Подготовка данных к анализу [M.005]

Информационный подход к анализу базируется на различных алгоритмах извлечения закономерностей из исходных данных, результатом работы которых являются модели. Таких алгоритмов довольно много, но они не способны гарантировать качественное решение. Никакой, даже весьма изощренный, метод сам по себе не даст хорошего результата, так как критически важным является качество исходных данных. Чаще всего именно оно становится причиной неудачи. Несмотря на то что существуют специальные методы очистки данных, понимание и соблюдение **принципов сбора и подготовки данных** значительно облегчит построение моделей и позволит получить хорошие результаты.

Особенности данных, накопленных в компаниях

Данные, которые накапливают предприятия и организации в базах данных и прочих источниках (так называемые *бизнес-данные*), имеют свои особенности. Рассмотрим их.

Бизнес-данные редко накапливаются специально для решения задач анализа. Предприятия и организации собирают данные для коммерческих целей: ведения учета, проведения финансового анализа, составления отчетности, принятия решений и т. п. Этим бизнес-данные отличаются от экспериментальных данных, которые собираются для исследовательских целей. Основными потребителями бизнес-данных обычно являются лица, принимающие решения в компаниях.

Бизнес-данные, как правило, содержат ошибки, аномалии, противоречия и пропуски. Это следствие того, что компании не собирают данные с целью анализа. В них появляются ошибки различной природы, что снижает качество данных.

С точки зрения анализа объемы хранимых данных очень велики. Современные базы данных содержат мегабайты и гигабайты информации. Для ресурсоемких алгоритмов анализа данных таблицу объемом 50 тыс. записей можно считать большой, поэтому при построении моделей важно применять процедуры сэмплинга, сокращения записей и отбора информативных признаков либо использовать специальные масштабируемые алгоритмы, способные работать на больших наборах данных.

Отмеченные особенности бизнес-данных влияют как на сам процесс анализа, так и на подготовку и систематизацию данных.

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 в конкретной компании. 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 каждого 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.

Рассмотрим эти принципы на примере формализации данных при решении задачи прогнозирования спроса. На этапе описания факторов, влияющих на продажи, и выдвижения гипотез полезно составить таблицу факторов и их значимости (таблица 1).

Таблица 1 — Факторы, влияющие на продажи, и их значимость

Показатель	Экспертная оценка значимости (1–100)
Сезон	100
День недели	80
Объем продаж за предыдущие недели	100
Объем продаж за аналогичный период прошлого года	95
Рекламная кампания	60
Маркетинговые мероприятия	40
Качество продукции	30
Рейтинг бренда	25
Отклонение цены от среднерыночной	60
Наличие данного товара у конкурентов	15

При создании подобной таблицы следуют принципам 1–3 формализации данных. Далее необходимо определить способ представления данных и оценить стоимость их сбора. К таблице добавятся еще два столбца (таблица 2). И уже после этого можно принимать решение о том, какие факторы включать в анализ, а какими пренебречь. Очевидно, что все легкодоступные показатели с высокой экспертной значимостью требуется включать в рассмотрение. А фактором Качество продукции, например, можно пренебречь: по мнению экспертов, он малозначим, а стоимость его сбора велика.

Таблица 2 — Факторы, влияющие на продажи с оценками экспертов

Показатель	Экспертная оценка значимости (1–100)	Способ представления	Экспертная оценка сложности получения
Сезон	100	Число	низкая
День недели	80	Дата	низкая
Объем продаж за предыдущие недели	100	Число	низкая
Объем продаж за аналогичный период прошлого года	95	Число	низкая
Рекламная кампания	60	Число	средняя
Маркетинговый бюджет	40	Число	средняя
Качество продукции	30	Строка (плохое/хорошее/отличное)	высокая
Рейтинг бренда	25	Строка (известный/малоизвестный и т. д.)	средняя
Отклонение цены от среднерыночной	60	Число	средняя
Наличие данного товара у конкурентов	15	Логическое (да/нет)	средняя

Методы сбора данных

Есть несколько методов сбора необходимых для анализа данных.

1. **Получение из учетных систем.** Обычно в учетных системах есть различные механизмы построения отчетов и экспорта данных, поэтому извлечение нужной информации из них чаще всего относительно несложная операция.
2. **Получение данных из косвенных источников информации.** О многих показателях можно судить по косвенным признакам, и этим нужно воспользоваться. Например, можно оценить реальное финансовое положение жителей определенного региона следующим образом. В большинстве случаев имеется несколько товаров, предназначенных для выполнения одной и той же функции, но отличающихся по цене: товары для бедных, средних и богатых. Если получить отчет о продажах товара в интересующем регионе и проанализировать пропорции, в которых продаются товары для бедных, средних и богатых, то можно предположить, что чем больше доля дорогих изделий из одной товарной группы, тем более состоятельны в среднем жители данного региона.
3. **Использование открытых источников.** Большое количество данных присутствует в таких открытых источниках, как статистические сборники, отчеты корпораций, опубликованные результаты маркетинговых исследований и пр.
4. **Приобретение аналитических отчетов у специализированных компаний.** На рынке работает множество компаний, которые профессионально занимаются сбором данных и предоставлением результатов клиентам для последующего анализа. Собираемая информация обычно предоставляется в виде различных таблиц и сводок, которые с успехом можно применять при анализе. Стоимость получения подобной информации чаще всего относительно невысока.

5 **Проведение собственных маркетинговых исследований и аналогичных мероприятий по сбору данных.** Этот вариант сбора данных может быть достаточно дорогостоящим, но в любом случае он существует.

6 **Ввод данных вручную.** Данные вводятся по различного рода экспертным оценкам сотрудниками организации. Такой метод является наиболее трудоемким.

Методы сбора информации существенно отличаются по стоимости и необходимому времени, поэтому следует соизмерять затраты с результатами. Возможно, от сбора некоторых данных придется отказаться, но факторы, которые эксперты оценили как наиболее значимые, нужно собрать обязательно несмотря на стоимость этих работ, либо вообще не проводить анализ.

Данные должны быть собраны в единую таблицу в формате MS Excel, DBase, в текстовые файлы с разделителями или в набор таблиц в любой реляционной СУБД (системе управления базами данных), то есть должны быть представлены в структурированном виде. Кроме того, необходимо унифицировать представление данных: один и тот же объект должен везде описываться одинаково.

Информативность данных

Одной из распространенных ошибок при сборе данных из структурированных источников является стремление взять для анализа как можно больше признаков, описывающих объекты. Между тем предварительная оценка данных, которая проводится визуально при помощи таблиц и базовой статистической информации по набору данных, существенно помогает в определении информативности признаков с точки зрения анализа.

Среди **неинформативных** признаков выделяются четыре типа:

- признаки, содержащие *только одно значение* (рисунок 1.а);
- признаки, содержащие *в основном одно значение* (рисунок 1.б);
- признаки с *уникальными значениями* (рисунок 1.в);
- *признаки, между которыми имеет место сильная корреляция,* — в этом случае для анализа можно взять один столбец (рисунок 1.г).

Признак
1
1
1
1
1
1
1
1
1
1
1
1
1

а)

Признак
1
1
1
1
1
0
1
1
1
1
1
1

б)

№ паспорта
0936-866096
8355-512943
8017-098471
2762-945535
0459-997701
6291-817248
0094-883508
6385-082612
9290-732300
7022-736158
3127-709332
4179-171975

в)

Пол	Gender
Жен	0
Жен	0
Жен	0
Муж	1
Муж	1
Жен	0
Жен	0
Муж	1
Муж	1
Жен	0
Жен	0
Муж	1

г)

Рисунок 1 – Примеры неинформативных признаков

Признаки, содержащие в основном одно значение, не всегда могут быть неинформативными, многое зависит от целей анализа. Например, при решении задачи анализа отклонений такие признаки могут существенно повлиять на построение моделей.

Требования к данным

Аналитические инструменты пытаются построить модели на основе предложенных данных, поэтому чем ближе данные к действительности, тем лучше. Необходимо понимать: модель не может «знать» о том, что находится за пределами собранных для анализа данных. Например, если при создании системы диагностики больных использовать только сведения о больных людях, то модель не будет знать о существовании в природе здоровых людей.

Существуют требования к минимальным объемам данных для возможности построения моделей на их основе. В зависимости от представления данных и решаемой задачи эти требования различны.

Для временных рядов, которые относятся к упорядоченным данным, требования следующие. Если для моделируемого бизнес-процесса (например, продажи) характерна сезонность/цикличность, то необходимо иметь данные хотя бы за один полный сезон/цикл с возможностью варьирования интервалов (понедельное, помесячное и т. д.). Максимальный горизонт прогнозирования зависит от объема данных: данные за 1,5 года — прогноз возможен максимум на 1 месяц; данные за 2–3 года — на 2 месяца.

Для неупорядоченных данных требования следующие.

- Количество примеров (прецедентов) должно быть значительно больше количества факторов.
- Желательно, чтобы данные покрывали как можно больше ситуаций реального процесса.
- Пропорции различных примеров (прецедентов) должны примерно соответствовать реальному процессу.

Транзакционные данные. Анализ транзакций целесообразно производить на большом объеме данных, иначе могут быть выявлены статистически необоснованные правила. Алгоритмы поиска ассоциативных связей способны быстро перерабатывать огромные массивы данных. Примерное соотношение между количеством объектов и объемом данных следующее:

- 300–500 объектов — не менее 10 тыс. транзакций;
- 500–1000 объектов — более 300 тыс. транзакций.