**«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 внешние критерии оценки качества кластеризации данных, применимые к различным типам кластеризации. | | | | | | | | | |
| Содержание ВКР:  введение, классификация алгоритмов кластеризации, примеры алгоритмов, критерии оценки качества кластеризации, сравнение критериев кластеризации, 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 | 07.04 – 21.04 |
| 2 | Реализация алгоритмов кластеризации | 22.04 – 10.01 |
| 3 | Реализация методов оценки качества кластеризации | 11.01 – 16.02 |
| 4 | Сравнение методов оценки качества кластеризации | 17.02 – 19.02 |
| 5 | Оформление пояснительной записки | 10.01 – 24.02 |
| 6 | Оформление иллюстративного материала | 24.02 – 26.02 |

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

**РЕФЕРАТ**

Пояснительная записка 51 стр., 20 рис., 5 табл., 7 ист., 0 прил.

Объектом исследования являются внешние критерии оценки качества кластеризации.

Цель работы – реализовать различные внешние критерии оценки качества кластеризации, подходящие для различных типов кластеризации.

Для тестирования методов оценки качества кластеризации были реализованы алгоритмы кластеризации различных типов. Затем были реализованы различные внешние методы оценки качества кластеризации, подходящие к типам реализованных алгоритмов кластеризации. Все алгоритмы кластеризации и критерии качества были реализованы на языке программирования C# в среде разработки Visual Studio. В результате выполнения работы были получены модификации внешних критериев, пригодные к таким типам кластеризации, как итеративная, плотностная, иерархическая, нечёткая и вероятностная кластеризация.

**ABSTRACT**

The work is devoted to realization of different external criterions of clustering quality grade. To testing these criterions the different clustering algorithms was implemented. Then, various external methods for assessing the quality of clustering, suitable for the types of implemented clustering algorithm, were implemented. All clustering algorithms and quality criteria were implemented in the C # programming language in the Visual Studio development environment. As a result of the work, modifications of external criteria were obtained. These modifications are suitable for such types of clustering as iterative, density, hierarchical, fuzzy and probabilistic clustering algorithms.

**содержание**

[ВВЕДЕНИЕ 7](#_Toc480748618)

[1. СОВРЕМЕННОЕ СОСТОЯНИЕ ВОПРОСА 8](#_Toc480748619)

[1.1. Общие сведения 8](#_Toc480748620)

[1.2. История вопроса 8](#_Toc480748621)

[1.3. Современные проблемы 10](#_Toc480748622)

[1.4. Пути решения проблем 11](#_Toc480748623)

[2. КЛАССИФИКАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ 12](#_Toc480748624)

[2.1. Итеративная кластеризация 12](#_Toc480748625)

[2.2. Плотностная кластеризация 12](#_Toc480748626)

[2.3. Иерархическая кластеризация 12](#_Toc480748627)

[2.4. Нечёткая кластеризация 12](#_Toc480748628)

[2.5. Вероятностная (статистическая) кластеризация 13](#_Toc480748629)

[3. ПРИМЕРЫ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ 14](#_Toc480748630)

[3.1. Пример итеративной кластеризации: k-means 14](#_Toc480748631)

[3.2. Пример плотностной кластеризации: DBSCAN 14](#_Toc480748632)

[3.3. Пример иерархической кластеризации: метод ближайшего соседа 14](#_Toc480748633)

[3.4. Пример нечёткой кластеризации: c-means 14](#_Toc480748634)

[3.5. Пример вероятностной кластеризации: EM-алгоритм 14](#_Toc480748635)

[4. КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ 19](#_Toc480748636)

[4.1. Критерии оценки качества кластеризации, взятые за основу для создания модификаций для различных типов кластеризации 19](#_Toc480748637)

[2.3. Расчёт времени работы методов 31](#_Toc480748638)

[4. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ 33](#_Toc480748639)

[3.1. Значения индексов оценки качества кластеризации низкого, среднего и высокого качества 33](#_Toc480748640)

[3.2. Значения индексов оценки качества кластеризации, свидетельствующие о том, что кластеризация проведена качественно 34](#_Toc480748641)

[3.3. Графики времени выполнения методов оценки качества кластеризации 39](#_Toc480748642)

[3.4. Определение наилучшего метода оценки качества кластеризации 45](#_Toc480748643)

[5. ТЕХНИКО-ЭКОНОМИЧЕСКОЕ ОБОСНОВАНИЕ 47](#_Toc480748644)

[4.1. Детализированный план работ 47](#_Toc480748645)

[4.2. Расчёт ожидаемой длительности работ 47](#_Toc480748646)

[4.3. Расчёт ставки заработной платы 48](#_Toc480748647)

[4.4. Расчёт расходов на заработную плату 48](#_Toc480748648)

[4.5. Расчёт отчислений на страховые взносы 49](#_Toc480748649)

[4.6. Расчёт затрат на содержание и эксплуатацию оборудования 49](#_Toc480748650)

[4.6. Совокупная величина затрат 50](#_Toc480748651)

[ЗАКЛЮЧЕНИЕ 51](#_Toc480748652)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 52](#_Toc480748653)

# ВВЕДЕНИЕ

Актуальность данной работы определяется тем, что существует необходимость определять корректность разбиения множеств на кластеры с целью проверки того, что полученному результату можно доверять. На сегодняшний день существуют различные коэффициенты, расчёт которых позволяет оценить качество кластеризации. Однако они разработаны для оценки качества итеративной кластеризации, и не подходят для оценки качества кластеризации других типов. Основной целью работы является разработка модификаций методов, пригодных к использованию на различных типах кластеризации. В результате работы различные модификации методов будут реализованы и протестированы.

# 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. КЛАССИФИКАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

## 2.1. Итеративная кластеризация

Итеративные методы кластеризации формируют кластеры до тех пор, пока не будет выполнено правило остановки. Данные методы могут использовать два подхода. Первый заключается в определении кластеров как наиболее плотных областей в пространстве кластеризуемых объектов, т.е. кластера располагаются в местах, где имеется “большое сгущение точек”. Второй подход заключается в минимизации меры различия объектов.

## 2.2. Плотностная кластеризация

Плотностные алгоритмы кластеризации позволяют разбивать исходное множество на кластеры произвольной формы. Большинство других алгоритмов создают кластеры, близкие по форме к сферическим, так как минимизируют расстояние до центра кластера. Идея алгоритмов плотностной кластеризации состоит в том, что внутри каждого кластера наблюдается типичная плотность объектов, которая заметно выше, чем плотность снаружи кластера, а также плотность в областях с шумом ниже плотности любого из кластеров. Для оценки плотности обычно для каждой точки вводится окрестность определённого радиуса, после чего для каждой окрестности определяется количество точек, попавших в окрестность.

## 2.3. Иерархическая кластеризация

Алгоритмы иерархической кластеризации используются в тех случаях, когда предполагается наличие вложенных групп (кластеров различного порядка). Алгоритмы иерархической кластеризации можно поделить на две группы: агломеративные и дивизимные. Агломеративные алгоритмы объединяют мелкие кластеры в более крупные. Дивизимные алгоритмы, наоборот, основаны на дроблении кластеров на более мелкие кластеры.

## 2.4. Нечёткая кластеризация

Нечёткие алгоритмы кластеризации позволяют одному и тому же объекту принадлежать нескольким кластерам. При этом каждому объекту ставится в соответствие набор вещественных чисел, показывающих степень принадлежности данного объекта к каждому из кластеров.

## 2.5. Вероятностная (статистическая) кластеризация

Вероятностные алгоритмы основаны на предположении о том, что кластеры описываются некоторым семейство вероятностных распределений. Тогда задача кластеризации сводится к разделению смеси распределений. В качестве распределений чаше всего берут сферические гауссовские распределения.

# 3. ПРИМЕРЫ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

Для тестирования разработанных модификаций внешних критериев оценки качества кластеризации были разработаны алгоритмы кластеризации, принадлежащие к различным типам. Данные алгоритмы представлены ниже.

## 3.1. Пример итеративной кластеризации: k-means

Для

## 3.2. Пример плотностной кластеризации: DBSCAN

Для

## 3.3. Пример иерархической кластеризации: метод ближайшего соседа

Для

## 3.4. Пример нечёткой кластеризации: c-means

Для

## 3.5. Пример вероятностной кластеризации: EM-алгоритм

Название алгоритма происходит от слов “*expectation-maximization*”, что переводится как “*ожидание-максимизация*”. В основе идеи EM-алгоритма лежит предположение о том, что данные в каждом кластере подчиняются закону нормального распределения. С учетом этого предположения можно определить математическое ожидание и дисперсию, которые соответствуют закону распределения элементов в кластере. Тогда задача будет заключаться в определении вероятностей принадлежности наблюдения к каждому кластеру. Очевидно, что наблюдение должно быть отнесено к тому кластеру, для которого данная вероятность выше.

Мы предполагаем, что в каждом кластере элементы имеют многомерное нормальное распределение. Его плотность для *q*-мерного вектора *x*=(**,** **,** …**,** ) может быть записана в виде:

где - ковариационная матрица размером *q*×*q*,  
*μ - q*-мерный вектор математических ожиданий.

Алгоритм предполагает, что данные подчиняются смеси многомерных нормальных распределений, а их плотность равна

где - вес i-го кластера;

- плотность распределения i-го кластера.

Алгоритм завершает свою работу, когда разница значений логарифмической функции правдоподобия на текущем шаге и на предыдущем меньше заданного допустимого отклонения. Логарифмическая функция правдоподобия, вычисляется по формуле:

где – количество объектов,

общая плотность распределения.

Алгоритм получает на вход число кластеров k, множество из n объектов q-мерного пространства Y={}, допустимое отклонение для логарифмического правдоподобия , максимальное число итераций Q.

В процессе работы алгоритм формирует следующие матрицы:

1. Матрица математических ожиданий кластерных распределений С размером q\*k,
2. Матрица ковариаций кластерных распределений R размером q\*q,
3. Матрица весов кластеров W размером k\*1.
4. Матрица вероятностей принадлежности объектов к кластерам X размером n\*k.

EM-алгоритм работает следующим образом:

1. Инициализация: установка начальных значений матриц C, R и W.

Обычно матрица С заполняется случайным образом либо на основе информации об исходном множестве, матрица R считается единичной, а матрица W заполняется значениями .

1. Пока изменение логарифмического правдоподобия *∆llh* ≥ ε и не достигнуто максимальное число итераций *Q*, выполнить шаги E и M.

Далее представлен псевдокод шагов E и M.

Шаг E:

Для i, изменяющегося от 1 до n

Для j, изменяющегося от 1 до k

Конец цикла по j

Конец цикла по i

Шаг M:

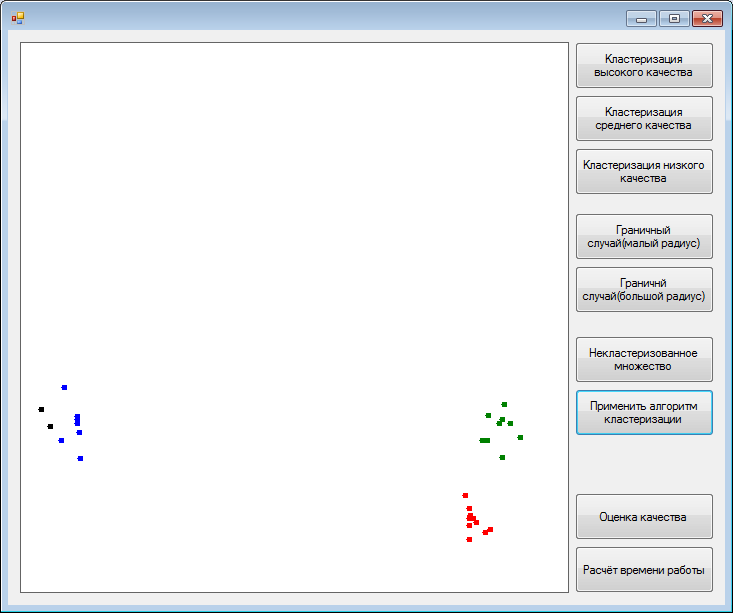
Для j, изменяющегося от 1 до k

Для i, изменяющегося от 1 до n

Конец цикла по i

Конец цикла по j

Принадлежность к кластерам определяется на основе матрицы вероятностей X: объект будет отнесён к тому кластеру, вероятность принадлежность к которому больше. Так как вероятности принадлежности некоторых объектов ко всем кластерам в процессе запуска реализованного алгоритма оказались равными нулю, такие объекты было решено не относить ни к одному кластеру. Результат работы реализованного алгоритма представлен на рисунке 1.



*Рисунок 1. Результат работы реализованного EM-алгоритма .*

Среди преимуществ EM-алгоритма можно выделить следующие:

* Мощная статистическая основа
* Линейное увеличение сложности при росте объема данных
* Устойчивость к шумам и пропускам в данных
* Возможность построения желаемого числа кластеров
* Быстрая сходимость при удачной инициализации

Однако алгоритм имеет и ряд недостатков:

* Предположение о нормальности всех измерений данных не всегда выполняется.
* При неудачной инициализации сходимость алгоритма может оказаться медленной. Кроме этого, алгоритм может остановиться в локальном минимуме и дать квазиоптимальное решение.

# 4. КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

В данной работе исследуются исключительно внешние критерии оценки качества кластеризации. Под внешними критериями понимаются критерии, использующую информацию о том, как должна быть проведена кластеризация. Данная информация представлена в виде набора классов. Класс представляет собой группу объектов, которые должны принадлежать одному кластеру.

## 4.1. Критерии оценки качества кластеризации, взятые за основу для создания модификаций для различных типов кластеризации

1. F1-мера(F1-meassure)

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:

Значения данных индексов лежат в интервале от нуля до единицы, при этом в случае наилучшей кластеризации значения равны единице, а в случае наихудшей – нулю.

## 4.2. Модификации для итеративной кластеризации

Для каждого из методов оценки качества кластеризации было рассчитано

## 4.3. Модификации для плотностной кластеризации

Для каждого из методов оценки качества кластеризации было рассчитано

## 4.4. Модификации для иерархической кластеризации

Для каждого из методов оценки качества кластеризации было рассчитано

## 4.5. Модификации для нечёткой кластеризации

Для каждого из методов оценки качества кластеризации было рассчитано

## 4.6. Модификации для вероятностной кластеризации

Особенностью реализованного в рамках данной работы алгоритма вероятностной кластеризации – EM-алгоритма, является то, что некоторые объекты не были отнесены ни к одному кластеру.

F1-мера может быть применена к алгоритмам вероятностной кластеризации без каких либо изменений.

Для использования индексов Rand, Jaccard и FM необходимо переопределить обозначения SS, SD, DS и DD, после чего данные индексы можно будет использовать для оценки качества вероятностной кластеризации.

Под SS будем понимать количество пар объектов, которые удовлетворяют нижеперечисленным условиям:

1. Объекты пары принадлежат одному кластеру либо оба объекта не были отнесены ни к одному кластеру.
2. Объекты пары принадлежат одному классу либо оба объекта не были отнесены ни к одному классу.

Под SD будем понимать количество пар объектов, которые удовлетворяют нижеперечисленным условиям:

1. Объекты пары принадлежат одному кластеру либо оба объекта не были отнесены ни к одному кластеру.
2. Объекты пары принадлежат разным классам либо один из объектов не был отнесён ни к одному классу, а второй объект был отнесён к какому-либо классу.

Под DS будем понимать количество пар объектов, которые удовлетворяют нижеперечисленным условиям:

1. Объекты пары принадлежат разным кластерам либо один из объектов не был отнесён ни к одному кластеру, а второй объект был отнесён к какому-либо кластеру.
2. Объекты пары принадлежат одному классу либо оба объекта не были отнесены ни к одному классу.

Под DD будем понимать количество пар объектов, которые удовлетворяют нижеперечисленным условиям:

1. Объекты пары принадлежат разным кластерам либо один из объектов не был отнесён ни к одному кластеру, а второй объект был отнесён к какому-либо кластеру.
2. Объекты пары принадлежат разным классам либо один из объектов не был отнесён ни к одному классу, а второй объект был отнесён к какому-либо классу.

# 4. 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 |
| Индекс Маулика-Бандиопадия | 943683,87 | 29842,09 | 671,58 |

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, нормализованная Hubert’s Г статистика и индекс Маулика-Бандиопадия.

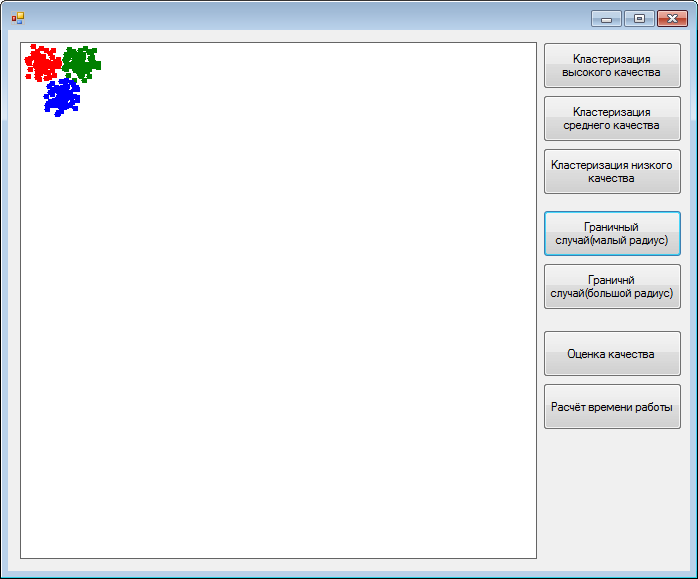
## 3.2. Значения индексов оценки качества кластеризации, свидетельствующие о том, что кластеризация проведена качественно

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

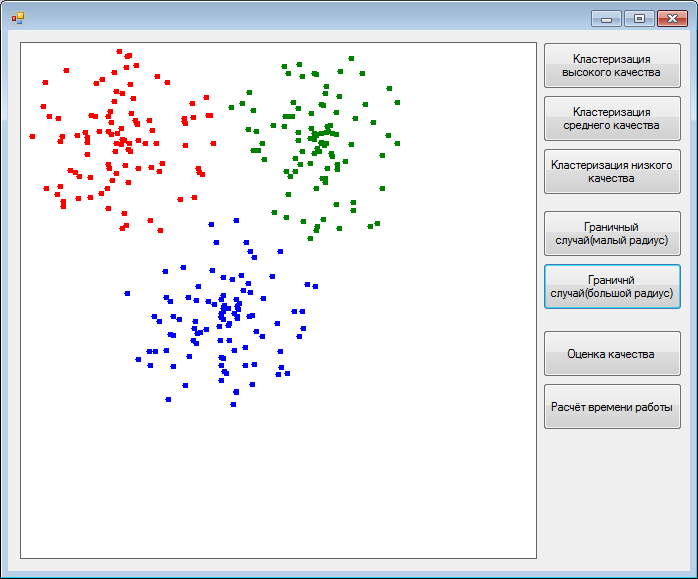
1. Индекс RMSSTD, Hubert’s Г статистика, модифицированная Hubert’s Г статистика

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.*

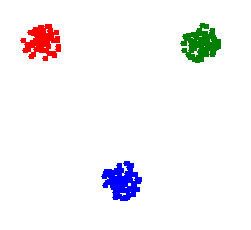
В качестве примера, подтверждающего это, можно рассмотреть множество, представленное на рисунке 7. Значение индекса RMSSTD для данного множества равно 6.



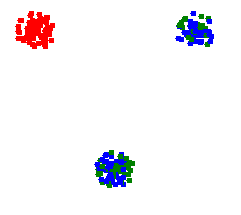
*Рисунок 8. Пример кластеризованного множества с большим размером кластеров.*

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, нормализованная Hubert’s Г статистика и индекс Маулика-Бандиопадия. На каждом фиксированном множестве большее значение вышеперечисленных Hubert’s Г статистик и индекса Маулика-Бандиопадия 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 |

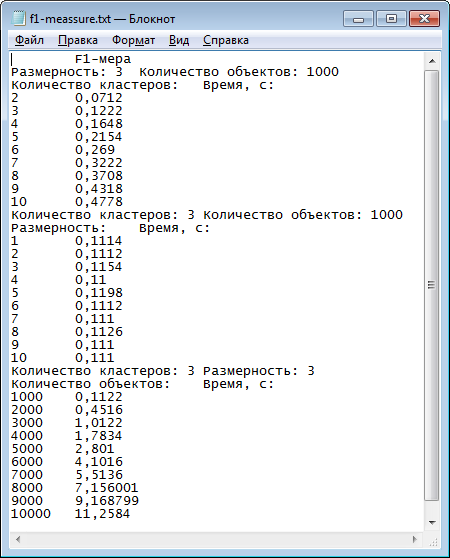
При увеличении качества кластеризации все вышеперечисленные индексы монотонно изменяются. При этом часть индексов монотонно возрастает, а часть – монотонно убывает. Для того чтобы определить поведение индекса при изменении качества кластеризации, можно обратиться к таблице 1, в которой перечислены значения 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 | <0,6 |
| 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 12-20 представлены сами графики. В связи с тем, что различия между значениями времени выполнения различных методов достаточно велики, размещение всех графиков на одном рисунке приводит к тому, что различия между наиболее быстрыми методами становятся незаметными. Поэтому графики наиболее медленных методов были размещены на отдельных рисунках.



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

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

*Рисунок 13. Графики зависимости времени от количества кластеров для остальных индексов с рисунка 12.*

*Рисунок 14. Графики зависимости времени от количества кластеров для более быстрых методов с рисунка 13.*

*Рисунок 15. Графики зависимости времени от размерности пространства.*

*Рисунок 16. Графики зависимости времени от размерности пространства для остальных индексов с рисунка 15.*

*Рисунок 17. Графики зависимости времени от размерности пространства для более быстрых методов с рисунка 16.*

*Рисунок 18. Графики зависимости времени от количества точек.*

*Рисунок 19. Графики зависимости времени от количества точек для остальных индексов с рисунка 18.*

*Рисунок 20. Графики зависимости времени от количества точек для более быстрых методов с рисунка 19.*

## 3.4. Определение наилучшего метода оценки качества кластеризации

В качестве наилучшего метода выберем самый быстрый. Для этого обратимся к графикам из предыдущего подраздела. По графика видно, что самым быстрым методом является расчёт индекса RMSSTD. Однако данный индекс обладает недостатком: по его значению нельзя определить, качественно ли проведена кластеризация. Расчёт данного индекса может помочь найти наилучший способ разбиения какого-либо конкретного множества на кластеры, так как на любом фиксированном множестве меньшее значение индекса RMSSTD будет соответствовать более качественной кластеризации. Поэтому расчёт индекса RMSSTD можно применять только в ситуации, когда какое-либо множество кластеризуется различными способами и требуется опредлить, какой из способов даёт лучший результат. Например, если требуется определить оптимальное число кластеров, на которые следует разбить множество.

Так как индекс RMSSTD позволяет только сравнивать кластеризованные множества между собой, найдём лучший из индексов, значения которых позволяют сказать, качественно ли проведена кластеризация. Для этого обратимся к графикам из предыдущего подраздела. Лучшим временем среди графиков зависимости от количества кластеров обладает индекс RS(за исключением уже рассмотренного индекса RMSSTD). Лучшим временем среди графиков зависимостей от размерности и количества объектов является индекс Девиса-Болдуина. Чтобы понять, какой из индексов лучше, занесём информацию о местах графиков времени вычисления индексов в таблицу 4.

*Таблица 4.Места графиков времени расчёта индекса RS и индекса Девиса-Болдуина.*

|  |  |  |  |
| --- | --- | --- | --- |
| Индексы | Место среди графиков зависимости от количества кластеров | Место среди графиков зависимости от размерности пространства | Место среди графиков зависимости от количества объектов |
| Индекс RS | 2 | 3 | 4 |
| Индекс Девиса-Болдуина | 4 | 2 | 2 |

По таблице 4 видно, что индекс Девиса-Болдуина занимает второе место в двух случаях и четвёртое в одном. Индекс RS также занимает четвёртое место в одном из случаев. Но в одном из случаев индекс RS занимает третье место. Поэтому индекс Девиса-Болдуина следует считать наилучшим из индексов, по значениям которых можно понять, качественно ли проведена кластеризация.

Теперь найдём лучшую меру сходства. Меры сходства позволяют оценить качество кластеризации на основе сравнения с информацией о том, как должна быть проведена кластеризация. В отличие от большинства индексов, меры сходства оценивают не структуру кластеров, а правильность кластеризации с точки зрения человека. Для работы мер сходства необходима информация о том, к какому кластеру следовало бы отнести каждый объект. В данной работе были рассмотрены такие меры сходства, как F1-мера, индексы Rand, Jaccard, FM. По графикам видно, что наиболее быстрыми в вычислении индексами являются индексы Jaccard и FM. Так как при вычислении индекса Jaccard используется меньшее количество арифметических операций, предпочтительней использовать именно его.

# 5. 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 5.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

В результате выполнения работы были реализованы различные методы оценки качества кластеризации, основанные на вычислении индексов. Затем были найдены интервалы, принадлежность индексов к которым означает, что кластеризация проведена качественно. Также были построены графики зависимости времени выполнения реализованных методов от количества кластеров, размерности пространства и количества объектов множества. На основании графиков был проведён сравнительный анализ методов. Результаты работы соответствуют поставленной цели. Так как работа содержит описание алгоритмов работы методов, а также графики, на которых изображено время работы алгоритмов в зависимости от различных параметров, результаты работы можно использовать для выбора метода оценки качества кластеризации. Информация о том, при каком значении индекса кластеризацию следует считать успешной, позволяет оценить качество кластеризации по рассчитанному значению индекса. Исходя из графиков и особенностей методов, можно сделать вывод, что наиболее предпочтительным является расчёт индекса Девиса-Болдуина. Для того, чтобы можно было считать кластеризацию качественной, значение индекса Девиса-Болдуина должно быть меньше 0,6. В случае, если требуется сравнить несколько различных способов кластеризации одного и того же множества, лучше использовать не индекс Девиса-Болдуина, а более быстрый в вычислении индекс RMSSTD. Меньшее значение индекса RMSSTD будет говорить о том, что кластеризация проведена более качественна. Если требуется оценить качество кластеризации не на основе структуры кластеризованного множества, а на основе степени соответствия результата кластеризации ожидаемому результату, следует использовать так называемые меры сходства. Для вычисления мер сходства используется информация о группах среди исходного множества, на которые следует разделить множество. Самой быстрой в вычислении мерой сходства оказался индекс Jaccard.

# 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
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).