УДК 004.334.7

А. Н. Рязанцев

ТЕХНОЛОГИИ KDD И DATA MINING

В настоящее время большое значение стало придаваться аналитическим системам, которые способствуют автоматизации процессов анализа и накопления информации, формированию готовых решений и многому другому. В данной работе рассмотрены технологии KDD и Data Mining, позволяющие существенно повысить коэффициент полезного действия от их внедрения.

Ключевые слова: технология, анализ, бизнес, методы, модели, Data Mining, KDD.

UDC 004.334.7

A.N. Ryazantsev

KDD TECHNOLOGIES AND DATA MINING

Now, analytical systems which promote analysis processes automation and information accumulation, formation of ready decisions and many other things are of great value. In the given work KDD technologies and Data Mining are considered, which essentially allow raising efficiency due to their introduction.

Keywords: technology, analysis, business, methods, models, Data Mining, KDD.

В настоящее время наблюдается довольно большой экспоненциальный рост рынка систем бизнес-аналитики. Они позволяют автоматизировать процессы анализа накопленной первичной информации, строить аналитические модели и на их базе получать готовые решения, а также применять их на практике. Еще одной важной частью аналитических систем является простота, эффективность и автоматизм.

Рассмотрим две современные и наиболее распространенные технологии, в основе которых лежит данная концепция, - это технология Data Mining и KDD – Knowledge Discovery in Databases.

Хотя существует довольно много бизнес-задач, но почти все они могут решаться по одной методике. Эта методика зародилась еще в 1989 г. и получила название Knowledge Discovery in Databases — извлечение знаний из баз данных. С помощью этой технологии описываются не конкретные алгоритмы или математический аппарат, а последовательность действий, которую нужно выполнить для обнаружения полезного знания. Эта методика не зависит от предметной области. Технология KDD включает в себя этапы выбора информативных признаков, построения моделей, подготовки данных, очистки, постобработки и интерпретации полученных результатов. Главной частью этого процесса являются методы Data Mining, которые позволяют обнаружить закономерности и знания (рис. 1).

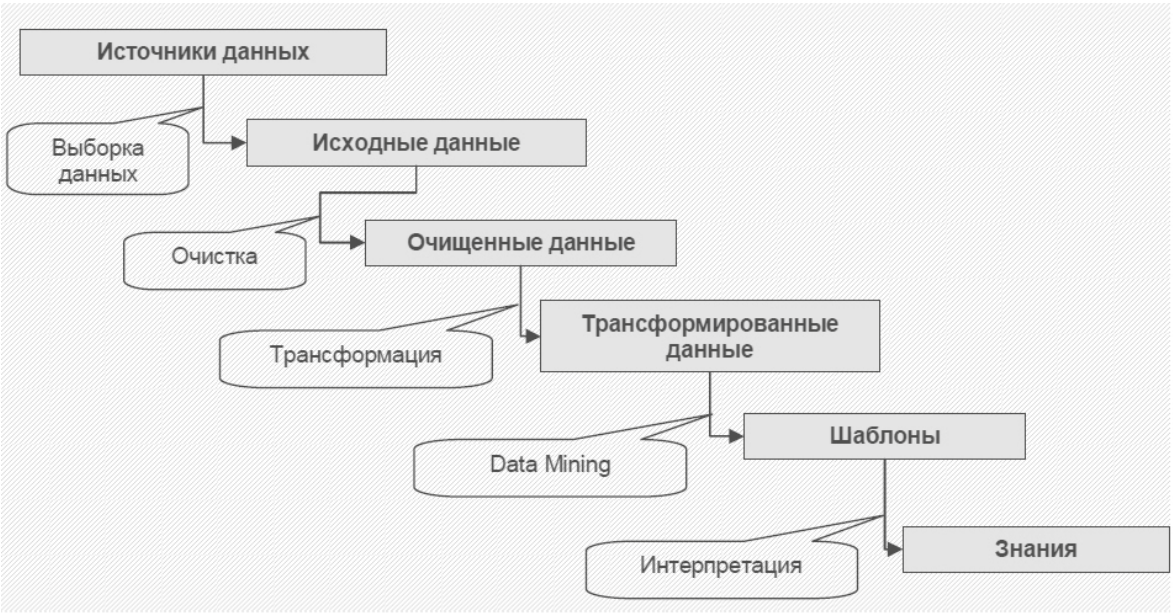

Рисунок 1 - Этапы KDD

Технология Knowledge Discovery in Databases — это 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.

Рассматривая технологию KDD, можно отметить, что первый этап состоит из процесса выборки данных. Здесь первым шагом в анализе является получение исходной выборки. После чего на основе отобранных данных строятся новые модели. Здесь требуется активное участие экспертов для выдвижения гипотез и отбора факторов, которые в свою очередь влияют на анализируемый процесс. Желательно, чтобы данные были уже собраны и консолидированы. Очень важны удобные механизмы для подготовки выборки: запросы, фильтрация данных и сэмплинг. Чаще всего в качестве источника используют специализированное хранилище данных, консолидирующее всю необходимую для анализа информацию.

На втором этапе производится очистка данных. Реальные данные требуют обработки для получения более полезных знаний. Необходимость в предварительной обработке при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Да же более того, эта задача может представлять самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. К задачам очистки данных относятся такие как: заполнение пропусков, подавление аномальных значений, сглаживание, исключение дубликатов и противоречий [1].

На третьем этапе рассматривается 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 можно отнести: скользящее окно, приведение типов, выделение временных интервалов, квантование, сортировка, группировка и пр.

Следующий четвертый этап это Data Mining. На этом этапе идет процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний.

На пятом, последнем этапе осуществляется 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.

Рассмотрим более подробно технологию Data Mining (рус. добыча данных, интеллектуальный анализ данных, глубинный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин введён Григорием Пиатецким-Шапиро в 1989 году. 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.

Существует несколько классификаций задач Data Mining. Из них выделят четыре базовых класса задач:

1. Классификация – установление зависимости дискретной выходной переменной от входных переменных.
2. Регрессия – установление зависимости непрерывной выходной переменной от входных переменных.
3. Кластеризация – 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 будет следовать событие Y [2].

Кроме перечисленных задач, еще выделяют отбор значимых признаков (feature selection), анализ отклонений (deviation detection), анализ связей (link analysis), 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 определенного класса объекта по его признакам, при этом множество классов, к которым может быть отнесен объект, известно заранее. В задачах регрессии выходная переменная является непрерывной, то есть множеством действительных чисел, например сумма продаж (рис. 2). К задаче регрессии сводится, в частности, прогнозирование временного ряда на основе исторических данных.

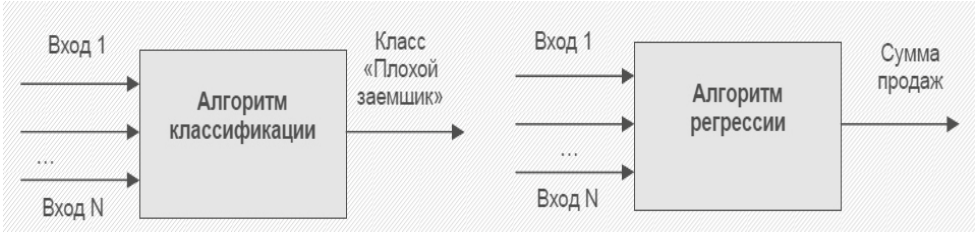

Рисунок 2 - Иллюстрация задачи классификации и задачи регрессии

Кластеризация отличается от классификации тем, что выходная переменная не требуется, а число кластеров, в которые необходимо сгруппировать все множество данных, может быть неизвестным.

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

– не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты;

– число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием;

– результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом;

Кластеризация указывает только на схожесть объектов, и не более того. Для объяснения образовавшихся кластеров необходима их дополнительная интерпретация (рис. 3).

Можно выделить несколько типов входных данных кластеризации:

К первому типу относятся признаковое описание объектов. Каждый объект описывается набором своих характеристик, называемых признаками. Признаки могут быть числовыми или нечисловыми.

Ко второму типу относят матрицу расстояний между объектами. Каждый объект описывается расстояниями до всех остальных объектов обучающей выборки.

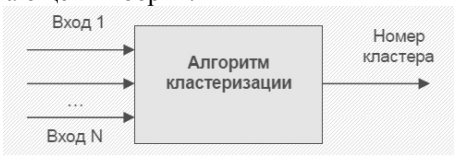

Рисунок 3 - Иллюстрация задачи кластеризации

Классификация используется если заранее известны классы, 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 довольно 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 [3].

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

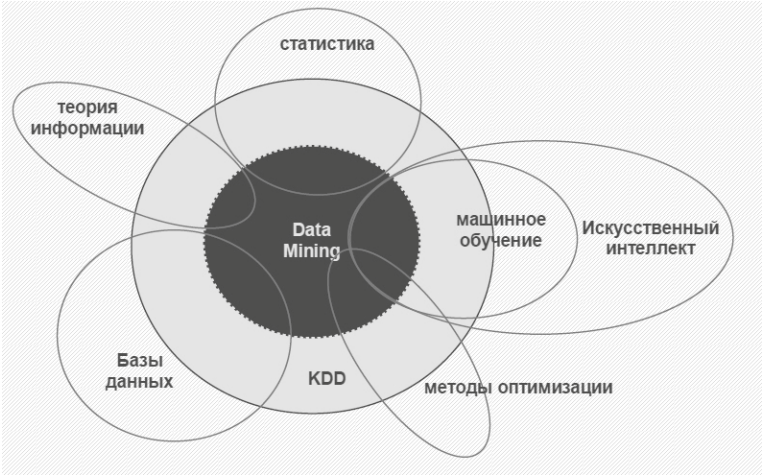

Рисунок 4 - Мультидисциплинарный характер Data Mining

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.

Машинное обучение (machine learning) — это 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. Выше были описаны базовые блоки, из которых собирается практически любое бизнес-решение. Рисунок 5 иллюстрирует некоторые популярные бизнес-задачи, которые решаются алгоритмами Data Mining.

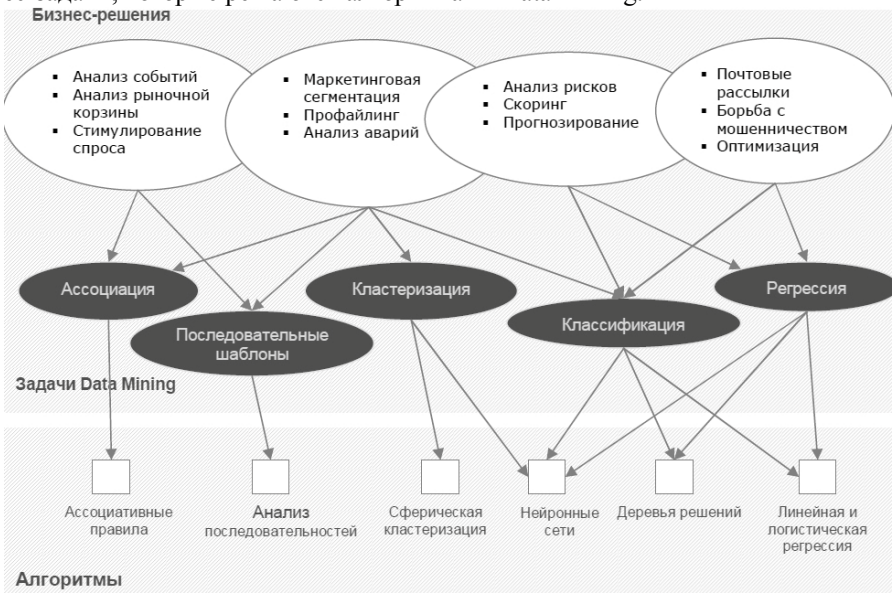

Рисунок 5 - От бизнес-решений к алгоритмам Data Mining

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.

Сфера применения Data Mining ничем не ограничена - она везде, где имеются какие-либо данные. Но в первую очередь, методы Data Mining сегодня, мягко говоря, заинтриговали коммерческие предприятия, развер-

тывающие проекты на основе информационных хранилищ данных (Data Warehousing). Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Например, известны сообщения об экономическом эффекте, в 10-70 раз превысившем первоначальные затраты от 350 до 750 тыс. дол. Известны сведения о проекте в 20 млн. дол., который окупился всего за 4 месяца. Другой пример - годовая экономия 700 тыс. дол. за счет внедрения Data Mining в сети универсамов в Великобритании.

Таким образом, технологии KDD и Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе. Так что можно смело говорить о том, что у таких технологий, как KDD и Data Mining, очень высокий потенциал развития, который приносит пользу, во много раз превышающую стоимости внедрения этих технологий, что, в свою очередь, приводит к быстрому росту и развитию компании.

Список литературы:

1. Фрегат-корпорация. Система автоматизации учета и управления предприятием. Анализ данных [Электронный ресурс]. - Режим доступа: http://www.frigat.ru/131/. Дата обращения 23.03.2012.

2. Разработка ПО для анализа и оптимизации деятельности агентства [Электронный ресурс]. - Режим доступа: http://www.masters.donntu.edu.ua/2007/fvti/kravchenko/diss/index.htm. Дата обращения 23.03.2012.

3. Комплексный подход к использованию средств бизнес-анализа [Электронный ресурс]. - Режим доступа: http://www.slideshare.net/alexgolder/bi-20110713. Дата обращения 23.03.2012.

4. Интеллектуальный анализ данных и извлечение знаний из данных [Электронный ресурс]. - Режим доступа: http://otherreferats.allbest.ru/emodel/00034033_0.html. Дата обращения 23.03.2012.

Рязанцев Александр Николаевич
студент 4 курса факультета управления
Орловский государственный институт экономики и торговли
e-mail:persik101@rambler.ru

Научный руководитель
Соколова Наталья Николаевна
к.э.н., доцент кафедры менеджмента
Орловского государственного института экономики и торговли.
e-mail:persik101@rambler.ru