УДК 004.8:004.032.26

В.А. Самитова

Харьковский национальный университет радиоэлектроники, Харьков

ОТОБРАЖЕНИЕ ПОРЯДКОВЫХ ХАРАКТЕРИСТИК В ЦИФРОВУЮ ШКАЛУ НА ОСНОВЕ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ

Рассмотрена задача кластеризациии данных, заданных в порядковой шкале, в условиях реального времени. Для классификации предложено использовать метод рекуррентной нечеткой кластеризации, основанный на замене данных их порядково цифровым отображением.

Ключевые слова: *кластеризация, порядковая шкала, FCM, рекуррентная кластеризация, порядково-цифровое отображение.*

Введение

Задача кластеризации многомерных наблюдений является неотъемлемой частью интеллектуального анализа данных (Data Mining), при этом в стандартной ее постановке предполагается, что каждый вектор-образ из выборки наблюдений может принадлежать только одному кластеру. Более естественной представляется ситуация, когда обрабатываемый вектор признаков с различными уровнями принадлежности может относиться сразу к нескольким классам. Данная ситуация рассматривается в рамках нечеткого кластерного анализа, широко используемого в настоящее время во множестве приложений, связанных с медициной, биологией, экономикой, социологией, образованием, видеообработкой и т.п.

Определение принадлежности объекта к тому или иному кластеру, прежде всего, связано с вычислением схожести (similarity) между объектами. Схожесть между двумя наблюдениями определяется суммированием (агрегацией) схожести между всеми значениями характеристик этих наблюдений.

© В.А. Самитова

Данные в выборке могут быть представлены в различных шкалах: числовой, порядковой, номинальной. В случае данных, представленных в числовой шкале, схожесть между объектами определяется расстоянием между двумя значениями соответствующих характеристик. Схожесть между объектами в номинальной шкале может быть выражена как 0 или 1, в зависимости от того, совпадают ли значения характеристик наблюдений или нет.

В случае порядковых шкал в существующих методах кластеризации данных таких, как k-средних [1, 2], «Fuzzy C-means» [3, 4], EM- алгоритм [5, 6], чаще всего определение схожести основано на замене лингвистических переменных их рангами. Однако в большинстве случаев данных подход оказывается некорректным, так как предполагает равенство расстояний между соседними числовыми рангами, что не всегда соответствует действительности. Более естественным представляется подход, развиваемый Р.К. Брауэром [7] и основанный на максимизации функции правдоподобия. В [8-10] было предложено осуществлять фаззификацию исходных данных на основе анализа распределения частот появления конкретных лингвистических переменных, предполагалось, что эти распределения подчиняются закону Гаусса. Ограничением этого подхода является предположение о гауссовом распределении исходных данных, что во многих приложениях не выполняется, а также способ вычисления правдоподобия для порядковых переменных. *В данной статье предлагается метод, в основе которого лежит замена порядковых характеристик их порядково-цифровым отображением.*

Результаты исследований

1. Правдоподобие и вероятность. Существует несколько основных подходов к кластеризации данных – иерархический, метрический, итерационный и т.п. [14]. Итерационная кластеризация применяется во многих областях, при этом алгоритм в цикле находит лучшие кластеры, к которым могут принадлежать наблюдения. Исходной информацией для решения задачи является выборка наблюдений, сформированная из N n-мерных векторов признаков $X = \{x_1, x_2, ..., x_j, ..., x_N\}$, где $j = 1,...,N$. На основе X рассчитывается матрица нечеткого разбиения $W = \left[w_{i,j} \right]_{c \times N}$, удовлетворяющая условиям:

$$\sum_{i=1}^{c} w_{i,j} = 1;\ 0 < w_{i,j} < 1, \quad \forall i \in \overline{1,c},\ j \in \overline{1,N}, \qquad (1)$$

где c – количество кластеров; $w_{i,j}$ – степень принадлежности j-го объекта i-му кластеру.

Предположим, что каждый объект в X имеет одинаковый тип свойств, и одно из этих свойств представлено в порядковой шкале. Пусть $L = \{l_1, ..., l_m\}$ – множество возможных значений данной порядковой характеристики, удовлетворяющее свойству $l_1 < l_2 < ... < l_m$. Для каждого значения l_s пусть существует подмножество объектов $X_s \subseteq X$, включающее в себя l_s. *Подобие (similarity)* между двумя значениями l_s и l_t, отражающее матрицу W, определяется как среднее *подобий* между объектами в X_s и X_t соответственно [11].

$$\begin{aligned} &\mathrm{sim}(W, l_s, l_t) \triangleq \mathrm{sim}(W, X_s, X_t), \\ &\forall s = 1...m; \quad t = 1...m, \quad s \neq t, \end{aligned} \qquad (2)$$

где *подобие* между двумя непересекающимися подмножествами X определяется следующим образом:

$$\forall A, B \subseteq X \ э\ A \cap B \neq \varnothing,$$

$$\mathrm{sim}(W, A, B) \triangleq \sum_{x \in A; y \in B} \mathrm{sim}(W, x, y) \Big/ \left(|A||B| \right) \qquad (3)$$

где $\mathrm{sim}(W, x, y)$ – *подобие* между $x \in X$ и $y \in X$.

Подобие между двумя объектами x и y в X, отражающее W, определено *контекстно-ориентированной близостью (context-based proximity)* между x и y

$$\mathrm{sim}(W, x, y) \triangleq \mathrm{prox}(W, x, y), \qquad (4)$$

а контекстно-ориентированная близость между x_j и x_k в отношении W, которая используется в (4), определяется следующим образом:

$$\mathrm{prox}(U, x_j, x_k) \triangleq \sum_{i=1}^{c} \min(w_{i,j}, w_{i,k}) \qquad (5)$$

Рассмотрим три последовательных значения l_{r-1}, l_r и l_{r+1}. При $\mathrm{sim}(W, l_{r-1}, l_r) > \mathrm{sim}(W, l_r, l_{r+1})$ можно в общем сказать, что l_r ближе к l_{r-1}, чем к l_{r+1} для всех $r = 2,...,m-1$ в данном множестве объектов X. Введем для удобства два дополнительных значения характеристики l_0 и l_{m+1} такие, что $l_0 < l_1$, а $l_m < l_{m+1}$. Теперь *порядково-цифровое отображение* g для порядковых значений характеристики в данном множестве объектов X определяется следующим образом:

$$\begin{aligned} &g(l_0) = 0, \\ &g(l_p) = \sum_{s=1}^{t} \frac{1 - \mathrm{sim}(W, l_{s-1}, l_s)}{\sum_{r=1}^{m+1} (1 - \mathrm{sim}(W, l_{r-1}, l_r))}, \forall p = 1...m \\ &g(l_{m+1}) = 1, \end{aligned} \qquad (6)$$

$$\begin{aligned} &\mathrm{sim}(W, l_0, l_1) \triangleq \mathrm{sim}(W, X_1, X - X_1), \\ \text{где}\ \ &\mathrm{sim}(W, l_{p-1}, l_p) \triangleq \mathrm{sim}(W, X_s, X_t), \forall p = 2...m, \\ &\mathrm{sim}(W, l_m, l_{m+1}) \triangleq \mathrm{sim}(W, X_m, X - X_m). \end{aligned}$$

2. Метод пакетной фаззи-кластеризации. Алгоритмы, основанные на оптимизации целевой функции, предполагают решение проблемы кластеризации путем оптимизации специального критерия кластеризации и являются наиболее обоснованными

с математической точки зрения. Исходной информацией для решения задачи является выборка наблюдений $X = \{x_1, x_2, ..., x_j, ..., x_N\}$, где $j = 1,...,N$, $x_j = \{x_j^l\}, l = 1,...,m$ – ранг конкретного значения лингвистической переменной для j-го объекта, подлежащего кластеризации.

Результатом работы алгоритма является разбиение исходного массива данных X на с непересекающихся классов (кластеров) с вычислением уровня принадлежности $w_{i,j}$ j-го вектора признаков i-му кластеру. Предлагаемый алгоритм имеет достаточно близкую алгоритмическую структуру к алгоритму «Fuzzy C-means» (FCM) [3]. Задача кластеризации для количественных характеристик решается путем минимизации целевой функции:

$$Q = \sum_{i=1}^{c}\sum_{j=1}^{N} w_{i,j}^{\beta} \left\| x_j - v_i \right\|^2 \qquad (7)$$

при ограничениях:

$$w_{i,j} \geq 0, \forall i = 1,...,c; \forall j = 1,...,N,$$
$$\sum_{i=1}^{c} w_{i,j} = 1, \forall j = 1,...,N, \sum_{j=1}^{N} w_{i,j} > 0, \forall i \in \overline{1,c}, \qquad (8)$$

где $w_{i,j}$ - уровень принадлежности j-го наблюдения к i-му кластеру, β - неотрицательный параметр фаззификации.

Вводя в рассмотрение функцию Лагранжа

$$L_S\left(w_{ij}, c_i, \lambda_j\right) =$$
$$= \sum_{j=1}^{N}\sum_{i=1}^{c} w_{ij}^{\beta} d^2(x_j, c_i) + \sum_{j=1}^{N} \lambda_j \left(\sum_{i=1}^{c} w_{ij} - 1 \right) = \qquad (9)$$
$$= \sum_{j=1}^{N} (\sum_{i=1}^{c} w_{ij}^{\beta} d^2(x_j, c_i) + \lambda_j (\sum_{i=1}^{c} w_{ij} - 1)),$$

(здесь λ_j - неопределенные множители Лагранжа) и решая систему уравнений Каруша- Куна- Таккера, приходим к решению:

$$\begin{cases} \dfrac{\partial L(w_{ij}, c_i, \lambda_j)}{\partial w_{ij}} = 0, \\ \nabla_{c_i} L(w_{ij}, c_i, \lambda_j) = 0, \\ \dfrac{\partial L(w_{ij}, c_i, \lambda_j)}{\partial \lambda_j} = 0, \end{cases} \qquad (10)$$

где искомые переменные могут быть получены следующим образом:

$$w_{ij} = \frac{(d^2(x_j, c_i))^{1/(1-\beta)}}{\sum_{t=1}^{c} (d^2(x_j, c_t))^{1/(1-\beta)}}, \forall t \in \overline{1,c}; \forall j \in \overline{1,N}, \qquad (11)$$

$$c_i = \sum_{j=1}^{N} w_{ij}^{\beta} x_j \Big/ \sum_{j=1}^{N} w_{ij}^{\beta}, \qquad (12)$$

$$\lambda_j = -\left(\sum_{i=1}^{c} (\beta d^2(x_j, c_i))^{1/(1-\beta)} \right)^{1-\beta}. \qquad (13)$$

3. Рекуррентная нечеткая кластеризация. В ряде задач, таких как обработка речи, web mining, медицинская диагностика, обработка сигналов датчиков в робототехнике и т.п. часто необходима обработка данных в реальном времени. В связи с чем целесообразным является использование рекурсивных процедур кластеризации.

Анализируя уравнение (11), можно заметить, что задача поиска седловой точки Лангранжиана может быть сведена к решению последовательности задач поиска седловой точки локальных модификаций функции Лагранжа.

В связи с этим мы будем использовать для расчета уровня принадлежности локальную модификацию Лангранжиана:

$$L_S\left(w_{ij}, c_i, \lambda_j\right) = \sum_{i=1}^{c} w_{ij}^{\beta} d^2(x_j, c_i) + \lambda_j (\sum_{i=1}^{c} w_{ij} - 1). \qquad (14)$$

Используя процедуру оптимизации Эрроу-Гурвица-Удзавы получаем алгоритм вида:

$$w_{ij} = (d^2(x_j, c_{ij}))^{1/(1-\beta)} \Big/ \sum_{t=1}^{c} (d^2(x_j, c_{tj}))^{1/(1-\beta)}, \qquad (15)$$

$$c_{i,j+1} = c_{ij} - \eta_j \nabla_{c_i} L_j(w_{ij}, c_{ij}, \lambda_j) = c_{ij} - \eta_j w_{ij}^{\beta} \times$$
$$\times d(x_{j+1}, c_{ij}) \nabla_{c_i} d(x_{j+1}, c_{ij}), \qquad (16)$$

где η_j - параметр шага обучения, c_{ij} - центроид i-го кластера, рассчитанного для выборки данных из j наблюдений.

Процедура (15), (16) по структуре достаточно близка к алгоритму нечеткого конкурентного обучения Чанга-Ли [12], а когда параметр фаззификации $\beta = 2$ – к градиентному алгоритму нечеткой кластеризации Парка-Дэггера [13]:

$$w_{ij} = \| x_j - c_{ij} \| \Big/ \sum_{t=1}^{c} \| x_j - c_{tj} \|^{-2}, \qquad (17)$$

$$c_{i,j+1} = c_{ij} + \eta_j w_{ij}^2 (x_{j+1} - c_{ij}). \qquad (18)$$

Используя порядково-цифровое отображение с помощью рекуррентной нечеткой кластеризации, получаем алгоритм кластеризации порядковых данных, состоящий из следующих шагов:

Инициализация:

1. Инициализируем все порядково-цифровые отображения равномерными интервалами.

2. Трансформируем данное множество объектов $X = \{x_1, ..., x_N\}$ в X^*, заменяя все порядковые характеристики порядково-цифровыми отображениями.

3. Инициализируем центроиды кластеров $c_i, \forall i = 1...c$ случайными значениями.

Повтор:

1. Вычисляются функции принадлежности $w_{ij}, \forall i = 1...c; j = 1...N$ с помощью формулы (17).

2. Вычисляются центроиды кластеров $c_i, \forall i = 1...c$ с помощью формулы (18).

3. Для каждого порядкового значения вычисляется порядково-цифровое отображение с помощью формулы (6).

4. Трансформируем данное множество объектов X в X^* с новыми порядково-цифровыми отображениями.

Вычисления производятся итерационно, пока не будет выполнено условие остановки алгоритма.

4. Численное моделирование. Для проверки работоспособности предложенного алгоритма были использованы известные выборки данных из UCI репозитория Wine data и Iris data. Поскольку данные выборки не содержат данных в порядковой шкале, одна из характеристик была искусственно переведена к нужному формату. Проводилось сравнение предложенного алгоритма со следующими методами: FCM и рекуррентным алгоритмом кластеризации Парка-Дэггера.

В результате работы алгоритмов были получены результаты, представленные в табл. 1.

Таблица 1

Уровень ошибок алгоритмов

Выборки	FCM	Park-Dagher	OMR FCM*
Iris	7,4%	6,9%	7,2%
Wine	3,8%	3,9%	4%

***** *OMR – Ordinal Mapping Recursive*

Выводы

Рассмотрен алгоритм рекуррентной нечеткой кластеризации, основанный на порядково-цифровом отображении для обработки многомерных наблюдений, заданных в порядковой шкале. В основе подхода лежит отображение лингвистических переменных в числовую шкалу и модификация известного метода нечетких с-средних. Рассмотренный алгоритм позволяет работать с данными в порядковой шкале в режиме реального времени и прост в численной реализации.

Список литературы

1. *MacQueen Z.B. Some Methods of Classification and Analysis of Multivariate Observations / Z.B. MacQueen // Berkely Symposium on Mathematical Statistics and Probability. – 1967.*
2. *Lloyd S.P. Least Squares Quantization in PCM / S.P. Lloyd // IEEE Transactions on Information Theory. – 1982. – Vol. IT-28. – P. 129-137.*
3. *Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – N.Y.:Plenum Press, 1981. – 272 p.*
4. *Jang J.-Sh. R., Sun Ch.-T., Mizutani E., Neuro-Fuzzy and Soft Computing. – Upper Saddle River, NJ: Prentice Hall, 1997. - 614 p.*
5. *Dempster A.P. Maximum-Likelihood from Incomplete Data via the EM Algorithm / A.P. Dempster, N.M. Laird // Journal of the Royal Statistical Society. – 1977. – Vol. B. – P. 1-38.*
6. *Zhong S. A Unified Framework for Model-based Clustering / S. Zhong, J. Ghosh // Journal of Machine Learning Research. – 2003. – vol. 4. – P. 1001-1037.*
7. *Mahnhoon L., Brouwer R.K., Likelihood based fuzzy clustering for data sets of mixed features / L. Mahnhoon, R.K. Brouwer // IEEE Symp. on Foundations of Comput. Intell. FOCI 2007. – 2007. – P. 544-549.*
8. *Brouwer R.K. A feedforward neural network for mapping vectors to fuzzy sets of vectors / R.K. Brouwer, W. Pedrycz // Proc. Int. Conf. on Artificial Neural Networks and Neural Information Processing ICANN/ICOMIP 2003. – Istanbul, Turkey, 2003. – P.45-48.*
9. *Butkiewicz B.S. Robust fuzzy clustering with fuzzy data / B.S. Butkiewicz // Lecture Notes in Computer Science. – V.3528. – Berlin – Heidelberg: Springer –Verlag, 2005. – P. 76-82.*
10. *Brouwer R.K. Fuzzy set covering of a set of ordinal attributes without parameter sharing / R.K. Brouwer // Fuzzy Sets and Systems. – 2006. – 157, № 13. – P.1775 – 1786.*
11. *Mahnhoon L. Mapping of Ordinal Feature Values to Numerical Values through Fuzzy Clustering / L. Mahnhoon // IEEE Fuzzy Systems. – 2008. – P.732-737.*
12. *Chung F.L. Fuzzy competitive learning. Neural Networks / F.L. Chung, T. Lee. – 1994. – P. 539-552.*
13. *Park D.C. Gradient based fuzzy c-means (GBFCM) algorithm / D.C. Park, I. Dagher // IEEE Int. Conf. on Neural Networks. – 1984. – P. 16.26-16.31.*
14. *Hoeppner F. Fuzzy-Clusteranalysen / F. Hoeppner, F. Klawonn, R. Kruse. – Braunschweig:Vieweg, 1997. – 280S.*

Поступила в редколлегию 16.04.2015

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

ВІДОБРАЖЕННЯ ПОРЯДКОВИХ ХАРАКТЕРИСТИК В ЧИСЕЛЬНУ ШКАЛУ НА ОСНОВІ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ

В.О. Самітова

Розглядається задача кластеризації даних, що задані в порядковій шкалі, в умовах реального часу. Для класифікації запропоновано метод рекурентної нечіткої кластеризації даних, що базується на заміні спостережень їх порядково-цифровим відображенням.

Ключові слова: *кластеризація, порядкова шкала, FCM, рекурентна кластеризація, порядково-цифрове відображення.*

FUZZY CLUSTERIZATION OF DATA IN ORDINAL SCALE BASED ON MAPPING ORDINAL FEATURE VALUES INTO NUMERICAL VALUES

V.A. Samitova

Fuzzy clusterization of data in ordinal scale based on mapping ordinal feature values into numerical values in real time is considered.

Keywords: *clusterization, index scale, FCM, recurrent clusterization, index-digital reflection.*