Random Forest for Gene Selection and Microarray Data Classification

Kohbalan Moorthy and Mohd Saberi Mohamad

Artificial Intelligence and Bioinformatics Research Group,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia
kohbalan@gmail.com, saberi@utm.my

Abstract. A random forest method has been selected to perform both gene selection and classification of the microarray data. The goal of this research is to develop and improve the random forest gene selection method. Hence, improved gene selection method using random forest has been proposed to obtain the smallest subset of genes as well as biggest subset of genes prior to classification. In this research, ten datasets that consists of different classes are used, which are Adenocarcinoma, Brain, Breast (Class 2 and 3), Colon, Leukemia, Lymphoma, NCI60, Prostate and Small Round Blue-Cell Tumor (SRBCT). Enhanced random forest gene selection has performed better in terms of selecting the smallest subset as well as biggest subset of informative genes through gene selection. Furthermore, the classification performed on the selected subset of genes using random forest has lead to lower prediction error rates compared to existing method and other similar available methods.

Keywords: Random forest, gene selection, classification, microarray data, cancer classification, gene expression data.

1 Introduction

Since there are many separate methods available for performing gene selection as well as classification [1]. The interest in finding similar approach for both has been an interest to many researchers. Gene selection focuses at identifying a small subset of informative genes from the initial data in order to obtain high predictive accuracy for classification. Gene selection can be considered as a combinatorial search problem and therefore can be suitably handled with optimization methods. Besides that, gene selection plays an important role preceding to tissue classification [2], as only important and related genes are selected for the classification. The main reason to perform gene selection is to identify a small subset of informative genes from the initial data before classification in order to obtain higher prediction accuracy. Classification is an important question in microarray experiments, for purposes of classifying biological samples and predicting clinical or other outcomes using gene expression data. Classification is carried out to correctly classify the testing samples according to the

D. Lukose, A.R. Ahmad, and A. Suliman (Eds.): KTW 2011, CCIS 295, pp. 174–183, 2012.

class. Therefore, performing gene selection antecedent to classification would severely improve the prediction accuracy of the microarray data. Many uses single variable rankings of the gene relevance and random thresholds to select the number of genes, which can only be applied to two class problems. Random forest can be used for problems arising from more than two classes (multi class) as stated by [3].

Random forest is an ensemble classifier which uses recursive partitioning to generate many trees and then combine the result. Using a bagging technique first proposed by [4], each tree is independently constructed using a bootstrap sample of the data. Classification is known as discrimination in the statistical literature and as supervised learning in the machine learning literature, and it generates gene expression profiles which can discriminate between different known cell types or conditions as described by [1]. A classification problem is said to be binary in the event when there are only two class labels present [5] and a classification problem is said to be a multiclass classification problem if there are at least three class labels.

An enhanced version of gene selection using random forest is proposed to improve the gene selection as well as classification in order to achieve higher prediction accuracy. The proposed idea is to select the smallest subset of genes with the lowest out of bag (OOB) error rates for classification. Besides that, the selection of biggest subset of genes with the lowest OOB error rates is also available to further improve the classification accuracy. Both options are provided as the gene selection technique is designed to suit the clinical or research application and it is not restricted to any particular microarray dataset. Apart from that, the option for setting the minimum no of genes to be selected is added to further improve the functionality of the gene selection method. Therefore, the minimum number of genes required can be set for gene selection process.

2 Materials and Methods

There are ten datasets used in this research, which are Leukemia, Breast, NCI 60, Adenocarcinoma, Brain, Colon, Lymphoma, Prostate and SRBCT (Small Round Blue Cell Tumor). For breast, there are two types used which has 78 samples and 96 sample each with different class. Five of the datasets are two class and others are multi class.

The microarray datasets used are mostly multi class datasets. The description of the datasets such as the no of genes, no of patients, the dataset class and also the reference of the related paper for the dataset has been listed in the Table 1.

Table 1. Main characteristics of the microarray datasets used

Dataset Name	Genes	Patients	Classes	Reference
Adenocarcinoma	9868	76	2	[6]
Brain	5597	42	5	[7]
Breast2	4869	77	2	[8]
Breast3	4869	95	3	[8]
Colon	2000	62	2	[9]
Leukemia	3051	38	2	[10]
Lymphoma	4026	62	3	[11]
NCI60	5244	61	8	[12]
Prostate	6033	102	2	[13]
SRBCT	2308	63	4	[14]

2.1 Standard Random Forest Gene Selection Method

In the standard random forest gene selection, the selection of the genes is done by using both the backward gene elimination and the selection based on the importance spectrum. The backward elimination is done for the selection of small sets of non redundant variables, and the importance spectrum for the selection of large, potentially highly correlated variables. Random forest gene elimination is carried out using the OOB error as minimization criterion, by successfully eliminating the least important variables. It is done with the importance information returned from random forest.

Using the default parameters stated by [3], all forests that result from iterative elimination based on fraction.dropped value, a fraction of the least importance variables used in the previous iteration is examined. The default fraction.dropped value is 0.2 which allows for relatively fast operation is consistent with the idea of an aggressive gene selection approach, and increases the resolution as the number of variables considered becomes smaller. By default, the gene importance is not recalculated at each step as according to [15], since severe over fitting resulting from recalculating of gene importance. The OOB error rates from all the fitted random forests are examine after fitting all the forests. The solution with the smallest number of genes whose error rate is within standard errors of the minimum error rate of all forests is selected. The standard error is calculated using the expression for a binomial error count as stated below. The p resembles the true efficiency and n is the sample size.

$$Standard\ error\ = \sqrt{p(1-p) * \frac{1}{n}} \tag{1}$$

The standard random forest gene selection method performs gene selection by selecting the smallest subset of genes with average out of bag (OOB) error rates between the smallest number of variables which is two and the subset with the number of variables that has the lowest OOB error rates. This strategy can lead to solutions with fewer genes but not the lowest OOB error rates.

2.2 Improvement Made to the Random Forest Gene Selection Method

Few improvements have been made to the existing random forest gene selection, which includes automated dataset input that simplifies the task of loading and processing of the dataset to an appropriate format so that it can be used in this software. Furthermore, the gene selection technique is improved by focusing on smallest subset of genes while taking into account lowest OOB error rates as well as biggest subset of genes with lowest OOB error rates that could increase the prediction accuracy. Besides that, additional functionalities are added to suite different research outcome and clinical application such as the range of the minimum required genes to be selected as a subset. Integration of the different approaches into a single function with parameters as an option allows greater usability while maintaining the computation time required.

Automated Dataset Input Function. In the current R package for random forest gene selection, the dataset format for input as well as processing is not mentioned and

cause severe confusion to the users. Besides that, the method for inputting the dataset which is mostly in text file format required further processing to cater to the function parameters and format for usability of the gene selection process. Therefore, an automated dataset input and formatting functions has been created to ease the access of loading and using the dataset of the microarray gene expression based on text files input. The standard dataset format used for this package has two separate text files, which are data file and class file. These files need to be inputted into the R environment before further processing can be done. The method and steps for the automated dataset input is described in Figure 1. The steps have been created as an R function which is included inside the package and can be used directly for the loading of the dataset. The function takes two parameters, which are the data file name with extension and class file name with extension.

Step 1: Input data name and class name.
Step 2: Error checking for valid file name, extension and file existence.
Step 3: Read data file into R workspace.
Step 4: Data processing.
Step 5: Transpose data.
Step 6: Read length of class/sample.
Step 7: Read class file into R workspace.
Step 8: Create class factor.
Step 9: Load both data and class for function variable access.

Fig. 1. Steps required for the automated dataset input and formatting in R environment

Selection of Smallest Subset of Genes with Lowest OOB Error Rates. The existing method performs gene selection based on random forest to select smallest subset of genes while compromising on the out of bag (OOB) error rates. The subset of genes is usually small but the OOB error rates are not the lowest out of all the possible selection through backward elimination. Therefore, enhancement has been made to improve the prediction accuracy by selecting the smallest subset with the lowest OOB error rates. Hence, lower prediction error rates can be achieved for classification of the samples. This technique is implemented in the random forest gene selection method. During each subset selection based on backward elimination, the mean OOB error rate and standard deviation OOB error rate are tracked at every loop as the less informative genes are removed gradually. Once the loop terminates the subset with the smallest number of variables and lowest OOB error rates are selected for classification.

```
While backward elimination process = TRUE
        If current OOB error rates < = previous OOB error rates
                Set lowest error rate as current OOB error rates
                Set no of variables selected
        End If
End While
```

Fig. 2. Method used for tracking and storing the lowest OOB error rates

The subset of genes is located based on the last iteration with the smallest OOB error rates. During the backward elimination process, the number of selected variables decreases as the iteration increases.

Selection of Biggest Subset of Genes with Lowest OOB Error Rates. Another method for improving the prediction error rates is by selecting the biggest subset with the lowest OOB error rates. This is due to the fact that any two or more subsets with different number of selected variables with same lowest error rates indicates the informative genes level are the same but the relation of the genes that contribute to the overall prediction is not the same. So, having more informative genes can increase the classification accuracy of the sample.

The technique applied for the selection of biggest subset of genes with the lowest OOB error rates are similar to the smallest subset of genes with the lowest OOB error rates, except the selection is done by picking the first subset with the lowest OOB error rates from all the selected subset which has the lowest error rates. If there is more than one subset with lowest OOB error rates, the selection of the subset is done by selecting the one with highest number of variables for this method. The detailed process flow for this method can be seen in the Figure 3. This technique is implemented to assist researches that require filtration of genes for reducing the size of microarray dataset while making sure that the numbers of informative genes are high. This is achieved by eliminating unwanted genes as low as possible while achieving highest accuracy in prediction.

```
While looping all the subset with lowest OOB error rate
        If Current no of selected genes >= Previous no of selected genes
                Set Biggest subset = Current number of selected genes
        End If
End While
```

Fig. 3. Method used for selecting the biggest subset of genes with lowest OOB error rates

Setting the Minimum Number of Genes to be Selected. Further enhancement is made to the existing random forest gene selection process by adding an extra functionality for specifying the minimum number of genes to be selected in the gene selection process that is included into the classification of the samples. This option allows flexibility of the program to suite the clinical research requirements as well as other application requirement based on the number of genes needed to be considered for classification.

The input for the minimum number of genes to be selected during the gene selection process is merged with the existing functions as an extra parameter input that has a default value of 2. The selected minimum values are used during the backward elimination process which takes place in determining the best subset of genes based on out of bag (OOB) error rates. At each time of a loop for selecting the best subset of genes, random forest backward elimination of genes is carried out by removing the unwanted genes gradually at each loop based on the *fraction.dropped* values selected. Therefore, as the no of loop increases, the no of genes in the subset

decreases leaving the most informative genes inside the subset, as less informative genes are removed. The minimum no of genes specified is checked at each loop and if the total number of genes for a subset is less than the specified value, the loop is terminated leaving behind all the subsets.

```
While backward elimination process
        If length of variables <= to minimum required variables
                Break
```

Fig. 4. Method used for terminating the loop once the desired number of variables achieved

Access Additional Functions through Parameters Input. The automated dataset input have been created as a separate function to facilitate the loading and processing of the dataset to suite the data format required by the gene selection. The choice for selecting the smallest or biggest subset of genes with the lowest OOB error rates is integrated in a single function inside the random forest gene selection process and can be access by specifying the parameters. The minimum required variables can also be set by specifying the value at the function parameter. The default option is to select the smallest subset of genes and the minimum number of selected variables is two.

2.3 Performance Measurement

For gene selection using random forest, backward elimination using OOB error rates is used as the final set of genes is selected based on the lowest out of bag (OOB) error rates as random forest returns a measure of error rate based on the out-of-bag cases for each fitted tree. The classification performance of the microarray data using random forest is measured using .632 bootstrap method [16]. In this method, the prediction error rates obtained is used to compare the performance of the random forest in classification where lower error rates means higher prediction accuracy.

In the .632 bootstrap, accuracy is estimated as followed. Given a dataset of size n, a bootstrap sample is created by sampling n instances uniformly from the data (with replacement). Since the dataset is sampled with replacement, the probability of any given instance not being chosen after n samples is:

$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} \approx 0.368 \tag{2}$$

The expected number of distinct instances from the original dataset appearing in the test set is thus 0.632. The accuracy estimate is derived by using the bootstrap sample for training and the rest of the instances for testing. Given a number *b*, the number of bootstrap samples, let $c0_i$ be the accuracy estimate for bootstrap sample *i*. The .632 bootstrap estimates are defined as:

$$acc_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0.632 \cdot c0_i + 0.368 \cdot acc_s) \tag{3}$$

Where acc_s is the resubstitution error estimate on the full dataset (the error on the training set). The assessment method used has been able to populate and list the overall performance of the algorithm with other similar algorithms and techniques through prediction error rates calculation comparison.

3 Results and Discussion

In this section, the full result of all the options used is compared. For the bar chart, the result for each dataset is plotted against the accuracy, therefore the higher the values the lower is the error rates. Based on the Figure 5, the enhanced random forest gene selection performs better compared to standard method. Though, different options have different effects to the dataset being tested. Most of the datasets tested showed larger improvement in terms of accuracy achieved for classification when the subset of genes selected is larger.

However, some datasets with smaller subset of genes outperformed the larger subset of genes. This could be due to the effect of the informative genes, as more informative genes contribute to better classification accuracy. For example in leukemia dataset, selecting bigger subset or setting the minimum no of genes more than 10 has reduced the prediction accuracy as its possibility of low no of informative genes. The highest accuracy achieved for this dataset is by selecting smallest subset of genes which has only two genes selected as the subset. Hence, the gene selection options vary according to the dataset used.

Based on the three different options presented for the enhanced random forest gene selection, the first option which is selection of smallest subset of genes based on lowest OOB error rates is suitable for Breast 2 and Leukemia dataset as it provided the highest accuracy compared to other options. The second option using selection of biggest subset of genes based on lowest OOB error rates is suitable for Brain, Breast 3, Colon, Lymphoma, Prostate and SRBCT as it manage to achieve highest accuracy for these datasets using this option. Whereas, the third option which performs selection of smallest subset of genes based on lowest OOB error rates with minimum selected genes set to ten is suitable for Adenocarcinoma and NCI60 dataset as the accuracy achieved for these datasets is highest compared to other options.

The highest accuracy achieved for Adenocarcinoma dataset is 0.8371, for Brain dataset is 0.8197, Breast 2 dataset is 0.6718, Breast 3 dataset is 0.6682, Colon dataset is 0.8757, Leukemia dataset is 0.9418, Lymphoma dataset is 0.9620, NCI60 dataset is 0.7271, Prostate dataset is 0.9446 and SRBCT dataset is 0.9761. The huge improvement achieved in terms of the error rates differences between the standard random forest gene selection method and enhanced random forest gene selection method is from NCI60 dataset, where the differences of the error rates is 0.0801.

Further comparison with other available methods such as Diagonal Linear Discriminant Analysis (DLDA), K nearest neighbor (KNN) and Support Vector Machines (SVM) with Linear Kernel has been done as well. The comparison results have been included as supplementary page and can be downloaded at this link: http://www.mediafire.com/?y6m9edecsjd88xg.

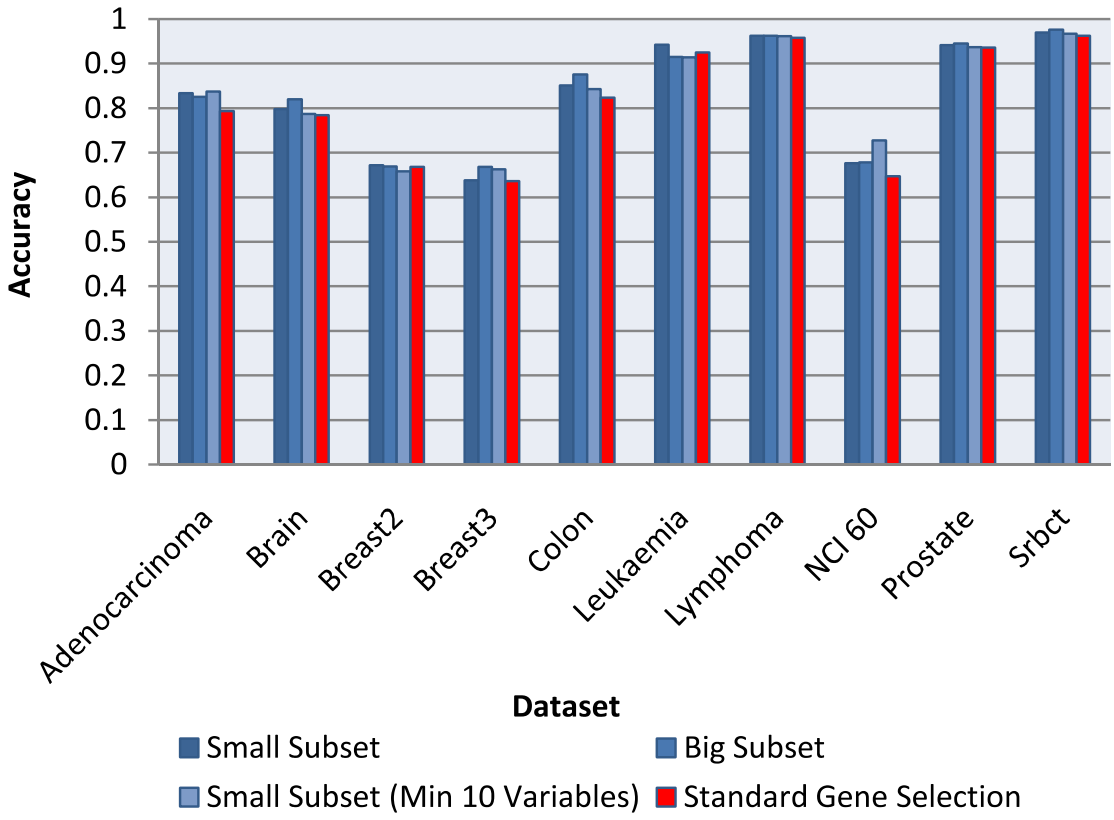

Fig. 5. Comparison between enhanced variables selection with three different options against standard gene selection method. A higher value indicates lower error rates

4 Conclusion

The proposed enhanced random forest gene selection has been tested with ten datasets and the outcome is as presented in the result and discussion section. There is improvement in terms of prediction accuracy for all datasets compared to the standard random forest gene selection. The gene selection plays an important role prior to classification and the way of selecting the subset of genes based on the type of dataset is also important in order to obtain lower error rates for classification. The option for selecting the smallest subset or bigger subset as well as setting the minimum required number of genes is the key factor in achieving higher accuracy in classification. For future works, additional options and functions can be integrated to suit other research works as well as clinical test by adding the features for selecting the range of genes or subset size to be selected for classification. This can be done by allowing the user to set the minimum number of genes as well as the maximum number of genes according to the requirement of the study. Hence, this enhanced random forest gene selection method provides the flexibility in determining the range of the genes in the subset as to how small or big the subset of genes required.