# **Data mining**

Полноценным процессом сбора, отбора, очистки, модификации и извлечения информации и называется Data mining. Обычно этот метод используется для больших массивов данных, а результатом являются знания.

Интеллектуальный анализ данных считается мультидисциплинарной сферой, которая берёт своё начало во многих науках, а именно, прикладная статистика, теория БД, искусственный интеллект, машинное обучение и другие.

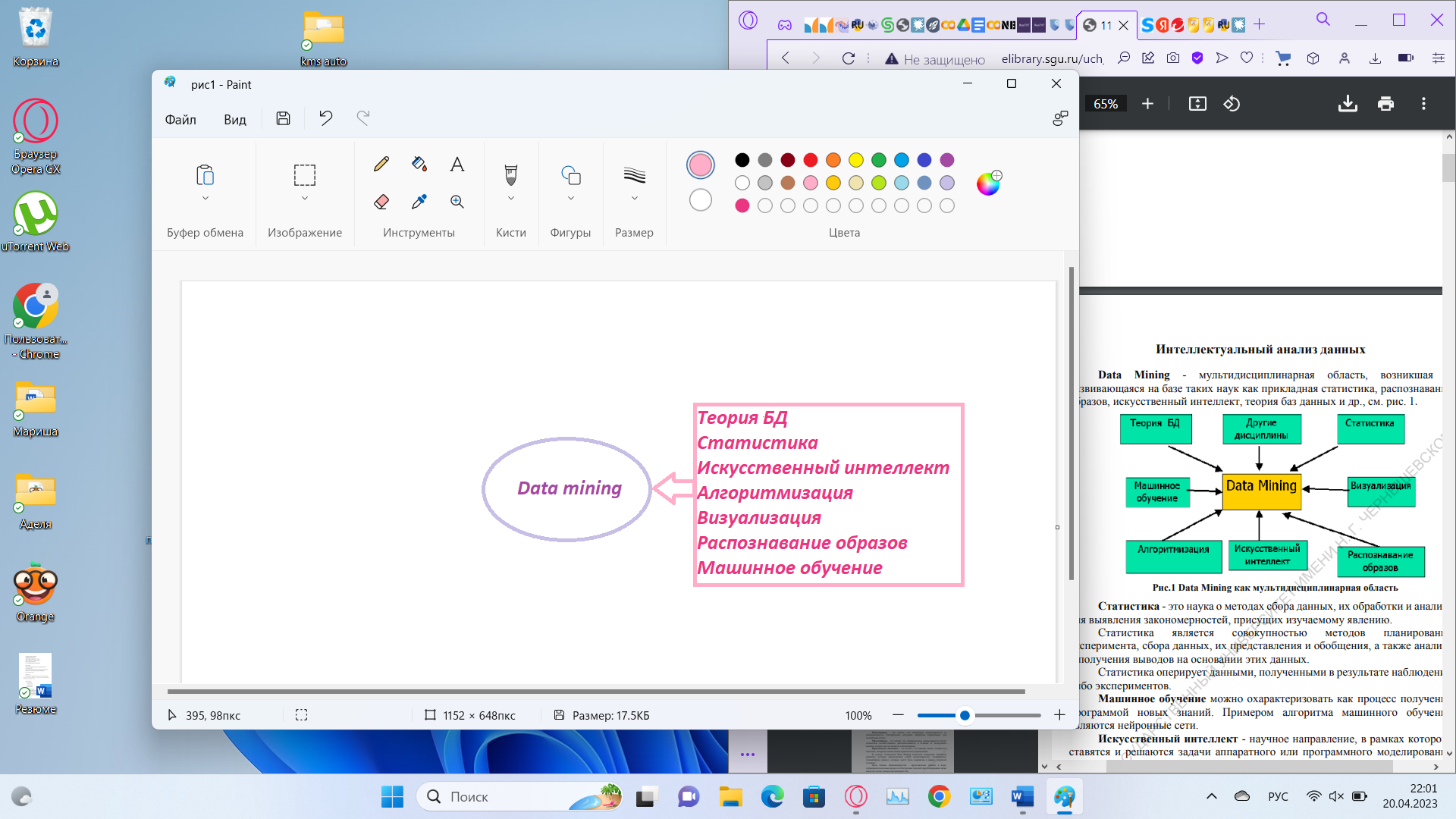


Рисунок 1 - Data mining - мультидисциплина

Процесс интеллектуального анализа включает в себя несколько этапов, которые характеризуются элементами сравнения, типизации, классификации, обобщения и повторения.

Обычно этот процесс можно разделить на следующие этапы:

* исследование области знания;
* определение цели;
* подготовка данных;
* формирование модели;
* анализ и оценка модели;
* выбор модели;
* использование модели;
* корректировка и модернизация модели.

Подготовка данных – необходимый этап Data mining, в котором оценивается качество данных, производится очистка данных или их преобразование.

В этом случае, данные можно разделить на данные высокого качества и низкого (грязные данные): полные, достоверные и актуальные данные называют данными высокого качества; пропущенные, недостоверные или бессмысленные в практическом плане данные называют грязными.

Преимущественные виды некачественных данных:

* пропущенные значения;
* повторы данных;
* шумы и выбросы.

Пропущенные значения можно обработать следующими способами:

* удалить из анализа объекты с пропущенными значениями;
* вычислить приблизительное значение на месте пропущенного;
* пренебрегать пропущенными значениями во время анализа;
* изменить пропущенные значения на вероятные.

Повторы данных или дубликаты – записи с одинаковыми значениями каждого атрибута, которые либо полностью исключаются, либо остаётся одна уникальная запись.

Выбросы – определённо отличные значения в наборе данных. Довольно часто сравнивают результаты, полученные с использованием данных выбросов и без, что может точно показать, являются ли данные выбросы информативными.

Рассчитанные результаты на основе грязных данных не могут быть достоверными, поэтому и применятся очистка данных, целью которой в целом является улучшение качества данных.

Рассмотрим методы интеллектуального анализа. Значительная часть аналитических методов Data mining – математические алгоритмы и методы, а также Data mining сочетает в себе последние достижения сферы информационных технологий.

В зависимости от того, сохраняются данные после анализа или дистиллируются для последующих действий методы интеллектуального анализа делятся на два типа:

В первом случае исходные данные в явном виде используются при применении прогностического моделирования, а также анализа исключений, в которых используются результаты свободного поиска. К методам этого разряда относят кластерный анализ, метод ближайшего и k-ближайшего соседа и рассуждение по аналогии.

Во-втором же случае (поиск и использование формализованных закономерностей) образцы информации формируются на основе исходных данных и модифицируются в формальные конструкции, которые в свою очередь зависят от метода Data mining. В первом разряде отсутствуют вышеперечисленные процессы, которые выполняются во время свободного поиска. К методам второго разряда относят логические методы, методы визуализации и кросс-табуляции, а также методы, основанные на уравнениях.

Рассмотрим отдельно каждую группу методов:

Логические методы (методы логической индукции): нечёткие запросы, анализы, символьные правила, генетические алгоритмы и деревья решений. Эта группа методов позволяет сформировать полученные закономерности наиболее понятном виде для человека. На выходе методов-правил могут быть как дискретные, так и непрерывные переменные. Следует отметить, что деревья решений и правила различны только по своим возможностям, хотя реализация правил затрагивает более медленные алгоритмы, нежели индукция деревьев решений. К тому же деревья решений при необходимости рассматриваются в качестве наборов символьных правил, которые можно получить, генерируя правило от корня дерева до его терминальной вершины.

Методы кросс-табуляции: агенты, доверительные (баесовские) сети, кросс-табличная визуализация.

Эти методы также реализуют главную задачу Data mining – находит шаблоны, хотя метод кросс-табличной визуализации не обеспечивает самостоятельный поиск закономерностей.

Методы на основе уравнений: статистические методы, нейронные сети. Эта группа методов представляет полученные закономерности в виде уравнений, что ограничивает их работу численными переменными. Основная задача, которая нуждается при решении в методах этой группы, - задача прогнозирования. Методы статистического анализа: корреляция рядов динамики, гармонический анализ, корреляционно-регрессионный анализ и другие.

Классификация задач Data mining:

Следует отметить весь спектр задач, которые входят в основу Data mining: классификация, кластеризация, ассоциации и последовательность, прогнозирование, определение отклонений, оценивание, анализ связей, визуализация и подведение итогов.

Задачи интеллектуального анализа делят на группы, представленные ниже:

* обучение с учителем: классификация, оценка, прогнозирование;
* обучение без учителя: кластеризация;
* другие: ассоциации и последовательность, определение отклонений, анализ связей, визуализация и подведение итогов.

Также задачи интеллектуального анализа могут быть разделены на группы описательных и прогнозирующих задач.

В качестве результата описательной задачи получают шаблон, который описывает данные, подлежащие модификации. Описательные задачи характеризуют данные и сравнивают их. Прогнозирующие задачи предполагают создание моделей и предсказывание новых свойств.