295f Finding Distinct Clusters in Gene Expression Data Using Similarity in Principal Component Subspaces

Sudhakar Jonnalagadda and Rajagopalan Srinivasan

Introduction

Clustering is widely used in gene expression data analysis to identify groups of similarly expressed (co-expressed) genes. Several clustering techniques have been proposed and successfully employed with gene expression data [1]. However, all the clustering techniques need some kind of user interference for the successful identification of clusters. The widely used hierarchical clustering displays the series of nested clusters as a tree, i.e. dendrogram, and leaves the task of selecting clusters to the user. Partition approaches such as k-means, fuzzy c-means, model-and neural network-based techniques, which are commonly used with gene-expression data, can automatically identify clusters if the user specifies the number of clusters a priori. Finding the number of clusters in a dataset is called the cluster validation problem.

While a wide variety of cluster validation methods have been proposed in data mining literature, very few have been applied to gene expression data. Recently, Bolshakova and Azuaje (2003) [2] recommended three general cluster validation methods — Silhouette index, Dunn's index, and Davies-Bouldin index — for gene expression data. These indices evaluate the partitions generated using clustering algorithm and find the ‘best' partition based on intra- and inter-cluster distances. We have previously shown the shortcomings of such distance based methods for gene expression data [3]. In this paper, we propose a general method that automatically finds the maximum number of ‘distinct' clusters in gene expression data.

Proposed methodology

The proposed method for finding distinct clusters is based on the definition of cluster: objects within the cluster are similar to one another (homogeneity) while being dissimilar to objects in other clusters (separation or distinctness). A partition with distinct, homogenous clusters is more preferable to other partitions. Here we use the Principal Component Analysis (PCA) similarity metric to identify such a partition.

The proposed method identifies, from a set of candidate partitions, the one with the maximal number of distinct clusters. Initially a number of candidate partitions are generated, for example, by using different clustering techniques and/or by specifying different number of clusters, k, in each partition. Then each partition is evaluated for ‘distinctness' of clusters by measuring the similarity of each pair of clusters in that partition. PCA is used to characterize each cluster in the given partition by its dominant eigenvectors that describe the correlation between the constituent genes. Similarity between each pair of clusters is measured as the angle between their Principal Component subspaces. A cluster is deemed to be ‘distinct' if it shows low similarity to all other clusters in that partition. The method then assigns each candidate partition a cumulative measure of the distinctness of all the clusters, called the Net Principal Subspace Information (NEPSI) Index. A candidate partition with the highest NEPSI index value has the maximal number of distinct clusters and is selected as the ‘best' partition.

Case study

In this paper, we evaluate the proposed method using two gene expression datasets. We use two different clustering algorithms techniques – k-means and model-based clustering – for generation of partitions. For the first dataset, we also report the comparison of the ‘distinct' clusters identified by the proposed method with the expert's classification available for that dataset. This comparison clearly shows the ability of the proposed method to meet the expert's classification. For the other dataset, we use the Gene Ontology (GO) terms to verify that the clusters are enriched with functionally relevant genes by collecting the hierarchy of GO terms for all three categories: Process, Function, and Cellular components. Since clusters are generally expected to be enriched with functionally relevant genes, this analysis provides clear means to evaluate the proposed method.

References

[1] Jiang ,D., Tang, C. and Zhang, A. (2004) Cluster analysis for gene expression data: A Survey. IEEE transactions on knowledge and data engineering, 16, 1370-1386.

[2] Bolshakova,N. and Azuaje, F. (2003) Cluster validation techniques for genome expression data. Signal Processing, 83, 825-833.

[3] Sudhakar, J., Rajagopalan, S (2004) An information theory approach for validating clusters in microarray data. Presented in Intelligent Systems for Molecular Biology (ISMB 2004).