Data Mining and Machine Learning

Priyanka Pandey

SBLMS Education Institute, Rajapuri, New Delhi - 110059

Address for correspondence:
E-mail: priyankapand@gmail.com

We have analyzed the problems of building binary classifiers for cancer classification Based on gene expression.with the development of micro array technology Scientists. Can now measure the expression levels of thousands of gene simultaneously in one single experiment. That later enabled the biologists in collecting gene expression for a large number of samples. One of urgent issues in the use of micro array data is to develop methods for characterizing samples based on gene expression. To evaluate the effectiveness of the cancer classification methods, two criteria mutational may be used, i.e the classification accuracy and the number of gene used by the classifier. For a cancer classifier, the fewer the genes used,lower the computational burden. a reduced number of gene can significantly increase the classification accuracy because of the reduction or the absence of irrelevant genes actingas"noise" for the classifier. Perhaps more importantly, once smaller subsets of genes are identified as revelent to a particular cancer, it helps biomedical researchers focus on these genes that contribute to the development of the cancer.therefore, finding the smallest gene subsets that can ensure highly reliable classification results become a problem of both theoretical and practical importance. it is proposed to address this issue in this thesis. Support vector machines is the commonly used classification method used for many classification problems that provided better classification accuracy. Here we compare the classification accuracy with various feature selection methods for the two data sets in our experiment. We have also observed the variation of accuracy with the number of genes a given dataset.