АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

Пусть имеется множество произвольных объектов x_i. Кластеризация - это объединение объектов в группы (кластеры) на основе схожести признаков для объектов одной группы и отличий между группами. Большинство алгоритмов кластеризации не опираются на традиционные для статистических методов допущения; они могут использоваться в условиях почти полного отсутствия информации о законах распределения данных.

Задача кластеризации состоит в разбиении объектов из множества x_i на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров. В метрическом пространстве "схожесть" обычно определяют через расстояние. Расстояние может рассчитываться как между исходными объектами x_i, так и от этих объектов к объектам - прототипам кластеров. Обычно координаты прототипов заранее неизвестны - они находятся одновременно с разбиением данных на кластеры.

Существует множество методов кластеризации, которые можно классифицировать на четкие и нечеткие. Четкие методы кластеризации разбивают исходное множество объектов на несколько непересекающихся подмножеств. При этом любой объект из x_i принадлежит только одному кластеру. Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью. Нечеткая кластеризация во многих ситуациях более "естественна", чем четкая, например, для объектов, расположенных на границе кластеров.

Методы кластеризации также классифицируются по тому, определено ли количество кластеров заранее или нет. В последнем случае количество кластеров определяется в ходе выполнения алгоритма на основе распределения исходных данных.

В данной лекции рассматривается так называемый алгоритм горной кластеризации, который не требует задания количества кластеров. Метод предложен Р. Ягером и Д. Филевым в 1993 г.

Пусть $D(x_i, x_j)$ - функция, определяющая расстояние от объекта x_i до объекта x_j исходного множества объектов.

На первом шаге горной кластеризации определяют точки, которые могут быть центрами кластеров. Если о центрах кластеров заранее ничего не известно, то в качестве потенциальных центров удобно принять сами объекты. Обозначим множество потенциальных центров кластеров Z_h. Если центрами кластеров считаются сами объекты, то тогда $Z_h = x_k$.

На втором шаге для каждой такой точки рассчитывается значение потенциала, показывающего возможность формирования кластера в ее окрестности. Чем плотнее расположены объекты в окрестности потенциального центра кластера, тем выше значение его потенциала. После этого итерационно выбираются центры кластеров среди точек с максимальными потенциалами. Потенциал центров кластеров рассчитывается по следующей формуле:

$$P_1(Z_h) = \sum_{k=1}^{M} \exp(-\alpha D(Z_h, x_k)),$$

где M - количество объектов x_i, exp – функция экспоненты, α - положительная константа, характеризующая масштаб расстояний между объектами. В простейших случаях удобно полагать, что α равно единице, делённой на среднее расстояние между объектами.

В случае, когда объекты кластеризации заданы двумя признаками (точки на плоскости), графическое изображение распределения потенциала будет представлять собой поверхность, напоминающую горный рельеф. Отсюда и название - горный метод кластеризации.

На третьем шаге алгоритма в качестве центров кластеров выбирают координаты "горных" вершин. Для этого центром первого кластера V_1 назначают точку с наибольшим потенциалом. Обычно, наивысшая вершина окружена несколькими достаточно высокими пиками. Поэтому назначение центром следующего кластера точки с максимальным потенциалом среди оставшихся вершин привело бы к выделению большого числа близко расположенных центров кластеров. Чтобы выбрать следующий центр кластера необходимо вначале исключить влияние только что найденного кластера. Для этого значения потенциала для оставшихся возможных центров кластеров пересчитывается следующим образом: от текущих значений потенциала вычитают вклад центра только что найденного кластера. Перерасчет потенциала происходит по формуле:

$$P_2(Z_h) = P_1(Z_h) - P_1(V_1) \cdot \exp(-\beta \cdot D(Z_h, V_1)),$$

где β - положительная константа, характеризующая масштаб размера одного кластера (чем значение β меньше, тем больше размер кластера). В простейших случаях её можно принять равной α.

Центр второго кластера V_2 определяется как точка с максимальным значением потенциала $P_2(Z_h)$. Затем снова пересчитывается значение потенциалов:

$$P_3(Z_h) = P_2(Z_h) - P_2(V_2) \cdot \exp(-\beta \cdot D(Z_h, V_2)).$$

Итерационная процедура пересчета потенциалов и выделения центров кластеров продолжается до тех пор, пока максимальное значение потенциала превышает некоторый порог. В простейших случаях этот порог можно считать равным половине значения максимального потенциала $P_1(V_1)$.

Точка x_i считается принадлежащим кластеру V_k, если расстояние до него $D(x_i, V_k)$ меньше, чем расстояние до других кластеров. В этом случае все точки будут принадлежать каким-нибудь кластерам. Если же принять, что x_i считается принадлежащим кластеру V_k, если расстояние до него $D(x_i, V_k)$ меньше, чем некоторое пороговое значение, то могут остаться точки, не принадлежащие ни одному из найденных кластеров.

Ниже приводятся рисунки, иллюстрирующие основные идеи метода горной кластеризации.

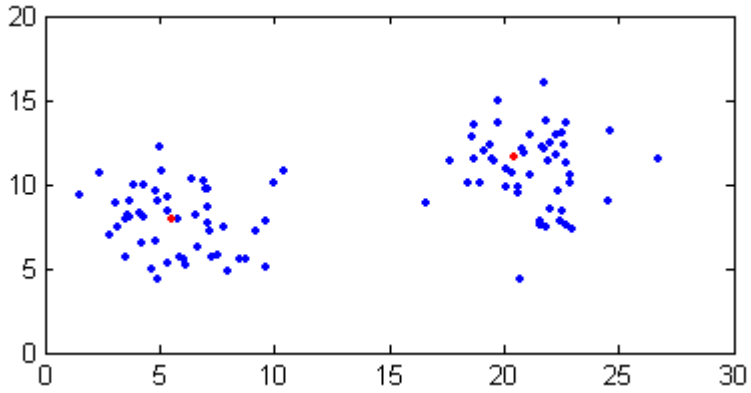

Рисунок 1. Точки на плоскости (синие) и найденные центры кластеров (красные)

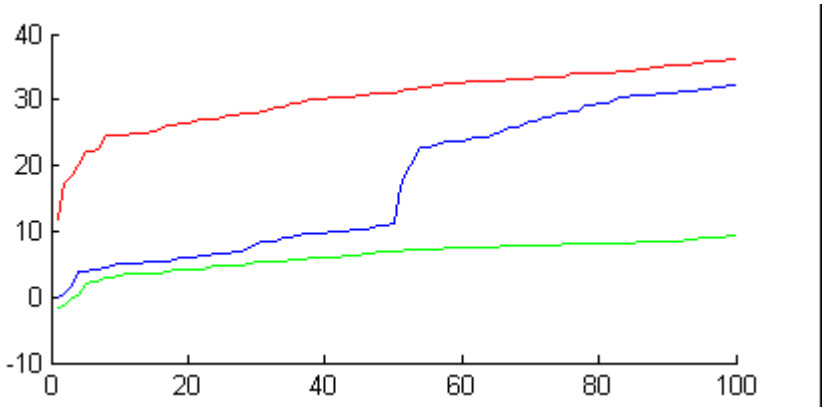

Рисунок 2. Упорядоченные значения потенциалов P1 (красная линия), P2 (синяя), P3 (зеленая).

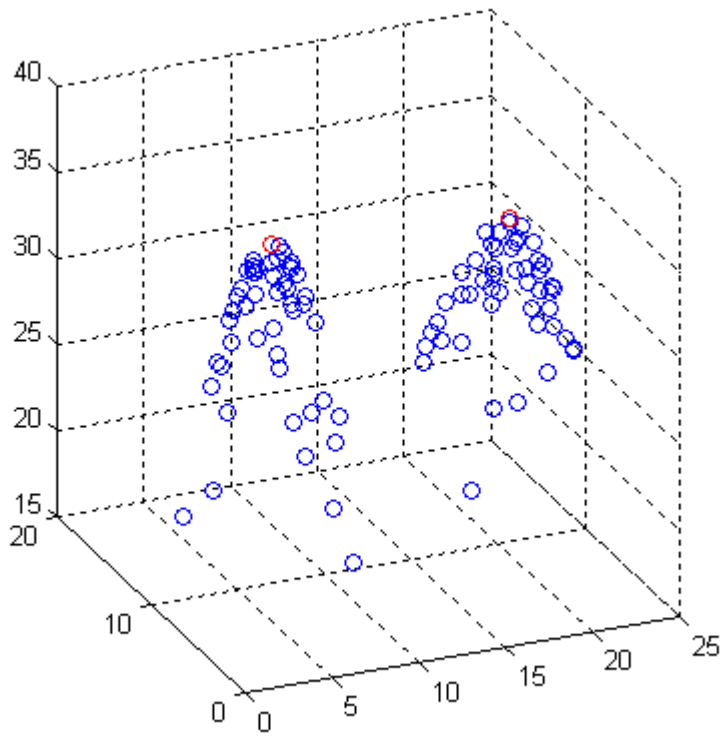

Рисунок 3. Значения потенциалов P1 точек (синим цветом) и значения потенциалов P1 центров кластеров (красным)

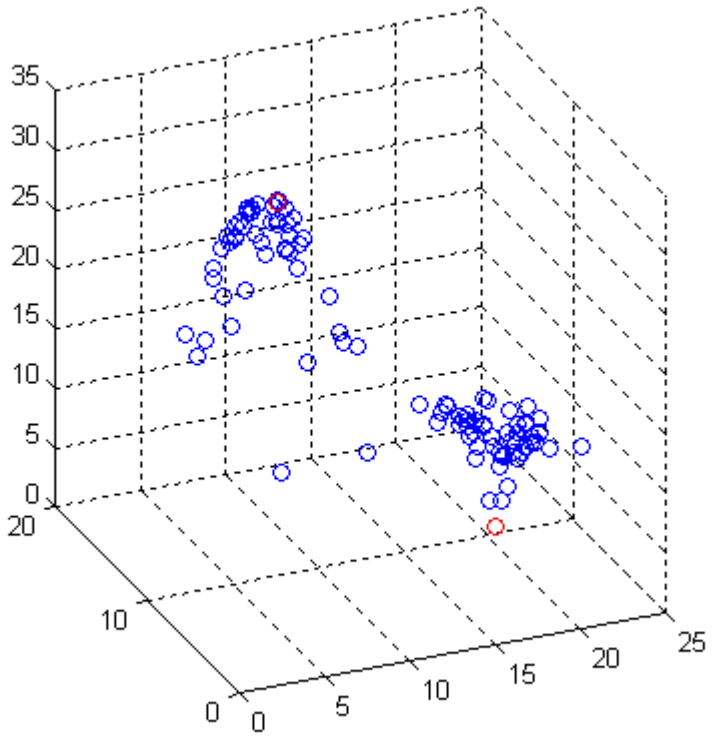

Рисунок 4. Значения потенциалов P2 точек (синим цветом) и значения потенциалов P2 центров кластеров (красным)

ЗАДАНИЕ К ЛАБОРАТОРНОЙ РАБОТЕ

Разработать программу, реализующую следующие функции:

1. Генерация случайных точек на плоскости вокруг трёх центров кластеризации (как на рисунке 1)
2. Определение потенциалов точек на первом шаге алгоритма кластеризации (как на рисунке 3)
3. Упорядочивание по возрастанию потенциалов точек, (как на рисунке 2)
4. Определение центра первого кластера, сравнение его с исходной точкой.
5. Реализация остальных шагов алгоритма, определение центров 2 и 3 кластера с визуализацией процесса (как на рисунках 1, 2, 3, 4).