A Feature Selection Based on One-Way-Anova for Microarray Data Classification

[*1]Arowolo, M.O., [1]Abdulsalam, S.O., [2]Saheed, Y.K. and [1]Salawu, M.D.

[1]Department of Computer Science, Kwara State University, Malete, Nigeria
[2]Department of Physical Sciences, Al-Hikmah University, Ilorin, Nigeria

Received: August 9, 2016; Revised: November 30, 2016; Accepted: December 8, 2016

Abstract

High dimensionality of microarray data and expressions of thousands of features in a much smaller number of samples is a challenge affecting the applicability of the analytical results. However Support Vector Machine (SVM) has been commonly used in the classification of microarray datasets, yet the problem of high dimensionality of the feature space of data still exist. This study deals with the reduction of gene expression data into a minimal subset of genes, by introducing feature selection, to greatly reduce computational burden and noise arising from irrelevant genes that can perform a classification of cancer from microarray data using machine learning. Various statistical theory and Machine Learning (ML) algorithms to select important features, remove redundant and irrelevant features have been proposed, but it is unclear how these algorithms respond to conditions like small sample-sizes. This paper presents combination of Analysis of Variance (ANOVA) for feature selection; to reduce high data dimensionality of feature space and SVM algorithms technique for classification; to reduce computational complexity and effectiveness. Computational burden and noise arising from redundant and irrelevant features are eliminated. It reduces gene expression data to a lesser number of genes rather than thousands of genes, which can drop the cost for cancer testing significantly. The proposed approach selects most informative subset of features for classification to obtain a high performance accuracy, sensitivity, specificity and precision.

Key words: Gene expressions, Microarray, One-Way-ANOVA, Support Vector Machines

1.0 Introduction

Deoxyribonucleic acid (DNA) microarrays suggest the ability to observe the expression of thousands of genes in a single experiment, a significant relevance of microarray gene expression is cancer classification [1]. With microarray analysis, researchers are be able to classify various diseases according to different expression levels and tumor cells, to find out the relationship between genes, and to classify the critical genes in the improvement of disease [2,3]. A major task of microarray classification is to build a classifier from past microarray gene expression data, and then use the classifier to classify future coming data. Due to the rapid development of DNA microarray analysis, gene selection and classification techniques are being computed for better use of classification algorithm in microarray gene expression data. The study of large gene expression data sets is becoming a challenge in cancer classification. Therefore gene selection is one of the significant characteristics. Efficient feature selection can drastically relieve computational load of subsequent classification tasks, and it can give a much lesser and more compact gene set with no loss of classification [2]. In classifying microarray data, the main objective of feature selection is to search for the genes, which maintains the maximum amount of information about the class and minimize the classification error. Data mining falls into either supervised or unsupervised classes [4].

The present study promises to enhance beneficial scope to cancer patients by diagnosing cancer varieties with improved accuracy. SVM is efficiently applied to the cancer classification problems [5]. SVM is a learning machine used as a tool for data classification, functions approximation e.t.c. due to its generalization ability and has establish achievement in many applications. Quality of SVM is that it reduces the upper bound of generalization error through maximizing the margin between separating hyper plane and dataset. SVM has an added advantage of automatic model selection, such that both the optimal number and locations of the basic functions are automatically obtained during training. The performance of SVM largely depends on the kernel [6].

***Corresponding Author: Tel: +234(0)8032284439, E-mail: olliray2002@yahoo.com**

In this study, a microarray Gene Expression Cancer Diagnosis is carried out using Colon cancer dataset [7]. Feature selection based on One-Way-ANOVA method was used to fetch vital numbers of gene subsets which eliminates redundant and irrelevant features. As an efficient technique, feature selection is frequently used in the analysis of microarray data to assess the significance of treatment effects and to select vital genes that enhances the performance of classification [10]. The classification of the gene subsets is carried out using SVM. This system based on One-Way-ANOVA proposes effective classification accuracy by combinational technique with reduced implementation complexity compared to the SVM-Based model.

2.0 Materials and Methods

2.1 Data Set and Methodology

The study was carried out using colon cancer dataset [7]. It contains DNA microarray gene expression data with, 2000 features and 62 samples. The computer configuration exploited for the purpose of comparative study uses iCore 2 processor, 4 GB RAM size, 64-bit System and MATLAB 2015a as the implementing tools.

The proposed approach involves two methods. The first method selects all genes in the colon cancer dataset and trains them for classification using a SVM-Radial Basis Function (SVM-RBF) for classification. The second method uses One-Way ANOVA method for feature selection, it reduces the data by selecting a subset out of the whole dataset, and classification was carried out on the reduced dataset using SVM-RBF. Our results are then evaluated in terms of accuracy, sensitivity, specificity and precision (Fig. 1).

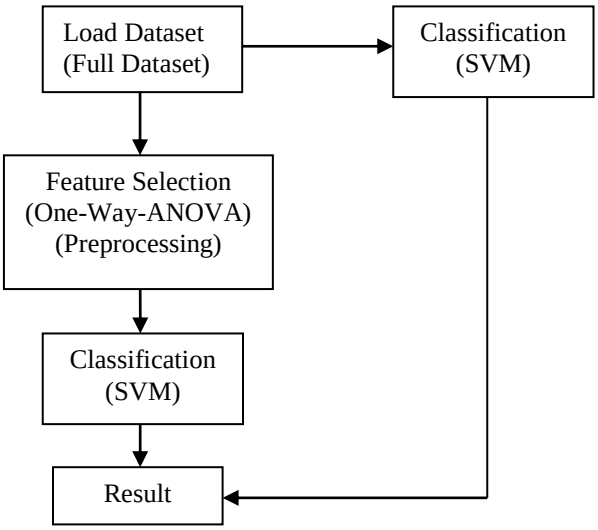

Fig. 1: Technique Workflow

2.2 Feature Selection

Statistical analyses have become more sophisticated and appropriate application of tests has improved over the past decades [8]. Several authors have incorrectly used t-test instead of ANOVA to compare features of datasets [9]. Although advancements have been made but errors still occur in the analysis of microarray. One-Way-ANOVA method for feature selection is a technique that analyzes the experimental data such that, one or more response variables are calculated under various conditions identified by one or more classification variables. It is often used in the analysis of data and drawing interesting information based on P-Value. ANOVA is a robust technique; it assumes all sample of a data to be distributed in general, having equal variance and independent [6]. This paper chooses the approach of One-Way-ANOVA, it performs analysis by comparing the given sample dataset and returns a single p-value, which is significant. In this paper the p-value is set at 0.05, any value lesser than 0.05 is effective, while any value greater than this value is non-significant. It uses the p-values to rank the important features with small values and the sorted numbers of features are used for further processing.

2.3 Support Vector Machine (SVM)

The selected feature results are classified with the use of SVM for classification. SVM is a supervised and constructive learning procedure based on statistical learning theory [11]. It is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise.

Assuming $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set with $x_{1i} \epsilon R^d$ and y_i is the corresponding target class. SVM can be reformulated as [10]:

Maximize: $J = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T, x_j)$ (1)

Subject to; $\sum_{i=1}^{n} \alpha_i y_i = 0 \ and \ \alpha_i \geq 0, i = 1,2, \ldots, n$ (2)

An advantage of SVM is the improvement of generalization performance by appropriate selection of kernel to specific significance and application [10]. The common kernels that are used in SVM are given as follows [11]:

Linear Kernel: $k(x_b x_j) = x_i . x_j$ (3)

Polynomial Kernels: $k(x_b x_j) = (\gamma(x_i x_j) + r)^d, r \geq 0, \gamma > 0$ (4)

Radial Basis Function (RBF): $k(x_i x_j) = \exp(-||x_i, x_j||^2 / 2\sigma^2) \ where \ \sigma > 0$ (5)

The RBF-kernel function method is suitable in solving classification problems; it has a sound theoretical foundation, it finds the best classification function to distinguish between members of the training data and it is less prone to over fitting [13]. The kernel function is used to solve the problem through analysis of the relationship among the data and creating complex divisions in the space [14].

3.0 Results and Data Analysis

The proposed approach has been evaluated on colon cancer microarray; the results are presented using the performance metrics of classifiers, in terms of classification accuracy, time, sensitivity, specificity and precision. The terms are defined as follows:

Sensitivity=TP/ (TP+FN) % (6)

Specificity = TN/ (TN+FN) % (7)

Accuracy = (TP+TN)/ (TP+TN+FP+FN) % (8)

Precision = TP/(TP+FP)% (9)

Where:
TP (True Positives) = correctly classified positive cases,
TN (True Negative) = correctly classified negative cases,
FP (False Positives) = incorrectly classified negative cases,
FN (False Negative) = incorrectly classified positive cases.

Sensitivity (true positive fraction) is the probability that a diagnostic test is positive, given that the person has the disease.
Specificity (true negative fraction) is the probability that a diagnostic test is negative, given that the person does not have the disease.
Accuracy is the probability that a diagnostic test is correctly performed.
Precision (Positive Predictive Value) is how many of the positively classified were relevant.

3.1 Classification of Results Using SVM-Based Framework

The proposed approach exploits microarray gene datasets of colon cancer for classification. The performance of the method is studied by classification using SVM; the details of achieved result of the experiment are shown in the confusion matrix (Fig. 2).

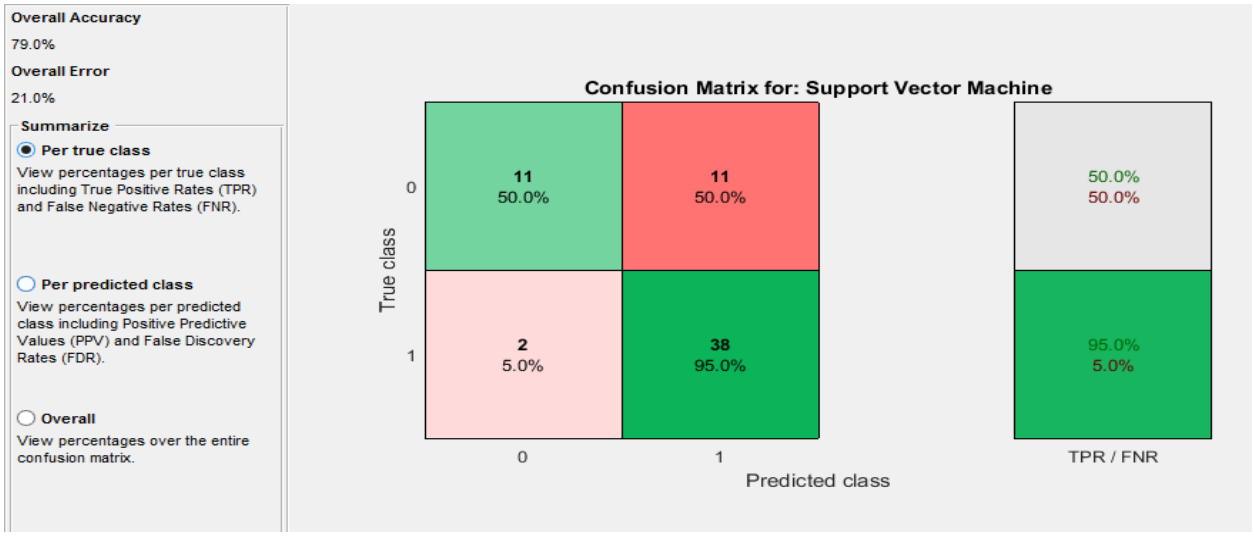

Fig. 2: Confusion Matrix of SVM for Colon cancer Dataset

TP=38 FP=11 FN=2 TN=11

3.2 Classification of Results Using One-Way-ANOVA-SVM Method

The One-Way-ANOVA method achieves necessary higher value in all the datasets on maximization parameters such as accuracy, sensitivity, specificity, and precision when compared to the classification-based method. The confusion matrix for the proposed classification using One-way-ANOVA-based for colon cancer dataset is shown in Fig. 3.

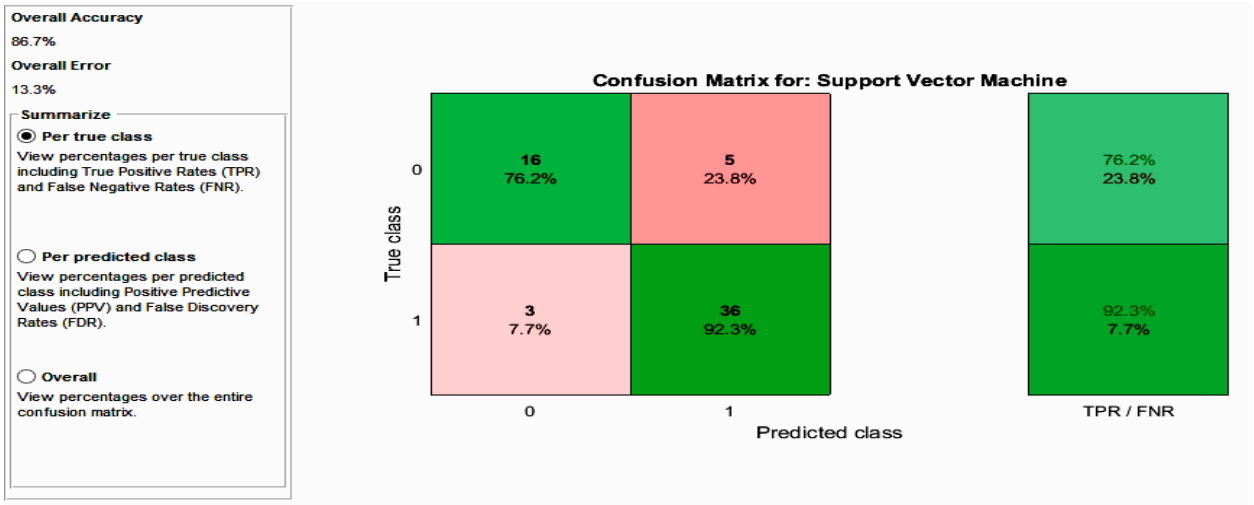

Fig. 3: Confusion Matrix of Proposed Classification, One-Way-ANOVA-Based Classification for Colon cancer Dataset

TP=36 FP=5 FN=3 TN=16

3.3 Comparative Analysis between One-Way-ANOVA-Based Method and SVM-based Method

Table 1 presents a comparative analysis between the two methods in terms of various measures such as accuracy, sensitivity, specificity and precision. The One-Way-ANOVA method achieves necessary higher value in the datasets parameters such as accuracy, sensitivity, specificity and precision when compared to the SVM-Based method. The proposed method is however relatively less effective than SVM-based method in terms of sensitivity function.

Table 1: Performance evaluation of proposed, One-Way-ANOVA–based and SVM–based methods (represented with Performance metrics)

S/No	Performance Metrics	*SVM-Based Method*	*One-Way-ANOVA-Based Method*
1	Accuracy (%)	*79.03*	*86.67*
2	Sensitivity (%)	*95.00*	*92.31*
3	Specificity (%)	*50.00*	*76.19*
4	Precision (%)	*77.55*	*87.91*

4.0 Discussion

Combination of feature selection with classification models in microarray technology play an important role in diagnosing and predicting diseases in medical research. A feature selection method (One-Way-ANOVA) for finding the most significant features is proposed. The classification accuracy rate achieved by the proposed One-way-ANOVA method using SVM classifier is 86.67% .The numbers of features are reduced from 2000 to minimum of 416 features. Experimental results on the Colon cancer datasets clearly indicate that the proposed technique has better performance compared to the SVM-Based method.

5.0 Conclusion and Recommendation

ANOVA is an effective ranking method for finding the smallest gene subsets to achieve accurate cancer classification. The result obtained with the colon dataset for gene combination gives a good separation for uniqueness. Finally, we obtained very good accuracy compared to SVM-Based method. It is therefore recommended to apply the proposed approach on more microarray data for confirmation and verification of its performance, as well as more classifiers and techniques.

References

[1] Hala, M.A., Ghada, H.B., and Yousef, A. (2013). A Study of Cancer Microarray Gene Expression Profile, In: Proceedings of the World Congress on Engineering: Objectives and Approaches, London, pp. 1324-1329.

[2] Nadir, O.F., Otham, I. and Ahmed, H.O. (2014). A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. Research Journal of Applied Sciences, Engineering and Technology, Vol. 7, pp. 625-638.

[3] Anne-Sophie, C., Alessandra, R., Pierre, T., Gilles, D. and Alain, H. (2004). The Operons, A Criterion To Compare The Reliability Of Transcriptome Analysis Tools: ICA Is More Reliable Than ANOVA, PLS and PCA. Journal of Computational Biology and Chemistry, Vol. 28, No. 1, pp.3-10.

[4] Vaidya, M. and Kulkarni, P.S. (2014). Innovative Technique for Gene Selection in Microarray Based on Recursive Cluster Elimination and Dimension Reduction for Cancer Classification. International Journal of Research in Advance Engineering, Vol.1, No. 6, pp. 209-225.

[5] Bharathi, A. and Natarajan, A.M. (2011). Cancer Classification using Support Vector Machines and Relevance Vector Machine based on Analysis of Variance Features. Journal of Computer Science, Vol. 7, No. 9, pp. 1393-1399.

[6] Santhosh, S.B. and Sasikala, S. (2010). Multicategory Classification Using Support Vector Machine for Microarray Gene Expression Cancer Diagnosis. Global Journal of Computer Science and Technology, Vol. 10, No. 15, pp. 31-37.

[7] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. *et al*. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, In: Proceedings of the National Academy of Sciences: USA, Vol. 96, No. 12, pp. 6745–6750.

[8] Lillian, S. and Charles, E. (2008) Analysis of Variance: Is There a Difference in Means and What Does It Mean. Journal of Surgical Research, Vol. 144, No. 1, pp. 158–170.

[9] Kurichi, J.E. and Sonnad, S.S. (2006). Statistical methods in the surgical literature. Journal of the American College of Surgeons, Vol. 202, pp. 476-484.

[10] Mallika, R. and Saravanan, V. (2014). An SVM based Classification Method for Cancer Data using Minimum Microarray Gene Expressions. World Academy of Science, Engineering and Technology, Vol.62, No. 99 pp. 543-547.

[11] Vapnik, V.N. (1998). Statistical Learning Theory. New York: John Wiley & Sons.

[12] Manju, B. and Agrawal, R.K. (2011). Optimal Decision Tree Based Multi-class Support Vector Machine. Informatica, Vol. 35, pp. 197-209.

[13] Hetal, B, and Amit, G. (2012). Comparative Study of Training Algorithms for Supervised Machine Learning. International Journal of Soft Computing and Engineering, Vol. 2, No. 4, pp. 137-151.

[14] Isabelle, G., Jason, W., Stephen, B. and Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning, Vol. 46, pp. 389-422.