Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 8; October, 2014 pp. 67-69
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/jbaer.html

Effective Fuzzy Based Clustering in Complex Database

S.R. Kannan[1], R. Devi[2], S. Ramathilagam[3]

[1, 2]Dept. of Mathematics, Pondicherry University, Pondicherry
[3]Dept. of Mathematics, Periyar Govt. Arts College, Cuddalore, Tamil Nadu

Abstract: **The most widely used fuzzy c-means in complex data clustering including automated image segmentation is incapable in clustering nonlinear clusters because of its Euclidian norm to measure the similarity between the data points. Further it fails to incorporate any information about spatial context, and neighborhood information to cluster the dataset into meaningful subgroups. To overcome the drawbacks, this paper formulates suitable novel fuzzy soft computing techniques with effective cluster center initialization in clustering more general shaped complex data structure of real world problems. The algorithms of this paper are obtained by incorporating the kernel induced distance function, entropy functions, weighted distance measure, and neighborhood terms based spatial context. Experimental results on Yeast dataset indicate that the proposed method is effective and more robust in clustering the complex datasets.**

Keywords: **Fuzzy C-Means, Kernel induced distance, Entropy Terms, Complex database.**

1. INTRODUCTION

Data analysis is the process of getting useful patterns and information from raw data. Clustering is a particular step in this process involving the application of specific algorithms for extracting information from data. Clustering divides the data set into several clusters. The potential of clustering methods to expose the underlying structures in data can be exploited in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining [1, 4, 13]. Clustering algorithms can be classified into main two categories: hard clustering algorithms and fuzzy clustering algorithms. The hard clustering algorithms yield exhaustive partitions of the dataset into non-empty and pair wise disjoint subsets. Due to the uncertain nature of many practical real world problems, hard partition is not suitable to cluster the real world data. Fortunately, the fuzzy set theory (Prof. Zadeh (1965)) [14] was given the concept of partial membership degrees to data elements by membership functions. The idea of using fuzzy set theory for clustering is an extensive tool to cluster the data elements with fuzzy memberships and it is known as fuzzy clustering. The most known method of fuzzy clustering is the Fuzzy c-Means method (FCM), initially proposed by Dunn [5] and generalized by Bezdek [1]. Fuzzy c-Means allows gradual memberships of data elements to clusters in [0, 1]. The membership value of the data element indicates the strength of its association in that cluster. Although the conventional Fuzzy c-Means algorithm works well on real world data, still it has severe limitations [3]. Hence researchers have invented modified fuzzy c-means algorithms [2, 6, 7, 8, 10, 12, 15] in order to deal the data objects with different noises for facing the real world problems. But the above algorithms are failed to improve the accuracy in clustering the dataset which have corrupted by heavy noise, like measurement error, and data transmitter error [11].

This paper introduces the new algorithms to alleviate these drawbacks by using nonlinear transformations from the original pattern space into a higher dimensional feature space with properties of kernel functions [7]. The proposed methods work well with the dataset which has affected by the noises such as measurement error, faulty equipment, and data transmission in dataset.

The rest of this paper is organized as follows. In section 2 this paper presents proposed algorithm. Section 3 reports the experimental results on Yeast Dataset. Section 4 gives the conclusion of this paper.

2. PROPOSED ALGORITHM

The modified objective function from the standard fuzzy c-means [1] is given by

$$J(U,V)=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}\left\|\psi(x_i)-\psi(v_k)\right\|^2 \qquad (1)$$

where ψ stands as map $x->\psi(x)\in F$, $x\in X$. The common ground of Kernel based FCM is to map the input data element into a feature space with higher dimension via a nonlinear transformation and then perform FCM in that feature space. And the distance function can be expressed using in product space as $\|\psi(x_i)-\psi(v_k)\|^2=\langle\psi(x_i),\psi(x_i)\rangle+\langle\psi(v_k),\psi(v_k)\rangle-2\langle\psi(x_i),\psi(v_k)\rangle$, where i = 1, 2, . . ., n, and k = 1, 2, . . ., c. We adopt Hyper Tangent Function to evaluate distance, that is, the $\psi(x_i,v_k)$ express as Hyper Tangent function $\psi(x_i,v_k)=1-\tanh\left(\frac{-\|x_i-v_k\|^2}{w_k}\right)$

where w_k is the weighted mean distance in cluster *k*, and is given by $w_k=\left\{\frac{\sum_{i=1}^{n}u_{ik}\|x_i-v_k\|^2}{\sum_{i=1}^{n}u_{ik}}\right\}^{\frac{1}{2}}$ (2)

Using the expression (2) we obtained $\psi(x_i,x_i)$ = 1 and $\psi(v_k,v_k)$ = 1, so the distance function can be rewritten as

$$\|\psi(x_i)-\psi(v_k)\|^2=2(1-\psi(x_i,v_k)) \tag{3}$$

From equation (1) & (3), we have the kernelized fuzzy c-means given by

$$J(U,V)=2\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^2.(1-\psi(x_i,v_k)) \tag{4}$$

In order to cluster effectively the more complicated dataset which have corrupted by the noises such as measurement error, faulty equipment, and data transmission, the Renyi's entropy fuzzy c-means based hyper tangent kernel algorithm [KEFCM$_{wd}$] is introduced as

$$J(U,V)=2\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^2.(1-\psi(x_i,v_k))+\frac{1}{|1-z|}\sum_{i=1}^{n}\sum_{k=1}^{c}\ln u_{ik}^{z} \tag{5}$$

Here *z* is the resolution parameter. The KEFCM$_{wd}$ objective function is optimized to obtain effective membership grades to the objects which are closed to their prototypes. Using the Lagrange multiplier to the objective function of KEFCM$_{wd}$, the equation for obtaining prototypes and membership grades are calculated.

Optimizing Equation (5)

$$u_{ik}=\frac{4(1-\psi(x_i,v_k))+\frac{z}{|1-z|}}{\sum_{j=1}^{c}4(1-\psi(x_i,v_j))+\frac{z}{|1-z|}} \tag{6}$$

The general equation is used to attain membership grades for data elements for getting meaningful groups. The accuracy of clustering results mainly depends on the cluster centers. Now optimizing the objective function of KEFCM$_{wd}$, this paper obtains the equations for updating the prototypes.

$$v_k=\frac{\sum_{i=1}^{n}\frac{u_{ik}^2}{w_k}\psi(x_i,v_k)\left(1+\tanh\left(\frac{-\|x_i-v_k\|^2}{w_k}\right)\right)x_i}{\sum_{i=1}^{n}\frac{u_{ik}^2}{w_k}\psi(x_i,v_k)\left(1+\tanh\left(\frac{-\|x_i-v_k\|^2}{w_k}\right)\right)} \tag{8}$$

3. EXPERIMENTAL RESULTS ON YEAST

Yeast Dataset [9] is composed of 1484 data, each data has 9 attributes (8 predictive, 1 name). The eight predictive are: 1) McGeoch's method for signal sequence recognition; 2) von Heijne's method for signal sequence recognition; 3) score of the ALOM membrane spanning region prediction program; 4) score of discriminant analysis of the amino acid content of the N-terminal region of mitochondrial and non-mitochondrial proteins; 5) Presence of "HDEL" substring; 6) peroxisomal targeting signal in the C-terminus; 7) score of discriminate analysis of the amino acid content of vacuolar and extracellular proteins; and 8) score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins. The data has 10 classes.

Kernel Fuzzy C-Means (KFCM) clustering results based on 10 classes in yeast data are plotted in Fig. 1. The divided 10 classes by KFCM are visualized in Fig. 2. The results of proposed KEFCM$_{wd}$ method based on 10 classes in yeast dataset are shown in Figs.3-4. The objects in 10 classes of yeast dataset by Proposed KEFCM$_{wd}$ are given in Fig.3. The captured size of 10 classes of yeast dataset by proposed KEFCM$_{wd}$ are shown in Fig.4. The clustering accuracy of KFCM, proposed KEFCM$_{wd}$ on clustering 10 classes in yeast database are listed in Table-1. This paper shows from Table-1, the proposed method improve the clustering accuracy than the KFCM algorithm because of weighted distance with Renyi's entropy.

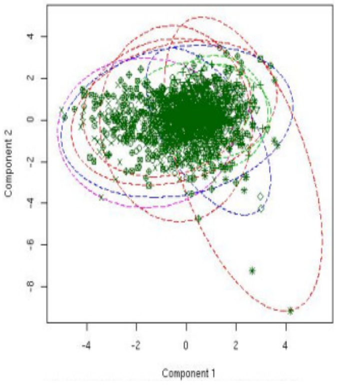

Fig 1. Size of Clusters by KFCM

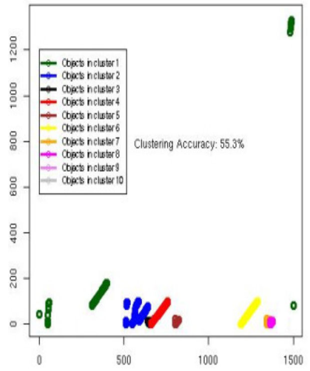

Fig 2. Reallocated 1484 Data by KFCM

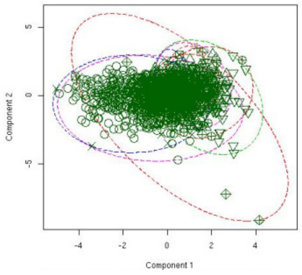

Fig3. Size of Clusters by KEFCM$_{wd}$

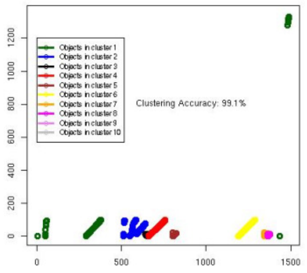

Fig 4. Reallocated 1484 Data by KEFCM$_{wd}$

Table1. Silhouette average values in clustering Yeast Dataset

	KFCM	KEFCM$_{wd}$
Accuracy	55.3 %	99.1%

Table 2: Error Matrix on Yeast Dataset

	KFCM	KEFCM$_{wd}$
Accuracy	51 %	98.7%

The Error Matrix Table-2 gives the accuracy between reference classes and the obtained classes in yeast dataset by the methods involved in this experiment study. From Table-2, the best clustering accuracy was obtained for proposed methods during the experiment on yeast dataset with ten clusters.

4. CONCLUSIONS

This paper has focused on the problem of clustering the objects of complex dataset, especially the dataset which is affected by measurement error and data transmitter error. This paper has proposed novel fuzzy clustering algorithm by incorporating kernel induced distance function, entropy functions, weighted distance measure and neighborhood terms based spatial context. Through the experimental work on Yeast dataset, this paper evaluated the proposed method is capable in clustering the dataset which corrupted by measurement error and data transmitter error. The clustering accuracies of the proposed methods have shown the effectiveness of the proposed method in clustering the objects of the data which are corrupted by measurement error and data transmitter error.

ACKNOWLEDGMENTS

This work was supported by Indo Taiwan Project, DST India and NSC Taiwan.

REFERENCES

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

[2] Cheng H. D., Chen J., and J. Li, “Thresholding selection based on fuzzy c-partition entropy approach, ” Pattern Recognit., Vol 31, pp. 857–870, (1998).

[3] Dao-Qiang Zhang, Song-Can Chen, Clustering Incomplete Data Using Kernel-Based Fuzzy C-means Algorithm. Neural Processing Letters 18: 155–162 (2003).

[4] Dubes.R.C and Jain.A.K, Clustering methodology in Exploratory Data Analysis. In M.C.Yovits, ” editor, Advances in Computers, Academic press, New York, pp. 113-225 (1980).

[5] J. Dunn, “A fuzzy relative of the Isodata process and its use in detecting compact, well-separated clusters”, Journal of Cybernetics, 3(3), pp. 32–57, 1973.

[6] Jawahar C. V., Biswas P. K., and Ray A. K., “Investigations on fuzzy thresholding based on fuzzy clustering, ” Pattern Recognition, Vol 30, pp.1605-1613, (1997).

[7] Kanzawa, Y. Endo, Y. Miyamoto, S., Fuzzy classification function of entropy regularized fuzzy c-means algorithm for data with tolerance using kernel function, page 350-355, Granular Computing, 2008. GrC 2008, IEEE Xplore

[8] Makoto Yasuda, Deterministic and Simulated Annealing Approach to Fuzzy C-means Clustering, International Journal of Innovative Computing, Information and Control, Vol. 5, no. 12(B), pp. 4981-4991, 2009.

[9] Plant, C, Boehm, C, Tilg, B., & Baumgartner, C, Enhancing instance-based classification with local density: A new algorithm for classifying unbalanced biomedical data. Bioinformatics, 22, 981-988, (2006).

[10] Prabhjot Kaur, Pallavi Gupta, Review and Comparison of Kernel Based Fuzzy Image Segmentation Techniques, I.J. Intelligent Systems and Applications, 2012, 7, 50-60.

[11] ShengDun Hu; KinTak, U., A Novel Video Steganography Based on Non-uniform Rectangular Partition, IEEE International Conference on Computational Science and Engineering CSE/I-SPAN/IUCC 2011, DOI:

[12] Wen-Feng Kuo, Chi-Yuan Lin and Wei-Yen Hsu: Medical Image Segmentation Using the Combination of Watershed and FCM Clustering Algorithms", International Journal of Innovative Computing, Information and Control, Volume 7, Number 9, pp. 5255-5267, September 2011

[13] Yong Y., Chongxun Z., Pan L., A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding Measurement Science Review, Volume 4, Section 1, 11-19, (2004).

[14] Zadeh, L.A., Fuzzy sets., Inform. and Control 8, 338.353 (1965)

[15] Zhao, A. M. N. Fu, and Yan H., “A technique of three level thresholding based on probability partition and fuzzy 3-partition, ” IEEE Trans. Fuzzy Systems, Vol 9, pp. 469–479, (2001)