Supervised Clustering of Genes

Marcel Dettling, Seminar für Statistik, ETH Zürich, CH-8092 Switzerland

Abstract. This paper presents a new method for supervised clustering of gene expression data that are equipped with categorical response variables. It searches for groups of genes with coherent average expression patterns by directly incorporating the response variables into the clustering process. The output of the procedure is very promising for medical diagnostics and functional genomics, since several gene clusters with clear separation of the response categories, good predictive potential as well as reasonable stability and relevance may be found.

Keywords. Microarray, Gene Expression, Discrimination, Classification.

1 Introduction

The recently developed microarray technology allows the simultaneous monitoring of up to several thousands of genes and is attested an enormous scientific potential. We assume to have a thoroughly preprocessed gene expression matrix, which is equipped with additional categorical response labels, describing e.g. cancer types. An important challenge in this setting is to find groups of genes which show optimal discrimination of cancer types, in order to a) accurately predict the phenotypes of new individuals for medical diagnostics and b) gain insights into biological processes and reveal how the genome works. Our new approach for finding such gene groups *directly* incorporates the response variables into the clustering process. The algorithm is a greedy stagewise procedure and relies on improving an empirical objective function that measures the strength for cancer type discrimination.

2 The Algorithm

Given a gene expression profile $X \in \mathbb{R}^p$ and its associated response variable $Y \in \{0, 1\}$, we model the conditional probability for class membership as

$$\mathrm{P}\left[Y = 1 | X = x\right] = f\left(x_{\mathcal{C}_1}, x_{\mathcal{C}_2}, \ldots, x_{\mathcal{C}_k}\right), \tag{1}$$

where $f(\cdot)$ is a nonlinear function and $x_{\mathcal{C}_i} \in \mathbb{R}$ denotes a "representative" value of gene cluster $\mathcal{C}_i$, defined as $x_{\mathcal{C}_i} = \frac{1}{|\mathcal{C}_i|} \sum_{r \in \mathcal{C}_i} s_r x_r$ with $s_r \in \{-1, 1\}$. Even by using such a simple group value, finding the optimal partition of thousands of genes into a few clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ is highly nontrivial and a direct, exhaustive search of the optimal solution from (1) is impracticable due to combinatorial complexity. We thus rely on optimizing an empirical

objective function which measures a cluster's strength for phenotype discrimination to find a similar partition as in (1). A quick and efficient *score* is Wilcoxon's two sample test statistic. Due to its discrete range we refine it by the continuous *margin*, the size of the gap between the two classes. Our objective function thus consists of *score* which gets first priority, and *margin* which is regarded with second priority to achieve uniqueness. The clustering process is initiated with the single gene that optimizes the objective function. We proceed by merging the current cluster in a greedy stepwise fashion with single genes, such that the objective function of the new cluster average is optimized by every acceptance. The merging is repeated until the objective function can no longer be improved. We then continue with a backward pruning stage, where genes are excluded step by step such that the objective function is optimized by every single removal.

3 Numerical Results

The algorithm was tried on 6 different datasets describing the gene expression of cancer patients. It yielded very promising output on all data since the cluster expression $x_{\mathcal{C}}$ always perfectly discriminated the cancer classes with zero *scores* and strongly positive *margins*, see figure 1. To dispel all doubt whether the clusters could be noise artifacts we ran the algorithm on random data obtained by Y shuffling. As evident from figure 1, the *margin* values from the original data are much better, which corresponds to a "for sure" rejection of the hypothesis that the clusters are artifacts.

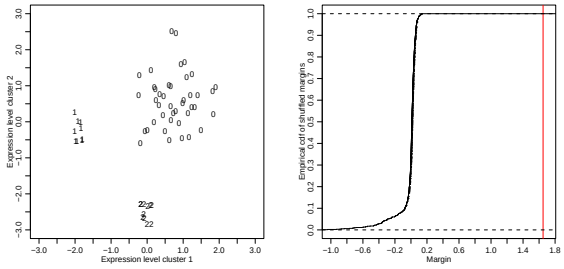

Fig. 1. 2-dimensional projection of patients suffering from 3 prevalent adult lymphoid malignancies with cluster expression profiles (left); and ecdf of *margins* based on 1000 response shuffling trials for the same data. The vertical line is the *margin* obtained from the original data (right).

The stability of the gene clusters under similar input is critical for interpretation in functional genomics. Bootstrapping showed that the algorithm can be judged as reasonably stable, yielding clusters containing 3–9 genes (with low variability) from a small core of less than 0.5% of all genes. An important task in medical diagnostics is to determine a new patient's cancer type. By using our cluster expression profiles with the simple 1-nearest-neighbor classifier, we observed excellent predictive potential, often better than for sophisticated classifiers based on single genes.