**НАРЗУЛЛАЕВ Д.З., ТУРСУНОВ А.Т.,**
**БАЙДУЛЛАЕВ А.С.**

# ОБРАБОТКА РАЗНОТИПНОЙ ИНФОРМАЦИИ В ЗАДАЧАХ АНАЛИЗА ДАННЫХ
## (Монография)

Ташкент – 2022

**УДК: 338.2.001.76**
**ББК 74.5**
**М-27**

**Нарзуллаев Д.З., Турсунов А.Т., Байдуллаев А.С. Обработка разнотипной информации в задачах анализа данных.** –Т.: «ILMIY - TEXNIKA AXBOROTI – PRESS NASHRIYOTI», 2022, 152 стр.

***Рецензенты:***

**Нишанов А.Х. -** доктор технических наук, профессор

**Абдурахманов Б.А. -** кандидат физико-математических наук, доцент

**ISBN 978-9943-4719-0-0**

В монографии рассматриваются методы и алгоритмы анализа данных при машинной обработке экспериментальной информации разнотипной природы. Исследуются разновидности и модели программного обеспечения прикладной статистики. Даётся обзор существующих методов преобразования типов признаков, указываются преимущества и недостатки этих методов.

Монография предназначена научным исследователям и специалистам, преподавателям и студентам, руководителям и инженерно-техническим работникам, заинтересованным в области рассматриваемых вопросов.

**УДК: 338.2.001.76**
**ББК 74.5**

***Утверждено к печати научным советом Ташкентского фармацевтического института 29 декабря 2021 года, протокол №5.***

**ISBN 978-9943-4719-0-0**

**О Г Л А В Л Е Н И Е**

# ВВЕДЕНИЕ

Современному бизнесу необходимы аналитические средства. Потребность постоянно уменьшать издержки производства, оптимизировать складские запасы, исследовать рынок и прогнозировать его развитие поддерживают интерес к подобным технологиям. На Западе без консультации с аналитиками не решается ни один серьезный вопрос. Однако следует заметить, что во многих областях четкие алгоритмы мирно сосуществуют с интуитивными догадками. Правильное использование современных аналитических инструментов может дать хорошие результаты, поэтому в последнее время интерес к разработкам в этой области проявляют многие компании, желающие повысить свои показатели. Проблема заключается в том, что большая часть аналитического инструментария рассчитана на крупные корпорации и стоит баснословных денег. Как правило, это специализированные технологии, поддерживающие работу с огромными массивами данных и требовательные к аппаратной платформе. Для малого и среднего бизнеса практически ничего нет, хотя потребность в анализе они также испытывают. Требования к подобным системам продиктованы здравым смыслом: реализация современных механизмов анализа данных, способность работать с выборками в сотни тысяч записей, интеграция с офисными приложениями, стабильная работа на офисных персональных компьютерах (ПК), возможность использования для анализа данных из разнородных источников,

доступная цена, простота использования. Именно эти предпосылки мы брали за основу при создании системы анализа данных SAD. Потребность в создании такой системы возникла в результате проведения многочисленных маркетинговых исследований, проводимых в Национальной компании экспортно-импортного страхования «Узбекинвест» для уточнения стратегии её развития в вопросах непрерывного улучшения качества и потребительских свойств страховых услуг, совершенствования страховых технологий и освоения новых видов страхования, повышения экономического роста и максимизации прибыли.

Из истории известно, что типовые примеры раннего этапа применения статистических методов описаны в Ветхом Завете. С математической точки зрения они сводились к подсчетам числа попаданий значений наблюдаемых признаков в определенные градации. В дальнейшем результаты стали представлять в виде таблиц и диаграмм, как это и сейчас делает Госкомстат.

Сразу после возникновения теории вероятностей (Паскаль, Ферма, 17 век) вероятностные модели стали использоваться при обработке статистических данных. Например, изучалась частота рождения мальчиков и девочек, было установлено отличие вероятности рождения мальчика от 0,5, анализировались причины того, что в парижских приютах эта вероятность не та, что в самом Париже, и т.д. Имеется достаточно много публикаций по истории теории вероятностей, однако наболее известной из них является работа академика Украинской АН Б.В.Гнеденко, включившего в

последнее издание своего курса [1] главу по истории математики случайного числа.

В 1794 г. (по другим данным - в 1795 г.) К.Гаусс разработал метод наименьших квадратов, один из наиболее популярных ныне статистических методов, и применил его при расчете орбиты астероида Церера - для борьбы с ошибками астрономических наблюдений [2]. В 19 веке заметный вклад в развитие практической статистики внёс бельгиец Кетле, на основе анализа большого числа реальных данных показавший устойчивость относительных статистических показателей, таких, как доля самоубийств среди всех смертей [3]. Статистические методы управления качеством, сертификации и классификации продукции сейчас весьма актуальны [4].

Современный этап развития прикладной статистики можно отсчитывать с 1900 г., когда англичанин К.Пирсон основан журнал "Biometrika". Первая треть XX в. прошла под знаком параметрической статистики. Изучались методы, основанные на анализе данных из параметрических семейств распределений, описываемых кривыми семейства Пирсона. Наиболее популярным было нормальное (гауссово) распределение. Для проверки гипотез использовались критерии Пирсона, Стьюдента, Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи планирования эксперимента.

Разработанная в первой трети XX в. теория называется параметрической статистикой, поскольку ее основной объект изучения - это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами. Как правило, нельзя указать каких-либо веских причин, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением, и т.д. Однако в подавляющем большинстве реальных ситуаций подобных моделей нет, и приближение реального распределения с помощью кривых из семейства Пирсона или его подсемейств - чисто формальная операция. Именно из таких соображений критиковал параметрическую статистику академик С.Н.Бернштейн в 1927 г. в своем докладе на Всероссийском съезде математиков [5].

Согласно классификации статистических методов, принятой в [6,7], прикладная статистика делится на следующие четыре области:

- статистика числовых (случайных) величин;
- многомерный статистический анализ;

– статистика временных рядов и случайных процессов;

– статистика объектов нечисловой природы;

Первые три из этих областей являются классическими. Остановимся на четвертой, только еще входящей в массовое сознание специалистов. Ее именуют также статистикой нечисловых данных или попросту нечисловой статистикой.

Исходный объект в математической статистике - это выборка. В вероятностной теории статистики выборка - это совокупность независимых одинаково распределенных случайных элементов. В классической математической статистике элементы выборки - это числа. В многомерном статистическом анализе - вектора. А в нечисловой статистике элементы выборки - это объекты нечисловой природы, которые нельзя складывать и умножать на числа. Другими словами, объекты нечисловой природы лежат в пространствах, не имеющих векторной структуры.

Примерами объектов нечисловой природы являются:

- значения качественных признаков, т.е. результаты кодировки объектов с помощью заданного перечня категорий (градаций);
- упорядочения (ранжировки) экспертами образцов продукции (при оценке ее технического уровня и конкурентоспособности) или заявок на проведение научных работ (при проведении конкурсов на выделение грантов);

- классификации, т.е. разбиения объектов на группы, сходные между собой (кластеры);
- толерантности, т.е. бинарные отношения, описывающие сходство объектов между собой, например, сходства тематики научных работ, оцениваемого экспертами с целью рационального формирования экспертных советов внутри определенной области науки;
- результаты парных сравнений или контроля качества продукции по альтернативному признаку ("годен" - "брак"), т.е. последовательности из 0 и 1;
- слова, предложения, тексты;
- ответы на вопросы экспертной, маркетинговой или социологической анкеты, часть из которых носит количественный характер (возможно, интервальный), часть сводится к выбору одной из нескольких подсказок, а часть представляет собой тексты.

С начала *70-х* годов под влиянием запросов прикладных исследований в технических, медицинских и социально-экономических науках активно развивается статистика объектов нечисловой природы, известная также как статистика нечисловых данных или нечисловая статистика. Большую роль в развитии этого направления сыграл основанный в 1973 г. научный семинар "Экспертные оценки и анализ данных". В 60-е годы научное сообщество стало интересоваться методами экспертных оценок. Как следствие, началось знакомство с конкретными

математизированными теориями, связанными с этими методами. Речь идет о репрезентативной теории измерений, ставшей известной в нашей стране по статье П.Суппеса и Дж.Зинеса в сборнике [8] и книге И.Пфанцагля [9], о теории нечеткости Л.А.Заде [10], теории парных сравнений, описанной в монографии Г.Дэвида [11].

В течение 70-х годов на основе запросов теории экспертных оценок (а также социологии, экономики, техники и медицины) развивались конкретные направления статистики объектов нечисловой природы. Были установлены связи между конкретными видами таких объектов, разработаны для них вероятностные модели [12].

Следующий этап - выделение статистики объектов нечисловой природы в качестве самостоятельного направления в прикладной статистике, ядром которого являются методы статистического анализа данных произвольной (разнотипной) природы. Программа развития этого нового научного направления впервые была сформулирована в статье [13]. Реализация этой программы была осуществлена в 80-е годы. Для работ этого периода характерна сосредоточенность на внутренних проблемах нечисловой статистики.

К 90-м годам статистика объектов нечисловой природы с теоретической точки зрения была достаточно хорошо развита, основные идеи, подходы и методы были разработаны и изучены математически, в частности, доказано достаточно много теорем.

Однако она оставалась и остаётся на сегодняшний день недостаточно апробированной на практике. Это связано как с ее сравнительной молодостью, так и с общеизвестными особенностями организации науки в 80-е годы, когда отсутствовали достаточные стимулы к тому, чтобы теоретики занялись широким внедрением своих результатов. И в 90-е годы наступило время от математико-статистических исследований перейти к применению полученных результатов на практике. Эта тенденция хорошо отражена в монографиях [14,15,17], материалах международной конференции "Управление большими системами" [16].

Следует отметить, что в статистике объектов нечисловой природы, как и в других областях прикладной математической статистики и прикладной математики вообще, одна и та же математическая схема может с успехом применяться и в технических исследованиях, и в медицине, и в социологии, и для анализа экспертных оценок, а потому ее лучше всего формулировать и изучать в наиболее общем виде, для объектов произвольной природы.

Перейдём к изложению основных идей статистики объектов нечисловой природы.

В чем принципиальная новизна нечисловой статистики? Для классической математической статистики характерна операция сложения. При расчете выборочных характеристик распределения (выборочное среднее арифметическое, выборочная дисперсия и

др.), в регрессионном анализе и других областях этой научной дисциплины постоянно используются суммы. Математический аппарат - законы больших чисел, Центральная предельная теорема и другие теоремы - нацелены на изучение сумм. В нечисловой же статистике нельзя использовать операцию сложения, поскольку элементы выборки лежат в пространствах, где нет операции сложения. Методы обработки нечисловых данных основаны на принципиально ином математическом аппарате - на применении различных расстояний в пространствах объектов нечисловой природы. Другой подход к решению данной задачи заключается в сведении исходной информации разнотипной природы к одному виду.

Кратко рассмотрим несколько идей, развиваемых в статистике объектов нечисловой природы для данных, лежащих в пространствах произвольного вида. Решаются классические задачи описания данных, оценивания, проверки гипотез - но для неклассических данных, а потому неклассическими методами.

Первой обсудим проблему определения средних величин. В рамках репрезентативной теории измерений удается указать вид средних величин, соответствующих тем или иным шкалам измерения [18]. В классической математической статистике средние величины вводят с помощью операций сложения (выборочное среднее арифметическое, математическое ожидание) или упорядочения (выборочная и теоретическая медианы). В пространствах произвольной природы средние значения нельзя

определить с помощью операций сложения или упорядочения. Теоретические и эмпирические средние приходится вводить как решения экстремальных задач. Для теоретического среднего это - задача минимизации математического ожидания (в классическом смысле) расстояния от случайного элемента со значениями в рассматриваемом пространстве до фиксированной точки этого пространства (минимизируется указанная функция от этой точки). Для эмпирического среднего математическое ожидание берется по эмпирическому распределению, т.е. берется сумма расстояний от некоторой точки до элементов выборки и затем минимизируется по этой точке. При этом как эмпирическое, так и теоретическое средние как решения экстремальных задач могут быть не единственным элементом пространства, а состоять из множества таких элементов, которое может оказаться и пустым. Тем не менее удалось сформулировать и доказать законы больших чисел для средних величин, определенных указанным образом, т.е. установить сходимость эмпирических средних к теоретическим.

Оказалось, что методы доказательства законов больших чисел допускают существенно более широкую область применения, чем та, для которой они были разработаны. А именно, удалось изучить асимптотику решений экстремальных статистических задач, к которым, как известно, сводится большинство постановок прикладной статистики [19]. В частности, кроме законов больших чисел установлена и состоятельность оценок минимального контраста, в том числе

оценок максимального правдоподобия и робастных оценок. К настоящему времени подобные оценки изучены также и в интервальной статистике.

В статистике в пространствах произвольной природы большую роль играют непараметрические оценки плотности, используемые, в частности, в различных алгоритмах регрессионного, дискриминантного, кластерного анализов. В нечисловой статистике предложен и изучен ряд типов непараметрических оценок плотности в пространствах произвольной природы, в частности, доказана их состоятельность, изучена скорость сходимости и установлен примечательный факт совпадения наилучшей скорости сходимости в произвольном случае с той, которая имеет быть в классической теории для числовых случайных величин.

Дискриминантный, кластерный, регрессионный анализы в пространствах произвольной природы основаны либо на параметрической теории - и тогда применяется подход, связанный с асимптотикой решения экстремальных статистических задач - либо на непараметрической теории - и тогда используются алгоритмы на основе непараметрических оценок плотности.

Для проверки гипотез могут быть использованы статистики интегрального типа, в частности, типа омега-квадрат. Известно, что предельная теория таких статистик, построенная первоначально в классической постановке [20], приобрела естественный вид именно для пространств произвольного вида

[21], поскольку при этом удалось провести рассуждения, опираясь на базовые математические соотношения, а не на те частные (с общей точки зрения), что были связаны с конечномерным пространством.

Для анализа нечисловых, в частности, экспертных данных весьма важны методы классификации. С другой стороны, наиболее естественно ставить и решать задачи классификации, основанные на использовании расстояний или показателей различия, в рамках статистики объектов нечисловой природы. Это касается как распознавания образов с учителем (другими словами, дискриминантного анализа), так и распознавания образов без учителя (т.е. кластерного анализа).

Статистические методы анализа нечисловых данных особенно хорошо приспособлены для применения в экономике, социологии и экспертных оценках, поскольку в этих областях от 50% до 90% данных являются нечисловыми.

Вопросы программной реализации методов анализа занимают особое место при создании автоматизированных систем научных исследований (АСНИ). Конечной целью разработчиков программных средств является создание экспертных систем анализа данных, позволяющих наряду с быстрой и эффективной выдачей конечных результатов оказывать помощь исследователям предметных областей в интерпретации этих результатов.

Целью работы является: разработка методов и алгоритмов оцифровки неколичественных признаков, создание на основе

классических и предложенных в работе методов диалоговой интегрированной системы анализа данных, ориентированной на проведение научных исследований конечным пользователем.

Для достижения поставленной цели в работе решаются следующие задачи:

- Создание диалоговой интегрированной системы анализа данных для решения основных задач обработки экспериментальной информации;
- Разработка алгоритмов поиска целочисленных числовых меток градаций неколичественных признаков;
- Создание на основе классических и разработанных алгоритмов системы анализа данных для решения в АСНИ задач обработки разнотипной экспериментальной информации.

Научная новизна результатов работы:

– Предложена структура диалоговой интегрированной системы анализа данных, позволяющей на уровне специалиста в области анализа данных формировать заключения о природе исследуемого явления на основе исходной информации;

– Предложена структура экспертной системы анализа данных.

Практическая ценность работы заключается в следующем:

– Создано программное обеспечение разработанных методов и алгоритмов, ориентированное на пользователя-непрограммиста, являющегося специалистом в определённой предметной области;

– Разработана система анализа данных SAD для ПК, позволяющая решать задачи исследования взаимозависимостей признаков, распознавания образов, автоматической классификации, регрессионного анализа.

Основные научные результаты.

Апробация работы. Основные научные результаты работы обсуждались и получили положительную оценку на научных семинарах и совещаниях в Ташкентском фармацевтическом институте, а также на различных международных и республиканских конференциях.

.

# ГЛАВА I. Методы и алгоритмы анализа данных при машинной обработке экспериментальной информации разнотипной природы

В данной главе рассматриваются методы и алгоритмы анализа данных при машинной обработке экспериментальной информации разнотипной природы. Исследуются разновидности и модели программного обеспечения прикладной статистики. Даётся обзор существующих методов преобразования типов признаков, указываются преимущества и недостатки этих методов. Формулируются цель и задачи работы.

## I.1. Общая проблематика

Научные исследования проводятся в различных направлениях, таких как дедуктивные построения, экспериментальные исследования и обработка их результатов, подбор и ознакомление с необходимыми публикациями, патентами на изобретения и т.п. Кибернетическое моделирование возникло с появлением и широким внедрением в нашу жизнь ЭВМ, при этом эксперименты проводятся не с самими объектами исследований, а с их описаниями на том или ином понятном машине языке. В этом случае исследования приобретают эволюционный характер, так как по мере накопления информации об изучаемом явлении происходит уточнение выбранной гипотезы.

Системы автоматизации, имеющие дело с перечисленными выше экспериментами, получили общее название систем

автоматизации научных исследований (АСНИ) [22]. В настоящее время развитие типовых средств АСНИ происходит в двух направлениях – разработка специализированных вычислительных комплексов и систем для разного рода исследований и ориентация математического обеспечения этих систем на полное решение задач определённых этапов, например, сбор и архивирование данных, математическая обработка, интерпретация результатов обработки и др. В функции АСНИ входит автоматизация практически всех участков работы экспериментатора: измерений, настройки аппаратуры, планирования эксперимента, создание и хранение банков данных первичной экспериментальной информации.

Одной из важнейших частей АСНИ является статистическое программное обеспечение - универсальный и мощный инструмент решения фундаментальных задач обработки экспериментальных данных с целью повышения их достоверности и информационного сжатия в научных исследованиях, а также задач анализа, прогноза и управления для сложных экономических, научно-технических и технологических систем. Решение на современном уровне целого ряда актуальных задач в таких важнейших направлениях, как экономический анализ, социология, геология, автоматизация и управление сложными технологическими процессами, невозможно без широкого применения и использования статистического программного обеспечения (СПО) или программного обеспечения прикладной статистики. В свою

очередь, прикладная статистика определяется как самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приёмы, математические методы и модели, предназначенные для сбора, стандартной записи, автоматизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов [18]. Прикладная статистика широко используется в различных направлениях науки, техники, медицины и других областях благодаря её постоянно развивающемуся программному обеспечению.

В настоящее время происходит интенсивное развитие СПО, позволяющее решать основные задачи анализа и обработки экспериментальных данных. Сводные данные о современном СПО представлены в [22-26]. На сегодняшний день практически все программные продукты по обработке экспериментальных данных реализованы для ПЭВМ типа IBM PC и их можно отнести к одному из следующих видов.

**Независимые программы** были наиболее распространены в 60-х годах прошлого столетия. Такие программы разрабатывались, как правило, программистами-одиночками и являлись реализацией частных случаев обработки данных. Здесь необходимо отметить жёсткую структуру, плохую совместимость с другими программами и, как следствие, недолговечность данного вида СПО. Этот вид СПО практически недоступен исследователю-непрограммисту ввиду того, что он должен

вручную настраивать параметры программы, входные и выходные данные. До сих пор в нашей стране и за рубежом распространяется огромное число одиночных статистических программ различной сложности. Как правило, эти программы используются самими разработчиками для доказательства эффективности работы разработанных алгоритмов и методов анализа экспериментальных данных.

В результате стандартизации отдельных компонентов независимых программ (ввод-вывод, согласование параметров) оформления их в виде подпрограмм, а также организации сервисных средств возникли **библиотеки программ** (БП), реализующих статистическую обработку данных в пределах одного или нескольких разделов обработки информации. БП – самая динамичная часть системы программного обеспечения, содержащая основной объём знаний ЭВМ. Программы БП обычно не содержат операторов ввода-вывода, ориентированы на пользователя-программиста и используются, как правило, вместе с языком программирования высокого уровня. Пользователю БП самому приходится составлять управляющую программу для вызова необходимых модулей, ввода исходной информации и вывода конечных результатов. Вопросам конструирования библиотек программ посвящена публикация [27], где сказано, что одним из основных вопросов, стоящих перед разработчиками БП, является создание качественной документации. Этот вопрос включает в себя выделение всех категорий пользователей, на

которых ориентирована библиотека, и предоставлении каждой из них той и только той информации, которая нужна данной категории пользователей. При этом пользователями считаются не только исследователи предметных областей, но и программисты, работающие над созданием других библиотек программ.

В процессе своей работы программы БП должны либо выдавать результат, либо печатать сообщение о недопустимости обработки данных с выдачей на печать причин такого решения.

Основным недостатком БП является тот факт, что для решения одной конкретной задачи статистической обработки данных требуются значительные затраты труда программистов, однако БП могут служить основой для создания других более совершенных и мощных видов статистического программного обеспечения.

Следующим видом СПО являются **пакеты прикладных программ** (ППП), определяемые как совокупность программ для решения типичных задач в различных прикладных областях. ППП характеризуются наличием:

- специализированного входного языка управления заданиями, позволяющим управлять вводом-выводом данных, функциями манипулирования данными;
- управляющей программы, генерирующей конкретную технологическую цепочку на основе входного задания;
- библиотеки функциональных модулей.

Эксплуатационные характеристики отдельных программ ППП непрерывно совершенствуются на основе информации, полученной от пользователей. Сведения об изменениях, внесённых в программу или пакет в целом, сообщаются всем пользователям, что особенно важно при огромных затратах на разработку прикладных программ и их эксплуатацию.

К этому виду СПО относятся пакеты БИМ-М [28], ОТЭКС [29], ППСА [30], а также большое множество других ППП статистического анализа. Современные ППП имеют широкие возможности в части общения с системами управления базами данных (СУБД), такими как ORACLE, MS SQL SERVER и другими. Таким образом, пользователи получают возможность в зависимости от своих запросов формировать гибкие проблемно-ориентированные ППП, включающие общестатистический ППП, а также, в случае необходимости, специализированную СУБД.

**Системы анализа данных (САД)**, как и ППП, предназначены для решения наиболее часто встречающихся задач обработки и анализа данных. САД определяется как единая многофункциональная программа, реализующая ввод-вывод данных, архивацию данных, функции редактирования и генерации технологических цепочек анализа входной информации.

Отсутствие у большинства пользователей, представителей конкретных предметных областей, глубокой математической подготовки и достаточного навыка в программировании обуславливает следующие требования к созданию САД [31]:

- Осуществление в процессе работы системы встроенной методики по анализу данных, позволяющей исследователю в ходе обработки информации осуществить “вычислительный эксперимент” по проверке своих гипотез и предположений об исследуемом процессе или явлении;
- Возможность решения широкого круга задач по анализу данных с использованием минимального набора наиболее эффективных алгоритмов их решения;
- Простота обращения к системе, возможность общения с системой на языке предметной области без применения специальных языков программирования или управления заданиями.

В дополнение к вышеуказанным требованиям необходимо отметить, что САД должна иметь возможность извлечения анализируемых данных их баз данных и, соответственно, должна быть интегрирована с имеющимися системами управления базами данных. На сегодняшний день это требование является наиболее актуальным, поскольку экспериментальные данные, как правило, заносятся в заранее подготовленные базы данных с чётко определённой структурой.

Ядром САД является управляющая программа (внутренний монитор), позволяющая работать исследователю либо в автоматическом режиме (при этом все функции по выбору соответствующих методов для решения конкретной задачи анализа данных берёт на себя монитор), либо в интерактивном

режиме. Монитор системы не исключает возможности совместного использования интерактивного (активный диалог) и пассивного (пассивный диалог) режимов. Структура САД представлена на рис. I.1.

В [31] отмечается, что для наиболее эффективного использования системы со стороны исследователей, представляющих различные предметные области и не являющихся специалистами по анализу данных и программированию, внутренний монитор должен обеспечивать взаимосвязанную работу всех модулей системы согласно определённой методике, позволяющей исследователю в ходе обработки информации осуществить “вычислительный эксперимент” по проверке своих гипотез и предположений об исследуемом явлении. Для осуществления такой методики САД должна отвечать следующим основным требованиям:

- Обеспечивать гибкость монитора системы, перенастраиваемого в ходе “вычислительного эксперимента“ по анализу данных;
- Наряду с реализацией основных и наиболее важных методов анализа данных осуществлять выполнение различных процедур по коррекции и преобразованию исходной таблицы экспериментальных данных (ТЭД).

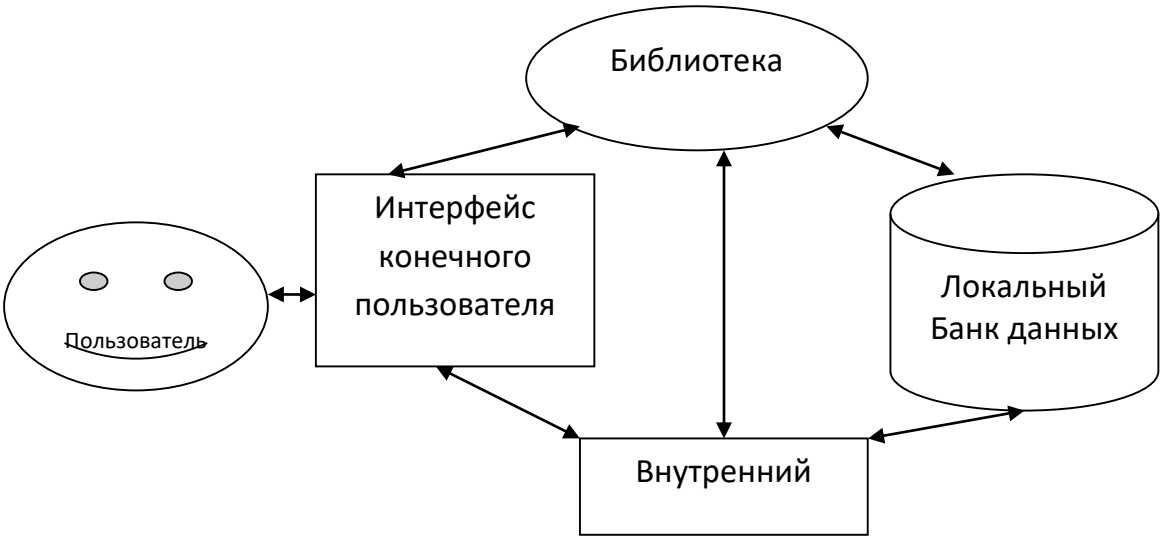


Рис. I.1. Структура системы анализа данных.

Интерфейс конечного пользователя обеспечивает интерактивное взаимодействие с системой пользователя-непрограммиста на языке предметной области без использования специальных языков программирования.

Локальный банк данных (ЛБД) предназначен для хранения исходных данных, промежуточных результатов и управляющей информации. Через него осуществляется связь программных модулей системы. При этом обрабатывающие модули ядра (библиотека модулей) являются информационно независимыми и обмениваются информацией через общую, специально организованную на внешнем носителе область памяти, физически представляющую ЛБД. При работе системы последовательность подключения обрабатывающих модулей, а, следовательно, и последовательность ввода-вывода в ЛБД вычисляемых промежуточных результатов будет зависеть от выбранной

исследователем стратегии проведения анализа данных. Необходимо отметить, что ЛБД системы является лишь вспомогательным средством межмодульного обмена. В то же время, как было отмечено выше, большие объёмы экспериментальных данных, предназначенных для обработки на ЭВМ, как правило, хранятся в базах данных, существующих самостоятельно и независимо от систем анализа данных. В этом случае необходимо создавать интегрированные программные средства, которые бы позволили в рамках одной системы объединить функции хранения, представления и поиска информации, традиционно возлагаемые на СУБД, с функциями анализа и выявления эмпирических закономерностей из имеющихся данных.

Использование интегрированного подхода к анализу данных [32] даёт возможность пользователю применять различные методы анализа данных, находить оптимальные параметры задачи, сопоставлять выходную информацию. В интегрированных САД обработка данных ведётся поэтапно, на каждом этапе уточняются закономерности, присущие анализируемой информации. В основу интегрированного подхода к анализу данных положена идея о том, чтобы САД была интегрирована по методам и средствам анализа данных с целью наилучшего использования навыков и знаний конечного пользователя.

Перечислим основные положения данного подхода [32,33]:

•По области своего применения САД должна быть предметно-ориентированной. Такая ориентация достигается за счёт максимально возможного использования профессиональных знаний пользователя и разработки некоторых специализированных программных модулей, учитывающих специфику предметной области;

•САД должна быть интегрированной по методам и средствам обработки информации, а также по совокупности обрабатываемых данных. Интеграция средств обработки означает разработку систем, совмещающих в себе функции хранения, обработки данных и принятия решений на основе использования универсальных пакетов программ обработки данных, машин баз данных, средств отображения и визуализации. Интеграция методов означает применение специальной последовательности алгоритмов обработки данных, позволяющей достигать достоверности получаемого результата за счёт применения к одним и тем же данным множества допустимых некоррелированных алгоритмов, поэтапно уточняющих решение. В свою очередь, интеграция данных обеспечивается совместной обработкой в САД доступных типов и структур данных при решении конкретных прикладных задач;

•САД должна быть максимально приближена к конечному пользователю, который не является программистом и не является специалистом в области анализа данных. Пользователю должны быть предоставлены удобные средства общения с САД,

позволяющие в то же время продуктивно использовать не формализуемые профессиональные знания.

**Экспертные системы анализа данных** (ЭСАД) ориентируются, как правило, не на универсальные статистические задачи обработки экспериментальной информации, а на конкретные приложения, такие как задачи социологических исследований, медицинской диагностики, обработки результатов геологических исследований и т.п. Здесь в полной мере реализуется идея интегрированного подхода к анализу данных. В литературе [34,35] экспертные системы определяются как вычислительные системы, в которые включены знания специалистов о некоторой конкретной предметной области и которая в пределах этой области способна принимать экспертные решения. В [36] предложена структура ЭСАД, в состав которой входят (см. рис. I.2):

- Банк данных, состоящий из реальных данных, записанных в каком-либо виде, либо имеющихся в базе данных. Во втором случае для извлечения анализируемой информации из базы данных требуется использование СУБД;
- Банк алгоритмов – алгоритмическое ядро системы – совокупность всех обрабатывающих модулей, использование которых позволяет сделать выводы о природе изучаемого явления;
- База знаний, включающая в себя экспертные знания пользователя и экспертные знания разработчика ЭСАД;

•Система управления работой обрабатывающих алгоритмов (внутренний монитор), реализующая определённую стратегию обработки данных, предложенную разработчиками;

•Лингвистический процессор, с помощью которого осуществляется настройка ЭСАД на задачи конкретной предметной области. В этом блоке происходит формализация задачи на языке моделей анализа данных и интерпретация результатов на языке предметной области.

Таким образом, экспертная система анализа данных – это вариант системы анализа данных, ориентируемый на конкретные прикладные задачи, решение которых основывается на широком использовании современных статистических процедур и методологии общего анализа данных в сочетании с эвристическими и формальными методами обработки информационных сообщений. Эти сообщения включаются в базу знаний, являющуюся одним из основных элементов ЭСАД и содержащую в себе как знания в области анализа данных, так и знания в исследуемой предметной области. Экспертные системы анализа данных представляют собой новый тип специализированных пакетов прикладных программ, характерной чертой которых является наличие программных средств для ведения базы знаний, тематически ориентированной на определённый класс задач. В качестве примера чисто статистической экспертной системы, ориентированной на решение задач регрессионного анализа, можно привести проект

REX, реализованный на основе построения специальной диалоговой настройки над универсальным статистическим пакетом S [37].

Экспертные системы принято относить к одной из основных форм высшего уровня интеллектуализации прикладного программного обеспечения. Их создание связано с разработкой методов и средств формализации и ввода знаний в компьютерные системы (блоки 3 и 6 на рис.I.2) и манипулирования введёнными знаниями.

В [38] отмечается, что необходимо остановиться ещё на одном факторе, стимулирующем развитие работ в области создания ЭСАД. Всё возрастающие объёмы информации, требующие грамотной статистической обработки, и огромное количество СПО, в основном в виде специальных пакетов и библиотек находятся в явном дисбалансе с относительно медленно растущей численностью квалифицированных специалистов в области прикладной статистики. В результате большое число неспециалистов в области статистического анализа данных используют СПО независимо от того, получили ли они одобрение специалистов по прикладной статистике. В конечном счёте неграмотное использование СПО приводит к неверным результатам и дискредитации аппарата прикладной статистики. Включение опыта специалистов по прикладной статистике в базу знаний ЭСАД, в первую очередь в области разведочного анализа данных, выбора подходящих моделей и нужной

последовательности применяемых методов, интерпретации промежуточных и конечных результатов статистического анализа, позволит ослабить развитие упомянутого процесса неквалифицированного использования СПО.

Создатели большинства известных к настоящему времени ЭСАД [39] ставили перед собой решение следующих вопросов:

- выдачу подсказок по существующим литературным, методическим и программным материалам, относящимся к специфике исследуемой задачи;
- генерирование советов в выработке адекватных исходных допущений о природе обрабатываемых данных и в выборе общего вида модели;
- предложение ограниченного количества подходящих методов анализа данных с пояснением особенностей применения этих методов;
- автоматическая реализация на ЭВМ технологической цепочки методов и алгоритмов для решения поставленной задачи;
- оказание помощи в интерпретации промежуточных и конечных результатов статистического анализа;
- оказание помощи в выборе форм представления результатов проведённого статистического анализа.

Основной круг пользователей, на которые рассчитаны подобные ЭСАД, это прикладные статистики и математики разного уровня квалификации, а также специалисты различных предметных областей, обладающие вероятностно-статистической

подготовкой в объёме экономического или технического вуза. К таким ЭСАД можно отнести серию методо-ориентированных статистических экспертных систем МОСЭС, разработанных в Центральном экономико-математическом институте и совместном предприятии «Диалог» (Москва).

На начальной стадии обработки данных необходимо определение структуры данных (СД) в целях изучения взаимоотношений между исследуемыми объектами [18,40]. В свою очередь, целью анализа СД является построение математической модели исследуемого явления в виде функционального, статистического или иного описания. Такая модель позволяет заменить экспериментальные данные как способ представления явления на некоторый более общий закон, из которого исходные данные вытекают уже как частный случай. Следовательно, СД можно рассматривать как пространственное выражение закономерностей, которые эти данные представляют. В свою очередь анализ СД предполагает два необходимых этапа:

• разбиение исходного множества объектов на непересекающиеся классы – на этом этапе используются методы распознавания образов или автоматической классификации;

• поиск законов, с помощью которых можно описать поведение объектов в каждом из классов – здесь используются методы регрессионного анализа.

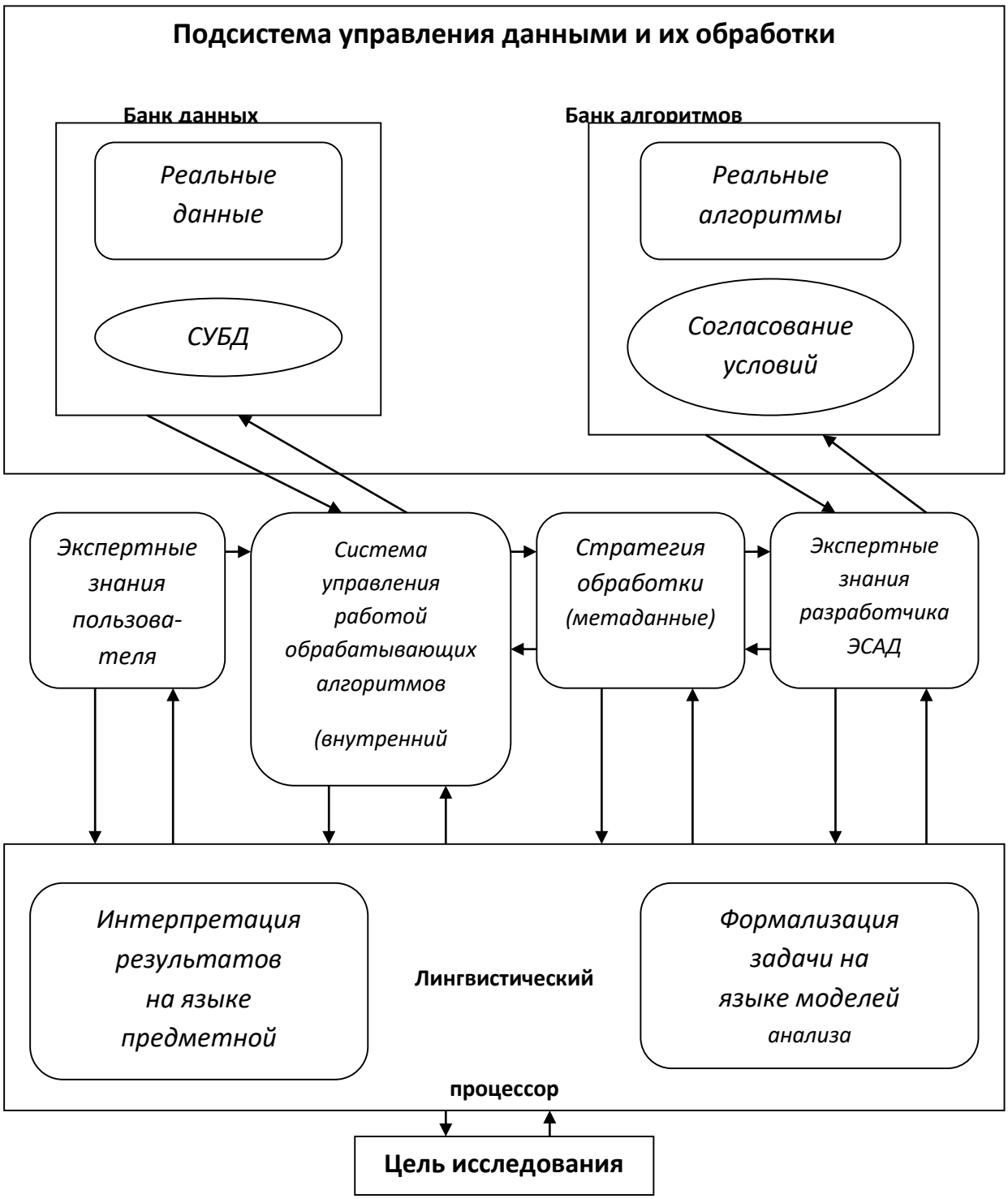


Рис I.2. Структура экспертной системы анализа данных.

Рассмотрим более подробно задачу распознавания образов.

Пусть исходная информация представлена в виде таблицы экспериментальных данных (ТЭД) «объект-признак (свойство)» размером N*p, где N – число объектов-строк, а p – число признаков-столбцов в ТЭД:

$$\mathbf{X}=(X_1, X_2, \ldots, X_N), \qquad (I.1)$$

в которой $X_i=(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)})'$ – вектор значений анализируемых признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, зарегистрированных на i-м обследованном объекте.

Разбиение исходного множества из N объектов на непересекающиеся классы при наличии обучающей выборки (набора объектов, принадлежность которых к образам указана) является основной целью решения задачи распознавания образов. При этом необходимо использовать некоторое решающее правило, которое бы с минимальным числом ошибок классифицировало вновь появляющиеся наблюдения.

В основе большинства статистических методов распознавания образов лежит идея о том, что классифицируемый объект считается принадлежащим к тому образу, плотность объектов которого наивысшая в данной точке пространства описания. Если плотности распределения для каждого из k образов известны, т.е. заданы описывающие их функции:

$$\begin{aligned} f_1 &= f_1(x^{(1)}, x^{(2)}, \ldots, x^{(p)}), \\ f_2 &= f_2(x^{(1)}, x^{(2)}, \ldots, x^{(p)}), \\ &\ldots\ldots\ldots\ldots\ldots \\ f_k &= f_k(x^{(1)}, x^{(2)}, \ldots, x^{(p)}), \end{aligned} \qquad (I.2)$$

то проблема решается подстановкой в каждую из формул (I.2) значений признаков классифицируемого объекта и вычислением величин $f_1$, $f_2$, …, $f_k$. Если распознаваемые объекты из различных образов появляются с одинаковой частотой, то номер большей из $f_i$ и есть номер образа, к которому скорее всего принадлежит классифицируемый объект. Такая процедура распознавания называется байесовской и обеспечивает минимальную вероятность неправильной классификации.

На практике функции плотности образов известны очень редко и в распоряжении исследователя имеются лишь обучающие выборки. В этом случае общая идея, на которой строятся алгоритмы распознавания, состоит в восстановлении или оценивании функций плотности по обучающим выборкам и процедура распознавания проводится описанным выше способом.

Перечислим некоторые наиболее распространённые подходы к решению задачи распознавания:

1. Параметрические методы. Здесь вид функций плотности образов предполагается известным – обычно соответствующим нормальному закону распределения, но неизвестными считаются параметры каждой функции – математическое ожидание, корреляционная матрица и дисперсии всех признаков, которые и подлежат оцениванию по обучающим выборкам. Однако основная сложность их применения в том, что принимаемая гипотеза о виде плотности может оказаться неверной и, соответственно,

результаты решения задачи распознавания оказываются ошибочными.

2. Непараметрические методы, основанные на локальных оценках плотности. Оценив плотность для каждого из образов в той точке, где располагается распознаваемый объект, можно принять решение о его принадлежности, отнеся его к образу с наивысшим значением оценки плотности. Позволяя проводить классификацию, эти методы не обеспечивают возможности явного описания решающего правила – разделяющей образы поверхности в виде функции от значения признаков.

3. Непараметрические методы, основанные на заданности вида разделяющих функций. В этом случае оцениванию по обучающим выборкам подлежат параметры этих функций при предположении, что разделяющие поверхности линейны. Таким образом, оцениваются не сами плотности образов, а только границы, на которых они равны друг другу и которые однозначно определяют правило классификации объектов. Эти методы наиболее важны в случае, когда требуется получить явное описание решающего правила.

Все вышеописанные подходы к решению задачи распознавания образов рассчитаны на наличие в исходной ТЭД признаков одного типа – количественных. Поэтому, в процессе предварительной обработки данных может возникнуть необходимость в преобразовании типов признаков. Это связано с тем, что очень часто исходная ТЭД содержит признаки, которые

могут быть измерены в различных шкалах. Наиболее распространёнными на практике являются следующие типы шкал: наименований, порядка, количественные. В свою очередь, по отношению к соответствующим шкалам обычно различают следующие типы признаков [29, 40]:

- Классификационные (номинальные) признаки измеряются в шкале наименований. Эта шкала позволяет разбить объекты на группы (классы), однородные по свойству данного признака *x*. При этом никакого упорядочения групп признаком *x* не вводится. Если все возможные классы заранее известны, то говорят о категоризированной номинальной шкале (шкале категорий), а сами классы называются категориями признака *x*. Каждому значению признака *x* можно присвоить определённый код (градацию) – число, которое играет роль имени значения признака или соответствующего этому значению класса эквивалентности. Объекты, входящие в один класс эквивалентности, считаются неразличимыми по значениям данного признака. Примером некатегоризированного номинального признака является название (или номер) страны, профессия, происхождение и т.д.

- Качественные (ординальные, порядковые) признаки измеряются в шкале порядка. Эта шкала отличается от классификационной наличием отношения линейного порядка для заданных классов эквивалентности. Признак *x*, измеренный в ординальной шкале, позволяет упорядочивать (по признаку *x*) объекты по степени проявления свойства, описываемого этим

признаком, но не даёт количественной меры для его выражения. Ординальная шкала может быть категоризированной или некатегоризированной. Если по рассматриваемому ординальному признаку объекты можно разделить на заранее известное число классов, то говорят о наличии категоризированной шкалы. Примером такой шкалы может служить оценка на экзамене. Ординальная шкала будет некатегоризированной, например, при ранжировании объектов по степени проявления какого-либо свойства, точная количественная мера для которого не определена. Каждой градации категоризированного ординального признака можно присвоить числовой код (метку) таким образом, чтобы порядок чисел соответствовал порядку его значений.

- Количественные признаки измеряются в шкалах интервалов, отношений, разностей. Такие признаки могут принимать непрерывный ряд значений из некоторого диапазона допустимых значений, либо допустимые значения можно пронумеровать (например, количество голов крупного рогатого скота, приходящегося на ферму). Разделение количественных признаков на непрерывные и дискретные до некоторой степени условно, поскольку из-за ограничений точности измерения даже показателей, непрерывных по своей природе (таких, например, как длина, масса), любой показатель может считаться как дискретный.

При решении реальных задач во многих случаях результаты исследований представлены таблицей экспериментальных данных, включающей одновременно значения признаков

различного типа. Такие исходные данные называют разнотипными, а рассматриваемое пространство – пространством разнотипных признаков. Шкальные значения каждого из типов признаков допускают лишь определённые операции и преобразования [8, 9]. Например, для количественных признаков допустимыми являются все математические операции и преобразования, для качественных – отношения порядка и сравнения, для классификационных – только отношения эквивалентности. Из сказанного следует, что определённая шкала допускает вычисление определённого набора статистических характеристик. Вследствие этого многие классические математико-статистические методы обработки данных, такие как дискриминантный, факторный, регрессионный анализ, оказываются неприменимыми для всего объёма информации разнотипной природы и используются только для обработки экспериментальных данных, измеренных в количественной шкале. Для анализа исходных данных, измеренных в шкалах наименований и порядка, используются методы, основанные на использовании порядковых статистик и различных мер связи категоризированных переменных [41, 42]. Результаты анализа разнотипных данных интерпретируются независимо друг от друга и это мешает пониманию исследуемого явления как единого целого. Закономерности и свойства изучаемых объектов исходной ТЭД, которые отражаются в связях между разнотипными показателями, оказываются не выявленными. Таким образом,

конечные результаты обработки разнотипных данных не отражают всего многообразия внутренних связей и искажают сущность исследуемого явления.

Рассмотрим наиболее распространённые методы обработки разнотипных данных.

## I.2. Методы преобразования типов признаков

Из сказанного в п. I.1 следует вывод о том, что при наличии в исходной эмпирической матрице, представленной в виде (I.1), признаков, измеренных в различных шкалах, на начальном этапе анализа данных необходимым условием является преобразование типов признаков. Здесь и далее под понятием «преобразование типов признаков» будем рассматривать переход от различных типов признаков к одному виду: количественному, качественному или классификационному.

Рассмотрим основные этапы обработки данных, присущие статистическому анализу экспериментальной информации, полученной в результате проведения опытов в исследуемой предметной области [31, 43, 44]:

1. Постановка задачи. На этом этапе решаются следующие задачи:

1.1. Определение цели исследования;

1.2. Определение состава данных;

1.3. Сбор данных;

1.4. Формализация данных (шкалирование).

2. Ввод данных в обработку. Здесь можно выделить следующие пункты:

2.1. Вод данных в память ЭВМ и работа с архивом данных;

2.2. Формирование задания для обработки.

3. Качественный анализ. На этом этапе необходимо выполнение трёх задач, а именно:

3.1. Преобразование типов признаков (при наличии разнотипных признаков);

3.2. Определение простейших характеристик данных;

3.3. Визуализация данных;

3.4. Анализ структуры данных.

4. Количественное описание данных. Один из важнейших этапов анализа данных, включающий в себя:

4.1. Выбор модели данных;

4.2. Выполнение обработки.

5. Интерпретация результатов, включающая:

5.1. Анализ результатов;

5.2. Принятие решений.

Из сказанного следует, что этапу обработки данных предшествует процесс шкалирования, т.е. представление значений свойств объектов исходного эмпирического множества в формализованной числовой форме, являющийся предметом изучения теории измерений [8, 9]. Применение процесса шкалирования приводит к получению из эмпирической таблицы

экспериментальных данных ТЭД, представленной в формализованной форме и готовой к вводу в память ЭВМ.

Из теории измерений известно, что для шкальных значений каждого из рассмотренных в п. I.1 типов признаков разрешены лишь определённые операции и преобразования, т. е. определённая шкала допускает вычисление ограниченного набора статистических характеристик. К примеру, для шкал наименований единственной координатой положения является мода, определяющая наиболее часто встречающееся в данной совокупности чисел значение. Характеристиками положения центра для качественных признаков являются мода и медиана, т.е. такое значение $\alpha$, для которого выполняется равенство:

$$P(x<\alpha) = P(x>\alpha),$$

где P - знак вероятности.

В случае количественных шкал используются арифметические операции, а подходящей мерой положения центра является среднее значение.

Обобщая сказанное можно сделать следующее заключение: поскольку каждая шкала обеспечивает вычисление определённого набора статистических характеристик, многие классические методы прикладной статистики оказываются неприменимыми для всего объёма разнотипной информации и используются только для обработки количественных данных.

В настоящее время развитие методов анализа разнотипных данных происходит в двух направлениях:

❖ Методы, позволяющие использовать признаки различных типов без сведения их к однотипным шкалам. Здесь можно выделить следующие исследования:

➢ подход, изложенный в [45], в котором вводится понятие “меры сходства” между исследуемыми объектами с учётом типа признака;

➢ использование класса логических решающих правил, которые строятся на основании элементарных высказываний и задаются на объектах эмпирическими отношениями, определёнными для различных признаков. При этом близость или сходство объектов заменяются понятием эквивалентности с точки зрения сходства значений логических функций [46].

❖ Методы, основанные на сведении различных типов признаков к однотипным. Эти методы можно разделить на следующие группы:

➢ использующие многомерные дискретные распределения, которые сводятся к поиску независимых распределений. Эти распределения аппроксимируют исходное дискретное распределение, образованное статистически независимыми дискретными переменными [47];

➢ основанные на представлении свойств средствами языка бинарных отношений, при котором строится матрица между объектами, называемая матрицей отношений. Степень проявления связи между объектами по заданному признаку характеризуется элементами такой матрицы. При этом осуществляется поиск

матриц отношений, наиболее точно описывающих исходную систему признаков [48];

- представляющие классификационные признаки в форме булевских матриц. Предполагается, что любое взаимнооднозначное преобразование может быть реализовано применением линейного оператора с соответствующей булевской матрицей [48]. В рамках такого подхода решаются в основном задачи описания и конструирования количественных факторов.

Недостатком вышеописанных методов является тот факт, что при переходе от качественных и количественных признаков к номинальным теряется информативность этих признаков. К примеру, диапазон измерения количественного признака, выбор которого неоднозначен, делится на интервалы группирования, соответствующие градациям нового классификационного признака. Для качественных признаков теряется отношение порядка.

Подход к анализу разнотипных данных, использующий методы преобразования качественных и номинальных признаков в количественные, основан на принципе оцифровки. Оцифровка неколичественных переменных проводится с целью их дальнейшего, наиболее эффективного использования наряду с имеющимися количественными признаками в классических статистических моделях [49-51]. Важным моментом в процедуре оцифровки является выбор критерия поиска числовых меток для градаций неколичественных переменных и методы его

оптимизации, подробно исследованные в [49]. Стохастические процедуры оцифровки признаков, измеренных в шкалах порядка, предложены в работе [52]. Обобщение результатов в этом направлении приведено в [18].

Среди недостатков методов оцифровки, препятствующих их широкому распространению, необходимо перечислить следующие:

- при использовании методов оцифровки не учитываются отношения, присущие исходным эмпирическим свойствам;
- методы оцифровки, описанные выше, не позволяют достаточно полно учитывать влияние количественных переменных, входящих в пространство описания наряду с качественными и номинальными (классификационными) признаками. Поэтому количественные переменные либо вовсе исключаются из рассмотрения – при этом нарушается адекватность описания исходных данных, либо сводятся к номинальному виду, вследствие чего теряется информативность признаков (переход от ‘сильной шкалы’ к ‘слабой шкале’);
- в результате необоснованного применения методов оцифровки дальнейшая обработка данных приводит к получению бессодержательной информации.

Устранению вышеизложенных недостатков посвящена работа [53], в которой процесс оцифровки рассматривается как самостоятельная исследовательская задача. Процедура перехода от менее “сильных” шкал к более “сильным” в рамках решения

определённой задачи анализа данных называется методом преобразования типов признаков и базируется на следующих основных положениях:

- последовательное усиление шкалы;
- при обогащении шкалы принимаются во внимание априорные предположения исследователя конкретной предметной области о скрытых отношениях, не отражённых в исходной системе признаков.

Перейдём к содержательной постановке задачи и изложению основных положений метода преобразования типов признаков в рамках [53]. Предварительно введём некоторые обозначения и сокращения, а именно:

- э.с.о. – эмпирическая система с отношениями;
- ч.с.о. – числовая система с отношениями;
- $C=\langle C, V\rangle$ - э.с.о.;
- $B = \langle B^1, U\rangle$ - ч.с.о. ;
- $C$ – эмпирическое множество;
- $B^1$ – числовое множество;
- $V$ – отношение, заданное в эмпирическом множестве;
- $U$ – отношение, заданное в числовом множестве;
- $c_i$ – элементы множества $C$, соответствующие градациям признака, который измерен в шкале $\Psi$;
- $g$ – число градаций неколичественного признака;
- $\{\xi\}$ – множество всех изоморфных отображений $\xi: C \to B$, сопоставляющих каждой градации $c_i \in C$ некоторое число $a_i \in \xi(c_i)$

из $B^1$, называемое числовой меткой градации неколичественного признака;

○ $B^* = \langle B^1, U^* \rangle$ ч.с.о. более богатая, чем B, $U \subset U^*$.

Для $B^* = \langle B^1, U^* \rangle$ можно рассмотреть э.с.о. $C^* = \langle C, \xi^{-1}U^* \rangle$, причём $V \subset \xi^{-1}U^*$. Тогда взаимнооднозначное отображение $\xi$ будет отвечать условием изоморфизма э.с.о. $C^*$ на ч.с.о. $B^*$.

Определение. Изоморфизм $\xi$ э.с.о. $C^*$ в ч.с.о. $B^*$ будем называть императивной шкалой, порождённой неколичественной шкалой $\boldsymbol{\Psi}$ при расширении U до $U^*$.

Таким образом, $\{\xi\}$ можно рассматривать как множество шкал более сильных, чем $\boldsymbol{\Psi}$.

Если имеется некоторый признак *x*, то через $x^{\xi}$ обозначим признак, измеренный в императивной шкале $\xi$, порождённой $\boldsymbol{\Psi}$. Преобразование типа признака означает переход от признака *x*, измеренного в шкале одного типа, к признаку $x^{\xi}$, измеренному в шкале другого типа.

В общем случае критерий, на основе которого производится присвоение числовых меток, может быть представлен в виде некоторой функции $F$: $\{\xi\} \to B^1$. Выбирается та шкала

$\xi^*$, для которой $F(\xi^*) = \min_{\xi \in \{\xi\}} F(\xi)$.

Императивные шкалы могут быть как качественными, так и количественными. В первом случае для системы $B^*$ вместо $B^1$ может быть выбрано любое линейно-упорядоченное множество, например отрезок натурального ряда чисел $\boldsymbol{K}_g = \{1, 2, \ldots, g\}$.

Отображение ξ будет присваивать каждой *i*-й градации значение целочисленной метки $d_i$ из $\boldsymbol{K}_g$, называемое рангом градации. Тогда отображение ξ будет задаваться перестановками рангов $\boldsymbol{D} \in \boldsymbol{P}^{\boldsymbol{g}}$, где $\boldsymbol{D}=(d_1, d_2, \ldots, d_g)$, $\boldsymbol{P}^{\boldsymbol{g}}$ – множество перестановок g натуральных чисел. В таком случае говорят о порядковой императивной шкале.

В общем случае задача поиска оптимальных императивных шкал при обработке разнотипных данных, представленной матрицей $\mathbf{X}=[\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]$ размерности $n * p$, где $\mathbf{X}_1$ – подматрица, имеющая $l$ количественных переменных $x^{(1)}, \ldots, x^{(l)}$, $\mathbf{X}_2$ – имеет качественные признаки с индексами $l+1, \ldots, m$, $\mathbf{X}_3$ – подматрица из **X**, имеющая только классификационные признаки с индексами $m+1, \ldots, p$, может быть рассмотрена как задача минимизации функции многих переменных:

для количественных императивных шкал

$$F(\mathbf{X}_1, \bar{a}_{l+1}, \ldots, \bar{a}_p) \to \min, \qquad \text{(I.3)}$$

где $\bar{a}_{l+1} \in B^{gl+1}, \ldots, \bar{a}_p \in B^{gp}$;

для императивных шкал порядка

$$F(\mathbf{X}_1, \mathbf{X}_2, D_{m+1}, \ldots, D_p) \to \min, \qquad \text{(I.4)}$$

где $D_{m+1} \in \boldsymbol{P}^{\boldsymbol{gm+1}}, \ldots, D_p \in \boldsymbol{P}^{\boldsymbol{gp}}$.

В [53] предлагается обобщённая формула для выбора императивных порядковых шкал в виде:

$$\frac{\bar{\mathbf{d}}^{\mathrm{T}} \mathbf{V} \bar{\mathbf{d}}}{\bar{\mathbf{d}}^{\mathrm{T}} \mathbf{Z} \bar{\mathbf{d}}} \to \min, \qquad \text{(I.5)}$$

где $d \in \boldsymbol{P}^{\boldsymbol{g}}$, $\boldsymbol{P}^{\boldsymbol{g}}$ - множество перестановок g натуральных чисел;

$$\frac{\bar{a}^{\mathrm{T}}\mathbf{V}\bar{a}}{\bar{a}^{\mathrm{T}}\mathbf{Z}\bar{a}} \rightarrow \min, \qquad \text{(I.6)}$$

где $\bar{a} \in B^g$.

Конкретные выражения для матриц **V** и **Z** зависят от типа решаемой задачи.

При условии, когда g>6, общее число возможных перестановок из g целочисленных рангов оказывается довольно большим (g!) и применение метода полного перебора для получения оптимальной императивной шкалы, удовлетворяющей критерию (I.5), становится неэффективным вследствие огромных затрат машинного времени. Поэтому оптимальную перестановку рангов в [53] предлагается искать с использованием метода последовательных приближений. При этом начальная перестановка $\bar{\mathbf{d}}_0$ выбирается произвольным способом и конечный результат зависит именно от $\bar{\mathbf{d}}_0$. Таким образом для различных начальных значений получаются различные локальные минимумы (I.5).

При оптимизации функций (I.5) и (I.6) вычисляется обратная ковариационная матрица и определитель ковариационной матрицы. Как известно, при общем числе признаков p, принимающем значения больше восьми, возникает сложность при вычислении обратной ковариационной матрицы и определителя. Это связано с тем обстоятельством, что при вычислениях получаются очень близкие к нулю значения, которые ЭВМ принимает как машинный нуль. Вследствие этого искажаются

конечные результаты оцифровки и, соответственно, конечные результаты обработки данных.

Из вышеизложенного, а также исходя из [86], где даётся критический анализ современного состояния прикладной статистики и обсуждаются тенденции развития статистических методов, следует необходимость дальнейшего развития методов преобразования типов признаков.

## I.3. Цель и задачи работы

Целью работы является: разработка методов преобразования типов признаков в задачах анализа данных разнотипной природы, создание системы анализа данных с использованием как классических методов обработки данных, так и предложенных в данной работе методов и алгоритмов.

Для достижения поставленной цели в работе решаются следующие основные задачи:

- разработка методов и алгоритмов поиска целочисленных числовых меток градаций признаков нечисловой природы;
- программная реализация разработанных методов и алгоритмов;
- создание на основе классических методов анализа данных системы обработки экспериментальной информации с подключением модулей, реализующих разработанные в данной работе методы и алгоритмы преобразования типов признаков.

## Выводы по главе I.

1. Проведён анализ современного состояния методов обработки нечисловой информации. Показано, что эти методы обладают рядом недостатков: потеря информативности при переходе от количественных признаков к неколичественным, сложность алгоритмической реализации, искажение конечных результатов обработки больших массивов информации. Предлагаются пути устранения этих недостатков и дальнейшего развития методов обработки нечисловой информации. Задача анализа данных разнотипной природы рассмотрена как самостоятельная и названа преобразованием типов признаков.

2. Существующее на сегодняшний день программное обеспечение классифицировано на типы. Утверждается, что современная тенденция развития программных средств прикладной статистики заключается в создании экспертных систем анализа данных, ориентированных как опытных пользователей, так и на новичков.

3. Сформулированы цель и задачи работы.

# ГЛАВА II. Разработка методов и алгоритмов анализа данных разнотипной природы

## II.1. Разработка алгоритмов преобразования типов признаков для задач исследования парных взаимозависимостей

В данном параграфе будут рассмотрены алгоритмы преобразования типов признаков для задач исследования взаимозависимостей признаков.

Предварительно исходную матрицу **X** из (I.1) представим в виде матрицы

$$\mathbf{X}=[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}] \qquad \text{(II.1)}$$

размерностью N**p*, где $\mathbf{X_1}$ – подматрица, имеющая *l* количественных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(l)}$, $\mathbf{X_2}$ – имеет только качественные признаки с индексами $i=\overline{l+1,m}$, и $\mathbf{X_3}$ – подматрица из **X**, имеющая только классификационные признаки с индексами $i=\overline{m+1,p}$. Будем рассматривать задачу поиска императивных шкал порядка как задачу минимизации функции многих переменных, представленную формулой (I.4). В соответствие с изложенным в п. I.2, в этом случае исследуется качественная императивная шкала и в качестве числового множества $B^1$ выбирается любое линейно-упорядоченное множество, например, отрезок натурального ряда чисел $\boldsymbol{K}_g=\{1, 2, \ldots, g\}$. При этом отображение ξ будет присваивать каждой *i*-й градации значение целочисленной метки $d_i$ из $\boldsymbol{K}_g$, называемое рангом градации. Тогда отображение ξ

будет задаваться перестановками рангов $\boldsymbol{D} \in \boldsymbol{P^g}$, где $\boldsymbol{D}=(d_1, d_2, \dots, d_g)$,

$\boldsymbol{P^g}$ – множество перестановок g натуральных чисел.

Применим обобщённую формулу критерия поиска императивных шкал (I.5) для задачи исследования парных взаимозависимостей признаков.

Пусть исходная ТЭД представлена в виде матрицы (II.1). Анализ взаимозависимостей признаков заключается в вычислении и интерпретации матрицы $\boldsymbol{R}=(r_{ij})$ размерностью $p \times p$, элементами которой являются выборочные оценки показателей парной взаимосвязи признаков определённого типа. На рис. II.1 показано, как формируется $\boldsymbol{R}$ при наличии в ТЭД разнотипных признаков. Подматрицы $\boldsymbol{R_{11}}, \boldsymbol{R_{22}}, \boldsymbol{R_{33}}$ состоят из мер связи соответственно для количественных, качественных и классификационных признаков, элементы же подматриц $\boldsymbol{R_{12}}$ $\boldsymbol{(R_{21})}$, $\boldsymbol{R_{13}}$ $\boldsymbol{(R_{31})}$, $\boldsymbol{R_{23}}$ $\boldsymbol{(R_{32})}$ характеризуют взаимосвязь признаков различного типа.

Необходимо отметить, что для признаков каждого типа существует много различных показателей взаимосвязи. В работе [54] делается вывод, что большинство методов измерения связей основано либо на принципе ковариации, либо на принципе взаимной сопряжённости. В случае использования принципа ковариации заключение о наличии связи

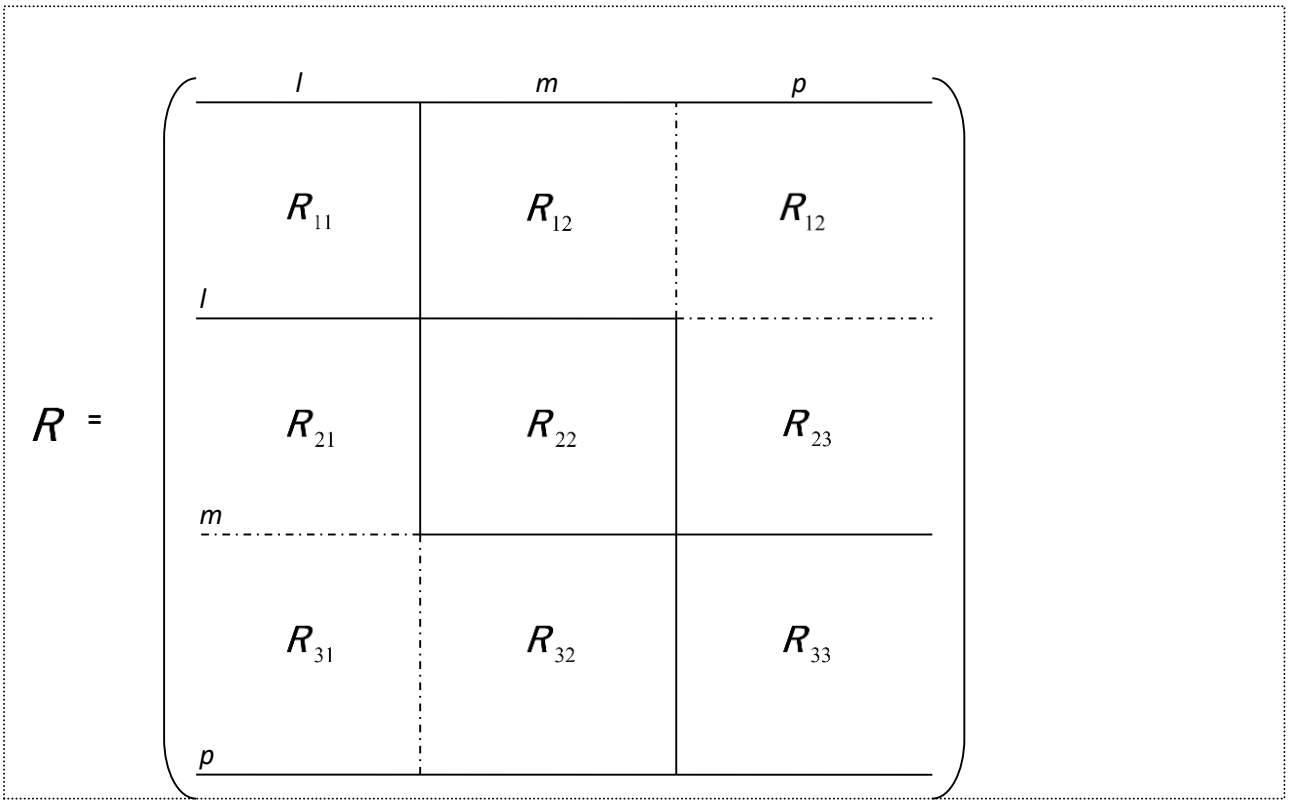


Рис. II.1. Матрица показателей парной взаимосвязи разнотипных признаков.

между признаками делается тогда, когда увеличение численных значений одного признака сопровождается устойчивым увеличением или уменьшением другого. В математическом отношении задача сводится к вычислению величины ковариации, т.е. сопутствующего изменения численных значений признаков, и последующего нормирования этой величины. В нашем случае по данному принципу построены подматрицы $\boldsymbol{R}_{11}$ и $\boldsymbol{R}_{22}$ для количественных и качественных признаков. В качестве меры связи можно рассматривать как обычный коэффициент корреляции, так и ранговые коэффициенты корреляции (ρ-коэффициенты) Кендала, Спирмана, Стьюарта [55, 56]. Формула обобщённого коэффициента корреляции для вычисления как обычных, так и

ранговых коэффициентов приводится в [57]. В [54] отмечается возможность использования обычного коэффициента корреляции как унифицированной меры связи для количественных и качественных признаков. В связи с этим элементы подматрицы $\boldsymbol{R}_{22}$ рассчитываются по формуле вычисления обычного коэффициента корреляции по значениям рангов.

Меры связи классификационных признаков, формирующих элементы подматрицы $\boldsymbol{R}_{33}$, построены на принципе взаимной сопряжённости ($\chi^2$-коэффициенты). К таковым относятся коэффициенты Пирсона, Крамера, Чупрова, показатели Q и Ф [18, 55, 58].

Для вычисления мер связи между разнотипными признаками будем также использовать обычный коэффициент корреляции. Это делается в целях усиления слабой шкалы и предотвращения потери информации [59, 60].

Предположим, что признак $x_1$ – количественный, а признак $x_2$ – классификационный. Требуется выявить не отражённые неколичественной шкалой отношения, которые бы позволили говорить о согласованном изменении значений рассматриваемых признаков. Другими словами, требуется определить императивную шкалу, порождённую исходной неколичественной шкалой и позволяющей преобразовать неколичественный признак $x_2$ в количественный $x_2^{\xi}$. Для решения данного вопроса рассмотрим множество $\{\xi\}$ императивных шкал, порождённых исходной неколичественной шкалой. Тогда значение

коэффициента корреляции $r_{12}$, вычисленное для $x_1$ и $x_2^{\xi}$ и измеренное в некоторой шкале ξ, будет характеризовать степень тесноты линейной зависимости между $x_1$ и $x_2^{\xi}$, которую обеспечивает данная императивная шкала. Коэффициент корреляции является мерой точности прогноза количественного признака [56] и поэтому $r_{12}$ оценивает качество прогноза значений $x_1$, производимого по числовым меткам $\xi(c_i)$ с использованием регрессии

$$x_1 = ax_2^{\xi} + b. \qquad \text{(II.2)}$$

Нас будет интересовать императивная шкала $\xi^*$, при которой достигается максимальное значение коэффициента корреляции $r_{12}$ и тем самым обеспечивается наиболее точный прогноз значений признака по линии регрессии (II.2). Из сказанного можно сделать вывод, что при анализе парной взаимозависимости разнотипных признаков критерием выбора количественной императивной шкалы является максимизация $r_{12}$:

$$1 - r_{12} \to \min. \qquad \text{(II.3)}$$

Формулу (II.3) применительно к задаче оптимизации (I.4) можно записать в виде:

$$1 - r_{12}^2(x_1, D) \to \min,\ D \in P^g, \qquad \text{(II.4)}$$

где $x_1$ - значения признака, измеренного в количественной или ранговой шкале.

Обобщая сказанное на многомерный случай, можно сделать вывод о том, что использование метода преобразования типов признаков для рассматриваемой задачи анализа данных позволит

применить обычные коэффициенты корреляции для вычисления показателей взаимосвязи в подматрицах $\boldsymbol{R_{12}}$ ***(R21)***, $\boldsymbol{R_{13}}$ ***(R31)***, $\boldsymbol{R_{23}}$ ***(R32)***. Далее коэффициенты корреляции из ***R***, вычисленные для признаков различного типа с использованием императивных шкал, могут быть применены и при построении многомерных моделей.

Перейдём теперь к аналитическому преобразованию критерия (II.3) для построения алгоритма преобразования типов признаков. Для этого рассмотрим случай, когда ТЭД размерностью N**(l+1)* представлена $l$ количественными признаками$x_i$, $i = \overline{1,l}$ и неколичественным признаком $x_{l+1}$ с количеством градаций, равным $g$.

Пусть $\bar{a}^{\mathrm{T}} = (a_1, \ldots, a_g)$ некоторый набор числовых меток для признака$x_{l+1}$; $n_i$- количество объектов, обладающих *i*-й градацией признака$x_{l+1}$; $m_k$ - выборочное среднее *k*-го признака, $k = \overline{1,l}$; $m_k^i$ - частное выборочное среднее *k*-го признака в *i*-й градации *l+1*-го признака. Используя выражения, полученные в [61], вычислим элементы *l+1*-й строки и *l+1*-го столбца матрицы ковариаций $\underline{\Lambda}$**:**

$$\lambda_{k,l+1} = \frac{1}{N}\sum_{i=1}^{g} a_i\, n_i (m_k^i - m_k), k = \overline{1,l};$$

$$\lambda_{l+1,l+1} = \frac{1}{N}\sum_{i=1}^{g} a_i\, n_i \left(a_i - \frac{1}{N}\sum_{i=1}^{g} a_i\, n_i\right).$$

Пусть $\mu_k^i = \frac{n_i}{N}(m_k^i - m_k)$. Тогда получим

$$\lambda_{k,l+1} = \lambda_{l+1,k} = \sum_{i=1}^{g} a_i \mu_k^i. \qquad \text{(II.5)}$$

Обозначим через $\boldsymbol{T}=[t_{ij}]$, $i, j=\overline{1,g}$, где

$$t_{ij}=\begin{cases} -\frac{n_i n_j}{N}, i \neq j, \\ \frac{n_i}{N^2}(N-n_i), i=j. \end{cases}$$

Тогда в матричной форме для $\lambda_{l+1,l+1}$ справедливо равенство:

$$\lambda_{l+1,l+1}=\bar{a}^T \boldsymbol{T}\, \bar{a}. \quad \text{(II.6)}$$

Обозначим через $\underline{\Lambda}_{ij}$ - подматрицу матрицы $\underline{\Lambda}$, полученную удалением из $\underline{\Lambda}$ $i$-й строки и $j$-го столбца; $\underline{\Lambda}_{ij,\,pq}$ - удалением $i$-й и $p$-й строк, $i$-го и $q$-го столбцов; $\Lambda_{ij}$ - дополнительный минор элемента $\lambda_{ij}$ в определителе $\Lambda$; $\Lambda_{ij,pq}$ - определитель матрицы $\underline{\Lambda}_{ij,pq}$. Предполагая, что $\operatorname{rang}(\Lambda_{l+1,l+1})=l$, определим матрицы:

$\boldsymbol{M}=[\mu_k^i]$ размерностью $g \times l$,

$$\boldsymbol{C}=\boldsymbol{M}\underline{\Lambda}_{l+1,l+1}^{-1}\boldsymbol{M}^T, \quad \text{(II.7)}$$

$$\boldsymbol{S}=\boldsymbol{T}-\boldsymbol{C}. \quad \text{(II.8)}$$

В [61] доказано равенство:

$$\Lambda=\Lambda_{l+1,l+1}\bar{a}^T \boldsymbol{S}\, \bar{a}, \quad \text{(II.9)}$$

которое является искомым представлением определителя в виде явной зависимости от числовых меток. Здесь определитель матрицы $\underline{\Lambda}$, включающий оцифрованную переменную, вычисляется на основании матрицы ковариаций $\Lambda_{l+1,l+1}$ только для количественных переменных. Это позволяет получить такие

критерии поиска императивных шкал, в которых влияние количественных признаков учитывается естественным образом без их предварительного сведения к номинальным.

Можно показать, что матрицы ***T***, ***C*** и ***S*** вырождены. Вычислим значение квадратичной формы $\bar{a}^T \boldsymbol{S}\, \bar{a}$:

$$\Lambda = \sigma_1^2(1-r_{12}^2)\ \ldots\ \sigma_l^2(1-r_{l,1\ldots l-1}^2)\,\sigma_{l+1}^2(1-r_{l+1,1\ldots l}^2) = \Lambda_{l+1,l+1}\,\sigma_{l+1}^2(1-r_{l+1,1\ldots l}^2),$$

откуда

$$\bar{a}^T \boldsymbol{S}\, \bar{a} = \sigma_{l+1}^2\,(1-r_{l+1,1\ldots l}^2). \qquad \text{(II.10)}$$

Перейдём к получению выражения для функций критериев выбора императивных шкал для задачи анализа взаимозависимостей признаков. В этой задаче анализа данных для преобразования типов признаков предлагается использовать в качестве критерия выбора императивных шкал минимизацию величины $1\text{-}r_{12}^2$, где $r_{12}^2$ - коэффициент корреляции между количественным признаком $x_1$ и оцифрованным неколичественным признаком $x_2^{\xi}$. Используя (II.5) и (II.6) для $r_{12}$, можно получить следующее соотношение:

$$r_{12}^2 = \frac{\bar{a}^T \boldsymbol{C} \bar{a}}{\bar{a}^T \boldsymbol{T} \bar{a}},$$

где $\boldsymbol{C} = \bar{\mu}\lambda_{11}^{-1}\bar{\mu}^T$, $\bar{\mu}^T = [\mu_1^1,\ \mu_1^2,\ \ldots,\ \mu_1^g]$. Из этих соотношений получим:

$$1\text{-}r_{12}^2 = \frac{\bar{a}^T \boldsymbol{H} \bar{a}}{\bar{a}^T \boldsymbol{T} \bar{a}}, \qquad \text{(II.11)}$$

где ***H=T-C.***

Таким образом критерий поиска числовых меток при исследовании парных зависимостей имеет вид

$$\frac{\bar{a}^T \boldsymbol{H} \bar{a}}{\bar{a}^T \boldsymbol{T} \bar{a}} \to \min, \qquad \text{(II.12)}$$

$\bar{a} \in \mathrm{B}^g$.

Перейдём к решению оптимизационной задачи (II.12) при условии, что необходимо найти оптимальные целочисленные числовые метки. В этом случае задача (II.12) примет вид:

$$\frac{\bar{D}^T \boldsymbol{H} \bar{D}}{\bar{D}^T \boldsymbol{T} \bar{D}} \to \min, \qquad \text{(II.13)}$$

где $D \in P^g$.

Вначале предположим, что число g градаций классификационного признака $x_{l+1}$ не превышает шести, g≤6. В этом случае каждой ранговой императивной шкале будет соответствовать перестановка из g целочисленных рангов $\bar{D}^T = (d_1, \ldots, d_g)$, $d_i \in \mathrm{K_g}$. Так как количество всех перестановок, равное g!, для рассматриваемого случая сравнительно невелико, то оптимальное значение критерия (II.13) предлагается искать методом полного перебора всех перестановок. Определим в этом случае способ получения всех перестановок из заданного числа градаций классификационного признака $x_{l+1}$, равного g.

Зададимся начальным числом $k_1$=1. Добавляя к этому значению второе число $k_2$=2 сначало на первое место, потом на второе, получим 2! Перестановок, т.е. {2,1}, {1,2}. На третьем шаге получим 3! перестановок, на шаге с номером g получим искомые g! перестановок.

Вычислительный алгоритм поиска оптимальной перестановки целочисленных рангов градаций при $g \le 6$, используемый при определении императивной шкалы порядка, состоит из следующих шагов:

Шаг 1. Определение всех $g!$ перестановок по предложенной выше схеме.

Шаг2. Выбор начальной перестановки градаций $\overline{D}_0=(1, 2, \ldots, g)$ и вычисление значения $F(\overline{D}_0)=\frac{\overline{D}_0^T \boldsymbol{H} \overline{D}_0}{\overline{D}_0^T \boldsymbol{T} \overline{D}_0}$; $F_{min}=F(\overline{D}_0)$.

Шаг 3. Переход к следующей перестановке $\overline{D}_1$. Вычисление значения $F(\overline{D}_1)$. Если $F(\overline{D}_1)<F_{min}$, то $F_{min}= F(\overline{D}_1)$.

Шаг 4. Если просмотрены все $g!$ перестановок, то переход к шагу 5. Иначе переход к шагу 3.

Шаг 5. Вектор$\overline{D}_k$, соответствующий минимальному значению $F_{min}$ и есть искомая императивная шкала порядка.

Рассмотрим теперь случай, когда $g>6$. При этом условии общее число возможных перестановок из $g$ целочисленных рангов оказывается довольно большим ($g!$) и применение метода полного перебора для получения оптимальной императивной шкалы, удовлетворяющей критерию (II.13), становится неэффективным вследствие огромных затрат машинного времени. Поэтому оптимальную перестановку рангов в [53] предлагается искать с использованием метода последовательных приближений. При этом начальная перестановка $\overline{D}_0$ выбирается произвольным способом и конечный результат зависит именно от $\overline{D}_0$. Таким

образом для различных начальных значений получаются различные локальные минимумы (II.13).

Для устранения этого недостатка предлагается поступать следующим образом. Определим прежде всего способ вычисления начального приближения $\overline{D}_0$, для чего решим вспомогательную задачу.

Предположим, что минимум (II.13) необходимо найти на множестве действительных чисел, т.е. расширим множество допустимых решений до $\mathrm{B}^g$ и задача поиска оптимального вектора будет иметь вид (II.12).

Отношение (II.12) называется обобщённым отношением Релея. Известно [62], что для обычного отношения Релея вида

$$\frac{(\mathbf{A}x,x)}{(x,x)} \tag{II.14}$$

справедливы неравенства

$$v_1 \leq \frac{(\mathbf{A}x,x)}{(x,x)} \leq v_n \ ,$$

а также соотношение

$$v_1 = \min_{x \neq 0} \frac{(\mathbf{A}x,x)}{(x,x)} \tag{II.15}$$

где $x$ - любой ненулевой вектор, $v_1, \ldots, v_n$ - собственные числа матрицы $\mathbf{A}$, а $v_1$ - минимальное из этих чисел.

Зададимся условием центрированности вектора $\bar{a}$, т.е.

$$\frac{1}{N}\sum_{i=1}^{g} a_i \, n_i = 0.$$

В этом случае матрица $\boldsymbol{T}$ для задач анализа взаимозависимости признаков примет диагональный вид с элементами $n_i/N$ и поэтому будет невырожденной. Тогда отношение (II.12) можно свести к виду (II.14), так как

$$\frac{\bar{a}^T \boldsymbol{H} \bar{a}}{\bar{a}^T \boldsymbol{T} \bar{a}} = \frac{(\boldsymbol{H}^T \bar{a}, \bar{a})}{(\boldsymbol{T}^T \bar{a}, \bar{a})} = \frac{(\boldsymbol{T}^{T^{-1/2}} \boldsymbol{H}^T \boldsymbol{T}^{T^{-1/2}} \bar{b}, \bar{b})}{(\bar{b}, \bar{b})},$$

где $\bar{b} = \mathrm{T}^{T^{-1/2}} \bar{a}$.

Обозначим $\boldsymbol{B}$=$\mathrm{T}^{T^{-1/2}} H^T \mathrm{T}^{T^{-1/2}}$. Тогда имеем

$$\frac{\bar{a}^T \boldsymbol{H} \bar{a}}{\bar{a}^T \boldsymbol{T} \bar{a}} = \frac{(\boldsymbol{B}\bar{b}, \bar{b})}{(\bar{b}, \bar{b})}.$$

Используя для данного равенства соотношение (II.15), можно найти собственный вектор $\overline{b_0}$, соответствующий минимальному собственному числу матрицы $\boldsymbol{B}$. Далее, искомый вектор вычисляется следующим образом:

$$\bar{a} = B^{-1/2} \bar{b}.$$

Однако, процесс нахождения матрицы $B^{-1/2}$ является довольно громоздким, поэтому введём ещё одно дополнительное ограничение:

$$\bar{a}^T T \bar{a} = 1. \qquad \text{(II.16)}$$

Тогда задача минимизации (II.12) сведётся к поиску условного минимума квадратичной формы

$$\bar{a}^T H \bar{a} \to \min, \qquad \text{(II.17)}$$

при условии (II.16).

Для решения задачи (II.17) используем метод множителей Лагранжа, по которому оптимальный вектор $\bar{a}$ обращает в нуль первую производную функции Лагранжа

$$F(\bar{a}) = \bar{a}^T H\bar{a} - v\,\bar{a}^T T\bar{a},$$

где $v$ - вводимый множитель Лагранжа.

Имеем:

$$\frac{\partial F(\bar{a})}{\partial \bar{a}} = 2\,\boldsymbol{H}\bar{a} - 2v\,\boldsymbol{T}\bar{a} = 0,$$

или получим систему линейных однородных уравнений

$$\boldsymbol{H}\bar{a} - v\,\boldsymbol{T}\bar{a} = 0. \qquad \text{(II.18)}$$

Умножим это уравнение слева на $\bar{a}$ и, учитывая (II.16), получим

$$\bar{a}^T H\bar{a} = v, \qquad \text{(II.19)}$$

т.е. значение (II.12) определяется введённым множителем Лагранжа при условии (II.16). Следовательно, выбрав в качестве $v$ минимальное собственное число и решив систему уравнений (II.18), получим вектор $\bar{a}$, минимизирующий (II.12).

Умножив (II.18) слева на $T^{-1}$, получим

$$T^{-1}\,\boldsymbol{H}\bar{a} = v\,\bar{a},$$

т.е. в этом случае искомое оптимальное решение (II.12) определяется собственным вектором матрицы $\boldsymbol{T}^{-1}\,\boldsymbol{H}$, соответствующим минимальному собственному числу $v$.

Обозначим через $\bar{a}_0$ найденный оптимальный вектор, минимизирующий (II.12).

Упорядочим компоненты вектора $\bar{a}_0$ по возрастанию, т.е.

$$a_{i_1} \le a_{i_2} \le \ldots a_{i_g},$$

где $i_j$ - индекс, указывающий на место компонента $a_{i_j}$ вектора $\bar{a}_0$, принимает значения из множества $\{1,2, \ldots, g\}$.

Присвоим компонентам вектора $\overline{D}_0$ целочисленные значения по следующей схеме:

$$d_{i_1} = a_{i_1} = 1;\ d_{i_2} = a_{i_2} = 2;\ \ldots\ d_{i_{1g}} = a_{i_g} = g;$$

Получим в итоге некоторый вектор

$$\overline{D}_0=(d_1,\ldots,d_g),$$

который будем рассматривать в качестве начальной перестановки рангов, используемой при определении императивной шкалы порядка методом последовательных приближений, суть которого излагается ниже.

Обозначим числитель и знаменатель отношения (II.13) через $M_1$ и $M_2$ соответственно. При транспозиции $k$-й и $l$-й градаций приращение числителя и знаменателя вычисляется по формулам [53]:

где $\begin{cases} \Delta M_1=\bar{c}\bar{d},\ \bar{c}=(c_1,\ldots,c_g), \\ 2(d_k-d_l)(h_{il}-h_{ik}),\ i\neq k,l; \end{cases}$

$$c_i = (d_k - d_l)(h_{ll} - h_{kk}),\ i = k,l;$$

$$\Delta M_2 = \overline{f}\,\overline{d},\ \overline{f} = (f_1,\ldots,f_g),$$

где $f_i = \begin{cases} 2(d_k - d_l)(t_{il} - t_{ik}),\ i \neq k,l; \\ (d_k - d_l)(t_{ll} - t_{kk}),\ i = k,l; \end{cases}$

Очевидно, что при условии

$$\frac{1+\Delta M_1 / M_1}{1+\Delta M_2 / M_2} < 1 \qquad \text{(II.20)}$$

транспозиция рангов $k$-й и $l$-й градаций уменьшит значение (II.13).

Опишем теперь вычислительный алгоритм поиска оптимальной перестановки целочисленных рангов градаций, минимизирующей отношение (II.13) для случая g>6:

Шаг 1. Определение начальной перестановки рангов градаций $\overline{D}_0$ по способу, описанному выше.

Шаг 2. Выбор некоторой отличной от предыдущей итерации пары градаций $k$ и $l$.

Шаг 3. Определение $\Delta M_1$ и $\Delta M_2$ при транспозиции рангов $d_k$ и $d_l$. Вычисление значения отношения (II.20).

Шаг 4. При выполнении неравенства (II.20) получаем новую перестановку $\overline{D}_1$. В противном случае – переход к шагу 5.

Шаг 5. Проверка условия остановки. Выход осуществляется в том случае, если транспозиция любой пары рангов полученного упорядочения не приводит к уменьшению (II.13). Иначе – переход к шагу 2.

Таким образом, описанный выше алгоритм позволяет найти оптимальную перестановку рангов $\overline{D}$, минимизирующую отношение (II.13) для признака $x^{l+1}$. В случае наличия нескольких неколичественных признаков те же операции проделываются с признаками $x^{l+2}, \ldots, x^{p}$.

## II.2. Разработка алгоритмов преобразования типов признаков для задачи распознавания образов

Вначале остановимся на принципах классификации методов распознавания образов.

Распознаванием образов называются задачи построения и применения формальных операций над числовыми или символьными отображениями объектов реального или идеального мира, результаты решения которых отражают отношения эквивалентности между этими объектами. Отношения эквивалентности выражают принадлежность оцениваемых объектов к каким-либо классам, рассматриваемым как самостоятельные семантические единицы.

При построении алгоритмов распознавания классы эквивалентности могут задаваться исследователем, который пользуется собственными содержательными представлениями или использует внешнюю дополнительную информацию о сходстве и различии объектов в контексте решаемой задачи. Тогда говорят о “распознавании с учителем” [63-65]. В противном случае, т.е. когда автоматизированная система решает задачу классификации без привлечения внешней обучающей информации, говорят об автоматической классификации или “распознавании без учителя”. Большинство алгоритмов распознавания образов требует привлечения весьма значительных вычислительных мощностей, которые могут быть обеспечены только высокопроизводительной компьютерной техникой.

Различные авторы дают различную типологию методов распознавания образов. Одни авторы различают параметрические,

непараметрические и эвристические методы, другие - выделяют группы методов, исходя из исторически сложившихся школ и направлений в данной области. Например, существует следующая типология методов распознавания образов:

- методы, основанные на принципе разделения;
- статистические методы;
- методы, построенные на основе “потенциальных функций”;
- методы вычисления оценок (голосования);
- методы, основанные на исчислении высказываний, в частности на аппарате алгебры логики.

Подобная типология методов распознавания с той или иной степенью детализации встречается во многих работах по распознаванию. В то же время известные типологии не учитывают одну очень существенную характеристику, которая отражает специфику способа представления знаний о предметной области с помощью какого-либо формального алгоритма распознавания образов. Д.А.Поспелов (1990) выделяет два основных способа представления знаний [70]:

1. Интенсиональное представление - в виде схемы связей между атрибутами (признаками).

2. Экстенсиональное представление - с помощью конкретных фактов (объекты, примеры).

Интенсиональное представление фиксируют закономерности и связи, которыми объясняется структура данных. Применительно

к диагностическим задачам такая фиксация заключается в определении операций над атрибутами (признаками) объектов, приводящих к требуемому диагностическому результату. Интенсиональные представления реализуются посредством операций над значениями атрибутов и не предполагают произведения операций над конкретными информационными фактами (объектами).

В свою очередь, экстенсиональные представления знаний связаны с описанием и фиксацией конкретных объектов из предметной области и реализуются в операциях, элементами которых служат объекты как целостные системы.

Описанные выше два фундаментальных способа представления знаний позволяют предложить следующую классификацию методов распознавания образов:

- Интенсиональные методы распознавания образов - методы, основанные на операциях с признаками.
- Экстенсиональные методы распознавания образов - методы, основанные на операциях с объектами.

В приводимой ниже классификации основное внимание уделено формальным методам распознавания образов и поэтому опущено рассмотрение эвристического подхода к распознаванию, получившего полное и адекватное развитие в экспертных системах. По поводу этого подхода ограничимся лишь несколькими замечаниями.

Эвристический подход основывается на трудно формализуемых знаниях и интуиции исследователя. В этом подходе исследователь сам определяет, какую информацию и каким образом нужно использовать для достижения требуемого эффекта распознавания.

Интенсиональные методы.

Отличительной особенностью интенсиональных методов является то, что в качестве элементов операций при построении и применении алгоритмов распознавания образов они используют различные характеристики признаков и их связей. Такими элементами могут быть отдельные значения или интервалы значений признаков, средние величины и дисперсии, матрицы связи признаков и т. п., над которыми производятся действия, выражаемые в аналитической или конструктивной форме. При этом объекты в данных методах не рассматриваются как целостные информационные единицы, а выступают в роли индикаторов для оценки взаимодействия и поведения своих атрибутов.

Группа интенсиональных методов распознавания образов обширна, и ее деление на подклассы носит в определенной мере условный характер.

Методы, основанные на оценках плотностей распределения значений признаков.

Эти методы распознавания образов заимствованы из классической теории статистических решений, в которой объекты

исследования рассматриваются как реализации многомерной случайной величины, распределенной в пространстве признаков по какому-либо закону. Они базируются на байесовской схеме принятия решений, апеллирующей к априорным вероятностям принадлежности объектов к тому или иному распознаваемому классу и условным плотностям распределения значений вектора признаков. Данные методы сводятся к определению отношения правдоподобия в различных областях многомерного пространства признаков.

Группа методов, основанных на оценке плотностей распределения значений признаков имеет прямое отношение к методам дискриминантного анализа. Байесовский подход к принятию решений и относится к наиболее разработанным в современной статистике так называемым параметрическим методам, для которых считается известным аналитическое выражение закона распределения (в данном случае нормальный закон) и требуется оценить лишь небольшое количество параметров (векторы средних значений и ковариационные матрицы).

К этой группе относится и метод вычисления отношения правдоподобия для независимых признаков. Этот метод, за исключением предположения о независимости признаков (которое в действительности практически никогда не выполняется), не предполагает знания функционального вида

закона распределения. Поэтому его можно отнести к непараметрическим.

Другие непараметрические методы, применяемые тогда, когда вид кривой плотности распределения неизвестен и нельзя сделать вообще никаких предположений о ее характере, занимают особое положение. К ним относятся известные метод многомерных гистограмм, метод “k-ближайших соседей, метод евклидова расстояния, метод потенциальных функций и др., обобщением которых является метод, получивший название “оценки Парзена” [70]. Эти методы формально оперируют объектами как целостными структурами, но в зависимости от типа задачи распознавания могут выступать и в интенсиональной и в экстенсиональной ипостасях.

Непараметрические методы анализируют относительные количества объектов, попадающих в заданные многомерные объемы, и используют различные функции расстояния между объектами обучающей выборки и распознаваемыми объектами [70]. Для количественных признаков, когда их число много меньше объема выборки, операции с объектами играют промежуточную роль в оценке локальных плотностей распределения условных вероятностей и объекты не несут смысловой нагрузки самостоятельных информационных единиц. В то же время, когда количество признаков соизмеримо или больше числа исследуемых объектов, а признаки носят качественный или дихотомический характер, то ни о каких

локальных оценках плотностей распределения вероятностей не может идти речи. В этом случае объекты в указанных непараметрических методах рассматриваются как самостоятельные информационные единицы (целостные эмпирические факты) и данные методы приобретают смысл оценок сходства и различия изучаемых объектов.

Таким образом, одни и те же технологические операции непараметрических методов в зависимости от условий задачи имеют смысл либо локальных оценок плотностей распределения вероятностей значений признаков, либо оценок сходства и различия объектов.

Методы, основанные на предположениях о классе решающих функций.

В данной группе методов считается известным общий вид решающей функции и задан функционал ее качества. На основании этого функционала по обучающей последовательности ищется наилучшее приближение решающей функции. Самыми распространенными являются представления решающих функций в виде линейных и обобщенных нелинейных полиномов. Функционал качества решающего правила обычно связывают с ошибкой классификации.

Основным достоинством методов, основанных на предположениях о классе решающих функций, является ясность математической постановки задачи распознавания, как задачи поиска экстремума. Решение этой задачи нередко достигается с

помощью каких-либо градиентных алгоритмов. Многообразие методов этой группы объясняется широким спектром используемых функционалов качества решающего правила и алгоритмов поиска экстремума. Обобщением рассматриваемых алгоритмов, к которым относятся, в частности, алгоритм Ньютона, алгоритмы перцептронного типа и др., является метод стохастической аппроксимации. В отличие от параметрических методов распознавания успешность применения данной группы методов не так сильно зависит от рассогласования теоретических представлений о законах распределения объектов в пространстве признаков с эмпирической реальностью. Все операции подчинены одной главной цели - нахождению экстремума функционала качества решающего правила. В то же время результаты параметрических и рассматриваемых методов могут быть похожими. Как показано выше, параметрические методы для случая нормальных распределений объектов в различных классах с равными ковариационными матрицами приводят к линейным решающим функциям. Отметим также, что алгоритмы отбора информативных признаков в линейных диагностических моделях, можно интерпретировать как частные варианты градиентных алгоритмов поиска экстремума.

Логические методы.

Логические методы распознавания образов базируются на аппарате алгебры логики и позволяют оперировать информацией, заключенной не только в отдельных признаках, но и в сочетаниях

значений признаков. В этих методах значения какого-либо признака рассматриваются как элементарные события [70]. В самом общем виде логические методы можно охарактеризовать как разновидность поиска по обучающей выборке логических закономерностей и формирование некоторой системы логических решающих правил (например, в виде конъюнкций элементарных событий), каждое из которых имеет собственный вес. Группа логических методов разнообразна и включает методы различной сложности и глубины анализа. Для дихотомических (булевых) признаков популярными являются так называемые древообразные классификаторы, метод тупиковых тестов, алгоритм “Кора” и другие. Более сложные методы основываются на формализации индуктивных методов Д.С.Милля. Формализация осуществляется путем построения квазиаксиоматической теории и базируется на многосортной многозначной логике с кванторами по кортежам переменной длины [70].

Алгоритм “Кора”, как и другие логические методы распознавания образов, является достаточно трудоемким, поскольку при отборе конъюнкций необходим полный перебор. Поэтому при применении логических методов предъявляются высокие требования к эффективной организации вычислительного процесса, и эти методы хорошо работают при сравнительно небольших размерностях пространства признаков и только на мощных компьютерах.

Лингвистические (структурные) методы.

Лингвистические методы распознавания образов основаны на использовании специальных грамматик порождающих языки, с помощью которых может описываться совокупность свойств распознаваемых объектов. Для различных классов объектов выделяются непроизводные (атомарные) элементы (подобразы, признаки) и возможные отношения между ними. Каждый объект представляется совокупностью непроизводных элементов, “соединенных” между собой теми или иными способами или, другими словами, “предложением” некоторого “языка”. Путем синтаксического анализа (грамматического разбора) “предложения” устанавливается его синтаксическая “правильность” или, что эквивалентно, - может ли некоторая фиксированная грамматика (описывающая класс) породить имеющееся описание объекта. Грамматический разбор производится так называемым “синтаксическим анализатором”, который представляет полное синтаксическое описание объекта в виде дерева грамматического разбора, если объект является синтаксически правильным (принадлежит классу, описываемому данной грамматикой). В противном случае, объект либо отклоняется, либо подвергается анализу с помощью других грамматик, описывающих другие классы объектов. Известны бесконтекстные, автоматные и другие типы грамматик. Однако задача восстановления (определения) грамматик по некоторому множеству высказываний (предложений - описаний объектов), порождающих данный язык, является трудно формализуемой.

Экстенсиональные методы.

В методах данной группы, в отличие от интенсионального направления, каждому изучаемому объекту в большей или меньшей мере придается самостоятельное диагностическое значение. По своей сути эти методы близки к клиническому подходу, который рассматривает людей не как проранжированную по тому или иному показателю цепочку объектов, а как целостные системы, каждая из которых индивидуальна и имеет особенную диагностическую ценность [70]. Основными операциями в распознавании образов с помощью обсуждаемых методов являются операции определения сходства и различия объектов. Объекты в указанной группе методов играют роль диагностических прецедентов. При этом в зависимости от условий конкретной задачи роль отдельного прецедента может меняться в самых широких пределах от главной до весьма косвенного участия в процессе распознавания. В свою очередь, условия задачи могут требовать для успешного решения участия различного количества диагностических прецедентов от одного в каждом распознаваемом классе до полного объема выборки, а также разных способов вычисления мер сходства и различия объектов. Этими требованиями объясняется дальнейшее разделение экстенсиональных методов на подклассы.

Метод сравнения с прототипом.

Это наиболее простой экстенсиональный метод распознавания. Он применяется, например, тогда, когда

распознаваемые классы отображаются в пространстве признаков компактными геометрическими группировками. В таком случае обычно в качестве точки - прототипа выбирается центр геометрической группировки класса (или ближайший к центру объект). Для классификации неизвестного объекта находится ближайший к нему прототип, и объект относится к тому же классу, что и этот прототип. Очевидно, никаких обобщенных образов классов в данном методе не формируется. В качестве меры близости могут применяться различные типы расстояний. Часто для дихотомических признаков используется расстояние Хэмминга, которое в данном случае равно квадрату евклидова расстояния. При этом решающее правило классификации объектов эквивалентно линейной решающей функции.

Указанный факт следует особо отметить. Он наглядно демонстрирует связь прототипной и признаковой репрезентации информации о структуре данных. Пользуясь приведенным представлением, можно, например, любую традиционную измерительную шкалу, являющуюся линейной функцией от значений дихотомических признаков, рассматривать как гипотетический диагностический прототип. В свою очередь, если анализ пространственной структуры распознаваемых классов позволяет сделать вывод об их геометрической компактности, то каждый из этих классов достаточно заменить одним прототипом который, фактически эквивалентен линейной диагностической модели.

На практике, конечно, ситуация часто бывает отличной от описанного идеализированного примера. Перед исследователем, намеревающимся применить метод распознавания, основанный на сравнении с прототипами диагностических классов, встают непростые проблемы. Это, в первую очередь, выбор меры близости (метрики), от которого может существенно измениться пространственная конфигурация распределения объектов. И, во-вторых, самостоятельной проблемой является анализ многомерных структур экспериментальных данных. Обе эти проблемы особенно остро встают перед исследователем в условиях высокой размерности пространства признаков, характерной для реальных задач.

Метод k-ближайших соседей.

Метод k-ближайших соседей для решения задач дискриминантного анализа был впервые предложен еще в 1952 году. Он заключается в следующем.

При классификации неизвестного объекта находится заданное число (k) геометрически ближайших к нему в пространстве признаков других объектов (ближайших соседей) с уже известной принадлежностью к распознаваемым классам. Решение об отнесении неизвестного объекта к тому или иному диагностическому классу принимается путем анализа информации об этой известной принадлежности его ближайших соседей, например, с помощью простого подсчета голосов.

Первоначально метод k-ближайших соседей рассматривался как непараметрический метод оценивания отношения правдоподобия. Для этого метода получены теоретические оценки его эффективности в сравнении с оптимальным байесовским классификатором. Доказано, что асимптотические вероятности ошибки для метода k-ближайших соседей превышают ошибки правила Байеса не более чем в два раза.

Как отмечалось выше, в реальных задачах часто приходится оперировать объектами, которые описываются большим количеством качественных (дихотомических) признаков. При этом размерность пространства признаков соизмерима или превышает объем исследуемой выборки. В таких условиях удобно интерпретировать каждый объект обучающей выборки, как отдельный линейный классификатор. Тогда тот или иной диагностический класс представляется не одним прототипом, а набором линейных классификаторов. Совокупное взаимодействие линейных классификаторов дает в итоге кусочно-линейную поверхность, разделяющую в пространстве признаков распознаваемые классы. Вид разделяющей поверхности, состоящей из кусков гиперплоскостей, может быть разнообразным и зависит от взаимного расположения классифицируемых совокупностей.

Также можно использовать другую интерпретацию механизмов классификации по правилу k-ближайших соседей. В ее основе лежит представление о существовании некоторых

латентных переменных, абстрактных или связанных каким-либо преобразованием с исходным пространством признаков. Если в пространстве латентных переменных попарные расстояния между объектами такие же, как и в пространстве исходных признаков, и количество этих переменных значительно меньше числа объектов, то интерпретация метода k-ближайших соседей может рассматриваться под углом зрения сравнения непараметрических оценок плотностей распределения условных вероятностей. Приведенное здесь представление о латентных переменных близко по своей сути к представлению об истинной размерности и другим представлениям, используемым в различных методах снижения размерности.

При использовании метода k-ближайших соседей для распознавания образов исследователю приходится решать сложную проблему выбора метрики для определения близости диагностируемых объектов. Эта проблема в условиях высокой размерности пространства признаков чрезвычайно обостряется вследствие достаточной трудоемкости данного метода, которая становится значимой даже для высокопроизводительных компьютеров. Поэтому здесь так же, как и в методе сравнения с прототипом, необходимо решать творческую задачу анализа многомерной структуры экспериментальных данных для минимизации числа объектов, представляющих диагностические классы.

Следует заметить, что уменьшение числа объектов в обучающей выборке (диагностических прецедентов) является недостатком данного метода, т.к. уменьшает представительность обучающей выборки.

Алгоритмы вычисления оценок (голосования).

Принцип действия алгоритмов вычисления оценок (АВО) состоит в вычислении приоритете (оценок сходства), характеризующих “близость” распознаваемого и эталонных объектов по системе ансамблей признаков, представляющей собой систему подмножеств заданного множества признаков. В отличие от всех ранее рассмотренных методов алгоритмы вычисления оценок принципиально по-новому оперируют описаниями объектов. Для этих алгоритмов объекты существуют одновременно в самых разных подпространствах пространства признаков. Класс АВО доводит идею использования признаков до логического конца: поскольку не всегда известно, какие сочетания признаков наиболее информативны, то в АВО степень сходства объектов вычисляется при сопоставлении всех возможных или определенных сочетаний признаков, входящих в описания объектов [70].

Используемые сочетания признаков (подпространства) называются опорными множествами или множествами частичных описаний объектов. Вводится понятие обобщенной близости между распознаваемым объектом и объектами обучающей выборки (с известной классификацией), которые называют

эталонными объектами. Эта близость представляется комбинацией близостей распознаваемого объекта с эталонными объектами, вычисленных на множествах частичных описаний. Таким образом, АВО является расширением метода k-ближайших соседей, в котором близость объектов рассматривается только в одном заданном пространстве признаков.

Еще одним расширением АВО является то, что в данных алгоритмах задача определения сходства и различия объектов формулируется как параметрическая и выделен этап настройки АВО по обучающей выборке, на котором подбираются оптимальные значения введенных параметров. Критерием качества служит ошибка распознавания. Параметры АВО задаются в виде значений порогов и (или) как веса указанных составляющих.

Теоретические возможности АВО превышают или, по крайней мере, не ниже возможностей любого другого алгоритма распознавания образов, так как с помощью АВО могут быть реализованы все мыслимые операции с исследуемыми объектами. Но, как это обычно бывает, расширение потенциальных возможностей наталкивается на большие трудности их практического воплощения. На практике применение АВО для решения высокоразмерных задач сопровождается введением каких-либо эвристических ограничений и допущений. В частности, известен пример использования АВО в

психодиагностике, в котором апробирована разновидность ABO, фактически эквивалентная методу k-ближайших соседей.

Коллективы решающих правил.

Так как различные алгоритмы распознавания проявляют себя по-разному на одной и той же выборке объектов, то закономерно встает вопрос о синтетическом решающем правиле, адаптивно использующем сильные стороны этих алгоритмов. В синтетическом решающем правиле применяется двухуровневая схема распознавания. На первом уровне работают частные алгоритмы распознавания, результаты которых объединяются на втором уровне в блоке синтеза. Наиболее распространенные способы такого объединения основаны на выделении областей компетентности того или иного частного алгоритма. Простейший способ нахождения областей компетентности заключается в априорном разбиении пространства признаков исходя из профессиональных соображений конкретной науки (например, расслоение выборки по некоторому признаку). Тогда для каждой из выделенных областей строится собственный распознающий алгоритм. Другой способ базируется на применении формального анализа для определения локальных областей пространства признаков как окрестностей распознаваемых объектов, для которых доказана успешность работы какого-либо частного алгоритма распознавания.

Самый общий подход к построению блока синтеза рассматривает результирующие показатели частных алгоритмов

как исходные признаки для построения нового обобщенного решающего правила. В этом случае могут использоваться все перечисленные выше методы интенсионального и экстенсионального направлений в распознавании образов. Эффективными для решения задачи создания коллектива решающих правил являются логические алгоритмы типа “Кора” и алгоритмы вычисления оценок (АВО), положенные в основу так называемого алгебраического подхода, обеспечивающего исследование и конструктивное описание алгоритмов распознавания, в рамки которого укладываются все существующие типы алгоритмов [70].

Таким образом, задачу распознавания образов можно отнести к задаче описания неколичественных параметров. Далее будем рассматривать критерии выбора императивных шкал в моделях распознавания, использующих линейные разделяющие (дискриминантные) функции. По приведённой выше классификации эти модели распознавания относятся к группе интенсиональных методов.

Для задачи распознавания образов, постановка которой изложена в параграфе I.1, императивные шкалы предполагается задавать такими числовыми метками, для которых минимизировалось бы следующее отношение рассеиваний:

$$J = \frac{\Lambda_W}{\Lambda_T},$$

где

$$\Lambda_T = \Lambda = \Lambda_{l+1,l+1} \bar{a}^T \boldsymbol{S}\, \bar{a},$$

матрица $\boldsymbol{S}$ вычисляется по формуле (II.8). Для вычисления $\Lambda_W$ обозначим $\underline{\Lambda}_W = (\lambda_{ij}^W)$. Тогда

$$\lambda_{ij}^W = \frac{1}{N} \sum_{p=1}^{Q} N_p \lambda_{ij}^{(p)},$$

где $i, j = \overline{1,l}$; Q – количество обучающих выборок (ОВ); $N_p$ - количество объектов в $p$-й обучающей выборке;

$$\lambda_{k,l+1}^W = \lambda_{l+1,k}^W = \frac{1}{N} \sum_{i=1}^{g} a_i \sum_{p=1}^{Q} N_p \mu_k^{(p)i},$$

где $\mu_k^{(p)i}$ параметр $\mu_k^i$, определённый по $p$-й обучающей выборке.

Обозначим через $m_k^{(p)i}$ и $m_k^{(p)}$ частное и общее средние значения $k$-го признака по $p$-й ОВ; $n_i^{(p)}$ и $t_{ij}^{(p)}$ - количество объектов в $i$-й градации и значение параметра $t_{ij}$ из (II.6), определённое по $p$-й ОВ.

Введём следующие параметры:

$$\mu_k^{*i} = \frac{1}{N} \sum_{p=1}^{Q} N_p \mu_k^{(p)i} = \frac{1}{N} \sum_{p=1}^{Q} N_p \frac{n_i^{(p)}}{N_p} (m_k^{(p)i} - m_k^{(p)}) = \frac{1}{N} \sum_{p=1}^{Q} (n_i^{(p)} m_k^{(p)i} - n_i^{(p)} m_k^{(p)}) =$$

$$\frac{n_i}{N} \left( m_k^i - \frac{1}{n_i} \sum_{p=1}^{Q} n_i^{(p)} m_k^{(p)} \right);$$

$$t_{ij}^* = \frac{1}{N} \sum_{p=1}^{Q} N_p t_{ij}^{(p)},$$

причём можно показать, что

$$\sum_{i=1}^{g} t_{ij}^* = 0; \qquad \text{(II.21)}$$

$$\sum_{i=1}^{g} \mu_k^{*i} = 0. \qquad \text{(II.22)}$$

Тогда получим:

$$\lambda^{W}_{k,l+1} = \sum_{i=1}^{g} a_i \ \mu_k^{*i}, \ \ k = \overline{1,l};$$

$$\lambda^{W}_{l+1,l+1} = \sum_{i=1}^{g}\sum_{j=1}^{g} a_i a_j \ t^{*}_{ij}.$$

Преобразовав определитель $\Lambda_W$, как и $\Lambda$, получим:

$$\Lambda_W = \Lambda_{Wl+1,l+1} \ \bar{a}^T \boldsymbol{W} \bar{a}, \qquad \text{(II.23)}$$

где

$$\boldsymbol{W} = (\omega_{ij}) = \boldsymbol{T}^* - \boldsymbol{M}^* \ \underline{\Lambda}^{-1}_{Wl+1,l+1} \ \boldsymbol{M}^{*T},$$

$$\boldsymbol{T}^* = (t^*_{ij}), \ \boldsymbol{M}^* = (\mu_k^{*i}), \ i, j = \overline{1,g}; \ k = \overline{1,l}.$$

Из (II.21) и (II.22) следует равенство:

$$\sum_{i=1}^{g} \omega_{ki} = 0, \qquad \text{(II.24)}$$

тогда отношение рассеиваний можно записать в виде:

$$\boldsymbol{J} = \frac{\Lambda_{Wl+1,l+1}}{\Lambda_{l+1,l+1}} \ \frac{\bar{a}^T \boldsymbol{W} \bar{a}}{\bar{a}^T \boldsymbol{S} \bar{a}}.$$

Поскольку $\Lambda_W \geq 0$ и выполняется (II.24), то квадратичная форма $\bar{a}^T \boldsymbol{W} \bar{a}$ является неотрицательной. Первый сомножитель в последнем равенстве не зависит от набора $\bar{a}$ и критерий выбора императивных шкал для задачи распознавания образов можно записать в виде:

$$\frac{\bar{a}^T \boldsymbol{W} \bar{a}}{\bar{a}^T \boldsymbol{S} \bar{a}} \to \min, \qquad \text{(II.25)}$$

где $\bar{a} \in \boldsymbol{B}^g$.

При условии, что необходимо найти оптимальные целочисленные числовые метки, задача (II.25) примет вид:

$$\frac{\overline{D}^T \boldsymbol{W} \overline{D}}{\overline{D}^T \boldsymbol{S} \overline{D}} \to \min, \qquad \text{(II.26)}$$

где $D \in \mathrm{P}^g$.

Для оптимизации критерия (II.26) используется подход, изложенный в п. II.1.

Таким образом, рассматривается возможность использования неколичественных признаков $x^i$, $i = \overline{l+1, p}$ для увеличения точности линейной модели распознавания. Линейные разделяющие функции [66] могут быть построены только для количественных признаков и применение признаков $x^i$, $i = \overline{l+1, p}$ допустимо только после их преобразования в количественные с использованием императивных шкал $\xi_i$, $i = \overline{l+1, p}$. С целью уменьшения уровня ошибки распознавания при поиске императивных шкал оптимизируется функция критерия рассеивания $\boldsymbol{J} = \frac{\Lambda_W}{\Lambda_T}$, вычисленной с учётом оцифрованных признаков.

## II.3. Разработка алгоритмов преобразования типов признаков для задач множественного линейного регрессионного анализа

Исследование зависимости одного из признаков, входящих в ТЭД, от набора других является одной из важных прикладных задач обработки экспериментальных данных в АСНИ.

Исследование зависимостей относится к задачам описания**,** которые для количественных признаков решаются методами множественного линейного регрессионного анализа.

В линейный регрессионный анализ входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных

$X = (x_1 , ..., x_p)$ и $Y = (y_1 ,..., y_m)$.

Предполагается, что $X$ - независимые переменные (факторы, объясняющие переменные) влияют на значения $Y$ - зависимых переменных (откликов, объясняемых переменных). По имеющимся эмпирическим данным $(X_i , Y_i)$, $i = 1, ..., n$ требуется построить функцию $f(X)$, которая приближенно описывала бы изменение $Y$ при изменении $X$:

$Y \approx f(X)$.

Предполагается, что множество допустимых функций, из которого подбирается $f(X)$, является параметрическим:

$f(X) = f(X, \theta)$,

где $\theta$ - неизвестный параметр (вообще говоря, многомерный). При построении $f(X)$ будем считать, что

$$Y = f(X, \theta) + \varepsilon,$$

где первое слагаемое - закономерное изменение $Y$ от $X$, а второе - $\varepsilon$ - случайная составляющая с нулевым средним; $f(X, \theta)$ является условным математическим ожиданием $Y$ при условии известного $X$ и называется регрессией Y по X**.**

Рассмотрим простую линейную регрессию. Пусть $X$ и $Y$ одномерные величины; обозначим их $x$ и $y$, а функция $f(x, \theta)$ имеет вид $f(x, \theta) = A + bx$, где $\theta = (A, b)$. Относительно имеющихся наблюдений $(x_i, y_i)$, $i = 1, ..., n$, полагаем, что

$$y_i = A + bx_i + \varepsilon_i ,$$

где $\varepsilon_1, ..., \varepsilon_n$ - независимые (ненаблюдаемые) одинаково распределенные случайные величины. Можно различными методами подбирать “лучшую” прямую линию. Широко используется метод наименьших квадратов. Построим оценку параметра $\theta = (A, b)$ так, чтобы величины

$$e_i = y_i - f(x_i, \theta) = y_i - A - bx_i ,$$

называемые остатками, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - A - bx_i)^2 = \min \text{ по } (A, b).$$

Чтобы упростить формулы, положим в $x_i = x_i - \bar{x} + \bar{x}$; получим:

$$y_i = a + b(x_i - \bar{x}) + \varepsilon_i , \quad i = 1, ..., n,$$

где $\bar{x} = \sum_{i=1}^{n} x_i / n$, $a = A + b\bar{x}$.

Сумму $\sum_{i=1}^{n} \left(y_i - a - b(x_i - \bar{x})\right)^2$ минимизируем по $(a,b)$, приравнивая нулю производные по $a$ и $b$; получим систему линейных уравнений относительно $a$ и $b$. Ее решение $(\hat{a}, \hat{b})$ легко находится:

$$\hat{a} = \bar{y}, \qquad \text{где} \quad \bar{y} = \sum_{i=1}^{n} y_i / n,$$

$$\hat{b} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} .$$

Приведём некоторые свойства оценок. Нетрудно показать, что если $M\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$ , то:

1) $M\hat{a} = a, \ M\hat{b} = b$, т.е. оценки несмещенные;

2) $D\hat{a} = \sigma^2 / n, \ \mathrm{D}\hat{b} = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$ ;

3) $cov\,(\hat{a}, \hat{b}) = 0$;

Если дополнительно предположить нормальность распределения $\varepsilon_i$ , то

4) оценки $\hat{a}$ и $\hat{b}$ нормально распределены и независимы;

5) остаточная сумма квадратов

$$Q^2 = \sum_{i=1}^{n} \left[ y_i - \hat{a} - \hat{b}(x_i - \bar{x}) \right]^2$$

независима от $(\hat{a}, \hat{b})$, а $Q^2 / \sigma^2$ распределена по закону хи-квадрат $\chi^2_{n-2}$ с $n$-2 степенями свободы.

Оценка для $\sigma^2$ и доверительные интервалы**.** Свойство 5) дает возможность несмещённо оценивать неизвестный параметр $\sigma^2$ величиной

$$s^2 = Q^2 / (n\text{-}2).$$

Поскольку $s^2$ независима от $\hat{a}$ и $\hat{b}$ , отношения

$$\sqrt{n}\,\frac{\hat{a}-a}{s} \quad \text{и} \quad \frac{\hat{b}-b}{s_b}, \quad \text{где } s_b = s \Big/ \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

имеют распределение Стьюдента с ($n$-2) степенями свободы, и потому доверительные интервалы для $a$ и $b$ таковы:

$$|\hat{a}-a| \le t_p \frac{s}{\sqrt{n}}, \quad |\hat{b}-b| \le t_p s_b,$$

где $t_p$ - квантиль уровня $(1 + P_Д) / 2$ распределения Стьюдента с $n$ - 2 степенями свободы, $P_Д$ - коэффициент доверия.

Проверка гипотезы о коэффициенте наклона**.** Проверим гипотезу $H$: $b = 0$. Если 0 не входит в доверительный интервал для $b$, т.е.

$$(b(\,/\,s_b > t_p,$$

то гипотезу $H$ следует отклонить; уровень значимости при этом $\alpha = 1 - P_Д$.

Другой способ проверки гипотезы $H$ состоит в вычислении статистики

$$F = \frac{\hat{b}^2 / D\hat{b}}{Q^2 /(\sigma^2(n-2))} = \frac{\hat{b}^2}{s_b^2},$$

распределенной, если гипотеза $H$ верна, по закону $F$ (1, $n$ – 2) Фишера с числом степеней свободы 1 и $n$ – 2. Если

$$F > F_{1-\alpha},$$

где $F_{1-\alpha}$ - квантиль уровня $1 - \alpha$ распределения $F$ (1, $n$ - 2), то гипотеза $H$ отклоняется с уровнем значимости $\alpha$..

Вариация зависимой переменной и коэффициент детерминации**.**

Рассмотрим вариацию (разброс) $T_{ss}$ значений $y_i$ относительно среднего значения $\bar{y}$

$$T_{ss} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Обозначим $\hat{y}_i$ предсказанные с помощью функции регрессии значения $y_i$: $\hat{y} = \hat{a} + \hat{b}x_i$. Сумма $R_{ss}$, вычисляемая по формуле

$$R_{ss} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 ,$$

означает величину разброса, которая обусловлена регрессией (ненулевым значением наклона $\hat{b}$). Сумма $E_{ss}$

$$E_{ss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

означает разброс за счет случайных отклонений от функции регрессии. Оказывается,

$$T_{ss} = R_{ss} + E_{ss} ,$$

т.е. полный разброс равен сумме разбросов за счет регрессии и за счет случайных отклонений. Величина $R_{ss} / T_{ss}$ - это доля вариации значений $y_i$, обусловленной регрессией (т.е. доля закономерной изменчивости в общей изменчивости). Статистика

$$R^2 = R_{ss} / T_{ss} = 1 - E_{ss} / T_{ss}$$

называется коэффициентом детерминации. Если $R^2 = 0$, это означает, что регрессия ничего не дает, т.е. знание $x$ не улучшает предсказания для $y$ по сравнению с тривиальным $\hat{y}_i = \bar{y}$. Другой

крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной прямой. Чем ближе к 1 значение $R^2$ , тем лучше качество подгонки.

Перейдём теперь к постановке задачи множественного регрессионного анализа. Пусть исходная таблица экспериментальных данных состоит из $N$ объектов и $p$, причём имеются признаки различных типов (II.1). Рассмотрим модель множественной линейной регрессии для количественных признаков $x_1, x_2, \ldots, x_l$. Пусть в качестве зависимой переменной выделен признак $x_1$ и его необходимо оценить линейной комбинацией признаков $x_2, \ldots, x_l$, рассматриваемых как независимые переменные. Тогда множественный линейный регрессионный анализ [67, 68] сводится к поиску такой гиперплоскости

$$x_1 = \beta_{12} x_2 + \beta_{13} x_3 + \ldots + \beta_{1l} x_l + \beta_{10},$$

для которой выполняется соотношение

$$\sum_{i=1}^{N} (x_1^i - \beta_{12} x_2^i - \ldots - \beta_{1l} x_l^i - \beta_{10})^2 = min. \qquad \text{(II.27)}$$

Используемая в анализе данных в качестве модели описания одного признака из таблицы экспериментальных данных посредством линейной комбинации остальных, множественная линейная регрессия связана с классической статистической моделью регрессии как условного среднего [55, 56, 69]. В этой модели значения признаков рассматриваются как выборки объёмом $N$ для случайных величин с произвольным

распределением, имеющим конечные моменты второго порядка. В этом случае выражение (II.27) определяет по методу наименьших квадратов плоскость среднеквадратической регрессии, которая является наилучшим линейным приближением к поверхности регрессии

$$x_1/x_2, \ldots, x_l = f(x_2, \ldots, x_l),$$

где $x_1/x_2, \ldots, x_l$ - условное среднее значение для $x_1$ при фиксированных $x_2, \ldots, x_l$.

Наряду с моделью регрессии как условного среднего в классической статистике рассматривается также модель регрессии как безусловного среднего. Здесь случайной величиной является лишь зависимая переменная, а независимые переменные считаются детерминированными. Эти две модели регрессии отличаются только статистическими свойствами оценок параметров плоскости регрессии, в то время как вычислительные аспекты совпадают как для классических статистических моделей, так и для линейной регрессии, рассматриваемой в анализе данных. Поэтому все аналитические выражения для определения плоскости регрессии по разнотипным данным могут быть использованы и при построении любой из двух классических моделей.

Определим параметры уравнения регрессии [56] по элементам ковариационной матрицы

$$\underline{\Lambda} = \begin{pmatrix} \lambda_{11}\lambda_{12}.....\lambda_{1l} \\ \lambda_{21}\lambda_{22}....\lambda_{2l} \\ ..... \\ \lambda_{l1}\lambda_{l2}.....\lambda_{ll} \end{pmatrix}$$

где $\lambda_{ij}$ - ковариации $i$-го и $j$-го признаков.

Аналогично п. II.1 обозначим через $\Lambda$ определитель матрицы $\underline{\Lambda}$, а через $\Lambda_{ij}$ - дополнительный минор элемента $\lambda_{ij}$ в определителе $\Lambda$. В этом случае неизвестные коэффициенты регрессии вычисляются по следующим формулам:

$$\beta_{1i} = \frac{(-1)^{i+j}\Lambda_{1i}}{\Lambda_{11}}, \quad \beta_{10} = \overline{x_1} - \sum_{i=2}^{l} \beta_{1i}\overline{x_i},$$

где $\overline{x_i}$, $i = \overline{1,l}$ - средние значения признаков, вычисленные по значениям из ТЭД.

Величину остаточной дисперсии $\sigma^2_{1,2,...,l}$ определяет обращающаяся в минимум сумма квадратов отклонений (II.27) действительных значений признака $x_1$ от задаваемой коэффициентами равенства плоскости. Остаточная дисперсия позволяет судить о точности оценки значения $x_1$ посредством линейной комбинации признаков $x_2$, … , $x_l$ и определяется выражением:

$$\sigma^2_{1,2,...,l} = \frac{1}{N}\sum_{i=1}^{N}(x_1^i - \beta_{12}x_2^i - ... - \beta_{1l}x_l^i - \beta_0)^2 = \frac{\Lambda}{\Lambda_{11}}. \qquad \text{(II.28)}$$

Другой характеристикой качества линейной аппроксимации зависимости $x_1$ от $x_2$, … , $x_l$ является множественный коэффициент корреляции, вычисляемый по формуле:

$$r_{1,2,\ldots,l} = \sqrt{1 - \frac{\sigma^2_{1,2,\ldots,l}}{\sigma^2_1}}\,. \qquad \text{(II.29)}$$

Коэффиентом детерминации называется квадрат коэффициента корреляции $r_{1,2,\ldots,l}$. Этот коэффициент характеризует долю полной дисперсии $\sigma_1^2$, объясняемую плоскостью регрессии $x_1$ по $x_2$, … , $x_l$. При наличии частных коэффициентов корреляции остаточная дисперсия вычисляется по формуле:

$$\sigma^2_{1,23\ldots l} = \sigma^2_1 (1 - r^2_{12})(1 - r^2_{13,2}) \ldots (1 - r^2_{1l,23,\ldots,l-1}). \qquad \text{(II.30)}$$

Выражение (II.30) можно переписать в виде:

$$\sigma^2_{1,23\ldots l} = \sigma^2_{1,23\ldots l-1}(1 - r^2_{1l,23,\ldots,l-1}). \qquad \text{(II.31)}$$

Соотношение (II.31) показывает, что представление $x_1$ линейной комбинацией величин $x_2$, … , $x_{l-1}$ может улучшено введением следующей величины $x_l$ в случае, если $r_{1l,23,\ldots,l-1} \neq 0$. Таким образом, последовательное дополнение модели линейной регрессии новыми зависимыми переменными, коррелированными с $x_1$, с каждым шагом всё более уточняет линейную аппроксимацию зависимости и чем больше таких переменных принято во внимание, тем точнее окажется представление зависимой переменной. Из сказанного следует, что для нахождения наиболее точного описания целесообразным является использование в регрессионной модели качественных и классификационных признаков $x_l$, $i = \overline{l+1, p}$.

Известно, что классическую модель линейной регрессии можно рассматривать только при наличии в ТЭД количественных признаков. Использование же признаков $x_l$, $i=\overline{l+1,p}$ в этой модели допустимо после их предварительного преобразования в количественные путём введения императивных шкал. В этом случае для определения параметров плоскости среднеквадратической регрессии $x_1$ по всем $x_l$, $i=\overline{2,p}$ необходимо доопределить матрицу $\underline{\Lambda}$, дополнив её строками и столбцами ковариаций неколичественных признаков. При построении многомерной модели для каждого неколичественного признака необходимо найти одну единственную императивную шкалу, которая используется при определении ковариаций со всеми остальными признаками из таблицы экспериментальных данных.

При решении задачи множественного линейного регрессионного анализа количественного признака $x_1$ по всем $x_l$, $i=\overline{2,p}$ императивные шкалы для неколичественных признаков $x_l$, $i=\overline{l+1,p}$ следует задавать по числовым меткам $\bar{a}_{l+1}, \ldots, \bar{a}_p$ в соответствии со следующим критерием:

$$\sigma^2_{1,23\ldots p}(X_1, \bar{a}_{l+1}, \ldots, \bar{a}_p) \to min, \qquad \text{(II.32)}$$

где $\bar{a}_{l+1} \in \mathrm{B}^{g_{l+1}}, \ldots, \bar{a}_p \in \mathrm{B}^{g_p}$,

или в случае поиска целочисленных числовых меток из множеств $\mathrm{K}_{g_i}$

$$\sigma^2_{1,23\ldots p}(X_1, X_2, D_{m+1}, \ldots, D_p) \to min, \qquad \text{(II.33)}$$

где $D_{m+1} \in \mathrm{P}^{g_{m+1}}, \ldots, D_p \in \mathrm{P}^{g_p}$.

Аналогично п. II.1 можно получить равенство:

$$\Lambda_{11} = \Lambda_{1.1,l+1,l+1} \bar{a}^T \mathrm{S}' \bar{a}, \tag{II.34}$$

где

$$\mathrm{S}' = \mathrm{T} - \mathrm{M} \Lambda^{-1}_{1.1,l+1.l+1} \mathrm{M}^{\mathrm{T}},$$

$$\bar{a}^T \mathrm{S}' \bar{a} = \sigma^2_{l+1}(1 - r^2_{l+1.2\ldots l}). \tag{II.35}$$

В выражение (II.28) подставим (II.9) и (II.34):

$$\sigma^2_{1,2,\ldots,l+1} = \frac{\Lambda}{\Lambda_{11}} = \sigma^2_{1,23\ldots l} \frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{S}' \bar{a}}. \tag{II.36}$$

Критерий поиска императивной шкалы для задачи регрессионного анализа примет вид:

$$\frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{S}' \bar{a}} \to min, \tag{II.37}$$

где $\bar{a} \in \mathrm{B}^g$.

Из (II.31) и (II.36) вытекает следующее равенство:

$$\frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{S}' \bar{a}} = 1 - r^2_{1l+1.23,\ldots,l}. \tag{II.38}$$

Императивная шкала, определяемая для модели линейной регрессии критерием (II.37), обеспечивает максимальное значение частного коэффициента корреляции количественного признака $x_1$ и оцифрованного признака $x^{\xi}_{l+1}$.

Предположим, что в качестве зависимой переменной рассматривается неколичественный признак $x^{\xi}_{l+1}$. Тогда числовые

метки, задающие для $x_{l+1}$ императивную шкалу, должны минимизировать величину $1-r^2_{l+1.1...l}$. Можно показать, что в этом случае критерий поиска императивной шкалы для неколичественного признака $x_{l+1}$ будет иметь вид:

$$\frac{\bar{a}^T S \bar{a}}{\bar{a}^T T \bar{a}} \to min, \qquad \text{(II.39)}$$

где $\boldsymbol{T}=[t_{ij}]$, $i, j=\overline{1,g}$,

$$t_{ij} = \begin{cases} -\dfrac{n_i n_j}{N}, \ i \neq j, \\ \dfrac{n_i}{N^2}(N-n_i), \ i = j, \end{cases}$$

$\bar{a} \in \mathrm{B}^g$.

В случае поиска целочисленных числовых меток критерии (II.37) и (II.39) примут вид:

$$\frac{\bar{D}^T \boldsymbol{S} \bar{D}}{\bar{D}^T \boldsymbol{S}' \bar{D}} \to min \text{ и } \frac{\bar{D}^T \boldsymbol{S} \bar{D}}{\bar{D}^T \boldsymbol{T} \bar{D}} \to min, \qquad \text{(II.40)}$$

где $\bar{D} \in \mathrm{P}^g$.

Для оптимизации критериев (II.40) также используется подход, изложенный в п. II.1.

Таким образом, использование методов преобразования типов признаков для задачи множественного линейного регрессионного анализа позволяет учесть влияние неколичественных признаков из таблицы экспериментальных данных для получения более точного описания зависимости

одного из признаков от остальных. При этом не исключается возможность использования классификационного или качественного признака в качестве зависимого.

## II.4. Экспериментальное исследование разработанных алгоритмов

В данном параграфе исследуется эффективность работы алгоритмов преобразования типов признаков, предложенных в предыдущих параграфах настоящей главы.

Эффективность работы алгоритмов преобразования типов признаков оценивалась на примере решения задачи распознавания образов. В качестве начальной информации были взяты хорошо изученные данные о трёх сортах ирисов [71]. Каждый сорт ирисов был представлен 50 цветками, у которых первоначально измерялись по четыре количественных признака, характеризующих длину и ширину лепестка, а также длину и ширину чашелистика. В целях проверки эффективности алгоритмов преобразования типов признаков четвёртый признак был преобразован в классификационный с числом градаций, равным 10. При этом каждый сорт представлялся по крайней мере двумя градациями и соответствующие этим градациям интервалы были приняты равными по величине. Присвоение числовых значений классификационному признаку производилось с целью перемешивания объектов, принадлежащих каждому из трёх сортов.

В целях решения задачи распознавания образов для каждого из трёх сортов ирисов произвольным образом была составлена обучающая выборка, включающая 25 объектов. Остальные 25 объектов каждого сорта составляли контрольную выборку.

Задача распознавания образов решалась для трёх массивов данных:

A. Исходные данные взяты непосредственно из [71] и представлены четырьмя количественными признаками;

B. Исходные данные представлены тремя количественными и одним классификационным признакам, полученным описанным выше способом. Количественные признаки также взяты из [71];

C. К массиву B был применён алгоритм поиска целочисленных меток для задачи распознавания образов, описанный в п. II.2.

Задача распознавания неклассифицированных ирисов при трёх указанных вариантах начальных данных решалась с помощью алгоритма RECHKA. Этот алгоритм входит в состав системы анализа данных, описание которой приведено в главе III настоящей работы. Распознавание неклассифицированных ирисов происходит с использованием трёх решающих правил: “дальнего соседа” (a), “средней связи” (b) и “ближнего соседа” (c) [40]. Исследования проводились на персональном компьютере. Результаты компьютерной обработки приведены в таблице II.1. Здесь на пересечении массива и решающего правила указано число неверно распознанных объектов.

| Решающее | a | b | c |
|---|---|---|---|
| A | 2 | 6 | 8 |
| B | 14 | 19 | 21 |
| C | 3 | 8 | 10 |

Таблица II.1.

Из таблицы II.1 видно, что наименьшее число ошибок получается при работе алгоритмов “дальнего соседа” и “средней связи”. Это обстоятельство объясняется тем, что алгоритмы классификации с использованием решающих правил “дальнего соседа” и “средней связи” распознают в основном объекты, образы которых представлены группировками эллипсоидной или сферической формы. В случае использования массива B число неверно распознанных объектов значительно возрастает, что объясняется сильным искажением исходной структуры данных при способе оцифровки, описанным выше, поскольку объекты, принадлежащие одинаковым обучающим выборкам, сильно перемешаны с объектами из других обучающих выборок.

Анализ таблицы II.1 показывает, что при рассмотрении массива C число неверно распознанных объектов при использовании всех трёх решающих правил значительно уменьшается по сравнению с вариантом B. Причём и в этом случае использование решающих правил “дальнего соседа” и “средней связи” обеспечивает лучшее качество распознавания. Это обстоятельство объясняется тем, что критерий выбора

императивных шкал для задачи распознавания обеспечивает объединение объектов обучающей выборки в компактные группы именно сферических или эллипсоидных форм.

Таким образом, применение разработанных алгоритмов преобразования типов признаков позволяет существенно повысить адекватность модели анализа данных по сравнению с исходными данными.

**Выводы по главе II.**

1. Предложены методы и алгоритмы поиска целочисленных меток градаций неколичественных признаков для различных задач анализа данных;
2. Даётся сравнительный анализ разработанных алгоритмов оцифровки и подхода, рассмотренного в [53];
3. Приведено экспериментальное исследование разработанных алгоритмов.

## ГЛАВА III. Разработка и применение прикладного программного обеспечения для решения задач анализа данных.

В настоящее время в связи с широким распространением вычислительной техники и повышением ее мощности актуальным становится вопрос эффективного ее применения для решения различных задач моделирования, прогнозирования, классификации и идентификации в различных областях науки и техники: экологии, климатологии, метеорологии, биологии,

геологии, океанографии и т.д. Особенность данных проблемных областей заключается в малом числе теоретически обоснованных и хорошо согласующихся с реальными данными вычислительных моделей. Поэтому прикладные задачи часто решаются на основе моделей, построенных по таблицам экспериментальных данных. При этом проблему представляет как сложность учета всех факторов, влияющих на ситуацию в конкретных территориях, так и сложность сбора территориально распределенной информации. В связи с этим часто приходится обрабатывать неполную информацию при наличии дублирующих друг друга либо малоинформативных признаков.

Если имеется таблица экспериментальных данных, где каждому объекту (строке таблицы) соответствует набор значений (столбцов) его независимых и зависимых признаков, то все задачи классификации и прогнозирования для такой таблицы можно свести к четырем классическим постановкам:

1. Распознавание образов (предсказание для объекта значения некоторого его целевого признака, выраженного в шкале наименований).

2. Предсказание значения числового (порядкового или количественного) признака для объекта.

3. Динамическое прогнозирование значения числового признака объекта, использующее временные измерения значений этого же признака (анализ временных рядов).

4. Автоматическая группировка объектов.

Каждая из перечисленных постановок сводится к единой задаче заполнения пропусков в таблице экспериментальных данных. При автоматической группировке объектов в таблицу добавляется новый столбец, содержащий информацию о разбиении всего множества объектов на группы похожих. Для иных постановок прогнозируются неизвестные значения признаков у тех объектов, где имеется пропущенная информация. Для этого требуется нахождение зависимостей в таблице экспериментальных данных. При этом появляются следующие специфические особенности:

- Таблица данных априорно является неполной, поскольку невозможно в общем случае описать все независимые и зависимые признаки, существенные для моделирования объекта или процесса. Это связано и с нашим ограниченным представлением о моделируемом объекте, и с ограничениями на возможность проведения тех или иных измерений.

- Задачи приходится решать при высокой априорной неопределенности, когда практически ничего неизвестно о виде функций распределения вероятностей в пространстве признаков. Всякое "сильное" предположение (о нормальности или унимодальности распределения, некоррелированности признаков и т.д.) ставит вопрос об адекватности предлагаемого вида действительному.

- При изучении сложных объектов возникают большие трудности при задании исходной системы признаков для их

описания. Поэтому в признаковом пространстве может быть много "дублирующих" и "шумящих" признаков. В результате проблема выбора наиболее информативной подсистемы признаков приобретает важное значение, поскольку уменьшение числа признаков часто улучшает качество решения (и сокращает экономические и временные затраты на измерения или сбор информации). Желательно иметь возможность определения значимости каждого признака для принятия решения и выделения минимально необходимого набора базовых признаков для прогнозирования целевого признака с заданной точностью.

- Для описания объектов используются признаки, измеренные в разных шкалах и, возможно, разнотипные - количественные (выраженные в шкалах интервалов, отношений и абсолютных значений), порядковые (шкалы порядка, частичного порядка, рангов, баллов) или номинальные (шкалы наименований).

- В связи со сложностью проведения измерений, отказом датчиков, историческими причинами в таблице могут отсутствовать некоторые значения исходных и целевых признаков у отдельных объектов. В данных всегда присутствуют ошибки разной природы, шум, а также имеются противоречия отдельных измерений друг другу. За исключением простых случаев, искажения в данных не могут быть устранены полностью

- Классификация объектов проблемной области, вводимая человеком, может не совсем точно отражать существующую в

проблемной области естественную кластерную структуру объектов, что создает дополнительные трудности.

Эти особенности не связаны только с перечисленными выше проблемными областями, но могут встретиться везде, где используется построение зависимостей по таблице экспериментальных данных. Также не зависит от проблемной области возможность сведения любой задачи прогнозирования и классификации к задаче "правдоподобного" заполнения пропусков в таблице.

Для обработки эмпирических данных традиционно используются классические методы математической статистики [6,18,24]. Можно получить для отсутствующих значений их условные математические ожидания (условия - значения других величин, описывающих конкретную ситуацию) и характеристики разброса - доверительные интервалы. Для решения задачи классификации методы математической статистики строят разделяющие поверхности между классами в признаковом пространстве. Однако достоверное статистическое оценивание требует либо очень большого объема известных данных, либо очень сильных предположений о виде функций распределения, и работает обычно только при нормальных или близких к нормальным функциях распределения. Поэтому при вычислении условного математического ожидания требуется проверять гипотезу о распределении эмпирических данных по нормальному закону или использовать аппарат непараметрической статистики,

восстанавливающей оценки плотностей распределения вероятностей.

Как отмечено в п. I.1, проникновение вычислительной техники в различные области знаний, наличие большого числа пользователей, предоставляющих для обработки на ЭВМ данные научного эксперимента, вызвали появление большого количества разработок по созданию программных комплексов анализа и обработки информации: SPSS, ППСА, ОТЭКС и др. В то же время широкому распространению существующего программного обеспечения препятствует ряд факторов, главными из которых являются:

- отсутствие возможностей системного использования программ из конкретного статистического пакета для решения задач анализа данных;
- сложность использования программного обеспечения для пользователей-непрограммистов;
- сложность в интерпретации полученных результатов обработки данных.

Существование первого и второго факторов объясняется тем, что большинство из существующих программных средств имеют пакетную или библиотечную организацию. Для успешного использования такого программного обеспечения от пользователя требуется хорошее знание как применяемых методов анализа данных, так и языков программирования. С точки зрения пользователя – специалиста в конкретной предметной области и

не имеющего достаточного опыта в использовании ЭВМ, такое программное обеспечение представляет собой неупорядоченное множество алгоритмов, использование которого зависит от правильности выбора схемы обработки данных.

Выходом из подобной ситуации является создание системной организации программного обеспечения по обработке данных. Такая система программ должна представлять собой совокупность процедур, вызов которых осуществляется единой управляющей программой.

В следующих параграфах настоящей главы рассмотрены принципы организации вычислений в системе анализа данных САД, вопросы реализации САД на персональной ЭВМ, а также использование САД в научных и практических исследованиях

## III.1. Принципы организации вычислений в системе анализа данных САД.

Для определения статистической модели изучаемого явления в настоящее время существует большое количество методов и алгоритмов обработки данных. Однако эта модель, как правило, отражает лишь какую-то одну сторону реального явления, оставляя в тени другие не менее важные особенности изучаемого объекта. Для преодоления этого недостатка при применении методов статистического анализа необходимым условием является комплексное использование различных методов, при котором могут быть учтены различные нюансы каждого из используемых подходов к обработке реальных

экспериментальных данных. Методика комплексного применения методов для решения отдельных классов задач статистического анализа данных позволяет получить наиболее достоверные конечные результаты обработки и включает в себя следующее:

- Сравнительный анализ и выбор методов, решающих сходные содержательные задачи;
- Глубокий анализ каждого выбранного метода как средства познания реальных явлений;
- Содержательную интерпретацию результатов применения каждого из методов статистического анализа данных.

Далее описывается система анализа данных САД, в которой сделана попытка реализации описанного выше подхода к обработке реальных экспериментальных данных из различных областей науки и техники. К данной системе были предъявлены следующие основные требования:

- Система должна решать широкий круг задач по анализу данных с использованием минимального набора наиболее эффективных алгоритмов их решения;
- В систему должна быть встроена методика анализа данных, позволяющая исследователю предметной области в ходе обработки осуществить вычислительный эксперимент по проверке своих гипотез и предположений об исследуемом явлении;
- Система должна быть простой в использовании с возможностью обращения к ней на языке предметной области без

применения специальных языков программирования или управления заданиями;

- Автоматическая организация процесса обработки данных и связей с модулями системы;
- Ведение банка данных пользователя и составление отчета о результатах проделанного анализа;
- Диалоговый режим работы пользователя с системой;
- Система должна обладать необходимой скоростью вычислений и представления результатов;
- Результаты обработки данных должны быть представлены в форме, позволяющей произвести содержательную интерпретацию полученных результатов с целью проведения дальнейшего анализа.

Необходимо отметить, что работа над повышением степени интеллектуальности статистического программного обеспечения преследует цель уменьшить ошибки при эксплуатации программ, предоставив пользователю в автоматическом режиме необходимую консультацию по правильной постановке задачи, выбору подходящего статистического инструментария, по умению обойти встречающиеся на пути статистического анализа типичные "ловушки", по правильной интерпретации результатов анализа и т. п. [72-75].

Как отмечено в [76], скорость работы статистического программного обеспечения важна для комфортной эксплуатации и косвенно отражает трудоемкость его разработки. Кроме того,

пакет с высоким быстродействием заметно уменьшает число необходимых персональных компьютеров, а это может вылиться в существенную экономию средств.

Степень интеллектуальности системы в первую очередь предполагает организацию такого режима работы, при котором пользователь имеет достаточно квалифицированное статистическое ассистирование в ходе всего процесса статистического анализа, т. е. при выяснении природы анализируемых данных, при выборе подходящих моделей и методов, их увязывании в технологическую цепочку, при интерпретации результатов и т. п. При этом основные показатели вовсе не обязательно связаны с наличием в системе подходящей экспертной системы. Речь идет о развитой системе компьютерной консультационной поддержки (по статистике), охватывающей различные стадии решения задачи:

 ориентирование пользователя в существующих литературных источниках по применяемым статистическим методам, а также обеспечение его подсказками по используемой терминологии, понятиям, существующим решениям аналогичных задач;

 помощь в постановке задачи, подробный предварительный анализ исходных данных с акцентированием внимания пользователя на их генезисе и особенностях;

 подбор подходящего вида модели и технологической цепочки обрабатывающих модулей;

 описание набора типичных статистических "ловушек" и способов, как их избежать;

 помощь в интерпретации промежуточных и финальных результатов статистического анализа;

 предложение направлений дальнейшего исследования.

В основе организации системы анализа данных САД лежит концепция «вычислительный эксперимент», впервые выдвинутая применительно к решению задач численного анализа [77, 78]. При этом предполагалось, что увеличение достоверности решения задачи может быть достигнуто за счёт использования различных методов численного анализа. Более широко концепция «вычислительный эксперимент» подразумевает предоставление конечному пользователю ЭВМ и её программного обеспечения в качестве некоторой «экспериментальной установки» для испытания гипотез и предложений пользователя об исследуемом процессе или явлении. При этом в обработке данных варьируется не только и не столько их объём, сколько исходные предположения, гипотезы (описание данных, типы признаков, формы группировок и другие трудноформализуемые знания пользователя), что приводит к такому процессу обработки, который предоставляет конечному пользователю статистического программного обеспечения наиболее полную картину исследуемого явления с различных позиций.

На основе этого положения каждый этап обработки данных в системе САД определяется различными исходными

предпосылками и интуитивными знаниями пользователя исследуемой предметной области. Однократное обращение к системе даёт лишь решение при определённых предпосылках, в то время как при изменении гипотез и структуре данных в работу по анализу включаются другие цепочки алгоритмов и методов, которые, взаимно перекрываясь, позволяют пользователю последовательно уточнять свои знания об изучаемых данных и, соответственно, об исследуемом явлении.

Для осуществления такой методики средства и методы обработки в системе должны обеспечивать не только решение основных и наиболее важных задач анализа данных, но и реализацию различных процедур по коррекции и преобразованию исходной информации, что особенно важно в контексте решения задач выбора информативного описания и отбора признаков.

На рис. 3.1 представлены основные этапы разработанной методики интегрированного анализа данных, базирующейся на методике «вычислительный эксперимент» [31, 43, 44].


Рис.3.1. Основные этапы анализа данных.

Перейдём теперь к рассмотрению структурной схемы анализа данных в системе САД, которая представлена на рис.3.2. Решение задачи анализа данных исследуемой предметной области осуществляется последовательно в виде замкнутого цикла, где каждый такой цикл представляет собой этап ведения «вычислительного эксперимента» применительно к обработке экспериментальной информации. После получения конечных результатов обработки и их интерпретации пользователем выбирается режим прохождения цикла, а именно, выбор цели исследования или конкретного метода решения определённой задачи. В результате повторного прохождения цикла осуществляется проверка гипотез и предположений об изучаемом явлении. Таким образом, описанная выше методика обработки данных позволяет решать широкий круг задач статистического анализа данных путём применения минимального количества наиболее эффективных методов и алгоритмов их решения.

В системе САД используются как классические методы анализа количественных данных [6, 18, 40], так и новые методы по исследованию и преобразованию типов признаков неколичественных данных, разработанные в рамках описанного в гл.II подхода [53, 79-83]. Во многих реальных задачах объекты описываются большим числом признаков, что существенно затрудняет процесс вычислений и влечет за собой неоправданные расходы машинного времени. Кроме зтих причин, обуславливающих необходимость сокращения исходного

пространства описания экспериментальных данных, имеет место еще одно важное обстоятельство. Начальное представление о разбиении исходного множества наблюдений на классы можно получить на основе визуализации данных. Однако, решение этой задачи также часто затрудняется из-за наличия большого числа характеристик объекта, без которых, по мнению исследователя предметной области, невозможно получить полное представление об изучаемом явлении. Хотя во многих практических задачах число признаков, определяющих различие объектов разных классов, невелико. Поэтому одной из самых важных задач предварительной обработки данных является задача снижения размерности исходного признакового пространства. Кроме того, в процессе предварительной обработки данных при наличии в ТЭД признаков разных типов возникает необходимость в преобразовании типов признаков. Это связано с тем, что очень часто исходная ТЭД содержит признаки, которые могут быть измерены в различных шкалах.

Как было отмечено в гл. II, определенная шкала допускает вычисление определенного набора статистических характеристик. По этой причине многие традиционно используемые для обработки данных математико-статистические методы, такие как регрессионный,факторный, дискриминантный анализы, оказываются неприменимыми для всего объема разнотипной информации и используются только для обработки количественных данных.

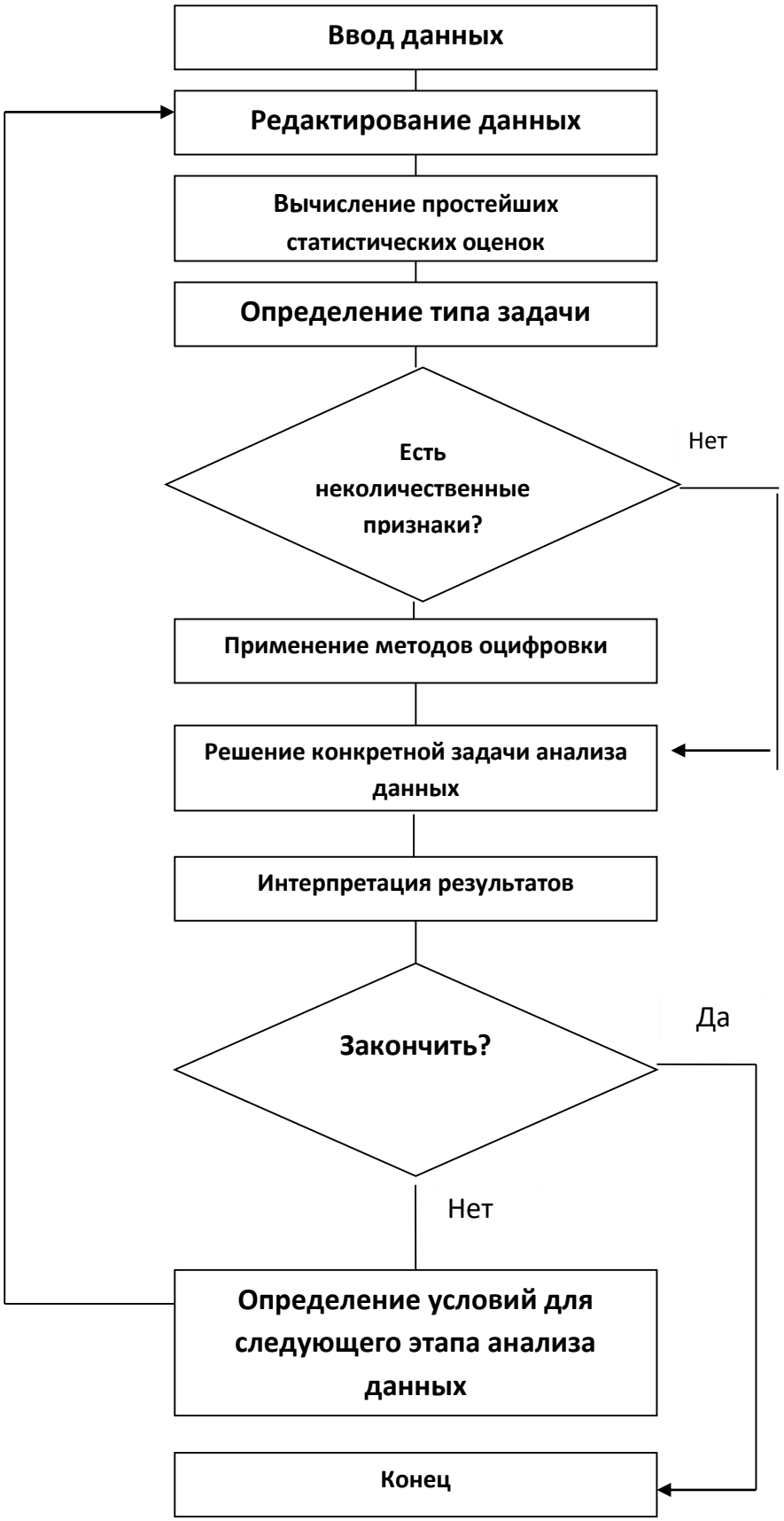


Рис.3.2. Структурная схема анализа данных в системе САД.

Статистический анализ качественных и номинальных признаков основан, в свою очередь, на использовании порядковых статистик и различных мер связи категоризированных переменных. Вследствие этого в большинстве существующих библиотек и систем анализа данных для различных типов признаков используются разные методы статистической обработки. Результаты анализа разнотипных данных по таким признакам интерпретируются независимо друг от друга, что мешает пониманию исследуемого явления как единого целого. Кроме того, могут быть не выявлены те закономерности и свойства изучаемых объектов, которые отражаются в связях между разнотипными показателями. Таким образом, отсутствие при работе с САД возможности совместного анализа разнотипных показателей приводит к тому, что некоторые классы экспериментальных данных вообще не могут быть обработаны на ЭВМ, результаты же, полученные после независимого анализа признаков различных типов, не отражают всего многообразия внутренних связей исследуемого явления и поэтому не представляют практического интереса.

Разработанная САД отвечает следующим требованиям:

- Максимально приближена к пользователям-непрограммистам и не специалистам в области анализа и обработки данных;

- Имеет возможность как автоматического выбора имеющихся в банке алгоритмов методов анализа данных, так и реализации алгоритмических цепочек по запросу пользователя;
- Осуществляет интегрированный подход к анализу данных;
- Имеет в банке алгоритмов статистические методы сжатия и обработки информации для обработки больших массивов данных;
- Имеет возможность работы с архивом данных и промежуточных результатов;
- Имеет возможность редактирования исходных данных, в частности, исключения из ТЭД неинформативных признаков или объектов, заполнения пропусков в ТЭД, выделения по запросу пользователя подтаблиц из исходной ТЭД, применения к выбранным признакам или объектам встроенных функций в целях нормировки или стандартизации данных;
- Имеет возможность определения простейших характеристик исходной информации;
- Осуществляет визуализацию данных;
- Имеет возможность решения основных задач анализа данных: распознавания образов, автоматической классификации, регрессионного и дисперсионного анализа;
- Имеет возможность преобразования типов признаков с целью применения всего арсенала классических методов анализа данных, рассчитанных на однотипные признаки;

- Имеет возможность дополнения банка алгоритмов алгоритмами и программами с целью использования новейших достижений в области прикладной статистики;
- Интегрируется с другими ППП ПС для получения наиболее достоверных конечных результатов;
- Имеет удобную и информативную функцию HELP для уточнения некоторых понятий и определений, используемых в ППП ПС;
- Максимально независима от типов ЭВМ и операционных систем.

Таким образом, в рассматриваемую САД включены методы совместного анализа признаков различного типа, позволяющие выявить для номинальных и ранговых признаков скрытые количественные отношения и перейти к измерению этих признаков в более сильных шкалах. Преобразование типов признаков позволяет использовать для обработки разнотипных данных классические методы анализа данных.

## III.2. Реализация САД на персональной ЭВМ

Рассмотрим способ построения архива данных применительно к интегрированной системе обработки разнотипных данных САД. Понятие архива данных [84] включает в себя две составляющие:

- внешний архив, в котором хранятся файлы данных и результаты решения различных задач;

- внутренний архив – локальный банк данных (ЛБД) – для хранения промежуточных и окончательных результатов по задаче, решаемой в данный момент.

В [84] отмечается, что архив – это важнейшая часть системы обработки данных, через который обмениваются информацией все модули системы и в него попадают все промежуточные результаты анализа. В архиве также хранятся результаты обработки целых задач, решение которых проводилось в несколько сеансов.

Рассмотрим более подробно организацию локального банка данных в системе САД. В целях хранения двумерных массивов информации (исходная таблица экспериментальных данных, матрица корреляций, матрица нагрузок на главные компоненты), представленных вещественными числами, был организован файл *bank1.dat*. Для хранения одномерных массивов вещественных чисел (вектора средних значений и дисперсий, максимальных и минимальных значений по признакам) в системе анализа данных используется файл *bank2.dat*. Соответственно для хранения одномерных массивов целочисленных данных (вектор принадлежности того или иного объекта к определённому классу при решении задачи распознавания образов, вектора целочисленных цифровых меток неколичественных признаков) в системе САД используется файл *bank3.dat*. Все эти вышеописанные файлы образуют ЛБД, обмен с которым

осуществляется посредством процедур *Change, Change1 и Change2* с тремя формальными параметрами:

- режим обмена с ЛБД (ввод или вывод);
- номер массива в ЛБД;
- идентификатор (имя) массива.

Запись, обновление и считывание информации в процедурах *Change* осуществляется стандартными операторами системы быстрой разработки приложений Delphi.

Предлагаемая организация обмена и хранения данных в системе САД обеспечивает простой доступ к различным фрагментам рабочих массивов.

Особое внимание в рассматриваемой системе анализа данных отводится этапу редактирования исходных данных (см. рис.3.2), поскольку существуют самые распространённые ошибки при ручном вводе данных в память ЭВМ [85]:

- ошибки копирования, возникающие из-за неправильного чтения рукописных цифр;
- множественные ошибки копирования, которые появляются при неправильном копировании нескольких цифр в одном и том же числе;
- перестановочные ошибки, возникающие от перестановки соседних цифр;
- ошибки сдвига, появляющиеся от пропуска или вставления цифр. Обычно это бывает, когда имеется достаточно длинный ряд нулей.

Как указано в [85], в 85% случаях ошибок встречаются ошибки копирования, далее идут перестановочные ошибки, двойные перестановочные ошибки.

Исправление ошибок, допущенных при сборе данных, составлении ТЭД и ввода её в ЭВМ, - трудоёмкая и ответственная операция, требующая длительного времени и огромного терпения. Обычно чистка данных занимает до половины всего времени решения задачи предметной области. В связи с этим система САД предоставляет пользователю разнообразные формы распечатки или вывода на экран монитора исходных и промежуточных данных, например:

- распечатать всю исходную ТЭД;
- распечатать подтаблицу из ТЭД, образованную признаками и объектами с указанными номерами;
- распечатать описание всех объектов, у которых значение заданного признака равно (или не равно) указанному;
- распечатать значение заданного признака и объекта с заданным номером;
- распечатать указанную строку исходной ТЭД;
- распечатать указанный столбец исходной ТЭД.

Часто в ТЭД даются контрольные суммы по строкам и столбцам. В этом случае система осуществляет проверку поступивших в ЭВМ исходных данных.

При редактировании исходных данных производятся исправления обнаруженных ошибок или изменения исследуемой

таблицы экспериментальных данных. Редактирование предоставляет следующие возможности:

1. исключить из рассмотрения неинформативные признаки;
2. проанализировать только часть ТЭД или отдельную группу объектов;
3. заменить указанное значение в ТЭД на новое;
4. заменить в указанном признаке все значения, равные заданному числу, на новое;
5. исключить из рассмотрения некоторое значение в указанном месте ТЭД;
6. оставить для дальнейшей обработки (или исключить из обработки) часть таблицы, определяемую объектами или признаками с указанными номерами;
7. Исключить из обработки объекты, у которых значение заданного признака равно (или не равно) указанному числу.

При внесении любых изменений в исходную ТЭД в ходе редактирования в локальном банке данных сохраняется как начальная таблица, так и обновлённая. По окончании работы с обновлённой ТЭД можно в этом же сеансе продолжить работу с исходной таблицей.

Другой важной функцией редактирования исходной информации является преобразование данных, под которой понимается любая замена исходных статистических данных другими, выполненная с целью улучшения, упрощения, удобства при следующей обработке. Обычно преобразование связано с

частичной потерей статистической информации. Число преобразованных данных может быть меньшим, равным или большим числа исходных данных. чаще всего на практике осуществляют преобразование исходного распределения к нормальному распределению. Найти распределение, порождающее нормальное распределение, обычно непросто. Иногда данные сами могут подсказать соответствующую замену. Для некоторых типов переменных используются стандартные замены, например, для некоторых экономических показателей – логарифмическая замена переменных. Иногда тип преобразования может подсказать эмпирическая плотность распределения. Например, резко асимметричная диаграмма с большим правым «хвостом» говорит о логнормальном или хи-квадрат распределении, поэтому в качестве преобразования может использоваться логарифмирование или извлечение квадратного корня. Функции преобразования данных в системе САД выполняются процедурой *RED*.

При взаимодействии с архивом данных лучше всего использовать диалоговый режим работы системы, который позволяет осуществить принцип получения информации в момент возникновения необходимости в ней.

Перейдём теперь к изложению особенностей реализации системы на персональных ЭВМ.

Одной из важнейших и достойных внимания программистов сторон персональных ЭВМ является возможность составления

интерактивных (диалоговых) программ, дающих возможность пользователям наиболее полно использовать возможности систем анализа данных, пакетов прикладных программ и других программных средств статистической обработки данных. Это же обстоятельство позволяет создавать программные продукты, близкие по своим возможностям к экспертным системам, что является конечной целью разработчиков программных систем.

При создании на персональном ЭВМ программного обеспечения, реализующего предложенный в настоящей работе подход к анализу данных разнотипной природы, возникают две основные задачи:

- программная реализация предложенных алгоритмов в виде комплекса процедур;
- организация совместного функционирования этих процедур с целью осуществления методики, предложенной в данной работе.

Для решения первой задачи была выбрана системы быстрой разработки приложений Delphi, которая является алгоритмическим языком программирования высокого уровня, позволяющим работать как с числовой информацией, так и с данными текстового типа. Система быстрой разработки приложений Delphi имеет возможность связи с большинством функционирующих и используемых на сегодняшний день систем управления базами данных, что позволяет извлекать в случае необходимости необходимую информация из этих баз данных.

При необходимости исследователь предметной области имеет возможность работы с системой анализа данных как с библиотекой программ. В этом случае пользователь сам составляет управляющую программу для комплексного решения своей задачи, используя при этом собственную стратегию обработки данных.

Персональные ЭВМ предоставляют широкие возможности для реализации диалогового режима взаимодействия с пользователем. В системе САД диалог реализован в форме меню, причём в самом начале работы пользователю предоставляется возможность выбора либо интерактивного, либо автоматического режимов взаимодействия. В последнем случае обработка производится в соответствии со стратегией, заложенной в системе.

Каждый модуль системы в процессе совместного функционирования системы может использовать результаты работы других модулей. Как было сказано выше, хранение исходной информации и промежуточных результатов осуществляется локальным банком данных.

В рассматриваемой системе важное значение придаётся задачам визуализации исходных данных и снижения размерности исходного пространства описания признаков. В качестве осей координат плоскости, на которую проецируются исходные данные, выбираются пары признаков либо первые два главных компонентов или общих фактора. Здесь необходимо отметить, что поскольку главные компоненты определяются на основе

дисперсии облака объектов, то они оказываются чувствительными к абсолютным значениям признаков. Например, если один из признаков имеет значения от единиц до тысяч, а другие – от нуля до десятков, то независимо от структуры данных первый признак будет отождествляться с первой главной компонентой, поскольку дисперсия вдоль его оси будет наибольшей. Поэтому метод главных компонент применяют для нормированных данных, когда интервалы изменения всех признаков примерно одинаковы.

Для решения задач автоматической классификации и распознавания образов в системе используется несколько процедур, реализующих методы «ближнего соседа», «средней связи» и «дальнего соседа», а также процедуры дискриминантного анализа.

Обобщая сказанное необходимо отметить, что система САД предназначена для решения в пространстве разнотипных признаков задач анализа данных: исследования парных взаимозависимостей, регрессионного анализа, автоматической классификации, распознавания образов, снижения размерности исходного признакового пространства описания. Все процедуры системы оформлены в виде модулей и могут работать автоматически как единая система, либо как самостоятельные модули.

В следующем параграфе иллюстрируется работа системы анализа данных применительно к реальным данным.

## III.3. Использование САД в научных и практических исследованиях

В данном параграфе приводятся результаты обработки данных социологического опроса клиентов Национальной компании экспортно-импортного страхования «Узбекинвест» для уточнения стратегии её развития в вопросах непрерывного улучшения качества и потребительских свойств страховых услуг, совершенствования страховых технологий и освоения новых видов страхования, повышения экономического роста и максимизации прибыли.

Результаты опроса были представлены в виде таблицы экспериментальных данных (ТЭД) «объект-признак». Случайным образом было выбрано 450 анкет. Вопросы анкеты соответствовали 9 разнотипным признакам, представленным в результирующей ТЭД объёмом 450х9 и имели следующее содержание:

- Количественные признаки:
  - “возраст”;
  - “стаж работы”;
  - “какую сумму рассчитываете потратить на страхование”;
- Качественные признаки:
  - “оценка работы сотрудников компании”;
  - “оценка цены страхового полиса «Узбекинвест»”.
- Классификационные признаки:

- “вид страхового полиса”;
- “откуда узнали о деятельности компании”;
- “причина приобретения страхового полиса”;
- “каким образом приобретён страховой полис”.

Для решения данной социологической задачи необходимо было провести анализ статистической взаимозависимости признаков, а также построить различные диаграммы по каждому признаку в отдельности.

При обработке возник ряд проблем, обусловленных необходимостью совместной обработки разнотипных данных.

Рассмотрим процесс обработки вышеописанной таблицы экспериментальных данных. В соответствии с п.I.2 на первом этапе были определены цель исследования, состав данных, осуществлён сбор исходной информации на основе анкетных данных и проведена формализация данных (шкалирование).

На втором этапе исходные данные были введены в ЭВМ и проведена работа с архивом данных. В результате этого действия были восстановлены пропущенные данные и осуществлён визуальный анализ исходной информации. Затем было сформировано задание для обработки, а именно, система анализа данных была настроена на решение задачи анализа статистической взаимозависимости признаков.

На третьем этапе обработки данной задачи был проведён качественный анализ, включающий в себя преобразование типов

признаков и определение простейших характеристик исходных данных.

В работах [85-105] рассмотрена возможность применения системы анализа данных в различных областях науки и техники.

**Выводы по главе III.**

1. Показано, что все задачи классификации и прогнозирования для такой таблицы экспериментальных данных можно свести к четырем классическим постановкам;
2. Приведены факторы, препятствующие широкому распространению существующего программного обеспечения прикладной статистики;
3. Приведены основные требования к разрабатываемым системам анализа данных;
4. Перечислены этапы анализа данных и структурная схема разработанной система анализа данных.

# Заключение

Целью представленной работы являлась разработка методов преобразования типов признаков в задачах анализа данных разнотипной природы, создание системы анализа данных с использованием как классических методов обработки данных, так и предложенных в данной работе методов и алгоритмов.

В работе получены следующие основные результаты:

1. Разработаны методы и алгоритмы преобразования типов признаков для различных задач анализа экспериментальных данных разнотипной природы;

2. Предложен алгоритм поиска императивных шкал порядка градаций неколичественных признаков;

3. Предложены структура системы анализа данных и экспертной системы анализа данных, решающей на уровне специалиста различные задачи анализа данных;

4. Разработаны требования к созданию программных средств прикладной статистики;

5. На основе вышеуказанных требований и разработанных методов преобразования типов признаков, а также классических методов анализа данных создана система анализа данных;

6. Реализована методика обработки экспериментальных данных разнотипной природы с использованием разработанной системы анализа данных в режиме диалогового взаимодействия;

7. Предложенные в диссертационной работе методы, алгоритмы и программное обеспечение были использованы в практике работ при обработке реальных исходных данных в различных научно-исследовательских институтах и производственных объединениях.

# ЛИТЕРАТУРА

1. Гнеденко Б.В. Курс теории вероятностей: Учебник. - Изд. 6-е, перераб. и доп. - М.: Наука, Гл. ред. физ.-мат. лит., 1988. – 448 с.

2. Клейн Ф. Лекции о развитии математики в 19 столетии. Часть 1 М. -Л.:

Объединенное научно-техническое издательство НКТП СССР, 1937. - 432 с.

3. Плошко Б.Г., Елисеева И.И. История статистики: Учеб. пособие. - М.: Финансы и статистика. 1990. - 295 с.

4. Орлов А.И. / Заводская лаборатория. 1997. Т.63. ц 3. С,55-62.

5. Бернштейн С.Н. В сб.: Труды Всероссийского съезда математиков в Москве 27 апреля - 4 мая 1927 г. - М.-Л.: ГИЗ, 1928. С.50-63.

6. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. - М.: ВНИИСтандартизации, 1987. - 64 с.

7. Орлов А.И. / Заводская лаборатория. 1990. Т.56. ч. 3. С.76-83.

8. Суппес П., Зинес Дж. - В сб.: Психологические измерения, -М: Мир,1967. С. 9-110.

9. Пфанцагль И. Теория измерений. – М.: Мир, 1976. 166 с.

10. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. - М.: Мир, 1976. 168 с.

11. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978. 144 с.

12. Орлов А.И. / Заводская лаборатория. 1995. Т.61. ц 5. С.43-51.

13. Орлов А.И. - В сб.: Экспертные оценки. Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. С.17-33.

14. Ларичев О.И., Мошкович Е.М. Качественные методы принятия решений. Вербальный анализ решений. - М.: Наука, 1996. 208 с.

15. Литвак Б.Г. Экспертные оценки и принятие решений. – М.: Патент, 1996. 271 с.

16. Управление большими системами. Материалы международной научно-практической конференции (22-26 сентября 1997 г., Москва, Россия). Общая редакция - Бурков В,Н., Новиков Д.А. - М.: СИНТЕГ, 1997. 432 с.

17. Орлов А.И. Устойчивость в социально-экономических моделях. – М.: Наука, 1979. - 296 с.

18. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. - М.: Финансы и статистика, 1983. – 471 с.

19. Орлов А.И. / Доклады АН СССР. 1974. Т.219. ц 4. С.808-811.

20. Орлов А.И. - В сб.: Вероятностные процессы и их приложения. - М.: МИЭМ, 1989. С.118- 123.

21. Глушков В.М. Основы безбумажной информатики.- М.: Наука, 1987.- 552 с.

22. Сильвестров Д.С. Программное обеспечение прикладной статистики. – М.: Финансы и статистика, 1988. – 240 с.

23. Сергиенко И.В., Парасюк И.Н. Пакеты программ для статистической обработки данных // Численные методы механики сплошной среды. – Новосибирск, 1981. - №3. – Т.12. – С. 114-160.

24. Дайитбегов Д.М. и др. Математическое обеспечение статистической обработки данных. – М.: МЭСИ, 1978. – 135 с.

25. Дайитбегов Д.М., Калмыкова О.В., Черепанов А.И. Программное обеспечение статистической обработки данных. – М.: Финансы и статистика, 1984. – 192 с.

26. Петрович М.Л. Анализ программного обеспечения по прикладной статистике (обзор) // Заводская лаборатория. – 1985. – Т.51. - №10. С. 47-56.

27. Богомолов Н.А., Ламзина Д.В. Системное наполнение библиотеки математической статистики НИВЦ МГУ // Вопросы конструирования библиотек программ. – М.: Изд-во МГУ, 1985. – С. 66-75.

28. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. –М: Финансы и статистика, 1989. 191 с.

29. Загоруйко Н.Г., Ёлкин В.Н., Емельянов С.В., Лбов Г.С. Пакет прикладных программ ОТЭКС. – М: Финансы и статистика, 1986. – 158 с.

30. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. Пакет ППСА. – М: Финансы и статистика, 1986. – 232 с.

31. Алексеев А.И., Калягина Л.В., Нарзуллаев Д.З., Никифоров А.М. Многофункциональная система анализа экспериментальных данных на ЕС ЭВМ // Информационное обеспечение систем автоматизации. Л.: Наука, 1986.- С. 47-55.

32. Александров В.В., Фазылов Ш.Х. Интегрированный подход к анализу данных // Автоматизация данных на основе информационно-вычислительной сети. – Л.: ЛНИВЦ АН СССР, 1985. – С. 4-10.

33. Фазылов Ш.Х., Нарзуллаев Д.З., Жуманазаров С.С. О путях развития программных средств обработки данных // Олий Ўқув юртлари ахбороти. Техника фанлари. – Тошкент, 2000, № 3. – С. 22-25.

34. Хейес-Рот Ф., Уотерман Д., Ленат Д. Построение экспертных систем. – М.: Мир, 1987. – 440 с.

35. Форсайт Р. Экспертные системы. Принципы работы и примеры. – М.: Радио и связь, 1987.- 222 с.

36. Нарзуллаев Д.З., Никифоров А.М. О развитии системы анализа данных до уровня экспертной // Информационные проблемы автоматизации. – Л.: ЛИИАН СССР, 1988. – С. 178-189.

37. Gale W.A., Pregibon D. An Expert System for Regression Analysis // Computer Science and Statistics Proc. Of the 14 Symp. On the Interface. – N.-Y.: Springer – Verlag, 1982. – P.110-117.

38. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. - М.: Финансы и статистика, 1989. - 608 с.

39. Hahn G.J. More Intelligent Statistical Software and Statistical Expert Systems: Future Directions (with Comment by P.F. Villeman and J.W. Tukey) // Amer. Stat. – Vol. 39. 1. - P. 1-16.

40. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. - Л.: Наука, 1983.-208 с.

41. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ: Пер. с англ. – М.: Мир, 1982. – 488 с.

42. Дид Э. Методы анализа данных. – М.: Финансы и статистика, 1985. – 357 с.

43. Алексеев А.И., Нарзуллаев Д.З., Никифоров А.М., Фазылов Ш.Х. Интегрированная система обработки разнотипных данных СИТО-ЕС. Инструкция для пользователя. – Л.: ЛИИАН, 1987. – 21 с.

44. Горский Н.Д., Фазылов Ш.Х. Анализ данных: основные этапы и вычислительный эксперимент. – Ташкент. 1987. – 28 с. Деп. в ВИНИТИ 18 авг. 1987 г., № 6057-В87.

45. Воронин Ю.А. Введение мер сходства и связи для решения геолого-географических задач // ДАН СССР, 1071. Т. 199, № 5. – С. 1011-1014.

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

47. Plackett R.L. The analysis of categorical data. London: Griffints. 1974. 159 p.

48. Миркин Б.Г. Анализ качественных признаков и структур. – М.: Статистика, 1980. – 320 с.

49. Енюков И.С. Методы оцифровки неколичественных переменных // Алгоритмы и программное обеспечение прикладного статистического анализа. – М.: Наука, 1980. – С. 309-315.

50. Загоруйко Н.Г., Ёлкина В.М., Лбов Г.С. Алгоритмы обнаружения эмпирических закономерностей. Новосибирск: Наука, 1985. – 110 с.

51. Мешалкин Л.Д. Присвоение числовых значений качественным признакам // Статистические проблемы управления. Вып. 14. – Вильнюс, 1976. – С.49-55.

52. Хованов И.В. Математические основы теории шкал измерения качества. Л.: ЛГУ, 1983. – 137 с.

53. Никифоров А.М., Фазылов Ш.Х. Методы и алгоритмы преобразования типов признаков в задачах анализа данных. – Ташкент, Фан, 1988. – 131 с.

54. Статистические методы анализа информации в социологических исследованиях / Под ред. Осипова Г.В. М.: Наука, 1979.- 319 с.

55. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973.- 900 с.

56. Крамер Г. Математические методы статистики.- М.: Мир, 1975.- 648 с.

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

58. Миркин Б.Г. Анализ качественных признаков. - М.: Статистика, 1976.- 166 с.

59. Базара М., Шетти К. Нелинейное программирование. Теория и алгоритмы. – М.: Мир, 1982. – 583 с.

60. Горский Н.Д. Оценка элементов корреляционной матрицы для данных, представленных шкалами разных типов // Алгоритмы и системы автоматизации исследований и проектирования. – М.: Наука, 1980. – С. 100-105.

61. Никифоров А.М. Обобщение модели классического регрессионного анализа на неколичественные данные // Препринт ЛНИВЦ АН СССР, № 17. Л., 1981. – 20 с.

62. Райс Дж. Матричные вычисления и математическое обеспечение. Пер. с англ. – М.: Мир, 1984.-320 с.

63. Дуда Р., Харт П. Распознавание образов и анализ сцен. – М: Мир, 1976.-511 с.

64. Распознавание образов в социальных исследованиях. Под ред. Н.Г. Загоруйко,Т.И. Заславской.-Новосибирск: Наука, 1968.-195 с.

65. Загоруйко Н.Г. Методы распознавания и их применение. М.: Сов. Радио, 1972.-207 с.

66. Уилкс С. Математическая статистика. М.: Наука, 1967.-632 с.

67. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1 и 2. М.: Финансы и статистика, 1982.

68. Diday E. et al. E'lements d'analyse de donnes, Paris: Dunod, 1982.-402 p.

69. Рао С.Р. Линейные статистические методы и их применение. М.: Наука, 1968.-547 с.

70. Поспелов Д.А. Ситуационное управление. Теория и практика. М.: Наука, 1986.

71. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды.- М.: Наука, 1976. – 736 с.

72. Айвазян С.А. Программное обеспечение персональных ЭВМ по статистическому анализу данных//Компьютер и экономика: экономические проблемы компьютеризации общества. М.: Наука, 1991. С.91-107.

73. Aivasyan S.A. Model and Method-Oriented Intelligent Software for Statistical Data Analysis System. Berlin: Springer Verlag, 1987. C. 153-158.

74. Прикладная статистика: Классификация и снижение размерности: Справ. изд. М., 1989. 607 с.

75. Айвазян С.А. Интеллектуализированные инструментальные системы в статистике и их роль в построении проблемно-ориентированных систем поддержки принятия решений// Обозрение проблем прикладной математики. Том 4, #2. М.: Наука; Изд-во "ТВП", 1997.

76. С. А. Айвазян, В. С. Степанов. Инструменты статистического анализа данных. Мир ПК, №8, 1997 г.

77. Самарский А.С. Современная прикладная математика и вычислительный эксперимент. – Коммунист, 1983. - №18. – С. 31-42.

78. Жаблон К., Симон Ж.К. Применение ЭВМ для численного моделирования в физике. – М.: Наука, 1983. – 235 с.

79. Абдурахманов М.А., Нарзуллаев Д.З., Фазылов Ш.Х. Развитие методов и программного обеспечения анализа данных применительно к исследованиям жизненного цикла продукции // Материалы I Международной научно-технической конференции «IPI (CALS)-2003. Информационные технологии в управлении жизненным циклом». Санкт-Петербург, 2003.-С.42.

80. Фазылов Ш.Х., Нарзуллаев Д.З., Абдурахманов М.А. О результатах анализа данных социологического опроса клиентов НКЭИС «Узбекинвест». Материалы VIII Санкт-Петербургской Международной конференции «Региональная информатика-2002», часть I. Санкт-Петербург, 2002.-С.163.

81. Абдурахманов М.А., Нарзуллаев Д.З. Об одном подходе к обработке разнотипных данных. Материалы Международной конференции "Инновация-2002". Ташкент, 2002.-С.258.

82. Нарзуллаев Д.З., Фазылов Ш.Х. Организация архива данных в интегрированной системе обработки разнотипных данных на ЕС ЭВМ. – ВИНИТИ, №5379-1387. – 10 с.

83. Джадд Д.Р. Работа с файлами. – М.: Мир, 1982. – 112 с.

84. Орлов А.И. Прикладная статистика XXI в. – Журнал «Экономика XXI века», 2000, No.9, с.3-27.

85. Narzullaev D Z, Shadmanov K K, Baidullaev A S, Rajabov E and Tursunov A T 2021 Automated farm management system in Uzbekistan IOP Conf. Series: Earth and Environmental Science 723 (2021) 032036.

86. Narzullaev, D.Z., Abdurakhmanov, B.A., Baydullaev, A.S., Ilyasov, S.T., Shadmanov, K.K. Transformation of types of signs for a task of the regression analysis. IOP Conference Series: Materials Science and Engineering, 2020, 862(5), 052065

87. Ilhamov, K.S., Narzullaev, D.Z., Ilyasov, S.T., Abdurakhmanov, B.A., Shadmanov, K.K. Model of a turbulent flow of a two-phase liquid with an uneven distributed phase concentration in a horizontal pipe. IOP Conference Series: Materials Science and Engineering, 2021, 1047(1), 012021

88. Baydullaev, A.S., Mamatkulov, Z.U., Samigova, N.H., Shadmanov, K.K., Narzullaev, D.Z. Principles of internet education for students on the subject information technology and mathematical

modeling of processes on the basis of moodle distance learning system. Journal of Physics: Conference Series, 2021, 2001(1), 012024.

89. Ilhamov, K.S., Arifjanov, A.M., Narzullaev, D.Z., Abdurakhmanov, B.A., Shadmanov, K.K. Determination of the formation model of the phase concentration field along the section of the turbulent flow of hydrotransport using information technologies. Journal of Physics: Conference Series, 2021, 2001(1), 012010.

90. Нарзуллаев Д.З., Шадманов К.К., Гулманов М.А. Интеллектуальная модель всесторонней оценки факторов риска состояния здоровья и спортивной формы высококвалифицированных спортсменов // МАТЕРИАЛЫ Республиканской научно-практической конференции «НАУКА И ИННОВАЦИИ В СОВРЕМЕННЫХ УСЛОВИЯХ УЗБЕКИСТАНА», Часть I, г. Нукус, 20 мая 2020 г.- С. 30.

91. Нарзуллаев Д.З., Керимов Ф.А., Умаров В.Д. Факторы риска состояния здоровья и спортивной подготовленности спортсменов // «Ўзбекистон республикасида жисмоний маданият ва спорт тизимининг замонавий асосларини яратиш» мавзусидаги илмий-амалий конференция материаллари.Андижон, 21 апрель 2020 й. – С. 162-165.

92. Нарзуллаев Д.З., Ильясов Ш.Т., Тойчиев А.Х. Информационные системы в спорте // «Ўзбекистон республикасида жисмоний маданият ва спорт тизимининг замонавий асосларини яратиш» мавзусидаги илмий-амалий

конференция материаллари.Андижон, 21 апрель 2020 й. – С. 168-170.

93. Керимов Ф.А., Умаров Д.Х., Нарзуллаев Д.З., Мадрахимов Ш.Ф., Тойчиев А.Х., Шадманов К.К. Информационная модель всесторонней оценки факторов риска состояния здоровья и спортивной формы высококвалифицированных спортсменов. Журнал «Фан-спортга», 2020. №4. - С. 19-28.

94. Нарзуллаев Д.З., Шадманов К.К., Ялгашева Ш.У., Кадиров М.А. Информационные системы в спорте. «Science and education» Scientific journal. ISSN 2181-0842. Volume 1, issue 2. May 2020. – С. 328-335.

95. Нарзуллаев Д.З., Шадманов К.К., Ильясов Ш.Т. Информационные технологии в практической и фундаментальной химии // Международная научно-практическая конференция «Современное состояние фармацевтической отрасли: проблемы и перспективы». 13 ноябр 2020. – С. 17.

96. Нарзуллаев Д.З. Преобразование типов признаков при анализе разнотипных данных в наукометрии // «Innovatsion faoliyatning rivojlantirishda ilmiy-tehnik axbortning o’rni» Xalqaro ilmiy-tehnikaviy anjumani materiallari. Toshkent, 2012. С. 242-245.

97. Нарзуллаев Д.З., Копалов С.У., Шодиев А.А. К вопросу организации электронной реферативной базы патентоспособных результатов интеллектуальной деятельности // Proceedings of the

Tashkent International Innovation Forum. Part 2. Tashkent, 2017. С. 8-11.

98. Нарзуллаев Д.З., Халмурадов Л., Абсаттарова С. Факторы риска состояния здоровья и спортивной подготовленности спортсменов // Научно-педагогические школы в сфере физической культуры и спорта. Материалы Международного научно-практического конгресса, посвященного 100-летию ГЦОЛИФК. Под общей редакцией А.А. Передельского. Москва, 2018. С. 118-122.

99. Керимов Ф.А., Нарзуллаев Д.З., Мадрахимов Ш.Ф. Применение информационно-коммуникационных технологий в спорте // Жисмоний тарбия ва спорт муаммолари. Халкаро илмий-амалий анжумани туплами. 19-20 апрел 2019 йил. Карши, 2019. – С. 262-263.

100. Нарзуллаев Д.З., Султонмуродов Д., Усмонов А., Абрайкулов А., Шадманов К.К. Определение терминов «информационная система», «информационная модель», «автоматизированная система управления» в задачах автоматизации фермерских хозяйств Узбекистана // Передовые инновационные разработки. Перспективы и опыт использования, проблемы внедрения в производство. Сборник научных статей по итогам восьмой международной научной конференции (30 сентября 2019 г.). С. 155-160.

101. Нарзуллаев Д.З., Усмонов А. Применение информационно-коммуникационных технологий в сельском

хозяйстве // Особенности инновационного развития российской науки. Материалы международной научно-практической конференции. Москва-Сочи, 14-21 марта 2019 г. С. 63-66.

102. Нарзуллаев Д.З., Шадманов К.К., Ражабов Э.Э., Самигова Н.Х. Информационные технологии в фармацевтике // Материалы республиканской научно-практической конференции с участием международных ученых «Современное состояние фармацевтической отрасли: проблемы и перспективы», Ташкент, 15-16 ноября 2019 г. С. 38-40.

103. Нарзуллаев Д.З., Шадманов К.К., Усмонов А. Базы данных и базы знаний при автоматизации фермерских хозяйств. Журнал «Теория и практика современной науки».-2019.- №10(52). С. 103-107.

104. Нарзуллаев Д.З., Шадманов К.К., Турсунов А.Т., Каршибоев Ш.У. Системный подход к проблемам математического моделирования // МАТЕРИАЛЫ Республиканской научно-практической конференции «Наука и инновации в современных условиях Узбекистана», Часть I, г. Нукус, 20 мая 2020 г.- С. 43-44.

105. Керимов Ф.А., Нарзуллаев Д.З. Применение статистических методов в спорте. Ташкент: «Фан ва технологиялар», 2014, 388 с.

**НАРЗУЛЛАЕВ Д.З., ТУРСУНОВ А.Т.,
БАЙДУЛЛАЕВ А.С.**

# ОБРАБОТКА РАЗНОТИПНОЙ ИНФОРМАЦИИ В ЗАДАЧАХ АНАЛИЗА ДАННЫХ (Монография)

**Редактор:**

Нарзуллаев Д.З.

**Компьютерная верстка:**

Урманова Д.А.

Изд.лиц. AI№ 283, 11.01.16. Разрешено в печать 2022.
Формат 60×84 $^{1}/_{16}$. Гарнитура “Times New Roman”.
Цифровая печать. Усл.п.л. 9.5. Изд.л 9,75.
Тираж 100 экз. Заказ № 14/03



г. Ташкент. Мирадабский район улица Фаргона йули 222/7