На правах рукописи

Катаева Алина Владимировна

ИЗВЛЕЧЕНИЕ И НЕИЗБЫТОЧНОЕ ПРЕДСТАВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ В МНОГОМЕРНЫХ ДАННЫХ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Красноярск 2019

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Сибирский федеральный университет», г. Красноярск

Научный руководитель: доктор физико-математических наук, доцент, **Быкова Валентина Владимировна**

Официальные оппоненты: **Ильев Виктор Петрович**, доктор физико-математических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «Омский государственный университет им. Ф. М. Достоевского», кафедра прикладной и вычислительной математики, профессор

Жилина Наталья Михайловна, доктор технических наук, доцент, «Новокузнецкий государственный институт усовершенствования врачей. Филиал федерального государственного бюджетного образовательного учреждения дополнительного профессионального образования «Российская медицинская академия непрерывного профессионального образования» Министерства здравоохранения Российской Федерации», кафедра медицинской кибернетики и информатики, профессор

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет»

Защита состоится «25» июня 2019 года в 16.00 часов на заседании диссертационного совета Д 212.099.22, созданного на базе Сибирского федерального университета по адресу: 660074, г. Красноярск, ул. Киренского, 26, ауд. УЛК 112.

С диссертацией можно ознакомиться в библиотеке и на сайте Сибирского федерального университета по адресу http://www.sfu-kras.ru

Автореферат разослан «__» мая 2019 г.

Ученый секретарь
диссертационного совета  Покидышева Людмила Ивановна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Современные методы интеллектуального анализа данных ориентированы на исследование многомерных и разнотипных данных с целью выявления знаний в виде закономерностей. Значительный вклад в развитие интеллектуального анализа данных внесли российские ученые: Ю.И. Журавлев (алгебраическая теория распознавания), Г.С. Лбов (логические методы распознавания), К.В. Рудаков (алгебраический синтез корректных алгоритмов), В.Н. Вапник, А.Я. Червоненкис (статистическая теория обучения), Н.Г. Загоруйко (когнитивный подход, FRiS-функции), С.О. Кузнецов, М.И. Забежайло (оценки сложности ДСМ-процедур) и др.

Средством описания причинно-следственных закономерностей в многомерных данных, представленных матрицей «объект–признак», служат ассоциативные правила, которые отражают, какие признаки, события или явления появляются вместе и насколько часто это происходит. Широкий интерес к этому классу закономерностей начался со статьи R. Agrawal, T. Imielinski, A. Swami, опубликованной в 1993 году, и с тех пор ежегодно появляются несколько сотен публикаций, содержащих новые методы и алгоритмы извлечения ассоциативных правил. Для многих приложений наиболее значимы строгие ассоциативные правила – ассоциативные правила с единичной достоверностью. Например, они широко востребованы при решении задач клинической диагностики. В национальном проекте «Электронное здравоохранение», утвержденном Президиумом Совета при Президенте Российской Федерации по стратегическому развитию и приоритетным проектам (протокол № 9 от 25.10.2016 г.) отмечается, что для повышения эффективности оказания медицинской помощи гражданам необходимо широкое внедрение в организации здравоохранения новейших лечебно-диагностических информационных технологий, базирующихся на интеллектуальном анализе данных.

В настоящее время практическое применение ассоциативных правил (АП) во многом ограничивается проблемой размерности. Число АП, извлекаемых современными методами анализа данных, часто достигает несколько десятков тысяч. Это существенно усложняет их интерпретацию и снижает степень доверия пользователя к полученным результатам. Для решения данной проблемы применяются два подхода: фильтрация с помощью мер значимости и когнитивный подход. Меры значимости позволяют численно оценивать достоверность и поддержку АП и предъявлять пользователю только те из них, для которых значения мер значимости превышают заданные пороговые значения. Когнитивный подход предполагает создание базисов как «сжатых» форм представления множества искомых АП. Между тем, оба подхода не исключают появление в результирующем множестве избыточных правил. Ассоциативное правило принято считать избыточным, если его удаление из множества выявленных правил не приводит к потере информации об ассоциациях между анализируемыми данными. Формальное определение избыточности предполагает уточнение, какая именно информация не должна быть утеряна. Для строгих АП такой информацией, прежде всего, служит уровень или порог поддержки – величина, характеризующая минимальную представительность этих правил в анализируемых данных.

Степень разработанности темы исследования. На сегодняшний день наиболее развиты методы формирования базисов строгих АП. В них под базисом понимается минимальное в некотором смысле множество строгих АП с заданным уровнем поддержки. Особого внимания заслуживают методы и алгоритмы построения канонического и минимаксного базисов, основанные на алгебраическом подходе, разработанном группой ученых под руководством Р. Вилле и известном в литературе как анализ формальных понятий. Канонический базис или базис Дюкена-Гига формируется из минимального числа строгих АП правил, рекуррентно описываемых в терминах псевдосодержаний. Этот базис достаточно полно изучен в работах B. Ganter, V. Duquenne, S. Rudolph, С.О. Кузнецова, С.А. Объедкова. Минимаксный базис создается из строгих АП, имеющих минимальную посылку и максимальное следствие. Именно такие АП интересны для клинической диагностики, поскольку каждое из них может определять минимальный набор диагностических признаков и максимальный набор последствий заболевания. Другой аргумент в пользу выбора минимаксного базиса для клинической диагностики – это наличие хорошо апробированных практикой алгоритмов его построения. В их числе различные версии алгоритма Close, представленные и изученные в работах M.J. Zaki, C.-J. Hsiao, T. Uno, T. Asai, Y. Uchida, H. Arimura.

Эксперименты показали, что канонические и минимаксные базисы могут содержать избыточность, устранение которой – это дополнительный шаг, позволяющий сокращать число строгих АП, предъявляемых пользователю для интерпретации. С этой целью представляет интерес использование выводимостей Армстронга. Известно, что строгие АП подчиняются шести выводимостям Армстронга, которые позволяют порождать из одних правил другие правила. Однако в общем случае выводимости Армстронга не гарантируют сохранение заданного уровня поддержки (далее кратко сохранение поддержки). Как отмечали в своих работах N. Pasquier, Y. Bastide, R. Taouil и L. Lakhal, именно этим ограничивалось применение выводимостей Амстронга для базисов строгих АП. Поэтому актуальны исследования выводимостей Амстронга с помощью анализа формальных понятий и выявление среди них тех, которые сохраняют поддержку АП, и с помощью которых можно устранять избыточность в минимаксном базисе при его построении, а далее при необходимости порождать из него строгие АП с сохранением поддержки.

Цель и задачи. Целью диссертационной работы является повышение эффективности анализа данных при решении задач клинической диагностики путем установления для строгих ассоциативных правил набора выводимостей, гарантирующих сохранение поддержки, и разработка на их основе математического и программного обеспечения.

Поставленная цель достигается путем решения следующих задач:

1. Установить свойства строгих ассоциативных правил и получить набор выводимостей, гарантирующих сохранение поддержки этих правил. Разработать и теоретически обосновать метод построения неизбыточного минимаксного базиса строгих ассоциативных правил.

2. Разработать алгоритм, реализующий метод построения неизбыточного минимаксного базиса строгих ассоциативных правил.

3. Сформировать набор средств снижения размерности матрицы «объект–признак», позволяющих уменьшать число искомых ассоциативных правил.

4. Разработать программное обеспечение, реализующее алгоритмы выявления строгих ассоциативных правил, построения неизбыточного минимаксного базиса, а также снижения размерности матрицы «объект–признак».

5. Провести экспериментальные исследования по оценке результативности разработанных метода, алгоритмов и программ на медицинских данных.

Научная новизна.

1. Разработан и теоретически обоснован новый метод построения неизбыточного минимаксного базиса строгих ассоциативных правил. В отличие от существующих метод позволяет устранять ту избыточность в минимаксном базисе, которые не способны удалять другие методы, сохраняя при этом поддержку строгих ассоциативных правил.

2. Разработан новый алгоритм извлечения строгих ассоциативных правил и представления их в форме неизбыточного минимаксного базиса. Алгоритм расширяет возможности известного алгоритма Close путем включения в него процедур по удалению из искомого множества зависимостей тех ассоциативных правил, которые распознаны как избыточные, без дополнительного обращения к анализируемому набору данных.

Методы исследования. Для решения поставленных в работе задач использовались методы анализа формальных понятий, статистические методы и методы объектно-ориентированного программирования.

Теоретическая значимость работы. Предложенный в работе метод построения неизбыточного минимаксного базиса быть использован для дальнейшего развития раздела интеллектуального анализа данных, связанного с извлечением закономерностей в данных и устранением избыточности в их представлении.

Практическая значимость работы. Применение результатов диссертационной работы в практическом здравоохранении позволяет повысить уровень информатизации клинической работы врачей, содействует верной и оперативной диагностике заболеваний. Результаты диссертационной работы могут быть также применены для тех приложений, где требуется высокая степень достоверности установленных ассоциативных правил и важна их «сжатая» форма представления, например, в информационной безопасности и анализе компьютерных сетей.

Положения, выносимые на защиту.

1. Доказательство выводимостей Армстронга с помощью анализа формальных понятий и установление среди них тех выводимостей, которые сохраняют поддержку строгих ассоциативных правил.

2. Метод построения неизбыточного минимаксного базиса строгих ассоциативных правил.

3. Алгоритм формирования неизбыточного минимаксного базиса строгих ассоциативных правил, устраняющего избыточность из минимаксного базиса в процессе его построения без дополнительного обращения к анализируемому набору данных.

Степень достоверности и апробация результатов работы. Достоверность результатов работы подтверждается строгими математическими доказательствами основных положений, а также численными экспериментами на реальных медицинских данных. Результаты диссертационных исследований докладывались и обсуждались на Республиканской научно-практической конференции «Статистика и ее применения» (Ташкент, 2017), Всероссийской конференции «Компьютерная безопасность и криптография» SIBECRIPT'17 (Красноярск, 2017), Региональной научно-практической конференции, посвященной 140-летию профессора В.Ф. Войно-Ясенецкого (Красноярск, 2017), XVII Международной конференции им. А.Ф. Терпугова «Информационные технологии и математическое моделирование» (Томск, 2018), Международной научно-практической конференции «Вопросы современных технических наук» (Екатеринбург, 2018), Международной конференции «X Сибирский конгресс женщин-математиков» (Красноярск, 2018), научных семинарах кафедры высшей и прикладной математики Сибирского федерального университета и кафедры медицинской кибернетики и информатики Красноярского государственного медицинского университета.

Результаты диссертационного исследования переданы в КГБУЗ «Красноярский краевой наркологический диспансер № 1», КГБУЗ «Краевая клиническая больница» для использования в научных исследованиях и клинической практике. Получены свидетельства о государственной регистрации программ для ЭВМ № 2018611317 от 01.02.2018, № 2018611886 от 08.02.2018.

Личный вклад автора в получении результатов, изложенных в диссертации. Основные результаты, составляющие новизну диссертационной работы, получены лично автором. Обсуждение метода, алгоритмов, результатов численных экспериментов и подготовка публикаций осуществлялись совместно с научным руководителем и соавторами опубликованных работ.

Публикации. По результатам диссертационных исследований опубликовано 12 печатных работ, из них 5 – в журналах, рекомендованных ВАК [1–5], 5 – в других изданиях [6–10], получено 2 свидетельства о государственной регистрации программ для ЭВМ [11, 12].

Структура и объем диссертации. Работа состоит из введения, четырех глав и заключения. Основной текст диссертации содержит 100 страниц, изложение иллюстрируется 13 рисунками и 10 таблицами. Библиографический список включает 134 источника.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, определены цель и задачи исследования, указаны научная новизна, практическая и теоретическая значимость выполненных исследований.

В первой главе рассматриваются особенности многомерных и разнотипных данных, характерных для клинической диагностики, исследуются различные виды зависимостей между данными и существующие методы их извлечения. Приводятся этапы лечебно-диагностического процесса, при реализации которых целесооб-

разно применение методов интеллектуального анализа данных, в том числе выявления ассоциативных правил и наборов диагностически значимых признаков.

Во второй главе содержатся основные результаты диссертационного исследования, связанные с доказательством выводимостей Армстронга методами анализа формальных понятий и установлением среди них тех выводимостей, которые сохраняют поддержку строгих ассоциативных правил, а также с разработкой метода и алгоритма построения неизбыточного минимаксного базиса. В ней решаются задачи 1 и 2 диссертационного исследования. Основные результаты второй главы опубликованы в работах [1, 3, 6, 7, 11].

Приведем основные определения и обозначения анализа формальных понятий, применяемые в диссертационной работе. Пусть для предметной области определены два непустых конечных множества G и M объектов и признаков соответственно. Пусть задано также непустое бинарное отношение $I \subseteq G \times M$. Тройку $K = (G, M, I)$ принято называть формальным контекстом (или просто контекстом) предметной области. Существование в I пары (g, m), $g \in G$ и $m \in M$, означает, что объект g имеет признак m и наоборот, признак m присущ объекту g. Если $A \subseteq G$ и $B \subseteq M$ – произвольные подмножества, то следующая пара операторов, называемых отображениями Галуа,

$$A' = \{m \in M: \forall g \in A \ \ (g, m) \in I\},$$
$$B' = \{g \in G: \forall m \in B \ \ (g, m) \in I\}$$

задает соответствие между частично упорядоченными множествами $(2^G, \subseteq)$ и $(2^M, \subseteq)$. Для отображений «$'$» характерны свойства антимонотонности и экстенсивности: для любых $B_1, B_2 \subseteq M$, если $B_1 \subseteq B_2$, то $(B_2)' \subseteq (B_1)'$; всегда $B_1 \subseteq (B_1)''$, где $(B_1)'' = ((B_1)')' \subseteq M$. Кроме того, $(B_1 \cup B_2)' = B_1' \cap B_2'$.

Двойное применение отображения «$'$» определяет оператор замыкания «$''$» на $(2^G, \subseteq)$ или $(2^M, \subseteq)$ в алгебраическом смысле. Принято говорить, что множество B'' является замыканием для $B \subseteq M$ в $K = (G, M, I)$. Множество B'' трактуется как набор признаков, которые всегда появляются в объектах контекста $K = (G, M, I)$ вместе с признаками из B, причем это множество является наибольшим по включению. Если $B = B''$, то множество признаков $B \subseteq M$ называется замкнутым относительно оператора «$''$» в контексте $K = (G, M, I)$.

Ассоциативным правилом на множестве признаков контекста $K = (G, M, I)$ называется упорядоченная пара множеств $r = (X, Y)$, $X, Y \subseteq M$. Традиционно ассоциативное правило $r = (X, Y)$ записывается в виде $X \Rightarrow Y$, а множества X и Y называются посылкой (или причиной) и заключением (или следствием) соответственно. Применительно к заданному контексту $K = (G, M, I)$ ассоциативное правило $X \Rightarrow Y$ количественно характеризуется двумя числовыми функциями: поддержкой $\delta(X \Rightarrow Y)$ и достоверностью $\gamma(X \Rightarrow Y)$. Эти функции выражаются в терминах анализа формальных понятий следующим образом.

Поддержка $\delta(X)$ множества $X \subseteq M$ в контексте $K = (G, M, I)$ определяет частоту встречаемости объектов, имеющих признаки X, и вычисляется как отношение числа $|X'|$ объектов, которым присущи признаки X, к общему числу $|G|$ объектов, представленных в этом контексте: $\delta(X) = |X'| / |G|$.

В силу антимонотонности отображения «′», поддержка множества признаков также удовлетворяет свойству антимонотонности: для любых X, $Y \subseteq M$ при $X \subseteq Y$ верно $\delta(Y) \le \delta(X)$. Для любого множества $X \subseteq M$ поддержка замыкания X'' всегда совпадает с поддержкой множества X: $\delta(X'') = \delta(X)$. Множество признаков $X \subseteq M$ называется частым в $K = (G, M, I)$, если его поддержка больше или равна заданному пороговому значению $\delta_0 \in [0, 1]$. Если одновременно $\delta(X) \ge \delta_0$ и $X = X''$, то X называется частым замкнутым множеством признаков в $K = (G, M, I)$. Частые множества и частые замкнутые множества признаков – основа существующих алгоритмов поиска ассоциативных правил в заданном контексте.

Поддержкой ассоциативного правила $X \Rightarrow Y$ в контексте $K = (G, M, I)$ называется величина

$$\delta(X \Rightarrow Y) = \delta(X \cup Y) = |(X \cup Y)'| / |G|, \qquad (1)$$

указывающая, какая доля объектов этого контекста представлена признаками $X \cup Y$. Достоверность $\gamma(X \Rightarrow Y)$ ассоциативного правила $X \Rightarrow Y$ – это отношение числа объектов, обладающих признаками $X \cup Y$, к числу объектов, которым свойственны только признаки X:

$$\gamma(X \Rightarrow Y) = |(X \cup Y)'| / \ |X'| = \delta(X \Rightarrow Y) / \delta(X) . \qquad (2)$$

Заметим, что достоверность определяется формулой (2) только для тех ассоциативных правил $X \Rightarrow Y$, для которых $\delta(X) \ne 0$. Если $\delta(X) = 0$, то согласно антимонотонности поддержки $\delta(X \cup Y) = 0$. В этом случае полагается $\gamma(X \Rightarrow Y) = 1$. Равенство $\gamma(X \Rightarrow Y) = 0$ возможно при $\delta(X) \ne 0$ и $\delta(X \cup Y) = 0$. Это соответствует ситуации, когда множество признаков X имеет в контексте $K = (G, M, I)$ ненулевую поддержку, однако X и Y одновременно не появляются в объектах данного контекста: $X' \ne \varnothing$, $(X \cup Y)' = X' \cap Y' = \varnothing$.

Ассоциативное правило $X \Rightarrow Y$ называется минимаксным в $K = (G, M, I)$, если для $K = (G, M, I)$ не существует другого ассоциативного правила $X^* \Rightarrow Y^*$ такого, что $X^* \subseteq X$, $Y \subseteq Y^*$ и верны равенства:

$$\delta(X^* \Rightarrow Y^*) = \delta(X \Rightarrow Y), \ \gamma(X^* \Rightarrow Y^*) = \gamma(X \Rightarrow Y).$$

Таким образом, минимаксное ассоциативное правило можно трактовать как присущую $K = (G, M, I)$ причинно-следственную закономерность с минимальной по включению посылкой X и максимальным по включению следствием Y.

Пусть заданы контекст $K = (G, M, I)$ и δ_0, γ_0 – вещественные числа из $[0, 1]$. Говорят, что $X \Rightarrow Y$ является (δ_0, γ_0)-ассоциативным правилом формального контекста $K = (G, M, I)$, если выполняются два условия:

$$\delta_0 \le \delta(X \Rightarrow Y) \le 1, \qquad (3)$$

$$\gamma_0 \le \gamma(X \Rightarrow Y) \le 1. \qquad (4)$$

Величины δ_0 и γ_0 играют роль пороговых значений для поддержки и достоверности соответственно. При $\delta_0 = 0$ условие (3) отражает естественные границы поддержки: $0 \le \delta(X) \le 1$. При $\gamma_0 = 1$ условие (4) вырождается в равенство $\gamma(X \Rightarrow Y) = 1$ и $(\delta_0, 1)$-ассоциативное правило называется строгим.

Задача извлечения ассоциативных правил формулируется так.

Заданы контекст $K = (G, M, I)$ и δ_0, γ_0 – вещественные числа из [0, 1].
Требуется найти для $K = (G, M, I)$ множество AR всех (δ_0, γ_0)-ассоциативных правил.

В общем случае при различных значениях δ_0 и γ_0 множество AR состоит из различных (δ_0, γ_0)-ассоциативных правил, отвечающих условиям (3) и (4). Известно, что в худшем случае число всех (δ_0, γ_0)-ассоциативных правил контекста $K = (G, M, I)$ экспоненциально зависит от $|M|$. При поиске правил это приводит к значительным вычислительным затратам и, что самое важное, затрудняет интерпретацию полученного множества AR. В настоящее время проблема размерности множества AR решается переходом к базисам этого множества. Заметим, что такой переход не изменяет сложность рассматриваемой задачи, ее перечислительный комбинаторный характер, а лишь позволяет представить результирующее множество AR в «сжатой» форме, удобной для интерпретации.

В главе 2 диссертации представлен новый метод построения неизбыточного минимаксного базиса строгих ассоциативных правил, опирающийся на доказанные в работе леммы 1–6 и теорему 7.

Рассмотрим строгое ассоциативное правило $X \Rightarrow Y$. В силу (1)–(2) его поддержка всегда совпадает с поддержкой посылки: $\delta(X \Rightarrow Y) = \delta(X)$. Если $\delta(X) \geq \delta_0$ в контексте $K = (G, M, I)$, то в этом контексте также $\delta(X \Rightarrow Y) \geq \delta_0$. Если результатом какого-либо преобразования строгого ассоциативного правила $X \Rightarrow Y$ является строгое ассоциативное правило с поддержкой не менее $\delta(X)$, то говорят, что такое преобразование сохраняет поддержку.

Лемма 1. Пусть в формальном контексте $K = (G, M, I)$ множество $X \subseteq M$ имеет поддержку $\delta(X) \geq \delta_0$. Тогда для контекста $K = (G, M, I)$ при любом $Y \subseteq X$ всегда справедливо строгое ассоциативное правило $X \Rightarrow Y$ с поддержкой $\delta(X) \geq \delta_0$.

Лемма 2. Если для контекста $K = (G, M, I)$ справедливо строгое ассоциативное правило $X \Rightarrow Y$ с поддержкой $\delta(X)$, то при любом $Z \subseteq M$ для этого контекста также справедливо строгое ассоциативное правило $X \cup Z \Rightarrow Y$ с поддержкой $\delta(X \cup Z) \leq \delta(X)$.

Лемма 2 отражает возможность пополнения посылки строгого ассоциативного правила, но без гарантии сохранения поддержки. Определен частный случай, когда расширение посылки сохраняет поддержку строгого ассоциативного правила: если для контекста $K = (G, M, I)$ верно строгое ассоциативное правило $X \Rightarrow Y$ с поддержкой $\delta(X)$, то при любом $Z \subseteq Y$ для этого контекста также справедливо строгое ассоциативное правило $X \cup Z \Rightarrow Y$ с поддержкой $\delta(X)$.

Лемма 3. Пусть в контексте $K = (G, M, I)$ множество $X \subseteq M$ имеет поддержку $\delta(X) \geq \delta_0$. Если для контекста $K = (G, M, I)$ справедливы строгие ассоциативные правила $X \Rightarrow Y$ и $X \Rightarrow Z$, то для этого контекста также справедливо строгое ассоциативное правило $X \Rightarrow Y \cup Z$ с поддержкой $\delta(X) \geq \delta_0$.

В данном случае $\delta(X \Rightarrow Y \cup Z) = \delta(X) \geq \delta_0$, т. е. свойство аддитивности строгих ассоциативных правил, заключающееся в объединении следствий при совпадении посылок и отраженное в лемме 3, гарантирует сохранение поддержки.

Лемма 4. Если для $K = (G, M, I)$ справедливо (δ_0, γ_0)-ассоциативное правило $X \Rightarrow Y$, то при любых $Z \subseteq Y$ и $Y \neq \varnothing$ для этого контекста также справедливо (δ_0, γ_0)-ассоциативное правило $X \Rightarrow Z$. В частном случае, при $\gamma(X \Rightarrow Y) = 1$ всегда верно $\gamma(X \Rightarrow Z) = 1$ и $\delta(X \Rightarrow Z) = \delta(X) \geq \delta_0$

Лемма 4 отражает свойство проективности ассоциативных правил: правую часть всякого (δ_0, γ_0)-ассоциативного правила можно «расщепить» до отдельного признака, сохраняя при этом поддержку и достоверность в заданных границах.

Для строгих ассоциативных правил леммы 3–4 констатируют равноценность различных форм их записи: представление $X \Rightarrow Y \cup Z$ эквивалентно представлению $X \Rightarrow Y$ и $X \Rightarrow Z$, при этом $\delta(X \Rightarrow Y \cup Z) = \delta(X \Rightarrow Y) = \delta(X \Rightarrow Z) = \delta(X)$.

Лемма 5. Если для контекста $K = (G, M, I)$ справедливы строгие ассоциативные правила $X \Rightarrow Y$ и $Y \Rightarrow W$ и $\delta(X) \geq \delta_0$, то какими бы ни были подмножества $X, Y, W \subseteq M$ для этого контекста также справедливо строгое ассоциативное правило $X \Rightarrow W$ с поддержкой $\delta(X) \geq \delta_0$.

Таким образом, согласно лемме 5 поддержка результирующего правила $X \Rightarrow W$ совпадает с поддержкой правила $X \Rightarrow Y$, играющего роль начала транзитивной цепочки строгих ассоциативных правил.

Лемма 6. Если для контекста $K = (G, M, I)$ справедливы строгие ассоциативные правила $X \Rightarrow Y$ и $Y \cup Z \Rightarrow W$, то какими бы ни были $X, Y, Z, W \subseteq M$ для контекста $K = (G, M, I)$ также справедливо строгое ассоциативное правило $X \cup Z \Rightarrow W$ с поддержкой $\delta(X \cup Z) \leq \delta(X)$.

Лемма 6 свидетельствует, что свойство псевдотранзитивности не гарантирует для результирующего ассоциативного правила сохранение поддержки. Для псевдотранзитивности всегда $\delta(X \cup Z \Rightarrow W) = \delta(X \cup Z)$, а согласно антимонотонности поддержки $\delta(X \cup Z) \leq \delta(X)$.

Леммы 1–6 позволяют сформулировать следующую теорему.

Теорема 7. Для любого контекста $K = (G, M, I)$ и произвольных $X, Y, Z, W \subseteq M$ справедливы следующие свойства строгих ассоциативных правил:

D_1. Рефлексивность: $X \Rightarrow Y, Y \subseteq X$.

D_2. Пополнение: если $X \Rightarrow Y$, то $X \cup Z \Rightarrow Y$.

D_3. Аддитивность: если $X \Rightarrow Y$ и $X \Rightarrow Z$, то $X \Rightarrow Y \cup Z$.

D_4. Проективность: если $X \Rightarrow Y$ и $Z \subseteq Y$, то $X \Rightarrow Z$.

D_5. Транзитивность: если $X \Rightarrow Y$ и $Y \Rightarrow W$, то $X \Rightarrow W$.

D_6. Псевдотранзитивность: если $X \Rightarrow Y$ и $Y \cup Z \Rightarrow W$, то $X \cup Z \Rightarrow W$.

Выводимости D_1, D_3, D_4, D_5 гарантируют сохранение поддержки, а их применение к $(\delta_0, 1)$-ассоциативным правилам неизменно приводит к $(\delta_0, 1)$-ассоциативным правилам.

Доказательства лемм 1–6 базируются на свойствах отображений Галуа, свойствах замкнутых множеств и следующем очевидном предложении: достоверность ассоциативного правила $X \Rightarrow Y$ относительно контекста $K = (G, M, I)$ равна единице тогда и только тогда, когда $X' \subseteq Y'$ (или $Y \subseteq X''$). Указанные в теореме 7 свойства строгих ассоциативных правил соответствуют выводимостям Армстронга. Доказательства того, что D_1, D_3, D_4, D_5 гарантируют сохранение поддержки, позволяют использовать данные выводимости при построении базисов для множества (δ_0, 1)-ассоциативных правил.

Пусть AR – множество всех (δ_0, 1)-ассоциативных правил формального контекста $K = (G, M, I)$. Строгое ассоциативное правило $X \Rightarrow Y$ выводится (или следует) из AR с сохранением уровня поддержки δ_0, если оно может быть получено из AR с помощью выводимостей D_1, D_3, D_4, D_5 и $\delta(X \Rightarrow Y) \geq \delta_0$. Этот факт будем обозначать $AR \models_{\delta_0} X \Rightarrow Y$. Строгое ассоциативное правило $X \Rightarrow Y$ избыточное в AR, если

$$AR \setminus \{X \Rightarrow Y\} \models_{\delta_0} X \Rightarrow Y. \qquad (5)$$

Множество (δ_0, 1)-ассоциативных правил неизбыточное, если оно не содержит избыточных строгих ассоциативных правил. Множество строгих ассоциативных правил назовем неизбыточным минимаксным базисом множества AR, если оно неизбыточное и состоит только из минимаксных строгих ассоциативных правил.

Суть предлагаемого метода: применение положений теоремы 7, касающихся выводимостей D_1–D_6 и специфики D_1, D_3, D_4, D_5, при построении минимаксного базиса (δ_0, 1)-ассоциативных правил для устранения избыточности в смысле (5) и формирование минимаксного базиса через генераторы частых замкнутых наборов признаков, аналогично тому, как это осуществляется в алгоритме Close. Кроме теоремы 7 теоретическим обоснованием предлагаемого метода является корректность алгоритма Close, доказанная M. Zaki и C. Hsiao.

Разработанный в диссертации алгоритм MClose – модификация алгоритма Close, направленная на удаление из результирующего множества AR избыточных (δ_0, 1)-ассоциативных правил. Причем распознавание избыточности осуществляется в процессе формирования AR. На вход алгоритма MClose подается исходный контекст $K = (G, M, I)$ и пороговое значение поддержки δ_0. Алгоритма MClose извлекает все (δ_0, 1)-ассоциативные правила исходного контекста и представляет их в форме неизбыточного минимаксного базиса.

В алгоритме MClose воспроизводятся основные действия алгоритма Close, направленные на пошаговое извлечение генераторов частых замкнутых наборов признаков и построение минимаксных строгих ассоциативных правил. Множество $\rho \subseteq M$ называется генератором замкнутого множества признаков $X \subseteq M$, $X = X''$, если $\rho'' = X$ и не существует другого множества $\tau \subseteq M$ такого, что $\tau \subset \rho$ и $\tau'' = X$. Если $|\rho| = k$, то ρ является k-генератором. Алгоритм MClose основывается также на равенстве $\delta(X'') = \delta(X)$ и свойстве сохранения поддержки (δ_0, 1)-ассоциативного правила при пополнении его посылки в отмеченном выше частном случае леммы 2.

Изначально искомое множество строгих ассоциативных правил *AR* считается пустым и $k = 1$. На первом шаге в качестве k-генераторов рассматриваются все одноэлементные подмножества множества *M*. Если $\delta(\rho_k'') \geq \delta_0$, то по ρ_k''строится минимаксное строгое ассоциативное правило $\rho_k \Rightarrow \rho_k'' \setminus \rho_k$ и добавляется в *AR*. Далее создаются кандидаты в $(k + 1)$-генераторы для следующей итерации. Каждый кандидат в $(k + 1)$-генераторы создается путем объединения двух k-генераторов, имеющих одинаковые первые $k - 1$ признаков. Проверяется, вложен ли найденный кандидат в ρ_k''. Если вложен, то он исключается из рассмотрения. После нахождения всех $(k + 1)$-генераторов осуществляется переход к следующей итерации. Алгоритм завершает работу, когда исчерпаны все генераторы (рисунок 1).

Алгоритм MClose

Вход: исходный контекст $K = (G, M, I)$, пороговое значение поддержки δ_0

```
1: begin
2:   AR ← ∅
3:   k ← 1
4:   while ρ_k ≠ ∅
5:       Gen-Closure (ρ_k)
6:       if δ(ρ_k) ≥ δ_0
7:           if ρ_k ≠ ρ_k''
8:               ρ_k^+ ← SX(ρ_k)
9:           end if
10:          if ρ_k^+ ≠ ρ_k''
11:              AR ← AR ∪ (ρ_k ⇒ ρ_k'' \ ρ_k)
12:          end if
13:      end if
14:      Gen-Generator (k + 1)
15:      k ← k + 1
16:  end while
17:  Non-Redundancy (AR)
18: end
```

Выход: *AR* – неизбыточный минимаксный базис $(\delta_0, 1)$-ассоциативных правил

Рисунок 1 – Пошаговое описание алгоритма MClose

В алгоритме MClose процедуры Gen-Closure и Gen-Generator выполняют вычисление замыканий и генераторов. Процедура SX осуществляет проверку условия (5). Процедура Non-Redundancy выполняет дополнительный просмотр результирующего множества *AR* с целью обнаружения оставшихся избыточных ассоциативных правил. Заметим, что оперативное удаление избыточных правил сдерживает рост мощности *AR* и не требует дополнительного обращения к анализируемому набору данных, представленному контекстом $K = (G, M, I)$. Все действия, выполняемые алгоритмом MClose и несвойственные алгоритму Close, выполняются за полиномиальное время относительно $|M|$.

Численные эксперименты, проведенные на реальных медицинских данных и представленные в главе 4 диссертации, показали, что алгоритм MClose по времени работы сопоставим с алгоритмом Close, при этом алгоритм MClose существенно уменьшает мощность минимаксного базиса, формируемого алгоритмом Close.

В третьей главе решается задача 3 диссертационного исследования. Основные результаты третьей главы опубликованы в работах [2, 4, 12].

В подразделе 3.1 формулируется задача FEATURES SELECTION (селекция признаков). Показано, что решение задачи FEATURES SELECTION путем конструирования обобщенных признаков (например, методами факторного анализа или методом экстремальной группировки признаков) приводит к трудностям интерпретации полученных результатов. Сделан вывод о том, что в медицинские аналитические системы клинической диагностики целесообразно включение методов, обладающих хорошей объяснительной способностью. Например, такими являются статистические методы Шеннона и Кульбака. Эти методы позволяют выполнить отбор признаков на основе заданной меры информативности. В подразделе 3.2 показано, что набор типичных представителей исходного множества объектов целесообразно формировать с помощью функции конкурентного сходства (FRiS-функции), введенной и изученной Н.Г. Загоруйко и его учениками Н.А. Борисовой, О.А. Кутненко, В.В. Дюбановым, Е.Н. Павловским и др.

В подразделе 3.3 описан алгоритм ELIMINATION, реализующий методы Шеннона и Кульбака, аппарат FRiS-функций, а также процедуры классификации и оценки качества классификации на основе ROC-анализа. Основное назначение ELIMINATION – снижение размерности матрицы «объект–признак» с целью уменьшения числа искомых ассоциативных правил, извлекаемых из этой матрицы. Другое назначение алгоритма ELIMINATION, вне зависимости от того, будет ли в дальнейшем осуществляться поиск ассоциативных зависимостей, – это получение новых медицинских знаний таких как, нахождение диагностически значимых признаков заболевания и типичных клинических случаев, а также решение задач клинический диагностики, сводимых к задачам классификации. В алгоритме ELIMINATION для классификации применяется известный метод ближайшего соседа, в котором решающим правилом является простое голосование.

В четвертой главе решаются задачи 4, 5 диссертационного исследования. Основные результаты четвертой главы опубликованы в работах [5, 8–12].

В подразделе 4.1 приведено описание разработанного комплекса программ, в котором реализованы разработанные в диссертации метод и алгоритмы выявления строгих ассоциативных правил и их «сжатого» представления в виде неизбыточного минимаксного базиса, снижения размерности матрицы «объект–признак». Комплекс программ создан для проведения экспериментальных исследований на реальных медицинских данных с целью оценки результативности предложенных средств.

Комплекс программ имеет модульную структуру и не привязан к каким-либо конкретным базам медицинских данных (таблица 1). Он может служить программной основой для создания медицинских аналитических систем клинической диагностики. Для расширения разработанного комплекса программ до самостоятельной медицинской аналитической системы необходимо создание информационной составляющей, ориентированной на конкретный лечебно-диагностический процесс и определенные нозологические формы заболеваний, с включением в нее базы данных пациентов и справочной медицинской информации.

Таблица 1 – Состав и функции программных модулей

Название модуля	Идентификатор модуля	Выполняемые функции
Модуль ввода исходных данных	INPUT	Осуществляет ввод признаковых описаний пациентов, добавление и удаление информации о пациентах в базу данных
Модуль предобработки исходных данных	EDIT_DATA	Осуществляет предобработку данных: шкалирование признаков, фильтрацию данных по пациентам и признакам, добавление и удаление признаков, выявление и удаление дубликатов строк или столбцов матрицы «объект–признак» и др.
Модуль формирования множества информативных признаков и типичных представителей заданных классов объектов	ELIMINATION	Производит расчет информативности признаков по выбранному методу (Шеннона или Кульбака). Выполняет классификацию признакового описания пациента по целевому признаку и отбор признаков с помощью показателей ROC-анализа. Находит набор типичных представителей классов объектов с помощью FRiS-функции
Модуль извлечения ассоциативных правил и построения неизбыточного минимаксного базиса	ASSOCIATION_RULES	Выполняет поиск ассоциативных правил с помощью выбранных алгоритмов (Apriori, Close или MClose). Строит неизбыточный минимаксный базис строгих ассоциативных правил с помощью алгоритма MClose. Позволяет осуществлять экспертную группировку исходных признаков
Модуль визуализации и интерпретации полученных результатов	OUTPUT	Выводит результирующие данные в стандартные форматы и визуальные формы
Головной модуль		Осуществляет взаимодействие модулей и реализует пользовательский интерфейс

В подразделе 4.2 диссертационной работы приведены результаты анализа диагностики наркозависимости с применением ассоциативных правил. Использовалась база медицинских данных Красноярского краевого наркологического диспансера (222 пациента, 64 признака).

Диагностика наркотического опьянения предполагает оценку психического состояния пациента, его соматовегетативных и неврологических признаков, и определения препаратов, возможно принятых пациентом. Существуют модели состояния пациента, обусловленные употреблением отдельных наркотических веществ или их комбинаций, по которым еще до проведения лабораторных исследований врач может определить, что именно принял пациент. Однако иногда реальное состояние пациента не совпадает с имеющимися моделями, поскольку оно вызвано принятием неизвестного препарата или ранее не встречающейся комбинации препаратов. В этой ситуации врачу-наркологу могут быть полезны строгие ассоциативные зависимости, выявленные из медицинских данных пациентов, проходивших ранее лечение в наркологическом диспансере.

Строгие ассоциативные правила, описывающие зависимости между принимаемыми препаратами и возможными последствиями от их употребления, определяют, чаще всего, известные модели состояния пациента. Зависимости между наблюдаемыми симптомами и возможными наборами препаратов позволяют выявлять ранее известные и новые комбинации психоактивных веществ. Уменьшение числа строгих ассоциативных правил, предъявляемых врачу, позволяет оперативно проводить лечение пациентов.

При выполнении численных экспериментов сравнивались алгоритмы Close и MClose по размеру построенных минимаксных базисов и времени работы. При сравнении использовался также известный алгоритм Apriori, который работает с частыми множествами и не использует базисы ассоциативных правил. Эксперименты выполнялись на компьютере с процессором Intel®Core™ i7-720QM Processor (6M Cache, 1.60 GHz) и ОЗУ размером 4 ГБ. В таблице 2 представлены результаты численных экспериментов, полученные при $\delta_0 = 0{,}1$.

Таблица 2 – Сравнение алгоритмов на медицинских данных по наркозависимости

Количество пациентов	Apriori		Close		MClose	
	Число извлеченных АП	Время вычислений, мс	Размер минимаксного базиса	Время вычислений, мс	Размер неизбыточного минимаксного базиса	Время вычислений, мс
50	2583	10560	32	4803	5	4034
100	2049	8700	28	3264	4	3056
150	553	3920	46	1176	8	1075
222	122	1154	**17**	867	**5**	695

Из результатов численных экспериментов следует, что в сравнении с алгоритмом Apriori, формирующие базисы алгоритмы Close и MClose на несколько порядков уменьшают число строгих ассоциативных правил, предъявляемых врачу для интерпретации. При этом алгоритм MClose в несколько раз уменьшает мощность минимаксного базиса, формируемого алгоритмом Close. Время работы алгоритма MClose сопоставимо со временем работы алгоритма Close.

Полученный в результате экспериментов неизбыточный минимаксный базис для объектно-признаковых описаний 222 наркозависимых содержит пять строгих ассоциативных правил (таблица 3). По мнению врачей-наркологов, он отражает известные модели состояния пациента, а именно часто встречающиеся зависимости между принимаемыми препаратами и возможными последствиями от их применения.

Таблица 3 – Неизбыточный минимаксный базис при $\delta_0 = 0{,}1$

Посылка ассоциативного правила		Следствие ассоциативного правила	Поддержка ассоциативного правила
{Этанол}	$\Rightarrow$	{Страх, Угроза жизни}	0,316
{Галоперидол}	$\Rightarrow$	{Возбуждение}	0,226
{MDPV, ТГК}	$\Rightarrow$	{Попытки подбросить наркотики}	0,181
{Пирровалерон, ТГК}	$\Rightarrow$	{Попытки подбросить наркотики}	0,181
{Пол ж, ТГК}	$\Rightarrow$	{Тревога}	0,181

Для установления предполагаемого набора принятых препаратов врач-нарколог задает набор показателей состояния пациента, полученных по результатам дифференциальной диагностики. Этот набор рассматривается программой в качестве причины, для которой формируются возможные комбинации принятых препаратов и предоставляются врачу-наркологу для анализа и принятия врачебного решения.

В подразделе 4.3 диссертационной работы представлены результаты численных экспериментов по оценке результативности алгоритмов снижения размерности матрицы «объект–признак» на двух задачах клинической диагностики: определение минимального набора признаков для распознавания множественной лекарственной устойчивости (МЛУ) возбудителя туберкулеза легких; нахождение диагностически значимых признаков для выявления сепсиса. При анализе использовались базы медицинских данных Красноярского краевого противотуберкулезного диспансера (779 пациентов, 26 признаков) и гнойно-септического центра Краевой клинической больницы (200 пациентов, 16 признаков).

Для выявления у пациента МЛУ стандартными диагностическими средствами затрачивается от 20 до 90 дней с момента выявления заболевания, что ведет к снижению эффективности лечения на начальном этапе. Сепсис – это нарушение функций органов человека, вызванное реакцией организма на инфекцию. Поскольку клиническое течение сепсиса протекает стремительно, от врача требуется оперативность в постановке диагноза. В результате экспериментов были выявлены 6 (из 26) наиболее информативных признаков распознавания МЛУ и 10 (из 16) диагностически значимых признаков распознавания сепсиса. По мнению врачей, полученные результаты не противоречат клинической практике и способствуют своевременному и правильному лечению туберкулеза легких и сепсиса.

В заключении сформулированы основные результаты и выводы, полученные в диссертационной работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Установлены свойства строгих ассоциативных правил и получен набор выводимостей D_1, D_3, D_4, D_5, гарантирующих сохранение поддержки (леммы 1–6, теорема 7). На их основе разработан и теоретически обоснован метод построения неизбыточного минимаксного базиса строгих ассоциативных правил.

2. Разработан алгоритм MClose формирования неизбыточного минимаксного базиса строгих ассоциативных правил, расширяющий возможности известного алгоритма Close путем включения в него процедур по удалению из искомого множества зависимостей тех ассоциативных правил, которые распознаны как избыточные, без дополнительного обращения к анализируемому набору данных. Численные эксперименты показали, что алгоритм MClose по времени работы сопоставим с алгоритмом Close, при этом алгоритм MClose существенно уменьшает мощность минимаксного базиса, формируемого алгоритмом Close.

3. Сформирован набор средств снижения размерности матрицы «объект–признак», позволяющий уменьшить число искомых ассоциативных правил и обладающих хорошей объяснительной способностью для практикующих врачей. К ним отнесены статистические методы Шеннона и Кульбака и FRiS-функции.

4. Разработан комплекс программ, реализующий алгоритмы извлечения строгих ассоциативных правил и их «сжатого» представления в виде неизбыточного минимаксного базиса, снижения размерности матрицы «объект–признак». Разработанная версия комплекса программ не привязана к каким-либо конкретным базам медицинских данных и может служить программной основой при создании медицинских аналитических систем клинической диагностики, ориентированных на конкретные нозологические формы заболеваний.

5. Выполнены численные эксперименты на реальных базах медицинских данных (по наркозависимости, множественной лекарственной устойчивости возбудителя туберкулеза легких, сепсису). Результаты экспериментов показали высокую результативность разработанных метода, алгоритмов и программ.

Применение результатов выполненного диссертационного исследования в практическом здравоохранении способствует повышению эффективности анализа данных при решении задач клинической диагностики.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

В изданиях, рекомендованных ВАК:

1. Быкова В.В., **Катаева А.В.** О неизбыточном представлении минимаксного базиса строгих ассоциативных правил // Прикладная дискретная математика. – 2016. – № 2 (36). – С. 113-126 (индексируется **Scopus)**.

2. Быкова В.В., **Катаева А.В.** Методы и средства анализа информативности признаков при обработке медицинских данных // Программные продукты и системы. – 2016. – № 2 (114). – С. 172-178.

3. Быкова В.В., **Катаева А.В.** Сжатое представление строгих ассоциативных правил в анализе данных // Программные продукты и системы. – 2017. – № 2 (30). – С. 187-192.

4. Наркевич А.Н., Виноградов К.А., Быкова В.В., **Катаева А.В.** Сокращение признакового пространства в анализе множественной лекарственной устойчивости возбудителя у больных туберкулезом легких // Врач и информационные технологии. – 2018. – № 2. – С. 48-57.

5. Виноградов К.А., Наркевич А.Н., **Катаева А.В.**, Пичугина Ю.А., Афанасьева Н.А. Средства интеллектуальной поддержки принятия решений в диагностике и лечении наркозависимых // Врач и информационные технологии. – 2018. – № 4. – С. 20-26.

В других изданиях и материалах конференций:

6. Быкова В.В., **Катаева А.В.** Алгоритм построения неизбыточного минимаксного базиса строгих ассоциативных правил // Прикладная дискретная математика. Приложение (труды Всероссийской конференции «Компьютерная безопасность и криптография» SIBECRIPT’17). – 2017. – № 10. – С. 154-157.

7. **Катаева А.В.** Минимизация базиса строгих ассоциативных правил на основе замкнутых множеств // Материалы республиканской научно-практической конференции «Статистика и ее применения». Ташкент. – 2017. – С. 77-83.

8. **Катаева А.В.** Интеллектуальная поддержка принятия решений в диагностике и лечении наркозависимых // Материалы XVII Международной конференции им. А.Ф. Терпугова «Информационные технологии и математическое моделирование» (ИТММ-2018). – Томск: Изд-во НТЛ, 2018. – С. 185-192.

9. Афанасьева Н.А., Березовская М.А., Коробицина Т.В., Пичугина Ю.А., Арапиев Ю.А., Виноградов К.А., Быкова В.В., **Катаева А.В.** Опыт применения метода математического моделирования психотических расстройств при сочетанном употреблении современных синтетических психоактивных веществ // Сибирский вестник психиатрии и наркологии. – 2018.– № 4 (101). – С. 28-34.

10. **Катаева А.В.**, Бахтина Ж.А. Применение телемедицинских технологий для диагностики и мониторинга сепсиса // Материалы Международной научно-практической конференции «Вопросы современных технических наук: свежий взгляд и новые решения». – Екатеринбург: ННИЦРОН, 2018. – № 5. – С. 10-13.

Свидетельства о государственной регистрации программы для ЭВМ:

11. **Катаева А.В.** Выявление ассоциативных правил и построение неизбыточного минимаксного базиса. Свидетельство о государственной регистрации программы для ЭВМ № 2018611317. Зарегистрировано в Реестре программ для ЭВМ 01.02.2018.

12. **Катаева А.В.** Программа сокращения признакового пространства на основе алгоритмов классификации и ROC – анализа. Свидетельство о государственной регистрации программы для ЭВМ № 2018611886. Зарегистрировано в Реестре программ для ЭВМ 08.02.2018.