На правах рукописи

СОТНЕЗОВ Роман Михайлович

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

01.01.09 – Дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва - 2012

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, доцент Дюкова Елена Всеволодовна

Официальные оппоненты: Обухов Юрий Владимирович, доктор физико-математических наук, профессор, Федеральное государственное бюджетное учреждение науки Институт радиотехники и электроники им. В.А. Котельникова Российской академии наук, заведующий лабораторией.

Вялый Михаил Николаевич, кандидат физико-математических наук, Федеральное государственное бюджетное учреждение науки Вычислительный центр им. А.А. Дородницына Российской академии наук, старший научный сотрудник.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Научно-исследовательский институт системных исследований Российской академии наук.

Защита состоится «___» _____ 2012 г. в _____ часов на заседании диссертационного совета Д002.017.02 при Федеральном государственном бюджетном учреждении науки Вычислительном центре им. А.А. Дородницына Российской академии наук по адресу: 119333, Москва, ул. Вавилова, 40, конференц-зал.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН

Автореферат разослан «___» _____ 2012 г.

Ученый секретарь
диссертационного совета Д002.017.02
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.

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.

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 основой логического синтеза распознающих процедур.

Представляет интерес использование конструкций алгебраического подхода для построения корректных распознающих процедур на базе произвольных наборов элементарных классификаторов, т.е. элементарных классификаторов необязательно являющихся корректными. Идея алгебро-логического синтеза корректных распознающих процедур предложена в [1].

Однако вопросы, связанные с практическим применением логического корректора, в [1] не исследовались.

Логический анализ данных в распознавании особенно эффективен в случае дискретной (целочисленной) информации низкой значности, например, бинарной. Веществennозначная информация часто рассматривается как целочисленная высокой значности. Поэтому актуальной является задача корректного понижения значности исходных целочисленных данных.

Практическое использование логических процедур распознавания напрямую связано со снижением их вычислительной сложности. При большой размерности обучающей выборки возникает необходимость рассматривать труднорешаемые дискретные задачи перечисления решений. Это задачи поиска покрытий булевых и целочисленных матриц и построения нормальных форм логических функций. Е.В. Дюковой (1977 г.) 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. Обоснование данного подхода базируется на получении асимптотик для типичного числа решений каждой из задач Z и Z_1. Подход хорошо зарекомендовал себя на практике.

К настоящему моменту асимптотически оптимальные алгоритмы поиска тупиковых покрытий булевых и целочисленных матриц (построения нормальных форм логических функций) предложены для достаточно узкого класса задач. В тех случаях, когда не удается построить асимптотические оптимальные алгоритмы, имеет смысл предъявлять более слабые требования к эффективности алгоритма.

При предварительном анализе обучающей выборки и конструировании логических процедур распознавания часто возникают дискретные оптимизационные задачи. Для их решения наряду с методами, имеющими

теоретическое обоснование, необходимо разрабатывать эвристические подходы, дающие хорошее приближенное решение.

1. Дюкова Е.В., Журавлев Ю.И., Рудаков К.В. 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. Были выделены следующие основные направления исследований.

2. Разработка новых подходов к повышению эффективности решения задачи распознавания по прецедентам методами логического и алгебро-логического анализа данных.
   2.1. Построение и исследование новых моделей корректных распознающих процедур на базе произвольных элементарных классификаторов.
   2.2. Развитие методов корректного понижения значности целочисленных данных в задачах распознавания. Использование новых критериев качества перекодирования.
3. Получение новых результатов, касающихся снижения вычислительной сложности логического и алгебро-логического анализа данных в распознавании.
   3.1. Построение генетических алгоритмов, эффективно решающих оптимизационные задачи, возникающие при алгебро-логическом синтезе распознающих процедур и в задачах корректного понижения значности целочисленной информации.
   3.2. Построение и обоснование эффективных алгоритмов поиска тупиковых покрытий булевых и целочисленных матриц для случаев, в которых не удается построить асимптотически оптимальные алгоритмы. Получение аналогичных результатов для задач построения нормальных форм логических функций.
   3.3. Получение новых асимптотических оценок для количественных характеристик множества покрытий булевых и целочисленных матриц.

Получение аналогичных оценок для нормальных форм логических функций.

Научная новизна. Разработаны и успешно апробированы на реальных задачах новые модели распознающих процедур, основанные на построении логических корректоров. Для снижения вычислительной сложности моделей использован генетический подход.

С использованием генетического подхода построены новые алгоритмы корректного понижения значности исходной целочисленной информации. Показано, что применение этих алгоритмов позволяет повысить качество распознавания алгоритма голосования по представительным наборам, сконструированного по перекодированным данным, существенно не увеличивая вычислительных затрат.

Получены асимптотики для типичного числа тупиковых покрытий и типичной длины тупикового покрытия целочисленной матрицы в случае, когда число столбцов матрицы не превосходит числа её строк. Аналогичные результаты получены для соответствующих количественных характеристик множества максимальных конъюнкций двузначной логической функции, заданной множеством нулей, при условии, что число переменных, от которых зависит функция, не превосходит числа ее нулей.

Введено понятие асимптотически эффективного алгоритма для труднорешаемой перечислительной задачи поиска тупиковых покрытий целочисленной матрицы. Доказана асимптотическая эффективность алгоритмов построения тупиковых покрытий целочисленной матрицы, основанных 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 двузначной 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 функций. Применяются методы построения покрытий булевых и целочисленных матриц, а также методы получения асимптотик для типичных значений количественных характеристик множеств покрытий булевых целочисленных матриц.

Теоретическая и практическая ценность. Результаты, полученные в диссертационной работе, могут быть использованы в теоретических и практических исследованиях, касающихся построения эффективных реализаций для логических процедур распознавания. Эти результаты могут быть также использованы при разработке спецкурсов по распознаванию образов и дискретной математики, преподаваемых в госуниверситетах для студентов математических специальностей. Эффективность предложенных подходов подтверждена решением реальных задач.

Апробация работы. Основные положения и результаты диссертации докладывались на следующих 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 из списка, рекомендованного ВАК РФ. Основные результаты, полученные 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, заключения и списка литературы из 68 наименований. Общий объем работы – 112 страниц.

СОДЕРЖАНИЕ РАБОТЫ

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

В главе 1 рассмотрена задача построения корректных 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 корректным наборам элементарных классификаторов классов. При конструировании этих алгоритмов использован генетический подход. Проведено тестирование построенных алгоритмов на реальных прикладных задачах.

Задача распознавания по прецедентам рассматривается в стандартной постановке для целочисленной информации. 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, используемых при построении логических процедур распознавания, является понятие элементарного классификатора.

Пусть $H = \{x_{j_1}, \dots, x_{j_r}\}$ - набор из r различных признаков, $r \le n$, 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

$$B_{(H,\sigma)}(S) = \begin{cases} 1, & \text{если } a_{j_i} = \sigma_i, i = 1,2, \dots, r, \\ 0, & \text{в противном случае.} \end{cases}$$

Пусть $U = \{(H_1, \sigma_1), \dots, (H_q, \sigma_q)\}$ - набор эл.кл, $S \in M$. Положим $\omega_U(S) = \left(B_{(H_1,\sigma_1)}(S), \dots, B_{(H_q,\sigma_q)}(S)\right)$.

Определение 1.1.1. 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 $S' \in K$, $S'' \notin K$ выполняется неравенство

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

Определение 1.4.1. Корректный набор эл.кл. U с корректирующей функцией $F_{U,K}$ называется *монотонным корректным* для класса K, если $F_{U,K}$ – монотонная булева функция и $F_{U,K}(\omega_U(S')) = 1$ для любого обучающего объекта S' класса K.

(Монотонный) корректный набор эл.кл. U класса K называется тупиковым, если любое его подмножество не является (монотонным) корректным для K. (Монотонный) корректный набор эл.кл. U класса K 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, $K \in \{K_1, \dots, K_l\}$, распознающий алгоритм A конструирует некоторое подмножество $W_A(K)$ множества корректных наборов эл.кл. класса K. Распознавание объекта S производится на основе вычисления оценок принадлежности этого объекта к классам $K_1, \dots, K_l$. Оценки вычисляются следующим образом.

Пусть $R(T, K)$ – множество обучающих объектов из T, принадлежащих K, $R(T, \overline{K})$ – множество обучающих объектов из T, не принадлежащих K.

Случай 1. Множество $W_A(K)$ состоит из корректных наборов эл.кл. класса K, не обязательно являющихся монотонными. Пусть $U \in W_A(K)$, $|U| = q$. Положим

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

Тогда оценка принадлежности объекта S к классу K имеет вид

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

где суммирование проводится по всем (U, S') из $W_A(K) \times R(T, K)$.

Случай 2. Множество $W_A(K)$ состоит только из монотонных корректных наборов эл.кл. класса K. Пусть $U \in W_A(K)$, $|U| = q$. Положим

$$\delta_U^2(S', S) = \begin{cases} 1, \text{если } \omega_U(S) \succcurlyeq \omega_U(S'), \\ 0, \text{иначе,} \end{cases}$$

где $\omega_U(S) \succcurlyeq \omega_U(S')$, $\omega_U(S) = (a_1, \dots, a_q)$, $\omega_U(S') = (a_1', \dots, a_q')$, означает, что $a_i \geq a_i'$ при $i = 1,2, \dots, q$.

В этом случае оценка принадлежности объекта S к классу K имеет вид

$$\Gamma_2(S, K) = \frac{1}{|R(T, K)||W_A(K)|} \sum_{(U, S')} \delta_U^2(S', S),$$

где суммирование проводится по всем (U, S') из $W_A(K) \times R(T, K)$.

Разработаны и реализованы распознающие алгоритмы А1, А2, А3, А4, работающие по описанной выше схеме. В качестве базисных алгоритмов используются одноэлементные эл.кл., то есть эл.кл. вида (x_j, a), где a – допустимое значение признака x_j. Для построения корректных наборов эл.кл. используются генетические алгоритмы, каждый из которых запускается $10 - 15$ раз для получения наилучшего результата. Особями популяции в генетических алгоритмах являются корректные наборы эл.кл.

Алгоритмы А1 и А2 решают задачу построения коллектива корректных наборов эл.кл., в котором каждый набор обладает хорошей распознающей способностью. Алгоритм А1 решает указанную задачу в случае 2, алгоритм А2 в случае 1.

Для построения коллектива корректных наборов эл.кл. исходная выборка разбивается случайным образом на две подвыборки T_0 и T_1 таких, что $|T_0|/|T_1| = 4$. Подвыборка T_0 используется для построения корректных наборов эл.кл., а подвыборка T_1 для оценки качества распознавания построенных корректных наборов эл.кл. Пусть $Q_0 = R(T_0, K)$, $Q_1 = R(T_1, K)$, $\overline{Q_1} = R(T_1, \overline{K})$. Оценка распознающей способности корректного набора эл.кл. имеет следующий вид

$$\tau_{A,K}(U) = \frac{1}{|Q_1|} \sum_{(S,S') \in Q_0 \times Q_1} \delta_U(S, S') - \frac{1}{|\overline{Q_1}|} \sum_{(S,S') \in Q_0 \times \overline{Q_1}} \delta_U(S, S'),$$

где $\delta_U(S, S') = \delta_U^2(S, S')$ для А1 и $\delta_U(S, S') = \delta_U^1(S, S')$ для А2.

Алгоритмы А3 и А4 решают задачу построения одного корректного набора эл.кл. по мощности близкого к минимальному соответственно в случаях 2 и 1. Функция приспособленности особи в генетическом алгоритме – мощность корректного набора эл.кл.

Тестирование разработанных алгоритмов проводилось на реальных задачах. Наилучшие результаты показал алгоритм А1.

В главе 2 рассмотрена задача корректного понижения значности данных, которая возникает в случае применения логических процедур распознавания к целочисленной информации высокой значности.

Эта задача ставится следующим образом. По обучающей выборке T строится специальная булева матрица L_T, строкам которой соответствуют пары обучающих объектов из разных классов, а столбцы 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 в диссертационной работе разработаны генетические алгоритмы поиска оптимальной корректной перекодировки исходной информации, которые описаны в главе 3. В качестве особей используются кодирующие покрытия, в качестве функций приспособленности используется один из двух функционалов:

$$f_1(H) = \sum_{j \in R_2(H)} c_j\,; \quad f_2(H) = \frac{1}{|H|} \sum_{j \in R_1(H)} \frac{1}{c_j},$$

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. Ставится задача минимизации функционалов $f_1(H)$ и $f_2(H)$.

Проведено тестирование разработанных алгоритмов на реальных задачах. Показано, что данная методика позволяет повысить качество распознавания алгоритма голосования по представительным наборам, сконструированного по перекодированным данным, существенно не увеличивая вычислительных

затрат. Предыдущие методики были основаны на других критериях качества кодирующего покрытия и требовали чрезвычайно больших 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 столбцов, входящих в покрытие, 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 показали, 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* (А.А. Агеев, 2004 г.). Сравнение на большом числе случайных матриц показало, что генетические алгоритмы, как правило, превосходят по точности решения как градиентный алгоритм, так и алгоритм *General*, что говорит об их практической ценности.

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 длина выдаваемого покрытия. При этом в случаях, когда уменьшение средней длины покрытия не происходит, наблюдается уменьшение времени поиска лучшего решения.

Модификации разработанных генетических алгоритмов использованы в главе 1 и 2 для задачи поиска корректного набора эл.кл., близкого к минимальному, задачи построения коллектива корректных наборов эл.кл. с хорошей распознающей способностью и задачи построения оптимальной корректной перекодировки.

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 двузначной логической функции. Введено понятие асимптотически эффективного алгоритма для задачи перечисления тупиковых покрытий целочисленной матрицы. Показана асимптотическая эффективность алгоритма поиска тупиковых покрытий целочисленной матрицы, основанного на перечислении с полиномиальной задержкой «совместимых наборов» столбцов этой матрицы. Получены аналогичные результаты для задач построения нормальных форм логических функций.

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* $\sigma-$*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 $(\beta_1,..,\beta_r) \in E_k^r$, в которой $\beta_p \neq \sigma_p$ и $\beta_j = \sigma_j$ при $j \in \{1,2,\dots,r\}\backslash\{p\}$, т.е. L^Hсодержит $\sigma-$*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 σ −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

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

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

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

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

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

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

$f_n \approx g_n, n \to \infty$, 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 4.2.1. 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

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

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

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

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

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

Теорема 4.2.2. Если $m \le k^{n^\beta}, \beta < 1/2$, то при $n \to \infty$ для почти всех матриц L из M_{mn}^k справедливо

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

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

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

Замечание 1. До настоящего момента для рассматриваемого случая $n \le m \le k^{n^\beta}$, $\beta < 1/2$, была известна лишь асимптотика $\log_k |B(L)|$ (Е.В. Дюкова, 2007 г.).

Замечание 2. Оценки, приведенные в утверждении 1 теоремы 4.2.2 впервые получены Е.В. Дюковой (2002 г.). В данной работе удалось получить более простое доказательство этого утверждения.

Приведены аналоги теорем 4.2.1 и 4.2.2 для количественных характеристик неприводимых покрытий булевой матрицы и $(0,0,\dots,0)-$ совместимых наборов столбцов булевой матрицы.

Таким образом, показано, что при $n \le m \le k^{n^\beta}$, $\beta < 1/2$ 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, длины которых принадлежит интервалу $[r_1(k), r_2(k)]$. Также показано, что при $n \le m \le k^{n^\beta}$, $\beta < 1/2$ 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 равно при $n \to \infty$, во-первых, числу совместимых наборов столбцов с длиной не меньше, чем $r_2(k)$, и, во-вторых, числу σ – подматриц с рангом $r_2(k)$.

В разделе 4.3 приведены аналогичные оценки для количественных характеристик множества максимальных конъюнкций двузначной логической функции F от n переменных с m нулями.

В разделе 4.4 введено понятие асимптотически эффективного алгоритма для задачи перечисления тупиковых покрытий целочисленной матрицы.

Показана асимптотическая эффективность алгоритмов поиска покрытий из $B(L)$, $L \in M_{mn}^{k}$, основанных на переборе (перечислении) с полиномиальной задержкой множества совместимых наборов столбцов матрицы L или некоторого его подмножества.

Эффективность алгоритмов решения дискретных перечислительных задач принято оценивать сложностью шага, т.е. сложностью построения очередного решения. Говорят, что алгоритм работает с полиномиальной задержкой, если каждый его шаг выполняется за полиномиальное от размера задачи время. На данный момент алгоритм с полиномиальной задержкой, перечисляющий все тупиковые покрытия целочисленной матрицы, не построен и неизвестно существует ли он.

В диссертационной работе рассматривается другой подход к построению алгоритмов для труднорешаемых перечислительных дискретных задач, который основан на понятии асимптотически эффективного алгоритма.

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 $Q(L)$ с полиномиальной задержкой, т.е. на каждом шаге строится очередной набор из $Q(L)$ и при этом выполняется не более d элементарных операций, где d ограничено сверху полиномом от m и n. Пусть $N_A(L)$ - 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 размера матрицы время. Если построенный набор столбцов принадлежит $B(L)$, то дополнительно проверяется, был ли этот набор построен ранее алгоритмом A. Требуется, чтобы данная проверка осуществлялась за полиномиальное от размера матрицы время.

Алгоритм A является *асимптотически эффективным*, если $N_A(L) \lesssim$ $\lesssim p(m,n) \times |B(L)|$ при $n \to \infty$ для почти всех матриц L из M_{mn}^{k}, где $p(m,n)$ – полином от m и n. Если $p(m,n) = 1$, то алгоритм A называется *асимптотически оптимальным*.

Пусть алгоритм A строит множество тупиковых покрытий $B(L)$ матрицы $L \in M_{mn}^k$ путем перечисления (без повторений) с полиномиальной задержкой множества $U(L)$ или некоторого его подмножества, $N_A(L)$ - число шагов алгоритма A. Известно, что в случае $m^\alpha \leq n \leq k^{m^\beta}$, $\alpha > 1$, $\beta < 1$, алгоритм A является асимптотически оптимальным. В диссертационной работе доказана следующая

Теорема 4.4.4. Если $n \leq m \leq k^{n^\beta}$, $\beta < 1/2$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо

$$\frac{N_A(L)}{|B(L)|} \lesssim (\log_k m)^{k^2}.$$

Приведен аналог теоремы 4.4.4 для алгоритма поиска неприводимых покрытий булевой матрицы, основанного на перечислении с полиномиальной задержкой $(0,0,\dots,0)$ −совместимых наборов столбцов булевой матрицы.

Приведен также аналог теоремы 4.4.4 для алгоритма построения сокращенной дизъюнктивной нормальной формы двузначной логической функции от n переменных с m нулями, основанного на перечислении с полиномиальной задержкой 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., каждый из которых обладает хорошей распознающей способностью.
2. Разработаны и исследованы алгоритмы корректного понижения значности целочисленных данных в задачах распознавания. Рассмотрены вопросы снижения вычислительной сложности этих алгоритмов. 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. 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 двузначной логической функции.
5. Введено понятие асимптотически эффективного алгоритма поиска тупиковых покрытий целочисленной матрицы (максимальных конъюнкций двузначной логической функции).
6. 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 двузначной логической функции, основанных на перечислении с полиномиальной задержкой неприводимых конъюнкций этой 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.