**Федеральное агентство связи**

**Ордена Трудового Красного Знамени**

**федеральное государственное бюджетное**

**образовательное учреждение высшего образования**

**«Московский технический университет связи и информатики»**

**Кафедра “Информатика”**

**Курсовая работа по дисциплине «Основы программирования»**

**На тему “технология Data Mining”**

Выполнил: студент гр. БСТ2002

Демидов К. С.

Проверил: доц. Гуриков С. Р.

**Москва 2020 г.**

# Содержание

[Содержание](#_uwr8ppveo3f6) 1

[Техническое задание](#_do0f9h2sig5h) 4

[Глава 1. Теоретическая часть](#_t11odtng7h2p) 7

[1.1 Понятие Data Mining и классификация технологических методов](#_79inc512z5fg) 7

[1.2 Задачи Data Mining](#_dyi4psqxgv85) 10

[1.3 Процессы Data Mining](#_cqvqmpvk170z) 22

[1.4 Рынок инструментов Data Mining](#_m84o9r7ktb2x) 24

[Заключение](#_brmyanwrlu7c) 28

[Список используемых источников](#_kavm96ima3tn) 29

Оглавление научитесь собирать

**Введение**

**ЗДЕСЬ И ДАЛЕЕ НАДО ЧИТАТЬ ГОСТ В ЧАСТИ ОФОРМЛЕНИЯ ЗАГОЛОВКОВ**

**Актуальность темы**

В наше время существует множество технологий в IT-сфере, но самая важная - **Data Mining**. Она является важнейшей технологией, так как её используют во многих сферах деятельности человека, например, в банковском деле, страховании, телекоммуникации и т.д.Главной целью этой технологии является поиск в больших объемах данных неочевидных, объективных и полезных на практике закономерностей. Выравнивание??

**Таким образом,** одно из важнейших назначений **Data Mining** состоит в наглядном представлении результатов вычислений, что позволяет использовать инструментарий технологии людьми без специальной математической подготовки.

**Подытожим, технология Data Mining** актуальна в наше время, так как, **во-первых**, она используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные, **во-вторых**, банки разрабатывают и внедряют модели интеллектуального анализа данных для прогнозирования способности заемщика брать и погашать кредиты и **в-третьих**, благодаря **Data Mining** можно обнаруживать и предотвращать мошеннические транзакции.

Я НЕ ВИЖУ, ЧТО ИСПОЛЬЗОВАЛСЯ ОБРАЗЕЦ ВВЕДЕНИЯ ГОВОРИЛИ НА ЧТО ОБРАЩАТЬ ВНИМАНИЕ

**Объект исследования**

Объектом исследования, проведенного в рамках данной работы, является технология Data Mining.

**Предмет исследования**

Предметом исследования является изучение методов Data Mining.

**Цели работы и задачи исследования**

Цель работы заключается в том, чтобы раскрыть основные свойства технологии Data Mining. Поставленная цель определила следующие основные задачи исследования:

1. Исследовать основные теоретические положения.
2. Разработать программный проект.
3. Изучить методы, задачи, а также процесс Data Mining.
4. Исследовать инструменты применения технологии Data Mining.

СКОЛЬКО ЗАДАЧ ДОЛЖНО БЫТЬ НА СЕГОДНЯ???

**ЧТО ВООБЩЕ ПРО НИХ ГОВОРИЛИ**

**Методы исследования**

Для решения поставленных задач были использованы теоретические методы исследования. Теоретическую основу исследования составили труды в области технологии Data Mining Чубуковой И. А.

# Техническое задание

# 1.1. Основание для разработки

Задание выполнено в соответствии с заданием полученным с кафедры “Информатика” Федерального агентства связи ордена Трудового Красного Знамени федерального государственного бюджетного образовательного учреждения высшего образования «Московский технический университет связи и информатики» и утвержденное научным руководителем доцентом кафедры “Информатика” К. П. Н. Гуриковым С. Р. 2 октября 2020 года. ДО ТЕХ ПОР ПОКА БУДЕТ БЕЗДУМНОЕ КОПИРОВАНИЕ У ДРУГИХ, ФАЙЛ БУДЕТ ВОЗВРАЩАТЬСЯ

# 1.2 Назначения разработки

Разработанный программный продукт предназначен для ознакомления с основными теоретическими положениями по теме курсовой работы и проверке знаний пользователя.

# 1.3 Требования к программе

# 1.3.1 Требования к функциональным характеристикам

Разработанный программный продукт должен обеспечивать выполнение следующих функций:

1. Возможность ознакомления с теоретическим материалом по теме курсовой работы
2. Возможность вывода результатов исследования для пользователя

# 1.3.2 Требования к надежности

Разработанное программное обеспечение должно иметь:

1. Устойчивую работу в соответствии с алгоритмом программы.
2. Выдавать сообщения об ошибках
3. Поддерживать диалоговый режим в рамках предоставляемых пользователю возможностей

**1.3.3 Требования к составу и параметрам технических средств**

1. Процессор с тактовой частотой не ниже 1,8 ГГц. Рекомендуется использовать как минимум двухъядерный процессор
2. 2 ГБ ОЗУ; рекомендуется 8 ГБ ОЗУ (минимум 2,5 ГБ при выполнении на виртуальной машине)
3. Место на жестком диске: до 210 ГБ (минимум 800 МБ) свободного места в зависимости от установленных компонентов; обычно для установки требуется от 20 до 50 ГБ свободного места
4. Скорость жесткого диска: для повышения производительности установите Windows и Visual Studio на твердотельный накопитель (SSD)
5. Видеоадаптер с минимальным разрешением 720p (1280 на 720 пикселей); для оптимальной работы Visual Studio рекомендуется разрешение WXGA (1366 на 768 пикселей) или более высокое

**1.3.4 Требования к информационной программной совместимости**

Программа должна легко устанавливаться, функционировать и корректно работать при наличии следующего программного обеспечения:

1. Операционная система Windows 10
2. Microsoft Visual Studio 2019

**1.3.5 Требования к транспортированию и хранению**

Программа предоставляется на usb-флеш-накопителе. Программная документация поставляется в электронном виде.

**1.3.6 Требования к программной документации**

В ходе разработки программы должны быть подготовлены:

1. Текст программы
2. Описание программы
3. Методика испытаний
4. Руководство пользователя

**1.3.7 Стадия и этапы разработки**

****

|  |  |  |
| --- | --- | --- |
| **Название** | **Срок сдачи** | **Отчётность** |
| Утверждение темы | 02.10.2020 |  |
| Написание введения | 8.10.2020 |  |
| Написание 1 главы и тех.задания | 29.10.2020 |  |

# Глава 1. Теоретическая часть

# 1.1 Понятие Data Mining и классификация технологических методов

Термин **Data Mining** получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Данную технологию достаточно точно определяет Григорий Пиатецкий-Шапиро - один из основателей этого направления: “ Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Суть и цель технологии можно охарактеризовать следующим образом: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей ”.

В технологии Data Mining объединены строго формализованные методы и методы формального анализа, т.е. количественный и качественный анализ данных.

**Метод** (англ. Method) - это норма или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Абсолютно все методы технологии разделяют по принципу работы с исходными обучающими данными.

1. **Непосредственное использование данных, или сохранение данных**

В этом случае исходные данные хранятся в явном детализированном виде и используются непосредственно на стадиях прогностического моделирования и/или анализа исключений. Главный недостаток этой группы методов заключается в том, что при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

**Примеры методов данной группы:** кластерный анализ, метод ближайшего соседа, рассуждение по аналогии.

1. **Выявление и использование формализованных закономерностей, или дистилляция шаблонов.**

При использовании данного метода один образец информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого метода Data Mining. Данный процесс выполняется на стадии свободного поиска, главным отличием данного метода от первого является то, что у первой группы методов данная стадия отсутствует. На стадиях прогностического моделирования и анализа исключений используются результаты стадии свободного поиска, они значительно компактнее самих баз данных.

**Примеры методов данной группы:** логические методы, методы визуализации, методы кросс-табуляции, методы, основанные на уравнениях.

Рассмотрим свойства методов технологии Data Mining. Различные методы данной технологии характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств. Среди основных свойств и характеристик методов Data Mining рассмотрим следующие: точность, масштабируемость, интерпретируемость, проверяемость, трудоемкость, гибкость, быстрота и популярность.

Что же такое масштабируемость?

**Масштабируемость** - свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности при добавлении к ней вычислительных ресурсов.

В таблице 2 приведена сравнительная характеристика самых распространенных методов.



|  |  |  |  |
| --- | --- | --- | --- |
| Алгоритм | Точность | Масштабируемость | Интерпретируемость |
| Классические методы | Нейтральная | Высокая | Высокая/  Нейтральная |
| Нейронные сети | Высокая | Низкая | Низкая |
| Методы визуализации | Высокая | Очень низкая | Высокая |
| Полиномиальные нейронные сети | Высокая | Нейтральная | Низкая |
| K-ближайшего соседа | Низкая | Очень низкая | Высокая/  нейтральная |

# 

# 1.2 Задачи Data Mining

В основу технологии положена концепция шаблонов, представляющих собой закономерности. В результате их обнаружения решаются задачи данной технологии. Различным типам закономерностей, которые могут быть выражены в форме, понятной человеку, соответствуют определенные задачи Data Mining. Задачи технологии иногда называют закономерностями или техниками. Единого мнения какие задачи следует относить к Data Mining нет, поэтому большинство авторитетных источников перечисляют следующие: классификация, кластеризация, прогнозирование, ассоциация, визуализация, анализ и обнаружение отклонений, оценивание, анализ связей, подведение итогов.

Рассмотрим подробнее каждую задачу.

1. **Классификация ( Classification )**

Наиболее простая и распространенная задача Data Mining. В результате решения данной задачи обнаруживаются признаки, характеризующие группы объектов исследуемого набора данных. По этим признакам новый объект можно отнести к тому или иному классу.

Для решения задачи **классификации** могут использоваться методы: ближайшего соседа, k-ближайшего соседа, байесовские сети, индукция деревьев решений, нейронные сети.

1. **Кластеризация ( Clustering )**

Кластеризация является наследником идеи классификации, но она является более сложной задачей. Её особенность заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

Пример метода решения задачи **кластеризации**: выделение групп людей на основе графа связей в социальных сетях

1. **Ассоциация ( Associations )**

В ходе решения задачи поиска ассоциативных правил находятся закономерности между связанными событиями в наборе данных. Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный алгоритм решения задачи поиска **ассоциативных правил** - алгоритм Apriori. **Алгоритм Apriori** использует стратегию [поиска в ширину](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%B8%D1%81%D0%BA_%D0%B2_%D1%88%D0%B8%D1%80%D0%B8%D0%BD%D1%83) для подсчёта объектов и использует функцию генерации кандидата, которая основана на свойстве нисходящего замыкания поддержки.

1. **Последовательность ( Sequence )**

Последовательность позволяет найти временные закономерности между транзакциями. Её целью является установление закономерностей между событиями, связанными во времени. Иными словами, последовательность определяется высокой вероятностью цепочка связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным шагом, равным нулю.

Существует **правило последовательности**: после события X через определенное время произойдет событие Y.

Пример последовательности. После покупки квартиры жильцы в 70% случаев в течение двух недель приобретают стиральную машину, а в течение двух месяцев в 30% случаев приобретается пылесос. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом.

1. **Прогнозирование ( Forecasting )**

В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

1. **Оценивание ( Estimation )**

Задача оценивания сводится к предсказанию непрерывных значений признака.

1. **Визуализация ( Visualization )**

В результате визуализации создается графический образ данных, которые анализируют. Для решения данной задачи используются графические методы, которые показывают наличие закономерностей в данных.

Пример методов визуализации: представление данных в 2D и 3D измерениях, подведение итогов.

Согласно классификации по стратегиям, задачи т**ехнологии Data Mining** подразделяются на следующие группы:

1. **Обучение с учителем**
2. **Обучение без учителя**
3. **Другие**

Категория **обучение с учителем** представлена следующими задачами: классификация, оценка, прогнозирование.

Категория **обучение без учителя** представлена задачей кластеризации. В категорию другие входят задачи, не включенные в предыдущие две стратегии.

Задачи Data Mining, в зависимости от используемых моделей, могут быть дескриптивными и прогнозирующими. В соответствии с этой классификацией, задачи технологии представлены группами описательных и прогнозирующих задач. Эти задачи описывают общую концепцию анализируемых данных, определяют информативные, итоговые, отличительные особенности данных. Концепция описательных задач подразумевает характеристики сравнение наборов данных.

Для начала рассмотрим задачу и цели метода классификации, введем понятие этой задачи.

**Классификация** - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация требует соблюдения данных правил:

1. В каждом акте деления необходимо применять только одно основание
2. Деление должно быть соразмерным, то есть общий объем видовых понятий должен равняться объему делимого родового понятия
3. Члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться

Различают несколько видов классификаций:

1. **Вспомогательную классификацию**, которая производится по внешнему признаку и служит для придания множеству предметов нужного порядка
2. **Естественную классификацию**, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий классификация может быть: **простой** и **сложной**.

Задачей классификации обычно называют предсказание категориальной зависимой переменной на основе выборки непрерывных и/или категориальных переменных.

Цель процесса классификации состоит в том, чтобы построить модель, которая использует прогнозирующие атрибуты в качестве входных параметров и получает значение зависимого атрибута. Процесс классификации заключается в разбиении множества объектов на классы по определенному критерию.

Рассмотрим данную задачу на примере. Например у нас есть база данных швейной фабрики с информацией о возрасте, доходе за год, также мы имеем рекламный материал двух видов: одежды среднего класса и премиальной. **Следовательно**, определены два класса клиентов: класс 1 и класс 2. Данная база данных приведена в таблице 3.



|  |  |  |  |
| --- | --- | --- | --- |
| Код клиента | Возраст | Доход | Класс |
| 1 | 16 | 5 | 1 |
| 2 | 18 | 150 | 2 |
| 3 | 38 | 15 | 1 |
| 4 | 31 | 900 | 2 |
| 5 | 27 | 300 | 2 |

Рассмотрим цели и задачи метода кластеризации.

**Задача кластеризации** сходна с задачей классификации, является её логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены. Кластеризация предназначена для разбиения совокупности объектов на однородные группы. Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

**Целью кластеризации** является поиск существующих структур.

**Подытожим,** несмотря на кажущуюся похожесть этих задач, решаются они разными способами и при помощи разных методов, прежде всего различие данных задач в исходных данных.

Для начала рассмотрим задачи прогнозирования. Данный вид задач решается в самых разнообразных областях человеческой деятельности, например, в науке, экономике, в производстве и др. Прогнозирование является достаточно важным элементом организации управления.

**Задача прогнозирования** является самой сложной из всех задач технологии Data Mining, так как требует тщательного исследования исходного набора данных и методов, которые подходят для анализа.

Основой для прогнозирования служит историческая информация, которая хранится в базе данных в виде временных рядов.

Точность прогноза оказывает колоссальное влияние на прогнозирующую систему. Точность прогноза напрямую зависит от того сколько у системы ресурсов.

Существует классификация наиболее распространенных ошибок:

1. Средняя ошибка ( CO )
2. Средняя абсолютная ошибка ( САО )
3. Сумма квадратов ошибок ( SSE )
4. Относительная ошибка ( ОО )

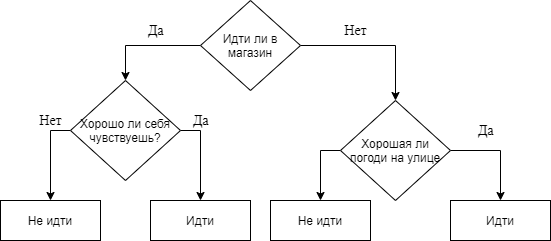
Рассмотрим **задачи визуализации**.

В результате применения данной задачи создается графический образ данных. Применение визуализации помогает в процессе анализа данных увидеть аномалии, структуры и тренды.

Задача визуализации имеет огромное преимущество, оно заключается в том, что у неё практически полностью отсутствует необходимость в специальной подготовке пользователя.

Наиболее популярным методом решения задач классификации и прогнозирования является метод под названием “Деревья решений”. В самом простом виде дерево решений - это способ представления правил в иерархической, последовательной структуре. Основа такой структуры - ответы "Да" или "Нет" на ряд вопросов.

На рисунке 1 представлен пример дерева решений “Идти в магазин или нет”.





В этом примере представляется работа бинарных деревьев. Данный вид деревьев является самым простым и частным случаем деревьев решения.

Деревья решений позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания, а также они дают возможность извлекать правила из базы данных на естественном языке.

Данный метод относится к группе граничных методов, группа определяет классы при помощи границ областей. При данном методе решаются задачи бинарной классификации. Также в основе метода опорных векторов лежит понятие плоскостей решений.

Главной целью этого метода является найти плоскость, которая разделяет два множества объектов. Также метод отыскивает **опорные векторы.**

**Опорные векторы -** это объекты множества, которые лежат на границе областей.

Рассмотрим следующий метод, а именно метод “Ближайшего соседа” относится к классу методов, чья работа основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная идентифицируется.

Данный метод относится к категории “Обучение без учителя”, так как является самообучающейся технологией, благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением примеров улучшаются.

Выделим преимущества и недостатки данного метода.

Преимущества:

1. Простота использования полученных результатов
2. Решения не уникальны для конкретной ситуации, возможно их использование для других случаев
3. Целью поиска является не гарантированно верное решение, а лучшее из возможных.

Недостатки:

1. Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт
2. Существует сложность выбора метрики
3. Вычислительная трудоемкость
4. Данный метод может решать только задачи небольшой размерности по количеству классов и переменных

В свою очередь с помощью метода “Ближайшего соседа” решаются задачи классификации и регрессии.

Нейронные сети также часто используются для решения каких-либо задач. Разберём что это такое.

**Нейронная сеть** - математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей.

Область применения **нейронной сети**  достаточно обширная, например, в распознавании лиц, для прогнозирования чего-либо и т.д. В **Data Mining** нейронные сети используются в данных задачах:

1. Классификации

Например: распознать текст, речь.

1. Прогнозировании

Например: решить задачу пропущенных значений

1. Кластеризации

Например: сжатие информации путем уменьшения размерности данных. Именно задачи кластеризации решаются с помощью **самоорганизующихся карт Кохонена.**

Сети, которые называются картами Кохонена - одна из разновидностей нейронных сетей, их главное отличие заключается в том, что они используют неконтролируемое обучение.

Главной идеей сетей Кохонена - аналогия со свойствами человеческого мозга. Данная сеть очень проста, в отличие от многослойной нейронной сети. Она представляет собой всего лишь два слоя, а именно **входной** и **выходной.** Элементы самоорганизующейся карты располагаются в некотором двумерном пространстве. Сеть Кохонена обучается благодаря методу приближений, при этом заключается не в минимизации ошибки, а в подстройке внутренних параметров нейронной сети для наибольшего совпадения с входными данными. Данные карты могут применяться в моделировании, прогнозировании и сжатии информации.

Впервые такой термин, как “Кластерный анализ” ввёл математик Роберт Трион в 1939 году и данный метод включает в себя более 100 различных алгоритмов.

Главными отличиями кластерного анализа от задач классификации является то, что он не требует априорных предположений о наборе данных и не накладывает ограничение на представление исследуемых объектов, а также позволяет анализировать показатели различных типов данных.

Существует четыре группы задач кластерного анализа и обычно при практическом использовании кластерного анализа одновременно решается несколько задач.

Перечислим все группы задач кластерного анализа:

1. Разработка типологии и классификации
2. Исследование полезных концептуальных схем группирования объектов
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения,действительно ли типы , выделенные тем или иным способом, присутствуют в имеющихся данных.

Также методы кластерного анализа можно разделить на две основные группы, а именно на **иерархические** и **неиерархические.**

Данный метод имеет свои недостатки, а именно:

1. Результаты кластеризации зависят от критериев разбиения совокупности исходных данных
2. При понижении размерности данных могут возникнуть определенные искажения
3. За счет обобщений могут потеряться некоторые индивидуальные характеристики объектов

Основной задачей технологии **Data Mining** является ассоциация, а главной целью поиска **ассоциативных правил** является нахождение закономерностей между связанными событиями в базах данных.

Ассоциативные правила часто встречаются в розничной торговле, оформлении каталогов, а также в маркетинге.

Основными характеристиками ассоциативного правила являются поддержка и достоверность правила.

Самый известный алгоритм поиска ассоциативных правил - Apriori. Его работа состоит из следующих этапов:

1. Формирование кандидатов
2. Подсчёт кандидатов

Данный алгоритм используется для частого анализа наборов элементов и изучения правил ассоциации в реляционных базах данных.

Разновидности алгоритма Apriori, являющиеся его оптимизацией, предложены для сокращения количества сканирований базы данных, а также количества наборов-кандидатов или того и другого. **Таким образом**, были предложены следующие разновидности алгоритма Apriori: AprioriTID и AprioriHybrid.

Существуют также способ визуализации. К данному способу относят: графики, диаграммы, таблицы списки и т.д.

Традиционные методы визуализации принято применять в следующих задачах:

1. Предоставлять пользователю информацию в наглядном виде
2. Компактно описывать закономерности, присущие исходному набору данных
3. Снижать размерность или сжимать информацию
4. Восстанавливать пробелы в наборе данных
5. Находить шумы и выбросы в наборе данных

Существует достаточно большое количество различных способов представления моделей, но самым “ценным” является **графическое представление**, так как в большинстве случаев пользователь не является специалистом в моделировании, чаще всего он эксперт в своей предметной области.

Следующий способ представления модели - представление её в интуитивном виде**.** При использовании данного способа пользователь может понимать, что происходит внутри модели. **Таким образом**, он принимает непосредственное участие в процессе.

В результате использования средств визуализации должны быть получены наглядные и выразительные, ясные и простые изображения, за счет использования разнообразных средств: цвета, контраста, границ, пропорций, масштаба и т.д.

# 1.3 Процессы Data Mining

Процесс Data Mining является своего рода исследованием. Этот процесс состоит из определенных этапов, включающих элементы сравнения, типизации, классификации, обобщения, абстрагирования и повторения. Существует три этапа процесса Data Mining, рассмотрим их подробнее.

**Этап 1. Анализ предметной области**

Данный этап самый важный, так как от него зависит успех будущей разработки приложений Data Mining. На данном этапе исследователю необходимо пронаблюдать за свойствами объектов с целью выявления и оценки важных закономерных отношений между показателями данных свойств.

**Этап 2. Постановка задачи**

Постановка задачи в Data Mining включает следующие шаги:

1. Формулировка задачи
2. Формализация задачи

Постановка задачи включает в себя описание статического, а также динамического поведения исследуемых объектов.

**Этап 3. Подготовка данных**

Цель данного этапа заключается в том, чтобы разработать базу данных для Data Mining. Данный этап состоит из трёх этапов.

1. Определение и анализ требований к данным
2. Сбор данных
3. Предварительная обработка данных

Изучим процесс очистки данных .На данном этапе выполняется выявление и удаление ошибок и несоответствий в данных с целью улучшения качества. Так как вся некорректная, дублирующаяся или утраченная информация может привести к результату неправильной статистики, **следовательно**, очистка данных является обязательной процедурой.

Существует две классификации инструментов очистки и редактирования данных

1. Средства анализа и модернизации данных

Средства анализа и модернизации данных служат для того, чтобы выявлять ошибки, а также для определения несоответствий и необходимых очищающих преобразований

1. Специальные средства очистки

Данные средства очистки обычно имеют дело с конкретными областями, например, с именами и адресами. Преобразования в основном либо обеспечиваются заранее, либо пользователем в интерактивном режиме.

Следующий процесс - процесс построения и использования модели. Данный процесс представляет собой построение модели и изучение ее свойств, которые подобны наиболее важным свойствам исследуемых объектов. Таким образом, при помощи моделирования изучаются свойства объектов путем исследования соответствующих свойств построенных моделей.

Этот процесс широко применяется при использовании методов технологии Data Mining, так как при использовании моделей Data Mining обнаруживается полезная и неизвестная информация, которую используют для принятия решений.

Построение моделей Data Mining осуществляется для того, чтобы исследовать или изучить моделируемый объект, процесс или решение для получения новых знаний и в будущем для принятия верного решения. Использование моделей помогает определить самое верное и выгодное решение в конкретной ситуации

# 1.4 Рынок инструментов Data Mining

Инструменты Data Mining - это наборы программных средств, с помощью которых выполняется подготовка данных и обеспечиваются алгоритмы их интеллектуального анализа, а также осуществляются процессы машинного обучения.

Инструменты Data Mining чаще всего используются маркетологами, маркетинговыми аналитиками, специалистами по анализу данных, а также специалистами по предотвращению мошеннических действий. В таблице 4 представлен список 10 самых популярных инструментов Data Mining



|  |  |
| --- | --- |
| Уровень популярности | Название инструмента |
| 1 | SAS Data Mining |
| 2 | Teradata |
| 3 | R-Programming |
| 4 | BOARD |
| 5 | Dundas |
| 6 | Inetsoft |
| 7 | H3O |
| 8 | Qlik |
| 9 | RapidMiner |
| 10 | Oracle Bl |

В наше с вами время на рынке программного обеспечения Data Mining существует огромное множество разнообразных продуктов, которые относятся к инструментам данной технологии, но наиболее популярная группа инструментов содержит следующие категории:

1. Наборы инструментов
2. Классификация данных
3. Кластеризация и сегментация
4. Инструменты статистического анализа
5. Анализ текстов
6. Извлечение отклонений
7. Инструменты визуализации

Рынок программного обеспечения Data Mining представлен множеством инструментов и на нем идет постоянная конкурентная борьба за потребителя, но в то же время отмечается отставание существующего программного обеспечения от теоретических разработок в связи со сложностью программной реализации некоторых новых теоретических разработок методов, а также алгоритмов Data Mining.

Изучим самые распространённые инструменты Data Mining, а именно SAS Enterprise Miner, PolyAnalyst, а также Oracle Data Mining и Deductor.

SAS Enterprise Miner - интегрированный компонент системы SAS, созданный специально для выявления в огромных массивах данных информации, которая необходима для принятия решений. Разработанный для поиска и анализа глубоко скрытых закономерностей в данных SAS, Enterprise Miner включает в себя методы статистического анализа, соответствующую методологию выполнения проектов Data Mining и графический интерфейс пользователя. Важной Особенностью SAS Enterprise Miner является его полная интеграция с программным продуктом SAS Warehouse Administrator, предназначенным для разработки и эксплуатации информационных хранилищ, и другими компонентами системы SAS. Разработка проектов Data Mining может выполняться как локально, так и в архитектуре клиент-сервер.

Система PolyAnalyst предназначена для автоматического и полуавтоматического анализа числовых баз данных и извлечения из сырых данных практически полезных знаний. PolyAnalyst находит многофакторные зависимости между переменными в базе данных, автоматически строит и тестирует многомерные нелинейные модели, выражающие найденные зависимости, выводит классификационные правила по обучающим примерам, находит в данных многомерные кластеры, строит алгоритмы решений.

Oracle Data Mining является опцией или модулем в Oracle Enterprise Edition. Опция Oracle Data Mining предназначена для анализа данных методами, относящимися к технологии извлечения знаний, или Data Mining.

Oracle Data Mining поддерживает все этапы технологии извлечения знаний, включая постановку задачи, подготовку данных, автоматическое построение моделей, анализ и тестирование результатов, использование моделей в реальных приложениях

Deductor Studio - это программа которая предназначена для анализа информации из различных источников данных. Она реализует функции импорта, обработки, визуализации и экспорта данных. Deductor Studio может функционировать без хранилища данных, получая информацию из любых других источников, но наиболее оптимальным является их совместное использование.

Выбор инструментального средства Data Mining и способа его внедрения должен проводиться в соответствии с конкретными целями и задачами, также должны учитываться уровни финансовых возможностей компании, квалификация пользователей и целый ряд других факторов, потому что внедрение Data Mining почти всегда требует серьезных финансовых затрат. Также следует не только учитывать задачи, которые стоят перед компанией сегодня, но и рассчитывать на возможность возникновения новых задач в будущем.

# 

# Заключение

В первой главе были рассмотрены основные теоретические положения, а также изучены методы, задачи процессы и самые популярные инструменты технологии Data Mining.

В итоге можно сделать вывод, что технология Data Mining необходима, потому что благодаря ей решаются достаточно сложные задачи в разных сферах деятельности человека.

# Список используемых источников

1. Чубукова, И. А. Data Mining : учебное пособие / И. А. Чубукова. — 3-е изд. — Москва, Саратов : Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020. — 469 c. — ISBN 978-5-4497-0289-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/89404.html (дата обращения: 29.10.2020)