УДК 004.021

АЛГОРИТМ КЛАСТЕРИЗАЦИИ МЕТОДОМ *K*-СРЕДНИХ

Стремоус М.А.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Ролич О.Ч. – канд.техн.наук, доцент, доцент кафедры ПИКС

Аннотация. Целью данной работы является реализация алгоритма кластеризации методом *k*-средних. В результате разработана программа, в которой реализован данный метод на примере множества точек на плоскости.

Ключевые слова: кластеризация, метод *k*-средних

Введение. Кластеризация – группировка множества объектов на подмножества (далее кластеры) таким образом, чтобы представители одного кластера были схожи друг с другом больше, чем с представителями других кластеров по какому-либо критерию. На данный момент кластеризация играет важную роль, так как она используется для классификации объектов, сжатия данных, обнаружения шумов и ряда других задач.

Основная часть. Метод *k*-средних – это итеративный алгоритм, основанный на минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров.

Суть рассматриваемого алгоритма кластеризации заключается в следующем: данное изначально множество разбивается на заранее известное количество кластеров *k*. При каждой итерации алгоритм стремится минимизировать среднеквадратичное отклонение точек каждого кластера. Данная идея реализовывается благодаря перевычислению на каждой итерации центра масс для каждого кластера, полученного на предыдущем шаге. После перевычисления все элементы множества вновь разбиваются на кластеры, относя себя к тому, центр которого оказался ближе. Данные действия завершаются, если после перевычисления центра масс кластера он совпал с центром масс кластера, полученным на предыдущей итерации [4].

Данный алгоритм реализован на языке программирования *Python*. Случайным образом задаются координаты для 16 точек, а также координаты начальных центров двух кластеров. Для большей наглядности центры кластеров отмечены отличительным знаком (красная звезда) [2]. Начальное распределение точек и центров кластеров представлены на рисунке 1:

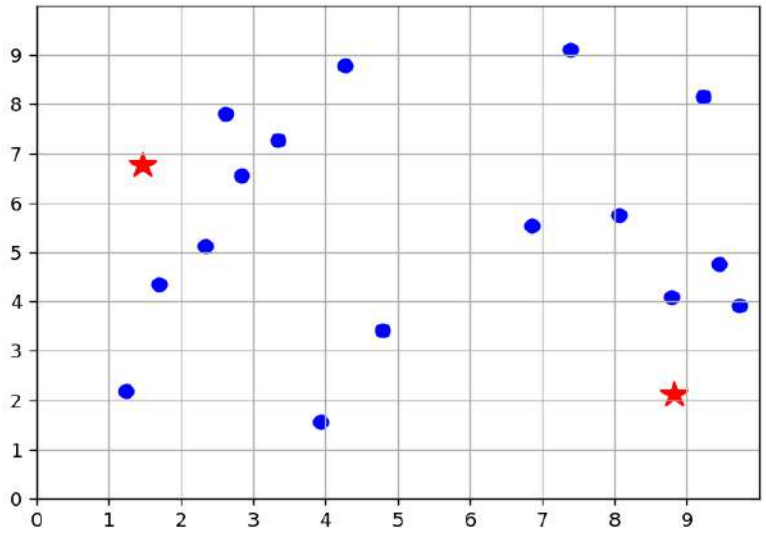

Рисунок 1 – Начальное распределение точек множества и центров кластеров

После проведения первой итерации наблюдается картина, которая представлена на рисунке 2:

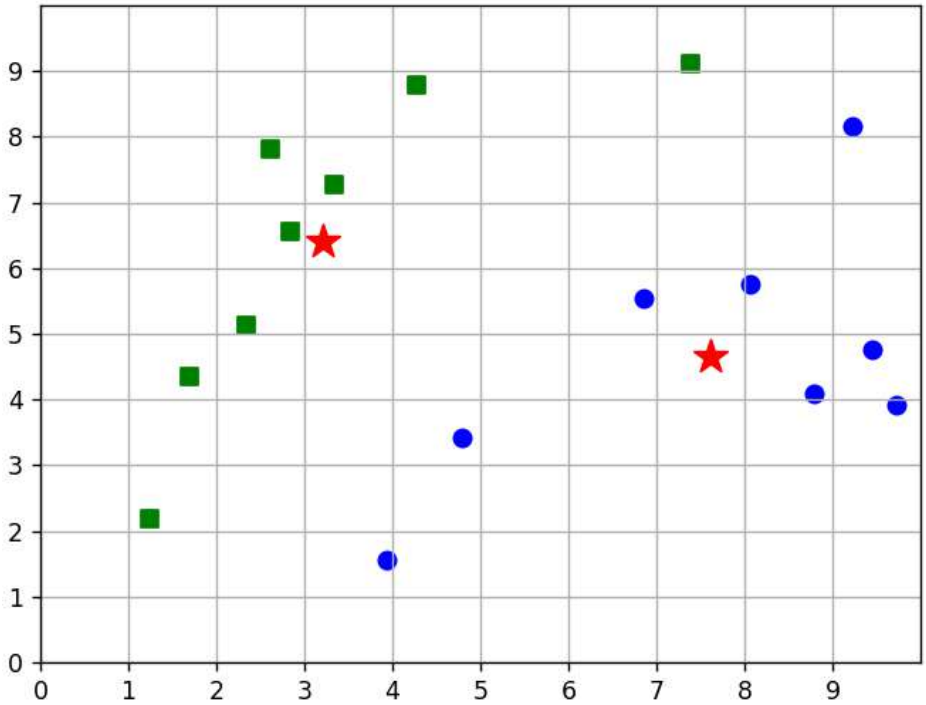

Рисунок 2 – Результат проведения первой итерации

После проведения второй итерации наблюдается картина, которая представлена на рисунке 3:

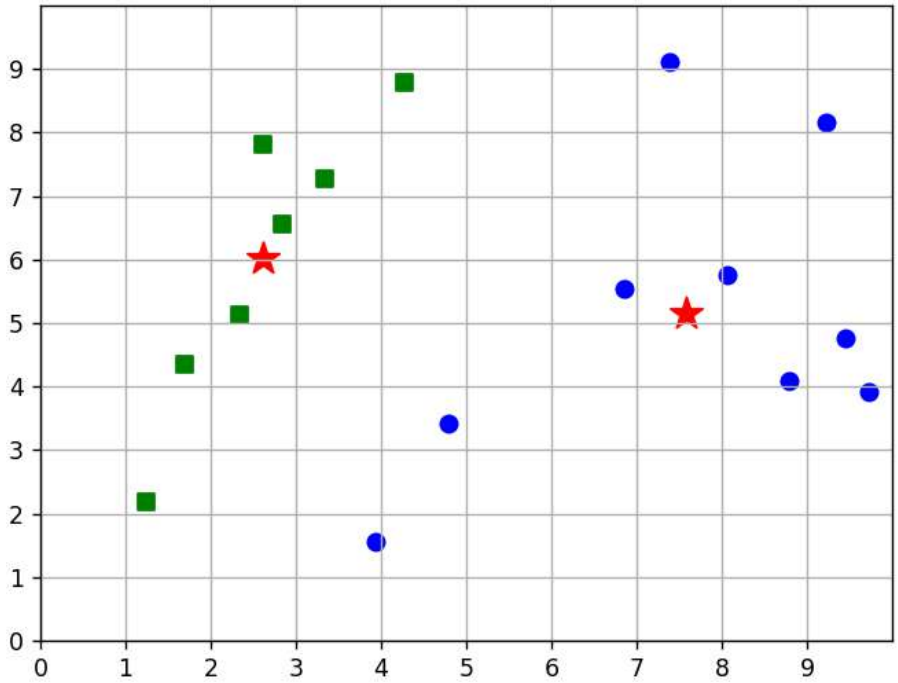

Рисунок 3 – Результат проведения второй итерации

Спустя 4 итерации получен результат, который представлен на рисунке 4:

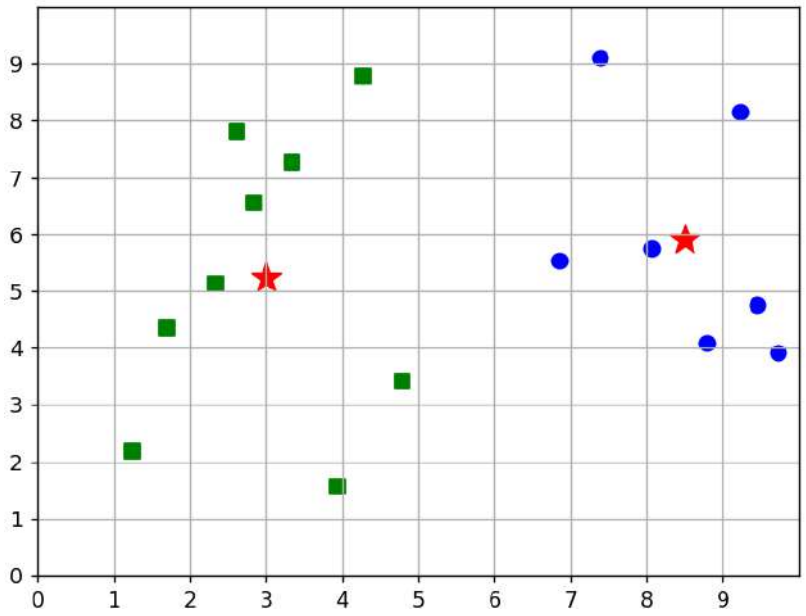

Рисунок 4 – Результат кластеризации методом *k*-средних

Для большей наглядности элементы множества, относящиеся к разным кластерам, разграничены разными цветами.

Стоит отметить, что в данном методе кластеризации в качестве метрики используется Евклидово расстояние. Количество кластеров заранее задаётся исследователем, а качество кластеризации зависит от первоначального разбиения [3].

Заключение. Реализованный алгоритм позволяет кластеризировать множество точек на плоскости методом *k*-средних. В результате формируются хорошо воспринимаемые кластеры с выделенными центрами, которые соответствуют центрам масс каждого кластера.

Список литературы

1. *Документация Python [Электронный ресурс]. – Режим доступа: https://www.python.org/doc/ – Дата доступа: 28.03.2022.*

2. *Python. Scatter - график разброса [Электронный ресурс]. – Режим доступа: https://pyprog.pro/mpl/mpl_scatter.html – Дата доступа: 28.03.2022.*

3. *Кластеризация: метод k-средних [Электронный ресурс]. – Режим доступа: http://statistica.ru/theory/klasterizatsiya-metod-k-srednikh/ – Дата доступа: 28.03.2022.*

4. *Кластеризация [Электронный ресурс]. – Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=Кластеризация – Дата доступа: 28.03.2022.*

UDC 004.021

K-MEANS CLUSTERING ALGORITHM

Stremous M.A.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Rolich O.Ch. – PhD, assistant professor, associate professor of the department of ICSD

Annotation. The purpose of this work is the implemention of the *k*-means clustering algorithm. As a result, there was a program in which this method is implemented using the example of a set of points on a plane.

Keywords: clustering, k-means method