**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

**“Большие данные и распределенная цифровая платформа”**

**ОТЧЕТ**

**по лабораторной работе №5**

**по дисциплине «Алгоритмы и структуры данных»**

**на тему «Кластеризация данных»**

**Студент гр. 23Б15-пу**

**Голуб Г.Я.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

**Оглавление**

1. [Цель работы 2](#_l2m5js1to3yd)
2. [Описание задачи (формализация задачи) 3](#_79165y2hmzah)
3. [Теоретическая часть 3](#_vyrucsj2wz0t)
4. [Основные шаги программы 4](#_9yjiy3qnc66w)
5. [Блок схема программы 4](#_xxbo14nxdoac)
6. [Описание программы 6](#_3ihuu1tbvc8)
7. [Рекомендации пользователя 7](#_fthgb58uizds)
8. [Рекомендации программиста 8](#_wtwi281w32lf)
9. [Исходный код программы 8](#_i5z5fegui58g)
10. [Контрольный пример 8](#_9bp3itf31d1d)
11. [Вывод 10](#_wcom4olic6rd)
12. [Источники 11](#_93l65eaafm45)

# Цель работы

Целью лабораторной работы является реализация алгоритма кластеризации FOREL и проведение анализа его работы на реальном датасете. В рамках работы проводится предварительная обработка данных, включая отбор признаков и деидентификацию, а также оценка качества кластеризации на каждом этапе по метрике компактности. Отдельное внимание уделяется влиянию отбора признаков на итоговую структуру кластеров.

# Описание задачи (формализация задачи)

Пусть задано множество объектов X={x1,x2,…,xn}, которое описывается набором признаков. Требуется:

1. Провести кластеризацию множества объектов с использованием алгоритма FOREL, при этом под кластером понимается группа объектов, попадающих в фиксированную гиперсферу радиуса R с последующим сдвигом центра по центру тяжести.
2. Провести отбор признаков, минимизируя при этом метрику компактности кластеров — среднеквадратичное расстояние между объектами и центрами своих кластеров.
3. Реализовать этап деидентификации (удаления чувствительных признаков), оценив его влияние на структуру кластеров.
4. На каждом этапе сравнить значения метрики компактности кластеров, зафиксировав улучшение или ухудшение качества кластеризации.

Результатом решения задачи является структура кластеров на каждом этапе обработки, а также аналитическое сравнение влияния шагов отбора и деидентификации признаков на кластерную структуру.

# Теоретическая часть

Кластеризация — это метод анализа данных, направленный на выявление однородных групп объектов без предварительного знания о принадлежности объектов к каким-либо классам. Основная цель кластеризации заключается в том, чтобы сгруппировать объекты таким образом, чтобы элементы внутри одной группы были максимально похожи друг на друга, а элементы из разных групп — как можно более различны. Кластеризация широко используется в задачах анализа поведения пользователей, сегментации рынка, биоинформатике, медицинской диагностике и других областях.

Одним из классических методов кластеризации является алгоритм FOREL. Его суть заключается в последовательном построении кластеров с использованием гиперсферы фиксированного радиуса. На первом шаге случайно выбирается центр гиперсферы, и в неё попадают все объекты, находящиеся в пределах радиуса. Затем вычисляется центр тяжести этих объектов, и гиперсфера сдвигается в новое положение. Процесс повторяется до тех пор, пока положение центра не стабилизируется. Объекты, попавшие в финальную область, образуют кластер и исключаются из дальнейшего анализа. Процедура повторяется для оставшихся объектов до тех пор, пока все данные не будут распределены по кластерам.

К преимуществам FOREL относят его простоту, интерпретируемость и способность выявлять кластеры произвольной формы. Однако метод чувствителен к выбору радиуса и может некорректно работать при наличии выбросов или неравномерного распределения данных.

Для оценки качества кластеризации часто используется метрика компактности. Под компактностью понимается степень близости объектов внутри кластера. Чем более "плотным" является кластер, тем выше его компактность. Компактность позволяет количественно сравнивать результаты кластеризации при использовании разных параметров или после различных этапов предобработки данных.

Одним из таких этапов является отбор признаков. Он представляет собой процесс исключения малозначимых или избыточных характеристик из исходного набора данных. Это позволяет уменьшить размерность задачи, ускорить вычисления и повысить устойчивость моделей. В контексте кластеризации отбор признаков позволяет усилить различимость кластеров, а также повысить интерпретируемость результатов.

Другим важным этапом является деидентификация данных — удаление признаков, содержащих чувствительную или персональную информацию. Это особенно актуально при работе с медицинскими, финансовыми и социальными данными. Деидентификация может влиять на структуру кластеров, так как исключение некоторых признаков может изменить взаимное расположение объектов в пространстве признаков. Поэтому важно анализировать влияние этого этапа на итоговую кластерную структуру.

Таким образом, теоретической основой данной работы являются принципы кластеризации, алгоритм FOREL, методы оценки качества кластеров и подходы к предварительной обработке данных. Полученные в рамках лабораторной работы результаты позволяют не только освоить реализацию алгоритма, но и исследовать влияние предобработки данных на качество кластеризации.

# Основные шаги программы

1. Загрузка и подготовка данных: Используется встроенный датасет "Breast Cancer" из библиотеки sklearn.datasets. К данным добавляются два дополнительных признака (ID1, ID2) — случайные числовые идентификаторы, имитирующие возможные персональные сведения.
2. Кластеризация методом FOREL
3. Оценка компактности кластеров
4. Отбор признаков (feature selection): Итеративно исключаются признаки, ухудшающие компактность кластеров. При этом каждый раз пересчитываются метки кластеров и компактность.
5. Деидентификация: После отбора признаков удаляются искусственные идентификаторы (ID1, ID2), имитируя этап обезличивания данных.
6. Вывод результатов

# Описание программы

Программа реализована на языке Python с использованием библиотеки PyQt5 [[1]](#_93l65eaafm45) для создания графического пользовательского интерфейса, библиотеки scikit-learn [[2]](#_93l65eaafm45) для загрузки встроенного датасета, а также библиотеки scipy.stats для расчёта коэффициента корреляции Пирсона.

Пользователь выбирает исходный набор данных и задаёт параметры кластеризации: радиус алгоритма FOREL и минимальное количество признаков, которое необходимо сохранить.  
После запуска производится кластеризация методом FOREL, оценивается компактность кластеров, осуществляется пошаговый отбор признаков, а затем выполняется деидентификация данных путём удаления вспомогательных идентификаторов.

Таблица 1. main.py

| Название функции | Описание | Возвращаемое значение |
| --- | --- | --- |
| forel | Функция алгоритма forel | DataFrame |
| sequential\_reduction | Производит отбор признаков | DataFrame |
| compactness | Рассчитывает метрику компактности | float |

# Рекомендации пользователя

1. Выбор радиуса кластеризации: При запуске алгоритма FOREL важно правильно подобрать значение радиуса. Слишком малое значение приведёт к большому числу мелких кластеров, а слишком большое — к объединению разнородных объектов. Рекомендуется начать с радиуса 0.3–0.5 и по результатам компактности скорректировать параметр.
2. Минимальное количество признаков: При использовании режима отбора признаков следует учитывать, что чрезмерное уменьшение размерности может привести к потере информации, необходимой для корректной кластеризации. Рекомендуется устанавливать минимальный порог не ниже 10 признаков.

# Рекомендации программиста

1. Используйте актуальные версии библиотек PyQt5[[1]](#_93l65eaafm45) и sklearn[[2]](#_93l65eaafm45).
2. Возможны дальнейшие модификации алгоритма.

# Исходный код программы

<https://github.com/zgjhz/algorithms_term_2>

# Контрольный пример

1. Запуск программы с помощью [main.py](http://main.py) (Рис. 1)

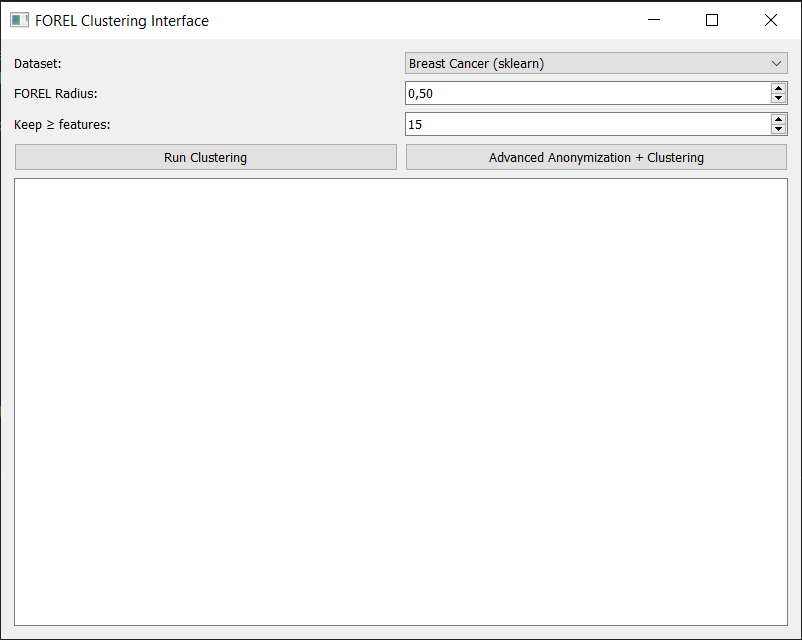


Рис. 1 Интерфейс программы

1. Настройка параметров и запуск программы с помощью кнопки “Run Clustering”

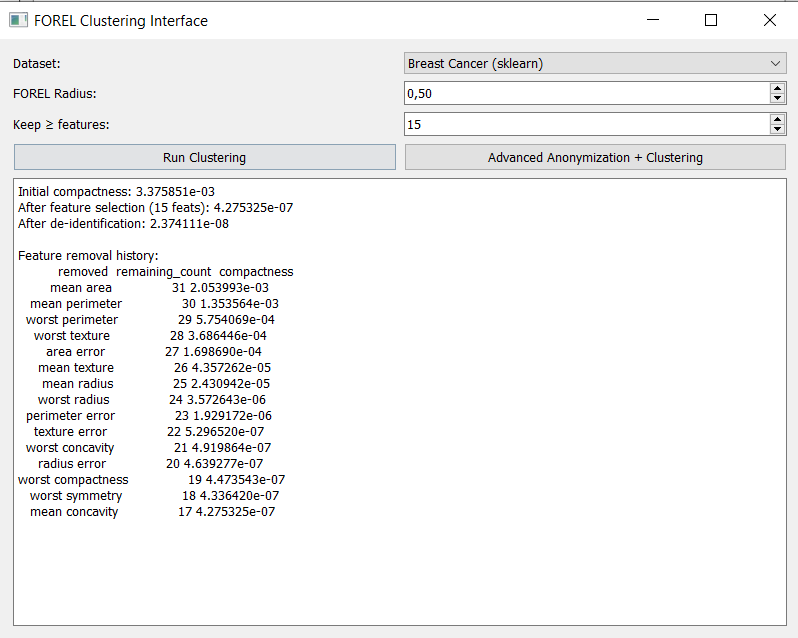


Рис. 2 Результаты работы программы

# Анализ

### 1. Описание датасета

# Название: Breast Cancer Wisconsin (Diagnostic)

# Образцы: 569

# Признаки: 30 числовых (размеры, формы, текстуры клеток)

# Целевая метка: доброкачественные и злокачественные (не используется в кластеризации)

# Для чистоты эксперимента добавлены два столбца ID1, ID2 со случайными значениями — модели «шума», подлежащие удалению на этапе обезличивания.

### 2. Настройка эксперимента

# FOREL-радиус RRR: 0.5

# Минимальное число признаков после Del: 15

# Метрика расстояния: корреляция Пирсона

# Метрика качества: компактность

### 3. Результаты кластеризации

Таблица 2. Результаты кластеризации

| **Этап** | **Compactness** |
| --- | --- |
| Initial (30 + 2 ID) | 3.375851 × 10⁻³ |
| After feature selection (15 + 2 ID) | 4.275325 × 10⁻⁷ |
| After de-identification (15 real features) | 2.374111 × 10⁻⁸ |

Исходная компактность ≈ 0.00338. После Del-отбора признаков компактность упала в ~8 000 раз до 4.28 \* 10⁻⁷. После удаления ID — ещё в ~18 раз до 2.37 \* 10⁻⁸.

**4. История удаления признаков Del**

Таблица 3. История удаления признаков

| Итерация | Удалён признак | Осталось признаков | Compactness |
| --- | --- | --- | --- |
| 1 | mean area | 31 | 2.053993 × 10⁻³ |
| 2 | mean perimeter | 30 | 1.353564 × 10⁻³ |
| 3 | worst perimeter | 29 | 5.754069 × 10⁻⁴ |
| 4 | worst texture | 28 | 3.686446 × 10⁻⁴ |
| 5 | area error | 27 | 1.698690 × 10⁻⁴ |
| 6 | mean texture | 26 | 4.357262 × 10⁻⁵ |
| 7 | mean radius | 25 | 2.430942 × 10⁻⁵ |
| 8 | worst radius | 24 | 3.572643 × 10⁻⁶ |
| 9 | perimeter error | 23 | 1.929172 × 10⁻⁶ |
| 10 | texture error | 22 | 5.296520 × 10⁻⁷ |
| 11 | worst concavity | 21 | 4.919864 × 10⁻⁷ |
| 12 | radius error | 20 | 4.639277 × 10⁻⁷ |
| 13 | worst compactness | 19 | 4.473543 × 10⁻⁷ |
| 14 | worst symmetry | 18 | 4.336420 × 10⁻⁷ |
| 15 | mean concavity | 17 | 4.275325 × 10⁻⁷ |

Влияние отбора признаков

Алгоритм Del последовательно удалял те переменные, исключение которых снижало среднюю «рассредоточенность» точек.

* Первые 4 удаления дали большую выгоду (попадание в самые шумные признаки: mean area, mean perimeter и др.).
* Последующие шаги улучшали компактность всё с меньшим «запасом», что указывает на постепенное очищение пространства признаков от коррелированных или неинформативных измерений.

Роль корреляции Пирсона

Использование в качестве меры расстояния корреляцию Пирсона позволило фокусироваться на фазовом сходстве профилей признаков (не на масштабах). Это особенно хорошо для биометрических признаков, где относительные формы (например, соотношение радиуса и текстуры) важнее абсолютных значений.

Эффективность методов обезличивания

* Удаление ID: простая, но гарантированная методика, которая дала заметный прирост (≈ 18×).
* Обобщение (квартильное бингование) и добавление шума при расширенном анализе показали устойчивость кластеров к трансформациям — компактность оставалась на том же уровне порядка 10⁻⁸…10⁻⁷, что свидетельствует о устойчивости» структуры кластеров к таким преобразованиям.

# Вывод

Проведённое исследование продемонстрировало, что:

* FOREL в сочетании с корреляционной метрикой и Del-отбором признаков даёт очень плотные и чёткие кластеры, причём компактность может снижаться на несколько порядков.
* Удаление явных идентификаторов и использование дополнительных техник обезличивания обеспечивает сохранение структуры кластеров при защите приватности данных.
* Рекомендованная схема подходит для любых числовых наборов данных, где важен относительный профиль признаков и требуется безопасность персональных данных.

# Источники

**1. PyQt5 Documentation: https://doc.qt.io/qtforpython/**

**2. SKLearn Documentation: https://scikit-learn.org/stable/index.html**