Задачи Data Mining. Информация и знания. Методы и стадии Data Mining

Понятие Data Mining

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

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.

Понятие Data Mining

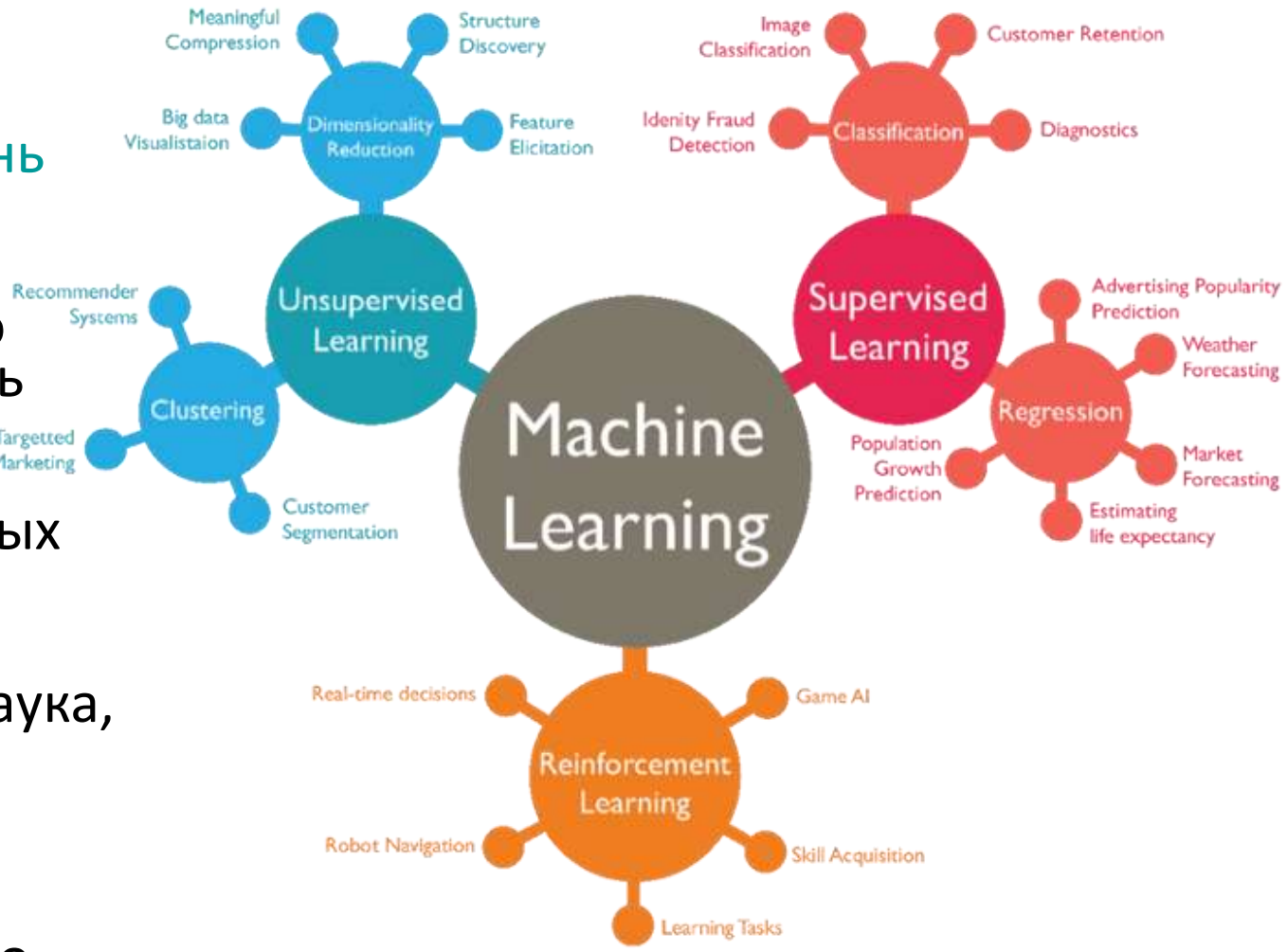
- **Неочевидных** - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

- **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.

Понятие Машинного обучения

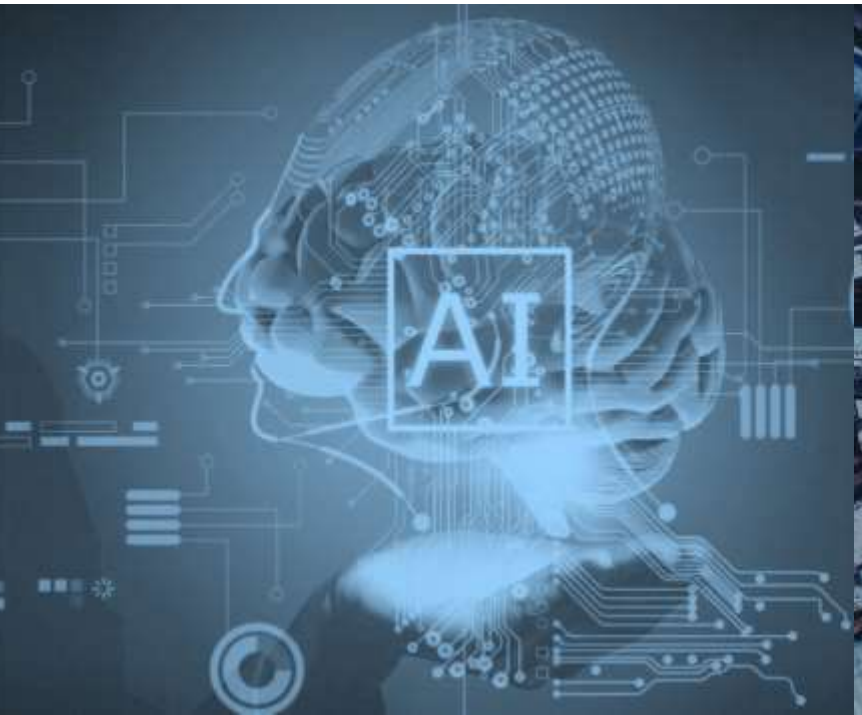
- Единого определения машинного обучения на сегодняшний день нет.
- **Машинное обучение** можно охарактеризовать как процесс получения программой новых знаний. в 1996
- «**Машинное обучение** - это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы» /Митчелл/

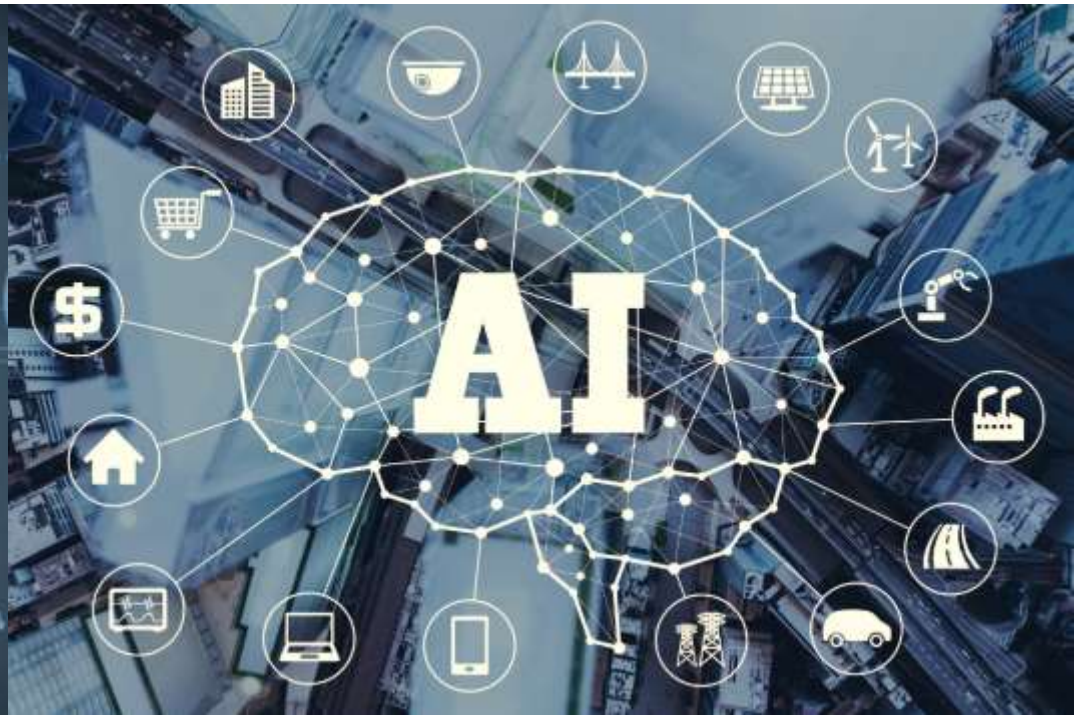


Одним из наиболее популярных примеров алгоритма машинного обучения являются нейронные сети.

Понятие Искусственного интеллекта

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





Отличия Data Mining от других методов анализа данных

- *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 неочевидные 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.
- OLAP больше подходит для понимания ретроспективных данных, **Data Mining опирается на ретроспективные данные** для получения ответов на вопросы о будущем.

Перспективы технологии Data Mining

- **выделение типов предметных областей с соответствующими им эвристиками**, формализация которых облегчит решение соответствующих задач Data Mining, относящихся к этим областям;
- **создание формальных языков и логических средств**, с помощью которых будут формализованы рассуждения и автоматизация которых станет инструментом решения задач Data Mining в конкретных предметных областях;
- **создание методов Data Mining**, **способных** не только извлекать из данных закономерности, но и **формировать некие теории, опирающиеся на эмпирические данные**;
- **преодоление существенного отставания возможностей инструментальных средств Data Mining** от теоретических достижений в этой области.

ЗАДАЧИ *DATA MINING*

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

- В технологии *Data Mining* гармонично следующие: *классификация*, *кластеризация*, *прогнозирование*, *ассоциация*, *визуализация*, *анализ* и обнаружение отклонений, *оценивание*, *анализ связей*, *подведение итогов*.

Классификация (Classification)

Наиболее простая и распространенная задача *Data* Mining.

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. Особенность кластеризации - 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*: *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)

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.

Последовательность определяется высокой вероятностью цепочки связанных во времени событий.
А*ссоциация* является частным случаем *последовательности* с временным шагом, равным нулю.

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)

Цель решения данной задачи - обнаружение и *анализ данных*, наиболее отличающихся от общего *множества* данных, выявление так называемых нехарактерных шаблонов.

Оценивание (Estimation)

Задача *оценивания* сводится к предсказанию непрерывных значений признака.

Анализ связей (Link Analysis)

Задача нахождения зависимостей в наборе данных.

Визуализация
(Visualization, Graph Mining)

В результате *визуализации* создается графический образ анализируемых данных. Для решения задачи *визуализации* используются графические методы, показывающие наличие закономерностей в данных.

Пример методов *визуализации* - *представление* данных в 2D и *3D* измерениях.

Классификация задач Data Mining

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

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

Категория *обучение с учителем* представлена: *классификация*, оценка, *прогнозирование*.
Категория *обучение без учителя* представлена задачей кластеризации.

Связь понятий

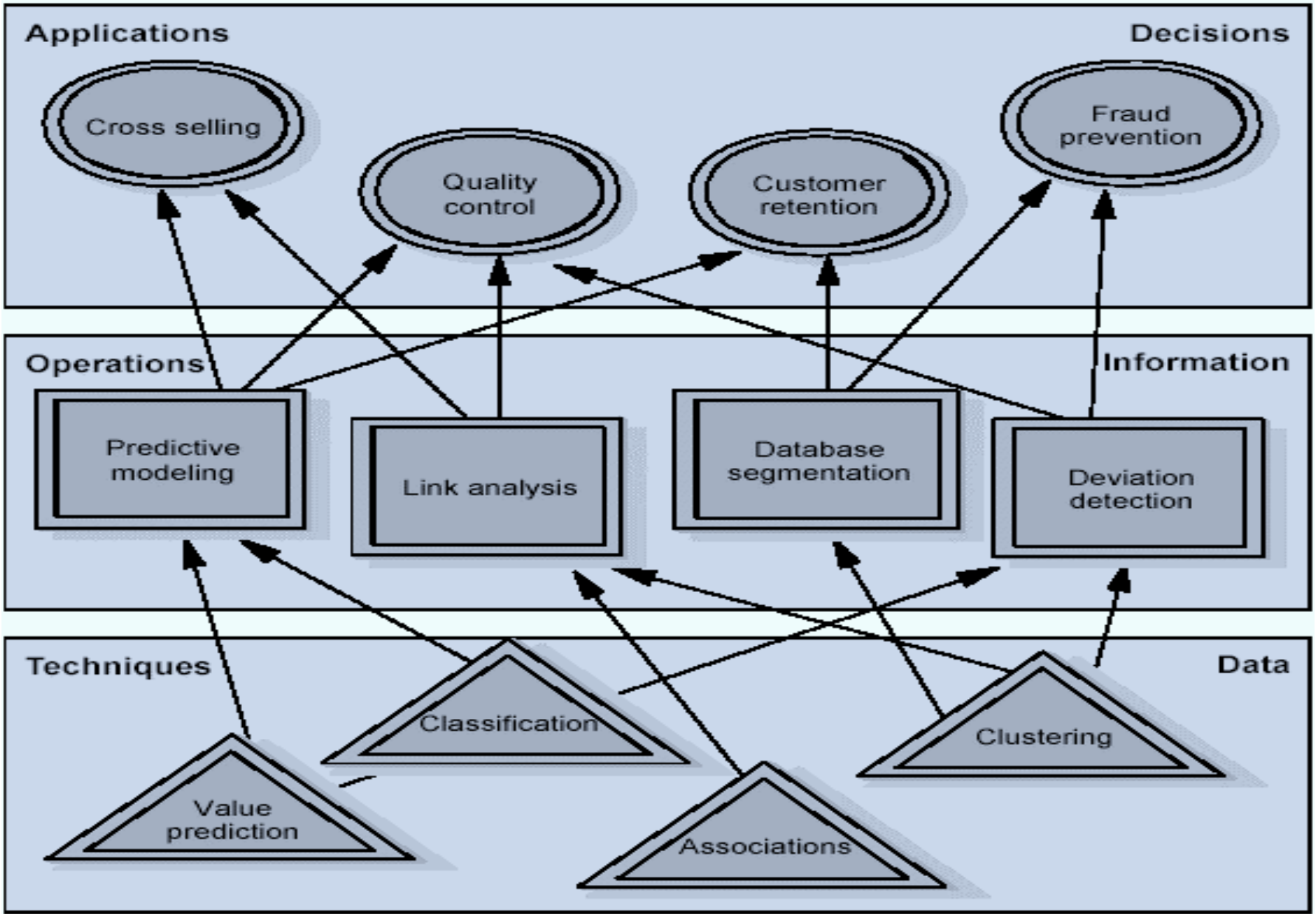
Главная ценность Data Mining - это практическая направленность данной технологии, путь от сырых данных к конкретному знанию, от постановки задачи к готовому приложению, при поддержке которого можно принимать решения.
Два потока:

1) ДАННЫЕ - ИНФОРМАЦИЯ - ЗНАНИЯ И РЕШЕНИЯ

2) ЗАДАЧИ - ДЕЙСТВИЯ И МЕТОДЫ РЕШЕНИЯ – ПРИЛОЖЕНИЯ

Эти потоки являются "двумя сторонами одной медали"

ОТ ЗАДАЧИ К ПРИЛОЖЕНИЮ



От задачи к приложению

- **Верхний - уровень приложений** - является уровнем бизнеса, на нем менеджеры принимают решения.

Приведенные примеры приложений: перекрестные продажи, *контроль* качества, удерживание клиентов.

- **Средний - уровень действий** - уровень *информации*, именно на нем выполняются действия *Data* Mining;

На рисунке действия: *прогностическое моделирование*, *анализ связей*, *сегментация* данных и другие.

- **Нижний - уровень определения *задачи*** *Data Mining*, которую необходимо решить применительно к данным, имеющимся в наличии;

Приведены задачи предсказания числовых значений, *классификация*, *кластеризация*, *ассоциация*.

Информация

Информация (лат. informatio) - любые сообщения о чем-либо; сведения, являющиеся объектом хранения, переработки и передачи (например генетическая информация);

в математике (кибернетике) - количественная мера устранения неопределенности (энтропия), мера организации системы;

в теории *информации* - раздел кибернетики, изучающий количественные закономерности, которые связаны со сбором, передачей, преобразованием и вычислением *информации*.

Информация

Информация - любые, неизвестные ранее сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций, для которых существует содержательная ***интерпретация*.**

Операции: восприятие, передача, преобразование, хранение и использование.

Свойства информации

- Полнота *информации*.
- Достоверность *информации*
- Ценность *информации*.
- Адекватность *информации*.
- Актуальность *информации*.
- Ясность *информации*.
- Доступность *информации*.
- Субъективность *информации*.

Понятие *информации* следует рассматривать только **при наличии источника и получателя** *информации*, а также канала связи между ними.

Требования, предъявляемые к информации

- ✓ Динамический характер *информации*.

  Информация существует только в момент взаимодействия данных и методов, т.е. в момент информационного процесса. Остальное время она пребывает в состоянии данных.

- ✓ Адекватность используемых методов.

  Информация возникает и существует в момент диалектического взаимодействия объективных данных и субъективных методов.

Знания

Знания - совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача.

По определению Денхема Грэя: «*Знания* - это абсолютное использование *информации* и данных, совместно с потенциалом практического опыта людей, способностями, идеями, интуицией, убежденностью и мотивациями».

Свойства:

Структурированность.

Удобство доступа и усвоения.

Лаконичность.

Непротиворечивость.

Процедуры обработки.

Одно из главных свойств знаний - возможность их передачи другим и способность делать выводы на их основе.

Сопоставление и сравнение понятий

- понятие *Data Mining* переводится на русский язык при помощи этих же трех понятий: как добыча **данных**, извлечение ***информации***, раскопка **знаний**.
- *Информация*, в отличие от данных, имеет смысл.
- Понятия "*информация*" и "*знания* ", с философской точки зрения, являются понятиями более высокого уровня, чем "данные", которое возникло относительно недавно.

Понятие " *информации* " непосредственно связано с сущностью процессов внутри информационной системы, тогда так понятие "*знание*" скорее ориентировано на качество процессов. Понятие "*знание*" тесно связано с процессом *принятия решений*

Это части одного потока: у истока его находятся **данные**, в процессе передачи которых возникает ***информация***, и в результате использования *информации*, при определенных условиях, возникают ***знания*.**

Выводы

- для получения ценных знаний необходимы качественные процедуры обработки.
- Процесс перехода от данных к *знаниям* занимает много времени и стоит дорого.
- Технология *Data* Mining с её мощными и разнообразными алгоритмами является инструментом, при помощи которого, продвигаясь вверх *по информационной пирамиде*, мы можем получать действительно качественные и ценные *знания*.

Основная особенность *Data Mining*

- это **сочетание** широкого математического инструментария (от классического статистического анализа до новых кибернетических *методов*);
- в технологии *Data Mining* гармонично объединились строго формализованные *методы* и *методы*

неформального анализа, т.е. **количественный и качественный *анализ* данных**.

Методы и алгоритмы Data Mining

- *искусственные нейронные сети*,
- деревья решений,
- символьные правила,
- *методы* ближайшего соседа и k-ближайшего соседа,
- *метод опорных векторов*,
- байесовские сети,
- линейная регрессия,
- корреляционно-регрессионный *анализ*;
- иерархические *методы* кластерного анализа,
- неиерархические *методы* кластерного анализа,
- *методы* поиска ассоциативных правил, в том числе *алгоритм* Apriori;
- метод ограниченного перебора,
- *эволюционное программирование* и генетические *алгоритмы*,
- разнообразные *методы* визуализации данных
- … и множество других *методов*.

Процесс *Data Mining*

СВОБОДНЫЙ ПОИСК →

ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ →

→ АНАЛИЗ ИСКЛЮЧЕНИЙ

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 методов Data Mining

Технологические методы Data Mining

Статистические методы Data mining

Кибернетические методы Data Mining

Классификация технологических методов Data Mining

Все *методы Data Mining* подразделяются на две большие группы по принципу работы с исходными обучающими данными.

В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после *Data Mining* либо они дистиллируются для последующего использования.

Технологические методы Data Mining

1. ***Непосредственное использование данных***,
или *сохранение данных*.
В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях *прогностического моделирования* и/или *анализа исключений*.

Проблема этой группы *методов* - могут возникнуть сложности анализа сверхбольших баз данных.

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

Технологические методы Data Mining

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

При технологии ***дистилляции шаблонов*** один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого *метода Data Mining*.

Этот процесс выполняется на стадии *свободного поиска*, у первой же группы *методов* данная стадия в принципе отсутствует.

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

Статистические методы Data mining

Арсенал статистических *методов Data Mining* классифицирован на четыре группы *методов*:

- Дескриптивный анализ и описание исходных данных.
- *Анализ связей* (корреляционный и регрессионный анализ, *факторный анализ*, *дисперсионный анализ*).
- Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ и др.).
- Анализ временных рядов (*динамические модели* и прогнозирование).

Кибернетические методы Data Mining

- ***искусственные нейронные*** *сети* (распознавание, кластеризация, прогноз);
- ***эволюционное программирование*** (в т.ч. *алгоритмы* метода группового учета аргументов);
- **генетические *алгоритмы*** (оптимизация);
- ***ассоциативная память*** (поиск аналогов, прототипов);
- **нечеткая логика**;
- **деревья решений**;
- **системы обработки экспертных знаний**.

Классификация по задачам Data Mining.

- Описательные *методы* служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.
- К методам, направленным на получение описательных результатов, относятся итеративные *методы* кластерного анализа, в том числе: *алгоритм* k-средних, k-медианы, иерархические *методы* кластерного анализа, *самоорганизующиеся карты* Кохонена и другие.

Классификация по задачам Data Mining.

- Прогнозирующие *методы* используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.
- К методам, направленным на получение прогнозирующих результатов, относятся такие *методы*: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод *опорных векторов* и др.

Свойства методов Data Mining

Среди основных свойств и характеристик *методов Data Mining* рассматривают следующие:

- **точность**,
- ***масштабируемость***,
- **интерпретируемость**,
- **проверяемость**,
- **трудоемкость**,
- **гибкость**,
- **быстрота** и
- **популярность**.

Выводы

- Каждый из *методов* имеет свои сильные и слабые стороны.
- Но **ни один метод**, какой бы не была его оценка с точки зрения присущих ему характеристик, **не может** обеспечить решение **всего спектра** задач *Data Mining*.
- Большинство инструментов *Data Mining*, **реализуют сразу несколько *методов***, например, деревья решений, индукцию правил и визуализацию, или же нейронные сети, *самоорганизующиеся карты* Кохонена и визуализацию.