УДК 519.681

О.В. Серая

Национальный технический университет «ХПИ», Харьков

НЕЧЕТКАЯ ЗАДАЧА КЛАСТЕРНОГО АНАЛИЗА

Рассмотрена задача кластеризации для случая, когда координаты центров кластеров и объектов группирования определены нечеткими числами с заданными функциями принадлежности. Предложена простая и легко реализуемая вычислительная процедура четкого распределения объектов по кластерам, основанная на выборе для каждого из объектов наиболее предпочтительного кластера.

Ключевые слова: *кластерный анализ, нечеткие координаты объектов, функции принадлежности нечетких чисел, степень предпочтения одного нечеткого числа перед другими, итерационная процедура.*

Введение

Постановка проблемы и анализ литературы. Кластерный анализ занимает одно из центральных мест среди декомпозиционных методов анализа данных. Теоретическая основа кластеризации определяется совокупностью методов, технологий и реализующих их алгоритмов, предназначенных для отыскания некоторого разбиения исследуемой совокупности объектов на подмножества похожих между собой объектов. При этом к результатам кластеризации обычно предъявляют следующие требования:

- каждый кластер должен содержать объекты с близкими значениями свойств или признаков;
- множество всех кластеров должно быть исчерпывающим, то есть содержать все объекты исследуемой совокупности;
- никакой объект из их совокупности не должен одновременно принадлежать разным кластерам.

Для решения задачи кластеризации разработано множество методов, однотипная концепция построения которых хорошо просматривается на примере метода k-средних, одного из наиболее часто используемых на практике [1]. Пусть имеется некоторое множество из n объектов, которые следует распределить по m кластерам. Для каждого кластера каким-либо образом определяется типичный представитель или центр кластера. Затем для каждого из объектов вычисляется «расстояние» до каждого из центров, после чего объект относят к тому из кластеров, «расстояние» до центра которого является наименьшим. Процедура отнесения объектов к кластерам может быть усовершенствована, если собственно присоединение осуществлять не сразу после сравнения расстояний от этого объекта до центров кластеров, а после того, как будут просмотрены все объекты, не рассмотренные и не присоединенные ранее. По окончанию разбиения всех объектов на кластеры осуществляется пересчет положения центров и описанная процедура повторяется до выполнения некоторого естественного критерия останова. Реальные задачи кластеризации усложняются в связи с тем, что при решении задач структуризации большинства сложных систем формируемые классы объектов размыты по своей природе. Эта размытость проявляется в том, что переход от принадлежности объекта к какому-либо кластеру до его непринадлежности скорее постепенен, нежели скачкообразен. Поэтому для каждого объекта вопрос состоит не в том, принадлежит ли этот объект к данному кластеру, а в том, до какой степени объект принадлежит каждому из кластеров. Формальное описание соответствующей задачи имеет следующий вид.

Предполагается, что каждый объект может быть представлен точкой в некотором q-мерном пространстве признаков (контролируемых параметров) и, таким образом, каждому из них ставится в соответствие вектор $X_j = \{x_{j1},...,x_{jq}\}$, $j = 1,...,n$. Пусть требуемое число кластеров известно и равно m. На первой итерации алгоритма устанавливается некоторое исходное нечеткое разбиение объектов на эти m кластеров, которое описывается следующим образом. Для каждого объекта, положение которого определяется соответствующим вектором X_j, $j = 1,...,n$, задается набор $\mu_k(X_j)$, $k = 1,...,m$, степеней принадлежности этого объекта каждому из кластеров. Для выбранного нечеткого разбиения рассчитываются координаты центров тяжести кластеров по формуле

$$a_{kp} = \sum_{j=1}^{n} \left(\mu_k(X_j)\right)^{\beta} x_{jp} \Big/ \sum_{j=1}^{n} \left(\mu_k(a_i)\right)^{\beta}, \quad (1)$$

$$k = 1,2,...,m,\ p = 1,2,...,q,$$

где a_{kp} – p-ая координата центра k-го кластера; β – некоторый параметр (> 1), называемый экспоненциальным весом. Затем с использованием (1) формируется новое нечеткое разбиение объектов на кластеры, характеризуемое новым набором функций принадлежности $\mu'_k(X_j)$, который рассчитывается как

$$\mu'_k(x_j) = \left[\sum_{l=1}^{m}\left(\sqrt{\sum_{p=1}^{q}\left(x_{jp}-a_{kp}\right)^2} \Big/ \sqrt{\sum_{p=1}^{q}\left(x_{jp}-a_{lp}\right)^2}\right)^{2/(\beta-1)}\right]^{-1}, \quad (2)$$

$$k = 1,2,...,m,\ j = 1,2,...,n.$$

Соотношение (2) в частном случае, когда $\beta = 2$, легко трактуется. Действительно, при этом

$$\mu'_k(x_j) = \left[\sum_{l=1}^{m}\left(\sum_{p=1}^{q}(x_{jp} - a_{kp})^2\right) \Big/ \left(\sum_{p=1}^{q}(x_{jp} - a_{lp})^2\right)\right]^{-1} =$$

$$= \left[\sum_{l=1}^{m}\frac{\rho_{jk}^2}{\rho_{jl}^2}\right]^{-1} = \left[\rho_{jk}^2\sum_{l=1}^{m}\frac{1}{\rho_{jl}^2}\right]^{-1} == \left[\frac{1}{g_{jk}^2}\sum_{l=1}^{m}g_{jl}^2\right]^{-1} = \frac{g_{jk}^2}{\sum_{l=1}^{m}g_{jl}^2}, \quad (3)$$

где ρ_{jk}^2 – мера удаленности j-го объекта от центра k-го кластера; $g_{jk}^2 = 1/\rho_{jk}^2$ – мера близости j-го объекта к центру k-го кластера. Понятно, что из (3) следует $\sum_{k=1}^{m}\mu'_k(x_j) = 1$, $j = 1,2,...,n$.

Далее, для полученного нечеткого разбиения по формуле (1) рассчитываются новые координаты центров тяжести объектов и процедура повторяется.

Этот алгоритм был предложен Дж. Данном [2] в 1974г., а затем Дж. Бездек [3] доказал его сходимость. Описанный метод кластеризации широко используется. Следует, однако, заметить, что на практике достаточно часто возникает необходимость решения задачи кластеризации в иной постановке. Так, например, в задачах классификации, медицинской и технической диагностики число центров кластеризации заранее известно, однако их положение определено нечетко. Точно так же нечетко заданы описания объектов. Теперь задача состоит в отыскании четкого разбиения объектов на кластеры в условиях нечетких исходных данных. Рассмотрим метод решения такой задачи.

Цель статьи. Поставим задачу построения процедуры, обеспечивающей получение четкого разбиения нечетко заданных объектов на кластеры, положение центров группирования которых также заданы нечетко. При этом из множества возможных вариантов организации процедуры группирования выберем метод, реализующий концепцию алгоритма k-средних.

Постановка задачи. Пусть для каждого из m кластеров (возможных диагнозов) заданы функции принадлежности $\mu(a_{kp})$, где a_{kp} - нечеткое значение p-го контролируемого параметра, соответствующего центру k-го кластера, $k = 1,2,...,m$, $p = 1,2,...,q$. Предположим далее, что в результате выполнения соответствующих измерений определены нечеткие координаты положения объектов кластеризации (диагностики), задаваемые совокупностью функций принадлежности $\mu(x_{jp})$, $j = 1,2,...,n$, $p = 1,2,...,q$. Задача состоит в получении четкого разбиения объектов по кластерам, наилучшим в некотором выбранном смысле.

Основные результаты

Для описания функций принадлежности $\mu(a_{kp})$, $\mu(x_{jp})$, $k = 1,2,...,m$, $j = 1,2,...,n$, $p = 1,2,...,q$, используем функции $(L-R)$ типа [4]:

$$\mu(a_{kp}) = \begin{cases} L\left(\left(\bar{a}_{kp} - a_{kp}\right)/\alpha_{kp}^{(a)}\right), & a_{kp} \le \bar{a}_{kp}; \\ R\left(\left(a_{kp} - \bar{a}_{kp}\right)/\beta_{kp}^{(a)}\right), & a_{kp} > \bar{a}_{kp}; \end{cases} \quad (4)$$

$$\mu(x_{jp}) = \begin{cases} L\left(\left(\bar{x}_{jp} - x_{jp}\right)/\alpha_{jp}^{(x)}\right), & x_{jp} \le \bar{x}_{jp}; \\ R\left(\left(x_{jp} - \bar{x}_{jp}\right)/\beta_{jp}^{(x)}\right), & x_{jp} > \bar{x}_{jp}, \end{cases} \quad (5)$$

где $\bar{a}_{kp}$ – модальное значение p-го параметра для k-го центра кластеризации, $k = 1,2,...,m$; $\bar{x}_{jp}$ – модальное значение p-го контролируемого параметра j-го объекта, $j = 1,2,...,n$; $\alpha_{kp}^{(a)}$, $\alpha_{jp}^{(x)}$, $\beta_{kp}^{(a)}$, $\beta_{jp}^{(x)}$ – левые и правые коэффициенты нечеткости в описаниях (4), (5). Эти функции могут быть заданы, например, следующим образом

$$\mu(a_{kp}) = \begin{cases} \exp\left(-\left(\bar{a}_{kp} - a_{kp}\right)^2 \Big/ \left(2\alpha_{kp}^2\right)\right), & a_{kp} \le \bar{a}_{kp}; \\ \exp\left(-\left(a_{kp} - \bar{a}_{kp}\right)^2 \Big/ \left(2\beta_{kp}^2\right)\right), & a_{kp} > \bar{a}_{kp}. \end{cases}$$

Предлагаемая процедура кластеризации состоит в следующем. Для каждого объекта рассчитывается нечеткое расстояние до каждого из центров кластеров и соответствующая функция принадлежности. Затем полученные функции принадлежности используются для отыскания кластера, имеющего по отношению к рассматриваемому объекту наивысшую степень предпочтения. Выполним формальное описание процедуры. Для пары (k-й кластер, j-й объект) введем нечеткое значение квадрата расстояния от центра кластера до объекта

$$\rho_{kj}^2 = \sum_{p=1}^{q}\left(a_{kp} - x_{jp}\right)^2.$$

Получим функцию принадлежности нечеткого числа ρ_{kj}^2. При проведении расчетов используем известные соотношения для результатов выполнения операций над нечеткими числами $(L-R)$ типа [4, 5].

Пусть $A_{LR} = \langle a_1, \alpha_1, \beta_1 \rangle$, $B_{LR} = \langle a_2, \alpha_2, \beta_2 \rangle$ - нечеткие числа $(L-R)$ типа. Тогда параметры нечеткого $C_{LR} = A_{LR} + B_{LR} = \langle a, \alpha, \beta \rangle$ вычисляются по формулам: $a = a_1 + a_2$, $\alpha = \alpha_1 + \alpha_2$, $\beta = \beta_1 + \beta_2$; параметры $C_{LR} = A_{LR} - B_{LR} = \langle a, \alpha, \beta \rangle$ вычисляются по формулам: $a = a_1 - a_2$, $\alpha = \alpha_1 + \beta_2$, $\beta = \alpha_2 + \beta_1$; параметры $C_{LR} = A_{LR} \cdot B_{LR} = \langle a, \alpha, \beta \rangle$ –по формулам: $a = a_1 a_2$, $\alpha = |a_1|\alpha_2 + |a_2|\alpha_1$, $\beta = |a_1|\beta_2 + |a_2|\beta_1$.

В соответствии с этим параметры нечетких чисел $\Delta_{kjp} = a_{kp} - x_{jp}$, $\Delta^2_{kjp} = \left(a_{kp} - x_{jp}\right)^2$, $\rho^2_{jk} = \sum_{p=1}^{q} \Delta^2_{kjp}$ определяется соотношениями:

$$\overline{\Delta}_{kjp} = \overline{a}_{kp} - \overline{x}_{jp},\ \alpha'_{kjp} = \alpha^{(a)}_{kp} + \beta^{(x)}_{jp},$$

$$\beta'_{kjp} = \alpha^{(x)}_{jp} + \beta^{(a)}_{kp};$$

$$\overline{\Delta}^2_{kjp} = \left(\overline{\Delta}_{kjp}\right)^2,\ \alpha_{kjp} = 2\left|\overline{\Delta}_{kjp}\right|\left(\alpha^{(a)}_{kp} + \beta^{(x)}_{jp}\right),$$

$$\beta_{kjp} = 2\left|\overline{\Delta}_{kjp}\right|\left(\alpha^{(x)}_{jp} + \beta^{(a)}_{kp}\right);$$

$$\overline{p}^2_{jk} = \sum_{p=1}^{q}\left(\overline{\Delta_{kjp}}\right),\ \alpha_{kj} = \sum_{p=1}^{q}\alpha_{kjp},\ \beta_{kj} = \sum_{p=1}^{q}\beta_{kjp}. \quad (6)$$

При этом функция принадлежности нечеткого значения квадрата расстояния от k-го центра до j-го объекта имеет вид

$$\mu\left(x_{jp}\right) = \begin{cases} L\left(\left(\left(\overline{\rho^2_{kj}}\right) - \rho^2_{kj}\right)\Big/\alpha_{kj}\right), & \rho^2_{kj} \le \left(\overline{\rho^2_{kj}}\right); \\ R\left(\left(\left(\overline{\rho^2_{kj}}\right) - \rho^2_{kj}\right)\Big/\beta_{kj}\right), & \rho^2_{kj} > \left(\overline{\rho^2_{kj}}\right). \end{cases} \quad (7)$$

В результате реализации описанной процедуры для каждого из объектов будут получены функции принадлежности m нечетких чисел, отображающих «расстояния» до центров соответствующих кластеров. Эти числа теперь необходимо сравнить между собой, выбирая то из них, для которого степень предпочтения по отношению ко всем остальным будет наименьшей. Это число будет определять кластер, «ближайший» по отношению к рассматриваемому объекту. Процедура сравнения нечетких чисел традиционна [4,5]. Пусть задана совокупность нечетких чисел $z_1, z_2, ..., z_m$ и их функции принадлежности $\mu(z_1), \mu(z_2), ..., \mu(z_m)$. Оценка степени предпочтения нечеткого числа z_k перед нечетким числом z_l осуществляется по формуле:

$$\eta\left(\mu(z_k), \mu(z_l)\right) = \sup_{z_k > z_l} \min\left\{\mu(z_k), \mu(z_l)\right\}, \quad (8)$$

$$k, l \in \{1, 2, ..., m\}.$$

С использованием (8) выбор нечеткого числа с наименьшей степенью предпочтения по отношению к другим числам совокупности трудностей не вызывает. При этом номер кластера k^*, к которому будет присоединен очередной объект определяется как

$$k^* = \arg\min_k \min_l \left\{\eta\left(\mu(z_k), \mu(z_l)\right)\right\},. \quad (9)$$

$$k, l \in \{1, 2, ..., m\}.$$

Заметим теперь, что если для каждого из объектов представляет интерес распределение степеней принадлежности к каждому из кластеров, то оно может быть получено с использованием (2):

$$\mu_k\left(x_j\right) = \left[\sum_{l=1}^{m}\left(\sqrt{\sum_{p=1}^{q}\overline{\Delta}^2_{kjp}}\Big/\sqrt{\sum_{p=1}^{q}\overline{\Delta}^2_{ljp}}\right)^{2/(\beta-1)}\right]^{-1},$$

$$k = 1, 2, ..., m,\ j = 1, 2, ..., n.$$

Описанная вычислительная процедура итерационно повторяется до выполнения какого-либо естественного критерия качества.

Выводы

Таким образом, рассмотрена технология нечеткой кластеризации объектов в ситуации, когда положение объектов в многомерном пространстве параметров объектов задано нечетко. Соответствующая вычислительная процедура проста и легко реализуется.

Список литературы

1. *Дюран Б. Кластерный анализ: пер. с франц./ Б. Дюран, П. Оделл. – М.: Статистика, 1977. – 128 с.*
2. *Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters / J.C. Dunn // Journal on cybernetics. – 1974. – Vol. 3. – P. 32-37.*
3. *Bezdek J.C. Some recent applications of fuzzy c-means in pattern recognition and image processing / J.C. Bezdek // IEEE Workshop Lang. Autom. – 1983. – P. 247-252.*
4. *Dubois D. Theorie des possibilities / D. Dubois, H. Prade. – Paris, 1988. – 283 p.*
5. *Раскин Л.Г. Нечеткая математика / Л.Г. Раскин, О.В. Серая. – Х.: Парус, 2008. – 352 с.*

Поступила в редколлегию 4.12.2009

Рецензент: д-р техн. наук, проф. Л.Г. Раскин, Национальный технический университет «ХПИ», Харьков.

НЕЧІТКА ЗАДАЧА КЛАСТЕРНОГО АНАЛІЗУ

О.В. Сіра

Розглянуто задачу кластеризації для випадку, коли координати центрів кластерів і об'єктів групування визначені нечіткими числами із заданими функціями приналежності. Запропонована проста обчислювальна процедура чіткого розподілу об'єктів, що легко реалізовується, по кластерах, заснована на виборі для кожного з об'єктів найбільш переважного кластера.

Ключові слова: *кластерний аналіз, нечіткі координати об'єктів, функції приналежності нечітких чисел, ступінь переваги одного нечіткого числа перед іншими, ітераційна процедура.*

FUZZY TASK OF CLUSTER ANALYSIS

O.V. Sira

The task of clusterization is considered for a case, when the coordinates of centers of clusters and objects of grouping are certain fuzzy variables with the set membership functions. Simple and easily realized calculable procedure of the clear distributing of objects is offered on clusters, based on a choice for each of objects of the most preferable cluster.

Keywords: *cluster analysis, unclear co-ordinates of objects, functions of belonging of unclear numbers, degree of preference of one unclear number before other, iterative procedure.*