**Отчет**

**о научно-исследовательской работе**

**на тему: «Разработка моделей и методов интеллектуального анализа больших данных»**

направление: Информатика и вычислительная техника

профиль: Системы автоматизированного проектирования

Выполнил:

студент группы 18ВВм2

Матюнин А. Р.

Проверил:

д.т.н., профессор Финогеев А. Г.

СОДЕРЖАНИЕ

[**Введение 3**](#_Toc534788955)

[**1. Выбор темы исследования. Постановка цели и задач НИР. Определение объекта и предмета исследования. Обоснование актуальности выбранной темы. 3**](#_Toc534788956)

[1.1. Тема научно-исследовательской работы 3](#_Toc534788957)

[1.2. Объект и предмет исследования 3](#_Toc534788958)

[1.3. Обоснование актуальности исследования 3](#_Toc534788959)

[1.4. Задачи по научно-исследовательской работе 4](#_Toc534788960)

[**2. Основная часть 6**](#_Toc534788961)

[2.1. Описание методов исследования 6](#_Toc534788964)

[2.2. Анализ литературных источников 6](#_Toc534788965)

[2.3. Сравнение существующих методов анализа больших данных 9](#_Toc534788966)

[**3. Заключение 12**](#_Toc534788967)

[**4. Список использованной литературы 13**](#_Toc534788968)

# Введение

Научно-исследовательская работа (НИР) проводилась на протяжении всего первого семестра обучения.

Цель работы: закрепление, расширение, углубление и систематизация знаний, полученных при изучении общепрофессиональных и специальных дисциплин согласно теме научно-исследовательской работы.

# Выбор темы исследования. Постановка цели и задач НИР. Определение объекта и предмета исследования. Обоснование актуальности выбранной темы.

## Тема научно-исследовательской работы

Для проведения научно-исследовательской работы была выбрана тема "Разработка моделей и методов интеллектуального анализа больших данных".

## Объект и предмет исследования

Объектом исследования является постоянно увеличивающейся объем часто неструктурированных данных, которые требуют анализа при помощи электронно-вычислительных машин.

Предметом исследования является сравнение алгоритмов интеллектуального анализа больших данных с минимальным потреблением как временных, так и аппаратных ресурсов.

## Обоснование актуальности исследования

На данный момент, проблема анализа больших данных стоит очень остро, так как объем данных постоянно увеличивается, а вместе с ним и время их обработки. Эффективная обработка данных не только одним человеком, но и группой людей становится невозможной.

При анализе данных с помощью методологий интеллектуального анализа данных можно построить описательные и предсказательные модели на основе нескольких алгоритмов. Полученные результаты могут выявить скрытые, неявные закономерностей в данных. Экономические, физические и другие системы описываются определенными законами, алгоритмы на основе анализа данных таких систем имеют определенные ограничения. Теоретическая база прикладной статистики также оказалась неэффективной для решения задачи анализа больших данных.

До сих пор не выявлено универсального алгоритма анализа больших данных, который подходил бы для решения абсолютно любой задачи и был удовлетворительным по затратам времени и аппаратных ресурсов.

Таким образом, актуальность исследования подтверждается следующими фактами:

* непрерывное экспоненциальное увеличение объема данных;
* доказательная база об эффективности внедрения анализа больших данных при существующих алгоритмах на предприятиях;
* в отличие от методов математической статистики, методы интеллектуального анализа больших данных обладают большей гибкостью и точностью;
* проблема исключения "выколотых" статистически не значимых точек алгоритмами интеллектуального анализа больших данных;
* невозможность точного утверждения правильности построенной модели через какой-либо промежуток времени при добавлении любого числа новых данных.

## Задачи по научно-исследовательской работе

В результате анализа темы и предмета исследования были поставлены следующие задачи:

1. Сравнение алгоритмов интеллектуального анализа больших данных, выделение критериев сравнения;
2. Выделение наиболее популярных алгоритмов интеллектуального анализа больших данных, анализ причин популярности, сравнение с результатами п. 1.
3. Из результатов п. 1 и п. 2 выделить наиболее перспективные варианты для дальнейшего исследования: анализ математической базы, сферы применения, ограничения применения, формат входных и выходных данных, слабые стороны каждого из алгоритмов;
4. Анализ возможности компиляции сильных сторон математических баз алгоритмов, возможные решения слабых сторон алгоритмов.

В ходе дальнейшего исследования каждая из задач может быть пересмотрена, убрана или добавлена новая.

# Основная часть



## Описание методов исследования

Сравнение существующих моделей и методов алгоритмов интеллектуального анализа больших данных предполагается проводить на основе существующих результатов работы данных моделей и методов.

Проведение исследования на отдельно выделенном кластере электронно-вычислительных машин с большими данными не предполагается, однако не исключается возможность моделирования такого кластера для получения приблизительных результатов.

## Анализ литературных источников

На данный момент, существует множество как отечественных, так и зарубежных статей о больших данных(BigData). Четко выявлены принципы работы с большими данными.

Прежде всего, стоит отметить, что в настоящее время объемы данных увеличиваются во времени не линейно, а по экспоненте. Также анализ данных дает конкурентные преимущества предприятиям, повышает их эффективность путем наиболее быстрого реагирования на изменяющиеся условия рынка [1].

Основные принципы работы с BigData [1]:

1. Горизонтальная масштабируемость: так как данных может быть много, то и система, в которой они хранятся должна быть расширяемой. Если объем данных вырос в 2 раза, то и количество кластеров увеличивается в 2 раза.
2. Отказоустойчивость: горизонтальная масштабируемость подразумевает тот факт, что машин в кластере большое количество. И естественно эти машины будут по тем или иным причинам выходить из строя. Методы работы с BigData должны учитывать этот фактор и продолжать работать без видимых потерь.
3. Локальность данных: в больших системах данные распределены на большом количестве машин. Важным вопросом в проектировании BigData стоит принцип локальности данных, обработке информации там же, где она хранится.

Как в отечественных, так и в зарубежных компаниях проводились исследования внедрения технологий анализа больших данных для управления бизнесом, присутствует доказательная база об эффективности внедрения данных технологий.

Аналитика больших данных позволит более точно оценивать различные вопросы бизнеса, а 38% компаний планируют использовать эту технологию для улучшения анализа продаж и возможностей, предоставляемых рынком. Более 60% надеются, что аналитика больших данных поможет им укрепить ресурсы своего социального маркетинга [2].

Согласно отчету McKinsey Institute «Большие данные: новый рубеж для инноваций, конкуренции и производительности» термин «большие данные» относится к наборам данных, размер которых превосходит возможности типичных баз данных (БД) по занесению, хранению, управлению и анализу информации [3].

Внедрение технологий Больших Данных на 5–7% увеличило эффективность использования ресурсов – труда, основных производственных фондов, энергии и т. п. и на 7–9% обеспечило рост объемов продаж. Для среднего бизнеса показатели оказались в полтора – два раза выше. Причем, следует отметить, что данные получены в условиях, когда мировая экономика испытывает на себе последствия глубочайшего финансово-экономического кризиса и экономический рост составляет в лучшем случае 1–2% [3].

Различными учеными часто выпускаются статьи, в которых предлагаются улучшения каких-либо алгоритмов анализа больших данных.

Существует большое количество методов анализа информации - методы математической статистики, кластеризации, классификации, в том числе «комитеты» алгоритмов, когда результаты, полученные с помощью нескольких алгоритмов, усредняются для построения конечной модели. Перечисленные методы с разной степенью точности могут описывать закономерности в данных. В современных условиях ввиду большого количества накапливающейся информации происходит усложнение ее структуры [4].

В этих условиях для увеличения эффективности анализа больших объемов информации возникает необходимость в новых нетрадиционных методах обработки информации (таких как интеллектуальный анализ данных) посредством автоматизированных познавательных процедур с использованием баз фактов и баз знаний, автоматического порождения гипотез, процедуры объяснения исходного состояния баз фактов с целью оправдания и принятия гипотезы [5].

Задачи интеллектуального анализа данных [6]:

* Задача классификации заключается в том, что для каждого варианта определяется категория или класс, которому он принадлежит;
* Задача регрессии во многом схожа с задачей классификации, но в ходе ее решения производится поиск шаблонов для определения числового значения. Иными словами, предсказываемый параметр здесь, как правило, число из непрерывного диапазона;
* Задача прогнозирования базируется на основании имеющихся значений числовой последовательности;
* Задача кластеризации - заключается в делении множества объектов на группы (кластеры) схожих по параметрам.
* Задача определения взаимосвязей, также называемая задачей поиска ассоциативных правил, заключается в определении часто встречающихся наборов объектов среди множества подобных наборов.
* Анализ отклонений позволяет отыскать среди множества событий те, которые существенно отличаются от нормы.

Таким образом, для исследования существует множество литературных источников, а именно научных журналов, монографий, статей и книг как отечественных, так и зарубежных авторов.

## Сравнение существующих методов анализа больших данных

Будет рассматриваться модель обработки больших данных Data Mining, так как именно она предполагает исследование и обнаружение при помощи вычислительных ресурсов (алгоритмами, средствами искусственного интеллекта) неявных связей между данными.

По результатам анализа темы для исследования были выделены следующие методы анализа:

* Линейная регрессия
* Нейронные сети
* Методы визуализации
* Деревья решений
* Байесовские сети
* Метод k-ближайших соседей

Для всех методов Data Mining обязательно, чтобы одним из входных параметров являлся массив входных данных.

Для каждого из методов можно привести краткий список входных параметров:

* Линейная регрессия
  + массив регрессоров(факторов, независимых переменных);
  + свободный член линейной регрессии;
  + ограничения для результирующей переменной.
* Нейронные сети
  + данные, часто представленные базой данных;
  + ограничения для результирующей переменной;
  + критерии остановки обучения.
* Методы визуализации
  + данные;
  + тип исходных данных (количественные или качественные признаки);
  + тип выходной модели.
* Деревья решений
  + данные;
  + атрибут, условие, располагается в корневом узле дерева;
  + алгоритм построения дерева.
* Байесовские сети
  + данные;
  + количество наблюдений(используется для вероятностного вывода).
* Метод k-ближайших соседей
  + данные;
  + число соседей (максимальное число ближайших по критерию точек входной выборки).

Сравнение методов производится по следующим критериям:

* точность;
* масштабируемость;
* интерпретируемость;
* проверяемость;
* трудоемкость;
* гибкость;
* быстрота;
* популярность.

Для предварительного анализа методов была взята сравнительная характеристика со стороннего источника [7]. Оценка каждого из критериев методов анализа больших данных производилась по шкале от 1 до 8. Результаты оценки представлены на рисунке 1.

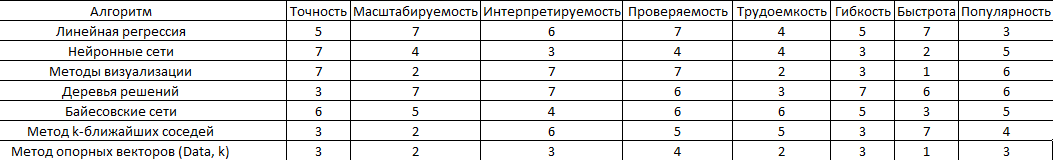


Рисунок 1 – Таблица с оценками критериев.

Выбор наиболее подходящих для последующего анализа методов (алгоритмов) производился при помощи многокритериальной оптимизации с использованием следующих методов:

* метод выделения главного критерия;
* метод лексикографического упорядочения критериев;
* метод последовательных уступок;
* метод свертывания векторного критерия;
* метод идеальной точки.

Практически каждый метод предлагает на выходе разные решения, что не позволяет, применив несколько методов к одной задаче, получить однозначный точный ответ, какой метод лучше применить или разработать с применением теоретической базы.

Однако можно выделить три наиболее часто встречаемых в ответах метода:

* нейронные сети;
* линейная регрессия;
* деревья решений.

# Заключение

В рамках научно-исследовательской работы решены следующие задачи:

1. Выбрана тема исследования;
2. Проведена постановка цели и задач ВКР. Определены объект и предмет исследования. Обоснована актуальность выбранной темы;
3. Проведен обзор литературы по теме больших данных и методам их анализа;
4. Проведено сравнение наиболее используемых категорий методов анализ больших данных и выделение среди них наиболее значимых для дальнейшего исследования при помощи многокритериальной оптимизации.

Дальнейшее исследование следует сосредоточить на изучении математической базы методов анализа больших данных в первую очередь тех, которые были выявлены в результате многокритериальной оптимизации.

# Список использованной литературы

**Публикации в научных журналах, сборниках и СМИ**

1. Веретенников А. В. BigData: анализ больших данных сегодня [Текст] // Молодой ученый. — 2017. — №32. Казань: ООО «Издательство Молодой ученый» — С - 9-12;
2. Виджайан Д. Анализ больших данных: проблемы и решения [Текст] // Сети/Network world. — 2011. — №4. ООО «Издательство «Открытые системы»;

**Научная и учебная литература**

1. Синяк Н. Г., Каклаускас А., Зинькина Д. В., Хабиб А., Шариф Н., Бондаренко А. В. Использование больших данных в оценке экономического потенциала организаций [Текст] // Вильнюсский технический университет им. Гедиминаса. — 2015. — С - 284-285;
2. Бритков В.Б., Булычев А.В. Методы анализа больших объемов слабоструктурированной информации [Текст] // Информационные технологии и вычислительные системы. — 2010. — С - 36-44;
3. Ю.М. Арский, В.К. Финн. Принципы конструирования интеллектуальных систем. Информационные технологии и вычислительные системы. [Текст] // — М.: 2008. — С - 4-37;

**Источники на электронных носителях**

1. Интеллектуальный анализ данных: базовые понятия. [Электронный ресурс] / НОУ ИНТУИТ – Режим доступа: https://www.intuit.ru/studies/courses/2312/612/lecture/13260, свободный. (дата обращения 30.12.2018 г.).
2. Обзор методов Data Mining [Электронный ресурс]. – Режим доступа: http://ai–news/2017/08/obzor\_metodov\_data\_mining\_intellektualnyj\_analiz\_dannyh.html, свободный. (дата обращения 13.11.2018 г.).