A combinatory algorithm of univariate and multivariate gene selection.

ABSTRACT

Microarray technology has provided the means to monitor the expression levels of a large number of genes simultaneously. Constructing a classifier based on microarray data has emerged as an important problem for diseases such as cancer. Difficulty arises from the fact that the number of samples are usually less than the number of genes which may interact with one another. Selection of a small number of significant genes is fundamental to correctly analyze the samples. Gene selection is usually based on univariate or multivariate methods. Univariate methods for gene selection cannot address interactions among multiple genes, a situation which demands the multivariate methods [1], [2]. In this paper, we considered new parameters which come up from singular value decomposition and present a combination algorithm for gene selection to integrate the univariate and multivariate approaches and compare it with gene selection based on correlation coefficient with binary output classes to analyze the effect of new parameters. Repeatability of selected genes is evaluated by external 10-fold cross validation whereas SVM and PLR classifiers are used to classify two well known datasets for cancers. We calculated the misclassification error in training samples and independent samples of two datasets (breast cancer and Leukemia). The results show that the mean of misclassification error of training samples in 100 iteration are almost equal in two algorithms but our algorithm have the better ability to classify independent samples.

Keyword: Singular value decomposition; Penalized logistic regression; Gene selection; Multivariate analysis; Algorithm.