**«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 |  |  |  | Иванов И.И. |
|  | *(Уч. степень, уч. звание)* | *подпись* |  |  |

Санкт-Петербург

2016**ЗАДАНИЕ**

**на выпускную квалификационную работу**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Утверждаю | | | | | | | |
|  | | Зав. кафедрой МОЭВМ | | | | | | | |
|  | | \_\_\_\_\_\_\_\_\_\_\_\_ Кринкин К.В. | | | | | | | |
|  | | «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | | | | | |
| Студент | Федоров А.М. | | | | |  | Группа | 3381 | |
| Тема работы: Реализация различных методов оценки качества кластеризации данных | | | | | | | | | |
| Место выполнения ВКР: МОЭВМ | | | | | | | | | |
| Исходные данные (технические требования):  Реализовать методы оценки качества кластеризации и провести их анализ | | | | | | | | | |
| Содержание ВКР:  Кратко перечисляются основные разделы ВКР | | | | | | | | | |
| Перечень отчетных материалов: пояснительная записка, иллюстративный материал, иные отчетные материалы | | | | | | | | | |
| Дополнительные разделы: указывается наименование дополнительного раздела | | | | | | | | | |
|  | | | | | | | | | |
| Дата выдачи задания | | | | Дата представления ВКР к защите | | | | | |
| «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | | «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. | | | | | |
|  | | | |  | | | | | |
| Студент | | |  | | Федоров А.М. | | | |
| Руководитель доцент | | |  | | Жукова Н.А. | | | |
| *(Уч. степень, уч. звание)* | | |  | |  | | | |
| Консультант | | |  | | Иванов И.И. | | | |
| *(Уч. степень, уч. звание)* | | |  | |  | | | |

**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 | 00.00 – 00.00 |
| 2 | Генерация кластеризованных множеств | 00.00 – 00.00 |
| 3 | Реализация методов оценки качества кластеризации | 00.00 – 00.00 |
| 4 | Анализ методов оценки качества кластеризации | 00.00 – 00.00 |
| 5 | Оформление пояснительной записки | 00.00 – 00.00 |
| 6 | Оформление иллюстративного материала | 00.00 – 00.00 |

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

**РЕФЕРАТ**

Пояснительная записка 16 стр., 3 рис., 1 табл., 00 ист., 00 прил.

Объектом исследования являются методы оценки качества кластеризации.

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

Кратко (в 10-12) строк описать основное содержание работы, методы исследования (разработки), полученные результаты.

Для тестирования реализованных методов оценки качества кластеризации была написана программа, генерирующая множество точек на плоскости, разбитых на кластеры. При этом рассмотрены случаи различного качества кластеризации точек: высокий, средний и низкий уровни качества. Были реализованы такие методы оценки качества, как расчёт F1-меры,

**ABSTRACT**

The work is devoted to realization and assessment of different methods of clustering quality grade. To testing methods the program for generation of clustering sets was created. The clustering sets has different level of quality: high, medium and low. Following quality grade methods were implemented: F1-meassure,

**содержание**

[ВВЕДЕНИЕ 7](#_Toc471758235)

[1. СОВРЕМЕННОЕ СОСТОЯНИЕ ВОПРОСА 8](#_Toc471758236)

[1.1. Общие сведения 8](#_Toc471758237)

[1.2. История вопроса 8](#_Toc471758238)

[1.3. Современные проблемы 10](#_Toc471758239)

[1.4. Пути решения проблем 11](#_Toc471758240)

[2. ОПИСАНИЕ ИССЛЕДОВАНИЯ 12](#_Toc471758241)

[2.1. Исследуемые методы оценки качества кластеризации 12](#_Toc471758242)

# ВВЕДЕНИЕ

На сегодняшний день существуют различные коэффициенты, расчёт которых позволяет оценить качество кластеризации. К таким коэффициентам относится F1-мера, … Актуальность оценки кластеризации определяется необходимостью оценивать корректность разбиения на кластеры с целью проверки того, что полученному результату можно доверять. Основной целью работы является реализация и анализ различных методов оценки качества кластеризации. В результате работы различные методы будут реализованы и протестированы, а также измерено время их выполнения на множествах различных размерностей, что позволит определить, какие из методов являются наилучшими.

# 1. СОВРЕМЕННОЕ СОСТОЯНИЕ ВОПРОСА

## 1.1. Общие сведения

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

Задачи кластеризации:

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

Кластеризация может использоваться в таких областях, как:

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

## 1.2. История вопроса

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

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

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

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

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

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

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

## 1.3. Современные проблемы

1. Проблема обоснования качества результатов кластеризации.

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

1. Сложность формализации некоторых областей

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

1. Проблема анализа большого количества разнотипных признаков.

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

1. Проблема поиска оптимального решения.

Многие алгоритмы не гарантируют, что найденное решение является оптимальным.

## 1.4. Пути решения проблем

1. Проблема обоснования качества результатов кластеризации.

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

1. Проблема анализа большого количества признаков.

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

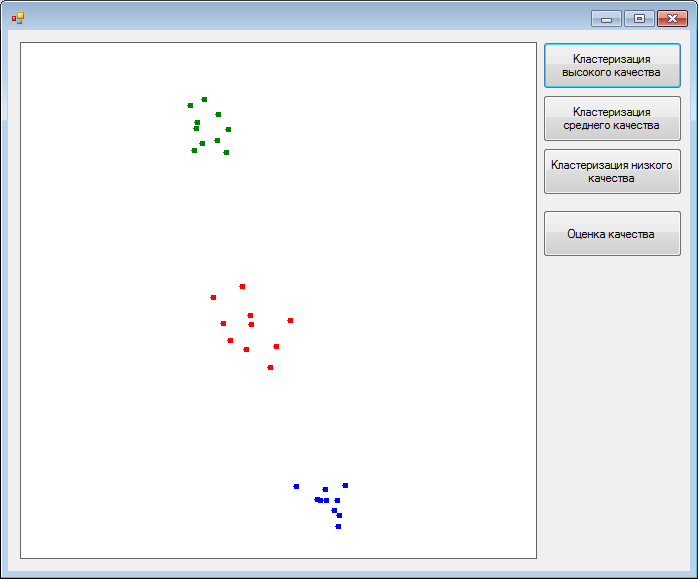
1. Проблема поиска оптимального решения.

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

# 2. ОПИСАНИЕ ИССЛЕДОВАНИЯ

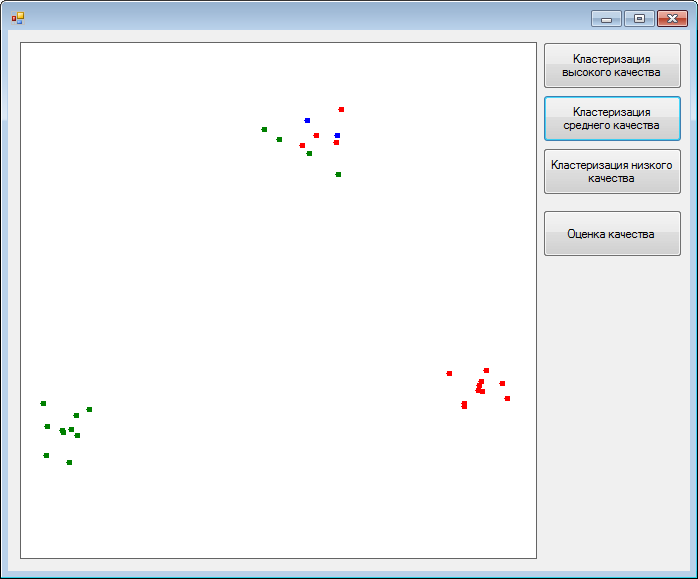
## 2.1. Описание генерации кластеризованных множеств

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



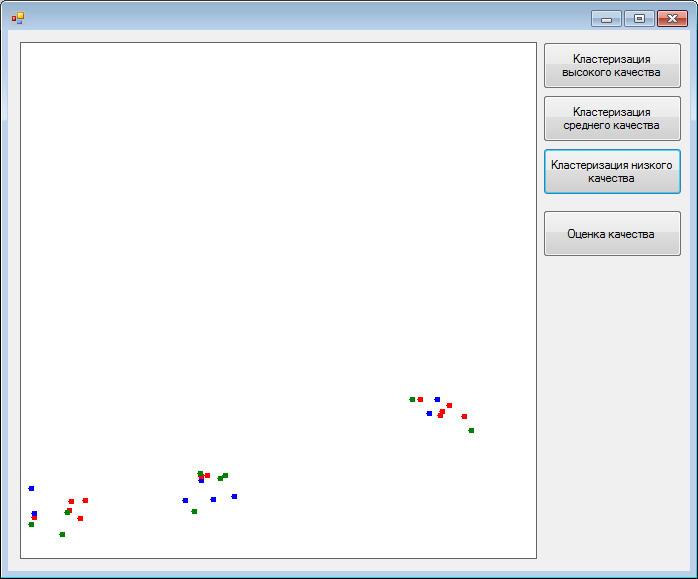
*Рисунок 1.Пример сгенерированного кластеризованного множества высокого качества.*

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



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

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



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

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

## 2.2. Исследуемые методы оценки качества кластеризации

Были реализованы и исследованы следующие методы:

1. F1-мера

Для применения F1-меры необходима информация о классах кластеризуемых объектов. В качестве примера можно рассмотреть программную систему, распознающую автомобили и пешеходов. В данном случае имеются два класса: автомобили и пешеходы. При проведении кластеризации логично ожидать, что в результате получится два кластера, один из которых будет содержать только автомобили, а второй – только пешеходов. Для определения степени соответствия результата кластеризации ожидаемому результату и применяется расчёт F1-меры.

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

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

где – количество таких объектов , таких что и ; ;

Тогда F1-мера для кластера и класса будет равна:

F1-мера для всего кластеризуемого множества равна:

# 3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

## 3.1. Рассчитанные значения коэффициентов оценки качества кластеризации

*Таблица 1.Значения коэффициентов оценки качетсва кластеризации..*

|  |  |  |  |
| --- | --- | --- | --- |
| Коэффициент | Значение для случая кластеризации высокого качества | Значение для случая кластеризации среднего качества | Значение для случая кластеризации низкого качества |
| F1-мера | 1 | 0,8 | 0,48 |