Chapter XI
Clustering Methods for Gene-Expression Data

L.K. Flack
University of Queensland, Australia

G.J. McLachlan
University of Queensland, Australia

ABSTRACT

Clustering methods are used to place items in natural patterns or convenient groups. They can be used to place genes into clusters to have similar expression patterns across the tissue samples of interest. They can also be used to cluster tissues into groups on the basis of their gene profiles. Examples of the methods used are hierarchical agglomerative clustering, k-means clustering, self organizing maps, and model-based methods. The focus of this chapter is on using mixtures of multivariate normal distributions to provide model-based clusterings of tissue samples and of genes.

INTRODUCTION

DNA microarrays are collections of microscopic DNA spots arrayed on a solid surface. Each of these DNA spots will hybridize with a particular target RNA or DNA sequence. Optical measurements are made of fluorophores attached to the target RNA or DNA. DNA microarrays allow us to simultaneously read expression levels of expression levels on thousands of genes. They and other high throughput measurement methods bring many new opportunities in data analysis, but they also create difficulties in taking advantage of this amount of data.

A variety of multivariate methods have been used to look for relationships among the genes and tissue samples. Cluster analysis has been one of the most frequently used of these methods. It has been useful in the discovery of gene function and of groups of interconnected biological processes; see Eisen et al. (1998) for examples.

In medical applications, we are usually interested in the supervised and unsupervised grouping of tissue samples on the basis of the genes expressed. In the latter context, the intent is to identify what subtypes of cancer or other diseases exist, with the aim of assigning patients to these subgroups in order to aid their prognosis and therapy. In biological studies, we are usually interested in partitioning the genes into clusters in which the genes display similar patterns of gene expression across the relevant tissue samples (or cell lines). Genes in the same cluster are perhaps likely to be part of the same biological pathway or otherwise related.

It can be seen there are two distinct but related clustering problems with microarray data. One problem concerns the clustering of the tissues on the basis of the genes; the other concerns the clustering of the genes on the basis of the tissues. This duality in cluster analysis is quite common.

The aim of clustering is to put items into groups so that they are more similar to each other than they are to members of other clusters. One of the difficulties of clustering is that the notion of clustering is vague. A useful way to think about the different clustering procedures is in terms of the shape of the clusters. The majority of the existing clustering methods assume that a similarity or distance measure or metric is known *a priori*; often the Euclidean metric is used. But clearly, it would be more appropriate to use a metric that depends on the shape of the clusters. As pointed out by Coleman et al. (1999), the difficulty is that the shape of the clusters is not known until the clusters have been found, and the clusters cannot be effectively identified unless the shapes are known.

We will give a brief overview of clustering before we describe its application to microarray data. More detailed accounts of clustering can be found in the many books on this topic; for example, Everitt (1993), Hartigan (1975), and Kaufman and Rousseeuw (1990).

SOME HEURISTIC CLUSTERING METHODS

In cluster analysis, we wish to group a number (n) of entities into a smaller number (g) of groups on the basis of measurements of some variables associated with each entity. We let $y_j = (y_{1j}, \ldots, y_{pj})^T$ be the observation or feature vector of p measurements $y_{1j}, \ldots, y_{pj}$ made on the jth entity ($j = 1, \ldots, n$) to be clustered. In discriminant analysis the data belong to g known classes and we wish to create an allocation rule to allow us to assign an unclassified entity to one of these classes on the basis of its feature vector.

In cluster analysis, we have no prior knowledge of group membership or structure, except possibly the number of classes. Clustering can have either or both of two aims. We might wish to split the data into several groups with no implication that these groups are a natural division of the data. We might do this for the sake of convenience or mathematical tractability. In this case intergroup boundaries do not necessarily have to be in regions of the feature space with a relatively low density of points. The feature space will be divided into contiguous and at least in some sense compact regions. This is sometimes called dissection or segmentation. Alternatively, we might wish to find a natural subdivision of the entities into groups. In this case the clusters will be regions of the feature space with a relatively high density of points separated by regions with relatively low densities of points. Sometimes the distinction between the two aims is stressed. But often it is not made, particularly as most methods for finding natural clusters are also useful for segmenting the data.

Clustering methods can be categorized as hierarchical or nonhierarchical. With a hierarchical clustering method every cluster obtained is a split or merger of clusters obtained at the previous stage. Hierarchical clustering methods can be agglomerative, starting with $g = n$ clusters or divisive starting