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 №11

на тему «Качество кластеризации»

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

Челябинск 2021

ЗАДАНИЕ

Для набора данных из задания 8 подберите оптимальное количество кластеров с помощью двух любых приемов из следующего множества: метод локтя, кросс-валидация, силуэтный коэффициент, визуализация матрицы схожести.

Постройте диаграммы, подтверждающие полученные результаты.

СОДЕРЖАНИЕ

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: 11 clustering quality

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).

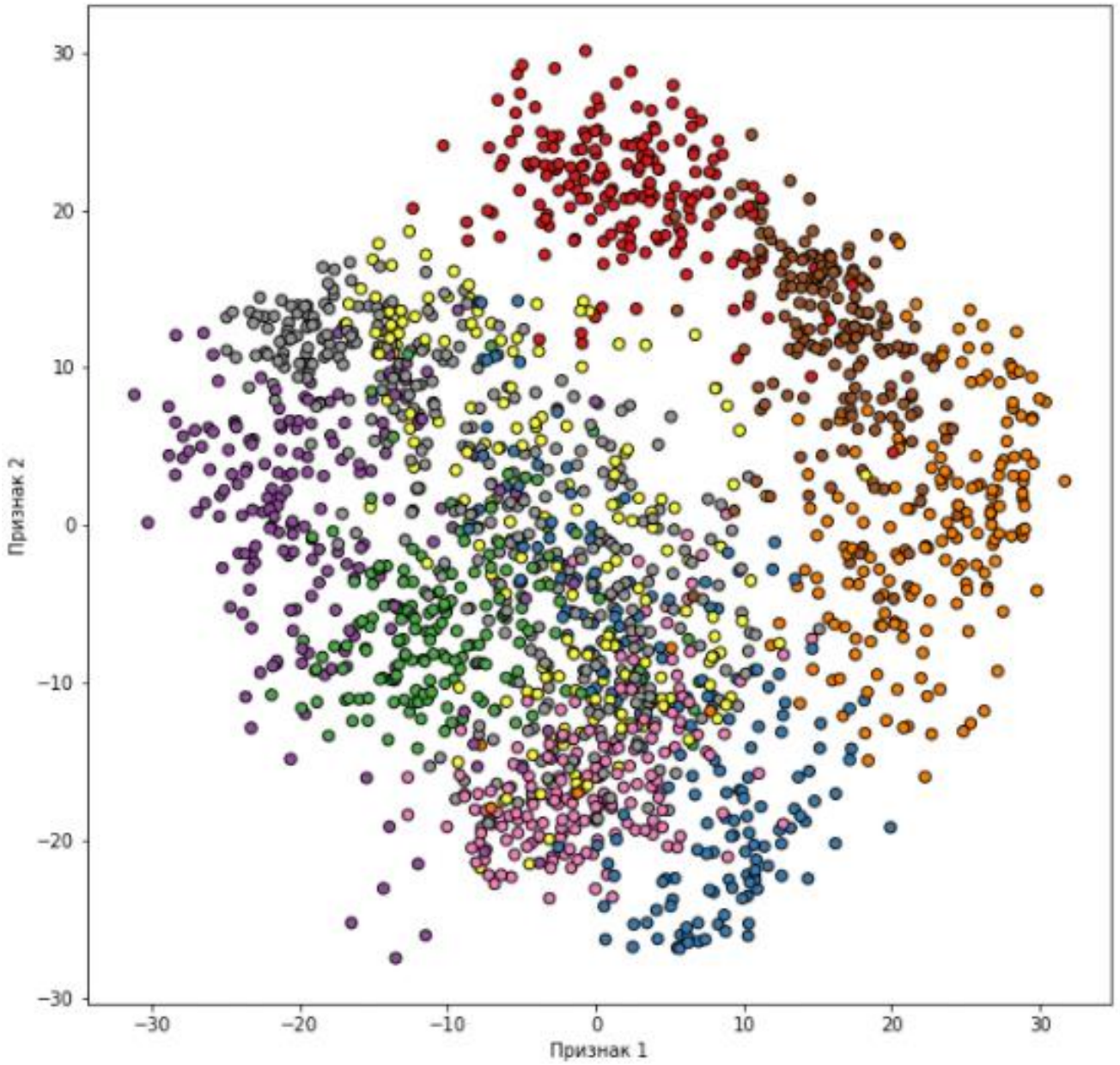

Рис. 1. Набор данных с рукописными цифрами

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.

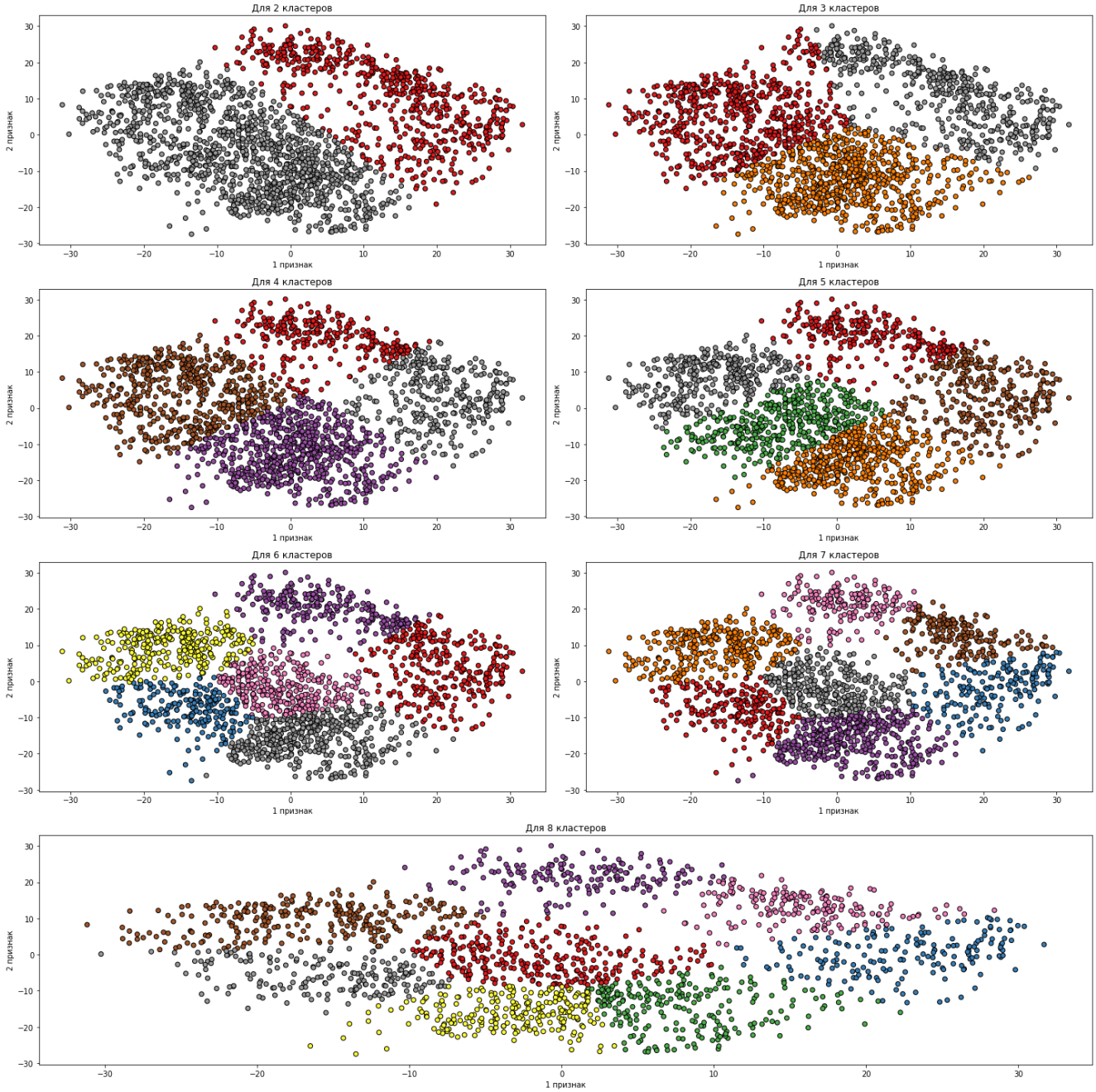

Рис. 2. Результат кластеризации первого набора с помощью алгоритма k-Means

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

Для понимания качества кластеризации используем приемы метод локтя и силуэтный коэффициент.

На рисунке 3 изображена диаграмма для метода локтя.

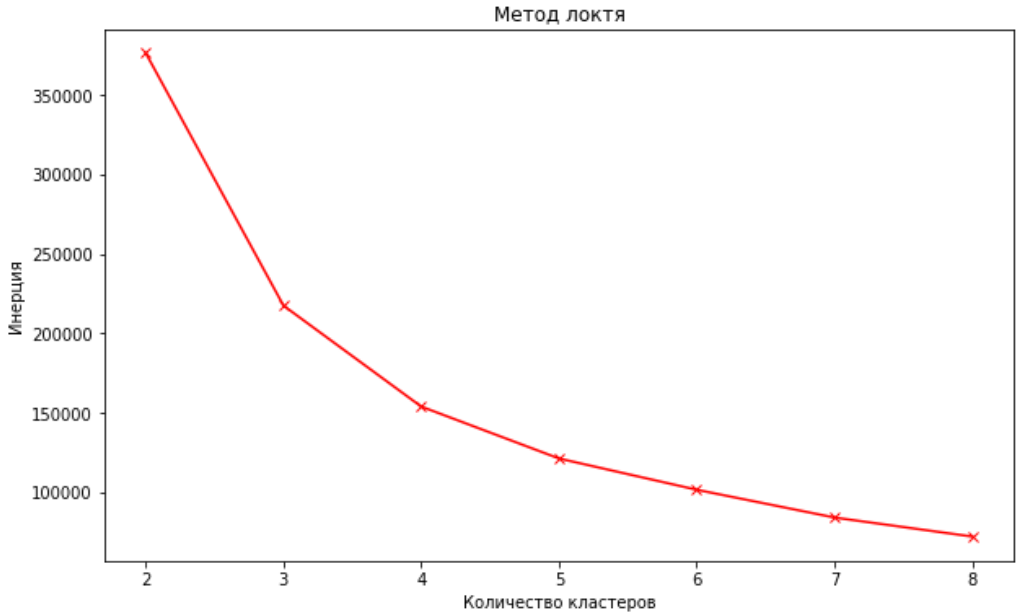

Рис. 3. Метод локтя

Из рисунка видно, что искажение продолжает уменьшаться после 8 кластера, из чего можно сделать вывод, что количество кластеров больше 8 может быть оптимальным для данного набора данных.

На рисунке 4 изображена диаграмма для силуэтного коэффициента.

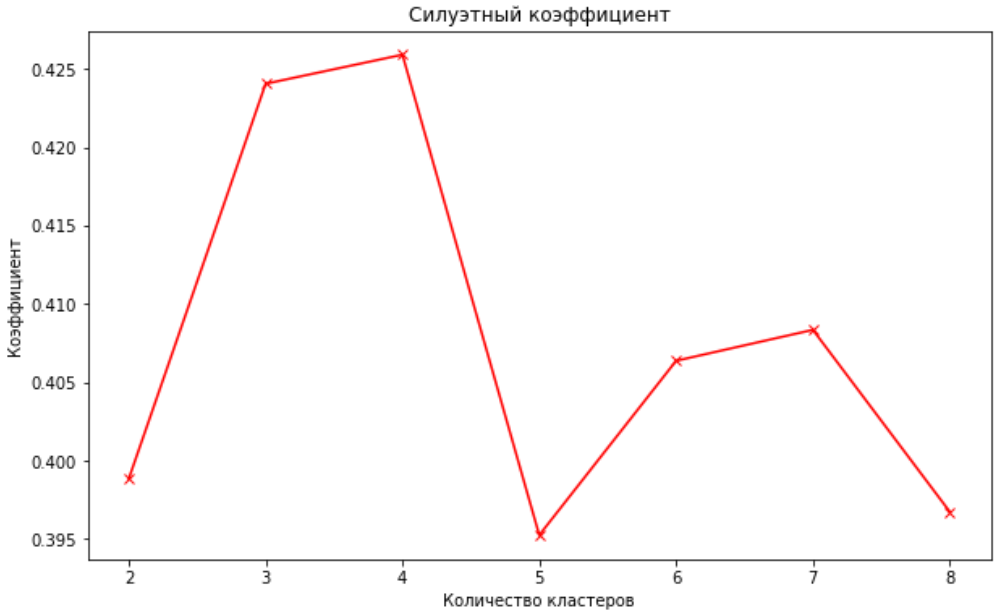

Рис. 4. Метод силуэтного коэффициента

Из рисунка видно, что наибольший силуэтный коэффициент получается при 4 кластерах.