МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
И ПРОДОВОЛЬСТВИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

ГЛАВНОЕ УПРАВЛЕНИЕ ОБРАЗОВАНИЯ, НАУКИ И КАДРОВ

Учреждение образования
«БЕЛОРУССКАЯ ГОСУДАРСТВЕННАЯ
ОРДЕНОВ ОКТЯБРЬСКОЙ РЕВОЛЮЦИИ
И ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
СЕЛЬСКОХОЗЯЙСТВЕННАЯ АКАДЕМИЯ»

В. И. Буць

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Курс лекций
для магистрантов, обучающихся
по специальности 1-25 80 01 Экономика

Горки
БГСХА
2021

УДК 004.896(075.8)
ББК 72я73
Б90

Рекомендовано методической комиссией
экономического факультета 23.10.2019 (протокол № 2)
и Научно-методическим советом БГСХА
27.11.2019 (протокол № 3)

Автор:
доктор экономических наук, доцент *В. И. Буць*

Рецензенты:
доктор экономических наук, профессор *А. Г. Ефименко*;
доктор экономических наук, доцент *Г. О. Читая*

Буць, В. И.

Б90 Технологии интеллектуального анализа данных : курс лекций / В. И. Буць. – Горки : БГСХА, 2021. – 105 с. : ил.
ISBN 978-985-882-110-4.

Отражены основные теоретические и практические направления, что дает возможность получить разносторонние знания о содержании и сущности методической базы технологии интеллектуального анализа данных, современном состоянии и тенденциях развития математического моделирования, программном обеспечении, важных составляющих современных информационных технологий.

Для магистрантов, обучающихся по специальности 1-25 80 01 Экономика.

УДК 004.896(075.8)
ББК 72я73

ISBN 978-985-882-110-4

ВВЕДЕНИЕ

С 1960-х гг. информационно-коммуникационные технологии (ИКТ) последовательно эволюционировали от простых систем обработки файлов до сложных, мощных систем управления базами данных (БД). Исследования в области БД с 1970-х гг. смещались от ранних иерархических и сетевых баз данных к реляционным системам управления базами данных (СУБД), инструментам моделирования данных, а также к вопросам индексирования и организации данных. Пользователи получили гибкий и удобный интерфейс доступа к данным с помощью языков запросов (типа SQL), пользовательские интерфейсы, управление транзакциями и т. п. Технология баз данных, начиная с середины 1980-х гг., характеризовалась популяризацией, широким внедрением и концентрацией исследовательских усилий на новые, все более мощные СУБД. Эффективные методы онлайновой обработки транзакций (*online transaction processing – OLTP*) внесли большой вклад в развитие технологий анализа данных, которые окончательно выделились в отдельную научную дисциплину в начале XXI в.

С интеллектуальным анализом данных (ИАД) тесно связаны два англоязычных термина – *Knowledge Discovery in Databases (KDD)* и *Data Mining.* Они развиваются в рамках направления «бизнес-аналитика» – это инструменты, используемые для преобразования, хранения, анализа, моделирования и доставки информации в ходе работы над задачами, связанными с принятием решений на основе фактических данных. При этом с помощью этих средств лица, принимающие решения, должны при использовании подходящих технологий получать нужные сведения и в нужное время. Термин «*KDD*», что можно перевести как «обнаружение знаний в базах данных», возник в конце 1980-х гг. на основе концепции разведочного анализа данных, предложенной Дж. Тьюки в 1962 г.

Под ним подразумевается не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для обнаружения полезного знания. Данный исследовательский процесс не зависит от предметной области; это набор атомарных операций, комбинируя которые, можно получить нужное решение. *KDD* включает в себя этапы подготовки данных, выбора информативных признаков, очистки, построения моделей, постобработки

и интерпретации полученных результатов.

За построение моделей отвечают методы *Data Mining* – обнаружение сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин был введен Г. Пятецким-Шапиро в 1989 г. Английское словосочетание «Data Mining» не получило устоявшегося перевода на русский язык. В литературе используются следующие варианты перевода: добыча данных, интеллектуальный анализ данных, глубинный анализ данных, посев информации, извлечение данных, интеллектуальный анализ данных. Некоторые исследователи считают неудачными большинство вариантов перевода («добыча данных» – разве добывают данные, а не знания?; «интеллектуальный анализ» – а что тогда «неинтеллектуальный» анализ?) и оперируют прямыми англоязычными терминами.

Методы ИАД с помощью алгоритмов машинного обучения итеративно подбирают модель, в определенном смысле наилучшим образом описывающую исходные данные. В этом смысле машинное обучение близко к непараметрической идентификации, которая предполагает, что нужно в ходе решения определить модель и дать оценку ее параметрам. Реализуется конструктивный подход к построению моделей, базирующийся на индуктивной теории и опирающийся на идею возможности описания данных с использованием рядов примитивов на основе их селекции по определенным критериям. В настоящее время к методам непараметрической идентификации можно отнести большинство методов *Data Mining*.

Целью учебной дисциплины «Технологии интеллектуального анализа данных» является изучение магистрантами теоретических и практических основ формирования аналитических данных посредством выполнения операции очищения локальных баз предприятия, применения статистических методов и других сложных алгоритмов. Задачи учебной дисциплины: выработать у магистрантов способности применять методы математики, теории управления и системного анализа; сформировать навыки использования аналитических, вычислительных и системно-аналитических методов для решения прикладных задач системного анализа экономических процессов.

Учебная дисциплина «Технологии интеллектуального анализа данных» является государственным компонентом и входит в модуль информационных технологий в экономике. Освоение учебной дисципли-

ны базируется на компетенциях, приобретенных ранее обучающимися магистрантами при изучении дисциплин «Высшая математика», «Информатика», «Эконометрика, экономико-математические методы и модели». Учебная дисциплина является основой изучения таких дисциплин, как «Прогнозирование национальной экономики», «Основы информационных технологий». В результате изучения дисциплины магистрант должен закрепить и развить следующую профессиональную компетенцию, предусмотренную в учебном плане МД-25-01-9-19у от 27 марта 2019 г.: УПК-5 – быть способным осуществлять анализ данных для решения экономических, управленческих, научно-исследовательских задач.

1. МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ТЕХНОЛОГИЙ И МОДЕЛЕЙ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

1.1. Сущность интеллектуального анализа данных в агроэкономических исследованиях.

1.2. Основные задачи интеллектуального анализа данных по типам производимой информации.

1.3. Модели интеллектуального анализа данных.

1.1. Сущность интеллектуального анализа данных в агроэкономических исследованиях

Интеллектуальный анализ данных (ИАД) – исследование данных, использующее методы искусственного интеллекта и ориентированное на придание системе свойств искусственного интеллекта [14, 16, 25].

Аналитик имеет дело и с документами, и с табличными значениями, которые также принято называть фактографическими.

Под единичным фактом (E_k) принято понимать описание некоторого события. В формализованном виде для этого применяется следующая запись:

$$E_k = \{a_j, t, x_1, x_2,..., x_m\},$$

где a_j – идентификатор (имя) объекта;

t – время измерения;

x_i – значение i-й характеристики объекта, $i = 1,\ldots, m$.

Интеллектуальный анализ данных выступаетв качестве инструмента принятия оптимальных хозяйственных решений. В настоящее время мы являемся свидетелями активного развития технологии интеллектуального анализа данных (ИАД или Data Mining – извлечение данных), появление которой связано в первую очередь с необходимостью аналитической обработки сверхбольших объемов информации, накапливаемой в современных хранилищах данных. Возможность использования хорошо известных методов математической статистики и машинного обучения для решения задач подобного рода открыло новые возможности перед аналитиками, исследователями, а также теми, кто принимает решения – менеджерами и руководителями. Аналитическая обработка сверхбольших объемов информации может проводиться с использованием комбинирования моделей (рис. 1.1) [12, 23].

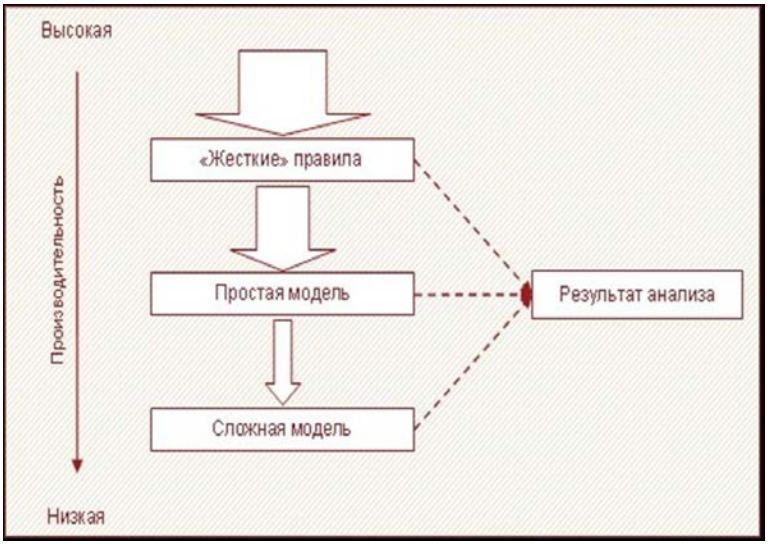

Рис. 1.1. Комбинирование моделей, прогон данных через сито моделей

Взаимосвязи между агроэкономическими событиями выявляются среди больших объемов данных: состояние окружающей среды; состояние почвенного покрова; состояние растений и животных; состояние материально-технической базы; экономическое и финансовое состояние [12, 27, 44].

1.2. Основные задачи интеллектуального анализа данных по типам производимой информации

Задача классификации заключается в том, что для каждого варианта определяется категория или класс, которому он принадлежит. В качестве примера можно привести оценку кредитоспособности потенциального заемщика: назначаемые классы здесь могут быть «кредитоспособен» и «некредитоспособен». Необходимо отметить, что для решения задачи необходимо, чтобы множество классов было известно заранее и было бы конечным и счетным [25].

Прогнозирование новых значений агроэкономических данных имеет следующие особенности:

1. Мир аграрной экономики населен агропродовольственными си-

стемами (АПС), эксплуатирующими продуктивные свойства растений и животных, использующими земельные и потребляющими расходуемые ресурсы, труд и знания, выпускающими сельскохозяйственную продукцию и предоставляющими услуги, формирующими социальную среду на селе; АПС образуют виды, отличающиеся ареалом распространения, организационноправовыми формами, специализацией, размером и другими важными признаками.

2. Учитываются имеющиеся тенденции (тренды), сезонность, другие факторы.

3. Методика прогнозирования основывается на одновременном использовании комплекса методов: трендового, адаптивного, корреляционно-регрессионного и имитационного моделирования. Это позволит получить достоверные и многовариантные результаты.

1.3. Модели интеллектуального анализа данных

Комплекс моделей интеллектуального анализа данных, составляющих структуру учебного цикла, представлен на рис. 1.2.

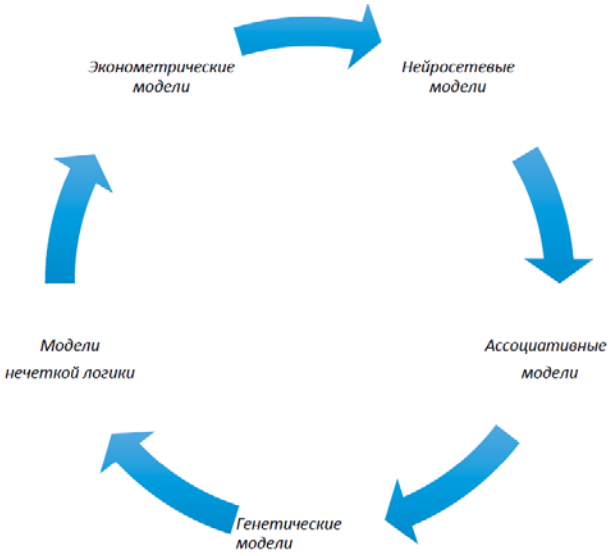

Рис. 1.2. Учебный цикл моделей интеллектуального анализа данных

Эконометрические модели.

Регрессия – один из двух методов прогнозирования. Данный метод использует имеющиеся фактические значения величин для прогнозирования будущих на основании трендов и имеющейся статистики [8].

Различие между регрессией и временными рядами состоит в том, что временные ряды предсказывают значения переменных, зависящих от времени. Время в данном случае может содержать иерархии (рабочая неделя, календарная неделя, период, праздники, сезоны, интервалы дат).

Нейросетевые модели.

Нейросетевые технологии в настоящее время предоставляют широкие возможности для решения задач прогнозирования, обработки сигналов и распознавания образов. По сравнению с традиционными методами математической статистики, классификации и аппроксимации эти технологии обеспечивают достаточно высокое качество решений при меньших затратах. Они позволяют выявлять нелинейные закономерности в сильно зашумленных неоднородных данных, дают хорошие результаты при большом числе входных параметров и обеспечивают адекватные решения при относительно небольших объемах данных.

Ассоциативные модели.

Задача ассоциативной модели заключается в том, чтобы для каждого условия, состоящего из подмножества значений входных признаков заданного наблюдения, сформировать следствие из набора доступных переменных класса, такое, чтобы результирующее правило имело максимальную поддержку или достоверность.

Генетические модели.

Генетический алгоритм представляет собой метод, отражающий естественную эволюцию методов решения проблем, и в первую очередь задач оптимизации. Генетические алгоритмы – это процедуры поиска, основанные на механизмах естественного отбора и наследования. В них используется эволюционный принцип выживания наиболее приспособленных особей. Они отличаются от традиционных методов оптимизации несколькими базовыми элементами. В частности, генетические алгоритмы выполняют следующие функции (рис. 1.3).

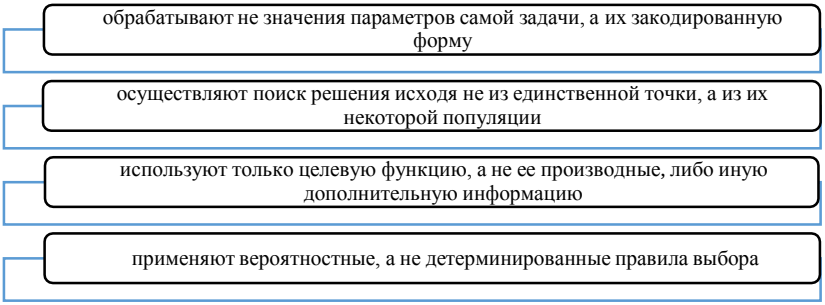

Рис. 1.3. Функции генетических алгоритмов в интеллектуальном анализе данных

Модели нечеткой логики.

Нечеткая система – это система, особенностью описания которой является:

нечеткая спецификация параметров;

нечеткое описание входных и выходных переменных системы;

нечеткое описание функционирования системы на основе продукционных «ЕСЛИ…, ТО…» правил.

Нечеткий логический вывод – это аппроксимация зависимости «входы…выход» на основе лингвистических высказываний типа «ЕСЛИ…ТО» и операций над нечеткими множествами.

2. ПРОЦЕСС ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

2.1. Определение и сущность процесса интеллектуального анализа данных.

2.2. Источники агроэкономических данных.

2.3. Построение структуры агроэкономических данных.

2.1. Определение и сущность процесса интеллектуального анализа данных

Интеллектуальный анализ данных (Data Mining) – вычислительный процесс обнаружения закономерностей в больших объемах данных с участием методов на пересечении искусственного интеллекта, машинного обучения, статистики и баз данных. Общая цель процесса интел-

лектуального анализа данных – извлечение знаний из набора данных и преобразование их в понятую для дальнейшего использования структуру [33]. Термин является модным словом, и им часто злоупотребляют для обозначения любой формы больших объемов данных или обработки информации (сбор, добыча, хранение, анализ и вычисление статистических характеристик). Данный термин также ассоциируют с любой компьютерной системой поддержки принятия решений, в том числе искусственного интеллекта, машинного обучения и бизнес-аналитики. Этот термин следует использовать в тех случаях, когда имеет место «обнаружение чего-нибудь нового».

Традиционно выделяются следующие этапы в процессе интеллектуального анализа данных:

1. Изучение предметной области, в результате чего формулируются основные цели анализа.

2. Сбор данных.

3. Предварительная обработка данных:

а) очистка данных – исключение противоречий и случайных «шумов» из исходных данных;

б) интеграция данных – объединение данных из нескольких возможных источников в одном хранилище;

в) преобразование данных. На данном этапе данные преобразуются в форму, подходящую для анализа. Часто применяются агрегация данных, дискретизация атрибутов, сжатие данных и сокращение размерности.

4. Анализ данных. В рамках данного этапа применяются алгоритмы интеллектуального анализа с целью извлечения паттернов.

5. Интерпретация найденных паттернов. Данный этап может включать визуализацию извлеченных паттернов, определение действительно полезных паттернов на основе некоторой функции полезности.

6. Использование новых знаний.

Однако процесс интеллектуального анализа данных может отличаться в зависимости от того, данные какой области человеческой деятельности мы анализируем. Межотраслевой стандартный процесс интеллектуального анализа данных (Crisp-DM) определяет шесть этапов (рис. 2.1).

Рис. 2.1. Этапы интеллектуального анализа данных

На первом, втором и третьем этапах выполняется осмысление поставленной задачи и уточнение целей, которые должны быть достигнуты методами Data Mining. Важно правильно сформулировать цели и выбрать необходимые для их достижения методы, так как от этого зависит дальнейшая эффективность всего процесса.

Четвертый этап – это, собственно, применение методов Data Mining. Сценарии этого применения могут быть самыми различными и включать сложную комбинацию разных методов, особенно если используемые методы позволяют проанализировать данные с разных точек зрения.

Следующий этап – проверка построенных моделей. Очень простой и часто используемый способ заключается в том, что все имеющиеся данные, которые необходимо анализировать, разбиваются на две группы. Как правило, одна из них большего размера, другая – меньшего. На большей группе, применяя те или иные методы Data Mining, получают модели, а на меньшей – проверяют их. По разнице в точности между тестовой и обучающей группами можно судить об адекватности построенной модели.

Последний этап – интерпретация полученных моделей человеком в целях их использования для принятия решений, добавление получившихся правил и зависимостей в базы знаний и т. д. Этот этап часто подразумевает использование методов, находящихся на стыке технологии Data Mining и технологии экспертных систем. От того, насколько эффективным он будет, в значительной степени зависит успех решения поставленной задачи.

2.2. Источники агроэкономических данных

Сведения, представленные в отчетности сельскохозяйственных организаций, создают комплексную и взаимоувязанную картину о деятельности организаций сельского хозяйства, ее масштабах, направлениях, об активах, капитале и обязательствах предприятия. Все эти данные образуют обширную базу для проведения экономического, в том числе финансового, анализа, охватывающего всю деятельность исследуемой организации и ее источники [6, 12].

Исследование бухгалтерской отчетности сельскохозяйственных организаций позволило по каждой форме выделить цель проведения анализа и определить процедуры, которые необходимо выполнить для ее достижения.

Источники могут быть разного характера, к основным из них относятся: источники учета; внеучетные источники; плановые и нормативные источники; прочие источники информации. *Учетные источники информации* включают в себя бухгалтерскую и налоговую отчетность компании, внутреннюю управленческую отчетность, данные оперативного учета и отдельно выбранные данные. Под *внеучетными источниками* информации понимают данные, полученные из официальных документов: законов, нормативно-правовых актов, договоров и соглашений, судебных решений и т. п. Информация из документов вневедомственных ревизий, проведения аудита, материалы, полученные в ходе налоговых проверок, материалы периодически проводимых совещаний руководства компании, плановые источники включают в себя разного рода финансовые планы, бизнес-планы, прогнозные расчеты.

Система экономической информации – важная составная часть хозяйственного руководства – основа анализа. Имеется двоякая связь предмета экономического анализа с информацией.

С одной стороны, анализ выступает как потребитель информации, с

другой – как ее источник. В переработанном и обобщенном виде информация используется руководителями и специалистами для принятия управленческих решений. Кроме того, предмет анализа выполняет также контрольные и распорядительные функции по отношению к самой информации, вызывая изменение ее объема, содержания, качества, адреса и цели использования [43].

До настоящего времени в развитии двух смежных наук – бухгалтерского учета и экономического анализа – ведущая роль принадлежала бухгалтерскому учету. Экономистам-аналитикам приходилось буквально «подстраиваться» под те сведения, которые заключали в себе различные виды учетной документации. Содержание анализа во многом определялось организацией учета, применяемыми в нем формами, счетами, выходными данными. Однако в действительности главная задача двух смежных наук заключается в определении путей повышения эффективности производства на основании подготовки и изучения данных о деятельности предприятия, и приоритет в ее решении принадлежит экономическому анализу. Бухгалтерский учет должен поставлять информацию, необходимую для аналитической работы, «подстраиваться» при этом под нужды анализа. Это позволит не только успешнее вести аналитическую работу, но и избежать перенасыщенности учетных документов теми сведениями, которые имеют для анализа второстепенное значение или вообще не имеют никакого значения.

2.3. Построение структуры агроэкономических данных

При проектировании информационной базы данных необходимо решить, какие атрибуты будут служить в качестве первичного ключа для каждой таблицы, если таковые будут. Первичный ключ (PK) – это уникальный идентификатор для данного объекта. С его помощью вы можете выбрать данные конкретного клиента, даже если знаете только это значение [3, 10].

Атрибуты, выбранные в качестве первичных ключей, должны быть уникальными, неизменяемыми и для них не может быть задано значение NULL (они не могут быть пустыми).

Теперь, когда данные преобразованы в таблицы, нужно проанализировать связи между ними. Сложность базы данных определяется количеством элементов, взаимодействующих между двумя связанными таблицами. Определение сложности помогает убедиться в том, что вы разделили данные на таблицы наиболее эффективно.

Каждый объект может быть взаимосвязан с другим с помощью одного из трех типов связи (рис. 2.2).

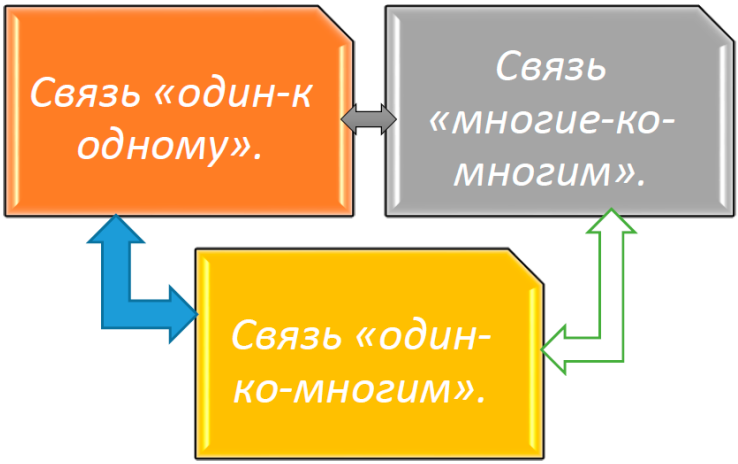

Рис. 2.2. Виды связей между объектами

Связь «один-к-одному». Когда существует только один экземпляр объекта A для каждого экземпляра объекта B, то говорят, что между ними существует связь «один-к-одному» (часто обозначается 1:1).

Связь «один-ко-многим». Эта связь возникает тогда, когда запись в одной таблице связана с несколькими записями в другой.

Связь «многие-ко-многим». Когда несколько объектов одной таблицы могут быть связаны с несколькими объектами другой, то говорят, что они имеют связь «многие-ко-многим» (M:N).

После предварительного проектирования базы данных можно применить правила нормализации, чтобы убедиться, что таблицы структурированы правильно.

В то же время не все базы данных необходимо нормализовать. В целом базы с обработкой транзакций в реальном времени (OLTP) должны быть нормализованы.

Базы данных с интерактивной аналитической обработкой (OLAP) позволяющие проще и быстрее выполнять анализ данных, могут быть более эффективными с определенной степенью денормализации. Основным критерием здесь является скорость вычислений. Каждая форма или уровень нормализации включает правила, связанные с нижними формами.

3. ОСНОВНЫЕ ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

3.1. Оперативный анализ данных посредством OLAP-систем.
3.2. Прогнозирование показателей.
3.3. Интеллектуальный анализ текстовой информации.

3.1. Оперативный анализ данных посредством OLAP-систем

Разработка, построение и внедрение хранилищ данных – это дорогостоящая и трудоемкая задача, которая зависит:

- от уровня информатизации бизнес-процессов компании;
- объема и структуры данных;
- требования к скорости выполнения запроса;
- характера решаемых аналитических задач.

Сокращение сроков построения по имеющимся заданиям можно достичь посредством различных типов OLAP (рис. 3.1) [4, 13].

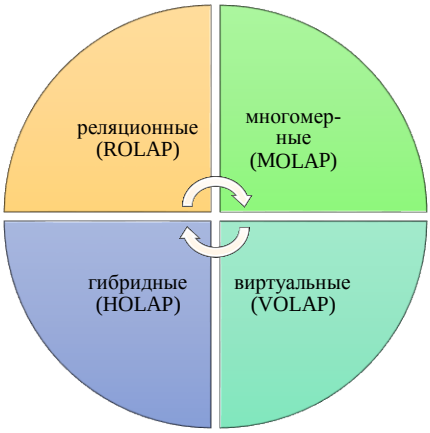

Рис. 3.1. Типы OLAP

Реляционные хранилища данных (РХД) используют классическую реляционную модель, характерную для оперативных регистрирующих OLTP-систем. Данные хранятся в реляционных таблицах, но образуют специальные структуры, эмулирующие многомерное представление

данных. Такая технология обозначается аббревиатурой ROLAP – Relational OLAP.

На логическом уровне различают две схемы построения РХД – «звезда» и «снежинка».

MOLAP – реализует многомерное представление данных на физическом уровне в виде многомерного куба.

Основное назначение многомерных хранилищ данных (МХД) – поддержка систем, ориентированных на аналитическую обработку данных, поскольку такие хранилища лучше справляются с выполнением сложных нерегламентированных запросов.

Многомерная модель данных, лежащая в основе построения многомерных хранилищ данных, опирается на концепцию многомерных кубов, или гиперкубов. Они представляют собой упорядоченные многомерные массивы, которые также часто называют OLAP-кубами (аббревиатура OLAP расшифровывается как On-Line Analytical Processing – оперативная аналитическая обработка). Технология OLAP представляет собой методику оперативного извлечения нужной информации из больших массивов данных и формирования соответствующих отчетов.

Если данные, поступающие с OLTP-систем, имеют большой объем, высокую степень детализации, а для анализа используются в основном обобщенные данные, то используют гибридные хранилища данных (ГХД).

Главным принципом построения ГХД является то, что детализированные данные хранятся в реляционной структуре (ROLAP), которая позволяет хранить большие объемы данных, а агрегированные – в многомерной (MOLAP), позволяющей увеличить скорость выполнения запросов (поскольку при выполнении аналитических запросов уже не требуется вычислять агрегаты).

Виртуальным хранилищем данных называется система, которая работает с разрозненными источниками данных и эмулирует работу обычного хранилища данных, извлекая, преобразуя и интегрируя данные непосредственно в процессе выполнения запроса [35, с. 156, 163, 168].

В основе многомерного представления данных лежит их разделение на две группы – измерения и факты. Измерения – это категориальные атрибуты, наименования и свойства объектов, участвующих в некотором бизнес-процессе. Значениями измерений являются наименования товаров, названия фирм-поставщиков и покупателей, Ф. И. О.

людей, названия городов и т. д. Измерения могут быть и числовыми, если какой-либо категории (например, наименованию товара) соответствует числовой код, но в любом случае это данные дискретные, т. е. принимающие значения из ограниченного набора. Измерения качественно описывают исследуемый бизнес-процесс.

Факты – это данные, количественно описывающие бизнес-процесс, непрерывные по своему характеру, т. е. они могут принимать бесконечное множество значений. К ним относятся: цена товара или изделия, их количество, сумма продаж или закупок, зарплата сотрудников, сумма кредита, страховое вознаграждение и т. д.

3.2. Прогнозирование показателей

Совершенствование алгоритмов интеллектуального анализа данных позволяет решать задачи прогнозной аналитики более эффективными способами. Ансамбли моделей – одно из активно развивающихся направлений, особенно в тех задачах, где прогностическая точность более важная, чем интерпретируемость модели.

Задача прогнозирования требует тщательного исследования исходного набора данных и методов, подходящих для анализа. Она включает решение таких подзадач, как выбор модели прогнозирования, анализ точности построенного прогноза. Ансамблевые модели позволяют сочетать прогнозы нескольких базовых моделей с целью уменьшения ошибок прогнозирования и повышения обобщающей способности моделей [2, 7, 18].

Прогнозирование определяет наиболее общие показатели перспективного развития, выявляет тенденции и альтернативные пути этого развития. Методы математической статистики используются главным образом для проверки заранее сформулированных гипотез, а также для анализа, составляющего основу оперативной аналитической обработки данных, что является недостаточным условием для полной оценки прогнозирования. Для обработки слабоструктурированных данных необходимо применение специфических методов интеллектуального анализа данных.

Основной проблемой при решении задачи прогнозирования является получение разумно точных прогнозов для будущих данных при анализе имеющейся информации. Если методы интеллектуального анализа данных не обеспечивают достаточной точности прогнозирования, эффективной альтернативой использованию методов прогнозирования является внедрение ансамблей моделей.

Под *прогнозированием* обычно понимается моделирование непрерывных значений, в отличие от задачи классификации, связанной с получением дискретных прогнозов.

Набор данных обычно состоит из векторов признаков, где каждый вектор представляет собой описание объекта с использованием набора показателей.

В Data Mining категориальные и числовые величины обрабатываются при помощи соответствующих алгоритмов.

Задачи классификации и прогнозирования сводятся к определению значения зависимой переменной объекта по его независимым переменным. Если зависимая переменная принимает количественные значения, то говорят о задаче прогнозирования, в противном случае – о задаче классификации. Задача прогнозирования может считаться одной из наиболее сложных задач Data Mining, она требует тщательного исследования исходного набора данных и методов, подходящих для анализа. При создании алгоритмов машинного обучения разработчики сталкиваются с такими проблемами, как вычислительные затраты на реализацию алгоритма, неясность построенных моделей для пользователя, а также неточность результатов. Большинство исследователей сосредоточиваются на повышении точности прогнозирования, поэтому оценка моделей часто рассматривается именно с этой точки зрения.

Эффективной альтернативой использованию единственного метода прогнозирования является объединение прогнозов из нескольких разных моделей. Комбинация нескольких прогнозов в большом количестве случаев существенно снижает общие ошибки прогнозирования, превосходя отдельные компоненты [30, 32].

3.3. Интеллектуальный анализ текстовой информации

Современные достижения в области информационных технологий позволили за короткий промежуток времени скопить в хранилищах данных различных организаций большие объемы информации, которая содержит скрытую информацию в виде знаний, поэтому задача аналитической обработки больших объемов информации становится весьма актуальной. Центральное место в процессах интеллектуального анализа естественно-языковой информации занимают следующие технологии: Data Mining, Text Mining и Semantic Web Ontology (язык OWL – Ontology Web Language), задачей которых является получение ранее неизвестных либо не выявленных знаний и закономерностей фактов в

больших хранилищах данных. Источниками исходной информации для аналитической обработки могут являться базы знаний разных типов. Например, огромнейшая база знаний Internet, хранилища данных различных организаций. Значительная часть информации в этих источниках представлена в виде естественных текстов, процесс аналитической обработки которых требует создания принципиально новых моделей, методик и систем интеллектуального анализа информации. Задача аналитической обработки естественных текстов является достаточно сложной и в общем случае связана с построением сложных интеллектуальных информационных систем (ИИС). Однако необходимо взять во внимание тот факт, что информационно-аналитические структурные элементы и компоненты образовательного процесса гуманитарных вузов не нуждаются в извлечении всех закономерностей из естественных текстов в силу специфичности познавательной деятельности студентов. Вследствие чего можно сделать вывод о том, что необходимость построения модели естественного текста, реализующей глубинный семантический анализ текста, не имеет смысла. Поэтому первоочередной задачей интеллектуального анализа текстовой информации в базах знаний экспертной системы является создание унифицированного формата метазнаний [26, 28].

4. СТАТИСТИЧЕСКИЕ МЕТОДЫ

4.1. Дескриптивный анализ и описание исходных данных.

4.2. Анализ связей (корреляционный и регрессионный, факторный и дисперсионный).

4.3. Анализ временных рядов.

4.1. Дескриптивный анализ и описание исходных данных

Дескриптивные (описательные) статистики являются базовым и наиболее общим методом анализа данных. Наряду с частотами, дескриптивный анализ предполагает расчет различных описательных статистик. Соответствуя своему названию, они предоставляют основную информацию о полученных данных. Уточним, использование конкретной статистики зависит от того, в каких шкалах представлена исходная информация. Номинальная шкала используется для фиксации объектов, не имеющих ранжированного порядка (пол, место жительства, предпочитаемая марка и т. д.). Для подобного рода массива

данных нельзя рассчитать какие-либо значимые статистические показатели, кроме моды – наиболее часто встречающегося значения переменной. Несколько лучше в плане анализа ситуация обстоит с порядковой шкалой. Здесь становится возможным, наряду с модой, расчет медианы – значения, разбивающего выборку на две равные части. Наиболее богатыми на все возможные статистики являются количественные шкалы, которые представляют собой ряды числовых значений, имеющих равные интервалы между собой и поддающихся измерению. В данном случае становятся доступными следующие информационные меры: среднее, размах, стандартное отклонение, стандартная ошибка среднего. Конечно, язык цифр является довольно «сухим» и для многих весьма непонятным. По этой причине дескриптивный анализ дополняется визуализацией данных путем построения различных диаграмм и графиков, как, например, гистограмм, линейных, круговых или точечных диаграмм [1, 22].

Таблицы сопряженности – это средство представления распределения двух переменных, предназначенное для исследования связи между ними. Таблицы сопряженности можно рассматривать как частный тип дескриптивного анализа. В них также является возможным представление информации в виде абсолютных и относительных частот, графическая визуализация в виде гистограмм или точечных диаграмм. Наиболее эффективно таблицы сопряженности проявляют себя в определении наличия взаимосвязи между номинальными переменными.

Для более точного выявления наличия связи между переменными используют разные статистические критерии. Наиболее часто применяются такие, как: критерий хи-квадрат (χ^2); коэффициент сопряженности; критерий лямбда; коэффициент R Спирмена; критерий корреляции Пирсона и др. [19, 20, 36, 42].

4.2. Анализ связей (корреляционный и регрессионный, факторный и дисперсионный)

Исследователя нередко интересует, как связаны между собой две или большее количество переменных в одной или нескольких изучаемых выборках. *Корреляционные связи* – это вероятностные изменения, которые можно изучать только на представительных выборках методами математической статистики. Корреляционная связь и корреляционная зависимость часто используются как синонимы. Зависимость подразумевает влияние, связь – любые согласованные изменения, ко-

торые могут объясняться сотнями причин. Корреляционные связи не могут рассматриваться как свидетельство причинно-следственной зависимости, они свидетельствуют лишь о том, что изменениям одного признака, как правило, сопутствуют определенные изменения другого. Корреляционная зависимость – это изменения, которые вносят значения одного признака в вероятность появления разных значений другого признака.

Методы регрессионного анализа рассчитаны главным образом на случай устойчивого нормального распределения, в котором изменения от опыта к опыту проявляются лишь в виде независимых испытаний. Выделяются различные формальные задачи регрессионного анализа. Они могут быть простыми или сложными по формулировкам, математическим средствам и трудоемкости [8, 9]. Первая задача – выявить факт изменчивости изучаемого явления при определенных, но не всегда четко фиксированных условиях. Ранее мы уже решали эту задачу с помощью параметрических и непараметрических критериев. Вторая задача – выявить тенденцию как периодическое изменение признака. Сам по себе этот признак может быть зависим или не зависим от переменной-условия (он может зависеть от неизвестных или неконтролируемых исследователем условий). Но это не важно для рассматриваемой задачи, которая ограничивается лишь выявлением тенденции и ее особенностей [11].

Метод дисперсионного анализа для связанных выборок применяется в тех случаях, когда исследуется влияние разных градаций фактора или разных условий на одну и ту же выборку испытуемых данных. Градаций фактора должно быть не менее трех. Главными целями *факторного анализа* являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т. е. классификация переменных. Поэтому факторный анализ используется или как метод сокращения данных, или как метод классификации. Для выявления наиболее значимых факторов и, как следствие, факторной структуры, наиболее оправданно применять *метод главных компонентов*. Суть данного метода состоит в замене коррелированных компонентов некоррелированными факторами.

4.3. Анализ временных рядов

В отличие от анализа случайных выборок, анализ временных рядов основывается на предположении, что последовательные значения в

файле данных наблюдаются через равные промежутки времени (тогда как в других методах нам не важна и часто не интересна привязка наблюдений ко времени) [5]. Существуют две основные цели анализа временных рядов: определение природы ряда и прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям). Обе эти цели требуют, чтобы модель ряда была идентифицирована и, более или менее, формально описана. Как только модель определена, вы можете с ее помощью интерпретировать рассматриваемые данные (например, использовать в вашей теории для понимания сезонного изменения цен на товары, если занимаетесь экономикой). Как и большинство других видов анализа, анализ временных рядов предполагает, что данные содержат систематическую составляющую (обычно включающую несколько компонент) и случайный шум (ошибку), который затрудняет обнаружение регулярных компонент. Большинство методов исследования временных рядов включает различные способы фильтрации шума, позволяющие увидеть регулярную составляющую более отчетливо.

Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей. *Тренд* представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени. *Сезонная составляющая* – это периодически повторяющаяся компонента. Оба эти вида регулярных компонент часто присутствуют в ряде одновременно. Не существует «автоматического» способа обнаружения тренда во временном ряде. Однако если тренд является монотонным (устойчиво возрастает или устойчиво убывает), то анализировать такой ряд обычно нетрудно. Если временные ряды содержат значительную ошибку, то первым шагом выделения тренда является сглаживание. Периодическая и сезонная зависимость (сезонность) представляет собой другой общий тип компонент временного ряда.

Методы анализа временных рядов нередко делят на два класса: анализ в частотной области и анализ во временной области. Первый основывается на спектральном анализе и с недавних пор вейвлетном анализе, и может рассматриваться в качестве не использующих модели методов анализа, хорошо подходящих для исследований на этапе разведки. Методы анализа во временной области также имеют безмодельное подмножество, состоящее из кросс-корреляционного анализа и автокорреляционного анализа, но именно здесь появляются частично и полностью определенные модели временных рядов.

5. НЕЙРОСЕТЕВЫЕ МОДЕЛИ

5.1. Нейросетевые модели программного и аппаратного исполнения.

5.2. Нахождение наилучшего приближения функции, заданной конечным набором входных значений.

5.3. Виды нейронных сетей.

5.1. Нейросетевые модели программного и аппаратного исполнения

Нейронные сети (Neural Networks) – это модели биологических нейронных сетей мозга, в которых нейроны имитируются относительно простыми, часто однотипными, элементами (искусственными нейронами). Нейронная сеть может быть представлена направленным графом с взвешенными связями, в котором искусственные нейроны являются вершинами, а синаптические связи – дугами. Нейронные сети широко используются для решения разнообразных задач [6].

Среди областей применения нейронных сетей – автоматизация процессов распознавания образов, прогнозирование, адаптивное управление, создание экспертных систем, организация ассоциативной памяти, обработка аналоговых и цифровых сигналов, синтез и идентификация электронных цепей и систем.

С помощью нейронных сетей можно, например, предсказывать объемы продаж изделий, показатели биржевого рынка, выполнять распознавание сигналов, конструировать самообучающиеся системы.

Модели нейронных сетей могут быть программного и аппаратного исполнения. Мы будем рассматривать сети первого типа.

Если выражаться простым языком, *слоистая нейронная сеть* представляет собой совокупность нейронов, которые составляют слои. В каждом слое нейроны между собой никак не связаны, но связаны с нейронами предыдущего и следующего слоев. Информация поступает с первого на второй слой, со второго на третий и т. д.

Среди задач Data Mining, решаемых с помощью нейронных сетей, будем рассматривать следующие (рис. 5.1) [45].

Классификация (обучение с учителем). Примеры задач классификации: распознавание текста, распознавание речи, идентификация личности.

Прогнозирование. Для нейронной сети задача прогнозирования мо-

жет быть поставлена таким образом: найти наилучшее приближение функции, заданной конечным набором входных значений (обучающих примеров). Например, нейронные сети позволяют решать задачу восстановления пропущенных значений.

Кластеризация (обучение без учителя). Примером задачи кластеризации может быть задача сжатия информации путем уменьшения размерности данных. Задачи кластеризации решаются, например, самоорганизующимися картами Кохонена.

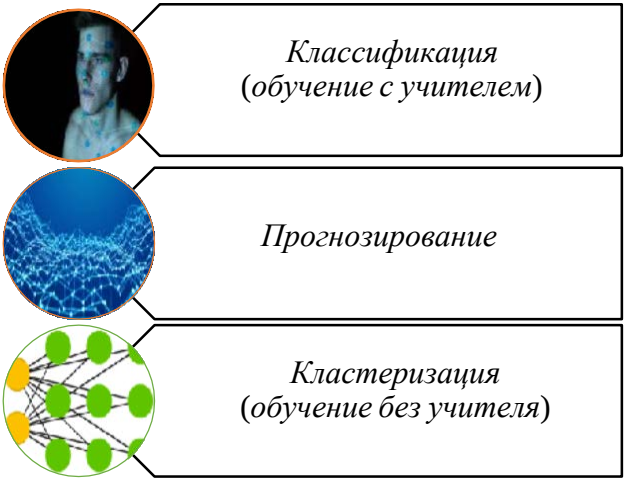

Рис. 5.1. Задачи интеллектуального анализа данных с помощью нейронных сетей

5.2. Нахождение наилучшего приближения функции, заданной конечным набором входных значений

Машинное обучение – это процесс, в результате которого машина (компьютер) способна показывать поведение, которое в нее не было явно заложено (запрограммировано). В общем случае задача обучения нейронной сети сводится к нахождению некой функциональной зависимости $Y = F(X)$, где X – входной, а Y – выходной векторы. В общем случае такая задача, при ограниченном наборе входных данных, имеет бесконечное множество решений. Для ограничения пространства поиска при обучении ставится задача минимизации целевой функции ошибки нейронной сети, которая находится по методу наименьших квадратов [25].

Типовые процедуры обучения нейронных сетей могут быть применены для настройки ANFIS-сети, так как в ней используются только дифференцируемые функции. Обычно применяется комбинация градиентного спуска в виде алгоритма обратного распространения ошибки и метода наименьших квадратов. Алгоритм обратного распространения ошибки настраивает параметры антецедентов правил, т. е. функций принадлежности. Методом наименьших квадратов оцениваются коэффициенты заключений правил, так как они линейно связаны с выходом сети. Каждая итерация процедуры настройки выполняется в два этапа. На первом этапе на входы подается обучающая выборка и по невязке между желаемым и действительным поведением сети итерационным методом наименьших квадратов находятся оптимальные параметры узлов четвертого слоя. На втором этапе остаточная невязка передается с выхода сети на входы и методом обратного распространения ошибки модифицируются параметры узлов первого слоя. При этом найденные на первом этапе коэффициенты заключений правил не изменяются. Итерационная процедура настройки продолжается, пока невязка превышает заранее установленное значение. Для настройки функций принадлежностей кроме метода обратного распространения ошибки могут использоваться и другие алгоритмы оптимизации, например метод Левенберга – Марквардта.

Еще одним простым, но эффективным методом обучения нейронной сети, который широко используется и в других моделях, является метод максимального правдоподобия. Суть метода заключается в максимизации функции правдоподобия. Как правило, в качестве данной функции используют логарифм. Однако, независимо от того, какая выбрана функция, дальнейший алгоритм сводится к методу градиентного спуска. По мнению автора, нет необходимости детально описывать сам метод, так как он является достаточно распространенным и его использование в нейронных сетях ничем не отличается от использования в других моделях анализа данных.

5.3. Виды нейронных сетей

Нейронная сеть Хопфилда (*Hopfield network, HN*) – это полносвязная нейронная сеть с симметричной матрицей связей (рис. 5.2). Во время получения входных данных каждый узел является входом, в процессе обучения он становится скрытым, а затем становится выходом. Сеть обучается так: значения нейронов устанавливаются в соот-

ветствии с желаемым шаблоном, после чего вычисляются веса, которые в дальнейшем не меняются. После того как сеть обучилась на одном или нескольких шаблонах, она всегда будет сводиться к одному из них (но не всегда к желаемому). Она стабилизируется в зависимости от общей «энергии» и «температуры» сети. У каждого нейрона есть свой порог активации, зависящий от температуры, при прохождении которого нейрон принимает одно из двух значений (обычно –1 или 1, иногда 0 или 1).

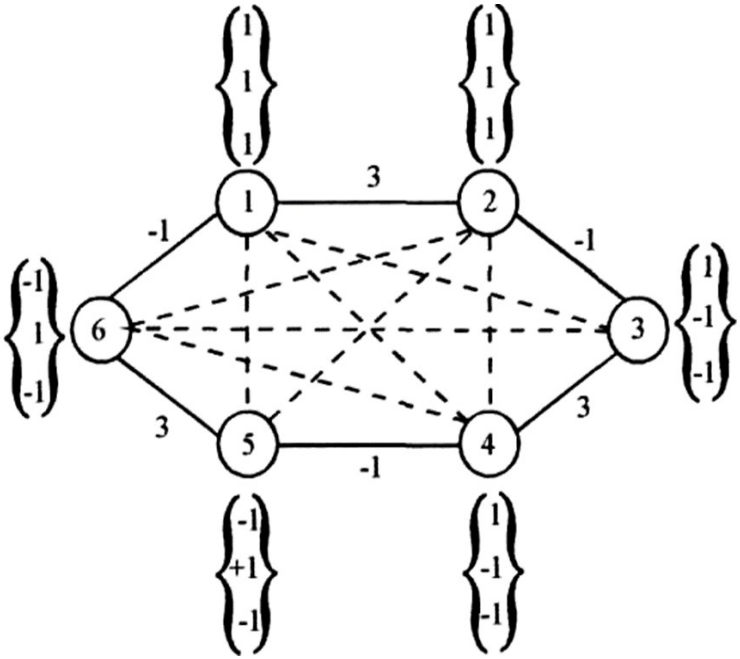

Рис. 5.2. Нейронная сеть Хопфилда

Самоорганизующиеся карты Кохонена – мощный самообучающийся механизм кластеризации, позволяющий отобразить результаты в виде компактных и удобных для интерпретации двумерных карт. Данный обработчик используется для поиска закономерностей в больших массивах данных. Это позволяет проводить разведочный анализ данных, отличающийся от классических статистических процедур, в ходе которых проверяется некоторый набор выдвинутых гипотез (рис. 5.3).

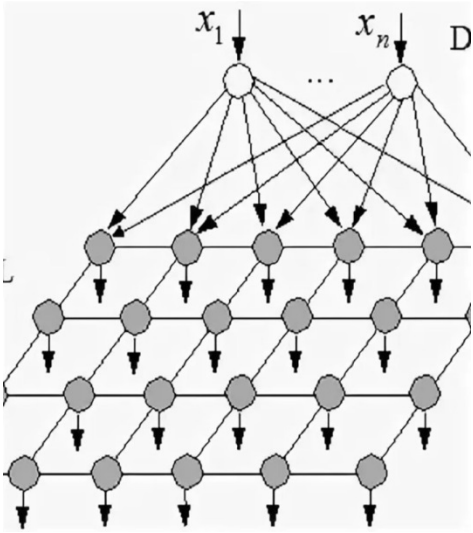

Рис. 5.3. Нейронная сеть Кохонена

Искусственная нейронная сеть Элмана, известная также как Simple Recurrent Neural Network, состоит из трех слоев – входного (распределительного) слоя (x_t), скрытого (h_t) и выходного (обрабатывающих) слоев (y_t). При этом скрытый слой имеет обратную связь сам на себя (рис. 5.4).

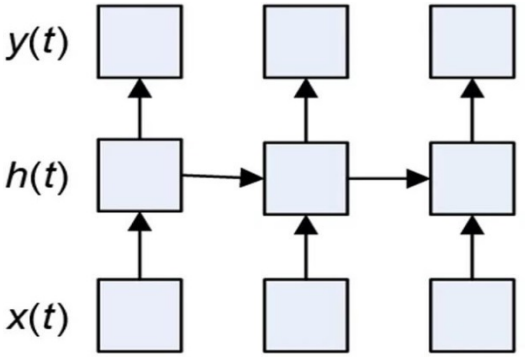

Рис. 5.4. Нейронная сеть Элмана

Искусственная нейронная сеть Хэмминга используется для решения задач классификации бинарных входных векторов (рис. 5.5). В основе ее работы лежат процедуры, направленные на выбор в качестве решения задачи классификации одного из эталонных образов, наиболее близкого к поданному на вход сети зашумленному входному образу, и отнесение данного образа к соответствующему классу. Для оценки меры близости к каждому классу используется критерий, учитывающий расстояние Хэмминга – количество различающихся переменных у зашумленного и эталонного входных образов.

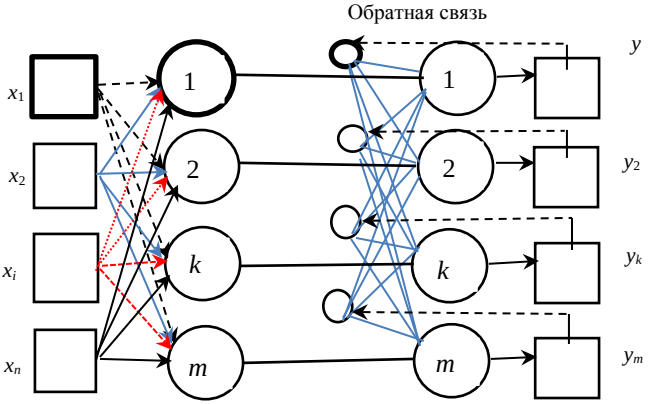

Рис. 5.5. Нейронная сеть Хэмминга

Вероятностные нейронные сети (PNN). Вид нейронных сетей для задач классификации, где плотность вероятности принадлежности классам оценивается посредством ядерной аппроксимации (рис. 5.6). Один из видов так называемых байесовых сетей (Speckt, 1990; Patterson, 1996; Bishop, 1995).

Обобщенно-регрессионная нейронная сеть (GRNN). Вид нейронной сети, где для регрессии используется ядерная аппроксимация.

Линейные сети по своей структуре аналогичны персептронам и отличаются лишь функцией активации. Для линейной нейронной сети используется правило обучения Видроу – Хоффа.

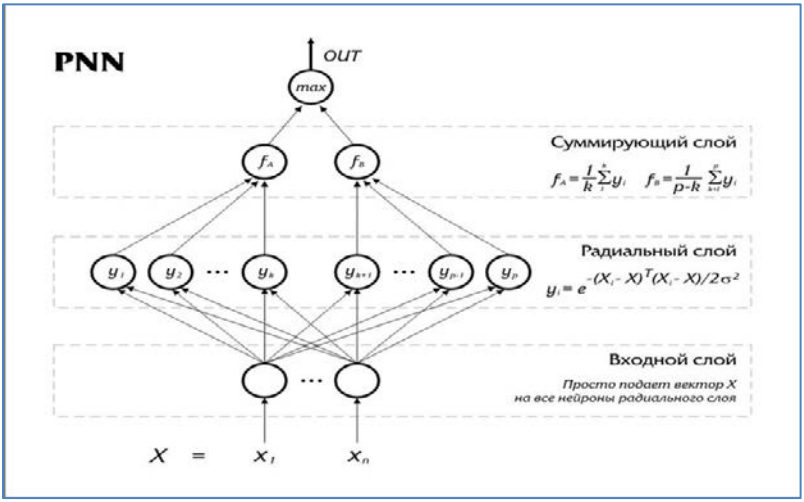

Рис. 5.6. Нейронная сеть PNN

6. МЕТОДЫ КЛАССИФИКАЦИИ: ДЕРЕВО РЕШЕНИЙ

6.1. Деревья решающих правил, классификации и регрессии.

6.2. Алгоритмы конструирования деревьев решений.

6.3. Основные характеристики алгоритмов построения деревьев решений.

6.1. Деревья решающих правил, классификации и регрессии

Решающее дерево (Decision tree) – решение задачи обучения с учителем, основанное на том, как решает задачи прогнозирования человек. В общем случае – это *k*-ичное дерево с решающими правилами в нелистовых вершинах (узлах) и некоторым заключением о целевой функции в листовых вершинах (прогнозом). Решающее правило – некоторая функция от объекта, позволяющая определить, в какую из дочерних вершин нужно поместить рассматриваемый объект. В листовых вершинах могут находиться разные объекты: класс, который нужно присвоить попавшему туда объекту (в задаче классификации), вероятности классов (в задаче классификации), непосредственно значение целевой функции (в задаче регрессии). Чаще всего на практике используются двоичные решающие деревья [34].

Обычно для построения дерева выбирается целое семейство решающих правил. Чтобы найти среди них оптимальное для каждого конкретного узла, требуется ввести некоторый критерий оптимальности. Для этого вводят некоторую меру $I(t)$ измерения того, насколько разбросаны объекты (регрессия) или перемешаны классы (классификация) в некотором узле t. Эта мера называется критерием информативности. Деревья решений (decision trees) относятся к числу самых популярных и мощных инструментов Data Mining, позволяющих эффективно решать задачи классификации и регрессии. В отличие от методов, использующих статистический подход, таких как классификатор Байеса, линейная и логистическая регрессия, деревья решений основаны на машинном обучении и в большинстве случаев не требуют предположений о статистическом распределении значений признаков. В основе деревьев решений лежат решающие правила вида «Если..., то...», которые могут быть сформулированы на естественном языке. Поэтому деревья решений являются наиболее наглядными и легко интерпретируемыми моделями. В основе работы деревьев решений лежит процесс рекурсивного разбиения исходного множества наблюдений или объектов на подмножества, ассоциированные с классами. Разбиение производится с помощью решающих правил, в которых осуществляется проверка значений атрибутов по заданному условию. *Рекурсивными* называются алгоритмы, которые работают в пошаговом режиме, при этом на каждом последующем шаге используются результаты, полученные на предыдущем шаге.

6.2. Алгоритмы конструирования деревьев решений

Классификационная модель, представленная в виде дерева решений, является интуитивной и упрощает понимание решаемой задачи. Результат работы алгоритмов конструирования деревьев решений, в отличие, например, от нейронных сетей, представляющих собой «черные ящики», легко интерпретируется пользователем. Это свойство деревьев решений не только важно при отнесении к определенному классу нового объекта, но и полезно при интерпретации модели классификации в целом.

Дерево решений позволяет понять и объяснить, почему конкретный объект относится к тому или иному классу. Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов (независимых переменных). На вход алгоритма можно подавать

все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева.

В сравнении, например, с нейронными сетями это значительно облегчает пользователю работу, поскольку в нейронных сетях выбор количества входных атрибутов существенно влияет на время обучения. Точность моделей, созданных при помощи деревьев решений, сопоставима с другими методами построения классификационных моделей (статистические методы [38], нейронные сети).

Разработан ряд масштабируемых алгоритмов, которые могут быть использованы для построения деревьев решения на сверхбольших базах данных; масштабируемость здесь означает, что с ростом числа примеров или записей базы данных время, затрачиваемое на обучение, т. е. построение деревьев решений, растет линейно.

CART (classification and regression trees) – это аббревиатура, обозначающая методы классификации и регрессии с использованием дерева решений. Это методика обучения, основанная на деревьях решений, которая возвращает классификационные или регрессионные деревья (рис. 6.1).

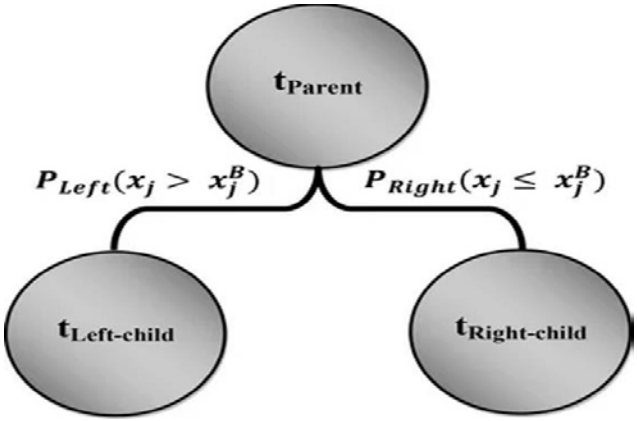

Рис. 6.1. Дерево решений по алгоритму CART

Алгоритм разработан в целях построения так называемых бинарных деревьев решений, т. е. тех деревьев, каждый узел которых при разбиении «дает» только двух потомков. Грубо говоря, алгоритм действует путем разделения на каждом шаге множества примеров ровно

напополам – по одной ветви идут те примеры, в которых правило выполняется (правый потомок), по другой – те, в которых правило не выполняется (левый потомок). Таким образом, в процессе «роста» на каждом узле дерева алгоритм проводит перебор всех атрибутов и выбирает для следующего разбиения тот, который максимизирует значение показателя, вычисляемого по математической формуле и зависящего от отношений числа примеров в правом и левом потомке к общему числу примеров.

Алгоритм C4.5 строит классификатор в форме дерева решений. Ему нужно передать набор уже классифицированных данных. А что такое классификатор? Классификатор – это инструмент, применяемый в Data Mining, который использует классифицированные данные и на их основании пытается предсказать, к какому классу стоит отнести новые данные. Вот отличия C4.5 от других систем, использующих деревья решений: во-первых, C4.5 использует приток информации при создании дерева решений; во-вторых, хотя другие системы также прореживают ветви дерева решений, C4.5 использует однопроходное прореживание, чтобы избежать переобучения. Отсечение ветвей улучшает модель; в-третьих, C4.5 может работать с дискретными и непрерывными значениями. Он делает это, ограничивая диапазоны и устанавливая пороги данных, обращая непрерывные данные в дискретные.

6.3. Основные характеристики алгоритмов построения деревьев решений

ID3. В основе этого алгоритма лежит понятие информационной энтропии, т. е. меры неопределенности информации (обратной мере информационной полезности величины). Для того чтобы определить следующий атрибут, необходимо подсчитать энтропию всех неиспользованных признаков относительно тестовых образцов и выбрать тот, для которого энтропия минимальна. Этот атрибут и будет считаться наиболее целесообразным признаком классификации.

C4.5. Этот алгоритм – усовершенствование предыдущего метода, позволяющее, в частности, «усекать» ветви дерева, если оно слишком сильно «разрастается», а также работать не только с категориальными атрибутами, но и с числовыми. В общем-то, сам алгоритм выполняется по тому же принципу, что и его предшественник; отличие состоит в возможности разбиения области значений независимой числовой переменной на несколько интервалов, каждый из которых будет являться

атрибутом. В соответствии с этим исходное множество делится на подмножества. В конечном счете, если дерево получается слишком большим, возможна обратная группировка – нескольких узлов в один лист. При этом поскольку перед построением дерева ошибка классификации уже учтена, она не увеличивается.

Среди прочих методов Data Mining метод дерева принятия решений имеет несколько достоинств:

– прост в понимании и интерпретации. Люди способны интерпретировать результаты модели дерева принятия решений после краткого объяснения;

– не требует подготовки данных. Прочие техники требуют нормализации данных, добавления фиктивных переменных, а также удаления пропущенных данных;

– способен работать как с категориальными, так и с интервальными переменными.

7. КЛАСТЕРНЫЙ АНАЛИЗ

7.1. Понятие кластера в интеллектуальном анализе данных.
7.2. Методы кластерного анализа.
7.3. Факторный анализ.

7.1. Понятие кластера в интеллектуальном анализе данных

Кластеризация – один из ключевых типов закономерностей, выявляемых методами интеллектуального анализа данных. Кластеризацию в контексте интеллектуального анализа обычно понимают как разделение целого множества на некоторое количество подмножеств по заранее неизвестным признакам, причем объекты внутри каждого из кластеров должны быть близки между собой по одному или нескольким признакам, доступным для интерпретации. Методы кластеризации могут оказаться полезными в самых разных отраслях экономики. В первую очередь речь идет об областях массового обслуживания. Банки, операторы мобильной связи, страховые организации – лишь некоторые экономические объекты, для которых объективное разделение множества потенциальных клиентов на разумно определяемые группы может привести к существенному положительному результату. Объектами сегментации могут выступать и другие экономические объекты, например товары, контрагенты, ценные бумаги, транзакции.

Простейшим методом кластеризации является визуализация. Однако она применима лишь тогда, когда число значимых для кластеризации факторов ограничено. Не составляет особого труда выделить кластеры на двухмерной диаграмме; иногда удается разглядеть кластеры на объемной трехмерной диаграмме. Но увеличение размерности пространства изучаемых образцов делает визуальные методы невозможными, что приводит к необходимости использования иных инструментов. Целый ряд таких инструментов кластеризации был разработан в рамках концепции интеллектуального анализа данных. Так, в составе аналитической платформы Deductor Studio компании Basegroup Labs представлена кластеризация методами k-means, методами g-means, а также нейросетевыми методами на основе самоорганизующихся карт Кохонена. Вместе с тем на практике инструменты многомерной кластеризации находят довольно ограниченное применение. Одна из причин состоит в том, что разные методы кластеризации часто приводят к разным результатам на одних и тех же массивах данных. Это в некоторой мере подрывает авторитет технологий в глазах практиков, не всегда способных оценить результаты кластеризации с точки зрения качества разделения.

Отсутствие единого понимания сущности кластера во многом вызвано неоднозначным и нередко расширенным его применением в различных науках (например, кластеры как элементы кластерной политики), а также в практике информационных технологий (кластеры как система компьютеров или как единица хранения данных в некоторых файловых системах). Разницу в подходах необходимо учитывать при использовании инструментов интеллектуального анализа данных, так как результат может отличаться от представлений пользователя [46, 49].

7.2. Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы: иерархические и неиерархические (рис. 7.1).

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

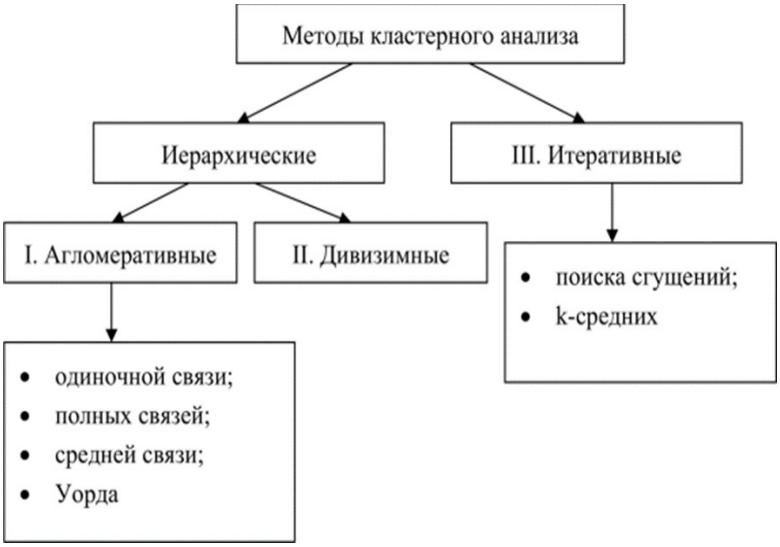

Рис. 7.1. Методы кластеризации

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос – как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

Метод ближнего соседа или одиночная связь. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов.

В результате работы этого метода кластеры представляются длинными «цепочками» или «волокнистыми» кластерами, «сцепленными вместе» только отдельными элементами, которые случайно оказались ближе остальных друг к другу (рис. 7.2).

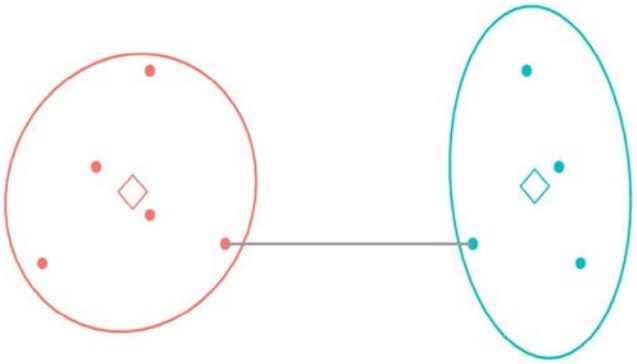

Рис. 7.2. Метод ближнего соседа

Метод наиболее удаленных соседей или полная связь. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. «наиболее удаленными соседями»). Данный метод хорошо использовать тогда, когда объекты действительно происходят из различных «рощ». Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является «цепочечным», то этот метод не следует использовать (рис. 7.3).

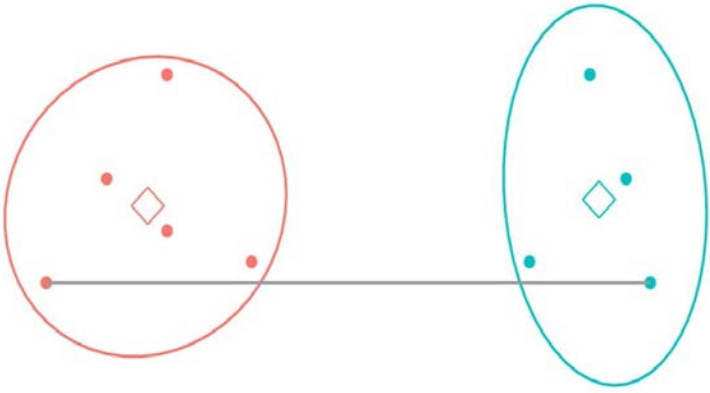

Рис. 7.3. Метод дальнего соседа

Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения (Ward, 1963). В отличие от других методов кластерного анализа для оценки расстояний между кластерами здесь используются методы дисперсионного анализа. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т. е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и создание кластеров малого размера (рис. 7.4).

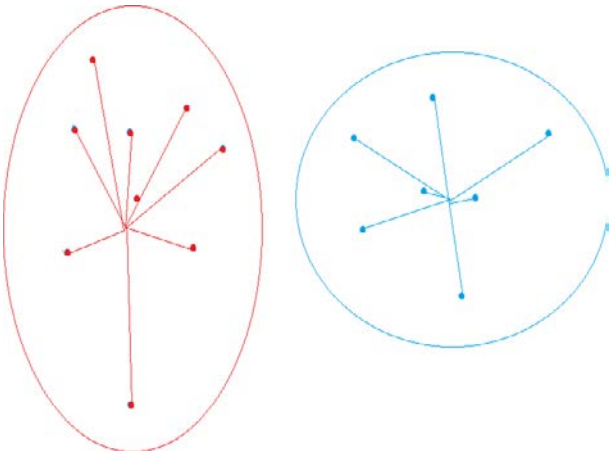

Рис. 7.4. Метод Варда

Метод невзвешенного попарного среднего (метод невзвешенного попарного арифметического среднего – unweighted pair-group method using arithmetic averages, UPGMA).

В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если объекты действительно происходят из различных «рощ», в случаях присутствия кластеров «цепочного» типа, при предположении неравных размеров кластеров.

Метод взвешенного попарного среднего (метод взвешенного попарного арифметического среднего – weighted pair-group method using arithmetic averages, WPGMA). Этот метод похож на метод невзвешен-

ного попарного среднего, разница состоит лишь в том, что здесь в качестве весового коэффициента используется размер кластера (число объектов, содержащихся в кластере).

Этот метод рекомендуется использовать именно при наличии предположения о кластерах разных размеров.

Невзвешенный центроидный метод (метод невзвешенного попарного центроидного усреднения – unweighted pair-group method using the centroid average). В качестве расстояния между двумя кластерами в этом методе берется расстояние между их центрами тяжести.

Взвешенный центроидный метод (метод взвешенного попарного центроидного усреднения – weighted pair-group method using the centroid average, WPGMC) (Sneath, Sokal, 1973). Этот метод похож на предыдущий, отличие его состоит в том, что для учета разницы между размерами кластеров (числом объектов в них) используются веса. Этот метод предпочтительно использовать в случаях, если имеются предположения относительно существенных отличий в размерах кластеров.

Кроме иерархических методов классификации существует группа итеративных методов кластерного анализа. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.). Как и в иерархическом кластерном анализе, в итеративных методах существует проблема определения числа кластеров. В общем случае их число может быть неизвестно. Не все итеративные методы требуют первоначального задания числа кластеров, но позволяют, используя несколько алгоритмов, меняя либо число образуемых кластеров, либо установленный порог близости для объединения объектов в кластеры, добиваться наилучшего разбиения по задаваемому критерию качества.

К группе итеративных методов принадлежит *метод k-средних*. Существуют две модификации метода *k*-средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, вторая – лишь после того, как будет завершен просмотр всех данных. Одним из итеративных методов классификации, не требующих задания числа кластеров, является *метод поиска сгущений*. Данный метод требует вычисления матрицы расстояний, затем выбирается объект, который является первоначальным центром первого кластера. Выбор такого объекта может быть произвольным, а может основываться на предварительном анализе точек и их окрестностей.

Выбранная точка принимается за центр гиперсферы заданного ра-

диуса R. Определяется совокупность точек, попавших внутрь этой сферы, и для них вычисляются координаты центра (вектор средних значений признаков). Далее рассматривается гиперсфера такого же радиуса, но с новым центром, и для совокупности попавших в нее точек опять рассчитывается вектор средних значений, который принимается за новый центр сферы, и т. д. Когда очередной пересчет координат центра сферы приводит к такому же результату, как на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер и из дальнейшего процесса кластеризации исключаются. Для всех оставшихся точек процедуры повторяются.

7.3. Факторный анализ

При обработке данных может возникать задача совмещения факторного и кластерного анализа. Эта задача возникает обычно в тех случаях, когда сначала исходная информация преобразуется путем выделения факторов, а затем на основе преобразованных данных производится поиск в изучаемой совокупности отделимых друг от друга классов. При проведении сегментирования факторный анализ используется прежде всего как метод сжатия данных, т. е. сокращения большого количества переменных. Переменные, которые могут быть использованы для сегментирования с применением кластерного анализа, сокращаются до некоторого основного набора составных переменных (факторов), которые затем и используются при кластеризации. Необходимость такого сокращения связана не только с желанием исследователя «ускорить» процедуру кластерного анализа, но и с некоторыми важными соображениями:

– если в кластерный анализ включаются несколько переменных, связанных с описанием одинаковых или близких характеристик (например, параметров товара), то эти характеристики получают гораздо больший вес. Поскольку расстояния вычисляются исходя из разностей между наблюдениями по каждой переменной, то несколько связанных переменных окажут большее влияние на результаты. Достаточно очевидной эта ситуация становится при рассмотрении гипотетического примера, когда в кластерном анализе участвуют две совершенно идентичные переменные. В этом случае двукратно усиливается воздействие измеряемой этими переменными характеристики на конечный результат;

– важной причиной использования факторного анализа перед про-

ведением кластеризации является четкость и простота интерпретации. Исследователю гораздо проще понять кластерное решение, основывающееся на анализе 5–6 факторов (если у них имеется осмысленная интерпретация), чем решение для 50–60 переменных.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) предложен Тьяном Зангом и его коллегами. Благодаря обобщенным представлениям кластеров, скорость кластеризации увеличивается, алгоритм при этом обладает большим масштабированием. В этом алгоритме реализован двухэтапный процесс кластеризации. В ходе первого этапа формируется предварительный набор кластеров. На втором этапе к выявленным кластерам применяются другие алгоритмы кластеризации, пригодные для работы в оперативной памяти. Можно привести следующую аналогию, описывающую этот алгоритм. Если каждый элемент данных представить себе, как бусину, лежащую на поверхности стола, то кластеры бусин можно «заменить» теннисными шариками и перейти к более детальному изучению кластеров теннисных шариков. Число бусин может оказаться достаточно велико, однако диаметр теннисных шариков можно подобрать таким образом, чтобы на втором этапе можно было, применив традиционные алгоритмы кластеризации, определить действительную сложную форму кластеров (рис. 7.5).

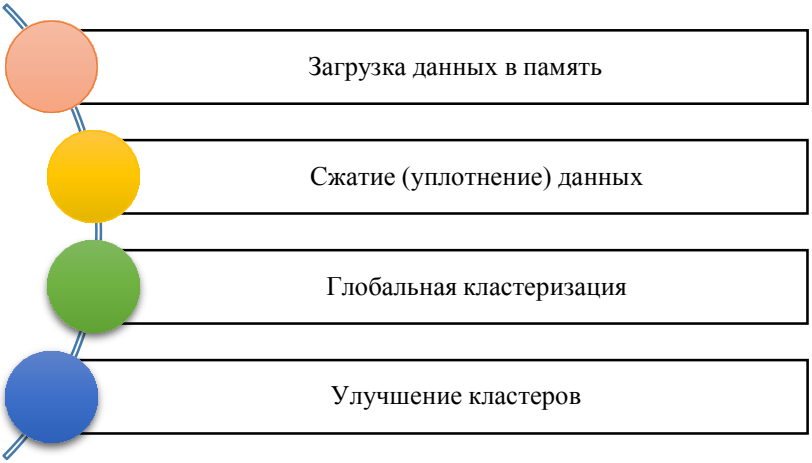

Рис. 7.5. Алгоритм BIRCH

Из рис. 7.5 следует, что этот алгоритм включает следующие фазы.

Фаза 1. Загрузка данных в память. Построение начального кластерного дерева (CF Tree) по данным (первое сканирование набора данных) в памяти. Подфазы основной фазы происходят быстро, точно, практически нечувствительны к порядку. Алгоритм построения кластерного дерева (CF Tree):

Кластерный элемент состоит из тройки чисел (*N, LS, SS*), где *N* – количество элементов входных данных, входящих в кластер, *LS* – сумма элементов входных данных, *SS* – сумма квадратов элементов входных данных.

Кластерное дерево – это взвешенно сбалансированное дерево с двумя параметрами: B – коэффициент разветвления, *T* – пороговая величина. Каждый нелистьевой узел дерева имеет не более чем B вхождений узлов следующей формы: [CF_i, *Child*], где *i* = 1, 2, …, B; *Child* – указатель на *i*-й дочерний узел.

Каждый листьевой узел имеет ссылку на два соседних узла. Кластер, состоящий из элементов листьевого узла, должен удовлетворять следующему условию: диаметр или радиус полученного кластера должен быть не более пороговой величины *T*.

Фаза 2 (необязательная). Сжатие (уплотнение) данных. Сжатие данных до приемлемых размеров с помощью перестроения и уменьшения кластерного дерева с увеличением пороговой величины *T*.

Фаза 3. Глобальная кластеризация. Применяется выбранный алгоритм кластеризации на листьевых компонентах кластерного дерева.

Фаза 4 (необязательная). Улучшение кластеров. Использует центры тяжести кластеров, полученные в фазе 3, как основы. Перераспределяет данные между «близкими» кластерами. Данная фаза гарантирует попадание одинаковых данных в один кластер.

WaveCluster представляет собой алгоритм кластеризации на основе волновых преобразований. В начале работы алгоритма данные обобщаются путем наложения на пространство данных многомерной решетки. На дальнейших шагах алгоритма анализируются не отдельные точки, а обобщенные характеристики точек, попавших в одну ячейку решетки. В результате такого обобщения необходимая информация умещается в оперативной памяти. На последующих шагах для определения кластеров алгоритм применяет волновое преобразование к обобщенным данным. Главные особенности WaveCluster приведены ниже (рис. 7.6).

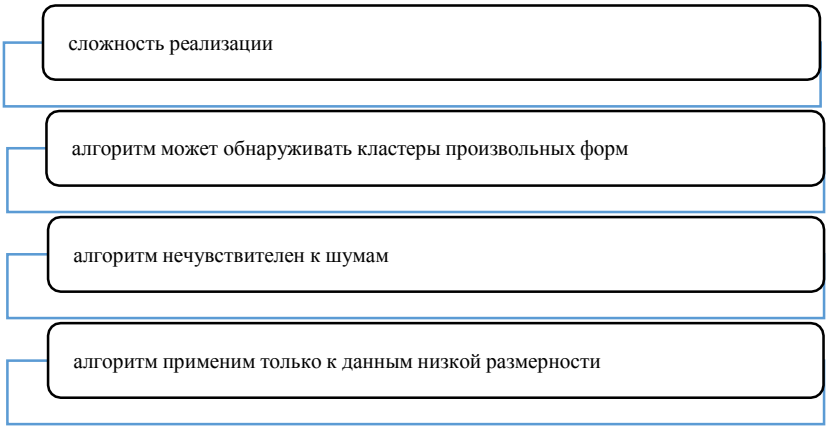

Рис. 7.6. Алгоритм WaveCluster

Алгоритм CLARA был разработан Kaufmann и Rousseeuw в 1990 г. для кластеризации данных в больших базах данных. Данный алгоритм строится в статистических аналитических пакетах, например таких как S+. Изложим кратко суть алгоритма. Алгоритм CLARA извлекает множество образцов из базы данных (рис. 7.7).

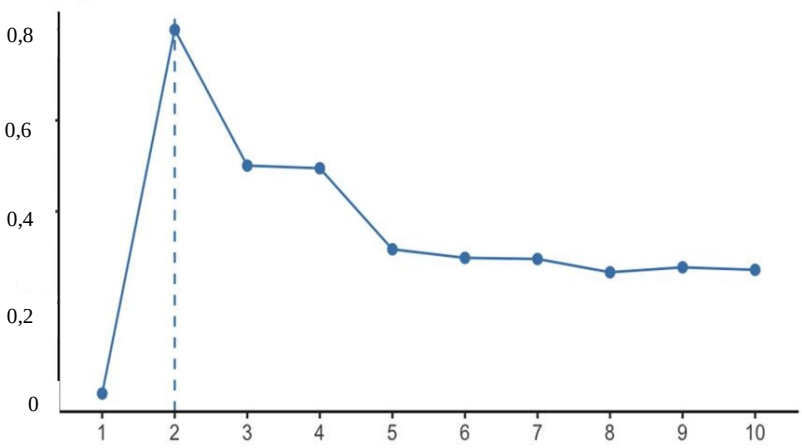

Рис. 7.7. Алгоритм CLARA

Кластеризация применяется к каждому из образцов, на выходе алгоритма предлагается лучшая кластеризация. Для больших баз данных этот алгоритм эффективнее, чем алгоритм PAM. Эффективность алгоритма зависит от выбранного в качестве образца набора данных. Хорошая кластеризация на выбранном наборе может не дать хорошую кластеризацию на всем множестве данных.

Алгоритм Clarins (Clustering Large Applications based upon RANdomized Search) формулирует задачу кластеризации как случайный поиск в графе. В результате работы этого алгоритма совокупность узлов графа представляет собой разбиение множества данных на число кластеров, определенное пользователем.

«Качество» полученных кластеров определяется при помощи критериальной функции. Алгоритм Clarins сортирует все возможные разбиения множества данных в поисках приемлемого решения. Поиск решения останавливается в том узле, где достигается минимум среди предопределенного числа локальных минимумов.

Среди новых масштабируемых алгоритмов также можно отметить *алгоритм CURE* – алгоритм иерархической кластеризации и *алгоритм DBScan*, где понятие кластера формулируется с использованием концепции плотности (density).

Основным недостатком алгоритмов BIRCH, Clarins, CURE, DBScan является то обстоятельство, что они требуют задания некоторых порогов плотности точек, а это не всегда приемлемо. Эти ограничения обусловлены тем, что описанные алгоритмы ориентированы на сверхбольшие базы данных и не могут пользоваться большими вычислительными ресурсами. Существует ряд сложностей, которые следует продумать перед проведением кластеризации:

1) сложность выбора характеристик, на основе которых проводится кластеризация. Необдуманный выбор приводит к неадекватному разбиению на кластеры и, как следствие, – к неверному решению задачи;

2) сложность выбора метода кластеризации. Этот выбор требует неплохого знания методов и предпосылок их использования. Чтобы проверить эффективность конкретного метода в определенной предметной области, целесообразно применить следующую процедуру: рассматривают несколько априори различных между собой групп и перемешивают их представителей между собой случайным образом.

Далее проводится кластеризация для восстановления исходного разбиения на кластеры. Доля совпадений объектов в выявленных и исходных группах является показателем эффективности работы метода;

3) проблема выбора числа кластеров. Если нет никаких сведений относительно возможного числа кластеров, необходимо провести ряд экспериментов и, в результате перебора различного числа кластеров, выбрать оптимальное их число.

8. АССОЦИАТИВНЫЕ ПРАВИЛА

8.1. Ассоциативные правила в интеллектуальном анализе данных.
8.2. Методы поиска ассоциативных правил.
8.3. Нечеткие ассоциативные правила.

8.1. Ассоциативные правила в интеллектуальном анализе данных

В настоящее время наблюдается переизбыток так называемых неструктурированных данных, в которых каждая единица хранения не может быть представлена конечным числом признаков (атрибутов). Такие данные могут содержать, например, информацию о товарах, купленных одним покупателем у предприятия розничной торговли.

В связи с этим возникают задачи:

– сокращения объемов неструктурированных данных путем удаления избыточных транзакций, исключение которых из дальнейшего рассмотрения не повлияет на качество синтезируемых правил и моделей;

– выявления интересных правил, позволяющих извлекать новые знания на основе имеющихся неструктурированных данных;

– построения моделей на основе больших массивов неструктурированных данных для решения практических задач прогнозирования, классификации и кластеризации данных.

Для обработки больших массивов неструктурированных данных и решения указанных задач целесообразно использовать методы поиска ассоциативных правил, позволяющие выявлять новые закономерности вида «если условие, то действие» в имеющихся данных и синтезировать на их основе интерпретабельные базы правил, понятные экспертам в прикладных областях [21, 23, 53].

В настоящее время предложено достаточно большое количество видов ассоциативных правил, каждый из которых целесообразно применять для решения определенного класса задач. Поэтому актуальными являются обзор и классификация ассоциативных правил для даль-

нейшего их применения с целью решения практических задач интеллектуального анализа данных.

Поиск ассоциативных правил – это метод извлечения данных для изучения корреляций и взаимосвязи между переменными в базе данных. В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Суть задачи заключается в определении часто встречающихся наборов объектов в большом множестве таких наборов.

Данная задача является частным случаем задачи классификации. Первоначально она решалась при анализе тенденций в поведении покупателей в супермаркетах. Анализу подвергались данные о совершаемых ими покупках, которые покупатели складывают в корзину. Это послужило причиной второго часто встречающегося названия – анализ рыночных корзин.

При анализе часто вызывает интерес последовательность производящих событий. При обнаружении закономерностей в таких последовательностях можно с некоторой долей вероятности предсказывать появление событий в будущем, что позволяет принимать более правильные решения.

Такая задача является разновидностью задачи поиска ассоциативных правил и называется анализом последовательностей. Основным отличием задачи анализа последовательностей от задачи поиска ассоциативных правил является установление отношения порядка между исследуемыми наборами. Данное отношение может быть определено разными способами. При анализе последовательности событий, происходящих во времени, объектами таких наборов являются события, а отношение порядка соответствует хронологии их появления.

8.2. Методы поиска ассоциативных правил

Ассоциативные правила представляют собой механизм нахождения логических закономерностей между связанными элементами (событиями или объектами).

Выделяют три вида правил:

полезные правила, содержащие действительную информацию, которая ранее была неизвестна, но имеет логическое объяснение;

тривиальные правила, содержащие действительную и легко объяснимую информацию, отражающую известные законы в исследуемой области, и поэтому не приносящие какой-либо пользы;

непонятные правила, содержащие информацию, которая не может быть объяснена (такие правила или получают на основе аномальных исходных данных, или они содержат глубоко скрытые закономерности, и поэтому для интерпретации непонятных правил нужен дополнительный анализ). Поиск ассоциативных правил обычно выполняют в два этапа:

в пуле имеющихся признаков A находят наиболее часто встречающиеся комбинации элементов T;

из этих найденных наиболее часто встречающихся наборов формируют ассоциативные правила.

Алгоритм AIS. Первый алгоритм поиска ассоциативных правил, называвшийся AIS (предложенный Agrawal, Imielinski and Swami), был разработан сотрудниками исследовательского центра IBM Almaden в 1993 г. С этой работы начался интерес к ассоциативным правилам; на середину 90-х гг. прошлого века пришелся пик исследовательских работ в этой области, и с тех пор каждый год появляется несколько новых алгоритмов. В алгоритме AIS кандидаты множества наборов генерируются и подсчитываются «на лету» во время сканирования базы данных.

Алгоритм SETM. Создание этого алгоритма было мотивировано желанием использовать язык SQL для вычисления часто встречающихся наборов товаров. Как и алгоритм AIS, SETM также формирует кандидатов «на лету», основываясь на преобразованиях базы данных. Чтобы использовать стандартную операцию объединения языка SQL для формирования кандидатов, SETM отделяет формирование кандидатов от их подсчета. Неудобство алгоритмов AIS и SETM – излишнее генерирование и подсчет слишком многих кандидатов, которые в результате не оказываются часто встречающимися. Для улучшения их работы был предложен алгоритм Apriori.

Алгоритм Apriori. На первом шаге алгоритма подсчитываются 1-элементные часто встречающиеся наборы. Для этого необходимо пройтись по всему набору данных и подсчитать для них поддержку, т. е., сколько раз встречается в базе. Следующие шаги будут состоять из двух частей: генерации потенциально часто встречающихся наборов элементов (кандидатов) и подсчета поддержки для кандидатов (рис. 8.1).

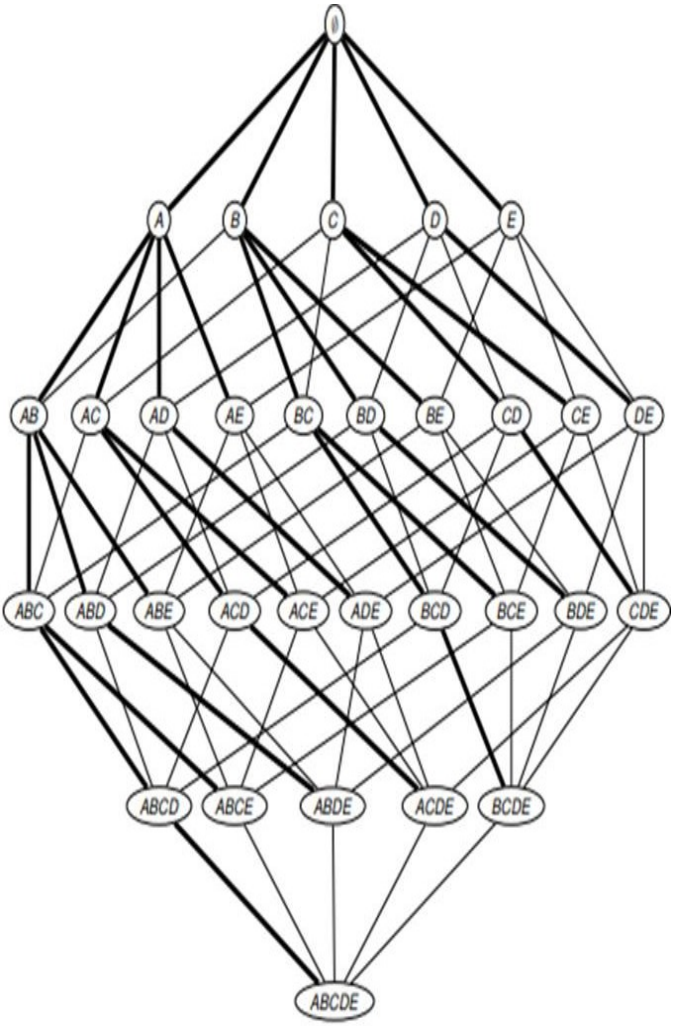

Рис. 8.1. Алгоритм Apriori

Некоторыми авторами были предложены другие алгоритмы поиска ассоциативных правил, целью которых также было усовершенствование алгоритма Apriori. Один из них – *алгоритм DHP*, также называемый *алгоритмом хеширования* (J. Park, M. Chen and P. Yu, 1995).

В основе его работы лежит вероятностный подсчет наборов кандидатов, осуществляемый для сокращения числа подсчитываемых кандидатов на каждом этапе выполнения алгоритма Apriori. Сокращение обеспечивается за счет того, что каждый из *k*-элементных наборов кандидатов помимо шага сокращения проходит шаг хеширования. В алгоритме на *k*-1 этапе во время выбора кандидата создается так называемая хеш-таблица. Каждая запись хеш-таблицы является счетчиком всех поддержек *k*-элементных наборов, которые соответствуют этой записи в хеш-таблице. Алгоритм использует эту информацию на этапе *k* для сокращения множества *k*-элементных наборов кандидатов. После сокращения подмножества, как это происходит в Apriori, алгоритм может удалить набор кандидатов, если его значение в хеш-таблице меньше порогового значения, установленного для обеспечения.

К другим усовершенствованным алгоритмам относятся PARTITION, DIC, алгоритм выборочного анализа.

Алгоритм PARTITION (A. Savasere, E. Omiecinski and S. Navathe, 1995). Этот алгоритм разбиения (разделения) заключается в сканировании транзакционной базы данных путем разделения ее на непересекающиеся разделы, каждый из которых может уместиться в оперативной памяти. На первом шаге в каждом из разделов при помощи алгоритма Apriori определяются «локальные» часто встречающиеся наборы данных. На втором подсчитывается поддержка каждого такого набора относительно всей базы данных. Таким образом, на втором этапе определяется множество всех потенциально встречающихся наборов данных.

Алгоритм DIC, Dynamic Itemset Counting (S. Brin, R. Motwani, J. Ullman and S. Tsur, 1997). Алгоритм разбивает базу данных на несколько блоков, каждый из которых отмечается так называемыми начальными точками (start point), и затем циклически сканирует базу данных.

8.3. Нечеткие ассоциативные правила

Нечеткие ассоциативные правила (fuzzy associative rules) – инструмент для извлечения из баз данных закономерностей, которые формулируются в виде лингвистических высказываний. Здесь введены специальные понятия нечеткой транзакции, поддержки и достоверности нечеткого ассоциативного правила (рис. 8.2). Гибридизация методов интеллектуальной обработки информации – девиз, под которым про-

шли 90-е гг. у западных и американских исследователей. В результате объединения нескольких технологий искусственного интеллекта появился специальный термин – «мягкие вычисления» (soft computing), который ввел Л. Заде в 1994 г.

Один из самых простых способов сделать алгоритм интеллектуального анализа ассоциативных правил более «нечетким» – выполнить «нарезку» (сделать так называемые альфа-сечения – alpha-cuts) в заданных точках в диапазоне между 0,0 и 1,0, где элементы со значением функции принадлежности выше уровня отсечения считаются «четкими» (что соответствует полной принадлежности), а элементы ниже уровня отсечения считаются элементами без принадлежности. Результатом являются уровни достоверности для правил при различных значениях уровня отсечения. Это один из «самых нечетких» методов с результатом в виде двумерного графа, а не единственного значения достоверности.

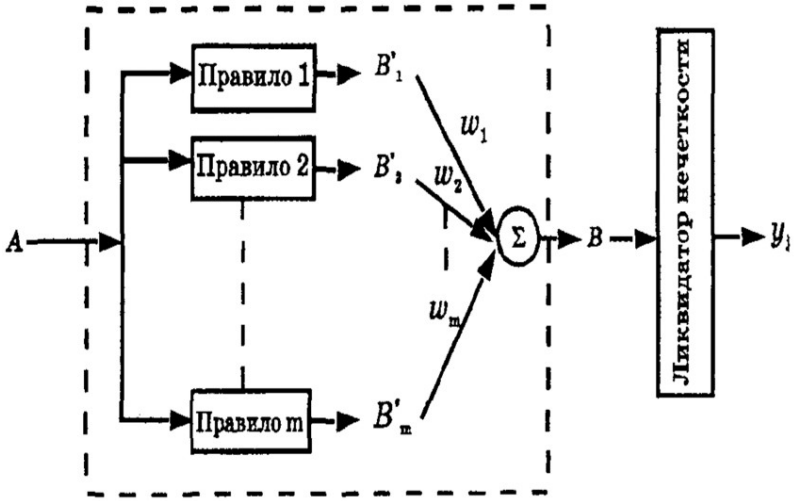

Рис. 8.2. Архитектура системы нечетких правил

В 1998 г. группа исследователей из Китайского университета Гонконга (Chinese University of Hong Kong), в том числе Ада Вай-чи (Ada Wai-chee) и Мень Хонь Вон (Man Hon Wong), разработали методику для поиска нечетких множеств в заданной базе данных. Научный до-

клад под названием «Finding Fuzzy Sets for the Mining of Fuzzy Association Rules for Numerical Attributes» (Поиск нечетких множеств с целью интеллектуального анализа нечетких ассоциативных правил для численных атрибутов), представлявший их работу, был написан специально для международной конференции «Intelligent Data Engineering and Automated Learning Conference» (Интеллектуальная инженерия данных и автоматизированное обучение) в 1998 г. Эта работа представляет особый интерес, поскольку устраняет требование о необходимости заранее знать значения или функции нечеткой принадлежности. Недостаток методики состоит в том, что она использует довольно сложный алгоритм кластеризации, известный как CLARINS.

9. ГЕНЕТИЧЕСКИЕ МОДЕЛИ

9.1. Генетика экономических процессов.
9.2. Основные понятия генетических алгоритмов.
9.3. Эволюционное моделирование.

9.1. Генетика экономических процессов

Рассмотрим статику, динамику и генетику экономических процессов. Течение хозяйственной жизни страны пронизано большим количеством разнообразных экономических процессов. Государственное регулирование экономики должно опираться на понимание и умелое использование закономерностей ее функционирования [29].

Экономика может рассматриваться в трех разрезах – с точки зрения статики, динамики и генетики. Статика раскрывает структуру экономики, взаимодействие между отдельными ее составными элементами в условиях сравнительно плавного эволюционного развития. Динамика показывает качественные перемены в экономике при переходе от фазы к фазе экономического цикла или при смене циклов. Генетика исследует причины и последствия качественных перемен в экономике, механизмы наследственности, изменчивости и отбора в этой сфере общества. Можно выделить несколько общих закономерностей цикличногенетической динамики экономики как целостной и сложной по структуре системы. Рыночная экономика развивается неравномерно, циклично, последовательно, проходя через фазы стабильного развития, кризиса, депрессии, оживления, подъема, а также среднесрочных, долгосрочных (Кондратьевских), сверхдолгосрочных (цивилизационных) циклов (рис. 9.1).

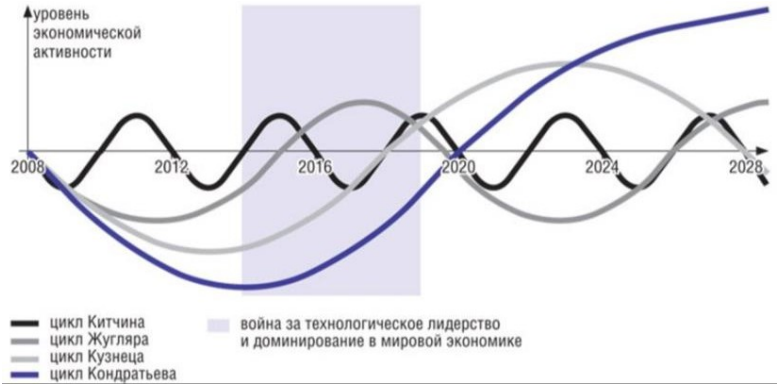

Рис. 9.1. Типы экономических циклов

Представления о возможности планомерно-равномерной, бескризисной экономической динамики оказались иллюзией. Экономика периодически переходит от одного относительно равновесного, устойчивого состояния к другому равновесному состоянию через переходный кризисный период, характеризующийся усилением неустойчивости, спадом производства, хаотичными переменами, перестройкой структуры экономики. При этом наследственное генетическое ядро экономической системы (или надсистемы – при смене систем) очищается от устаревших элементов и обогащается дополнительным содержанием в результате наследственной изменчивости, позволяющей адаптироваться к новым условиям развития общества. Эти полезные изменения происходят на основе отбора – стихийного или сознательного (целенаправленного) – из множества возможных вариаций.

Экономическая эволюция – процесс системного порядка, который охватывает все уровни (от нано- до мега-) проявлений изменчивости, наследственности и необратимости. Это процесс непрерывного изменения степени сложности субъектов и объектов, их функций и связей, их организации и институции, средств и методов привлечения из среды для преобразования в новые общественно полезные и необходимые формы вещества и энергии [51].

9.2. Основные понятия генетических алгоритмов

Генетический алгоритм представляет собой метод, отражающий естественную эволюцию решения проблем задач оптимизации.

При его описании используются определения, заимствованные из генетики. Например, говоря о популяции особей, в качестве базовых понятий применяются ген, хромосома, генотип, фенотип, аллель. Также используются соответствующие этим терминам определения из технического лексикона, в частности цепь, двоичная последовательность, структура. Популяция – это конечное множество особей. Особи, входящие в популяцию, в генетических алгоритмах представляются хромосомами с закодированными в них множествами параметров задачи, т. е. решений, которые иначе называются точками в пространстве поиска.

Хромосомы (цепочки или кодовые последовательности) – это упорядоченные последовательности генов. Ген (также называемый свойством, знаком или детектором) – атомарный элемент генотипа, в частности хромосомы. Каждая хромосома может кодировать только один параметр задачи. Но так как в одном генотипе может быть несколько хромосом, то один генотип может закодировать несколько параметров. Например, генотип включает в себя две хромосомы, каждая хромосома состоит из трех генов, значит, первые три гена генотипа кодируют один параметр, а вторые три гена – второй параметр (рис. 9.2).

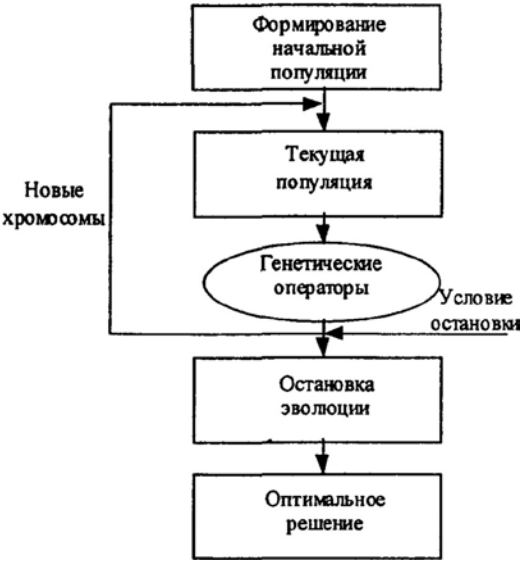

Рис. 9.2. Генетический алгоритм

Очень важным понятием в генетических алгоритмах считается функция приспособленности, иначе называемая *функцией оценки*. Она представляет собой меру приспособленности данной особи в популяции. Эта функция играет важнейшую роль, поскольку позволяет оценить степень приспособленности конкретных особей в популяции и выбрать из них наиболее приспособленные (имеющие наибольшие значения функции приспособленности) в соответствии с эволюционным принципом «выживания сильнейших» (лучше всего приспособившихся). Функция приспособленности также получила свое название непосредственно из генетики. Она оказывает сильное влияние на функционирование генетических алгоритмов и должна иметь точное и корректное определение. В задачах оптимизации функция приспособленности, как правило, оптимизируется (точнее говоря, максимизируется) и называется целевой санкцией. В задачах минимизации целевая функция преобразуется и проблема сводится к максимизации. В теории управления функция приспособленности может принимать вид функции погрешности, а в теории игр – стоимостной функции. На каждой итерации генетического алгоритма приспособленность каждой особи данной популяции оценивается при помощи функции приспособленности, и на этой основе создается следующая популяция особей, составляющих множество потенциальных решений проблемы, например, задачи оптимизации.

9.3. Эволюционное моделирование

Потребность в прогнозе и адекватной оценке последствий осуществляемых человеком мероприятий (особенно негативных) приводит к необходимости моделирования динамики изменения основных параметров системы, динамики взаимодействия открытой системы с ее окружением (ресурсы, потенциал, условия, технологии и т. д.), с которым осуществляется обмен ресурсами в условиях враждебных, конкурентных, кооперативных или же безразличных взаимоотношений. Здесь необходимы системный подход, эффективные методы и критерии оценки адекватности моделей, которые направлены не только (не столько) на максимизацию критериев типа «прибыль», «рентабельность», но и на оптимизацию отношений с окружающей средой. Если критерии первого типа важны, например, для кратко- и среднесрочного прогнозирования и тактического администрирования, то второго типа – для средне- и долгосрочного прогноза, стратегического админи-

стрирования. При этом необходимо выделить и изучить достаточно полную и информативную систему параметров исследуемой системы и ее окружения, разработать методику введения мер информативности и близости состояний системы. Важно отметить, что при этом некоторые критерии и меры могут часто конфликтовать друг с другом. Многие такие социально-экономические системы можно описывать с единых позиций, средствами и методами единой теории – эволюционной.

При эволюционном моделировании процесс моделирования сложной социально-экономической системы сводится к созданию модели ее эволюции или к поиску допустимых состояний системы, к процедуре (алгоритму) отслеживания множества допустимых состояний (траекторий) [31, 41, 55]. При этом актуализируются такие атрибуты биологической эволюционной динамики (ниже в скобках приведены возможные социально-экономические интерпретации этих атрибутов для эволюционного моделирования), как, например, на рис. 9.3.

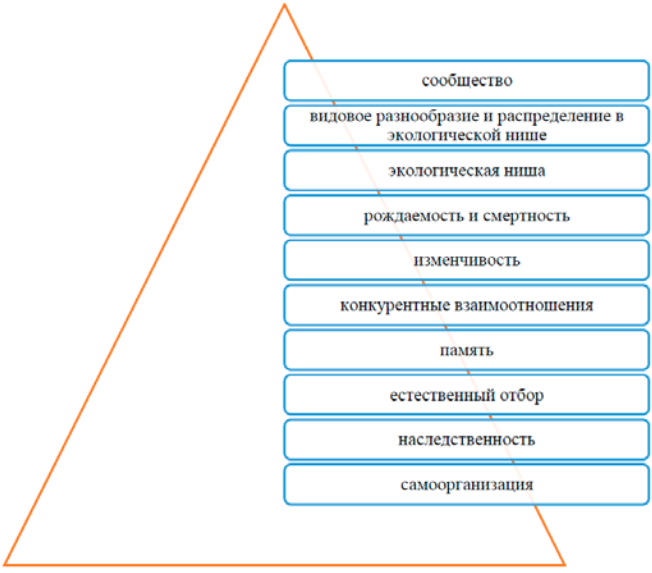

Рис. 9.3. Атрибуты эволюционного моделирования

Из рис. 9.3 представлены такие атрибуты эволюционного моделирования, как:

сообщество (корпорация, корпоративные объекты, субъекты, окружение);

видовое разнообразие и распределение в экологической нише (типы распределения ресурсов, структура связей в данной корпорации);

экологическая ниша (сфера влияния и функционирования, эволюции на рынке, в бизнесе);

рождаемость и смертность (производство и разрушение); изменчивость (экономической обстановки, ресурсов); конкурентные взаимоотношения (рыночные отношения); память (способность к циклам воспроизводства); естественный отбор (штрафные и поощрительные меры);

наследственность (производственные циклы и их предыстория); регуляция (инвестиции);

самоорганизация и стремление системы в процессе эволюции максимизировать контакт с окружением в целях самоорганизации, возврата на траекторию устойчивого развития и др.

При исследовании эволюции системы необходима ее декомпозиция на подсистемы с целью обеспечения следующих функций (рис. 9.4).

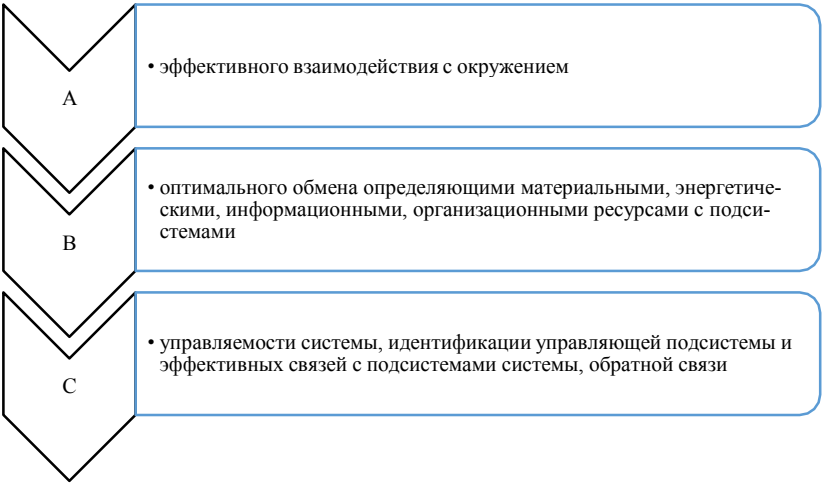

Рис. 9.4. Функции декомпозиции системы эволюционного моделирования

Одной из важнейших экономических задач, реализующей концепцию эволюционного моделирования, является транспортная задача.

Имеется три ее модификации: задача Монжа – Канторовича (рис. 9.5); задача Ордена; задача Хичкока.

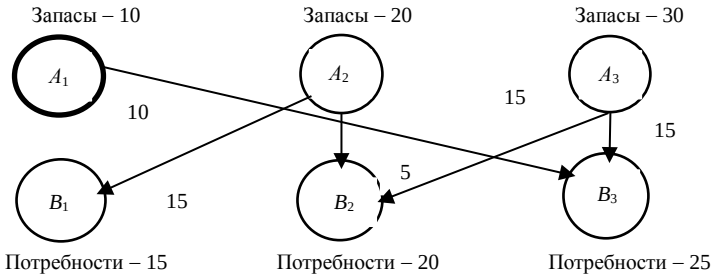

Рис. 9.5. Пример решения задачи Монжа – Канторовича в виде графа

Задача Монжа – на заданной площади размещено множество объемов товаров (A_1, A_2, A_3), которые необходимо перевезти на другую заданную площадь (B_1, B_2, B_3) с минимизацией транспортных затрат.

Задача Ордена – распределение товаров задано, и в некоторых местах спрос превышает предложение, а в некоторых – наоборот. При этом задана стоимость перевозок единицы продукта. Необходимо при минимальных затратах удовлетворить спрос.

Задача Хичкока – задано множество портов отправления и назначения. Дана матрица количества перевозок, числа кораблей в каждом из портов назначения, количество товаров, необходимых в разных портах назначения. Нужно найти план перевозок с минимальными издержками.

10. НЕЧЕТКАЯ ЛОГИКА

10.1. Нечеткая логика, как обобщение Аристотелевой логики.

10.2. Нечеткие высказывания и нечеткие модели систем.

10.3. Математическое обеспечение системы оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику.

10.1. Нечеткая логика, как обобщение Аристотелевой логики

Нечеткая логика – это обобщение традиционной Аристотелевой логики на случай, когда истинность рассматривается как лингвистическая переменная, принимающая значения типа: «очень истинно», «бо-

лее-менее истинно», «не очень ложно» и т. п. Указанные лингвистические значения представляются нечеткими множествами.

Лингвистической называется переменная, принимающая значения из множества слов или словосочетаний некоторого естественного или искусственного языка. Множество допустимых значений лингвистической переменной называется *терм-множеством* (рис. 10.1).

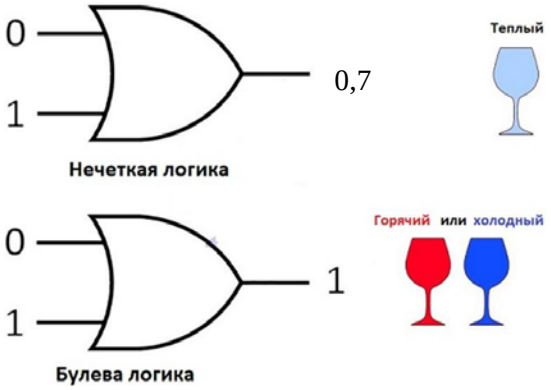

Рис. 10.1. Сравнение четкой и нечеткой логики

Задание значения переменной словами, без использования чисел, для человека более естественно. Ежедневно мы принимаем решения на основе лингвистической информации типа: «очень высокая температура»; «длительная поездка»; «быстрый ответ»; «красивый букет»; «гармоничный вкус» и т. п. Психологи установили, что в человеческом мозге почти вся числовая информация вербально перекодируется и хранится в виде лингвистических термов. Понятие лингвистической переменной играет важную роль в нечетком логическом выводе и в принятии решений на основе приближенных рассуждений. Формально, лингвистическая переменная определяется следующим образом.

Определение. *Лингвистическая переменная* задается пятеркой $\langle x, T, U, G, M \rangle$, где x – имя переменной; T – терм-множество, каждый элемент которого (терм) представляется как нечеткое множество на универсальном множестве; UG – синтаксические правила, часто в виде грамматики, порождающие названия термов; M – семантические правила, задающие функции принадлежности нечетких термов, порожденных синтаксическими правилами G.

Нечеткие логические операции. Вначале кратко напомним основные положения обычной (булевой) логики. Рассмотрим два утверждения A и B, каждое из которых может быть истинным или ложным, т. е. принимать значения «1» или «0». Для этих двух утверждений всего существует $2^{2^2} = 16$ различных логических операций, из которых содержательно интерпретируются лишь пять: И($\wedge$), ИЛИ ($\vee$), исключающее ИЛИ ($\oplus$), импликация ($\Rightarrow$) и эквивалентность ($\Leftrightarrow$).

В многозначной логике логические операции могут быть заданы таблицами истинности. В нечеткой логике количество возможных значений истинности может быть бесконечным, следовательно, в общем виде табличное представление логических операций невозможно. Однако в табличной форме можно представить нечеткие логические операции для ограниченного количества истинностных значений, например для терм-множества {«истинно», «очень истинно», «не истинно», «более-менее ложно», «ложно»}.

В результате выполнения логических операций часто получается нечеткое множество, которое не эквивалентно ни одному из ранее введенных нечетких значений истинности. В этом случае необходимо среди нечетких значений истинности найти такое, которое соответствует результату выполнения нечеткой логической операции в максимальной степени. Другими словами, необходимо провести так называемую *лингвистическую аппроксимацию*, которая может рассматриваться как приближение эмпирического распределения стандартными функциями случайных величин.

Определение. *Нечеткой базой знаний* называется совокупность нечетких правил «Если…то», определяющих взаимосвязь между входами и выходами исследуемого объекта. Обобщенный формат нечетких правил такой: **Если** *посылка правила,* **то** *заключение правила.*

В узком смысле нечеткая логика – это логическое исчисление, являющееся расширением многозначной логики. В ее широком смысле, который сегодня является преобладающим в использовании, *нечеткая логика* равнозначна теории нечетких множеств. С этой точки зрения, *нечеткая логика* в узком смысле является разделом нечеткой логики в широком смысле.

Наверное, самым впечатляющим у человеческого интеллекта является способность принимать правильные решения в условиях неполной и нечеткой информации. Построение моделей приближенных размышлений человека и использование их в компьютерных системах представляет в настоящее время одну из важнейших проблем науки.

10.2. Нечеткие высказывания и нечеткие модели систем

Нечеткими высказываниями будем называть высказывания следующего вида:

1. Высказывание <*b* есть *b*'>, где *b* – наименование лингвистической переменной, *b*' – ее значение, которому соответствует нечеткое множество на универсальном множестве *X*, например, высказывание <*давление большое*> предполагает, что лингвистической переменной *«давление»* придается значение *«большое»*, для которого на универсальном множестве *X* переменной *«давление»* определено соответствующее данному значению «*большое*» нечеткое множество.

2. Высказывание <*b* есть *mb*'>, где *m* – модификатор, которому соответствуют слова *ОЧЕНЬ, БОЛЕЕ ИЛИ МЕНЕЕ, МНОГО БОЛЬШЕ* и др., например, <*давление очень большое*>, <*скорость много больше средней*> и др.

3. Составные высказывания, образованные из приведенных выше высказываний видов 1 и 2 и союзов *И, ИЛИ, ЕСЛИ…, ТО..., ЕСЛИ, ТО..., ИНАЧЕ*. Логико-лингвистические методы описания систем основаны на том, что поведение исследуемой системы описывается на естественном (или близком к естественному) языке в терминах лингвистических переменных. Входные и выходные параметры системы рассматриваются как лингвистические переменные, а качественное описание процесса задается совокупностью высказываний следующего вида:

L_1: если $<A_1>$ то $<B_1>$,
L_2: если $<A_2>$ то $<B_2>$,
……
L_k: если $<A_k>$ то $<B_k>$,

где $<A_i>$, i = 1, 2, …, k – составные нечеткие высказывания, определенные на значениях входных лингвистических переменных, а $<B_i>$, i = = 1, 2, …, k – высказывания, определенные на значениях выходных лингвистических переменных.

С помощью правил преобразования дизъюнктивной и конъюнктивной форм описание системы можно привести к такому виду:

L_1: если $<A_1>$ то $<B_1>$,
L_2: если $<A_2>$ то $<B_2>$,
……
Lk: если $<A_k>$ то $<B_k>$,

где A_1, A_2,…, A_k – нечеткие множества, заданные на декартовом произведении X универсальных множеств входных лингвистических переменных, а B_1, B_2,…, B_k – нечеткие множества, заданные на декартовом произведении Y универсальных множеств выходных лингвистических переменных.

Совокупность импликаций $\{L_1, L_2,\ldots, L_k\}$ отражает функциональную взаимосвязь входных и выходных переменных и является основой построения нечеткого отношения XRY, заданного на произведении $X´Y$ универсальных множеств входных и выходных переменных. Если на множестве X задано нечеткое множество A, то композиционное правило вывода $B = A \cdot R$ определяет на Y нечеткое множество B с функцией принадлежности $mB(Y) = V_x\,(mA(X)\ LmR(X, Y))$.

Таким образом, композиционное правило вывода в этом случае задает закон функционирования нечеткой модели системы.

10.3. Математическое обеспечение системы оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику

Математическая база создания информационных систем (ИС) новых поколений формируется на основе эффективного сочетания накопленной системы знаний с новыми подходами и парадигмами искусственного интеллекта (ИИ). Среди них важная роль принадлежит методам и моделям, обеспечивающим формализацию и интеграцию знаний, механизм логического вывода, поиск решений и выдачу практических рекомендаций. Наряду с традиционными методами «инженерии знаний здесь находит применение концепция «мягких вычислений» (рис. 10.2).

Принцип формализации нечеткой информации в мультипроцессорной вычислительной среде позволяет осуществлять параллельные цепочки нечеткого вывода в непрерывно изменяющихся условиях динамики объекта и внешней среды. Использование этого принципа в ИС поддержки принятия решений дает следующие преимущества:

открывает перспективы программной реализации сложных моделей представления и обработки нечеткой системы знаний;

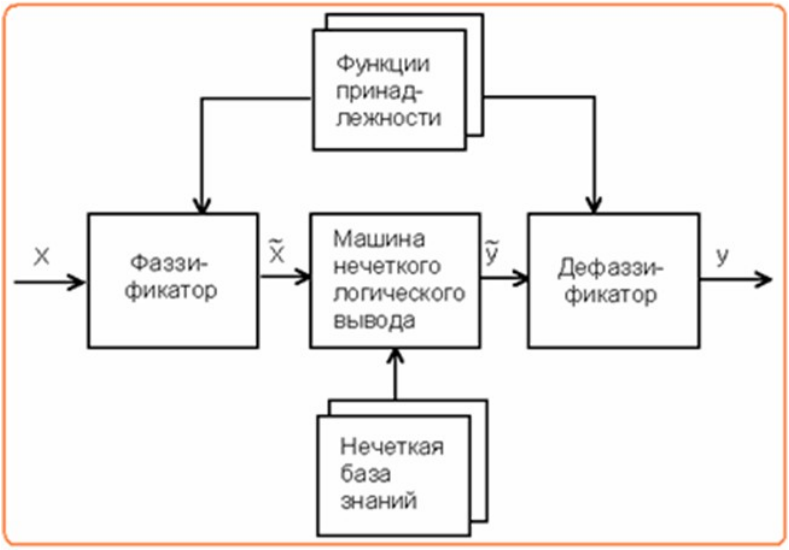

Рис. 10.2. Концепция «мягких» вычислений

обеспечивает функционирование комплекса в режиме реального времени и сокращает расходы на разработку аппаратного обеспечения механизма нечеткого вывода;

устраняет трудности решения задач при распараллеливании вычислительного процесса с существенной нерегулярностью вычислений, характерной для задач поддержки принятия решений, реализуемых на базе интегрированных комплексов.

Недерминированность выводов. Практически во всех системах искусственного интеллекта знания накапливаются фрагментарно, и нельзя априори определить цепочку логических выводов, в которых они будут использоваться. Последовательность действий при поиске решения заранее не может быть определена, и необходимо методом проб и ошибок выбрать некую цепочку выводов, а в случае неуспеха организовать перебор с возвратами для поиска другой цепочки и т. д.

Многозначность. О многозначности знаний говорят тогда, когда один и тот же элемент знаний (понятие, символ, звук, изображение и т. п.) может быть интерпретирован по-разному. Многозначность интерпретации – обычное явление при понимании естественных языков и распознавании изображений и речи. Ненадежными являются те знания,

представить которые двумя значениями – истина или ложь – невозможно или трудно. В современной физике и технике такую ненадежность представляют вероятностью, подчиняющейся законам Байеса. Но при обработке знаний было бы нелогично иметь дело со степенью надежности, приписанной знаниям изначально, и поэтому применяются специфические методы работы с ненадежными знаниями.

Неполнота знаний. Полностью описать окружающий мир чрезвычайно сложно. Содержимое базы знаний по любой предметной области является неполным, поскольку можно (хотя и трудно) перечислить все верные знания в данной области, но невозможно перечислить и разумно определить неверные знания. Поэтому целесообразно в базе знаний определять только заведомо верные знания, а любые утверждения, которые не определены, относить к ложным. Это называется гипотезой закрытого мира.

Нечеткие экспертные системы представляют знания в форме нечетких продукций и совокупности лингвистических переменных. Основу представления лингвистической переменной составляет терм с функцией принадлежности. Способ обработки знаний – это вывод по нечетким продукциям. Особенностью конкретной нечеткой ЭС являются способы извлечения функций принадлежности, которые сводятся либо к статистическим методам построения из гистограммы, либо к вариантам метода экспертных оценок.

Мягкой экспертной системой (ЭС) будем называть нечеткую ЭС, которая обладает следующими особенностями (рис. 10.3).

Мягкая ЭС для извлечения знаний использует статистические данные, которые интерпретирует как обучающие выборки для нечетких нейронных сетей. Мягкая ЭС представляет знания как совокупность лингвистических переменных (функций принадлежности), нечетких продукций и обученных нейронных сетей, функций свертки критериев при многокритериальном выборе. Редукция множества нечетких продукций выполняется с помощью генетических алгоритмов. Мягкая ЭС сочетает шаги вывода по нечетким продукциям с шагами вывода на базе онтологий.

Таким образом, если мягкими называют вычисления, сочетающие теорию нечетких систем, нейронные сети, вероятностные рассуждения и генетические алгоритмы и обладающие синергическим эффектом, то мягкой называют ЭС, сочетающую перечисленные теории ради того же эффекта взаимного усиления.

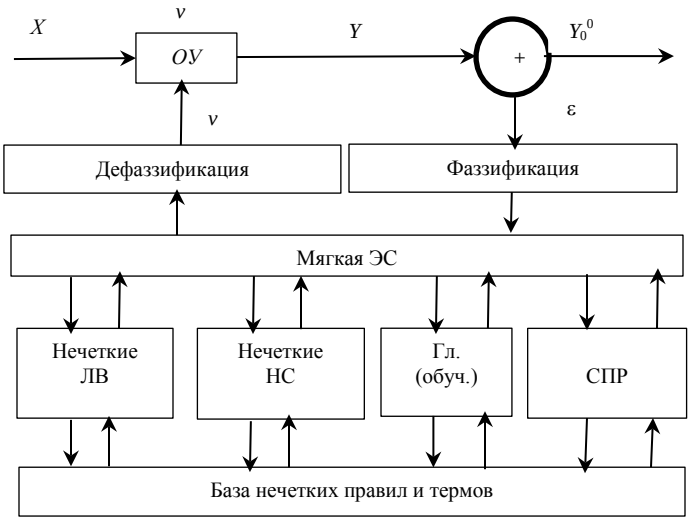

Рис. 10.3. Мягкая экспертная система

11. ДОКУМЕНТАЛЬНЫЕ ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

11.1. Документальный поиск.
11.2. Библиографический поиск.

11.1. Документальный поиск

Документальный информационный поиск, поиск документов (document retrieval) – вид информационного поиска, связанный с процессами нахождения и выдачи документов.

Задача *документального информационного поиска или отбора документов* (ОД) сводится к тому, чтобы, не прочитывая текстов множества документов (T_i), по каким-то внешним описательным признакам выбирать из этого множества такие документы (D_i), которые по смыслу соответствуют информационному запросу (C_i). Для этого каждый документ снабжается поисковым образом – характеристикой (P_i), в которой кратко и однозначно выражается основное смысловое содержание документа. В виде такой же краткой и однозначной записи – поискового предписания – должен быть сформулирован и информационный запрос (рис. 11.1).

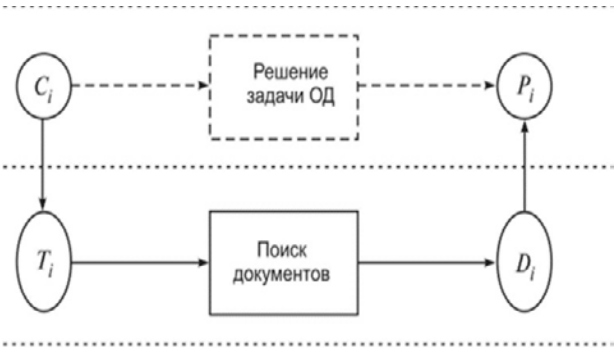

Рис. 11.1. Процедура поиска документов

Благодаря этому процедура информационного поиска может быть сведена к простому сопоставлению поисковых образов документов с заданным поисковым предписанием. Если поисковый образ документа в установленной степени совпадает с поисковым предписанием, то считается, что этот документ отвечает на информационный запрос [28].

Если в *документальном информационном поиске* обратная связь с абонентом служит для уточнения информационных потребностей и коррекции запроса для поиска в существующем банке документов, то обратная связь в фактографии, кроме коррекции запроса, главным образом должна служить для уточнения тематики при формировании баз данных АФИПС. Необходимость в постоянной актуализации фактографической информации в отношении существующих информационных потребностей является характерным признаком фактографических ИПС.

Очевидно, что описанный метод *документального информационного поиска* может быть применен как в одноконтурных, так и в двухконтурных информационно-поисковых системах (ИПС), т. е. в ИПС с одним или двумя ЗУ – активным и пассивным.

Человек в силу определенной ограниченности своих возможностей более не может, пользуясь лишь традиционными библиотечно-библиографическими методами и средствами, удовлетворительно решать задачу *документального информационного поиска*. Возникла острая практическая потребность в создании машинных ИПС, которые могли бы значительно эффективнее, чем человек, выполнять формально-логические операции информационного поиска.

Информационный поиск (Information retrieval) – процесс поиска неструктурированной документальной информации, удовлетворяющей информационные потребности, и наука об этом поиске.

Термин «информационный поиск» был впервые введен Кельвином

Муром в 1948 г. в его докторской диссертации, опубликован и употребляется в литературе с 1950 г.

Виды поиска:

1) полнотекстовый поиск – поиск по всему содержимому документа. Пример полнотекстового поиска – любой интернет-поисковик, например www.yandex.ru, www.google.com. Как правило, полнотекстовый поиск для ускорения поиска использует предварительно построенные индексы. Наиболее распространенной технологией для индексов полнотекстового поиска являются инвертированные индексы;

2) поиск по метаданным – это поиск по неким атрибутам документа, поддерживаемым системой: название документа, дата создания, размер, автор и т. д. Пример поиска по реквизитам – диалог поиска в файловой системе (например, MS Windows);

3) поиск изображений – поиск по содержанию изображения. Поисковая система распознает содержание фотографии (загружена пользователем или добавлен URL изображения). В результатах поиска пользователь получает похожие изображения. Так работают поисковые системы Polar Rose, Picollator и др. (рис. 11.2).

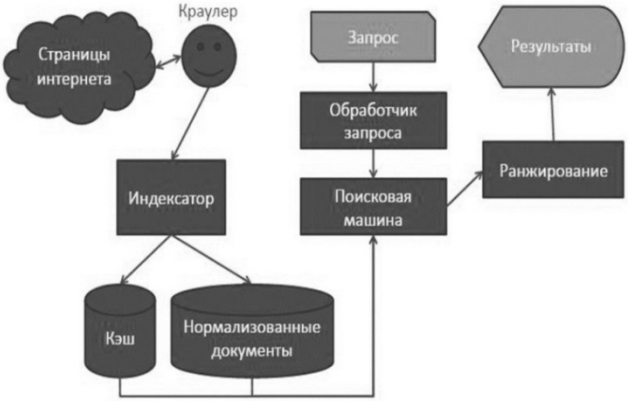

Рис. 11.2. Общая структура системы информационного поиска

Библиографический поиск – информационный поиск, осуществляемый на основании библиографических данных.

Документальный поиск – информационный поиск, при котором объектом поиска являются документы; процесс поиска в хранилище информационно-поисковой системы первичных документов или в базе данных вторичных документов, соответствующих запросу пользователя. Существует два вида документального поиска (рис. 11.3).

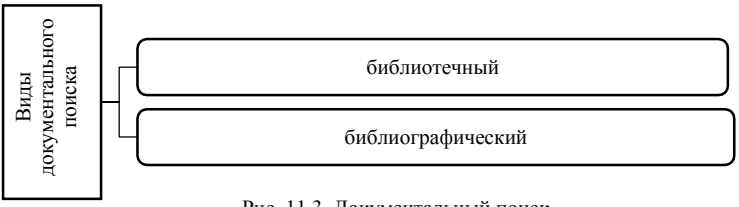

Рис. 11.3. Документальный поиск

На рис. 11.3 представлены два основных вида документального поиска:

1) *библиотечный*, направленный на нахождение первичных документов;

2) *библиографический*, направленный на нахождение сведений о документах, представленных в виде библиографических записей.

Фактографический поиск – информационный поиск, при котором отыскиваемая информация имеет характер конкретных фактических сведений (в отличие от документального поиска, позволяющего получить сведения лишь об источниках информации); процесс поиска фактов, соответствующих информационному запросу (рис. 11.4).

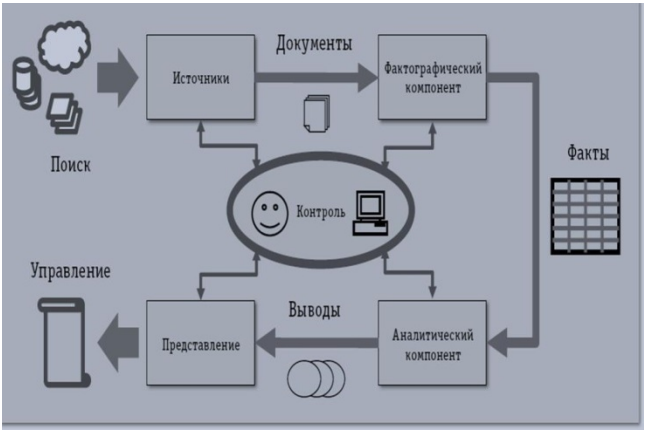

Рис. 11.4. Функциональная схема факто-аналитической программной системы

К фактографическим данным относятся сведения, извлеченные из документов как первичных, так и вторичных и получаемые непосредственно из источников их возникновения.

Различают два вида:

1) *документально-фактографический*, заключается в поиске в документах фрагментов текста, содержащих факты;

2) *фактологический* (описание фактов), предполагающий создание новых фактографических описаний в процессе поиска путем логической переработки найденной фактографической информации.

11.2. Библиографический поиск

Библиографический поиск – это информационный поиск, т. е. отбор библиографируемых документов из числа выявленных с целью их последующей библиографической обработки (записи), осуществляемый на основании библиографических данных (рис. 11.5).

В библиографии выделяют три основные цели информационного поиска:

1. Поиск необходимых сведений об источнике и установление его наличия в системе других источников. Ведется путем разыскания библиографической информации и библиографических пособий (информационных изданий), специально создаваемых для более эффективного поиска и использования информации (литературы, книги).

2. Поиск самих информационных источников (документов и изданий), в которых есть или может содержаться нужная информация.

3. Поиск фактических сведений, содержащихся в литературе, книге, например, об исторических фактах и событиях, о технических характеристиках машин и процессов, о свойствах веществ и материалов, о биографических данных из жизни и деятельности писателя, ученого и т. п.

Ниже приведены методы библиографического поиска.

Сплошной метод. При сплошном методе «библиограф для осуществления поставленной задачи обследует сплошь и без пропусков все наличие имеющихся пособий и источников...»

Выборочный метод. Более рациональный и реальный путь поиска литературы – выборочный метод, т. е. «ограниченно сплошной». В литературе его часто называют также «эпизодический метод».

Интуитивный метод. Индивидуальный подход к поиску необходимых источников на основе предположения либо базовых знаний, с учетом конкретизации по какому-либо типу (автор произведений, жанр, издательство).

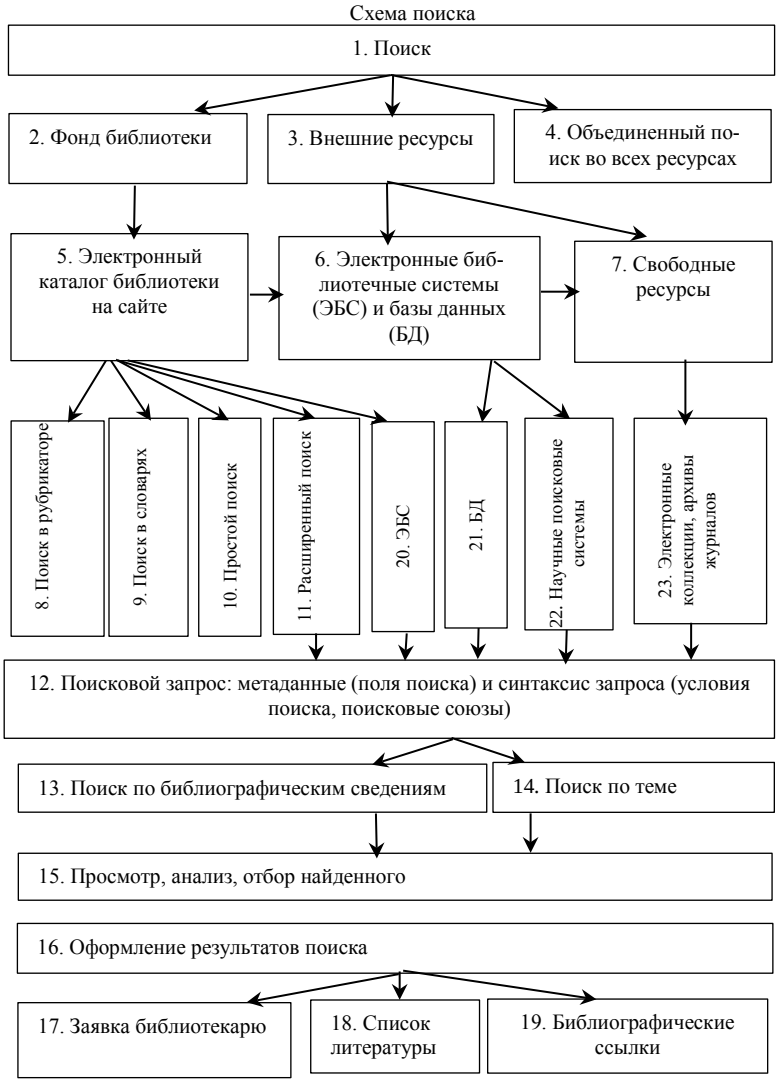

Рис. 11.5. Схема библиографического поиска

12. СИСТЕМЫ, ОСНОВАННЫЕ НА ЗНАНИЯХ

12.1. Понятие интеллектуальной системы.

12.2. Инженерия знаний.

12.3. Использование систем, основанных на знаниях, в оптимизации параметров функционирования объектов и процессов агропромышленного производства.

12.1. Понятие интеллектуальной системы

В 1950 г. британский математик Алан Тьюринг опубликовал в журнале «Mind» свою работу «Вычислительная машина и *интеллект*», в которой описал тест для проверки программы на интеллектуальность. Он предложил поместить исследователя и программу в разные комнаты и до тех пор, пока исследователь не определит, кто за стеной – человек или *программа*, считать поведение программы разумным. Это было одно из первых определений интеллектуальности, т. е. А. Тьюринг предложил называть интеллектуальным такое поведение программы, которое будет моделировать разумное поведение человека (рис. 12.1).

Рис. 12.1. Этапы проектирования системы искусственного интеллекта

С тех пор появилось много *определений интеллектуальных систем* (*ИС*) и *искусственного интеллекта* (*ИИ*). Сам термин *ИИ (AI* –

Artificial Intelligence) был предложен в 1956 г. на семинаре в Дартмутском колледже (США). Приведем некоторые из этих определений. Д. Люгер в своей книге определяет «***ИИ*** как область компьютерных наук, занимающуюся исследованием и автоматизацией разумного поведения» [14, 39, 40].

Адаптивная система – это система, которая сохраняет работоспособность при непредвиденных изменениях свойств управляемого объекта, целей управления или окружающей среды путем смены *алгоритма функционирования*, программы поведения или поиска оптимальных, в некоторых случаях просто эффективных, решений и состояний. Традиционно, по способу адаптации различают самонастраивающиеся, самообучающиеся и *самоорганизующиеся системы*.

Под *алгоритмом* будем понимать последовательность заданных действий, которые однозначно определены и выполнимы на современных ЭВМ за приемлемое время для решаемой задачи.

Под ***ИС*** будем понимать *адаптивную систему*, позволяющую строить программы целесообразной деятельности по решению поставленных перед ними задач на основании конкретной ситуации, складывающейся на данный момент в окружающей их среде.

Интеллектуальные робототехнические системы (*ИРС*) содержат переменную, настраиваемую модель внешнего мира и реальной исполнительной системы с объектом управления. Цель и управляющие воздействия формируются в *ИРС* на основе *знаний* о внешней среде, объекте управления и на основе моделирования ситуаций в реальной системе.

О каких признаках интеллекта уместно говорить применительно к *интеллектуальным системам*? *ИС* должна уметь в наборе фактов распознать существенные, *ИС* способны из имеющихся фактов и *знаний* сделать выводы не только с использованием *дедукции*, но и с помощью аналогии, индукции и т. д. Кроме того, *ИС* должны быть способны к самооценке – обладать рефлексией, т. е. средствами для оценки результатов собственной работы. С помощью подсистем объяснения *ИС* может ответить на вопрос, почему получен тот или иной результат. Наконец, *ИС* должна уметь обобщать, улавливая сходство между имеющимися фактами.

12.2. Инженерия знаний

Инженерия знаний представляет собой совокупность моделей, методов и технических приемов, нацеленных на создание систем, кото-

рые предназначены для решения проблем с использованием знаний. Фактически инженерия знаний – это теория, методология и технология, которые охватывают методы добычи, анализа, представления и обработки знаний экспертов [15, 17].

Представление знаний, их обработка и использование, рассматриваемые применительно к конкретной прикладной области, являются предметом инженерии знаний.

На высоком уровне процесс инженерии знаний состоит:

- из извлечения знаний – преобразования сырых знаний в организованные;

- внедрения знаний – преобразования организованных знаний в реализованные.

С областью инженерии знаний тесно связано понятие искусственного интеллекта (ИИ).

Сущностью искусственного интеллекта можно считать научный анализ и автоматизацию интеллектуальных функций человека. Однако для большинства проблем общая реальность – трудность их машинного воплощения. Исследования по ИИ позволили утвердиться во мнении, что подлинно необходимым для решения проблем являются знания экспертов, т. е. если создать систему, способную запоминать и использовать знания экспертов, то она найдет применение в практической деятельности.

Инженерия знаний тесно связана со всем процессом разработки интеллектуальных информационных систем в целом и экспертных систем (ЭС) в частности – от возникновения замысла до его реализации и совершенствования (рис. 12.2).

В конце 1960-х – начале 1970-х гг. под руководством Э. Фейгенбаума в Стэнфордском университете США была создана система DENDRAL, а позднее – MYCIN. Поскольку данные системы накапливают в памяти компьютера знания экспертов и используют эти знания для решения проблем, извлекая их при необходимости из памяти, то они получили название экспертных, а профессор Э. Фейгенбаум, являющийся одним из создателей экспертных систем, выдвинул для данной области техники название «инженерия знаний». Слово «engineering» в английском языке означает искусную обработку предметов, изобретение или создание чего-либо. Следовательно, работу по оснащению программ специальными экспертными знаниями из проблемной области, выполняемую человеком либо компьютером (программой), также можно назвать *инженерией знаний*.

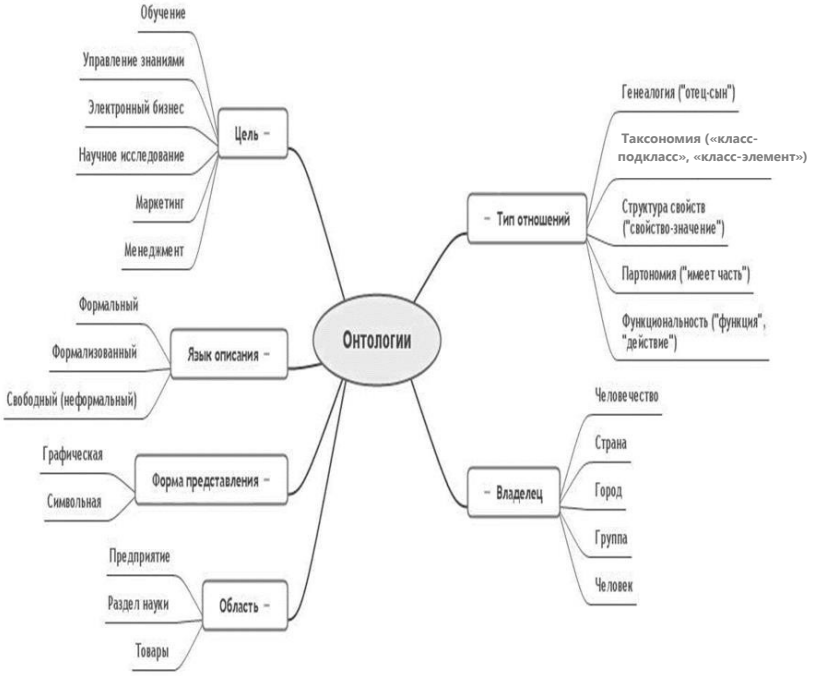

Рис. 12.2. Онтологии в инженерии знаний

12.3. Использование систем, основанных на знаниях, в оптимизации параметров функционирования объектов и процессов агропромышленного производства

Системный анализ и имитационные модели посевов обеспечивают альтернативный метод управления сельскохозяйственным производством. Используя модели развития и роста посевов, можно разработать сценарии управления, провести анализ стратегий для оценки влияния изменения климата в сельскохозяйственном производстве и агротехнологической адаптации. Исходя из заданных входных параметров, системы обеспечивают разработку и выдачу инженерно-технологического проекта получения урожая на конкретном поле, формируя последовательное описание всех технологических процессов и операций для получения запланированного урожая культур.
Экспертные системы в сельском хозяйстве осуществляют:

- планирование программ агротехнических мероприятий для конкретных полей, на которых будут выращиваться культуры;

- определение параметров управления, срок проведения операций, их характеристики и условия воспроизводства;

- коррекцию информационной базы проектирования согласно новым представлениям о технологии обработки;

- выдачу обоснованных рекомендаций;

- автоматизацию системы оперативного управления технологическим процессом возделывания сельскохозяйственных культур системами экономических расчетов.

Применение систем позволяет улучшить, ускорить и удешевить процесс проектирования и обеспечить получение рекомендаций, адекватных свойствам конкретного посева, поля, оборудования.

Разработанные программные алгоритмы и программные комплексы позволят:

- проектировать технологию выращивания сельскохозяйственных культур в целом и фрагментарно;

- планировать агроприемы на различный временной период;

- обеспечить расчеты действительно возможного урожая и затрат на его получение;

- установить нормы и сроки проведения поливов;

- управлять режимом подпитки почвы.

Приведем для сравнения две экспертные системы DSSAT – система поддержки агротехнологических решений, разработанная группой специалистов-инициаторов из трех американских штатов: Флорида, Мичиган, Гавайи, и система поддержки принятия решений при производстве сельскохозяйственной продукции «Геомир», которая начала работу с 80-х гг. и в настоящее время является единственной в Российской Федерации компанией, которая предлагает агропромышленным и другим предприятиям разработку и «сдачу под ключ» информационно-аналитических систем для эффективного производства.

DSSAT – система поддержки агротехнологических решений. Направлена на проверку правильности решений, которые принимает фермер: от выбора культур на разных участках земли к использованию различных технологий, выбора промежутков времени, планирования посевов, полива, использования удобрений.

Геомир – система поддержки принятия решений при производстве сельскохозяйственной продукции. В 2004 г. на 32-м Международном салоне инноваций и изобретений в Швейцарии «Геомир» представила свой проект «Система поддержки принятия решений при производстве

сельскохозяйственной продукции», стала победителем в разделе «Сельское хозяйство» и была награждена Золотой медалью Салона.

Для управления сельскохозяйственным предприятием, производящим продукцию растениеводства, необходима объективная информация относительно размера и состояния сельхозугодий. Большой объем пространственной и атрибутивной информации качественно можно обрабатывать и анализировать только при помощи специального программного обеспечения, учитывающего как пространственную привязку, так и специальные сведения о полях.

Имеющиеся в хозяйстве картографические материалы можно условно разделить на три группы:

- землеустроительные (планы внутрихозяйственного землеустройства);
- грунтовые (почвенные карты);
- агрохимические или агрохимические картограммы (содержания гумуса, подвижного фосфора, подвижного калия, pH) [27].

В полном варианте агрономическая ПС должна включать многослойную электронную карту хозяйства и атрибутивную базу данных истории полей, с учетом предварительных агротехнических мероприятий. Количество тематических слоев электронной карты зависит от сложности ландшафтно-экологических условий и уровня интенсификации агротехнологий (определяется по параметрам урожайности и объем затрат на гектар). В общем виде электронная карта полей должна включать:

- мезорельєф (с показом мезоформ рельефа, форм склонов);
- крутизну склонов;
- экспозицию склонов (теплые, холодные, нейтральные);
- микрорельеф (с показом контуров с преобладанием тех или иных форм микрорельефа, имеющих агрономическое значение);
- микроклимат;
- уровень грунтовых вод, их минерализацию и состав;
- почвообразовательные и подстилающие породы;
- микроструктуру почвенного покрова (почвенная карта);
- содержание гумуса в почве;
- обеспеченность подвижными формами элементов минерального питания растений и микроэлементами;
- значение pH почв;
- физические свойства почв;
- загрязнения тяжелыми металлами, радионуклидами и другими токсикантами;

- эродированность почв, эрозионную опасность и другие виды физической деградации (оползни, сели и т. п.);

- переувлажнения и заболачивания почв;

- растительный покров с оценкой состояния природных кормовых угодий;

- лесную растительность с оценкой состояния естественных лесов и лесных насаждений;

- распределение полезных видов животных, птиц, полезных энтомофагов, оценку их территориального влияния;

- фитосанитарное состояние посевов.

ПС Панорама АГРО предназначена для комплексной автоматизации управления сельскохозяйственным предприятием в отрасли растениеводства и обеспечивает решение двух взаимосвязанных задач: управление аграрными технологиями и мониторинг подвижных технических средств компании на основе GPS/ГЛОНАСС навигации.

Основные функции программы:

1. Ведение нормативно-справочной информации. Ведение паспортов полей с привязкой к году урожая.

2. Общие сведения о земельных угодьях (параметры поля; сведения о севообороте; механический состав почв; агрохимический состав почв; сведения о фитосанитарном состоянии; привязка к карте земельных угодий; привязка к карте инфраструктуры предприятия).

Управление электронной картой:

- составом слоев электронной карты;

- составом растров;

- составом матриц;

- подключение атласа карт.

Создание и редактирование электронной карты – создание контуров полей по изображению карты или космоснимки. Расчеты по карте. Обработка навигационных данных и контроль перемещений автотранспорта и специальной техники (ведение списка объектов мониторинга; визуализация перемещений объектов мониторинга на фоне карты; в режиме реального времени; в режиме прокрутки истории; расчет и отображение показателей мониторинга; анализ показателей мониторинга на графиках).

Формирование и анализ событий, происходящих с объектами мониторинга, по данным и показателям: датчиков; взаимного расположения объектов мониторинга; маршрутов и геозон.

Планирование и учет перемещений автотранспорта и специальной техники:

- ведение реестра маршрутов и геозон;
- ведение реестра пунктов (транспортных узлов);
- планирование работ водителей и механизаторов;
- автоматизированный учет выполненных работ;
- создание и редактирование карты маршрутов и геозон.

Встроенная подсистема управления графом дорог:

- определение минимального пути между двумя произвольными точками;
- определение минимального пути между двумя объектами карты;
- определение минимального пути между объектом мониторинга и указанным объектом карты;
- определение минимального пути между объектом мониторинга и пунктом назначения – транспортным узлом.

Обработка результатов полевых измерений, данных дистанционного зондирования и обновление карты земельных угодий:

- использование возможностей Google для обновления карты;
- загрузка данных из файлов формата SHAPE;
- загрузка данных от автопилотов;
- редактирование карты на основе треков объектов мониторинга.

Построение тематических карт отдельных показателей земельных угодий на основании сведений, представленных в паспортах полей:

- автоматическое создание цветовых картограмм;
- автоматическое создание карт условных знаков;
- управление составом отображаемых данных.

Планирование и учет технологических операций в соответствии с установленным севооборотом:

- составление базовой технологической карты культуры;
- привязка технологической карты к полям в соответствии с севооборотом;
- планирование технологических операций;
- автоматизированный учет технологических операций.

План-фактный анализ технологических операций:

- агротехнические мероприятия;
- внесение удобрений;
- внесение мелиорантов;
- внесение средств защиты растений.

Формирование отчетов и статистических справок.

Отчеты по технике: парк техники; парк объектов мониторинга; парк навесного оборудования.

Отчеты по выполненным работам: оперативный учет; фактические работы механизаторов; обработанная площадь; расход топлива за период (автомобили); расход топлива за период (специальная техника); пользовательский отчет.

Отчеты по полям:

- паспорт поля;
- структура земель и пашни;
- валовой сбор культур;
- урожайность по хозяйству;
- высев семян;
- внесение средств защиты растений;
- внесение мелиорантов;
- внесение удобрений;
- выполненные технологические операции.

Отчеты по событиям объектов мониторинга:

- ведение ресурсов системы и разграничение доступа;
- обмен данными с внешними программами.

13. СТРУКТУРНЫЙ АНАЛИЗ СИСТЕМ

13.1. Функционально-структурный анализ систем и его особенности.

13.2. Использование структурного анализа систем в технологиях интеллектуального анализа данных.

13.3. Когнитивный анализ и моделирование сложных ситуаций.

13.1. Функционально-структурный анализ систем и его особенности

В процессе функционально-структурного анализа изучаются алгоритмы функционирования и состав как самой исследуемой системы, так и ее подсистем и элементов. Цель функционально-структурного анализа заключается в определении и уточнении следующих факторов:

- закономерностей и алгоритмов функционирования системы в целом, ее подсистем и элементов;
- взаимодействий и взаимовлияний подсистем и элементов;
- пространства состояний системы;
- совокупности управляемых и неуправляемых параметров системы;
- состава системы;

• связей между подсистемами и элементами системы;
• структурных свойств системы и ее подсистем.

Выполнение этих операций должно завершаться выработкой заключения об оптимальности алгоритмов функционирования и структуры системы и выработкой рекомендаций и путей совершенствования системы (рис. 13.1).

Рис. 13.1. Структура системного анализа

Структура системы описывает общую ее конфигурацию. Разделение системы на части (блоки) осуществляется исходя из целей исследования и требуемой детализации описания структуры и выполняемых функций. Структура системы обычно изображается в виде так называ-

емых структурных схем благодаря их наглядности и информативности. Однако эти схемы практически не поддаются формализации, что позволило бы не только наглядно изобразить структуру, но и выполнить аналитические исследования. Инструментом, обеспечивающим решение последней упомянутой задачи, является представление структурной схемы в виде соответствующего графа. Это позволит использовать хорошо развитый аппарат теории графов для анализа и оценки структурных схем.

Структурные свойства системы определяются характером отношений между частями (блоками) системы. По этому признаку системы подразделяются на иерархические, многосвязные, смешанные. В иерархических структурах компоненты системы упорядочены по степени важности. Среди них имеются структуры строгой и нестрогой иерархии. Структуры строгой иерархии имеют следующие особенности:

• имеется только один главный управляющий компонент с числом связей не менее двух;

• исполнительные компоненты имеют только одну связь с компонентом вышележащего уровня и не менее двух связей с компонентами нижнего уровня.

В структурах нестрогой иерархии могут иметь место связи через один или несколько уровней иерархии. Особенностью иерархических структур является отсутствие горизонтальных связей на одном уровне иерархии, что, в общем-то, является некоторой абстракцией. В реальных иерархических системах такие связи присутствуют. В неиерархических (многосвязных) структурах компоненты могут быть как исполнительными, так и управляющими, причем каждый компонент взаимодействует более чем с одним компонентом.

По степени определенности связей между частями (блоками) системы обычно различают детерминированные, вероятностные и хаотические структуры. В детерминированных структурах связь между частями системы описывается определенными функциональными зависимостями, в вероятностных – случайными величинами с известными законами распределения вероятностей. Для хаотических структур характерно отсутствие ограничений, части системы связываются случайно без установленных законов распределения случайностей.

13.2. Использование структурного анализа систем в технологиях интеллектуального анализа данных

Предметно-ориентированные аналитические системы очень разнообразны. Наиболее широкий подкласс таких систем, получивший рас-

пространение в области исследования финансовых рынков, носит название «технический анализ». Он представляет собой совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля, основанных на различных эмпирических моделях динамики рынка. Эти методы часто используют несложный статистический аппарат, но максимально учитывают сложившуюся в своей области специфику (профессиональный язык, системы различных индексов и пр.) [47, 48].

В результате научных экспериментов и компьютерного моделирования постоянно накапливаются большие объемы данных, которые организуются в электронные информационные ресурсы: базы и хранилища данных, электронные информационные и вычислительные системы, научные центры данных. Такие информационные ресурсы становятся местом накопления, хранения, верификации, извлечения, использования и распространения профессиональных и корпоративных знаний. Эффективное развитие науки и высоких технологий требует интенсивной обработки и анализа фундаментальных знаний, накопленных в различных исследовательских организациях, что приводит к потребности в развитии информационных технологий накопления, извлечения и анализа предметно-ориентированных профессиональных знаний на основе разработки универсальных и специализированных моделей организации и представления научных данных и знаний в электронных ресурсах.

Таким образом, создание информационных ресурсов в Интернете, предназначенных для сбора, хранения, верификации, извлечения, распространения и производства новых профессиональных знаний в различных предметных областях, является актуальной научной и научно-практической задачей.

В национальном стандарте США [3] термин «производство знаний» определяется:

как разработка и обеспечение новых знаний (JECD 1966:2);

обстоятельства, при которых люди, группы людей и организации успешно генерируют новые знания и практики (OECD 2000:39).

Под *производством знаний* понимается извлечение новых знаний из эмпирических данных в рамках компьютерной системы с участием человека и использованием методов прикладного искусственного интеллекта (рис. 13.2).

Рис. 13.2. Аналитические инструменты извлечения знаний и уровни анализа

С учетом требования анализа и производства знаний в программно-технологическую архитектуру предметно-ориентированных систем научной осведомленности предъявляются следующие дополнительные программные компоненты:

1. Компонент фактографических научных баз данных, содержащих экспериментальные или модельные данные: фундаментальные константы, числовые и лингвистические характеристики химических или физических процессов.

2. Компонент интеллектуального анализа данных. Поскольку новые инструменты научных исследований обладают исключительной точностью, увеличиваются точность и качество фактографических данных. Для анализа таких данных с целью нахождения тонких эффектов, упущенных в предыдущих исследованиях, требуется набор алгоритмов, позволяющий проводить сложный анализ данных.

3. Компонент производства новых знаний, который позволяет строить прогнозы значений физических или химических процессов и оценивать значение фундаментальных характеристик материалов. Это создает предпосылки для встраивания в предметно-ориентированные системы научной осведомленности элементов прикладного искусственного интеллекта, например экспертных систем для производства новых знаний и их сохранения в системе.

4. Компонент распространения профессиональных знаний (дистанционного обучения). Наличие такого компонента в системе делает ее более привлекательной для использования и распространения предметно-ориентированных знаний, а также служит привлечению заинтересованного круга профессиональных пользователей к производству новых знаний.

Предметно-ориентированные системы научной осведомленности могут быть созданы в научных проектах меньшего масштаба. При хорошо организованной и спроектированной системе метаданных они легко могут быть интегрированы в более крупные системы научной осведомленности с учетом территориальной распределенности последних. Такие системы могут стать основой для разработки методов получения, анализа и обработки экспертной информации в научной деятельности. Для разработки и создания предметно-ориентированных систем научной осведомленности целесообразно использовать технологии мультиагентных систем. Компонентом такой системы становится интеллектуальный агент, который можно представить в виде веб-приложения, наделенного искусственным интеллектом и расположенного за некоторым внешним порталом. При этом сам агент ориентирован на обработку научных данных в узкоспециализированном разделе предметной области. При наличии протокола взаимодействия между такими агентами система научной осведомленности в целом строится поэтапно.

13.3. Когнитивный анализ и моделирование сложных ситуаций

Сложности анализа процессов и принятия управленческих решений в таких областях как экономика, социология, экология и т. п. обусловлены рядом особенностей, присущих этим областям, а именно:

многоаспектностью происходящих в них процессов (экономических, социальных и т. п.) и их взаимосвязанностью; в силу этого невозможно вычленение и детальное исследование отдельных явлений – все происходящие в них явления должны рассматриваться в совокупности;

отсутствием достаточной количественной информации о динамике процессов, что вынуждает переходить к качественному анализу таких процессов;

изменчивостью характера процессов во времени и т. д. [54, 56].

В силу указанных особенностей экономические, социальные и тому подобные системы называются слабоструктурированными. Под *текущей ситуацией* понимается состояние слабоструктурированной системы в рассматриваемый момент времени. Число факторов в ситуации может измеряться десятками. И все они вплетены в паутину меняющихся во времени причин и следствий.

Увидеть и осознать логику развития событий на таком многофак-

торном поле крайне трудно. А ведь постоянно приходится отвечать (нередко – незамедлительно) на вопросы типа: «Что нужно сделать (на какие факторы повлиять), чтобы улучшить состояние ситуации?», «Что будет с ситуацией через такое-то время, если ничего не предпринимать?», «Какие из предпринимаемых мероприятий будут эффективнее в плане достижения поставленной цели?» и пр.

На такие вопросы можно успешно ответить, если использовать компьютерные средства познавательного (когнитивного) моделирования ситуаций. Подобные средства в экономически развитых странах применяются уже десятки лет, помогая предприятиям выжить и развить бизнес, а властям готовить хорошие нормативные документы [44].

Специфика применения средств когнитивного моделирования – в их ориентированности на конкретные условия развития ситуации в той или иной стране, регионе, городе, городке, поселке (политическая и экономическая устойчивость, ментальность населения и власти, хаотичность информационной сферы, открытость рынка, полнота нормативной базы и пр.).

Исходным понятием в когнитивном моделировании сложных ситуаций является понятие когнитивной карты ситуации. При анализе конкретной ситуации пользователь обычно знает или предполагает, какие изменения базисных факторов являются для него желательными. Факторы, представляющие наибольший интерес для пользователя, назовем целевыми. Это выходные факторы когнитивной карты. Когнитивная карта некоторой экономической ситуации представлена на рис. 13.3.

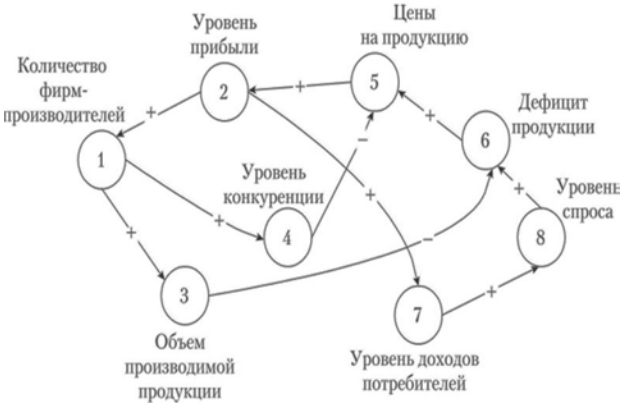

Рис. 13.3. Когнитивная карта некоторой экономической ситуации

Когнитивная карта экономической ситуации представляет собой ориентированный взвешенный граф, в котором:

– вершины взаимно однозначно соответствуют базисным факторам ситуации, в терминах которых описываются процессы в ситуации. Множество первоначально отобранных базисных факторов может быть верифицировано с помощью технологии Data Mining, позволяющей отбросить «избыточные» факторы, «слабо связанные» с «ядром» базисных факторов;

– определяются непосредственные взаимосвязи между факторами путем рассмотрения причинно-следственных цепочек, описывающих распространение влияний от каждого фактора на другие факторы. Считается, что факторы, входящие в посылку «если...» цепочки «если..., то...», влияют на факторы следствия «то...» этой цепочки, причем это влияние может быть либо усиливающим (положительным), либо тормозящим (отрицательным), либо переменного знака в зависимости от возможных дополнительных условий.

Когнитивная карта отображает лишь факт наличия влияний факторов друг на друга. В ней не отражаются ни детальный характер этих влияний, ни динамика изменения влияний в зависимости от изменения ситуации, ни временные изменения самих факторов. Учет всех этих обстоятельств требует перехода на следующий уровень структуризации информации, отображенной в когнитивной карте, т. е. к когнитивной модели. На этом уровне каждая связь между факторами когнитивной карты раскрывается до соответствующего уравнения, которое может содержать как количественные (измеряемые) переменные, так и качественные (неизмеряемые) переменные. При этом количественные переменные входят естественным образом в виде их численных значений. Каждой же качественной переменной ставится в соответствие совокупность лингвистических переменных, отображающих различные состояния этой качественной переменной (например, покупательский спрос может быть слабым, умеренным, ажиотажным и т. п.), а каждой лингвистической переменной соответствует определенный числовой эквивалент на шкале [0,1]. По мере накопления знаний о процессах, происходящих в исследуемой ситуации, становится возможным более детально раскрывать характер связей между факторами. Здесь существенную помощь может оказать использование процедур Data Mining [50, 52].

Задача выработки решений по управлению процессами в ситуации состоит в том, чтобы обеспечить желательные изменения целевых

факторов, что и является целью управления. Цель считается корректно заданной, если желательные изменения одних целевых факторов не приводят к нежелательным изменениям других целевых факторов.

В исходном множестве базисных факторов выделяется совокупность так называемых управляющих факторов («входных» факторов когнитивной модели), через которые подаются управляющие воздействия в модель. Управляющее воздействие считается согласованным с целью, если оно не вызывает нежелательных изменений ни в каком из целевых факторов.

При корректно заданной цели управления и при наличии управляющих воздействий, согласованных с этой целью, решение задачи управления не вызывает особых трудностей (даже при нелинейной когнитивной модели ситуации со знаковыми постоянными влияниями факторов друг на друга). Нахождение условий для обеспечения целенаправленного поведения в ситуации является весьма непростой задачей, требующей специального рассмотрения.

Формально когнитивная модель ситуации, как и когнитивная карта, может быть представлена графом, однако каждая дуга в этом графе представляет уже некую функциональную зависимость между соответствующими базисными факторами, т. е. когнитивная модель ситуации представляется функциональным графом.

14. ПЕРСПЕКТИВНЫЕ МЕТОДОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

14.1. Понятие искусственного интеллекта.

14.2. Методы и средства искусственного интеллекта.

14.3. Условия достижения интеллектуальности (гипотеза Ньюэлла – Саймона, тест Тьюринга).

14.1. Понятие искусственного интеллекта

Интеллект (от лат. *intellectus* – ощущение, восприятие, разумение, понимание, понятие, рассудок), или ум, – качество психики, состоящее из способности приспосабливаться к новым ситуациям, способности к обучению и запоминанию на основе опыта, пониманию и применению абстрактных концепций и использованию своих знаний для управления окружающей средой. Интеллект – это общая способность к познанию и решению трудностей, которая объединяет все познавательные

способности человека: ощущение, восприятие, память, представление, мышление, воображение. В начале 1980-х гг. ученые в области теории вычислений Барр и Файгенбаум предложили следующее определение искусственного интеллекта (ИИ):

«Искусственный интеллект – это область информатики, которая занимается разработкой интеллектуальных компьютерных систем, т. е. систем, обладающих возможностями, которые мы традиционно связываем с человеческим разумом, – понимание языка, обучение, способность рассуждать, решать проблемы и т. д.».

Позже к ИИ стали относить ряд алгоритмов и программных систем, отличительным свойством которых является то, что они могут решать некоторые задачи так, как это делал бы размышляющий над их решением человек. Основные свойства ИИ – это понимание языка, обучение и способность мыслить и, что немаловажно, действовать.

ИИ – комплекс родственных технологий и процессов, развивающихся качественно и стремительно, например:

обработка текста на естественном языке;

машинное обучение;

экспертные системы;

виртуальные агенты (чат-боты и виртуальные помощники);

системы рекомендаций.

16 апреля 2019 г. стало известно, что подкомитет ISO/IEC по стандартизации в области искусственного интеллекта поддержал предложение Технического комитета «Кибер-физические системы», созданного на базе РВК, о разработке стандарта «Artificial intelligence. Concepts and terminology» на русском языке в дополнение к базовой английской версии.

Терминологический стандарт «Artificial intelligence. Concepts and terminology» является основополагающим для всего семейства международных нормативно-технических документов в области искусственного интеллекта. Кроме терминов и определений, данный документ содержит концептуальные подходы и принципы построения систем с элементами AI, описание взаимосвязи AI с другими сквозными технологиями, а также базовые принципы и рамочные подходы к нормативно-техническому регулированию искусственного интеллекта. По итогам заседания профильного подкомитета ISO/IEC в Дублине эксперты ISO/IEC поддержали предложение делегации из России о синхронной разработке терминологического стандарта в сфере AI не только на английском, но и на русском языке. Документ был утвержден в 2021 г.

В ходе заседания эксперты ISO/IEC также поддержали разработку проекта международного документа Information Technology – Artificial Intelligence (AI) – Overview of Computational Approaches for AI Systems, в котором Россия выступает в качестве соредактора. Документ предоставляет обзор современного состояния систем искусственного интеллекта, описывая основные характеристики систем, алгоритмы и подходы, а также примеры специализированных приложений в области AI. Разработкой этого проекта документа займется специально созданная в рамках подкомитета рабочая группа 5 «Вычислительные подходы и вычислительные характеристики систем Искусственного интеллекта» (SC 42 Working Group 5 «Computational approaches and computational characteristics of AI systems»). Искусственный интеллект в широком смысле делят на четыре области (рис. 14.1).

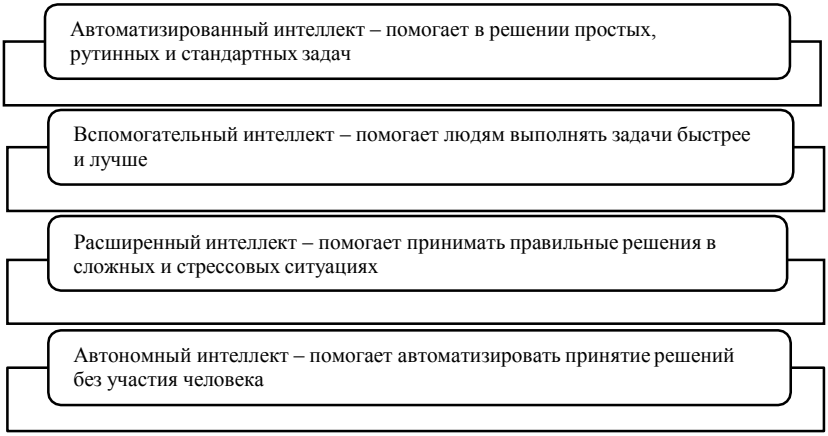

Рис. 14.1. Области искусственного интеллекта

14.2. Методы и средства искусственного интеллекта

Основные методы (подходы) к использованию средств искусственного интеллекта представлены на рис. 14.2.

Логический подход. Основой для логического подхода служат булева алгебра и ее логические операторы (в первую очередь знакомый всем оператор IF «ЕСЛИ»). Свое дальнейшее развитие булева алгебра получила в виде исчисления предикатов, в котором она расширена за счет введения предметных символов, отношений между ними, кванто-

ров существования и всеобщности. Практически каждая система ИИ, построенная на логическом принципе, представляет собой машину доказательства теорем. При этом исходные данные хранятся в базе данных в виде аксиом, а правила логического вывода – как отношения между ними.

Рис. 14.2. Основные методы и средства искусственного интеллекта

Для большинства логических методов характерна большая трудоемкость, поскольку во время поиска доказательства возможен полный перебор вариантов. Поэтому данный подход требует эффективной реализации вычислительного процесса, и хорошая работа обычно гарантируется при сравнительно небольшом размере базы данных.

Методы самоорганизации и эволюционный подход. Можно отметить следующие принципы самоорганизации математических моделей:

принцип ***неокончательных решений*** (предложен Д. Габором (1972) и заключается в необходимости сохранения достаточной свободы выбора нескольких лучших решений на каждом шаге самоорганизации);

принцип ***внешнего дополнения*** (базируется на теореме К. Геделя (Нагель, Ньюмен, 1970) и заключается в том, что только внешние критерии, основанные на новой информации, позволяют синтезировать истинную модель объекта, скрытую в зашумленных экспериментальных данных);

принцип ***массовой селекции*** (предложен А. Г. Ивахненко и указывает наиболее целесообразный путь постепенного усложнения самоорганизующейся модели, с тем чтобы критерий ее качества проходил через свой минимум).

Для возникновения самоорганизации необходимо иметь исходную структуру, механизм случайных ее мутаций и критерии отбора, благодаря которому мутация оценивается с точки зрения полезности для улучшения качества системы. То есть при построении этих систем ИИ исследователь задает только исходную организацию и список переменных, а также критерии качества, формализующие цель оптимизации, и правила, по которым модель может изменяться (самоорганизовываться или эволюционировать). Причем сама модель может принадлежать самым различным типам: линейная или нелинейная регрессия, набор логических правил или любая другая модель.

Можно выделить следующие подклассы самоорганизующихся моделей:

модели, реализующие полиномиальные алгоритмы, обобщением которых явился *метод группового учета аргументов* (*МГУА*);

модели, основанные на вероятностных методах самоорганизации и грамматике конечных стохастических автоматов;

исследование структуры сложной системы и решение задач восстановления уравнений (физических законов), описывающих разомкнутый объект по небольшому количеству экспериментальных точек.

Принцип массовой селекции, используемый в алгоритмах МГУА, как и многие другие идеи кибернетики, заимствует действующие природные механизмы и схематически повторяет агротехнические методы селекции растений или животных, например:

высевается некоторое количество семян и в результате опыления образуются сложные наследственные комбинации;

селекционеры выбирают некоторую часть растений, у которых интересующее их свойство выражено больше всего (эвристический критерий);

семена этих растений собирают и снова высевают для образования новых, еще более сложных комбинаций;

через несколько поколений селекция останавливается, и ее результат является оптимальным;

если чрезмерно продолжать селекцию, то наступит «инцухт» – вырождение растений (т. е. существует оптимальное число поколений и оптимальное количество семян, отбираемых в каждом из них).

Эволюционное моделирование (Фогель с соавт., 1969; Букатова, 1979; Букатова с соавт., 1991) представляет собой существенно универсальный способ построения прогнозов макросостояний системы в условиях, когда полностью отсутствует апостериорная информация, а априорные данные задают лишь предысторию этих состояний. Общая схема алгоритма эволюции выглядит следующим образом (рис. 14.3).

Рис. 14.3. Алгоритм эволюции

Таким образом, из рис. 14.3 следует, что процесс эволюционного моделирования протекает в несколько этапов:

задается исходная организация системы (в эволюционном моделировании в этом качестве может фигурировать, например, конечный детерминированный автомат Мили (Растригин, Марков, 1976; Букатова, 1979);

проводят случайные «мутации», т. е. изменяют случайным образом текущий конечный автомат;

отбирают для дальнейшего «развития» ту организацию (тот автомат), которая является «лучшей» в смысле некоторого критерия, например максимальной точности предсказания последовательности значений макросостояний экосистемы.

Критерий качества модели в этом случае мало чем отличается, например, от минимума среднеквадратической ошибки на обучающей последовательности метода наименьших квадратов (со всеми вытекающими отсюда недостатками). Однако, в отличие от адаптации, в эволюционном программировании структура решающего устройства мало меняется при переходе от одной мутации к другой, т. е. не происходит перераспределения вероятностей, которые бы закрепляли мутации, приведшие к успеху на предыдущем шаге. Поиск оптимальной структуры происходит в большей степени случайным и нецеленаправленным, что затягивает процесс поиска, но обеспечивает наилучшее приспособление к конкретным изменяющимся условиям.

Структурный подход и нейросетевое моделирование. Под *структурным подходом* подразумеваются попытки построения систем ИИ путем моделирования структуры человеческого мозга. В последние десять лет впечатляет феномен взрыва интереса к структурным методам самоорганизации – *нейросетевому моделированию*, которое успешно применяется в самых различных областях – бизнесе, медицине, технике, геологии, физике, т. е. везде, где нужно решать задачи прогнозирования, классификации или управления (Горбань, 1990, 1998; Уоссермен, 1992; Васильев с соавт., 1997). Описаны и широко распространяются нейросетевые расширения к популярным пакетам прикладных программ (Горбань, Россиев, 1996; Дьяконов, Круглов, 2001), что делает процесс проектирования интеллектуальных систем доступным любой домохозяйке с персональным компьютером [37].

Способность нейронной сети к обучению впервые была исследована Дж. Маккалоком и У. Питтом, когда в 1943 г. вышла их работа «Логическое исчисление идей, относящихся к нервной деятельности». В ней была представлена модель нейрона и сформулированы принципы построения искусственных нейронных сетей.

Крупный толчок к развитию нейрокибернетики дал американский нейрофизиолог Ф. Розенблатт, предложивший в 1962 г. свою модель нейронной сети – персептрон (Розенблатт, 1965; Минский, Пейперт, 1971). Воспринятый первоначально с большим энтузиазмом, персептрон вскоре подвергся интенсивным нападкам со стороны крупных научных авторитетов. И хотя подробный анализ их аргументов показывает, что они оспаривали не совсем тот персептрон, который предлагал Розенблатт, крупные исследования по нейронным сетям были свернуты почти на десять лет. И когда в журнале «Успехи физических наук» стали появляться статьи, связанные с фазовыми переходами в

нейронных системах, корректоры упорно исправляли в этих статьях слово «нейрон» на слово «нейтрон».

Значительную роль в общем подъеме интереса к нейропроблемам сыграла теория, предложенная Дж. Хопфилдом (Hopfield, 1982). Она буквально заворожила на продолжительное время физиков-теоретиков. И хотя, с точки зрения нейро-теоретиков и технологов, эта теория мало что дала, возбужденные ей аналогии и каскады головокружительных вычислений доставили немало эстетических радостей адептам науки. Более того, по аллитерации «нейрон»-«нейтрон» возникло модное в ту пору сочетание «нейронная бомба» и нейросетевые исследования стали финансироваться в рамках исследовательских программ всех родов войск США. Не исключено, что на вооружении каких-то стран уже имеются нейронные снаряды-камикадзе, чей нейросетевой «интеллект» направлен на уничтожение каких-то конкретных целей…

Другой важный класс нейронных систем был введен на рассмотрение финном Т. Кохоненом (1982). У этого класса красивое название: «самоорганизующиеся отображения состояний, сохраняющие топологию сенсорного пространства». Теория Кохонена активно использует *теорию адаптивных систем*, которую развивал на протяжении многих лет академик РАН Я. З. Цыпкин (1968, 1984).

Весьма популярна в настоящее время во всем мире оценка возможностей обучающихся систем, в частности нейронных сетей, основанная на *теории размерности*, созданной в 1966 г. советскими математиками В. Н. Вапником и А. Я. Червоненкисом (1974). Еще один класс нейроподобных моделей представляют сети с обратным распространением ошибок, в развитии современных модификаций которых ведущую роль сыграл профессор А. Н. Горбань и возглавляемая им красноярская школа нейроинформатики. Большую научную и популяризаторскую работу проводит Российская ассоциация нейроинформатики под руководством президента В. Л. Дунина-Барковского.

В основе всего нейросетевого подхода лежит идея построения вычислительного устройства из большого числа параллельно работающих простых элементов – формальных нейронов. Эти нейроны функционируют независимо друг от друга и связаны между собой однонаправленными каналами передачи информации. Ядром нейросетевых представлений является идея о том, что каждый отдельный нейрон можно моделировать довольно простыми функциями, а вся сложность мозга, гибкость его функционирования и другие важнейшие качества определяются связями между нейронами. Предельным выражением

этой точки зрения может служить лозунг: «*структура связей – все, свойства элементов – ничто*».

Нейронные сети (НС) – очень мощный метод моделирования, позволяющий воспроизводить чрезвычайно сложные зависимости, *нелинейные* по свой природе. Как правило, нейронная сеть используется тогда, когда неизвестны предположения о виде связей между входами и выходами (хотя, конечно, от пользователя требуется какой-то набор эвристических знаний о том, как следует отбирать и подготавливать данные, выбирать нужную архитектуру сети и интерпретировать результаты).

14.3. Условия достижения интеллектуальности (гипотеза Ньюэлла – Саймона, тест Тьюринга)

Несмотря на обилие методов искусственного интеллекта, часть которых была рассмотрена выше, теорией явно не определено, что именно считать необходимыми и достаточными условиями достижения интеллектуальности. На этот счет существует ряд гипотез, среди которых можно выделить следующие.

Гипотеза Ньюэлла – Саймона, формулировка которой выглядит следующим образом: физическая символическая система имеет необходимые и достаточные средства для того, чтобы производить осмысленные действия. Другими словами, без символических вычислений невозможно выполнять осмысленные действия, а способность выполнять символические вычисления вполне достаточна для того, чтобы быть способным выполнять осмысленные действия. Независимо от того, справедлива ли эта гипотеза, символические вычисления стали реальностью, и полезность этой парадигмы для программирования трудно отрицать.

Алленом Ньюэллом и Гербертом Саймоном было высказано предположение о том, что физические символьные системы (PSS) способны к интеллектуальному поведению, и что интеллектуальное поведение общего характера требует физической символьной системы. Согласно такому предположению, допускается наличие определенного типа связи между PSS и интеллектом (интеллектуальным продуктом), проявляющейся в виде регулярности и закономерности формирования определенных интеллектуальных процессов, явлений и событий.

Основанием для выдвижения такого предположения явились результаты работы компьютерной программы General Problem Solver

(GPS) (буквально – Общий решатель задач), полученные в ходе ее выполнения. Результаты работы GPS были квалифицированы как факты (эмпирические доводы, полученные в результате эксперимента, «интерсубъективно» подтверждающие проявление интеллекта), запротоколированы и положены в основу ряда подтверждающих наблюдений. В итоге предположение о том, что физическая символьная система (PSS) обладает необходимыми и достаточными средствами для организации общей интеллектуальной деятельности, приняло форму научной эмпирической гипотезы, которая известна как Гипотеза Ньюэлла – Саймона о физической символьной системе.

Ньюэлл и Саймон отметили, что не смогли найти никакого способа, чтобы продемонстрировать связь между символьными системами и интеллектом чисто логическим путем, поэтому гипотеза приобрела эмпирический характер, а не теоретический. Но, для теоретиков очевидно, что в действительности отсутствие способа теоретического доказательства обусловлено «феноменальностью» зависимой переменной, так как в философии и науке пока не решен вопрос о том, что именно считать интеллектом.

И, вероятно, именно этот факт стал главным препятствием для проведения теоретического исследования, которое проводится с использованием технологий и формальных методов теоретического исследования. Ньюэлл и Саймон должны были понимать, что в результате теоретического исследования ожидалось бы в первую очередь целостное представление (теория) о целой области интеллекта, а во вторую очередь – подтверждение предположения о взаимосвязи PSS и интеллекта (теорема), но не наоборот.

Но, нельзя не отметить объективную новизну и практическую значимость (PSSH-NS). Следует признать, что Ньюэллу и Саймону удалось найти то удивительное в известном, увидеть неочевидное и поставить вопрос, на который (уже более полувека) не могут дать ответ.

Также сложно не согласиться с тем, что PSSH-NS возникла как ответ на вызов окружающей реальности, которая постоянно ставит человека лицом к лицу с проблемами формирования (моделирования и имитации) искусственной интеллектуальной деятельности (автоматизации интеллектуальной деятельности).

Гипотеза Ньюэлла – Саймона о физической символьной системе (PSSH-NS) является научной эмпирической гипотезой (эмпирическим обобщением), а не теоремой. PSSH-NS прошла «классический путь развития эмпирической гипотезы» и, как любая другая эмпирическая

гипотеза (эмпирическое предположение), подлежит экспериментальной проверке.

Тест Тьюринга – мысленный эксперимент, предложенный в качестве критерия и конструктивного определения интеллектуальности. За почти 70 лет со времен первой публикации процедура прохождения претерпевала изменения, однако суть теста Тьюринга остается прежней.

Кратко ее можно выразить следующим образом: если, общаясь с человеком и машиной, экспериментатор не сможет определить, кто из них кто, значит, машиной тест пройден. Иными словами, идея теста заключается в том, что компьютер своими ответами должен убедить собеседника (он же судья) в своей человечности. По мнению Тьюринга, это свидетельствует о способности искусственного интеллекта мыслить и должно стать основанием для признания его разумности (рис. 14.4).

Рис. 14.4. Тест Тьюринга

Тест Алана Тьюринга является эмпирическим. Это значит, что он основан на опыте, наблюдениях, данных, полученных опытным путем. Идея данного теста возникла из салонной игры (игры для вечеринок того времени) – Imitation Game (Игра в имитацию). В ней участвовали как минимум три человека: женщина, мужчина и «судья» (любого пола). Мужчина и женщина уходили в разные комнаты и оттуда переда-

вали третьему игроку записочки. По ним нужно было определить, в какой комнате представитель какого пола находится. При этом они старались запутать «судью»: женщина могла выдавать себя за мужчину и наоборот.

Конечно, чтобы тест состоялся, судья не должен видеть собеседника, слышать его голос и т. д. В противном случае эксперимент явно будет провален, но это не будет связано с интеллектуальными возможностями машины. Как правило, формой общения выбирается электронная переписка. В изначальной версии теста человек общался с двумя субъектами – другим человеком и машиной. Чуть позже Тьюринг видоизменил прохождение – перед ИИ ставилась задача убедить в своей разумности ряд судей, которые, в свою очередь, общались с несколькими людьми и несколькими машинами. Это в том числе позволяет избежать субъективности в оценках и снизить риск простого угадывания.

Количество подопытных машин и людей в современных версиях теста разнится, как и время их общения. Судья может говорить со своими виртуальными собеседниками, о чем пожелает: вопросы теста Тьюринга не имеют ограничений. Для машины это представляет дополнительную сложность. Чтобы выполнить такое задание, компьютерная программа должна не просто понимать человеческий язык, но и давать естественные ответы по самым разным темам, отделяя важную информацию от несущественной для того или иного направления беседы.

Одним из главных недостатков теста видится то, что фактически перед машиной ставится задача запутать, обмануть человека. Говорит ли это о том, что мы может признать мыслящими и разумными только тех, кто умеет обманывать и манипулировать? Этот вопрос, скорее, лежит в области философии. Тем более что в теории прошедший тест Тьюринга робот должен хорошо имитировать, повторять действия человека, а не запутывать судью. На практике же с тестом лучше других справлялись «манипуляторы» – например, те, кто допускал опечатки в ответах. Машин даже специально этому обучали, чтобы их переписка выглядела естественнее. Еще одна распространенная уловка компьютера: умолчать о чем-либо, дать неполный ответ на вопрос или совсем сослаться на незнание. Иначе искусственный интеллект можно вычислить по тому, что он «слишком умный».

7 июня 2014 г. суперкомпьютер по имени Eugene попытался воссоздать интеллект тринадцатилетнего подростка – Евгения Густмана.

В тестировании, организованном Школой системной инженерии при Университете Рединга (Великобритания), участвовали пять суперкомпьютеров. Испытание представляло собой серию пятиминутных письменных диалогов. Разработчикам программы удалось подготовить бота ко всем возможным вопросам и даже обучить его собирать примеры диалогов через Twitter. Кроме того, инженеры наделили героя ярким характером. Притворяясь тринадцатилетним мальчиком, виртуальный «Евгений Густман» не вызывал сомнений у экспертов. Они поверили в то, что мальчик может не знать ответы на многие вопросы, ведь уровень знаний у среднего ребенка существенно ниже, чем у взрослых. При этом его правильные и точные ответы списывали на необычную эрудицию и начитанность. В тесте участвовали 25 «скрытых» людей и пять чат-ботов. Каждый из 30 судей провел по пять чат-сессий, пытаясь определить реальную природу собеседника. Для сравнения, в традиционном ежегодном конкурсе программ искусственного интеллекта на премию Лёбнера участвуют всего четыре программы и четыре скрытых человека.

ЗАКЛЮЧЕНИЕ

Современные информационные технологии позволяют эффективно собирать и накапливать большой объем разнородных агроэкономических данных, грамотный и всесторонний анализ которых является необходимым для проведения полноценного исследования. Стандартные методы математической и статистической обработки данных нередко не позволяют обнаружить существующие в данных нетривиальные и заранее непредсказуемые закономерности, для выявления которых разрабатываются и используются методы интеллектуального анализа данных. Таким образом, наличие обязательного курса по интеллектуальному анализу данных поможет развитию научно-исследовательских компетенций у обучающихся магистрантов. Для приобретения навыков в области анализа данных для не ИТ-специалистов рекомендуется сосредоточиться на демонстрации возможностей интеллектуального анализа данных с помощью существующих инструментов (в различных практиках использовались Weka, SASTM Enterprise Miner, надстройки Excel, Matlab и др.).

Интеллектуальный анализ данных является синтетической областью, цель которой часто ограничивается нахождением в автоматическом режиме закономерностей (моделей и отношений), скрытых в базе

(массиве) данных, которые не всегда могут быть найдены общеизвестными статистическими методами, хотя исторически сложилось так, что мощный арсенал методов прикладной статистики стал первым направлением развития средств интеллектуального анализа данных. Значимость статистических методов для интеллектуального анализа данных достаточно велика. Несмотря на то, что они разработаны в рамках статистической парадигмы, эти методы все же способны решать часть задач интеллектуального анализа данных, но они не позволяют генерировать новые гипотезы в автоматизированном режиме. Таким образом, формулировка гипотез остается прерогативой исследователя и производится «вручную», без помощи информационных систем.

Классическим примером применения статистических методов в процессе ИАД является проведение кластерного анализа, когда исследователь только задает классифицирующие признаки, но заранее не может предположить ни состав, ни объем кластеров. Таким образом, априорные предположения исследователя неполны, рабочая гипотеза не может быть четко сформулирована. Более того, при проведении кластерного анализа приходится перебирать «вручную» гипотезы о количестве кластеров.

Хотя прикладная статистика и входит в математический инструментарий интеллектуального анализа данных, но она является только частью этого инструментария. В основу большинства методов ИАД положена концепция шаблонов (паттернов) и зависимостей, отражающих многоаспектные взаимоотношения в данных. Поиск паттернов может производиться автоматическими методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей (что является обязательным в рамках статистической парадигмы).

Важной особенностью интеллектуального анализа данных является то, что он позволяет более полно использовать способности человека, освобождая его не только от рутинных вычислений, но даже от «рутинной» формулировки гипотез (естественно, при наличии «сильной» интеллектуальной системы, оснащенной «хорошим» математическим аппаратом, позволяющим реализовать методологию генерации и отбора наиболее интересных гипотез). Однако ИАД не решает задачи за аналитика, а всего лишь служит инструментом, который способствует поиску нетривиальных решений содержательных задач. Для того чтобы не профанировать методы интеллектуального анализа данных, их

следует понимать, знать достоинства и недостатки, границы применения каждого из них. Кроме того, ИАД предполагает совместное использование различных методов и алгоритмов при исследовании одного и того же социального феномена, реализуя таким способом принцип триангуляции в процессе анализа эмпирических данных. В данном контексте особое значение приобретают математические знания, навыки компьютерной реализации различных методов анализа данных и корректной интерпретации полученных результатов.

В интеллектуальном анализе данных обсуждают два основных подхода извлечения практически полезных знаний – *дедуктивный* (на основе некоторой априори сформулированной гипотезы, от общего к частному) и *индуктивный* (на основе известных *паттернов*, от частного к общему).

Дедуктивный подход к исследованию данных предполагает наличие некоторой сформулированной гипотезы, подтверждение или опровержение которой после анализа данных позволяет получить некоторые частные сведения.

Индуктивный подход к исследованию данных позволяет сформулировать (скорректировать существующую) гипотезу и найти с ее помощью новые пути аналитических решений.

Для поиска значимых закономерностей порой требуется совместное попеременное использование индуктивного и дедуктивного подходов, при котором формируется такая среда, в которой модели не нужно быть исключительно статической или эмпирической. Вместо этого, модель непрерывно тестируется, модифицируется и улучшается до тех пор, пока не будет достаточно усовершенствована.

ВОПРОСЫ ДЛЯ САМОПРОВЕРКИ

1. Какие тренды информационно-коммуникационных технологий способствовали развитию Data Mining?
2. Приведите примеры применения методов Data Mining для решения практических задач.
3. Какие области человеческой деятельности наиболее и наименее подходят для их анализа методами Data Mining?
4. Что понимается под Data Mining и Big Data? Почему возникла такая терминология?
5. В чем состоит суть индуктивных и дедуктивных подходов в Data Mining?

6. Каковы основные этапы интеллектуального анализа данных?

7. Какие классификации методов Data Mining существуют? Приведите примеры.

8. В чем заключается предварительная обработка данных и какова ее цель? Какие подходы при этом применяются?

9. В чем заключается оптимизация признакового пространства? Какие методы с трансформацией и без трансформации пространства применяют и в чем их отличия?

10. В чем заключается метод классификации? Какие подходы для его реализации могут быть использованы и в чем состоит их суть?

11. Что такое неконтролируемая классификация и какие методы применяют для ее реализации?

12. В чем заключается суть метода машины опорных векторов и в чем его преимущество перед аналогами?

13. Как работают деревья принятия решений? Какие их разновидности существуют? Каковы пределы применимости этого метода?

14. Что такое регрессия? Какие подходы применяют для ее реализации?

15. Как работают ассоциативные алгоритмы?

16. Как работают алгоритмы последовательной ассоциации?

17. Что такое обнаружение аномалий? Приведите примеры применения этого подхода и методы его реализации.

18. Что такое визуализация и какие инструменты ее реализации существуют?

19. Какие инструменты, модели и технологии существуют в настоящее время для реализации высокопроизводительных вычислений? Какие критерии эффективности при этом используют?

20. Приведите примеры коммерческих многофункциональных систем и свободно распространяемых решений, реализующих инструментарий Data Mining. Их сравнительные характеристики.

21. Архитектуры и особенности функционирования информационных систем, реализующих методы Data Mining как сервис.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айзек, М. П. Графика, формулы, анализ данных в Excel. Пошаговые примеры / М. П. Айзек, М. В. Финков. – Санкт-Петербург: Наука и техника, 2019. – 386 с.

2. Аксень, Э. М. Стохастическое моделирование макроэкономической динамики / Э. М. Аксень; Белорус. гос. экон. ун-т. – Минск, 2011. – 326 с.

3. Бенгфорт, Б. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка / Б. Бенгфорт. – Санкт-Петербург: Питер, 2019. – 368 с.

4. Бергер, А. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / А. Бергер. – Санкт-Петербург: BHV, 2007. – 928 с.

5. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA: учеб. пособие для вузов. + CD / В. П. Боровиков. – Москва: РиС, 2015. – 288 с.

6. Буць, В. И. Методология построения социально-экономических индексов управления ресурсосбережением / В. И. Буць. – Горки: БГСХА, 2009. – 168 с.

7. Винстон, У. Бизнес-моделирование и анализ данных. Решение актуальных задач с помощью Microsoft Excel / У. Винстон. – Санкт-Петербург: Питер, 2006. – 320 с.

8. Воскобойников, Ю. Е. Регрессионный анализ данных в пакете MATHCAD + CD / Ю. Е. Воскобойников. – Санкт-Петербург: Лань, 2011. – 224 с.

9. Горяинова, Е. Р. Прикладные методы анализа статистических данных: учеб. пособие / Е. Р. Горяинова, А. Р. Панков, Е. Н. Платонов. – Москва: ИД ГУ ВШЭ, 2012. – 310 с.

10. Дайитбегов, Д. М. Компьютерные технологии анализа данных в эконометрике: монография / Д. М. Дайитбегов. – Москва: Вузовский учебник, ИНФРА-М, 2013. – 587 с.

11. Есаулов, И. Г. Регрессионный анализ данных в пакете Mathcad: учеб. пособие / И. Г. Есаулов. – Санкт-Петербург: Лань П, 2016. – 224 с.

12. Ефремов, А. А. Использование оболочки данных для оценки сравнительной эффективности функционирования сельскохозяйственных организаций / А. А. Ефремов // Вестн. Могилев. гос. ун-та им. А. А. Кулешова. Сер.: В. Математика. Физика. Биология. – 2016. – № 49 (1). – С. 189–191.

13. Железко, Б. А. Теория и практика построения информационно-аналитических систем поддержки принятия решений: монография / Б. А. Железко, А. Н. Морозевич. – Минск: Армита-Маркетинг, Менеджмент, 1999. – 143 с.

14. Змитрович, А. И. Интеллектуальные информационные системы / А. И. Змитрович. – Минск: ТетраСистемс, 1997. – 368 с.

15. Информационные технологии и вычислительные системы: Обработка информации и анализ данных. Программная инженерия. Математическое моделирование. Прикладные аспекты информатики / под ред. С. В. Емельянова. – Москва: Ленанд, 2015. – 104 с.

16. Искусственный интеллект и принятие решений: Интеллектуальный анализ данных. Моделирование поведения. Когнитивное моделирование. Моделирование и управление / под ред. С. В. Емельянова. – Москва: Ленанд, 2012. – 108 с.

17. Кабаков, Р. R в действии. Анализ и визуализация данных в программе R / Р. Кабаков. – Москва: ДМК, 2016. – 588 с.

18. Калинина, В. Н. Анализ данных. Компьютерный практикум (для бакалавров) / В. Н. Калинина, В. И. Соловьев. – Москва: КноРус, 2017. – 240 с.

19. Кацко, И. А. Практикум по анализу данных на компьютере / И. А. Кацко, Н. Б. Паклин. – Москва: КолосС, 2009. – 278 с.

20. Козлов, А. Ю. Статистический анализ данных в MS Excel: учеб. пособие / А. Ю. Козлов, В. С. Мхитарян, В. Ф. Шишов. – Москва: ИНФРА-М, 2018. – 80 с.

21. Крянев, А. В. Метрический анализ и обработка данных / А. В. Крянев, Г. В. Лукин, Д. К. Удумян. – Москва: Физматлит, 2012. – 308 с.

22. Кулаичев, А. П. Методы и средства комплексного анализа данных: учеб. пособие / А. П. Кулаичев. – Москва: Форум, 2018. – 160 с.

23. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман. – Москва: ДМК, 2016. – 498 с.

24. Маккинли, У. Python и анализ данных / У. Маккинли. – Москва: ДМК, 2015. – 482 с.

25. Макшанов, А. В. Технологии интеллектуального анализа данных: учеб. пособие / А. В. Макшанов, А. Е. Журавлев. – Санкт-Петербург: Лань, 2018. – 212 с.

26. Малинецкий, Г. Г. Проблемы математической истории: Основания, информационные ресурсы, анализ данных / Г. Г. Малинецкий, А. В. Коротаев. – Москва: ООО «КД «Либроком», 2009. – 256 с.

27. Мамонтов, В. Г. Химический анализ почв и использование аналитических данных. Лабораторный практикум: учеб. пособие / В. Г. Мамонтов. – Санкт-Петербург: Лань, 2019. – 328 с.

28. Марманис, Х. Алгоритмы интеллектуального Интернета. Передовые методики сбора, анализа и обработки данных / Х. Марманис, Д. Бабенко. – Москва: Символ, 2011. – 480 с.

29. Марчук, Г. И. Геронтология in silico: становление новой дисциплины. Математические модели, анализ данных и вычислительные эксперименты / Г. И. Марчук. – Москва: БИНОМ. Лаборатория знаний, 2009. – 535 с.

30. Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью R (черно-белые графики) / С. Э. Мастицкий. – Москва: ДМК, 2015. – 496 с.

31. Математические модели социально-экономических процессов. Динамические системы. Управление рисками и безопасностью. Оптимизация, идентификация, теория игр. Обработка и анализ изображений и сигналов. Интеллектуальный анализ данных и распознавание образов // Труды ИСА РАН / под ред. С. В. Емельянова. – Москва: Красанд, 2013. – 128 с.

32. Миркин, Б. Г. Введение в анализ данных. Учебник и практикум / Б. Г. Миркин. – Люберцы: Юрайт, 2016. – 174 с.

33. Модели и методы интеллектуального анализа данных [Электронный ресурс]: учеб.-метод. пособие / сост. О. А. Попова. – Режим доступа: https://www.docme.ru/doc/1155653/ 882.modeli-i-metody-- intellektual._ nogo-analiza- dannyh---u... – Дата доступа: 06.04.2019.

34. Мусаев, А. А. Интеллектуальный анализ данных [Электронный ресурс]: учеб. пособие / А. А. Мусаев. – Режим доступа: http://sa.technolog.edu.ru/repository/ iad_iadl.pdf. – Дата доступа: 06.04.2019.

35. Нархид, Н. Apache Kafka. Потоковая обработка и анализ данных / Н. Нархид. – Санкт-Петербург: Питер, 2019. – 320 с.

36. Наследов, А. Д. IMB SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных / А. Д. Наследов. – Санкт-Петербург: Питер, 2013. – 416 с.

37. Наследов, А. Д. Математические методы психологического исследования. Анализи интерпретация данных: учеб. пособие / А. Д. Наследов. – Санкт-Петербург: Речь, 2012. – 392 с.

38. Ниворожкина, Л. И. Статистические методы анализа данных: учебник / Л. И. Ниворожкина, С. В. Арженовский, А. А. Рудяга. – Москва: Риор, 2018. – 320 с.

39. О развитии цифровой экономики [Электронный ресурс]: Декрет от 21 декабря 2017 г. № 8. – Режим доступа: http://president.gov.by/ru/. – Дата доступа: 06.04.2019.

40. Об утверждении стратегии Республики Беларусь в сфере интеллектуальной собственности на 2012–2020 годы [Электронный ресурс]: постановление Совета Министров Респ. Беларусь от 21 марта 2018 г. № 208 // Нац. правовой Интернет-портал Респ. Беларусь. – Режим доступа: http://www.pravo.by/document/. – Дата доступа: 06.04.2019.

41. Орлов, А. И. Организационно-экономическое моделирование: в 3 ч. / А. И. Орлов. – Москва: МГТУ им. Н. Э. Баумана, 2012. – Ч. 3: Статистические методы анализов данных. – 623 с.

42. Панкратова, Е. В. Анализ данных в программе SPSS для начинающих социологов / Е. В. Панкратова, И. Н. Смирнова, Н. Н. Мартынова. – Москва: Ленанд, 2018. – 200 с.

43. Петрунин, Ю. Ю. Информационные технологии анализа данных: учеб. пособие / Ю. Ю. Петрунин. – Москва: КДУ, 2010. – 292 с.

44. Пилипук, А. В. Механизм и модели конкурентного функционирования / А. В. Пилипук // Современная конкуренция. – 2016. – Т. 10, № 3 (57). – С. 119–142.

45. Рафалович, В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович. – Москва: SmartBook, 2018. – 352 с.

46. Резник, Г. А. Методы многомерного анализа статистических данных: учеб. пособие / Г. А. Резник. – Москва: Финансы и статистика, 2008. – 400 с.

47. Романко, В. К. Статистический анализ данных в психологии: учеб. пособие / В. К. Романко. – Москва: Изд-во «БИНОМ. Лаборатория знаний», 2013. – 312 с.

48. Сидняев, Н. И. Теория планирования эксперимента и анализ статистических данных: учеб. пособие для магистров / Н. И. Сидняев. – 2-е изд., перераб. и доп. – Люберцы: Юрайт, 2016. – 495 с.

49. Симчера, В. М. Методы многомерного анализа статистических данных / В. М. Симчера. – Москва: Финансы и статистика, 2008. – 400 с.

50. Сирота, А. А. Методы и алгоритмы анализа данных и их моделирование в MATLAB / А. А. Сирота. – Санкт-Петербург: BHV, 2016. – 384 с.

51. Скобцов, В. Ю. Интеллектуальный анализ данных: генетические алгоритмы [Электронный ресурс]: учеб.-метод. пособие / В. Ю. Скобцов, Н. В. Лапицкая, С. Н. Нестеренков. – Режим доступа: https://ru.b- ok.org/book/3637836/3bca77. – Дата доступа: 06.04.2019.

52. Тюрин, Ю. Н. Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Акаров. – Москва: МЦНМО, 2016. – 368 с.

53. Форман, Дж. Много цифр: Анализ больших данных при помощи Excel / Дж. Форман. – Москва: Альпина Паблишер, 2019. – 461 с.

54. Чашкин, Ю. Р. Математическая статистика. Анализ и обработка данных: учеб. пособие / Ю. Р. Чашкин; под ред. С. Н. Смоленского. – Ростов-на-Дону: Феникс, 2010. – 236 с.

55. Чесноков, С. В. Детерминационный анализ социально-экономических данных / С. В. Чесноков. – Минск: Книжный дом «Либроком», 2013. – 168 с.

56. Яцков, Н. Н. Интеллектуальный анализ данных [Электронный ресурс]: пособие / Н. Н. Яцков. – Режим доступа: http://elib.bsu.by/handle/ 123456789/114127?mode= full. – Дата доступа: 06.04.2019.

СОДЕРЖАНИЕ

Введение 3
1. Методологические основы технологий и моделей интеллектуального анализа данных 6
2. Процесс интеллектуального анализа данных 10
3. Основные технологии интеллектуального анализа данных 16
4. Статистические методы 20
5. Нейросетевые модели 24
6. Методы классификации: дерево решений 30
7. Кластерный анализ 34
8. Ассоциативные правила 45
9. Генетические модели 51
10. Нечеткая логика 57
11. Документальные информационно-поисковые системы 64
12. Системы, основанные на знаниях 70
13. Структурный анализ систем 78
14. Перспективные методологии искусственного интеллекта 86
Заключение 98
Вопросы для самопроверки 100
Библиографический список 102

Учебное издание

Буць Владимир Иванович

ТЕХНОЛОГИИ
ИНТЕЛЛЕКТУАЛЬНОГО
АНАЛИЗА ДАННЫХ

Курс лекций

Редактор *Е. В. Ширалиева*
Технический редактор *Н. Л. Якубовская*

Подписано в печать 30.06.2021. Формат $60\times84^{1}/_{16}$. Бумага офсетная.
Ризография. Гарнитура «Таймс». Усл. печ. л. 6,28. Уч.-изд. л. 5,74.
Тираж 35 экз. Заказ .

УО «Белорусская государственная сельскохозяйственная академия».
Свидетельство о ГРИИРПИ № 1/52 от 09.10.2013.
Ул. Мичурина, 13, 213407, г. Горки.

Отпечатано в УО «Белорусская государственная сельскохозяйственная академия».
Ул. Мичурина, 5, 213407, г. Горки.