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

# Clustering Gene Expression Data using Quad Tree based Expectation Maximization Approach

Leela Rani.P
Assistant Professor
Sri Venkateswara College of Engineering

Rajalakshmi.P
Assistant Professor
Sri Venkateswara College of Engineering

## ABSTRACT
In molecular biology, micro arrays are employed in monitoring the expression levels of genes simultaneously. Arrays are used in the domains of gene expression, genome mapping, toxicity, pathogen identification and other biological applications. Clustering is a useful technique for grouping gene expression data. In clustering, similar gene expression data will be grouped together for identifying relationships between the genes. Clustering of gene expression data is a useful tool for identifying co-expressed genes and biologically relevant grouping of genes, which is an important research area in Bioinformatics. In this paper, a Quad Tree based Expectation Maximization (EM) algorithm has been applied for clustering gene expression data. Quad Tree is used to initialize the cluster centroids. With these centroids, EM is used to group the data efficiently. Expectation Maximization is used to compute maximum likelihood estimates given incomplete samples. Silhouette refers to a method of interpretation and validation of clusters. This measure provides a representation of how well each object lies within its cluster.Experimental results have shown that Quad Tree based Expectation Maximization algorithm finds compact clusters when compared to K-Means algorithm.

## General Terms
Bioinformatics, gene expression data, Clustering, Partition.

## Keywords
Clustering, Quad Tree, Expectation Maximization Algorithm, K-Means, Silhouette measure, Similarity.

## 1. INTRODUCTION
A micro array [9] is a glass slide on which DNA molecules are fixed on an ordered manner at specific locations called spots. The spots are printed on the glass slide by photolithography. Using micro arrays, the expression of many genes in a single reaction can be assessed quickly. This enables the researchers to explore the genetic causes of abnormalities occurring in the functioning of the human body. Data mining tasks like Clustering and Classification are used to extract useful knowledge from micro array data[9]. Many techniques can be applied to analyze micro array data, which can be grouped in four categories: classification, feature selection, clustering and association rules.

Micro array data sets are normally very large. It will be better if the dataset is condensed into those genes that are best distinguished between the two classes - normal vs. diseased. The reduction produces a list of genes whose expression is likely to change known as differentially expressed genes. Identification of differential gene expression is the first task in micro array analysis. There are two methods for micro array data analysis - clustering and classification.

Clustering [5] is the unsupervised approach to classify data into groups of genes. The goal of clustering in micro array technology is to group genes or experiments into clusters according to a similarity measure. Genes that share a similar expression pattern under various conditions may imply co-regulations or relations in functional pathways. They can be grouped in two categories -partitioning and hierarchical algorithms. Classification [5] is a supervised learning. Given a set of pre-classified examples, the classifier learns to assign an unseen test case to one of the classes. The most used classification algorithms used in the micro array[9] analysis belong to four categories: decision tree, Bayesian classifiers, neural networks and support vector machines.

## 2. RELATED WORK
Clustering is an important tool in gene expression data analysis. Different clustering methods usually produce different solutions. K-Means [1] is a prototype-based partitional clustering method. The number of clusters to be identified is specified by the user. The basic idea is to choose random cluster centers for each cluster. Each remaining point is considered and the similarity with all cluster centers is calculated using a distance measure. The point is assigned to the nearest cluster. When this process is done, a new center will be calculated for each cluster using the points in it. For each cluster, the mean value will be calculated for all the points in that cluster and it will be set as the new center. The process must start over again with the new *k* centroids. The *k* centroids change the locations step by step until no more changes are done. The demerit of K-Means[1] is that the quality of clusters depends on the initially selected centroids. Randomness in centroid selection does not result in compact clusters.
There is no solution to find the optimal number of clusters for any given data set in K-Means[7]. A simple approach is to compare the results with different k clusters and choose the best one according to a given criterion.

The initial cluster centers are found using a quad tree based algorithm [2]. A quad tree is a tree data structure in which each internal node has exactly four children. Quad trees are most often used to partition a two dimensional space by subdividing it into four quadrants. The regions may be square or rectangular. The cluster centers, thus found, serve as input to the clustering algorithms. All forms of Quad trees[2] share some common features:

1 They decompose space into adaptable cells / buckets.

2 Each cell has a maximum capacity. When maximum capacity is reached, the bucket splits.

3 The tree directory follows the spatial decomposition of the Quad tree.

Expectation Maximization is a type of model based clustering method. It attempts to optimize the fit [8] between the given data and some mathematical model. This method is based on the assumption that the data are generated by a mixture of underlying probability distributions. The EM algorithm is an extension of the K-Means algorithm [1]. Each cluster can be represented mathematically by a parametric probability distribution. The entire data is a mixture of these distributions. The Expectation Maximization algorithm[4] is a popular iterative refinement algorithm that can be used for finding parameter estimates. It can be viewed as an extension of the K-Means paradigm, which assigns an object to the cluster with which it is most similar based on the mean of the cluster. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. Hence in addition to working towards minimizing the Euclidean distance, it also takes into account the probability of membership of each data point to each cluster.

# 3. OVERVIEW OF THE PAPER

This paper is organized as follows. Section 4 describes the proposed work. Section 5 deals with the experimental results. Section 6 deals with conclusion. Section 7 deals with future work.

# 4. PROPOSED WORK

## 4.1 Overview

In gene expression clustering, a micro array [9] gene expression measurement for thousands of genes under varying conditions will be considered. The goal is to group the observed expression vectors into distinct clusters of related genes. The existing approaches cluster data sets using K-Means and Expectation Maximization [7] Algorithm. The basic idea in K-Means is to choose random cluster centers for each cluster. The demerit of K-Means is that the quality of clusters depends on the initially selected centroids. In the proposed work, the cluster centers found out using Quad Tree and provided as input to the Expectation Maximization algorithm.

## 4.2 K-Means algorithm

The main objective in cluster analysis is to group objects that are similar in one cluster and separate objects that are dissimilar by assigning them to different clusters. K-Means clustering classifies object to a pre-defined number of clusters, which is given by the user (assume *K* clusters). The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other [1]. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids. The Euclidean distance is calculated between two multi-dimensional data points *X* and *Y*. The K-Means [7] method aims to minimize the sum of squared distances between all points and the cluster center. This procedure consists of the following steps, as described below.

K-Means clustering algorithm [3]
**Input**: D = {$d_1$, $d_2$, $d_3$, ..., $d_n$} // Set of n data points.
K - Number of desired clusters
**Output**: A set of K clusters.

1 Arbitrarily choose *k* data points from *D* as initial centroids
2 Repeat
Assign each point $d_i$ to the cluster which has the closest Centroid
Calculate the new mean for each cluster
Until convergence criteria is met.

The demerit of K-Means is that the quality of clusters depends on the initially selected centroids. Randomness in centroid selection does not result in compact clusters.

## 4.3 Quad Tree with Expectation Maximization algorithm

Quad tree is used to find the cluster centers. The parameters used in the algorithm are described below.
**LOW** -User defined threshold for minimum number of data points in a sub bucket.

**HIGH**-User defined threshold for maximum number of data points in a sub bucket.

**White leaf bucket**- A sub bucket having less than LOW number of data points of the parent bucket.

**Black leaf bucket** - A sub bucket having more than HIGH number of data points of the parent bucket.

The initialization step of the Quad tree algorithm is given below [2].

1 Initialize LOW & HIGH values.
2 Fix a parameter along the X axis.
3 Fix another parameter along the Y axis.
4 For each cluster, repeat the following steps
4.1 Find the minimum and maximum x and y co–ordinates.
4.2 Calculate midpoint using the values obtained in previous step.
4.3 Partition the spatial area into four sub regions based on the midpoint.
4.4 Plot the points and label partitions as white leaf buckets or black leaf buckets.
4.5 The white leaf buckets are left untouched.
4.6 The centers of each black leaf bucket is calculated.
4.7 The mean of all the center points obtained in the previous step is calculated.
4.8 The computed mean gives the centroid for that cluster.

After implementing Quad tree on the data set, the resulting centroids are fed as input to the Expectation Maximization [4] algorithm. The steps are briefed below.

1 Input the centroids obtained using the quad tree algorithm as the initial cluster centers to the next step.
2 Compute distance between each data point and each centroid using distance formula

$$\text{Distance} = \left|(x_2 - x_1)\right| + \left|(y_2 - y_1)\right|$$

3 Assign weights for each combination of data point and cluster based on the probability of membership of a data point to a particular cluster.

4 Repeat

4.1 (re) assign each data point to the cluster with which it has highest weight / highest probability of association.

4.2 If a data point belongs to more than one cluster with the same probability, then (re)assign the data point to the cluster based on minimum distance.

4.3 Update the cluster means for every iteration

5 Until clustering converges.

# 5. EXPERIMENTAL RESULTS

## 5.1 Data sets

The data sets that are considered for clustering in this paper are serum data and yeast cell data.

In the Serum data [10], the genes are listed according to their cluster order along with their Genbank Accession number and Clone IDs. Gene names with the SID prefix are not sequence verified. The expression changes are given as the ratio of the expression level at the given time-point to the expression level in serum-starved fibroblasts. All ratios are normalized to time zero. An updated database of these results is available at http://genome-www.stanford.edu/serum

The yeast gene expression data [10] is available at http://rana.stanford.edu/software/demo.txt

## 5.2 Silhouette Coefficient

Silhouette coefficient [6] combines cohesion and separation.The silhouette value for each point is a measure of how similar that point is to other points in its own cluster compared to points in other clusters. It ranges from -1 to +1.

Silhouette coefficient[6] is computed following the three steps.

1 For the $i^{th}$ object, calculate its average distance to all other objects in its cluster. This value is denoted as $a_i$.

2 For the $i^{th}$ object and any cluster not containing the object, calculate the object's average distance to all the objects in the given cluster. Find the minimum of all values in all the clusters. This value is denoted as $b_i$.

3 For the $i^{th}$ object, the silhouette coefficient is computed using the formula

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

A negative value for Silhouette coefficient is undesirable because this means that $a_i$ is greater than $b_i$. The Silhouette coefficient should be positive ($a_i<b_i$) and $a_i$ should be as close to 0 as possible. The coefficient assumes the maximum value of 1 when $a_i=0$. The average coefficient is computed by taking tha average of the Silhouette coefficients of points belonging to the cluster.

K-Means algorithm is implemented on the above mentioned data sets and the cluster quality is measured using the silhouette's coefficient[6].

Quad tree algorithm is invoked on the data sets to find out the initial cluster centers. The initial centroids are given as input to the Expectation Maximization algorithm. After the Expectation Maximization algorithm converges, the final number of clusters is derived. The cluster quality is measured using the Silhouette's coefficient.

## 5.3 Performance Evaluation on Serum data

K-Means and Quad tree based Expectation Maximization algorithms are applied on Serum data set. The initial number of clusters assumed is 10.

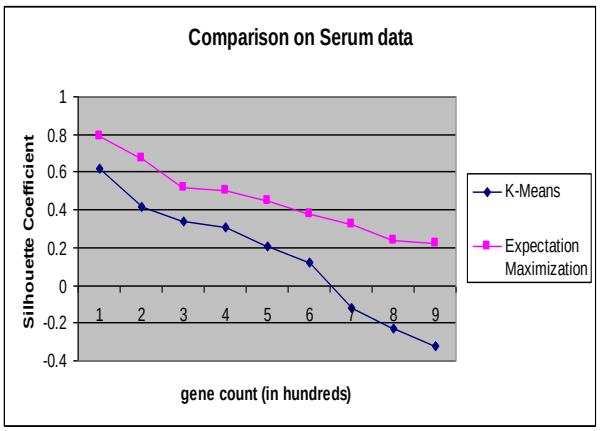


**Figure 1: Performance Evaluation on Serum data**

Figure 1 shows that the Silhouette's coefficient is better in Quad tree based Expectation Maximization algorithm than in K-Means.

## 5.4 Performance Evaluation on Yeast cell data

K-Means and Quad tree based Expectation Maximization algorithms are applied on Yeast cell data set. The initial number of clusters assumed is 30.

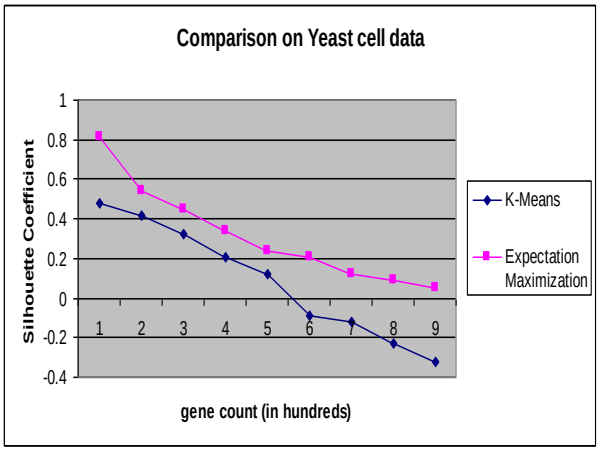


**Figure 2: Performance Evaluation on Yeast data**

Figure 2 shows that the Silhouette's coefficient is better in Quad tree based Expectation Maximization algorithm than in K-Means.

## 6. CONCLUSION

The demerit of K-Means is that the quality of clusters depends on the initially selected centroids. This demerit of K-Means was overcome in this paper by using Quad tree for finding the initial centers and using those centers with Expectation Maximization algorithm. Cluster quality is measured using Silhouette coefficient. The overall cluster quality is superior in the proposed work, when compared to the most popular K-Means algorithm.

Combining the Quad Tree based EM gives a clustering that fits the number of clusters and tries to make them compact. The proposed work also proves to be faster in clustering as opposed to K-Means. K-means does not guarantee convergence. EM guarantees elegant convergence.

The proposed work is implemented on Serum data and Yeast cell data. The quality of the clusters is evaluated by calculating Silhouette coefficient. The results have shown that Quad tree based EM approach yields better quality clusters as compared to K-Means. The performance evaluation done on Serum data set and Yeast cell data set is shown as individual charts in Figure 1 and 2.

## 7. FUTURE WORK

This proposed approach can be extended to include a Hyper Quad tree based EM clustering. The Hyper Quad tree is used as a replacement to the Quad tree approach, to obtain precise cluster centers. Hyper Quad trees are expected to give better cluster centers than the Quad Tree.

## 8. REFERENCES

[1] Bashar Al-Shboul and Sung-Hyon Myaeng, "Initializing K-Means using Genetic Algorithms", World Academy of Science, Engineering and Technology 54, 2009.

[2] P.S. Bishnu and V. Bhattacherjee, "A New Initialization method for K-Means using Quad Tree," Proc of National. conf. on Methods and Models in Computing, JNU, New Delhi, pp. 73-81, 2008.

[3] T.Chandrasekhar, K.Thangavel and E.Elayaraja**, "**Performance Analysis of Enhanced Clustering Algorithm for Gene Expression data", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[4] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B,39(1):1–38.

[5] J. Han and M. Kamber , Data mining Concepts and techniques, 2nd edition, Morgan Kaufmann Publishers, pp. 401-404, 2007.

[6] Moh'd Belal Al- Zoubi and Mohammad al Rawi, "An Efficient Approach for Computing Silhouette Coefficients". Journal of Computer Science 4 (3): 252-255, 2008.

[7] G.Nathiya, S.C.Punitha, M.Punithavalli, "An Analytical Study on Behavior of Clusters Using K-Means, EM and K* Means Algorithm", (IJCSIS) International Journal of Computer Science and Information Security,Vol. 7, No 3, 2010.

[8] Osama Abu Abbas, Computer Science Department, Yarmouk University, Jordan, "Comparisons between data clustering algorithms" The international Arab Journal of Information Technology,Vol.5,No.3,July 2008.

[9] Sunnyvale, Schena M," Microarray biochip technology"., CA: Eaton Publishing; 2000.

[10] Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, and Patrick O. Brown, "The Transcriptional Program in the Response of Human Fibroblasts to Serum", www.sciencemag.org, Science Vol.283,1 January 1999.