающих процедур (алгебро-логический подход). Вопросы, касающиеся построения и исследования корректных процедур распознавания с использованием

алгебро-логического подхода являются актуальными, поскольку ранее эти вопросы в должной степени не рассматривались.

Практическое использование логических процедур распознавания напрямую связано со снижением их вычислительной сложности. При большой размерности обучающей выборки возникает необходимость рассматривать труднорешаемые дискретные задачи. Это задачи преобразования нормальных форм логических функций и поиска покрытий булевых и целочисленных матриц.

Е.В. Дюковой предложен подход к решению указанных перечислительных задач, основанный на понятии асимптотически оптимального алгоритма. Показано, что при определенных условиях почти всегда исходную задачу Z можно заменить на более простую задачу Z_1, эффективно решаемую и такую, что, во-первых, множество решений задачи Z_1 содержит множество решений задачи Z, и, во-вторых, с ростом размера задачи Z число ее решений асимптотически равно числу решений задачи Z_1. Данный подход хорошо зарекомендовал себя при решении практических задач. В тех случаях, когда не удается построить асимптотические оптимальные алгоритмы для задач поиска тупиковых покрытий булевых и целочисленных матриц (преобразования нормальных форм логических функций), имеет смысл предъявлять более слабые требования к эффективности алгоритма.

При конструировании логических процедур распознавания и предварительного анализа обучающей выборки часто возникают дискретные оптимизационные задачи. Для их решения наряду с методами имеющими теоретическое обоснование необходимо разрабатывать эвристические подходы дающие хорошее приближенное решение.

1. Е.В. Дюкова, Ю.И. Журавлев, К.В. Рудаков. Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // Ж. вычисл. матем. и матем. физ. 1996 Т. 36 № 8 С. 215-223.

Цели и задачи диссертационной работы. Целью диссертационной работы является развитие логического и алгебро-логического анализа данных.

В рамках поставленной задачи были выделены следующие основные направления исследований.

1. Разработка новых подходов к повышению эффективности решения задачи распознавания по прецедентам методами логического и алгебро-логического анализа данных.
   1.1. Построение и исследование новых моделей распознающих процедур на базе произвольных элементарных классификаторов.
   1.2. Развитие методов корректного понижения значности целочисленных данных в задачах распознавания.
2. Получение новых результатов, касающихся снижения вычислительной сложности логических процедур распознавания.
   2.1. Построение генетических алгоритмов, эффективно решающих оптимизационные задачи, возникающие при построении логических процедур распознавания на базе элементарных классификаторов.
   2.2. Построение и обоснование эффективных алгоритмов поиска тупиковых покрытий булевых и целочисленных для случаев, когда не удается построить асимптотически оптимальные алгоритмы. Получение аналогичных результатов для задачи синтеза сокращенной дизъюнктивной нормальной формы логической функции.
   2.3. Усовершенствование техники нахождения асимптотических оценок числа решений труднорешаемых дискретных задач. Получение новых асимптотических оценок для количественных характеристик множества покрытий целочисленной матрицы (количественных характеристик дизъюнктивной нормальной формы логической функции).

Научная новизна. Развит алгебро-логический подход к синтезу процедур распознавания. Разработана оригинальная модель распознающих процедур, основанная на построении коллектива логических корректоров. Для снижения вычислительной сложности модели использован генетический подход. Модель успешно апробирована на реальных задачах.

Разработаны новые методы понижения значности исходной целочисленной информации с сохранением разбиения обучающего множества объектов на классы.

Развиты методы получения асимптотических оценок количественных характеристик сокращённой дизъюнктивной нормальной формы логической функции. Решена технически сложная задача получения асимптотик для типичных значений числа тупиковых покрытий и длины тупикового покрытия целочисленной матрицы в случае, когда число столбцов матрицы не превосходит числа её строк.

Аналогичные результаты получены для количественных характеристик множества максимальных конъюнкций двузначной логической функции, заданной конъюнктивной нормальной формой (КНФ), в случае, когда число переменных в КНФ не превосходит числа элементарных дизъюнкций.

Доказана асимптотическая эффективность алгоритмов построения тупиковых покрытий целочисленной матрицы, основанного на перечислении с полиномиальной задержкой «совместимых» наборов столбцов в случае, когда число столбцов матрицы не превосходит числа её строк. Аналогичный результат получен для алгоритмов поиска максимальных конъюнкций логической функции, основанных на перечислении с полиномиальной задержкой «неприводимых» конъюнкций этой функции.

Методы исследования. В работе используется аппарат дискретной математики, в частности алгебры логики, теории дизъюнктивных нормальных форм логических функции. Применяются методы построения покрытий булевых и целочисленных матриц, а также методы получения асимптотик для типичных значений количественных характеристик

множеств неприводимых покрытий булевой матрицы и тупиковых покрытий целочисленной матрицы.

Теоретическая и практическая ценность. Результаты, полученные в диссертационной работе, могут быть использованы в теоретических исследованиях, касающихся построения эффективных реализаций для моделей логических процедур распознавания. Эффективность предложенных подходов подтверждена решением практических задач.

Апробация работы. Основные положения и результаты диссертации докладывались на 7 конференциях:

1. 9th International Conference on Pattern Recognition and Image Analysis: new Information Technologies (PRIA-9-2008), Нижний Новгород, сентябрь 2008 г.
2. Восьмая Международная конференция «Дискретные модели в теории управляющих систем», Москва, апрель 2009 г.
3. Всероссийская конференция «Математические методы распознавания образов» ММРО-14, Суздаль, сентябрь 2009 г.
4. Восьмая Международная конференция «Интеллектуализация обработки информации - 2010», Республика Кипр, Пафос, октябрь 2010 г.
5. Second International Conference «Classification, Forecasting, Data Mining», Болгария, Варна, июнь 2010 г.
6. Всероссийская конференция «Математические методы распознавания образов» ММРО-15, Петрозаводск, сентябрь 2011 г.
7. Научная конференция «Ломоносовские чтения», г. Москва, МГУ, ноябрь 2011 г.

Результаты работы докладывались и обсуждались на научных семинарах Учреждения Российской академии наук Вычислительный центр им. А.А. Дородницына РАН и кафедры Математических методов прогнозирования

факультета вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова.

Публикации. По теме диссертации опубликовано 11 статей, в том числе 4 статьи в изданиях, входящих в перечень ведущих рецензируемых журналов и изданий, рекомендованных ВАК для публикации основных результатов диссертации на соискание ученой степени доктора и кандидата наук. Описания основных результатов, полученных в диссертации, включались в научные отчеты по проектам РФФИ 07-01-00516-а, 10-01-00770-а и в отчеты по грантам президента РФ по поддержке ведущих научных школ НШ №5294.2008.1 и НШ №7950.2010.1.

Структура и объем работы. Диссертация состоит из введения, 4 глав и списка литературы из 67 наименований. Общий объем работы – 103 страницы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы и обсуждается круг проблем, возникающих при построении логических процедур распознавания, перечисляются основные цели диссертационной работы, приводится краткое изложение результатов, полученных в работе.

В главе 1 рассмотрена задача построения процедур распознавания на базе произвольных элементарных классификаторов. Введены понятия корректного набора элементарных классификаторов класса и монотонного корректного набора элементарных классификаторов класса. Разработаны и исследованы алгоритмы распознавания, основанные на голосовании по коллективу (монотонных) наборов элементарных классификаторов Проведено тестирование построенных алгоритмов на реальных прикладных задачах.

Задача распознавания по прецедентам рассматривается в стандартной постановке. Исследуется некоторое множество объектов M, про которое известно, что оно представимо в виде объединения непересекающихся подмножеств (классов) $K_1, \dots, K_l$. Объекты множества M описываются

набором целочисленных признаков $x_1, \dots, x_n$, каждый из которых имеет конечное число допустимых значений. В качестве исходной информации дано множество объектов $T = \{S_1, \dots, S_m\}$ из M, о которых известно каким классам они принадлежат (обучающая выборка). Требуется по предъявленному набору значений признаков $x_1, \dots, x_n$, описывающему некоторый объект S из M, определить класс, к которому относится объект S.

Одним из основных понятий, используемое при построении логических процедур распознавания, является понятие элементарного классификатора.

Пусть $H = \{x_{j_1}, \dots, x_{j_r}\}$ - набор из r различных признаков, $r \leq n$, и пусть $\sigma = (\sigma_1, \dots, \sigma_r)$, σ_i – допустимое значение признака x_{j_i} при $i = 1,2, \dots, r$. Пара (H, σ) называется *элементарным классификатором* (эл.кл.). Близость объекта $S = (a_1, \dots, a_n)$ из M и эл.кл. (H, σ) оценивается величиной

$$B_{(H,\sigma)}(S) = \begin{cases} 1, & \text{если } a_{j_i} = \sigma_i, \\ 0, & \text{в противном случае.} \end{cases}$$

Пусть $U = \{(H_1, \sigma_1), \dots, (H_q, \sigma_q)\}$ - набор эл.кл, $S \in M$. Двоичный набор $\left(B_{(H_1,\sigma_1)}(S), \dots, B_{(H_q,\sigma_q)}(S)\right)$ представляет собой вектор классификации объекта S набором U и обозначается как $\omega_U(S)$.

Определение 1. Набор эл.кл. U называется корректным для класса K, $K \in \{K_1, \dots, K_l\}$, если существует функция алгебры логики $F_{U,K}$ такая, что для любых двух объектов S' и S'' из обучающей выборки, таких что $S' \in K$, $S'' \in \overline{K}$ выполняется неравенство

$$F_{U,K}(\omega_U(S')) \neq F_{U,K}(\omega_U(S''))$$

(здесь и далее $\overline{K} = (K_1 \cup \dots \cup K_l) \backslash K$).

Функция $F_{U,K}$ называется логическим корректором класса K.

Определение 2. Корректный набор эл.кл. U называется монотонным корректным для класса K, если функция $F_{U,K}$ – монотонная и для любого обучающего объекта S' класса K значение функции $F_{U,K}(\omega_U(S'))$ равно 1.

Монотонная функция $F_{U,K}$ называется монотонным логическим корректором класса K.

(Монотонный) корректный набор эл.кл. U называется тупиковым, если любое его подмножество не является (монотонным) корректным для K. (Монотонный) корректный набор эл.кл. U называется минимальным, если не существует (монотонного) корректного набора эл.кл. класса K меньшей мощности.

Далее рассматриваются распознающие алгоритмы, работающие по следующей схеме. Для каждого класса K, $K \in \{K_1, \dots, K_l\}$ конструируется некоторое подмножество $W_A(K)$ множества корректных наборов эл.кл. класса K. Распознавание объекта S производится на основе оценок принадлежности этого объекта к классам $K_1, \dots, K_l$. Эти оценки вычисляются следующим образом.

Случай 1. Множество $W_A(K)$ состоит из корректных наборов эл.кл., не обязательно являющихся монотонными.

Тогда оценка принадлежности объекта S к классу K имеет вид

$$\Gamma^*(S,K) = \frac{1}{|W_A(K)|} \sum_{U \in W_A(K)} \sum_{S' \in T \cap K} \delta_U(S', S),$$

где $\delta_U(S', S) = 1$, если $\omega_U(S) = \omega_U(S')$, и $\delta_U(S', S) = 0$ иначе.

Случай 2. Множество $W_A(K)$ состоит только из монотонных корректных наборов эл.кл. Пусть

$$\delta_U^*(S', S) = \begin{cases} 1, \text{если } \omega_U(S) \succcurlyeq \omega_U(S'), \\ 0, \text{иначе,} \end{cases}$$

где обозначение $a' \succcurlyeq a''$ для двоичных векторов $a' = (a'_1, \dots, a'_n)$ и $a'' = (a''_1, \dots, a''_n)$ означает, что $a'_i \geq a''_i$ при $i = 1,2, \dots, n$.

Оценка принадлежности объекта S к классу K имеет вид

$$\Gamma_1^*(S,K) = \frac{1}{|W_A(K)|} \sum_{U \in W_A(K)} \sum_{S' \in T \cap K} \delta_U^*(S', S).$$

На основе генетического подхода разработаны и реализованы алгоритмы А1, А2, А3, А4, конструирующие корректные наборы эл.кл. Данные алгоритмы являются модификациями генетических алгоритмов

описанных в главе 3. В качестве особей популяции в разработанных алгоритмах используются корректные наборы эл.кл.

Алгоритмы А1 и А2 решает задачу конструирования, соответственно, коллектива монотонных корректных наборов эл.кл. и коллектива просто корректных наборов эл.кл. с высокой распознающей способностью. Функция приспособленности корректного набора эл.кл. – оценка распознающей способности этого набора, имеющая следующий вид

$$\tau_{A,K}(U) = \frac{1}{|T_1 \cap K|} \sum_{S \in T_0 \cap K} \sum_{S' \in T_1 \cap K} \delta_U(S, S') - \frac{1}{|T_1 \cap \overline{K}|} \sum_{S \in T_0 \cap K} \sum_{S' \in T_1 \cap \overline{K}} \delta_U(S, S'),$$

где $T = T_0 \cup T_1,\ T_0 \cap T_1 = \emptyset$. Выборка T_0 используется для построения (монотонных) корректных наборов эл.кл., а выборка T_1 для оценки качества распознавания построенных (монотонных) корректных наборов эл.кл.

Алгоритмы А3 и А4 решают задачу конструирования, соответственно, одного монотонного корректного набора эл.кл. и просто корректного набора эл.кл., по мощности близких к минимальному. Функция приспособленности корректного набора эл.кл. – мощность набора.

Тестирование разработанных алгоритмов проводилось на реальных задачах. Показано, что алгоритм А1 является наилучшим по качеству распознавания.

В главе 2 развиты вопросы применения логических процедур распознавания в случае веществennозначной информации и целочисленной информации высокой значности. Рассмотрена задача корректного понижения значности данных.

Задача ставится следующим образом. По обучающей выборке T строится специальная булева матрица L_T, столбцы которой разбиты на n групп, где n – число признаков. Требуется построить кодирующее покрытие – набор столбцов матрицы L_T, который, во-первых, является покрытием матрицы L_T, и, во-вторых, содержит хотя бы один столбец из каждой группы. Каждое кодирующее покрытие определяет некоторую корректную

перекодировку исходной информации, то есть такое преобразование обучающей информации, при котором объекты из разных классов остаются различимыми.

Встает вопрос о выборе наилучшей в смысле качества распознавания корректной перекодировки. Полный перебор всех перекодировок является трудоемким в вычислительном плане вследствие большого размера матрицы L_T. Для сокращения перебора разработаны генетические алгоритмы поиска оптимальной корректной перекодировки исходной информации, которые являются модификациями алгоритмов поиска минимального покрытия, описанных в главе 3. В качестве особей используются кодирующие покрытия, в качестве функций приспособленности один из двух функционалов:

$$f_1(H) = \sum_{j \in R_2(H)} c_j,$$

$$f_2(H) = \frac{1}{|H|} \sum_{j \in R_1(H)} c_j,$$

здесь H - кодирующее покрытие, $c_j, j \in \{1,2,\dots,n\}$, - число единиц в $j-$ом столбце матрицы L_T, $R_1(H)$ - множество номеров столбцов L_T, входящих в H, $R_2(H)$ - множество номеров столбцов матрицы L_T, не входящих в H.

Проведено тестирование разработанных алгоритмов на реальных данных и сравнение с другими алгоритмами перекодирования данных. Показано что эта методика позволяет повысить качество распознавания алгоритма голосования по представительным наборам, сконструированного по перекодированным данным, существенно не увеличивая вычислительных затрат.

В главе 3 рассмотрена задача поиска минимального покрытия булевой матрицы. Данная задача относится к классу *NP*-полных, в связи с чем, известные алгоритмы поиска точного решения имеют экспоненциальную вычислительную сложность и малопригодны на практике. Для задач больших

размерностей ищутся приближенные решения. Как правило, хорошие результаты дает градиентный алгоритм. Однако в ряде случаев, например, на матрицах разреженных по числу единиц, качество решения, выдаваемого градиентным алгоритмом, резко ухудшается. Поэтому актуальными являются вопросы разработки быстро работающих эвристик, дающих хорошие приближенные решения для сложных задач.

Для задачи поиска минимального покрытия булевой матрицы разработаны два генетических алгоритма: алгоритм с бинарным представлением задачи и алгоритм с целочисленным представлением задачи. В первом случае для описания покрытия матрицы (особи популяции) используется бинарный вектор, во втором – целочисленный. Оба алгоритма осуществляют поиск минимального покрытия среди неприводимых покрытий. В качестве оценки пригодности решения (функции приспособленности) использован вес соответствующего покрытия. Предложены нестандартные операторы скрещивания, учитывающие веса используемых столбцов и значения функций приспособленности особей-родителей, а также операторы мутации с переменным числом мутируемых генов. Число мутируемых генов $k(t)$ на шаге t, возрастает с развитием популяции и определяется по формуле

$$k(t) = k_0\left(1 - \frac{1}{C \cdot t + 1}\right),$$

где k_0 – число мутируемых генов на последнем шаге алгоритма, C – параметр, регулирующий скорость изменения числа мутируемых генов.

Эффективность построенных алгоритмов оценена на тестовых задачах, содержащихся в электронной библиотеке *OR Library*. Эти задачи состоят из 65 разреженных по числу единиц матриц, разбитых на 11 классов. Результаты тестирования показывают, что хотя бы один из алгоритмов находит оптимальное решение в 61 задаче. В четырех оставшихся задачах лучшее найденное покрытие отличается по весу от оптимального на единицу.

Проведено сравнение построенных в работе генетических алгоритмов с двумя алгоритмами, имеющими теоретические оценки точности. Первым алгоритмом является градиентный алгоритм, в качестве второго алгоритма выбран один из вариантов алгоритма *General*. Сравнение на большом числе случайных матриц показало, что генетический алгоритм, как правило, превосходит по точности решения как градиентный алгоритм, так и алгоритм *General*, что говорит о практической ценности разработанных алгоритмов.

Разработанные генетические алгоритмы адаптированы для многопроцессорных комплексов с различными схемами обмена информацией между процессорами. Предложен следующий подход к распараллеливанию алгоритмов. На каждом вычисляющем процессоре запускается генетический алгоритм со своим набором входных параметров. Через определенное количество шагов между вычисляющими процессорами осуществляется обмен сообщениями о найденных решениях.

Сравнение параллельных реализаций генетических алгоритмов проводилось по следующим параметрам: средняя длина полученного покрытия для каждого конкретного числа вычисляющих процессоров и среднее время поиска лучшего решения. Было выявлено, что при возрастании числа процессоров, как правило, уменьшается средняя величина выдаваемого покрытия. При этом в случаях, когда уменьшение средней длины покрытия не происходит, наблюдается уменьшение времени поиска лучшего решения.

В главе 4 получены асимптотики типичного числа тупиковых покрытий целочисленной матрицы и типичной длины тупикового покрытия в случае, когда число столбцов матрицы не превосходит числа ее строк. Приведены аналогичные результаты для типичного числа максимальных конъюнкций и типичного ранга максимальной конъюнкций логической функции. Показана асимптотическая эффективность алгоритма, основанного на перечислении «совместимых наборов» столбцов.

В разделе 4.1 вводятся основные понятия и описываются результаты в данной области, полученные ранее другими исследователями.

Пусть M_{mn}^{k} - множество всех матриц размера $m \times n$ с элементами из $\{0,1,\dots,k-1\}$, $k \geq 2$; E_k^r, $k \geq 2$, $r \leq n$, – множество всех наборов вида $(\sigma_1,..,\sigma_r)$, где $\sigma_i \in \{0,1,\dots,k-1\}$, $i = 1,2,\dots,n$.

Пусть $L \in M_{mn}^{k}$, H - набор столбцов матрицы L, $\sigma \in E_k^r$, $\sigma = (\sigma_1,..,\sigma_r)$. Набор столбцов H называется тупиковым σ −покрытием, если выполнены следующие два условия: 1) подматрица L^H матрицы L, образованная столбцами набора H, не содержит строку σ, 2) для каждого $p \in \{1,2,\dots,r\}$ подматрица L^Hсодержит по крайней мере одну из строк вида $(\beta_1,..,\beta_r) \in E_k^r$, где $\beta_p \neq \sigma_p$ и $\beta_j \neq \sigma_j$ при $j \in \{1,2,\dots,r\}\backslash\{p\}$, т.е. L^Hсодержит σ −подматрицу.

Если выполнено только условие 1), то набор столбцов H называется $\sigma -$ покрытием матрицы L.

Если выполнено только условие 2), то набор столбцов H называется $\sigma -$ совместимым набором столбцов матрицы L.

Нетрудно видеть, что понятие (тупикового) $(0,0,\dots,0)$-покрытия булевой матрицы совпадает с понятием (неприводимого) покрытия булевой матрицы. Отметим, что $(0,0,\dots,0)$-подматрица булевой матрицы является единичной подматрицей.

Пусть $L \in M_{mn}^{k}$. Положим $S(L,\sigma), \sigma \in E_k^r$ - множество всех -подматриц матрицы L, $C(L,\sigma)$, $\sigma \in E_k^r$ - множество всех σ −покрытий матрицы L, $B(L,\sigma), \sigma \in E_k^r$, - множество всех тупиковых σ −покрытий матрицы L, $U(L,\sigma), \sigma \in E_k^r$, - множество всех σ −совместимых наборов столбцов матрицы L.

Пусть далее

$S_r(L) = \bigcup_{\sigma \in E_k^r} S(L,\sigma)$, $S(L) = \bigcup_{r=1}^{n} S_r(L)$,

$C_r(L) = \bigcup_{\sigma \in E_k^r} C(L,\sigma)$, $C(L) = \bigcup_{r=1}^{n} C_r(L)$,

$B_r(L) = \bigcup_{\sigma \in E_k^r} B(L,\sigma)$, $B(L) = \bigcup_{r=1}^{n} B_r(L)$,

$U_r(L) = \bigcup_{\sigma \in E_k^r} U(L,\sigma)$, $U(L) = \bigcup_{r=1}^{n} U_r(L)$,

$r_1(k) = [\log_k m - \log_k \ln \log_k m - 1]$,

$r_2(k) =]\log_k m + c[$, $c = \log_k \log_k m + \log_k \log_k \log_k n$,

$r_1 = r_1(2), r_2 = r_2(2)$,

$p_r(k) = \exp(-mk^{-r})\,(1 - \exp(-mk^{-r}))^r$,

$f_n \approx g_n, n \to \infty$, означает, что $f_n = g_n(1 + \delta_n)$, где $\delta_n \to 0$ при $n \to \infty$;

$f_n \lesssim g_n, n \to \infty$, означает, что $f_n \le g_n(1 + \delta_n)$, где $\delta_n \to 0$ при $n \to \infty$;

$|V|$ - мощность множества V.

В разделе 4.2 доказаны следующие теоремы.

Теорема 4.2.3. Если $n \le m \le k^{n^\beta}$, $\beta < 1/2$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо

$$1)|B(L)| \approx \sum_{r=r_1(k)}^{r_2(k)} |B_r(L)| \approx \sum_{r=r_1(k)}^{r_2(k)} C_n^r k^r p_r(k)\,;$$

$$2)\ |U(L)| \approx \sum_{r=r_1(k)}^{r_2(k)} |U_r(L)| \approx \sum_{r=r_1(k)}^{r_2(k)} C_n^r k^r (1 - \exp(-mk^{-r}))^r\,;$$

$$3)\ \sum_{r \ge r_2(k)} |B_r(L)| \approx \left|B_{r_2(k)}(L)\right| \approx \sum_{r \ge r_2(k)} |S_r(L)| \approx \left|S_{r_2(k)}(L)\right| \approx$$

$$\approx \sum_{r \ge r_2(k)} |U_r(L)| \approx \left|U_{r_2(k)}(L)\right| \approx C_n^{r_2(k)} k^{r_2(k)} p_{r_2(k)}(k) \approx$$

$$\approx C_n^{r_2(k)} k^{r_2(k)} \left(1 - \exp\left(-mk^{-r_2(k)}\right)\right)^{r_2(k)}.$$

Теорема 4.2.4. Если $m \le k^{n^\beta}$, $\beta < 1/2$, то при $n \to \infty$ для почти всех матриц L из M_{mn}^k справедливо

$$1)\ \sum_{r \le r_1(k)} |B_r(L)| \approx \left|B_{r_1(k)}(L)\right| \approx \sum_{r \le r_1(k)} |C(L)| \approx \left|C_{r_1(k)}(L)\right| \approx$$

$$\approx C_n^{r_1(k)} k^{r_1(k)} p_{r_1(k)} \approx C_n^{r_1(k)} k^{r_1(k)} \exp\left(-mk^{-r_1(k)}\right);$$

$$2)\ \sum_{r \le r_1(k)} |U_r(L)| \approx \left|U_{r_1(k)}(L)\right| \approx C_n^{r_1(k)} k^{r_1(k)}.$$

Замечание. Оценки, приведенные в п. 1 теоремы 4.2.4 не являются новыми. Впервые эти оценки получены в работе Дюковой Е.В. (2002 г.). Однако в данной работе удалость получить более простое доказательство утверждения 1 теоремы 4.2.4.

Приведены аналоги теоремы 4.2.3 и 4.2.4 для неприводимых покрытий булевой матрицы.

Оценки, полученные в разделе 4.2, использованы в разделе 4.3 для получения асимптотик типичного числа максимальных конъюнкций и типичного ранга максимальной конъюнкции монотонной булевой функции F от n переменных, заданной конъюнктивной нормальной формой из m элементарных дизъюнкций.

Конъюнкция называется допустимой для F, если пересечение ее интервала истинности с множеством нулей функции F пусто. Конъюнкция называется неприводимой для F, если при удалении из нее хотя бы одного сомножителя увеличивается пересечение ее интервала истинности с множеством нулей функции F. Конъюнкция называется максимальной для F, если она допустимая и неприводимая.

Показано, что при $n \le m \le 2^{n^\beta}$, $\beta < 1/2$ ранги почти всех максимальных конъюнкций для почти всех монотонных булевых функций F от n переменных, заданной конъюнктивной нормальной формой из m элементарных дизъюнкций, принадлежат интервалу $[r_1, r_2]$. Число максимальных конъюнкций функции F с рангом, не меньше, чем r_2, асимптотически равно при $n \to \infty$ числу неприводимых конъюнкций функции F с рангом, не меньше, чем r_2. Число максимальных конъюнкций функции F с рангом, не превосходящим r_1, асимптотически равно при $n \to \infty$ числу допустимых конъюнкций функции F с рангом, не превосходящим r_1.

В разделе 4.4 показана асимптотическая эффективность класса алгоритмов поиска тупиковых покрытий целочисленной, в частности булевой, матрицы, основанных на переборе (перечислении) некоторого подмножества совместимых наборов столбцов этой матрицы.

Эффективность алгоритмов перечисления принято оценивать сложностью шага, то есть сложностью построения очередного элемента перечисляемого множества. Говорят, что алгоритм работает с

полиномиальной задержкой, если каждый его шаг выполняется за полиномиальное от размера задачи число элементарных операций. Под элементарной операцией понимается просмотр одного элемента рассматриваемой матрицы.

Пусть $Q(L)$ – конечная последовательность наборов столбцов матрицы L из M_{mn}^{k}, содержащая множество $B(L)$. Предполагается, что некоторые элементы в $Q(L)$ могут повторяться. Пусть алгоритм A строит с полиномиальной задержкой последовательность $Q(L)$, $N_A(L)$ - число шагов алгоритма A (число элементов в $Q(L)$). При построении очередного элемента из $Q(L)$ алгоритм A проверяет его на принадлежность $B(L)$. Очевидно, что такая проверка может быть осуществлена за полиномиальное от размеров матрицы число элементарных операций. Если построенный элемент принадлежит $B(L)$, то дополнительно проверяется, что этот элемент не был ранее построен алгоритмом A. На алгоритм A налагается условие, чтобы данная проверка также осуществлялась за полиномиальное от размеров матрицы число элементарных операций.

Алгоритм A является асимптотически оптимальным, если $N_A(L) \approx |B(L)|$ при $n \to \infty$ для почти всех матриц L из M_{mn}^{k}.

В работе введено понятие асимптотически эффективного алгоритма.

Алгоритм A является асимптотически эффективным, если $N_A(L) \approx d(m,n) \times |B(L)|$ при $n \to \infty$ для почти всех матриц L из M_{mn}^{k}, где $d(m,n)$ – полином от m и n.

Пусть алгоритм A^* строит множество тупиковых покрытий $B(L)$ матрицы $L \in M_{mn}^{k}$ путем перечисления совместимых наборов столбцов этой матрицы, $N_{A^*}(L)$ - число шагов алгоритма A^*. Доказана следующая теорема.

Теорема 4.4.2. Пусть $m^{\alpha} \le n \le k^{m^{\beta}}$, $\alpha > 1$, $\beta < 1/2$, то для почти всех матриц L из M_{mn}^{k} при $n \to \infty$ справедливо

$$\frac{N_{A^*}(L)}{|B(L)|} \lesssim 2\log_{\mathrm{k}}^{4} m.$$

Аналогично теореме 4.4.2 доказана асимптотическая эффективность алгоритмов поиска максимальных конъюнкций булевой функции, основанных на перечислении неприводимых конъюнкций этой функции.

ЗАКЛЮЧЕНИЕ

1. Разработаны и исследованы две модели алгоритмов распознавания, основанные на построении корректных наборов эл.кл. В первой модели строится один корректный набор эл.кл. с минимальной мощностью, во второй конструируется коллектив корректных наборов эл.кл. с хорошей распознающей способностью.
2. Разработаны и исследованы алгоритмы корректного понижения значности целочисленных данных, позволяющие эффективно решать задачи поиска оптимальной корректной перекодировки. Предложены и исследованы различные критерии качества корректных перекодировок.
3. Разработаны и исследованы генетические алгоритмы, эффективно решающие следующие задачи: поиск минимального покрытия булевой матрицы, построение минимального по сложности корректного набора эл.кл., конструирование коллектива корректных наборов эл.кл. с хорошей распознающей способностью, поиск оптимальной корректной перекодировки.
4. Введено понятие асимптотически эффективного алгоритма поиска тупиковых покрытий целочисленной матрицы (максимальных конъюнкций булевой функции).
5. Доказана асимптотическая эффективность алгоритмов поиска тупиковых покрытий целочисленной матрицы, основанных на перечислении с полиномиальной задержкой совместимых наборов столбцов этой матрицы. Аналогичный результат получен для алгоритмов поиска максимальных конъюнкций булевой функции, основанных на перечислении неприводимых конъюнкций этой функции.

6. Показано, что при $n \leq m$, для почти всех матриц L из M_{mn}^{k} при $n \to \infty$ число всех тупиковых покрытий и число всех совместимых наборов столбцов матрицы L асимптотически равно, соответственно, числу тупиковых покрытий и числу совместимых наборов столбцов с длинами из интервала $[r_1(k), r_2(k)]$. Аналогичный результат получен для задач преобразования нормальных форм булевой функции.
7. Показано, что при $n \leq m$, для почти всех матриц L из M_{mn}^{k} при $n \to \infty$ число совместимых наборов столбцов с длиной не больше, чем $r_1(k)$ асимптотически совпадает с числом наборов столбцов матрицы с длиной не больше, чем $r_1(k)$. Аналогичный результат получен для задач преобразования нормальных форм булевой функции.
8. Показано, что при $n \leq m$, для почти всех матриц L из M_{mn}^{k} при $n \to \infty$ число тупиковых покрытий с длиной не меньше, чем $r_2(k)$, асимптотически совпадает при $n \to \infty$ с числом совместимых наборов столбцов с длиной не меньше, чем $r_2(k)$, и с числом σ – подматриц с рангом не меньше, чем $r_2(k)$. Аналогичный результат получен для задач преобразования нормальных форм булевой функции.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Sotnezov R.M. Genetic Algorithms in Problems of Discrete Optimization and Recognition // 9th International Conference “Pattern Recognition and Image Analysis: New Information Technologies” (PRIA-9-2008): Conference Preceedings. Vol. 2. – Nizhni Novgorod, 2008. P. 173-175.
2. Сотнезов Р.М. Генетические алгоритмы в задаче о покрытии. Сборник тезисов лучших дипломных работ 2008 года. М. Издательский отдел факультета ВМиК МГУ, 2008, с. 73-74.
3. Sotnezov R.M. Genetic Algorithms for Problems of Logical Data Analysis in Discrete Optimization and Image Recognition // Pattern Recognition and Image Analysis, 2009, Vol. 19, No 3, pp. 469-477

4. Дюкова Е.В., Сизов А.В., Сотнезов Р.М. Об одном методе построения приближённого решения для задачи о покрытии // Доклады 14-й Всероссийской конференции «Математические методы распознавания образов». М.: МАКС Пресс, 2009. С. 241-243.
5. Сотнезов Р.М. Генетические алгоритмы в задаче о покрытии // Восьмая международная конференция «Дискретные модели в теории управляющих систем». Москва, 2009 г. Электронный сборник материалов конференции. С.179-183 (http://dmconf.ru/dm8/proceedings.pdf).
6. Дюкова Е.В., Сотнезов Р.М. О сложности дискретных задач перечисления // Докл. Акад. Наук. 2010. Т. 143. №1. С. 11-13.
7. Djukova E.V., Zhuravlev Yu.I., Sotnezov R.M. Synthesis of Corrector Family with High Recognition Ability // New Trends in Classification and Data Mining. Sofia, 2010. – P. 32-39.
8. Дюкова Е.В., Сотнезов Р.М. О сложности перечисления элементарных классификаторов в логических процедурах распознавания // Интеллектуализация обработки информации: 8-я международная конференция. Кипр, г. Пафос, 17-23 октября 2010 г.: Сборник докладов. – М.: МАКС Пресс, 2010. – С. 43-46.
9. Дюкова Е.В., Сотнезов Р.М. Асимптотические оценки числа решений задачи дуализации и ее обобщений // Ж. вычисл. матем. и матем. физ. 2011. Том 51, № 8. С. 1531-1540.
10. Djukova E.V., Zhuravlev Yu.I., Sotnezov R.M. Construction of an Ensemble of Logical Correctors on the Basis of Elementary Classifiers // Pattern Recognition and Image Analysis, 2011, Vol. 21, No. 4, pp. 599–605.
11. Дюкова Е.В., Сизов А.В., Сотнезов Р.М. О корректном понижении значности данных в задачах распознавания // Доклады Всероссийской конференции «Математические методы распознавания образов» (ММРО-15), г. Петрозаводск, 11-17 сентября 2011 г. С. 80-83.