Abstract No: STEM 12 Science, Technology, Engineering and Mathematics

An overview of parametric and non-parametric gene selection methods in classification approaches in microarray data

Gayathri Y. K. K. M. K.[1*], Napagoda N. A. D. N.[1]

Cancer-causing genetic changes can be inherited from parents and incompatible lifestyles. Breast, lung, colorectal, liver, cervix uteri, thyroid, and skin cancers are the most common cancers worldwide. The root causes of a cancer can be identified by using DNA microarrays. Microarray has been established as an efficient tool for gene expression profiling. Gene expression experiments help to understand the evolution of gene regulation in different organisms. The identification of differentially expressed genes and a marker genes selection are essential steps in cancer gene classification using gene expression data. Therefore, most researchers used parametric and non-parametric methods to reduce the dimension of the microarray data to detect the optimal affected genes of the disease. The parametric test must be used with a few assumptions, and non-parametric methods do not weigh the underlying assumptions of the probability distribution based on the population. The primary goal of our research is to compare parametric and non-parametric gene selection methods with various classification approaches in order to identify the best method(s) for recognizing the genes that contribute to cancer diseases such as breast, leukemia, lung, colon tumor, skin, and prostate cancer, etc. Therefore, gene selection and classification methods are used to improve the performance of the selected genes and detect the accuracy of ranked genes. In past studies, both parametric (the Two-sample t-test, Welch t-test, Euclidean distance, Analysis of Variance (ANOVA), Bayesian) and non-parametric approaches (Wilcoxon rank-sum test, Random permutations, Wilcoxon Mann Whitney (WMW) test, Significance Analysis of Microarray (SAM)) have been employed. Furthermore, the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), decision tree, Naïve Bayes, and Linear Discriminant Analysis (LDA) are well-known classification techniques, and the confusion matrix with accuracy, sensitivity, and specificity can be used to evaluate the performance of classification methods. The Two-sample t-test, WMW test, and SAM methods can be recognized as the most established gene selection methods in past studies. Furthermore, the SVM algorithm is a very effective classification technique even with structured, semi-structured, and unstructured cancer data. In addition to that, the usage of Two sample T-tests with SVM and SAM with SVM methods is simple and less time-consuming. Therefore, our study included an overview of powerful parametric and non-parametric gene selection approaches with suitable classification methods that can be used to identify the top affected genes with a high accuracy rate for identifying cancer genes. The results of past studies have found that the presence of a few top-rank genes effectively supports the identification of disease cells with appropriate, high prediction accuracy. Moreover, the results can be used for diagnostic purposes in clinical practices and patient well-being. The future study can be enhanced by applying a few hybrids of parametric and non-parametric gene selection algorithms with a suitable classification method on cancer datasets to identify differentially expressed cancer genes.

Keywords: Classification, Gene selection, Microarray, Non-parametric, Parametric

[1] Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka
* minushigayathri@gmail.com