International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

EFFECTIVE UNSUPERVISED WAY OF CLUSTERING: ANALYZING COMPLEX DATABASE

S. R. Kannan

Department of Mathematics, Pondicherry Central University, (India)

ABSTRACT

This paper introduces some novel fuzzy c-means techniques to analyze the real world database for finding meaningful subgroups without any prior information about the data objects. The proposed methods have been implemented on synthetic dataset, and IRIS dataset to show the performance of the methods. The effectiveness of proposed methods have been accounted through running time, number of iterations, visual segmentation effects and clustering accuracy. The superiority of the proposed methods has been tested through Error Matrix.

Keywords: Clustering, Fuzzy C-Means, Complex Data Structure.

I. 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 similarity measure uses for distinguishing the difference between the data objects, the two data points belong to the same group if they are close to each other, and they are evidently from different groups if the distance between objects is specifically large. 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 [2, 4]. 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. A hard partition forces the full assignment of data point to exactly one of the clusters. Such hard assignment of data elements to cluster can be inadequate in presence of data points that are equally distant to two or more clusters. Due to the uncertain nature of many practical real world problems, hard partition is not suitable to cluster the real world data. Therefore the combination of fuzzy concept with clustering process deals the real life data elements extremely well than hard clustering algorithms, because fuzzy clustering algorithm allows partial membership in different clusters with reasonable membership grades. Even though there are lots of benefits while using FCM algorithm, still it has considerable drawbacks such as the result of clustering process deteriorates while noise and outliers exist in the dataset and it works well only on spherical shaped dataset not in general shaped dataset. To eradicate these shortcomings we proposed improved version of FCM algorithms. The new algorithms have been proposed using nonlinear transformations from the original pattern space into a higher dimensional feature space with properties

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

of kernel functions [3]. Using properties of kernel functions to inner product in the original space this paper tries to map the space into higher dimensional feature space. Therefore, the new fuzzy c-means leads to cluster general shaped dataset. The proposed methods elegantly locate the correct distance between cluster centers and data points using kernel induced distance. At the initial stage of the paper we propose novel kernel fuzzy c-means based entropy method algorithm and it drives equation for membership and updating centers from proposed novel fuzzy c-means. Secondly it introduces novel kernel fuzzy c-means with strong fuzzy parameter [6] based regularized terms to obtain strengthen memberships for objects in clusters. The rest of this paper is organized as follows. In Section 2, this paper briefs the basic fuzzy c-means. Section 3 contains the development of the proposed algorithms. The experimental results on Synthetic Dataset, and IRIS Dataset of the proposed clustering methods are reported in Section 4. Finally, conclusion and discussion are presented in Section 5.

II. FUZZY C-MEAN ALGORITHM

The idea of basic Fuzzy C-Means (FCM) has invented by Bezdek [1] in 1981. FCM provides a technique that how to group data points that populate some multidimensional space into a specific number of different clusters. Fuzzy c-means is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It is based on minimization of the following objective function

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

where c represents the number of clusters, n represents the number of unlabeled data points to be clustered, u_{ik} represents the fuzzy membership of the i^{th} data point in the k^{th} cluster, $V=\{v_k\}_{k=1}^{c}\subset R^N$ is the centroids of the clusters, and m represents the fuzzy parameter, $m = 2$.

The cluster centroids are obtained by minimizing the Eq. (1) with respect to V equal to zero:

$$v_k=\frac{\sum_{i=1}^{n}u_{ik}^{m}x_i}{\sum_{i=1}^{n}u_{ik}^{m}}\quad k=1,2,3,...,c$$

.....(2)

Minimizing the Eq. (1) with respect to U and setting it equal to zero, results in the following equation for fuzzy memberships:

$$u_{ik}=\left(\sum_{j=1}^{c}\left(\frac{\|x_i-v_k\|}{\|x_i-v_j\|}\right)^{\frac{2}{m-1}}\right)^{-1},\ k=1,2,3,...,c;\ i=1,2,3,...,n \qquad \text{..... (3)}$$

Eq. (3) is used to update the fuzzy membership values in an iterative framework. Note that when a data point in the solution domain has a maximum fuzzy membership measure within a cluster, this data point is considered to belong to this cluster**.** The following termination criterion is used to stop the calculation: $\|V^{(l+1)}-V^{(l)}\|<\varsigma$ .....(4) where $V^{(l+1)}$ and $V^{(l)}$ are the vector of cluster centroids at $(l+1)^{th}$ and $(l)^{th}$ iterations, and ς is a given tolerance bound.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

The above clustering procedures done subject to the following conditions:

$$\sum_{k=1}^{c} u_{ik} = 1 \;\; i = 1,2,3,\ldots,n$$

$$\sum_{k=1}^{c} u_{ik} > 0, \;\; k = 1,2,3,\ldots,c$$

III. PROPOSED FUZZY C-MEANS ALGORITHMS

3.1. Novel Kernel Fuzzy C-Means based Entropy Method [NKFCM_E]

At the initial stage, this paper introduces the NKFCM_E by the following objective function of fuzzy c-mean:

$$J(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik} \left\| \phi(x_i) - \phi(v_k) \right\|^2 + \delta^{-1} \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik} \log u_{ik} \qquad \ldots\ldots\ldots \quad (5)$$

where ϕ stands as map $x -> \phi(x) \in F$, and $x \in X$.

This paper considered the σ^2 variance of given data. We have novel kernel fuzzy c-means based entropy regularized function for considering ill-posed database as

$$J(U,V) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}.(1 - \phi(x_i, v_k)) + \delta^{-1} \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik} \log u_{ik} \qquad \ldots. (7)$$

The role of parameter δ is similar to the role of *m* in the fuzzy c-means. Here, the smaller δ, the fuzzier the solutions.

3.1.1 Membership & Prototype of NKFCM_E

The NKFCM_E objective function is minimized to have high membership grades to objects which are close to their prototypes and to obtain low membership grades to the objects when the objects are far from their prototypes.

In general, we have the membership from (7) as

$$u_{ik} = \frac{\{\exp[-2\delta(1 - \phi(x_i, v_k))]\}}{\sum_{j=1}^{c} \{\exp[-2\delta(1 - \phi(x_i, v_j))]\}} \qquad \ldots\ldots. (8)$$

The general equation is used to obtain membership grades for objects in data for finding meaningful groups. The precision of clustering results mainly depends on the cluster centers. Now minimizing the following objective function, this paper obtains the equations for updating the prototypes of our NKFCM_E.

$$v_k = \frac{\sum_{i=1}^{n} u_{ik} (\phi(x_i, v_k))(\phi'(x_i, v_k))x_i}{\sum_{i=1}^{n} u_{ik} (\phi(x_i, v_k))(\phi'(x_i, v_k))} \qquad \ldots\ldots \quad (9)$$

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

3.2 Novel Kernel Fuzzy C-Means with Strong Fuzzy Parameter based Entropy Method [NKFCM_E$_p$]

In order to have high memberships to object which is almost close or in equal distance of two prototypes of different clusters, a new objective function of fuzzy c-means with fuzzy parameter based entropy is introduced. To obtain more effective membership equation, the objective function of *NKFCM_E$_p$* as:

$$J(U,V)=2\left(\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}(1-\phi(x_i,v_k))+\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}(1-\phi(\eta_i,v_k))\right) \quad \ldots. (10)$$

Here η_i is the average value of the neighboring elements of x_i

We have the membership

$$u_{ik}=\frac{((1-\phi(x_i,v_k))+(1-\phi(\eta_i,v_k)))^{-\frac{1}{m-1}}}{\sum_{j=1}^{c}((1-\phi(x_i,v_j))+(1-\phi(\eta_i,v_j)))^{-\frac{1}{m-1}}} \quad \ldots..(11)$$

The general equation is used to obtain membership grades for objects in data for finding meaningful groups. Now minimizing the objective function, we obtain the equation for updating prototypes.

$$v_k=\frac{\sum_{i=1}^{n}u_{ik}^{m}(\phi'(x_i,v_k))(\phi(x_i,v_k))x_i+\sum_{i=1}^{n}u_{ik}^{m}(\phi'(\eta_i,v_k))(\phi(\eta_i,v_k))\eta_i}{\sum_{i=1}^{n}u_{ik}^{m}(\phi'(x_i,v_k))(\phi(x_i,v_k))+\sum_{i=1}^{n}u_{ik}^{m}(\phi'(\eta_i,v_k))(\phi(\eta_i,v_k))} \quad \ldots..(12)$$

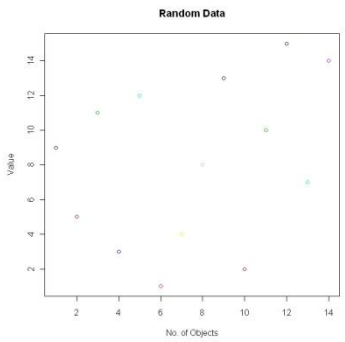
IV. EXPERIMENTAL RESULTS

4.1 Experimental results on artificial image

This subsection describes the experimental results on artificial image which is generated by random data given in Fig. 1. There are a total of three algorithms used in this section, i.e., standard FCM, NKFCM_E and NKFCM_E$_p$ for showing the performance of proposed methods. Note the value of σ in NKFCM_E and NKFCM_E$_p$ is a very important effect on performances of the algorithms. So far, the value of σ is constructed by choosing randomly, but this paper sets the value of σ by using Standard Deviation of the given data. In addition, we set the parameters β from dense of each cluster in each iteration of algorithm. This paper sets the termination parameter $\varepsilon = 0.001$ in all the experiments. Further we initialize the initial prototypes or centers of clusters using Method for Initializing Centers of Clusters for the experiments.

First experiment of this paper introduces the standard FCM algorithm to an artificial image which is generated by random data in Fig. 1. The artificial image includes two classes is given in Fig. 2. The results of standard FCM are given in Table 1 and in Fig. 3. Table 1 lists the memberships obtained for each object in final iteration of standard FCM and Fig. 3 gives the segmentation result of standard FCM. The standard FCM takes 14 iterations to converge the ε value.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346



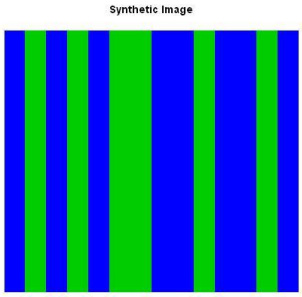
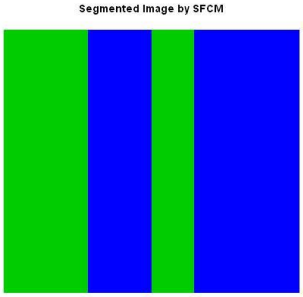


Fig.1. Random Data **Fig. 2. Corrupted Image**



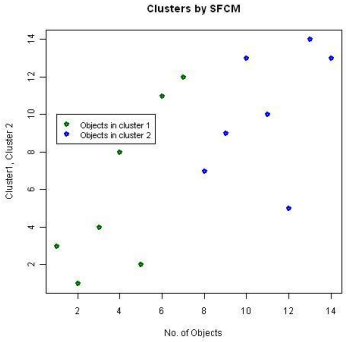


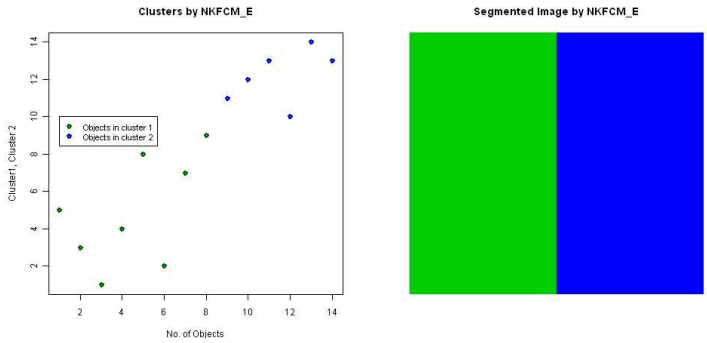
Fig3. Image by Standard FCM **Fig. 4. Clusters by Standard FCM**

Table1. Memberships of Final Iteration of Standard FCM

S.No	Objects	Membership for cluster 1	Membership for cluster 2
	9	0.30030063	0.69969937
	5	0.34530039	0.65469961
	11	0.69613438	0.30386562
	3	0.66657954	0.33342046
	12	0.72356954	0.27643046
	1	0.74302715	0.25697285
	4	0.67859250	0.32140750
	8	0.65052479	0.34947521
	13	0.21246849	0.78753151
	2	0.79441036	0.20558964
	10	0.32046422	0.67953578
	15	0.23221660	0.76778340
	7	0.39721622	0.60278378
	14	0.27771354	0.72228646

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Now this paper introduces the proposed NKFCM_E to cluster the artificial image into two clusters in order to test its effect on performance. This paper sets the initial cluster centers by using Prototypes knowledge Method. Fig. 5(a) and (b) show the results of NKFCM_E on synthetic image. It is observed from Fig. 5(b) that the NKFCM_E reduces the misclassification in colors, and it achieves better memberships to the objects for a particular cluster than standard FCM which are listed in Table 2. The algorithm obtains the results after five iterations of the algorithm.



(a) **(b)**

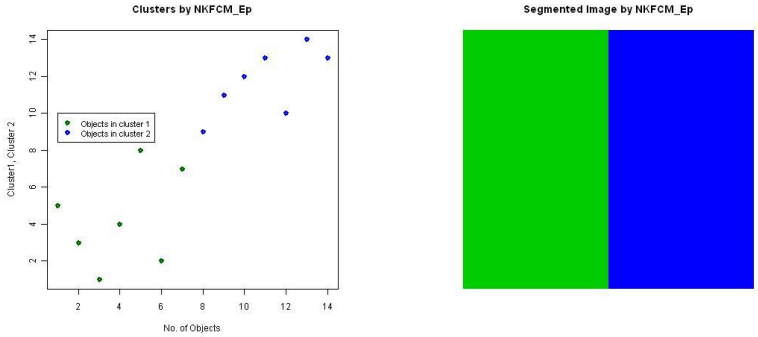
Fig.5 (a) Clusters by NKFCM_E (b) Image by NKFCM_E

Table2. Memberships of Final Iteration of NKFCM_E

S.No	Objects	Membership for cluster 1	Membership for cluster 2
1.	9	0.97969937	0.02030063
2.	5	0.98530039	0.01469961
3.	11	0.01613438	0.98386562
4.	3	0.97657954	0.02342046
5.	12	0.02356954	0.97643046
6.	1	0.98302715	0.01697285
7.	4	0.97859250	0.02140750
8.	8	0.97552479	0.02447521
9.	13	0.01246849	0.98753151
10.	2	0.98441036	0.01558964
11.	10	0.02046422	0.97953578
12.	15	0.02221660	0.97778340
13.	7	0.98721622	0.01278378
14.	14	0.02771354	0.97228646

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Finally this subsection implements the proposed NKFCM_E$_p$ on synthetic image and it sets the parameter $\beta = 2$. To test its effect on performance, we have initialized two cluster centers using Prototypes knowledge Method and divided the synthetic image into two partitions. Figs. 6(a) and (b) show the results of NKFCM_E$_p$. In NKFCM_E$_p$, the additional term β uses to improve the effect on performance of this method. From Fig. 6(b), it is observed that the numbers of misclassified objects reduced much and there are no improper ordered colors appear on image and it obtains strengthen membership to place the object in a particular cluster which are listed in Table3. It can also be seen from Fig. 6(b) that the NKFCM_E$_p$ is superior to the standard FCM algorithms. According to Fig. 5(b) and 6(b), under experimental approach of synthetic image, the proposed methods have achieved much better performance than standard FCM.



(a) **(b)**

Fig. 6 (a) Clusters by NKFCM_Ep (b) Image by NKFCM_Ep

Table 3. Memberships of Final Iteration of NKFCM_Ep

S.No	Objects	Membership for cluster 1	Membership for cluster 2
	9	0.01969937	0.98030063
	5	0.99630039	0.00369961
	11	0.00613438	0.99386562
	3	0.98757954	0.01242046
	12	0.00356954	0.99643046
	1	0.99302715	0.00697285
	4	0.99759250	0.00240750
	8	0.99552479	0.00447521
	13	0.00346849	0.99653151
	2	0.99441036	0.00558964
	10	0.01046421	0.98953579
	15	0.01021660	0.98978340
	7	0.99721622	0.00278378
	14	0.01071354	0.98928646

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Fig. 7 shows the comparison of the convergence of termination value of each iteration in the experiment of clustering the synthetic data into two clusters. The difference in clusters centers between current and previous iterations for reaching termination value are plotted in the Fig. 7 for showing the strength of proposed methods in obtaining results within short number of iterations.

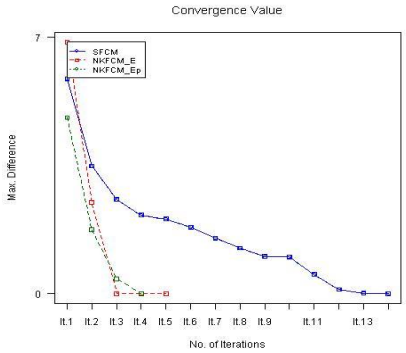
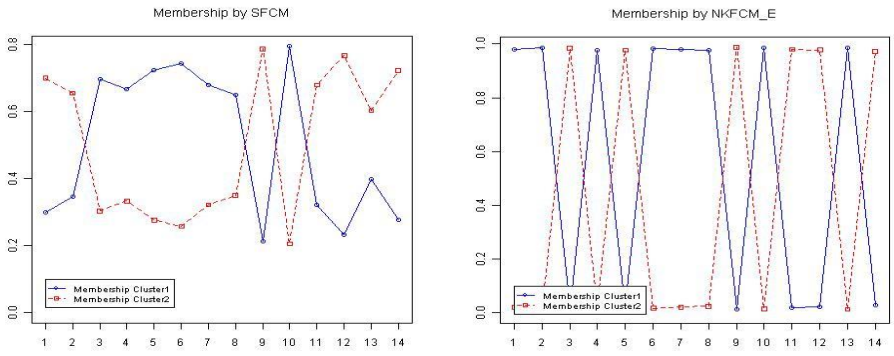


Fig. 7 Comparison of convergence of termination value under synthetic image by NKFCM_E & NKFCM_E_p

This paper has shown that the NKFCM_E and NKFCM_E_p have been converged the termination value within few numbers of iterations, and SFCM has taken large number of iterations to converge the termination value from Fig. 7.

In order to evaluate the effect of membership equations of proposed methods in obtaining memberships to objects on clustering data into appropriate clusters, the resulted memberships of this experimental study on synthetic image given in Fig. 8(a-c). It is observed from Fig. 8(a) that there is no wide difference between the memberships of the objects between the first and second clusters, it is because of poor distance measure of SFCM. From Fig. 8(b-c) it is observed that the proposed methods have wide difference in between the membership values for the objects for first and second clusters.



IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

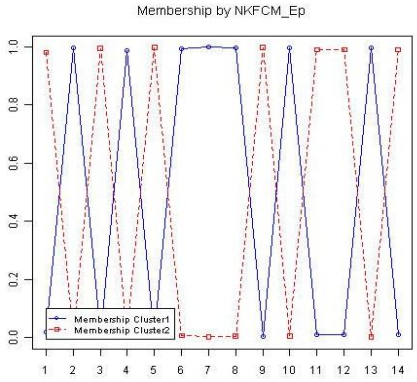


Fig. 8 Comparison of membership (a) by SFCM (b) NKFCM_E (c) NKFCM_E_p

Finally Table 4 shows the comparison of the number of iteration, running time, and clustering accuracy during the experiment of standard FCM, NKFCM_E and NKFCM_E_p, on synthetic image. The standard FCM takes 14 iterations to complete the experimental work on synthetic image for clustering it into two partitions, but the proposed methods have taken less number of iterations to complete the algorithms. It is clear from the all above observations; the proposed methods give better clustering results, clustering accuracy [5], and high memberships for clustering the data into two groups. Further the proposed methods require less running time, and less number of iterations to complete the experimental works.

Table 4. Comparison of Iteration Count, Running Time and clustering accuracy

	No. of Iterations	No. of clusters	Running Time	Clustering Accuracy
Standard FCM	14	2	1Minute	57%
NKFCM_E	4	2	4 Seconds	98.5%
NKFCM_E_p	4	2	3 Seconds	98.999%

4.2 Experimental results with IRIS Dataset

In order to investigate the effects of the proposed algorithms, we run it on 150 IRIS dataset [7] for clustering it into three clusters. We used all the attributes Petal length, Petal width, Special width and Special length for our experimental work, since these all carry the most information about the classes of the iris flowers. Now the Standard Fuzzy C-Means has divided the IRIS data into three different clusters for three different classes. Fig. 9 shows the size of classes or clusters and their silhouette value obtained by Standard fuzzy c-means algorithm. The reallocated 150 IRIS data into three clusters based on partitioned results of SFCM is given in Fig. 10. Figs. 11 - 12 show the size of three clusters and their silhouette value obtained by proposed algorithms. The reallocation of IRIS data into three clusters has done based on the partitioned results of proposed methods NKFCM_E and NKFCM_E_p and given in Figs. 13 – 14 respectively.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

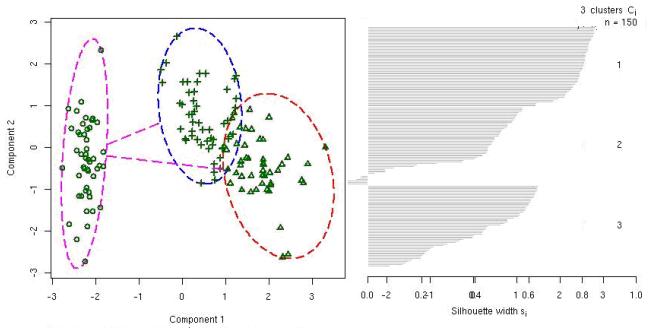


Fig.9 Obtained size of clusters and silhouette value by SFCM

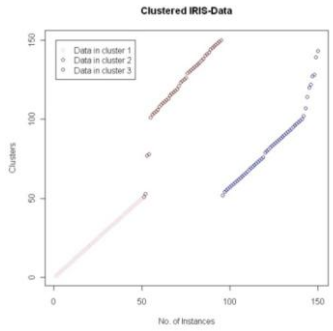


Fig.10 Three Clusters of 150 IRIS dataset by SFCM

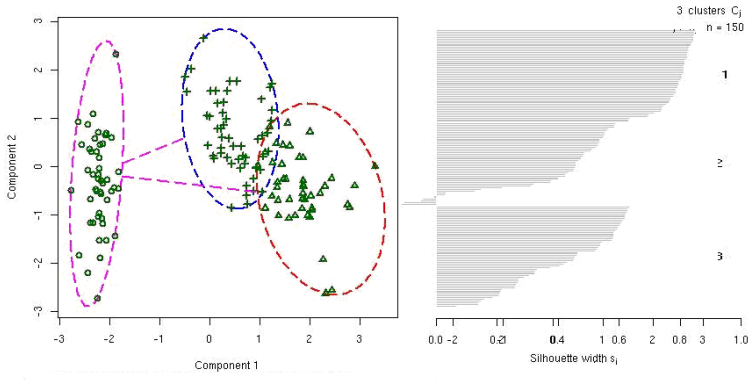


Fig.11 Obtained size of clusters and silhouette value by NKFCM_E

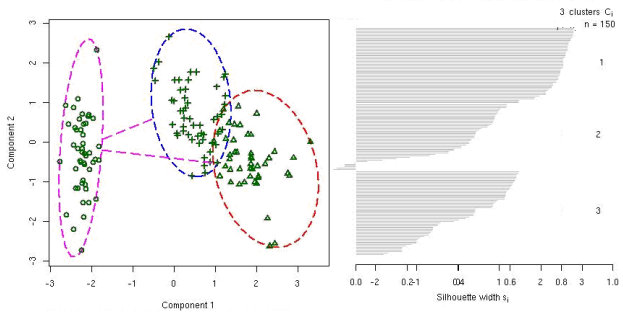
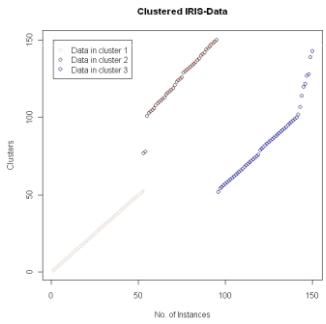


Fig.12 Obtained size of clusters and silhouette value by NKFCM_E_p

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346



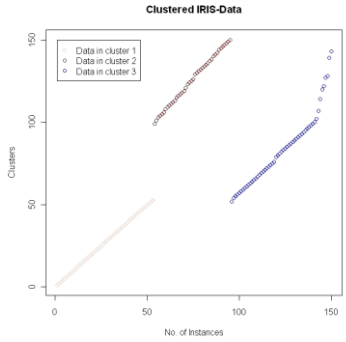


Fig.13 Allocated 150 dataset by NKFCM_E **Fig.14 Allocated 150 dataset by NKFCM_E_p**

Table 5 gives the cluster validation using Silhouette width on clustering IRIS data into three clusters during the experimental work of all three algorithms with 150 IRIS data. As shown in Table 5, the best clustering performance was obtained Using silhouette width with experimental results of three clustering algorithms. The highest silhouette average width has obtained for the clustering results during the experiment on IRIS data using NKFCM_E_p. It can be also seen from the Table 5 and cluster allocation plot Fig. 13-14 that the proposed algorithms not only getting good silhouette width for correctly partitions the data, it also separated well the three clusters in 150 IRIS Data.

Table 5: Comparison of Iteration Count, Running Time and clustering accuracy

	IRIS Data									
	No. of Objects in Cluster 1	SW	No. of Objects in Cluster2	SW	No. of Objects in Cluster3	SW	ASW	Accuracy	Running Time	No. of Iterations
SFCM	49	0.65	50	0.38	51	0.43	0.49	49%	1.4minutes	49
NKFCM_E	49	0.84	46	0.83	54	0.88	0.85	85%	5seconds	7
NKFCM_E_p	50	0.90	46	0.89	52	0.88	0.89	89%	5seconds	7

Table 6: Error Matrix

	SFCM	*NKFCM_E*	*NKFCM_E_p*
Accuracy	47 %	99.1%	99.2%

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

The error matrix Table 6 gives the accuracy of reference classes and the obtained classes in IRIS data by the methods involved in this experiment study. From Table 6, the best clustering accuracy was obtained for Proposed NKFCM_E_p during the experiment on IRIS data with three clusters with average silhouette width.

V. CONCLUSION

This paper introduced novel kernel fuzzy c-means for clustering real world database. To show the effectiveness of the proposed methods the Synthetic dataset and IRIS datasets are used. This paper has been proved the superiority of the proposed methods through the experimental results of cluster validation using silhouette width, error Matrix, running time, number of iterations and well separated clusters.

Acknowledgment: This work was supported by Indo Taiwan Collaborative Research Project.

REFERENCES

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

[2] Hassanien, A. E. (2004). Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer. Journal of the American Society for Information Science and Technology, 55(11), 954–962.

[3] 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

[4] Polat, K., & Gunes, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17(4), 694–701.

[5] Rousseeuw PJ (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[6] Tamer M., Ayvaz, Halil Karahan., Mustafa Aral M., Aquifer parameter and zone structure estimation using kernel-based fuzzy c-means clustering and genetic algorithm, Journal of Hydrology 343, 240– 253, (2007).

[7] UCI Benchmark repository: a huge collection of artificial and real world data sets, University of California Irvine. http://www.ics.uci.edu/~mlearn