Sindh Univ. Res. Jour. (Sci. Ser.) Vol.46 (2):123-128 (2014)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Automatic Generation of Fuzzy Membership Functions based on K-Means and EM Clustering

I. MALA[++], P. AKHTAR*, A. R. MEMON, T. J. ALI**

Faculty of Engineering, Science and Technology, Hamdard University, Karachi

Received 8th December 2013 and Revised 12th February 2014

Abstract: Fuzzy partitioning is based on expert opinion in generating fuzzy membership function. Automatic generation of fuzzy partitioning can be made for a given dataset, using different clustering algorithms. We have developed a fuzzy partitioning algorithm, based on clustering algorithm. For medical dataset, automatic generation of fuzzy membership function is done using K-Means and Expected Maximization (EM) clustering algorithms. According to our results, K-Means clustering is more accurate for fuzzy partitioning as compared to EM clustering based on our developed fuzzy partitioning algorithm.

Keywords: fuzzy partitioning algorithm, medical expert opinion, mean and standard deviation

1. INTRODUCTION

Definition of fuzzy set is done by its membership function. Any expert or group of experts determines this by knowledge acquisition. After defining the fuzzy sets, its associated membership functions are elaborated. The shape of any membership function depends on the application. Most problems of fuzzy logic comprises of triangular shape where the problem is linear in nature. These parameters are usually based on the domain person experience and/or are generated automatically. Many triangular membership functions applications are not accurately represented in linguistic terms and statistical approach is used by experts to automatically generate shapes. Membership function can be defined either by manual or automatic technique.

Many approaches can be considered for automatic generation of membership functions. In automatic generation, the expert is completely or partially removed from the process or is only used at the initial phase to guess membership function and after that values are fined tuned automatically. Modern soft computing techniques (fuzzy, neural or genetic algorithms) are more appropriate as compared to other methods or approaches.

Different methods are used to develop fuzzy membership function. In this research paper, fuzzy membership function is developed automatically in triangular shape based on K-Means and EM clustering algorithms. The dataset used is diabetes from the UCI repository and the generated fuzzy membership function is compared with that of the medical expert. Both K-Means and EM clustering algorithms results are generated in WEKA tool which is a data mining tool, developed in opened Java based system (Tiwari, *et al.*, 2012).

The paper is organized as follows. A literature review of K-Means and EM clustering is presented in Section 2. Section 3 discusses the fuzzy partitioning methodology. Section 4 presents results and discussion of fuzzy partitioning algorithm. The last section gives conclusion and suggests future research direction.

2. LITERATURE REVIEW

Data mining is knowledge mining or search for hidden patterns of knowledge from huge data sets. Clustering is one of the data mining techniques used for partitioning a set of data objects into subsets. The objects in each cluster are similar to one another, and dissimilar to objects in other clusters (Soni and Ganatra, 2012). The partition is performed automatically by clustering algorithm. Hence, clustering is useful technique to locate unknown groups within data not previously located.

Clustering is unsupervised classification or segmentation. Categorization of clustering is dependent on different methods like partitioning, hierarchical, density based, grid based, and neural network etc. (Rai and Singh, 2010). Both K-Means and EM algorithms depends on varying input parameters which are used to find natural clusters within given data sets. K-Means algorithms typically converge to a solution very quickly as opposed to other clustering algorithms. Expected maximization (EM) is statistical probabilistic model which comes under the category of unsupervised clustering where most of the parameters are not known. EM algorithms rely very heavily on the assumption that

[++]Correspondence author : Idris Mala idrismala@yahoo.com cell +92-346-2767766
*National University of Science and Technology, Karachi
**HITEC University, Taxila Cantt.

the initial partition values are close to the natural clusters of the given data.

i. K-Means Algorithm

Data is partitioned into K sets using the K-Means algorithm (MacQueen, 1967). The results are a set of k centers, each located at the centroid of the data for which it is the closest center. For the membership function, each data point belongs to its nearest center, forming a Voronoi partition of the data. The objective function that the K-Means algorithm optimizes is

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 , \quad \textbf{(1)}$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ in Equation (1) is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centers (Kumar, *et al.*, 2011).

The objective function gives an algorithm which minimizes the within-cluster variance (the squared distance between each center and its assigned data points). K-Means has a hard membership function, and a constant weight function that gives all data points equal importance (Hamerly and Alkan, 2002). K-Means is easy to understand and implement, making it a popular algorithm for clustering.
K-Means is easy to implement in various problems of market segmentation, computer vision, geo-statistics, astronomy and such other fields.

ii. Expectation-Maximization (EM) algorithm

EM is iterative algorithm and is a probabilistic model based approach to solve clustering problems. Different application for which EM algorithm is widely used is pattern recognition, computer vision, and speech processing (Han and Kamber, 2012). Clustering of data in EM algorithm is done by different manner than K-means. Unlike distance based or hard membership algorithms (such as K-Means), EM is used to construct proper statistical model of the data as being an appropriate optimization algorithm (Bradley, *et al.*, 1998). EM also works in similar manner like K-means by initially defining the numbers of clusters that are desired. EM initializes with values for unknown (hidden) variables. It converges to local maxima as it uses maximum, around the initial values.

The important factor in EM is selecting initial value which is based on some criteria. EM algorithm is based on three steps based on an iterative technique.

1) Initialize hidden variables

2) Estimates the unobserved variables with respect to the known variables
3) Compute the maximum likelihood for the unobserved data and then finally check for the stop condition

EM clustering algorithm calculates probabilities of cluster membership based on one or more probability distributions (Nasser, *et al.*, 2006). It then maximizes the overall probability or likelihood of data, given the (final) clusters. EM clustering is used for data clustering in machine learning and computer vision.

3. FUZZY PARTITIONING METHODOLOGY

Data partitioning can be developed using clustering algorithm, one of the technique of data mining (Verma, *et al.*, 2012). Clustering algorithm calculates the mean and standard deviation for each cluster. The input to clustering algorithm is dataset and number of clusters required of the given data. The output of each algorithm is mean and standard deviation of each cluster.

The algorithm for fuzzy partition works as follows:

Fuzzy Partitioning Algorithm

1) Read the data of attribute selected
2) Find min and max of the data
3) Select a clustering algorithm and apply to given attribute
4) Select k = 3 (3 clusters)
5) Get mean ($\bar{x}$) and standard deviation (σ) of all 3 clusters
6) let $\bar{x}_1$, σ_1 for low
   let $\bar{x}_2$, σ_2 for medium
   let $\bar{x}_3$, σ_3 for high
7) Low will be (min, $\bar{x}_1$, $\bar{x}_1 + \sigma_1$)
8) Medium will be ($\bar{x}_2 - \sigma_2$, $\bar{x}_2$, $\bar{x}_2 + \sigma_2$)
9) High will be ($\bar{x}_3 - \sigma_3$, $\bar{x}_3$, max)

The number of clusters is kept fixed that is three clusters for each data attribute, that is low, medium and high. **(Fig. 1)** show the fuzzy membership function developed using the fuzzy partition algorithm in which different clustering algorithm will result in different fuzzy partitioning. The dataset taken for this research is PIMA Indian Diabetes Dataset (Sigillito, 2011), where three attributes are of interest, which are sugar, age and blood pressure. Fuzzy membership function are calculated for these attribute using K-Means and EM clustering algorithm based on the above fuzzy partitioning algorithm.

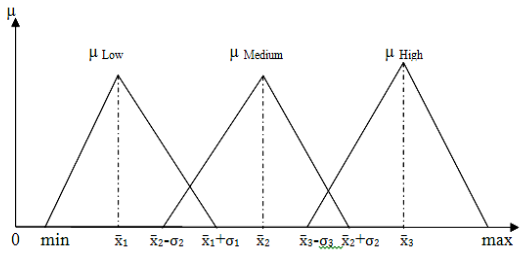

Fig.1. Fuzzy membership function using fuzzy partitioning algorithm

The membership function developed by the above algorithm based on different clustering technique is of triangular shape. We also have membership function based on medical expert opinion as shown in **(Table 1)**. A comparison between expert fuzzy membership and once calculated by our fuzzy partition algorithm will give analyses of how good are the automatic generating membership functions.

3. RESULTS AND DISCUSSION

The results of fuzzy partition algorithm are based on K-Means and EM clustering technique. The fuzzy membership function developed by the algorithm is compared with medical expert opinion as given in **(Table 1)**. Since the given values do not have overlapping membership function so assuming 50 % value to be overlapping, therefore each side will have 25 % extra.

Table 1. Categorical Partitioning of PIMA diabetic data set attributes based on Medical Experts opinion (Kargewoda, at el., 2012)

Attributes Name	***Partitioned Data***
Sugar	{low, medium, high} < 95 , 95 – 150 , >150
Age	{low, medium, high} { 20 - 34 , 35 – 46 , > 46 }
Diastolic Blood Pressure (BP)	{low, medium, high} { < 70 , 70 – 100 , > 100 }

Consider 0 to 95 for sugar and then let add 25 % of 95 to 150 which is 14. Therefore the low membership function of sugar will be 0, 95 and 95 + 14 = 109. The **(Table 2)** shows complete analysis for calculating membership function of attribute sugar, for all three linguistic value of low, medium and high. The diagram of fuzzy membership function is shown in **(Fig. 2, 3, 4)**.

Similarly, same calculations are made for age and blood pressure. **(Table 2)** also shows the membership function of age and blood pressure based on expert advice. Similarly **(Fig. 2, 3, 4)** also shows fuzzy partition diagram of both attributes.

Table 2. Fuzzy membership function calculation of Sugar, age and blood pressure Based on medical expert opinion

Sugar			
LOW	0	95	109
MED	95	123	150
HIGH	136	150	199
age			
LOW	20	29	37
MED	32	40	50
HIGH	43	55	81
Bp			
LOW	0	70	77
MED	70	85	100
HIGH	93	100	122

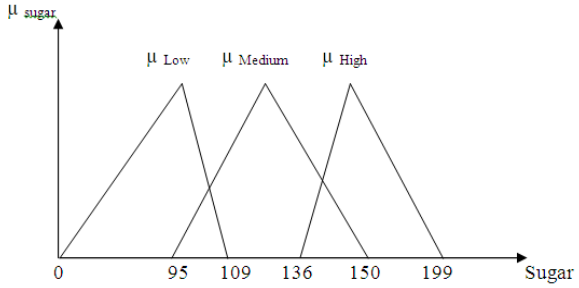

Fig.2. Fuzzy partition for sugar based on medical expert opinion

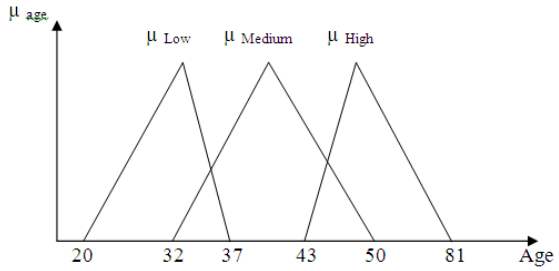

Fig.3. Fuzzy partition for age based on medical expert opinion

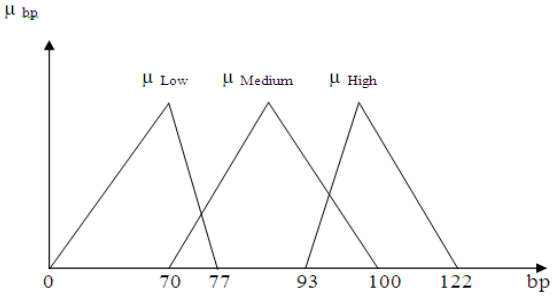

Fig.4. Fuzzy partition for blood pressure based on medical expert opinion

The diabetes dataset is applied to the fuzzy partition algorithm. The clustering technique used is K-Means and the results are shown in **(Table 3)**. From the results, fuzzy membership functions based on our fuzzy partition algorithm is calculated as shown in **(Table 4, 5, 6, and 7)**.

Table 3. K-Means Clustering Algorithm results (k = 3 clusters)

Attribute	*Cluster 1*	*Cluster 2*	*Cluster 3*
Sugar			
Mean	106	121	141
Std. Deviation	23	29	32
Age			
Mean	25	37	48
Std. Deviation	4	11	10
Blood Pressure			
Mean	65	70	77
Std. Deviation	18	21	13

Table 4. Fuzzy membership function calculation of sugar, age, and blood pressure based on fuzzy partition algorithm based on K-Means clustering

Sugar			
LOW	0	106	129
MED	92	121	150
HIGH	110	141	199
age			
LOW	21	25	29
MED	27	37	47
HIGH	38	48	81
Bp			
LOW	0	65	83
MED	49	70	91
HIGH	64	77	122

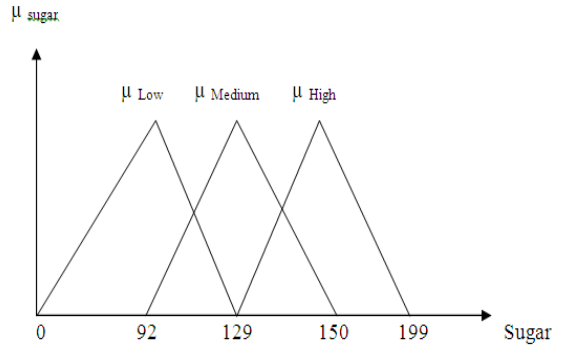

Fig.5. Fuzzy partition for sugar based on K-Means clustering technique using Fuzzy Partition Algorithm

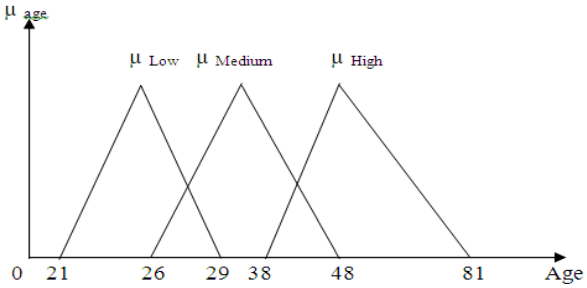

Fig.6. Fuzzy partition for age based on K-Means clustering technique using Fuzzy Partition Algorithm

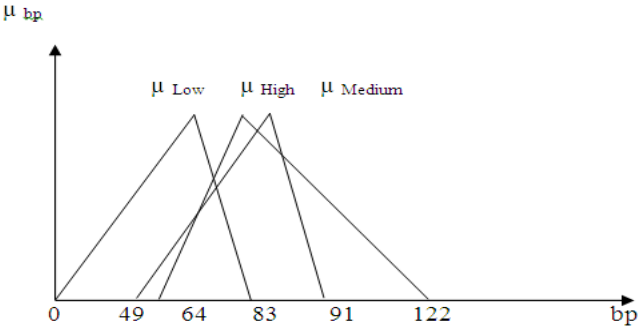

Fig.7. Fuzzy partition for blood pressure based on K-Means clustering technique using Fuzzy Partition Algorithm

Similarly for the same diabetes data and the clustering technique used is expected maximization (EM) the results are shown in **(Table 5)**. From the results, fuzzy membership functions based on our fuzzy partition algorithm is calculated as shown in **(Table 6) (Fig. 8, 9, 10)**.

Table 5. EM Clustering Algorithm results (k = 3 clusters)

Attribute	*Cluster 1*	*Cluster 2*	*Cluster 3*
Sugar			
Mean Std. Deviation	105 18	110 40	133 32
Age			
Mean Std. Deviation	24 3	30 8	39 11
Blood Pressure			
Mean Std. Deviation	30 35	66 10	77 10

Table 6. Fuzzy membership function calculation of sugar, age, and blood pressure based on fuzzy partition algorithm based on EM clustering

Sugar			
LOW	0	105	123
MED	70	110	150
HIGH	101	133	199
age			
LOW	21	24	27
MED	23	30	38
HIGH	27	39	81
Bp			
LOW	0	30	65
MED	56	66	76
HIGH	66	76	122

(Table 7) shows results of accuracy of K-Means and EM based on mean, comparing with medical expert opinion. Sugar attribute is comparable as both have accuracy above 95 %, but as compared to age and

blood pressure where EM accuracy decreases and is lower by 13 % with K-Means case of age and 16 % in case of blood pressure. Therefore, K-Means clustering gives better results than the EM clustering when considering the mean of the clusters.

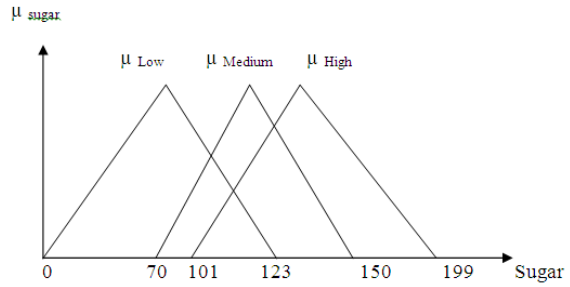

Fig.8. Fuzzy partition for sugar based on EM clustering technique using Fuzzy Partition Algorithm

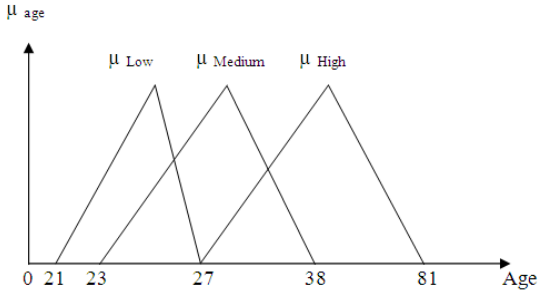

Fig.9. Fuzzy partition for age based on EM clustering technique using Fuzzy Partition Algorithm

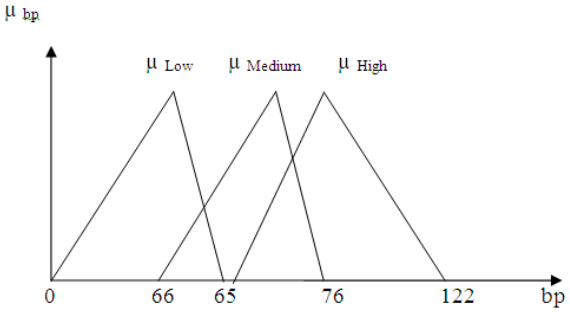

Fig.10. Fuzzy partition for blood pressure based on EM clustering technique using Fuzzy Partition Algorithm

Table 7. Accuracy of Fuzzy partition algorithm based on mean

Attributes	Medical Expert SD	K-Means Clustering Mean Accuracy	EM Clustering Mean Accuracy
Sugar	122	122 100.00%	116 95.08%
Age	41	36 88.71%	31 75.61%
Blood Pressure	85	70 83.14%	57 67.06%

Standard deviation accuracy as shown in **(Table 8)** is more or less similar for both K-Means and EM, and is greater than 95 %. Both techniques results are comparable with the medical expert opinion analyses of standard deviation.

Table 8. Accuracy of Fuzzy partition algorithm based on standard deviation

Attributes	Medical Expert SD	K-Means Clustering SD	K-Means Clustering Accuracy	EM Clustering SD	EM Clustering Accuracy
Sugar	51.66	50.59	97.93%	51.67	99.98%
Age	16.76	17.25	97.08%	17.49	95.64%
Blood Pressure	32.26	31.19	96.68%	31.51	97.68%

(Table 9) calculates the difference percentage of fuzzy partition data regarding left, medium and right side values as compared to medical expert fuzzy membership function. The variation is from 5.62 % to 1.46 % in case of K-Means clustering and 4.56 % to 0.51 % in case of EM clustering, which shows the results of EM clustering are better than K-Means.

Table 9. Error of difference with medical expert opinion and Fuzzy partition algorithm based on clustering technique

Attributes	K-Means Clustering Left	K-Means Clustering Center	K-Means Clustering Right	EM Clustering Left	EM Clustering Center	EM Clustering Right
Sugar	4.19%	1.99%	1.46%	4.33%	1.81%	0.51%
Age	3.86%	3.77%	2.18%	4.56%	4.17%	2.18%
Blood Pressure	5.58%	5.62%	1.67%	4.19%	5.42%	2.01%

Overall comparison of K-Means and EM clustering in our Fuzzy Partition Algorithm results that K-Means technique provides better mean of clusters, and results are more comparable to medical expert opinion than EM but as far as difference is values are concerned, EM provides better results. Both techniques can be used depending on the requirement of specific domain, and data can vary with change in domain. This analysis is made for diabetic domain of PIMA Indian diabetes data from UCI repository.

Similar results have been shown by K-Means and EM clustering algorithms when tested for software quality and performance, and was found that both clustering algorithm performed better results in terms of accuracy compared to Self-Organization Map (SOM) and hierarchical clustering algorithms (Abbas, 2008). EM algorithm is used for 3D medical imaging in MRI and results provided useful clinical measurement (Lynch, *et al.*, 2007). K-mean and EM performance are comparable but EM fails in high dimensional data set (Alldrin, *et al.*, 2003).

We have shown similar results as discussed above in fuzzy partitioning algorithm using the K-Means clustering and EM clustering algorithms where K-Means provide comparable results to medical expert opinion than EM clustering because of its local maxima problem which results in deviation of the mean of clusters. The fuzzy partitioning algorithm can be used for heart dataset for generating fuzzy membership function for attribute like blood pressure, heart rate, age and cholesterol etc. with our developed FDSL tool (Mala, *et al.,* 2013) for fuzzy relational database management system.

4. CONCLUSION

Automatic generation of fuzzy membership function using our fuzzy partitioning algorithm is demonstrated. The result of diabetes dataset is first developed using fuzzy membership function by medical expert opinion. Similarly, fuzzy membership function is generated by our fuzzy portioning algorithm for both K-Means and EM clustering. The results compared with medical expert shows that K-Means performs better when accuracy is measured based on mean of clusters and is comparable with EM clustering in case of accuracy in case of standard deviation. EM clustering results is better by 2 to 3 percent compared to K-Means when considering the variations in data in terms of low, medium and high values but overall both performed equally better and data is varied only by 5 %. K-Means clustering is to be used in this case of diabetic dataset.

Automatic partitioning in case of other medical domain like heart dataset or liver disorder dataset etc. would be future direction and different clustering technique used to get better results and should be comparable to medical expert opinion.

REFERENCES:

Abbas, O. B., (2008) "Comparison between data clustering algorithms". The International Arab Journal of Information Technology. Vol. 5, No. 3. 320-325

Alldrin, N., A. Smity, D. Turnbull, (2003). "Clustering with EM and K-Means," [Online], Available: http://cseweb.ucsd.edu/~atsmith/project1_253.pdf [6 September 2012]

Bradley, P. S., U.M. Fayyad, C. A. Reina. (1998) "Scaling EM (Expectation-Maximization) Clustering to Large Databases". Microsoft Research Technical Report MSR-TR-98-35, Redmond, WA.

Hamerly, G., C. Alkan. (2002) "Alternatives to the k-means algorithm that find better clusterings". Proceedings of the eleventh international conference on Information and knowledge management. 600-607, CIKM-2002. November 4-9, McLean VA.

Han, J., M. Kamber, J. Pei. (2012) "Data Mining Concepts and Techniques (3rd Edition)". The Morgan Kaufmann Series in Data Management Systems.

Karegowda, A., G. Punya, M. A. Jayaram, A. S. Munjunath. (2012) "Rule based Classification for Daibetic Patients using Cascaded K-Means and Decision Tree C4.5'. International Journal of Computer Application. Vol. **(45):**. Issue. 12. 45-50.

Kumar, K. N., G. N. Kumar, C. V. Reedy. (2011) "Partition Algorithms – A Study and Emergence of Mining Projected Clusters in High-Dimensional Dataset". International Journal of Computer Science and Telecommunications. Vol. **(2):** Issue 4. 34-47.

Lynch, M. D. Ilea, K. Robinson, O. Ghita, P. F. Whelan. (2007) "Automatic seed initialization for the expectation-maximization algorithm and its application in 3D medical imaging". Journal of Medical Engineering & Technology. Vol. **(31):**. No. 5. 332-340.

MacQueen, J. B. (1967) "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1: 281-297.

Mala, I., P. Akhtar, A. R. Memon, T. J. Ali (2013) "FDSL Tool : An Approach of Fuzzy Relational Database Management System". Life Science Journal. Vol. **(10):**.Issue. 2. 1606-1612.

Nasser, S., R. Alkhalidi, G. Vert. (2006) "A Modified Fuzzy K-means Clustering using Expectation Maximization". IEEE International Conference on Fuzzy Systems. 231-235. Vancouver BC.

Rai, P., S. Singh. (2010) "A survey of clustering techniques". International Journal of Computer Application. Vol. **(7):**. 10. 1-5.

Sigillito, V., (2011) "Pima Indian Dataset", UCI Dataset Repository. The Johns Hopkins University. URL: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes [28 February 2011]

Soni, N., A. Ganatra. (2012) "Categorization of Several Clustering Algorithms from Different Perspective: A Review". International Journal of Advanced Research in Computer Sci. and Software Eng. Vol. **(2):** 8. 63-68.

Tiwari, M., M. B. Jha, O. Yadav. (2012) "Performance analysis of Data Mining algorithms in Weka". IOSR Jour. of Computer Eng. (IOSRJCE). Vol. **(6):** 3. 32-41.

Verma, M., M. Srivastava, N. Chack, A. K. Diswar, N. Gupta. (2012) "A Comparative Study of Various Clustering Algorithms in Data Mining". International Journal of Engineering Research and Application. Vol. **(2):** 3. 1379-1384.