TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки: 09.03.04 «Программная инженерия»
Отделение информационных технологий

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Сравнение методов кластеризации на экспериментальных данных проекта «ATLAS»

УДК 004.93'14:519.2:004.6

Студент

Группа	ФИО	Подпись	Дата
8К51	Белоусова Алёна Романовна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР ТПУ	Чердынцев Евгений Сергеевич	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Основная часть»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ведущий программист ОИТ ИШИТР ТПУ	Губин Максим Юрьевич			

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСТН ШБИП ТПУ	Подопригора Игнат Валерьевич	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП ТПУ	Винокурова Галина Федоровна	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР ТПУ	Чердынцев Евгений Сергеевич	к.т.н.		

Томск – 2019 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа: Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность): 09.03.04 «Программная инженерия»
Отделение школы (НОЦ): Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Чердынцев Е.С.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8К51	Белоусовой Алёне Романовне

Тема работы:

Сравнение методов кластеризации на экспериментальных данных проекта «ATLAS»	
Утверждена приказом директора (дата, номер)	14.05.2019 г. №3733/с

Срок сдачи студентом выполненной работы:	13.06.2019 г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Работа направлена на сравнение методов кластеризации и поиска наиболее эффективной технологии анализа описания вычислительных задач по обработке данных эксперимента «ATLAS». Исходными данными к работе являются набор данных, описывающий обработанные в системе задачи и техническое задание на разработки.
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования,*	1. Рассмотрение понятий и методов машинного обучения; 2. Рассмотрение предметной области; 3. Анализ исходных данных; 4. Обработка данных с помощью различных методов кластеризации;

конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).	5. Вычисление параметров качества кластеризации; 6. Выявление наиболее эффективного метода.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Графики распределения данных в двумерном пространстве 2. Графики кластеризации данных 3. Матрица SWOT 4. Таблица трудозатрат на выполнение проекта 5. Линейный график работ 6. Презентация Microsoft PowerPoint

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Подопригора Игнат Валерьевич
Социальная ответственность	Винокурова Галина Федоровна

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель / консультант (при наличии):

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР ТПУ	Чердынцев Евгений Сергеевич	к.т.н.		
Ведущий программист ОИТ ИШИТР ТПУ	Губин Максим Юрьевич			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8К51	Белоусова Алёна Романовна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8К51	Белоусовой Алёне Романовне

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Бакалавриат	Направление/специальность	09.03.04 Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Амортизационные затраты на спецоборудование – 4822,4 рублей;* *Затраты на основную и дополнительную з/п – 74236,04 + 11135,41 рублей;* *Затраты на отчисление во внебюджетные фонды – 23904 рубля;* *Накладные расходы – 17315,83 рубля.*
2. *Нормы и нормативы расходования ресурсов*	
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения*	*Технология QuaD;* *SWOT-анализ.*
2. *Планирование и формирование бюджета научных исследований*	*Структура работ в рамках научного исследования* *Определение трудоемкости выполнения работ и разработка графика проведения научного исследования* *Бюджет проекта*
3. *Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования*	*Оценка сравнительной эффективности проекта*
Перечень графического материала *(с точным указанием обязательных чертежей)***:**	
1. *Оценка конкурентоспособности технических решений* 2. *Матрица SWOT* 3. *График проведения и бюджет НИ* 4. *Оценка ресурсной, финансовой и экономической эффективности НИ*	
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСТН ШБИП ТПУ	Подопригора Игнат Валерьевич	Кандидат экономических наук		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8К51	Белоусова Алёна Романовна		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа: Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность): 09.03.04 «Программная инженерия»
Уровень образования: Бакалавр
Отделение школы (НОЦ): Отделение информационных технологий
Период выполнения: (осенний / весенний семестр 2018 /2019 учебного года)

Форма представления работы:

Бакалаврская работа

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	13 июня 2019 г.

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
02.05.2019	*Раздел 1. Теоретические основы*	*20*
20.05.2019	*Раздел 2. Применение методов кластеризации*	*50*
25.05.2019	*Раздел 3. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение*	*15*
28.05.2019	*Раздел 4. Социальная ответственность*	*15*

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР ТПУ	Чердынцев Евгений Сергеевич	к.т.н.		

Консультант (при наличии)

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ведущий программист ОИТ ИШИТР ТПУ	Губин Максим Юрьевич			

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР ТПУ	Чердынцев Евгений Сергеевич	к.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8К51	Белоусовой Алёне Романовне

Школа	ИШИТР	Отделение (НОЦ)	ОИТ
Уровень образования	Бакалавриат	Направление/специальность	09.03.04 Программная инженерия

Тема ВКР:

Сравнение методов кластеризации на экспериментальных данных проекта «ATLAS»	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Объект исследования – методы кластеризации, применяемые для анализа данных проекта «ATLAS». Рабочее место – рабочий стол с персональным компьютером в общем помещении.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	– Рабочее место соответствует нормам ГОСТа 12.2.032 – 78 – Организация рабочего места с электронно-вычислительной-машиной регулируется СанПиНом 2.2.2/2.4.1340 – Организация труда в течение рабочего времени соответствует Трудовому Кодексу РФ
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	– Отклонение показателей микроклимата – Отсутствие или недостаток естественного света – Монотонность труда – Повышенный уровень электромагнитных излучений – Опасность поражения электрическим током – Повышенный уровень шума
3. Экологическая безопасность:	Анализ негативного влияния на окружающую среду: утилизация макулатуры, оргтехники, светодиодных ламп и батареек
4. Безопасность в чрезвычайных ситуациях:	Возможные чрезвычайные ситуации: пожар
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП ТПУ	Винокурова Галина Федоровна	Кандидат технических наук, доцент		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8К51	Белоусова Алёна Романовна		

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код результата	Результат обучения (выпускник должен быть готов)
Р1	Применять базовые и специальные естественнонаучные и математические знания в области информатики и вычислительной техники, достаточные для комплексной инженерной деятельности.
Р2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач.
Р3	Ставить и решать задачи комплексного анализа, связанные с созданием аппаратно-программных средств информационных и автоматизированных систем, с использованием базовых и специальных знаний, современных аналитических методов и моделей.
Р4	Разрабатывать программные и аппаратные средства (системы, устройства, блоки, программы, базы данных и т. п.) в соответствии с техническим заданием и с использованием средств автоматизации проектирования.
Р5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания аппаратных и программных средств информационных и автоматизированных систем.
Р6	Внедрять, эксплуатировать и обслуживать современные программно-аппаратные комплексы, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды.
Р7	Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности.

Р8	Владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности.
Р9	Эффективно работать индивидуально и в качестве члена группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации.
Р10	Демонстрировать знания правовых, социальных, экономических и культурных аспектов комплексной инженерной деятельности.
Р11	Демонстрировать способность к самостоятельному обучению в течение всей жизни и непрерывному самосовершенствованию в инженерной профессии.

Реферат

Выпускная квалификационная работа содержит: 80 страниц, 19 рисунков, 21 таблицу, 6 приложений и 20 источников.

Ключевые слова: программирование, разработка, анализ данных, машинное обучение, машинное обучение без учителя, кластеризация.

Объектом исследования данной работы является применение методов кластеризации для анализа данных проекта «ATLAS».

Целью работы является реализация методов машинного обучения для анализа описания вычислительных задач по обработке данных эксперимента и поиск наиболее эффективного из них.

В процессе исследования проводился анализ данных, описывающих результаты экспериментов проекта и исследованы методы машинного обучения без учителя.

В результате исследования были выделены 4 метода кластеризации и применены к имеющимся данным. Из этих методов был выбран тот, который за минимальным промежуток времени, относительно других, распределяет все данные по оптимальному количеству кластеров.

Степень внедрения: планируется внедрение в течении следующего полугода.

Область применения: методы машинного обучения без учителя, кластеризация.

Содержание

Введение

В настоящее время, эксперименты по физике высоких энергий являются наиболее актуальным видом физических исследований, вносящих неоспоримый вклад в фундаментальную науку. Как правило, результатом таких экспериментов является огромное количество данных, зарегистрированных детекторами ускорителя заряженных частиц. Обработка такого объема данных требует больших вычислительных мощностей, в связи с чем создаются распределенные системы обработки данных, содержащие большое количество суперкомпьютеров. Однако, использование такого оборудования обходится дорого, поэтому необходима организация рационального планирования обработки данных, во избежание простоя оборудования и неравномерного распределения задач по обработке, среди суперкомпьютеров системы обработки данных. Одним из этапов такой обработки является построение и выделение группы схожих объектов, изучение их особенностей и построение для каждой группы отдельной модели. Данный этап упрощает дальнейшую обработку данных, что сокращает время работы с такими данными.

Целью данный работы является применение методов машинного обучения для анализа описания вычислительных задач по обработке данных эксперимента ATLAS.

Для достижения поставленной цели, выдвинуты следующие задачи:

- Рассмотрение понятий и методов машинного обучения
- Рассмотрение предметной области
- Анализ исходных данных
- Обработка данных с помощью различных методов кластеризации
- Выявление наиболее эффективного метода

Глава 1. Теоретические основы

Целью данной главы является рассмотрение основных понятий кластеризации, которые использованы в рамках данной работы. Проводятся описания методов кластеризации и их оценки.

1.1. Основные понятия кластеризации

Кластеризация – это метод анализа данных, который дает возможность распределить большое количество информации в подгруппы (кластеры), не имея предварительных сведений о принадлежности к группе[1]. Каждый кластер, образующийся при анализе, обозначает группу объектов, обладающих определенной степенью подобия, при этом больше отличающихся от объектов в других кластерах. Использование кластеризации необходимо для структурирования информации и получения содержательных связей внутри данных. На рисунке 1 продемонстрировано применение кластеризации для размещения немаркированных данных в три разные группы на основе подобия их признаков x_1 и x_2:

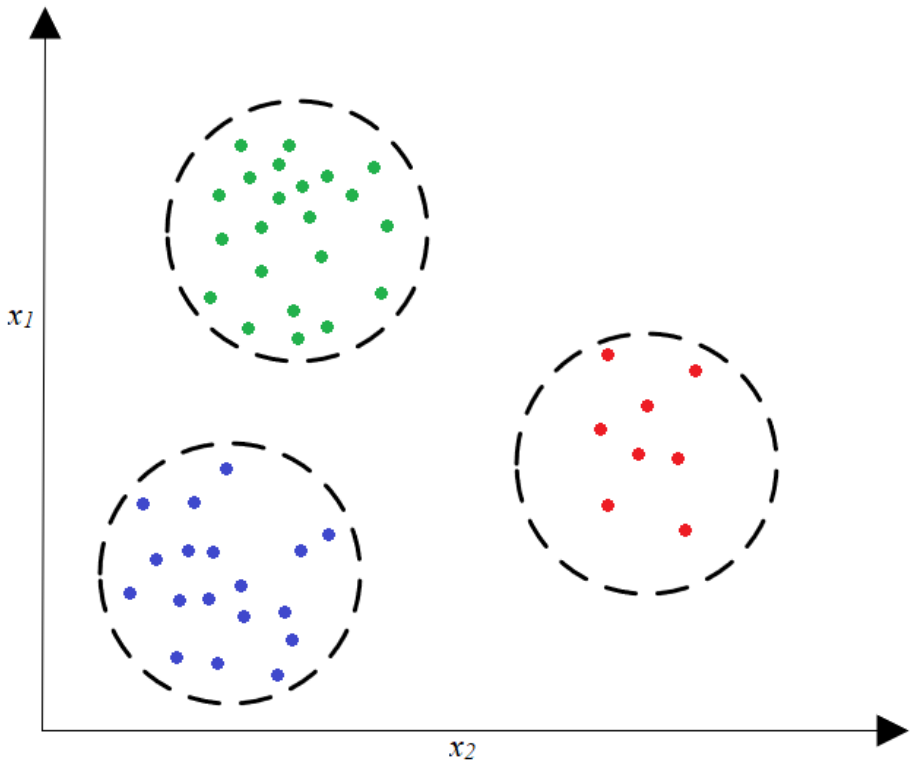


Рисунок 1. Пример кластеризации

В кластеризации можно выделить несколько шагов [2]:

Шаг 1. Выделение выборки объектов для анализа;

Шаг 2. Определение и нормализация некоторого множества переменных, в котором элементы будут оцениваться в выборке;

Шаг 3. Определение степени сходства между элементами;

Шаг 4. Применение метода кластерного анализа для создания групп схожих элементов;

Шаг 5. Демонстрация результатов кластеризации.

При получении некорректного результата, неверно отражающего действительность, возможны внесения изменений: выбор иного метода кластеризации и корректировка выбранной метрики до момента получения наиболее благоприятного результата.

1.2. Метрики расстояний

Первоначально для каждого объекта необходимо сформировать вектор характеристик, с целью определения дальнейшей схожести с другими объектами. В большинстве случаев таким вектором выступают числовые значения. Однако, встречаются алгоритмы, способные обрабатывать категорийные (качественные) характеристики.

Как только определяется вектор характеристик для объекта, проводится нормализация, которая способствует получению равнозначных значений при расчете расстояния между элементами. Первоначальные значения элементов приводят к определенному диапазону, наиболее удобному для вычислений. Например, [0; 1] или [-1; 1].

После этапа нормализации к каждому объекту назначают пару и вычисляют степень их схожести – в случае кластеризации, расстояния между объектами. Данный этап может осуществляться посредством множества метрик [3]. Ниже приведены наиболее распространённые из них:

1. Евклидово расстояние

Данная метрика используется при реализации многих методов кластеризации. Евклидово расстояние представляет собой геометрическое расстояние в многомерном пространстве. Для расчета евклидова расстояния между двумя элементами используется следующая формула:

$$p(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2}, \quad (1)$$

2. Квадрат евклидово расстояния

В случае более отдаленных друг от друга объектов, с целью придать наибольший вес возможно использование квадрата евклидова расстояния. Крайне важно использовать данную мерку для стандартизированных переменных. Для его расчета, значение евклидова расстояния, рассмотренное в предыдущем пункте, возводят в квадрат:

$$p(x, y) = \sum_i^n (x_i - y_i)^2, \quad (2)$$

3. Расстояние Минковского

Метрика расстояние Минковского имеет также название метрика степенного расстояния. Степенное расстояние применяется в случае, когда необходимо в несколько раз уменьшить или увеличить вес для размерности соответствующих элементов, крайне отличающихся друг от друга. Формула степенного расстояния:

$$p(x, y) = \sqrt[r]{\sum_i^n (x_i - y_i)^p}, \quad (3)$$

Где r и p – параметры, определяемые индивидуально, исходя от различия данных. Параметр r влияет на прогрессивное взвешивание больших расстояний между объектами, а значение параметра p способно изменить постепенное взвешивание разностей по отдельным координатам. Если оба параметра – r и p равны двум, то это расстояние совпадает с расстоянием Евклида.

4. Расстояние Чебышева

Эту метрику рекомендовано использовать в случае, когда требуется определить различие двух объектов по какой-либо одной координате. Рассматриваемое расстояние соответствует значению максимального расстояния между соответствующими координатами объектов. Данная метрика является неточным измерением различия, так как значительная часть имеющейся информации не рассматривается.

$$p(x, y) = \max_i |x_i - y_i|, \quad (4)$$

5. Манхэттенское расстояние

Также известно, как расстояние городских кварталов. Результаты вычисления по этой метрике зачастую совпадают с результатами вычисления Евклидова расстояния. Но в случае этого расстояния уменьшается влияния отдельных выбросов (больших разностей), поскольку их не возводят в квадрат. Расстояние городских кварталов рассчитывается как среднее разностей по координатам. Для расчета такого расстояния между двумя точками используют следующую формулу:

$$p(x, y) = \sum_i^n |x_i - y_i|, \quad (5)$$

1.3. Метод К-средних

Метод К-средних (K-means) – наиболее популярный и простой метод кластеризации.

Данный алгоритм направлен на минимизацию суммарного квадратичного отклонения точек кластеров от их центров:

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2, \quad (6)$$

Где, k соответствует количеству кластеров, S_i – полученные кластеры, i = 1, 2, …, k, а μ_i – центры масс всех векторов x из кластера S_i.

Каждый кластер представлен прототипом, который может быть либо центроидом (средним) подобных точек с непрерывными признаками, либо медоидом (наиболее представительной или наиболее часто встречающейся точкой) в случае категориальных признаков. В то время как алгоритм К-средних очень хорошо выполняет идентификацию кластеров сферической формы, один из недостатков этого алгоритма кластеризации состоит в том, что нам нужно указывать число кластеров k. Некорректный выбор числа k может привести к плохой кластеризующей способности [4].

Алгоритм К-средних можно состоит из четырех этапов:

1. Случайно выбрать из точек образцов к центроидов, как исходных центров кластеров
2. Назначить каждый образец самому ближайшему к ней центроиду
3. Переместить каждый центроид в центр образцов, которые были ему назначены
4. Повторять шаги 2 и 3, пока назначения кластеров не перестанут изменяться, либо не будет достигнут заданный пользователями допуск или максимальное число итераций.

Данный алгоритм показывает хорошие результаты при работе с данными, но стоит брать во внимании и минусы его реализации.

Основные сложности алгоритма К-средних:

1. Необходимо указывать заранее число кластеров;
2. Гарантируется лишь достижение одного из локальных минимумов, а не глобального минимума суммарного квадратичного отклонения;
3. Результат напрямую зависит от выбора исходных центров групп элементов(кластеров), их оптимальный выбор заранее неизвестен.

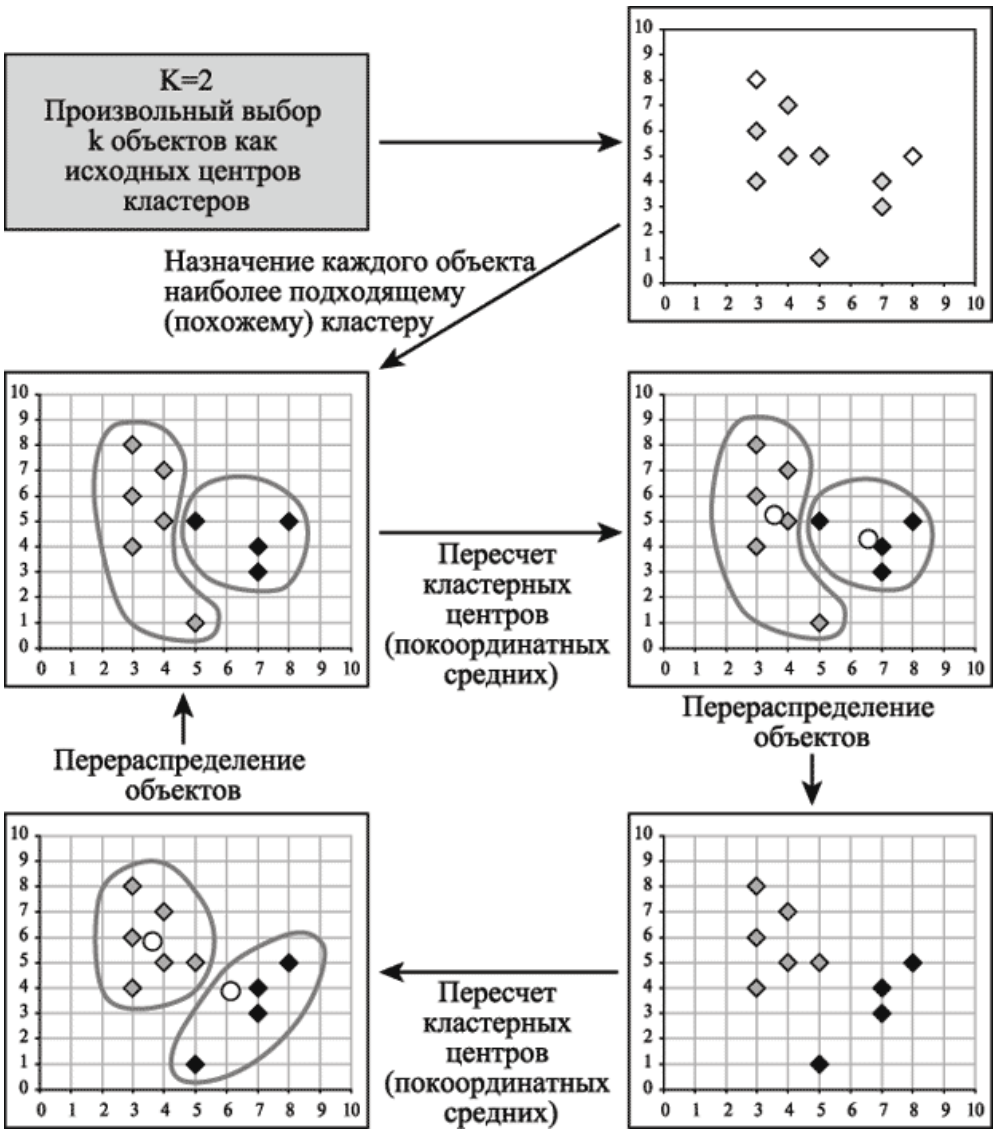


Рисунок 2. Схема алгоритма К-Средних

Так как для использования данного метода необходимо указывать число кластеров, следует использовать методы для нахождения этого параметра.

1.3.1. Метод локтя

Одна из основных трудностей кластеризации состоит в том, что мы не знаем точного ответа. В исследуемом наборе данных нет установленных данных о метках классов, позволяющих применять методы. Поэтому для количественного определения качества кластеризации нам нужно использовать внутренние метрики – такие как внутрикластерная SSE

(искажение или инерция) – для сравнения качества разных кластеризаций по методу к-средних [1].

Основываясь на внутрикластерной SSE, мы можем применить графический инструмент, метод локтя, для оценки оптимального числа k кластеров для поставленной задачи. Интуитивно можно сказать, что если k увеличивается, то искажение уменьшается. Это вызвано тем, что образцы будут ближе к центроидам, которым они назначены. В основе метода локтя лежит идея, которая состоит в том, чтобы идентифицировать значение k в точке, где искажение начинает увеличиваться быстрее всего, что станет понятнее, если построить график искажения для разных значений k.

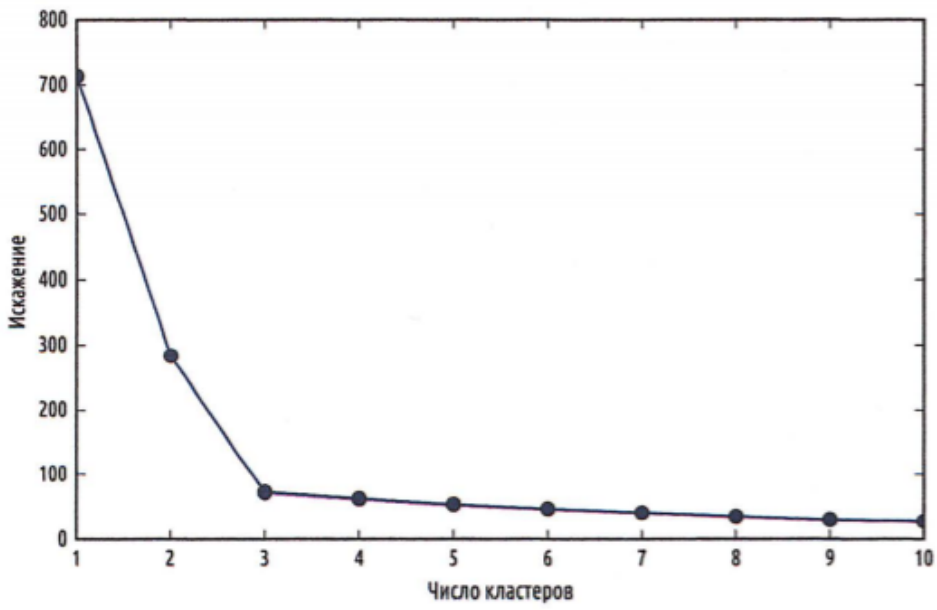


Рисунок 3. Пример использования метода локтя

Из рисунка видно, что для данной ситуации оптимальное количество кластеров – три.

1.3.2. Организация кластеров в виде иерархичного дерева

Еще одним способом определить количество кластеров является построение дендограммы. Дендограмма представляет собой графическое представление последовательности объединения и разделения кластеров и

определяет степень их близости. Число шагов разделения или слияния групп элементов(кластеров) пропорционально количеству уровней дендограммы.

Дендограммой является иерархичное дерево, построенное опираясь на матрицу степени близости. Иерархическое дерево наглядно показывет связи между объектами из заданного множества. При его создании применяют матрицу сходства или различия, определяющую степень сходства объектов множества. Чаще всего ими выступают агломеративные методы [5].

Для построения матрицы сходства (различия) необходимо задать меру расстояния между двумя кластерами.

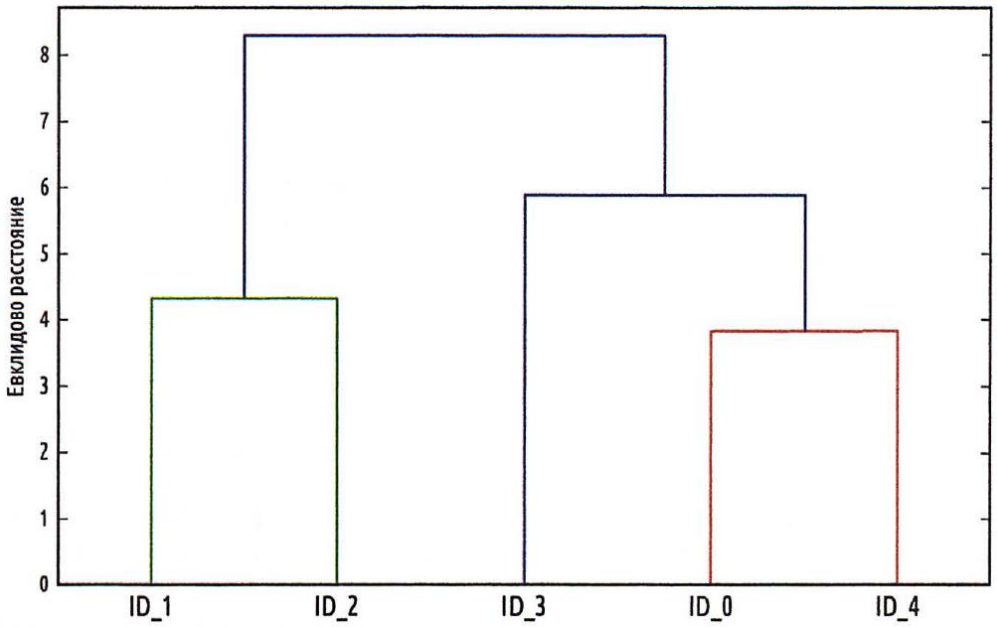


Рисунок 4. Пример использования дендограммы

Такая древовидная диаграмма резюмирует разные кластеры, сформированные при кластеризации. Так из рисунка 4 видно, что, основываясь на евклидовой метрике расстояния, образцы ID_O и ID_4, а затем ID_1и ID 2 являются самыми похожими. Из этого следует, что оптимальное количество кластеров – 3.

1.3.3. Метрика Силуэт

В отличие от описанных выше метрик, данная не предполагает знание истинных меток объектов, а оценивает качество кластеризации, рассматривая результат кластеризации и неразмеченную выборку. Первоначально метрика силуэт задается отдельно для каждого элемента [6]. Данная метрика для элемента множества вычисляется по следующей формуле:

$$s = \frac{b-a}{\max(a,b)}. \quad (7)$$

Где a – среднее расстояние от рассматриваемого элемента до элементов из того же кластера, а b – среднее расстояние от элемента до элементов из близкорасположенного кластера.

Силуэтом выборки выступает величина, соответствующая среднему значению силуэта одного элемента из рассматриваемой выборки, что демонстрирует, в какой мере среднее расстояние до элементов других кластеров отличается от среднего расстояния до элементов собственного кластера. Величины, полученные при реализации данной метрики, варьируются в диапазоне от -1 до 1. Чем ближе найденное значение к 1, тем более точно выделены кластеры (высокая плотность элементов), а значение близкое к -1 соответствует разрозненной кластеризации, если же значение близко к 0, то такие кластеры пересекаются и накладываются друг на друга. Из этого следует, что чем ближе значение силуэта к единице (чем больше силуэт), тем более компактны и плотно сгруппированы кластера, а значит более точно выражены.

Данная метрика позволяет выбрать оптимальное число кластеров – выбирается число кластеров, при котором значение силуэта ближе к единице. В отличие от предыдущих метрик, силуэт напрямую зависит от формы кластеров, и достигает оптимального значения при выпуклых кластерах,

получаемых с помощью алгоритмов, основанных на восстановлении плотности распределения.

1.4. Метод DBSCAN

Метод кластеризации DBSCAN – это плотностный алгоритм интеллектуального анализа данных без учителя, предназначенный для анализа больших данных с присутствием шума. Если иные алгоритмы, основанные на применении плоского разбиения, минимизируют расстояние от элемента до центра кластера и стремятся к созданию кластеров сферической формы, то DBSCAN может формировать кластеры различной формы[6].

Понятие плотности в DBSCAN определяется как число точек внутри указанного радиуса ε. В DBSCAN каждому образцу (точке) назначается специальная метка, при этом используются следующие критерии:

- Точка рассматривается как корневая, если указанное количество окрестных точек(MinPts) попадает в переделы указанного радиуса;
- Граничная точка – это точка, имеющая соседей меньше, чем MinPts в пределах ε, но лежащая в пределах радиуса корневой точки;
- Все остальные точки рассматриваются, как шумовые.

После маркировки точек как корневых, граничных, либо шумовых алгоритм DBSCAN можно представить в двух простых шагах:

1. Сформировать для каждой корневой точки отдельный кластер, либо связную группу корневых точек (корневые точки являются связными, если они расположены не дальше, чем ε);
2. Назначить каждую граничную точку кластеру соответствующей корневой точки.

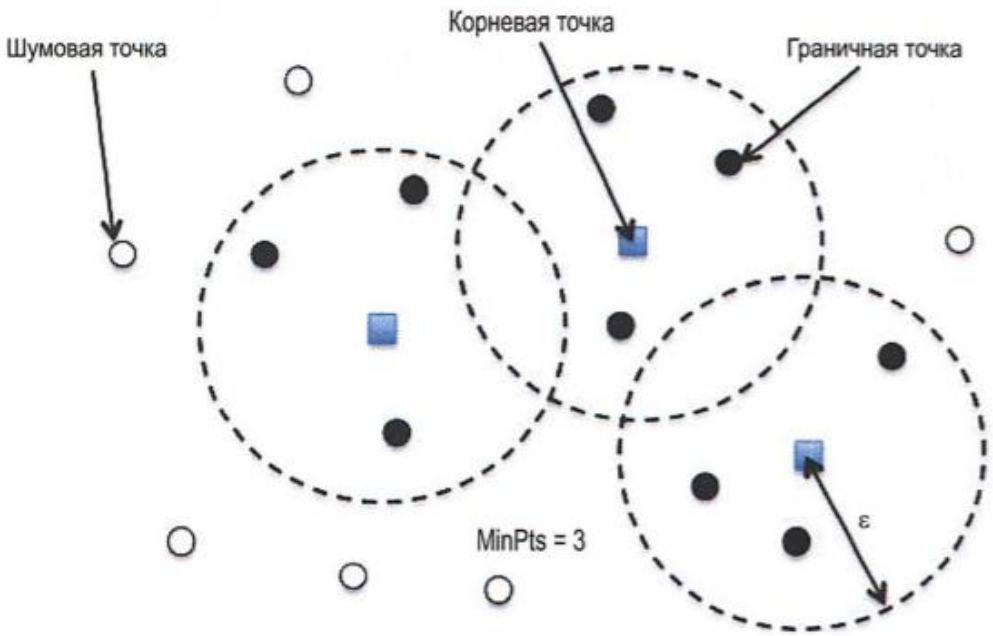


Рисунок 5. Графическое представление метода DBSCAN

Одно из основных преимуществ использования алгоритма DBSCAN состоит в том, что он не делает допущения о сферичной форме кластеров, как в алгоритме К-средних. Кроме того, алгоритм DBSCAN отличается тем, что он с необходимостью не назначает каждую точку кластеру и одновременно способен удалять шумовые точки [1].

1.5. Метод BIRTH

Сбалансированное итеративное сокращение и кластеризация с помощью иерархий (англ. BIRCH – balanced iterative reducing and clustering using hierarchies) – это алгоритм интеллектуального анализа данных без учителя, представляющий из себя иерархическую кластеризацию большого размера данных. Самым главным преимуществом метода выступает динамическая кластеризация данных по мере их поступления с целью получения оптимального результата для имеющихся ресурсов: временных рамок и памяти. Данный алгоритм зачастую требует одного прохода по базе данных [9].

Каждое решение кластеризации BIRCH локально и осуществляется без просмотра всех точек данных и существующих на текущий момент кластеров.

Метод работает на наблюдениях, пространство данных которых обычно не однородно заполнено и не каждая точка данных одинаково важна. Метод позволяет использовать всю доступную память для получения наиболее точных возможных подкластеров при минимизации цены ввода/вывода. Метод является инкрементальным и не требует наличия полного набора данных.

Кластеризация BIRCH предусматривает двухэтапный процесс кластеризации. Данный метод предназначен для кластеризации данных больших размеров. Но, как и ранее рассмотренные алгоритмы, он способен обрабатывать только числовые данные.

К основным достоинствам метода можно отнести:

1. Возможность применения при единственном сканировании входного набора данных;
2. Выделяет область данных с большой плотность, как единый кластер;
3. Двухэтапная кластеризация;
4. Учитывает, что данные могут неодинаково распределены по пространству;
5. Работает на ограниченном объеме памяти;
6. Является локальным алгоритмом;
7. Кластеризация больших объемов данных.

Из главных недостатков можно выделить:

1. Необходимо задавать пороговые значения;
2. Лучше выделяет кластеры сферической формы;
3. Работает лишь с числовыми данными.

Описание алгоритма:

Этап 1. Имеющиеся данные загружаются в память. Первоначальное построение в памяти кластерного дерева по элементам, полученным в результате первичного сканирования набора данных. Все шаги в данном этапе практически не чувствительны к порядку и выполняются в кратчайшее время.

Кластерное дерево строится по следующему алгоритму:

Кластерный объект содержит в себе три числа: количество элементов входных данных, их сумма и сумма квадратов элементов входных данных. Кластерное дерево (CT) представляет собой взвешенное сбалансированное дерево, содержащее два параметра: пороговую величину (a) и коэффициент разветвления (b).

Каждый нелистьевой узел дерева имеет не более чем b вхождений узлов следующей формы: [CT, m], где i = 1, 2, …, m, а m – указатель на i-й дочерний узел.

Каждый листьевой узел ссылается на два соседних узла. Кластер состоящий из элементов листьевого узла должен удовлетворять следующему условию: диаметр или радиус полученного кластера должен быть не более пороговой величины.

Этап 2 (необязательный). При необходимости выполняется сжатие данных (уплотнение). Данные сжимаются до приемлемых размеров при помощи уменьшения и перестроения кластерного дерева с увеличением пороговой величины.

Этап 3. Выполнение глобальной кластеризации. На листьевых компонентах кластерного дерева применяется выбранный алгоритм кластеризации.

Этап 4 (необязательный). При необходимости выполняется улучшение кластеров. Полученные в 3 этапе центра тяжести кластеров

используются как основы и происходит перераспределение данных между схожими т.е. близкими кластерами. Данный этап обеспечивает попадание одинаковых элементов в один кластер.

1.6. Метод Уорда

Данный алгоритм интеллектуального анализа данных ориентирован на группирование близко расположенных друг к другу кластеров. В алгоритме изначально каждому объекту соответствует свой собственный кластер, далее на каждом этапе алгоритма пара кластеров с наименьшим минимальным приращением функции расстояния между ними объединяется в один единый кластер [11].

Метод Уорда основан на применение методов дисперсионного анализа для оценки расстояния между элементами и кластерами, что крайне отличает его от других методов кластеризации. Прирост суммы квадратов расстояний элементов до образованного при объединении центра кластера принимается в качестве расстояния между кластерами. В каждом этапе алгоритма находятся и объединяются два кластера, имеющие минимальное увеличение дисперсии. Метод показывает высокую эффективность на данных с близко расположенными кластерами.

Алгоритм Уорда минимизирует сумму квадратов для любых двух кластеров, которые могут быть сформированы на каждом шаге. Данный метод стремится создавать кластеры маленького размера, но при этом является довольно эффективным.

1.7. Оценка качества кластеризации

Задача оценки качества кластеризации должна соответствовать двум требованиям. Во-первых, такие оценки не должны зависеть от самих значений меток, а только от самого разбиения выборки. Во-вторых, не всегда известны истинные метки объектов, поэтому также нужны оценки, позволяющие оценить качество кластеризации, используя только неразмеченную выборку.

Выделяют внешние и внутренние метрики качества. Внешние используют информацию об истинном разбиении на кластеры, в то время как внутренние метрики не используют никакой внешней информации и оценивают качество кластеризации, основываясь только на наборе данных. Оптимальное число кластеров обычно определяют с использованием внутренних метрик[2].

1. Adjusted Rand Index (ARI)

Предполагается, что известны истинные метки объектов. Данная мера не зависит от самих значений меток, а только от разбиения выборки на кластеры. Пусть a – число объектов в выборке, n – число пар объектов, имеющих одинаковые метки и находящихся в одном кластере, а b – число пар объектов, имеющих различные метки и находящихся в разных кластерах. Тогда формула будет иметь следующий вид:

$$RI = \frac{2(a-b)}{n(n-1)}. \quad (8)$$

То есть это доля объектов, для которых эти разбиения (исходное и полученное в результате кластеризации) "согласованы". Rand Index (RI) выражает схожесть двух разных кластеризаций одной и той же выборки. Чтобы этот индекс давал значения близкие к нулю для случайных кластеризаций при любом n и числе кластеров, необходимо нормировать его. Формула Adjusted Rand Index:

$$ARI = \frac{RI-E[RI]}{\max(RI)-E[RI]}. \quad (9)$$

Эта мера симметрична, не зависит от значений и перестановок меток. Таким образом, данный индекс является мерой расстояния между различными разбиениями выборки. ARI принимает значения в диапазоне [-1;1]. Отрицательные значения соответствуют "независимым" разбиениям на кластеры, значения, близкие к нулю, случайным разбиениям, и

положительные значения говорят о том, что два разбиения схожи (совпадают при ARI = 1).

2. Adjusted Mutual Information (AMI)

Данная мера очень похожа на ARI. Она также симметрична, не зависит от значений и перестановок меток. Определяется с использованием функции энтропии, интерпретируя разбиения выборки, как дискретные распределения (вероятность отнесения к кластеру равна доле объектов в нём). Индекс MI определяется как взаимная информация для двух распределений, соответствующих разбиениям выборки на кластеры. Интуитивно, взаимная информация измеряет долю информации, общей для обоих разбиений: насколько информация об одном из них уменьшает неопределенность относительно другого.

Аналогично ARI определяется индекс AMI, позволяющий избавиться от роста индекса MI с увеличением числа классов. Он принимает значения в диапазоне [0;1]. Значения, близкие к нулю, говорят о независимости разбиений, а близкие к единице – об их схожести (совпадении при AMI = 1).

Глава 2. Применение методов кластеризации

2.1. Обзор инструментальных средств

Для анализа данных и построения моделей машинного обучения использовался язык программирования Python 3.6. В качестве среды разработки использовалась интерактивная веб оболочка для языка Python – Jupyter Notebook. Данная среда позволяет объединять код, текст и диаграммы в одном файле, и распространять их для других пользователей, что делает ее незаменимым инструментом для выполнения работ, требующих постоянного обсуждения результатов участниками проекта.

Помимо стандартных библиотек языка Python, были использованы такие библиотеки как: Pandas, NumPy, MatPlotLib, Sklearn, FastParquet, SciPy и DateTime. Библиотека NumPy предназначена для работы с большими многомерными массивами данных, а также содержит большое количество математических функция для операций над ними. Pandas предоставляет специальные структуры данных, такие как DataFrame и Series, также операции для манипулирования числовыми таблицами и временными рядами. Библиотека SciPy содержит математические функции для обработки данных, не реализованные в NumPy. Библиотека Sklearn содержит большое количество различных алгоритмов машинного обучения, в том числе методы кластеризации и метрики их оценки. Также были использованы библиотека DateTime, которая предназначена для работы со временным типом данных. Библиотека MatPlotLib служит для работы с двумерными графиками [10].

2.2. Описание предметной области

Работа направлена на сравнение методов кластеризации и поиска наиболее эффективной технологии анализа описания вычислительных задач по обработке данных эксперимента ATLAS. Целью данный работы является применение методов машинного обучения для анализа описания вычислительных задач по обработке данных эксперимента ATLAS. Данная

система отвечает за обработку данных для групп ученых-физиков, а также за предварительную обработку и анализ данных.

Работа состоит из нескольких последовательных этапов:

1. Первичная обработка данных эксперимента
2. Применение различных методов кластеризации к обработанному набору данных
3. Выявление наиболее эффективного метода

2.3. Первичная обработка входных данных

Данные эксперимента, с которыми будет проводится данное исследование, представлены в файле разрешением «.csv» и описывают вычислительные задачи по обработке данных. Файл представляет собой набор данных и состоит не только из числовых параметров, поэтому для последующей кластеризации необходимо извлечь только числовые значения [8].

Извлечение требуемых значений состояло из следующих шагов:

1. Открытие и чтения файла с данными
2. Приведение данных к табличному виду
3. Поиск числовых столбцов
4. Удаление всех нечисловых столбцов
5. Замена элементов с пустыми значениями на 0.
6. Приведение данных к одной шкале

После выполнения всех этих шагов данные готовые для дальнейшей работы с ними – кластеризации. В первом этапе кластеризации будут подвергнуты данные по первым двум параметрам, а на следующем кластеризация будет проведена по всем имеющимся параметрам.

На рисунке 5 представлены первые 10000 выборок полученных данных по первым двум признакам, с которыми в последствии и будут реализованы методы кластеризации.

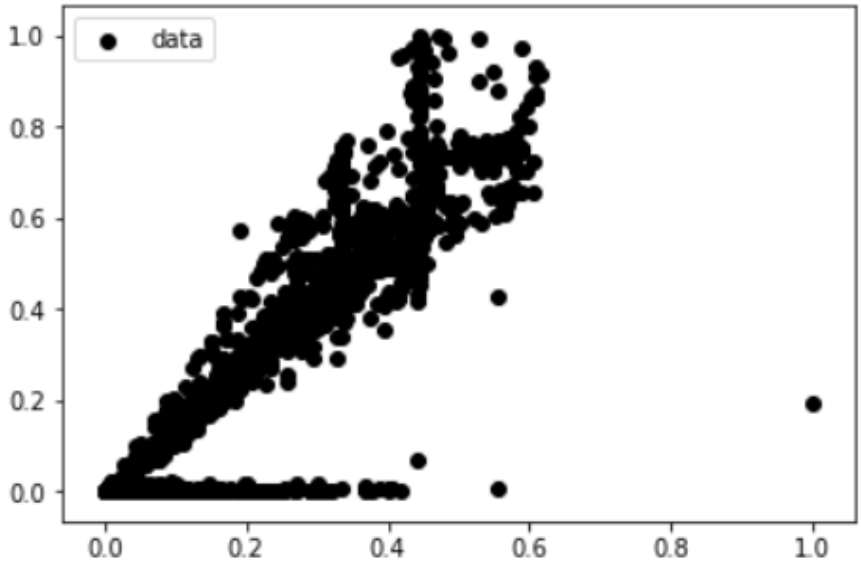


Рисунок 6. Первые 10000 выборок из файла

Реализация данной обработки представлена в приложении А на языке программирования Python.

Также при помощи алгоритмов снижения размерности был построен график для всех параметров (рис. 7).

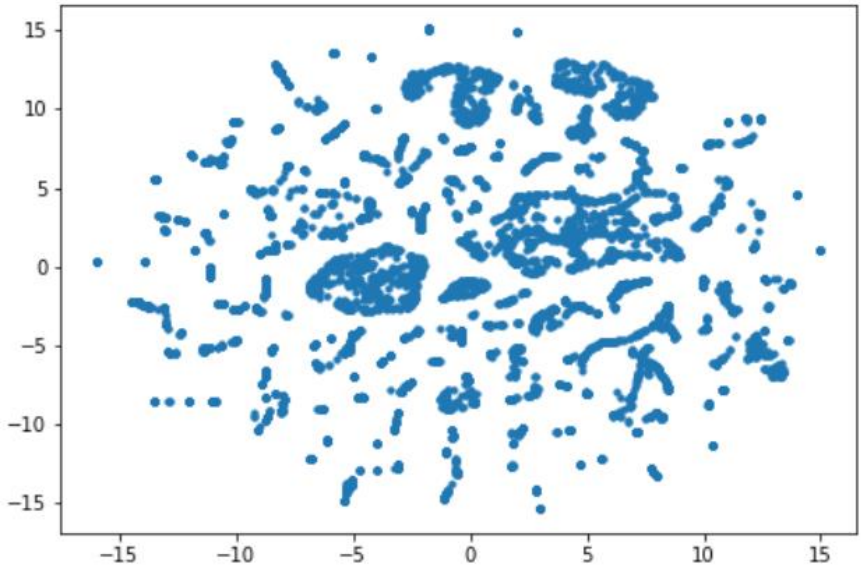


Рисунок 7. Представление всех данных в графическом виде

2.4. Метод локтя

Метод «локтя» рассматривает характер изменения разброса с увеличением числа групп. Объединив все объекты в одну группу, мы имеем наибольшую внутрикластерную дисперсию. На каком-то этапе можно усмотреть, что снижение этой дисперсии замедляется – на графике это происходит в точке, называемой “локтем” (родственник “каменистой осыпи” для анализа главных компонент).

Ниже на рисунке 8 представлены результаты применения данного метода на исходных данных.

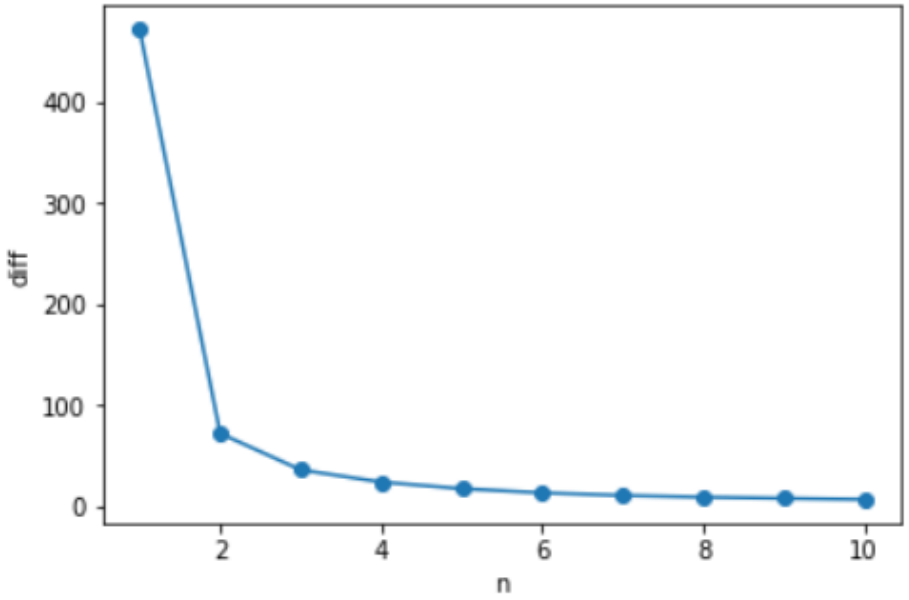


Рисунок 8. Применение метода локтя для 2-х параметров

На вертикальной оси отображается искажение, а вертикальная отображает число кластеров. Из рисунка видно, что оптимальное число кластеров для данного набора будет 3. Такое количество кластеров и будет применено в последующих методах кластеризации. Применение данного метода отображено в приложении Б.

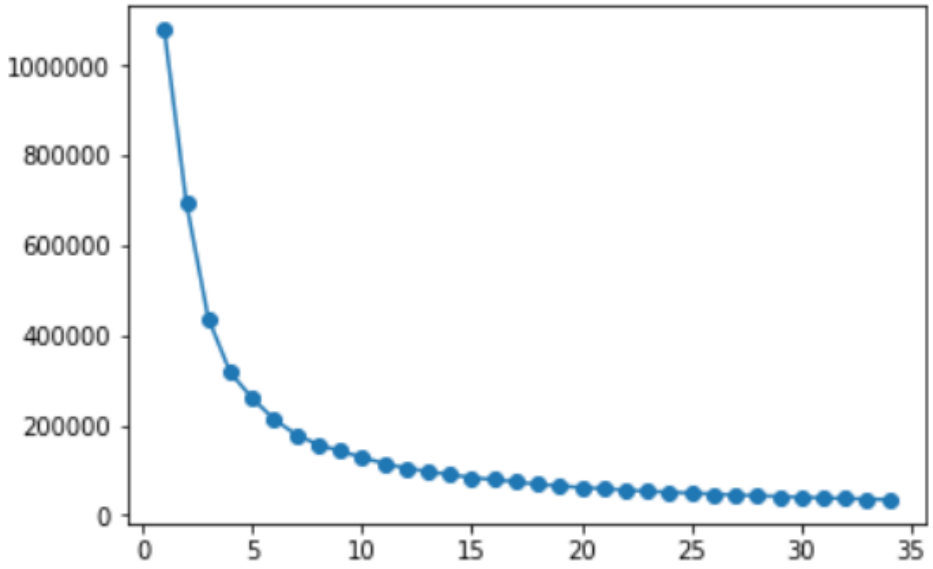


Рисунок 9. Применение метода локтя на всех данных

На рисунке 9 изображен график применения метода для всех имеющихся параметров. В сравнении с 8 рисунком, количество кластеров увеличилось в несколько раз. Оптимальное количество кластеров – 14.

2.5. Метод Силуэт

Также количество кластеров можно определить при помощи коэффициента Силуэт.

Высший балл по коэффициенту Силуэт относится к модели с определенными кластерами. Коэффициент силуэта определяется для каждого образца и состоит из двух партитур:

- среднее расстояние между образцом и всеми другими точками в одном классе.
- среднее расстояние между образцом и всеми другими точками в следующем ближайший кластер.

Ниже на рисунке изображен вывод на консоль значение коэффициента Силуэта для всех имеющихся данных.

```
For n_clusters=2, The Silhouette Coefficient is 0.024660585448145866
For n_clusters=3, The Silhouette Coefficient is 0.027399791404604912
For n_clusters=4, The Silhouette Coefficient is 0.0004715860995929688
For n_clusters=5, The Silhouette Coefficient is -0.042945899069309235
For n_clusters=6, The Silhouette Coefficient is -0.04548322781920433
For n_clusters=7, The Silhouette Coefficient is -0.030850326642394066
For n_clusters=8, The Silhouette Coefficient is -0.062292225658893585
For n_clusters=9, The Silhouette Coefficient is -0.04902235418558121
For n_clusters=10, The Silhouette Coefficient is 0.0019855822902172804
For n_clusters=11, The Silhouette Coefficient is -0.010138427838683128
For n_clusters=12, The Silhouette Coefficient is 0.000409809552365914
For n_clusters=13, The Silhouette Coefficient is -0.04149959608912468
For n_clusters=14, The Silhouette Coefficient is 0.017118748277425766
For n_clusters=15, The Silhouette Coefficient is -0.029498156160116196
For n_clusters=16, The Silhouette Coefficient is -0.05043131113052368
For n_clusters=17, The Silhouette Coefficient is -0.03703857958316803
For n_clusters=18, The Silhouette Coefficient is -0.049699265509843826
```

Рисунок 10. Вывод программы при использовании Силуэта

Количество оптимальное количество кластеров выбирается таким образом, чтобы значение коэффициента силуэта было максимально приближено к 1. Из десятого рисунка видно, что наибольшее значение коэффициента принимает при количестве кластеров равным 14.

2.6. Организация кластеров в виде иерархичного дерева

Иерархическая кластеризация представляет собой алгоритм, который строит иерархию кластеров. Данный алгоритм начинает работу с того, что каждому экземпляру данных сопоставляется свой собственный кластер. Затем два ближайших кластера объединяются в один и так далее, пока не будет образован один общий кластер.

Результат иерархической кластеризации может быть представлен с помощью дендрограммы. Рассмотрим этот тип кластеризации на примере первых двух параметров данных на рисунке 11.

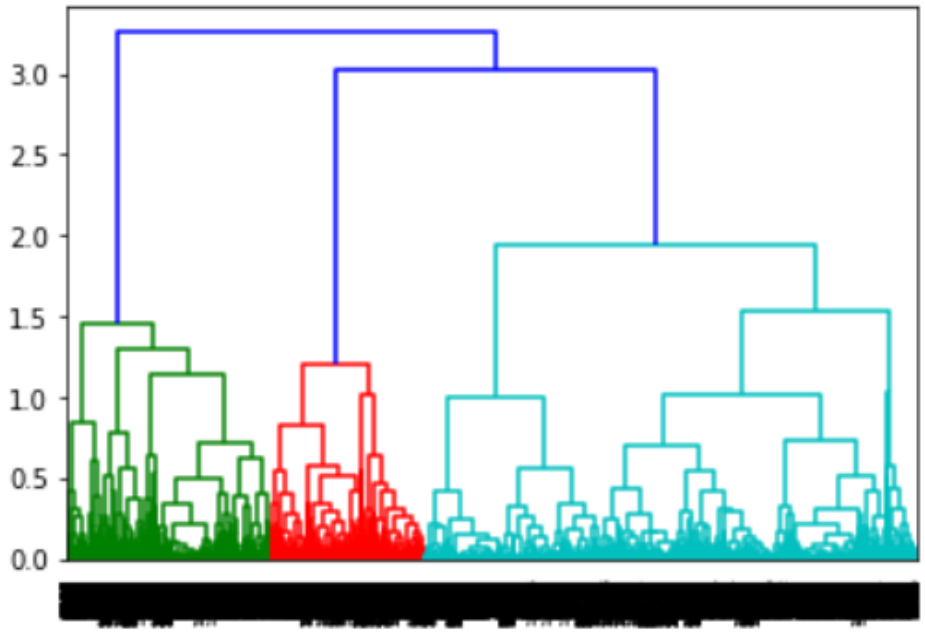


Рисунок 11. Применение дендограммы

Можно видеть, что в результате иерархической кластеризации данных естественным образом произошло разбиение на три кластера, обозначенных на рисунке различным цветом. При этом исходно число кластеров не задавалось.

2.7. Метод К-средних

Применение алгоритма К-средних предполагает ввод числа кластеров. А это значит, что данных параметр нужно определить заранее. В предыдущем пункте, с помощью метода локтя было определено оптимальное число кластеров: для первых двух параметров данных – 3, для всех данных – 14.

Алгоритм К-средних можно состоит из четырех этапов:

1. Случайно выбрать из точек образцов к центроидов, как исходных центров кластеров
2. Назначить каждый образец самому ближайшему к ней центроиду
3. Переместить каждый центроид в центр образцов, которые были ему назначены

Повторять шаги 2 и 3, пока назначения кластеров не перестанут изменяться, либо не будет достигнут заданный пользователями допуск или максимальное число итераций.

На рисунке 12 продемонстрирован график данной кластеризации для имеющихся данных по первым двум признакам.

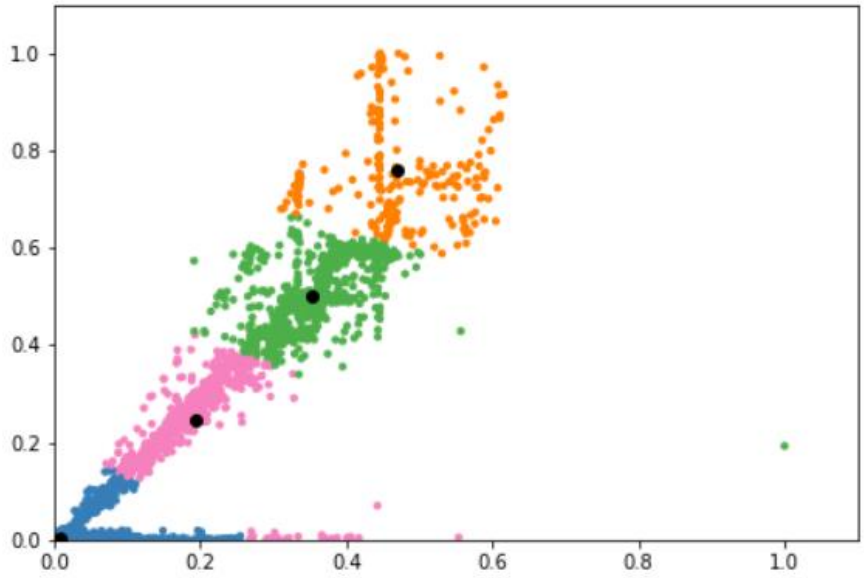


Рисунок 12. Кластеризация методом К-средних

Черным цветов обозначены центроиды кластеров. Из рисунка видно, что было определено три кластера, каждый кластер выделен различным цветом. Данная кластеризация производилась лишь по первым двум параметрам, как и все последующие. Код данного метода реализованный на языке программирование python представлен в приложении В.

Распределение данных всех имеющихся параметров по кластерам с помощью рассматриваемого метода представлено на рисунке 13. Количество кластеров изначально было определено с помощью метрик Силуэта и Локтя: обе метрики достигали наилучшего результата при количестве кластеров равном 14.

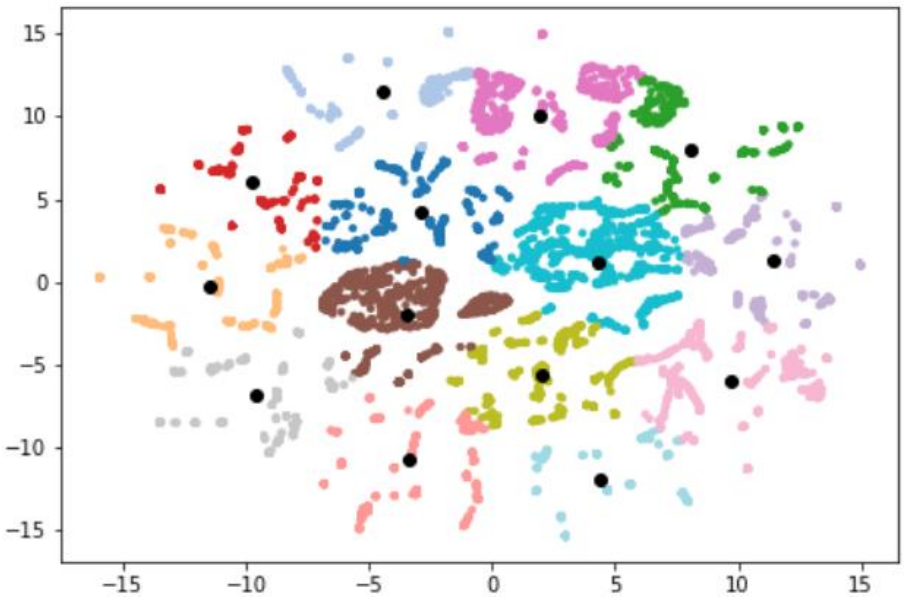


Рисунок 13. Результаты кластеризации К-средних

2.8. Метод DBSCAN

Данный метод, в отличие от предыдущего, не предполагает знание количество кластеров, но в свою очередь крайне чувствителен к выбору значения радиуса ε и числу точек внутри указанного радиуса ε. Данные параметры необходимо оптимизировать с целью получения наилучшего результата. На рисунке 14 представлен график данного метода с выбором радиуса равного 0,26 и числу точек в радиусе равному 1.

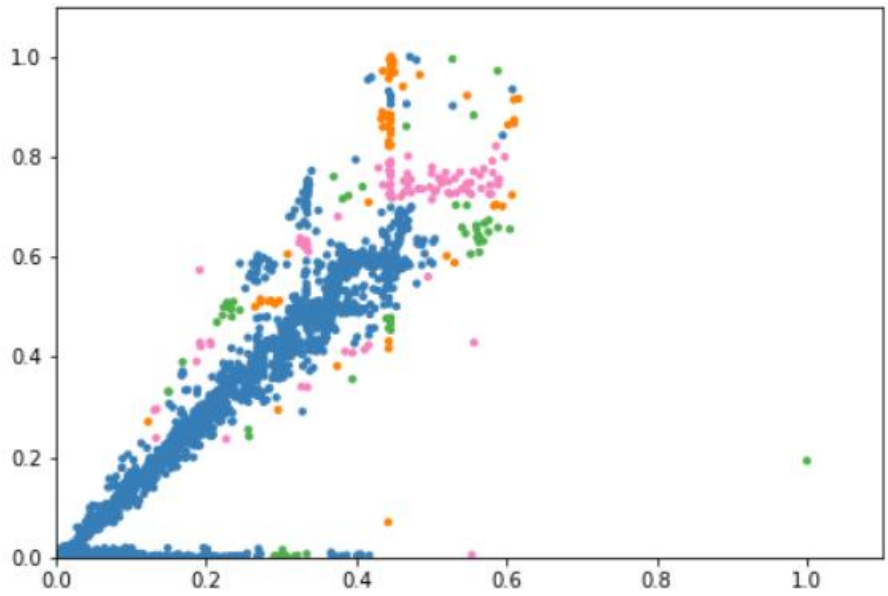


Рисунок 14. Применение метода DBSCAN

Из рисунка видно, что было определено три кластера, каждый кластер выделен различным цветом. Самый обширный кластер значений выделен синим цветом, а наименьший зеленым.

Так как данный метод крайне чувствителен к выбору параметров радиуса и числа точек, то его успешность его результата зависит только от правильности ручного подбора, что усложняет работу с данным методом.

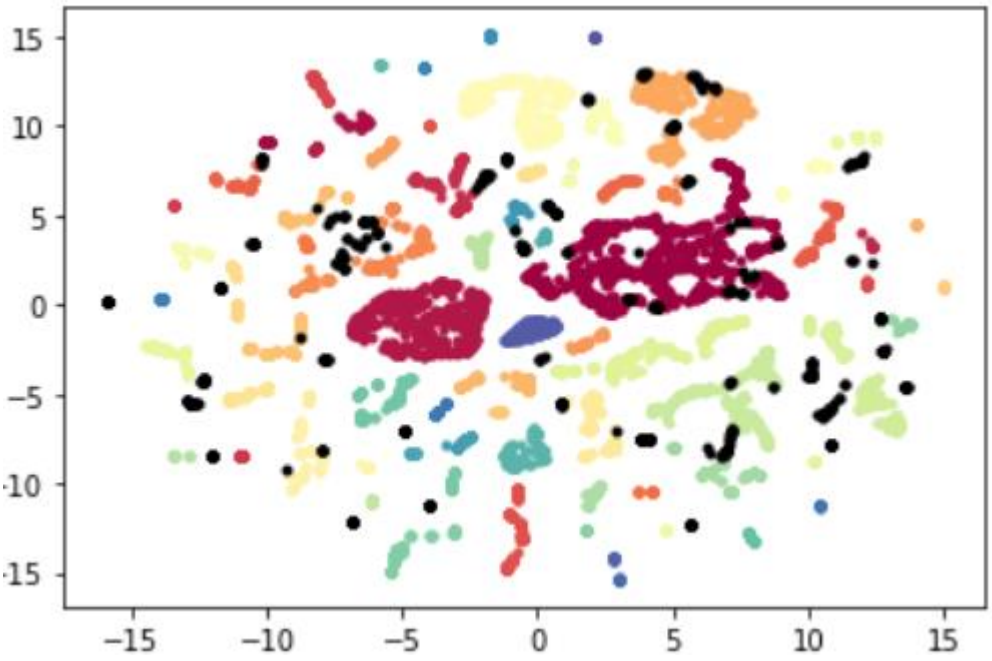


Рисунок 15. Кластеризация DBSCAN

На рисунке 15 показана кластеризация данным методом по всем параметрам. Точки черного цвета соответствуют шумовым точкам – не входящим в состав кластера. Результат кластеризации напрямую зависит от выбора плотности и минимальных точек вхождения, поэтому крайне важно тщательно подбирать каждый из этих параметров.

Код данного метода реализованный на языке программирование python представлен в приложении Г.

2.9. Метод BIRTH

Алгоритм BIRCH берёт в качестве входа набор из N точек данных, представленный как вещественные вектора, и желаемое число кластеров K. Алгоритм разбит на четыре фазы, вторая из которых не обязательна.

Первая фаза строит CF дерево точек данных, высоко сбалансированную древесную структуру.

На втором шаге алгоритм просматривает все листья в начальном CF-дереве, чтобы построить меньшее CF-дерево путём удаления выпадений и группирования переполненных подклассов в большие подклассы. Этот шаг в исходном представлении BIRCH помечен как необязательный.

На третьем шаге используется существующий алгоритм для кластеризации всех листов. После этого шага получаем набор кластеров, которые содержат главные схемы распределения в данных. Однако могут существовать небольшие локальные неточности, которые могут быть обработаны необязательным шагом 4. На шаге 4 центры тяжести кластеров, полученных на шаге 3, используются как зародыши и точки перераспределения точек данных для получения нового набора кластеров. Шаг 4 обеспечивает также возможность отбрасывания выбросов. То есть точка, которая слишком далека от ближайшего зародыша, может считаться выбросом.

На рисунке 16 представлен результат работы данного алгоритма.

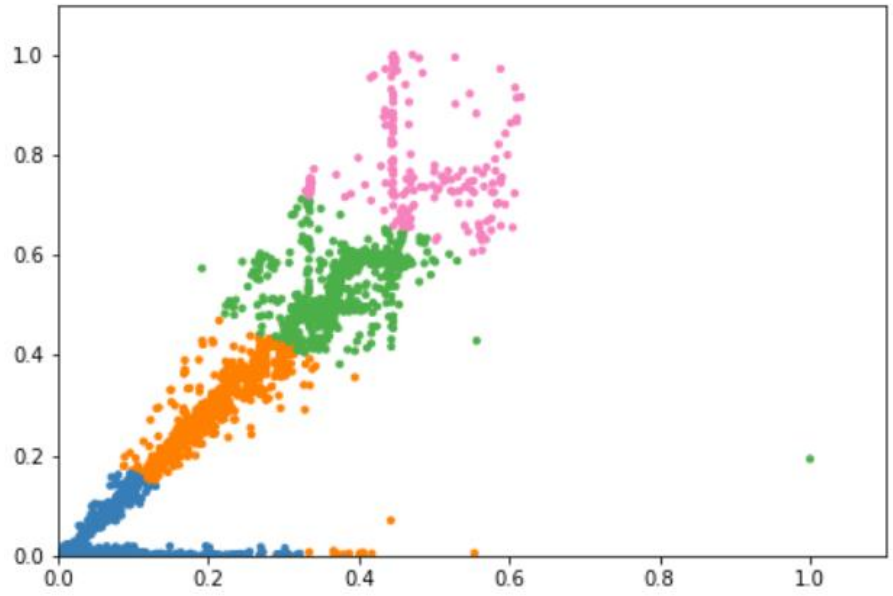


Рисунок 16. Применение метода BIRCH

С помощью алгоритма были найдены и окрашены 4 различных кластера. Код данного алгоритма описан в приложении Д.

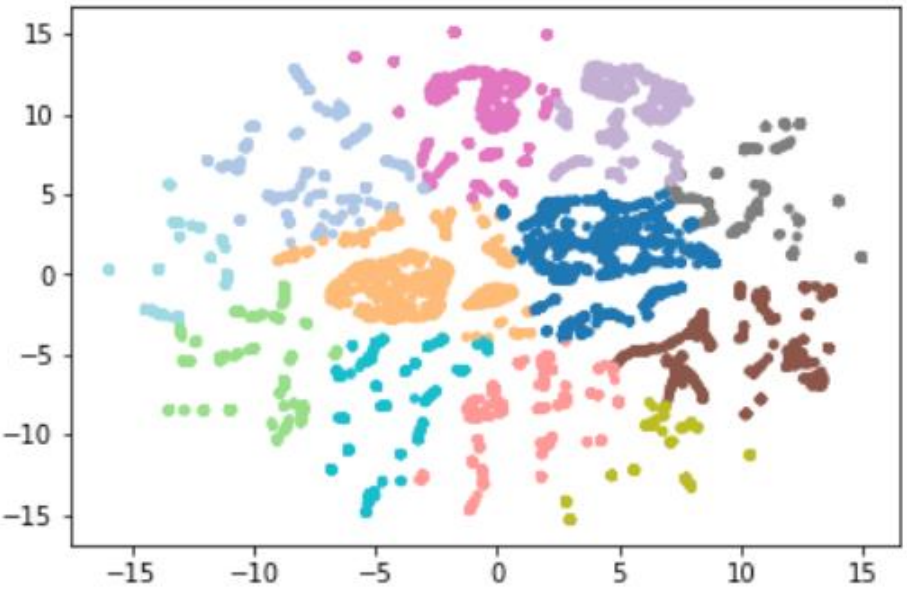


Рисунок 17. Применение метода BIRCH

Для всех имеющихся данных метод распределил 12 кластеров, графическое представление изображено на рисунке 17. Данный метод в кратчайшие сроки за один проход по данным распределил все элементы по кластерам.

2.10. Метод Уорда

В отличие от других методов кластерного анализа, для оценки расстояний между кластерами здесь используются методы дисперсионного анализа. В качестве расстояния между кластерами берётся прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению дисперсии. Этот метод применяется для задач с близко расположенными кластерам.

Метод минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера. В данном методе, как и в предыдущем, не требуется вводить количество кластеров.

Ниже приведен график метода кластеризации Уорда полученный по первым двум параметрам имеющихся данных.

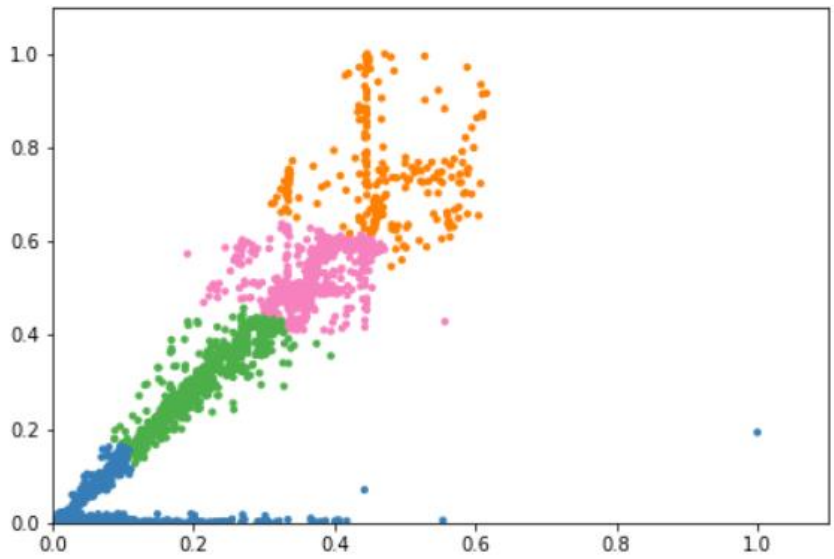


Рисунок 18. Применение метода Уорда.

Из рисунка 18 видно, что было определено четыре кластера, каждый кластер выделен различным цветом. Реализация данного метода на языке программирования python представлена приложении Е.

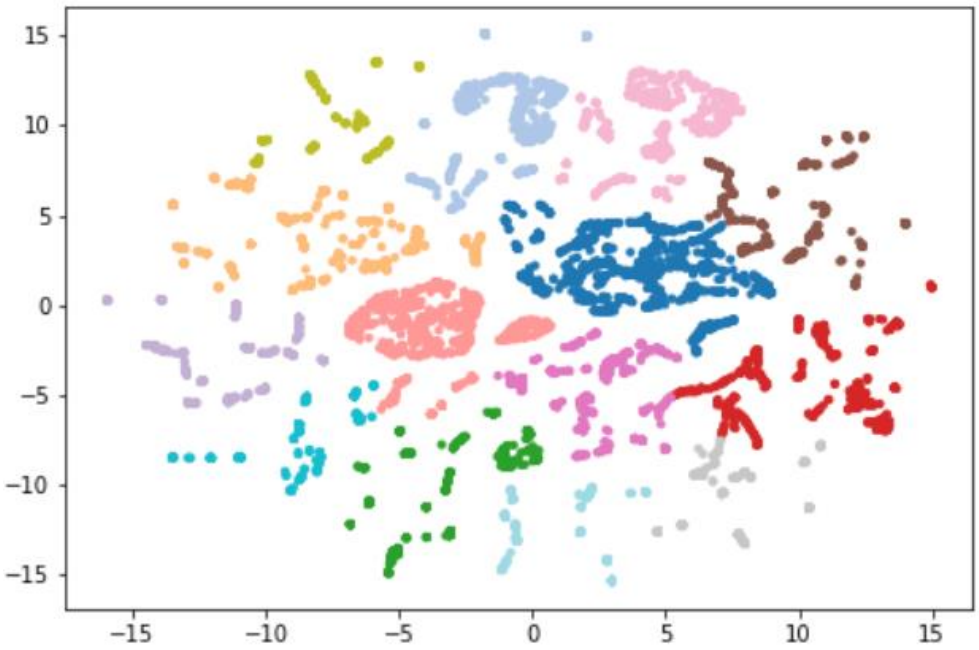


Рисунок 19. Применение метода Уорда.

При применении данного алгоритма ко всем имеющимся данным были сформированы 14 кластеров. Данное распределение представлено на рисунке 19.

Сравнивая визуально графики всех реализованных методов, можно сделать вывод, что кластеризация методом Уорда наиболее эффективна для имеющихся данных, но распределение по кластерам происходит за более длительный промежуток времени, чем выше рассмотренные.

Глава 3. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Экономическое обоснование необходимо при осуществлении любой научно-исследовательской деятельности. Оно позволяет определить перспективность исследования, оценить его сильные и слабые стороны, выявить уровень востребованности и конкурентоспособности.

Целью данного раздела является конструирование и внедрение конкурентоспособного проекта, соответствовать современным потребностям в сферах ресурсосбережения и ресурсоэффективности.

В ходе выполнения данного раздела будут определены стратегии реализации и развития проекта, рассчитана себестоимость, проведено планирование научно-исследовательской работы, оценен коммерческий потенциал и перспективы.

3.1. Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

Данный раздел является крайне важной частью работы, так как содержит в себе анализ экономической эффективности проекта. Главной целью данного раздела является экономическое обоснование разработки модулей, а также определение и расчет денежных и трудовых затрат на их создание.

3.1.1. Потенциальные потребители результатов исследования

Суть данного проекта заключается в сравнении различных методов кластеризации на экспериментальных данных проекта «ATLAS» для поиска наиболее эффективного метода, который позволит за минимальный период времени обработать большой объем данных и подготовить их для дальнейшего анализа.

Результаты данного исследования являются уникальными для данных проекта «ATLAS» и будут иметь другие значения для других организаций, поэтому продажа реализации данных методов не рассматривается.

3.1.2. Технология QuaD

Технология QuaD (Quality Advisor) предназначена для количественной оценки качественных характеристик, описывающих качество и перспективность разработки, а также для помощи в принятии решения о целесообразности вкладывания денежных средств в научно-исследовательскую работу [13]. В таблице 1 представлена эта технология относительно рассматриваемого проекта.

Таблица 1 – Оценочная карта сравнения конкурентных технических решений (разработок)

Критерии оценки	Вес критерия	Баллы	Макс. балл	Относительное значение (3/4)	Средне-взвешенное значение (5*2)
1	2	3	4	5	6
Показатели оценки качества разработки					
Простота эксплуатации	0,06	70	100	0.7	0.042
Повышение производительности	0,09	95	100	0.95	0.0855
Потребность в вычислительных мощностях	0,09	90	100	0.9	0.081
Скорость обработки данных	0,094	100	100	1	0.094
Использование ресурсов	0,084	95	100	0.95	0.0798
Надежность	0,094	85	100	0.85	0.0799
Перспективность внедрения	0,09	80	100	0.8	0.072
Эффективность методов	0,094	97	100	0.97	0.09118
Унифицированность	0,06	45	100	0.45	0.027
Безопасность	0,094	75	100	0.75	0.0705
Долговечность	0,08	75	100	0.75	0.06
Тех. поддержка после внедрения	0,07	87	100	0.87	0.0609

Оценка качества и перспективности по технологии QuaD определяется по формуле:

$$П_{ср} = \sum В_i * Б_i, \quad (10)$$

Где $П_{ср}$ – средневзвешенное значение показателя качества и перспективности научной разработки;

$В_i$ – вес показателя (в долях единицы);

$Б_i$ – средневзвешенное значение i-го показателя.

Таким образом, $П_{ср}$ = 84%, из чего можно сделать вывод, что разработка является перспективной.

3.1.3. SWOT-анализ

SWOT-анализ заключается в исследовании внешней и внутренней среды проекта и показывает его слабые и сильные стороны [12].

Таблица 2 – SWOT-анализ

	Сильные стороны научно-исследовательского проекта: С1. Сокращение времени обработки данных. С2. Простота использования. С3. Гарантия получения результата обработки данных. С4. Увеличение эффективности работы с данными больших размеров. С5. Техническая поддержка научной разработки после его внедрения	**Слабые стороны научно-исследовательского проекта:** Сл1.Узкая направленность. Сл2. Ошибки при обработке информации. Сл3. Потребность в больших вычислительных мощностях.
Возможности: В1. Стойкий спрос инициаторов проекта на разработку В2. Спрос на дальнейшее развитие проекта		
Угрозы: У1. Нехватка вычислительных ресурсов. У2. Отсутствие возможности расширения функционала.		

Таблица 3 – Интерактивная матрица проекта

		Сильные стороны проекта					Слабые стороны проекта		
		С1	С2	С3	С4	С5	Сл1	Сл2	Сл3
Возможности проекта	В1	+	+	+	+	+	-	+	0
	В2	+	+	0	+	+	-	+	0
Угрозы проекта	У1	+	-	0	0	-	-	-	+
	У2	-	0	-	-	+	0	0	+

Таблица 4 – Итоговая матрица SWOT-анализа

	Сильные стороны научно-исследовательского проекта: С1. Сокращение времени обработки данных. С2. Простота использования. С3. Гарантия получения результата обработки данных. С4. Увеличение эффективности работы с данными больших размеров. С5. Техническая поддержка научной разработки после его внедрения.	**Слабые стороны научно-исследовательского проекта:** Сл1.Узкая направленность. Сл2. Ошибки при обработке информации. Сл3. Потребность в больших вычислительных мощностях.
Возможности: В1. Стойкий спрос инициаторов проекта на разработку. В2. Спрос на дальнейшее развитие проекта.	**Направления развития:** В1С1С2С3С4С5 – сильные стороны проекта способствуют стойкому спросу инициаторов проекта на данную разработку. В1С1С2С4С5 – дальнейшее улучшение и развитие системы положительно влияет на дальнейший спрос на продукт.	**Сдерживающие факторы:** В1Сл2 и В2Сл2 – ошибки при обработке данных влекут за собой последующие ошибки при работе с информацией, что искажает работу с данными, а это снижает спрос на разработку проекта и ее дальнейшее развитие.
Угрозы: У1. Нехватка вычислительных ресурсов. У2. Отсутствие возможности расширения функционала.	**Угрозы развития:** У1С1 – нехватка вычислительных ресурсов увеличивает время обработки данных. У2С5 – при отсутствии возможности расширения функционала дальнейшая техническая поддержка проекта будет неактуальна.	**Уязвимости:** У1Сл3 – нехватка вычислительных ресурсов увеличивает потребность в больших вычислительных мощностях. У2Сл3 – при маленьких вычислительных ресурсах невозможно расширение функционала, которое требует большие ресурсы.

В процессе проведения SWOT-анализа были обозначены и подробно рассмотрены сильные и слабые стороны данного проекта, что позволило увидеть угрозы и уязвимости. Благодаря этому появилась возможность для планирования требуемых изменений, которые позволят максимально минимизировать слабые стороны.

3.2. Планирование научно-исследовательских работ

3.2.1. Структура работ в рамках научного исследования

Для реализации конкретного проекта необходимо определить и согласовать способ организации занятости каждого из участников, что достигается с помощью построения линейного графика работ.

В таблице 5 представлена загруженность исполнителей проекта на всех этапах выполнения проекта. Для реализации данной работы необходимы два участника разработки: научный руководитель (НР) и студент (С).

Таблица 5 – Перечень этапов, работ и распределение их исполнителей.

Основные этапы	**№ раб.**	**Содержание работы**	**Должность исполнителя**
Разработка задания на НИР	1	Постановка целей и задач, получение исходных данных	НР, С
Выбор направления исследований	2	Подбор и изучение материалов по тематике	НР, С
	3	Выбор методов кластеризации	НР, С
	4	Разработка календарного плана	НР, С
	5	Обсуждение литературы	НР, С
	6	Проведение анализа предметной области	С
Проведение ОКР			
Экспериментальные исследования	7	Выбор технических средств	С
	8	Предварительная обработка данных	С
	9	Применение методов кластеризации	С
Обобщения и оценка результатов	10	Анализ результатов: выбор наиболее эффективного метода	С
Оформление отчета по НИР	11	Оформление расчетно-пояснительной записки	С

3.2.2. Определение трудоемкости выполнения работ

Расчет трудоемкости выполнения работ является важным этапом, так как трудовые затраты в большинстве случаях образуют основную часть стоимости разработки. Данный расчет был определен опытно-статистическим экспертным методом.

Для расчета среднего значения трудоемкости $t_{\text{ож}\,i}$ применяется следующая формула:

$$t_{\text{ож}i} = \frac{3t_{\min i}+2t_{\max i}}{5}, \quad (11)$$

Где $t_{\text{ож}i}$ – ожидаемая трудоемкость выполнения i-ой работы чел.-дн.;

$t_{\min i}$ – минимально возможная трудоемкость выполнения заданной i-ой работы, чел.-дн.;

$t_{max\,i}$ – максимально возможная трудоемкость выполнения заданной i-ой работы, чел.-дн.

Ниже приведен пример, по которому рассчитываются значения ожидаемой трудоемкости для каждого этапа.

$$t_{\text{ож1}} = \frac{3 * 3 + 2 * 5}{5} = 3{,}8 \text{ дней.}$$

Продолжительность каждой работы в рабочих днях рассчитывается исходя из ожидаемой трудоемкости работ по следующей формуле:

$$T_{p_i} = \frac{t_{\text{ож}i}}{\text{Ч}_i}, \quad (12)$$

Где T_{pi} – продолжительность одной работы, раб. дн.;

$t_{\text{ож}\,i}$ – ожидаемая трудоемкость выполнения одной работы чел.-дн.;

Ч_i – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

$$T_{p_1} = \frac{3{,}8}{2} = 1{,}9$$

T_{p_1} равен 3,8 дней, оставшиеся значения рассчитаны аналогично.

3.2.3. Разработка графика проведения научного исследования

Для построения графика выполнения научно-исследовательских работ необходимо перевести длительность этапов из рабочих дней в календарные. Для этих расчетов следует воспользоваться формулой:

$$T_{к_i} = T_{p_i} * k_{кал}, \quad (13)$$

Где $T_{к_i}$ – продолжительность выполнения *i*-ой работы в календарных днях;

T_{p_i} – продолжительность выполнения *i*-ой работы в рабочих днях;

$k_{кал}$ – коэффициент календарности.

Для расчета коэффициента календарности необходимо использовать следующую формулу:

$$k_к = \frac{Т_{кал}}{Т_{кал} - Т_{вых} - Т_{пр}}, \quad (14)$$

Где $Т_{кал}$ – количество календарных дней в году ($Т_{кал}$ = 365 дней);

$Т_{вых}$ – количество выходных дней в году ($Т_{вых}$ = 52 дня);

$Т_{пр}$ – количество праздничных дней в году ($Т_{пр}$ = 14 дней).

Коэффициент календарности $Т_{кал}$ равен 1,221. Продолжительность первого этапа в календарных днях $T_{к_1}$ = 4,64 ≈ 5 дней, оставшиеся значения рассчитаны аналогично.

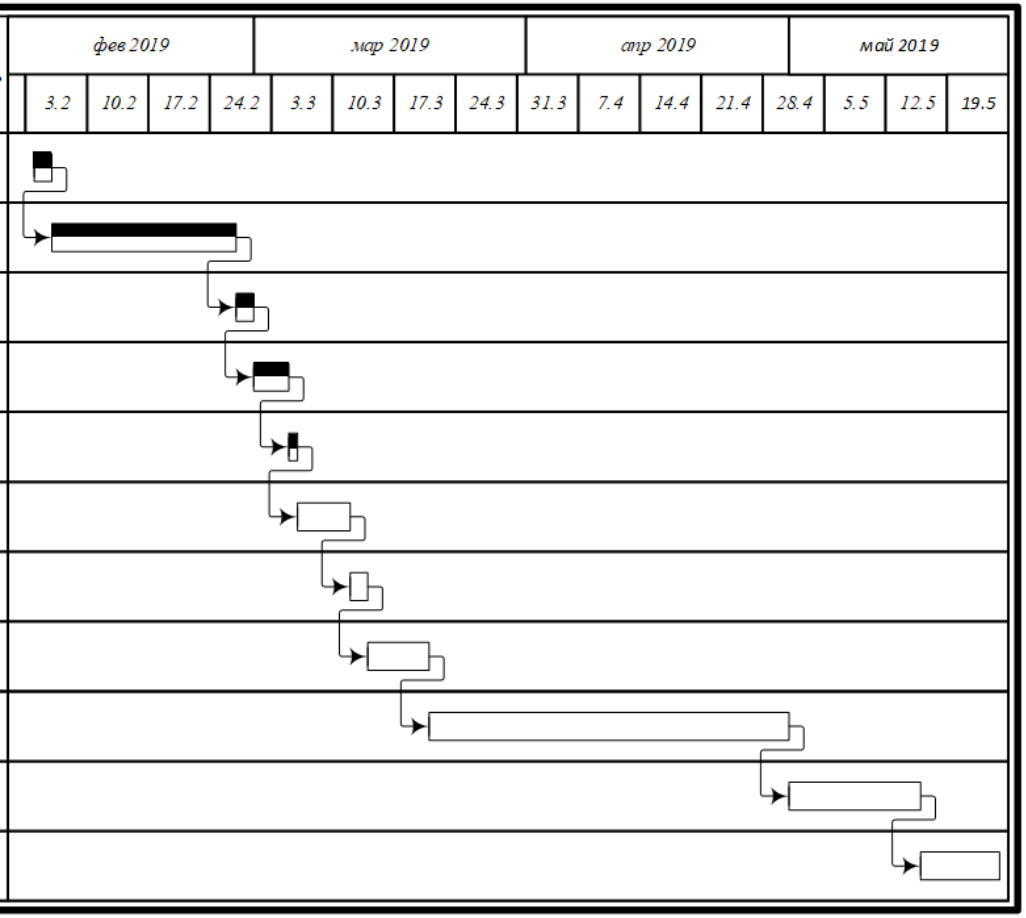
Все рассчитанные значения по трудозатратам представлены в таблице 6. На основе рассчитанных значений из таблицы 6 строится календарный график-план, представленный в таблице 7.

Таблица 6 – Трудозатраты на выполнение проекта

Название работы	**Трудоемкость работ**			**Исполнители**	**Длительность работ в рабочих днях**	**Длительность работ в календарных днях**
	t_{min}, чел-дни	t_{max}, чел-дни	$t_{ож\,i}$, чел-дни			
Постановка целей и задач, получение исходных данных	3	5	3,8	НР, С	2	2
Подбор и изучение материалов по тематике	30	32	30,8	НР, С	15	19
Выбор методов кластеризации	3	5	3,8	НР, С	2	2
Разработка календарного плана	3	6	4.2	НР, С	2	3
Обсуждение литературы	1	3	1,8	НР, С	1	1
Проведение анализа предметной области	3	6	4,2	С	4	5
Выбор технических средств	1	3	1,8	С	2	2
Предварительная обработка данных	4	7	5.2	С	5	6
Применение методов кластеризации	25	35	29	С	29	35
Анализ результатов: выбор наиболее эффективного метода	8	15	10.8	С	11	13
Оформление расчетно-пояснительной записки	5	10	7	С	7	9

Таблица 7 – Календарный план-график работ

ИД	Название задачи	Исполнители	Длительность
1	Постановка целей и задач, получение исходных данных	НР, С	2д
2	Подбор и изучение материалов по тематике	НР, С	15д
3	Выбор методов кластеризации	НР, С	2д
4	Разработка календарного плана	НР, С	2д
5	Обсуждение литературы	НР, С	1д
6	Проведение анализа предметной области	С	4д
7	Выбор технических средств	С	2д
8	Предварительная обработка данных	С	5д
9	Применение методов кластеризации	С	29д
10	Анализ результатов: выбор наиболее эффективного метода	С	11д
11	Оформление расчетно-пояснительной записки	С	7д



3.3. Бюджет научно-технического исследования (НТИ)

Для точного подсчета и планирования бюджета исследования необходимо провести полный анализ всех видов расходов, которые связаны с его выполнением. Во время формирования бюджета исследования следует использовать данные группировки затрат по статьям:

- материальные затраты НТИ;
- затраты на специальное оборудование для научных работ;
- основная заработная плата исполнителей темы;
- дополнительная заработная плата исполнителей темы;
- отчисления во внебюджетные фонды (страховые отчисления);
- накладные расходы.

3.3.1. Расчет материальных затрат НТИ

Данная статья состоит из стоимости всех материалов, которые используются при разработке. Для расчета материальных затрат необходимо использовать следующую формулу [13]:

$$З_м = (1 + k_T) * \sum_{i=1}^{m} Ц_i * N_{расх\,i}, \quad (15)$$

Где m – количество видов материальных ресурсов;

$N_{расх\,i}$ – количество материальных ресурсов i-го вида, планируемых к использованию при выполнении научного исследования;

$Ц_i$ – цена приобретения единицы i-го вида потребляемых материальных ресурсов;

k_T – коэффициент, учитывающих транспортно-заготовительные расходы (15%).

Материальные затраты, необходимы для данной разработки, представлены в таблице 8.

Так как при выполнении данной работы использовалось свободно распространяемое ПО, таблица содержит лишь один пункт – персональный компьютер, на котором проводилось все исследование.

Таблица 8 – Материальные затраты

Наименование	Ед. измерения	Количество	Цена за ед., руб	Затраты на материалы, руб
Персональный компьютер	шт	1	44000	50600
Итого				50600

3.3.2. Расчет затрат на специальное оборудование для научных (экспериментальных) работ

Данная статья состоит из затрат, связанных с приобретением спецоборудования, необходимого для проведения научных работ.

В ходе выполнения исследования использовалось имеющееся оборудование, поэтому его стоимость учитывается в калькуляции в виде амортизационных отчислений. Амортизационные отчисления можно рассчитать по формуле:

$$З_{ам} = \frac{(Ц_i * Н_а)}{100\%}, \quad (16)$$

Где $З_{ам}$ – ежедневная сумма амортизационных отчислений;

$Ц_i$ – норма амортизационных отчислений (%), которая в соответствии с Налоговым кодексом РФ определяется по формуле:

$$Н_а = \frac{1}{Т_{п.и.}} * 100\%, \quad (17)$$

Где $Т_{п.и.}$ – срок полезного использования объекта (в днях), определяется в соответствии с классификацией основных средств, включаемых в амортизационной группы. Персональный компьютер относится к второй амортизационной группе, где срок полезного использования от двух до трех лет. Установлен срок полезного использования 730 дней.

$Н_а = \frac{1}{730} * 100\% = 0{,}137$ %;

$З_{ам} = \frac{(44000*0{,}137)}{100\%} = 60{,}28$ руб.

Срок реализации проекта = 80 д.

Амортизация за период = $З_{ам}$ * Срок реализации проекта = 60,28 *80 = 4822,4 руб.

Таблица 9 – Расчет бюджета затрат на приобретение спецоборудования для научных работ

Наименование оборудования	Количество	$Т_{п.и.}$	$Н_а$	$З_{ам}$	Амортизация за весь период
Персональный компьютер	1 шт.	730	0,137	60,28	4822,4

3.3.3. Основная заработная плата исполнителей темы

Данная часть состоит из основной заработной платы работников, непосредственно занятых выполнением исследования, (включая премии, доплаты) и дополнительной заработной платы:

$$З_{зп} = З_{осн} + З_{доп}, \quad (18)$$

Где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата.

Основную заработную плату руководителя от предприятия можно рассчитать по формуле:

$$З_{осн} = З_{дн} \cdot T_р, \quad (19)$$

Где $З_{осн}$ – основная заработная плата одного работника;

$Т_р$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн.;

$З_{дн}$ – среднедневная заработная плата работника, руб. Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м * М}{F_л}, \quad (20)$$

где $З_м$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года, М =10,4 месяца при отпуске 48 раб. дней, 6-дневная неделя;

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн. Баланс рабочего времени представлен в таблице 10.

Таблица 10 – Баланс рабочего времени (для 6-дневной недели)

Показатели рабочего времени	Дни
Календарные дни	365
Нерабочие дни (праздники/выходные)	66
Потери рабочего времени (отпуск/невыходы по болезни)	56
Действительный годовой фонд рабочего времени	243

Месячный должностной оклад работника:

$$З_м = З_{тс} * (1 + k_{пр} + k_д) * k_р, \quad (21)$$

Где $З_{тс}$ – заработная плата по тарифной ставке, руб.;

$k_{пр}$ – премиальный коэффициент, равный 0,3 (т.е. 30% от $З_{тс}$);

$k_д$ – коэффициент доплат и надбавок в данной работе равен 15%;

$k_р$ – районный коэффициент, равный 1,3 (для Томска).

Расчёт основной заработной платы представлен в таблице 11.

Таблица 11 – Расчет основной заработной платы

Исполнитель	$З_{тс}$, руб	$k_{пр}$	$k_{д}$	$k_{р}$	$З_{м}$, руб	$Т_{р}$, раб. дни	$З_{осн}$, руб
Студент	5 045,2	0,3	0,2	1,3	9 838.14	80	34248,25
Научный руководитель	21 760	0,3	0,2	1,3	42 432	22	39987,78
Итого:							74236,04

3.3.4. Дополнительная заработная плата исполнителей темы

Расчет дополнительной заработной платы производится по следующей формуле:

$$З_{доп} = k_{доп} * З_{осн}, \quad (22)$$

где $k_{доп}$ – коэффициент дополнительной заработной платы (15% от заработной платы).

Расчет дополнительной заработной платы представлен в таблице 12.

Таблица 12 – Расчет дополнительной заработной платы

Исполнители	$З_{осн}$, руб.	$k_{доп}$	$З_{доп}$, руб.
Студент	34248,25	0,15	5137,24
Научный руководитель	39987,78	0,15	5998,17
Итого:			11135,41

3.3.5. Отчисления во внебюджетные фонды (страховые отчисления)

Величина отчислений во внебюджетные фонды находится по следующей формуле:

$$З_{внеб} = k_{внеб} * (З_{осн} + З_{доп}), \quad (23)$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и т.д.). Расчёт отчислений во внебюджетные фонды представлен в таблице 13.

Таблица 13 – Расчет отчислений во внебюджетные фонды

Исполнители	$З_{осн}$, тыс. руб.	$K_{внеб}$	$З_{доп}$, тыс. руб	$З_{внеб}$, руб.
Студент	34,25	0,28	5,14	11027,94
Научный руководитель	39,99	0,28	5,99	12876,0
Итого:				23904

3.3.6. Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов. Их величина определяется по следующей формуле:

$$З_{накл} = \text{сумма статей } (1 \div 5) \cdot k_{нр}, \quad (24)$$

где $k_{нр}$ – коэффициент, учитывающий накладные расходы ($k_{нр}$ = 16%).

$З_{накл} = 164697,85 \cdot 0,16 = 26351,66$ руб.

3.3.7. Формирование бюджета затрат научно-исследовательского проекта

По итогам расчётов можно составить полный бюджет затрат на реализацию проекта (таблица 14). Так как затраты по соответствующим статьям для всех вариантов использования равны, в таблице приведены общие значения.

Таблица 14 – Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб	Примечание
1. Материальные затраты НТИ	50600	Пункт 3.4.1
2. Затраты на специальное оборудование для научных работ	4822,4	Пункт 3.4.2
3. Затраты по основной заработной плате исполнителей темы	74236,04	Пункт 3.4.3
4. Затраты по дополнительной заработной плате исполнителей темы	11135,41	Пункт 3.4.4
5. Отчисления во внебюджетные фонды	23904	Пункт 3.4.5
6. Накладные расходы	26351,66	16% от суммы ст. 1-5
7. Бюджет затрат НТИ	191049,5	Сумма всех статей

3.4. Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования

В ходе выполнения данного раздела ВКР были выявлены сильные и слабые стороны разработки, также проанализированы факторы внешней среды. Это позволило сформировать стратегии дальнейшей работы с проектом, которые показывают заинтересованность инициаторов проекта в его реализации и дальнейшем развитии.

Кроме того, была подсчитана общий бюджет затрат на реализацию проекта, который составил 191049,5 руб. Срок реализации проекта по расчет составит 8о дней. Также оценка качественных параметров проекта показала,

что его разработка будет перспективной – средневзвешенный показатель качества составил 84%.

Разработка решает проблему рационального планирования загрузки вычислительных задач, из чего можно сделать вывод об необходимости реализации проекта, вне зависимости от его себестоимости, так как инициаторами проекта не было установлено ограничений на сумму его реализации.

Глава 4. Социальная ответственность

Данная работа была выполнена при помощи персонального компьютера в помещении с естественным и искусственным освещением, оборудованном компьютерными столом и креслом.

Разработка данного программного модуля, не оказывает негативного воздействия на окружающую среду, но при этом работа на персональном компьютере оказывает негативное воздействие на здоровье человека. Данное воздействие осуществляется за счет монитора и системного блока компьютера.

4.1. Правовые и организационные вопросы обеспечения безопасности

Большая часть данной работы была выполнена при помощи персонального компьютера. По этой причине необходимо соблюдать режим работы и отдыха. Данный вид работы относится к группе В, в соответствии с СанПиНом 2.2.2/2.4.2732-10, так как он заключается в творческой работе в режиме взаимодействия с ПК. Ниже (таблица 1) представлено время перерывов исходя из продолжительности работы.

Таблица 15 – Время регламентированных перерывов при работе с ПК

Категория работы с ПК	Уровень нагрузки за рабочую смену			Суммарное время регламентированных перерывов, мин	
	Группа А, количество знаков	Группа Б, количество знаков	Группа В, ч	8-ми часовая работа	12-ти часовая работа
I	до 20000	до 15000	до 2	50	80
II	до 40000	до 30000	до 4	70	110
III	до 60000	до 40000	до 6	90	140

С целью повышения работоспособности сотрудника, рекомендовано выполнять работу над заданием совмещая и чередуя работу с использованием ПК и без него. Также советуется применять индивидуальный подход с

ограничением времени работы с ПК, если несмотря на соблюдение всех санитарно-гигиенических и эргономических требований, у работника с ПК возникает зрительный дискомфорт и другие неблагоприятные субъективные ощущения.

Если работа требует сильного напряжения зрительного внимания и высокого уровня концентрации, а при этом периодическое переключение внимания на другие виды трудовой деятельности без использования персонального компьютера невозможно, следует внедрять перерывы на 10-15 мин каждые 45-60 мин работы. Кроме того, с целью снижения утомления зрительного анализатора, нервно-эмоционального напряжения, а также предотвращения развития утомления, во время регламентированных перерывов возможно выполнение комплексы специальных упражнений, благоприятно сказывающихся на организме человека.

4.2. Производственная безопасность

Все возможные опасные и вредные производственные факторы, которые могут оказывать какое-либо негативное влияние на организм человека можно разделить на несколько групп воздействия: физическое, химическое, психофизическое и биологическое.

Так как химические и биологические факторы воздействия не оказывают весомого влияния на организм исполнителя данной работы, то ниже будут рассмотрены лишь физические факторы.

В таблице 2 наглядно представлены все возможные опасные и вредные факторы, которые можно выявить при выполнении данного проекта. А также приведена их классификация в соответствии с нормативными документами [15].

Таблица 16 – Возможные опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Проектирование	Программирование	Эксплуатация	
1. Отклонение показателей микроклимата	+	+	+	1. ГОСТ 12.0.003.2015 2. ГОСТ 12.1.006-84 3. ГОСТ 12.1.003.2014 4. СанПин 2.2.4.548-96 5. СанПин 2.2.1/2.1.1.1278-03 6. СанПин 2.2.2/2.4.1.2732-10 7. СанПин 2.2.4.1191-03 8. СанПин 2.2.4.1340-03 9. СНиП 2.04.05-91 10. СН 2.2.4/2.1.8.562-96 11. НПБ 105-95
2. Превышение уровня шума	+	+	+	
3. Превышение уровня электромагнитных излучений	+	+	+	
4. Отклонение показателей освещенности рабочего места	+	+	+	

4.3. Анализ опасных и вредных факторов

4.3.1. Микроклимат рабочего помещения

Для анализа микроклимата рабочего помещения необходимо оценить следующие параметры: влажность и скорость движения воздуха, а также температуру. Данные факторы оказывают существенное влияние на человеческий организм и определяют микроклимат, по этой причине необходимо соблюдать соответствие данных показателей санитарным нормам.

Оптимальные микроклиматические условия – это комбинация значений характеристик микроклимата, которая обеспечивает ощущение теплового комфорта на протяжении всей 8-ми часовой рабочей смены при наименьшем напряжении механизмов терморегуляции.

В связи с тем, что работа исполнителя проводилась сидя за столом, с использованием компьютера и сопровождалась незначительным количеством физических нагрузок, то в соответствии с СанПин 2.2.4.548-96, такой вид

работ можно отнести к легкой физической работе (Iа) [16]. В таблице 3 приведены оптимальные значения характеристик микроклимата.

Таблица 17 – Оптимальные значения характеристик микроклимата

Период года	Категория работ по уровню энергозатрат, Вт	Температура воздуха, °С	Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	Iа (до 139)	22-24	21-25	60-40	0,1
Теплый	Iа (до 139)	23-25	22-26	60-40	0,1

Допустимые микроклиматические условия – это комбинация значений характеристик микроклимата, которые установлены соответственно критериям допустимого функционального и теплового состояния человека на протяжении 8-ми часовой рабочей смены. Соблюдение этих условий не вызывает весомых нарушений состояния здоровья и повреждений организма, но возможно появление легкого дискомфорта, понижение работоспособности и ухудшение самочувствия. Допустимые величины показателей устанавливают в случаях, если по технологическим требованиям, техническим и экономическим причинам не могут быть достигнуты оптимальные значения. Ниже представлены допустимые значения (таблица 4).

Таблица 18 – Допустимые значения характеристик микроклимата

Период года	Категория работ по уровню энергозатрат, Вт	Температура воздуха, °С		Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с	
		Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин			Для диапазона температур воздуха ниже оптимальных величин	Для диапазона температур воздуха выше оптимальных величин
Холодный	Iа (до 139)	20,0-21,9	24,1-25,0	19,0-26,0	15-15	0,1	0,1
Теплый	Iа (до 139)	21,0-22,9	25,1-28,0	20,0-29,0	15-15	0,1	0,2

Для помещения, в котором разрабатывался данный модуль, параметры микроклимата регулировались системой центрального отопления.

4.3.2. Производственные шумы

Шум – это звуковые колебания, находящиеся в диапазоне слышимых частот, которые способны оказывать негативное воздействие на здоровье и безопасность работника[14].

На рабочем месте сотрудника шум оказывает раздражающее влияние на человека, ускоряет и увеличивает его утомляемость, что при выполнении работы, требующей высокой концентрации, приводит к увеличению времени выполнения и росту ошибок. Длительное воздействие шума на работника вызывает нарушение слуха, а высокий уровень вредности от непредвиденных высокоинтенсивных шумов: кратковременные (удары, взрывы и т.д.) шумы способны привести к острым нейросенсорным и физическим повреждениям.

В рассматриваемом рабочем пространстве основным источником шума является персональный компьютер, с помощью которого выполняется данный проект. Согласно требованиям СН 2.2.4/2.1.8.562-96, при выполнении легкой физической работы в рабочих помещениях, уровни шума на рабочих местах не могут превышать предельно допустимых значений, т.е. 80 дБА [19]. Данное условие соблюдено.

4.3.3. Электромагнитное излучение

При работе с персональным компьютером организм человека подвергается воздействию электромагнитного излучения, которое в свою очередь оказывает негативное влияние на организм.

Электромагнитное поле (ЭМП), создаваемое персональным компьютером, является совокупностью электрического и магнитного полей, порождающих друг друга. Величина действия данного поля напрямую зависит

от величин размера облучаемого тела, напряженности электрического поля (Е), магнитного поля (Н), потока энергии и частоты колебаний.

Согласно СанПиН 2.2.4.1191-03 время пребывания человека в зоне воздействия электромагнитного излучения (Т) зависит от значений напряженности электрического поля [14]. Количество часов допустимого пребывания в рабочей зоне определяется по следующей формуле:

$$\mathrm{T} = \frac{50}{\mathrm{E}} - 2, \qquad (25)$$

Например, работа в условиях облучения электрическим полем с напряженностью 10-12 кВ/м не может продолжаться более 3 часов. При напряженности до 5 кВ/м, что соответствует персональному компьютеру, разрешается присутствие людей в течение 8 часов.

Ниже в таблице 5 подробно представлено какое время допустимо для пребывания работника в зоне воздействия электромагнитного поля, зависящее от напряженности поля и магнитной индукции.

Таблица 19 – Допустимое время пребывания в зоне действия ЭМП

Время воздействия за рабочий день, мин	Условия воздействия			
	Общее		Локальное	
	Предельно допустимый уровень напряженности, кА/м	Предельно допустимый уровень магнитной индукции, мТл	Предельно допустимый уровень напряженности, кА/м	Предельно допустимый уровень магнитной индукции, мТл
0-10	24	30	40	50
11-60	16	20	24	30
61-480	8	10	12	15

Также существуют временные допустимые уровни (ВДУ) электромагнитных полей регламентирующиеся СанПиН 2.2.2/2.4.2732-10. Ниже в таблице 6 можно более подробно увидеть временные допустимые уровни ЭМП, которые создаются персональным компьютером.

Таблица 20 – Временные допустимые уровни ЭМП

Наименование параметров		ВДУ ЭМП
Напряженность электрического поля	в диапазоне частот 5 Гц - 2 кГц	25 В/м
	в диапазоне частот 2 кГц - 400 кГц	2,5 В/м
Плотность магнитного потока	в диапазоне частот 5 Гц - 2 кГц	250 нТл
	в диапазоне частот 2 кГц - 400 кГц	25 нТл
Электростатический потенциал экрана видеомонитора		500 В

Чтобы обезопасить себя от воздействия ЭМП необходимо выполнять следующие условия:

- использовать сертифицированный ПК;
- увеличить расстояние от облучаемого тела до источника излучения (минимальное расстояние – 50см);
- делать перерывы в работе с компьютером.

4.3.4. Производственное освещение

Производственное освещение – это процесс использования искусственных и естественных источников света для обеспечения зрительной работоспособности в производственных помещениях. Необходимо соблюдать правильный режим освещения в помещении, так как его отклонение от нормы может оказывать отрицательное влияние на организм человека.

В соответствии с гигиеническими требованиями освещение данного типа помещения, необходимо оборудовать искусственными и естественными источниками света [18]. Рабочее место требуется располагать боковой стороной к световым проемам, а естественный свет должен падать на рабочее пространство преимущественно слева. Величина минимального значения уровня искусственного освещения необходимо удерживать в размере не менее 300 лк. В свою очередь, величина коэффициента пульсации освещенности должна быть в размере свыше 5%. Освещенность поверхности рабочего стола

необходимо варьировать в пределах 300-500 лк. Яркость светильников общего освещения и светящихся поверхностей (в поле зрения) не может превышать 200 кд/м^2 [19].

В рассматриваемом помещении имеется сочетание естественного и искусственного освещений. Для искусственного освещения используются 1 светильник, расположенный в центре потолка комнаты и имеющий 5 светодиодных ламп, и одна настольная светодиодная лампа. Также в помещении находится один оконный проем, и рассматриваемое рабочее место располагается слева от него.

Таблица 21 – Параметры систем искусственного и естественного освещения

Наименование рабочего места	Тип светильника и источника света	Коэффициент естественной освещенности, %		Освещенность при совмещенной системе, лк	
		Фактически	Норм. значение	Фактически	Норм. значение
Помещение для работы с ПК	LED B35-9W	-	0,7	1010 лк	300÷500 лк

4.3.5. Электробезопасность

Электрическая безопасность – это система технических средств и организационных предприятий, направленная на предотвращения опасного и вредного воздействий на работника от электрического тока, статического электричества и электрической дуги [20].

В помещениях для работы с персональными компьютерами, электричество представляет особую опасность, так как сам ПК находится под напряжением. Увеличивают риск поражения электрическим током в помещении следующие факторы: превышение относительной влажности (более 75%) и температуры (более 35°С), одновременное соприкосновение с металлическим элементом, соединенным с землей и металлическим корпусом

электрооборудования, а также наличие токопроводящих полов и пыли. В связи с этим, опасность поражения человека электрическим током зависит от соблюдения правил электробезопасности и правильного размещения оборудования.

Причины поражения током: короткое замыкание в высоковольтных блоках, прикосновение к токоведущим частям, прикосновение нетоковедущим частям и поверхностям, под напряжением.

В рассматриваемом помещении используется персональная электронно-вычислительная машина и другие устройства, использующие электрический ток, поэтому нужно следовать следующим мерам предосторожности:

1. Убедиться в исправности оборудования перед использованием;
2. При обнаружении неисправностей, не предпринимать никаких самостоятельных действий по исправлению и незамедлительно сообщить ответственному за оборудование о неисправности;
3. Необходимо содержать рабочее место свободным от лишних предметов.

При возникновении несчастного случая необходимо незамедлительно освободить пострадавшего от действия электрического тока, обращаясь в скорую помощь, оказать ему необходимую помощь.

4.4. Экологическая безопасность

Для типа помещения, в котором выполнялся описываемый проект, мероприятия по экологической безопасности сводятся к утилизации отходов жизнедеятельности человека и бытового мусора. В данных помещениях основными отходами являются канцелярские принадлежности, бумага, продовольственные отходы, батарейки, энергосберегающие лампы и прочее [15].

В связи с таким разнообразием отходов необходимо принять меры по сортированию и сбору мусора в зависимости от его происхождения перед устранением. Данное внедрение позволит сильно упростить процесс переработки отходов для вторичного использования, а также избежать его горения и гниения в окружающей среде. Стоит отметить, что отходы, которые имеют статус опасный для окружающей среды, установленный в Федеральном Классификационном Каталоге Отходов (ФККО), необходимо устранять при помощи специальных утилизирующих компаний.

4.5. Безопасность в чрезвычайных ситуациях

Для рассматриваемого помещения самым распространенным видом чрезвычайных ситуаций является пожар, по этой причине в данном разделе будет подобно рассмотрена пожарная безопасность.

Состояние производственного помещения, при котором с выявленной вероятностью исключается возможность возникновения и развития пожара, а также воздействия на людей его опасных факторов и создаются условия для защиты материальных ценностей, называется – пожарная безопасность.

4.5.1. Основные правила пожарной безопасности помещения

Для рассматриваемого помещения работнику необходимо соблюдать следующие правила пожарной безопасности:

1. Использование исправного оборудование с действующим техническим паспортом;
2. По окончании работы с электрическими устройствами работнику отключить и обесточить устройства;
3. Рабочее место содержать в чистоте и порядке, очищать стол от излишних бумаг и предметов;
4. Ежедневно освобождать помещение от мусора;
5. Соблюдать Правила технической эксплуатации и устройства электроустановок при использовании электрооборудования

6. Не использовать неисправное электрооборудование;
7. Обеспечить свободный доступ к электрооборудованию и электрощитам;
8. Проводить уборку помещения и чистку оборудования после отключения всех электроприборов их сети;
9. Подключать электрические приборы к сети через исправный сетевой фильтр;
10. Проверять исправность штепсельных розеток и электрошнуров, а в случае неисправности не использовать.
11. Проверка и измерение сопротивления изоляции в осветительных и силовых сетях минимум один раз в год.

1.7.1. Мероприятия по предупреждению и устранению пожаров

Чтобы предупредить возникновение пожара помещение следует оборудовать средствами связи и средствами тушения пожара (огнетушителями, стендом с противопожарным инвентарем, ящиком с песком), эксплуатировать только исправную электрическую проводку электрооборудования и осветительных приборов. Каждому работнику необходимо предоставлять информацию о месте нахождения средств связи, средств пожаротушения и номерах телефонов для сообщения о пожаре. Необходимо обучать и проверять работников на предмет умения пользоваться средствами пожаротушения. Для этого возможно внедрение регулярного проведения инструктажа работников о пожаробезопасности, с последующим тестированием.

В рассматриваемом помещении имеются переносной углекислотный огнетушитель, использование которого способно предотвратить первичное загорание, электрические щит, способный обесточить все электроприборы, а также все электрические приборы подключаются в сети посредством сетевых фильтров, помогающий бороться со скачками электрического тока,

инструкция противопожарной безопасности и таблички с указанием номеров телефонов для сообщения о пожаре.

4.5.2. Действие сотрудников в случае пожара

При обнаружении пожара работнику необходимо незамедлительно сообщить об случившемся в пожарную службу по телефону 01 и кратко сообщить: фамилию и имя сотрудника, адрес возгорания, предмет и причину возгорания, наличие опасности для людей (есть или нет);

Следующим шагом необходимо продублировать информацию о возгорании руководителю, обесточить с помощью электрощита помещение, начать тушение пожара огнетушителем и подручными средствами. Работнику необходимо следовать распоряжениям руководителя, а после организованно покинуть здание. В случае невозможности выхода из здания (сильное задымление, огонь перекрыл выход, высокая температура) необходимо плотно закрыть дверь помещения и уплотнить плотной тканью (при возможности намочить) щель между дверью и полом, а также вентиляционные отверстия, после открыть окно и ждать пожарных. Следует усвоить, что во время задымления более чистый воздух над полом, и использовать данную информацию при эвакуации и ожидании помощи.

Выводы по разделу

В процессе выполнения данной главы было определено, что помещение, в котором выполнялась данная выпускная квалификационная работа, соответствует установленным требованиям по следующим параметрам: микроклимат рабочего помещения, уровень шума, уровень электромагнитного излучения, показатели освещенности рабочего места.

Так же стоит отметить, что данный тип помещений находится в зоне риска возникновения пожара. По этой причине в помещении проведены меры по пресечению подобных ситуаций, а также проводятся информирование сотрудников о правилах противопожарной безопасности.

Заключение

В процессе выполнения данной работы были достигнуты следующие задачи:

- Изучены понятия и методы машинного обучения;
- Изучена предметная область;
- Применены четыре различных метода кластеризации к имеющимся данным;
- Выявлен наиболее эффективный метод.

Каждый из разработанных методов имеет свои преимущества и недостатки, относительно других методов. Но наиболее эффективными оказались методы Уорда и BIRCH, которые наиболее точно отразили количество кластеров и их расположение.

Метод Уорда требует для выполнения больших временных затрат, чем метод BIRCH, по этой причине последний был выбран наиболее эффективным методом.

Также хорошо проявил себя метод К-средних, но для его использования необходимо вводить количество кластеров. Для этого потребовалось использовать дополнительные метрики качества, которые позволили вычислить оптимальное количество кластеров, для имеющихся данных.

Список используемой литературы

1. Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения [Электронный ресурс] / Рашка С. – ДМК Пресс, 2017.
2. Юре, Л. Анализ больших наборов данных [Электронный ресурс] / Юре Л., Ананд Р., Джеффри Д. У.; Пер. с англ. Слинкин А.А. – ДМК Пресс, 2016.
3. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: учебное пособие / А. А. Барсегян [и др.]. – 2-е изд. – СПб: БХВ-Петербург, 2008.
4. Открытый курс машинного обучения. Тема 7. Обучение без учителя: PCA и кластеризация. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/ods/blog/325654/
5. Обзор алгоритмов кластеризации данных. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/101338/
6. Машинное обучение. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/319288/
7. Интересные алгоритмы кластеризации, часть вторая: DBSCAN. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/322034/
8. Data cleaning with python. [Электронный ресурс]. – Режим доступа: https://www.datacamp.com/courses/cleaning-data-in-python
9. Л. П. Коэльо, В. Ричард Построение систем машинного обучения на языке Python. - М.: ДМК Пресс, 2016.
10. Язык программирования Python 3 [Электронный ресурс] / Python 3 для начинающих. – URL: https://pythonworld.ru/
11. Уэс, М. . Python и анализ данных [Электронный ресурс] / Уэс М. ; Пер. с англ. Слинкин А.А.. — ДМК Пресс, 2015. — 482 с.. — Книга из коллекции ДМК Пресс

12. SWOT-Анализ. 5 Главных Правил, Которых Стоит Придерживаться. [Электронный ресурс]. – Режим доступа: https://geniusmarketing.me/lab/swot-analiz-5-glavnyx-pravil-kotoryx-stoit-priderzhivatsya/

13. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение: учебно-методическое пособие / И.Г. Видяев, Г.Н. Серикова, Н.А. Гаврикова, Н.В. Шаповалова, Л.Р. Тухватулина З.В. Криницына; Томский политехнический университет. − Томск: Изд-во Томского политехнического университета, 2014.

14. СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы (с изменениями на 21 июня 2016 года).

15. ГОСТ 12.2.032-78. Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования.

16. СанПиН 2.2.4.548-96. Гигиенические требования к микроклимату производственных помещений.

17. СНиП 23-05-95*. Естественное и искусственное освещение (с Изменением N 1).

18. СанПиН 2.2.1/2.1.1.1278-03. Гигиенические требования к естественному, искусственному и совмещенному освещению жилых и общественных зданий (с изменениями на 15 марта 2010 года).

19. СанПиН 2.2.4/2.1.8.055-96. Электромагнитные излучения радиочастотного диапазона (ЭМИ РЧ). Санитарные правила и нормы.

20. ГОСТ 12.1.038-82. Система стандартов безопасности труда (ССБТ). Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов (с Изменением N 1).

```
mms = MinMaxScaler()
#открытие файла с данными и чтение
with open('14296407.csv', 'r') as f:
    data = f.readlines()

#приведение файла к табличному виду
fixed_df = pd.read_csv('14296407.csv',
                       sep=',', index_col='id')

#Поиск числовых столбцов
df = fixed_df._get_numeric_data()

#Замена NaN на 0
df=df.fillna(0)

x = df.iloc[0: 10000, [1, 2]].values
# Приведение к одной шкале
x = mms.fit_transform(x)
# Отображение первых 10000 выборок.
plt.scatter(x[0:10000, 0], x[0:10000, 1], color='black', label='data')
plt.legend(loc='upper left')
plt.show()
```

```
dist = []
for i in range(1, 11):
    km = KMeans( n_clusters=i, init='k-means++', n_init=10,
                 max_iter=10000, random_state=0 )
    km.fit(x)
    dist.append(km.inertia_)

#Построение графика
plt.plot(range(1, 11), dist, marker='o')
plt.xlabel('n')
plt.ylabel('diff')
plt.show()
```

```
km = KMeans (     n_clusters=3,
                  init='random',
                  n_init=10,
                  max_iter=10000,
                  tol=1e-04,
                   random_state=0 )
y_km = km.fit_predict(x)
plt.scatter ( x[y_km == 0, 0], x[y_km == 0, 1], s=50, c='lightgreen',
              marker='s', label='c1' )
plt.scatter( x[y_km == 1, 0], x[y_km == 1, 1], s=50, c='blue', marker='v',
             label='c2' )
plt.scatter( x[y_km == 2, 0], x[y_km == 2, 1], s=50, c='red', marker='*',
             label='c3' )
centers = km.cluster_centers_

#Построение графика
plt.scatter( centers[0:3, 0], centers[0:3, 1], c='black', label='centers' )
plt.legend(loc='upper left')
plt.grid()
plt.show()
```

```
db = DBSCAN( eps=0.026, min_samples=1, metric='euclidean' )

y_db = db.fit_predict(x)


#Построение графика

plt.scatter( x[y_db == 0, 0], x[y_db == 0, 1], c='lightblue', marker='o', s=10,
             label='c1' )

plt.scatter( x[y_db == 1, 0], x[y_db == 1, 1], c='red', marker='s', s=10,
             label='c2' )

plt.scatter( x[y_db == 2, 0], x[y_db == 2, 1], c='blue', marker='*', s=10,
             label='c3' )

plt.legend(loc='upper left')

plt.show()
```

```
brc = Birch(branching_factor=10000, n_clusters=None, threshold=0.55,
            compute_labels=True)
brc.fit(pca_2d)
y_brc = brc.predict(pca_2d)
# Построение графика
plt.subplots_adjust(left=.02, right=.98, bottom=.001, top=.96, wspace=.05,
                    hspace=.01)
colors = np.array(list(islice(cycle(['#377eb8', '#ff7f00', '#4daf4a',
                                     '#f781bf', '#a65628', '#984ea3',
                                     '#999999', '#e41a1c', '#dede00']),
                              int(max(y_brc) + 1))))
# Назначаем черный цвет для выбросов (если они есть)
colors = np.append(colors, ["#000000"])
plt.scatter(pca_2d[:, 0], pca_2d[:, 1], s=10, color=colors[y_brc])

plt.show()
```

```
dist = []
for i in range(2, 20):
    ward = AgglomerativeClustering(n_clusters=i)
    ward.fit(df)
    dist.append(metrics.silhouette_score(df, ward.labels_))
n_clusters = dist.index(max(dist))
ward = AgglomerativeClustering(n_clusters=n_clusters + 2)
y_wr = ward.fit_predict(x)

#Построение графика
plt.scatter( x[y_wr == 0, 0], x[y_wr == 0, 1], s=50, c='lightgreen', marker='s',
             label='c1' )
plt.scatter( x[y_wr == 1, 0], x[y_wr == 1, 1], s=50, c='blue', marker='v',
             label='c2' )
plt.scatter( x[y_wr == 2, 0], x[y_wr == 2, 1], s=50, c='red', marker='*',
             label='c3' )
plt.scatter( x[y_wr == 3, 0], x[y_wr == 3, 1], s=50, c='pink', marker='o',
             label='c3' )
plt.legend(loc='upper left')
plt.show()
```