Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/120408

A Novel Feature Selection Algorithm for Coronary Artery Disease Prediction

K. Uma Maheswari[1,*] *and A. Valarmathi*[2]

[1] Department of Information Technology, Anna University, BIT Campus, Tiruchirappalli–620024, India
[2] Department of Computer Applications, Anna University, BIT Campus, Tiruchirappalli–620024, India

Received: 12 Apr. 2018, Revised: 19 Jun. 2018, Accepted: 20 Jun. 2018
Published online: 1 Jul. 2018

Abstract: Cardiovascular disease is the predominant cause of death throughout the world. In the present era, many diseases are caused by gene transformation. Asian Indians have a higher chance of having cardiovascular disease as compared to any other global population. Identifying relevant and candidate genes for classification of samples is a tedious task when dealing with gene expression data analysis. The objective of this paper is to find the relevant genes responsible for causing coronary artery disease. In this paper we developed a novel feature selection algorithm based on fold change and p-value. Instead of selecting genes randomly our proposed method selects the top ranking candidate genes responsible for coronary artery disease. The selected differentially expressed genes from the feature selection phase are evaluated using the proposed ensemble classifier. The classifier used in this work are support vector machine, neural network and naïve bayes. The proposed framework is validated by experiments on three publicly available microarray datasets. The results clearly show that the proposed ensemble classifier performs better when compared to other classifiers. The selected candidate genes are used for carrying out diagnostic tests and for classifying the patients, which reduces the cost and also improves the accuracy.

Keywords: Coronary Artery Disease, Support Vector Machine, Naïve bayes, Neural Network, Candidate genes, Ensemble Classifier.

1 Introduction

Cardiovascular disease is one of the main cause of death in human life, and is subjective by both environmental and genetic factors [1]. With the advancements in microarray tools and technologies it is possible to predict and diagnose heart disease using microarray DNA data by analyzing blood cells itself. It has also proven that blood cells can also provide useful genomic information for cardiovascular ailments. The World Health Organization(WHO)has estimated that mortality rate due to heart disease is about 12 million globally [2]. 25% of deaths occurs due to heart disease for people in the age group of 25 to 69 years. Genomics is a method to study thousands of genes in an organism all at one time. A microarray is a huge collection of spots that contain massive amounts of compressed data. Each spot (one gene) of a microarray contains a unique DNA sequence.

The DNA microarray generates gene expression data. A microarray database is a repository which contains microarray gene expression data. Microarrays are the latest technology to find the expression levels of many genes at the same time in tissues. Significant information can be extracted from these genes by using machine learning techniques. Clinical microarray data can be analyzed from different perspectives.

One is classification; that is in order to make predictions and using clustering for finding out different classes. Next one is about performing dimensionality reduction which is otherwise known as feature extraction. A microarray gene expression data set can be represented in a matrix form, in which each row represents the gene and each column represents the sample. The cell of the matrix is the measured expression level of a particular gene in a sample. By analyzing the gene expression data the genes which are responsible for causing diseases are identified.

The microarray gene expression dataset has large set of genes and lesser set of samples. In order to overcome this problem, feature selection is needed. Feature selection is one of the preprocessing step used for eliminating irrelevant features for classification.

[*] Corresponding author e-mail: umaravi03@gmail.com

© 2018 NSP
Natural Sciences Publishing Cor.

Feature selection is used in many fields because it reduces the severe effects of curse of dimensionality and improves model interpretability by singling out the most relevant features.

Thanyaluk et al. [3] developed the computational method which performed robustly and accurately. Z-Score analysis was used to identify finest set of candidate genes that categorize the data. Ali Anaissi et al. [4] designed BIRF (Balanced Iterative Random forest algorithm) for selecting the most significant genes for identifying the disease with gene expression microarray data. He showed that BIRF approach outperforms the state of the art methods. Ramon et al. [5] used random forest for classifying microarray data and also proposed a new gene selection method which is based on random forest. Yanshi et al. [6] used moderated t-test to figure out the differentially expressed genes.

Feng Yang et al. [7] discussed and analyzed multi-criterion fusion for feature selection for data which is of high dimension and developed a feature subset selection algorithm for MCF-RFE and showed that MCF-RCE is better in classification performance. Jihong Liu et al. [8] proposed a hybrid method for feature selection which is the combination of both filter, wrapper method and showed that the hybrid approach performs better classification.

Revathy et al. [9] proposed a new method of GA-SVM which is a wrapper approach for ranking the genes and also used for classification. Lin-kai Luo et al. [12] proposed an algorithm ISVM-RCE to find feature genes and showed that it reduces the time for selecting candidate genes. Han-Yu Chuang et al. [17] described a method for selecting relevant genes and showed that RAC(Ranking and Combination analysis) was robust-efficient approach for identifying relevant genes in microarray gene data.

Inaki Inza et al. [20] compared the filter and wrapper approach for gene selection. Smitha et al. [24] described framework called Clustering-based feature selection and concluded that the framework produces optimal feature subset by eliminating irrelevant features and showed the framework improved classification performance. Huijuan et al. [26] proposed a new hybrid feature selection algorithm which combines mutual information maximization and the adaptive GA which reduces the dimensionality problem of gene expression data.

Milos et al. [28] proposed a temporal minimum redundancy and maximum relevance feature selection approach capable of handling multivariate temporal data and achieved improvement in accuracy. Sai Prasad et al. [32] introduced a distributed feature selection strategy that can be applied for complex high dimensional datasets and also increased the performance of the classification.

The objective of this work is to find the disease causing genes for Coronary Artery Disease and to show that feature selection increases the performance of the classifier. This paper is ordered as follows: Section 2 provides an overview of methods used, Section 3 explains the proposed methodology and Section 4 represents experimental results and the conclusion is given in Section 5.

2 Methods

This section gives a detailed overview of the methods used for finding informative genes and regarding ensemble classification.

2.1 Data Collection

The gene expression dataset was taken from the GEO website. GEO is a publicly available functional genomics data repository. The three publicly available gene expression datasets GSE12288, GSE9820, GSE20681 are used.

2.2 R Language

R is the open source language and Rstudio is the Integrated Development Environment (IDE). It is used to do manipulation and analysis of various data in the datasets. Various plots can be made using R language and it is utilized for software development activities in data mining, machine learning and in various fields. It is an effective, extensible and ample environment for various statistical computations and graphics. The key features of R language are that it supports user-created R packages and we can import data containing variety of file formats such as CSV (Comma Separated Values), XML(), binary files.

R language has various data structures. It includes matrices, arrays, vectors, lists and data frames. There are many packages available for R and we can use the package whenever we are in need by using library (package name) command.

2.3 Hierarchical Clustering

Hierarchical clustering is used for identifying groups in the dataset. It is used to form a cluster tree to represent data. Dendrogram is a tree diagram which shows relationships between similar sets of data. In hierarchical clustering, the data objects can be represented in hierarchy form which is useful for visualization and data summarization. Hierarchical clustering can be categorized into probabilistic methods, algorithmic methods and bayesian methods. In bioinformatics, dendrogram is used to show clustering between genes or samples.

© 2018 NSP
Natural Sciences Publishing Cor.

2.4 Log Transformation of Data

Log transformation makes highly-skewed distributions less skewed. It makes the relationships clear and makes the pattern more visible. It describes the relationships between logs and the geometric mean. It is used for normalizing the data. Gene expression levels are heavily slanted in linear scale. In order to make highly skewed distributions less skewed log transformation is used. Genes with low expression levels has values in the range of 0 to 1 and other half has values in the range of 1 to positive infinity. Log transformation makes the data makes more symmetrical and therefore a parametric statistical test provides answer with more accuracy and relevancy.

2.5 Fold Change (Biological Significance)

Fold change represents a measure. This describes how a quantity changes from a starting value to end value. The $\log_2 FC$ is calculated by taking the expression levels of each genes across two conditions (i.e, treatment and control).

$$\begin{aligned}\log_2 \text{ fold-change } &= \log_2(FC)\\ &= \log_2 \left\{ \begin{matrix}\text{of ratio of treatment}\\ \text{and control data}\end{matrix} \right\}\\ &= \log_2 \text{ (treatment / control)}\end{aligned}$$

$$\text{Fold Change} = 2\left|\log_2 \frac{\text{avg } T}{\text{avg } N}\right|$$

2.6 *t*-test (Statistical significance)

t-test is one of the parametric method, it is used to verify the two groups by comparing the mean values. It is used when the samples fulfill the conditions like normality, equal variance and independence. Independent *t*-test and paired *t*-test are the two types of *t*-test. Here, we used independent *t*-test because it involves comparison of two groups that are independent to one another. *p*-value is used to measure the strength of evidence against the null hypothesis. We choose the features that has *p*-values less than 0.05. Sample means, and sample variances of Y_{jk} for gene j have the following two conditions:

$$\bar{Y}_j(1) = \frac{\sum_{k=1}^{K_1} Y_{jk}}{K_1}$$

$$\bar{Y}_j(2) = \frac{\sum_{k=K_1+1}^{K_1+K_2} Y_{jk}}{K_2}$$

and

$$S_j^2(1) = \frac{\sum_{k=1}^{K_1}(Y_{jk} - \bar{Y}_j(1))^2}{K_1 - 1}$$

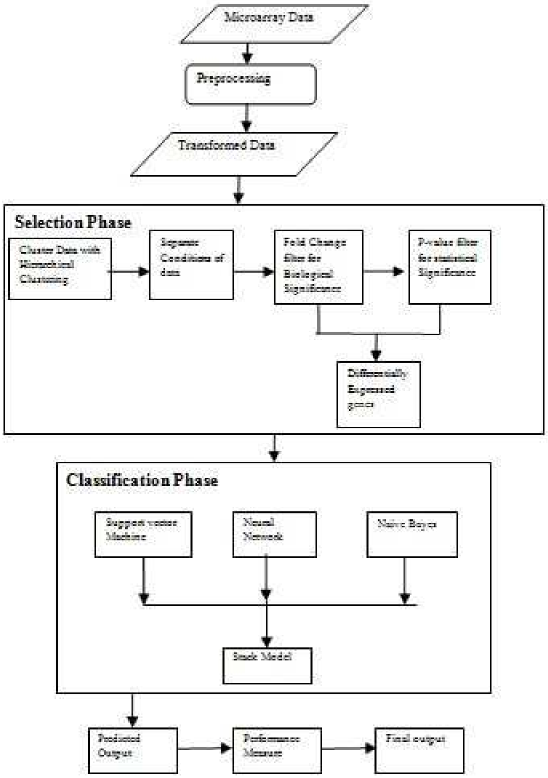
Microarray Data
Preprocessing
Transformed Data
Selection Phase
Cluster Data with Hierarchical Clustering
Separate Conditions of data
Fold Change filter for Biological Significance
P-value filter for statistical Significance
Differentially Expressed genes
Classification Phase
Support vector Machine
Neural Network
Naive Bayes
Stack Model
Predicted Output
Performance Measure
Final output

Fig. 1: Architecture of the proposed approach

$$S_j^2(2) = \frac{\sum_{k=K_1+1}^{K_1+K_2}(Y_{jk} - \bar{Y}_j(2))^2}{K_2 - 1}$$

The formula for t-statistic is given as

$$Z_j = \frac{\bar{Y}_j(1) - \bar{Y}_j(2)}{\sqrt{\frac{S_j^2(1)}{K_1} + \frac{S_j^2(2)}{K_2}}} \tag{1}$$

3 Proposed Methodology

The objective of this work is to identify the differentially-expressed genes i.e. marker genes which are responsible for coronary artery disease. The proposed method has two main phases. Feature selection using biological significance and statistical significance is done in the first phase and ensemble of classifiers using stacking is done in the second phase. Fig. 1 shows the clear view of proposed methodology.

The microarray data is taken as input from the GEO website and the microarray data is preprocessed and log transformation of the data is performed.

In the first phase i.e., selection phase, the conditions of the two classes like control and case is separated. For each group, mean values are calculated and then

© 2018 NSP
Natural Sciences Publishing Cor.

maximum of means is determined. Fold change for biological significance is calculated and statistical significance is calculated using t-test. Fold change Filter and p-value filter is applied for filtering the genes and then filter values are combined to get the final differentially-expressed genes. In second phase, i.e, the classification phase, ensemble of classifiers, i.e., stack model is built by combining the predictions of three classifiers. Then the output is validated in terms of performance measures like accuracy, sensitivity, and specificity.

3.1 Feature Selection using Biological and Statistical Significance

Feature subset selection works by removing features that are not relevant. The feature selection method is used for selecting the top genes which are responsible for causing CAD. In this algorithm, fold change and t-test are mainly used for measuring the gene relevancy for samples having two classes(with CAD and without CAD). In all the previous work they used either the fold change or t-test for identifying genes which are differentially expressed, but in our proposed methodology, we combine the t-test and fold-change results for finding differentially-expressed genes.

In general t-test selects genes with less standard deviation values and fold-change chooses genes with large differences between healthy and diseased conditions. From that, we conclude that our selected genes improves classification performance. The following steps are used for identifying differentially-expressed genes from the gene expression dataset which is shown in Fig. 2.

3.2 Ensemble of classifiers using stacking

Ensemble technique is used to develop a predictive model by combining various models. In this proposed methodology three classifiers: support vector machine, neural network and naïve bayes are integrated to produce a stacking model which is an ensemble technique. The differentially-expressed genes that are obtained in the feature selection phase are given as input to the classifiers and the performance of the individual classifiers is recorded and the three classifiers are integrated to produce an ensemble model. The results show that the ensemble model performs better when compared to individual classifiers. The classification phase is shown in the Fig. 1.

3.2.1 Differential Expression

A gene is supposed to be differentially-expressed if there is any difference in read counts among two experimental conditions which is statistically significant.

```
Procedure $FSA(f,G,S)$
  // Initialization
  DE[] ← { }
      // Differentially Expressed genes
  $D = \{f,G,S\}$
      // Input data set, $f = \{f_1,\dots,f_n\}$,
      // $G = \{G_1,G_2,\dots.G_m\}$,
      // $S = \{1,\dots,k\}$, $k = 2$
      // Normalization (log transformation)
  $\mathrm{norm}(G) \leftarrow G_n$
  $\log_2(G) \leftarrow G_n$
      // Clustering
  $HC \leftarrow \{hc_1,hc_2,\dots,hc_k\}$
    // Hierarchical Clustering
    // Partitioning data
  $G \leftarrow G_c + G_t$
    // $G_c$–Control gene, $G_t$–treatment gene
    // Finding mean
  $m_c = \sum_i xc_i/nc$ // for controlled genes
  $m_t = \sum_i xt_i/nt$ // for treatment genes
    // Max of mean
  $\arg\max_i(m_c,m_t) \leftarrow m_{\max}$
    // Fold change
  $F_c = m_c - m_t$
    // statistical test
  for each gene $G_i \in G$ do
    $t = (m_c - m_t)/\sqrt{S_c^2/n_c + S_t^2/n_t}$
    // calculate $t$ value and also $p$–value
  end for
  // select differentially expressed genes
  $F_f = \mathrm{abs}(F_c) > F_T$ // $F$ threshold
  $F_P = P_r < P_T$ // $P_T = 0.01$
  return DE // DE[] $\leftarrow \{G_1,\dots,G_n\}$
```

Fig. 2: Pseudo-code of the proposed feature selection algorithm

We have: $C_1,\dots,C_m$: Normal samples (Control)
$T_1,\dots,T_n$: Diseased samples (Treatment)

We look for: genes with significant differences between C and T and compare values of gene X from group C with those of group T,

$$C = \{c_1,\dots,c_m\}$$
$$T = \{t_1,\dots,t_n\}$$

Fold change and t-test are used. Using the proposed feature subset selection algorithm, the 25 genes are found to be differentially-expressed out of 22,282 for the dataset GSE12288. 31 genes are differentially-expressed for the dataset GSE9820 and 26 genes are differentially-expressed for GSE20861.

3.2.2 Ensemble Classifiers

The classification is the process of predicting the classes among the huge amount of dataset by using some

© 2018 NSP
Natural Sciences Publishing Cor.

Table 1: Results of differentially-expressed genes for the microarray datasets

Microarray Dataset	Number Genes	No. of Differentially expressed genes
GSE12288	22282	25
GSE9820	20589	31
GSE20861	45015	26

machine learning algorithms. Classification is the most important technique in microarray technology. This technique is used for prediction of classes among the genes or samples. Prediction plays an important role in biomedical field for disease stage prediction and drug discovery. An ensemble of classifiers is a set of classifiers whose individual predictions are combined to classify new samples. The ensemble method is used to build a new predictive model and also improves prediction performance.

3.2.3 Support Vector Machine

SVM is a learning model that is supervised in nature. It is used for regression and classification. SVM classifier is used to classify the control and CAD groups based on gene expression levels from microarray data. SVM learns from the training dataset and makes correct predictions on the testing data. SVM is popular because its performance is good on high dimensional datasets. The formula for SVM classifier is given as:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0.1-y_i(w\cdot x_i-b)\right)\right]+\lambda\|w\|^2 \tag{2}$$

Here, w–denotes the normal vector for hyper plane
λ–represents the tradeoff between increasing the margin-size and
x_i-denotes support vectors.

In our classifier, λ produces very hard margin to separate the variables. Here, hard margin is used because our variables are linearly separable. These are the conditions that the points must lie on the correct side of the margin is given in the following equations:

$$\mathbf{w}\cdot\mathbf{x}_i-b\geq 1, \qquad \text{if } y_i=1 \quad \text{or} \tag{3}$$

$$\mathbf{w}\cdot\mathbf{x}_i-b\leq -1, \qquad \text{if } y_i=-1 \tag{4}$$

3.2.4 Naïve Bayes

Naïve bayes is one of the algorithm used for classification. It is based on bayes theorem, it is based on the principle that every feature is classified without considering any other feature values. The formula for finding posterior probability is given as:

$$P(C/X)=P(X/C)\cdot P(C)/P(X) \tag{5}$$

where, $P(X/C)$–likelihood, $P(C)$–class probability and $P(X)$–predictor probability

Table 2: Statistics of heart disease dataset

Datasets	Sample size	Number of Genes	Number of class
GSE12288	222	22282	2
GSE9820	153	20589	2
GSE20861	198	45015	2

3.2.5 Neural Network

It is a best model for classifying new data. One of its important feature is to adjust the weights of every feature in order to predict the correct class in learning phase itself. It handles noisy data in a better way and also classify patterns without any training. It mainly has three rules or steps. That includes: first step is due to high complexity of data in input and output numbers in respective layers, it introduces hidden layer. The second step is if the process or work is divided into number of small modules, it introduces equal number of hidden layers. The third step is setting upper bound which is based on the number of nodes in training data.

The activation a_l^j of the jth neuron in the lth layer is related to the activations in the $(l-1)$th layer which is given as:

$$a_l^j=\sigma\left(\sum_k w_l^{jk}a_{l-1}^k+b_i^j\right) \tag{6}$$

Here, l–layer,
w_l^{jk}–represents weight matrix of layer l in jth row and kth column,
b_i^j–bias vector in the l layer and
a_l^j–activation vector in the l layer.

4 Experimental Results and Analysis

4.1 Data Source

The gene expression dataset was taken from the Gene Expression Omnibus. Gene expression data are usually presented in a matrix form.

GSE12288 dataset has 222 samples and 22,282 genes. Patients who went for coronary angiography are selected based on the Duke CAD index. 110 patients are affected with CAD whose CAD$_i$ value is greater than 23 and 112 patients without CAD whose CAD$_i=0$.

GSE9820 dataset has 153 samples and 20,859 genes. 86 patients are affected with severe triple-vessel CAD and 67 patients are not affected with CAD.

GSE20681 dataset has 198 samples and 45,015 genes. Expression profiling of all blood cells from patients are taken before cardiac catheterization. From 198 samples, 99 patients have $\geq$ 50% stenosis by QCA-Quantitative coronary angiography, 99 patients have luminal stenosis of $<$ 50% by QCA.

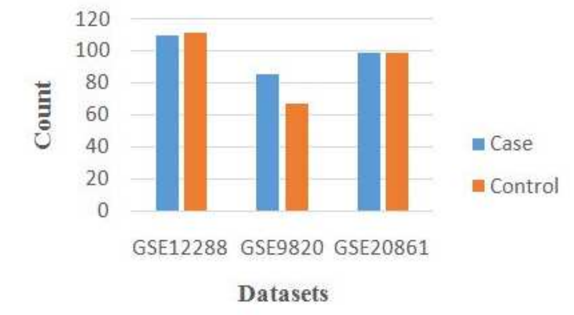

Fig. 3: Overview of the datasets of coronary artery disease

The statistics of gene dataset is tabulated in Table 2. It gives an overview of the three microarray datasets such as sample size, number of genes, and number of classes. All the datasets have control and case classes.

The overview of the coronary artery disease datasets is shown in Fig. 3. It gives information about the number of cases and controls in the gene dataset. The number of cases(diseased) are shown in blue colour and the number of controls(healthy) are shown in orange colour.

4.2 Heat Map

Heat map is used for visualization to present gene expression data, here the individual values contained in the matrix are represented as colors. It is in the form of an array where rows denote genes and columns denote samples. The heat map gives a visual summary of gene expression data. In heat map, the darker colors indicate low activity, and brighter colors indicate high activity. The difference in colors along with the row is called the expression pattern of the genes associated with them. If the genes are randomly ordered, it is difficult to interpret patterns in the heat map. Therefore, the genes are sorted so that the two genes having similar expression patterns are nearer together.

Fig. 4 shows the analysis of differentially-expressed genes between patients with coronary artery disease and controls for the dataset GSE12288. Out of 22,282 genes the 25 genes are differentially-expressed and it is shown in the heat map. For the dataset GSE9820, 31 genes are found to be differentially-expressed and for the dataset GSE20681, 26 genes are differentially-expressed.

4.3 Volcano Plot

For analyzing micro array datasets volcano plot is used, which gives an overview of interesting genes.

The *x*-axis represents the log fold change between the two conditions. The *y*-axis shows negative $\log_{10}$ of the *p*-values from the statistical test of the comparison. This shows data points with small *p* values which are highly significant, appears at the top of the plot. Fig. 5 shows the volcano plot for GSE12288. Each dot on the plot represents the individual gene. The differentially-expressed genes are shown in blue and red color. The blue color dots characterize the up-regulated genes and the red color dots characterize the down-regulated genes and the black color dots characterize non-significant genes. In the volcano plot of GSE12288, 7 up-regulated(blue color) genes are present and 18 down regulated(red color) genes are present, so totally 25 genes are differentially-expressed.

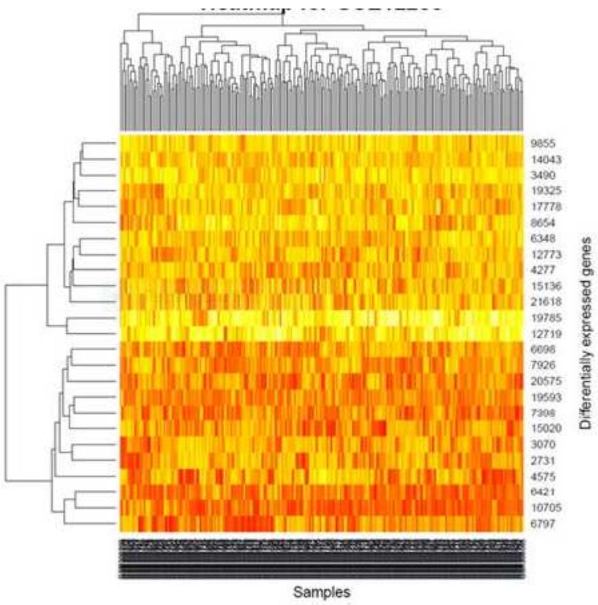

Fig. 4: Heat map analysis of differentially-expressed genes for patients with CAD and controls for the dataset GSE12288

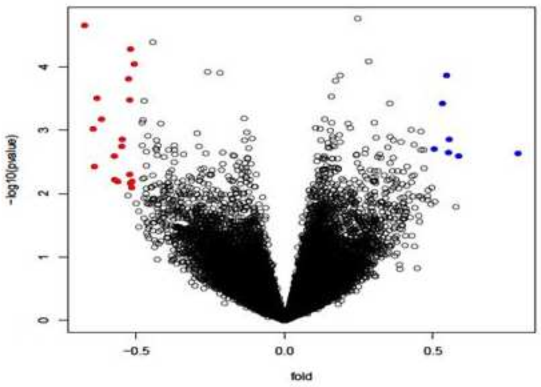

Fig. 5: Volcano plot of GSE12288

4.4 Stacking

Stacking is one method of ensemble technique. It comes under supervised machine learning. Stacking Model combines the prediction of different classifiers in order to get improved accuracy by combining the predictions of various models in order to produce optimal model. In this work, the three classifiers such as support vector machine, naïve bayes and neural network are used and their individual predictions are obtained and the predictions of these classifiers are combined using stacking for improving the performance measures like accuracy, sensitivity, and specificity. Table 3 shows that the proposed stacked combo model shows better accuracy, sensitivity, and specificity when compared to other classifiers.

4.5 Performance Measures

The performance of the work is evaluated by using some measures given below.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$
$$\text{Specificity} = TN/(TN+FP)$$
$$\text{Sensitivity} = TP/(TP+FN)$$

where,
TP =No. of samples classified as true while they were true
TN =No. of samples classified as false while they were false
FN =No. of samples classified as false while they were true
FP =No. of samples classified as true while they were false

The proposed algorithm is tested on GSE12288, GSE9820, GSE20681 datasets. The developed method was implemented using *R*. The differentially-expressed genes were given as input the classifiers like SVM, naïve bayes and neural network. Performance of the stacked combo model is compared with the individual models. Table 3 shows the accuracy, specificity, and sensitivity measures for different classifiers.

From Table 3 it is clear that the stacked combo model is better when compared to other models(SVM, NB, Neural Network)in terms of accuracy, sensitivity and specificity.

The graph in Fig. 6 shows the comparison of performance measures for the dataset GSE12288 for the different classifiers. The accuracy is shown in blue color, sensitivity in dark red color and specificity in green color. For GSE12288 dataset the accuracy is 98.25%, sensitivity is 96.43%, and specificity is 100% for stacked combo model and it is clear that stacked model performs better.

The graph in Fig. 7 shows the comparison of performance measures for the dataset GSE9820 for the different classifiers and it is evident that stacked model performs better in terms of accuracy, sensitivity and specificity. For GSE9820 dataset the accuracy is 98.11%, sensitivity is 100% and specificity is 96.15% for stacked combo model.

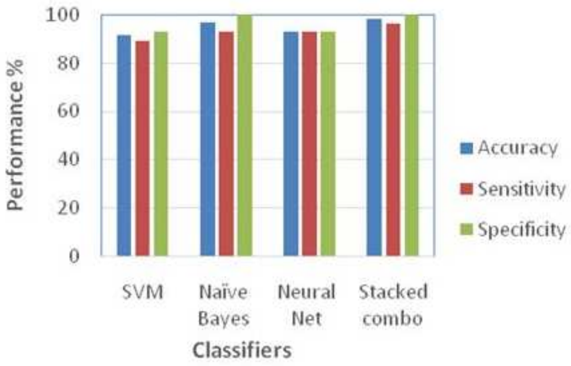

Fig. 6: Comparison of performance measures–GSE12288

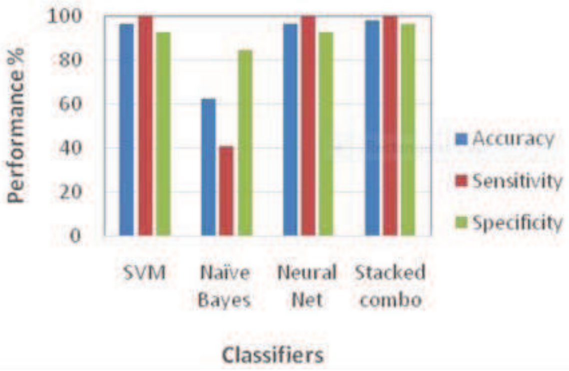

Fig. 7: Comparison of performance measures–GSE9820

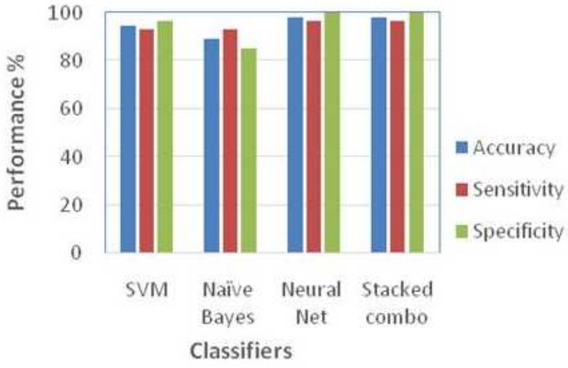

Fig. 8: Comparison of performance measures–GSE20681

The graph in Fig. 8 shows the comparison of performance measures for the dataset GSE20681 for the different classifiers. For GSE20681, dataset the accuracy is 98.18%, sensitivity is 96.43%, and specificity is 100% for stacked combo model. From the graph it is clear that the performance of stacked model is good.

© 2018 NSP
Natural Sciences Publishing Cor.

Table 3: Accuracy measure for various classifiers

Dataset	Total Features	Classifiers	Selected Features	Accuracy	Sensitivity	Specificity
GSE12288	22,282	SVM	25	91.22	89.29	93.1
		Naïve bayes		96.49	92.86	100
		Neural Net		92.98	92.86	93.1
		Stacked combo		98.25	96.43	100
GSE9820	20,589	SVM	31	96.22	100	92.31
		Naïve bayes		62.26	40.74	84.62
		Neural Net		96.23	100	92.31
		Stacked combo		98.11	100	96.15
GSE20681	45,015	SVM	26	94.55	92.86	96.3
		Naïve bayes		89.09	92.86	85.19
		Neural Net		98.18	96.43	100
		Stacked combo		98.18	96.43	100

5 Conclusion

In this work, we propose a novel feature selection algorithm to identify the marker genes responsible for causing coronary artery disease and also to show that the feature selection increases the accuracy of prediction of the classifiers. The analysis of microarray gene expression data is a challenging task due to high dimensionality, redundancy, noisy, and small sample size. Feature selection is used for removing the curse of dimensionality problem. In this work, we utilized three public microarray datasets to identify the genes which are differentially-expressed in patients with coronary artery disease and control. The proposed feature selection algorithm gives suggestions on the biomarkers for the early detection and prevention of coronary artery disease for the treatment based on the gene functionality. Subsequently, different classifiers like support vector machines, naïve bayes, neural network and stacked combo model are used to predict the diagnosis of patients with increased accuracy, sensitivity, and specificity with minimum number of attributes. Also, the observation shows that the stacked combo model outperforms the other three methods. In future, this work can be extended by incorporating fuzzy learning models and genetic algorithms to evaluate coronary artery disease prediction.

References

[1] N. Kazmi and T.R. Gaunt, Diagnosis of Coronary Heart Diseases using Gene Expression Profiling; Stable Coronary Artery Disease, Cardiac Ischemia with and without Myocardial Necrosis, PLOS ONE, **11(3)**, e10149475 (2016).

[2] R. Suganya, S. Rajaram, A. Sheik Abdullah and V. Rajendran, A Novel Feature Selection method for predicting heart diseases with Data Mining Techniques, Asian Journal of Information Technology, **15(8)**, 1314–1321 (2016).

[3] T. Jirapech-Umpai and Stuart Aitken, Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics, **6**, 148 (2005).

[4] Ali Anaissi, Paul J. Kennedy, Madhu Goyal and Daniel R. Catchpoole, A balanced iterative random forest for gene selection from microarray data. BMC Bioinformatics, **14**, 261 (2013).

[5] Ramon Díaz-Uriarte, Sara Alvarez de Andres, Gene selection and classification of microarray data using random forest. BMC Bioinformatics, **7**, 3 (2006).

[6] Yan Shi, Sijin Yang, Man Luo, Wei-Dong Zhang and Zun-Ping ke, Systematic analysis of coronary artery disease datasets revealed the potential biomarker and treatment target, Oncotarget, **8(33)**, 54583–54591 (2017).

[7] Feng Yang and K.Z. Mao, Robust Feature Selection for Microarray Data Based on Mulitcriterion Fusion, IEEE/ACM Transactions on Computational Biology and Bioinformatics, **8(4)**, 1080–1092 (2011).

[8] Jihong Liu and Guoxiong Wang, A Hybrid Feature Selection Method for Data Sets of Thousands of Variables, IEEE, **2**, 288–291 (2010).

[9] N. Revathy and R. Balasubramanian, GA-SVM wrapper approach for gene ranking and classification using expressions of very few genes, Journal of Theoretical and Applied Information Technology, **40(2)**, 113–119 (2012).

[10] Michael P.S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., and David Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, PNAS, **97(1)**, 262–267 (2000).

[11] Joseph S. Verducci, Vincent F. Melfi, Shili Lin, Zailong Wang, Sashwati Roy and Chandan K. Sen, Microarray analysis of gene expression: considerations in data mining and statistical treatment, Physiol Genomics, **25**, 355–363 (2006).

[12] Lin-kai Luo, Deng-Feng Huang, Ling-Jun Ye, Qi-Feng Zhou, Giu-Fang Shao and Hong Peng, Improving the computational Efficiency of Recursive Cluster Elimination for Gene Selection, IEEE/ACM Transactions on Computational Biology and Bioinformatics, **8(1)**, 122–129 (2011).

[13] Jaison Bennet, Chilambuchelvan Ganaprakasam, and Nirmal Kumar, A Hybrid Approach for Gene Selection and Classification using Support Vector Machine, The International Arab Journal of Information Technology, **12(6A)**, 695–700 (2015).

[14] Jing Sun and Kalpdrum Passi, Chakresh Kumar Jain, Improved Microarray Data Analysis using Feature Selection Methods with Machine Learning Methods, IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2016).

[15] Wei Pan, A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, Bioinformatics, **18(4)**, 546–554 (2002).

[16] Trey Ideker, Vesteinn Thorsson, Andrew F. Siegel, Leroy E. Hood, Testing for Differentially-Expressed Genes by Maximum-Likelihood Analysis of Microarray Data, Journal of Computational Biology, **7(6)**, 805–817 (2000).

[17] Han-Yu Chuang, Hong Fang Liu, Stuart Brown, Cameron McMunn-Coffran, Cheng-Yan Kao and D. Frank. Hsu, Identifying Significant Genes from Microarray Data, Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering, 358–365, (2004).

[18] Prathima Arvind,Shanker Jayashree, SriKarthika Jambunathan, Jiny Nair and Vijay V. Kakkar, Understanding gene expression in coronary artery disease through global profiling, network analysis and independent validation of key candidate genes, Journal of Genetics, **94(4)**, 601–610 (2015).

[19] Shafa Mahajan, Abhishek, Shailendra Singh, Review On Feature Selection Approaches Using Gene Expression Data, Imperial Journal of Interdisciplinary Research (IJIR), **2(3)**, 356–364 (2016).

[20] Inaki Inza, Pedro Larranga, Rosa Blanco, Antonio J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artificial Intelligence in Medicine, **31**, 91–103 (2004).

[21] Fatemeh Vafaee Sharbaf, Sara Mosafer, Mohammad Hossein Moattar, A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization, Genomics, **107**, 231–238 (2016).

[22] Kohbalan Moorthy and Mohd Saberi Mohamad, Random forest for gene selection and microarray data classification, Bioinformation, **7(3)**, 142–146 (2011).

[23] Zena M. Hira and Duncan F. Gillies, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, Advances in Bioinformatics, Article ID 198363, **2015**, 13 pages (2015).

[24] Smita Chormunge and Sudarson Jena, Clustering-Biased Feature Selection Framework for Microarray Data, International Journal of Performability Engineering, **13(4)**, 383–389 (2017).

[25] Sen Liang, Anjun Ma, Sen Yang,Yan Wang, Qin Ma, A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis, Computational and Structural Biotechnology Journal, **16**, 88–97 (2018).

[26] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, Zhigang Gao, A Hybrid Feature Selection Algorithm for Gene Expression Data Classification, Neurocomputing, **256**, 56–62 (2017).

[27] Mehrab Ghanat Bari,Sirajul Salekin and Jianqiu(Michelle) Zhang, A Robust and Efficient Feature-Selection Algorithm for Microarray Data, Molecular Informatics, **36**, 1600099 (2017).

[28] Milos Radovic, Mohmed Ghalwash, Nenad Filipovic and Zoran Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinformatics, **18**, 9 (2017).

[29] T. Sai Sujana, N. Madhu Sudana Rao, Raja Sekar Reddy, An Efficient Feature Selection using Parallel Cuckoo Search and Naïve Bayes Classifier, International Conference on Networks and Advances in Computational Technologies, 167–172 (2017).

[30] S. Jayanthi and C.R. Rene Robin, A Survey on different classification methods for microarray data analysis, Advances in Environmental Biology, **11(5)**, 13–18 (2017).

[31] Bibhuprasad Sahu, A Combo Feature Selection Method(Filter + Wrapper) for Microarray Gene Classification, International Journal of Pure and Applied Mathematics, **118(16)**, 389–401 (2018).

[32] Si Prasad Pothuraju and M. Sreedevi, Distributed feature selection(DFS) strategy for microarray gene expression data to improve the classification performance, Clinical Epidemiology and Global Health (2018).

[33] Michail Tsagris,Vincenzo Lagani and loannis Tsamardinos, Feature Selection for high-dimensional temporal data, BMC Informatics, **19**, 17 (2018).

[34] Saleh M. Abu-Soud and Sufyan Almajali,ILA-3, An Inductive Learning Algorithm with a new Feature Selection Approach, SEAS Transactions on Systems and Control, **13**, 171–185 (2018).

[35] Farzana Kabir Ahmad, Yuhanis Yusof, Nooraini Yusoff, Filter-Based Gene Selection Method for Tissues Classification on Large Scale Gene Expression Data, International Journal of Engineering and Technology, **7(2.15)**, 68–71 (2018).

K. Uma Maheswari completed B.E. (Computer science) in Institute of Road and Transport Technology, Bharathiar University, Tamilnadu in the year 1998 and M. Tech in Software Engineering in the year 2008 at Bharathidasan Institute of Technology, Trichy. Currently she is working as Assistant Professor in the Department of Information Technology, Anna University Chennai, BIT Campus, Trichy. She has 18 years of teaching experience and currently pursuing her Ph. D. She has published many research papers in reputed journals, international and national conferences. Her areas of interest include Data Mining, Bio-informatics, Algorithms and Software Engineering.

A. Valarmathi is working as Assistant Professor & Head, Department of MCA, Anna University, Trichy, Tamil Nadu, India. She earned her Doctorate in 2013 from Anna University Chennai. Her areas of interest include Mobile computing, Wireless Networks and Bioinformatics. She joined as an Assistant Professor in Anna University, BIT Campus, Tiruchirappalli from 2009. She has 12 years of teaching experience. Presently, she is guiding 10 research scholars and guided for more than 100 MCA & M.Tech. students. She is actively involved in research besides teaching. At present, she is handling Consultancy projects worth of rupees more than 50 lakhs. She has published around 100 papers in International journals and conferences. She has authored many books. She is presently the reviewer for many reputed journals such as Inderscience, Springer and ELSEVIER, etc.