Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 1, Number 1; October, 2014 pp. 17-20
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

Robust Fuzzy Clustering Algorithms in Real World Database

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

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

***Abstract:** The widely used standard fuzzy c-means often fails in clustering nonlinear and high-dimensionality dataset because of its Euclidean based distance to measure the similarity between the objects. To enhance the robustness and clustering the non-Euclidean complex structured dataset, this paper proposes effective fuzzy c-means by incorporating the kernel induced distance function, entropy functions, threshold parameter, and neighborhood terms. In order to reduce the computational time and to avoid incorrect results, this paper proposes new prototype knowledge method. There are total of seven real life databases used to show the efficiency of proposed methods during experimental study of this paper and the superiority of the proposed methods is proved through the clustering accuracy using Silhouette method, Error Matrix, number of iteration and the strength of the membership.*

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

1. INTRODUCTION

Clustering plays vital rule in data mining technique that has a broad range of applications [5, 6, 8] in many areas such as medicine, business management, biology, pattern recognition and image analysis etc. Clustering is the process of separating a dataset of objects into different subsets such that the objects in each subset or cluster are similar to each other and dissimilar to the objects in other clusters. In essence, clustering represents objects in dataset by few clusters, and hence, it models data by its subgroups. Recently, Fuzzy clustering technique is extensively applied in analyzing dataset, since it does not require any prior information about the objects of data. Further it assigns each object in all the clusters with meaningful membership grades.

The standard Fuzzy *c*-means (FCM) algorithm is introduced by Bezdek [2], and then researchers have introduced modified FCM algorithms [4, 9, 10, 14] in order to nonlinear and high-dimensionality dataset. Though the modified fuzzy c-means algorithms are performed well than fuzzy c-means algorithm, they suffer in clustering the data which have heavy noise and high-dimensionality. The modified algorithms have incapable in updating prototypes and assigning correct memberships to the object belong to the data which has high-dimensionality and corrupted by heavy noise. In order to rectify the above said drawbacks in modified fuzzy c-means this paper introduces effective fuzzy clustering methods by incorporating the nonlinear transformation properties of kernel functions [13], entropy terms, thresholding parameter, neighborhood terms, and Lagrangian multipliers method [12]. The threshold parameter with fuzzy c-means is capable to have appropriate membership to the objects for their clusters in a dataset which have heavy noise and the entropy terms along with neighborhood terms [1, 3] of fuzzy c-means is capable in clustering the dataset which has affected by the noises such as measurement error, faulty equipment, and data transmission error. Initially this paper proposes entropy fuzzy c-means based hyper tangent kernel algorithm with thresholding parameter [KFCM_E_t] and it obtains equation for membership, updating centers by minimizing the algorithms. Secondly, this paper introduces entropy fuzzy c-means based hyper tangent with neighborhood term [KFCM_E_{ng}] to have desirable membership to the objects for placing them into appropriate clusters. Further this paper proposes a mathematical initialization method for initializing centers in order to speed up the algorithms.

The rest of this paper is organized as follows. We present our proposed algorithm in section 2. The experimental results on artificial data and real datasets of the proposed clustering methods are reported in Section 3. Finally, conclusion is presented in Section 4.

2. PROPOSED ALGORITHM

The modified objective function from the standard fuzzy c-means [2] is given by

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

where $\chi : X->F$. 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

$$\|\chi(x_i)-\chi(v_k)\|^2 = \langle\chi(x_i),\chi(x_i)\rangle+\langle\chi(v_k),\chi(v_k)\rangle-2\langle\chi($$

, where i = 1, 2, . . . , n, and k = 1, 2, . . . , c. We adopt Hyper Tangent Function to evaluate distance, that is, the $\chi(x_i, v_k)$ express using Hyper Tangent function as

$$\chi(x_i, v_k) = 1-\tanh\left(\frac{-\|x_i-v_k\|^2}{\sigma^2}\right) \qquad (2)$$

Here the second argument v_k is a center of clusters chosen from the given dataset X. This paper considered the σ^2 variance of given data. Using the expression (2) we obtained $\chi(x_i, x_i)$ = 1 and $\chi(v_k, v_k)$ = 1, so the distance function can be rewritten as $\|\chi(x_i)-\chi(v_k)\|^2 = 2(1-\chi(x_i, v_k))$ (3)

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

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

In order to cluster effectively the more complicated dataset which have corrupted by the noises such as measurement error, faulty equipment, and data transmission error, the entropy fuzzy c-means based hyper tangent kernel algorithm with thresholding parameter [KFCM_E$_t$] is introduced as

$$J(U,V,\lambda) = 2\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}(1-\chi(x_i, v_k)) + \frac{\gamma}{2}\sum_{i=1}^{n}\sum_{k=1}^{c}\frac{u_{ik}^{m}}{|x_i-\alpha_k|}\ln\alpha_k \quad (5)$$

Here α_k is the average value of data elements in the kth cluster and γ is the resolution parameter. The value of γ chosen from median of each classes. The KFCM_E$_t$ objective function is optimized to have high membership grades to the objects which are closed to their prototypes and to obtain low membership grades to the objects when the objects are far from their prototypes. By using the Lagrange multiplier, optimizing the objective function of KFCM_E$_t$, the equation for obtaining prototypes and membership grades are calculated. Optimizing Equation (5) under the membership constraint

$$u_{ik} = \frac{\left(\dfrac{1}{2(1-\chi(x_i, v_k)) + \dfrac{\gamma}{2|x_i-\alpha_k|}\ln\alpha_k}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\dfrac{1}{2(1-\chi(x_i, v_j)) + \dfrac{\gamma}{2|x_i-\alpha_j|}\ln\alpha_j}\right)^{\frac{1}{m-1}}} \qquad (6)$$

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. Optimizing the above objective function with respect to v_k, we have

$$v^t{}_k = \frac{\sum_{i=1}^{n}u_{ik}^{m}\chi(x_i, v_k^{t-1})\left(1+\tanh\left(\dfrac{-\|x_i-v_k^{t-1}\|^2}{\sigma^2}\right)\right)x_i}{\sum_{i=1}^{n}u_{ik}^{m}\chi(x_i, v_k^{t-1})\left(1+\tanh\left(\dfrac{-\|x_i-v_k^{t-1}\|^2}{\sigma^2}\right)\right)} \qquad (7)$$

Where v_k^{t-1} represents the k^{th} cluster center of the previous t-1^{th} iteration.

3. EXPERIMENTAL RESULTS ON WINE DATASET

The Wine data is widely used by researchers in comparing various classifiers [7, 11]. The wine data is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivators. The Wine data of three classes with 13 constituents consists of total 178 instances which are used for the purpose of experimental works. From the visualization of Wine data with all three cultivators in Fig. 1, this paper concluded that there is no resemblance between the 13 constituents.

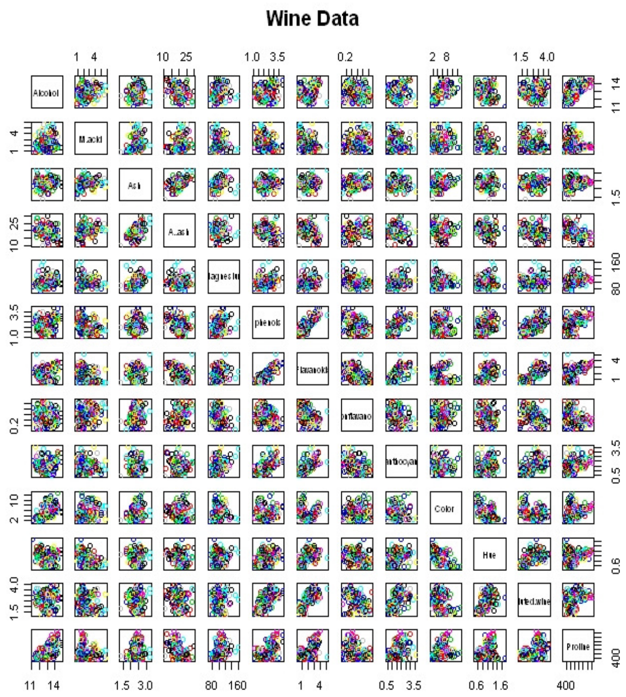

Fig. 1: 178 Wine Data

There are total three algorithms involved in this subsection, i.e., Standard FCM, and Proposed KFCM_E$_t$. The first experiment starts with Standard FCM on Wine and Fig.2 shows the size of obtained three clusters on 178 wine data with value of Silhouette Width [Sil.Wid.] of each cluster for representing the performance of clustering validity. The reallocated 178 wine dataset into three clusters using the experimental results of SFCM is given in Fig. 3 for getting the difference in the actual three classes.

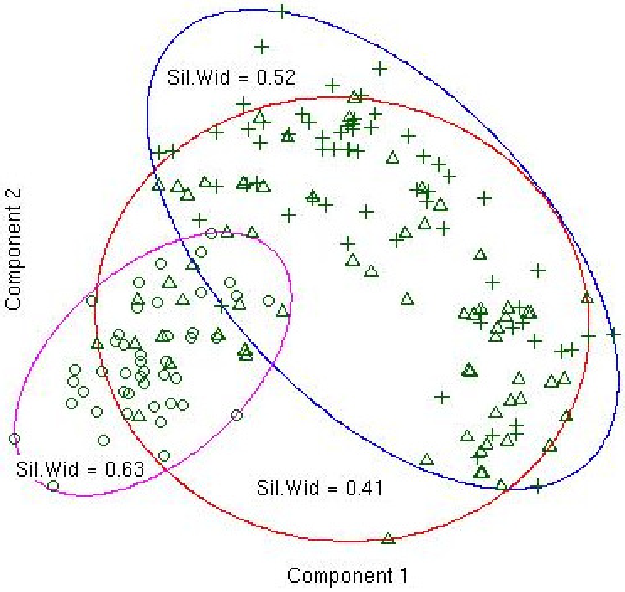

Fig.2: Obtained size of clusters and silhouette value by SFCM

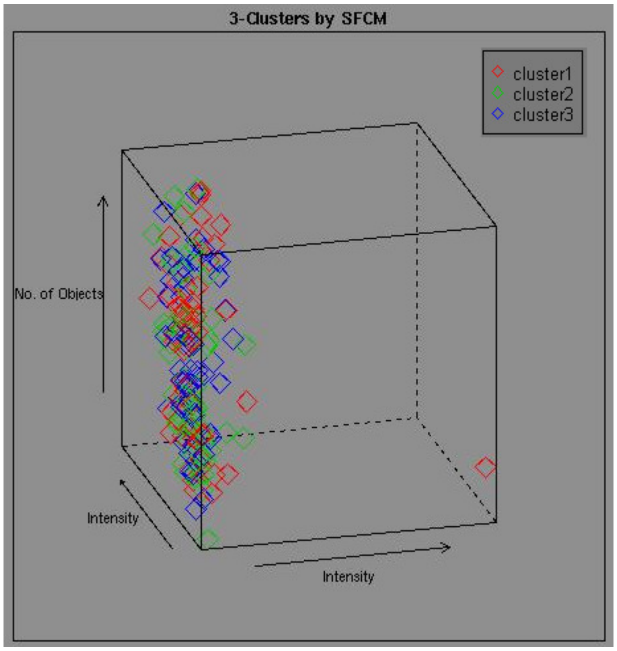

Fig. 3: Three Clusters of 178 Wine dataset by SFCM

In the second and third experiments we have introduced our proposed methods on Wine dataset and the obtained size of three clusters and their silhouette widths are given in Fig. 4. The reallocated data into three clusters based on partitioned results of proposed KFCM_E$_t$ is given in Fig. 5.

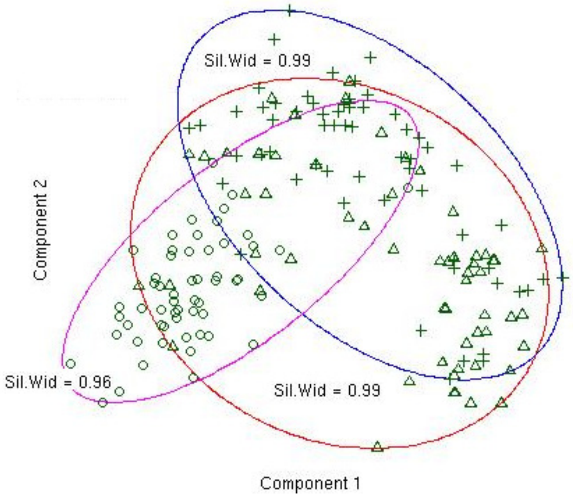

Fig. 4: Obtained size of clusters and silhouette value by KFCM_E$_t$.

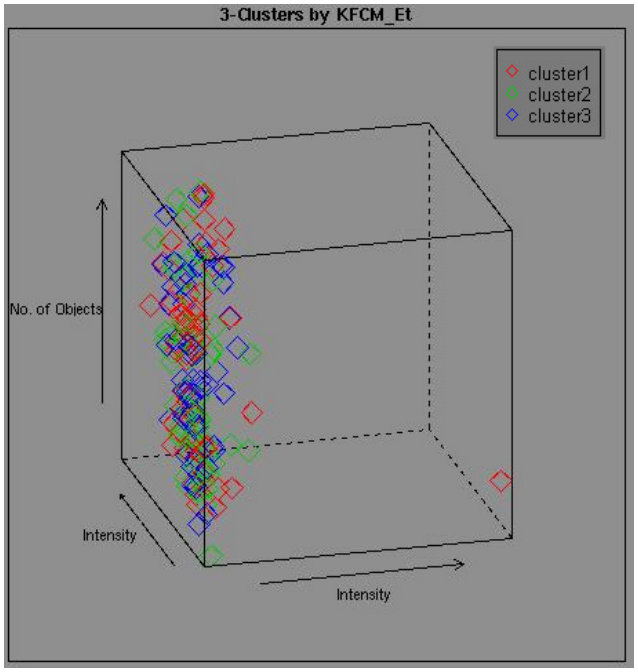

Fig. 5: Allocated 178 dataset by KFCM_E$_t$.

Table 1 gives the cluster validation using Silhouette width of the three algorithms on clustering Wine dataset into three different classes. From Table 1, the best clustering validity was obtained for Proposed KFCM_E$_t$ during the experiment on wine data with three clusters.

Table 1: Cluster Results on Wine Dataset. Silhouette Width [SW], Clustering Accuracy [CA], Minutes [M], Seconds[S], Running Time [RT] and Iterations [Its

Methods						
	No. of Objects in cluster(c)			**CA**	**RT**	**Its**
	C1	**C2**	**C3**			
SFCM	57	58	63	52%	2M	53
KFCM_E$_t$	48	64	66	98%	7S	7

4. CONCLUSION

This paper has presented the robust KFCM_E$_t$ for effective data clustering in general shaped and complex structured dataset. In order to speed up the algorithms; Centers Knowledge Method has been introduced by this paper. The effectiveness of the proposed methods has been validated using experimental study on Synthetic Image, and Wine Dataset. Experimental results on convergence of termination value, number of iterations, and updating centers have shown that the proposed KFCM_E$_t$ algorithm is superior to the corresponding existed algorithm. Overall, this paper has concluded that the proposed algorithms are most appropriate to cluster the complex dataset without any prior information of the objects.

5. ACKNOWLEDGEMENT

This work was supported by UGC-MRP Project (Ref.No.42-29/2013 (SR)), India.

REFERENCES

[1] Ahmed M. N, Yamany S. M, Mohamed. N, Farag. A. A, Moriarty T, A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data, IEEE Trans. Med. Imaging., Vol. 21, pp. 193-199 (2002).

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

[3] Chen S.C, Zhang D.Q., Robust Image Segmentation Using FCM With Spatial Constraints Based on New Kernel-Induced Distance Measure, IEEE Trans. Syst., Man, Cybern. B., 34(4) (2004), 1907-1916

[4] Cheng H. D et al., Thresholding selection based on fuzzy c-partition entropy approach, Pattern Recognit., Vol 31, pp. 857–870, (1998).

[5] Cover, T. M., Geomeasureal and Statistical Properties of Systems of Linear Inequalities in Pattern Recognition. Electron. Computing, Vol. EC-14, pp. 326-334 (1965).

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

[7] Erin J. Bredenstenier, Kristin P. Bennett, Multicategory Classification by Support Vector Machines, Computational Optimization and Applications, Vol.12, pp. 53–79 (1999).

[8] Girolami, M.: Mercer -Based Clustering in Feature Space. IEEE Transactions on Neural Networks, Vol. 13, Issue 3, pp. 780-784 (2002).

[9] Hsu-Shen Tsai et al., A Kernel-based Fuzzy C-means Algorithm with partition index maximization, Proceedings of 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp.391-394.

[10] 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).

[11] Jennifer G. Dy, Carla E. Brodley, Feature Selection for Unsupervised Learning, Journal of Machine Learning Research Vol.5, pp.845–889 (2004).

[12] Kanzawa et al., Fuzzy classification function of entropy regularized fuzzy c-means algorithm for data with tolerance using kernel function, In proceeding of the 2008 IEEE International Conference on Granular Computing, pp. 350-355, 2008.

[13] M¨uller K R, Mika S, R¨atsch G, Tsuda K, and Sch¨olkopf B, An introduction to kernel based learning algorithms, IEEE Transactions on Neural Networks, Vol. 12, No. 2, pp. 181–201 (2001).

[14] 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).