АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ (ПРОДОЛЖЕНИЕ)

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. Кластеризацию проводят для объектов с количественными (числовыми), качественными или смешанными признаками. В этой лабораторной рассматривается кластеризация только для объектов с количественными признаками. Исходной информацией для кластеризации является матрица наблюдений:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & & & \\ x_{M1} & x_{M2} & \cdots & x_{Mn} \end{bmatrix},$$

каждая строчка которой представляет собой значения n признаков одного из M объектов кластеризации.

Задача кластеризации состоит в разбиении объектов из X 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 (строчками матрицы X), так и от этих объектов к прототипу 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 X на несколько непересекающихся подмножеств. При этом любой объект из X 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 этой лабораторной рассмотрим алгоритм c-средних, разбивающий данные на наперед заданное число кластеров.

Нечеткие кластера опишем следующей матрицей нечеткого разбиения:

$F = f_{ki},\ f_{ki} \in [0,1],\ k = \overline{1,M}\ ,\ i = \overline{1,c}\ ,$

где M – число объектов, c – количество кластеров, к – я строчка матрицы $F = f_{ki}$ содержит степени принадлежности объекта $(x_{k1}, x_{k2}, \ldots, x_{kn})$ к кластерам $A_1, A_2, \ldots, A_c$. При этом имеют место два условия:

$\sum_{i=1,c} f_{ki} = 1$ для всех к;

$0 < \sum_{k=1,M} f_{ki} < M$ для всех I (то есть не может быть пустых кластеров и все объекты не могут принадлежать одному кластеру).

Нечеткое разбиение позволяет просто решить проблему объектов, расположенных на границе двух кластеров - им назначают степени принадлежностей равные 0.5. Недостаток нечеткого разбиения проявляется при работе с объектами, удаленными от центров всех кластеров. Удаленные объекты имеют мало общего с любым из кластеров, поэтому интуитивно хочется назначить для них малые степени принадлежности. Однако, по условию (12.7) сумма их степеней принадлежностей такая же, как и для объектов, близких к центрам кластеров, т.е. равна единице. Для устранения этого недостатка можно использовать возможностное разбиение, которое требует, только чтобы произвольный объект из X принадлежал хотя бы одному кластеру с некоторой степенью, то есть

$0 < \sum_{i=1,c} f_{ki} < 1.$

Для оценки качества нечеткого разбиения используется такой критерий разброса:

$\sum_{i=1,c} \sum_{k=1,M} (f_{ki})^m \|V_i - X_k\|^2$, где

$V_i = \frac{\sum_{k=1,M} (f_{ki})^m X_k}{\sum_{k=1,M} (f_{ki})^m}$ - центры нечетких кластеров, $m = [1,\infty)$ - экспоненциальный вес, определяющий нечеткость, размазанность кластеров.

Предложено множество алгоритмов нечеткой кластеризации, основанных на минимизации критерия разброса. Нахождение матрицы нечеткого разбиения F с минимальным значение критерия разброса представляет собой задачу нелинейной оптимизации, которая может быть решена разными методами. Наиболее известный и

часто применяемый метод решения этой задачи алгоритм нечетких с-средних, в основу которого положен метод неопределенных множителей Лагранжа. Он позволяет найти локальный оптимум, поэтому выполнение алгоритма из различных начальных точек может привести к разным результатам.

Этапы алгоритма.

1. Установить параметры алгоритма: c - количество кластеров; m - экспоненциальный вес; ε - параметр останова алгоритма.
2. Случайным образом сгенерировать матрицу нечеткого разбиения F
3. Рассчитать центры кластеров $V_i = \frac{\sum_{k=1,M} (f_{ki})^m X_k}{\sum_{k=1,M} (f_{ki})^m}$, $i = \overline{1,c}$
4. Рассчитать расстояния между объектами X_k и центрами кластеров: $D_{ki} = \sqrt{\|X_k - V_i\|^2}$, $k = \overline{1,M}$, $i = \overline{1,c}$.
5. Пересчитать элементы матрицы нечеткого разбиения $f_{ki} = \frac{1}{\left(D_{ik}^2 \sum_{j=1,c} \frac{1}{D_{jk}^2} \right)^{1/(m-1)}}$, $k = \overline{1,M}$, $i = \overline{1,c}$.
6. Рассчитать новые центры кластеров $V_i^* = \frac{\sum_{k=1,M} (f_{ki})^m X_k}{\sum_{k=1,M} (f_{ki})^m}$, $i = \overline{1,c}$
7. Проверить условие $\left\| V_i^* - V_i \right\| < \varepsilon$ для всех $i = \overline{1,c}$ (то есть расстояние между старыми и новыми значениями центров кластеров меньше заданного порогового значения ε). Если условие истинно – то конец, найденные значения V_i^* есть искомые центры кластеров, если ложное – переход на шаг 4.

В приведенном алгоритме самым важным параметром является количество кластеров c. Его выбирают исходя из априорной информации о данных.

Вторым параметром алгоритма кластеризации является экспоненциальный вес (m). Чем больше m, тем конечная матрица нечеткого разбиения F становится более

"размазанной", и при $m \to \infty$ она примет вид $f_{ki} = 1/c$, что является очень плохим решением, т. к. все объекты принадлежат ко всем кластерам с одной и той же степенью. Кроме того, экспоненциальный вес позволяет при формировании координат центров кластеров усилить влияние объектов с большими значениями степеней принадлежности и уменьшить влияние объектов с малыми значениями степеней принадлежности. На сегодня не существует теоретически обоснованного правила выбора значения экспоненциального веса. Обычно устанавливают $m = 2$.

Ниже приведены два рисунка, поясняющие идеи алгоритма нечеткой кластеризации с-средних.

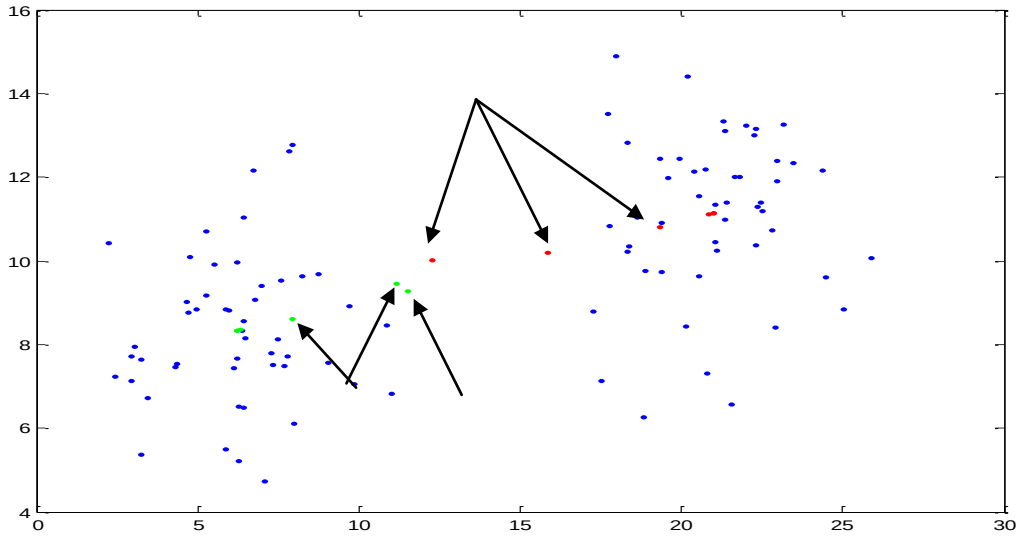

Рисунок 1. Точки на плоскости (синие) и последовательно найденные центры их кластеров, отмеченные стрелками (красные и зеленые точки).

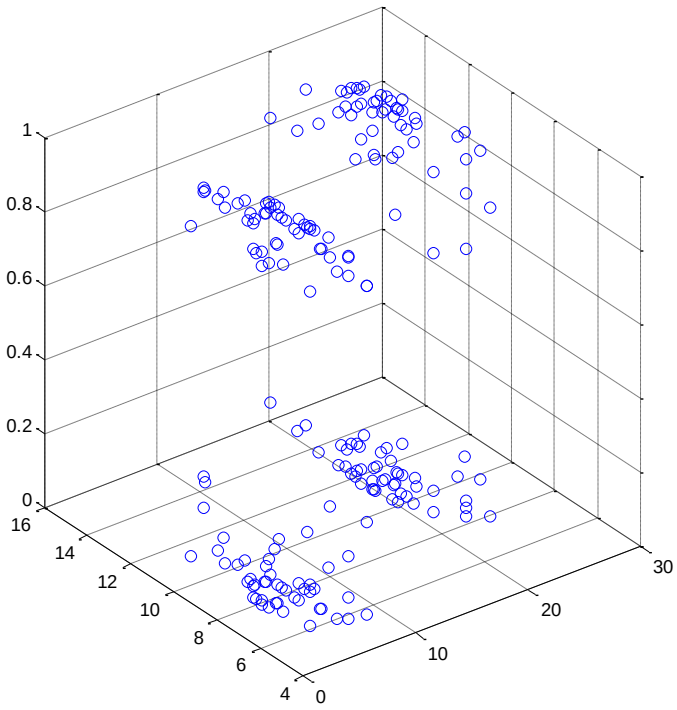

Рисунок 2. Точки и соответствующие им значения степени принадлежности к кластерам

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 центров кластеров и степени принадлежности точек к кластерам алгоритмом с-средних (понадобится 5-7 итераций).