Feature selection for genomic data

Paola Cerchiello[1], Silvia Figini[2]

[1] University of Pavia, paola.cerchiello@eco.unipv.it
[2] University of Pavia, silviafigini@eco.unipv.it

Abstract. Building predictive models for genomic mining requires feature selection, as an essential preliminary step to reduce the large number of available variable. Feature selection in the process of select a generally smaller subset of variables (features) that can be considered the best, from a statistical point of view, with respect to the employed model for the analysis. In gene expression microarray data, being able to select a few number of important genes not only makes data analysis efficient but also helps their biological interpretation. Microarray data have typically several thousands of genes (features) but only tens of samples. Problems which can occur due to the small sample size have not been addressed well in the literature. Our aim is to discuss some issues on feature selection applied to microarray data in order to select the most important genes from a predictive point of view.

Keywords: Feature selection, Gene expression, Marker Selection, Kruskal-Wallis test, Model Assessment, Predictive models.

1 Introduction

Many authors discuss the problem of selecting relevant features, and the problem of selecting relevant samples on data sets containing large amounts of irrelevant information. For large data sets, we can usually choose only a few of the most relevant features to build a model to classify the data. The resulting model will be at least as good as the one built from all the features. Hence it is often useful to select a subset of features of a data set to describe the data. In this paper, we focus on data sets with many features and a few samples. This paper is structured as follows: in Section 2 we present a review of statistical methods for feature selection. In Section 3 we describe our proposed method for feature selection and in Section 4 our proposed predictive models. Finally in Section 5 we present the application of our methods to the available data.

2 Feature selection

The basic feature selection problem is an optimization problem, with a performance measure for each subset of features to measure its ability to classify the samples. The problem is to search through the space of feature subsets to identify the optimal or near-optimal ones with respect to the performance measure. Feature selection is generally an empirical process that is performed prior to, or jointly with, the parameter estimation process. Many successful feature selection algorithms have been devised. Yang and Honavar (1997) classify many existing approaches into three groups: exhaustive search, heuristic search, and randomized search. Another common way to classify feature selection algorithms is determined by how the learning method is integrated into the algorithm, see e.g. Xing, Jordan and Karp (2001).Yang et al. (2000), Forman (2003) and Golub et al.(2000). DNA microarrays have been used by biologists to monitor the level of gene expression of thousands of genes in different biological tissues. Microarray technologies produce gene expression patterns that provide dynamic information about cell functions. These information can be used to investigate complex interaction within the cell. In this context, data mining methods can be used to determine co-regulated genes and suggest biomarkers for specific diseases, or to ascertain and summarize the set of genes responding to a certain level of stress in an organism.
Being gene expression data typically high-dimensional, they need appropriate statistical features