УДК 519.7 : 004.8

Е.В. Бодянский, В.А. Опанасенко, А.Н. Слипченко

Харьковский национальный университет радиоэлектроники

НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ДАННЫХ, ЗАДАННЫХ В ПОРЯДКОВОЙ ШКАЛЕ

В статье предложен алгоритм нечеткой кластеризации многомерных наблюдений, компонентами которых являются лингвистические упорядоченные переменные, заданные своими рангами.

классификация данных, алгоритм нечеткой кластеризации многомерных наблюдений

Введение

Задача кластеризации многомерных наблюдений достаточно часто встречается во многих приложениях, связанных с интеллектуальным анализом данных (Data Mining). Традиционный подход к решению этой задачи предполагает, что каждое наблюдение может относиться только к одному кластеру, хотя более естественной представляется ситуация, когда обрабатываемый вектор признаков с различными уровнями вероятности (возможности, принадлежности, достоверности и т.п.) может принадлежать сразу нескольким классам. Данная ситуация является предметом рассмотрения нечеткого (фаззи) кластерного анализа, который к настоящему времени достаточно широко используется в различных приложениях [1 – 4].

Ситуация существенно усложняется, когда исходные данные заданы не в числовой, а в ранговой (порядковой) шкале, что часто встречается в социологии, медицине, образовании и т.п. В одномерном случае такая информация задается в виде последовательности упорядоченных лингвистических переменных

$$x^1, x^2, ..., x^j, ..., x^m, 1 < ... < j-1 < j < j+1 < ... < m,$$

где x^j – собственно лингвистическая переменная; j – соответствующий ранг.

Характерным примером является традиционная система оценок в образовании типа «плохо», «удовле-

творительно», «хорошо», «отлично». Заметим, что в своей повседневной деятельности человек гораздо чаще пользуется порядковой шкалой, нежели числовой.

Для решения задач кластеризации данных на порядковой шкале наиболее простым представляется подход, основанный на замене лингвистических переменных их рангами, однако в большинстве случаев этот прием оказывается некорректным, поскольку предполагает равенство расстояний между соседними числовыми рангами. Интуитивно ясно, что при оценке знаний учащихся, расстояние между «плохо» и «удовлетворительно» гораздо больше, чем расстояние между «удовлетворительно» и «хорошо». Немало аналогичных примеров легко найти и в медицине.

Более естественным представляется подход, основанный на фаззификации исходных данных и дальнейшем использовании методов нечеткой кластеризации. При этом исходный набор лингвистических переменных $x^1, x^2, ..., x^j, ..., x^m$ заменяется множеством функций принадлежности

$$\mu_1(x), \mu_2(x), ..., \mu_m(x),$$

заданных на интервале $[0,1]$. Такой прием был использован в [5], где с помощью "Fuzzy C-means" (FCM) алгоритма [1] производилась кластеризация не исходных данных, а параметров, описывающих соответствующие им функции принадлежности, хотя сам способ определения этих параметров не указан.

Более естественным представляется подход, развиваемый Р.К. Брауэром [6, 7] и основанный на анализе распределения частот появления конкретных значений лингвистических переменных. Ограничением этого подхода является предположение о гауссовом распределении исходных данных, что во многих приложениях не выполняется.

Целью настоящей работы является синтез алгоритма нечеткой кластеризации многомерных данных, описываемых векторами, чьи компоненты суть лингвистические переменные, заданные в порядковой шкале. Исходной информацией для решения задачи является выборка наблюдений, сформированная из N n-мерных векторов признаков

$$X = \{x(1), x(2), ..., x(k), ..., x(N)\}, \; k = 1, 2, ...N,$$

где $x(k) = \{x_i^j(k)\}$, $i = 1, 2, ...n$; $j = 1, 2, ...m$ – ранг конкретного значения лингвистической переменной по i-й координате n-мерного пространства для k-го объекта, подлежащего кластеризации.

Результатом работы алгоритма является разбиение исходного массива данных X на m классов (кластеров) с вычислением уровня принадлежности $w_j(k)$ k-го вектора признаков j-му кластеру.

1. Фаззификация исходных данных

Процесс фаззификации последовательности ранговых лингвистических переменных рассмотрим на примере одномерной выборки $x(1), x(2), ..., x(N)$, где каждому из наблюдений $x(k)$ может быть приписан один из рангов j, $j = 1, 2, ..., m$.

Пусть значение $x(k)$, соответствующее j-му рангу, встречается в выборке N_j раз. Тогда можно ввести в рассмотрение относительные частоты появления j-го ранга $f_j = \dfrac{N_j}{N}$ и кумулятивные частоты

$$F_1 = \frac{f_1}{2}; \; F_j = \frac{f_j}{2} + \sum_{l=1}^{j-1} f_l, \; j = 2, 3, ..., m,$$

при этом естественно выполняется условие

$$\sum_{i=1}^{m} f_i = 1.$$

На основе кумулятивных частот формируются центры набора функций принадлежности $\mu_j(x)$ так, как это показано на рис. 1, при этом для вычисления центров удобно воспользоваться рекуррентным соотношением

$$c_1 = 0,5f_1; \; c_j = c_{j-1} + 0,5(f_{j-1} + f_j), \; j = 2, 3, ..., m,$$

а сами функции принадлежности задать в форме

$$\mu_1(x) = 1, \; x \in [0, c_1],$$

$$\mu_j(x) = \begin{cases} \dfrac{x - c_{j-1}}{c_j - c_{j-1}}, \; x \in [c_{j-1}, c_j]; \\ \dfrac{c_{j+1} - x}{c_{j+1} - c_j}, \; x \in [c_j, c_{j+1}]; \; \mu_m(x) = 1, \; x \in [c_m, 1] \\ 0, \; x \notin [c_{j-1}, c_{j+1}]; \end{cases}$$

хотя, конечно, возможно использование функций иного вида с конечным носителем

$$\operatorname{supp} \mu_j(x) = [c_{j-1}, c_{j+1}].$$

Рассмотрим далее две соседние функции принадлежности $\mu_j(x)$ и $\mu_{j+1}(x)$ (рис. 2). Используя понятие α-разреза в виде $A_\alpha = \{x \in X : \mu(x) \ge \alpha\}$, можно ввести области влияния двух соседних рангов (на рисунке заштрихованы) в форме

$$\begin{cases} A_j^R = \left\{ \begin{array}{c} x \in [c_j, c_j + 0,5f_j] : \mu_j(x) \ge \alpha_j^R = \\ = 1 - 0,5\, f_j \big/ \left(c_{j+1} - c_j\right) \end{array} \right\}; \\ A_{j+1}^L = \left\{ \begin{array}{c} x \in [c_{j+1} - 0,5f_{j+1}, c_{j+1}] : \\ : \mu_{j+1}(x) \ge \alpha_{j+1}^L = 1 - 0,5\, f_j \big/ \left(c_{j+1} - c_j\right) \end{array} \right\}. \end{cases} \quad (1)$$

Такой способ задания функций принадлежности автоматически обеспечивает разбиение Руспини, т.е. выполнение условия

$$\sum_{j=1}^{m} \mu_j(x) = 1,$$

где R и L обозначают правую или левую стороны соседних функций принадлежности. При попадании некоторого наблюдения в область влияния конкретного ранга, можно говорить о «четкой» принадлежности к данному рангу.

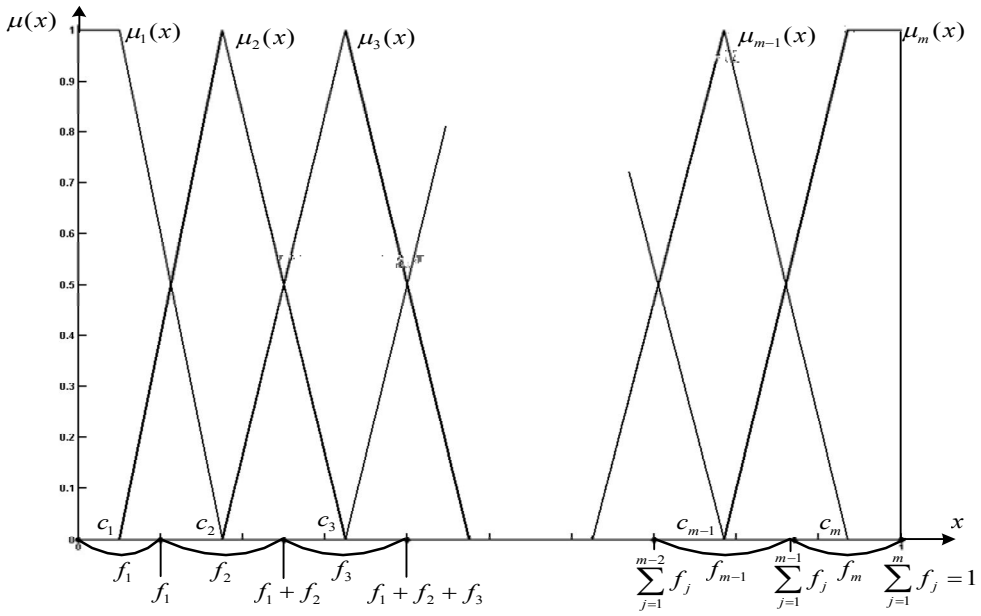

Рис. 1. Функции принадлежности для выборки ранговых переменных

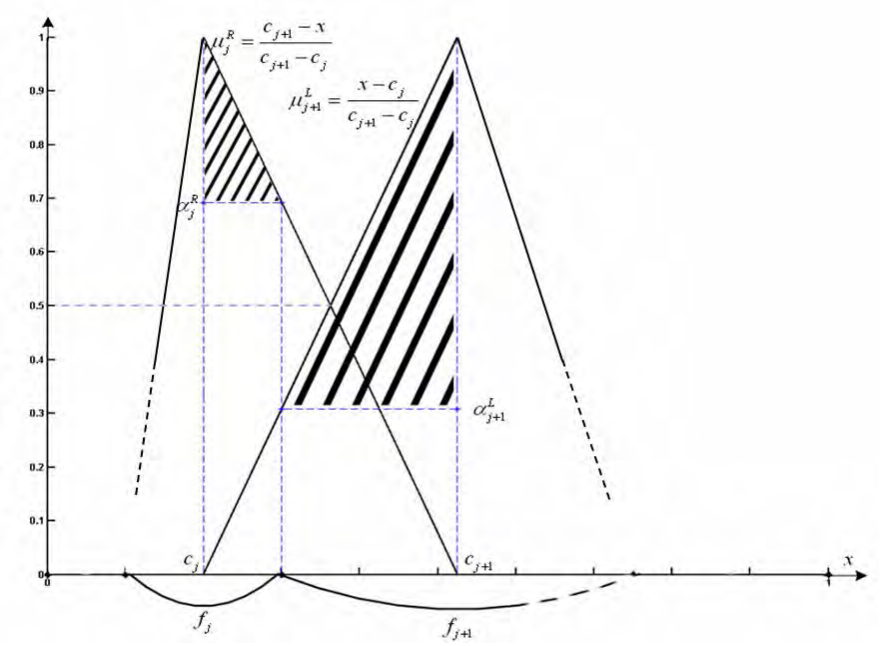

Рис. 2. Области влияния соседних рангов

2. Алгоритм кластеризации

Поскольку кластеризации подлежит выборка многомерных наблюдений-векторов, аналогично предыдущему необходимо провести фаззификацию по каждой из координат n-мерного пространства признаков. При этом формируется nm функций принадлежности с центрами c_{ij} так, как это показано на двумерном примере, приведенном на рис. 3. Здесь же приведен объект x(k), подлежащий кластеризации, с координатами

$$x(k) = \begin{pmatrix} x_1^2(k) = "\text{удовлетворительно}" \equiv a \\ x_2^4(k) = "\text{отлично}" \equiv e \end{pmatrix},$$

при этом, как видно из рисунка, срабатывают функции принадлежности $\mu_{12}(x_1), \mu_{13}(x_1), \mu_{14}(x_1), \mu_{22}(x_2), \mu_{23}(x_2), \mu_{24}(x_2)$ так, что немедленно принять решение о принадлежности x(k) к одному из классов «плохо», «удовлетворительно», «хорошо», «отлично», весьма затруднительно (рис. 3). Процесс нечеткой кластеризации ранговых переменных проведем на этом же примере, который иллюстрируется рис. 4. После формирования nm функций принадлежности (в примере $2 \times 4 = 8$) в рассмотрение вводится n-мерные векторы-центроиды кластеров

$$c_j = (c_{1j}, c_{2j}, ..., c_{nj})^T,\ j = 1, 2, ..., m$$

(в примере $c_1 = (c_{11}, c_{12})^T, c_2 = (c_{21}, c_{22})^T,$

$$c_3 = (c_{31}, c_{32})^T, c_4 = (c_{41}, c_{42})^T)$$

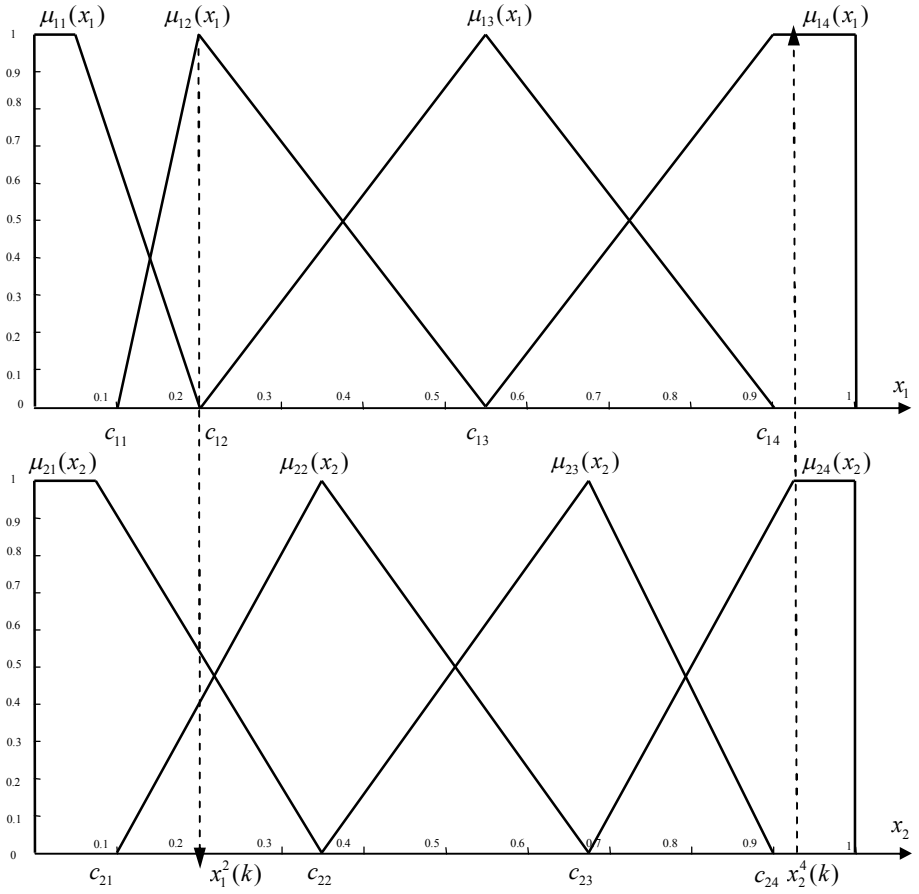

Рис. 3. Функции принадлежности двумерного пространства признаков

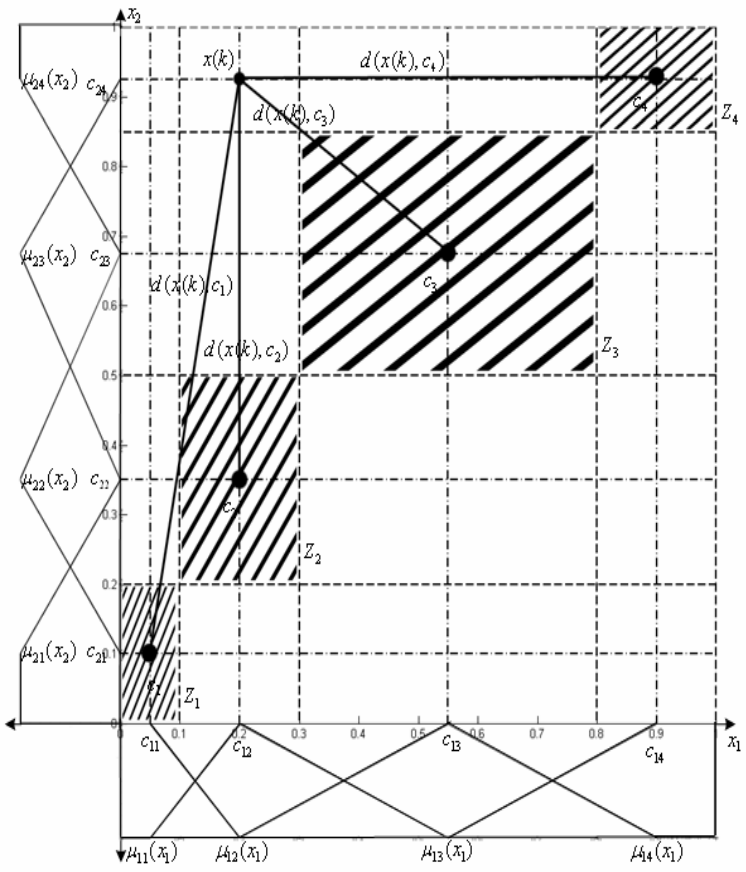

Рис. 4. Нечеткая кластеризация ранговых переменных

со своими областями влияния, описываемыми соотношениями (1) (в примере заштрихованные области), при попадании в которые, можно говорить о четкой принадлежности объекта x(k) конкретному кластеру. Здесь же приведен классифицируемый объект $x(k)=(e,a)^T$, который после фаззификации представлен в числовой форме с координатами c_{12} и c_{24}.

Далее вычисляются расстояния между x(k) и всеми центроидами c_j $d(x(k),c_j)=\| x(k)-c_j \|$ (в примере

$$d(x(k),c_1)=\| x(k)-c_1 \|;$$
$$d(x(k),c_2)=\| x(k)-c_2 \|,$$
$$d(x(k),c_3)=\| x(k)-c_3 \|;$$
$$d(x(k),c_4)=\| x(k)-c_4 \|),$$

после чего уровни принадлежности $w_j(k)$ вектора x(k) j-му кластеру можно определить согласно FCM-алгоритму [1] в виде

$$w_j(k)=\frac{\| x(k)-c_j \|^{-2}}{\sum_{l=1}^{m}\| x(k)-c_l \|^{-2}}=\frac{d^{-2}(x(k),c_j)}{\sum_{l=1}^{m}d^{-2}(x(k),c_l)}. \quad (2)$$

Недостатком оценки (2) является то, что в результате (кроме случаев, когда

x(k) попадает в область влияния одного из центроидов) объект «размазывается» по всем существующим кластерам, что в ранговой шкале ведет к потере физического смысла.

Так рассматриваемый объект $x(k)=(e,a)^T$ с ненулевым уровнем принадлежности может относиться и к кластеру «плохо», что, конечно же, бессмысленно. В связи с этим представляется целесообразным после вычисления всех расстояний $d(x(k),c_j)$ провести их ранжирование по возрастанию и выбрать наименьшее $d_{min\,min}(x(k),c_j)$ и следующее за ним $d_{min}(x(k),c_l)$. Далее можно воспользоваться формулой (2) с той разницей, что в расчет принимаются только два наименьших расстояния. В результате $x(k)$ будет принадлежать двум соседним кластерам с центроидами c_j и c_{j+1} (или c_{j-1}) с некоторыми уровнями принадлежности $w_j(k)$ и $w_{j+1}(k)$ (или $w_{j-1}(k)$).

Таким образом, алгоритм нечеткой кластеризации многомерных наблюдений, заданных на порядковой шкале, реализуется в виде последовательности следующих этапов.

1. Вычисление относительных f_j и накопленных F_j частот по выборке

$$x(1),x(2),...,x(k),...,x(N).$$

2. Фаззификация исходной выборки лингвистических переменных построением nm функций принадлежности $\mu_{ij}(x_i),\ j=\overline{1,m};\ i=\overline{1,n}$ и m векторов-центроидов $c_j=(c_{1j},...,c_{nj})^T$ формируемых кластеров.

3. Построение областей влияния Z_j центроидов c_j в виде ортотопа с ребрами $c_{ij}\pm 0,5f_{ij}$.

4. Проверка возможности четкой кластеризации в виде: если $x(k)\in Z_j$, то данное наблюдение однозначно классифицируется, т.е. $w_j(k)=1$ и $w_l(k)=0$ для всех остальных $l\neq k$.

5. Если предыдущее условие не выполняется, производится расчет всех расстояний

$$d(x(k),c_j)=\|x(k)-c_j\|.$$

6. Выделение двух наименьших расстояний $d_{min\,min}(x(k),c_j)$ и $d_{min}(x(k),c_l)$, где l может принимать значение или $j-1$, или $j+1$.

7. Расчет уровней принадлежности $x(k)$ к двум соседним кластерам в виде

$$w_j(k)=\frac{d^{-2}_{min\,min}(x(k),c_j)}{d^{-2}_{min\,min}(x(k),c_j)+d^{-2}_{min}(x(k),c_l)};$$

$$w_l(k)=\frac{d^{-2}_{min}(x(k),c_l)}{d^{-2}_{min\,min}(x(k),c_j)+d^{-2}_{min}(x(k),c_l)}.$$

3. Результаты экспериментальных исследований

Для проверки работоспособности предложенного алгоритма были взяты данные об успеваемости студентов потока на одном из факультетов Харьковского национального университета радиоэлектроники. Набор данных содержит оценки по шести предметам для 103 человек.

Статистический анализ показал, что для каждой из переменных (дисциплин) гипотеза о том, что оценки имеют нормальный закон распределения, не подтверждается. Таким образом, нельзя использовать подход, предложенный Р.К. Брауэром.

В результате работы алгоритма были определены центроиды для каждого из рангов (оценок) по каждой из переменных. Далее с помощью описанного выше алгоритма классификации были определены степени принадлежности каждому из классов.

Заключение

В статье предложен алгоритм нечеткой кластеризации многомерных наблюдений, компонентами которых являются лингвистические упорядоченные переменные, заданные своими рангами. В основе алгоритма лежит фаззификация исходной выборки и дальнейшее использование введенной модификации «Fuzzy C-means» - процедуры. Алгоритм прост в вычислительном отношении и позволяет ясно интерпретировать результаты расчетов.

Список литературы

1. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. – N.Y.:Plenum Press, 1981. – 272 p.

2. Jang J.-Sh. R., Sun Ch.-T., Mizutani E. Neuro-Fuzzy and Soft Computing. – Upper Saddle River, NJ: Prentice Hall, 1997. – 614 p.

3. Hoeppner F., Klawonn F., Kruse R. Fuzzy Cluster-analyse. – Braunschweig:Vieweg, 1999. – 280 S.

4. Nelles O. Nonlinear System Identification. – Berlin: Springer, 2001. – 785 p.

5. Butkiewicz B.S. Robust fuzzy clustering with fuzzy data // Lecture News in Computer Science. – V.3528. – Berlin – Heidelberg: Springer –Verlag, 2005. – P. 76-82.

6. Brouwer R.K., Pedrycz W. A feed forward neural network for mapping vectors to fuzzy sets of vectors // Proc.Int.Conf. on Artificial Neural Networks and Neural Information Processing ICANN/ICOMIP 2003. – Istanbul, Turkey, 2003. – P. 45-48.

7. Brouwer R.K. Fuzzy set covering of a set of ordinal attributes without parameter sharing // Fuzzy Sets and Systems. – 2006. – 157. – № 13. – P. 1775-1786.

Поступила в редколлегию 2.04.2007

Рецензент: д-р техн. наук, проф. В.А. Филатов, Харьковский национальный университет радиоэлектроники, Харьков.