

IDENTIFICATION OF SIGNIFICANT FEATURES USING RANDOM FOREST FOR HIGH DIMENSIONAL MICROARRAY DATA

ARPITA NAGPAL *, VIJENDRA SINGH

Department of Computer Science and Engineering, School of Engineering and Technology
The NorthCap University, Gurugram, Haryana, India
*Corresponding Author: arpitanagpal@ncuindia.edu

Abstract

Feature subset selection for microarray data aims at reducing the number of genes so that useful information can be extracted from the samples. At the same time, selecting the relevant genes (features) from the high dimensional data can improve the classification accuracy of the learning algorithm. This paper proposes a feature selection algorithm, which is fit for high dimensional and small sample size microarray data. Feature selection is performed in two phases. In the first phase, Random Forest is used to identifying the importance of each feature, so that the features with high relevance can be given priority over less relevant ones. In the second phase, feature clustering is performed around the relevant features to yield the reduced feature set. A statistical method is used to create the clusters that aid in giving the genes specifically representing the disease. The effectiveness of the proposed algorithm has been compared with three state-of-the-art feature selection algorithms viz. Fast-Correlation Based Filter (FCBF), a Fast Clustering-Based Feature Selection Algorithm (FAST) and Random Forest (RF) on nine real-world cancer microarray datasets. Empirically, the algorithms have been evaluated through three well-known classifiers viz. probability based Naïve Bayes, Tree-based C4.5, and the Instance-based *IB1*. The stated result shows that the proposed algorithm can be helpful in finding the smaller set of features for cancer microarray datasets with better classification accuracy.

Keywords: Classification, Feature selection, High dimensional data, Microarray data, Random forest,

1. Introduction

In recent years, many researches had been conducted in the medical community on microarray data. Various data analysis techniques had been applied to the classification of cancer microarray data [1, 2]. While dealing with cancer microarray data, researchers have to face the greatest challenge of the high dimensionality of features. When the comparison between the small numbers of samples is made with a large number of genes, then it leads to the curse of dimensionality [3]. Such a large number of genes/features can lead to the problem of over-fitting while using these features as an input to machine learning tasks such as clustering, classification, etc. It also raises the classifiers complexity, training time and execution time. Availability of redundant and irrelevant features can also affect the classifier's performance. Therefore, there is a great need to remove these irrelevant and redundant features. Feature subset selection approach is one of an important concept to remove these irrelevant and redundant features. This feature selection approaches further decreases the data acquisition cost and learning time, which, results in improving cancer diagnosis.

There are various feature selection methods, i.e., filter, wrapper, hybrid and embedded method, depending upon the evaluation criterion involved. Wrapper method involves the learning algorithm within the search method, but it is computationally expensive. Filter method is independent of the learning algorithm, so they are computationally less expensive [4]. Filter feature selection algorithms have been widely investigated by various researchers and are broadly categorized as Univariate and Multivariate features [4-9].

A representative of various univariate methods is t-test, Pearson correlation, mutual information [6, 10], distance [11], correlation measure [12], model-based ranking. Univariate methods give a better understanding of data as it determines the strength of each feature individually with the response variable. They are simple to compute and their results are easily interpreted. However, such methods do not take into account the relationship between the features when dealing with high dimensional data. This problem is referred to as the problem of redundancy. This leads to difficulty in interpreting the results and deciding, which features are more important. In contrast, multivariate feature selection methods are introduced, which, captures these relationships among the selected set of features [9, 10].

The representatives of multivariate filter methods are FCBF [5], FAST [10]. They give a good subset of features, but it is difficult to decide a threshold at which, relevant features should be kept. They also use the concept of Information gain, which is always biased towards the features with unique values. One of the drawbacks of FCBF is that it takes into account the pair-wise correlation between the features. FAST algorithm [10] creates clusters, which, uses minimum Spanning tree to remove the problem of redundancy. This leads to an increased computational cost.

Unlike other algorithms, we propose an algorithm that avoids the threshold problem while optimizing the computational cost. The aim of this method is not just to reduce the features but also to determine that the selected features are the most important ones. Hence, this paper uses Random Forest (RF) [13] algorithm for feature selection. The principle of Random Forests is to combine many binary decision trees built using several bootstrap samples coming from the learning

sample L and choosing randomly at each node a subset of explanatory variables, X [14]. Random forest performs excellently where the number of variables is much larger than the samples [15]. Exploiting this property of Random Forest, it can be easily used for microarray datasets. It can become part of the “standard tool-box” of methods for class prediction and feature selection with microarray data [16].

Random Forest variable importance score has become very useful in many applications in the biomedical domain. Diaz-Uriarte et al. [16] have worked on classification of microarray data and shown that Random Forest has comparable performance to other classification methods, including Diagonal Linear Discriminant Analysis (DLDA), K-Nearest Neighbour (KNN), and support vector machine (SVM). However, if we need to obtain the smallest possible set of genes, this method is not preferable. Yao et al. [17] have used the variable importance score obtained from the Random Forest to rank the features and its result is evaluated with the average accuracy of SVM classifiers. Yang et al. [18], has proposed a method RM and SVMFS based on the fact that the classification performance of Random Forest and SVM classifier is similar. Here, Random Forest variable importance score is used and features are removed accordingly. Then the new obtained features are evaluated using SVM Classifier. In all these algorithms, Random Forest was directly used as a classifier to evaluate the features. Some of them have directly used variable importance score to segregate features [17, 19].

The proposed method uses a feature importance measure obtained from the Random Forest algorithm (RF). An importance score is obtained as the average increase in Mean Squared Error (MSE) of prediction when replacing the variable in the model with a random variable. The importance score allows getting a natural threshold for deciding, which features are relevant to the response (class) variable. All the features with negative or Zero in the importance score are considered as irrelevant ones and hence are removed. To further reduce the features, a statistical test has been applied, which analyse the preference of the feature directly with the help of paired comparison between the features. A significance level of 0.5 has been considered. The features are arranged in order of their importance score starting from the most important feature. They are clustered till the features Preference Value lies within the confidence interval. As the Preference Value exceeds the confidence interval, the cluster gets changed. Again, a new cluster is formed and the procedure is followed for all the relevant features. Finally, from each cluster, one representative feature is picked out to form a feature subset. These informative set of features forms the feature subset across the entire dataset.

The structure of the paper is as follows. Section 2 provides the framework of the proposed algorithm. Section 3 discusses the experimental results established on the publicly available cancer microarray datasets. Finally, conclusion and future directions are given in section 4.

2. Proposed feature selection technique

In the previous section, a new algorithm for feature selection has been discussed. Grounded on this, this section presents the proposed algorithm for feature subset selection for microarray data in two phases. The first phase is feature elimination using RF and the second phase is clustering the redundant features via a statistical test for further reducing the dimensions. Both these phases are hereby elaborated.

2.1. Feature elimination using Random Forest

This phase removes irrelevant features using Random Forest algorithm. As discussed in the previous section, the Random Forest algorithm evaluates the importance score for each feature, which is obtained as the percentage increase in Mean Square Error (% increase MSE) for each variable. This importance score indicates the percentage of decrease in classification accuracy if there would be the exclusion of variables from the set of variables. Our aim is to increase the classification accuracy. So, to achieve this, features are arranged in descending order of their MSE. This implies that the feature affecting the accuracy is at top of the list and hence they cannot be removed from the feature set. As described above, negative values signify that the feature is not important and hence can be removed. The features with value zero are also removed, as zero in MSE has no importance. The features with a higher value are kept on the list and are positive values. While moving down the list, features with zero and negative values are removed.

2.2. Clustering based dimension reduction

To further reduce the dimension of features, in the second phase a statistical significance testing measure has been adopted. This measure creates the clusters of the features. The statistical test applied here assesses whether the two features are statistically different from each other or not. The features found statistically same are clustered together. This is performed by initially finding the strength of each relevant feature found in the first phase and then comparing each feature relative to the highest strength feature. The strength of each feature (gene) is defined as follows:

Definition 1 (Strength of feature): Strength of a feature f_i is the weightage obtained by f_i relative to all the features of a dataset. It is shown in Eq. (1).

$$\text{Strength} [f_i] = RFimp_score_{[f_i]} * \frac{100}{\sum(RFimp_score)} \quad (1)$$

where, RF_{imp_score} is the value of the importance score obtained from the Random Forest algorithm.

When there is a large number of features, RF scores are very close to each other. To distinguish between them, the strength of each feature is calculated. In the experiment, we use the strength variable and calculate Preference Value between two features, which is the chance of preferring feature one gene over other. This is defined as follows:

Definition 2 (Preference Value): Preference Value of a feature $f_{(i+1)}$ is defined by Eq. (2).

$$PV[i + 1] = \text{Strength} [i + 1] / (\text{Strength} [i + 1] + \text{Val} [i]) \quad (2)$$

where, PV [i+1] is the Preference Value of the feature, Val [i] is the final value of the previous feature. The final value in Val for the first feature is same as its strength but is changes subsequently depending upon the comparisons performed.

For an instance, a feature with the strength of 11.2 is compared with a feature of the strength of 5.8 and has a chance of being ranked higher 66% of the time. The task of finding the preference of each gene corresponding to the previous one is a

pairwise comparison, which, expresses a preference between two alternatives. It can be explained by an analogy, suppose there are two features 'a' and 'b', the following can be the possible comparisons between them:

- " $a > b$ " means there is an expected win of 'a' over 'b' and the Preference Value is calculated as $a/(a+b)$.
- " $b > a$ " means there is an expected win of b over a and the Preference Value is calculated as $b/(a+b)$.

If both features a and b are equally preferred, then " $a=b$ " and Preference Value is calculated as $b/(b+b)$. This value is always equal to 0.5.

This test considers a significance level of 0.5. This means that if two features are compared then they both have an equal chance of preference. Confidence Interval (ci) is calculated by subtracting the significance level by $1/\sqrt{n}$ (where n is the number of a feature in the dataset). $1/\sqrt{n}$ is considered as the estimator of the sample.

2.3. Algorithm

The proposed algorithm is illustrated in Algorithm 1. The input given to the algorithm are all the features given in the dataset $D = \{f_1, f_2, \dots, f_n\}$ and Class C. Initially the algorithm finds the importance score of each gene using RF algorithm. These genes are arranged in descending order according to their score in an array, $RF_{imp_score} []$. The irrelevant feature removal is performed in step 7 where features with strength greater than zero are kept in a relevant array and others are removed. The feature with the highest strength is on the top of the list and is assigned to the first cluster. For each feature, Preference Value (PV) and final value (Val) is calculated. The PV and the final value for the first feature is the same as its strength and is in the first cluster (line algorithm1:10). The next step of the algorithm calculates PV for all features except the first feature as per definition 2. If this PV value of the next feature is less than ci value then it is added to next cluster and its value (Val) is its strength, otherwise, it is assigned to the previous cluster and its value (Val) is the value (Val) of the previous feature. Similarly, it is performed between all the features in the relevant set.

In algorithm 1, the value in the array 'Val' represents the significance of the features. The features having the same value in array 'Val' are placed in one cluster, which, represents each of them is equally significant. So, any one of them can be selected from that cluster. The output of the algorithm is the 'j' number of clusters. Henceforth, we pick one representative feature from each cluster C to form a feature subset of the complete dataset.

To elucidate the statistical procedure, the algorithm is explained below by taking an example of the results found on the Leukemia dataset. The dataset contains 7129 original genes. This implies that, $n=7129$ and hence confidence interval, $ci=0.4881$. The steps 1 to 4 of the algorithm finds the importance score of each feature by applying the RF algorithm with a class variable as dependent and all other features as independent ones.

After finding the importance score of each feature, the features are arranged in descending order and their strength is calculated according to definition 1. Table 1 depicts the first six features of the leukemia dataset arranged in order. Gene *a4008* is the most relevant feature found so it is kept as one of the features in the final subset. Its final value in Val is same as its strength and is placed in the first cluster. PV of the second gene *a4379* is calculated according to definition 2, $PV=0.465$. This value is less than ci so this gene constitutes to be a part of the second cluster and its final value in Val is same as its strength. Now PV for the third gene is calculated, which is 0.499. This value is greater than ci , so it is in the same cluster, i.e., the second cluster. Similarly, PV for each gene is calculated. According to the calculations of PV value, the third cluster starts forming from gene *a5543* when its PV value is less than ci . Table 1 clearly depicts that Genes *a4379*, *a1834*, *a5171* are placed in the second cluster. Genes *a5543*, *a976* are placed in the third cluster. This procedure is followed for all the relevant genes/features and clusters are formed.

```

Algorithm 1
Input: D ( $f_1, f_2, \dots, f_n, C$ ), the given dataset
Output: Clusters ( $j$  number of clusters)
Begin:

1. Grow random forest algorithm
2. Importance score= values in % inc MSE variable found by random forest algorithm.
3. RFimp_score[ ]  $\leftarrow$  Arrange importancescore in descending order.
4. For  $i=1$  to  $n$ 
5. Find the Strength of each feature 'i' using Eq. (1)
6. Endfor
7. Relevant [ ]  $\leftarrow$  Features with Strength > 0
8. Calculate  $ci=0.5-(1/\sqrt{n})$  where 'n' is total number of features in dataset'
9.  $j=1$ 
10. Assign, Val[1] = Strength[1] and Relevant[1] = cluster Cj
11. For  $i= 1$  to size of (Relevant)
12. Find PV [i+1] using Eq. (2)
13. If  $PV[i+1] < ci$ 
14.  $j = j+1$ 
15. Val[i+1] = Strength[i+1]
16. Associate Relevant[i+1], to cluster Cj
17. Else Val[i+1] =Val[i]
18. Associate Relevant[i+1], to cluster Cj
19. Endif
20. Endfor
21. End;
    
```

Table 1. A snapshot of most relevant six features of leukemia dataset depicting the statistical procedure.

| Feature | RFimp_score | Strength | PV | Val |
|--------------|-------------|-----------|-----------|-----------|
| <i>a4008</i> | 1.969578827 | 1.2139736 | | 1.2139736 |
| <i>a4379</i> | 1.718147118 | 1.0590006 | 0.4659097 | 1.0590006 |
| <i>a1834</i> | 1.714946004 | 1.0570276 | 0.4995338 | 1.0590006 |
| <i>a5171</i> | 1.657751682 | 1.0217752 | 0.4910549 | 1.0590006 |
| <i>a5543</i> | 1.613976581 | 0.9947939 | 0.4843687 | 0.9947939 |
| <i>a976</i> | 1.600142241 | 0.9862669 | 0.4978479 | 0.9947939 |

3. Experimental Results

3.1. Experimental setup and dataset

The experiments have been performed on nine cancer microarray datasets of high dimensions. The description of this dataset has been presented in Table 2. Some datasets have two classes and some may have more than two classes. Colon, leukemia and prostate datasets were obtained from Kent Ridge Biomedical Dataset data repository. For SRBCT, Khan et al. dataset have been used [20].

Before actually using the data in the experiments, a preprocessing procedure has been applied to them. Datasets of the breast and endometrium contained null values. The features containing more than 30% missing values have been left out. Other null values were replaced with the class wise mean of their respective attributes. Thus, 3000 features were left for endometrium and 5000 for breast cancer data [21]. Data for colon, SRBCT, leukemia and melanoma were used as it is. For other datasets, we adopted the technique suggested by Yang et al. [22] and Ramaswamy et al. [23]. For the prostate dataset, floor value of 100 and a ceiling value of 16000 with a variation of the max/min ratio as 5 and Max-Min difference of 50 were used to filter the values. For CNS-v1, Colon-I, the intensity threshold value as floor and ceiling between 20-16000 with max/min ratio as 5, 3 and max-min difference of 500,100 was used respectively. All datasets were normalized using z-score normalization before using them in experiments.

Table 2. Dataset description.

| Dataset | Instances | Original features | Pre-processed features | No. of class |
|-------------|-----------|-------------------|------------------------|--------------|
| Colon_I | 37 | 22883 | 8826 | 2 |
| Prostrate | 102 | 12600 | 5966 | 2 |
| Breast | 97 | 24482 | 5000 | 2 |
| Colon | 62 | 2000 | 2000 | 2 |
| SRBCT | 83 | 2308 | 2308 | 4 |
| Endometrium | 42 | 8872 | 3000 | 4 |
| Leukemia | 72 | 7129 | 7129 | 3 |
| Melanoma | 38 | 8076 | 8076 | 3 |
| CNS-v1 | 34 | 7129 | 2277 | 2 |

Random Forest uses two important parameters; they are *mtry*, the number of input variables randomly chosen to generate tree and *ntree*, the number of trees in the forest. Diaz-Uriarte and de Andres [16] have experimentally shown that changing the values of these parameters has a negligible effect on a number of genes selected. They have suggested that default values are a good option, with just a little variation in time of execution. Henceforth, in our experiment, we have used the default values for these parameters. ($ntree=500$, $mtry=\sqrt{n}$, n is the number of features).

The parameters used to compare both the algorithms are the number of features, runtime and classification accuracy. Since the runtime of the algorithm is machine dependent, hence all the algorithms are implemented and compared on the same machine.

The classification accuracy is calculated using the 10-fold cross-validation strategy for the training and testing set on all the datasets in Weka [24]. For each

classification algorithm, average classification accuracy found under each algorithm and each dataset has been obtained.

3.2. Result analysis

We have compared the performance of the proposed algorithm, RFST with the existing feature selection algorithms FCBF [5] and FAST [10]. They are implemented again so as to compare the results obtained from other algorithms on the same machine. We also compared the obtained results of RFST with important variables selected from the Random Forest (RF) algorithm using a package VSURF [25]. Genuer et al. [25] have proved that the results obtained using VSURF package on the SRBCT dataset are comparable to RF algorithm given by Diaz-Uriarte and de Andres [16].

For result estimation, three performance metrics viz., Number of selected features, runtime and classification Accuracy has been selected. Accuracy (*Acc*) is obtained as given by Eq. (3).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

where, TP, TN, FP and FN represent True Positive, True Negative, False positive and False Negative values respectively. While comparing the algorithms, an algorithm with higher classification accuracy is considered as the best algorithm. However, the algorithm with the same accuracy is compared based on a number of genes/features and the algorithm with less number of selected features is the best performer.

Table 3 represents a number of features selected and runtime on different datasets for four algorithms. We can say that different algorithms select a different subset with different size. FCBF in select only one feature for Endometrium datasets. Figure 1 depicts the graphical view of a number of features selected by the four feature selection algorithms. Based on observations in Table 3 and Fig. 1 it is clear that RFST algorithm selects less number of features than FCBF and FAST in all the datasets except for prostate dataset where FAST selects 13 less number of features. As compared to variables selected using Random Forest (RF), RFST selects less number of genes in all datasets except for prostate and melanoma.

Table 3. Number of feature selected and runtime for four feature selection algorithms.

| Dataset | Number of feature selected | | | | Runtime (in min.) | | | |
|----------------|----------------------------|--------------|--------------|--------------|-------------------|---------------|---------------|--------------|
| | FCBF | FAST | RF | RFST | FCBF | FAST | RF | RFST |
| Colon I | 15 | 16 | 63 | 11 | 32.66 | 50.146 | 30.6 | 20.4 |
| Prostrate | 77 | 19 | 19 | 32 | 39.86 | 840 | 1020 | 25.5 |
| Breast | 99 | 16 | 42 | 12 | 33.41 | 41.75 | 23.3 | 22.46 |
| Leukemia | 12 | 15 | 16 | 10 | 150.64 | 114.1 | 67.6 | 60 |
| Colon | 34 | 14 | 26 | 17 | 14.724 | 166 | 59.2 | 41 |
| Endometrium | 1 | 12 | 14 | 8 | 102.5 | 32.83 | 60 | 25.3 |
| Melanoma | 43 | 29 | 7 | 25 | 114.37 | 160.69 | 127.5 | 93.6 |
| SRBCT | 10 | 5 | 13 | 8 | 8.603 | 70.6 | 15.9 | 6.8 |
| CNS-v1 | 35 | 20 | 21 | 19 | 3.62 | 67.8 | 27.5 | 3.7 |
| Average | 36.22 | 16.22 | 24.55 | 15.77 | 55.60 | 171.55 | 159.07 | 33.20 |

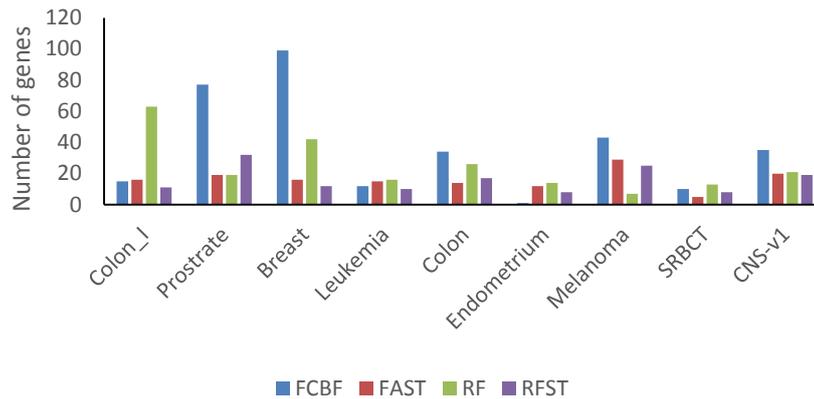


Fig. 1. Variation in number of selected genes for all four feature algorithms.

From Table 3, it can be observed that the proposed algorithm RFST performs better in terms of runtime. As average runtime is 22.4%, 138.4%, 125.9% less than that of FCBF, FAST and RF.

Table 4 gives the classification accuracy values on each dataset for the Naïve Bayes classifier. The best classification accuracy achieved for each dataset is shown in bold for a given filter method and a classifier.

Table 4. Accuracy obtained for Naïve Bayes classifier with four feature selection algorithms.

| Dataset | Accuracy (10 fold) | | | | |
|------------------------------|--------------------|-------|--------------|-------|--------------|
| | Full set | FCBF | FAST | RF | RFST |
| Colon_I (two class) | 81.08 | 42.4 | 65.64 | 97.29 | 100 |
| Prostrate (two class) | 62.74 | 83.33 | 93.77 | 82.35 | 72.54 |
| Breast (two class) | 59.79 | 59.79 | 65.71 | 69.07 | 71.13 |
| CNS-v1 (two class) | 73.29 | 74.70 | 75.23 | 87.05 | 88.23 |
| Colon (two class) | 51.61 | 64.51 | 95.08 | 75.80 | 95.48 |
| Average (two class) | 65.70 | 64.94 | 79.08 | 82.31 | 85.47 |
| Melanoma (three class) | 47.36 | 50 | 42.67 | 50 | 63.15 |
| Leukemia (three class) | 64.44 | 72.77 | 100 | 98.61 | 100 |
| Average (three class) | 55.9 | 61.38 | 71.33 | 74.30 | 81.57 |
| SRBCT (four class) | 86.74 | 83.13 | 89.23 | 97.59 | 98.79 |
| Endometrium (four class) | 80.95 | 52.38 | 72.31 | 80.33 | 80.95 |
| Average (four class) | 83.84 | 67.76 | 80.77 | 88.96 | 89.87 |

As described, Table 4 indicates the obtained accuracy for Naïve Bayes with three feature selection algorithms and Fig. 2 describes the variation in classification accuracy with Naïve Bayes. In Table 4 and Fig. 2, some important results have been noticed as explained below:

- In most of the cases, when compared with the original data (full set), all the algorithms have improved the accuracy of Naïve Bayes. Except in the case of four class datasets, the accuracy of the full set is 16.08%, 3.07%, which is more than the FCBF and FAST respectively. However, the RF algorithm has 5.12% and RFST algorithm has 6.03% increase in accuracy than the full set.

- For two class datasets, the proposed RFST algorithm has improved the classification accuracy by an average of 20.53%, 6.39%, 3.16% than FCBF, FAST, RF respectively.
- For three class datasets, RFST algorithm is better at 20.19%, 10.24%, 7.27% than FCBF, FAST and RF respectively. The FAST and RFST algorithm has 100% accuracy for the leukemia dataset.
- In the case of four class dataset, RFST is more effective by an average of 22.11%, 9.1%, 0.91% than FCBF, FAST and RF respectively.

It is clear from Fig. 2, that apart from prostate data, in all other datasets the proposed RFST algorithm shows improvement in classification accuracy found using Naïve Bayes classifier.

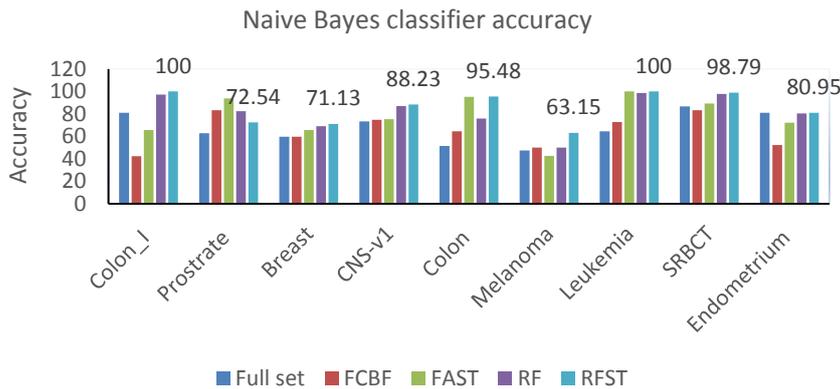


Fig. 2. Variation in terms of classification accuracy with Naïve Bayes classifier.

As described, Table 5 gives the classification accuracy of *C4.5* classifier and Table 6 depicts the accuracy of *IB1* classifier. This indicates some important points as explained below:

- Compared to the full dataset, the accuracy of *C4.5* has been improved by FAST, RF, RFST by 2.91%, 6.38%, 7.02% respectively in case of two class dataset, 12.8%, 3.94%, 17.25% respectively in case of three class dataset and 8.47%, 11.39%, 13.17%, respectively in case of four class dataset.
- For two class datasets, RFST has improved the classification accuracy of *C4.5* by 8.07%, 4.11%, 0.64% than FCBF, FAST and RF respectively. The accuracy of RFST in case of *IB1* classifier is more by 15.94 %, 9.47% and 1.18 % than FCBF, FAST and RF respectively.
- In case of *C4.5* classifiers, for three class datasets, RFST algorithm is better by 30.2%, 4.45% and 13.31% than FCBF, FAST and RF. In case of *IB1* classifier, RFST algorithm is 41.27%, 12.11% and 9.36% better than FCBF, FAST and RF respectively. RFST and FAST have obtained 100 percent accuracy for the leukemia dataset.
- For four class datasets, RFST ranks 1 with a margin of 11.39%, 4.7% and 1.78% more accuracy than FCBF, FAST and RF when *C4.5* classifier is

considered. In case of *IB1* classifier, it wins by a margin of 17.36%, 18.95%, 1.43% than FCBF, FAST and RF.

As discussed in Table 3, FCBF in datasets Endometrium, Melanoma, SRBCT select only one feature. From Tables 4 and 5, it can be seen that the accuracy of these datasets is very less, even less than the accuracy of the full dataset. So, it can be concluded that features should be reduced to an appropriate size, over reducing them to a size of one, does not lead to good performance.

Figures 3 and 4 further depicts the comparisons in terms of classification accuracy for all nine microarray datasets with two classifiers, *C4.5* and *IB1* respectively. From this comparison, two important points can be noticed as mentioned:

- With *C4.5* classifiers, the accuracy of RFST algorithm is highest among all other algorithms for all datasets except for prostate data.
- With the *IB1* classifier, RFST performed well for all datasets except with little decrease in accuracy than RF for SRBCT data.

Table 5. Accuracy obtained for *C4.5* classifier with four feature selection algorithms.

| Dataset | Accuracy (10 fold) <i>C4.5</i> | | | | |
|------------------------------|--------------------------------|-------|--------------|--------------|--------------|
| | Full set | FCBF | FAST | RF | RFST |
| Colon_I (two class) | 94.59 | 89.18 | 70.14 | 97.20 | 97.29 |
| Prostrate (two class) | 75.49 | 89.21 | 89.87 | 81.37 | 80.39 |
| Breast (two class) | 57.73 | 52.57 | 58.68 | 58.04 | 60.82 |
| CNS-v1 (two class) | 70.58 | 67.01 | 73.19 | 82.35 | 85.29 |
| Colon (two class) | 69.35 | 64.51 | 90.4 | 80.64 | 79.03 |
| Average (two class) | 73.54 | 72.49 | 76.45 | 79.92 | 80.56 |
| Leukemia (three class) | 94.44 | 52.77 | 100 | 94.44 | 100 |
| Melanoma (three class) | 50 | 65.78 | 70.04 | 57.89 | 78.95 |
| Average (three class) | 72.22 | 59.27 | 85.02 | 76.16 | 89.47 |
| SRBCT (four class) | 73.49 | 72.28 | 77.69 | 86.74 | 85.54 |
| Endometrium (four class) | 57.14 | 61.90 | 69.88 | 66.66 | 71.42 |
| Average (four class) | 65.31 | 67.09 | 73.78 | 76.7 | 78.48 |

Table 6. Accuracy obtained for *IB1* classifier with four feature selection algorithms.

| Dataset | Accuracy (10 folds) | | | | |
|------------------------------|---------------------|-------|------------|------------|--------------|
| | Full set | FCBF | FAST | RF | RFST |
| Colon_I (two class) | 78.37 | 81.08 | 80.47 | 97.20 | 97.29 |
| Prostrate (two class) | 86.27 | 78.43 | 86.55 | 90.19 | 90.19 |
| Breast (two class) | 58.76 | 67.01 | 73.19 | 74.28 | 75.60 |
| CNS-v1 (two class) | 76.47 | 70.58 | 81.21 | 87.05 | 91.17 |
| Colon (two class) | 62.90 | 58.06 | 66.12 | 80.25 | 80.64 |
| Average (two class) | 72.55 | 71.03 | 77.50 | 85.79 | 86.97 |
| Melanoma (three class) | 50 | 47.36 | 62.63 | 73.68 | 86.84 |
| Leukemia (three class) | 83.33 | 56.94 | 100 | 94.44 | 100 |
| Average (three class) | 66.66 | 52.15 | 81.32 | 84.06 | 93.42 |
| SRBCT (four class) | 83.13 | 84.33 | 73.33 | 100 | 100 |
| Endometrium (four class) | 90.47 | 59.52 | 67.33 | 75.71 | 78.57 |
| Average (four class) | 86.8 | 71.92 | 70.33 | 87.85 | 89.28 |

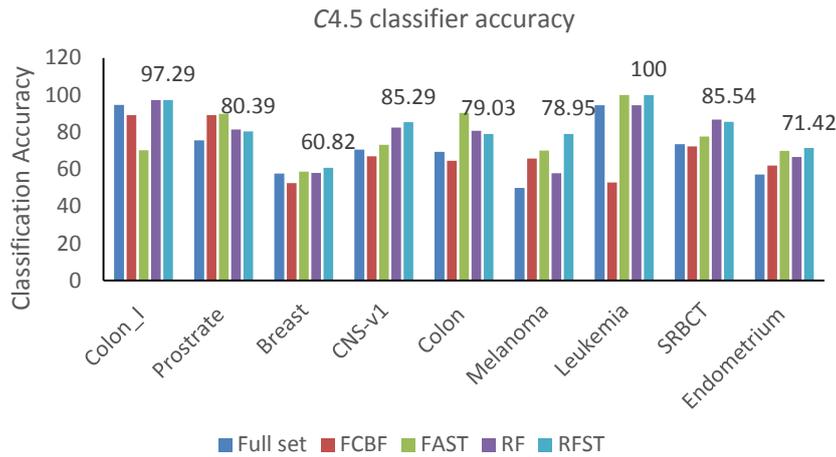


Fig. 3. Variation in terms of classification accuracy with C4.5 classifier.

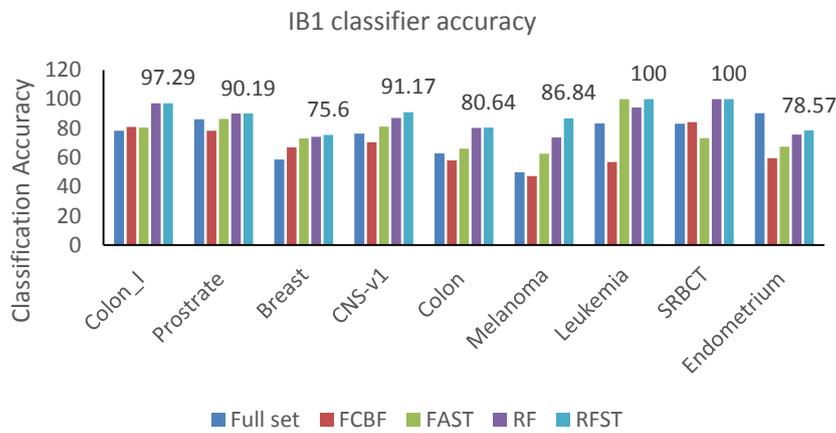


Fig. 4. Variation in terms of classification accuracy with IB1 classifier.

Table 7 lists the comparison of RFST with existing state of art algorithms in terms of a number of features and classification accuracy. It represents the best average accuracy of each filter algorithm with the number of features being represented within the parenthesis. The RFST is found to be better than the existing algorithm as it utilizes the property of the Random Forest algorithm for high dimensional data. The hybridization of RF score along with a statistical test is a better option than other filter or hybrid algorithms.

3.3. Statistical significance of the result

The purpose of statistical significance testing is to determine whether the two feature selection methods are different or not. One of the tests used is a Friedman test. The Friedman test [26] is a non-parametric approach. It can be used as a

measure to compare the rank of ‘ k ’ algorithms over ‘ d ’ datasets. It provides a test of significance for data with ranks less than 6. If the value for k is greater than 5, the level of significance or rank of the algorithm can be seen in the χ^2 distribution table. Treat the data as $\{x_{ij}\}d \times k$ matrix, where ‘ d ’ is a number of datasets called blocks and k is a number of columns, which has different algorithms.

$$M = \frac{12}{dk(k+1)} X \sum R_j^2 - 3d(k+1) \quad (4)$$

Table 7. Comparison of RFST with other state of art algorithms.

| | Colon | Leukemia | SRBCT | Colon_I | Breast | Prostrate | CNS-V1 |
|---------------------------------|----------------------|--------------------|-------------------|----------------|----------------------|---------------|----------------------|
| RFST+Naïve Bayes | 95.48 (17) | 100 (10) | | 100(11) | | | |
| BBF+SVM [27] | 90.32 (12) | 100 (3) | | | | | |
| MI+SVM [28] | 74.19 (23) | | | | | | |
| MRMR-ABC+SVM [29] | 94.17 (20) | | 96.30 (10) | | | | |
| RFST+C4.5 | | 100 (10) | | | | | |
| Seeja and Sweta [30] | | 94.11 (200) | | | | | |
| MGS-CM+SVM [31] | | 90.97 | | | | | |
| t-test+SVM [32] | | 96.88 | | | | | |
| SVM based t-statistics+SVM [32] | | 98.12 | | | | | |
| SVM+RFE [32] | | 97.88 | | | | | |
| MDS+SVM [33] | | 96.75 | | | | 94.32 | |
| POS+KNN [34] | | 99.5 (1) | | | 66.8 (11) | | |
| POS+SVM [34] | | 99.5 (1) | 99.7 (8) | | 68.7 (22) | | |
| Alonso-Gonzales et al. [35] | | 78.28 (100) | | | | | 75.49 (100) |
| RFST+IB1 | | | 100 (8) | | 75.60 (12) | 90.19 (32) | 91.17 (19) |
| GS2+SVM [36] | | | 100 (96) | | | | |
| Gs1+SVM [36] | | | 98.8 (34) | | | | |
| Ftest+SVM [36] | | | 100 (78) | | | | |
| Mortazavi and Moattar [37] | | | 97.27 | | | 90.12 | |
| POS+KNN [34] | | | 99.5 (22) | | | | |

The null hypothesis of the Friedman test considered here is that there is no difference between the feature selection algorithms based on accuracies for all three classifiers. Decision rule then says that, reject the null hypothesis if $M >$ critical value. If this hypothesis gets rejected, then some posthoc test is required to compare the performance. The posthoc test can be Nemenyi test [38] or Bonferron’s Dunn test.

- As stated by Nemenyi test, two classifiers perform differently if the corresponding average ranks ($R_x - R_y$ where R_x and R_y are the average ranks of algorithms x and y respectively) differ by at least the Critical Difference (CD).

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}} \tag{5}$$

where k is the number of algorithms, N is the number of datasets and q_{α} is based on the studentized range statistic divided by $\sqrt{2}$.

The above-described Friedman test has been applied on a number of the feature selected from each of the four different algorithms found in Table 3. The null hypothesis of Friedman test is rejected when $n=9$, $k=4$ and $\alpha =1\%$. Therefore, to find algorithms that have significant differences a Nemenyi posthoc test has been applied. Critical distance is found to be 0.9851. Figure 5 indicates the result and it is found that at critical distance 0.9851, the number of the feature selected of RFST is statistically smaller than those of RF and FCBF and there is no evidence to indicate statistical difference between RFST and FAST.

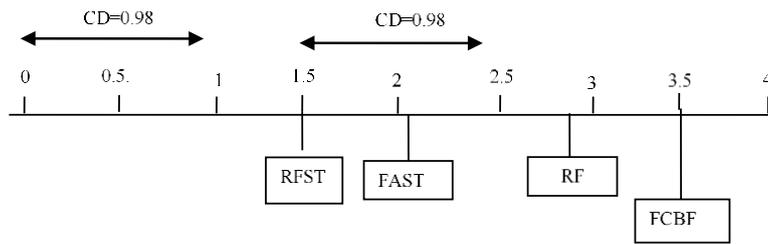


Fig. 5. Number of feature selected comparison between each of the algorithms.

We have also individually performed three Friedman tests on each classifier with four feature selection algorithms to explore the comparison of algorithms based on classification accuracies found in Table 4, 5 and 6. Here, we compare with the critical value at $\alpha=1\%$, $d=9$ and $k=4$. The null hypothesis gets rejected in each test and we can say that all feature selection algorithms are performing differently. Further, to compare the two algorithms Nemenyi test is performed. The value of the critical distance is found to be 0.9851. Two algorithms are performing differently if average rank differs by at least the critical difference. Figure 6 shows the results when compared with Naïve Bayes Classifier. We observe that the accuracy of Naïve Bayes with RFST algorithm is statistically better than RF, FAST and FCBF. However, there is no statistical accuracy difference between RF and FAST algorithms.

With similar calculations, it was found that the accuracy of C4.5 with RFST is statistically better than those with FCBF. However, there is no statistical accuracy difference between RFST and RF. We observe that the accuracy of IB1 classifier with RFST is statistically better than that with FAST and FCBF.

Table 8 lists the gene accession numbers for some of the features obtained as the best features in regard to maximum accuracy and a minimum number of genes. These are the highly frequent genes delivering maximum classification accuracy.

In all aspects, we can conclude that Random Forest algorithm when combined with a statistical test, can generate accuracy performance of the classifiers with a minimum number of features when applied to some of the cancer microarray datasets. Therefore, RFST is a promising approach when solving feature selection for cancer data classification problems.

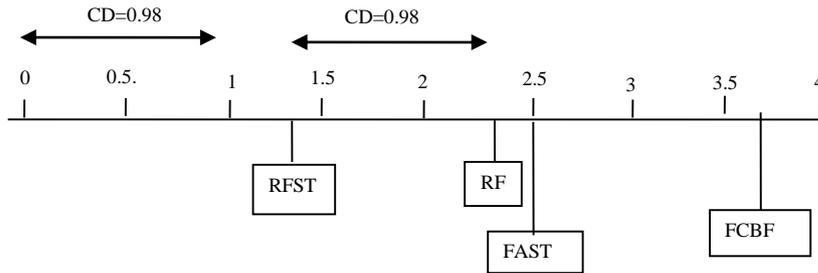


Fig. 6. Ranking of feature selection algorithms based on Naïve Bayes accuracy with Nemenyi test.

Table 8. Accession number of microarray dataset.

| Dataset | Gene List |
|--------------------|--|
| Prostate | AFFX.Thr.5_at,AFFX.HSAC07.00351_M_st,AFFX.YEL024w.RIP1_at,31380_at,31381_at,31388_at,31397_at,31415_at,31491_s_at,31494_at,31502_at,31515_at,31542_at,31553_at,31556_at,31563_at,31567_at,31625_at,31650_g_at,31652_at,31663_at,31675_s_at,31753_at,31784_at,31920_at,31933_r_at,31955_at,32009_at,32017_at,32409_at,32455_s_at,32471_at |
| Colon | Hsa.8147,Hsa.692,Hsa.37937,Hsa.1832,Hsa.831,Hsa.692,Hsa.3306,Hsa.878,Hsa.7165,Hsa.8125,Hsa.6039,Hsa.710,Hsa.1410,Hsa.103,Hsa.852,Hsa.6782,Hsa.9103 |
| Breast | Contig48393_RC AI830996, Contig25534_RC, AI200787, N65982, AI075917, AA972327, AI075662,AA758685, AI125186, W73245,AI679537,AA928010 |
| SRBCT | 234398, 950680, 845453, 37553, 244154, 1434948, 154790, 26021 |
| Endometrium | 159811,161106,168380,159764,159776,169140,168623, 161302 |
| Leukemia | affx-trpnx-m_at,affx-trpnx-3_at,affxhumisgf3a/m97935_5_at,affxhumisgf3a/m97935_ma_at,31376_at,40659_at,39365_i_at, 34734_at,32084_at,34904_at, 41439_at |
| Melanoma | 34304, 291503, 470621, 244307, 488431, 347220, 30966, 194342, 291706, 503819, 502664, 141768, 140966, 418126, 950482, 325090, 139689, 448190, 37491, 297392, 199610, 796398, 266361, 682528, 123953 |
| Colon_I | 201986_at, 204664_at, 207059_at, 200769_s_at, 206524_at, 201961_s_at, 209153_s_at, 207310_s_at, 206022_at, 202427_s_at, 202572_s_at |

4. Conclusion

This paper presents a new algorithm, RFST (Random Forest Statistical Test), which is used for feature selection of high dimensional microarray data. The algorithm takes the advantage of Random Forest as it covers the impact of each feature individually as well as in multivariate interaction with other features in the dataset. The proposed RFST algorithm filters out irrelevant features using the importance score obtained from Random Forest. Then, it assigns the statistically different features to different clusters. Finally, to remove redundant features, a representative from each cluster is chosen, which, results in the selection of a small number of features from high dimensional data.

In other domains, FAST and FCBF algorithms must be performing well with high dimensional data but in microarray domain, the proposed RFST technique performs much better. Additionally, it does not require a threshold value to segregate relevant features. The experiments have been carried out on the proposed method (RFST), variable selection performed using Random Forest (RF) as well as with commonly used feature selection algorithms Fast-Correlation Based Filter

(FCBF) and FAST on nine publicly available cancer microarray datasets. Results obtained confirms that the RFST algorithm is more promising than FCBF, FAST and RF, in terms of a number of the feature selected, runtime and classification accuracy for Naïve Bayes, *C4.5* and *IB1*. Friedman test, which is a Non-Parametric test, depicts that the RFST method differs significantly from the other three algorithms. In addition, the RFST algorithm's accuracy is comparable with the other state-of-art methods in all the datasets.

Nomenclatures

| | |
|-------------------|--|
| <i>Acc</i> | Classification accuracy |
| C_i | A single i^{th} cluster C |
| ci | Confidence Interval |
| fi | i^{th} features from all the features. |
| k | Number of algorithms used in statistical test |
| RF_{imp_score} | Importance score obtained from Random Forest algorithm |

Greek Symbols

| | |
|------------|--|
| α | Confidence level in Nemenyi test |
| q_α | Student rang statistic for infinite degrees of freedom |

Abbreviations

| | |
|------|---|
| FAST | Fast Clustering Based Feature Selection Algorithm |
| FCBF | Fast Correlation-Based Filter |
| MSE | Mean Square Error |
| PV | Preference Value |
| RF | Random Forest Algorithm |

References

1. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; Bloomfield, C.D.; Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531-537.
2. Khan, J.; Wei, J.S.; Ringner, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; and Meltzer, P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6), 673-679.
3. Bellman, R.E. (2015). *Adaptive control processes: A guided tour*. New Jersey: Princeton University Press.
4. Kohavi, R.; and John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2), 273-324.
5. Yu, L., and Liu, H. (2003, August). Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the Twentieth International Conference in Machine Learning*. Washington D.C., 856-863.
6. Huang, J.; Cai, Y.; and Xu, X. (2006). A filter approach to feature selection based on mutual information. *Proceedings of the 5th IEEE International Conference on Cognitive Informatics*. Beijing, China, 84-89.

7. Fu, L.M.; and Fu-Liu, C.S. (2005). Evaluation of gene importance in microarray data based upon probability of selection. *BMC Bioinformatics*, 6, 11 pages.
8. Risinger, J.I.; Maxwell, G.L.; Chandramouli, G.V.; Jazaeri, A.; Aprelikova, O.; Patterson, T.; and Berchuck, A.; and Barrett, J.C. (2003). Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. *Cancer Research*, 63(1), 6-11.
9. Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4), 537-550.
10. Song, Q.; Ni, J.; and Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 1-14.
11. Kira, K.; and Rendell, L.A. (1992). The feature selection problem: Traditional methods and a new algorithm. *Proceedings of the Tenth National Conference on Artificial Intelligence*. San Jose, California, 129-134.
12. Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning*. Stanford, California, 359-366.
13. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
14. Genuer, R.; Michel, V.; Eger, E.; and Thirion, B. (2010). Random forests based feature selection for decoding fMRI data. *Proceedings Compstat*, 267, 1-8.
15. Biau, G.; and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197-227.
16. Díaz-Uriarte, R.; and de Andres, S.A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 13 pages.
17. Yao, D.; Yang, J.; Zhan, X.; Zhan, X.; and Xie, Z. (2015). A novel random forests-based feature selection method for microarray expression data analysis. *International Journal of Data Mining and Bioinformatics*, 13(1), 84-101.
18. Yang, J.; Yao, D.; Zhan, X.; and Zhan, X. (2014). Predicting disease risks using feature selection based on random forest and support vector machine. *Proceedings of the 10th International Symposium on Bioinformatics Research and Applications*. Zhangjiajie, China, 1-11.
19. Genuer, R.; Poggi, J.-M.; and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225-2236.
20. Khan, J.; Wei, J.S.; Ringner, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; and Meltzer, P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6), 673-679.
21. Sardana, M.; Agrawal, R.K.; and Kaur, B. (2015). An incremental feature selection approach based on scatter matrices for classification of cancer microarray data. *International Journal of Computer Mathematics*, 92(2), 277-295.
22. Yang, K.; Cai, Z.; Li, J.; and Lin, G. (2006). A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7, 16 pages.
23. Ramaswamy, S.; Tamayo, P.; Rifkin, R.; Mukherjee, S.; Yeang, C.H.; Angelo, M.; Ladd, C.; Reich, M.; Litulippe, E.; Mesirov, J.P.; Poggio, T.; Gerald, W.; Loda, M.; Lander, E.S.; and Golub, T.R. (2001). Multiclass cancer diagnosis

- using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26), 15149-15154.
24. Witten, I.H.; Frank, E.; Hall, M.A.; and Pal, C.J. (2016). *Data mining: Practical machine learning tools and techniques (4th ed.)*. San Francisco: Morgan Kaufmann.
 25. Genuer, R.; Poggi, J.-M.; and Tuleau-Malot, C. (2015). VSURF: An R package for variable selection using random forests. *The R Journal*, 7(2), 19-33.
 26. Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1), 86-92.
 27. Zhang, J.-G.; and Deng, H.-W. (2007). Gene selection for classification of microarray data based on the Bayes error. *BMC Bioinformatics*, 8, 9 pages.
 28. Vanitha, C.D.A.; Devaraj, D.; and Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. *Procedia Computer Science*, 47, 13-21.
 29. Alshamlan, H.; Badr, G.; and Alohal, Y. (2015). MRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Article ID604910. *BioMed Research International*, 2015, 15 pages.
 30. Seeja, K.R.; and Shweta. (2011). Microarray data classification using support vector machine. *International Journal of Biometrics and Bioinformatics*, 5(1), 10-15.
 31. Salem, D.A.; Seoud, R.A.A.; and Ali, H.A. (2011). MGS-CM: a multiple scoring gene selection technique for cancer classification using microarrays. *International Journal of Computer Applications*, 36(6), 30-37.
 32. Mundra, P.A.; and Rajapakse, J.C. (2010). Gene and sample selection for cancer classification with support vectors based t-statistic. *Neurocomputing*, 73(13-15), 2353-2362.
 33. Qi, Y.; Sun, H.; Sun, Q.; and Pan, L. (2011). Ranking analysis for identifying differentially expressed genes. *Genomics*, 97(5), 326-329.
 34. Mahmoud, O.; Harrison, A.; Perperoglou, A.; Gul, A.; Khan, Z.; Metodiev, M. V.; and Lausen, B. (2014). A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. *BMC Bioinformatics*, 15, 20 pages.
 35. Alonso-González, C.J.; Moro-Sancho, Q.I.; Simon-Hurtado, M.A.; and Varela-Arrabal, R. (2012). Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications*, 39(8), 7270-7280.
 36. Yu, H.L.; Gao, S.; Qin, B.; and Zhao, J. (2011). Multiclass microarray data classification based on confidence evaluation. *Genetics and Molecular Research*, 11(2), 1357-1369.
 37. Mortazavi, A.; and Moattar, M.H. (2016). Robust feature selection from microarray data based on cooperative game theory and qualitative mutual information. Article ID 1058305. *Advances in Bioinformatics*, 2016, 16 pages.
 38. Nemenyi, P. (1963). *Distribution-free multiple comparisons*. Ph.D. Thesis, Princeton University.