Прикладные методы обработки данных

Лектор: Цыганова Мария Сергеевна

Понятие анализа данных

В литературе существует множество определений понятия «анализ данных».

В широком смысле

анализ данных – это процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

Как правило, для анализа данных используются различные математические методы

Эволюция анализа данных

Согласно одной из классификаций, анализ данных вырос из задач прикладной математики.

Исторически первыми сложились классические подходы к анализу данных:

- вычислительная математика;
- идентификация моделей.

Эволюция анализа данных: вычислительная математика

Вычислительная математика решает задачу определения некоторых характеристик изучаемого объекта или явления по известным значениям других его характеристик.

При этом:

- модель объекта считается известной;
- зависимости между характеристиками описываются аналитическими выражениями в виде уравнений или систем уравнений и неравенств.

Проблемы при решении таких задач связаны, как правило, с большими объемами вычислений и накоплением погрешности (из-за округления чисел)

Эволюция анализа данных: идентификация моделей

Постановка задачи:

известен набор переменных, влияющих на целевую характеристику, и общий вид зависимости между характеристиками;

параметры модели (коэффициенты, показатели степеней и т. п.) неизвестны – для их определения используются данные наблюдений (значения целевых характеристик при различных значениях переменных).

Эволюция анализа данных: идентификация моделей

В процессе решения подбираются такие значения параметров, при которых модель позволяет по известным значениям входных характеристик с заданной точностью определить значения целевых (выходных) характеристик.

Математические методы:

- корреляционный анализ;
- регрессионный анализ;
- факторный анализ;
- численные методы оптимизации
- др.

Классический подход к анализу данных

Для решения задач выбираются готовые математические модели.

Они проверяются на основе имеющихся данных.

Основной недостаток:
за адекватность модели рассматриваемому явлению исследователь не отвечает

Классический подход к анализу данных

Современные задачи анализа данных:

точный вид «истинной» модели, как правило, неизвестен;

часто неизвестен ни характер связей между переменными, ни даже исчерпывающий перечень самих переменных.

«В сущности, все модели неправильные, но некоторые из них полезны».
Дж. Бокс (британский статистик)

Классический подход к анализу данных

На любом этапе развития прикладной математики возникают реальные задачи, для решения которых нет готовых математических моделей, и нет времени для их создания.

Поиск новых парадигм анализа данных.

Концепция разведочного анализа

Автор – Джон Тьюки (John Tukey).

Время появления – 1962 г.

Основная идея:

требуется первичный (разведочный) этап анализа данных – Exploratory Data Analysis (EDA) – для выявления признаков закономерностей и выдвижения гипотез.

Важнейший элемент разведочного анализа – широкое использование визуального представления многомерных данных (графиков, схем, таблиц и т. п.)

Концепция разведочного анализа

Наглядное представление позволяет увидеть закономерности в данных и выдвинуть гипотезы.

После работ Тьюки в статистике получили широкое распространение *ящичная диаграмма* («ящик с усами», box plot) и *диаграмма рассеяния* (scatterplot).

Концепция разведочного анализа

Разведочный анализ – в большей степени подход, чем теория.

Это синтез детерминированных, стохастических и эвристических подходов к анализу данных наблюдений.

Концепция разведочного анализа

Предполагается три этапа анализа данных:

Разведочный анализ

Цель – выявление внутренних закономерностей в данных для формирования рабочих гипотез о связях между переменными (при отсутствии априорной информации о таких связях)

Подтверждающий (конфирматорный) анализ

Проверка соответствия сформулированных гипотез имеющимся эмпирическим данным; вычисление итоговых статистических оценок моделей и определение их погрешностей

Итоговый анализ

Экспертный анализ результатов и их обобщение

Концепция разведочного анализа

Предполагается три этапа анализа данных:

Разведочный анализ

Цель – выявление внутренних закономерностей в данных для формирования рабочих гипотез о связях между переменными (при отсутствии априорной информации о таких связях)

Подтверждающий (конфирматорный) анализ

Проверка соответствия сформулированных гипотез имеющимся эмпирическим данным; вычисление итоговых статистических оценок моделей и определение их погрешностей

Итоговый анализ

Экспертный анализ результатов и их обобщение

Концепция разведочного анализа

Если результаты разведочного анализа говорят в пользу некоторой модели, то правильность этой модели можно проверить, применив ее к новым (не участвовавшим в разведочном анализе) данным.

Отсюда вышли процедуры разделения данных на обучающую и тестовую выборки в методах Data Mining

Современное понимание анализа данных

Задача анализа явлений, для которых нет готовых математических моделей.

Есть:

наборы экспериментальных данных («входы» – «выходы» или даже только «входы») в виде массивов или таблиц.

Основной предмет анализа – конструирование моделей и определение их параметров.

Исследователь несет ответственность за привнесение эвристических гипотез о формах зависимостей, параметры предполагаемых распределений и т. п.

Современное понимание анализа данных

Важно:

концепция «моделей от данных» требует внимательного отношения к качеству самих исходных данных (ошибочные, зашумленные, противоречивые данные могут привести к моделям и выводам, не имеющим никакого отношения к реальному процессу).

В современном анализе важную роль играют *интеграция*, *подготовка* и *очистка* данных.

1.2 Формы представления данных

Структурированность данных

По степени структурированности различают следующие типы данных:

- структурированные,
- слабоструктурированные,
- неструктурированные.

Доступ к данным

Организация хранения данных (как структурированных, так и неструктурированных) связана с обеспечением доступа к ним.

Под доступом понимается возможность выделения элемента (множества элементов) данных среди других элементов по каким-либо признакам с целью выполнения некоторых действий над выделенными элементами.

Структурированные данные

Структурированные данные отражают отдельные факты предметной области.

Структурированные данные – данные, определенным образом упорядоченные и организованные с целью обеспечения возможности применения к ним некоторых действий (например, визуального или компьютерного анализа).

Это основная форма представления сведений в БД.

Структурированные данные

Самая распространенная модель хранения структурированных данных – таблица.

Данные упорядочиваются в двумерную структуру, состоящую из столбцов и строк.

В ячейках таблицы – элементы данных (числа, символы, логические значения).

ab Код	12 Возраст	ab Пол	ab Состоит в браке	12 Иждивенцы	9.0 Доход	9.0 Опыт работы	9.0 Срок проживания	9.0 Неде
1	28	женский	Да	0	9 000,00	9,00	7,00	
2	39	мужской	Да	1	13 500,00	17,00	6,00	
3	31	мужской	Нет	2	7 000,00	11,00	3,00	
4	34	мужской	Нет	1	10 200,00	15,00	2,00	
5	46	женский	Да	2	8 500,00	20,00	8,00	
6	30	женский	Да	2	9 500,00	12,00	30,00	
7	47	мужской	Нет	2	7 900,00		6,00	
8	33	мужской	Нет	2	12 600,00	15,00	23,00	
9	22	мужской	Нет	0	34 000,00	4,00	19,00	
10	30	мужской	Да	1	33 000,00	10,00	8,00	

Неструктурированные данные

Неструктурированные данные непригодны для обработки напрямую методами анализа; они подвергаются специальным процедурам структуризации.

Например:

при анализе текста структурирование может состоять в формировании из исходного текста таблицы частот встречаемости слов.

Дальнейшая обработка такой таблицы – методами работы со структурированными данными.

Слабоструктурированные данные

Слабоструктурированные данные – это данные, для которых определены некоторые правила и форматы, но в самом общем виде.

Например:

строка с адресом, ФИО и т. п.

Такие данные проще преобразовать к структурированной форме, но без процедуры преобразования они также непригодны для анализа.

Слабоструктурированные данные

Пример.

625003 г. Тюмень, ул. Перекопская, 15А

Поле	Значение
Индекс	625003
Город	Тюмень
Улица	Перекопская
Дом	15А
Корпус	

Подавляющее большинство методов анализа применимо только к структурированным данным.

В дальнейшем будут рассматриваться структурированные данные.

1.3 Основные типы шкал измерений значений признаков

Измерительная шкала

Шкала (измерительная шкала) – это знаковая система, для которой задано отображение (операция измерения), ставящее в соответствие реальным объектам тот или иной элемент (значение) шкалы.

Формально: шкалой называют кортеж

$$<X, \phi, Y>,$$

где X – множество реальных объектов,

ϕ – отображение,

Y – множество элементов (значений) знаковой системы.

Измерительная шкала

Шкалы классифицируются по типам измеряемых данных, которые определяют допустимые для данной шкалы отношения, в том числе те, что соответствуют математическим преобразованиям значений шкалы.

Номинальная шкала (шкала наименований)

Это качественная шкала, по которой объектам x_i или классам эквивалентных объектов сопоставляется некоторый знак (шкальное значение).

Шкала $\phi : X \to Y$ называется *шкалой наименований*, если она единственна с точностью до взаимно-однозначного преобразования.

Номинальная шкала (шкала наименований)

Название шкалы объясняется тем, что шкальные значения используются только как имена объектов (знаки, которые, в частности, могут быть цифрами).

Кроме сравнения на совпадение, любые арифметические действия над этими знаками бессмысленны.

Цифры, используемые в качестве шкальных значений, не следует рассматривать как числа (например, цифра «2» не больше, чем цифра «1» и не есть 1+1).

Номинальная шкала (шкала наименований)

Примеры: номера автомашин, телефонов,
номера групп (Вуз),
гербы, флаги, коды государств, регионов,
коды городов, лиц, объектов и т. п.

Измерение в шкале наименований – это определение принадлежности объекта к тому или иному классу и обозначение этого факта с помощью соответствующего знака.

Если каждый класс состоит из одного объекта, то шкала наименований используется для различения объектов.

Номинальная шкала (шкала наименований)

Для обработки результатов измерений в шкале наименований могут быть использованы операции:

- число наблюдений k–го класса n_k ;
- относительная частота появления класса $p_k^* = \frac{n_k}{n}$;
- мода $k_{\max} = \arg\max_k \left(p_k^*\right)$;
- статистические тесты на относительных частотах (например, χ^2).

Порядковая (ранговая) шкала

Шкала называется *ранговой* (*шкалой порядка*), если она единственна с точностью до монотонно возрастающего преобразования.

Порядковый тип шкал допускает не только различение объектов, как номинальный тип, но и используется для упорядочения объектов по измеряемым свойствам.

Порядковая (ранговая) шкала

Примеры.

1. Нумерация очередности (позиции в рейтинге).
2. Воинские звания.
3. Шкала силы ветра по Бофорту.
   Сила ветра определяется по волнению моря:
   0 – штиль, 4 – умеренный ветер, 6 – сильный ветер,
   10 – шторм (буря), 12 – ураган.
4. Шкала магнитуд землетрясений по Рихтеру –
   12-балльная шкала для оценки энергии
   сейсмических волн в зависимости и последствий
   прохождения их по данной территории.

Шкала интервалов

Пусть упорядочение объектов можно выполнить настолько точно, что известны «расстояния» между любыми двумя из них.

Пусть все измерения выражены в единицах, хотя и произвольных, но одинаковых по всей длине шкалы.

Следствие: независимость отношения двух интервалов от того, в какой из шкал эти интервалы измерены (какова единица длины и какое значение принято за начало отсчёта)

$$\frac{\varphi_1(x_1)-\varphi_1(x_2)}{\varphi_1(x_3)-\varphi_1(x_4)} = \frac{\varphi_2(x_1)-\varphi_2(x_2)}{\varphi_2(x_3)-\varphi_2(x_4)}.$$