Title

Two-Stage Multinomial Classification with ensemble classifier of SVM and RF for High-Throughput Data

Authors

Abhijeet R Patil[1] and Sangjin Kim[2, 3]

[1]Computational Science Program, [2]Department of Mathematical Sciences and [3]Border Biomedical Research Center, The University of Texas at EL Paso, El Paso, TX

Abstract

Recently, high dimensional classification methods have been popularly studied with the advent of high-throughput technology. However, Selection of best single algorithm is difficult amongst several popular competing algorithms for various reasons. This paper proposes a two-stage approach for multi-class classification with filtering of variables and applying those variables which pass the filter to a proposed ensemble classification. In the first stage we use marginal statistical tests to filter informative variables based on familywise error rate correction. These variables are input to the proposed method, ensemble of random forest (RF) and support vector machines (SVM) combined which is most popular as nonparametric methods to predict classification at second stage. The nonparametric methods are less sensitive to highly correlated data structure. The proposed ensemble method is implemented in R and utilizes accuracy, sensitivity and specificity as metrics for determining performance. We show that our proposed method for multinomial classification has better performance of prediction on the test set of samples compared to individual algorithm of RF and SVM which tries to directly find the optimal classifier based on performance of individual classifiers from the training set of samples in the application of several high dimensional cancer gene expression data.