TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 15, No. 2, August 2015, pp. 373 ~ 380
DOI: 10.11591/telkomnika.v15i2.8183

Two Level Clustering for Quality Improvement using Fuzzy Subtractive Clustering and Self-Organizing Map

Erick Alfons Lisangan[1], Aina Musdholifah[2], Sri Hartati[3]
[1]Universitas Atma Jaya Makassar, Makassar, Indonesia
[2,3]Universitas Gadjah Mada, Yogyakarta, Indonesia
E-mail: erick_lisangan@lecturer.uajm.ac.id

Abstract

Recently, clustering algorithms combined conventional methods and artificial intelligence. FSC-SOM is designed to handle the problem of SOM, such as defining the number of clusters and initial value of neuron weights. FSC find the number of clusters and the cluster centers which become the parameter of SOM. FSC-SOM is expected to improve the quality of FSC since the determination of the cluster centers are processed twice i.e. searching for data with high density at FSC then updating the cluster centers at SOM. FSC-SOM was tested using 10 datasets that is measured with F-Measure, entropy, Silhouette Index, and Dunn Index. The result showed that FSC-SOM can improve the cluster center of FSC with SOM in order to obtain the better quality of clustering results. The clustering result of FSC-SOM is better than or equal to the clustering result of FSC that proven by the value of external and internal validity measurement.

Keywords*: clustering, fuzzy substractive clustering, self-organizing map*

1. Introduction

Clustering is one of the most important research issues in the domain of data mining and very useful for many applications, such as marketing, industrial engineering, biology, medicine, and image processing [1]. Clustering divides data into a homogeneous groups called clusters. Each cluster consists of data that have a greater similarity between the other data in their own cluster as compared with data in other cluster [2].

The efforts to make improvements cluster models, such as the optimal number of clusters and the best clustering results still continuing because the methods that has been developed is heuristic [3]. Recently, clustering algorithms combined conventional methods and artificial intelligence, like neural network, genetic algorithm, fuzzy set theory, and evolutionary programming. Combining two clustering methods, sometimes called two level clustering, have been certified to be more powerful than the individual methods. Two level clustering is proposed to improve partitional method, e.g. k-Means or Fuzzy C-Means, that sensitive to the initial cluster center and difficult to determine the number of clusters [4].

Self-Organizing Map (SOM) is clustering algorithm that apply the concept of neural network and can be used for data visualization [5]. Generally, clustering algorithms tries to group data by maximize the inter-cluster and minimize the intra-cluster [6]. SOM perform to group data with a different characteristic that is maintaining the relationship of neighborhood in data [7]. The advantage of SOM is resistance to the data noise [8]. But the disadvantage of SOM is the structure of the neural network and the number of neurons in the Kohonen layer must be defined first [8]. SOM is implemented to produce protocluster in two level clustering [4, 7], [10-11]. Then, the second clustering algorithms group the protocluster at the second level. The research about using SOM at the second level is not found so far.

Fuzzy Subtractive Clustering (FSC) can solve the disadvantage of SOM by using data point as a candidate of the cluster center [12]. A data point with the highest density will be defined as a cluster center [13]. FSC is implemented to initialize the number of cluster and cluster center in two level clustering and combine with FCM, also called Hybrid Fuzzy Clustering [14], [15]. FCM can not ensure the unique clustering result because number of cluster must be defined first and the initial of cluster centers is selected [15].

Received May 22, 2015; Revised July 3, 2015; Accepted July 17, 2015

In this research, a new methos is proposed for two levels clustering by using FSC and SOM. FSC is used to find the number of clusters by searching data point with the highest density will be the cluster center. The result of FSC is the number of clusters and cluster centers then will be the initial weight of SOM. Then, SOM will ameliorate the cluster centers of FSC and is expected to improve the quality of clustering by FSC.

2. The Proposed Algorithm

Fuzzy Subtractive Clustering (FSC) is proposed by Stephen Chiu (1994) where finding the number of clusters based on density of each data point. The data point with the highest number of neigbors or highest density will be chosen as the cluster center and the density value of the cluster centers will be reduce so that can not be chosen again. The algorithm will find another data point with the highest number of neigbors or highest density to be another cluster center [16].

Self-Organizing Map (SOM) is proposed by Teuvo Kohonen (1982) and widely used as a method to reduce the dimension of data and clustering [17]. SOM is a type of neural network that is trained using unsupervised learning to produce a representation of data into a map, such as 1D [18]. In this research, we used 1D map in feature map or output layer of SOM. The numbers of neuron in input layer have the same amount with the number of attribute (*j*) of dataset. Similarly, the numbers of neuron in output layer have the same amount with the number of cluster (*k*) that result best quality from each dataset.

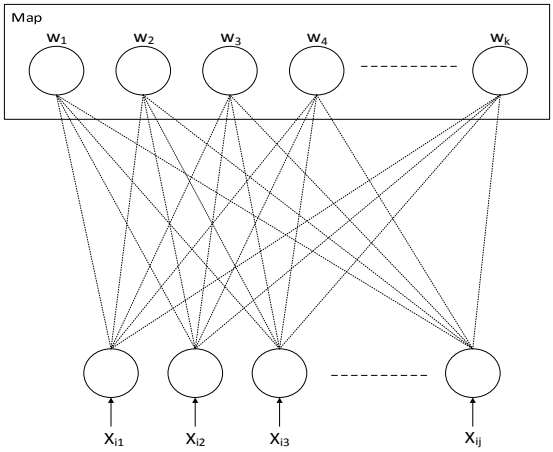

Figure 1. SOM Architecture

FSC-SOM is proposed to solve the disadvantage of SOM that need some parameter, i.e. the number of neuron in output layer and the initial weights of neurons that determined randomly. Futhermore, FSC-SOM is expected to improve the quality of clustering result from FSC. At the first level, FSC is implemented to estimate the number of clusters and find the cluster centers that become the parameter of SOM. At the second level, SOM will ameliorate the cluster centers of FSC with the purpose to improve the quality of clustering by FSC.

We can define two level clustering algorithm of FSC-SOM into five main process, as follows:

1. Initialitation.
   a. Dataset with data point X_{ij} where *i* is *i*-th data point of *n* data and *j* is *j*-th attribute of *m* attribute in dataset.
   b. Initialize the parameter, i.e. *r* (radius), *reject ratio, accept ratio*, *q* (squash factor)*, α* (*learning rate*), *maxEpoch* (maximum epoch), and *ε* (threshold).
2. Data normalitation using Min-Max Normalization.
3. Cluster Estimation
   a. Calculate the density value of each data point (D_i) using Formula (1).

$$D_i = \sum_{k=1}^{n} e^{-4(\sum_{j=1}^{m}\left[\frac{X_{ij}-X_{kj}}{r}\right]^2)} \quad (1)$$

b. Find the data point with the highest density value and set it become the candidate cluster center.
c. Calculate ratio of canditate cluster center (*R*) by divide it with density value of first canditate cluster center.
d. Checking the eligibility of the candidate cluster center with this following conditions:
   i. If *R* > *accept ratio* then the candidate cluster center can be accepted as cluster center, otherwise check the second condition,
   ii. If *R* > *reject ratio* then calculate the sum of the ratio and distance between the candidate cluster center and predefined cluster centers, otherwise cluster estimation process is stopped because there is no data point can be the candidate cluster center (step 4).

   If the sum is greater than or equal to 1 then the candidate cluster center can be accepted as cluster center, otherwise the data point cannot be accepted as cluster center and set the density value of it become 0.
e. If the canditate cluster center can be accepted become the new cluster center then increament the number of cluster (*k*) and reduce the density value of each data point around the new cluster center (*c*) using Formula (2) then back to step 3b.

$$D_i^{'} = D_i - D_c * exp\left(-\frac{\|X_i - X_c\|}{\left(\frac{q*r}{2}\right)^2}\right) \quad (2)$$

4. Usage FSC
   a. After the process of estimation cluster is completed, then calculate membership function of each cluster for each data point using Formula (3).

$$\mu_{ki} = e^{-\sum_{j=1}^{m}\frac{\left(x_{ij}-c_{kj}\right)^2}{2\sigma_j^2}} \quad (3)$$

Where the sigma value of attribut *j* (δ_j) can be calculate using Formula (4), $XMin_j$ and $XMax_j$ is the minimum and maximum value for *j*-th attribute.

$$\sigma_j = \frac{r*(XMax_j - XMin_j)}{\sqrt{8}} \quad (4)$$

   b. For each data point, find the highest membership function of each cluster. Cluster with the highest membership function for each data point indicate that the data point get in that cluster. After that, calculate the quality of clustering result using F-Measure (F_{fsc}) using Formula (12).
5. Learning
   a. Calculate the distance value between each neuron weight (*w*) and each data point X_i using Formula (5).

$$D_{ik} = \sum_{j=1}^{m}(w_{kj} - X_{ij})^2 \quad (5)$$

   b. Find winner neuron that is nearest neuron from *i*-th data point.
   c. Update the weight of winner neuron and neurons around the winner neuron based on the neighborhood value in *t*-th epoch (*d(t)*) using Formula (6).

$$w_{kj} = w_{kj} + \alpha * \left(X_{ij} - w_{kj}\right) \quad (6)$$

Repeat step 5a if there is data point that have not been calculate the distance with each *neuron*, otherwise go to step 5d.

d. Modify the value of learning rate (*α*) and neighborhood value (*d*) using Formula (7) and (8) then incremant the value of *epoch*.

$$\alpha(t) = \alpha_o * \left(1 - \frac{t}{T}\right) \quad (7)$$

$$d(t) = d_o * \left(1 - \frac{t}{T}\right) \quad (8)$$

e. Convergence condition
   i. Find the nearest neuron for each data point indicate that the data point get in that cluster. After that, calculate the quality of clustering result using F-Measure ($F_{fsc\text{-}som}$) using Formula (12).
   ii. Check convergence condition, if the difference between $F_{fsc\text{-}som}$ and F_{fsc} is more than ε or maximum epoch has been reached then FSC-SOM process is stopped, otherwise back to step 5a.

3. Research Method

3.1. Dataset

In this research, we use 10 dataset from UCI Machine Learning (URL: http://archive.ics.uci.edu/ml/) to test our proposed method with one level clustering, i.e. FSC and SOM. Table 1 show about testing dataset that we used in this research and detail about the dataset, i.e. the number of data point, attribute, and class they have. For dataset wine and glass, the real number of class is 7 but there is 1 class that does not have a member so we define that they have 6 class.

Table 1. Testing Dataset (UCI Machine Learning Repository)

Dataset	Data Point	Attribute	Class
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6*
WDBC	569	30	2
CMC	1473	9	3
Yeast	1484	8	10
Optical Digit	5620	64	10
Statlog	6435	36	6*
Thyroid	7200	21	3
Magic Gamma	19020	10	2

3.2. Cluster Evaluation

There are 3 approaches to study the validity of the clustering results, which is based on external criteria, internal, and relative [19]. The validity of external criteria is done by evaluating the clustering results with predefined structure in a dataset. The measuring instrument validity based on external criteria is F-measure and entropy. The validity of internal criteria is done by evaluating the clustering results with utilize vector dataset information. The measuring instrument validity of internal criteria is Silhouette index and Dunn index.

3.2.1. F-Measure

F-Measure is used to calculate the precision and recall between the clustering results with true class. F-Measure for each cluster *r* can be calcuted using Formula (9).

$$F(r,s) = \frac{2*n(r,s)}{n_r + n_s} \quad (9)$$

Where *n(r,s)* is the number of member that is in cluster *r* and *s*, n_r is the number of member that is in cluster *r*, and n_s is the number of member that is in cluster *s*.

The overall of F-Measure (F) from the clustering result can be calculated using Formula (10). The greater value of the F-measure then the better clustering results is obtained [20].

$$F = \sum_{r=1}^{k} \frac{n_r}{n} \max\{F(r,s)\} \quad (10)$$

3.2.2. Entropy

Entropy is used to measure how much the homogeneity of the cluster or distribution of cluster members in each cluster [21]. The lower value of entropy is more homogeneous clusters and the quality of clustering results is getting better.

$$E(S_r) = -\sum_{i=1}^{k} \frac{n_i^r}{n_r} log \frac{n_i^r}{n_r} \quad (11)$$

Where *k* is the number of cluster, n_i^r is the number of data from cluster *i* that get in cluster *r*. The overall of entropy (E) can be calculated using Formula (12).

$$E = \sum_{r=1}^{k} \frac{n_r}{n} E_r \quad (12)$$

3.2.3. Silhouette Index

Silhouette Index or Silhouette Coefficient is a normalize summation index [22] that combines both cohession and separation terms [6, 23].

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (13)$$

Where cohession (*a(i)*) is measured by calculating the average distance of all data point in a cluster and separation (*b(i)*) is measured by calculating the average distance of each data point in a cluster with its nearest cluster. *a(i)* and *b(i)* can be calculated using Formula (14) and (15).

$$a(i) = \frac{\sum d(i,j)}{n_{C_i}}, i,j \in C_i \quad (14)$$

$$b(i) = min_{C_k \neq C_i} \left\{ \frac{\sum d(i,k)}{n_{C_k}} \right\}, i \in C_i \; dan \; k \in C_k \quad (15)$$

Where *d(i,j)* is the distance between *i*-th and *j*-th data point, n_{Ci} and n_{Ck} is the number of data point in *i*-th and *k*-th cluster.

Silhouette width (*s(i)*) from each data point is used to calculate Silhoutte Index (S) using Formula (16) where *n* is the number of data point. The range of Silhouette Index is [-1, 1]. The greater its value then the better quality of clustering results is achieved.

$$S = \frac{1}{n} \sum s(i) \quad (16)$$

3.2.4. Dunn Index

Dunn Index (D) is proposed by Dunn [24] measure the ratio between the smallest intercluster distance with the largest intracluster distance. Dunn index is used to to identify clusters that are compact and well separated [6].

$$D = min_{i \in c} \left\{ min_{j \in c, i \neq j} \left\{ \frac{d(i,j)}{max_{k \in c}(d(k))} \right\} \right\} \quad (17)$$

Where *i, j,* and *k* is cluster from the clustering result, *d(i,j)* is the intercluster distance between cluster *i* and *j*, *d(k)* is intracluster distance from cluster *k*. The larger value of Dunn Index showed the better clustering results are obtained [19].

4. Results and Analysis

FSC need 4 parameter, i.e. radius (*r*), reject ratio, accept ratio, and squash factor (*q*). Choosing the value of accept ratio and reject ration can affect the clustering result [16]. If accept ratio is too large then too little data point can be accepted as cluster center. Whereas if reject ratio is too small then too much cluster centers can be resulted. The recommended value of each parameter, i.e. accept ratio=0.5, reject ratio=0.15, and q=1.5 [16].

The value of *r* is different for each dataset because the resolution of each dataset is different each other. In this experimental result, the optimal value of *r* is the value of *r* that can produce the highest value of F-Measure and produce the number of cluster about 2 clusters from the real number of cluster for each dataset. Table 2 show the optimal value of *r* for each dataset and the number of cluster that can be produced.

Table 2. The Optimal Value of *r*

Dataset	r	True Class	Predefined Class
Iris	0.45	3	3
Wine	0.9	3	3
Glass	0.145	6	8
WDBC	0.5	2	2
CMC	1.1	3	2
Yeast	0.16	10	10
Optical Digit	2.2	10	10
Statlog	0.65	6	7
Thyroid	0.5	3	4
Magic Gamma	0.7	2	2

The value of learning rate (α) and maximum epoch (maxEpoch) is α=0.4 and *maxEpoch*=50 that is the best combination in [25]. The threshold value (ε) is 0.7 for FSC-SOM because there is convergence condition that compare the difference between $F_{fsc\text{-}som}$ and F_{fsc} in learning process at the second level.

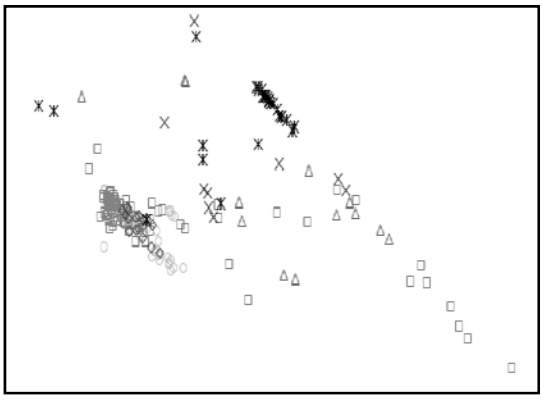

Figure 2. Visualization of glass dataset

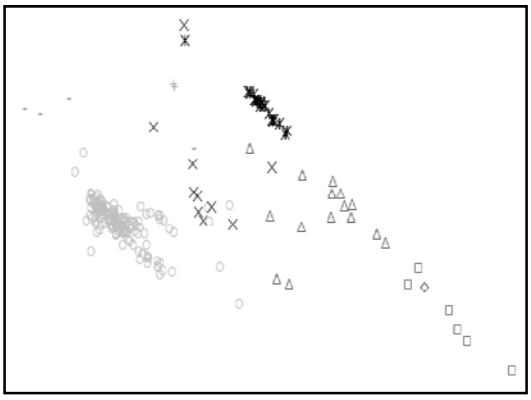

Figure 3. Visualization of FSC-SOM result for glass dataset

The performance of FSC-SOM can be seen in Table 3 where the meaning of checkmark is the quality of clustering result by FSC-SOM greater than or equal to the clustering result of another algorithms, i.e. FSC and SOM. There is 4 cluster validity measurements to compare the proposed algorithm with another algorithm, i.e. F-Measure, Entropy, Silhouette Index, and Dunn Index. The result show that the quality of clustering result by FSC-SOM at least equal to the quality of clustering result by FSC for all dataset and all cluster validity either external or internal validity.

Whereas, the quality of clustering result by FSC-SOM is greater than or equal to the quality of clustering result by SOM for some dataset and different cluster validity. The quality of

clustering result by FSC-SOM based on the precision and recall of true class using F-Measure is greater than or equal to SOM in 7 datasets. The quality of clustering result by FSC-SOM based on homogeneity of the cluster using entropy is greater than or equal to SOM in 9 datasets. The quality of clustering result by FSC-SOM based on the ratio between the average distance of intracluster distance and the average distance of intercluster using Silhouette Index is greater than or equal to SOM in 8 datasets. The quality of clustering result by FSC-SOM based on the ratio between the smallest distances of intercluster with largest distance of intracluster using Dunn Index is greater than or equal to SOM in 8 datasets.

Table 3. The Performance of FSC-SOM

Dataset	F-Measure		Entropy		Silhouette		Dunn	
	FSC	SOM	FSC	SOM	FSC	SOM	FSC	SOM
Iris	√	√	√	√	√	√	√	√
Wine	√	√	√	√	√	√	√	√
.Glass	√		√	√	√	√	√	√
WDBC	√	√	√	√	√	√	√	√
CMC	√	√	√		√		√	√
Yeast	√		√	√	√	√	√	√
Optical Digit	√	√	√	√	√	√	√	√
Statlog	√		√	√	√	√	√	
Thyroid	√	√	√	√	√	√	√	
Magic Gamma	√	√	√	√	√		√	√

5. Conclusion

FSC can handle the problem of SOM through defining the parameter of SOM, i.e. the number of cluster and inital value of neuron's weight. SOM also can ameliorate the cluster centers that are defined by FSC so the better quality of clustering can be achieved. The clustering result of FSC-SOM is better than or equal to the clustering result of FSC that proven by the value of external and internal validity measurement. Futhermore, the clustering result of FSC-SOM is better than the clustering result of SOM for some datasets.

Future work will be involved with using another method to update the value of learning rate and neighborhood in SOM, e.g. Gaussian or Heuristic and using another method to get the best combination of SOM's parameter, i.e. value of learning rate, maximum epoch, and threshold.

References

[1] Yang C, Chi S. *An Ant-Based Self-Organizing Feature Maps Algorithm*. 5th Workshop On Self-Organizing Maps. Paris. 2005.
[2] Gu L, Lu X. *Semi-supervised Subtractive Clustering by Seeding*. 9th International Conference on Fuzzy Systems and Knowledge Discovery. Sichuan. 2012; 1: 738-741.
[3] Santosa B. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu. 2007.
[4] Chi S, Yang C. A Two-stage Clustering Method Combining Ant Colony SOM and K-means. *Journal of Information Science and Engineering.* 2008; 24(1): 1445-1460.
[5] Luo B, Tang X. *Using Self-Organizing Map for Ideas Clustering of Group Argumentation*. The 11th International Symposium on Knowledge and Systems Sciences. Xi'an. 2010; 1: 1-6.
[6] Mushdolifah A, Hashim SZM. *Triangular Kernel Nearest Neighbor Based Clustering for Pattern Extraction in Spatio-Temporal Database*. Intelligent Systems Design and Applications (ISDA), Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference. Cairo. 2010; 1: 67-73.
[7] Sarlin P, Eklund T. *Fuzzy Clustering of the Self-Organizing Map: Some Applications on Financial Time Series*. Advances in Self-Organizing Maps - 8th International Workshop, WSOM 2011. Espoo. 2011; 1: 40-50.
[8] Silva B, Marques N. *Feature Clustering with Self-Organizing Maps and An Application to Financial Time-Series for Portfolio Selection*. Proceedings of the International Conference on Fuzzy Computation and International Conference on Neural Computation. Valencia. 2010; 1: 301-309.
[9] Mokris I, Forgac R. *Decreasing the Feature Space Dimension by Kohonen Self-Organizing Maps*. 2nd Slovakian – Hungarian Joint Symposium on Applied Machine Intelligence. Budapest. 2004.

[10] Tarek KM, Farouk B. *Kohonen Maps Combined to Fuzzy C-means, a Two Level Clustering Approach. Application to Electricity Load Data.* Self Organizing Maps - Applications and Novel Algorithm Design. 2011; 1: 541-558.
[11] Souza JR, Ludermir TB, Almeida LM. *A Two Stage Clustering Method Combining Self-Organizing Maps and Ant K-Means.* Artificial Neural Networks – ICANN 2009. Limassol. 2009; 5768: 485-494.
[12] Abdullah AG, Feranie S. Development of Short Term Load Forecasting Based On Fuzzy Subtractive Clustering. 2014. https://www.researchgate.net/publication/228933118_DEVELOPMENT_OF_SHORT_TERM_LOAD_FORECASTING_BASED_ON_FUZZY_SUBTRACTIVE_CLUSTERING.
[13] Sastria G, Liong C, Hashim I. *Application of Fuzzy Subtractive Clustering for Enzymes Classification*. Applied Computing Conference (ACC '08). Istanbul. 2008; 1: 304-309.
[14] Han L, Chen G. *HFCT: A Hybrid Fuzzy Clustering Method for Collaborative Tagging*. 2007 International Conference on Convergence Information Technology. Gyeongju. 2007; 1: 1389-1394.
[15] Yang Q, Zhang D, Tian F. *An Initialization Method for Fuzzy C-Means Algorithm Using Subtractive Clustering*. 2010 Third International Conference on Intelligent Networks and Intelligent Systems. Shenyang. 2010; 1: 393-396.
[16] Chiu SL. Fuzzy Model Identification Based On Cluster Estimation. *Journal of Intelligent and Fuzzy Systems*. 1994; 2(3): 267-278.
[17] Camastra F, Vinciarelli A. Machine Learning for Audio, Image and Video Analysis. London: Springer-Verlag. 2007.
[18] Rojas R. Neural Networks: A Systematic Introduction. Berlin: Springer-Verlag. 1996.
[19] Rendón E, Abundez I, Arizmendi A, Quiroz EM. Internal versus External cluster validation indexes. *International Journal of Computers and Communications*. 2011; 5(1): 27-34.
[20] Chen Y, Qin B, Liu T, Li S. The Comparison of SOM and K-means for Text Clustering. *Computer and Information Science*. 2010; 3(2): 268-274.
[21] Zhao Y, Karypis G. Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Machine Learning*. 2004; 55: 311–331.
[22] Arbelaitz, O, Gurrutxaga I, Muguerza J, Perez JM, Perona I. An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recognition*. 2013; 46: 243-256.
[23] Rousseeuw JP. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*. 1987; 20: 53-65.
[24] Dunn J. Well Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*. 1974; 4(1): 95-104.
[25] Chaudary V, Bhatia RS, Ahlawat AK. A Constant Learning Rate Self-Organizing Map (CLRSOM) Learning Algorithm. *Journal of Information Science and Engineering.* 2015; 31(1): 387-397.