

Identification of differentially expressed genes by unsupervised learning method.

P.Venkatesan¹ and Jamal Fathima .J.I²

1.Department of statistics National Institute For Research In Tuberculosis, Indian Council For Medical Research,Chennai,India. Venkaticmr@gmail.com

2.-Research Scholar -Department of statistics National Institute For Research In Tuberculosis, Indian Council For Medical Research,Chennai,India,jf_ayub@yahoo.co.in

ABSTRACT

Microarrays are one of the latest breakthroughs in experimental molecular biology that allow monitoring of gene expression of tens of thousands of genes in parallel. Micro array analysis include many stages. Extracting samples from the cells, getting the gene expression matrix from the raw data, and data normalization which are low level analysis. Cluster analysis for genome-wide expression data from DNA micro array data is described as a high level analysis that uses standard statistical algorithms to arrange genes according to similarity patterns of expression levels. This paper presents a method for the number of clusters using the divisive hierarchical clustering, and k-means clustering of significant genes. The goal of this method is to identify genes that are strongly associated with disease in 12607 genes. Gene filtering is applied to identify the clusters.

k-means shows that about four to seven genes or less than one percent of the genes account for the disease group which are the outliers, more than seventy percent falls as undefined group. The hierarchical clustering dendrogram shows clusters at two levels which shows again less than one percent of the genes are differentially expressed.

1. INTRODUCTION

Cluster analysis is a high level analysis .It is traditionally used in phylogenetic research and has been adopted in micro array analysis.DNA micro array experiment allows us to measure the expression levels of thousands of genes simultaneously under various conditions. The primary objective of supervised clustering (Carr et al 1997,Cho et al 1998,Szabho et al 2002) is to predict the functions of unknown genes by comparing the gene expression patterns of unknown genes. Gene clustering is one of the widely used statistical tools in microarray data analysis. Due to the high dimensionality of the data set; data mining methods are applied to summarize the information for a synthetic interpretation.Sandrine Dudoit and Rebert Gentleman(2002).

Unsupervised learning technique aims in detecting the relation between the tissues or genes and within them(Hastie et al 2002). Thus in cluster analysis similarity between samples or genes have been identified. Unsupervised clustering is otherwise called hierarchical clustering, with samples /gene clustering within clusters (Eisen et al 2000).Grouping of genes having similar expression pattern is finding at most interest in micro array data analysis.

Many clustering procedures have shown success in micro array gene clustering, most of them belong to the family of heuristic clustering algorithm, which are based on the assumption that the whole set of micro array data is a finite mixture of a certain type of distributions with different parameter. The commonly used are hierarchical clustering (Carr et al 1997),k-means(Tavazoie et al1999),model based clustering (Das Gupta and Raftery 1998),Support Vector Machine(SVM)(Brown et al 2000).

2.MATERIAL AND METHODS:

The main goal is to find clusters of samples or clusters of genes such that observations within a cluster are more similar to each other then they are in different clusters. Gene clustering has two main goals: (i) Understanding the gene function. (ii) Discovery of co-regulated genes. This paper concentrates on the second goal. To achieve this we compare the genes having similar expression profiles in a set of conditions among different tissues or different times into clusters. Class discovery depends on the gender or tissue type or by the combination of both. In micro array data analysis class discovery process is dominated by sample characteristics –patients gender or tissue type. In Our method we construct clusters of samples in the data set

by k-means/median applying Euclidean and Pearson distance measure to find the number of clusters. Let x_1, X_2, \dots, X_p be the p-dimensional vector where each X_i is a n-dimension, that is it gives the estimate of i^{th} gene of p^{th} sample. We take the log2 ratio of the pre value where pre value is obtained as the ratio of tumour to normal tissue. The data corresponding to the log intensity ratios of i^{th} gene in j^{th} experiment is defined as :

$x_{ij} = \log(\text{prevalue})$, where each x_i is a p-dimensional vector $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. Expression data are analysed in a matrix form, each row represents a gene and each column a sample. The matrix entries x_{ij} corresponding to the expression values of i^{th} gene in the j^{th} sample where $(i=1,2,\dots,12607 \text{ genes})$ and $j=1,\dots,4$. Sample subsets). many software's are available for performing cluster analysis on the matrix 'x'. To perform analysis the matrix X is transformed to $X^T=Y$, such that $y_{ij} = x_{ji}$. (1)

2.1. GENE FILTERING

As the data size is huge, gene filtering plays a vital role in removing the noise or unexpressed genes. Gene filtering can be done at a threshold of a statistic. An arbitrary threshold based on a value for t-statistic can be applied to filter put a certain percentage of the genes. In addition to identifying the groups of related genes, hierarchical clustering helps in screening out the uninteresting genes. Removal of uninteresting genes is especially important in micro array data where the number of gene is very large.

2.2.K-means clustering:

- (i)'k' specified in advance
- (ii) Select a set of k-points as clusters seed these seeds represent a first guess at the centroids of the k-means.
- (ii)Assign each individual observation to the clusters defined whose centroid is nearest. The centroids are recalculated for the clusters receiving the new object and for the clusters losing the object.
- (iii)Repeat (i)-(ii) until no changes occur in the cluster composition.

K-means clustering does not give ordering of objects within a cluster. The final assignment depends on the initial selection of seed points. As the number of clusters 'k' is changed, the cluster grouping may change. K-means clustering assigns each gene to only one clusters, genes assigned to the same clusters

may not be necessarily have similar expression patterns.

2.3. Hierarchical Clustering

Given a set of data the hierarchical clustering algorithm aims in finding the groups or clusters in the data. There are two general forms of hierarchical clusters may be either agglomerative (bottom-up) or divisive (top-bottom). In agglomerative approach each data point initially forms a cluster and the two closest cluster are merged at each step. In divisive method one large cluster which contains all the data points and splits off a cluster at each step. Cluster analysis depends on some measurement of similarity or dissimilarity between two vectors to be clustered. Commonly used method in cluster analysis (BenDor et al 1999, Tavazoie et al 1999, Brazma and Vilo 2000, Horimoto and Toh 2001) have defined the dissimilarity between two objects, i and j as their (i)Euclidean distance, (ii)Linear correlation co-efficient.. Here we have adopted Divisive method for clustering and the distance measure is the Euclidean distance given by D_{ij}

(i)Euclidean distance: $D_{ij} = \sqrt{\sum_{j=1}^n (x_{ij} - x_{jk})^2}$ (2)

(ii)Distance based on correlation: $R_{ij} = \frac{\sum_j (x_{ii} - \bar{x}_i)(\bar{x}_{ji} - \bar{x}_j)}{\sqrt{\sum_l (x_{il} - \bar{x}_i)^2} \sqrt{\sum_j (\bar{x}_{ji} - \bar{x}_j)^2}}$ (3)

With 'n' data points the hierarchical clustering merges the two closest clusters providing a nested sequence of clusters. The clustering results are displayed in a dendrogram. Nodes or clusters forming lower on the dendrogram are closest together and the upper nodes represent the clusters that are far apart. Each data point beginning as a single cluster as a single cluster, the leaves (the terminating node at the bottom of the dendrogram). Each node represent a data point, while the nodes in the interior point of a cluster represent cluster of more than one data point. The node of the cluster dendrogram represents the entire data point as a single cluster.

Hierarchical clustering has different approaches-single linkage, complete linkage and average linkage (Gorden 1999). In single linkage the distance between any two clusters of points is the smallest distance between any point in cluster one and two. In complete linkage the largest distance between any point between the two clusters is taken as the inter

cluster distance. Average linkage takes the average of all pair wise distances between the points in the first cluster and second cluster (Sokal and Michener 1958). Ward (1963) examines the sum of squared distance from each point to the mean of its cluster and merges the two clusters and is adopted to obtain the clusters in this work.

3. Micro array database:

The database consists of log ratios of gene expression values of breast cancer FBN1 cell expression profiles-cDNA micro array (G4100A)fibrillin1,organism:Homosapiens with ID Ref no 6775502.Data consist of 12607 unique clones .The expression profiling of breast cancer cell lines HCC1954 and MDA-

MB-436 in reference to mammary epithelial cells.

4. Application to micro array data:

Data studied was obtained from NCBI GENE OMNIBUS. The log intensities of 12607 genes in breast cancer data set. Data filtering is carried out by leaf and stem method. We have made a comparative study based on k-means clustering hierarchical clustering

4.1. K-means clustering

By applying k- means clustering on the log ratio intensities one can see that 5 to 12% of the genes accounts for the total variation or differentially expressed.. The results obtained from k-means clustering for k=3, 4, 5 are given in the following table.Distance measure: Euclidean distance.

TABLE

| Cluster number | Cluster size | Gene Percentage in each cluster |
|----------------|--------------|---------------------------------|
| K=3 | | |
| 1 | 2 | 1.3* |
| 2 | 8 | 5.2*** |
| 3 | 145 | 93.5 |
| K=4 | | |
| 1 | 2 | 1.3* |
| 2 | 3 | 1.9** |
| 3 | 15 | 9.7 |
| 4 | 135 | 87.1 |
| K=5 | | |
| 1 | 2 | 1.3* |
| 2 | 3 | 1.9** |
| 3 | 10 | 6.5 |
| 4 | 21 | 13.5 |
| 5 | 119 | 76.8 |

*Gene id - {16990, 8822}

** Gene id - {938, 2859, 18236}

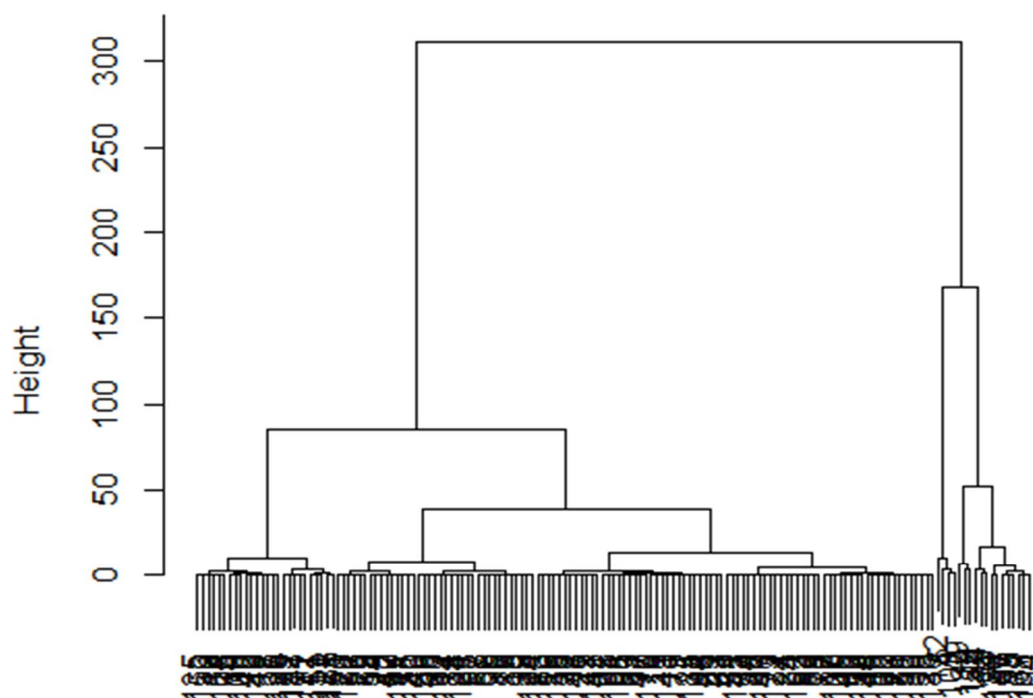
*** Gene id - {938, 2859, 12298, 1268, 13256, 13258, 2610, 18236}

The results of table shows that genes with gene id: { 938, 2859, 18236,16990, 8822} are the most expressive genes or the genes which accounts for maximum variability. Altogether out of 12607 genes five genes are most differentially expressed or significant. As the cluster number is increased the within variance

decreases and between variance increases leading to lesser error within the clusters. Hence the percentage of misclassification is 84.4% or 85% approximately.

4.2. Hierarchical clustering: The cluster dendrogram by Ward’s method is obtained for the gene expression data. The results of which are shown below.

Cluster Dendrogram



d
hclust (*, "ward")

The cluster dendrogram of hierarchical clustering is obtained by divisive method. The sample of 12607 genes are considered as one cluster which splits up to different clusters step by step applying the Euclidean distance. From the dendrogram we can see if the cluster tree is cut at height=250 we get two cluster where the second cluster accounts for less than ten percent of genes and the second cluster for the remaining. If the tree is cut at height =1500 the number of clusters are three. Cluster two and three accounts for less than twenty percentage of genes and the remaining forms one large cluster.

5. Summary

We have illustrated here how k-means clustering performs better when compared to hierarchical clustering. Though hierarchical clustering gives a pictorial representation the dendrogram which is easier to identify the clusters it becomes difficult when the size of the observations (genes) is very large. From

the results of the k-means and Hierarchical clustering one can see that out of the 12607 genes less than one percent that is a total of five to fifteen genes only are differentially expressed or significant. K-means clustering gives a better interpretation of the clusters compared to hierarchical as the numbers of genes belonging to the different clusters are clearly identified where hierarchical clustering gives a pictorial representation (a dendrogram) of the different clusters. When the data is very large (huge gene size) identification of the genes belonging to each cluster becomes difficult.

References:

1. **Baldi .P and Hatfield G.W. (2002)**, DNA Micro arrays and Gene expression from Experiments to Data Analysis and Modelling. Cambridge University Press
2. **Bishop C. M. ,(2006)**, Pattern Reorganization and machine learning -Springer
3. **Brown PO, Botstein D,(1999)**, Exploring the new world of the genome with DNA

microarrays. *Nat Genet*;21(1 Suppl):33-7. Review.

4. **Cuverhouse Robert, Duncan Jillard Shannon William (2003)**, Analyzing Micro array Data using Cluster Analysis- *Pharmogenomics* 41-51.

5. **Darlene R,Goldstein ,Debashis Ghosh and Erin M Conlon(2002)**,Statistical Issues in Clustering-*Statistica Sinica* 12219-240.

6. **Dudoit S, Gentleman RC, Quackenbush J,(2003)**, Open source software for the analysis of Microarray data. *Bio techniques*. Mar;Suppl:45-51.

7. **Everitt,B.S.(1993)**,Cluster Analysis. Arnold,London.

8. **Eisen M,Spellman P,Brown Po et al (1998)**, Cluster analysis and display of genome-wide expression patterns.*Proc.National Academic Science* 95;14863-14868.

9.**Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.,Coller, H., Loh, M., Downing, J., and Caligiuri, M.**

(1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–536.

10. **Hartigan J.A .,Mohanty.S.(1992)**, The Run test for Multimodality .*Journal of Classification* ,9:63-70.

11. **Hartigan.J,Wong.M (1979)**,A K-means clustering algorithm, *Applied Statistics* 28,pg100-108

12. **Hastie T ,Tibshirani R et al(2001)**,The elements of statistical learning ,Spinger ,New York City.

13. **Levy Belitska: (2006)**,a generalized clustering problem-Published by The Berkeley Electronic press.

14. **Yeung .K.Y.,Fraley .C,Murna .A,Raftery A.E and Ruzzo W.L(2001)**,Model Based clustering and data transformations for gene expression data; *Bioinformatics* ,17(10)977-987.

15. **Ward J.(1963)**,Hierarchical groupings to optimize an objective function .*Journal of the American Statistical Association* 58:234-44