An IMPROVED FCM-POSSIBLE CLUSTERING ALGORITHM FOR INTERVAL DATA

LI Qing, LUO Jianlu, TAN Xiaodong, DENG Xiaoyan, LU Bing
Department of Electronic Technology, Officers College of CAPF, Chengdu 610213, Sichuan, China
email: 6900624@qq.com

Abstract

In order to overcome the disadvantages of Fuzzy c-Means(FCM) and possible clustering algorithm in the practical application of interval data, an improved FCM-possible clustering algorithm for interval data is proposed in this paper by combing their merits. The improved clustering algorithm introduce the possibility theory into the clustering problem of interval data, by relaxing the constraints of the sample membership and modifying IFCM algorithm's objective function The results of simulation experiments and the average CR index analysis show that: For the cluster problems containing poorly representative sample data such as noise and outliers, the improved FCM-possible clustering algorithm proposed in this paper is much better than the FCM algorithm and possible clustering algorithm, which can effectively reduce the influence on the clustering result by the noise .

Keywords: Interval data; Improved FCM-Possible Clustering Algorithm; Average CR index

Introduction

With the continuous development of the computer technology and the needs of practical problems, clustering algorithm based on the objective function gradually becomes the mainstream of fuzzy clustering, fuzzy c means clustering algorithm is one of the most widely used among them [1]. However, it is difficult to evaluate and describe the property of the clustering objective precisely in the practical application due to the uncertain effect of objective things such as complexity, randomness, and fuzziness. Therefore, interval data are usually used to describe the property information. And clustering analysis of interval data gives rise to extensive attention and study [2] [3] [4] [5]. Francisco et al. [6] discussed two kinds of fuzzy c means clustering algorithm of interval data based on the Euclidean distance of the interval data. Fan Jiulun et al. [7] proposed FCM algorithm with two kinds of the interval value data in virtue of the classical FCM algorithm. Furthermore, Gao Xinbo et al. [8] put forward an improved

algorithm which can control the influence of interval size on the clustering results by introducing one weighting factor to generate an adaptive distance.

These algorithms are proposed by expanding the standard FCM algorithm to the interval data. In the practical application, there exist some defects and deficiencies similar to the standard FCM algorithm. For example, the iterative process is easy to fall into local minimum point, and the clustering effect is affected by isolated points or the noise data, etc.

The possibility clustering algorithm of interval data

In this paper, based on the Fuzzy c-Means(FCM) clustering algorithm of the interval data, the improved FCM-possible clustering algorithm of interval data by relaxing the normalized constraints of the sample membership and modifying the objective function of the algorithm is proposed in this paper. Through using the synthetic data to carry out simulation experiments, the algorithm proposed in this paper can reduce the influence of isolated points or the noise data on the clustering effect, and has certain advantages.

According to the deficiency that the IFCM algorithm is extremely sensitive to the noise data in the application, in this paper, we improve IFCM algorithm from the relaxing sample membership constraints and modifying objective function, and propose the possibility clustering algorithm of one kind of interval data to reduce the influence of the isolated points or noise data on the clustering effect.

On the one hand, the IFCM algorithm has the normalized constraints for each sample point, which makes the sample membership not only about the clustering prototype, but also affected by the other clustering prototypes. Moreover, that must meet the sum of the membership equaling to 1.

Usually, when the data set includes the isolated points or noise data, theoretically, the membership degree of the noise data affiliated with each clustering should be small. But the membership degree of noise data is high because of the limitation of the normalization conditions, so the clustering results can't reflect the actual situation. Hence, IPFCM should firstly relax constraints of the membership of the sample，not require the sum of the membership equaling to 1, and just need $\max_i \mu_{ki} > 0$

That the constraints equation is revised as：

$$U_{\text{IPCM}} = \left\{ \begin{array}{l} U = (\mu_{ki}) \mid \mu_{ki} \in [0,1], \forall i, k; \\ 0 < \sum_{k=1}^{n} \mu_{ki} < n, \forall i; \max_i \mu_{ki} > 0, \forall k \end{array} \right\} \quad (1)$$

Especially, applying the membership based on the typicality can automatically reduce the influence on the clustering effect for the sample data of poorer representation including the noise and isolated points, etc.

On the other hand, the possibility theory and penalty factor are introduced into the clustering problems of the interval data, and the objective function is revised.

$$J_{\text{IPCM}}(U,V,\delta) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ki})^m\sum_{j=1}^{p}[(x_{kj}^- - v_{ij}^-)^2 + (x_{kj}^+ - v_{ij}^+)^2] + \sum_{i=1}^{c}\delta_i\sum_{k=1}^{n}(1-u_{ki})^m \quad (2)$$

Where $\delta_i > 0$ is the penalty factor, which decides the scope size of one kind of the clustering.

Similar to the IFCM algorithm, it can obtain the fuzzy dividing matrix and the iterative formula of clustering prototype by solving optimization problem: $\min_{U\in U_{\text{IPCM}},V} J_{\text{IPCM}}(U,V,\delta)$.

And, we can obtain the iterative formula of clustering prototype as：

$$v_{ij}^- = \frac{\sum_{k=1}^{n}(u_{ki})^m x_{kj}^-}{\sum_{k=1}^{n}(u_{ki})^m} \quad (3)$$

$$v_{ij}^+ = \frac{\sum_{k=1}^{n}(u_{ki})^m x_{kj}^+}{\sum_{k=1}^{n}(u_{ki})^m} \quad (4)$$

,

$$\mu_{ki} = \frac{1}{1+\left(\frac{\sum_{j=1}^{p}[(x_{kj}^- - v_{ij}^-)^2 + (x_{kj}^+ - v_{ij}^+)^2]}{\delta_i}\right)^{1/(m-1)}} \quad (5)$$

Therefore, the specific steps of IPCM algorithm are as follows.

Step one Initialization: Preset clustering number c, $(1 \le c \le n)$; set the weighted index $m \ge 1$; set iteration termination error $\varepsilon > 0$;

Step two: Divide matrix $\overline{U}$ in initialization , and estimate δ_i;

Step three: Calculate $\overline{V} = \arg\min_{V} J(\overline{U},V)$ and $\overline{U} = \arg\min_{U\in U_{\text{IPCM}}} J(U,\overline{V})$;

Step four: Estimate δ_i again;

Step five: If $\overline{U}$ or $\overline{V}$ weaken, the algorithm stop; or return to Step 3.

Experiment and Result

Example 1: Synthetic data sets are generated by reference to the literature [10]. Considering that two-dimensional spatial data sets contain 500 spatial data of three classes, of which the sample points (z_1, z_2) obey independent two-dimensional normal distribution. The samples of three clustering in the data sets are generated randomly and join white noise, which is shown in Figure 1.

The above spatial data sets generates interval data sets and is denoted by X_{500}, of which γ_1 and γ_2 represent the width and height of interval data set, respectively. We carry out the experiment repeatedly to ensure the validity and stability of the experiment, and the value interval of γ_1 and γ_2 generate randomly in the [1,8]，[1,16]，[1,24]，[1,36], respectively.

Based on Monte Carlo simulation, we use CR index to evaluate the clustering performance of IPCM algorithm. We suppose that $U = \{u_1, u_2, ..., u_R\}$ is a division of IPCM algorithm, $V = \{v_1, v_2, ..., v_C\}$ is a transcendental division, CR index is defined as

$$\mathrm{CR} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} C_2^{n_{ij}} - \left(C_2^n\right)^{-1} \sum_{i=1}^{R} C_2^{n_{i.}} \sum_{j=1}^{C} C_2^{n_{.j}}}{\frac{1}{2}\left(\sum_{i=1}^{R} C_2^{n_{i.}} + \sum_{j=1}^{C} C_2^{n_{.j}}\right) - \left(C_2^n\right)^{-1} \sum_{i=1}^{R} C_2^{n_{i.}} \sum_{j=1}^{C} C_2^{n_{.j}}} \quad (6)$$

Where n_{ij} shows the number of clustering objects belonging to both u_i and v_j; $n_{i.}$ represents the number of clustering objects belonging to u_i; $n_{.j}$ represents the number of clustering objects belonging to v_j; n is the total number of clustering objects. Obviously, $-1 \le \mathrm{CR} \le 1$ When the CR is closer to 1, it indicates that performance of clustering algorithm is better, otherwise that is worse.

Figure 2 shows interval data sets X_{500} generated by spatial datasets (value interval of γ_1, γ_2 generated randomly in [1,8]).

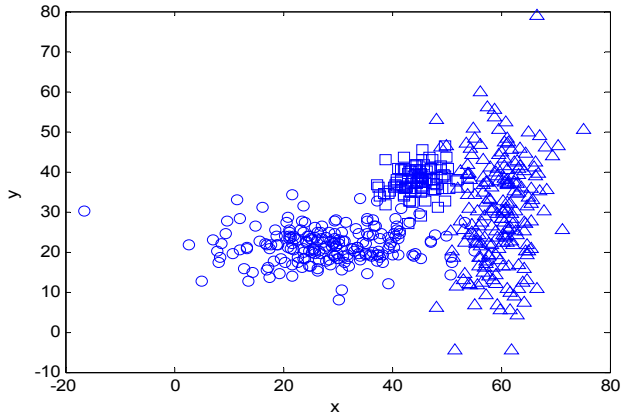

Figure 1. The spatial data set (containing white noise) randomly generated.

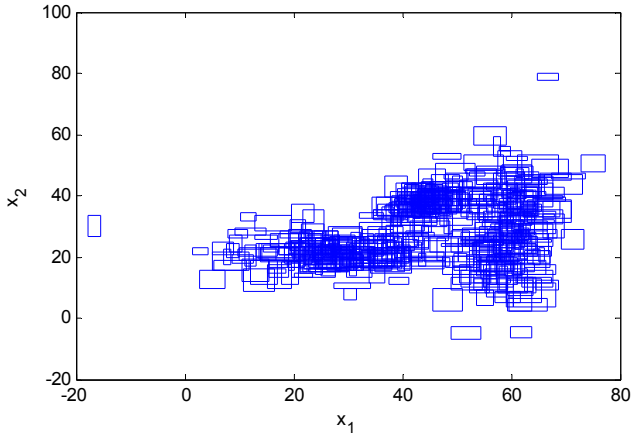

Figure 2. Synthetic interval data set X_{500}.

Choosing $m = 2$, $c = 3$, $\varepsilon = 0.001$, value interval of γ_1 and γ_2 is generated randomly in [1, 8], [1, 16], [1, 24], [1, 36], respectively, and then the synthetic interval datasets are obtained. And the average CR index values are calculated by using Monte Carlo simulation to repeat the experiment 250 times with the IFCM algorithm and IPCM algorithm, respectively, shown in Table 3. From Table 3, for the four kinds of generating situations of $[\gamma_1, \gamma_2]$, the average CR index values obtained by the IPCM algorithm are considerably greater than those obtained by the IFCM algorithm in this paper, which shows IPCM algorithm embodies better superiority in the clustering of noise data set. However, by comparing the computational efficiency of the algorithm, it is found that the average number of iterations of IPCM algorithm in this article is greater than that of IFCM algorithm. That is because the penalty factor δ_i needs to be re-evaluated with constantly updated matrix division and clustering prototype in IPCM algorithm.

Table 1 Comparison of average CR index value and calculation efficiency.

$[\gamma_1, \gamma_2]$	Average CR index value		Average time of iteration	
	IFCM algorithm	IPCM algorithm	IFCM algorithm	IPCM algorithm
[1,8]	0.473	0.815	22	54
[1,16]	0.481	0.798	20	56
[1,32]	0.448	0.807	19	53
[1,36]	0.406	0.763	19	50

Conclusion

In this paper, the author presents a possibility clustering algorithm for interval data to solve problems of interval clustering data. Compared with IFCM algorithm, IPCM algorithm can effectively reduce the impact of noise and isolated points on clustering effect. However, we find that in the simulation experiment IPCM algorithm is sensitive to initial values and easy to produce consistent clustering. Therefore, it is necessary to further explore the initialization problems of IPCM algorithm and the consistency clustering problems in the future study.

References

[1] SADAAKI MIYAMOTO, HIDETOMO ICHIHASHI, KATSUHIRO HONDA. Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications [M]. Berlin: Springer, 2008.

[2] Renato. M. C. R. de Souza,De Carvalho, F. A. T. Clustering of interval data based on city-block distances[J]. Pattern Recognition Letters, 2004, 25: 353-365.

[3] De Carvalho, F. A. T. , Renato. M. C. R. de Souza, Marie Chavent and Yves Lechevallier. Adaptive Hausdorff distances and dynamic clustering of symbolic data[J].Pattern Recognition Letters, 2006, 27: 167-179.

[4] Meng Guangwu, Zhang Xingfang, Zheng Yalin. Clustering method based on interval-valued fuzzy sets [J]. Journal of Engineering Mathematics, 2001, 28(02):69-73.

[5] Xu Jianjiang, Xu Baowen. Parallel fuzzy clustering algorithm to Interval data [J] Southeast University (Natural Science), 2003, 33 (04): 406-409.

[6] Francisco de A.T. de Carvalho. Fuzzy c-means clustering methods for symbolic interval data[J]. Pattern Recognition Letters, 2007, 28(4): 423-437.

[7] Fan Jiulun, Pei Jihong,Xie Weixin. Interval-valued fuzzy c- means clustering algorithm[C]. Baoding, China: Fuzzy Set Theory and Applications-Chinese Fuzzy Mathematics and Fuzzy Systems Committee the Ninth Annual Meeting of Selected Papers in 1998, Hebei University Press, 1998: 602-604 .

[8] Gao Xinbo, Fan Jiulun, Xin Weixin. New Interval-valued fuzzy c- means clustering algorithm [J] Xi'an University of Electronic Science and Technology, 1999, (05): 604-609.
[9] Zhang Weibin, Liu Wenjiang. Fuzzy c -means clustering algorithm of interval data [J]. Computer Engineering, 2008, 34 (11): 26-28.
[10] Yue Mingdao. The new Fuzzy c means clustering algorithm of interval data [J] Computer Engineering and Applications, 2011, 47 (13): 157-160.