International Journal of Advanced Studies in Computer Science and Engineering

IJASCSE, Volume 3, Issue 1, 2014

Clustering Breast Cancer Dataset using Self-Organizing Map Method

M. Thenmozhi
Assistant Professor
Dept. of Computer Science
Mahendra Arts & Science College, kaalipatti, Tamilnadu, India

I. Dhanalakshmi
Assistant Professor
Dept. of Computer Science
Vivekananda Arts & Science, College Thiruchencode, Tamilnadu, India

S. Krishnaveni
Assistant Professor
Dept. of Computer Science
Mahendra Arts & Science College, kaalipatti, Tamilnadu, India

R. Jeevambal
Assistant Professor
Dept. of Computer Science
Mahendra Arts & Science College, kaalipatti, Tamilnadu, India

Abstract: Data mining is the sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Breast cancer is one of the major causes of death among women. Small clusters of micro calcifications appearing as collection of white spots on mammograms show an early warning of breast cancer. Early detection performed on X-ray mammography is the key to improve breast cancer diagnosis.

Clustering is a technique to group together a set of items having similar characteristics. In the clustering process can classified into different types.Partitioning clustering is the one of the clustering methods. In this paper, an attempt is made to develop an SOM clustering algorithm method for breast cancer database. The algorithm works faster so and compared with the traditional k meansand enhanced 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 SOM clustering algorithm perform well and comparatively better than the traditional K-Means and enhanced K-Means algorithm for clustering breast cancer databases.

Keywords: Breast cancer, Self-Organizing Map K-Means, Neural network, Clustering.

I. INTRODUCTION

A. Neural Network and Clustering Analysis

Artificial neural network[5] [1]

The concept of ANN is basically introduced from the topic of biology where neural network plays an important and key role in human body. In human body work is done with the help of 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.

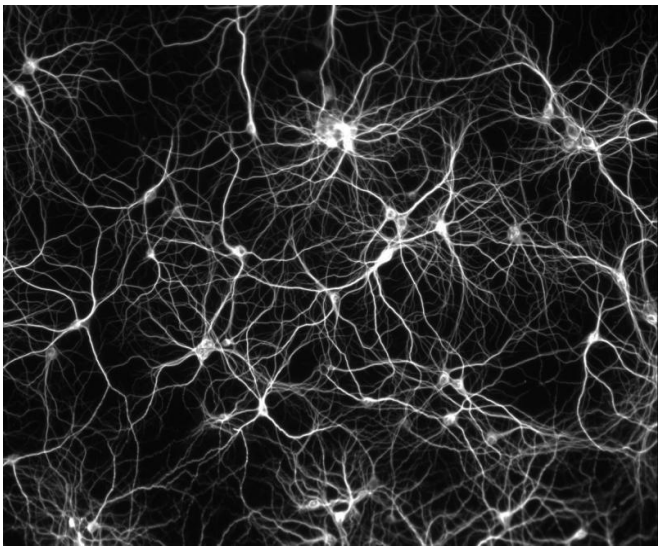

Figure 1. Neural Network in Human Body [11]

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].

The whole process of receiving and sending signals is done in particular manner like a neuron receives signals from other neuron through dendrites. The Neuron transmit signals at spikes of electrical activity through a long thin stand known as an axon and an axon splits this signals through synapse and send it to the other neurons [3].

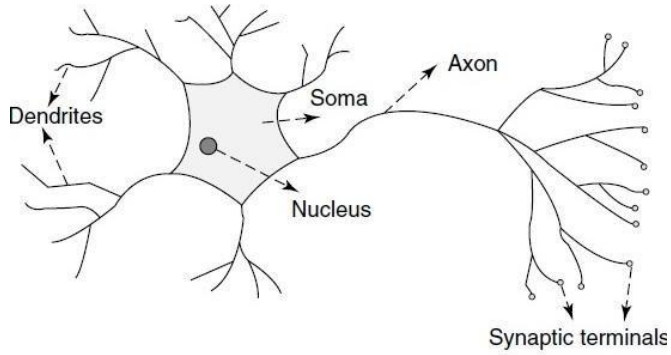

Figure 2. ***Human Neuron***

An Artificial Neuron is mainly 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.[4], [6]. Similar to biological neuron Artificial Neural Network also have neurons which are artificial and they also receive inputs from the other elements or other artificial neurons and then after the inputs are weighted and added, the result is then transformed by a transfer function into the output. The transfer function may be anything like Sigmoid, hyperbolic tangent functions or a step.

This paper is organized as follows. Section II presents an overview of Clustering techniques and its method. Section III describes performance of Experimental analysis and results discussion. Section IV presents conclusion and future work.

II. CLUSTERING TECHNIQUE

Partitional clustering [12] 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 structure of a dendogram is computationally prohibitive. A problem accompanying the use of a Partitional algorithm is the choice of the number of desired output clusters. The Partitional technique usually produce clusters by optimizing a criterion function defined either locally or globally.

A. K-Means Algorithm:

K-Means [9], [8], [13] is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group is done. At this point, it is need to re-calculate k new centroids as centres of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has been generated. As a result of this loop it may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. The computational complexity of the original K-Means algorithm is very high, especially for large Data sets. The algorithm is composed in the following steps:

Algorithm 1: K-Means Clustering Algorithm

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.

Methods

1. Arbitrarily choose *k* data points from *D* as initial centroids;
2. Assign each point d_i to the cluster which has the closest centroid;
3. Calculate the new mean for each cluster;
4. Repeat **step 2** and **step 3** until convergencecriteria is met.

B. Enhanced K-Means Algorithm

Enhanced K-Means [7], [14]algorithm is one of the clustering algorithms based on the K-Means algorithm calculating with initial centroid selection method instead of selection centroid randomly. This algorithm is same as normal K-Means algorithm but differ in selecting the initial centroid to improve the clustering process. The Enhanced k-means algorithm uses initial centroid to decrease the effects of irrelevant attributes and reflect the semantic information of objects. Enhanced K-Means algorithms are iterative and use hill-climbing to find an optimal solution (clustering), and thus usually converge to a local minimum.

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.The Enhanced K-Means clustering working procedure is given in the following algorithm step.

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***

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

5. ***Calculate*** $\sum_{i=1}^{n} M_{ij}^2{}_{(j=1,..,n)}$ .... ***(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.

C. Self-Organizing Map

A Self-Organizing Map [6], [7], or SOM, is a neural clustering technique. It is more stylish than Kmeans in terms of presentation and not only clusters the data points into groups, but also presents the relationship between the clusters in a two-dimensional space. The SOM concept is outlined in Figure 3 and the algorithm is presented in Figure 4.

.

Basic concepts of SOM

The input vectors are connected to an array of neurons (usually 1 dimensional (a row) or 2 dimensional (a rectangular lattice))

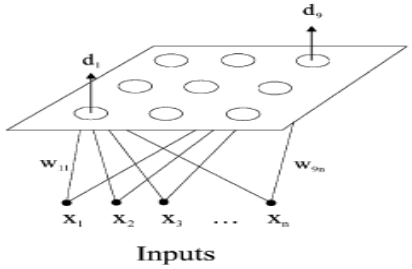

Figure 3. Self-Organizing map concepts

When an input is presented, certain region of the array will fire and the weights connecting the inputs to that region will be strengthened.

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.

Once the learning is complete, similar inputs will "fire" the same regions. In this way, similar inputs patterns can be identified and grouped together or clustered.

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

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

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:**

$$w_j(t+1) = w_j(t) + \eta(t)(i_l - w_j(t))$$

5) Increment *t* End while

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Data Set Information: (Breast cancer)

Thebreast cancer dataset can be download for the website **www.ucirepository.com.**Which contains 698 patients details with 10 attributes values which are listed and explained in the below section.

B. Cluster Validity Measures and Techniques

A lot of criteria have been developed for determining cluster validity. Now we present a validity criterion based on a validity function which identifies compact and separate partitions without assumptions as to the number of substructures inherent in the data. This function depends on the data set, geometric distance measure, distance between cluster centroids and more importantly on the partition generated by any K-Means algorithm used. The function is mathematically justified via its relationship to a well-defined hard clustering validity function, the separation index for which the condition of uniqueness has already been established. The performance of this validity function compares favorably to that of several others.In a generalization of **Davies-Bouldin validity index** is discussed and measure the compactness and separation of clusters.

1. 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,

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

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,

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

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

C. Breast Cancer Detection And Comparative Study Methods

1. Breast cancer Detection

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.

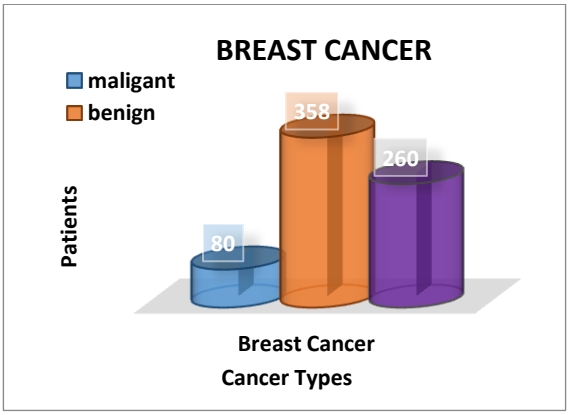

Figure 4. Clustering breast cancer dataset

From the above the figure 4, 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 are benign and malignant with normal pain patient. The SOM clustering algorithms are very clear to cluster the cancer data and the following section are describe the improvement of SOM clustering algorithms.

2. Comparative Study on clustering algorithms:

In this paper, there are three dissimilar clustering algorithms were implemented on breast cancer data collected from the website. Table I to V 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.

a) Distance Function:

In the SOM algorithm, to calculate the distance between the data object and the centroid with the help of the two major distance functions are

- Euclidean Distance
- Manhattan Distance

Euclidean distance or **Euclidean metric** is the "ordinary" distance between objects and centroids, which can be proven by repeated application of the Pythagorean Theorem. By using this formula as distance, the associated norm is called the **Euclidean norm**.

$$d_{ij} = ||x_i - c_j||^2$$

The taxicab metric is also known as **rectilinear distance**, L_1 **distance** or L^1 **norm**, **city block distance**, **Manhattan distance**, or **Manhattan length**, with corresponding variations in the name of the geometry. The last name alludes to the grid layout of most streets on the island of Manhattan, which causes the shortest path a car could take between two points in the city to have length equal to the points' distance in taxicab geometry.

$$d_{ij} = |x_i - c_j|$$

Here, dij is the distance between the object in x and the j^{th}cluster

In this paper, we calculate the DB index value for the SOM algorithm by using the two different distance function called Euclidean and Manhattan function. The values are depicted in the below Table I.

TABLE I. Comparative Analysis of various Distance functions

S.No	Dataset	Davis Bouldin index	
		Euclidean Distance	Manhattan Distance
1	50	0.3966	0.525
2	100	0.4132	0.568
3	150	0.4088	0.566
4	200	0.4311	0.569
5	250	0.4314	0.571
6	300	0.4149	0.579
7	350	0.4122	0.569
8	400	0.4145	0.571
9	450	0.4128	0.572
10	500	0.4314	0.576

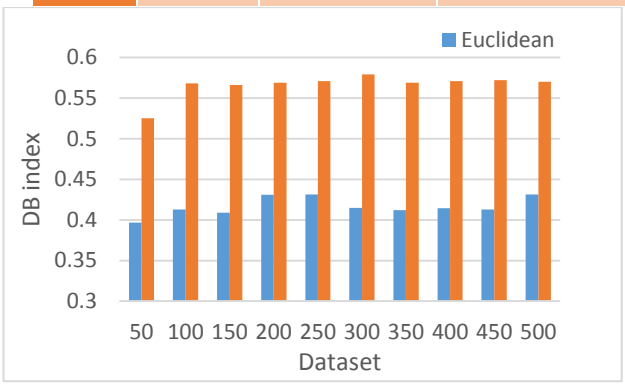

Figure 5. DB index for various distance functions.

From the figure 5, we compared that the various distance function called Euclidean and Manhattan distance for the SOM algorithm. Here we use 500 data are used in the clustering process. Consequently, Davies-Bouldin index value for the SOM algorithm using Euclidean distance function obtained a small DB Index value then Manhattan distance function for a good clustering.

b) 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 II

TABLE II. Parameter Tuning

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

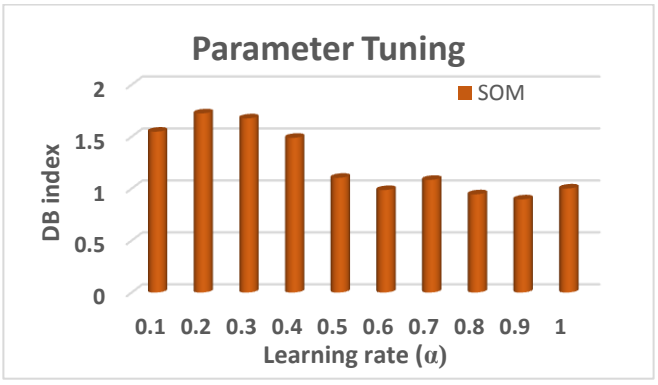

Figure 6. Parameter Tuning

From the figure 6, 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.

c) 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 III

TABLE III. Performance on DB index for different Clusters

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	0.571	0.848
7	21	0.415	0.671
8	24	0.606	0.711
9	27	0.548	0.654
10	30	0.874	0.987

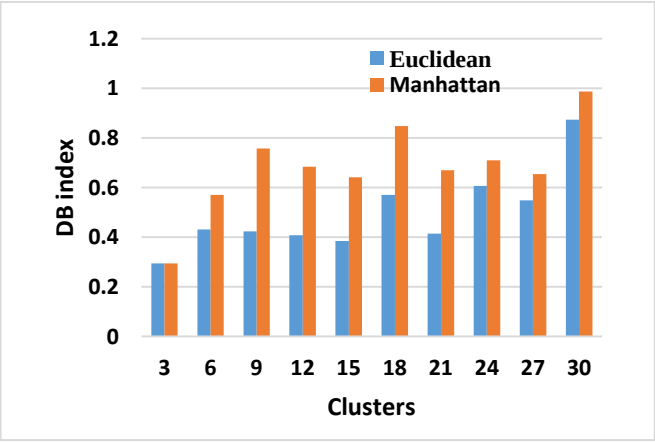

Figure 7. Comparison of DB index for different distance function with different clusters

In the figure 7, 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.

d) 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 IV.

TABLE IV. Performance of Different clustering algorithm

S. No	Clusters	Davis Bouldin Index		
		K-Means	*Enhanced K-Means*	*SOM*
1	2	1.545	1.546	0.957
2	4	1.531	1.321	1.348
3	6	1.424	1.345	1.548
4	8	1.492	1.512	1.348
5	10	1.493	1.391	1.654
6	12	1.548	1.404	1.451
7	14	1.496	1.452	1.568
8	16	1.483	1.542	1.458
9	18	1.568	1.399	2.451
10	20	1.511	1.353	2.689

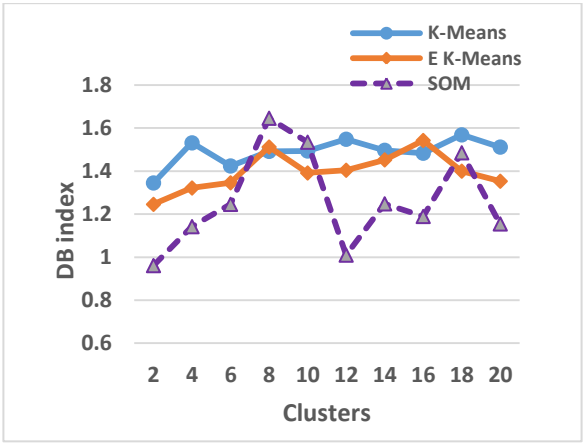

Figure 8. Performance on various clustering algorithms.

From the figure 8, 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 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.

e) Execution Time

The execution time for the three algorithms are already discussed in the section II, the execution time of the K-Means, Enhanced K-Means and SOM clustering algorithms are calculated with various cluster values and listed in the below Table V

TABLE V. Performance Analysis

S. No	Clusters	Execution Time (seconds)		
		K-Means	*Enhanced Kmeans*	*SOM*
1	3	1.584	1.348	0.900
2	6	1.784	1.871	1.348
3	9	1.957	1.254	1.548
4	12	2.139	1.741	1.348
5	15	1.981	1.575	1.654
6	18	2.934	2.096	1.451
7	21	3.871	2.861	1.568
8	24	2.941	2.078	1.458
9	27	4.922	3.072	2.451
10	30	4.861	3.342	2.689

Figure 9. Execution Time chart for K-Means and Ekmeans and SOM.

From the figure 9, the execution time of the K-Means Enhance K-Means and SOM clustering algorithms are shown, the SOM clustering algorithm obtain the minimum execution time for most of the different cluster values due to this algorithm converged with small number of iterations, hence the SOM clustering algorithm is better than the traditional K-Means algorithm and Enhanced Kmeans for breast cancer data set.

IV. CONCLUSION

In this research, the clustering methods and clustering algorithm K-Means, enhanced K-Means and Self-Organizing map is studied well and compared with one another. The different clustering algorithms are executed with the breast cancer dataset. The first and most important conclusion that can be drawn from this study is that SOM is less prone to local optima than K-Means. During our tests it is quite evident that the search space is better explored by SOM. This is due to the effect of the neighborhood parameter which forces units to move according to each other in the early stages of the process. On the other hand, K-Means gradient orientation forces a premature convergence which, depending on the initialization, may frequently yield local optimum solutions. It is important to note that there are certain conditions that must be observed in order to render robust performances from SOM. First it is important to start the process using a high learning rate and neighborhood radius, and progressively reduce both parameters to zero. 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 factor 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 values and also the SOM clustering method deliver the result within the minimum execution time than other clustering algorithm, 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] Christos Stergiou and DimitriosSiganos, "*Neural Networks*".

[4] Eldon Y. Li, "Artificial Neural Networks and their Business Applications", Taiwan, 1994.

[5] Feed Back Network from website *http://www.idsia.ch/juergen/rnn.html*.

[6] Freitas AA and Lavington SH. "*Mining Very Large Databases with Parallel Processing",* Kluwer, 1998.

[7] Hae-Sang Park, Jong-Seok Lee, and Chi-Hyuck Jun, "K-means-like Algorithm for K-medoids Clustering and Its Performance".

[8] Hae-Sang Park, Jong-Seok Lee, and Chi-Hyuck Jun, "K-means-like Algorithm for K-medoids Clustering and Its Performance".

[9] HARTIGAN, J. and WONG, M. 1979. Algorithm AS136: "A k-means clustering algorihm". Applied Statistics, 28, 100-108.

[10] HARTIGAN, J. and WONG, M., Algorithm AS136: "A kmeans clustering algorithm". Applied Statistics, 28, 100-108, 1979

[11] Image of a Neuron form website http://transductions.net/ 2010/02/04/313/neurons.

[12] Jiawei Han and MichelineKamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 2001.

[13] SauravjoytiSarmah and Dhruba K. Bhattacharyya. May 2010 "An Effective Technique for Clustering Incremental Gene Expression data", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3.

[14] Velmurugan T and T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms," *Journal of Computer Science*, vol. 6, no. 3, 2010.