***Авторы:***

***Мошкович Софья Михайловна***

студентка

***Клементьева Светлана Сергеевна***

студентка

***Матюшонок Александр Геннадьевич***

студент

***Научный руководитель:***

***Додонов Михаил Витальевич***

канд. пед. наук, доцент, преподаватель

ФГАОУ ВО «Самарский государственный

аэрокосмический университет

им. академика С.П. Королёва (НИУ)»

г. Самара, Самарская область

# АЛГОРИТМЫ РЕАЛИЗАЦИИ DATA MINING В БАЗАХ ДАННЫХ

***Аннотация****: в данной работе авторами рассмотрены и проанализированы различные методы Data Mining. Описан алгоритм реализации приложения, осуществляющий методы данного анализа.*

***Ключевые слова****: кластеризация, ассоциативные связи, анализ данных, Data Mining, базы данных, итерации.*

*Основные определения*:

*Data Mining* (Высокоинтеллектуальный исследование данных) – данная методика нахождения необходимых взаимосвязей в сыром потоке данных, для нахождения нужного нетривиального решения.

## *Введение*

Многочисленные фирмы в течение долгого периода накапливают большой объём данных, рассчитывая, то что все они несомненно помогут им в принятии верных решений. Предположим, узнать, то что в тот или иной определенный период потребитель приобрёл некоторый продукт в торговом центре 123 – не

так ужи трудно. Но тут необходимы знания – знания о том, то что, к примеру, торговые точки 123 и 130 реализуют продукт X в несколько раз быстрее, чем прочие торговые точки. В данном случае мы можем использовать различные алгоритмы, анализировать данные и получать результаты, которые будут благоприятно сказываться на прибылях компании. Таким образом в Data Mining (DM) находится необходимый набор процедур обнаружения таких кластеров необходимой информации о коммерческой стороне.

Рассмотрим область применения Data Mining. Данный анализ помогает улучшить работу предприятия, т.к. при применении данного анализа исследователи могут дать более точную оценку результатов событий, происходящих в фирме. Data Mining имеет широкое применение в различных областях человеческой деятельности, таких как оптовая и розничная торговля, здравоохранение, сфера образования, промышленные производства и пр.

*Алгоритм C4.5*

Первый из них это C4.5 – один из наиболее популярных алгоритмов построения деревьев решений. Этот метод обрабатывает входные данные таким образом, чтобы определить их классовую принадлежность. Если говорить конкретнее, то во входных данных каждый объект должен иметь набор атрибутов, на основе которых алгоритм определит к какому классу его можно отнести.

Алгоритм, опираясь на обучающую выборку примеров C4.5 строит дерево, постепенно разделяя множество на подмножества с разными значениями атрибутов. Затем полученные подмножества делятся дальше, но уже проверяется различие иного атрибута. Процедура продолжается, пока в порожденном множестве не окажется примеры из одного класса, либо оно не окажется пустым.

Такую систему можно применять для принятия решений, если каждому классу сопоставить решение, действие которое будет применено к каждому объекту в нем.

Недостатки алгоритма состоят в том, что он неприменим для нечеткой логики (когда примеры принадлежат к классу с некоторой вероятностью), а также, что ему необходима начальная выборка примеров. Но тем не менее, деревья решений просто интерпретируются и имеют большую скорость работы.

*Метод к-средних*

Создает к-групп из набора данных таким образом, чтобы объекты группы были наиболее однородными. Это широко используемая техника кластерного анализа для исследования предоставленного набора данных. Для начала разберёмся, что такое кластерный анализ. Кластерный анализ – это набор алгоритмов, разработанных для формирования групп таким образом, чтобы объекты группы были наиболее схожи друг с другом и отличались от элементов, не входящих в группу. Кластер(объединение нескольких схожих элементов, которые могут рассматриваться как отдельная единица, обладающая определенными свойствами) и группа – являются синонимами в вопросах кластерного анализа.

Рассмотрим пошаговый алгоритм выполнения метода к-средних:

1. Метод к-средних выбирает позиции из многомерного пространства, которые будут представлять к-кластеры. Данные элементы называются центрами тяжести.

2. Каждый элемент мы расположим на самое ближайшее расстояние к одной из точек. Благодаря этой итерации создаётся несколько групп.

3. Теперь у нас есть к-кластеров, и каждый объект – это член какого-то из них.

4. Метод к-средних, учитывая положение элементов кластера, находит центр каждого из к-кластеров.

5. Вычисленный центр становится новым центром тяжести кластера.

6. Поскольку центр тяжести переместился, элементы вероятно могут сместиться оказаться ближе к другим центрам тяжести. Таким образом может произойти смена кластера,.

7. Шаги 2–6 повторяются до тех пор, пока центр тяжести не перестанут изменяться и группы не стабилизируется. Это называется сходимостью.

*Метод Apriori*

Алгоритм Apriori находит ассоциативные связи и применяется по отношению к каждому элементу базы данных, которая содержит большое количество различных транзакций.

Ассоциативные правила – это техника, применяемая в data mining для изучения соотношений и отношений между элементами базы данных.

Приведём пример использования ассоциативные правил. Предположим, мы имеем базу данных транзакций производимых ежедневно в супермаркете. Как вариант, такая база представляет собой большую таблицу, в ней каждая строка – является номером определённой транзакции, а каждый столбик -это отдельные покупки.

Благодаря использованию метода Apriori мы можем определить товары, купленные вместе – то есть установить ассоциативные правила.

Таким образом мы можем определить товары, которые часто покупают вместе. Главная цель маркетинга – заставить покупателей выбирать и покупать больше единиц товара. Связанные единицы называются наборами.

Приведём бытовой пример. Зубная паста и зубная щётка находятся на прилавке магазина рядом. Между ними есть четкая ассоциативная связь. Это называется двухэлементным набором. Когда база имеет достаточно большие объёмы, обнаружить и учесть все взаимосвязи гораздо сложней, особенно трудоёмки случаи, когда мы имеете дело с трёхэлементными или наборами с большим количеством элементов. Именно в таких ситуациях отлично подходит метож Apriori.

Разберёмся в работе данного алгоритма. Перед тем, как начать его описание, необходимо определить три величины:

1. Первым делом нужно установить размер набора. Из скольких элементов состоит набор: двух, трёх или более?

2. Затем обозначить поддержку – это количество транзакций, входящих в набор, разделенное на общее количество транзакций. Набор, который равен поддержке, является самым часто встречаемым набором.

3. Последний пункт – это определить достоверность, то есть условную вероятность того, что нужный объект будет расположен в одной группе с другими объектами. Для наглядости,можно привести пример описанный выше : зубная паста в нашей корзине имеют большую вероятность (около 70%) оказаться в одной корзине с зубной пастой.

Метод Apriori определяется последовательностью трёх итераций:

1. Объединение. Мы просматриваем базу данных и определяем частоту вхождение определнных элементов.

2. Отсечение. Те группы, которые удовлетворяют условиям поддержки и достоверности, переходят к следующему шагу с наборами, состоящими из двух компонентов.

3. Повторение. Предшествующие две итерации повторяются для каждого элемента набора, до тех пор пока не будет повторно получен ранее определенный размер.

Apriori обычно рассматривается как самообучающийся алгоритм, поэтому его часто применяют для обнаружения важных элементов и требуемых отношений.

В настоящее время часто применяется модификация метода Apriori, способная проводить классификацию маркированных элементов. Apriori хорош тем, что он прост в реализации, понятен в объяснении и имеет большое количество модификаций.

Значительным минусоми алгоритма является то, что в процессе реализации алгоритм трати большое количество ресурсов и в следствие этого производимые итерации могут производиться в течение долгого времени.

Данный метод широко применим. Существует большое количество реализаций Apriori. Самые часто используемые: ARtool, Weka и Orange.

Метод опорных векторов (SVM – Support vector machine) – это тоже алгоритм, использующийся для задач классификации, но в отличие от C4.5, вместо деревьев он использует гиперплоскости.

Таким образом, если исходное множество примеров можно разделить на 2 класса некоторой линией, то последующие объекты будут разделены на классы соответственно по одну и другую сторону этой линии. Разделяющая линия при этом должна быть такой, чтобы расстояние от нее до каждого объекта была максимальной, тогда линия будет оптимальной. Но не всегда можно построить опорную линию, в этом случае поступают так: элементы множества помещают в пространство более высокой размерности так, чтобы там они были разделимы, затем ищут оптимальную гиперплоскость в новом пространстве.

Перенося этот метод на многомерные множества, можно распределять данные и на большее количество классов.

SVM также требует начальное множество примеров, к тому же плохо интерпретируется, но достоинством может быть то, что это достаточно быстрый метод и вполне точный.

*Алгоритм создания приложения, осуществляющего Data Mining*

Естественно, в наш век современных технологий, хотелось бы автоматизировать данный алгоритм. Далее мы опишем, как создать приложение, реализующее Data Mining.

Для того, чтобы создать данный продукт, следует пошагово выполнить следующий алгоритм:

1. Выяснить объёмы проекта, характеризующие, какие сведения следует получить в результате. Немаловажно, для того чтобы план был ориентирован на реализацию необходимых предпринимательских задач.

2. Разработать базу данных для *Data Mining*. Нужная информация может быть расположена в нескольких базах, иногда часть информации хранится не в электронной форме. Данные из различных баз необходимо консолидировать и устранить несоответствия. На самом деле развитие технологии баз данных уже не требует применения алгоритмов DM к отдельной витрине данных. Фактически, эффективный анализ требует корпоративного Хранилища данных, что с точки зрения вложений обходится дешевле, чем использование отдельных витрин. Отметим, что по мере внедрения DM-проектов в масштабе предприятия

количество пользователей растет, все чаще возникает необходимость в доступе к крупным инфраструктурам данных. Современное Хранилище предоставляет не только эффективный способ хранения всех корпоративных данных и устраняет необходимость в использовании других витрин и источников, но и становится идеальной основой для *Data Mining* проектов. Репозиторий данных предприятия обеспечивает согласованные и актуальные данные о клиентах. Внедряя *Data Mining* функции в Хранилище, компании сокращают расходы в двух направлениях. В этом случае, во-первых, уже не нужно приобретать и обслуживать дополнительное оборудование для *Data Mining*. Во-вторых, компании не нужно переносить данные из Хранилища в специальные источники для DM-проектов, при этом экономятся время и материальные ресурсы.

3. Еще один важный момент – очистка данных. Здесь подразумевается проверка на целостность и обработка отсутствующих значений. Точность методов *Data Mining* зависит от качества информации, лежащей в основе.

4. Заметим, что первые два этапа могут занять половину (а то и больше) времени, отведенного на весь проект.

5. Дать количественные оценки элементам данных. Какого человека можно назвать «расточительным»: того, кто тратит 50 или 300 долларов в неделю? Имеет ли смысл группировать стиральные машины и духовки вместе или стоит их рассматривать по отдельности? Сотрудничество с экспертами в предметной области поможет решить подобные вопросы и выделить элементы данных, которые несут максимальный смысл с точки зрения бизнеса.

6. Применить алгоритмы *Data Mining* для определения отношений между данными. И не исключено, что для выявления нужных зависимостей придется использовать несколько различных алгоритмов. Одни из них подойдут на первых этапах процесса, другие на более поздних. В определенных случаях имеет смысл запустить несколько алгоритмов параллельно, чтобы проанализировать данные с разных точек зрения.

7. Исследовать соотношения, выявленные на предыдущих этапах, на применимость в масштабах проекта. На этом этапе может потребоваться помощь

эксперта в предметной области. Он определит, являются ли те или иные отношения слишком специфичными или слишком общими, и укажет, в каких областях следует продолжить анализ.

8. Представить результаты в виде отчета, в котором будут перечислены все интерпретируемые отношения. Такой отчет принесет только одномоментную выгоду, тогда как приложение, позволяющее эксперту творчески подходить к выявлению отношений, гораздо полезнее. Поэтому фирма-поставщик должна не только научить клиента методике поиска зависимостей в данных, но и обратить особое внимание на обучение работе с самой программой.

*Заключение*

Данный анализ широко применим в различных сферах человеческой деятельности. Различные модификации позволяют оптимизировать нахождение необходимой информации в огромных базах данных. А автоматизации только улучшит работу различных фирм.

***Список литературы***

1. Глушаков С.В. Базы данных / С.В. Глушаков, Д.В. Ломотько. – М.: Харьков: Фолио, 2000. – 504 с.

2. Администрирование баз данных в операционной системе UNIX. – М.: СПб: ЦКТиП Газпром.

3. Lewalle J. Введение в анализ данных с применением непрерывного вейвлет-преобразования / Lewalle J. – М., 1998. – 742 с.