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

to discern possible patterns and to identify mechanisms that govern the activation of genes in a organism.
In order to find the best predictive models, we have reduced the number of input variables by means of feature selection. In particular, in this paper we present a method for feature selection based on marker selection, see e.g. Mott (2003) compared to another approach based on nonparametric techniques Cerchiello Gindici (2006).

2.1 Marker selection

In high dimensional data sets, the identification of irrelevant inputs is more difficult than the identification of redundant ones. A good strategy considers two sequential steps: first of all reducing redundancy and then tackling irrelevancy in a lower dimension space.
Marker selection approach, see e.g. R. Mott (2003), is based on the structure of the genes. Consider the general representation of the frequency distribution of a qualitative variable with K levels. Null heterogeneity, holds when all the observations assume the same level. That is if $p_i = 1$ for a certain i, and $p_i = 0$ for the other k-1 levels. Maximum heterogeneity, holds when the observations are uniformly distributed amongst the k levels, that is $p_i = 1/k$ for all $i = 1, ..., k$. Heterogeneity measures can be extended and applied to gene expressions. As a measure of genes diversity, the entropy (E) can be calculated using:

$$E - -\sum_{k=1}^{m} p_i log p_i,$$

where p_i is the probability of gene i being activated, and K the number of genes. Wanting to obtain a 'normalised' index, which assumes values in the interval [0,1], E can be rescaled by its maximum value, obtaining the following relative index of heterogeneity:

$$E' = \frac{E}{log(K)},$$

In order to select the most predictive genes, genes are sequentially subdivided in groups (as in a divisive cluster analysis algorithm). The previous entropy is calculated for each chosen subset. It will continuously increase starting from 0 up to a maximum of E. Grouping and, hence, gene marker selection is stopped when a suitable threshold is reached (e.g. 0.95).
If s is a subset of t we have that $E(s) \prec E(t) \prec E$. The difference $E(t)$ - $E(s)$ is a good measure of how nested subsets compare in describing the data. A sequence of marker subsets $s_1 ... s_k$ generates a monotonic sequence of optimal approximations (as measured by their entropy) to the gene structure of the data. The probability of detecting an association between a marker and a diseased phenotype decreases with the distance between the marker and the actual position of the gene responsible for the phenotype. Thus, the probability of detecting disease linkage can be maximized by choosing markers as closely spaced as possible. This procedure is naturally related to principal component analysis and can be used as an alternative method for eliminating redundant dimensions.
This type of variable clustering is able to efficiently find groups of variables that are as much correlated as possible among themselves and as much uncorrelated as possible with respect to variables contained in other clusters. If the second eigenvalue for the cluster is greater than a specified threshold, the cluster is split into two different dimensions. The reassignment of variables to clusters occurs in two phases. The first is a nearest component sorting phase, similar in principle to the nearest centroid sorting algorithms described by Anderberg (1973). During each iteration the cluster components are computed and each variable is assigned to the component with which it has the highest squared correlation. The second phase involves a search algorithm in which each variable is tested to see if assigning it to a different cluster increases the amount of variance explained. If a variable is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested.

2.2 Nonparametric feature selection

On of the nonparametric method employed in this contribution is the chi-square selection criterion in the case of binary targets. This criterion 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.

We stress out the the previous one is not the only employed feature selection method. In order to obtain a valid comparison on the topic under analysis, we introduce the Kruskal-Wallis test (see e.g. Conover, 1971), the non parametric version of ANOVA analysis representing a simple generalization of the Wilcoxon test for two independent samples, as well. On the basis of K independent samples $n_1, \ldots, n_k$,in this context represented by the different status of the tissues (normal vs malignant) present in the analysis , a unique big sample is created by means of fusion of the originals k samples. The above result is ordered from the smaller value (frequency) to the bigger one, and a rank is assigned to each one. Finally R_i is calculated, that is the mean of the ranks of the observations in the -th sample. The statistic is:

$$KW = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{K} n_i (R_i - \frac{N+1}{2})^2}{1 - \frac{[\sum_{j=1}^{g} (t_i^3 - t_i)]}{(N^3 - N)}} \tag{1}$$

where the denominator factor is needed when there are tied observations (as it is typical in gene expression application)and the null hypothesis, that all the gene expression distributions are the same, is rejected if:

$$KW > X^2_{K-1} \tag{2}$$

We also recall some elements regarding classification or decision tree methods that will be employed as feature selection method as well. They are a good choice when the task of the analysis is classification or prediction of outcomes and the goal is to generate rules that can be easily understood and explained. In particular in genomic mining field they represent trees in which internal nodes are labelled by genes names, branches departing from them are labelled by tests on the weight that the gene has in the analyzed patient, and finally leafs represent category (normal, malignant). This kind of classifier categorizes a test row (a patient) p_j by recursively testing for the wights that the gene labelling the internal nodes have in vector p_j, until a leaf node is reached; the label if this node is then assigned to p_j. As the Kruskal-Wallis test, the decision tree is a method belonging to the family of non parametric models, so that we do not have to choose a distribution for the gene present in the dataset, that constitutes a not simple problem in this kind of application.

In this particular context the CHAID tree (see, e.g. Breiman, 1984) has been employed. CHAID (Chi Squared Automatic Interaction Detector), as we said before, represents a valid and simple non parametric method to investigate the dependency of a response variables on several explanatory variables. The name itself suggests that the measure employed for the evaluation of such a dependency is the known chi square index on the basis of which the homonymous test is derived. The purpose of the procedure is to split a set of observations in a way that, the subgroups (final leaves) differ significantly with respect to the designated criterion. The segments derived by CHAID are mutually exclusive and exhaustive which means that the segments do not overlap and each observation of the sample is contained in exactly one segment. A CHAID tree can be employed every time a categorical dependent variable and a set of categorical independent variables (continuous may be used as well but the first step of the analysis is to categorize them) are present in the database.

The procedure used by the CHAID tree is quite elaborated and is divided in two distinct phases:the first one is dedicated to the evaluation of the best number of categories within each variable by means of a strategy of 'merge and divide', based on the p-values derived by the below

multiple chi-square test. Once completed the above step, another phase starts aiming at the location of the best independent variable as first predictor of the target variable. Once again a chi square test is employed and is computed for each contingency table derived from the intersection of the target variable with every single response variable. As a consequence the first split falls on the variable that presents the higher computed chi square test and the lowest p-value, i.e. the stronger association between them. Once completed the first level of the tree, the procedure starts again considering all the available variables included the first level one. Thereby the procedure is recursive and it stops once satisfied one of the selected stopping rules.
The methodology here proposed combines the main elements of both the models above. First of all we applied the Kruskal-Wallis test to the gene expression available dataset , in order to select genes characterized by distribution profiles more heterogeneous and different from each other. This part of the analysis is useful to eliminate genes which values repeat in a constant or semi-constant way along the group variable (malignant or not tissue). The test is repeated 226 times (as many as the genes present in the dataset)each time on a different variable obtaining a list of test statistic and associated p-values. From the comparison between the latter and the α choose (1 %), 108 variables (genes) are selected in the sense that they reject the null hypothesis of similar distribution in the two different groups (tissues status). Even if the dimension has been strongly reduced, another step of selection is needed in the direction of keeping only the most relevant genes satisfying the initial requirement of best discrimination.
The genes selected as shown above ,are combined, as we will see later, with genes located by the application of the decision tree. In fact the decision tree is applied to the same data set previously employed and during every phase, the recursive algorithm selects features that better distinguish tissues status. After the application of the two models we have two different sets of genes, one selected with the Kruskal-Wallis test, and another one located with the decision tree. Now it is useful to remember the objective of this analysis: first of all the dimensionality reduction to obtain just genes which expression reveal a real influence on the status. In the table we report only genes in common between the test and the decision tree based on Chi square measure. As a consequence we finally obtain a very small set of genes that will be compared to the marker selection derived ones.

3 Methodological Proposal

Our methodological proposal aims at the comparison of the above explained feature selection methods by means of an opportune predictive model. We want to investigate which feature selection method is better in choosing genes most related to the target variable that in this context is represented by the tissue status (malignant vs normal). Among the several available predictive techniques, we decided to employ decision tree models which impurity measure is the Gini index. We assess this chose model on the basis of the confusion matrix.

4 Application

In this section we describe a real application based on gene expressions. The available data is a sub-set taken from a large database (GeneExpress) and analyzed by S. Young et al (1997). The 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). The original database contains variables measured on a continous scale with both positive and negative values. In order to apply the statistical feature selection methods shown before, we recodified the variables on a categorical scale with values ranging between [-6; +5].We have also applied label '1' to malignant tissues and '0' to normal tissues (binary values of the target variable).
For sake of simplicity, we refer the reader to Figini and Giudici (2006) for the detailed descriptive

data analysis. First of all we have applied the feature selection methods described above on the modified dataset as we have explained so far.

For what concerns marker selection, we use the entropy as a measure of genes diversity that attains a maximum if all genes are present in equal quantities. If only a subset s of genes was typed then some of the original might become indistinguishable and hence will be merged. The sequence of genes subset generates a monotonic sequence of optimal approximations (as measured by their entropy) to the structure of the data. This method has been implemented in a software code using simple recursive search algorithm, which generates and evaluates all possible genes subsets. In this way we select the variables that are most important to explain the patient disease. The genes selected by marker selection approach are (in decreasing order of importance):

- **Gene149**
- **Gene119**
- Gene130
- Gene217
- Gene154
- Gene51
- Gene2
- Gene173
- Gene56
- Gene28

  Gene30
- **Gene15**
- Gene8
- Gene7

  Gene121
- Gene3
- Gene53
- Gene199
- Gene207
- Gene144
- Gene5
- Gene1

The methodology based on non parametric feature selection, explained above consists of two sequential steps:

  Kruskal-Wallis test to the gene expression available dataset
- CHAID tree applied to the same dataset.

This part of the analysis is useful to eliminate genes which values repeat in a constant or semi-constant way along the group variable (malignant or normal tissue).
The test is repeated 226 times (as many as the genes present in the dataset)each time on a different variable obtaining a list of test statistic and associated p-values.
The genes selected as shown above, are combined, with genes located by the application of the decision tree. In particular we consider as significant only genes in common between the test and the CHAID decision tree. As a consequence we finally obtain a very small set of genes:

- **Gene149**
- **Gene15**
- **Gene119**
- Gene185
- Gene193

As the reader can simply notice there are just three genes in common between the procedure, in fact the marker selection procedure tends to select a higher number of feature, thereby it is useful to evaluate which reduced set is the best in terms of prediction power. In order to better compare the two selections (the first one based on marker selection and the second one based on Kruskall wallis test), we have run two classification tree models with the same settings and compared the resulting goodness of fit measures like as misclassification rate and confusion matrix.
The confusion matrix (see,e.g. Giudici, 2003) is used as an indication of the properties of a classification (discriminator) rule. It contains the number of elements that have been correctly or incorrectly classified for each class. On its main diagonal we can see the number of observations that have been correctly classified for each class while the off-diagonal elements indicate the number of observations that have been incorrectly classified. We classify the observations of a validation dataset in four possible categories: the observations predicted as events and effectively such; the observations predicted as events and effectively non events; the observations predicted as non events and effectively events; the observations predicted as non events and effectively such. We have built two different decision trees, one based on genes selected by the marker selection approach and another one based on genes located with non parametric approach. They are both based on the same measure of impurity (Gini index) and the required minimal number of observations for a split is equal to 30. In order to select the better feature selection method in Table 1 we show the two confusion matrix.

Frequency	Pred marker=0	Pred marker=1	Pred KW-CHAID=0	Pred KW-CHAID=1
Obs marker=0	34	12	-	-
Obs marker=1	16	39	-	-
Obs KW-CHAID=0	-	-	33	12
Obs KW-CHAID=1	-	-	17	39

Table 1. Confusion Matrix of marker selection and Kruskal-Wallis CHAID selection

Table 2. Misclassification Errors

Misclassification errors	
Marker Select	27%
K-W CHAID	28%

As we can infer from the confusion matrix results, the two proposed feature selection methods are comparable and quite similar in term of misclassification error. By the way the number of selected genes are slightly different:

- 7 genes from marker feature selection approach
- 5 genes from Kruskal-Wallis CHAID selction

According to one of the most important principle in model selection field, known as parsimony rule, we should prefer the second approach. However more analysis are needed in order to extend this results. We would like to improve this contribution focusing on stratified feature selection, taking into account the biological composition for the available tissue types.

5 Acknowledgment

This work has been supported by MIUR PRIN FUNDS "Data Mining for e-business approaches", 2004-2006 and by MUSING 2006 contract number 027097, 2006-2010.
The paper is the results of a close collaboration between the two authors.

References

1. Anderberg, M.R.: Cluster analysis for applications, New York Academic Press, (1973).
2. Breiman L., Friedman J.H., Olshen R., and Stone C. J.: Classification and regression trees, Wadsworth, Belmont(1984).
3. Cerchiello P., Giudici P.,: A non parametric method to identify unknown authors, Technical Report number 187, (2006).
4. Conover W. J., : Practical nonparametric statistics, Wiley, New York(1971).
5. Figini S., Giudici P.,: Building predictive models for feature selection in genomic mining , Tchnical Report number 184 (2006).
6. Forman G., :An Extenisve empirical study of feature selection metrics for text classification. In Journal of Machine Learning Research, 3, 1289 1306(2003).
7. Giudici P.: Applied data mining, Wiley, (2003).
8. L. Liu, D. M. Hawkins, S. Ghosh, Young S.: Robust Singular Value Decomposition Analysis of Microarray Data, (2000).
9. Mott, R.: Marker selection, University of Oxford,(2003).
10. Slonim, D.K., Tamayo, P., Mesirov, J., Golub, T., and Lander, E.: Class prediction and discovery using gene expression data. Proceedings of the 4th Annual International Conference on Computational Molecular Biology, 263–272, (2000) .
11. Yang, J., and Honavar, V.: Feature subset selection using a genetic algorithm. Proceedings of the Genetic Programming Conference, 380 385,(1997).
12. Xing, E., Jordan, M., and Karp, R.: Feature selection for high-dimensional genomic microarray data. International Conference on Machine Learning, (2001).