

A nonparametric algorithm for identifying informative genes from microarray data

Fazil Alioglu Aliev, Ph.D.
Baskent University, Faculty of Engineering,
Ankara, Turkey

Abstract:

Microarray data routinely contain gene expression levels of thousands of genes. In the context of medical diagnostics, an important problem is to find the genes that are correlated with given phenotypes. These genes may reveal insights to biological processes and may be used to predict the phenotypes of new samples. In most cases, while the gene expression levels are available for a large number of genes, only a small fraction of these genes may be informative in classification with statistical significance.

We introduce a new nonparametric test statistics based on order statistics. Since our test operates in rank-transformed data, it appears to be a robust choice for microarray data, which are often nonnormal and contain outliers. The exact and asymptotic distributions of this statistics are computed using advanced combinatorics techniques under null hypothesis assumption. Our statistics assigns a score to each gene based on samples with known classes. With respect to our method, we can find a small set of genes, which are informative of their class, and subsequent analysis can be carried out with this set. We study the properties of new algorithm and apply it to the simulated data. The comparison with the known nonparametric tests as the Wilcoxon rank sum test and nonparametric t-test are presented.

Other possible applications to medical data, two-sample problem and also modifications and extensions of new statistics will also be discussed.