Луньков А.Д., Харламов А.В.

Интеллектуальный анализ данных

Учебно-методическое пособие

Часть I

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Введение

Учебно-методическое пособие состоит из трех частей. Первая теоретическая часть курса – лекции по интеллектуальному анализу данных. Теоретическая часть базируется на электронной книге И.А.Чубуковой Data Mining. Вторая часть представляет задания и рекомендации по их выполнению на практических занятиях, в основе которых лежит использование аналитической платформы Dedactor Academic в целях иллюстрации работы методов интеллектуального анализа данных. Третья часть курса посвящена самостоятельной работе студентов, в основе которой лежит самостоятельная разработка и реализация алгоритмов некоторых инструментов Data Mining с использованием системы MatLab. Для построения модельных данных рекомендуется использовать табличный процессор Gnumeric. Все рекомендуемые программные средства находятся в свободном доступе.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО

Интеллектуальный анализ данных

Data Mining - мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др., см. рис. 1.

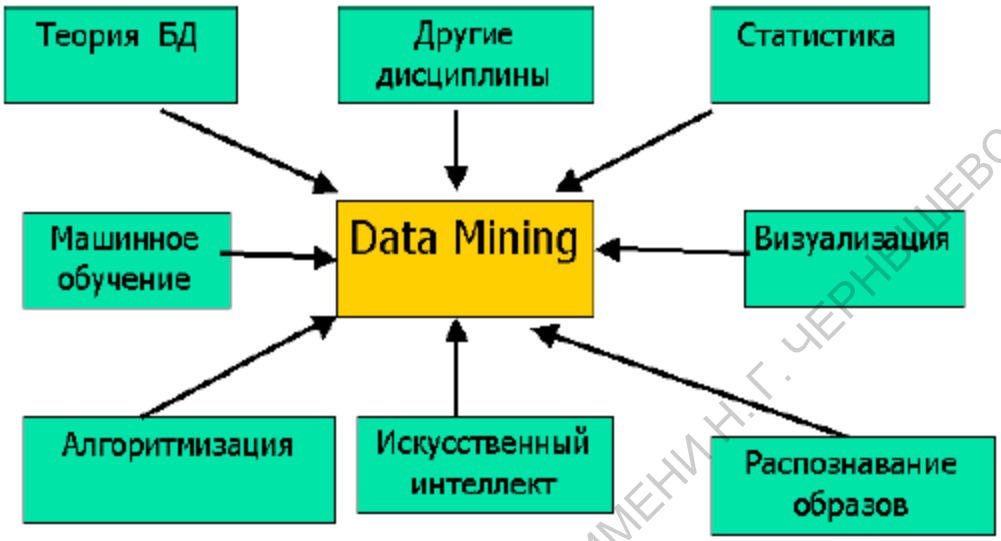

Рис.1 Data Mining как мультидисциплинарная область

Статистика - это наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению.

Статистика является совокупностью методов планирования эксперимента, сбора данных, их представления и обобщения, а также анализа и получения выводов на основании этих данных.

Статистика оперирует данными, полученными в результате наблюдений либо экспериментов.

Машинное обучение можно охарактеризовать как процесс получения программой новых знаний. Примером алгоритма машинного обучения являются нейронные сети.

Искусственный интеллект - научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования видов человеческой деятельности, традиционно считающихся интеллектуальными.

Искусственным интеллектом называют свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека.

Data Mining - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей.

Технология Data Mining - это процесс обнаружения в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Технология Data Mining предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Неочевидных - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных - это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных - это значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

В основу технологии Data Mining положена концепция шаблонов (patterns), которые представляют собой закономерности, свойственные подвыборкам данных, которые могут быть выражены в форме, понятной человеку.

Цель поиска закономерностей - представление данных в виде, отражающем искомые процессы. Построение моделей прогнозирования также является целью поиска закономерностей.

Средства Data Mining, в отличие от статистических, не требуют наличия строго определенного количества ретроспективных данных. Эта особенность может стать причиной обнаружения недостоверных, ложных моделей и, как результат, принятия на их основе неверных решений. Необходимо осуществлять контроль статистической значимости обнаруженных знаний.

Данные

В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты. Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.

Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки. Иными словами, данные - это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

Набор данных и их атрибутов

На рис.2 представлена двухмерная таблица, представляющая собой набор данных. По горизонтали таблицы располагаются атрибуты объекта или его признаки. По вертикали таблицы - объекты.

Объект описывается как набор атрибутов. Объект может быть представлен, как строка таблицы.

Атрибут - свойство, характеризующее объект. Атрибут также называют переменной, измерением, характеристикой. Он может быть представлен, как поле таблицы.

В результате перехода от общих категорий к конкретным величинам (операционализации понятий), получается набор переменных изучаемого понятия.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Переменная (variable) - свойство или характеристика, общая для всех изучаемых объектов, проявление которой может изменяться от объекта к объекту.

	Атрибуты				
Объекты	Код клиента	Возраст	Семейное положение	Доход	Класс
	1	18	Single	125	1
	2	22	Married	100	1
	3	30	Single	70	1
	4	32	Married	120	1
	5	24	Divorced	95	2
	6	25	Married	60	1
	7	32	Divorced	220	1
	8	19	Single	85	2
	9	22	Married	75	1
	10	40	Single	90	2

Рис.2 Двухмерная таблица «объект-атрибут»

Значение (value) переменной является проявлением признака.

При анализе данных, как правило, нет возможности рассмотреть всю совокупность объектов. Изучение очень больших объемов данных является дорогостоящим процессом, требующим больших временных затрат, а также неизбежно приводит к ошибкам, связанным с человеческим фактором.

Вполне достаточно рассмотреть некоторую часть всей совокупности, то есть выборку, и получить интересующую информацию на ее основании.

Однако размер выборки должен зависеть от разнообразия объектов, представленных в генеральной совокупности. В выборке должны быть представлены различные комбинации и элементы генеральной совокупности.

Генеральная совокупность (population) - вся совокупность изучаемых объектов, интересующая исследователя.

Выборка (sample) - часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Параметры - числовые характеристики генеральной совокупности.

Статистики - числовые характеристики выборки.

Часто исследования основываются на гипотезах. Гипотезы проверяются с помощью данных. Гипотеза это предположение относительно параметров совокупности объектов, которое должно быть проверено на ее части.

Гипотеза - частично обоснованная закономерность знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

Допустим, существует гипотеза, что **зависимая переменная** (продолжительность жизни) изменяется в **зависимости** от некоторых причин (качество питания, образ жизни, место проживания и т.д.), которые и являются **независимыми переменными.**

Однако переменная изначально не является зависимой или независимой. Она становится таковой после формулировки конкретной гипотезы. Зависимая переменная в одной гипотезе может быть независимой в другой.

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

В процессе подготовки данных измеряется не сам объект, а его характеристики. Измерение осуществляется в некоторой шкале.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

Переменные могут являться **числовыми** данными либо **символьными.**

Числовые данные, в свою очередь, могут быть дискретными и непрерывными.

Дискретные данные являются значениями признака, общее число которых не более чем счетно.

Непрерывные данные - данные, значения которых находятся в некотором интервале.

Шкалы

Обычно рассматривают следующие шкалы измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

Номинальная шкала (nominal scale) - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Для этой шкалы применимы только такие операции: равно\не равно.

Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

Шкала измерений дает возможность ранжировать значения переменных. Измерения же в порядковой шкале содержат информацию только о порядке следования величин, но не позволяют сказать "насколько одна величина больше другой", или "насколько она меньше другой".

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Для этой шкалы применимы только такие операции: равно\не равно, больше\меньше.

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Эта шкала позволяет находить разницу между двумя величинами, обладает свойствами номинальной и порядковой шкал, а также позволяет определить количественное изменение признака.

Для этой шкалы применимы только такие операции: равно\не равно, больше\меньше, сложения\вычитания.

Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.

Для этой шкалы применимы операции: равно\не равно, больше\меньше, сложения\вычитания), умножения\деления.

Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории, обычно это «да-нет» (частный случай номинальной шкалы).

Пример использования разных шкал для измерений свойств различных объектов, в данном случае температурных условий, приведен в таблице данных, изображенной на рис.3.

Номер объекта	Профессия (номинальная шкала)	Средний балл (интервальная шкала)	Образование (порядковая шкала)
1	слесарь	22	среднее
2	ученый	55	высшее
3	учитель	47	высшее

Рис.3 Множество измерений свойств различных объектов

Типы наборов данных

Данные, состоящие из записей

Наиболее часто встречающиеся данные - данные, состоящие из записей (record data). Это табличные данные, матричные данные, документальные данные, транзакционные или операционные.

Табличные данные состоят из записей, каждая из которых состоит из фиксированного набора атрибутов.

Транзакционные данные представляют собой особый тип данных, где каждая запись, являющаяся транзакцией, включает набор значений.

Пример транзакционной базы данных, содержащей перечень покупок клиентов магазина, приведен на рис.4.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Транза́кция — минимальная логически осмысленная операция, которая имеет смысл и может быть совершена только полностью.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Рис.4 Пример транзакционных данных

Графические данные.

Примеры графических данных: WWW-данные; молекулярные структуры; графы (рис.5); карты.

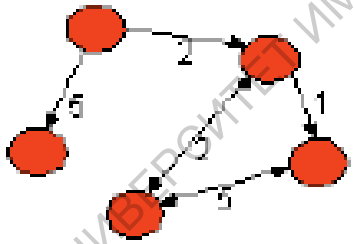

Рис.5 Пример графа

С помощью карт, например, можно отследить изменения объектов во времени и пространстве, определить характер их распределения на плоскости или в пространстве.

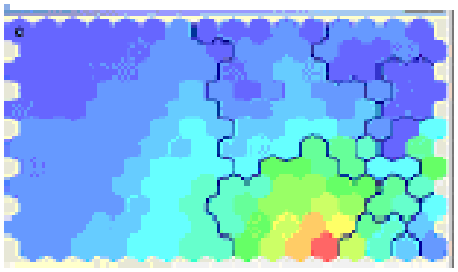

Рис.6 Пример данных типа «Карта Кохонена»

Преимуществом графического представления данных является большая простота их восприятия, чем, например, табличных данных. Пример карты Кохонена (модель нейронных сетей), представлен на рис.6.

Форматы хранения данных.

Одна из основных особенностей данных современного мира состоит в том, что их становится очень много.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Возможны четыре аспекта работы с данными: определение данных, вычисление (измерение), манипулирование (форматное преобразование) и обработка (сбор, передача и др.).

При манипулировании данными используется структура данных типа "файл". Файлы могут иметь различные форматы.

Наиболее распространенным форматом хранения данных для Data Mining выступают базы данных.

Базы данных (основные положения).

База данных (Database) - это особым образом организованные и хранимые в электронном виде данные.

Данные организованы неким конкретным способом, способным облегчить их поиск и доступ к ним для одного или нескольких приложений. Также такая организация данных предусматривает наличие минимальной избыточности данных.

Базы данных являются одной из разновидностей информационных технологий, а также формой хранения данных.

Целью создания баз данных является построение такой системы данных, которая бы не зависела от программного обеспечения, применяемых технических средств и физического расположения данных в ЭВМ. Построение такой системы данных должно обеспечивать непротиворечивую и целостную информацию. При проектировании базы данных предполагается многоцелевое ее использование. База данных в простейшем случае представляется в виде системы двумерных таблиц.

Схема данных - описание логической структуры данных, специфицированное на языке описания данных и обрабатываемое СУБД (системой управления базами данных).

Схема пользователя - зафиксированный для конкретного пользователя один вариант порядка полей таблицы.

Системы управления базами данных (СУБД).

Система управления базой данных - это программное обеспечение, контролирующее организацию, хранение, целостность, внесение изменений, чтение и безопасность информации в базе данных.

СУБД (Database Management System, DBMS) - представляет собой оболочку, с помощью которой при организации структуры таблиц и заполнения их данными получается та или иная база данных.

Система управления **реляционными базами данных** (Relational Database Management System) - это СУБД, основанная на реляционной модели данных.

В реляционной модели данных любое представление данных сводится к совокупности реляционных таблиц (двумерных таблиц особого типа). Системы управления реляционными базами данных используются для построения хранилищ данных (предметно-ориентированных информационных баз данных).

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

СУБД имеет программные, технические и организационные составляющие.

Программные средства включают систему управления, обеспечивающую ввод-вывод, обработку и хранение информации, создание, модификацию и тестирование базы данных.

Внутренними языками программирования СУБД являются языки четвертого поколения (C, C++, Pascal, Object Pascal). С помощью языков БД создаются приложения, базы данных и интерфейс пользователя, включающий экранные формы, меню, отчеты.

К базам данных, а также к СУБД предъявляются такие требования как:

1. высокое быстродействие;
2. простота обновления данных;
3. независимость данных;
4. возможность многопользовательского использования данных;
5. безопасность данных;
6. стандартизация построения и эксплуатации БД (фактически СУБД);
7. адекватность отображения данных соответствующей предметной области;
8. дружелюбный интерфейс пользователя.

СУБД отвечает за обработку запросов к базе данных и получение ответа. Способы хранения данных могут быть различными: модель данных может быть как реляционной, так и многомерной, сетевой или иерархической.

Классификация видов данных.

Реляционные данные - это данные из реляционных баз - таблиц (логическая модель данных).

Многомерные данные - это данные, представленные в кубах OLAP (аналитическая обработка в реальном времени).

Измерение (dimension) или ось - в многомерных данных - это собрание данных одного и того же типа, что позволяет структурировать многомерную базу данных.

По критерию постоянства своих значений в ходе решения поставленной задачи данные могут быть:

1) переменными - данные, которые изменяют свои значения в процессе решения задачи;
2) постоянными - данные, которые сохраняют свои значения в процессе решения задачи (математические константы, координаты неподвижных объектов) и не зависят от внешних факторов;
3) условно-постоянными - данные, которые могут иногда изменять свои значения, но эти изменения не зависят от процесса решения задачи, а определяются внешними факторами.

Данные, в зависимости от тех функций, которые они выполняют, могут быть: **справочными, оперативными, архивными.**

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Следует различать данные за период и точечные данные. Эти различия важны при проектировании системы сбора информации, а также в процессе измерений.

Данные за период характеризуют некоторый период времени.

Точечные данные представляют значение некоторой переменной в конкретный момент времени.

Данные бывают первичными и вторичными.

Вторичные данные - это данные, которые являются результатом определенных вычислений, примененных к **первичным данным**. Вторичные данные, как правило, приводят к ускоренному получению ответа на запрос пользователя за счет увеличения объема хранимой информации. **Первичные данные** получают непосредственным измерением (наблюдением).

Метаданные.

Метаданные (Metadata) - это данные о данных.

В состав метаданных могут входить: каталоги, справочники, реестры.

Метаданные содержат сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.

Метаданные, применяемые при управлении хранилищем, содержат информацию, необходимую для его настройки и использования.

Различают бизнес-метаданные и оперативные метаданные.

Бизнес-метаданные содержат бизнес-термины и определения, принадлежность данных и правила оплаты услуг хранилища.

Оперативные метаданные - это информация, собранная во время работы хранилища данных, она включает:

1) происхождение перенесенных и преобразованных данных;
2) статус использования данных (активные, архивированные или удаленные);
3) данные мониторинга, такие как статистика использования, сообщения об ошибках и т.д.

Метаданные хранилища обычно размещаются в репозитории (обычно данные в репозитории хранятся в виде файлов, доступных для дальнейшего распространения по сети). Это позволяет использовать метаданные совместно с различными инструментами и процессам при проектировании, установке, эксплуатации и администрировании хранилища.

Начальный этап процесса Data Mining

Процесс Data Mining состоит из определенных этапов, включающих элементы сравнения, типизации, классификации, обобщения, абстрагирования, повторения. Процесс Data Mining неразрывно связан с процессом принятия решений. Процесс Data Mining строит модель, а в процессе принятия решений эта модель эксплуатируется.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Традиционный процесс Data Mining включает следующие этапы:

- анализ предметной области;
- постановка задачи;
- подготовка данных;
- построение моделей;
- проверка и оценка моделей;
- выбор модели;
- применение модели;
- коррекция и обновление модели.

Этап 1. Анализ предметной области.

Исследование - это процесс познания определенной предметной области, объекта или явления с определенной целью.

Процесс исследования заключается в наблюдении свойств объектов с целью выявления и оценки важных, с точки зрения субъекта-исследователя, закономерных отношений между показателями данных свойств.

Предметная область - это мысленно ограниченная область реальной действительности, подлежащая описанию или моделированию и исследованию.

Предметная область состоит из объектов, различаемых по свойствам и находящихся в определенных отношениях между собой или взаимодействующих каким-либо образом.

Предметная область - это часть реального мира, она содержит как существенные, так и не значащие данные, с точки зрения проводимого исследования. Существенность данных зависит от выбора предметной области.

В процессе изучения предметной области должна быть создана ее модель. Знания из различных источников должны быть формализованы при помощи каких-либо средств. Это могут быть текстовые описания предметной области или специализированные графические нотации. Существует большое количество методик описания предметной области: например, методика структурного анализа SADT и основанная на нем IDEF0, диаграммы потоков данных Гейна-Сарсона, методика объектно-ориентированного анализа UML и другие.

Модель предметной области описывает процессы, происходящие в предметной области, и данные, которые в этих процессах используются. От того, насколько верно смоделирована предметная область, зависит успех дальнейшей разработки приложения Data Mining.

Этап 2. Постановка задачи.

Постановка задачи Data Mining включает следующие шаги:

- формулировка задачи;
- формализация задачи.

Постановка задачи включает также описание статического и динамического поведения исследуемых объектов.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Описание статики подразумевает описание объектов и их свойств. При описании динамики описывается поведение объектов и те причины, которые влияют на их поведение. Динамика поведения объектов часто описывается вместе со статикой.

Иногда этапы анализа предметной области и постановки задачи объединяют в один этап.

Этап 3. Подготовка данных.

Цель этапа: разработка базы данных для Data Mining.

Подготовка данных является важнейшим этапом, от качества выполнения которого зависит возможность получения качественных результатов всего процесса Data Mining. Рассмотрим шаги этого этапа.

1. Определение и анализ требований к данным.

На этом шаге осуществляется так называемое моделирование данных, т.е. определение и анализ требований к данным, которые необходимы для осуществления Data Mining. При этом изучаются вопросы распределения данных (географическое, организационное, функциональное); вопросы доступа к данным, которые необходимы для анализа, необходимость во внешних и/или внутренних источниках данных; а также аналитические характеристики системы (измерения данных, основные виды выходных документов, последовательность преобразования информации и др.).

2. Сбор данных.

Наличие в организации хранилища данных делает анализ проще и эффективней, его использование, с точки зрения вложений, обходится дешевле, чем использование отдельных баз данных или витрин данных. Однако далеко не всегда имеются хранилища данных. В этом случае источником для исходных данных являются оперативные, справочные и архивные БД, т.е. данные из существующих информационных систем.

Также для Data Mining может потребоваться информация из информационных систем внешних источников, бумажных носителей, а также знания экспертов или результаты опросов.

На этом шаге осуществляется кодирование некоторых данных.

При определении необходимого количества данных следует учитывать, являются ли данные упорядоченными или нет.

Если данные упорядочены и имеют дело с временными рядами, желательно знать, включает ли такой набор данных сезонную/цикличную компоненту. В случае присутствия в наборе данных сезонной/цикличной компоненты, необходимо иметь данные как минимум за один сезон/цикл.

Если данные не упорядочены, то есть события из набора данных не связаны по времени, в ходе сбора данных следует соблюдать следующие правила.

- Недостаточное количество записей в наборе данных может стать причиной построения некорректной модели. С точки зрения статистики, точность модели увеличивается с увеличением количества исследуемых данных.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

- Возможно, некоторые данные являются устаревшими или описывают какую-то нетипичную ситуацию, и их нужно исключить из базы данных.
- Алгоритмы, используемые для построения моделей на сверхбольших базах данных, должны быть масштабируемыми.
- При использовании многих алгоритмов необходимо определенное (желательное) соотношение входных переменных и количества наблюдений. Количество записей (примеров) в наборе данных должно быть значительно больше количества факторов (переменных).
- Набор данных должен быть репрезентативным и представлять как можно больше возможных ситуаций. Пропорции представления различных примеров в наборе данных должны соответствовать реальной ситуации.

3. Предварительная обработка данных.

Анализировать можно как качественные, так и некачественные данные. Результат будет достигнут и в том, и в другом случае. Для обеспечения качественного анализа необходимо проведение предварительной обработки данных, которая является необходимым этапом процесса Data Mining.

Данные, полученные в результате сбора, должны соответствовать определенным критериям качества. Таким образом, можно выделить важный подэтап процесса Data Mining - **оценивание качества данных.**

Качество данных (Data quality) - это критерий, определяющий полноту, точность, своевременность и возможность интерпретации данных.

Данные могут быть высокого качества и низкого качества, последние - это так называемые грязные или "плохие" данные.

Данные высокого качества - это полные, точные, своевременные данные, которые поддаются интерпретации. Такие данные обеспечивают получение качественного результата: знаний, которые смогут поддерживать процесс принятия решений.

Данные низкого качества, или грязные данные - это отсутствующие, неточные или бесполезные данные с точки зрения практического применения (например, представленные в неверном формате, не соответствующем стандарту).

Наиболее распространенные виды грязных данных:

- пропущенные значения;
- дубликаты данных;
- шумы и выбросы.

Пропущенные значения (Missing Values).

Некоторые значения данных могут быть пропущены в связи с тем, что:

- данные вообще не были собраны (например, при анкетировании скрыт возраст);

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

- некоторые атрибуты могут быть неприменимы для некоторых объектов (например, атрибут "годовой доход" неприменим к ребенку).

Обработка пропущенных данных.

1. Исключить объекты с пропущенными значениями из анализа.
2. Рассчитать новые значения для пропущенных данных.
3. Игнорировать пропущенные значения в процессе анализа.
4. Заменить пропущенные значения на возможные значения.

Дублирование данных (Duplicate Data).

Набор данных может включать продублированные данные, т.е. дубликаты. **Дубликатами** называются записи с одинаковыми значениями всех атрибутов. Наличие дубликатов в наборе данных может являться способом повышения значимости некоторых записей. Такая необходимость иногда возникает для особого выделения определенных записей из набора данных. Однако в большинстве случаев, продублированные данные являются результатом ошибок при подготовке данных.

Существует два варианта обработки дубликатов. При первом варианте удаляется вся группа записей, содержащая дубликаты. Этот вариант используется в том случае, если наличие дубликатов вызывает недоверие к информации, полностью ее обесценивает. Второй вариант состоит в замене группы дубликатов на одну уникальную запись.

Шумы и выбросы.

Выбросы - резко отличающиеся объекты или наблюдения в наборе данных. Шумы и выбросы являются достаточно общей проблемой в анализе данных. Выбросы могут как представлять собой отдельные наблюдения, так и быть объединенными в некие группы. Задача аналитика - не только их обнаружить, но и оценить степень их влияния на результаты дальнейшего анализа. Если выбросы являются информативной частью анализируемого набора данных, используют робастные методы и процедуры.

Достаточно распространена практика проведения двухэтапного анализа - с выбросами и с их отсутствием - и сравнение полученных результатов.

Различные методы Data Mining имеют разную чувствительность к выбросам, этот факт необходимо учитывать при выборе метода анализа данных. Также некоторые инструменты Data Mining имеют встроенные процедуры очистки от шумов и выбросов.

Визуализация данных позволяет представить данные, в том числе и выбросы, в графическом виде.

Очевидно, что результаты Data Mining на основе грязных данных не могут считаться надежными и полезными, очевидно необходима очистка данных.

Очистка данных (data cleaning, data cleansing или scrubbing) занимается выявлением и удалением ошибок и несоответствий в данных с целью улучшения качества данных.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Специальные средства очистки обычно имеют дело с конкретными областями - в основном это имена и адреса - или же с исключением дубликатов. Преобразования обеспечиваются либо в форме библиотеки правил, либо пользователем в интерактивном режиме. Преобразования данных могут быть автоматически получены с помощью средств согласования схемы.

Метод очистки данных должен удовлетворять ряду критериев.

1. Метод должен выявлять и удалять все основные ошибки и несоответствия, как в отдельных источниках данных, так и при интеграции нескольких источников.

2. Метод должен поддерживаться определенными инструментами, чтобы сократить объемы ручной проверки и программирования, и быть гибким в плане работы с дополнительными источниками.

3. Очистка данных не должна производиться в отрыве от связанных со схемой преобразования данных, выполняемых на основе сложных метаданных.

4. Функции маппирования для очистки и других преобразований данных должны быть определены декларативным образом и подходить для использования в других источниках данных и в обработке запросов.

5. Инфраструктура технологического процесса должна особенно интенсивно поддерживаться для Хранилищ данных, обеспечивая эффективное и надежное выполнение всех этапов преобразования для множества источников и больших наборов данных.

Этапы очистки данных

В целом, очистка данных включает следующие этапы

Этап № 1. Анализ данных.

Подробный анализ данных необходим для выявления подлежащих удалению видов ошибок и несоответствий. Здесь можно использовать как ручную проверку данных или их шаблонов, так и специальные программы для получения метаданных о свойствах данных и определения проблем качества.

Этап № 2. Определение порядка и правил преобразования данных.

В зависимости от числа источников данных, степени их неоднородности и загрязненности, данные могут требовать достаточно обширного преобразования и очистки. Иногда для отображения источников общей модели данных используется трансляция схемы; для Хранилищ данных обычно используется реляционное представление. Первые шаги по очистке могут уточнить или изменить описание проблем отдельных источников данных, а также подготовить данные для интеграции. Дальнейшие шаги должны быть направлены на интеграцию схемы/данных и устранение проблем множественных элементов, например, дубликатов. Для Хранилищ в процессе работы по определению ETL (*Extract, Transform, Load*) должны быть определены методы контроля и поток данных, подлежащий преобразованию и очистке.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Преобразования данных, связанные со схемой, так же как и этапы очистки, должны, насколько возможно, определяться с помощью декларативного запроса и языка маппирования, обеспечивая, таким образом, автоматическую генерацию кода преобразования. К тому же, в процессе преобразования должна существовать возможность запуска написанного пользователем кода очистки и специальных средств.

Этапы преобразования могут требовать обратной связи с пользователем по тем элементам данных, для которых отсутствует встроенная логика очистки.

Этап № 3. Подтверждение.

На этом этапе определяется правильность и эффективность процесса и определений преобразования. Это осуществляется путем тестирования и оценивания, например, на примере или на копии данных источника, - чтобы выяснить, необходимо ли как-то улучшить эти определения. При анализе, проектировании и подтверждении может потребоваться множество итераций, например, в связи с тем, что некоторые ошибки становятся заметны только после проведения определенных преобразований.

Этап № 4. Преобразования.

На этом этапе осуществляется выполнение преобразований либо в процессе ETL для загрузки и обновления Хранилища данных, либо при ответе на запросы по множеству источников.

Этап № 5. Противоток очищенных данных.

После того как ошибки отдельного источника удалены, загрязненные данные в исходных источниках должны замениться на очищенные, для того чтобы улучшенные данные попали также в унаследованные приложения и в дальнейшем при извлечении не требовали дополнительной очистки. Для Хранилищ очищенные данные находятся в области хранения данных.

Такой процесс преобразования требует больших объемов метаданных (схем, характеристик данных уровня схемы, определений технологического процесса и др.). Для согласованности, гибкости и упрощения использования в других случаях, эти метаданные должны храниться в репозитории (хранилище) на основе СУБД. Для поддержки качества данных подробная информация о процессе преобразования должна записываться как в репозиторий, так и в трансформированные элементы данных, в особенности информация о полноте и свежести исходных данных и происхождения информации о первоисточнике трансформированных объектов и произведенных с ними изменениях.

Методы и стадии Data Mining

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические и неиерархические методы кластерного анализа, методы поиска ассоциативных правил; метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и проч.

Большинство аналитических методов, используемые в технологии Data Mining – это известные математические алгоритмы и методы. Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств.

Следует отметить, что большинство методов Data Mining были разработаны в рамках теории искусственного интеллекта.

Метод (method) представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Алгоритм (algorithm) - точное предписание относительно последовательности действий (шагов), преобразующих исходные данные в искомый результат.

Понятие алгоритма появилось задолго до создания электронных вычислительных машин. Сейчас алгоритмы являются основой для решения многих прикладных и теоретических задач в различных сферах человеческой деятельности, в большинстве - это задачи, решение которых предусмотрено с использованием компьютера.

Классификация стадий Data Mining

Data Mining может состоять из двух или трех стадий.

Стадия 1. Свободный поиск (Discovery)

На стадии свободного поиска осуществляется исследование набора данных с целью поиска скрытых закономерностей. Предварительные гипотезы относительно вида закономерностей здесь не определяются.

Закономерность (law) - существенная и постоянно повторяющаяся взаимосвязь, определяющая этапы и формы процесса становления, развития различных явлений или процессов.

Система Data Mining на этой стадии определяет шаблоны, для получения которых в системах OLAP, например, аналитику необходимо обдумывать и создавать множество запросов. Здесь же аналитик освобождается от такой работы - шаблоны ищет за него система. Особенно полезно применение данного подхода в сверхбольших базах данных, где уловить закономерность

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

путем создания запросов достаточно сложно, для этого требуется перепробовать множество разнообразных вариантов.

Свободный поиск представлен следующими действиями:

- выявление закономерностей условной логики (conditional logic);
- выявление закономерностей ассоциативной логики (associations and affinities);
- выявление трендов и колебаний (trends and variations).

В результате свободного поиска закономерностей система сформирует набор логических правил "если ..., то ...". При этом требуется задать целевую переменную.

Действия системы в рамках стадии свободного поиска, выполняются при помощи:

- индукции правил условной логики (задачи классификации и кластеризации, описание в компактной форме близких или схожих групп объектов);
- индукции правил ассоциативной логики (задачи ассоциации и последовательности и извлекаемая при их помощи информация);
- определения трендов и колебаний (исходный этап задачи прогнозирования).

На стадии свободного поиска также может осуществляться валидация закономерностей, т.е. проверка их достоверности на части данных, которые не принимали участие в формировании закономерностей. Такой прием разделения данных на обучающее и проверочное множество часто используется в методах нейронных сетей и деревьев решений.

Стадия 2. Прогностическое моделирование (Predictive Modeling).

На данном этапе используются результаты работы первой стадии. Здесь обнаруженные закономерности применяются непосредственно для прогнозирования.

Прогностическое моделирование включает такие действия:

- предсказание неизвестных значений (outcome prediction);
- прогнозирование развития процессов (forecasting).

В процессе прогностического моделирования решаются задачи классификации и прогнозирования.

При решении задачи классификации результаты работы первой стадии (индукции правил) используются для отнесения нового объекта, с определенной уверенностью (вероятностью), к одному из известных, предопределенных классов на основании известных значений.

При решении задачи прогнозирования результаты первой стадии (определение тренда или колебаний) используются для предсказания неизвестных (пропущенных или же последующих) значений целевой переменной (переменных).

Отметим, что свободный поиск (стадия 1) раскрывает общие закономерности. Он по своей природе индуктивен. Закономерности, полученные на этой стадии, формируются от частного к общему. В результате

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

получают некоторое общее знание о некотором классе объектов на основании исследования отдельных представителей этого класса.

Прогностическое моделирование (стадия 2), напротив, дедуктивно. Закономерности, полученные на этой стадии, формируются от общего к частному и единичному. Здесь получают новое знание о некотором объекте или же группе объектов на основании:

- знания класса, к которому принадлежат исследуемые объекты;
- знание общего правила, действующего в пределах данного класса объектов.

Следует отметить, что полученные закономерности, а точнее, их конструкции, могут быть прозрачными, т.е. допускающими толкование аналитика (рассмотренные выше правила), и непрозрачными, так называемыми "черными ящиками". Типичный пример последней конструкции - нейронная сеть.

Стадия 3. Анализ исключений (forensic analysis)

На третьей стадии Data Mining анализируются исключения или аномалии, выявленные в найденных закономерностях.

Для выявления отклонений (deviation detection) необходимо определить норму, которая рассчитывается на стадии свободного поиска.

Для отклонений от нормы, возможны два варианта трактовок. Первый из них - существует некоторое логическое объяснение отклонений, которое также может быть оформлено в виде правила. Второй вариант - это ошибки исходных данных. В этом случае стадия анализа исключений может быть использована в качестве очистки данных.

Классификация методов Data Mining

Все методы Data Mining подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после Data Mining либо они дистиллируются для последующего использования.

1. Непосредственное использование данных, или сохранение данных.

В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключений. Проблема этой группы методов - при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

2. Выявление и использование формализованных закономерностей, или дистилляция шаблонов.

При технологии дистилляции шаблонов один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

формальные конструкции, вид которых зависит от используемого метода Data Mining. Этот процесс выполняется на стадии свободного поиска, у первой же группы методов данная стадия в принципе отсутствует.

На стадиях прогностического моделирования и анализа исключений используются результаты стадии свободного поиска, они значительно компактнее самих баз данных.

Напомним, что конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками").

Методы этой группы: логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях.

Логические методы, или методы логической индукции, включают: нечеткие запросы и анализы; символьные правила; деревья решений; генетические алгоритмы.

Методы этой группы являются, пожалуй, наиболее интерпретируемыми - они оформляют найденные закономерности, в большинстве случаев, в достаточно прозрачном виде с точки зрения пользователя. Полученные правила могут включать непрерывные и дискретные переменные. Деревья решений могут быть легко преобразованы в наборы символьных правил путем генерации одного правила по пути от корня дерева до его терминальной вершины. Деревья решений и правила фактически являются разными способами решения одной задачи и отличаются лишь по своим возможностям. Кроме того, реализация правил осуществляется более медленными алгоритмами, чем индукция деревьев решений.

Методы кросс-табуляции: агенты, баесовские (доверительные) сети, кросс-табличная визуализация. Последний метод не совсем отвечает одному из свойств Data Mining - самостоятельному поиску закономерностей аналитической системой. Однако, предоставление информации в виде кросс-таблиц обеспечивает реализацию основной задачи Data Mining - поиск шаблонов, поэтому этот метод можно также считать одним из методов Data Mining.

Методы на основе уравнений. Методы этой группы выражают выявленные закономерности в виде математических выражений - уравнений. Они могут работать лишь с численными переменными, и переменные других типов должны быть закодированы соответствующим образом. Это несколько ограничивает применение методов данной группы, тем не менее они широко используются при решении различных задач, особенно задач прогнозирования. Основные методы данной группы: статистические методы и нейронные сети. Статистические методы наиболее часто применяются для решения задач прогнозирования. Существует множество методов статистического анализа данных, среди них, например, корреляционно-регрессионный анализ, корреляция рядов динамики, выявление тенденций динамических рядов, гармонический анализ.

Другая классификация разделяет все многообразие методов Data Mining на две группы: статистические и кибернетические методы. Эта схема

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

разделения основана на различных подходах к обучению математических моделей.

Статистические методы Data mining.

Эти методы представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
- выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и др.);
- динамические модели и прогноз на основе временных рядов.

Арсенал статистических методов Data Mining классифицирован на четыре группы методов:

1. Дескриптивный анализ и описание исходных данных.
2. Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы Data Mining.

Второе направление Data Mining - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие методы:

- искусственные нейронные сети (распознавание, кластеризация, прогноз);
- эволюционное программирование;
- генетические алгоритмы (оптимизация);
- ассоциативная память (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний.

Методы Data Mining также можно классифицировать по задачам Data Mining. В соответствии с такой классификацией выделяем две группы. Первая из них – это подразделение методов Data Mining на решающие **задачи сегментации** (т.е. задачи классификации и кластеризации) и **задачи прогнозирования**.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

В соответствии со второй классификацией по задачам методы Data Mining могут быть направлены на получение описательных и прогнозирующих результатов.

Описательные методы служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.

К методам, направленным на получение описательных результатов, относятся итеративные методы кластерного анализа, в том числе: алгоритм k-средних, k-медианы, иерархические методы кластерного анализа, самоорганизующиеся карты Кохонена, методы кросс-табличной визуализации, различные методы визуализации и другие.

Прогнозирующие методы используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие методы: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод опорных векторов и др.

Задачи Data Mining

В основу технологии Data Mining положена концепция шаблонов, представляющих собой закономерности. В результате обнаружения этих, скрытых от невооруженного глаза закономерностей решаются задачи Data Mining. Различным типам закономерностей, которые могут быть выражены в форме, понятной человеку, соответствуют определенные задачи Data Mining.

Задачи (tasks) Data Mining иногда называют закономерностями (regularity) или техниками (techniques).

Обычно выделяют следующие: классификация, кластеризация, прогнозирование, ассоциация, визуализация, анализ и обнаружение отклонений, оценивание, анализ связей, подведение итогов.

Классификация (Classification). Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу. Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks).

Кластеризация (Clustering). Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы. Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Ассоциация (Associations). В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Отличие ассоциации от двух предыдущих задач Data Mining поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно. Наиболее известный алгоритм решения задачи поиска ассоциативных правил – алгоритм Apriori.

Последовательность (Sequence), или последовательная ассоциация (sequential association). Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов (sequential pattern). Правило последовательности: после события X через определенное время произойдет событие Y.

Прогнозирование (Forecasting). В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Определение отклонений или выбросов (Deviation Detection), анализ отклонений или выбросов. Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

Оценивание (Estimation). Задача оценивания сводится к предсказанию непрерывных значений признака.

Анализ связей (Link Analysis) - задача нахождения зависимостей в наборе данных.

Визуализация (Visualization, Graph Mining). В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных. Пример методов визуализации - представление данных в 2-D и 3-D измерениях.

Подведение итогов (Summarization) - задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

Классификация задач Data Mining

Согласно классификации по стратегиям, задачи Data Mining подразделяются на следующие группы:

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

- обучение с учителем;
- обучение без учителя;
- другие.

Категория обучение с учителем представлена следующими задачами Data Mining: классификация, оценка, прогнозирование. Категория обучение без учителя представлена задачей кластеризации. В категорию «другие» входят задачи, не включенные в предыдущие две стратегии.

Задачи Data Mining, в зависимости от используемых моделей, могут быть дескриптивными и прогнозирующими.

В соответствии с этой классификацией, задачи Data Mining представлены группами описательных и прогнозирующих задач.

В результате решения **описательных** (descriptive) задач аналитик получает шаблоны, описывающие данные, которые поддаются интерпретации. Эти задачи описывают общую концепцию анализируемых данных, определяют информативные, итоговые, отличительные особенности данных. Концепция описательных задач подразумевает характеристику и сравнение наборов данных. Характеристика набора данных обеспечивает краткое и сжатое описание некоторого набора данных. Сравнение обеспечивает сравнительное описание двух или более наборов данных.

Прогнозирующие (predictive) основываются на анализе данных, создании модели, предсказании тенденций или свойств новых или неизвестных данных.

Достаточно близким к вышеупомянутой классификации является подразделение задач Data Mining на следующие: исследования и открытия, прогнозирования и классификации, объяснения и описания.

Автоматическое исследование и открытие (свободный поиск). Пример задачи: обнаружение новых сегментов рынка. Для решения данного класса задач используются методы кластерного анализа.

Прогнозирование и классификация. Пример задачи: предсказание роста объемов продаж на основе текущих значений. Методы: регрессия, нейронные сети, генетические алгоритмы, деревья решений.

Задачи **классификации** и **прогнозирования** составляют группу так называемого индуктивного моделирования, в результате которого обеспечивается изучение анализируемого объекта или системы. В процессе решения этих задач на основе набора данных разрабатывается общая модель или гипотеза.

Объяснение и описание. Пример задачи: характеристика клиентов по демографическим данным и историям покупок. Методы: деревья решения, системы правил, правила ассоциации, анализ связей.

В интерпретации обобщенной модели аналитик получает новое знание. Группировка объектов происходит на основе их сходства.

Главная ценность Data Mining - это практическая направленность данной технологии, путь от сырых данных к конкретному знанию, от постановки задачи к готовому приложению, при поддержке которого можно

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

принимать решения. Рассмотрим подробнее задачи классификации и кластеризации.

Задачи классификации и кластеризации

Задача классификации

Классификация является наиболее простой и одновременно наиболее часто решаемой задачей Data Mining. Ввиду распространенности задач классификации необходимо четкое понимания сути этого понятия.

Приведем несколько определений.

Классификация - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация - упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Классификация требует соблюдения следующих правил:

- в каждом акте деления необходимо применять только одно основание;
- деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- деление должно быть последовательным.

Различают:

- вспомогательную (искусственную) классификацию, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
- естественную классификацию, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений.

Последняя является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий классификация может быть:

- простой - деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "А и не А");

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

- сложной - применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Под классификацией будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

Классификация - это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы. Таким образом, для проведения классификации должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или объект (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).

Классификация относится к стратегии обучения с учителем (supervised learning), которое также именуют контролируемым или управляемым обучением.

Задачей классификации часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Например, можно предсказать, кто из клиентов фирмы является потенциальным покупателем определенного товара, а кто - нет, кто воспользуется услугой фирмы, а кто - нет, и т.д. Этот тип задач относится к задачам бинарной классификации, в них зависимая переменная может принимать только два значения (например, да или нет, 0 или 1).

Другой вариант классификации возникает, если зависимая переменная может принимать значения из некоторого множества предопределенных классов. Например, когда необходимо предсказать, какую марку автомобиля захочет купить клиент. В этих случаях рассматривается множество классов для зависимой переменной.

Классификация может быть одномерной (по одному признаку) и многомерной (по двум и более признакам).

Процесс классификации.

Цель процесса классификации состоит в том, чтобы построить модель, которая использует прогнозирующие атрибуты в качестве входных параметров и получает значение зависимого атрибута. Процесс классификации заключается в разбиении множества объектов на классы по определенному критерию.

Классификатором называется некая сущность, определяющая, какому из предопределенных классов принадлежит объект по вектору признаков.

Для проведения классификации с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат классификации. Таким описанием в нашем случае выступает база данных.

Каждый объект (запись базы данных) несет информацию о некотором свойстве объекта.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое.

Обучающее множество (training set) - множество, которое включает данные, использующиеся для обучения (конструирования) модели. Такое множество содержит входные и выходные (целевые) значения примеров. Выходные значения предназначены для обучения модели.

Тестовое (test set) множество также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки работоспособности модели.

Процесс классификации состоит из двух этапов: конструирования модели и ее использования.

1. Конструирование модели: описание множества предопределенных классов.
   - Каждый пример набора данных относится к одному предопределенному классу.
   - На этом этапе используется обучающее множество, на нем происходит конструирование модели.
   - Полученная модель представлена классификационными правилами, деревом решений или математической формулой.

2. Использование модели: классификация новых или неизвестных значений.
   - Оценка правильности (точности) модели.
     a) Известные значения из тестового примера сравниваются с результатами использования полученной модели.
     b) Уровень точности - процент правильно классифицированных примеров в тестовом множестве.
     c) Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.
   - Если точность модели допустима, возможно использование модели для классификации новых примеров, класс которых неизвестен.

Процесс классификации, а именно, конструирование модели и ее использование, представлен на рис.7 и рис.8.

Методы, применяемые для решения задач классификации.

Для классификации используются различные методы. Основные из них:

- классификация с помощью деревьев решений;
- байесовская (наивная) классификация;
- классификация при помощи искусственных нейронных сетей;
- классификация методом опорных векторов;
- статистические методы, в частности, линейная регрессия;
- классификация при помощи метода ближайшего соседа;
- классификация CBR-методом;
- классификация при помощи генетических алгоритмов.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

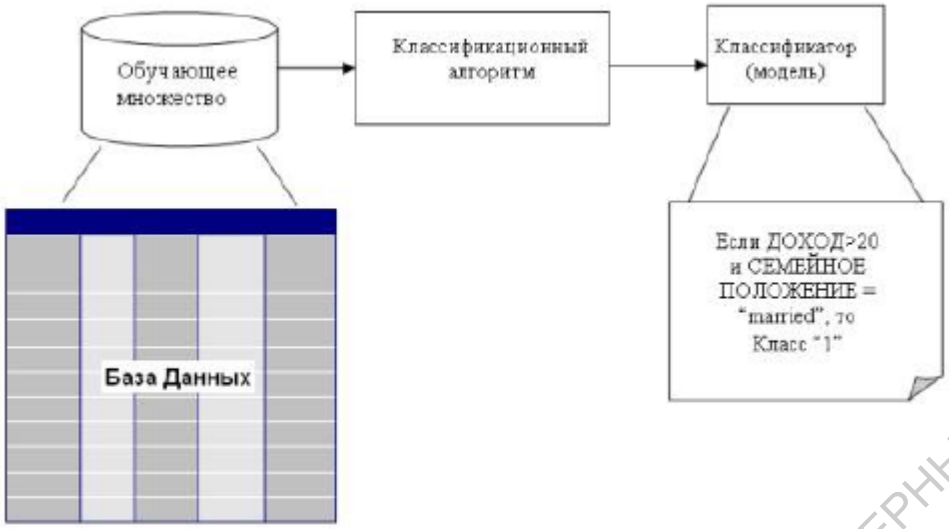

Рис.7 Процесс классификации. Конструирование модели

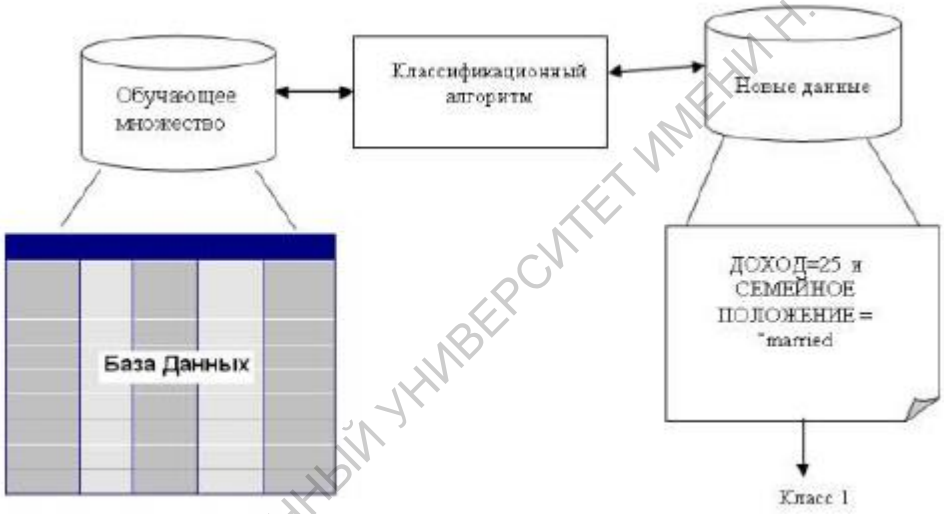

Рис.8. Процесс классификации. Использование модели

Точность классификации: оценка уровня ошибок.

Оценка точности классификации может проводиться при помощи кросс-проверки. **Кросс-проверка** (Cross-validation) - это процедура оценки точности классификации на данных из тестового множества, которое также называют кросс-проверочным множеством. Точность классификации тестового множества сравнивается с точностью классификации обучающего множества. Если классификация тестового множества дает приблизительно такие же результаты по точности, как и классификация обучающего множества, считается, что данная модель прошла кросс-проверку.

Разделение на обучающее и тестовое множества осуществляется путем деления выборки в определенной пропорции, например, обучающее множество - две трети данных и тестовое - одна треть данных. Этот способ следует использовать для выборок с большим количеством примеров. Если же выборка имеет малые объемы, рекомендуется применять специальные методы, при использовании которых обучающая и тестовая выборки могут частично пересекаться.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Задача кластеризации

Задача кластеризации сходна с задачей классификации и является ее логическим продолжением. Отличие состоит в том, что классы изучаемого набора данных заранее не предопределены.

Синонимами термина "кластеризация" являются "автоматическая классификация", "обучение без учителя" и "таксономия".

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

Цель кластеризации - поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Само понятие "кластер" определено неоднозначно: в каждом исследовании свои "кластеры". Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Вопрос, задаваемый аналитиками при решении многих задач, состоит в том, как организовать данные в наглядные структуры, т.е. развернуть таксономии. Наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. Для решения экономических задач кластеризация длительное время мало использовалась из-за специфики экономических данных и явлений.

На рис. 5.7 схематически представлены задачи классификации и кластеризации.

Кластеры могут быть непересекающимися, или эксклюзивными (non-overlapping, exclusive), и пересекающимися (overlapping).

Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры "цепочного" типа, когда кластеры представлены длинными "цепочками", кластеры удлиненной формы и т.д., а некоторые методы могут создавать кластеры произвольной формы.

Различные методы могут стремиться создавать кластеры определенных размеров (например, малых или крупных) либо предполагать в наборе данных наличие кластеров различного размера. Некоторые методы кластерного анализа особенно чувствительны к шумам или выбросам, другие - менее.

В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма. Данные особенности следует учитывать при выборе метода кластеризации.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

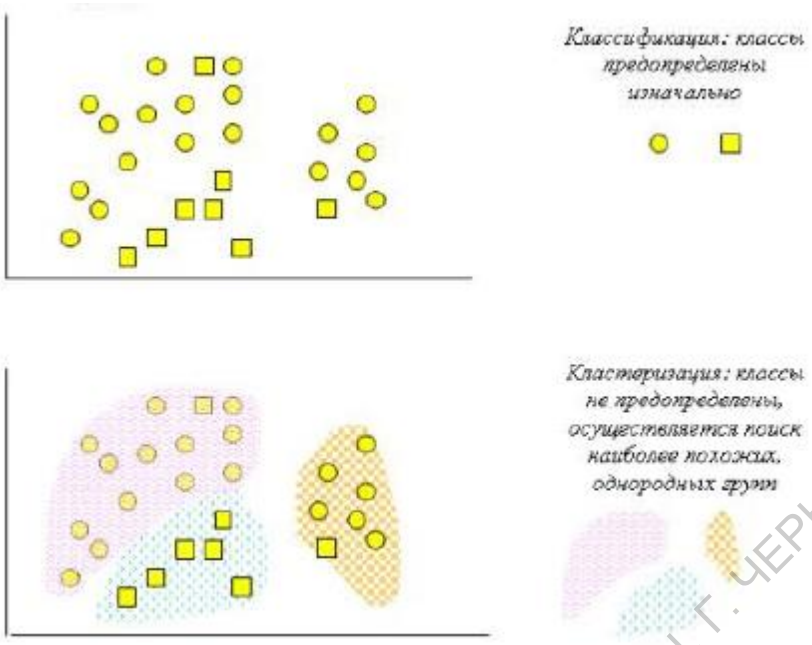

Рис.9 Сравнение задач классификации и кластеризации

На сегодняшний день разработано более сотни различных алгоритмов кластеризации.

Приведем краткую характеристику подходов к кластеризации.

- Алгоритмы, основанные на разделении данных (Partitioning algorithms), в т.ч. итеративные:
  - разделение объектов на k кластеров;
  - итеративное перераспределение объектов для улучшения кластеризации.
- Иерархические алгоритмы (Hierarchy algorithms):
  - агломерация: каждый объект первоначально является кластером, кластеры, соединяясь друг с другом, формируют больший кластер и т.д.
- Методы, основанные на концентрации объектов (Density-based methods):
  - основаны на возможности соединения объектов;
  - игнорируют шумы, нахождение кластеров произвольной формы.
- Грид-методы (Grid-based methods):
  - квантование объектов в грид-структуры.
- Модельные методы (Model-based):
  - использование модели для нахождения кластеров, наиболее соответствующих данным.

Оценка качества кластеризации.

Оценка качества кластеризации может быть проведена на основе следующих процедур:

- ручная проверка;

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

- установление контрольных точек и проверка на полученных кластерах;
- определение стабильности кластеризации путем добавления в модель новых переменных;
- создание и сравнение кластеров с использованием различных методов.

Разные методы кластеризации могут создавать разные кластеры, и это является нормальным явлением. Однако создание схожих кластеров различными методами указывает на правильность кластеризации.

Процесс кластеризации.

Процесс кластеризации зависит от выбранного метода и почти всегда является итеративным. Он может стать увлекательным процессом и включать множество экспериментов по выбору разнообразных параметров, например, меры расстояния, типа стандартизации переменных, количества кластеров и т.д. Полученные результаты требуют дальнейшей интерпретации, исследования и изучения свойств и характеристик объектов для возможности точного описания сформированных кластеров.

Задача прогнозирования

Задачи прогнозирования решаются в самых разнообразных областях человеческой деятельности, таких как наука, экономика, производство и множество других сфер. Прогнозирование является важным элементом организации управления как отдельными хозяйствующими субъектами, так и экономики в целом.

Развитие методов прогнозирования непосредственно связано с развитием информационных технологий, в частности, с ростом объемов хранимых данных и усложнением методов и алгоритмов прогнозирования, реализованных в инструментах Data Mining.

Прогнозирование (forecasting), в широком понимании этого слова, определяется как опережающее отражение будущего. Целью прогнозирования является предсказание будущих событий.

Прогностика (prognostics) - теория и практика прогнозирования. Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем. Таким образом, решение задачи прогнозирования требует некоторой обучающей выборки данных.

Прогнозирование - установление функциональной зависимости между зависимыми и независимыми переменными. Типичной в сфере маркетинга является задача прогнозирования рынков (market forecasting). В результате решения данной задачи оцениваются перспективы развития конъюнктуры определенного рынка, изменения рыночных условий на будущие периоды, определяются тенденции рынка (структурные изменения, потребности покупателей, изменения цен).

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

В самых общих чертах решение задачи прогнозирования сводится к решению таких подзадач:

- выбор модели прогнозирования;
- анализ адекватности и точности построенного прогноза.

Прогнозирование сходно с задачей классификации.

Многие методы Data Mining используются для решения задач классификации и прогнозирования: линейная регрессия, нейронные сети, деревья решений (которые иногда так и называют - деревья прогнозирования и классификации).

При решении обеих задач используется двухэтапный процесс построения модели на основе обучающего набора и ее использования для предсказания неизвестных значений зависимой переменной.

Различие задач классификации и прогнозирования состоит в том, что в первой задаче предсказывается класс зависимой переменной, а во второй - числовые значения зависимой переменной, пропущенные или неизвестные (относящиеся к будущему).

Прогнозирование и временные ряды

Основой для прогнозирования служит историческая информация, хранящаяся в базе данных в виде временных рядов.

Существует понятие Data Mining временных рядов (Time-Series Data Mining). **Временной ряд** - последовательность наблюдаемых значений какого-либо признака, упорядоченных в неслучайные моменты времени.

Типичный пример временного ряда - данные биржевых торгов.

Анализ временного ряда осуществляется с целью:

- определения природы ряда;
- прогнозирования будущих значений ряда.

В процессе определения структуры и закономерностей временного ряда предполагается обнаружение: шумов и выбросов, тренда, сезонной компоненты, циклической компоненты. Определение природы временного ряда может быть использовано как своеобразная "разведка" данных. Знание аналитика о наличии сезонной компоненты необходимо, например, для определения количества записей выборки, которое должно принимать участие в построении прогноза.

Шумы и выбросы усложняют анализ временного ряда. Существуют различные методы определения и фильтрации выбросов, дающие возможность исключить их с целью более качественного анализа.

Основными составляющими временного ряда являются тренд и сезонная компонента. Тренд является систематической компонентой временного ряда, которая может изменяться во времени.

Трендом называют неслучайную функцию, которая формируется под действием общих или долговременных тенденций, влияющих на временной ряд. Автоматического способа обнаружения трендов во временных рядах не существует. Но если временной ряд включает монотонный тренд (т.е.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

отмечено его устойчивое возрастание или устойчивое убывание), анализировать временной ряд в большинстве случаев нетрудно.

Существует большое разнообразие постановок задач прогнозирования, которое можно подразделить на две группы: прогнозирование односерийных рядов и прогнозирование мультисерийных, или взаимовлияющих, рядов.

Группа прогнозирование односерийных рядов включает задачи построения прогноза одной переменной по ретроспективным данным только этой переменной, без учета влияния других переменных и факторов.

Группа прогнозирования мультисерийных, или взаимовлияющих, рядов включает задачи анализа, где необходимо учитывать взаимовлияющие факторы на одну или несколько переменных.

Кроме деления на классы по односерийности и многосерийности, ряды также бывают сезонными и несезонными.

Последнее деление подразумевает наличие или отсутствие у временного ряда такой составляющей как сезонность, т.е. включение сезонной компоненты. **Сезонная составляющая** временного ряда является периодически повторяющейся компонентой временного ряда.

Свойство сезонности означает, что через примерно равные промежутки времени форма кривой, которая описывает поведение зависимой переменной, повторяет свои характерные очертания.

Ряд можно считать несезонным, если при рассмотрении его внешнего вида нельзя сделать предположений о повторяемости формы кривой через равные промежутки времени. Иногда по внешнему виду кривой ряда нельзя определить, является он сезонным или нет.

Существует понятие сезонного мультиряда. В нем каждый ряд описывает поведение факторов, которые влияют на зависимую (целевую) переменную.

При сборе данных и выборе факторов для решения задачи по прогнозированию в таких случаях следует учитывать, что влияние объемов продаж товаров друг на друга здесь намного меньше, чем воздействие фактора сезонности.

Очень часто тренд и сезонность присутствуют во временном ряде одновременно.

Принято различать циклическую компоненту и сезонную. Продолжительность цикла, как правило, больше, чем один сезонный период, циклы, в отличие от сезонных периодов, не имеют определенной продолжительности.

При выполнении каких-либо преобразований понять природу временного ряда значительно проще, такими преобразованиями могут быть, например, удаление тренда или сглаживание ряда.

Перед началом прогнозирования необходимо ответить на следующие вопросы:

1. Что нужно прогнозировать?

2. В каких временных элементах (параметрах)?

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

3. С какой точностью прогноза?

При ответе на первый вопрос, мы определяем переменные, которые будут прогнозироваться. Это может быть, например, уровень производства конкретного вида продукции в следующем квартале, прогноз суммы продажи этой продукции и т.д. При выборе переменных следует учитывать доступность ретроспективных данных, предпочтения лиц, принимающих решения, окончательную стоимость Data Mining. Часто при решении задач прогнозирования возникает необходимость предсказания не самой переменной, а изменений ее значений.

При решении второго вопроса задачи прогнозирования определяют следующие параметры:

- период прогнозирования;
- горизонт прогнозирования;
- интервала прогнозирования.

Период прогнозирования - основная единица времени, на которую делается прогноз.

Горизонт прогнозирования - это число периодов в будущем, которые покрывает прогноз.

Интервал прогнозирования - частота, с которой делается новый прогноз. Интервал прогнозирования может совпадать с периодом прогнозирования.

При выборе параметров необходимо учитывать, что горизонт прогнозирования должен быть не меньше, чем время, которое необходимо для реализации решения, принятого на основе этого прогноза. Только в этом случае прогнозирование будет иметь смысл.

С увеличением горизонта прогнозирования точность прогноза, как правило, снижается, а с уменьшением горизонта - повышается.

Можно улучшить качество прогнозирования, уменьшая время, необходимое на реализацию решения, для которого осуществляется прогноз, и, следовательно, уменьшив при этом горизонт и ошибку прогнозирования.

При выборе интервала прогнозирования следует выбирать между двумя рисками: вовремя не определить изменения в анализируемом процессе и высокой стоимостью прогноза. При длительном интервале прогнозирования возникает риск не идентифицировать изменения, произошедшие в процессе, при коротком - возрастают издержки на прогнозирование. При выборе интервала необходимо также учитывать стабильность анализируемого процесса и стоимость проведения прогноза.

Точность прогноза.

Точность прогноза, требуемая для решения конкретной задачи, оказывает большое влияние на прогнозирующую систему. Ошибка прогноза зависит от используемой системы прогноза. Чем больше ресурсов имеет такая система, тем больше шансов получить более точный прогноз. Однако прогнозирование не может полностью уничтожить риски при принятии решений. Поэтому всегда учитывается возможная ошибка прогнозирования.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Точность прогноза характеризуется ошибкой прогноза.

Наиболее распространенные виды ошибок:

1. Средняя ошибка (СО). Она вычисляется простым усреднением ошибок на каждом шаге. Недостаток этого вида ошибки - положительные и отрицательные ошибки аннулируют друг друга.
2. Средняя абсолютная ошибка (САО). Она рассчитывается как среднее абсолютных ошибок. Если она равна нулю, то мы имеем совершенный прогноз. В сравнении со средней квадратической ошибкой, эта мера "не придает слишком большого значения" выбросам.
3. Сумма квадратов ошибок (SSE), среднеквадратическая ошибка. Она вычисляется как сумма (или среднее) квадратов ошибок. Это наиболее часто используемая оценка точности прогноза.
4. Относительная ошибка (ОО). Предыдущие меры использовали действительные значения ошибок. Относительная ошибка выражает качество подгонки в терминах относительных ошибок.

Виды прогнозов.

Прогноз может быть краткосрочным, среднесрочным и долгосрочным.

Краткосрочный прогноз представляет собой прогноз на несколько шагов вперед, т.е. осуществляется построение прогноза не более чем на 3% от объема наблюдений или на 1-3 шага вперед.

Среднесрочный прогноз - это прогноз на 3-5% от объема наблюдений, но не более 7-12 шагов вперед; также под этим типом прогноза понимают прогноз на один или половину сезонного цикла. Для построения краткосрочных и среднесрочных прогнозов вполне подходят статистические методы.

Долгосрочный прогноз - это прогноз более чем на 5% от объема наблюдений. При построении данного типа прогнозов статистические методы практически не используются, кроме случаев очень "хороших" рядов, для которых прогноз можно просто "нарисовать".

Задача визуализации

Визуализация - это инструментарий, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения.

В результате использования визуализации создается графический образ данных. Применение визуализации помогает в процессе анализа данных увидеть аномалии, структуры, тренды. В задачах классификации и кластеризации, для иллюстрации распределения объектов в двухмерном пространстве также применяется визуализация.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Применение визуализации является более экономичным и простым способом выявления закономерностей (на первом этапе исследования). Линия тренда или скопления точек на диаграмме рассеивания позволяет аналитику намного быстрее определить закономерности и прийти к нужному решению. Таким образом, здесь идет речь об использовании в Data Mining не символов, а образов.

Визуализации данных может быть представлена в виде: графиков, схем, гистограмм, диаграмм и т.д.

Кратко роль визуализации можно описать такими ее возможностями:

- поддержка интерактивного и согласованного исследования;
- помощь в представлении результатов;
- наглядность, чтобы создавать зрительные образы и осмысливать их.

При этом необходимо иметь в виду, что неграмотная визуализация может приводит к неверным заключениям, которые обычно выявляются при дальнейшем анализе.

Основы анализа данных

На первом этапе статистического анализа данных, как правило, вычисляют описательную статистику, проводят расчет корреляций и строят линейную регрессию.

Описательная статистика (Descriptive statistics) - техника сбора и суммирования количественных данных, которая используется для превращения массы цифровых данных в форму, удобную для восприятия и обсуждения.

Цель описательной статистики - обобщить первичные результаты, полученные в результате наблюдений и экспериментов.

В состав описательной статистики входят такие характеристики: среднее; стандартная ошибка; медиана; мода; стандартное отклонение; дисперсия выборки; эксцесс; асимметричность; интервал; минимум; максимум; сумма; счет.

Центральная тенденция.

Измерение центральной тенденции заключается в выборе числа, которое наилучшим способом описывает все значения признака набора данных. Рассмотрим две характеристики этого измерения, а именно: среднее значение и медиану. Главная цель среднего - представление набора данных для последующего анализа, сопоставления и сравнения.

Значение среднего легко вычисляется и может быть использовано для последующего анализа. Оно может быть вычислено для данных, измеряемых по интервальной шкале, и для некоторых данных, измеряемых по порядковой шкале. **Среднее значение** рассчитывается как среднее арифметическое набора данных: сумма всех значений выборки, деленная на объем выборки. "Сжимая" данные таким образом, теряется много информации.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Среднее значение очень информативно и позволяет делать вывод относительно всего исследуемого набора данных.

Свойства среднего.

1. При расчете среднего не допускаются пропущенные значения данных.
2. Среднее может вычисляться только для числовых данных и для дихотомических шкал.
3. Для одного набора данных может быть рассчитано одно и только одно значение среднего.

Информативность среднего значения переменной высока, если известен ее доверительный интервал. **Доверительным интервалом** для среднего значения является интервал значений вокруг оценки, где с данным уровнем доверия находится "истинное" среднее популяции.

Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Ширина доверительного интервала зависит от размера выборки и от разброса данных.

С увеличением размера выборки точность оценки среднего возрастает. С увеличением разброса значений выборки надежность среднего падает. Если размер выборки достаточно большой, качество среднего увеличивается независимо от выполнения предположения нормальности выборки.

Медиана - точная середина выборки, которая делит ее на две равные части по числу наблюдений.

Обязательным условием нахождения медианы является упорядоченность выборки. Таким образом, для нечетного количества наблюдений медианой выступает наблюдение с номером (n+1)/2, где n - количество наблюдений в выборке. Для четного числа наблюдений медианой является среднее значение наблюдений n/2 и (n+2)/2.

Некоторые свойства медианы.

1. Для одного набора данных может быть рассчитано одно и только одно значение медианы.
2. Медиана может быть рассчитана для неполного набора данных, для этого необходимо знать номера наблюдений по порядку, общее количество наблюдений и несколько значений в середине набора данных.

Характеристики вариации данных.

Наиболее простыми характеристиками выборки являются максимум и минимум.

Минимум - наименьшее значение выборки.

Максимум - наибольшее значение выборки.

Размах - разница между наибольшим и наименьшим значениями выборки.

Дисперсия - среднее арифметическое квадратов отклонений значений от их среднего.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Стандартное отклонение - квадратный корень из дисперсии выборки - мера того, насколько широко разбросаны точки данных относительно их среднего.

Эксцесс показывает "остроту пика" распределения, характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение (пик заострен). Отрицательный эксцесс обозначает относительно сглаженное распределение (пик закруглен). Если эксцесс существенно отличается от нуля, то распределение имеет или более закругленный пик, чем нормальное, или, напротив, имеет более острый пик (возможно, имеется несколько пиков). Эксцесс нормального распределения равен нулю.

Асимметрия или асимметричность показывает отклонение распределения от симметричного. Если асимметрия существенно отличается от нуля, то распределение несимметрично, нормальное распределение абсолютно симметрично. Если распределение имеет длинный правый хвост, асимметрия положительна; если длинный левый хвост - отрицательна.

Выбросы (outliers) - данные, резко отличающиеся от основного числа данных. При обнаружении выбросов перед исследователем стоит дилемма: оставить наблюдения - выбросы либо от них отказаться. Второй вариант требует серьезной аргументации и описания. Полезным будет провести анализ данных с выбросами и без и сравнить результаты.

Следует помнить, что при применении классических методов статистического анализа, которые, как правило, не являются робастными (устойчивыми), наличие выбросов в наборе данных приводит к некорректным результатам. Если набор данных относительно мал, исключение данных, которые считаются выбросами, может заметно повлиять на результаты анализа.

Наличие выбросов в наборе данных может быть связано с появлением так называемых "сдвинутых" значений, связанных с систематической ошибкой, ошибок ввода, ошибок сбора данных и т.д. Иногда к выбросам могут относиться наименьшие и наибольшие значения набора данных.

Корреляционный анализ

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Корреляционный анализ дает возможность установить, ассоциированы ли наборы данных по величине. Для определения наличия взаимосвязи между двумя свойствами используется коэффициент корреляции. Тесноту связи определяют по величине коэффициента корреляции, который может принимать значения от -1 до +1 включительно.

Обычно вычисляют коэффициент корреляции Пирсона, который отражает степень линейной зависимости между двумя множествами данных:

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}$$

.

Здесь x и y - значения признаков, n - число пар данных.

Варианты связи, характеризующие наличие или отсутствие линейной связи между признаками:

- большие значения из одного набора данных связаны с большими значениями другого набора (положительная корреляция) - наличие прямой линейной связи;
- малые значения одного набора связаны с большими значениями другого (отрицательная корреляция) - наличие отрицательной линейной связи;
- данные двух диапазонов никак не связаны (нулевая корреляция) - отсутствие линейной связи.

Любая зависимость между переменными обладает двумя важными свойствами: величиной и надежностью. Чем сильнее зависимость между двумя переменными, тем больше величина зависимости и тем легче предсказать значение одной переменной по значению другой переменной. Величину зависимости легче измерить, чем надежность.

Надежность зависимости характеризует, насколько вероятно, что эта зависимость будет снова найдена на других данных. С ростом величины зависимости переменных ее надежность обычно возрастает.

Регрессионный анализ

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Рассмотрим кратко этапы регрессионного анализа.

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.

2. Определение зависимых и независимых (объясняющих) переменных.

3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.

4. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).

5. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии)

6. Оценка точности регрессионного анализа.

7. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.

8. Предсказание неизвестных значений зависимой переменной.

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та, где оно меньше нуля, - к другому классу.

Рассмотрим основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной.

Установление формы зависимости.

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

- положительная линейная регрессия (выражается в равномерном росте функции);
- положительная равноускоренно возрастающая регрессия;
- положительная равнозамедленно возрастающая регрессия;
- отрицательная линейная регрессия (выражается в равномерном падении функции);
- отрицательная равноускоренно убывающая регрессия;
- отрицательная равнозамедленно убывающая регрессия.

Описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии.

Определение функции регрессии.

Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов. Функция регрессии определяется в виде математического уравнения того или иного типа.

Оценка неизвестных значений зависимой переменной.

Решение этой задачи сводится к решению задачи одного из типов:

1. Оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции.
2. Оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

Рассмотрим некоторые предположения, на которые опирается регрессионный анализ.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Предположение линейности, т.е. предполагается, что связь между рассматриваемыми переменными является линейной. Так диаграмма рассеивания (изображения точек (x,y) с координатами переменных) обычно отображает вид зависимости. Если же на диаграмме рассеивания переменных видно явное отсутствие линейной связи, т.е. присутствует нелинейная связь, следует использовать нелинейные методы анализа.

Предположение о нормальности остатков. Оно допускает, что распределение разницы предсказанных и наблюдаемых значений является нормальным. Для визуального определения характера распределения можно воспользоваться гистограммами остатков.

При использовании регрессионного анализа следует учитывать его основное ограничение. Оно состоит в том, что регрессионный анализ позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей.

Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений.

Уравнение регрессии.

Уравнение регрессии выглядит следующим образом: Y=a+b*X.

При помощи этого уравнения переменная Y выражается через константу «a» и угол наклона прямой (или угловой коэффициент) «b», умноженный на значение переменной X. Константу «a» также называют свободным членом, а угловой коэффициент - коэффициентом регрессии или B-коэффициентом.

Как правило всегда наблюдается определенный разброс наблюдений относительно регрессионной прямой.

Остаток - это отклонение отдельной точки (наблюдения) от линии регрессии (предсказанного значения).

При оценивании уравнения регрессии обычно вычисляются следующие показатели регрессионной статистики.

Величина **R-квадрат**, называемая также мерой определенности (коэффициента детерминации), характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала [0;1]. В большинстве случаев значение R-квадрат находится между этими значениями, называемыми экстремальными.

Если значение R-квадрата близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение R-квадрата, близкое к нулю, означает плохое качество построенной модели.

Множественный R - коэффициент множественной корреляции - выражает степень зависимости независимых переменных (X) и зависимой переменной (Y).

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Множественный R равен квадратному корню из коэффициента детерминации, эта величина принимает значения в интервале от нуля до единицы. В простом линейном регрессионном анализе множественный R равен коэффициенту корреляции Пирсона.

Направление связи между переменными определяется на основании знаков (отрицательный или положительный) коэффициентов регрессии (коэффициента «b»).

Если знак при коэффициенте регрессии - положительный, связь зависимой переменной с независимой будет положительной. В нашем случае знак коэффициента регрессии положительный, следовательно, связь также является положительной.

Если знак при коэффициенте регрессии - отрицательный, связь зависимой переменной с независимой является отрицательной (обратной).

Оценив уравнение регрессии, можно решать задачу прогнозирования, которая сводится к вычислению значений Y=a+b*X по заданным значениям X при известных значениях параметров.

Если функция регрессии определена, интерпретирована и обоснована, и оценка точности регрессионного анализа соответствует требованиям, можно считать, что построенная модель и прогнозные значения обладают достаточной надежностью.

Прогнозные значения, полученные таким способом, являются средними значениями, которые можно ожидать.

Деревья решений

Метод **деревьев решений** (decision trees) является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод Data Mining также называют деревьями решающих правил, деревьями классификации и регрессии.

Если зависимая, т.е. целевая переменная принимает дискретные значения, при помощи метода дерева решений решается задача классификации.

Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т.е. решает задачу численного прогнозирования.

В наиболее простом виде дерево решений - это способ представления правил в иерархической, последовательной структуре. Основа такой структуры - ответы "Да" или "Нет" на ряд вопросов.

На рис.10 приведен пример дерева решений, задача которого - ответить на вопрос: "Играть ли в гольф?" Чтобы решить задачу, т.е. принять решение, играть ли в гольф, следует отнести текущую ситуацию к одному из известных классов (в данном случае - "играть" или "не играть"). Для этого требуется ответить на ряд вопросов, которые находятся в узлах этого дерева, начиная с его корня.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

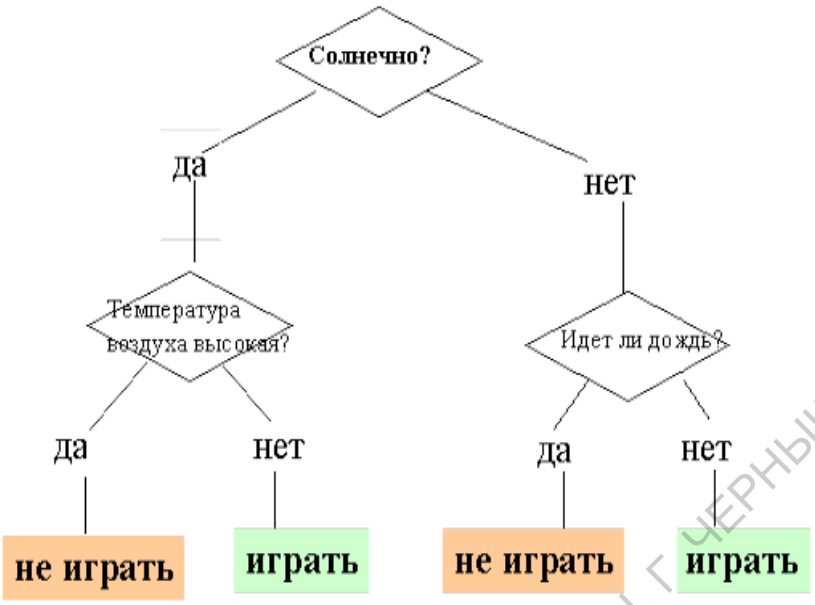

Рис.10 Дерево решений «Играть ли в гольф?»

Первый узел дерева "Солнечно?" является узлом проверки, т.е. условием. При положительном ответе на вопрос осуществляется переход к левой части дерева, называемой левой ветвью, при отрицательном - к правой части дерева. Таким образом, внутренний узел дерева является узлом проверки определенного условия. Далее идет следующий вопрос и т.д., пока не будет достигнут конечный узел дерева, являющийся узлом решения. Для нашего дерева существует два типа конечного узла: "играть" и "не играть" в гольф.

В результате прохождения от корня дерева (иногда называемого корневой вершиной) до его вершины решается задача классификации, т.е. выбирается один из классов - "играть" и "не играть" в гольф.

Целью построения дерева решения в данном случае является определение значения категориальной зависимой переменной.

В рассмотренном примере решается задача бинарной классификации, т.е. создается дихотомическая классификационная модель. Пример демонстрирует работу так называемых бинарных деревьев.

В узлах бинарных деревьев ветвление может вестись только в двух направлениях, т.е. существует возможность только двух ответов на поставленный вопрос ("да" и "нет"). Бинарные деревья являются самым простым, частным случаем деревьев решений. В остальных случаях, ответов и, соответственно, ветвей дерева, выходящих из его внутреннего узла, может быть больше двух.

Рассмотрим более сложный пример. База данных, на основе которой должно осуществляться прогнозирование, содержит следующие ретроспективные данные о клиентах банка, являющиеся ее атрибутами: возраст, наличие недвижимости, образование, среднемесячный доход, вернул ли клиент вовремя кредит. Задача состоит в том, чтобы на основании перечисленных выше данных (кроме последнего атрибута) определить, стоит ли выдавать кредит новому клиенту.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Такая задача решается в два этапа: построение классификационной модели и ее использование.

На этапе построения модели, собственно, и строится дерево классификации или создается набор неких правил. На этапе использования модели построенное дерево, или путь от его корня к одной из вершин, являющийся набором правил для конкретного клиента, используется для ответа на поставленный вопрос "Выдавать ли кредит?"

Правилом является логическая конструкция, представленная в виде "если:… то:…".

На рис.11 приведен пример дерева классификации, с помощью которого решается задача "Выдавать ли кредит клиенту?". Она является типичной задачей классификации, и при помощи деревьев решений получают достаточно хорошие варианты ее решения.

Внутренние узлы дерева (возраст, наличие недвижимости, доход и образование) являются атрибутами описанной выше базы данных. Эти атрибуты называют прогнозирующими, или атрибутами расщепления (splitting attribute). Конечные узлы дерева, или листы, именуются метками класса, являющимися значениями зависимой категориальной переменной "выдавать" или "не выдавать" кредит. Каждая ветвь дерева, идущая от внутреннего узла, отмечена предикатом расщепления. Последний может относиться лишь к одному атрибуту расщепления данного узла.

Характерная особенность предикатов расщепления: каждая запись использует уникальный путь от корня дерева только к одному узлу-решению. Объединенная информация об атрибутах расщепления и предикатах расщепления в узле называется критерием расщепления (splitting criterion).

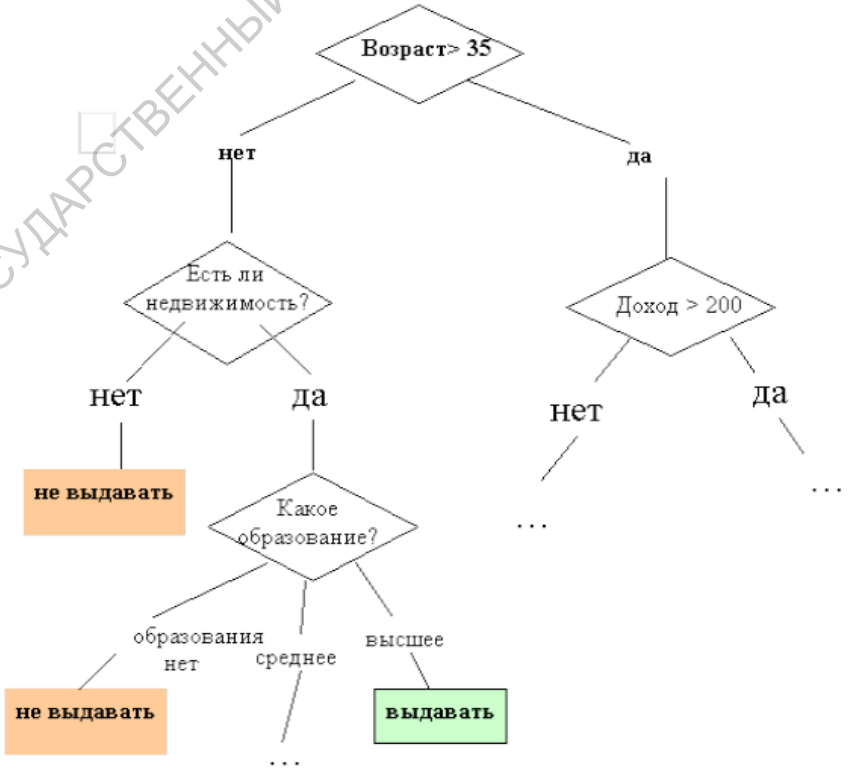

Рис.11 Дерево решений «Выдавать ли кредит?»

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

На рис.11 изображено одно из возможных деревьев решений для рассматриваемой базы данных. Например, критерий расщепления "Какое образование?", мог бы иметь два предиката расщепления и выглядеть иначе: образование "высшее" и "не высшее". Тогда дерево решений имело бы другой вид.

Таким образом, для данной задачи (как и для любой другой) может быть построено множество деревьев решений различного качества, с различной прогнозирующей точностью. Качество построенного дерева решения весьма зависит от правильного выбора критерия расщепления.

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации.

Преимущества деревьев решений.

Интуитивность 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.

Деревья решений дают возможность извлекать правила из базы данных на естественном языке. Пример правила: Если Возраст > 35 и Доход > 200, то выдать кредит.

Деревья решений позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания.

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, нейронные сети).

Быстрый процесс обучения. На построение классификационных моделей при помощи алгоритмов конструирования деревьев решений требуется значительно меньше времени, чем, например, на обучение нейронных сетей.

Большинство алгоритмов конструирования деревьев решений имеют возможность специальной обработки пропущенных значений.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Многие классические статистические методы, при помощи которых решаются задачи классификации, могут работать только с числовыми данными, в то время как деревья решений работают и с числовыми, и с категориальными типами данных.

Многие статистические методы являются параметрическими, и пользователь должен заранее владеть определенной информацией, например, знать вид модели, иметь гипотезу о виде зависимости между переменными, предполагать, какой вид распределения имеют данные. Деревья решений, в отличие от таких методов, строят непараметрические модели.

Таким образом, деревья решений способны решать такие задачи Data Mining, в которых отсутствует априорная информация о виде зависимости между исследуемыми данными.

Процесс конструирования дерева решений.

Рассматриваемая задача классификации относится к стратегии обучения с учителем, иногда называемого индуктивным обучением. В этих случаях все объекты тренировочного набора данных заранее отнесены к одному из предопределенных классов.

Алгоритмы конструирования деревьев решений состоят из этапов "построение" или "создание" дерева (tree building) и "сокращение" дерева (tree pruning). В ходе создания дерева решаются вопросы выбора критерия расщепления и остановки обучения (если это предусмотрено алгоритмом). В ходе этапа сокращения дерева решается вопрос отсечения некоторых его ветвей.

Критерий расщепления.

Процесс создания дерева происходит сверху вниз, т.е. является нисходящим. В ходе процесса алгоритм должен найти такой критерий расщепления, иногда также называемый критерием разбиения, чтобы разбить множество на подмножества, которые бы ассоциировались с данным узлом проверки. Каждый узел проверки должен быть помечен определенным атрибутом. Существует правило выбора атрибута: он должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса или же были максимально приближены к такому разбиению. Последняя фраза означает, что количество объектов из других классов, так называемых "примесей", в каждом классе должно стремиться к минимуму.

Существуют различные критерии расщепления. Наиболее известные - мера энтропии и индекс Gini.

В некоторых методах для выбора атрибута расщепления используется так называемая мера информативности подпространств атрибутов, которая основывается на энтропийном подходе и известна под названием "мера информационного выигрыша" (information gain measure) или мера энтропии.

Другой критерий расщепления, предложенный Брейманом (Breiman) и др., реализован в алгоритме CART и называется индексом Gini. При помощи

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

этого индекса атрибут выбирается на основании расстояний между распределениями классов.

Если дано множество T, включающее примеры из n классов, индекс Gini, т.е. gini(T), определяется по формуле:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

где T - текущий узел, p_j - вероятность класса j в узле T, n - количество классов.

Чем больше частных случаев описано в дереве решений, тем меньшее количество объектов попадает в каждый частный случай. Такие деревья называют "ветвистыми" или "кустистыми", они состоят из неоправданно большого числа узлов и ветвей, исходное множество разбивается на большое число подмножеств, состоящих из очень малого числа объектов. В результате "переполнения" таких деревьев их способность к обобщению уменьшается, и построенные модели не могут давать верные ответы.

В процессе построения дерева, чтобы его размеры не стали чрезмерно большими, используют специальные процедуры, которые позволяют создавать оптимальные деревья, так называемые деревья "подходящих размеров".

Какой размер дерева может считаться оптимальным? Дерево должно быть достаточно сложным, чтобы учитывать информацию из исследуемого набора данных, но одновременно оно должно быть достаточно простым. Другими словами, дерево должно использовать информацию, улучшающую качество модели, и игнорировать ту информацию, которая ее не улучшает.

Тут существует две возможные стратегии. Первая состоит в наращивании дерева до определенного размера в соответствии с параметрами, заданными пользователем.

Определение этих параметров может основываться на опыте и интуиции аналитика, а также на некоторых "диагностических сообщениях" системы, конструирующей дерево решений.

Вторая стратегия состоит в использовании набора процедур, определяющих "подходящий размер" дерева. Процедуры, которые используют для предотвращения создания чрезмерно больших деревьев, включают: сокращение дерева путем отсечения ветвей; использование правил остановки обучения.

Следует отметить, что не все алгоритмы при конструировании дерева работают по одной схеме. Некоторые алгоритмы включают два отдельных последовательных этапа: построение дерева и его сокращение; другие чередуют эти этапы в процессе своей работы для предотвращения наращивания внутренних узлов.

Остановка построения дерева.

Рассмотрим правило остановки. Оно должно определить, является ли рассматриваемый узел внутренним узлом, при этом он будет разбиваться дальше, или же он является конечным узлом, т.е. узлом решением.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Остановка - такой момент в процессе построения дерева, когда следует прекратить дальнейшие ветвления.

Один из вариантов правил остановки - "**ранняя остановка**" (prepruning), она определяет целесообразность разбиения узла. Преимущество использования такого варианта - уменьшение времени на обучение модели. Однако здесь возникает риск снижения точности классификации. Поэтому имеет смысл вместо остановки использовать «отсечение».

Второй вариант остановки обучения - **ограничение глубины** дерева. В этом случае построение заканчивается, если достигнута заданная глубина.

Еще один вариант остановки - задание **минимального количества примеров**, которые будут содержаться в конечных узлах дерева. При этом варианте ветвления продолжаются до того момента, пока все конечные узлы дерева не будут чистыми или будут содержать не более чем заданное число объектов.

Сокращение дерева или отсечение ветвей.

Решением проблемы слишком ветвистого дерева является его сокращение путем отсечения (pruning) некоторых ветвей.

Качество классификационной модели, построенной при помощи дерева решений, характеризуется двумя основными признаками: точностью распознавания и ошибкой.

Точность распознавания рассчитывается как отношение объектов, правильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Ошибка рассчитывается как отношение объектов, неправильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Отсечение ветвей или замену некоторых ветвей поддеревом следует проводить там, где эта процедура не приводит к возрастанию ошибки. Процесс проходит снизу вверх, т.е. является восходящим.

Деревья, получаемые после отсечения некоторых ветвей, называют усеченными. Если такое усеченное дерево все еще не является интуитивным и сложно для понимания, используют извлечение правил, которые объединяют в наборы для описания классов.

Каждый путь от корня дерева до его вершины или листа дает одно правило. Условиями правила являются проверки на внутренних узлах дерева.

Алгоритмы.

На сегодняшний день существует большое число алгоритмов, реализующих деревья решений: CART, C4.5, CHAID, CN2, NewId, ITrule и другие.

Алгоритм CART (Classification and Regression Tree), как видно из названия, решает задачи классификации и регрессии. Атрибуты набора данных могут иметь как дискретное, так и числовое значение. Алгоритм CART предназначен для построения бинарного дерева решений.

Особенности алгоритма CART:

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

- функция оценки качества разбиения;
- механизм отсечения дерева;
- алгоритм обработки пропущенных значений;
- построение деревьев регрессии.

Каждый узел бинарного дерева при разбиении имеет только двух потомков, называемых дочерними ветвями. Дальнейшее разделение ветви зависит от того, много ли исходных данных описывает данная ветвь. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров на две части. Правая его часть (ветвь right) - это та часть множества, в которой правило выполняется; левая (ветвь left) - та, для которой правило не выполняется.

Оценка качества разбиения, которая используется для выбора оптимального правила, основана на индексе Gini. Эта оценочная функция основана на идее уменьшения неопределенности в узле. Допустим, есть узел, и он разбит на два класса. Максимальная неопределенность в узле будет достигнута при разбиении его на два подмножества по 50% примеров, а максимальная определенность - при разбиении на 100% и 0% примеров.

Правила разбиения. Алгоритм CART работает с числовыми и категориальными атрибутами. В каждом узле разбиение может идти только по одному атрибуту. Если атрибут является числовым, то во внутреннем узле формируется правило вида $x_i \leq c$, Значение «c» в большинстве случаев выбирается как среднее арифметическое двух соседних упорядоченных значений переменной x_i обучающего набора данных. Если же атрибут относится к категориальному типу, то во внутреннем узле формируется правило $x_i \epsilon V(x_i)$, где $V(x_i)$ - некоторое непустое подмножество множества значений переменной x_i в обучающем наборе данных.

Механизм отсечения. Этим механизмом, имеющим название minimal cost-complexity tree pruning, алгоритм CART принципиально отличается от других алгоритмов конструирования деревьев решений. В рассматриваемом алгоритме отсечение - это некий компромисс между получением дерева "подходящего размера" и получением наиболее точной оценки классификации. Метод заключается в получении последовательности уменьшающихся деревьев, но деревья рассматриваются не все, а только "лучшие представители".

Перекрестная проверка (V-fold cross-validation) является наиболее сложной и одновременно оригинальной частью алгоритма CART. Она представляет собой путь выбора окончательного дерева, при условии, что набор данных имеет небольшой объем или же записи набора данных настолько специфические, что разделить набор на обучающую и тестовую выборку не представляется возможным.

Итак, основные характеристики алгоритма CART: бинарное расщепление, критерий расщепления - индекс Gini, алгоритмы minimal cost-complexity tree pruning и V-fold crossvalidation, принцип "вырастить дерево, а

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

затем сократить", высокая скорость построения, обработка пропущенных значений.

Алгоритм C4.5 строит дерево решений с неограниченным количеством ветвей у узла. Данный алгоритм может работать только с дискретным зависимым атрибутом и поэтому может решать только задачи классификации. C4.5 считается одним из самых известных и широко используемых алгоритмов построения деревьев классификации.

Для работы алгоритма C4.5 необходимо соблюдение следующих требований:

1. Каждая запись набора данных должна быть ассоциирована с одним из предопределенных классов, т.е. один из атрибутов набора данных должен являться меткой класса.
2. Классы должны быть дискретными. Каждый пример должен однозначно относиться к одному из классов.
3. Количество классов должно быть значительно меньше количества записей в исследуемом наборе данных.

Алгоритм C4.5 медленно работает на сверхбольших и зашумленных наборах данных.

Оба рассмотренных алгоритма являются робастными, т.е. устойчивыми к шумам и выбросам данных.

Алгоритмы построения деревьев решений различаются следующими характеристиками:

- вид расщепления - бинарное (binary), множественное (multi-way);
- критерии расщепления - энтропия, Gini, другие;
- возможность обработки пропущенных значений;
- процедура сокращения ветвей или отсечения;
- возможности извлечения правил из деревьев.

Ни один алгоритм построения дерева нельзя априори считать наилучшим или совершенным, подтверждение целесообразности использования конкретного алгоритма должно быть проверено и подтверждено экспериментом.

Методы классификации и прогнозирования

Метод опорных векторов

Метод опорных векторов (Support Vector Machine - SVM) относится к группе граничных методов. Она определяет классы при помощи границ областей. При помощи данного метода решаются задачи бинарной классификации.

В основе метода лежит понятие плоскостей решений.

Плоскость (plane) решения разделяет объекты с разной классовой принадлежностью.

На рис.12 приведен пример, в котором участвуют объекты двух типов. Разделяющая линия задает границу, справа от которой - все объекты типа

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

brown (коричневый - прямоугольный), а слева - типа yellow (желтый - овальный). Новый объект, попадающий направо, классифицируется как объект класса brown или - как объект класса yellow, если он расположился по левую сторону от разделяющей прямой. В этом случае каждый объект характеризуется двумя измерениями.

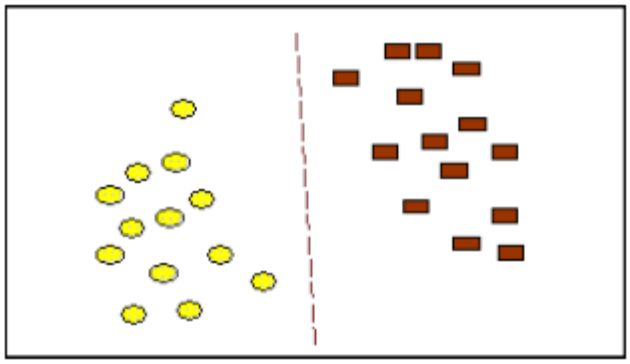

Рис.12 Разделение классов прямой линией

Цель метода опорных векторов - найти плоскость, разделяющую два множества объектов. Метод отыскивает образцы, находящиеся на границах между двумя классами, т.е. опорные вектора.

Опорными векторами называются объекты множества, лежащие на границах областей. Классификация считается хорошей, если область между границами пуста.

Линейный SVM.

Решение задачи бинарной классификации при помощи метода опорных векторов заключается в поиске некоторой линейной функции, которая правильно разделяет набор данных на два класса.

Задачу можно сформулировать как поиск функции f(x), принимающей значения меньше нуля для векторов одного класса и больше нуля - для векторов другого класса. В качестве исходных данных для решения поставленной задачи, т.е. поиска классифицирующей функции f(x), дан тренировочный набор векторов пространства, для которых известна их принадлежность к одному из классов. Семейство классифицирующих функций можно описать через функцию f(x). Гиперплоскость определена вектором «a» и значением «b», т.е. f(x)=ax+b. Решение данной задачи проиллюстрировано на рис.13.

В результате решения задачи, т.е. построения SVM-модели, найдена функция, принимающая значения меньше нуля для векторов одного класса и больше нуля – для векторов другого класса. Для каждого нового объекта отрицательное или положительное значение определяет принадлежность объекта к одному из классов.

Наилучшей функцией классификации является функция, для которой ожидаемый риск минимален. Понятие **ожидаемого риска** в данном случае означает ожидаемый уровень ошибки классификации.

Напрямую оценить ожидаемый уровень ошибки построенной модели невозможно, это можно сделать при помощи понятия эмпирического риска.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Однако следует учитывать, что минимизация последнего не всегда приводит к минимизации ожидаемого риска. Это обстоятельство следует помнить при работе с относительно небольшими наборами тренировочных данных. **Эмпирический риск** - уровень ошибки классификации на тренировочном наборе.

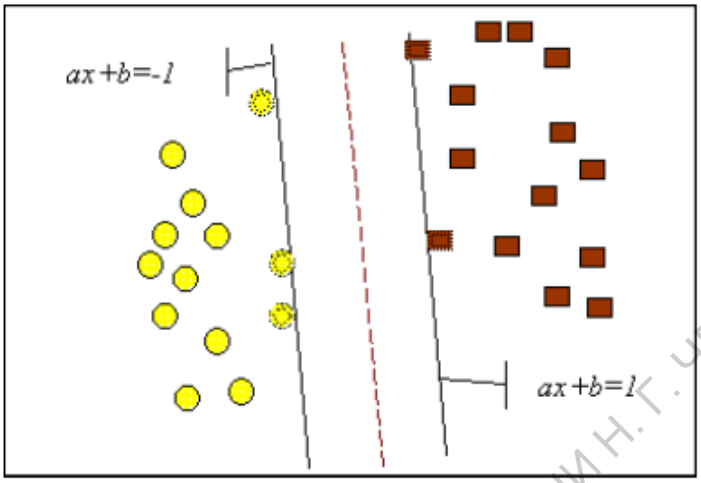

Рис.13 Линейный SVM

Таким образом, в результате решения задачи методом опорных векторов для линейно разделяемых данных получают функцию классификации, которая минимизирует верхнюю оценку ожидаемого риска.

Одной из проблем, связанных с решением задач классификации рассматриваемым методом, является то обстоятельство, что не всегда можно легко найти линейную границу между двумя классами.

В таких случаях один из вариантов - увеличение размерности, т.е. перенос данных из плоскости в трехмерное пространство, где возможно построить такую плоскость, которая идеально разделит множество образцов на два класса. Опорными векторами в этом случае будут служить объекты из обоих классов, являющиеся экстремальными.

Таким образом, при помощи добавления так называемого оператора ядра и дополнительных размерностей, находятся границы между классами в виде гиперплоскостей.

Сложность построения SVM-модели заключается в том, что чем выше размерность пространства, тем сложнее с ним работать. Один из вариантов работы с данными высокой размерности - это предварительное применение какого-либо метода понижения размерности данных для выявления наиболее существенных компонент, а затем использование метода опорных векторов.

Как и любой другой метод, метод SVM имеет свои сильные и слабые стороны, которые следует учитывать при выборе данного метода.

Недостаток метода состоит в том, что для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

Достоинство метода состоит в том, что для классификации методом опорных векторов, в отличие от большинства других методов, достаточно небольшого набора данных. При правильной работе модели, построенной на

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

тестовом множестве, вполне возможно применение данного метода на реальных данных.

Метод опорных векторов позволяет:

- получить функцию классификации с минимальной верхней оценкой ожидаемого риска (уровня ошибки классификации);
- использовать линейный классификатор для работы с нелинейно разделяемыми данными, сочетая простоту с эффективностью.

Метод "ближайшего соседа"

Метод "ближайшего соседа" ("nearest neighbour") относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.

Например, при рассмотрении нового клиента банка, его атрибуты сравниваются со всеми существующими клиентами данного банка (доход, возраст и т.д.). Множество "ближайших соседей" потенциального клиента банка выбирается на основании ближайшего значения дохода, возраста и т.д.

При таком подходе используется термин "k-ближайший сосед" ("k-nearest neighbour"). Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей". Поскольку не всегда удобно хранить все данные, иногда хранится только множество "типичных" случаев. В таком случае используемый метод называют рассуждением по аналогии (Case Based Reasoning, CBR), рассуждением на основе аналогичных случаев, рассуждением по прецедентам.

Прецедент - это описание ситуации в сочетании с подробным указанием действий, предпринимаемых в данной ситуации.

Подход, основанный на прецедентах, условно можно поделить на следующие этапы:

- сбор подробной информации о поставленной задаче;
- сопоставление этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;
- выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов;
- адаптация выбранного решения к текущей проблеме, если это необходимо;
- проверка корректности каждого вновь полученного решения;
- занесение детальной информации о новом прецеденте в базу прецедентов.

Таким образом, вывод, основанный на прецедентах, представляет собой такой метод анализа данных, который делает заключения относительно данной ситуации по результатам поиска аналогий, хранящихся в базе прецедентов.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Данный метод по своей сути относится к категории "обучение без учителя", т.е. является "самообучающейся" технологией, благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением примеров улучшаются.

Преимущества метода "ближайшего соседа".

1. Простота использования полученных результатов.
2. Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
3. Целью поиска является не гарантированно верное решение, а лучшее из возможных.

Недостатки метода "ближайшего соседа".

1. Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, - в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы.
2. Существует сложность выбора меры "близости" (метрики). От этой меры главным образом зависит объем множества записей, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза. Также существует высокая зависимость результатов классификации от выбранной метрики.
3. При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.
4. Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

Рассмотрим подробно принципы работы метода k-ближайших соседей для решения задач классификации и регрессии (прогнозирования).

Решение задачи классификации новых объектов.

Эта задача схематично изображена на рис.14. Примеры (известные экземпляры) отмечены знаком "+" или "-", определяющим принадлежность к соответствующему классу ("+" или "-"), а новый объект, который требуется классифицировать, обозначен красным кружочком. Новые объекты также называют точками запроса.

Цель заключается в оценке (классификации) отклика точек запроса с использованием специально выбранного числа их ближайших соседей. Другими словами, надо узнать, к какому классу следует отнести точку запроса: как знак "+" или как знак "-".

Для начала рассмотрим результат работы метода k-ближайших соседей с использованием одного ближайшего соседа. В этом случае отклик точки запроса будет классифицирован как знак плюс, так как ближайшая соседняя точка имеет знак плюс.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Теперь увеличим число используемых ближайших соседей до двух. На этот раз метод k-ближайших соседей не сможет классифицировать отклик точки запроса, поскольку вторая ближайшая точка имеет знак минус и оба знака равноценны (т.е. победа с одинаковым количеством голосов).

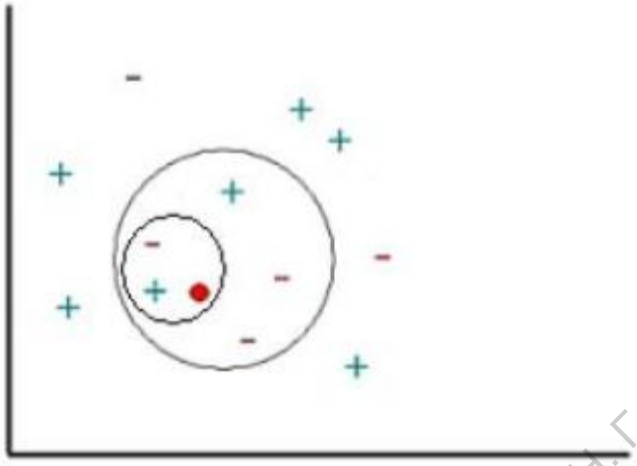

Рис.14 Классификация объектов множества при разном значении параметра k

Увеличив число используемых ближайших соседей до 5 получим окрестность точки запроса (серая окружность). Так как в области содержится 2 точки со знаком "+" и 3 точки со знаком "-", алгоритм k-ближайших соседей присвоит знак "-" отклику точки запроса.

Решение задачи прогнозирования.

Далее рассмотрим принцип работы метода k-ближайших соседей для решения задачи регрессии. Регрессионные задачи связаны с прогнозированием значения зависимой переменной по значениям независимых переменных набора данных.

Рассмотрим график, показанный на рис.15. Изображенный на ней набор точек (зеленые прямоугольники) получен по связи между независимой переменной x и зависимой переменной y (кривая красного цвета). Задан набор зеленых объектов (т.е. набор примеров); мы используем метод k-ближайших соседей для предсказания выхода точки запроса X по данному набору примеров (зеленые прямоугольники).

Сначала рассмотрим в качестве примера метод k-ближайших соседей с использованием одного ближайшего соседа (k=1). Берем набор примеров (зеленые прямоугольники) и выделяем из их числа ближайший к точке запроса X. Для этого случая ближайший пример - точка (x_4;y_4). Выход x_4 (т.е. y_4), таким образом, принимается в качестве результата предсказания выхода X (т.е. Y). Следовательно, для одного ближайшего соседа можем записать: выход Y равен y_4 ($Y=y_4$).

Далее рассмотрим ситуацию, когда k равно двум, т.е. рассмотрим двух ближайших соседей. В этом случае выделяют уже две ближайшие к X точки. На графике это точки y_3 и y_4 соответственно. Вычислив среднее их выходов, записываем решение для Y в виде $Y=(y_3 + y_4)/2$.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Решение задачи прогнозирования осуществляется путем переноса описанных выше действий на использование произвольного числа ближайших соседей таким образом, что выход Y точки запроса X вычисляется как среднеарифметическое значение выходов k- ближайших соседей точки запроса.

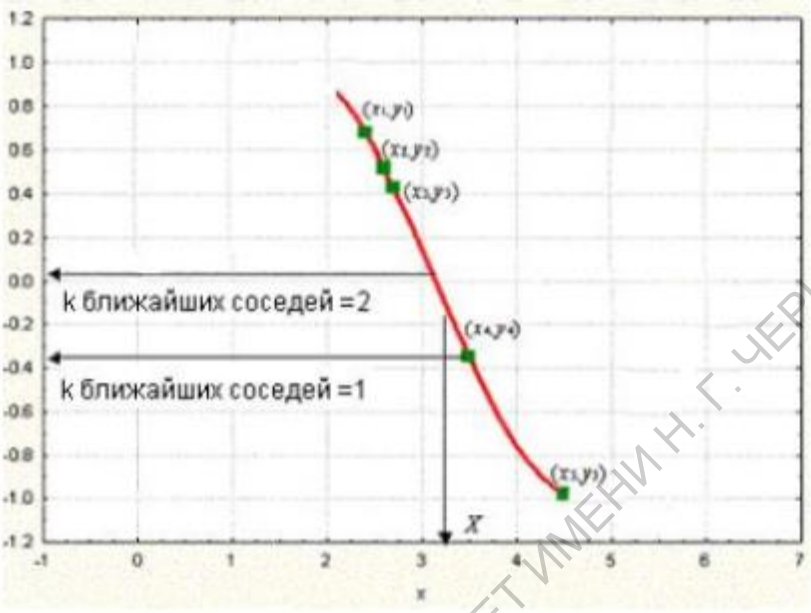

Рис.15 Решение задачи прогнозирования при разных значениях параметра k

Независимые и зависимые переменные набора данных могут быть как непрерывными, так и категориальными. Для непрерывных зависимых переменных задача рассматривается как задача прогнозирования, для дискретных переменных - как задача классификации.

Предсказание в задаче прогнозирования получается усреднением выходов k-ближайших соседей, а решение задачи классификации основано на принципе "по большинству голосов".

Критическим моментом в использовании метода k-ближайших соседей является выбор параметра k. Он один из наиболее важных факторов, определяющих качество прогнозной либо классификационной модели.

Если выбрано слишком маленькое значение параметра k, возникает вероятность большого разброса значений прогноза. Если выбранное значение слишком велико, это может привести к сильной смещенности модели. Таким образом, должно быть выбрано оптимальное значение параметра k. То есть это значение должно быть настолько большим, чтобы свести к минимуму вероятность неверной классификации, и одновременно, достаточно малым, чтобы k соседей были расположены достаточно близко к точке запроса.

Здесь k рассматривается как сглаживающий параметр, для которого должен быть найден компромисс между силой размаха (разброса) модели и ее смещенностью.

Оценка параметра k методом кросс-проверки.

Один из вариантов оценки параметра k - проведение кросс-проверки.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Основная идея метода - разделение выборки данных на v-"складок". V-"складки" - это случайным образом выделенные изолированные подвыборки.

По фиксированному значению k строится модель k-ближайших соседей для получения предсказаний на v-м сегменте (остальные сегменты при этом используются как примеры) и оценивается ошибка классификации. Для регрессионных задач наиболее часто в качестве оценки ошибки выступает сумма квадратов, а для классификационных задач удобней рассматривать точность (процент корректно классифицированных наблюдений).

Далее процесс последовательно повторяется для всех возможных вариантов выбора v. По исчерпании v-"складок" (циклов), вычисленные ошибки усредняются и используются в качестве меры устойчивости модели (т.е. меры качества предсказания в точках запроса).

Вышеописанные действия повторяются для различных k, и значение, соответствующее наименьшей ошибке (или наибольшей классификационной точности), принимается как оптимальное в смысле метода кросс-проверки.

Другой вариант выбора значения параметра k - самостоятельно задать его значение. Однако этот способ следует использовать, если имеются обоснованные предположения относительно возможного значения параметра, например, предыдущие исследования сходных наборов данных.

Метод k-ближайших соседей показывает достаточно неплохие результаты в самых разнообразных задачах.

Байесовская классификация

Альтернативные названия: байесовское моделирование, байесовская статистика, метод байесовских сетей. Изначально байесовская классификация использовалась для формализации знаний экспертов в экспертных системах, сейчас баесовская классификация также применяется в качестве одного из методов Data Mining. Так называемая наивная классификация или наивно-байесовский подход (naive-bayes approach) является наиболее простым вариантом метода, использующего байесовские сети. При этом подходе решаются задачи классификации, результатом работы метода являются так называемые "прозрачные" модели.

"Наивной" классификация называется потому, что исходит из предположения о взаимной независимости признаков.

Свойства наивной классификации:

1. Использование всех переменных и определение всех зависимостей между ними.

2. Наличие двух предположений относительно переменных:

- все переменные являются одинаково важными;
- все переменные являются статистически независимыми, т.е. значение одной переменной ничего не говорит о значении другой.

Большинство других методов классификации предполагают, что перед началом классификации вероятность того, что объект принадлежит тому или иному классу, одинакова; но это не всегда верно.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Отмечают такие достоинства байесовских сетей как метода Data Mining:

- в модели определяются зависимости между всеми переменными, это позволяет легко обрабатывать ситуации, в которых значения некоторых переменных неизвестны;
- байесовские сети достаточно просто интерпретируются и позволяют на этапе прогностического моделирования легко проводить анализ по сценарию "что, если";
- байесовский метод позволяет естественным образом совмещать закономерности, выведенные из данных, и, например, экспертные знания, полученные в явном виде;
- использование байесовских сетей позволяет избежать проблемы переучивания (overfitting), то есть избыточного усложнения модели, что является слабой стороной многих методов (например, деревьев решений и нейронных сетей).

Наивно-байесовский подход имеет следующие недостатки:

- перемножать условные вероятности корректно только тогда, когда все входные переменные действительно статистически независимы; хотя часто данный метод показывает достаточно хорошие результаты при несоблюдении условия статистической независимости, но теоретически такая ситуация должна обрабатываться более сложными методами, основанными на обучении байесовских сетей;
- невозможна непосредственная обработка непрерывных переменных - требуется их преобразование к интервальной шкале, чтобы атрибуты были дискретными; однако такие преобразования иногда могут приводить к потере значимых закономерностей;
- на результат классификации в наивно-байесовском подходе влияют только индивидуальные значения входных переменных, комбинированное влияние пар или троек значений разных атрибутов здесь не учитывается [43]. Это могло бы улучшить качество классификационной модели с точки зрения ее прогнозирующей точности, однако, увеличило бы количество проверяемых вариантов.

Нейронные сети

Идея нейронных сетей родилась в рамках теории искусственного интеллекта, в результате попыток имитировать способность биологических нервных систем обучаться и исправлять ошибки.

Нейронные сети (Neural Networks) - это модели биологических нейронных сетей мозга, в которых нейроны имитируются относительно простыми, часто однотипными, элементами (искусственными нейронами).

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

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.

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.

Модели нейронных сетей могут быть программного и аппаратного исполнения. Будем рассматривать сети первого типа. 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, рассматриваются такие:

1. Классификация (обучение с учителем): распознавание текста, распознавание речи, идентификация личности.

2. Прогнозирование: найти наилучшее приближение функции, заданной конечным набором входных значений (обучающих примеров). Например, нейронные сети позволяют решать задачу восстановления пропущенных значений.

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

Элементы нейронных сетей.

Искусственный нейрон (формальный нейрон) - элемент искусственных нейронных сетей, моделирующий некоторые функции биологического нейрона. Главная функция искусственного нейрона - формировать выходной сигнал в зависимости от сигналов, поступающих на его входы.

В самой распространенной конфигурации входные сигналы обрабатываются адаптивным сумматором, затем выходной сигнал сумматора поступает в нелинейный преобразователь, где преобразуется функцией активации, и результат подается на выход (в точку ветвления).

Общий вид искусственного нейрона приведен на рис.16.

Нейрон характеризуется текущим состоянием и обладает группой синапсов - однонаправленных входных связей, соединенных с выходами других нейронов. Нейрон имеет аксон - выходную связь данного нейрона, с которой сигнал (возбуждения или торможения) поступает на синапсы следующих нейронов.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

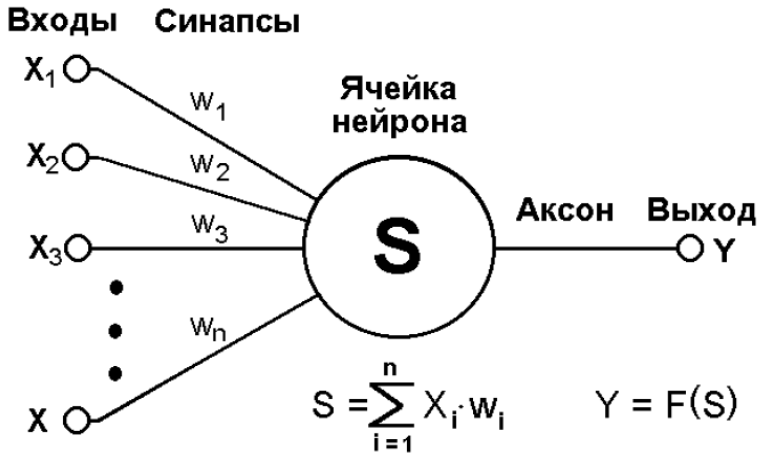

Рис.16 Искусственный нейрон

Каждый синапс характеризуется величиной синаптической связи (ее весом w_i). Текущее состояние нейрона определяется как взвешенная сумма его входов:

$$s = \sum_{i=1}^{n} x_i \cdot w_i$$

Выход нейрона есть функция его состояния: y=f(s).

Активационная функция, которую также называют характеристической функцией, - это нелинейная функция, вычисляющая выходной сигнал формального нейрона.

Часто используемые активационные функции:

1. Жесткая пороговая функция.
2. Линейный порог.
3. Сигмоидальная функция.

Выбор активационной функции определяется спецификой поставленной задачи либо ограничениями, накладываемыми некоторыми алгоритмами обучения.

Нелинейный преобразователь - это элемент искусственного нейрона, преобразующий текущее состояние нейрона (выходной сигнал адаптивного сумматора) в выходной сигнал нейрона по некоторому нелинейному закону (активационной функции).

Точка ветвления (выход) - это элемент формального нейрона, посылающий его выходной сигнал по нескольким адресам и имеющий один вход и несколько выходов. На вход точки ветвления обычно подается выходной сигнал нелинейного преобразователя, который затем посылается на входы других нейронов.

Архитектура нейронных сетей.

Нейронные сети могут быть синхронные и асинхронные.

В **синхронных нейронных сетях** в каждый момент времени свое состояние меняет лишь один нейрон.

В **асинхронных** - состояние меняется сразу у целой группы нейронов, как правило, у всего слоя.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Можно выделить две базовые архитектуры - слоистые и полносвязные сети. Ключевым в слоистых сетях является понятие слоя.

Слой - один или несколько нейронов, на входы которых подается один и тот же общий сигнал.

Слоистые нейронные сети - нейронные сети, в которых нейроны разбиты на отдельные группы (слои) так, что обработка информации осуществляется послойно. В слоистых сетях нейроны i-го слоя получают входные сигналы, преобразуют их и через точки ветвления передают нейронам (i+1) слоя. И так до k-го слоя, который выдает выходные сигналы для интерпретатора и пользователя. Число нейронов в каждом слое не связано с количеством нейронов в других слоях, может быть произвольным.

В рамках одного слоя данные обрабатываются параллельно, а в масштабах всей сети обработка ведется последовательно - от слоя к слою. К слоистым нейронным сетям относятся, например, многослойные персептроны, сети радиальных базисных функций, когнитрон, некогнитрон, сети ассоциативной памяти.

Сигнал не всегда подается на все нейроны слоя. В когнитроне, например, каждый нейрон текущего слоя получает сигналы только от близких ему нейронов предыдущего слоя.

Слоистые сети, в свою очередь, могут быть однослойными и многослойными.

Однослойная сеть - сеть, состоящая из одного слоя.

Многослойная сеть - сеть, имеющая несколько слоев.

В многослойной сети первый слой называется входным, последующие - внутренними или скрытыми, последний слой - выходным. Таким образом, промежуточные слои - это все слои в многослойной нейронной сети, кроме входного и выходного. Входной слой сети реализует связь с входными данными, выходной - с выходными.

Нейроны могут быть входными, выходными и скрытыми.

Входной слой организован из входных нейронов (input neuron), которые получают данные и распространяют их на входы нейронов скрытого слоя сети. Скрытый нейрон (hidden neuron) - это нейрон, находящийся в скрытом слое нейронной сети. Выходные нейроны (output neuron), из которых организован выходной слой сети, выдает результаты работы нейронной сети.

В полносвязных сетях каждый нейрон передает свой выходной сигнал остальным нейронам, включая самого себя. Выходными сигналами сети могут быть все или некоторые выходные сигналы нейронов после нескольких тактов функционирования сети. Все входные сигналы подаются всем нейронам.

Обучение нейронных сетей.

Перед использованием нейронной сети ее необходимо обучить.

Процесс обучения нейронной сети заключается в подстройке ее внутренних параметров под конкретную задачу.

Алгоритм работы нейронной сети является итеративным, его шаги называют эпохами или циклами.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Эпоха - одна итерация в процессе обучения, включающая предъявление всех примеров из обучающего множества и, возможно, проверку качества обучения на контрольном множестве.

Процесс обучения осуществляется на обучающей выборке.

Обучающая выборка включает входные значения и соответствующие им выходные значения набора данных. В ходе обучения нейронная сеть находит некие зависимости выходных полей от входных.

Таким образом, ставится вопрос - какие входные поля (признаки) необходимо использовать. Первоначально выбор осуществляется эвристически, далее количество входов может быть изменено.

Сложность может вызвать вопрос о количестве наблюдений в наборе данных. Количество необходимых наблюдений зависит от сложности решаемой задачи. При увеличении количества признаков количество наблюдений возрастает нелинейно, эта проблема носит название "проклятие размерности". При недостаточном количестве данных рекомендуется использовать линейную модель.

Необходимо определить количество слоев в сети и количество нейронов в каждом слое, назначить такие значения весов и смещений, которые смогут минимизировать ошибку решения. Веса и смещения автоматически настраиваются таким образом, чтобы минимизировать разность между желаемым и полученным на выходе сигналами, которая называется **ошибка обучения**.

Ошибка обучения для построенной нейронной сети вычисляется путем сравнения выходных и целевых (желаемых) значений. Из полученных разностей формируется функция ошибок.

Функция ошибок - это целевая функция, требующая минимизации в процессе управляемого обучения нейронной сети. С помощью функции ошибок можно оценить качество работы нейронной сети во время обучения. Например, сумма квадратов ошибок.

От качества обучения нейронной сети зависит ее способность решать поставленные перед ней задачи.

Переобучение нейронной сети.

При обучении нейронных сетей часто возникает серьезная трудность, называемая проблемой переобучения (overfitting).

Переобучение, или чрезмерно близкая подгонка - излишне точное соответствие нейронной сети конкретному набору обучающих примеров, при котором сеть теряет способность к обобщению.

Переобучение возникает в случае слишком долгого обучения, недостаточного числа обучающих примеров или переусложненной структуры нейронной сети. Переобучение связано с тем, что выбор обучающего (тренировочного) множества является случайным. С первых шагов обучения происходит уменьшение ошибки. На последующих шагах с целью уменьшения ошибки (целевой функции) параметры подстраиваются под особенности обучающего множества. Однако при этом происходит

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

"подстройка" не под общие закономерности ряда, а под особенности его части - обучающего подмножества. При этом точность прогноза уменьшается.

Один из вариантов борьбы с переобучением сети - деление обучающей выборки на два множества (обучающее и тестовое).

На обучающем множестве происходит обучение нейронной сети. На тестовом множестве осуществляется проверка построенной модели. Эти множества не должны пересекаться. С каждым шагом параметры модели изменяются, однако постоянное уменьшение значения целевой функции происходит именно на обучающем множестве. При разбиении множества на два мы можем наблюдать изменение ошибки прогноза на тестовом множестве параллельно с наблюдениями над обучающим множеством. Какое-то количество шагов ошибки прогноза уменьшается на обоих множествах. Однако на определенном шаге ошибка на тестовом множестве начинает возрастать, при этом ошибка на обучающем множестве продолжает уменьшаться. Этот момент считается концом реального или настоящего обучения, с него и начинается переобучение.

Прогноз на тестовом множестве является проверкой работоспособности построенной модели. Ошибка на тестовом множестве может являться ошибкой прогноза, если тестовое множество максимально приближено к текущему моменту.

Модели нейронных сетей.

Рассмотрим наиболее простые модели нейронных сетей: однослойный и многослойный персептрон.

Однослойный персептрон (персептрон Розенблатта) - однослойная нейронная сеть, все нейроны которой имеют жесткую пороговую функцию активации.

Однослойный персептрон имеет простой алгоритм обучения и способен решать лишь самые простые задачи. Классический пример такой нейронной сети - однослойный трехнейронный персептрон - представлен на рис.17.

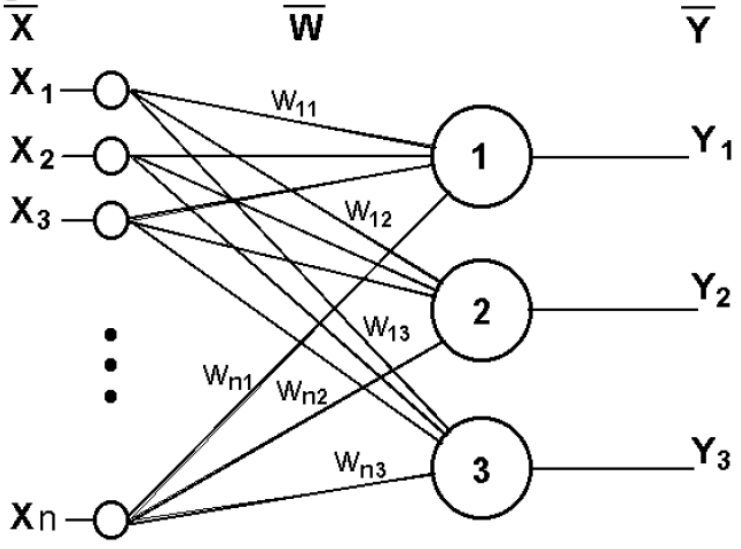

Рис. 17 Однослойный трехнейронный персептрон

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Сеть, изображенная на рисунке, имеет n входов, на которые поступают сигналы, идущие по синапсам на 3 нейрона. Эти три нейрона образуют единственный слой данной сети и выдают три выходных сигнала.

Многослойный персептрон (MLP) - нейронная сеть прямого распространения сигнала (без обратных связей), в которой входной сигнал преобразуется в выходной, проходя последовательно через несколько слоев.

Первый из таких слоев называют входным, последний - выходным. Эти слои содержат так называемые вырожденные нейроны и иногда в количестве слоев не учитываются. Кроме входного и выходного слоев, в многослойном персептроне есть один или несколько промежуточных слоев, которые называют скрытыми.

В этой модели персептрона должен быть хотя бы один скрытый слой. Присутствие нескольких таких слоев оправдано лишь в случае использования нелинейных функций активации.

Пример двухслойного персептрона представлен на рис.18.

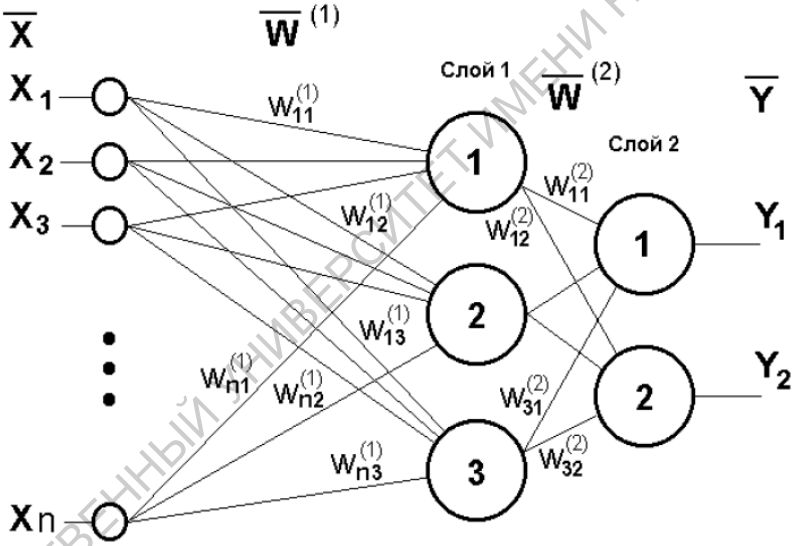

Рис. 18 Двухслойный перцептрон

Сеть, изображенная на рисунке, имеет n входов. На них поступают сигналы, идущие далее по синапсам на 3 нейрона, которые образуют первый слой. Выходные сигналы первого слоя передаются двум нейронам второго слоя. Последние, в свою очередь, выдают два выходных сигнала.

Метод обратного распространения ошибки (Back propagation, backprop) – алгоритм обучения многослойных персептронов, основанный на вычислении градиента функции ошибок. В процессе обучения веса нейронов каждого слоя нейросети корректируются с учетом сигналов, поступивших с предыдущего слоя, и невязки каждого слоя, которая вычисляется рекурсивно в обратном направлении от последнего слоя к первому.

Программное обеспечение, имитирующее работу нейронной сети, называют нейросимулятором либо нейропакетом.

Большинство нейропакетов включают следующую последовательность действий:

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

1. Создание сети (выбор пользователем параметров либо одобрение установленных по умолчанию).
2. Обучение сети.
3. Выдача пользователю решения.

Самоорганизующиеся карты Кохонена.

Классификация нейронных сетей

Одна из возможных классификаций нейронных сетей - по направленности связей.

Нейронные сети бывают с обратными связями и без обратных связей.

Сети без обратных связей.

1. Сети с обратным распространением ошибки.

Сети этой группы характеризуются фиксированной структурой, итерационным обучением, корректировкой весов по ошибкам.

2. Другие сети (когнитрон, неокогнитрон, другие сложные модели).

Преимуществами сетей без обратных связей является простота их реализации и гарантированное получение ответа после прохождения данных по слоям. Недостатком этого вида сетей считается минимизация размеров сети – нейроны многократно участвуют в обработке данных.

Меньший объем сети облегчает процесс обучения.

Сети с обратными связями.

1. Сети Хопфилда (задачи ассоциативной памяти).
2. Сети Кохонена (задачи кластерного анализа).

Преимуществами сетей с обратными связями является сложность обучения, вызванная большим числом нейронов для алгоритмов одного и того же уровня сложности.

Недостатки этого вида сетей - требуются специальные условия, гарантирующие сходимость вычислений.

Другая классификация нейронных сетей: сети прямого распространения и рекуррентные сети.

Сети прямого распространения:

1. Персептроны.
2. Сеть Back Propagation.
3. Сеть встречного распространения.
4. Карта Кохонена.

Рекуррентные сети.

1. Сеть Хопфилда.
2. Сеть Элмана - сеть, состоящая из двух слоев, в которой скрытый слой охвачен динамической обратной связью, что позволяет учесть предысторию наблюдаемых процессов и накопить информацию для выработки правильной стратегии

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

управления. Эти сети применяются в системах управления движущимися объектами.

Характерная особенность таких сетей - наличие блоков динамической задержки и обратных связей, что позволяет им обрабатывать динамические модели.

Нейронные сети могут обучаться с учителем или без него.

При обучении с учителем для каждого обучающего входного примера требуется знание правильного ответа или функции оценки качества ответа. Такое обучение называют управляемым. Нейронной сети предъявляются значения входных и выходных сигналов, а она по определенному алгоритму подстраивает веса синаптических связей. В процессе обучения производится корректировка весов сети по результатам сравнения фактических выходных значений с входными, известными заранее.

При обучении без учителя раскрывается внутренняя структура данных или корреляции между образцами в наборе данных. Выходы нейронной сети формируются самостоятельно, а веса изменяются по алгоритму, учитывающему только входные и производные от них сигналы. Это обучение называют также неуправляемым. В результате такого обучения объекты или примеры распределяются по категориям, сами категории и их количество могут быть заранее не известны.

Подготовка данных для обучения.

При подготовке данных для обучения нейронной сети необходимо обращать внимание на следующие существенные моменты.

Количество наблюдений в наборе данных. Следует учитывать тот фактор, что чем больше размерность данных, тем больше времени потребуется для обучения сети.

Работа с выбросами. Следует определить наличие выбросов и оценить необходимость их присутствия в выборке. Обучающая выборка должна быть представительной (репрезентативной), не должна содержать противоречий, так как нейронная сеть однозначно сопоставляет выходные значения входным.

Нейронная сеть работает только с числовыми входными данными, поэтому важным этапом при подготовке данных является преобразование и кодирование данных.

При использовании на вход нейронной сети следует подавать значения из того диапазона, на котором она обучалась. Например, если при обучении нейронной сети на один из ее входов подавались значения от 0 до 10, то при ее применении на вход следует подавать значения из этого же диапазона или близлежащие.

Существует понятие нормализации данных. Целью нормализации значений является преобразование данных к виду, который наиболее подходит для обработки, т.е. данные, поступающие на вход, должны иметь числовой тип, а их значения должны быть распределены в определенном диапазоне. Нормализатор может приводить дискретные данные к набору уникальных индексов либо преобразовывать значения, лежащие в произвольном

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

диапазоне, в конкретный диапазон, например, [0..1]. Нормализация выполняется путем деления каждой компоненты входного вектора на длину вектора, что превращает входной вектор в единичный.

Выбор структуры нейронной сети.

Выбор структуры нейронной сети обуславливается спецификой и сложностью решаемой задачи. В большинстве случаев выбор структуры нейронной сети определяется на основе объединения опыта и интуиции разработчика. Однако существуют основополагающие принципы, которыми следует руководствоваться при разработке новой конфигурации:

1. возможности сети возрастают с увеличением числа ячеек сети, плотности связей между ними и числом выделенных слоев;

2. введение обратных связей наряду с увеличением возможностей сети поднимает вопрос о динамической устойчивости сети;

3. сложность алгоритмов функционирования сети (в том числе, например, введение нескольких типов синапсов - возбуждающих, тормозящих и др.) также способствует усилению мощи НС.

Вопрос о необходимых и достаточных свойствах сети для решения того или иного рода задач представляет собой целое направление нейрокомпьютерной науки. Так как проблема синтеза нейронной сети сильно зависит от решаемой задачи, дать общие подробные рекомендации затруднительно. Очевидно, что процесс функционирования НС, то есть сущность действий, которые она способна выполнять, зависит от величин синаптических связей, поэтому, задавшись определенной структурой НС, отвечающей какой-либо задаче, разработчик сети должен найти оптимальные значения всех переменных весовых коэффициентов (некоторые синаптические связи могут быть постоянными).

Карты Кохонена

Самоорганизующиеся карты (Self-Organizing Maps, SOM).

Сети, называемые картами Кохонена, - это одна из разновидностей нейронных сетей, однако они принципиально отличаются от рассмотренных выше, поскольку используют неконтролируемое обучение. При таком обучении обучающее множество состоит лишь из значений входных переменных, в процессе обучения нет сравнивания выходов нейронов с эталонными значениями. Можно сказать, что такая сеть учится понимать структуру данных.

Основной принцип работы сетей - введение в правило обучения нейрона информации относительно его расположения.

В основе идеи сети Кохонена лежит аналогия со свойствами человеческого мозга. Кора головного мозга человека представляет собой плоский лист и свернута складками. Таким образом, можно сказать, что она обладает определенными топологическими свойствами (участки, ответственные за близкие части тела, примыкают друг к другу и все

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

изображение человеческого тела отображается на эту двумерную поверхность).

Задачи, решаемые при помощи карт Кохонена.

Самоорганизующиеся карты могут использоваться для решения таких задач, как моделирование, прогнозирование, поиск закономерностей в больших массивах данных, выявление наборов независимых признаков и сжатие информации.

Наиболее распространенное применение сетей Кохонена - решение задачи классификации без учителя, т.е. кластеризации.

При такой постановке задачи дается набор объектов, каждому из которых сопоставлена строка таблицы (вектор значений признаков). Требуется разбить исходное множество на классы, т.е. для каждого объекта найти класс, к которому он принадлежит.

В результате получения новой информации о классах возможна коррекция существующих правил классификации объектов.

Наиболее распространенным применением карт Кохонена являются разведочный анализ данных и обнаружение новых явлений.

Разведочный анализ данных. Сеть Кохонена способна распознавать кластеры в данных, а также устанавливать близость классов. Таким образом, пользователь может улучшить свое понимание структуры данных, чтобы затем уточнить нейросетевую модель. Если в данных распознаны классы, то их можно обозначить, после чего сеть сможет решать задачи классификации. Сети Кохонена можно использовать и в тех задачах классификации, где классы уже заданы, - тогда преимущество будет в том, что сеть сможет выявить сходство между различными классами.

Обнаружение новых явлений. Сеть Кохонена распознает кластеры в обучающих данных и относит все данные к тем или иным кластерам. Если после этого сеть встретится с набором данных, непохожим ни на один из известных образцов, то она не сможет классифицировать такой набор и тем самым выявит его новизну.

Обучение сети Кохонена.

Сеть Кохонена, в отличие от многослойной нейронной сети, очень проста; она представляет собой два слоя: входной и выходной. Ее также называют самоорганизующейся картой. Элементы карты располагаются в некотором пространстве, как правило, двумерном. Сеть Кохонена изображена на рис.19.

Сеть Кохонена обучается методом последовательных приближений. В процессе обучения таких сетей на входы подаются данные, но сеть при этом подстраивается не под эталонное значение выхода, а под закономерности во входных данных. Начинается обучение с выбранного случайным образом выходного расположения центров.

В процессе последовательной подачи на вход сети обучающих примеров определяется наиболее схожий нейрон (тот, у которого скалярное произведение весов и поданного на вход вектора минимально). Этот нейрон

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

объявляется победителем и является центром при подстройке весов у соседних нейронов. Такое правило обучения предполагает "соревновательное" обучение с учетом расстояния нейронов от "нейрона-победителя".

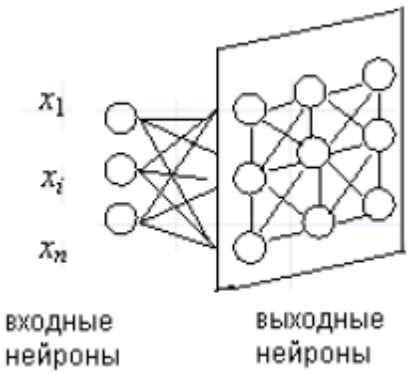

Рис.19 Сеть Кохонена

Обучение при этом заключается не в минимизации ошибки, а в подстройке весов (внутренних параметров нейронной сети) для наибольшего совпадения с входными данными.

Основной итерационный алгоритм Кохонена последовательно проходит ряд эпох, на каждой из которых обрабатывается один пример из обучающей выборки. Входные сигналы последовательно предъявляются сети, при этом желаемые выходные сигналы не определяются. После предъявления достаточного числа входных векторов синаптические веса сети становятся способны определить кластеры. Веса организуются так, что топологически близкие узлы чувствительны к похожим входным сигналам.

В результате работы алгоритма центр кластера устанавливается в определенной позиции, удовлетворительным образом кластеризующей примеры, для которых данный нейрон является "победителем". В результате обучения сети необходимо определить меру соседства нейронов, т.е. окрестность нейрона-победителя. Окрестность представляет собой несколько нейронов, которые окружают нейрон- победитель.

Сначала к окрестности принадлежит большое число нейронов, далее ее размер постепенно уменьшается. Сеть формирует топологическую структуру, в которой похожие примеры образуют группы примеров, близко находящиеся на топологической карте.

Полученную карту можно использовать как средство визуализации при анализе данных. В результате обучения карта Кохонена классифицирует входные примеры на кластеры (группы схожих примеров) и визуально отображает многомерные входные данные на плоскости нейронов.

Уникальность метода самоорганизующихся карт состоит в преобразовании n-мерного пространства в двухмерное. Применение двухмерных сеток связано с тем, что существует проблема отображения пространственных структур большей размерности. Имея такое представление данных, можно визуально определить наличие или отсутствие взаимосвязи во входных данных.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Нейроны карты Кохонена располагают в виде двухмерной матрицы, раскрашивают эту матрицу в зависимости от анализируемых параметров нейронов. На рис.20 приведен пример карты Кохонена. Группа объектов, обозначенная красным цветом имеет наибольшие значения рассматриваемого показателя, группа объектов, обозначенная синим цветом - наименьшие значения.

Карты Кохонена (как и географические карты) можно отображать в двухмерном виде, тогда карта раскрашивается в соответствии с уровнем выхода нейрона или в трехмерном виде.

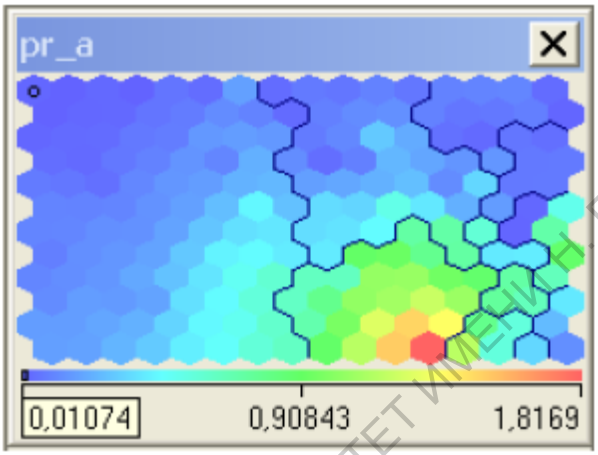

Рис.20 Пример карты Кохонена

В результате работы алгоритма получают такие карты:

- карта входов нейронов;
- карта выходов нейронов;
- специальные карты.

Координаты каждой карты определяют положение одного нейрона. Так, координаты [15:30] определяют нейрон, который находится на пересечении 15-го столбца с 30-м рядом в матрице нейронов.

Карта входов нейронов.

Веса нейронов подстраиваются под значения входных переменных и отображают их внутреннюю структуру. Для каждого входа рисуется своя карта, раскрашенная в соответствии со значением конкретного веса нейрона. При анализе данных используют несколько карт входов. На одной из карт выделяют область определенного цвета - это означает, что соответствующие входные примеры имеют приблизительно одинаковое значение соответствующего входа. Цветовое распределение нейронов из этой области анализируется на других картах для определения схожих или отличительных характеристик.

Карта выходов нейронов.

На карту выходов нейронов проецируется взаимное расположение исследуемых входных данных. Нейроны с одинаковыми значениями выходов образуют кластеры – замкнутые области на карте, которые включают нейроны с одинаковыми значениями выходов.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Специальные карты - это карта кластеров, матрица расстояний, матрица плотности попадания и другие карты, которые характеризуют кластеры, полученные в результате обучения сети Кохонена.

Важно понимать, что между всеми рассмотренными картами существует взаимосвязь – все они являются разными раскрасками одних и тех же нейронов. Каждый пример из обучающей выборки имеет одно и то же расположение на всех картах.

Методы кластерного анализа

Опишем понятие "кластер" с математической точки зрения, а также рассмотрим методы кластерного анализа. Кластерный анализ включает в себя более 100 различных алгоритмов. В отличие от задач классификации, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). Все переменные должны измеряться в сравнимых шкалах.

Кластерный анализ может применяться к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

№ примера	признак X	признак Y
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Рис.21 Набор данных A

Задачи кластерного анализа можно объединить в следующие группы:

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Как правило, при практическом использовании кластерного анализа одновременно решается несколько из указанных задач.

Рассмотрим пример процедуры кластерного анализа.

Допустим, мы имеем набор данных A, состоящий из 14-ти примеров, у которых имеется по два признака X и Y. Данные по ним приведены рис.21.

Данные в табличной форме не носят информативный характер. Представим переменные X и Y в виде диаграммы рассеивания, изображенной на рис.22

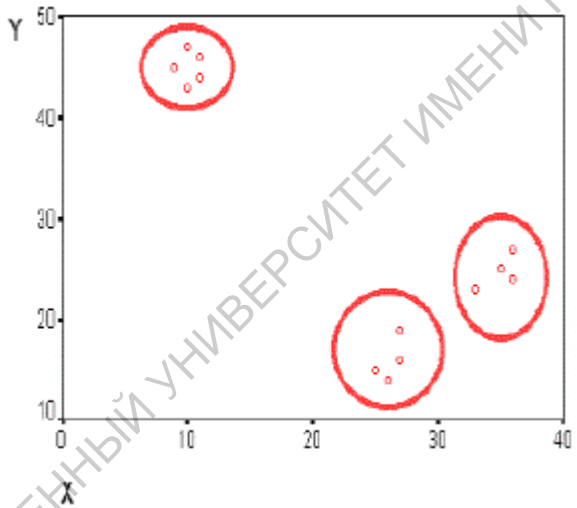

Рис.22 Диаграмма рассеивания переменных X и Y

На рисунке мы видим несколько групп "похожих" примеров. Примеры (объекты), которые по значениям X и Y "похожи" друг на друга, принадлежат к одной группе (кластеру); объекты из разных кластеров не похожи друг на друга.

Критерием для определения схожести и различия кластеров является расстояние междуточками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ - вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Центр кластера - это среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от центра кластера.

Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. Второе предположение - правильность выбора масштаба или единиц измерения признаков.

Выбор масштаба в кластерном анализе имеет большое значение. Рассмотрим пример.

Представим себе, что данные признака X в наборе данных A на два порядка больше данных признака У (значения переменной X находятся в диапазоне от 100 до 700, а значения переменной У - в диапазоне от 0 до 1).

Тогда, при расчете величины расстояния между точками, отражающими положение объектов в пространстве их свойств, переменная, имеющая большие значения, т.е. переменная X, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной У. Таким образом из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками.

Эта проблема решается при помощи предварительной стандартизации переменных.

Стандартизация (standardization) или **нормирование** (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных.

Два наиболее распространенных способа:

- деление исходных данных на среднеквадратичное отклонение соответствующих переменных;
- вычисление Z-вклада или стандартизованного вклада.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Наряду со стандартизацией переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

В ходе экспериментов возможно сравнение результатов, полученных с учетом экспертных оценок и без них, и выбор лучшего из них.

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

- иерархические;
- неиерархические.

Каждая из групп включает множество подходов и алгоритмов.

Используя различные методы кластерного анализа, можно получить различные решения для одних и тех же данных. Это считается нормальным явлением.

Рассмотрим иерархические и неиерархические методы подробно.

Иерархические методы кластерного анализа.

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

Иерархические агломеративные методы (Agglomerative Nesting, AGNES).

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA).

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде дендрограммы показан на рис.23.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах Data Mining, которые позволяют решать задачи достаточно большой размерности.

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

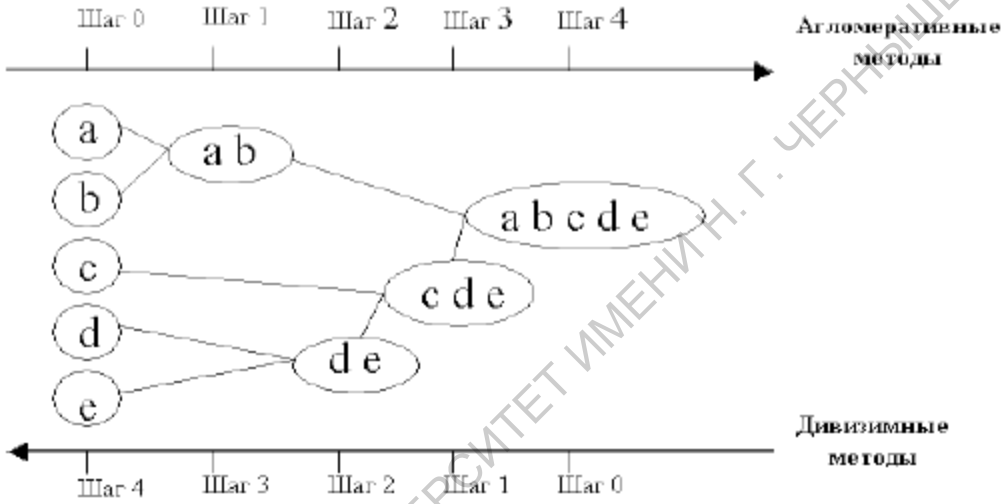

Рис.23 Дендрограмма агломеративных и дивизимных методов

Преимуществом иерархических методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа.

Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально.

Меры сходства.

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. **Евклидово расстояние** - наиболее популярная мера сходства.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Квадрат евклидова расстояния. Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом евклидова расстояния путем возведения в квадрат стандартного евклидова расстояния.

Манхэттенское расстояние (расстояние городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием. Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат.

Расстояние Чебышева. Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

Процент несогласия. Это расстояние вычисляется, если данные являются категориальными.

Методы объединения или связи.

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.

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"). Метод хорошо использовать, 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.

Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. 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

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и "стремится" создавать кластеры малого размера.

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 невзвешенного 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). Этот метод похож на предыдущий, разница состоит в том, что для учета разницы между размерами кластеров (числе объектов в них), используются веса. Этот метод предпочтительно использовать в случаях, если имеются предположения относительно существенных отличий в размерах кластеров.

Определение количества кластеров.

Существует проблема определения числа кластеров. Иногда можно априорно определить это число. Однако в большинстве случаев число кластеров определяется в процессе агломерации/разделения множества объектов.

Процессу группировки объектов в иерархическом кластерном анализе соответствует постепенное возрастание коэффициента, называемого критерием Е. Скачкообразное увеличение значения критерия Е можно определить как характеристику числа кластеров, которые действительно существуют в исследуемом наборе данных. Таким образом, этот способ сводится к определению скачкообразного увеличения некоторого коэффициента, который характеризует переход от сильно связанного к слабо связанному состоянию объектов.

Итеративные методы кластерного анализа.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов.

Алгоритм k-средних (k-means).

Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Описание алгоритма

1. Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр. Выбор начальных центроидов может осуществляться следующим образом:

- выбор k-наблюдений для максимизации начального расстояния;
- случайный выбор k-наблюдений;
- выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

2. Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

- кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

На рис.24 приведен пример работы алгоритма k-средних для k, равного двум.

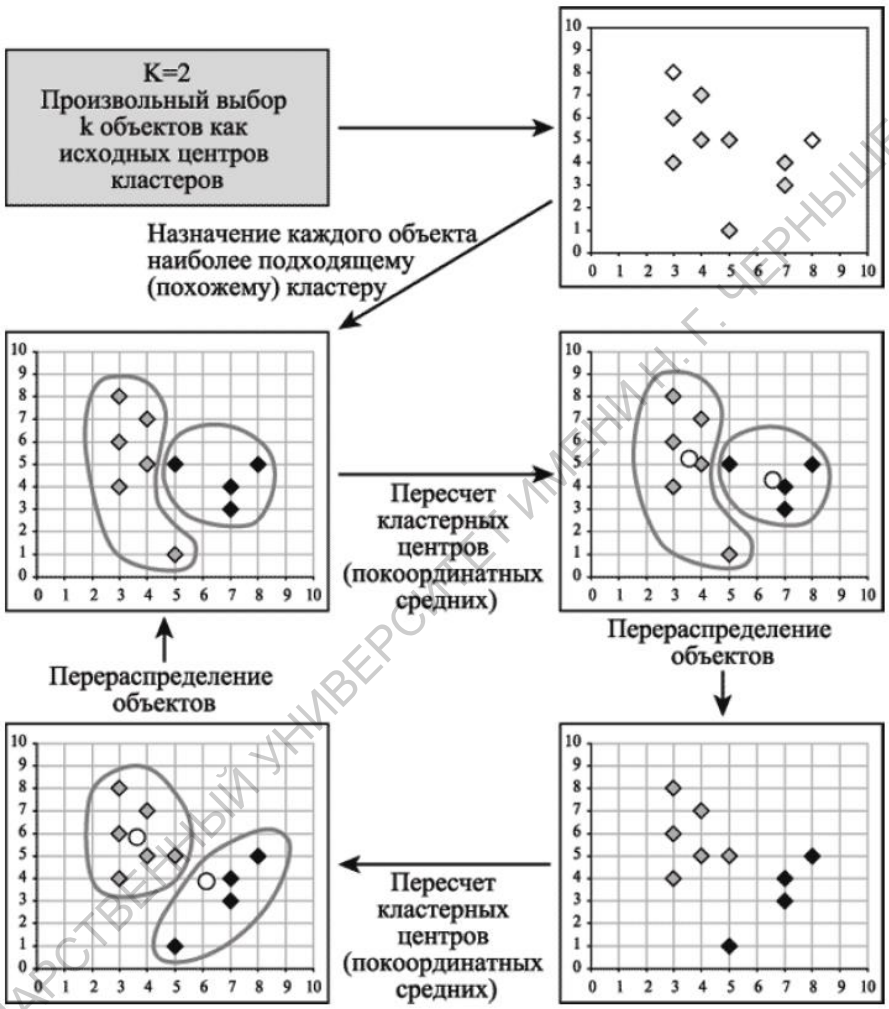

Рис.24 Пример работы алгоритма k-средних (k=2)

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

Проверка качества кластеризации.

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга).

Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Достоинства алгоритма k-средних:

- простота использования;

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

- быстрота использования;
- понятность и прозрачность алгоритма.

Недостатки алгоритма k-средних:

- алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;
- алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

Алгоритм PAM (partitioning around Medoids).

PAM является модификацией алгоритма k-средних, алгоритмом k-медианы (k-medoids). Алгоритм менее чувствителен к шумам и выбросам данных, чем алгоритм k-means, поскольку медиана меньше подвержена влияниям выбросов. PAM эффективен для небольших баз данных, но его не следует использовать для больших наборов данных.

Предварительное сокращение размерности.

Более понятные и прозрачные результаты кластеризации могут быть получены, если вместо множества исходных переменных использовать некие обобщенные переменные или критерии, содержащие в сжатом виде информацию о связях между переменными. Т.е. возникает задача понижения размерности данных. Она может решаться при помощи различных методов; один из наиболее распространенных - факторный анализ.

Факторный анализ - это метод, применяемый для изучения взаимосвязей между значениями переменных.

Факторный анализ преследует две цели:

- сокращение числа переменных;
- классификацию переменных - определение структуры взаимосвязей между переменными.

Соответственно, факторный анализ может использоваться для решения задач сокращения размерности данных или для решения задач классификации.

Критерии или главные факторы, выделенные в результате факторного анализа, содержат в сжатом виде информацию о существующих связях между переменными. Эта информация позволяет получить лучшие результаты кластеризации и лучше объяснить семантику кластеров. Самим факторам может быть сообщен определенный смысл.

При помощи факторного анализа большое число переменных сводится к меньшему числу независимых влияющих величин, которые называются факторами.

Фактор в "сжатом" виде содержит информацию о нескольких переменных. В один фактор объединяются переменные, которые сильно коррелируют между собой. В результате факторного анализа отыскиваются такие комплексные факторы, которые как можно более полно объясняют связи между рассматриваемыми переменными.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

На первом шаге факторного анализа осуществляется стандартизация значений переменных. Факторный анализ опирается на гипотезу о том, что анализируемые переменные являются косвенными проявлениями сравнительно небольшого числа неких скрытых факторов.

Факторный анализ - это совокупность методов, ориентированных на выявление и анализ скрытых зависимостей между наблюдаемыми переменными. Скрытые зависимости также называют латентными.

Один из методов факторного анализа - метод главных компонент - основан на предположении о независимости факторов друг от друга.

Выбирая между иерархическими и неиерархическими методами, необходимо учитывать следующие их особенности.

Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации. Ценой, которую приходится платить за эти достоинства метода, является слово "априори". Необходимо заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации.

Если нет предположений относительно числа кластеров, рекомендуют использовать иерархические алгоритмы. Однако если объем выборки не позволяет это сделать, возможный путь - проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты. За счет такого "варьирования" результатов достигается достаточно большая гибкость кластеризации.

Иерархические методы, в отличие от неиерархических, отказываются от определения числа кластеров, а строят полное дерево вложенных кластеров.

Сложности иерархических методов кластеризации: ограничение объема набора данных; выбор меры близости; негибкость полученных классификаций.

Преимущество этой группы методов в сравнении с неиерархическими методами – их наглядность и возможность получить детальное представление о структуре данных.

При использовании иерархических методов существует возможность достаточно легко идентифицировать выбросы в наборе данных и, в результате, повысить качество данных.

Эта процедура лежит в основе двухшагового алгоритма кластеризации. Такой набор данных в дальнейшем может быть использован для проведения неиерархической кластеризации.

Иерархические методы не могут работать с большими наборами данных, а использование некоторой выборки, т.е. части данных, могло бы позволить применять эти методы.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Результаты кластеризации могут не иметь достаточного статистического обоснования. С другой стороны, при решении задач кластеризации допустима нестатистическая интерпретация полученных результатов, а также достаточно большое разнообразие вариантов понятия кластера. Такая нестатистическая интерпретация дает возможность аналитику получить удовлетворяющие его результаты кластеризации, что при использовании других методов часто бывает затруднительным.

В связи с появлением сверхбольших баз данных, появились новые требования, которым должен удовлетворять алгоритм кластеризации. Основное из них - это масштабируемость алгоритма (масштабируемость здесь означает, что с ростом объемов данных время, затрачиваемое на обучение, и кластеризацию, растет линейно).

Отметим также другие свойства, которым должен удовлетворять алгоритм кластеризации: независимость результатов от порядка входных данных; независимость параметров алгоритма от входных данных.

Разработаны алгоритмы, в которых методы иерархической кластеризации интегрированы с другими методами. К таким алгоритмам относятся: BIRCH, CURE, CHAMELEON, ROCK.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).

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.

Аналогия, описывающая этот алгоритм. Если каждый элемент данных представить себе как бусину, лежащую на поверхности стола, токластеры 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.

Алгоритм WaveCluster.

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

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

алгоритм применяет волновое преобразование к обобщенным данным. Главные особенности WaveCluster:

- сложность реализации;
- алгоритм может обнаруживать кластеры произвольных форм;
- алгоритм не чувствителен к шумам;
- алгоритм применим только к данным низкой размерности.

Алгоритм CLARA (Clustering LARge Applications).

Алгоритм CLARA извлекает множество образцов из базы данных. 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.

Алгоритмы Clarans, CURE, DBScan.

Алгоритм Clarans (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 Clarans 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, Clarans, CURE, DBScan является то, 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.

Методы поиска ассоциативных правил

Целью поиска ассоциативных правил (association rule) является нахождение закономерностей между связанными событиями в базах данных.

Впервые задача поиска ассоциативных правил (association rule mining) была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).

Рыночная корзина - это набор товаров, приобретенных покупателем в рамках одной отдельно взятой транзакции.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Транзакция - это множество событий, которые произошли одновременно. Регистрируя все бизнес-операции в течение всего времени своей деятельности, торговые компании накапливают огромные собрания транзакций. Каждая такая транзакция представляет собой набор товаров, купленных покупателем за один визит. Полученные в результате анализа шаблоны включают перечень товаров и число транзакций, которые содержат данные наборы. Транзакционная или операционная база данных (Transaction database) представляет собой двумерную таблицу, которая состоит из номера транзакции (TID) и перечня покупок, приобретенных во время этой транзакции.

TID - уникальный идентификатор, определяющий каждую сделку или транзакцию. Пример транзакционной базы данных, состоящей из покупательских транзакций, приведен на рис.25.

TID	Приобретенные покупки
100	Хлеб, молоко, печенье
200	Молоко, сметана
300	Молоко, хлеб, сметана, печенье
400	Колбаса, сметана
500	Хлеб, молоко, печенье, сметана

Рис.25 Пример транзакционной базы данных

В таблице первая колонка (TID) определяет номер транзакции, во второй колонке таблицы приведены товары, приобретенные во время определенной транзакции.

На основе имеющейся базы данных нужно найти закономерности между событиями, то есть покупками.

Часто встречающиеся шаблоны или образцы.

Допустим, имеется транзакционная база данных D. Присвоим значениям товаров переменные (рис.25).

Хлеб=a, Молоко=b, Печенье=c, Сметана=d, Колбаса=e, Конфеты=f.

Рассмотрим набор товаров (Itemset), включающий, например, {Хлеб, молоко, печенье}. Выразим этот набор с помощью переменных: abc={a,b,c}

Этот набор товаров встречается в нашей базе данных три раза, т.е. поддержка этого набора товаров равна 3:

$$SUP(abc)=3.$$

При минимальном уровне поддержки, равной трем, набор товаров abc является часто встречающимся шаблоном.

Поддержкой называют количество или процент транзакций, содержащих определенный набор данных. Для данного набора товаров

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

поддержка, выраженная в процентном отношении, равна 50%. Поддержку иногда также называют обеспечением набора.

Таким образом, набор представляет интерес, если его поддержка выше определенного пользователем минимального значения (min support). Эти наборы называют часто встречающимися (frequent).

Характеристики ассоциативных правил.

Ассоциативное правило имеет вид: "Из события А следует событие В".

В результате такого вида анализа мы устанавливаем закономерность следующего вида: "Если в таранзакции встретился набор товаров (или набор элементов) А, то можно сделать вывод, что в этой же транзакции должен появиться набор элементов В)". Установление таких закономерностей дает нам возможность находить очень простые и понятные правила, называемые ассоциативными.

Основными характеристиками ассоциативного правила являются поддержка и достоверность правила. Рассмотрим правило "из покупки молока следует покупка печенья" для данных рис.25.

Правило имеет поддержку s, если s% транзакций из всего набора содержат одновременно наборы элементов А и В или, другими словами, содержат оба товара.

Молоко - это товар А, печенье - это товар В. Поддержка правила "из покупки молока следует покупка печенья" равна 3, или 50%.

Достоверность правила показывает, какова вероятность того, что из события А следует событие В. Правило "Из А следует В" справедливо с достоверностью γ, если γ процентов транзакций из всего множества, содержащих набор элементов А, также содержат набор элементов В.

Число транзакций, содержащих молоко, равно четырем, число транзакций, содержащих печенье, равно трем, достоверность правила равна 75%.

Границы поддержки и достоверности ассоциативного правила.

При помощи использования алгоритмов поиска ассоциативных правил можно получить все возможные правила вида "Из А следует В", с различными значениями поддержки и достоверности. Однако в большинстве случаев, количество правил необходимо ограничивать заранее установленными минимальными и максимальными значениями поддержки и достоверности.

Если значение поддержки правила слишком велико, то в результате работы алгоритма будут найдены правила очевидные и хорошо известные. Слишком низкое значение поддержки приведет к нахождению очень большого количества правил, которые, возможно, будут в большей части необоснованными, но не известными и не очевидными. Таким образом, необходимо определить такой интервал, "золотую середину", который с одной стороны обеспечит нахождение неочевидных правил, а с другой - их обоснованность.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Если уровень достоверности слишком мал, то ценность правила вызывает серьезные сомнения. Например, правило с достоверностью в 3% только условно можно назвать правилом.

Методы поиска ассоциативных правил

Алгоритм AIS. В алгоритме AIS кандидаты множества наборов генерируются и подсчитываются "на лету", во время сканирования базы данных.

Алгоритм SETM. Как и алгоритм AIS, SETM также формирует кандидатов "на лету", основываясь на преобразованиях базы данных. Чтобы использовать стандартную операцию объединения языка SQL для формирования кандидата, SETM отделяет формирование кандидата 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. Работа данного алгоритма состоит из нескольких этапов, каждый из этапов состоит из следующих шагов:

- формирование кандидатов;
- подсчет кандидатов.

Формирование кандидатов (candidate generation) - этап, на котором алгоритм, сканируя базу данных, создает множество i-элементных кандидатов (i - номер этапа). На этом этапе поддержка кандидатов не рассчитывается.

Подсчет кандидатов (candidate counting) - этап, на котором вычисляется поддержка каждого i-элементного кандидата. Здесь же осуществляется отсечение кандидатов, поддержка которых меньше минимума, установленного пользователем (min_sup). Оставшиеся i-элементные наборы называем часто встречающимися.

Рассмотрим работу алгоритма Apriori на примере базы данных D. Иллюстрация работы алгоритма приведена на рис.26. Минимальный уровень поддержки равен 3.

На первом этапе происходит формирование одноэлементных кандидатов. Далее алгоритм подсчитывает поддержку одноэлементных наборов. Наборы с уровнем поддержки меньше установленного, то есть 3, отсекаются. В нашем примере это наборы e и f, которые имеют поддержку, равную 1. Оставшиеся наборы товаров считаются часто встречающимися одноэлементными наборами товаров: это наборы a, b, c, d.

Далее происходит формирование двухэлементных кандидатов, подсчет их поддержки и отсечение наборов с уровнем поддержки, меньшим 3. Оставшиеся двухэлементные наборы товаров, считающиеся часто встречающимися двухэлементными наборами ab, ac, bd, принимают участие в дальнейшей работе алгоритма.

Если смотреть на работу алгоритма прямолинейно, на последнем этапе алгоритм формирует трехэлементные наборы товаров: abc, abd, bcd, acd,

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

подсчитывает их поддержку и отсекает наборы с уровнем поддержки, меньшим 3. Набор товаров abc может быть назван часто встречающимся.

Однако алгоритм Apriori уменьшает количество кандидатов, отсекая - априори - тех, которые заведомо не могут стать часто встречающимися, на основе информации об отсеченных кандидатах на предыдущих этапах работы алгоритма.

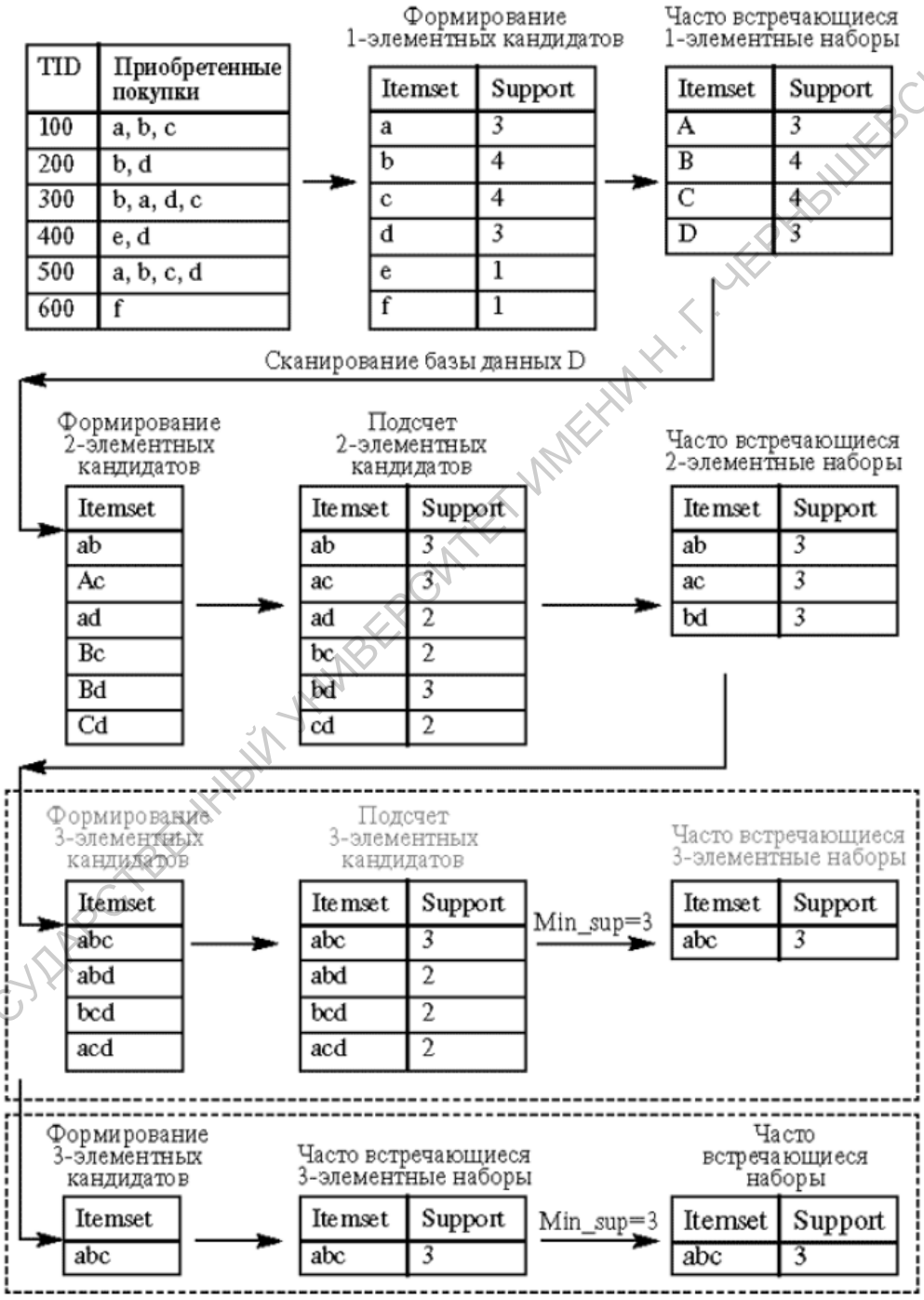

Рис.26 Алгоритм Apriori

Отсечение кандидатов происходит на основе предположения о том, что у часто встречающегося набора товаров все подмножества должны быть часто встречающимися.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Если в наборе находится подмножество, которое на предыдущем этапе было определено как нечасто встречающееся, этот кандидат уже не включается в формирование и подсчет кандидатов.

Так наборы товаров ad, bc, cd были отброшены как нечасто встречающиеся, алгоритм не рассматривал товаров abd, bcd, acd.

При рассмотрении этих наборов формирование трехэлементных кандидатов происходило бы по схеме, приведенной в верхнем пунктирном прямоугольнике. Поскольку алгоритм априори отбросил заведомо нечасто встречающиеся наборы, последний этап алгоритма сразу определил набор abc как единственный трехэлементный часто встречающийся набор (этап приведен в нижнем пунктирном прямоугольнике).

Алгоритм Apriori рассчитывает также поддержку наборов, которые не могут быть отсечены априори. Это так называемая негативная область (negative border), к ней принадлежат наборы-кандидаты, которые встречаются редко, их самих нельзя отнести к часто встречающимся, но все подмножества данных наборов являются часто встречающимися.

Разновидности алгоритма Apriori.

В зависимости от размера самого длинного часто встречающегося набора алгоритм Apriori сканирует базу данных определенное количество раз. Разновидности алгоритма Apriori, являющиеся его оптимизацией, предложены для сокращения количества сканирований базы данных, количества наборов-кандидатов или того и другого. Были предложены следующие разновидности алгоритма Apriori: AprioriTID и AprioriHybrid.

AprioriTid. Особенность этого алгоритма в том, что база данных D не используется для подсчета поддержки кандидатов набора товаров после первого прохода. С этой целью используется кодирование кандидатов, выполненное на предыдущих проходах. В последующих проходах размер закодированных наборов может быть намного меньше, чем база данных, и таким образом экономятся значительные ресурсы.

AprioriHybrid. Анализ времени работы алгоритмов Apriori и AprioriTid показывает, что в более ранних проходах Apriori добивается большего успеха, чем AprioriTid; однако AprioriTid работает лучше Apriori в более поздних проходах. Кроме того, они используют одну и ту же процедуру формирования наборов-кандидатов. Основанный на этом наблюдении, алгоритм AprioriHybrid предложен, чтобы объединить лучшие свойства алгоритмов Apriori и AprioriTid. AprioriHybrid использует алгоритм Apriori в начальных проходах и переходит к алгоритму AprioriTid, когда ожидается, что закодированный набор первоначального множества в конце прохода будет соответствовать возможностям памяти. Однако, переключение от Apriori до AprioriTid требует вовлечения дополнительных ресурсов.

Алгоритм DHP, также называемый алгоритмом хеширования. В основе его работы - вероятностный подсчет наборов-кандидатов, осуществляемый для сокращения числа подсчитываемых кандидатов на каждом этапе выполнения алгоритма Apriori. Сокращение обеспечивается за счет того, что

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

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-кандидат, если его значение в хеш-таблице меньше порогового значения, установленного для обеспечения.

Алгоритм PARTITION. 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). Алгоритм разбивает базу данных на несколько блоков, каждый из которых отмечается так называемыми "начальными точками" (start point), и затем циклически сканирует базу данных.

Методы визуализации

Визуализация традиционно рассматривалась как вспомогательное средство при анализе данных, однако сейчас все больше исследований говорит о ее самостоятельной роли.

Традиционные методы визуализации могут находить следующее применение:

- представлять пользователю информацию в наглядном виде;
- компактно описывать закономерности, присущие исходному набору данных;
- снижать размерность или сжимать информацию;
- восстанавливать пробелы в наборе данных;
- находить шумы и выбросы в наборе данных.

Визуализация инструментов Data Mining

Каждый из алгоритмов Data Mining использует определенный подход к визуализации. В ходе использования каждого из методов Data Mining, а точнее, его программной реализации, получают некие визуализаторы, при помощи которых удается интерпретировать результаты, полученные в результате работы соответствующих методов и алгоритмов.

Для деревьев решений это визуализатор дерева решений, список правил, таблица сопряженности.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Для нейронных сетей в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.

Для карт Кохонена: карты входов, выходов, другие специфические карты.

Для линейной регрессии в качестве визуализатора выступает линия регрессии.

Для кластеризации: дендрограммы, диаграммы рассеивания.

Диаграммы и графики рассеивания часто используются для оценки качества работы того или иного метода.

Все эти способы визуального представления или отображения данных могут выполнять одну из функций:

- являются иллюстрацией построения модели (например, представление структуры (графа) нейронной сети);
- помогают интерпретировать полученный результат;
- являются средством оценки качества построенной модели;
- сочетают перечисленные выше функции (дерево решений, дендрограмма).

Визуализация моделей Data Mining

Первая функция (иллюстрация построения модели), по сути, является визуализацией Data Mining модели. Существует много различных способов представления моделей, но графическое ее представление дает пользователю максимальную "ценность". Модель Data Mining должна быть представлена на естественном языке или, хотя бы, содержать минимальное количество различных математических и технических элементов.

Таким образом, доступность является одной из основных характеристик модели Data Mining. Несмотря на это, существует и такой распространенный и наиболее простой способ представления модели, как "черный ящик". В этом случае пользователь не понимает поведения той модели, которой пользуется. Однако, несмотря на непонимание, он получает результат - выявленные закономерности. Классическим примером такой модели является модель нейронной сети.

Другой способ представления модели - представление ее в интуитивном, понятном виде. Понимание модели ведет к пониманию ее содержания. Классическим примером является дерево решений.

Кроме понимания, такие модели обеспечивают возможность взаимодействовать с моделью, задавать ей вопросы и получать ответы. Примером такого взаимодействия является средство "что, если". При помощи диалога "система-пользователь" можно получить понимание модели.

Функции, которые помогают интерпретировать и оценить результаты построения Data Mining моделей, это всевозможные графики, диаграммы, таблицы, списки и т.д.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Примерами средств визуализации, при помощи которых можно оценить качество модели, являются диаграмма рассеивания, таблица сопряженности, график изменения величины ошибки.

Диаграмма рассеивания представляет собой график отклонения значений, прогнозируемых при помощи модели, от реальных. Эти диаграммы используют для непрерывных величин. Визуальная оценка качества построенной модели возможна только по окончанию процесса построения модели.

Таблица сопряженности используется для оценки результатов классификации. Такие таблицы применяются для различных методов классификации. Оценка качества построенной модели возможно только по окончанию процесса построения модели.

График изменения величины ошибки. График демонстрирует изменение величины ошибки в процессе работы модели. Например, в процессе работы нейронных сетей пользователь может наблюдать за изменением ошибки на обучающем и тестовом множествах и остановить обучение для недопущения "переобучения" сети. Здесь оценка качества модели и его изменения может оцениваться непосредственно в процессе построения модели.

Примерами средств визуализации, которые помогают интерпретировать результат, являются: линия тренда в линейной регрессии, карты Кохонена, диаграмма рассеивания в кластерном анализе.

Методы визуализации

Методы визуализации, в зависимости от количества используемых измерений, принято разделять на две группы:

- представление данных в одном, двух и трех измерениях;
- представление данных в четырех и более измерениях.

Представление данных в одном, двух и трех измерениях.

К этой группе методов относятся хорошо известные способы отображения информации, которые доступны для восприятия человеческим воображением. Практически любой современный инструмент Data Mining включает способы визуального представления из этой группы.

В соответствии с количеством измерений это могут быть следующие способы:

- одномерное (univariate) измерение, или 1-D;
- двумерное (bivariate) измерение, или 2-D;
- трехмерное или проекционное (projection) измерение, или 3-D.

При использовании двух- и трехмерного представления информации пользователь имеет возможность увидеть закономерности набора данных:

- кластерную структуру и распределение объектов на классы (например, на диаграмме рассеивания);
- топологические особенности;
- наличие трендов;

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

- информацию о взаимном расположении данных;
- существование других зависимостей, присущих исследуемому набору данных.

Если набор данных имеет более трех измерений, то возможны следующие варианты:

- использование многомерных методов представления информации (они рассмотрены ниже);
- снижение размерности до одно-, двух- или трехмерного представления.

Существуют различные способы снижения размерности, один из них - факторный анализ. Для снижения размерности и одновременного визуального представления информации на двумерной карте используются самоорганизующиеся карты Кохонена.

Представление данных в 4 + измерениях.

Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации.

Наиболее известные способы многомерного представления информации:

- параллельные координаты;
- "лица Чернова";
- лепестковые диаграммы.

Параллельные координаты.

В параллельных координатах переменные кодируются по горизонтали, вертикальная линия определяет значение переменной. Пример набора данных, представленного в декартовых координатах и параллельных координатах, дан на рис.27.

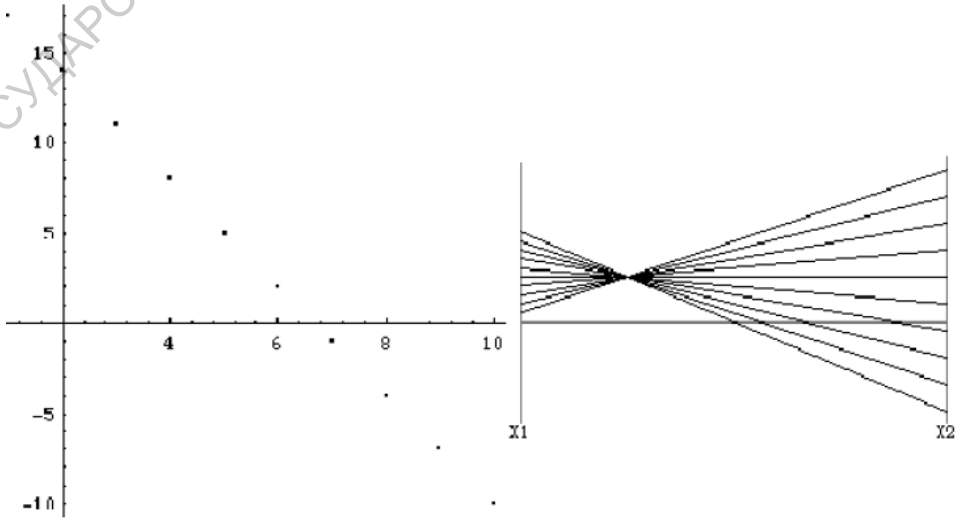

Рис.27 Набор данных в декартовых координатах и в параллельных координатах

"Лица Чернова".

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Основная идея представления информации в "лицах Чернова" состоит в кодировании значений различных переменных в характеристиках или чертах человеческого лица. Пример такого "лица" приведен на рис.28.

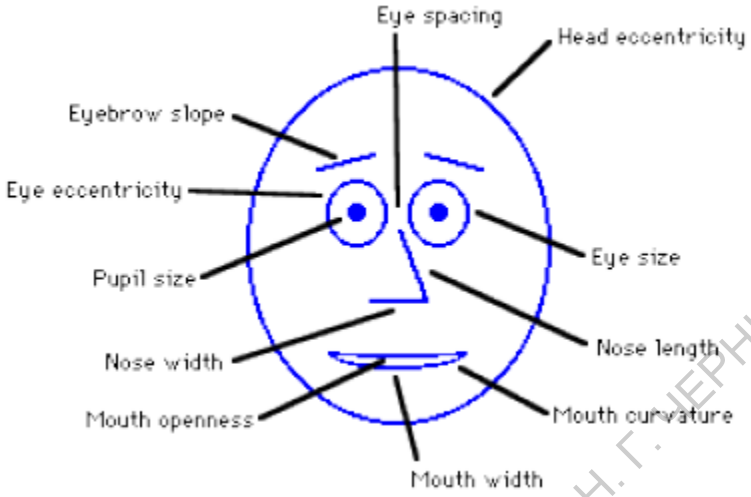

Рис.28 "Лицо Чернова"

Для каждого наблюдения рисуется отдельное "лицо". На каждом "лице" относительные значения переменных представлены как формы и размеры отдельных черт лица (например, длина и ширина носа, размер глаз, размер зрачка, угол между бровями). Анализ информации при помощи такого способа отображения основан на способности человека интуитивно находить сходства и различия в чертах лица.

Перед использованием методов визуализации необходимо: проанализировать, следует ли изображать все данные или же какую-то их часть; выбрать размеры, пропорции и масштаб изображения; выбрать метод, который может наиболее ярко отобразить закономерности, присущие набору данных.

Среди двухмерных и трехмерных средств наиболее широко известны линейные графики, линейные, столбиковые, круговые секторные и векторные диаграммы.

При помощи **линейного графика** можно отобразить тенденцию, передать изменения какого-либо признака во времени. Для сравнения нескольких рядов чисел такие графики наносятся на одни и те же оси координат.

Гистограмму применяют для сравнения значений в течение некоторого периода или же соотношения величин.

Круговые диаграммы используют, если необходимо отобразить соотношение частей и целого, т.е. для анализа состава или структуры явлений. Составные части целого изображаются секторами окружности. Секторы рекомендуют размещать по их величине: вверху - самый крупный, остальные - по движению часовой стрелки в порядке уменьшения их величины. Круговые диаграммы также применяют для отображения результатов факторного анализа, если действия всех факторов являются однонаправленными. При этом каждый фактор отображается в виде одного из секторов круга.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Выбор того или иного средства визуализации зависит от поставленной задачи (например, нужно определить структуру данных или же динамику процесса) и от характера набора данных.

Качество визуализации.

Современные аналитические средства, в том числе и Data Mining, немыслимы без качественной визуализации. В результате использования средств визуализации должны быть получены наглядные и выразительные, ясные и простые изображения, за счет использования разнообразных средств: цвета, контраста, границ, пропорций, масштаба и т.д.

В связи с ростом требований к средствам визуализации, а также необходимости сравнивания их между собой, в последние годы был сформирован ряд принципов качественного визуального представления информации.

Принципы Тафта (Tufte's Principles) графического представления данных высокого качества гласят:

- предоставляйте пользователю самое большое количество идей, в самое короткое время, с наименьшим количеством чернил на наименьшем пространстве;
- говорите правду о данных.

Основные принципы компоновки визуальных средств представления информации можно сформулировать следующим образом:

1. Принцип лаконичности.
2. Принцип обобщения и унификации.
3. Принцип акцента на основных смысловых элементах.
4. Принцип автономности.
5. Принцип структурности.
6. Принцип стадийности.
7. Принцип использования привычных ассоциаций и стереотипов.

Принцип лаконичности говорит о том, что средство визуализации должно содержать лишь те элементы, которые необходимы для сообщения пользователю существенной информации, точного понимания ее значения или принятия (с вероятностью не ниже допустимой величины) соответствующего оптимального решения.

Кроме того средство визуализации должно обладать высокой надежностью и скоростью.

Представление пространственных характеристик.

Отдельным направлением визуализации является наглядное представление пространственных характеристик объектов. В большинстве случаев такие средства выделяют на карте отдельные регионы и обозначают их различными цветами в зависимости от значения анализируемого показателя. На рис.29 приведен пример такой визуализации. Карта представлена в виде графического интерфейса, отображающего данные в виде трехмерного ландшафта произвольно определенных и позиционированных форм (столбчатых диаграмм, каждая с индивидуальными высотой и цветом).

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО

Такой способ позволяет наглядно показывать количественные и реляционные характеристики пространственно-ориентированных данных и быстро идентифицировать в них тренды.

Рис.29 Ландшафтный визуализатор

Литература и программные ресурсы.

1. Паклин Н.Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009.
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. 2-е изд.– СПб.: БХВ – Петербург, 2008.
3. Кацко И.А., Н.Б. Паклин. Практикум по анализу данных на компьютере. – М.: КолосС, 2009.
4. Дюк В.А., Самойленко А.П. Data Mining: учебный курс. – СПб.: Питер, 2001.
5. Хайкин С. Нейронные сети: полный курс. 2-е. изд. / Пер. с англ. – М.: Издательский дом «Вильямс», 2006.
6. Ханк Д.Э., Уичерн Д.У., Райте А.Дж. Бизнес-прогнозирование. 7-е изд. / Пер. с англ. – М.: Издательский дом «Вильямс», 2003.
7. Дубров А. М., Мхитарян В. С., Трошин Л. И. Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2000.
8. Андрейчиков А.В., Андрейчикова О.Н. Интеллектуальные информационные системы. – М.: ФиС, 2004.
9. Кравченко Т.К., Перминов Г.И. Информационная технология процесса принятия экономических решений. - М.: ГУ-ВШЭ., 2005.
10. http://portal.tpu.ru:7777/SHARED/a/AAPONOMAREV/metod/Tab/lections%20data%20mining.pdf.
11. http://www.basegroup.ru/deductor/.
12. http://www.mathworks.com/.
13. http://www.gnumeric.org/.

САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО