**«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* |  |  |
| Консультанты |  |  |  | Леньков А.Н. |
|  |  | *подпись* |  |  |
|  |  |  |  | Яценко И.В. |
|  |  | *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):  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 | | доцент | |  | |  | Жукова Н.А. | | | | |
|  | | *(Уч. степень, уч. звание)* | | *подпись* | |  |  | | | | |
| Консультант | |  | |  | |  | Леньков А.Н. | | | | |
|  | |  | | *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 – 10.01 |
| 3 | Реализация методов оценки качества кластеризации | 30.04 – 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 |
| 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.

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 модельная 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 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](#_Toc482303216)

[Общие сведения 7](#_Toc482303217)

[История кластеризации 7](#_Toc482303218)

[Современные проблемы кластеризации 9](#_Toc482303219)

[Пути решения проблем 10](#_Toc482303220)

[Цель и задачи работы 10](#_Toc482303221)

[1. АНАЛИЗ СУЩЕСТВУЮЩИХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ И КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ 12](#_Toc482303222)

[1.1. Классификация алгоритмов кластеризации 12](#_Toc482303223)

[1.2. Примеры алгоритмов кластеризации 13](#_Toc482303224)

[1.3. Критерии оценки качества кластеризации 20](#_Toc482303225)

[2. РАЗРАБОТКА КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА 22](#_Toc482303226)

[2.1. Модификации для итеративной кластеризации 22](#_Toc482303227)

[2.2. Модификации для плотностной кластеризации 22](#_Toc482303228)

[2.3. Модификации для иерархической кластеризации 23](#_Toc482303229)

[2.4. Модификации для нечёткой кластеризации 23](#_Toc482303230)

[2.5. Модификации для модельной кластеризации 23](#_Toc482303231)

[3. РЕАЛИЗАЦИЯ КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА 24](#_Toc482303232)

[3.1. Описание генератора исходных множеств 24](#_Toc482303233)

[3.2. Реализация алгоритмов кластеризации 26](#_Toc482303234)

[3.3. Представление информации о классах 29](#_Toc482303235)

[3.4. Реализация модификаций внешних критериев оценки качества 30](#_Toc482303236)

[4. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ 31](#_Toc482303237)

[5. ТЕХНИКО-ЭКОНОМИЧЕСКОЕ ОБОСНОВАНИЕ 31](#_Toc482303238)

[4.1. Детализированный план работ 31](#_Toc482303239)

[4.2. Расчёт ожидаемой длительности работ 31](#_Toc482303240)

[4.3. Расчёт ставки заработной платы 32](#_Toc482303241)

[4.4. Расчёт расходов на заработную плату 32](#_Toc482303242)

[4.5. Расчёт отчислений на страховые взносы 33](#_Toc482303243)

[4.6. Расчёт затрат на содержание и эксплуатацию оборудования 33](#_Toc482303244)

[4.6. Совокупная величина затрат 34](#_Toc482303245)

[ЗАКЛЮЧЕНИЕ 35](#_Toc482303246)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 36](#_Toc482303247)

# ВВЕДЕНИЕ

## 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.

## 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.

## Современные проблемы кластеризации

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.

## Пути решения проблем

1. Проблема обоснования качества результатов кластеризации.

Внешние критерии позволяют максимально полно учитывать имеющиеся знания об объектах.

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.

## Цель и задачи работы

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. Классификация алгоритмов кластеризации

1. Итеративная кластеризация[8].

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.[8]

1. Плотностная кластеризация[13].

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[13].

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.

## 1.2. 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.[9]

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.[10] Результатом работы алгоритма является последовательностей объединений кластеров, которая может быть представлена в виде дерева, называемого дендрограммой.

1. Пример нечёткой кластеризации: с-means

Азван[11]

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.[12]

## 1.3. 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)[14]

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

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

## 2.1. Модификации для итеративной кластеризации

Все вышеперечисленные критерии оценки качества 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 кластеру.

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

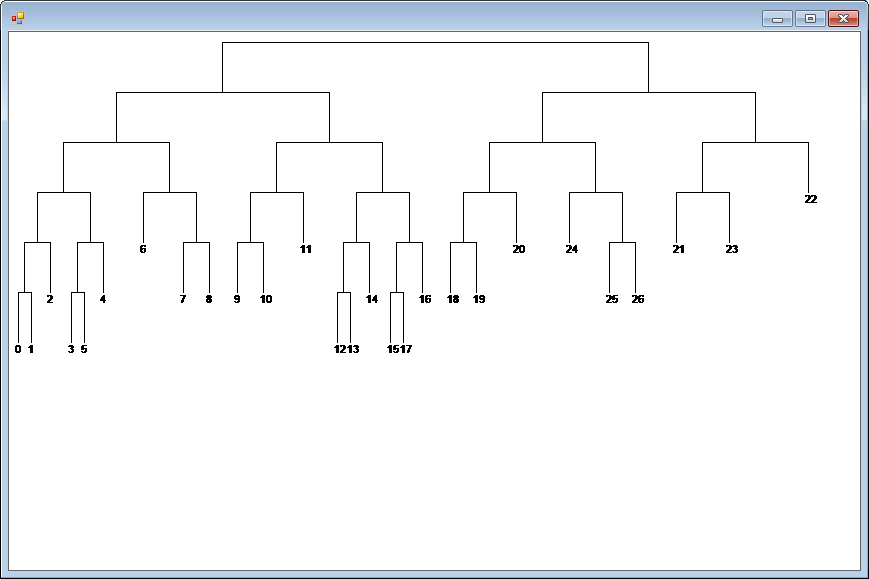
Особенностью иерархической кластеризации является наличие вложенных кластеров. В связи с этим классы, характеризующие то, как должна быть проведена кластеризация, также должны допускать вложенность. Для описания структуры множества классов введём понятие уровня класса.

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 метрик рассчитывается отдельно для каждого уровня класса, после чего считается среднее арифметическое от всех рассчитанных значений, которое и является значением метрики для всего кластеризованного множества.

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 исходный критерий для классов первого уровня и соответствующих им кластеров.
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

Для каждого

## 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.

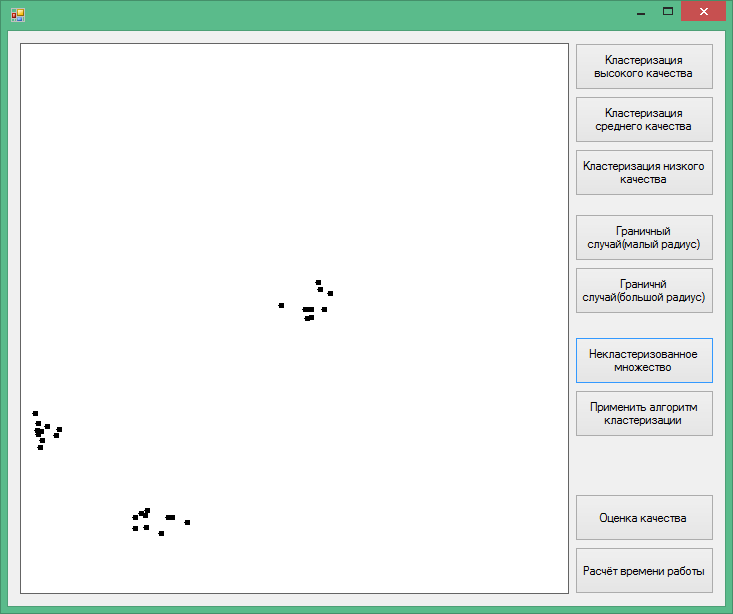
# 3. РЕАЛИЗАЦИЯ КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ДЛЯ ГРУПП АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА

## 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.

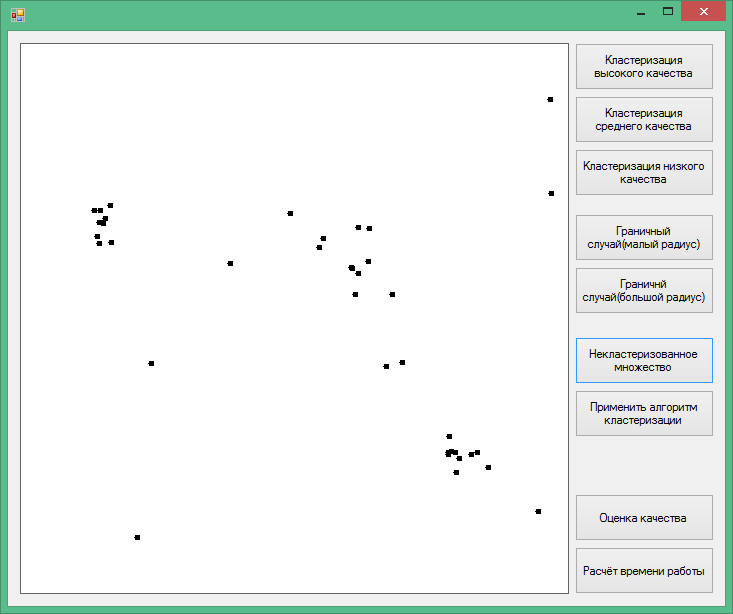
Генератор случайным образом выбирает центры и радиусы трёх окружностей, после чего внутри данных окружностей случайным образом генерируются по 10 точек в каждой. Результат генерации в таком режиме представлен на рисунке 1.



*Рисунок 1. Исходное множество, сгенерированное в режиме без шума.*

1. Режим с шумом.

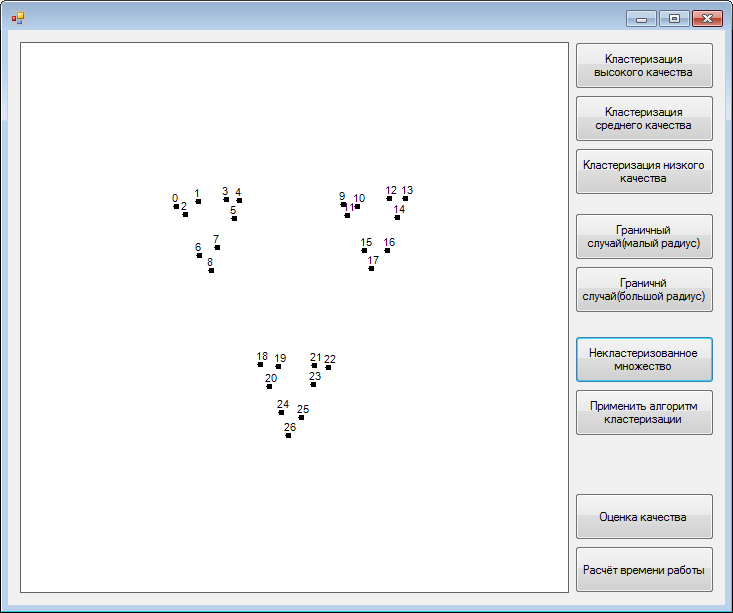
Генератор вначале создаёт точки аналогично режиму без шума, а затем добавляет 10 случайных точек, которые и составляют шум. Результат такой генерации представлен на рисунке 2.



*Рисунок 2. Исходное множество, сгенерированное в режиме с шумом.*

1. Режим для иерархической кластеризации.

Так как иерархическая кластеризация предполагает наличие вложенных кластеров, генаратор создаёт вложенные группы точек. Результат такой генерации представлен на рисунке 3.



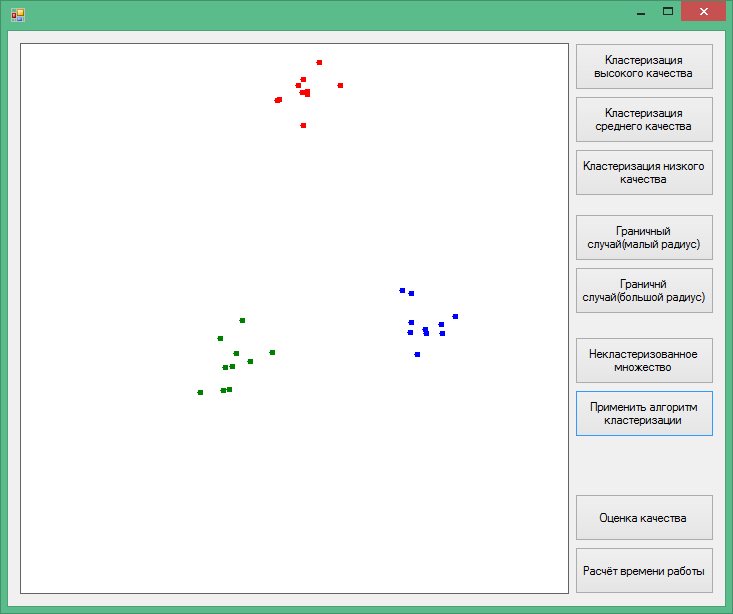
*Рисунок 3. Исходное множество, сгенерированное в режиме для иерархической кластеризации.*

## 3.2. Реализация алгоритмов кластеризации

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

1. Итеративный алгоритм: k-means.

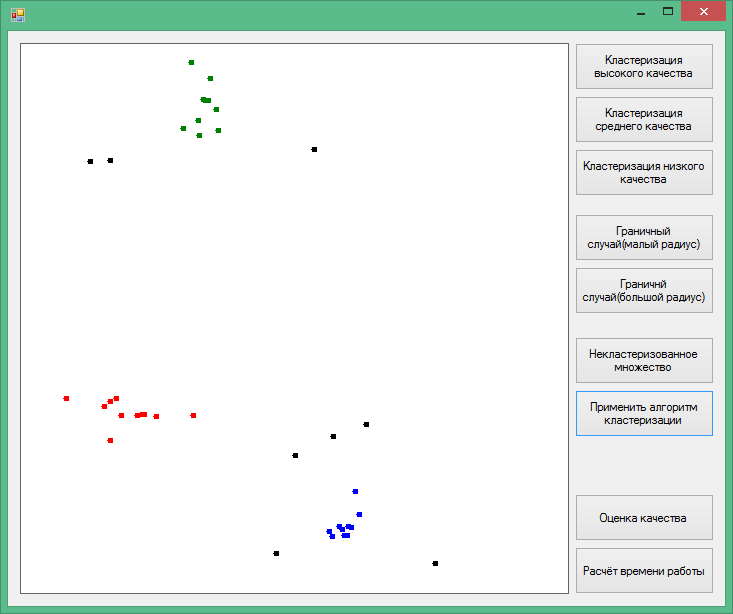
Результат применения данного алгоритма к множеству без шума представлен на рисунке 5.



*Рисунок 5. Результат работы реализованного алгоритма k-means.*

1. Плотностной алгоритм: DBSCAN.

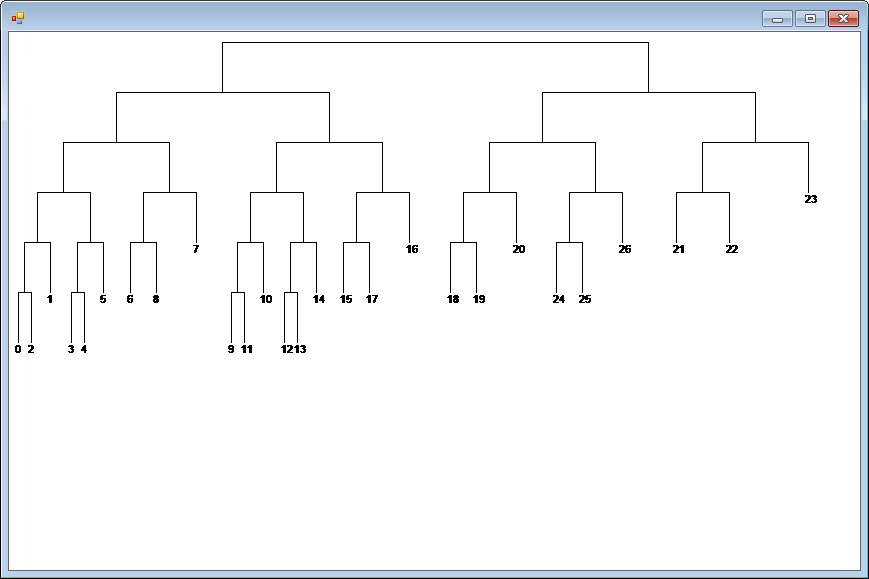
Результат применения данного алгоритма к множеству c шумом представлен на рисунке 6.



*Рисунок 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 рисунке 7.



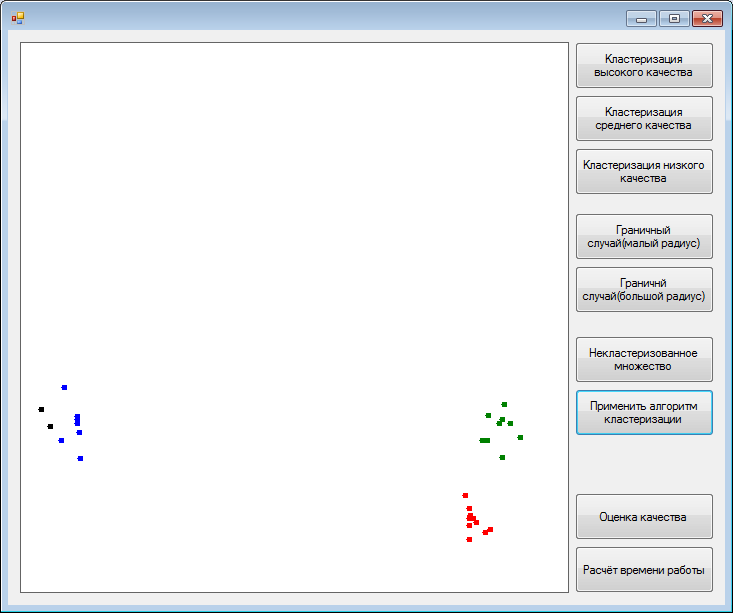
*Рисунок 7. Результат работы реализованного метода ближайшего соседа.*

1. Нечёткий алгоритм: с-means.

Результат применения данного алгоритма к множеству для нечёткой кластеризации представлен на рисунке 8.

1. Модельный алгоритм: EM-алгоритм.

Результат применения данного алгоритма к множеству без шума представлен на рисунке 9.



*Рисунок 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 классу.

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 1.

*Таблица 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 |

## 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

# 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

# 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: 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).

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