Multiclass classification system for large-scale cancer genomics data

Sebastian Student[1] **and Krzysztof Fujarewicz**[1]
[1] Silesian University of Technology, Poland

Recent studies suggest that molecular profiles may represent a promising alternative for clinical cancer classification. With the recent developments in biology it is feasible to measure the expression of thousands of molecular tissue biomarkers. For example we can use data of gene expression measured by DNA microarrays or RNA-Seq technique, DNA methylation levels measured by DNA methylation microarrays or protein and phosphoprotein levels measured by reverse phase protein array. Molecular-based approaches have opened the possibility of investigating the molecular changes in many diseases, for example to classify different cancer subtypes. A big problem in applying large-scale genomic and proteomic data for classification problem is the dimension of this data (Nguyen et al.,2002). In most cases standard statistical methodology does not work well when in the classified data are more variables than samples. The key aspect is to build a good tool for classification problems that use only the small subset of most important features (Student et al., 2012). In this paper we describe multiclass feature selection and classification system using RNA-Seq gene expression data, DNA methylation levels data and protein and phosphoprotein expression data. We have used different dataset combinations and compared the feature sets selected in the system and also the classification accuracy. As the selection method we have used a gene selection methods based on different methods: Multivariate Partial Least Squares (MPLS) (Hskuldsson et al., 1988), GS method (Yang et al., 2006) and method based on t-test. In the standard way PLS algorithm is used only for dimension reduction and not for selecting significant genes. The new idea is to use PLS method for multiclass feature selection in different ways: multiclass approach, and also as binary selections that use one versus rest (OvR) and one versus one (OvO) methods. In order to classify cancer data samples the support vector machines (SVM) and linear discriminant analysis (LDA) technique is used. Bootstrap resampling used in this study give us the opportunity to sort the probes and features by the importance in classification problem, for example using the Bootstrap Based Feature Ranking BBFR plot (Fujarewicz et al., 2005). For error estimation we have used balanced bootstrap 632+ (variance-reducing bootstrap 632+) which is relevant for small sample data sets. In this study we have used publicly available data from The Cancer Genome Atlas Data Portal (TCGA). Our results have shown that our feature selection and classification system used with large-scale data can improve significantly the classification accuracy rate in combined genomic and proteomic data. Calculations were carried out using the computer cluster Ziemowit (http://www.ziemowit.hpc.polsl.pl) funded by the Silesian BIO-FARMA project No. POIG.02.01.00-00-166/08

in the Computational Biology and Bioinformatics Laboratory of the Biotechnology Centre in the Silesian University of Technology.
This work was supported by grant ZIS - BK219/Rau1/2014,t.3 Silesian University of Technology in Gliwice.

Keywords: Multiclass classification, Feature selection, Suport Vector Machines, RNA-Seq

References:

Student S., Fujarewicz K. (2012). Stable feature selection and classification algorithms for multiclass microarray data. *Biology Direct 7:33*.

Nguyen DV., Rocke DM. (2002). Tumor classification by partial least squares using microarray gene expression data *Bioinformatics 18(1)*, *39-50*.

Hskuldsson A. (1998). PLS regression methods *J Chemom 1988*, *2(3)*, *211-228*.

Yang K., Cai Z., Li J., Lin G. (1998). A stable gene selection in microarray data analysis. *BMC Bioinf*, *7:228*.

TCGA The Cancer Genome Atlas Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways *Nature 455*, *10611068*.

Fujarewicz K. (2007). A multigene approach to differentiate papillary thyroid carcinoma from benign lesions: gene selection using bootstrap-based Support Vector Machines *Endocrine - Related Cancer 14*, *809-826*.