ЛЕКЦИЯ 4. Понятие анализа данных

Анализ данных широкое понятие. Сегодня существуют десятки его определений. В самом общем смысле анализ данных это исследования, связанные с обсчетом многомерной системы данных, имеющей множество параметров. В процессе анализа данных исследователь производит совокупность действий с целью формирования определенных представлений о характере явления, описываемого этими данными как правило, для анализа данных используются различные математические методы.

Анализ данных нельзя рассматривать только как обработку информации после ее сбора. Анализ данных это прежде всего средство про верки гипотез и решения задач исследователя.

Известное противоречие между ограниченными познавательными способностями человека и бесконечностью Вселенной заставляет нас использовать модели и моделирование, тем самым упрощая изучение интересующих объектов, явлений и систем,

Слово «модель» (лат, modelium) означает «мера», «способ», «сходство с какой-то вещью». Построение моделей универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других задач. Основная цель моделирования в том, что модель должна достаточно хорошо отображать функционирование моделируемой системы.

Модель – объект или описание объекта, системы для замещения (при определенных условиях, предположениях, гипотезах) одной системы (то есть оригинала) другой системой для лучшего изучения оригинала или воспроизведения каких-либо его свойств.

Моделирование – универсальный метод получения, описания и использования знаний. Применяется в любой профессиональной деятельности.

Таким образом, анализ данных тесно связан с моделированием.

Аналитический подход к моделированию. Модель в традиционном понимании представляет собой результат отображения одной структуры (изученной) на другую (малоизученную). Так, отображая физическую систему (объект) на математическую (например, математический аппарат уравнений), получим физико-математическую модель системы, или математическую модель физической системы. Любая модель строится и исследуется при определенных допущениях, гипотезах. Делается это обычно с помощью математических методов.

Пример. Рассмотрим экономическую систему. Величина ожидаемого спроса s на будущий месяц $(t+1)$ рассчитывается на основе формулы $s(t+1) = [s(t) + s(t_1) + s(t_2)]/3$. то есть как среднее от продаж за предыдущие три месяца. Это простейшая математическая модель прогноза продаж. При построении этой модели были приняты следующие гипотезы.

Во-первых, годовая сезонность в продажах отсутствует.

Во-вторых, на величину продаж не влияют никакие внешние факторы: действия конкурентов, макроэкономическая ситуация и т. д.

Использовать такую модель легко: имея данные о продажах за предыдущие месяцы, по формуле мы получим прогноз на будущий месяц. Такой подход к моделированию в литературе называют аналитическим. Аналитический подход к моделированию базируется на том, что исследователь при изучении системы отталкивается от модели (рис. 1.1).

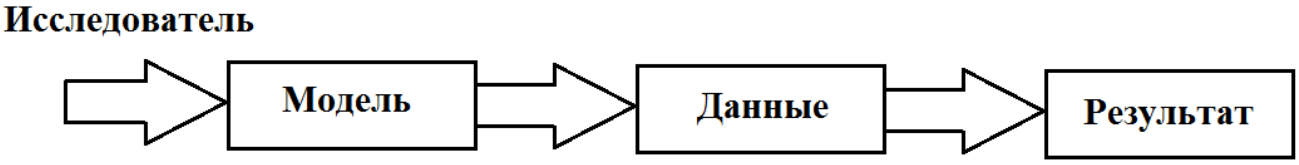

Рис. 1.1. Движение от модели к результату

В этом случае он по тем или иным соображениям выбирает подходящую модель. Как правило, это теоретическая модель, закон, известная зависимость, представленная чаще всего в функциональном виде (например, уравнение, связывающее выходной параметр у с входными воздействиями x_1, x_2). Варьирование входных параметров на выходе даст результат, который моделирует поведение системы в различных условиях.

Результат моделирования может соответствовать действительности, а может и не соответствовать. В последнем случае исследователю ничего не остается, кроме как выбрать другую модель или другой метод ее исследования. Новая модель, возможно, будет более адекватно описывать рассматриваемую систему.

При аналитическом подходе не модель «подстраивается» под действительность, а мы пытаемся подобрать существующую аналитическую модель таким образом, чтобы она адекватно отражала реальность.

Модель всегда исследуется каким-либо методом (численным, качественным и т. п.). Поэтому выбор метода моделирования часто означает выбор модели.

Информационный подход к моделированию.

При использовании традиционного аналитического подхода неизбежно возникнут проблемы из-за несоответствия между методами анализа и

реальностью, которую они призваны отражать. Существуют трудности, связанные с формализацией бизнес процессов. Здесь факторы, определяющие явления, столь многообразны и многочисленно, их взаимосвязи так «переплетены, что почти никогда не удается создать модель, удовлетворяющую таким же условиям, Простое наложение известных аналитических методов, законов, зависимостей на изучаемую картину реальности не принесет успеха.

В сложности и слабой формализации бизнес процессов главным образом «виноват» человеческий фактор, поэтому бывает трудно судить о характере закономерностей априори (а иногда и апостериори (знание, полученное из опыта в противоположность априори), после реализации какого-либо математического метода). С одинаковым успехом описывать эти закономерности могут различные модели. Использование разных методов для решения одной и той же задачи нередко приводит исследователя к противоположным выводам. Какой метод выбрать? Получить ответ на подобный вопрос можно, лишь глубоко проанализировав как смысл решаемой задачи, так и свойство используемого математического аппарата.

Поэтому в последние годы получил распространение информационный подход к моделированию, ориентированный на использование данных, Его цель освобождение аналитика от рутинных операций и возможных сложностей в понимании и применении современных математических методов.

При информационном подходе реальный объект рассматривается как «черный ящик», имеющий ряд входов и выходов, между которыми моделируются некоторые связи. Иными словами, известна только структура модели (например, нейронная сеть, линейная регрессия), а сами параметры модели «подстраиваются» под данные, которые описывают поведение объекта. Для корректировки параметров модели используется обратная связь – отклонение результата моделирования от действительности, а процесс настройки модели часто носит итеративный (то есть цикличный) характер (рис, 1.2).

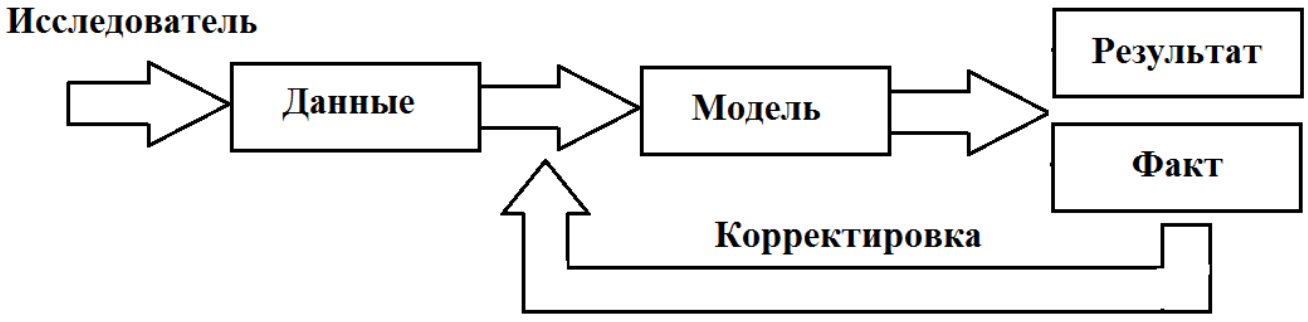

Рис. 1.2. Построение модели от данных

Таким образом, при информационном подходе отправной точкой являются данные, характеризующие исследуемый объект, и модель «подстраивается» под действительность.

Модели, полученные с помощью информационного подхода, учитывают специфику моделируемою объекта, явления, в отличие от аналитического подхода. Для бизнеса процессов последнее качество очень важно, поэтому информационный подход лег в основу большинства современных промышленных технологий и методов анализа данных: Knowledge Discovery in Databases (KDD – извлечение знаний из баз данных), Data Mining (DM – добыча данных, интеллектуальный анализ данных, глубинный анализ данных), машинного обучения.

Однако концепция моделей от данных требует тщательного подхода к качеству исходных данных, поскольку ошибочные, аномальные и зашумленные данные могут привести к моделям и выводам, не имеющим никакого отношения к действительности. Поэтому в информационном моделировании важную роль играют консолидация данных, их очистка и обогащение.

Модель, построенная на некотором множестве данных, описывающих реальный объект или систему, может оказаться не работающей на практике, поэтому в информационном моделировании используются специальные приемы: разделение данных на обучающее и тестовое множества, оценка обучающей и обобщающей способностей модели, проверка предсказательной силы модели.

В дальнейшем, говоря об анализе данных, мы будем предполагать использование именно информационного подхода, поскольку данные могут быть представлены в различной форме, круг рассмотрения будет ограничен областью структурированных данных. Инструментальной поддержкой процесса построения моделей на основе информационного подхода выступают современные технологии анализа данных KDD (извлечение знаний из баз данных) и Data Mining, а средством построения прикладных решений в области анализа.

Этапы моделирования. Построение моделей универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других важных задач. Но самое главное: полученные таким образом знания можно тиражировать.

Тиражирование знаний – совокупность методологических и инструментальных средств создания моделей. которые обеспечивают конечным пользователям возможность использовать результаты моделирования для принятия решений без необходимости понимания методик, при помощи которых эти результаты получены.

Процесс построения моделей состоит из нескольких шагов (рис, 1.3),

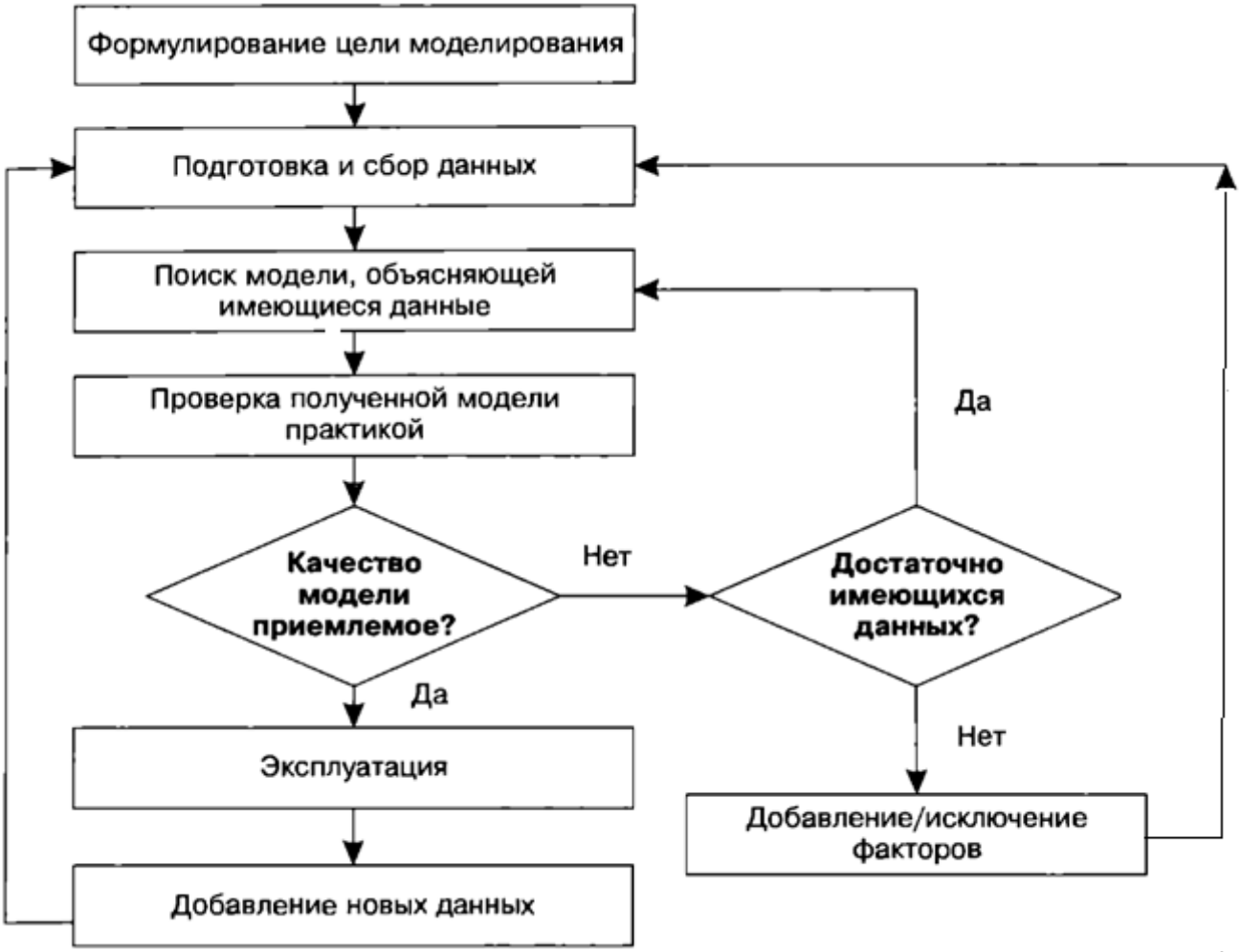

Рис. 1.3. Процесс построения модели

Формулирование цели моделирования. При построении модели следует отталкиваться от задачи, которую можно рассматривать как получение ответа на интересующий заказчика вопрос. Например, в розничной торговле к таким вопросам относятся следующие:

- Какова структура продаж за определенный период?
- Какие клиенты приносят наибольшую прибыль?
- Какие товары продаются или заказываются вместе?
- Как оптимизировать товарные остатки на складах?

В этом случае можно говорить о создании модели прогнозирования продаж, модели выявления ассоциаций и т. д, Данный этап также называют анализом проблемной ситуации.

Подготовка и сбор данных. Информационный подход к моделированию основан на использовании данных, подготовить и систематизировать которые – отдельная задача, Принципам подготовки данных, а также их очистке и обогащение Вы должны были проходить по предмету «Базы данных».

Поиск модели. После сбора и систематизации данных переходят к поиску модели, которая объясняла бы имеющиеся данные, позволила бы добиться эмпирически обоснованных ответов на интересующие вопросы. В промышленном анализе данных предпочтение отдается самообучающимся алгоритмам, машинному обучению, методам Data Mining.

Если построенная модель показывает приемлемые результаты на практике (например, в тестовой эксплуатации), ее запускают в промышленную эксплуатацию. Так, при тестовой эксплуатации скоринговой модели (система оценки клиентов, в основе которой заложены статистические методы), рассчитывающей кредитный рейтинг клиента и принимающей решение о выдаче кредита, каждое решение может подтверждаться человеком кредитным экспертом. При запуске скоринга в промышленную эксплуатацию человеческий фактор удаляется теперь решение принимает только компьютер,

Если качество модели неудовлетворительное, то процесс построения модели повторяется, как это показано на рис. 1.3.

Моделирование позволяет получать новые знания, которые невозможно извлечь каким-либо другим способом. Кроме того, полученные результаты представляют собой формализованное описание некого процесса, вследствие чего поддаются автоматической обработке. Однако результаты, полученные при использовании моделей, очень чувствительны к качеству данных, к знаниям аналитика и экспертов и к формализации самого изучаемого процесса. К тому же почти всегда имеются случаи, не укладывающиеся ни в какие модели.

На практике подходы комбинируются. Например, визуализация данных наводит аналитика на некоторые идеи, которые он пробует проверить при помощи различных моделей, а к полученным результатам применяются методы визуализации.

Полнофункциональная система анализа не должна замыкаться на применении только одного подхода или одной методики. Механизмы визуализации и построения моделей должны дополнять друг друга. Максимальную отдачу можно получить, комбинируя методы и подходы к анализу данных.

Структурированные данные.

Данные, описывающие реальные объекты, процессы и явления, могут быть представлены в различных формах и иметь разные тип и вид.

Данные сведения. которые характеризуют систему. явление. процесс или объект. представленные в определенной форме и предназначенные для дальнейшего использования.

По степени структурированности выделяют следующие формы представления данных:

- неструктурированные;
- структурированные;
- слабоструктурированные.

к неструктурированным относятся данные, произвольные по форме, включающие тексты и графику, мультимедиа (видео, речь, аудио). Эта форма представления данных широко используется, например, в Интернете, а сами данные представляются пользователю в виде отклика поисковыми системами.

Структурированные данные отражают отдельные факты предметной области. Структурированными называются данные, определенным образом упорядоченные и организованные с целью обеспечения возможности применения к ним некоторых действий (например, визуального или машинного анализа). Это основная форма представления сведений в базах данных.

Организация того или иного вида хранения данных (структурированных или неструктурированных) связана с обеспечением доступа к ним. Под доступом понимается возможность выделения элемента данных (или множества элементов) среди других элементов по каким-либо признакам с целью выполнения некоторых действий над элементом.

Одной из самых распространенных моделей хранения, структурированных данных, является таблица.

Неструктурированные данные непригодны для обработки напрямую методами анализа данных, поэтому такие данные подвергаются специальным приемам структуризации, причем сам характер данных в процессе структуризации может существенно измениться.

Например, в анализе текстов при структурировании из исходного текста может быть сформирована таблица с частотами встречаемости слов, и уже такой набор данных будет обрабатываться методами, приводными для структурированных данных.

Слабоструктурированные данные – это данные, для которых определены некоторые правила и форматы, но в самом общем виде. Например, строка с адресом, строка в прайс-листе, ФИО и т. п. В отличие от неструктурированных, такие данные с меньшими усилиями преобразуются к структурированной форме, однако без процедуры преобразования они тоже непригодны для анализа.

Подавляющее большинство методов анализа данных работает только с хорошо структурированными данными, представленными в табличном виде.

Формализация данных.

При сборе данных нужно придерживаться следующих принципов:

1. Абстрагироваться от существующих информационных систем и имеющихся в наличии данных. Большие объемы накопленных данных совершенно не говорят о том, что их достаточно для анализа. Необходимо отталкиваться от задачи и подбирать данные для ее решения, а не брать имеющуюся информацию. К примеру, при построении моделей прогноза продаж опрос экспертов показал, что на спрос очень влияет цветовая характеристика товара. Анализ имеющихся данных продемонстрировал, что информация о цвете товарной позиции отсутствует в учетной системе. Значит, нужно каким-то образом добавить эти данные, иначе не стоит рассчитывать на хороший результат использования моделей,

2. Описать все факторы, потенциально влияющие на анализируемый процесс/объект. Основным инструментом здесь становится опрос экспертов и людей, непосредственно владеющих проблемной ситуацией. Необходимо максимально использовать знания экспертов о предметной области и, полагаясь на здравый смысл, постараться собрать и систематизировать максимум возможных предположений и гипотез.

3. Экспертно оценить значимость каждою фактора. Эта оценка не является окончательной, она будет отправной точкой. В процессе анализа вполне может выясниться, что фактор, который эксперты посчитали очень важным, таковым не является, и наоборот, незначимый, с их точки зрения, фактор может оказывать значительное влияние на результат.

4. Определить способ представления информации число, дата, да/нет, категория (то есть тип данных). Определить способ представления, то есть формализовать некоторые данные, просто. Например, объем продаж в рублях – это определенное число. Но довольно часто бывает непонятно, как представить фактор. Чаще всего такие проблемы возникают с качественными характеристиками. Например, на объемы продаж влияет качество товара. Качество сложное понятие, но если этот показатель действительно важен, то нужно придумать способ его формализации. Скажем, качество можно определять по количеству брака на тысячу единиц продукции либо оценивать экспертно, разбив на несколько категорий отлично/хорошо/удовлетворительно/плохо,

5. Собрать все легкодоступные факторы. Они содержатся в первую очередь в источниках структурированной информации учетных системах, базах данных и т, п.

6. Обязательно собрать наиболее значимые, с точки зрения экспертов, факторы. Вполне возможно, что без них не удастся построить качественную модель.

7. Оценить сложность и стоимость сбора средних и наименее важных по значимости факторов, Некоторые данные легкодоступны, их можно извлечь из существующих информационных систем. Но есть информация, которую непросто собрать, например, сведения о конкурентах, поэтому необходимо оценить, во что обойдется сбор данных, Сбор данных не является самоцелью. Если информацию получить легко, то, естественно, нужно ее собрать. Если сложно, то необходимо соизмерить затраты на ее сбор и систематизацию с ожидаемыми результатами.

Технологии KDD и Data Mining.

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

Несмотря на разнообразие бизнес задач, почти все они могут решаться по единой методике. Эта методика, зародившаяся в 1989 т., 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.

Knowledge Discovery in Databases (KDD) – это процесс поиска полезных знаний в "сырых" данных. KDD включает в себя вопросы: подготовки данных, выбора информативных признаков, очистки данных, применения методов Data Mining (DM), постобработки данных и интерпретации полученных результатов. Безусловно, «сердцем» всего этого процесса являются методы Data Mining, позволяющие обнаруживать знания.

Этими знаниями могут быть правила, описывающие связи между свойствами данных (деревья решений), часто встречающиеся шаблоны (ассоциативные правила), а также результаты классификации (нейронные сети) и кластеризации данных (карты Кохонена) и т.д.

Рассмотрим последовательность шагов, выполняемых в процессе KDD (рис. 1.4).

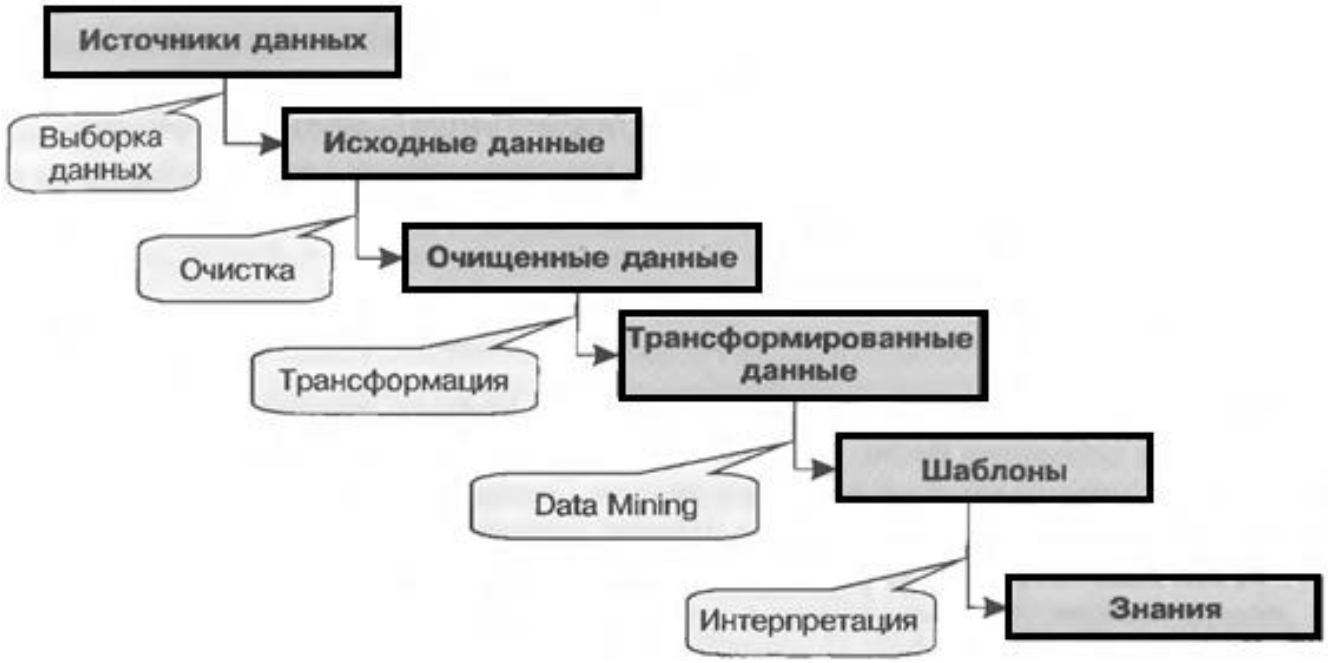

Рис. 1.4. Этапы KDD

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. 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 метода 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.

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

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 (извлечение знаний из баз данных) 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.

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 У. Такие правила называются ассоциативными. Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее называют анализом рыночной корзины. Если же нас интересует последовательность про исходящих 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 У.

Кроме перечисленных задач, часто выделяют анализ отклонений, анализ связей, отбор значимых признаков, 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 (рис. 1.5). К задаче регрессии сводится, в частности, прогнозирование временного ряда на основе исторических данных.

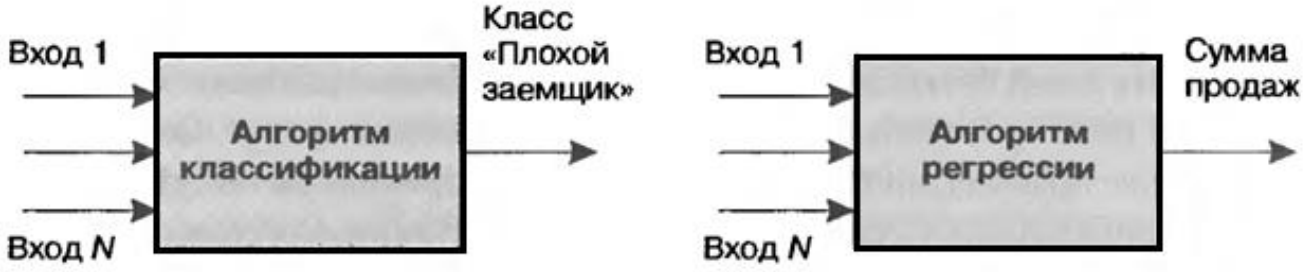

Рис. 1.5. Иллюстрация задачи классификации и задачи регрессии

Кластеризация отличается от классификации тем, что выходная переменная не требуется, а число кластеров, в которые необходимо сгруппировать все множество данных, может быть неизвестным. 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 (рис. 1.6).

Рис. 1.6. Иллюстрация задачи кластеризации

Перечислим наиболее известные способы применения этих задач.

Классификация используется, если заранее известны классы, например, при отнесении нового товара к той или иной товарной группе, отнесении клиента к какой-либо категории (при кредитовании 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 под ходить к каждому индивидуально, поэтому их удобно объединять в группы сегменты с однородными признаками. Выделять сегменты можно по нескольким группам признаков, например, по сфере деятельности или 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 при лавках, 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, при меняемых в этих дисциплинах.

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) – 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», – называемой 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.

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.

Задача консолидации

Ценность и достоверность знаний, полученных в результате интеллектуального анализа бизнес данных, зависит не только от эффективности используемых аналитических методов и алгоритмов, но и от того, насколько правильно подобраны и подготовлены исходные данные для анализа.

Поэтому, прежде чем приступать к анализу данных, необходимо выполнить ряд процедур, цель которых доведение данных до приемлемого уровня качества и информативности, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов.

Консолидация - комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

Консолидация данных является начальным этапом реализации любой аналитической задачи или проекта. В основе консолидации лежит процесс сбора и организации хранения данных в виде, оптимальном с точки зрения их обработки на конкретной аналитической платформе или решения конкретной аналитической задачи. Сопутствующими задачами консолидации являются оценка качества данных и их обогащение.

Основные критерии оптимальности с точки зрения консолидации данных:

- обеспечение высокой скорости доступа к данным;
- компактность хранения;
- автоматическая поддержка целостности структуры данных;
- контроль непротиворечивости данных.

Источники данных

Ключевым понятием консолидации является источник данных объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи. Необходимо, чтобы используемая аналитическая платформа могла осуществлять доступ к данным из этого объекта непосредственно либо после их преобразования в другой формат. В противном случае очевидно, что объект не может считаться источником данных.

Аналитические приложения, как правило, не содержат развитых средств ввода и редактирования данных, а работают с уже сформированными выборками. Таким образом, формирование массивов данных для анализа в большинстве случаев ложится на плечи заказчиков аналитических решений.

Основные задачи консолидации данных

В процессе консолидации данных решаются следующие задачи:

- выбор источников данных;
- разработка стратегии консолидации;
- оценка качества данных;
- обогащение;

- очистка;
- перенос в хранилище данных.