Санкт-Петербургский государственный университет
аэрокосмического приборостроения
Кафедра информационных технологий предпринимательства

Задачи Data Mining. Информация и знания

Лектор

- Блюм Владислав Станиславович
- e-mail: vladblum7@gmail.com

Аннотация

- В лекции кратко описана основная суть задач Data Mining и их классификация. Подробно рассмотрены понятия "информация", "знания", а также дано сопоставление и сравнение этих понятий.

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.

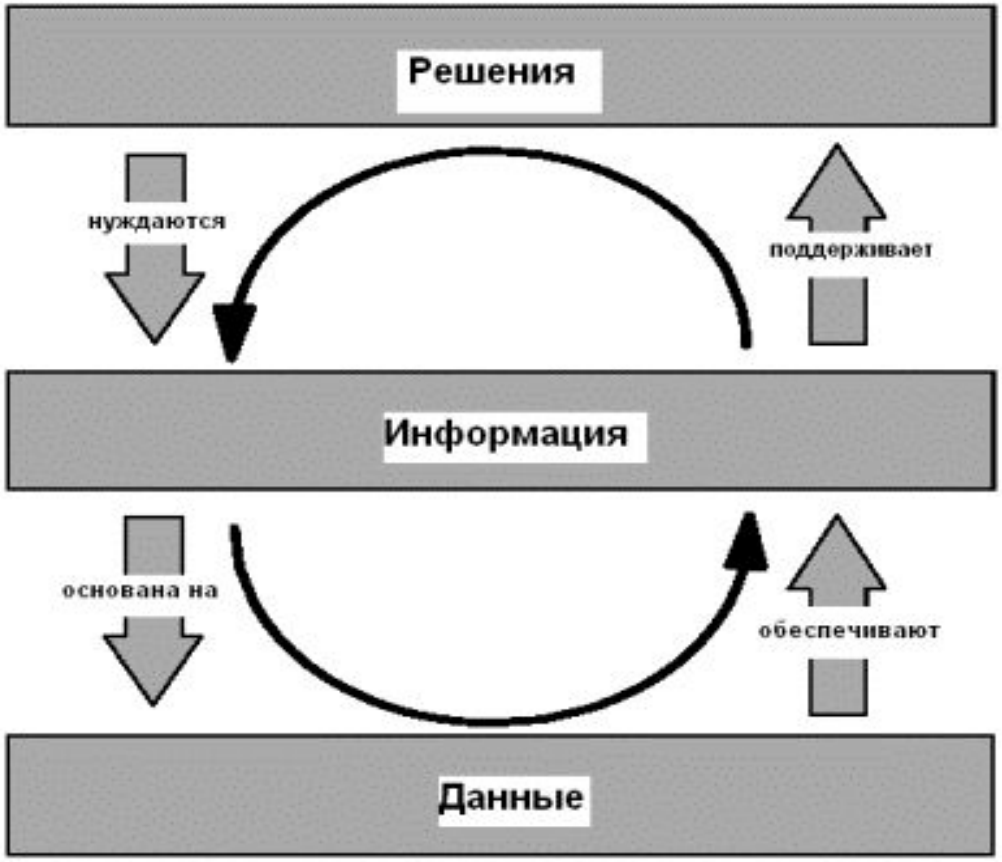
Два потока:

ДАННЫЕ - ИНФОРМАЦИЯ - ЗНАНИЯ И РЕШЕНИЯ

ЗАДАЧИ - ДЕЙСТВИЯ И МЕТОДЫ РЕШЕНИЯ – ПРИЛОЖЕНИЯ

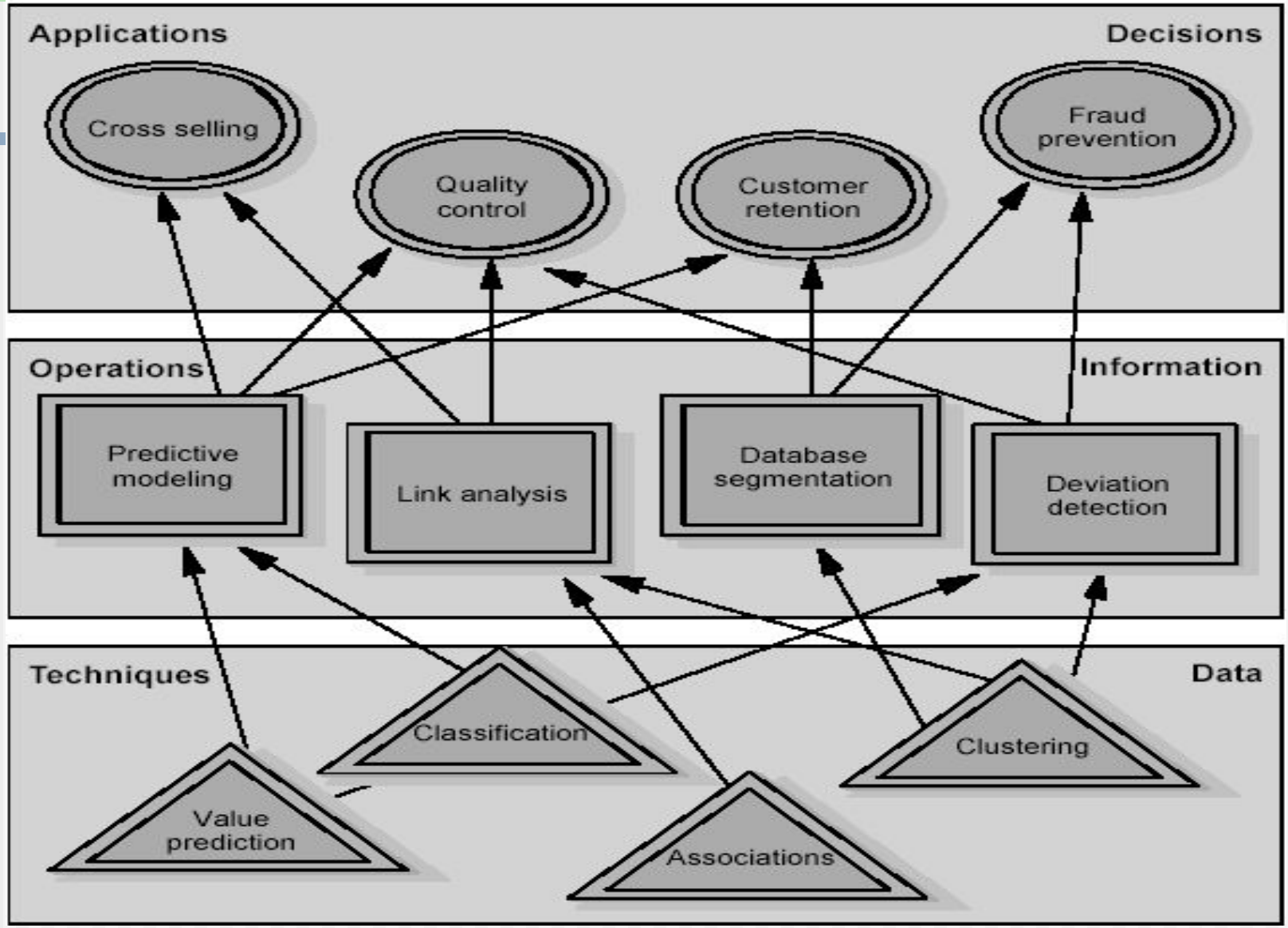
Эти потоки являются "двумя сторонами одной медали"

От данных к решениям (*информационная пирамида*)



ИАД. Анализ данных. Блюм В.С. 2016 г.

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.

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

Свойства информации

1. Полнота *информации*.
2. Достоверность *информации*
3. Ценность *информации*.
4. Адекватность *информации*.
5. Актуальность *информации*.
6. Ясность *информации*.
7. Доступность *информации*.
8. Субъективность *информации*.

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".

Знания имеют определенные 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.

Сопоставление и сравнение понятий

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*.