Казанский
Государственный
Медицинский
Университет

Особенности медико-биологических данных. Способы получения. Структура.

Лакман Ирина Александровна

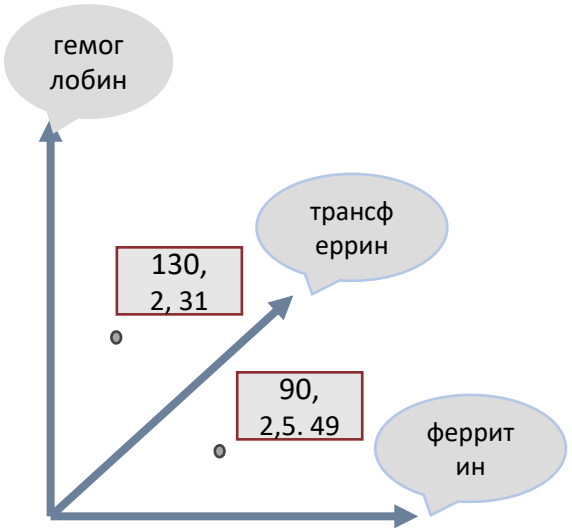
Казань, 2024

КГМУ

План лекции

- Объекты и признаки
- Типы данных
- Структура данных
- *Представление данных*
- Большие данные
- Стандарты работы с данным, в том числе в клинической медицине

Объекты и признаки



- машинном обучении под *объектом* понимают то, зачем происходит наблюдение. Поэтому предполагают:

Объект ≡ Наблюдение ≡ Случай ≡ Ситуация ≡ Явление ≡ Образец ≡ Прецендент

- Пример объектов: пациенты, препараты, терапии, лечебные мероприятия
- *Признаком* называется результат измерения некоторой характеристики объекта.

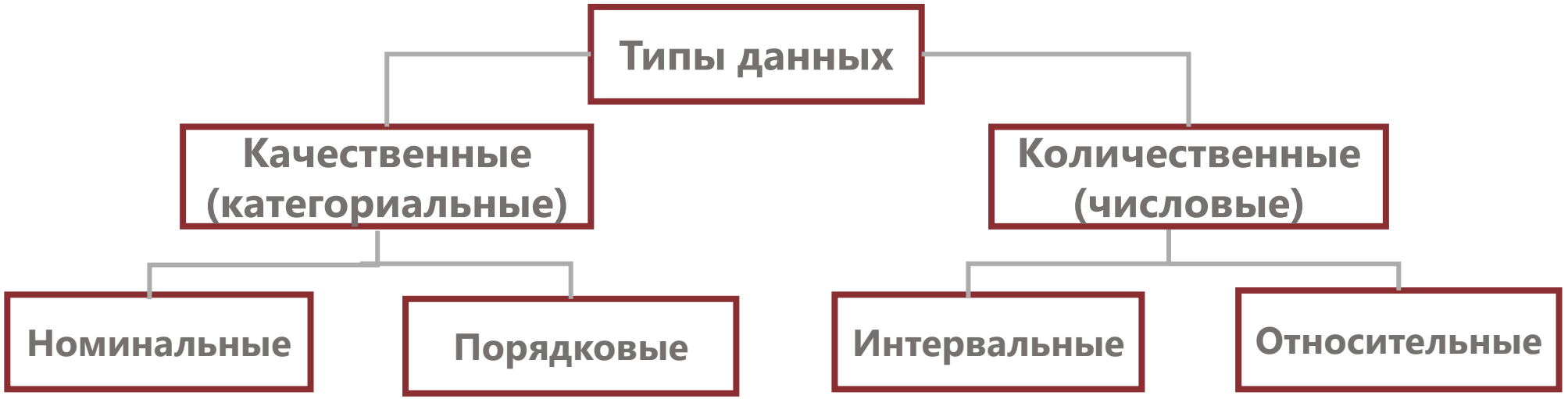
Признак ≡ Характеристика ≡ Фактор наблюдения ≡ Ковариата ≡ Feature

- Пример *признакового описания объекта* : пациенты, врачи, мероприятия
- *Признаковое описание объекта* —вектор, составленный из значений признаков данного объекта
- Пример признакового описания пациента: возраст, пол, диагноз, ИМТ, стадия ГБ
- Пример признакового описания терапии: препарат, способ введения, доза

Типы данных

Определение: Данные это значение, присвоенное определенному наблюдению или измерению.

Определение: Данные, используемые для описания интересующего аспекта совокупности, называют параметром или единицами совокупности (явления).



Качественные (атрибутивные) признаки не поддаются количественному (числовому) выражению.

Количественные признаки можно разделить на прерывные (дискретные) и непрерывные.

Структура данных

Скаляр, точка, величина (Value) – представляется одним числом или логической переменной (TRUE / FALSE) или словом/текстовым выражением

Вектор данных (строка или столбец) (5 7 21,1 6) (“Алексей” “Мария” “Ренат”)

Матрица – объединение строк и столбцов – таблица значений

$$\begin{matrix} 5{,}0 & 2{,}3 & 6{,}0 \\ 0{,}3 & 9{,}5 & 8{,}1 \\ 2{,}2 & 4{,}5 & 4{,}2 \end{matrix}$$

5,0	2,3	6,0
0,3	9,5	8,1
2,2	4,5	4,2

Датафрейм -это набор данных в виде именованных столбцов, то есть таблицы, в которых столбцы имеют наименование и определение типа данных.

Датасет (набор данных) — это набор строго типизированных структурированных данных, является расширением Датафрейма.

id	ФИО	гемоглобин	АГ
1	ИОН	110	Нет
2	МИР	125	Да
3	СКА	141	

Структурированная и неструктурированная информация

Структурированная информация – машиночитаемые данные, легко извлекаемые, имеющие определенный формат.

Данные холтер-мониторирования, данные, снятые с анализаторов по результатам лабораторно-клинических обследований.

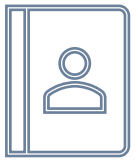
Полуструктурированная информация – частично структурированные данные, данные, которые имеет некоторые последовательные и определенные характеристики.

Данные по иерархии вложенности данных

Неструктурированная информация – полностью необработанные данные (изображения, звук, текст, видео).

Результаты ЭКГ, снимки при эндоскопии, текстовое заключение клинического психолога.

Для извлечения данных из неструктурированной информации применяют технологию **извлечения признаков (Feature Extraction)** , то есть генерацию признаков.



Представление данных

Наблюдения (observations)

Переменные (variables)

Пациент	Уровень биомаркера	Число стентированных сосудов	Систолическое давление	Пол
ММТ	21	6	110	муж
ИИР	22.8	4	93	жен
САП	21.4	6	110	муж
МХГ	14.3	8	245	муж
ААГ	24.4	4	62	муж
ОНТ	22.8	4	95	жен
ААВ	19.2	6	123	муж
ЛИА	17.8	6	123	муж
ПДВ	16.4	8	180	жен

Категориальные данные

Число стентированных сосудов

	Количество
4	3
6	4
8	2

Количественные данные

Уровень биомаркера

Мин	14.3
Среднее	20.0
Медиана	21.0
Макс	24.4

Основные статистики:

- Минимальное и максимальное
- Среднее и медиана
- Среднеквадратическое отклонение
- Квартили

Перекрестные данные

Перекрёстные данные (cross-sectional data) – это данные, которые приходят от разных групп или людей в один и тот же момент времени.

Регион	Население	Доля населения моложе 50 лет	Доля привитых от кори	Плотность населения
Москва	12387573	0,419	84,8	4925,92
Санкт-Петербург	5367912	0,408	86	3837,41
Самарская область	3188276	0,379	81,8	59,42
Челябинская область	3484395	0,355	75,3	39,26
Тюменская область	1508737	0,39	76,5	9,48
Татарстан	3896456	0,402	81,8	57,46

Временные ряды

Под временным рядом понимают последовательность некоторых наблюдений, упорядоченных во времени.

Они обычно собираются с фиксированными интервалами, например, ежедневно, ежемесячно, ежегодно и т.д.

Особенность формирования набора данных в виде временного ряда: *присвоении каждому наблюдению метки – вехи времени*.

Дата	Ежесуточное число заболевших гриппом	Общее число заболевших
20.02.2024	88	1212
21.02.2024	173	1385
22.02.2024	330	1715
23.02.2024	340	2055
24.02.2024	374	2429
25.02.2024	335	2764
26.02.2024	568	3332
27.02.2024	956	4288
28.02.2024	1085	5373
29.02.2024	1416	6789
01.03.2024	1775	8564

Панельные данные

Регион	Год	Число лесных пожаров	Площадь лесных земель, пройденная пожарами, Га
Ханты-Мансийский АО	2016	454	8036
Ямало-Ненецкий АО	2016	407	49367
Иркутская область	2016	1212	709035
Ханты-Мансийский АО	2017	387	53924
Ямало-Ненецкий АО	2017	284	171602
Иркутская область	2017	1209	917380

Источник: Федеральная служба государственной статистики (Росстат)

Панельные данные состоят из наблюдений одних и тех же экономических единиц, которые осуществляются в последовательные периоды времени.

У панельных данных три измерения:
- Объекты
- Переменные
- Время

Панельные данные могут быть сбалансированы, когда все объекты наблюдаются во все периоды времени, или несбалансированными, когда некоторые объекты наблюдаются не во все периоды времени.

Цензурированные данные

ФИО пациента	Период под наблюдением	Наступило событие (1) или нет (0)	Цензурированное событие
ИОЖ	24	1	
МСА	15	0	+
ВАН	60	0	Не полное
ОМК	60	1	
АОС	60	0	Не полное
ИНЕ	20	1	

Здесь время наблюдения – 60 месяцев
Цензурирование справа – то есть пациент выбыл из наблюдения и мы не знаем произошло ли событие или нет

В анализе выживаемости могут быть использованы цензурированные данные. Существует несколько направлений цензурирования:

- Цензурирование слева (left truncation): отсутствуют данные о начале исследования.
- Цензурирование справа (right censoring): событие (отказ, смерть) для объекта не наблюдается в течение периода исследования.
- Интервальное цензурирование (interval censoring): нет возможности установить точную дату наступления события.

Большие данные

Согласно ГОСТ Р ИСО/МЭК 20546-2021 БОЛЬШИЕ ДАННЫЕ

большие данные (big data): Большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.

облачные вычисления (cloud computing): Парадигма для предоставления возможности сетевого доступа к масштабируемому и эластичному пулу общих физических или виртуальных ресурсов с предоставлением самообслуживания и администрированием по требованию.

кластер (в распределенной обработке данных) (cluster): Совокупность функциональных устройств, находящихся под общим управлением. Здесь имеется ввиду вычислительный кластер.

данные (data): Представление информации в формальном виде, пригодном для передачи, интерпретации или обработки.

аналитика данных (data analytics): Составное понятие, охватывающее получение, сбор, проверку и обработку данных, включая их количественную оценку, визуализацию и интерпретацию.

Большие данные

Согласно ГОСТ Р ИСО/МЭК 20546-2021 БОЛЬШИЕ ДАННЫЕ

база данных (database): Совокупность данных, организованная в соответствии с концептуальной структурой, в которой описываются характеристики этих данных и взаимосвязи между представляемыми ими сущностями для одной или нескольких областей применения.

модель данных (data model): Схема данных, структурированная в базе данных в соответствии с формальными описаниями в информационной системе и требованиями используемой системы управления базой данных.

обработка данных (data processing): Систематическое выполнение операций с данными

1 Арифметические или логические операции с данными, объединение или сортировка данных или такие операции с текстом, как редактирование, сортировка, объединение, хранение, извлечение, отображение или печать.

2 Термин "обработка данных" не должен использоваться в качестве синонима для термина "обработка информации".

наука о данных (data science): Извлечение практических знаний из данных посредством исследования или создания и проверки гипотез.

массив данных (data set, dataset): Идентифицируемая совокупность данных, к которой можно получить доступ или скачать в одном или нескольких форматах.

тип данных (data type, datatype): Совокупность объектов данных установленной структуры и набора допустимых операций над этими объектами

Стандарты работы с данными

CRISP-DM (от английского Cross-Industry Standard Process for Data Mining) — межотраслевой стандартный процесс исследования данных.
Методология, первая версия которой была представлена в Брюсселе в марте 1999 года, а пошаговая инструкция опубликована в 2000 году.
CRISP-DM описывает жизненный цикл исследования данных, состоящий из 6 фаз.

Стандарты работы с данными

разметка [аннотация] данных (data labeling): Этап обработки структурированных и неструктурированных данных, в процессе которого данным (в том числе текстовым документам, фото- и видеоизображениям) присваиваются идентификаторы, отражающие тип данных (классификация данных), и (или) осуществляется интерпретация данных для решения конкретной задачи, в том числе с использованием СИИ

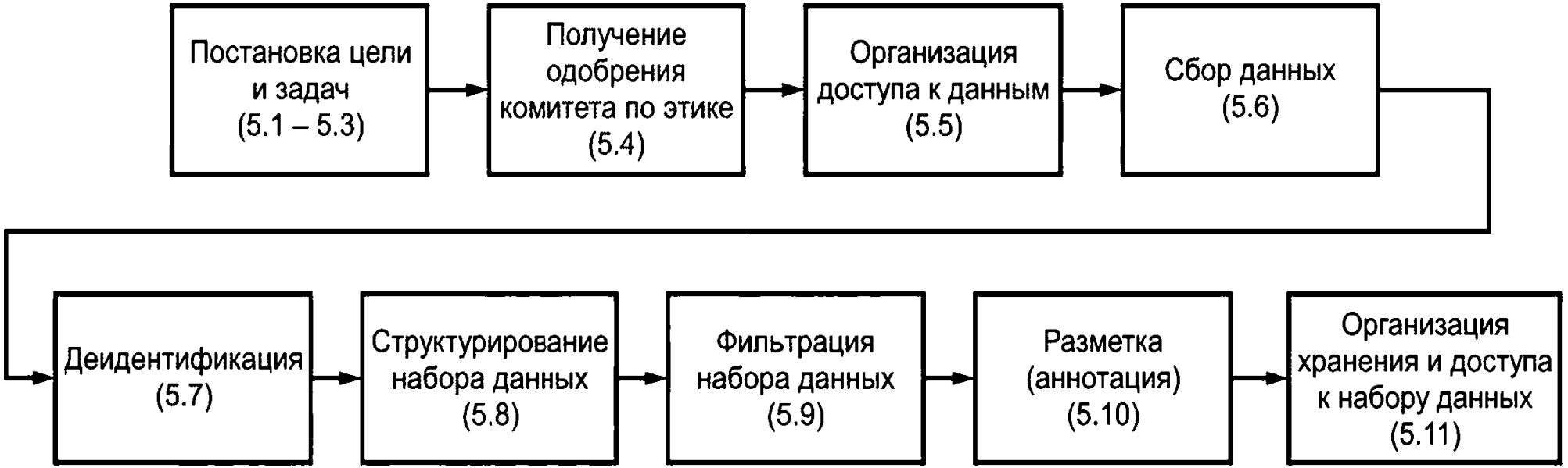
ретроспективная разметка (или естественная) (retrospective annotation): Сбор данных в соответствии с указанными метаданными, перечень которых выбирают в соответствии с поставленной целью формирования набора данных

проспективная разметка (prospective annotation): Сбор данных в соответствии с поставленной целью формирования набора данных, а также проведение дополнительных манипуляций с элементами

сбор данных (data collection): Процесс объединения данных, поступающих из одного или более источников, в целях их использования при обучении и тестировании СИИ

Стандарт РФ ГОСТ Р 59921.5-2022 «Системы искусственного интеллекта в клинической медицине. Часть 5. **Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов**»

Стандарты работы с данными



Стандарт РФ ГОСТ Р 59921.5-2022 «Системы искусственного интеллекта в клинической медицине. Часть 5. **Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов**»

Спасибо за внимание!