Лабораторная работа № 4

«Сравнение методов кластеризации»

Цель работы

Освоить методы построения и сравнения различных алгоритмов кластеризации: центроидных, иерархических и плотностных. Изучить методы определения оптимального числа кластеров и оценки качества кластеризации.

Задачи

1. Изучить основы методов кластеризации и их применение для анализа данных
2. Освоить построение центроидных методов (K-Means)
3. Изучить методы иерархической кластеризации
4. Освоить плотностные методы кластеризации (DBSCAN)
5. Изучить методы определения оптимального числа кластеров
6. Сравнить качество различных алгоритмов кластеризации

Исходные данные (выбрать один из датасетов)

- Вариант 1: Iris Plants Dataset
- Вариант 2: Wine Recognition Dataset
- Вариант 3: Breast Cancer Wisconsin Dataset
- Вариант 4: Synthetic blobs (генерируемые данные)

Описание выбранного датасета

- Название датасета: [указать выбранный вариант]
- Описание: [краткое описание данных]
- Количество наблюдений: [указать количество]
- Количество признаков: [указать количество]
- Тип признаков: [указать типы - непрерывные, категориальные]
- Истинное число кластеров (если известно): [указать количество]

Методика выполнения

Часть 1: Подготовка данных и предварительный анализ

1. Загрузка и предобработка данных
2. Анализ распределения данных
3. Исследование корреляций между признаками
4. Визуализация данных в 2D/3D пространстве
5. Масштабирование признаков

Часть 2: Определение оптимального числа кластеров

1. Метод локтя (Elbow Method) для K-Means
2. Анализ силуэтов (Silhouette Analysis)
3. Индекс Калински-Харабаша (Calinski-Harabasz Index)
4. Индекс Дэвиса-Болдина (Davies-Bouldin Index)

Часть 3: Центроидные методы

1. Построение модели K-Means с оптимальным числом кластеров
2. Визуализация результатов кластеризации

Часть 4: Иерархические методы

1. Построение дендрограммы
2. Агломеративная кластеризация с различными функциями связи:
   - Single linkage
   - Complete linkage
   - Average linkage
   - Ward's method
3. Сравнение качества различных методов связи
4. Визуализация дендрограмм и результатов кластеризации

Часть 5: Плотностные методы

1. Построение модели DBSCAN
2. Подбор параметров ε и min_samples
3. Анализ k-distance графика для выбора ε
4. Оценка качества плотностной кластеризации
5. Визуализация результатов и обнаруженных шумовых точек

Часть 6: Сравнительный анализ

1. Сравнение всех алгоритмов по метрикам качества
2. Анализ стабильности алгоритмов
3. Сравнение времени выполнения
4. Анализ чувствительности к параметрам
5. Формулирование практических рекомендаций

Метрики оценки качества

Для сравнения алгоритмов использовать следующие метрики:

- **Silhouette Score** - средняя мера качества кластеризации
- **Calinski-Harabasz Index** - отношение межкластерной дисперсии к внутрикластерной
- **Davies-Bouldin Index** - среднее сходство между кластерами
- **Adjusted Rand Index (ARI)** - мера сходства с истинными метками (если известны)
- **Время выполнения** - время работы алгоритма

Требования к отчету

Содержание отчета

1. Титульный лист
2. Цель и задачи работы
3. Описание выбранного датасета
4. Методика исследования
5. Результаты по каждой части:
    - Код реализации
    - Таблицы с результатами
    - Графики и визуализации
    - Статистические выводы
6. Сравнительный анализ всех алгоритмов
7. Выводы и практические рекомендации

Визуализация

- Графики метода локтя и анализа силуэтов
- Дендрограммы для иерархической кластеризации
- K-distance график для DBSCAN
- Визуализация результатов кластеризации в 2D/3D
- Сравнение метрик качества на графиках

Аналитические таблицы

Сводная таблица результатов

Модель	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**	**ARI**	**Время (с)**
K-Means					
DBSCAN					
Single					
Complete					
Average					
Ward					

Таблица 1: Сравнение метрик качества алгоритмов кластеризации

Результаты определения числа кластеров

Метод	**Оптимальное k**	**Значение метрики**	**Интерпретация**
Elbow Method			
Silhouette Analysis			
Calinski-Harabasz			
Davies-Bouldin			

Таблица 2: Результаты методов определения оптимального числа кластеров