УДК 004.032.26

Е.В.Бодянский[1], В.А.Самитова[2]
[1-2]ХНУРЭ, г. Харьков, Украина

ВОЗМОЖНОСТНАЯ НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ МАССИВОВ КАТЕГОРИАЛЬНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ЧАСТОТНЫХ ПРОТОТИПОВ И МЕР НЕСХОДСТВА

В статье рассмотрены алгоритмы нечеткой кластеризации данных, представленных в категориальной шкале. Классические алгоритмы кластеризации сталкиваются с определенными трудностями в процессе обработки подобных наблюдений, поскольку в таких данных отсутствует понятие расстояния. Предложен метод возможностной нечеткой кластеризации категориальных данных на основе частотных прототипов и мер несходства.

КАТЕГОРИАЛЬНЫЕ ДАННЫЕ, КАТЕГОРИАЛЬНАЯ ШКАЛА, ВОЗМОЖНОСТНАЯ НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ, ЧАСТОТНЫЕ ПРОТОТИПЫ, МЕРЫ НЕСХОДСТВА

Введение

Задача кластеризации массивов многомерных данных часто встречается во многих приложениях, связанных с интеллектуальным анализом данных (DataMining), а для ее решения может быть использовано множество различных подходов, методов, алгоритмов [1-7]. Суть этой задачи состоит в том, что исходный массив данных, описываемый многомерными векторами-образами, в режиме самообучения должен быть разбит на однородные принятом смысле группы-кластеры. Традиционный подход к решению задачи кластеризации опирается на предположение, что каждый вектор-образ может принадлежать только одному классу, а это означает, что формируемые в многомерном пространстве признаков кластеры не пересекаются. Исходной информацией для данной задачи является выборка наблюдений, сформированная из N n-мерных векторов-признаков

$$X = \{x(1), x(2), ..., x(k), ..., x(N)\} \subset R^n,$$

заданных, как правило, либо в интервальной шкале, либо в шкале отношений, при этом $x(k) = (x_1(k), ..., x_j(k), ..., x_n(k))^T$, а между двумя векторами $x(k)$, $x(q)$ может быть рассчитано расстояние в некоторой метрике, как правило, евклидовой. Результатом кластеризации является разбиение исходного массива наблюдений на r непересекающихся классов, при этом предполагается, что как значения r, N, так и параметры собственно процедуры кластеризации заданы априорно и не меняются в процессе обработки информации.

Более сложная ситуация возникает в случае, когда кластеры взаимно пересекаются, что приводит к тому, что любое наблюдение может принадлежать сразу нескольким кластерам. Эта ситуация является предметом рассмотрения нечеткого (fuzzy) кластерного анализа [8-10] и многие используемые здесь методы являются обобщением «четких» процедур на «нечеткий» случай. Результатом нечеткой кластеризации также является разбиение исходного массива данных на r пересекающихся кластеров, однако при этом дополнительно рассматриваются уровни принадлежности $u_i(k)$ k-го вектора признаков i-му кластеру, $i = 1, 2, ..., r$.

При этом полученные результаты существенным образом зависят от параметра, именуемого «фаззификатором», который задает уровень «размытости» границ между нечеткими кластерами.

Во многих практических задачах, возникающих в WebMining, TextMining, MedicalDataMining и т.п., достаточно часто возникает ситуация, когда признаки $x_j(k)$ заданы не в числовой, а в категориальной (номинальной) шкале, при этом каждый такой признак может принимать конечное значение «имен» $x_j^l(k)$, где $j = 1, 2, ..., n$; $l = 1, 2, ..., m_j$; $k = 1, 2, ..., N$. Понятно, что в этой ситуации традиционные методы не работают в силу отсутствия самого понятия «расстояние» в категориальной шкале. В принципе, данные описываемые в номинальной шкале, без особых проблем могут быть трансформированы в бинарную шкалу, однако при этом резко возрастает размерность пространства признаков, что существенно усложняет решение задачи из-за возникновения эффектов «проклятья размерности», а в нечетком случае – «концентрации норм».

В связи с этим в [11-13] предлагается вместо традиционного евклидова расстояния, лежащего в основе классического метода k-средних, использовать «несходство» (dissimilarity) между векторами-образами, а вместо стандартных средних – моды отдельных признаков.

При этом несходство между двумя векторами $x(k)$ и $x(q)$ может быть описано с помощью выражения:

$$d(x(k), x(q)) = \sum_{j=1}^{n} \delta(x_j(k), x_j(q)), \quad (1)$$

где $\delta(x_j(k), x_j(q)) = \begin{cases} 0, if \ \ x_j(k) = x_j(q), \\ 1, if \ \ x_j(k) \neq x_j(q), \end{cases}$

при этом, если $x(k) = x(q)$, то $d(x(k), x(q)) = 0$, а при полном несовпадении компонент этих векторов $d(x(k), x(q)) = n$, т.е. $0 \le d(x(k), x(q)) \le n$.

В этом случае в качестве прототипов-центроидов кластеров используются наиболее часто встречающиеся в данном кластере значения компонент наблюдений – моды.

И хотя метод k -мод, являясь «ближайшим родственником» k -средних, нагляден и прост в численной реализации, его использование ограничивается тем фактом, что мода каждого из кластеров не единственна, что не позволяет получить устойчивое решение.

1. Модифицированный метод k -мод

Для преодоления отмеченного недостатка в [13] в качестве прототипов кластеров категориальных данных предложено использовать не обычные моды, а, так называемые, «представители» (representatives), учитывающие и значения частот появления отдельных значений признаков.

Пусть i -й кластер содержит N_i наблюдений $x(k)$ так, что $Cl_i = \{x(1), x(2), ..., x(N_i)\} \subset R^n$, $\sum_{i=1}^{r} N_i = N$.

При этом вектор-прототип этого кластера может быть представлен в виде $c_i = (c_{i1}, c_{i2}, ..., c_{in})^T$, а для каждой компоненты c_{ij} может быть рассчитана частота появления соответствующего значения признака в кластере в виде

$$f_{ij} = \frac{N_{ij}}{N_i}, \tag{2}$$

где N_{ij} - число появлений признака x_j в Cl_i.

В связи с тем, что каждый признак x_j может принимать только конечное число значений $x_j^l, l = 1, 2, ..., m_j$, выражение (2) можно так же переписать в виде $f_{ij}^l = \frac{N_{ij}^l}{N_i}$.

Тогда в качестве меры несходства между прототипом c_i и наблюдением $x(k)$ вместо (1) используется оценка

$$d(c_{ij}, x(k)) = \sum_{j=1}^{n} \sum_{l_j=1}^{m_j} f_{ij}^l \delta(c_{ij}, x_j(k)). \tag{3}$$

Понятно, что (3) также лежит в интервале $0 \le d(c_i, x(k)) \le n$.

Авторами [13] показано, что использование меры несходства (3) позволяет максимально приблизить задачу кластеризации категориальных данных к нахождению стандартных k -средних путем минимизации целевой функции

$$E(u_i(k), c_i) = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i(k) d(c_i, x(k)), \tag{4}$$

$$\sum_{i=1}^{r} u_i(k) = 1, \; u_i(k) \in \{0, 1\},$$

при этом если $x(k)$ принадлежит Cl_i, то $u_i(k) = 1$ и $u_i(k) = 0$ в противоположном случае.

Собственно, процесс кластеризации реализуется в форме последовательности следующих шагов.

1. Некоторым достаточно произвольным образом задаются r начальных прототипов c_i, $i = 1, 2, ..., r$.

2. Приписать наблюдение $x(k)$ к Cl_i, если $d(c_i, x(k)) < d(c_t, x(t)), \forall t = 1, 2, ..., r; t \ne i$.

3. Рассчитать моды-прототипы для всех кластеров Cl_i и соответствующие им частоты f_{ij}^l.

4. Рассчитать N_r оценок несходства новых прототипов со всеми $x(k)$.

5. Продолжать до тех пор, пока не стабилизируются прототипы.

Дополнительно к мере несходства (3) можно ввести в рассмотрение оценку «сходства» (similarity) в виде

$$0 \le sim(c_i, x(k)) = 1 - \frac{d(c_i, x(k))}{n} \le 1. \tag{5}$$

Это значение может служить простейшей оценкой уровня нечеткой принадлежности в случае возможного перекрытия формируемых кластеров, т.е.

$$sim(c_i, x(k)) = u_i(k).$$

2. Нечеткая кластеризация категориальных данных

В теории и практике нечеткой кластеризации количественных переменных наибольшее распространение получил метод нечетких c - средних (FCM) Дж. Бездека [8], основанный на минимизации целевой функции:

$$E(u_i(k), c_i) = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) \|x(k) - c_i\|^2 \tag{6}$$

при ограничениях

$$\sum_{i=1}^{r} u_i(k) = 1, \; 0 \le \sum_{k=1}^{N} u_i(k) \le N, \; u_i(k) \in [0, 1], \tag{7}$$

где β - неотрицательный параметр фаззификации (fuzzifier).

Минимизация (6) при ограничениях (7) с помощью стандартной техники нелинейного программирования приводит к известному результату:

$$\begin{cases} c_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}, \\ u_i(k) = \dfrac{(\|x(k)-c_i\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{t=1}^{r}(\|x(k)-c_t\|^2)^{\frac{1}{1-\beta}}}. \end{cases} \quad (8)$$

В [14-16] были введены модификации стандартного FCM, позволяющие обрабатывать векторы наблюдений, образованные категориальными переменными. Так, в [16] показано, что использование меры несходства (3) приводит к оценке уровня принадлежности наблюдений $x(k)$ кластеру Cl_i вида:

$$u_i(k) = \frac{d^{\frac{1}{1-\beta}}(c_i, x(k))}{\sum\limits_{t=1}^{r} d^{\frac{1}{1-\beta}}(c_t, x(k))}, \quad (9)$$

по сути совпадающей со вторым соотношением (8). Для расчета же мод-прототипов вектор $x(k)$ приписывается к кластеру Cl_i, для которого

$$u_i(k) > u_t(k), \forall t = 1,2,...,c; t \neq i. \quad (10)$$

Таким образом, процесс нечеткой кластеризации реализуется аналогично предыдущему в форме последовательности следующих шагов.

1. Некоторым достаточно произвольным образом задаются r начальных прототипов c_i, $i = 1,2,...,r$.

2. Рассчитать N_r оценок несходства (3) для каждого Cl_i и каждого $x(k)$.

3. Рассчитать уровни принадлежности каждого $x(k)$ каждому Cl_i согласно выражению (9).

4. Приписать наблюдение $x(k)$ к Cl_i в соответствии с условием (10).

5. Рассчитать моды-прототипы для всех кластеров Cl_i и соответствующие им частоты f_{ij}^l.

6. Рассчитать N_r оценок несходства новых прототипов со всеми $x(k)$.

7. Продолжать до тех пор, пока не стабилизируются прототипы.

Как видно, данный подход принципиально отличается от стандартного FCM, в связи с чем логично его распространить и на случай, когда объем обрабатываемой выборки N заранее не фиксируется, а растет с течением времени [17,18].

Несмотря на эффективность и широкое распространение FCM, ему присущ и существенный недостаток, который можно пояснить простым примером. Пусть сформировано два кластера с прототипами c_1 и c_2, и пусть на обработку поступило наблюдение $x(k)$, не принадлежащее ни одному из кластеров, однако в смысле несходства (1) равноотстоящее от обоих прототипов. Тогда, в силу первого ограничения (7) это наблюдение с равными уровнями принадлежности будет приписано обоим классам в соответствии с оценкой (9).

Этого недостатка лишен метод возможностных c-средних (PCM) [19], порождаемыйминимизацией целевой функции

$$E(u_i(k), c_i) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k)\|x(k)-c_i\|^2 + \sum_{i=1}^{r}\tau_i\sum_{k=1}^{N}(1-u_i(k))^{\beta}, \quad (11)$$

где $\tau_i > 0$ определяет расстояние между $x(k)$ и c_i, на котором уровень принадлежности $u_i(k)$ принимает значение $0,5$.

Минимизация (11) по c_i, $u_i(k)$ и τ_i ведет к результату:

$$\begin{cases} c_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}, \\ u_i(k) = \dfrac{1}{1+(\frac{\|x(k)-c_i\|^2}{\tau_i})^{\frac{1}{\beta-1}}}, \\ \tau_i = (\sum\limits_{k=1}^{N} u_i^{\beta}(k))^{-1}(\sum\limits_{k=1}^{N} u_i^{\beta}(k)\|x(k)-c_i\|^2), \end{cases} \quad (12)$$

который в случае номинальных переменных приобретает вид:

$$\begin{cases} u_i(k) = (1+\dfrac{d(c_i, x(k))}{\tau_i})^{\frac{1}{1-\beta}}, \\ \tau_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k)d(c_i, x(k))}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}. \end{cases} \quad (13)$$

Оценка (13) с вычислительной точки зрения несколько сложнее (9), однако имеет меньше недостатков, присущих FCM.

Сам же процесс возможностной нечеткой кластеризации реализуется в форме последовательности шагов, совершенно аналогичной той, что была описана выше.

Выводы

Рассмотрена задача нечеткой кластеризации категориальных переменных на основе мер несходства и сходства. Введена модификация метода возможностных *c* - средних, обладающая рядом преимуществ перед соответствующей модификацией метода нечетких *c* - средних. Предложенный метод прост в численной реализации и может найти применение при решении задач интеллектуального анализа данных, когда исходная информация задана в номинальных шкалах.

Список литературы: 1. Jain, A.K. Algorithms for Clustering Data / A.K. Jain, R.C. Dubes. – Englewood Cliffs, N.J.:Prentice Hall, 1988. - 318p. **2.** Kaufman, L. Finding Groups in Data: An Introduction to Cluster Analysis / L. Kaufman, P.J. Rousseeuw. – N.Y.: John Wiley & Sons, Inc., 1990. – 342 p. **3.** Han, J. Data Mining:Concepts and Techniques / J. Han, M. Kamber. – San Francisco: Morgan Kaufmann, 2006. – 800 p. **4.** Gan, G. Data Clustering: Theory, Algorithms, and Applications / G. Gan, C. Ma, J. Wu. – Philadelphia:SIAM, 2007. – 466 p. **5.** Abonyi, J. Cluster Analysis for Data Mining and System Identification / J. Abonyi, B. Feil. – Basel: Birkhäuser, 2007. – 303 p. **6.** Olson, D.L. Advanced Data Mining Techniques / D.L. Olson, D. Dursun. – Berlin: Springer, 2008. – 180 p. **7.** Aggarwal, C.C. Data Clustering: Algorithms and Applications / C.C. Aggarwal, C.K. Reddy. – Boca Raton:CRC Press, 2014. – 648 p. **8.** Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – N.Y.:Plenum Press, 1981. – 272 p. **9.** Hoeppner, F. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition / F. Hoeppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester: John Wiley & Sons, 1999. – 289 p.**10.** Bezdek, J.C. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J. Keller, R. Krisnapuram, N. Pal. – N.Y.: Springer Science + Business Media, Inc. – 2005. – 776 p. **11.** Huang, Zh. Extensions to the k-means algorithm for clustering large data sets with categorical values / Zh. Huang // Data Mining and Knowledge Discovery. – 1998. – 2. – №2. – P.283-304. **12.** He, Z. Improving k-modes algorithm considering frequencies of attribute values in mode / Z. He, S. Deng, X. Xu // Computational Intelligence and Security. – 2005. – V.3801 of the series «Lecture Notes in Computer Science». – P. 157-162. **13.** Lei, M. An improved k-means algorithm for clustering categorical data / M. Lei, P. He, Zh. Li // J. of Communications and Computer. – 2006. – 3. – №8. – P. 20-24. **14.** Huang, Zh. A fuzzy k-modes algorithm for clustering categorical data / Zh. Huang, M.K. Ng // IEEE Trans on Fuzzy Systems. – 1999. – 7. – №4. – P. 446-452. **15.** Kim, D.W. Fuzzy clustering of categorical data using fuzzy centroids / D.W. Kim, K.H. Lee, D. Lee // Pattern Recognition Letters. – 2004. -25. – P. 1263-1271. **16.** Lee, M. Fuzzy p-mode prototypes: A generalization of frequency-based cluster prototypes for clustering categorical objects / M. Lee // Computational Intelligence and Data Mining. – Nashville, TN, 2009. – P. 320-323. **17.** Bodyanskiy, Y. Recursive fuzzy clustering algorithms / Y. Bodyanskiy, V. Kolodyazhniy, A. Stephan // Proc. 10th East–West Fuzzy Colloquium. – Zittau, Germany. – 2002. – P. 276-283. **18.** Bodyanskiy, Y. Computational intelligence techniques for data analysis / Y. Bodyanskiy // Lecture Notes in Informatics. – P-72. – Bonn: GI. – 2005. – P. 15–36. **19.** Krishnapuram, R. A possibilistic approach to clustering / R. Krishnapuram, J. Keller // IEEE Trans. on Fuzzy Systems. – 1993. – 2. - №1. – P.98-110.

Поступила в редколлегию 24.12.2015

УДК 004.032.26

Можливістна нечітка кластеризація масивів категоріальних даних із використанням частотних прототипів та мір несхожості / Е.В.Бодянський, В.О. Самітова// Біоніка інтелекту: наук.-техн. журнал. – 2016. –№ 1(86). – С. 72-75.

Стаття присвячена дослідженням методів нечіткої кластеризації масивів категоріальних даних. Проводиться детальний опис методу нечіткої кластеризації на основі частотних прототипів, а також методу, в основі якого лежать міри несхожості. Розглянуті переваги та недоліки розглянутих методів. Запропонована модифікація методу можливістних *c* - середніх, яка має низку переваг при роботі з категоріальними даними.

Бібліогр.: 19 найм.

UDC 004.032.26

Possibilistic fuzzy c-mean clustering method for categorical fata using frequency-based cluster prototypes and dissimilarity measure / Y. Bodyanskiy, V. Samitova // Bionicaintellecta. – 2016. – N 1 (86). – P. 72-75.

Article is devoted to researches of categorical data fuzzy clustering. The detailed description of clustering method using frequency-based cluster prototypes and dissimilarity measure is given. The modification of possibilistic fuzzy c-means clustering algorithm for categorical data possessing some advantages is proposed.

Ref: 19 items.