**«Санкт-Петербургский государственный электротехнический университет**

**«ЛЭТИ» им. В.И.Ульянова (Ленина)»**

**(СПбГЭТУ «ЛЭТИ»)**

|  |  |  |
| --- | --- | --- |
| **Направление** | 01.03.02 – Прикладная математика и информатика | |
| **Факультет** | КТИ | |
| **Кафедра** | МОЭВМ | |
| *К защите допустить* |  | |
| Зав. кафедрой |  | Кринкин К.В. |

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

БАКАЛАВРА

Тема: РЕАЛИЗАЦИЯ РАЗЛИЧНЫХ МЕТОДОВ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ ДАННЫХ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент |  |  |  | Федоров А.М. |
|  |  | *подпись* |  |  |
| Руководитель | доцент |  |  | Жукова Н.А. |
|  | *(Уч. степень, уч. звание)* | *подпись* |  |  |

Санкт-Петербург

2016**ЗАДАНИЕ**

**на выпускную квалификационную работу**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Утверждаю | | | | | | | |
|  | | Зав. кафедрой МОЭВМ | | | | | | | |
|  | | \_\_\_\_\_\_\_\_\_\_\_\_ Кринкин К.В. | | | | | | | |
|  | | «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | | | | | |
| Студент | Федоров А.М. | | | | |  | Группа | 3381 | |
| Тема работы: Реализация различных методов оценки качества кластеризации данных | | | | | | | | | |
| Место выполнения ВКР: МОЭВМ | | | | | | | | | |
| Исходные данные (технические требования):  Реализовать методы оценки качества кластеризации и провести их анализ | | | | | | | | | |
| Содержание ВКР:  введение, современное состояние вопроса, описание исследования, результаты исследования, технико-экономическое обоснование, заключение, список использованных источников | | | | | | | | | |
| Перечень отчетных материалов: пояснительная записка, иллюстративный материал | | | | | | | | | |
| Дополнительный раздел: технико-экономическое обоснование | | | | | | | | | |
|  | | | | | | | | | |
| Дата выдачи задания | | | | Дата представления ВКР к защите | | | | | |
| «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | | «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | | | |
|  | | | |  | | | | | |
|  | | | |  | | | | | |
| Студент | | |  | | Федоров А.М. | | | |
| Руководитель доцент | | |  | | Жукова Н.А. | | | |

**календарный план выполнения**

**выпускной квалификационной работы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Утверждаю | | | |
|  | | Зав. кафедрой МОЭВМ | | | |
|  | | \_\_\_\_\_\_\_\_\_\_\_\_ Кринкин К.В. | | | |
|  | | «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | |
| Студент | Федоров А.М. | |  | Группа | 3381 |
| Тема работы: Реализация различных методов оценки качества кластеризации данных | | | | | |

|  |  |  |
| --- | --- | --- |
| № п/п | Наименование работ | Срок выполнения |
| 1 | Обзор литературы по теме работы | 03.01 – 07.01 |
| 2 | Генерация кластеризованных множеств | 08.01 – 10.01 |
| 3 | Реализация методов оценки качества кластеризации | 11.01 – 00.00 |
| 4 | Анализ методов оценки качества кластеризации | 00.00 – 00.00 |
| 5 | Оформление пояснительной записки | 00.00 – 00.00 |
| 6 | Оформление иллюстративного материала | 00.00 – 00.00 |

|  |  |  |
| --- | --- | --- |
| Студент |  | Федоров А.М. |
| Руководитель доцент |  | Жукова Н.А. |
| *(Уч. степень, уч. звание)* |  |  |

**РЕФЕРАТ**

Пояснительная записка 48 стр., 6 рис., 1 табл., 00 ист., 00 прил.

Объектом исследования являются методы оценки качества кластеризации.

Цель работы – реализовать различные методы оценки качества кластеризации и провести их анализ.

Для тестирования реализованных методов оценки качества кластеризации была написана программа, генерирующая множество точек на плоскости, разбитых на кластеры. При этом рассмотрены случаи различного качества кластеризации точек: высокий, средний и низкий уровни качества. Были реализованы различные методы оценки качества кластеризации. Затем было рассчитано время выполнения оценки качества с использованием различных методов в зависимости от таких параметров как количество кластеров, размерность пространства и количество объектов. По полученным значениям были построены графики, на основании которых было проведено сравнение эффективности методов.

**ABSTRACT**

The work is devoted to realization and assessment of different methods of clustering quality grade. To testing methods the program for generation of clustering sets was created. The clustering sets has different level of quality: high, medium and low. Then different methods of clustering quality grade were realized. Time of executing of this methods was calculated depending on such parameters as clusters count, space dimension and objects count. Based on calculated values of time graph were built. Then efficiency comparison of methods was made based on graphs.

**содержание**

[ВВЕДЕНИЕ 7](#_Toc473472274)

[1. СОВРЕМЕННОЕ СОСТОЯНИЕ ВОПРОСА 8](#_Toc473472275)

[1.1. Общие сведения 8](#_Toc473472276)

[1.2. История вопроса 8](#_Toc473472277)

[1.3. Современные проблемы 10](#_Toc473472278)

[1.4. Пути решения проблем 11](#_Toc473472279)

[2. ОПИСАНИЕ ИССЛЕДОВАНИЯ 12](#_Toc473472280)

[2.1. Описание генерации кластеризованных множеств 12](#_Toc473472281)

[2.2. Исследуемые методы оценки качества кластеризации 17](#_Toc473472282)

[2.3. Расчёт времени работы методов 24](#_Toc473472283)

[3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ 26](#_Toc473472284)

[3.1. Значения индексов оценки качества кластеризации низкого, среднего и высокого качества 26](#_Toc473472285)

[3.2. Значения индексов оценки качества кластеризации, свидетельствующие о том, что кластеризация проведена качественно 26](#_Toc473472286)

[3.3. Графики времени выполнения методов оценки качества кластеризации 26](#_Toc473472287)

[4. ТЕХНИКО-ЭКОНОМИЧЕСКОЕ ОБОСНОВАНИЕ 28](#_Toc473472288)

[4.1. Детализированный план работ 28](#_Toc473472289)

[4.2. Расчёт ожидаемой длительности работ 28](#_Toc473472290)

[4.3. Расчёт ставки заработной платы 29](#_Toc473472291)

[4.4. Расчёт расходов на заработную плату 29](#_Toc473472292)

[4.5. Расчёт отчислений на страховые взносы 30](#_Toc473472293)

[4.6. Расчёт затрат на содержание и эксплуатацию оборудования 30](#_Toc473472294)

[4.6. Совокупная величина затрат 31](#_Toc473472295)

[ЗАКЛЮЧЕНИЕ 32](#_Toc473472296)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 33](#_Toc473472297)

# ВВЕДЕНИЕ

Актуальность данной работы определяется тем, что существует необходимость определять корректность разбиения множеств на кластеры с целью проверки того, что полученному результату можно доверять. Также некоторые алгоритмы кластеризации разбивают множество на конкретное количество кластеров, и для определения этого количества можно перебирать различные варианты и определять наилучший из них, используя методы оценки качества кластеризации. На сегодняшний день существуют различные коэффициенты, расчёт которых позволяет оценить качество кластеризации. Основной целью работы является реализация методов, основанных на расчёте данных коэффициентов, с целью их анализа. В результате работы различные методы будут реализованы и протестированы, а также измерено время их выполнения при различных значениях количества кластеров, размерности пространства и количества объектов, что позволит определить, какие из методов являются наилучшими.

# 1. СОВРЕМЕННОЕ СОСТОЯНИЕ ВОПРОСА

## 1.1. Общие сведения

Кластеризация (кластерный анализ) – совокупность математических методов, выполняющих разбиение заданной выборки объектов на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Задачи кластеризации:

1. Понимание данных – разбиение выборки на группы позволяет упростить дальнейшую обработку данных и принятие решений, применяя к каждому кластеру свой метод анализа.
2. Сжатие данных – можно сократить исходную выборку, оставим по одному представителю от каждого кластера
3. Обнаружение новизны – выделение объектов, которые не удаётся присоединить ни к одному из кластеров.

Кластеризация может использоваться в таких областях, как:

1. Информатика – упрощение работы с информацией, визуализация данных, сегментация изображений, интеллектуальный поиск.
2. Экономика – анализ рынков и финансовых потоков, выделение закономерностей на фондовых биржах.
3. Маркетинг – сегментация рынков, анализ поведения потребителей
4. Лингвистика – восстановление эволюционного древа языков
5. Астрономия – выделение групп звёзд и галактик, автоматическая обработка снимков космоса.

## 1.2. История вопроса

Кластеризации предшествовала классификация – принцип распределения однородных объектов по группам по какому-либо общему признаку. Одна из первых работ, упорядочивающих процесс классификации – теория классификации и систематизации, предложенная французским ботаником Огюстеном Декандолем в 1813 году для классификации растений. Декандоль ставил своей целью описать и классифицировать все виды растений. Именно для этого им была разработана система классификации, в соответствии с которой каждое растение должно принадлежать к серии таксонов последовательно соподчинённых рангов (вид, род, семейство, класс, отдел), где таксон – это группа объектов, связанных общностью признаков. Данная теория получила название таксономия. Первоначально она использовалась исключительно в биологии, но позже она нашла применение и в других науках, имеющих дело со множествами иерархически организованных объектов.

Начало развития кластерного анализа как самостоятельной дисциплины относится к первой половине XX века. Одной из первых публикаций по теме кластеризации является статья польского антрополога Яна Чекановского, написанная в 1911 году. В этой статье он выдвигал идею о “структурной классификации”, содержащую основную мысль кластерного анализа - выделение групп близких объектов, а также некоторые способы выделения таких групп, которые легли в основу более поздних алгоритмов.

Следующим важным шагом в развитии кластеризации стал “метод корреляционных плеяд”, разработанный советским гидробиологом П.В. Терентьевым в 1925 году. Данный метод предназначался для кластерного анализа признаков, а не объектов. Идея метода заключалась в представлении признаков в виде вершин графа. Затем вершины соединяются рёбрами тогда и только тогда, когда коэффициент корреляции превосходит пороговое значение. Связанные компоненты такого графа и есть кластеры признаков.

Сам термин “кластерный анализ” был впервые введён и использован только в 1939 году английским учёным Трионом.

Большинство современных методов кластеризации были предложены в 1960-е годы. Это время характеризуется огромным количеством публикаций. В качестве основных можно перечислить работы следующих авторов: Г. Болла и Д. Холла, Д. Мак-Кина – по методам k-средних; Р.Сокала и Д. Снита, Г. Ланса и У. Уильямса, Н. Джардайна и Р. Сибсона – по иерархическим процедурам; Д. Роджеса и Т. Танимото, Э.М. Бравермана, А.А. Дорофеюка, И.Б. Мучника – по процедурам типа последовательного формирования кластеров и диагонализации. Эти и многие другие авторы сформировали математическую базу для применения кластерных методов в различных областях науки.

1970 годы характеризуются значительно меньшим количеством работ по созданию новых алгоритмов и методов. В это время осуществляются многочисленные попытки осмысления существующих алгоритмов и способов их применения.

После 1970-1980-х годов число публикаций по кластерному анализу падает, а его развитие продолжается в более тесном контакте с теми областями, где он используется. Основное направление развития в этот период – создание спецификаций существующих алгоритмов, наиболее подходящих для решения конкретных задач.

## 1.3. Современные проблемы

1. Проблема обоснования качества результатов кластеризации.

Процесс группировки носит субъективный характер. Один и тот же набор объектов может классифицироваться по-разному в зависимости от прикладной области, степени полноты информации об объектах и т.д.

1. Сложность формализации некоторых областей

Построение математических моделей некоторых областей затруднено, что приводит к тому, что использование алгоритмов расщепления смеси распределений (таких как EM-алгоритм) становится проблематичным.

1. Проблема анализа большого количества разнотипных признаков.

В случае разнотипного пространства возникает проблема определения в нём метрики. С другой стороны, даже в пространстве однотипных признаков при увеличении их числа точки могут стать неразличимыми. Так, расстояния от некоторой точки до ближайшей и наиболее удалённой точки могут практически совпадать в пространствах большой размерности.

1. Проблема поиска оптимального решения.

Многие алгоритмы не гарантируют, что найденное решение является оптимальным.

## 1.4. Пути решения проблем

1. Проблема обоснования качества результатов кластеризации.

Необходимо разрабатывать и применять методы оценки качества кластеризации, позволяющие максимально полно учитывать имеющиеся знания об объектах, например расчёт F1-меры.

1. Проблема анализа большого количества признаков.

Для проведения кластеризации в пространстве признаков высокой размерности можно воспользоваться алгоритмом CLIQUE, адаптированным под кластеризацию данных высокой размерности. Метод основан на том предположении, что если в многомерном пространстве данных распределение объектов не равномерно, то проекция региона плотности в подпространство с меньшей размерностью будет частью региона плотности в этом подпространстве. Алгоритм CLIQUE производит кластеризацию многомерного пространства данных следующим образом: пространство данных разбивается на не пересекающиеся ячейки фиксированного размера, среди них идентифицируются плотные ячейки – такие, плотность объектов данных в которых превышает заданное пороговое значение. Далее из найденных ячеек формируется пространство, в котором могут существовать плотные ячейки большей размерности. Процесс начинается с одномерных пространств (описанная процедура выполняется для каждого измерения) с последующим переходом к подпространствам более высокой размерности.

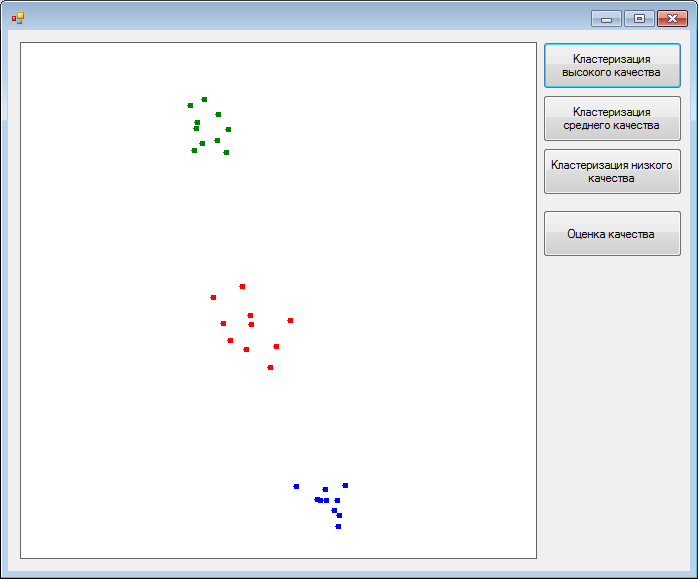
1. Проблема поиска оптимального решения.

Генетические (эволюционные) алгоритмы и нейронные сети позволяют найти оптимальное решение.

# 2. ОПИСАНИЕ ИССЛЕДОВАНИЯ

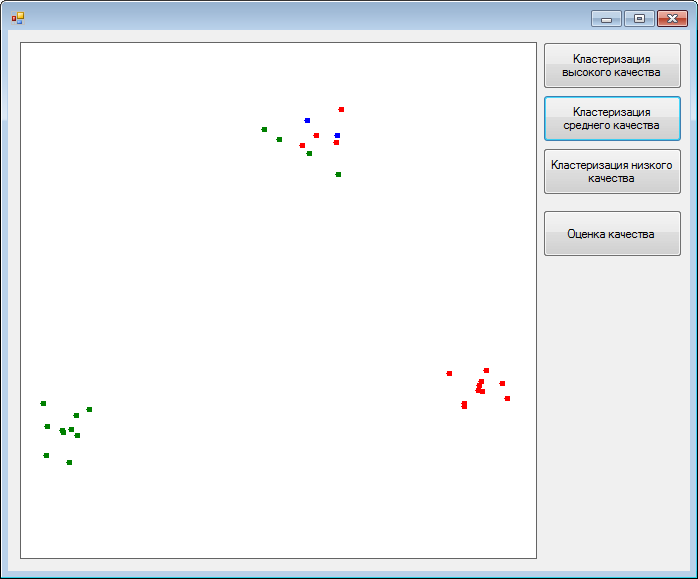
## 2.1. Описание генерации кластеризованных множеств

Для исследования методов оценки качества кластеризации была написана программа, генерирующая кластеризованные множества различного качества. Генерация работает следующим образом: случайным образом выбираются три точки на плоскости, затем внутри кругов с центрами в этих точках случайным образом генерируются 30 точек – по 10 в каждом круге.



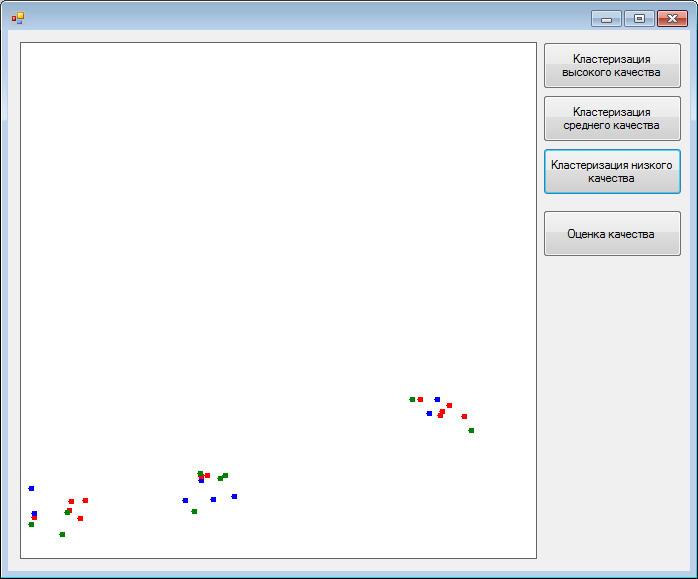
*Рисунок 1.Пример сгенерированного кластеризованного множества высокого качества.*

При выборе генерации кластеризованного множества высокого качества точки, сгенерированные внутри одного круга, относятся к одному кластеру. Пример такого множества представлен на рисунке 1.



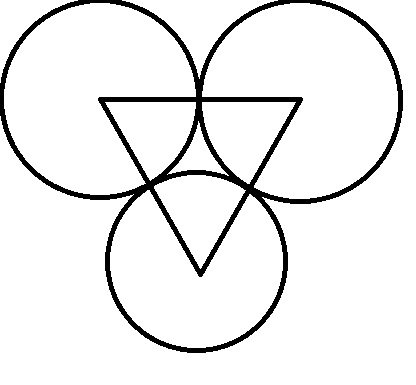
*Рисунок 2. Пример сгенерированного кластеризованного множества среднего качества.*

При выборе среднего качества кластеризации точки, сгенерированные внутри двух кругов, относятся к двум кластерам аналогично случаю высокого качестства кластеризации. Для точек, сгенерированных внутри третьего круга, принадлежность к кластерам выбирается случайным образом. Пример представлен на рисунке 2.



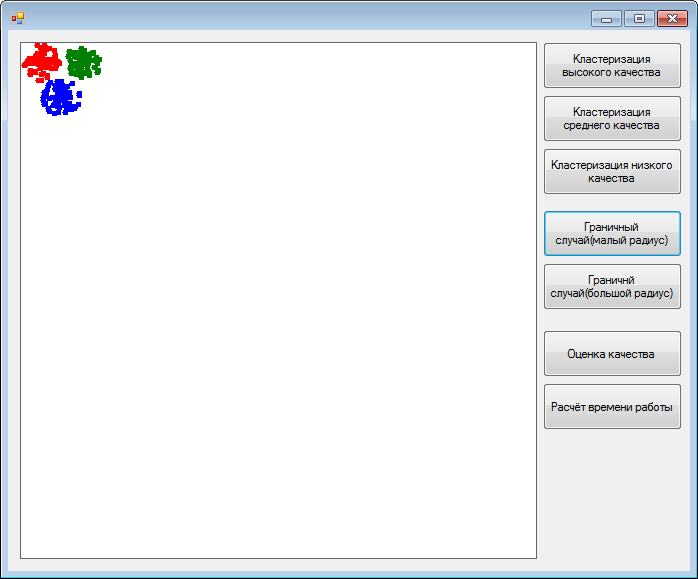
*Рисунок 3. Пример сгенерированного кластеризованного множества низкого качества.*

При выборе низкого качества кластеризации принадлежность каждой точки к какому-либо кластеру определяется случайным образом. Пример представлен на рисунке 3.

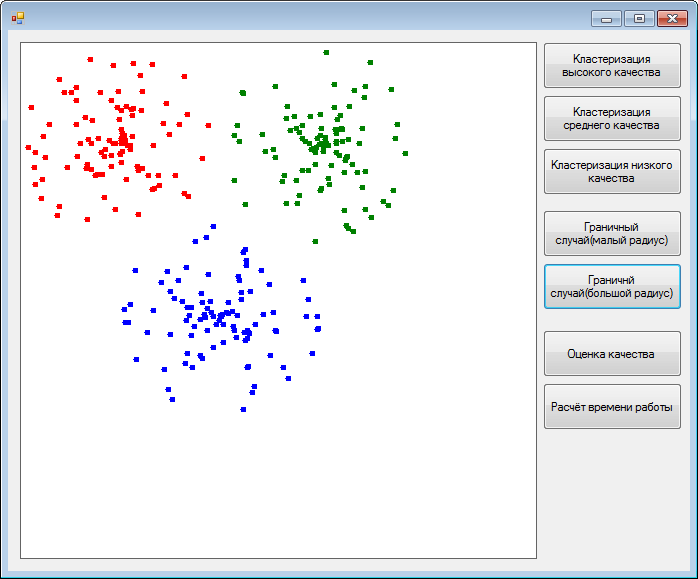


*Рисунок 4. Схема расположения кластеров в граничном случае.*

Для определения оптимальных значений исследуемых в работе индексов оценки качества кластеризации была добавлена возможность генерации кластеров в граничном случае. В качестве граничного случая понимается такой случай, в котором проведённую кластеризацию можно считать качественной, но дальнейшее увеличение размеров кластеров либо их сближение приведёт к их слиянию. Схема расположения кластеров в граничном случае представлена на рисунке 4. Три кластера, каждый из которых состоит из 100 точек, генерируются внутри трёх кругов. Их центры расположены в вершинах равностороннего треугольника, а радиусы равны половине длины стороны треугольника. Для того, чтобы убедится, что исследуемые индексы слабо изменяются при увеличении размеров кластеров в случае сохранения их утрукткры, была реализована генерация двух вариантов граничных случаев, которые отличаются только размерами кластеров. Примеры сгенерированных множеств для двух этих случаев представлены на рисунках 5 и 6.



*Рисунок 5. Пример сгенерированного кластеризованного множества в граничном случае с малым размером кластеров.*



*Рисунок 6. Пример сгенерированного кластеризованного в граничном случае с большим размером кластеров.*

## 2.2. Исследуемые методы оценки качества кластеризации

Были реализованы и исследованы следующие методы:

1. F1-мера(F1-meassure)

F1-мера является мерой сходства. Это значит, что для её расчёта используется информация о том, как должна быть проведена кластеризация. Данная информация представляется в виде набора классов. Класс представляет собой группу объектов, которые должны принадлежать одному кластеру. В качестве примера можно рассмотреть программную систему, распознающую автомобили и пешеходов. В данном случае имеются два класса: автомобили и пешеходы. При проведении кластеризации логично ожидать, что в результате получится два кластера, один из которых будет содержать только автомобили, а второй – только пешеходов. Для определения степени соответствия результата кластеризации ожидаемому результату и применяется расчёт мер сходства, в том числе и F1-меры.

При расчёте F1-меры от сгенерированного кластеризованного множества в качестве классов будут взяты множества точек, сгенерированных в одном круге.

Для расчёта F1-меры необходимо рассчитать precession и recall для каждого кластера и каждого класса по следующим формулам:

где – количество объектов , таких что и ; ;

Тогда F1-мера для кластера и класса будет равна:

F1-мера для всего кластеризуемого множества равна:

F1-мера лежит в интервале от нуля до единицы, при этом в случае наилучшей кластеризации F1-мера равна единице, а в случае наихудшей – нулю.

1. Индексы Rand, Jaccard, FM

Индексы Rand, Jaccard, FM являются мерами сходства. В качестве классов будут взяты множества точек, сгенерированных в одном круге.

Введём следующие обозначения:

* + SS – количество пар объектов (), в которых и принадлежат одному кластеру и одному классу
  + SD – количество пар объектов (), в которых и принадлежат одному кластеру, но разным классам
  + DS – количество пар объектов (), в которых и принадлежат разным кластерам, но одному классу
  + DD – количество пар объектов (), в которых и принадлежат разным кластерам и разным классам

Индекс Rand вычисляется по следующей формуле:

Индекс Jaccard вычисляется по следующей формуле:

Индекс FM вычисляется по следующей формуле:

1. Критерий Calinski-Harabasz(Calinski-Harabasz criterion)

Для расчёта данного критерия используются средние квадраты расстояний между элементами для всего множества и для каждого кластера. Обозначим средний квадрат расстояний для всего множества как , а для кластера как

Для расчёта критерия Calinski-Harabasz необходимо рассчитать сумму квадратов расстояний внутри кластеров по формуле:

где c – количество кластеров, – количество элементов в i-ом кластере

Также необходимо рассчитать сумму квадратов расстояний между группами по формуле:

где с – количество кластеров, N – количество элементов во всём множестве,

Тогда критерий Calinski-Harabasz вычисляется по формуле:

1. Индекс Данна(Dunn index)

Для расчёта данного критерия используются межкластерное расстояние r и диаметр кластера diameter, рассчитываемые по формулам:

где и – кластеры, – расстояние между и

Тогда индекс Данна вычисляется по формуле:

где с – количество кластеров; , – кластеры

По данной формуле виден недостаток индекса Данна: при больших значениях диаметра одного из кластеров и при малых значениях диаметров остальных кластеров значение индекса будет нехарактерно низким, в связи с тем, что в знаменателе формулы находится значение наибольшего диаметра.

1. Индекс Девиса-Болдуина (Davies-Bouldin index)

Для расчёта данного индекса используется мера разброса внутри кластера, определяемая по следующей формуле:

где – кластер, – количество объектов в кластере, – центр кластера,

значение p зависит от метрики пространства(p=2 для Евклидова пространства)

Также при расчёте индекса Девиса-Болдуина используется мера различия между кластерами, рассчитываемая по следующей формуле:

где dim – размерность пространства, и – k-ые координаты центров i-го и j-го кластера, значение p зависит от метрики пространства(p=2 для Евклидова пространства)

Определим коэффициент , который будет показывать, насколько качественна структура кластеров и . Пусть меньшее значение соответствует более качественной структуре. Более качественная структура достигается при наибольшем значении меры различия между кластерами и при наименьших значениях мер разбросов внутри кластеров и . Следовательно, должен обладать следующими свойствами:

* + Если и , то
  + Если и , то

Формула, обладающая данными свойствами, имеет следующий вид:

Теперь определим , рассчитываемое по данной формуле:

Тогда индекс Дэвиса-Болдуина будет равен:

где – количество кластеров

1. Индекс Scatter-Distance

Для расчёта индекса Scatter-Distance используются значения дисперсии на всём множестве и на каждом из кластеров. Дисперсия представляет собой вектор в пространстве кластеризуемого множества:  
   
где dim – размерность пространства  
Координата i вектора дисперсии на всём множестве вычисляются по формуле:  
где n – количество элементов множества,   
 – i-ая координата k-ого элемента множества,   
 – i-ая координата центра множества

Координата i вектора дисперсии на кластере j вычисляются по формуле:  
где – количество элементов кластера j,   
 – i-ая координата k-ого элемента кластера,   
 – i-ая координата центра кластера j  
Средний разброс в кластерах определяется следующей формулой:  
где с – количество кластеров, – вектор дисперсии для j-ого кластера,

– вектор дисперсии для всего множества  
Отделимость кластеров определяется следующей формулой:

где – расстояние между кластерами i и j, с – количество кластеров

Индекс Scatter-Distance вычисляется по следующей формуле:

где – взвешивающий коэффициент

1. Индекс Scatter-Density

Для расчёта индекса Scatter-Density используются значения дисперсии на всём множестве и на каждом из кластеров. Дисперсия представляет собой вектор в пространстве кластеризуемого множества:  
   
где dim – размерность пространства  
Координата i вектора дисперсии на всём множестве вычисляются по формуле:  
где n – количество элементов множества,   
 – i-ая координата k-ого элемента множества,   
 – i-ая координата центра множества

Координата i вектора дисперсии на кластере j вычисляются по формуле:  
где – количество элементов кластера j,   
 – i-ая координата k-ого элемента кластера,   
 – i-ая координата центра кластера j  
Средний разброс в кластерах определяется следующей формулой:  
где с – количество кластеров, – вектор дисперсии для j-ого кластера,

– вектор дисперсии для всего множества  
Определим функцию f следующей формулой:

где – объекты кластеризуемого множества,

– расстояние между и , – количество кластеров,

– вектор дисперсии для j-ого кластера

Функция плотности для двух кластеров определяется следующей формулой:

где и – кластеры,

– середина отрезка, соединяющая центры кластеров и

Функция плотности для одного кластера определяется следующей формулой:

где – кластер, – центр кластера

Плотность между кластерами определяется следующей формулой:

где с – количество кластеров

1. Индекс RMSSTD(Root – Mean – Square Standard Deviation)

Для расчёта индекса Scatter-Density используется сумма квадратов расстояний между центрами и элементами кластеров. Она рассчитывается по следующей формуле:

где – количество кластеров, – кластер, – расстояние между элементом кластера и центром кластера

Индекс RMSSTD рассчитывается по следующей формуле:

где – размерность пространства, – количество элементов,

с – количество кластеров

1. Hubert’s Г статистика(Hubert’s gamma statistic)

Количество пар элементов множества задаётся следующей формулой:

где N – количество элементов множества

Для расчёта Hubert’s Г статистики используется матрица близости(proximity matrix). В данной работе в качестве матрицы близости принята матрица расстояний между элементами. Каждый элемент такой матрицы , находящийся в i-ой строке j-ого столбца, является расстоянием между i-ым и j-ым элементом.

Также при расчёте используется функция Y(i, j), заданная следующей формулой:

Тогда Hubert’s Г статистика вычисляется по следующей формуле:

где – количество объектов, – элемент матрицы близости

1. Модифицированная Hubert’s Г статистика(Modified Hubert’s gamma statistic)

Количество пар элементов множества задаётся следующей формулой:

где N – количество элементов множества

Для расчёта модифицированной Hubert’s Г статистики используется матрица близости(proximity matrix). В данной работе в качестве матрицы близости принята матрица расстояний между элементами D. Каждый элемент такой матрицы , находящийся в i-ой строке j-ого столбца, является расстоянием между i-ым и j-ым элементом.

Также при расчёте используется матрица расстояний между центрами кластеров Q. Каждый элемент такой матрицы , находящийся в i-ой строке j-ого столбца, является расстоянием между центрами кластеров, к которым принадлежат i-ый и j-ый элемент.

Тогда модифицированная Hubert’s Г статистика вычисляется по следующей формуле:

где – количество объектов, – элемент матрицы близости,

– элемент матрицы расстояний между центрами кластеров

1. Нормализованная Hubert’s Г статистика(Normalized Hubert’s gamma statistic)

Количество пар элементов множества задаётся следующей формулой:

где N – количество элементов множества

Определим матрицу близости D и матрицу расстояний между центрами кластеров Q так же, как и для модифицированной Hubert’s Г статистики.

Тогда нормализованная Hubert’s Г статистика вычисляется по следующей формуле:

где – количество объектов, – элемент матрицы близости,

– элемент матрицы расстояний между центрами кластеров,

и - значения дисперсий для элементов матриц D и Q соответственно,

и - средние значения элементов матриц D и Q соответственно

1. Индекс RS(R Squared index)

Индекс RS основан на сравнении расстояний внутри кластеризованного множества с расстояниями внутри кластеров. При качественной кластеризации кластеры должны существенно различаться, поэтому расстояния внутри кластеров должны быть меньше расстояний внутри всего множества. Таким образом, RS индекс определяет степень непохожести кластеров друг на друга.

Для расчёта индекса RS необходимо рассчитать сумму квадратов расстояний для всего множества и сумму квадратов расстояний внутри кластеров.

Сумма квадратов расстояний для всего множества рассчитывается по следующей формуле:

где N – количество элементов множества,

– расстояние между i-ым элементом множества и центром множества

Сумма квадратов расстояний внутри кластеров рассчитывается по следующей формуле:

где c – количество кластеров, – количество элементов в i-ом кластере,

– расстояние между j-ым элементом i-ого кластера и центром i-ого кластера

Индекс RS вычисляется по формуле:

## 2.3. Расчёт времени работы методов

Для каждого из методов оценки качества кластеризации было рассчитано время выполнения в зависимости от количества кластеров, размерности пространства и количества объектов. Методы выполнялись на специально сгенерированных множествах. При генерации множеств координаты объектов выбирались случайным образом из множества целых чисел от 0 до 1000. Принадлежность к кластерам также выбиралась случайным образом. Для анализа времени выполнения такого метода определения качества кластеризации, как расчёт F1-меры, также генерировалась информация о принадлежности объектов классам, при этом принадлежность каждого объекта к классу определялась случайным образом, а количество классов совпадало с количеством кластеров.

Были рассмотрены следующие случаи:

1. Размерность: 3, количество объектов: 1000, количество кластеров: 2, 3, ….10
2. Количество кластеров: 3, количество объектов: 1000, размерность: 1, 2, ….10
3. Количество кластеров: 3, размерность: 3, количество объектов: 1000, 2000, ….10000

Для каждого случая время работы метода было посчитано пять раз, после чего было рассчитано среднее значение, которое затем было записано в текстовый файл. Для каждого из методов был сформирован свой текстовый файл, который содержал значения времени во всех вышеперечисленных случаях. Затем полученные файлы были открыты в Microsoft Excel, и по ним были построены следующие графики:

1. Графики зависимости времени выполнения всех методов от количества кластеров в одной системе координат
2. Графики зависимости времени выполнения всех методов от размерности пространства в одной системе координат
3. Графики зависимости времени выполнения всех методов от количества объектов в одной системе координат

# 3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

## 3.1. Значения индексов оценки качества кластеризации низкого, среднего и высокого качества

При расчёте значений индексов измерения проводились 5 раз, при этом после каждого измерения кластеризованное множество генерировалось заново. Затем были рассчитаны средние значения, которые и представлены в таблице 1.

*Таблица 1.Значения индексов оценки качества кластеризации в случаях кластеризации различного качества .*

|  |  |  |  |
| --- | --- | --- | --- |
| Индекс | Значение для случая кластеризации высокого качества | Значение для случая кластеризации среднего качества | Значение для случая кластеризации низкого качества |
| F1-мера | 1 | 0,8 | 0,48 |
| Индекс Rand | 1 | 0,78 | 0,57 |
| Индекс Jaccard | 1 | 0,52 | 0,2 |
| Индекс FM | 1 | 0,68 | 0,33 |
| Критерий Calinski-Harabasz | 1233,21 | 29,68 | 0,86 |
| Индекс Данна | 1,67 | 0,02 | 0,01 |
| Индекс Девиса-Болдуина | 0,3 | 0,97 | 9,06 |
| Индекс Scatter-Distance | 0,018 | 0,26 | 0,89 |
| Индекс Scatter-Density | 0,056 | 0,49 | 2,08 |
| Индекс RMSSTD | 14,97 | 60,8 | 73,36 |
| Hubert’s Г статистика | 150,53 | 133,16 | 124,89 |
| Модифицированная Hubert’s Г статистика | 43071,55 | 36259,71 | 7661,4 |
| Нормализованная Hubert’s Г статистика | 2,66 | 2,43 | 0,3 |
| Индекс RS | 0,97 | 0,33 | 0,03 |
| Индекс Силуэта | 0,78 | 0,11 | -0,092 |

Стоит отметить, что один из вышеперечисленных индексов: индекс RMSSTD, может иметь одно и то же значение на разных множествах при различном качестве кластеризации. Так, значение данного индекса, равное 73,36, не обязательно означает то, что кластеризация проведена плохо. Но, для каждого конкретного множества большее значение индекса RMSSTD означает то, что кластеризация проведена менее качественно. Таким же недостатком обладают Hubert’s Г статистика, модифицированная Hubert’s Г статистика и нормализованная Hubert’s Г статистика.

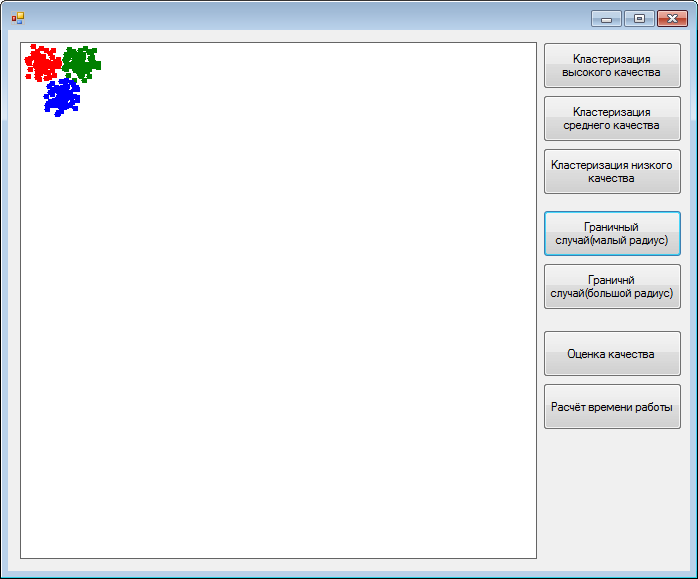
## 3.2. Значения индексов оценки качества кластеризации, свидетельствующие о том, что кластеризация проведена качественно

1. Меры сходства

В данной работе были исследованы такие меры сходства, как F1-мера, индекс Rand, индекс Jaccard, индекс FM. Если допустима ошибка в n%, то кластеризацию можно считать качественной, если значение данных индексов больше или равно

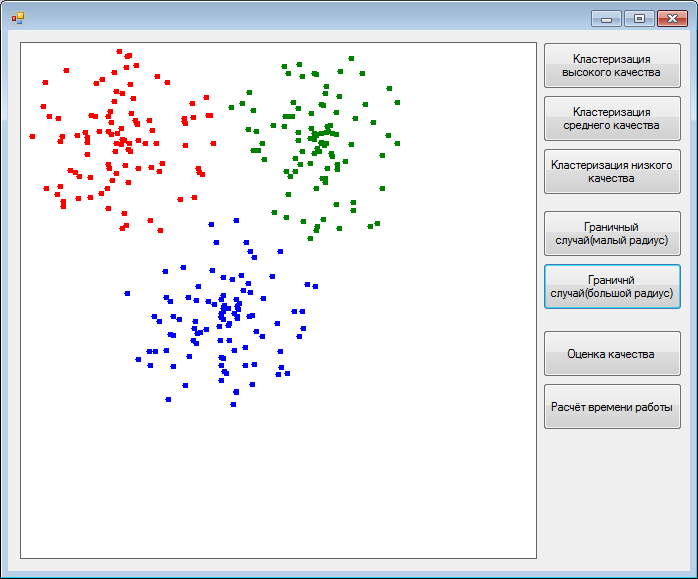
1. Индекс RMSSTD, Hubert’s Г статистика, модифицированная Hubert’s Г статистика

Одно и то же значение индекса RMSSTD на одном множестве может означать качественную кластеризацию, а на другом – некачественную. Также верно, что два множества, кластеризованные одинаково хорошо, могут иметь значения индекса RMSSTD, значительно отличающиеся друг от друга.

**

*Рисунок 7. Пример кластеризованного множества с малым размером кластеров.*

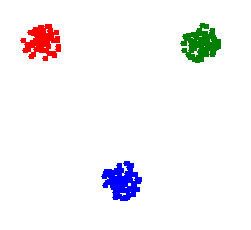
В качестве примера, подтверждающего это, можно рассмотреть множество, представленное на рисунке 7. Значение индекса RMSSTD для данного множества равно 6.



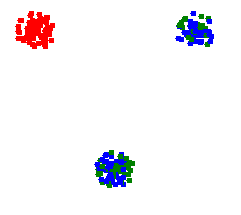
*Рисунок 8. Пример кластеризованного множества с большим размером кластеров.*

При увеличении размеров кластеров с сохранением их структуры и взаимного расположения будет получено множество, изображённое на рисунке 8. Значение индекса RMSSTD для данного множества будет равно 33. Таким образом, значение индекса увеличилось в 5,5 раз, хотя качество кластеризации практически не изменилось. Из вышеперечисленного можно сделать вывод о том, что не существует какого-либо конкретного значения данного индекса, которое позволило бы ответить на вопрос “Качественно ли проведена кластеризация?”. Данный индекс можно применять, если есть необходимость кластеризовать конкретное множество различными способами, и требуется определить, какой из способов даёт наилучший результат. На каждом фиксированном множестве, меньшее значение индекса RMSSTD будет означать, что кластеризация проведена более качественно.

Аналогичным недостатком обладают Hubert’s Г статистика, модифицированная Hubert’s Г статистика и нормализованная Hubert’s Г статистика. На каждом фиксированном множестве большее значение вышеперечисленных Hubert’s Г статистик будет означать, что кластеризация проведена более качественно. Стоит отметить, что во время выполнения работы было обнаружено, что нормализованная Hubert’s Г статистика принимает максимальное значение не в оптимальном случае, если множество представляет собой точки, находящиеся внутри кругов с центрами в вершине равностороннего треугольника (или в незначительном отдалении от вершин).



*Рисунок 9. Оптимальный случай кластеризации для множества точек в кругах с центрами в вершинах равностороннего треугольника.*



*Рисунок 10. Случай, при котором значение нормализованной Hubert’s Г статистики принимает значение, большее, чем в оптимальном случае.*

Например, на множестве, изображённом на рисунке 9, нормализованная Hubert’s Г статистика принимает значение 6 , а на множестве на рисунке 10 – значение 8 . Таким образом, несмотря на то, что наилучшая кластеризация представлена на рисунке 9, нормализованная Hubert’s Г статистика принимает большее значение в случае, представленном на рисунке 10. На других множествах подобное явление не наблюдалось.

1. Остальные индексы

Для определения граничных значений индексов, были проведены расчёты значений индексов на сгенерированных кластеризованных множествах в граничных случаях. При расчёте измерения проводились 5 раз для случаев больших и малых кластеров, при этом после каждого измерения кластеризованное множество генерировалось заново. Затем были рассчитаны средние значения, которые и представлены в таблице 2.

*Таблица 2.Граничные значения индексов оценки качества кластеризации.*

|  |  |  |
| --- | --- | --- |
| Индекс | Случай малых кластеров | Случай больших кластеров |
| Критерий Calinski-Harabasz | 651,9 | 603,08 |
| Индекс Данна | 0,14 | 0,09 |
| Индекс Девиса-Болдуина | 0,54 | 0,57 |
| Индекс Scatter-Distance | 0,21 | 0,2 |
| Индекс Scatter-Density | 0,26 | 0,31 |
| Индекс RS | 0,83 | 0,81 |
| Индекс Силуэта | 0,64 | 0,62 |

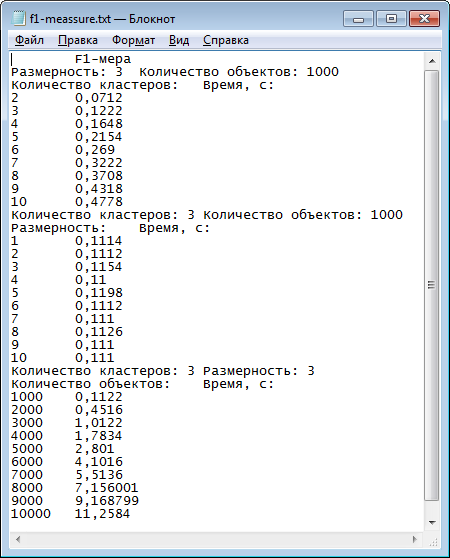
При увеличении качества кластеризации все вышеперечисленные индексы монотонно изменяются. При этом часть индексов монотонно возрастает, а часть – монотонно убывает. Для того чтобы определить поведение индекса при изменении качества кластеризации, можно обратиться к таблице 1, в которой перечислены значения индексов на кластеризованных множествах различного качества. Округлив граничные значения из таблицы 2 и воспользовавшись информацией из таблицы 1, можно определить, при каких значениях индексов кластеризацию следует считать качественной. Данные значения перечислены в таблице 3.

*Таблица 3.Значения индексов оценки качества кластеризации, свидетельствующие о том, что кластеризация проведена качественно.*

|  |  |
| --- | --- |
| Индекс | Значение |
| Критерий Calinski-Harabasz | >500 |
| Индекс Данна | >0,05 |
| Индекс Дэвиса-Болдуина | <0,7 |
| Индекс Scatter-Distance | <0,25 |
| Индекс Scatter-Density | <0,4 |
| Индекс RS | >0,8 |
| Индекс Силуэта | >0,6 |

## 3.3. Графики времени выполнения методов оценки качества кластеризации

Графики строились с помощью Microsoft Excel по текстовым файлам. Содержимое такого файла для F1-меры представлено на рисунке 11. На рисунках 12-16 представлены сами графики.



*Рисунок 11. Содержимое файла, содержащего значения времени расчёта F1-меры.*

*Рисунок 12. Графики зависимости времени от количества кластеров.*

*Рисунок 13. Графики зависимости времени от размерности пространства.*

*Рисунок 14. Графики зависимости времени от количества точек.*

*Рисунок 15. Графики зависимости времени вычисления мер сходства от количества кластеров.*

*Рисунок 16. Графики зависимости времени вычисления мер сходства от количества точек.*

# 4. ТЕХНИКО-ЭКОНОМИЧЕСКОЕ ОБОСНОВАНИЕ

## 4.1. Детализированный план работ

1. Составление технического задания
2. Обзор литературы
3. Разработка генератора кластеризованных множеств
4. Реализация методов оценки качества кластеризации
5. Расчёт времени работы реализованных методов в зависимости от различных параметров кластеризованного множества
6. Построение графиков зависимости времени от параметров
7. Оформление пояснительной записки
8. Оформление иллюстративного материала

## 4.2. Расчёт ожидаемой длительности работ

Для расчёта продолжительности каждой работы будем использовать расчётный метод. Для этого воспользуемся следующей формулой:

где – ожидаемая длительность j-ой работы, и – наименьшая и наибольшая по мнению эксперта длительность работы

Произведём расчёты для каждой из работ:

1. Составление технического задания
2. Обзор литературы
3. Разработка генератора кластеризованных множеств
4. Реализация методов оценки качества кластеризации
5. Расчёт времени работы реализованных методов в зависимости от различных параметров кластеризованного множества
6. Построение графиков зависимости времени от параметров
7. Оформление пояснительной записки
8. Оформление иллюстративного материала

## 4.3. Расчёт ставки заработной платы

В качестве месячной заработной платы примем заработную плату инженера, равную 45000 рублей. [4]

Для определения дневной ставки заработной платы разделим заработную плату за месяц на количество рабочих дней в месяце (21 день). В результате получим 2142,858 рубля.

## 4.4. Расчёт расходов на заработную плату

Расходы на основную заработную плату определяются по формуле:

где T – время, затраченное на исследование, С – ставка

Для получения T сложим ожидаемые длительности всех работ. В результате получим 72,6 дня.

Таким образом, рублей

Расходы на дополнительную заработную плату определяются по формуле:

где – расходы на основную заработную плату,

– норматив дополнительной заработной платы, равный 14 %

Подставив необходимые значения в формулу, получим:

рублей

Суммарные расходы на заработную плату составят 177351,5 рублей.

## 4.5. Расчёт отчислений на страховые взносы

Отчисления на страховые взносы на обязательное социальное, пенсионное и медицинское страхование определяются по формуле:

где – основная заработная плата,

– дополнительная заработная плата,

– норматив отчислений на страховые взносы на обязательное социальное, пенсионное и медицинское страхование, равный 30 % [5]

Подставив необходимые значения в формулу, получим:

рублей

## 4.6. Расчёт затрат на содержание и эксплуатацию оборудования

Для выполнения данной работы используется персональный компьютер. Для определения затрат на его эксплуатацию воспользуемся тарифом для потребителей, приравненных к населению. Есть три вида таких тарифов: одноставочный тариф; тариф, дифференцированный по двум зонам суток и тариф, дифференцированный по трём зонам суток. Будем рассматривать тариф, дифференцированным по двум зонам суток. В таком случае размер тарифа в дневную зону будет равен 4,29 руб./кВт\*час.[6]

Компьютер средней мощности потребляет 250 Вт\*час энергии. Переведя данное значение в кВт\*час, получим 0,25 кВт\*час. Умножив данное значение на тариф, получим затраты на эксплуатацию компьютера в течение одного часа: 1,073 рубля. Для того, чтобы найти затраты за день, затраты за час необходимо умножить на 8. В результате получим 8,584 рубля.

Компьютер используется во всех работах, за исключением составления технического задания. Сложив ожидаемые длительности всех работ, за исключением составления технического задания, получим 70 дней. Данное время является временем эксплуатации компьютера. Перемножив его на затраты за день получим общие затраты на содержание и эксплуатацию оборудования. В результате получим 600,88 рублей.

## 4.6. Совокупная величина затрат

Для расчёта совокупной величины затрат, связанных с проведением исследования, занесём рассчитанные значения в таблицу 2.

*Таблица 2.Смета затрат на ВКР.*

|  |  |
| --- | --- |
| Наименование статьи | Сумма, руб |
| Расходы на оплату труда | 177351,5 |
| Отчисления на социальные нужды | 53205,45 |
| Расходы на содержание и эксплуатацию оборудования | 600,88 |
| Итого затрат | 231157,83 |

# ЗАКЛЮЧЕНИЕ

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Gan, Guojun, Chaoqun Ma, Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, ASA, Alexandria, VA, 2007.С
2. А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. - 336 с.
3. Котов Александр, Красильников Николай. Кластеризация данных. [Электронный ресурс]. URL: http://yury.name/internet/02ia-seminar-note.pdf (дата обращения: 16.01.2017).
4. Trud.com. Обзор статистики зарплат профессии инженер в Санкт-Петербурге [Электронный ресурс].

URL: http://sankt-peterburg.trud.com/salary/865/3670.html (дата обращения: 16.01.2017).

1. Главбух. Взносы в ПФР, ФОМС, ФСС. [Электронный ресурс].

URL: http://www.glavbukh.ru/rubrika/157 (дата обращения: 17.01.2017).

1. Петроэлектросбыт. Тарифы на электроэнергию по СПб. [Электронный ресурс]. URL: https://pes.spb.ru/for\_customers/electricity\_tariffs/electricity\_tariffs\_for\_st\_petersburg/ (дата обращения: 17.01.2017).
2. Enargys. Потребление электроэнергии компьютерами разной мощности. [Электронный ресурс].

URL: http://enargys.ru/potreblenie-elektroenergii-kompyuterami/ (дата обращения: 17.01.2017).