Методы кластеризации: четкие и нечеткие.

Кластерный анализ – это метод разделение данных на классы объектов, в каждом из которых объекты имеют похожие свойства. Сам по себе кластерный анализ - это не конкретный алгоритм, а общая задача, которую необходимо решить с помощью различных алгоритмов. Не существует объективно «правильного» алгоритма кластеризации. Наиболее подходящий алгоритм кластеризации необходимо выбирать экспериментально, в зависимости от набора данных, если только не существует математической причины предпочесть один алгоритм другому.

Методы кластеризации можно разделять по способу обработки данных, по способу анализа данных, по масштабируемости, по времени выполнения и так далее. Методы по способу анализа данных, в свою очередь, разделяют на четкие (или традиционные) и нечеткие. К четким алгоритмам относятся те алгоритмы, в которых каждый объект данных принадлежит одному кластеру. К нечетким алгоритмам кластеризации относятся те, в которых каждый объект данных принадлежит более одному кластеру с определенной степенью.

В 1965 году L. Zadeh [21] представил аксиоматическую структуру – нечеткое множество. Первое применение теории нечетких множеств к кластерному анализу было сделано в 1969 году E.H. Ruspini [16]. Однако только после появления работ J.C. Dunn [9] и J.C. Bezdek [4] теория нечетких множеств приобрела актуальность для кластерного анализа и теории распознавания образов.

В 1993 году R. Krishnapuram и J.M. Keller в работе [11] представили возможностный алгоритм кластеризации. Из этой работы также выросла целая ветвь алгоритмов кластеризации.

Методы кластеризации относятся к классу задач обучения без учителя и широко используются в теории распознавания образов, анализе изображений, поиске информации, биоинформатике, сжатии данных, компьютерной графике и машинном обучении.

В данной лекции мы подробно рассмотрим некоторые четкие и нечеткие алгоритмы кластеризации и расскажем о преимуществах и недостатках каждого. Начнем с хорошо известного EM-алгоритма.

1 EM-алгоритм.

EM-алгоритм – это итеративный метод нахождения оценок максимального правдоподобия параметров статистической модели, когда модель зависит от скрытых ненаблюдаемых переменных. Каждая итерация алгоритма состоит двух шагов. На шаге ожидания (expectation) вычисляется ожидаемое значение функции правдоподобия с использование текущей оценки параметров. На шаге максимизации (maximization) вычисляются параметры, максимизирующие ожидаемую функцию максимального правдоподобия, найденную на шаге ожидания. Впервые такое название алгоритма появилось в работе [7], хотя подобная итерационная процедура рассматривалась гораздо раньше многими авторами, например, A. G. McKendrick [14] и М. И. Шлезингером [3].

1.1 Общее описания алгоритма. [1]

Пусть X и Y – случайные величины, принимающие значения в пространствах $\mathbb{R}^n$ и $\mathbb{R}^m$ соответственно, где $n, m \geq 1$. Пусть θ – параметр из некоторого множества Θ произвольной природы. Плотность совместного распределения $(n+m)$-мерного случайного вектора (X, Y) обозначим

$$f_\theta(x, y),\ x \in \mathbb{R}^n,\ y \in \mathbb{R}^m,\ \theta \in \Theta.$$

Условная плотность случайной величины Y при условии $X = x$ определяется как

$$f_\theta(y|x) = \frac{f_\theta(x, y)}{f_\theta^X(x)},\ \ y \in \mathbb{R}^m, \tag{1}$$

где

$$f_\theta^X(x) = \int\limits_{\mathbb{R}^m} f_\theta(x, y)\mu_Y(dy)$$

является маргинальной плотностью случайной величины X относительно меры μ_Y. Выражение (1) имеет смысл, если $f_\theta^X(x) \neq 0$. Аналогично определяется условная плотность случайной величины X при условии $Y = y$:

$$f_\theta(x|y) = \frac{f_\theta(x,y)}{f_\theta^Y(y)}, \;\; x \in \mathbb{R}^n. \tag{2}$$

Выражение (2) имеет смысл, если маргинальная плотность случайной величины Y относительно меры μ_X

$$f_\theta^Y(y) = \int\limits_{\mathbb{R}^n} f_\theta(x,y)\mu_X(dx) \neq 0.$$

В качестве мер μ_x и μ_Y рассматривается либо мера Лебега, либо другая считающая мера (формальный эквивалент количества элементов множества). Из соотношений (1), (2) вытекает, что

$$f_\theta(x,y) = f_\theta(y|x)f_\theta^X(x) = f_\theta(x|y)f_\theta^Y(y). \tag{3}$$

Будем считать, что случайная величина X имеет смысл наблюдаемых данных, в то время как скрытая (ненаблюдаемая) случайная величина Y играет вспомогательную роль. Зная совместную плотность $f_\theta(x,y)$ и значение x наблюдаемой величины X можно формально определить *полную* функцию правдоподобия

$$L(\theta; x, y) = f_\theta(x,y), \; \theta \in \Theta, \tag{4}$$

как функцию параметра θ. При этом

$$L(\theta; x) = f_\theta^X(x), \; \theta \in \Theta, \tag{5}$$

функция правдоподобия параметра θ при неполных данных.

Цель EM-алгоритма – найти значения параметра θ, максимизирующее функции (4) или (5) при неизвестном значении Y или, другими словами, найти оценки максимального правдоподобия параметра θ. Процедура EM-алгоритма состоит из вычисления последовательности значений $\{\theta^{(m)}\}$, $m \geq 1$ параметра θ. Если задано некоторое значение $\theta^{(m)}$, то вычисление следующего значения $\theta^{(m+1)}$ можно разделить на два этапа. Опишем эти этапы.

1. (E-этап) Определим функцию $Q(\theta, \theta^{(m)})$, как условное математическое ожидание логарифма полной функции правдоподобия при известном значении наблюдаемой компоненты X:

$$Q(\theta, \theta^{(m)}) = E_{\theta^{(m)}}(\log f_\theta(X,Y)|X). \tag{6}$$

   В этом определении θ является аргументом, а $\theta^{(m)}$ и X параметрами. При известном значении $X = x$ символ $E_{\theta^{(m)}}$ означает среднее значение случайной величины Y относительно условного распределения $f_{\theta^{(m)}}(y|x)$, то есть:

$$Q(\theta, \theta^{(m)}) = \int_{\mathbb{R}^m} (\log f_\theta(X,Y)) f_{\theta^{(m)}}(y|x)\mu_Y(dy).$$

2. (M-этап) На этом этапе вычисляется

$$\theta^{(m+1)} = \arg\max_\theta Q(\theta, \theta^{(m)}). \tag{7}$$

   Далее выбирается метрика $\rho(\cdot,\cdot)$ и фиксируется малое положительное ε. Итерационный процесс останавливается на m-ом шаге, если $\rho(\theta^{(m)}, \theta^{(m+1)}) < \varepsilon$.

Отметим свойство монотонности EM-алгоритма, которое впервые было установлено в работе [3]. Однако этого недостаточно, чтобы утверждать, что последовательность оценок параметров, построенная EM-алгоритмом, гарантировано сходится к локальному максимуму функции правдоподобия. Чтобы установить такую сходимость, приходится предполагать, что рассматриваемые распределения удовлетворяют дополнительным условиям регулярности, и, в частности, условиям гладкости (подробнее смотри в [1]). Одновременно монотонность EM-алгоритма свидетельствует о его сильной зависимости от выбора начального (стартового) приближения.

1.2 Применение EM-алгоритма к задаче разделения смесей вероятностных распределений. [1]

Задача поиска наиболее правдоподобных оценок параметров смесей вероятностных распределений является одним из самых популярных приложений EM-алгоритма. Предполагается, что данные в каждом кластере подчиняются определенному закону распределения. Для наглядности будем рассматривать смеси одномерных распределений. В рамках данной задачи плотность распределения наблюдаемой случайной величины X имеет вид

$$f_\theta^X(x) = \sum_{i=1}^{k} p_i \psi_i(x; t_i), \tag{8}$$

где $k \geq 1$ – известное натуральное число, $\psi_1, \dots \psi_k$ – известные плотности распределения, $\theta = (p_1, \dots, p_k, t_1, \dots, t_k)$ – неизвестный параметр, причем каждое $p_i \geq 0$ и $p_1 + \dots + p_k = 1$, t_i, $i = 1, \dots, k$, – многомерные параметры. Плотности $\psi_1, \dots, \psi_k$ будем называть *компонентами* смеси (8), а $p_1, \dots, p_k$ – *весами* соответствующих компонент.

Задачей разделения смеси (8) называется задача статистического оценивания параметров $\theta = (p_1, \dots, p_k, t_1, \dots, t_k)$ по известным реализациям случайной величины X.

Предположим, что имеется независимая выборка значений $x = (x_1, \dots, x_n)$ случайной величины X. В рамках модели (8) логарифм классической (неполной) функции правдоподобия параметра θ имеет вид:

$$\log L(\theta; x) = \log \prod_{j=1}^{n} f_\theta^X(x_j) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} p_i \psi_i(x_j; t_i) \right).$$

Непосредственный поиск точки максимума этой функции затруднителен. Однако, если мы будем трактовать наблюдения x, как *неполные*, то функцию правдоподобия можно записать в более удобном виде.

Предположим, что наряду с наблюдаемой случайной величиной X задана ненаблюдаемая случайная величина Y со значениями $(y_1, \dots, y_n)$, где $y_j \in \{1, 2, \dots, k\}$ содержат информацию о номерах компонент, в соответствии с которыми "генерируется" наблюдение $x = (x_1, \dots, x_n)$. Будем предполагать, что пары значений (x_j, y_j) являются стохастически независимыми реализациями пары случайных величин (X, Y).

Совместную плотность случайных величин X и Y, как и раньше, обозначим $f_\theta(x, y)$. Так как дискретная случайная величина Y абсолютно непрерывна относительно считающей меры и принимает значения $i = 1, 2, \dots, k$, то ее маргинальная плотность равна

$$f_\theta^Y(i) = p_i, \ i = 1, 2, \dots, k,$$

а условная плотность случайной величины X при фиксированном значении $Y = i$ равна

$$f_\theta(x|i) = \psi_i(x; t_i).$$

Поэтому, если бы значения $y = (y_1, \dots, y_n)$ были известны, то логарифм полной функции правдоподобия имел бы вид:

$$\log L(\theta; x, y) = \log \prod_{j=1}^{n} f_\theta(x_j, y_j) = \sum_{j=1}^{n} \log f_\theta(x_j, y_j) = \sum_{j=1}^{n} \log \left(f_\theta(x_j|y_j) f_\theta^Y(y_j) \right) = \sum_{j=1}^{n} \log p_{y_j} + \sum_{j=1}^{n} \log \psi_{y_j}(x_j; t_{y_j}).$$

После некоторых преобразований (подробнее в [1]), условное математическое ожидание логарифма полной функции правдоподобия при фиксированных значениях $x = (x_1, \dots, x_n)$ наблюдаемой случайной величины X имеет вид:

$$Q(\theta, \theta^{(m)}) = \sum_{l=1}^{k} \sum_{j=1}^{n} f_\theta(l|x_j) \log p_l + \sum_{l=1}^{k} \sum_{j=1}^{n} f_\theta(l|x_j) \log \psi_l(x_j; t_l). \tag{9}$$

Для поиска максимума функции (9) по $\theta = (p_1, \dots, p_k, t_1, \dots, t_k)$ можно максимизировать слагаемые в правой части (9) независимо друг от друга, так как они зависят от разных параметров: первое зависит только от весов $p_1, \dots, p_k$, а второе - только от параметров $t_1, \dots, t_k$ компонент смеси. Учитывая ограничение

$$\sum_{l=1}^{k} p_l = 1,$$

с помощью метода неопределенных множителей Лагранжа находим значение $\theta = (p_1, \ldots, p_k, t_1, \ldots, t_k)$, доставляющие максимум функции (9). Подразумевая, что значения $\theta^{(m)} = (p_1^{(m)}, \ldots, p_k^{(m)}, t_1^{(m)}, \ldots, t_k^{(m)})$ параметра θ на m-ой итерации известны, находим $p_1^{(m+1)}, \ldots, p_k^{(m+1)}$ на $m+1$-итерации EM-алгоритма.

Отметим, что в прикладных работах EM-алгоритм чаще всего применяется к исследованию модели (8), где $\psi_i(x; t_i)$– плотность нормального распределения вероятностей. Однако, именно эта модель не удовлетворяет условиям, гарантирующим правильную работу EM-алгоритма. А именно, сходимость EM-алгоритма доказана при обязательном условии ограниченности логарифма функции правдоподобия. Для смесей нормальных распределений указанное условие, вообще говоря, не выполняется. Также наличие большого числа локальных максимумов логарифма функции правдоподобия для модели (8) с большим числом ($k \geq 2$) нормальных компонент приводит к большой неустойчивости по отношению к начальному приближению и исходным данным.

2 Алгоритм k-средних.

Алгоритм k-средних разбивает n наблюдений на $k \leq n$ кластеров, при этом каждое наблюдение принадлежит тому кластеру, к центру которого оно ближе всего. Термин "k-средних" впервые появился в работе [13]. Алгоритм прост в реализации, но в тоже время требует больших вычислительных ресурсов. Он похож на EM-алгоритм, применяемый для разделения смеси гауссиан. Оба они используют итеративный подход уточнения, а кластерные центры используются для моделирование данных. Однако алгоритм k-средних стремится найти не пересекающиеся кластеры, имеющие сферическую форму, в то время, как EM-алгоритм позволяет кластерам пересекаться и иметь любую форму, так как функции распределения, используемые в EM-алгоритме, имеют дисперсию и ковариацию.

Пусть задано множество наблюдений $X = (x_1, \ldots, x_n)$, где $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. Требуется разбить множество наблюдений X на k непересекающихся кластеров $S_1, \ldots, S_k$, $S_i \cap S_j = \emptyset$, $i \neq j$, $\bigcup_{i=1}^{k} S_i = X$, таким образом, чтобы минимизировать сумму квадратов расстояний от каждой точки кластера до его центра (центра масс кластера), что равносильно поиску

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2, \tag{10}$$

где μ_i – центры кластеров S_i, $i = 1, \ldots, k$, $\|x - \mu_i\|^2 = \sum\limits_{x \neq y \in S_i} (x - \mu_i)(\mu_i - y)$.

Процедура алгоритма k-средних состоит из следующих шагов. Случайным образом выбираются центры кластеров $\mu_1^{(1)}, \ldots, \mu_k^{(1)}$. Далее происходит итерация между двумя шагами.

1. Каждое наблюдение x_p присваивается тому кластеру, центр которого ближе всего к наблюдению, то есть
$$S_i^{(t)} = \{x_p : \ \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2, \text{ для любого } j = 1 \ldots, k\},$$
где x_p присваивается только одному $S_i^{(t)}$, даже если его можно отнести к двум и более кластерам.

2. Перевычисление центров кластеров для уже присвоенных различным кластерам наблюдений:
$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j.$$
Алгоритм останавливается, когда $m_i^{(t)} = m_i^{(t+1)}$ для любого i.

Заметим, что число кластеров k необходимо знать заранее. Неправильный выбор k может привести к плохим результатам. Поэтому перед применением алгоритма k-средних важно выполнять диагностическую проверку для определения количества кластеров в наборе данных. Также к недостатками алгоритма k-средних можно отнести зависимость от выбора исходных центров кластеров, чувствительность к шуму, а также сходимость только к локальным минимумам. При этом данный алгоритм прост в применении, поэтому часто используется в качестве этапа предварительной обработки.

3 Алгоритм нечеткой кластеризации (FCM).

Алгоритмом нечеткой кластеризацией (FCM) называется кластеризация, в которой каждое из n наблюдений может принадлежать сразу нескольким кластерам с разной степенью принадлежности. Таким образом, данные, расположенные на границах кластеров, не обязаны полностью принадлежать одному кластеру, а могут быть в составе многих кластеров со степенью частичной принадлежности от 0 до 1.

В 1965 году Lotfi Zadeh [21] представил аксиоматическую структуру - нечеткое множество. Нечеткое множество было задумано чтобы разобраться с проблемой распознавания образов в контексте неточно определенных категорий. Подробнее об этом в нашей первой части серии лекций по Теории возможностей. В 1969 году E.H. Ruspini [16] опубликовал статью, которая стала основой для большинства алгоритмов нечеткой кластеризации. Он впервые применил теорию нечетких множеств к кластерному анализу. Однако, только после появления работ J.C. Bezdek [4] и J.C. Dunn [9] алгоритмы нечеткой кластеризации стали важной вехой в теории кластерного анализа, так как была четко установлена актуальность теории нечетких множеств для кластерного анализа и распознавания образов.

Алгоритм нечеткой кластеризации очень похож на алгоритм k-средних. Пусть задано множество наблюдений $X = (x_1, \ldots, x_n)$, где $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. Требуется разбить множество наблюдений X на c нечетких кластеров $(S_1, \ldots, S_c)$ с центрами $(\beta_1, \ldots, \beta_c) = \beta$ таким образом, чтобы минимизировать функцию потерь

$$\arg\min_{(U,\beta)} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^m \|x_i - \beta_j\|^2, \tag{11}$$

где $w_{ij} \in [0,1]$ – степень принадлежности элемента x_i кластеру S_j с центром β_j, которая удовлетворяет ограничениям

$$w_{ij} \in [0,1] \text{ для всех } i, j, \tag{12}$$

$$0 < \sum_{j=1}^{n} w_{ij} < n \text{ для всех } i, \tag{13}$$

$$\sum_{i}^{c} w_{ij} = 1 \text{ для всех } j. \tag{14}$$

Число $m \in [1, +\infty)$ в функции (11) – экспоненциальный вес, определяющий нечеткость кластеров. Из необходимых условий локального экстремума получаем:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - \beta_j\|}{\|x_i - \beta_k\|} \right)^{\frac{2}{m-1}}}, \; i = 1, \ldots, n, \; j = 1, \ldots, c, \tag{15}$$

$$\beta_j = \frac{\sum_{i=1}^{n} w_{ij}^m x_i}{\sum_{i=1}^{n} w_{ij}^m}, \; j = 1, \ldots, c. \tag{16}$$

Ограничение (14) обобщает соотношение, первоначально предложенное L.A. Zadeh [21], для определения степени принадлежности любой точки x нечеткому множеству S и его дополнению S'. А именно, дополнение S' к нечеткому множеству S определяется равенством $f_{S'}(x) = 1 - f_S(x)$, где $f_S(x) : X \to [0,1]$ – характеристическая функция нечеткого множества S, которая ставит в соответствие каждой точке из X действительное число из отрезка $[0,1]$(см. первую часть лекций по теории возможностей). Из-за своего сходства с аддитивным законом вероятностей, соотношение (14) часто называют вероятностным ограничением. Однако, закон (14) описывает природу классифицируемого набора данных, а не статистическое предположение о случайном процессе, который генерирует набор данных.

Процедура алгоритма нечеткой кластеризации состоит из следующих шагов. Случайным образом сгенерировать матрицу нечеткого разбиения $W^{(1)} = \{w_{ij}^{(1)}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, c$. Вычислить центры кластеров по формуле (16). Далее происходит итерации между шагами.

1. Расчитать расстояние от каждого наблюдения x_i до центров кластеров β_j, то есть $\|x_i - \beta_j\|$;
2. Пересчитать элементы матрицы нечеткого разбиения W по формуле (15);

3. Перевычислить центры кластеров по формуле (16) для новых элементов матрицы W из пункта 2;
4. Сравнить $W^{(t+1)}$ с $W^{(t)}$, где t – номер итерации. Если $\|W^{(t+1)} - W^{(t)}\| < \varepsilon$ (для заданного ε), то останавливаемся, иначе – переходим на первый шаг.

Число кластеров c, как и в случае алгоритма k-средних, необходимо знать заранее. Чем больше экспоненциальный вес m, тем более "размазанной" становится конечная матрица нечеткого разбиения W. При $m \to \infty$ элементы матрицы W будут равны $1/c$. Это будет говорить о том, что все наблюдаемые элементы принадлежат всем кластерам с одной и той же степенью $1/c$. При $m \to 1$ элементы матрицы W будут сходится либо к 0, либо к 1, что говорит о четком разделении наблюдаемых элементов на кластеры, а функция минимизации будет совпадать с функцией минимизации в алгоритме k-средних. Кроме того, экспоненциальный вес m позволяет при вычислении 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 $m = 2$.

Несмотря на успехи алгоритма нечеткой кластеризации в более естественном разделении данных на кластеры, проблема чувствительности к шуму осталась. Достаточно рассмотреть простой пример. Пусть у нас есть два кластера. Если наблюдения x_k равноудалены от центров обоих кластеров, то в независимости от удаленности от этих центров, их степени принадлежности из условия (14) совпадают и равны 0.5. Поэтому шумовым точкам, находящимся далеко, но равноудаленно от центров двух кластеров, тем не менее может быть присвоено равное членство в обоих кластерах, тогда как кажется гораздо более естественным, чтобы такие точки имели очень низкую степень принадлежности или даже не принадлежали ни к какому-либо кластеру.

К недостатками алгоритма нечеткой кластеризации также можно отнести зависимость от выбора значений степеней принадлежности w_{ij} на начальном этапе, а также сходимость к локальным минимумам.

Форма кластеров в любом алгоритме кластеризации определяется функцией, которая исследуется на минимум, в которой в свою очередь участвует расстояние, индуцирующее топологическую метрику в $\mathbb{R}^d$. Поэтому в алгоритме нечеткой кластеризации кластеры могут принимать форму, близкую к сферической, как и в случае алгоритма k-средних. Чтобы преодолеть это ограничение алгоритмы нечеткой кластеризации пошли в своем развитии по следующим направлениям.

Первое направление относится к алгоритмам, основу которых составила работа D.E. Gustafson and W.C. Kessel [10]. В ней предложено заменить норму расстояния $\|x_i - \beta_j\|^2$ в функции, исследуемой на минимум, на альтернативную норму $\|x_i - \beta_j\|^2_{A_j} = (x_i - \beta_j)^T A_j (x_i - \beta_j)$, где A_j – симметричные положительно-определенные матрицы и $\det A_j = \rho_j$, $\rho_j > 0$. Таким образом, каждый кластер принимает форму, которая заложена в A_j. Также они признали, что необходимые условия для поиска локального минимума, имеют сходство с уравнениями максимального правдоподобия в EM-алгоритме, применяемом для разделения смеси гауссиан.

Второе направление, по которому развивались нечеткие алгоритмы, появилось из работы [5], в которой нечеткие кластеры имеют форму линии. Из этой работы выросло целое направление разных алгоритмов нечеткой кластеризации, в которых центры кластеров заменяются на более общие структуры, типа линий, плоскостей, гиперкубов и так далее.

4 Возможностный алгоритм кластеризации (PCM).

R. Krishnapuram и J.M. Keller [11] предложили идею ослабления ограничения (14) путем добавления второго члена в функции (11), что позволило решить проблему с шумовыми точками.

Пусть задано множество наблюдений $X = (x_1, \ldots, x_N)$, где $x_i \in \mathbb{R}^d$, $i = 1, \ldots, N$, $\beta = (\beta_1, \ldots, \beta_C)$ – центры кластеров, d_{ij}^2 – расстояние от точки x_i до центра β_j, а $U = \{u_{ij}\}$ – матрица размером $C \times N$, элементы которой являются характеристическими значениями элемента x_j по отношению к кластеру S_i. Требуется разбить множество наблюдений X на C нечетких кластеров $S_1, \ldots, S_C$, таким образом, чтобы минимизировать функцию потерь

$$\arg\min_{(U,\beta)} \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2 + \sum_{i=1}^{C} \eta_i \sum_{j=1}^{N} (1 - u_{ij})^m, \tag{17}$$

где $m \in [1, +\infty)$ – экспоненциальный вес, η_i – подходящие положительные числа. На характеристические значения u_{ij} взамен (12)–(14) накладываются следующие ограничения:

$$u_{ij} \in [0, 1] \text{ для всех } i, j, \tag{18}$$

$$0 < \sum_{j=1}^{N} u_{ij} \leq N \text{ для всех } i, \tag{19}$$

$$\max_i u_{ij} > 0 \text{ для всех } j. \tag{20}$$

Минимизация функции (17) предполагает, чтобы в первом слагаемом расстояние от точки x_i до центра кластера β_j было как можно меньше, в то время, как во втором слагаемом u_{ij} должно быть как можно ближе к 1. Если бы в (17) отсутствовало бы второе слагаемое, то без ограничения вида (14) на u_{ij}, минимизация функции приводила бы к тривиальному решению $u_{ij} = 0$ для всех i, j.

Заметим, что строки и столбцы матрицы $U = \{u_{ij}\}$ независимы друг от друга. Поэтому минимизацию функцию (17) можно свести к минимизации CN независимых функций

$$u_{ij}^m d_{ij}^2 + \eta_i (1 - u_{ij})^m. \tag{21}$$

Согласно необходимым условиям локального экстремума, получаем:

$$u_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^2}{\eta_i}\right)^{\frac{1}{m-1}}}, \; i = 1, \ldots C, \; j = 1, \ldots, N, \tag{22}$$

$$\beta_j = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}, \; i = 1, \ldots, C. \tag{23}$$

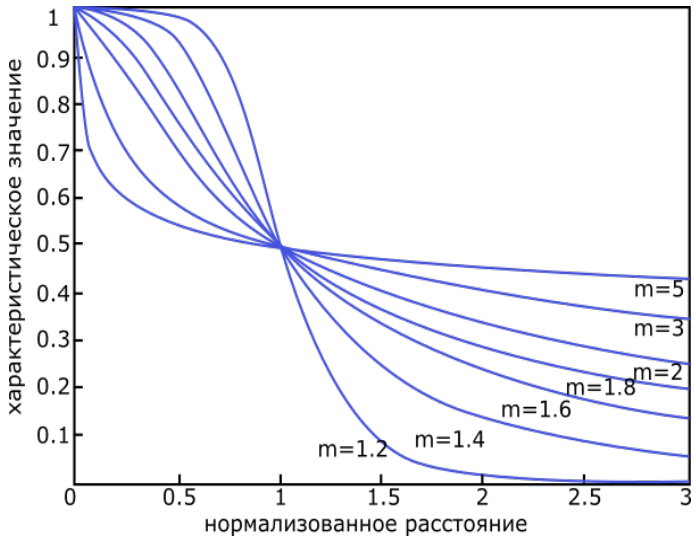
Элементы матрицы $U = \{u_{ij}\}$ сильно зависит от выбора параметра η_i. Если η_i маленькое, то и u_{ij} маленькое. Если η_i большое, то u_{ij} также большое. Также η_i определяет степень, с которой второе слагаемое в (17) сравнимо с первым. Если оба слагаемых в (17) равновесны, то η_i должно быть порядка d_{ij}^2. R. Krishnapuram и J.M. Keller предложили следующие соотношения для η_i ([11], [12]):

$$\eta_i = \frac{\sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} u_{ij}^m}, \; i = 1, \ldots, C, \tag{24}$$

$$\eta_i = \frac{\sum_{u_{ij} > \alpha} d_{ij}^2}{\sum_{u_{ij} > \alpha} 1}, \; i = 1, \ldots, C, \tag{25}$$

где $0 < \alpha < 1$. Параметр η_i может быть фиксированным для всех итераций алгоритма, если кластеры имеют похожую форму. В общем случае η_i меняется в каждой итерации алгоритма, что может привести к неустойчивости, так как необходимые условия локального экстремума получены для фиксированного η_i. Поэтому часто сначала применяют алгоритм нечеткой кластеризации для инициализации u_{ij}, далее вычисляют η_i по формуле (24), после чего применяют возможностный алгоритм кластеризации, в котором η_i вычисляется по формуле (25).

Значение m играет важную роль в определении характеристических значений u_{ij}. На следующем рисунке видно, что при $m \to 1$ характеристические значения u_{ij} стремятся к нулю для тех точек x_j, для которых d_{ij}^2 больше, чем η_i. При $m \to \infty$ характеристические значения перестают стремится к нулю. Значение $m = 2$ дает хорошие результаты в алгоритме нечеткой кластеризации. Однако, в возможностном алгоритме для такого значения m характеристические функции убывают не достаточно быстро для больших значения d_{ij}^2. Поэтому более подходящий выбор $m = 1.5$ [12].



Процедура возможностного алгоритма кластеризации выглядит следующим образом. Генерируем элементы матрицы $U = \{u_{ij}\}$. Вычисляем кластерные центры по формуле (23). Далее происходит итерация между шагами:

1. Рассчитать расстояние d_{ij} от каждого наблюдения x_i до центров кластеров β_j;
2. Вычислить η_i по формуле (24) или (25);
3. Пересчитать элементы матрицы U по формуле (22);
4. Перевычислить центры кластеров β_j по формуле (23) для новых элементов матрицы U из пункта 3;
5. Сравнить $u_{ij}^{(t+1)}$ с $u_{ij}^{(t)}$, где t – номер итерации. Если $\|u_{ij}^{(t+1)} - u_{ij}^{(t)}\|^2 < \varepsilon$ (для заданного ε), то останавливаемся, иначе – переходим на первый шаг.

Предложенный R. Krishnapuram и J.M. Keller возможностный алгоритм – это только некоторая реализация общей идеи возможностного подхода. Возможностный подход означает, что характеристическое значение точки по отношению к кластеру представляет собой возможность точки принадлежать кластеру.

Так как минимизация функции (17) сводится к минимизации CN независимых функций (21) , то могут возникнуть совпадающие кластеры. Это проблема типична для функций, которые можно выразить, как сумму независимых функций. Причина кроется не в плохом выборе второго слагаемого в (17), а скорее в отсутствии подходящих ограничений на u_{ij}. С одной стороны ограничение (14) в алгоритме нечеткой кластеризации слишком сильное – оно заставляет шумовые точки принадлежать одному или нескольким кластерам с достаточно высокими степенями принадлежности. С другой стороны, ограничение (19) в возможностном алгоритме слишком слабое, так как матрица U сильно зависит от выбора параметров m и η_i. И хотя возможностный алгоритм кластеризации более робастный к шуму, так как шумовые точки будут принадлежать кластерам с маленькими характеристическими значениями, платить за это придется совпадающими кластерами.

Чтобы преодолеть проблему чувствительности к шуму, а также проблему совпадающих кластеров, было предложено несколько алгоритмов. Например, в работе [15] была предложена возможностно-нечеткая модель кластеризации (PFCM), в которой функция, исследуемая на минимум, включала и характеристические значения u_{ij}, и степени принадлежности w_{ij}. Однако этот алгоритм по-прежнему сталкиваются с проблемами инициализации и выбора параметров модели. Еще один алгоритм, предложенный в [23], основан на идее, что на начальном этапе все наблюдаемые данные являются центрами кластеров. Затем происходит процедура автоматического слияния точек специальным образом в соответствии с исходной структурой данных. При этом число кластеров находится автоматически с сохранением робастности алгоритма. Тот факт, что все точки используются в качестве начальных центров кластеров, является серьезной проблемой при масштабировании этого алгоритма для больших объемов данных и высокой размерности.

Конечно, все направления по развитию алгоритмов кластеризации, основанных на теории нечетких множеств L. Zadeh, здесь описать мы не можем. Но некоторые постарались донести в простой и понятной форме.

Обзор по этой теме можно посмотреть [18].

Литература

[1] В.Ю. Королев, EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. ИПИ РАН Москва, 2007.

[2] Ю.П. Пытьев, Возможность. Элементы теории и применения. М.: Эдиториал УРСС, 2000.

[3] М. И. Шлезингер, О самопроизвольном распознавании образов. – в сб.: Читающие автоматы. "Наукова думка", Киев, 1965.

[4] J. C. Bezdek, Fuzzy Mathematics in Pattern Classification, Ph.D. thesis, Cornell Univ., Ithaca, NY, 1973.

[5] J. C. Bezdek, R. Gunderson, R. Ehrlich, and T. Meloy, On the extension of fuzzy k-means algorithms for the detection of linear clusters, in Proc. IEEE Conf. Decision and Control, 1978, pp. 1438–1443.

[6] A.P. Dempster, Upper and Lower Probabilities Induced by a Multivalued Mapping, Ann. Math. Statist., V.38, 1967.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B, V. 39, No. 1, 1977.

[8] D. Dubois and H. Prade, Possibility Theory, Probability Theory and Multiple-valued Logics: A Clarification, Annals of Mathematics and Artificial Intelligence 32, 2001.

[9] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, J. Cybern. 3 ,1974.

[10] D. E. Gustafson and W. C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in Proc. IEEE Conf. Decision and Control, pp. 761–766. 1978.

[11] R. Krishnapuram and J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (2) 1993.

[12] R. Krishnapuram and J. M. Keller, The Possibilistic C-Means Algorithm: Insights and Recommendations, IEEE Trans. Fuzzy Syst. 4(3) 1996.

[13] J.B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol.1: Statistics, University of California Press.,1967.

[14] A. G. McKendrick, Applications of mathematics to medical problems. – Proceedings of the Edinburgh Mathematical Society, 1926, vol. 44, p. 98-130.

[15] N.R. Pal, K. Pal, J. M. Keller and J. C. Bezdek, A Possibilistic Fuzzy C-Means Clustering Algorithm, IEEE Trans. Fuzzy Syst. 13(4) 2005.

[16] E. H. Ruspini, A new approach to clustering, Inf. Control, vol. 15, no. 1, 1969.

[17] E. H. Ruspini, On the semantics of fuzzy logic, Int. J. Approx. Reason., vol. 5, no. 1, 1991.

[18] E. H. Ruspini, J. C. Bezdek and J. M. Keller, Fuzzy Clustering: A Historical Perspective, IEEE Computational Intelligence Magazine, 14(1) 2019.

[19] L. J. Savage, The foundations of statistics. Dover Publication, Inc., New York, 1972.

[20] G. Shafer, A mathematical theory of evidence, Princeton University Press, 1976.

[21] L.A. Zadeh, Fuzzy sets, Information and Control, V.8, 1965.

[22] L.A. Zadeh, Calculus of fuzzy restrictions in: L.A. Zadeh, K.S. Fu, K. Tanaka and M. Shimura, eds., Fuzzy Sets and Their Applications to Cognitive and Decision Processes, Academic Press, New York, 1975.

[23] M.-S. Yang and C.-Y. Lai, A robust automatic merging possibilistic clustering method, IEEE Trans. Fuzzy Syst., 19(1) 2011.