A Comparative Study on Self-Organization Map Clustering Method Using Breast Cancer Dataset

R. Prabu[1] M. Sudha[2]
[1]M.Phil Research scholar [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Muthayammal College of Arts & Science, Tamilnadu India

Abstract— Artificial neural networks (ANNs) are computational models inspired by an animal's central nervous systems (in particular the brain), and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature. Clustering is a technique to group together a set of items having similar characteristics. In the clustering process can classified in to different types. In those types, partitioning clustering is the one of the clustering methods. In this thesis, an attempt is made to develop an neural network based clustering algorithm in partitioning clustering method for yeast database, The algorithm works faster so and compared with the traditional k means clustering algorithm and tested the performance of the different clustering algorithm with different cluster centroid values and also finding the optimal cluster center to improve the clustering process. The experimental results shows that the enhanced neural networking based clustering algorithm perform well and comparatively better than the traditional k means clustering algorithm for clustering yeast databases.

Key words: Clustering, Breast Cancer Dataset

I. INTRODUCTION

A. *Data Mining*

Data Mining has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "The science of extracting useful information from the large data sets or databases". Data mining is an extraction of hidden predictive information from large databases [14]. These tools can include statistical models, mathematical algorithms, and machine learning methods. Data mining has become increasingly common in both the public and private sectors. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research.

II. NEURAL NETWORK AND CLUSTERING ANALYSIS

A. *Artificial Neural Network*

Neural Network is just a web of inter connected neurons which are millions and millions in number. With the help of this interconnected neurons all the parallel processing is done in human body and the human body is the best example of Parallel Processing. A neuron is a special biological cell that process information from one neuron to another neuron with the help of some electrical and chemical change. It is composed of a cell body or soma and two types of out reaching tree like branches: the axon and the dendrites. The cell body has a nucleus that contains information about hereditary traits and plasma that holds the molecular equipment's or producing material needed by the neurons[2].

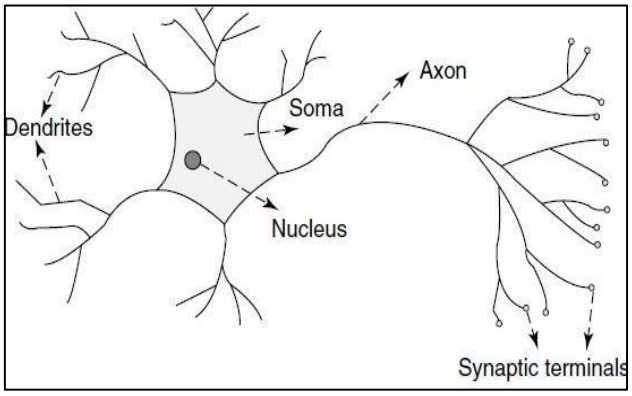

Fig. 1: Human Neuron

An Artificial Neuron is basically an engineering approach of biological neuron. It have device with many inputs and one output. ANN is consist of large number of simple processing elements that are interconnected with each other and layered also.[6], [4]

B. *ANN Characteristics*

Basically Computers are good in calculations that basically takes inputs process then and after that gives the result on the basis of calculations which are done at particular Algorithm which are programmed in the software's but ANN improve their own rules, the more decisions they make, the better decisions may become the Characteristics are basically those which should be present in intelligent System like robots and other Artificial Intelligence Based Applications.

There are six characteristics of Artificial Neural Network which are basic and important for this technology which are showed with the help of diagram.

C. *Network Structure*

The Network Structure of ANN should be simple and easy. There are basically two types of structures recurrent and non-recurrent structure. The Recurrent Structure is also known as Auto associative or Feedback Network [33] and the Non Recurrent Structure is also known as Associative or feed forward Network.[6],[4],[33],[24]. In Feed forward Network, the signal travel in one way only but in Feedback Network, the signal travel in both the directions by introducing loops in the network. The Figures are given below which shows the direction of signals in both the network structures Feed forward and feedback.

D. Self-Organizing Maps

So far we have looked at networks with supervised training techniques, in which there is a target output for each input pattern, and the network learns to produce the required outputs. We now turn to unsupervised training, in which the networks learn to form their own classifications of the training data without external help. To do this we have to assume that class membership is broadly defined by the input patterns sharing common features, and that the network will be able to identify those features across the range of input patterns.One particularly interesting class of unsupervised system is based on competitive learning, in which the output neurons compete amongst themselves to be activated, with the result that only one is activated at any one time. This activated neuron is called a winner-takes all neuron or simply the winning neuron. Such competition can be induced/implemented by having lateral inhibition connections (negative feedback paths) between the neurons. The result is that the neurons are forced to organize themselves. For obvious reasons, such a network is called a Self-Organizing Map (SOM).

E. Organization of the Mapping

We have points x in the input space mapping to points I(x) in the output space: Each point I in the output space will map to a corresponding point w(I) in the input space.

F. Kohonen Networks

We shall concentrate on the particular kind of SOM known as a Kohonen Network. This SOM has a feed-forward structure with a single computational layer arranged in rows and columns. Each neuron is fully connected to all the source nodes in the input layer:

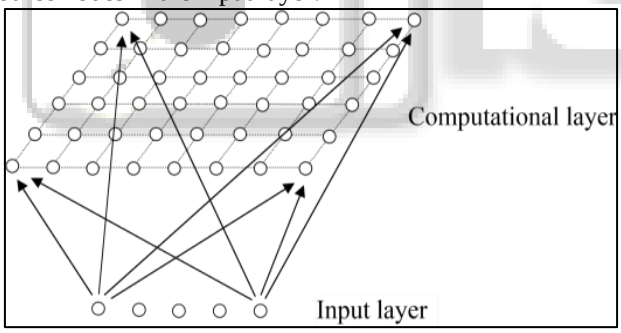

Fig. 2: Kohonen network

Clearly, a one dimensional map will just have a single row (or a single column) in the computational layer.

G. Overview of the SOM Algorithm

We have a spatially continuous input space, in which our input vectors live. The aim is to map from this to a low dimensional spatially discrete output space, the topology of which is formed by arranging a set of neurons in a grid. Our SOM provides such a nonlinear transformation called a feature map. The stages of the SOM algorithm can be summarized as follows:

- Initialization – Choose random values for the initial weight vectors wj.
- Sampling – Draw a sample training input vector x from the input space.
- Matching – Find the winning neuron I(x) with weight vector closest to input vector.
- Updating – Apply the weight update equation $\Delta^{w}_{ji} = \eta(t)\, ^{T}_{j,I(x)}(t)\, (^{x}_{i} - w_{ji})$.
- Continuation – keep returning to step 2 until the feature map stops changing.

III. PARTITIONAL CLUSTERING TECHNIQUES

A. Partitional Algorithms

Partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such as dendogram produced by a hierarchical technique. Partitional methods have advantages in applications involving large data sets for which the construction of a dendogram is computationally prohibitive.

B. K-Means Algorithm

The K-means [9], [10], [23], [24] method aims to minimize the sum of squared distances between all points and the cluster Centre. This procedure consists of the following steps, as described by Tou and Gonzalez [97].

1) Choose K initial cluster centres z_1 (1), z_2 (1)…z_k (1).
2) At the k-th iterative step, distribute the samples {x} among the K clusters using the relation
3) For all i=1, 2…K; I ≠ j; where C_j (k) denotes the set of samples whose cluster centre is z_j (k).
4) Compute the new cluster centres z_j (k+1), j=1, 2…K such that the sum of the squared distances from all points in C_j (k) to the new cluster centre is minimized. The measure which minimizes this is simply the sample mean of C_j (k).
5) If z_j (k+1) =z_j (k) for j=1, 2…K then the algorithm has converged and the procedure is terminated.
6) Otherwise go to step 2.

C. Working Principle

The K-Means algorithm working principles are clearly explained in the following algorithm steps.

K-Means Algorithm[10]

Input:D = {d1, d2, d3, ..., di, ..., dn } // Set of n data points.

k =Number of desired clusters

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest centroids.

Steps:

1) Initialize the number of clusters k.
2) Randomly selecting the centroids in the given data set $(c_1 c_2 \ldots c_k)$.
3) Compute the distance between the centroids and objects using the Euclidean Distance equation.

$$d_{ij} = \left|\left|x_i - c_k\right|\right|^2$$

4) Update the centroids.
5) Stop the process when the new centroids are nearer to old one Otherwise, go to step-3.

The K-Means algorithm is an iterative procedure for clustering the objects which requires an initial classification of the data. The K-Means algorithm proceeds as follows:

- It computes the centre of each cluster, and then computes new partitions by assigning every object to the cluster whose centre is the closest to that object.
- This cycle is repeated during a given number of iterations or until the assignment has not changed during one iteration.

D. Enhanced K-Means Algorithm

In the Enhance K-means method, first, it determines the initial cluster centroids by using the equation which is given in the following algorithm 2. The Enhance K-Means algorithm is improved by selecting the initial centroids manually instead of selecting centroids by randomly. It selects 'K' objects and each of which initially represents a cluster mean or centroids. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar based on the 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) Enhanced k-means Algorithm [10]

Input: a set of n data points and the number of clusters (K)
Output: centroids of the K clusters
Steps:

1) Initialize the number of clusters k.
2) Selecting the centroids $(c_1 c_2 \dots c_k)$ by initial centroid selection method in the data set.
3) Using Euclidean distance as a dissimilarity measure, compute the distance between every pair of all objects as follow.

$$d_{ij} = \sqrt{\sum_{a=}^{p} (X_{ia} - X_{ja})^2} \quad i\&j = 1, \dots, n; \quad (1)$$

4) Calculate M_{ij} to make an initial guess at the centres of the clusters

$$M_{ij} = \frac{d_{ij}}{\sum_{i=1}^{n} d_{ij}} \quad i\,\&j = 1, \dots, n. \qquad (2)$$

5) Calculat $\sum_{i=1}^{n} M_{ij}^2{}_{(j=1,..,n)}$ .... (3)at each object and sort them in ascending order.
6) Select K objects having the minimum value as initial cluster centroids which are determined by the above equation. Arbitrarily choose k data points from D as initial centroids.
7) Find the distance between the centroids using the Euclidean Distance equation.
   $d_{ij} = \left|\left| w_{.} * (x_i - c_k) \right|\right|^2$
8) Update the centroids using this equation.
9) Stop the process when the new centroids are nearer to old one. Otherwise, go to step-4.

The Enhanced K-Means algorithm is used to clustering the objects. Using this algorithm we can also selecting the initial centroids manually instead of randomly and clustering the data in the dataset. The dataset are taken from www.ucirepository.com website.

E. Self-Organization Map

A Self-Organizing Map [6, 7], or SOM, is a neural clustering technique. It is more sophisticated than Kmeans in terms of presentation; it not only clusters the data points into groups, but also presents the relationship between the clusters in a two-dimensional space.

The input vectors are connected to an array of neurons (usually 1 dimensional (a row) or 2 dimensional (a rectangular lattice))When an input is presented, certain region of the array will fire and the weights connecting the inputs to that region will be strengthened.

1) During Learning Process

- The weight connecting the input space to the winning neuron are strengthened
- The weights of neurons in the "neighbourhood" of the winning neuron are also strengthened.

2) SOM clustering Algorithm:

- Select output layer network topology
  Initialize current neighborhood distance, D(0), to a positive value
- Initialize weights from inputs to outputs to small random values
- Let t = 1
- While computational bounds are not exceeded do

1) Select an input sample
2) Compute the square of the Euclidean distance of d_{ij} From weight vectors (W_j) associated with each output node

$$w_j = \sum_{k=1}^{n} (i_{l,k} - w_{i,k}(t))^2 \quad (2)$$

3) Select output node j that has a weight vector with minimum value from step 2.*
4) Update weights to all nodes within a topological distance given by D(t) from j, using the weight update rule below:*
5) Increment t

End while

IV. Experimental Results and Discussion

A. Data Set Information: (Breast cancer)

The breast cancer dataset can be download for the website www.ucirepository.com and http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

B. Cluster Validity Measures and Techniques

Many criteria have been developed for determining cluster validity. All of which have a common goal to find the clustering which results in compact clusters which are well separated. Clustering validity is a concept that is used to evaluate the quality of clustering results.

C. Davies-Bouldin Validity Index

This index (Davies and Bouldin, 1979) is a function of the ratio of the sum of within-cluster scatter to between-cluster separation If dp_i is the dispersion of the cluster Pi, and dv_{ij}denotes the dissimilarity between two clusters Pi and Pj, then a cluster similarity matrix FR = { FR_{ij}, (i, j) = 1; 2 …..C} is defined as:

$$FR_{ij} = \frac{dp_i + dp_j}{dv_{ij}}$$

The dispersion dpi can be seen as a measure of the radius of Pi,

$$dp_i = \left(\frac{1}{n_i} \sum_{x \sum p_i} ||\, x - V_i|| \quad ^2 \right) \quad ^{\frac{1}{2}}$$

Where $n_i$is the number of objects in the i^{th} cluster.
V_i is the centroid of the i^{th} cluster.
dv_{ij}describes the dissimilarity between P_i and P_j,

$$dv_{ij} = ||V_i - V_j|| \quad ^2$$

The corresponding DB index is defined as:

$$DB_{FR} = \frac{1}{c}\sum_{i=1}^{c} FR_i$$

Here, c is the number of clusters.Hence the ratio is small if the clusters are compact and far from each other. Consequently, Davies-Bouldin index will have a small value for a good clustering

D. Breast Cancer Detection and Comparative Study Methods

The breast cancer dataset can be classified by using clustering algorithm called SOM clustering which classified the 698 data into 80 data as normal pain patients and 358 patients affected in Benign Breast cancer and 260 patients affected in Malignant breast cancer.

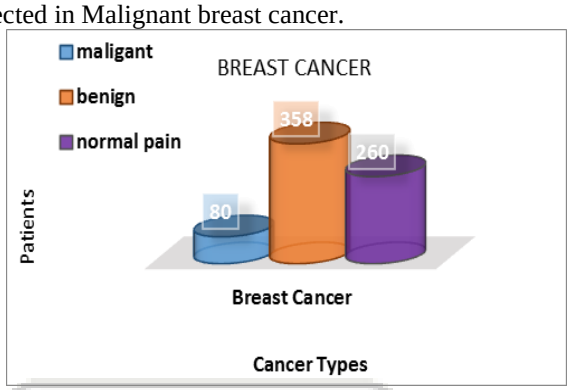

Fig. 3: Clustering breast cancer dataset

From the above the figure 3, it clearly shows that the breast cancer dataset can be classified with the help of the SOM clustering algorithm. The SOM clustering algorithm correctly clustering cancer dataset of 698 patients including cancer and non-cancer patients by two different types of breast cancer is benign and malignant with normal pain patient. The SOM clustering algorithms are very clear to cluster the cancer data and the following section is describing the improvement of SOM clustering algorithms.

E. Comparative Study on Clustering Algorithms

In this thesis, there are three different clustering algorithms were implemented on breast cancer data collected from the website. Table 1 to 4 provides the results obtained for the various algorithms described in this thesis. The values obtained for the Davies-Bouldin validity index are specified in the table. Clustering results have a set of patients which is in the different types of breast cancer.

F. Parameter Tuning

Learning rate is the one of the parameter in self-organisation map clustering algorithm, the SOM clustering algorithm can improve by tuning the parameter learning rate (α). The learning rate parameter can be tuned from the value 0.1 to 1.0 and execute the SOM clustering algorithm with breast cancer dataset and the obtained DB index value for the different parameter values are depicted in the below table 1.

S. no	clusters	Learning rate (α)	DAVIS-BOULDIN INDEX
1	10	0.1	1.548
2		0.2	1.724
3		0.3	1.676
4		0.4	1.489
5		0.5	1.104
6		0.6	0.986
7		0.7	1.084
8		0.8	0.945
9		0.9	0.896
10		1	1.001

Table 1 Parameter Tuning

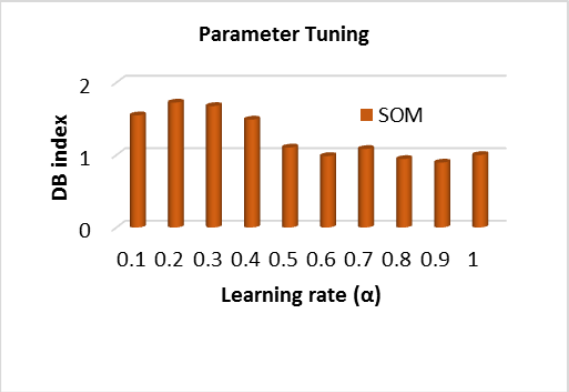

Fig. 4: Parameter Tuning

From the fig 4. it clearly shows that the SOM clustering algorithm is executed with the cancer dataset and increasing the learning rate from0.1 to 1.0, but the learning rate parameter from 0.5 to 0.9 obtain minimum DB index value for SOM clustering algorithm, in particular the learning rate 0.9 obtain minimum DB index than all other learning rate value in SOM, hence the learning rate 0.9 for SOM clustering algorithm produce better results than other parameters values.

G. Clusters Analysis

In the SOM clustering algorithm, we modify the number of clusters by 3 at each time, and then we obtain the different DB index values for the different distance function. In this process the number of data in cancer dataset are remain constant. The various DB index values and cluster centre are depicted in the following table 2.

S. no	Clusters	DAVIS-BOULDIN INDEX	
		Euclidean distance	Manhattan distance
1	3	0.294	0.295
2	6	0.431	0.571
3	9	0.424	0.758
4	12	0.408	0.684
5	15	0.385	0.642
6	18	0.57	0.848
7	21	0.415	0.67
8	24	0.606	0.71
9	27	0.548	0.654
10	30	0.874	0.987

Table 2: Performance on DB index for different Clusters

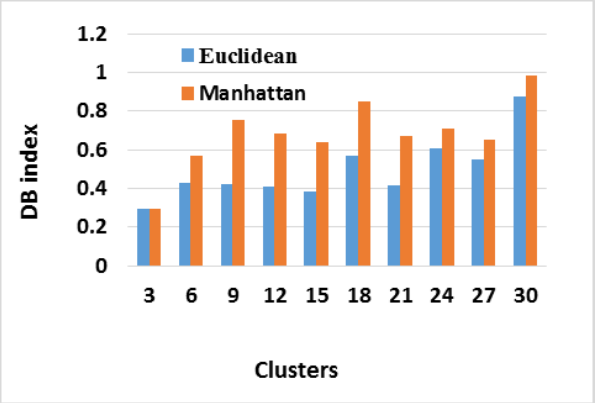

Fig. 5: Comparison of DB index for different distance function with different clusters.

In the figure5, there are two different distance functions are used to calculate the DB index value for the SOM algorithm. Here the cluster values differ from 3 to 30.We compared the DB index value for SOM algorithm using Euclidean function and Manhattan function, but the Euclidean distance function yields the better result for most of the cluster values. Hence the Euclidean distance function is the better suitable than Manhattan distance function for the SOM clustering algorithm.

H. Performance Analysis

The SOM Clustering algorithm is compared with the K-Means, Enhanced K-Means clustering algorithm by selecting initial centroids manually instead of selecting the centroids randomly. The algorithm is executed by setting different cluster values from 2 to 20 and the obtained DB index values of various clustering algorithms is depicted in below table 3.

S. No	Clusters	Davis Bouldin Index		
		K-Means	Enhanced K-Means	SOM
1	2	1.5451	1.5465	0.957
2	4	1.5315	1.3217	1.348
3	6	1.4241	1.3457	1.548
4	8	1.4921	1.5121	1.348
5	10	1.4938	1.3915	1.654
6	12	1.5483	1.4042	1.451
7	14	1.4969	1.4523	1.568
8	16	1.4835	1.5426	1.458
9	18	1.5687	1.3999	2.451
10	20	1.5119	1.3532	2.689

Table 3: Performance of Different clustering algorithm

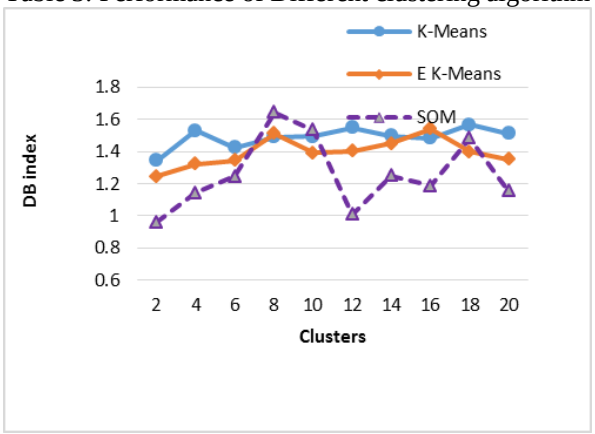

Fig. 6: Performance on various clustering algorithms.

From the fig 6 it shows that we compared the performance of the three clustering algorithm called K-Means, Enhanced K-Means and SOM algorithm with the help of DB index values, the values are depicted in the table and the comparison are done in the above chart. From the chart 5.4.4, clearly we identify that the performance of the SOM clustering algorithm produce better result than the K-Means algorithm and Enhanced Kmeans. Hence the SOM clustering algorithm best suitable for clustering process.

V. Conclusion

In this study, the clustering methods and clustering algorithm K-Means, enhanced K-Means and self-organization map is studied well and compared with one another. The different clustering algorithms are executed with the breast cancer dataset. During our tests it is quite evident that the search space is better explored by SOM. The SOM clustering algorithm is executed with two different distance function, the experimental results shows that the Euclidean distance function produce better clustering results than Manhattan distance function. The learning rate in the SOM is one of the major factors for the clustering process. In this thesis the SOM clustering algorithm is executed by changing the learning rate from 0.1 to 1.0, but the learning rate 0.9 for SOM clustering algorithm obtained better results than other learning rates. The three different clustering algorithm K-Means, Enhanced K-Means and SOM are executed by breast cancer dataset with different cluster values and the clustering results are validated by the Davis Bouldin index, the SOM clustering algorithm obtained the minimum DB index for the most of the cluster, hence the SOM method is better suitable for clustering the breast cancer dataset than K-Means and Enhanced K-Means clustering methods. The SOM clustering algorithm is enhanced by selecting the initial centroids in systematic method and validated by the different index method is our future wok.

References

[1] Ajith Abraham, "Artificial Neural Networks", Stillwater,OK, USA, 2005.
[2] Anil K Jain, Jianchang Mao and K.M Mohiuddin, "Artificial Neural Networks: A Tutorial", Michigan State university, 1996.
[3] Carlos Gershenson, "Artificial Neural Networks for Beginners", United kingdom.
[4] Christos Stergiou and Dimitrios Siganos, "Neural Networks".
[5] Davies & Bouldin, 1979. Davies, D.L., Bouldin, D.W., (2000) "A cluster separation measure." IEEE Trans.Pattern Anal. Machine Intell., 1(4), 224-227.
[6] Eldon Y. Li, "Artificial Neural Networks and their Business Applications", Taiwan, 1994.
[7] Freitas AA and Lavington SH. "Mining Very Large Databases with Parallel Processing", Kluwer, 1998.
[8] Girish Kumar Jha, "Artificial Neural Network and its Applications",IARI New delhi.
[9] Hae-Sang Park, Jong-Seok Lee, and Chi-Hyuck Jun, "K-means-like Algorithm for K-medoids Clustering and Its Performance".

[10] HARTIGAN, J. and WONG, M., Algorithm AS136: "A kmeans clustering algorithm". Applied Statistics, 28, 100-108, 1979
[11] Herve Debar, Monique Becker and Didier Siboni " A Neural Network Component for an Intrusion Detection System", Les Ulis Cedex France, 1992,
[12] Howard Demuth and Mark Beale, "Neural Network Toolbox", with the help of matlab, user guide version 4.
[13] Image of a Neuron form website http://transductions.net/ 2010/02/04/313/neurons.

