Pathway-based analysis with Support Vector Machine (SVM-LASSO) for gene selection and classification

ABSTRACT

Genomic knowledge has become a popular research field in bioinformatics biological process that providing further biological process information. Many methods have been done to address the issues of high data throughput due to increased use of microarray technology. However, it is still not able to determine the appropriate diseases accurately. This is because of existing noninformative genes that could be included in the analysis of context-specific data like cancer gene expression data, which affect the classification performance. This study proposed a pathway-based analysis for gene classification. Pathway-based analysis enables handling microarray data in order to improve biological interpretation of the analysis outcome. Secondly, Support Vector Machine with Least Absolute Shrinkage and Selection Operator algorithm (SVM-LASSO) is proposed, which to find informative genes for each pathway to ensure efficient gene selection and classification in every pathway. Experiments are done using lung cancer dataset and breast cancer dataset that widely used in cancer classification area. A stratified 10-fold cross validation is implemented to evaluate the performance of the proposed method in terms of accuracy, specificity, and sensitivity. Moreover, biological validation has been done on the selected genes based on biological literature and biological databases. Next, the results from the proposed methods are compared with the previous study throughout all the data sets in terms of performance. As a conclusion, this research finding can contribute in biology area especially in cancer classification area.

Keyword: Genomic knowledge; Gene analysis; Microarray technology; Pathway-based analysis; Support vector machine; LASSO; 10-fold cross-validation; Cancer classification