Volume 5, Issue 2, February 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Performance Analysis of Self-Organizing Neural Network-Based Clustering

Kadijath Thahira, Jovita Vani Sequeira, Sameema, Zahid Ansari

Dept. of Computer Science &Engineering

P.A. College of Engineering,

Mangalore, India

Abstract*— Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention. Data mining is the method of analyzing the large amounts of data stored in data warehouses. We can perform data analysis, classification, clustering etc. of huge data by using different algorithms. It is important to evaluate the performance of various clustering techniques because the application of different clustering techniques generally results in different sets of cluster formation. The performance can be evaluated in terms of accuracy and validity of the clusters, and also the time required to generate them, using appropriate performance measures. In this paper, we have analysed the performance of Self-Organizing neural network based clustering and k-Means clustering using Matrix Laboratory tool, MATLAB. These techniques are tested against the various datasets. Finally, their performance results are compared and presented.

***Keywords*— Clustering, K-Means Clustering, Neural Network Based Clustering, Self-Organizing Map (SOM), Sum of Square Errors (SSE), Xie-Beni Index, Dunn's Index, Silhouette Index**

I. INTRODUCTION

Data mining is generally the process of examining data from different aspects and summarizing it into valuable information. There are number of data mining software's for analysing the data. They allow users to examine the data from various angles, categorize it, and summarize the relationships identified.

Cluster analysis is [1] one of the major data analysis methods which are widely used for many practical applications in emerging areas like Bioinformatics ([2], [3]). Cluster analysis is the grouping of a set of patterns into clusters based on similarity. Clustering is a distribution of data into groups of related objects. Each group, called cluster, consists of objects that are related to each other and not related to the objects in other groups. If data is represented by fewer clusters, it accomplishes simplification, but it loses certain fine details. Data is modelled by its clusters [4].

The K-means algorithm ([1], [5]-[8]) is useful in producing clusters for many practical applications. But for the large datasets, the computational complexity of the K-Means algorithm is very high. Furthermore, because of random choice of initial centroids, this algorithm results in various kinds of clusters.

The Neural Networks is widely applied in the data mining. The neural networks have complex structure, long training time, and difficult to understand, but it is most applicable for noisy data.

An artificial neural network (ANN), often called as "neural network" (NN), is a computational model based on biological neural networks. In other words, it is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation [9]. Self-Organizing Map (SOM), introduced by Teuvo Kohonen, is one of the best known unsupervised natural learning algorithms [10].

In this paper, we have analysed the performance of neural network based SOM clustering and K-Means clustering by considering the SSE, Xie-Beni Index, Dunn's Index, and Sihouette Index.

The remainder of the paper is organized as follows. Section II presents literature review. Section III presents overview of K-Means clustering and Self-Organizing Map. Section IV describes the various clustering validity measures. Experimental results and discussion are described in Section V. In section VI conclusion about our analysis work is presented.

II. LITERATUREGE REVIEW

Obula Reddy, Dr.Maligela Ussenaiah [11] described that Cluster analysis is unsupervised learning method that constitutes a cornerstone of an intelligent data analysis process. It is useful for the exploration of inter-relationships among a collection of patterns, by organizing into homogeneous clusters. It is called as unsupervised learning because no priori labelling of some patterns is available to use in categorizing others and inferring the cluster structure of the whole data. Intra-connectivity is a measure of the density. A high intra-connectivity means a good clustering arrangement, because the instances grouped within the same cluster are highly dependent on each other. Inter-connectivity is a measure of the connectivity between distinct clusters. A low degree of interconnectivity is advantageous because it indicates that individual clusters are largely independent of each other.

According to them, cluster analysis is a difficult problem because of many factors 1. Effective similarity measures, 2.Criterion functions, 3. Algorithms come into play in devising a well-tuned clustering technique for a given clustering problem. It is well known that no clustering method can adequately handle all sorts of cluster structures i.e. shape, size, and density. Sometimes, the quality of the clusters that are found can be improved by preprocessing the given data.

Martin Hynar, Michal Burda, and Jana Sarmanov [12] described the comparisons between K-Means method and SOM neural network and between the class method and gng neural network and they said that utilizing neural networks (with competitive learning) is a good idea in the clustering domain. The most important feature of such neural networks is their natural ability to find dense areas in the input space. The extension of the basic SOM algorithm to dynamically reflect relations in the data (possibility of the net growth) makes neural networks even much more interesting.

Amanpreet KaurToor, Amarpreet Singh [13] described a comparative study where they found that Kohonen SOM gives the better performance as compared to K-Means with minimum error rate or high accuracy, minimum computation time on same data set and parameters.

Usually, Self Organizing Feature Map (SOFM) can be initialized using random values for the weight vectors. Abdel-Badeeh M. and others [14] presented a different approach for initializing SOFM. This approach will use K-Means algorithm as an initialization step. This method is an efficient method for forming a topologically ordered feature map and they found that the proposed technique speeds up the conventional SOFM training algorithm.

A comparison among some non-hierarchical and hierarchical clustering algorithms including SOM (Self-Organization Map)neural network and Fuzzy c-means method is shown by Sueli A. Mingoti, Joab O. Lima[15]. The Data taken for this analysis is simulated considering correlated and uncorrelated variables, non-overlapping and overlapping clusters with and without outliers. The results showed that Fuzzy c-means had a good performance but SOM did not perform well. The traditional hierarchical clustering and K-Means methods presented similar performance but better than the SOM performance.

III. METHODOLOGY

In this paper, we compared the performance of K-Means and Neural network based clustering i.e., SOM.

A. K-Means Clustering Algorithm

The K-Means algorithm is undoubtedly the most widely used partition clustering algorithm [16]. Its popularity can be attributed to several reasons. First, it is conceptually simple and easy to implement. Virtually all data mining software's include an implementation of it. Second, it is versatile, i.e., almost every aspect of the algorithm (initialization, distance function, termination criterion, etc.) can be modified. Third, it has linear time complexity.

On the other hand, K-Means has several significant disadvantages. First, it requires the number of clusters, K, to be specified priori. Second, it can only detect compact, hyper spherical clusters that are well separated. Third, due to its utilization of the squared Euclidean distance, it is sensitive to noise and outlier points since even a few such points can significantly influence the means of their respective clusters.

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low [1]. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

The K-Means algorithm works as follows: First, it randomly selects K of the objects, each of which initially represents a cluster mean or center. For each of the remaining object, the object is assigned to the cluster to which it is the most similar, based on distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges [1].

Algorithm: K-Means clustering algorithm for partitioning, where each cluster's center is represented by the mean value of the data points in that cluster.
Input :K : the number of clusters.

D: a data set containing n objects.

Output: A set of K clusters.
Method:
(1) Arbitrarily choose K objects from D as the initial cluster centers.
(2) **Repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **Until** all or many of the objects are most similar to the newly assigned cluster based on their mean value;

Fig. 1 K-Means clustering algorithm

B. Self Organizing Map (SOM)

Kohonen's self-organizing map (SOM), is a competitive unsupervised learning neural network that uses a neighbourhood lateral interaction function to discover the topological structure hidden in the data set. It is an

unsupervised learning which has both visualization and clustering properties. SOM produces a similarity graph of input data, which is much similar to the self-organizing characteristics of human brain. It consists of a finite set of models that approximate the open set of input data. The basic idea of SOM is to map the data patterns onto a n-dimensional grid of neurons or units [17]. The grid forms what is known as the output space, as opposed to the input space where the data patterns are. This mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space, and vice-versa [17]. So as to allow an easy visualization, the output space is usually 1 or 2 dimensional grid of neurons or units. Algorithm [17] is given as follows:

Algorithm: SOM based clustering algorithm for partitioning, where each cluster's center is represented by the mean value of the data points in that cluster.
Input: Training data: vectors, X, a Vectors of length *n:* ($x_{p,1}$, ...,$x_{p,i}$, ... , ... , $x_{p,n}$)
Outputs: A vector, Y, of length *m*: (y_1 , y_2 , ... , y_i , ... , ... , y_m)
Method:
(1) Let X be the set of n training patterns x_1, x_2,..x_n
W be a p×q grid of units w_{ij} where i and j are their coordinates on thatgrid
α be the learning rate, assuming values in [0,1], initialized to a given initial learning rate
r be the radius of the neighbourhood function h(w_{ij}, w_{mn}, r), initialized to a given initial radius
(2) **Repeat**
for k=1 to n
for all w_{ij}, w_{ij}∈W, calculate d_{ij}= || x_k-w_{ij}||
(3) select the unit that minimizes d_{ij} as the winnerw_{winner}
(4) update each unit w_{ij}∈W:w_{ij}=w_{ij}+ αh (w_{winner}, w_{ij}, r) || x_k– w_{ij}||
(5) decrease the value of α and r
until α reaches 0

Fig. 2 SOM - based clustering algorithm

IV. VALIDITY INDICES

We executed Self Organization Map (SOM), i.e., neural network based clustering algorithm and K-Means clustering algorithm on the four datasets shown in Table I. For analyzing the patterns discovered, we computed the sum of square errors (SSE) for each cluster and also for set of clusters. We also calculated the time required to run K-Means and SOM algorithm in MATAB. In order to compare how well two algorithms can be accomplished on different data sets and also to find which cluster is best, we evaluated our results based on some of the cluster validity indices. They are, i)Xie-beni Index, ii)Dunn's Index, iii) Silhouette Index.

A. Xie-Beni Index

Xie-Beni Index:

Xie-Beni [18], XB, it is also called as the compactness and separation validity function. This index is used to identify the most appropriate number of clusters. The XB can be calculated using the following formula:

$$XB=\frac{\sum_1^N \sum_{k=1}^M u_{ik}^2 \|x_i-C_k\|^2}{N \min\{\|C_t-C_s\|^2\}} \quad t \neq s$$

Where
u_{ik} is the membership degree of the i^{th} data point belonging to the k^{th}cluster,
x_i is the i^{th} data point
C is the centroid.
If the value of XB is lesser, it indicates that the cluster is compact and well separated.

B. Dunn's Validity Index

Dunn's Index [19], (DI) is used to find the solid and well- separated clusters. It means that the inter-cluster distances should be more and intra cluster distances should be less. The Dunn's cluster can be calculated using the below formula:

$$DI=\left[\min_{i \in c}\left[\min_{j \in c, j \neq i}\left[\frac{d(C_i,C_j)}{\max\{\Delta C_l\}}\right]\right]\right]$$

Where

$d(C_i, C_j)$ is the distance between the clustersC_i and C_j

$d(C_i, C_j)$ =min$\{d(x_i,x_j), where\ x_i \in C_i, x_j \in C_j\}$

$\Delta C_l = max\{d(x_i,x_j), where\ x_i, x_j \in C_i\}$

Higher value of DI implies that the cluster formation is better for that particular cluster.

C. Silhouette Validity Index

Silhouette technique is used to compare the similarity between data points in the same cluster and the data points in the different cluster[20]. Silhouette index value ranges from -1 to 1. The Silhouette value for the i^{th} data point S_i, can be calculated as follows:

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

Where

a_i is the average distance from the i^{th} data point to other data point in the same cluster,

b_i is the minimum average distance between the i^{th} data point and the data points in the different cluster.

When is largest, i.e., closest to 1, it indicates that data points are well clustered. If is closest to zero, it means that data point lies at equidistant from both the cluster, so it can be assigned to other cluster. But if is closest to -1, then it indicates that the data point lies much closer to the other cluster. For the best clustering, the average silhouette value should be larger.

V. RESULTS AND DISCUSSIONS

Following sections A, B gives the description about datasets used in our analysis work and comparisons of SOM algorithm and K-means algorithm.

A. Dataset Description

We have considered four data sets (Table I) which are available in MATLAB for comparing the performance of SOM based clustering and k-Means clustering. The four datasets are fisher iris dataset, cancer dataset, glass data set and wine dataset.

TABLE I FOUR DATASETS

Dataset name	No. of samples	No.of attributes
iris_dataset	150	4
cancer_dataset	699	9
glass_dataset	214	9
wine_dataset	178	13

Iris dataset is a 4x150 matrix of four attributes of 1000 flowers. It has three class labels i.e. setosa, versicolor, virginica. The four attributes are Sepal length in cm, Sepal width in cm, Petal length in cm, Petal width in cm.

The cancer dataset is a 699x9 matrix which is categorized as either benign or malignant. The nine attributes are Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal Adhesion, Single epithelial cell size, Bare nuclei, Bland chromatin, Normal nucleoli, Mitoses.

The glass dataset is a 214x9 matrix which is categorized either as Window glass or Non-window glass. The nine attributes are Refractive index, Sodium, Magnesium, Aluminium, Silicon, Potassium, Calcium, Barium, and Iron.

The Wine dataset is a 178x13 matrix which is categorized based on three winery's. There are thirteen attributes, which are Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Non flavonoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline.

B. Comparison of K-Means and SOM Algorithms

In order to compare the K-Means and SOM algorithms, we considered four data sets, which we have described above in section III.C. To find out how accurate is the cluster formation, we computed the clustering error, SSE(Sum of Squared Error) for each of the dataset. We say that the performance of clustering is better if the value of the SSE is less. We also calculated the time required to run the algorithm for different number of clusters. We can see in Table IIthat,the sum of squared error (SSE) is less in case of SOM compared to K-means ain most of the cases. But the time required for running the SOM algorithm is more compared to the time required to run K-Means algorithm.

The sum of squared error for different datasets and different number of clusters for K-Means and SOM techniques is shown in Table II. From the table, we conclude that the SSE is less for SOM based clustering.

The K-Means and SOM algorithms result in various cluster formation for different number of clusters. So, to find out the number of clusters which results in better cluster formation, we analyzed our result based on the validity indices discussed in section IV. In Table III, various values of validity indices are shown for K-Means and SOM algorithms for different datasets and number of clusters. They are further described in below sections:

1) *Evaluation of the cluster formation based on Xie-Beni Validity Index:* Fig.3, Fig. 4, Fig. 5, &Fig. 6 shows the graph plots of Xie-Beni validity index as a function of number of clusters for the iris dataset, cancer dataset, glass dataset, and wine dataset respectively. XB is mainly used to find the appropriate number of clusters. From the below graphs, For example, we can see that the XB value is lesser for SOM based Clustering.

From Fig. 3, we can see that the Xie-Beni index value plotted for fisher iris dataset is lesser for SOM algorithm for number of clusters 10, 15, 20, 25.

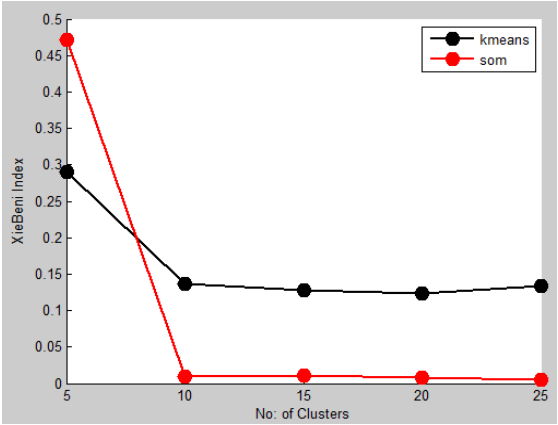

Fig. 3 Xie-Beni Index Vs No. of Clusters for fisher iris dataset

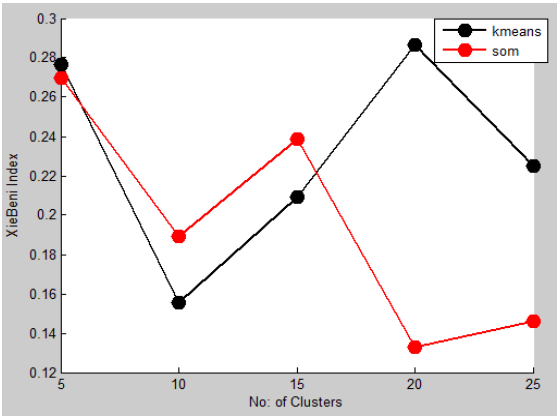

Fig. 4 Xie-Beni Index Vs No. of Clusters for cancer dataset

From Fig. 4, we can see that the Xie-Beni index value plotted for cancer dataset is lesser for SOM algorithm for number of clusters 5, 20, 25.

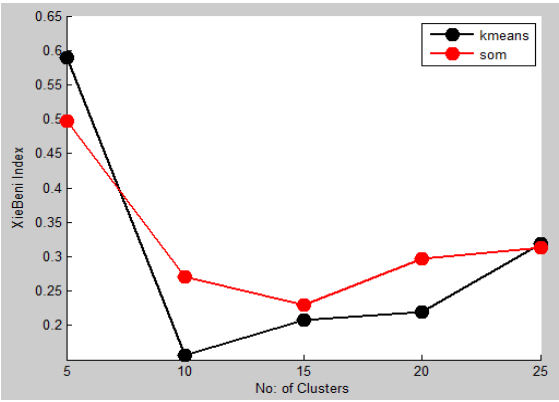

Fig. 5 Xie-Beni Index Vs No. of Clusters for glass dataset

From Fig. 5, Xie-Beni index value plotted for glass dataset is lesser for SOM algorithm for the clusters 5, 25.

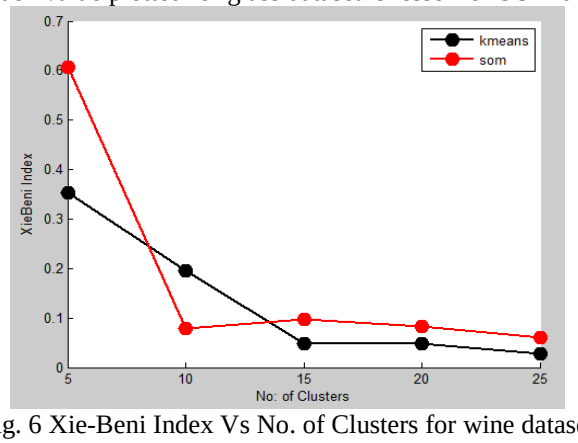

Fig. 6 Xie-Beni Index Vs No. of Clusters for wine dataset

From Fig. 6, we can see that the Xie-Beni index value plotted for wine dataset is lesser for SOM algorithm for the clusters 5, 25.

TABLE III COMPARISON OF K-MEANS AND SOM CLUSTERING METHODS

Dataset	Method	No.of. Clusters	Xie-Beni Index	Dunn's Index	Silhouette Index	SSE	Executi-on Time
iris_dataset	K-Means	5	0.2903	0.0823	0.6125	49.8623	0.0066
		10	0.1364	0.1206	0.4296	30.0195	0.0088
		15	0.1279	0.0990	0.4225	22.4564	0.0078
		20	0.1230	0.0990	0.4260	18.2212	0.0134
		25	0.1345	0.1491	0.4201	12.2908	0.0101
	SOM	5	0.4712	0.0823	0.4712	4.0823	1.8140
		10	0.0095	0.1089	0.6230	7.8883	2.4027
		15	0.0105	0.1539	0.4618	5.5983	1.9799
		20	0.0087	0.1658	0.4556	5.5952	1.8977
		25	0.0056	0.1658	0.4500	4.0000	2.2467
cancer_dataset	K-Means	5	0.2766	0.0512	0.3668	139.6495	0.0268
		10	0.1556	0.0634	0.2837	112.4151	0.0193
		15	0.2095	0.0636	0.2581	96.5658	0.0928
		20	0.2866	0.0645	0.2537	81.7629	0.0356
		25	0.2249	0.0750	0.1994	79.4644	0.0245
	SOM	5	0.2698	0.0512	0.3682	58.4963	4.1300
		10	0.1893	0.0602	0.2687	32.8623	4.9416
		15	0.2389	0.0619	0.2721	30.5535	4.9834
		20	0.1329	0.0655	0.2581	44.6437	5.1925
		25	0.1462	0.0698	0.1946	38.7979	6.3871
glass_dataset	K-Means	5	0.5888	0.0265	0.6018	400.2590	0.0172
		10	0.1564	0.0300	0.3324	242.5576	0.0105
		15	0.2080	0.0322	0.3961	187.7034	0.0092
		20	0.2198	0.0331	0.3969	131.1502	0.0330
		25	0.3186	0.0493	0.4216	111.0840	0.0171
	SOM	5	0.4971	0.0413	0.5940	15.6356	1.8307
		10	0.2703	0.0284	0.4441	13.7761	1.7503
		15	0.2295	0.0385	0.3795	9.2988	1.8017
		20	0.2963	0.0421	0.3937	7.5393	1.9157
		25	0.3128	0.0318	0.4176	16.2410	1.9935
wine_dataset	K-Means	5	0.3531	0.3531	0.0161	1.0099e+06	1.0098
		10	0.1947	0.1947	0.0359	2.4578e+05	0.2458
		15	0.0471	0.0471	0.0289	1.4837e+05	0.1484
		20	0.0485	0.0485	0.0432	1.0012e+05	0.1001
		25	0.0271	0.0271	0.0354	9.0701e+04	0.0907
	SOM	5	0.6069	0.0342	0.7136	18.7710	1.9602
		10	0.0787	0.0338	0.6977	7.3938	2.1986
		15	0.0966	0.0391	0.6734	2.0186	1.7939
		20	0.0823	0.0369	0.6551	11.4325	1.8392
		25	0.0608	0.0366	0.6405	5.8093	1.9532

2) *Evaluation of the cluster formation based on Dunn's Validity Index:* Fig.7, Fig. 8, Fig. 9,& Fig. 10 shows the graph plot of Dunn's validity index as a function of number of clusters for the iris dataset, cancer dataset, glass dataset, and wine dataset respectively. The main aim of Dunn's validity is to maximize the inter cluster distances. From the below graphs, we can see that the Dunn's index value is maximum for SOM based clustering.

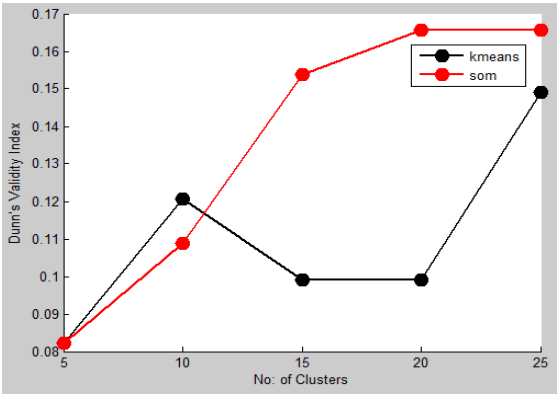

Fig. 7 Dunn's Index Vs No. of Clusters for iris dataset

From Fig. 7, we can see that the Dunn's index value plotted for iris dataset is larger for SOM algorithm for number of clusters 15, 20, 25.

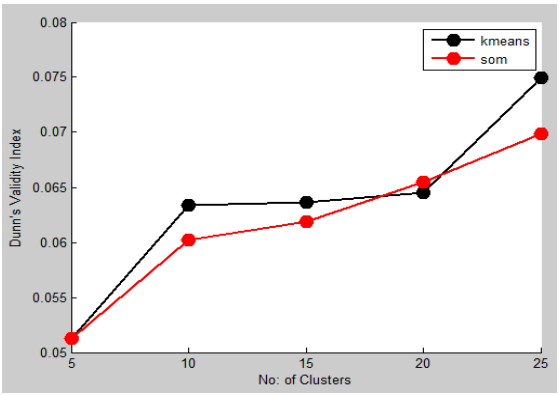

Fig. 8 Dunn's Index Vs No. of Clusters for cancer dataset

From Fig. 8, we can see that the Dunn's index value plotted for cancer dataset is larger for SOM algorithm for number of clusters 20.

From Fig. 9, we can see that the Dunn's index value plotted for glass dataset is larger for SOM algorithm for number of clusters 5, 15, 20.

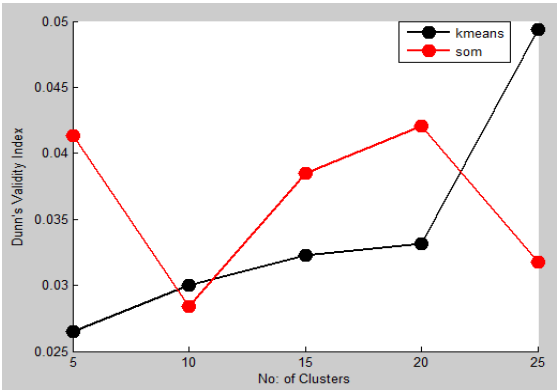

Fig. 9 Dunn's Index Vs No. of Clusters for glass dataset

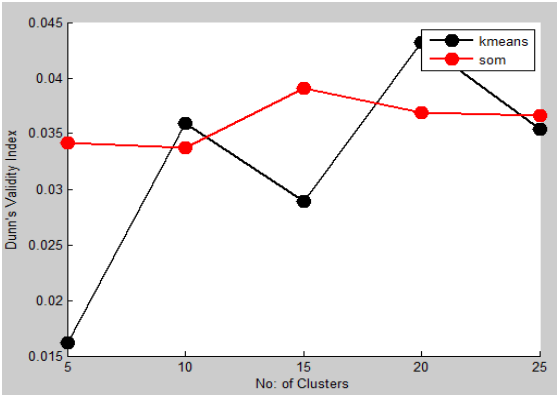

Fig. 10 Dunn's Index Vs No. of Clusters for wine dataset

From Fig. 10, we can see that the Dunn's index value plotted for wine dataset is larger for SOM algorithm for number of clusters 5, 15, 25.

3) *Evaluation of the cluster formation based on Silhouette Index:* Fig.11, Fig. 12, Fig. 13, &Fig. 14 shows the graph plot of Silhouette index as a function of number of clusters for the iris dataset, cancer dataset, glass dataset, and wine dataset respectively. Silhouette values lies between -1 and 1. If the value is closest to 1 means, the data point is assigned to the appropriate cluster. In our analysis we found that all the values between 0 and 1 for k-means and SOM based algorithm. From the below graphs we can see that the Silhouette index value is closest to 1 for SOM based clustering.

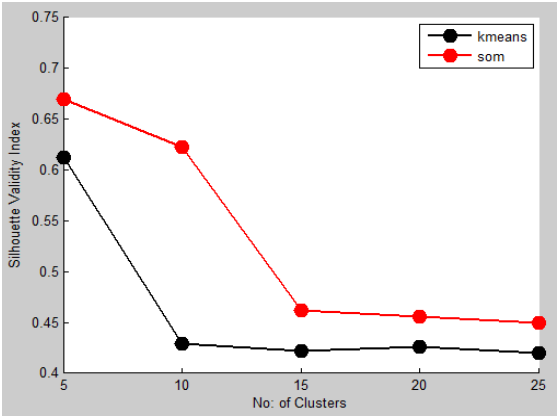

Fig. 11 Silhouette Index Vs No. of Clusters for iris dataset

From Fig. 11, we can see that the Silhouette index value plotted for iris dataset is closer to1 for SOM algorithm for the clusters 5, 10.

From Fig. 12, we can see that the Silhouette index value plotted for cancer dataset is closer to1 for SOM algorithm for number of clusters 5.

From Fig. 13, we can see that the Silhouette index value plotted for cancer dataset is closer to1 for SOM algorithm for number of clusters 5.

From Fig. 14, we can see that the Silhouette index value plotted for cancer dataset is closer to1 for SOM algorithm for the clusters 5, 10, 15.

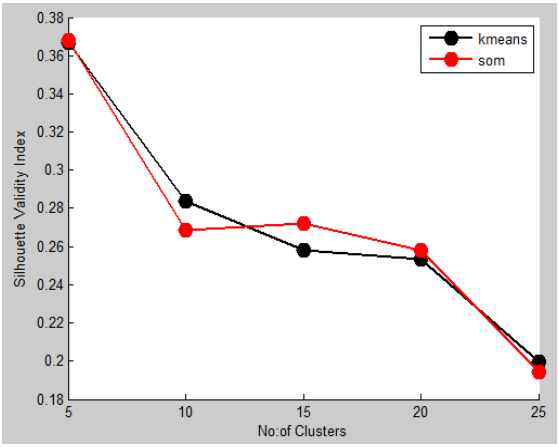

Fig. 12 Silhouette Index Vs No. of Clusters for cancer dataset

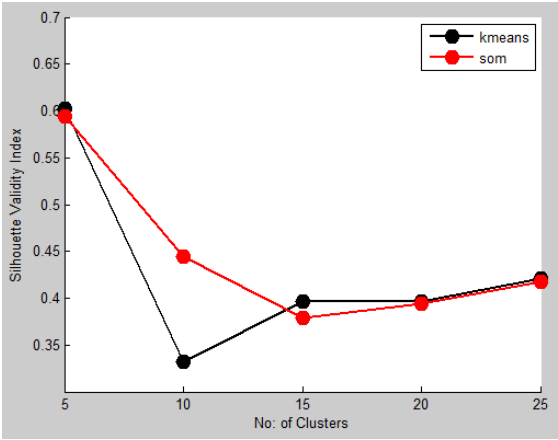

Fig. 13 Silhouette Index Vs No. of Clusters for glass dataset

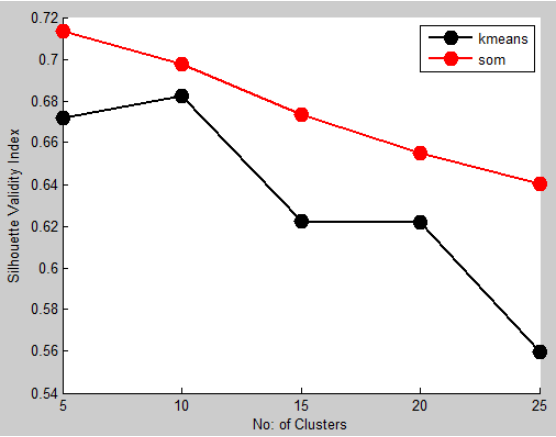

Fig. 14 Silhouette Index Vs No. of Clusters for wine dataset

VI. CONCLUSIONS

In this paper, we have compared the performance of SOM based clustering and K-Means clustering. Performances of both algorithms are compared in terms of SSE and Various Validity measures. Also we compared the time required to run the algorithm for four different datasets for different number of clusters. We found the following observations in our analysis work:

- We found that SSE is less for neural network based clustering compared to K-Means clustering in most of the cases. We can conclude that the accuracy of SOM is better by considering the SSE.
- We found that Xie-Beni Index for SOM method is less for certain number of clusters.
- We conclude that Dunn's index and Silhouette value is better for SOM.
- We found that Silhouette Index is near to 1 for SOM method in most of the cases.

From the above points we can conclude that the performance of SOM based clustering is better than the K-Means clustering.

REFERENCES

[1] Jiawei Han, and Micheline Kambler., *Data mining concepts and techniques*, Morgan Kaufmann Publishers, An imprint of Elsevier, 2006.

[2] Amir Ben-Dor, Ron Shamir, and Zohar Yakini., "Clustering Gene Expression patterns," Journal of Computational Biology, 6(/43):281-297, 1999.

[3] Daxin Jiang, Chum Tong, and Aidong Zhang, "Cluster analysis for gene expression data," IEEE Transactions on Data and Knowledge Engineering, 16(11):1370-1386, 2004.

[4] Han J., and Kamber M., "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2001.

[5] Margaret H. Dunham, "Data Mining- Introductory and Advanced Concepts," Pearson Education, 2006.

[6] McQueen J, "Some methods for classification and analysis ofmultivariate observations," Proc. 5th Berkeley Symp. Math.Statist.Prob., (1):281–297, 1967.

[7] Pang-Ning Tan, Michael Steinback, and Vipin Kumar, "Introduction to Data Mining," Pearson Education, 2007.

[8] Stuart P. Lloyd, "Least squares quantization in pcm," IEEE Transactions on Information Theory, 28(2): 129-136.

[9] Dr.Yashpal Singh, and Alok Singh Chauhan., "Neural Networks in Data Mining," Journal of Theoretical and Applied Information Technology, Volume 5. No 1, 2009.

[10] Kohonen, T., "Self-Organizing Maps," Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995 .

[11] B.G.Obula Reddy, and Dr. Maligela Ussenaiah., "Literature Survey On Clustering Techniques," IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 1 (July-Aug. 2012), PP 01-12.

[12] Martin Hynar, Michal Burda, and Jana Sarmanov, "Unsupervised clustering with growing self-organizing neural network – a comparison with non-neural approach," Department of Computer Science, V˘SB – Technical University of Ostrava 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic.

[13] AmanpreetKaur Toor1, and Amarpreet Singh "Analysis of Clustering Algorithm Based on Number of Clusters, Error Rate, Computation Time and Map Topology on Large Data Set," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 2, Issue 6, November – December 2013 ISSN 2278-6856.

[14] Abdel-Badeeh M. Salem, Mostafa M. Syiam, and Ayad F. Ayad, "Improving Self-Organizing Feature Map (Sofm) Training Algorithm Using K-Means Initialization," Computer Science Department, Faculty of Computer & Information Sciences Ain Shams University, Cairo, Egypt.

[15] Sueli A. Mingoti, and ,Joab O. Lima, "Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms," European Journal of Operational Research 174 (2006) 1742–1759.

[16] M. EmreCelebi, Hassan A. Kingravi, and Patricio A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm," Expert Systems with Applications 40.1 (2013): 200-210.

[17] Fernando Bacao, Victor Lobo, and Marco Painho, "Self-organizing Maps as Substitutes for K-Means Clustering," Computational Science–ICCS 2005.Springer Berlin Heidelberg, 2005.476-483.

[18] Xie, Xuanli Lisa, and Gerardo Beni, "A Validity measure for Fuzzy Clustering," IEEE Transactions on Pattern Analysis and machine Intelligence, Vol.13, No4, 1991.

[19] J.C. Dunn., "Well separated clusters and optimal fuzzy partitions," .J.Cybern. 4.95-104. 1974.

[20] Peter J. Rousseeuw, "Silhouettes:a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and applied Mathematics 20 (1987) 53-65, North-Holland.

[21] Zahid Ansari, M.F. Azeem, Waseem Ahmed, A. Vinaya Babu., "Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions," *World of Computer Science and Information Technology (WCSIT)* ISSN: 2221-0741 Vol. XXX, No. XXX, 2011.

