Katarzyna Stąpor, PhD, DSc
Institute of Computer Science, Silesian University of Technology
Paweł Błaszczyk, MSc, Adrian Brueckner, MSc
Institute of Mathematics, Silesian University

Discriminant analysis and multiple hypothesis testing for the classification of high-dimensional microarray data

We have developed the new method to analyze data from DNA microarray experiments which is composed of: 1) variable selection, i.e. the identification of 'marker' genes that characterize the different tumor classes using multiple hypothesis testing procedure; 2) the discriminant method—the support vector machines (SVM) for learning decision functions.

Gene expression data on p genes for n tumor mRNA samples may be summarized by an nxp matrix $X = (X_{ij})$, where x_{ij} denotes the expression level of gene (variable) j in mRNA sample (observation) i. Each observation i is a gene expression profile: $x_i = (x_{i1}, \dots, x_{ip})$. In our method, each variable is taken individually and a relevance score, a p-value, is computed. Microarray data analysis is an extreme multiple testing situation, since p hypotheses H_j are tested simultaneously, where H_j denote a null hypothesis of equal treatment and control mean expression levels for gene j. The t-statistic for each gene j is:

$$t_j = \frac{\overline{x}_{2j} - \overline{x}_{1j}}{\sqrt{\frac{s_{1j}^2}{n_1} + \frac{s_{2j}^2}{n_2}}},$$

where $\overline{x}_{1j}$, s_{1j}^2, $\overline{x}_{2j}$, s_{2j}^2 denote the average expression level of gene j and the variance of gene j's expression level in the n_1 control and n_2 treatment hybridizations. We are not assuming that the t-statistics actually follow a t-distribution, rather we use permutation to estimate their distribution. In our method, p-values are computed using the Benjamini-Hochberg [1] correction procedure to control the false discovery rate. The variables are then ranked according to their p-value score and SVM classifier [2] is used to select the best amount of genes among the top-ranking genes.

Having a training set $S = (x_i, y_i, 1 \leq i \leq N)$ composed of the examples $x_i \in R^n$, each belonging to a class labeled by $y_i \in \{1, -1\}$, the decision function of the SVM classifier [2] is as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} y_i \alpha_i K(x_i, x) + b\right),$$

where N_s is the number of support vectors, α_i, b are constants, all determined through the numerical optimization during learning, $K(.)$ is a kernel function.

Our method was demonstrated on the data set composed of gene expression in two types of acute leukemias from recently published cancer gene expression studies as well as on the simulated data sets.

References

[1] Y. Benjamini, Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. R. Statist. Soc. 57 (1995), 289–300.

[2] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York 1998.