# A SURVEY ON THE CLUSTERING METHODS FOR THE IDENTIFICATION OF COEXPRESSED GENES

Navya S Bhaskar[1], Aswathy Devi T.[2]

[1] M.Tech Signal Processing Student, Department of ECE, LBSITW, Kerala, India
[2] Assistant Professor, Department of ECE, LBSITW, Kerala, India

## ABSTRACT

*Bioinformatics is an interdisciplinary area of collecting and analyzing gene codes .Gene clustering is one of the first steps in gene expression analysis. The main aim of clustering is to subdivide a set of data in such a way that similar data fall into same cluster whereas dissimilar items fall in different clusters. The main application is Gene-expression profiling, medical diagnosis – cancers, cardio vascular disorders, genetic disorders, biomedicine, drug discovery and development and tumor classification. This paper is the comparative study of different types of clustering techniques like k-means, fuzzy c means, self organizing map, hierarchical clustering. For the precise clustering of genes it is important to choose optimal number of genes and feature selection from microarray dataset. So dimensionality reduction is necessary in gene clustering. The gene expression data is mainly analyzed in two ways, supervised and unsupervised. Clustering techniques have proven to be helpful to understand gene function, cellular processes, and subtypes of cells. The different methods of measuring similarity are by considering Euclidean distance, pearson correlation and Manhattan distance etc .There is no one-size-fits all solution to clustering. It is ambiguous of how good clusters look like. So validity of clusters should be done in the sense of biological significance.*

**Keyword**: - *Clustering, Euclidean distance, Co expression, Dendograms*

## 1. INTRODUCTION

[1,14]Due to invention of DNA (Deoxyribonucleic acid) microarray technology, it has become feasible to examine the expression level of thousands of genes at a time during their different ongoing biological processes and across collection of related samples. Different application areas of microarray technology are gene expression profiling, medical diagnosis, bio-medicine. Usually, during the biological experiment, and at different time points, gene expression values are measured. A microarray gene expression data structure is defined as 2D matrix A [fij] of size c x t, where c represents a gene and t represents a time point. Each element fij tells about the expression level of ith gene at the jth time point. To depict the set of genes exhibiting similar expression profile, clustering or unsupervised learning is in general used. Clustering, also termed as unsupervised learning, is the procedure of grouping the data items into different partitions or clusters in such a way that data items belonging to same group are similar to each other according to some criteria of similarity and dissimilar to each other according to same criteria [1]. In supervised classification, actual class labels of some data points are available. The main problem of supervised classification is to generate the labeled data, which is both time consuming and expensive. Traditional approaches of genomic research concentrates on local examination and collection of data for a single gene. However, use of microarray technology enables monitoring of expression levels of tens of thousands of genes simultaneously. There are mainly two different types of microarray experiments psychological changes.

During microarray experiment, a large number DNA sequences (genes, cDNA clones, or co-expressed sequence tags) are judged under multiple conditions. Examples of conditions are time series during a biological process (e.g., the Yeast cell cycle) or a collection of different tissue samples (e.g., normal versus cancerous tissues). In general no distinction is made among DNA sequences; which are uniformly termed as ‘‘genes’’. Similarly all kinds of experimental conditions are termed as ‘‘samples’’. [6]Microarray technology has been successfully applied for solving problems from many areas like medical diagnosis, bio-medicine, gene expression profiling, etc. In general, the gene expression values during a biological experiment are determined at different time points. Some noise and missing values can be present in the original gene expression matrix determined from the scanning process. There may be some systematic variations arising from the experimental procedure. Data pre-processing is a necessary step before application of any clustering technique. After this data normalization is done. Thereafter any clustering technique can be applied on the gene expression data.[4,5]

## 2. BRIEFING OF THE DIFFERENT CLUSTERING TECHNIQUES FOR THE IDENTIFICATION OF CO EXPRESSED GENES.

### 2.1 K mean clustering

[12]The k-Means algorithm is one of the simplest unsupervised learning algorithms that answer the well known clustering problem. The procedure follows a simple and calm method to classify a given data set through a certain number of cluster, assume k clusters) static a priori. The k-Means algorithm can be run multiple times to decrease the complexity of grouping data. The k-Means is a simple algorithm that has been modified to many problem areas and it is a noble candidate to work for a randomly generated data points. Cluster analysis organizes data (a collection of patterns, each design could be a direction measurements) by abstracting and an underlying structure. The algorithm is composed of the following steps:

Step 1: Residence k points into the space represented by the objects that are being clustered. These points represent initial group centroid.

Step 2: Allocate each item to the group that has the closest centroid.

Step 3: When all objects have been given, recalculate the positions of the k centroid.

Step 4: Repeat Steps 2 and 3 until the centroid no longer move.

The algorithm is also significantly sensitive to the initial randomly selected cluster centers. This is proved by more than a few times in this recent as well as in the past research; recurring problem has to do with the initialization of the algorithm. The k-Means is a simple algorithm. The K-means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise [7]

### 2.2 Hierarchical clustering

[10] [11] The hierarchical methods group training data into a tree of clusters. This tree structure called dendrogram (Fig. 1). It represents a sequence of nested cluster which constructed top-down or bottom-up. The root of the tree represents one cluster, containing all data points, while at the leaves of the tree, there are n clusters, each containing one data point. By cutting the tree at a desired level, a clustering of the data points into disjoint groups is obtained. Hierarchical clustering algorithms divide into two categories: Agglomerative and Divisive. Agglomerative clustering executes in a bottom–top fashion, which initially treats each data point as a singleton cluster and then successively merges clusters until all points have been merged into a single remaining cluster. Divisive clustering, on the other hand, initially treats all the data points in one cluster and then split them gradually until the desired number of clusters is obtained. To be specific, two major steps are in order. The first one is to choose a suitable cluster to split and the second one is to determine how to split the selected cluster into two new clusters.

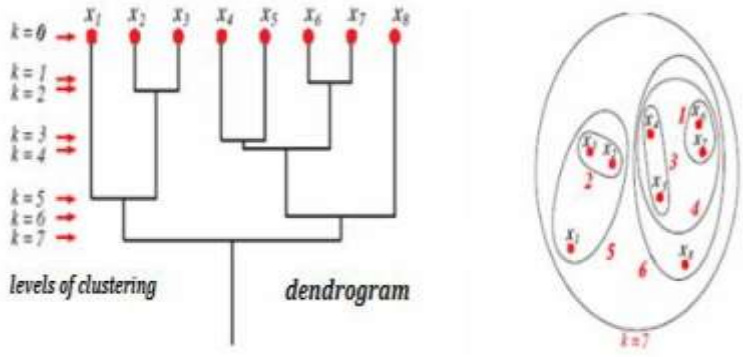


**Figure 1 Dendrogram**

### 2.3 Fuzzy C mean clustering

The FCM employs fuzzy subdividing such that a given data point can belong to some groups with the degree of belongingness stated by membership ratings between 0 and 1. However, FCM still uses a cost function that is to be minimized while trying to partition the data set. The membership matrix U is allowed to have elements with values between 0 and 1. However, the summary of grades of belongingness of a data point to all clusters is always equal to unity. [9]

Step 1: Initialize the membership matrix U with random values between 0 and 1 such that the constraints are satisfied in the equation base.

Step 2: Calculate c fuzzy cluster centers, 1, ,i c i = c , using the final equation .

Step 3: Compute the cost function.

Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Compute a new U …. Go to step 2.

### 2.4Self organizing map

[8]Self organizing Map (SOM) is used for visualization and analysis of high-dimensional datasets. SOM facilitate presentation of high dimensional datasets into lower dimensional ones, usually 1-D, 2-D and 3-D. It is an unsupervised learning algorithm, and does not require a target vector since it learns to classify data without supervision. A SOM is formed from a grid of nodes or units to which the input data are presented. Every node is connected to the input, and there is no connection between the nodes. The Self-Organizing Map (SOM) was developed by Kohonen on the basis of a single layered neural network. SOFMs were developed by observing how neurons work in the brain and in ANNs The firing of neurons impact the firing of other neurons that are near it Neurons that are far apart seem to inhibit each other Neurons seem to have specific non-overlapping tasks The term self-organizing indicates the ability of these NNs to organize the nodes into clusters based on the similarity between them Those nodes that are closer together are more similar than those that are far apart The most common example of a SOFM is the Kohonen Self organizing map It is used extensively in commercial data mining products to perform clustering There is one input layer and one special layer which produces output values that compete Multiple outputs are created and the best one is chosen This extra layer is not technically either a hidden layer or an output layer, so we refer to it here as the competitive layer Nodes in this layer are viewed as a two-dimensional grid of nodes Each input node is connected to each node in this grid[21]

## 3. PERFORMANCE COMPARISON OF CLUSTERING METHODS

[21]This paper also presents the comparative study on the above mention clustering algorithms based on their accuracy and demerits. The outcome of study shows that the clustering algorithm gives better accuracy for k-means and then other algorithm in clustering. Comparison is on table 1 [13]

**Table 1** comparison of different clustering methods.

| Sl no | AUTHOR NAME | ALGORITHM | OUTCOME | DEMERITS |
|---|---|---|---|---|
| 1 | A.Dharmarajan.,T. Velmurugan | K mean | Time complexity is high. | Difficult to compare a quality of cluster |
| 2 | Jacob Goldberger., Tamir Tassa K.Sasirekha,P.Baby | hierarchical | Accurate Result | Times process is slow |
| 3 | Nikhil R. Pal, Kuhu Pal, James M. Keller, James C. Bezdek | Fuzzy c means | High times complexity with best result | Iteration process is expensive |
| 4 | Jyrki Joutsensalo and Antti Miettinen, Tamayo, P | som | High Accuracy. | Computational process is high. |

## 4. CONCLUSIONS

Clustering algorithms are useful for identifying biologically relevant groups of genes and sample clustering techniques are essential in the data mining process to reveal natural structure and identifying pattern in the data sets. From the above context identified that, for microarray clustering techniques k-means algorithm is used widely. The future research direction towards the gene ontology (GO) terms which comes under microarray gene expression data and k-means clustering algorithm is hybridized with other algorithms to achieve quality and accuracy.

## 5. REFERENCES

[1] Mann AK, Kaur n. Survey paper on clustering techniques. Ijsetr. 2013 apr; 2 (4):803–6.
[2] Kaufman, L. and Rousseeuw, P.J. Finding Groups in Data: an Introduction to Cluster Analysis. JohnWiley and Sons, 1990
[3] Brazma, Alvis and Vilo, Jaak. Mini review: Gene expression data analysis. Federation of European Biochemical societies, 480:17–24, June 2000.36
[4] Siedow, J. N. Meeting report: Making sense of microarrays. Genome Biology,2(2):reports 4003.1– 4003.2, 2001.
[5] Derisi, J.l. , Iyer, V.R., and brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, pages 680–686, 1997.
[6] R.Sharan, R.Elkon, R.Shamir, Cluster Analysis and its Applications to Gene Expression Data.
[7] In A.Dharmarajan.,T.Velmurugan k-means algorithm.
[8] Jyrki Joutsensalo and Antti Miettinen,Tamayo, P. And others, interpreting patterns of gene expression with self organizing Maps, pnas 96, p.2907--2912, 1999

[9] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek fuzzy c-means algorithm Springer, Berlin, Heidelberg, 1995.
[10] Jacob Goldberger., Tamir Tassa Hierarchical clustering algorithm 2nd edution springer-verlage 1998.
[11] K.Sasirekha, P.Baby.,Agglomerative algorithm 2nd edution springer-verlage 1998.
[12] Hartigan j.a clustering algorithm wiley,new york Hartigan j.a and Wong m.a(1979) a k-means clustering.
[13] S uma,porkodi,a survey on clusterin techiques of microarray co expressed genes ISSN: 2321-8169 Volume: 4 Issue: 3 335 – 341