ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Subhalakshmi et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

TRACKING OF CHNGES IN CANCER DATA USING SOFTCLUSTERING

Chatti Subbalakshmi[1*], G Rama Krishna[2], and S Krishna Mohan Rao[3]

[1]*Guru Nanak Institutions Technical Campus, Dept. of CSE, Hyderabad, Telangana, INDIA*
[2]*K L University, Dept. of CSE, Vijayawada, AP, INDIA*
[3]*Sidhartha Engineering College, Dept. of CSE, Hyderabad, Telangana, INDIA*

ABSTRACT

Data mining is process of extracting the knowledge from large amount of data and its methods are being used efficiently in biological and biomedical applications. Soft computing is for intelligent management systems and its components are fuzzy logic, evaluation computing and genetic algorithms. In recent years, combinations of data mining with soft computing approaches are more suitable for bio-informatics. Clustering is a process of unsupervised learning algorithm in data mining which can be implemented using soft computing approaches. Fuzzy Clustering or soft clustering is based on fuzzy set theory. It groups the data based on partial membership function and it assigns data point more than one cluster. It is used in many biomedical databases, like gene expression, protein sequences; image processing and image segmentation is main step for detection of cancer. In this paper, we proposed dynamic fuzzy clustering algorithm to identify changes in cluster structure. We applied on Wisconsin breast cancer data is collected periodically grouped into eight with class (benign / malignant). We executed fuzzy c-means clustering in individual groups and also executed dynamic fuzzy clustering by incrementally adding instances of groups. We presented our observation of results in both cases which can be support the analysis of cancer state between the instances of class.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 16th –Aug-2015

KEY WORDS

Data mining; soft computing; breast cancer data; dynamic fuzzy clustering;

***Corresponding author:** **Email:** subbalakshmichatti@gmail.com; **Tel.:** +91-40-9032312260

INTRODUCTION

In Data analysis, data mining plays major role in many application databases like business intelligence, science and engineering, bio-informatics, medical analytics. Data mining is part of "Knowledge Discovery in Databases" (KDD) [1] and it is computational process of finding patterns in large databases. It is set of methods with association of artificial intelligence, machine learning, statistics and database systems to handle different types of data. The main data mining tasks are supervised and unsupervised learning methods and the most frequently used unsupervised learning method is cluster analysis [2]; process of grouping most similar objects into clusters depend on similarity measures.

There are many clustering models are present in literature, connectivity models, centroid models, density models, graph-based models. All these algorithms are called hard clustering or exclusive clustering, as they assign data point to only one cluster as it uses conventional crisp set theory. The many complex applications in biology, medicine, the humanities, management sciences require mathematics and analytical methods with uncertain and unpredictable. But these algorithms do not handles application databases with uncertainty, imprecision, partial truth and approximation characteristics, hence soft computing approaches are introduced in cluster analysis to support these databases. The components of soft computing are Fuzzy logic, Neural networks [3], Evolutionary computation [4] and Support Vector Machines [5] and has been most frequently applied successfully in Bioinformatics and Biomedicine in recent years [6].

Fuzzy logic deals with approximate and value ranges in between 0 and 1and it was introduced with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh [7, 8]. It had been applied to several areas, from control areas to intelligence systems. The fuzzy clustering or soft clustering uses fuzzy set theory for grouping data objects and it assigns partial membership value for each data objects to each cluster. Fuzzy c-means is centroid based fuzzy clustering algorithm and it uses degree of membership value to cluster data point [9, 10].

THE IIOAB JOURNAL
ISSN: 0976-3104

In this paper, we have used soft clustering method, i.e. Fuzzy c-means is to group the cancer data. Fuzz c-means takes number of clusters in prior to execution and it has to update as data change. The objective of this paper is, we consider the problem of clustering on cancer data set and identifying the changes in the data as data is added periodically. For that, we selected Wisconsin breast cancer database is collected from UCI repository which was obtained from University of Wisconsin hospitals, Madison from Dr. William H Walberg. They collected breast cancer instances continuous eight months and grouped into eight. Each instance defined by nine features and one class (benign / malignant). We implemented fuzzy c-means clustering on individual groups of instances and then applied dynamic fuzzy c-means algorithm by incrementally adding instances of groups in R data mining software.

The paper is organized as, related work is given in section 2, dynamic fuzzy clustering in section 3, results are given in section 3 and conclusion is mentioned in section 4.

RELATED WORK

Fuzzy clustering

The fuzzy clustering defined by the fuzzy set theory and it is a process of grouping the objects by allowing the concept of partial membership, in which each object can belong to multiple clusters. For all data object, it assigns the membership values between 0 to 1 represents fit in for each cluster and the sum of the membership values of each data objects to all clusters must be 1. The high membership value shows more likely that data object belongs to that cluster. The most widely used fuzzy clustering algorithm is Fuzzy C-Means (FCM) [11].

Given a set of n data objects, p_k=p_1, p_2, p_3,, p_n the algorithm minimizes a weighted within group of sum of squared error an objective function shown in equation (1).

$$J = \sum_{i=1}^{n}\sum_{k=1}^{c} \mu_{ik}^{m} \, |p_i - v_k|^2 \qquad (1)$$

Where J is objective function, n is number of data objects, c is cluster number, μ_{ik} membership value, p_i is data object, v_k is center of cluster k, m is fuzziness factor value always greater than one. The center of k^{th} cluster can be calculated using equation (2) as,

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{m} p_i}{\sum_{i=1}^{n} \mu_{ik}^{m}} \qquad (2)$$

The fuzzy membership value can calculated using equation (3) as,

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (|p_i - v_k| / |p_i - v_j|)^{\frac{2}{m-1}}} \qquad (3)$$

The fuzzy c-means is objective function-based clustering method, which takes cluster number determined before the execution of algorithm. The deficiency of algorithm is, it does not identify noise and outliers as it uses sum of square error objective function. The second deficiency FCM, defined by Krishnapuram and Keller is that due to the constraint on membership shows as degree of sharing, but not as degree of possibility of a point belonging to a class. Mainly it deals with similarity between perfectly described objects, i.e. all feature values are exactly known and it does not deal with the uncertainty included by missing or incorrect data.

Cluster validity

The internal cluster validity can be done by using Silhouette cluster validity index is defined as [12, 13]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (4)$$

Where *a (i)* is average dissimilarity between data point (*i*) *and* all other data within the similar cluster and *b(i)* is the minimum average dissimilarity of *i* to all other cluster. The positive value of s(i) indicates the correct clustering and negative value shows the incorrect clustering.

THE IIOAB JOURNAL
ISSN: 0976-3104

DYNAMIC FUZZY CLUSTERING

The main objective of dynamic fuzzy clustering is to revise the initial parameters of the algorithm when new data is added. Fuzzy c-means clustering algorithm requires number of clusters as an input value and this value is defined by data size. When data size changes number of clusters might be changes. Therefore, for each new incoming data cycle of dynamic fuzzy clustering as illustrate in **Figure– 1**.

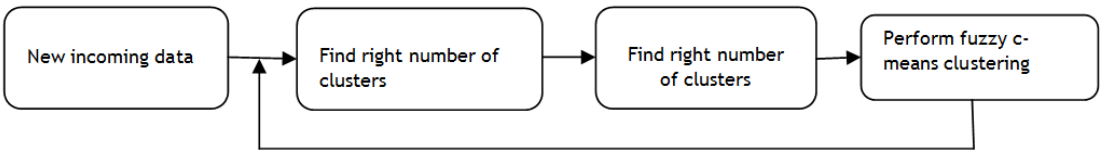

Fig: 1. Cycle of dynamic fuzzy clustering algorithm.

The general steps of algorithm are given as:

Initial clustering:

Step1: find the size of data set *n* for $D_{initial}$;
Step 2: for n= 1 to n-1/2
Calculate cluster average silhouette width S for each cluster;
Step 3: find c with maximum cluster average silhouette width;
Step 4: execute fuzzy c-means clustering algorithm;

For each new incoming data D_{new}

Step 5: add new data with old data as,
$D = D_{inital} + D_{new}$
Repeat step1 to step4
In this algorithm, we are adding the data incrementally and finding the cluster changes in data.

RESULTS

We have used Wisconsin breast cancer database is collected from UCI repository which was obtained from University of Wisconsin hospitals, Madison from Dr. William H Walberg. They collected samples periodically and made into eight groups as:

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated database on 15 July 1992)
Each instance consists of 9 attributes and one class (benign/malignant) with range of values as:

#	Attribute	Domain
1.	Clump Thickness	1 - 10
2.	Uniformity of Cell Size	1 - 10
3.	Uniformity of Cell Shape	1 - 10
4.	Marginal Adhesion	1 - 10
5.	Single Epithelial Cell Size	1 - 10
6.	Bare Nuclei	1 - 10
7.	Bland Chromatin	1 - 10
8.	Normal Nucleoli	1 - 10
9.	Mitoses	1 – 10
10.	Class	2 for benign and 4 for malignant

THE IIOAB JOURNAL
ISSN: 0976-3104

Results of fuzzy c-means on individual groups of cancer data

For each group of data, we executed fuzzy c-means clustering after finding right number of clusters using silhouette cluster validity index. Our observations are for all individual eight groups of data resulting only two right clusters which indicate only two class values (benign/malignant) exits. From the results of fuzzy c-means on group1 data, some of the instance's membership values partially belongs to benign and malignant. But in next seven groups, the Dunn's partition coefficient indicates, there is no overlapping instance which shows instance belongs to any one class. From the result of eight groups, we have identified that silhouette width gradually increasing, i.e. instances are very close to each other with same characteristic. The comparative results between each group are given in **Table–1**, the centers of eight group clusters are given **Table–2** and for each group fuzzy c-means cluster centers are shown in **Figure –2**.

Table: 1. Results of fuzzy c-means clustering algorithm for eight groups of cancer data

	Group 1 with 367 instances	Group 2 with 70 instances	Group 3 with 31 instances	Group 4 with 17 instances	Group 5 With 48 Instances	Group 6 With 49 instances	Group 7 With 31 Instances	Group 8 With 86 instances
Clus Avg Sil width	0.69340811	0.87437541	0.8765545	0.9372129	0.8434955	0.8871365	0.7294244	0.8878145
Right c value	2	2	2	2	2	2	2	2
Dunn_coeff	0.8041438	0.9035008	0.904723	0.9583406	0.897402	0.9212774	0.8312131	0.9178934
Belgian	207	14	8	3	11	09	19	74
Malignant	160	56	23	14	37	40	12	12

Table: 2. Fuzzy c-means cluster centers of eight groups of cancer instances

	Cluster centers	Clump thickness	Uniformity cell size	Uniformity cell shape	Marginal adhesion	Single Epithelial Cell Size	Bare nuclei	Bland chromatin	Normal nucleoli	Mitoses
Group 1	Cluster1	2.160705	2.838931	1.428396	1.54586	1.368898	2.23415	1.551031	2.635899	1.456337
	Cluster 2	3.898318	7.367233	6.570298	6.643348	5.579655	5.680445	8.065509	5.63092	6.320495
Group 2	Cluster1	2.015703	2.707998	1.578677	1.718243	1.390675	2.164551	1.185944	1.909195	1.315732
	Cluster 2	3.831131	8.082565	8.120871	7.860144	5.168349	5.772975	8.372444	6.79263	5.386354
Group 3	Cluster1	3.9925	7.772094	6.34029	6.367802	6.603301	5.100012	8.420575	8.448891	7.173032
	Cluster 2	2.061884	3.835446	1.178616	1.313346	1.882224	2.008565	1.580559	1.127177	1.069119
Group 4	Cluster1	3.997523	7.093996	8.31591	8.357888	8.232472	6.374376	8.301134	8.709004	8.441122
	Cluster 2	2.003191	4.294436	1.218299	1.290711	1.355207	1.791213	1.012295	1.150615	1.009591
Group 5	Cluster1	3.947392	6.252951	7.346187	7.155194	7.63559	4.293177	8.736356	6.53448	4.95016
	Cluster 2	2.063439	3.20521	1.211239	1.190368	1.313407	1.837813	1.287205	1.704877	1.044458
Group 6	Cluster1	2.017287	3.464124	1.237472	1.303173	1.165077	1.956136	1.217999	2.229199	1.198536
	Cluster 2	3.977982	7.076954	7.993649	7.721948	6.686502	4.84257	8.063018	7.784607	6.933361
Group 7	Cluster1	2.088273	3.903984	1.265267	1.599191	1.55003	2.040429	1.259187	1.774887	1.156555
	Cluster 2	3.99606	6.438842	7.514016	7.442704	7.437029	5.106654	7.721833	7.746245	5.172701
Group 8	Cluster1	2.013597	2.863381	1.216343	1.392772	1.305293	2.085176	1.162901	1.657168	1.114918

THE IIOAB JOURNAL
ISSN: 0976-3104

	Cluster 2	3.93992	5.889319	9.011659	8.241414	6.810015	5.551078	5.622764	7.155445	3.05486

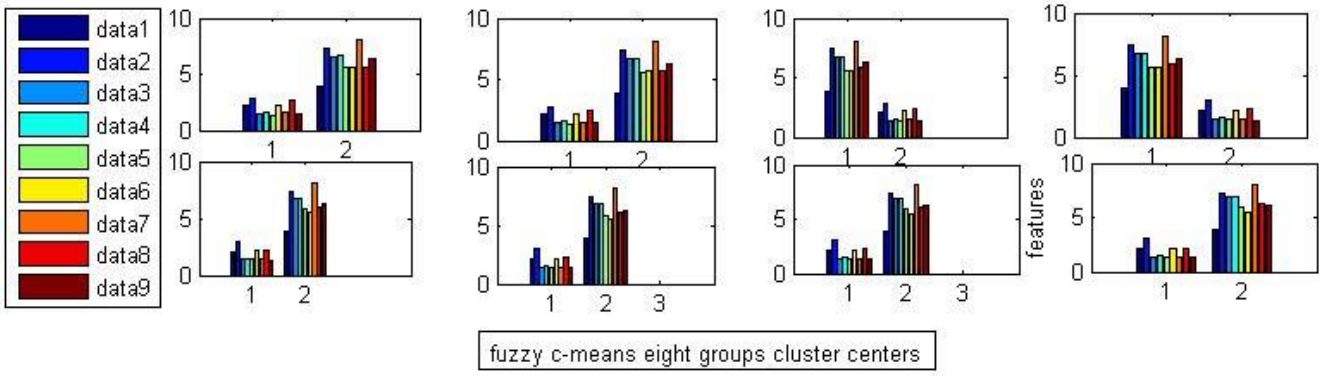

Fig: 2. Fuzzy c-means cluster centers.

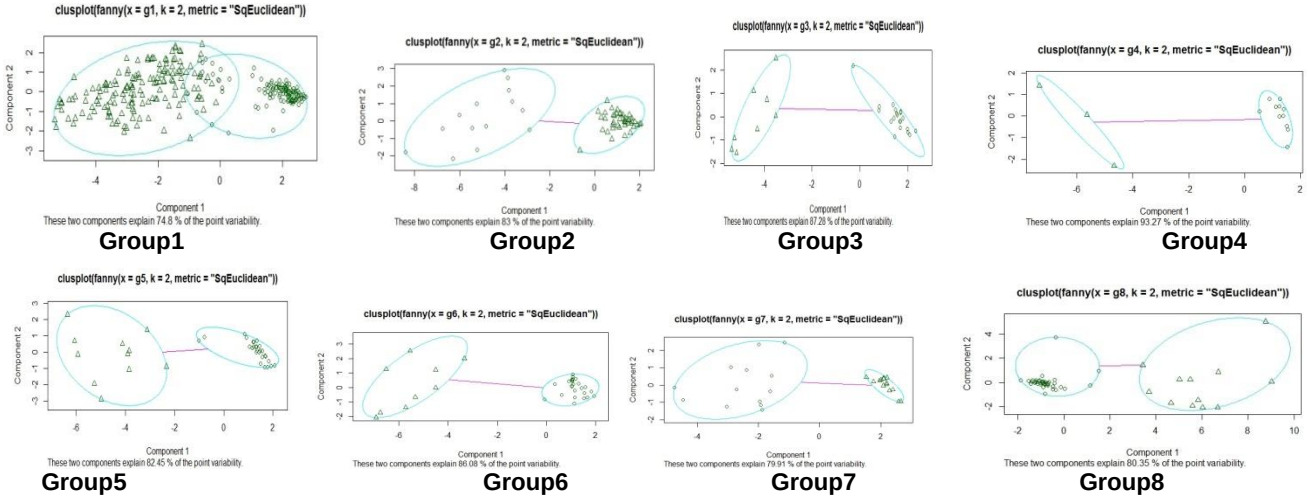

Fig: 3. The outputs of fuzzy c-means cluster points.

Results of dynamic fuzzy clustering algorithm

We have executed initial clustering on group1 data with 316 instances and incrementally we added remaining group of data in the next cycles. The results are given **Table– 3** and clustering points

Table: 3. Results of dynamic fuzzy c-means clustering algorithm

	Cycle-1	Cycle-2	Cycle-3	Cycle-4	Cycle-5	Cycle-6	Cycle-7	Cycle-8
Data size	367	437	468	485	533	582	613	697
Clus Avg Sil width	0.69340811	0.7204372	0.72818879	0.73489439	0.7443762	0.7564407	0.7547007	0.7687378
Right c value	2	2	2	2	2	2	2	2
Dunn_coeff	0.8041438	0.8187395	0.8222193	0.825793	0.8315249	0.838496	0.8375065	0.8455845
Belgian	207	263	286	301	338	378	396	469
Malignant	160	174	182	184	195	204	217	228

THE IIOAB JOURNAL
ISSN: 0976-3104

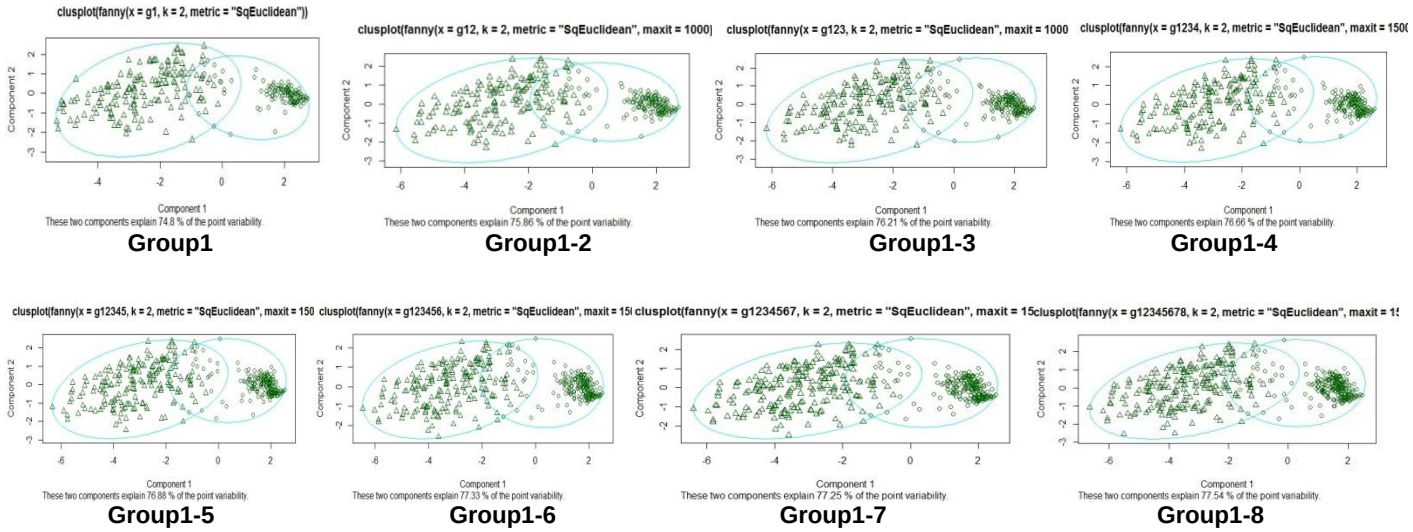

Fig: 4. For each cycle of dynamic fuzzy c-means cluster points.

Table: 4. Dynamic fuzzy c-means cluster centers of eight groups of cancer instances

	Cluster centers	Clump thickness	Uniformity cell size	Uniformity cell shape	Marginal adhesion	Single Epithelial Cell Size	Bare nuclei	Bland chromatin	Normal nucleoli	Mitoses
Group1	Cluster1	2.160703	2.838925	1.428393	1.545856	1.368896	2.234148	1.551025	2.635896	1.456334
	Cluster 2	3.898317	7.367231	6.570284	6.643336	5.579643	5.680435	8.065507	5.630912	6.320481
Group 2	Cluster1	2.129815	2.810183	1.459899	1.582505	1.373803	2.219787	1.473415	2.4792	1.425762
	Cluster 2	3.892513	7.433734	6.707197	6.747639	5.537787	5.68313	8.089441	5.734023	6.247491
Group 3	Cluster1	3.897594	7.450822	6.688592	6.727644	5.597415	5.651275	8.105787	5.878557	6.303923
	Cluster 2	2.12504	2.890006	1.44051	1.563888	1.413542	2.205181	1.483058	2.374672	1.399523
Group 4	Cluster1	3.899821	7.438435	6.727208	6.764603	5.658186	5.668592	8.114747	5.939162	6.355033
	Cluster 2	2.120565	2.96046	1.431706	1.553691	1.411136	2.187144	1.464081	2.318261	1.382631
Group 5	Cluster1	2.114216	2.987723	1.408163	1.513945	1.40153	2.148456	1.443685	2.249238	1.344811
	Cluster 2	3.902375	7.367682	6.754794	6.778681	5.773704	5.585881	8.15409	5.975062	6.275716
Group 6	Cluster1	2.104058	3.040041	1.3894	1.492056	1.376849	2.127336	1.41911	2.247866	1.3284
	Cluster 2	3.906056	7.353648	6.818226	6.82056	5.813291	5.557155	8.14977	6.058938	6.309007
Group 7	Cluster1	2.104699	3.08218	1.387464	1.500128	1.386206	2.125459	1.414027	2.229006	1.322869
	Cluster 2	3.911752	7.302867	6.859551	6.859199	5.917126	5.529439	8.137364	6.16135	6.248692
Group 8	Cluster1	2.091352	3.049704	1.365231	1.486661	1.37489	2.12169	1.375652	2.142978	1.290061
	Cluster 2	3.912348	7.235365	6.968129	6.930414	5.972788	5.531019	8.01619	6.215421	6.08946

DISCUSSIONS

In The Wisconsin breast cancer database consists of eight groups of cancer instances with one class. These instances are already classified into benign or malignant. In our experiment, initially we have done fuzzy c-means on individual groups. For that, we removed the class filed and we applied fuzzy c-means clustering on nine features. From the results it shows that all groups of instance are clustered into only two groups and it was similar to classification according database. We identified that in group1 instances having overlapping, i.e. some instances partially belongs both groups. But in next groups does not have overlapping instances and they almost belong to one group. The cancer instances in overlapping portion between classes, it is difficult to find state of

THE IIOAB JOURNAL
ISSN: 0976-3104

patient. In next step, we executed dynamic fuzzy c-means where we added gradually groups of instances and every time we checked the number of clusters. In all cycles, instances are grouped into two and overlapping between them. We identified the changes in class centers and silhouette widths. With these results, one can identify the similarity between the breast cancer patients within the same group and other group patients.

CONCLUSION

We consider the problem of tracking of changes in cancer data using soft clustering algorithm. The clustering can be performed using hard computing or soft computing approaches. Hard clustering is not efficient to handle impression, uncertainty, partial truth and approximation data set. Soft computing approaches successful in handling this type of data and it supports many complex applications. We apply the fuzzy c-means clustering or soft clustering on Wisconsin breast cancer data on individual groups, after that for each cycle incrementally added group cases and apply dynamic fuzzy c-means clustering algorithm. We projected results in each case which can be useful for finding the comparison between group instances. This method can be implemented using any other soft clustering methods to improve the results.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Fayyad Usama, Piatetsky-Shapiro, Gregory Smyth, Padhraic. [1996] From Data Mining to Knowledge Discovery in Databases.

[2] Data Mining Curriculum ACM SIGKDD. 2006-04–30.

[3] Ferreira C. [2006] Designing Neural Networks Using Gene Expression Programming, Applied Soft Computing Technologies: *The Challenger of Complexity*, 517–536, *Springer-Verlag*.

[4] AE Eiben and M Schoenauer. [2002*] Evolutionary computing, Information Processing Letters*, 82(1):1–6,

[5] Zadeh Lotfi A.[1994] Fuzzy Logic, Neural Network, and Soft Computing, Communication of the ACM, March 1994, 37 (3): 77–84.

[6] Yudong Zhang, Saeed Balochian, Vishal Bhatnagar. [2014] "Emerging Trends in Soft Computing Models in Bioinformatics and Biomedicine". *The Scientific World Journal* 3.doi:10.1155/2014/683029.

[7] Fuzzy Logic. *Stanford Encyclopedia of Philosophy*. 2006-07–23.

[8] Zedeh LA. [1965] Fuzzy Set . *Information and Control* 8 (3): 338–353.

[9] J Bezdek, S Pal. [1992] Fuzzy models for pattern recognition, IEEE press, New York ,

[10] H- J Zimmermam, Fuzzy set theory and it applications, [1991].

[11] Nock R, Nielsen F. [2006] On Weighting Clustering, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 28 (8): 1–13.

[12] Peter J Rousseeuw. [1987] "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*20: 53–65.

[13] Chatti Subbalakshmi. [2015] A method to find optimal number of clusters based on fuzzy silhouette on dynamic dataset. *Elsevier Science Direct*, 2015,doi: 10.1016 /j.procs.2015.02.030.