**«Санкт-Петербургский государственный электротехнический университет**

**«ЛЭТИ» им. В.И.Ульянова (Ленина)»**

**(СПбГЭТУ «ЛЭТИ»)**

|  |  |  |
| --- | --- | --- |
| **Направление** | 01.03.02 – Прикладная математика и информатика | |
| **Профиль** | Без профиля | |
| **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 | к.т.н., доцент |  |  | Жукова Н.А. |
|  | *(Уч. степень, уч. звание)* | *подпись* |  |  |
| Консультанты | к.э.н., доцент |  |  | Ширяева Т.П. |
|  | *(Уч. степень, уч. звание)* | *подпись* |  |  |
|  | к.т.н., доцент |  |  | Яценко И.В. |
|  | *(Уч. степень, уч. звание)* | *подпись* |  |  |

Санкт-Петербург

2017**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):  Разработать и реализовать внешние критерии оценки качества кластеризации данных, применимые к различным типам кластеризации. | | | | | | | | | | |
| Содержание ВКР:  введение, анализ существующих алгоритмов кластеризации и критериев оценки качества кластеризации, разработка критериев оценки качества для групп алгоритмов кластерного анализа, реализация критериев оценки качества для групп алгоритмов кластерного анализа, экспериментальные исследования, 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. | | | | |
|  | |  | | *подпись* | |  |  | | | | |
| Руководитель | | к.т.н., доцент | |  | |  | Жукова Н.А. | | | | |
|  | | *(Уч. степень, уч. звание)* | | *подпись* | |  |  | | | | |
| Консультант | | к.э.н., доцент | |  | |  | Ширяева Т.П. | | | | |
|  | | *(Уч. степень, уч. звание)* | | *подпись* | |  |  | | | | |

**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 | 24.04 – 27.04 |
| 2 | Реализация алгоритмов кластеризации | 27.04 – 20.05 |
| 3 | Реализация методов оценки качества кластеризации | 30.04 – 21.05 |
| 4 | Сравнение методов оценки качества кластеризации | 22.05 – 23.05 |
| 5 | Оформление пояснительной записки | 02.05 – 23.05 |
| 6 | Оформление иллюстративного материала | 24.02 – 26.02 |
| 7 | Предзащита | 10.06 |

|  |  |  |
| --- | --- | --- |
| 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.

КЛАСТЕРИЗАЦИЯ, ОЦЕНКА КАЧЕСТВА КЛАСТЕРИЗАЦИИ, КЛАССИФИКАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ, ВНЕШНИЕ КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА

Объектом исследования являются внешние критерии оценки качества кластеризации.

Целью работы является разработка и реализация модификаций внешних критериев оценки качества, пригодных к использованию на различных типах 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 модельная кластеризация.

**ABSTRACT**

The work is devoted to development and realization of different external criterions of clustering quality grade. To testing these criterions the different clustering algorithms was implemented. Then, various modifications of external methods for assessing the quality of clustering, suitable for the types of implemented clustering algorithm, were developed and 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](#_Toc484276800)

[1. АНАЛИЗ СУЩЕСТВУЮЩИХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ И КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ 8](#_Toc484276801)

[1.1. Общие сведения 8](#_Toc484276802)

[1.2. История кластеризации 8](#_Toc484276803)

[1.3. Классификация алгоритмов кластеризации 10](#_Toc484276804)

[1.4. Примеры алгоритмов кластеризации 11](#_Toc484276805)

[1.5. Критерии оценки качества кластеризации 19](#_Toc484276806)

[1.6. Выводы 22](#_Toc484276807)

[2. РАЗРАБОТКА КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА 25](#_Toc484276808)

[2.1. Модификации для итеративной кластеризации 25](#_Toc484276809)

[2.2. Модификации для плотностной кластеризации 25](#_Toc484276810)

[2.3. Модификации для иерархической кластеризации 26](#_Toc484276811)

[2.4. Модификации для нечёткой кластеризации 29](#_Toc484276812)

[2.5. Модификации для модельной кластеризации 31](#_Toc484276813)

[2.6. Сравнение модификаций 31](#_Toc484276814)

[3. РЕАЛИЗАЦИЯ КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА 34](#_Toc484276815)

[3.1. Описание генератора исходных множеств 34](#_Toc484276816)

[3.2. Реализация алгоритмов кластеризации 38](#_Toc484276817)

[3.3. Представление информации о классах 43](#_Toc484276818)

[3.4. Реализация модификаций внешних критериев оценки качества 44](#_Toc484276819)

[4. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ 46](#_Toc484276820)

[4.1. Описание экспериментов 46](#_Toc484276821)

[4.2. Эксперименты для модификаций для плотностной кластеризации. 46](#_Toc484276822)

[4.3. Эксперименты для модификаций для иерархической кластеризации. 47](#_Toc484276823)

[4.4. Эксперименты для модификаций для нечёткой кластеризации. 48](#_Toc484276824)

[5. ТЕХНИКО-ЭКОНОМИЧЕСКОЕ ОБОСНОВАНИЕ 50](#_Toc484276825)

[5.1. Концепция экономического обоснования 50](#_Toc484276826)

[5.2. Трудоёмкость и календарный план выполнения проекта 50](#_Toc484276827)

[5.3. Смета затрат на выполнение проекта 51](#_Toc484276828)

[5.4. Себестоимость выполнения ВКР 54](#_Toc484276829)

[5.5. Выводы 54](#_Toc484276830)

[ЗАКЛЮЧЕНИЕ 55](#_Toc484276831)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 56](#_Toc484276832)

# ВВЕДЕНИЕ

Объектом исследования являются внешние критерии оценки качества кластеризации.

Предметом исследования является применимость внешних критериев к алгоритмам кластеризации различных типов.

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.

Целью работы является разработка модификаций методов, пригодных к использованию на различных типах кластеризации. В результате работы различные модификации методов будут реализованы и протестированы.

В рамках выполнения работы будут выполнены следующие задачи:

1. Реализация алгоритмов кластеризации различных типов.
2. Реализация внешних критериев оценки качества кластеризации.
3. Разработка и реализация модификаций внешних критериев, пригодных для оценки качества результатов кластеризации различных типов.
4. Сравнительный анализ разработанных модификаций внешних критериев и исходных критериев.

# 1. АНАЛИЗ СУЩЕСТВУЮЩИХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ И КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

## 1.1. Общие сведения

Кластеризация (кластерный анализ) – совокупность методов, выполняющих выделение групп объектов, называемых кластерами, таким образом, чтобы объекты одного кластера были похожи, а объекты разных кластеров существенно различались.[1]

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 кластеризации

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. Классификация алгоритмов кластеризации

1. Итеративная кластеризация[6].

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.[6]

1. Плотностная кластеризация[11].

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.

1. Иерархическая кластеризация[2].

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, начиная с исходных элементов. Дивизимные алгоритмы начинают с одного кластера, содержащего все элементы, после чего расщепляют кластеры на более мелкие.[2]

1. Нечёткая кластеризация[3].

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.[1]

1. Модельная кластеризация[11].

Модельные алгоритмы основаны на предположении о том, что кластеры описываются некоторой математической моделью. Тогда задача кластеризации сводится к максимизации сходства модели и набора кластеров. При этом часто используют аппарат математической статистики.

## 1.4. Примеры алгоритмов кластеризации

Далее приведены примеры алгоритмов кластеризации к каждому типу согласно вышепредставленной классификацией алгоритмов.

1. Пример итеративной кластеризации: k-means

Алгоритм k-means стремится минимизировать суммарное квадратичное отклонение объектов кластеров от центров этих кластеров:

где k – количество кластеров, – i-ый кластер, – вектор координат j-ого объекта, – вектор координат центра i-ого кластера.

Достижение глобального минимума V не гарантируется, так как алгоритм может завершить свою работу в одном из локальных минимумов. Результат кластеризации зависит от выбора начальных центров кластеров. Число кластеров должно быть известно заранее.

Алгоритм начинает свою работу с выбора начальных центров масс кластеров. Они могут быть выбраны любым образом, например случайно либо на основе анализа исходных данных.

Затем на каждой итерации объекты разбиваются на кластеры в соответствии с тем, какой объект оказался ближе к центру масс по выбранной метрике, после чего для каждого кластера перевычисляется центр масс.

Алгоритм завершается, когда при перевычислении центров кластеров не происходит их изменения.

1. Пример плотностной кластеризации: DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise - основанная на плотности пространственная кластеризация множеств с шумом) – алгоритм кластеризации, созданный в 1996 году Мартином Эстером, Хансом-Питером Кригелем, Йоргом Сандером и Сяоэй Ксу. DBSCAN проводит кластеризацию, основываясь на плотности: группируются те точки, которые расположены в плотных областях (областях, в которых у каждой точки много соседей). При этом точки, лежащие в областях с низкой плотностью, помечаются как шум.[7]

DBSCAN делит точки кластеризуемого множества на три группы: основные точки, достижимые точки и шум.

Точка p называется основной точкой, если по крайней мере minPts точек находятся на расстоянии не более от точки p (включая саму точку p). Данные точки называются непосредственно достижимыми из p.

Параметры minPts и выбираются пользователем.

Точка q достижима из p, если существует путь , , …, , где и , такой, что непосредственно достижима из .

Все точки, недостижимые из какой-либо точки называются шумом.

Кластеры формируются вокруг основных точек и состоят из всех точек, достижимых из основной.

Алгоритм DBSCAN начинается с произвольной начальной точки. Определяется количество точек в -окрестности начальной точки. Если это количество оказалось меньше minPts, точка помечается как шум, при этом в дальнейшем эта точка может быть отнесена к какому-либо кластеру, если окажется в -окрестности основной точки.

В противном случае начальная точка и её -окрестность относится к первому кластеру. Также к первому кластеру относятся точки из -окрестностей точек, отнесённых к первому кластеру, при условии, что эти -окрестности содержат не меньше minPts точек.

Затем алгоритм повторяется для другой точки, которая не была помечена как шум и не была отнесена к какому-либо кластеру, при этом точки будут относится не к первому кластеру, а ко второму.

Алгоритм продолжает повторяться до тех пор, пока все точки не будут помечены как шум или отнесены к какому-либо кластеру, при этом на каждом шаге точки добавляются к следующему кластеру.

Псевдокод алгоритма:

DBSCAN(исходное множество, eps, MinPts)

{

**номер\_кластера**= 0

для каждой точки P из исходного множества

{

если P была посещена

перейти к следующему шагу цикла

пометить P как посещённую точку

**соседние\_точки** = НайтиТочки(P, eps)

если размер(**соседние\_точки**) < MinPts

пометить P как шум

иначе

{

**номер\_кластера** = **номер\_кластера**+1

СформироватьКластер(P, **соседние\_точки**, **номер\_кластера**, eps, MinPts)

}

}

}

СформироватьКластер(P, соседние\_точки, номер\_кластера, eps, MinPts)

{

добавить P в кластер с номером **номер\_кластера**

для каждой точки P' из множества **соседние\_точки**

{

если P' не посещена

{

пометить P' как посещённую точку

**соседние\_точки'** = НайтиТочки(P', eps)

**if** размер(**соседние\_точки'**) >= MinPts

**соседние\_точки** = **соседние\_точки** **соседние\_точки'**

}

если P' не отнесена к какому-либо кластеру

добавить P' к кластеру с номером **номер\_кластера**

}

}

НайтиТочки(P, eps)

вернуть все точки в eps-окрестности P (включая P)

Среди преимуществ DBSCAN можно выделить следующие:

* + Нет необходимости указывать количество кластеров
  + Возможность находить кластеры произвольной формы
  + Возможность поиска шума

Однако алгоритм имеет и ряд недостатков:

* + Алгоритм не может быть применён к кластерам с большими различиями в плотностях, так как параметры и minPts не могут быть выбраны надлежашим образом для всех кластеров
  + Для корректной работы алгоритма необходимо правильно выбрать параметры и minPts, основываясь на исходном множестве

1. Пример иерархической кластеризации: метод ближайшего соседа (метод одиночной связи)

Перед началом работы алгоритма каждый объект представляют собой кластер. Затем строится матрица расстояний между кластерами. Расстояние между двумя кластерами считается равным минимальному расстоянию между объектами, один из которых принадлежит первому кластеру, а другой – второму кластеру. В матрице ищется минимальное расстояние, не равное нулю, соответствующее расстоянию между двумя наиболее близкими кластерами. Данные кластеры объединяются в один кластер, после чего матрица расстояний перерасчитывается. Объединение кластеров продолжается до тех пор, пока все кластера не будут объединены в один.[8] Результатом работы алгоритма является последовательностей объединений кластеров, которая может быть представлена в виде дерева, называемого дендрограммой.

1. Пример нечёткой кластеризации: с-means

Алгоритм c-means минимизирует следующую функцию :

где m – любое положительное число, большее 1; N – количество объектов; С – количество кластеров; – степень принадлежности i-ого объекта j-ому кластеру; – i-ый объект; – центр j-ого кластера.

На каждом шаге алгоритма обновляются степени принадлежности и центры кластеров. Шаги продолжаются до тех пор, пока максимальная разница между степенями принадлежности, рассчитанными на текущем шаге, и степенями принадлежности на предыдущем шаге, не окажется меньше некоторого выбранного малого числа Алгоритм находит локальный минимум функции .

Алгоритм состоит из следующих шагов:

1. Инициализация матрицы степеней принадлежности U. Элементы матрицы должны лежать в интервале от нуля до единицы, при этом сумма элементов в каждой строке матрицы должна быть равна единице. Матрица может быть заполнена как- угодно, например случайным образом.
2. Расчёт центров кластеров по формуле:
3. Обновление матрицы степеней принадлежности по формуле:
4. Если , то алгоритм завершает свою работу, иначе осуществляется переход на шаг 2.

На выходе алгоритма будет получена матрица степеней принадлежности объектов к кластерам U. Каждая i-ая строка такой матрицы содержит степени принадлежности i-ого объекта ко всем кластерам.[9]

1. Пример модельной 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.

Среди преимуществ EM-алгоритма можно выделить следующие:

* + 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.[10]

## 1.5. 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 представляется в виде набора классов. Класс представляет собой группу объектов, которые должны принадлежать одному кластеру.

За основу для создания модификаций для различных типов кластеризации были взяты следующие внешние критерии оценки качества:

1. F1-мера(F1-meassure)[12]

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(Rand statistic), Jaccard(Jaccard coefficient), FM(Folkes and Mallows index)[1]

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

В результате анализа литературы была обнаружена модификация индекса Rand для нечёткой кластеризации[13]. Данная модификация производит сравнение степеней принадлежности объектов к кластерам со степенями принадлежности объектов к классам.

Для определения сходства между объектами согласно множеству кластеров P определяется величина , заданная следующей формулой:

где – вектор степеней принадлежности объекта x ко всем кластерам, - вектор степеней принадлежности объекта ко всем кластерам, - метрика L1 для разности векторов и .

Для определения сходства между объектами согласно множеству классов Q определяется величина , заданная следующей формулой:

где – вектор степеней принадлежности объекта x ко всем классам, - вектор степеней принадлежности объекта ко всем классам, - метрика L1 для разности векторов и .

Далее вводится понятие меры удалённости d(P, Q), определяемое следующей формулой:

где P – множество кластеров, Q – множество классов, n – количество объектов исходного множества.

Тогда предложенная автором модификация рассчитывается по следующей формуле:

.

## 1.6. Выводы

В результате анализа литературы было выяснено, что большинство внешних критериев были разработаны для итеративной кластеризации. Также была обнаружена одна модификация внешнего критерия, разработанная для нечёткой кластеризации.

В качестве основы для модификаций были выбраны такие внешние меры, как F1-мера, индексы Rand, Jaccard и FM. Все они были разработаны для итеративной кластеризации. Также все эти критерии равны 1 в случае наилучшей кластеризации и 0 в случае наихудшей кластеризации.

Итеративная кластеризация предполагает, что каждый объект должен быть отнесён ровно к одному кластеру. Плотностная и модельная кластеризации допускают наличие шума – множества объектов, которые не принадлежат ни к одному кластеру. F1-мера вычисляется по формуле . Наличие шума будет приводить к увеличению N, при этом остальные параметры формулы не будут меняться. Таким образом, в случае наилучшей кластеризации значение F1-меры будет меньше 1, при условии, что во множестве присутствует шум. Так как F1-мера в случае наилучшей кластеризации должна равняться 1, наличие шума приводит к некорректному результату. Поэтому F1-мера не применима к плотностной и модельной кластеризации.

Иерархическая и нечёткая кластеризации допускают принадлежность одного объекта к нескольким кластерам. Рассмотрим расчёт F1-меры для случая наилучшей кластеризации при условии, что некоторые объекты принадлежат более чем к одному классу. Принадлежность объектов к нескольким классам приведёт к увеличению для классов, которым принадлежат эти объекты. При этом значение N и не изменится. Это приведёт к тому, что значение F1-меры станут больше 1. Так как F1-мера должна лежать в интервале от 0 до 1, наличие объектов, принадлежащих более чем одному кластеру, приводит к некорректному результату, что говорит о неприменимости F1-меры к иерархической и нечёткой кластеризации.

Расчёты индексов Rand, Jaccard и FM основаны на подсчёте количества пар, удовлетворяющих некоторым условиям. Данные количества обозначаются следующим образом:

* + 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.

При этом SS и DD соответствует парам, правильно отнесённым к кластерам, а SD и DS соответствуют ошибкам. Таким образом, в случае наилучшей кластеризации SD и DS должны равняться нулю.

Рассмотрим случай плотностной или модельной кластеризации. Пусть имеется некоторая пара объектов (). Пусть принадлежит какому-либо классу, а не принадлежит ни к одному классу. Пусть и не принадлежат ни к одному из кластеров. Данная пара является ошибочной, так как принадлежность к классам не совпадает с принадлежностью к кластерам. Но при этом она не будет учтена ни в SD, ни в DS. Это приведёт к пропуску ошибки, что приведёт к тому, что индексы Rand, Jaccard и FM окажутся больше, чем должны быть. Поэтому данные индексы не применимы к алгоритмам плотностной и модельной кластеризации.

Рассмотрим случай иерархической или нечёткой кластеризации. Пусть исходное множество состоит из двух объектов: и . При этом во множестве существует три кластера: кластер, содержащий оба объекта; кластер, содержащий только ; кластер, содержащий только . Пусть множество классов полностью совпадает со множеством кластеров. Данный случай соответствует наилучшей кластеризации, поэтому значения SD и DS должны равняться нулю. Рассмотрим пару (). Объекты данной пары принадлежат одному кластеру. Также объекты данной пары принадлежат разным классам. Поэтому данная пара удовлетворяет условиям для SD. В связи с этим SD равен 1, а не 0, что говорит о том, что кластеризация проведена не наилучшим образом. Полученное противоречие говорит о неприменимости индексов Rand, Jaccard и FM к иерархической и нечёткой кластеризации.

# 2. РАЗРАБОТКА КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА

## 2.1. Модификации для итеративной кластеризации

Все критерии, которые были взяты за основу для модификаций, могут быть применены к алгоритмам итеративной кластеризации без каких-либо изменений.

## 2.2. Модификации для плотностной кластеризации

Так как при использовании плотностной кластеризации вводится понятие шума, критерии оценки качества необходимо модифицировать.

Для использования F1-меры необходимо в формуле переопределить N. Под N будем понимать не размер исходного множества, а количество объектов исходного множества, отнесённых к какому-либо классу.

Для использования индексов Rand, Jaccard и FM необходимо переопределить обозначения SS, SD, DS и DD.

Под SS будем понимать количество пар объектов, которые удовлетворяют двум 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.

## 2.3. Модификации для иерархической кластеризации

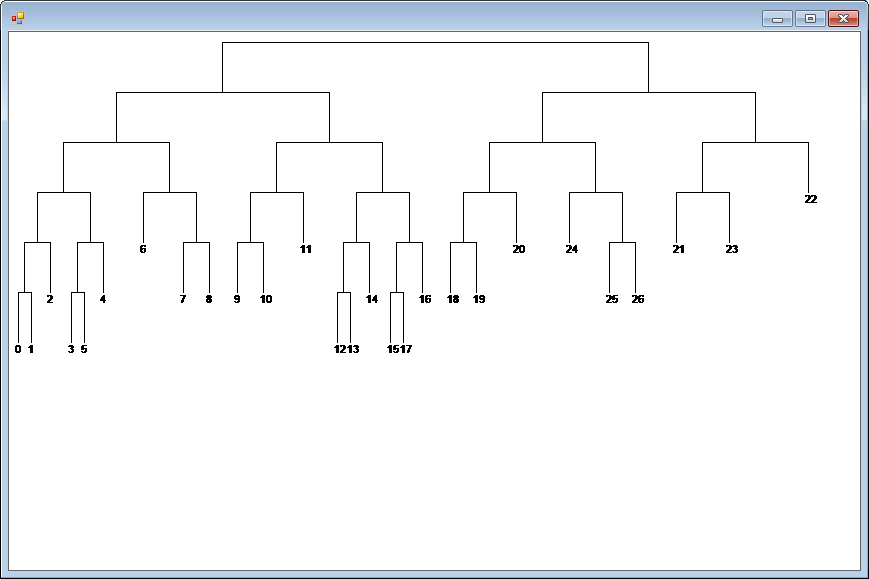
Особенностью иерархической кластеризации является наличие вложенных кластеров. В связи с этим классы, характеризующие то, как должна быть проведена кластеризация, также должны допускать вложенность. Для описания структуры множества классов введём понятие уровня класса.

Уровень класса определяется количеством классов, которые его содержат. Классы, не содержащиеся ни в одном из других классов, являются классами первого уровня. Классы, содержащиеся в n других классах, являются классами уровня n+1. Таким образом, исходное множество делится на классы первого уровня, которые могут делиться на кластеры второго уровня, которые в свою очередь также могут делиться на кластеры третьего уровня. Деление может продолжаться аналогичным образом дальше.

Если мы разобьём исходное множество на классы уровней 1, 2, …, N, то каждому объекту исходного множества будет поставлено в соответствие N классов уровней 1, 2, …, N, которые его содержат. Таким образом, классы могут быть заданы в виде множества списков классов, к которым принадлежит каждый из объектов.

Так как результатом иерархической кластеризации обычно является дендрограмма, необходимо разработать метод сравнения множества классов с дендрограммой.

Дендрограмма представляет собой дерево, которое показывает, как делились либо объединялись кластеры в процессе работы алгоритма кластеризации. Пример дендрограммы приведён на рисунке 1.



*Рисунок 1.Пример дендрограммы.*

Дендрограмма показывает все кластеры, созданные в процессе работы алгоритма кластеризации, а также их вложенность относительно друг друга. Кластерами, состоящими из одного элемента, являются все листья дерева. Также кластером являются все листья-потомки любого узла дерева. Так, на рисунке 1 кластерами являются следующие множества: {0}, {0, 1, 2}, {0, 1, 2, 3, 4, 5}, множество всех объектов. А множество {3, 4, 5, 6, 7, 8} не является кластером, так как не существует узла дерева, такого, чтобы множество всех его листьев-потомков совпадало с данным множеством.

Таким образом, дендрограмма описывает множество кластеров. Данное множество можно сравнить с множеством классов. При этом для сравнения не подойдут исходные внешние меры, так как они не предполагают наличие вложенных кластеров и классов. Поэтому возникает потребность в их модификации.

Модификация рассматриваемых внешних критериев заключается в следующем: значение критериев рассчитывается отдельно для каждого уровня класса, после чего считается среднее арифметическое от всех рассчитанных значений, которое и является значением модифицированного критерия для всего кластеризованного множества. При этом для F1-меры значения критериев для каждого уровня следует рассчитывать по исходной формуле, а для индексов Rand, Jaccard и FM необходимо переопределить обозначения DS и DD.

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.

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.

При расчёте внешних критериев для уровня n происходит сравнение классов уровня n с соответствующими им кластерами. Под соответствующим классу кластером будем понимать кластер из множества кластеров, построенного по дендрограмме, наиболее похожий на данный класс.

Для определения похожести кластера и класса используется расчёт следующего коэффициента:

где – количество объектов, которые есть и в классе, и в кластере, – размер класса, – размер кластера.

Данный коэффициент равен единице, если кластер и класс полностью совпадают, и равен нулю, если у кластера и класса нет ни одного общего элемента. При этом большее значение коэффициента говорит о большем сходстве между кластером и классом.

Таким образом, для расчёта модифицированных критериев оценки качества кластеризации необходимо выполнить следующие шаги:

1. Сформировать кластеры, соответствующие каждому классу первого уровня на основе дендрограммы.
2. Рассчитать критерий для классов первого уровня и соответствующих им кластеров. Если рассчитывается модифицированная F1-мера, то на данном шаге необходимо рассчитывать исходную F1-меру. Если рассчитываются модификации индексов Rand, Jaccard и FM, то необходимо в исходных индексах переопределить обозначения DS и DD.
3. Повторить шаги 1-2 для следующего уровня, если он есть.
4. Рассчитать среднее значение для значений, полученных на шаге 2. Полученный результат и будет значением модифицированного критерия.

## 2.4. Модификации для нечёткой кластеризации

Особенностью нечёткой кластеризации является то, что один объект может принадлежать нескольким кластерам одновременно. Обычно результатом нечёткой кластеризации является матрица степеней принадлежностей каждого объекта к каждому кластеру. На сравнении матриц степеней принадлежности для классов и кластеров основана существующая модификация индекса Rand для нечёткой кластеризации, описанная в подразделе 1.5.

В рамках данной работы будут разработаны модификации, работающие не со степенями принадлежности, а со списками кластеров, к которым принадлежит каждый объект.

В список кластеров каждого объекта будем включать кластер, степень принадлежности к которому данного объекта является наибольшей. Также в список кластеров каждого объекта будем включать все кластеры, степень принадлежности к которым будет отличаться от наибольшей для данного объекта менее чем на заранее выбранную малую величину .

Список кластеров для каждого объекта задаёт множество кластеров. Для задания множества классов введём список классов, к которым принадлежит каждый объект.

Тогда для использования F1-меры необходимо в формуле переопределить N. Под N будем понимать не размер исходного множества, а сумму размеров всех классов.

Для использования индексов Rand, Jaccard и FM необходимо переопределить обозначения SS, SD, DS и DD.

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

существует кластер, к которому принадлежат оба объекта пары;

существует класс, к которому принадлежат оба объекта пары.

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

существует кластер, к которому принадлежат оба объекта пары;

не существует класс, к которому принадлежат оба объекта пары.

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

не существует кластер, к которому принадлежат оба объекта пары;

существует класс, к которому принадлежат оба объекта пары.

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

не существует кластер, к которому принадлежат оба объекта пары;

не существует класс, к которому принадлежат оба объекта пары.

## 2.5. Модификации для модельной кластеризации

Так как вероятности принадлежности некоторых объектов ко всем кластерам в процессе запуска реализованного алгоритма модельной кластеризации – EM-алгоритма, оказались равными нулю, такие объекты было решено не относить ни к одному кластеру.

Объекты, не отнесённые ни к одному кластеру, являются шумом, который возникает также при использовании плотностной кластеризации. Поэтому для оценки качества модельной кластеризации можно использовать модификации, разработанные для плотностной кластеризации.

## 2.6. Сравнение модификаций

Значения всех полученных модификаций лежат в интервале от 0 до 1, при этом 0 соответствует наихудшей кластеризации, а 1 – наилучшей. Оказалось, что многие разработанные модификации подходят к разным типам кластеризации. Применимость модификаций F1-меры представлена в таблице 2.1, а применимость модификаций индексов Rand, Jaccard и FM представлена в таблице 2.2.

*Таблица 2.1.Применимость модификаций F1-меры к кластеризации различных типов.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Модификации | Итеративная кластеризация | Плотностная кластеризация | Иерархическая кластеризация | Нечёткая кластеризация | Модельная кластеризация |
| Исходный критерий | + | - | - | - | - |
| Модификация для плотностной кластеризации | + | + | - | - | + |
| Модификация для иерархической кластеризации | - | - | + | - | - |
| Модификация для нечёткой кластеризации | + | + | + | + | + |

*Таблица 2.2.Применимость модификаций индексов Rand, Jaccard и FM к кластеризации различных типов.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Модификации | Итеративная кластеризация | Плотностная кластеризация | Иерархическая кластеризация | Нечёткая кластеризация | Модельная кластеризация |
| Исходный критерий | + | - | - | - | - |
| Модификация для плотностной кластеризации | + | + | - | - | + |
| Модификация для иерархической кластеризации | - | - | + | - | - |
| Модификация для нечёткой кластеризации | + | + | - | + | + |

В результате сравнения модификаций оказалось, что модификация F1-меры, разработанная для нечёткой кластеризации, применима к кластеризации всех рассматриваемых типов.

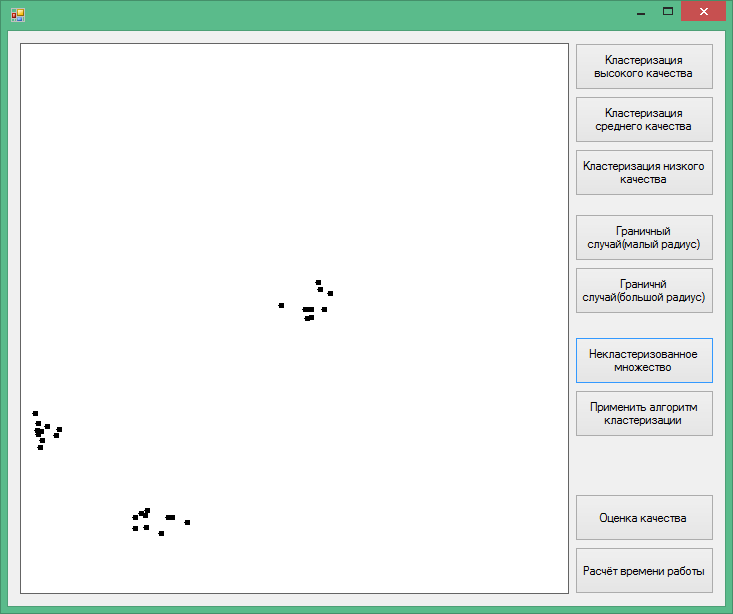
# 3. РЕАЛИЗАЦИЯ КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА

## 3.1. Описание генератора исходных множеств

Для тестирования реализованных модификаций критериев оценки качества кластеризации необходимо наличие кластеризованных множеств. Для их создания необходимо применить какой-либо алгоритм кластеризации к исходному множеству. Для создания исходных множеств был написан генератор исходных множеств. Так как ожидается применение алгоритмов кластеризации из разных групп, возникает необходимость генерировать разные множества. Для обеспечения такой возможности были реализованы следующие режимы работы генератора:

1. Режим без шума.

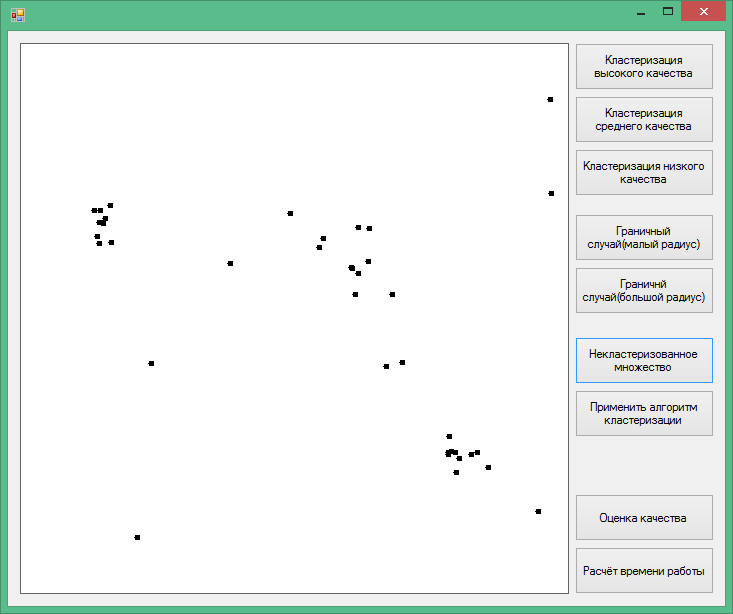
Генератор случайным образом выбирает центры и радиусы трёх окружностей, после чего внутри данных окружностей случайным образом генерируются по 10 точек в каждой. Результат генерации в таком режиме представлен на рисунке 3.1.



*Рисунок 3.1. Исходное множество, сгенерированное в режиме без шума.*

1. Режим с шумом.

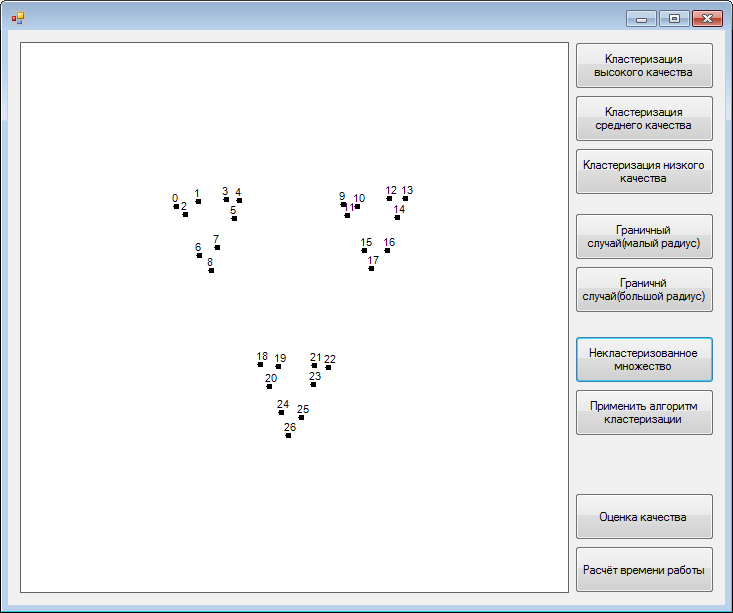
Генератор вначале создаёт точки аналогично режиму без шума, а затем добавляет 10 случайных точек, которые и составляют шум. Результат такой генерации представлен на рисунке 3.2.



*Рисунок 3.2. Исходное множество, сгенерированное в режиме с шумом.*

1. Режим для иерархической кластеризации.

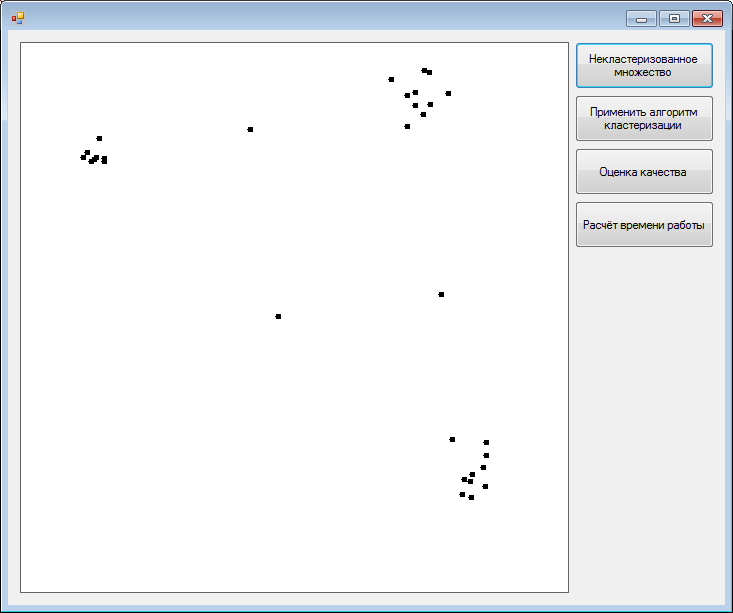
Так как иерархическая кластеризация предполагает наличие вложенных кластеров, генератор создаёт вложенные группы точек. Результат такой генерации представлен на рисунке 3.3.



*Рисунок 3.3. Исходное множество, сгенерированное в режиме для иерархической кластеризации.*

1. Режим для нечёткой кластеризации.

Генератор вначале создаёт точки аналогично режиму без шума. Затем генератор добавляет три точки, находящиеся на серединах отрезков, соединяющих центры окружностей, внутри которых происходила генерация точек: первую точку между 1-ой и 2-ой окружностью, вторую – между 1-ой и 3-ей окружностью, третью – между 2-ой и третьей окружностью. Результат такой генерации представлен на рисунке 3.4.



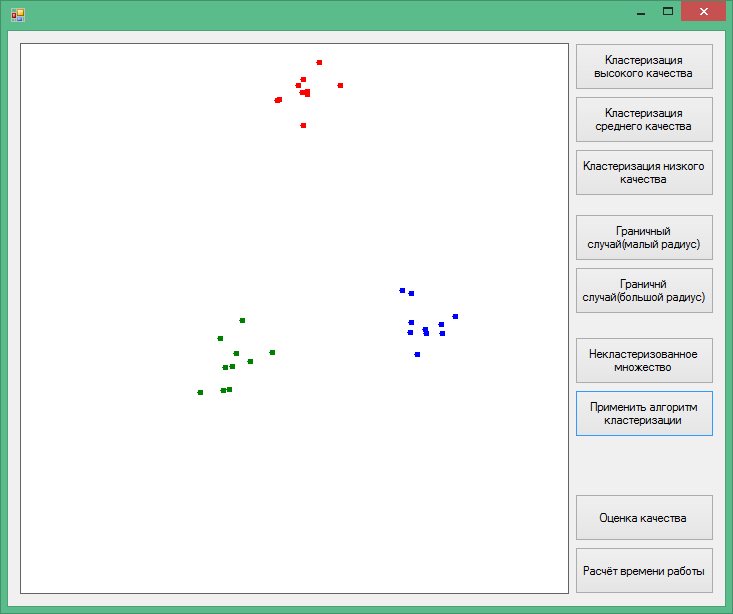
*Рисунок 3.4. Исходное множество, сгенерированное в режиме для нечёткой кластеризации.*

## 3.2. Реализация алгоритмов кластеризации

Для тестирования реализованных модификаций критериев оценки качества кластеризации были реализованы следующие алгоритмы кластеризации:

1. Итеративный алгоритм: k-means.

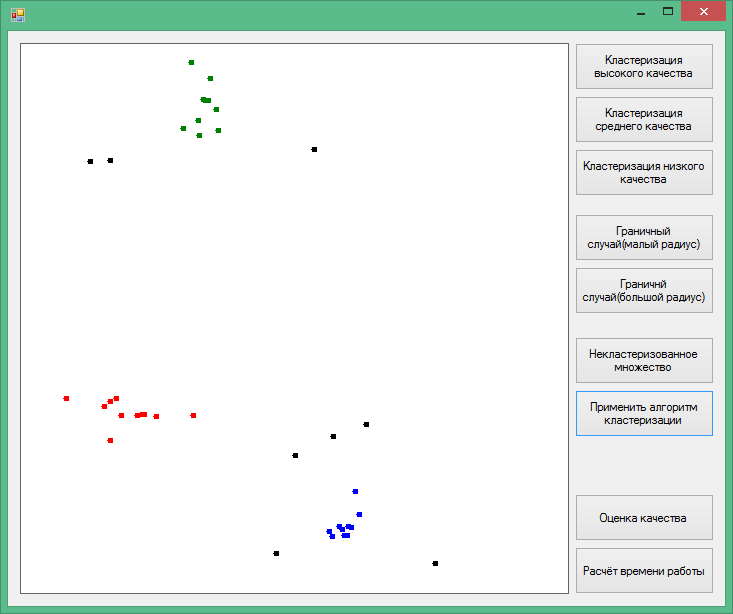
Результат применения данного алгоритма к множеству без шума представлен на рисунке 3.5.



*Рисунок 3.5. Результат работы реализованного алгоритма k-means.*

1. Плотностной алгоритм: DBSCAN.

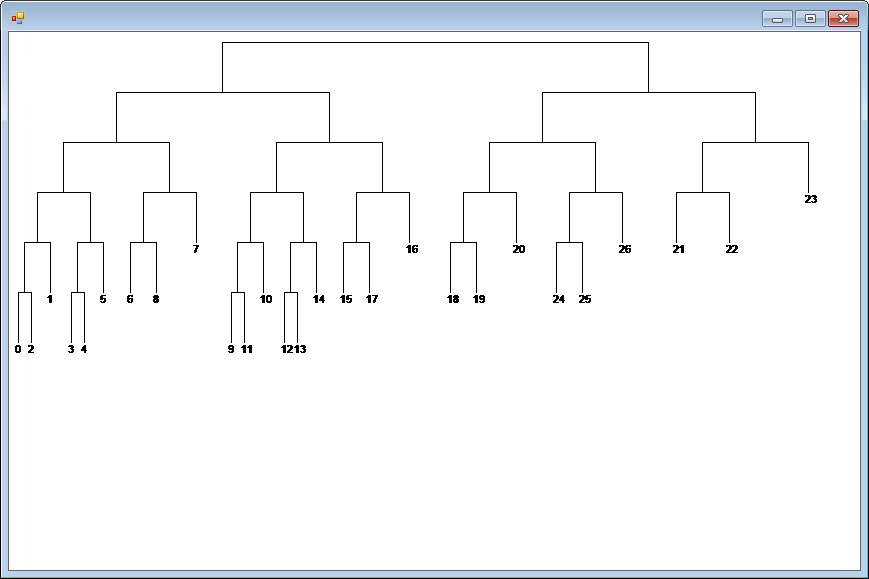
Результат применения данного алгоритма к множеству c шумом представлен на рисунке 3.6.



*Рисунок 3.6. Результат работы реализованного алгоритма DBSCAN.*

1. Иерархический алгоритм: метод ближайшего соседа.

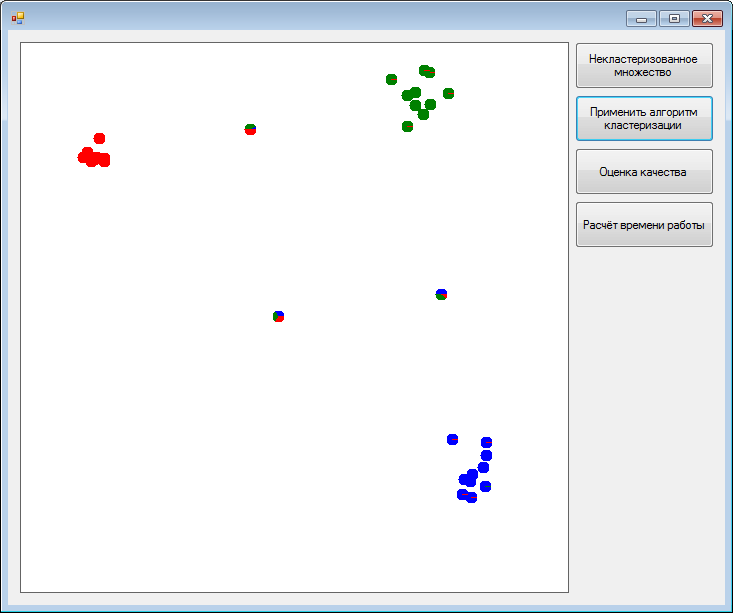
Результатом метода ближайшего соседа является дендрограмма – дерево, показывающее, как объединялись кластеры во время работы алгоритма. Результат для множества, изображённого на рисунке 3, представлен на рисунке 3.7.



*Рисунок 3.7. Результат работы реализованного метода ближайшего соседа.*

1. Нечёткий алгоритм: с-means.

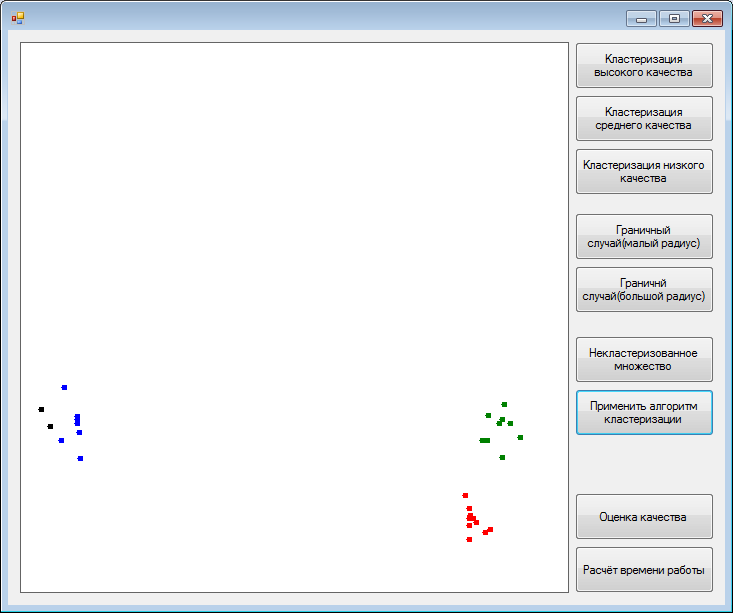
Результатом метода c-means является матрица степеней принадлежности. Каждая i-ая строка такой матрицы содержит степени принадлежности i-ого объекта ко всем кластерам. Для визуализации степеней принадлежности каждый объект был представлен кругом, состоящим из трёх секторов различных цветов. Угол каждого из секторов пропорционален степени принадлежности объекта к одному из кластеров, при этом сектора одного цвета соответствуют одному кластеру. Результат применения данного алгоритма к множеству для нечёткой кластеризации представлен на рисунке 3.8.



*Рисунок 3.8. Результат работы реализованного алгоритма c-means.*

1. Модельный алгоритм: EM-алгоритм.

Результат применения данного алгоритма к множеству без шума представлен на рисунке 3.9.



*Рисунок 3.9. Результат работы реализованного EM-алгоритма.*

## 3.3. Представление информации о классах

Так как в данной работе разрабатываются внешние критерии, возникает необходимость в дополнительной информации о том, как должна быть проведена кластеризация. Данная информация представлена в виде классов.

Классы формируются по-разному в зависимости от режима генерации исходного множества. Описание формирования классов при различных режимах представлено ниже:

1. Режим без шума.

Каждая группа точек, сгенерированных внутри одной окружности, считается отдельным классом.

1. Режим с шумом.

Аналогично режиму без шума, каждая группа точек, сгенерированных внутри одной окружности, считается отдельным классом. Точки, составляющие шум, не относятся ни к одному из классов.

1. Режим для иерархической кластеризации.

Классы для иерархической кластеризации имеют два уровня, при этом кластеры второго уровня вложены в кластеры первого уровня. Объекты множества для иерархической кластеризации пронумерованы в соответствии с рисунком 3. Принадлежность объектов к классам двух уровней представлена в таблице 3.1.

*Таблица 3.1. Принадлежность объектов к классам для случая иерархической кластеризации.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Номер объекта | 0-2 | 3-5 | 6-8 | 9-11 | 12-14 | 15-17 | 18-20 | 21-23 | 24-26 |
| Номер класса первого уровня | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 |
| Номер класса второго уровня | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

1. Режим для нечёткой кластеризации.

Аналогично режиму без шума, каждая группа точек, сгенерированных внутри одной окружности, считается отдельным классом. Точка, сгенерированная между первой и второй кластерами, относится к первому и второму классу одновременно. Оставшиеся две точки аналогичным образом относятся к двум классам: одна – к первому и третьему классу, другая – ко второму и третьему классу.

## 3.4. Реализация модификаций внешних критериев оценки качества

Модификации внешних критериев были реализованы в виде классов на языке C#, конструкторы которых принимают в качестве аргументов информацию о разбиении исходного множества на кластеры и информацию о классах. Классы, реализующие критерии, содержат метод, возвращающий значение критерия, вычисленное с помощью информации, переданной в конструктор.

# 4. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

## 4.1. Описание экспериментов

Эксперименты проводились над множеством сайтов, поделённым на три класса: фишинговые сайты, подозрительные сайты и нефишинговые сайты[14]. Цель экспериментов – проверить информацию о применимости разработанных модификаций внешних критериев из таблиц 2.1 и 2.2. Все разработанные модификации принимают значения от 0 до 1, при этом большее значение говорит о том, что кластеризация проведена более качественно. При проведении экспериментов в качестве множества классов берутся классы из множества сайтов. Рассматриваются следующие случаи:

1. Идеальная кластеризация.

При идеальной кластеризации множество кластеров совпадает с множеством классов. Рассматриваемые модификации критериев в данном случае должны равняться единице.

1. Кластеризация с помощью алгоритма кластеризации.

В данном случае к множеству сайтов применяется один из рассматриваемых алгоритмов кластеризации. Полученное в результате кластеризации множество и является множеством кластеров.

1. Случайная кластеризация.

В данном случае объекты разбиваются по кластерам случайным образом.

Каждая разработанная модификация внешнего критерия применяется к случаю наилучшей кластеризации, случаям кластеризации с помощью алгоритмов кластеризации тех типов, к которым применима данная модификация в соответствии с таблицами 2.1 и 2.2, а также к случаю случайной кластеризации.

## 4.2. Эксперименты для модификаций для плотностной кластеризации.

Результаты экспериментов для модификаций для плотностной кластеризации представлены в таблице 4.1.

*Таблица 4.1. Результаты экспериментов для модификаций для плотностной кластеризации.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Случаи | F1-мера | Индекс Rand | Индекс Jaccard | Индекс FM |
| Идеальная кластеризация | 1 | 1 | 1 | 1 |
| Итеративная кластеризация: k-means | 0,62 | 0,61 | 0,35 | 0,52 |
| Плотностная кластеризация: DBSCAN | 0,59 | 0,6 | 0,32 | 0,49 |
| Модельная кластеризация: EM-алгоритм | 0,66 | 0,65 | 0,35 | 0,53 |
| Случайная кластеризация без шума | 0,47 | 0,5 | 0,3 | 0,47 |
| Случайная кластеризация с шумом | 0,38 | 0,52 | 0,23 | 0,38 |

В идеальном случае значения всех модификаций оказались равными 1. Значения в случаях применения алгоритмов кластеризации оказались больше, чем в случаях случайной кластеризации.

## 4.3. Эксперименты для модификаций для иерархической кластеризации.

Результаты экспериментов для модификаций для иерархической кластеризации представлены в таблице 4.2.

*Таблица 4.2. Результаты экспериментов для модификаций для иерархической кластеризации.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Случаи | F1-мера | Индекс Rand | Индекс Jaccard | Индекс FM |
| Идеальная кластеризация | 1 | 1 | 1 | 1 |
| Иерархическая кластеризация: метод ближайшего соседа | 0,6 | 0,5 | 0,3 | 0,47 |
| Случайная иерархическая кластеризация | 0,6 | 0,5 | 0,3 | 0,47 |

В идеальном случае значения всех модификаций оказались равными 1. Значения в случае применения алгоритма кластеризации и в случае случайной кластеризации оказались равными, что говорит о том, что качество кластеризации в данных двух случаях одинаково. Высокие значения для случайной кластеризации можно объяснить тем, что при иерархической кластеризации формируется большое количество вложенных кластеров, и при расчёте модификаций критериев для иерархической кластеризации классы сравниваются с наиболее похожими на данные классы кластерами. Так как количество кластеров намного больше, чем количество классов, наиболее похожие на классы кластеры обладают высокой степенью сходства с классами. Поэтому в случае иерархической кластеризации качество случайной кластеризации может совпадать с качеством кластеризации с помощью алгоритма кластеризации.

## 4.4. Эксперименты для модификаций для нечёткой кластеризации.

Результаты экспериментов для модификаций для нечёткой кластеризации представлены в таблице 4.3.

*Таблица 4.3. Результаты экспериментов для модификаций для нечёткой кластеризации.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Случаи | F1-мера | Индекс Rand | Индекс Jaccard | Индекс FM |
| Идеальная кластеризация | 1 | 1 | 1 | 1 |
| Итеративная кластеризация: k-means | 0,62 | 0,61 | 0,35 | 0,52 |
| Плотностная кластеризация: DBSCAN | 0,59 | 0,6 | 0,32 | 0,49 |
| Иерархическая кластеризация: метод ближайшего соседа | 0,62 | - | - | - |
| Нечёткая кластеризация: c-means | 0,71 | 0,59 | 0,48 | 0,68 |
| Модельная кластеризация: EM-алгоритм | 0,66 | 0,65 | 0,35 | 0,53 |
| Случайная кластеризация без шума | 0,47 | 0,5 | 0,3 | 0,47 |
| Случайная кластеризация с шумом | 0,38 | 0,52 | 0,23 | 0,38 |
| Случайная иерархическая кластеризация | 0,6 | - | - | - |
| Случайная нечёткая кластеризация | 0,46 | 0,49 | 0,34 | 0,52 |

В идеальном случае значения всех модификаций оказались равными 1. Значения в случаях применения алгоритмов кластеризации оказались больше, чем в соответствующих случаях случайной кластеризации.

# 5. ТЕХНИКО-ЭКОНОМИЧЕСКОЕ ОБОСНОВАНИЕ

## 5.1. Концепция экономического обоснования

Кластеризация (кластерный анализ) – совокупность методов, выполняющих выделение групп объектов, называемых кластерами, таким образом, чтобы объекты одного кластера были похожи, а объекты разных кластеров существенно различались.

Кластеризация позволяет разбить объекты на группы, что упрощает работу с данными. Для оценки качества кластеризации может применяться расчёт различных критериев, по значениям которых можно судить о качестве проведённой кластеризации. Среди данных критериев можно выделить внешние критерии, которые используют информацию о том, как должна быть проведена кластеризация. Разработке внешних критериев и посвящена данная работа.

Алгоритмы кластеризации могут быть разбиты на несколько типов: итеративная, плотностная, иерархическая, нечёткая и модельная кластеризации. Большинство внешних критериев были разработаны для итеративной кластеризации и не подходят к кластеризации других типов.

Целью работы является разработка внешних критериев оценки качества, применимых к кластеризации различных типов, на основе внешних критериев для итеративной кластеризации.

## 5.2. Трудоёмкость и календарный план выполнения проекта

Трудоёмкость и календарный план представлены в таблице 5.1.

*Таблица 5.1. Трудоёмкость и календарный план выполнения проекта.*

|  |  |  |
| --- | --- | --- |
| Наименование работ | Трудоёмкость, чел./дни | |
| Руководитель | Дипломант |
| Составление технического задания | 3 | - |
| Обзор литературы | - | 5 |
| Реализация алгоритмов кластеризации | 1 | 18 |
| Реализация методов оценки качества кластеризации | - | 11 |
| Сравнение методов оценки качества кластеризации | - | 2 |
| Оформление пояснительной записки | 2 | 18 |
| Итого | 5 | 54 |

## 5.3. Смета затрат на выполнение проекта

Себестоимость проекта будем определять по фактическим затратам. Расчёт себестоимости осуществляется по следующим статьям:

1. Материалы с учётом транспортно-заготовительных расходов;
2. Издержки на амортизацию;
3. Основная и дополнительная заработная плата;
4. Страховые взносы в государственные внебюджетные фонды;
5. Накладные расходы.

В качестве месячной заработной платы дипломанта примем заработную плату инженера, равную 37000 рублей.[4] Для определения дневной ставки заработной платы дипломанта разделим заработную плату за месяц на количество рабочих дней в месяце (21 день). В результате получим 1761,91 рубля/день.

В качестве месячной заработной платы руководителя примем 60000 рублей. Для определения дневной ставки заработной платы руководителя разделим заработную плату за месяц на количество рабочих дней в месяце (21 день). В результате получим 2857,14 рубля/день.

Расходы на основную заработную плату определяются по формуле:

где k – количество исполнителей, – время, затраченное i-ым исполнителем на исследование, – ставка i-ого исполнителя.

Значения возьмём из таблицы 5.1. Подставив значения в формулу, получим 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.

К статье “Страховые взносы в государственные внебюджетные фонды” относят отчисления на обязательное социальное, пенсионное и медицинское страхование. Страховые взносы 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

Затраты на материалы будем определять по формуле:

где l – индекс материала, – норма расхода l-ого материала на единицу продукции, – цена приобретения единицы l-ого материала, – норма транспортно-заготовительных расходов.

Будем считать, что норма транспортно-заготовительных расходов составляет 15%.

Затраты на сырьё и материалы представлены в таблице 5.2.

*Таблица 5.2. Материалы.*

|  |  |  |  |
| --- | --- | --- | --- |
| Материал | Количество | Цена, руб. | Сумма, руб. |
| Бумага, пачка | 1 | 230,00 | 230,00 |
| Чернила для принтера | 4 | 140,00 | 560,00 |
| Канцелярские товары |  |  | 300,00 |
| ИТОГО | | | 1090,00 |
| Транспортные расходы (15 %) | | | 163,50 |
| ВСЕГО | | | 1253,50 |

Величина амортизационных отчислений по i-ому основному средству за год определяются следующей формулой:

где – первоначальная стоимость i-ого основного средства, – годовая норма амортизации i-ого средства.

При выполнении работы использовались такие основные средства, как компьютер и принтер.

Рассчитаем величину амортизационных отчислений за год. Первоначальная стоимость компьютера – 20000 рублей, принтера – 4000 рублей. Срок службы компьютера и принтера – 5 лет. Будем считать, что годовая норма амортизации составляет 20%. Подставив данные значения в формулу, получим следующие значения амортизационных отчислений за год: по компьютеру – 4000 рублей, по принтеру – 800 рублей.

Суммарная амортизация за год составит 4800 рублей. Для определения амортизации за день разделим амортизацию за год на количество рабочих дней в году – 247. В результате получим 19,43 рубля. Амортизация за время работы над проектом составит (54+5)\*19,43=1146,37 рубля.

Накладные расходы вычисляются по формуле:

где – норматив накладных расходов, – затраты на основную заработную плату, – затраты на дополнительную заработную плату.

Норматив накладных расходов составляет 40%. Подставим значения в формулу. Получим:

рублей

## 5.4. Себестоимость выполнения ВКР

Себестоимость выполнения ВКР представлена в таблице 5.3.

*Таблица 5.3.Затраты на выполнение ВКР.*

|  |  |
| --- | --- |
| Статья затрат | Сумма, руб. |
| Материалы | 1253,5 |
| Издержки на амортизацию | 1146,37 |
| Основная заработная плата | 109428,59 |
| Дополнительная заработная плата | 15320,00 |
| Страховые взносы | 37424,58 |
| Накладные расходы | 49899,44 |
| ИТОГО | 214472,48 |

## 5.5. Выводы

В данном разделе была рассчитана себестоимость разработки и реализации внешних критериев оценки качества кластеризации, применимых к алгоритмам различных типов.

В ходе анализа статей затрат было выяснено, что максимальные расходы приходятся на заработную плату.

Себестоимость проекта составляет 214472 руб. 48 коп. В результате выполнения проекта были разработаны методы, позволяющие оценить качество кластеризации различных типов.

# ЗАКЛЮЧЕНИЕ

В

# 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, И.И. Холод, М.Д. Тесс, С.И. Елизаров. Анализ данных и процессов. СПб.: БХВ-Петербург, 2009
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: 23.05.2017).

1. Главная книга. Страховые взносы в 2017 году. [Электронный ресурс].

URL: http://glavkniga.ru/situations/k501873 (дата обращения: 23.05.2017).

1. НОУ ИНТУИТ. Методы кластерного анализа. Итеративные методы. [Электронный ресурс]. URL: http://www.intuit.ru/studies/courses/6/6/lecture/184%25253Fpage% (дата обращения: 07.05.2017).
2. Хабрахабр. Плотностный алгоритм кластеризации пространственных данных с присутствием шума – DBSCAN. [Электронный ресурс]. URL: https://habrahabr.ru/post/143151/ (дата обращения: 07.05.2017).
3. Портал искусственного интеллекта. Метод ближайшего соседа или метод одиночной связи. [Электронный ресурс]. URL:

http://www.aiportal.ru/articles/autoclassification/single-link.html (дата обращения: 07.05.2017).

1. A tutorial on Clustering Algorithms. Fuzzy C-Means Clustering. [Электронный ресурс]. URL:

https://home.deib.polimi.it/matteucc/Clustering/tutorial\_html/cmeans.html (дата обращения: 07.05.2017).

1. BaseGroup Labs. EM – масштабируемый алгоритм кластеризации. URL:

https://basegroup.ru/community/articles/em (дата обращения: 07.05.2017).

1. **Хабрахабр**. Обзор алгоритмов кластеризации числовых пространств данных. URL: https://habrahabr.ru/post/164417/ (дата обращения: 11.05.2017).
2. Elena Sivogolovko. The influence of data quality on Clustering Outcomes. IOS Press BV, 2013
3. Sonia Amodio, Antonio d’Ambrosio, Carmela Iorio, Roberta Siciliano. Adjusted Concordance Index, an extension of the Adjusted index to fuzzy partitions. [Электронный ресурс]. URL: https://arxiv.org/pdf/1509.00803.pdf (дата обращения: 27.05.2017).
4. Machine Learning Repository. Website Phishing Data Set. [Электронный ресурс]. URL: http://archive.ics.uci.edu/ml/datasets/Website+Phishing (дата обращения: 02.06.2017).