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 №9

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Плотностная кластеризация

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

Челябинск 2021

1. Задание

Выполните кластеризацию набора 2-х или 3-мерных данных с помощью алгоритма DBSCAN (предполагается, что полученные кластеры не будут выпуклыми), используя различные значения параметров $MinPts$ (из интервала 3..9) и Eps. Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

Выполните кластеризацию зашумленного набора данных из задания 8 с помощью алгоритма DBSCAN, используя различные значения параметров $MinPts$ (из интервала 3..9) и Eps. Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

2. Краткие сведения о наборах данных

Использованные наборы данных:

Mobile Price Classification (https://www.kaggle.com/iabhishekofficial/mobile-price-classification). Набор содержит информацию о различных параметрах мобильного телефона, а также ценовую категорию (от 0 до 3). После приведения количества параметров к 2 с помощью PCA, телефоны образуют группы, напоминающие выпуклые кластеры.

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 были использованы библиотеки scikit-learn и scikit-learn-extra, включающие в себя множество алгоритмов для анализа данных.

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

Каталог для задания: 9. DenCluster

4. Визуализация показателей качества

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

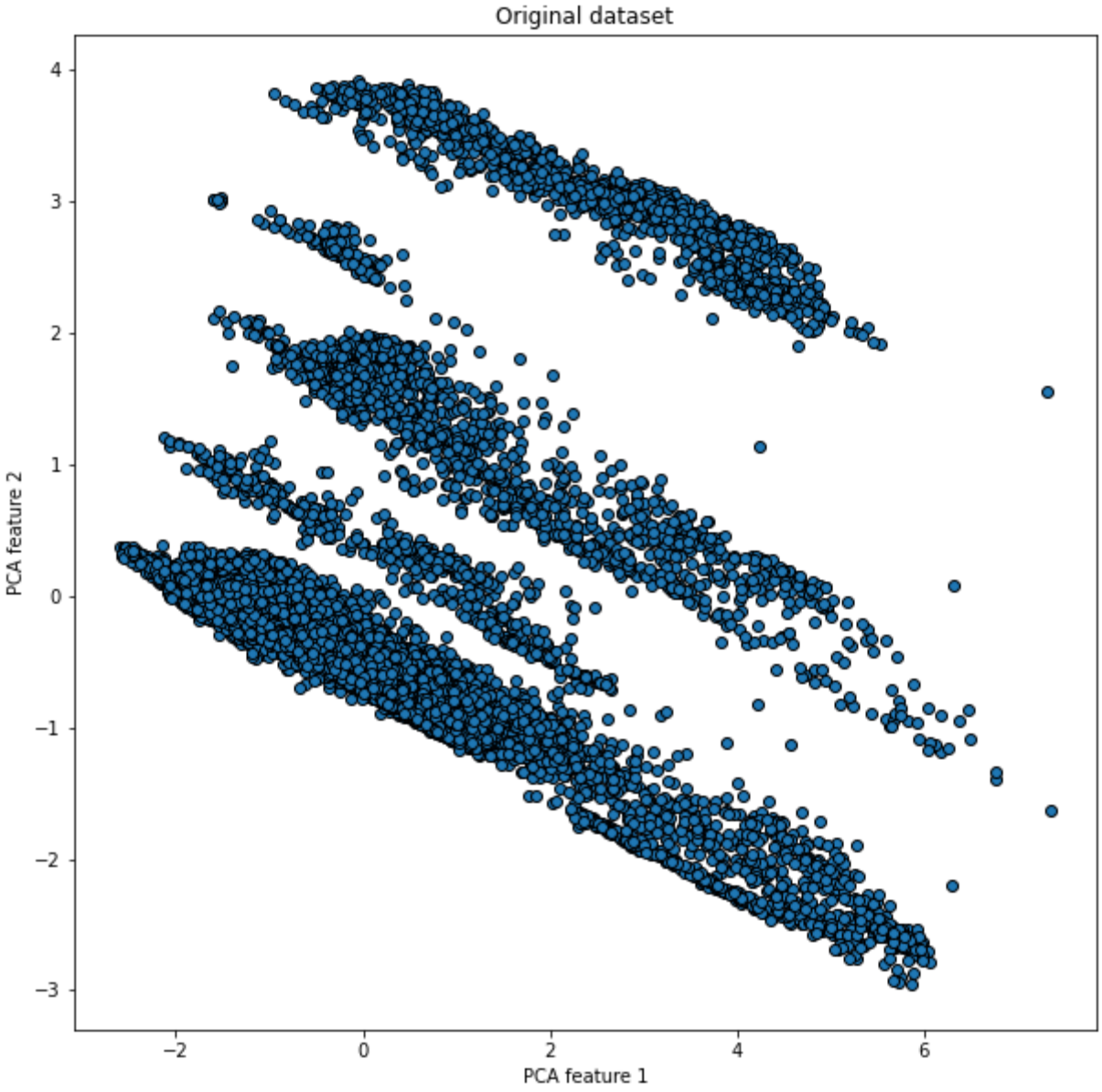

Рис. 1. Визуализация второго набора данных

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

Далее была выполнена кластеризация выбранного набора с помощью алгоритма DBSCAN со значениями параметра min_samples от 4 до 16 с шагом 2 и параметра eps от 0.1 до 0.2 с шагом 0.05. Результаты кластеризации приведены на рис. 2.

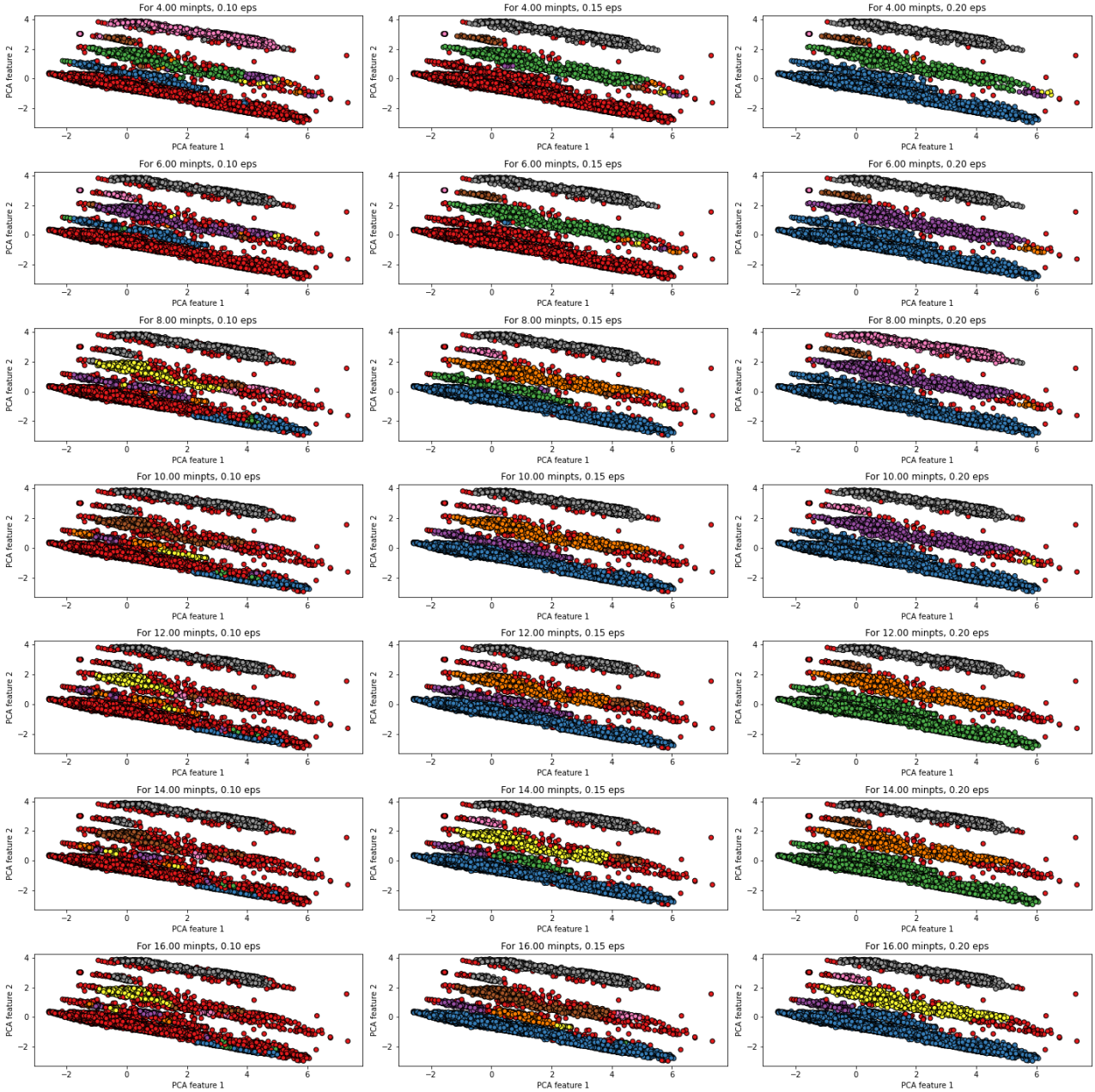

Рис. 2. Результаты кластеризации второго набора данных с помощью DBSCAN

Как видно из рисунка, наиболее удачными параметрами оказались min_samples = 8 и eps = 0.15. В целом, алгоритм DBSCAN показывает заметно лучшие результаты на этом наборе данных, чем k-Means в предыдущем задании.

Далее был загружен первый набор данных (о телефонах). Его визуализация приведена на рис. 3.

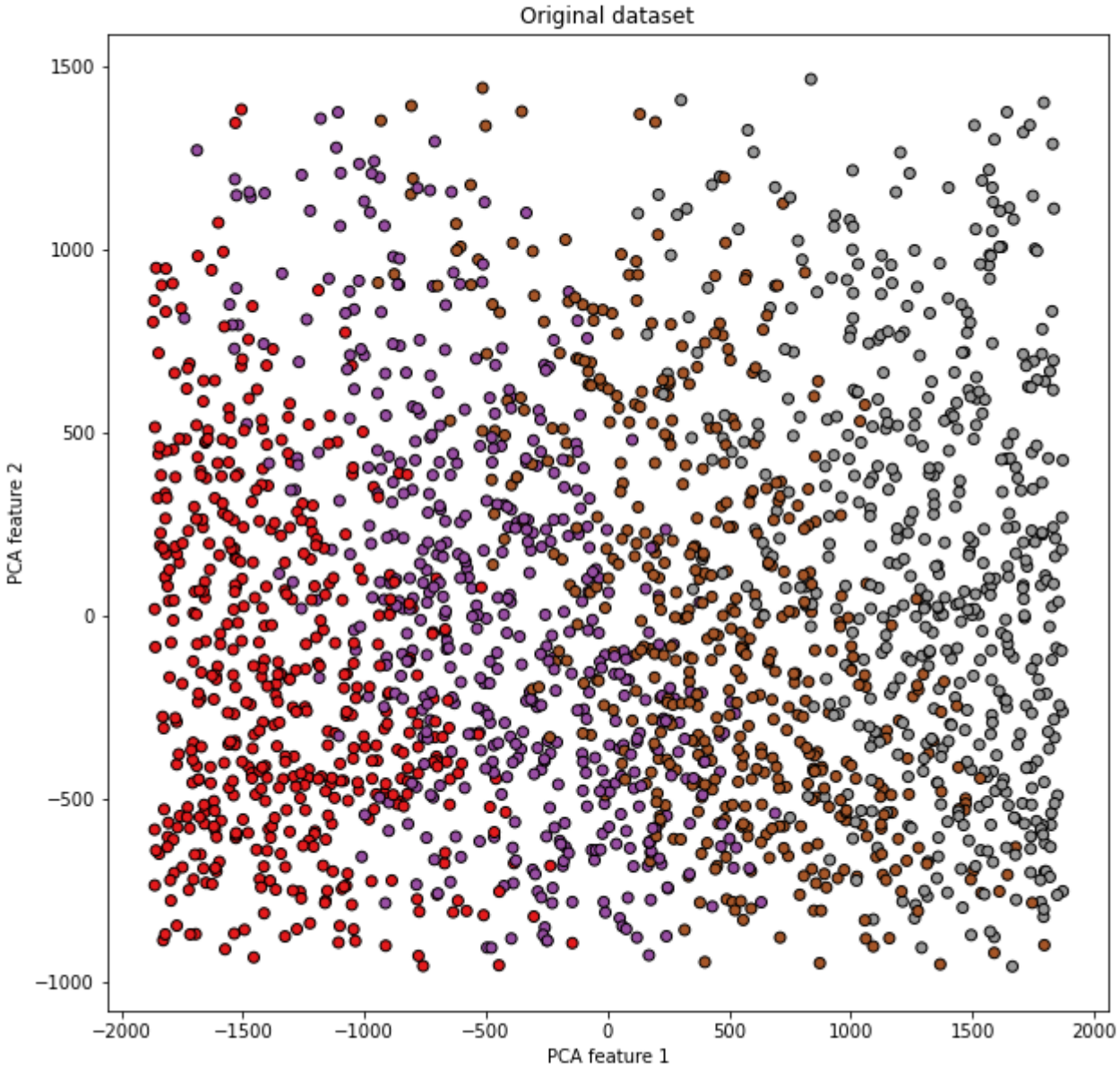

Рис. 3. Визуализация первого набора данных

На рисунке можно увидеть приблизительную форму кластеров.

Далее метод DBSCAN был применён для кластеризации набора данных из задания 8. При этом были использованы значения min_samples от 4 до 16 с шагом 2 и eps от 60 до 120 с шагом 30. Результаты кластеризации приведены на рис. 4.

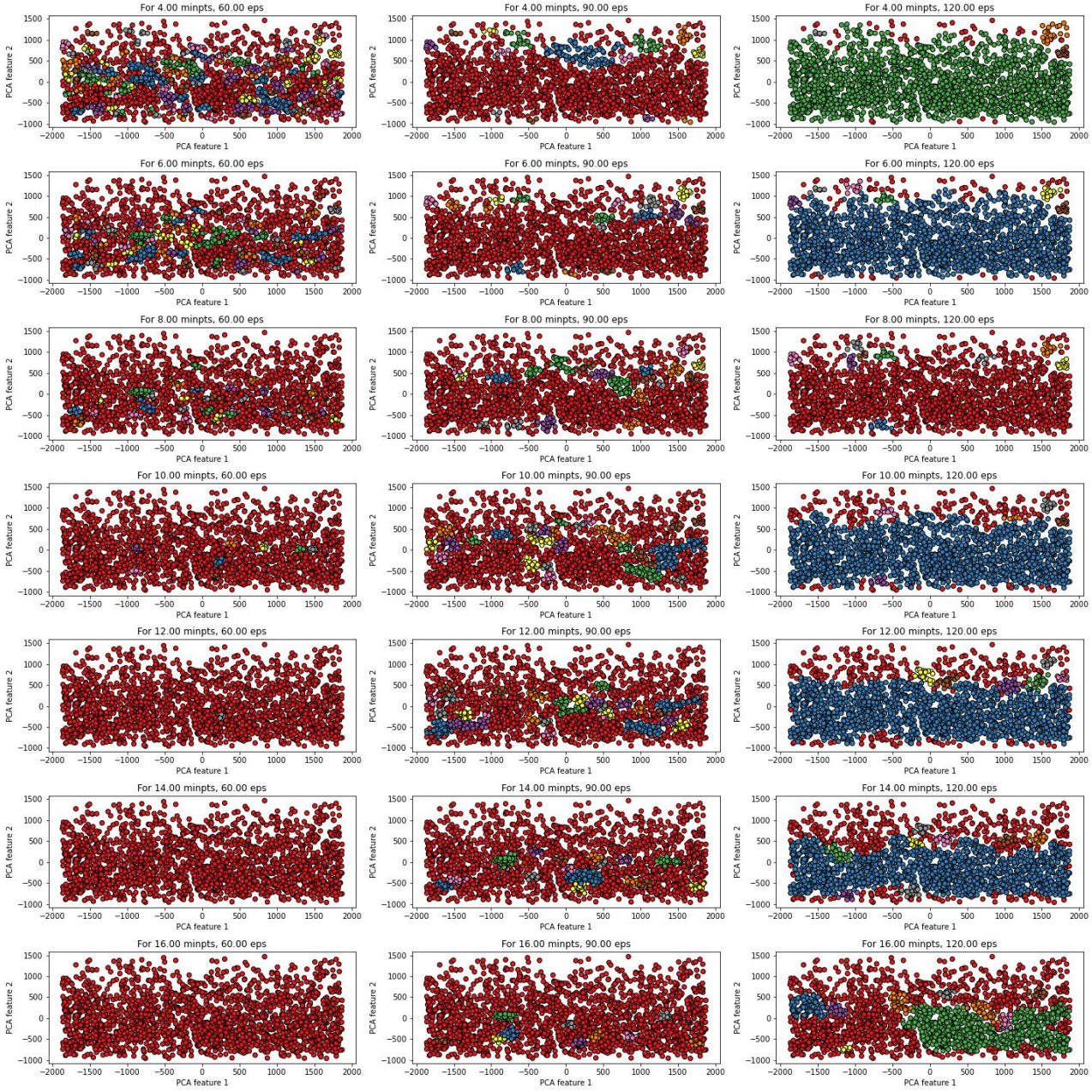

Рис. 4. Результаты кластеризации первого набора данных с помощью DBSCAN

Как можно увидеть из рисунка, хотя результаты кластеризации заметно меняются при изменении комбинаций параметров, ни одна из использованных комбинаций не привела к получению кластеров, близких к нужной форме.

Таким образом, алгоритм DBSCAN удобно применять для кластеризации наборов, в которых есть плотные кластеры, в том числе вытянутые кластеры. Для кластеризации неплотных наборов, в которых сложно выделить кластеры этот алгоритм не подходит.