Clustering Techniques for Revealing Gene Expression Patterns

Crescenzio Gallo
Department of Clinical and Experimental Medicine, University of Foggia, Italy

Vito Capozzi
Department of Clinical and Experimental Medicine, University of Foggia, Italy

INTRODUCTION

The possible applications of modeling and simulation in the field of bioinformatics are very extensive, ranging from understanding basic metabolic paths to exploring genetic variability. Molecular biologists need robust computational tools to determine models that can learn to recognize DNA and amino acid sequences and assign protein structures to certain sequences. Experimental results carried out with DNA microarrays allow researchers to measure expression levels for thousands of genes simultaneously, across different conditions and over time. A key step in the analysis of gene expression data is the detection of groups of genes that manifest similar expression patterns. In this Article we describe the main clustering algorithms developed for analyzing gene expression data, comparing their results with the classification deriving by the application of unsupervised neural networks.

In the analysis of gene expression data of particular interest is the search for correlated patterns, which is typically done by clustering analysis. DNA microarray technologies (Lockhart *et al*., 1996) allow the monitoring of thousand genes quickly and efficiently. These technologies have introduced new rules for the exploration of an organism with a genome wide-ranging vision. In particular, the study of gene expression of a complete genome (such as that of *Saccharomyces cerevisiae*) is now possible. Studies have also been developed (Perou *et al*., 1999) through the use of DNA microarrays until the complete mapping of the human genome. The production of targeted drugs and identification of drugs are other areas that can significantly benefit from these techniques.

One problem inherent the use of DNA microarray technology is the huge amount of data available, the analysis of which is a significant problem per se. Several approaches are used in the analysis of gene expression data, grouped in two areas: clustering and classification. Clustering is a purely data-driven activity that uses only data from the study or experiment to group together measurements. Classification, in contrast, uses additional data, including heuristics, to assign measurements to groups. Among these, commonly statistical methods applied to microarray data are Hierarchical Clustering (Sneath & Sokal, 1973) and (Unsupervised) Neural Networks (Herrero *et al*., 2001): The identification of the optimal method for the analysis of these data is still a topic of discussion.

In this Article we examine some methods for gene co-expression analysis, such as "correlation graphs" and supervised-unsupervised clustering methods. The next section is a brief exposition of the underlying background of clustering techniques. Then we detail the clustering algorithm based on correlation graphs. Next we examine the application of supervised and unsupervised techniques. The Article ends with some final considerations and further research directions.

BACKGROUND

Cluster Analysis

Cluster analysis (Duda *et al*., 2001; Jain & Dubes, 1998; Jain *et al*., 1999) can be summarized as follows. Let n experimental outcomes $x_i \in \mathbb{R}^m$ ($i=1...n$; each point has m components): the objective is to identify

DOI: 10.4018/978-1-4666-5888-2.ch042

the underlying structure of the data, partitioning the n points into k clusters in order to group in the same cluster points "closer" to each other than to points belonging to different clusters.

In the above statement no clear definition exists for "closer" points, and this depends on the resolution at which the data are viewed. The last issue is typically addressed by generating a tree of clusters (a dendrogram), whose number and structure depend on the resolution that is used. Two of the most common methods of clustering gene expression data are hierarchical clustering and *k*-means clustering (Geraci *et al*., 2009; Jain *et al*., 1999).

Hierarchical clustering is the most used, and produces a representation of the data with the most similar patterns grouped in a hierarchy of subsets. This method, however, suffers from considerable problems when applied to data containing a significant amount of noise, revealing itself of low applicability. In this case the solutions may not be unique and be data-order dependent. Mathematically, hierarchical clustering involves computing a matrix of all distances for each expression measurement in the study, merging and averaging the values of the closest nodes, and repeating the process until all nodes are merged into a single node.

K-means clustering involves generating cluster centers in *n*-dimensions and computing the distance of each data point from each of the cluster centers. The data points are assigned to their closest cluster center. A new cluster position is then computed by averaging the data points assigned to the cluster center. The process is repeated until the positions of the cluster centers stabilize.

Clustering microarray gene expression data is useful because it may provide insight into gene function. For example, if two genes are expressed in the same way, they may be functionally related. In addition, if a gene's function is unknown, but it is clustered with genes of known function, the gene may share functionality with the genes of known function. Similarly, if the activity of genes in one cluster consistently precedes activity in a second cluster, the genes in the two may be functionally related. For example, genes in the first cluster may regulate activity of genes in the second cluster.

Unsupervised Neural Networks

Artificial Neural Networks can be used not only for prediction but also for data classification. Unlike regression problems, where the goal is to produce a particular output value for a given input, classification problems require labeling of all data as belonging to one of n known classes. These classification models are typical cases of supervised networks, in which it is a priori possible to associate data to clusters and to train the neural network for the classification of further data.

However, there are cases in which the association is not known before. It is then up to the neural network to independently find the associative structures between data, grouping them in clusters. These neural networks are known as unsupervised or self-organizing (associative memories).

There is not — therefore — a *correct* output, as there is no erroneous output. This involves careful observation and analysis by the researcher. In fact, after an unsupervised network has been trained, it must be tested in many directions as it is important to understand what are the data structures resulting from the neural network itself.

An unsupervised network can be compared to a hypersurface in the *n*-dimensional space, with *m* "valleys" or minima corresponding to the desired shape of the outputs. The presentation of an input $X_i=\{x_{i1},...,x_{in}\}$ to the network is equivalent to putting a weight in the corresponding point on the surface, thus making it fall into the nearest valley, corresponding to $Y_i=\{y_{i1},...,y_{in}\}$. Obviously, an input X_i' "similar" to X_i — according to some similarity criterion — should fall into the same depression, drawing by association (hence the term "associative memories") the same output Y_i (Cammarata, 1990).

Practically, an unsupervised neural network consists of a number of codebook vectors Y_i, which constitute the center of each cluster. They are the same size of the input space, and their components are the parameters of the unsupervised network. The codebook vectors are the network's neurons. Some parameters help in training the neural network. The position of the vectors should be adjusted so that the average Euclidean