Robust Cross-Entropy Fuzzy Clustering Algorithm Postprint

Authors: Yao Lan, Yan Hanbing, Wei Zefeng

Date: 2018-06-19T00:00:00+00:00

Abstract

To address the problems of noise sensitivity and susceptibility to local minima in the fuzzy C-means clustering algorithm, a fuzzy clustering algorithm based on cross-entropy is proposed. This algorithm redefines the objective function of the traditional FCM algorithm by introducing cross-entropy, utilizes cross-entropy to measure the differences among sample membership degrees, and employs the Lagrange multiplier method and Lambert W function to solve the resulting optimization problem; furthermore, it analyzes the distribution of the sample partition matrix and identifies noise samples based on the distribution characteristics. Experimental results on artificial datasets and standard datasets with added noise demonstrate that the algorithm improves the anti-interference capability of the traditional FCM algorithm, exhibits stronger robustness, and achieves high accuracy in noise sample identification.

Full Text

Preamble

Robust Cross-Entropy Fuzzy Clustering Algorithm

Yao Lan, Yan Hanbing, Wei Zefeng

(School of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: To address the sensitivity of the fuzzy C-means clustering algorithm to noise and its tendency to converge to local minima, this paper proposes a fuzzy clustering algorithm based on cross-entropy. The algorithm redefines the objective function of the traditional FCM algorithm by introducing cross-entropy to measure the differences between sample membership degrees. The Lagrange multiplier method and Lambert W function are employed to solve the optimization problem of the objective function. Additionally, the distribution of the sample partition matrix is analyzed to identify noise samples based on its characteristics. Experimental results on synthetic and standard datasets with added noise demonstrate that the proposed algorithm enhances the anti-interference capability of the traditional FCM algorithm, exhibits stronger robustness, and achieves higher accuracy in noise sample identification.

Keywords: fuzzy clustering; cross-entropy; fuzzy C-means clustering; clustering performance

0 Introduction

Since Ruspini first introduced fuzzy theory into cluster analysis in 1969, fuzzy-theory-based clustering methods have attracted widespread attention from researchers, leading to the development of numerous fuzzy clustering algorithms. Among these, the fuzzy C-means clustering algorithm (FCM) has become a classical method in fuzzy clustering due to its simple design, wide applicability, and ease of implementation, finding extensive use across various domains. However, theoretical studies and experiments have revealed that FCM suffers from inherent limitations, including extreme sensitivity to noisy data and a propensity to converge to local minima.

To overcome these issues, researchers have proposed many algorithms by modifying the objective function and relaxing the membership constraints. Representative algorithms include the possibilistic clustering algorithm (PCM) and its improved version IPCM, the possibilistic fuzzy C-means algorithm (PFCM), the weighted fuzzy C-means algorithm (WFCM) and related methods, clustering algorithms based on noise models, and enhanced fuzzy partition clustering algorithms. These approaches have improved upon FCM's limitations to varying degrees, enhancing its practical applicability.

Inspired by information theory' s use of entropy to measure system information content, scholars have introduced entropy into FCM algorithms, developing entropy-based fuzzy clustering methods. These approaches endow the sample membership calculation formula with Gaussian distribution characteristics, thereby exhibiting stronger noise resistance. The maximum entropy clustering algorithm (MEC) was proposed using an entropy function as a component of the objective function, though this method remains sensitive to outliers, whose interference often causes severe deviation of cluster centers. The robust maximum entropy clustering algorithm (RMEC) was subsequently developed to improve outlier resistance. Building upon fuzzy entropy, an intuitionistic fuzzy C-means clustering algorithm was proposed by redefining the objective function with hesitation degrees, which shows some effect on noise suppression but with limited performance and high computational complexity. The relative entropy fuzzy C-means clustering algorithm (REFCM) was introduced using relative entropy to replace entropy. Other researchers have proposed fuzzy clustering algorithms based on generalized entropy, solved using augmented Lagrangian methods and Hopfield neural networks, as well as generalized entropy-based possibilistic fuzzy

C-means clustering algorithms that can more accurately obtain cluster centers for noisy data.

However, entropy only addresses individual variables and cannot effectively measure the differences between two random distributions, nor does it satisfy symmetry. To address these limitations, this paper introduces cross-entropy into the traditional FCM objective function, proposing a cross-entropy fuzzy C-means clustering algorithm (CEFCM). Cross-entropy effectively measures the differences between membership degrees of various sample classes. The resulting algorithm ensures that during objective function optimization, the obtained partition matrix makes each sample' s membership degrees influenced not only by distance but also constrained by cross-entropy, preventing the iteration process from falling into local minima. The objective function consists of two terms: the original FCM objective function and a cross-entropy term that essentially measures the degree of difference among sample membership degrees. When the first term is minimized and the second term is maximized, the objective function reaches its minimum, indicating that all samples are closest to their cluster centers with concentrated intra-class data, while the differences among class memberships are maximized, resulting in clear inter-class separation and an optimal partition.

Furthermore, the final partition matrix theoretically exhibits distinct distributions for data samples versus noise samples: data samples have relatively large membership values for one class and smaller values for others, while noise samples have uniformly low membership values for all classes. Based on these distribution characteristics, samples with consistently low membership values can be screened out as noise samples, enabling noise identification.

1 Fuzzy Cross-Entropy

Based on Kullback' s definition of cross-entropy in information theory, for two fuzzy sets A and B in the universe of discourse X, the symmetric form of fuzzy cross-entropy is defined as:

$$D(A,B)=\sum_{i=1}^{N}\left[\mu_A(x_i)\ln\frac{\mu_A(x_i)}{\mu_B(x_i)}+\mu_B(x_i)\ln\frac{\mu_B(x_i)}{\mu_A(x_i)}\right]$$

where $\mu_A(x_i)$ and $\mu_B(x_i)$ represent the membership degrees of element x_i in X belonging to A and B, respectively. Cross-entropy is a convex function that attains a minimum value of zero when the two distributions are identical. Although cross-entropy is not a true geometric distance, it proves highly effective for measuring differences between fuzzy sets.

2.1 Basic Idea of the Algorithm

Clustering is the process of distinguishing and classifying a dataset according to the differences among data objects. Greater differences facilitate the weakening of irrelevant information while preserving relevant information, thereby producing good clustering results. Accordingly, this paper introduces fuzzy cross-entropy, which measures differences between fuzzy sets, into the FCM objective function, proposing the CEFCM algorithm. The constructed objective function is:

$$\min J(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^2 d_{ij}^2 + \theta \sum_{j=1}^{N}\sum_{i=1}^{c}\sum_{\substack{k=1\\k\neq i}}^{c}\mu_{ij}\ln\frac{\mu_{ij}}{\mu_{kj}}$$

subject to:

$$\sum_{i=1}^{c}\mu_{ij} = 1, \quad \forall j; \quad 0 < \mu_{ij} < 1, \quad \forall i,j$$

where d_{ij}^2 is the Euclidean distance from the j-th sample to the i-th cluster center, μ_{ij} is the membership degree of the j-th sample to the i-th cluster center, c is the number of cluster centers, N is the number of samples, and θ is the cross-entropy adjustment coefficient determining the influence of cross-entropy.

This objective function comprises two terms: the first term is the original FCM objective function, while the second term is the cross-entropy function that essentially measures the degree of difference among membership degrees of various samples. By fusing these two terms, the CEFCM algorithm ensures that during objective function optimization, the membership degrees of each sample in the obtained partition matrix are influenced not only by distance but also constrained by cross-entropy, preventing the iteration process from easily falling into local minima. Clearly, when the first term is minimized and the second term is maximized, the objective function reaches its minimum value, indicating that all samples are at the minimum distance from their cluster centers with concentrated intra-class data, while the differences among class memberships are maximized, resulting in clear inter-class separation and an optimal partition.

Additionally, the final partition matrix theoretically shows that the membership degree distributions of data samples and noise samples are mutually distinct: data samples necessarily have relatively large membership values for one class and smaller values for others, whereas noise samples have uniformly low membership values for all classes. Therefore, based on the differences in membership distribution characteristics between data samples and noise samples, samples with consistently low membership values can be screened from the sample set as noise samples, completing noise identification.

2.2 Algorithm Description

For the objective function presented above, this section provides the specific derivation and algorithm steps. Using the Lagrange multiplier λ_j, the constrained optimization function is constructed as:

$$L = \sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^2 d_{ij}^2 + \theta\sum_{j=1}^{N}\sum_{i=1}^{c}\sum_{\substack{k=1\\k\neq i}}^{c}\mu_{ij}\ln\frac{\mu_{ij}}{\mu_{kj}} - \sum_{j=1}^{N}\lambda_j\left(\sum_{i=1}^{c}\mu_{ij}-1\right)$$

Since the membership values μ_{ij} of samples to their cluster centers are mutually independent, the minimization problem of the objective function J with respect to μ_{ij} can be decomposed into minimization problems with respect to each membership function μ_{ij}.

The first necessary condition is obtained by taking the partial derivative of L with respect to μ_{ij}:

$$\frac{\partial L}{\partial \mu_{ij}} = 2\mu_{ij}d_{ij}^2 + \theta\left(\sum_{\substack{k=1\\k\neq i}}^{c}\ln\frac{\mu_{ij}}{\mu_{kj}} + \sum_{\substack{k=1\\k\neq i}}^{c}(1-\frac{\mu_{kj}}{\mu_{ij}})\right) - \lambda_j = 0$$

Equation (6) cannot be solved directly. Considering the consistency of the value range, we substitute y_{ij} for μ_{ij} and y_{kj} for μ_{kj}, rewriting equation (6) as:

$$2y_{ij}d_{ij}^2 + \theta\left(\sum_{\substack{k=1\\k\neq i}}^{c}\ln\frac{y_{ij}}{y_{kj}} + \sum_{\substack{k=1\\k\neq i}}^{c}(1-\frac{y_{kj}}{y_{ij}})\right) - \lambda_j = 0$$

Let $K = \sum_{\substack{k=1\\k\neq i}}^{c} y_{kj}$. Solving equation (8) for y_{ij} yields:

$$y_{ij} = -\frac{\theta K}{2d_{ij}^2}W\left(-\frac{2d_{ij}^2}{\theta K}\exp\left(-\frac{2d_{ij}^2}{\theta K}\right)\exp\left(-\frac{\lambda_j}{\theta K}\right)\right)$$

where $W(\cdot)$ is the Lambert W function satisfying $W(Z)\exp(W(Z)) = Z$.

Therefore, the membership degree can be expressed as:

$$\mu_{ij} = \frac{W\left(-\frac{2d_{ij}^2}{\theta K}\exp\left(-\frac{2d_{ij}^2}{\theta K}\right)\exp\left(-\frac{\lambda_j}{\theta K}\right)\right)}{\sum_{k=1}^{c}W\left(-\frac{2d_{kj}^2}{\theta K}\exp\left(-\frac{2d_{kj}^2}{\theta K}\right)\exp\left(-\frac{\lambda_j}{\theta K}\right)\right)}$$

2.3 Parameter Selection

The constraint condition $\sum_{i=1}^{c} \mu_{ij} = 1$ requires that equation (10) be non-negative, which in turn requires that the expression inside the Lambert W function must be non-negative. For the function $W(Z)$, when $Z \geq 0$, we have $W(Z) \geq 0$. In this paper, we require:

$$-\frac{2d_{ij}^2}{\theta K} \exp\left(-\frac{2d_{ij}^2}{\theta K}\right) \exp\left(-\frac{\lambda_j}{\theta K}\right) \geq 0$$

Since $\mu_{ij} \geq 0$ for all i, j, and $\sum_{k=1,k\neq i}^{c} \mu_{kj} \geq 0$, we have $K \geq 0$. The condition that each sample' s membership degree to cluster centers must be less than 1, i.e., $\mu_{ij} \leq 1$ for all i, j, leads to:

$$\frac{W(\cdot)}{\sum_{k=1}^{c} W(\cdot)} \leq 1$$

Based on equation (14), we can determine the value range of λ_j during each iteration:

$$\lambda_j \geq -\theta \ln\left(\sum_{k=1,k\neq i}^{c} \mu_{kj} \exp\left(-\frac{2d_{ij}^2}{\theta \sum_{k=1,k\neq i}^{c} \mu_{kj}}\right)\right)$$

The value of λ_j is independent of μ_{ij}.

2.4 Algorithm Steps

Based on the above discussion, the computational steps of the CEFCM algorithm are as follows:

a) Determine the number of clusters c and the cross-entropy coefficient θ, as well as the iteration threshold ε.

b) Initialize the membership degrees μ_{ij} and cluster centers $v_i^{(0)}$, set $t = 1$.

c) Calculate the distances d_{ij} between samples and each cluster center.

d) Determine the Lagrange multiplier λ_j according to equation (14).

e) Update the membership degrees μ_{ij} using equation (10).

f) Update the cluster centers v_i using equation (11).

g) If $\max \|v_i^{(t)} - v_i^{(t-1)}\| < \varepsilon$, stop the algorithm; otherwise, set $t = t + 1$ and return to step c).

For cluster centers v_i, the general method obtains the iterative update expression by solving $\frac{\partial L}{\partial v_i} = 0$. Considering that d_{ij} represents Euclidean distance in this paper, the update formula is:

$$v_i = \frac{\sum_{j=1}^{N} \mu_{ij}^2 x_j}{\sum_{j=1}^{N} \mu_{ij}^2}$$

2.5 Algorithm Time Complexity Analysis

Observing equation (12), we find that λ_j and the membership functions are interdependent and cannot be solved directly. The time complexity of the classic FCM algorithm is $O(Ncp)$, where N represents the dataset size, c is the number of clusters, and p is the dataset dimensionality. For the proposed CEFCM algorithm, due to the introduction of cross-entropy in the objective function, its time complexity becomes $O(Nc^2p+Nc^2 \log c)$. Although the CEFCM algorithm increases computational overhead, its clustering performance is significantly improved.

3 Experimental Results and Analysis

To validate the performance of the proposed CEFCM algorithm, this section conducts experiments using two different datasets. The first experiment uses a synthetic dataset to test the CEFCM algorithm's performance, while the second experiment employs the IRIS standard dataset for analysis and evaluation.

In the experiments, clustering validity indices are used to compare algorithm performance. Since literature [21] indicates that validity indices based on the membership matrix are more suitable for practical applications with noisy sample data, this paper calculates compactness and separation based on the membership matrix. Compactness is defined as:

$$C = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij}^2$$

where u_{ij} represents the membership degree of the j-th sample to the i-th class. When all samples have high membership degrees to a particular class, the compactness is greater, indicating tighter clustering results.

Overall separation is defined as:

$$S = \min_{i \neq k} S_{ik}, \quad \text{where } S_{ik} = \sum_{j=1}^{N} \min(\mu_{ij}, \mu_{kj})$$

S_{ik} represents the separation between the i-th and k-th classes. When the partition between two classes is clear, samples must have membership degrees to one class significantly larger than to others; therefore, smaller S_{ik} values indicate clearer clustering results. Overall separation measures the separation between the two most ambiguous classes. Smaller overall separation indicates greater distinction between classes.

The synthetic dataset used in this section is generated following the method in literature [16], consisting of four classes of data in 2D Euclidean space following Gaussian distributions, with each class containing 100 data points. The first class has center (1,2) and covariance matrix $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$; the second class has center (-1,-2) and covariance matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; the third class has center (-3,5) and covariance matrix $\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$, as shown in Figure 1(a). The fourth class consists of noise samples placed at the tail of the dataset, with center (-3,5) and covariance matrix $\begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$, as shown in Figure 1(b).

[Figure 1: see original paper] shows the synthetic dataset sample data: (a) original dataset, (b) noisy dataset.

Table 1 presents the performance comparison of the FCM, REFCM, and proposed CEFCM algorithms. The results show that the proposed algorithm achieves higher compactness and smaller separation, indicating that CEFCM produces better clustering results for noisy samples, with clearer sample partitioning and demonstrably strong noise resistance and robustness.

[Figure 2: see original paper] compares the partition matrices obtained by three clustering algorithms on the dataset from Figure 1(b): (a) FCM, (b) REFCM, (c) CEFCM. All three algorithms produce partition matrices where data samples have high membership values for one class and low values for others, while noise samples have low membership values, consistent with the theoretical conclusion that data samples and noise samples exhibit distinct membership distribution characteristics. Notably, Figure 2(c) shows more pronounced distribution features, with data samples having higher membership values to their respective centers and noise samples having lower values, indicating more reasonable data partitioning by the CEFCM algorithm and significantly reduced noise impact on the clustering process. Moreover, the more distinct the distribution characteristics between data and noise samples, the more prominent their differences, facilitating the screening of samples with uniformly low membership values as noise and achieving higher noise identification rates.

3.2 Test on Noisy IRIS Dataset

This section employs the renowned IRIS dataset as test data. The IRIS dataset consists of 150 samples in four-dimensional space, divided into three classes with 50 samples each. Literature [22] provides the actual class centers as: $p_1 = (5.00, 3.42, 1.46, 0.24)$, $p_2 = (5.93, 2.77, 4.26, 1.32)$, $p_3 = (6.58, 2.97, 5.55, 2.02)$.

Table 2 shows the results of various algorithms on the noisy IRIS dataset. Under noise interference, FCM, PFCM, and WFCM exhibit large numbers of misclassified samples and significant center deviations, indicating high noise sensitivity and poor robustness. IPCM and the method from literature [8] simultaneously achieve better misclassification rates and cluster centers, demonstrating strong robustness. The proposed method achieves the smallest number of misclassified samples and highest accuracy, with center deviation slightly lower than literature [8], indicating fewer misclassifications and more accurate cluster centers, thus achieving high clustering accuracy.

4 Conclusion

To address the noise sensitivity problem of the FCM algorithm, this paper proposes a fuzzy clustering algorithm based on fuzzy cross-entropy. Experimental results demonstrate that the algorithm exhibits strong robustness to noise, achieves fewer misclassifications and more accurate cluster centers, and produces superior clustering results compared to existing algorithms such as FCM and PFCM. Additionally, the algorithm can effectively perform noise identification simultaneously with clustering. Future research will focus on further improvements, including parameter θ selection and sample feature extraction.

References

[1] Ruspini E H. A new approach to clustering [J]. Information & Control, 1969, 15(1): 22-32.

[2] Bezdek J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1981.

[3] Krishnapuram R, Keller J. A possibilistic approach to clustering [J]. IEEE Trans on Fuzzy Systems, 1993, 1(2): 98-110.

[4] Zhang Jiangshe, Leung Yiu wing. Improved possibilistic C-means clustering algorithms [J]. IEEE Trans on Fuzzy Systems, 2004, 12(2): 209-217.

[5] Pal N R, Pal K, Keller J M, et al. A possibilistic fuzzy C-means clustering algorithm [J]. IEEE Trans on Fuzzy Systems, 2005, 13(4): 517-530.

[6] Askari S, Montazerin N, Zarandi M H F. Generalized possibilistic fuzzy C-means with novel cluster validity indices for clustering noisy data [J]. Applied Soft Computing, 2017, 53: 262-283.

[7] Gao Xinbo, Li Jie, Ji Hongbing. A multi-threshold image segmentation algorithm based on weighting fuzzy C-means clustering and statistical test [J]. Acta Electronica Sinica, 2004, 32(4): 661-664.

[8] Liu Bing, Xia Shixiong, Zhou Yong, et al. A sample-weighted possibilistic fuzzy clustering algorithm [J]. Acta Electronica Sinica, 2012, 40(2): 371-375.

[9] Wang Lina, Wang Jiandong, Li Tao, et al. Cluster' s feature weighting fuzzy clustering algorithm integrating rough sets and shadowed sets [J]. System Engineering and Electronics, 2013, 35(8): 1769-1776.

[10] Zhao Xuemei, Li Yu, Zhao Quanhua. A fuzzy clustering image segmentation algorithm with double neighborhood system combined with Markov Gaussian model [J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(4): 615-623.

[11] Zhu Lin; Chung Fulai; Wang Shitong. Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions [J]. IEEE Trans on Systems Man & Cybernetics Part B Cybernetics, 2009, 39(3): 578-591.

[12] Jiang Yizhang, Deng Zhaohong, Wang Jun, et al. Transfer generalized fuzzy C-means clustering algorithm with improved fuzzy partitions by leveraging knowledge [J]. Pattern Recognition and Artificial Intelligence, 2013, 26(10): 975-984.

[13] Li Ruiping, Mukaidono, M. A Maximum-entropy approach to fuzzy clustering [C] //Proc of the 4th IEEE International Conference on Fuzzy Systems, Piscataway, NJ: IEEE Press, 1995: 2227-2232.

[14] Wang Shitong, Chung Fulai, Deng Zhaohong, et al. Robust maximum entropy clustering algorithm with its labeling for outliers [J]. Soft Computing, 2006, 10(7): 555-563.

[15] Chaira T. A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images [J]. Applied Soft Computing, 2011, 11(2): 1711-1717.

[16] Zarinbal M, Zarandi M H F, Turksen I B. Relative entropy fuzzy C-means clustering [J]. Information Sciences, 2014, 260(1): 74-97.

[17] Li Kai, Cao Zhe. A Fuzzy clustering algorithm with generalized entropy based on neural network [J]. Acta Electronica Sinica, 2016, 44(8): 1881-1887.

[18] Askari S, Montazerin N, Zarandi M H F, et al. Generalized entropy based possibilistic fuzzy C-means for clustering noisy data and its convergence proof [J]. Neurocomputing, 2017, 219(5): 186-202.

[19] Kullback S. Information Theory and Statistics [M]. New York: John Wiley, 1959.

[20] Long Min, Zhou Tiejun. A survey of properties and application of the Lambert W function [J]. Journal of Hengyang Normal University, 2011, 32(6): 38-40.

[21] Piao Shangzhe, Chaomurilige, Yu Jian. Cluster validity indexes for FCM clustering algorithm [J]. Pattern Recognition & Artificial Intelligence, 2015,

28(5): 452-461.

[22] Hathaway R J, Bezdek J C. Nerf C-means: Non-Euclidean relational fuzzy clustering [J]. Pattern Recognition, 1994, 27(3): 429-437.

Note: Figure translations are in progress. See original paper for figures.