Интеллектуальный анализ данных на транспорте

Лекция 1

2022

Введение

Интеллектуальный анализ данных

(глубинный анализ данных)

— это собирательное название, используемое для обозначения совокупности методов обнаружения в данных новых знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Термин «интеллектуальный анализ данных» эквивалентен английскому термину «data mining»

Термин введён Григорием Пятецким-Шапиро в 1989 году

KDnuggets.com.

Г. Пятецкий - Шапиро считается основателем двух направлений:

«data mining» = «интеллектуальный анализ данных»

и

«knowledge discovery in data» = «обнаружение знаний в базах данных».

https://en.wikipedia.org/wiki/Gregory_Piatetsky-Shapiro

Далее мы будем использовать термины как синонимы:

data mining

интеллектуальный анализ данных

ИАД

KDD

knowledge discovery in data

извлечение данных

обнаружение знаний в базах данных

Методы ИАД

- Методы классификации, моделирования и прогнозирования,
- Построение деревьев решений,
- Искусственные нейронные сети,
- Генетические алгоритмы,
- Эволюционное программирование,
- Ассоциативная память,
- Нечёткая логика.

К методам data mining относят *статистические методы* :

- дескриптивный (описательный) анализ,
- корреляционный и регрессионный анализ,
- факторный анализ,
- дисперсионный анализ,
- компонентный анализ,
- дискриминантный анализ,
- анализ временных рядов,
- анализ выживаемости,
- анализ связей

Выбор конкретного метода зависит от задачи

- Статистические методы предполагают некоторые априорные представления об анализируемых данных
- Это несколько расходится с первоначальными целями *data mining*
- Основная цель - обнаружение ранее неизвестных нетривиальных и практически полезных знаний

- Одно из важнейших назначений методов data mining состоит в наглядном представлении результатов вычислений (визуализация)
- Это позволяет использовать инструментарий data mining людьми, не имеющими специальной математической подготовки.

- Применение статистических методов анализа данных требует хорошего владения теорией вероятностей и математической статистикой
- Методы data mining лежат на стыке баз данных, статистики и искусственного интеллекта.

Первоначально задача ИАД ставится следующим образом:

имеется достаточно крупная база данных;

предполагается, что в базе данных находятся некие «скрытые знания».

Необходимо разработать методы обнаружения знаний, скрытых в больших объёмах исходных «сырых» данных.

В текущих условиях глобальной конкуренции именно найденные закономерности (найденные знания) могут быть источником дополнительного конкурентного преимущества.

Что означает «скрытые знания»?

ранее неизвестные — то есть такие знания, которые должны быть новыми, а не подтверждающими какие-то ранее полученные сведения

Что означает «скрытые знания»?

нетривиальные — то есть такие, которые нельзя просто так увидеть при непосредственном визуальном анализе данных или при вычислении простых статистических характеристик

Что означает «скрытые знания»?

практически полезные — то есть такие знания, которые представляют ценность для исследователя или потребителя;

Что означает «скрытые знания»?

доступные для интерпретации — то есть такие знания, которые легко представить в наглядной для пользователя форме и легко объяснить в терминах предметной области.

Эти требования во многом определяют суть методов ИАД и то,

в каком виде и

в каком соотношении в технологии data mining используются системы управления базами данных, статистические методы анализа и методы искусственного интеллекта.

- Методы data mining могут быть применены как для работы с <u>большими данными</u>, так и для обработки сравнительно малых объемов данных

(полученных, например, по результатам отдельных экспериментов, либо при анализе данных о деятельности компании)

- Критерием достаточного количества данных может быть как область исследования, так и применяемый алгоритм анализа

Знания, добываемые методами data mining, принято представлять в виде *закономерностей (паттернов)*.

В качестве таких выступают:

- ассоциативные правила;
- деревья решений;
- кластеры;
- математические функции.

Задачи, решаемые методами data mining, принято разделять на описательные (*descriptive*) и предсказательные (*predictive*).

В *описательных* задачах самое главное — это дать наглядное описание имеющихся скрытых закономерностей,

в *предсказательных* задачах на первом плане стоит вопрос о предсказании тех случаев, для которых данных ещё нет.

К *описательным* задачам относятся:

- поиск ассоциативных правил или паттернов (образцов);
- группировка объектов, кластерный анализ;
- построение регрессионной модели.

К *предсказательным* задачам относятся:

классификация объектов (для заранее заданных классов);

регрессионный анализ, анализ временны́х рядов.

- Перед использованием алгоритмов data mining необходимо произвести подготовку набора анализируемых данных.
- Так как ИАД может обнаружить только присутствующие в данных закономерности, исходные данные с одной стороны должны иметь достаточный объём, чтобы эти закономерности в них присутствовали, а с другой — быть достаточно компактными, чтобы анализ занял приемлемое время.
- Чаще всего в качестве исходных данных выступают хранилища или витрины данных.
- Далее данные фильтруются. Фильтрация удаляет выборки с шумами и пропущенными данными.

Интеллектуальный анализ данных на транспорте

Лекция 2

2022

Процесс интеллектуального анализа данных

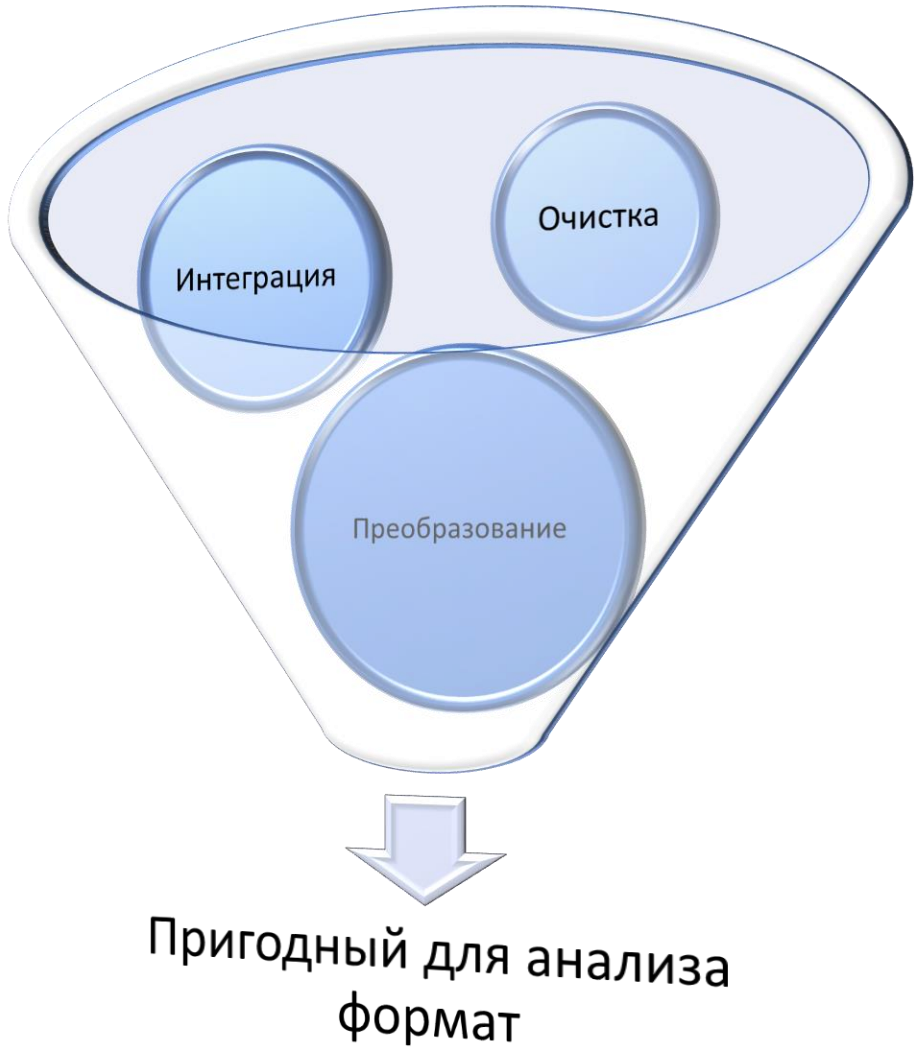
1. Изучение предметной области. В результате формулируются основные цели анализа.
2. Сбор данных. Формируется база данных или датасет (data set).
3. Предварительная обработка данных. Данные приводятся к виду, удобному для следующих этапов.

Процесс интеллектуального анализа данных

4. Анализ данных. Применяются алгоритмы интеллектуального анализа с целью извлечения паттернов.
5. Интерпретация найденных паттернов. Данный этап может включать визуализацию извлеченных паттернов, определение действительно полезных паттернов на основе некоторой функции полезности.
6. Использование новых знаний.

Предварительная обработка данных

a) Очистка данных – исключение противоречий и случайных "шумов“ из исходных данных
b) Интеграция данных – объединение данных из нескольких возможных источников в одном хранилище
c) Преобразование данных



Предварительная обработка данных

На данном этапе данные преобразуются к форме, подходящей для анализа.

Часто применяется:

- агрегация данных
- дискретизация атрибутов
- сжатие данных
- сокращение размерности

- Очистка данных занимается обнаружением и удалением ошибок и несоответствия данных в целях повышения качества данных.
- Проблемы качества данных могут происходить из разных источников, таких как системы федеративных баз данных, информационные системы на базе Интернета или просто из-за ошибочного ввода данных

Качество данных

ГОСТ Р ИСО/ТС 8000-1-2009 **КАЧЕСТВО ИНФОРМАЦИОННЫХ ДАННЫХ**

Качество данных - это показатель того, насколько набор данных подходит для достижения конкретной цели и насколько он надежен для принятия надежных решений.

Качество данных

Характеристики данных, из которых складывается их качество:

- Доступность
- Точность
- Взаимосвязанность
- Полнота
- Непротиворечивость
- Однозначность
- Релевантность
- Своевременность

Качество данных

- Доступность

У аналитика должен быть доступ к данным.

Это предполагает не только разрешение на их получение, но также наличие соответствующих инструментов, обеспечивающих возможность их использовать и анализировать.

В настоящее время «аналитик данных» – это отдельная профессия

Качество данных

- Точность

Данные должны отражать истинные значения или положение дел. Например, показания неправильно настроенного термометра, ошибка в дате рождения или устаревший адрес – это все примеры неточных данных.

Данные должны быть собраны и обработаны с необходимой степенью детализации

Качество данных

- Взаимосвязанность

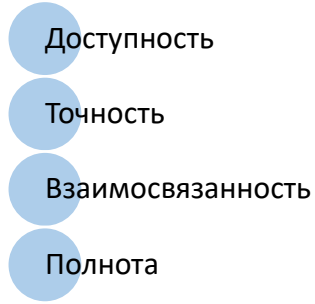
Должна быть возможность точно связать одни данные с другими.

Например, заказ клиента должен быть связан с информацией о нем самом, с товаром или товарами из заказа, с платежной информацией и информацией об адресе доставки.

Этот набор данных обеспечивает полную картину заказа клиента.

Взаимосвязь обеспечивается набором идентификационных кодов или ключей, связывающих воедино информацию из разных частей базы данных.

Качество данных



- Полнота

Под неполными данными может подразумеваться как отсутствие части информации , так и полное отсутствие единицы информации

например, в сведениях о клиенте не указано его имя

или в результате ошибки при сохранении в базу данных потерялась вся информация о клиенте

Качество данных

- Непротиворечивость

Данные должны быть согласованными.

Например, адрес конкретного клиента в одной базе данных должен совпадать с адресом этого же клиента в другой базе.

При наличии разногласий один из источников следует считать основным или вообще не использовать сомнительные данные до устранения причины разногласий.

Качество данных

- Однозначность

Каждое поле, содержащее индивидуальные данные, имеет определенное, недвусмысленное значение.

Четко названные поля в совокупности со словарем базы данных помогают обеспечить качество данных.

Качество данных

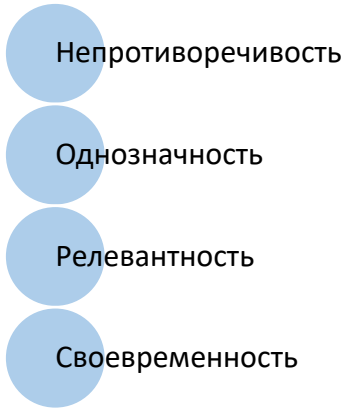
- Релевантность

Степень соответствия найденных данных информационным нуждам пользователя

Данные зависят от характера анализа.

Если нужно изучать транспортные потоки, то не обязательно изучать транзакции при покупках в одном отдельно взятом магазине.

Качество данных



- Своевременность

Между сбором данных и их доступностью для использования в аналитической работе всегда проходит время.

На практике это означает, что аналитики получают данные как раз вовремя, чтобы завершить анализ к необходимому сроку.

При задержке данные становятся практически бесполезными (при сохранении издержек на их хранение и обработку), их можно использовать только в целях долгосрочного стратегического планирования и прогнозирования.

Качество данных

Дополнительно выделяют

- Надежность данных

Данные должны быть одновременно

полными

(то есть содержать все сведения, которые вы ожидали получить)

и точными

(то есть отражать достоверную информацию).

Качество данных

Ошибка всего в одном из этих аспектов может привести к тому, что данные окажутся частично или полностью непригодными к использованию или, хуже того, будут казаться достоверными, но приведут к неправильным выводам.

Качество данных

Проблемы с качеством данных присутствуют в единичных коллекции данных, такие как файлы и базы данных, например, из-за неправильного написания при вводе данных, отсутствующей информации или другие неверные данные.

Когда необходимо интегрировать несколько источников данных, например, в хранилищах данных, объединить системы баз данных или глобальные информационные веб-системы, потребность в очистке данных возрастает существенно.

Это связано с тем, что источники часто содержат избыточные данные в разных представлениях.

Чтобы обеспечить доступ к точным и непротиворечивым данным, объединение различных представлений данных и устранение дублирующейся информации становится необходимо.

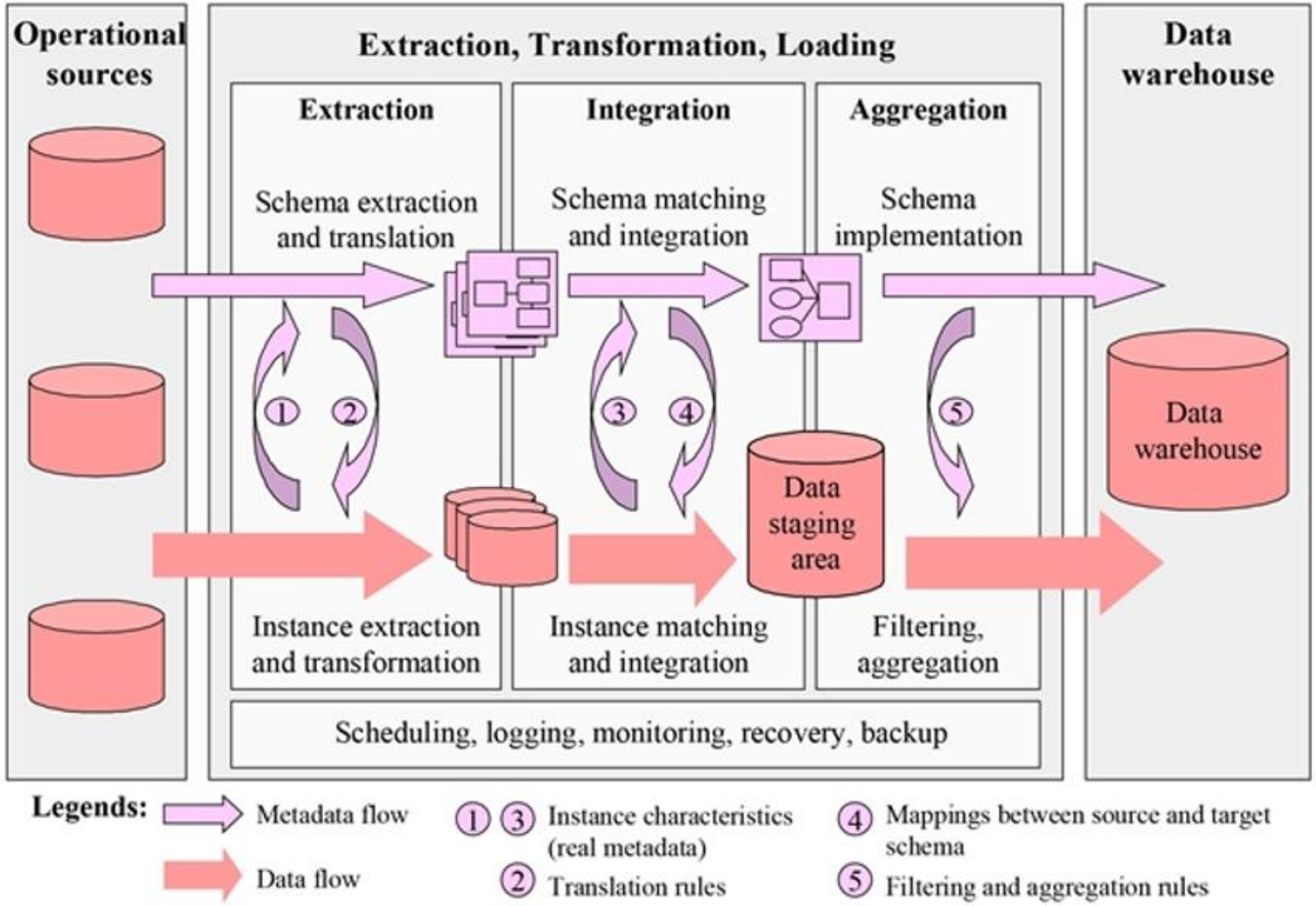


Figure 1. Steps of building a data warehouse: the ETL process

Проблема качества данных

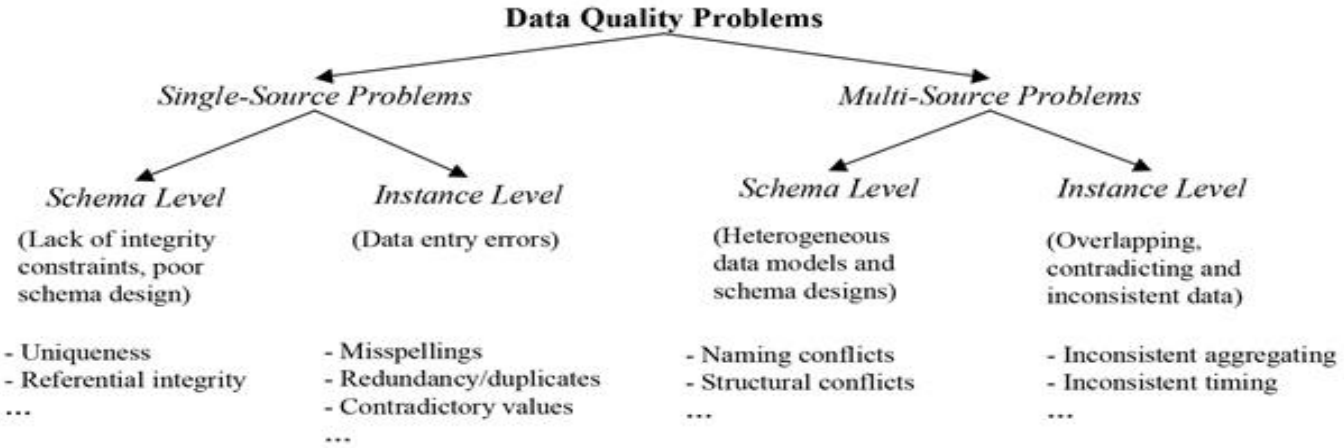


Figure 2. Classification of data quality problems in data sources

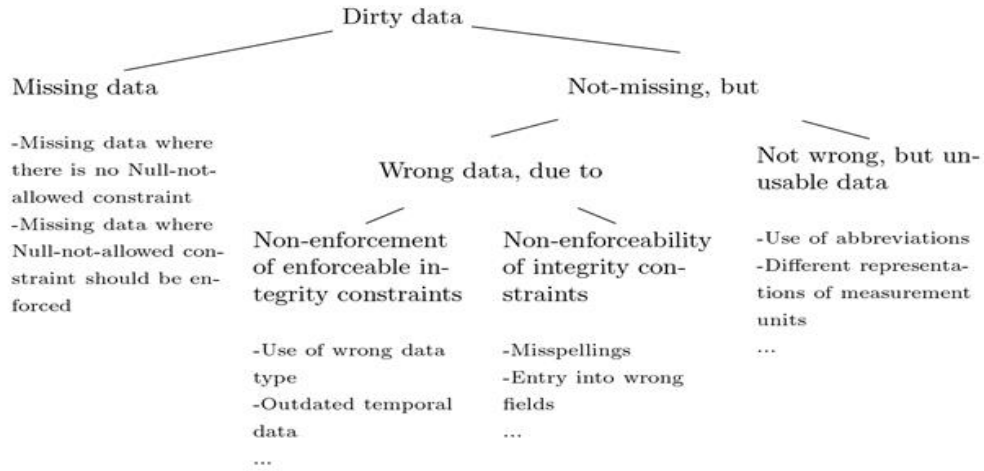
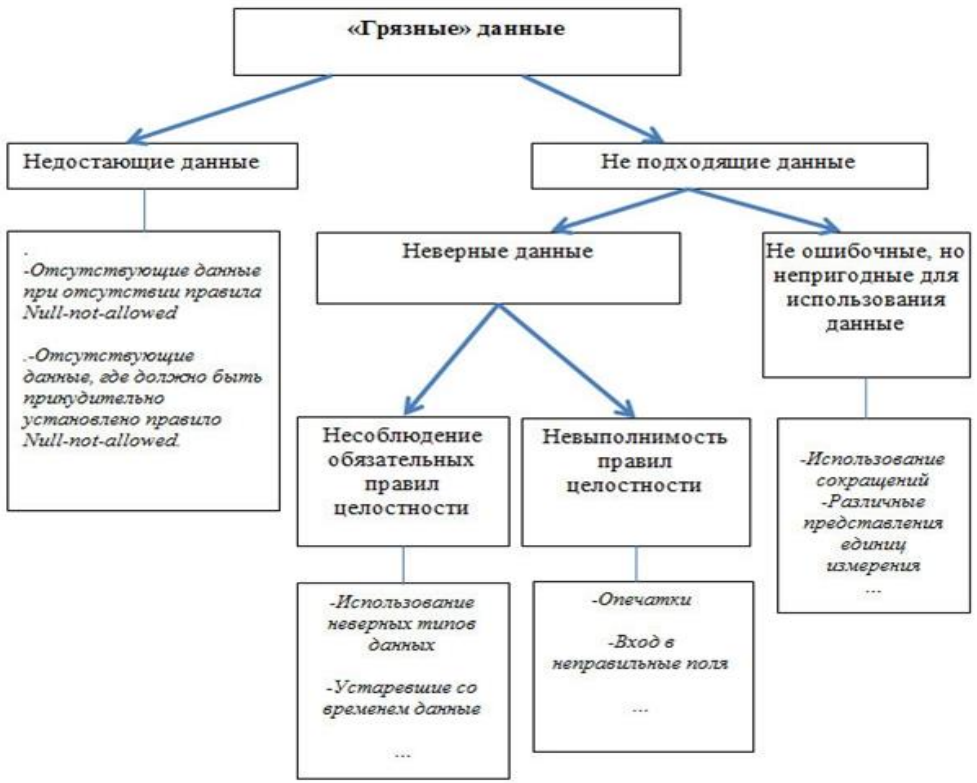


Fig. 2. Classification of dirty data by Kim et al. [2]



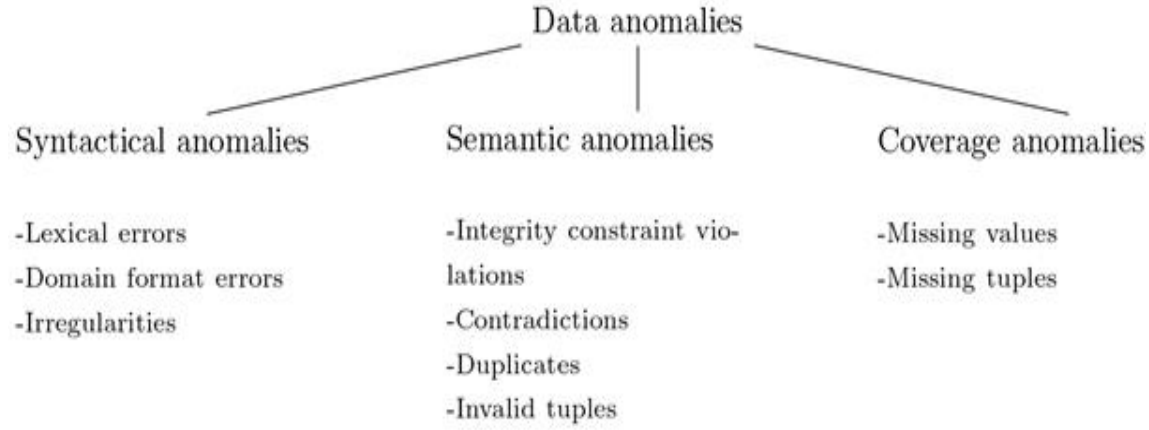


Fig. 3. Classification of data anomalies by Müller and Freytag [3]

Этапы очистки данных

- Анализ данных
- Определение рабочего процесса преобразования и правил сопоставления
- Преобразование данных
- Проверка
- Трансформация
- Обратный поток очищенных данных

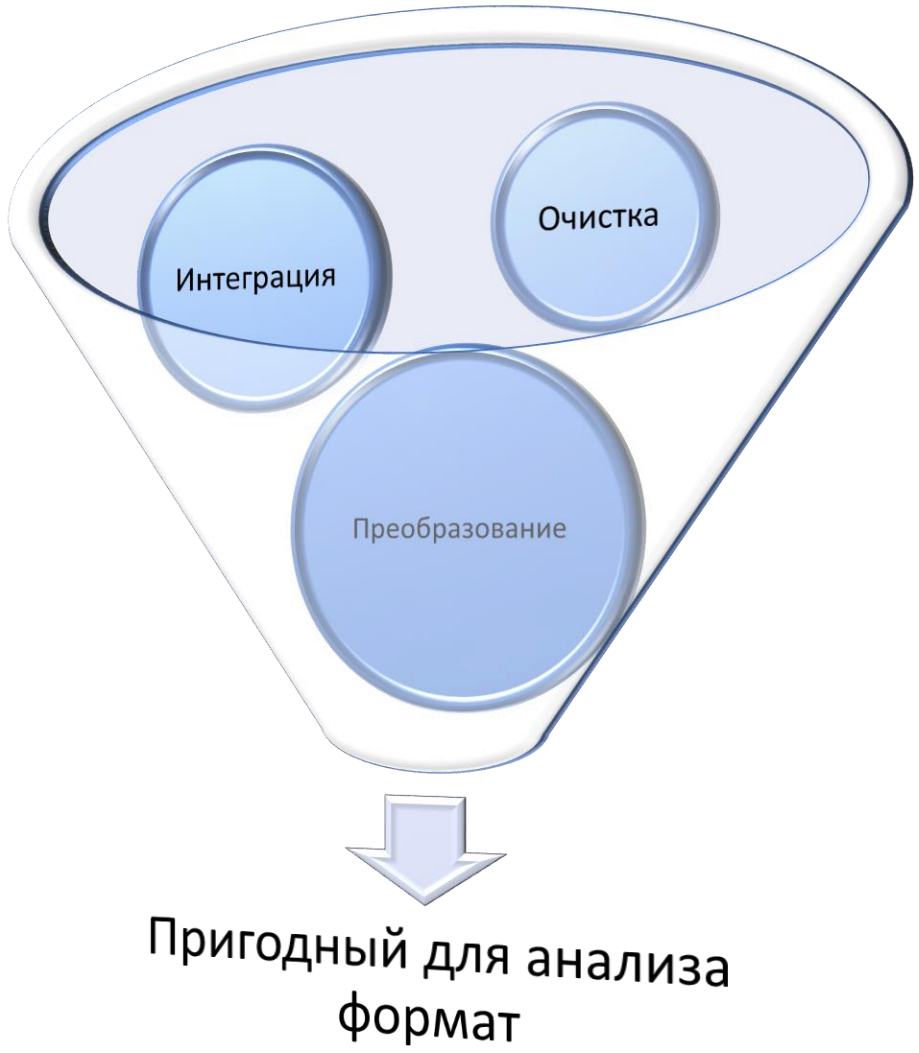
Интеллектуальный анализ данных на транспорте

Лекция 3

2022

Предварительная обработка данных

a) Очистка данных – исключение противоречий и случайных "шумов"из исходных данных
b) Интеграция данных – объединение данных из нескольких возможных источников в одном хранилище
c) Преобразование данных



Очистка данных

- Проверка
- Трансформация
- Обратный поток очищенных данных

Очистка данных включает различные методы, основанные на проблемах и типах данных.

Различные методы могут быть применены с каждым атрибутом данных, но имеют свои собственные особенности применения.

В целом, неверные данные либо удаляются, либо исправляются, либо присваиваются.

Очистка данных

- Проверка

Требуются определения преобразований, которые должны быть сделаны в процессе работы с данными.

Правильность и эффективность рабочего процесса следует тестировать и оценивать.

Тестирование проводится на некотором образце или копии исходных данных, чтобы при необходимости улучшить определения.

Может потребоваться несколько итераций этапов анализа, проектирования и проверки.

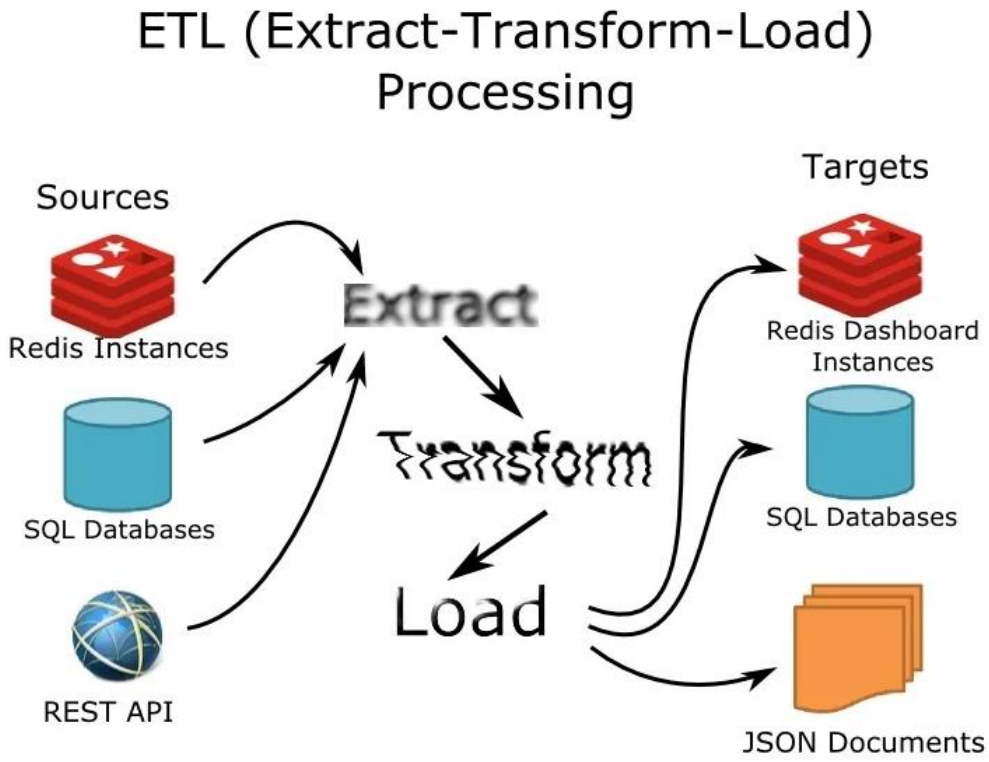
Некоторые ошибки в данных становятся очевидными только после применения преобразований.

Очистка данных

- Трансформация

Выполнение шагов преобразования либо путем запуска рабочего процесса для загрузки и обновления хранилища данных, либо во время ответа на запросы из нескольких источников.

ETL - *Extract, Transform, Load* — дословно «извлечение, преобразование, загрузка»



Очистка данных

- Обратный поток очищенных данных

После удаления ошибок очищенные данные также должны заменять «грязные данные» в исходных источниках,

чтобы избежать повторения работы по очистке для будущего извлечения данных.

Для хранилищ данных очищенные данные доступны из области подготовки данных

Нерелевантные данные — Irrelevant data

Нерелевантные данные — это те, которые на самом деле не нужны и не вписываются в контекст проблемы, которую мы пытаемся решить.

- Например, если бы мы анализировали данные об общем состоянии здоровья населения, номер телефона не был бы необходим.
- **Только если** вы уверены, что часть данных не важна, вы можете удалить ее.
- В противном случае изучите матрицу корреляции между атрибутами объекта.
- И даже если вы не заметили никакой корреляции, вы должны спросить кого-то, кто является экспертом в предметной области. Вы никогда не знаете, что функция, которая кажется неактуальной, может быть очень актуальной с точки зрения предметной области, например с клинической точки зрения.

Дубликаты — Duplicates

- **Дубликаты** — это значения, которые повторяются в вашем наборе данных (повторяться могут как транзакции, так и сущности в различных справочниках).

Дублирование часто случается, когда, например, данные объединены из разных источников

Пользователь может дважды нажать кнопку «Отправить», думая, что форма фактически не была отправлена.

Распространенным симптомом является случай, когда два пользователя имеют одинаковый идентификационный номер.

И поэтому они просто должны быть удалены.

Тип преобразования — Type conversion

- Числа должны храниться в виде числовых типов данных.
- Дата должна храниться в виде объекта даты

Категориальные значения могут быть преобразованы в/из чисел при необходимости.

Предупреждение: значения, которые нельзя преобразовать в указанный тип, следует преобразовать в специальное значение NA с отображением **предупреждения**. Это указывает на неправильное значение, которое должно быть исправлено.

Синтаксические ошибки — Syntax errors

Удалить пробелы:

следует удалить лишние пробелы в начале или конце строки.

Исправьте опечатки:

строки могут быть введены разными способами,
и неудивительно, что могут быть ошибки.
Для исправления ошибок с помощью встроенных алгоритмов
разрабатывают правила.
Правила могут отличаться в разных предметных областях.

- Например, “lisbon” можно ввести как “lisboa”, “lisbona”, “Lisbon”, и т.д.
- City Distance **from** "lisbon"
- lisbon 0
- lisboa 1
- Lisbon 1
- lisbona 2
- london 3
- ...
- Если это так, то мы должны заменить все значения, которые означают одно и то же, на одно уникальное значение. В этом случае замените первые 4 строки на «lisbon».

Стандартизация данных

- Поместить каждое значение в один и тот же стандартизированный формат.

Для строк убедитесь, что все значения указаны в нижнем или верхнем регистре.

Для числовых значений убедитесь, что все значения имеют определенную единицу измерения.

Высота, например, может быть в метрах и сантиметрах.

Для дат, версия для США отличается от европейской версии.

Запись даты в качестве метки времени (количество миллисекунд) не совпадает с записью даты в качестве объекта даты.

Масштабирование / Преобразование — Scaling / Transformation

- Масштабирование означает преобразование данных таким образом, чтобы они соответствовали определенному масштабу, например 0–100 или 0–1.

Например, баллы по экзамену студента могут быть пересчитаны в процентах (0–100) вместо среднего балла (0–5).

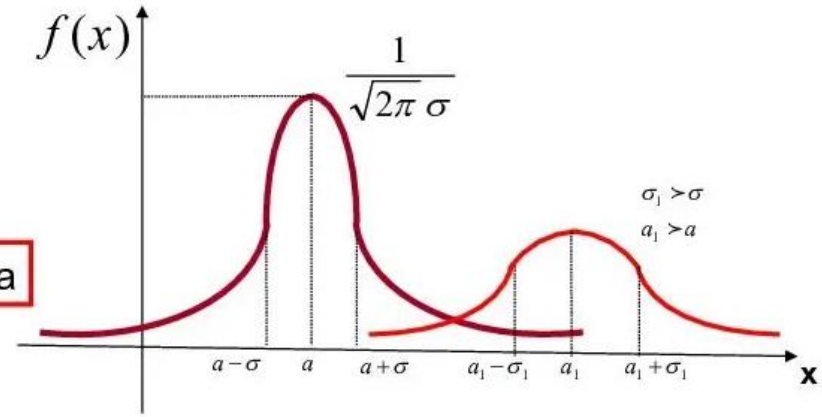
Масштабирование также может выполняться для данных, которые имеют разные единицы измерения.

Нормализация — Normalization

- Актуально для случайных величин, имеющих нормальное распределение при использовании статистических методов анализа.

Цель состоит в том, чтобы преобразовать данные так, чтобы они имели стандартное нормальное распределение.

- **График плотности распределения.**



- **Нормированное распределение.**

$$a = 0\,,\ \sigma = 1 \quad \Rightarrow \quad f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \varphi(x)$$

Недостающие значения — Missing values

Учитывая тот факт, что отсутствующие значения неизбежны,
возникает вопрос:
что делать, когда мы сталкиваемся с ними?
Игнорирование пропущенных данных — это то же самое,
что создание дыр в лодке — она будет тонуть.
Есть три, или, возможно, больше, способов справиться с ними.

Недостающие значения — Missing values

— Первый способ.

Удаление — Drop.

Если пропущенные значения в столбце встречаются редко и
происходят случайным образом, то самое простое и прямое решение —
отбросить наблюдения (строки) с пропущенными значениями.

Если большинство значений столбца отсутствуют и встречаются случайным образом,
типичным решением является удаление всего столбца.

Это особенно полезно при проведении статистического анализа,
поскольку заполнение пропущенных значений может привести
к неожиданным или необъективным результатам.

Недостающие значения — Missing values

— Второй способ.
Внести значение — Impute.
Это значит рассчитать недостающее значение на основе других наблюдений.
Есть много способов сделать это.
1. Используются **статистические значения,** такие как среднее значение, медиана.
В нормально распределенных данных можно сгенерировать все значения, которые находятся в пределах 2 стандартных отклонений от среднего.

Недостающие значения — Missing values

2. Использовать **линейную регрессию**.

На основании существующих данных можно рассчитать линию наилучшего соответствия между двумя переменными

Стоит отметить, что модели линейной регрессии чувствительны к выбросам.

3. Hot-deck: Копирование значений из других похожих записей. Это полезно, только если у вас достаточно доступных данных. И это может быть применено к числовым и категориальным данным.

Можно использовать случайный подход, где мы заполняем отсутствующее значение **случайным** значением.

Недостающие значения — Missing values

— Третий способ. Flag.

Некоторые утверждают, что заполнение пропущенных значений приводит к потере информации, независимо от того, какой метод вменения мы использовали.

Это потому, что утверждение о том, что данные отсутствуют, само по себе информативно, и алгоритм должен знать об этом. В противном случае мы просто усиливаем шаблон, уже существующий другими функциями.

Это особенно важно, когда пропущенные данные не случаются случайно.

В то время как **категориальные данные** могут быть заполнены, скажем, «Отсутствует»: новая категория, которая сообщает, что этот фрагмент данных отсутствует.

!!!

— Принимать во внимание …

Отсутствующие значения не совпадают со значениями по умолчанию. Например, ноль можно интерпретировать как отсутствующий или по умолчанию, но не оба.

Отсутствующие значения не являются «неизвестными». Проведенное исследование, в котором некоторые люди не помнят, подвергались ли они издевательствам в школе или нет, должно рассматриваться и обозначаться как неизвестное и не пропущенное.

Каждый раз, когда мы отбрасываем или приписываем значение, мы теряем информацию. Таким образом, пометка может прийти на помощь.

Выпадающие значения — Outliers

- Это значения, которые значительно отличаются от всех других наблюдений.

Любое значение данных, которое находится на расстоянии более (1,5 * IQR) от квартилей Q1 и Q3, считается выбросом.

Выбросы невиновны, пока их вина не доказана ☺

С учетом сказанного, их не следует удалять, если для этого нет веских причин.

Например, можно заметить некоторые странные, подозрительные значения, которые вряд ли произойдут, и поэтому решает удалить их. Тем не менее, они заслуживают расследования, прежде чем удалить.

Стоит также отметить, что некоторые модели, такие как линейная регрессия, очень чувствительны к выбросам. Другими словами, выбросы могут отбросить модель, из которой собрана большая часть данных.

Ошибки в записях и между наборами данных — In-record & cross-datasets errors

Эти ошибки возникают из-за наличия двух или более значений в одной строке или между наборами данных, которые противоречат друг другу.

Например, ребенок не может быть женат.

Заработная плата работника не может быть меньше рассчитанных налогов.

0 автомобилей не могут иметь скорость движения больше 0.

Та же идея применима к связанным данным в разных наборах данных.

Проверка (Верификация) — Verifying

- Когда все сделано, следует проверить правильность, повторно проверив данные и убедившись, что они соблюдают правила и ограничения.
- Например, после заполнения отсутствующих данных они могут нарушить любое из правил и ограничений.
- Это может потребовать некоторой ручной коррекции, если не возможно иначе.

Составление отчетов — Reporting

- Отчет о том, насколько «здоровы» данные, одинаково важен для очистки.
- Как упоминалось ранее, программные пакеты или библиотеки могут генерировать отчеты о внесенных изменениях, какие правила были нарушены и сколько раз.
- Помимо регистрации нарушений, следует учитывать причины этих ошибок. Почему они произошли в первую очередь?

Независимо от того, насколько надежен процесс проверки и очистки, данные будут «загрязняться» по мере поступления новых.

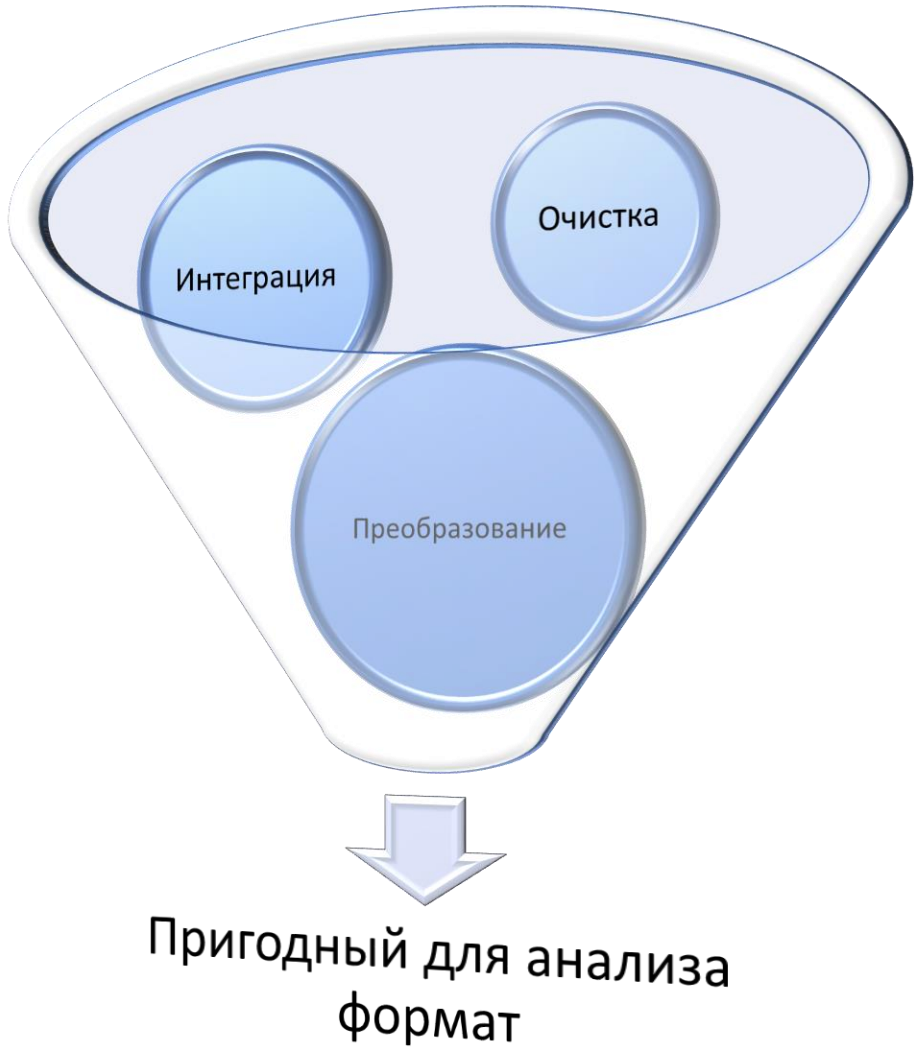
Интеллектуальный анализ данных на транспорте

Лекция 4

2022

Предварительная обработка данных

a) Очистка данных – исключение противоречий и случайных "шумов"из исходных данных
b) Интеграция данных – объединение данных из нескольких возможных источников в одном хранилище
c) Преобразование данных



Интеграция данных

- Интеграция данных в информационных системах понимается как обеспечение единого унифицированного интерфейса для доступа к некоторой совокупности неоднородных независимых источников данных

Интеграция данных

для пользователя информационные ресурсы всей новой интегрированной совокупности источников - это новый единый источник информации

- Система, обеспечивающая пользователю такие возможности, называется системой интеграции данных

Интеграция данных

Интегрируемыми источниками данных могут быть

- традиционные системы баз данных, поддерживающие различные модели данных (реляционные, объектные, объектно-реляционные, графовые и т.п.),
- разнообразные унаследованные системы,
- репозитории,
- веб-сайты,
- файлы структурированных данных.

Интеграция данных

- состав множества источников может быть наперед заданным или динамически пополняемым, источники данных могут обладать неизменным или обновляемым содержанием.

Интеграция данных

Первые шаги в этой области относятся еще к середине 70-х гг., когда начались разработки распределенных систем баз данных и когда сформировались более четкие представления

- о многоуровневой архитектуре систем баз данных,
- о моделях данных как инструменте моделирования реальности и
- об отображении моделей данных.

Интеграция данных

- Позднее несколько более общая форма этой задачи была связана с созданием
- мультибаз и федеративных баз данных,
- хранилищ данных,
- различных репозиториев информационных ресурсов, а также
- веб-приложений.

Уровни интеграции данных

Системы интеграции данных могут обеспечивать интеграцию данных на

- физическом,
- логическом и
- семантическом уровне.

Уровни интеграции данных

- Интеграция данных на физическом уровне с теоретической точки зрения является наиболее простой задачей и сводится к конверсии данных из различных источников в требуемый единый формат их физического представления.

Уровни интеграции данных

- Интеграция данных на логическом уровне предусматривает возможность доступа к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление с учетом структурных и, возможно, поведенческих (при использовании объектных моделей) свойств данных.

Уровни интеграции данных

- интеграция данных на семантическом уровне обеспечивает поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области.

Уровни интеграции данных

- Источники данных могут обладать различными свойствами, существенными для выбора методов интеграции данных
- Они могут поддерживать представление данных в терминах той или иной модели данных, могут быть статическими или динамическими и т.п.
- Множество источников интегрируемых данных может быть однородным или неоднородным относительно характеристик, соответствующих используемому уровню интеграции.

Подходы к интеграции данных

- Возможны два подхода к интеграции данных – виртуальное или актуальное (материализованное) представление интегрированных данных.
- При виртуальном подходе создается механизм доступа, который при обработке пользовательского запроса порождает данные в требуемом представлении непосредственно из источников данных.

Полное материализованное представление интегрированных данных в терминах единого пользовательского интерфейса при этом не поддерживается.

Виртуальный подход чаще всего применяется при использовании часто обновляемых источников данных.

- При актуальном подходе на стадии интеграции формируется полное материализованное представление интегрированных данных, отчужденное от исходных источников и сосуществующее с ними.

Именно это представление данных используется для обработки пользовательских запросов. Такой подход используется, в частности, в хранилищах данных.

Задачи интеграции данных

- Разработка архитектуры системы интеграции данных.
- Создание интегрирующей модели данных, являющейся основой единого пользовательского интерфейса в системе интеграции.
- Разработка методов отображения моделей данных и построение отображений в интегрирующую модель для конкретных моделей, поддерживаемых отдельными источниками данных.
- Интеграция метаданных, используемых в системе источников данных.
- Преодоление неоднородности источников данных.
- Разработка механизмов семантической интеграции источников данных.

Инструменты интеграции данных

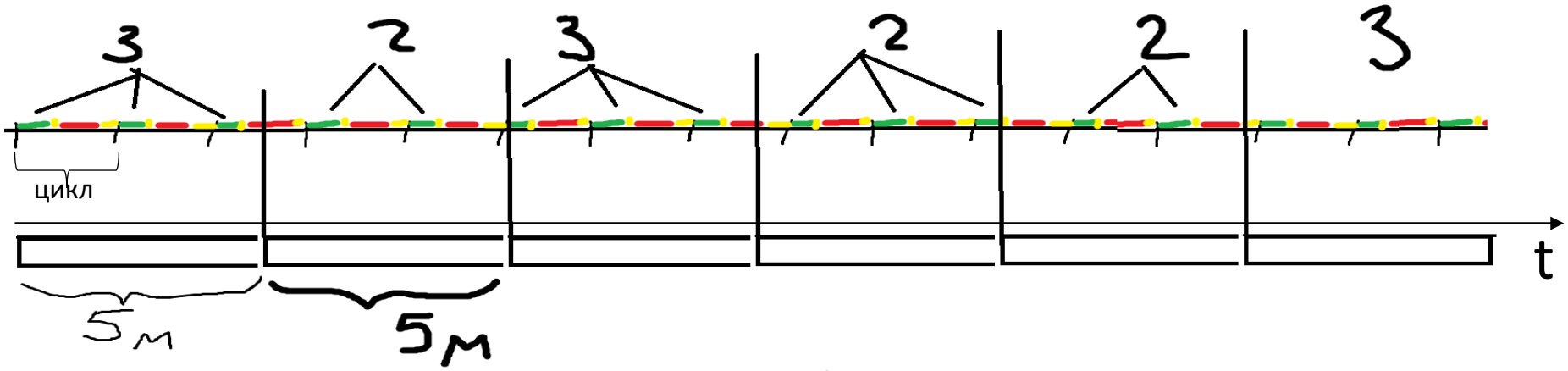
К числу основных средств, используемых для обеспечения интеграции информационных ресурсов, относятся

- конверторы данных,
- интегрирующие модели данных,
- механизмы отображения моделей данных,
- объектные адаптеры (Wrappers),
- посредники (Mediators),
- онтологические спецификации,
- средства интеграции схем и интеграции онтологических спецификаций
- архитектура, обеспечивающая взаимодействие средств, используемых в конкретной системе интеграции ресурсов.

Светофорные циклы + детекторы

Светофорные циклы

Детекторы



Интеллектуальный анализ данных на транспорте

Лекция 6

2022

Статистические методы ИАД

К методам data mining относят *статистические методы* :

- дескриптивный (описательный) анализ,
- корреляционный и регрессионный анализ,
- факторный анализ,
- дисперсионный анализ,
- компонентный анализ,
- дискриминантный анализ,
- анализ временных рядов,
- анализ выживаемости,
- анализ связей

Выбор конкретного метода зависит от задачи

Дескриптивный (описательный) анализ

Это анализ с помощью самых простых статистических характеристик.

Выбор статистической характеристики зависит от шкалы измерения исследуемой величины.

Шкалы измерений :

- Номинальная
- Порядковая
- Интервальная
- Относительная

качественные

Например, если измерить атрибут «Транспорт» в номинальной шкале, то она будет выглядеть так: 1 – автобус; 2 – такси; 4 – велосипед.

«Отлично», «Очень хорошо», «Хорошо», «Плохо», «Очень плохо»

количественные

Cnt0	счетчик коротких до 3 м
Cnt1	счетчик легковых от 3 до 5,5 м
Cnt2	счетчик от 5,5 до 7 м
Cnt3	счетчик от 7 до 10 м
Cnt4	счетчик от 10 до 15 м
Cnt5	счетчик свыше 15 м

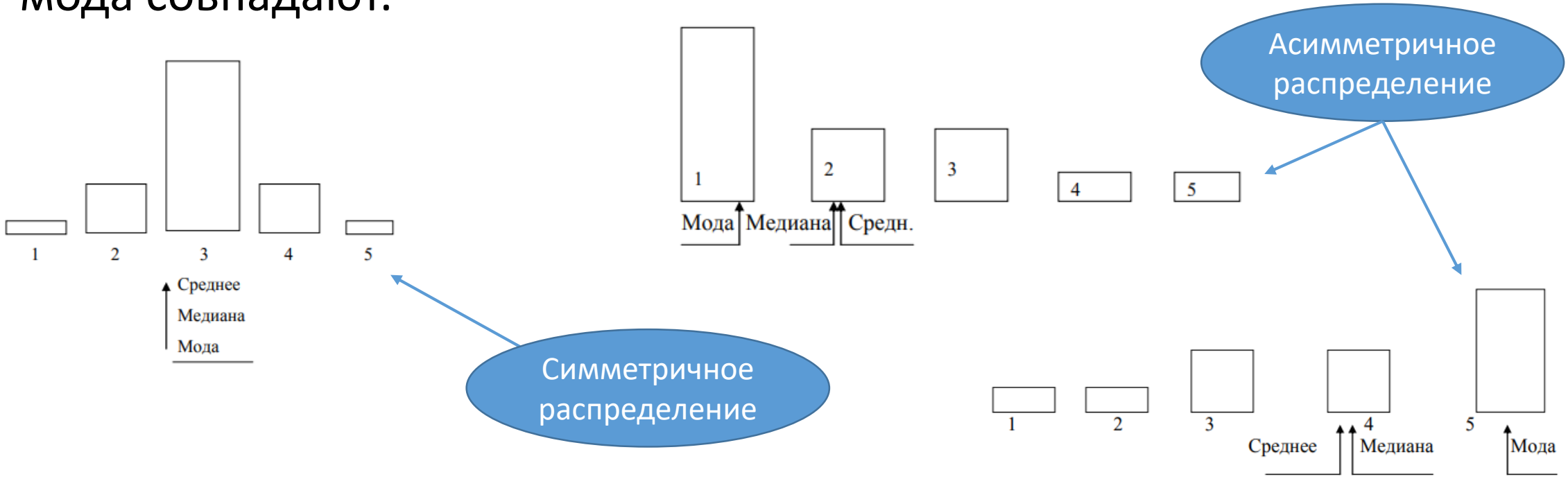
Время, скорость, температура и прочие физические величины

Дескриптивный (описательный) анализ

Шкала / Описательная характеристика	Номинальная	Порядковая	Интервальная	Относительная
Распределение частот	+	+	+	+
Доля	+	+	+	+
Процент	+	+	+	+
Пропорция	+	+	+	+
Мода	+	+	+	+
Медиана		+	+	+
Среднее			+	+

Дескриптивный (описательный) анализ

- Соотношение среднего, моды и медианы. Среднее, мода и медиана дают различное видение характеристик ряда. Распределение будет симметричным, если среднее, медиана и мода совпадают.



Корреляционный анализ

измеряется теснота связи между двумя или более переменными

- **Линейный коэффициент корреляции**

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}. \qquad \overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t, \overline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t$$

Корреляционный анализ

- **Коэффициент ранговой корреляции Кендалла**

Применяется для выявления взаимосвязи между количественными или качественными показателями, если их можно ранжировать (упорядочить)

Значения показателя X выставляют в порядке возрастания и присваивают им ранги.

Ранжируют значения показателя Y и рассчитывают коэффициент

$$\tau = \frac{2S}{n(n-1)},$$

$S=P-Q$

P - суммарное число наблюдений, следующих за текущими наблюдениями с **большим** значением рангов Y

Q - суммарное число наблюдений, следующих за текущими наблюдениями с **меньшим** значением рангов Y.

Корреляционный анализ

- **Коэффициент ранговой корреляции Спирмена**

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d_i – разности рангов X и Y

Корреляционный анализ

- **Множественный коэффициент корреляции**

Характеризует тесноту линейной корреляционной связи между одной случайной величиной и некоторым множеством случайных величин.

$$\rho^2_{\xi_1 \bullet \xi_2,\ldots,\xi_k} = 1 - \frac{|R|}{R_{11}}$$

$|R|$ - определитель корреляционной матрицы

R_{11} – алгебраическое дополнение элемента r_{11}

Элементы корреляционной матрицы r_{ij} – это линейные коэффициенты корреляции

Корреляционный анализ

Общая классификация корреляционных связей:

- Сильная или тесная при $|r_{xy}| > 0,70$;
- Средняя при $0,50 < |r_{xy}| < 0,69$;
- Умеренная при $0,30 < |r_{xy}| < 0,49$;
- Слабая при $0,20 < |r_{xy}| < 0,29$;
- Очень слабая при $|r_{xy}| < 0,19$.

Корреляционный анализ

Проверка значимости коэффициента корреляции с помощью распределения Стьюдента

Вычисляется коэффициент

$$t = |r_{xy}| \sqrt{\frac{n-2}{1-r_{xy}^2}}$$

Если $t<t_{кр}(\alpha, n-2)$, то коэффициент корреляции значим.

Регрессионный анализ

Корреляционный анализ непосредственно связан с регрессионным анализом.

- Регрессионный анализ – это набор статистических методов исследования влияния одной или нескольких независимых переменных X_1, X_2,…,X_n на зависимую переменную Y.

При регрессионном анализе:

- Определяют коэффициенты регрессионной модели и оценивают их значимость
- Определяют коэффициент корреляции и оценивают его значимость
- Оценивают адекватность регрессионной модели

Интеллектуальный анализ данных на транспорте

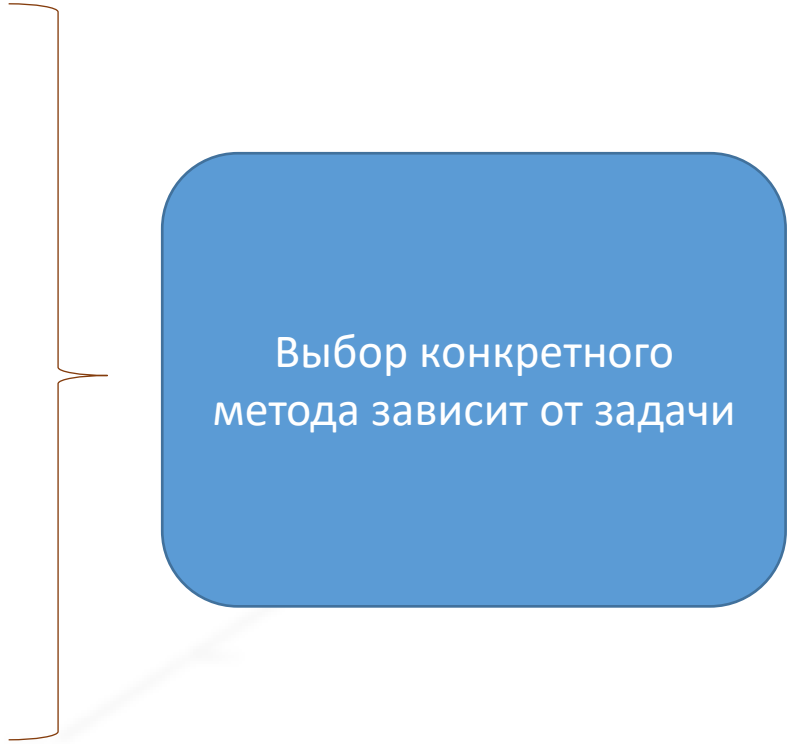
Лекция 7

2022

Статистические методы ИАД

К методам data mining относят *статистические методы* :

- дескриптивный (описательный) анализ,
- корреляционный и регрессионный анализ,
- факторный анализ,
- дисперсионный анализ,
- компонентный анализ,
- дискриминантный анализ,
- анализ временных рядов,
- анализ выживаемости,
- анализ связей



Факторный анализ

- **Факторный анализ** – это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводят к меньшему количеству независимых влияющих величин, называемых факторами:
  - в один фактор объединяются переменные, сильно коррелирующие между собой.
  - переменные из разных факторов слабо коррелируют между собой.

- **Факторный анализ классифицирует признаки** (переменные), описывающие наблюдения.
- **Фактор** –скрытая переменная, конструируемая таким образом,

чтобы можно было объяснить корреляцию между набором имеющихся переменных.

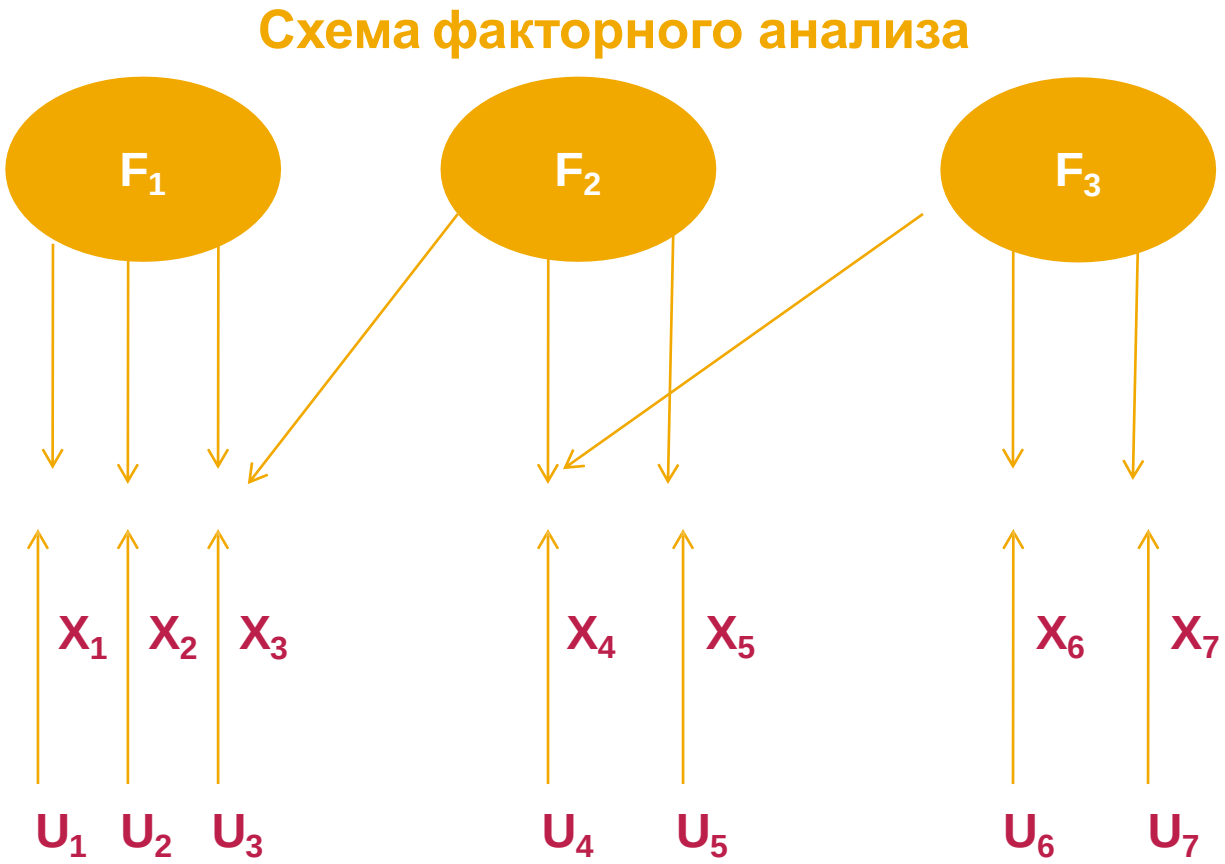
- **Концепция факторного анализа** заключается в «сжатии» информации.

Факторный анализ

- **Цель факторного анализа** – сокращение числа переменных на основе их классификации и определения структуры взаимосвязей между ними.

Благодаря **сокращению числа переменных** вместо исходного набора переменных появляется возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа взаимосвязанных переменных.

Факторный анализ



F – **общие факторы**, каждый из которых влияет на определенную совокупность переменных

X – **переменные**, фиксируемые на основании ответов

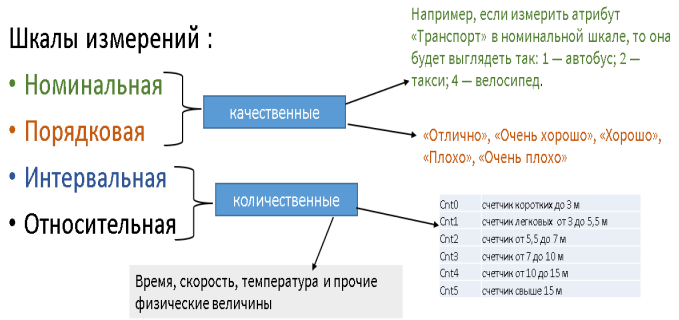
U – **уникальные факторы**, каждый из которых влияет только на одну переменную

Факторный анализ

Обязательные условия проведения факторного анализа

- Все признаки должны быть **количественными переменными**.
- Число наблюдений должно быть минимум **в два раза больше** числа переменных.
- Выборка должна быть **однородна**.
- Исходные переменные должны быть распределены **симметрично**.
- Номинальные переменные должны быть переведены в **дихотомические**

(переменные, имеющие только две категории).



Факторный анализ

Процедура факторного анализа состоит из четырех основных стадий:

1. Вычисление корреляционной матрицы для всех переменных, участвующих в анализе.
2. Извлечение факторов.
3. Выбор факторов и вращение факторов для создания упрощенной структуры.
4. Интерпретация факторов.

Дисперсионный анализ

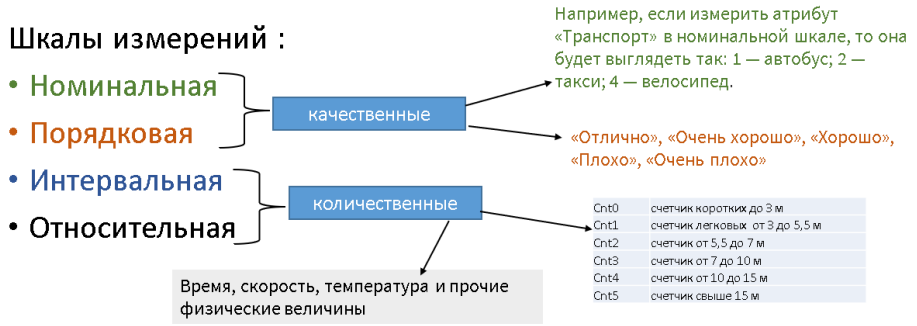
- **Дисперсионный анализ** — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях

В отличие от t-критерия, позволяет сравнивать средние значения трёх и более групп.

Разработан Р. Фишером для анализа результатов экспериментальных исследований.

В литературе также встречается обозначение ANOVA (*ANalysis Of VAriance*)

Дисперсионный анализ



Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных, обычно именуемых факторами, на зависимую переменную.

Зависимые переменные представлены значениями абсолютных шкал (шкала отношений).

Независимые переменные являются номинальными (шкала наименований), то есть отражают групповую принадлежность, и могут иметь два или более значения (типа, градации или уровня).

Дисперсионный анализ

В зависимости от типа и количества переменных различают:

- однофакторный и многофакторный дисперсионный анализ (одна или несколько независимых переменных);
- одномерный и многомерный дисперсионный анализ (одна или несколько зависимых переменных);
- дисперсионный анализ с повторными измерениями (для зависимых выборок);
- дисперсионный анализ с постоянными факторами, случайными факторами, и смешанные модели с факторами обоих типов;

Дисперсионный анализ

- Математическая модель дисперсионного анализа представляет собой частный случай основной линейной модели. Пусть с помощью методов $A_j \ (1 \leq j \leq m)$

производится измерение нескольких параметров

$$x_i \ (1 \leq i \leq n)$$

чьи точные значения — $\mu_i \ (1 \leq i \leq n)$

В таком случае результаты измерений различных величин различными методами можно представить как:

Дисперсионный анализ

$$x_{i,j} = \mu_i + a_{i,j} + e_{i,j},$$

где:

- $x_{i,j}$ — результат измерения i-го параметра по методу A_j;
- μ_i — точное значение i-го параметра;
- $a_{i,j}$ — систематическая ошибка измерения i-го параметра в группе по методу A_j;
- $e_{i,j}$ — случайная ошибка измерения i-го параметра по методу A_j.

Дисперсионный анализ

- Вычисляются дисперсии следующих случайных величин:

$$x_{i,j}$$

$$x_{i,j} - x_{i,*} - x_{*,j} + x_{*,*}$$

$$x_{i,*}$$

$$x_{*,j}$$

$$x_{*,j} = \frac{1}{n} \sum_i x_{i,j},$$

$$x_{i,*} = \frac{1}{m} \sum_j x_{i,j},$$

$$x_{*,*} = \frac{1}{nm} \sum_{i,j} x_{i,j}$$

Дисперсионный анализ

Дисперсии

$$s^2 = \frac{1}{nm}\sum_i\sum_j (x_{i,j} - x_{*,*})^2$$

$$s_0^2 = \frac{1}{nm}\sum_i\sum_j (x_{i,j} - x_{i,*} - x_{*,j} + x_{*,*})^2$$

$$s_1^2 = \frac{1}{n}\sum_i (x_{i,*} - x_{*,*})^2$$

$$s_2^2 = \frac{1}{m}\sum_j (x_{*,j} - x_{*,*})^2$$

$$s^2 = s_0^2 + s_1^2 + s_2^2$$

Дисперсионный анализ

- Процедура дисперсионного анализа состоит в определении соотношения систематической (межгрупповой) дисперсии к случайной (внутригрупповой) дисперсии в измеряемых данных.
- В качестве показателя изменчивости используется сумма квадратов отклонения значений параметра от среднего.
- Соотношение межгрупповой и внутригрупповой дисперсий имеет F-распределение (распределение Фишера) и определяется при помощи (F-критерия Фишера):

$$F_{df_{\mathrm{bg}},df_{\mathrm{wg}}} = \frac{MS_{\mathrm{bg}}}{MS_{\mathrm{wg}}}$$

Дисперсионный анализ

$$F_{df_{\text{bg}},df_{\text{wg}}} = \frac{MS_{\text{bg}}}{MS_{\text{wg}}}$$

$$MS_{\text{bg}} = \frac{SS_{\text{bg}}}{J-1} \qquad SS_{\text{bg}} = \sum_{i=1}^{n_j}(M_j - M)^2$$

$$MS_{\text{wg}} = \frac{SS_{\text{wg}}}{N-J} \qquad SS_{\text{wg}} = \sum_{i=1}^{n_j}(x_{i,j} - M_j)^2$$

M_j - среднее ***j*** – ой группы

M - среднее совокупности N - объём полной выборки J — количество групп.

$df_{\text{bg}} = J - 1$ $df_{\text{wg}} = N - J$

Дисперсионный анализ

Исходные положения дисперсионного анализа :

- нормальное распределение значений изучаемого признака в генеральной совокупности;
- равенство дисперсий в сравниваемых генеральных совокупностях; случайный и независимый характер выборки.

Дисперсионный анализ

Нулевой гипотезой в дисперсионном анализе является утверждение о равенстве средних значений.

При отклонении нулевой гипотезы принимается альтернативная гипотеза о том, что не все средние равны, то есть имеются, по крайней мере, две группы, отличающиеся средними значениями

При наличии трёх и более групп для определения различий между средними применяются t-тесты или метод контрастов