На правах рукописи

Прокофьев Пётр Александрович

Корректное распознавание по прецедентам: построение логических корректоров общего вида и вычислительные аспекты

Специальность 05.13.17 — Теоретические основы информатики

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2016

Работа выполнена в Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Научный руководитель: доктор физико-математических наук, доцент
Дюкова Елена Всеволодовна

Официальные оппоненты:

доктор физико-математических наук, профессор
Двоенко Сергей Данилович,
Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет», профессор,

кандидат технических наук,
Игнатов Дмитрий Игоревич,
Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики»», доцент.

Ведущая организация: Федеральное государственное бюджетное учреждение науки «Институт математики и механики им. Н. Н. Красовского» Уральского отделения Российской академии наук

Защита состоится 6 октября 2016 г. в 14 часов на заседании диссертационного совета Д 002.073.05 на базе Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» по адресу: 119333, Москва, ул. Вавилова, дом 40, конференц-зал.

С диссертацией можно ознакомиться в библиотеке Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» и на сайте `web.frccsc.ru`.

Автореферат разослан «____» _____ 2016 г.

Ученый секретарь
диссертационного совета
Д 002.073.05, д.ф.-м.н., профессор В. В. Рязанов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Центральной проблемой, на которую нацелена диссертационная работа, является корректное распознавание по прецедентам. Исследуется множество объектов, которое может быть разбито на конечное число классов. О характере этого разбиения можно судить только по обучающей выборке (конечному набору прецедентов). Каждый объект может быть представлен в виде числового вектора, полученного в результате наблюдения или измерения определённых характеристик объекта. Такие характеристики называются признаками. Требуется построить алгоритм распознавания, который по предъявленному признаковому описанию объекта определяет, к какому классу следует отнести этот объект. Алгоритм распознавания, безошибочно классифицирующий прецеденты, называется корректным. Важным показателем качества корректного алгоритма распознавания является его обобщающая способность (частота ошибок на объектах, не участвующих в обучении).

В случае целочисленных признаков задача корректного распознавания достаточно эффективно решается методами *логического подхода*[1]. Базовым для этого подхода является понятие *элементарного классификатора* (эл.кл.) — элементарной конъюнкции, заданной на признаковых описаниях объектов. Говорят, что эл.кл. выделяет некоторый объект, если он принимает значение 1 на признаковом описании этого объекта. Традиционно при построении логических алгоритмов распознавания используются корректные эл.кл. Эл.кл. называется корректным для некоторого класса, если совокупность выделяемых им прецедентов является подмножеством либо этого класса, либо объединения остальных классов. Если все прецеденты, выделяемые корректным эл.кл., принадлежат одному классу, то такой эл.кл. называется представительным набором. Известно, что алгоритмы голосования по представительным наборам наиболее успешно применяются для задач распознавания с признаками небольшой значности (под значностью признака понимается число его допустимых значений). В этом случае, как правило, удаётся найти достаточное количество информативных представительных наборов.

Проблемными для классических логических алгоритмов распознавания являются задачи с вещественными признаками и целочисленными признаками большой значности. Для повышения эффективности решения таких задач применяются следующие методики: 1) ищутся логические закономерности (понятие логической закономерности обобщает понятие эл.кл. на случай вещественных признаков)[2]; 2) вещественные признаки трактуются как целочисленные высокой значности и выполняется корректная

[1] *Журавлев Ю. И., Дмитриев А. Н., Кренделев Ф. П.* О математических принципах классификации предметов и явлений // Дискретный анализ, Сб. научн. тр. Т. 7. Ин-т математики СО АН СССР Новосибирск, 1966. С. 3—15 ; *Вайнцвайг М. Н.* Алгоритм обучения распознаванию образов «Кора». М.: Советское радио, 1973. С. 110—116 ; *Дюкова Е. В., Песков Н. В.* Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // ЖВМ и МФ. 2002. Т. 42, № 5. С. 741—753.

[2] *Ковшов Н. В., Моисеев В. Л., Рязанов В. В.* Алгоритмы поиска логических закономерностей в задачах распознавания // ЖВМ и МФ. 2008. Т. 48, № 2. С. 329—344.

перекодировка признаков с целью понижения их значности[3]; 3) строятся корректные алгоритмы распознавания на базе произвольных, не обязательно корректных эл.кл. (*алгебро-логический подход*)[4].

В основе алгебро-логического синтеза распознающих алгоритмов лежат понятия и методы двух подходов: логического и алгебраического. *Алгебраический подход*, развиваемый школой Ю. И. Журавлёва[5], применяется, когда требуется скорректировать работу нескольких различных алгоритмов, каждый из которых безошибочно классифицирует лишь часть обучающих объектов. Цель коррекции — сделать так, чтобы ошибки одних алгоритмов были скомпенсированы другими, и качество результирующего алгоритма оказалось лучше, чем каждого из базовых алгоритмов в отдельности.

В работе Е. В. Дюковой, Ю. И. Журавлёва, К. В. Рудакова вводится понятие *корректного набора эл.кл.*, которое впоследствии становится основным для алгебро-логического подхода[6]. Алгоритмы распознавания, основанные на голосовании по корректным наборам эл.кл., называются *логическими корректорами*. Фактически эл.кл. выступают в роли базовых распознающих алгоритмов и корректируются булевыми функциями. Основной задачей этапа обучения логических корректоров является поиск корректных наборов эл.кл. с хорошей распознающей способностью. Каждый корректный набор эл.кл. однозначно соответствует покрытию булевой матрицы, построенной специальным образом по обучающей выборке. При большой значности признаков приходится обрабатывать матрицы, размер которых экспоненциально зависит от объема обучающей информации. Поэтому возникает проблема применения логических корректоров на практике.

В работе Е. В. Дюковой, Ю. И. Журавлёва, Р. М. Сотнезова разработаны первые практические модели логических корректоров[7]. Для снижения вычислительных затрат предложено использовать эл.кл. ранга 1 и поиск корректных наборов эл.кл. с распознающей способностью, близкой к максимальной, осуществлять генетическим алгоритмом. Установлено, что логические корректоры с монотонными корректирующими функциями (монотонные логические корректоры) имеют более высокую обобщающую способность, чем с произвольными.

Проведённые в работе М. М. Любимцевой эксперименты показывают, что на прикладных задачах с большой значностью признаков монотонные логические корректоры опережают классические логические алгоритмы распознавания[8]. В случае небольшой значности признаков ситуация обратная. По-видимому, ограничение, налагаемое на

---

[3]Обработка вещественнозначной информации логическими процедурами распознавания / Е. В. Дюкова [и др.] // Искусственный интеллект. НАН Украины. 2004. № 2. С. 80—85.

[4]*Дюкова Е. В., Журавлев Ю. И., Рудаков К. В.* Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // ЖВМ и МФ. 1996. Т. 36, № 8. С. 216—223.

[5]*Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5—68.

[6]*Дюкова Е. В., Журавлев Ю. И., Рудаков К. В.* Указ. соч.

[7]*Djukova E. V., Zhuravlev Y. I., Sotnezov R. M.* Construction of an ensemble of logical correctors on the basis of elementary classifiers // Pattern Recognition and Image Analysis. 2011. Т. 21, № 4. С. 599—605.

[8]*Любимцева М. М.* Логические корректоры в задачах распознавания // Сборник тезисов лучших дипломных работ факультета ВМК МГУ 2014 года — М: МАКС ПРЕСС. 2014. С. 47—49.

ранг эл.кл., не позволяет построить в последнем случае логические корректоры с хорошей обобщающей способностью.

Актуальной задачей является расширение границ применимости алгебро-логического подхода за счёт построения и исследования новых, более совершенных моделей логических корректоров. Перспективным направлением является использование других семейств корректирующих функций, отличных от семейства монотонных булевых функций и множества всех булевых функций. Также необходимо разработать методику обучения логических корректоров, позволяющую с небольшими вычислительными затратами получать высокое качество распознавания.

Трудности вычислительного характера, возникающие при реализации как классических логических алгоритмов распознавания, так и логических корректоров, связаны с необходимостью решать известные своей сложностью дискретные задачи. Среди этих задач главной считается *дуализация*. Это задача перечисления неприводимых покрытий булевой матрицы. Говорят, что алгоритм дуализации имеет полиномиальную задержку, если каждый его шаг (построение очередного решения) осуществляется за время, полиномиально зависящее от размера входа[9]. Вопрос о полиномиальной разрешимости дуализации поставлен более 40 лет назад, однако до сих пор ответ на этот вопрос не найден. В зарубежной литературе наибольшее распространение получил инкрементальный принцип построения алгоритмов дуализации, и в этом направлении лучшим теоретическим результатом считается построение инкрементальных алгоритмов, квазиполиномиальная сложность которых обоснована «в худшем» случае[10]. Однако на реальных задачах наилучшие результаты показывают так называемые асимптотически оптимальные алгоритмы дуализации, имеющие теоретическое обоснование эффективности «в среднем»[11].

Цель данной работы — развитие методов алгебро-логического подхода к корректному распознаванию по прецедентам, а именно построение логического корректора общего вида, позволяющего в определенной степени повысить качество распознавания и снизить вычислительные затраты этапа обучения; разработка конструкций асимптотически оптимальных алгоритмов дуализации для решения задач большого размера.

Решена следующая группа задач.

1. Обобщено понятие корректного набора эл.кл. Описана схема логического корректора общего вида. Выявлено место классических логических алгоритмов распознавания и ранее построенных логических корректоров в этой схеме.

2. Разработана и исследована более совершенная модель логического корректора с корректирующими функциями из семейства, отличного от семейства монотонных булевых функций и множества всех булевых функций.

[9] *Johnson D., Yannakakis M., Papadimitriou C.* On generating all maximal independent sets // Information Processing Letters. 1988. Т. 27, № 3. С. 119—123.

[10] *Fredman M. L., Khachiyan L.* On the complexity of dualization of monotone disjunctive normal forms // Journal of Algorithms. 1996. Т. 21, № 3. С. 618—628.

[11] *Дюкова Е. В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. 1997. Т. 233, № 4. С. 527—530 ; *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // Журнал вычислительной математики и математической физики. 2004. Т. 44, № 3. С. 562—572.

3. Разработана методика повышения качества распознавания и скорости обучения логических корректоров. Проведено экспериментальное обоснование эффективности предложенной методики.

4. Модифицированы конструкции ряда асимптотически оптимальных алгоритмов дуализации с целью снижения времени их работы. Экспериментально показано превосходство построенных алгоритмов дуализации по сравнению с другими известными алгоритмами дуализации.

Методы исследования. Применялись методы дискретной математики, алгебры, математической логики, анализа алгоритмов и вычислительной сложности. Экспериментальное исследование проводилось с использованием программно-алгоритмического комплекса, разработанного автором.

Научная новизна. В работе строится логический корректор общего вида, для описания которого используется язык предикатов. Вводятся понятия корректного и представительного предиката. Каждый предикат однозначно определяется некоторым набором эл.кл. и корректирующей функцией этого набора.

Впервые решается важная методологическая задача обобщения логического и алгебро-логического синтеза корректных алгоритмов распознавания. Предложенная в работе схема синтеза корректных алгоритмов распознавания может быть использована для описания как классических логических распознающих алгоритмов, так и ранее построенных логических корректоров.

В рамках общей схемы построена новая модель практического логического корректора POLAR, голосующего по предикатам специального вида и имеющего поляризуемую корректирующую функцию. Булева функция называется *поляризуемой*, если она по каждой переменной либо монотонно не возрастает, либо монотонно не убывает. Семейство монотонных булевых функций содержится в семействе поляризуемых булевых функций. Ранее поляризуемые функции общего вида в качестве корректирующих не использовались.

Предложена новая методика снижения вычислительных затрат и повышения качества распознавания логических корректоров. На этапе обучения логического корректора семейства голосующих предикатов формируются итеративно по принципу *бустинга*[12]. Снято ограничение на ранг эл.кл., и поиск корректных наборов эл.кл. осуществляется в рамках *локальных базисов классов* — предварительно построенных корректных наборов, состоящих из информативных эл.кл. Разработаны итеративные алгоритмы формирования «хороших» локальных базисов. Вообще говоря, идея применения локальных базисов в алгебраическом подходе впервые встречается в работах К. В. Воронцова[13].

В диссертационной работе построен ряд новых асимптотически оптимальных алгоритмов дуализации, в основе которых лежит следующий подход. Исходная перечислительная задача Z заменяется на более «простую» перечислительную задачу Z_1,

[12] Boosting the margin: a new explanation for the effectiveness of voting methods / R. E. Schapire [и др.] // Annals of Statistics. 1998. Т. 26, № 5. С. 1651—1686.

[13] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распознавания // ЖВМ и МФ. 1998. Т. 38, № 5. С. 870—880.

имеющую тот же вход и решаемую с полиномиальной задержкой. При этом, во-первых, множество решений задачи Z_1 содержит множество решений задачи Z, и во-вторых, почти всегда с ростом размера входа число решений задачи Z_1 асимптотически равно числу решений задачи Z. Теоретическое обоснование данного подхода базируется на получении асимптотик для типичного числа решений каждой из задач Z и Z_1.

Таким образом, в отличие от «точного» алгоритма с полиномиальной задержкой асимптотически оптимальному алгоритму разрешено делать «лишние» полиномиальные шаги. Лишний шаг — это построение такого решения задачи Z_1, которое либо было найдено ранее, либо построено впервые, но не является решением задачи Z. Проверка того, является ли выполненный шаг лишним должна осуществляться за полиномиальное время от размера задачи.

Работу асимптотически оптимального алгоритма дуализации A на входной матрице L наглядно можно представить в виде обхода в глубину дерева решений $T_A(L)$. Корнем дерева $T_A(L)$ является пустой набор, остальным вершинам соответствуют наборы столбцов матрицы L. Построение висячей вершины связано либо с получением неприводимого покрытия матрицы L, либо с завершением «лишнего» шага алгоритма. Если вершина H не является висячей, то каждая её дочерняя вершина образуется добавлением к H в точности одного столбца.

Построенные в работе асимптотически оптимальные алгоритмы дуализации являются лидерами по скорости счёта. Снижение вычислительных затрат достигается за счёт сокращения общего числа вершин дерева решений. Ранее при построении асимптотически оптимальных алгоритмов основные усилия по уменьшению времени счёта направлялись на сокращение числа висячих вершин дерева решений (числа лишних шагов). При этом, как правило, усложнялся шаг алгоритма.

Теоретическая значимость. Построена общая схема алгебро-логического синтеза корректных алгоритмов распознавания, основанной на голосовании по предикатам, каждый из которых является композицией некоторого корректного набора эл.кл. и его корректирующей функции. Предложен метод построения предикатов специального вида. Исследованы свойства этих предикатов.

Получены теоретические оценки скорости сходимости бустинг-алгоритма формирования семейств голосующих предикатов. На каждой итерации ищется предикат, наилучшим образом компенсирующий ошибки ранее построенных предикатов. Качество добавляемого предиката оценивается функционалом «взвешенной» информативности. Поиск предиката с максимальной информативностью сведён к специальной задаче дискретной оптимизации, обобщающей ряд известных задач[14]. Решение поставленной задачи в общем случае представляет теоретический и практический интерес.

На значительном объеме тестовых данных, включающих разнотипные модельные и прикладные задачи, проведено сравнение новых и ранее построенных асимптотически оптимальных алгоритмов дуализации с другими известными алгоритмами. Подобное

[14] *Peleg D.* Approximation algorithms for the label-cover max and red-blue set cover problems // Journal of Discrete Algorithms. 2007. Т. 5, № 1. С. 55—64 ; *Miettinen P.* On the positive–negative partial set cover problem // Information Processing Letters. 2008. Т. 108, № 4. С. 219—221.

экспериментальное обоснование асимптотически оптимального подхода до сих пор не проводилось.

Рассмотрена задача поиска ветви дерева решений $T_A(L)$, началом которой является некоторая фиксированная внутренняя вершина, а концом — висячая вершина, соответствующая решению дуализации. Доказано, что эта задача NP-полна. Данный результат объясняет, почему не увенчались успехом предпринимаемые ранее попытки избавиться от лишних шагов в асимптотически оптимальных алгоритмах дуализации, основанных на обходе в глубину дерева решений $T_A(L)$.

Практическая значимость. Разработанные распознающие алгоритмы позволяют решать широкий класс прикладных задач, в которых объекты могут быть представлены целочисленными признаковыми описаниями. К таким задачам относятся компьютерный анализ речи, распознавание изображений, медицинская диагностика и пр. Как уже отмечалось, дуализация является одной из центральных дискретных перечислительных задач. К дуализации могут быть сведены многие задачи, возникающие при логическом анализе данных, к числу которых, помимо распознавания по прецедентам, относятся кластерный анализ, построение ассоциативных правил, составление расписаний и пр. Построенные в работе алгоритмы дуализации, согласно экспериментам, позволяют за приемлемое время решать достаточно большие прикладные задачи.

На защиту выносятся следующие результаты.

1. Создание общей схемы синтеза логических корректоров, которая может быть использована для описания классических логических алгоритмов распознавания и ранее построенных логических корректоров.

2. Построение практического логического корректора POLAR с поляризуемой корректирующей функцией.

3. Разработка методики повышения качества распознавания и скорости обучения логических корректоров, в основе которой лежат построение локальных базисов классов и формирование семейств голосующих предикатов по принципу бустинга.

4. Построение асимптотически оптимальных алгоритмов дуализации AO1M, AO1K, AO2M, AO2K, RUNC, RUNC-M, PUNC и экспериментальное исследование границ применимости этих алгоритмов в зависимости от типа и размера входа.

Достоверность полученных результатов подтверждается доказательствами сформулированных утверждений и теорем, а также результатами экспериментов, проведённых автором.

Апробация работы. Основные положения и результаты диссертации докладывались на конференциях «Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции (RCDL-2011)» (г. Воронеж, 2011 г.), «Математические методы распознавания образов (ММРО-15)» (г. Петрозаводск, 2011 г.), «Интеллектуализация обработки информации (ИОИ-9)» (Черногория, г. Будва,

2012 г.), «Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013)» (г. Самара, 2013 г.), «Математические методы распознавания образов (ММРО-16)» (г. Казань, 2013 г.), «Интеллектуализация обработки информации (ИОИ-10)» (Греция, о. Крит, 2014 г.), «Математические методы распознавания образов (ММРО-17)» (г. Светлогорск, 2015 г.) и на семинаре отдела Интеллектуальных систем ВЦ РАН им. А.А. Дородницына в июне 2015 г.

Публикации. По тематике исследований опубликовано 15 научных работ, в том числе 5 статей в журналах, рекомендованных ВАК.

Структура работы Работа состоит из введения, трёх глав, заключения и списка литературы из 83 наименований. Материал изложен на 100 страницах.

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во **введении** обосновывается актуальность исследований в области алгебро-логического синтеза алгоритмов распознавания и алгоритмов решения дискретных перечислительных задач. Формулируются цель, задачи работы, положения, выносимые на защиту, указываются научная новизна, теоретическая и практическая значимость результатов, приводятся сведения о структуре диссертации и апробации полученных результатов.

Глава 1. Корректное распознавание по прецедентам

В данной главе даётся обзор основных подходов к построению корректных логических алгоритмов распознавания, а именно, логического, оптимизационного, алгебраического и алгебро-логического. Обобщается понятие корректного набора эл.кл., являющееся базовым для алгебро-логического подхода. Описывается общая схема синтеза логических корректоров. Строится новый практический логический корректор POLAR с поляризуемой корректирующей функцией.

1.1. Основные подходы к решению задачи корректного распознавания по прецедентам

Рассматривается стандартная постановка задачи распознавания по прецедентам. Исследуется множество объектов M, которое может быть представлено в виде объединения непересекающихся подмножеств $K_1, ..., K_l$, называемых классами. Объекты из M описываются системой целочисленных признаков $(x_1, \ldots, x_n)$, то есть каждый объект S может быть представлен вектором $(x_1(S), \ldots, x_n(S))$, в котором j-я координата равна значению признака x_j для объекта S. Задано множество объектов $T = \{S_1, \ldots, S_m\}$ из M, и для каждого объекта $S_i \in T$ известен номер класса, которому он принадлежит. Объекты из T называются *прецедентами* или *обучающими объектами*. Требуется по обучающей выборке T построить *алгоритм распознавания*, то есть алгоритмически реализовать отображение $A_T : M \to \{0, 1, \ldots, l\}$, ставящее в соответствие каждому объекту из M номер класса или принимающее значение 0

в случае отказа от распознавания. Алгоритм распознавания называется *корректным*, если он не ошибается на обучающих объектах. Качество работы A_T на объектах, не входящих в T, характеризует обобщающую способность A_T. Представляет интерес синтез корректных алгоритмом распознавания с хорошей обобщающей способностью.

1.2. Общая схема построения логического корректора

Определяются понятия корректного и представительного предиката класса. С использованием этих понятий описывается общая схема алгебро-логического синтеза корректных процедур распознавания. В рамках предложенной схемы рассматриваются классические логические распознающие алгоритмы и ранее построенные логические корректоры.

1.2.1. Понятия корректного предиката как обобщение понятия корректного набора элементарных классификаторов

Пусть $H = (x_{j_1}, \ldots, x_{j_r})$ — набор различных признаков и $\sigma = (\sigma^1, \ldots, \sigma^r)$ — набор, в котором σ^q — допустимое значение признака x_{j_q}, $q \in \{1, \ldots, r\}$. Пара (H, σ) определяет эл.кл. ранга r. Эл.кл. (H, σ) выделяет объект S, если признаковое подописание $H(S) = (x_{j_1}(S), \ldots, x_{j_r}(S))$ совпадает с вектором σ.

Пусть имеется набор эл.кл. $U = \big((H_1, \sigma_1), \ldots, (H_d, \sigma_d)\big)$. Объекту S из M ставится в соответствие вектор $U(S)$ длины d, j-я координата которого равна 1, если эл.кл. (H_j, σ_j) выделяет объект S, иначе — 0. Вектор $U(S)$ называется *откликом* набора эл.кл. U на объекте S.

Набор эл.кл. $U = \big((H_1, \sigma_1), \ldots, (H_d, \sigma_d)\big)$ называется *корректным для класса* K, если существует булева функция $F(t_1, \ldots, t_d)$ такая, что $F(U(S_i)) \neq F(U(S_t))$ для любой пары прецедентов $S_i \in K$ и $S_t \notin K$. Корректный для K набор эл.кл. U называется *монотонным*, если существует монотонная булева функция F такая, что $F(U(S_i)) > F(U(S_t))$ для любой пары прецедентов $S_i \in K$ и $S_t \notin K$.

Предикат $B : M \to \{0, 1\}$ называется *корректным для* K, $K \in \{K_1, \ldots, K_l\}$, если множество прецедентов, на которых B равен 1, является подмножеством либо K, либо $\overline{K}$ (здесь и далее для $K \subseteq M$ через $\overline{K}$ обозначается $M \setminus K$). Корректный для класса K предикат B называется *представительным для* K, если существует хотя бы один прецедент $S_i \in K$ такой, что $B(S_i) = 1$.

Пусть $U = \big((H_1, \sigma_1), \ldots, (H_d, \sigma_d)\big)$ — набор эл.кл. и F — булева функция от d переменных. Через $F(U)$ обозначается предикат, значение которого на $S \in M$ равно $F(U(S))$. Набор эл.кл. U называется *корректным (представительным) для* K с *корректирующей* функцией F, если предикат $F(U)$ является корректным (представительным) для K.

1.2.2. Алгоритм голосования по корректным предикатам

На этапе обучения для каждого класса K строятся два семейства Z_K и $Z_{\overline{K}}$ предикатов вида $F(U)$, где U — набор эл.кл., F — булева функция. Предикаты из Z_K являются представительными для K. Предикаты из $Z_{\overline{K}}$ корректны для K, но не являются представительными для K. Предикату B приписывается вес $\alpha_B > 0$. *Распознавание* осуществляется взвешенным голосованием по корректным предикатам, построенным на этапе обучения. При распознавании объекта S для каждого класса K вычисляется оценка $\Gamma(S, K)$ принадлежности объекта S классу K,

$$\Gamma(S, K) = \sum_{B \in Z_K} \alpha_B B(S) - \sum_{B \in Z_{\overline{K}}} \alpha_B B(S). \qquad (1)$$

Объект S относится к тому классу K, для которого оценка $\Gamma(S, K)$ имеет наибольшее значение. Если таких классов несколько, то алгоритм отказывается от распознавания и возвращает 0. Корректность алгоритма распознавания обеспечивается за счёт корректности каждого предиката, участвующего в голосовании (утверждение 1.1 в тексте диссертации).

1.2.3. Классические логические алгоритмы распознавания и ранее построенные логические корректоры в рамках схемы голосования по корректным предикатам

Показано, что предложенная в 1.2.2 схема может быть использована для описания как классических логические алгоритмов распознавания, основанных на голосовании по корректным эл.кл., так и ранее построенных логических корректоров. Понятия, на которых базируются эти алгоритмы, определяются на языке предикатов.

Эл.кл. (H, σ) называется *корректным для класса* K, если множество прецедентов, выделяемых эл.кл. (H, σ) является подмножеством либо K, либо $\overline{K}$. Корректный эл.кл. (H, σ) называется представительным набором класса K, если он выделяет хотя бы один прецедент из K. Набор признаков H называется *тестом*, если для любого класса K и любого прецедента $S_i \in K$ эл.кл. $(H, H(S_i))$, является представительным набором класса K.

Утверждение 1.2. *Эл.кл. (H, σ) корректен (является представительным набором) для класса K тогда и только тогда, когда предикат $[H(S) = \sigma]$ является корректным (представительным) для K (здесь и далее через $[p]$ обозначается предикат, принимающий значение 1 в случае, когда выражение p истинно, и 0 — в противном случае). При этом корректирующей функцией для набора эл.кл. $U = \big((H, \sigma)\big)$ относительно класса K является функция $F(t_1) = t_1$.*

В утверждении 1.2, фактически, даются определения корректного эл.кл. и представительного набора на языке предикатов.

Утверждение 1.4. *Пусть H — тест. Тогда для любого класса K существует монотонный корректный для K набор эл.кл. U такой, что для любого прецедента $S_i \in K$*

выполняется тождество $[H(S_i) = H(S)] \equiv [U(S_i) \preccurlyeq U(S)]$, $\forall S \in M$ *(здесь и далее* $(\alpha_1, \ldots, \alpha_d) \preccurlyeq (\beta_1, \ldots, \beta_d)$ *означает, что* $\alpha_i \leqslant \beta_i, \forall i \in \{1, \ldots, d\}$ *).*

Утверждение 1.4 показывает, что алгоритм голосования по тестам является логическим корректором специального вида с монотонной корректирующей функцией.

Утверждение 1.5. *Пусть* U *— монотонный корректный для класса* K *набор эл.кл. Тогда для любого прецедента* $S_i \in K$ *существует представительный для* K *набор* (H, σ) *такой, что выполняется тождество* $[H(S) = \sigma] \equiv [U(S_i) \preccurlyeq U(S)]$, $\forall S \in M$.

Утверждение 1.5 показывает, что монотонный логический корректор является алгоритмом голосования по специального вида семействам представительных наборов. Поэтому для задач с небольшой значностью признаков не удаётся построить монотонный логический корректор, превосходящий по качеству классические алгоритмы голосования по представительным наборам.

1.3. Логический корректор POLAR с поляризуемой коректирующей функцией

Предлагается новый практический логический корректор POLAR, имеющий в качестве корректирующей поляризуемую булеву функцию. В данном корректоре в роли голосующих предикатов выступают так называемые поляризуемые предикаты. Построение этих предикатов сводится к поиску покрытий специальной булевой матрицы. Рассматривается задача поиска голосующих предикатов с наибольшей информативностью.

1.3.1. Поляризуемые предикаты

Пусть $R = (r_1, \ldots, r_d)$ — набор бинарных отношений на множестве $\{0, 1\}$ (например, $r_1(x, y) = [x \leqslant y]$, $r_2(x, y) = [x = y]$, $r_3(x, y) = [x \geqslant y]$ и т.д.), $U = \big((H_1, \sigma_1), \ldots, (H_d, \sigma_d)\big)$ — набор эл.кл. и G, $G \subseteq T$, — набор прецедентов. Для бинарных векторов $\alpha = (\alpha_1, \ldots, \alpha_d)$ и $\beta = (\beta_1, \ldots, \beta_d)$ вводится обозначение $R(\alpha, \beta) = r_1(\alpha_1, \beta_1) \wedge \ldots \wedge r_d(\alpha_d, \beta_d)$. Исследуются предикаты вида

$$B_{(U,R,G)}(S) = \bigvee_{S_i \in G} R(U(S_i), U(S)). \qquad (2)$$

Формулируются условия, при которых предикат $B_{(U,R,G)}$ является представительным для K и набор эл.кл. U имеет поляризуемую корректирующую функцию (утверждения 1.7–1.10 в тексте диссертации). Представительные для K предикаты вида (2), удовлетворяющие указанным условиям, называются *поляризуемыми*. Через $\mathcal{P}_K$ обозначается множество поляризуемых предикатов, представительных для класса K.

1.3.2. Алгоритм голосования по поляризуемым предикатам

Строится логический корректор POLAR, основанный на голосовании по поляризуемым предикатам. *На этапе обучения* для каждого класса K формируются два семейства Z_K и $Z_{\overline{K}}$ корректных для K поляризуемых предикатов, $Z_K \subset \mathcal{P}_K$, $Z_{\overline{K}} \subset \mathcal{P}_{\overline{K}}$.

Предикату B приписывается вес $\alpha_B > 0$. *Распознавание* осуществляется взвешенным голосованием по предикатам, построенным на этапе обучения. Возможны два режима распознавания: базовый и аддитивный.

1. *В базовом режиме* оценка $\Gamma(S, K)$ принадлежности объекта S классу K вычисляется по формуле (1).

2. *В аддитивном режиме* для распознаваемого объекта S и каждого построенного предиката $B_{(U,R,G)}$ вычисляется

$$\gamma(S, B_{(U,R,G)}) = \frac{1}{|G|} \sum_{S_i \in G} R(U(S_i), U(S)).$$

Затем для каждого класса K вычисляется оценка $\Gamma(S, K)$ принадлежности объекта S классу K, имеющая вид

$$\Gamma(S, K) = \sum_{B \in Z_K} \alpha_B \gamma(S, B) - \sum_{B \in Z_{\overline{K}}} \alpha_B \gamma(S, B).$$

1.3.3. Сведение задачи построения поляризуемых предикатов к поиску покрытий булевой матрицы

Основная задача этапа обучения логического корректора ΡOLAR — поиск информативных предикатов из $\mathcal{P}_K$. Эта задача сводится к поиску покрытий булевой матрицы. Покрытием булевой матрицы L называется набор её столбцов J такой, что в подматрице, составленной из столбцов набора J, в каждой строке есть хотя бы один единичный элемент. Через $L(R, C)$ обозначается подматрица матрицы L, составленная из её строк R и столбцов C.

Множество троек (H, σ, r), где (H, σ) — эл.кл. и r — бинарное отношение на $\{0, 1\}$, обозначается через $\mathcal{V}^*$. Строится булева матрица L_T по следующему правилу. Каждой строке матрицы L_T сопоставляется пара обучающих объектов $(S_i, S_t) \in T \times T$. Столбцы матрицы L_T имеют один из двух типов. Каждому столбцу первого типа соответствует тройка $(H, \sigma, r) \in \mathcal{V}^*$. Элемент матрицы L_T, расположенный на пересечении строки (S_i, S_t) и столбца (H, σ, r), равен $1 - r\left([H(S_i) = \sigma], [H(S_t) = \sigma]\right)$. Каждому столбцу второго типа соответствует прецедент $S_j \in T$. Элемент матрицы L_T, расположенный на пересечении строки (S_i, S_t) и столбца S_j, равен $[i = j]$. Матрицу, построенную по указанному правилу, принято называть *матрицей сравнения*. Через L_K, $K \in \{K_1, \ldots, K_l, \overline{K}_1, \ldots, \overline{K}_l\}$, обозначается подматрица $L_T\big((T \cap K) \times (T \setminus K), \mathcal{V}^* \cup (T \cap K)\big)$. Доказывается

Утверждение 1.11. *Пусть* $U{=}((H_1, \sigma_1), \ldots, (H_d, \sigma_d))$ *— набор эл.кл.,* $R{=}(r_1, \ldots, r_d)$ *— набор бинарных отношений из* $\{[x \leqslant y], [x \geqslant y]\}$ *и* G *— набор прецедентов класса* K.

Предикат $B_{(U,R,G)}$ *является корректным для* K *тогда и только тогда, когда набор столбцов* $J = \{(H_1, \sigma_1, r_1), \ldots, (H_d, \sigma_d, r_d)\} \cup ((T \cap K) \setminus G)$ *является покрытием матрицы* L_K.

1.3.4. Поиск поляризуемых предикатов с наибольшей информативностью

Вводятся функционалы информативности голосующих предикатов. Ставится задача поиска поляризуемых предикатов с наибольшей информативностью, которая сводится к решению специальных дискретных оптимизационных задач.

С каждым прецедентом S_i связывается неотрицательный вес w_i, характеризующий цену ошибки на объекте S_i. В базовом режиме работы логического корректора POLAR информативность предиката B относительно класса K оценивается значением функционала $I(B,K) = P(B,K) - N(B,K)$, где $P(B,K) = \sum_{S_i \in K} w_i B(S_i)$, $N(B,K) = \sum_{S_i \notin K} w_i B(S_i)$. В аддитивном режиме используется функционал $\hat{I}(B_{(U,R,G)},K) = \hat{P}(B_{(U,R,G)},K) - \hat{N}(B_{(U,R,G)},K)$, где $\hat{P}(B_{(U,R,G)},K) = \sum_{S \in G} P(B_{(U,R,\{S\})},K)$ и $\hat{N}(B_{(U,R,G)},K) = \sum_{S \in G} N(B_{(U,R,\{S\})},K)$.

Пусть G^+ — набор прецедентов класса K и G^- — набор прецедентов из $\overline{K}$. Обозначим через $\mathcal{P}_K(G^+,G^-)$ семейство поляризуемых предикатов таких, что $G \subseteq G^+$ и не существует двух объектов $S_i \in G$ и $S_t \in G^-$, для которых выполняется равенство $R(U(S_i),U(S_t)) = 1$.

Задача 1.3. *Пусть даны булевы матрицы $L_0, L_1, \ldots, L_d$ и ненулевые веса $\alpha_1, \ldots, \alpha_d$. Каждая матрица имеет n столбцов. Требуется найти (неприводимое) покрытие J матрицы L_0 такой, что сумма весов матриц, не покрытых набором J, максимальна.*

Задача 1.4. *Пусть даны две булевы матрица L_0 и L' с n столбцами. Для каждой строки i матрицы L' задан ненулевой вес β_i. Требуется найти (неприводимое) покрытие J матрицы L_0 такое, что сумма весов строк матрицы L', не покрытых набором J, максимальна.*

Поиск предиката B из $\mathcal{P}_K(G^+,G^-)$, обладающего максимальной информативностью $I(B,K)$, сводится к решению дискретной оптимизационной задачи 1.3. При использовании функционала $\hat{I}(B,K)$ решается задача 1.4.

Заметим, что ряде работ рассматриваются задачи, являющиеся частными случаями задачи 1.4, например, Red-Blue Set Cover Problem[15] и Positive–Negative Partial Set Cover Problem[16]. Исследование задач 1.3 и 1.4 в приведённых постановках автору не известны. В настоящей работе для их решения используется метод ветвей и границ на базе алгоритмов дуализации из третьей главы.

Глава 2. Методы повышения эффективности логических корректоров

Во данной главе разрабатывается методика повышения скорости обучения и качества распознавания логических корректоров. Семейства голосующих предикатов строятся итеративно по принципу *бустинга*. Поиск голосующих предикатов осуществляется

[15] *Peleg D.* Approximation algorithms for the label-cover max and red-blue set cover problems // Journal of Discrete Algorithms. 2007. Т. 5, № 1. С. 55—64 ; On the red-blue set cover problem / R. D. Carr [и др.] // in: Proc. 11th ACM-SIAM Symp. on Discrete Algorithms. 2000. С. 345—353.

[16] *Miettinen P.* On the positive-negative partial set cover problem // Information Processing Letters. 2008. Т. 108, № 4. С. 219—221.

в рамках *локальных базисов классов* — предварительно формируемых корректных наборов, состоящих из информативных эл.кл. Эффективность предложенной методики тестируется на реальных данных.

2.1. Итеративное формирование семейств голосующих предикатов по принципу бустинга

Строится бустинг-алгоритм обучения логического корректора POLAR, работающего в базовом режиме распознавания.

Пусть $S_i \in T$, y_i — номер класса, которому принадлежит S_i, K — класс, A_t — логический корректор, голосующий по предикатам, построенным за t, $t \geqslant 0$, итераций. Через $\Gamma_t(S_i, K)$ обозначается оценка за отнесение объекта S_i к классу K, вычисляемая логическим корректором A_t, через $Q(A_t)$ обозначается число ошибок и отказов алгоритма A_t на обучающей выборке. Рассматривается функционал

$$\hat{Q}(A_t) = \sum_{y=1}^{l} \sum_{S_i \notin K_y} \exp(-M_t(S_i, K_y)), \text{ где } M_t(S_i, K_y) = \Gamma_t(S_i, K_{y_i}) - \Gamma_t(S_i, K_y).$$

Известно, что для любого $t \geqslant 0$ выполняется неравенство $Q(A_t) \leqslant \hat{Q}(A_t)$. Предикат, добавляемый в семейство голосующих предикатов логического корректора A_t на итерации $t+1$, минимизирует значение $\hat{Q}(A_{t+1})$.

Бустинг-алгоритм обучения логического корректора POLAR.

Параметры: t_{max} — число итераций, $\delta > 0$ — параметр выбора пространства поиска предикатов.

При инициализации взять $Z_{K_1} = \ldots = Z_{K_l} = Z_{\overline{K}_1} = \ldots = Z_{\overline{K}_l} = \varnothing$.

Пусть произведено t, $t \geqslant 0$, итераций. На итерации $t+1$ выполняется следующее.

1. Для каждого класса K и каждого прецедента S_i вычислить вес

$$w_t(S_i, K) = \frac{1}{\hat{Q}(A_t)} \begin{cases} \sum_{K_y \neq K} \exp(-M_t(S_i, K_y)), & S_i \in K, \\ \exp(-M_t(S_i, K)), & S_i \notin K \end{cases}$$

(если прецедент S_i принадлежит классу K, то вес $w_t(S_i, K)$ характеризует «трудность отделения» объекта S_i от прецедентов из $\overline{K}$ логическим корректором A_t, иначе вес $w_t(S_i, K)$ указывает насколько «трудно» прецедент S_i отличить от прецедентов класса K).

2. Выбрать класс K и семейство поляризуемых предикатов $P_{\mathcal{K}}(G^+, G^-)$ такие, что в $P_{\mathcal{K}}(G^+, G^-)$ существует предикат B, для которого $J_t^*(B, K) > \delta$, где $J_t^*(B, K) = \sqrt{P_t(B, K)} - \sqrt{N_t^*(B, K)}$, $P_t(B, K) = \sum_{S_i \in K} w_t(S_i, K) B(S_i)$, $N_t(B, K) = \sum_{S_i \notin K} w_t(S_i, K) B(S_i)$ и

$$N_t^*(B, K) = \begin{cases} N_t(B, K), & N_t(B, K) > 0, \\ \frac{1}{2m}, & \text{иначе.} \end{cases}$$

3. Найти в $\mathcal{P}_K(G^+, G^-)$ предикат B с наибольшей информативностью $I_t(B,K)$=$P_t(B,K) - N_t(B,K)$.

4. Добавить предикат B в семейство Z_K с весом

$$\alpha_B = \frac{1}{2}\ln\frac{P_t(B,K)}{N_t^*(B,K)}.$$

5. Если $t+1 \neq t_{max}$, то перейти к следующей итерации. ■

Теорема 2.3. *Пусть бустинг-алгоритм обучения логического корректора POLAR запускается с параметрами*

$$t_{max} > \frac{\ln(m(l-1))}{\delta^2} \text{ и } \delta < \frac{1}{\sqrt{l}} - \frac{1}{\sqrt{2m}}.$$

Тогда в результате его работы строит корректный распознающий алгоритм.

Использование описанного бустинг-алгоритма позволяет повысить качество распознавания за счёт 1) настройки весов предикатов, по которым осуществляется голосование, и 2) построения семейств, состоящих из существенно различающихся предикатов. Кроме этого, повышает скорость обучения, так как для поиска голосующих предикатов используется не вся матрица сравнения, а лишь её небольшая подматрица.

2.2. Локальные базисы классов

Поиск голосующих предикатов осуществляется в рамках *локальных базисов классов* — предварительно формируемых корректных наборов, состоящих из информативных эл.кл. Локальный базис определяет состав столбцов матрицы сравнения, строящейся для поиска поляризуемых предикатов. Предлагается алгоритм формирования «хороших» локальных базисов из эл.кл. произвольного ранга.

Набор $\mathcal{V}_K = \{(H_1,\sigma_1,r_1),\ldots,(H_d,\sigma_d,r_d)\}$ троек из $\mathcal{V}^*$ называется *локальным базисом класса* K, если не существует двух прецедентов $S_i \in K$ и $S_t \notin K$, для которых выполняется равенство $R(U(S_i),U(S_t)) = 1$, где $U = ((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ и $R = (r_1,\ldots,r_d)$. Ясно, что $\mathcal{V}_K$ является локальным базисом класса K тогда и только тогда, когда подматрица, составленная из столбцов $\mathcal{V}_K$ матрицы L_K, не имеет нулевых строк, то есть для этой подматрицы существует покрытие.

Набор $\mathcal{V} \subseteq \mathcal{V}^*$, являющийся локальным базисом для каждого из классов $K_1,\ldots,K_l$, называется *локальным базисом задачи*. Например, набор $\mathcal{V}_1$, состоящий из троек $(H,\sigma,r) \in \mathcal{V}^*$ таких, что (H,σ) имеет ранг 1 и отношение r принадлежит $\{[x \leqslant y],[x \geqslant y]\}$, является локальным базисом задачи.

Предлагается универсальный метод построения локального базиса класса, состоящего из эл.кл. произвольного ранга. Рассматривается задача распознавания с двумя классами K и $\overline{K}$. Строится семейство эл.кл. C_K и каждому эл.кл. $(H,\sigma) \in C_K$ присваивается ненулевой вес $\alpha_{(H,\sigma)}$. В результате получается распознающий алгоритм

$$A_T^K(S) = \operatorname{sign} \sum_{(H,\sigma)\in C_K} \alpha_{(H,\sigma)}[H(S) = \sigma],$$

где $\mathrm{sign}(x)$ — функция «знак», возвращающая 1, при $x > 0$, -1, при $x < 0$, и 0, при $x = 0$. Алгоритм A_T^K считается корректным в случае, когда $A_T^K(S_i) = 1$, $\forall S_i \in K$, и $A_T^K(S_i) = -1$, $\forall S_i \notin K$. По взвешенному семейству C_K строится набор $\mathcal{V}_K$ такой, что каждому эл.кл. (H, σ) из C_K однозначно соответствует тройка $(H, \sigma, r) \in \mathcal{V}_K$, в которой $r = [x \leqslant y]$, при $\alpha_{(H,\sigma)} > 0$, и $r = [x \geqslant y]$, при $\alpha_{(H,\sigma)} < 0$. Метод обосновывается справедливостью следующего утверждения.

Утверждение 2.4. *Если распознающий алгоритм A_T^K корректен, то набор $\mathcal{V}_K$, построенный по взвешенному семейству эл.кл. C_K является локальным базисом класса K. Причём упорядоченный набор, составленный из эл.кл. семейства C_K, является корректным для K и имеет поляризуемую корректирующую функцию.*

2.3. Реализация и экспериментальное исследование логических корректоров POLAR

Реализованы 4 модификации логического корректора POLAR, формирующие семейства голосующих предикатов бустингом и отличающиеся стратегией формирования локальных базисов классов:

- POLAR-1 использует локальный базис задачи $\mathcal{V}_1$;
- POLAR-2 строит локальный базис задачи бустингом над эл.кл.;
- POLAR-3 на каждой итерации строит или обновляет локальный базис класса алгоритмом голосования по представительным наборам;
- POLAR-4 на каждой итерации строит или обновляет локальный базис класса бустингом над эл.кл.

Логические корректоры POLAR-1 – POLAR-4 тестируются на прикладных задачах из репозитория UCI. Результаты тестирования говорят о практической применимости новых логических корректоров, которые опережают по качеству распознавания ранее построенные логические корректоры и классические логические алгоритмы распознавания почти на всех тестовых задачах. При этом за приемлемое время осуществляется обучение на больших объёмах данных с большой значностью признаков. Наиболее быстрыми являются POLAR-3 и POLAR-4.

Глава 3. Новые асимптотически оптимальные алгоритмы дуализации

В данной главе рассматривается одна из центральных дискретных перечислительных задач — дуализация. Даётся обзор основных подходов её решения, среди которых выделяется подход к построению асимптотически оптимальных алгоритмов. Алгоритмы, построенные в рамках этого подхода классифицируются на два типа. Строятся новые асимптотически оптимальные алгоритмы первого типа AO1K, AO1M, AO2K и AO2M, и второго типа RUNC, RUNC-M и PUNC. Новые и ранее построенные асимптотически оптимальные алгоритмы дуализации экспериментально исследуются на большом объеме разнотипных данных.

3.1. Задача дуализации и подходы к ее решению

Пусть $L = \|a_{ij}\|$ — булева матрица размера $m \times n$. Набор столбцов H матрицы L называется покрытием, если подматрица L^H матрицы L, образованная столбцами набора H, не содержит строки вида $(0, 0, \ldots, 0)$. Покрытие матрицы L называется неприводимым, если любое его собственное подмножество не является покрытием L. Через $\mathcal{P}(L)$ обозначается множество неприводимых покрытий L. Требуется построить (перечислить) множество $\mathcal{P}(L)$.

В основе асимптотически оптимальных алгоритмов дуализации лежит

Критерий USM unit submatrix. *Набор H из r столбцов матрицы L является неприводимым покрытием тогда и только тогда, когда выполняются следующие два условия: 1) подматрица L^H матрицы L, образованная столбцами набора H, не содержит строки вида $(0, 0, \ldots, 0)$; 2) подматрица L^H содержит каждую из строк вида $(1, 0, 0, \ldots, 0, 0), (0, 1, 0, \ldots, 0, 0), \ldots, (0, 0, 0, \ldots, 0, 1)$, то есть с точностью до перестановки строк содержит единичную подматрицу Q порядка r.*

При выполнении условия 1) единичная подматрица Q является максимальной, в том смысле, что она не содержится в других единичных подматрицах. Набор столбцов, удовлетворяющий условию 2), называется *совместимым*. Максимальная единичная подматрица порождает максимальный совместимый набор столбцов, то есть такой совместимый набор, который не содержится ни в каком другом совместимом наборе.

Асимптотически оптимальные алгоритмы дуализации классифицируются на два типа. Алгоритм первого типа перечисляет подмножество множества максимальных единичных подматриц матрицы L и может совершать лишние шаги, связанные с повторным построением решений. Примерами алгоритмов первого типа служат алгоритмы АО1 и АО2[17]. Алгоритм второго типа перечисляет без повторений подмножество множества максимальных совместимых наборов столбцов. Примерами алгоритмов второго типа является алгоритм ОПТ[18], и алгоритмы MMCS, RS[19].

Алгоритмы дуализации MMCS и RS появились в сравнительно недавних публикациях K. Murakami и T. Uno. Авторы на значительном объёме разнотипных данных протестировали разработанные ими алгоритмы, а также алгоритмы, имеющие иную конструкцию. Алгоритмы MMCS и RS показали наилучшие результаты.

В указанных публикациях утверждается, что в основе алгоритмов MMCS и RS лежит изобретённый авторами новый принцип построения алгоритмов дуализации, называемый условием «crit». Однако это условие, сформулированное с помощью понятий теории гиперграфов, эквивалентно критерию USM. Таким образом, принципиальная схема работы алгоритмов, использующих условие «crit», не отличается от схемы рабо-

[17] *Дюкова Е. В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. 1997. Т. 233, № 4. С. 527—530 ; *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // Журнал вычислительной математики и математической физики. 2004. Т. 44, № 3. С. 562—572.

[18] *Дюкова Е. В., Инякин А. С.* Асимптотически оптимальное построение тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. 2008. Т. 17. С. 235—246.

[19] *Murakami K., Uno T.* Efficient algorithms for dualizing large-scale hypergraphs // Discrete Applied Mathematics. 2014. Т. 170. С. 83—94.

ты отечественных асимптотически оптимальных алгоритмов без повторяющихся шагов, построенных значительно раньше в работах Е. В. Дюковой и её учеников.

Проведённое нами дополнительное тестирование показало, что алгоритмы MMCS и RS опережают алгоритмы ОПТ и АО2, поскольку почти всегда строят деревья решений из меньшего числа вершин, чем отечественные алгоритмы. Следовательно, одним из способов сокращения времени работы асимптотически оптимального алгоритма дуализации является «оптимизация структуры» строящегося им дерева решений.

3.2. Асимптотически оптимальные алгоритмы дуализации первого типа

Приводится схема работы асимптотически оптимального алгоритма первого типа. В рамках этой схемы описываются алгоритмы АО1, АО2 и строятся их модификации АО1К, АО1М, АО2К, АО2М, в которых сокращаются вычислительные затраты за счёт уменьшения общего числа вершин дерева решений.

Говорят, что столбец j *покрывает* строку i матрицы L, если $a_{ij} = 1$. Пусть H — набор столбцов матрицы L. Говорить, что набор H покрывает строку i, если существует столбец $j \in H$, покрывающий строку i.

Через $E(L)$ обозначается $\{(i,j) : a_{ij} = 1, i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}\}$. Два элемента (i,j) и (t,l) из $E(L)$ называются *совместимыми*, если $a_{il} = 0,\ a_{tj} = 0$. Набор Q элементов из $E(L)$ называется *совместимым*, если любые два различных элемента (i,j) и (t,l) из Q совместимы. Совместимый набор Q называется *максимальным*, если Q не является подмножеством другого совместимого набора элементов из $E(L)$.

Пусть Q — совместимый набор элементов из $E(L)$. Столбец l называется *запрещённым* для Q, если существует элемент $(i,j) \in Q$ такой, что столбец l покрывает строку i. В противном случае говорят, что столбец l *совместим* с набором Q.

Набор B, $B \subseteq E(L)$, *порождает* набор столбцов $H(B) = \{j\colon \exists (i,j) \in B\}$. Говорят, что строка i матрицы L *покрыта* набором B элементов из $E(L)$, если набор столбцов $H(B)$ покрывает строку i. Совместимый набор Q называется *покрывающим*, если все строки матрицы L покрыты набором Q. Набор столбцов H является неприводимым покрытием матрицы L тогда и только тогда, когда найдётся покрывающий набор Q, для которого $H(Q) = H$.

Покрывающий набор $Q = \{(i_1,j_1),\ldots,(i_r,j_r)\}$ называется *верхним*, если для любого покрывающего набора $Q' = \{(t_1,j_1),\ldots,(t_r,j_r)\}$ верны неравенства $t_u \geqslant i_u$, $u \in \{1,\ldots,r\}$. Для любого неприводимого покрытия H существует единственный верхний набор Q, такой, что $H(Q) = H$. Таким образом, задача построения $\mathcal{P}(L)$ сводится к перечислению верхних наборов элементов из $E(L)$.

Задача 3.1. Вход: L — *булева матрица*, Q — *совместимый набор элементов из* $E(L)$. Выход: 1, *если существует верхний набор* $Q'\colon Q \subseteq Q'$, *и* 0 — *иначе*.

Теорема 3.3. *Задача 3.1 является* NP*-полной.*

Теорема 3.3 показывается, что при $\mathsf{P} \neq \mathsf{NP}$ не существует асимптотически оптимального алгоритма первого типа, не делающего лишних шагов и перечисляющего неприводимые покрытия без повторений с полиномиальной задержкой.

3.3. Асимптотически оптимальные алгоритмы дуализации второго типа

Даётся схема асимптотически оптимального алгоритма дуализации второго типа. В рамках схемы описываются ранее построенные алгоритмы ОПТ, RS и MMCS, а также новые алгоритмы RUNC, RUNC-M и PUNC. Для этого вводятся дополнительные понятия и обозначения.

Пусть H — набор столбцов матрицы L. Строка i матрицы L называется *опорной* для пары (H, j), $j \in H$, если $a_{ij} = 1$ и $a_{il} = 0$, $l \neq j$, $l \in H$. Множество опорных строк для (H, j) обозначается через $S(H, j)$. Набор H является совместимым тогда и только тогда, когда для каждого (H, j), $j \in H$, множество $S(H, j)$ не пусто.

Столбец j матрицы L называется *запрещённым* для набора столбцов H, если существует столбец $l \in H$ такой, что столбец j покрывает все опорные для (H, l) строки. В противном случае говорят, что столбец j *совместим* с набором H.

Пусть $L(R, C)$ — подматрица матрицы L. Число $v_j(R) = \sum_{i \in R} a_{ij}$, $j \in C$, называется *весом* столбца j в подматрице $L(R, C)$. Число $w_i(C) = \sum_{j \in C} a_{ij}$, $i \in R$, называется *весом* строки i в подматрице $L(R, C)$. При $w_i(C) = 0$ ($v_j(R) = 0$) строка $i \in R$ (столбец $j \in C$) называется *нулевой* (*нулевым*) в $L(R, C)$.

Для обозначения связи некоторого объекта X с текущей вершиной H дерева решений используется запись $X[H]$.

Алгоритмы RUNC и RUNC-M. На шаге 1 на итерации 1 выбирается строка i матрицы L (в алгоритме RUNC $i = 1$; алгоритм RUNC-M ищет строку i с минимальным весом в матрице L), строится набор столбцов $C[\varnothing]$, покрывающих строку i, и строится подматрица $L(R[\varnothing], D[\varnothing])$ путем последовательного удаления из матрицы L охватывающих строк и нулевых столбцов. Далее корень становится текущей вершиной, и происходит переход к следующей итерации.

Пусть на шаге $s, s \geqslant 1$, на итерации $t, t \geqslant 1$, текущей стала вершина H. Тогда на итерации $t + 1$ выполняется следующее.

1. Если $C[H] = \varnothing$, то происходит переход к следующему шагу. В противном случае берётся первый по порядку столбец $j \in C[H]$, столбец j удаляется из $C[H]$ и из $D[H]$. Строится вершина $H' = H \cup \{j\}$.

2. Если столбец j покрывает все строки, не покрытые набором H, то результатом шага становится неприводимое покрытие H', и происходит переход к следующему шагу.

3. В противном случае в подматрице $L(R[H], D[H])$ выбирается строка i, не покрытая столбцом j (алгоритм RUNC использует строку с наименьшим номером, алгоритм RUNC-M ищет строку i с наименьшим весом $w_i(D[H])$).

4. Формируется набор $C[H']$ покрывающих строку i столбцов подматрицы $L(R[H], D[H])$, и строится подматрица $L(R[H'], D[H'])$ путем удаления из подматрицы $L(R[H], D[H])$ покрытых столбцом j строк и запрещённых для H' столбцов.

5. Текущей вершиной становится H', и происходит переход к следующей итерации.

Пусть результатом шага $s, s \geqslant 1$, является набор H. Тогда на шаге $s+1$ на итерации 1 среди вершин ветки дерева, соединяющей корень с вершиной H, ищется ближайшая к H вершина H' такая, что $C[H'] \neq \varnothing$. Если вершина H' найдена, то она становится текущей вершиной, и происходит переход к следующей итерации. В противном случае алгоритм завершает работу. ■

Сложность шага алгоритма RUNC (RUNC-M) равна $\mathcal{O}(mnq)$, $q = \min\{m, n\}$. Для работы алгоритма дополнительно требуется $\mathcal{O}(m+n)$ памяти. Фактически, алгоритмы RUNC и RUNC-M различаются способом выбора строки i, определяющей состав дочерних вершин в дереве решений для новой построенной вершины H. Экспериментально установлено, что в большинстве случаев RUNC-M является эффективнее RUNC, поскольку его дерево решений имеет существенно меньше вершин, что компенсирует вычислительные затраты поиска строки с минимальным весом. Алгоритм RUNC-M можно назвать «жадным» алгоритмом, пытающимся минимизировать число внутренних вершин дерева решений.

3.4. Экспериментальное исследование асимптотически оптимальных алгоритмов дуализации

Для тестирования эффективности предложенных в работе алгоритмов проведена серия экспериментов на случайных булевых матрицах, формируемых по ранее предложенной методике[20], а также на других модельных данных и прикладных задачах[21]. Алгоритмы AO1, AO1K, AO1M, AO2, AO2K, AO2M, ОПТ, RUNC, RUNC-M и PUNC реализованы на языке C++ и доступны в Интернет по адресу `http://sourceforge.net/p/logicalanalyze/code/HEAD/tree/trunk/dualization/`. Исходные коды программ алгоритмов MMCS и RS взяты из `http://research.nii.ac.jp/~uno/dualization.html`.

Результаты счёта показывают, что среди задач есть такие, на которых лидируют ранее построенные алгоритмы RS, MMCS, ОПТ, AO1 а также задачи, где лидируют новые алгоритмы AO1M, RUNC, RUNC-M и PUNC. Тестирование на прикладных задачах показывает, что лучшим является алгоритм RUNC-M, преимущество которого особенно очевидно на входных матрицах большого размера.

[20] *Дюкова Е. В., Инякин А. С.* Указ. соч.
[21] *Murakami K., Uno T.* Указ. соч.

ЗАКЛЮЧЕНИЕ

1. Предложена общая схема синтеза логических корректоров, голосующих по корректным предикатам. Показано, что схема логического корректора общего вида может быть использована для описания классических логических алгоритмов распознавания и ранее построенных логических корректоров.

2. Построен практический логический корректор POLAR с поляризуемой корректирующей функцией.

3. Разработана методика повышения качества распознавания и скорости обучения логических корректоров, основанная на построении локальных базисов классов и итеративном формировании семейств голосующих предикатов по принципу бустинга.

4. Построены новые асимптотически оптимальные алгоритмы дуализации AO1M, AO1K, AO2M, AO2K, RUNC, RUNC-M, PUNC, в которых снижение времени счёта достигается за счёт уменьшения общего числа вершин дерева решений. Показано, что построенные алгоритмы достаточно быстро обрабатывают булевы матрицы большого размера.

Одним из дальнейших направлений исследований видится обобщение методов алгебро-логической коррекции на случай, когда в задаче распознавания на множествах значений признаков определены отношения частичного порядка. Практический интерес представляют частичные порядки, являющиеся цепями, антицепями, полурешётками, решётками или лесами. При выполнении коррекции потребуются эффективные перечислительные алгоритмы, для решения задач, обобщающих дуализацию.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в журналах, рекомендованных ВАК

1. *Djukova E. V., Lyubimtseva M. M., Prokofjev P. A.* Logical correctors in recognition // Pattern Recognition and Image Analysis. — 2014. — Т. 24, № 3. — С. 358—364.
2. *Djukova E. V., Prokofjev P. A.* Asymptotically optimal dualization algorithms // Computational Mathematics and Mathematical Physics. — 2015. — Т. 55, № 5. — С. 891—905.
3. *Djukova E. V., Prokofjev P. A.* Models of Recognition Procedures with Logical Correctors // Pattern Recognition and Image Analysis. — 2013. — Т. 23, № 2. — С. 235—244.
4. *Дюкова Е. В., Прокофьев П. А.* Об асимптотически оптимальном перечислении неприводимых покрытий булевой матрицы // Прикладная дискретная математика. — 2014. — № 1. — С. 96—105.
5. *Прокофьев П. А.* Классификация фрагментов текстов с описанием зависимостей правилами на интерпретируемом экспертами языке // Вестник ВГУ, серия: Системный анализ и информационные технологии. — 2012. — № 1. — С. 174—178.

Статьи в прочих журналах

6. *Дюкова Е. В., Любимцева М. М., Прокофьев П. А.* Об алгебро-логической коррекции в задачах распознавания по прецедентам // Машинное обучение и анализ данных. — 2013. — Т. 1, № 6. — С. 705—713.
7. *Дюкова Е. В., Прокофьев П. А.* Построение и исследование новых асимптотически оптимальных алгоритмов дуализации // Машинное обучение и анализ данных. — 2014. — Т. 1, № 8. — С. 1048—1067.
8. *Дюкова Е. В., Журавлёв Ю. И., Прокофьев П. А.* Методы повышения эффективности логических корректоров // Машинное обучение и анализ данных. — 2015. — Т. 1, № 11. — С. 1555—1583.

Тезисы докладов на конференциях

9. *Djukova E. V., Lyubimtseva M. M., Prokofjev P. A.* Logical correctors in recognition problems // 11th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013). Samara, September 23-28, 2013. Т. 1. — Samara: IPSI RAS, 2013. — С. 82—83.
10. *Васильев В. Г., Прокофьев П. А.* Извлечение информации из текста с автоматическим построением правил // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. XIII Всероссийская научная конференция RCDL'2011. Воронеж, 19–22 октября 2011 г.: труды конференции. — Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011. — С. 358—364.
11. *Прокофьев П. А.* Дискретный подход при извлечении информации из текста с автоматическим построением правил (текстовых запросов) // Математические методы распознавания образов: XV Всероссийская конференция, г. Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М.: МАКС Пресс, 2011. — С. 585—588.
12. *Дюкова Е. В., Прокофьев П. А.* Методы обучения логических процедур распознавания, основанных на семействах корректных наборов элементарных классификаторов // Интеллектуализация обработки информации: IX Международная конференция. Черногория, г. Будва, 2012: Сборник докладов. — М.: Торус Пресс, 2012. — С. 67—70.
13. *Дюкова Е. В., Любимцева М. М., Прокофьев П. А.* Логические корректора в задачах классификации по прецедентам // Математические методы распознавания образов: XVI Всероссийская конференция, г. Казань, 6–12 сентября 2013 г.: Тезисы докладов. — М.: Торус Пресс, 2013. — С. 7.
14. *Дюкова Е. В., Журавлёв Ю. И., Прокофьев П. А.* Вопросы эффективности логических корректоров // Математические методы распознавания образов: Тезисы докладов XVII Всероссийской конференции с международным участием, г. Светлогорск, 2015 г. — М.: Торус Пресс, 2015. — С. 70—71.
15. *Дюкова Е. В., Прокофьев П. А.* Новые асимптотически оптимальные алгоритмы дуализации // Интеллектуализация обработки информации: X Международная конференция. Греция, о. Крит, 4–11 октября, 2012: Тезисы докладов. — М.: Торус Пресс, 2014. — С. 50—51.