Feature Selection in Predictive Genomic Mining

Selezione delle Caratteristiche nel Genomic Mining

Silvia Figini Paolo Giudici
University of Pavia – Data Mining Laboratory silviafigini@tiscali.it

Keywords : Boosting, Feauture Selection, Marker Selection, Predictive Methods

1. Introduction

Microarray technologies produce gene expression patterns that provide dynamic informations about cell functions. These informations can be used to investigate complex interaction within the cell. In this contest, 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. A typical question in genomic mining, see e.g. Speed 2003, is in fact "which gene is the most similar to which" in terms of gene expression. Another important aspect is the correlation between gene expressions and malignant samples. Gene expression data being typically high-dimensional, it needs appropriate statistical features to discern possible patterns and to identify mechanisms that govern the activation of genes in a organism.

2. The problem

Our data set is composed by 112.896 gene expressions ordered into 224 columns and 504 rows. Columns represent a set of 224 genes, rows correspond to 504 samples, covering 8 tissue types - adipose tissue, breast, colon, kidney, liver, lung, ovary and prostate - both normal (249 samples) and malignant (255 samples).
Measured values of the gene expression data have been put in bins and markers as being "1" (highly expressed) or "0" (not-highly expressed). We have also applied label "1" to malignant tissues and "0" to normal tissues. The goal of the study is to create a valid predictive model to diagnose malignant tissues, based on the observation of gene expressions.

3. Methodology of analysis

First, we have seen the problem of the correlation between gene expression and malignant samples as a logistic regression problem with a categorical predictor variables – genes – and a response binary variable being the sample's status "1" (malignant) or "0" (normal). These results have been compared to tree-based methods. Then we used association rules to analyse genes resulting from these predictive

methods. Mining of association rules, in fact, had already been successfully applied on microarray data by using the *A priori* algorithm. Associations rules can be used to express associations between cell environmental effects and gene expressions, to diagnose a profiled cancer sample, or to analyse drug treatment effects. In order to find the best predictive models, we have reduced the number of inputs by setting the status of input variables that are not related to the target as rejected. We have compared chi-square selection and a new approach on variable selection, based on marker selection.

The chi-square selection criterion is available for binary targets. This criterion also provides a fast preliminary variable assessment and facilitates the rapid development of predictive models with large volumes of data.

Variable selection, based on chi-square, is performed using binary variable splits for maximizing the chi-square values of a contingency table. Each level of the ordinal or nominal variables are decomposed into binary variables.

Marker selection approach, see e.g. R. Mott, 2003, is based on the structure of the genes. The probability of detecting an association between a marker and a diseased phenotype decreases with distance between the marker and the actual position of the gene responsible for the phenotype. Thus, one can maximize the probability of detecting disease linkage by choosing markers as closely spaced as possible.

In general, the decrease in probability of detecting association is not linear with the distance; this probability tends to be relatively constant across patches of the genome known as "haplotype blocks".

The locations of genetic markers are known prior to project initiation. In general, the number of known genetic markers exceeds the number necessary and/or affordable for a project. Suppose we have a set of K markers typed across n tissues, producing $2n$ haplotypes, that are assumed to be known. Suppose haplotype with label i occurs with frequency f_i in the sample. We are interested in finding subsets of markers which capture as much of the haplotypes diversity as possible, weighted by the populations frequencies of the haplotypes.

In our talk, for marker selection we use the entropy as a good measure of haplotypic diversity that attains a maximum if all haplotypes are present in equal quantities. If only a subset s of markers were typed then some of the original haplotypes might become indistinguishable and hence will be merged. The sequence of marker subset $s_1 s_2 ..s_k$ generates a monotonic sequence of optimal approximations (as measured by their entropy) to the haplotypic structure of the data. The sequence need not be unique. This method has been implemented in a program using simple recursive search algorithm which generates and evaluates all possible marker subsets.

We compare marker selections based on the entropy, with a marker selection based on a procedure that divides a set of variables into either disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster, which may be either the first principal component or the centroid component. The procedure tries to maximize the sum across clusters of the variance of the original variables that is explained by the cluster components. Either the correlation or the covariance matrix can be analyzed. If correlations are used, all variables are treated as equally important. If covariances are used, variables with larger variances have more importance in the analysis.

In this approach we select the variables that are most important to explain the patient disease.

In our applications it turns out that marker selection based on the entropy and the second approach, give equal results in term of gene selection.
With the previous approach on gene selection, we compare a set of predictive models on the basis of confusion matrix, different indexes (error rate, accuracy, sensitivity, specificity) and some graphs (lift chart and ROC Curve).
In order to understand the relationships between genes we use models based on link analysis. In this paper, link analysis is the examination of the linkages between genes in a sample of human tissues and consider all indirect sequences of any order up to a maximum of 10.
Finally, we compare the best model with a new model derived from boosting method. Boosting method is based on re-weighted re-sampling developed from a weak learning algorithm (Freund and Schapire, 1997). The weights in the re-sampling are increased for those observations most often misclassified in the previous models. Therefore, the distribution of weights observation is based on the model performance of the previous samples. Boosting models are created by re-sampling the training data and fitting a separate model for each sample. Boosting adaptively re-weights each training observation. In our paper we show that it is possible to improve the previous results (logistic regression and tree-based model) using boosting method.

References

Bauer E. and Kohavi R. (1998), An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants, in: *Machine Learning*.

Breiman L. (1996), Bagging Predictors, in: *Machine Learning*.

Clemen W.W. (1995), Combining Forecasts: A Review and Annotated Bibliography, in: *International Journal of Forecasting*.

Giudici P.(2003), *Applied data mining*, Wiley.

Hastie T. and Tibshirani R. and Friedman J. (2001), *The elements of statistical learning*, Springer.

Hand D. and Mannila H. and Smyth P.(2001), *Principles of Data mining*, MIT Press.

Lange W. (2002), *Mathematical and statistical methods for genetic analysis*, New York, Springer.

Speed T. (2003), *Statistical analysis of gene expression microarray data*, New York, Chapman & Hall.