# Booster in High Dimensional Data Classification

HyunJi Kim, Byong Su Choi, and Moon Yul Huh

**Abstract**—Classification problems in high dimensional data with a small number of observations are becoming more common especially in microarray data. During the last two decades, lots of efficient classification models and feature selection (FS) algorithms have been proposed for higher prediction accuracies. However, the result of an FS algorithm based on the prediction accuracy will be unstable over the variations in the training set, especially in high dimensional data. This paper proposes a new evaluation measure Q-statistic that incorporates the stability of the selected feature subset in addition to the prediction accuracy. Then, we propose the Booster of an FS algorithm that boosts the value of the Q-statistic of the algorithm applied. Empirical studies based on synthetic data and 14 microarray data sets show that Booster boosts not only the value of the Q-statistic but also the prediction accuracy of the algorithm applied unless the data set is intrinsically difficult to predict with the given algorithm.

**Index Terms**—High dimensional data classification, feature selection, stability, *Q*-statistic, Booster

## 1 Introduction

The presence of high dimensional data is becoming more common in many practical applications such as data mining, machine learning and microarray gene expression data analysis. Typical publicly available microarray data has tens of thousands of features with small sample size and the size of the features considered in microarray data analysis is growing. The statistical classification of the data with huge number of features and small sample size (undersampled problem) presents an intrinsic challenge [29]. A striking result has been found that the simple and popular Fisher linear discriminant analysis can be as poor as random guessing as the number of features gets larger [7], [16].

As was reported in [14], [59], most of the features of high dimensional microarray data are irrelevant to the target feature and the proportion of relevant features or the percentage of up-regulated or down-regulated genes compared with appropriate normal tissues is only $2\% \sim 5\%$. Finding relevant features simplifies learning process and increases prediction accuracy. The finding, however, should be relatively robust to the variations in training data, especially in biomedical study, since domain experts will invest considerable time and efforts on this small set of selected features. Hence, the proposed selection should provide them not only with the high predictive potential but also with the high stability in the selection [40].

### 1.1 Previous Studies

There have been lots of researches on the FS during the last two decades, and the research continues to be still one of the hot topics in machine learning area [1], [19], [26], [31], [33], [53], [58], [62], [67]. One often used approach is to first discretize the continuous features in the preprocessing step and use mutual information (MI) to select relevant features [13], [41], [49], [69]. This is because finding relevant features based on the discretized MI is relatively simple while finding relevant features directly from a huge number of the features with continuous values using the definition of relevancy is quite a formidable task.

Methods used in the problems of statistical variable selection such as forward selection, backward elimination and their combination can be used for FS problems. Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward elimination method since it is impractical to implement backward elimination process with huge number of features.

A serious intrinsic problem with forward selection is, however, a flip in the decision of the initial feature may lead to a completely different feature subset and hence the stability of the selected feature set will be very low although the selection may yield very high accuracy [10], [43], [77]. This is known as the stability problem in FS [40]. The research in this area is relatively a new field [3], [15], [24], [28], [30], [40], [47], and devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research.

### 1.2 A New Proposal for Feature Selection

This paper proposes *Q*-statistic to evaluate the performance of an FS algorithm with a classifier. This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm.

The basic idea of Booster is to obtain several data sets from original data set by resampling on sample space. Then FS algorithm is applied to each of these resampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm.


- *H. Kim is with Korea Fair Trade Mediation Agency, Korea. E-mail: hyunjikim123@gmail.com.*
- *B. S. Choi is with the Department of Multimedia, Hansung University, Korea. E-mail: cbs@hansung.ac.k.r.*
- *M. Yul Huh is with the Department of Statistics, SungKyunKwan University, Korea. E-mail: moon.huh@gmail.com.*


*Manuscript received 28 May 2014; revised 27 June 2015; accepted 2 July 2015. Date of publication 21 July 2015; date of current version 3 Dec. 2015. Recommended for acceptance by G. Karypis. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2015.2458867*

Empirical studies show that the Booster of an algorithm boosts not only the value of $Q$-statistic but also the prediction accuracy of the classifier applied.

Several studies based on resampling technique have been done to generate different data sets for classification problem [6], [34], [56], [65], [68], [73], and some of the studies utilize resampling on the feature space [6], [34], [65], [73]. The purposes of all these studies are on the prediction accuracy of classification without consideration on the stability of the selected feature subset.

The paper is organized as follows. Section 2 describes the pre-processing steps to find weakly relevant features based on t-test and to remove irrelevant features based on MI. Section 3 introduces a new evaluation criterion $Q$-statistic and investigates its properties. Section 4 gives Booster algorithm and some theoretical backgrounds are provided. Then, Section 5 presents the results of the experimentation based on synthetic data and 14 microarray data sets. Conclusion is given in Section 6.

## 2 Preprocessing Step

FS in high dimensional data needs preprocessing process to select only relevant features or to filter out irrelevant features. Relevancy of a feature is defined as follows.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ be a set of $p$ features and let $Y$ be the target feature taking one of $g$ possible classes. Then a feature $X_i$ is defined to be *strongly relevant* iff the following is satisfied [37], [40], [77]:

$$P[Y|X_i, \mathbf{X}_{-i}] \neq P[Y|\mathbf{X}_{-i}], \tag{1}$$

where $\mathbf{X}_{-i} = \mathbf{X} - \{X_i\}$ for $i = 1, \ldots, p$.

A feature $X_i$ is defined to be *weakly relevant* iff there exists a feature subset $\mathbf{X}'_{-i} \subset \mathbf{X}_{-i}$ such that the following is satisfied:

$$P[Y|X_i, \mathbf{X}_{-i}] = P[Y|\mathbf{X}_{-i}] \tag{2}$$

$$\text{and } P[Y|X_i, \mathbf{X}'_{-i}] \neq P[Y|\mathbf{X}'_{-i}]. \tag{3}$$

A feature $X_i$ is defined to be *irrelevant* iff the following is satisfied:

$$P[Y = j|X_i, \mathbf{X}'_{-i}] = P[Y = j|\mathbf{X}'_{-i}], \forall \mathbf{X}'_{-i} \subseteq \mathbf{X}_{-i}. \tag{4}$$

An efficient FS algorithm should not include *redundant* features in the selection.

A feature $X_i$ is defined to be *redundant* if it is weakly relevant and has a Markov blanket $\mathbf{M}_i$ within the current set $\mathbf{G} \subset \mathbf{X}$. $\mathbf{M}_i$ is a Markov blanket of $X_i \notin \mathbf{M}_i$ if the following is satisfied [42]:

$$P[\mathbf{X} - \mathbf{M}_i - \{X_i\}, Y|X_i, \mathbf{M}_i] \tag{5}$$

$$= P[\mathbf{X} - \mathbf{M}_i - \{X_i\}, Y|\mathbf{M}_i]. \tag{6}$$

Hence, $X_i$ is removed from $\mathbf{G} \subset \mathbf{X}$ when there exists $\mathbf{M}_i$ of $X_i$ within the current set $\mathbf{G}$. That is, the redundant features are removed from $\mathbf{G}$.

### 2.1 Finding Weakly Relevant Features by t-Test

When preprocessing is performed on the original numeric data, t-test or F-test has been conventionally applied to reduce feature space in the preprocessing step [16], [25], [35], [44], [74]. We will show that the t-test will remove irrelevant features.

Assume $g = 2$, or there are two classes and let $\mu_j = E[X|Y = j], j = 1, 2$. Then the two sample t-test is to test the equality of the two means, or $H_0 : \mu_1 = \mu_2$. Under normality, this is equivalent to the test of the equality of two distributions $f(x|Y = 1)$ and $f(x|Y = 2)$. Hence, under the null hypothesis, $X$ and $Y$ are independent of each other, or $f(x|y) = f(x)$. From Bayes rule, we have the following:

$$f(x|y) = \frac{f(x, y)}{f(y)} = \frac{f(y|x)f(x)}{f(y)}. \tag{7}$$

Since $f(x|y) = f(x)$ under the null hypothesis, we have $f(y|x) = f(y)$ under the null hypothesis. Hence, when the null hypothesis is rejected, we have $f(x|y) \neq f(x)$ with the size of type I error $\alpha$, where $\alpha = P[reject\ H_0|H_0\ is\ true]$. This shows that the feature $X$ is relevant with probability $1 - \alpha$ by setting $\mathbf{X}'_{-i} = \emptyset$ in the definition of (3).

For the problems with more than two classes, F-test can be used instead. The null hypothesis in this case is $H_0 : \mu_1 = \cdots = \mu_g, g > 2$. Under the null hypothesis, $Y$ and $X$ are independent. Hence, the features selected by the F-test are also weakly relevant with the size of type I error $\alpha$.

### 2.2 Removing Irrelevant Features by Discretization

It has been observed that MI is an equivalent concept to feature relevance [9], [20], [39], [49]. The MI between two continuous random variables $X$ and $Y$ having marginal and joint densities $f(x)$, $f(y)$ and $f(x, y)$, respectively, is defined as follows:

$$I(Y; X) = \int \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dxdy. \tag{8}$$

Similarly, the MI between the two discrete random variables $X$ and $Y$ with marginal and joint probability mass functions $f(x)$, $f(y)$ and $f(x, y)$, respectively, is defined as follows:

$$I(Y; X) = \sum \sum f(x, y) \log \frac{f(x, y)}{f(x)f(y)}. \tag{9}$$

The basic concept and properties of MI can be found in [13].

The MI estimation with numerical data involves density estimation of high dimensional data. Although much researches have been done on multivariate density estimation [8], [45], [55], [57], high dimensional density estimation with small sample size is still a formidable task.

The MI estimation based on discretized data is straightforward. In this respect, lots of researches on FS algorithms work on discretized data and huge number of researches have been done in the area of discretization. Most of the recent successful FS algorithms based on discretized data [27], [46], [50], [62], [77] utilized the well known minimum description length principle (MDLP) method [18] for discretization. Hence, this paper also uses the MDLP method for discretization.

When a feature has only one distinct value after discretization, it has MI $= 0$ with the target feature and it does not contribute any information to the target feature.

Hence, this feature is considered as an irrelevant feature and is filtered out.

Some authors suggested to filter out the feature $X_i$ if $I(Y;X_i) < \delta$ where $\delta$ corresponds to the value of the $w^{\text{th}}$ largest $I(Y;X_i)$, $w = \sqrt{p}\log(p)$, where $w$ denotes the integer not exceeding [32], [62], [77].

## 3 A New Evaluation Criterion $Q$-Statistic

Several studies have been done on the measure of the stability of the selected feature subset [40], [43], [61], [75]. A simple and straightforward measure for the similarity of a set of sequences of features $V_1, V_2, \ldots, V_h$, for a given set size $h$, is given as follows [40]:

$$K(V_1,\ldots,V_h) = \frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{j=i+1}^{h} T_{ij}, \tag{10}$$

where $T_{ij}$ is the Tanimoto distance between two sets $V_i$ and $V_j$ which are defined as follows:

$$T_{ij} = \frac{|V_i \cap V_j|}{|V_i \cup V_j|}, \tag{11}$$

where $|A|$ is the cardinality of a set $A$.

Another measure is suggested by [43] and it considers the correction for chance in selecting the feature set in addition to the cardinality of the intersection of two feature sets. It is given as follows:

$$U(V_1,\ldots,V_h) = \frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{j=i+1}^{h} U_{ij}, \tag{12}$$

where $U_{ij} = \frac{rp-u^2}{u(p-u)}$, $r = |V_i \cap V_j|$, $u = |V_i| = |V_j|$ and $p$ is the total number of features in the original full feature set.

The measure $U$, however, is targeted to the evaluation of different feature selectors based on the wrapper method with equal prefixed size of selected features. For this measure, an FS algorithm is applied to each data set to find the set of first $u$ features giving the highest accuracies.

This paper considers the filter approach for FS. For filter approach, the selection of features is performed independently of a classifier and the evaluation of the selection is obtained by applying a classifier to the selected features. The evaluation of FS in this paper is based on both the accuracy of the classifier and the stability of the selection. For this, we propose Q-statistic as follows by modifying the Tanimoto measure of (10) and (11):

$$Q(V_1,\ldots,V_h,a_1,\ldots,a_h) = \frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{j=i+1}^{h} Q_{ij}. \tag{13}$$

$Q_{ij}$ is defined as follows:

$$Q_{ij} = \frac{\sqrt{a_i a_j}|V_i \cap V_j|}{|V_i \cup V_j|}, \tag{14}$$

where $a_i$ and $a_j$ are the accuracies of a classifier based on the feature sets $V_i$ and $V_j$, respectively.

From the definition, $K \geq Q$. When a classifier gives perfect prediction accuracies ($a_i = 1$ for all $i$), $K = Q$. We can consider two special cases. First case is $V_i = V_j = V$ for all $i, j$. This gives $Q = a$ where $a$ is the accuracy of a classifier with the same $V$. Second case is $V_i \cap V_j = \emptyset$ for all $i, j$. In this case, $Q = 0$.

If an FS algorithm provides a feature subset whose accuracy is almost perfect but gives very small $Q$-statistic, the algorithm is not preferable based on the $Q$-statistic criterion.

## 4 Booster

Booster is simply a union of feature subsets obtained by a resampling technique. The resampling is done on the sample space. Assume we have training sets and test sets. For Booster, training set $\mathcal{D}$ is divided into $b-$partitions $\mathcal{D}_i, i = 1,\ldots,b$ such that $\mathcal{D} = \cup_{i=1}^{b}\mathcal{D}_i$. From these $b$ $\mathcal{D}_i$'s, we obtain $b$ training subsets $\mathcal{D}_{-i}$ such that $\mathcal{D}_{-i} = \mathcal{D} - \mathcal{D}_i, i = 1,\ldots,b$. To each of these $b$ generated training subsets, an FS algorithm $s$ is applied to obtain the corresponding feature subsets $V_i, i = 1,\ldots,b$. The subset selected by the Booster of $s$ is $V^* = \cup_{i=1}^{b} V_i$.

Booster needs an FS algorithm $s$ and the number of partitions $b$. When $s$ and $b$ are needed to be specified, we will use notation $s$-Booster$_b$. Hence, $s$-Booster$_1$ is equal to $s$ since no partitioning is done in this case and the whole data is used. When $s$ selects relevant features while removing redundancies, $s$-Booster$_b$ will also select relevant features while removing redundancies.

We now give a proof that $V^*$ will cover more relevant features in probability than the relevant features obtained from the whole data set $\mathcal{D}$. Since $V^* \supseteq V_i$ for any $i$, we have $P[v \in V^*] \geq P[v \in V_i]$ for any relevant feature $v \in V$. Since the data set $\mathcal{D}_{-i}$ is a random sample from the data $\mathcal{D}$, $V_i$ obtained from $\mathcal{D}_{-i}$ will have the same distributional property as $V_{\mathcal{D}}$ from the whole data $\mathcal{D}$. Hence, $P[v \in V^*] \geq P[v \in V_i] = P[v \in V_{\mathcal{D}}]$.

**Algorithm 1.** $s$-Booster$_b$

**Input:** Data set $\mathcal{D}$, FS algorithm $s$, number of partitions $b$
**Output:** selected feature subset $V^*$

```
Split D into b-partitions D_i, i = 1, .., b.
V* = ∅
for i = 1 to b do
    D_{-i} = D - D_i # remove D_i from D
    V_i ← s(D_{-i}) # obtain V_i by applying s on D_{-i}
    V* = V* ∪ V_i
end for
return V*
```

From the above result, we can observe that if the selected subsets $V_1,\ldots,V_b$ obtained by $s$ consist only of the relevant features where redundancies are removed, $V^*$ will include more relevant features where redundancies are removed. Hence, $V^*$ will induce smaller error of selecting irrelevant features. However, if $s$ does not completely remove redundancies, $V^*$ may result in the accumulation of larger size of redundant features.

The number of partitions $b$ plays the key factor for Booster. Larger $b$ will find more relevant features but may include more irrelevant features, and also may induce more redundant features. This is because no FS algorithm can select all relevant features while removing all irrelevant features and redundant features. Another problem with larger $b$ is more computing burden. In contrast, too small $b$ may

fail to include valuable (strong) relevant features for classification. We will investigate this problem in more detail in the next section and will suggest appropriate choice of $b$.

## 5 Experimentation

Our experimentation first filters out irrelevant features or selects weakly relevant features by the preprocessing methods described in Section 2. Three preprocessing methods explained in Section 2 are applied here, and the size of the subset of features left out after preprocessing is equal to $N = min(p_t, p_D, p_L)$ where $p_t$ is the number of features having p-value $< 0.05$ by t-test or F-test, $p_D$ is the number of features with more than two distinct values after discretization, $p_L$ is the number of preprocessed features left out by the $\delta$ criterion explained in the Section 2.2. When $N$ is decided, the preprocessed data set will consist of the features having the first $N$ largest MI's with the target, and this data set will be the input data for the Booster algorithm 1.

Three FS algorithms considered in this paper are minimal-redundancy-maximal-relevance (mRMR) [50], Fast Correlation-Based Filter (FCBF) [77], and Fast clustering-bAased feature Selection algoriThm (FAST) [62]. All three methods work on discretized data. For mRMR with large $p$ ($p$ >5,000), the size of the selection $m$ is fixed to 50 after extensive experimentations. Smaller size ($m < 30$) gives lower accuracies and lower values of $Q$-statistic while larger size ($m = 100$) gives not much improvement than $m = 50$.

The background of our choice of the three methods is that FAST is the most recent one we found in the literature and the other two methods are well known for their efficiencies. FCBF and mRMR explicitly include the codes to remove redundant features. Although FAST does not explicitly include the codes for removing redundant features, they should be eliminated implicitly since the algorithm is based on minimum spanning tree. Our extensive experiments supports that the above three FS algorithms are at least as efficient as other algorithms including CFS [27] and Relief [46].

For convenience, we will use the notation FAST-Booster, FCBF-Booster, and mRMR-Booster for the Booster of the corresponding FS algorithm.

To obtain the value of $Q$-statistic, we need a classifier. This paper considers three classifiers: Support Vector Machine (SVM) [12], k-Nearest Neighbors algorithm (KNN) [2], and Naive Bayes classifier (NB) [38]. We will first consider choosing the appropriate number of partitions $b$ for Booster. Then we will evaluate the relative performance efficiency of $s$-Booster over the original FS algorithm $s$ based on the prediction accuracy and $Q$-statistic.

To evaluate the efficiencies of the three FS algorithms—FAST, FCBF, and mRMR—and their corresponding Boosters, we apply $k$-fold cross validation. For this, $k$ training sets and their corresponding $k$ test sets are generated. For each training set, Booster is applied to obtain $V^*$. Classification is performed based on the training set with the selection $V^*$, and the test set is used for prediction accuracy. This process is repeated for the $k$ pairs of training-test sets, and the value of the $Q$-statistic is computed. In this paper, $k = 5$ is used. The flow of the evaluation process is given in Algorithm 2.

TABLE 1
Results from the Synthetic Data

| b | FAST | FCBF | mRMR |
|---|---|---|---|
| 1 | 8.93 | 9.32 | 10 |
| 2 | 10.47 | 11.79 | 11.31 |
| 3 | 13.10 | 13.47 | 14.94 |
| 5 | 17.27 | 17.62 | 18.64 |
| 10 | 19.46 | 19.73 | 20.47 |

*Average size of the feature subsets selected by the three Boosters with $b = 1, 2, 3, 5,$ and 10. $m$ for mRMR is set to 10.*

**Algorithm 2.** Evaluation process of FS

**Input:** FS algorithm $s$,
number of folds $k$, original data set $\mathcal{D}$ and $k$-folded data subsets $\mathcal{D}_i$, $i = 1, \ldots, k$.
1: **for** $i = 1$ to $k$ **do**
2: $\mathcal{D}_{-i} = \mathcal{D}$ - $\mathcal{D}_i$ # *apply $\mathcal{D}_{-i}$ to $s$-Booster$_5$*
3: $V_i^* \leftarrow f$-Booster$_5(\mathcal{D}_{-i})$
4: $a_i \leftarrow Classifier(\mathcal{D}_i)$
5: **end for**
6: Q $\leftarrow$ compute Q using $k$-pairs of $(V_i^*, a_i)$

### 5.1 Synthetic Data

In this section, Monte Carlo experimentation is performed to evaluate the usefulness of $Q$-statistic and to show the efficiency of the Booster in FS process. For the generation of synthetic data, we follow the approach of the works on generating microarray data [14], [59] and the method used here is following the works by [16], [17].

Synthetic data consists of $1,200$ features with two classes and 30 samples from each class. Each feature $X$ is generated from the following statistical model:

$$X = \begin{cases} \mu + \epsilon & \text{from class 1} \\ \epsilon & \text{from class 2.} \end{cases}$$

$\mu$ is from the following distribution:

$$\mu \sim (1 - c)\delta_0 + \frac{1}{2} c e^{-2|x|}, \tag{15}$$

where $c \in (0, 1)$ is a constant, $-\infty < x < \infty$ and $\epsilon$ is from a mixture of random noise of distributions as defined in [16]. $c$ is the percentage of significant (relevant) features among the $1,200$ features. For our data, we set $c = 0.02$. Hence, among the 1,200 features, 2 percent or 24 genes are relevant. This process is repeated 1,000 times to obtain the results.

Tables 1 and 2 give the summary of the Monte Carlo experiments for various choices of the Booster size $b$: $b = 1, 2, 3, 5,$ and 10. Table 1 gives the increase of the average size of the selected feature subsets as the Booster size $b$ increases, and Table 2 gives the accuracies and the $Q$-statistics of the three FS algorithms (FAST, FCBF, and mRMR) based on the three classifiers (SVM, KNN, and NB) for different values of $b$.

#### 5.1.1 Choice of $b$ for $s$-Booster$_b$

Table 1 shows that the average $|V^*|$ increases rapidly to 18 as $b$ increases to 5 but does not increase much after $b = 5$.

TABLE 2
Results from the Synthetic Data

| | | FAST | | | FCBF | | | mRMR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | b | SVM | KNN | NB | SVM | KNN | NB | SVM | KNN | NB |
| accuracy(%) | 1 | 95.3 | 94.2 | 96.8 | 95.3 | 94.8 | 97.2 | 97.2 | 95.9 | 97.4 |
| | 2 | 94.3 | 94.5 | 96.0 | 94.1 | 94.8 | 96.3 | 95.9 | 95.4 | 96.8 |
| | 3 | 96.7 | 95.7 | 97.8 | 96.6 | 95.9 | 97.8 | 97.3 | 96.5 | 98.2 |
| | 5 | **97.7** | 96.6 | 98.8 | **97.6** | 96.7 | **99.0** | **98.3** | 96.8 | 99.0 |
| | 10 | **97.7** | **96.8** | **99.1** | 97.5 | **97.0** | **99.0** | **98.3** | **97.1** | **99.1** |
| 100 × K | 1 | 58.5 | 58.5 | 58.5 | 63.6 | 63.6 | 63.6 | 71.5 | 71.5 | 71.5 |
| | 2 | 78.6 | 78.6 | 78.6 | 80.2 | 80.2 | 80.2 | **99.6** | **99.6** | **99.6** |
| | 3 | 85.2 | 85.2 | 85.2 | 87.9 | 87.9 | 87.9 | 98.6 | 98.6 | 98.6 |
| | 5 | 91.3 | 91.3 | 91.3 | 93.3 | 93.3 | 93.3 | 96.6 | 96.6 | 96.6 |
| | 10 | **92.0** | **92.0** | **92.0** | **93.8** | **93.8** | **93.8** | 95.9 | 95.9 | 95.9 |
| 100 × Q | 1 | 53.5 | 52.2 | 55.2 | 57.9 | 57.3 | 60.4 | 67.7 | 65.9 | 68.2 |
| | 2 | 69.9 | 70.1 | 72.5 | 71.0 | 72.0 | 74.3 | 91.7 | 90.6 | 93.4 |
| | 3 | 79.7 | 78.1 | 81.6 | 82.1 | 80.8 | 84.3 | **93.3** | **91.9** | **95.1** |
| | 5 | 87.1 | 85.0 | 89.1 | 88.9 | 87.2 | 91.2 | **93.3** | 90.5 | 94.6 |
| | 10 | **87.8** | **86.2** | **90.3** | **89.1** | **88.3** | **92.0** | 92.6 | 90.5 | 94.2 |

*Average accuracy, K-statistic and Q-statistic from $s$-Booster$_b$ with $b = 1, 2, 3, 5$, and 10 for the three FS algorithms and the three classifiers. Each value of the table is the average based on 1,000 repetitions. $s$-Booster$_1$ is the same as the original FS algorithm $s$. $m$ for mRMR is set to 10.*

The plots in Fig. 1 show that the $Q$-statistics of FAST-Booster and FCBF-Booster increase rapidly until $b = 5$ and remains stable after $b = 5$. The $Q$-statistic of mRMR-Booster, however, reaches its maximum at $b = 3$, and remains stable after $b = 3$. From these results, $b = 5$ is a safe choice for $s$-Booster$_b$.

### 5.1.2 Outperformance of mRMR-Booster

Also noted in Fig. 1 is the outperformance of mRMR-Booster over the other two Boosters, FAST-Booster and FCBF-Booster. The choice of $m$ is crucial for mRMR, where $m$ denotes the initial number of features for the algorithm. One option is to set $m = c \times p$ where $c$ is the proportion of the significant genes in $p$ features. It has been noted that generally $2 \sim 5$ percent genes are significant in microarray data [14].

For the synthetic data, $c$ was set to 0.02 and hence $m = 0.02 \times 1,200 = 24$ is suggested. When the size of the feature set is large, this rule will set $m$ too large which results in the burden of computation and interpretation. With Booster, however, we may set smaller value to $m$ because $V^* = \cup_{i=1}^{b} V_i$ has the effect of setting larger value to $m$. Monte-Carlo experiment supports this. $m = 10$ is set to obtain each $V_i$, and Table 1 shows that mRMR-Booster$_5$ gives average $|V^*| = 18.64$. The table also shows that the average $Q$-statistic is $0.65 \sim 0.68$ from mRMR with $m = 10$, and the average $Q$-statistic is $0.9 \sim 0.95$ from mRMR-Booster$_5$.

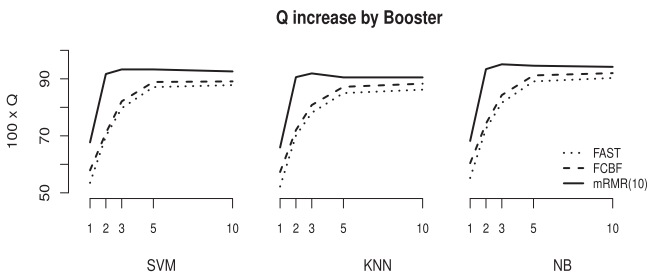



Fig. 1. Accuracy and $Q$-statistic from $s$-Booster$_b$ with $b$=1, 2, 3, 5, and 10. The values are the averages across 1,000 repetitions.

## 5.2 Real Data

Fourteen microarray data sets are considered for experiments. These are all high dimensional data sets with small sample sizes and large number of features. Among the 14 data sets, five data sets have the number of classes ($g$) larger than 2. They are summarized in Table 3. The number of features ranges from 457 to 24,482 and the sample sizes are in the range of $47 \sim 248$.

### 5.2.1 Choice of $b$ for Booster

Table 4 gives the average number of selected features ($|V^*|$) for the 14 data sets with $b = 1, 2, 3, 5, 10$, and 20. It shows the trend that the increase of $|V^*|$ is rapid up to $b = 5$ and $|V^*|$ remains relatively stable after $b = 5$.

Specifically, for mRMR-Booster, $m$ for mRMR is set to 50. Hence $|V^*| = 50$ when $b = 1$. When mRMR is boosted five times, or when mRMR-Booster$_5$ is applied, average $|V^*|$ of the mRMR-Booster$_5$ is 93. When $b > 5$, average $|V^*|$ remains almost the same. Hence, $b = 5$ is a reasonable choice as was the case with the synthetic data.

Now, we consider the effect of $b$ on the change of the accuracy and the $Q$-statistic for all the combinations of the three FS algorithms and the three classifiers. Table 5 summarizes the average accuracies and the $Q$-statistics of the 14 data sets for $b = 1, 2, 3, 5, 10$, and 20. Fig. 2 graphically summarizes Table 5. The figure has two plots. The left hand side plot gives the change of the accuracy and the right hand side plot gives the change of the $Q$-statistic depending on the values of $b$. From the Table 5 and the Fig. 2, we can observe that the improvement of accuracy and $Q$-statistic is rapid from $b = 1$ to $b = 5$, and is slow or decreases after $b = 5$. Hence, $b = 5$ is suggested for the Booster$_b$.

We can also observe that classification methods do not have much effect on the prediction accuracy and the stability of the selected feature subset. The $Q$-statistic of

TABLE 3
14 Data Sets Used in this Paper

| ID | Dataset | n | $p$ | $g$ | $p_L$ | $p_t$ | $p_D$ |
|---|---|---|---|---|---|---|---|
| D1 | B-cell1 [4] | 47 | 4,026 | 2 | 527 | 902 | 2,264 |
| D2 | coloncancer [5] | 62 | 2,000 | 2 | 340 | 478 | 135 |
| D3 | embryonal-tumours [51] | 60 | 7,129 | 2 | 749 | 521 | 69 |
| D4 | leukemia [22] | 72 | 7,129 | 2 | 749 | 2,046 | 1,012 |
| D5 | lungcancer [23] | 181 | 12,533 | 2 | 1,056 | 4,857 | 4,937 |
| D6 | prostate [60] | 136 | 12,600 | 2 | 1,060 | 5,430 | 2,185 |
| D7 | breastcancer [70] | 97 | 24,481 | 2 | 1,581 | 2,264 | 756 |
| D8 | GLI-85 [21] | 85 | 22,283 | 2 | 1,494 | 7,307 | 3,545 |
| D9 | SMK-CAN-187 [64] | 187 | 19,993 | 2 | 1,400 | 4,961 | 1,815 |
| D10 | tissue-specific DNA (christensen) [11] | 217 | 1,413 | 3 | 273 | 1,328 | 1,312 |
| D11 | TOX-171 [54] | 171 | 5,748 | 4 | 656 | 3,484 | 1,537 |
| D12 | multiple tissues (su) [66] | 102 | 5,565 | 4 | 643 | 4,659 | 3,446 |
| D13 | breastcancer (sorlie) [63] | 85 | 456 | 5 | 131 | 359 | 160 |
| D14 | leukemia (yeoh) [76] | 248 | 12,625 | 6 | 1,061 | 6,360 | 2,660 |

*$n$: the number of samples; $p$: the number of features including target feature; $p_L$: the number of features after filtering using the $\delta$ criterion explained in the Section 2.2, $p_t$: the number of features after filtering based on t-test or F-test, $\alpha = 0.05$; $p_D$: the number of features with more than two distinct values after discretization; $g$: the number of class categories.*

mRMR-Booster is far better than the $Q$-statistic of the other two algorithms for all $b$ considered in this paper.

#### 5.2.2 Efficiency of Booster

Tables 6 and 7 give detailed results of the accuracies and the $Q$-statistics for all combinations of the three FS algorithms and three classifiers. Tables 8 and 9 give the rate of the increase of accuracy and $Q$-statistic by the Booster with $b = 5$. From now on, $b = 5$ is the default value assigned to a Booster if there is no ambiguity.

Fig. 3 graphically presents the effect of $s$-Booster on accuracy and $Q$-statistic against the original $s$'s. Classifier used here is NB. Separate plots are drawn for the data sets with $g = 2$ and $g > 2$. Upper two plots are for the comparison of the accuracies and the lower two plots are for the comparison of the $Q$-statistics: $y$-axis is for $s$-Booster and $x$-axis is for $s$. Hence, if a point lies above $y = x$ line, $s$-Booster is more efficient than $s$. Since three FS algorithms are considered for each of the 14 data sets, there are 42 cases in each plot.

#### 5.2.3 Booster Boosts Accuracy

Tables 6 and 8 demonstrate that mRMR-Booster improves accuracy considerably: overall average accuracy increases from $0.91$ to $0.96$. One interesting point to note here is that mRMR-Booster is more efficient in boosting the accuracy of the original mRMR when it gives low accuracies. Table 6 shows that data sets giving three lowest accuracies with mRMR are D7, D9, and D11: 0.76, 0.71, and 0.62, respectively. Table 8, however, shows that these three data sets give highest increase rates of accuracies with mRMR-Booster: 1.23, 1.13, and 1.24, respectively.

From the two tables, we can observe that FAST-Booster also improves accuracy, but not as high as mRMR. For FCBF-Booster, overall average accuracy remains the same, but the average accuracy for the data sets with $g > 2$ decreases by $0.4$ percent.

TABLE 4
Average Size of the Feature Subset from $s$-Booster$_b$ for the Three FS Algorithms with $b = 1, 2, 3, 5, 10,$ and $20$

| b | FAST | FCBF | mRMR |
|---|---|---|---|
| 1 | 32.8 | 63.6 | 50.0 |
| 2 | 47.0 | 101.2 | 68.4 |
| 3 | 58.7 | 122.2 | 80.4 |
| 5 | 74.0 | 144.9 | 93.1 |
| 10 | 79.1 | 156.0 | 97.3 |
| 20 | 79.4 | 158.1 | 98.4 |

#### 5.2.4 Booster Boosts $Q$-Statistic

Table 7 shows that mRMR is outstanding in its performance on the $Q$-statistic over FCBF and FAST as we have already noticed with the synthetic data. Overall average is 0.44: 0.38 for the data sets with $g = 2$ and 0.57 for the data sets with $g > 2$.

FCBF gives poor performance on $Q$-statistic in contrast to its high performance on accuracy. Overall average is 0.28: 0.20 for the data sets with $g = 2$ and 0.42 for the data sets with $g > 2$.

FAST gives quite poor performance on $Q$-statistic. The highest value for the data sets with $g = 2$ is 0.28 (D6), and most of the values are below 0.1.

Fig. 3 graphically demonstrates that Booster improves the $Q$-statistic for all the cases considered except the case with the data set D6.

The improvement by Booster is generally more significant for the data sets with $g = 2$ than for the data sets with $g > 2$. This is because of the fact that the $Q$-statistic from original FS algorithm gives higher value for $g > 2$ than for $g = 2$.

Now, consider the improvement of the $Q$-statistic by mRMR-Booster. From Table 9, the rate of overall increase is 1.40: 1.53 for the data sets with $g = 2$ and 1.16 for the data sets with $g > 2$. Specifically, for mRMR-Booster, overall average $Q$-statistic is 0.62: 0.581 for the data sets with $g = 2$ and 0.661 for the data sets with $g > 2$.

An interesting observation is that the $Q$-statistic for D7 is extremely low by mRMR: 0.075, 0.077, and 0.075 for SVM, KNN, and NB, respectively (Table 7). Table 9, however, shows that mRMR-Booster gives extremely high

TABLE 5
Accuracy and $Q$-Statistic from $s$-Booster$_b$ for the Three FS Algorithms and the Three Classifiers with $b = 1, 2, 3, 5, 10,$ and $20$

| | | FAST | | | FCBF | | | mRMR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | b | SVM | KNN | NB | SVM | KNN | NB | SVM | KNN | NB |
| accuracy (%) | 1 | 87.9 | 87.9 | 88.8 | 94.3 | 92.8 | 94.9 | 90.2 | 89.8 | 90.6 |
| | 2 | 89.8 | 88.6 | 91.0 | 93.7 | 92.4 | 93.5 | 93.5 | 92.1 | 93.5 |
| | 3 | 90.6 | 89.9 | **91.8** | 94.4 | **93.9** | 94.0 | 94.2 | 93.1 | 94.1 |
| | 5 | 90.9 | 89.8 | 91.5 | 94.6 | 93.3 | **95.3** | **94.9** | 92.9 | **94.4** |
| | 10 | 90.8 | 90.3 | **91.8** | 94.9 | 93.5 | 94.7 | 94.1 | 92.9 | 94.2 |
| | 20 | **91.5** | **90.7** | **91.8** | **95.1** | 93.4 | 94.8 | 93.9 | **93.3** | 93.8 |
| 100 × Q | 1 | 16.4 | 16.2 | 16.8 | 27.6 | 26.8 | 27.7 | 43.7 | 43.2 | 44.4 |
| | 2 | 17.6 | 17.4 | 18.4 | 29.3 | 28.7 | 29.7 | 44.1 | 43.1 | 44.7 |
| | 3 | 20.4 | 20.1 | 21.1 | 33.7 | 33.7 | 33.8 | 48.5 | 47.6 | 49.0 |
| | 5 | 22.1 | 21.7 | 22.7 | 38.2 | 37.7 | 39.2 | **54.6** | 52.7 | 54.7 |
| | 10 | 23.6 | **23.7** | **24.4** | **40.7** | **39.7** | **40.9** | 54.1 | **53.2** | **54.9** |
| | 20 | **23.7** | 23.6 | 24.1 | 40.5 | 39.4 | 40.8 | 53.7 | 53.0 | 54.0 |

*Each value is the averages over the 14 data sets.*

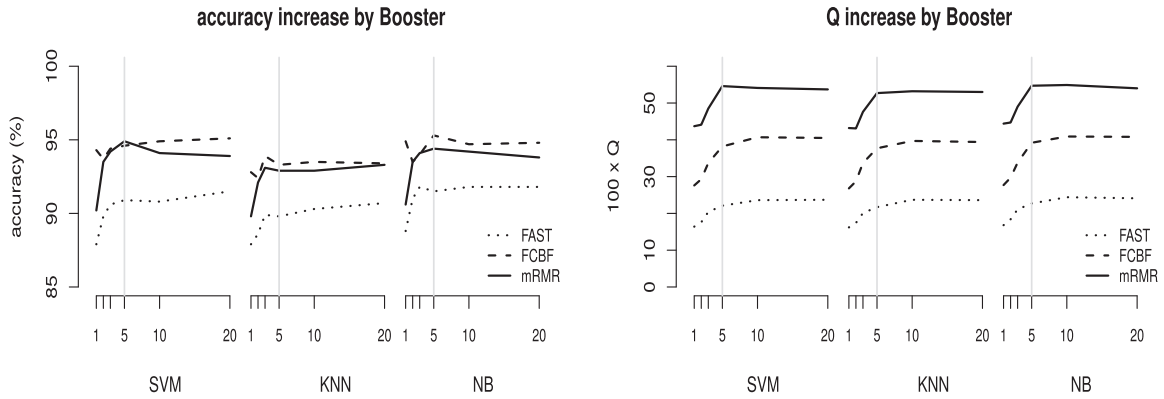



Fig. 2. Accuracy and $Q$-statistic of $s$-Booster$_b$ for $b = 1, 2, 3, 5, 10,$ and 20 ($x$-axis). Each value is the average over the 14 data sets. $s$-Booster$_1$ is $s$. The grey vertical line is for $b = 5$.

TABLE 6
Accuracies Obtained by the Three Classifiers Based on the Features Selected by the Three FS Algorithms: FAST, FCBF, and mRMR

| | FAST | | | FCBF | | | mRMR | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | SVM | KNN | NB | SVM | KNN | NB | SVM | KNN | NB |
| D1 | 0.77 | 0.86 | 0.84 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| D2 | 0.85 | 0.85 | 0.89 | **0.90** | **0.90** | **0.90** | 0.89 | **0.90** | 0.87 |
| D3 | 0.80 | 0.80 | 0.82 | 0.93 | 0.88 | 0.92 | 0.93 | 0.90 | **0.95** |
| D4 | **1.00** | **1.00** | **1.00** | 0.97 | 0.97 | 0.99 | **1.00** | 0.99 | 0.99 |
| D5 | **0.99** | **0.99** | **0.99** | **0.99** | **0.99** | **0.99** | **0.99** | **0.99** | **0.99** |
| D6 | 0.84 | 0.84 | 0.84 | 0.93 | 0.91 | 0.91 | **0.95** | 0.94 | 0.94 |
| D7 | 0.94 | 0.91 | 0.93 | 0.94 | 0.87 | **0.95** | 0.76 | 0.77 | 0.76 |
| D8 | **0.96** | 0.95 | 0.95 | 0.93 | 0.94 | 0.95 | 0.92 | 0.93 | 0.93 |
| D9 | 0.76 | 0.76 | 0.72 | 0.80 | 0.74 | **0.85** | 0.76 | 0.76 | 0.71 |
| D10 | 0.99 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.99 | 0.99 | 0.99 |
| D11 | 0.58 | 0.60 | 0.61 | 0.87 | 0.85 | **0.90** | 0.66 | 0.62 | 0.68 |
| D12 | 0.99 | 0.95 | **1.00** | **1.00** | **1.00** | **1.00** | 0.99 | 0.94 | 0.99 |
| D13 | 0.86 | 0.80 | 0.86 | **0.98** | 0.95 | 0.94 | 0.91 | 0.88 | 0.93 |
| D14 | 0.97 | **0.98** | **0.98** | 0.96 | 0.97 | **0.98** | 0.89 | 0.95 | 0.94 |
| average g = 2 | 0.88 | 0.89 | 0.89 | 0.93 | 0.91 | 0.94 | 0.91 | 0.91 | 0.91 |
| average $g > 2$ | 0.88 | 0.87 | 0.89 | 0.96 | 0.95 | 0.97 | 0.89 | 0.88 | 0.91 |
| overall average | 0.88 | 0.88 | 0.89 | 0.94 | 0.93 | 0.95 | 0.90 | 0.90 | 0.91 |

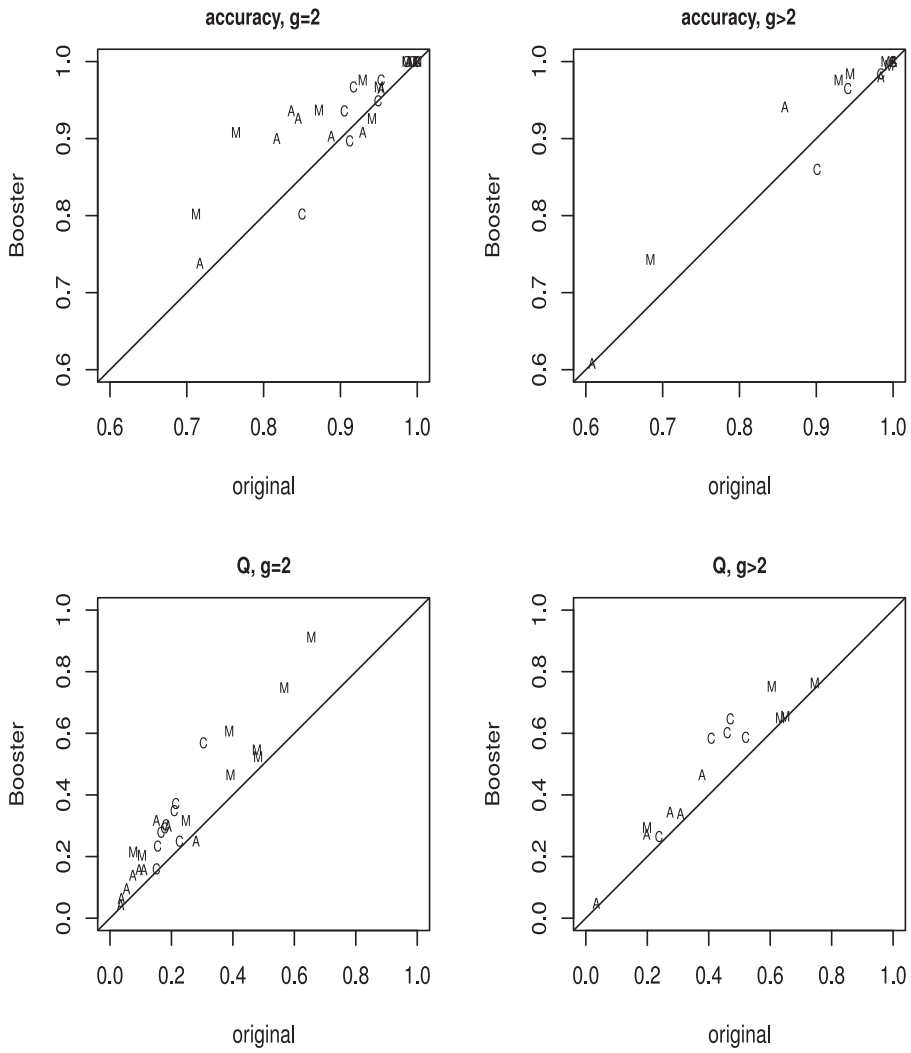



Fig. 3. Comparison of $s$-Booster$_5$ over $s$ for prediction accuracy and $Q$-statistic. Plots are drawn separately for the data sets with $g = 2$ and $g > 2$. Classifier used is NB. "A" stands for FAST, "C" for FCBF, and "M" for mRMR.

TABLE 7
$Q$-Statistics Obtained by the Three Classifiers Based on the Features Selected by the Three FS Algorithms: FAST, FCBF, and mRMR

| | FAST | | | FCBF | | | mRMR | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | SVM | KNN | NB | SVM | KNN | NB | SVM | KNN | NB |
| D1 | 0.05 | 0.06 | 0.05 | 0.18 | 0.18 | 0.18 | **0.39** | **0.39** | **0.39** |
| D2 | 0.10 | 0.10 | 0.11 | 0.21 | 0.21 | 0.21 | 0.40 | **0.42** | 0.39 |
| D3 | 0.14 | 0.14 | 0.15 | 0.32 | 0.28 | 0.30 | 0.63 | 0.59 | **0.66** |
| D4 | 0.10 | 0.10 | 0.10 | 0.17 | 0.17 | 0.18 | **0.49** | 0.48 | 0.48 |
| D5 | 0.19 | 0.19 | 0.19 | 0.21 | 0.21 | 0.21 | **0.57** | **0.57** | **0.57** |
| D6 | 0.28 | 0.28 | 0.28 | 0.23 | 0.23 | 0.23 | **0.49** | 0.48 | 0.48 |
| D7 | 0.08 | 0.07 | 0.07 | 0.16 | 0.14 | **0.17** | 0.08 | 0.08 | 0.08 |
| D8 | 0.04 | 0.04 | 0.04 | 0.15 | 0.15 | 0.15 | 0.24 | **0.25** | **0.25** |
| D9 | 0.04 | 0.04 | 0.04 | 0.13 | 0.12 | **0.15** | 0.12 | 0.12 | 0.10 |
| D10 | 0.37 | 0.38 | 0.38 | 0.47 | 0.47 | 0.47 | 0.74 | 0.74 | **0.75** |
| D11 | 0.03 | 0.03 | 0.03 | 0.22 | 0.21 | **0.24** | 0.18 | 0.16 | 0.20 |
| D12 | 0.30 | 0.28 | 0.31 | 0.46 | 0.46 | 0.46 | **0.65** | 0.59 | **0.65** |
| D13 | 0.20 | 0.17 | 0.20 | 0.44 | 0.42 | 0.41 | 0.58 | 0.55 | **0.61** |
| D14 | 0.27 | 0.27 | 0.27 | 0.50 | 0.51 | 0.52 | 0.56 | **0.64** | 0.63 |
| average g $= 2$ | 0.11 | 0.11 | 0.11 | 0.20 | 0.19 | 0.20 | 0.38 | 0.37 | 0.38 |
| average $g > 2$ | 0.23 | 0.23 | 0.24 | 0.42 | 0.41 | 0.42 | 0.54 | 0.53 | 0.57 |
| overall average | 0.16 | 0.15 | 0.16 | 0.28 | 0.27 | 0.28 | 0.44 | 0.43 | 0.44 |

TABLE 8
Ratio of the Accuracy from $s$-Booster$_5$ to the Accuracy from $s$

| | FAST | | | FCBF | | | mRMR | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | SVM | KNN | NB | SVM | KNN | NB | SVM | KNN | NB |
| D1 | **1.18** | 1.09 | 1.12 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 |
| D2 | 1.06 | 1.02 | 1.02 | 1.05 | **1.07** | 1.03 | 1.05 | 1.00 | **1.07** |
| D3 | **1.15** | 1.04 | 1.10 | 1.00 | 1.00 | 1.05 | 1.04 | 1.00 | 1.02 |
| D4 | 1.00 | 1.00 | 1.00 | **1.03** | **1.03** | 1.01 | 1.00 | 1.00 | 1.01 |
| D5 | **1.01** | **1.01** | **1.01** | 1.00 | 1.00 | **1.01** | 1.00 | **1.01** | **1.01** |
| D6 | **1.07** | 1.05 | 1.10 | 0.98 | 1.03 | 0.98 | 0.99 | 1.00 | 0.98 |
| D7 | 0.97 | 1.01 | 0.98 | 1.01 | 1.02 | 1.00 | **1.23** | 1.16 | 1.19 |
| D8 | 1.01 | 1.00 | 1.01 | **1.05** | 1.01 | 1.02 | **1.05** | 1.04 | 1.05 |
| D9 | 1.06 | 0.97 | 1.03 | 1.05 | 1.02 | 0.94 | 1.10 | 1.06 | **1.13** |
| D10 | **1.00** | **1.00** | **1.00** | 0.99 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| D11 | 1.03 | 1.05 | 1.00 | 0.97 | 0.91 | 0.95 | **1.24** | 1.17 | 1.09 |
| D12 | 1.01 | 1.03 | 1.00 | 1.00 | 0.99 | 1.00 | 1.01 | **1.05** | 1.01 |
| D13 | 1.03 | 1.07 | 1.10 | 0.94 | 0.99 | 1.03 | 1.04 | **1.08** | 1.05 |
| D14 | 0.97 | 0.99 | 1.00 | 0.98 | 1.00 | 1.00 | **1.08** | 1.02 | 1.04 |
| average g $= 2$ | 1.06 | 1.02 | 1.04 | 1.02 | 1.02 | 1.01 | 1.05 | 1.03 | 1.05 |
| average $g > 2$ | 1.008 | 1.029 | 1.018 | 0.975 | 0.978 | 0.996 | 1.074 | 1.066 | 1.038 |
| overall average | 1.04 | 1.02 | 1.03 | 1.00 | 1.01 | 1.00 | 1.06 | 1.04 | 1.05 |

improvement on the $Q$-statistic for all three cases. It shows that the increase rate for the three classifiers is $305$, $273$, and $285$ percent, respectively.

All the codes in this paper are programmed in R [52]. Booster and FAST codes are programmed by the authors, mRMR is from [36], FCBF is implemented in Weka [72], SVM and NB are from [48] and KNN is from [71]. The computing burden of Booster depends upon the FS algorithm applied. The choice of $b = 5$ consumes five times more computing time of the original algorithm.

## 6 Conclusion

This paper proposed a measure $Q$-statistic that evaluates the performance of an FS algorithm. $Q$-statistic accounts both for the stability of selected feature subset and the prediction accuracy. The paper proposed Booster to boost the performance of an existing FS algorithm. Experimentation with synthetic data and 14 microarray data sets has shown that the suggested Booster improves the prediction accuracy and the $Q$-statistic of the three well-known FS algorithms: FAST, FCBF, and mRMR. Also we have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and $Q$-statistic. Especially, the performance of mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and $Q$-statistic.

It was observed that if an FS algorithm is efficient but could not obtain high performance in the accuracy or the $Q$-statistic for some specific data, Booster of the FS algorithm will boost the performance. However, if an FS algorithm itself is not efficient, Booster may not be able to obtain high performance. The performance of Booster depends on the performance of the FS algorithm applied.

TABLE 9
Ratio of the $Q$-Statistic from $s$-Booster$_5$ to the $Q$-Statistic from $s$

| | FAST | | | FCBF | | | mRMR | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | SVM | KNN | NB | SVM | KNN | NB | SVM | KNN | NB |
| D1 | **1.92** | 1.63 | 1.77 | 1.66 | 1.66 | 1.66 | 1.19 | 1.14 | 1.19 |
| D2 | 1.51 | 1.47 | 1.43 | 1.79 | **1.86** | 1.74 | 1.50 | 1.35 | 1.56 |
| D3 | **2.32** | 1.90 | 2.10 | 1.67 | 1.69 | 1.87 | 1.44 | 1.34 | 1.39 |
| D4 | 1.66 | 1.66 | 1.66 | **1.69** | **1.69** | 1.65 | 1.11 | 1.11 | 1.14 |
| D5 | 1.56 | 1.56 | 1.58 | 1.66 | 1.66 | **1.67** | 1.31 | 1.32 | 1.32 |
| D6 | 0.86 | 0.85 | 0.89 | 1.10 | **1.20** | 1.11 | 1.11 | 1.12 | 1.09 |
| D7 | 1.85 | 2.05 | 1.88 | 1.71 | 1.76 | 1.68 | **3.05** | 2.73 | 2.85 |
| D8 | **1.76** | 1.72 | 1.73 | 1.59 | 1.48 | 1.51 | 1.29 | 1.25 | 1.28 |
| D9 | 1.35 | 1.12 | 1.24 | 1.32 | 1.23 | 1.06 | 1.85 | 1.71 | **1.96** |
| D10 | 1.23 | 1.23 | 1.23 | 1.35 | **1.38** | **1.38** | 1.02 | 1.03 | 1.02 |
| D11 | 1.52 | 1.62 | 1.41 | 1.14 | 1.02 | 1.11 | **1.93** | 1.72 | 1.48 |
| D12 | 1.13 | 1.17 | 1.10 | **1.31** | 1.28 | **1.31** | 1.01 | 1.09 | 1.01 |
| D13 | 1.21 | 1.32 | 1.37 | 1.20 | 1.32 | **1.43** | 1.22 | 1.32 | 1.24 |
| D14 | 1.19 | 1.24 | **1.26** | 1.08 | 1.14 | 1.13 | 1.11 | 0.99 | 1.03 |
| average g $= 2$ | 1.64 | 1.55 | 1.59 | 1.58 | 1.58 | 1.55 | 1.54 | 1.45 | 1.53 |
| average $g > 2$ | 1.25 | 1.32 | 1.27 | 1.22 | 1.23 | 1.27 | 1.26 | 1.23 | 1.16 |
| overall average | 1.50 | 1.47 | 1.47 | 1.45 | 1.46 | 1.45 | 1.44 | 1.37 | 1.40 |

If Booster does not provide high performance, it implies two possibilities: the data set is intrinsically difficult to predict or the FS algorithm applied is not efficient with the specific data set. Hence, Booster can also be used as a criterion to evaluate the performance of an FS algorithm or to evaluate the difficulty of a data set for classification.

## Acknowledgments

The authors are grateful to the anonymous reviewers for their comments that improved the quality of our paper. This research was financially supported by Hansung University. Moon Yul Huh is the corresponding author of this paper.

## References

[1] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[2] D. Aha and D. Kibler, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.

[3] S. Alelyan, "On feature selection stability: A data perspective," PhD dissertation, Arizona State Univ., Tempe, AZ, USA, 2013.

[4] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. M. Izidore, S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. H. Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[5] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci.*, vol. 96, no. 12, pp. 6745–6750, 1999.

[6] F. Alonso-Atienza, J. L. Rojo-Alvare, A. Rosado-Muñoz, J. J. Vinagre, A. Garcia-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1956–1967, 2012.

[7] P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[8] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *The Ann. Statist.*, vol. 38, no. 5, pp. 2916–2957, 2010.

[9] G. Brown, A. Pocock, M. J. Zhao, and M. Lujan, "Conditional likelihood maximization: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 27–66, 2012.

[10] C. Kamath, *Scientific data mining: a practical perspective*, Siam, 2009.

[11] B. C. Christensen, E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, H. H. Nelson, M. R. Karagas, J. F. Padbury, R. Bueno, D. J. Sugarbaker, R.-F. Yeh, J. K. Wiencke, and K. T. Kelsey, "Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context," *PLOS Genetics*, vol. 5, no. 8, e1000602, 2009.

[12] C. Corinna and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Series in Telecommunications and Signal Processing), 2nd ed. Hoboken, NJ, USA: Wiley, 2002.

[14] D. Dembele, "A flexible microarray data simulataion model," *Microarrays*, vol. 2, no. 2, pp. 115–130, 2013.

[15] D. Dernoncourt, B. Hanczar, and J. D. Zucker, "Analysis of feature selection stability on high dimension and small sample data," *Comput. Statist. Data Anal.*, vol. 71, pp. 681–693, 2014.

[16] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.

[17] J. Fan, P. Hall, and Q. Yao, "To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied?," *J. Am. Statist. Assoc.*, vol. 102, no. 480, pp. 1282–1288, 2007.

[18] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," *Artif. Intell.*, vol. 13, no. 2, pp. 1022–1027, 1993.

[19] A. J. Ferreira and M. A. T. Figueiredo, "Efficient feature selection filters for high dimensional data," *Pattern Recog. Lett.*, vol. 33, no. 13, pp. 1794–1804, 2012.

[20] B. Franay, G. Doquire, and M. Verleysen, "Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification," *Neurocomputing*, vol. 112, pp. 64–78, 2013

[21] W. A. Freije, F. E. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, and L. M. Liau, "Gene expression profiling of gliomas strongly predicts survival," *Cancer Res.*, vol. 64, no. 18, pp. 6503–6510, 2004.

[22] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Am. Assoc. Advancement Sci.*, vol. 286, no. 5439, pp. 531–537, 1999.

[23] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res.*, vol. 62, no. 17, pp. 4963–4967, 2002.

[24] G. Gulgezen, Z. Cataltepe, and L. Yu, "Stable and accurate feature selection," in *Proc. Mach. Learn. Knowl. Discovery Databases*, pp. 455–468, 2009.

[25] G. Guoan, G. G. Tarek, H. Chiang-Ching, G. T. Dafydd, A. S. Kerby, M. G. T. Jeremy, L. R. K. Sharon, E. M. David, J. G. Thomas, D. I. Mark, B. O. Mark, M. H. Samir, and G. B. David, "Proteomic analysis of lung adenocarcinoma: Identification of a highly expressed set of proteins in tumors," *Clinical Cancer Res.*, vol. 8, no. 7, pp. 2298–2305, 2002.

[26] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[27] M. A. Hall, "Correlation-based feature selection for machine learning," Department of Computer Science, PhD dissertation, The Univ. of Waikato, Hamilton, New Zealand, 1999.

[28] Y. Han and L. Yu, "A variance reduction framework for stable feature selection," *Statist. Anal. Data Mining*, vol. 5, no. 5, pp. 428–445, 2012.

[29] T. Hastie, *The Elements of Statistical Learning*, New York, NY, USA: Springer, 2009.

[30] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Comput. Biol. Chem.*, vol. 34, no. 4, pp. 215–225, 2010.

[31] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings Bioinf.*, vol. 9, no. 2, pp. 102–118, 2008.

[32] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, and W. Pedrycz, "Measuring relevance between discrete and continuous features based on neighborhood mutual information," *Expert Syst. with Appl.*, vol. 38, no. 9, pp. 10737–10750, 2011.

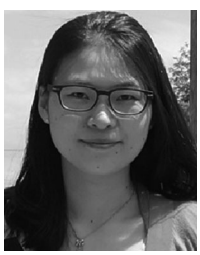
[33] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognit.*, vol. 42, no. 3, pp. 409–424, 2009.

[34] W. L. Hung, M. S. Yang, and D. H. Chen, "Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1317–1325, 2008.

[35] P. Jafari and F. Azuaje, "An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors," *BMC Med. Inf. Decision Making*, vol. 6, no. 27, 2006.

[36] N. D. Jay, S. Papillon-Cavanagh, C. Olsen, G, Bontempi, and B. Haibe-Kains, "mRMRe: An R package for parallelized mRMR ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, 2012.

[37] G. H. John, R. Kohavi, and K. Pfleger," Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf. Mach. Learn.*, vol. 94, pp. 121–129, 1994.
[38] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, pp. 338–345, 1995.
[39] R. V. Jorge and A. E. Pablo, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014.
[40] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.
[41] I. Kojadinovic, "Relevance measures for subset variable selection in regression problems based on k-additive mutual information," *Comput. Statist. Data Anal.*, vol. 49, no. 4, pp. 1205–1227, 2005.
[42] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Mach. Learn.*, pp. 284–292, 1996.
[43] L. I. Kuncheva, "A stability index for feature selection," in *Proc. Artif. Intell. Appl.*, pp. 421–427, 2007.
[44] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics Series*, vol. 13, pp. 51–60, 2002.
[45] H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman, "Forest density estimation," *The J. Mach. Learn. Res.*, vol. 12, pp. 907–951, 2011.
[46] R. S. Marko, and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, no. 1–2, pp. 23–69, 2003.
[47] N. Meinshausen and P. Buhlmann, "Stability selection," *J. Roy. Statist. Soc.: Series B (Statist. Methodol.)*, vol. 72, no. 4, pp. 417–473, 2010.
[48] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "e1071: Misc functions of the department of statistics (e1071), TU Wien," 2012.
[49] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 261–274, 2008.
[50] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[51] L. P. Scott, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.
[52] R Core Team. (2014). R: A Language and environment for statistical computing, r foundation for statistical computing, Vienna, Austria [Online]. Available: http://www.R-project.org/
[53] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[54] S. A. Sajan, J. L. Rubenstein, M. E. Warchol, and M. Lovett, "Identification of direct downstream targets of Dlx5 during early inner ear development," *Human Molecular Genetics*, vol. 20, no. 7, pp. 1262–1273, 2011.
[55] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Hoboken, NJ, USA: Wiley, 2009.
[56] H. Silva and A. Fred, "Feature subspace ensembles: A parallel classifier combination scheme using feature selection," *Multiple Classifier Syst.*, vol. 4472, pp. 261–270, 2007.
[57] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Boca Raton, FL, USA: CRC Press, 1986.
[58] C. Sima, S. Attoor, U. Braga-Neto, J. Lowey, E. Suh, and E. R. Dougherty, "Impact of error estimation on feature selection," *Pattern Recognit.*, vol. 38, no. 12, pp. 2472–2482, 2005.
[59] S. Singhal, C. G. Kyvernitis, S. W. Johnson, L. R. Kaiser, M. N. Liebman, and S. M. Albelda, "Microarray data simulator for improved selection of differentially expressed genes," *Cancer Biology and therapy*, vol. 2, no. 4, pp. 383–391, 2003.
[60] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.
[61] P. Somol and J. Novovicova, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1921–1939, Nov. 2010.
[62] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.
[63] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Nat. Acad. Sci.*, vol. 98, no. 19, pp. 10869–10874, 2001.
[64] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y. M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M. E. Lenburg, and J. S. Brody, "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nature Med.*, vol. 13, no. 3, pp. 361–366, 2007.
[65] J. Stefanowski, "An experimental study of methods combining multiple classifiers-diversified both by feature selection and bootstrap sampling," *Issues Representation Process. Uncertain Imprecise Inf.*, Akademicka Oficyna Wydawnicza, Warszawa, pp. 337–354, 2005.
[66] A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch, "Large-scale analysis of the human and mouse transcriptomes," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 7, pp. 4465–4470, 2002.
[67] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.
[68] K. M. Ting, J. R. Wells, S. C. Tan, S. W. Teng, and G. I. Webb, "Feature-subspace aggregating: Ensembles for stable and unstable learners," *Mach. Learn.*, vol. 82, no. 3, pp. 375–397, 2011.
[69] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control Comput.*, pp. 368–377, 1999.
[70] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, R. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.
[71] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York, NY, USA: Springer, 2003.
[72] I. Witten and E. Frank, *Data Mining-Practical Machine Learning Tools and Techniques with JAVA Implementations*, San Mateo, CA, USA: Morgan Kaufmann, 2000.
[73] T. Windeatt, M. Prior, N. Effron, and N. Intrator, "Ensemble-based feature selection criteria," in *Proc. MLDM Posters*, pp. 168–182, 2007.
[74] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.
[75] W. Yao and Q. Wang, "Robust variable selection through MAVE," *Comput. Statist. Data Anal.*, vol. 63, pp. 42–49, 2013.
[76] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, and L. Wong JR, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.
[77] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *The J. Mach. Learn. Res.*, vol. 5, no. 2, pp. 1205–1224, 2004.



**HyunJi Kim** received the PhD degree in statistics from SungKyunKwan University, Korea, in 2014. She is currently a researcher at Korea Fair Trade Mediation Agency. The research was carried out while she was researcher at the Research Institute of Applied Statistics, SungKyunKwan University. Her research interests include high-dimensional data analysis, feature selection, data mining, machine learning, and classification.

**Byong Su Choi** received the MSc and PhD degrees in computational statistics from SungKyunKwan University in Seoul, Korea, in 1982 and 1991, respectively. He is currently a professor in the Department of Multimedia Engineering, Hansung University, Seoul, Korea. His research interests include multimedia processing, data engineering, and statistical computing.

**Moon Yul Huh** graduated from Seoul National University and received the PhD degree from Southern Methodist University in statistics. He is a professor Emeritus in the Department of Statistics, SungKyunKwan Universityr of Seoul and the research was carried out while he was visiting the Department of Statistics and Probability at Michigan State University for one year. He served as the president of The Korean Statistical Society. His major interests are in multivariate statistics, statistical learning, and data visualization.