International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 2– No.3, May 2012 – www.ijais.org

Kohonen Self Organizing Map with Modified K-means clustering For High Dimensional Data Set

Madhusmita Mishra
Depart. of Comp. Sci. & Engineering
V. S .S. University of Technology, Burla
Sambalpur, Orissa, India

H.S. Behera
Depart. of Comp. Sci. & Engineering
V. S .S. University of Technology, Burla
Sambalpur, Orissa, India

ABSTRACT
Since it was first proposed, it is amazing to notice how K-Means algorithm has survive over the years. It has been one among the well known algorithms for data clustering in the field of data mining. Day in and day out new algorithms are evolving for data clustering purposes but none can be as fast and accurate as the K-Means algorithm. But in spite of its huge speed, accuracy and simplicity K-Means has suffered from some of its own problem. Such as, the exact number of cluster is not known prior to clustering. The other thing that is causing problem is that it is quite sensitive to initial centroids. Not just that, K-Means fails to give optimum result when it comes to clustering high dimensional data set because its complexity tends to make things more complicated when more number of dimensions are added. In Data Mining this problem is known as "Curse of High Dimensionality". Here in our paper we proposed a new Modified K-Means algorithm that will overcome the problem faced by the standard K-Means algorithm. We proposed the use of Kohonen Self Organizing Map (KSOM) so as to visualize exact number of clusters before clustering and genetic algorithm is applied for initialization. The Kohonen Self Organizing Map (KSOM) with Modified K-Means algorithm is tested on an iris data set and its performance is compared with other clustering algorithm and is found out to be more accurate, with less number of classification and quantization errors and can be applied even for high dimensional dataset.

Keywords

K-Means, Kohonen Self Organizing Map, Genetic algorithm, curse of Dimensionality, classification error

1. INTRODUCTION
Data mining is the process of extracting useful information from a collection of data's in a large database. One can view cluster analysis as one of the major task for the process of data mining to be successful. It is used in many applications such as pattern recognition, medical purpose, web documentation, business purposes, scientific purposes and so on. Clustering can be defined as the process of organizing objects into groups such that the objects in a group will be similar to one another and different from the objects in other groups. A good clustering will produce a very high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and also its ability to discover some or all of the hidden patterns.

K-means is among the commonly used partitioning based clustering method that tries to find a specified number of clusters (k), represented by their centroids, by minimizing the sum of square error function. It is very simple and fast but is very sensitive to initial positions of cluster centers. Dash et al[1] found out that the complexity of original K-means algorithm is very high, especially for large data sets because the distance calculation increases exponentially with increase in dimensions. Usually only a small number of dimensions are relevant to certain clusters; the irrelevant one may produce noise and mask the real clusters to be discovered. Moreover when dimensionality increases, data usually become increasingly sparse, this can affect the quality of clustering. Hence, attribute reduction or dimensionality reduction is is an essential task for dataset having many attributes. Dimensionality reduction can be defined as transforming of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data. There are of two types

1. Feature Selection which aims at finding subset of the most representative features according to some objective function in discrete space.

2. Feature Extraction/ Feature Reduction algorithms aim to extract features by projecting the original high dimensional data into a lower-dimensional space through algebraic transformations. It finds the optimal solution of a problem in a continuous space, but with computational complexity. PCA is among the commonly used feature reduction method in terms of minimizing the reconstruction error.

Traditional K-means algorithm does not work well for high dimensional data and results may not be accurate most of the time due to noise and outliers associated with original data. Also the computational complexity increases rapidly as the dimension increases. Moreover the exact number of clusters cannot be determined and that it is very sensitive to initial centroids. Hence to improve the performance we proposed KSOM with Modified K-Means algorithm, basically SOM for dimension reduction and determining the number of clusters followed by GA for initialization of the enhanced K-Means. It is found out that the approach gives better accuracy and better performance in terms of speed. Below we gave a brief description of the proposed algorithms.

2. RELATED WORKS
There have been many works done on improving the performance and efficiency of k-means clustering.. A hybridized K-Means clustering approach for high dimensional data set was proposed by Dash, et al [1] where PCA was used for dimensional reduction and for finding the initial centroids a new method is employed that is by finding the mean of all the data sets divided in to k different sets in ascending order. Hs Behera et al [2] proposed another paper an improved hybridized k-means clustering algorithm (IHKMCA) for high

dimensional dataset making used of Canonical Variate analysis and Genetic Algorithm or initialization of the algorithms. Juha Vesanto and Esa Alhoniemi [3] proposed the used of Self Organizing Map for clustering whereby Self Organizing Map(SOM) was used for clustering purpose .J. Vesanto [4] in his another work proposed the SOM based data visualization to visualize the data's using SOM.

M.N.M and Ehsan Moheb [5] proposed the hybridized self organizing map for overlapping clusters. Geoff Bohling [6] give a brief idea on dimensionality reduction and the various technique that can be applied for dimensional reduction. For improving the performance of K-Means clustering M Yedla et al[7] proposed an enhanced K-Means algorithm with improved initial center by the distance from the origin. Fahim A M et al [8].proposed an efficient method for assigning data points to clusters. Zhang Chen et al [9] proposed the initial centroids algorithm based on K-Means that have avoided alternate randomness of initial centroids. Bashar Al Shboul et.al [10] proposed an efficient way of initializing K-Means clustering by using Genetic algorithm thus there by solve the problem of randomly initializing the centroids.

3. MATERIALS AND METHODS

3.1 K-Means Algorithm

One of the simplest and widely used partitioning based, nonhierarchical clustering methods is the K-Means. For any given set of numeric dataset X and an integer number k, the K-means algorithm searches for a partition of X into k clusters that minimizes the within groups sum of squared errors. The K-means algorithm starts by initializing the k cluster centers. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers. The steps of the K-means algorithm are written below:

1. Initialization: choose randomly K input vectors (data points) to initialize the clusters.
2. Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.
3. Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster.
4. Stopping rule: repeat steps 2 and 3 until no more change in the value of the means

3.2 Principal Component Analysis

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The main objectives of PCA are:

1. Identify new meaningful underlying variables;
2. To reduce the dimensionality of the data set.

The mathematical background lies in covariance matrix and "Eigen analysis": In PCA a dataset is normalized by subtracting the mean attribute value from all the attributes in a particular dimension. Then covariance of the normalized matrix is calculated . Then eigen vector and eigen value of the covariance matrix is calculated. The eigenvector associated with the largest Eigen value is used to determine first principal component. The eigenvector associated with the second largest Eigen value helps in determining the second principal component. In here we used the second objective, in that case the covariance matrix of the data set is defined as follows:

$$\text{Cov}(x,y)= \frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{(n-1)}$$

where $\bar{x}$ is the mean of the data (n equals the number of objects in the data set). Principal Component Analysis (PCA) is based on the projection of correlated high-dimensional data onto a hyperplane. This mapping uses only the first few q nonzero eigenvalues and the corresponding eigenvectors of the Covariance matrix. The matrix formed by taking the first eigen vectors is called feature vector. Then the feature vector is applied to normalized data set to find the reduced data set.

3.3 Self Organizing Map

Kohonen Self Organizing Feature Maps, or SOMs provide a way of representing multidimensional data in much lower dimensional spaces - usually one or two dimensions. This process, of reducing the dimensionality of vectors, is essentially a data compression technique known as vector quantization. In addition, the Kohonen technique creates a network that stores information in such a way that any topological relationships within the training set are maintained. One of the most interesting aspects of SOMs is that they learn to classify data without any external supervision whatsoever. It consists of neurons or map units, each having a location in a continuous multi- dimensional measurement space as well as in a discrete two dimensional data collection is repeatedly presented to the SOM until a topology preserving mapping from the multi dimensional measurement space into the two dimensional output space is obtained. This dimensionality reduction property of the SOM makes it especially suitable for data visualization.

There are 'm' cluster unit, arranged in a one or two dimensional array and the input signals are n-tuples. The weight vector for a cluster unit is the exemplar of the input patterns associated with that cluster. In self organizing process, the cluster unit whose weight vector matches the input pattern closely is selected as the winner. The winning and the neighboring units update their weights.

$$c = \min_i \|X - m_i\|$$

In which m_i is the location of the ith map unit in the measurement space and c is the index of the winner map unit in the output grid of SOM.
After the winner search, the locations of the map units in the measurement space are updated according to the rule:

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$$

In which $0<\alpha(t)<1$ is a learning rate factor and $h_{ci}(t)$ is usually the Gaussian neighborhood function.

$$h_{ci}(t) = \exp\left(-\frac{\|r_c - r_i\|}{2\sigma^2(t)}\right)$$

Where r_c is the location of the winner unit and r_i is the location of the ith map unit in the discrete output grid of SOM. The learning rate factor $\alpha(t)$ and the radius $\sigma(t)$ are monotonically decreasing functions of time t.

As mention earlier the original K-Means algorithm does not work well for high dimensions. And we have seen some of its weaknesses, such as sensitive to initialization, unknown number of clusters needed, and complexity problem. So to overcome its entire problem we proposed the Kohonen K-Mean. Where we basically apply KSOM on the dataset for reducing the dimension keeping intact the topological structure of the data. The KSOM, not only it reduce the dimension but it gives us a clear confirmation on the number of clusters. To the resulting reduced data set we then applied the GA for obtaining the initial centroid and finally the data's are group into cluster using the modified K-Means algorithm.

3.4 Genetic Algorithm

It is an optimization algorithm and can be used in any algorithm to optimize the result. The algorithm begins by creating a random initial population. The algorithm then creates a sequence of new populations. At each step, the algorithm uses the individuals in the current generation to create the next population. To create the new population, the algorithm performs the following steps:

a. Stores each member of the current population by computing its fitness value.

b. Scales the raw fitness scores to convert them into a more usable range of values.

c. Selects members, called parents, based on their fitness.

d. Some of the individuals in the current population that have lower fitness are chosen and are passed to the next population.

e. Produces children from the parents(members having best fitness value). Children are produced either by making random changes to a single parent—*mutation*—or by combining the vector entries of a pair of parents—*crossover*.

f. Replaces the current population with the children to form the next generation.

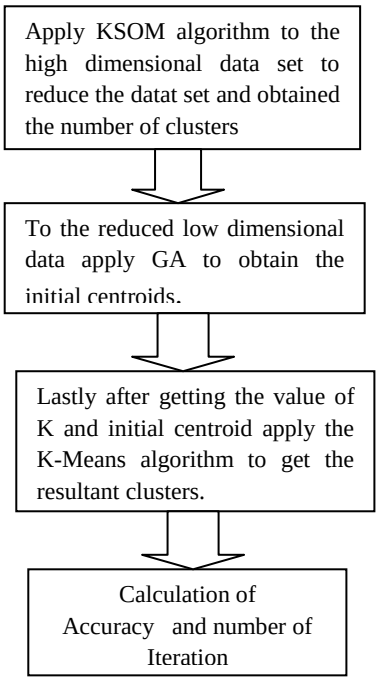

Fig 1: Block diagram of KSOM

4. PROPOSED ALGORITHM

Input: X = {d1, d2,……..,dn} // set of n data items

Step1. Each node's weights are initialized.

Step2 A vector is chosen at random from the set of training data and presented to the lattice.

Step3. Every node is examined to calculate which one's weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU). Euclidean Distance is used to find similarity

$$D_{ij} = \sqrt{\sum_{k=1}^{N}(x_{ik} - x_{jk})^2}$$

Step 4. The radius of the neighborhood of the BMU is now calculated. This is a value that starts large, typically set to the 'radius' of the lattice, but diminishes each time-step. Any nodes found within this radius are deemed to be inside the BMU's neighborhood.

Step 5. Each neighboring node's (the nodes found in step 4) weights are adjusted to make them more like the input vector. The closer a node is to the BMU; the more its weights get altered.

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$$

Step 6. Repeat step 2 for N iterations.

4.1 Genetic Algorithm

```
Input: A resulting reduced data set W from KSOM
Output: a set of K variables  for initial centroids
      t=0;
      Initialize P(t);
      Evaluate P(t);
      While not (termination condition)
       Begin
       t=t+1;
      Select P(t) from P(t-1);
      Recombine pairs in P(t);
      Mutate P(t);
      Evaluate P(t);
      End
```

4.2 Modified K-Means Algorithm

Input:
W //set of n data points.
K // number of desired clusters
Output: a set of K clusters.
Steps:
K initial centroids from genetic algorithm.
Iterative process:
Assign each point ai, to the cluster which has the closest centroid. Calculate the new mean for each cluster UNTIL the convergence criteria is met.

5. EXPERIMENTAL ANALYSIS

The experimental analysis is performed on an iris data set which we can get from UCI Repository of Machine Learning Databases. The data set contains 5 dimensions of three types of flower species setosa,virsicolor and virginica. Based on the length and width of sepal and petal we are to cluster these different flower species.

Step1:

To start with the first stage is using SOM on the iris dataset mainly for two purposes. One for dimensional reduction and the other to visualize and analyze the exact number of cluster needed. On applying KSOM we found out the quantization error and the topographic error to be 0.393 and 0.013 which gives us the clear cut idea of the original data and how much the transformation from high dimension to low dimension has affected the data's. The other important aspect that SOM got is that it not only does reduce the dimension of the data set but it also group together data's of similar properties close to one another. This unique property of SOM is what makes it so powerful in terms of defining the number of clusters.

From the Figure.2 below the U-Matrix and component planes and labels are shown. Clearly some of the conclusions that can be drawn from the fig below are:

There are essentially two clusters

1. PetalL and PetalW are highly correlated
2. SepalL is somewhat correlated to PetalL and PetalW
3. One cluster corresponds to the Setosa species and exhibits small petals and short but wide sepals
4. The other cluster corresponds to Virginica and Versicolor
5. Such that Versicolor has smaller leaves (both sepal and petal) than Virginica
6. Inside both clusters, SepalL and SepalW are highly correlated

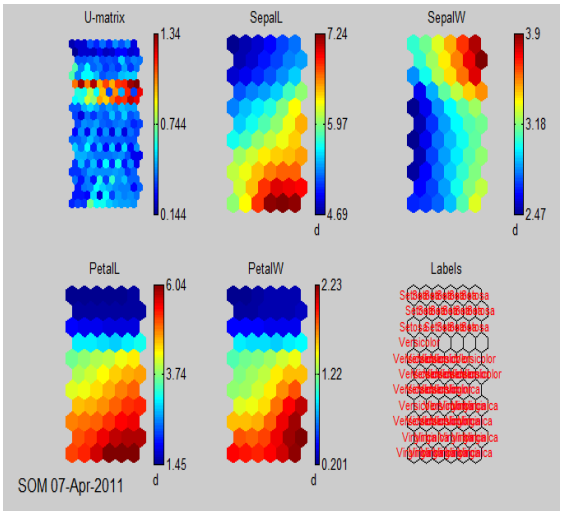

Fig 2: Visual inspection of the map

Now from the above U matrix the distance information can be extracted and projection can be done to visualize the inter relation of data's and confirms the number of clusters. The projection given in fig 2 below confirms the existence of two clusters seem to divide the Virginica flowers into two classes, the difference being in the size of sepal leaves. From the third figure we can get a detail description on the various things we have observed before. Original data points are in the upper triangle, map prototype values on the lower triangle, and histograms on the diagonal: black for the data set and red for the map prototype values .
This distributions of single and pairs of variables both in the data and in the map shows a lot of information. For example there are two clusters: 'Setosa' (blue, dark green) and Virginica','Versicolor' (light green, yellow, reds).

- The PetalL and PetalW have a high linear correlation (see subplots 4,3 and 3,4)

- SepalL is correlated (at least in the bigger cluster) with PetalL and PetalW (in subplots 1,3; 1,4; 3,1 and 4,1)

- SepalL and SepalW have a clear linear correlation, but it is slightly different for the two main clusters (ii subplots 2,1 and 1,2)

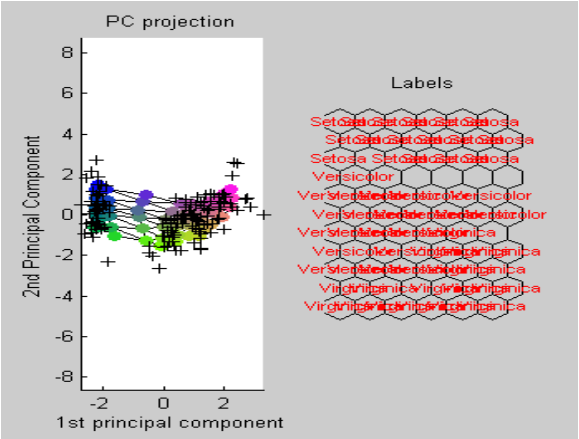

Fig 3: Principal Component Projection of map

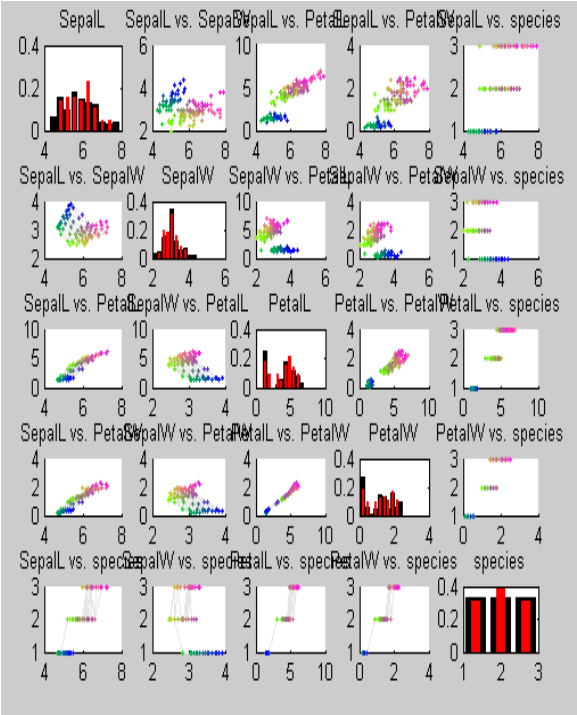

Fig 4: Confirmation of number of clusters

The above figure-4 confirmed the existence of two clusters.

Step 2:

After dimension reduction and determining the K value for K-Means clustering the initialization is done using Genetic Algorithm (GA) which gives the best fitted value compared to all the values as initial centroid.

Step 3:

Lastly but not the least from the K value obtained from step 1 and initial centroid outputted by step 2 the reduced dataset is clustered using the modified K-Means algorithm.

The results obtained are listed in the table below and the clusters are shown in the fig below. The result is compared with some of the well known high dimensional clustering algorithm and found out to have better performance in terms of accuracy and speed.

Table 1. comparison of Kohonen K-Means with PCA and Sammon mapping

Algorithms	Quantization Error	Number of Iteration	Accuracy in %
Kohonen K-Means	0.393	7	91.25
PCA	0.77	8	78.11
Sammon mapping	0.73	8	81.37

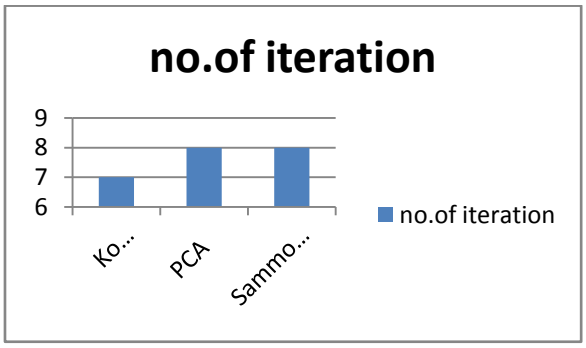

Fig 5: Comparison of Time complexity

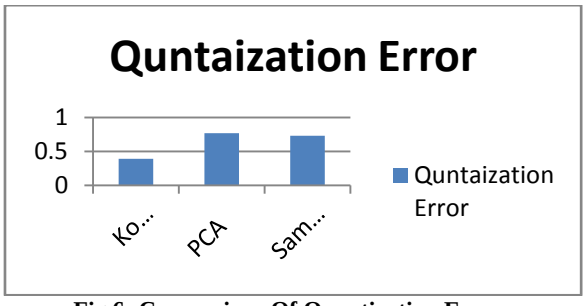

Fig 6: Comparison Of Quantization Error

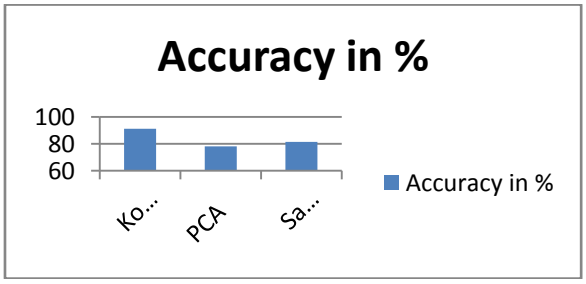

Fig 7: Comparison of Accuracy

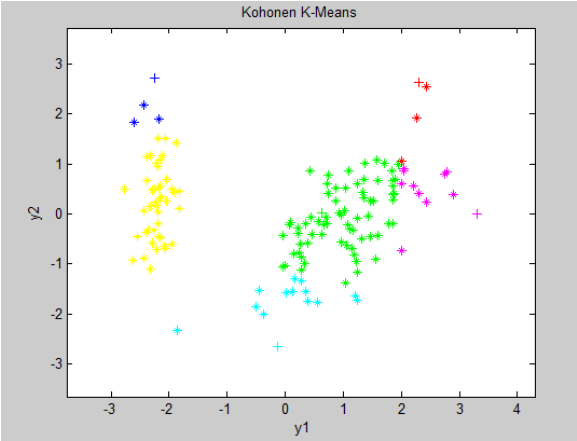

Fig -8 Clusters using Kohonen K-Means

6. CONCLUSION

In this paper we have proposed the Kohonen K-Means algorithm using KSOM for dimension reduction and for determining number of clusters and GA is used for optimization of centroid obtained from KSOM with modified K-Means algorithm. Not only we found out that it has better performance but also that it omits many of the problem that standard K-Means algorithm faced such as unknown number of clusters and the sensitivity to initial centroid. Further research can be done to use a more accurate method in finding initial number of centroid and use better optimization technique for good clustering purposes.

7. REFERENCES

[1] Dash, R. et.al , "A Hybridized k-Means Clustering Algorithm for High Dimensional Dataset", International Journal of Engineering, Science and Technology, vol. 2, No. 2, pp.59-66,2010.

[2] Behera, H. S. et al, "An improved hybridized k-means clustering algorithm(IHKMCA) for high dimensional dataset and it's performance analysis" International journal of Computer science & Engineering,Vol-3 no-2,pp 1183-1190,2011.

[3] Vesanto, J. and Alhoniemi, E., "Clustering of the Self-Organizing Map", IEEE Transactions on Neural Networks, Vol. 11, No. 3, May 2000, pp. 586-600.

[4] Vesanto J., "SOM-based data visualization methods", Intell, Data Analysis, vol. 3, No. 2, pp. 111-126, 1999.

[5] M.N.M and Moheb, E., "Hybrid Self Organizing Map for Overlapping Clusters", International Journal of Signal Processing, Image Processing and Pattern Recognition,pp-11-20.

[6] Bohling, J., "Dimension Reduction And Cluster Analysis", EECS 833, 6 March 2006.

[7] Yedla, M. et al, "Enhancing K means algorithm with improved initial center", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , pp- 121-125,2010.

[8] Fahim A. M., et al, "An efficient k-means with good initial starting points", Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol. 2, No. 19, pp. 47-57,2009.

[9] Zhang, C., Xia, S., et al, "K-means Clustering Algorithm with Improved Initial Center," Second International Workshop on Knowledge Discovery and Data Mining, wkdd, pp.790-792,2009.

[10] Bashar Al Shboul et.al "Initializing K-Means Clustering Algorithm by using Genetic Algorithm" , World Academy of Science, Engineering and Technology 54 2009.