Cluster Analysis of Gene Expression Data

Alan Wee-Chung Liew
Griffith University, Australia

Ngai-Fong Law
The Hong Kong Polytechnic University, Hong Kong

Hong Yan
City University of Hong Kong, Hong Kong
University of Sydney, Australia

INTRODUCTION

Important insights into gene function can be gained by gene expression analysis. For example, some genes are turned on (expressed) or turned off (repressed) when there is a change in external conditions or stimuli. The expression of one gene is often regulated by the expression of other genes. A detail analysis of gene expression information will provide an understanding about the inter-networking of different genes and their functional roles.

DNA microarray technology allows massively parallel, high throughput genome-wide profiling of gene expression in a single hybridization experiment [Lockhart & Winzeler, 2000]. It has been widely used in numerous studies over a broad range of biological disciplines, such as cancer classification (Armstrong et al., 2002), identification of genes relevant to a certain diagnosis or therapy (Muro et al., 2003), investigation of the mechanism of drug action and cancer prognosis (Kim et al., 2000; Duggan et al., 1999). Due to the large number of genes involved in microarray experiment study and the complexity of biological networks, clustering is an important exploratory technique for gene expression data analysis. In this article, we present a succinct review of some of our work in cluster analysis of gene expression data.

BACKGROUND

Cluster analysis is a fundamental technique in exploratory data analysis (Jain & Dubes, 1988). It aims at finding groups in a given data set such that objects in the same group are similar to each other while objects in different groups are dissimilar. It aids in the discovery of gene function because genes with similar gene expression profiles can be an indicator that they participate in related cellular processes. Clustering of genes may suggest possible roles for genes with unknown functions based on the known functions of some other genes in the same cluster. Clustering of gene expression data has been applied to, for example, the study of temporal expression of yeast genes in sporulation (Chu et al., 1998), the identification of gene regulatory networks (Chen, Filkov, & Skiena, 1999), and the study of cancer (Tamayo et al., 1999).

Many clustering algorithms have been applied to the analysis of gene expression data (Sharan, Elkon, & Shamir, 2002). They can be broadly classified as either hierarchical or partition-based depending on how they group the data. Hierarchical clustering is further subdivided into agglomerative methods and divisive methods. The former proceed by successive merging of the N objects into larger groups, whereas the latter divide a larger group successively into finer groupings. Agglomerative techniques are more common in hierarchical clustering.

Hierarchical clustering is among the first clustering technique being applied to gene expression data (Eisen et al., 1998). In hierarchical clustering, each of the gene expression profile is considered as a cluster initially. Then, pairs of clusters with the smallest distance between them, are merged together to form a single cluster. This process is repeated until there is only one cluster left. The hierarchical clustering algorithm arranges the gene expression data into a hierarchical tree structure known as a dendrogram, which allows easy visualization and interpretation of results. However, the hierarchical tree cannot indicate the optimal number of clusters in the data. The user has to interpret the tree topologies and identify branch

points that segregate clusters of biological relevance. In addition, once a data is assigned to a node in the tree, it cannot be reassigned to a different node even though it is later found to be closer to that node.

In partition-based clustering algorithms, such as K-means clustering (Jain & Dubes, 1988), the number of clusters is arbitrarily fixed by the users at start. Setting the correct number of clusters can be a difficult problem and many heuristics are used. The basic idea of K-means clustering is to partition the data into a predefined number of clusters such that the variability of the data within each cluster is minimized. Clustering is achieved by first generating *K* random cluster centroids, then alternatingly updating the cluster assignment of each data vector and the cluster centroids. The Euclidean distance is usually employed in K-means clustering to measure the closeness of a data vector to the cluster centroids. However, such distance metric inevitably imposes an ellipsoidal structure on the resulting clusters. Hence, data that do not conform to this structure are poorly clustered by the K-means algorithm.

Other approach to clustering includes model-based approach. In contrast to model-free partition-based algorithms, model-based clustering uses certain distribution models for clusters and attempts to optimize the fit between the data and the model. Each cluster is represented by a parametric distribution, like a Gaussian, and the entire data set is modeled by a mixture of these distributions. The most widely used clustering method of this kind is the one based on a mixture of Gaussians (McLachlan & Basford, 1988; Yeung et al., 2001).

CLUSTERING OF GENE EXPRESSION DATA

Binary Hierarchical Clustering (BHC): In Szeto et al. (2003), we proposed the BHC algorithm for clustering gene expression data based on the hierarchical binary subdivision framework of Clausi (2002). Consider the dataset with three distinct classes as shown in Fig.1. The algorithm starts by assuming that the data consists of one class. The first application of binary subdivision generates two clusters A and BC. As the projection of class A and class BC have a large enough Fisher criterion on the A-BC discriminant line, the algorithm splits the original dataset into two clusters. Then, the binary subdivision is applied onto each of the two clusters. The

Figure 1. The binary subdivision framework. (a) Original data treated as one class, (b) Partition into two clusters, A and BC. (c) Cluster A cannot be split further, but cluster BC is split into two clusters, B and C. (d) Both cluster B and C cannot be split any more, and we have three clusters A, B, and C. (Figure adopted from Clausi (2002))