Вестник Сыктывкарского университета.
Серия 1: Математика. Механика. Информатика.
Выпуск 3 (32). 2019

УДК 004.8

КЛАСТЕРИЗАЦИЯ ИЗОБРАЖЕНИЯ МЕТОДОМ K-СРЕДНИХ

Н. О. Котелина, Б. Р. Матвийчук

В работе рассматривается задача кластеризации данных методом k-средних на примере растрового изображения. Решением задачи будет служить программа, реализующая метод k-средних и в качестве результата работы выдающая изображения, разбитые на k кластеров. Оценивается качество кластеризации.

Ключевые слова: метод k-средних, кластеризация, кластер.

1. Введение

1.1. Основные определения

Кластеризация — это разбиение элементов некоторого множества на группы по принципу схожести. Эти группы принято называть *кластерами*.

Объект — элементарная группа данных, с которой оперируют алгоритмы кластеризации.

Каждый объект описывается *вектором характеристик* [1]:

$$X = \{x_1, x_2, \ldots, x_m\}.$$

Компоненты x_i являются отдельными характеристиками объекта. Как правило, это количественные признаки (координаты для точки, цветовые компоненты для цвета, пол, возраст для человека), но существуют алгоритмы, которые работают с качественными признаками (цвет, статус, воинское звание и т. д.).

Количество характеристик m определяет размерность пространства характеристик. Множество, состоящее из всех векторов характеристик, будем обозначать :

$$A = \{X_1, X_2, \ldots, X_n\}.$$

Отметим, что для корректной работы алгоритмов кластеризации характеристики следует нормализовать, то есть привести к одному диапазону.

Кластер — подмножество близких друг к другу объектов из A.

Расстояние $\rho(x_i, x_j)$ *между объектами* x_i *и* x_j — результат применения выбранной метрики в пространстве характеристик [1].

Для ускорения процесса кластеризации можно уменьшить размерность пространства характеристических векторов (например, методом главных компонент), то есть выделить наиболее важные свойства объектов. Уменьшение размерности в ряде случаев позволяет визуально оценивать результаты кластеризации.

1.2. Виды метрик

Метрика выбирается в зависимости:

1) от пространства, в котором расположены объекты;

2) неявных характеристик кластеров.

Евклидово расстояние

$$\rho(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \quad (1)$$

Одна из наиболее известных и часто применяемых метрик. Представляет собой расстояние между точками в n-мерном пространстве.

Квадрат евклидова расстояния

$$\rho(x_i, y_i) = \sum_{i=1}^{n} (x_i - y_i)^2. \quad (2)$$

Для придания больших весов более отдаленным друг от друга объектам можно воспользоваться квадратом евклидова расстояния [2].

Манхеттенское расстояние или «расстояние городских кварталов»

$$\rho(x_i, y_i) = \sum_{i=1}^{n} |x_i - y_i|^2. \quad (3)$$

В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния Евклида. Однако для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат [2].

Расстояние Чебышева

$$\rho(x_i, y_i) = \max |x_i - y_i|. \quad (4)$$

Это расстояние стоит использовать, когда необходимо определить два объекта как «различные», если они отличаются хотя бы по одному измерению [2].

Степенное расстояние

$$\rho(x_i, y_i) = \sqrt[r]{\sum_{i=1}^{n} (x_i - y_i)^p}. \tag{5}$$

Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Здесь r и p — параметры, определяемые пользователем. Параметр p ответствен за постепенное взвешивание разностей по отдельным координатам, параметр r ответствен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра r и p равны двум, то это расстояние совпадает с расстоянием Евклида [2], [3].

2. Метод k-средних

Метод k-средних — один из популярных итеративных методов кластеризации данных. Он быстр и эффективен в применении.

Рассмотрим алгоритм на примере растрового изображения. В качестве объектов будут выступать пиксели изображения, а в качестве характеристик — их цвет.

Описание алгоритма:

1. Выбирается число k — количество кластеров.

2. Далее случайным образом из заданного изображения выбирается k точек. На первом шаге эти точки будут считаться «центрами» кластеров. Каждому кластеру соответствует один центр.

3. Все точки изображения распределяются по кластерам. Вычисляется расстояние от точки до каждого центра кластера, и точку

относят к тому кластеру, расстояние до центра которого будет наименьшим.

4. Когда все точки изображения распределены по кластерам, происходит пересчет центров кластеров. В качестве нового центра кластера берется среднее арифметическое всех точек, принадлежащих кластеру [4].

Пункты 3 и 4 повторяются до тех пор, пока не будет выполнено условие в соответствии с некоторым критерием остановки:

- кластерные центры стабилизировались, то есть все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

В методе k-средних ставится цель минимизировать *полную внутриклассовую дисперсию:*

$$V = \sum_{i=0}^{k} \sum_{X_j \in C_i} (X_j - \mu_i)^2, \tag{6}$$

где X_j — векторы характеристик, k — количество кластеров, C_i — кластеры, μ_i — центры кластеров. Описанный выше алгоритм не гарантирует нахождения наилучшего решения. Чтобы уменьшить зависимость от неудачного выбора центров, алгоритм часто прогоняют несколько раз с различными начальными центрами, а затем выбирают решение с наименьшей дисперсией V [5].

3. Постановка задачи

Дано произвольное растровое изображение формата *bmp*. Задается количество кластеров, на которые будет разбито данное изображение. Требуется произвести кластеризацию заданного изображения методом k-средних.

Поскольку поставлена задача кластеризации изображения, то объектами кластеризации будут пиксели изображения, а вектором характеристик будет выступать цвет пикселя в трехмерном пространстве RGB. В качестве метрики возьмем евклидово расстояние.

Решением задачи служит программа, реализованная на языке программирования C++ в среде программирования CodeGear RAD Studio 2007. Визуализация кластеров в трехмерном пространстве RGB реализована при помощи языка программирования Python.

4. Результаты

В качестве исходного изображения возьмем стандартное изображение «Лена». «Лена» — название стандартного тестового изображения, широко используемого в научных работах для проверки и иллюстрации алгоритмов обработки изображений (сжатия, шумоподавления и т. д.).

1. Дано стандартное изображение «Лена» в градациях серого, формата bmp, размером 512x512 пикселей (рис. 1). Количество кластеров $k = \mathbf{5}$. Поскольку изображение в градациях серого, то все компоненты R, G, B равны и в качестве метрики можно выбрать $|V1 - V2|$, где V1 и V2 — яркости пикселей.

   Как показано на рис. 2, в качестве результата работы программа

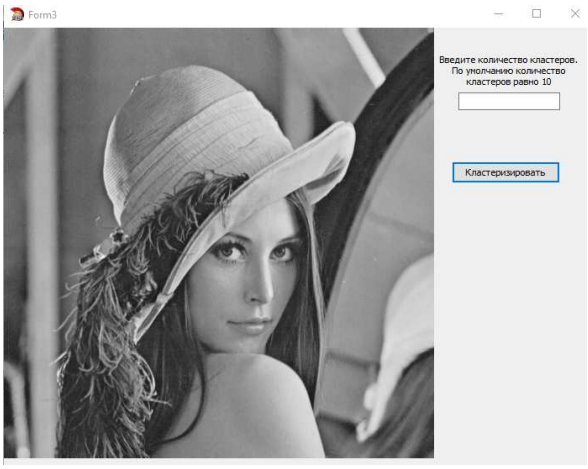

Рис. 1. Демонстрация входных данных программы

выдает:

- кластеризованное изображение;
- первые k центров в формате (RGB), которые были выбраны случайным образом;
- новые центры в формате (RGB), которые были получены в результате кластеризации.

2. Дано стандартное цветное изображение «Лена» формата bmp, размером 220x220 пикселей (рис. 3). Количество кластеров $k = \mathbf{10}$.

   Как показано на рис. 4, в качестве результата работы программа выдает:

   - кластеризованное изображение;
   - первые k центров в формате (RGB), которые были выбраны случайным образом;

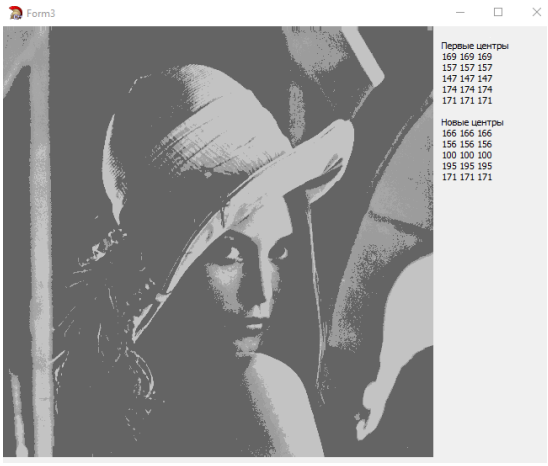

Рис. 2. Демонстрация выходных данных программы

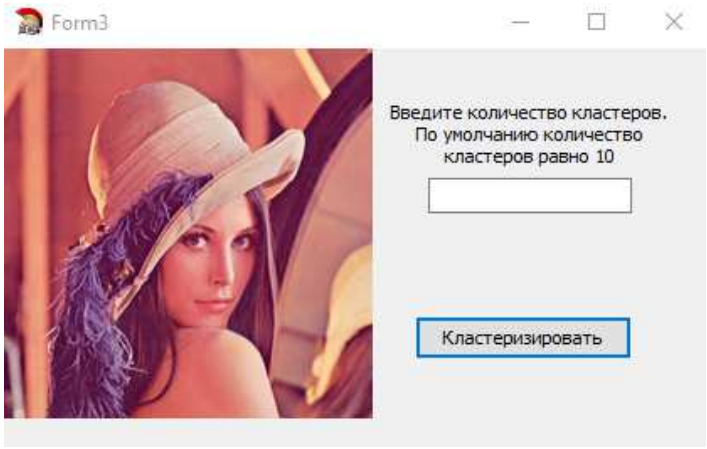

Рис. 3. Демонстрация входных данных программы

- новые центры в формате (RGB), которые были получены в результате кластеризации.

На рис. 5 приведен график, демонстрирующий представление объектов исходного изображения в пространстве RGB до кластеризации.

На рис. 6 показан график, демонстрирующий представление объ-

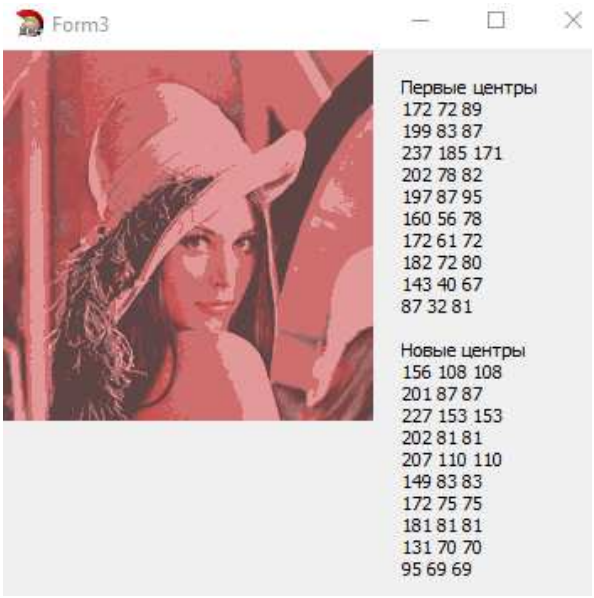

Рис. 4. Демонстрация выходных данных программы

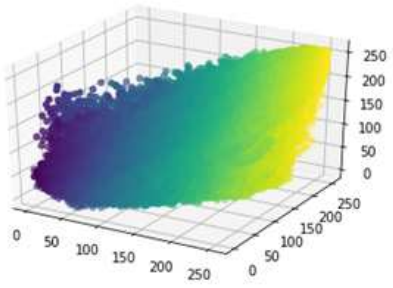

Рис. 5. График объектов в пространстве RGB до кластеризации

ектов исходного изображения в пространстве RGB после кластеризации.

5. Анализ результатов

Метод кластеризации данных k-средних является эффективным и быстрым методом. Однако у него есть свои недостатки. Одним из недостатков метода k-средних является необходимость задавать количество

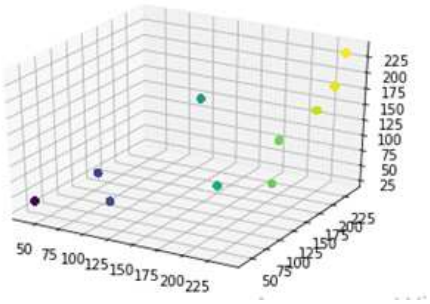

Рис. 6. График объектов в пространстве RGB после кластеризации

кластеров.

Однако главным недостатком данного метода является выбор центров кластеров случайным образом. Поскольку центры кластеров выбираются случайным образом, то результаты работы программы, запущенной несколько раз для одного и того же изображения, будут отличаться.

Так, например, для изображения с рис. 3 программа была запущена несколько раз, и в результате были получены соответствующие внутриклассовые дисперсии: 1504941; 1418464; 1561644; 1386604. Таким образом, наилучшее решение было получено при четвертом запуске программы.

Результаты алгоритма изменятся, если вместо выбора центров кластеров случайным образом предпочтительнее находить наиболее удаленные друг от друга центры.

Список литературы

1. **Котов А., Красильников Н.** Кластеризация данных. М., 2006.

16 с.

2. **Чубукова И. А.** Data Mining. М.: Бином, 2008. 326 с.

3. Обзор алгоритмов кластеризации данных. URL: https://habr.com/ru/post/101338/ (дата обращения: 02.12.2019).

4. **Тюрин А. Г., Зуев И. О.** Кластерный анализ, методы и алгоритмы кластеризации // *Вестник МГТУ МИРЭА №12. М.: Изд-во МГТУ, 2014. 12 с.*

5. **Ян Эрик Солем.** Программирование компьютерного зрения на языке Python / пер. с англ. А. А. Слинкин М.: ДМК Пресс, 2016. 312 с.

Summary

Kotelina N. O., Matwiichuck B. R. Image clustering by k-means

The paper deals with the problem of data clustering by the k-means method on the example of a raster image. The solution of the problem will be a program that implements the k-means method and as a result of the work, produces images divided into k clusters. The quality of clustering is estimated.

Keywords: k-means method, clustering, cluster.

References

1. **Kotov A., Krasilnikov N.** *Klasterizatsiya dannykh* (Data clustering), M., 2006, 16 p.

2. **Chubukova I. A.** *Data Mining*, M.: Binom, 2008, 326 p.

3. *Obzor algoritmov klasterizatsii dannykh* (Overview of data clustering algorithms), URL: tt https://habr.com/en/post/101338/ (date of the application: 12.02.2019).

4. **Tyurin A. G., Zuev I. O.** Klasternyy analiz, metody i algoritmy klasterizatsii (Cluster analysis, methods and algorithms of clustering), *Vestnik MGTU MIREA*, No 12, M.: Publishing house of MSTU, 2014, 12 p.

5. **Ian Eric Solem** *Programmirovaniye komp'yuternogo zreniya na yazyke Python* (Programming computer vision in Python), M.: DMK Press, 2016, 312 p.

Для цитирования: Котелина Н. О., Матвийчук Б. Р. Кластеризация изображения методом k-средних // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика. 2019. Вып. 3 (32). С. 101–112.*

For citation: Kotelina N. O., Matwiichuck B. R. Image clustering by k-means, *Bulletin of Syktyvkar University. Series 1: Mathematics. Mechanics. Informatics*, 2019, 3 (32), pp. 101–112.

СГУ им. Питирима Сорокина *Поступила 10.12.2019*