Что такое Data Mining?

Что такое Data Mining?

В литературе переводится по-разному

- Добыча данных (калька)
- Интеллектуальный анализ данных (а бывает неинтеллектуальный?)
- Искусственный интеллект (перевод в стиле школы времен АСУ)
- Поиск закономерностей
- ...

- Независимо от перевода смысл одинаков (в большинстве случаев):

Это средство превратить данные в знания

Мало прока от строки в таблице, говорящей, что в день A клиент B приобрел товар C в магазине D на сумму ... в кол-ве ... и т.д.

Однако просмотрев миллионы подобных строк можно заметить, например:

что товар C в магазине D расходится лучше, чем в других торговых точках,

что клиент B проявляет покупательскую активность в дни A

что товар C1 продается в основном с товаром C

...

Эти знания уже можно непосредственно использовать в бизнесе

Почему мы сегодня говорим о технологии Data Mining?

- За последние два десятилетия реляционные БД на предприятиях накопили грандиозные объемы данных в самых различных областях и приложениях
  - ERP, CRM, Inventory, финансы, …
  - Просто журналы посещений, наконец
- Для чего реально использовались эти данные?
  - Выпустили пару раз отчетность на их основе, потом сагрегировали, заархивировали и забыли?
  - Лежат мертвым грузом вместо того, чтобы работать и приносить прибыль
  - Data Mining – средство их «оживить» и заставить работать

- Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с Data Mining встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in databases) 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.

- Данные имеют неограниченный объем
- Данные являются разнородными (количественными, качественными, текстовыми)
- Результаты должны быть конкретны и понятны
- Инструменты для обработки сырых данных должны быть просты в использовании

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Какие факторы лучше всего предсказывают несчастные случаи?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)?	Какие характеристики отличают клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Какие схемы покупок характерны для мошенничества с кредитными карточками?


Уровни знаний,
извлекаемых из данных
Аналитические
инструменты
Технологии
«сверху-вниз»
ПОВЕРХНОСТНЫЙ
Язык простых
запросов
НЕГЛУБОКИЙ
Оперативная
аналитическая
обработка
СКРЫТЫЙ
Data Mining
«Раскопка данных»
Технологии
«снизу-вверх»

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

- ***Data Mining*** - это процесс выделения из *данных* неявной и неструктурированной информации и представления ее в виде, пригодном для использования.
- ***Data Mining*** - это процесс выделения, исследования и моделирования больших объемов *данных* для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (определение SAS Institute).
- ***Data Mining*** - это процесс, цель которого - обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых *данных* с использованием методик распознавания образцов плюс применение статистических и математических методов (определение Gartner Group).

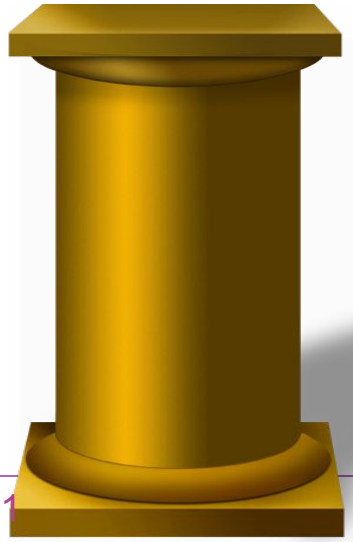
Три принципа в основе DM

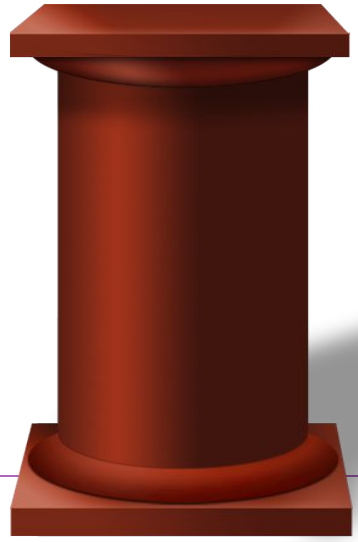
- Иными словами, Data Mining – это анализ данных с целью отыскания в них типовых образцов или стереотипных изменений, скрытых от нас по причине невозможности держать в голове такое количество данных и анализировать такое количество взаимосвязей между ними

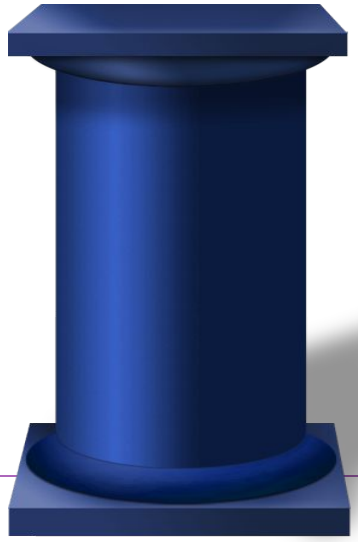
Исследование данных

Отыскание образцов

Предсказание результатов







Задачи Data Mining

1. 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).

2. 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.

3. 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.

4. 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.

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

5.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.

6. 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.

7. ***Оценивание (Estimation)***

Задача оценивания сводится к предсказанию непрерывных значений признака.

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

9. 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.

10.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:

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

Задачи Data Mining, в зависимости от используемых моделей, могут быть

дескриптивными и прогнозирующими.

Применение Data Mining для решения бизнес-задач.

- Основные направления: банковское дело, финансы, страхование, CRM, производство, телекоммуникации, электронная коммерция, маркетинг, фондовый рынок и другие.
- Применение Data Mining для решения задач государственного уровня. Основные направления: поиск лиц, уклоняющихся от налогов; средства в борьбе с терроризмом.
- Применение Data Mining для научных исследований. Основные направления: медицина, биология, молекулярная генетика и генная инженерия, биоинформатика, астрономия, прикладная химия, исследования, касающиеся наркотической зависимости, и другие.
- Применение Data Mining для решения Web-задач. Основные направления: поисковые машины (search engines), счетчики и другие

Банковское дело

- Задача привлечения новых клиентов банка
- Другие задачи сегментации клиентов
- Задача управления ликвидностью банка
- Задача выявления случаев мошенничества с кредитными картами

- Страхование
- Электронная коммерция

Основные задачи Data Mining в промышленном производстве :

· комплексный системный анализ производственных ситуаций;
· краткосрочный и долгосрочный прогноз развития производственных ситуаций;
· выработка вариантов оптимизационных решений;
· прогнозирование качества изделия в зависимости от некоторых параметров технологического процесса;
· обнаружение скрытых тенденций и закономерностей развития производственных процессов;
· прогнозирование закономерностей развития производственных процессов;
· обнаружение скрытых факторов влияния;
· обнаружение и идентификация ранее неизвестных взаимосвязей между производственными параметрами и факторами влияния;
· анализ среды взаимодействия производственных процессов и прогнозирование изменения ее характеристик;
· выработку оптимизационных рекомендаций по управлению производственными процессами;
· визуализацию результатов анализа, подготовку предварительных отчетов и проектов
допустимых решений с оценками достоверности и эффективности возможных реализаций.

Вот список задач фондового рынка, которые можно решать при помощи технологии Data Mining :

· прогнозирование будущих значений финансовых инструментов и индикаторов по их прошлым значениям;

· прогноз тренда (будущего направления движения - рост, падение) финансового инструмента и его силы (сильный, умеренно сильный и т.д.);

· выделение кластерной структуры рынка, отрасли, сектора по некоторому набору характеристик;

· динамическое управление портфелем;

· прогноз волатильности;

· оценка рисков;

· предсказание наступления кризиса и прогноз его развития;

· выбор активов и др.

Web Mining

- Web Mining можно перевести как "добыча данных в Web".
- Здесь можно выделить два основных направления: Web Content Mining и Web Usage Mining.

Web Content Mining

В этом направлении, в свою очередь, выделяют два подхода:

- подход, основанный на агентах,
- подход, основанный на базах данных.

Подход, основанный на агентах (Agent Based Approach), включает такие системы:

· интеллектуальные поисковые агенты (Intelligent Search Agents);

· фильтрация информации / классификация;

· персонифицированные агенты сети.

Примеры систем интеллектуальных агентов поиска:

- Harvest (Brown и др., 1994),
- FAQ-Finder (Hammond и др., 1995),
- Information Manifold (Kirk и др., 1995),
- OCCAM (Kwok and Weld, 1996), and ParaSite (Spertus, 1997),
- ILA (Information Learning Agent) (Perkowitz and Etzioni, 1995),
- ShopBot (Doorenbos и др., 1996).

- Подход, основанный на базах данных (Database Approach), включает системы:

· многоуровневые базы данных;

· системы web-запросов (Web Query Systems);

Примеры систем web-запросов:

- W3QL (Konopnicki и Shmueli, 1995),
- WebLog (Lakshmanan и др., 1996),
- Lorel (Quass и др., 1995),
- UnQL (Buneman и др., 1995 and 1996),
- TSIMMIS (Chawathe и др.., 1994).

Второе направление Web Usage Mining подразумевает обнаружение закономерностей в действиях пользователя Web-узла или их группы.

Анализируется следующая информация:

· какие страницы просматривал пользователь;

· какова последовательность просмотра страниц.

Анализируется также, какие группы пользователей можно выделить среди общего их числа на основе истории просмотра Web-узла.

Web Usage Mining включает следующие составляющие:

- предварительная обработка;
- операционная идентификация;
- инструменты обнаружения шаблонов;
- инструменты анализа шаблонов.

Плюсы и минусы Web Usage Mining

Плюсы

Web Usage Mining имеет ряд преимуществ, что делает эту технологию привлекательной для корпораций, в том числе государственных учреждений[13]:

- Эта технология позволила электронной торговле создать персонализированный маркетинг, который в конечном итоге привел к увеличению объемов торговли.
- Государственные учреждения используют эту технологию для классификации угроз и для борьбы с терроризмом.
- Прогнозирование возможностей горнодобывающей промышленности может принести пользу обществу путём выявления преступной деятельности.
- Компании могут установить более тесные взаимоотношениями с клиентами, предоставляя им именно то, что им нужно.
- Компании могут лучше понять потребности клиента и быстрее реагировать на потребности клиентов.
- Компании могут найти, привлечь и удержать клиентов, сэкономить на себестоимости продукции за счет использования приобретенного понимания требований заказчика.
- Компании повышают рентабельность за счет целевого ценообразования на основе созданных профилей.

Минусы

- Самый критикуемый этический вопрос, связанный с Web Usage Mining, является вопрос о вторжении в частную жизнь. Защита считается потерянной, когда полученная информация об отдельном пользователе используется или распространяется без их ведома и согласия. Полученные данные будут проанализированы и кластеризованы в форме профилей или будут анонимными до кластеризации без создания личных профилей. Таким образом, эти приложения де-индивидуализируют пользователя, судя о них только по их щелчками мыши[14].
- Другой важной проблемой является то, что компании по сбору данных могут их использовать для совершенно разных целей, что существенно нарушает интересы пользователей.
- Растущая тенденция использования персональных данных в качестве товара призывает владельцев веб-сайтов к торговле этими данными, расположенными на их сайтах.
- Некоторые алгоритмы интеллектуального анализа могут использовать спорные атрибуты, как пол, раса, религия или сексуальная ориентация. Эти методы могут быть против анти-дискриминационного законодательства.

Задачи Web Mining можно подразделить на такие категории:

· Предварительная обработка данных для Web Mining.

· Обнаружение шаблонов и открытие знаний с использованием ассоциативных правил,

временных последовательностей, классификации и кластеризации;

· Анализ полученного знания.

Text Mining

Text Mining охватывает новые методы для выполнения семантического анализа текстов, информационного поиска и управления. Синонимом понятия Text Mining является KDT (Knowledge Discovering in Text - поиск или обнаружение знаний в тексте).

В отличие от технологии Data Mining, которая предусматривает анализ упорядоченной в некие структуры информации, технология Text Mining анализирует большие и сверхбольшие массивы неструктурированной информации.

Программы, реализующие эту задачу, должны некоторым образом оперировать естественным человеческим языком и при этом понимать семантику анализируемого текста. Один из методов, на котором основаны некоторые Text Mining системы, - поиск так называемой подстроки в строке

- Извлечение понятий
- Ответ на запросы
- Тематическое индексирование
- Поиск по ключевым словам

Call Mining

Среди разработчиков новой технологии Call Mining ("добыча" и анализ звонков) -

компании CallMiner, Nexidia, ScanSoft, Witness Systems.

В технологии Call Mining разработано два подхода

- на основе преобразования речи в текст
- на базе фонетического анализа.

Классификация стадий Data Mining

Стадия 1. Выявление закономерностей (свободный поиск).

Стадия 2. Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование).

Стадия 3. Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

Классификация стадий Data Mining

Стадия 1. Выявление закономерностей (свободный поиск).

Стадия 2. Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование).

Стадия 3. Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

СВОБОДНЫЙ ПОИСК (в том числе ВАЛИДАЦИЯ) ->

-> ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ->

-> АНАЛИЗ ИСКЛЮЧЕНИЙ

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

Свободный поиск представлен такими **действиями**:

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

"**Если возраст < 20 лет и** желаемый уровень вознаграждения > 700 условных единиц, то в 75% случаев соискатель ищет работу программиста"

или

"**Если возраст >35 лет и желаемый уровень** вознаграждения > 1200 условных единиц, то в 90% случаев соискатель ищет руководящую работу".

Целевой переменной в описанных правилах выступает профессия.

При задании другой целевой переменной, например, возраста, получаем такие правила:

"**Если соискатель ищет руководящую работу и его стаж > 15 лет, то возраст соискателя >** 35 лет в 65 % случаев".

Описанные действия, 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).

2. Прогностическое моделирование (Predictive Modeling)

Прогностическое моделирование включает такие **действия**:

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

Зная, что соискатель ищет руководящую работу и его стаж > 15 лет, на 65 % можно быть уверенным в том, что возраст соискателя > 35 лет.

Или же, если возраст соискателя > 35 лет и желаемый уровень вознаграждения > 1200 условных единиц, на 90% можно быть уверенным в том, что соискатель ищет руководящую работу.

3. Анализ исключений (forensic analysis)

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

Действие, выполняемое на этой стадии, - выявление отклонений (deviation detection). Для выявления отклонений необходимо определить норму, которая рассчитывается на стадии свободного поиска.

Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных 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.


Распознавание образов
Нейросети
Визуализация данных
Статистика
Data Mining
Экспертные системы
Эффективные вычисления
Информационный поиск
Хранилища данных
Теория баз данных
Оперативная аналитическая обработка

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.

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.
- В качестве примеров наиболее мощных и распространенных статистических пакетов можно назвать SAS (компания SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics),STATISTICA, STADIA и другие.

Пакеты прикладных программ "Статистические методы анализа»

- **МЕЗОЗАВР** - "Статистические методы анализа временных рядов"
- **МЕЗОЗАВР-ЭКОНОМЕТРИКА**
- **САНИ** - "Статистические методы анализа нечисловой информации"
- **КЛАССМАСТЕР** - "Статистические методы классификации и дискриминантного анализа"


Clementine
PRW
DataMIND
GainSMARTS
Intelligent Miner for data
MINESET
M1
Model 1
DARWIN
ModelQuest Enterprise
pcOLPARS
CART
wizwhy
NeuroShell 2
S-PLUS
COGNOS
TOOLS THAT BUILD BUSINESS
See5

Деревья решений (decision trees)

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining.

Иногда этот метод Data Mining также называют деревьями решающих правил, деревьями классификации и регрессии.

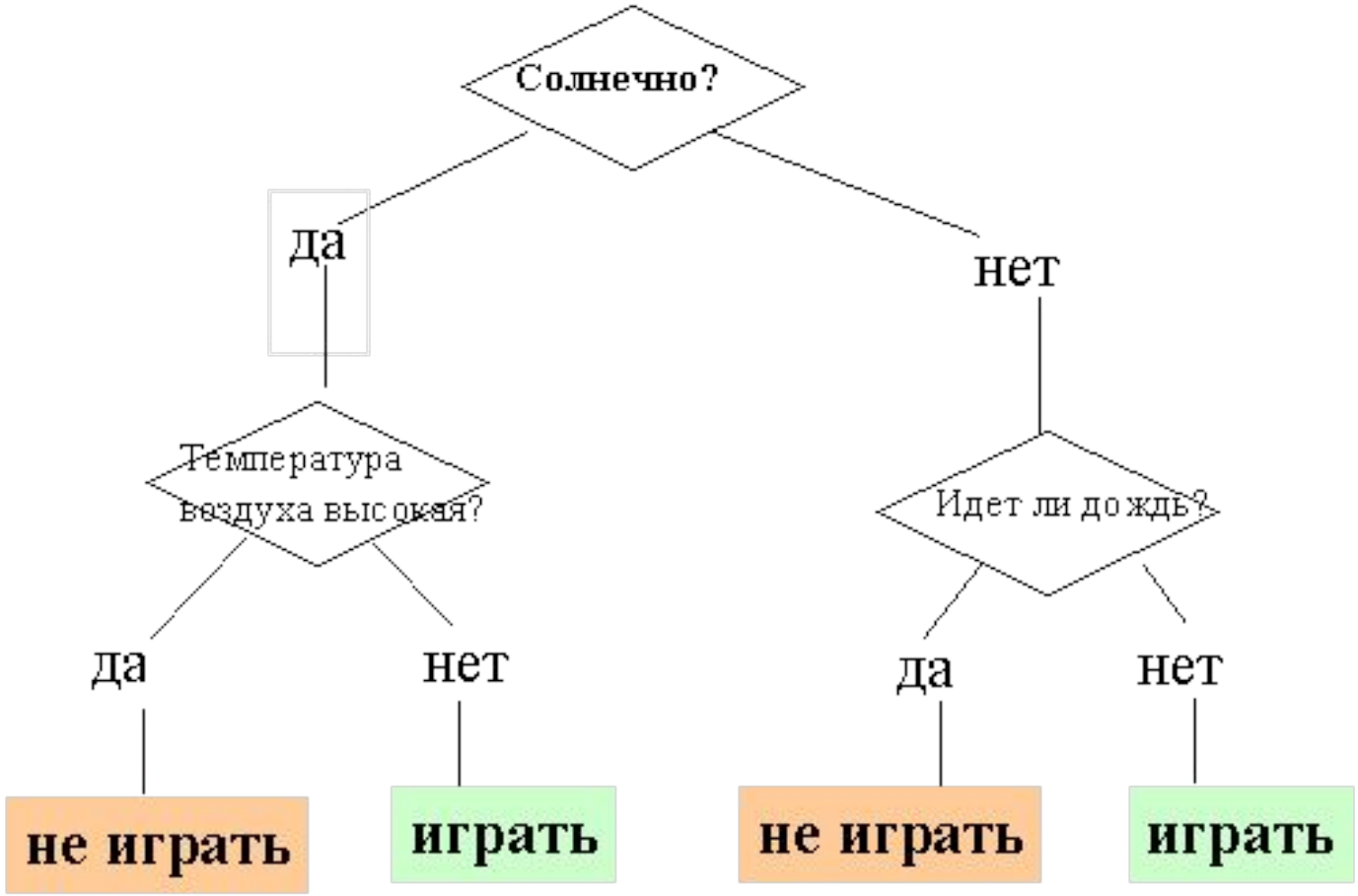
Как видно из последнего названия, при помощи данного метода решаются задачи классификации и прогнозирования.

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.

Впервые деревья решений были предложены Ховилендом и Хантом (Hoveland, Hunt) в конце 50-х годов прошлого века.

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


Солнечно?
да
нет
Температура воздуха высокая?
Идет ли дождь?
да
нет
да
нет
не играть
играть
не играть
играть

Они создают иерархическую структуру классифицирующих правил типа "ЕСЛИ... ТО..." (if-then), имеющую вид дерева.

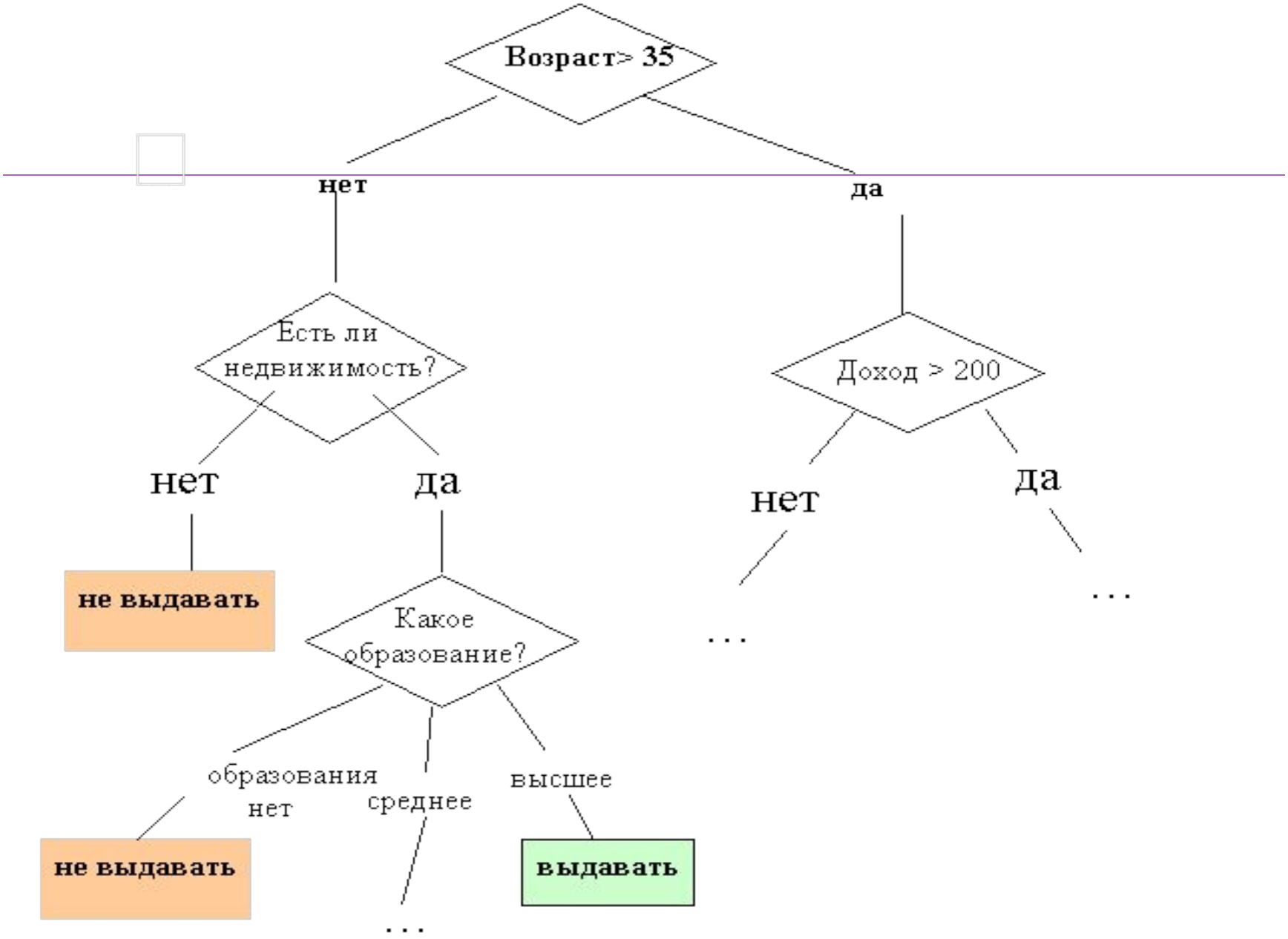
Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня.

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.

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.


Возраст> 35
нет
да
Есть ли недвижимость?
Доход > 200
нет
да
нет
да
не выдавать
Какое образование?
...
образования нет
среднее
высшее
не выдавать
выдавать
...

Как мы видим, 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)

Преимущества деревьев решений

1. **Интуитивность деревьев решений**
2. **Точность** моделей
3. **Быстрый процесс обучения**.

Алгоритмы

На сегодняшний день существует большое число алгоритмов, реализующих деревья решений:

- CART,
- C4.5,
- CHAID,
- CN2,
- NewId,
- Itrule
- и другие.

Алгоритм CART

Алгоритм CART (Classification and Regression Tree), как видно из названия, решает задачи классификации и регрессии. Он разработан в 1974-1984 годах четырьмя профессорами статистики - Leo Breiman (Berkeley), Jerry Friedman (Stanford), Charles Stone (Berkeley) и Richard Olshen (Stanford).

Атрибуты набора данных могут иметь как дискретное, так и числовое значение.

Алгоритм CART предназначен для построения бинарного дерева решений.

Другие особенности алгоритма CART:

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

Алгоритм C4.5

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:

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

Последняя версия алгоритма - алгоритм C4.8 - реализована в инструменте Weka как J4.8 (Java). Коммерческая реализация метода: C5.0, разработчик RuleQuest, Австралия.

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

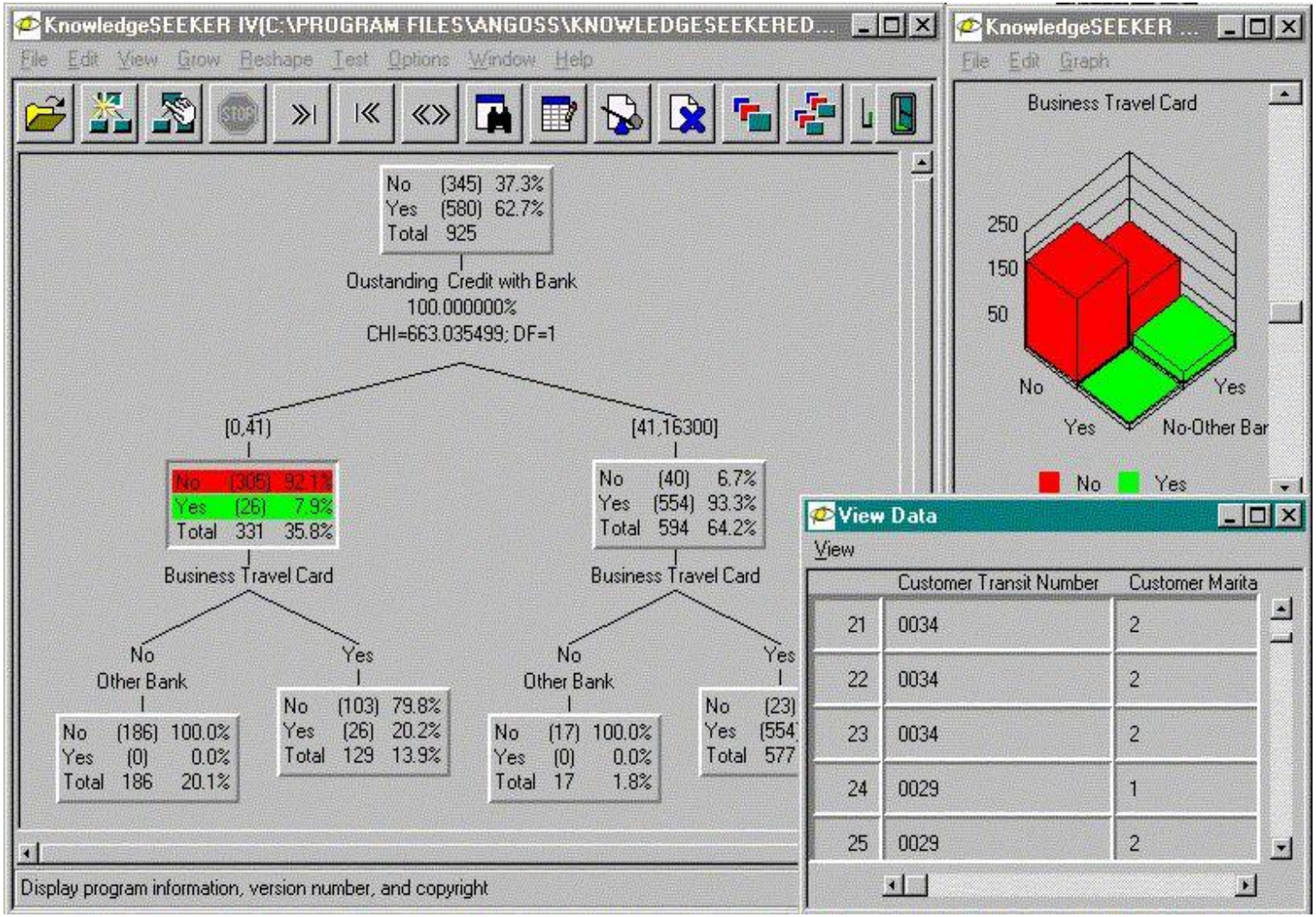
Большинство систем **Data mining** используют метод деревьев решений.

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.


Clementine
PRW
DataMIND
GainSMARTS
Intelligent Miner for data
MINESET
M1
Model 1
DARWIN
ModelQuest Enterprise
pcOLPARS
CART
wizwhy
NeuroShell 2
S-PLUS
COGNOS
TOOLS THAT BUILD BUSINESS
See5


KnowledgeSEEKER IV(C:\PROGRAM FILES\ANGOSS\KNOWLEDGESEEKERED...
File Edit View Grow Reshape Test Options Window Help
No (345) 37.3%
Yes (580) 62.7%
Total 925
Oustanding Credit with Bank
100.000000%
CHI=663.035499; DF=1
[0,41)
[41,16300]
No (305) 92.1%
Yes (26) 7.9%
Total 331 35.8%
No (40) 6.7%
Yes (554) 93.3%
Total 594 64.2%
Business Travel Card
Business Travel Card
No
Other Bank
Yes
No
Other Bank
Yes
No (186) 100.0%
Yes (0) 0.0%
Total 186 20.1%
No (103) 79.8%
Yes (26) 20.2%
Total 129 13.9%
No (17) 100.0%
Yes (0) 0.0%
Total 17 1.8%
No (23)
Yes (554
Total 577
Display program information, version number, and copyright
KnowledgeSEEKER ...
File Edit Graph
Business Travel Card
250
150
50
No
Yes
Yes
No-Other Bar
No
Yes
View Data
View
Customer Transit Number
Customer Marita
21 0034 2
22 0034 2
23 0034 2
24 0029 1
25 0029 2

Метод опорных векторов

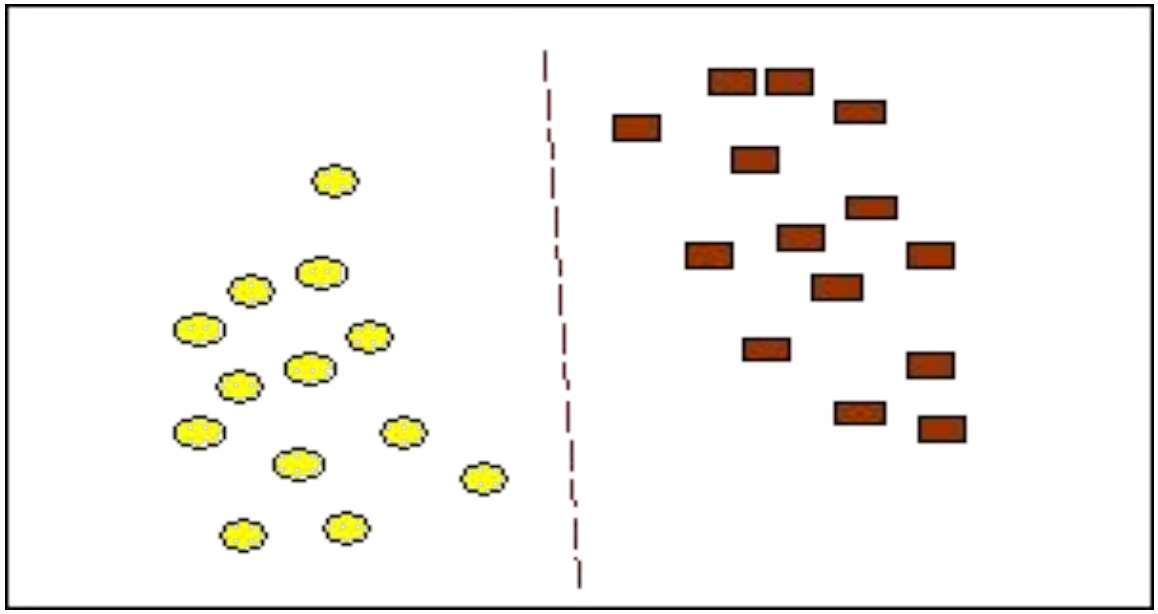
Метод опорных векторов (Support Vector Machine - SVM) относится к группе граничных методов. Она определяет классы при помощи границ областей.

При помощи данного метода решаются задачи бинарной классификации.

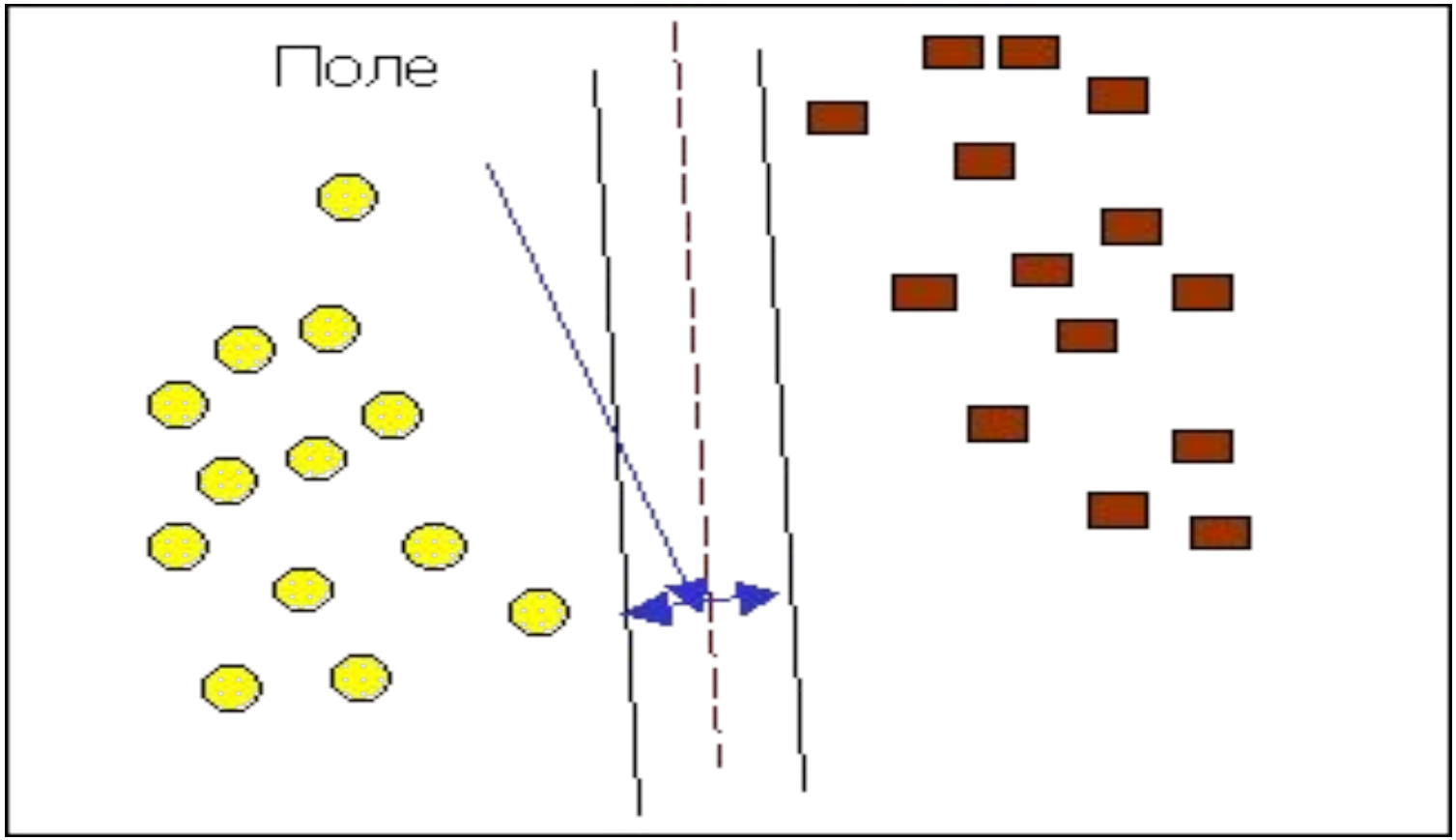
В основе метода лежит понятие плоскостей решений.

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

Метод опорных векторов



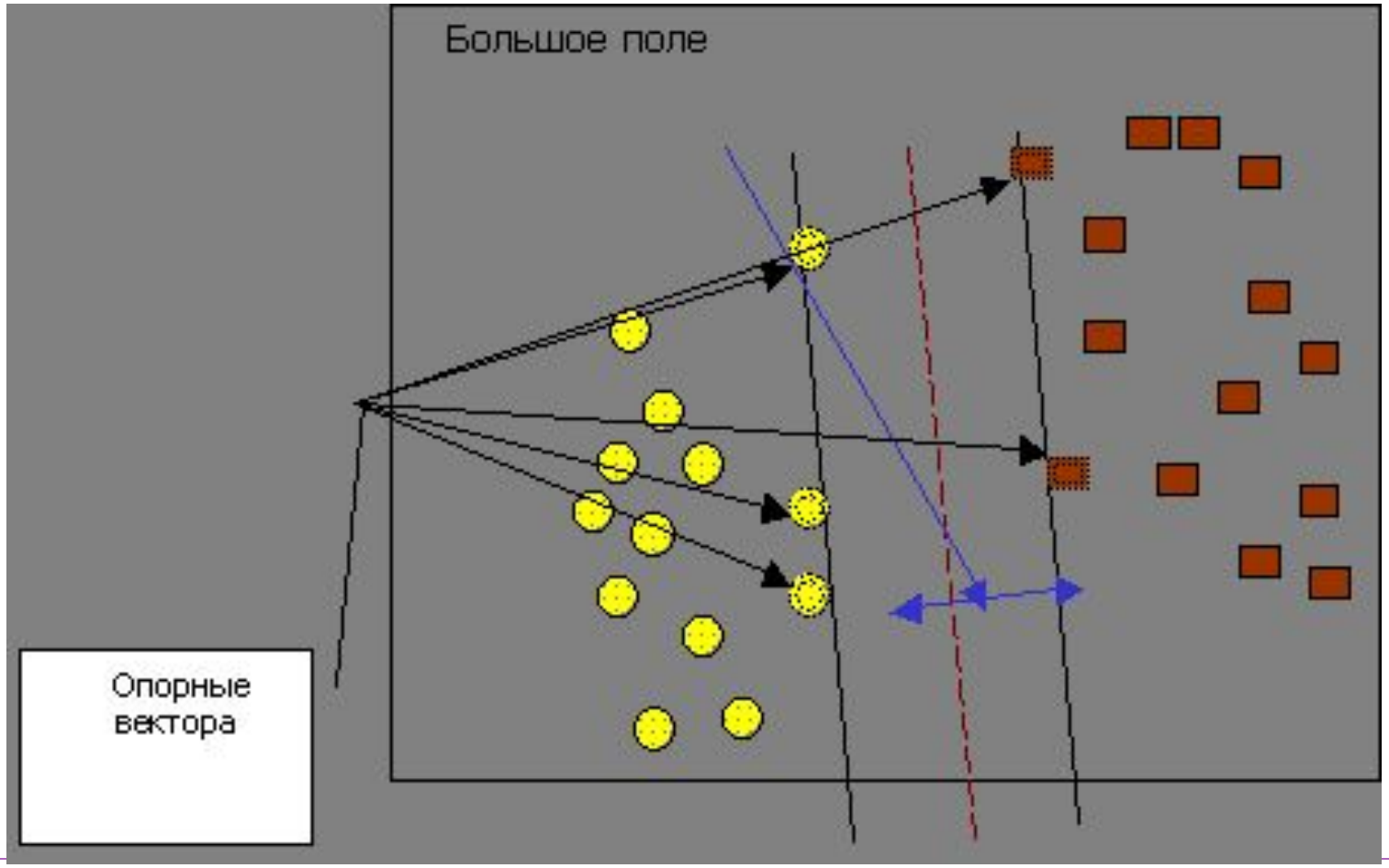
Плоскость (plane) решения разделяет объекты с разной классовой принадлежностью.



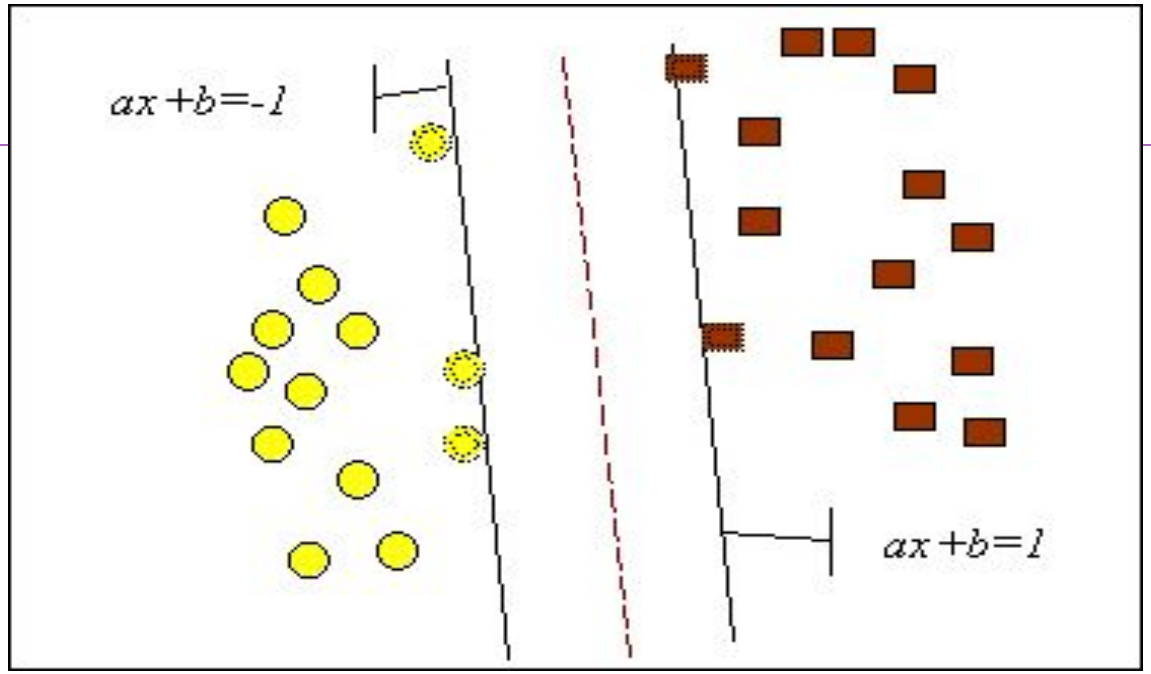
Метод отыскивает образцы, находящиеся на границах между двумя классами, т.е. опорные вектора

Опорными векторами называются объекты множества, лежащие на границах областей.

Классификация считается хорошей, если область между границами пуста.



- 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.



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.

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:

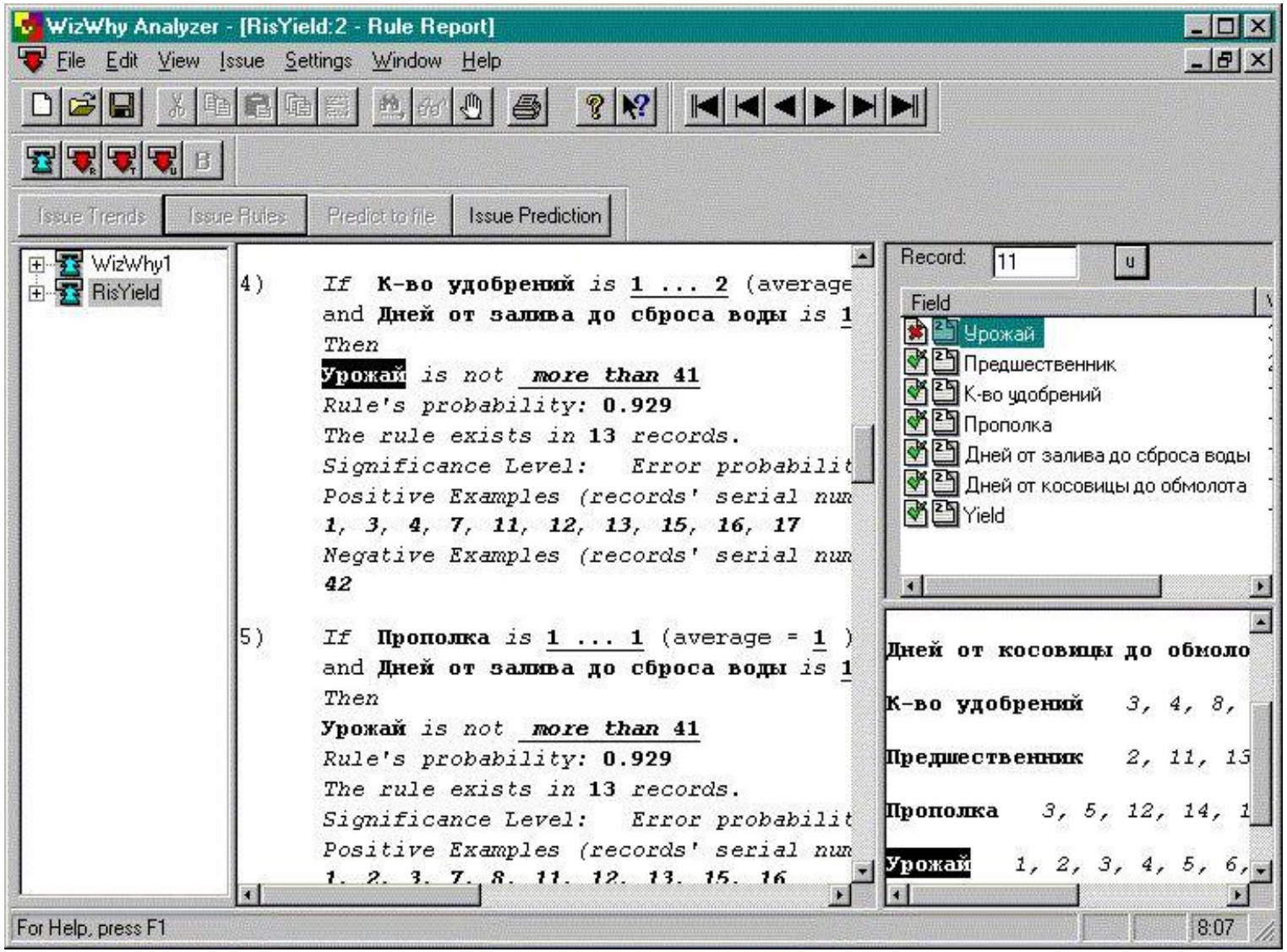
X = a; X < a; X a; a < X < b и др., где X - какой либо параметр, "a" и "b" - константы.

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.

WizWhy

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.).

Программа просматривает заданную базу данных и, собрав статистику, отыскивает правила и закономерности, которым подчиняются сведения, собранные в базе. Само собой, правила отыскиваются только там, где они действительно есть.
WizWhy, проанализировав базу данных, дает возможность пользователю заняться предсказаниями и прогнозами. Человек вводит значения известных ему параметров, а WizWhy, основываясь на обнаруженных ею в базе закономерностях, выдает наиболее вероятные значения недостающих параметров.


WizWhy Analyzer - [RisYield:2 - Rule Report]
File Edit View Issue Settings Window Help
Issue Trends
Issue Rules
Predict to file
Issue Prediction
WizWhy1
RisYield
4) If К-во удобрений is 1 ... 2 (average
and Дней от залива до сброса воды is 1
Then
Урожай is not more than 41
Rule's probability: 0.929
The rule exists in 13 records.
Significance Level: Error probabilit
Positive Examples (records' serial num
1, 3, 4, 7, 11, 12, 13, 15, 16, 17
Negative Examples (records' serial num
42
5) If Прополка is 1 ... 1 (average = 1)
and Дней от залива до сброса воды is 1
Then
Урожай is not more than 41
Rule's probability: 0.929
The rule exists in 13 records.
Significance Level: Error probabilit
Positive Examples (records' serial num
1, 2, 3, 7, 8, 11, 12, 13, 15, 16
Record: 11
u
Field
Урожай
Предшественник
К-во удобрений
Прополка
Дней от залива до сброса воды
Дней от косовицы до обмолота
Yield
Дней от косовицы до обмоло
К-во удобрений 3, 4, 8,
Предшественник 2, 11, 13
Прополка 3, 5, 12, 14, 1
Урожай 1, 2, 3, 4, 5, 6,
For Help, press F1
8:07

- WizWhy обнаруживает и математические, и логические закономерности. Допустим, что вам известны погода, температура воздуха, настроение сослуживцев и результат тиража спортлото. Предположим, вы собрали из этих сведений базу данных за несколько месяцев. WizWhy найдет в ней неочевидные человеческому взгляду закономерности там, где они и в самом деле существуют. Как уже говорилось, WizWhy можно использовать для предсказаний, основывающихся на обнаруженных закономерностях. Думаю, что в результате исследований такой базы данных вы с легкостью определите, кто из ваших сослуживцев регулярно проигрывает в спортлото, кто особенно чувствителен к жаре или холоду, кто какую погоду предпочитает и так далее. Вы не только извлечете эти знания из базы, но и сможете применить их для прогнозов.

WizWhy может стать незаменимым инструментом аналитика. Очевидны возможности ее применения в геологии, в медицинской диагностике, социальных исследованиях, при анализе клиентуры в банковских и финансовых учреждениях, в маркетинговых исследованиях и тому подобном.
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.

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

Метод "ближайшего соседа" или системы рассуждений на основе аналогичных случаев

Идея систем рассуждений по аналогии (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.

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

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

Преимущества метода

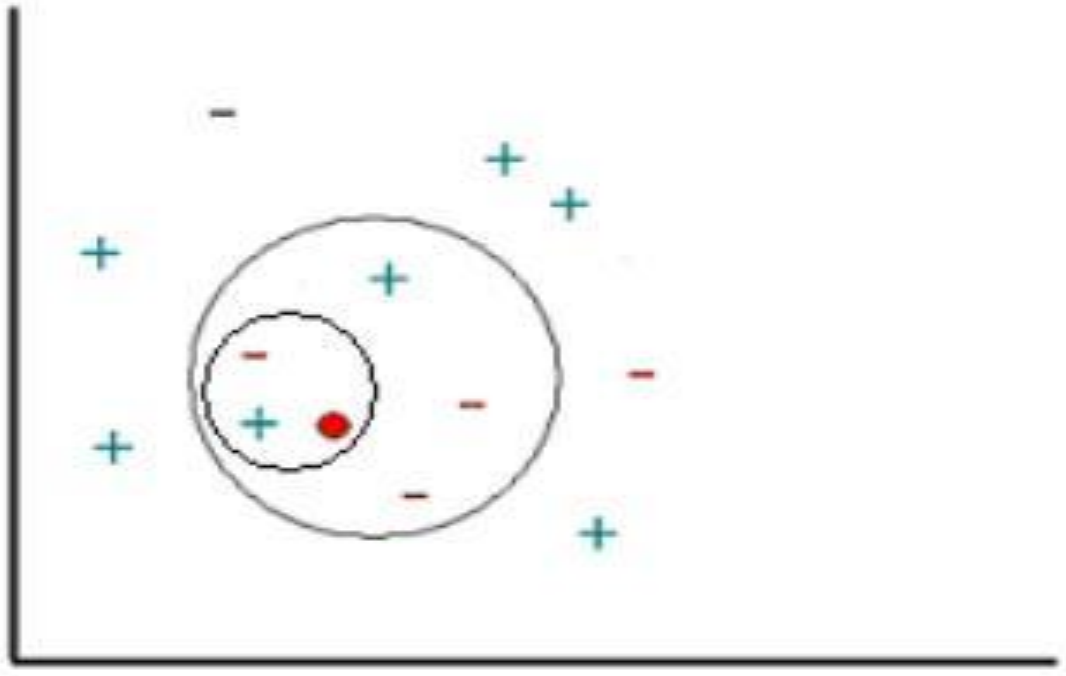
- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных.

С помощью данного метода решаются задачи классификации и регрессии.

Недостатки метода "ближайшего соседа"

- 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.
- 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
- При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

Классификация объектов множества при разном значении параметра k



Байесовская классификация

Байесовская сеть (или *байесова сеть*, *байесовская сеть доверия*, англ. *Bayesian network, belief network*) — графическая вероятностная модель, представляющая собой множество переменных и их вероятностных зависимостей.

Например, байесовская сеть может быть использована для вычисления вероятности того, чем болен пациент по наличию или отсутствию ряда симптомов, основываясь на данных о зависимости между симптомами и болезнями.

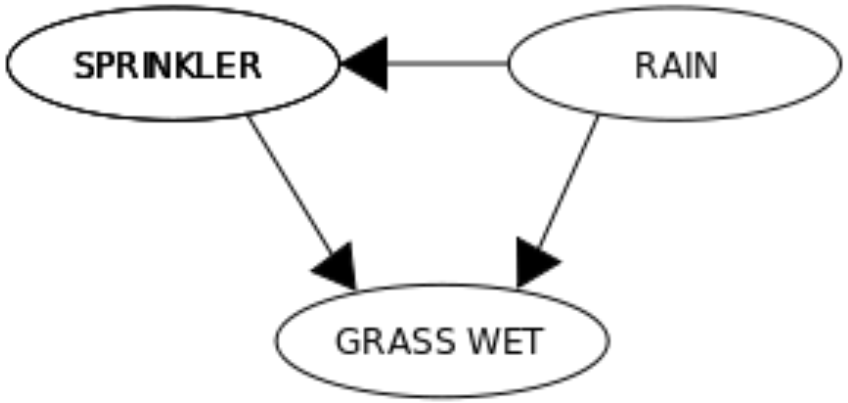
Математический аппарат байесовых сетей создан американским ученым Джудой Перлом, лауреатом Премии Тьюринга (2011).

- Предположим, что может быть две причины, по которым трава может стать мокрой (GRASS WET): сработала дождевальная установка, либо прошел дождь. Также предположим, что дождь влияет на работу дождевальной машины (во время дождя установка не включается). Тогда ситуация может быть смоделирована проиллюстрированной Байесовской сетью. Все три переменные могут принимать два возможных значения: T (правда — true) и F (ложь — false).
- Совместная вероятность функции:

$$\mathrm{P}(G, S, R) = \mathrm{P}(G|S, R)\mathrm{P}(S|R)\mathrm{P}(R)$$

- где имена трех переменных означают:
- *G = Трава мокрая (Grass wet),*
- *S = Дождевальная установка (Sprinkler),*
- *R = Дождь (Rain).*

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



	RAIN
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

суммируя переменные:

- Модель может ответить на такие вопросы как «Какова вероятность того, что прошел дождь, если трава мокрая?» используя формулу условной вероятности и суммируя переменные:

$$\mathrm{P}(R = T \mid G = T) = \frac{\mathrm{P}(G = T, R = T)}{\mathrm{P}(G = T)} = \frac{\sum_{S \in \{T,F\}} \mathrm{P}(G = T, S, R = T)}{\sum_{S,R \in \{T,F\}} \mathrm{P}(G = T, S, R)}$$

$$= \frac{(0.99 \times 0.01 \times 0.2 = 0.00198_{TTT}) + (0.8 \times 0.99 \times 0.2 = 0.1584_{TFT})}{0.00198_{TTT} + 0.288_{TTF} + 0.1584_{TFT} + 0_{TFF}} \approx 35.77\%.$$

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

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 ..

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.

Классификация нейронных сетей

Одна из возможных классификаций нейронных сетей - по направленности связей.

Нейронные сети бывают с обратными связями и без обратных связей.

1. Сети без обратных связей

- Сети с обратным распространением ошибки.

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

- 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

2. Сети с обратными связями

- Сети Хопфилда (задачи ассоциативной памяти).
- Сети Кохонена (задачи кластерного анализа).

Особенностью 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.

1. Сети прямого распространения

- Персептроны.
- Сеть Back Propagation.
- Сеть встречного распространения.
- Карта Кохонена.

2. Рекуррентные сети.

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

- Сеть Хопфилда.
- 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 управления.

Эти сети применяются в системах управления движущимися объектами.

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.

Выбор структуры нейронной сети

Существуют **принципы**, которыми следует руководствоваться при разработке новой конфигурации:

- возможности сети возрастают с увеличением числа ячеек сети, плотности связей между ними и числом выделенных слоев;
- введение обратных связей наряду с увеличением возможностей сети поднимает вопрос о динамической устойчивости сети;
- сложность алгоритмов функционирования сети (в том числе, например, введение нескольких типов синапсов - возбуждающих, тормозящих и др.) также способствует усилению мощи НС.

Карты Кохонена

- В результате работы алгоритма получаются следующие карты:
  карта входов нейронов — визуализирует внутреннюю структуру входных данных путем подстройки весов нейронов карты. Обычно используется несколько карт входов, каждая из которых отображает один из них и раскрашивается в зависимости от веса нейрона. На одной из карт определенным цветом обозначают область, в которую включаются приблизительно одинаковые входы для анализируемых примеров.
  карта выходов нейронов — визуализирует модель взаимного расположения входных примеров. Очерченные области на карте представляют собой кластеры, состоящие из нейронов со схожими значениями выходов.
  специальные карты — это карта кластеров, полученных в результате применения алгоритма самоорганизующейся карты Кохонена, а также другие карты, которые их характеризуют

Карты Кохонена

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.

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.

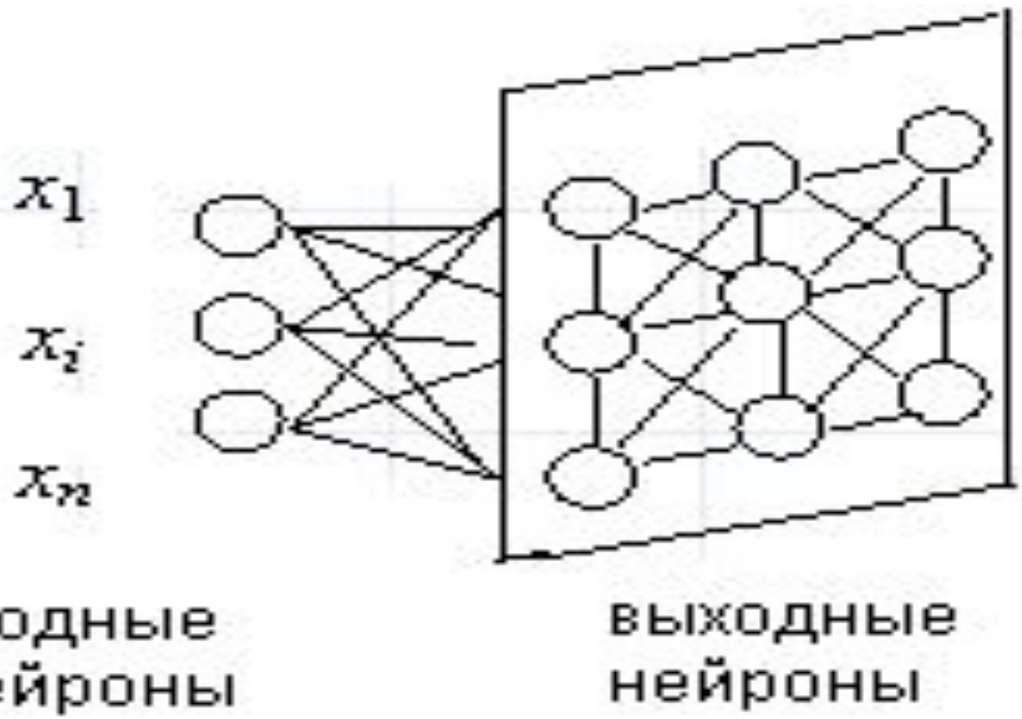
Обнаружение новых явлений.

- 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.

Карты Кохонена

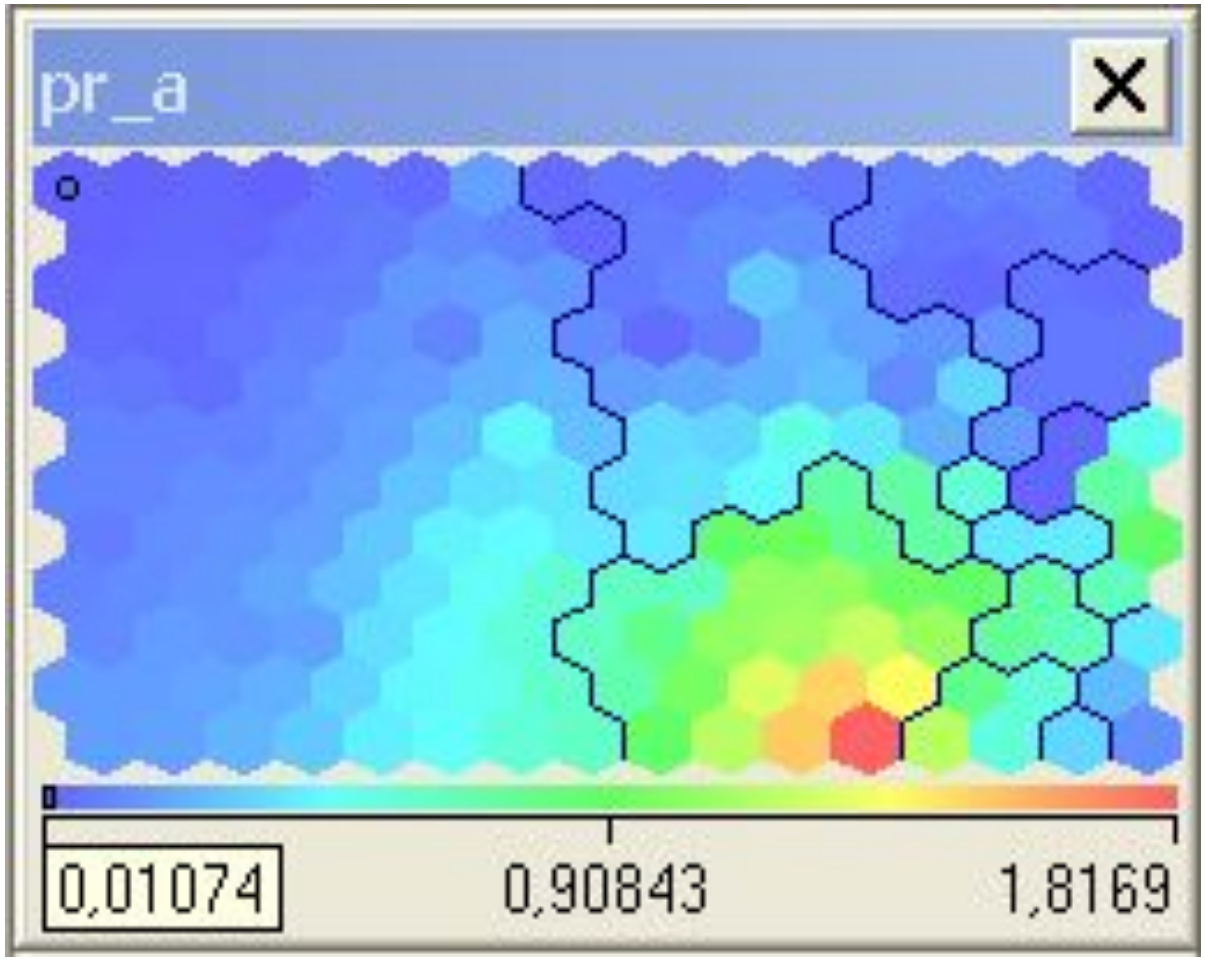
Самоорганизующаяся карта состоит из компонентов, называемых узлами или нейронами. Их количество задаётся аналитиком.

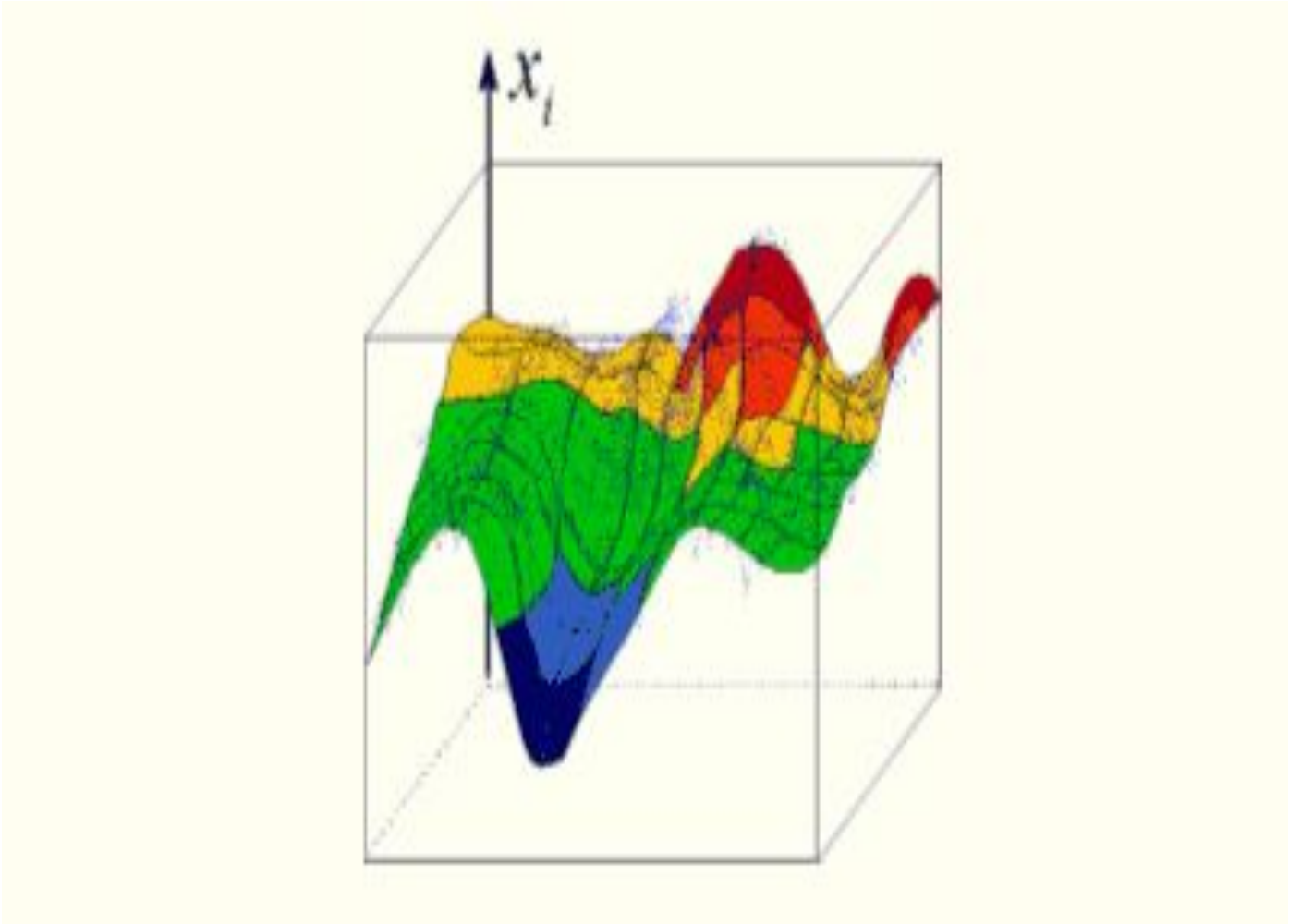
Каждый из узлов описывается двумя векторами. Первый — т. н. вектор веса *m*, имеющий такую же размерность, что и входные данные. Второй — вектор *r*, представляющий собой координаты узла на карте. Карта Кохонена визуально отображается с помощью ячеек прямоугольной или шестиугольной формы; последняя применяется чаще, поскольку в этом случае расстояния между центрами смежных ячеек одинаковы, что повышает корректность визуализации карты.

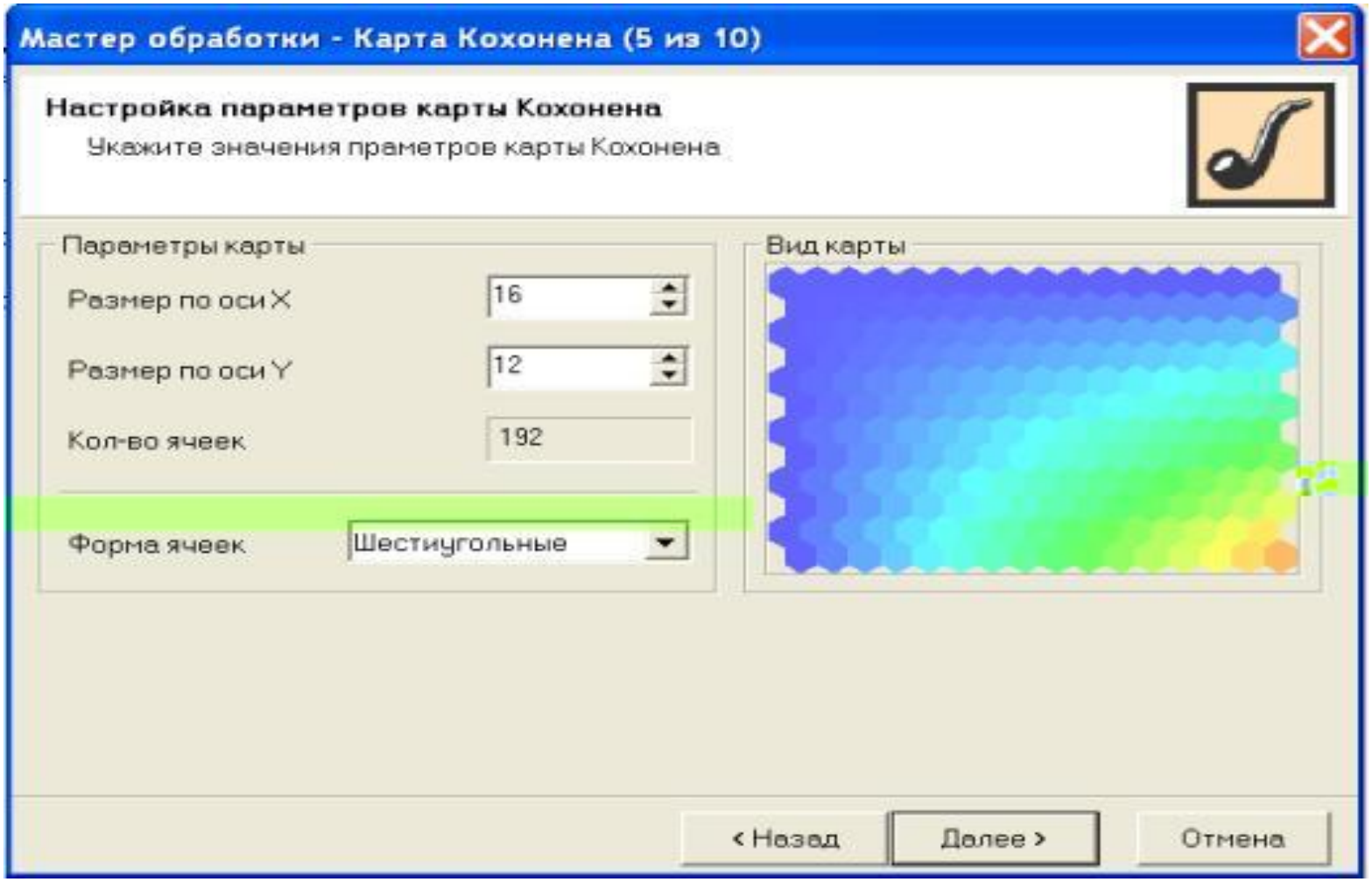

x_1
x_i
x_n
входные нейроны
выходные нейроны

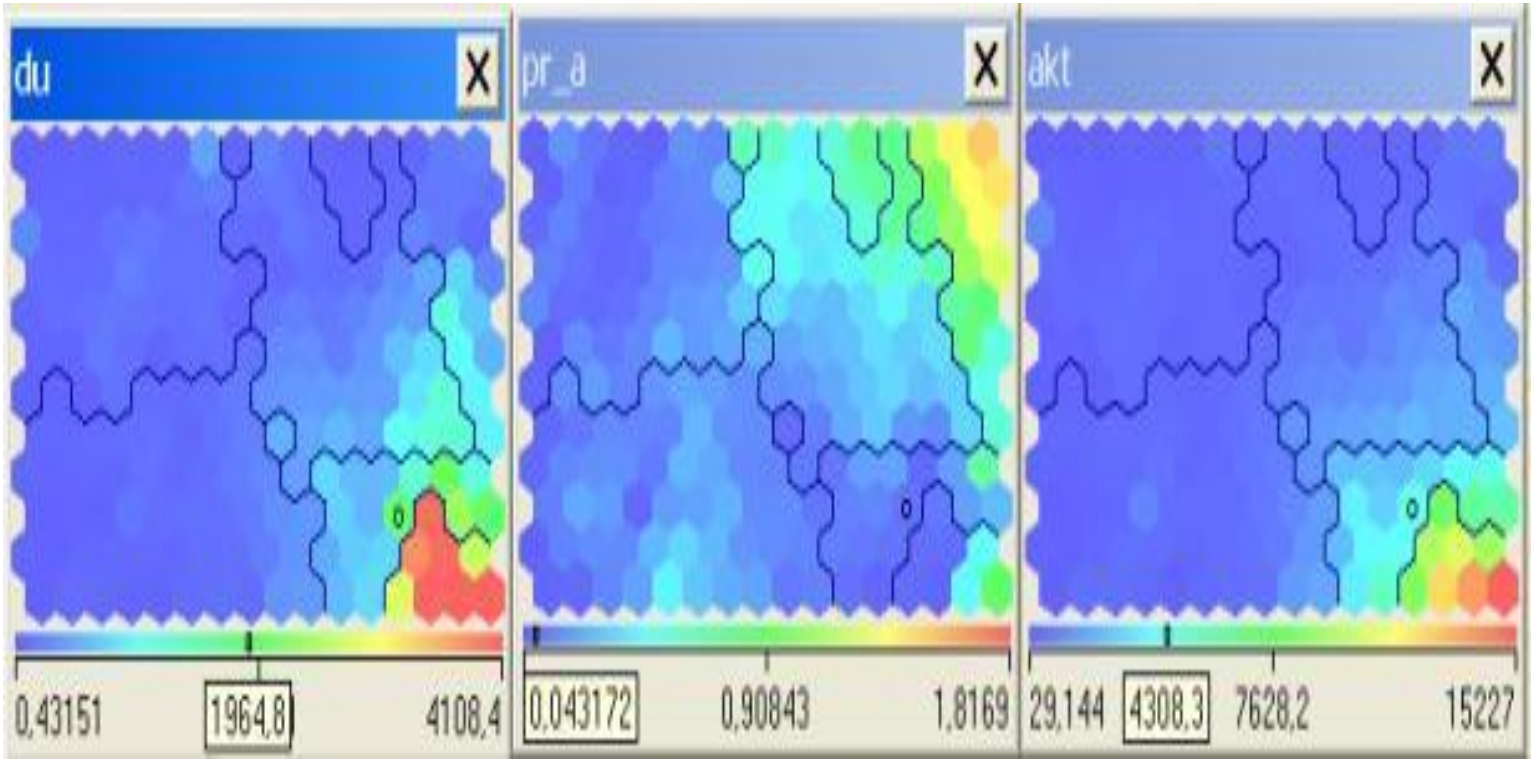
Карты Кохонена

Изначально известна размерность входных данных, по ней некоторым образом строится первоначальный вариант карты. В процессе обучения векторы веса узлов приближаются к входным данным. Для каждого наблюдения (семпла) выбирается наиболее похожий по вектору веса узел, и значение его вектора веса приближается к наблюдению. Также к наблюдению приближаются векторы веса нескольких узлов, расположенных рядом, таким образом если в множестве входных данных два наблюдения были схожи, на карте им будут соответствовать близкие узлы. Циклический процесс обучения, перебирающий входные данные, заканчивается по достижении картой допустимой (заранее заданной аналитиком) погрешности, или по совершении заданного количества итераций. Таким образом, в результате обучения карта Кохонена классифицирует входные данные на кластеры и визуально отображает многомерные входные данные в двумерной плоскости, распределяя векторы близких признаков в соседние ячейки и раскрашивая их в зависимости от анализируемых параметров нейронов.


pr_a
0,01074
0,90843
1,8169


x_i


Мастер обработки - Карта Кохонена (5 из 10)
Настройка параметров карты Кохонена
Укажите значения праметров карты Кохонена
Параметры карты
Размер по оси X
16
Размер по оси Y
12
Кол-во ячеек
192
Форма ячеек
Шестиугольные
Вид карты
< Назад
Далее >
Отмена

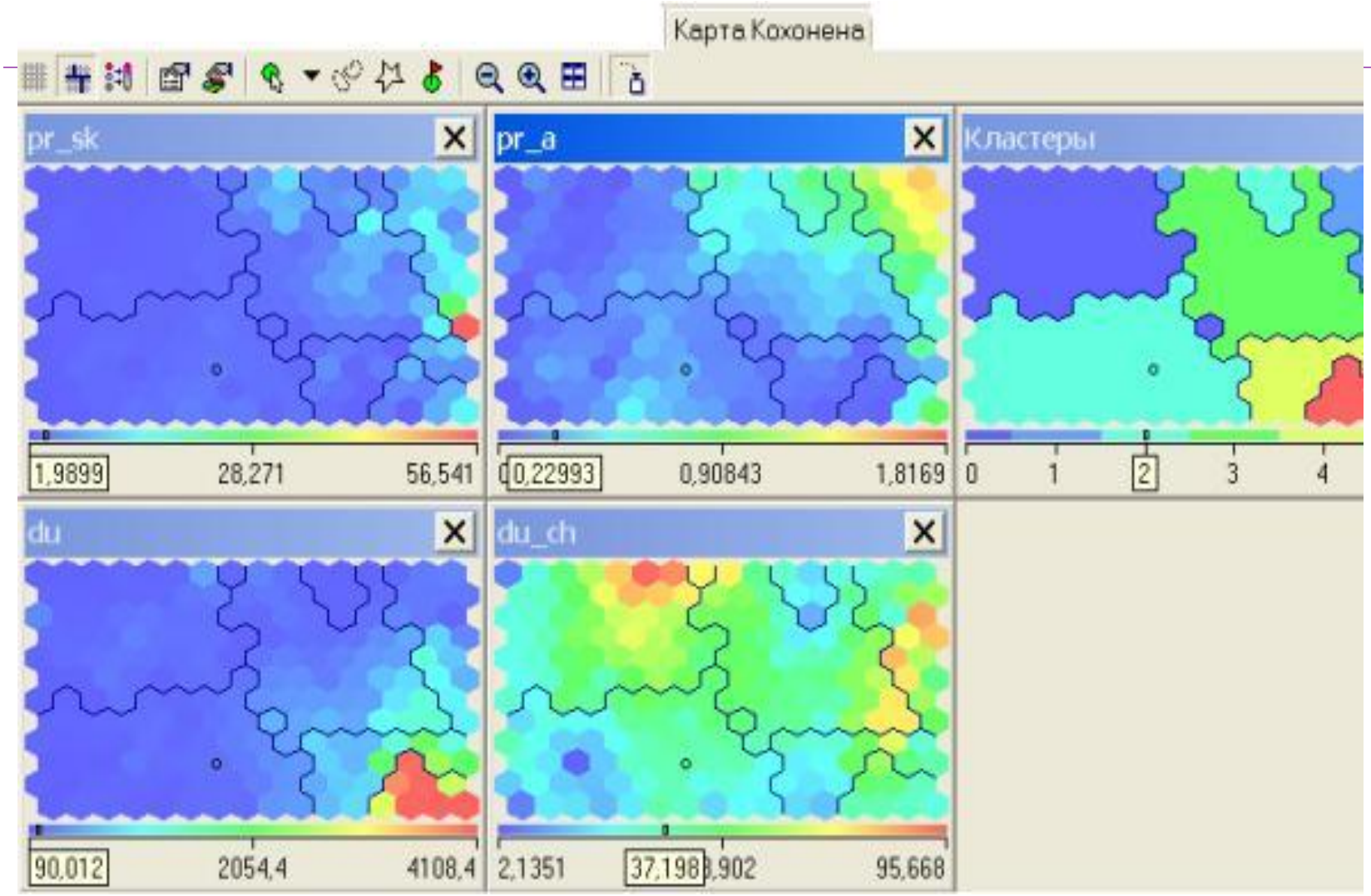

du
0,43151
1964,8
4108,4
pr_a
0,043172
0,90843
1,8169
akt
29,144
4308,3
7628,2
15227

В результате работы алгоритма получаются следующие карты:

карта входов нейронов — визуализирует внутреннюю структуру входных данных путем подстройки весов нейронов карты. Обычно используется несколько карт входов, каждая из которых отображает один из них и раскрашивается в зависимости от веса нейрона. На одной из карт определенным цветом обозначают область, в которую включаются приблизительно одинаковые входы для анализируемых примеров.

карта выходов нейронов — визуализирует модель взаимного расположения входных примеров. Очерченные области на карте представляют собой кластеры, состоящие из нейронов со схожими значениями выходов.

специальные карты — это карта кластеров, полученных в результате применения алгоритма самоорганизующейся карты Кохонена, а также другие


Карта Кохонена
pr_sk
1,9899
28,271
56,541
pr_a
0,22993
0,90843
1,8169
Кластеры
0
1
2
3
4
du
90,012
2054,4
4108,4
du_ch
2,1351
37,198
95,668

Полиномиальная нейросеть

$$Z_{17} = 3.1 + 0.4a - .15b^2 + 0.9bc - 0.62abc + 0.5c^3$$

Методы кластерного анализа. Иерархические методы

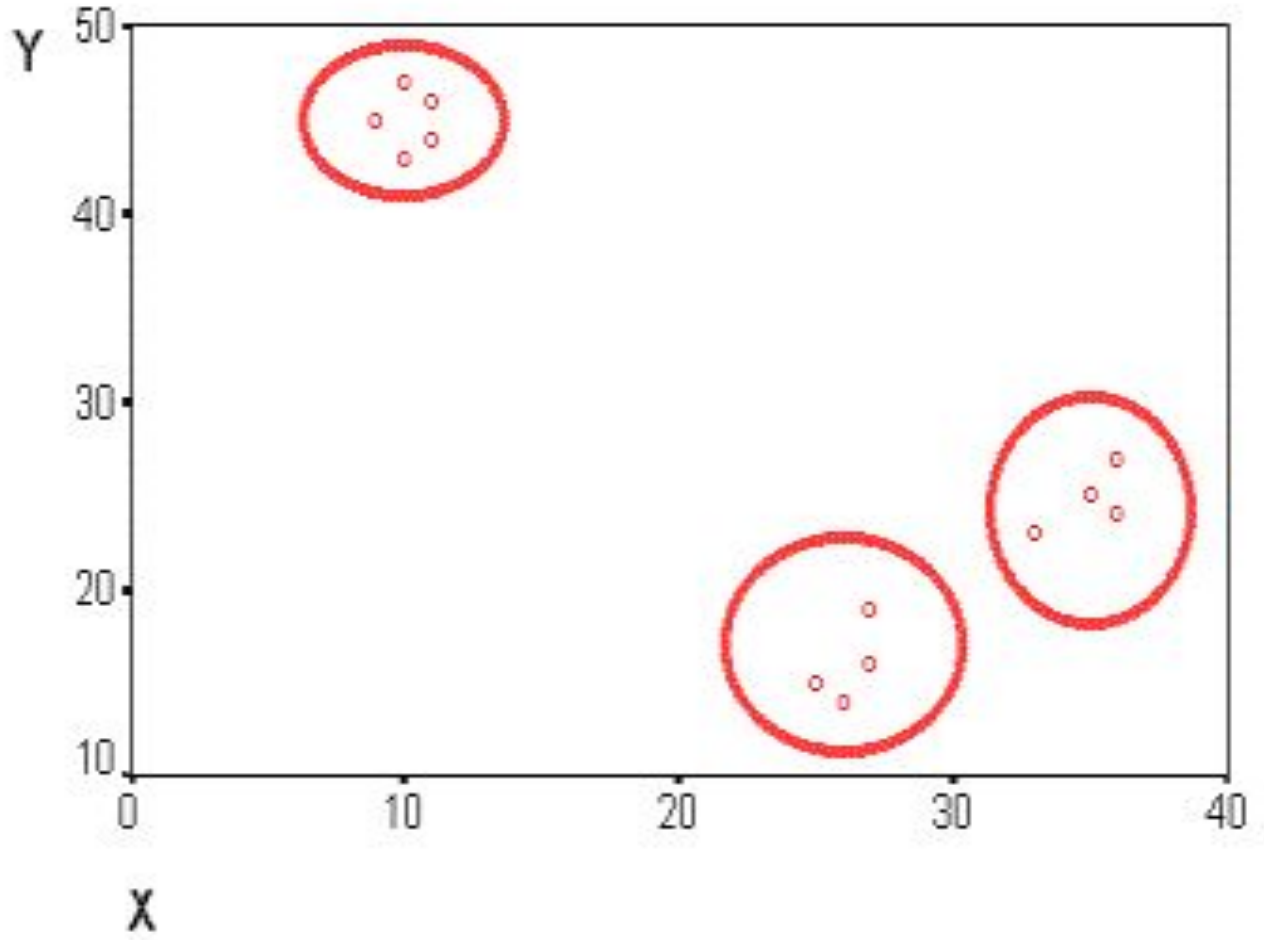
Задачи кластерного анализа можно объединить в следующие группы:

- Разработка типологии или классификации.
- Исследование полезных концептуальных схем группирования объектов.
- Представление гипотез на основе исследования данных.
- 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.

Таблица 13.1. Набор данных A

№ 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


Y
50
40
30
20
10
0
10
20
30
40
X

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, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

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)

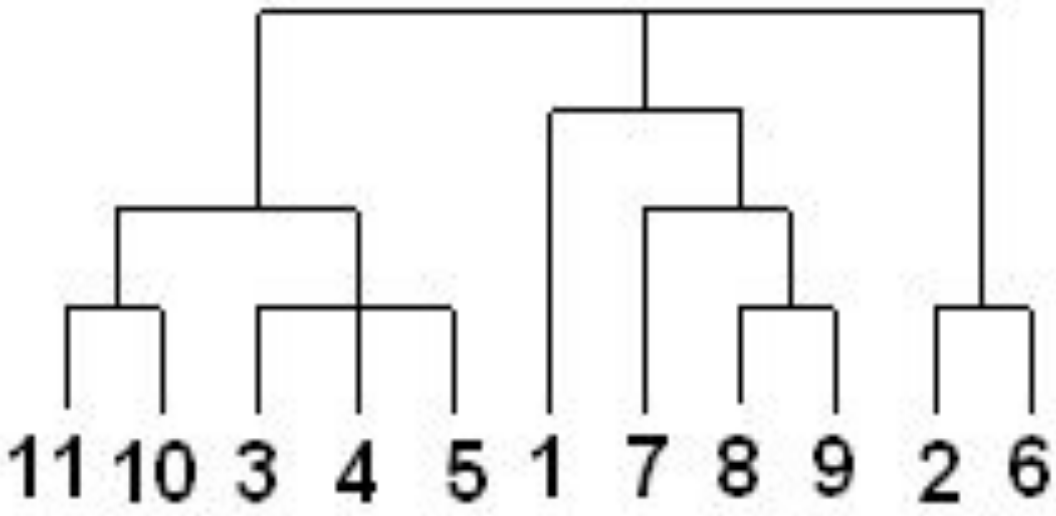
Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

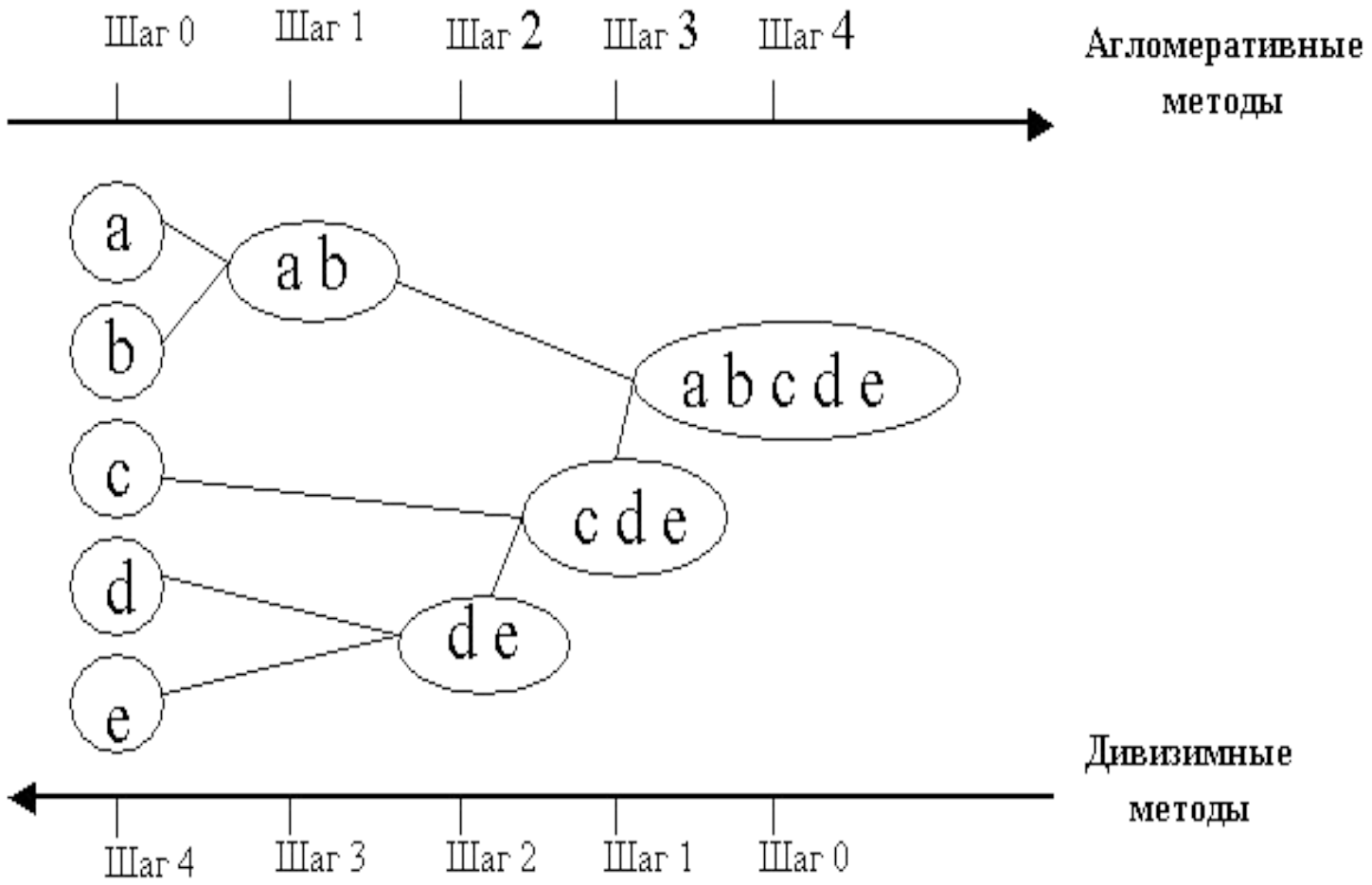
В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

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.


11 10 3 4 5 1 7 8 9 2 6


Шаг 0
Шаг 1
Шаг 2
Шаг 3
Шаг 4
Агломеративные методы
a
b
a b
c
d
e
c d e
d e
a b c d e
Дивизимные методы
Шаг 4
Шаг 3
Шаг 2
Шаг 1
Шаг 0

Методы поиска ассоциативных правил

Ассоциативное правило имеет вид: "Из события A следует событие B".

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

Основными характеристиками ассоциативного правила являются поддержка и достоверность правила.

Алгоритм AIS.

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

Алгоритм SETM

- Создание этого алгоритма было мотивировано желанием использовать язык SQL для вычисления часто встречающихся наборов товаров. Как и алгоритм AIS, SETM также формирует кандидатов "на лету", основываясь на преобразованиях базы данных. Чтобы использовать стандартную операцию объединения языка SQL для формирования кандидата, SETM отделяет формирование кандидата от их подсчета.

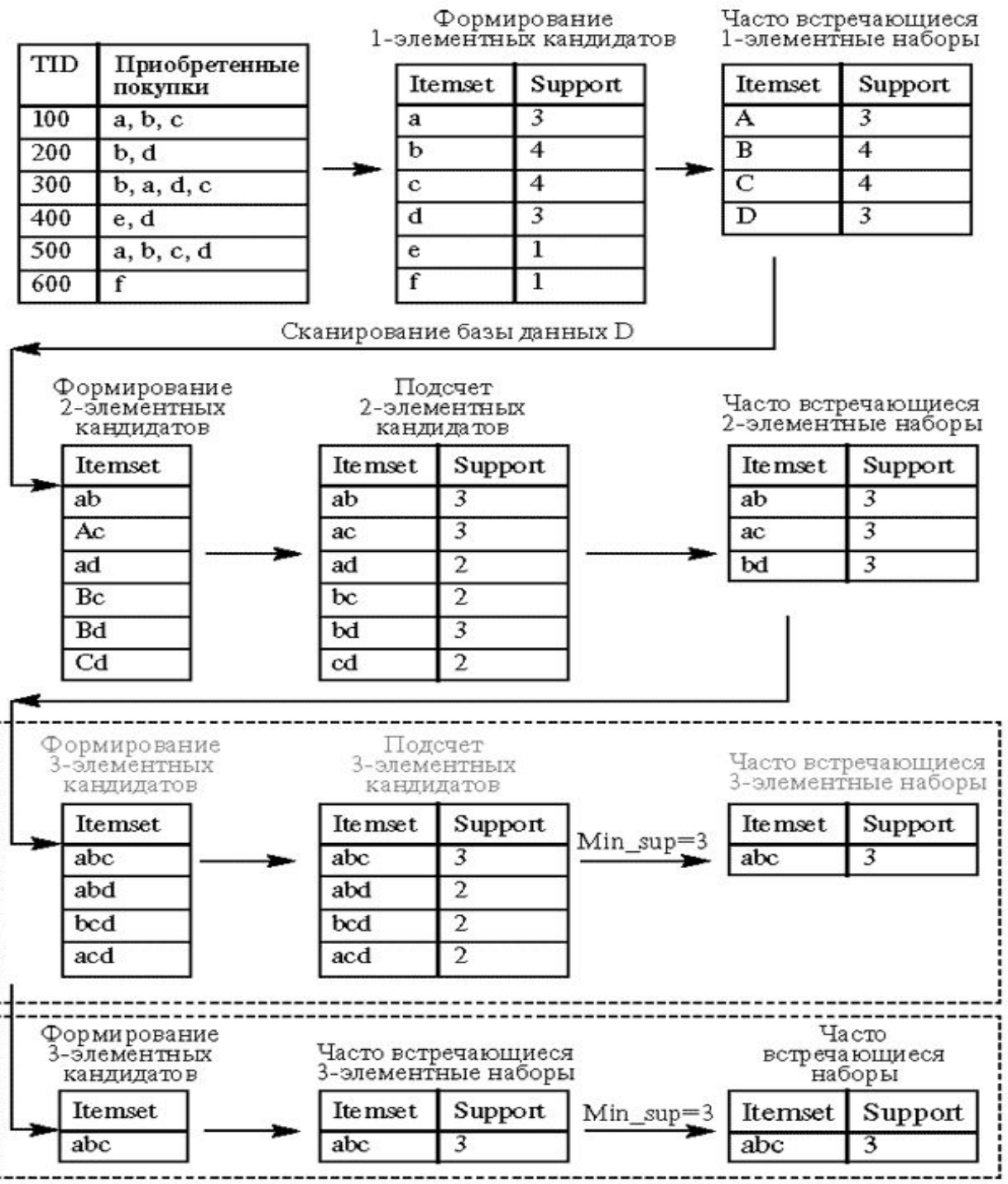
Неудобство алгоритмов AIS и SETM - излишнее генерирование и подсчет слишком многих кандидатов, которые в результате не оказываются часто встречающимися. Для улучшения их работы был предложен алгоритм Apriori

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.


TID
Приобретенные покупки
100 a, b, c
200 b, d
300 b, a, d, c
400 e, d
500 a, b, c, d
600 f
Формирование 1-элементных кандидатов
Itemset Support
a 3
b 4
c 4
d 3
e 1
f 1
Часто встречающиеся 1-элементные наборы
Itemset Support
A 3
B 4
C 4
D 3
Сканирование базы данных D
Формирование 2-элементных кандидатов
Itemset
ab
Ac
ad
Bc
Bd
Cd
Подсчет 2-элементных кандидатов
Itemset Support
ab 3
ac 3
ad 2
bc 2
bd 3
cd 2
Часто встречающиеся 2-элементные наборы
Itemset Support
ab 3
ac 3
bd 3
Формирование 3-элементных кандидатов
Itemset
abc
abd
bcd
acd
Подсчет 3-элементных кандидатов
Itemset Support
abc 3
abd 2
bcd 2
acd 2
Min_sup=3
Часто встречающиеся 3-элементные наборы
Itemset Support
abc 3
Формирование 3-элементных кандидатов
Itemset
abc
Часто встречающиеся 3-элементные наборы
Itemset Support
abc 3
Min_sup=3
Часто встречающиеся наборы
Itemset Support
abc 3

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

Интересная особенность этого алгоритма - то, что база данных D не используется для подсчета поддержки кандидатов набора товаров после первого прохода.

- ***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

- Один из них - **алгоритм DHP**, также называемый алгоритмом хеширования (J. Park, M. Chen and P. Yu, 1995 год).
- **PARTITION алгоритм** (A. Savasere, E. Omiecinski and S. Navathe, 1995 год).
- **Алгоритм DIC**, Dynamic Itemset Counting (S. Brin R. Motwani, J. Ullman and S. Tsur, 1997 год).

Системы для визуализации многомерных данных

К способам визуального или графического представления данных относят графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты и т.д.

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.

Методы визуализации

Методы визуализации, в зависимости от количества используемых измерений, принято классифицировать 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 представления 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.

Следует заметить, что наиболее естественно человеческий глаз воспринимает двухмерные представления информации.

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

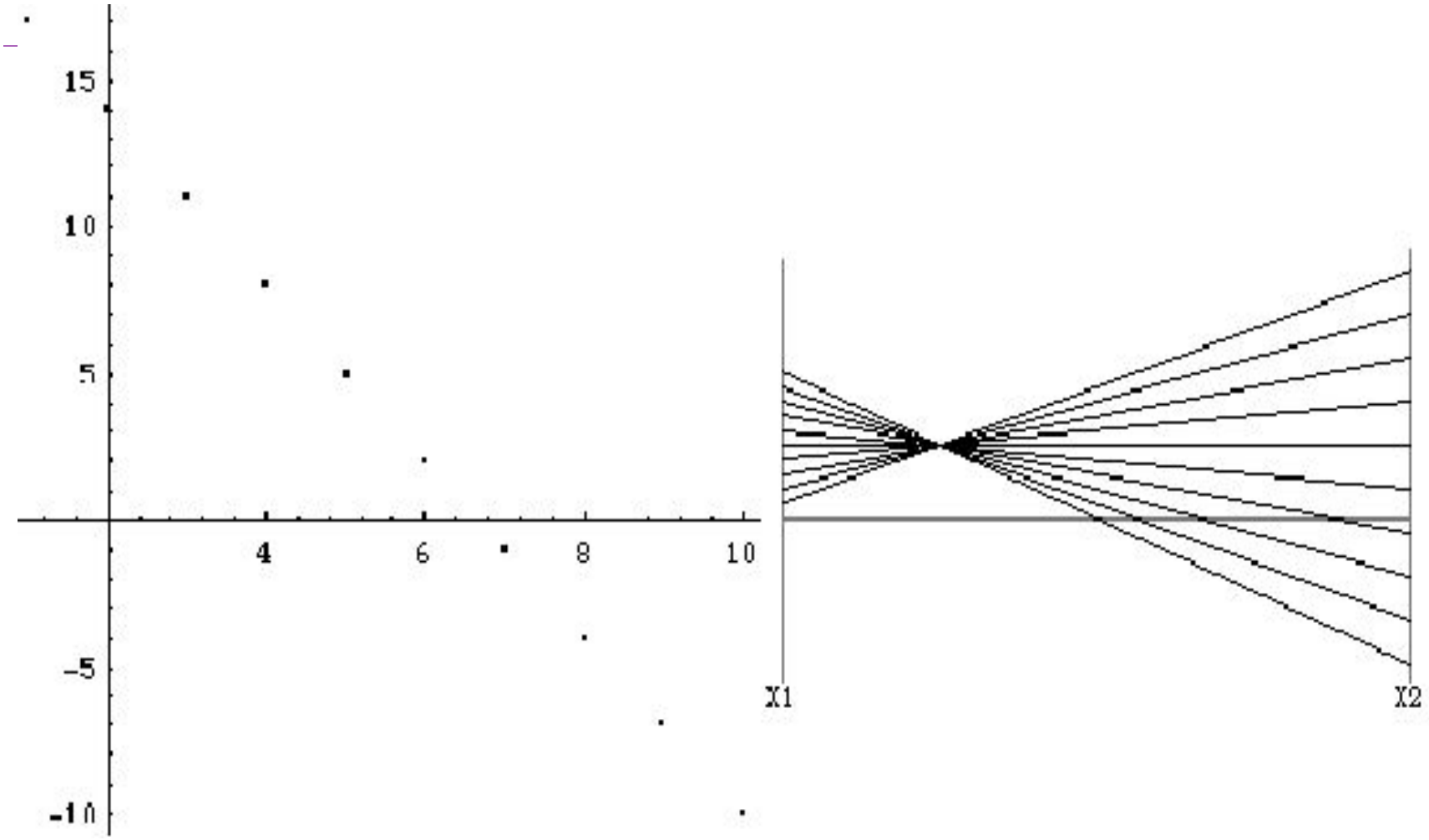
- его кластерную структуру и распределение объектов на классы (например, на диаграмме рассеивания);
- топологические особенности;
- наличие трендов;
- информацию о взаимном расположении данных;
- существование других зависимостей, присущих исследуемому набору данных.

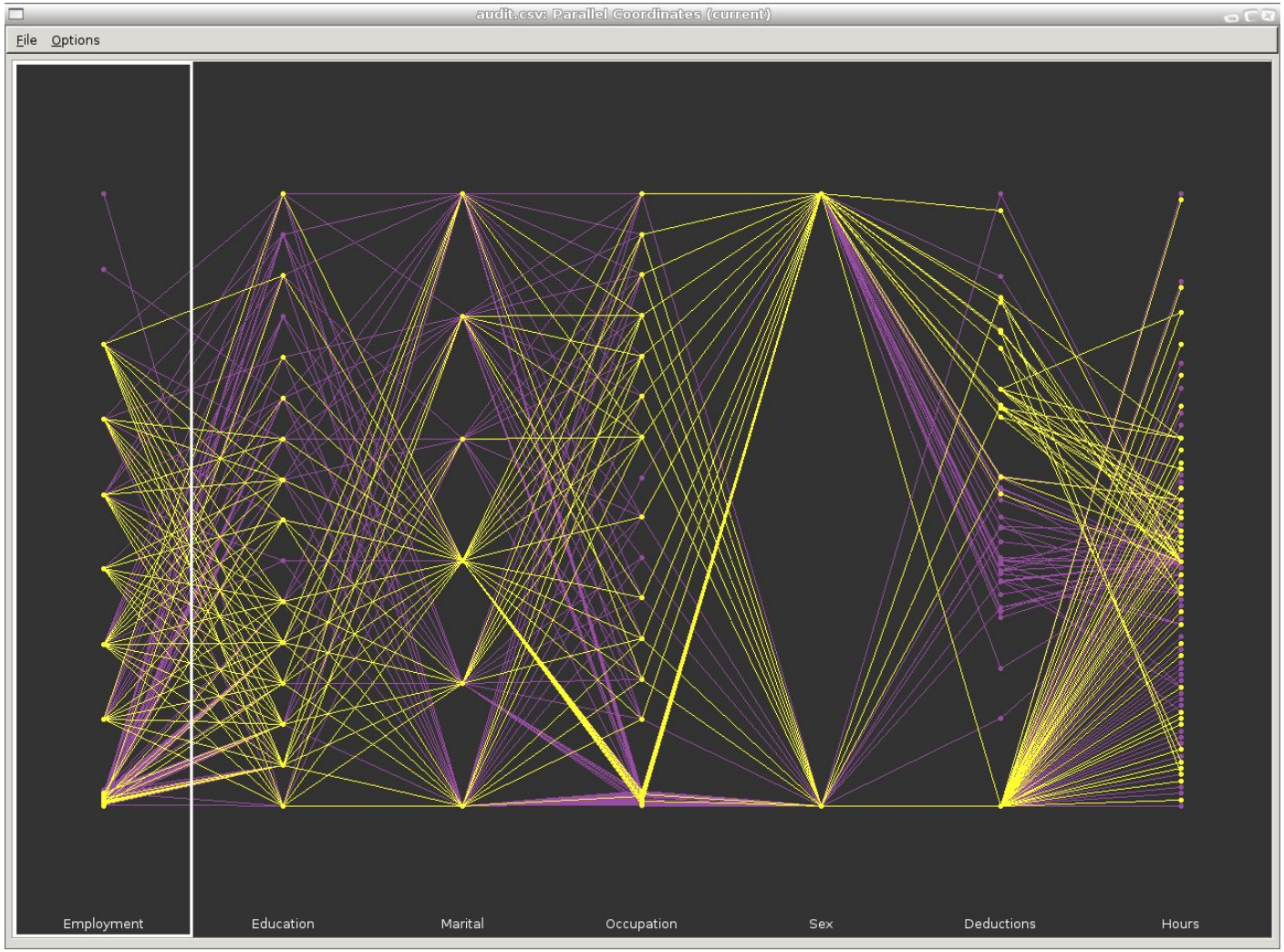
Представление данных в 4 х измерениях

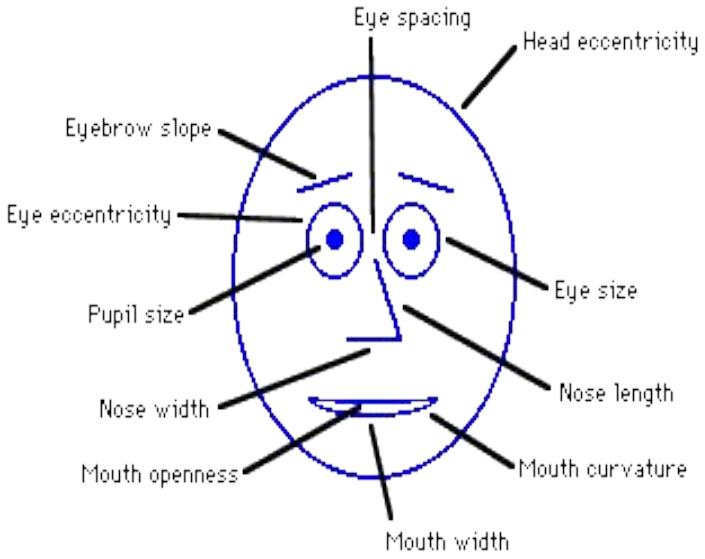
Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации.

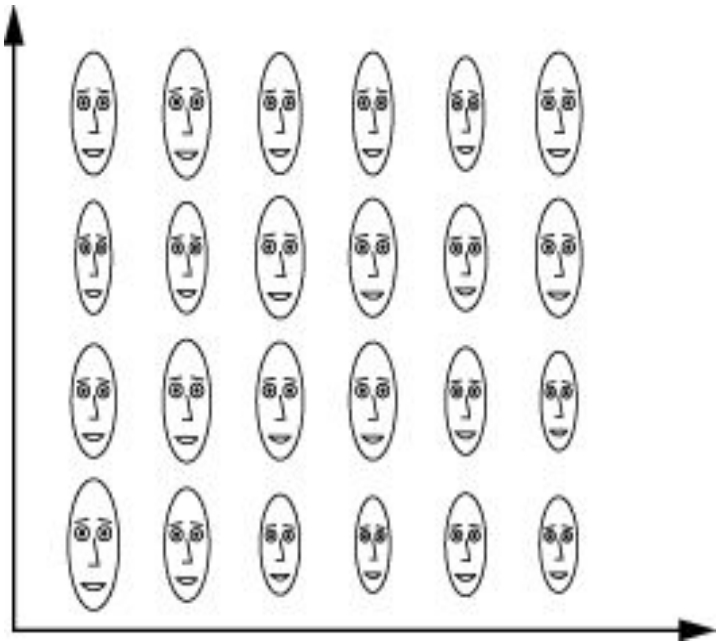
Наиболее известные способы многомерного представления информации:

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


15
10
5
-5
-10
4
6
8
10
X1
X2

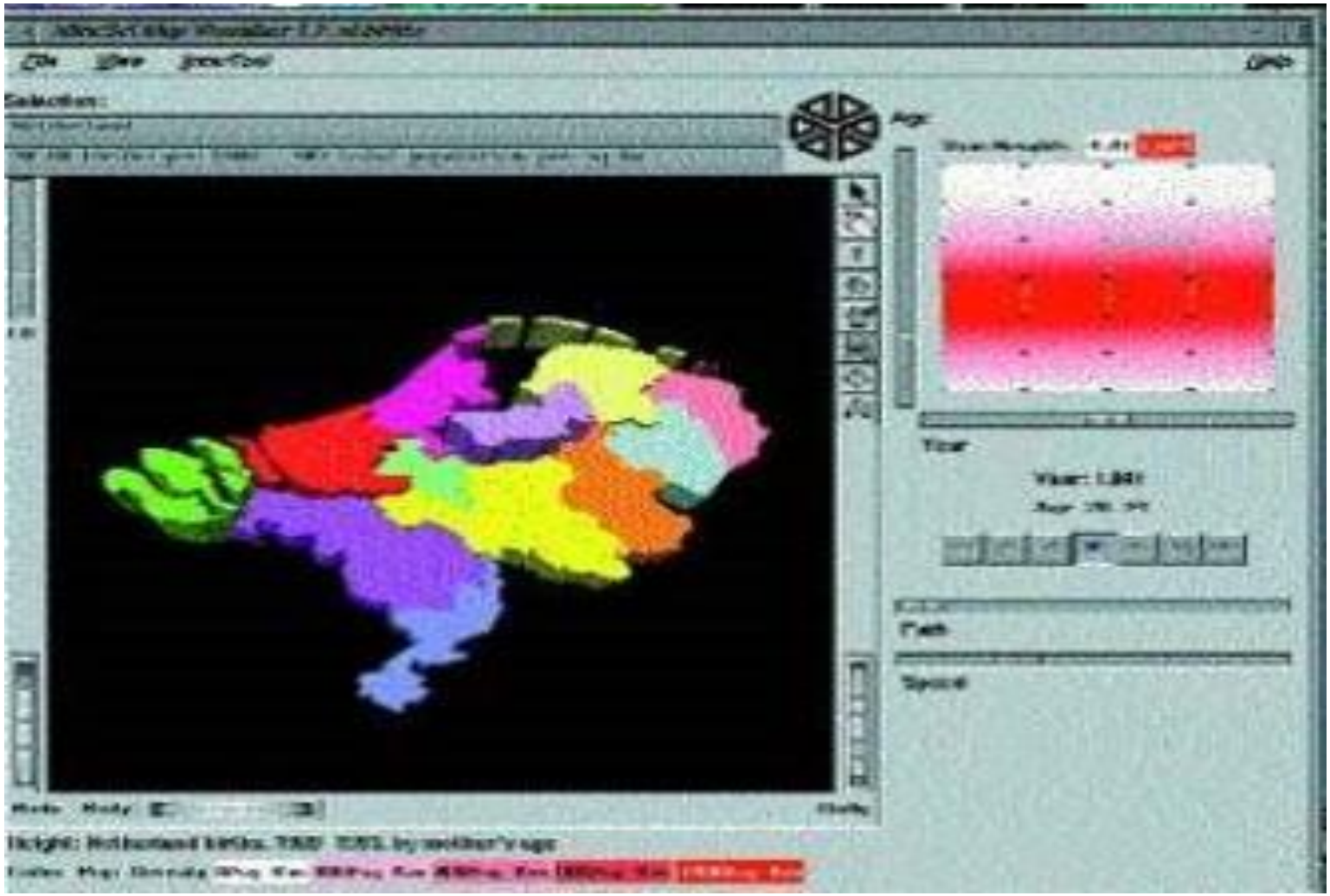

audit.csv: Parallel Coordinates (current)
File
Options
Employment
Education
Marital
Occupation
Sex
Deductions
Hours

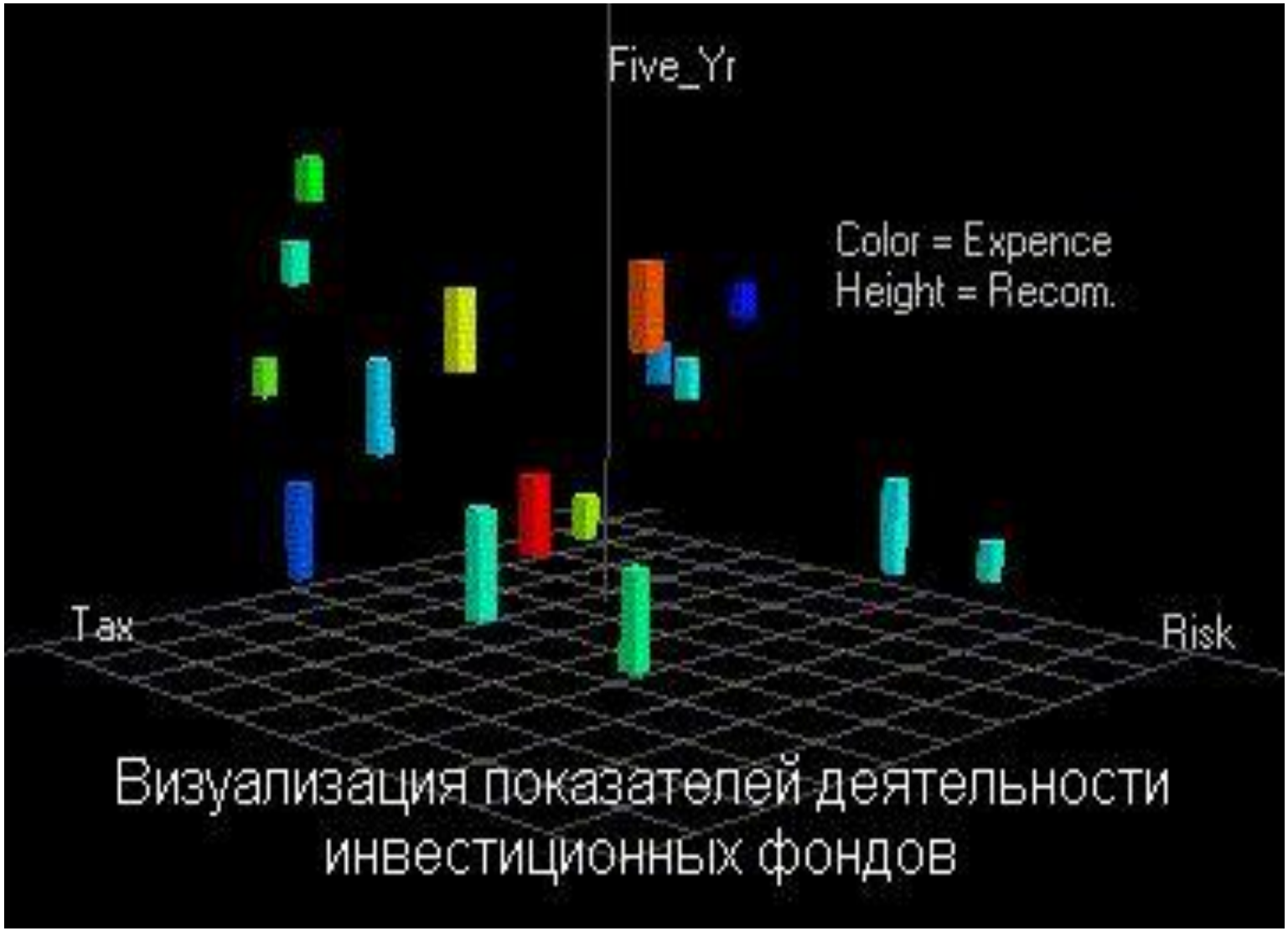

Eye spacing
Head eccentricity
Eyebrow slope
Eye eccentricity
Eye size
Pupil size
Nose length
Nose width
Mouth openness
Mouth curvature
Mouth width



Основные принципы компоновки визуальных средств представления информации:

- Принцип лаконичности.
- Принцип обобщения и унификации.
- Принцип акцента на основных смысловых элементах.
- Принцип автономности.
- Принцип структурности.
- Принцип стадийности.
- Принцип использования привычных ассоциаций и стереотипов





Визуализация показателей деятельности инвестиционных фондов