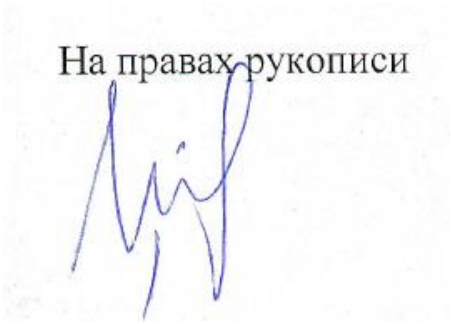
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»



Згуральская Екатерина Николаевна

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ПОИСКА СКРЫТЫХ ЗАКОНОМЕРНОСТЕЙ В БАЗАХ ДАННЫХ ПРИМЕНЕНИЕМ ИНТЕРВАЛЬНЫХ МЕТОДОВ НА ПРИМЕРАХ В ПРОМЫШЛЕННОСТИ И ДРУГИХ ОБЛАСТЯХ

Специальность 05.13.01 –Системный анализ, управление и обработка информации (информационные технологии и промышленность)

ДИССЕРТАЦИЯ
на соискание ученой степени кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Крашенинников В. Р.

Ульяновск – 2021

Содержание

ВВЕДЕНИЕ

В диссертации исследуется использование интервальных методов для поиска скрытых закономерностей в базах данных в предметных областях.

Актуальность избранной темы. Использование цифровых технологий для управления в научной, производственной и социальной сферах является одним из главных факторов инновационного развития современного общества. Важную роль для совершенствования цифровых технологий играют информационные модели, основанные на знаниях. Как правило, неявные знания содержатся в базах и хранилищах данных в форме скрытых закономерностей. Отсюда возникает задача выявления этих закономерностей. Идёт поиск путей повышения эффективности управления с учётом системных связей и новых знаний о функционировании объектов в предметных областях. Большой вклад в развитие системного анализа, управления и обработки информации внесли российские и зарубежные ученые: Вапник В.Н., Воронцов К. В., Граничин О.Н., Гудфеллоу Я., Дюк В.А., Журавлёв Ю.И., Загоруйко Н.Г., Пятецкий-Шапиро Г. и др.

К числу основных проблем построения информационных моделей в слабо структурированных предметных областях является высокая комбинаторная сложность алгоритмов для поиска логических закономерностей.

Важное значение имеет выбор способов предобработки данных для уменьшения комбинаторной сложности алгоритмов интеллектуального анализа данных, разработка новых методов оценки обобщающей способности алгоритмов распознавания и отбор информативных наборов признаков в описании допустимых объектов.

В рамках метода логической геометрии, разработанного Дюком В.А., предложено поиск логических закономерностей производить в окрестности указанного объекта через снижение размерности исходного пространства с использованием линейного отображение его на плоскость. Такое отображение существенно искажает структуру отношений объектов в исходном пространстве, что сильно ограничивает возможности для интерпретации результатов анализа на плоскости.

Метод опорных векторов SVM, предложенный Вапником В.Н., очень чувствителен к наличию шумовых объектов в обучающей выборке. Жадная стратегия обучения алгоритма метода рассматривает шумовые объекты как граничные, что оказывает существенное влияние на снижение обобщающей способности метода.

Селекция обучающих выборок в работах Загоруйко Н.Г. через поиск минимального покрытия эталонами реализуется путём ручной настройки параметров алгоритма FRIS STOLP. Кроме того, не решена проблема численного решения обнаружения начала переобучения.

Основным инструментарием в диссертации для поиска закономерностей в данных являются интервальные методы. С помощью этих методов упорядоченные значения признака (исходного или латентного) требуется разбить на интервалы так, чтобы каждый интервал содержал как можно больше значений признака объектов одного класса и как можно меньше значений признака объектов других классов. С этим требованием согласуются два используемых в диссертации критерия качества, в которых оценивается степень однородности, устойчивости значений признака по интервалам. Оптимальность разбиения понимается в смысле экстремума значения критерия, выбор которого определяется спецификой задачи.

Следует отметить, что иногда данные могут иметь пропуски, то есть у некоторых объектов имеются значения не всех признаков, что усложняет задачу поиска закономерностей вообще, и построения границ интервалов в частности. Кроме того, имеются трудности, связанные с выбором методов преобразования шкал измерений признаков с минимальной потерей информации и селекции обучающих выборок на данных с большой размерностью. Эти вопросы рассматриваются в диссертации.

Поиск оптимального разбиения (то есть границ интервалов) представляет сложную задачу, решить которую простым перебором практически невозможно. Поэтому тема данной диссертации является актуальной, так как ее основной задачей является нахождение способов снижения вычислительной сложности

поиска оптимального разбиения. Для этого разработан рекурсивный алгоритм, который позволяет вычислять границы интервалов при отсутствии измеренных значений некоторых признаков в описании части объектов, а также способы предобработки данных (численный алгоритм вычисления экстремума критерия разбиения значений признака на непересекающиеся интервалы) и методы отбора информативных наборов признаков по выборке данных в целом и для формирования собственного пространства объекта.

Показано, что применение полученных в диссертации результатов позволяет находить закономерности интервальными методами даже в очень больших объемах данных (с возможными пропусками) при приемлемых вычислительных затратах.

Обнаруженные закономерности можно использовать для решения практических задач системного анализа, управления и обработки информации.

Эффективность разработанных алгоритмов проиллюстрирована примерами обработки данных при обнаружении неисправностей ультразвуковых расходомеров жидкости, классификации изображений, медицинской диагностике сердечно-сосудистых заболеваний и анализе причин, повлиявших на продолжительность срока выживаемости у больных хроническим лимфолейкозом.

Цель. Повышение эффективности поиска скрытых закономерностей по базам и хранилищам данных и многообразиям структур отношений объектов как нового знания из предметных областей за счёт применения интервальных методов.

Объект исследования. Базы (хранилища) данных из предметных областей.

Методология и методы диссертационного исследования. В диссертационной работе использованы методы интеллектуального анализа данных, нечёткой логики, дискретной оптимизации.

Задачи:

1. Разработать численный алгоритм разделения значений признаков в описании допустимых объектов классов на непересекающиеся интервалы с

использованием предобработки данных при числе интервалов, равном числу классов. Оценить сложность алгоритмов при использовании и без использования предобработки данных.

2. Разработать способ оценивания устойчивости разбиения значений признаков в границах непересекающихся интервалов для выборки данных из двух классов при числе интервалов, больше либо равном двум. Значение устойчивости является обобщающим показателем доминирования представителей объектов классов по каждому интервалу.

3. Разработать способ отбора информативных наборов признаков по выборке данных в целом и для формирования собственного пространства объекта. Исследовать результаты отбора для принятия решений о наличии и виде неисправностей по данным калибровки ультразвуковых расходомеров жидкости [16].

4. Разработать рекомендации по выбору правил для распознавания объектов, формируемых с использованием интервальных методов. Исследовать эффективность такого выбора правил на примерах данных по сегментации изображений из базы [14] и данных по медицинской диагностике больных хроническим лимфолейкозом.

Научные новизна. В диссертационной работе впервые получены следующие **результаты:**

1. Разработан численный алгоритм вычисления экстремума критерия качества разбиения значений признака на непересекающиеся интервалы с использованием предобработки данных. Показано, что оценка сложности алгоритма с использованием предобработки значительно ниже, чем у алгоритма без предобработки. Описан способ выбора границ интервалов при условии, что число различных значений признака равно числу классов.
2. Предложен способ отбора информативных наборов разнотипных признаков для описания объектов класса, новизна которого заключается в применении рекурсивного алгоритма для упорядочивания признаков по отношению

информативности с использованием предобработки данных путём формирования матрицы близости по парам признаков.

3. Разработаны способы использования интервальных методов в рамках информационных моделей, основанных на знаниях.

а) синтезированы латентные признаки, эффективность принятия решений по котором с точки зрения истинности гипотезы о компактности выше, чем по исходным признакам, используемым для их синтеза;

б) способ отбора информативного набора разнотипных признаков для собственного пространства объекта и значение оценки его по этому набору;

в) способ выбора границ между классами как логических закономерностей в форме полуплоскостей;

г) способ формирования *if…then* правил, отбираемых по значениям устойчивости разбиения признака на непересекающиеся интервалы, для классификации объектов;

д) способ вычисления обобщённых оценок объектов по нелинейным преобразованиям признаков с использованием значений функции принадлежности к классам.

Полученные **результаты соответствуют** следующим пунктам **паспорта специальности 05.13.01. – «Системный анализ, управление и обработка информации (**информационные технологии и промышленность**)»**, а именно:

п.5 - разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации (Результаты 1 и 2 Заключения);

п.13 - методы получения, анализа и обработки экспертной информации (Результат 3 Заключения).

Теоретическая значимость диссертационной работы заключается в разработке новых методов предобработки данных для уменьшения комбинаторной сложности алгоритмов интеллектуального анализа данных.

Практическая значимость. Обнаружение скрытых закономерностей с помощью интервальных методов позволяет повысить обобщающую способность

алгоритмов распознавания и обосновывать процесс принятия решений в технических и других системах.

Степень достоверности полученных результатов обосновывается корректным применением математического аппарата, подтверждается вычислительными экспериментами и результатами практического использования.

Основные положения, выносимые на защиту.

– Разработанный численный алгоритм для разбиения значений признаков в описании объектов классов на непересекающиеся интервалы с применением предобработки данных требует значительно меньших вычислительных ресурсов, чем алгоритм без предобработки.

– Показано, что поиск оптимальной эвристики для отбора информативных наборов разнотипных признаков целесообразно проводить на основе результатов минимального покрытия обучающей выборки объектами-эталонами. Для оптимальной эвристики среднее число объектов выборки, описываемое информативным набором и притягиваемое одним эталоном минимального покрытия, имеет максимальное значение и лучшую обобщающую способность при распознавании по алгоритму «ближайший сосед».

– Разработанный рекурсивный алгоритм позволяет вычислять границы интервалов и их число при частичном отсутствии измеренных значений признаков в описании части объектов классов.

– Синтез латентных признаков по операциям умножения и деления значений исходных признаков позволяет увеличить внутриклассовое сходство и межклассовое различие в процессе принятия решений.

Апробация результатов. Основные положения диссертации докладывались на конференциях: V Международная конференция и молодежная школа «*Информационные технологии и нанотехнологии*» (г. Самара, 2019 г.), VI Международная конференция и молодежная школа «*Информационные технологии и нанотехнологии*» (г. Самара, 2020 г.), XI Всероссийская научно-практическая конференция «*Современные проблемы проектирования, производства и эксплуатации радиотехнических систем*» (г. Ульяновск, 2019 г.),

I Всероссийская научно-техническая конференция «*Теоретические и практические аспекты развития отечественного авиастроения*» (г. Ульяновск, 2012 г.), V Всероссийская научно-техническая конференция «*Теоретические и практические аспекты развития отечественного авиастроения*» (г. Ульяновск, 2018 г.), Международная конференция «*Инфокоммуникационные и вычислительные технологии в науке, технике и образовании*» (г. Ташкент, 2004 г.)

Публикации. По теме диссертационной работы опубликовано 16 печатных работ, из них 4 в изданиях из перечня ВАК, 2 Scopus, 1 патент на изобретение.

Внедрение результатов. Результаты диссертационной работы внедрены в гематологическом отделении Государственного учреждения здравоохранения «Ульяновская областная клиническая больница».

Сведения о личном вкладе автора. Постановка задач исследования осуществлялась совместно с научным руководителем. Все основные теоретические и практические исследования проведены автором диссертационной работы самостоятельно. Подготовка к публикации некоторых результатов проводилась совместно с соавторами, вклад соискателя был определяющим.

Структура диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы (наименований) и приложения. Работа изложена на 124 страницах, включающих 12 рисунков и 31 таблицу. Список использованной литературы включает в себя 101наименование.

ГЛАВА 1. Интервальные методы в анализе данных

В данной главе дается формулировка основной решаемой в диссертации задачи, обзор литературных источников в области выявления скрытых закономерностей и применения интервальных методов, описание и обоснование вычислительных алгоритмов для поиска оптимального разбиения значений признаков на непересекающиеся интервалы по двум критериям из [52, 17] (для удобства изложения будем идентифицировать их как CR1 и CR2). В качестве примера приведён вычислительный эксперимент для алгоритма сегментации изображений [14].

1.1. Постановка задачи

Пусть задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, содержащее представителей l непересекающихся классов $K_1,\ldots,K_l$. Описание объектов производится с помощью набора из n разнотипных признаков $X_n = (x_1, \ldots, x_n)$, $\delta(\delta < n)$ из которых измеряются в номинальной, $n - \delta$ в интервальной шкалах. Допускается наличие пропусков и повторяющихся значений в данных. Считается, что задан критерий $F(*)$ для разбиения значений каждого количественного признака (как исходного, так и латентного) на непересекающиеся интервалы. Латентные признаки могут представлять комбинации из номинальных и количественных признаков. Требуется определить значения границ l интервалов, при которых $F(*)=extr$. Совокупность таких границ будем называть оптимальным решением задачи.

При вычислении CR1 число интервалов равно числу непересекающихся классов. Значения границ определяются как произведение внутриклассового сходства и межклассового различия.

Для критерия CR2 число классов равно 2, число интервалов больше или либо равно 2 [52, 17]. Для вычисления границ интервалов, число которых изначально неизвестно, используются «разность частот встречаемости значений признаков (как исходных, так и латентных) в описании объектов двух классов. Значения признаков на числовой оси образуют последовательность кластеров (интервалов)»

[52, 17]. Не существует двух соседних кластеров, в которых доминировали бы представители (по частоте встречаемости) одного класса.

Потребность в разработке нескольких вычислительных (машинных) алгоритмов для реализации процесса оптимизации критериев CR1, CR2 связана с:

– наличием пропусков и повторяющихся значений в данных [90];

– ограничением на число различных значений признаков и непересекающихся классов;

– наличием условий существования разбиения на интервалы.

Реализация интервальных методов ориентирована на поиск скрытых закономерностей в данных. С точки зрения ИАД результаты поиска должны быть [11]:

– «раннее неизвестны;

– нетривиальны;

– практически полезны;

– легко интерпретируемы человеком».

Разбиение значений признаков на непересекающиеся интервалы для вычисления параметров распознающих алгоритмов и выбора описаний допустимых объектов производится с целью [3, 47]:

– снижения размерности пространства (решения проблемы проклятия размерности) путём отбора информативных наборов признаков;

– формирования решающих правил для распознавания;

– формирование процесса принятия решения в слабо формализованных предметных областях.

Алгоритмы решения некорректных задач с использованием методов ИАД, как правило, характеризуются огромной комбинаторной сложностью [37, 38]. Для решения проблемы (уменьшения сложности) в [35] предлагается использовать предобработку данных и некоторые эвристики. Эффективность использования предобработки [51, 53, 66, 90] демонстрируется через оценки сложности алгоритмов разбиения признаков на непересекающиеся интервалы.

1.2. Интервальные методы для обнаружения закономерностей в данных

В этом разделе дается обзор интервальных методов для обнаружения закономерностей. Закономерности – это отношения (явные и скрытые) между свойствами (признаками) объектов. Одним из способов анализа отношений является разбиение на интервалы значений количественных признаков и поиск по ним закономерностей.

Основным направлением современных исследований является разработка и обоснование новых эвристик для методов ИАД, позволяющих производить поиск скрытых закономерностей по базам и хранилищам данных из плохо структурированных предметных областей. Характерной спецификой решения задач в этих областях является как большое количество объектов, так и большая размерность признакового пространства (задачи Bigdata). Есть потребность в разработке специальных методов предобработки данных, форма представления которых адаптируется для реализации уже имеющихся алгоритмов ИАД.

В теории управления наблюдается повышенный интерес к использованию интеллектуальных встроенных систем [31, 32]. Выполнение большей части «проектов» по разработке таких систем отводится ИАД. Процессы «добычи знаний» и принятие управленческих решений находятся в неразрывной связи друг с другом [92].

При моделировании с целью извлечения знаний интервальные методы используются для поиска логических закономерностей по базам (хранилищам) данных. Вычислительные эксперименты по полигонам задач, хранящимся в репозиториях, дают возможность для сравнения алгоритмов по разным критериям. По результатам тестирования на точность предпочтение отдается алгоритмам, использующим нелинейное преобразование признакового пространства для дискриминантных функций и локальные метрики объектов при распознавании по прецедентам [45, 8].

С целью повышения обобщающей способности алгоритмов разрабатываются способы селекции обучающихся выборок. Предложено несколько эвристик,

подбор параметров для которых производится в процессе вычислительного эксперимента [49].

Разрабатываются методы доказательства истинности гипотезы, что в окрестности каждого допустимого объекта существует свое логическая закономерность [36]. Доказательство истинности гипотезы основывается на отображении (визуализации) описаний объектов из исходного пространства на плоскость. Для визуального анализа наличия логических закономерностей в базах данных был предложен метод «локальной геометрии» [22, 24]. Решение проблемы комбинаторной сложности алгоритмов поиска логических закономерностей в рамках метода выглядит следующим образом. Любой объект выборки может рассматриваться как независимый классификатор. Для указанного объекта строится собственное (локальное) пространство признаков, в котором определяется индивидуальная мера его сходства и различия с другими объектами.

«В рамках метода локальной геометрии» [22, 24] используется селекция выборки путём исключения из описания объектов признака или группы признаков. Затем с помощью алгоритма визуализации получают отображение структуры выборки в виде точечного скопления. «Эксперт принимает (субъективное) решение об информативности группы признаков» сравнивая полученные изображения. Основным принципом для сохранения или исключения признаков из описания объектов выборки является «принцип визуального группирования» [22].

Интервальные методы анализа также использовались для визуализации объектов, которые были разбиты на два непересекающихся класса [13]. В форме вычислительного эксперимента было представлено доказательство о равносильности двух числовых шкал для проекции на них описаний объектов по наборам разнотипных признаков. Значения на шкалах представляли обобщённые оценки объектов классов, вычисляемые по стохастическому алгоритму.

Параллельно с развитием методов ИАД интенсивно разрабатываются средства программного обеспечения как для персональных компьютеров и суперкомпьютеров, так и для встроенных систем. Созданы библиотеки с

алгоритмами классификации, визуализации данных, которые входят в Matlab, SPSS, Statistica, SAS Enterprise Miner, Rapid Miner и многие другие популярные пакеты прикладных программ. Разработан и стал доступным для использования ряд хранилищ данных (UCI Machine Learning Repository, GEMLeR, StatLib, KDD cups и др.), по которым можно тестировать работоспособность эвристических алгоритмов при решении практических задач.

Выбор и применение математического аппарата в информационных моделях самым существенным образом зависят от предметной области [22]. Анализ данных в предметных областях существенно ограничен из-за отсутствия теоретического обоснования использования средств контроля за отношениями между объектами, связанными с изменением размерности признакового пространства (проклятие размерности). Проверка на адекватность используемой информационной модели реальному положению вещей определяют в конечном итоге практическую востребованность принимаемых в её рамках решений для слабоструктурированных предметных областей.

Разрабатываются методы выбора пространства из латентных признаков в описании объектов с целью повышения обобщающей способности алгоритмов распознавания [78, 91]. При синтезе латентных признаков из исходных применяются правила иерархической агломеративной группировки, для вычисления мер близости между группами используются интервальные методы. В основе доказательства единственности числа групп (латентных признаков) и состава исходных признаков в них лежит принцип динамического программирования.

Для разбиения значений признаков на интервалы применялись различные эвристики. В ряде случаев число интервалов считалось изначально известным, либо разбиение строилось на предположении, что известна природа среды данных. Интервальные методы использовались для обоснования алгоритмов распознавания по правилам и прецедентам.

При реализации алгоритма линейного дискриминанта Фишера граница (порог между проекциями объектов на числовую ось) между двумя классами

выбиралась из «предположений о нормальном распределении данных выборки» [34]. При выборе порога [60] по критерию для проверки истинности утверждения «Каждый интервал содержит представителей одного класса» удалось повысить обобщающую способность алгоритма без каких–либо предположений о природе среды.

В.Н. Вапником и А. Я. Червоненкисом [25, 26] было доказано, что с ростом размерности признакового пространства увеличивается вероятность корректного разделения классов выборки объектов.

Преимуществом использования дискриминантных функций является отсутствие обучающей выборки для принятия решения. Использование обобщённых функций для этих целей приводит к резкому увеличению размерности пространства в описании объектов. Смысл термина «проклятие размерности» по [34] выражает бесперспективность реализации машинных алгоритмов при относительно небольшой размерности исходного признакового пространства.

Для формирования процесса интуитивного принятия решения было разработано несколько способов получения латентных признаков из исходных разнотипных признаков. Латентные признаки выбирались (использовались) в качестве атрибутов в узлах деревьев решений. Разработано два метода (линейный и нелинейный) для формирования латентных признаков с помощью правил иерархической агломеративной группировки. В основу правил была заложена проверка отношений между значениями признаков на числовой оси с использованием интервальных методов [90].

Постановка задачи о выборе собственного признакового пространства объекта впервые была описана в [7]. Объект рассматривался как центр гипершара, от которого вычислялись расстояния по его локальной метрике до всех объектов выборки. Процесс отбора исходных признаков для собственного пространства был связан с частотным анализом последовательности меток классов объектов, упорядоченных по расстояниям от центра гипершара.

Одной из целей перехода в новое признаковое пространство является визуализация данных. При наличии нескольких методов для визуализации открытым оставался вопрос оценки качества этого перехода. В методе локальной геометрии [35] данные отображались на плоскости двух первых главных компонент [20] для определения наиболее перспективного объекта, относительно которого строилось локальное пространство признаков. В зависимости от выдвигаемых экспертом–исследователем гипотез производился выбор последующих объектов. Одной из таких гипотез могло быть утверждение о количестве объектов класса, лежащих за границами выделяющихся точечных скоплений. Выбор последующих центральных объектов проводилось в соответствии с целью исследования, например, в качестве цели могло быть выбрано изучение объектов, расположенных за границами выделяющихся точечных скоплений.

Идея использования границ интервалов для кодирования признаков содержится в [94]. Реализуется процедура проверки истинности гипотезы о том, что признаки можно дискретизировать через отношение относительных частот встречаемости объектов обучающих выборок из двух классов в границах интервалов. Предполагается, что по значениям частот встречаемости строится интерполяция функций. Для этих функций выполняется свойство унимодальности или монотонности. Каждому выделенному интервалу ставится в соответствие кодовое число. Значения числа определяется через тип признака. Есть особенности представления бинарных признаков. В этом случае выделенные интервалы предлагается использовать в качестве самостоятельных признаков [22].

Н.Г. Загоруйко [45, 46] выдвигались идеи использование компактности объектов классов для оценки обобщающей способности алгоритмов распознавании. Предлагалась производить селекцию обучающих выборок путём отбора эталонных объектов. Часть эталонов рассматривалась как шумовые объекты, которые следовало удалить из выборки.

По результатам экспериментальных исследований к числу лучших с точки зрения точности и обобщающей способности относятся алгоритмы метода

опорных векторов (SVM). Успех SVM объясняется использованием ядерных функций для решения задач классификации и регрессии.

Благодаря ядерным функциям происходит искусственное расширение признакового пространства (переход в спрямляющее пространство), в котором объекты могут быть линейно разделимы. Теоретическое обоснование такого разделения в так называемом VC пространстве содержится в работах Вапника [25, 26]. Практически эти идеи были реализованы в методе SVM, который занимает первые места среди алгоритмов по показателю обобщающей способности.

Существенным недостатком метода SVM является отсутствие формальных правил для выбора ядерных функций при решении конкретных задач. Используются разные эвристики, предпочтительность которых является сомнительной из-за отсутствия строгих критериев отбора.

В [44] описывается технология разведочного анализа данных в обучающих выборках. Технология базируется на поиске в обучающей выборке информативных подмножеств объектов. Информативные подмножества позволяют оценивать различие объектов из разных классов. По результатам разведочного анализа объекты рассматриваются как типичные (наиболее представительные) и нетипичные для своего класса. Информация о типичности (нетипичности) востребована для селекции выборок данных

Согласно [75] признаки в описании объектов делятся на типичные и нетипичные. Такое разделение даётся на основе оценки типичности значений каждого признака. Если частота встречаемости признака во всех классах относительно мала (такие признаки нельзя характеризовать как значимые), то он относится к нетипичным («шумящим»). Граничный по определяемой мере близости объект в ряде случаев может рассматриваться как нетипичный. Для такого объекта велика вероятность совпадение его описания с представителями противоположных классов. Показатель типичности предлагается определять по результатам проверки обобщающей способности алгоритма. Наиболее распространённым на практике методом проверки является кросс-валидация.

Аппарат дискретной математики, «в частности булевой алгебры, теории дизъюнктивных нормальных форм, теории покрытий булевых и целочисленных матриц» [78] используется при поиске информативных наборов признаков. Так, в «алгоритмах вычисления оценок, разработанных Журавлевым Ю.И. и его учениками, находятся оценки ансамблей признаков, которые являются обобщениями коэффициентов информативности», рассмотренных в [41, 42, 43].

«Разбиение значений количественных показателей на интервалы широко применяется в различных алгоритмах анализа данных» [51]. «В прикладной статистике значения количественных признаков, как правило, разбивается на заранее заданное число интервалов. Примером тому служит построение гистограмм, децильного и процентильного распределений» [51].

Можно различать статистические методы по разбиению признаков на интервалы таким образом: деления признаков на равные или неравные интервалы. К числу методов разбиения на равные интервалы можно отнести гистограммы, децильные разбиения и т.д. В работе [25] критерией разбиения на интервалы основан на анализе плотности распределения вероятностей.

Широко используемый на практике метод гистограмм позволяет увидеть закономерности, трудно различимые в простой таблице с набором чисел, благодаря графическому представлению имеющейся количественной информации. Метод чаще всего используется при проведении разведочного анализа данных.

Выбор критерия для оптимального разбиения является существенным для построения гистограммы. Критерий является средством для поиска компромисса между снижением детализации оценки плотности распределения при увеличении числа интервалов и понижением точности её значения при уменьшении интервалов. Гистограммы применяются в основном для визуализации данных на начальном этапе статистической обработки.

Одной из графических форм представления отношений между объектами является децильный коэффициент. Например, показатель дифференциации внутреннего валового продукта (ВВП), выражающий соотношение между ВВП на

душу населения у 7 наиболее высокоразвитых стран и 7 самых слаборазвитых стран мира.

Во многих процедурах статистического анализа данных используется процентное распределение по определенному признаку относительно другого показателя. Распределения на интервалы упорядоченных значений признаков производится на основе определённых критериев. К числу наиболее известных из них относится коэффициент Джини.

Коэффициент Джини (Gini) изначально рассматривался как статистический показатель дифференциации населения отдельно взятой страны или области по определяемому или заданному признаку. Например, по уровню потребления на душу населения. Множество допустимых значений коэффициента принадлежит интервалу [0;1]. Выводы о равномерности распределения показателя делают по его близости к 0.

Наиболее часто в современных экономических расчётах уровень годового дохода берётся в качестве признака для анализа. Для таких случаев коэффициент Джини рассматривается как макроэкономический показатель для анализа разброса денежных доходов населения. Коэффициент служит индикатором равномерности распределения доходов по разным группам населения. Применяется в различных моделях прогнозирования в социологии, экономики, стратегического планирования.

Коэффициент Джини [84] рассчитывается как отношение площадей двух фигур. Конфигурация первой фигуры образованна кривой Лоренца и кривой равенства, второй фигуры – кривыми равенства и неравенства. При расчёте необходимо соблюдать последовательность действий. Сначала вычисляется площадь первой фигуры, а затем её значение делится на площадь второй.

Задача разбиения на интервалы рассматривалась и в теории распознавания образов с учителем. В [25] описан метод, реализация которого основывается на предположениях о законе распределения и числе интервалов. Метод является эвристическим, для разбиения на интервалы используется мера неопределённости

принадлежности объекта к тому или иному классу энтропии, допускается отсутствие разбиения [51].

Разбиение значений признака на непересекающиеся интервалы с учётом заданной классификации объектов описано в [25]. Решение о выборе границ интервалов производится по критерию минимизации энтропии. Энтропия оценивает степень неопределённости выбора.

Производится начальное разбиение значений признака на относительно большое число τ интервалов. Последующие действия заключаются в склеивании соседних интервалов и уменьшения значения τ. Конечной целью вычислений является минимизация энтропии по τ.

Предложенный в [25] алгоритм позволяет найти такое разбиение значений признака на конечное число интервалов при заданной классификации объектов, которое обеспечивает минимальную (или близкую к минимальной) оценку энтропии. Недостатком приведённого алгоритма является отсутствие критерия вычисления значений границ интервалов, начальное разбиение производится произвольным образом, а окончательное разбиение является локально оптимальным.

Использование численных методов оптимизации позволяет подбирать параметры модели, при которых алгоритмы распознавания допускают наименьшее число ошибок на заданной обучающей выборке. Метод, осуществляющий подгонку моделей распознавания и прогнозирования под выборку, получил название минимизации риска [26]. Увеличение сложности модели не всегда является благом, так как «оптимальные» алгоритмы начинают хорошо подстраиваться под конкретные данные, в том числе под масштабы измерений обучающей выборки и погрешность самой модели.

В теории искусственных нейронных сетей (ИНС) сложность модели распознавания выражается через способность к обобщению. В форме вычислительного эксперимента доказана [8] связь размерности признакового пространства и способности алгоритмов распознавания к обобщению. Требуется, чтобы алгоритмы ИНС не только хорошо решали задачу на обучении, но и были

способны также хорошо принимать решение на объектах, которые они не видели в процессе обучения. Этим целям служат разработки новых методов ИАД, позволяющих получать новые знания о решаемой задаче [21, 23, 58, 61, 69, 72] и использовать их, в том числе, и для повышения точности алгоритмов ИНС [53] для произвольных допустимых объектов.

Разбиение на интервалы является составной частью большинства вычислительных методов, используемых для поиска скрытых закономерностей в данных [55, 57] и повышения качества принятия решения в слабо формализованных предметных областях. В [18, 65, 79, 78] интервальные методы использовались для построения обобщённых оценок по заданным и определяемым наборам признаков, для выбора аргументов в узлы дерева решений, вычисления значений функций принадлежности к нечётким множествам. Другим существенным приложением интервальных методов был выбор латентных признаков в описании объектов через синтез нелинейных комбинаций исходных признаков по правилам иерархической агломеративной группировки [91].

В [35, 36] постулируется тезис, что наилучший результат при поиске логических закономерностей можно получить при мелком разбиении (в рамках компьютерных ограничений) признаков на интервалы. Проблема «первого шага» (сегментация признаков) является общей проблемой для традиционных методов. Из-за желания ограничить перебор вариантов для поиска логических закономерностей исследователи вводят специальные ограничения для разбиения признаков на интервалы. По этой причине использование алгоритмов поиска *if…then* правил приводит к ошибке уже на первом шаге их реализации.

Истинность описанного выше тезиса есть смысл рассматривать лишь для неклассифицированных данных. Для доказательства своего тезиса в [35, 36] использовался комплекс тестов. Комплекс тестов использовался для доказательства ограниченности возможностей известных методов ИАД для решения тривиальных задач. Из числа известных были приведены методы для реализации деревьев решений, эвристики для неполного перебора всех

возможных вариантов. Результаты, полученные по этим методам, по большей части можно характеризовать как ложные закономерности.

Использование численных методов для получения оптимальных решений по критериям CR1 и CR2 решает проблему выбора первого шага при поиске логических закономерностей.

1.3. Критерий для разбиения значений признаков на число интервалов, равное числу классов объектов

В теории распознавания образов разбиение объектов на классы базируется на гипотезе о компактности. Согласно этой гипотезе, «близкие» объекты должны лежать в одном классе. Требуется специальное уточнение (разъяснение) понятий близость и компактность объектов.

Не существует единого общепринятого определения понятия «компактность» [46, 48]. Компактность предполагает наличие границы между областями признакового пространства с описанием объектов из разных классов.

В [64, 8] была определена мера компактности объектов непересекающихся классов и выборки в целом со значениями в [0;1]. Вычисление меры основано на анализе отношений между объектами выборки по заданной метрике $\rho(x, y)$. При анализе использовалось разбиение объектов каждого из классов $K_1,\ldots,K_l$, $l\geq 2$, $K_d \cap CK_d = \emptyset$ на непересекающиеся группы $G_{d1},\ldots,G_{dp}$, $p\geq 1$ по отношению связанности «↔». Для любых двух объектов $S_i, S_j \in K_d \cap G_{dt}$, $1\leq t\leq p$ существует путь $S_i \leftrightarrow S_u \leftrightarrow \ldots \leftrightarrow S_j$, $S_u \in G_{dt}$. Мера компактности объектов по классу K_d вычисляется как $\Theta_d = \frac{\sum_{j=1}^{p}|G_{dj}|^2}{|K_d|^2}$, а по выборке в целом как $R(\rho) = \frac{\sum_{i=1}^{l}|K_i|\Theta_i}{m}$.

В одномерном случае можно явно установить границу (в виде значения порога) между классами, так как существует направление (по отношению меньше, больше) на числовой оси. Например, значение порога применяется при распознавании по линейному дискриминанту Фишера [34]. В многомерном случае [64] при вычислении меры компактности объектов классов и выборки в целом используется отношение связанности объектов. Отношение связанности

определяется через множество гипершаров, в пересечении которых содержатся граничные объекты классов по заданной метрике.

Есть потребность в отображении данных для анализа в пространство R^1, R^2. Открываются возможности для визуализации отношений между объектами на прямой и на плоскости.

При выборе признакового пространства для описания объектов возникает вопрос: Какие исходные признаки можно использовать для синтеза латентных? В [90, 13] были предложены правила иерархической агломеративной группировки исходных признаков для нелинейного отображения их значений в описании объектов на числовую ось. В основе правил для попарного объединения признаков при группировке лежит вычисление компактности объектов классов по латентному признаку в границах непересекающихся интервалов.

В качестве данных для интервальных методов рассматривались значения расстояний между объектами. Такие данные использовались при:

– отборе собственного признакового пространства объекта и вычисления его индекса [63, 68, 71];

– выборе латентных признаков по правилам иерархической агломеративной группировки [78]. В форме вычислительного эксперимента было доказано, что латентные признаки образуют упорядоченную по отношению информативности последовательность.

Значения границ интервалов использовалось при выборе порога для реализации линейных дискриминантных функций [60, 62]. Различие такого порога от порога, вычисляемого для линейного дискриминанта Фишера, заключается в отказе от всяких предположений о природе среды данных. Для метода Фишера считается, что данные распределены по нормальному закону. Эффект от использования интервальных методов для вычисления порога заключается в повышении обобщающей способности алгоритмов относительно метода Фишера.

Использование критерия CR1 основывается на проверке следующей гипотезы: «*Существует разбиение на непересекающихся интервалов значений*

количественного признака, произведение мер межклассового различия и внутриклассового сходства для которых равно единице». Идеальный случай при разбиении интерпретируется так:

– в границах каждого интервала лежат все значения объектов одного класса;

– значения критерия CR1 равно 1.

Обозначим через *I, J* множество номеров соответственно количественных и номинальных признаков в описании допустимых объектов, $|I| + |J| = n, I \neq \emptyset$. Поскольку в критерии CR1 используются только количественные признаки (исходные и латентные), то для простоты изложения ограничимся множеством *I.*

Пусть α_{ij} $(\alpha_{ij} > 1)$ – число объектов из класса K_i не имеющих пропусков по x_j, $j \in I$, и число уникальных значений признака в $m_j = \sum_{i=1}^{l} a_{ij}$ объектах больше или равно l. Упорядоченное множество измеренных значений признака x_j, $j \in I$ разобьём на непересекающиеся интервалы $(c_{2k-1}, c_{2k}], c_{2k-1} < c_{2k}, k = \overline{1,l}$,, каждый из которых считается градацией номинального признака.

Пусть u_i^p– множество измеренных значений признака x_j, $j \in I$ класса K_i в интервале $\left(c_{2p-1}, c_{2p}\right]$, $A=(a_0,\ldots,a_l)$, $a_0=0$, $a_l=m_j$, m_j – число объектов без пропусков $(2l \le m_j \le m)$ по x_j, a_p – порядковый номер элемента упорядоченной по возрастанию последовательности $r_1, \ldots, r_{m_j}$ значений x_j из E_0, определяющий правую границу интервала $c_{2p} = r_{a_p}$.

Критерий CR1 [72]

$$\left(\frac{\sum_{p=1}^{l}\sum_{i=1}^{l} u_i^p(u_i^p-1)}{\sum_{i=1}^{l}|\alpha_{ij}|(|\alpha_{ij}|-1)}\right)\left(\frac{\sum_{p=1}^{l}\sum_{i=1}^{l} u_i^p(m_j-|\alpha_{ij}|-\sum_{t=1}^{l} u_t^p+u_i^p)}{\sum_{i=1}^{l}|\alpha_{ij}|(m_j-|\alpha_{ij}|)}\right) \to \max_{\{A\}} \qquad (1.1)$$

позволяет вычислять оптимальные значения границ интервалов $\{(c_{2p-1}, c_{2p}]\}$, $p=1,\ldots,l$ и использовать их для определения градаций количественного признака в номинальной шкале измерений. По выражению в правой скобке в (1.1) вычисляется мера внутриклассового сходства, в левой – межклассового различия.

Смысл критерия CR1 сводится к проверке истинности гипотезы о компактности применительно к количественным (исходным и латентным) признакам [68]. По значению критерия можно упорядочивать признаки по их

вкладу в процесс принятия решения. Очевидно, что признак с большим значением критерия (большим вкладом) будет «лучше», чем с меньшим. В идеале, если значение критерия равно единице, то классификацию можно выполнять по одному признаку (исходному или латентному).

Оптимальное решение задачи является ответом на вопросы:

1. Насколько компактно расположены значения объектов классов на числовой оси?
2. Где проходит границы между объектами классов при компактном размещении?

В основе поиска $F(*)=extr$ лежит вычисление значений мер межклассового сходства и межклассового различия.

Актуальной является решение проблемы синтеза латентных признаков из исходных, с помощью которых можно было бы добиться «идеальной» классификации. В качестве инструмента для такого синтеза предлагается использовать критерий (1.1) в рамках вычислительного эксперимента.

Критерий для вычисления оптимальных границ интервалов используется (см. §3.1) для выбора информативных наборов разнотипных признаков с максимально выраженной независимостью [53]. Свойство независимости признаков используется при синтезе искусственных нейронных сетей с минимальной конфигурацией [39, 72, 70, 73] и для формирования процесса интуитивного принятия решений [68].

Критерий для вычисления оптимальных границ интервалов использовался для выбора информативных наборов разнотипных признаков для решения различных прикладных задач [19, 88, 89, 6, 76].

1.4. Критерий для разбиения на интервалы с доминированием значений признака объектов одного из классов выборки

Препятствием для применения методов прикладной статистики в разведочном анализе данных является разнотипность шкал измерений признаков в описании допустимых объектов из слабо структурированных предметных

областей. Разнотипность шкал измерений не является препятствием для использования методов ИАД. Эти методы востребованы для поиска скрытых закономерностей из баз и хранилищ данных слабо структурированных предметных областей. При доказательстве истинности выдвигаемых экспертами гипотез чаще всего применяется классификация объектов. В качестве нового знания о предметной области может использоваться результаты анализа структуры отношений объектов классов и формы конфигурации границ классов [20, 93, 5].

Для извлечения информации о структуре отношений объектов классов применялись разные способы. Как правило, главной компонентой этих способов являлись меры близости. Например, выводы о форме конфигурации границ классов делались на основе результатов корректного распознавания объектов классов линейными и нелинейными решающими функциями [34, 95].

Знания о структуре отношений объектов в выборке данных можно получить через показатели устойчивости объектов в непересекающихся классах. Устойчивость рассматривается как мера структурного разнообразия объектов. Обоснование этой меры проводилось в рамках непараметрических методов классификации [5] таких как *k* ближайших соседей и парзеновское окно.

Устойчивость характеризует свойства объектов классов в локальных областях признакового пространства. Знания этих свойств применяются для:

– выбора аномальных объектов классов;

– обоснования корректного распознавания объектов выборки;

– отбора объектов для минимального покрытия выборки эталонами при распознавании алгоритмом ближайший сосед.

Выбор метрики в качестве меры расстояния важным моментом для формирования многообразия устойчивости объектов классов [5]. Исследовались и другие подходы для анализа данных. Существует альтернатива использованию метрик для вычисления устойчивости в разнотипном признаковом пространстве. Например, устойчивость размещения каждого из объектов – эталонов локально - оптимального покрытия $\Pi_j = \{S^1, \dots, S^p\}, p > 1$ классов обучающей выборки в

ИНС с минимальной конфигурацией [73] определялась на множестве Π_j методом скользящего экзамена. Методом скользящего экзамена определялась долю некорректно распознанных объектов выборки при удалении эталона из покрытия. Результаты анализа устойчивости нашли применение для селекции обучающих выборок.

В [72] разработан метод вычисления границ непересекающихся интервалов количественных признаков. В границах интервалов доминируют (по частоте встречаемости) значения объектов одного из двух классов. С помощью этого метода реализовано вычисление обобщённых оценок объектов (латентных признаков) в разнотипном признаковом пространстве и меры их устойчивости.

Мера устойчивости в ИНС служит в качестве показателя обобщающей способности нейронов сети на объектах, которых алгоритм «*не видел*» в процессе обучения. Значение меры является индикатором структуры отношений объектов классов в локальных областях признакового пространства.

Латентные признаки, вычисляемые как обобщённые оценки объектов классов, применялись при моделировании процессов и явлений в слабо формализованных предметных областях. Примером такой модели является гомеостатическое равновесия в живых организмах [1].

Рассматривается множество M допустимых объектов, разбитое на l непересекающихся подмножеств (классов) $K_1,\ldots,K_l$. Считается, что представители классов заданы через выборку (подмножество M) объектов $E_0 = \{S_1, \ldots, S_m\}$. Объекты выборки описываются с помощью n разнотипных признаков, множества допустимых значений ξ из которых измеряются в интервальных шкалах, $n - \xi$ в номинальной.

Вычисление устойчивости объектов по значениям исходных и латентных признаков производится относительно отдельных классов [73]. Необходимость организации вычислительного процесса в форме двухклассовой задачи распознавания с объектами из K_t и $CK_t = M \backslash K_t$, $t=1,\ldots,l$ связана с тем, что:

– значение любого количественного признака (исходного и латентного) относительно. Объекты каждого из классов противопоставляются объектам

противоположных классов (например, класс исправного и класс неисправного оборудования);

– отсутствуют наборы аналитических функций для восстановления зависимостей в пространстве разнотипных признаков.

Требуется:

– на множестве допустимых значений каждого из количественных признаков определить разбиение на минимальное число непересекающихся интервалов, в границах которых доминируют значения объектов класса K_t или $CK_t = M \backslash K_t$, $t=1,\ldots,l$;

– вычислить значения меры устойчивости разбиения на интервалы признаков объектов E_0 относительно класса K_t, $t=1,\ldots,l$.

Обозначим через I, J множество номеров соответственно количественных и номинальных (качественных) признаков $X(n)=\{x_1,\ldots,x_n\}$ в описании допустимых объектов, $|I| + |J| = n$. Для удобства выкладок будем рассматривать два класса объектов K_1 и K_2.

Произведём разбиение на интервалы для каждого количественного признака, в границах которых доминируют значения объектов класса K_t или K_{3-t}, $t=1,2$. Для этого упорядочим значения c–го признака ($c \in I$) по возрастанию

$$r_{c_1}, r_{c_2}, \ldots, r_{c_m}. \tag{1.2}$$

Согласно определяемому ниже критерию последовательность (1.2) разбивается на $\tau_c, (\tau_c \geq 2)$ непересекающихся интервалов $\left[r_{c_u}, r_{c_v}\right]^i$, $1 \leq u, u \leq v \leq m$, $\mathrm{i} = \overline{1, \tau_c}$. Значения, лежащие в интервале $\left[r_{c_u}, r_{c_v}\right]^i$, далее могут рассматриваться как градация номинального признака.

Пусть $d_t^i(u,v), d_{3-t}^i(u,v)$ – количество представителей соответственно классов K_t, K_{3-t} интервале $\left[r_{c_u}, r_{c_v}\right]^i$. Для рекурсивной процедуры выбора значений r_{c_u}, r_{c_v} будем использовать критерий CR2 [65]:

$$\left| \frac{d_t^i(u,v)}{|E_0 \cup K_t|} - \frac{d_{3-t}^i(u,v)}{|E_0 \cup K_{3-t}|} \right| \to \max. \tag{1.3}$$

Границы первого интервала $\left[r_{c_u}, r_{c_v}\right]^1$ на последовательности (1.2) вычисляются по максимуму критерия (1.3). Аналогичным образом определяются границы для $\left[r_{c_u}, r_{c_v}\right]^p, p > 1$ на значениях (1.2), не вошедших в $\left[r_{c_u}, r_{c_v}\right]^1, \dots, \left[r_{c_u}, r_{c_v}\right]^{p-1}$. Критерием останова процедуры служит покрытие всех значений (1.2) непересекающимися интервалами [57].

Обозначим через

$$\eta_{1i}(t) = \frac{d_t^i(u,v)}{|E_0 \cup K_t|}, \eta_{2i}(t) = \frac{d_{3-t}^i(u,v)}{|E_0 \cup K_{3-t}|}$$

результаты оптимального разбиения по (1.3) для каждого интервала $\left[r_{c_u}, r_{c_v}\right]^i$, $i = \overline{1, \tau_c}$. Количественно доминирование [57] выражается через значения функции принадлежности $f_t(i) \in [0,1]$ класса K_t ,t=1,2.

Значение функции принадлежности c–го признака к K_1 по интервалу $\left[r_{c_u}, r_{c_v}\right]^i$определим как

$$f_1(i) = \frac{\eta_{1i}}{\eta_{1i}+\eta_{2i}}. \quad (1.4)$$

С учётом того, что $f_t(i)$=1 – $f_{3-t}(i)$, t=1,2, устойчивость признака по множеству интервалов разбиения вычисляется как

$$U(c) = \frac{1}{m}\sum_{\{[r_u,r_v]^i\}} \begin{cases} f_t(i)(v-u+1), & f_t(i) > 0.5, \\ \big(1-f_t(i)\big)(v-u+1), & f_t(i) < 0.5, \end{cases} \quad (1.5)$$

и выражает степень однородности (не перемешанности) значений j-го признака объектов в границах интервалов доминирования ($r_u = r_{c_u},\ r_v = r_{c_v}$), определяемых по (1.3,1.4). Если (в идеале) в границах интервалов лежат значения признака одного класса, то $U(c) = 1$.

Визуальная интерпретация границ интервалов, полученных по (1.3), показана на рис. 1.1, где (u1, v1), (u2, v2), … – индексы упорядоченной последовательности (1.2). Нетрудно заметить, что не существует двух соседних интервалов, в которых доминировали бы представители одного класса.

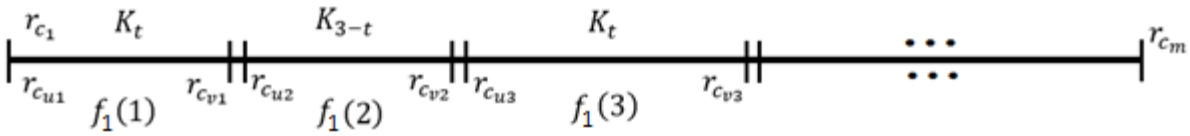


Рисунок 1.1. – Разбиение упорядоченных значений признака на интервалы

1.4.1. Модификация критерия (1.3) для случая наличия пропусков в данных

Рассмотрим модификацию критерия (1.3) для случая наличия пропусков в данных. С учётом пропусков критерий (1.3) примет вид

$$\left|\frac{d_t^i(u,v)}{T_p^c} - \frac{d_{3-t}^i(u,v)}{T_{3-p}^c}\right| \to \max, \qquad (1.6)$$

где T_p^c, T_{3-p}^c – количество значений признака $x_c \in X(n)$ без пропусков у объектов E_0 соответственно из классов K_p и $K_{3\text{-}p}$. Естественным условием для реализации (1.6) является:

– число различных значений признака больше или равно 2;

– значения $T_p^c > 0, T_{3-p}^c > 0$.

С учётом пропусков в данных значение устойчивости (см. (1.5)) будет выглядеть так

$$U(c) = \frac{1}{\mu}\sum_{\{[r_u,r_v]^i\}} \begin{cases} f_t(i)(v-u+1), & f_t(i) > 0.5, \\ \big(1-f_t(i)\big)(v-u+1), & f_t(i) < 0.5, \end{cases} \qquad (1.7)$$

где $\mu = T_p^c + T_{3-p}^c$.

Примером формирования латентного признака из двух исходных, один из которых измеряется в количественной, а другой в номинальной шкале, может быть следующий. Пусть x_i, $x_j \in X(n)$, $i \in J$, $j \in I$ и признак x_i имеет 2 градации. Тогда для получения латентного признака в виде произведения $x_i x_j$ значения признака x_i нужно выбирать из {-1,1}.

Разбиения на интервалы по (1.3) и (1.6) дают возможность для наглядного представления знаний в виде дизъюнкций элементарных конъюнкций. Элементарные конъюнкции нужны для проверки принадлежности значения признака к одному из интервалов. Запись правила для отнесения объекта классу

K_t, t=1,2 может иметь такой вид: $a_1 \le x_i \le b_1$ or $a_2 \le x_i \le b_2$ or ….or $a_{\eta-1} \le x_i \le b_{\eta-1}$, где $a_j, b_j, j \in \{1, \eta\}$ – границы интервалов, η – число непересекающихся интервалов.

Значения устойчивости по (1.5) или (1.7) служат индикатором для использования разбиения на интервалы в качестве нового знания. В диссертации рекомендуется считать результаты анализа новым знанием при значении устойчивости из [0.9,1] и числе интервалов не больше 4. Как указывается в [36] эффективность системы для поиска if…then правил зависит от:

– затрат вычислительных ресурсов для обнаружения закономерностей;

– заданной точности правил для принятия решений.

Проблема повышения эффективности [35, 36] является целью для разработки новых и обоснования имеющихся подходов к решению задачи поиска логических закономерностей в данных.

1.5. Вычислительный эксперимент

Для вычислительного эксперимента [57] с целью апробации предложенных в диссертации способов поиска скрытых закономерностей были использованы данные по сегментации изображений [14] из UCI Machine Learning Repository. Выборка [14] состоит из 2300 объектов (изображений, сделанных на улице) по 19 признаков. Изображения этой выборки были вручную сегментированы (разбиты) экспертами на 7 классов (кирпич, небо, листва, цемент, окно, дорога, трава.). Эти тестовые данные предназначены для оценивания эффективности методов поиска сегментов различных типов на изображениях.

В диссертации при проведении эксперимента выбирался один класс объектов (K_1), все остальные объекты считались принадлежащими классу K_2. Результаты разбиения на интервалы по (1.3) и устойчивости по (1.5) при выборе в качестве объектов K_1 изображений «кирпич» приведены в табл.1.1.

Таблица 1.1. Результаты разбиения на интервалы при выборе в качестве класса K_1 изображения «кирпич»

№	Название признака	Границы Интервалов	Значение функции принадлежности к K_1	Устойчивость разбиения по (1.5)
1	region-centroid-col (столбец центрального пикселя области)	[1, 151]	0.5987	0.6557
		[152, 254]	0.2539	
2	region-centroid-row (строка центрального пикселя области)	[11, 50]	0.1533	0.7889
		[51, 149]	0.6607	
		[150, 251]	0	
3	region-pixel-count (количество пикселей в области = 9)	Нет	0	0
4	short-line-density-5 (результаты алгоритма экстракции линии, контраст, меньше или равный 5)	[0, 0]	0.4863	0.5222
		[0.1111, 0.3333]	0.5856	
5	short-line-density-2 (Результаты алгоритма экстракции линии, контраст больше 5)	[0,0]	0.5089	0.5214
		[0.1111, 0.2222]	0.1714	
6	vedge-mean (Среднее значение измерения контраста по горизонтали, используется как детектор вертикального края)	[0,0.2777]	0.1923	0.6116
		[0.2778, 0.6111]	0.750769	
		[0.6111, 29.2222]	0.4305	
7	vegde-sd (Стандартное отклонение измерения контраста по горизонтали, используется как детектор вертикального края)	[0, 0.0333]	0.2857	0.6181
		[0.0333, 0.4333]	0.6797	
		[0.4333, 991.718]	0.4102	
8	hedge-mean (Среднее значение измерения контраста вертикально смежных пикселей, используется для определения горизонтальной линии)	[0, 0.3333]	0.1046	0.6259
		[0.3333, 2.9444]	0.5662	
		[3, 44.7222]	0.2434	

№	Название признака	Границы Интервалов	Значение функции принадлежности к K_1	Устойчивость разбиения по (1.5)
9	hdge-sd (Среднее отклонение измерения контраста вертикально смежных пикселей, используется для определения горизонтальной линии)	[-1.5e-008, 0.0296]	0	0.5981
		[0.0296, 0.4444]	0.6661	
		[0.4554, 1386.33]	0.4406	
10	intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	[0, 3.8889]	0.0179	0.8860
		[3.9259, 28.6296]	0.7443	
		[28.7407, 143.444]	0	
11	rawred-mean (среднее значение по области значения R)	[0,5.3333]	0.0956	0.8903
		[5.4444, 26.1111]	0.7685	
		[26.3333, 137.111]	0	
12	rawblue-mean (среднее значение по области значения B)	[0, 4.6667]	0.0453	0.8525
		[4.7778, 36.2222]	0.7207	
		[36.3333, 150.889]	0.0298	
13	rawgreen-mean (среднее значение по области значения G)	[0, 1.6667]	0	0.9103
		[1.7778, 20.6667]	0.7794	
		[20.7778, 142.556]	0.0104	
14	exred-mean (избыток красного: (2R - (G + B)))	[-49.6667, -5.6667]	0.0790	0.8952
		[-5.5556, 7.2222]	0.8327	
		[9.8889, 9.8889]	0	
15	exblue-mean (избыток синего: (2B - (G + R)))	[-12.4444, 0.5556]	0.0316	0.8365
		[0.6667,23]	0.7494	
		[23.1111, 82]	0.1342	
16	exgreen-mean (избыток зеленого: (2G - (R + B)))	[-33.8889, -19.8889]	0.0933	0.8148
		[-19.7778, -6.3333]	0.6918	
		[-6.2222, 24.6667]	0.0441	

№	Название признака	Границы Интервалов	Значение функции принадлежности к K_1	Устойчивость разбиения по (1.5)
17	value-mean (среднее значение: трехмерное нелинейное преобразования RGB)	[0,5.3333]	0	0.8588
		[5.4444, 36.2222]	0.7230	
		[36.3333, 150.889]	0.0298	
18	saturatoin-mean (среднее значение насыщенности нелинейного преобразования RGB)	[0, 0.3679]	0.0052	0.9034
		[0.3688, 0.6170]	0.8057	
		[0.6176, 1]	0.1699	
19	hue-mean (среднее значение оттенка нелинейного преобразования RGB)	[-3.0442, -1.8905]	0.0190	0.9825
		[-1.8884, -0.5709]	0.9716	
		[-0.0049, 2.9125]	0	

По результатам из табл. 1.1. устойчивость по (1.5) больше 0.9 у признаков с номерами 13, 18, 19. Согласно выше указанным рекомендациям, границы интервалов этих признаков можно использовать в качестве нового знания об объектах класса K_1 (кирпич). Например, при формировании if …then правил в базах знаний. Значение $U(3)=0$ по (1.5), так как не существует интервалов (для признака region-pixel-count) в границах которых доминируют по (1.5) представители одного из двух классов.

Наиболее значимые результаты по устойчивости разбиения признаков на интервалы при использовании других способов классификации изображений приведены в табл. 1.2.

Таблица 1.2. Устойчивость разбиений признаков на интервалы

Название класса K_1	**Название признаков**	**Границы интервалов**	**Значение функции принадлежности к K_1**	**Устойчивость разбиения по (1.5)**
Небо	intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	[0, 79.6667]	1	1
		[86.2963, 143.444]	0	
	rawred-mean (среднее значение по области значения R)	[0, 65]	1	0.9999
		[70.6667,137.111]	0.0005	
	rawblue-mean (среднее значение по области значения B)	[0,91.4444]	1	1
		[112.111,150.889]	0	
	rawgreen-mean (среднее значение по области значения G)	[0,66.4444]	1	0.9999
		[76.1111,142.556]	0.0005	
	value-mean (среднее значение: трехмерное нелинейное преобразования RGB)	[0, 91.4444]	1	1
		[112.111, 150.889]	0	
	19.hue-mean	[-3.0442,-2.4590]	0.75	0.9154
		[-2.45498,-2.1669]	0.1320	
		[-2.1662,2.9125]	0.9358	

Название класса K_1	**Название признаков**	**Границы интервалов**	**Значение функции принадлежности к K_1**	**Устойчивость разбиения по (1.5)**
Путь	region-centroid-row (строка центрального пикселя области)	[11, 158]	1	0.9828
		[159, 210] 171	0.0795	
		[211,251]	1	
	intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	[0, 29.5926]	0.9782	0.9576
		[29.6296, 63.5185]	0.1172	
		[63.6296, 143.444]	1	
	rawblue-mean (среднее значение по области значения B)	[0, 36.2222]	0.9735	0.9491
		[36.3333, 79.5556]	0.1325	
		[79.6667, 150.889]	1	
	rawgreen-mean (среднее значение по области значения G)	[0, 26.4444]	0.9688	0.9461
		[26.5556,55.6667]	0.1340	
		[56,142.556] 374	1	
	value-mean (среднее значение: трехмерное нелинейное преобразования RGB)	[0,36.2222] 1323	0.9735	0.9491
		[36.3333,79.5556]	0.1325	
		[79.6667,150.889]	1	
	saturatoin-mean (среднее значение насыщенности нелинейного преобразования RGB)	[0, 0.262093]	0.8569	0.9283
		[0.262981, 0.32108]	0.1053	
		[0.321139, 1]	0.9652	

Название класса K_1	**Название признаков**	**Границы интервалов**	**Значение функции принадлежности к K_1**	**Устойчивость разбиения по (1.5)**
Трава	region-centroid-row (строка центрально-го пикселя области)	[11,155]	0.9784	0.9405
		[156,251]	0.1518	
	intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	[0, 6.3704]	0.9345	0.9112
		[6.4074, 25.963]	0.1909	
		[26, 143.444]	0.9769	
Трава	rawred-mean (среднее значение по области значения R)	[0, 3.8889]	1	0.9157
		[4, 4.2222]	0.3023	
		[4.3333, 4.7778]	1	
		[4.8889, 21.2222]	0.2	
		[21.3333, 137.111]	0.9726	
	rawblue-mean (среднее значение по области значения B)	[0, 4.1111]	0.9745	0.9026
		[4.2222, 25.4444]	0.2187	
		[25.5556, 150.889]	0.9757	
	rawgreen-mean (среднее значение по области значения G)	[0, 8.2222]	0.9813	0.9190
		[8.3333, 33.4444]	0.2003	
		[33.5556, 142.556]	0.9863	
	exblue-mean (избыток синего: (2B - (G + R)))	[-12.444, -0.2222]	0.0006	0.9319
		[0, 82]	0.9218	
	exgreen-mean (избыток зеленого: (2G - (R + B)))	[-33.8889, 0.3333]	0.9910	0.9921
		[1.7778, 24.6667]	0.0010	
	hue-mean (среднее значение оттенка нелинейного преобразования)	[-3.0442, -3.0442]	1	0.9948
		[-3.0151, -3.0151]	0	
		[-2.8700, 0]	0.9940	
		[1.2871, 2.9125]	0	

Использование в качестве K_1 изображений «Листва», «Цемент», «Окно» не дало закономерностей с устойчивостью разбиения признаков, значения которых принадлежат интервалу [0,9;1]. Для класса «Трава» устойчивые закономерности были получены для 8 признаков из 19.

Результаты разбиения на интервалы по (1.2) могут служить основой для формирования лингвистических правил и наполнения баз знаний. Для формирования правил имеет значение такие показатели как количество интервалов разбиения, значения функции принадлежности, устойчивость разбиения. Очевидно, чем меньше интервалов доминирования и выше устойчивость разбиения, тем сильнее выражена закономерность на конкретном признаке в классе. Эта особенность интервального метода может быть использована для проведения ранжирования количественных показателей в информационных моделях прикладных задач [73]. Для поиска и удаления шумящих признаков предлагается использовать значения рангов.

Покажем возможность обнаружения скрытых закономерностей с помощью (1.6), (1.7) при наличии пропусков в данных. Номера объектов из [14], содержащих пропуски, определим по датчику случайных чисел. Будем использовать разделение выборки на класс «трава» (K_1) и другие (K_2) с количеством пропущенных значений 10%, 20%, 30%. Для демонстрации выбраны признаки из табл. 1.2 со значением устойчивости (1.5) в (0.9;1]. В табл. 1.3 приведены результаты разбиения на интервалы при количестве пропусков, равном 10%.

Таблица 1.3. Результаты разбиения при количестве пропусков, равном 10%

Признаки	**Интервалы**	**Значение функции принадлежности к K_1**	**Значение (1.7)**
region-centroid-row (строка центрального пикселя области)	[11,155]	0.9760	0.9389
	[156,251]	0.1506	
intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	[0,6.37037]	0.9385	0.9134
	[6.4074, 25.963]	0.1855	
	[26, 143.444]	0.9746	
rawred-mean (среднее значение по области значения R)	[0,3.88889]	1.0	0.9133
	[4,21.2222]	0.2027	
	[21.3333,136.889]	0.9700	
rawblue-mean (среднее значение по области значения B)	[0,4.1111]	0.9709	0.9070
	[4.22222,26.1111]	0.2190	
	[26.2222,150.889]	0.9890	
rawgreen-mean (среднее значение по области значения G)	[0,6.7778]	1.0	0.9224
	[7,33.4444]	0.1977	
	[33.5556,142.556]	0.9848	
exblue-mean (избыток синего: (2B - (G + R)))	[-12.4444,-0.2222]	0.0006	0.9300
	[0,79.4444]	0.9193	
exgreen-mean (избыток зеленого: (2G - (R + B)))	[-33.8889,0.3333]	0.9933	0.9941
	[1.7778,24.6667]	0.0011	
hue-mean (среднее значение оттенка нелинейного преобразования RGB)	[-3.0442,0]	0.9933	0.9943
	[1.28706,2.91248]	0.0	

Сравнительный анализ результатов из табл. 1.2 и табл. 1.3. показывает:

– близость значений устойчивости по (1.5) и (1.7) по всем признакам;

– различие по числу интервалов разбиения у признаков 11.rawred-mean, 15.exblue-mean, 19.hue-mean.

Различие в числе интервалов объясняется наличием «шумящих» значений признаков в описании объектов классов. Из анализа результатов следует, что для обнаружения скрытых закономерностей показатели устойчивости (1.5) и (1.7) имеют приоритетное значение, чем число непересекающихся интервалов разбиения. Влияние 20% и 30% пропусков в данных на результаты устойчивости (1.7) показаны в табл. 1.4.

Таблица 1.4. Значение устойчивости при количестве пропусков 20% и 30%

Признаки	**Пропуски 20%**		**Пропуски 30%**	
	Число интервалов	**Значение (1.7)**	**Число интервалов**	**Значение (1.7)**
region-centroid-row (строка центрального пикселя области)	2	0.9389	2	0.9409
intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	3	0.9134	3	0.9166
rawred-mean (среднее значение по области значения R)	3	0.9133	3	0.9121
rawblue-mean (среднее значение по области значения B)	3	0.9070	5	0.9097
rawgreen-mean (среднее значение по области значения G)	3	0.9224	3	0.9204
exblue-mean (избыток синего: (2B - (G + R)))	2	0.9300	2	0.9275
exgreen-mean (избыток зеленого: (2G - (R + B)))	2	0.9941	2	0.9924
hue-mean (среднее значение оттенка нелинейного преобразования RGB)	2	0.9942	2	0.9926

Анализ результатов табл. 1.2, 1.3, 1.4 оставляет открытой проблему измерения границ интервалов и выбору по ним логической закономерности в форме параллепипеда для реализации *if…then* правил. Вычисление границ интервалов по (1.3) и (1.6) производится жадным алгоритмом. Свойством жадного

алгоритма подстраиваться под структуру данных объясняется различие интервалов при отсутствии или наличии пропусков.

Для обоснования использования закономерностей в качестве *if…then* правил при числе интервалов больше либо равном 3 предлагается делить выборку на два непересекающихся подмножества E_{ob} и E_k. Результаты разбиения на интервалы на E_{ob} применяются в качестве правила для подсчёта числа ошибок на E_k. Приемлемым считается разбиение, отвечающее следующим требованиям:

– значение устойчивости на E_{ob} принадлежит (0.9;1];

– число ошибок на E_k минимально.

Таким образом, на данных по сегментации изображений продемонстрировано утверждение, что показатели устойчивости имеют более приоритетное значение для обнаружения скрытых закономерностей, чем число непересекающихся интервалов.

Выводы по главе 1

1. Для поиска скрытых закономерностей с учётом пропусков в данных предложено использовать два критерия разбиения упорядоченных значений количественных признаков на интервалы в описании допустимых объектов непересекающихся классов. При обнаружении закономерностей используются значения: произведения внутриклассового сходства и межклассового различия при числе интервалов, равном числу классов; устойчивость разбиения при числе интервалов больше либо равном 2, определяемая по отношению доминирования в каждом интервале представителей одного из двух классов. Принадлежность значения устойчивости к (0.9; 1.0] является условием формирования логических выражений для *if…then* правил;

2. Показатели устойчивости (1.5) и (1.7) имеют более приоритетное значение для обнаружения скрытых закономерностей, чем число непересекающихся интервалов, полученных по критериям (1.3) и (1.6). При отсутствии и наличии пропусков значений признаков у 10% объектов, представляющих описание изображений, максимальное отклонение устойчивости между (1.5) и (1.7) было во

втором знаке, максимальная разница между числом интервалов разбиения равнялась 3.

ГЛАВА 2. Выбор методов принятия решений

В данной главе разрабатываются способы снижения сложности алгоритмов вычисления границ интервалов и реализация отбора информативных признаков на их основе за счет предобработки данных и иерархической агломеративной группировки признаков.

Основной проблемой для поиска скрытых закономерностей в данных является комбинаторная сложность алгоритмов. Для решения этой проблемы предлагается использовать предобработку данных с учётом особенностей представления описаний объектов классов. К особенностям представления относятся: наличие пропусков и повторяющихся значений в данных; использование различных способов синтеза латентных признаков из исходных, замеренных в интервальной и номинальной шкалах.

В [36] рассматривалось введение контекстно-зависимых локальных метрик в качестве мер расстояния от указанных объектов выборки. На основе мер расстояния предлагалось решать задачу поиска логических закономерностей. Средством для решения задачи поиска были выбраны методы линейной алгебры и интерактивной графики. Визуализация данных открывала возможности для селекции объектов выборки. Например, часть объектов по результатам визуального анализа рассматривалась как шумовые.

Несмотря на декларируемую простоту, в методе локальной геометрии [36] не учитываются разнотипность шкал измерений. Поиск логических закономерностей на модельных примерах демонстрируется на пространства относительно малой (10≤) размерности.

2.1. Сложность реализации алгоритмов по критерию (1.1)

В теории информатики сложность алгоритма определяется как функция оценки затрат ресурсов для получения результата в зависимости от размера входных данных. Как правило, в качестве ресурсов используется время для расчёта в условных единицах и объём требуемой памяти. Время определяется

количеством элементарных операций, необходимых для решения задачи, которое зависит как от объема входных данных, так и от значений самих данных.

В [51] рассмотрены оценки сложности алгоритмов выбора оптимальных границ интервалов разбиения значений признаков в задачах распознавания с учителем по критерию (1.1). Было показано, какой выигрыш получает алгоритм оптимизации критерия, использующий предобработку данных и не использующий её.

Теоретическая оценка сложности в [51] определена для случая с несовпадающими значениями признака. В [90] рассмотрен случай, когда часть данных не измерена («пропуски в данных») и в которых встречаются повторяющиеся значения.

В реальной практике чаще всего приходится иметь дело с Bigdata (большие данные). Большие данные обычно определяют, как сумму структурированных и неструктурированных данных постоянно растущих объемов. Практический интерес представляют методы обработки таких данных в распределенной вычислительной сети. Имеет значение учет многообразия форматов и источников данных. Технология обработки направлена на получение результатов доступных для интерпретации их пользователем. Интерпретируемость результатов можно повысить через создание дружественного интерфейса и использования терминов из баз данных предметных областей [98].

Особенность работы со слабо структурированными данными проявляется в том, что из них трудно формировать выборки для уже имеющихся методов анализа. Из-за комбинаторной сложности алгоритмов методов практически невозможно получить результаты за приемлемое время [98].

Наличие методик обработки больших данных проявляется в том, что из них трудно формировать выборки для уже имеющихся методов анализа. Из-за комбинаторной сложности алгоритмов методов используются различные эвристики для того чтобы получить результаты за приемлемое время. Практическая ценность этих методик выражается в возможности обнаружения закономерностей с минимальными затратами вычислительных ресурсов.

Предполагается, что ценность полученных результатов может быть определена по специальным критериям качества.

Если данные имеют большой размер (как правило, больше 1000 объектов), то вычисление экстремума критерия (1.1) становится практически неприемлемым, т.к. объём вычислений растет экспоненциально.

Поскольку проверка наличия пропусков и повторяющихся значений в данных связана с затратами дополнительных ресурсов [90], имеет смысл показать, при каком их количественном соотношении достигается эффект снижения сложности алгоритмов. Учет таких соотношений необходим при принятии решений и выборе эффективных алгоритмов в информационных системах. Исследование этой проблемы предлагается проводить через оценки сложности алгоритмов.

Первая теоретическая оценка сложности алгоритмов $F\left(l,m\right)$ для вычисления экстремума (1.1) была сделана в [51]. При вычислении оценки предполагалось, что при описании объектов нет пропусков (неизмеренных значений) в данных и все значения каждого признака отличны друг от друга. Оценка сложности представляет произведение количества разбиений значений признака на непересекающиеся интервалы и операций, необходимых для определения значений u_p^t (частоты встречаемости представителей класса K_p в t–ом интервале) при каждом из них. Количество непересекающихся интервалов в (1.1) определяется как

$$\psi\left(l,m\right)=lC_{m-1}^{l-1}. \tag{2.1}$$

При числе классов l=2 $\psi\left(l,m\right)=2\left(m-1\right)$. Оценка сложности вычислялась как

$$F\left(l,m\right)=\left(2m\left(\frac{1+l}{2}\right)+m\right)\psi\left(l,m\right)=m\left(2+l\right)\psi\left(l,m\right), \tag{2.2}$$

в которой значение $m(2+l)$ представляет количество элементарных операций для подсчёта частоты встречаемости представителей класса в интервале.

Обозначим через $q=q_1+q_2$– сумму повторяющихся q_1 и пропущенных q_2 значений признака $x_j \in X(n)$, $j \in I$. При вычислении максимального числа интервалов нужно будет рассматривать значение q как отдельный параметр в модифицированной формуле (2.1)

$$\psi(l,m,q)=lC_{m-1-q}^{l-1}. \tag{2.3}$$

При $q\neq 0$ q_2 с учётом (2.3) вычисление оценки сложности будет таким:

$$F(l,m,q)=\begin{cases} m(2+l)\psi(l,m,q)+2m, & q_1=0, q_2>0; \\ m(2+l)\psi(l,m,q)+C_{m-1-q}^{l-1}, & q_1>0, q_2=0; \\ m(2+l)\psi(l,m,q)+2m+C_{m-1-q}^{l-1}, & q_1>0, q_2>0. \end{cases} \tag{2.4}$$

Для уменьшения комбинаторной сложности вычислений экстремума (1.1) в [51] предложено проводить предобработку данных. С учётом постановки задачи в процессе предобработки по упорядоченной последовательности значений признака $r_1,\ldots,r_{m_j}$, $m_j \le m$ формируется целочисленная матрица вида:

$$D=\begin{pmatrix} d_{10}d_{11}\ldots d_{1m} \\ \ldots\ldots\ldots\ldots\ldots \\ d_{l0}d_{l1}\ldots d_{lm} \end{pmatrix}, \tag{2.5}$$

в которой индекс столбца элемента d_{pi}, $p=1,\ldots,l$, $i=1,\ldots,m$ соответствует объекту $S \in E_0$ со значением признака r_i. Элементы матрицы (2.5) вычисляется как

$$d_{pi}=\begin{cases} 0, i=0, \\ d_{p,i-1}+g(p,i), i>0, \end{cases} \text{ где } g(p,i)=\begin{cases} 1, S\in K_p, \\ 0, S\notin K_p. \end{cases}$$

Число представителей u_t^p класса K_p, $p=1,\ldots,l$, $t=1,\ldots,l$ в интервале $[c_1,c_2]$ при $t=1$ и $(c_{2t-1},c_{2t}]$ определяются как

$$u_t^p=d_{p\nu}-d_{p\eta}, \tag{2.6}$$

где $\eta=a_{t-1}$, $\nu=a_t$, $c_{2t-1}=r_{j\eta}$, $c_{2t}=r_{j\nu}$.

В [90] предобработка данных по (2.5), (2.6) даёт такие оценки сложности

$$R(l,m,q)=\begin{cases} l\psi(l,m,q)+2m, & q_1=0, q_2>0; \\ l\psi(l,m,q)+C_{m-1-q}^{l-1}, & q_1>0, q_2=0; \\ l\psi(l,m,q)+2m+C_{m-1-q}^{l-1}, & q_1>0, q_2>0. \end{cases} \tag{2.7}$$

По результатам вычислительных экспериментов в [90] было показано, что при числе классов $l>2$ и заданном числе объектов m можно определить минимальное значение q, начиная с которого сложность алгоритма начинает уменьшаться относительно «идеального случая». Считается, что в идеальном случае нет ни пропусков в данных ни повторяющихся значений признака и соответственно нет затрат вычислительных ресурсов на их обнаружение. Это свойство может быть учтено при разработке машинных алгоритмов для решения многих прикладных задач, наличие и отсутствие пропусков в данных или повторяющихся значений в которых изначально известно.

Демонстрация [51] на тестовом примере одного из вариантов разбиения на интервалы несовпадающих значений количественного признака по критерию (1.1) при m=16, числе классов l=3 и мощности классов $|K_1|$=6, $|K_2|$=6, $|K_3|$=4 показана на рис. 2.1.

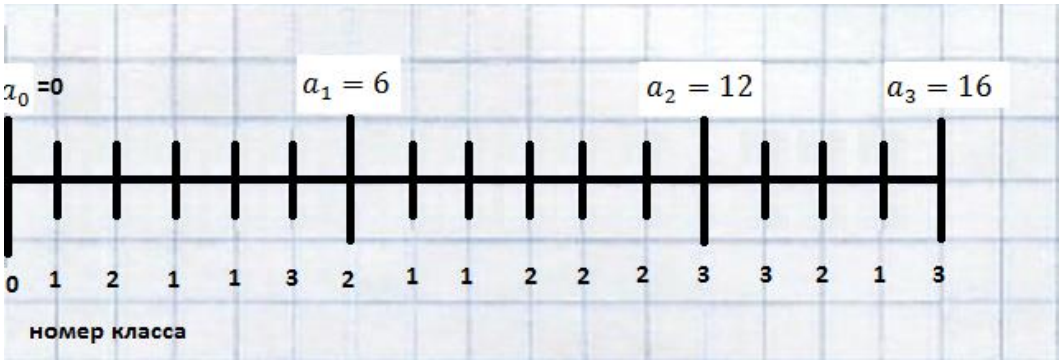


Рисунок 2.1 – Вариант разбиения значений признака на интервалы

Результаты предобработки упорядоченных данных тестового примера (см. рис.2.1) в виде значений матрицы (2.5) выглядят так

$$D=\begin{pmatrix} 0 & 1 & 1 & 2 & 3 & 3 & 3 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 6 & 6 \\ 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 4 & 5 & 5 & 5 & 6 & 6 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 3 & 3 & 3 & 4 \end{pmatrix}.$$

Согласно (2.1) максимальное число вариантов разбиения значений признака на интервалы $\psi(3,16)=2\times13\times14=364$, сложность алгоритма без предобработки $F(3,16)=16\times(2+3)\times364=29520$ с предобработкой (учитывая вычисление (2.7))

$3\times 364+3\times 16=1140$. Три варианта разбиения данных из тестового примера на интервалы представлены в табл. 2.1.

Таблица 2.1. Варианты разбиения на интервалы данных тестового примера

№ п/п	a_1	a_2	**Значение критерия (1.1)**
1	1	2	0.1944
2	2	8	0.3452(оптим.)
3	6	12	0.2146

Очевидно, что при оптимальном разбиении (a_1=2 и a_2=8) нет ни одного интервала, содержащего все значения признака объектов одного класса. При использовании описанной выше предобработки стало возможным вычисление критерия (1.1) для интервалов и весов латентных (явно неизмеримых) признаков. Под весом здесь понимается оптимальное значение критерия (1.1). Примерами латентных признаков может служить $x_i x_j$, $x_i x_j^{-1}$, где $x_{i,}x_j \in X(n)$ и $i,j\in I$. Рассматриваются только линейные зависимости, так как любая непрерывная функция разлагается в ряд Тейлора. Высокие значения весов латентных признаков (как правило, ближе или равные 1) служать основанием для построения моделей интуитивного принятия решений. На практике латентные признаки часто используются в форме различных индексов.

Вес признака по критерию (1.1) содержит в себе важную информацию об его информативности [51]. Однако при отборе информативных наборов признаков нельзя полностью полагаться только на их упорядочение по значениям весов, то есть руководствоваться принципом «чем больше вес, тем признак более информативный в наборе». В расчёт идёт и такой фактор как взаимная коррелированность признаков. Такая задача рассматривалась в [53], где исследовался вопрос отбора наборов информативных разнотипных признаков и их влияние на эффективность реализации искусственных нейронных сетей.

К числу особенностей реализации алгоритмов оптимизации критерия (1.1) можно отнести случай, когда число отличных друг от друга значений равно числу классов. Предлагается следующий алгоритм решения проблемы. Пусть a_1, a_2,…,a_l

– упорядоченные в порядке возрастания значения признака для задачи с $K_1,\ldots,K_l$ непересекающимися классами объектов. Значения l–1 границы $[a_1,c_1]$ $(c_1, a_2]$ $(c_{l-1}, a_l]$ определяются как $c_i = \frac{a_i + a_{i+1}}{2}$, i=1,…,l-1. Оценка сложности алгоритма по (2.7) для (1.1) будет минимальной. Для её вычисления необходимо будет подсчитать число объектов классов $K_1,\ldots,K_l$, принимающих значения $a_1,a_2,\ldots,a_l$.

2.2. Поиск закономерностей по границам интервалов

На проблемы поиска и обоснования логических закономерностей ориентированы работы многих исследователей [28, 67, 80, 82, 86, 87]. Например, с позиций геометрического подхода Дюк В.А. для анализа структуры логических закономерностей в [36] предложил использовать локальные метрики. В качестве инструментов для анализа рекомендуются методы снижения размерности признакового пространства и правила иерархической агломеративной группировки. Комбинация этих методов даёт возможность получать наглядные визуальные представления о наличии логических закономерностей. По визуальному представлению можно делать выводы о геометрической структуре закономерностей. Результаты вычислительных экспериментов позволяют проводить поиск нового знания в хранилищах данных. Например, деревья, получаемые с помощью правил агломеративной иерархической группировки [40, 54], отображают структуру логических закономерностей. Иерархия структуры при выборе логических правил позволяет формировать из них понятия и метапонятия.

С позиций геометрического подхода в [35] рассмотрены варианты представления логических правил в терминах нечётких множеств. Нечёткость логических правил в методе локальной геометрии реализуется через операции на нечётких интервалах. Определена мера расстояния от объекта до логического правила для набора из количественных признаков. Мера расстояния вычисляется через смещение границ интервалов, описываемых элементарными логическими событиями. Автор приводит свою интерпретацию нечеткости. Традиционное

представление функции нечёткости определялось на основе субъективных оценок экспертов. Вместо этого в [36] предлагается использовать в качестве значений функции эмпирические распределения расстояний объектов выборки до логического правила.

К числу основных форм логических закономерностей относятся параллелипипед, гипершар и полуплоскость. Рассмотрим пример двухклассовой задачи распознавания. Значения количественного (исходного или латентного) признака могут быть использованы для выбора порога [62] с целью разделения объектов классов как $R = \frac{c_2 + b}{2}$, где $[c_1,\ c_2]$ $(c_2,\ c_3]$ разбиение на интервалы по (1.1), $b>c_2$ – значение ближайшего к c_2 признака в описании объектов обучающей выборки. Никаких предположений о природе среды при этом не делается. Например, для линейного дискриминанта Фишера [34] таким предположением является распределение данных по нормальному закону.

О возможности использования разбиения по (1.5) для представления решающих правил в форме параллелограмма указывалось в § 1.3. Оценка компактности объекта по системе гипершаров с применением (1.1) описана в [7]. Компактность объекта $S \in E_0 \cap K_p$, p=1,2 рассматривалась как средство отбора собственного набора из разнотипных признаков для принятия решений.

Для иллюстрации комбинированного использования критериев (1.1) и (1.5) с целью разбиения признаков на интервалы рассмотрим тестовый пример. Пример приводится на рис.2.2.

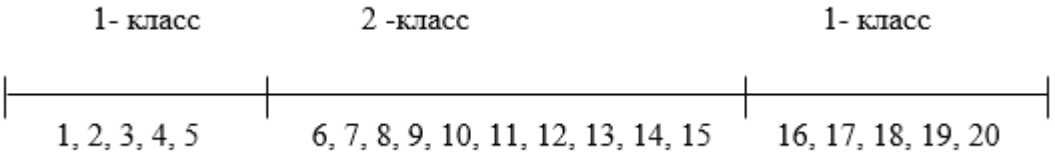


Рисунок 2.2. – Тестовый пример разбиения на непересекающиеся интервалы

Оптимальное значение критерия (1.1) равно 0.3611 при разбиении на интервалы [1,5], (5,20], что указывает на относительно плохую разделимость

(компактность) по признаку объектов классов. Устойчивость (1.5) по результатам разбиения на 3 интервала по критерию (1.3) равна 1 и демонстрирует наличие хорошо обусловленных кластеров по значениям признака.

Разбиение на непересекающиеся интервалы по критериям (2.2), (2.4) служат источником новых знаний при анализе данных из слабо структурированных предметных областей. Этим свойством критериев можно воспользоваться для поиска скрытых закономерностей в базах (хранилищах) данных.

Логические закономерности можно искать как между признаками, так и между объектами. В качестве одного из средств для реализации такого поиска были названы иерархические методы группировки [36, 40]. Реальных предложений по конкретным методам не было.

2.2.1. Скрытые закономерности на многообразии отношений между объектами

Пусть на наборе признаков $X(h) \subset X(n)$, $1 \le h \le n$ определена метрика $\rho(x, y)$. Объект $S \in E_0 \cap K_p$, $p=1,2$ рассматривается как центр гипершара, от которого по упорядоченному множеству объектов $\{S, S^1, \ldots, S^{m-1}\} = E_0$ формируется последовательность вложенных друг в друга гипершаров с радиусами

$$\rho(S,S) < \rho(S,S^1) \le \rho(S,S^2) \le \ldots \le \rho(S,S^{m-1}). \qquad (2.8)$$

Геометрическая интерпретация формирования упорядоченного множества $\{S, S^1, \ldots, S^{m-1}\}$ показана на рис. 2.3.

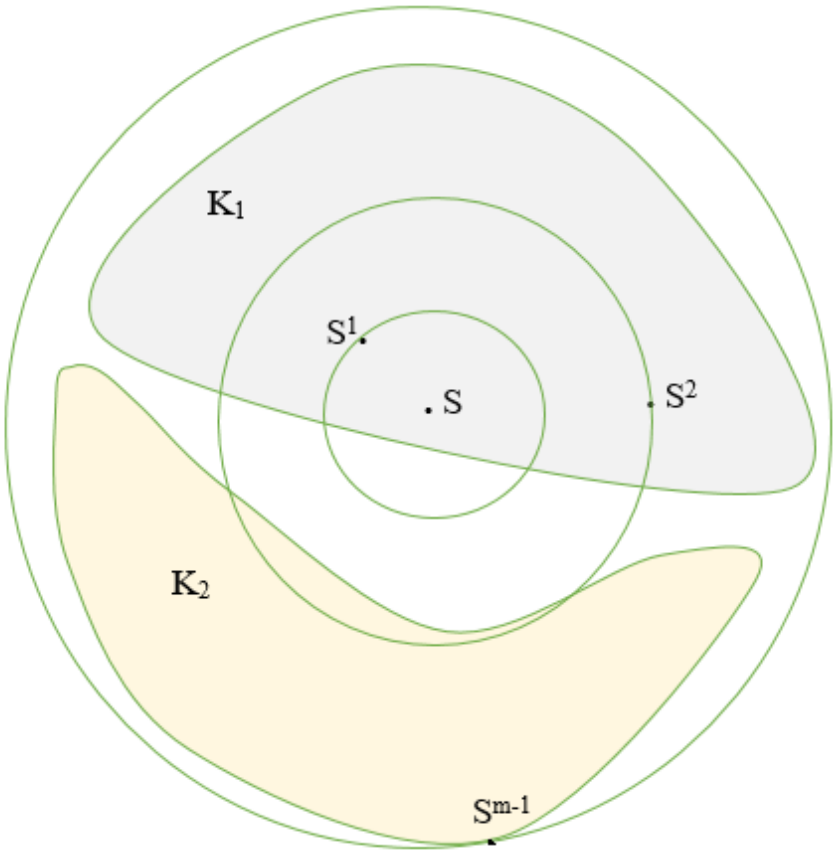


Рисунок 2.3. – Последовательность вложенных гипершаров

Значения границ интервалов каждого объекта $S \in E_0 \cap K_p$, p=1,2 на E_0, вычисляемые по критерию (1.1) на (2.8), используется как средство для отбора информативного набора разнотипных признаков из $X(n)$.

Существует разделение алгоритмов поиска оптимального набора признаков на три категории [10]. Это экспоненциальные, последовательные и рандомизированные алгоритмы. Для отбора информативного набора признаков в работе предлагается использовать последовательный алгоритм. На каждом шаге отбора к текущему набору добавляется один новый признак. Выбор начального состава набора рассматривается как отдельная проблема.

Пусть по (1.1) на (2.8) определены границы интервалов $[c_1, c_2]$, $(c_2, c_3]$, $c_1=\rho(S,S)=0$. Оценка компактности объекта $S \in K_p$ по набору признаков $X(h) \subset X(n)$ вычисляется как

$$\varphi(S,X(h))=\theta_1(1-\theta_2), \qquad (2.9)$$

где

$$\theta_1 = \frac{\left|\left\{S^i \in K_p \middle| \rho\left(S, S^i\right) \in [c_1, c_2]\right\}\right|}{\left|K_p\right|}, \quad \theta_1 = \frac{\left|\left\{S^i \in K_{3-p} \middle| \rho\left(S, S^i\right) \in [c_1, c_2]\right\}\right|}{\left|K_{3-p}\right|}.$$

Для сокращения комбинаторной сложности при поиске логических закономерностей в [36] было предложено использовать иерархические методы группировки. Отмечалось важность выбора первого шага для поиска закономерностей.

Специфика реализации методов иерархической агломеративной группировки зависит от используемым в них правил. Пусть $\varphi(X(h), S_i)$ – оценка компактности (2.9) объекта $S_i \in E_0$ на $X(h)$. При формировании набора $X(h+1)$ из $X(h)$ необходимо вычислить

$$R\left(X\left(h+1\right)\right) = \frac{1}{m}\sum_{i=1}^{m}\begin{cases}1, \varphi\left(X\left(h+1\right), S_i\right) \geq \varphi\left(X\left(h\right), S_i\right), \\ 0, \varphi\left(X\left(h+1\right), S_i\right) < \varphi\left(X\left(h\right), S_i\right).\end{cases} \tag{2.10}$$

Условием (правилом) для включения признака $x_j \in X(n) \backslash X(h)$ в $X(h+1)$ является:

$$R\left(X\left(h\right) \cup \left\{x_j\right\}\right) > \frac{1}{2} \quad и \quad R\left(X\left(h\right) \cup \left\{x_j\right\}\right) = \max_{x_i \in X(n) \backslash X(h)} R\left(X\left(h\right) \cup \left\{x_i\right\}\right). \tag{2.11}$$

Группа (набор) $X(h)$ считается сформированной, если не существует признака $x_j \in X(n) \backslash X(h)$, для которого выполняется (2.11).

Одной из особенностей методов Bigdata является анализ выборок данных, в которых число признаков больше либо равно числа объектов. Количество групп, на которое разбивается набор признаков $X(n)$ по (2.11) изначально неизвестно. Экспериментально доказано [78], что информативность каждой последующей группы признаков при использовании мажоритарных правил в иерархической агломеративной группировке меньше информативности предыдущей. Последовательность формирования групп определяется по принципу динамического программирования. По этой причине состав признаков из первой группы рассматривается в качестве информативного набора.

В качестве первого шага при отборе информативного набора признаков в диссертации предлагается выбирать подмножество $Y \subset X(n)$, состоящее из

одного или двух признаков. Подмножество *Y* должно удовлетворять следующему требованию:

$$B(Y)=\max_{\{i,j\}\in X(n)}\sum_{d=1}^{m}\left|\left\{S_j\in K_p\,\middle|\,\rho(S_j,S_d)<R,\qquad R=\min_{S_c\in K_{3-p}}\rho(S_c,S_d)\right\}\right|. \tag{2.12}$$

Геометрическая интерпретация выбора первого шага по (2.12) показана на рис. 2.4.

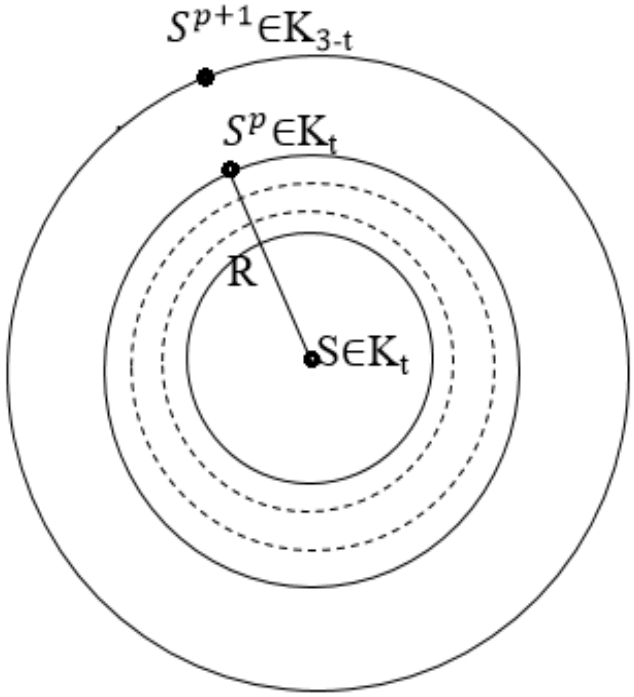


Рисунок 2.4. – Геометрическая интерпретация выбора первого шага

2.2.2. Вычислительный эксперимент

Для иллюстрации применения разработанных в диссертации алгоритмов рассмотрим задачу диагностики расходомеров. Расходомеры – это устройства, используемые для измерения объёмного или массового расхода жидкости. Существуют много причин, приводящих к ошибочным показаниям устройства, для исправления которых необходима повторная калибровка [4]. Тестовая база данных [16] служит основой для апробации методов решения задачи диагностики состояния расходомера, которая сводится к задаче классификации. Выборка данных состоит из 4-х классов: K1 (исправный), K2 (впрыск газа), K3 (дефекты

установки), K4 (восковая депиляция). В базе данных имеется 180 объектов, описываемых 43 признаками.

Необходимо:

- выбрать минимальный набор признаков, при котором точность классификации была бы выше, чем по набору исходных признаков;
- определить набор признаков, значения которых являются причиной неисправности конкретного устройства.

При поиске скрытых закономерностей в диссертации рассматриваются два варианта разбиения объектов на классы: исходный по 4 классам, описанным выше и на два K_1 – объекты с правильными показателями (исправные) и K_2 – объекты с 3 видами неисправностей. В качестве эвристики для отбора диагностических показателей по первому варианту используется упорядочение признаков по (1.1). Упорядоченный по (1.1) набор признаков приводится в табл. 2.2.

Таблица 2.2. Упорядоченный по (1.1) набор признаков

Название признака	**Значение веса**
Коэффициент усиления на начале 4 пути	0.6464
Коэффициент усиления на конце 4 пути	0.6464
Уровень сигнала на начале 4 пути	0.6268
Уровень сигнала на конце 4 пути	0.6268
Коэффициент усиления на начале 2 пути	0.6134
Коэффициент усиления на конце 2 пути	0.6134
Скорость потока на 4 пути	0.6108
Качество сигнала на конце 4 пути	0.5928
Качество сигнала на начале 4 пути	0.5878
Время прохождения в начале 4 пути	0.5706
Скорость звука на 4 канале	0.5598
Время прохождения в конце 4 пути	0.5587
Симметрия	0.5391
Коэффициент усиления на начале 3 пути	0.5388
Коэффициент усиления на конце 3 пути	0.5388
Качество сигнала на начале 3 пути	0.5329
Коэффициент усиления на начале 1 пути	0.5317

Название признака	**Значение веса**
Коэффициент усиления на конце 1 пути	0.5317
Качество сигнала на конце 2 пути)	0.5296
Качество сигнала на конце 3 пути)	0.5264
Коэффициент профиля	0.5223
Поперечный поток	0.5212
Уровень сигнала на конце 3 пути	0.5117
Уровень сигнала на начале 3 пути	0.5113
Качество сигнала на начале 1 пути	0.5025
Качество сигнала на конце 1 пути	0.4570
Качество сигнала на начале 2 пути	0.4331
Уровень сигнала на начале 1 пути	0.4271
Уровень сигнала на начале 2 пути	0.4249
Уровень сигнала на конце 2 пути	0.4225
Уровень сигнала на конце 1 пути	0.4114
Время прохождения на конце 1 пути	0.3549
Время прохождения на начале 1 пути	0.3515
Скорость звука на 1 канале	0.3464
Время прохождения на конце 3 пути	0.3406
Время прохождения на конце 2 пути	0.3369
Время прохождения на начале 2 пути	0.3335
Время прохождения на начале 3 пути	0.3277
Скорость звука на 3 канале	0.3205
Скорость звука на 2 канале	0.3195
Скорость потока на 1 пути	0.3125
Скорость потока на 2 пути	0.3067
Скорость потока на 3 пути	0.2994

Для признака «Коэффициент усиления на обоих концах каждого из четырех путей (начало путь 3)» (см. табл. 2.2) были получены такие границы 4 интервалов: [-1.0..-0.7] (-0.7..25.2167] (25.2167..45.0] (45.0..45.1].

К множеству количественных признаков применим дробно-линейное отображение их значений в интервал [0;1]. Такое отображение проводится с

целью сделать масштабы измерений признаков унифицированными. В качестве меры расстояния между объектами S_u, $S_v \in E_0$ ($S_c=(a_{c1},\ldots,a_{cn})$, $c=1,\ldots,m$) для отбора информативных признаков будем использовать метрику Журавлёва

Первой парой признаков на данных [16], включённой в информативный набор по (2.12), была (x_{29}, x_{33}). Процесс пошагового отбора признаков по (2.11) демонстрируется в табл. 2.3.

Таблица 2.3. Пошаговый отбор информативных признаков по (2.11)

Число признаков *h* в наборе	**Добавлен признак в *X*(*h*-1)**	**Значение R(*h*) по (2.10)**
3	x_{15}	0.8555
4	x_{30}	0.8555
5	x_{34}	0.9944
6	x_{16}	0.9833
7	x_{17}	0.8666
8	x_{18}	0.7944
9	x_{13}	0.6055
10	x_{14}	0.6777
11	x_{31}	0.5222
12	x_{27}	0.5833

Результатом пошагового отбора (см. табл. 2.3) является информативный набор признаков $X(12)=(x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{27}, x_{29}, x_{30}, x_{31}, x_{33}, x_{34})$.

Для демонстрации изменения обобщающей способности алгоритмов при отборе информативных наборов признаков по (2.8) используем критерий, отличный от (2.9)

$$\varphi(S, X(h)) = \max_{1\le i\le m-1}\left(\frac{d_t(i)}{|K_t|} - \frac{d_{3-t}(i)}{|K_{3-t}|}\right), \qquad (2.13)$$

где $d_t(i)$ ($d_{3-t}(i)$) – частота встречаемости объектов класса K_t (K_{3-t}) в подпоследовательности $S^1,\ldots,S^i$ из (2.8). Процесс отбора информативных признаков по (2.13) показан в табл. 2.4. Первая пара признаков в набор выбрана по (2.12).

Таблица 2.4. Пошаговый отбор информативных признаков по (2.13)

Число признаков *h* в наборе	**Добавлен признак в *X*(*h*-1)**	**Значение R(*h*) по (2.10)**
3	x_{30}	0.8777
4	x_{34}	0.9888
5	x_{15}	0.9777
6	x_{16}	0.8555
7	x_{17}	0.85
8	x_{18}	0.7833
9	x_{13}	0.7166
10	x_{31}	0.8888
11	x_{32}	0.8
12	x_{14}	0.7166
13	x_{27}	0.6611
14	x_{0}	0.5666
15	x_{1}	0.8
16	x_{19}	0.6944
17	x_{28}	0.7
18	x_{2}	0.7055
19	x_{37}	0.7111
20	x_{42}	0.6888
21	x_{25}	0.6555
22	x_{41}	0.6555
23	x_{20}	0.5777
24	x_{21}	0.5611
25	x_{10}	0.5277
26	x_{22}	0.5111
27	x_{35}	0.5055

Результатом пошагового отбора (см. табл. 2.4) является информативный набор признаков $X(27)=(x_0, x_1, x_2, x_{10}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{25}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{33}, x_{34}, x_{35}, x_{37}, x_{41,} x_{42})$.

В результате вычислительного эксперимента были получены два набора информативных признаков X(12) и X(27). Чтобы сравнивать наборы признаков, полученных по разным эвристикам, в диссертации предлагается использовать оценки обобщающей способности алгоритмов распознавания по правилу ближайший сосед.

2.3. Оценка обобщающей способности алгоритмов

Последовательные алгоритмы поиска информативных наборов признаков [10] обладают существенным недостатком: они не гарантируют, что при отборе достигнут глобальный экстремум функции качества. Как правило, в теории распознавания при оценке качества используется обобщающая способность алгоритмов. Распространённым методом для оценки обобщающей способности является кросс-валидация. Точность распознавания определяется на объектах, которых алгоритм «не видел» в процессе обучения. Для этого на выборке объектов производится последовательность разбиений её на две части, используемые с целью обучения и контроля. Недостатками метода кросс-валидации являются:

– вероятность переобучения классификатора;

– неполное использование выборки для обучения;

– реализация метода представляет неэффективную с точки зрения вычислений процедуру.

Предлагается метод оценки обобщающей способности алгоритмов, базирующийся на вычислении компактности объектов классов [8]. При реализации метода не требуется, как при кросс-валидации производить разделение выборки на «обучение» и «контроль».

Во введении к главе II описана идея анализа кластерной структуры для вычисления меры компактности непересекающихся классов и выборки в целом по отношению связанности объектов по заданной метрике. В [8] такое разбиение на группы рассматривается как предобработка данных для поиска минимального покрытия обучающей выборки объектами–эталонами. Сокращение

комбинаторной сложности алгоритма происходило за счёт поиска эталонов покрытия по каждой группе в отдельности. Было показано, что обобщающая способность алгоритмов зависит от удаления из выборки шумовых объектов. Меры компактности служат средством анализа изменений в структуре выборки при удалении шумовых объектов.

Множество шумовых объект [8] рассматривается как подмножество граничных объектов классов по заданной метрике. Множество граничных объектов $B \subset E_0$ определяется как

$$B = \left\{ S \in E_0 \left| \rho(S_i, S) = \min_{S_i \in K_j, S_d \in CK_j} \rho(S_i, S_d) \right. \right\}.$$

Объект $S \in B \cap K_j$, $j=1,\ldots,l$ принадлежит множеству шумовых объектов D_j класса K_j, если

$$\left| \left\{ S_i \in E_0 \left| \rho(S_i, S) = \min_{S_i \in CK_j, S_d \in K_j} \rho(S_i, S_d) \right. \right\} \right| > \left| \left\{ S_i \in K_j \left| \rho(S_i, S) < \min_{S_i \in K_j, S_d \in CK_j} \rho(S_i, S_d) \right. \right\} \right|. \quad (2.14)$$

Для проверки условия (2.14) нужно определить:

– число объектов из CK_j, для которых $S \in B \cap K_j$ является ближайшим по метрике $\rho(x, y)$;

– количество выполненных неравенств вида $\rho(S_i, S) < \min_{S_i \in K_j, S_d \in CK_j} \rho(S_i, S_d)$ для объектов $S_i \in K_j$.

Поиск шумовых объектов по (2.14) имеет смысл, если число $\left| K_j \cap B \right| > \frac{\max_{1 \le i \le l} |K_i|}{|K_j|}$, $j=1,..,l$. Шумовые объекты необходимы для проведения селекции обучающих выборок. Удаление шумовых объектов приводит к повышения обобщающей способности алгоритмов распознавания. Алгоритму даётся возможность делать ошибки на определяемых наборах объектов. В качестве таковых в нашем случае рассматриваются объекты из $\bigcup_{i=1}^{l} D_i$. Геометрическая интерпретация обнаружения шумового объекта показана на рис. 2.5.

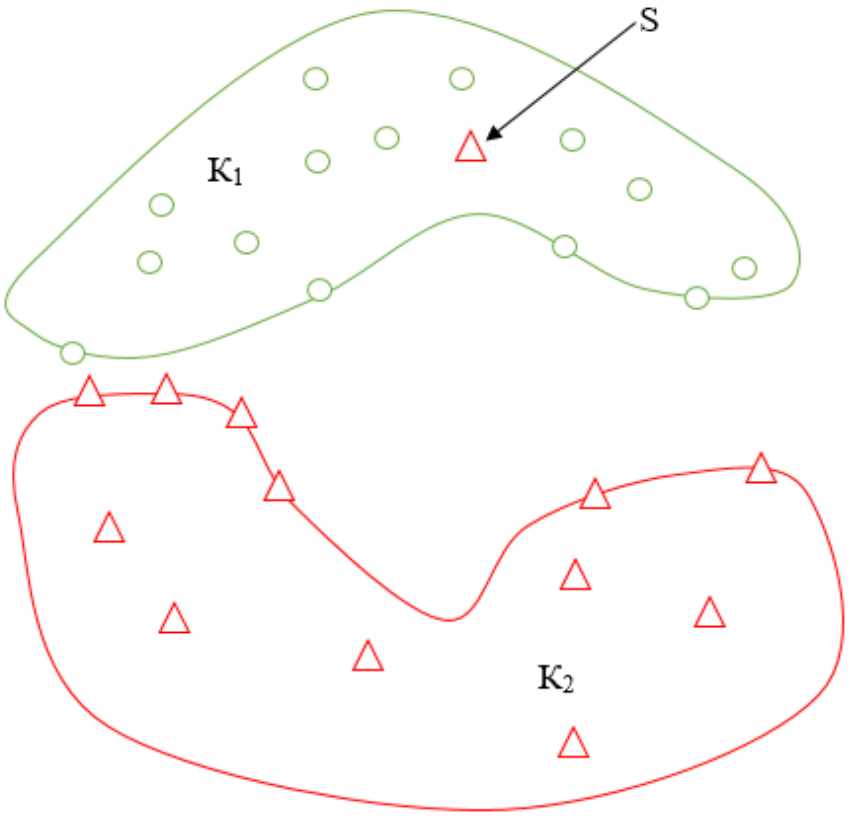


Рисунок 2.5. – Объект S является граничным для класса K_2 и шумовым для объектов класса K_1

Свойство (отношение) связанности объектов классов по системе гипершаров, в пересечении которых содержатся граничные объекты, позволяет получить разбиение выборки на минимальное число непересекающихся групп по алгоритму из [8]. По отношению связанности объектов выборка данных в признаковом пространстве разбивается на ряд областей. Эти области являются отправной точкой для поиска минимального покрытия выборки объектами–эталонами. Практически каждая область может содержать от одного и более эталонов. Процедуре поиска минимального покрытия предшествует удаление шумовых объектов по (2.14). Цель удаления – исключить выбор шумовых объектов в качестве эталонов.

Пусть представители класса $K_i \cap \left(E_0 \setminus \bigcup_{j=1}^{l} D_j\right)$, $i=1,\ldots,l$ разделены на минимальное число μ непересекающихся групп объектов $G_{i1},\ldots,G_{i\mu}$ по алгоритму из [8], $m_{ij} = \left|G_{ij}\right|, j = 1,\ldots,\mu, \sum_{j=1}^{\mu} m_{ij} = m_i$. Обозначим через $R_S=\rho(S,S^*)$ расстояние от объекта $S \in K_t$ до ближайшего объекта S^* из противоположного к K_t

класса ($S^* \in CK_t$), через δ – минимальное число непересекающихся групп объектов классов на $E_{ob} = E_0 \setminus \bigcup_{j=1}^{l} D_j$.

Порядок выполнения алгоритма поиска минимального покрытия объектами-эталонами выборки E_{ob} определяется следующим образом. Упорядочим объекты каждой группы $G_u \cap K_t, u = 1, \dots, \delta, t = 1, \dots, l$ по множеству значений $\{R_S\}_{S \in G_u}$. Для определения сходства между $S \in G_u, u = 1, \dots, \delta$ и произвольным допустимым объектом S' используется локальная метрика $d(S,S')=\rho(S,S')/R_S$. Решение о принадлежности S' к одному из классов $K_1,\dots,K_l$ или отказе от распознавания принимается по правилу: $S' \in K_t$ если

$$d(S_\mu, S') = \min_{S_j \in E_{ob}} d(S_j, S') \text{ и } S_\mu \in K_t \text{ и } d(S_\mu, S') \neq \min_{S_j \in CK_t \cap E_{ob}} d(S_j, S'). \quad (2.15)$$

Согласно принципу *последовательное исключение,* используемого в процессе поиска покрытия, выборка E_{ob} делится на два подмножества: множество эталонов E_{ed} и контрольное множество E_k, $E_{ob} = E_{ed} \cup E_k$. В начале процесса $E_{ed}=E_{ob}$, $E_k=\varnothing$. Чтобы выбрать объект $S \in G_u$ для удаления его из E_{ed} используется значение R_S. Для этого множество $\{R_S\}_{S \in G_u}$, $u=1,\dots,\delta$ предварительно упорядочивается.

В основе поиска минимального числа эталонов лежит идея проверки условия, при котором алгоритм распознавания по (2.15) остаётся корректным (без ошибок распознающим объекты) на E_{ob}.

Будем считать, что нумерация групп объектов [8] отражает порядок $|G_1| \geq \dots \geq |G_\delta|$ и по группе G_p, $p=1,\dots,\delta$ не производился отбор объектов-эталонов. Объекты группы G_p упорядочиваются по значениям меры расстояния от них до ближайшего объекта от них объекта из противоположного класса. Удаление объекта $S \in G_p$ из E_{ed} может нарушить корректность алгоритма распознавания по (2.14). В этом случае объект S остаётся в составе множества E_{ed}.

Показателем компактности выборки E_0 при использовании правила (2.15) является среднее число объектов E_{ob}, притягиваемых одним эталоном минимального покрытия из E_{ed}

$$\Omega(E_0, \rho) = \frac{|E_{ob}|}{|E_{ed}|}. \quad (2.16)$$

2.3.1. Вычислительный эксперимент

Рассмотрим количественную оценку компактности (2.16) (см. табл. 2.5) структуры отношений объектов классов K_1, K_2, K_3, K_4 по наборам признаков $X(43)$, $X(25)$, $X(17)$, полученных из табл. 2.2. Из наборов $X(25)$ и $X(18)$ удалены соответственно 18 и 26 признаков с относительно малыми значения весов (1.1). В скобках указано число объектов-эталонов из классов K_1, K_2, K_3, K_4.

Таблица 2.5. Значения компактности (2.16) по 3-м наборам признаков

Пространство	**Число объектов**		**Компактность по (2.16)**
	Шумовых	**Эталонов**	
Все	34	15(6+2+6+1)	7.8948
x_0, x_1, x_2, x_6, x_{10}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{33}, x_{34}, x_{41}, x_{42}	32	16(6+3+6+1)	7.6055
x_1, x_6, x_{10}, x_{17}, x_{18}, x_{23}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{33}, x_{34}, x_{41}, x_{42}	39	15(6+2+6+1)	7.3633

Процесс упорядочения признаков по весам (1.1) можно рассматривать как одну из эвристик для формирования информативных наборов из них для принятия обоснованных решений. Близость показателей компактности (2.16) на $X(43)$ и $X(25)$, равных соответственно 7.8948 и 7.6056, указывают на целесообразность снижения размерности исходного пространства.

Анализ компактности (2.16) объектов классов K_1(исправное состояние) и K_2 (наличие неисправностей) проводился на исходном наборе (43 признака), наборах, полученным по (2.9) (17 признаков) и по (2.13) (21 признак). Результаты анализа представлены в таблице 2.6.

Таблица 2.6. Значения компактности (2.16) по информативным наборам признаков

Наборы Признаков	**Число объектов**		**Компактность по (2.16)**
	Шумовых	**Эталонов**	
Все признаки	34	8(5+3)	14.8027
x_9, x_{10}, x_{13}, x_{14}, x_{15}, x_{16}, x_{25}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{39}, x_{40}, x_{41}, x_{42}	3	29(3+26)	6.0017
x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{39}, x_{40}, x_{41}, x_{42}	32	7(3+4)	17.3841

Результаты вычисления меры компактности (2.16) на наборе $X(21)= (x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{39}, x_{40}, x_{41}, x_{42})$ (см. табл. 2.6) были выше чем на наборе $X(43)$. Следствием из этого является целесообразность сокращения числа измеряемых показателей более чем в 2 раза для достижения лучшей обобщающей способности алгоритмов классификации по правилу (2.15). Максимальное значение компактность (2.16) на классах K_1 и K_2 из табл. 2.6 в 2 раза выше чем при классификации по 4 классам из табл. 2.5.

Рассмотрим случай попарного разделения 3 видов (классов) неисправностей. Интерес представляет выбор информативных наборов признаков для достижения относительно высокой обобщающей способности по решающему правилу (2.15) для каждого класса. Для значений (2.16) по каждому виду неисправности использовалось три набора признаков: исходный и информативные, при отборе которых использовались критерии (2.9) и (2.13). Результаты анализа демонстрируются в табл. 2.7.

Таблица 2.7. Результаты анализа компактности по (2.16) для 3 видов неисправностей

Вид неисправности	**Набор признаков**	**Значение по (2.16)**
Впрыск газа	Все признаки	20.5116
	$x_1, x_3, x_7, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{19}, x_{20}, x_{21}, x_{22}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{35}, x_{37}$	30.281
	$x_1, x_7, x_8, x_{11}, x_{12}, x_{13}, x_{14}, x_{19}, x_{20}, x_{24}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{35}, x_{37}, x_{38}, x_{39}$	30.281
Дефекты установки	Все признаки	62.5155
	$x_0, x_1, x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{33}, x_{34}, x_{35}, x_{36}, x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{42}$	41.0232
	$x_0, x_1, x_2, x_3, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x15, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{33}, x_{34}, x_{35}, x_{36}, x_{37}, x_{38}, x_{39}, x_{41}, x_{42}$	41.677
Восковая депиляция	Все признаки	62.5155
	$x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{23}, x_{24}, x_{25}, x_{26}, x_{33}, x_{34}, x_{35}, x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{42}$	63.5038
	$x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{23}, x_{24}, x_{25}, x_{26}, x_{33}, x_{34}, x_{35}, x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{42}$	63.5038

Самая высокая обобщающая способность (см. табл. 2.7) со значением (2.16), равном 63.5039, показана на неисправности *восковая депиляция* на наборе из 36 признаков.

Задача диагностики расходомеров рассматривалась в [4]. В [4] авторы используют новую технику контролируемого уменьшения линейной размерности алгоритм (LDR). Предлагаемый алгоритм проецирует исходные данные на *(K - 1)* - мерное подпространство, где *K* – количество классов. Как отмечают авторы [4] у алгоритма LDR имеются недостатки:

- алгоритм требует построения $(K^2 + K - 4) / 2$ классификаторов, что может быть довольно затратным с точки зрения вычислений для набора данных, содержащего слишком много классов;
- алгоритм основан на предположении о нормальности распределения данных в каждом из K-классов.

В [4] также утверждается, что снижение размерности с помощью метода главных компонент (PCA) и линейного дискриминанта Фишера (LDR) приводит к потере точности классификации при диагностике неисправностей расходомеров. Результаты обучения метрического алгоритма распознавания [8] доказывают эффективность снижения размерности пространства через меру компактности (2.16). Расхождение в результатах объясняется тем, что в LDR используется распознавание по правилам, в метрических алгоритмах – по прецедентам [27], что подтверждает важность выбора методов в математических моделях для управления процессом диагностики неисправностей технических устройств. Для диагностики можно использовать меньшее количество измеряемых показателей, гарантировать более высокую точность классификации, снижать материальные затраты на проведение калибровки устройств.

2.4. Выбор латентных признаков для обоснования процесса интуитивного принятия решения

Большой интерес представляет отыскание латентных, то есть скрытых признаков, которые могут быть информативными при классификации, что и составляет одну из задач настоящего исследования. Считается, что для разбиения значений количественного признака (как исходного, так и латентного) на непересекающиеся интервалы используются критерии (1.1) и (1.4). Латентные признаки могут представлять комбинации из номинальных и количественных признаков.

Пусть задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, содержащее представителей d непересекающихся классов $K_1,\ldots,K_d$. Описание объектов производится с помощью набора из n разнотипных признаков $X(n)$, δ ($\delta<n$) из которых

измеряются в номинальной, n–δ в интервальных шкалах. Допускается наличие пропусков и повторяющихся значений в данных. Считается, что на E_0 определена процедура формирования латентных признаков по множеству $\{(x_i, x_j)\}$, $(x_i, x_j) \subset X(n)$.

Требуется определить:

- границы интервалов и значения критерия (1.1) на исходных и латентных признаках;
- число интервалов, значения их границ и устойчивость разбиения исходных признаков по критерию (1.4).

Многообразие способов формирования латентных признаков и критериев для разбиения их значений на непересекающиеся интервалы необходимо для поиска скрытых закономерностей по базам данных предметных областей.

Латентные признаки из набора $X(n)$ будем формировать в виде комбинаций $x_i * x_j$ и x_i / x_j. Если число градаций номинального признака равно числу непересекающихся классов объектов, то им всегда можно поставить в соответствие набор целых чисел $a_1,\ldots,a_d$, где $a_i \neq 0$, $i=1,\ldots,d$ и $a_{j+1}-a_j = const$, где $j=1,\ldots,d-1$. Каждый непересекающийся интервал по (1.1) будет представлен одним значением. Например, при числе градаций равной 2, удобной для вычисления формой представления является выбор значений из {-1,1}.

Для вычисления весов номинальных признаков (как и для вычисления компактности количественных признаков по (1.1)) используется произведение внутриклассового сходства и межклассового различия.

Обозначим через g_{1c}^j, g_{2c}^j – количество значений градации $j \in \{1,\ldots,p_c\}$ признака $x_c \in X(n)$ в описании объектов соответственно класса K_1 и K_2. Межклассовое различие по признаку x_c определяется как величина:

$$\lambda_c = 1 - \frac{\sum_{j=1}^{p_c} g_{1c}^j g_{2c}^j}{|K_1||K_2|}. \qquad (2.17)$$

Степень однородности (мера внутриклассового сходства) β_c значений градаций признака по классам K_1, K_2 вычисляется по формуле:

$$D_{dc} = \begin{cases} (|K_d| - l_{dc} + 1)(|K_d| - l_{dc}), p_c > 2, \\ |K_d|(|K_d| - 1), p_c \leq 2; \end{cases},$$

где l_{dc} –

$$\beta_c = \begin{cases} \frac{\sum_{1c}^{p_c} g_{1c}^j \left(g_{1c}^j - 1\right) + g_{2c}^j \left(g_{2c}^j - 1\right)}{D_{1d} + D_{2d}}, D_{1d} + D_{2d} > 0, \\ 0, D_{1d} + D_{2d} = 0. \end{cases} \quad (2.18)$$

С помощью (2.17), (2.18) вес признака $x_c \in X(n)$ в номинальной шкале аналогично (2) определяется как произведение внутриклассового сходства и межклассового различия

$$\nu_c = \beta_c \lambda_c. \quad (2.19)$$

Множество допустимых значений весов признаков, вычисляемых по (2.19), принадлежит интервалу [0,1].

2.4.1. Вычислительный эксперимент

Для иллюстрации целесообразности поиска скрытых закономерностей по латентным признакам проведем вычислительный эксперимент. Рассмотрим результаты разбиения количественных признаков на непересекающиеся интервалы по критериям (1.1), (1.4) на выборке данных сердечно-сосудистых заболеваний из репозитория [15]. Номинальные признаки x_2, x_6, x_9 имеют две градации (т.е. число градаций признака равно числу классов). Компактность количественных признаков из $X(13)$ и границы интервалов по (1.1) приводятся в таблице 2.8.

Таблица 2.8. Границы интервалов и значения компактности по (1.1)

	Название признака	**Границы интервалов**	**Компактность**
x_1	Возраст	[29..54] (54..77]	0.2871
x_4	Покоящееся кровяное давление	[94..135] (135..200]	0.2548
x_5	Холесторальная сыворотка в мг./дл.	[126..252] (252..564]	0.2684
x_8	Достигнутый максимальный сердечный ритм	[71..147] (147..202]	0.3413
x_{10}	Oldpeak = депрессия ST, вызванная упражнениями относительно покоя	[0..1.6] (1.6..6.2]	0.3177
x_{11}	Наклон пикового упражнения ST сегмент	(1..2] (2..3]	0.3246
x_{12}	Количество основных сосудов (0-3), окрашенных флоусопой	[0..1] (1..3]	0.3772

В таблице 2.9 представлены значения всех шести номинальных признаков.

Таблица 2.9. Веса номинальных признаков

Признак	**Название признака**	**Вес**
x_2	Пол	0.2727
x_3	Тип боли в груди	0.3203
x_6	Уровень сахара в крови натощак> 120 мг./дл.	0.1873
x_7	Результаты электрокардиографии покоя	0.2762
x_9	Упражнение индуцированной стенокардии	0.3453
x_{13}	thal: 3 = нормальный; 6 =фиксированный дефект; 7 = обратимый дефект	0.4193

Как видно из табл. 2.8 и табл. 2.9, компактность количественных по (1.1) и значения весов номинальных признаков сильно отличаются от идеала. Значение количественного признака в границах непересекающегося интервала по (1.1) можно рассматривать как градацию (номер интервала) в номинальной шкале измерений. Вес признака при таком описании объектов в номинальной шкале будет совпадать со значением компактности по критерию (1.1).

Число непересекающихся интервалов и устойчивость разбиения по критерию (1.4) приводится в табл. 2.10.

Таблица 2.10. Устойчивость признаков и границы интервалов по критерию (1.4)

Признак	**Границы интервалов**	**Устойчивость**
x_1	[29, 54], [55, 70], [71, 76], [77, 77]	0.6571
x_4	[94, 122], [123, 200]	0.5585
x_5	[126, 160], [164, 174], [175, 245], [246, 353], [354, 394], [407, 409], [417, 564]	0.6309
x_8	[71, 147], [148, 194], [195, 195], [202, 202]	0.7030
x_{10}	[0, 0.8], [0.9, 6.2]	0.6957
x_{12}	[0, 0], [1, 3]	0.7316

Как видно из табл. 2.8 и табл. 2.10, относительно высокие значения компактности получены по признаку x_{12}.

Разбиение на два интервала латентных признаков, полученных по операциям умножения и деления значений исходных признаков, приводится в табл. 2.11.

Таблица 2.11. Границы интервалов для латентных признаков и значения компактности по (1.1)

Латентный признак	**Границы интервалов**	**Компактность**
x_4*x_9	[-192..105] (105..200]	0.3552
x_8*x_9	[-202..-115] (-115..186]	0.3718
x_{10}*x_{11}	[1..3.3] (3.3..21.6]	0.3684
x_2/x_8	[-0.0104..0.0067] (0.0067..0.0140]	0.3597
x_8/x_{10}	[16.8182..66.6667] (66.6667..202]	0.3726
x_8/x_{11}	[32..75] (75..202]	0.3555
x_9/x_4	[-0.0106..-0.0062] (-0.0062..0.0106]	0.3504
x_9/x_8	[-0.0140..0.0061] (0.0061..0.0113]	0.3523
x_{10}/x_8	[0.0049..0.0149] (0.0149..0.0594]	0.3726
x_{11}/x_8	[0.0049..0.0132] (0.0132..0.0312]	0.3555

Анализ результатов из таблицы 2.11 и таблицы 2.8 показывает целесообразность поиска скрытых закономерностей по латентным признакам,

компактность которых выше, чем каждого из исходных признаков, входящих в их состав. По результатам эксперимента самая высокая оценка компактности 0.4193 получена (см. табл. 2.9) по номинальному признаку x_{13}.

Выводы по главе 2

1. Предложена предобработка значений количественных (исходных и латентных) признаков в описании объектов для вычисления (1.1) с целью снижения сложности вычислений. Получена оценка сложности алгоритмов вычисления экстремума критерия (1.1) в описании объектов до и после предобработки. Основанием для использования предобработки служит снижение сложности вычисления алгоритмов после её проведения;

2. Описан численный алгоритм выбора границ непересекающихся интервалов по (1.1) на основе предобработки данных. На тестовом примере показано, что сложность вычисления по этому алгоритму на порядок ниже чем по алгоритму без использования предобработки;

3. Описано правило иерархической агломеративной группировки для пошагового объединения признаков в информативный набор с использованием мер компактности объекта (2.9) и (2.13). Правило для включения признака в информативный набор содержит последовательную проверку двух отношений: значение меры компактности по радиусам вложенных гипершаров после включения стало больше; у абсолютного большинства объектов выборки значение меры компактности увеличилось. Получены информативные наборы признаков при разбиении неисправностей ультразвукового расходомера на 3 класса и объединении их в один класс;

4. Доказана эффективность диагностики неисправностей ультразвукового расходомера по 21 признаку из 43. Значение меры компактности (2.16) по 21 признаку было равно 17.3841, на 43 признаках – 14.8027;

5. Разбиение на интервалы по критерию (1.1) позволяет вычислять компактность данных как по исходным, так и латентным признакам. Мера

компактности определяет структуру отношений между объектами классов. Для критерия (1.1) эти отношения рассматриваются на числовой прямой. Латентный признак определяет скрытую закономерность, если значение критерия по нему выше чем по любому исходному признаку, входящего в его состав.

ГЛАВА 3. Формирование описаний объектов выборок данных

В данной главе рассматриваются методы оценки структуры отношений объектов для выбора собственного признакового пространства объекта и селекции обучающих выборок [39, 53, 97, 100].

Вычисление собственного пространства каждого объекта выборки даёт возможность для сравнения объектов между собой по заданной метрике и определяемому набору признаков, например, позволяет определить набор признаков, значения которых являются причиной неисправности конкретного устройства.

С целью повышения обобщающей способности алгоритмов предлагаются способы селекции обучающихся выборок. Селекция проводится с использованием трёх методов:

– сокращение размерности признакового пространства;

– отбор шумовых объектов;

– минимальное покрытие обучающей выборки объектами–эталонами.

Для доказательства эффективности селекции предлагается использовать меру компактности (2.16) и оценку затрат ресурсов на принятие решения об отнесении произвольного допустимого объекта к одному из непересекающихся классов.

3.1. Компактность объектов классов по определяемым наборам признаков

В настоящее время существует несколько методов для вычисления меры компактности [71, 29]. Меры компактности используются для оценивания структуры отношений объектов в ограниченных областях признакового пространства. Размерность пространства является важным качественным и количественным параметром для выбора мер компактности. Например, в одномерном случае используют интервальные методы, в многомерном – среднее число объектов, притягиваемых эталоном минимального покрытия выборки.

В одномерном случае на числовой оси можно производить сравнение объектов по значениям их исходных и латентных признаков, используя отношения «больше», «меньше» или «равно».

При вычислении меры компактности в многомерном случае в [8, 71] применялось отношение связанности объектов по подмножеству (оболочке) граничных объектов непересекающихся классов. На основе этого отношения производилось разбиение объектов на непересекающиеся группы. Связанность объектов S_i, S_j рассматривалась, как свойство логических закономерностей в форме гипершаров, центрами которых они являлись. Объекты S_i и S_j считались связанными, если в пересечении их гипершаров содержались объекты оболочки. Любую пару объектов (S_i, S_j) из одной группы всегда можно соединить цепочкой из связанных объектов. В идеале все объекты класса представляют одну группу из связанных объектов. Графическая иллюстрация отношений связанности объектов показана на рис. 3.1.

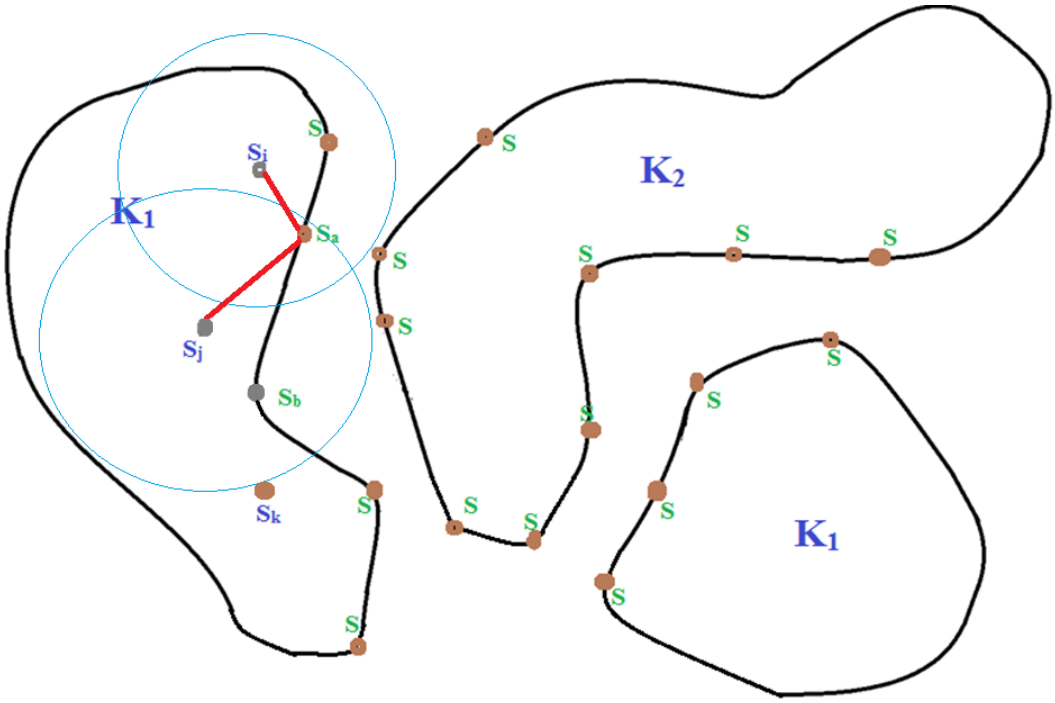


Рисунок 3.1. – Связанность объектов группы $S_i \leftrightarrow S_j$ по объекту оболочки класса S_a.

В диссертации исследуются структуры отношений между описаниями объектов классов на числовой оси. В качестве инструмента для исследования предлагаются меры компактности, вычисляемые по результатам разбиения значений признаков (исходных и латентных) на непересекающиеся интервалы. Значения мер компактности используются для поиска скрытых закономерностей в

данных. Скрытые закономерности [77, 96, 101] могут рассматриваться как новое знание, полученное в рамках информационных моделей для слабо структурированных предметных областей.

Реализация практически всех известных алгоритмов машинного обучения резко ухудшается при большой размерности данных [33]. В [3] была дана геометрическая интерпретация возникновения эффекта проклятия размерности. Объясняется это тем, что количество возможных конфигураций (наборов) множества признаков возрастает экспоненциально с увеличением числа признаков [30].

Проклятие размерности характерно для разных разделов информатики, но особенно часто проявляется в машинном обучении. Проблема проклятия размерности связана с соотношением числа признаков к числу объектов на обучении. Комбинаций наборов признаков на $X(n)$ может быть значительно больше мощности обучающих примеров [75-77].

В большинстве традиционных алгоритмов машинного обучения просто предполагается, что классификация для нового объекта должна быть примерно такой же, как для ближайшего объекта обучения. При достаточном числе примеров из обучающей выборки от алгоритма обучения легко добиться правильного обобщения.

Размытость отношений [50] отдельного объекта выборки зависит от используемого набора признаков и заданной метрике. Исследовать размытость можно по усреднённым величинам компактности объектов в классе и выборке в целом с применением частотного анализа.

Покажем, что для вычисления компактности объектов в выборке можно использовать критерии отличные от (2.9). Пусть определены последовательность радиусов гипершаров (2.8) с центром в $S_d \in K_p$ и $c_p(i)$, $c_{3\text{-}p}(i)$, $i \le m$ – число объектов из K_p, $K_{3\text{-}p}$ в гипершаре с радиусом $\rho(S_d, S^i)$. Для вычисления компактности объекта $S_d \in K_p$ предлагаются критерии

$$z_d(X(r)) = \max_{1 \le i \le m} \left(\frac{c_p(i)}{|K_p|} - \frac{c_{3-p}(i)}{|K_{3-p}|} \right), \qquad (3.1)$$

$$f_d(X(r)) = \frac{1}{\mu}\max_{1\le i\le \mu}\left(c_p(i) - c_{3-p}(i)\right), \qquad (3.2)$$

где $\mu=|K_p|$.

Очевидно, что значение, вычисленное по (3.1), (3.2) ограничены сверху 1. Значения критериев, равные 1, означает, что все объекты лежат в гипершаре с центром в определяемом объекте. Отметим, что мегапонятия, которые могут быть сформированы иерархическим агломеративным алгоритмом по правилу (2.11) и критериям (3.1), (3.2), в общем случае могут отличаться от мегапонятий по критерию (2.9).

Интервальные методы могут служить инструментом для вычисления собственного пространства каждого объекта выборки. Это даёт возможность для сравнения объектов между собой по заданной метрике $\rho(x,y)$ и определяемому набору признаков $X(k)\subset X(n)$, $k\le n$. Идея использования компактности (устойчивости) объекта $S_d\in K_p$ описана в §2.3.

Приведём описание рекурсивного алгоритма для формирования собственного пространства объекта. Пусть для объекта $S_d\in K_p$, p=1,2 расстояния до объектов выборки E_0 представлены в виде упорядоченной последовательности (2.8). Обозначим через P – множество индексов признаков из $X(n)$. Реализация алгоритма по шагам будет следующей.

Шаг 1. $P=\{1,\ldots,n\}$. $Y=\{x_i\}_{i\in P}$. Вычислить crit=$\varphi(S_d, Y)$ по (2.9). $Z=Y$;

Шаг 2. **Цикл** по $i,j\in P$. $b_{ij}=\varphi(S_d, Y\backslash\{x_i,x_j\})$. Конец **цикла**;

Шаг 3. A=0. **Цикл** по $i,j\in P$. Если $A< b_{ij}$ то $i1=i$, $j1=j$, $A=b_{ij}$. Конец **цикла**;

Шаг 4. $Y=Y\backslash\{x_{i1},x_{j1}\}$. $P=P\backslash\{i1,j1\}$. Если $A>$crit, то crit=A, $Z=Y$;

Шаг 5. Если $|P|>2$, то идти 2;

Шаг 6. Вывод Z;

Шаг 7. Конец.

3.1.1. Вычислительный эксперимент

Рассмотрим снова данные диагностики неисправностей 8-лучевого жидкостного ультразвукового расходомера (USM) [16] для определения информативных наборов признаков полученный по рекурсивному алгоритму на данных, показанному в табл. 3.1. Выборка содержит 87 экземпляров (объектов) описываемых 37 признаками и разбита на 2 класса по состоянию работоспособности: K_1 – исправное, K_2 – дефекты установки. Приводятся результаты по объектам из K_2. В качестве меры расстояния между описаниями объектов используется метрика Журавлёва. Значения количественных признаков пронормированы в [0;1].

Таблица 3.1. Отбор информативных наборов признаков по рекурсивному алгоритму

Номер объекта/объектов	**Набор признаков**	**Значение критерия (2.9)**
35	$x_0, x_{11}, x_{12}, x_{21}$	0.6445
36	$x_0, x_3, x_4, x_5, x_6, x_7, x_8, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{21}, x_{30}, x_{31}, x_{32}, x_{35}$	0.3857
37	$x_0, x_{11}, x_{12}, x_{13}, x_{14}, x_{21}$	0.7489
38	x_{11}, x_{21}	0.6527
39	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{21}$	0.717
40	$x_0, x_{11}, x_{13}, x_{14}, x_{16}, x_{17}, x_{24}, x_{30}$	0.7912
41	$x_0, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{30}$	0.7582
42	$x_0, x_{11}, x_{12}, x_{13}, x_{21}$	0.7494
43	$x_0, x_3, x_4, x_5, x_6, x_8, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{21}$	0.5802
44	$x_2, x_3, x_4, x_6, x_7, x_{10}, x_{11}, x_{12}, x_{13}, x_{20}, x_{21}, x_{24}, x_{32}$	0.4395
45	x_0	0.6164
46	$x_2, x_3, x_6, x_7, x_{10}, x_{11}, x_{24}, x_{33}$	0.3956
47	x_0	0.5983
48	x_0	0.6538
49	x_{10}	0.3758
50	$x_{11}, x_{12}, x_{13}, x_{15}, x_{16}, x_{17}, x_{18}, x_{27}$	0.5714
51	$x_{11}, x_{12}, x_{13}, x_{16}, x_{17}, x_{22}$	0.6065

Номер объекта/объектов	**Набор признаков**	**Значение критерия (2.9)**
52	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{22}$	0.6714
53	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{21}, x_{22}$	0.6456
54	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{22}$	0.6714
55	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{21}, x_{22}$	0.6824
56	$x_{11}, x_{12}, x_{13}, x_{15}, x_{16}, x_{22}$	0.6769
57	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{22}$	0.6461
58	$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{22}, x_{29}$	0.6307
59	x_{12}	0.5813
60	$x_2, x_3, x_4, x_6, x_{11}, x_{12}, x_{13}, x_{14}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{23}, x_{25}, x_{31}, x_{32}, x_{33}$	0.4395
61	$x_2, x_3, x_6, x_7, x_{10}, x_{12}, x_{13}, x_{32}, x_{33}$	0.3758
62-64	$x_2, x_3, x_4, x_6, x_{10}, x_{11}, x_{32}, x_{33}$	0.3956
65	$x_2, x_3, x_6, x_7, x_{10}, x_{12}, x_{32}, x_{33}$	0.3956
66	x_{10}	0.3758
67	$x_2, x_3, x_4, x_6, x_7, x_{10}, x_{23}, x_{34}$	0.3857
68	$x_0, x_1, x_2, x_3, x_6, x_7, x_{11}, x_{12}, x_{13}, x_{15}, x_{16}, x_{17}, x_{18}, x_{23}, x_{25}, x_{26}, x_{27}, x_{29}, x_{30}, x_{31}$	0.3824
69	$x_0, x_{11}, x_{12}, x_{13}, x_{16}$	0.6043
70	$x_0, x_{11}, x_{12}, x_{13}, x_{16}$	0.6043
71	$x_0, x_{11}, x_{12}, x_{14}, x_{15}$	0.6318
72	$x_0, x_{11}, x_{12}, x_{14}, x_{16}, x_{24}$	0.7664
73	$x_0, x_3, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{21}$	0.6373
74	x_0	0.6538
75	$x_0, x_{12}, x_{14}, x_{16}, x_{21}$	0.7664
76	$x_0, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{30}$	0.8005
77	$x_0, x_{11}, x_{12}, x_{13}, x_{14}, x_{24}$	0.8087
78	$x_0, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{24}, x_{25}$	0.8087
79	$x_0, x_{12}, x_{13}, x_{14}, x_{24}$	0.8159
80	$x_0, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{20}, x_{21}, x_{22}, x_{24}$	0.7208
81	$x_0, x_1, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{21}, x_{22}, x_{24}, x_{33}$	0.6857
82-86	x_0	0.6538

Описанный выше рекурсивный алгоритм реализует пошаговое удаление одного или двух признаков из $X(n)$ для получения информативного набора объекта. Так же, как и в [36] утверждается, что процесс формирования собственного пространства путём пошагового добавления признаков зависит от выбора первого шага. При отборе информативного набора признаков объекта $S_d \in K_p$ предлагается выбирать подмножество $Y \subset X(n)$, состоящее из одного или двух признаков. Подмножество Y должно удовлетворять следующим требованиям:

$\forall\, S_i \in K_{3\text{-}p}$ $\rho(S_j,S_d)>0$ на $Y \subset X(n)$;

$$\left|\left\{ S_j \in K_p \middle| \rho(S_j, S_d) < R, \qquad R = \min_{S_c \in K_{3-p}} \rho(S_c, S_d) \right\}\right| = \max_{E_0} . \qquad (3.3)$$

Первое требование связано с отсутствием как минимум одного объекта из $K_{3\text{-}p}$, описание которого совпадает с $S_d \in K_p$, p=1,2, второе с максимальным количеством объектов одного с S_d класса в гипершаре с радиусом R и центром в S_d. Результаты отбора информативных признаков с учётом (3.3) показаны в табл. 3.2.

Таблица 3.2. Отбор информативных наборов признаков с учётом (3.3)

Номер объекта/объектов	**Набор признаков**	**Значение критерия (2.9)**
35	x_0, x_{12}, x_{19}, x_{21}	0.6445
36	x_0, x_{12}, x_{14}, x_{21}	0.6593
37	x_0, x_{11}, x_{12}, x_{13}, x_{19}, x_{21}	0.7489
38	x_{19}, x_{21}	0.6527
39	x_0, x_{12}, x_{19}, x_{21}	0.7324
40	x_{19}, x_{24}	0.7494
41	x_{12}, x_{14}, x_{21}	0.7010
42	x_0, x_{11}, x_{19}, x_{21}, x_{24}	0.7324
43	x_{13}, x_{14}, x_{21}, x_{24}	0.6813
44	x_{11}, x_{14}, x_{21}, x_{24}	0.6983
45	x_0, x_{30}	0.6214
46-47	x_0, x_{30}	0.6054

Номер объекта/объектов	**Набор признаков**	**Значение критерия (2.9)**
48	x_0, x_{30}	0.6681
49	x_0, x_{21}	0.5983
50	x_{14}, x_{17}	0.5571
51	x_{19}	0.5571
52	$x_{11}, x_{12}, x_{13}, x_{14}, x_{19}, x_{22}$	0.6714
53	$x_{11}, x_{13}, x_{14}, x_{21}$	0.6428
54	x_{14}	0.5813
55	x_{13}, x_{14}	0.6214
56	x_{14}	0.6181
57-59	x_{19}	0.5813
60	x_{14}, x_{18}	0.5813
61	x_{17}, x_{18}	0.5307
62	x_{14}, x_{19}	0.4675
63-65	$x_2, x_6, x_{18}, x_{23}, x_{26}, x_{32}, x_{33}$	0.3956
66	$x_2, x_{20}, x_{23}, x_{32}, x_{33}$	0.3923
67	$x_2, x_{31}, x_{32}, x_{33}$	0.3829
68	x_{12}, x_{14}, x_{24}	0.6329
69	x_{14}, x_{24}, x_{30}	0.6428
70	x_{14}, x_{21}, x_{24}	0.6571
71	x_{12}, x_{14}, x_{24}	0.6857
72	$x_0, x_{12}, x_{14}, x_{21}, x_{24}$	0.7747
73	$x_{14}, x_{21}, x_{24}, x_{33}$	0.7307
74	x_0, x_{27}	0.6538
75	x_{14}, x_{21}, x_{24}	0.8005
76	x_{14}, x_{24}	0.7664
77	$x_{12}, x_{14}, x_{24}, x_{30}, x_{33}$	0.7615
78	$x_{12}, x_{14}, x_{24}, x_{34}$	0.7494
79	x_{14}, x_{24}	0.7835
80	x_{14}, x_{24}	0.7736
81	$x_{13}, x_{14}, x_{21}, x_{24}$	0.7384
82	$x_{11}, x_{12}, x_{16}, x_{21}, x_{24}$	0.7032
83-86	x_0, x_{22}	0.6538

Особенности вычисления логических закономерностей в форме гипершаров допускают равные значения (2.9) (учитываются только количество объектов *своего* класса в гипершарах). В табл. 3.1 это объекты № 72 и № 75, в табл. 3.2 – объекты № 54 и № 57.

В (2.9) реализуется мультипликативная форма вычисления значения критерия, в (3.1) – аддитивная (разность частот встречаемости объектов классов). Аналогичные табл. 3.2 результаты по критерию (3.1) приводится в табл. 3.3.

Таблица 3.3. Отбор информативных наборов признаков по (3.1)

Номер объекта	**Набор признаков**	**Значение критерия (3.9)**
35	x_{19}, x_{21}	0.5615
36	x_{14}, x_{21}	0.5615
37	$x_0, x_{11}, x_{12}, x_{13}, x_{19}, x_{21}$	0.7324
38	x_{19}, x_{21}	0.6175
39	$x_0, x_{12}, x_{19}, x_{21}$	0.7126
40	x_{19}, x_{24}	0.7318
41	x_{12}, x_{14}, x_{21}	0.6747
42	$x_0, x_{11}, x_{19}, x_{21}, x_{24}$	0.7126
43	$x_0, x_{12}, x_{14}, x_{15}, x_{21}$	0.6840
44	$x_{11}, x_{14}, x_{21}, x_{24}$	0.6741
45	x_0, x_{30}	0.5785
46	x_0, x_{30}	0.5593
47	x_0, x_{30}	0.5593
48	x_0, x_{30}	0.6450
49	x_0, x_{21}	0.5868
50	x_{17}	0.4928
51	x_{19}	0.4928
52	$x_{11}, x_{12}, x_{13}, x_{14}, x_{19}, x_{22}$	0.6467
53	x_{14}	0.5406
54	x_{14}	0.5417
55	x_{13}, x_{14}	0.5785
56	x_{14}	0.5796
57-59	x_{19}	0.5417

Номер объекта	Набор признаков	Значение критерия (3.9)
60	x_{14}, x_{18}	0.5417
61	x_{17}, x_{18}	0.4648
62	x_{14}, x_{19}	0.3686
63-66	x_{2}, x_{32}	0.2280
67	x_{2}, x_{20}, x_{23}, x_{32}, x_{33}	0.2934
68-69	x_{12}, x_{14}, x_{24}	0.6087
70	x_{14}, x_{21}, x_{24}	0.6274
71	x_{12}, x_{14}, x_{24}	0.6659
72	x_{0}, x_{12}, x_{14}, x_{21}, x_{24}	0.7609
73	x_{14}, x_{21}, x_{24}, x_{33}	0.7307
74	x_{0}, x_{27}	0.6538
75	x_{14}, x_{21}, x_{24}	0.7895
76	x_{14}, x_{24}	0.7510
77	x_{14}, x_{24}	0.7038
78	x_{14}, x_{24}	0.7230
79	x_{14}, x_{24}	0.7703
80	x_{14}, x_{24}	0.7604
81	x_{13}, x_{14}, x_{21}, x_{24}	0.7219
82	x_{11}, x_{12}, x_{16}, x_{21}, x_{24}, x_{32}	0.6923
83-86	x_{0}, x_{22}	0.6538

В табл. 3.1, 3.2, 3.3 представлены отборы информативных признаков по двум пошаговым методам (удаление малоинформативного признака, добавление информативного признака). Интерес для исследования представляет анализ частоты встречаемости отдельных признаков в составе собственного пространства объектов из класса K_2 с показателями дефектов установки при диагностике неисправностей жидкостного ультразвукового расходомера. В табл. 3.4 представлены 10 значимых показателей при диагностике неисправностей по критериям, используемым в табл. 3.1, 3.2, 3.3.

Таблица 3.4. Перечень показателей при диагностике неисправностей

№ п/п	Название признака	Частота встречаемости в таблицах		
		3.3	3.4	3.5
1.	x_{12} -скорость звука на пути 2	34	12	9
2.	x_{11} - скорость звука на пути 1	33	6	5
3.	x_{13} - скорость звука на пути 3	29	6	4
4.	x_0 -коэффициент плоскостности	25	12	12
5.	x_{14} - скорость звука на пути 4	23	25	23
6.	x_{21} - усиление на пути 1 конец 2	14	17	16
7.	x_{17} - скорость звука на пути 7	13	2	2
8.	x_{24} - усиление на пути 3 конец 1	9	18	16
9.	x_{32} - усиление на пути 7 конец 1	6	3	3
10.	x_{30} - усиление на пути 6 конец 1	5	5	4

Наиболее часто встречаемыми показателями неисправностей в работе ультразвукового расходомера (см. табл. 3.4) являются значения скоростей звука на путях. Выбор собственного пространства объекта позволяет определить набор признаков, значения которых являются причиной неисправности конкретного устройства.

3.2. Селекция обучающих выборок через отбор информативных разнотипных признаков и минимальное покрытие объектами-эталонами

Под оценкой (мерой) сложности $R(X(k))$, $k \le n$ алгоритма принятия решения по допустимому объекту S будем понимать количество элементарных операций для его распознавания по набору признаков ($X(k)$). Рассматривается следующая задача построения упорядоченной по неубыванию сложности алгоритмов последовательности признаков.

Пусть задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, содержащее представителей l непересекающихся классов (подмножеств E_0) $K_1,\ldots,K_l$. Описание объектов произведено с помощью набора из n разнотипных признаков $X(n)=(x_1,\ldots,x_n)$, δ из которых измеряются в номинальной шкале, n - δ в

количественных шкалах, допускается отсутствие измеренных значений в данных. Требуется найти последовательность информативных наборов $X(n)$, $X(n\text{-}1),\ldots$, $X(k)$, $k\le n$ на которых значения меры сложности алгоритмов $R(X(n)),R(X(n\text{-}1)),\ldots,R(X(k))$ образуют невозрастающую последовательность $R(X(n))\ge R(X(n\text{-}1))\ge\ldots\ge R(X(k))$.

Для решения этой задачи обозначим через I, J множество номеров соответственно количественных и номинальных признаков в описании допустимых объектов, $|I|+|J|=n$.

Процесс селекции обучающих выборок соответствует идеям, описанным в [39, 70, 73]. Соответствие выражается в решении задачи о минимальном покрытии обучающей выборки E_0 объектами-эталонами множества $\Pi_j=\{S^1,\ldots,S^\alpha\}$, $\alpha\le m$, $\Pi_j\subset E_0$, j=1,2,…. Единственность множества $\Pi_j=\{S^1,\ldots,S^\alpha\}$ как мощности так и по составу гарантируется методом из [8].

Для введения меры близости на множестве разнотипных признаков и унификации шкал измерений предлагается предобработка данных. Известно, что от сильных (интервальных) шкал измерений всегда можно перейти к слабым шкалам.

Упорядоченное множество значений признака x_j, $j\in I$ разбивается на последовательность непересекающихся интервалов $(c_{2k\text{-}1}, c_{2k}]$, $c_{2k\text{-}1}<c_{2k}$, $k=1,\ldots l$. Значению признака в границах интервала ставится в соответствие градация номинального признака. Описание критерия (1.1) для определения значений границ интервалов $(c_{2k\text{-}1}, c_{2k}]$ имеется в главе I данной работы. Идеальным считается разбиение, при котором в границах интервала размещаются все значения признака объектов из одного класса.

Пусть u_j^p - мощность множества значений признака x_j, $j\in I$ у объектов класса K_i в границах интервала $(c_{2p\text{-}1}, c_{2p}]$, $A=(a_0,\ldots,a_l)$, $a_0=0$, $a_l=m$, a_p –номер по порядку элемента упорядоченной последовательности $r_{j1},\ldots,r_{jm}$ значений признака x_j у объектов из E_0, идентифицирующий правую границу интервала $c_{2p}=r_{a_p}$.

Критерий (1.1) позволяет вычислять оптимальные значения границ интервалов $\{(c_{2p-1}, c_{2p}]\}$ и использовать их для определения градаций количественного признака в номинальной шкале измерений. Процесс преобразования при этом оказывается неразрывным от классификации, вводимой на множестве объектов обучения, и может быть реализован с учётом пропусков в данных.

При наличии объектов, содержащих пропуски (не измеренные значения) в данных для анализа разнотипных признаков предлагается произвести предобработку по критерию (1.1). Результаты предобработки используются для вычисления значения вклада каждого признака x_p, $p \in I \cup J$ в принятие решения о разделении объектов классов как

$$\lambda_p = \frac{\sum_{i=1}^{l}\sum_{j=1}^{u_p} z_{pj}^{i}\left(z_{pj}^{i}-1\right)}{\sum_{i=1}^{l}\tau_i} - \frac{\sum_{i=1}^{l}\sum_{j=1}^{u_p} z_{pj}^{i}\overline{z_{pj}^{i}}}{\sum_{i=1}^{l} b_{ip}\overline{b_{ip}}}, \qquad (3.4)$$

где $z_{pj}^{i}, \overline{z_{pj}^{i}}$ - количество значений j –й градаций p –го признака соответственно класса K_i и его дополнения $CK_i = E_0 \backslash K_i$, u_p– число градаций p –го признака, l_{ip}– число градаций p –го признака в классе K_i, $b_{ip}, \overline{b_{ip}}$ -число значений p –го признака без пропусков соответственно в K_i и CK_i

$$\tau_i = \begin{cases} (b_{ip}-l_{ip}+1)(b_{ip}-l_{ip}), & \text{где } l_{ip} > l, \\ b_{ip}(b_{ip}-1), & \text{где } l_{ip} \le l. \end{cases}$$

Множество значений $\{\lambda_p\}$, вычисленных по (3.4), может быть использовано для предварительного удаления неинформативных признаков для случая с большой размерностью (от 500 и более) пространства в описании объектов. Кандидатами на удаление являются признаки с относительно малыми значениями λ_p.

3.2.1. Отбор информативных признаков с максимально выраженной независимостью

Отбор информативных наборов признаков является важным и наименее формализованным разделом дискриминантного анализа. Эффективность численной реализации критериев отбора во многом зависит от использования определенных (как правило, скрытых) закономерностей, наличие которых позволяет уменьшить комбинаторную сложность алгоритмов отбора. К сожалению, рамки диссертационного исследования не позволяют сделать исчерпывающий обзор по данной проблеме.

В настоящем исследовании для сведения к минимуму числа переборов использовалось упорядочение признаков по отношению сложности (в смысле, определённом во введении главы III) алгоритмов распознавания. Теоретически (в идеале) минимальный набор должны представлять независимые признаки. В практической реализации выбираются исходные (не искусственные) признаки, коррелированность между собой которых минимальна, т.е. признаки с максимально выраженной независимостью.

На множестве пар объектов (S_a=(x_{a1},…,x_{an}), S_b=(x_{b1},…,x_{bn})), определим функции:

$$g(a,b,i,j)=\begin{cases} 2, x_{ai}\neq x_{bi} \text{ и } x_{aj}\neq x_{bj} \\ 1, x_{ai}=x_{bi} \text{ или } x_{aj}=x_{bj} \\ 0, x_{ai}=x_{bi} \text{ и } x_{aj}=x_{bj}; \end{cases}$$

$$\alpha(a,b)=\begin{cases} 0, S_a, S_b \in K_i, i=\overline{1,l}, \\ 1, S_a \in K_i, S_b \in K_j, i\neq j. \end{cases}$$

Меру близости между парой номинальных признаков x_i, x_j на E_0 зададим как

$$b_{ij}=\begin{cases} \dfrac{\sum_{a=1}^{m}\sum_{b=1}^{m}\alpha(a,b)g(a,b,i,j)}{2\sum_{p=1}^{l}\left|K_p\right|\left(m-\left|K_p\right|\right)}, i\neq j \\ 0, i=j. \end{cases} \qquad (3.5)$$

Считается, что селекция обучающей выборки производится по трём методам так как это описано в начале главы III. По результатам селекции получено

минимальном покрытие множества E_0 набором объектов–эталонов $\Pi_j=\{S^1,\ldots,S^\alpha\}$, $\alpha\le m$, $\Pi_j\subset E_0$, $j=1,2,\ldots$.

Критерием для отбора информативного набора признаков $X(k)=(x_1,\ldots,x_k)$, $k\le n$ служит сложность алгоритма [53]

$$\min_{\{\Pi_j\}}\left|\Pi_j\right|k \to \min_{E_0},$$

где k – число признаков в наборе, Π_j –множество объектов минимального покрытия, получаемого на $X(k)$ и обеспечивающее корректное (без ошибок) распознавание на E_0 без шумовых объектов (2.14) с помощью (2.15). Ставится задача построения упорядоченной последовательности признаков с целью реализации направленного отбора информативных признаков. В форме вычислительного эксперимента доказывается утверждение, что удаление (в заданном порядке) признака из последовательности должно приводить к монотонному невозрастанию сложности

$$R(X_k) = \left|\Pi_j\right|k \tag{3.6}$$

алгоритмов распознавания.

Далее будем считать, что матрица $B=\{\ b_{ij}\ \}_{m\cdot n}$ построена по (3.5) и значения вкладов $\{\lambda_i\}_1^n$ вычислены по (3.4) на множестве разнотипных признаков $X(n)=(x_1,\ldots,x_n)$. Для определения информативного набора $X(k)=(x_1,\ldots,x_k)$, $k<n$ используются рекурсивные вычисления. Разработана процедура [53] для формирования упорядоченного по отношению информативности набора признаков

$$x_{i_1}, x_{i_2}, \ldots, x_{i_n}. \tag{3.7}$$

Пусть по матрице B определена пара (x_i, x_j), $(\lambda_i \ge \lambda_j)$ с максимальным значением b_{ij}. Эта пара помещается в начало (слева направо) набора (3.7). Строки и столбцы с номерами i и j. Рекурсивность вычислений выражается в том, что следующая пара признаков аналогичным образом определяется из усечённой матрицы B.

Эффективность поиска информативного набора обеспечивается за счёт ограниченного перебора признаков-кандидатов (справа налево) на

последовательное исключение из (3.7). Инвариантность к масштабам измерений признаков в последовательности (3.7) рассматривается как глобальное ограничение в моделях алгоритмов распознавания.

3.2.2. Вычислительный эксперимент

Эксперимент является продолжением исследования выборки данных по 4-канальному жидкостному расходомеру USM [16], начало которому положено в главе II. Параметры выборки: 180 объектов, разделённые на 4 непересекающихся класса (исправный; впрыск газа; эффекты установки; восковая депиляция) и описываемые 43 количественными признаками. По сравнению с главой II привносятся при планировании эксперимента следующие элементы анализа.

1. Формирование последовательности вида (3.7), инвариантной к масштабам измерений признаков.
2. На данных не используется нормирование.
3. Визуально отслеживается структура отношений объектов по значениям сырых признаков на разных наборах признаков по методу t-SNE [9].
4. Демонстрируется график изменения сложности алгоритмов по наборам признаков из последовательности вида (3.7).
5. Показаны особенности реализации метода минимального покрытия выборки объектами-эталонами при использовании одной метрики для всех объектов и локальной метрики для каждого.

Для формирования последовательности вида (3.7) применялся рекурсивный алгоритм к результатам предобработки данных по (1.1) и (3.5). Упорядоченная последовательность из 43 признаков по данным расходомера имела следующий вид:

$$x_{29}, x_{30}, x_{27}, x_{28}, x_{23}, x_{26}, x_{6}, x_{33}, x_{22}, x_{34}, x_{10}, x_{24}, x_{31}, x_{32}, x_{17}, x_{37}, x_{18}, x_{42}, x_{16}, x_{39}, x_{15}, x_{19}, x_{20}, x_{40}, x_{1}, x_{25}, x_{7}, x_{41}, x_{2}, x_{21}, x_{0}, x_{3}, x_{9}, x_{38}, x_{4}, x_{38}, x_{11}, x_{13}, x_{5}, x_{14}, x_{8}, x_{12}, x_{35}. \quad (3.7^*)$$

Наборы для эксперимента из (3.7*) формировались путём отбрасывания (справа–налево) по 5 признаков. Значения сложности алгоритмов по (3.6) приводится в табл. 3.5.

Таблица 3.5. Значения сложности алгоритмов по (3.6)

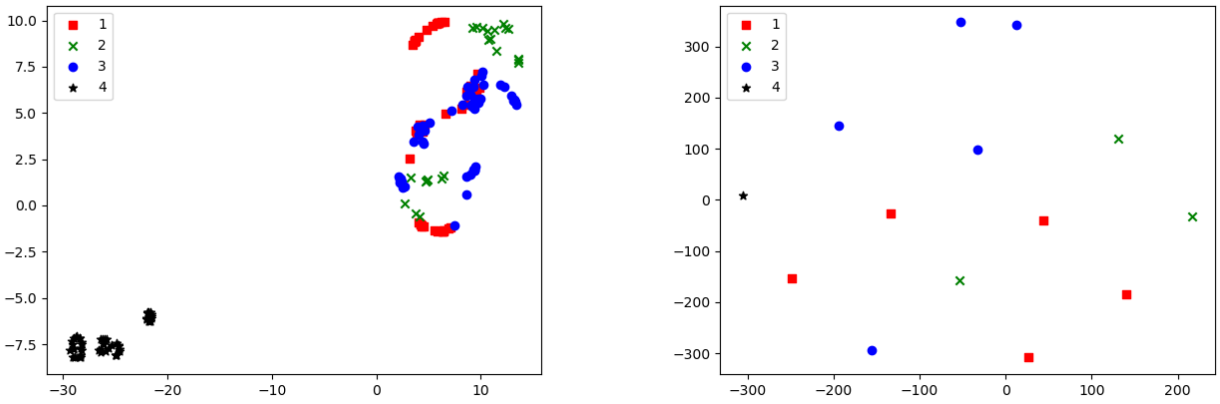
№	**Число признаков в наборе**	**Компактность на выборке по (2.16)**	**Количество объектов**		**Сложность алгоритма по (3.6)**
			Шумовых	Эталонных	
1	43	9.1683	28	14	602
2	38	9.1683	28	14	532
3	33	10.4167	30	12	396
4	28	10.1407	32	12	336
5	23	9.4111	26	14	322
6	18	5.6011	38	20	360
7	13	5.5225	39	20	260
8	8	5.6802	38	20	160
9	3	4.5125	28	36	108

Компактность выборки данных по (2.16) из табл. 3.5 больше чем при их нормировании (см. табл. 2.5). На 43 признаках это соответственно значения 9.1683 и 7.8948, число объектов минимального покрытия 14 при нормировании – 15. График изменения показателя сложности алгоритмов демонстрируется на рис. 3.2.

Рисунок 3.2. – Зависимость сложности алгоритмов от числа признаков

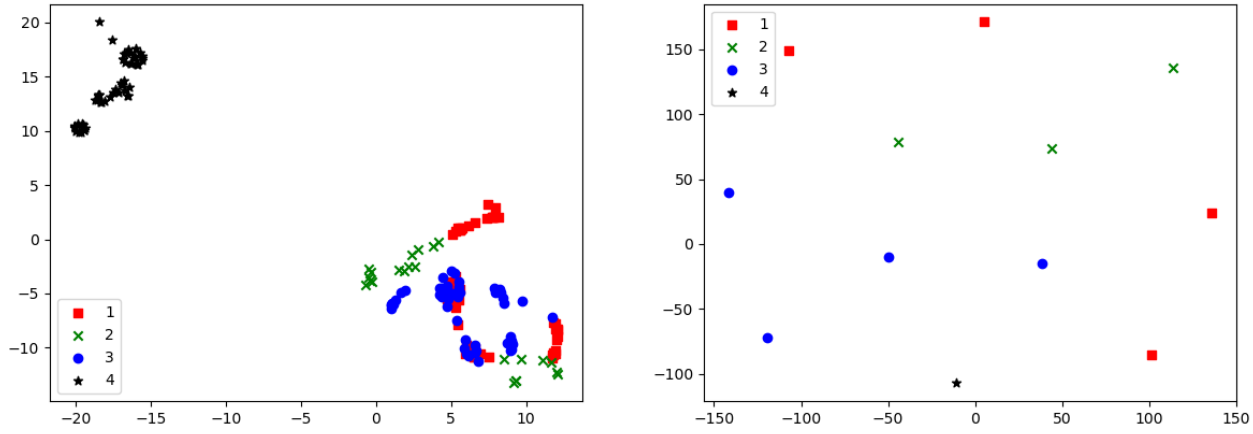
Для визуализации результатов вычислительного эксперимента на разных наборах признаков, приведенных в табл. 3.5, использовался алгоритм нелинейного снижения размерности t-SNE (t-distributed Stochastic Neighbor Embedding) [9]. В методе нет условия для определения «*реперных точек*» для привязки изображения на плоскости.

Визуализация приводится (см. рис. 3.3–3.5) относительно структуры отношений всех объектов выборки в пространстве $X(k)$, $k \le n$ и объектов–эталонов минимального покрытия по их локальной метрике, используемой в (2.15). Локальная метрика объекта $S_k \in K_t$, t=1,…,4 как в главе II вычислялась по формуле $\rho^*(S,S_k)= \rho(S,S_k)/b$, где b – минимальное расстояние от S_k до объекта из дополнения CK_t к классу K_t.



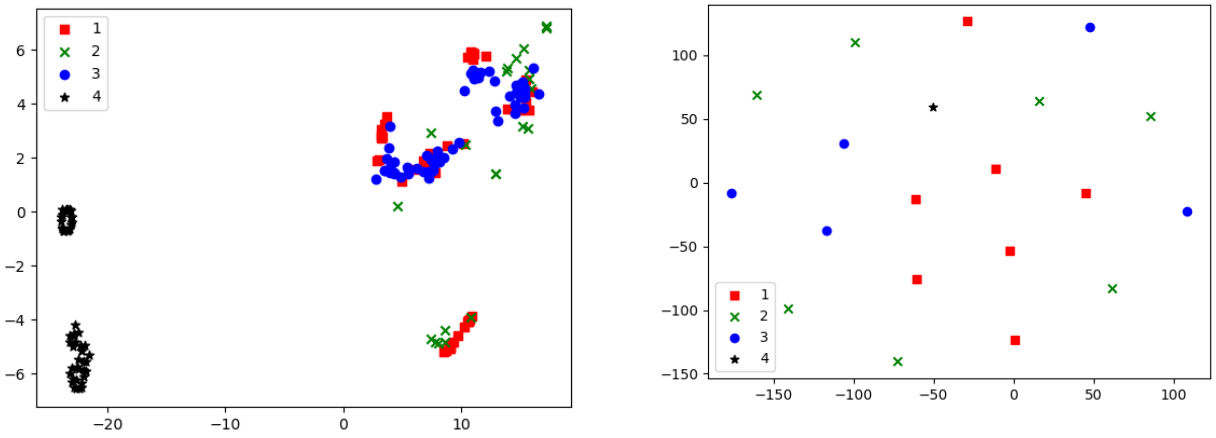
а) б)

Рисунок 3.3. – Визуальное представление объектов по 43 признакам: а–все объекты выборки; б–только объекты-эталоны



а) б)

Рисунок 3.4. – Визуальное представление объектов по 28 признакам: а–все объекты выборки; б– только объекты-эталоны



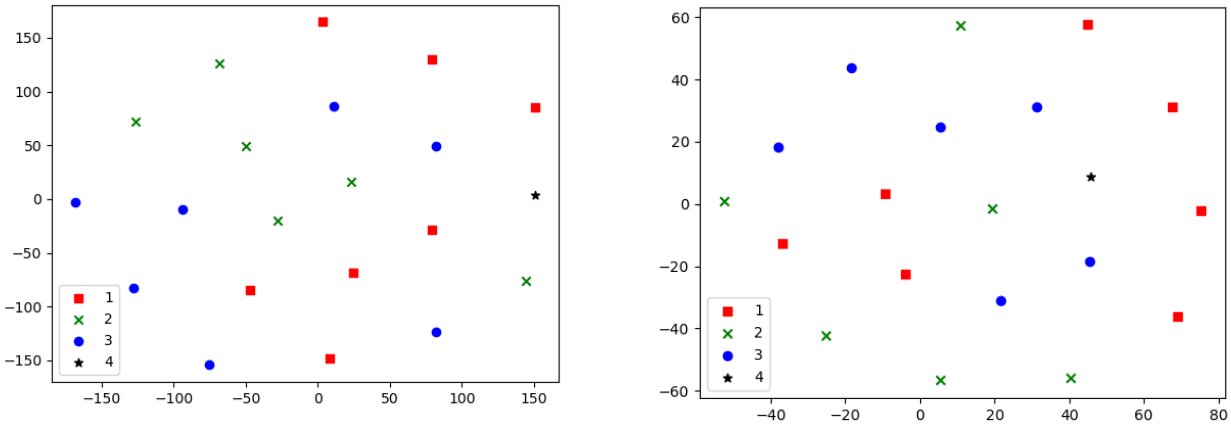
а) б)

Рисунок 3.5. – Визуальное представление объектов по 8 признакам: а–все объекты выборки; б–только объекты-эталоны

Представление объектов-эталонов классов на рис. 3.3 –3.5 порождают иллюзию о квазиравномерном их распределении в сыром признаковом пространстве. Такая особенность объясняется тем, что у каждого эталона есть «*своя*» локальная метрика, используемая при принятии решений по (2.15). На наборах из 43, 28 и 8 признаков (см. рис. 3.3–3.5) отчётливо (визуально) прослеживается отделимость класса K_4 с неисправностью «*восковая депиляция*».

Так как обобщающая способность алгоритмов [8] растёт с увеличением (2.16), то рекомендуется для классификации неисправностей использовать набор из 33 признаков (см. табл. 3.5) со значением меры компактности, равной 10.4116.

Как будет выглядеть минимальное покрытие выборки объектами–эталонами при единой (не локальной) мере расстояния демонстрируется на рис. 3.6 для наборов из 43 и 28 признаков.



а) б)

Рисунок 3.6. – Визуальное представление минимального покрытия объектами–эталонами без локальной метрики: а - по 43 признакам; б - по 28 признакам

В качестве примера покажем, что для набора из 28 признаков (см. рис. 3.6) без локальных метрик объектов показатели из табл. 3.5 будут выглядеть так:

– компактность (2.16) по выборке 6.0844;

– количество объектов шумовых 32, эталонов 20;

– сложность по (3.16) – 460.

В ходе вычислительного эксперимента была показана эффективность выбора эталонных объектов с локальной метрикой, данный результат также визуально отслеживается при использовании метода t-SNE.

3.3 Анализ причин, влияющих на общую выживаемость больных хроническим лимфолейкозом

В качестве материала для исследования использовались данные 123 пациентов с хроническим лимфолейкозом (ХЛЛ) A-C стадии по Binet в возрасте от 47 до 82 лет с известными значениями общей выживаемости ОВ, полученные в гематологическом отделении Ульяновской областной клинической больницы [56, 85, 59, 74].

С позиций системного анализа живой организм – это совокупность взаимодействующих, взаимосвязанных, взаимовлияющих функциональных систем, который можно описать комплексом медико-биологических показателей. На момент постановки диагноза регистрировался возраст пациента, рассчитывался индекс коморбидности Charlson, измерялись стандартные биохимические показатели: аланинаминотрансфераза (АЛТ), аспартатаминотрансфераза (АСТ), общий билирубин, непрямой билирубин, глюкоза, креатинин, мочевина, мочевая кислота, лактатдегидрогеназа (ЛДГ), показатель скорости клубочковой фильтрации (СКФ) по MDRD. При прохождении курса лечения регистрировалось количество сеансов химиотерапии и фактический показатель выживаемости в месяцах. В базу данных не включались больные с ВИЧ-инфекцией и с отличными от ХЛЛ онкологическими состояниями.

Необходимо:

Определить причины (медико-биологические показатели), которые влияют на продолжительность жизни у людей ХЛЛ, с целью принятия решений при выборе тактики лечения пациентов.

Анализ причин общей выживаемости предлагается делать через отбор информативных наборов признаков по таблице «объект–свойство», на основе которых врачи смогут прогнозировать отклонение срока фактической выживаемости пациента с ХЛЛ от срока ОВ, рассчитанной по стандартной системе стадирования [2, 12].

По гендерному принципу были сформированы две выборки данных. Выборка данных больных мужского пола состояла из 60 объектов (возраст 64,6 ± 9,0 лет), женского пола из 63 объектов (возраст 67,0 ± 8,4 лет). Объекты каждой из выборок были разделены на два непересекающихся класса K_1 (фактическая выживаемость меньше прогнозируемой ОВ) и K_2 (фактическая выживаемость больше или равна прогнозируемой ОВ). Разделение объектов выборок на классы показано в табл. 3.6.

Таблица 3.6. Разделение объектов выборок на классы по гендерному принципу

Пол	**Число объектов в классе**	
	K_1	K_2
Мужской	36	24
Женский	42	21

Предварительный (разведывательный) анализ данных на двух выборках пациентов с ХЛЛ проведён на основе разделения количественных признаков на интервалы по (1.1). Границы интервалов и значения критерия (1.1) показаны в табл. 3.7 и табл. 3.8.

Таблица 3.7. Разбиение количественных признаков на интервалы для пациентов мужского пола

Название признака	**Границы интервалов** $[c_1; c_2]$, $(c_2; c_3]$	**Значение критерия (1.1)**
Индекс коморбидности	[2.0..4.0] (4.0..9.0]	0.4012
Возраст	[46.0..63.0] (63.0..87.0]	0.3283
СКФ по MDRD	[42.0..76.0] (76.0..99.0]	0.8141
АЛТ	[6.5..16.8] (16.8..113.8]	0.3136
АСТ	[8.2..19.0] (19.0..70.7]	0.3805
Билирубин общий	[5.9..9.8] (9.8..46.4]	0.2785
Непрямой билирубин	[2.1..6.4] (6.4..35.1]	0.3064
Креатинин	[57.0..84.0] (84.0..177.0]	0.2717
Мочевина	[3.4..6.5] (6.5..11.6]	0.3003
Глюкоза	[3.04..5.11] (5.11..10.5]	0.3003
Количество курсов	[0.0..1.0] (1.0..4.0]	0.3268

Таблица 3.8. Разбиение количественных признаков на интервалы для пациентов женского пола

Название признака	**Границы интервалов $[c_1; c_2]$, $(c_2; c_3]$**	**Значение критерия (1.1)**
Индекс коморбидности	[3.0..5.0] (5.0..9.0]	0.3754
Возраст	[45.0..54.0] (54.0..80.0]	0.2969
СКФ по MDRD	[42.0..64.0] (64.0..96.0]	0.5620
АЛТ	[2.6..11.5] (11.5..135.3]	0.2676
АСТ	[3.8..18.0] (18.0..217.9]	0.3627
Билирубин общий	[6.0..10.0] (10.0..45.3]	0.3348
Непрямой билирубин	[3.5..7.9] (7.9..31.1]	0.3015
Креатинин	[65.0..81.0] (81.0..114.0]	0.3015
Мочевина	[2.5..5.8] (5.8..11.3]	0.3032
Глюкоза	[4.0..4.69] (4.69..8.7]	0.3465
Количество курсов	[0.0..2.0] (2.0..4.0]	0.2575

Как видно из табл. 3.7 и табл. 3.8, значение критерия (1.1) по признакам индекс коморбидности, СКФ по MDRD, возраст, АЛТ, АСТ, непрямой билирубин, количество курсов у пациентов мужского пола было выше, чем у пациентов женского пола.

Информативный набор, полученный по правилу иерархической агломеративной группировки (2.11) при выборе первой пары по (2.12) для пациентов мужского пола был представлен признаками СКФ по MDRD и АЛТ. Для пациентов женского пола первой парой признаков, выбранной по (2.12), было возраст и СКФ по MDRD. Последовательность объединения признаков иерархическим агломеративным алгоритмом по правилу (2.11) показана в табл. 3.9.

Таблица 3.9. Последовательность объединения признаков в информативный набор для пациентов женского пола

Номер шага	**Добавлен признак**	**Значение (2.11)**
1	Количество курсов	0.6825
2	Индекс коморбидности	0.6190
3	Билирубин общий	0.6349
4	Креатинин	0.5714
5	АСТ	0.5714

Анализ табл. 3.9 показывает, что по числу признаков в информативном наборе пациенты женского пола более чем в 3 раза превосходят пациентов мужского пола. Признак СКФ по MDRD является общим по результатам отбора на двух выборках данных.

Результаты отбора информативных признаков согласуются с результатами исследования [83], где сделан вывод о том, что эффективность лечения ХЛЛ зависит не только от вида химиотерапии, но и от функционального состоянием почек на момент диагностики заболевания.

В работе [99] утверждается, что в общем нет и не может быть общей нормы здоровья для всех. При обосновании индивидуальной нормы и выборе курса лечения необходимо брать в расчёт множество различных, в общем случае противоречивых критериев (врачебных рекомендаций). Например, различные ограничения на приём лекарственных препаратов, наличие специфических генетических маркеров у пациента и т.д. Решение проблемы предлагается через выбор методов для формирования информационной модели, в рамках которой можно делать прогноз ОВ с учётом явно и неявно определяемого множества критериев.

Анализ причин, повлиявших на продолжительность срока выживаемости у каждого больного предлагается делать по его индивидуальному информативному набору (признаков) симптомов и синдромов. Сходство между двумя больными может заключаться в том, что фактическая выживаемость у них меньше

прогнозируемого показателя ОВ, рассчитанного статистическими методами [2, 12]. Несмотря на такое сходство, причины низкой (относительно прогнозируемого показателя ОВ) выживаемости у них могут сильно различаться по индивидуальным наборам симптомов и синдромов. С точки зрения теории методов ИАД, обнаружение такого рода различий рассматривается как поиск скрытых закономерностей (новых знаний) в данных.

Отбор индивидуального набора признаков (собственного признакового пространства) позволяет вычислять максимальное значение (оценку) (2.9) для каждого объекта выборки. Оценки можно сравнивать, используя отношения «больше», «меньше» или «равно» и с их помощью давать медицинскую интерпретацию состояния пациентов. С учётом разделения данных по гендерному принципу для интерпретации выбраны разные пороговые значения оценок объектов. Для пациентов мужского пола - это значение равно 0.85, для пациентов женского пола 0.7. Оценки (2.9) пациентов мужского пола в собственном признаковом пространстве приводятся в табл. 3.10 и табл. 3.11.

Таблица 3.10. Оценки пациентов мужского пола из класса K_1 в собственном признаковом пространстве

Номер объекта	**Информативный набор признаков**	**Значение оценки по (2.9)**	**На сколько меньше прожил (мес.)**
0	СКФ по MDRD, непрямой билирубин	0.8888	80
16	Индекс коморбидности, СКФ по MDRD	0.8784	18
18	СКФ по MDRD, непрямой билирубин	0.8888	32
21	СКФ по MDRD, АЛТ	0.9051	4
22	СКФ по MDRD, АСТ	0.9444	16
26	СКФ по MDRD, АСТ	0.9317	53
31	Индекс коморбидности, СКФ по MDRD, глюкоза	0.8888	51
35	Индекс коморбидности, СКФ по MDRD	0.9051	50
36	СКФ по MDRD, АЛТ	0.8888	55
38	СКФ по MDRD	0.875	19

Номер объекта	**Информативный набор признаков**	**Значение оценки по (2.9)**	**На сколько меньше прожил (мес.)**
41	СКФ по MDRD, АЛТ, АСТ	0.9722	33
43	СКФ по MDRD, АЛТ, непрямой билирубин	0.9166	65
44	СКФ по MDRD, АЛТ, мочевина, глюкоза, количество курсов	0.8912	64

Таблица 3.11. Оценки пациентов мужского пола из класса K_2 в собственном признаковом пространстве

Номер объекта	**Информативный набор признаков**	**Значение оценки по (2.9)**	**На сколько больше прожил (мес.)**
1	СКФ по MDRD, АСТ	0.9317	6
2	СКФ по MDRD, АСТ	0.9317	41
3	Возраст, СКФ по MDRD	0.9722	111
13	СКФ по MDRD, АСТ	0.9317	150
19	Возраст, СКФ по MDRD	0.9317	75
28	Возраст, СКФ по MDRD	0.8657	74
34	СКФ по MDRD, АСТ	0.9317	36
47	Возраст, СКФ по MDRD	0.8912	54
48	СКФ по MDRD, АСТ	0.9317	48
55	СКФ по MDRD	0.8750	18

Как видно из табл. 3.10 и табл. 3.11 все информативные наборы объектов содержат признак СКФ по MDRD. Результаты вычисления оценок (2.9) пациентов женского пола в собственном признаковом пространстве приводится в табл. 3.12 и табл. 3.13.

Таблица 3.12. Оценки пациентов женского пола из класса K_1 в собственном признаковом пространстве

Номер объекта	**Информативный набор признаков**	**Значение оценки по (2.9)**	**На сколько меньше прожила (мес.)**
0	Индекс коморбидности, возраст, СКФ по MDRD, билирубин общий, непрямой билирубин, глюкоза	0.7517	21
2	Возраст, СКФ по MDRD, билирубин общий	0.7539	48
4	Индекс коморбидности, СКФ по MDRD, билирубин общий, креатинин, мочевина	0.738	75
5	СКФ по MDRD	0.7131	94
6	Индекс коморбидности, СКФ по MDRD, билирубин общий	0.7324	70
8	СКФ по MDRD	0.7131	49
9	Возраст, СКФ по MDRD, билирубин общий	0.7324	100
12	Индекс коморбидности, возраст, СКФ по MDRD, билирубин общий, непрямой билирубин, глюкоза	0.7517	21
14	Возраст, СКФ по MDRD, билирубин общий	0.7539	48
16	Индекс коморбидности, СКФ по MDRD, креатинин	0.7755	75
17	Возраст, СКФ по MDRD, билирубин общий	0.7324	94
20	СКФ по MDRD	0.7131	49
21	СКФ по MDRD	0.7131	100
26	СКФ по MDRD	0.7131	40
32	Индекс коморбидности, возраст, СКФ по MDRD, мочевина, расчетная выживаемость	0.7709	69
35	Индекс коморбидности, возраст, СКФ по MDRD, креатинин, глюкоза	0.7131	67
37	Возраст, СКФ по MDRD, билирубин общий, мочевина, глюкоза	0.7142	27
38	СКФ по MDRD, билирубин общий, креатинин, мочевина	0.7755	33
40	СКФ по MDRD	0.7131	40
54	Индекс коморбидности, СКФ по MDRD, непрямой	0.7131	24

Номер объекта	Информативный набор признаков	Значение оценки по (2.9)	На сколько меньше прожила (мес.)
	билирубин, глюкоза, расчетная выживаемость		
56	Возраст, СКФ по MDRD, АЛТ, мочевина	0.7539	17
57	Индекс коморбидности, СКФ по MDRD, билирубин общий	0.7131	8
58	Индекс коморбидности, СКФ по MDRD, билирубин общий, непрямой билирубин, креатинин, стадия	0.7539	24
60	Возраст, СКФ по MDRD, АЛТ, мочевина	0.7539	17

Таблица 3.13 Оценки пациентов женского пола из класса K_2 в собственном признаковом пространстве

Номер объекта	Информативный набор признаков	Значение оценки по (2.9)	На сколько больше прожила (мес.)
19	СКФ по MDRD	0.7131	28
29	Возраст, СКФ по MDRD, АЛТ	0.7324	24
31	СКФ по MDRD, АЛТ, АСТ, креатинин	0.7709	3
43	Возраст, СКФ по MDRD, АЛТ	0.7324	24
45	СКФ по MDRD, АЛТ, АСТ, креатинин	0.7709	2
55	СКФ по MDRD	0.7131	10
59	СКФ по MDRD	0.7131	12

Из анализа значений оценок (2.9) (см. табл. 3.10 и табл. 3.11) пациентов мужского пола из классов K_1 и K_2 следуют такие выводы. Высокие (близкие к 1) значения оценок являются следствием сильных межклассовых различий по описаниям пациентов в их собственном признаковом пространстве. Более слабые различия (относительно пациентов мужского пола) имеют место (см. табл. 3.12 и табл. 3.13) в описании по классам K_1 и K_2 пациентов женского пола.

По двум выборкам пациентов с ХЛЛ А-С стадии методами ИАД определено, что скорость клубочковой фильтрации по MDRD является самым информативным показателем для прогнозирования сроков отклонения реальной ОВ от расчетной по стандартной системе стадирования Binet. Возраст пациента на момент постановки диагноза существенного влияния на результаты прогноза не имеет.

У больных ХЛЛ стадии А-С в момент первичного обследования проводят забор периферической венозной крови и стандартными биохимическими методами измеряют уровень креатинина. Скорость клубочковой фильтрации (СКФ) пациентов мужского пола рассчитывается по формуле MDRD [81]:

$$\text{СКФ}=186\times\{[\text{креатинин в сыворотке (плазме)} + 88{,}4]^{-1{,}154}\} \times \text{возраст}^{-0{,}0203}. \quad (3.20)$$

Очевидно, что в (3.20) определена сложная нелинейная зависимость между двумя признаками. Каждый из этих признаков не имеет хорошо выраженной разделимости (см. табл. 3.7), а СКФ, получаемый как результат нелинейной зависимости, демонстрирует хорошую разделимость между классами со значением критерия (1.1), равном 0.8141.

Для выбора границы (порога) между классами использовалась формула $G=(c_2+b)/2$, где $b(b>c_2)$ – ближайшее к c_2 значение признака из $(c_2; c_3]$. По признаку СКФ по MDRD значение порога было определено равным 76.5 [56, 85]. Правило для принятия решения будет выглядеть так:

if СКФ по MDRD<76.5 ***then*** «пациент будет жить больше расчетного срока по стандартной системе стадирования Binet».

Для обоснования выбора порога G между классами вычислялась устойчивость (1.5) разбиения признака на непересекающиеся интервалы. При вычислении использовались значения функции принадлежности к интервалу $t(t=1,2)$ по классу K_i, $i=1,2$. Значение функции принадлежности (1.4) к интервалам по СКФ по MDRD приводятся в таблице 3.14.

Таблица 3.14. Значения функции принадлежности (1.4) по признаку СКФ по MDRD

Интервал	**Класс**	
	***K*1**	***K*2**
[42.0;76.0]	0.8888	0.1112
(76.0; 99.0]	0.0000	1.0000

Устойчивость разбиения по (1.5) для СКФ по MDRD равнялась 0.9277. Показатели СКФ для мужчин, проживших больше расчётных сроков (см. табл. 3.14), лежат в интервале (76.0; 99.0].

Замена исходных значений признака на значение функции принадлежности объектов к классам рассматривается как нелинейное преобразование. Предлагается новый порядок следования использовать для записи признака в номинальной шкале измерений и вычисления его веса. Целью перехода к представлению в номинальной шкале является вычисление обобщённых оценок объектов [65] и поиск скрытых закономерностей (новых знаний) в данных.

Для вычисления обобщённых оценок объектов E_0 будем использовать вклады признаков. Вклад признака x_c по градации $j \in \{1,\ldots,p_c\}$ определяется как

$$\eta_c(j) = \nu_c \left(\frac{\alpha_{cj}^1}{|K_1|} - \frac{\alpha_{cj}^2}{|K_2|} \right). \quad (3.21)$$

где $\alpha_{cj}^1, \alpha_{cj}^2$. – количество значений градации j признака x_c соответственно в классах K_1 и K_2, ν_c– вес признака x_c по (2.19). Обобщённая оценка объекта $S_r \in E_0$ по описанию в номинальной шкале измерений на наборе $X(n)$ и вкладам (3.21) вычисляется как

$$Z(S_r) = \sum_{i=1}^{n} \eta_i(\alpha_{ri}). \quad (3.22)$$

Пусть $d_{tc}(\mu)$, $d_{3\text{-}t,c}(\mu)$ – количество представителей классов K_t, $K_{3\text{-}t}$ в интервале $[r_u;\ r_v]^{\mu}$, вычисленного по (1.3), или градации номинального признака $\mu \in \{1,\ldots,p_c\}$. Суть нелинейных преобразований признаков сводится к замене их исходных значений на значения функции принадлежности объектов к классам. Значение функции принадлежности $f_c(\mu)$ к классу K_1 вычисляется как

$$f_c(\mu) = \frac{d_{1c}(\mu)/|K_1|}{d_{1c}(\mu)/|K_1| + d_{2c}(\mu)/|K_2|}. \quad (3.23)$$

Граница между объектами классов по (3.23) для $x_c \in X(n)$ определяется как

$$G_c = (s1 + s2)/2, \quad (3.24)$$

где $s2 = \max\{f_c(\mu) | 0.5 - f_c(\mu) > 0, \mu = 1,...,p_c\}$, и $s1 = \min\{f_c(\mu) | 1 - f_c(\mu) < 0.5, \mu = 1,...,p_c\}$. Значение градации a_{ic}, $c=1,\ldots,n$ для объекта $S_i=(x_{i1},\ldots,x_{in})$ по (3.24) определяется как

$$a_{ic} = \begin{cases} 1, x_{ic} = \mu, f_c(\mu) < G_c, \\ 2, x_{ic} = \mu, f_c(\mu) > G_c. \end{cases} \quad (3.25)$$

Признаковое пространство для описания объектов по (3.23) будем называть сырым, после использования (3.25) – унифицированным.

Далее демонстрация результатов экспериментов проводится на пациентах мужского пола. Об изменении весов признаков (2.19) можно получить информацию из табл. 3.15. В скобках указано число градаций сырых признаков.

Таблица 3.15. Устойчивость признаков и их веса в сыром и унифицированном пространстве

Название признака (число градаций)	**Веса признаков**		**Устойчивость по (1.5)**
	Сырых	**Унифицированных**	
Индекс коморбидности (2)	0.3820	0.3820	0.7330
Возраст (4)	0.3943	0.3628	0.7292
СКФ MDRD (2)	0.8142	0.8142	0.9278
АЛТ (6)	0.3921	0.3624	0.7728
АСТ (4)	0.4637	0.4266	0.7584
Билирубин общий (6)	0.3945	0.4266	0.7603
Непрямой билирубин (4)	0.3458	0.3829	0.7672
Креатинин (4)	0.3089	0.2714	0.6861
Мочевина (5)	0.3984	0.3424	0.7678
Глюкоза (5)	0.3728	0.3283	0.7026
Общая выживаемость	0.3692	0.2786	0.6268

Название признака	**Веса признаков**		**Устойчивость**
(9)			
Первый курс (5)	0.2986	0.3268	0.6787
Всего курсов (3)	0.3028	0.2657	0.6072

Как видно из табл. 3.16, явного преимущества по увеличению значений весов признаков по (2.19) при использовании нелинейного преобразования (3.25) нет. Самая большая устойчивость 0.9278 (близкая к 1) у признака СКФ MDRD.

Покажем преимущество с точки зрения компактности формирования пространства из унифицированных признаков перед сырыми [74]. Отражением этого преимущества является вычисление обобщенных оценок по (3.22). Для доказательства этого проверялась истинность утверждения (гипотезы) о наличии наборов признаков, по которым объекты классов по обобщённым оценкам будут без ошибок (корректно) разделены на числовой оси.

Результаты проверки истинности гипотезы по минимальному набору унифицированных признаков приводятся в табл. 3.16. Значение границы (порога) между классами по обобщённым оценкам вычислялось по (1.1) аналогично описанному выше для СКФ MDRD. Алгоритм вычисления обобщённых оценок не является жадным. Поэтому точность распознавания можно определять и по обучающей выборке.

Таблица 3.16. Результаты проверки корректности разделения на классы по обобщённым оценкам

Набор признаков	**Граница между классами**	**Точность в %**
Индекс коморбидности, СКФ MDRD, АСТ, Билирубин общий, Непрямой билирубин	0.5103	100.0

В итоге эксперимента выявлено, что при сравнительном анализе точность распознавания по обобщённым оценкам на наборе сырых признаков из табл. 3.16

уменьшилась с 100 до 98,33. Сравнение показателей точности служат доказательством эффективности использования нелинейных преобразований признаков по (3.25).

Выводы по главе 3

1. Описаны два подхода к формированию признакового пространства для описания объектов классов. В первом подходе рассматривается отбор информативных признаков для всей выборки объектов, во втором – для каждого объекта в отдельности;

2. Разработан рекурсивный алгоритм построения упорядоченной по отношению информативности последовательности признаков и критерий для вычисления сложности алгоритма классификации как произведение числа признаков на число объектов локально-оптимального покрытия выборки объектами-эталонами. Показано, что сложность алгоритма (см. табл. 3.5) без снижения компактности по (2.16) со значения 602 на 43 признаках уменьшилась до 322 при отборе 23 информативных признаков из упорядоченной последовательности;

3. При диагностике неисправностей ультразвукового расходомера (см. табл. 3.2) определено, что собственное пространство объекта 37 представлено 6 информативными признаками из 37 исходных. В этом пространстве получено максимальное значение оценки 0.7489 за объект. При значении оценки, равной 1, должна существовать последовательность вложенных гипершаров с центром в указанном объекте, содержащая все объекты класса с эффектами установки.

4. Разработана методика поиска скрытых закономерностей методами интеллектуального анализа данных для больных хроническим лимфолейкозом. При анализе использовались нелинейные преобразования признаков на основе значений функции принадлежности объектов к классам. Описана и обоснована последовательность преобразований признаков от исходного представления до значений в номинальной шкале измерений для вычисления обобщённых оценок объектов. Найдены логические закономерности в форме полуплоскостей,

пороговые значения для которых определены как по отдельным признакам, так и обобщённым оценкам объектов.

ЗАКЛЮЧЕНИЕ

В результате проведённых исследований получено несколько способов снижения вычислительных затрат при выявлении скрытых закономерностей в значениях признаков заданного множества объектов. Получаемый эффект от применения этих способов позволяет считать, что поставленная цель диссертации достигнута. Основными результатами диссертации являются следующие.

1. Разработан численный алгоритм разделения значений признаков в описании допустимых объектов классов на непересекающиеся интервалы с использованием предобработки данных при числе интервалов, равном числу классов. Эффект от применения предобработки заключается в сокращении вычислительных затрат, что дало возможность расширения объёмов, анализируемых данных как по числу признаков, так и по числу объектов. На тестовом примере из 16 объектов показано уменьшение оценки сложности вычислений с 29520 до 1140 соответственно без использования и при использовании предобработки. По экстремальному значению критерия на 43 признаках показаний калибровки 4-х канального ультразвукового расходомера для диагностики 3 видов неисправностей определено, что максимальный вклад в принятие решения о неисправностях дают 2 признака: коэффициенты усиления в начале и конце 4-го пути. Коэффициенты имели одинаковое значение критерия (компактности), равное 0.6464.

2. Получена оценка устойчивости разбиения значений признаков в границах непересекающихся интервалов для выборки данных из двух классов при числе интервалов больше либо равном двум. Множество допустимых значений оценки принадлежит (0.5;1]. Такая оценка является более важным показателем для поиска скрытых закономерностей, чем число интервалов и значения их границ. Устойчивость равна 1, если в каждом интервале представлены объекты только одного из классов. Показано, что устойчивость разбиения признаков на интервалы как при наличии до 30% пропусков в данных, так и при их отсутствии на выборке данных по сегментации изображений отличалась разбросом значений не более 5%.

3. Разработан способ отбора информативных наборов разнотипных признаков по выборке данных с использованием нескольких эвристик. Для сравнения эвристик применялся поиск минимального покрытия выборки объектами-эталонами. Лучшим считался набор признаков, при использовании которого среднее число объектов выборки, притягиваемых одним объектом-эталоном покрытия, было максимальным. Реализован алгоритм отбора информативных признаков объекта по максимальному значению произведения внутриклассового сходства и межклассового различия. Например, для данных по диагностике неисправностей 8-ми лучевого ультразвукового расходомера значение произведения в (0;1] интерпретировалось как индекс тяжести неисправности по индивидуальному набору показателей. Выбор от 2 до 6 информативных показателей и оценка по ним степени тяжести неисправности устройств уменьшают время для диагностики, сокращает материальные расходы на ремонт и ущерб от неверных значений показателей при измерении.

4. Разработаны следующие рекомендации по выбору правил для распознавания объектов с использованием интервальных методов.

– Признаки, значение устойчивости разбиения которых лежит в [0.9; 1] при числе интервалов не больше 4, предложено использовать для формирования *if…then* правил для распознавания. Например, для данных по сегментации изображений было отобрано 19 признаков, в границах разбиения, на интервалы которых сформированы *if …then* правила.

– При вычислении порога логических закономерностей в форме полуплоскостей, при числе интервалов и классов, равным двум, рекомендовано использовать значения границ интервалов.

– Для прогнозирования сроков выживаемости у больных хроническим лимфолейкозом мужского пола рекомендовано использовались обобщённые оценки объектов по унифицированным значениям пяти медико-биологических показателей. Унификация в {1,2} проводилась с помощью значений функции принадлежности объектов к классам, вычисляемых в границах интервалов, определяемых по (4).

Перспективы дальнейшей разработки темы:

- В диссертации для оценивания качества разбиения на интервалы использовано два критерия конкретного вида. В дальнейшем предполагается рассмотреть использование других критериев качества при применении и модернизации интервальных методов с целью извлечения скрытых закономерностей в базах данных.

- В дальнейших исследованиях интервальные методы будут использоваться при разработке интеллектуальных встроенных систем, в которых совмещены процессы «добычи знаний» и принятия управленческих решений. Например, при сертификации датчиков авиационной техники.

СПИСОК ЛИТЕРАТУРЫ

1. Adilova, F. T. The Approach to Individualized Teleconsultations of Patients with Arterial Hypertension / F.T.Adilova, N.A.Ignat’ev, Sh.F.Madrakhimov // Global Telemedicine and eHealth Updates: Knowledge Resources. - 2010. - Vol. 3. - P.372-376.
2. Binet, J. L. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis / J. L.Binet, A.Auquier, G.Dighiero // Cancer. -1981. - T. 48. - C.198-206.
3. Goodfellow, I. Deep Learning / I.Goodfellow, Y.Bengio, A.Courville. - Cambridge: MIT Press, 2016. - 652 p.
4. Gyamfi, K.S. Linear dimensionality reduction for classification via a sequential Bayes error minimization with an application to flow meter diagnostics [Электронный ресурс] / K. S.Gyamfi, J.Brusey, A.Hunt, E.Gaura // Expert Systems with Applications. - 2017. - Режим доступа: https://pure.coventry.ac.uk/ws/portalfiles/portal/13117856/MGLD_Revision.pdf
5. Ignat’ev, N.A. Knowledge Discovering from Clinical Data Based on Classification Tasks Solving / N. A. Ignat'ev, F.T. Adilova, G.R. Matlatipov, P.P. Chernysh // MediNFO. - 2001. - Pp. 1354-1358.
6. Ignat’ev, N.A. New approach neural networks designing: empirical study on acure myocardial infarction predicting / N.A.Ignat’ev, F.T.Adilova, E.H.Ignat’eva // В сборнике: Инфокоммуникационные и вычислительные технологии в науке, технике и образовании труды международной научной конференции. - 2004. - С. 451-454.
7. Ignatev, N.A. The Intelligent Health Index Calculation System / N.A.Ignatev, A.I.Mirzaev // Journal of Pattern recognition and Image Analysis, 2016, № 1, P. 73–77.
8. Ignatyev, N.A. Structure Choice for Relations between Objects in Metric Classification Algorithms / N.A.Ignatyev // Pattern Recognition and Image Analysis. - 2018. - Vol. 28, № 4. - Pp. 590-597.

9. Maaten, L. Visualizing High-Dimensional Data Using t-SNE. [Электронный ресурс] / L.J.P. van der Maaten, G.E.Hinton //Journal of Machine Learning Research. - 2008. - Режим доступа: http://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
10. Molina, L.C. Feature Selection Algoritms: Aervey And Experimental Evaluation / L.C.Molina, L.Belanche, A.Nebot // Proceedings of the 2002 IEEE International Conference on Data Mining. - 2002. - Pp. 306-313.
11. Piatetsky-Shapiro, G. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from ‘university” to “business” and “analytics” / G.Piatetsky-Shapiro // Data Mining and Knowledge Discovery. - 2007. - Vol. 15. - - Pp. 99-105.
12. Rai, K.R. Clinical staging of chronic lymphocytic leukemia / K.R.Rai, A.Sawitsky, EP.Cronkite // Blood. - 1975. - № 46. - Pp. 219-234.
13. Saidov, D.Y. Data visualization and its proof by compactness criterion of objects of classes / D.Y.Saidov // International Journal of Intelligent Systems and Applications (IJISA). - 2017. - Vol. 9, №. 8. - Pp. 51-58.
14. UCI repository of machine learning databases. Image Segmentation Data Set. [Электронный ресурс] - Режим доступа: http://archive.ics.uci.edu/ml/datasets/Image+Segmentation
15. UCI repository of machine learning databases. Statlog (Heart) Data Set. [Электронный ресурс] - Режим доступа: http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29
16. UCI repository of machine learning databases. Ultrasonic flowmeter diagnostics Data Set. [Электронный ресурс] - Режим доступа: http://archive.ics.uci.edu/ml/datasets/Ultrasonic+flowmeter+diagnostics#
17. Zguralskaya, E.N. Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness / E.N.Zguralskaya // Workshop Proceedings Information Technology and Nanotechnology (ITNT-2019). - 2019. - Pp. 283-289.

18. Абдримов, К.Р. Визуализация многомерных данных и свойств объектов в задачах классификации / К.Р.Абдримов, Ш.Ю.Нуржонов // Проблемы информатики и энергетики. - 2012. - №2. - С. 75-80.
19. Адылова, Ф.Т. Оценка степени тяжести хронической сердечной недостаточности с позиции биосимметрики / Ф.Т.Адылова, П.П.Черныш, Е.Н.Згуральская // Украинский журнал телемедицины и медицинской информатики. - 2008. - Т.6, №1. - С.42-47.
20. Айвазян, С.А. Прикладная статистика: Классификация и снижение размерности: Справочное издание / С.А.Айвазян, В.М.Бухштабер, И.С.Енюков, Л.Д.Мешалкин. - М.: Финансы и статистика, 1989. -608 с.
21. Арсеньев, С. Извлечение знаний из медицинских баз данных [Электронный ресурс] / С.Арсеньев // - Режим доступа: http://neural.narod.ru/Arsen.htm
22. Берестнева О.Г. Анализ структуры многомерных данных методом локальной геометрии / О.Г.Берестнева, Е.А.Муратова, А.Е.Янковская // Известия Томского политехнического университета. - 2003. - Т. 306. № 3. - С.19-24.
23. Берестнева, О.Г., Выявление скрытых закономерностей в сложных системах / О.Г.Берестнева, Я.С.Пеккер // Известия Томского политехнического университета. Управление, вычислительная техника и информатика. - 2009. - Т. 315, № 5. - С.138-143.
24. Борисова И.А. Методы решения задач распознавания образов комбинированного типа: автореф. дис. ... канд. тех. наук: 05.13.17 / Борисова Ирина Артемовна. - Новосибирск, 2008. - 23 с.
25. Вапник В. Н. Алгоритмы и программы восстановления зависимостей.- М.: Наука, 1984. – 816 с.
26. Вапник, В. Н. Восстановление зависимостей по эмпирическим данным / В.Н.Вапник - М.: Наука, 1979. - 447с.
27. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин) [Электронный ресурс] / К.В.Воронцов // - Режим доступа: http://www.ccas.ru/voron

28. Воронцов, К. В. Лекции по логическим алгоритмам классификации [Электронный ресурс] / К.В.Воронцов // - Режим доступа: www.MachineLearning.ru.

29. Воронцов, К.В. Профили компактности и выделение опорных объектов в метрических алгоритмах классификации / К.В.Воронцов, А.О.Колосков // Искусственный Интеллект. - 2006. - С. 30-33.

30. Гордеев, Э.Н. Задачи выбора и их решение/ Э.Н.Гордеев // Компьютер и задачи выбора - 1989. - С. 5- 48.

31. Граничин, О.Н. Рандомизация, усреднение и мультиагентные технологии в data mining и управлении / О.Н.Граничин // В сборнике: Лавровские чтения 2013. Материалы пленарных докладов всероссийской научной конференции по проблемам информатики. - 2013. С. 98-114.

32. Граничин, О.Н. Рандомизированные алгоритмы в задачах обработки данных и принятия решений / О.Н.Граничин // Системное программирование. - 2011. - Т.6, №1. - С.141-162.

33. Груман Г. Информационный потенциал больших данных [Электронный ресурс] / Г.Гурман // Технологический прогноз БОЛЬШИЕ ДАННЫЕ: как извлечь из них информацию. - 2010. - №3 - Режим доступа: https://4cio.ru/usercontent/1324/PwC_Technology-Forecast-Issue3%202010_rus.pdf

34. Дуда Р. Распознавание образов и анализ сцен / Р.Дуда, П.Харт - М.:Мир, 1976. - 512 с.

35. Дюк В. А. Методология поиска логических закономерностей в предметной области с нечеткой системологией: На примере клинико-экспериментальных исследований: автореф. дис. … д-ра техн. наук: 05. 13. 01/Дюк Вячеслав Анатольевич. - СПб., 2005.- 33 с.

36. Дюк В. А. Методология поиска логических закономерностей в предметной области с нечеткой системологией: На примере клинико-экспериментальных исследований: дис. … д-ра техн. наук: 05. 13. 01/Дюк Вячеслав Анатольевич. - СПб., 2005. - 309 с.

37. Дюк, В.А. Осколки знаний / В.А.Дюк // Экспресс-Электроника. - 2002. - №6 - С.60-65.
38. Дюк, В.А. Формирование знаний в системах искусственного интеллекта: геометрический подход / В.А. Дюк // Вестник Академии Технического Творчества. - 1996. - № 2. - С.46 -67.
39. Ешмуратов Ш.А. Прозрачность принятия решения при синтезе искусственных нейронных сетей с минимальной конфигурацией: дис. ... канд. тех. наук: 05.13.18 /Ешмуратов Шавкат Артыкбаевич. - Т., 2008. - 120 с.
40. Жамбю М. Иерархический кластер-анализ и соответствия / М.Жамбю. - М.: Финансы и статистика, 1988. - 342 с.
41. Журавлев, Ю.И. Об алгебраических методах в задачах распознавания и классификации / Ю.И.Журавлев // Распознавание, классификация, прогнозирование: Математические методы и их применение. - 1989. - № 1. - С.9-16.
42. Журавлев, Ю.И. Об алгебраическом подходе к решению задач распознования и классификации / Ю.И.Журавлев // Проблемы кибернетики. - 1978. - С. 5-68.
43. Журавлев, Ю.И. Распознавание образов и анализ изображений / Ю.И.Журавлев, И.Б.Гуревич; под общ. ред. Д.А.Поспелова. // Искусственный интеллект. Модели и методы: Справочник. - 1990. - С.149–190.
44. Журавлёв, Ю.И., Гуревич И.Б. Распознавание образов и анализ изображений / Ю.И.Журавлев, И.Б.Гуревич // Искусственный интеллект: Модели и методы. - 2000. - 310 с.
45. Загоруйко, Н. Г. Обучение распознаванию образов без переобучения / Н.Г.Загоруйко, О.А.Кутненко, А.О.Зырянов, Д.А.Леванов // Машинное обучение и анализ данных. - 2014. - Т. 17. - С. 891–901.
46. Загоруйко, Н. Г. Гипотезы компактности и λ-компактности в методах анализа данных / Н.Г.Загоруйко // Сибирский журнал индустриальной математики. - 1998. Т.1, №1. С. 114-126.
47. Загоруйко, Н.Г. Выбор информативных признаков для диагностики заболеваний по генетическим данным / Н.Г.Загоруйко, О.А.Кутненко,

И.А.Борисова, И.И.Дюбанов, А.О.Зырянов, Д.А.Леванов // Вавиловский журнал генетики и селекции. - 2014. - Т.18, № 4/2, - С.898-903.

48. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г.Загоруйко. - Новосибирск: Издательство Института Математики, 1999. - 270 с.
49. Загоруйко, Н.Г. Цензурирование обучающей выборки / Н.Г.Загоруйко, О.А.Кутненко // Вестник Томского гос. Университета. - 2013. - № 1(22). - С.66-73.
50. Зак, Ю.А. Принятие решений в условиях нечетких и размытых данных: Fuzzy-технологии / Ю.А.Зак. - М.: Книжный дом «ЛИБРОКОМ», 2013. - 352 с.
51. Згуральская, Е.Н. Алгоритм выбора оптимальных границ интервалов разбиения значений признаков при классификации / Е.Н.Згуральская // Известия Самарского научного центра РАН. - 2012. - Т.14, №4(3). - С.826-829.
52. Згуральская, Е.Н. Анализ структур отношений между описаниями объектов классов и оценки их компактности / Е.Н.Згуральская // В сборнике: Информационные технологии и нанотехнологии (ИТНТ-2019) труды V международной конференции и молодежной школы. - 2019. - С. 166-170.
53. Згуральская, Е.Н. Выбор информативных признаков для решения задач классификации с помощью искусственных нейронных сетей / Е.Н.Згуральская // Нейрокомпьютеры: разработка, применение. - 2012. - № 2. - С. 20-27.
54. Згуральская, Е.Н. Иерархический кластерный анализ данных и снижение размерности признакового пространства / Е.Н.Згуральская // В сборнике: Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. - 2015. - С. 220-222.
55. Згуральская, Е.Н. Поиск закономерностей по значениям количественных признаков с помощью детерминистических критериев разбиения на интервалы // В сборнике: Междисциплинарные исследования в области математического моделирования и информатики труды 3-й научно-практической конференции. - 2014. - С.199-203.

56. Згуральская, Е.Н. Поиск скрытых закономерностей в форме полуплоскостей интервальными методами / Е.Н.Згуральская // В сборнике: Современные проблемы проектирования, производства и эксплуатации радиотехнических систем труды XI всероссийской научно-практической конференции. - 2019. - С.249-251.
57. Згуральская, Е.Н. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей / Е.Н.Згуральская // Известия Самарского научного центра РАН. - 2018. - Т.20, № 4(3). - С.451-455.
58. Зиновьев, А.Ю. Визуализация многомерных данных / А.Ю.Зиновьев. Красноярск: КГТУ, 2000. - 180 с.
59. Игнатьев, Н.А Нелинейные преобразования признаков и поиск закономерностей на данных больных хроническим лимфолейкозом / Н.А.Игнатьев, Е.Н.Згуральская, М.В.Марковцева // В сборнике: Информационные технологии и нанотехнологии (ИТНТ-2020) труды VI международной конференции и молодежной школы. - 2020. - С. 123-128.
60. Игнатьев, Н.А. Анализ данных и принятие решений с помощью логических закономерностей в форме полуплоскостей / Н.А.Игнатьев, Д.Ю.Саидов // Известия Самарского научного центра РАН. - 2017. - Т.19, №4(2). - С.294-299.
61. Игнатьев, Н.А. Выбор минимальной конфигурации нейронных сетей / Н.А.Игнатьев // Вычислительные технологии. - 2001. - Т.6, № 1. - С.23-28.
62. Игнатьев, Н.А. Выбор параметров регуляризации для повышения обобщающей способности дискриминантных функций / Н.А.Игнатьев, Ш.Ю.Нуржонов // Известия Академии Вооруженных сил Республики Узбекистан. - 2014. - № 1(14). - С.81-87.
63. Игнатьев, Н.А. Выбор собственного пространства объекта с использованием нелинейных преобразований признаков / Н.А.Игнатьев // Информационные технологии. - 2018. - Т. 24, №10. - С.665-670.
64. Игнатьев, Н.А. Вычисление обобщенных оценок и иерархическая группировка признаков / Н.А.Игнатьев // Вестник Томского государственного университета. - 2015. - С. 31-38.

65. Игнатьев, Н.А. Вычисление обобщённых показателей и интеллектуальный анализ данных / Н.А.Игнатьев // Автоматика и телемеханика. - 2011. - №5 - С.183-190.
66. Игнатьев, Н.А. Вычисление сложности эффективных алгоритмов выбора оптимальных границ интервалов / Н.А.Игнатьев, Д.Ю.Саидов // Проблемы информатики и энергетики. - 2014. - №6. - С.35-40.
67. Игнатьев, Н.А. Извлечение явных знаний из разнотипных данных с помощью нейронных сетей / Н.А.Игнатьев // Вычислительные технологии. -2003. - Т.8, №2. - С.69-73.
68. Игнатьев, Н.А. Индексирование объектов по индивидуальным наборам информативных признаков / Н.А.Игнатьев // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2016. - № 4(37). - С.27-35.
69. Игнатьев, Н.А. Итеративный метод построения линейных оболочек и информативных множеств классов в задачах распознавания / Н.А.Игнатьев // Проблемы управления и информатики. - 2002. - № 3. - С.133-137.
70. Игнатьев, Н.А. О некоторых способах повышения прозрачности нейронных сетей / Н.А.Игнатьев, Ш.Ф.Мадрахимов // Вычислительные технологии. - 2003. - Т. 8, № 6. - С.31-37.
71. Игнатьев, Н.А. Отбор признаков в собственное пространство объекта на основе меры его компактности / Н.А.Игнатьев, А.И.Мирзаев // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. - 2019. - № 49. - С.55-62.
72. Игнатьев, Н.А. Синтез факторов в искусственных нейронных сетях / Н.А.Игнатьев // Вычислительные технологии. - 2005. - Т.10, №3. - С.32-38.
73. Игнатьев, Н.А. Устойчивость и обобщённые оценки классифицированных объектов в разнотипном признаковом пространстве / Н.А.Игнатьев, Ш.Ф.Мадрахимов // Вычислительные технологии. - 2011. - Т.16, № 2, - С.70-77.

74. Игнатьев, Н.А., Згуральская Е.Н., Марковцева М.В. Поиск скрытых закономерностей, влияющих на общую выживаемость больных, методами интеллектуального анализа данных / Н.А.Игнатьев, Е.Н.Згуральская, М.В.Марковцева // Искусственный интеллект и принятие решений. - 2020. - №3. - С.73-80.
75. Колесникова, С.И. Методы анализа информативности разнотипных признаков / С.И.Колесникова // Вестник Томского государственного университета. - 2009. - №1(6). - С.69-80.
76. Крашенинников, В.Р. Способ отбора информативных признаков для решения задачи классификации / В.Р.Крашенинников, Е.Н.Згуральская // REDS: Телекоммуникационные устройства и системы. - 2015. - Т. 5, № 4. - С. 324-327.
77. Лбов, Г.С. Методы обработки разнотипных экспериментальных данных / Г.С.Лбов. - 1981. - 160 с.
78. Мадрахимов, Ш.Ф. Выбор латентных признаков по результатам иерархической агломеративной группировки данных / Ш.Ф.Мадрахимов, Д.Ю.Саидов // Актуальные проблемы прикладной математики и информационных технологий. - 2016. - С. 88-91.
79. Мадрахимов, Ш.Ф. Построение нечётких правил вывода для диагностики нестабильности атеросклеротической бляшки / Ш.Ф.Мадрахимов, Г.А.Розыходжаева // Врач и Информационные технологии. - 2018. - № 3. - С.81-88.
80. Марухина О.В., Мокина Е.Е., Берестнева Е.В. Применение методов data mining для выявления скрытых закономерностей в задачах анализа медицинских данных [Электронный ресурс]/ О.В.Марухина, Е.Е.Мокина, Е.В.Берестнева // Фундаментальные исследования. – 2015. - №4. - Режим доступа: https://www.fundamental-research.ru.
81. Медицинский информационный сайт [Электронный ресурс] - Режим доступа: https://medqueen.com/medicina/diagnostika/diagnostika-statya/1966-skorost-klubochkovoy-filtracii-skf.html

82. Наследов, А.Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / А.Д.Наследов. - СПб.: Питер, 2005. - 416 с.

83. Никитина, А.К. Эффективность лечения и выживаемость больных хроническим лимфолейкозом в зависимости от почечной функции / А.К.Никитина, Н.О.Сараева // Забайкальский медицинский вестник. - 2014. - № 4. - С.122-127.

84. Носова, С.С. Economics: словарь современной экономической теории / С.С. Носова. - Москва: Русайнс, 2016. - 254 с.

85. Патент РФ 2725877. Способ прогнозирования общей выживаемости больных хроническим лимфолейкозом мужского пола в стадии А-С/ Марковцева М. В., Згуральская Е. Н. Бюл. № 19. – 3 с. Опубл. 07.07.2020

86. Переверзев-Орлов В.С. Советчик специалиста: опыт создания партнерской системы / В.С.Переверзев-Орлов. - М.: Наука, 1990. -133 с.

87. Потапов А. С. Технологии искусственного интеллекта – СПб: СПбГУ ИТМО, 2010. - 218 с.

88. Розыходжаева, Г.А. Изучение информативности параметров неинвазивных методов диагностики в качестве маркеров старения у больных ишемической болезнью сердца / Г.А.Розыходжаева, Е.Н.Игнатьева // Врач и информационные технологии. – М., 2006. - № 1 - С. 38-44.

89. Розыходжаева, Г.А. Сравнительный анализ мер информативности в прогнозе 5-летней смертности больных ИБС пожилого и старческого возраста / Г.А.Розыходжаева, Е.Н.Згуральская // В сборнике трудов международной конференции. Компьютерная медицина. - 2007.

90. Саидов, Д.Ю. Информационные модели на основе нелинейных преобразований признакового пространства в задачах распознавания: дисс. … д-ра физ.-мат. наук: 05.13.17/Саидов Дониер Юсупович. - Т., 2017. - 93 с.

91. Саидов, Д.Ю. Обобщающая способность алгоритмов распознавания с учётом нелинейности / Д.Ю.Саидов, Ш.А.Нуржанов // Проблемы информатики и энергетики. - 2016. - №1. - С.33-39.

92. Смагин, А.А. Разработка базы знаний для экспертной системы морского мониторинга / А.А.Смагин, С.В.Липатова, Е.С.Кукин // Автоматизация процессов управления. - 2009. - № 4. - С.31-39.

93. Субботин, С. А. Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов / С. А.Субботин // Математичні машини і системи. - 2010. - № 1. - С. 25-39.

94. Таранова, Н.Н. Метод адаптивного кодирования признаков / Н.Н.Таранова // В сборнике: Динамика систем. - 1995. - С. 54-70.

95. Ту, Дж. Принципы распознавания образов / Дж.Ту, Р.Гонсалес. - М.: Мир, 1978. - 416 с.

96. Убайдуллаева, Р.Т. Дифференциация менталитета студентов методами интеллектуального анализа данных / Р.Т.Убайдуллаева, Н.А.Игнатьев // Организация и самоорганизация интелегенции в современном российском обществе. - 2013. - С.400-407.

97. Царегродцев, В.Г. Производство полуэмпирических знаний из таблиц данных с помощью обучаемых искусственных нейронных сетей / В.Г.Царегродцев // Методы нейроинформатики. - 1998. - С.176-198.

98. Черняк, Л. Большие данные - новая теория и практика / Л.Черняк // Открытые системы. - 2011. - № 10.

99. Шумаков, В.И. Моделирование физиологических систем организма / В.И.Шумаков, В.Н.Новосельцев, М.П.Сахаров, Е.Ш.Штенголд. – М: Медицина. - 1971. - 352 с.

100. Юлдашов Р. У. Интеллектуальный анализ данных в нейроэкспертных системах и задачи прогнозирования: дисс. … канд. тех. наук: 05.13.17/Юлдашев Равшанбек Уринбаевич. - Т., 2011.- 107 с.

101. Янковская А.Е. Унификация разнотипных данных в интеллектуальных распознающих системах / А.Е.Янковская, Е.А.Муратова, О.Г. Берестнева // В сборнике: Труды Международной научно-практической конференции. Знание-Диалог-Решение (KDS–2001). - 2001. - С.661-668.

Приложение 1.

УТВЕРЖДАЮ
Главный врач ГУЗ «Ульяновская
областная клиническая больница»
_____ **Н.А. Манина**
«08» 08 2019 года

АКТ ВНЕДРЕНИЯ
результатов диссертационной работы
Згуральской Екатерины Николаевны

Результаты диссертационной работы «Применение интервальных методов для поиска скрытых закономерностей по описаниям объектов классов» нашли свое научное и практическое применение при решении задачи отбора информативных наборов признаков, на основе которых врачи смогут прогнозировать отклонение срока фактической выживаемости пациентов с хроническим лимфолейкозом (ХЛЛ) от срока общей выживаемости (ОВ), рассчитанной по стандартной системе стадирования Binet.

По гендерному принципу были сформированы две выборки данных. Каждая выборка данных была разделена на два класса пациентов, не доживших до расчётных сроков ОВ и проживших больше расчётных сроков. Результаты прогнозирования были получены алгоритмами методов отбора информативных признаков для всей выборки данных и по каждому пациенту в отдельности. Было доказано, что скорость клубочковой фильтрации по MDRD является самым информативным показателем для прогнозирования сроков отклонения реальной ОВ от расчетной по стандартной системе стадирования Binet. Возраст пациента на момент постановки диагноза существенного влияния на результаты прогноза не имеет.

Использование результатов вычисления информативных признаков в клинической практике позволит для каждого пациента с ХЛЛ объективно обосновывать отклонение реальных сроков ОВ от рассчитанных по стандартной системе стадирования Binet.

зав. гематологическим отделением
ГУЗ Ульяновской областной
клинической больницы Н.Б. Есефьева

Приложение 2.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ИЗОБРЕТЕНИЕ

№ 2725877

Способ прогнозирования общей выживаемости больных хроническим лимфолейкозом мужского пола в стадии А-С

Патентообладатель: ***Марковцева Мария Владимировна (RU)***

Авторы: ***Марковцева Мария Владимировна (RU),***
Згуральская Екатерина Николаевна (RU)

Заявка № **2020106661**
Приоритет изобретения **11 февраля 2020 г.**
Дата государственной регистрации в
Государственном реестре изобретений
Российской Федерации **07 июля 2020 г.**
Срок действия исключительного права
на изобретение истекает **11 февраля 2040 г.**

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев