УДК 519.6

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-СРЕДНИХ И ЕГО РЕАЛИЗАЦИЯ В СРЕДЕ MATLAB

Ооржак О.Е.
Научный руководитель: доцент, канд. физ.-мат.наукБаранова И.В.
Сибирский федеральный университет

Подкластеризациейпонимаетсяразбениезаданнойсовокупностиобъектов (данных) наразличныеподмножества, называемыекластерами, таким образом, чтобы кластеры были непересекающимися и состояли из схожих по свойствам объектов, при этом объекты разных классов отличались.Кластер – группа объектов, схожих между собой по определенным признакам. В качестве признаков рассматриваются некоторые количественные характеристики объектов.

Формализуем математически процесс кластеризации. Пусть дан набор данных $X_n = \{x_1, ..., x_n\} \subset X, x \in N$ и функция, определяющая степень сходства объектов, в большинстве случаев это функция расстояния между объектами $\rho(x_i, x_j)$.Требуется разбить последовательность X_nна непересекающиеся подмножества (кластеры) так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. Алгоритм кластеризации – это функция $A: X \to Y$ которая любому объекту $x \in X$ставит в соответствие метку кластера $y_i \in Y$. Чаще всего множество Y заранее не известно и дополнительной задачей является определение оптимального числа кластеров с точки зрения того или иного показателя качества кластеризации.

Одним из наиболее известных приложений кластеризации является классификация растений или животных в отдельные группы или виды. Упрощение исходных данных осуществляется заменой координаты каждой точки в кластере координатами описательной точки кластера («центра масс» кластера).

Наиболее популярным методом кластеризации является метод k-средних (k-meansmethod). Алгоритм данного метода стремится минимизировать сумму квадратов расстояний всех точек, входящихв кластерную область, до центра кластера. Число кластеров *K* заранее определяется. Алгоритмсостоит из следующих шагов.

Шаг 1. Выбираются *K*исходных центров кластеров $z_1(1)$, $z_2(1)$, ..., $z_K(1)$. Этот выбор производится произвольно, и обычно в качестве исходных центров используются первые *K* результатов выборки из заданного множества образов.

*Шаг 2.*На *k*-ом шаге итерации заданное множество образов {*x*} распределяется по *K*кластерам по следующему правилу: $x \in S_j(k)$, если

$$\|x - z_j(k)\| < \|x - z_i(k)\| \qquad (*)$$

для всех $i = 1, 2, .., K, i \neq j$, где $S_j(k)-$ множество образов, входящих в кластер с центром $z_j(k)$. В случае равенства в (*) решение принимается произвольным образом.

Шаг 3. На основе результатов шага 2 определяются новые центры кластеров $z_j(k+1)$, $j = 1, 2, ..., K,$ исходя из условия, что сумма квадратов расстояний между всеми образами, принадлежащими множеству $S_j(k)$, и новым центром кластера должна быть минимальной. Другими словами, новые центры кластеров $z_j(k+1)$

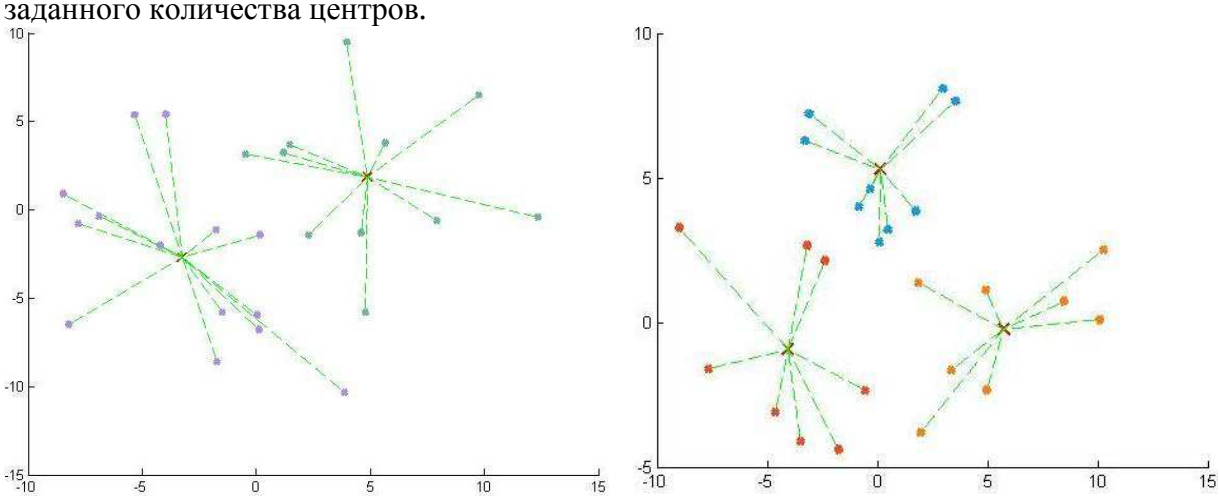
выбираются таким образом, чтобы минимизировать показатель качества $M_j = \sum_{x \in S_j(k)} \left\| x - z_j(k+1) \right\|^2$, $j = 1, 2, ..., K$.

Центр $z_j(k+1)$, обеспечивающий минимизацию показателя качества, является, в сущности, выборочным средним, определенным по множеству $S_j(k)$. Следовательно, новые центры кластеров определяются как $z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x$, $j = 1, 2, ..., K$, где N_j – число выборочных образов, входящих во множество $S_j(k)$. Очевидно, что название алгоритма «*k*-средних» определяется способом, принятым для последовательной коррекции назначения центров кластеров.

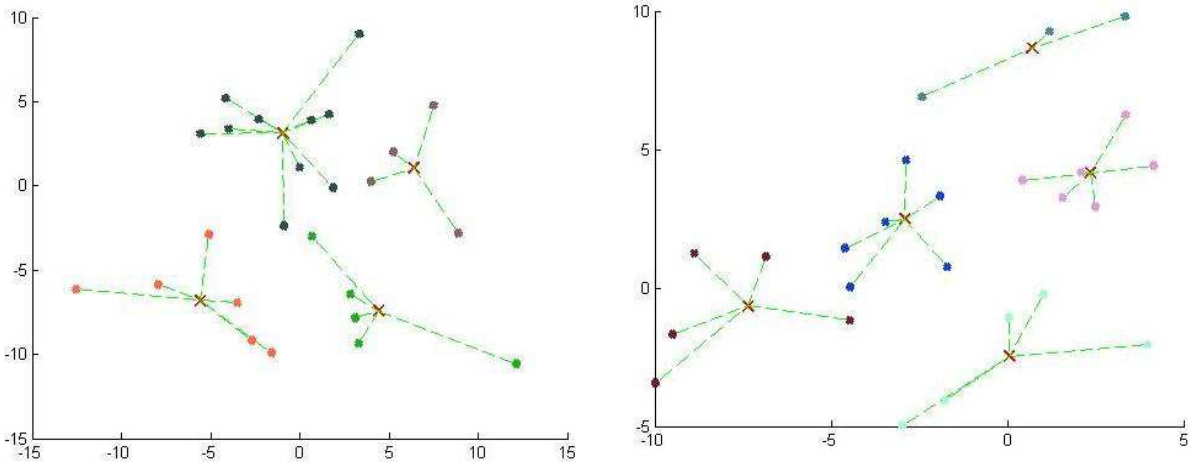
Шаг 4. Равенство $z_j(k+1) = z_j(k)$ при $j = 1, 2, ..., K$ является условием сходимости алгоритма, и при его достижении выполнение алгоритма заканчивается. В противном случае алгоритм повторяется от шага 2.

Иллюстрация результата работы алгоритма в MATLAB. В качестве среды визуализации алгоритма *k*-средних был выбран пакетMATLABверсии R2013a. На встроенномязыке программирования MATLABбыла написана программа циклического вычисления «центров масс» (заранее заданного количества, например, *K*=3) кластеров и определения расстояний(евклидова метрика)от этих центров до точек («образов»), случайно «разбросанных»в прямоугольной системе координат функцией randn(). В результате программыкаждыйкластер обретает свой цвет, центры кластеров обозначены крестиками, расстояния от точек до центров обозначены пунктиром.

На рис. 1 представлены результаты кластеризации для 25 случайных точек и заданного количества центров.



а) *б*)



в) *г*)

Рис. 1. Результаты кластеризации для 25 случайных точек и заданного количества центров:
а) *K*=2; *б*) *K*=3; *в*) *K*=4; *г*) *K*=5.

Недостатком алгоритма является его медленная работа на больших объемах данных.Также применениеметодаK-средних предполагает наличие гипотезы о наиболее вероятном количестве кластеров. Выбор числа K может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции. Хотя для этого алгоритма общее доказательство сходимости неизвестно [3], следует отметить, что при произвольном задании большого числа K, а такженеудачного выбора исходного положения начальных центров (например, расположить их очень близко кдруг другу в евклидовой метрике) алгоритм может не сойтись вовсе.Стандартная рекомендация – создать два кластера (K=2), затем 3, 4, 5 и т.д., и, сравнивая полученные результаты, выбрать оптимальный вариант.