Title:
Significant gene selection in cancer analysis and diagnosis by penalized logistic regression

Abstract:
Today, cancer is one of the main causes of mortality among human beings that microarray technology plays an important role in diagnosis, treatment, and classification of cancer tissue. The high number of genes and small number of samples have caused some problems in the design of classifiers. Thus, reducing the size, and selection of a small number of genes with high prediction accuracy for classifying cancer is a significant challenge in microarray data analysis. Microarrays are quick methods to simultaneously examine thousands of genes or proteins in a single experiment. With this technique, we can examine the change in gene expression in diseases, conditions, duration of treatment, etc. Studying gene expression during disease can provide new and better therapies. In this thesis, a new method for selecting genes affecting cancer is provided. At first, the correlation between genes and the class label are determined by Pearson correlation coefficient (PCC), and genes are divided into two groups, then genes probability of belonging to the label classes (cancer tissue and healthy tissue) are determined by Penalized logistic regression (PLR). Based on the average results of PLR genes are sorted in descending order and at last, 100 most influential and important breast cancer genes are selected to classify samples. The proposed model is based on two sets of microarray information of colon cancer and breast cancer. Some gene selection methods such as PCC, SNR, SVM-RFE with classifiers SVM (lin) and SVM (rbf) were also implemented and compared with the proposed model. In the proposed model, for assessing the selected genes, 10-fold external cross-validation and B632 + have been used. The results indicated an improvement in the classification performance and reducing the number of genes. In addition, the collection of selected genes was tested on independent samples in each microarray datasets. This shows high accuracy in classification.

Keywords:
gene selection, classification, penalized logistic regression (PLR)