На правах рукописи

Масляков Глеб Олегович

Корректная классификация над произведением частичных порядков

Специальность 1.2.3 — Теоретическая информатика, кибернетика

Автореферат

диссертации на соискание учёной степени

кандидата физико-математических наук

Москва — 2023

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

Научный руководитель:	**Дюкова Елена Всеволодовна** д.ф.-м.н., доцент, ФГУ ФИЦ ИУ РАН, главный научный сотрудник
Официальные оппоненты:	**Кузнецов Сергей Олегович**, д.ф.-м.н., профессор, ФГАОУ ВО НИУ ВШЭ, заведующий международной лабораторией интеллектуальных систем и структурного анализа
	Мокряков Алексей Викторович к.ф.-м.н., доцент, ФГБОУ ВО РГУ им. А.Н. Косыгина, заведующий кафедрой «Прикладная математика и программирование»
Ведущая организация:	Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет»

Защита состоится ___ _____ _____ года в _____ на заседании диссертационного совета 24.1.224.03 при Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

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

Автореферат разослан «___» _____ 2023 года.

Ученый секретарь
диссертационного совета 24.1.224.03
к.т.н. Рейер И.А.

Общая характеристика работы

Актуальность темы. В диссертации рассматривается одна из центральных задач машинного обучения — задача классификации на основе прецедентов.

Под прецедентной (обучающей) информацией понимается совокупность примеров изучаемых объектов, в которой каждый объект представлен в виде числового вектора, полученного на основе измерения или наблюдения ряда его параметров или характеристик, называемых признаками. Каждый пример (обучающий объект или прецедент) приписан к определённому классу объектов. Требуется на основе анализа обучающей информации построить алгоритм, позволяющий классифицировать новые, не входящие в обучающую выборку, объекты.

Главное достоинство логического подхода к задаче классификации на основе прецедентов — возможность получения результата при отсутствии дополнительных предположений вероятностного характера и при небольшом числе прецедентов. Считается, что каждый признак принимает ограниченное число допустимых значений, которые кодируются целыми числами. Для каждого признака задаётся бинарная функция близости между его значениями, что позволяет проводить сравнение описания распознаваемого объекта с описаниями прецедентов. Анализ прецедентной информации сводится к поиску в исходных данных специальных фрагментов описаний объектов, различающих объекты из разных классов. Найденные фрагменты имеют содержательное описание в терминах той прикладной области, в которой решается задача. По их наличию или, наоборот, отсутствию в описании распознаваемого объекта решается вопрос о его классификации. Большое внимание уделяется вопросам синтеза алгоритмов, безошибочно классифицирующих материал обучения. Такие алгоритмы называются корректными.

К наиболее известным направлениям логической классификации относятся Correct Voting Procedures или CVP (процедуры корректного голосования), впервые предложенные в отечественных работах, а также Logical Analysis of Data или LAD (логический анализ данных) и Formal Concept Analysis или FCA (анализ формальных понятий). Все три названных направления CVP, LAD и FCA имеют много общего. С другой стороны, каждый из подходов использует свою терминологию и демонстрирует некоторую оригинальность.

Фундаментальную роль в создании методов CVP сыграли работы С. В. Яблонского, в которых введено хорошо известное в дискретной математике понятие теста, и работы Ю. И. Журавлёва, опубликованные в 70-х

и 80-х годах прошлого века. Основы проблематики были заложены также в статьях российских учёных М. М. Бонгарда (1967 г.) и М. Н. Вайнцвайга (1973 г.). В дальнейшем направление CVP развивалось в работах отечественных и зарубежных авторов и наиболее существенное развитие получило в публикациях представителей школы Ю. И. Журавлёва.

Основополагающие идеи LAD и FCA принадлежат соответственно П. Хаммеру (1986 г.) и Р. Вилле (1982 г.).

В России методы LAD предложены практически параллельно с зарубежными авторами и, в основном, получили развитие в работах Ю. И. Журавлёва, В. В. Рязанова, О. В. Сенько и И. С. Масича.

В 1981 г. В. К. Финн предложил так называемый метод автоматического порождения гипотез (или ДСМ-метод), который позднее в 1990-х годах был адаптирован В. К. Финном и его учениками для задач машинного обучения. Позднее С. О. Кузнецов описал ДСМ-метод в терминах FCA. В России методы FCA представлены работами В. К. Финна, С. О. Кузнецова, М. И. Забежайло, Д. И. Игнатова и Д. В. Виноградова.

Пусть исследуемое множество объектов M представимо в виде объединения попарно не пересекающихся подмножеств (классов) $K_1, \ldots, K_l$, и пусть объекты из множества M описываются целочисленными признаками $x_1, \ldots, x_n$. Описание объекта S из M имеет вид $(a_1, \ldots, a_n)$, здесь a_j — значение признака x_j для объекта S.

Классические процедуры CVP базируются на поиске фрагментов описаний прецедентов, называемых корректными элементарными классификаторами. Элементарным классификатором **ЭК** называется пара (σ, H), где $H = \{x_{j_1}, \ldots, x_{j_r}\}$ — набор различных признаков, $\sigma = (\sigma_1, \ldots, \sigma_r)$ — набор, в котором σ_i значение признака x_{j_i}, $i = 1, 2, \ldots, r$. Близость между объектом $S = (a_1, \ldots, a_n)$ и **ЭК** (σ, H) оценивается функцией $B(S, \sigma, H)$, которая принимает значение 1, если $a_{j_i} = \sigma_i$ для всех $i = 1, 2, \ldots, r$, и 0 в противном случае. Если $B(S, \sigma, H) = 1$, то говорят, что объект S содержит **ЭК** (σ, H). **ЭК** (σ, H) корректен для класса K, $K \in \{K_1, \ldots, K_l\}$, если нельзя указать пару прецедентов, одновременно содержащих этот **ЭК**, причём один из них принадлежит классу K, а другой не принадлежит классу K.

В общем случае корректный **ЭК** (σ, H) по отношению к классу K может обладать одним из следующих двух свойств:

1) некоторые обучающие объекты из класса K содержат (σ, H);

2) ни один обучающий объект из класса K не содержит (σ, H).

Корректный **ЭК** первого типа называется *представительным* **ЭК** класса K. Корректный **ЭК** второго типа называется *покрытием* класса K.

На этапе обучения для каждого класса K классификатор A строит некоторое множество $C^A(K)$ корректных **ЭК** класса K. При распознавании произвольного объекта S из M найденные **ЭК** участвуют в процедуре голосования с целью вычисления общей оценки принадлежности объекта S этому классу.

В модели голосования по представительным **ЭК** множество $C^A(K)$ состоит из представительных **ЭК** класса K (необязательно всех). Оценка $\Gamma_1(S, K)$ принадлежности объекта S к классу K вычисляется по формуле

$$\Gamma_1(S, K) = \frac{1}{W} \sum_{(\sigma,H)\in C^A(K)} P_{(\sigma,H)} B(S, \sigma, H),$$

здесь $W = \sum_{(\sigma,H)\in C^A(K)} P_{(\sigma,H)}$, в качестве $P(\sigma, H)$ обычно берется число обучающих объектов из K, содержащих (σ, H). Объекту S присваивается класс с наивысшей оценкой. Если таких оценок несколько, то происходит отказ от классификации.

Аналогично устроена модель голосования по покрытиям класса. Однако в данной модели оценка $\Gamma_2(S, K)$ принадлежности объекта S к классу K вычисляется по формуле

$$\Gamma_2(S, K) = \frac{1}{W} \sum_{(\sigma,H)\in C^A(K)} P_{(\sigma,H)} \left(1 - B(S, \sigma, H)\right).$$

На практике хорошие результаты показывает алгоритм голосования по тупиковым корректным **ЭК** классов. Корректный для класса K **ЭК** (σ, H) называется *тупиковым*, если не является корректным любой **ЭК** (σ', H') такой, что $\sigma' \subset \sigma$, $H' \subset H$.

Направление LAD ориентировано на поиск так называемых *логических закономерностей* или *patterns*. Описание классификаторов обычно даётся в терминах логических функций. Логическая закономерность класса K — это представительный **ЭК** класса K. Наиболее информативными считаются логические закономерности, оптимальные с точки зрения некоторого заранее выбранного функционала. Например, ищутся логические закономерности, содержащиеся в наибольшем числе прецедентов (наибольшие логические закономерности).

В направлении FCA для задачи классификации ключевым термином является *положительная ДСМ-гипотеза*. Каждая положительная ДСМ-гипотеза класса K порождает в исходной обучающей выборке представительный **ЭК** (σ, H) класса K, обладающего следующим свойством: любой пред-

ставительный **ЭК** (σ', H') класса K такой, что $\sigma \subset \sigma'$, $H \subset H'$, содержится в меньшем числе прецедентов.

Таким образом, алгоритмы CVP, LAD и FCA ориентированы на поиск представительных **ЭК**, но каждое направление по разному определяет информативность **ЭК**. Это обуславливает и различие в методологии поиска требуемых представительных **ЭК**.

Направление CVP опирается на теорию труднорешаемых перечислительных задач. Алгоритмы LAD в значительной степени используют методы теории целочисленного программирования и при этом, как правило, строят небольшое количество представительных **ЭК**, что позволяет лучше интерпретировать полученные результаты. Схема работы на этапе распознавания алгоритма в LAD полностью аналогична схеме работы алгоритма голосования по представительным **ЭК**.

ДСМ-классификатор (FCA) действует более строго по сравнению с классификаторами из CVP и LAD. На первом этапе для каждого класса K строится некоторое множество представительных **ЭК** класса K, порождаемых положительными для класса K ДСМ-гипотезами. Объект S относится к классу K, если S содержит хотя бы один найденный **ЭК**, и не содержит ни одного **ЭК** из порождаемых положительными ДСМ-гипотезами для K', $K' \neq K$. В противном случае происходит отказ от классификации.

Настоящая работа посвящена развитию методов направления CVP.

При поиске тупиковых корректных **ЭК** возникает необходимость рассматривать сложные в вычислительном плане задачи, которые в теории алгоритмической сложности дискретных задач называют труднорешаемыми. Среди этих задач центральное место принадлежит монотонной дуализации — задаче построения сокращённой дизъюнктивной нормальной формы монотонной булевой функции, заданной конъюнктивной нормальной формой. Задача допускает гиперграфовую формулировку и матричную формулировку с использованием понятия неприводимого покрытия булевой матрицы. Труднорешаемость монотонной дуализации имеет два аспекта: экспоненциальный рост числа решений при увеличении размера задачи и сложность их нахождения (перечисления). Наиболее эффективными считаются алгоритмы с полиномиальным шагом (алгоритмы с полиномиальной задержкой). Полиномиальные алгоритмы построены лишь для некоторых частных случаев монотонной дуализации. В настоящее время сформировано два основных направления исследований.

Первое направление нацелено на построение так называемых инкрементальных алгоритмов монотонной дуализации, когда алгоритму разрешено просматривать решения, найденные на предыдущих шагах. При этом оценка

сложности шага алгоритма даётся для худшего случая (для самого сложного варианта задачи). В 1996 г. Л. Г. Хачияном и М. Л. Фридманом построен инкрементальный алгоритм монотонной дуализации с квазиполиномиальным шагом, определяемым фактически не только размером входа задачи, но и размером её выхода.

Второе направление основано на построении асимптотически оптимальных алгоритмов дуализации, впервые предложенных в 1977 г. Е. В. Дюковой. В этом случае алгоритму разрешено делать лишние полиномиальные шаги при условии, что их число почти всегда должно быть достаточно мало по сравнению с числом всех решений задачи. В рамках данного направления удалось построить алгоритмы монотонной дуализации, эффективные в типичном случае (эффективные для почти всех вариантов задачи). Асимптотически оптимальные алгоритмы используют матричную формулировку задачи монотонной дуализации и являются лидерами по скорости счёта, среди которых одним из наиболее быстрых является алгоритм RUNC-M (Е. В. Дюкова, П. А. Прокофьев 2015 г.). Существуют обобщения задачи монотонной дуализации на случай целочисленной матрицы (предложены Е. В. Дюковой в 1987 г.), основанные на модификации понятия неприводимого покрытия булевой матрицы, для которых также построены асимптотически оптимальные алгоритмы.

Современные прикладные задачи классификации не всегда могут быть описаны в рамках классической постановки, когда отдельные значения признака сравниваются с использованием простых отношений «равно» и «не равно». В ряде случаев возникает необходимость рассматривать более сложные отношения на множествах допустимых значений признаков. Например, когда на этих множествах заданы отношения частичных порядков и описания объектов представляют собой элементы декартова произведения конечных частично упорядоченных множеств.

Пусть $M = N_1 \times \cdots \times N_n$, где N_i — частично упорядоченное множество значений признака x_i, $i \in \{1, 2, \ldots, n\}$. Запись $a \preceq b$ ($b \succeq a$) означает, что b следует за a. Элементы a, b из частично упорядоченного множества N_i называются сравнимыми, если a следует за b или b следует за a. В противном случае a и b несравнимы. Если все элементы множества N_i попарно сравнимы, то множество N_i называется *линейно* упорядоченным или *цепью*. Если все различные элементы множества N_i попарно несравнимы, то множество N_i называется *антилинейно* упорядоченным или *антицепью*. Обозначим $a \prec b$, если $a \preceq b$ и $a \neq b$. Объект $S = (a_1, \ldots, a_n) \in M$ следует за объектом $S' = (b_1 \ldots b_n) \in M$, если a_i следует за b_i при $i = 1, 2, \ldots, n$. Таким образом,

на множестве объектов из M естественным образом возникает отношение частичного порядка.

Введём обобщённую функцию близости $\tilde{B}(S, \sigma, H)$ между объектом $S = (a_1, \ldots, a_n)$ и **ЭК** (σ, H), которая принимает значение 1, если $a_{j_i} \preceq \sigma_i$, $i = 1, 2, \ldots, r$, и 0 в противном случае. При построении процедур классификации, учитывающих частичную упорядоченность данных, требуется ввести более общие понятия корректного **ЭК**, представительного **ЭК** и покрытия класса, используя функцию близости $\tilde{B}$. Актуальным является построение асимптотически оптимальных методов поиска корректных **ЭК** общего вида, опирающихся на получение асимптотических оценок типичного числа и типичного ранга таких ЭК.

Как правило, результат классификации существенно зависит от того, какие частичные порядки заданы на множествах значений признаков. При этом выбор частичных порядков, обеспечивающих высокое качество классификации, путём полного перебора возможных вариантов бесперспективен в силу колоссальной вычислительной сложности. Важными являются вопросы синтеза вычислительно эффективных процедур выбора частичных порядков на множествах значений признаков, гарантирующих корректную классификацию.

Хорошо известно, что использование алгоритмов стохастических композиций приводит к улучшению качества классификации и повышению скорости обучения логических классификаторов. На данный момент, сильнейшими алгоритмами классификации являются стохастические композиции над решающими деревьями, такие как CatBoost или LGBM. Поэтому актуальным является создание стохастических композиций над логическими классификаторами в случае, когда информация представлена в виде декартова произведения конечных частично упорядоченных множеств.

Основной **целью** данной работы является обобщение процедур CVP на случай, когда данные представляют собой элементы декартова произведения конечных частично упорядоченных множеств. В классическом варианте порядок на множестве прецедентов фактически не установлен, так как отдельные значения каждого признака несравнимы между собой.

Рассмотрены следующие конкретные **задачи**.

1. Описание в рамках терминологии CVP единой схемы синтеза алгоритмов логической классификации, включающей все три названных направления, а именно CVP, LAD и FCA.

2. Обобщение классических понятий CVP на случай частично упорядоченных данных. Разработка и исследование моделей голосования по кор-

ректным **ЭК** общего вида. Создание алгоритмов поиска корректных **ЭК** общего вида на основе асимптотически оптимального решения задачи дуализации над произведением частичных порядков.

3. Разработка и исследование методов задания частичных порядков на множествах значений признаков, обеспечивающих корректность классификации. Создание процедуры линейного упорядочения множеств допустимых значений признаков, обеспечивающей высокое качество классификации при меньших временных затратах, но не гарантирующей корректность классификации.

4. Создание стохастических композиций построенных процедур классификации над произведением частичных порядков.

Методы исследования. Применялись методы дискретной математики, в частности алгебры логики, теории дизъюнктивных нормальных форм логических функций, методов построения покрытий булевых и целочисленных матриц, современных подходов к синтезу алгоритмов для дискретных перечислительных задач. Экспериментальное исследование проводилось с использованием программ, разработанных автором.

Основные положения, выносимые на защиту:

1. Разработана единая схема синтеза алгоритмов логической классификации с использованием терминологии Correct Voting Procedures (CVP).

2. Создана общая схема синтеза логических классификаторов в случае частично упорядоченных данных, которая может быть использована для описания классических логических алгоритмов классификации и предлагаемых в рамках данной работы моделей.

3. Предложена методика повышения качества классификации без потери корректности за счёт задания частичного порядка на множествах значений признаков.

4. Разработан асимптотически оптимальный алгоритм дуализации над произведением цепей RUNC-M+. Дано теоретическое обоснование предлагаемому алгоритму на основе матричной формулировки задачи дуализации над произведением частичных порядков.

5. Созданы методики повышения качества классификации путём синтеза стохастических композиций логических классификаторов.

6. Экспериментально подтверждено, что предлагаемые в рамках данной работы процедуры результативно применимы для решения проблемы классификации по прецедентам.

Научная новизна. Впервые создана единая схема синтеза логических процедур классификации по прецедентам, включающая направления CVP, LAD и FCA, и для направления CVP построены корректные логические классификаторы над произведением частичных порядков. Развит асимптотически оптимальный подход к труднорешаемым перечислительным дискретным задачам, возникающим на этапе обучения построенных классификаторов. Поставлена задача выбора «корректных» частичных порядков (гарантирующих корректность классификации) на множествах значений признаков, и намечены пути её решения. Предложена быстрая процедура выбора «некорректных» линейных порядков на множествах значений признаков. Разработаны и экспериментально исследованы практические модели стохастических композиций логических классификаторов над произведением частичных порядков. Все полученные результаты являются новыми.

Теоретическая и практическая значимость. Диссертационная работа содержит как теоретические, так и практические результаты.

Рассмотрены методологические аспекты алгоритмов логической классификации. Показано, что известные алгоритмы логической классификации, описанные в традициях разных научных направлений, могут быть синтезированы в рамках единой схемы.

Создана общая схема синтеза корректных логических алгоритмов классификации по прецедентам над произведением частичных порядков, согласно которой классические модели корректного голосования — это классификаторы над произведением антицепей. Показано, что в случае представления данных в виде декартова произведения конечных частичных порядков синтез процедур CVP сводится к решению задачи дуализации над произведением частичных порядков.

Дана матричная формулировка задачи дуализации над произведением частичных порядков, в рамках которой установлено, что в общем случае анализ прецедентной информации приводит к необходимости находить так называемые упорядоченные тупиковые покрытия целочисленной матрицы. На основе изучения метрических (количественных) свойств множества упорядоченных тупиковых покрытий целочисленной матрицы получена асимптотика типичного числа решений задачи дуализации над произведением цепей, и для этой задачи разработан асимптотически оптимальный алгоритм. Построены и реализованы асимптотически оптимальные классификаторы над произведением конечных цепей.

Исследована актуальная и ранее не изучавшаяся задача выбора на этапе предварительного анализа обучающей выборки «хороших» частичных порядков. Разработана «быстрая» процедура линейного упорядочения множеств допустимых значений признаков, эффективная по времени вычислений и позволяющая повысить качество классификации, но не гарантирующая корректность классификации. Поставлена задача выбора «корректных» частичных порядков на множествах допустимых значений признаков, и показано, что эта задача может быть решена на основе построения неприводимых покрытий специальной булевой матрицы.

На базе идей бэггинга и бустинга построены и экспериментально исследованы стохастические композиции логических классификаторов, использующих голосование по тупиковым представительным **ЭК** общего вида.

Степень достоверности и апробация работы. Достоверность полученных результатов обеспечивается доказательствами сформулированных утверждений и теорем, а также результатами экспериментов, проведённых автором. Результаты работы докладывались и обсуждались на следующих научных конференциях: «Математические методы распознавания образов (ММРО-18)» (Таганрог, 2017), «Дискретные модели в теории управляющих систем» (Подмосковье, 2018), «Математические методы распознавания образов (ММРО-19)» (Москва, 2019), «Конференция по искусственному интеллекту (КИИ-2019)» (Ульяновск, 2019), «Интеллектуализация обработки информации (ИОИ-13)» (Москва, 2020), «Информационные технологии и нанотехнологии (ИТНТ-2021)» (Самара, 2021), «Математические методы распознавания образов (ММРО-20)» (Москва, 2021), «Информационные технологии и нанотехнологии (ИТНТ-2023)» (Самара, 2023).

Результаты диссертационной работы включены в отчёты по двум проектам Российского фонда фундаментальных исследований: №16-01-00445 «Логический анализ данных в распознавании: обобщение классических подходов и вычислительные аспекты» и №19-01-00430 «Логический анализ частично упорядоченных целочисленных данных в задачах классификации и поиска ассоциативных правил».

Личный вклад. Все приведенные в диссертации результаты получены диссертантом лично при научном руководстве д.ф.-м.н. Е. В. Дюковой.

Публикации. По тематике работы опубликовано 15 научных работ. В изданиях ВАК опубликованы работы [2, 6, 8, 12]. Англоязычные версии работ [2, 8] опубликованы в журналах, индексируемых в Web Of Science Core Collection. В изданиях, индексируемых в Scopus, опубликованы работы [7, 9, 15] и англоязычная версия работы [13].

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и списка литературы. Полный объём диссертации 111 страниц, из них 97 страниц текста, включая 2 рисунка и 5 таблиц. Список литературы содержит 93 наименования на 11 страницах.

Содержание работы

Во введении сформулированы основные цели и задачи, описаны основные результаты и структура диссертационной работы.

В первой главе рассматривается задача классификации по прецедентам и даётся обзор трёх основных подходов к построению логических алгоритмов классификации, а именно CVP, LAD и FCA. Приводится описание единой схемы синтеза алгоритмов логической классификации, включающей направления CVP, LAD и FCA.

Показано, что каждое направление вводит специальный частичный порядок на множестве $\mathcal{P}(K)$ представительных **ЭК** класса K и в качестве наиболее информативных **ЭК** рассматривает максимальные относительно заданного порядка элементы множества $\mathcal{P}(K)$. Элемент частично упорядоченного множества называется *максимальным*, если за ним не следует ни один другой элемент из этого множества.

В частности, для процедур голосования по тупиковым представительным **ЭК** на множестве $\mathcal{P}(K)$ задаётся отношение частичного порядка $\preceq_1$, согласно которому **ЭК** $(\sigma_1, H_1) \in \mathcal{P}(K)$ следует за **ЭК** $(\sigma_2, H_2) \in \mathcal{P}(K)$ (т.е. $(\sigma_2, H_2) \preceq_1 (\sigma_1, H_1)$), если $\sigma_1 \subseteq \sigma_2$, $H_1 \subseteq H_2$. Доказано следующее

Утверждение 1.5.1. ***ЭК*** $(\sigma, H) \in \mathcal{P}(K)$ *является тупиковым представительным для класса* K *тогда и только тогда, когда* (σ, H) *— максимальный относительно частичного порядка* $\preceq_1$ *элемент множества* $\mathcal{P}(K)$.

Аналогичные утверждения доказаны для направлений LAD и FCA.

Во второй главе дано описание схемы синтеза процедур CVP для случая частично упорядоченных данных.

Понятия корректного **ЭК** класса K, представительного **ЭК** класса K и покрытия класса K переносятся на рассматриваемый случай заменой функции близости $B(S, \sigma, H)$ между объектом S и **ЭК** (σ, H) на функцию $\tilde{B}(S, \sigma, H)$. При этом предполагается, что частично упорядоченное множество значений признаков N_i, $i = 1, 2, \ldots, n$, содержит наибольший элемент k_i.

Пусть (σ, H) — **ЭК**, в котором $H = \{x_{j_1}, \ldots, x_{j_r}\}$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, $\sigma_i \in N_{j_i}$, $i = 1, 2, \ldots, r$. **ЭК** (σ, H) сопоставим набор $S_{(\sigma,H)} = (\gamma_1, \ldots, \gamma_n)$ из M, в котором $\gamma_t = \sigma_i$ при $t \in \{j_1, \ldots, j_r\}$, и $\gamma_t = k_t$ при $t \notin \{j_1, \ldots, j_r\}$.

Корректный для класса K **ЭК** назовём тупиковым, если любой другой **ЭК** (σ', H') такой, что $S_{(\sigma,H)} \preceq S_{(\sigma',H')}$, не является корректным **ЭК** класса K.

Через $R(K)$ обозначим множество прецедентов из класса K. $R(K)^+$ — множество объектов из M, которые следуют за хотя бы одним объектом из $R(K)$. Элемент $S \in M$ называется независимым от $R(K)$, если $S \in M \setminus R(K)^+$. Задача построения множества $I(R(K))$, содержащего все максимальные элементы множества $M \setminus R(K)^+$ известна как задача дуализации над произведением частичных порядков и относится к классу труднорешаемых.

Утверждение 2.2.1. *Покрытие (σ, H) класса K является тупиковым покрытием класса K тогда и только тогда, когда $S(\sigma, H) \in I(R(K))$.*

Пусть $\overline{K} = M \setminus K$. Будем рассматривать $\overline{K}$ как отдельный класс, т.е. будем считать, что есть всего два класса K и $\overline{K}$.

Утверждение 2.2.2. ***ЭК*** (σ, H) *является тупиковым представительным для класса K тогда и только тогда, когда $S(\sigma, H) \in I(R(\overline{K}))$ и $S(\sigma, H) \in R(K)^+$.*

Таким образом показано, что поиск тупиковых корректных **ЭК** общего вида сводится к решению задачи дуализации над произведением частичных порядков.

В общем случае существование представительных для класса K **ЭК** не гарантировано. Пусть $\tilde{\tilde{M}} = \tilde{\tilde{N}}_1 \times \cdots \times \tilde{\tilde{N}}_n$, $\tilde{\tilde{N}}_i$ совпадает с N_i, $i = 1, 2, \ldots, n$, но на $\tilde{\tilde{N}}_i$ задано обратное отношение порядка, т.е. $a \preceq b$ в $\tilde{\tilde{N}}_i$ тогда и только тогда, когда $b \preceq a$ в N_i.

Зададим отображение $\psi : M \to M \times \tilde{\tilde{M}}$ следующим образом. Отображение ψ переводит объект $S = (a_1, \ldots, a_n)$ из M в объект $\psi(S) = (a_1, \ldots, a_n, a_{n+1}, \ldots, a_{2n})$ из $M \times \tilde{\tilde{M}}$, в котором $a_{i+n} = a_i$ при $i \in \{1, 2, \ldots, n\}$, т.е. признаковое описание объекта S дублируется с обратным отношением порядка.

Пусть $\psi(A), A \subset M$, — образ A при отображении ψ. Имеет место следующая

Теорема 2.2.1. *Если классы множества M не пересекаются, то любой прецедент из класса $\psi(K)$ порождает тупиковый представительный* ***ЭК*** *класса $\psi(K)$.*

Согласно теореме 2.2.1 существует такое преобразование признакового описания множества M, которое обеспечивает корректность классификации.

Отметим, что этап обучения — это самый сложный в вычислительном плане этап логической классификации из-за необходимости решать задачу дуализации над произведением частичных порядков, число решений которой растёт экспоненциально с ростом размера входа задачи. Поэтому описанный метод преобразования признакового пространства применим только в случае небольшого числа признаков.

В данной главе дано описание практических моделей логической классификации над произведением частичных порядков, основанных на стохастической композиции над обобщёнными логическими классификаторами. Предлагаемые модели основаны на известных способах ансамблирования (бэггинг и бустинг), в которых в качестве базового классификатора использован алгоритм голосования по представительным **ЭК**. Эти модели не гарантируют корректность классификации, но демонстрируют высокое качество классификации на реальных задачах, что показало проведённое подробное экспериментальное исследование.

В третьей главе разработаны эффективные методы задания частичных порядков на множествах значений признаков, обеспечивающих корректность классификации. Описана быстрая процедура линейного некорректного упорядочения значений признаков и приведены результаты её тестирования на реальных задачах.

Пусть A — классификатор над произведением частичных порядков, строящий все тупиковые представительные **ЭК** класса K. Тогда справедлива

Теорема 3.1.1. *Алгоритм A классифицирует правильно объект S' из $R(K)$ тогда и только тогда, когда $S' \in M \setminus R(\overline{K})^+$.*

Частичный порядок на множестве M называется (A, K)-*корректным*, если алгоритм A правильно классифицирует каждый объект из $R(K)$. Построим булеву матрицу B_K. Каждой паре объектов (S', S''), где $S' \in R(K)$ и $S'' \in R(\overline{K})$, соответствует строка в матрице B_K. Каждому признаку x_j, $j \in \{1, 2, \ldots, n\}$, и каждой паре (a, b), $a \in N_j$, $b \in N_j$, $a \neq b$, соответствует столбец (x_j, a, b) матрицы B_K. Элемент матрицы B_K, расположенный на пе-

ресечении строки (S', S'') и столбца (x_j, a, b), равен 1, если значение признака x_j равно a и b у объектов S' и S'' соответственно.

Набор столбцов H матрицы B_K называется *покрытием*, если каждая строка матрицы B_K в пересечении хотя бы с одним из столбцов, входящих в H, дает 1. Покрытие матрицы B_K называется *неприводимым*, если любое его собственное подмножество покрытием не является.

С использованием утверждения теоремы 3.1.1 доказана

Теорема 3.3.1. *Частичный порядок, заданный на множестве M, является (A, K)-корректным тогда и только тогда, когда существует покрытие H матрицы B_K такое, что для любого $j \in \{1, 2, \ldots, n\}$ и для любых $a, b \in N_j$, $b \prec a$, столбец (x_j, a, b) не входит в H.*

Частичный порядок на множестве объектов из M называется *линейным (антилинейным) на множестве M*, если каждое множество N_j, $j = 1, 2, \ldots, n$ является цепью (антицепью).

Рассмотрим (A, K)-корректный линейный порядок на множестве M. Согласно теореме 3.3.1 существует покрытие H матрицы B_K такое, что для любого столбца $(x_j, a, b) \in H$, $a, b \in N_j$, не выполнено $b \prec a$. Поскольку множество N_j является цепью, то $a \prec b$. Поэтому справедливо

Следствие 3.3.1. *Линейный порядок на множестве M является (A, K)-корректным тогда и только тогда, когда существует покрытие H матрицы B_K такое, что $a \prec b$ для любого столбца (x_j, a, b) из покрытия H.*

Пусть на множестве M задан (A, K)-корректный антилинейный порядок. Тогда для всех $j = 1, 2, \ldots, n$ и для всех $a, b \in N_j$ не выполнено $b \prec a$. Следовательно, для любого покрытия матрицы B_K выполнены условия теоремы 3.3.1. Поэтому справедливо

Следствие 3.3.2. *Антилинейный порядок на множестве M является (A, K)-корректным для любого класса K из $\{K_1, \ldots, K_l\}$.*

С целью увеличения числа признаков с линейно упорядоченным множеством значений была разработана генетическая процедура поиска покрытия матрицы B_K, по длине близкого к минимальному. Необходимость использования генетического алгоритма обусловлена большими размерами матрицы B_K и её разреженностью по числу единиц.

Для линейного упорядочения множества значений отдельного признака использовался алгоритм топологической сортировки с линейным временем работы. В случае невозможности линейного упорядочения на множестве значений признака устанавливался антилинейный порядок.

Отметим, что каждое покрытие матрицы B_K может порождать несколько (A, K)-корректных частичных порядков, из-за чего описанная процедура выбора корректного частичного порядка имеет высокую степень неопределённости. Этого недостатка лишена описанная в данной главе процедура выбора линейного порядка на множестве M, основанная на оценке информативности значений отдельных признаков и не гарантирующая корректность классификации.

Пусть $\mu_{ij}^{(1)}(a)$ и $\mu_{ij}^{2}(a)$, $i \in \{1, 2, \ldots, l\}$, $j \in \{1, 2, \ldots, n\}$, $a \in N_j$, — соответственно доля прецедентов класса K_i и доля прецедентов не из класса K_i, у которых признак x_j принимает значение a. Величина $\mu_{ij}(a) = \mu_{ij}^{1}(a) - \mu_{ij}^{2}(a)$ служит мерой важности значения a признака x_j в классе K_i. Для каждого класса $K_i, i \in \{1, 2, \ldots, l\}$, и для каждого признака x_j, $j \in \{1, 2, \ldots, n\}$, зададим следующий линейный порядок: $\forall y, z \in N_j$, $y \preceq z$ тогда и только тогда, когда $\mu_{ij}(y) \geq \mu_{ij}(z)$. Описанная процедура обеспечивает высокое качество классификации и незначительное время счёта, что подтверждено результатами экспериментального исследования.

В четвёртой главе приведена матричная формулировка задачи дуализации над произведением частичных порядков. Показано, что данная задача сводится к поиску так называемых упорядоченных тупиковых покрытий целочисленной матрицы. Понятие упорядоченного тупикового покрытия целочисленной матрицы обобщает известное понятие тупикового покрытия целочисленной матрицы. Для дуализации над произведением цепей построен асимптотически оптимальный алгоритм RUNC-M+. Его теоретическое обоснование базируется на приведённой ниже теореме 4.2.1.

Пусть $P = P_1 \times P_2 \times \cdots \times P_n$, P_i — конечное частично упорядоченное множество с наибольшим элементом. Введём обозначения: $Q_1(x, P)$, $x \in P$, — множество всех элементов в P, непосредственно следующих за x ($Q_1(x, P) = \{y \in P : x \prec y,\ \forall a \in P : x \prec a \Rightarrow a \not\prec y\}$); $Q_2(x, y, P)$, $x \in P$, $y \in Q_1(x, P)$, — множество всех элементов в P, не предшествующих x и предшествующих y ($Q_2(x, y, P) = \{a \in P : a \not\preceq x, a \preceq y\}$).

Пусть M_{mn}^{k} — совокупность всех матриц размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$, $k \geq 2$; E_k^r, $r \leq n$, — множество всех наборов вида $(\sigma_1, \ldots, \sigma_r)$, в которых $\sigma_i \in \{0, 1, \ldots, k-1\}$, $k \geq 2$, при $i = 1, 2, \ldots, r$. Рассмотрим $\sigma \in E_k^r$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, $\sigma_i < k - 1$, $i = 1, 2, \ldots, r$. Через $Q_i(\sigma)$, $i \in \{1, 2, \ldots, r\}$,

обозначим множество наборов $(\beta_1, \ldots, \beta_r)$ в E_k^r, таких что $\beta_i = \sigma_i + 1$ и $\beta_j \leq \sigma_j$ при $j \in \{1, 2, \ldots, r\} \setminus \{i\}$.

Пусть H — набор из r различных столбцов матрицы $L \in M_{mn}^k$. Множество различных строк подматрицы матрицы L, образованной столбцами набора H, можно рассматривать как некоторое подмножество E^H наборов из E_k^r. Набор столбцов H называется упорядоченным тупиковым σ-покрытием матрицы L, если выполнены два следующих условия:

1) E^H не содержит набор $(\beta_1, \ldots, \beta_r) \in E_k^r$, в котором $\beta_j \leq \sigma_j$ при $j \in \{1, 2, \ldots, r\}$;

2) если $i \in \{1, 2, \ldots, r\}$, то E^H содержит хотя бы один набор из $Q_i(\sigma)$.

Если $L \in M_{mn}^2$ и набор столбцов H является упорядоченным тупиковым $(0, 0, \ldots, 0)$-покрытием матрицы L, то H — неприводимое покрытие матрицы L.

Заметим, что если $P_i,\ i \in \{1, 2, \ldots, r\}$, — конечная цепь и $x \in P_i$ не является наибольшим элементом в P_i, то множество $Q_1(x, P)$ состоит из одного элемента, обозначаемого далее через $x+1$, и следовательно, $Q_2(x, x+1, P) = \{x + 1\}$. Поэтому в случае произведения конечных цепей условие 1) из определения упорядоченного тупикового σ-покрытия превращается в следующее условие: для любого $i \in \{1, 2, \ldots, n\}$ подматрица L_R^H матрицы L_R, образованная столбцами из H, содержит строку $(\beta_1, \ldots, \beta_{i-1}, \sigma_{i+1}, \sigma_{i+1}, \ldots, \sigma_r)$, где $\beta_t \preceq \sigma_t$ при $t \neq i,\ t \in \{1, 2, \ldots, r\}$.

Возьмём элемент $x = (x_1, \ldots, x_n) \in P$, в котором компонента $x_{j_i} = \sigma_i, i \in \{1, 2, \ldots, r\}$, не является наибольшим элементом в P_{j_i}, а каждая из остальных компонент $x_j,\ j \in \{1, 2, \ldots, r\} \setminus \{j_1, \ldots, j_r\}$ — наибольший элемент в P_j. Положим $\sigma = (\sigma_1, \ldots, \sigma_r)$. Имеет место следующая

Теорема 4.1.1 Элемент x является максимальным независимым от R тогда и только тогда, когда набор столбцов матрицы L_R с номерами $j_1, \ldots, j_r$ является упорядоченным тупиковым σ-покрытием матрицы L_R.

Квадратную подматрицу порядка r матрицы $L \in M_{mn}^k$ назовем упорядоченной σ-подматрицей, если для множества E^H выполнено $E^H \cap Q_i(\sigma) \neq \varnothing$ при $i \in \{1, 2, \ldots, r\}$.

Обозначим: $\phi_d,\ d > 0$, — интервал $(\frac{1}{2}\log_d mn - \frac{1}{2}\log_d \log_d mn - \log_d \log_d \log_d n, \frac{1}{2}\log_d mn - \frac{1}{2}\log_d \log_d mn + \log_d \log_d \log_d n)$; $\Pi_r(\sigma) = (\sigma_1 + 1)^{r-1} \ldots (\sigma_r + 1)^{r-1}, \sigma \in E_{k-1}^r$.

Пусть $L \in M_{mn}^k$, $\sigma \in E_{k-1}^r$. Положим $B(L,\sigma)$ — множество всех упорядоченных тупиковых σ-покрытий матрицы L; $S(L,\sigma)$ — множество всех упорядоченных σ-подматриц матрицы L;

$$\Sigma_1(L) = \sum_{r=1}^{n} \sum_{\sigma \in E_{k-1}^r} |B(L,\sigma)|;$$

$$\Sigma_2(L) = \sum_{r=1}^{n} \sum_{\sigma \in E_{k-1}^r} |S(L,\sigma)|.$$

Теорема 4.2.1 *Если $m^\alpha \le n \le d^m$, $\alpha > 1$, $d = k/(k-1)$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо*

$$\Sigma_1(L) \sim \Sigma_2(L) \sim \sum_{r \in \phi_d} \sum_{\sigma \in E_{k-1}^r} \Pi_r(\sigma) C_n^r C_m^r r! k^{-r^2}$$

и длины почти всех упорядоченных тупиковых покрытий матрицы L принадлежат интервалу ϕ_d.

Из теоремы 4.2.1 следуют оценки типичных значений количественных характеристик множества неприводимых покрытий булевой матрицы:

Следствие 4.2.1. *Если $m^\alpha \le n \le 2^m$, $\alpha > 1$, то для почти всех матриц L из M_{mn}^2 при $n \to \infty$ справедливо*

$$\Sigma_1(L) \sim \Sigma_2(L) \sim \sum_{r \in \phi_2} C_n^r C_m^r r! 2^{-r^2},$$

и длины почти всех неприводимых покрытий матрицы L принадлежат интервалу ϕ_2.

В заключении приводятся положения диссертации, выносимые на защиту и задаются направления дальнейших исследований.

Список литературы включает 93 публикации.

Публикации автора по теме диссертации

1. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О дуализации над произведением частичных порядков. // Машинное обучение и анализ данных, 2017. — Том 3, Вып. 4. — С. 239–249.

2. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* Дуализация над произведением цепей: асимптотичекие оценки числа решений // Доклады Академии наук, 2018. — Т. 483, №2. — С. 130–133 (**ВАК**);
Djukova E.V., Masliakov G.O., Prokofjev P.A. Dualization Problem over the Product of Chains: Asymptotic Estimates for the Number of Solution // Doklady Mathematics, 2018. — Vol. 98, No. 3. — P. 564–567 (**WOS CORE**).

3. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О поиске максимальных независимых элементов частичных порядков (случай цепей) // Труды «Прикладная математика и информатика», 2018. — Т. 58. — С. 12–20.

4. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* Задача монотонной дуализации и её обобщение — дуализация над произведением цепей // Труды 10-й Международной конференции «Дискретные модели в теории управляющих систем». — Москва и Подмосковье 23-25 мая 2018 г. — Труды / Отв. ред. В. Б. Алексеев, Д. С. Романов, Б. Р. Данилов. Москва: МАКС Пресс, 2018. — С. 117–119.

5. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* Задача дуализации над произведением цепей: асимптотика типичного числа решений // Тезисы докладов 12-й Международной конференции «Интеллектуализация обработки информации (ИОИ-2018)». — Москва, Россия – Гаэта, Италия. М.: ТОРУС ПРЕСС, 2018. — С. 12–159.

6. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О числе максимальных независимых элементов частичных порядков (случай цепей) // Информатика и её применения, 2019. — Т. 13, Вып. 1. — С. 25–32 (**ВАК**).

7. *Djukova E.V., Masliakov G.O., Prokofjev P.A.* Finding Maximal Independent Elements of Products of Partial Orders (the Case of Chains) // Computational Mathematics and Modeling, 2019. — Vol. 30, Iss. 1. — P. 7–12 (**SCOPUS**).

8. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О логическом анализе данных с частичными порядками в задаче классификации по прецедентам // Ж. вычисл. матем. и матем. физ., 2019. — Т. 59, №9. — С. 1605–1616 (**ВАК**);
Djukova E.V., Masliakov G.O., Prokofjev P.A. On the Logical Analysis of Partially Ordered Data in the Supervised Classification Problem // Computational Mathematics and Mathematical Physics, 2019. — Vol. 59, Iss. 9. — P. 1542–1552 (**WOS CORE**).

9. *Djukova E.V., Masliakov G.O., Prokofjev P.A.* Logical Classification of Partially Ordered Data // 7th Russian Conference, RCAI 2019. — Ulyanovsk, Russia. October 21–25, 2019. — Proceedings. Editors: Kuznetsov, Sergei O., Panov, Aleksandr I. — P. 115–126 (**SCOPUS**).

10. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* Классификация над произведением частичных порядков // Тезисы докладов 19-й Всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2019)». — Москва, 26—29 декабря 2019 г. — С. 29–31.

11. *Бакланова А.О., Дюкова Е.В., Масляков Г.О.* Исследование зависимости качества классификации от выбора частичных порядков на множествах значений признаков // Тезисы докладов 13-й Международной конференции «Интеллектуализация обработки информации (ИОИ-2020)». — г. Москва. 2020. — С. 21–25.

12. *Дюкова Е.В., Масляков Г.О.* О выборе частичных порядков на множествах значений признаков в задаче классификации // Информатика и её применения, 2021. — Т. 15, Вып. 4. — С. 74–80 (**ВАК**).

13. *Дюкова Е. В., Масляков Г.О.* О корректной классификации над произведением частичных порядков // Сборник трудов по материалам VII Международной конференции и молодёжной школы (ИТНТ-2021), Т. 3. — Самара, 2021. — С. 34292;
Djukova E.V., Masliakov G.O. Correct classification over a product of partial orders // IEEE Proceedings of the VII International Conference on Information Technology and Nanotechnology (ITNT-2021). — Samara, Russia, 2021. — P. 1–5 (**SCOPUS**).

14. *Дюкова Е. В., Масляков Г.О.* Корректная классификация над произведением частичных порядков // Тезисы докладов 20-й Всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2021)». — Москва 7–10 декабря 2021 г. — С. 59–63.

15. *Djukova E.V., Masliakov G.O.* Correct classification over a product of partial orders // IEEE Proceedings of the IX International Conference on Information Technology and Nanotechnology (ITNT-2023). — Samara, Russia, 2023. — P. 1–5 (**SCOPUS**).