Федеральное государственное автономное образовательное
учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

На правах рукописи

Катаева Алина Владимировна

ИЗВЛЕЧЕНИЕ И НЕИЗБЫТОЧНОЕ ПРЕДСТАВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ В МНОГОМЕРНЫХ ДАННЫХ

Специальность 05.13.17 – Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д-р физ.-мат. наук, доцент
Быкова Валентина Владимировна

Красноярск 2019

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

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, отражающие, 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 – 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 [73].

В настоящее время практическое применение ассоциативных правил (АП) во многом ограничивается проблемой размерности [2, 7, 17, 22]. 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 установленные 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 [95, 116, 117].

Канонический базис (базис Дюкена-Гига) создается из минимального числа строгих ассоциативных правил, рекуррентно задаваемых в терминах псевдосодержаний. Этот базис достаточно полно изучен в работах B. Ganter, V. Duquenne, S. Rudolph, С.О. Кузнецова, С.А. Объедкова [95, 107–109, 112, 116, 120, 128].

Минимаксный базис формируется из строгих АП, имеющих минимальную посылку и максимальное следствие. Именно такие АП интересны для клинической диагностики, поскольку каждое из них может определять минимальный набор симптомов заболевания и максимальный набор признаков, задающих его последствия. 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 [93, 114, 127, 132].

Вычислительные 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 [83]. Известно, что строгие АП подчиняются шести выводимостям Армстронга, которые позволяют порождать из одних правил другие правила [60]. Однако в общем случае выводимости Армстронга не гарантируют сохранение заданного уровня поддержки (далее кратко сохранение поддержки). Как отмечали в своих работах J.L. Balcazar, N. Pasquier, Y. Bastide, R. Taouil и L. Lakhal, именно этим ограничивалось применение выводимостей Амстронга для базисов строгих АП [87,115]. 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 [11, 13, 14, 18, 19], 5 – в других изданиях [5, 12, 48–50], получено 2 свидетельства о государственной регистрации программ для ЭВМ [46, 47].

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

В первой главе диссертационной работы рассматриваются особенности многомерных и разнотипных данных, характерных для клинической диагностики. Исследуются различные виды зависимостей между данными и существующие методы их извлечения.

Во второй главе диссертационной работы 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 третьей главе представлен алгоритм ELIMINATION, предназначенный для снижения размерности матрицы «объект–признак» и реализующий методы Шеннона и Кульбака, аппарат FRiS-функций, а также процедуры классификации и оценки качества классификации на основе ROC-анализа.

В четвертой главе диссертационной работы описывается комплекс программ, в котором реализованы разработанные в диссертации метод и алгоритм выявления строгих ассоциативных правил и их «сжатого» представления в виде неизбыточного минимаксного базиса, а также алгоритм ELIMINATION. Приведены результаты численных экспериментов на реальных базах медицинских данных (по наркозависимости, множественной лекарственной устойчивости возбудителя туберкулеза легких, сепсису).

Глава 1 Зависимости между данными как основа повышения эффективности клинической диагностики

Данная глава носит вводный характер. В ней рассматриваются особенности многомерных и разнотипных данных, характерных для клинической диагностики. Исследуются различные виды зависимостей между данными и существующие методы их извлечения. Приводятся основные этапы лечебно-диагностического процесса, при реализации которых целесообразно применение методов интеллектуального анализа данных, в том числе методов выявления ассоциативных правил и наборов диагностически значимых признаков.

1.1 Методы интеллектуального анализа данных в медицинских аналитических системах клинической диагностики

Принятие врачебных решений традиционно производится на основе знаний и опыта врача, а также шаблонных моделей заболеваний и методик их лечения [7, 26, 27, 29, 32]. Наиболее важными решениями, принимаемыми врачом, являются:

- постановка диагноза на основе имеющихся данных о пациенте (социально-демографических, диагностических и др.);
- выбор эффективного метода лечения на основе первичных данных и информации о проводимой терапии, а также с учетом противопоказаний и индивидуальных особенностей пациента;
- прогнозирование эффективности проводимого лечения и возможном осложнении здоровья пациента.

Постановка конкретного диагноза, как правило, сводится к задачам классификации или кластеризации данных и определение состояния здоровья пациента на основе известных классов (моделей) того или иного заболевания

[3, 4, 33, 81, 88, 91, 103]. Выбор методики лечения и прогнозирование возможных осложнений также базируется на шаблонных методиках с учетом особенностей пациента.

Наиболее известными зарубежными и отечественными программными платформами и системами, предназначенными для поддержки лечебно-диагностических процессов и использующими методы интеллектуального анализа данных, являются IndiGO-Archimedes, Auminence, IBM Watson for Oncology, Botkin.AI, OncoFinder, CoBrain-Аналитика и другие.

IndiGO-Archimedes обрабатывает данные медицинских карт с помощью алгоритмов машинного обучения и формирует индивидуальные протоколы диагностики и лечения пациентов с учетом истории болезни. Прогнозирует риск сердечных приступов, диабетических кризов и т. д. [92].

Система Auminence разработана специалистами Кэмбриджа [33, 122]. Данная система дифференциальной диагностики, анализирует сведения о симптомах, выявляет закономерности и формирует диагностический план (checklist).

Система Watson for Oncology фирмы IBM разработана на основе когнитивных компьютерных технологий для постановки диагноза и выбора программы лечения онкологических заболеваний [105, 133]. Она производит анализ как структурированных данных, так и текстов на естественном языке. С 2013 года система используется в Мемориальном онкологическом центре им. Слоуна-Кеттеринга в Нью-Йорке.

Среди отечественных разработок необходимо отметить систему Botkin.AI для распознавания патологических проявлений в рентгенологических снимках, компьютерной томографии, маммограммах [7, 27, 33, 53]. Данная система разработана компанией ООО «Интеллоджик» г. Москва. Она обучена с помощью нейронных сетей диагностике онкологических заболеваний.

Система OncoFinder является совместной разработкой Первого Онкологического Научно-Консультационного Центра (Россия), Федерального на-

учно-клинического центра детской гематологии, онкологии и иммунологии им. Д. Рогачева (Россия), университета Летбридж (Канада), Калифорнийского Технологического Института (США) [27, 35, 86, 121, 124]. Система выполняет анализ внутриклеточных сигнальных путей и подбирает наилучшие терапевтические препараты при различных типах рака.

Информационно-аналитическая система Cobrain-Аналитика, разработчики Сколковский институт науки и технологий и Национальный медицинский исследовательский центр профилактической медицины (Россия), представляет собой централизованное хранилище клинических и параклинических данных пациентов с заболеваниями мозга [27, 35]. Система предназначена для диагностики и оценки состояния больного, а также помогает в последующей коррекции лечебных мероприятий.

Все рассмотренные выше программные средства (платформы и системы) используют методы интеллектуального анализа данных, в том числе методы извлечения закономерностей. Опыт практического применения подтверждает их результативность в аспекте повышения уровня информатизации клинической работы врачей. Однако относительно отечественных средств многими экспертами были выявлены факторы, сдерживающие их использование в практическом здравоохранении [26]. Так, из-за низкого качества данных о пациентах точность выявленных при интеллектуальном анализе данных закономерностей или паттернов (моделей, шаблонов, зависимостей) обычно не достаточно высока, чтобы использовать их в клинических условиях. Выдаваемые алгоритмами результаты зачастую являются плохо интерпретируемыми для врача. Основными характерными особенностями приведенных выше систем являются: узкая направленность (каждая из них предназначена для определенных нозологических форм заболеваний); локальность (большинство из них никак не интегрированы в информационные системы лечебно-профилактических учреждений); трудоемкость эксплуатации (требуют временных затрат и нагрузки на врача для внесения первичной информации о пациенте) [99, 130].

Существующие универсальные программные средства интеллектуального анализа данных, такие как WEKA, RapidMiner, SQL Server Data Mining, R, SPSS, STATISTICA требуют специальной подготовки пользователя и не достаточно учитывают специфику медицинских данных [2–4, 8, 15, 29, 129]. Перенос зарубежных платформ в отечественное практическое здравоохранение в полном объеме также невозможен в виду большой разницы в организации здравоохранения и государственной политики по импортозамещению.

Функционально медицинские аналитические системы клинической диагностики (МАСКД) ориентируются исключительно на информатизацию клинической работы врачей и опираются на конкретный электронный клинико-инструментальный «образ» пациента. Данный вид систем является принципиально отличным от других медицинских программ, носящих обучающий или справочный характер.

К настоящему времени медицинским сообществом сформулированы требования к вновь разрабатываемым МАСКД [17, 27, 35, 57, 58, 67]:

- максимально удобный и информативный пользовательский интерфейс, возможность внесения экспертных знаний в систему со стороны врача;
- инструменты для ввода данных должны быть просты в использовании, простые требования к представлению данных;
- возможность предобработки данных с участием врача-эксперта;
- возможность импорта данных из различных источников;
- возможность вывода зависимостей в виде, допускающем экспертный анализ (верификацию и интерпретацию) полученного результата;
- возможность экспорта результатов в различные форматы;
- возможность встраивания в существующие программные решения.

Важно отметить, что в МАСКД не должно быть функций по постановке диагноза и назначения лечения. Система должна лишь осуществлять информационную поддержку врача при принятии врачебных решений, но окончательное решение всегда остается за врачом.

Преимущество использования МАСКД заключается в возможности выявления закономерностей между данными, с которыми эксперт не встречался ранее, применения на практике методов интеллектуальной обработки медицинских данных для повышения эффективности клинической диагностики. Специфика подобных приложений проявляется в специфике данных, на основе которых принимаются решения.

1.2 Специфика медицинских данных

Залогом успешного применения методов и алгоритмов интеллектуального анализа данных в клинической диагностике является наличие информации, пригодной для извлечения из нее закономерностей. Анализ медицинских данных часто осложняется большим объемом, неоднородностью и сложной структурой этих данных [8, 65, 76].

Современные медицинские информационные системы аккумулируют большие объемы разнородной информации о пациентах, включая социально-демографические сведения и клинические данные, представляемые в виде электронных медицинских карт пациентов (ЭМК). Содержание ЭМК определено Министерством здравоохранения РФ (11 ноября 2013 г. № 18-1/1010) и представляет собой некоторую формализованную структуру данных, задающую признаковое описание пациента.

Выбор показателей (признаков), входящих в ЭМК, их сочетание определяется для каждого случая соответствующей нозологической формой или патологическим синдромом. Признаковые описания пациентов могут включать в себя значения несколько десятков и сотен разнотипных признаков. Такие описания служат основой формирования матриц «объект–признак», которые являются традиционным представлением информации в интеллектуальном анализе данных. Строки матрицы «объект–признак» – это признаковое описания объектов (пациентов), а столбцы этой матрицы соответствуют определенным признакам. В роли признаков выступают социально-

демографические показатели и показатели состояния здоровья пациента, которые могут быть различными по типу. Система признаков исследуемого множества пациентов рассматривается в качестве признакового пространства, а совокупность значений признаков отдельного пациента определяет его признаковое описание [2, 12, 41, 45, 47, 82].

Различают несколько типов признаков. Количественные признаки – это признаки, значения которых можно измерить в некоторой числовой шкале. Например, вес пациента. Качественные признаки измеряются в некоторых порядковых шкалах и употребляются для показателей, не имеющих числового выражения. Например, степень тяжести заболевания. Номинальные признаки определяются шкалой наименований. Например, группа крови пациента. Как правило, при анализе номинального признака каждое его отдельное значение рассматривают в качестве отдельного признака, принимающего значения 1 («да») или 0 («нет»).

Подавляющее большинство методов анализа ориентировано на числовые данные. Присвоение числовых значений качественным и номинальным признакам в анализе данных принято называть шкалированием [21, 43, 55, 62, 68, 80]. После шкалирования к качественным и номинальным признакам возможно применение различных методов численного анализа, включая статистические методы.

Существуют разные способы шкалирования [68, 69, 76, 79]. В общем случае шкалирование признака (качественного или номинального) возможно двумя способами:

- автоматический. Каждому уникальному значению признака присваивается уникальное количественное значение. Например, при шкалировании признака «пол» значению «женский» присваивается «0», а значению «мужской» ставится в соответствие «1»;
- ручное шкалирование, выполняемое экспертом. Все допустимые значения признака анализируются экспертом вручную, и каждой близкой группе значений присваивается уникальное значение. Например, при

шкалировании признака «обстоятельства травмы» значениям «ДТП» и «автодорожное происшествие» ставится в соответствие одно количественное значение «1».

При шкалировании медицинских данных роль эксперта особенно важна, так как эти данные зачастую содержат синонимы, ошибки и очень близкие значения, которые невозможно определить автоматически.

В некоторых случаях, например, при использовании методов извлечения ассоциативных правил, матрица «объект–признак» должна содержать только значения 1 или 0. Если элемент матрицы равен 1, то это интерпретируется как наличие у пациента соответствующего признака. Если элемент матрицы равен 0, то это трактуется как отсутствие этого признака. Для приведения номинального признака к бинарному типу его разделяют на систему бинарных признаков. Очевидно, что такое преобразование значительно увеличивает количество анализируемых признаков и целесообразно применять методы снижения признакового пространства, позволяющие производить анализ данных без потерь интересующих знаний [65, 79].

Для поиска закономерностей в медицинских данных необходимо иметь достаточный объём этих данных. Не менее важной проблемой при анализе медицинских данных является предоставление результатов их анализа в виде, удобном для верификации и интерпретации. Проблема «сжатого» представления результирующих данных должна решаться на этапе разработки медицинской аналитической системы клинической диагностики.

При разработке МАСКД следует учитывать все указанные выше особенности медицинских данных и выбирать современные методы интеллектуального анализа, направленные на исследование больших объемов многомерных, разнотипных данных с целью выявления в них закономерностей и, в частности, ассоциативных правил и наборов наиболее информативных признаков.

1.3 Виды зависимостей между данными и методы их извлечения

Любые знания о предметной области выражаются, чаще всего, в виде описания закономерностей, существующих между признаками. Например, в клинической практике наибольший интерес имеют следующие основные виды зависимостей [62, 63, 89]:

- причинно-следственные связи, определяющие сведения о патологиях, которые могли быть причинами того или иного заболевания;
- связи, позволяющие прогнозировать течение болезни и состояние здоровья пациента;
- ассоциативные связи, учитывающие, какие синдромы заболевания встречаются одновременно и как часто это происходит.

Такие зависимости, а также многие другие могут быть установлены с помощью методов интеллектуального анализа данных [2, 16, 21, 23, 31].

Если существует цепочка связанных во времени событий, то говорят об их последовательности или временной связи. Например, в течение определенного срока после употребления одного наркотического препарата с высокой степенью вероятности будет принят другой.

Традиционно классификация – это процесс отнесения некоторого объекта или явления к одному из классов [1, 21, 22, 70, 84, 99, 102]. Применительно к медицинским данным можно классифицировать пациентов по виду и степени тяжести заболевания (дифференциальная диагностика), по наличию или отсутствую множественной лекарственной устойчивости к микобактериям туберкулеза легких [71, 72].

Кластеризация в анализе данных – это выделение групп или классов в заданном множестве наблюдаемых объектов, при этом всякая группа должна содержать объекты, схожие между собой в большей степени, чем с объектами других выделяемых групп [52]. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Для медицинских данных кластеризация в большинстве случаев сводится к выделению групп «схожих» пациентов для изучения тех или иных воздействия на них (новых методов лечения или фармакологических препаратов) [16, 98].

Ассоциативные правила отражают, какие признаки, события или явления встречаются вместе и как часто это наблюдается [24, 25, 100, 101]. Например, в наркологии ассоциативные правила позволяют установить зависимости между принятыми препаратами и возможными последствиями от их употребления, а также между наблюдаемыми симптомами и возможными наборами принятых пациентом препаратов [19, 48].

Методы интеллектуального анализа данных, используемые для извлечения зависимостей в данных, можно разделить на несколько групп. Для задач прогноза, например, прогнозирования воздействия различных препаратов или течения болезни традиционно используются нейронные сети, метод наименьших квадратов, деревья принятия решений и логистическая регрессия [3, 15, 16, 23, 33]. При решении задачи диагностики на основе совокупности симптомов применяются метод главных компонент, факторный анализ, методы классификации, кластеризации и дискриминантного анализа [52]. Для поиска зависимостей применяется факторный анализ, алгоритмы выявления ассоциативных правил, байесовские классификаторы, нейросетевой подход и генетические алгоритмы [51, 68, 75, 77, 97, 104]. Следует отметить, что многие из перечисленных методов представляют результаты анализа в виде, плохо интерпретируемым для врача. Это ограничивает их применение в качестве математического обеспечения МАСКД.

Ассоциативные правила являются одним из хорошо изученных классов зависимостей, для которого уже разработаны много методов извлечения. Однако существующие на сегодняшний день методы и средства поиска ассоциативных правил часто приводят к значительному числу искомых ассоциативных правил, многие из которых являются избыточными.

Для решения данной проблемы применяются два подхода: фильтрация с помощью мер значимости [96] и когнитивный подход [40, 131]. Меры значимости позволяют оценивать достоверность и поддержку ассоциативных правил 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 правил [6, 115, 119]. Между тем, оба подхода не исключают появление в результирующем множестве избыточных правил.

В рамках диссертационного исследования предлагается использование когнитивного подхода в следующих аспектах: «сжатие выхода» путем построения неизбыточного минимаксного базиса ассоциативных правил; «сжатие входа» путем снижения размерности матрицы «объект–признак».

1.4 Выводы по главе 1

1. Опыт практического применения существующих зарубежных и отечественных программных систем, реализующих методы интеллектуального анализа данных для решения задач клинической диагностики, демонстрирует полезность применения этих методов для повышения уровня информатизации клинической работы врачей.

2. Существующие универсальные программные средства интеллектуального анализа данных требуют специальной подготовки пользователя и не достаточной степени учитывают специфику медицинских данных. Перенос зарубежных платформ в отечественное практическое здравоохранение в полном объеме также невозможен в виду большой разницы в организации здравоохранения и государственной политики по импортозамещению.

3. Современные МАСКД – результат эволюции экспертных медицинских систем. Они ориентируются исключительно на информатизацию клинической работы врачей и опираются на широкий спектр методов интеллекту-

альной обработки данных. К сожалению, имеется лишь небольшое число отечественных разработок МИС, включая МАСКД, реализация которых находится на уровне практической значимости и внедрения в здравоохранение.

4. В настоящее время основная медицинская информация о пациентах (социально-демографические сведения и клинические данные) аккумулируется в виде электронных медицинских картах, хранимых в базах данных медицинских информационных систем. Анализ медицинских данных значительно осложняется их многомерностью и разнотипностью.

5. При разработке МАСКД следует учитывать особенности медицинских данных и выбирать методы интеллектуального анализа, позволяющие выявлять в них закономерности, в том числе, ассоциативные правила и представлять их в виде удобном для интерпретации пользователем.

Глава 2 Построение неизбыточного минимаксного базиса строгих ассоциативных правил

Существующие методы извлечения ассоциативных правил основываются преимущественно на теории вероятностей и анализе формальных понятий [9, 15, 16, 20, 21, 28, 30, 31, 34, 36, 37, 42, 44, 61, 118]. Анализ формальных понятий является прикладной ветвью алгебраической теории решеток и математическим аппаратом, позволяющим формализовать все понятия, связанные с ассоциативными правилами, включая понятие избыточности [9, 10, 28, 116]. С помощью него представляется возможным формирование «сжатого» представления (базиса) результирующего множества ассоциативных правил.

Глава 2 содержит основные результаты диссертационного исследования, связанные с установлением для строгих ассоциативных правил набора выводимостей, гарантирующих сохранение поддержки найденных правил, разработкой метода и алгоритмов построения неизбыточного минимаксного базиса. Данные результаты опубликованы в работах [11, 12, 14, 49].

В подразделе 2.1 приведены основные сведения об ассоциативных правилах в аспекте анализа формальных понятий. В подразделе 2.2 исследована задача нахождения множества всех ассоциативных правил для заданной предметной области. Приведен обзор алгоритмов ее решения. Указаны проблемы, возникающие при извлечении ассоциативных правил. В подразделе 2.3 исследованы свойства строгих ассоциативных правил и приведен набор выводимостей, гарантирующих сохранение поддержки этих правил. Предложен метод построения неизбыточного минимаксного базиса строгих ассоциативных правил. Подраздел 2.4 посвящен описанию алгоритма построения неизбыточного минимаксного базиса строгих ассоциативных правил. В подразделе 2.5 описан прием, позволяющий существенно снизить количество сгенерированных ассоциативных правил с помощью экспертной группировки признаков и симптомов.

2.1 Анализ формальных понятий и ассоциативные правила

Приведем основные определения и обозначения анализа формальных понятий, применяемые в диссертационной работе [9, 10, 28, 95].

Пусть определены два непустые конечные множества G объектов и M признаков некоторой предметной области. Пусть также задано непустое бинарное отношение $I \subseteq G \times M$. Тройку $K = (G, M, I)$ называют формальным контекстом предметной области. Существование в I пары (g, m), $g \in G$ и $m \in M$, говорит о том, что объект g обладает признаком m и наоборот, признак m характерен для объекта g.

Пусть $g \in G$ и $m \in M$. Определим для g и m отображения φ и ψ:

$$\varphi(g) = \{m \in M : (g, m) \in I\},$$
$$\psi(m) = \{g \in G: (g, m) \in I\}.$$

Здесь $\varphi(g)$ – множество признаков, которые имеет объект g, а $\psi(m)$ – множество объектов, которым присущ признак m. Указанные выше отображения φ и ψ обобщаются на множества $A \subseteq G$ и $B \subseteq M$:

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

Значит, $\varphi(A)$ можно интерпретировать как множество признаков, которые являются общими для всех исследуемых объектов из A, а $\psi(B)$ – множество объектов, имеющих все признаки из B. Определенные выше отображения φ и ψ таковы, что при $A_1, A_2 \subseteq G$ и $B_1, B_2 \subseteq M$ справедливы равенства:

$$\varphi(A_1 \cup A_2) = \varphi(A_1) \cap \varphi(A_2),$$
$$\psi(B_1 \cup B_2) = \psi(B_1) \cap \psi(B_2).$$

Целесообразно положить $\varphi(\varnothing) = M$ и $\psi(\varnothing) = G$, т. е. пустое множество объектов имеет все признаки из M и всякому объекту присуще пустое множество признаков.

Используем для φ и ψ единое обозначение $(\cdot)'$. Тогда выражения для $\varphi(A)$, $\psi(B)$, $\varphi(A_1 \cup A_2)$ и $\psi(B_1 \cup B_2)$ представляются следующим образом:

$$A' = \cap_{g \in A} g', \tag{2.1}$$

$$B' = \cap_{m \in B} m', \tag{2.2}$$

$$(A_1 \cup A_2)' = A_1' \cap A_2', \tag{2.3}$$

$$(B_1 \cup B_2)' = B_1' \cap B_2'. \tag{2.4}$$

Полагается, что $\varphi(g) = \{g\}'$ и $\psi(m) = \{m\}'$.

Справедливы следующие утверждения, непосредственно вытекающие из определения отображений «′».

Утверждение 2.1. Для каждого формального контекста $K = (G, M, I)$ и всяких $B_1, B_2 \subseteq M$ верны свойства:

– антимонотонность: при $B_1 \subseteq B_2$ всегда $(B_2)' \subseteq (B_1)'$;

– экстенсивность: $B_1 \subseteq (B_1)''$, при этом $(B_1)'' = ((B_1)')' \subseteq M$.

Утверждение 2.2. Для каждого формального контекста $K = (G, M, I)$ и всяких $A_1, A_2 \subseteq G$ верны свойства:

– антимонотонность: при $A_1 \subseteq A_2$ всегда $(A_2)' \subseteq (A_1)'$;

– экстенсивность: $A_1 \subseteq (A_1)''$, при этом $(A_1)'' = ((A_1)')' \subseteq G$.

Известно [9, 28], что

$$((A')')' = (A'')' = A', ((B')')' = (B'')' = B'. \tag{2.5}$$

Двойное применение «′» определяет оператор замыкания «′′» на $(2^G, \subseteq)$ или $(2^M, \subseteq)$ в алгебраическом смысле [10]. Для него справедливы свойства:

– рефлексивность: для произвольного $B \subseteq M$ неизменно $B \subseteq B''$;

– монотонность: при $B_1 \subseteq B_2 \subseteq M$ всегда $(B_1)'' \subseteq (B_2)'' \subseteq M$;

– идемпотентность: для произвольного $B \subseteq M$ неизменно $(B'')'' = B''$.

Правильность этих свойств непосредственно следует из приведенных выше утверждений 2.1 и 2.2.

Пара множеств B_1, $B_2 \subseteq M$ называется эквивалентными в контексте $K = (G, M, I)$, если $B_1' = B_2'$. Данное отношение эквивалентности индуцирует соответствующим образом классы эквивалентности в 2^M. Поскольку согласно (2.5) для любого (не обязательно замкнутого) множества $B \subseteq M$ всегда $(B'')' = B'$, то множества B и B'' принадлежат одному классу эквивалентности. Причем B'' является наибольшим по мощности представителем этого класса. Следовательно, B'' можно интерпретировать как наибольшее по включению множество признаков, общих для всех объектов множества B' в пределах контекста $K = (G, M, I)$. В силу монотонности оператора замыкания для всякого множества $B \subseteq M$ его замыкание B'' является наименьшим по включению замкнутым множеством в $K = (G, M, I)$, содержащим B.

Множество B'' – это совокупность признаков, которые всегда встречаются в объектах формального контекста $K = (G, M, I)$ вместе с признаками из B. Если $B = B''$, то $B \subseteq M$ называется замкнутым множеством в $K = (G, M, I)$. Очевидно, что

$$(\varnothing)'' = \varphi(\psi(\varnothing)) = G',$$

где G' – совокупность таких признаков из M, которые свойственны всем объектам формального контекста $K = (G, M, I)$. Если $B' = \varnothing$, то неизменно

$$B'' = \varphi(\psi(B)) = \varphi(\varnothing) = M.$$

Если $B' \neq \varnothing$, то согласно (2.1)–(2.4) замыкание B'' можно определить по следующей формуле:

$$B'' = \cap_{g \in G} \{g' : B \subseteq g'\}, \qquad (2.6)$$

за единственный просмотр формального контекста $K = (G, M, I)$.

Пара множеств (A, B), $A \subseteq G$, $B \subseteq M$, таких, что $A' = B$ и $B' = A$, принято называть формальным понятием формального контекста $K = (G, M, I)$ с объемом A и содержанием B [95]. В (A, B) множества A и B неизменно замкнуты в формальном контексте $K = (G, M, I)$. Далее в ряде случаев определение «формальный» перед словом «контекст» будет опускаться.

Дадим определение ассоциативного правила и связанных с ним числовых функций, характеризующих меры его значимости, в терминах анализа формальных понятий.

Ассоциативным правилом на множестве M формального контекста $K = (G, M, I)$ традиционно называется упорядоченная пара множеств

$$r = (X, Y),\ X, Y \subseteq M.$$

Ассоциативное правило $r = (X, Y)$ принято записывать в виде $X \Rightarrow Y$, множество X называть посылкой (или причиной), а множество Y – заключением (или следствием). Часто полагают, что посылка и заключение ассоциативного правила являются непустыми непересекающимися множествами [20]. С формальных позиций эти ограничения не являются существенными.

Применительно к заданному формальному контексту $K = (G, M, I)$ любое ассоциативное правило $X \Rightarrow Y$ численно характеризуется функциями [2, 123, 134]:

$\delta(X \Rightarrow Y)$ – поддержка (или представительность),

$\gamma(X \Rightarrow Y)$ – достоверность (или значимость).

Эти функции выражаются через отображения «$'$» следующим образом.

Пусть заданы $K = (G, M, I)$ и $X \subseteq M$. Поддержка множества признаков X в $K = (G, M, I)$ обозначается через $\delta(X)$. Эта величина определяется как отношение количества $|X'|$ объектов с признаками из X, к общему числу $|G|$ объектов формального контекста $K = (G, M, I)$:

$$\delta(X) = |X'| / |G|. \qquad (2.7)$$

Таким образом, $\delta(X)$ определяет частоту встречаемости в формальном контексте $K = (G, M, I)$ тех объектов, которые имеют признаки из множества X. Из (2.7) следует, что для каждого $X \subseteq M$ значение $\delta(X)$ всегда находится в границах

$$0 \le \delta(X) \le 1. \qquad (2.8)$$

Очевидно, что чем ближе $\delta(X)$ к единице, тем больше объектов формального контекста $K = (G, M, I)$ имеют признаки X.

Согласно антимонотонности отображений «′», величина $\delta(X)$ также удовлетворяет свойству антимонотонности: для любых $K = (G, M, I)$ и $X, Y \subseteq M$ при $X \subseteq Y$ всегда справедливо неравенство:

$$\delta(Y) \leq \delta(X). \qquad (2.9)$$

Из (2.9) вытекает, что неизменно поддержка множества признаков меньше или равна поддержки любого из его подмножеств. Для произвольного множества признаков $X \subseteq M$ всегда

$$0 \leq \delta(M) \leq \delta(X) \leq \delta(\varnothing) = 1.$$

Множество признаков $X \subseteq M$ принято называть частым в $K = (G, M, I)$, если $\delta(X) \geq \delta_0$, где δ_0 – пороговое значение из интервала [0, 1]. Если одновременно $\delta(X) \geq \delta_0$ и $X = X''$, то X считается частым замкнутым множеством в $K = (G, M, I)$. Частые и частые замкнутые множества признаков являются основой многих существующих алгоритмов извлечения ассоциативных правил. Известно, что в общем случае количество частых замкнутых множеств равно количеству частых множеств признаков и экспоненциально зависит от $|M|$. Однако на практике это не всегда верно [60, 61].

Приведем и докажем известное утверждение, использование которого дает возможность при извлечении ассоциативных правил вместо частых множеств употреблять частые замкнутые множества признаков, и тем самым уменьшать на практике область поиска ассоциативных правил.

Утверждение 2.3. Для каждого формального контекста $K = (G, M, I)$ и всякого $X \subseteq M$ поддержка X'' всегда совпадает с поддержкой X:

$$\delta(X'') = \delta(X).$$

Доказательство. Для произвольного $X \subseteq M$ согласно из (2.5) и (2.7), справедливо равенство

$$\delta(X'') = |(X'')'| / |G| = |X'| / |G| = \delta(X),$$

которое доказывает правильность утверждения 2.3. □

Из утверждения 2.3 вытекает, что при $\delta(X) \geq \delta_0$ всегда верно $\delta(X'') \geq \delta_0$, т. е. замыкание частого множества признаков тоже является частым.

Поддержка $\delta(X \Rightarrow Y)$ ассоциативного правила $X \Rightarrow Y$ относительно формального контекста $K = (G, M, I)$ определяется формулой

$$\delta(X \Rightarrow Y) = \delta(X \cup Y) = |(X \cup Y)'| / |G|, \qquad (2.10)$$

и указывает долю объектов этого контекста, которым присущи признаки $X \cup Y$. Заметим, что согласно (2.3) верно равенство

$$(X \cup Y)' = X' \cap Y'.$$

Достоверность $\gamma(X \Rightarrow Y)$ ассоциативного правила $X \Rightarrow Y$ определяется отношением числа $|(X \cup Y)'|$ объектов, обладающих всеми признаками из $X \cup Y$, к числу $|X'|$ объектов, которым присущи лишь признаки X:

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

Данную величину можно также выразить формулой:

$$\gamma(X \Rightarrow Y) = \delta(X \Rightarrow Y) / \delta(X) = \delta(X \cup Y) / \delta(X). \qquad (2.11)$$

Заметим, что формула (2.11) верна только для ассоциативных правил $X \Rightarrow Y$, при $\delta(X) \neq 0$. При $\delta(X) = 0$ (когда контексте не содержит ни одного объекта, обладающего признаками X) согласно (2.8) и (2.9) также справедливо равенство $\delta(X \cup Y) = 0$. В таком особом случае полагается $\gamma(X \Rightarrow Y) = 1$. Исходя из (2.7)–(2.11), всегда находится в естественных границах

$$0 \leq \gamma(X \Rightarrow Y) \leq 1.$$

Чем ближе $\gamma(X \Rightarrow Y)$ к единице, тем с большей уверенностью можно сказать, что признаки Y появляются вместе с признаками X.

Ассоциативное правило $X \Rightarrow Y$ считается минимаксным в $K = (G, M, I)$, если для данного контекста нет другого ассоциативного правила $X^* \Rightarrow Y^*$ такого, что $X^* \subseteq X$, $Y \subseteq Y^*$ и справедливы равенства:

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

Таким образом, всякое минимаксное ассоциативное правило определяет причинно-следственную связь с минимальной по включению посылкой X и максимальным по включению следствием Y. Именно минимаксные ассоциативные правила интересны для клинической диагностики, поскольку каждое из них может определять минимальный набор диагностических признаков некоторого заболевания и максимальный набор последствий этого заболевания.

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

$$\delta_0 \le \delta(X \Rightarrow Y) \le 1, \quad (2.12)$$

$$\gamma_0 \le \gamma(X \Rightarrow Y) \le 1. \quad (2.13)$$

Величины δ_0 и γ_0 выполняют роль пороговых значений для поддержки и достоверности соответственно. При $\delta_0 = 0$ условие (2.12) отражает естественные границы поддержки. Данная ситуация свидетельствует о том, что нет ограничений на частоту появления признаков $X \cup Y$ в контексте $K = (G, M, I)$. При $\gamma_0 = 1$ условие (2.13) приводит к равенству $\gamma(X \Rightarrow Y) = 1$. В этом случае имеем $(\delta_0, 1)$-ассоциативное правило, которое называют строгим. Следовательно, строгое ассоциативное правило – правило с достоверностью 1 и любой ненулевой поддержкой. Ассоциативные правила с нулевой поддержкой и нулевой достоверностью не имеют практической ценности и поэтому далее не рассматриваются.

Из формул (2.10), (2.11) следуют два важных с алгоритмической точки зрения свойства достоверности ассоциативных правил:

$$\gamma(X \Rightarrow Z) = \gamma(X \Rightarrow Y) \cdot \gamma(Y \Rightarrow Z) \text{ при любых } X \subseteq Y \subseteq Z \subseteq M, \quad (2.14)$$

$$\gamma(X \Rightarrow Y) = \gamma(X \Rightarrow Y \cup Z) \text{ для любых } Z \subseteq X \subseteq M. \quad (2.15)$$

Формула (2.14) указывает способ расчета достоверности ассоциативного правила $X \Rightarrow Z$, полученного из транзитивной цепочки двух других правил $X \Rightarrow Y$ и $Y \Rightarrow Z$ при заданных условиях вложенности посылок и заключений. Формула (2.15) показывает случай, когда без изменения значений достоверности и поддержки можно пополнять заключение ассоциативного правила.

Следует отметить, что существует множество других мер значимости ассоциативных правил [96]. Многие из них можно вывести, используя значение поддержки в терминах анализа формальных понятий.

2.2 Задача извлечения ассоциативных правил и проблема числа правил

Задача извлечения ассоциативных правил формулируется следующим образом.

Заданы формальный контекст $K = (G, M, I)$ и δ_0, γ_0 – вещественные числа из интервала [0, 1].

Требуется для заданного контекста $K = (G, M, I)$ найти 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 (2.12) и (2.13). Известно, что 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. Такой переход не изменяет сложность рассматриваемой задачи, ее пере-

числительный комбинаторный характер, а лишь позволяет представить результирующее множество *AR* в «сжатой» форме, удобной для интерпретации.

К настоящему времени разработано большое число методов и алгоритмов выявления ассоциативных правил. Их обзор представлен в работах [85, 93, 101, 110, 113, 125, 126]. Первые алгоритмы извлечения ассоциативных правил использовали теоретико-вероятностный подход к генерации частых множеств или кандидатов (Candidate generation algorithms). Самым известным представителем этой группы является алгоритм Apriori, созданный в 1994 году [93, 103].

Алгоритм Apriori генерирует (δ_0, γ_0)-ассоциативные правила для любых δ_0 и γ_0 и основывается на свойстве антимонотонности функции поддержки. В процессе работы алгоритма Apriori происходит многократное сканирование контекста $K = (G, M, I)$ и вычисление поддержки для каждой новой генерации частых множеств. Эффективность алгоритма Apriori высокая, в случае если значение δ_0 близко к 1. При малых значениях δ_0 процесс генерации ассоциативных правил приводит к полному перебору всех подмножеств множества *M*. Практика показала, что данный алгоритм генерирует огромное количество избыточных правил. Число сгенерированных правил может в несколько раз превосходить размер контекста $K = (G, M, I)$.

С целью устранения многократного сканирования начальных данных и постоянной генерации наборов-кандидатов в 2000 году были предложены алгоритмы «наращивания часто встречаемых наборов» (Pattern growth algorithms). Одним из наиболее эффективных алгоритмов является Frequent Pattern Growth (FPG), который дает возможность уменьшить необходимое число сканирований исходных данных до двух [125].

В основе алгоритма FPG лежит предобработка начальных данных, в процессе которой они преобразуются в компактную древовидную структуру FP-tree (frequent pattern tree). При таком представлении данных не требуется производить вычисление поддержки для каждого частого набора. Тем не менее, известны примеры, когда построение дерева занимает еще больше вре-

мени, чем генерация частых наборов, а размер построенного дерева может превышать размер контекста $K = (G, M, I)$. Несмотря на то, что перечисленные алгоритмы основываются на свойстве антимонотонности, позволяющем отсечь достаточное количество вариантов, подобные алгоритмы вычислительно неэффективны на больших контекстах $K = (G, M, I)$.

Наибольшую эффективность по времени и памяти показали алгоритмы, базирующиеся на методах анализа формальных понятий, а именно замкнутых множествах. К таким алгоритмам относится алгоритм Close и его разновидности A-Close, CHARM, Close$^+$ [114, 127, 132]. Алгоритм Close извлекает только $(\delta_0, 1)$-ассоциативные правила из частых замкнутых наборов признаков. Переход от частых множеств к частым замкнутым множествам дает возможность сузить пространство поиска. Другим достоинством алгоритма Close и его многочисленных модификаций является формирование минимаксных ассоциативных правил.

Пример 2.1. Рассмотрим контекст $K = (G, M, I)$, представленный в таблицах 2.1 и 2.2, где $G = \{g_1, g_2, g_3, g_4, g_5\}$ – множество объектов, $M = \{a, b, c, d, e\}$ – множество признаков, I– матрица инцидентности. В таблице 2.1 и далее при написании множеств для краткости мы опускаем фигурные скобки, запятые между элементами этих множеств и располагаем элементы в лексикографическом порядке. Например, вместо $\{a, b, c\}$ записываем abc, а вместо $\{a, b\} \Rightarrow \{c\}$ кратко пишем $ab \Rightarrow c$.

Таблица 2.1 – Исходный контекст

Объекты	Признаки, присущие объектам
g_1	*acd*
g_2	*bce*
g_3	*abce*
g_4	*be*
g_5	*abce*

Таблица 2.2 – Матрица инцидентности исходного контекста

	a	*b*	*c*	*d*	*e*
g_1	1	0	1	1	0
g_2	0	1	1	0	1
g_3	1	1	1	0	1
g_4	0	1	0	0	1
g_5	1	1	1	0	1

Для данного контекста при $\delta_0 = 1/5$ и $\gamma_0 = 1$ алгоритм Apriori приводит к семнадцати строгим ассоциативным правилам:

$$AR = \{a \Rightarrow c,\ d \Rightarrow a,\ e \Rightarrow b,\ b \Rightarrow e,\ d \Rightarrow c,\ ab \Rightarrow c,$$
$$ae \Rightarrow b,\ ab \Rightarrow e,\ cd \Rightarrow a,\ ad \Rightarrow c,\ d \Rightarrow ac,\ ae \Rightarrow c,$$
$$ce \Rightarrow b,\ bc \Rightarrow e,\ ace \Rightarrow b,\ abe \Rightarrow c,\ abc \Rightarrow e\}.$$

Полученное множество *AR* содержит много избыточных правил в том смысле, что если их исключить, то они выводимы из оставшихся правил. Избыточными являются следующие ассоциативные правила:

$$\{d \Rightarrow c,\ ab \Rightarrow c,\ ae \Rightarrow b,\ ab \Rightarrow e,\ cd \Rightarrow a,\ ad \Rightarrow c,\ d \Rightarrow ac,$$
$$ae \Rightarrow c,\ ce \Rightarrow b,\ bc \Rightarrow e,\ ace \Rightarrow b,\ abe \Rightarrow c,\ abc \Rightarrow e\}.$$

Некоторые из представленных ассоциативных правил исключаются алгоритмом Close в процессе построения множества *AR*. В результате алгоритм Close извлекает лишь восемь минимаксных строгих ассоциативных правил:

$$AR = \{a \Rightarrow c,\ b \Rightarrow e,\ e \Rightarrow b,\ d \Rightarrow ac,$$
$$bc \Rightarrow e,\ ce \Rightarrow b,\ ab \Rightarrow ce,\ ae \Rightarrow bc\}.$$

Алгоритм Close находит минимаксные строгие ассоциативные правила по следующим частым замкнутым наборам признаков:

$$c,\ ac,\ be,\ bce,\ acd,\ abce.$$

Все эти правила имеют допустимые поддержки:

$$\delta(a \Rightarrow c) = 3/5,\ \delta(b \Rightarrow e) = 4/5,$$
$$\delta(e \Rightarrow b) = 4/5,\ \delta(d \Rightarrow ac) = 1/5,$$

$$\delta(bc \Rightarrow e) = 3/5,\ \delta(ce \Rightarrow b) = 3/5,$$

$$\delta(ab \Rightarrow ce) = 2/5,\ \delta(ae \Rightarrow bc) = 2/5. \square$$

Главным недостатком рассмотренных выше алгоритмов является большое число результирующих правил, которые существенно усложняют экспертный анализ выявленных (δ_0, γ_0)-ассоциативных правил. Фильтрация (δ_0, γ_0)-ассоциативных правил с помощью δ_0 и γ_0 сокращает число правил, но не решает проблему размерности полностью. После фильтрации остается большое число ассоциативных правил, при этом многие из них избыточные. 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 ассоциативных правил такой информацией, прежде всего, служит уровень или порог поддержки δ_0 – величина, характеризующая минимальную представительность этих правил в анализируемых данных.

2.3 Метод построения неизбыточного минимаксного базиса строгих ассоциативных правил

Докажем для строгих ассоциативных правил набор выводимостей, которые дают возможность из одних строгих ассоциативных правил выводить другие строгие ассоциативные правила (с сохранением или без сохранения поддержки). Для этого нам необходим критерий распознавания наличия в контексте $K = (G, M, I)$ строгого ассоциативного правила [95]. Приведем его с доказательством.

Утверждение 2.4. Достоверность ассоциативного правила $X \Rightarrow Y$ относительно контекста $K = (G, M, I)$ равна 1 тогда и только тогда, когда $X' \subseteq Y'$ (или $Y \subseteq X''$).

Доказательство. Согласно формуле (2.11) равенство $\gamma(X \Rightarrow Y) = 1$ верно тогда и только тогда, когда $\delta(X \cup Y) = \delta(X)$, или, то же самое, когда $(X \cup Y)' = X'$. В силу (2.4) равенство $X' \cap Y' = X'$ возможно тогда и только тогда, когда $X' \subseteq Y'$.

Теперь предположим, что $X' \subseteq Y'$. По утверждениям 2.1 и 2.2 всегда $Y \subseteq Y''$, а при $X' \subseteq Y'$ верно включение $Y'' \subseteq X''$. Отсюда $Y \subseteq X''$. Пусть $Y \subseteq X''$. Тогда в силу антимонотонности отображения «$'$» имеем включение $(X'')' \subseteq Y'$. С учетом (2.5) верно $X' \subseteq Y'$. Утверждение 2.4 доказано. □

Заметим, что утверждение 2.4 тривиальным образом выполняется для $X \Rightarrow X''$ во всяком контексте $K = (G, M, I)$ и при любом $X \subseteq M$.

Рассмотрим некоторые частные случаи утверждения 2.4, важные с точки зрения устранения избыточности во множестве строгих ассоциативных правил.

Случай 1: Ассоциативные правила вида $X \Rightarrow Y$ при любых $Y \subseteq X \subseteq M$.

В силу антимонотонности отображения «$'$» при $Y \subseteq X$ справедливо включение $X' \subseteq Y'$. Условие утверждения 2.4 выполняется, поэтому $\gamma(X \Rightarrow Y) = 1$. Таким образом, если в ассоциативном правиле заключение является подмножеством посылки, то такое правило имеет достоверность 1 в любом контексте $K = (G, M, I)$ с поддержкой $\delta(X)$. Подобные строгие ассоциативные правила не несут в себе информации о существенных отношениях между множествами признаков X и Y, кроме естественного отношения «целое и часть целого». Поэтому их следует считать тривиальными и не принимать во внимание. В частности, ассоциативные правила вида $\varnothing \Rightarrow \varnothing$, $M \Rightarrow Y$ и $X \Rightarrow \varnothing$ при любых $X, Y \subseteq M$ относятся к тривиальным строгим правилам.

Случай 2: Ассоциативные правила вида $\varnothing \Rightarrow Y$ при $\varnothing \neq Y \subseteq M$.

Для всякого формального контекста $K = (G, M, I)$ и $Y \subseteq M$ всегда $Y' \subseteq G$. Кроме того $\psi(\varnothing) = \varnothing' = G$ и $\delta(\varnothing) = 1$. Поэтому для ассоциативного правила $\varnothing \Rightarrow Y$ имеем равенства:

$$\gamma(\varnothing \Rightarrow Y) = \delta(\varnothing \Rightarrow Y) / \delta(X) = \delta(\varnothing \cup Y) / \delta(\varnothing) = \delta(Y).$$

Исходя из утверждения 2.4, равенство $\gamma(\varnothing \Rightarrow Y) = 1$ справедливо тогда и только тогда, когда $\varnothing' \subseteq Y'$. Поскольку $\psi(\varnothing) = \varnothing' = G$, то включение $\varnothing' \subseteq Y'$ верно лишь при $G = Y'$. Таким образом, строгое ассоциативное правило $\varnothing \Rightarrow Y$ при $Y \neq \varnothing$ имеет поддержку $\delta(\varnothing) = 1$ и отражает наличие жесткого ограничения на контекст $K = (G, M, I)$: все объекты, которые представлены в этом контексте, обязательно обладают множеством признаков Y.

Рассмотрим строгое ассоциативное правило $X \Rightarrow Y$. В силу (2.11) его поддержка всегда совпадает с поддержкой его посылки: $\delta(X \Rightarrow Y) = \delta(X)$. Если $\delta(X) \geq \delta_0$, то также $\delta(X \Rightarrow Y) \geq \delta_0$. Если после какого-либо преобразования правила $X \Rightarrow Y$, результирующее правило имеет поддержку не менее $\delta(X)$, то говорят, что такое преобразование сохраняет поддержку исходного правила.

Сформулируем и докажем для строгих ассоциативных правил набор выводимостей, которые позволяют из одних строгих ассоциативных правил вывести другие строгие ассоциативные правила (с сохранением или без сохранения поддержки).

Лемма 2.1. 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 $\delta(X) \geq \delta_0$.

Доказательство. При $Y \subseteq X$ в силу антимонотонности отображения «′» всегда $Y' \subseteq X'$. Тогда по утверждению 2.4 ассоциативное правило $X \Rightarrow Y$ является строгим. В силу (2.11) для него $\delta(X \Rightarrow Y) = \delta(X) \geq \delta_0$. □

Лемма 2.2. 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 $\delta(X \cup Z) \leq \delta(X)$.

Доказательство. Воспользуемся утверждением 2.4 и свойством монотонности оператора замыкания. Так как $X \Rightarrow Y$ являет строгим ассоциативным правилом в $K = (G, M, I)$, то $Y \subseteq X''$ и $\delta(X \cup Y) = \delta(X)$. При $X \subseteq X \cup Z$ верно включение $X'' \subseteq (X \cup Z)''$. Следовательно, $Y \subseteq (X \cup Z)''$. Это означает, что для $K = (G, M, I)$ справедливо строгое ассоциативное правило $X \cup Z \Rightarrow Y$ и для него $\delta(X \cup Z \cup Y) = \delta(X \cup Z)$. Отсюда с учетом антимонотонности поддержки имеем

$$\delta(X \cup Z) = \delta(X \cup Z \cup Y) \leq \delta(X \cup Y) = \delta(X).$$

Лемма 2.2 доказана. □

Лемма 2.2 отражает возможность пополнения посылки для строгого ассоциативного правила, но без гарантии сохранения поддержки. Особо следует отметить случай, при котором расширение посылки строгого ассоциативного правила сохраняет поддержку этого правила.

Следствие 2.1. Если для контекста $K = (G, M, I)$ справедливо 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 $\delta(X)$.

Доказательство. Если $\gamma(X \Rightarrow Y) = 1$, то по лемме 2.2 также $\gamma(X \cup Z \Rightarrow Y) = 1$. Значит, верны равенства

$$\delta(X \Rightarrow Y) = \delta(X \cup Y) = \delta(X),$$
$$\delta(X \cup Z \Rightarrow Y) = \delta(X \cup Z \cup Y) = \delta(X \cup Z).$$

Отсюда при $Z \subseteq Y$ имеем

$$\delta(X \cup Z \Rightarrow Y) = \delta(X \cup Z \cup Y) = \delta(X \cup Y) = \delta(X).$$ □

Лемма 2.3. 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 $\delta(X) \geq \delta_0$.

Доказательство. В данном случае целесообразно вновь воспользоваться необходимыми и достаточными условиями для строгих правил, которые определены в утверждении 2.4. Поскольку $X \Rightarrow Y$ и $X \Rightarrow Z$ являются строгими ассоциативными правилами в $K = (G, M, I)$, то необходимо выполняются включения $Y \subseteq X''$ и $Z \subseteq X''$. Значит, $Y \cup Z \subseteq X''$, что достаточно для выполнимости строгого ассоциативного правила $X \Rightarrow Y \cup Z$ в заданном контексте. В данном случае $\delta(X \Rightarrow Y \cup Z) = \delta(X) \geq \delta_0$, т. е. свойство аддитивности сохраняет поддержку исходных правил. Лемма 2.3 доказана. □

Очевидно, что свойства рефлексивности и пополнения не выполняются для произвольных (δ_0, γ_0)-ассоциативных правил. Однако, следующее свойство, называемое проективностью, справедливо для любых (δ_0, γ_0)-ассоциативных правил.

Лемма 2.4. 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 $X \Rightarrow Z$.

Доказательство. Поскольку $Z \subseteq Y$ и $Y \neq \varnothing$, то Y можно представить в виде $Y = Z \cup (Y \setminus Z)$. Исходя из условия (2.15) и свойства антимонотонности функции поддержки, имеем

$$\delta_0 \leq \delta(X \cup Y) = \delta(X \cup (Z \cup (Y \setminus Z))) =$$
$$= \delta((X \cup Z) \cup (Y \setminus Z)) \leq \delta(X \cup Z). \qquad (2.16)$$

Значит, $\delta(X \Rightarrow Z)$ удовлетворяет условию (2.15). Аналогичным образом полагая, что $\delta(X) \neq 0$, получаем

$$\gamma_0 \leq \gamma(X \Rightarrow Y) = \delta(X \Rightarrow Y) / \delta(X) =$$
$$= \delta(X \cup Y) / \delta(X) \leq \delta(X \cup Z) / \delta(X) = \gamma(X \Rightarrow Z).$$

Следовательно, $\gamma(X \Rightarrow Z)$ удовлетворяет условию (2.16). При $\delta(X) = 0$ (или то же самое $X' \neq \varnothing$) всегда $\gamma(X \Rightarrow Y) = 1$ при любом Y, в том числе и $Z \subseteq Y$. Лемма 2.4 доказана. □

Применительно к строгим ассоциативным правилам данная лемма доказывается тривиальным образом. Если для контекста $K = (G, M, I)$ справедливо строгое ассоциативное правило $X \Rightarrow Y$, то $Y \subseteq X''$. Значит, при любых $Z \subseteq Y$ и $Y \neq \varnothing$ справедливо включение $Z \subseteq X''$. Следовательно, $X \Rightarrow Z$ является строгим ассоциативным правилом в контексте $K = (G, M, I)$. Кроме того, верны равенства

$$\delta(X \Rightarrow Y) = \delta(X), \quad \delta(X \Rightarrow Z) = \delta(X).$$

Очевидно, что если $\delta(X) \geq \delta_0$, то $\delta(X \Rightarrow Z) \geq \delta_0$.

Лемма 2.4 отражает тот факт, что 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 2.3 и 2.4 свидетельствуют о равноценности различных эквивалентных форм записи этих правил.

Следствие 2.2. Представление строгого ассоциативного правила в виде $X \Rightarrow Y \cup Z$ эквивалентно его представлению в виде двух строгих ассоциативных правил $X \Rightarrow Y$ и $X \Rightarrow Z$, при этом

$$\delta(X \Rightarrow Y \cup Z) = \delta(X \Rightarrow Y) = \delta(X \Rightarrow Z) = \delta(X).$$

Лемма 2.5. 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 $\delta(X) \geq \delta_0$.

Доказательство. Воспользуемся утверждением 2.4 и свойством монотонности оператора замыкания. Если для контекста $K = (G, M, I)$ справедливы строгие ассоциативные правила $X \Rightarrow Y$ и $Y \Rightarrow W$, то верны включения

$$X' \subseteq Y' \text{ и } Y' \subseteq W'.$$

Следовательно, $X' \subseteq W'$. По утверждению 2.4 это условие является достаточным для выполнимости строгого ассоциативного правила $X \Rightarrow W$.

Для него $\delta(X \Rightarrow W) = \delta(X) \geq \delta_0$. Таким образом, поддержка результирующего правила $X \Rightarrow W$ совпадает с поддержкой правила $X \Rightarrow Y$, играющего роль начала транзитивной цепочки строгих ассоциативных правил. Лемма 2.5 доказана. □

Следующая лемма обобщает лемму 2.5 и определяет свойство псевдотранзитивности строгих ассоциативных правил.

Лемма 2.6. 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 $\delta(X \cup Z) \leq \delta(X)$.

Доказательство. Если для контекста $K = (G, M, I)$ справедливы строгие ассоциативные правила $X \Rightarrow Y$ и $Y \cup Z \Rightarrow W$, то

$$X' \subseteq Y' \text{ и } (Y \cup Z)' \subseteq W'.$$

Требуется доказать, что $(X \cup Z)' \subseteq W'$. По лемме 2.2 из строгого ассоциативного правила $X \Rightarrow Y$ вытекает справедливость строгого ассоциативного правила $(X \cup Z) \Rightarrow Y$. Тогда $(X \cup Z)' \subseteq Y'$. В силу антимонотонности отображения «′» всегда $(X \cup Z)' \subseteq Z'$. Отсюда, если $(X \cup Z)' \subseteq Y'$ и $(X \cup Z)' \subseteq Z'$, то верно включение

$$(X \cup Z)' \subseteq Y' \cap Z'.$$

По формуле (2.4) имеем $(Y \cup Z)' = Y' \cap Z'$. Отсюда, учитывая, что $(Y \cup Z)' \subseteq W'$, окончательно получаем

$$(X \cup Z)' \subseteq Y' \cap Z' = (Y \cup Z)' \subseteq W'. \square$$

К сожалению, свойство псевдотранзитивности не гарантирует для результирующего правила сохранение поддержки. Для него всегда

$$\delta(X \cup Z \Rightarrow W) = \delta(X \cup Z).$$

Поэтому по свойству антимонотонности поддержки $\delta(X \cup Z) \leq \delta(X)$.

Леммы 2.1–2.6 позволяют сформулировать следующую теорему [11].

Теорема 2.1. 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.

Указанные в теореме 2.1 свойства (или выводимости) D_1–D_6 соответствуют выводимостям Армстронга [66, 83]. Они дают возможность из некоторого множества строгих ассоциативных правил вывести многие другие строгие ассоциативные правила без дополнительного сканирования контекста. Доказательства того, что D_1, D_3, D_4, D_5 гарантируют сохранение поддержки, позволяют использовать данные выводимости при построении базисов для множества $(\delta_0, 1)$-ассоциативных правил.

Дадим определение избыточного строгого ассоциативного правила. Пусть AR–множество всех $(\delta_0, 1)$-ассоциативных правил формального контекста $K = (G, M, I)$. Будем говорить, что 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

$$AR \setminus \{X \Rightarrow Y\} \models_{\delta_0} X \Rightarrow Y. \qquad (2.17)$$

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 (2.17) и формирование минимаксного базиса через генераторы частых замкнутых наборов признаков, аналогично тому, как это осуществляется в алгоритме Close. Кроме теоремы 2.1 теоретическим обоснованием предлагаемого метода является корректность алгоритма Close, доказанная M. Zaki и C. Hsiao [132].

2.4 Алгоритм MClose построения неизбыточного минимаксного базиса строгих ассоциативных правил

Свойства (выводимости) D_1–D_6, приведенные в теореме 2.1, дают возможность из одного множества строгих ассоциативных правил выводить многие другие правила без дополнительного обращения к контексту. На первый взгляд, именно выводимости D_1–D_6 являются причиной экспоненциального числа возможных строгих ассоциативных правил для анализируемого контекста. С другой стороны, они позволяют строить неизбыточный минимаксный базис для множества строгих ассоциативных правил, описывающий связи между данными в компактной форме.

Для построения неизбыточного минимаксного базиса необходимо выполнить генерацию минимаксных строгих ассоциативных правил с помощью алгоритма Close и устранить среди них избыточные правила.

Разработанный в диссертации алгоритм MClose – модификация алгоритма Close, направленная на удаление из результирующего множества *AR*

избыточных $(\delta_0, 1)$-ассоциативных правил. Причем распознавание избыточности осуществляется в процессе формирования *AR*. Исходными данными для алгоритма MClose являются исходный контекст $K = (G, M, I)$ и пороговое значение δ_0. Алгоритма MClose извлекает все $(\delta_0, 1)$-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 следствии 2.1 (свойстве сохранения поддержки $(\delta_0, 1)$-ассоциативного правила при пополнении его посылки в отмеченном частном случае).

Суть алгоритма Close сводится к пошаговому извлечению генераторов и частых замкнутых наборов признаков [132]:

- вначале искомое множество *AR* является пустым и $k = 1$;
- первый шаг алгоритма заключается в применении всех одноэлементных подмножеств множества *M* в качестве k-генераторов. Замыкание ρ_k'' для генератора ρ_k вычисляется по формуле (2.3). Значение поддержки для ρ_k'' находится по формуле (2.6);
- если $\delta(\rho_k'') \geq \delta_0$, то по ρ_k'' создается минимаксное ассоциативное правило вида $\rho_k \Rightarrow \rho_k'' \setminus \rho_k$ и сохраняется в *AR*;
- затем по ρ_k'' создаются кандидаты в $(k + 1)$-генераторы для следующей итерации. Каждый подобный кандидат формируется путем объединения двух k-генераторов, обладающих одинаковыми первыми $k - 1$ признаками;

- далее осуществляется проверка, вложен ли созданный кандидат в ρ_k''. Если вложен, то этот кандидат исключается из дальнейшего рассмотрения;
- после нахождения всех возможных (k + 1)-генераторов осуществляется переход к следующей итерации;
- завершение работы алгоритма происходит, когда все генераторы исчерпаны.

Множество всех ассоциативных правил, выявленных в результате работы алгоритма Close, формирует минимаксный базис строгих ассоциативных правил контекста $K = (G, M, I)$. Алгоритм Close можно модифицировать таким образом, что в результирующее множество AR не будут попадать те строгие ассоциативные правила, которые являются заведомо избыточными.

Нахождение избыточных строгих ассоциативных правил во множестве AR основывается на проверке условия (2.17). При реализации этой проверки используется понятие замыкания множества признаков относительно множества AR. Алгоритм, осуществляющий такую проверку полиноминален относительно $|M|$ по времени. Замыканием множества $X \subseteq M$ относительно AR (обозначается X^+) называется множество всех признаков $m \in M$ таких, что верно логическое следование $AR \models_{\delta_0} X \Rightarrow m$. Отметим, что всегда $X^+ \subseteq M$. Из выводимостей D_1, D_3, D_4 следует верность следующего утверждения.

Утверждение 2.5. Следование $AR \models_{\delta_0} X \Rightarrow Y$ верно, если и только если $Y \subseteq X^+$.

Всегда $AR \models_{\delta_0} X \Rightarrow X^+$, $AR \models_{\delta_0} X \Rightarrow X^+ \setminus X$. Исходя из утверждения 2.5, для того чтобы убедится в справедливости (2.17), достаточно лишь вычислить X^+ относительно $AR \setminus \{X \Rightarrow Y\}$ и произвести проверку включения $Y \subseteq X^+$. В случае если $Y \subseteq X^+$, то ассоциативное правило $X \Rightarrow Y$ является избыточным в AR, иначе оно не является избыточным. □

Процедура SX построения X^+ целиком основывается на выводимостях D_1, D_3, D_4, D_5, гарантирующих сохранения поддержки, производится путем выполнения следующих действий:

- вначале полагается, что $X^+ = X$;
- далее производится просмотр всех правил из AR и пополнение замыкания по следующему принципу: если для $Y \Rightarrow Z \in AR$ справедливо включение $Y \subseteq X^+$, то множество Z добавляется к X^+.

Процесс повторяется пока изменяется X^+. Так как множества M и AR являются конечными, то и процесс построения X^+ конечен.

Для исключения добавления заведомо избыточного строгого ассоциативного правила в результирующее множество AR требуется каждый раз после нахождения ρ_k''производить следующие. Если посылка ρ_k выявленного ассоциативного правила не равна ρ_k'', то выполняется поиск замыкания ρ_k^+ относительно найденного множества AR. Если $\rho_k^+ = \rho_k''$, то минимаксное правило $\rho_k \Rightarrow \rho_k'' / \rho_k$ является избыточным (согласно утверждению 2.5), иначе оно включается в результирующее множество AR.

После окончания процесса генерации минимаксных строгих ассоциативных правил необходимо произвести дополнительное сканирование множества AR для выявления оставшихся избыточных правил. Такие строгие ассоциативные правила не являются избыточными по отношению к ранее найденным правилам, однако могут стать избыточными после поступления в AR новых ассоциативных правил. Эти действия выполняет процедура Non-Redundancy, используемая в алгоритме MClose.

В результате работы алгоритма множество AR является неизбыточным и в его состав входят только минимаксные строгие ассоциативные правила. Отметим, что оперативное исключение избыточных строгих ассоциативных правил замедляет рост мощности множества AR и уменьшает время работы алгоритма.

Описание алгоритма MClose приведено на рисунке 2.1. В алгоритме MClose процедуры Gen-Closure и Gen-Generator выполняют вычисление замыканий и генераторов. Они аналогичны одноименным процедурам классического алгоритма Close [132]. Все действия, выполняемые алгоритмом MClose и несвойственные алгоритму Close, выполняются за полиномиальное время относительно $|M|$.

Алгоритм MClose

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

1: ***begin***
2: $AR \leftarrow \varnothing$
3: $k \leftarrow 1$
4: ***while*** $\rho_k \neq \varnothing$
5: **Gen-Closure** (ρ_k)
6: ***if*** $\delta(\rho_k) \geq \delta_0$
7: ***if*** $\rho_k \neq \rho_k''$
8: $\rho_k^+ \leftarrow SX(\rho_k)$
9: ***end if***
10: ***if*** $\rho_k^+ \neq \rho_k''$
11: $AR \leftarrow AR \cup (\rho_k \Rightarrow \rho_k'' \setminus \rho_k)$
12: ***end if***
13: ***end if***
14: **Gen-Generator** ($k + 1$)
15: $k \leftarrow k + 1$
16: ***end while***
17: **Non-Redundancy** (AR)
18: ***end***

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

Рисунок 2.1 – Пошаговое описание алгоритма MClose

Пример 2.2. Для контекста, представленного в таблицах 2.1 и 2.2, при $\delta_0 = 1/5$ и $\gamma_0 = 1$ минимаксный базис состоит из восьми правил, в котором избыточными являются

$$bc \Rightarrow e,\ ce \Rightarrow b,\ ab \Rightarrow ce,\ ae \Rightarrow bc.$$

Неизбыточный минимаксный базис, построенный алгоритмом MClose, состоит всего из четырех минимаксных строгих ассоциативных правил:

$$AR = \{a \Rightarrow c,\ b \Rightarrow e,\ e \Rightarrow b,\ d \Rightarrow ac\}.$$

Следует отметить, что на основе *AR* можно получить любое строгое ассоциативное правило с помощью алгоритма построения X^+. Для нахождения максимального следствия для посылки *bc*, достаточно построить bc^+ относительно *AR*. В результате построения имеем $bc^+ = bce$. Отсюда $AR \models_{\delta_0} bc \Rightarrow e$. Поскольку $bc^+ = bc'' = bce$ и $e \in bc''$, то ассоциативное правило $bc \Rightarrow e$ является строгим с поддержкой $\delta(bc \Rightarrow e) = \delta(bc) = 3/5$. □

Для проверки эффективности алгоритмов Apriori, Close и MClose проводились численные эксперименты по сравнению числа сгенерированных ими строгих ассоциативных правил и времени их работы. Эксперименты проводились на компьютере с процессором Intel® Core™ i5 CPU& 2.30GHz и ОЗУ размером 4 ГБ на контекстах, сгенерированных случайным образом, обладающих различной размерностью и плотностью. Результаты проведенных экспериментов показаны в таблице 2.3. Для всех анализируемых контекстов $K = (G, M, I)$ в таблице указаны $|G|$ – количество объектов, $|M|$ –количество признаков, а $\sigma = n/(|G| \cdot |M|)$ – плотность контекста, где *n* отражает число элементов матрицы инцидентности *I*, равных единице. Контекст из 10000 объектов был сформирован многократным копированием контекста, состоящего из 500 объектов.

Таблица 2.3 – Сравнение алгоритмов Apriori, Close и MClose

Характеристика контекста $K = (G, M, I)$			Число извлеченных строгих ассоциативных правил / время, мс					
$\|G\|$	$\|M\|$	σ	Apriori		Close		MClose	
20	10	0,38	1797	17562	45	250	22	297
30	10	0,39	2029	18347	46	374	19	412
30	10	0,55	15438	187202	69	390	20	484
50	10	0,53	27769	375178	46	78	13	124
500	10	0,53	27769	376154	42	124	13	168
10000	10	0,53	27769	378400	42	671	13	872

Из таблицы 2.3 следует, что алгоритмы Close и MClose более эффективны, чем алгоритм Apriori и по количеству выявленных строгих ассоциативных правил, и по времени работы. Алгоритм MClose сопоставим с алгоритмом Close по времени работы, однако MClose позволяет значительно уменьшить мощность минимаксного базиса, который формирует алгоритм Close.

2.5 Экспертная группировка признаков как дополнительный прием сокращения числа ассоциативных правил

Ассоциативные правила, извлеченные с применением некоторого алгоритма, представляют ценность для врача, если они могут быть им интерпретируемы.

Все множество извлеченных ассоциативных правил обычно разделяют на три класса [134]: полезные, тривиальные, непонятные.

Полезные правила отражают ранее неизвестную информацию, которая имеет однозначное и логичное объяснение. Такие правила могут быть использованы в клинической практике для принятия врачебных решений.

Тривиальные правила содержат уже известную информацию, которую легко объяснить. Такие правила не дают новых знаний и, вероятнее всего,

уже знакомы врачу. В анализе медицинских данных такими правилами могут являться закономерные симптомы давно изученного заболевания, по которым уже много лет проводят диагностику.

Непонятные правила отражают ранее неизвестную информацию, которая не имеет объяснения. Такие правила чаще всего получаются на основе не связанных между собой значений.

Варьируя значениями δ_0 и γ_0, можно избавиться от тривиальных и непонятных закономерностей. Большое значение величины δ_0 обычно приводит к тому, что будут найдены только тривиальные правила. Чаще всего, наиболее интересные полезные правила выявляются именно при низком пороге поддержки, однако слишком низкое значение этой величины способствует поиску статистически необоснованных правил. Слишком маленькое значение γ_0 влечет за собой образование большого количества непонятных правил, а при больших значениях достоверности большинство правил могут являться тривиальными. Тем не менее, встречаются задачи, при решении которых требуется найти только строгие ассоциативные правила или редкие случаи с низким значением поддержки. Выбор значений δ_0 и γ_0 сильно зависит от решаемой клинической задачи и выполняется на основании рекомендаций врача-эксперта.

Число ассоциативных правил, предъявляемых врачу для интерпретации, можно существенно уменьшить путем внесения экспертных знаний о рассматриваемом множестве признаков. В данном случае полезной может быть любая экспертная группировка признаков.

Например, все множество признаков исходного контекста можно разделить на три группы:

- признаки, которые могут играть роль посылки;
- признаки, которые могут являться только следствием;
- признаки, которые могут выступать как в роли причины, так и в роли следствия.

Такое разделение признаков очень часто возможно при анализе медицинских данных, например, при определении зависимостей между наркотическими препаратами и симптомами, возникающими при их применении наркомании. Разделение множества признаков на группы врачом-экспертом перед началом анализа медицинских данных значительно сокращает количество ассоциативных правил и повышает качество интерпретации выявленных закономерностей.

2.6 Выводы по главе 2

1. Анализ формальных понятий является прикладной ветвью алгебраической теории решеток и математическим аппаратом, позволяющим формализовать понятия, связанные с ассоциативными правилами, а также формировать «сжатое» представление (неизбыточный минимаксный базис) результирующего множества строгих ассоциативных правил.

2. Задача нахождения всех (δ_0, γ_0)-ассоциативных правил в заданном контексте $K = (G, M, I)$ является труднорешаемой, поскольку число выявленных правил в худшем случае может экспоненциально зависеть от числа $|M|$ рассматриваемого множества признаков. При малых значениях δ_0 время решения задачи сопоставимо с $O(2^{|M|})$. Все это приводит к значительным затратам по времени и памяти при поиске ассоциативных правил, и затрудняет экспертный анализ полученных правил.

3. Фильтрация результирующего множества ассоциативных правил с помощью мер значимости в ряде случаев уменьшает число правил, но не решает проблему размерности полностью. После фильтрации остается большое число ассоциативных правил, при этом многие из них избыточные.

4. Существуют различные формальные определения избыточных ассоциативных правил и методы их устранения. Наиболее развиты методы устранения избыточности для строгих ассоциативных правил. Строгие ассоциа-

тивные правила являются наиболее важными для медицинской практики, эти правила имеют достоверность или значимость, равную единице. Устранение этой избыточности с помощью некоторого формального набора выводимостей дает возможность существенно сократить число ассоциативных правил, предъявляемых врачу для их верификации и интерпретации.

5. Доказанные в теореме 2.1 свойства (выводимости) D_1–D_6 дают возможность из одного множества строгих ассоциативных правил выводить многие другие множества строгих ассоциативных правил без дополнительного обращения к контексту. При этом выводимости D_1, D_3, D_4, D_5 сохраняют поддержку: результатом их применения к строгим ассоциативным правилам со значением поддержки не менее δ_0 всегда получаются строгие ассоциативные правила с тем же порогом поддержки.

6. Предложен метода построения неизбыточного минимаксного базиса строгих ассоциативных правил, который заключается в устранении в минимаксном базисе избыточности на основе выводимостей D_1, D_3, D_4, D_5.

7. Разработан алгоритм MClose, являющийся модификацией алгоритма Close. Данный алгоритм дает возможность существенно уменьшать мощность минимаксного базиса, формируемого алгоритмом Close, с сохранением уровня поддержки и достоверности ассоциативных правил без дополнительного сканирования исходного контекста.

8. Предложен прием, позволяющий уменьшить число ассоциативных правил, предъявляемых врачу для интерпретации, за счет привлечения экспертных знаний о рассматриваемом множестве признаков.

Глава 3 Средства снижения размерности матрицы «объект–признак»

Целью когнитивного подхода к анализу данных является систематизация и структуризация информации – представление результатов наблюдений в упрощенном, «сжатом» виде, доступном для понимания и дальнейшего использования. Сжать информацию можно за счет сокращения числа описывающих пациентов признаков. Из всего множества наблюдаемых признаков в этом случае выбираются наиболее значимые диагностические признаки. Другим вариантом «сжатия» данных является сокращение числа рассматриваемых пациентов путем перехода к небольшому количеству типичных электронных клинико-инструментальных «образов» пациентов, формируемых на основе базы клинических данных.

Глава 3 посвящена формированию набора эффективных для клинической диагностики средств снижения размерности матрицы «объект–признак», которые целесообразны для МАСКД. 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. [38–40, 90, 131].

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.

Результаты диссертационного исследования, представленные в главе 3, опубликованы в работах [13, 18, 47].

3.1 Снижение размерности признакового пространства

Проблема снижения размерности признакового пространства весьма актуальна для медицинских аналитических систем клинической диагностики. Существуют обстоятельства, обусловливающие необходимость перехода от большего количества показателей состояния здоровья пациента к значительно меньшему числу признаков. Во-первых, это дублирование информации из-за наличия связей между признаками. Во-вторых, небольшая информативность для отдельных признаков.

Следует отметить, что информативность признака – это понятие относительное. Одна и та же признаковая система может быть информативной при диагностики одного заболевания и совершенно не информативной при диагностике другого [1, 54–59]. Например, проводя дифференциальную диагностику заболевания почек целесообразно использовать одни признаки, а

для диагностики бронхиальной астмы другие. Снижение признакового пространства – это определение наиболее информативных признаков из всего множества наблюдаемых показателей здоровья пациентов. Для клинической диагностики весьма существенно, чтобы данные признаки обладали хорошей распознавательной способностью того или иного заболевания, т. е. были диагностически значимыми.

Задачу сокращения признакового пространства можно сформулировать как задачу комбинаторной оптимизации [78]. Пусть G – множество объектов, M – конечное множество признаков, присущих этим объектам. Для любого объекта $g \in G$ известно его признаковое описание.

Задана совокупность признаковых описаний объектов в виде матрицы «объект–признак» размера $|G| \times |M|$, $Inf(Z)$ – некоторая мера информативности подмножества $Z \subseteq M$.

Требуется найти подмножество $Z^* \subseteq M$ такое что

$$Inf(Z^*) = \max_{Z \subseteq M} \{Inf(Z)\}. \qquad (3.1)$$

В интеллектуальном анализе данных сформулированная задача называется FEATURES SELECTION (селекция признаков) [1, 63, 74, 106, 111]. Следует отметить, что задача FEATURES SELECTION имеет высокую вычислительную сложность, поскольку в худшем случае для анализа всех различных подмножеств $Z \subseteq M$ требуется $O(2^{|M|})$ времени. Задача FEATURES SELECTION обобщается путем задания преобразования $Z = F(M)$, дающего возможность из M формировать новое признаковое пространство Z, $|Z| < |M|$. В такой постановке задачу называют FEATURES EXTRACTION (конструирование или извлечение признаков) [1, 74]. При решении этой задачи формируется множество новых признаков на основе тех, что уже имеются в M. В простейшем случае $Z = F(M)$ – это линейное преобразование. Та или иная версия конкретизации задач FEATURES SELECTION и FEATURES EXTRACTION зависит от определения меры информативности в формуле

(3.1), а также класса допустимых преобразований $Z = F(M)$ для задачи FEATURES EXTRACTION.

Основными методами решения задачи FEATURES EXTRACTION являются методы факторного анализа и метод экстремальной группировки признаков.

Факторный анализ выделяет обобщенные признаки (или факторы), каждый из которых объединяет сразу нескольких исходных признаков. Одним из таких методов является метод главных компонент [1, 23, 52]. Суть этого метода состоит в поиске возможных линейных комбинаций признаков из M и конструирования на их основе пространства признаков $Z \subseteq M$ меньшего по мощности, информативность которого равнозначна информативности всех признаков из M. Основной недостаток метода главных компонент – полученные линейные комбинации исходных признаков, как правило, трудно интерпретируемы в клинической практике.

Суть метода экстремальной группировки заключается в вычислении корреляционной матрицы по исходной матрице «объект–признак» и разбиении множества M на группы таким образом, чтобы внутри одной группы между признаками была сильная корреляция, при этом между группами наблюдалась сравнительно слабая корреляция. Далее производится замена каждой такой группы одним равнодействующим признаком. Главным недостатком данного метода также как метода главных компонент является трудность трактовки полученных групп признаков [1, 38].

Приведенные выше методы решают задачу FEATURES EXTRACTION и основаны на конструировании обобщенных признаков. Однако из-за трудности или даже невозможности в ряде случаев интерпретировать полученные результаты в МАСКД лучше всего включать методы, обладающие хорошей объяснительной способностью. Такими являются статистические методы Шеннона и Кульбака [23, 33, 55]. Эти методы предназначены для решения более простой задачи FEATURES EXTRACTION – задачи отбора наиболее информативных признаков на основе заданной меры информативности. Они

основаны на простых процедурах вычисления меры информативности и составляют математический базис некоторых алгоритмов решения FEATURES SELECTION, применяемых с целью сокращения полного перебора в (3.1).

Приведем описание методов Шеннона и Кульбака [23, 33, 55]. Обозначим через T исходную матрицу «объект–признак». Оценка информативности признаков традиционно рассматривается для набора распознаваемых образов (или классов объектов). Предполагается, что объекты, входящие в различные классы, обладают одним и тем же набором признаков M, и каждый из них может входить только в один класс. Например, для двух классов матрица T представляется в виде матриц T_1 и T_2 с одним и тем же набором столбцов.

В этом случае задача FEATURES SELECTION формулируется следующим образом: для заданных T_1, T_2 и меры информативности $Inf(M)$ требуется вычислить $Inf(m)$ для каждого признака $m \in M$, и найти подмножество признаков из M, которое в наибольшей степени объясняют различие между заданными классами.

В методе Шеннона мерой информативности некоторого количественного признака m является средневзвешенное количество информации, свойственное анализируемому признаку. Значение информативности признака m вычисляется по формуле:

$$Inf(m) = 1 + \sum_{i=1}^{q} (P_i \cdot \sum_{k=1}^{2} p_{ik} \cdot \log_2(p_{ik})) . \qquad (3.2)$$

где q – число градаций признака; $k = 1, 2$ – номер класса; P_i – вероятность попадания значения признака в i-ю градацию

$$P_i = \frac{\sum_{k=1}^{2} ch_{ik}}{N}, \qquad (3.3)$$

где ch_{ik} – частота появления значения признака в i-ой градации для матрицы T_k, N – общее число признаков, входящих в T_1 иT_2; p_{ik} – вероятность появления значения признака в i-ой градации

$$p_{ik} = \frac{ch_{ik}}{ch_{i1} + ch_{i2}},\ k = 1, 2. \tag{3.4}$$

Отметим, что метод Шеннона позволяет получить меру информативности некоторого анализируемого признака в виде величины, которая вычисляется с помощью формул (3.2)–(3.4) и принимает значения в интервале от нуля до единицы. Чем ближе значение величины $Inf(m)$ к единице, тем больше информативность признака m, и наоборот, чем ближе значение $Inf(m)$ к нулю, тем меньше информативность этого признака.

В методе Кульбака мерой информативности некоторого количественного признака m является дивергенция Кульбака. Эта величина отражает степень расхождения между двумя классами:

$$Inf(m) = \sum_{i=1}^{q} (p_{i1} - p_{i2}) \cdot \log_2 \frac{p_{i1}}{p_{i2}}. \tag{3.5}$$

где q – число градаций признака; p_{ik} – вероятность попадания значения признака в i-ую градацию.

$$p_{ik} = \frac{ch_{ik}}{ch_{i1} + ch_{i2}},\ k = 1, 2 \tag{3.6}$$

где ch_{ik} – частота появления значения признака в i-ой градации матрицы T_k.

Метод Кульбака позволяет оценить информативность признака величиной, вычисляемой по формулам (3.5), (3.6). Чем выше значение величины $Inf(m)$, тем больше информативность признака m.

3.2 Снижение числа анализируемых объектов

Уменьшение размерности матрицы «объект–признак» можно также осуществить за счет сокращения числа рассматриваемых объектов путем перехода от всего множества объектов к небольшому количеству его типичных представителей. Считается, что все множество объектов разбито на классы.

Для поиска типичных и нетипичных случаев (пациентов, препаратов или симптомов) применяются предложенные Н.Г. Загоруйко и его учениками функции конкурентного сходства и различия (FRiS-функции) и алгоритмы, основанные на них [32, 38–40, 90, 131].

Специфика FRiS-функции состоит в том, что мерой сходства является относительная величина, зависящая не только от сходства некоторого объекта g с представителем a из класса T_1, но и от его различия с ближайшим к a представителем b из другого (конкурирующего) класса T_2.

FRiS-функция – мера, оценивающая сходство объекта g с объектом a в конкуренции с объектом b вычисляется следующим образом:

$$F(g, a \mid b) = (r(g, b) - r(g, a)) / (r(g, b) + r(g, a)),$$

где $r(g, a)$ – расстояние от объекта g до объекта a, $r(g, b)$ – расстояние от объекта g до объекта b. Значения функции $F(g, a \mid b)$ изменяются в интервале от 1 до –1. В случае если анализируемый объект g совпадал с представителем класса T_1, то $-r(g, a) = 0$ и $F = 1$. Это говорит о полном сходстве объекта g с представителем класса T_1 и о максимальном его отличии от представителя класса T_2.

Нахождение множества типичных представителей классов можно осуществить известным алгоритмом FRiS-Stolp. Данный алгоритм допускает любое конечное число классов. Алгоритм FRiS-Stolp подробно описан в работе [38]. Замена каждого класса множеством его типичных представителей позволяет снижать число строк матрицы «объект–признак». В клинической практике типичными представителями могут выступать пациенты с наиболее типичными симптомами.

3.3 Алгоритм ELIMINATION

Алгоритм ELIMINATION реализует методы Шеннона и Кульбака, аппарат FRiS-функций, а также процедуры классификации и оценки качества

классификации на основе ROC-анализа. Основное назначение данного алгоритма – снижение размерности матрицы «объект–признак» с целью уменьшения числа искомых ассоциативных правил, извлекаемых из этой матрицы.

Алгоритм ELIMINATION находит приближенное решение задачи FEATURES SELECTION за полиномиальное время. Алгоритм последовательно удаляет (elimination) из M наименее информативные признаки.

Задачи клинической диагностики – это по существу задачи классификации (например, «Здоровая почка» или «Имеются множественные кисты»). Поэтому условием окончания процесса удаления признаков служит качество классификации, оценка которого выполняется на обучающих выборках (на заданных матрицах T_1 и T_2). Алгоритм ELIMINATION для оценки информативности каждого отдельного признака использует метод Шеннона или Кульбака [47].

Алгоритм ELIMINATION выполняет следующие действия:

- вначале производится расчет информативности каждого признака из M, затем признаки сортируются в порядке убывания рассчитанного значения меры информативности и оформляются в виде списка Z;
- из списка Z последовательно снизу вверх удаляются наименее информативные признаки (на каждом шаге по одному признаку);
- на основе оставшихся признаков осуществляется бинарная классификация объектаg, выбранного из $T_1 \cup T_2$. Выбор g выполняется методом «скользящего окна»;
- после этого вычисляются показатели ROC-анализа для оценки качества выполненной классификации, т. е. оценки правильности распознавания класса для g. Если показатели ROC-анализа демонстрируют приемлемое качество классификации, то действие алгоритма останавливается. В противном случае из Z удаляется следующий признак, при этом ранее удаленные признаки в Z не возвращаются.

Существует большое количество методов классификации, обладающих различной сложностью и эффективностью [1, 76, 102, 118].

При выборе метода классификации следует учитывать тип данных, для которого применяется данный алгоритм. В анализе медицинских данных приходится иметь дело с разнотипными признаками. В этом случае целесообразно применение простейшего метода бинарной классификации – метода ближайшего соседа [1].Решение о том, к какому классу (первому или второму) следует отнести объектg, в методе ближайшего соседа принимается простым голосованием. При одинаковом количестве голосов происходит отказ от классификации.

Заметим, что определение класса принадлежности объекта g в алгоритме ELIMINATION вначале осуществляется по каждому отдельному признаку, а затем по всем признакам Z в целом: каждый признак голосует за тот класс, к которому необходимо отнести g, при равенстве голосов происходит отказ от классификации. Качество классификации оценивается процентами верного распознавания класса, ошибочного указания класса и отказов от классификации. Такие оценки позволяет получить ROC-анализ [23, 94].

В ROC-анализе исследуются верно классифицированные положительные и неверно классифицированные отрицательные случаи, при этом первые называются истинно положительными, вторые – ложно отрицательными.

Процент истинно положительных случаев (TruePositivesRate) определяется формулой:

$$TPR = 100\% \cdot TP / (TP + FN), \qquad (3.7)$$

где TP (TruePositives) – число истинно положительных случаев, FN (FalseNegatives) – положительные случаи, классифицированные как отрицательные (так называемые ложно отрицательные случаи). В алгоритме ELIMINATION значение TPR вычисленное по формуле (3.7), сравнивается с заданным пороговым значением τ: если $TPR \geq \tau$, то дальнейшее исключение признаков не целесообразно и действие алгоритма ELIMINATION останавливается; в противном случае удаление признаков продолжается.

Процент отказов R вычисляется следующим образом:

$$R = 100\% \cdot r / |T_1 \cup T_2|,$$

где r – количество отказов от классификации, $|T_1 \cup T_2|$– общее количество объектов, для которых была произведена классификация. При оценке качества классификации в алгоритме ELIMINATION используются также значения показателей чувствительности SE и специфичности SP.

Чувствительность отражает процент истинно положительных случаев и вычисляется по формуле (3.7). Показатель чувствительности важен в гипердиагностике – максимальном предотвращении пропуска больных. Специфичность показывает процент истинно отрицательных случаев, которые были правильно классифицированы:

$$SP = 100\% \cdot TN / (TN + FP),$$

где TN (TrueNegative) – число верно классифицированных отрицательных случаев, FP (FalsePositive) – отрицательные случаи, классифицированные как положительные (ложно положительные случаи). Чем выше значения чувствительности и специфичности, тем лучше качество классификации.

Таким образом, алгоритм ELIMINATION как комбинация известных алгоритмов, позволяет выбрать подмножество $Z^* \subseteq Z$ наиболее значимых диагностических признаков – подмножество признаков, позволяющих с заданным качеством классифицировать объекты из $T_1 \cup T_2$. При этом не возникают новые признаки, которые подлежат дополнительной интерпретации. Очевидно, что изменение обучающих выборок T_1 и T_2 может приводить к различным подмножествам Z^*.

В подразделе 4.3 диссертационной работы представлены результаты численных экспериментов по оценке результативности алгоритма ELIMINATION на двух задачах клинической диагностики: определение минимального набора признаков для распознавания множественной лекарственной устойчивости возбудителя туберкулеза легких; нахождение диагностически значимых признаков для выявления сепсиса.

3.4 Выводы по главе 3

1. Снизить размерность матрицы «объект–признак» можно путем селекции и отбора наиболее значимых диагностических (информативных) признаков и перехода от всего множества объектов к небольшому количеству их типичных представителей.

2. Задача FEATURES SELECTION (селекция признаков) имеет высокую вычислительную сложность и, как правило, решается методами конструирования обобщенных признаков, т. е. методами факторного анализа и методом экстремальной группировки признаков. Однако из-за трудности или даже невозможности в ряде случаев интерпретировать полученные результаты в МАСКД лучше всего включать методы, обладающие хорошей объяснительной способностью. Такими являются статистические методы Шеннона и Кульбака. Эти методы предназначены для решения более простой задачи FEATURES EXTRACTION – задачи отбора наиболее информативных признаков на основе заданной меры информативности.

3. Набор типичных представителей заданного множества объектов можно формировать с помощью FRiS-функций.

4. Применение в МАСКД комбинации статистических методов Шеннона и Кульбака и аппарата FRiS-функций уменьшает размерность матрицы «объект–признак», что является вспомогательным средством для сокращения числа извлекаемых ассоциативных правил.

Глава 4 Программное обеспечение и результаты экспериментальных исследований

Разработанные в диссертационной работе метод и алгоритмы выявления строгих ассоциативных правил и их «сжатого» представления в виде неизбыточного минимаксного базиса, снижения размерности матрицы «объект–признак» реализованы в виде комплекса программных модулей с целью проведения экспериментальных исследований на клинических данных и оценки результативности предложенных средств. В подразделе 4.1 представлен состав комплекса программных модулей и схема их взаимодействия.

В подразделе 4.2 приведены результаты анализа диагностики наркозависимости с применением ассоциативных правил и экспертной группировки признаков и симптомов. В подразделе 4.3 приведен анализ результативности алгоритма ELIMINATION, предназначенного для снижения размерности матрицы «объект–признак» на двух задачах клинической диагностики: определение минимального набора признаков для распознавания множественной лекарственной устойчивости возбудителя туберкулеза; нахождение диагностически значимых признаков для выявления сепсиса.

Результаты диссертационного исследования, представленные в главе 4, опубликованы в работах [5, 12, 13, 18, 19, 46–50].

4.1 Состав программных модулей и схема их взаимодействия

Функциональные возможности разработанного комплекса программных модулей определены существующей процедурой постановки диагноза. Диагностика заболеваний врачом, как правило, реализуется в три этапа [69].

Этап 1. Вначале анализируются жалобы, анамнез жизни и заболевания больного. Если жалобы и особенности течения болезни четко отвечают определенной нозологической единице, то на следующих этапах ее следует подтвердить. Если описанные больным симптомы встречаются при различных

заболеваниях, то определяется набор сходных болезней и диагноз может быть поставлен только после второго или даже третьего этапов. Если данные анамнеза не характерны ни для какого-либо определенного заболевания, то выполняется второй этап.

Этап 2 (дифференциальная диагностика). Производится непосредственное обследование больного: осмотр, пальпация, перкуссия, аускультация. Анализ полученной информации может приводить к различным выводам: диагноз полностью определился; круг заболеваний сузился; диагностической концепции пока нет.

Этап 3 (клиническая диагностика). Выполняются лабораторные и инструментальные обследования больного. Возможные исходы: предполагаемый диагноз подтверждается; диагноз остается неясным; необходимо дальнейшее наблюдение больного.

При таком механизме постановки диагноза не исключены врачебные ошибки. Информатизация клинической работы врачей и медицинские знания, извлеченные из накопленных баз клинических данных, способствуют снижению процента врачебных ошибок и назначению эффективного лечения. Разработанный комплекс программных модулей функционально ориентируется исключительно на информатизацию клинической работы врачей и опирается на конкретный электронный клинико-инструментальный «образ» пациента, формируемый на основе базы клинических данных.

Извлечение из клинических данных множества строгих ассоциативных зависимостей и построение его неизбыточного представления – одна из основных функций разработанных программных средств. Установленные зависимости дают возможность учитывать, например, на фоне каких причин может развиться рассматриваемое заболевание и с какими другими заболеваниями оно может встречаться совместимо, то есть какие заболевания (или синдромы) встречаются одновременно.

Следует отметить, что современная нозологическая номенклатура, которой должен оперировать врач, включает около 10 тысяч болезней и около 100 тысяч симптомов. В такой ситуации важна функция по снижению размерности матрицы «объект–признак» путем уменьшения количества анализируемых признаков и перехода к типичным представителям. Данная функция предусматривает возможность определения набора наиболее значимых диагностических признаков, а также выявления типичных пациентов со свойственным им набором симптомов заболевания. Кроме значимости этой функции с точки зрения клинической диагностики, также важно ее влияние на сжатие выходных данных – число ассоциативных правил, представляемых врачу для верификации и интерпретации.

Состав основных программных модулей и их функциональное назначение приведены в таблице 4.1, а схема их взаимодействия изображена на рисунке 4.1.

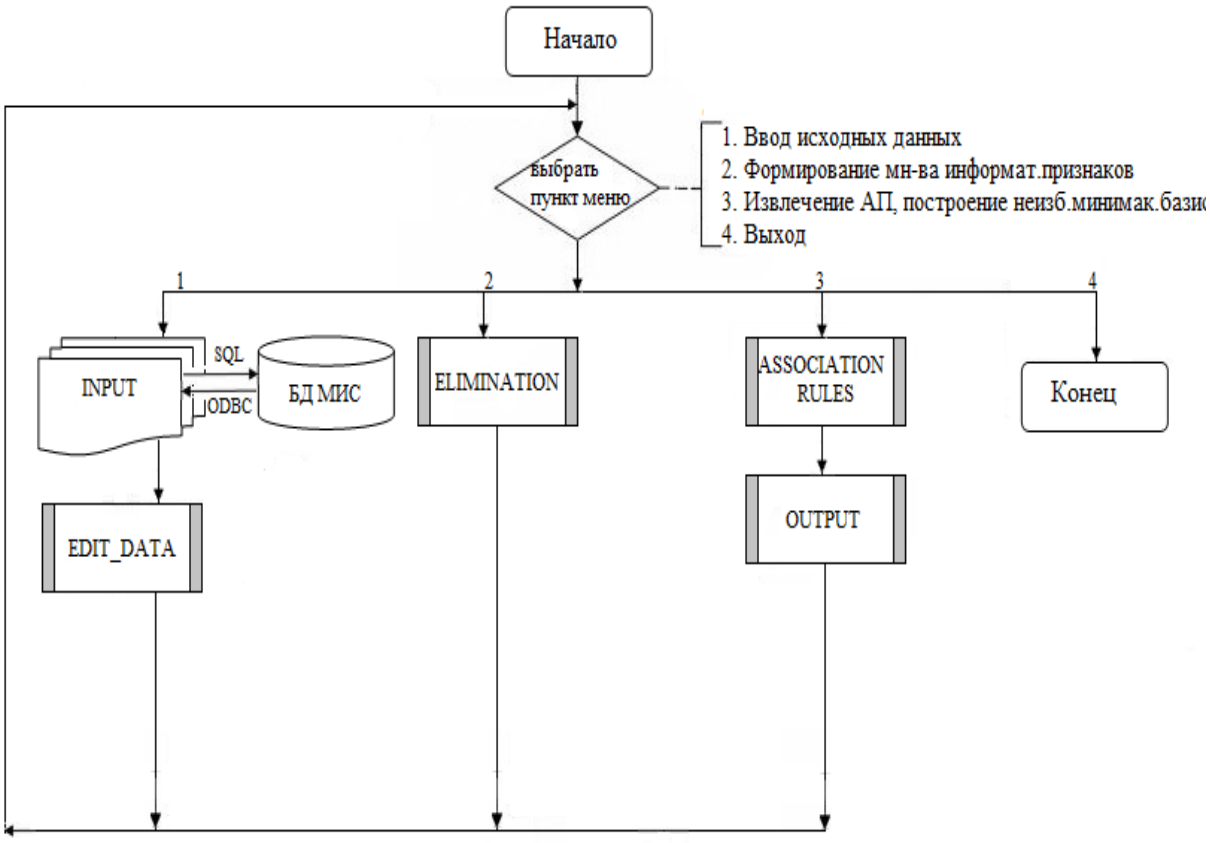


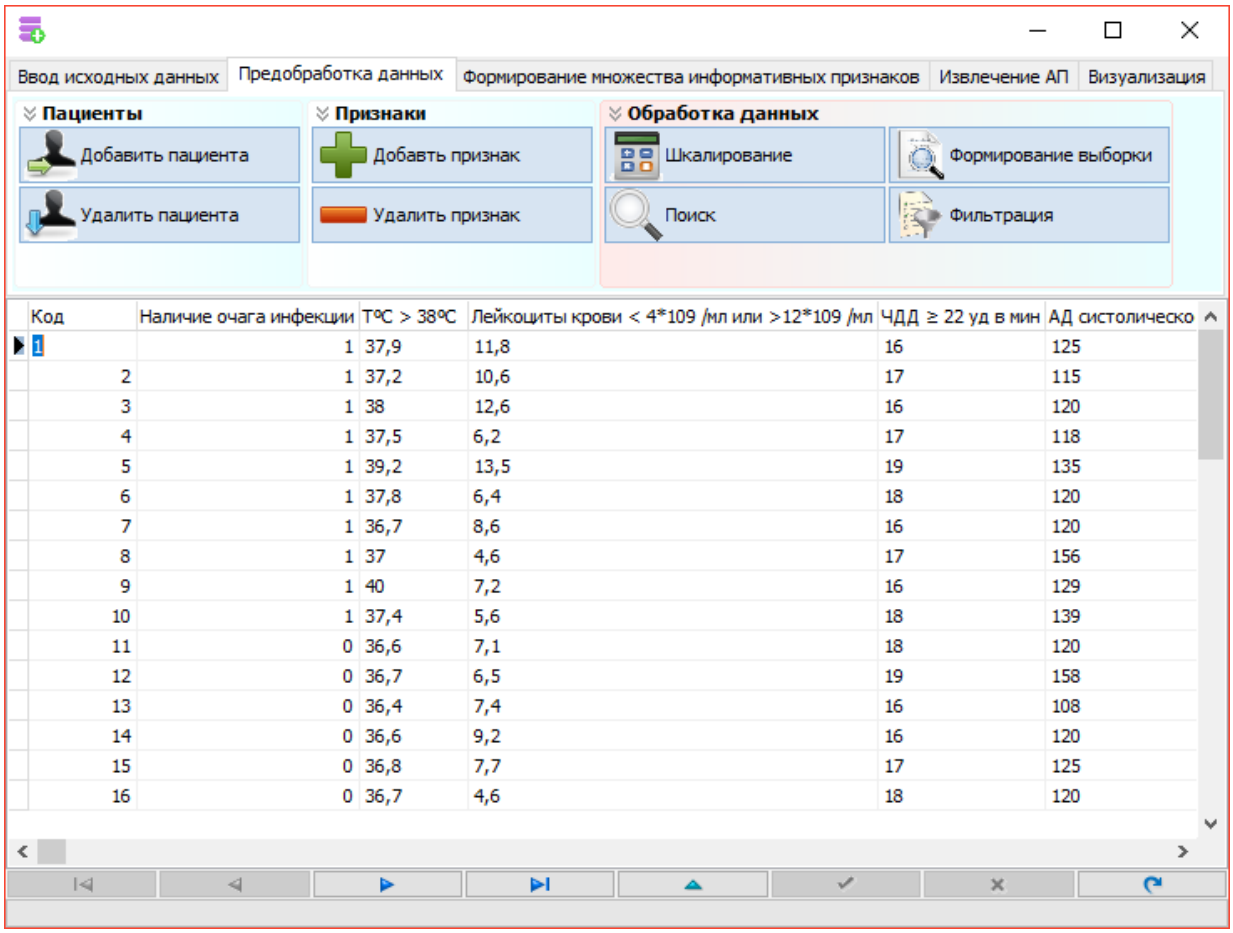
Рисунок 4.1 – Схема взаимодействия модулей

Созданный комплекс программных модулей может служить программной основой для создания медицинских аналитических систем клинической диагностики, ориентированных на конкретные нозологические формы заболеваний. Для этого достаточно формирование соответствующей базы данных и знаний. В реализованной версии комплекс программ жестко не привязан к каким-либо базам медицинских данных.

Таблица 4.1 – Состав программных модулей

Название модуля	Идентификатор модуля	Выполняемые действия
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, фильтрация 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	ELIMINATION	Производит расчет информативности признаков по выбранному методу (Шеннона или Кульбака). Выполняет оценку качества отбора признаков с помощью показателей ROC-анализа, классификацию признакового описания пациента по целевому признаку
Модуль извлечения ассоциативных правил и построения неизбыточного минимаксного базиса	ASSOCIATION_RULES	Выполняет поиск ассоциативных правил с помощью выбранных алгоритмов (Apriori, Closeили 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

Модули INPUT и EDIT_DATA осуществляют ввод, редактирование и предварительную обработку признаковых описаний пациентов. Предварительная обработка данных включает следующие действия: поиск, фильтрация и шкалирование данных, добавление и удаление признаков и пациентов, формирование обучающих выборок в виде матриц «объект–признак». Взаимодействие пользователя с этими модулями осуществляется через интерфейс, представленный на рисунке 4.2.



Код	Наличие очага инфекции	T°C > 38°C	Лейкоциты крови < 4*109 /мл или >12*109 /мл	ЧДД ≥ 22 уд в мин	АД систолическо
1	1	37,9	11,8	16	125
2	1	37,2	10,6	17	115
3	1	38	12,6	16	120
4	1	37,5	6,2	17	118
5	1	39,2	13,5	19	135
6	1	37,8	6,4	18	120
7	1	36,7	8,6	16	120
8	1	37	4,6	17	156
9	1	40	7,2	16	129
10	1	37,4	5,6	18	139
11	0	36,6	7,1	18	120
12	0	36,7	6,5	19	158
13	0	36,4	7,4	16	108
14	0	36,6	9,2	16	120
15	0	36,8	7,7	17	125
16	0	36,7	4,6	18	120

Рисунок 4.2 – Интерфейс для ввода и предобработки исходных данных

Модуль ELIMINATION является программной реализацией одноименного алгоритма, представленного в подразделе 3.3 диссертационной работы. В нем выполняется отбор информативных признаков. Интерфейс модуля ELIMINATION представлен на рисунке 4.3. В модуле допускается выбор метода расчета информативности признаков по методу Шеннона или методу Кульбака. Для определения диагностически значимых признаков врачу необ-

ходимо определить исходный состав признаков, задать целевой признак, по которому будет производиться разделение пациентов на группы, и выбрать метод расчета информативности.

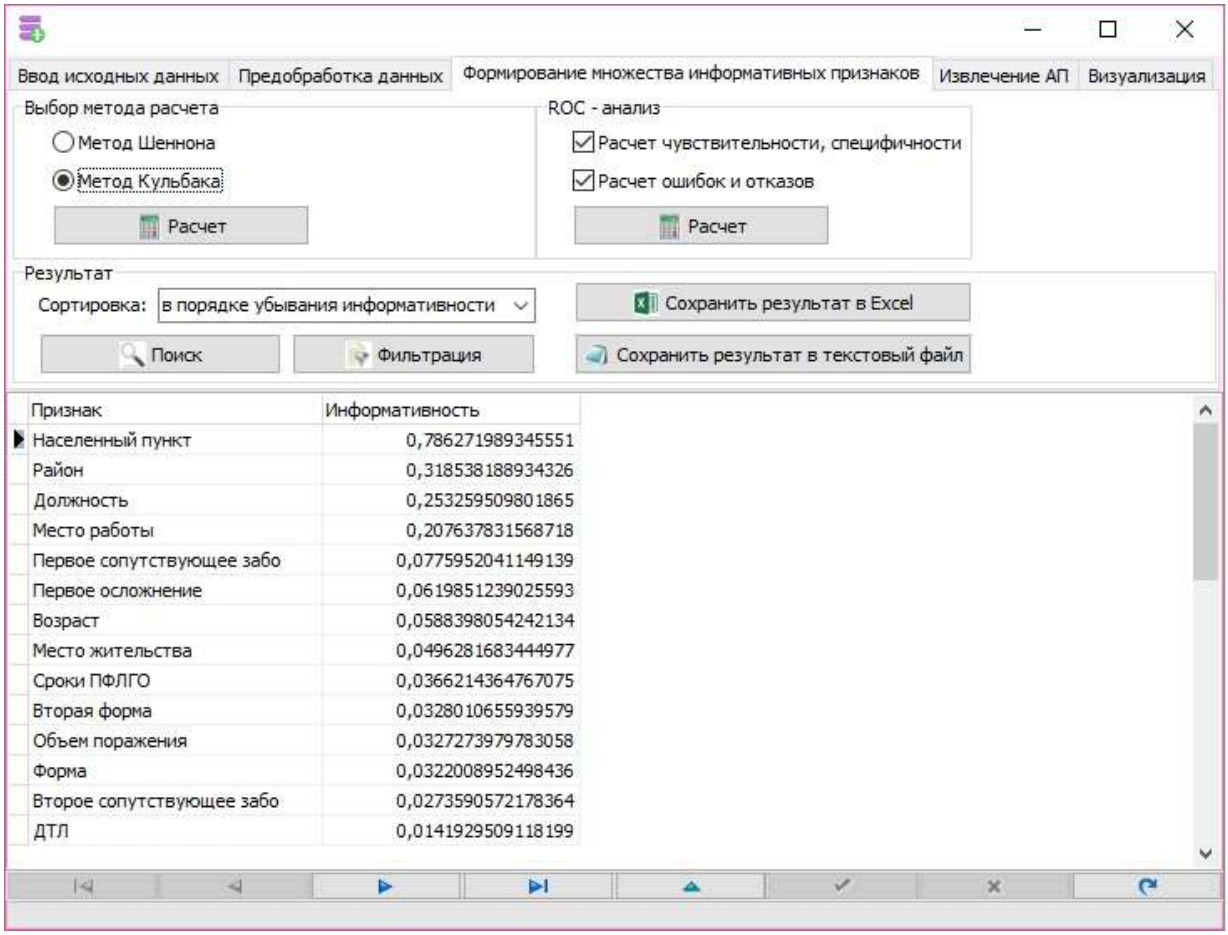


Рисунок 4.3 – Интерфейс модуля расчета информативности признаков

Модуль ASSOCIATION_RULES позволяет извлекать ассоциативные правила с помощью выбранных алгоритмов (Apriori, Close или MClose), строить неизбыточный минимаксный базис строгих ассоциативных правил с помощью алгоритма MClose, осуществлять экспертную группировку исходных признаков (рисунок 4.4). Детальное описание алгоритма MClose приведено в подразделе 2.4 диссертации.

Модуль OUTPUT выводит результирующие данные в стандартных форматах Microsoft Excel или TXT, а также различных графических формах (круговых диаграмм, гистограмм и пр.). Возможна печать полученных результатов в виде таблиц.

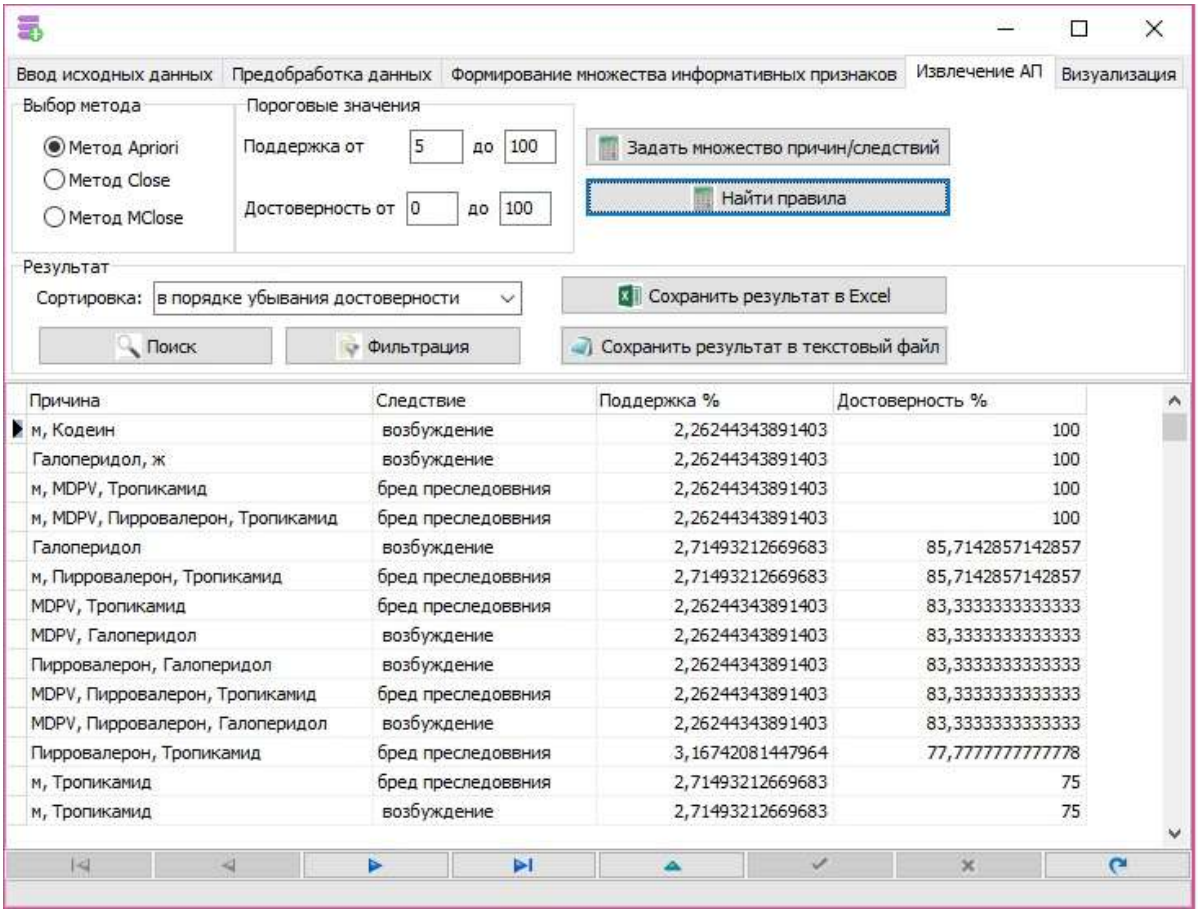


Рисунок 4.4 – Интерфейс модуля ASSOCIATION_RULES

Разработанный комплекс программных модулей реализован на языке программирования C++ в среде разработки Embarcadero RAD Studio XE 8 [64]. Для его работы необходим компьютер с операционной системой Windows XP и выше.

Минимальные технические требования: X86-совместимый процессор, объем ОЗУ 1 Гб, 10 Мб свободного места на жестком диске. Расширение разработанного комплекса модулей до самостоятельной медицинской аналитической системы потребует изменения технических характеристик в части увеличения вычислительных мощностей.

4.2 Анализ диагностики наркозависимости с применением ассоциативных правил

Правильная и своевременная диагностика состояния наркотического опьянения необходима для выбора лечебного воздействия, социальной реабилитации и предупреждения развития наркотической зависимости. Анализ признаков наркотического опьянения конкретного пациента предполагает оценку его психического состояния, соматовегетативных и неврологических признаков, и определения препаратов, возможно принятых пациентом. Существуют специфические симптомы (некоторые модели) употребления отдельных наркотических веществ, по которым еще до проведения лабораторных исследований можно определить, что именно принял пациент. Однако чаще всего реальное состояние пациента не совпадает с имеющимися моделями, поскольку это состояние обусловлено принятием нескольких наркотических препаратов. Кроме того, состав наркотических веществ постоянно расширяется и поэтому не существует постоянного набора моделей, отражающих поведение пациент в зависимости от принятого им препарата. Во всех этих ситуациях для правильной диагностики врачу-наркологу могут быть полезны строгие ассоциативные зависимости, выявленные из медицинских данных пациентов, проходивших ранее лечение в наркологическом диспансере.

Первая серия вычислительных экспериментов выполнялась с целью сравнения алгоритмов Apriori, Close и MClose по количеству сгенерированных ими строгих ассоциативных правил и времени их работы. Использовалась база медицинских данных пациентов, проходивших лечение в Красноярском краевом наркологическом диспансере № 1 в 2016 году (222 пациента, 64 признака). Состав рассматриваемых признаков, включая демографические признаки, применяемые препараты и наблюдаемые симптомы, приведен в таблице 4.2.

Таблица 4.2 – Признаки для генерации ассоциативных правил

Демографические признаки и принимаемые препараты	Наблюдаемые симптомы
1	2
Возраст, пол мужской, пол женский, ТГК, MDPV, Морфин, Этанол, Димедрол, Наркотических веществ не обнаружено, Карбокситетра канабинол, ПВП, Тропикамид, Фенобарбитал, Каннабидиол, Галоперидол, Декстрометорфан, Лидокаин, Метронидазол, 3,4 метилендиоксипиралиндинобутирофенон, Атропин, Карбамазепин, Кодеин, Доксиламин, Дектрорфан, Хлорфенамин, 6моноацетилморфин, Бензодеазепин, Прометазин, Амфетамин, Исследование не проводилось, Амитриптилин, Каннабиноиды, Пирровалерон, Другие вещества	Бред преследования, Страх, Повреждение предметов, Возбуждение, Угроза жизни (убить могут), Речь (ответы не по сути), Галлюцинации зрительные, Галлюцинации слуховые, Попытки подбросить наркотики, Воздействие током, Защита от преследователей (прячется), Защита от преследователей (нападает), Идеи самоуничижения, Агрессия к окружающим, Самоповреждения шантажные, Самоповреждения депрессивные, Тревога, Отсутствие сна, Нарушения сознания (дезориентировка), Отсутствие контакта, Психосенсорные расстройства, Отсутствие одежды или снимает, Идеи отравления, Суицидальные мысли, Сенестопатии вынимают или выпадают внутренности, На теле жучки, Судороги, Идеи величия, Двигательное возбуждение, Бег, Желание спрыгнуть с высоты

Число пациентов, исследуемых в ходе вычислительного эксперимента, постепенно увеличивалось от 50 до 222. В таблице 4.3 представлены результаты работы алгоритма Apriori при $\delta_0 = 0{,}1$. Были рассмотрены два случая: с разделением множества признаков на группы и без разделения. В первом случае посылками ассоциативных правил являлись только признаки из группы 1, а следствием – из группы 2, а во втором случае все множество исходных признаков могло встречаться и в посылках, и в следствиях. Разделение множества признаков было осуществлено врачом-наркологом: к группе 1 были отнесены демографические признаки и принимаемые препараты, к группе 2 – наблюдаемые симптомы.

Таблица 4.3 – Результаты работы алгоритма Apriori

Число пациентов	Без разделения множества признаков на группы		Число частых множеств	С разделением множества признаков на группы	
	Число строгих АП	Время вычислений, мс		Число строгих АП	Время вычислений, мс
1	2	4	5	6	7
50	466236	256110	9185	2583	10560
100	254381	107000	6237	2049	8700
150	30032	45000	4761	533	3920
222	6294	43000	3607	122	1154

Из таблицы 4.3 видно, что экспертная группировка признаков существенно уменьшает число извлекаемых строгих ассоциативных правил, предъявляемых врачу для интерпретации, время их нахождения зависит от числа частых множеств. В таблицах 4.4, 4.5 представлены результаты работы алгоритмов Apriori, Close и MClose, предназначенных для извлечения ассоциативных правил.

Таблица 4.4 – Сравнение алгоритмов по числу результирующих правил

Число пациентов	Apriori	Close	MClose
	Число строгих АП	Число строгих АП, составляющих минимаксный базис	Число строгих АП, составляющих неизбыточный минимаксный базис
50	2583	32	5
100	2049	28	4
150	553	46	8
222	122	17	5

Таблица 4.5 – Сравнение алгоритмов по времени работы

Число пациентов	Apriori	Close	MClose
	Время, мс	Время, мс на формирование минимаксного базиса	Время, мс на формирование неизбыточного минимаксного базиса
50	10560	4803	4034
100	8700	3264	3056
150	3920	1176	1075
222	1154	867	695

Если алгоритм Apriori генерирует ассоциативные правила с заданными пороговыми значениями поддержки и достоверности, то алгоритмы Close и MClose формируют минимаксный и неизбыточный минимаксный базис для строгих ассоциативных правил соответственно. Данные алгоритмы сравнивались по количеству сгенерированных строгих ассоциативных правил и по времени работы. Для всех алгоритмов пороговое значение поддержки равнялось $\delta_0 = 0{,}1$. Множество признаков было разделено на группы: к группе 1 были отнесены демографические признаки и принимаемые препараты, к группе 2 – наблюдаемые симптомы. Данные, представленные в таблицах 4.4 и 4.5, иллюстрируют рисунки 4.5 и 4.6.

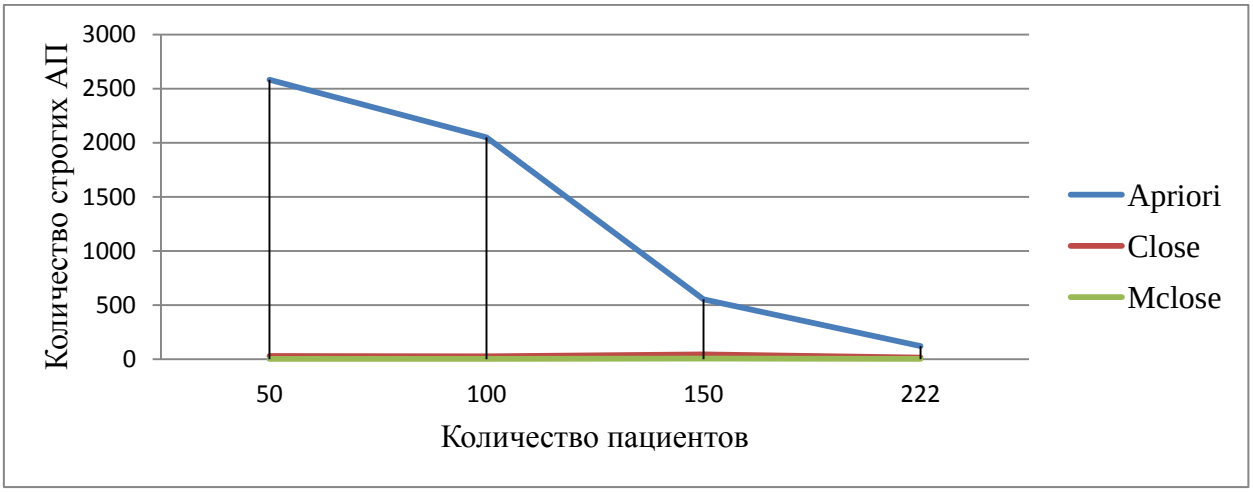


Рисунок 4.5 – Число результирующих строгих ассоциативных правил, формируемых сравниваемыми алгоритмами

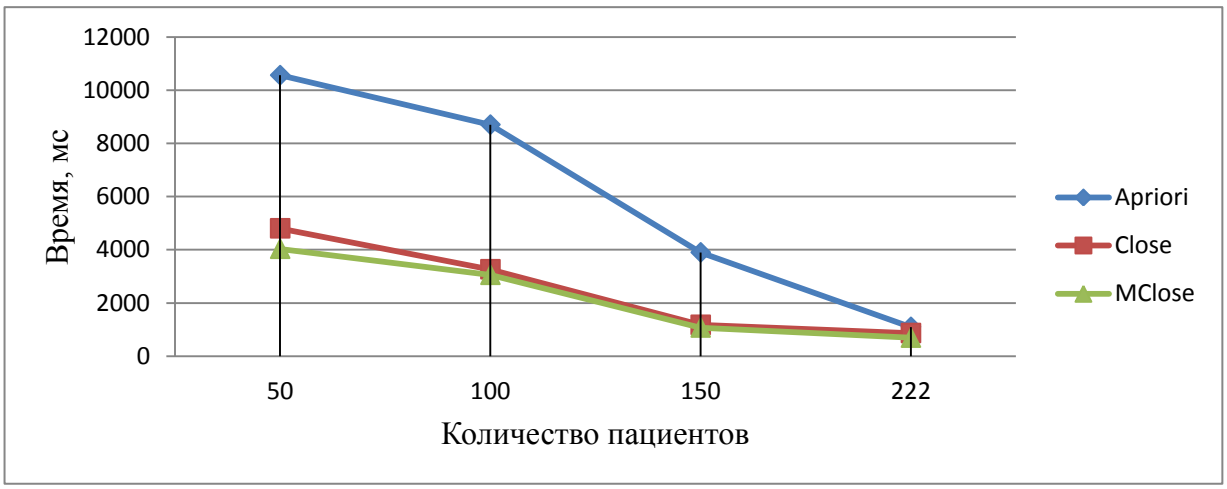


Рисунок 4.6 – Время, затраченное сравниваемыми алгоритмами для формирования множества результирующих строгих ассоциативных правил

Из представленных выше результатов экспериментов следует, что

– для исходного набора данных (222 пациента, 64 признака) алгоритм Apriori выявил 126 строгих ассоциативных правил;
– алгоритм Close построил минимаксный базис, состоящий из 17 правил (таблица 4.6);
– алгоритм MClose удалил из минимаксного базиса 12 избыточных и сформировал неизбыточный минимаксный базис из 5 правил (в таблице 4.6 они выделены подчеркиванием).

Таблица 4.6 – Минимаксный базис строгих ассоциативных правил

Посылка ассоциативного правила		Следствие ассоциативного правила	Поддержка
{Этанол}	⇒	**{Страх, Угроза жизни (убить могут)}**	0,221
{Галоперидол}	⇒	**{Возбуждение}**	0,217
{Галоперидол, MDPV}	⇒	{Возбуждение}	0,186
{Галоперидол, Пирровалерон}	⇒	{Возбуждение}	0,172
{Галоперидол, Пол ж}	⇒	{Возбуждение}	0,162
{MDPV, ТГК}	⇒	**{Попытки подбросить наркотики}**	0,149
{Пирровалерон, ТГК}	⇒	**{Попытки подбросить наркотики}**	0,144
{Пол ж, ТГК}	⇒	**{Тревога}**	0,113
{Галоперидол, MDPV, Пирровалерон}	⇒	{Возбуждение}	0,108
{Галоперидол, MDPV, Пол ж}	⇒	{Возбуждение}	0,11
{Галоперидол, Пирровалерон, Пол ж}	⇒	{Возбуждение}	0,104
{Каннабиноиды, MDPV, ТГК}	⇒	{Попытки подбросить наркотики}	0,186
{Каннабиноиды, Пирровалерон, ТГК}	⇒	{Попытки подбросить наркотики}	0,149
{Каннабиноиды, Пол ж, ТГК}	⇒	{Тревога}	0,113
{MDPV, Пирровалерон, ТГК}	⇒	{Попытки подбросить наркотики}	0,104
{Галоперидол, MDPV, Пирровалерон, Пол ж}	⇒	{Возбуждение}	0,162
{Каннабиноиды, MDPV, Пирровалерон, ТГК}	⇒	{Попытки подбросить наркотики}	0,108

С точки зрения врача-нарколога каждое выявленное алгоритмом MClose минимаксное ассоциативное правило из таблицы 4.6 в действительности определяет:

- известную зависимость между принимаемыми препаратами и возможными последствиями от их применения (специфическими симптомами). Такая зависимость подтверждает ранее известные знания. Например, {Галоперидол} ⇒ {Возбуждение};
- неизвестную ранее зависимость, которая имеет логическое объяснение. Например, {Пирровалерон, ТГК} ⇒ {Попытки подбросить наркотики}. Такая зависимость несет некоторые новые знания;
- неизвестную ранее зависимость, которая пока не имеет логическое объяснение и требует дополнительной экспертной оценки. Например, {Другие вещества, Пол ж ⇒ {MDPV, Возбуждение}.

Всякое избыточное минимаксное ассоциативное правило из таблицы 4.6 может получено с помощью процедуры SX, являющей составляющей алгоритма MClose, и выводимостей D_1, D_3, D_4, D_5.

Покажем это для строгого ассоциативного правила:

{Галоперидол, MDPV, Пирровалерон} ⇒ {Возбуждение}.

Действительно, согласно D_1 получаем:

{Галоперидол, MDPV, Пирровалерон} ⇒ {Галоперидол, MDPV, Пирровалерон}.

Так как {Галоперидол} ⊆ {Галоперидол, MDPV, Пирровалерон}, то по D_4 имеем:

{Галоперидол, MDPV, Пирровалерон} ⇒ {Галоперидол}.

Согласно D_5 верно:

если {Галоперидол, MDPV, Пирровалерон} ⇒ {Галоперидол} и **{Галоперидол} ⇒ {Возбуждение},** то

{Галоперидол, MDPV, Пирровалерон} ⇒ {Возбуждение}.

Для установления предполагаемого набора принятых препаратов врачу-наркологу требуется задать набор показателей состояния пациента, полученных по результатам дифференциальной диагностики. Этот набор рассматривается в качестве причины, для которой формируются возможные комбинации принятых препаратов. Например, на рисунке 4.7 показаны возможные комбинации принятых препаратов при симптоматике {Суицидальные мысли, Идеи отравления}.

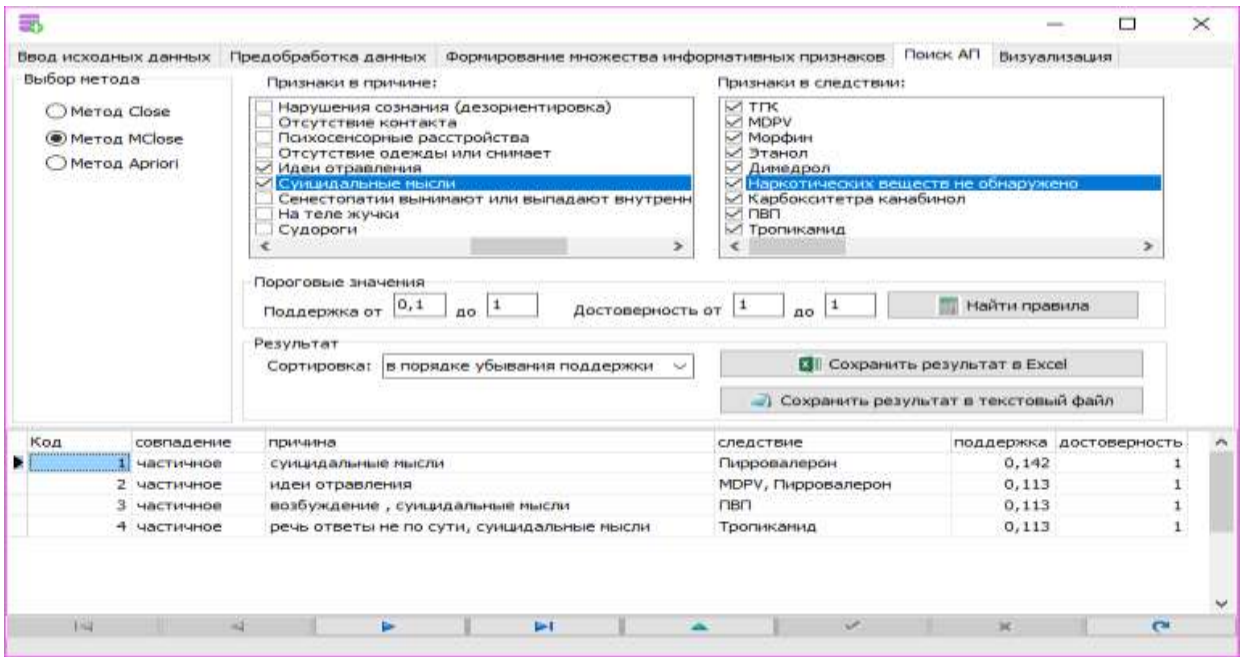


Рисунок 4.7 – Возможные комбинации принятых препаратов при заданной симптоматике

Разделение множества признаков на группы значительно сокращает количество ассоциативных правил и повышает качество интерпретации выявленных закономерностей.

4.3 Оценка результативности средств снижения размерности матрицы «объект–признак»

Вторая серия вычислительных экспериментов выполнялась по определению наиболее информативных признаков с целью определения множественной лекарственной устойчивости (МЛУ) возбудителя туберкулеза. Ис-

пользовалась база данных больных туберкулезом легких, впервые выявленных и проходивших стационарное лечение в Красноярском краевом противотуберкулезном диспансере № 1 за период 2008–2012 годы.

МЛУ – устойчивость микобактерий, вызывающих туберкулез легких, к двум основным противотуберкулезным препаратам. Для выявления у пациента МЛУ в настоящее время стандартными диагностическими средствами затрачивается от 20 до 90 дней с момента выявления заболевания, что ведет к снижению эффективности лечения на начальном этапе. Только после получения результатов исследования наличия МЛУ, лечение корректируется и препараты, к которым микобактерии туберкулеза имеют устойчивость, заменяются препаратами, к которым у возбудителя сохранена чувствительность.

Все множество пациентов T было разделено на две обучающие выборки T_1 и T_2, где T_1 – список пациентов, у которых были выделены микобактерии туберкулеза без МЛУ (всего 334 человека), T_2 – список пациентов, у которых были выделены микобактерии туберкулеза, обладающие МЛУ (всего 445 человек).

В анализе МЛУ традиционно рассматривают 26 признаков: обсеменение, объем поражения, ДТЛ, распад, уплотнение, инфильтрация, рассасывание, место работы, район, форма, населенный пункт, возраст, возрастная группа, первое осложнение, место жительства, должность, сроки ПФЛГО, второе сопутствующее заболевание, статус, первое сопутствующее заболевание, второе осложнение, вторая форма, путь выявления, третья форма, третье осложнение, пол.

Структура файла Tuberkulez.xls, в котором хранятся входные параметры для модуля INPUT, представлена на рисунке 4.8. Первая строка содержит названия признаков, далее содержится информация о 779 пациентах, характеризующихся 26 признаками.

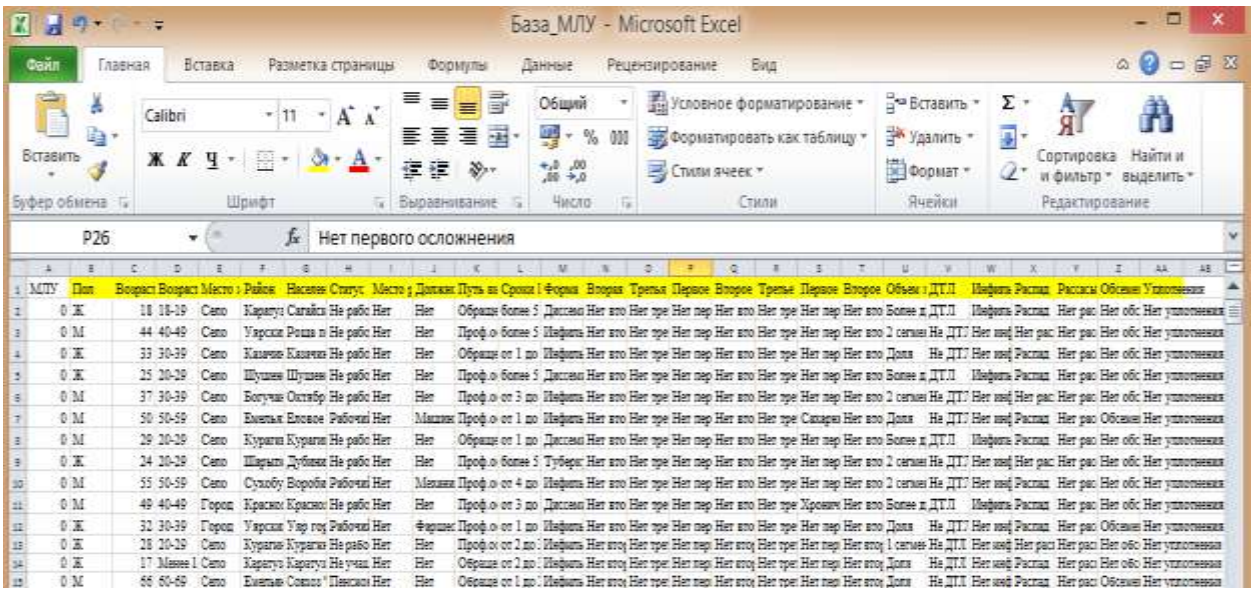


Рисунок 4.8 – Файл с исходными данными

Необходимо преобразовать исходные данные из XLS-формата к форме, подходящей для анализа данных (выполнить дискретизацию и кодирование атрибутов, очистку данных, сократить размерность). Для этого была выполнена процедура автоматического шкалирования с использованием программного модуля EDIT_DATA и вычислена мера информативности по методу Кульбака с использованием программного модуля ELIMINATION. Признаки ранжируются в порядке убывания значения меры информативности (рисунок 4.9). После этого отбирается несколько первых информативных признаков и проверяют качество бинарной классификации по выбранным признаками.

Для проверки качества классификации были выполнены вычисления показателей ROC-анализа: чувствительность, специфичность (рисунок 4.10). В качестве критерия оптимального отсечения выбрана максимальная суммарная чувствительность и специфичность. С учетом выбранного критерия, наиболее информативными (способствующими верной классификации на «нет МЛУ» и «имеет место МЛУ») являются признаки с номерами от 1 до 6 (выделены на диаграмме информативности признаков зеленым цветом). Признаки с номерами 7–26 имеют малую информативность.

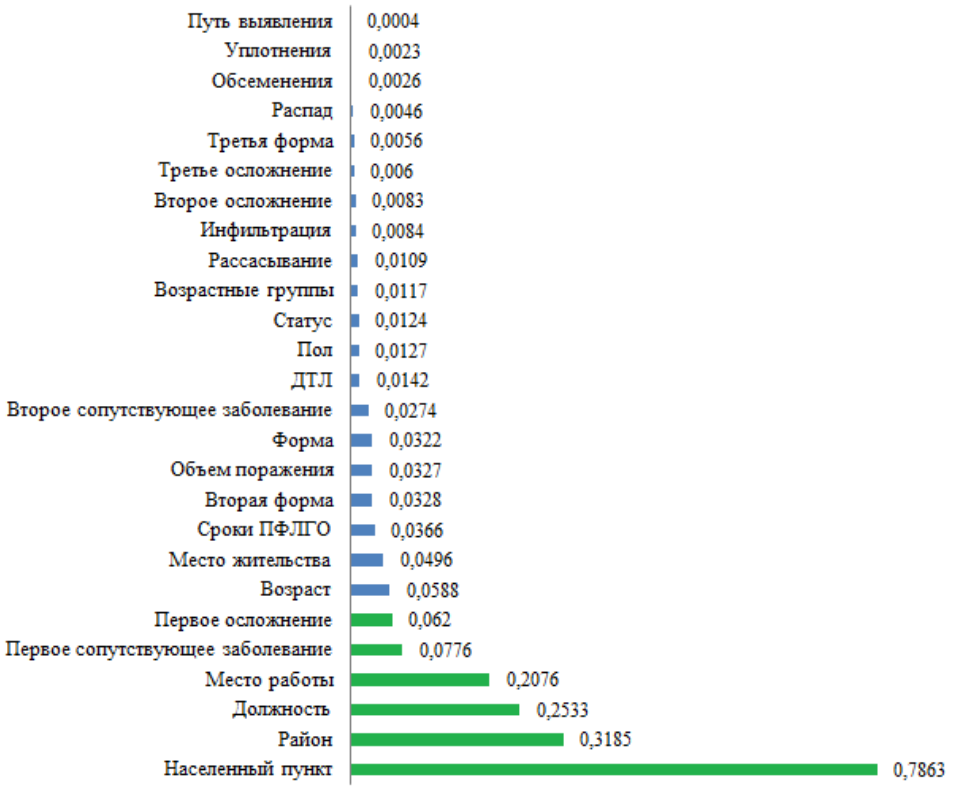
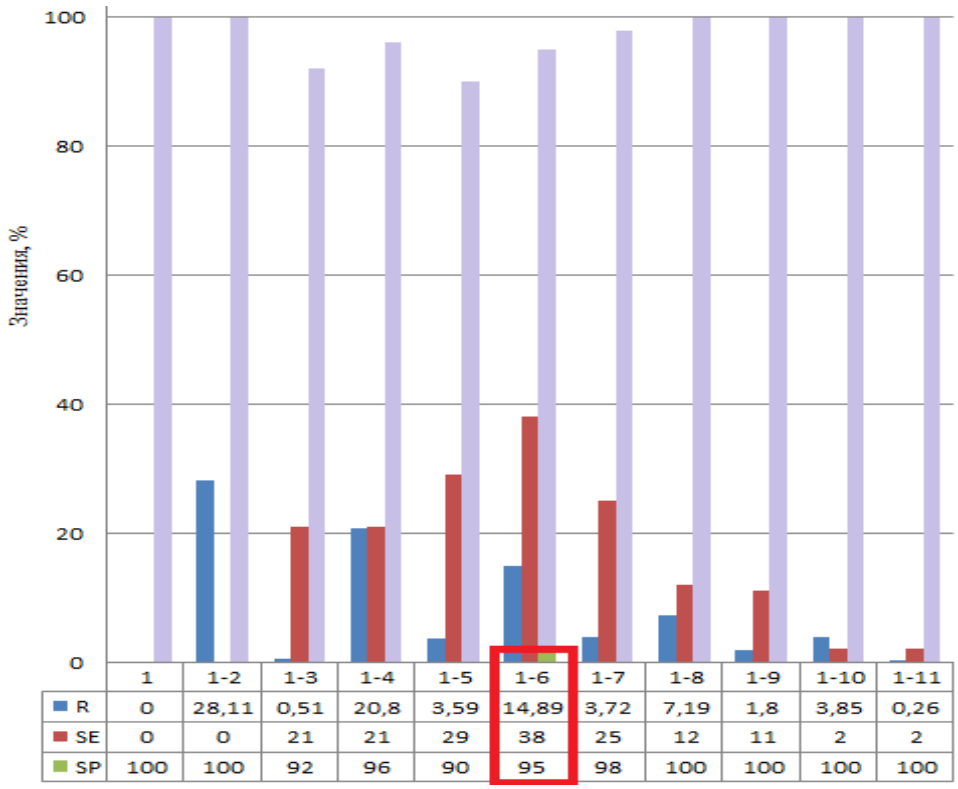


Рисунок 4.9 – Отсортированный список признаков по МЛУ

Максимум значений чувствительности и специфичности достигается при использовании первых шести признаков. При этом чувствительность равна 38 %, а специфичность равна 95 %. Это означает, что при анализе первых шести признаков для 36 % пациентов верно выявляется отсутствие МЛУ, а для 95 % пациентов верно устанавливается наличие МЛУ. Если анализ МЛУ осуществлять с учетом большего числа признаков, то значение чувствительности становится равным нулю при стопроцентной специфичности, т.е. верно классифицируются только отрицательные случаи.

Выяснилось, что наибольшей информативностью обладают признаки, составляющие социальный портрет пациента, а не признаки, описывающие состояние здоровья пациентов. Этот факт подтвержден врачами-экспертами.



	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11
R	0	28,11	0,51	20,8	3,59	14,89	3,72	7,19	1,8	3,85	0,26
SE	0	0	21	21	29	38	25	12	11	2	2
SP	100	100	92	96	90	95	98	100	100	100	100

Рисунок 4.10 – Показатели ROC-анализа по МЛУ

Для выявления типичных представителей множество пациентов делилось на группы по некоторым целевым признакам. Выявленная группа заменялась типичными представителями, тем самым уменьшалось число анализируемых пациентов (таблица 4.7).

Таблица 4.7 – Признаки для генерации ассоциативных правил

Целевой признак, по которому формировалась выборка	Количество пациентов, полученное после замены группы пациентов типичным представителем
Без сокращения	779
Район	61
Возраст	61
Населенный пункт	279
Статус	8
Место работы	63

Вторая серия вычислительных экспериментов выполнялась с целью определения множества наиболее значимых диагностических признаков для выявления наличия сепсиса. Сепсис – это нарушение функций органов, вызванное реакцией организма на инфекцию. Является одной из главных причин летальности у пациентов в критическом состоянии. Поскольку клиническое течение сепсиса может протекать стремительно, от врача требуется очень быстро поставить правильный диагноз.

База данных для решения второй задачи была сформирована из пациентов, проходивших стационарное лечение в Красноярском краевом гнойно-септическом центре в 2017–2018 годы. Исследовались две обучающие выборки T_1 и T_2, где T_1 – список пациентов, у которых не был выявлен сепсис (всего 100 человек), T_2 – список пациентов, у которых был выявлен сепсис (всего 100 человек). Признаковое описание пациентов в анализе сепсиса традиционно рассматривают 16 признаков. Эти признаки принимают числовые значения и не требуют проведения шкалирования. Для всех признаков была вычислена мера информативности по методу Шеннона относительно обучающих выборок T_1 и T_2. Затем признаки были отсортированы в порядке убывания значения меры информативности (рисунок 4.11).

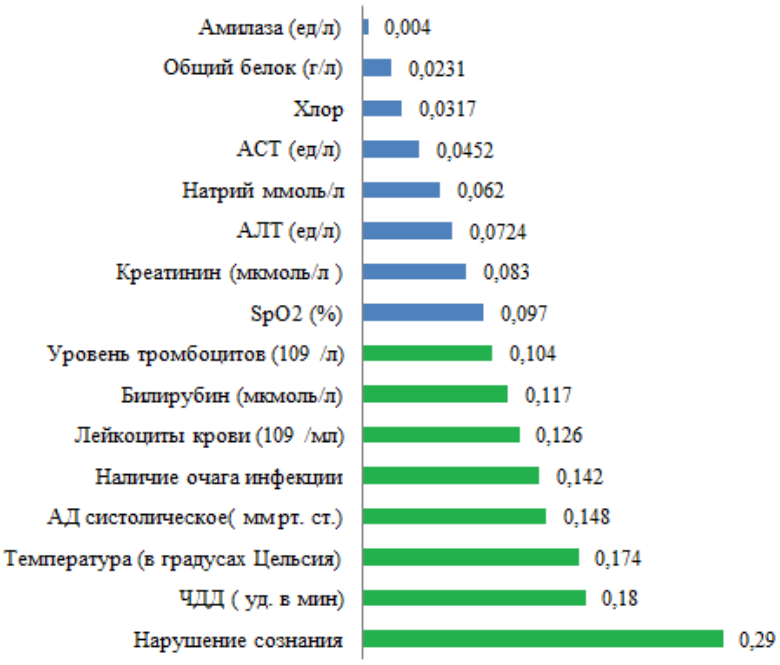
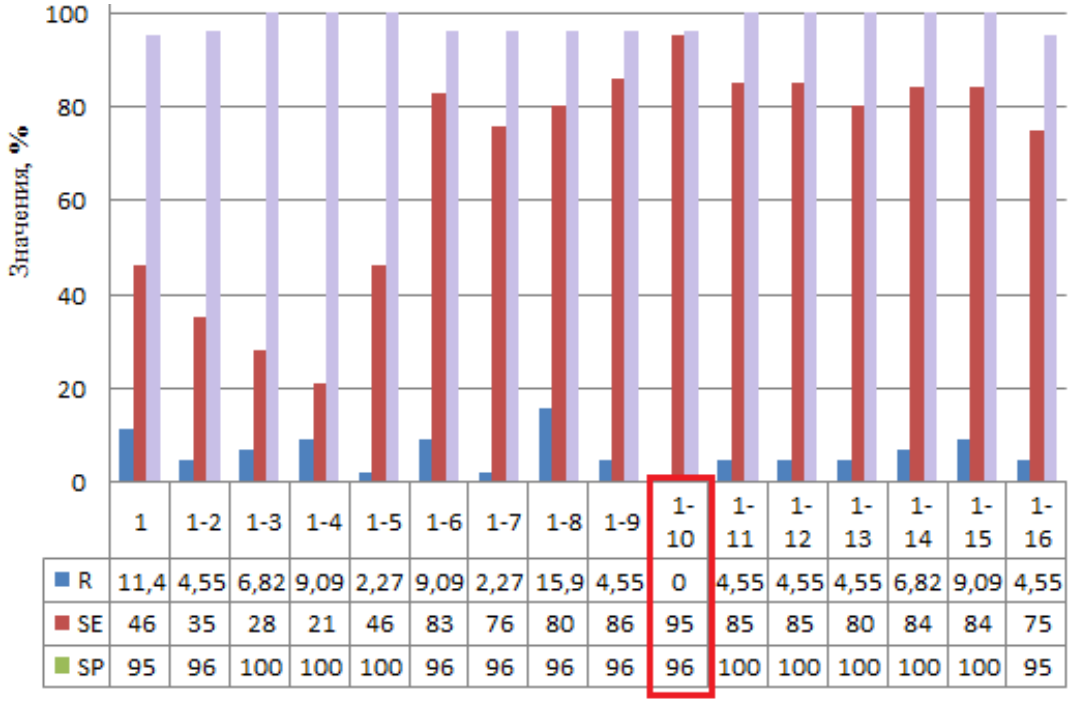


Рисунок 4.11 – Отсортированный список признаков по сепсису

Для проверки качества классификации были выполнены вычисления показателей ROC-анализа. Результаты этих вычислений представлены на рисунке 4.12.



	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	1-13	1-14	1-15	1-16
R	11,4	4,55	6,82	9,09	2,27	9,09	2,27	15,9	4,55	0	4,55	4,55	4,55	6,82	9,09	4,55
SE	46	35	28	21	46	83	76	80	86	95	85	85	80	84	84	75
SP	95	96	100	100	100	96	96	96	96	96	100	100	100	100	100	95

Рисунок 4.12 – Показатели ROC-анализа по сепсису

После вычисления показателей ROC-анализа можно судить о том, что максимальное качество классификации достигается при использовании признаков с номерами 1–10, поскольку при этом достигается максимум значений чувствительности и специфичности. В 95% случаев с использованием этих признаков можно верно классифицировать отсутствие сепсиса и в 96% верно определить наличие сепсиса.

Особого внимания заслуживают объекты, которые не удалось классифицировать или они классифицированы не верно. Значительные отличия свойств этих объектов от остальных объектов одного класса могут объясняться их уникальностью, однако чаще причина отличий состоит во влиянии не учитываемых факторов, например ошибки занесения данных. Такие объекты называют «шумом» и должны удаляться в результате предобработки данных. Не типичные объекты, не являющиеся «шумом» должны сохраняться и анализироваться врачом, либо как некий «особого» случай, либо в диаг-

нозе допущена ошибка. Выявлены и удалены 31 «зашумленный» объект в первой серии экспериментов и 15 – во второй.

Врачи-эксперты, предоставившие клинические данные, подтвердили, что найденный набор средств, реализованный в модуле ELIMINATION, эффективен для клинической диагностики. Применение модуля ELIMINATION позволяет снизить размерность матрицы «объект-признак», а полученные результаты не вызывают затруднений в интерпретации.

4.4 Выводы по главе 4

1. Проведены экспериментальные исследования на клинических данных, предоставленных специалистами медицинских учреждений Красноярского края и сотрудниками кафедры медицинской кибернетики и информатики Красноярского государственного медицинского университета имени профессора В.Ф. Войно-Ясенецкого, для оценки результативности разработанных метода, алгоритмов и программ.

2. Эксперименты показали, что минимальное количество строгих ассоциативных правил было получено с помощью алгоритма MClose и среди них нет избыточных правил. Любое изменение состава выборки существенно влияет на состав сгенерированных ассоциативных правил. Разделение множества признаков на группы значительно сокращает количество ассоциативных правил и повышает качество интерпретации выявленных закономерностей.

3. Использование на практике неизбыточного базиса строгих ассоциативных правил значительно упрощает процесс интерпретации причинно-следственных связей в задачах медицинской диагностики

ЗАКЛЮЧЕНИЕ

1. Установлены свойства строгих ассоциативных правил и получен набор выводимостей D_1, D_3, D_4, D_5, гарантирующих сохранение поддержки (леммы 2.1–2.6, теорема 2.1). 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 литературы

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. Справочное издание – М.: Финансы и статистика, 1989. – 607 с.

2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. Справочное издание – М.: Финансы и статистика, 1985. – 487 с.

3. Аксёнов С.В., Костин К.А., Иванова А.В., Liang J., Замятин А.В. Диагностика патологий по данным видеоэндоскопии с использованием ансамбля сверточных нейронных сетей // Современные технологии в медицине – 2018. – Т. 10. – № 2. – С. 7-19.

4. Андрющенко В.С., Углов А.С., Замятин А.В. Статистическая классификация иммуносигнатур для задач ранней диагностики заболеваний при значительном сокращении размерности признакового пространства // Современные технологии в медицине. – 2018. – Т. 10. – № 3. – С. 14-20.

5. 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.

6. Бабин М.А. О приближенном базисе импликаций // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2012. – № 8. – С. 20-23.

7. Баранов А. А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнева Е.А., Антонова Е.В., Смирнов В.И., Латышев А.В. Методы и средства комплексного интеллектуального анализа ме-

дицинских данных // Труды института системного анализа российской академии наук. – 2015. – Т. 65. – № 2. – С. 81-93.

8. Белышев Д.В., Кочуров Е.В. Анализ методов хранения данных в современных медицинских информационных системах // Программные системы: теория и приложения. – 2016. – Т. 7. – № 2. – С. 85-103.

9. Биркгоф Г. Теория решеток. – М.: Наука, 1984. – 568 с.

10. Биркгоф Г., Барти Т. Современная прикладная алгебра. – М.: Мир, 1976.– 400 с.

11. Быкова В.В., Катаева А.В. О неизбыточном представлении минимаксного базиса строгих ассоциативных правил // Прикладная дискретная математика. – 2017. – № 36. – С. 113-126.

12. Быкова В.В., Катаева А.В. Алгоритм построения неизбыточного минимаксного базиса строгих ассоциативных правил // Прикладная дискретная математика. Приложение. – 2017. – № 10. – С. 154-157.

13. 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.

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

15. Бююль А., Цефель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. – СПб.: ООО ДиаСофтЮП, 2005. – 608 с.

16. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов: статистические проблемы обучения. – М.: Наука, 1974. – 416 с.

17. Виноградов А.Н., Гулиев Я.И., Куршев Е.П., Малых В.Л. Перспективные направления исследований в области клинического моделирования,

управления и принятия решений // Врач и информационные технологии. – 2014. – № 5. – С. 48-59.

18. Виноградов К.Н., Быкова В.В., Наркевич А.Н., 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.

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

20. Витяев Е. Е., Демин А. В., Пономарев Д. К. Вероятностное обобщение формальных понятий // Программирование. – 2012. – Т. 38. – № 5. – С. 18-34.

21. Воронов К.В. Машинное обучение: курс лекций. – 2010. [Электронный ресурс] URL: http://www.machinelearning.ru (дата обращения 16.02. 2019).

22. Вьюгин В.В. Математические основы машинного обучения и прогнозирования – М.: МЦНМО, 2013. – 387 с.

23. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001. – 752 с.

24. Городецкий В.И., Тушканова О.Н. Ассоциативная классификация: аналитический обзор. Ч. 1 // Труды СПИИРАН. – 2015. – № 38. – С. 183-203.

25. Городецкий В.И., Тушканова О.Н. Ассоциативная классификация: аналитический обзор. Ч. 2 // Труды СПИИРАН. – 2015. – № 39. – С. 212-240.

26. Грибова В.В., Петряева М.В., Окунь Д.Б., Шалфеева Е.А. Онтология медицинской диагностики для интеллектуальных систем поддержки принятия решений // Онтология проектирования. – 2018. – Т. 8. – № 1 (27). – С. 58-73.

27. Гулиев Я.И. Основные аспекты разработки медицинских информационных систем // Врач и информационные технологии. – 2014. – № 5. – С. 10-19.

28. Гуров С. И. Булевы алгебры, упорядоченные множества, решетки: определения, свойства, примеры. – М.: Либроком, 2013. – 221 с.

29. Гусев А.В., Зарубина Т.В. Поддержка принятия врачебных решений в медицинских информационных системах медицинской организации // Врач и информационные технологии. – 2017. – № 2. – С. 60-62.

30. Долгов А.И. Алгоритмизация прикладных задач. – М.: Флинта, 2013. – 136 с.

31. Демин А.В., Витяев Е.Е. Разработка универсальной системы извлечения знаний "Discovery" и ее применение // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2009. – Т. 7. – № 1. – С. 73-83.

32. Дюбанов В.В., Руднев А.С., Павловский Е.Н., Зозуля Ю.В., Самочернова А.С., Сандер Д.С. Методы исследования операций и когнитивного анализа данных в решении задач лечебно-профилактических учреждений // Патология кровообращения и кардиохирургия. – 2011. – № 4. – С.77-82.

33. Дюк В., Эмануэль В. Информационные технологии в медико-биологических исследованиях. – СПб.: Питер, 2003. – 528 с.

34. Дюкова Е.В., Песков Н.В. Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Журнал вычислительной математики и математической физики. – 2002. – Т. 42. – № 5. – С. 741-753.

35. Ефименко И.В., Хорошевский В.Ф. Интеллектуальные системы принятия поддержки принятия решений в медицине: ретроспективный обзор состояния исследований и разработок и перспективы // Открытые семантические технологии проектирования интеллектуальных систем. – 2017. – № 7. – С. 251-260.

36. Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. – 1978. – Т. 33. – С. 5-68.

37. Журавлёв Ю. И., Рязанцев В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. – М.: Фазис, 2005. – 159 с.

38. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

39. Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А. Функции конкурентного сходства в алгоритмах распознавания комбинированного типа // Вестник Сибирского государственного аэрокосмического университета им. М.Ф. Решетнева. – 2010. – № 5 (31). – С. 19-21.

40. Загоруйко Н.Г. Когнитивный анализ данных. – Новосибирск: ИМ СО РАН, 2013. – 186 с.

41. Зайко Т.А., Олейник А.А., Субботин С.А. Ассоциативные правила в интеллектуальном анализе данных // Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование. – 2013. – № 39 (1012). – С. 82-96.

42. Игнатов Д.И., Кузнецов С.О. Бикластеризация объектно-признаковых данных на основе решеток замкнутых множеств // Труды 12-ой национальной конференции по искусственному интеллекту с международным участием. – 2010. – С. 175-182.

43. Каменский В. С. Методы и модели неметрического многомерного шкалирования (обзор) // Автоматика и телемеханика. – 1977. – № 8. – С. 118-156.

44. Кашницкий Ю.С. Методы замкнутых описаний в задаче классификации данных со сложной структурой: автореф. дис. … канд. техн. наук: 15.13.17. – М., 2018. – 25 с.

45. Карпов Л.Е., Юдин В.Н. Методы добычи данных при построении локальной метрики в системах вывода по прецедентам // Препринт Института системного программирования РАН. – 2006. – № 18. – С. 1-42.

46. 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 1.02.2018.

47. 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 8.02.2018.

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

49. 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.

50. 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.

51. Кедров С.А., Кузнецов С.О. Исследование групп пользователей Интернет-ресурсами методами анализа формальных понятий и разработки данных (DataMining) // Бизнес-информатика. – 2007. – № 1. – С. 45-51.

52. Ким Дж. О., Мюллер Ч.У., Клекка У.Р. Факторный, дискриминантный и кластерный анализ.– М.: Финансы и статистика, 1989. – 215 с.

53. Кобринский Б.А. Автоматизированные диагностические и информационно-аналитические системы в педиатрии // Русский медицинский журнал. – 1999. – Т. 7. – № 4. – С. 35-42.

54. Колесникова С. И. Методы анализа информативности разнотипных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2009. – №1(6). – С. 69-80.

55. Колесникова С.И., Янковская А.Е. Статистический подход к оцениванию зависимых признаков в интеллектуальных системах // Математические методы распознавания образов. – 2007. – Т. 13. – № 1. – С. 143-146.

56. Колесникова С.И., Янковская А.Е. Оценка значимости признаков для тестов в интеллектуальных системах // Известия РАН. Теория и системы управления.– 2008. – № 6 – С. 99-112.

57. Копаница Г., Цветкова Ж. Европейский опыт и пути развития информатизации системы здравоохранения // Врач и информационные технологии. – 2013. – № 1. – С. 49-53.

58. Космачёва И.М. Технологическая схема контроля обработки медицинских данных // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2012. – № 1. – С. 124-130.

59. Кривенко М.П. Критерии значимости отбора признаков классификации // Информатика и её применения. – 2016. – Т. 10. – № 3. – С. 32-40.

60. Кузнецов С.О. Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. – 2001. – № 10. – С. 3-27.

61. Кузнецов С.О. Методы теории решеток и анализа формальных понятий в машинном обучении // Новости искусственного интеллекта. – 2004. – № 3. – С. 19-31.

62. Кузнецова А.В., Сенько О.В. Возможности использования методов DataMining при медико-лабораторных исследованиях для выявления закономерностей в массивах данных // Врач и информационные технологии. – 2005. – № 2. – С. 49-56.

63. Лопач С.Н., Чубенко А.В., Бабич П.Н. Статические методы в медико-биологических исследованиях с использованием Excel. – Киев: МОРИОН, 2001. – 408 с.

64. Лафоре Р. Объектно-ориентированное программирование в С++. – СПб.: Питер, 2018. – 928 с.

65. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. – Новосибирск: Наука,1981. – 160 с.

66. МайерД. Теория реляционных баз данных. – М.: Мир, 1987. – 608 с.

67. Мещеряков Р.В., Балацкая Л.Н., Чойнзонов Е.Л. Специализированная информационная система поддержки деятельности медицинского учреждения // Информационно-управляющие системы. – 2012. – № 5. – С. 51-56.

68. Миркин Б. Г. Анализ качественных признаков и структур– М.: Статистика, 1980. – 317 с.

69. Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч. 2. Исследование медицинских технологических процессов на основе интеллектуального анализа данных. – М.: ФИЗМАТЛИТ, 2006. – 144 с.

70. Назаренко Г.И., Осипов, Г.С., Назаренко, А.Г., Молодченков А.И. Интеллектуальные системы в клинической медицине. Синтез плана лечения на основе прецедентов // Информационные технологии и вычислительные системы. – 2010. – № 1. – С. 24-35.

71. Наркевич А.Н., Виноградов К.А., Корецкая Н.М., Катаева А.В., Журбенко Е.О. Оценка информативности и отбор признаков при идентификации объектов на цифровых изображениях микроскопических препаратов,

окрашенных по методу Циля-Нильсена // В мире научных открытий. – 2017. – Т. 9. – № 4. – С. 106-121.

72. Наркевич А.Н., Плотников Д.В., Виноградов К.А., Катаева А.В. Сравнение методов отбора признаков для идентификации объектов на цифровых изображениях микроскопических препаратов // Инженерный вестник Дона. – 2018. – № 2 (49). – С. 23-33.

73. Паспорт приоритетного проекта «Совершенствование процессов организации медицинской помощи на основе внедрения информационных технологий». Приложение к протоколу президиума Совета при Президенте Российской Федерации по стратегическому развитию и приоритетным проектам от 25 октября 2016 г. № 9. [Электронный ресурс] URL: http://static.government.ru/media/files/9ES7jBWMiMRqONdJYVLPTyoVKYwgr4Fk.pdf. (дата обращения: 22.02.2019).

74. Платонов В.В., Семёнов П.О. Методы сокращения размерности в системах обнаружения сетевых атак // Проблемы информационной безопасности. Компьютерные системы. – 2012. – № 3. – С. 40-45.

75. Рао С.Р. Линейные статистические методы и их применения. – М.: ФИЗМАТЛИТ, 1968. – 548 с.

76. Самойленко Н.Э., Кувина В.Н., Кувин С.С. Комплексный анализ медицинских данных // Вестник Воронежского государственного технического университета. – 2009. – Т. 5. – № 9. – С. 114-118.

77. Стамат Д., Альгамди В., Шталь Д., Замятин А., Мюррей Р., Ди Форти М. Могут ли искусственные нейронные сети предсказать психиатрические состояния, связанные с употреблением каннабиса? // Достижения IFIP в области информационных и коммуникационных технологий. – 2018. – Т. 519. – С. 311-322.

78. Сигал И.Х., Иванова А.П. Введение в прикладное дискретное программирование: модели и вычислительные алгоритмы. – М.: ФИЗМАТЛИТ, 2016. – 304 с.

79. Терехина А.Ю. Методы многомерного шкалирования и визуализации данных (обзор) // Автоматика и телемеханика. – 1973. – № 7. – С. 80-94.

80. Трояновский В.М. Информационно-управляющие системы и прикладная теория случайных процессов. – М.: Гелиос АРВ, 2014. – 304 с.

81. Щекина Е.Н. Использование системного подхода для создания систем поддержки принятия решений в медицине // Вестник новых медицинских технологий. Электронное издание.– 2017. – № 1. – С. 356-364.

82. Янковская А.Е., Горбунов И.В., Черногорюк Г.Э. Влияние способа вычисления весовых коэффициентов признаков и построения безизбыточных безусловных диагностических тестов для гибридной интеллектуальной системы дифференциальной диагностики диссеминированных заболеваний легких // VII Всероссийская научно-практическая конференция «Нечеткие системы, мягкие вычисления и интеллектуальные технологии». – СПб., 2017. – Т. 2. – С. 191-200.

83. Armstrong W.W. Dependency structure of data bases relationships // Proceedings IFIP Congress. Geneva. – 1974. – P. 580-583.

84. Agrawal R., Imielinski T., Swami A. Database mining: A performance perspective // Special Issue on Learning and Discovery in Knowledge-Based Databases / Ed. by N. Cercone, M. Tsuchiya. – Washington, U.S.A.: Institute of Electrical and Electronics Engineers, 1993. – Vol. 6. – No. 5. – P. 914-925.

85. Agrawal R., Srikant R. Fast algorithms for mining association rules // Proceedings 20 th Int. Conf. Very Large Data Bases, VLDB. – Morgan Kaufmann, 1994. – P. 487–499.

86. Aikins J.S., Kunz J.C., Shortliffe E.H., Fallat R.J. PUFF: an expert system for interpretation of pulmonary function data // Computers and biomedical research. – 1983. – Vol. 16. – No. 3. – P. 199-208.

87. Balcazar J.L. Redundancy, deduction schemes, and minimum-size bases for association rules // Logical Methods in Computer Science. – 2010. – Vol. 6. – No. 2-3. – P. 1-33.

88. Barnett G.O., Cimino J.J., Hupp, J.A., Hoffer E.P. DXplain: an evolving diagnostic decision-support system // Jama, 1987. – Vol. 258. – No. 1. – P. 67-74.

89. Berner E.S. Clinical decision support systems: State of the art // Communications and Network. – 2009. – Vol. 9. – No. 4. – P. 4-26.

90. Borisova I.A., Zagoruiko N.G. Feature selection by using the FRiS function in the task of generalized classification // Pattern recognition and image analysis. – 2011. – Vol. 21. – No. 2. – P. 117-120.

91. Brossette S.E., Sprague A.P., Jones W.T., Moser S.A. A data mining system for infection control surveillance // Methods of information in medicine. – 2000. – Vol. 39 – No. 4 – P. 303-310.

92. Chen T. J., Chou L. F., Hwang S. J. Application of a data-mining technique to analyze co prescription patterns for antacids in Taiwan // Clinical Therapeutics. – 2003. – Vol. 25 – No. 9. – P. 2453-2463.

93. Ceglar A., Roddick J. Association Mining // ACM Computing Surveys. – 2006 – Vol. 38. – No. 2. – P. 1-42

94. Fawcett T. ROC graphs: Notes and practical considerations for researchers // Machine learning. – 2004. – Vol. 31 – No. 1. – P. 1-38.

95. Ganter B., Wille R. Formal Concept Analyses: mathematical foundations Springer Science and Business Media, 2012. – 314 p.

96. Geng L., Hamilton H. Interestingness measures for data mining: A survey // ACM Computing Surveys (CSUR). – 2006. – Vol. 38. – No. 3. – P. 9-41.

97. Gironi M., Saresella M., Rovaris M., Vaghi M., Nemni R., Clerici M., Grossi E. A novel data mining system points out hidden relationships between immunological markers in multiple sclerosis // Immunity Ageing. – 2013. – Vol. 10 – No. 1. [Электронный ресурс] URL: http://www.biomedcentral.com /content/pdf/1742–4933–10–1.pdf. (дата обращения: 28.09.2018).

98. Glinsky G.V. Gene expression profiling predicts clinical outcome of prostate cancer // Journal of clinical investigation. – 2004. – Vol. 113. – No. 6. – P. 913-923.

99. Harper P. R. A review and comparison of classification algorithms for medical decision making // Health Policy. – 2005. – Vol. 71. – No. 3. – P. 315-331.

100. Hidber C. Online association rule mining // SIGMOD Conf. – 1999. – P. 145-156.

101. Hipp J., Guntzer U., Nakhaeizadeh G. Algorithms for association rule mining – a general survey and comparison // SIGKDD Explorations. – 2000.– Vol. 2. – No. 1. – P. 58-64.

102. Hu H., Li J., Plank A., Wang H., Daggard G. A comparative study of classification methods for microarray data analysis // Proceedings of the fifth Australasian conference on Data mining and analytics, Australian Computer Society, Inc., 2006. –Vol. 61. – P. 33-37.

103. Ilayaraja M., Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm // Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on. – IEEE, 2013. – P. 194-199.

104. Kaytoue M., Kuznetsov S.O., Napoli A., Duplessis S. Mining gene expression data with pattern structures in formal concept analysis // Information Sciences. – 2011. – Vol. 181. – No. 10. – P. 1989-2001.

105. Kelly III J. E., Hamm S. Smart machines: IBM's Watson and the era of cognitive computing. – Columbia University Press, 2013. – 160 p.

106. Kira K., Rendell L.A. The feature selection problem: Traditional methods and a new algorithm // Proceedings of AAAI-92, (San Jose, CA). – 1992. – Vol. 2. – P. 122-126.

107. Kuznetsov S.O., Obiedkov S. A. Comparing performance of algorithms for generating concept lattices // Journal of Experimental Theoretical Artificial Intelligence. – 2002. – Vol. 14. – No. 2-3. – P. 189-216.

108. Kuznetsov S.O. On stability of a formal concept // Annals of Mathematics and Artificial Intelligence. – 2007. – Vol. 49. – No.1-4. – P. 101-115.

109. Kuznetsov S.O. Scalable knowledge discovery in complex data with pattern structures // International Conference on Pattern Recognition and Machine Intelligence. – Springer, Berlin, Heidelberg, 2013. – P. 30-39.

110. Mannila H., Toivonen H., Verkamo A. I. Efficient algorithms for discovering association rules // AAAI Workshop on Knowledge Discovery in Databases (KDD-94). – Seattle, Washington: AAAI Press, 1994. – P. 181-192.

111. Molina L.C., Belanche L., Nebot A. Feature Selection Algorithms: A Survey and Experimental Evaluation // Proceedings of the 2002 IEEE International Conference on Data Mining. – 2002. – P. 306-313.

112. Obiedkov S., Duquenne V. Attribute-incremental construction of the canonical implication basis // Annals of Mathematics and Artificial Intelligence. – 2007. – Vol. 49. – No. 1-4. – P. 77-99.

113. Park J. S., Chen M. S., Yu P.S. An effective hash-based algorithm for mining association rules // ACM. – 1995. – Vol. 24. – No. 2. – P. 175-186.

114. Pasquier N., Bastide Y., Taouil R., Lakhal L. Efficient mining of association rules using closed itemset lattices // Information systems. – 1999. – Vol. 24. – No. 1. – P. 25-46.

115. Pasquier N., Taouil R., Bastide Y., Stumme G., Lakhal L. Generating a condensed representation for association rules // Journal of intelligent information systems. – 2005. – Vol. 24. – No. 1. – P. 29-60.

116. Poelmans J., Ignatov D.I., Kuznetsov S.O., Dedene G. Formal concept analysis in knowledge processing: A survey on models and techniques // Expert systems with applications. – 2013. – Vol. 40. – No. 16. – P. 6601-6623.

117. Poelmans J., Ignatov D.I., Viaene S., Dedene G., Kuznetsov S.O. Text mining scientific papers: a survey on FCA-based information retrieval research // Industrial Conference on Data Mining. – Springer, Berlin, Heidelberg, 2012. – P. 273-287.

118. Prokasheva O., Onishchenko A., Gurov S. Classification methods based on formal concept analysis // FCAIR 2012-Formal Concept Analysis Meets Information Retrieval. – 2013. – P. 95-104.

119. Ryssel U., Distel F., Borchmann D. Fast algorithms for implication bases and attribute exploration using proper premises // Annals of Mathematics and Artificial Intelligence. – 2013. –Vol. 65. – P. 1-29.

120. Rudolph S. Some notes on pseudo-closed sets // International Conference on Formal Concept Analysis. – Springer, Berlin, Heidelberg, 2007. – P. 151-165.

121. Santos R.S., Malheiros S. M., Cavalheiro S., De Oliveira, J.P. A data mining system for providing analytical information on brain tumors to public health decision makers // Computer methods and programs in biomedicine. – 2013. – Vol. 109. – No. 3. – P. 269-282.

122. Shortliffe E. Computer-based medical consultations: MYCIN, Elsevier, 2012. – Vol. 2. – 236 p.

123. Srikant R., Agrawal R. Mining generalized association rules // Future generation computer systems. – 1997. – Vol. 13. – No. 2-3. – P. 161-180.

124. Stumme G., Taouil R., Bastide Y., Pasquier N., Lakhal L. Computing iceberg concept lattices with TITANIC // Data knowledge engineering. – 2002. – Vol. 42. – No. 2. – P. 189-222.

125. Sucahyo Y. G., Gopalan R. P. CT-ITL: Efficient frequent item set mining using a compressed prefix tree with pattern growth // Proceedings of the 14th Australasian database conference. – Australian Computer Society, Inc., 2003. – Vol. 17. – P. 95-104.

126. Toivonen H. Sampling large databases for association rules // Proceedings Int. Conf. Very Large Data Bases. – Morgan Kaufman, 1996. – P. 134-145.

127. Uno T., Asai T., Uchida Y., Arimura H. An efficient algorithm for enumerating closed patterns in transaction databases // LNCS, 2004. – Vol. 3245. – P. 16-31.

128. Van Der Merwe D., Obiedkov S., Kourie D. Addintent: A new incremental algorithm for constructing concept lattices // International Conference on Formal Concept Analysis. – Springer, Berlin, Heidelberg, 2004. – P. 372-385.

129. Yevtushenko S.A. System of data analysis "Concept Explorer" // Proceedings 7th National Conference on Artificial Intelligence (KII'00). – 2000. – P. 127-134.

130. Yoo I., Alafaireet P., Marinov M., Pena-Hernandez K., Gopidi R., Chang J. F., Hua L. Data mining in healthcare and biomedicine: a survey of the literature // Journal of medical systems. – 2012. – Vol. 36. – No. 4. – P. 2431-2448.

131. Zagoruiko N.G., Borisova I.A., Dyubanov V.V., Kutnenko O.A. Methods of recognition based on the function of rival similarity // Pattern recognition and image analysis. – 2008. – Vol. 18. – No. 1. – P. 1-6.

132. Zaki M. J., Hsiao C. J. Efficient algorithms for mining closed itemsets and their lattice structure // IEEE Transactions on Knowledge Data Engineering. – 2005. – No. 4. – P. 462-478.

133. Zauderer M.G., Gucalp A., Epstein A.S., Seidman A.D., Caroline A., Granovsky S., Petri J. Piloting IBM Watson Oncology within Memorial Sloan Kettering’s regional network // Journal of Clinical Oncology. – 2014.– Vol. 32. – P. 153-176.

134. Zhang C., Zhang S. Association rules mining: models and algorithms. – Springer-Verlag, 2002. – 240 p.

УТВЕРЖДАЮ
Главный врач КГБУЗ ККБ
Корчагин Егор Евгеньевич
_____
«20» 03 2019 г.

АКТ

использования научных результатов диссертационной работы Катаевой А.В. «Извлечение и неизбыточное представление закономерностей в многомерных данных»

Комиссия Краевого государственного бюджетного учреждения здравоохранения «Краевая клиническая больница» в составе Корчагина Егора Евгеньевича (главного врача), Черкашина Олега Андреевича (начальника отдела автоматизированных систем управления) и Масленникова Алексея Викторовича (ведущего специалиста отдела автоматизированных систем управления) рассмотрела вопрос об использовании результатов диссертационной работы Катаевой Алины Владимировны «Извлечение и неизбыточное представление закономерностей в многомерных данных» на соискание ученой степени канд. физ.-мат. наук по специальности 05.13.17 – Теоретические основы информатики, выполненной в Сибирском федеральном университете.

Результаты диссертационной работы Катаевой А.В. обладают актуальностью и представляют интерес для практического здравоохранения. Программы для ЭВМ «Выявление ассоциативных правил и построение неизбыточного минимаксного базиса», «Программа сокращения признакового пространства на основе алгоритмов классификации и ROC-анализа», разработанные Катаевой А.В. в рамках диссертации, переданы в отдел автоматизированных систем управления Краевой клинической больницы для встраивания в существующую медицинскую информационную систему.

Переданные программы рекомендованы к использованию в Краевом государственном бюджетном учреждении здравоохранения «Краевая клиническая больница» для проведения научных исследований и клинической диагностики в рамках мероприятий по выполнению национального проекта «Электронное здравоохранение», утвержденного Президиумом Совета при Президенте Российской Федерации (протокол № 9 от 25.10.2016 г.).

Начальник отдела АСУ КГБУЗ ККБ _____ О.А. Черкашин

Ведущий специалист отдела АСУ
КГБУЗ ККБ _____ А.В. Масленников

УТВЕРЖДАЮ

И.о. ректора ФГБОУ ВО КрасГМУ
им. проф. В.Ф. Войно-Ясенецкого
Минздрава России
Никулина Светлана Юрьевна

_____

«20» марта 2019 г.

АКТ

о внедрении в учебный процесс Федерального государственного бюджетного образовательного учреждения высшего образования «Красноярский государственный медицинский университет им. проф. В.Ф. Войно-Ясенецкого» Министерства здравоохранения Российской Федерации научных результатов диссертационной работы Катаевой А.В. «Извлечение и неизбыточное представление закономерностей в многомерных данных»

Программы для ЭВМ «Выявление ассоциативных правил и построение неизбыточного минимаксного базиса», «Программа сокращение признакового пространства на основе алгоритмов классификации и ROC-анализа», разработанные Катаевой Алиной Владимировной, и теоретические результаты кандидатской диссертации «Извлечение и неизбыточное представление закономерностей в многомерных данных», выполненной в Сибирском федеральном университете, внедрены в учебный процесс на кафедре медицинской кибернетики и информатики. Эти материалы используются при подготовке врачей-кибернетиков по специальности 30.05.03 – «Медицинская кибернетика», при изучении дисциплин «Информатика, медицинская информатика», «Клиническая кибернетика», «Методы интеллектуального анализа данных в медицине» и выполнении выпускных квалификационных работ.

Заведующий кафедрой медицинской
кибернетики и информатики,
д-р мед. наук, профессор _____ К.А. Виноградов