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 means and 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]
**Algorithm 1: K-Means Clustering Algorithm**

**Input:**

$D = \{d_1, d_2, d_3 \ldots d_n\}$ // 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 convergence criteria 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 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 Enhanced K-means method, first, it determines the initial cluster centroids by using the equation which is given in the following algorithm 2. The Enhanced 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 \((c_1c_2 \ldots c_k)\) 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.

\[
d_{ij} = \left( \sum_{a=1}^{p} (x_{ja} - x_{ja})^2 \right)^{1/2} \text{ for } i \& j = 1, \ldots, n; \quad (1)
\]

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

\[
M_{ij} = \frac{d_{ij}}{\sum_{i=1}^{n} d_{ij}} \text{ for } i \& j = 1, \ldots, n. \quad (2)
\]

5. Calculate \( \sum_{i=1}^{n} M_{ij}(j=1 \ldots n) \) 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.

\[
d_{ij} = \left| \left| w \ast (x_i - c_k) \right| \right|^2
\]

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](image)

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
- \( \text{Let } t = 1 \)
- 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

\[
W_j = \sum_{i=1}^{n} (i_{ik} - W_{i,k}(t))^2
\]

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:

\[
w_{ji}(t+1) = w_{ji}(t) + \eta(t)(i_j - w_{ji}(t))
\]

5) Increment \( t \) End while
III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Data Set Information: (Breast cancer)

The breast cancer dataset can be downloaded from 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 $d_{pi}$ is the dispersion of the cluster $Pi$, and $d_{vij}$ denotes the dissimilarity between two clusters $Pi$ and $Pj$, then a cluster similarity matrix $FR = \{ FR_{ij}, (i, j) = 1; 2 \ldots C\}$ is defined as:

$$FR_{ij} = \frac{d_{pi} + d_{pj}}{d_{vij}}$$

The dispersion $d_{pi}$ can be seen as a measure of the radius of $Pi$,

$$d_{pi} = \left( \frac{1}{n_i} \sum_{x \in P_i} \left\| x - V_i \right\|^2 \right)^{\frac{1}{2}}$$

Where $n_i$ is the number of objects in the $i^{th}$ cluster. $V_i$ is the centroid of the $i^{th}$ cluster.

$$d_{vij} = \left\| V_i - V_j \right\|^2$$

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, L1 distance or L1 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, \( d_{ij} \) 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>
<thead>
<tr>
<th>S.No</th>
<th>Dataset</th>
<th>Davis Bouldin index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Euclidean Distance</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.3966</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>0.4132</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
<td>0.4088</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>0.4311</td>
</tr>
<tr>
<td>5</td>
<td>250</td>
<td>0.4341</td>
</tr>
<tr>
<td>6</td>
<td>300</td>
<td>0.4149</td>
</tr>
<tr>
<td>7</td>
<td>350</td>
<td>0.4122</td>
</tr>
<tr>
<td>8</td>
<td>400</td>
<td>0.4145</td>
</tr>
<tr>
<td>9</td>
<td>450</td>
<td>0.4128</td>
</tr>
<tr>
<td>10</td>
<td>500</td>
<td>0.4314</td>
</tr>
</tbody>
</table>

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 (\( \alpha \)). 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>
<thead>
<tr>
<th>S. No</th>
<th>Clusters</th>
<th>Learning rate (( \alpha ))</th>
<th>Davies-Bouldin Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.1</td>
<td>1.548</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.2</td>
<td>1.724</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.3</td>
<td>1.676</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.4</td>
<td>1.489</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.5</td>
<td>1.104</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.6</td>
<td>0.986</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.7</td>
<td>1.084</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.8</td>
<td>0.945</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.9</td>
<td>0.896</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>1</td>
<td>1.001</td>
</tr>
</tbody>
</table>

From the figure 6, it clearly shows that the SOM clustering algorithm is executed with the cancer dataset and increasing
the learning rate from 0.1 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>
<thead>
<tr>
<th>S. No</th>
<th>Clusters</th>
<th>Euclidean distance</th>
<th>Manhattan distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.294</td>
<td>0.295</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.431</td>
<td>0.571</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>0.424</td>
<td>0.758</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>0.408</td>
<td>0.684</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>0.385</td>
<td>0.642</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>0.571</td>
<td>0.848</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>0.415</td>
<td>0.671</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>0.606</td>
<td>0.711</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>0.548</td>
<td>0.654</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>0.874</td>
<td>0.987</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>S. No</th>
<th>Clusters</th>
<th>Davis Bouldin Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>K-Means</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1.545</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1.531</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>1.424</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1.492</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>1.493</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>1.548</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>1.496</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>1.483</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>1.568</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>1.511</td>
</tr>
</tbody>
</table>

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 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>
<thead>
<tr>
<th>S. No</th>
<th>Clusters</th>
<th>Execution Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K-Means</td>
<td>Enhanced Kmeans</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1.584</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1.784</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>1.957</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>2.139</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>1.981</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>2.934</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>3.871</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>2.941</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>4.922</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>4.861</td>
</tr>
</tbody>
</table>

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


