Semi-Supervised Possibilistic *Fuzzy c-Means* Clustering Algorithm on Maximized Central Distance

Li-Liu
School of IOT Engineering
Jiangnan University
Wuxi, China
liuli_jiangnan@163.com

Xiao-Jun Wu
School of IOT Engineering
Jiangnan University
Wuxi, China
xiaojun _wu_jnu @163.com

***Abstract*—Abandoning the constraint conditions of memberships in traditional fuzzy clustering algorithms, such as Fuzzy C-Means (FCM), Possibilistic Fuzzy c-Means (PCM) is more robust in dealing with noise and outliers. A small amount of labeled patterns guiding the clustering process are easy to be obtained in practical applications. In this study, a novel semi-supervised clustering technique titled semi-supervised possibilistic clustering (sPCM) is proposed. Because the PCM algorithm is easy to fall into identical clusters, we introduce the center maximization to overcome this difficulty. The proposed algorithm makes distance between different classes as far as possible, which can avoid identical clusters. The experimental results demonstrate that the accuracy of the proposed sPCM algorithm has been improved, making algorithm more robust by inheriting the characteristics of PCM.**

Keywords-PCM, maximized central distance, semi-supervised clustering, robustness

I. INTRODUCTION

Clustering is the process of distinguishing and classifying physical or abstract objects according to the similarity between them. Traditional clustering algorithms can be divided into two categories: unsupervised clustering and semi-supervised clustering. The algorithms that make use of unlabeled data and available labeled patterns are termed semi-supervised clustering algorithms[1]. The clustering task has been applied in several problems, such as biological information processing [2], text processing [3], image processing[4] and edge detection[5]. Furthermore, the semi-supervised clustering algorithms can be divided into two types according to the use of the different ways of monitoring information: 1) pairwise constraint [5, 6, 7]; 2) label information of the sample data [8, 9, 10].

The development of semi-supervised clustering based on label information of the sample data is as follows. Pedrycz [9] firstly proposed semi-supervised fuzzy c-means clustering based on label information of the sample data in 1997, which successfully used the labeled samples information. In 2004, Zhang et al.[11] retained the objective function from the FCM but replaced the Euclidean distance metric with a Gaussian Kernel-based one. And only unlabeled patterns undergo supervised learning, which means the labeled patterns never gets updated. In 2008, Li[12] proposed an improved algorithm to avoid the redundant unsupervised learning of labeled patterns in Pedrycz. In 2009, ENDO et al. [13] trained both labeled and unlabeled patterns in both unsupervised and supervised fashion, and the supervised training function was entropy-regularized. Most improved algorithms are based on the classic semi-supervised FCM algorithm proposed by Pedrycz. However, the probability constraints of membership in semi-supervised FCM makes it sensitive to noise and exceptions points, which will affect the clustering performance seriously. In this paper, we propose a semi-supervised PCM algorithm based on PCM [14] by using a small amount of labeled information. The algorithm ignores restricted condition of membership, making the membership value of noise and outliners tend to be smaller value and enhancing the noise immunity of the clustering process.

In the meantime, because sPCM algorithm abandons constraint condition of membership, making each cluster has no contact with each other, which leads to identical clusters. The idea of central distance maximization which makes the distance between each cluster center is as far as possible is introduced in sPCM to avoid the identical cluster. The experiments show that proposed sPCM algorithm has a better clustering performance and robustness compared to the semi-supervised FCM algorithm.

II. THE IDEA OF MAXIMIZED CENTRAL DISTANCE

Making the distance between different cluster centers as far as possible can avoid identical clusters in iterative phase, which is the main meaning of the idea of maximized central distance. Fig.2-1 shows the conceptual of maximized central distance. The objective function of maximized central distance is:

$$\Delta = \beta \sum_{i=1}^{C} \sum_{\substack{h=1 \\ h \neq i}}^{C} \| v_i - v_h \|^2 \qquad (2\text{-}1)$$

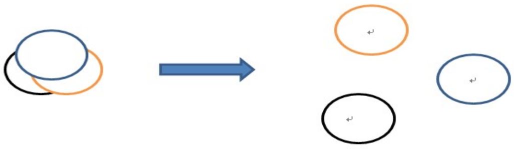


Fig.2-1. The conceptual display of maximized central distance.

III. SEMI-SUPERVISED POSSIBILISTIC CLUSTERING ALGORITHM ON MAXIMIZED CENTRAL DISTANCE（sPCM）

Given a data set $X=\{x_i \mid i=1,2,\cdots,N\}$, $x_i \in R^d$, C is the clustering number. Let $U=\{u_{ij} \mid i=1,2,\cdots,C,\ j=1,2,\cdots,N\}$ to be the membership matrix, where u_{ij} represents the membership degree of x_j corresponding to the i th cluster, $\hat{u}_{ij}$ is the typical value of labeled samples, $V=\{v_i \mid i=1,2,\cdots,C\}$ denotes C cluster centers, $d_{ik}=\|x_k-v_i\|$ represents the distance between x_j and v_i, m is the fuzzy weighted index.

Take account the full use of a small number of labeled information and to avoid the coincident cluster problem, we present semi-supervised possibilistic fuzzy c-means clustering algorithm based on maximized central distance to improve the performance of clustering.

The objective function of sPCM can be described as follows:

$$J=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}d_{ij}^{2}+\sum_{i=1}^{C}\eta_i\left(\sum_{j=1}^{N}\left(1-u_{ij}\right)^{m}\right)+ \alpha\sum_{i=1}^{C}\sum_{j=1}^{N}\left(u_{ij}-\hat{u}_{ij}\right)^{m}d_{ij}^{2}-\beta\sum_{i=1}^{C}\sum_{\substack{h=1\\h\neq i}}^{C}\|v_i-v_h\|^2 \qquad (3\text{-}1)$$

where the fuzzy weighted index $m=2$, α denotes a scaling factor used to maintain the balance between supervised and unsupervised component, β is the coefficient of center maximization item, η_i is a penalty factor. In (3-1), the first two items are PCM items, the third item is semi-supervised item, and the last item denotes maximized central distance.

Minimizing the objective function by Lagrangian multipliers, the updating equation of membership and cluster center can be expressed by:

$$u_{ij}=\frac{\alpha\hat{u}_{ij}d_{ij}^{2}+\eta_i}{\left(1+\alpha\right)d_{ij}^{2}+\eta_i} \qquad (3\text{-}2)$$

$$v_i=\frac{\sum_{j=1}^{N}u_{ij}^{2}x_j+\alpha\sum_{j=1}^{N}\left(u_{ij}-\hat{u}_{ij}\right)^{2}x_j-\beta\sum_{\substack{h=1\\h\neq i}}^{C}\left(v_h\right)}{\sum_{j=1}^{N}u_{ij}^{2}+\alpha\sum_{j=1}^{N}\left(u_{ij}-\hat{u}_{ij}\right)^{2}-\beta(C-1)} \qquad (3\text{-}3)$$

where $d_{ij}^{2}=\left(x_j-v_i\right)\left(x_j-v_i\right)^{T}$.

The algorithm of sPCM is described as follows.

1. Set the clustering number C, parameter $m>0$, $\eta_i>0$, $\alpha>0$, $\beta>0$,threshold $\varepsilon=0.001$, the maximal number of iterations *t_max=100*, randomly initialize cluster centers v_i and the typical values of labeled patterns $U\_label=\{\hat{u}_{ij}\}$.
2. Compute the partition matrix $U(t)$ by (3-2).
3. Compute the cluster center matrix $V(t)$ by (3-3).
4. Repeat step 2 to step 3, until the termination criterion is satisfied.

The properties of sPCM are:

(1) X_u denotes unlabeled pattern set, X_l denotes labeled pattern set, and $\alpha_0=\frac{|X_l|}{|X_u|}$, when $\alpha\geq\alpha_0$, the clustering performance of sPCM is optimal;

(2) when $|X_u|>>|X_l|$, the performance of sPCM degenerates into unsupervised PCM.

IV. EXPERIMENTAL RESULTS

In this section, numerical experiments are conducted on artificial and UCI standard data sets to investigate the performance of sPCM. The rand index (RI) and the normalized mutual information (NMI) are used for revaluating the performance of the proposed sPCM algorithm. Both RI and NMI take the value within the interval between 0 and 1. The higher the values, the better the clustering performance[15].

In order to reflect the fairness of the comparison, we fixed the parameters used in our experiments as follows: the maximal number of iterations *t_max*=100, parameter *m*=2, $\beta=0.01$ and the threshold $\varepsilon=0.001$.The principle of labeled patterns selected are as follows: assuming that the category properties of labeled patterns are known in advance, the membership of labeled pattern x_j is defined as $\hat{u}_{ij}=1$, and the membership of unlabeled pattern x_j is defined as $\hat{u}_{ij}=0$.

A. Experimental analysis of noise immunity

In order to support that the proposed algorithm has overcome the noise sensitivity of sFCM, we conduct an example with a simple artificial data set. We denote $\{x_1,x_2,\ldots,x_{10}\}$ by X_{10}, and $\{x_1,x_2,\ldots,x_{10},x_{11},x_{12}\}$ by X_{12}, x_{11} in Fig.4-1 is outliner and x_{12} in Fig.4-1 is noise point, X_{10} has two diamond shaped clusters with five points both on the left and right side of the axis. x_{11} and x_{12} are equidistant from all corresponding pairs of points in the two clusters. Designated x_1 and x_6 for the labeled samples, and the typical values of them are defined:

$\hat{u}_{11}=1$, $\hat{u}_{21}=0$, $\hat{u}_{16}=0$, $\hat{u}_{26}=1$.parameter $\alpha=1$, $\beta=0.01$, m=2, the number of cluster C=2. Figure 4-1 shows the distribution of original data set X_{12}. Table 4-1 shows how outliner and noise point affect partitions found by sFCM and sPCM.

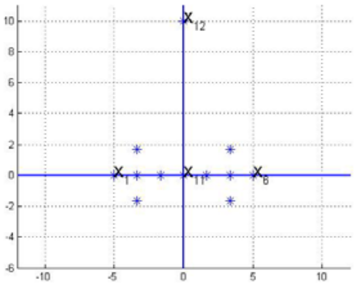


Fig. 4-1 the distribution of original data set X_{12}

Table 4-1 Memberships from sFCM and sPCM in X_{12}

	DATA		sFCM		sPCM	
Pt	x	y	Cluster1	Cluster2	Cluster1	Cluster2
1	-5	0	0.15	0.85	0.76	0.03
2	-3.34	1.67	0.12	0.88	0.30	0.05
3	-3.34	0	0. 10	0.90	0.79	0.05
4	-3.34	-1.67	0.18	0.82	0.30	0.05
5	-1.67	0	0.17	0.83	0.22	0.09
6	1.67	0	0.84	0.16	0.04	0.99
7	3.34	1.67	0.97	0.03	0.02	0.22
8	3.34	0	0.99	0.01	0.03	0.38
9	3.34	-1.67	0.91	0.09	0.02	0.21
10	5	0	0.94	0.06	0.01	0.12
11	0	0	**0.50**	**0.50**	**0.08**	**0.08**
12	0	10	**0.50**	**0.50**	**0.01**	**0.01**

It can be seen from Table 4-1 that the typicality values of outliner x_{11} and noise point x_{12} assigned by sFCM are both 0.50. This significantly affects the estimation of the cluster centers. The proposed sPCM algorithm gives very low memberships for the two points of x_{11} and x_{12} as desired. The reason is that the sPCM relaxes the column constraint $\sum_{i=1}^{C} u_{ij}=1, \forall j$ for fuzzy partitions, which has circumvented the counterintuitive results just displayed. As a result, the cluster centers are virtually unchanged. From the analysis above, sPCM prototypes are less influenced by the noise and outliners than sFCM.

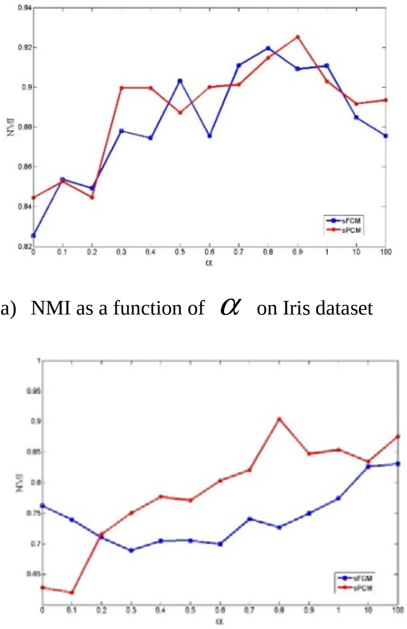
B. UCI datasets

The performance of the proposed sPCM algorithm has been evaluated and compared with four clustering algorithms using two UCI datasets. Iris data set contains 150 samples with 3 classes, and each sample has 4 feature values. Glass data set contains 214 samples with 3 classes, and each sample has 10 feature values. The parameters are set as: $\alpha=1$, $\beta=0.01$. Table 4-3 shows the performance comparison of four algorithms on Iris and Glass datasets, and the number of labeled patterns is 0.1 of the total number of patterns.

Table 4-3 the performance comparison of four algorithms on Iris and Glass datasets

Index		FCM	sFCM	PCM	sPCM
NMI(Iris)	Mean	0.8502	0.8994	0.8404	**0.9076**
	std	0.0990	0.0322	0.0672	**0.0187**
RI(Iris)	Mean	0.9101	0.9586	0.9226	**0.9633**
	std	0.0948	0.0140	0.0682	**0.0109**
NMI(Glass)	Mean	0.7621	0.8080	0.7216	**0.8131**
	std	0	0.0241	0.1478	**0.0417**
RI(Glass)	Mean	0.8545	0.8613	0.8369	**0.8991**
	std	0.1110	0.0122	0.2282	**0.0284**

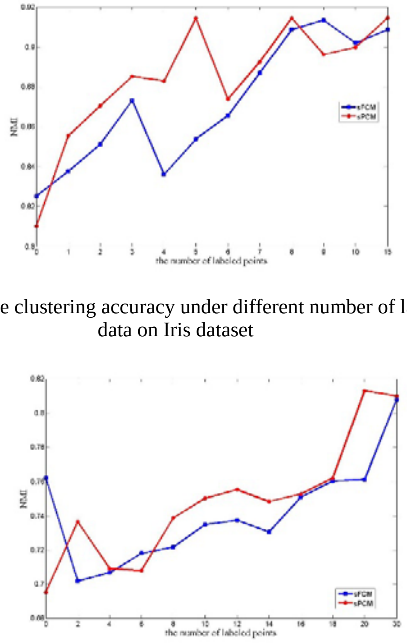
Fig. 4-5 shows NMI as a function of α on Iris and Glass datasets, and the number of labeled patterns is 0.1 of the total number of patterns. Fig. 4-6 shows the clustering accuracy under different number of labeled data on Iris and Glass datasets when $\alpha=1$.



(a) NMI as a function of α on Iris dataset

(b) NMI as a function of α on Glass dataset

Fig. 4-5 NMI as a function of α on Iris and Glass datasets



(a) the clustering accuracy under different number of labeled data on Iris dataset

(b) The clustering accuracy under different number of labeled data on Glass dataset

Fig. 4-6 the clustering accuracy under different number of labeled data on Iris and Glass datasets

Table 4-3 shows the performance comparison of four algorithms on Iris and Glass datasets. The clustering accuracy of semi-supervised sFCM and sPCM is much higher than unsupervised FCM and PCM due to the guiding role of a small amount of labeled information. By introducing the idea of center maximization, which makes distance between each class as far as possible, the performance of the sPCM is superior than sFCM and also effectively avoids identical clusters.

Fig. 4-5(a) shows NMI as a function of α on Iris dataset. The trends of sFCM and sPCM are same in the whole range, and the optimal values of the two algorithms are both obtained when $\alpha = 0.9$. In most cases, the clustering performance of sFCM is lower than sPCM with the same value of α . The clustering performance begins to decline with increasing of α .Fig. 4-5(b) shows NMI as a function of α on Glass dataset. The optimal value of sPCM is obtained when $\alpha = 0.8$, and the clustering performance begins to stabilize with increasing of α . The optimal value of sFCM is obtained when $\alpha = 10$, and the clustering performance begins to stabilize with increasing of α , but its clustering accuracy never exceeds sPCM.

Fig. 4-6(a) shows the clustering accuracy under different number of labeled data on Iris dataset. The clustering performance of sFCM is lower than sPCM with increasing of the number of labeled patterns in the most cases, and the clustering performance begins to stabilize with increasing of the number of labeled patterns. Fig. 4-6(b) shows the clustering accuracy under different number of labeled data on Glass dataset. The clustering performance of sFCM is higher than sPCM when the number of labeled patterns is zero. With increasing of the number of labeled patterns, the clustering performance of sPCM gradually exceeds sFCM, and the clustering accuracy tends to stabilize when the number of labeled patterns increased to a certain amount.

V. CONCLUSIONS

In practical applications, most of the datasets can get a small amount of labeled information easily. Many studies have shown that a small amount of labeled information is very valuable in guiding the clustering in the semi-supervised algorithms. Thus, a novel semi-supervised PCM algorithm sPCM is proposed in this paper. However, sPCM still prone to generate the coincident clusters like PCM. Introducing the idea of maximized central distance has successfully avoided above weaknesses. The experimental results indicate that the proposed sPCM algorithm has better clustering performance and is more robust than sFCM algorithm when the data set contains noise points and outliners. Even if for the data set without noise points and outliners, sPCM algorithm still has the same clustering performance as sFCM algorithm. Most data sets in real-world applications in the presence of noise and outliners, which verifies the proposed algorithm is more applicable. However, the run time of sPCM has increased, how to eliminate the weakness to gain better clustering results have not been solved yet.

REFERENCES

[1] ENDO Yasunori, HAMASUNA Yukihiro, et al. On Semi-Supervised Fuzzy c-Means Clustering. FUZZ-IEEE, Korea, 1119-1124, 2009.

[2] Tari L, Baral C, Kim S. Fuzzy C-Means Clustering with Prior Biological Knowledge. Journal of Biomedical Informatics, 74-81, 2009.

[3] Huang Ruizhang, Lam W. An active learning frame work for semi-supervised document clustering with language modeling. Data &Knowledge Engineering, 49 – 67, 2008.

[4] Chang Hong, Yeung D Y. Locally Linear Metric Adaptation with Application to Semi-Supervised Clustering and Image Retrieval. Pattern Recognition, 1253- 1264, 2006.

[5] Wagstaff K, Cardie C, Rogers S, et al. Constrained K -means Clustering with Background Knowledge . Proc of 18th International Conference on Machine Learning. San Francisco, USA, 577- 584, 2001.

[6] Demiriz A, Bennett K P, Embrechts M J. Semi-Supervised Clustering Using Genetic Algorithm s. Proc of the Artificial Neural Networks in Engineering Conference. New York, USA, 809- 814, 1999.

[7] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means Clustering with Background Knowledge. Proc of 18th International Conference on Machine Learning. San Francisco, USA, 577- 584, 2001.

[8] Bensaid A M., Bezdek J C. Clarke L.P. Partially Supervised Clustering for Image Segmentation , Pattern Recognition,

[9] 859-871, 2009.

[10] Pedrycz J C. Fuzzy Clustering with Partial Supervision, IEEE Transactions on Systems, Man and Cybernetics, 787-795, 1997.

[11] Basu S, Banerjee A. R. J. Mooney, Semi-supervised Learning by Seeding, Proc.ICML2002,19-26, 2002.

[12] D. Zhang, K. Tan, and S. Chen, “Semi-supervised kernel-based fuzzy c-means,” Lecture Notes in Computer Science: Neural Information Processing. 1229–1234, 2004.

[13] C. Li, L. Liu, and W. Jiang, “Objective function of semi-supervised fuzzy c-means clustering algorithm,” IEEE International Conference on Industrial Informatics. 737–742, 2008.

[14] Y. Endo, Y. Hamasuna, M. Yamashiro, and S. Miyamoto, “On semisupervised fuzzy c-means clustering,” IEEE International Conference on Fuzzy Systems.2009.

[15] A Possibilistic Fuzzy c-Means Clustering Algorithm. Nikhil R. Pal, Kuhu Pal, James M. Keller, et al. IEEE TRANSACTIONS ON FUZZY SYSTEMS. 517-530,2005.

Published by Atlantis Press, Paris, France.