МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

 **«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Экономический факультет**

**Кафедра экономики и управления инновационными системами**

**КУРСОВАЯ РАБОТА**

**ОСОБЕННОСТИ И ВОЗМОЖНОСТИ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ**

Работу выполнила Л.А. Фельд

(подпись, дата)

Направление 27.03.03 Системный анализ и управление

Направленность (профиль) Системный анализ и управление экономическими процессами

Научный руководитель:

канд. техн. наук, доц.Н.Ю. Нарыжная

(подпись, дата)

Нормоконтролер:

Н.Ю. Нарыжная

(подпись, дата)

Краснодар

2022

СОДЕРЖАНИЕ

[Введение 3](#_Toc122747170)

[1 Теоретические основы интеллектуального анализа данных 6](#_Toc122747171)

[1.1 Общая характеристика и этапы интеллектуального анализа данных 6](#_Toc122747172)

[1.2 Классификация методов Data Mining 9](#_Toc122747173)

[1.3 Области применения методов интеллектуального анализа данных 13](#_Toc122747174)

[2 Анализ особенностей и возможностей методов Data Mining в решении типовых задач 18](#_Toc122747175)

[2.1 Свойства, определяющие выбор метода Data Mining 18](#_Toc122747176)

[2.2 Метод искусственных нейронных сетей 20](#_Toc122747177)

[2.3 Методы поиска ассоциативных правил (Apriori, FP-Growth) 24](#_Toc122747178)

[2.4 Методы кластеризации (CURE, MST, k-средних и пр.) 29](#_Toc122747179)

[2.5 Сравнительная характеристика инструментов Data Mining 34](#_Toc122747180)

[Заключение 38](#_Toc122747181)

[Список использованных источников 40](#_Toc122747182)

[Приложение 43](#_Toc122747183)

# ВВЕДЕНИЕ

В наши дни организации собирают данные с очень поразительной скоростью. Источники этого огромного потока данных разнообразны. Он может исходить из транзакций по кредитным картам, общедоступных данных клиентов, данных банков и финансовых учреждений, а также данных, которые пользователи должны предоставить только для использования и загрузки приложения на свои ноутбуки, мобильные телефоны, планшеты и настольные компьютеры.

Хранение таких огромных объемов данных является нелегкой задачей. Таким образом, многие серверы реляционных баз данных постоянно строятся для этой цели. Онлайн-транзакционный протокол или системы OLTP также разрабатываются для хранения всего этого на различных серверах баз данных. Системы OLTP играют жизненно важную роль в обеспечении бесперебойного функционирования предприятий.

Теперь топ-менеджерам нужен доступ к фактам, основанным на данных, на которых они могут основывать свои решения. Именно здесь онлайн-аналитическая обработка или системы OLAP входят в картину. Хранилища данных и другие OLAP-системы строятся все больше и больше из-за этой самой потребности или топ-менеджеров. Появляется необходимость не только в наличии самих данных, но и в аналитике, связанной с ними, чтобы принимать оптимальные и выгодные решения. Системы OLTP и OLAP работают в тандеме.

Системы OLTP хранят все огромные объемы данных, которые генерируются ежедневно. Затем эти данные отправляются в системы OLAP для построения аналитики на основе данных. Она может помочь в принятии решений, основанных на знаниях, которые могут вывести компанию на новый уровень роста. Однако исследование данных никогда не должно происходить поверхностно. Компаниям нужно более детально анализировать данные, чтобы обогатиться знаниями, которые помогут им принимать лучшие решения для обеспечения успешности бизнеса. Все данные, которые просто скопились за время работы организации, по большому счету бесполезны, если она ничему не учится на них. Доступные данные настолько огромны, что люди чисто физически не способны их обработать и осмыслить. Интеллектуальный анализ данных или обнаружение знаний – это то, что как раз относительно недавно появилось для решения этой проблемы, что обуславливает актуальность и значимость данной темы.

Целью курсовой работы является исследование особенностей и возможностей наиболее часто используемых методов интеллектуального анализа данных.

Для достижения указанной цели необходимо выполнить следующие задачи:

* дать общую характеристику и определить этапы интеллектуального анализа данных;
* определить типы закономерностей, выявляемых с помощью интеллектуального анализа данных;
* изучить классификацию методов Data Mining;
* определить свойства, обуславливающие выбор метода интеллектуального анализа данных;
* провести анализ особенностей и возможностей методов Data Mining в решении типовых задач (включая метод нейронных сетей, методы кластерного анализа, методы поиска ассоциативных правил);
* исследовать инструменты, с помощью которых реализуются методы Data Mining для решения типовых задач.

Объектом исследования данной работы выступают непосредственно сами методы Data Mining. Предметисследования*,* в свою очередь*,* ‒ это возможности и особенности их использования.

Для решения поставленных задач были использованы следующие методы*:* системный и сравнительный анализ, синтез, систематизация, описание, алгоритмизация, структурирование, дедукция.

В качестве информационной базы исследования были использованы статьи и публикации отечественных и зарубежных аналитиков в сфере интеллектуального анализа данных. Теоретической базой работы послужили учебные пособия отечественных ученых по анализу данных, особенно касательно Data Mining.

Структура работы определена характером исследуемых в ней вопросов. Курсовая работа содержит: введение, два раздела, восемь подразделов, заключение, список использованных источников. Во введении обоснована актуальность работы, поставлена цель и задачи, обозначен объект и предмет данной работы. В первой главе рассмотрены теоретические аспекты интеллектуального анализа данных, обозначена классификация методов и типовые решаемые задачи. Во второй главе системный анализ классических современных методов интеллектуального анализа данных. В заключении подведены итоги и сделаны выводы исследования.

# 1 Теоретические основы интеллектуального анализа данных

# Общая характеристика и этапы интеллектуального анализа данных

Интеллектуальный анализ данных – это процесс, который помогает извлекать информацию из заданного набора данных для выявления тенденций, закономерностей и полезных данных. Целью использования интеллектуального анализа данных является принятие поддерживаемых данными решений на основе огромных наборов данных [10].

Интеллектуальный анализ данных работает совместно с прогностическим анализом, отраслью статистической науки, которая использует сложные алгоритмы, предназначенные для работы со специальной группой задач. Прогностический анализ сначала выявляет закономерности в огромных объемах данных, которые интеллектуальный анализ данных обобщает для прогнозов. Также интеллектуальный анализ данных служит уникальной цели, которая заключается в распознавании шаблонов в наборах данных для определения проблем, относящихся к определенной области.

Он делает это, используя сложный алгоритм для обучения модели для конкретной проблемы. Когда аналитик знает область проблемы, с которой имеет дело, он может использовать машинное обучение для моделирования системы, способной идентифицировать закономерности в наборе данных. При задействовании машинного обучение получается автоматизировать систему решения проблем в целом, и пропадает необходимость придумывать специальное ПО для решения каждой проблемы, с которой сталкивается аналитик.

Также интеллектуальный анализ данных можно определить как метод исследования шаблонов данных, которые относятся к определенным перспективам. Это помогает нам классифицировать эти данные в полезную информацию, которая затем накапливается и собирается для хранения на серверах баз данных, таких как хранилища данных, или используется чтобы помочь в принятии решений. Кроме того, он может быть использован для получения доходов и сокращения затрат среди других целей [11].

Итак, интеллектуальный анализ данных – это процесс поиска больших наборов данных для поиска закономерностей и тенденций, которые не могут быть найдены с помощью простых методов анализа. Он использует сложные математические алгоритмы для изучения данных, а затем оценки возможности событий, происходящих в будущем, на основе полученных результатов. Это также называется обнаружением знаний данных или KDD (Knowledge Discovery Data).

Интеллектуальный анализ данных используется организациями для извлечения конкретной информации из больших объемов данных для поиска решений своих бизнес-проблем. Он имеет возможность преобразовывать необработанные данные в знания, которые могут помочь компаниям расти, принимая лучшие решения. Интеллектуальный анализ данных имеет несколько типов, включая интеллектуальный анализ графических данных, интеллектуальный анализ текста, интеллектуальный анализ социальных сетей, веб-майнинг, а также интеллектуальный анализ аудио и видео.

После пониманияопределения интеллектуального анализа данных следует разобраться в самом процессе интеллектуального анализа данных. До того, как осуществляется фактический Data Mining, в подготовке к его реализации участвует несколько процессов (на рисунке 1 представлен весь процесс) [10].

1. Бизнес-исследование. Прежде чем начать интеллектуальный анализ, необходимо иметь полное представление о поставленных целях, доступных ресурсах и текущих сценариях в соответствии с имеющимися требованиями. Это поможет создать подробный план осуществления интеллектуального анализа данных, который эффективно достигает поставленных организацией целей.
2. Проверка качества данных. Поскольку данные собираются из различных источников, они должны быть проверены и сопоставлены, чтобы убедиться, что нет узких мест в процессе интеграции данных.

Рисунок 1 – Процесс интеллектуального анализа данных

Обеспечение качества помогает выявлять любые основные аномалии в данных, такие как отсутствующая интерполяция данных, сохраняя данные в отличной форме до того, как они подвергнутся интеллектуальному анализу.

1. Очистка данных. Считается, что 90% времени уходит на выбор, очистку, форматирование и анонимизацию данных перед майнингом.
2. Преобразование данных. Здесь задействованные процессы, состоящие из пяти подэтапов, которые делают данные готовыми к окончательным наборам данных. Она включает в себя:
* Сглаживание данных: здесь шум удаляется из данных. Шумные данные – это информация, которая была повреждена при передаче, хранении или манипуляции до такой степени, что она непригодна для анализа данных. Помимо потенциального искажения результатов любого исследования интеллектуального анализа данных, хранение зашумленных данных также увеличивает объем пространства, которое должно быть выделено для набора данных.
* Сводка данных. В этом процессе применяется агрегирование наборов данных.
* Обобщение данных: здесь данные обобщаются путем замены любых низкоуровневых данных концептуализацией более высокого уровня.
* Нормализация данных**:** здесь данные определяются в заданных диапазонах. Для работы интеллектуального анализа данных нормализация данных является обязательной. Это в основном означает изменение данных из их исходного формата в более подходящий для обработки. Целью нормализации данных является уменьшение или устранение избыточной информации.
* Data Attribute Construction: перед интеллектуальным анализом данных наборы данных должны находиться в наборе атрибутов.
1. Моделирование данных: для лучшей идентификации шаблонов данных в набор данных реализовано несколько математических моделей, основанных на нескольких условиях.
2. Интеллектуальный анализ данных (важный процесс, в котором применяются интеллектуальные методы для извлечения шаблонов данных).
3. Оценка паттернов (для выявления увлекательных паттернов, представляющих знания на основе некоторых показателей интересности).
4. Представление знаний (где методы представления и визуализации знаний используются для представления добытых знаний пользователю).

# Классификация методов Data Mining

Существует множество методов, используемых для интеллектуального анализа данных, но решающим шагом является выбор наиболее подходящего метода из них в соответствии с постановкой проблемы. Эти методы помогают прогнозировать будущее, а затем принимать соответствующие решения, а также помогает в анализе тенденций рынка и увеличении доходов компании [1, 9]. Примерами наиболее популярных методов представлены на рисунке 2.

Отметим, что каждый указанный метод используется для выборки конкретных данных, построения прогностической модели или для избавления от ошибок в БД. Поэтому для удобства рассмотрения и выделения характерных черт стоит разбить все существующие методы на классы, а классифицировать их можно согласно разным признакам.



Рисунок 2 – Основные методы (техники) интеллектуального анализа данных

Так, одним из способов разбиения методов является классификация по принципу работы с данными, она разбивает методы ИАД (интеллектуального анализа данных) на две большие категории:

1. Методы, которые непосредственно связаны с использованием (сохранением) данных. Данные в ходе обработки детализируются при построении прогностической модели или в ходе анализа исключений. Однако такие методы малоэффективны при работе с Big Data (большими данными). Методики этой категории применяются в формах кластерного анализа, метода ближайшего соседа, метода k-ближайшего соседа, рассуждений по аналогии.
2. Методы, которые дистиллируют шаблоны. Они формируют и применяют закономерности, имеющие упорядоченный вид, т.е. извлекают информацию из изначальных данных путем ее преобразования в определенную систематизированную конструкцию. Технологии этой группы представлены логическими, визуализирующими, кросс-табуляционными и базирующимися на уравнениях методами. Использование таких методов способствует эффективному задействованию полученных в ходе свободного поиска результатов (они более компактны по сравнению с базами данных) и преобразование этих сведений в понятные для пользователей закономерности.

Существует еще одна классификация методов ИАД согласно принципам применения математических моделей в обучении. Здесь выделяют методы:

1. статистические, в которых данные собираются, анализируются, исследуются и представляются для выявления закономерностей и тенденций. В качестве альтернативы это называется количественным анализом.

В статистике есть две основные подкатегории:

– Описательная статистика: целью описательной статистики является организация данных и определение основных характеристик этих данных; графики или цифры суммируют данные; среднее значение, с.о. (стандартное отклонение) и корреляция – вот некоторые из наиболее часто используемых описательных статистических методов.

– Логическая статистика: процесс составления выводов на основе теории вероятностей и обобщения данных; анализируя статистику выборки, можно вывести параметры о совокупностях и создать модели взаимосвязей внутри данных.

На самом деле, современные статистические методы, используемые в области интеллектуального анализа данных, обычно являются производными от обширного статистического инструментария, разработанного для решения проблем, возникающих в других областях. Эти методы преподаются в учебных программах по естественным наукам. Например, необходимо проверить несколько гипотез. Описанные выше гипотезы помогают нам оценить обоснованность наших усилий по интеллектуальному анализу данных при попытке сделать какие-либо выводы из изучаемых данных. При использовании более сложных статистических методов оценки и тестов эти проблемы становятся более выраженными.

1. Нестатистические (кибернетические), в которых используются основы компьютерной математики и технологии искусственного интеллекта. В число таких методов входят: эволюционное программирование, нейросети, системы обработки экспертных знаний. К кибернетическим методам также относятся ассоциативные правила, деревья решений, нечеткая логика, генетические алгоритмы.

Методы Data Mining также можно классифицировать по задачам Data Mining. В соответствии с такой классификацией выделяют две группы. Первая из них – это подразделение методов Data Mining на решающие задачи сегментации (т.е. задачи классификации и кластеризации) и задачи прогнозирования. Другими словами, методы Data Mining могут быть направлены на получение описательных и прогнозирующих результатов.

Описательные методы служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика. К методам, направленным на получение описательных результатов, относятся итеративные методы кластерного анализа, в том числе: алгоритм k-средних, k-медианы, иерархические методы кластерного анализа, самоорганизующиеся карты Кохонена, методы кросс-табличной визуализации, различные методы визуализации и другие.

Прогнозирующие методы используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных. К методам, направленным на получение прогнозирующих результатов, относятся такие методы: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод опорных векторов и др.

В таблице 1 представлены все вышеперечисленные классификации с примерами методов [8].

Таблица 1 – Классификация методов ИАД

|  |  |  |
| --- | --- | --- |
| Классификационный признак | Класс | Примеры |
| Принцип работы с данными | 1) использование (сохранение) данных | кластерный анализ, метод k-ближайшего соседа, рассуждения по аналогии |
| 2) дистилляция шаблонов | логические, визуализирующие, кросс-табуляционные и базирующимися на уравнениях методы |
| Принцип применения математических моделей в обучении | 1) статистические | корреляционный, регрессионный анализ, классификация |
| 2) кибернетические | анализ и прогноз на основе нейронных сетей,эволюционное программирование,генетические алгоритмы,метод группового учета аргументов |
| Решаемая задача | 1) сегментация | алгоритм k-средних, k-медианы, иерархические методы кластерного анализа, самоорганизующиеся карты Кохонена, методы кросс-табличной визуализации, различные методы визуализации |
| 2) прогнозирование | нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод опорных векторов |

Таким образом, при выборе метода стоит сначала определить класс, который будет актуален в решении конкретной задачи, а затем уже и сам метод.

# Области применения методов интеллектуального анализа данных

Рассмотрев классификацию методов ИАД, стоит рассмотреть их применение в различных секторах от здравоохранения до финансов и банковского дела, дабы понять возможности, которые открываются при их применении.

Ниже приведены лишь некоторые области применения интеллектуального анализа данных, в которых он наиболее востребован.

1. Здравоохранение.

Методы интеллектуального анализа данных могут полностью трансформировать систему здравоохранения. Он может быть использован для выявления лучших практик на основе данных и аналитики, которые могут помочь медицинским учреждениям сократить расходы и улучшить результаты лечения пациентов. Интеллектуальный анализ данных, наряду с машинным обучением, статистикой, визуализацией данных и другими методами, может быть использован для изменения ситуации. Он может пригодиться при прогнозировании пациентов разных категорий [13]. Это поможет пациентам получать последующие рекомендации и предложения, которые будут им актуальны. Интеллектуальный анализ данных также может помочь медицинским страховщикам выявлять мошеннические действия.

2. Образование.

Использование методов интеллектуального анализа данных в образовании все еще находится в зачаточной фазе. Он направлен на разработку методов, которые могут использовать данные, поступающие из образовательной среды, для изучения знаний. Цели, которым эти методы, как ожидается, будут служить, включают изучение того, как образовательная поддержка влияет на студентов, поддержку будущих потребностей студентов и продвижение науки об обучении среди других. Учебные заведения могут использовать эти методы не только для прогнозирования того, как студенты будут сдавать экзамены, но и для принятия точных решений. Обладая этими знаниями, эти учреждения могут больше сосредоточиться на своей педагогической педагогике.

3. Анализ рыночной корзины.

Это метод моделирования, который использует гипотезу в качестве основы. Гипотеза гласит, что если вы покупаете определенные продукты, то весьма вероятно, что вы также приобретете продукты, которые не принадлежат к той группе, у которой вы обычно покупаете. Розничные торговцы могут использовать эту технику, чтобы понять покупательские привычки своих клиентов [5]. Розничные торговцы могут использовать эту информацию, чтобы внести изменения в макет своего магазина и сделать покупки намного проще и менее трудоемкими для клиентов.

4. Управление взаимоотношениями с клиентами (CRM)

CRM включает в себя приобретение и удержание клиентов, повышение лояльности и использование клиентоориентированных стратегий. Каждому бизнесу нужны данные о клиентах, чтобы проанализировать их и использовать результаты таким образом, чтобы они могли построить долгосрочные отношения со своими клиентами. Интеллектуальный анализ данных может помочь им в этом. Приложения интеллектуального анализа данных в CRM включают в себя:

* Прогнозирование продаж: Предприятия могут лучше планировать потребности в пополнении запасов, анализируя тенденции с течением времени с использованием методов интеллектуального анализа данных. Он также помогает в финансовом управлении и управлении цепочками поставок и предлагает вам полный контроль над вашими собственными внутренними процессами.
* Сегментация рынка: учитывайте их предпочтения при создании рекламы и других маркетинговых материалов. С использованием методов интеллектуального анализа данных можно определить, какой сегмент рынка обеспечивает наилучшую отдачу от инвестиций. С этой информацией вы не будете тратить время или ресурсы на преследование потенциальных клиентов, которые не заинтересованы в покупке определенного продукта.
* Чтобы улучшить обслуживание бренда, удовлетворенность клиентов и лояльность клиентов, интеллектуальный анализ данных использует концепцию, известную как «кластер клиентов», которая опирается на информацию, которой делятся аудитории социальных сетей.

5. Инжиниринг производства.

Производственная компания во многом полагается на доступные ей данные или информацию. Интеллектуальный анализ данных может помочь этим компаниям в выявлении закономерностей в процессах, которые слишком сложны для понимания человеческим разумом. Они могут определять отношения, существующие между различными элементами проектирования на системном уровне, включая потребности клиентов в данных, архитектуру и портфель продуктов.

Интеллектуальный анализ данных также может оказаться полезным для прогнозирования общего времени, необходимого для разработки продукта, затрат, связанных с процессом, и ожиданий, которые компании могут иметь от конечного продукта [7].

Данные могут быть оценены, гарантируя, что производственная фирма обладает достаточными знаниями определенных параметров. Эти параметры распознают архитектуру продукта, правильный набор портфелей продуктов и требования клиентов. Эффективные возможности интеллектуального анализа данных в производстве и инжиниринге гарантируют, что разработка продукта завершится в оговоренные сроки и не превысит бюджет, выделенный изначально.

6. Финансы и банковское дело.

Банковская система стала свидетелем генерации огромных объемов данных с тех пор, как она подверглась цифровизации. Банкиры могут использовать методы интеллектуального анализа данных для решения проблем, с которыми сталкиваются предприятия, путем выявления корреляций и тенденций в рыночных издержках и деловой информации. Эта работа слишком сложна без интеллектуального анализа данных, так как объем данных, с которыми они имеют дело, слишком велик [14]. Менеджеры в банковском и финансовом секторах могут использовать эту информацию для приобретения, удержания и обслуживания клиента.

Анализ становится простым и быстрым благодаря выборке и распознаванию большого набора данных о клиентах. Отслеживание недоверчивых действий становится простым путем анализа таких параметров, как период транзакции, способ платежей, географическое местоположение, история активности клиентов и многое другое. Относительная мера клиента рассчитывается на основе этих параметров. Следовательно, его можно использовать в любой форме в зависимости от рассчитанных показателей.

Это далеко не все области применения ИАД, потому в таблице 2 представлена систематизация ключевых областей использования наиболее популярных методов Data Mining.

Таблица 2 – Примеры применения интеллектуального анализа данных

|  |  |  |  |
| --- | --- | --- | --- |
|  | Информационные технологии | Торговля | Финансовая сфера |
| Классификация |  |  | Оценкакредитоспособности |
| Регрессия |  |  | Оценка допустимого кредитного лимита |
| Прогнозирование |  | Прогнозирование продаж | Прогнозирование цен акции |
| Кластеризации |  | Сегментация клиентов | Сегментация клиентов |
| Определения взаимосвязей |  | Анализ потребительской корзины |  |
| Анализ последовательностей | Анализ переходов по страницам web-сайта |  |  |
| Анализ отклонений | Обнаружение вторжений в информационные системы |  | Выявление мошенничества с банковскими картами |

Стоит отметить, что одним из важнейших назначений методов интеллектуального анализа данных является наглядное представление результатов вычислений, что позволяет использовать инструментарий интеллектуального анализа данных людьми, не имеющими специальной математической подготовки. В то же время применение статистических методов анализа данных требует хорошего владения теорией вероятностей и математической статистикой.

# 2 Анализ особенностей и возможностей методов Data Mining в решении типовых задач

# 2.1 Свойства, определяющие выбор метода Data Mining

Различные методы Data Mining 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 (пригодность к использованию), трудоемкость, гибкость (разносторонность), быстрота и популярность [18]. Так как все характеристики, кроме масштабируемости, интуитивно понятны, то введем понятие масштабируемости – это способность метода или алгоритма справляться с растущим объемом работы, сохраняя текущий уровень производительности, за счет добавления вычислительных ресурсов.

В таблице 3 приведена сравнительная характеристика некоторых распространенных методов. Оценка каждой из характеристик проведена по следующим категориям в порядке возрастания: чрезвычайно низкая, очень низкая, низкая/нейтральная, нейтральная/низкая, нейтральная, нейтральная/высокая, высокая, очень высокая.

Как видно из рассмотренной таблицы, каждый из классических методов имеет свои сильные и слабые стороны. Но ни один метод не может обеспечить решение задачи Data Mining с наилучшим результатом по всем критериям.

Также дополнительным свойством при выборе метода может послужить его наличие в располагаемом аналитиком инструменте. Хоть обычно сам инструмент должен подбираться под конкретную задачу, однако есть инструменты, которые предусматривают решение широкого спектра задач, отчего аналитик предпочитает работу именно с ним [12].

Таблица 3 – Сравнительная характеристика методов Data Mining

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Алгоритм** | **Точность** | **Масштабируемость** | **Интерпретируемость** | **Пригодность к использованию** | **Трудоемкость** | **Разносторонность** | **Быстрота** | **Популярность, широта использования**  |
| **классические методы (линейная регрессия)** | нейтральная | высокая | высокая / нейтральная | высокая | нейтральная | нейтральная | высокая | низкая |
| **нейронные сети** | высокая | низкая | низкая | низкая | нейтральная | низкая | очень низкая | низкая |
| **методы визуализации** | высокая | очень низкая | высокая | высокая | очень высокая | низкая | чрезвычайно низкая | высокая / нейтральная |
| **деревья решений** | низкая | высокая | высокая | высокая / нейтральная | высокая | высокая | высокая / нейтральная | высокая / нейтральная |
| **полиномиальные нейронные сети** | высокая | нейтральная | низкая | высокая / нейтральная | нейтральная / низкая | нейтральная | низкая / нейтральная | нейтральная |
| **k-ближайшего соседа** | низкая | очень низкая | высокая / нейтральная | нейтральная | нейтральная / низкая | низкая | высокая | низкая |

В универсальных прикладных статистических пакетах (например, SPSS, SAS, STATGRAPHICS, Statistica, др.) реализуется широкий спектр разнообразнейших методов (как статистических, так и кибернетических). Следует учитывать, что для возможности их использования, а также для интерпретации результатов работы статистических методов (корреляционного, регрессионного, факторного, дисперсионного анализа и др.) требуются специальные знания в области статистики.

Однако такая универсальность того или иного инструмента часто накладывает определенные ограничения на его возможности. Преимуществом использования таких универсальных пакетов является возможность относительно легко сравнивать результаты построенных моделей, полученные различными методами. Такая возможность реализована, например, в пакете Statistica, где сравнение основано на так называемой «конкурентной оценке моделей». Эта оценка состоит в применении различных моделей к одному и тому же набору данных и последующем сравнении их характеристик для выбора наилучшей из них.

Для дальнейшего рассмотрения были выбраны такие распространенные методы ИАД, как нейронные сети, поиск ассоциативных правил и кластеризация. Наибольшее разнообразие имеют методы, связаны с кластеризацией данных, что порождает к ним все же чуть больший интерес. По этой причине в данной работе сделан акцент на вышеупомянутую типовую задачу Data Mining, а также на несколько более поверхностном уровне затронуты методы нейронных сетей и поиска ассоциативных правил, однако этого будет достаточно для выделения их возможностей и особенностей.

# 2.2 Метод искусственных нейронных сетей

Идея нейросетей заключается в моделировании (повторении) поведения различных процессов на основе исторической информации. Сама нейросеть представляет собой набор специальных математических функций с множеством параметров, которые настраиваются в процессе обучения на прошлых данных. Затем обученная нейросеть обрабатывает исходные реальные данные и выдает свой прогноз будущего поведения изучаемой системы. Суть нейросети заключается в стремлении подражать происходящим процессам. По своей структуре нейронная сеть аналогична мозгу человека и также способна к обучению [6].

Главное отличие нейросетевых моделей от кривых роста или регрессионных методов состоит в том, что если указанные методы подгоняют реальный процесс или явление под стандартную математическую функцию, то нейронные сети подбирают параметры системы уравнений, приводя ее к реальной жизни [16].

Схематически искусственная нейронная сеть состоит из слоя входных сигналов, выходного слоя и нескольких внутренних слоев (рисунок 2).



Рисунок 3 – Схема искусственной нейронной сети

При применении данного метода слой входных сигналов составят значения факторов, оказывающих наибольшее влияние на исследуемый показатель, т.е. значения независимых переменных. А в качестве выходного слоя используются значения прогнозируемых показателей (зависимых переменных), соответствующие имеющемуся набору значений входных переменных.

Процессы построения и обучения сети в программном комплексе, поддерживающем создание нейронных сетей, происходит следующим образом: на вход подаются значения входных переменных, а вид связи и весовые коэффициенты выбираются случайным образом, затем рассчитываются значения выходной переменной. Полученные значения сравниваются с реальными, после этого происходит корректировка весов и вида сети, направленная на уменьшение ошибки.

Производя последовательные итерации, подобные описанной выше, сеть обучается на исторических данных. Для наилучшего обучения все имеющиеся исходные данные разбиваются случайным образом на три выборки: обучающую, тестовую и контрольную.

Главным плюсом метода прогнозирования с помощью нейронных сетей является их:

1. способность к обучению, нейросетевые модели обучаемы и способны получать обоснованный результат на основании данных, которые не встречались в процессе обучения;
2. адаптивность, поскольку обученные действовать в определенной среде, они легко могут быть переучены для работы в условиях незначительных колебаний параметров среды, что имеет место в производственно-хозяйственной деятельности, например, строительных предприятий;
3. способность моделировать реальные связи между влияющими факторами и прогнозируемым процессом или явлением;
4. неплохие результаты на ограниченном количестве исходных данных; однако, в этом случае для достижения наилучших результатов необходимо периодически корректировать нейронную сеть при появлении новых исторических данных;
5. отсутствие проблемы размерности; нейронные сети могут выявлять зависимости даже при большом количестве переменных;
6. независимость от свойств входных данных;
7. возможность быть нелинейными, это особенно актуально, если у входных данных отсутствует функциональная зависимость от результата.

Минусом данного метода следует назвать:

1. сложность построения сети для конкретной задачи – нет стандартной схемы, что вынуждает в каждом случае выполнять конструирование с начала;
2. необходимость значительного опыта исследователя-прогнозиста, а также потребность в специальном программном обеспечении;
3. сложность интерпретации результатов обучения вследствие невозможности объяснения значений параметров элементов сети в терминах решаемой задачи.

Нейронные сети целесообразно использовать для решения плохо формализованных задач (которые требуют трудоёмких вычислений). К таким задачам относятся:

1. Прогнозирование. Способность к обобщению и выявлению скрытых зависимостей внутри элементов сети позволяет справиться с подобными задачами.
2. Классификация объектов анализа. Например, классификация клиентов по степени риска предоставления им займа.

Надо отметить, что, с одной стороны, нейронные сети являются уже достаточно изученным инструментом прогнозирования с разработанной методологией и алгоритмами выполнения работ, а с другой – довольно новым методом по сравнению, например, с регрессионными моделями. Так, в системе управления земельными ресурсами данный метод используется пока недостаточно [2]. Хотя именно здесь он может оказаться особенно эффективным, исходя из специфики земельных ресурсов. Во-первых, имеется в виду большая инерционность протекающих процессов. Во-вторых, инвестиции в недвижимость зачастую осуществляются на более длительный срок, чем в другие виды активов. В-третьих, это связано с тем, что рынок недвижимости менее подвержен влиянию психологических факторов, чем, например, фондовый рынок.

Следовательно, у участников земельных отношений возникает потребность в выполнении прогнозов на средне- и долгосрочный периоды, что делает неприменимыми традиционные методы прогнозирования и открывает широкие возможности для использования нейронных сетей.

# 2.3 Методы поиска ассоциативных правил (Apriori, FP-Growth)

Задача поиска ассоциативных правил относится к прикладным задачам Data Mining и является одной из основных, так как позволяет выявлять корреляцию между данными в объёмных базах данных.

Данный тип задач впервые был сформулирована и определен в ритейле, когда перед маркетологами возник вопрос, какую пользу для бизнеса можно извлечь из анализа большой базы транзакций кассовых операций. Если ранее анализировались только общие объёмы продаж, то анализ кассовых чеков покупателей открывал новые горизонты, так как позволял анализировать отдельные наборы продуктов, приобретаемые покупателями.

Первый алгоритм был создан группой разработчиков из IBM в 1993 году. Тогда же были сформулированы основные постулаты, которые легли в основу целой группы алгоритмов.

Прежде всего, найденные с помощью алгоритмов правила должны быть часто встречаемые. То есть они должны быть не случайны и в анализируемой базе данных должны повторяться как минимум определенное количество раз. Так сказать, подтверждаться на практике. И с точки зрения статистики выборка транзакций, содержащая такое правило, должна быть репрезентативной. Для соблюдения этого требования во всех алгоритмах поиска ассоциативных правил присутствует параметр минимальной поддержки (MinSup - minimal support), указывающий в долях соотношение частоты появления правила к общему числу транзакций в анализируемой выборке.

Как можно заметить, здесь алгоритмы поиска ассоциативных правил сильно отличаются от всех рассмотренных ранее алгоритмов. Если в других методах максимально используются все доступные данные, то алгоритмы поиска ассоциативных правил сразу отсеивают случайные (шумовые) признаки [19].

Второй параметр, используемый во всех алгоритмах поиска ассоциативных правил – минимальная степень доверия правилу (MinConf - minimal confidence), который также указывается в долях единицы. Для объяснения данного параметра надо сказать, что каждое правило состоит из 2-х частей: посыл и следствие. И посыл, и следствие могут состоять как из одного признака, так и из целого набора признаков. В общем случае правило звучит как «если верен посыл, то довольно часто будет и следствие».

Один из самых известных алгоритмов поиска ассоциативных правил – это алгоритм Apriori, который предложили Ракеш Агравал и Рамакришнан Срикант в 1994 году. В основе алгоритма лежит итерационный процесс поиска наиболее частых паттернов в базе данных с последующим извлечением правил из отобранных паттернов.

Чтобы лучше его понять, предлагается рассмотреть действие алгоритма на небольшом примере из 10 транзакций с 5-ю признаками (рисунок 4).



Рисунок 4 – Пример входных данных для поиска ассоциативных правил

Надо понимать, что все алгоритмы поиска ассоциативных правил работают с бинарными массивами. Поэтому, для начала работы приведенные выше данные были также представлены в виде бинарной таблицы. Также были введены в задачу константы минимальной поддержки 0.3 и минимальной степени доверия 0.7 (30% и 70% соответственно).

Из правой таблицы на рисунке 4 легко посчитать, что признак A встречается всего 2 раза и его поддержка составляет 0.2 или 20%. Аналогично посчитаны поддержки для остальных признаков: B – 0.6, C – 0.7, D – 0.8, E – 0.4. Как можно заметить, только признак А не удовлетворяет требованию минимальной поддержки, и его стоит исключить из дальнейшей обработки по свойству анти-монотонности.

Из оставшихся элементов составляются кандидаты в часто встречающиеся паттерны. На предыдущем шаге были определены часто встречающиеся признаки. И, согласно алгоритму, определены в кандидаты наборы из 2-х признаков: BC, BD, BE, CD, CE, DE.

Теперь необходимо повторно пройтись по всей базе данных и определить поддержку для каждого из отобранных кандидатов (приложение).

Далее итерационный процесс продолжается и на это раз составляются наборы-кандидаты из 3-х признаков. Для этого берутся отобранные на предыдущей итерации частые паттерны и объединяются пары, отличающиеся только одним элементом. Таким образом определяются 4 кандидата: BCD, BCE, BDE, CDE, и согласно алгоритму Apriori необходимо опять пройтись по всей базе данных и определить поддержки для всех новых кандидатов.

На последней итерации ввиду отсутствия новых кандидатов процесс поиска частых паттернов останавливается и осуществляется переход ко второму подпроцессу – определению правил из отобранных частых паттернов. Для этого отобранные паттерны разбиваются на посыл и следствие. После чего определяется степень доверия каждому правилу и сравнивается с минимальным требуемым уровнем доверия.

Правила строятся последовательно для каждого товара и набора. Так как на первом этапе были отсеяны все паттерны с признаком *A* (его поддержка ниже минимального требования), то определение правил начинается с признака *B*. Конечный результат применения метода Apriori представлен в таблице 4.

Таблица 4 – Полученные ассоциативные правила методом Apriori

| Паттерн | Посыл | Поддержка | Правило |
| --- | --- | --- | --- |
| BC (0.5) | B (0.6) | 0.83 | B -> C |
| CD (0.6) | D (0.8) | 0.75 | D -> C |
| CE (0.4) | E (0.4) | 1.00 | E -> C |
| BCD (0.4) | BD (0.4) | 1.00 | BD -> C |
| BCE (0.3) | BE (0.3) | 1.00 | BE -> C |
| CDE (0.3) | DE (0.3) | 1.00 | DE -> C |
| CD (0.6) | C (0.7) | 0.86 | C -> D |
| DE (0.3) | E (0.4) | 0.75 | E -> D |
| BCD (0.4) | BC (0.5) | 0.80 | BC -> D |
| CDE (0.3) | CE (0.4) | 0.75 | CE -> D |

Сейчас был рассмотрен один из наиболее известных алгоритмов поиска ассоциативных правил Apriori. Но надо сказать, что несмотря на его простоту и известность, сейчас его мало кто использует на практике. Дело в том, что, как можно заметить, наиболее узким местом рассмотренного метода является многократный проход по базе данных для оценки поддержек кандидатов в частые паттерны. С ростом объема анализируемых баз данных это становится все большей и большей проблемой. Она практически решена в следующем алгоритме, который требует только 2 прохода по базе данных не зависимо от её объема и количества анализируемых признаков.

Вариант решения вышеописанных проблем предлагается рассмотреть на примере одного из наиболее быстрых алгоритмов поиска ассоциативных правил FP-Growth (Frequent Pattern - Growth). Благодаря особенностям построения алгоритма, в процессе его выполнения полный перебор всех элементов обучающей выборки осуществляется только 2 раза. Других обращений к обучающей выборке алгоритмом не предусмотрено.

Как и выше рассмотренный алгоритм поиска ассоциативных правил, алгоритм FP-Growth условно можно разделить на 2 подзадачи:

1. Нахождение часто встречающихся паттернов. В рассматриваемом алгоритме называется построением FP-дерева.
2. Определение правил.

Начинается работа алгоритма с отсеивания случайных признаков. Для этого, аналогично предыдущему алгоритму, осуществляется первый проход по всей обучающей выборке и подсчитывается поддержка каждого признака. После чего удаляются все признаки, частота появления которых меньше минимальной поддержки.

Далее осуществляется построение (выращивание) FP-дерева. Для этого производится второй проход по обучающей выборке. При этом в каждой транзакции берутся только часто встречающиеся признаки выстроенные в порядке убывания поддержек и выстраиваем путь в нашем дереве. Таким образом, узел с максимальной поддержкой будет в корне дерева, а с минимальной поддержкой – будет его листом. При этом для каждого узла создается счетчик. И на первой итерации присваиваем значения счетчика равное 1 (или 1/N, где N - размер обучающей выборки). Цикл итераций повторяется до полного перебора всей обучающей выборки. Для приведенного выше примера получается FP-дерево, представленное на рисунке 5.



Рисунок 5 – Построенное FP-дерево по транзакциям

Вполне очевидно, что вначале процесса выращивания FP-дерева в большей части будут создаваться новые узлы. Но в процессе продвижения по обучающей выборке осуществляется приход ко все большему увеличению счетчиков существующих узлов без создания новых ответвлений. Изюминка данного алгоритма в том, что в процессе построения дерева можно сжать обучающую выборку до таких размеров, которыми легко можно оперировать в рамках оперативной памяти компьютера без обращения к базе данных.

Дальнейшая работа по определению правил осуществляется с FP-деревом, без обращения к исходной базе данных. Построение правил осуществляется для всех признаков в порядке возрастания их поддержек.

# 2.4 Методы кластеризации (CURE, MST, k-средних и пр.)

Кластеризация (или кластерный анализ) – это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Применение кластерного анализа сводится к следующим этапам [13]:

1. Отбор выборки объектов для кластеризации;
2. Определение множества переменных, по которым будут оцениваться объекты в выборке (при необходимости – нормализация значений переменных);
3. Вычисление значений меры сходства между объектами;
4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров);
5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Существует целое множество методов кластеризации, которые можно разбить на следующие группы согласно разным критериям (таблица 5).

Таблица 5 – Группы методов кластеризации по разным критериям

|  |  |
| --- | --- |
| Признак | Группы методов |
| По способу обработкиданных  | Иерархические | Агломеративные (AGNES) |
| Дивизимные (DIANA) |
| Неиерархические |
| По способу анализа данных | Четкие |
| Нечеткие |
| По количеству применений алгоритмов кластеризации | С одноэтапной кластеризацией |
| С многоэтапной кластеризацией |
| По возможности расширения объема обрабатываемых данных | Масштабируемые |
| Немасштабируемые |
| По времени выполнения кластеризации  | Потоковые (on-line |
| Не потоковые (off-line) |

Классической классификацией считается первая, представленная в таблице 5 (по способу обработки данных), по этой причине будет обзорно рассмотрено по одному методу, представляющему иерархические и неиерархические группы, а затем обозначим в таблице достоинства и недостатки с нерассмотренными методами для общего представления.

Начнем с агломеративного (AGNES – AGglomerative NESting) метода – Алгоритм CURE (Clustering Using REpresentatives). Он выполняет иерархическую кластеризацию с использованием набора определяющих точек для определения объекта в кластер. Его назначением является кластеризация очень больших наборов числовых данных. Ограничения метода заключаются в том, что он эффективен для данных низкой размерности, работает только на числовых данных [20].

Пошаговое описание алгоритма CURE следующее:

Шаг 1: Построение дерева кластеров, состоящего из каждой строки входного набора данных.

Шаг 2: Формирование «кучи» в оперативной памяти, расчет расстояния до ближайшего кластера (строки данных) для каждого кластера. При формировании кучи кластеры сортируются по возрастанию дистанции от кластера до ближайшего кластера. Расстояние между кластерами определяется по двум ближайшим элементам из соседних кластеров. Для определения расстояния между кластерами используются «манхеттенская», «евклидова» метрики или похожие на них функции.

Шаг 3: Слияние ближних кластеров в один кластер. Новый кластер получает все точки входящих в него входных данных. Расчет расстояния до остальных кластеров для новообразованного кластера. Для расчета расстояния кластеры делятся на две группы: первая группа – кластеры, у которых ближайшими кластерами считаются кластеры, входящие в новообразованный кластер, остальные кластеры – вторая группа. И при этом для кластеров из первой группы, если расстояние до новообразованного кластера меньше, чем до предыдущего ближайшего кластера, то ближайший кластер меняется на новообразованный кластер. В противном случае ищется новый ближайший кластер, но при этом не берутся кластеры, расстояния до которых больше, чем до новообразованного кластера. Для кластеров второй группы выполняется следующее: если расстояние до новообразованного кластера ближе, чем предыдущий ближайший кластер, то ближайший кластер меняется. В противном случае ничего не происходит.

Шаг 4: Осуществление перехода на шаг 3, если не получено требуемое количество кластеров.

В рамках дивизимных методов (DIANA – DIvisive ANAlysis) рассмотрим алгоритм MST (Algorithm based on Minimum Spanning Trees). Его назначение состоит в кластеризации больших наборов произвольных данных.

Представим краткое описание алгоритма:

Шаг 1: Построение минимального остовного дерева:

Связный, неориентированный граф с весами на ребрах G(V, E), в котором V – множество вершин (контактов), а E – множество их возможных попарных соединений (ребер), для каждого ребра (u,v) однозначно определено некоторое вещественное число w(u,v) – его вес (длина или стоимость соединения). Здесь можно применить, например, алгоритм Крускала: обход ребер по возрастанию весов. При условии безопасности ребра добавляем его к остновному дереву. Также стоит упомянуть возможность применения алгоритма Прима и Борувки.

Шаг 2: Разделение на кластеры. Дуги с наибольшими весами разделяют кластеры.

Как представителя метода неиерархической кластеризации рассмотрим алгоритм k-средних (k-means).

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Ограничения заключаются в том, что объем данных должен быть небольшим.

Приведем краткое описание алгоритма:

Этап 1. Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются «центрами» кластеров. Каждому кластеру соответствует один центр.

Этап 2. Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т. д., сравнивая полученные результаты.

Рассмотрев обзорно по одному методу из каждой группы, обозначим в таблице 6 достоинства и недостатки как приведенных выше методов, так и не рассмотренных методов для сравнения [4].

Таблица 6 – Достоинства и недостатки методов кластеризации

|  |  |  |
| --- | --- | --- |
| Метод | Достоинства | Недостатки |
| CURE  | Выполняет кластеризацию на высоком уровне даже при наличии выбросов, выделяет кластеры сложной формы и различных размеров, обладает линейно зависимыми требованиями к месту хранения данных и временную сложность для данных высокой размерности  | Есть необходимость в задании пороговых значений и количества кластеров  |
| BIRCH  | Двухступенчатая кластеризация, кластеризация больших объемов данных, работает на ограниченном объеме памяти, является локальным алгоритмом, может работать при одном сканировании входного набора данных, использует тот факт, что данные неодинаково распределены по пространству, и обрабатывает области с большой плотностью как единый кластер  | Работа с только числовыми данными, хорошо выделяет только кластеры выпуклой или сферической формы, есть необходимость в Задании пороговых значений  |
| MST  | Выделяет кластеры произвольной формы, в т.ч. Кластеры выпуклой и вогнутой форм, выбирает из нескольких оптимальных решений самое оптимальное  | Чувствителен к выбросам  |
| k-средних  | Простота использования; быстрота использования; понятность и прозрачность алгоритма  | Алгоритм слишком чувствителен к выбросам, которые могут искажать среднее; медленная работа на больших бд; необходимо задавать количество кластеров; невозможность применения алгоритма на данных, где имеются пересекающиеся кластеры  |
| PAM  | Простота использования; быстрота использования; понятность и прозрачность алгоритма, алгоритм менее чувствителен к выбросам в сравнении с k-means  | Необходимо задавать количество кластеров; медленная работа на больших базах данных  |
| CLOPE  | Высокие масштабируемость и скорость работы, а также качество кластеризации, что достигается использованием глобального критерия оптимизации на основе максимизации градиента высоты гистограммы кластера. Он легко рассчитывается и интерпретируется. Во время работы алгоритм хранит в ram небольшое количество информации по каждому кластеру и требует минимальное число сканирований набора данных. Clope автоматически подбирает количество кластеров, причем это регулируется одним единственным параметром – коэффициентом отталкивания  | Нет существенных недостатков |

Продолжение таблицы 6

|  |  |  |
| --- | --- | --- |
| Метод | Достоинства | Недостатки |
| Самоорганизующиеся карты Кохонена  | Используется универсальный апроксиматор – нейронная сеть, обучение сети без учителя, самоорганизация сети, простота реализации, гарантированное получение ответа после прохождения данных по слоям  | Работа только с числовыми данными, минимизация размеров сети, необходимо задавать количество кластеров  |
| Алгоритм HCM  | Легкость реализации, вычислительная простота  | Задание количества кластеров, отсутствие гарантии в нахождении оптимального решения  |
| Fuzzy C-means  | Нечеткость при определении объекта в кластер позволяет определять объекты, которые находятся на границе, в кластеры  | Вычислительная сложность, задание количества кластеров, возникает неопределенность с объектами, которые удалены от центров всех кластеров  |

Таким образом, все приведенные методы кластеризации обладают особенностями, обуславливающими степень их эффективности применения в зависимости от поставленной задачи и имеющихся данных.

# 2.5 Сравнительная характеристика инструментов Data Mining

Существует множество инструментов, помогающих процессу принятия решений на основе интеллектуального анализа данных, и выбор правильного инструмента является сложной задачей для специалистов по обработке данных или аналитиков данных. Наиболее частые вопросы состоят в следующем: сколько пользователей задействуют тот или иной инструмент, насколько легко его освоить, как он позиционируется на рынке, и если вы владелец бизнеса, то также встает вопрос о стоимости владения такими инструментами.

Ниже представлены несколько наиболее популярных инструментов интеллектуального анализа данных на сегодняшний день:

1. Python.

Python изначально разрабатывался как объектно-ориентированный язык программирования для разработки программного обеспечения и веб-разработки, а затем был усовершенствован для науки о данных. Python – это самый быстрорастущий язык программирования на сегодняшний день. Более того, это мощный инструмент анализа данных, обладающий отличным набором удобных библиотек для любого аспекта научных вычислений. Python – это бесплатное программное обеспечение с открытым исходным кодом, и его легко освоить [17].

За годы существования Python обзавелся множеством специализированных библиотек. Нас интересуют шесть из них:

* Pandas – отвечает за обработку данных;
* NumPy – работает с матрицами;
* Statsmodels – содержит основные статистические функции и модели;
* Sklearn и PyBrain – специализируются на машинном обучении;
* Matplotlib, отвечающий за визуализацию.

Кроме хорошо документированных библиотек, Python отличается гибкостью и понятным синтаксисом – благодаря последнему, он приятен в работе. Немаловажно и то, что у Python огромное сообщество пользователей, настоящих специалистов своего дела, поэтому язык не перестает развиваться.

1. R.

Появившийся в 1993 году язык R на сегодняшний день является стандартом в области анализа данных. R «заточен» под статистическую обработку данных, работу с графикой и алгоритмами машинного обучения. То есть, R - ведущий язык программирования для статистического моделирования, визуализации и анализа данных. Он в основном используется для статистического анализа, работы с Big Dara и машинного обучения.

R – это бесплатный язык программирования с открытым исходным кодом, который имеет множество улучшений в виде пользовательских пакетов. В то же время R требует некоторых практических знаний в области программирования от пользователя. Тем не менее, это отличный язык, когда дело доходит до синтаксиса и согласованности.

Также R является наилучшим выбором, когда дело доходит до EDA (исследовательский анализ данных) – это подход к анализу наборов данных для обобщения их основных характеристик, часто с помощью визуальных методов.

Манипулировать данными в R легко с помощью таких пакетов, как plyr, dplyr и tidy. R отлично подходит для визуализации и анализа данных с помощью таких пакетов, как ggplot, lattice, ggvis и др. Также у R есть огромное сообщество разработчиков для поддержки развития языка.

В целом, серьезному аналитику данных или специалисту по их обработке следует изучить как – R, так и Питон, поскольку оба они являются программным обеспечением с открытым исходным кодом (а открытый исходный код всегда является достоинством) и их можно скачать бесплатно. Обычно новичкам без какого-либо опыта программирования и для проведения некоторого анализ данных единожды (не на постоянной основе), определенно рекомендуется использовать язык R. В противном случае предлагается использовать оба варианта, а ключевые достоинства и недостатки представим в таблице 7.

Таблица 7 – Достоинства и недостатки языков R и Python

|  |  |  |
| --- | --- | --- |
|  | R | Python |
| Достоинства | - отлично подходит для создания прототипов;- отлично подходит для статистического анализа;- хорошая среда разработки (IDE) | - отлично подходит для написания сценариев и автоматизации различных конвейеров интеллектуального анализа данных;- легко интегрируется в производственный процесс;- может использоваться в различных подразделениях вашей команды разработчиков программного обеспечения;- библиотека scikit-learn отлично подходит для задач машинного обучения;- Ipython также является мощным инструментом для наглядного анализа и презентаций |
| Недостатки | - синтаксис может быть неясным;- документация библиотек не всегда удобна для пользователя;- сложнее интегрировать в производственный процесс | - не такой тщательный статистический анализа, как в R;- более быстрое обучение в сравнении с R, так как вы можете сделать гораздо больше с помощью Python |

1. Программные продукты (Weka, Rapidminer, STATISTICA Data Miner).

Weka – это комплекс алгоритмов и инструментов для анализа данных и построения прогностических моделей. Он характеризуется удобством интерфейса, возможностью предварительной очистки данных, поддерживает различные алгоритмы машинного обучения, выводит результаты работы в удобной форме и позволяет сравнивать их точность, представляет ход обработки данных в виде схем взаимосвязей.

Rapidminer – инструмент для решения задач, связанных с преобразованиями данных, статистическим анализом, машинным обучением и визуализацией результатов.

STATISTICA Data Miner – средство универсальной и всесторонней обработки данных, имеющее большой набор готовых решений, гибкий механизм управления и высокую эффективность разведочного анализа.

Однако стоит отметить, что ни один программный продукт не сможет предоставить столько свободы в рамках машинного обучения, обработки данных и визуализации, сколько это могут сделать языки программирования R и Python. Хотя Weka, например, как инструмент хорош в быстром получении предварительных результатов, а также в представлении результатов работы.

# ЗАКЛЮЧЕНИЕ

В заключение, стоит отметить, что интеллектуальный анализ данных в совокупности с Big Data (большими данными) и Machine Learning (машинным обучением) являются не далекими технологиями будущего, а уже сегодня находятся на вооружении множества компаний. Они позволяют эффективно решать задачи, которые поступают из абсолютно разных сфер бизнеса, от оптимизации систем поставок до увеличения вовлеченности аудитории к продукту и оценки стоимости машины, и т.д.

Цель данного исследования была определена как исследование особенностей и возможностей наиболее часто используемых методов интеллектуального анализа данных. В ходе данной работы были выполнены следующие задачи:

* дана общая характеристика и определены этапы интеллектуального анализа данных;
* определены типы закономерностей, выявляемых с помощью интеллектуального анализа данных;
* изучены классификации методов Data Mining;
* определены свойства, обуславливающие выбор метода интеллектуального анализа данных;
* проведен анализ особенностей и возможностей методов Data Mining в решении типовых задач (включая метод нейронных сетей, методы кластерного анализа, методы классификации);
* исследованы инструменты, с помощью которых реализуются методы Data Mining для решения типовых задач.

В результате проведенного исследования во первой главе была осуществлена систематизация теоретической информации касательно интеллектуального анализа данных. Data Mining был определен как процесс обнаружения новых знаний в уже имеющихся данных, обозначены этапы его осуществления от предварительной обработки данных до их представления. Также выделены такие основные классификации методов Data Mining, как, в первую очередь, по типу решаемой задачи, так и по принципу работы с данными (сохранение или их дистилляция). Стоит отметить, что в ходе исследования были обозначены области применения (информационные технологии, торговля, здравоохранение, образование, финансовая сфера), что указывает на широкую востребованность проведения интеллектуального анализа данных в довольно разнородных областях.

Во второй главе проведен анализ наиболее распространенных методов проведения интеллектуального анализа данных согласно решаемым задачам. Обозначены свойства, определяющие выбор метода (например, масштабируемость, интерпретируемость и точность). Рассмотрены инструменты, используемые при реализации методов и получено заключение о том, что все же лучшим вариантом будет использование языков R или Python, так как ни один программный продукт не сможет предоставить столько свободы в рамках машинного обучения, обработки данных и визуализации, сколько это могут сделать R и Python. Хотя Weka, например, как инструмент хорош в быстром получении предварительных результатов, а также в представлении результатов работы. Что касается непосредственно самих методов, то был сделан вывод о том, что по каждой задаче (будь то прогнозирование, или разбиение на классы) нельзя выделить какой-то один конкретный метод, который бы удовлетворял всевозможным критериям аналитика в решении поставленных задач. Так, например, часто в решении задачи кластеризации используется неиерархический метод k-средних ввиду своей простоты, быстроты и понятности использования, однако он слишком чувствителен к выбросам, которые могут искажать среднее, он медленно работает на больших базах данных, а также не применим на данных, где имеются пересекающиеся кластеры.

Таким образом, путем проведения системного анализа были исследованы особенности и возможности наиболее распространенных методов ИАД с акцентом на методы, применяемые в решении задач кластеризации, а также определены лучшие инструменты для их реализации – это R и Python.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Aleksey M. Classification of tasks of data mining and data processing in the economy // Baltic Journal of Economic Studies. 2017. №3. URL: https://cyberleninka.ru/article/n/classification-of-tasks-of-data-mining-and-data-processing-in-the-economy (дата обращения: 29.11.2022).
2. Gordana V., Slobodan S., Zorica A. Quality of research results in agro-economy by Data Mining // Економика пољопривреде. 2018. №4. URL: https://cyberleninka.ru/article/n/quality-of-research-results-in-agro-economy-by-data-mining (дата обращения: 18.11.2022).
3. Milosz M. Data mining as a modern method of data analysis // Известия КазГАСУ. 2018. №1 (9). URL: https://cyberleninka.ru/article/n/data-mining-as-a-modern-method-of-data-analysis (дата обращения: 04.11.2022).
4. Negmatov U.M. CLUSTERING ALGORITHMS IN DATA MINING // Вестник магистратуры. 2019. №2-2 (89). URL: https://cyberleninka.ru/article/n/clustering-algorithms-in-data-mining (дата обращения: 28.11.2022).
5. Vladimir G., Sergey A. Data mining techniques in Real-time Marketing // Proceedings of the Spring/Summer Young Researchers’ Colloquium on Software Engineering. 2012. №6. URL: https://cyberleninka.ru/article/n/data-mining-techniques-in-real-time-marketing (дата обращения: 29.10.2022).
6. Горбаченко, В. И.  Интеллектуальные системы: нечеткие системы и сети: учебное пособие для вузов / В. И. Горбаченко, Б. С. Ахметов, О. Ю. Кузнецова. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2022. – 105 с. – (Высшее образование). – ISBN 978-5-534-08359-0. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/514580 (дата обращения: 04.11.2022).
7. Горленко, О. А.  Дисперсионный анализ экспериментальных данных: учебное пособие для вузов / О. А. Горленко, Н. М. Борбаць, Т. П. Можаева. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2022. – 132 с. – (Высшее образование). – ISBN 978-5-534-14677-6. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/518202 (дата обращения: 20.10.2022).
8. Дюк В. А., Фомин В. В. Прогнозирование временных рядов на основе методов data mining // Известия СПбГТИ (ТУ). 2022. №13. URL: https://cyberleninka.ru/article/n/prognozirovanie-vremennyh-ryadov-na-osnove-metodov-data-mining (дата обращения: 04.12.2022).
9. Дядичев В.В., Ромашка Е.В., Голуб Т.В. Задачи и методы интеллектуального анализа данных // Геополитика и экогеодинамика регионов. 2017. №3. URL: https://cyberleninka.ru/article/n/zadachi-i-metody-intellektualnogo-analiza-dannyh (дата обращения: 20.11.2022).
10. Миркин, Б. Г.  Введение в анализ данных: учебник и практикум / Б. Г. Миркин. – Москва: Издательство Юрайт, 2022. – 174 с. – (Высшее образование). – ISBN 978-5-9916-5009-0. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/511121 (дата обращения: 22.10.2022).
11. Мхитарян В. С. Анализ данных : учебник для вузов // под редакцией В. С. Мхитаряна. – Москва: Издательство Юрайт, 2022. – 490 с. – (Высшее образование). – ISBN 978-5-534-00616-2. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/511020 (дата обращения: 24.10.2022).
12. Нечипорук Д. В. Особенности технологии Data Mining // Молодой исследователь Дона. 2017. №1 (4). URL: https://cyberleninka.ru/article/n/osobennosti-tehnologii-data-mining (дата обращения: 14.11.2022).
13. Новиков, Ф. А.  Символический искусственный интеллект: математические основы представления знаний: учебное пособие для вузов / Ф. А. Новиков. – Москва: Издательство Юрайт, 2023. – 278 с. – (Высшее образование). – ISBN 978-5-534-00734-3. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/512382 (дата обращения: 13.11.2022).
14. Слепнева Л. Д., Кривоберец В. Б. Использование инструментария data mining в управлении кредитными рисками // Экономика промышленности. 2020. №1-2 (62). URL: https://cyberleninka.ru/article/n/ispolzovanie-instrumentariya-data-mining-v-upravlenii-kreditnymi-riskami (дата обращения: 10.11.2022).
15. Станкевич, Л. А.  Интеллектуальные системы и технологии: учебник и практикум для вузов / Л. А. Станкевич. – Москва: Издательство Юрайт, 2022. – 397 с. – (Высшее образование). – ISBN 978-5-534-02126-4. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/511651 (дата обращения: 24.12.2022).
16. Тришечкин С. Н. Data Mining и метод нейронных сетей // Вестник науки и образования. 2019. №8-1 (62). URL: https://cyberleninka.ru/article/n/data-mining-i-metod-neyronnyh-setey (дата обращения: 14.11.2022).
17. Федоров, Д. Ю.  Программирование на языке высокого уровня Python: учебное пособие для вузов / Д. Ю. Федоров. – 4-е изд., перераб. и доп. – Москва: Издательство Юрайт, 2022. – 214 с. – (Высшее образование). – ISBN 978-5-534-15733-8. – Текст: электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/515076 (дата обращения: 14.11.2022).
18. Фридман О. В. Data Mining - методы и алгоритмы, краткий обзор // Труды Кольского научного центра РАН. 2021. №5 (12). URL: https://cyberleninka.ru/article/n/data-mining-metody-i-algoritmy-kratkiy-obzor (дата обращения: 05.11.2022).
19. Холод И. И., Мотовилов И. А. Визуализация модели ассоциативных правил data Mining // Естественные и математические науки в современном мире. 2019. №6 (41). URL: https://cyberleninka.ru/article/n/vizualizatsiya-modeli-assotsiativnyh-pravil-data-mining (дата обращения: 12.11.2022).
20. Щербакова Н. Г. Анализ IP-трафика методами Data Mining проблема кластеризации // Проблемы информатики. 2019. №1 (18). URL: https://cyberleninka.ru/article/n/analiz-ip-trafika-metodami-data-mining-problema-klasterizatsii (дата обращения: 04.12.2022).

# ПРИЛОЖЕНИЕ



Рисунок – Итерации поиска ассоциативных правил методом Apriori