Введение в анализ данных [M.001]

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 проверки 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 «меру», «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.

Определение

> *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.*

Определение

> *Моделирование — универсальный метод получения, описания и использования знаний. Применяется в любой профессиональной деятельности.*

По виду моделирования модели делят:

- на эмпирические — полученные на основе эмпирических фактов, зависимостей;
- теоретические — полученные на основе математических описаний, законов;
- смешанные, полуэмпирические — полученные на основе эмпирических зависимостей и математических описаний.

Нередко теоретические модели появляются из эмпирических, например, многие законы физики первоначально были получены из эмпирических данных.

Пример

Совокупность предприятий функционирует на рынке, обмениваясь товарами, сырьем, услугами, информацией. Если описать экономические законы, правила взаимодействия на рынке с помощью математических соотношений, например системы алгебраических уравнений, где неизвестными будут величины прибыли, получаемые от взаимодействия предприятий, а коэффициентами уравнения — значения интенсивности таких взаимодействий, то получится математическая модель экономической системы, то есть *экономико-математическая модель системы предприятий на рынке*.

Таким образом, анализ данных тесно связан с моделированием.

Отметим важные свойства любой модели.

- Упрощенность. Модель отображает только существенные стороны объекта и, кроме того, должна быть проста для исследования или воспроизведения.
- Конечность. Модель отображает оригинал лишь в конечном числе его отношений, и, кроме того, ресурсы моделирования конечны.
- Приближенность. Действительность отображается моделью грубо или приближенно.
- Адекватность. Модель должна успешно описывать моделируемую систему.
- Целостность. Модель реализует некоторую систему (то есть целое).
- Замкнутость. Модель учитывает и отображает замкнутую систему необходимых основных гипотез, связей и отношений.
- Управляемость. Модель должна иметь хотя бы один параметр, изменениями которого можно имитировать поведение моделируемой системы в различных условиях.

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 (рисунок 1). 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 *y* с входными воздействиями x_1, x_2…). Варьирование входных параметров на выходе даст результат, который моделирует поведение системы в различных условиях.

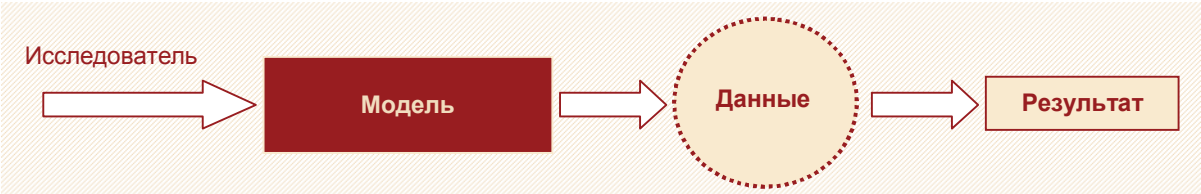

Рисунок 1 — Движение от модели к результату

Пример

Рассмотрим физическую систему. Тело массой m*, на которое воздействует сила* F*, скатывается по наклонной плоскости с ускорением* a*. Исследуя такие системы, Ньютон получил математическое соотношение:* F = ma*. Это* математическая модель физической системы*. При построении этой модели были приняты следующие гипотезы.*

- *Поверхность идеальна (то есть коэффициент трения равен нулю).*
- *Тело находится в вакууме (то есть сопротивление воздуха равно нулю).*
- *Масса тела неизменна.*
- *Тело движется с одинаковым постоянным ускорением в любой точке.*

При моделировании многих физических явлений мы используем закон Ньютона и делаем выводы.

Результат моделирования может соответствовать действительности, а может и нет. 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 в бизнесе неизбежно возникнут проблемы. Основным фактором, определяющим неблагополучие при использовании этих методов для решения бизнес-задач, является несоответствие между этими методами и реальностью, которую они призваны отражать. Существуют трудности, связанные с формализацией бизнес-процессов. Здесь факторы, определяющие явления, столь многообразны и многочисленны, 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, 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 (рисунок 2).

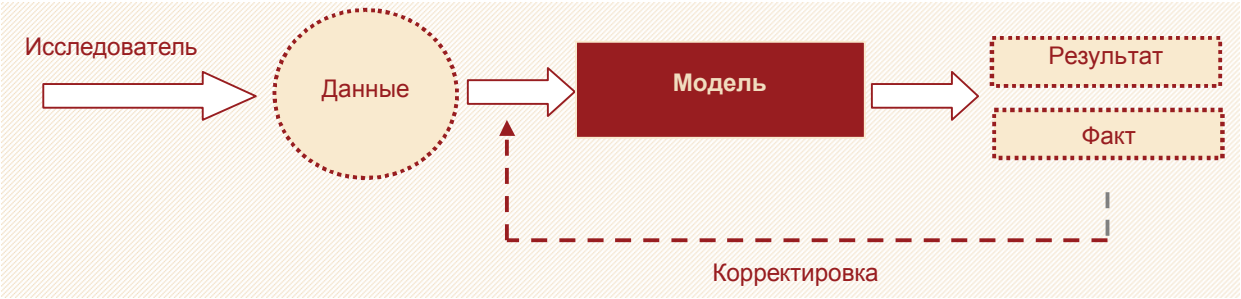

Рисунок 2 – Построение модели от данных

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

Если при аналитическом подходе мы можем выбрать модель, даже не имея никаких экспериментальных данных, характеризующих свойства системы, и начать ее использовать, то при информационном подходе без данных невозможно построить модель, так как ее параметры полностью определяются ими.

Пример

В банковском риск-менеджменте широко известна модель Дюрана для расчета рейтинга кредитоспособности заемщика, которая получила распространение в 40–50-е гг. XX в. На основе собственного опыта Дюран разработал балльную модель для оценки заемщика по совокупности его имущественных и социальных параметров (возраст, пол, профессия и т. д.). Преодолев некоторый порог, заемщик считался кредитоспособным. Эта модель представляет собой аналитическую зависимость y = f(*X*), где y – рейтинг, *X* – набор признаков заемщика.

Если перед современным российским банком встанет задача рассчитать рейтинг заемщика, банк может воспользоваться моделью Дюрана. Однако будет ли адекватной для современной российской действительности модель, разработанная в середине прошлого века на Западе? Естественно, не будет, так как она не учитывает связи между характеристиками российских заемщиков (возраст, образование, доход и т.д.) и дефолтностью по кредитам. Если же банк возьмет собственные данные по кредитным историям и на их основе построит модель, рассчитывающую рейтинг клиента, то, вполне вероятно, она окажется работоспособной.

В первом случае, когда мы брали модель Дюрана, мы использовали аналитический подход*. Во втором —* информационный*; для построения модели нам понадобились данные — кредитные истории заемщиков банка.*

Модели, полученные с помощью информационного подхода, учитывают специфику моделируемого объекта, явления в отличие от аналитического подхода. Для бизнес-процессов последнее качество очень важно, поэтому информационный подход лег в основу большинства современных промышленных технологий и методов анализа данных: KDD, Data Mining, 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 и Data Mining, а средством построения прикладных решений в области анализа – аналитические платформы.