Available online at www.sciencedirect.com

SciVerse ScienceDirect

Energy Procedia 13 (2011) 3924 – 3931

Energy Procedia

ESEP 2011: 9-10 December 2011, Singapore

A Sample-Weighted Robust Fuzzy C-Means Clustering Algorithm

Shi-xiong XIA, Xu-dong HAN*, Bing LIU, Yong ZHOU

School of Computer Science, China University of Mining and Technology, XuZhou, China

Abstract

The Fuzzy c-means clustering (FCM) algorithm is sensitive to noises and outliers and tends to get local optimal solutions. The Possibilistic Fuzzy C-Means (PFCM) clustering algorithm and the Improved Possibilistic C-Means (IPCM) clustering algorithm are resistant to noises but consume too much time because of the slow convergence and what is more, they are not robust enough to noises. To overcome these defects, this paper proposes a sample-weighted robust fuzzy c-means clustering algorithm which calculates the weight for each sample. The outliers will receive much lower weights and their effect on cluster centers is restrained and thus robustness is achieved. The experimental results show the effectiveness of the algorithm. In addition, the idea of calculating weights for samples can be applied to other fuzzy clustering algorithms to form new robust algorithms.

Keywords: sample-weighted; robust; fuzzy c-means clustering

1. Introduction

The fuzzy c-means (FCM)[1] clustering algorithm is one of the most popular fuzzy clustering algorithms and is widely used in pattern recognition and image processing and so on for its simplicity and low complexity. FCM is based on the least-squared error clustering criterion and the fuzzy memberships of each data point for all classes sum to 1 by the probabilistic constraint. But the fuzzy memberships do not always correspond to the intuitive concept of the degree of belong or compatibility and thus FCM is very sensitive to noises or outliers. To solve these problems exist in FCM, Krishnapuram and Keller have proposed the Possibilistic C-Means (PCM)[2] algorithm in which the constraint of fuzzy memberships in

* Tel.:+86-151-621-16313.
E-mail address: hanxudonghxd@126.com.

1876-6102 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Singapore Institute of Electronics
doi:10.1016/j.egypro.2011.11.563

FCM was abandoned, the concept of typicality was proposed and a new objective function was constructed. PCM is robust against noises to some extent, but tends to generates coincident clusters[3]. To overcome the disadvantages, Pal and so on proposed the Possibilistic Fuzzy C-Means (PFCM)[4], and Zhang and Leung proposed the Improved Possibilistic C-Means (IPCM)[5] algorithm. These two algorithms generate both fuzzy memberships and possibilistic memberships and solve the problem of coincidence well but are not robust enough in noisy circumstances.

To enhance the robustness to noises or outliers, this paper proposes a new robust fuzzy c-means clustering algorithm in which weights for samples are used. These weight values reflect the strength of fuzzy memberships of a sample in any cluster. The influence of the outliers and noises can be reduced in the clustering process by assigning lower weight values to them, so the noise resistance is achieved.

2. Fuzzy C-Means Clustering Algorithm

Given an unlabeled data set X={x_1, x_2,…,x_n }, find the partition of X into 1<c<n fuzzy subsets by minimizing the following objective function

$$J_{FCM}(X,U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d^{2}(v_i,x_j) \tag{1}$$

Here $0\leq u_{ij}\leq 1$,and $\sum_{i=1}^{c}u_{ij}=1$, j=1,2…n. d(v_i, x_j) denotes the distance of x_j and v_i, usually, we use the Euclidean distance. $m\in(1,\infty)$ and is constant. For m→1, we have for the membership degrees u_{ij}→0/1, so the classification tends to be crisp. If m→∞, then u_{ij}→1/c, where c is the number of clusters.

Solving the minimization problem (1) by the alternating optimization technique gives the following equations:

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}(\frac{d_{ij}}{d_{kj}})^{\frac{2}{m-1}}} \tag{2}$$

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_k}{\sum_{j=1}^{n}u_{ij}^{m}} \quad \forall i \tag{3}$$

It is easy to understand and implement FCM, and FCM has low complexity and runs fast but has a lot of drawbacks, such as that it is sensitive to initial values and noises, and easy to get a local optimum solution. This paper overcomes the defect of sensitivity to noises by calculating the weight values for samples and is able to get more accurate results.

3. Sample-Weighted Robust Fuzzy C-Means Clustering Algorithm

3.1. Methods of calculating the weight values for samples

The traditional clustering algorithms treat all samples equally, which means every sample plays the same role in the clustering process. In this way, the noises and outliers may have great impacts on the final clustering result which is not reasonable. To avoid this problem, [6] introduced the concept of weights for samples. By assigning small weight values to outliers, the participation degree of outliers was reduced. [6] applied this method to PCM and obtained good results, but might still generate coincident clusters.

There have been many ways to calculate the weight values for samples, such as the improved climbing algorithm proposed by Chiu and Cheng in [7]. Dave and Krishnapuram also discussed this method in [8]. Using this approach, the equation for the weight values becomes

$$\varphi_j = \sum_{k=1}^{n} \exp(-\alpha \left\| x_j - x_k \right\|^2) \tag{4}$$

Here α is a suitable constant and j=1, 2, …, n.

Another similar method for determining the weight values was proposed by Schneider in [6]. The weights were calculated using the equation

$$\varphi_j = \sum_{i=1}^{c} \exp(-\alpha \left\| x_j - v_i \right\|^2) \tag{5}$$

where $\alpha>0$ is a suitable constant and j=1,2,…,n. v_i is the prototype of the class i. The latter method takes the proximity of the cluster centers into account, but is sensitive to the centers. If the cluster centers and the actual centers differ a lot, the weight values may be inaccurate and meaningless. Good cluster centers are usually not easy to get, so the equation (4) is adopted in this paper.

3.2. The Description of the Sample-Weighted Robust Fuzzy C-Means Clustering Algorithm

The objective function in the sample-weighted robust fuzzy c-means clustering algorithm is as follows:

$$J(X,U,V) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \varphi_j d^2(x_j, v_i) \tag{6}$$

where ψ_j is the weight value of x_j, and is calculated by the equation (4).

To minimize equation (6) under the constraint of U, we have

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{m-1}}} \tag{7}$$

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} \varphi_j x_k}{\sum_{j=1}^{n} \varphi_j u_{ij}^{m}} \quad \forall i \tag{8}$$

The fully description of the sample-weighted robust fuzzy c-means clustering algorithm is as follows and is called the SWFCM algorithm.

Initialization

1) fix c, m, α, $n>c>1, +\infty>m>1, \alpha>0$;
2) Set iteration counter r=1and maximum iteration r_{max};
3) Set the threshold value ε;

```
4)  Initialize the cluster centers $V^{(0)}$ randomly;
5)  Calculate the weight value for each sample using the equation (4);
Repeat
    Step 1  Update membership $U^{(r-1)}$ by the equation (7);
    Step 2  Update membership $V^{(r)}$ by the equation (8);
    Step 3  Increment r;
Until  ($||V(r)-V(r-1)||< \varepsilon$) or $r> r_{max}$.
```

3.3. Time Complexity

We assume that N stands for the number of samples, C stands for the number of clusters and L for the iterations. The complexity for computing the weights is O(N*N). The complexity for computing the partition matrix is O(CN), and the complexity for computing the prototypes is O(CN) at each iteration. So the complexity of SWFCM is O(N*N)+O(CNL) +O(CNL)=O(N(N+CL)). The time complexity of FCM is O(NCL), The experiment below shows that SWFCM and FCM consume about the same time when N is not too big.

4. Experiments

In order to test the algorithm proposed by this paper, We implement FCM, SWFCM, PCM, IPCM and PFCM algorithms on some datasets and compare their clustering results to illustrate the robustness of SWFCM. For all datasets we set the following parameters: $\varepsilon=0.00001$, $r_{max}=200$, m=2, p=2, a=1, b=1, $\alpha=1.0/0.4$.

4.1. IRIS dataset

Table 1 shows the clustering results of FCM, SWFCM, PCM, PFCM and IPCM on IRIS dataset. The data in the column of Errors lists the numbers of data points which are classified wrongly and the number in the brackets is counted according to the classification made by the fuzzy memberships of the data points and the number outside the brackets is counted according to the classification made by the typicalities or the possibilistic memberships. The Center Deviation calculates the sum of the squared Euclidean distance between the result centers and the actual centers which reflects the accuracy of the result centers.

PCM nearly generates coincident clusters, and the number of errors is big and so as its Center Deviation. SWFCM consumes similar time with FCM and gets the smallest errors and the least center deviation which demonstrates the SWFCM is able to get good clustering results when the noises do not exist or the outliers are few.

Table 1. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on the IRIS dataset

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	24	0.015	(16)	0.08
SWFCM	27	0.015	(12)	0.05
PCM	37	0.024	50	1.37
PFCM	23	0.031	14(13)	0.04
IPCM	99	0.109	15(71)	0.16

Table 2. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on the IRIS dataset with 10 noisy points

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	24	0.015	(16)	0.20
SWFCM	28	0.016	12	0.05
PCM	22	0.016	50	1.36
PFCM	62	0.047	11(12)	0.60
IPCM	109	0.109	12(60)	0.12

Table 3. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on the IRIS dataset with 20 noisy points

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	34	0.015	(19)	0.90
SWFCM	33	0.016	12	0.05
PCM	21	0.031	57	1.22
PFCM	30	0.047	23(11)	0.48
IPCM	99	0.110	12(52)	0.12

Table 4. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on the IRIS dataset with 30 noisy points

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	35	0.015	(50)	1.52
SWFCM	23	0.016	12	0.05
PCM	20	0.031	54	1.27
PFCM	22	0.032	41(8)	0.31
IPCM	96	0.11	12(50)	0.11

Table 5. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on the IRIS dataset with 40 noisy points

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	33	0.016	(50)	1.54
SWFCM	29	0.016	12	0.05
PCM	20	0.031	49	1.40
PFCM	22	0.047	36(11)	0.44
IPCM	98	0.135	12(45)	0.11

4.2. IRIS dataset with noises

In order to test the robustness of the SWFCM, We add 10, 20, 30, 40 noisy points to the IRIS dataset and test the clustering results and draw some comparisons. The noisy points are random and uniformly distributed over the region [0,10]×[0,10]. Table 2,3,4,5 are the clustering results.

From the results in Table 2,3,4,5, we can conclude that many clustering algorithms are obviously influenced by noisy data points. FCM generates local minimum solutions and errors become more and the Center Deviation increases. PCM still generates coincident clusters, and the results are poor. In noisy circumstances the errors and the Center Deviation in PFCM and IPCM tend to grow and their running time is long. But SWFCM can obtain almost the same good results in difference noisy circumstances with the fewest errors and the least center deviation which shows the robustness to noisy points of SWFCM.

4.3. synthetic data sets

In order to test the effect of the proposed method on larger datasets, this paper does the experiment on the synthetic data sets. We generate a dataset which is nearly the same as that in [5]. There are three clusters and the data points in each cluster are normally distributed over two-dimensional space. Their respective means are (1, 0), (3, 0), (5, 0), and their respective covariance matrices are

$$\begin{bmatrix} 0.4 & 0.0 \\ 0.0 & 0.8 \end{bmatrix}, \begin{bmatrix} 0.6 & 0.0 \\ 0.0 & 1.2 \end{bmatrix} \text{and} \begin{bmatrix} 0.4 & 0.0 \\ 0.0 & 0.8 \end{bmatrix}.$$

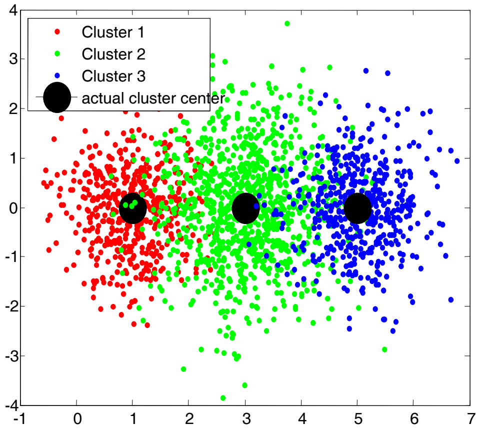

Fig. 1. Data points in X2000 data set. The actual cluster centers are (1, 0), (3, 0), (5, 0)

Table 6. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on X2000 data set

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	50	0.219	286	0.034
SWFCM	28	0.656	306	0.1124
PCM	48	0.391	1000	8.026
PFCM	23	0.363	332(336)	0.363
IPCM	117	1.453	313(335)	0.187

There are 1000 data points in the middle cluster and there are 500 data points in each of the other two clusters. Fig.1 also shows the actual cluster centers. The clustering results are shown in Table 6.

The results show that PCM still nearly generates coincident clusters, the errors and the Center Deviation in IPCM are good but consumes too much time. Other clustering algorithms work well.

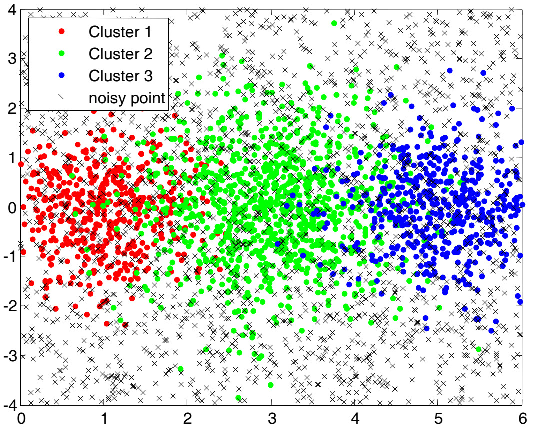

Fig. 2. Data points of X2000 and 2000 noisy points

Table 7. Clustering results of FCM, SWFCM(our proposed method), PCM, PFCM and IPCM on the data set of X2000 with 2000 noisy points

Algorithms	Iterations	Time(s)	Errors	Center Deviation
FCM	359	3.27	768	4.25
SWFCM	156	3.55	424	0.46
PCM	38	4.01	1000	8.00
PFCM	172	5.51	523(584)	1.99
IPCM	225	8.25	303(377)	0.14

Fig.2 shows a new data set constructed based on the X_{2000} by adding 2000 noisy points which are random and uniformly distributed over the region [0, 6] × [-4, 4]. The clustering results on the noisy X_{2000} are shown in Table 7.

The experimental results show that in noisy environment FCM and PCM generate more errors and more center deviations, and IPCM and SWFCM obtain fewer errors and less center deviation which reflects the robustness feature but IPCM consumes much longer time. FCM and SWFCM consume the shortest time but SWFCM generates much fewer errors and much less center deviation. On the whole SWFCM is robust enough against noisy points and consumes less time.

5. Conclusions

Many existing clustering algorithms are sensitive or are not robust enough to noises and outliers. To solve this problem, this paper proposed a sample-weighted robust fuzzy c-means which modified and improved the fuzzy c-means by adding the concept of sample weights. The experimental results on various datasets have shown the clustering algorithm proposed by this paper is very robust to noises and outliers and is able to get high clustering accuracy with not too much time. Future work includes applying the sample-weights into other fuzzy clustering algorithms.

Acknowledgment

The authors would like to thank postgraduates and teachers who discussed with them and provided them with many useful ideas and helpful suggestions. The authors have also gained a lot of guidance from senior teachers in their school and much encouragement from their friends. At this point, they want to express their sincere gratitude to all helpers!

References

[1] J C Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. New York : Plenum Press,1981.

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

[3] M Barni, V Cappellini. A Mecocci. Comments on "A possibilistic approach to clustering" [J]. IEEE Trans Fuzzy Systems, 1996, 4(3):393-396.

[4] N R Pal, K Pal, J C Bezdek. A possibilistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Systems, 2005, 13(4): 517-530.

[5] J S Zhang, Y W Leung. Improved Possibilistic C-Means Clustering Algorithms. IEEE Trans Fuzzy Systems, 2004, 2(12): 209-217.

[6] A Schneider. Weighted Possibilistic c-means Clustering Algorithms. IEEE Trans Fuzzy Systems, 2000, (1): 176-180.

[7] S Chiu, J J Cheng. Automatic role generation of fuzzy rule base for robot arm posture selection. Proceedings of NAFIPS Conf., San Antonio, TX,1994: 436-440.

[8] R Krishnapuram, R Dave. Robust clustering methods: A unified view. IEEE Trans Fuzzy Systems, 1997, (5): 270-293.

[9] J C Bezdek, J M Keller, R Krishnapuram, et al. Will the Real Iris data stand up? IEEE Trans Fuzzy System, 1999, 7(3): 368-369.