Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра «Системное программирование»

ОТЧЁТ

по лабораторной работе №10

на тему «Иерархическая кластеризация»

Выполнил
Студент группы КЭ-120
_____ Д.А. Снегирева
«____»_____2021 г.
Email: dashasneg@mail.ru

Челябинск 2021

ЗАДАНИЕ

Выполните иерархическую кластеризацию набора данных, используя различные меры схожести: Single linkage, Complete linkage, Group average, расстояние Уорда (Ward).

Выполните визуализацию полученных результатов в виде дендрограмм.

СОДЕРЖАНИЕ

ЗАДАНИЕ 2
1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ 4
2 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ..... 5

1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ

В данной работе был использован набор данных по рукописным цифрам UCI ML hand-written digits datasets (https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits), состоящий из десяти классов. Изображения рукописных цифр в наборе представлены матрицей 8 x 8 (интенсивности белого цвета для каждого пикселя). Далее эта матрица "разворачивается" в вектор длины 64, получается признаковое описание объекта. С помощью PCA размерность была снижена до 2 признаков.

Также, был использован набор данных Ирисы Фишера (https://archive.ics.uci.edu/ml/datasets/iris). После понижения размерности с помощью PCA до 2 признаков, данные образуют 3 группы, которые имеют вытянутую форму.

В качестве средств реализации были использованы библиотеки scikit-learn и scikit-learn-extra.

Репозиторий задания: https://github.com/DasHaSneg/BigDataMiningCourse

Каталог задания: 10 hierarchical clustering

2 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

На рисунке 1 приведен набор данных ирисы Фишера, сниженный до 2 признаков с помощью PCA. Можно заметить, что кластеры имеют вытянутую форму.

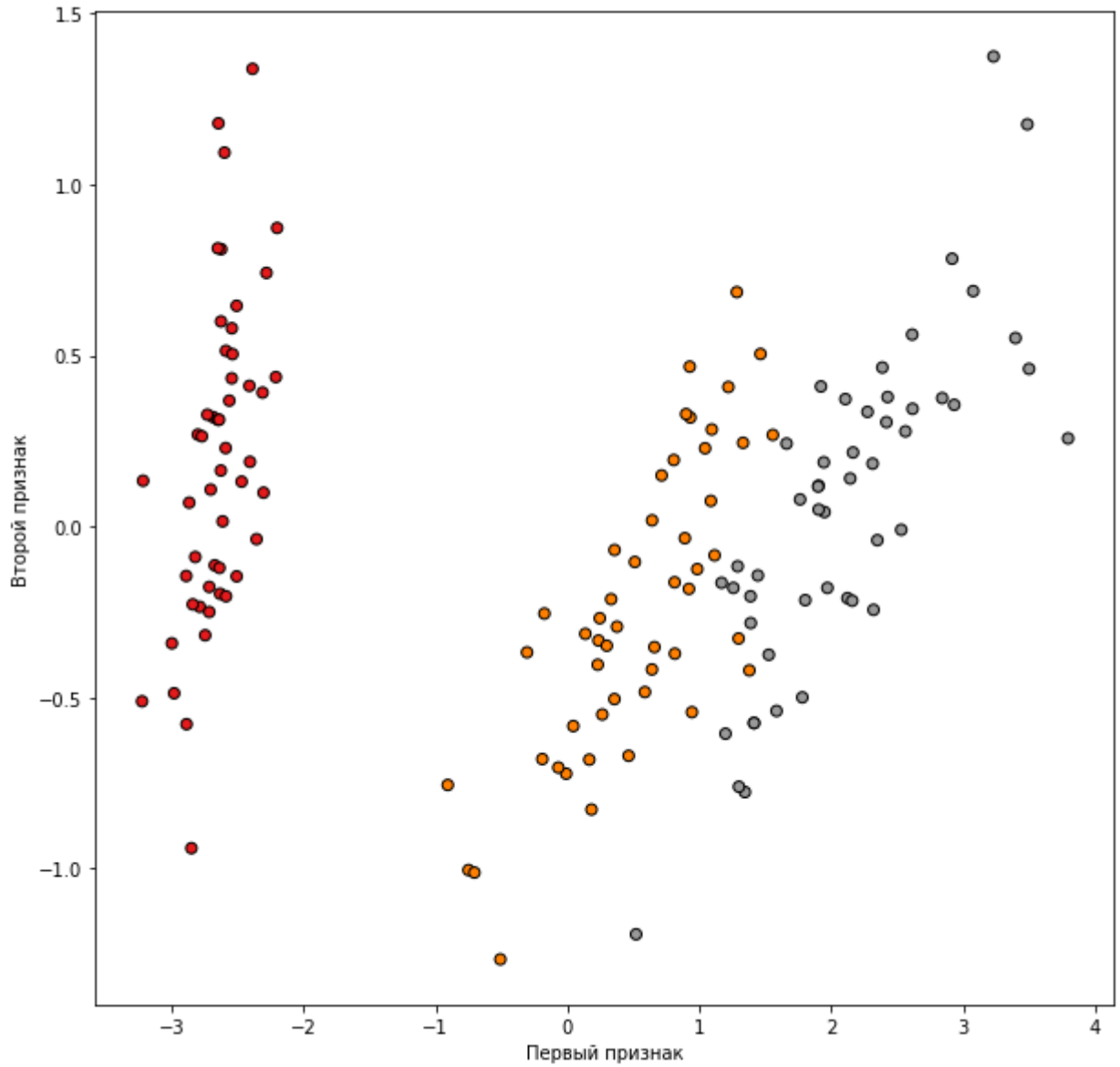

Рис. 1. Набор данных с ирисами Фишера

Далее была выполнена иерархическая кластеризация данного набора с использованием мер схожести Single linkage, Complete linkage, Group average, расстояние Уорда (Ward) и значением параметра n_clusters равным 3. Построенные дендрограммы и результаты кластеризации для каждой меры схожести соответственно приведены на рисунке 2.

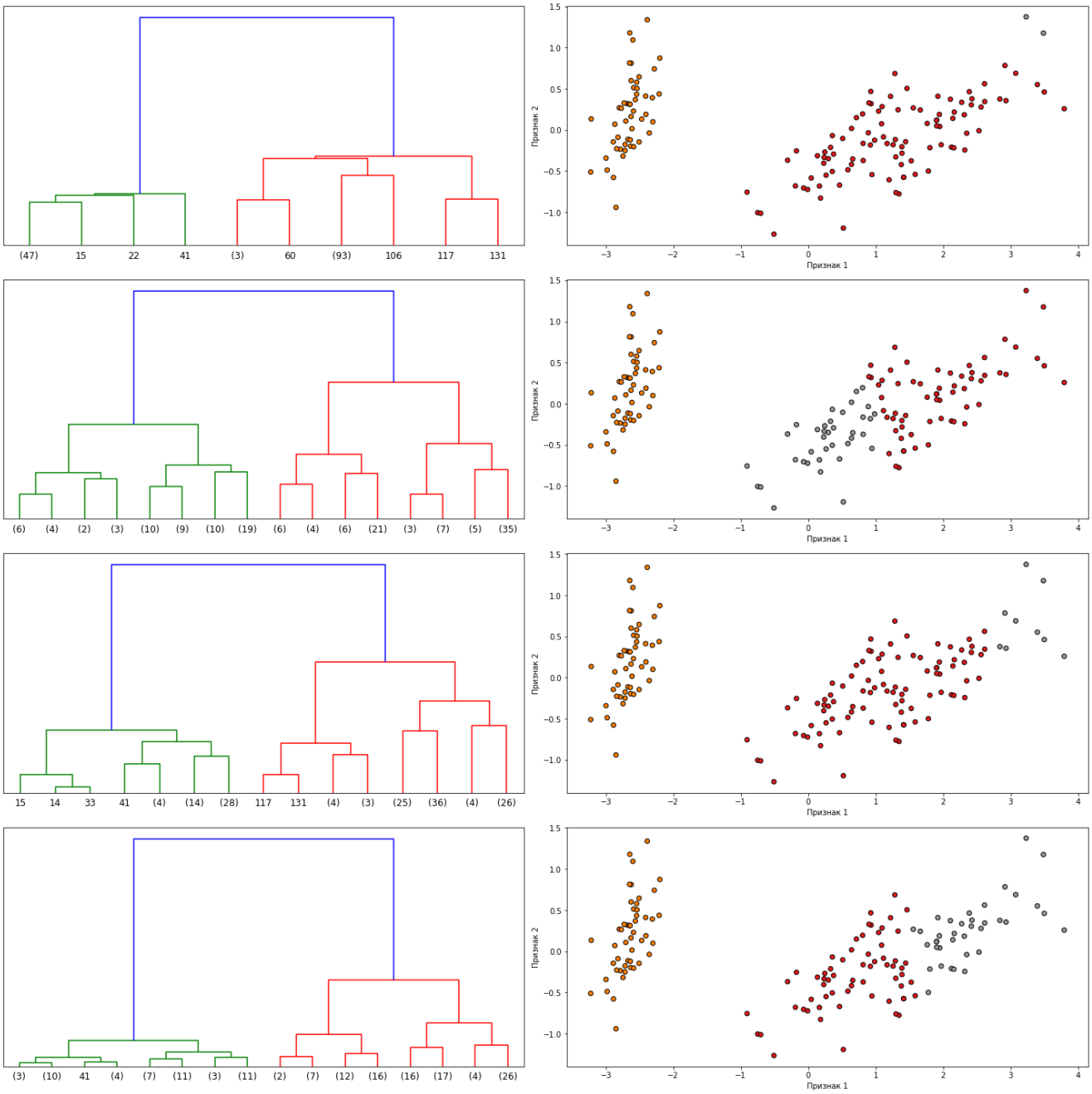

Рис. 2. Дендрограммы и результаты кластеризации для набора данных Ирисы Фишера

Из рисунка видно, что ни одна мера схожести не помогла достигнуть точного результата. Похожий результат наблюдается при использовании расстояния Уорда (Ward).