Технологии KDD и Data Mining [M.006]

Информационный подход к анализу получил распространение в таких методиках извлечения знаний, как **Knowledge Discovery in Databases (KDD)** и **Data Mining**. Сегодня на базе этих методик создается большинство прикладных аналитических решений в бизнесе и многих других областях.

Методика извлечения знаний

Несмотря на разнообразие бизнес-задач почти все они могут решаться по единой методике. Эта методика, зародившаяся в 1989 г., получила название Knowledge Discovery in Databases — извлечение знаний из баз данных. Она описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для обнаружения полезного знания. Методика не зависит от предметной области; это набор атомарных операций, комбинируя которые, можно получить нужное решение. KDD включает в себя этапы подготовки данных, выбора информативных признаков, очистки, построения моделей, постобработки и интерпретации полученных результатов. Ядром этого процесса являются методы Data Mining, позволяющие обнаруживать закономерности и знания (рисунок 1).

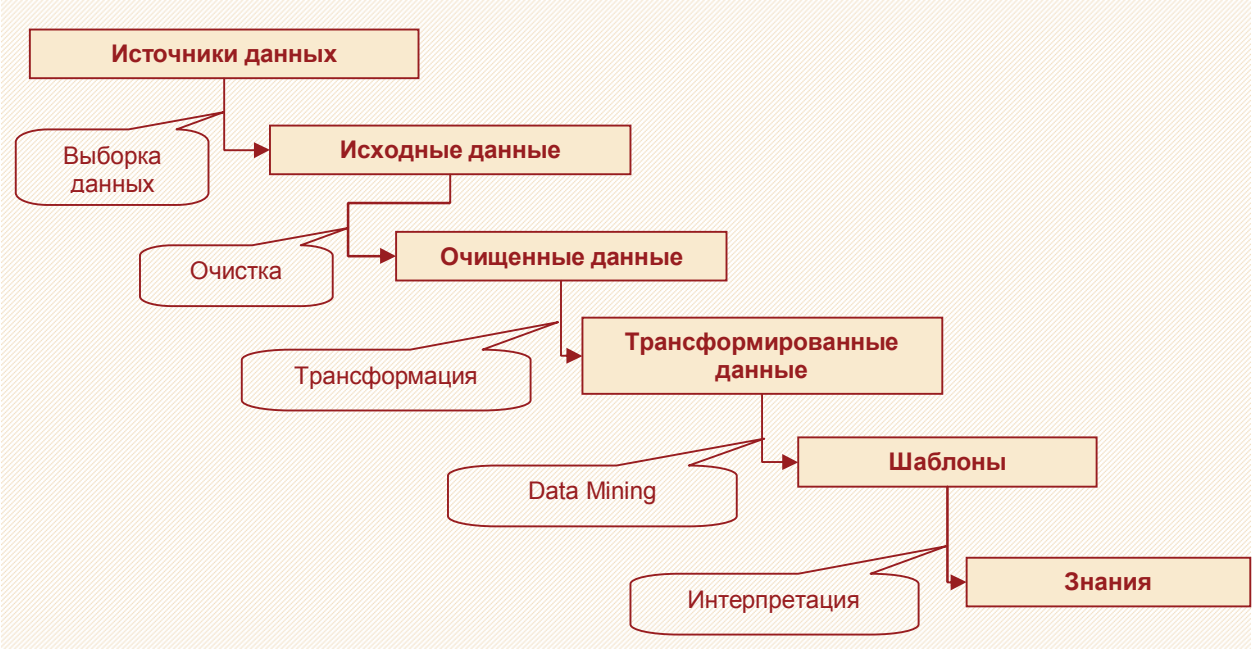

Рисунок 1 – Этапы KDD

Определение

> *Knowledge Discovery in Databases — процесс получения из данных знаний в виде зависимостей, правил, моделей, обычно состоящий из таких этапов, как выборка данных, их очистка и трансформация, моделирование и интерпретация полученных результатов.*

Кратко рассмотрим последовательность шагов, выполняемых на каждом этапе KDD.

Выборка данных. Первым шагом в анализе является получение исходной выборки. На основе отобранных данных строятся модели. Здесь требуется активное участие экспертов для выдвижения гипотез и отбора факторов, влияющих на анализируемый процесс. Желательно, чтобы данные были уже собраны и консолидированы. Крайне необходимы удобные механизмы подготовки выборки: запросы, фильтрация данных и **сэмплинг**. Чаще всего в качестве

источника рекомендуется использовать специализированное хранилище данных, консолидирующее всю необходимую для анализа информацию.

Очистка данных. Реальные данные для анализа редко бывают хорошего качества. Необходимость в предварительной обработке при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. К задачам очистки данных относятся: заполнение пропусков, подавление аномальных значений, сглаживание, исключение дубликатов и противоречий и пр.

Трансформация данных. Этот шаг необходим для тех методов, при использовании которых исходные данные должны быть представлены в каком-то определенном виде. Дело в том, что различные алгоритмы анализа требуют специальным образом подготовленных данных. Например, для прогнозирования необходимо преобразовать временной ряд при помощи скользящего окна или вычислить агрегированные показатели. К задачам трансформации данных относятся: скользящее окно, приведение типов, выделение временных интервалов, квантование, сортировка, группировка и пр.

Data Mining. На этом этапе строятся модели.

Интерпретация. В случае, когда извлеченные зависимости и шаблоны непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду. Для оценки качества полученной модели нужно использовать как **формальные методы**, так и знания аналитика. Именно аналитик может сказать, насколько применима полученная модель к реальным данным. Построенные модели являются, по сути, формализованными знаниями эксперта, а следовательно, их можно **тиражировать**. Найденные знания должны быть применимы и к новым данным с некоторой степенью достоверности.

Пример

Имеется сеть магазинов розничной торговли. Пусть требуется получить прогноз объемов продаж на следующий месяц. Первым шагом будет сбор истории продаж в каждом магазине и объединение ее в единый набор данных. Следующим шагом станет предобработка собранных данных: их группировка по месяцам, сглаживание кривой продаж, исключение из рассмотрения факторов, слабо влияющих на объемы продаж. Далее следует построить модель зависимости объемов продаж от выбранных факторов, после чего можно получить прогноз, подав на вход модели историю продаж. Зная прогнозное значение, его можно использовать, например, в приложениях для оптимизации товарных запасов.

Data Mining

Термин Data Mining дословно переводится как «добыча данных» или «раскопка данных» и имеет в англоязычной среде несколько определений.

Определение

Data Mining — обнаружение в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Зависимости и шаблоны, найденные в процессе применения методов Data Mining, должны быть нетривиальными и ранее неизвестными, например, сведения о средних продажах таковыми не являются. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

Нередко KDD отождествляют с Data Mining. Однако правильнее считать Data Mining шагом процесса KDD.

Data Mining — это не один метод, а совокупность большого числа различных методов обнаружения знаний. Существует несколько условных классификаций задач Data Mining. Мы будем говорить о четырех базовых классах задач.

1. **Классификация** – это установление зависимости *дискретной выходной переменной* от входных переменных.
2. **Регрессия** – это установление зависимости *непрерывной выходной переменной* от входных переменных.
3. **Кластеризация** – это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от других, которые вошли в другие кластеры.
4. **Ассоциация** – выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из события X следует событие Y. Такие правила называются *ассоциативными*. Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее называют анализом рыночной корзины (market basket analysis). Если же нас интересует последовательность происходящих событий, то можно говорить о *последовательных шаблонах* — установлении закономерностей между связанными во времени событиями. Примером такой закономерности служит правило, указывающее, что из события X спустя время t последует событие Y.

Кроме перечисленных задач, часто выделяют **анализ отклонений** (deviation detection), **анализ связей** (link analysis), **отбор значимых признаков** (feature selection), хотя эти задачи граничат с очисткой и визуализацией данных.

Задача классификации отличается от задачи регрессии тем, что в классификации на выходе присутствует переменная дискретного вида, называемая меткой класса. Решение задачи классификации сводится к определению класса объекта по его признакам, при этом множество классов, к которым может быть отнесен объект, известно заранее. В задаче регрессии выходная переменная является непрерывной — множеством действительных чисел, например сумма продаж (рисунок 2). К задаче регрессии сводится, в частности, прогнозирование временного ряда на основе исторических данных.

Рисунок 2 – Иллюстрация задачи классификации (слева) и задачи регрессии (справа)

Кластеризация отличается от классификации тем, что выходная переменная не требуется, а число кластеров, в которые необходимо сгруппировать все множество данных, может быть неизвестным. Выходом кластеризации является не готовый ответ (например, плохо/удовлетворительно/хорошо), а группы похожих объектов – кластеры. Кластеризация указывает только на схожесть объектов, и не более того. Для объяснения образовавшихся кластеров необходима их дополнительная интерпретация (рисунок 3).

Рисунок 3 – Иллюстрация задачи кластеризации

Перечислим наиболее известные применения этих задач в экономике.

Классификация используется, если заранее известны класс, например, при отнесении нового товара к той или иной товарной группе, клиента к какой-либо категории (при кредитовании – по каким-то признакам к одной из групп риска).

Регрессия используется для установления зависимостей между факторами. Например, в задаче прогнозирования зависимая величина – объемы продаж, а факторами, влияющими на нее, могут быть предыдущие объемы продаж, изменение курсов валют, активность конкурентов и т. д. Или, например, при кредитовании физических лиц вероятность возврата кредита зависит от личных характеристик человека, сферы его деятельности, наличия имущества.

Кластеризация может использоваться для сегментации и построения профилей клиентов. При достаточно большом количестве клиентов становится трудно подходить к каждому индивидуально, поэтому их удобно объединять в группы — сегменты с однородными признаками. Выделять сегменты можно по нескольким группам признаков, например по сфере деятельности, по географическому расположению. После кластеризации можно узнать, какие сегменты наиболее активны, какие приносят наибольшую прибыль, выделить характерные для них признаки. Эффективность работы с клиентами повышается благодаря учету их персональных предпочтений.

Ассоциативные правила помогают выявлять совместно приобретаемые товары. Это может быть полезно для более удобного размещения товара на прилавках, стимулирования продаж. Тогда человек, купивший пачку спагетти, не забудет купить к ней бутылочку соуса. Последовательные шаблоны могут использоваться при планировании продаж или предоставления услуг. Они похожи на ассоциативные правила, но в анализе добавляется временной показатель, то есть важна последовательность совершения операций. Например, если заемщик взял потребительский кредит, то с вероятностью 60 % через полгода он оформит кредитную карту.

Для решения вышеперечисленных задач используются различные методы и алгоритмы Data Mining. Ввиду того что Data Mining развивается на стыке таких дисциплин, как математика, статистика, теория информации, машинное обучение, теория баз данных, программирование, параллельные вычисления, вполне закономерно, что большинство алгоритмов и методов Data Mining были разработаны на основе подходов, применяемых в этих дисциплинах (рисунок 4).

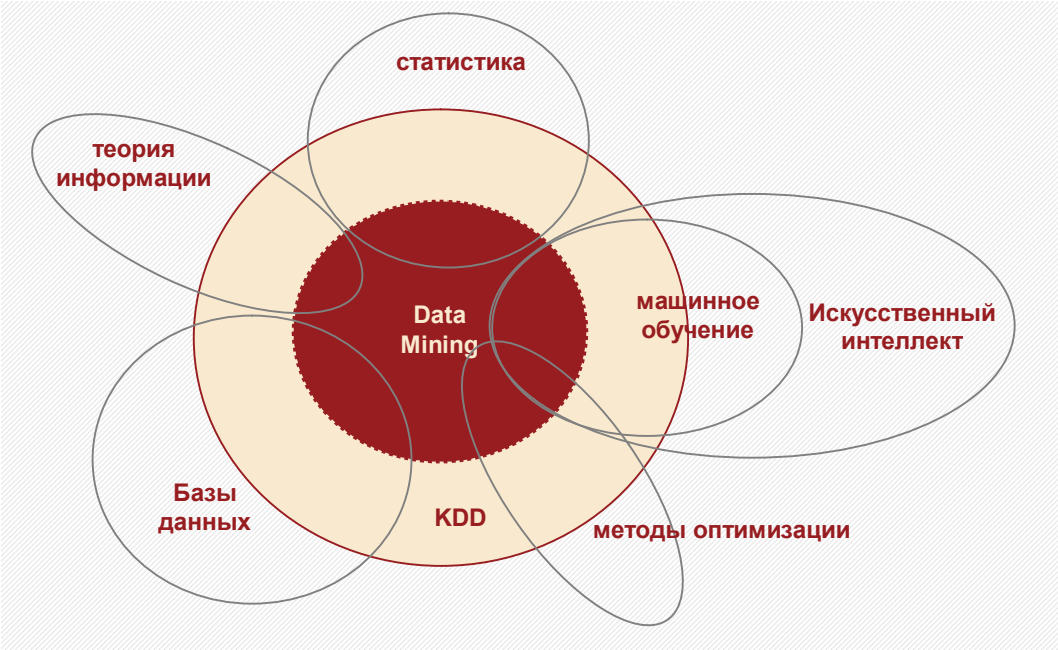

Рисунок 4 – Мультидисциплинарный характер Data Mining

В общем случае непринципиально, каким именно алгоритмом будет решаться задача, главное – иметь метод решения для каждого класса задач. На сегодняшний день наибольшее

распространение в Data Mining получили **методы машинного обучения**: деревья решений, нейронные сети, ассоциативные правила и т. д.

Определение

> *Машинное обучение (machine learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться на данных.*

Общая постановка задачи обучения следующая. Имеется множество *объектов* (ситуаций) и множество возможных *ответов* (откликов, реакций). Между ответами и объектами существует некоторая зависимость, но она неизвестна. Известна только конечная совокупность *прецедентов* — пар вида «объект — ответ», — называемая **обучающей выборкой**. На основе этих данных требуется обнаружить зависимость, то есть построить модель, способную для любого объекта выдать достаточно точный ответ. Чтобы измерить точность ответов, вводится критерий качества.

Решение подавляющего большинства бизнес-задач сводится к процессу KDD. Ранее были описаны базовые блоки, из которых собирается практически **любое бизнес-решение**. Рисунок 5 иллюстрирует некоторые популярные бизнес-задачи, которые решаются алгоритмами Data Mining.

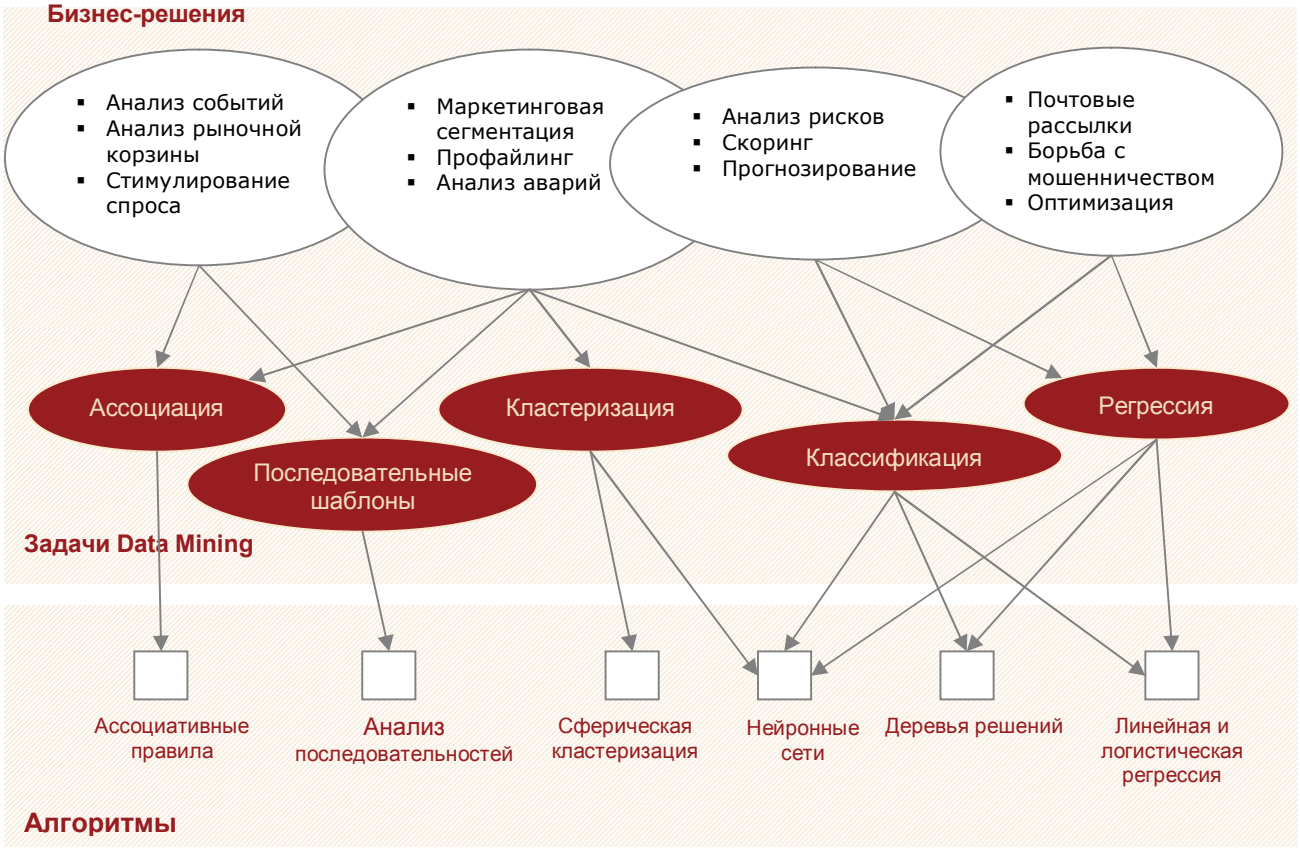

Рисунок 5 – От бизнес-решений к алгоритмам Data Mining

Отметим, что Data Mining не ограничивается алгоритмами решения упомянутых классов задач. Существует несколько современных подходов, которые «встраиваются» внутрь алгоритмов машинного обучения, придавая им новые свойства. Так, генетические алгоритмы призваны эффективно решать задачи оптимизации, поэтому их можно встретить в процедурах обучения нейронных сетей, карт Кохонена, логистической регрессии, при отборе значимых признаков. Математический аппарат нечеткой логики (fuzzy logic) также успешно включается в состав практически всех алгоритмов Data Mining; так появились нечеткие нейронные сети, нечеткие деревья решений, нечеткие ассоциативные правила. Объединение технологии хранилищ

данных и нечетких запросов позволяет аналитикам получать **нечеткие срезы**. И подобных примеров множество.

Причины распространения KDD и Data Mining

В KDD и Data Mining нет ничего принципиально нового. Специалисты в различных областях человеческого знания решали подобные задачи на протяжении нескольких десятилетий. Однако в последние годы интеллектуальная составляющая бизнеса стала возрастать, и для распространения технологий KDD и Data Mining были созданы все необходимые и достаточные условия.

1. Развитие технологий автоматизированной обработки информации создало основу для учета сколь угодно большого количества факторов и достаточного объема данных.
2. Возникла острая нехватка высококвалифицированных специалистов в области статистики и анализа данных. Поэтому потребовались технологии обработки и анализа, доступные для специалистов любого профиля за счет применения методов визуализации и самообучающихся алгоритмов.
3. Возникла объективная потребность в тиражировании знаний. Полученные в процессе KDD и Data Mining результаты являются формализованным описанием некоего процесса, а следовательно, поддаются автоматической обработке и повторному использованию на новых данных.
4. На рынке появились программные продукты, поддерживающие технологии KDD и Data Mining, – аналитические платформы. С их помощью можно создавать полноценные аналитические решения и быстро получать первые результаты.