Michał Marczyk
Institute of Automatic Control, Silesian University of Technology
e-mail: Michal.Marczyk@polsl.pl
Roman Jaksik
Institute of Automatic Control, Silesian University of Technology
Joanna Polańska
Institute of Automatic Control, Silesian University of Technology
Andrzej Polański
Institute of Informatics, Silesian University of Technology

Discriminative gene selection in low dose radiotherapy microarray data for radiosensitivity profile search

In radiotherapy total dose delivered to targeted tumor tissue is limited to minimize late side effects in normal tissue, which also limits its healing effect. Ability to adjust the dose to the individual patient radiosensitivity with the use of information given after low dose radiation will help in reducing the negative effects of radiotherapy while increasing the efficiency of cancer treatment. In most gene expression studies selection of significant features for sample classification is a common task. The main goal of this step is to discover the smallest possible set of genes that allows to achieve good predictive performance. However, in analysis of cancer patients radiosensitivity, differences between analyzed groups are hardly noticed. Also clinical observations indicate large variations between individuals within group, which provides a need to explore different methods of feature selection.

Examined data contain two groups of breast cancer patients showing clinical differences in their normal tissue late response to radiotherapy. Data pre-processing includes probe sets re-annotation using PLANdbAffy database, tRMA background correction, normalization and summarization. Preliminary data analysis and quality control pointed out strong batch effect, which was corrected using ComBat software.

To select significant genes, which can predict the status of the sample on the basis of the expression profile, we use statistical methods (t-test, modified Welch test, F-test) and recurrent feature replacement methods (Recursive Feature Elimination, fuzzy C-Means RFE). In statistical methods correction due to correlation between genes was applied. We perform comprehensive experiments to compare feature selection algorithms using two classifiers as SVM, with linear and nonlinear kernel, and Naive Bayes. The validation step was divided into 2 stages. Training pilot study patient set, which in opinion of clinicians was more informative, and testing set, which contained the rest of samples, were used to see if there exist gene signature related to radiosensitivity. Multiple random validation procedure using all data was later performed to prove generalizability of selected features.

As a result of applying the above described algorithms, it was possible to construct a classifier that could discriminate patients based on their late response to radiotherapy treatment with 25% error rate using SVM and nonlinear kernel. This result was proven through multiple random validation. When comparing methodologies of feature selection recruitment modified Welch test which deals with unequal variability of genes between groups performed best, however only with correction due to correlation.

This work was supported by the European Program FP6 - 036452, GENEPI-lowRT and Ministry of Science and Higher Education grant no N N519 647840.