Лекция 1

Data Mining – технология добычи данных

Технология Data Mining

- Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с Data Mining встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in data bases) и "интеллектуальный анализ данных". Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных.
- До начала 90-х годов, людям, не имевшем представления о распознавании образов и факторном анализе, казалось, не было особой нужды переосмысливать ситуацию в этой области. Все шло своим чередом в рамках направления, называемого прикладной статистикой. Теоретики проводили конференции и семинары, писали внушительные статьи и монографии, изобиловавшие аналитическими выкладками.
- Вместе с тем, практики всегда знали, что попытки применить теоретические разработки для решения реальных задач в большинстве случаев оказываются бесплодными. Но на озабоченность практиков до поры до времени можно было не обращать особого внимания — они решали главным образом свои частные проблемы обработки небольших локальных баз данных.

- В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информационной руды в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Что делать с этой информацией? Стало ясно, что без продуктивной переработки потоки сырых данных образуют никому не нужную свалку.
- Специфика современных требований к такой переработке следующие:
- Данные имеют неограниченный объем.
- Данные являются разнородными (количественными, качественными, текстовыми).
- Результаты обработки должны быть конкретны и понятны.
- Инструменты для обработки сырых данных должны быть просты в использовании.

- Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. Главная причина — *концепция усреднения по выборке*, приводящая к операциям над фиктивными величинами (типа средней температуры пациентов по больнице, средней высоты дома на улице, состоящей из дворцов и лачуг и т.п.). Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для “грубого” разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing, OLAP).
- В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих *фрагменты* многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные *подвыборкам данных*, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборке и виде распределений значений анализируемых показателей.

Таблица - Примеры формулировок задач при использовании методов OLAP и Data Mining

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)?	Имеются ли характерные черты клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?

Важное положение Data Mining — нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). Таким образом пришло понимание, что сырые данные (raw data) содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки.

- **Рисунок 1. Уровни знаний, извлекаемых из данных**

Литература

- 1. А.А. Барсегян «Методы и модели анализа данных: OLAP и Data Mining», Санкт-Петербург, изд-во БХВ-Петрбург, 2004 г. Книга – обзор технологий обработки данных, первая на русском языке.
- 2. Р.Г.Степанов. Технология Data Mining: Интеллектуальный Анализ Данных; 2008
- 3. И.А.Чубукова. Data Mining; 2008
- 4. Р.Гонсалес.Принципы распознавания образов Дж. Ту,; 1978

Определение Data Mining

- В целом технологию Data Mining достаточно точно определяет Григорий Пиатецкий-Шапиро — один из основателей этого направления:
- **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.
- Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей. 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. (Григорий Пиатецкий-Шапиро)

- Традиционные методы анализа данных (статистические методы) и OLAP в основном ориентированы на проверку заранее сформулированных гипотез (verification-driven data mining) и на "грубый" разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как одно из основных положений Data Mining - поиск неочевидных закономерностей. Инструменты Data Mining могут находить такие закономерности самостоятельно и также самостоятельно строить гипотезы о взаимосвязях. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, преимущество Data Mining по сравнению с другими методами анализа является очевидным.

Двухмерная таблица "объект-атрибут"

	Атрибуты				
	Код клиента	Возраст	Семейное положение	Доход	Класс
Объекты	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

в браке

разведенный

Основные понятия

- **Данные** - это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.
- **Объект** описывается как набор атрибутов. Объект также известен как запись, случай, пример, строка таблицы и т.д.
- **Атрибут** - свойство, характеризующее объект. Например: цвет глаз человека, температура воды и т.д. Атрибут также называют переменной, полем таблицы, измерением, характеристикой.
- **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.
- **Параметры** - числовые характеристики генеральной совокупности.
- **Статистики** - числовые характеристики выборки.
- **Гипотеза** - частично обоснованная закономерность знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

- **Измерение** - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.
- В процессе подготовки данных измеряется не сам объект, а его характеристики.
- **Шкала** - правило, в соответствии с которым объектам присваиваются числа. Существует пять типов шкал измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

Атрибуты

Многие инструменты Data Mining при импорте данных из других источников предлагают выбрать тип шкалы для каждой переменной и/или выбрать тип данных для входных и выходных переменных (символьные, числовые, дискретные и непрерывные). Пользователю такого инструмента необходимо владеть этими понятиями.

- Атрибуты (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 конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.
- Пример дискретных данных. Продолжительность маршрута троллейбуса (количество вариантов продолжительности конечно): 10, 15, 25 мин.
- **Непрерывные данные** - данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность.
- Пример непрерывных данных: температура, высота, вес, длина и т.д.

Шкалы

- Шкала - правило, в соответствии с которым объектам присваиваются числа.
- Существует пять типов шкал 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.**
- Номинальная шкала состоит из названий, категорий, имен для классификации и сортировки объектов или наблюдений по некоторому признаку.
- Пример такой шкалы: профессии, город проживания, семейное положение.
- Для этой шкалы применимы только такие операции: равно (=), не равно ($\neq$).

- **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".
- Пример такой шкалы: место (1, 2, 3-е),которое команда получила на соревнованиях, номер студента в рейтинге успеваемости(1-й,23-й,и т.д.), при этом неизвестно, насколько один студент успешней другого, известен лишь его номер в рейтинге.
- Для этой шкалы применимы только такие операции: равно (=), не равно ($\neq$), 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.
- Пример такой шкалы: температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1,26 раз выше.
- Номинальная и порядковая шкалы являются дискретными, а интервальная шкала - непрерывной, она позволяет осуществлять точные измерения признака и производить арифметические операции сложения, вычитания, умножения, деления.
- Для этой шкалы применимы только такие операции: равно (=), не равно ($\neq$), больше (>), меньше (<), операции сложения (+) и вычитания(-).

- **Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.**
- Пример такой шкалы: вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее.
- Цена на картофель в супермаркете выше в 1,2 раза, чем цена на базаре.
- Относительные и интервальные шкалы являются числовыми.
- Для этой шкалы применимы такие операции: равно (=), не равно ( ), больше (>), меньше (<), операции сложения (+) и вычитания(-),умножения (*) и деления (/).

- **Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории.**
- Пример такой шкалы: пол (мужской и женский).
- Пример использования разных шкал для измерений свойств различных объектов, в данном случае температурных условий, приведен в таблице данных

- Таблица - Множество измерений свойств различных объектов

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

Задачи анализа данных

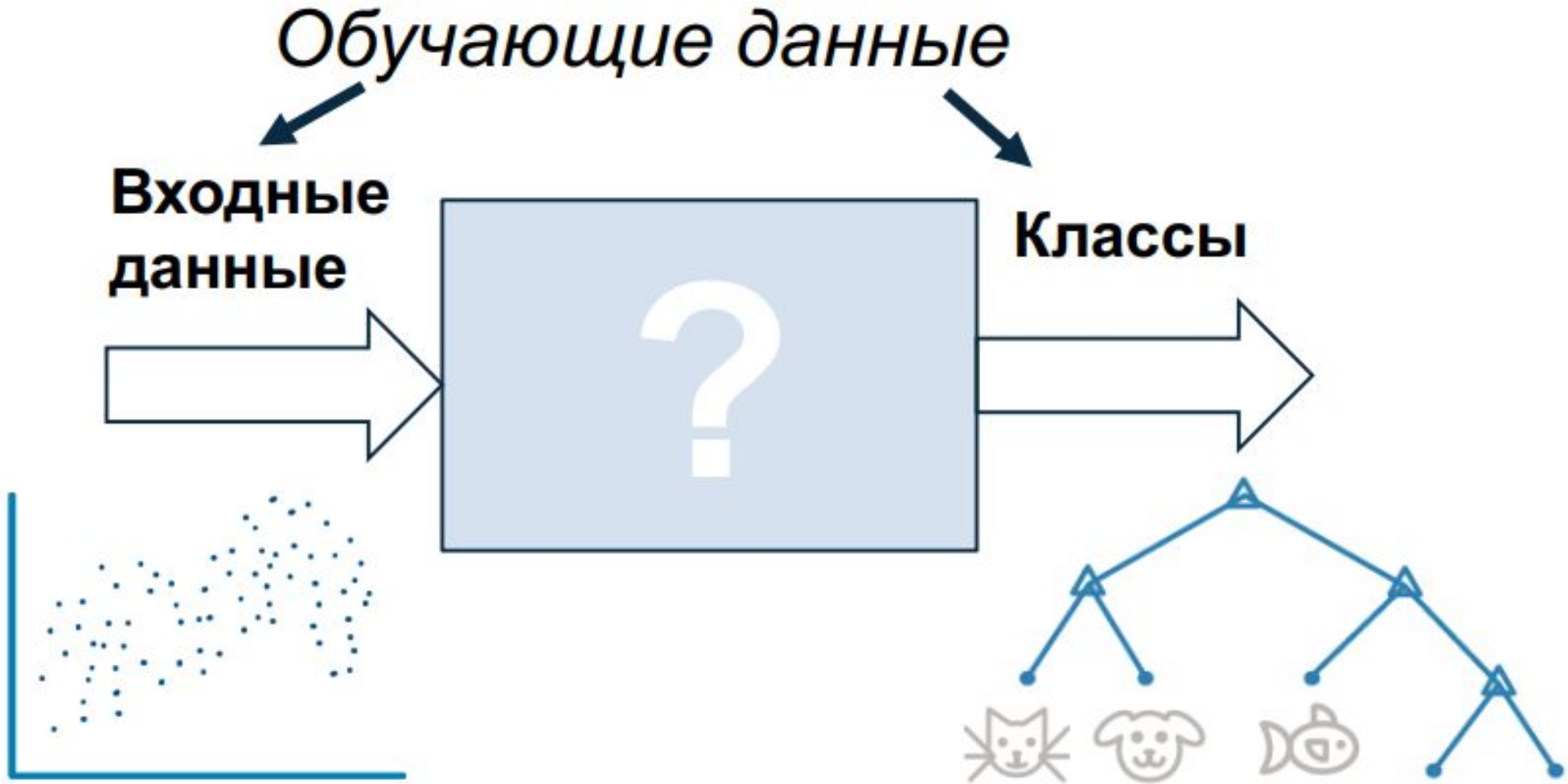
- **Классификация** (Classification) Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу. Методы решения. 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.

- **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, Секвенциальный анализ). 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. Пример. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (Customer Lifecycle Management).

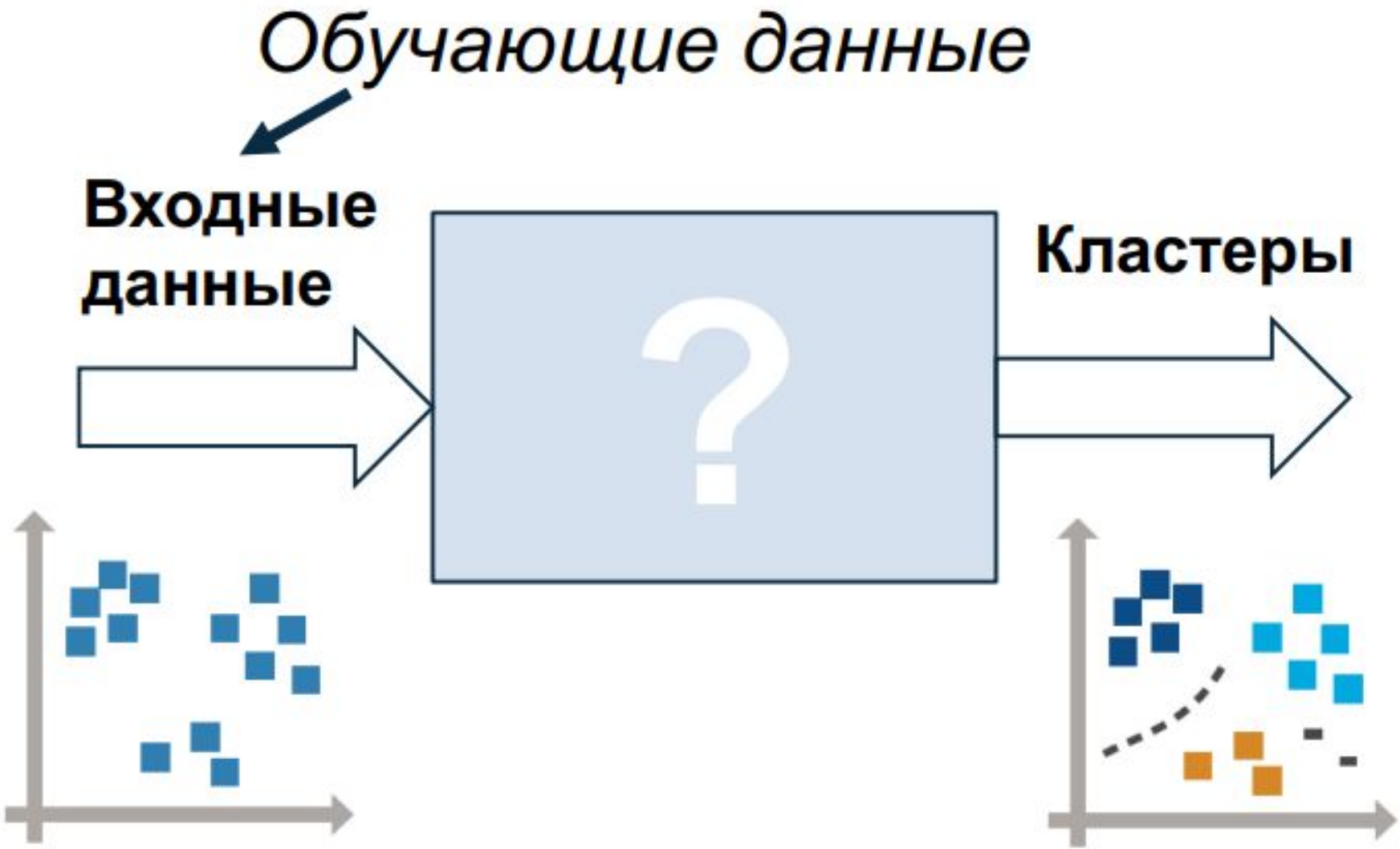
- **Прогнозирование** (Forecasting) В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.
- **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. Частный случай оценивания – регрессионный анализ.
- **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.
- Категория обучение с учителем представлена следующими задачами Data Mining: классификация, оценка, прогнозирование.
- Категория обучение без учителя представлена задачей кластеризации.

Алгоритмы машинного обучения

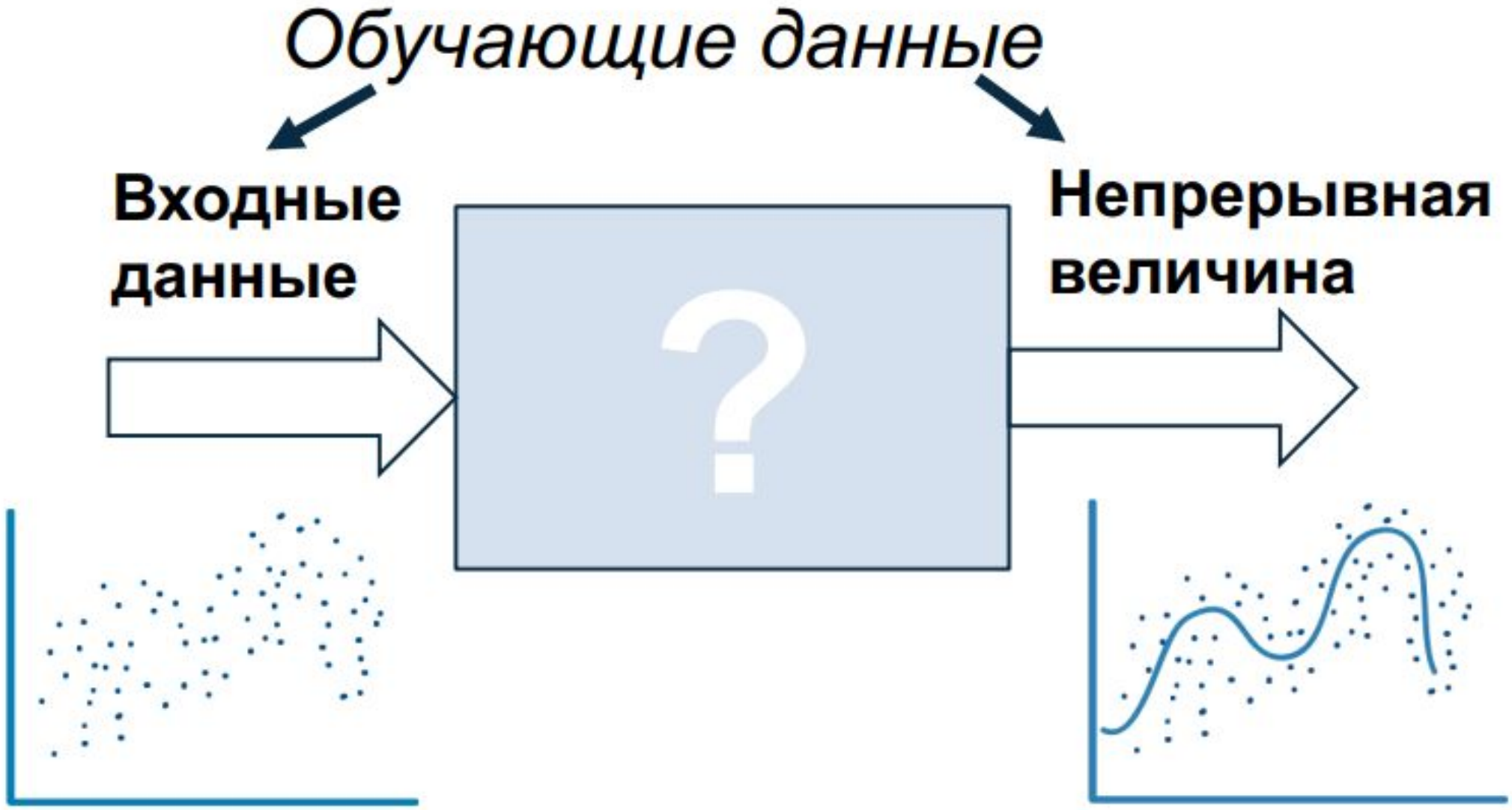
Классификация



Кластеризация



Регрессия



Сфера применения Data Mining

- Сфера применения Data Mining ничем не ограничена — она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня, мягко говоря, заинтриговали коммерческие предприятия, развертывающие проекты на основе информационных хранилищ данных (Data Warehousing). Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Например, известны сообщения об экономическом эффекте, в 10–70 раз превысившем первоначальные затраты от 350 до 750 тыс. дол. Известны сведения о проекте в 20 млн. дол., который окупился всего за 4 месяца. Другой пример — годовая экономия 700 тыс. дол. за счет внедрения Data Mining в сети универсамов в Великобритании.

- **Некоторые бизнес-приложения Data Mining**

- Розничная торговля
- Банковское дело
- Телекоммуникации
- Страхование
- Другие приложения в бизнесе

- **Розничная торговля**
- Предприятия розничной торговли сегодня собирают подробную информацию о каждой отдельной покупке, используя кредитные карточки с маркой магазина и компьютеризованные системы контроля. Вот типичные задачи, которые можно решать с помощью Data Mining в сфере розничной торговли:
- *анализ покупательской корзины* (анализ сходства) предназначен для выявления товаров, которые покупатели стремятся приобретать вместе. Знание покупательской корзины необходимо для улучшения рекламы, выработки стратегии создания запасов товаров и способов их раскладки в торговых залах.
- *исследование временных шаблонов* помогает торговым предприятиям принимать решения о создании товарных запасов. Оно дает ответы на вопросы типа "Если сегодня покупатель приобрел видеокамеру, то через какое время он вероятнее всего купит новые батарейки и пленку?"
- *создание прогнозирующих моделей* дает возможность торговым предприятиям узнавать характер потребностей различных категорий клиентов с определенным поведением, например, покупающих товары известных дизайнеров или посещающих распродажи. Эти знания нужны для разработки точно направленных, экономичных мероприятий по продвижению товаров.

- **Банковское дело**
- Достижения технологии Data Mining используются в банковском деле для решения следующих распространенных задач:
- *выявление мошенничества с кредитными карточками*. Путем анализа прошлых транзакций, которые впоследствии оказались мошенническими, банк выявляет некоторые стереотипы такого мошенничества.
- *сегментация клиентов*. Разбивая клиентов на различные категории, банки делают свою маркетинговую политику более целенаправленной и результативной, предлагая различные виды услуг разным группам клиентов.
- *прогнозирование изменений клиентуры*. Data Mining помогает банкам строить прогнозные модели ценности своих клиентов, и соответствующим образом обслуживать каждую категорию.

- **Телекоммуникации**
- В области телекоммуникаций методы Data Mining помогают компаниям более энергично продвигать свои программы маркетинга и ценообразования, чтобы удерживать существующих клиентов и привлекать новых. Среди типичных мероприятий отметим следующие:
- *анализ записей о подробных характеристиках вызовов.* Назначение такого анализа — выявление категорий клиентов с похожими стереотипами пользования их услугами и разработка привлекательных наборов цен и услуг;
- *выявление лояльности клиентов*. Data Mining можно использовать для определения характеристик клиентов, которые, один раз воспользовавшись услугами данной компании, с большой долей вероятности останутся ей верными. В итоге средства, выделяемые на маркетинг, можно тратить там, где отдача больше всего.

- **Страхование**
- Страховые компании в течение ряда лет накапливают большие объемы данных. Здесь обширное поле деятельности для методов Data Mining:
- *выявление мошенничества*. Страховые компании могут снизить уровень мошенничества, отыскивая определенные стереотипы в заявлениях о выплате страхового возмещения, характеризующих взаимоотношения между юристами, врачами и заявителями.
- *анализ риска*. Путем выявления сочетаний факторов, связанных с оплаченными заявлениями, страховщики могут уменьшить свои потери по обязательствам. Известен случай, когда в США крупная страховая компания обнаружила, что суммы, выплаченные по заявлениям людей, состоящих в браке, вдвое превышает суммы по заявлениям одиноких людей. Компания отреагировала на это новое знание пересмотром своей общей политики предоставления скидок семейным клиентам

Типы закономерностей

- Выделяют пять стандартных типов закономерностей, которые позволяют выявлять методы Data Mining: ассоциация, последовательность, классификация, кластеризация и прогнозирование (рис. 2).

Рисунок 2. Типы закономерностей, выявляемых методами Data Mining

Ассоциация имеет место в том случае, если несколько событий связаны друг с другом. Например, исследование, проведенное в супермаркете, может показать, что 65% купивших кукурузные чипсы берут также и "кока-колу", а при наличии скидки за такой комплект "колу" приобретают в 85% случаев. Располагая сведениями о подобной ассоциации, менеджерам легко оценить, насколько действенна предоставляемая скидка.

Если существует цепочка связанных во времени событий, то говорят о *последовательности*. Так, например, после покупки дома в 45% случаев в течение месяца приобретается и новая кухонная плита, а в пределах двух недель 60% новоселов обзаводятся холодильником.

- С помощью *классификации* выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил.
- *Кластеризация* отличается от классификации тем, что сами группы заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.
- Основой для всевозможных систем *прогнозирования* служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить найти шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

Классы систем Data Mining

- Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. (рис. 3). Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining. Многие из таких систем интегрируют в себе сразу несколько подходов. Тем не менее, как правило, в каждой системе имеется какая-то ключевая компонента, на которую делается главная ставка.

Рисунок 3. Data Mining — мультидисциплинарная область

Предметно-ориентированные аналитические системы

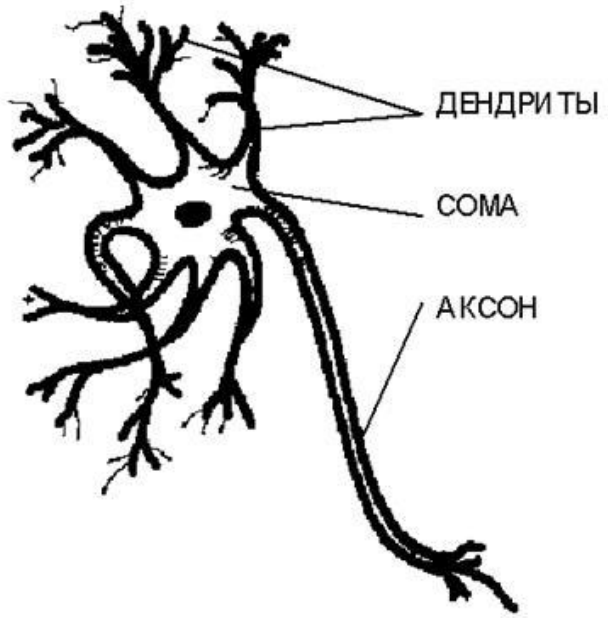
- Предметно-ориентированные аналитические системы очень разнообразны. Наиболее широкий подкласс таких систем, получивший распространение в области исследования финансовых рынков, носит название "технический анализ". Он представляет собой совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля, основанных на различных эмпирических моделях динамики рынка. Эти методы часто используют несложный статистический аппарат, но максимально учитывают сложившуюся своей области специфику (профессиональный язык, системы различных индексов и пр.). На рынке имеется множество программ этого класса. Как правило, они довольно дешевы (обычно $300–1000).

Статистические пакеты

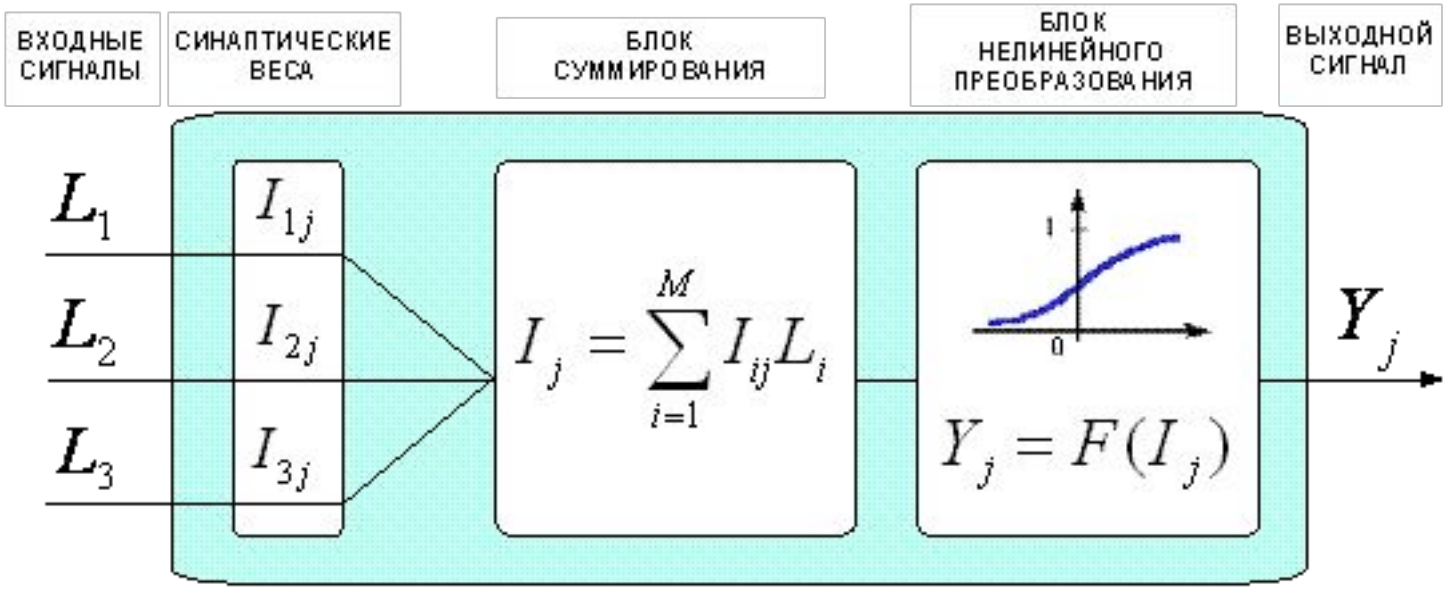
- Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. Но основное внимание в них уделяется все же классическим методикам — корреляционному, регрессионному, факторному анализу и другим.
- В качестве примеров наиболее мощных и распространенных статистических пакетов можно назвать SAS (компания SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA и другие.
- Недостатком систем этого класса считают требование к специальной подготовке пользователя. Также отмечают, что мощные современные статистические пакеты являются слишком "тяжеловесными" для массового применения в финансах и бизнесе. К тому же часто эти системы весьма дороги — от $1000 до $15000.
- Есть еще более серьезный принципиальный недостаток статистических пакетов, ограничивающий их применение в Data Mining. Большинство методов, входящих в состав пакетов опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А эти характеристики, как указывалось выше, при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами.

Нейронные сети

- Это большой класс систем, архитектура которых имеет аналогию с построением нервной ткани из нейронов. В одной из наиболее распространенных архитектур, многослойном перцептроне с обратным распространением ошибки, имитируется работа нейронов в составе иерархической сети, где каждый нейрон более высокого уровня соединен своими входами с выходами нейронов нижележащего слоя. На нейроны самого нижнего слоя подаются значения входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации и т. д. Эти значения рассматриваются как сигналы, передающиеся в следующий слой, ослабляясь или усиливаясь в зависимости от числовых значений (весов), приписываемых межнейронным связям. В результате на выходе нейрона самого верхнего слоя вырабатывается некоторое значение, которое рассматривается как ответ — реакция всей сети на введенные значения входных параметров. Для того чтобы сеть можно было применять в дальнейшем, ее прежде надо "натренировать" на полученных ранее данных, для которых известны и значения входных параметров, и правильные ответы на них. Тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам.

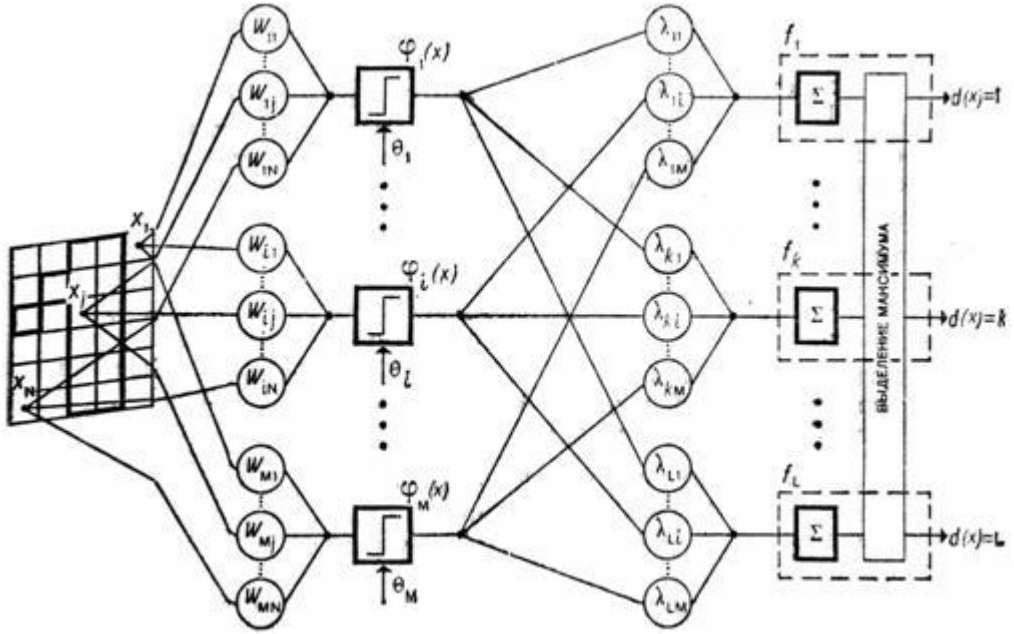


Структура биологического нейрона



Классическая модель нейрона Дж. Маккалоки и У. Питта

1943 году Дж. Маккалоки и У. Питт предложили формальную модель биологического нейрона как устройства, имеющего несколько входов (входные синапсы – дендриты), и один выход (выходной синапс – аксон)



- Рисунок 5. Нейросеть, реализующая двух-слойный персептрон

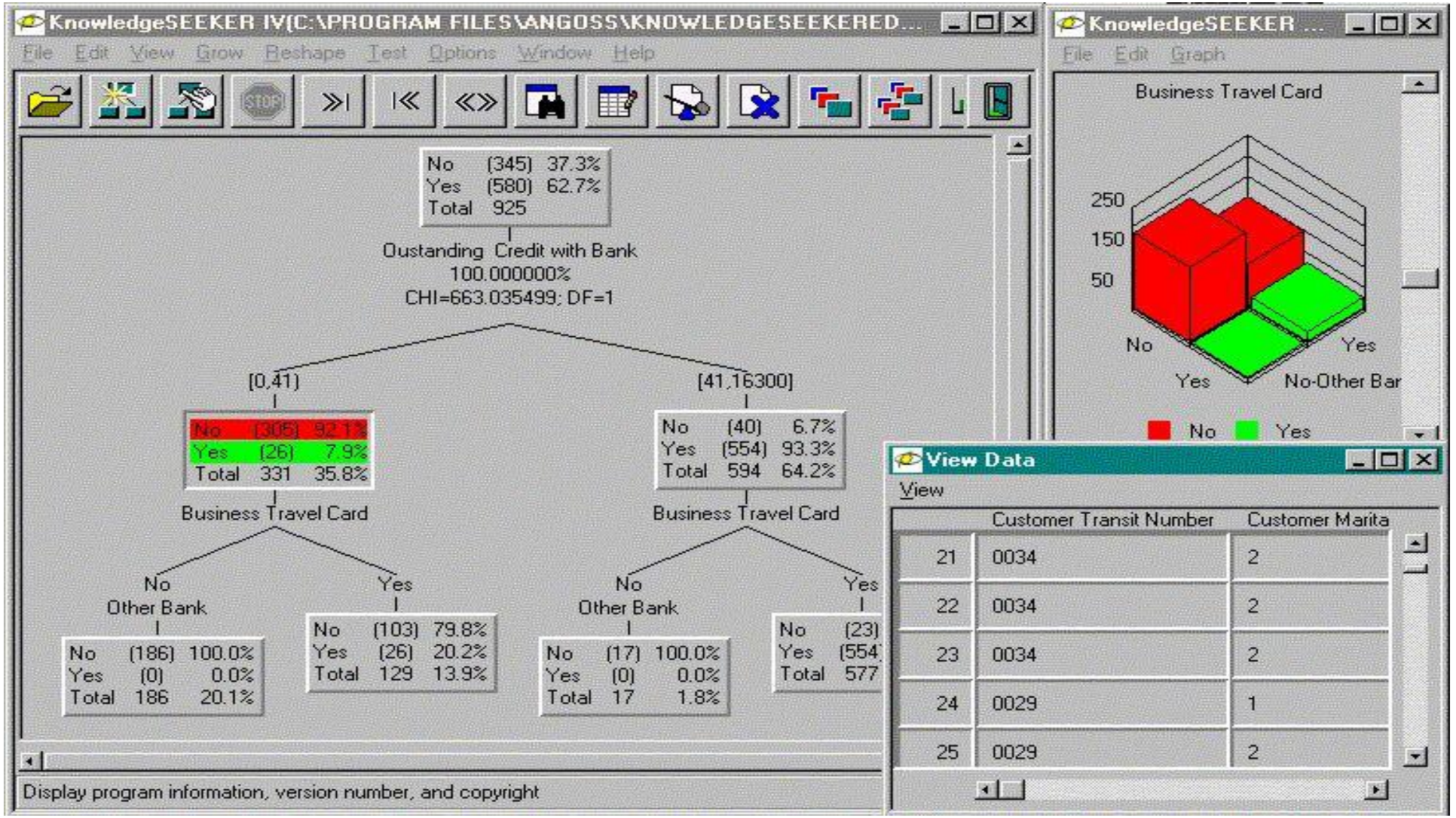
- Основным недостатком нейросетевой парадигмы является необходимость иметь очень большой объем обучающей выборки. Другой существенный недостаток заключается в том, что даже натренированная нейронная сеть представляет собой черный ящик. Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком (известные попытки дать интерпретацию структуре настроенной нейросети выглядят неубедительными – система “KINOsuite-PR”).
- Примеры нейросетевых систем — BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic). Стоимость их довольно значительна: $1500–8000.

Системы рассуждений на основе аналогичных случаев

- Идея систем case based reasoning — CBR — на первый взгляд крайне проста. Для того чтобы сделать прогноз на будущее или выбрать правильное решение, эти системы находят в прошлом близкие аналоги наличной ситуации и выбирают тот же ответ, который был для них правильным. Поэтому этот метод еще называют методом "ближайшего соседа" (nearest neighbour). В последнее время распространение получил также термин memory based reasoning, который акцентирует внимание, что решение принимается на основании всей информации, накопленной в памяти.
- Системы CBR показывают неплохие результаты в самых разнообразных задачах. Главным их минусом считают то, что они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт, — в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов CBR системы строят свои ответы.
- Другой минус заключается в произволе, который допускают системы CBR при выборе меры "близости". От этой меры самым решительным образом зависит объем множества прецедентов, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза.
- Примеры систем, использующих CBR, — KATE tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, США).

Деревья решений (decision trees)

- Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа "ЕСЛИ... ТО..." (if-then), имеющую вид дерева.



- Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид "значение параметра A больше x?". Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный — то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.
- Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить “лучшие” (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и “цепляют” фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода.
- Вместе с тем, большинство систем используют именно этот метод. Самыми известными являются See5/C5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Стоимость этих систем варьируется от 1 до 10 тыс. долл.

Генетические алгоритмы

- Data Mining не основная область применения генетических алгоритмов. Их нужно рассматривать скорее как мощное средство решения разнообразных комбинаторных задач и задач оптимизации. Тем не менее генетические алгоритмы вошли сейчас в стандартный инструментарий методов Data Mining.
- Первый шаг при построении генетических алгоритмов — это кодировка исходных логических закономерностей в базе данных, которые именуют хромосомами, а весь набор таких закономерностей называют популяцией хромосом. Далее для реализации концепции отбора вводится способ сопоставления различных хромосом. Популяция обрабатывается с помощью процедур репродукции, изменчивости (мутаций), генетической композиции. Эти процедуры имитируют биологические процессы. Наиболее важные среди них: случайные мутации данных в индивидуальных хромосомах, переходы (кроссинговер) и рекомбинация генетического материала, содержащегося в индивидуальных родительских хромосомах (аналогично гетеросексуальной репродукции), и миграции генов. В ходе работы процедур на каждой стадии эволюции получаются популяции со все более совершенными индивидуумами.

- Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой из них независимо, обмениваясь время от времени несколькими хромосомами. Существуют также и другие методы распараллеливания генетических алгоритмов.
- Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и далеко не гарантируют нахождения “лучшего” решения. Как и в реальной жизни, эволюцию может “заклинить” на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.
- Примером может служить система GeneHunter фирмы Ward Systems Group. Его стоимость — около $1000.

Эволюционное программирование

- Проиллюстрируем современное состояние данного подхода на примере системы PolyAnalyst — российской разработке, получившей сегодня общее признание на рынке Data Mining. В данной системе гипотезы о виде зависимости целевой переменной от других переменных формулируются в виде программ на некотором внутреннем языке программирования. Процесс построения программ строится как эволюция в мире программ (этим подход немного похож на генетические алгоритмы). Когда система находит программу, более или менее удовлетворительно выражающую искомую зависимость, она начинает вносить в нее небольшие модификации и отбирает среди построенных дочерних программ те, которые повышают точность. Таким образом система "выращивает" несколько генетических линий программ, которые конкурируют между собой в точности выражения искомой зависимости. Специальный модуль системы PolyAnalyst переводит найденные зависимости с внутреннего языка системы на понятный пользователю язык (математические формулы, таблицы и пр.).

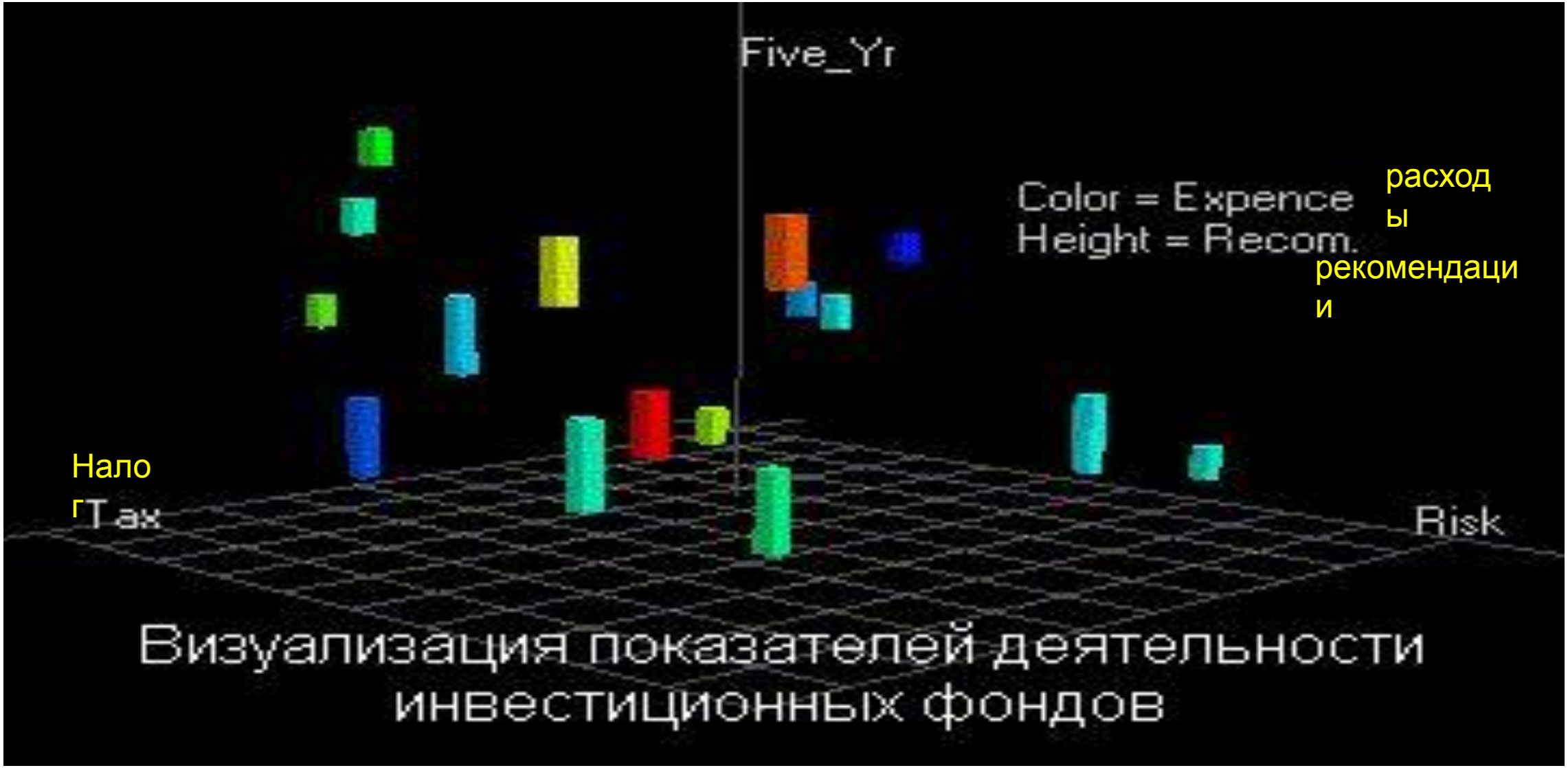
- Другое направление эволюционного программирования связано с поиском зависимости целевых переменных от остальных в форме функций какого-то определенного вида. Например, в одном из наиболее удачных алгоритмов этого типа — методе группового учета аргументов (МГУА) зависимость ищут в форме полиномов. В настоящее время из продающихся в России систем МГУА реализован в системе NeuroShell компании Ward Systems Group.
- Стоимость систем до $ 5000.

Алгоритмы ограниченного перебора

- Алгоритмы ограниченного перебора были предложены в середине 60-х годов М.М. Бонгардом для поиска логических закономерностей в данных. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей.
- Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: X = a; X < a; X > a; a < X < b и др., где X — какой либо параметр, “a” и “b” — константы. Ограничением служит длина комбинации простых логических событий (у М. Бонгарда она была равна 3). На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр.
- Наиболее ярким современным представителем этого подхода является система WizWhy предприятия WizSoft. Хотя автор системы Абрахам Мейдан не раскрывает специфику алгоритма, положенного в основу работы WizWhy, по результатам тщательного тестирования системы были сделаны выводы о наличии здесь ограниченного перебора (изучались результаты, зависимости времени их получения от числа анализируемых параметров и др.).
- Система WizWhy является на сегодняшний день одним из лидеров на рынке продуктов Data Mining. Это не лишено оснований. Система постоянно демонстрирует более высокие показатели при решении практических задач, чем все остальные алгоритмы. Стоимость системы около $ 4000, количество продаж — 30000.

Системы для визуализации многомерных данных

- В той или иной мере средства для графического отображения данных поддерживаются всеми системами Data Mining. Вместе с тем, весьма внушительную долю рынка занимают системы, специализирующиеся исключительно на этой функции. Примером здесь может служить программа DataMiner 3D словацкой фирмы Dimension5 (5-е измерение).
- В подобных системах основное внимание сконцентрировано на дружелюбности пользовательского интерфейса, позволяющего ассоциировать с анализируемыми показателями различные параметры диаграммы рассеивания объектов (записей) базы данных. К таким параметрам относятся цвет, форма, ориентация относительно собственной оси, размеры и другие свойства графических элементов изображения. Кроме того, системы визуализации данных снабжены удобными средствами для масштабирования и вращения изображений. Стоимость систем визуализации может достигать нескольких сотен долларов.



- Рисунок 8. Визуализация данных системой DataMiner 3D

Выводы

- 1. Рынок систем Data Mining экспоненциально развивается. В этом развитии принимают участие практически все крупнейшие корпорации. В частности, Microsoft непосредственно руководит большим сектором данного рынка (издает специальный журнал, проводит конференции, разрабатывает собственные продукты).
- 2. Системы Data Mining применяются по двум основным направлениям: 1) как массовый продукт для бизнес-приложений; 2) как инструменты для проведения уникальных исследований (генетика, химия, медицина и пр.). В настоящее время стоимость массового продукта от $1000 до $10000. Количество инсталляций массовых продуктов, судя по имеющимся сведениям, сегодня достигает десятков тысяч. Лидеры Data Mining связывают будущее этих систем с использованием их в качестве интеллектуальных приложений, встроенных в корпоративные хранилища данных.
- Несмотря на обилие методов Data Mining, приоритет постепенно все более смещается в сторону логических алгоритмов поиска в данных if-then правил. С их помощью решаются задачи прогнозирования, классификации, распознавания образов, сегментации БД, извлечения из данных “скрытых” знаний, интерпретации данных, установления ассоциаций в БД и др. Результаты таких алгоритмов эффективны и легко интерпретируются.

- Вместе с тем, главной проблемой логических методов обнаружения закономерностей является проблема перебора вариантов за приемлемое время. Известные методы либо искусственно ограничивают такой перебор (алгоритмы KOPA, WizWhy), либо строят деревья решений (алгоритмы CART, CHAID, ID3, See5, Sipina и др.), имеющих принципиальные ограничения эффективности поиска if-then правил. Другие проблемы связаны с тем, что известные методы поиска логических правил не поддерживают функцию обобщения найденных правил и функцию поиска оптимальной композиции таких правил. Удачное решение указанных проблем может составить предмет новых конкурентоспособных разработок.