



## ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

*Предварительное знание того, что хочешь  
сделать, дает смелость и легкость.*  
Д. Дидро

Технологии анализа данных

---

---

---

---

---

---

---

---

## Содержание

2

- Понятие и цели предварительной обработки данных
- Методы предварительной обработки данных

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Зачем нужна предварительная обработка данных

3

- Неполнота данных
  - Отсутствие значений
    - Место=" ", ГодРождения="n/a"
- "Шумы" (ошибки или аномалии) в данных
  - Зарплата=-10000
- Несогласованность данных
  - Возраст=70 и ДатаРождения="7-Окт-52"
  - В разных записях Категория ∈ {1, 2, 3} или {A, B, C}

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Почему данные содержат ошибки

4

- Неполные данные
  - "п/а" при сборе данных
  - Изменение точки зрения на корректность с момента сбора данных на момент анализа данных
  - Ошибки в программном/аппаратном обеспечении, человеческий фактор
- "Шумы" в данных
  - Ошибки в программном обеспечении сбора данных, человеческий фактор
  - Сбои при передаче данных
- Несогласованность данных
  - Много различных источников данных
  - Нарушения ФЗ (например, модификация связанных данных)
- Дублирование данных

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Почему предварительная обработка данных важна

5

- Нет качественных данных – нет качественных результатов их анализа
  - Качественные решения должны быть основаны на качественных данных
  - Хранилище данных нуждается в согласованной интеграции качественных данных
- Извлечение, очистка и трансформация данных составляют большую часть работы по построению хранилища данных

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Измерение качества данных

6

- Общепринятая шкала качества
  - Точность
  - Полнота
  - Согласованность
  - Поддержка времени
  - Правдоподобие
  - Дополняемость
  - Интерпретируемость
  - Доступность
- Другие категории
  - Привязка к контексту, репрезентативность, доступность

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Цели предварительной обработки

7

- Очистка данных (data cleaning)
  - Заполнить отсутствующие значения, сгладить зашумленные данные, определить или удалить аномалии, исправить несогласованность
- Интеграция данных
  - Интеграция баз данных, кубов данных, файлов
- Трансформация данных
  - Нормализация и агрегация
- Редукция данных
  - Усеченное представление, гарантирующее те же или сходные результаты аналитической обработки
- Дискретизация данных
  - Часть редукции – для числовых данных

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

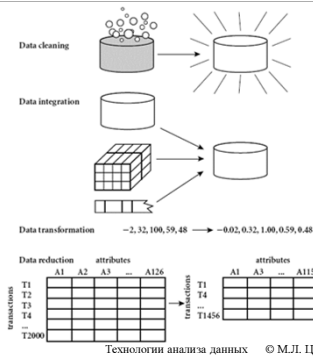
---

---

---

## Цели предварительной обработки

8



Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Обобщение данных

9

- Мотивация
  - Лучше понять данные: общая тенденция, вариативность, разброс
- Характеристики дисперсии
  - медиана, min, max, квантили и др.
- Численные измерения, соответствующие упорядоченным интервалам
  - Анализ упорядоченных интервалов
- Анализ разброса вычисленных мер
  - Перевод мер в числовые значения
  - Анализ OLAP-куба

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Измерение тенденции данных

10

- **Среднее (mean)**
  - арифметическое  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
  - взвешенное  $\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$
- **Медиана (median)**

$$\text{median}(a_1, \dots, a_n) = \begin{cases} a_{[n/2]}, & n - \text{нечетное} \\ (a_{[n/2]} + a_{1+[n/2]})/2, & n - \text{четное} \end{cases}$$
- **Мода (mode)**
  - значение, которое встречается наиболее часто
  - уни- и мультимодальность
  - эмпирическая формула:  $\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$
- **Средний уровень (midrange):**  $(\text{max} - \text{min})/2$

Технологии анализа данных © М.Л. Цимблер

---

---

---

---

---

---

---

---


---

---

## Измерение дисперсии данных

11

- **Среднеквадратичное отклонение**  $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$
- **Стандартное отклонение**  $s = \sqrt{\frac{n}{n-1} \sigma^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$
- **Разброс:**  $\text{max} - \text{min}$
- **Ящики с усами (box-and-whiskers plot)**
  - *k*-й перцентиль – число, отделяющее сверху *k*% расположенных в возрастающем порядке значений числового ряда
  - **Квартили**
    - Q1 – 25-й перцентиль
    - Q2 – медиана, 50-й перцентиль
    - Q3 – 75-й перцентиль
  - **Интерквартильный размах:**  $IQR = Q3 - Q1$
  - **Выбросы (аномалии):**  $1.5 * Q3$  выше Q3 или ниже Q1.



Технологии анализа данных © М.Л. Цимблер

---

---

---

---

---

---

---

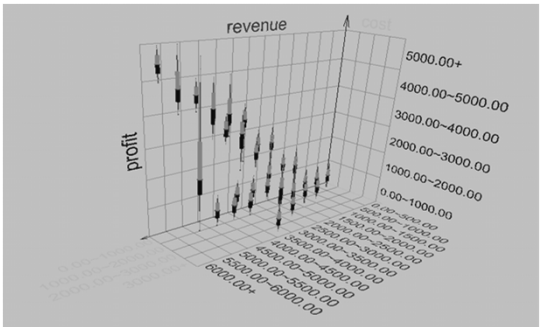
---

---

---

## Визуализация дисперсии данных

12



Технологии анализа данных © М.Л. Цимблер

---

---

---

---

---

---

---

---

---

---

## Очистка данных

13

- Важность
  - Очистка данных – одна из основных проблем построения хранилищ данных
- Основные цели
  - Заполнение отсутствующих значений
  - Определить аномалии и сгладить шумы
  - Исправить несогласованные данные
  - Устранить избыточность данных, вызванную интеграцией

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Обработка отсутствующих данных

14

- Игнорировать запись
- Ввести вручную (часто невозможно)
- Ввести автоматически
  - значение глобальной константы
    - значения UNKNOWN или N/A – не всегда хорошо
  - Ввести среднее или медиану
  - Ввести среднее или медиану класса (классы должны быть определены)
  - Ввести наиболее ожидаемую величину (используя моду, дерева решений или др.)

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Зашумленные данные

15

- Шум – случайное значение измеряемой переменной
- Причины некорректного значения
  - сбой при сборе данных
  - сбой при вводе данных
  - сбой при передаче данных
  - сбой при преобразовании данных
  - несогласованность имен данных
  - технологические лимиты (на размер записи и др.)
- Другие проблемы, требующие очистки
  - дублирование записей
  - неполные данные
  - несогласованные данные

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Обработка зашумленных данных

16

- Биннинг (binning)
  - Сортировка набора данных, затем разделение на равномошные поднаборы, затем сглаживание средним, медианой или граничными значениями поднабора
- Регрессия
  - Сглаживание путем подгонки данных под регрессионную функцию
- Кластеризация
  - Определение и удаление аномалий
- Совместная инспекция человека и компьютера
  - Определение подозрительных значений с помощью компьютера для их последующей проверки экспертом

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Биннинг

17

- Отсортированные данные:
  - 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Разделение данных на группы с одинаковой частотой
  - 4, 8, 9, 15
  - 21, 21, 24, 25
  - 26, 28, 29, 34
- Сглаживание с помощью среднего арифметического группы
  - 9, 9, 9, 9
  - 23, 23, 23, 23
  - 29, 29, 29, 29
- Сглаживание с помощью граничных значений группы
  - 4, 4, 4, 15
  - 21, 21, 25, 25
  - 26, 26, 26, 34

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

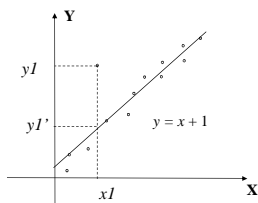
---

---

## Регрессия

18

- *Регрессия* – зависимость среднего значения какой-либо величины от некоторой другой величины или от нескольких величин.



Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

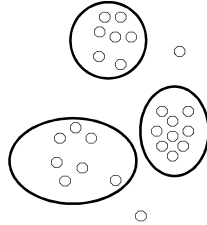
---

---

## Кластеризация

19

- *Кластеризация* – разбиение заданной выборки объектов на подмножества (кластеры) таким образом, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.



Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Несоответствие данных

20

- Верификация вводимых данных (проверка форматов и значений)
- Корректировка с использованием данных по внешним ссылкам

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Интеграция данных

21

- Основные проблемы
  - Идентификация сущностей
    - Как понять, что  $A.CustCode \equiv B.CustID \equiv C.CustNo$  ?
  - Избыточность
    - Значение некоторого атрибута выводимо из других атрибутов
    - Несоответствие в именах атрибутов
- Пути решения проблем
  - Использование метаданных
  - Использование корреляционного анализа
    - статистический анализ взаимосвязи двух или нескольких случайных величин

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Корреляционный анализ числовых данных

22

- Коэффициент корреляции Пирсона

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

- $r_{A,B} > 0$ : положительная корреляция А и В
- $r_{A,B} < 0$ : отрицательная корреляция А и В
- $r_{A,B} = 0$ : независимость А и В

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Корреляционный анализ нечисловых данных

23

- Тест  $\chi^2$   $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$

- Чем больше  $\chi^2$ , тем больше связь между величинами

- Пример

	<i>Играют в шахматы</i>	НЕ играют в шахматы	ВСЕГО
<i>Любят фантастику</i>	250 (90)	200 (360)	450
НЕ любят фантастику	50 (210)	1000 (840)	1050
ВСЕГО	300	1200	1500

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Трансформация данных

24

- Сглаживание
  - удаление шумов с помощью биннинга, регрессии или кластеризации
- Агрегация
  - суммирование частичных итогов (например, месячные продажи вместо ежедневных)
- Обобщение
  - замена частных значений более общими (например, "молодой", "среднего возраста", "пожилой" вместо значения возраста)
- Нормализация
  - приведение значений к одному заданному промежутку
- Создание новых атрибутов из имеющихся
  - например,  $SP.TotalSale = SP.Qty * P.Price$

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---



## Нормализация данных

25

- Мин-макс [ $new\_min_A, new\_max_A$ ]

$$v' = \frac{v - min_i}{max_i - min_i} (new\_max_i - new\_min_i) + new\_min_i$$

- Z-нормализация

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Десятичное масштабирование

$$v' = \frac{v}{10^j} \quad \text{где } j - \text{наименьшее целое, что } \max(|v'|) < 1$$

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Редукция данных

26

- Уменьшение количества строк (объектов)
- Уменьшение количества столбцов (атрибутов)
- Сжатие
- Дискретизация

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Уменьшение количества строк

27

- Агрегация
  - ▣ Использование более высокого уровня иерархии в измерениях (например, день → неделя → месяц → год)
- Использование моделей
  - ▣ Если данные подходят под некоторую модель, оцениваем параметры модели, сохраняем параметры, отбрасываем данные (за исключением аномалий)
- Использование гистограмм
  - ▣ Разделение данных на подгруппы, хранение представления подгрупп (сумма, количество и др.)
- Кластеризация
  - ▣ Разделение данных на подгруппы на основе расстояний между элементами, хранение представителей (центроидов) кластеров и аномалий
- Сэмплинг
  - ▣ Большое множество данных представляется своим существенно меньшим по мощности подмножеством, элементы которого выбираются случайным образом

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

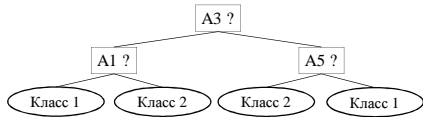
---

---

## Уменьшение количества столбцов

28

- Пошаговый прямой отбор
  - $\{\} \rightarrow \{A1\} \rightarrow \{A1, A3\} \rightarrow \{A1, A3, A5\}$
- Пошаговый обратный отбор
  - $\{A1, A2, A3, A4, A5\} \rightarrow \{A1, A3, A4, A5\} \rightarrow \{A1, A3, A5\}$
- Деревья решений
  - $\{A1, A2, A3, A4, A5\} \rightarrow \dots \rightarrow \{A1, A3, A5\}$



Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Сжатие данных

29

- Может принести выгоду, если алгоритму интеллектуального анализа не потребуется восстановление сжатых данных
- Сжатие без потерь
  - LZW, ZIP, DWT, ...
- Сжатие с потерями
  - JPEG, MPEG, ...

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Дискретизация

30

- Типы атрибутов
  - Номинальные – значения из неупорядоченного множества (например, Цвет, Профессия)
  - Ординальные – значения из упорядоченного множества (например, ВоинскоеЗвание)
  - Непрерывные – вещественные или целые числа
- Дискретизация
  - Разбиение промежутка значений непрерывного атрибута на интервалы
  - Уменьшение размера данных (границы/метки интервала заменяют значения)
  - Введение иерархий

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

---

---

## Дискретизация

31

- Биннинг
- Гистограммы
- Кластеризация
- Дискретизация на основе энтропии
  - Выбор точки разбиения интервала таким образом, чтобы минимизировать функцию энтропии
- Естественное разбиение
  - Правило 3-4-5

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Дискретизация на основе энтропии

32

- Пусть  $T$  – граница интервала  $S$ , разделяющая его на интервалы  $S_1$  и  $S_2$ . Тогда

$$I(S, T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2) \quad Entropy(S_1) = - \sum_{i=1}^n p_i \log_2(p_i)$$

- Выбирается граница  $T$ , которая минимизирует функцию энтропии.
- Процесс разбиения может повторяться до достижения определенного критерия останова

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Дискретизация по правилу 3-4-5

33

- Если интервал содержит 3, 6, 7 или 9 различных значений (по заданной значащей цифре), разбить его на 3 равных интервала
- Если интервал содержит 2, 4 или 8 различных значений (по заданной значащей цифре), разбить его на 4 равных интервала
- Если интервал содержит 1, 5 или 10 различных значений (по заданной значащей цифре), разбить его на 5 равных интервалов

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Дискретизация категориальных данных

34

- Порядок атрибутов на уровне схемы данных
  - Улица < Город < Область < Округ
- Спецификация иерархии для набора данных
  - { Челябинск, Златоуст, Магнитогорск } < Челябинская область
- Спецификация частичного порядка атрибутов
  - Улица < Город
- Автоматическая генерация иерархии из множества атрибутов на основе анализа количества уникальных значений
  - К-во уникальных значений: Округ > Область > Город > Улица
  - Исключения: ДеньНедели, Месяц, Квартал, Год

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---

## Заключение

35

- Предварительная обработка данных – важная задача для построения хранилищ данных и интеллектуального анализа данных.
- Нет качественных данных – нет качественных результатов их анализа.
- Предварительная обработка включает в себя
  - Очистка и интеграция данных
  - Редукция данных
  - Дискретизация данных

Технологии анализа данных © М.Л. Цымблер

---

---

---

---

---

---

---

---