International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:5, No:3, 2011

Similarity Measures and Weighted Fuzzy C-Mean Clustering Algorithm

Bainian Li, Kongsheng Zhang, and Jian Xu

Abstract—In this paper we study the fuzzy c-mean clustering algorithm combined with principal components method. Demonstratively analysis indicate that the new clustering method is well rather than some clustering algorithms. We also consider the validity of clustering method.

Keywords—FCM algorithm, Principal Components Analysis, Cluster validity.

I. FUZZY SET THEORETIC SIMILARITY MEASURES

ONE of the most important issues in recommender systems research is computing similarity between users, and between items (products, events, services, etc.). This in turns highly depends on the appropriateness and reliability of the methods of representation. The set-theoretic, proximity-based and logic-based are the three classes of measures of similarity. In fuzzy set and possibility framework, similarity of users or items is computed based on the membership functions of the fuzzy sets associated with the users or item features. Based on the work of Cross and Sudkamp [1], those similarity measures [2] that are relevant to item recommendation application are adapted.

For items I_j and I_k that are defined as $\{x_i, \mu_{x_i}(I_j)), i = 1, 2, \ldots, N\}$ and $\{x_i, \mu_{x_i}(I_k)), i = 1, 2, \ldots, N\}$, a similarity measure between I_j and I_k is denoted by $S(I_k, I_j)$, and the different similarity measures are defined as

$$S_1(I_k, I_j) = \frac{\sum_i \min(\mu_{x_i}(I_k), \mu_{x_i}(I_j))}{\sum_i \max(\mu_{x_i}(I_k), \mu_{x_i}(I_j))}, \quad (1)$$

$$S_2(I_k, I_j) = \frac{\sum_i \mu_{x_i}(I_k)\mu_{x_i}(I_j)}{\sqrt{\sum_i (\mu_{x_i}(I_k))^2}\sqrt{\sum_i (\mu_{x_i}(I_j)))^2}}, \quad (2)$$

$$S_3(I_k, I_j) = 1 - \frac{d_2(I_k, I_j)}{\max_i \{\mu_{x_i}(I_k), \mu_{x_i}(I_j)\}}, \quad (3)$$

$$S_4(I_k, I_j) = 1 - \frac{2}{Z_{I_k} + Z_{I_j}} d_2(I_k, I_j)^2, \quad (4)$$

where

$$d_2(I_k, I_j) = \sqrt{\sum_i (\mu_{x_i}(I_k) - \mu_{x_i}(I_j))^2},$$

$$Z_{I_a} = \sum_i (2\mu_{x_i}(I_a) - 1)^2 for\ a = k\ or\ j.$$

Corresponding author: Bainian Li is with the School of Statistic and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China.(e -mail: libainian49@163.com).

In this article we define a new similarity measures as below:

$$S(I_k, I_j) = \frac{\sum_i 2\mu_{x_i}(I_k)\mu_{x_i}(I_j)}{\sum_i (\mu_{x_i}(I_k))^2 + \sum_i (\mu_{x_i}(I_j))^2}. \quad (5)$$

Formula (5) has the following character
(a) Reflexive, i.e., for all I_k

$$S(I_k, I_k) = 1. \quad (6)$$

(b) Symmetric i.e., for all $I_k,\ I_j$,

$$S(I_k, I_j) = S(I_j, I_k). \quad (7)$$

(c) Transitive, i.e., if $\mu_{x_i}(I_k) < \mu_{x_i}(I_j) < \mu_{x_i}(I_m)$, then

$$S(I_k, I_j) \geq S(I_k, I_m). \quad (8)$$

By using formula (5) we can obtain a real symmetry matrix.

Let X be a real matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$$

Then

$R(X) = (r_{kj})_{p\times p}$ is a real symmetry matrix,

where

$$r_{kj} = \frac{2\sum_{i=1}^{n} x_{ik}x_{ij}}{\sum_{i=1}^{n} x_{ik}^2 + \sum_{i=1}^{n} x_{ij}^2}, (k, j = 1, 2, ..., p).$$

Furthermore, we have

Theorem 1: Let $X = (x_{ij})_{n\times p}$ be a real matrix, matrix Y is standardization of matrix X,
then

$$R(Y) = \text{corrcoef}(X),$$

where

$Y = (y_{ij})_{n\times p}, y_{ij} = (x_{ij} - \bar{x}_j)/s_j,\ \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$, corrcoef$(X)$ is correlation matrix

Proof:

$$\frac{2\sum_{i=1}^{n} y_{ik}y_{ij}}{\sum_{i=1}^{n} y_{ik}^2+\sum_{i=1}^{n} y_{ij}^2} = \frac{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)(x_{ij}-\bar{x}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$$

$$= 2\sum_{i=1}^{n}\frac{x_{ik}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}}\frac{x_{ij}-\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$$

$$= \sum_{i=1}^{n}\frac{(x_{ik}-\bar{x}_k)^2}{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}+\sum_{i=1}^{n}\frac{(x_{ij}-\bar{x}_j)^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$$

$$= 2\sum_{i=1}^{n}\frac{x_{ik}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}}\frac{x_{ij}-\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}+\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$$

$$= \sum_{i=1}^{n}\frac{x_{ik}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}}\frac{x_{ij}-\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)(x_{ij}-\bar{x}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}.$$

II. ORIGINAL FUZZY C-MEANS ALGORITHM

For a given data set $X=\{X_1,X_2,...,X_N\}\subset R^p$, FCM is an iterated process involving cluster center $C=(v_1,v_2,...,v_c)$ and membership matrix $U=(u_{ij}), i=1,2,...,c, j=1,2,...,n$, where u_{ij} denotes the grade of j-th object which belongs to center v_i.The process is listed as follows [3]:

Step 1: Given a positive integer c which can be decided by some rules. Initialize the membership matrix U by random uniform numbers in interval [0, 1].

Step 2: For $i=1,2,\cdots,c$, $j=1,2,...,n$, and $m>1$, we calculate the cluster center C,

$$v_i^{(l)}=\sum_{k=1}^{\mathrm{N}}(u_{ik}^{(l)})^m x_k/\sum_{k=1}^{\mathrm{N}}(u_{ik}^{(l)})^m \tag{9}$$

and new membership matrix

$$u_{ik}^{(\mathrm{l}+1)}=1/\sum_{j=1}^{\mathrm{c}}(d_{ik}/d_{jk})^{\frac{2}{m-1}} \tag{10}$$

and update the initialized fuzzy membership matrix which has the following character.

Step 3: Compute the objective function

$$J(U,V)=\sum_{k=1}^{\mathrm{N}}\sum_{i=1}^{\mathrm{c}}(u_{ik})^m(d_{ik})^2. \tag{11}$$

Step 4: Given $\varepsilon>0$, if $\max\{|u_{ik}^t-u_{ik}^{t-1}|\}<\varepsilon$, then the procedure ends, else go to step 2.

III. WEIGHTED FCM

Weighted FCM is the following programming:

$$\min J(U,V,c)=\sum_{k=1}^{\mathrm{N}}\sum_{i=1}^{\mathrm{c}}(u_{ik})^m(\sqrt{w_j}d_{ik})^2,$$

$$s.t\begin{cases}0\le u_{ik}\le 1, & 1\le i\le c, 1\le k\le N\\ \sum_{i=1}^{c}u_{ik}=1, & 1\le k\le N\\ 0<\sum_{k=1}^{N}u_{ik}<N, & 1\le i\le c\\ 1\le m\le\infty.\end{cases}$$

where w_j can be computed by three steps.

Berget et al [4] obtained a new modification and application of fuzzy c-mean methodology. Our method is as below:

Step 1: To perform a fuzzy similarity matrix $Y=(y_{ij})$, where

$$y_{ij}=2\sum_{i=1}^{n}x_{ik}\cdot x_{ij}/(\sum_{i=1}^{n}x_{ik}^2+\sum_{k=1}^{n}x_{ij}^2).$$

Step 2: Compute eigenvalue λ_i of matrix Y.

Step 3: Compute weight vector $w=(w_1,w_2,...,w_p)$, where

$$w_i=\lambda_i/\sum_{i=1}^{p}\lambda_i.$$

We next use the weighted FCM to deal with the partition problem of Iris database consisting of 150 samples and three classes in Fisher. Each sample has four features: sepal length, sepal width, and petal length and petal width. The error rates of four different methods for this data set are listed in Table 1. For the Iris databasethe Weighted FCM works very well. From Table 1, one can see the performance of weighted FCM is better than those of Wang et al [5], Zhang et al [6] and Hung et al [7] which resorted to bootstrap method.

By using MATLAB software, we obtain that when $2<m\le 2.5$, error rates=5/150. Meanwhile we have Table 1 and Fig.1.

IV. CLUSTER VALIDITY

The main purpose of studying cluster validity is to determine the optimal cluster number and partitions [8,9,10]. As mentioned earlier, the degree of variation can be quantized by computing the intra cluster errors. Some cluster validity indices available as below:

TABLE I
THE ERROR RATES OF FOUR METHODS FOR IRIS DATA

m			*1.5*		
Methods	Original FCM	Wang et al.	Hung et al.	Zhang et al.	Our method
Error rates	16/150	9/150	9/150	8/150	6/150
m			*2*		
Methods	Original FCM	Wang et al.	Hung et al.	Zhang et al.	Our method
Error rates	16/150	8/150	9/150	9/150	6/150
m			*5*		
Methods	Original FCM	Wang et al.	Hung et al.	Zhang et al.	Our method
Error rates	15/150	10/150	8/150	8/150	6/150
m			*10*		
Methods	Original FCM	Wang et al.	Hung et al.	Zhang et al.	Our method
Error rates	12/150	10/150	7/150	10/150	6/150

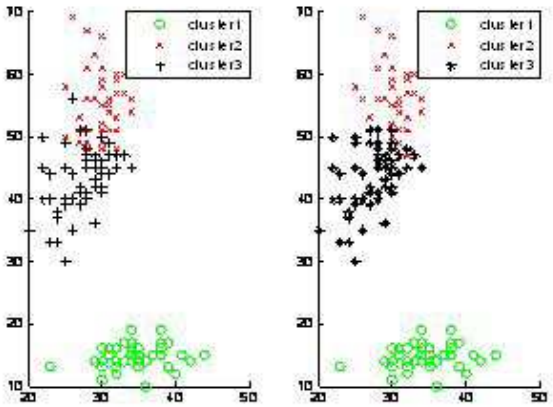

Fig. 1 Original FCM (left) and our method (right) for Iris data

(a) Bezdek

$$\mathrm{V_{PE}} = -\frac{1}{\mathrm{n}}\sum_{\mathrm{j=1}}^{\mathrm{n}}\sum_{\mathrm{i}}^{\mathrm{c}} u_{ij}\log_{\mathrm{a}}(u_{ij}). \quad (12)$$

In general, we find an optimal c^* by solving $\min_{2\le c\le n-1} V_{PE}$ to produce the best clustering performance for the data set X.

(b) Xie-Beni

$$\mathrm{V_{XB}} = \sum_{\mathrm{i=1}}^{\mathrm{c}}\sum_{\mathrm{j=1}}^{\mathrm{n}} u_{ij}^2||x_j - v_i||^2/\mathrm{n}\min_{i\neq k}||v_j - v_i||^2. \quad (13)$$

In general, we find an optimal c^* by solving $\min_{2\le c\le n-1} V_{XB}$ to produce the best clustering performance for the data set X.

(c) Kuyama & Sugeno

$$\mathrm{V_{FS}} = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m||x_j - v_i||^2 - \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m||v_j - \bar{v}||^2. \quad (14)$$

In general, we find an optimal c^* by solving $\min_{2\le c\le n-1} V_{FS}$ to produce the best clustering performance for the data set X, where $\bar{v} = \sum_{i=1}^{c} v_i/c$.

(d) Kwon

$$\mathrm{V_k} = [\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2||x_j - v_i||^2 + \tfrac{1}{c}||v_j - \bar{v}||^2]/\min_{i\neq k}||v_i - v_k||^2. \quad (15)$$

In general, we find an optimal c^* by solving $\min_{2\le c\le n-1} V_K$ to produce the best clustering performance for the data set X.

In this article we establish a new cluster validity rule as below:

$$V_{LI} = \sqrt{\frac{c+1}{c-1}}\sum_{j=1}^{n}[\max_{1\le i\le c}(-u_{ij}\log(u_{ij})) + \sqrt{(c\sum_{i=1}^{c} u_{ij}^2 - 1)/c(c-1)}],\ c>1. \quad (16)$$

where c is cluster number, $u = (u_{ij})_{c\times n}$ is the final fuzzy partition matrix (or membership function matrix). $-u_{ij}\log(u_{ij})$ is entropic of membership function matrix, $\sqrt{(c\sum_{i=1}^{c} u_{ij}^2 - 1)/c(c-1)}$ is variation coefficient of membership function matrix.

We find an optimal c^* by solving $\min_{2\le c\le n-1} V_{LI}$ to produce the best clustering performance for the data set X.

Cluster validity functions are often used to evaluate the performance of clustering in different indexes and even two different clustering methods. A lot of cluster validity criteria were proposed during the last 10 years. Most of them came from different studies dealing with the number of clusters.

To test validity indices, we conduct the Iris data sets and Wine data sets. Iris data sets are perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of Iris plant. One class is linearly separable from the other 2; the latter is not linearly separable from each other. Wine data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determines the quantity of 13 constituents found in each of the three types of wines.

The optimal number of cluster is shown in Table 2. From Table 2, we find that the cluster result of V_{LI} is the most excellent.

V. Conclusion

The weighted FCM algorithm constructed in this paper has the following characteristics:

(1) Misjudgment rate is low than other clustering method for some classical data;

(2) Better stability of our method for different index m. The superiority to other clustering methods suggests that we should adopt the new clustering algorithm.

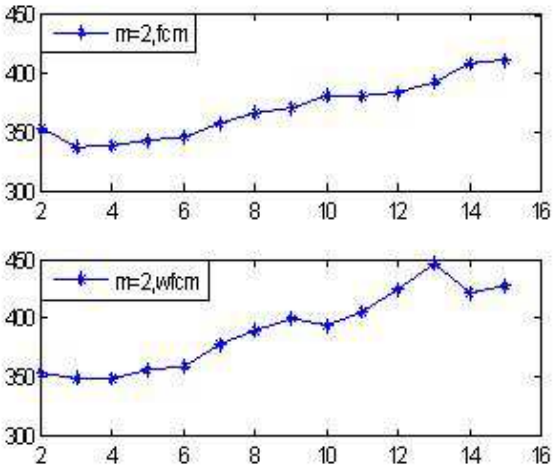

Fig. 2 Iris data cluster number c =3

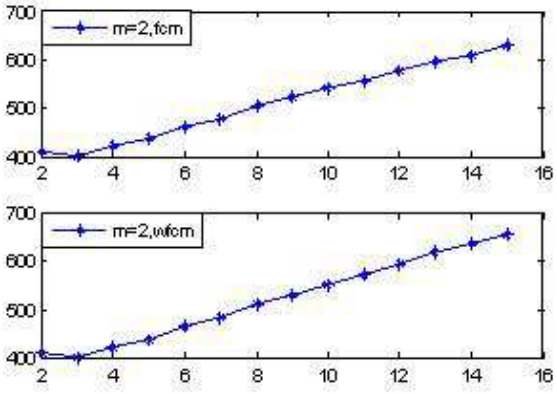

Fig. 3 Wine data cluster number c =3

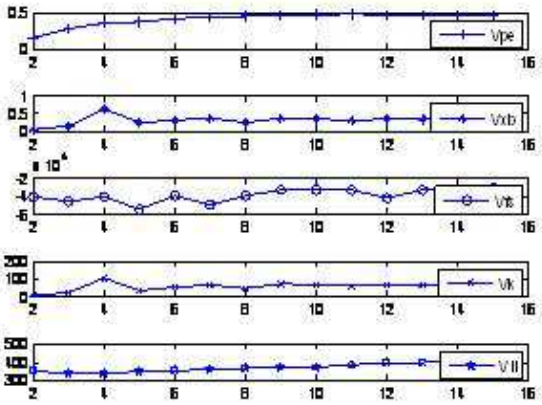

Fig. 4 Cluster validity index of Iris data(m =2)

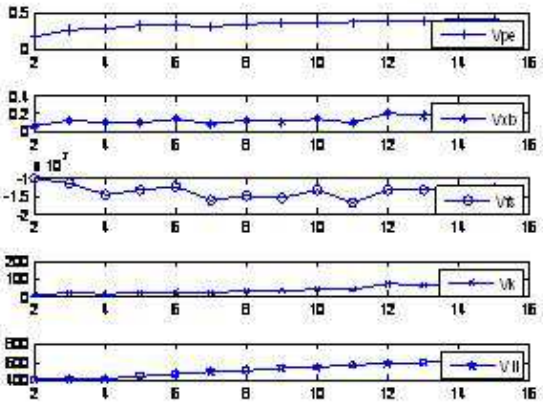

Fig. 5 Cluster validity index of Wine data(m =2)

TABLE II
OPTIONS CLUSTER NUMBERS FOR IRIS DATA AND WINE DATA

Criterion function	Number of Cluster (Iris data)	Number of Cluster (Wine data)	m
V_{PE}	2	2	
V_{XB}	2	2	
V_{FS}	5	7	1.5
V_{K}	2	2	
V_{LI}	3	2	
V_{PE}	2	2	
V_{XB}	2	2	
V_{FS}	5	11	2
V_{K}	2	2	
V_{LI}	3	3	
V_{PE}	2	2	
V_{XB}	2	2	
V_{FS}	5	6	2.1
V_{K}	2	2	
V_{LI}	4	3	
V_{PE}	2	2	
V_{XB}	2	2	
V_{FS}	5	4	2.5
V_{K}	2	2	
V_{LI}	5	3	

ACKNOWLEDGMENT

This work was supported by the science Foundation of Anhui Province (090416222,KJ2010B001), and the social science Foundation of Anhui Province (2010sk226).

REFERENCES

[1] V.V. Cross and T.A. Sudkamp, Similarity and Compatibility in Fuzzy Set Theory: assessment and Applications, Physica-Verlag, New York, 2002.
[2] M. Kalina, Derivatives of fuzzy functions and fuzzy derivatives, Tatra Mountains Mathematical Publications 12 (1997) 27-34.
[3] K.L.Wu and M.S.Yang. Alternative c-means clustering algorithms. Pattern Recognition. 2001,120:249-254.
[4] I.Berget, B.H.Mevi and T.Nas. New modifications and applications of fuzzy c-means methodology. Computational Statistics & Data Analysis. 2008,52:2403-2418.
[5] X.Z.Wang, Y.D.Wang and L.J.Wang. Improving fuzzy c-means clustering based on feature-weighted learning. Pattern Recognition Letters.2004, 25:1123-1132.
[6] K.S.Zhang, B.N.Li. New modification of fuzzy c-means clustering algorithm. In Cao BY, Zhang CY Proceedings of the Third Annual Conference on Fuzzy Information and Engineering .New York: Springer, 2009: 448-455.
[7] W.L.Hung, M.S.Yang and D.H.Chen. Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation. Pattern Recognition Letters,2008,29:1317-1325.
[8] J.J.Higgings. Introduction to Modern Nonparametric Statistics. Duxbury, Belmont, CA,2002.
[9] K.J.Zhu, S.H.Shu and J.L. Li. Optimal number of clusters and the best partion in fuzzy c-means. Systems, Engineering-Theory and Practice. 2005, 3:52-61.(in Chinese)
[10] Y.J.Zhang, W.N.Wang, X.N. Zhang and Y.Li. A cluster validity index for fuzzy clustering. Information Sciences. 2008,178:1205-1218.