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 №8

на тему «Разделительная кластеризация»

Выполнил
Студент группы КЭ-120
_____ Д.А. Снегирева
«____»_____2020 г.
Email: dashasneg@mail.ru

Челябинск 2021

ЗАДАНИЕ

1. Выполните кластеризацию набора 2-х или 3-мерных данных с помощью алгоритма k-Means (предполагается, что полученные кластеры будут выпуклыми), используя различные значения параметра k (из интервала 3..9). Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

2. Внесите шум в набор данных (случайным образом изменить определенную долю объектов набора: 1%, 3%, 5%, 10%; изменение может заключаться в добавлении/вычитании к/из одной/нескольких координат объекта случайного числа). Выполните кластеризацию зашумленного набора данных с помощью алгоритмов k-Means и k-Medoids (или PAM), используя различные значения параметра k (из интервала 3..9). Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

3. Выполните кластеризацию набора данных из задания 9 (с невыпуклыми кластерами) с помощью алгоритмов k-Means и k-Medoids (или PAM), используя различные значения параметра k (из интервала 3..9). Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

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: 8 separation clustering

2 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

На рисунке 1 приведен набор данных с рукописными цифрами, сниженный до 2 признаков с помощью PCA. Можно заметить, что даже на глаз рукописные цифры неплохо разделяются на кластеры (разные цвета точек означают разные кластеры).

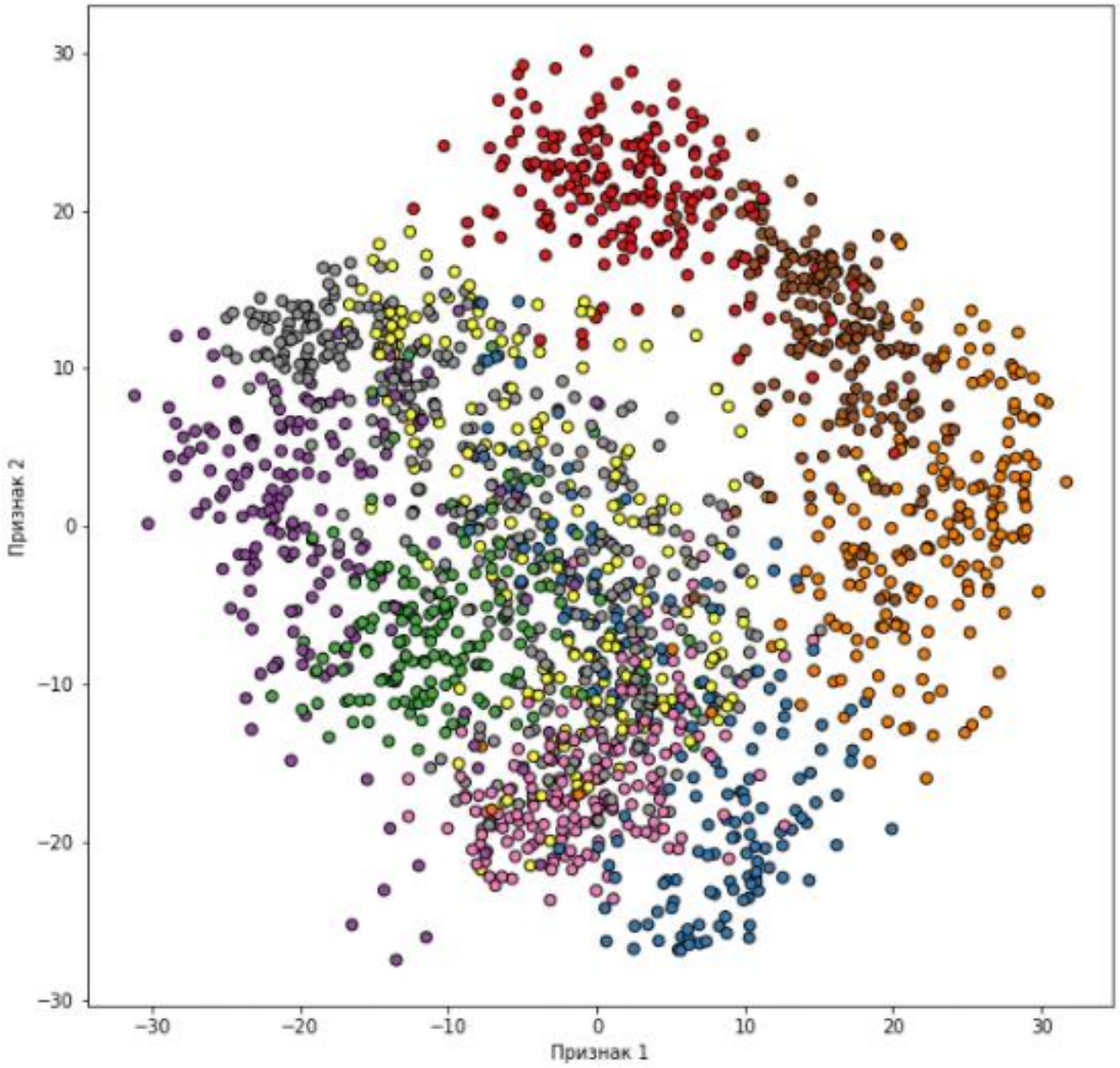

Рис. 1. Набор данных с рукописными цифрами

Затем была выполнена кластеризация набора данных с помощью алгоритма k-Means с использованием различных значений параметров k (из интервала от 3 до 9). Результаты приведены на рисунке 2.

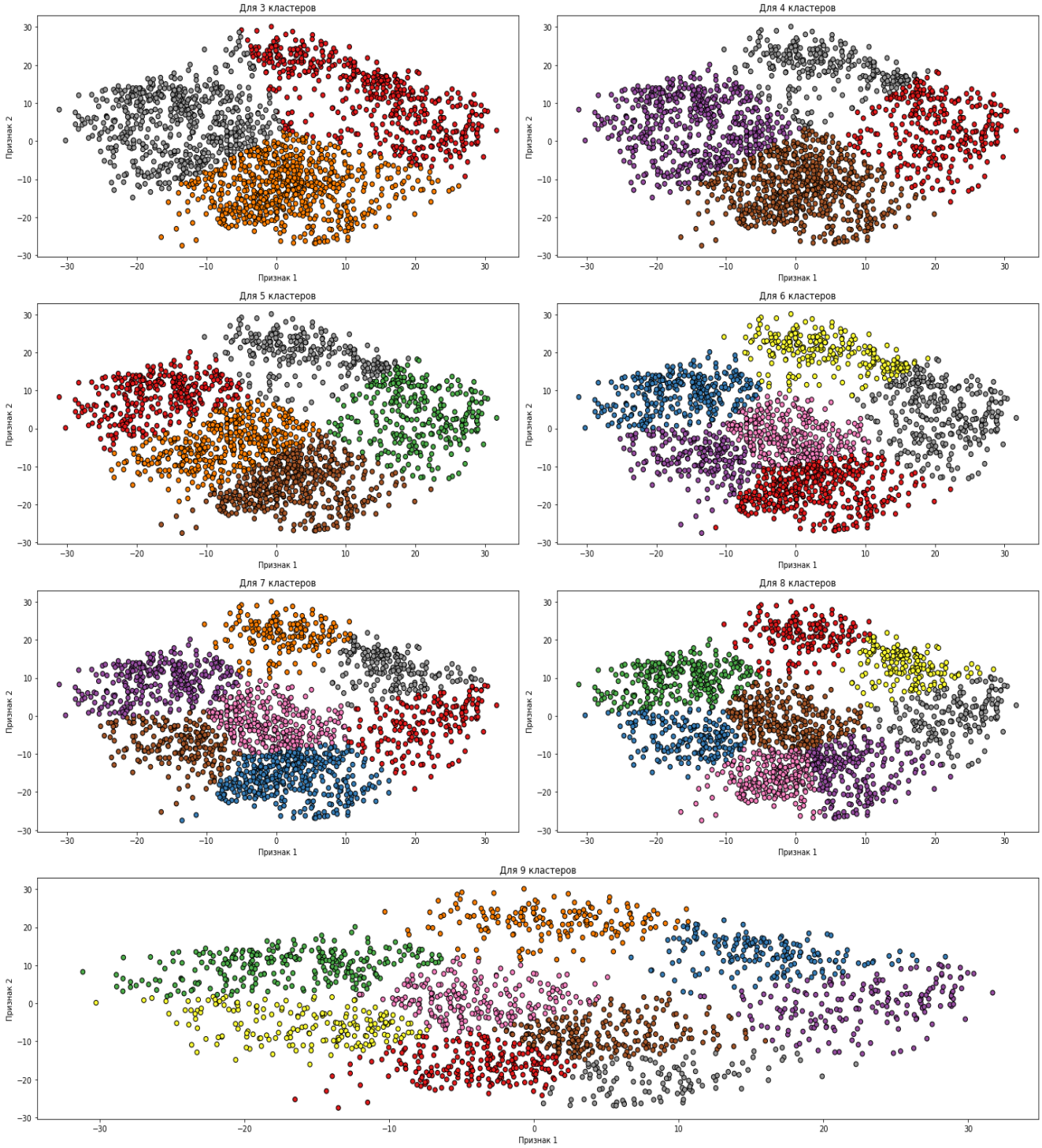

Рис. 2. Результат кластеризации первого набора с помощью алгоритма k-Means

Из рисунка видно, что результат полученный при использовании 8 кластеров наиболее похож на кластеры, полученные при PCA.

Затем был внесен шум в набор данных (случайным образом изменено 15% объектов набора на 10%) и была выполнена кластеризация набора данных с помощью алгоритма k-Means. Результат представлен на рисунке 3.

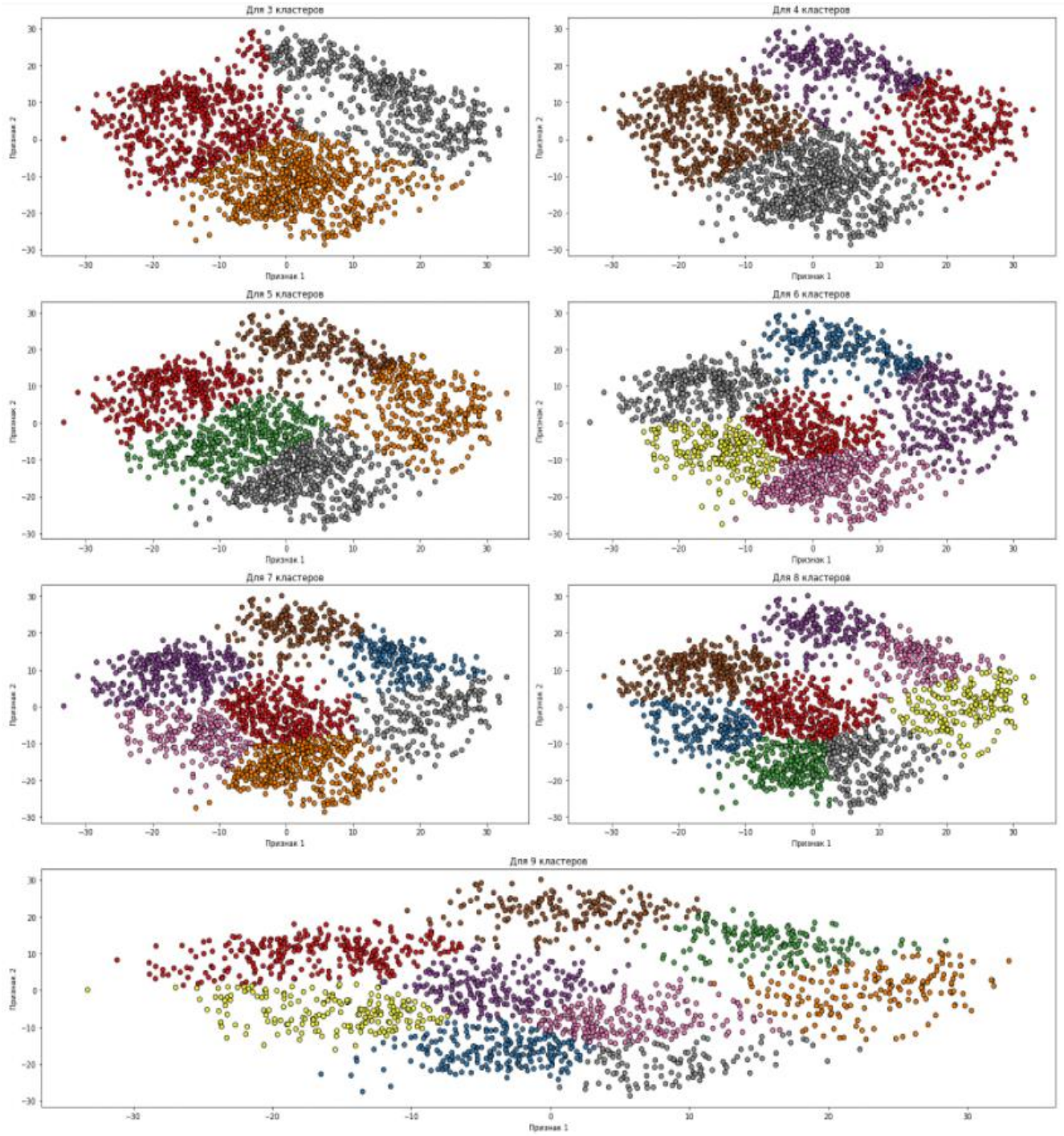

Рис. 3. Результат кластеризации зашумленного первого набора с помощью алгоритма k-Means

Из рисунка видно, что результат не очень сильно отличается от предыдущего.

Затем была выполнена кластеризация зашумленного набора данных с помощью алгоритма k-Medois. Результаты представлены на рисунке 4.

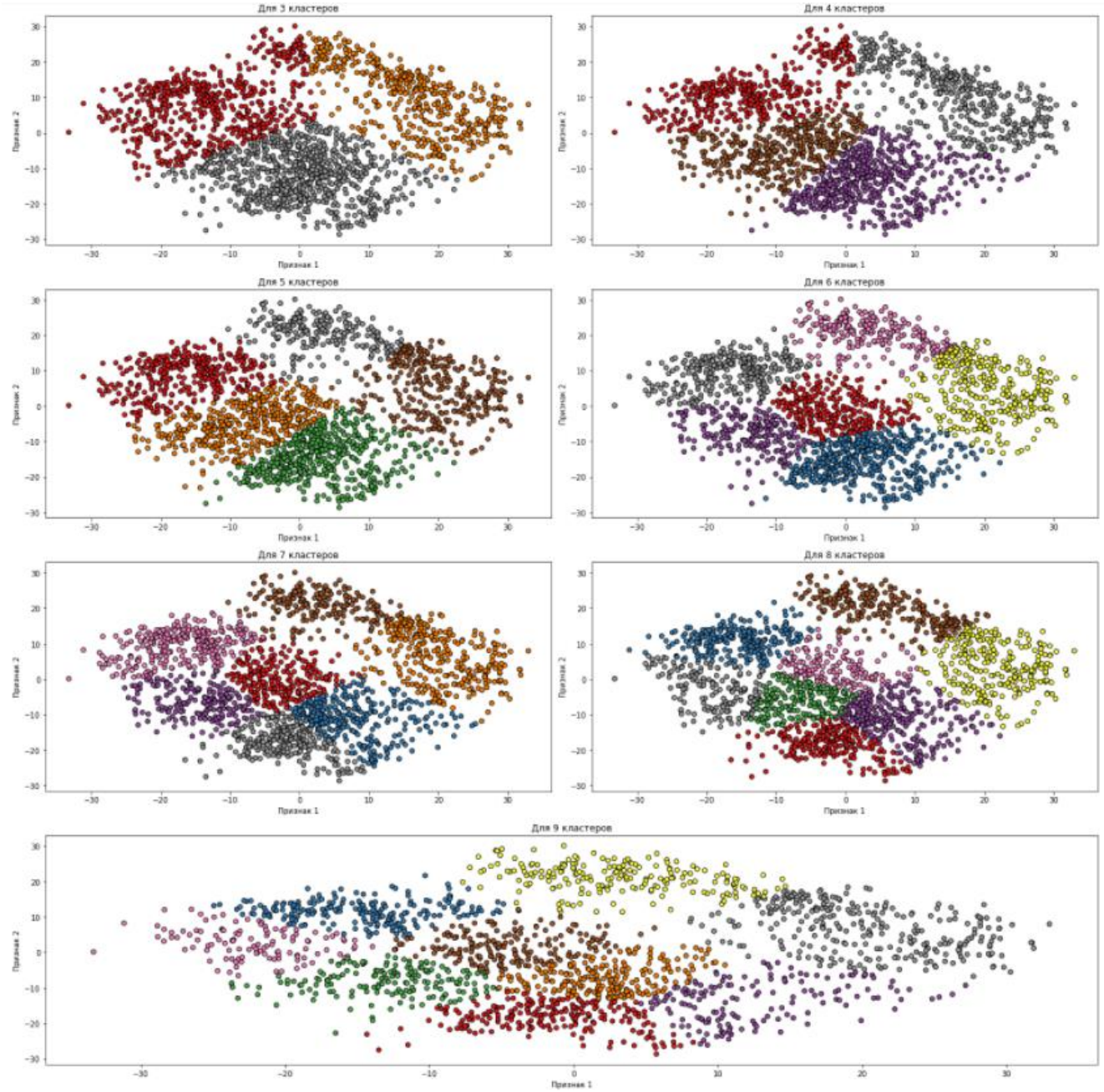

Рис. 4. Результат кластеризации зашумленного первого набора с помощью алгоритма k-Medois

Из рисунка видно, что кластеры отличаются от полученных раннее, наиболее заметно это различие при 7 кластерах.

На рисунке 5 приведен набор данных ирисы Фишера, сниженный до 2 признаков с помощью PCA. Можно заметить, что кластеры имеют вытянутую форму.

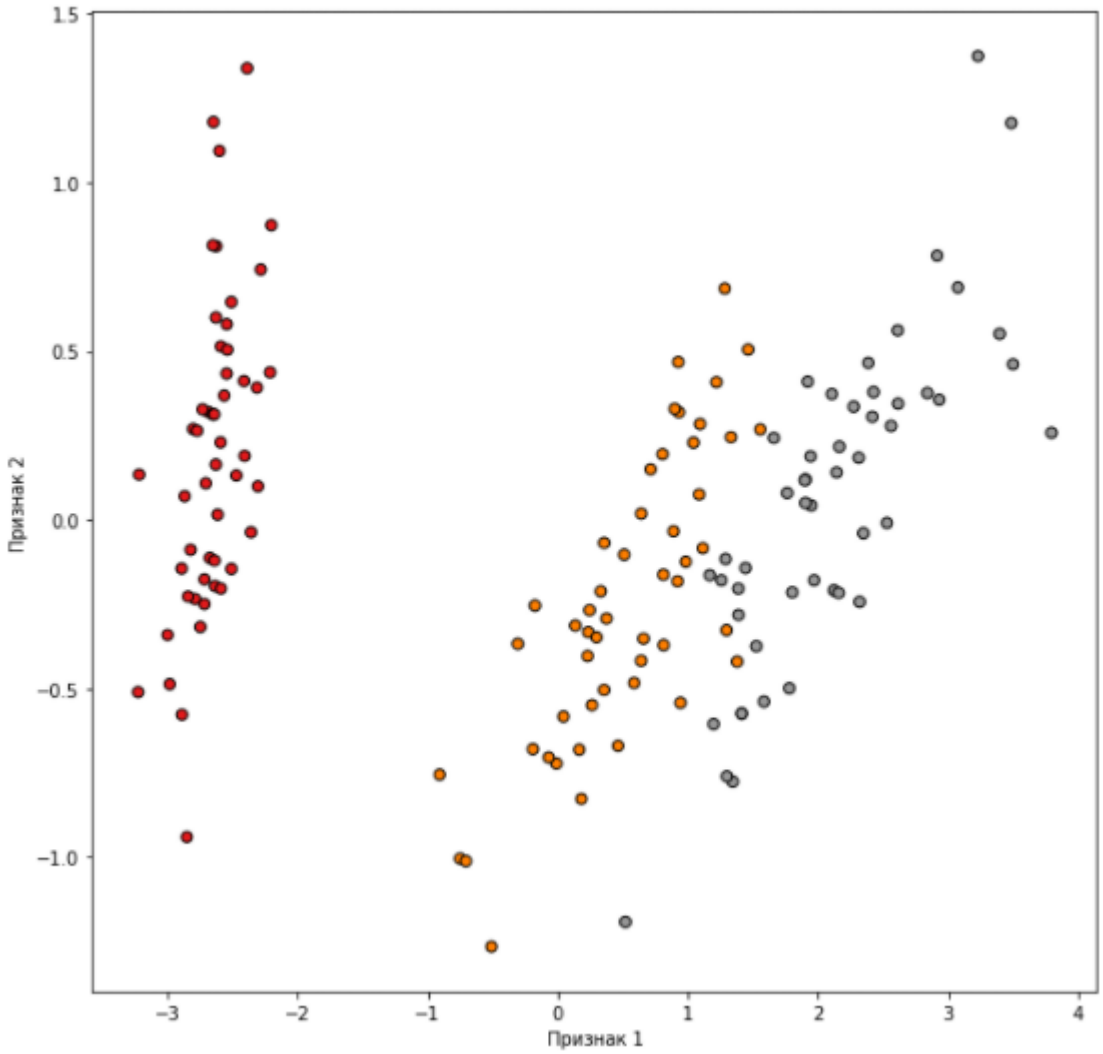

Рис. 5. Набор данных с ирисами Фишера

Затем была выполнена кластеризация набора данных с помощью алгоритма k-Means с использованием различных значений параметров k (из интервала от 3 до 9). Результаты приведены на рисунке 6.

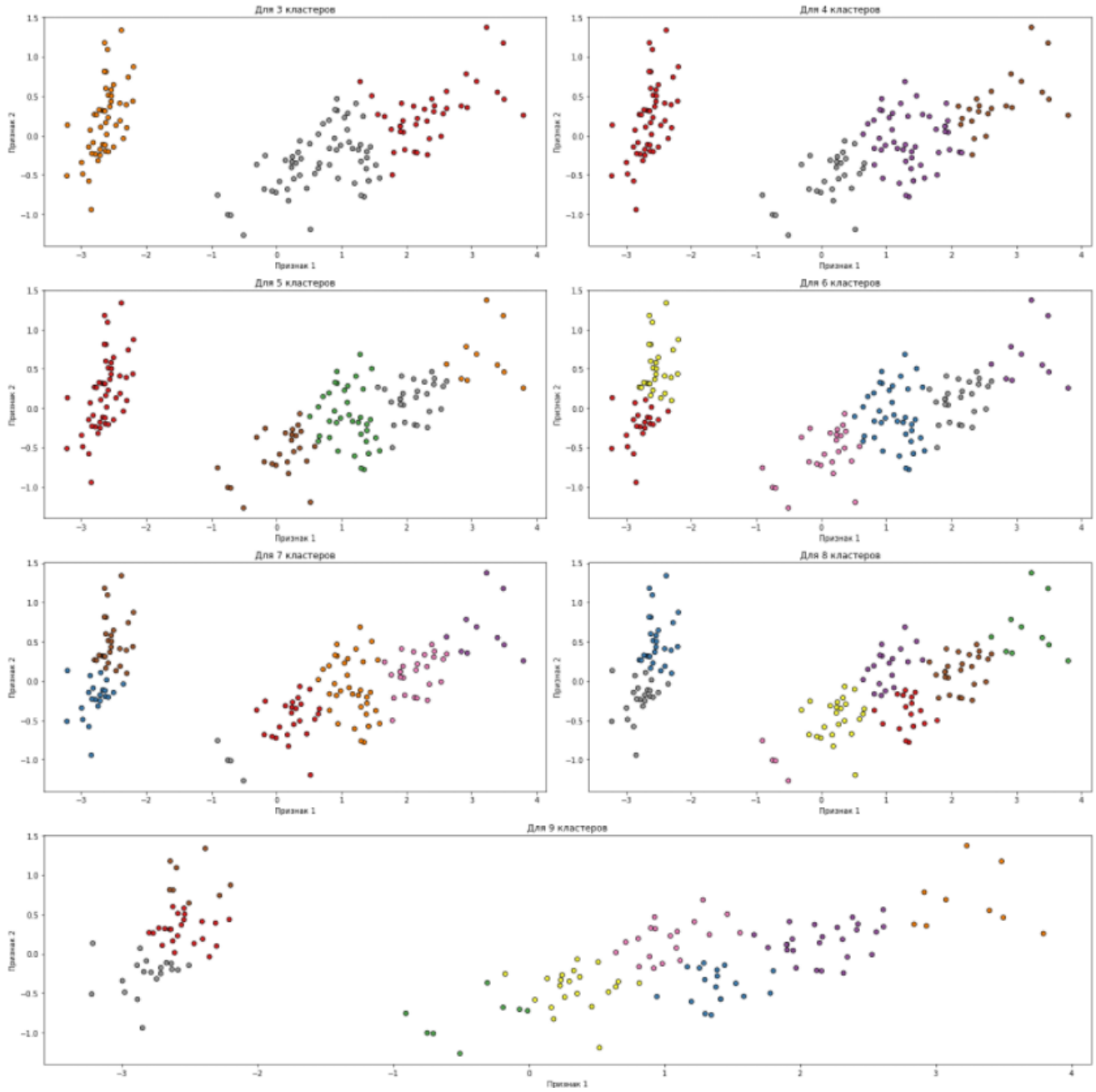

Рис. 6. Результат кластеризации второго набора с помощью алгоритма k-Means

Из рисунка видно, что алгоритм выделяет кластеры не совсем верно с увеличением количества кластеров.

Затем была выполнена кластеризация набора данных с помощью алгоритма k-Medois. Результаты представлены на рисунке 7.

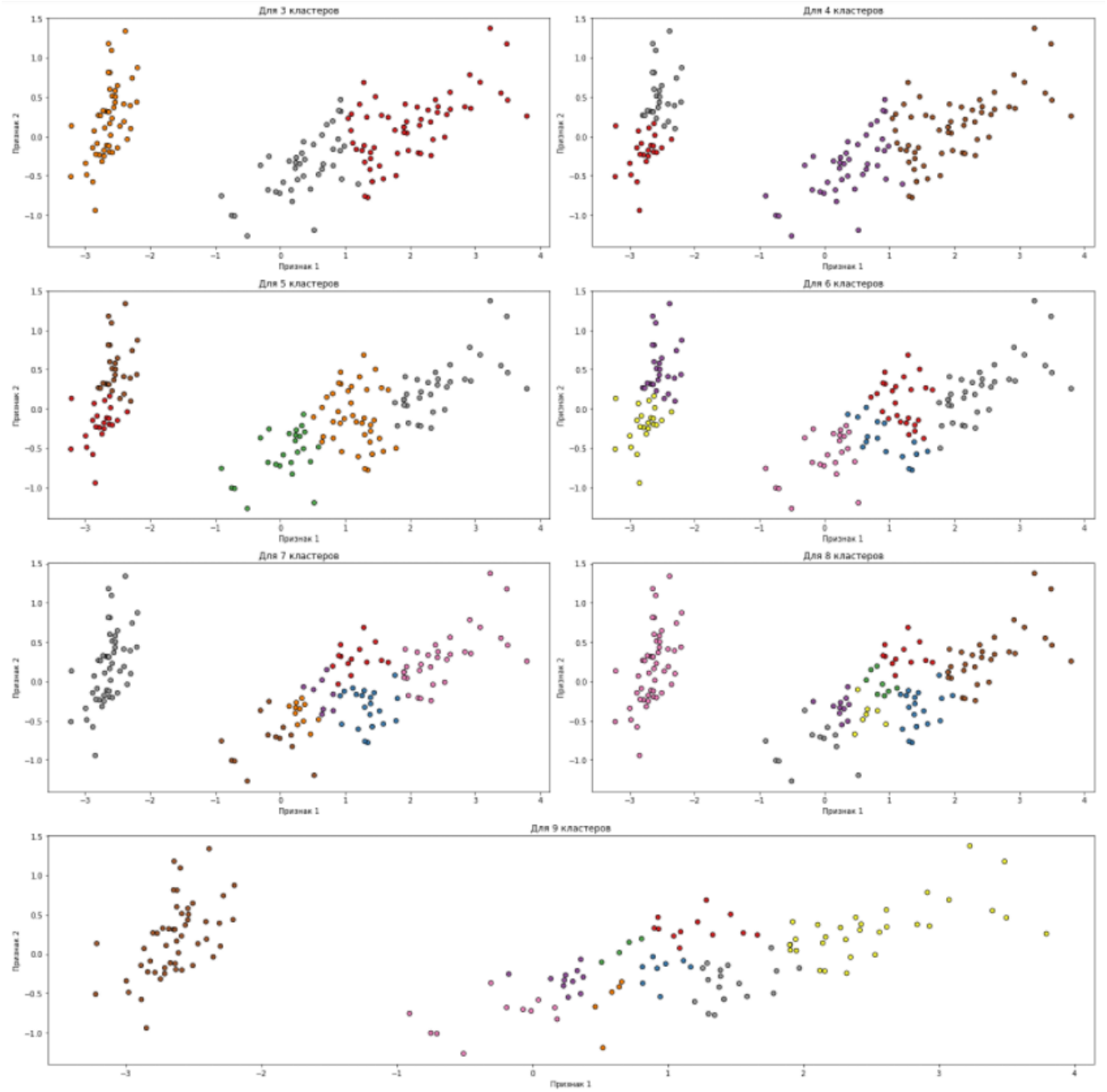

Рис. 7. Результат кластеризации второго набора с помощью алгоритма k-Medois

Из рисунка видно, что результаты отличаются, особенно для большего числа кластеров.