International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 03 – Issue 04, July 2014

A Weighted Sample's Fuzzy Clustering Algorithm With Generalized Entropy

Kai Li
Hebei university
School of mathematics and computer
Baoding, China
Email: Likai {at} hbu.edu.cn

Lijuan Cui
Hebei university
Library
Baoding, China

Abstract—**Combined with weight of samples and kernel function, fuzzy clustering method with generalized entropy is studied. Objective function for fuzzy clustering with generalized entropy based on sample weighting is obtained. Following that, fuzzy clustering algorithm with generalized entropy based on sample weighting is presented. In addition, by introducing kernel into the presented objective function, kernel fuzzy clustering algorithm with generalized entropy based on sample weighting is obtained. At the same time, some methods for determining weight of samples are analyzed. Aiming at some selected representative datasets, experiments are conducted to validate the effectiveness of presented algorithms above.**

Keywords-fuzzy clustering; generalized entropy; sample weighting; kernel

I. INTRODUCTION

Clustering is an unsupervised pattern classification method and it has been applied to pattern recognition, data analysis, biological information processing, data mining, and etc. Currently, researchers have proposed many clustering algorithms. They can be divided into five types: grid-based clustering analysis, model-based clustering analysis, division-based clustering analysis, density-based cluster analysis and hierarchical-based clustering analysis. On them, division-based cluster analysis (also known as cluster analysis method based on the objective function) is one of the commonly used methods, for example K-means clustering algorithm and Fuzzy C-means clustering algorithm. Theoretical and experimental studies show that these clustering algorithms have some shortcomings, such as data points in the data set with the same weight, the number of data points in each cluster with equal or approximately equal, shape of each cluster with spherical and no anti-noise performance. To solve these problems, researchers have proposed many different improved algorithms. Karayiannis [1] introduced entropy into fuzzy clustering and proposed fuzzy clustering algorithm based on maximum entropy. Following that, Li et al.[2], and Wagner et al.[3] combined the loss of function for data samples to cluster centers to propose maximum entropy clustering algorithm. Wei et al.[4] presented a bidirectional association fuzzy clustering network to solve the problem of fuzzy clustering. Pedrycz et al.[5] compared fuzzy C-means algorithms and their kernel fuzzy C-means algorithm. Zhang and Chen[6] presented fuzzy c-means and possible c-means algorithm based kernel. Then they modified BCFCM (Bias Corrected FCM) algorithm and proposed the KFCM algorithm with spatial constraints[7]. Yang[8] further studied kernel clustering algorithm with spatial correction to propose GKFCM, and was successfully applied to image segmentation. Kannan et al. [9], Swagatam et al.[10] studied image segmentation based on kernel fuzzy clustering. It can be seen that algorithms above does not consider different sample role in clustering process. To better improve the performance of clustering algorithm, some researchers presented fuzzy clustering method based on entropy weighting [12]. In this paper, fuzzy clustering with generalized entropy based on sample weighting is studied.

This paper is organized as follows. In section 2, we give an objective function about sample weighting for fuzzy clustering with generalized entropy and use Lagrange method to obtain membership of samples and centers of cluster. Then, fuzzy clustering algorithm with the generalized entropy based on sample weighting is given. In section 3, by introducing kernel into presented objective function, we obtain kernel fuzzy clustering algorithm with the generalized entropy based on sample weighting. In the section, we choose some commonly used datasets from UCI and artificial generated dataset to test the presented algorithms' performance. In the final section, conclusion is given.

II. OBJECTIVE FUNCTION AND FUZZY CLUSTERING WITH GENERALIZED ENTROPY BASED ON SAMPLE WEIGHTING

Let $X=\{x_1,x_2,\cdots,x_n\}$ be a data set, where $x_i\in R^s$, c is a positive integer greater than one and $m\ge 1$ is fuzzy index, $\mu_{ij}=\mu_i(x_j)>0$ is degree of membership for x_j belonging to ith cluster and $\sum_{i=1}^{c}\mu_{ij}=1$. U is a matrix which is composed of all μ_{ij}'s (i=1,2,…,c; j=1,2,…,n), V is a vector whose component consists of cluster center v_i(i=1,2,…,c). Objective function with generalized entropy's fuzzy clustering is represented as

$$J_G(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}{}^{m}\,\|x_j-v_i\|^2+\delta\sum_{j=1}^{n}(2^{1-m}-1)^{-1}(\sum_{i=1}^{c}\mu_{ij}{}^{m}-1)\,. \quad (1)$$

Based on (1), we obtain two algorithms GEFCM and KGEFCM [13]. As samples for data set have different importance in clustering process, here, weights of samples are introduced into objective function (1) above. So we obtain the following objective function on the basis of (1)

$$J_{WG}(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}w_j\mu_{ij}{}^{m}\,\|x_j-v_i\|^2+\delta\sum_{j=1}^{n}(2^{1-m}-1)^{-1}(\sum_{i=1}^{c}\mu_{ij}{}^{m}-1)\,, \quad (2)$$

where δ is an adjustable parameter, w_j is the weight of the jth data sample.

Based on objective function (2), we obtain the following optimization problem for fuzzy clustering with generalized entropy based on sample weighting

$$\begin{aligned}&\min\ J_{WG}(U,V)\\ &s.t.\ \sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,\cdots,n\end{aligned}\,. \quad (3)$$

For optimization problem (3), we use Lagrange approach to solve degree of membership for each sample and centers of clusters. So, Lagrange function L corresponding to (3) is written as

$$\begin{aligned}&L(U,V;\lambda_1,\cdots,\lambda_n)\\ &=\sum_{j=1}^{n}\sum_{i=1}^{c}w_j\mu_{ij}{}^{m}\,\|x_j-v_i\|^2+\delta\sum_{j=1}^{n}(2^{1-m}-1)^{-1}(\sum_{i=1}^{c}\mu_{ij}{}^{m}-1)\\ &+\lambda_1(\sum_{i=1}^{c}\mu_{i1}-1)+\lambda_2(\sum_{i=1}^{c}\mu_{i2}-1)+\cdots+\lambda_n(\sum_{i=1}^{c}\mu_{in}-1)\end{aligned}\,.$$

Here, we take the derivative of function L with respect to λ_j,μ_{ij} and v_i and let them equal to zero, namely

$$\frac{\partial L}{\partial\lambda_j}=\sum_{i=1}^{c}\mu_{ij}-1=0\ , \quad (4)$$

$$\frac{\partial L}{\partial\mu_{ij}}=w_j m\mu_{ij}^{m-1}\|x_j-v_i\|^2+m\delta(2^{1-m}-1)^{-1}\mu_{ij}^{m-1}+\lambda_j=0\ , \quad (5)$$

$$\frac{\partial L}{\partial v_i}=-2\sum_{j=1}^{n}\mu_{ij}^{m}(x_j-v_i)=0\,\cdot \quad (6)$$

From (4), (5) and (6), and using simple algebra operation, we obtain the following degree of membership for each sample and centers of clusters.

$$\begin{aligned}&\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\dfrac{w_j\|x_j-v_i\|^2+\delta(2^{1-m}-1)^{-1}}{w_j\|x_j-v_k\|^2+\delta(2^{1-m}-1)^{-1}}\right)^{\frac{1}{m-1}}}\,,\\ &i=1,2,\cdots,c;\ j=1,2,\cdots,n\end{aligned} \quad (7)$$

$$\begin{aligned}&v_i=\frac{\sum_{j=1}^{n}w_j\mu_{ij}^{m}x_j}{\sum_{j=1}^{n}w_j\mu_{ij}^{m}}\,.\\ &i=1,2,\cdots,c\end{aligned} \quad (8)$$

In the following, we give the sample weighting's fuzzy clustering algorithm with generalized entropy and refer it to as WGEFCM.

Step 1 Initialize c centers of clusters and assign m and δ.

Step 2 Calculate weights w_j for each sample in data set X according to the selected method.

Step 3 Compute degree of membership μ_{ij} for each sample according to (7).

Step 4 Compute center of cluster v_i for each cluster according to (8).

Step 5 Repeat step 3 to step 4 until the center of cluster v_i does not change.

III. KERNEL FUZZY CLUSTERING WITH GENERALIZED ENTROPY BASED ON SAMPLE WEIGHTING

Assume that Φ is a mapping from original space X to feature space and $k(x_j,v_i)$ is a kernel function. Here, let $\|x_j-v_i\|^2$ be replaced by $\|\Phi(x_j)-\Phi(v_i)\|^2$ in objective function (2). So we obtain objective function $J_{WGK}(U,V)$ based on kernel, namely

$$\begin{aligned}J_{WGK}(U,V)=&\sum_{j=1}^{n}\sum_{i=1}^{c}w_j\mu_{ij}{}^{m}\,\|\Phi(x_j)-\Phi(v_i)\|^2\\ &+\delta\sum_{j=1}^{n}(2^{1-m}-1)^{-1}(\sum_{i=1}^{c}\mu_{ij}{}^{m}-1)\end{aligned}\,. \quad (9)$$

Assume that $K(x_j,v_i)=e^{-\frac{\|x_j-v_i\|^2}{\sigma^2}}$ is Gauss kernel. As $\|\Phi(x_j)-\Phi(v_i)\|^2=K(x_j,x_j)+K(v_i,v_i)-2K(x_j,v_i)$ and $\|\Phi(x_j)-\Phi(v_i)\|^2=2-2K(x_j,v_i)$, objective function $J_{WGK}(U,V)$ is become as

$$\begin{aligned}J_{WGK}(U,V)=&2\sum_{j=1}^{n}\sum_{i=1}^{c}w_j\mu_{ij}{}^{m}\left(1-K\left(x_j,v_i\right)\right)\\ &+\delta\sum_{j=1}^{n}(2^{1-m}-1)^{-1}(\sum_{i=1}^{c}\mu_{ij}{}^{m}-1)\end{aligned}\,. \quad (10)$$

Thus, optimization problem for kernel fuzzy clustering with generalized entropy based on sample weighting is as follows:

$$\begin{aligned}&\min\ J_{WGK}(U,V)\\ &s.t.\ \sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,\cdots,n\end{aligned}\,. \quad (11)$$

According to same method above, we obtain degree of membership μ_{ij} and center of cluster v_i, namely

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{w_j(1-K(x_j,v_i))+\delta(2^{1-m}-1)^{-1}}{w_j(1-K(x_j,v_k))+\delta(2^{1-m}-1)^{-1}}\right)^{\frac{1}{m-1}}}, \tag{12}$$

$$j = 1,2,\cdots,n \qquad i = 1,2,\cdots,c$$

$$v_i = \frac{\sum_{j=1}^{n} w_j \mu_{ij}^{\ m} K\left(x_j, v_i\right) x_j}{\sum_{j=1}^{n} w_j \mu_{ij}^{\ m} K\left(x_j, v_i\right)}. \tag{13}$$

$$i = 1,2,\cdots,c$$

Kernel fuzzy clustering algorithm with generalized entropy based on sample weighting is given in the following and we call it as KWGEFCM.

Step 1 Initialize c centers of clusters and assign m, δ and σ.

Step 2 Calculate weight w_j for each sample in data set X according to the selected method.

Step 3 Compute degree of membership μ_{ij} for each sample according to (12).

Step 4 Compute center of cluster v_i for each cluster according to (13).

Step 5 Repeat step 3 to step 4 until the center of cluster v_i does not change.

IV. EXPERIMENT

A. *Determination of Weight for Each Sample*

In the following, we choose two methods to determine weight for each sample. One is based on similarity between samples [12]. That is to say that the greater distance for a sample to other samples, the smaller weight for this sample. Assume that weight for jth sample is w_j which is computed by (18):

$$w_j = \sum_{i=1, i\neq j}^{n}\left(1/d_{ji}\right) \Big/ \sum_{j=1}^{n}\sum_{i=1, i\neq j}^{n}\left(1/d_{ji}\right), \tag{18}$$

where d_{ji} is distance between sample x_j and sample x_i .

The other is on the basis of attributes of sample. Assume that the data set X has n data samples which are represented by x_i and each sample has s attributes, wherein the lth value of attribute for ith data sample is represented as x_{il}. Firstly, samples are normalized by following method

$$r_{jl} = \frac{x_{jl} - \min_l\left(x_{jl}\right)}{\max_l\left(x_{jl}\right) - \min_l\left(x_{jl}\right)},$$

where $\max_l(x_{jl})$ and $\min_l(x_{jl})$ represent maximum value and minimum value for lth attribute, respectively. Secondly, we compute entropy H_l for each attribute using (19) and weight a_l for each attribute.

$$H_l = -k\sum_{j=1}^{n} f_{jl}\ln f_{jl},\ a_l = \frac{1-H_l}{s-\sum_{l=1}^{s} H_l}, \tag{19}$$

where $k = \frac{1}{\ln n}$, $f_{jl} = r_{jl}\Big/\sum_{j=1}^{n} r_{jl}$, $0 \le a_l \le 1$, $\sum_{l=1}^{s} a_l = 1$.

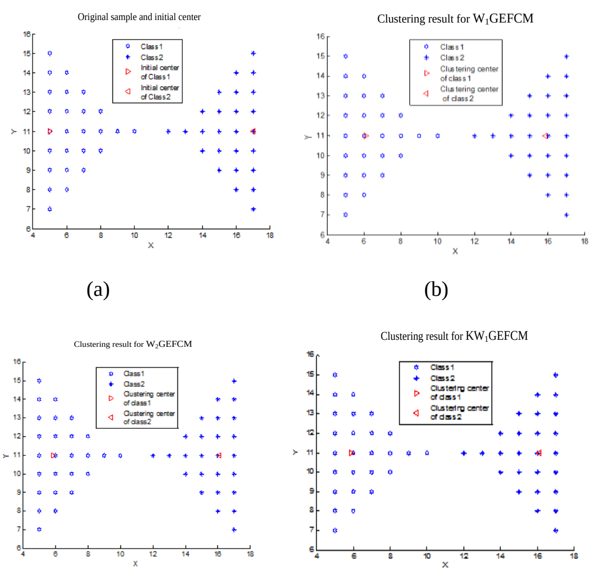
Finally, weight w_j for each sample is computed by using weight a_l for each attribute, namely

$$w_j = \frac{1}{m}\sum_{l=1}^{s} x_{jl} a_l, \quad j = 1,2,\cdots,n\ . \tag{20}$$

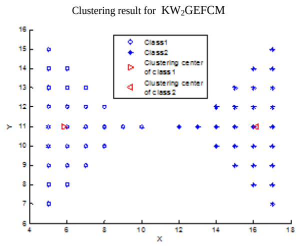
B. *Experimental Results and Analysis*

In order to verify the effectiveness of the proposed algorithm WGEFCM and KWGEFCM, we select five datasets from UCI data repository and artificial generated data set Butterfly. Here, to distinguish two different methods for determining weight of sample, we use subscript in symbol 'W' to denote different method. That is to say, W_1 and W_2 represent the computed weight by using (18) and (20), respectively. In addition, we choose two indexes to evaluate performance of clustering result. They are accuracy *ACC* and mutual information MI, respectively, where *ACC* is represented as $ACC = \frac{n-err}{n}\times 100\%$, n is number of sample in dataset X and *err* is number of misclassified sample. In addition, we choose two methods to initialize centers of clusters in experiment. One is randomly selected method from given dataset. The other is to use fixed method[14].

Firstly, we conduct the experimental study on data set Butterfly. Some experimental results are shown in Fig.1, where Fig. 1(a) is sample set for Butterfly and randomly chosen initial centers, Fig. 1(b)-(e) are clustering results for W_1GEFCM, W_2GEFCM, KW_1GEFCM and KW_2GEFCM, respectively. It is seen that for dataset Butterfly, these algorithms obtain better clustering results.



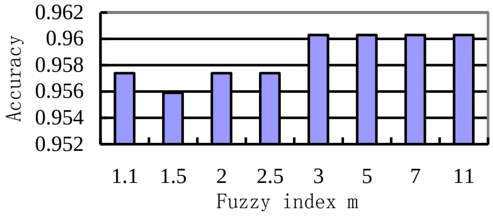
(c) (d)



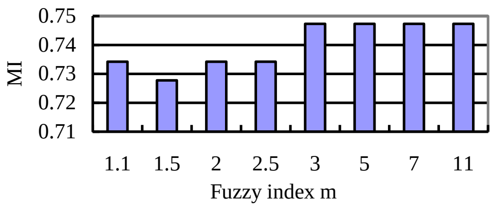
(e)

Figure 1. Clustering results using different algorithm for dataset Butterfly

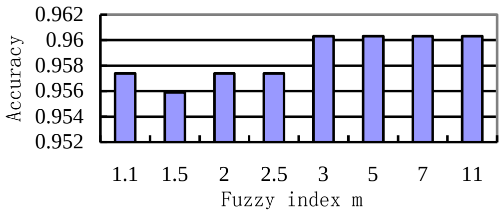
Secondly, we choose dataset Breast-w to conduct the detailed experimental study. In the experiment, parameter δ is fixed as certain value; fuzzy index m is taken as 1.1, 1.5, 2, 2.5, 3, 5, 7 and 11, respectively. Aimed at algorithms W_1GEFCM and W_2GEFCM, experimental results are seen in Fig. 2 and Fig. 3, where Fig. 2(a)-(b) and Fig. 3(e)-(f) are clustering results for algorithm W_1GEFCM about two evaluation indexes whereas Fig. 2(c)-(d) and Fig. 3(g)-(h) are clustering results for algorithm W_2GEFCM about two evaluation indexes. Moreover, it is noted that for two different initial center methods, we get same clustering results as seen in Fig. 2(a) and Fig. 2(c), Fig. 3(e) and Fig. 3(g).



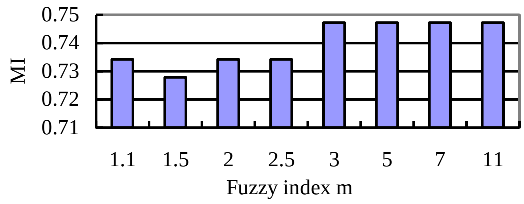
(a)



(b)

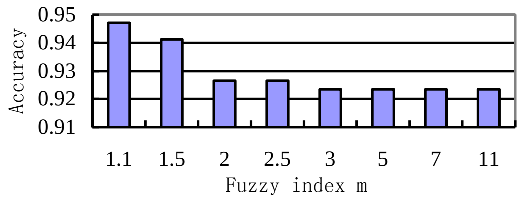


(c)

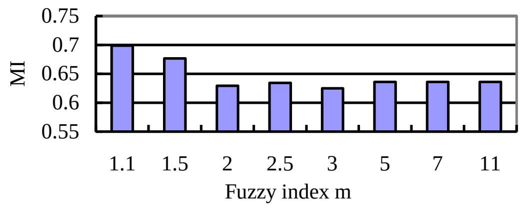


(d)

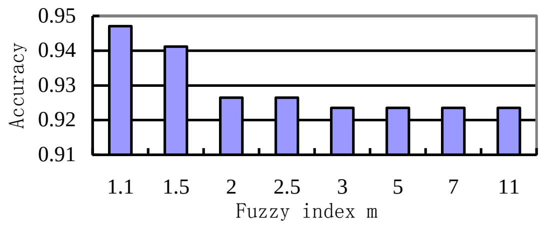
Figure 2. Relation between fuzzy index m, accuracy and mutual information(MI) about W_1GEFCM, where (a) and (b) are clustering results based on fixed chosen center's method whereas (c) and (d) are clustering results based on randomly chosen sample's method.



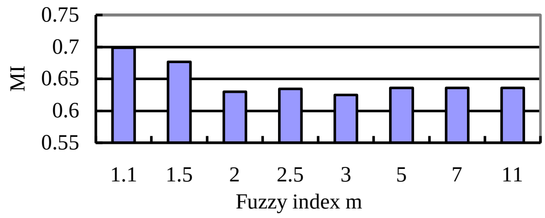
(e)



(f)



(g)

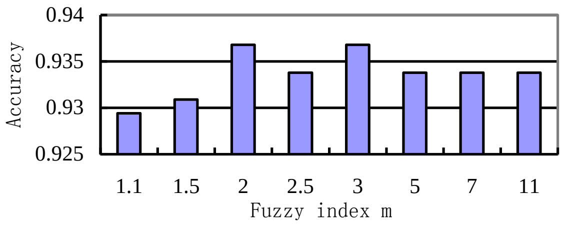


(h)

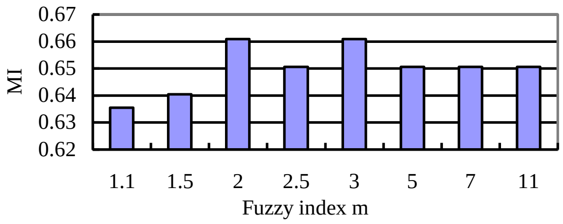
Figure 3. Relation between fuzzy index m, accuracy and mutual information (MI) about W_2GEFCM, where (e) and (f) are clustering results based on fixed chosen center's method whereas (g) and (h) are clustering results based on randomly chosen sample's method.

According to above method, we also study the clustering performance of KW_1GEFCM and KW_2GEFCM. Experimental results are seen in Fig. 4 and Fig. 5, where Fig. 4(a)-(b) and

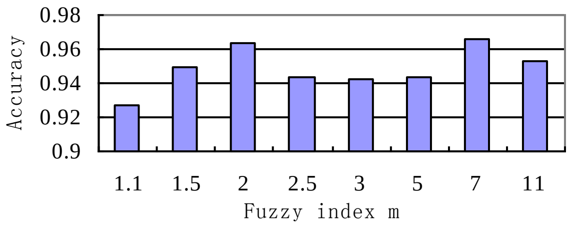
Fig. 5(e)-(f) are clustering results based on fixed chosen center's method whereas Fig. 4(c)-(d) and Fig. 5 (g)-(h) are clustering results based on randomly chosen sample's method. It is seen that for kernel clustering algorithms KWGEFCM, their clustering performance is different for different initial center's method. Especially, when using randomly method to determine initial centers, we obtain some better clustering results.



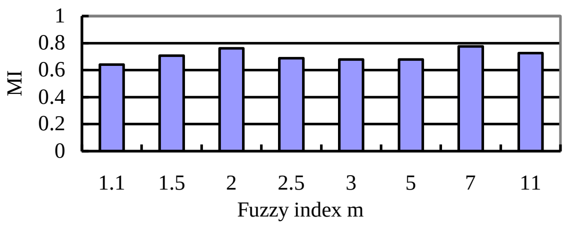
(a)



(b)

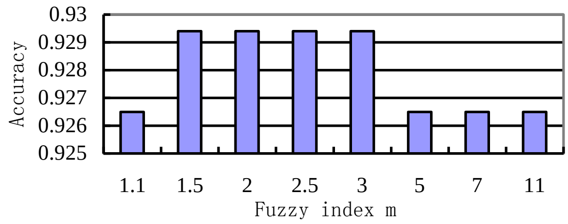


(c)

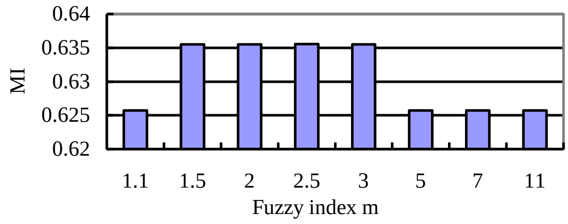


(d)

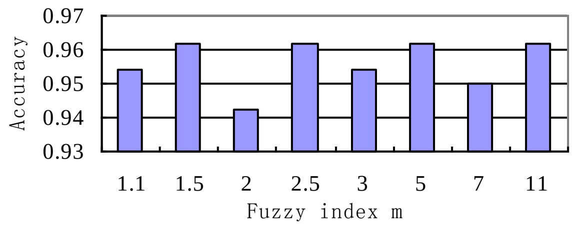
Figure 4 Relation between fuzzy index m, accuracy and mutual information(MI) about KW_1GEFCM, where (a) and (b) are clustering results based on fixed chosen center's method whereas (c) and (d) are clustering results based on randomly chosen sample's method.



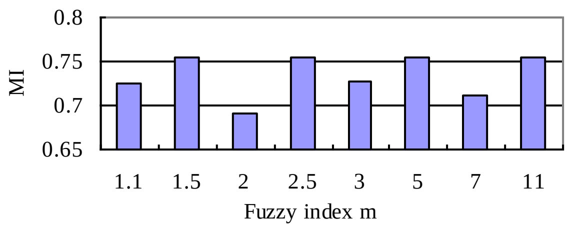
(e)



(f)



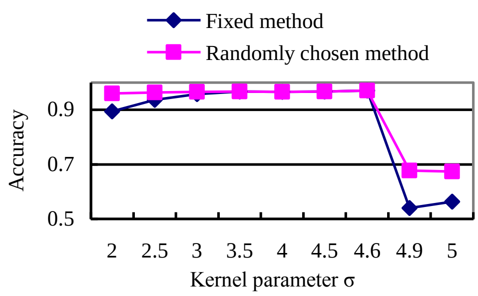
(g)



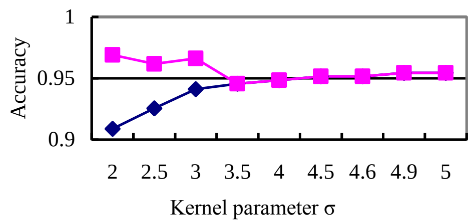
(h)

Figure 5. Relation between fuzzy index m, accuracy and mutual information(MI) about KW_2GEFCM, where (e) and (f) are clustering results based on fixed chosen center's method whereas (g) and (h) are clustering results based on randomly chosen sample's method.

Based on above analysis, in the following experiments, we further study performance of algorithm KWGEFCM. Here, we fix parameters m and δ and let kernel parameter σ take value 2, 2.5, 3, 3.5, 4, 4.5, 4.6, 4.9, and 5, respectively. Experimental results are seen in Fig. 6 and Fig. 7, where Fig. 6(a) and Fig. 6(b) are clustering result for KWGEFCM using two different sample weighting methods; Fig. 7(a) and Fig. 7(b) are mutual information about clustering result for KWGEFCM using two different sample weighting methods.

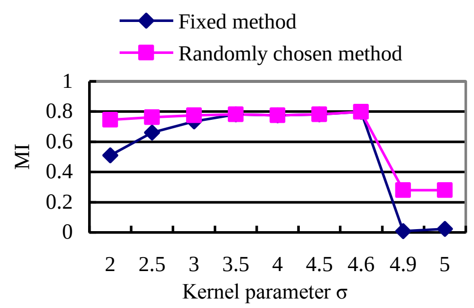


(a)

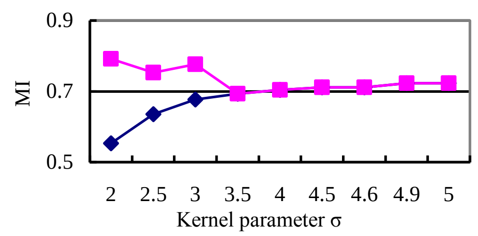


(b)

Figure 6. Relation between kernel parameter σ and accuracy for KW_1GEFCM(a) and KW_2GEFCM(b)



(a)



(b)

Figure 7. Relation between kernel parameter σ and mutual information (MI) for KW_1GEFCM(a) and KW_2GEFCM(b)

Besides, we study clustering performance for datasets Australian, Heart, Ionosphere and Iris. Experimental results are seen in Fig. 8. At the same time, we also give clustering

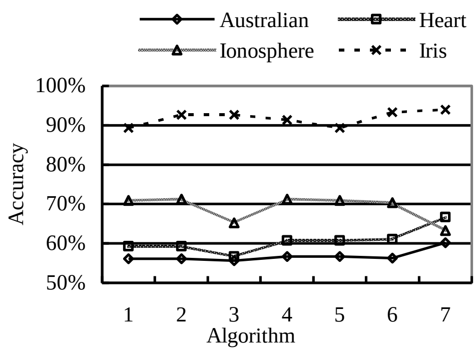


Figure 8. Performance of clustering using different sample weighting method and initialized center method

results for FCM, GEFCM and KGEFCM [13]. It is seen that we obtain better clustering results compared with FCM, GEFCM and KGEFCM in majority datasets. Note that in Fig. 8, digitals 1-7 are represented as different algorithm, namely 1 as FCM, 2 as GEFCM, 3 as KGEFCM, 4 as W_1GEFCM, 5 as W_2GEFCM, 6 as KW_1GECM and 7 as KW_2GEFCM.

V. CONCLUSIONS

Aiming at fuzzy clustering with generalized entropy, we study fuzzy clustering and its kernel fuzzy clustering with weights of samples. Objective function for fuzzy clustering method with generalized entropy based on weighted sample is obtained. Then, fuzzy clustering algorithm with generalized entropy based on weighted sample and its kernel fuzzy clustering algorithm are presented. We select some representative datasets and combine two methods for determining weight of sample to conduct experimental study. Experimental results show that the presented algorithms are effective.

ACKNOWLEDGMENT

This work is support by Natural Science Foundation of China (No. 61375075) and Nature Science Foundation of Hebei Province (No. F2012201014).

REFERENCES

[1] N. B. Karayiannis, "MECA : maximum entropy clustering algorithm," IEEE world congress on computational intelligence, In: Proceedings of the third IEEE conference on fuzzy systems,vol. 1, pp. 630-635,1994.

[2] R.P. Li, M. A Mukaidon,"Maximum entropy approach to fuzzy clustering," In: Proceedings of fouth IEEE international conference on fuzzy system,Yokohama,Japan, pp. 2227-2232,1995.

[3] D.Tran and M.Wagner,"Fuzzy entropy clustering," The Ninth IEEE International Conference on Fuzzy Systems, vol. 1, pp.152-157,2000.

[4] C.Wei C.Fahn, "The multisynapse neural network and its application to fuzzy clustering," IEEE transactions on neural networks,vol. 13, issue 3, pp. 600-618,2002.

[5] D. Graves,W. Pedrycz, "Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study," Fuzzy Sets and Systems, vol. 161, pp. 522–543, 2010.

[6] D. Q. Zhang, S. C. Chen, "Kernel-based fuzzy and possibilistic c-means clustering," In: Proceedings of the International Conference on Artificial Neural Networks (ICANN'03), Istanbul, Turkey, pp.122~125,2003.

[7] S. C. Chen, D. Q Zhang, "Robust image segmentation using FCM with spatial constrains based on new kernel-induced distance measure," IEEE Trans. Systems Man Cybernet. Pt. B,vol. 34, pp. 1907–1916 ,2004.

[8] M. S. Yang, H. S. Tsai, "A gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction," Pattern recognition letters, vol. 29,pp. 1713-1725,2008.

[9] S. R. Kannan, S. Ramathilagam, A. Sathya, et al. "Effective fuzzy c-means based kernel function in segmenting medical images," Computers in Biology and Medicine, vol. 40, pp. 572–579,2010.

[10] D. Swagatam, S. Sudeshna, "Kernel-induced fuzzy clustering of image pixels with an improved differential evolution algorithm," Information Sciences,vol. 180,pp. 1237-1256,2010.

[11] H. Q. Rong, Z.C. Li, "Automated variable weighting in k-means type clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, issue 5, pp. 657-668,2005.

[12] X. Su, X. Y.Wang, Z. Wang, Y. Y.Xiao, "A new fuzzy clustering algorithm based on entropy weighting," Journal of Computational Information Systems, vol.6, issue 10, pp.3319-3326,2010.

[13] K. Li, H. Y. Ma and Y. Wang, "Unified model of fuzzy clustering algorithm based on entropy and its application to image segmentation," Journal of computational information systems, vol. 7,issue 15, pp.5476-5483,2011.

[14] S. R.Kannan, S.Ramathilagam, A.Sathya, et al., "Effective fuzzyc-means based kernel function in segmenting medical images," Computers in Biology and Medicine, vol. 40, pp.572–579,2010.