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 системного программирования

ОТЧЁТ ПО ЗАДАНИЮ №10

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Иерархическая кластеризация

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

1. 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.

2. Краткие сведения о наборах данных

Использованный набор данных:

ECG Heartbeat Categorization Dataset (https://www.kaggle.com/shayanfazeli/heartbeat). Набор содержит информацию о сердцебиении здоровых пациентов и пациентов с различными формами аритмии. После приведения количества параметров к 2 с помощью PCA, пациенты образуют 5 групп, которые имеют вытянутую форму, препятствующую эффективной кластеризации с помощью алгоритма K-Means.

3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека scikit-learn, включающая в себя множество алгоритмов для анализа данных.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 10. HierCluster

4. Визуализация

На рис. 1 приведён набор данных о сердцебиении, приведённый к 2 измерениям с помощью PCA.

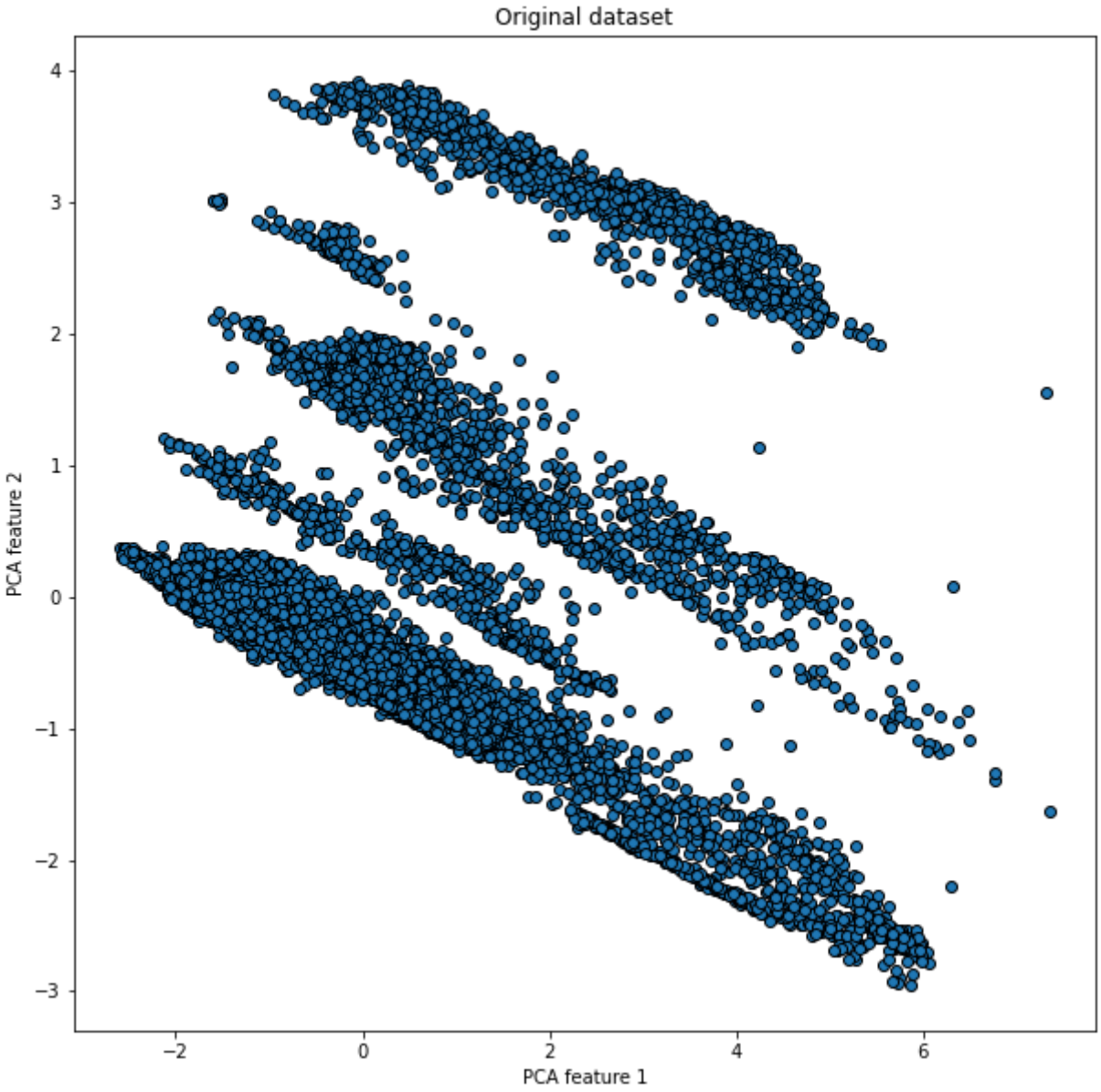

Рис. 1. Визуализация второго набора данных

Из рисунка можно увидеть, что данные о сердцебиении образуют 5 вытянутых кластеров.

Далее была выполнена иерархическая кластеризация выбранного набора с использованием мер схожести Single linkage, Complete linkage, Group average, расстояние Уорда (Ward). Значение параметра n_clusters было установлено на 5. Построенные дендрограммы и результаты кластеризации приведены на рис. 2.

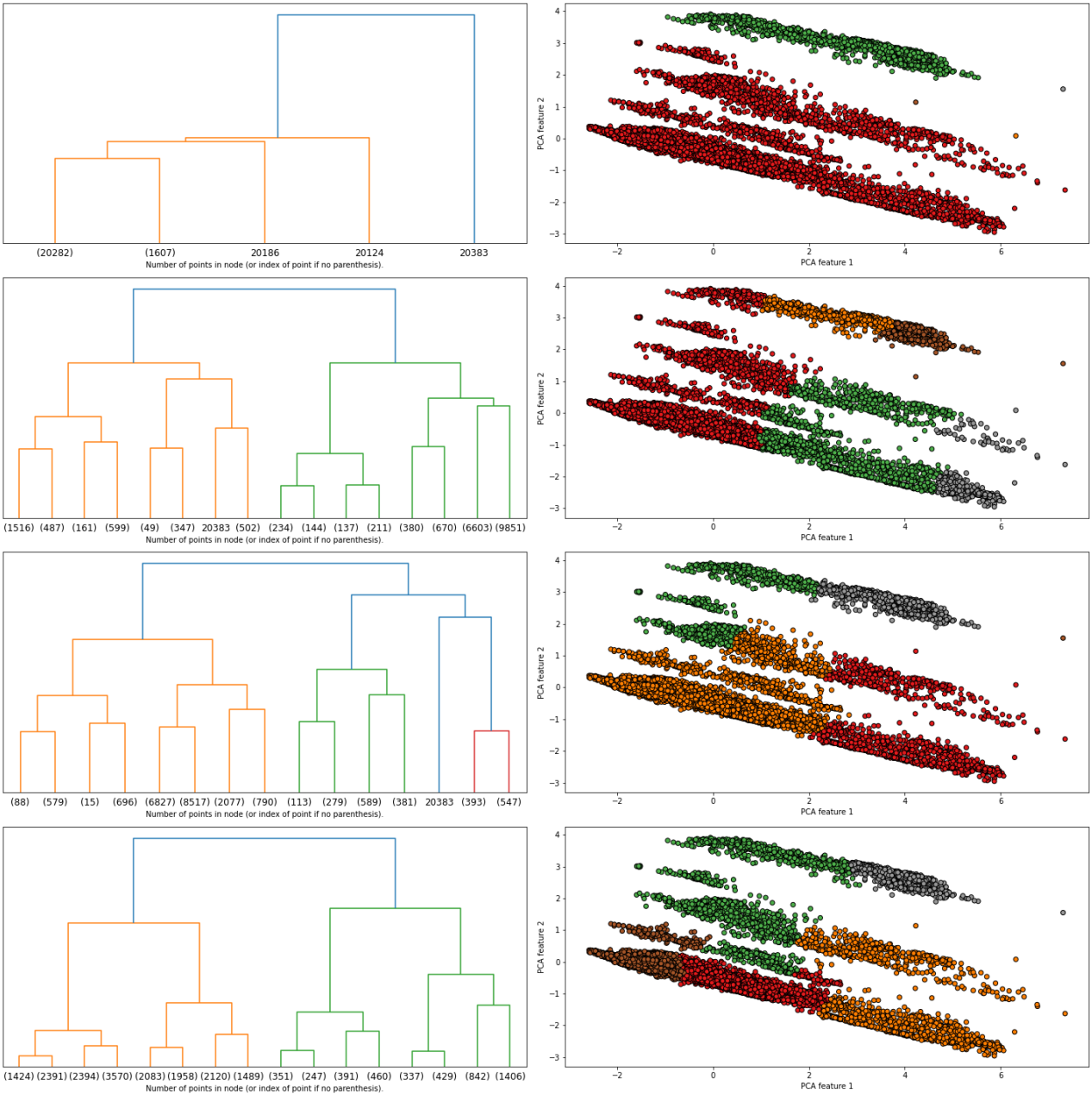

Рис. 2. Результаты кластеризации второго набора данных с помощью DBSCAN

Как видно из рисунка, иерархическая кластеризация не выделила верные кластеры ни с одной из использованных мер схожести. Почти все полученные результаты содержат разбиение реально существующих кластеров на части.

Таким образом, можно сделать вывод, что иерархическая кластеризация не очень хорошо подходит для выбранного набора данных.