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 №11

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Качество кластеризации

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

Челябинск 2021

1. 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.

2. Краткие сведения о наборах данных

Для получения простого набора данных нужной формы была использована функция sklearn.datasets.make_blobs. С помощью этой функции был сгенерирован набор данных из 2000 точек в двухмерном пространстве. Функция выбирает указанное количество центров (в данном случае 6) и генерирует выпуклые кластеры. Центры могут наложиться друг на друга, поэтому в полученном наборе может оказаться меньше 6 кластеров.

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: 11. ClusterQuality

4. Визуализация

На рис. 1 приведён сгенерированный набор данных.

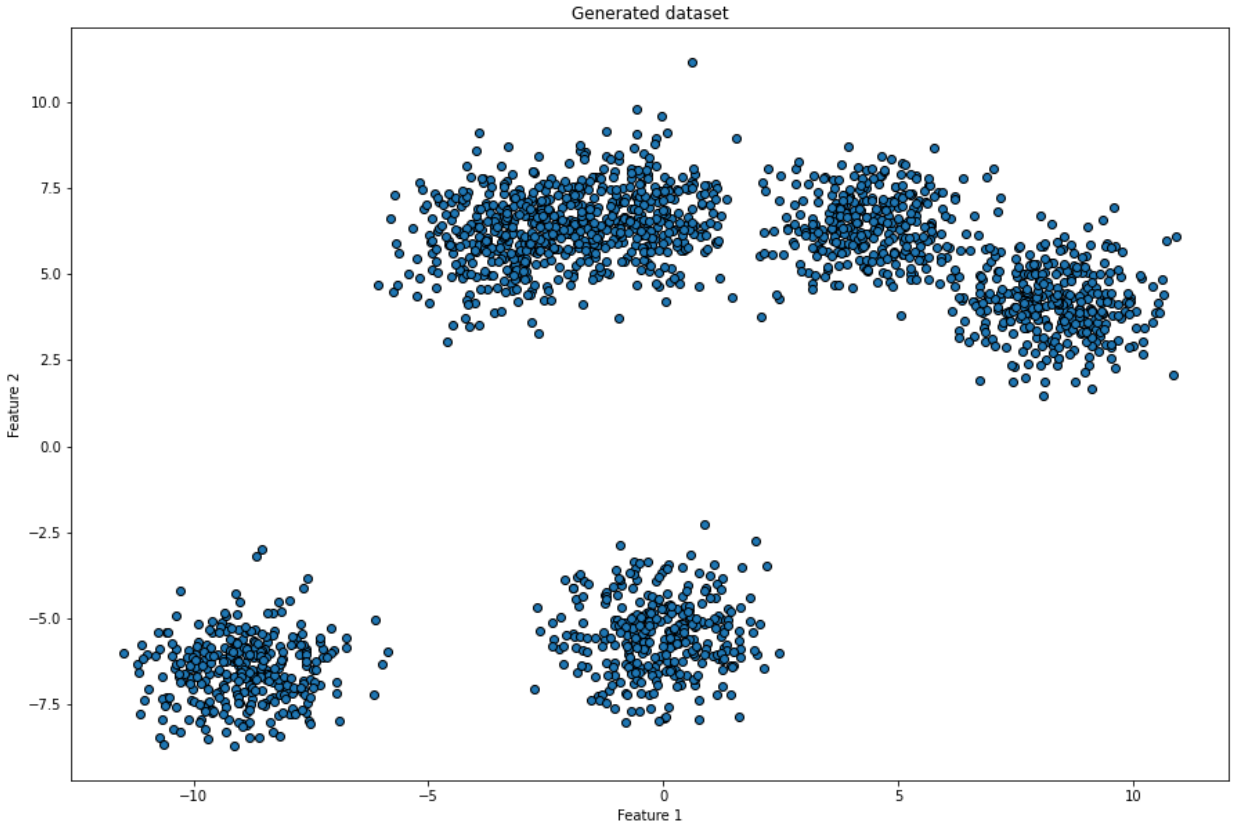

Рис. 1. Визуализация набора данных

Из рисунка можно увидеть, что некоторые из 6 сгенерированных кластеров.

Далее была выполнена кластеризация данного набора с помощью алгоритма k-Means с количеством кластеров от 3 до 9. Результаты кластеризации приведены на рис. 2.

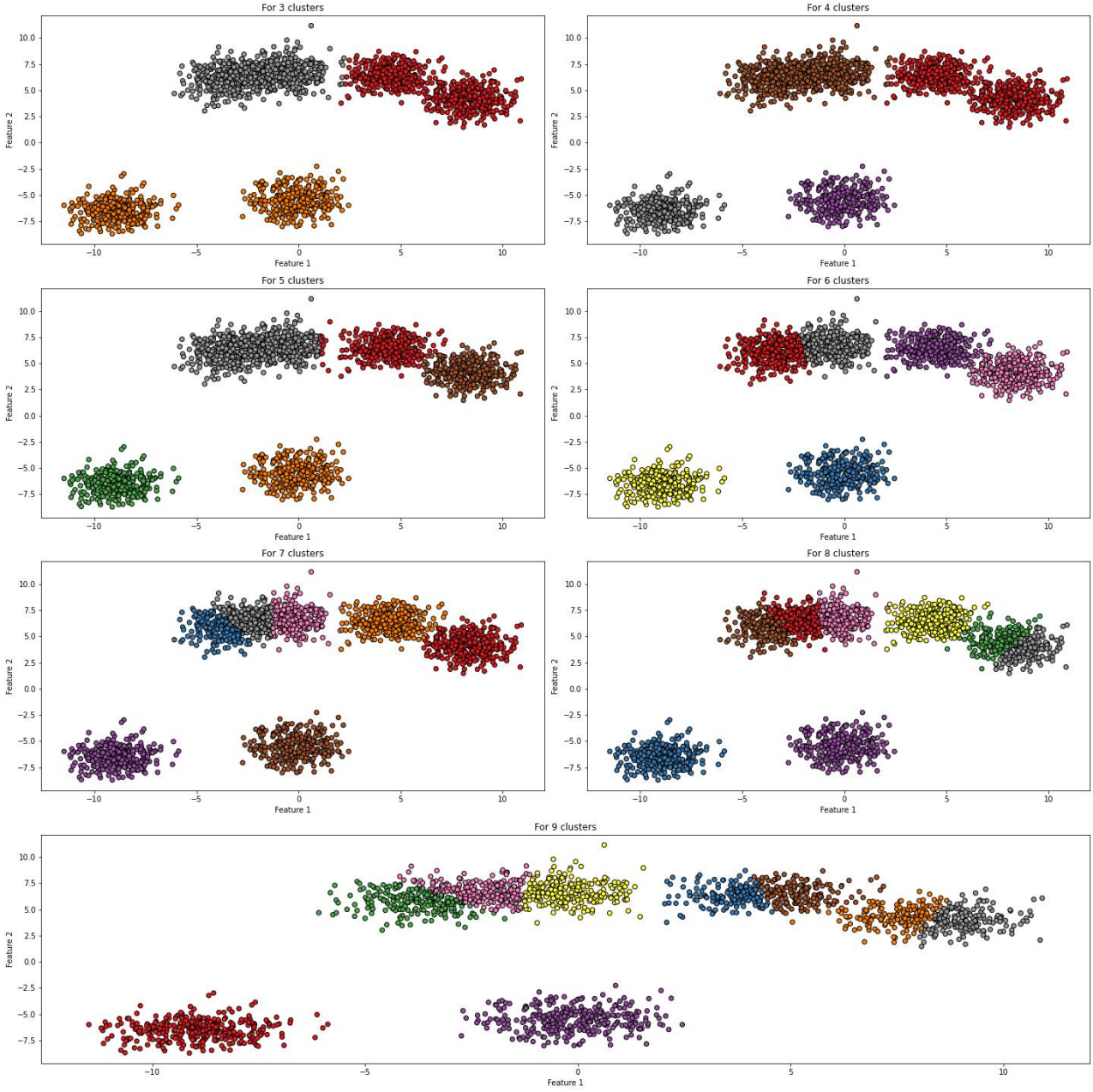

Рис. 2. Результаты кластеризации набора данных с помощью k-Means

Из рисунка можно увидеть, что наиболее оптимальные кластеризации были получены с количествами кластеров, равными 5 и 6.

Далее был построен график искажений, приведённый на рис. 3.

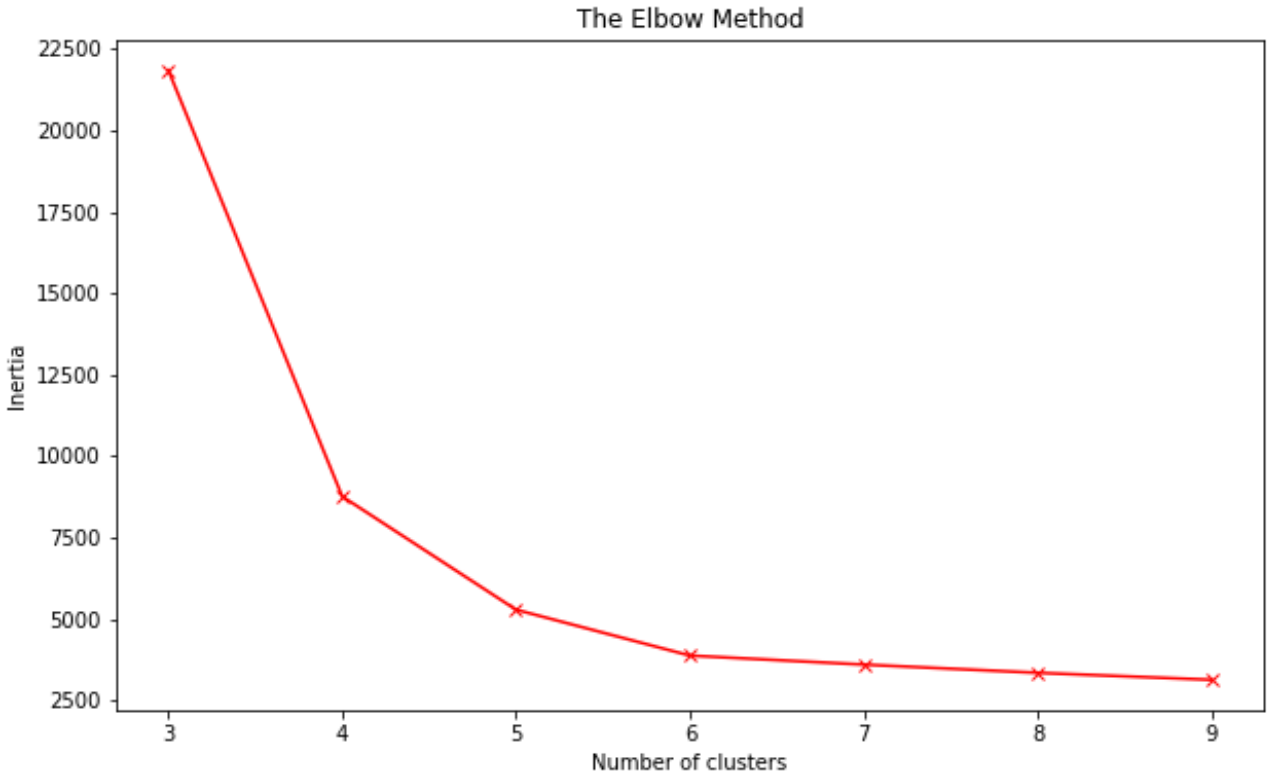

Рис. 3. График искажений для кластеризации

На рисунке нет чёткого локтя, но видно, что темпы уменьшения искажения становятся намного меньше после 6 кластеров, из чего можно заключить, что 6 кластеров являются оптимальным количеством для данноо набора данных.

Далее был построен график силуэтных коэффициентов, приведённый на рис. 4.

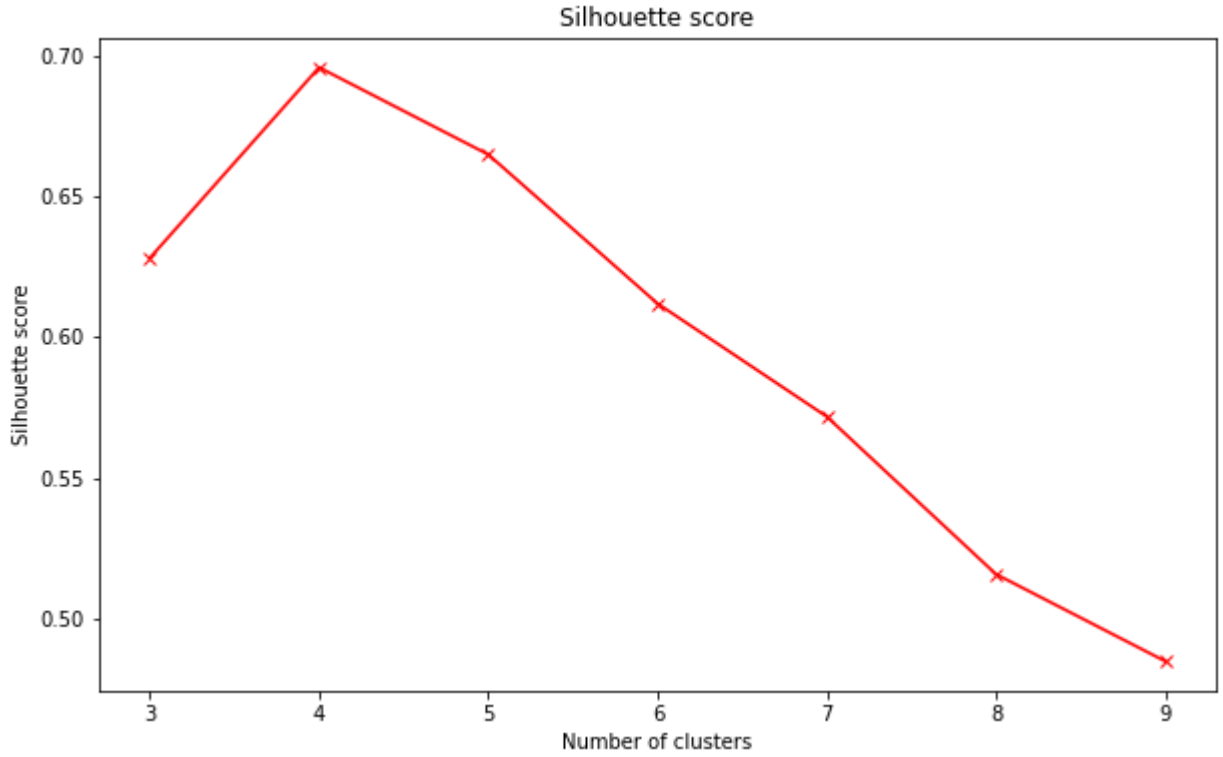

Рис. 4. График силуэтных коэффициентов для кластеризации

Из графика можно увидеть, что наибольший силуэтный коэффициент имеет кластеризация с 4 кластерами. По-видимому, причина этого в том, что в сгенерированно наборе данных есть два случая наложения кластеров друг на друга.

Таким образом, можно сделать вывод, что силуэтный коэффициент лучше подходит при оценке кластеризации кластеров разной формы, в то время как график искажений может, помимо этого, продемонстрировать максимально возможное количество кластеров при их наложении друг на друга.