Semi-Supervised MEC Clustering Algorithm on Maximized Central Distance

Ziyang Yao
College of Thing of Internet, Wuxi Professional College of Science and Technology, Wuxi, P.R. China
Email: yzy_wx@163.com

***Abstract*—In the field of pattern recognition, the traditional supervised learning methods and unsupervised learning methods are not always suitable for the practical applications. In some applications, the data obtained is neither no-information-given nor all-information-given. In addition, the data obtained usually contains some noises due to many interference factors in practical collection procedure and these noises are of great influence on the traditional clustering methods. In order to overcome the two problems mentioned above, based on the classical Maximal Entropy Clustering (MEC), we propose a semi-supervised MEC algorithm based on the maximized central distance and the compensation term for membership, i.e., CM-sSMEC algorithm. The experimental results on benchmarking UCI data sets show that it has a better performance than the traditional unsupervised clustering method.**

***Index Terms*—Semi-supervised, Maximum Center, Maximum Entropy Clustering, Noise Immunity**

I. INTRODUCTION

In the field of pattern recognition, data mining methods generally contain supervised learning methods and unsupervised learning methods. The traditional supervised learning methods is usually classified into the field of classification. The famous algorithm is the model of SVM and its related improved models[1-2]. In this study, the unsupervised learning method is focused on. The clustering technique is the main method of unsupervised learning, such as the classic k-means algorithm [3-6] and the FCM algorithm [7-12]. In 1995, a novel clustering algorithm called maximum entropy clustering algorithm (MEC) is proposed by fuzzy clustering technology[13]. The MEC has more clear physical meaning with concise mathematical expression compared woth the previous clustering algorithms. However, there still exit some respective problems in the MEC algorithm and other related improved algorithms. Since all known data information in the data collection process is not adequate, it makes the classification method wait for the processing of the data information adequate for modeling. As well known, the clustering algorithm does not need to know the sample information, such as the data labels. Therefore, when we face the above data mining task, the clustering algorithm is more adaptive than classification algorithm. Especially in the initial stage of data analysis, the data scale is very small and the data information is also very limited. The clustering algorithm can be used to process this situation, but the traditional clustering algorithms also ignored the limited known information which is very significant during the data analysis process.

Through the above analysis, unsupervised clustering algorithm which is more efficient to process data in the initial processing is used as the natural model in this study. And a specific clustering algorithm called MEC algorithm is studied. In recent years, some improved MEC algorithms have been proposed, such as MECA algorithm [14], FBACN algorithm[15]. Although these related works have some good practical values within a certain range of applications, they still belong to field of the unsupervised learning. Therefore, they do not use the known information effectively. In such scene, we propose a new compensation term of membership for the classical MEC. On the other hand, there must be some interference-information in the known data information obtained for the real-world environment, and the interference-information will affect the final clustering result. Therefore, a new mechanism of maximized central distance is proposed to maximize the distances of different cluster centers. It will weaken the impact of interference-information. By introducing these two novel items into the traditional MEC algorithm, we propose a robust semi-supervised MEC clustering algorithm based on maximized central distance, i.e., CM-sSMEC algorithm. The algorithm not only can effectively utilize the existing sample information, but also weaken the influence of interference information, which can effectively enhance the anti-interference ability of the algorithm.

II. RELATED WORK

A. Basic Principle Described of MEC Algorithm

Given a dataset $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots \boldsymbol{x}_N\}$, the objective function of the classic MEC algorithm is as follows:

$$J_{MEC}(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\left\|\boldsymbol{x}_j-\boldsymbol{v}_i\right\|^2+\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij} \quad (1)$$

$$\text{s.t. } \mu_{ij}\in[0,1], 1\le i\le C, 1\le j\le N, \sum_{i=1}^{C}\mu_{ij}=1$$

where C and N represent the clustering number and the number of the samples respectively; γ is a scale parameter; $d_{ij}^2=\left\|\boldsymbol{x}_j-\boldsymbol{v}_i\right\|^2$ represents the distance between

Corresponding author: Ziyang Yao

doi:10.4304/jcp.8.10.2570-2574

$\boldsymbol{x}_j$ and $\boldsymbol{v}_i$; u_{ij} represents the membership degree of $\boldsymbol{x}_j$ corresponding to the i th cluster; $\boldsymbol{U}$ denotes the partition matrix $\boldsymbol{U} \in R^{N\times C}$ constituted by μ_{ij}^m ; v_i is the cluster center of the i th cluster; $\boldsymbol{V}$ denotes the cluster center matrix constituted by $\boldsymbol{v}_i$.

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

$$\mu_{ij} = \frac{\exp(-\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2 / \gamma)}{\sum_{k=1}^{C} \exp(-\left\|\boldsymbol{x}_j - \boldsymbol{v}_k\right\|^2 / \gamma)} \quad (2)$$

$$i = 1,2,\ldots,C;\ j = 1,2,\ldots,N.$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{N} \mu_{ij} \boldsymbol{x}_j}{\sum_{j=1}^{N} \mu_{ij}} \quad i = 1,2,\ldots,C \quad (3)$$

B. The Algorithm of MEC

The algorithm of MEC can be described as follows.

Input: Given a dataset $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_N\}$, with the number of samples N .

Initialize: Set the clustering number C , the maximal number of iterations M , threshold ε , parameter γ ; randomly initialize membership matrix $\boldsymbol{U}(t)$ and set $t = 1, 1 \le t \le M$.

Iteration:

Step 1: Compute the cluster center $\boldsymbol{V}(t+1)$ by (2).

Step 2: Compute the partition matrix $\boldsymbol{U}(t+1)$ by (3).

Step 3: Repeat step 1 to step 2, until the termination criterion is satisfied.

Output: Output the partition matrix $\boldsymbol{U}$ and the cluster center $\boldsymbol{V}$.

C. Motivation

Question1 (*The low utilization rate of the sample information*): In the real-world, we can only obtain a small amount of sample information, neither no-information-given nor all-information-given. In this case, the unsupervised traditional MEC algorithm ignores the known information, as shown in Fig.1. In Fig.1, the colored part represents the known information, the red color and blue color represent different classes respectively. How to effectively use the known information is becoming the focus of attention.

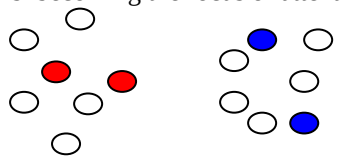

Figure 1. N The schematic diagram of known information

Question2(*The presence of interference of known information*): Due to the weak anti-interference ability of the current sensor, the collected samples usually contain noise, which indirectly causes the interference-information in the known information. Generally, the interference-information is defined as misclassified sample information. For example, sample $\boldsymbol{x}$ should belong to Class A, but the information obtained is assigned to Class B, as shown in Fig.2. Therefore, if this information is directly used into the clustering procedure the final result will be serious interferential.

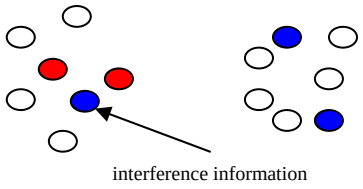

Figure 2. The schematic diagram of interference information

III. SEMI-SUPERVISED MEC ALGORITHM ON MAXIMIZED CENTRAL DISTANCE(CM-SSMEC)

For the lower utilization of known information (question 1), through the introduction of the compensation term to the membership the problem of low utilization of sample information will be solved. The structure form of the compensation term is as follow:

$$\Theta = \sum_{i=1}^{C} \sum_{j=1}^{N} (\mu_{ij} - \hat{\mu}_{ij}) \quad (4)$$

where the membership μ_{ij} is unknown and $\hat{\mu}_{ij}$ is known which plays a supervisory role. $\hat{\mu}_{ij}$ =0 represents the label of sample x_{ij} is unknown.

Meanwhile, for the problem of the known information with interference (question 2), in order to avoid the known information attracting the cluster centers overlapping, we introduce the maximum center item to maximize between-cluster separation, which will weaken the impact of noise-points. The objective function of maximum center item is defined as follows.

$$\Gamma = \eta \sum_{i=1}^{C} \sum_{\substack{h=1, \\ h\neq i}}^{C} \left\|\boldsymbol{v}_i - \boldsymbol{v}_h\right\|^2 \quad (5)$$

By introducing (4) and (5) into the MEC algorithm, the new objective function is presented as follows:

$$J_{\text{CM-sSMEC}}(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij} - \hat{\mu}_{ij}) d_{ij}^2 - \eta \sum_{j=1}^{N}(\mu_{ij} - \hat{\mu}_{ij}) \sum_{i=1}^{C} \sum_{\substack{h=1, \\ h\neq i}}^{C} \left\|\boldsymbol{v}_i - \boldsymbol{v}_h\right\|^2 + \gamma \sum_{i=1}^{C}\sum_{j=1}^{N} \mu_{ij} \ln \mu_{ij} \quad (6)$$

$$\text{s.t. } \mu_{ij} \in [0,1], \sum_{i=1}^{C} \mu_{ij} = 1, 1 \le i \le C, 1 \le j \le N$$

The symbolic meanings of the above equation are defined in Table I.

TABLE I.
SYMBOL DEFINITION

Symbol	Definition
C	the clustering number
N	the total number of samples
D	sample dimension
η	the coefficient of center maximization item
$\boldsymbol{x}_j$	the j th sample
μ_{ij}	the membership degree of x_j corresponding to the i th cluster
$\boldsymbol{v}_i$	the i th cluster center

A. Iterative Parameter Derivation

Minimizing the objective function by Lagrangian optimization, the updating equations of membership and cluster center can be expressed by Eqs. (7) and (8), respectivel.

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{N}(\mu_{ij}-\hat{\mu}_{ij})\boldsymbol{x}_j - \eta\sum_{j=1}^{N}(\mu_{ij}-\hat{\mu}_{ij})\sum_{\substack{h=1,\\ h\neq i}}^{C}\boldsymbol{v}_h}{\sum_{j=1}^{N}(\mu_{ij}-\hat{\mu}_{ij}) - \eta\cdot(C-1)\sum_{j=1}^{N}(\mu_{ij}-\hat{\mu}_{ij})} \tag{7}$$

$$\mu_{ij} = \hat{\mu}_{ij} + (1-\sum_{k=1}^{C}\hat{\mu}_{kj})\frac{\exp(-\frac{d_{ij}^2 - \eta\sum_{\substack{h=1,\\ h\neq i}}^{C}\|\boldsymbol{v}_i-\boldsymbol{v}_h\|^2}{\gamma})}{\sum_{k=1}^{C}\exp(-\frac{d_{kj}^2 - \eta\sum_{\substack{h=1,\\ h\neq i}}^{C}\|\boldsymbol{v}_k-\boldsymbol{v}_h\|^2}{\gamma})} \tag{8}$$

B. Description of CM-sSMEC Algorithm

Input:

Given a data set $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots\boldsymbol{x}_N\}$, with N samples.

Initialize:

Set the clustering number C, the maximal number of iterations M, threshold ε, parameter γ; randomly initialize membership matrix $\boldsymbol{U}(t)$ and $\boldsymbol{V}(t)$; set the number of the initial iteration t =1，and the maximal iteration number M. Given the known membership matrix $\hat{\boldsymbol{U}}$ and he coefficient of center maximization item η, $0<\eta<1$）.

Iteration:

Step 1: Compute the cluster center matrix $\boldsymbol{V}(t+1)$ by (7).

Step 2: Compute the partition matrix $\boldsymbol{U}(t+1)$ by (8).

Step 3: Repeat step 1 to step 2, until the termination criterion is satisfied.

Output: Output the partition matrix $\boldsymbol{U}$' and the cluster center $\boldsymbol{V}$'.

IV. EXPERIMENT

In this section, the proposed CM-sSMEC algorithm has been evaluated on several benchmarking UCI datasets. Two metrics, the rand index (RI) and the normalized mutual information (NMI) are used for evaluating the performance of the proposed CM-sSMEC algorithm. we compare the experimental results of several related algorithms, inclusing FCM, MEC and semi-supervised sSFCM [16].

TABLE II.

THE INTRODUCTION TO THE EVALUATION INDEXES

Index	Computational formula
NMI[9]	$NMI = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C}N_{i,j}\log N \cdot N_{i,j}/N_{i\cdot}N_j}{\sqrt{\sum_{i=1}^{C}N_i\log N_i/N \cdot \sum_{j=1}^{C}N_j\log N_j/N}}$
RI[9]	$RI = \frac{f_{00}+f_{11}}{N(N-1)/2}$

Both RI and NMI take the value within the interval between 0 and 1. The higher the values, the better the clustering performance. The experimental environment are shown in Table III.

TABLE III.
THE EXPERIMENTAL ENVIRONMENT

Hardware platform	Programming environment	
CPU：Core i5	Software version	System version
Frequency：2.7GHz Memory：2GB	MATLAB 7.0	Windows 7

We fix the parameters used in our experiments as follows: the maximal number of iterations M=200, threshold $\varepsilon = 1e-5$. η can be adjusted according to the data set, and need to meet $0<\eta<1$.

A. UCI Datasets

The ability to use known information of the proposed algorithm has been evaluated and compared with four clustering algorithms using two UCI datasets of Iris and wine (The specific description of the two data sets are shown in Table IV). In this experimental part, the proportions of the known sample information are 1%, 5%, 10%, 20% and 50%, respectively. Specific results are shown in Table V.

TABLE IV.
UCI DATA SET DESCRIPTION

Dataset	Number	Dimension	Number of clusters
Iris	150	4	3
Wine	178	13	3

The experimental results of Table V indicate that the proposed algorithm can efficiently use known information to guide learning. With the increasing known information, the advantages of the proposed algorithm become more obvious. When the known information is accounted for 50% of the total amount of data, the

clustering accuracy of the proposed algorithm is nearly 100%. The above observation is consistent with the actual situation. As we know, the guidance role is not obvious when the known information is limited, while when the known information is sufficient, the clustering problem has actually become a classification problem. We find that the only difference between clustering and classification is the size of the amount of known information received prior to the model. Data analysis problem is transformed into clustering problem in the face of a very small amount of known information. On the other hand, we also find that regardless of the known information is 1% or 50%, this information always can play a positive role in the clustering algorithm, and greatly enhance the accuracy of the algorithm. This study has effectively used this known information, and thus the accuracy of the proposed algorithm has be improved obviously than the other algorithms.

TABLE V.

THE PERFORMANCE COMPARISON IN THE DIFFERENT INFORMATION-RATE

Data	Information-rate	Index	FCM	MEC	sSFCM	CM-sSMEC
Iris	1%	NMI	0.8997	0.7387	0.9024	**0.9188**
		RI	0.9575	0.8797	0.9532	**0.9696**
	5%	NMI	0.8997	0.7387	0.9099	**0.9263**
		RI	0.9575	0.8797	0.9591	**0.9734**
	10%	NMI	0.8997	0.7387	0.9272	**0.9464**
		RI	0.9575	0.8797	0.9600	**0.9764**
	20%	NMI	0.8997	0.7387	0.9578	**0.9742**
		RI	0.9575	0.8797	0.9716	**0.9880**
	50%	NMI	0.8997	0.7387	0.9715	**1**
		RI	0.9575	0.8797	0.9842	**1**
Wine	1%	NMI	0.8336	0.7844	0.8405	**0.8643**
		RI	0.9331	0.9128	0.9360	**0.9498**
	5%	NMI	0.8336	0.7844	0.8547	**0.8785**
		RI	0.9331	0.9128	0.9460	**0.9598**
	10%	NMI	0.8336	0.7844	0.8811	**0.9049**
		RI	0.9331	0.9128	0.9488	**0.9726**
	20%	NMI	0.8336	0.7844	0.9113	**0.9348**
		RI	0.9331	0.9128	0.9577	**0.9815**
	50%	NMI	0.8336	0.7844	0.9309	**1**
		RI	0.9331	0.9128	0.9534	**1**

TABLE VI.

THE PERFORMANCE COMPARISON IN THE SAME INFORMATION-RATE AND DIFFERENT INTERFERENCE INFORMATION

Information_rate-Data	Error-information-rate	Index	FCM	MEC	sSFCM	CM-sSMEC
20%-Iris	1%	NMI	0.8997	0.7387	0.9414	**0.9684**
		RI	0.9575	0.8797	0.9673	**0.9903**
	5%	NMI	0.8997	0.7387	0.9349	**0.9619**
		RI	0.9575	0.8797	0.9645	**0.9815**
	10%	NMI	0.8997	0.7387	0.9255	**0.9525**
		RI	0.9575	0.8797	0.9607	**0.9787**
	20%	NMI	0.8997	0.7387	0.9016	**0.9286**
		RI	0.9575	0.8797	0.9592	**0.9762**
20%-Wine	1%	NMI	0.8336	0.7844	0.9039	**0.9294**
		RI	0.9331	0.9128	0.9584	**0.9839**
	5%	NMI	0.8336	0.7844	0.9012	**0.9267**
		RI	0.9331	0.9128	0.9536	**0.9791**
	10%	NMI	0.8336	0.7844	0.8762	**0.9117**
		RI	0.9331	0.9128	0.9472	**0.9757**
	20%	NMI	0.8336	0.7844	0.8622	**0.8977**
		RI	0.9331	0.9128	0.9429	**0.9684**

B. Experimental Analysis of Noise Immunity

In order to validate the anti-interference of the proposed algorithm, experiment analysis has been conducted on the two data sets with the rate of known information is 20% and the interference information (noise) is respectively 1%, 5%, 10%, and 20%. The detailed results are as shown in Table VI. The experimental results show that the proposed algorithm is more robust in the face of data containing the interference information. With the increase of the interference information, the accuracy of the proposed algorithm has a

smaller downward trend, but the decline is less than the increase of the interference information. It can be seen from the experimental result of Wine, the interference information increasing from 1% to 20%, and the accuracy of NMI index only fell from 92.94% to 89.77% with a decrease of about 3%. The above analysis shows that the proposed method has a good anti-interference capability.

V. CONCLUSIONS

In order to solve the weaknesses of the traditional cluster analysis methods that the known information for data analysis is not used well, the CM-sSMEC algorithm is proposed by integrating a new compensation term of membership and maximized the central distance into the classic MEC algorithm. Experimental results on Iris and Wine datasets show that the proposed method is better than the traditional clustering methods. It can not only effectively use known information to guide learning, but also to reduce the impact of interference information on clustering accuracy, thus making the CM-sSMEC algorithm more applicable in the practical application.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments that have greatly improved the quality of our manuscript in many ways.

REFERENCES

[1] F. Cai, V. Cherkassky, “Generalized SMO Algorithm for SVM-Based Multitask Learning”. *IEEE Trans- actions on Neural Networks and Learning Systems*, vol.23 no.6, pp. 997~1003, 2012

[2] K. P. Lin, M. S. Chen, “On the Design and Analysis of the Privacy- Preserving SVM Classifier,” *IEEE Trans-actions on Knowledge and Data Engineering*, vol. 23 no.11, pp.1704~1717, 2011

[3] S. Yu, etc, “Optimized Data Fusion for Kernel k-Means Clustering”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.34 no.5, pp.1031~1039, 2012

[4] J. Y. Xie, S. Jiang, W. X. Xie, X. B. Gao, “An Efficient Global K-means Clustering Algorithm,” *Journal of Computers*, vol. 6 no.2, pp.271-279, 2011

[5] J. Wu, J. Xia, J. Chen, Z. Cui, “Moving Object Classification Method Based on SOM and K-means,” *Journal of Computers*, vol.6 no.8, pp.1654-1661, 2011

[6] T. Li, Y. Chen, J. Zhang, “Logistics Service Provider Segmentation Based on Improved FCM Clustering for Mixed Data,” *Journal of Computers*, vol.7 no.11, pp.2629-2633, 2012

[7] L. O. Hall, D. B. Goldgofd, “Convergence of the Single-Pass and Online Fuzzy C-Means Algorithms,” *IEEE Transactions on Fuzzy Systems*, vol.9 no.4, pp.792~794, 2011

[8] L. Zhu, etc, “Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions,” *IEEE Transactions on Systems Man and Cybernetics*, vol.39 no.3, pp.578~591, 2009

[9] Z. H. DENG, etc, “Enhanced soft subspace clustering integrating within-cluster and between-cluster information,” *Pattern Recognition*, vol.43 no.3, pp.767~781, 2010

[10] H. Jiang, J. Gu, Y. Liu, F. Ye, H. Xi, M. Zhu, “Study of Clustering Algorithm based on Fuzzy C-Means and Immunological Partheno Genetic,” *Journal of Software*, vol.8 no.1, pp.134-141,2013

[11] Q. Niu, X. Huang, “An improved fuzzy C-means clustering algorithm based on PSO,” *Journal of Software*, vol.6 no.5, pp.873-879, 2011

[12] Q. Zhao, “The Study on Rotating Machinery Early Fault Diagnosis based on Principal Component Analysis and Fuzzy C-means Algorithm,” *Journal of Software*, vol.8, no.3, pp.709-715, 2013

[13] R. P. Li, M. Mukaidon, “A maximum entropy approach to fuzzy clustering,” *Proc. of the 4th IEEE Int’l Conf. on Fuzzy System*. Yokohama: IEEE, 1995. pp.2227~2232

[14] N. B. Karayiannis, “MECA: Maximum entropy clustering algorithm,” *IEEE World Congress on Computational Intelligence, Vol 1. Proc. of the 3rd IEEE Conf. on Fuzzy Systems, Vol 2.* Orlando: IEEE, 1994. pp.630~635

[15] C. Wei, C. Fahn, “The multisynapse neural network and its application to fuzzy clustering,” *IEEE Transactions on Neural Networks*, vol.13 no.3, pp.600~618, 2002

[16] Y. Endo, Y. Hamasuna, M. Yamashiro, and S. Miyamoto, “On semi-supervised fuzzy c-means clustering,” *IEEE International Conference on Fuzzy Systems, 2009.* pp.1119~1124

Ziyang Yao received the master's degree from Jiangnan University of China in 2005. He is currently working at Wuxi Professional College of science and technology. His research interests include technology of the Internet of things and data mining.