

На правах рукописи



Драгунов Никита Аркадьевич

**Поиск частых (нечастых) элементов декартова произведения
конечных частичных порядков и приложения**

Специальность 1.2.3 – «Теоретическая информатика, кибернетика»

АВТОРЕФЕРАТ

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2026

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук.

Научный руководитель:

Дюкова Елена Всеволодовна, доктор физико-математических наук, доцент, главный научный сотрудник ФИУ ИУ РАН.

Официальные оппоненты:

Гордеев Эдуард Николаевич, доктор физико-математических наук, профессор, профессор ФГАОУ ВО «Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)».

Двоенко Сергей Данилович, доктор физико-математических наук, профессор, профессор ФГБОУ ВО «Тульский государственный университет».

Ведущая организация:

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики»».

Защита состоится «__» _____ 2026 г. в __ часов на заседании диссертационного совета 24.1.224.03 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу: 119333, Россия, Москва, ул. Вавилова, д. 42.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук и на сайте https://www.frccsc.ru/diss-council/00207305/diss/list/dragunov_na.

Автореферат разослан «__» _____ 2026 г.

Учёный секретарь
диссертационного совета 24.1.224.03,
кандидат технических наук



Рейер Иван Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Задачи поиска частых и нечастых элементов в данных являются одними из центральных задач анализа информации и имеют широкий спектр приложений, в частности, в дискретной математике, машинном обучении и при нахождении ассоциативных правил. Во многих прикладных областях данные естественным образом упорядочены, однако традиционные алгоритмы решения рассматриваемых задач порядок не учитывают. В связи с этим возникает необходимость разработки эффективных методов поиска частых и нечастых элементов для случая, когда данные представлены в виде декартова произведения конечных частично упорядоченных множеств.

Конечное множество P называется *частично упорядоченным*, если на элементах этого множества задано бинарное отношение $\preceq$, обладающее свойствами:

- рефлексивности: $x \preceq x, \forall x \in P$;
- антисимметричности: если $x \preceq y$ и $y \preceq x$, то $x = y, \forall x \in P, \forall y \in P$;
- транзитивности: если $x \preceq y$ и $y \preceq z$, то $x \preceq z, \forall x \in P, \forall y \in P, \forall z \in P$.

Если $x \preceq y, x \in P, y \in P$, то говорят, что элемент y *следует* за элементом x или что x *предшествует* y . Элементы x и y называются *сравнимыми*, если либо y следует за x , либо x следует за y . Иначе x и y *несравнимы*. Множество P – *цепь* (*антицепь*), если любые два различные элемента в P сравнимы (несравнимы). Элемент x из P называется *максимальным* (*минимальным*), если не существует другого элемента из P , следующего за x (предшествующего x).

Задачи поиска частых и нечастых элементов в частично упорядоченных данных рассматриваются в следующей постановке. Дана база данных $D(P)$, представляющая собой совокупность некоторых не обязательно различных элементов из P . Элементы базы данных называются *транзакциями*. *Частотой* элемента $x \in P$ называется величина $\nu(x)$, равная числу транзакций в $D(P)$, следующих за x . Если P – антицепь, то $\nu(x)$ равна числу транзакций,

совпадающих с x . Задан некоторый порог s , $s \in \mathbb{N}$, $s \leq |D(P)|$. Здесь и далее $\mathbb{N}$ – это множество натуральных чисел, $|D(P)|$ – число транзакций в базе данных. Если $\nu(x) \geq s$, то элемент x называется s -частым. Иначе x называется s -нечастым элементом. Требуется по заданным $D(P)$ и s перечислить частые (нечастые) элементы.

На практике вместо перечисления всех частых (нечастых) элементов, как правило, строят более компактное множество максимальных частых (минимальных нечастых) элементов. Если $x \in P$ является s -частым элементом и в P не существует другого s -частого элемента, следующего за x , то x называется *максимальным s -частым* элементом. Если $x \in P$ является s -нечастым элементом и в P не существует другого s -нечастого элемента, предшествующего x , то x называется *минимальным s -нечастым* элементом. Далее в работе s -частые и s -нечастые элементы для краткости называются просто частыми и нечастыми элементами соответственно.

К поиску (максимальных) частых и (минимальных) нечастых элементов частично упорядоченного множества сводится ряд задач дискретной математики и машинного обучения. Имеются в виду важная в дискретной математике задача оптимальной расшифровки двузначной монотонной логической функций и методы логической классификации по прецедентам.

Традиционные алгоритмы решения названных задач вычислительно сложны, как правило, ориентированы на «худший случай» и, в основном, рассчитаны на работу с данными, в которых порядок на множестве рассматриваемых элементов не задан. Это существенно ограничивает их применимость на практике. В связи с чем актуальными являются вопросы разработки вычислительно эффективных алгоритмов, в основе которых лежат методы построения (максимальных) частых и (минимальных) нечастых элементов, а также обобщение этих алгоритмов для произвольных (или более сложных, чем антицепи) частичных порядков.

Степень разработанности темы. Задачи построения частых и нечастых элементов в данных впервые возникли в контексте анализа потребительской

корзины при поиске ассоциативных правил и наиболее изучены в случае бинарной информации. Одними из наиболее известных алгоритмов в этой области являются алгоритмы Apriori, DepthProject и FP-Growth.

В 2024 году К. Эльбассиони изучена важная с теоретической точки зрения задача совместного перечисления максимальных частых и минимальных нечастых элементов множества P при условии, что P представлено в виде декартова произведения конечных частично упорядоченных множеств $P_1, \dots, P_n$, $n \geq 2$ (запись $P = P_1 \times \dots \times P_n$). Считается, что элемент $y = (y_1, \dots, y_n)$ из P следует за элементом $x = (x_1, \dots, x_n)$ этого множества, если y_i следует за x_i при $i = \overline{1, n}$. Показано, что при решении указанной задачи появляется необходимость рассматривать одну из центральных труднорешаемых перечислительных проблем дискретной математики – дуализацию над произведением частичных порядков. Поясним сказанное.

Пусть R – некоторое непустое подмножество частично упорядоченного множества P . Через R^+ обозначается множество всех элементов из P следующих хотя бы за одним элементом из R , а через R^- – множество всех элементов из P предшествующих хотя бы одному элементу из R . Множество $I(R^+)$, состоящее из всех максимальных элементов множества $P \setminus R^+$, называется *максимальным независимым от R* . Множество $I(R^-)$, состоящее из всех минимальных элементов множества $P \setminus R^-$, называется *минимальным независимым от R* . Каждая из задач построения $I(R^+)$ и $I(R^-)$ называется дуализацией над произведением частичных порядков.

Особый интерес представляет случай, когда $P_i = \{0, 1, \dots, k-1\}$, $k \geq 2$, при $i = \overline{1, n}$, и в каждом P_i задан порядок $0 \leq 1 \leq \dots \leq k-1$. Тогда множество P называется *k -значным n -мерным кубом* со стандартным линейным порядком и обозначается далее через E_k^n . Множество E_2^n называется *булевым кубом*.

Обозначим через X_{max} и Y_{min} соответственно множества всех максимальных частых и минимальных нечастых элементов частичных порядков. Существует достаточно очевидный последовательный способ перечисления этих множеств, основанный на свойстве двойственности: $I(X_{max}^-) = Y_{min}$ и $I(Y_{min}^+) =$

X_{max} . То есть сначала находится одно из множеств X_{max} или Y_{min} , а затем второе множество строится путем дуализации первого. Экспериментальное исследование показывает, что подход является эффективным только в случае, когда одно из множеств частых или нечастых элементов существенно меньше другого.

Задача совместного перечисления максимальных частых и минимальных нечастых элементов частичных порядков имеет большое число практических приложений. В частности, она связана с важной в дискретной математике задачей расшифровки двузначной монотонной функции, которая в работе рассматривается в следующей постановке.

Исследуется функция f , определенная на элементах декартова произведения частично упорядоченных множеств $P = P_1 \times \dots \times P_n$, $n \geq 1$, и принимающая два значения 0 и 1. Функция f называется *монотонной*, если для любых двух элементов x и y из P таких, что $x \preceq y$, выполнено $f(x) \leq f(y)$. Монотонная функция f задается при помощи некоторого оператора B (оракула), который для любого $x \in P$ возвращает значение функции $f(x)$. Если $f(x) = 0$, то элемент x называется *нулем* функции f , если же $f(x) = 1$, то элемент x называется *единицей* функции f . Требуется путем «минимального» числа обращений к оператору B найти все нули и единицы функции f .

Традиционный подход к решению задачи расшифровки нацелен на случай функции, заданной на E_k^n , и основан на построении оптимального по Шеннону алгоритма. Согласно данному подходу, сложность алгоритма расшифровки следует оценивать числом обращений к оракулу B в «худшем случае» (то есть для самого сложного варианта задачи). Пусть V – множество всех двузначных монотонных функций, определенных на P ; A – некоторый алгоритм, выполняющий расшифровку функций из V . Обозначим через $t_A(f)$ общее число обращений к оператору B алгоритма A при расшифровке функции $f \in V$. Под сложностью алгоритма A по Шеннону (сложностью в худшем случае) понимается величина $\max_{f \in V} t_A(f)$, где максимум берется по всем функциям из V . Алгоритм называется *оптимальным по Шеннону* на V , если его сложность

минимальна среди всех алгоритмов, выполняющих расшифровку функций из V . Задача оптимальной по Шеннону расшифровки для монотонной функции, заданной на E_2^n (монотонная булева функция), поставлена В. К. Коробковым в 1965 году и решена Ж. Анселем в 1968 году. В 1976 г. В. Б. Алексеевым исследован случай функции, заданной на E_k^n , и предложен алгоритм, который является оптимальным по Шеннону в случае $k = 2$ и «близок» по сложности к оптимальному по Шеннону в случае $k > 2$.

Еще одним важным приложением методов поиска частых и нечастых элементов в данных является классификация по прецедентам. Задача рассматривается в стандартной для логического подхода постановке. Исследуется множество объектов M , которое представимо в виде объединения непересекающихся подмножеств $K_1, \dots, K_l$, называемых *классами*. Объекты из M описываются в системе числовых признаков $x_1, \dots, x_n$. Предполагается, что каждый признак x_i , $i = \overline{1, n}$, имеет ограниченное множество допустимых значений N_i , элементы которого кодируются целыми числами. На множестве N_i задается частичный порядок. Множество M фактически определяется как $M = N_1 \times \dots \times N_n$. Даны примеры объектов из каждого класса, называемые *прецедентами*. Требуется на основе анализа множества прецедентов построить алгоритм, умеющий по признаковому описанию объекта из M определять его класс.

Пусть $H = \{x_{j_1}, \dots, x_{j_r}\} \subseteq \{x_1, \dots, x_n\}$, $j_1 < \dots < j_r$, – набор из r различных признаков; $\sigma = (\sigma_1, \dots, \sigma_r)$, $\sigma_i \in N_{j_i}$, $i = \overline{1, r}$, – набор допустимых значений признаков из H . Пара (σ, H) называется *элементарным классификатором (ЭК) ранга r* . Объект $S = (a_1, \dots, a_n) \in M$ содержит ЭК (σ, H) , если $a_{j_i} \preceq \sigma_i$, $i = \overline{1, r}$.

Пусть $K \in \{K_1, \dots, K_l\}$, $\bar{K} = \{K_1, \dots, K_l\} \setminus \{K\}$, $R(K)$ и $R(\bar{K})$ – множество всех прецедентов из K и из $\bar{K}$ соответственно. ЭК (σ, H) называется *корректным* для класса K , если не существует пары прецедентов $S' \in R(K)$, $S'' \in R(\bar{K})$ таких, что S' и S'' содержат (σ, H) . Корректный ЭК (σ, H) называется

представительным для класса K , если хотя бы один прецедент из $R(K)$ содержит (σ, H) .

Отечественное направление логической классификации в основном представлено методами, именуемыми процедурами корректного голосования (Correct Voting Procedures или CVP). Фундаментальную роль в создании этого направления сыграли научные школы чл.-корр. РАН С. В. Яблонского и акад. РАН Ю. И. Журавлева. Направление развивается с 70-х годов 20-го века. В ряде публикаций отечественных и зарубежных авторов предложены и развиты два других направления логической классификации, а именно логический анализ данных (Logical Analysis of Data) и анализ формальных понятий (Formal Concept Analysis). Показано (Е. В. Дюкова, Г. О. Масляков, А. П. Дюкова, 2023), что в каждом из трёх направлений на этапе обучения логического классификатора для каждого класса K фактически задается некоторый свой частичный порядок на множестве представительных ЭК и строится множество максимальных относительно заданного порядка ЭК. При этом решаются сложные в вычислительном плане задачи.

Целью настоящей работы является разработка вычислительно эффективных методов логического анализа информации, представленной в виде декартова произведения частично упорядоченных конечных множеств (произведения частичных порядков), и применение этих методов в области дискретной математики и машинного обучения. Поставлены следующие **задачи**.

1. Разработать эффективный алгоритм совместного перечисления максимальных частых и минимальных нечастых элементов декартова произведения частично упорядоченных множеств.

2. Предложить ориентированный на «типичный случай» алгоритм расшифровки двузначной монотонной функции, определенной на декартовом произведении конечных частично упорядоченных множеств.

3. Построить быстрые алгоритмы логической классификации, базирующиеся на поиске часто встречающихся ЭК в описаниях прецедентов класса и ориентированные на обработку данных, представленных в виде

декартова произведения конечных частично упорядоченных множеств. Получить теоретические оценки типичного числа и ранга искомых ЭК.

Методология и методы исследования. В работе использованы методы дискретной математики, комбинаторики, машинного обучения, теории вычислительной сложности дискретных задач и теории вероятности. Экспериментальное исследование проведено с использованием программно-алгоритмического комплекса, разработанного автором.

Научная новизна работы заключается в следующем.

1. Предложен новый последовательно-совместный подход к перечислению множеств X_{max} максимальных частых и Y_{min} минимальных нечастых элементов декартова произведения частичных порядков.

2. На базе последовательно-совместного подхода к перечислению X_{max} и Y_{min} построен алгоритм асимптотически оптимальной расшифровки двузначной монотонной функции, заданной на декартовом произведении частично упорядоченных множеств.

3. На основе поиска частых элементов в данных разработаны новые быстрые алгоритмы логической классификации, базирующиеся на перечислении корректных ЭК специального вида, названных правильными представительными. ЭК ранга r называется правильным для класса K , если не менее чем r прецедентов класса K содержат этот ЭК. Для случая декартова произведения антицепей построен алгоритм REC перечисления правильных представительных ЭК. В дальнейшем была установлена асимптотическая оптимальность алгоритма REC (Е. В. Дюкова, 2025). Для случая декартова произведения цепей разработана его модификация REC+.

4. Развита техника получения оценок метрических (количественных) характеристик логических классификаторов. А именно, получены асимптотические оценки типичного числа и типичного ранга правильных ЭК для случаев антицепей (Теорема 1) и цепей (Теоремы 2 и 3). Тем самым дано теоретическое обоснование эффективности предложенных в работе классификаторов.

Теоретическая и практическая значимость работы. Предложенные в диссертации методы работают с данными, представленными в виде декартова произведения конечных частичных порядков. Алгоритмы программно реализованы и проверены на синтетических и реальных данных.

Алгоритм последовательно-совместного перечисления X_{max} и Y_{min} выигрывает у ранее известных последовательного и совместного методов в случае, когда мощности множеств частых и нечастых элементов не сильно отличаются. Алгоритм асимптотически оптимальной расшифровки двузначной монотонной функции работает в среднем быстрее ориентированного на худший случай алгоритма (В. Б. Алексеев, 1976) и требует меньшего числа обращений к оракулу. Классификаторы REC и REC+ требуют меньших вычислительных затрат, чем классические модели направления CVP, не уступая, а в ряде случаев превосходя их в качестве распознавания.

Полученные асимптотические оценки типичного числа и ранга правильных представительных ЭК (Теоремы 1–3) показывают, что число правильных представительных ЭК существенно меньше, чем число тупиковых ЭК, поиск которых ведётся в направлении CVP. Тем самым эффективность REC и REC+ подтверждена не только экспериментально, но и теоретически. Кроме того, полученные оценки позволяют обоснованно выбирать параметры алгоритмов (ранги искомых ЭК) на практике, не прибегая к перебору.

Положения, выносимые на защиту.

1. Последовательно-совместный подход к перечислению множеств максимальных частых и минимальных нечастых элементов декартова произведения частичных порядков. Доказательство корректности предлагаемого метода (Утверждения 1–3). Экспериментальное исследование подхода, подтверждающее его эффективность.

2. Алгоритм асимптотически оптимальной расшифровки двузначной монотонной функции, заданной на декартовом произведении частично упорядоченных множеств. Доказательство корректности предлагаемого метода

(Утверждения 4–6). Экспериментальное исследование алгоритма, подтверждающее его эффективность.

3. Быстрые логические классификаторы REC (для случая декартова произведения антицепей) и REC+ (для случая декартова произведения цепей), основанные на поиске правильных представительных элементарных классификаторов. Экспериментальное исследование качества и скорости работы классификаторов.

4. Асимптотические оценки типичного числа и типичного ранга правильных ЭК для случаев антицепей (Теорема 1) и цепей (Теоремы 2 и 3), позволяющие теоретически обоснованно выбирать параметры алгоритмов REC и REC+.

Степень достоверности полученных результатов обеспечивается доказательством сформулированных утверждений и теорем, а также результатами экспериментов, проведенных автором.

Апробация работы. Основные результаты работы доложены на конференциях «Математические методы распознавания образов (ММРО-19)» (Москва, 2019), «Информационные технологии и нанотехнологии (ИТНТ-2021)» (Самара, 2021), «Математические методы распознавания образов (ММРО-20)» (Москва, 2021), «Информационные технологии и нанотехнологии (ИТНТ-2022)» (Самара, 2022), «Интеллектуализация обработки информации (ИОИ-14)» (Москва, 2022), «Информационные технологии и нанотехнологии (ИТНТ-2025)» (Самарканд, 2025).

Результаты диссертационной работы включены в отчеты по проектам РФФИ 19-01-00430 «Логический анализ частично упорядоченных целочисленных данных в задачах классификации и поиска ассоциативных правил» и РНФ 24-21-00301 «Разработка новых методов корректной классификации и анализа данных на основе логического подхода».

Публикации. По теме диссертации опубликовано 11 научных работ, при этом две статьи имеют англоязычные версии. В журналах, рекомендованных

ВАК, опубликовано 5 статей, в изданиях, индексируемых в Web of Science Core Collection, – 2 статьи, в изданиях, индексируемых в Scopus, – 5 работ.

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

Соответствие паспорту специальности. Работа соответствует следующим пунктам паспорта научной специальности 1.2.3 «Теоретическая информатика, кибернетика»: п. 1 – «Теория информации»; п. 3 – «Теория сложности алгоритмов и вычислений»; п. 17 – «Анализ больших данных, обнаружение закономерностей в данных и их извлечение».

Работа состоит из введения, трех глав, заключения и списка литературы.

Список литературы включает 75 наименований. Общий объём диссертации составляет 106 страниц.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе предложен и исследован последовательно-совместный подход к перечислению максимальных частых и минимальных нечастых элементов декартова произведения частичных порядков, основанный на применении асимптотически оптимального алгоритма дуализации цепей RUNC-M+ (Е. В. Дюкова, Г. О. Масляков, П. А. Прокофьев, 2017). Корректность предложенного подхода базируется на Утверждениях 1-3.

Утверждение 1. Если $X \subset X_{\max}$, то $I(X^-)$ содержит хотя бы один частый элемент. Аналогично, если $Y \subset Y_{\min}$, то $I(Y^+)$ содержит хотя бы один нечастый элемент.

Утверждение 2. Если $X \subset X_{\max}$ и $y \in I(X^-)$ является нечастым элементом, то y – минимальный нечастый элемент. Аналогично, если $Y \subset Y_{\min}$ и $x \in I(Y^+)$ является частым элементом, то x – максимальный частый элемент.

Утверждение 3. Пусть $X \subseteq X_{\max}$, $Y \subseteq Y_{\min}$. Тогда $I(X^-) = Y$ (или $I(Y^+) = X$) тогда и только тогда, когда $X = X_{\max}$, $Y = Y_{\min}$.

Дадим краткое описание предлагаемого алгоритма. Положим $X_0 = \emptyset$. Строится последовательность $X_1 \subset X_2 \subset \dots \subset X_{max}$. На первом шаге $X_1 = \{x\}$, где x ищется простым обходом дерева решений в глубину. На шаге $i + 1$ ($i \geq 1$) решается задача дуализации множества $X_i \setminus X_{i-1}$. Пусть множество D есть результат дуализации $X_i \setminus X_{i-1}$. Согласно утверждению 1, множество D содержит частые элементы. Для каждого частого элемента из D находится один содержащий его максимальный частый элемент путем обхода дерева решений в глубину. Все найденные максимальные частые элементы, которых нет в множестве X_i , добавляются к X_i , и таким образом формируется X_{i+1} . Если же все найденные частые элементы уже содержатся в X_i , то решается задача дуализации множества X_i . Если в $I(X_i^-)$ нет частых элементов, то $I(X_i^-) = Y_{min}$, $X_i = X_{max}$ и алгоритм завершает работу. Иначе для каждого частого элемента из $I(X_i^-)$ находится один содержащий его максимальный частый элемент, который пополняет множество X_i , формируя в результате множество X_{i+1} .

Для случая данных, представленных в виде произведения конечных цепей, проведено экспериментальное сравнение последовательного, совместного и последовательно-совместного методов, а также независимого построения искомых множеств алгоритмом Apriori. Результаты показывают, что последовательно-совместный метод наиболее эффективен в случае, когда мощности множеств частых и нечастых элементов сопоставимы. При существенном различии мощностей этих множеств выигрывает последовательный метод. Наихудшие результаты продемонстрировал независимый подход на основе алгоритма Apriori, что подтверждает целесообразность применения алгоритмов дуализации.

Во второй главе показана связь между задачей совместного перечисления максимальных частых и минимальных нечастых элементов декартова произведения частичных порядков и задачей оптимальной расшифровки двузначной монотонной функции. На базе последовательно-совместного подхода к перечислению X_{max} и Y_{min} построен новый алгоритм расшифровки, нацеленный на «типичный случай». Экспериментально показано, что при

задании функции на E_k^n в случае больших значений k и n , предложенный алгоритм работает существенно быстрее классического алгоритма, ориентированного на «худший случай», и требует меньшего числа обращений к оператору B .

Установим связь между задачей совместного перечисления X_{max} и Y_{min} и задачей расшифровки функции f . Введем ряд необходимых определений. Ноль двужначной монотонной функции f называется *верхним*, если он не предшествует никакому другому нулю этой функции. Единица функции f называется *нижней*, если она не следует ни за какой другой единицей этой функции. Для расшифровки функции f достаточно построить множества всех ее верхних нулей и нижних единиц.

Пусть дана база данных $D(P)$ и порог $s \leq |D(P)|$, $s \in \mathbb{N}$. На множестве P определим двужначную монотонную функцию $f_{D,s}$, которая принимает значение 1 на частых элементах и 0 – на нечастых. Фактически, функция задаётся при помощи оператора B , который для произвольного $x \in P$ возвращает значение функции $f_{D,s}$ путем вычисления частоты встречаемости элемента x .

Нетрудно видеть, что нижние единицы (верхние нули) функции $f_{D,s}$ взаимно однозначно соответствуют минимальным нечастым (максимальным частым) элементам множества P . Таким образом, для решения задачи расшифровки произвольной двужначной монотонной логической функции может быть применен предложенный в работе последовательно-совместный алгоритм перечисления множеств X_{max} и Y_{min} . Корректность предложенного подхода базируется на Утверждениях 4-6.

Утверждение 4. Если $X \subset X_{up}$, то $I(X^-)$ содержит хотя бы один ноль функции $f_{D,s}$. Аналогично, если $Y \subset Y_{low}$, то $I(Y^+)$ содержит хотя бы одну единицу функции $f_{D,s}$.

Утверждение 5. Если $X \subset X_{up}$ и $y \in I(X^-)$ является единицей функции $f_{D,s}$, то y – нижняя единица функции $f_{D,s}$. Аналогично, если $Y \subset Y_{low}$ и $x \in I(Y^+)$ является нулем функции $f_{D,s}$, то x – верхний ноль функции $f_{D,s}$.

Утверждение 6. Пусть $X \subseteq X_{up}$, $Y \subseteq Y_{low}$. Тогда $I(X^-) = Y$ (или $I(Y^+) = X$) тогда и только тогда, когда $X = X_{up}$, $Y = Y_{low}$.

В ходе экспериментального исследования рассмотрен случай функции, заданной на E_k^n , и показана эффективность предлагаемого алгоритма: в случае больших значений k и n предложенный метод работает существенно быстрее ориентированного на «худший случай» алгоритма и требует меньшего числа обращений к оператору B . Экспериментально выявлены условия применимости каждого из подходов: традиционный алгоритм предпочтителен при небольших k и n (в этом случае алгоритм близок по сложности к оптимальному по Шеннону), тогда как последовательно-совместный алгоритм выигрывает при больших k и n .

В третьей главе предложены и реализованы алгоритмы логической классификации, основанные на поиске частых элементов в частично упорядоченных данных. Экспериментально показано, что предложенный алгоритм REC работает быстрее традиционных логических классификаторов в случае декартова произведения антицепей и не уступает им в качестве распознавания. Получены асимптотические оценки типичного числа и типичного ранга правильных представительных элементарных классификаторов, которые позволяют теоретически обоснованно выбирать параметры алгоритмов. Приведём указанные оценки.

Введем ряд необходимых понятий. Обозначим через L_K матрицу, строками которой являются признаковые описания прецедентов класса K . Квадратная подматрица матрицы L_K называется *правильной*, если она состоит из одинаковых строк. Правильные ЭК класса K порождаются правильными подматрицами матрицы L_K . Напомним, что ЭК ранга r называется правильным для класса K , если не менее чем r прецедентов класса K содержат этот ЭК.

Пусть m_1 – число прецедентов в $R(K)$; $M_{m_1 n}^k$ – множество всех матриц размера $m_1 \times n$ с элементами из $E_k = \{0, 1, \dots, k-1\}$; $S(L_K)$, $L_K \in M_{m_1 n}^k$, – множество правильных подматриц в матрице L_K ; ϕ_1 – интервал $(0, r_1)$, где $r_1 =$

$0.5 \log_k m_1 n - 0.5 \log_k \log_k m_1 n + \log_k \log_k \log_k n$; $b_n \sim c_n$, $n \rightarrow \infty$, означает, что $\lim_{n \rightarrow \infty} b_n / c_n = 1$.

Теорема 1. Если $n^a \leq m_1 \leq k^n$, $a > 1$, то для почти всех матриц L_K из $M_{m_1 n}^k$ справедливо

$$|S(L_K)| \sim \sum_{r \in \phi_1} C_n^r C_{m_1}^r k^{r-r^2} , \text{ при } n \rightarrow \infty ,$$

и порядки почти всех правильных подматриц из $S(L_K)$ принадлежат интервалу ϕ_1 .

Так как правильные представительные ЭК класса K порождаются правильными подматрицами из $S(L_K)$, то оценка типичного числа правильных подматриц является верхней асимптотической оценкой числа правильных представительных ЭК. Оценка типичного порядка правильной подматрицы свидетельствует о том, что в случае, когда число прецедентов существенно больше числа признаков, типичный ранг правильного представительного ЭК не превосходит r_1 .

В диссертационной работе разработана модификация алгоритма REC для работы с данными, представленными в виде совокупности элементов декартова произведения конечных цепей, названная REC+. Классификатор REC+ работает по схеме алгоритма REC. Эффективность предлагаемого алгоритма обоснована теоретически (Теоремы 2, 3) и экспериментально на реальных данных. Приводимые ниже в Теоремах 2 и 3 оценки получены в предположении, что признак x_j принимает значения из E_k , $k \geq 2$, $j = \overline{1, n}$, и элементы в E_k линейно упорядочены в порядке возрастания.

Пусть $\mathfrak{A}(L_K)$, где $L_K \in M_{m_1 n}^k$, – это множество всех правильных ЭК в $R(K)$; $d = k / (k - 1)$; $\sigma \in E_{k-1}^r$, $\sigma = (\sigma_1, \dots, \sigma_r)$; $Q_r(\sigma) = (\sigma_1 + 1)^r \times \dots \times (\sigma_r + 1)^r$; $r_2 =]\log_d m_1 + \log_d \log_d m_1[$, здесь $]q[$ – наименьшее целое, превосходящее q ; $r_3 =]0.5 \log_d m_1 n - 0.5 \log_d \log_d m_1 n + \log_d \log_d \log_d n [$; φ_2 – интервал $[1, r_2]$, φ_3 – интервал $[1, r_3]$; $b_n \lesssim c_n$, $n \rightarrow \infty$, означает, что $\lim_{n \rightarrow \infty} b_n / c_n \leq 1$.

Теорема 2. Если $n \leq m_1$, то для почти всех матриц L_K из $M_{m_1 n}^k$ верно

$$|\mathfrak{A}(L_K)| \lesssim \sum_{r \in \varphi_2} \sum_{\sigma \in E_{k-1}^r} Q_r(\sigma) C_n^r C_{m_1}^r k^{-r^2}, \text{ при } n \rightarrow \infty,$$

и ранги почти всех правильных ЭК из $\mathfrak{A}(L_K)$ принадлежат интервалу φ_2 .

Теорема 3. Если $m_1^\alpha \leq n \leq d^{m_1}$, $\alpha > 1$, то для почти всех матриц L_K из $M_{m_1 n}^k$ верно

$$|\mathfrak{A}(L_K)| \sim \sum_{r \in \varphi_3} \sum_{\sigma \in E_{k-1}^r} Q_r(\sigma) C_n^r C_{m_1}^r k^{-r^2}, \text{ при } n \rightarrow \infty,$$

и ранги почти всех правильных ЭК из $\mathfrak{A}(L_K)$ принадлежат интервалу φ_3 .

В Теоремах 2 и 3 получены оценки типичного числа и ранга правильных ЭК соответственно в случаях, когда число прецедентов превосходит число признаков, и, когда число признаков существенно больше числа прецедентов. В экспериментах ранг искомым правильных ЭК в алгоритме REC+ выбирается с использованием приведенных в Теоремах 2 и 3 оценок типичного ранга правильного ЭК.

Проведено экспериментальное сравнение алгоритма REC с алгоритмом голосования по тупиковым представительным ЭК из направления CVP. Показано, что REC работает существенно быстрее традиционного алгоритма из направления CVP и, как правило, превосходит его по качеству классификации. Сравнение с такими известными алгоритмами, как случайный лес и логистическая регрессия, показало, что REC в среднем не уступает по качеству указанным методам, а на ряде задач превосходит их.

Экспериментальное сравнение алгоритмов REC и REC+ на реальных задачах показало, что использование линейных порядков на множествах значений признаков позволяет повысить качество классификации. Полученное повышение качества достигается без привлечения дополнительной информации — линейные порядки строятся исключительно на основе анализа обучающей выборки.

ЗАКЛЮЧЕНИЕ

В настоящей диссертационной работе разработаны вычислительно эффективные методы логического анализа информации, представленной в виде декартова произведения конечных частично упорядоченных множеств. Сформулируем основные результаты, полученные в работе.

1. Предложен последовательно-совместный подход к перечислению множеств X_{max} максимальных частых и Y_{min} минимальных нечастых элементов декартова произведения частичных порядков. Доказана корректность предлагаемого метода (Утверждения 1–3). Экспериментально показано, что последовательно-совместный алгоритм наиболее эффективен в случае, когда мощности множеств частых и нечастых элементов примерно равны.

2. На базе последовательно-совместного подхода к поиску X_{max} и Y_{min} построен алгоритм асимптотически оптимальной расшифровки двузначной монотонной функции, заданной на декартовом произведении частично упорядоченных множеств. Доказана его корректность (Утверждения 4–6). Экспериментально показано, что в случае задания рассматриваемой функции на множестве E_k^n , предложенный алгоритм в среднем работает быстрее классического алгоритма (В. Б. Алексеев, 1976) и требует меньшего числа обращений к оракулу при больших n .

3. На основе поиска частых элементов в данных разработаны новые быстрые алгоритмы логической классификации, базирующиеся на перечислении корректных ЭК специального вида, названных правильными представительными. Для случая декартова произведения антицепей построен алгоритм REC перечисления правильных представительных ЭК. Для случая декартова произведения цепей разработана его модификация REC+. Получены асимптотические оценки типичного числа и типичного ранга правильных ЭК для случаев антицепей (Теорема 1) и цепей (Теоремы 2 и 3). Полученные оценки свидетельствуют о том, что число правильных представительных ЭК существенно меньше числа тупиковых представительных ЭК, поиск которых осуществляется в направлении CVP. Экспериментально подтверждено, что

алгоритм REC обучается существенно быстрее классической модели из CVP и не уступает ей по качеству классификации. Алгоритм REC+ превосходит алгоритм REC по качеству классификации на большинстве рассмотренных задач, что свидетельствует о целесообразности задания линейных порядков на множестве значений признаков.

Таким образом, в диссертационной работе разработаны вычислительно эффективные методы логического анализа частично упорядоченных данных. Указанные методы применены к задачам перечисления максимальных частых и минимальных нечастых элементов декартова произведения частичных порядков, расшифровки двузначных монотонных логических функций и классификации по прецедентам. Предложенные методы обоснованы теоретически и подтверждены экспериментально. Результаты работы опубликованы в 11 научных работах, из которых 5 статей опубликованы в журналах, рекомендованных ВАК, 2 статьи – в журналах, индексируемых в Web of Science Core Collection, и 5 работ – в изданиях, индексируемых в Scopus.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых научных журналах из перечня ВАК и в изданиях, индексируемых в Web of Science Core Collection и Scopus:

1. Драгунов, Н. А. О метрических (количественных) свойствах логических классификаторов / Н. А. Драгунов // Труды ИСА РАН. – 2024. – Т. 74, № 4. – С. 14–19.
2. Драгунов, Н. А. О поиске максимальных частых и минимальных нечастых наборов произведения частичных порядков / Н. А. Драгунов, Е. В. Дюкова // Информатика и её применения. – 2022. – Т. 16, вып. 1. – С. 82–87.
3. Драгунов, Н. А. Об одном подходе к расшифровке монотонной логической функции / Н. А. Драгунов, Е. В. Дюкова // Автоматика и телемеханика. – 2022. – Вып. 10. – С. 134–143.

Dragunov, N. A. One approach to monotone logical function decoding / N. A. Dragunov, E. V. Djukova // Automation and Remote Control. – 2022. – № 10. – P. 1600–1607.

4. Драгунов, Н. А. Правильные представительные элементарные классификаторы над произведением частичных порядков / Н. А. Драгунов, Е. В. Дюкова // Информатика и её применения. – 2025. – Т. 19, вып. 4. – С. 43–52.
5. Драгунов, Н. А. Логическая классификация на основе поиска правильных представительных элементарных классификаторов / Н. А. Драгунов, Е. В. Дюкова, А. П. Дюкова // Известия РАН. Теория и системы управления. – 2024. – № 4. – С. 86–92.

Dragunov, N. A. Logical classification based on finding regular representative elementary classifiers / N. A. Dragunov, E. V. Djukova, A. P. Djukova // J. Comput. Syst. Sci. Int. – 2024. – Vol. 63. – P. 634–641.

6. Dragunov, N. A. Supervised classification and finding frequent elements in data / N. A. Dragunov, E. V. Djukova, A. P. Djukova // Proc. 8th Int. Conf. Information Technology and Nanotechnology. – 2022. – P. 5.

Публикации в трудах международных и всероссийских конференций:

7. Драгунов, Н. А. Поиск минимальных нечастых и максимальных частых наборов в частично упорядоченных данных / Н. А. Драгунов, Е. В. Дюкова // Математические методы распознавания образов: тезисы докл. 19-й Всеросс. конф. с междунар. участием. – 2019. – С. 10–12.
8. Драгунов, Н. А. Асимптотически оптимальная расшифровка двужначной монотонной функции / Н. А. Драгунов, Е. В. Дюкова // Математические методы распознавания образов: тезисы докл. 20-й Всеросс. конф. с междунар. участием. – 2021. – С. 38–40.
9. Драгунов, Н. А. Поиск частых и нечастых элементов произведения частичных порядков и задача расшифровки двужначной монотонной функции / Н. А. Драгунов, Е. В. Дюкова // Информационные технологии и нанотехнологии: сб. тр. VII Междунар. конф. и молодёж. школы. – 2021. – С. 031852.
10. Драгунов, Н. А. Поиск частых элементов в данных и обучение по прецедентам / Н. А. Драгунов, Е. В. Дюкова // Интеллектуализация обработки информации: тезисы докл. 14-й Междунар. конф. – 2022. – С. 47–49.
11. Драгунов, Н. А. Классификация по прецедентам и поиск в данных частых элементов / Н. А. Драгунов, Е. В. Дюкова, А. П. Дюкова // Информационные технологии и нанотехнологии: сб. тр. VIII Междунар. конф. и молодёж. школы. – 2022. – С. 050712.