Robust Feature Selection for Microarray Data Based on Multicriterion Fusion

Feng Yang and K.Z. Mao

Abstract—Feature selection often aims to select a compact feature subset to build a pattern classifier with reduced complexity, so as to achieve improved classification performance. From the perspective of pattern analysis, producing stable or robust solution is also a desired property of a feature selection algorithm. However, the issue of robustness is often overlooked in feature selection. In this study, we analyze the robustness issue existing in feature selection for high-dimensional and small-sized gene-expression data, and propose to improve robustness of feature selection algorithm by using multiple feature selection evaluation criteria. Based on this idea, a multicriterion fusion-based recursive feature elimination (MCF-RFE) algorithm is developed with the goal of improving both classification performance and stability of feature selection results. Experimental studies on five gene-expression data sets show that the MCF-RFE algorithm outperforms the commonly used benchmark feature selection algorithm SVM-RFE.

Index Terms—Feature selection, multicriterion fusion, recursive feature elimination, robustness, classification.

1 Introduction

Feature selection plays an increasingly important role in machine learning and data mining with emerging of high-dimensional data such as microarray gene-expression data. Feature selection for gene-expression data, also known as gene selection, mainly serves two purposes. First, feature selection is to identify certain disease-related genes. Second, feature selection is to find a compact set of discriminative genes to build a pattern classifier with reduced complexity and improved generalization capabilities. Depending on the purpose of gene selection, two types of feature selection algorithms including ranking-based feature selection and set-based feature selection are employed in microarray gene-expression data analysis [1], [3], [18], [22], [26], [28], [33]. In ranking-based feature selection, features are evaluated on an individual basis, without considering inter-relationship between features in general, while set-based feature selection evaluates features based on their role in a feature set by taking into account dependency between features.

Gene selection might seem to be a straightforward application of feature selection techniques to gene-expression data, but the problem is not so simple. Due to high dimensionality (as high as a few thousands) and small sample size (as small as a few dozens) of gene-expression data, gene selection encounters problems that are not commonly seen in feature selection for conventional data having relatively low dimensionality and large sample size. One issue encountered is the validity of cross validation. Cross validation is a widely used error estimation method, but the study in [5] revealed that cross validation-based error estimation could exhibit large variance when applied to small sized data. Another problem that was often overlooked is the ties problem associated with classification error-based feature evaluation [45]. Under small sample size, classification error has too few possible values to distinguish different features, and this in turn results in selection uncertainty. Peaking phenomenon also becomes a non-neglectable problem under high dimensionality and small sample size [7], [21].

Robustness or stability of feature selection for high-dimensional and small-sized data also received attentions in recent years. Robustness or stability of a feature selection algorithm refers to its sensitivity to varying conditions such as perturbations of training data. If a feature selection algorithm lacks robustness, it might produce unrepeatable results even only a few samples are added to or deleted from the training data set. Even without perturbation of training data different feature selection algorithms usually produce different selection results. The inconsistent gene selection results thus produced could cause confusions to biological researchers and result in loss of their enthusiasm and confidence in applying machine learning techniques to solve biological problems.

In the literature, just a few work explores the robustness issue of feature selection. Kalousis et al. [23] examined three categories of existing stability measures in high-dimensional space and constructed stability profiles for some well-known feature selection algorithms; Somol and Novovicovae [38] proposed a new consistency measure with reduced computational complexity; Krizek et al. [27] developed an entropy-based measure for stability assessment; Gulgezen et al. [19] studied the stability and classification accuracy of Minimum Redundancy Maximum Relevance-based feature selection. Besides the above work that analyzes robustness of existing feature selection algorithms, some new algorithms aiming to improve robustness of feature selection have also been developed. For example, Saeys et al. [34] proposed an instance perturbation-based ensemble scheme for single

- *The authors are with the Division of Control and Instrumentation, School of Electrical and Electronic Engineering, College of Engineering, Nanyang Technological University, 50 Nanyang Avenue, S1-B4b-06, Biomedical Electronics Lab, Singapore 639798.*
E-mail: yang0159@e.ntu.edu.sg; ekzmao@ntu.edu.sg.

Manuscript received 5 Nov. 2009; revised 6 Apr. 2010; accepted 13 July 2010; published online 15 Oct. 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2009-11-0203.
Digital Object Identifier no. 10.1109/TCBB.2010.103.

1545-5963/11/$26.00 © 2011 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

feature selection; Yu et al. [44] proposed a general feature selection framework based on dense feature groups and developed Dense Relevant Attribute Group Selector (DRAGS); Loscalzo et al. [30] proposed a consensus group-based framework and developed Consensus Group Stable Feature Selection algorithm (CGS). Both DRAGS and CGS could achieve good stability without sacrificing classification performance.

In this study, we propose to improve robustness of feature selection through fusion of multiple criteria for feature evaluation. Based on this idea, a multicriterion fusion-based recursive feature elimination (MCF-RFE) algorithm is developed. As revealed in [23], [34], the various feature selection algorithms are sensitive even to a minor variation of training samples. We believe this might be due to inaccurate estimation of statistical parameters such as sample mean and standard deviation employed in the feature evaluation criterion. By using multiple criteria, the merit evaluation of features tends to be less sensitive to the inaccurate estimation of the statistical parameters, and hence, the robustness of the feature selection algorithm is improved. In addition, the proposed new algorithm alleviates the disagreement between different feature selection algorithms by getting their consensus, and hence, improves the credibility of the selected features. Experimental studies on five gene-expression data sets show that the new MCF-RFE algorithm produces feature subsets with good stability and classification accuracy.

The rest of this paper is organized as follows: In Section 2, the robustness issue is illustrated by a case study and a stability measure is introduced. Section 3 explains the basic idea of multicriterion fusion as well as basic fusion methods. In Section 4, the detailed new feature selection algorithm MCF-RFE is described. In Section 5, experimental results and discussions are presented. The last, Section 6, concludes this study.

2 Issue of Robustness in Gene Selection

2.1 Illustration of Robustness Issue in Fisher's Ratio-Based Feature Selection Under Training Data Perturbation

Ideally, a feature selection algorithm should be robust enough to produce very similar feature subsets even if the training data are subject to perturbations such as addition or deletion of a few samples. But this may not necessarily be true when the data are with high dimensionality and small sample size. To illustrate this point, the following experiment was conducted. First, a perturbation to training data was performed by randomly removing l samples from the training data to produce a perturbed training data. Second, Fisher's ratio-based feature selection algorithm (please refer to Section 4 for details) was applied to the perturbed training data to select top 10 features. Assume S_0 and S_l are the feature subsets selected using the original training data and the perturbed training data with l samples removed. The number of common features in S_0 and S_l reflects the sensitivity of the feature selection algorithm to training data perturbation. A feature selection algorithm is considered to be robust or insensitive to perturbation of training data if the features in S_0 and S_l are similar.

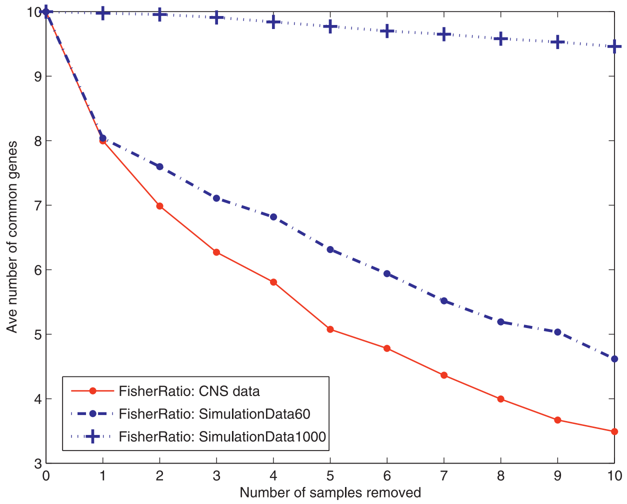


Fig. 1. Average number of common genes versus number of samples removed when a subset of 10 genes is selected.

To obtain reliable evaluation, the above experiment was repeated 300 times for each l value and the average number of common features in S_0 and S_l was calculated. Fig. 1 shows the results of Fisher's ratio-based feature selection on CNS gene-expression data, which has 60 samples and 7,129 genes (details of the data are given in Section 5). From Fig. 1, it is observed that when only 1 sample is removed, the number of common genes among the top 10 genes is reduced from 10 to 8, on average, and the number is further reduced to 5, when 5 samples are removed. This result tells us that the commonly used Fisher's ratio is actually sensitive to training data perturbation.

To further investigate the robustness of Fisher's ratio-based feature selection, experimental study on simulated data was also conducted. Two artificial data sets containing 7,129 features with 60 and 1,000 samples were generated, respectively, using the Matlab code in Guyon [16]. As shown in Fig. 1 (the curve "FisherRatio: SimulationData60"), the feature selection results of Fisher's ratio on the simulated data set with 60 samples are not stable. But for the data set with 1,000 training samples (see the curve "FisherRatio: SimulationData1000" in Fig. 1), the selection results are less sensitive to removal of training data. This result explains why robustness is not an issue in feature selection for data with large sample size. But for small sized data like gene-expression data, the robustness of feature selection is indeed an issue that should be considered seriously because such a minor variation/perturbation of training data is very likely to occur, in practice, such as addition of a few new cases or samples.

Fig. 1 shows the results of Fisher's ratio as an example, but our study found that almost all commonly used feature selection algorithms are prone to perturbation of small-sized training data. It is thus, necessary to develop robust feature selection algorithms to alleviate the problem.

2.2 Robustness Measures

The robustness (stability) of a feature selection algorithm can be evaluated based on its ability to select repeated features, given different batches of data under the same

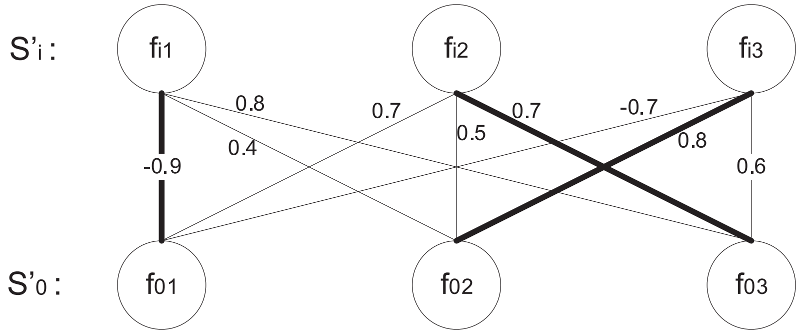


Fig. 2. Computation of sum of correlation.

distribution. Since the true distribution of real data is usually unknown and only a number of samples are available during the learning process, the different batches of data can be generated through resampling.

Assume S_i and S_0 denote feature subsets selected using the ith batch of resampled data and the full data, respectively. The similarity between the two feature subsets can be measured using Jaccard Index (or Tanimoto Index) [23], [34]

$$J_i(k) = \frac{|S_i \bigcap S_0|}{|S_i \bigcup S_0|}, \tag{1}$$

where k is the cardinality of S_i and S_0, $|S_i \cap S_0|$ is the number of common features between S_i and S_0 and $|S_i \cup S_0|$ is the total number of features without reduplication in S_i and S_0.

The above similarity measure takes into account only the common features between two feature subsets. But there exist a great number of features that are highly correlated in gene-expression data. In order to give a more general and precise measure of the similarity between two feature subsets, we employ the following similarity index $JC(\in [0,1])$ that takes into account the correlations (Pearson correlation in this study) between the different features of two feature subsets:

$$JC_i(k) = \frac{|S_i \bigcap S_0| + SC_i}{k}, \tag{2}$$

where SC_i is the sum of absolute correlation values between the dissimilar features from S_i and S_0.

The idea behind the index is illustrated in Fig. 2, where S_i' and S_0' are two feature subsets after removing the common features between S_i and S_0, and each node represents a feature and each edge denotes the correlation between the corresponding features. The final sum of correlations between feature subsets S_i' and S_0' is

$$\begin{aligned} SC_i &= |Corr(f_{i1}, f_{01})| + |Corr(f_{i3}, f_{02})| + |Corr(f_{i2}, f_{03})| \\ &= |-0.9| + |0.8| + |0.7| = 2.4, \end{aligned}$$

where $|Corr(f_{ij}, f_{0j})|$ is the absolute value of correlation between features f_{ij} and f_{0j}. The above similarity index resembles the one proposed in [44], where SC_i is calculated in an optimal way. For computational simplicity, SC_i is computed using the greedy search algorithm in this study.

Assume totally m batches of data are generated by resampling and m feature subsets are selected. The robustness or stability measure of the feature selection algorithm is defined as

$$\overline{JC}(k) = \frac{\sum_{i=1}^{m} JC_i(k)}{m}. \tag{3}$$

For the experiment study in Section 2.1, the robustness indices calculated by (3) are shown in Fig. 3, where the number of removed samples is fixed to 10. These results are in line with those in Fig. 1.

The above robustness index provides a sensible evaluation of the stability of a feature selection algorithm. But it should be noted that a robust feature selection may not necessarily guarantee good classification performance because the measure is independent of a classification model. In practice, both stability and classification performance should be considered when evaluating a feature selection algorithm because a stable but classification-ineffective selection result does not make any sense.

3 Robust Feature Selection Based on Fusion of Multiple Criteria

3.1 The Motivation of Using Multicriterion Fusion

In pattern classification, it is well acknowledged that combining or integrating multiple classifiers, especially uncorrelated weak ones could greatly improve the classification performance [10], [11], [24], [40]. Motivated by the

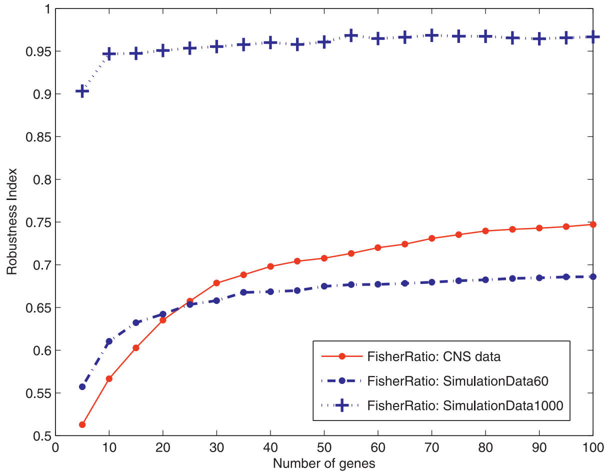


Fig. 3. Robustness index when 10 samples are removed in gene selection.

TABLE 1
Top 10 Genes Selected by Five Different Feature Selection Criteria on CNS Data Set with all 60 Samples as Training Data

	Top 10 genes									
Criterion	1	2	3	4	5	6	7	8	9	10
Fisher'ratio	327	348	2196	2695	2496	1352	1320	3320	844	3645
Relief	1431	1791	6165	1615	5978	3389	2912	43	1281	6891
ADC	4588	1320	3185	3783	327	2695	2854	1474	3731	4174
AW-SVM	4503	6252	3389	5812	2914	1697	2700	942	4546	4247
SVM-RFE	2671	5581	5389	2404	4576	5061	6555	2914	1478	2917

success of multiple classifier combination, in this study, we propose to improve the robustness of feature selection through integrating multiple feature selection criteria (algorithms). The reasons for integrating multiple criteria are manifold. First, different feature selection algorithms produce different feature subsets. Table 1 shows the top 10 genes (denoted by their serial numbers in the table) selected by five different feature selection criteria, including Fisher's ratio, Relief, ADC (asymmetric dependency coefficient), AW-SVM (absolute weight of SVM), and SVM-RFE, the details of which can be found in the following Section 3.2 and Section 4. Obviously, the selection results are mostly different. Among the different selection results, which result is superior and should be adopted? Actually, there is no agreed ways to decide. Integrating different "opinions" from multiple feature selection criteria to yield a consensus seems to be a reasonable solution. Second, a model built upon weak assumptions usually performs more robust than a model built upon stringent assumptions. The existing feature selection criteria are generally built upon certain assumption(s) of data distribution. But the distribution of the learning data is usually rather complicated (e.g., a mixture of many different distributions) and unknown. Even the distribution is precisely available, it may violate the assumptions at certain extent. A criterion that aggregates multiple feature selection criteria can help to weaken the assumptions, and consequently, improve the robustness. Third, feature subsets produced by different feature selection criteria may exhibit complementary effects because of the nonindependence among features, and thus, a fusion of these feature subsets may produce a better representation in feature space to describe the data. Fourth, each feature selection criterion usually has its own specific but restrained ability to search in the feature space, and thus, may be stuck at a local optimum, while fusion of multiple criteria utilizes and aggregates the search abilities of each of the criteria to obtain a wider "vision" that may help to get closer to a global optimal solution.

3.2 Basis Criteria

Many feature evaluation criteria have been proposed in the literature. However, it is unnecessary or impractical to use all of them in the criterion fusion. In this study, the criteria selected for fusion are named as *basis criteria*.

It was found that if two feature selection criteria produce similar results, fusion of such two criterion does not help. This finding provides us a guideline for basis criteria selection: basis criteria should exhibit diversity. The deeper reason for diversity is that the diverse results produced by the diverse multiple basis criteria complement each other. Another reason for using diverse basis criteria is to prevent the fusion result from being dominated by criteria that produce similar results.

For computational simplicity and performance diversity analyzed above, the basis criteria used in this study are Fisher's ratio, Relief, ADC (asymmetric dependency coefficient), and AW-SVM (absolute weight of SVM). Fisher's ratio is a univariate filter method evaluating each feature individually, while Relief is a multivariate filter method taking into account dependencies between features. ADC is a information theory-based filter method and AW-SVM is an embedded method that ranks features based on their corresponding coefficients in the SVM classifier. Details of the four basis criteria are presented below.

3.2.1 Fisher's Ratio

Fisher's ratio is an individual feature evaluation criterion that measures the discriminative power of a feature j by the ratio of interclass difference to intraclass spread

$$FR(j) = \frac{(\widehat{\mu}_{j1} - \widehat{\mu}_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \tag{4}$$

where $\widehat{\mu}_{jc}$ is the sample mean of feature j within class c and σ_{jc}^2 is the variance of feature j within class c, for $c = 1, 2$. The larger the FR value, the more discriminative the feature is.

3.2.2 Relief

Relief is a weight-based feature ranking method inspired by the instance-based learning [32]. It evaluates the "$Relevance$" of features through multiple iterations. At each iteration, a sample $\mathbf{x}$ is first randomly selected from the data set and its nearest-Hit $\mathbf{x}_{\mathrm{H}}$ (nearest neighbor from the same class) and nearest-Miss $\mathbf{x}_{\mathrm{M}}$ (nearest neighbor from other class) are identified, and the relevance of all features are then updated using the difference between the nearest-Hit and nearest-Miss as follows:

$$W(j) = W(j) - \frac{diff(j, \mathbf{x}, \mathbf{x}_{\mathrm{H}})}{n} + \frac{diff(j, \mathbf{x}, \mathbf{x}_{\mathrm{M}})}{n}, \tag{5}$$

where $W(j)$ is the relevance of feature j to the targets and it is initialized to zero; $diff(j, \mathbf{x}, \mathbf{x}')$ denotes the difference of feature j between samples $\mathbf{x}$ and $\mathbf{x}'$. For continuous features, $diff(j, \mathbf{x}, \mathbf{x}')$ is the actual difference normalized to interval $[0, 1]$

$$diff(j, \mathbf{x}, \mathbf{x}') = \frac{|x_j - x_j'|}{x_{j\max} - x_{j\min}}, \tag{6}$$

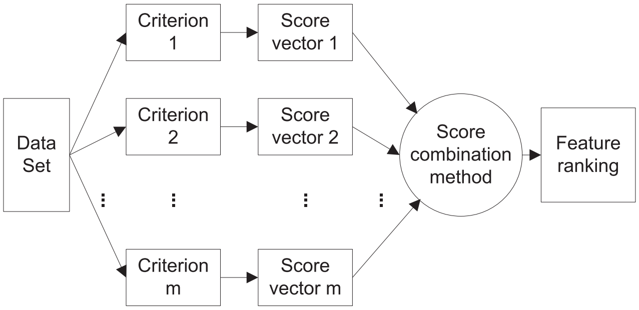


Fig. 4. Score-based multicriterion fusion.

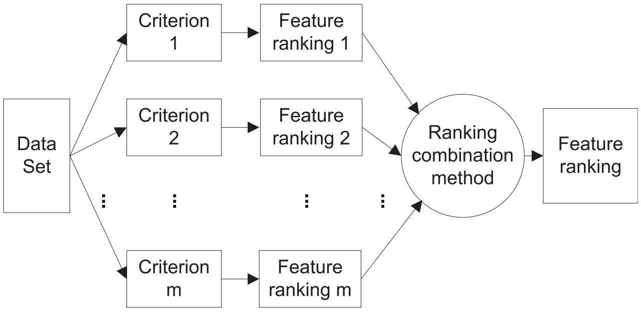


Fig. 5. Ranking-based multicriterion fusion.

where x_j and x'_j are the values of feature j in $\mathbf{x}$ and $\mathbf{x}'$, $x_{j\max}$ and $x_{j\min}$ are the maximal and minimal values of feature j in all samples.

The updating process in (5) repeats n times for a training data set with sample size n to get the final relevance.

3.2.3 ADC

The Asymmetric Dependency Coefficient (ADC) is a feature ranking method that measures the dependency of class label Y on a feature j (corresponding variable is denoted by X_j) using information gain [36]

$$ADC(Y, j) = \frac{MI(Y, X_j)}{H(Y)}, \tag{7}$$

where $H(Y)$ is the entropy of Y and $MI(Y, X_j)$ is the mutual information between label Y and feature j defined as

$$H(Y) = -\sum_y p(Y = y) \log p(Y = y), \tag{8}$$

$$H(X_j) = -\sum_x p(X_j = x) \log p(X_j = x), \tag{9}$$

$$MI(Y, X_j) = H(Y) + H(X_j) - H(Y, X_j). \tag{10}$$

3.2.4 AW-SVM

Support vector machine (SVM) is a popular classification algorithm suitable for high-dimensional data because of its insensibility to data dimensionality [9], [13]. There are several variants of the basic SVM. In this work, the *linear binary* SVM with soft margin is employed because of its good performance for gene-expression data. The linear SVM classifier is, in fact, a hyper plane defined by [42]:

$$\sum_{j=1}^{p} w_j x_j + b_0 = 0, \tag{11}$$

where p is the total number of features and w_j is the weight of feature j. The weight w_j indicates the importance of feature j, and hence, the absolute weight of SVM (AW-SVM) can be used to evaluate and rank the features.

3.3 Basic Fusion Methods

In this study, we used two fusion methods including score-based multicriterion fusion and ranking-based multicriterion fusion.

3.3.1 Score-Based Multicriterion Fusion

In score-based multicriterion fusion, each basis criterion first produces a score vector containing scores of all features, a score combination algorithm is then employed to aggregate the multiple score vectors into one consensus score vector, and a feature ranking procedure is finally performed to rank the features based on their consensus scores. The score-based multicriterion fusion procedure is illustrated in Fig. 4.

In score aggregating, it is essential to ensure that the scores produced by different basis criteria are comparable. Thus, score normalization should be done before score combination is performed. In this study, the scores produced by each basis criterion are normalized to the range of $[0, 1]$. Assume $\mathbf{u}_i$ is the score vector produced by basis criterion i, the score normalization is performed as follows:

$$\mathbf{u}_i' = \frac{\mathbf{u}_i - \mathbf{u}_{i\min}}{\mathbf{u}_{i\max} - \mathbf{u}_{i\min}}, \tag{12}$$

where $\mathbf{u}_{i\min}$ and $\mathbf{u}_{i\max}$ are the minimum and maximum values in vector $\mathbf{u}_i$.

For all the basis criteria, it is assumed that the larger the score, the better the feature. A simple yet effective score combination method is to take the average of the normalized scores

$$\mathbf{u} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{u}_i', \tag{13}$$

where m is the number of basis criteria used in fusion.

3.3.2 Ranking-Based Multicriterion Fusion

Ranking-based multicriterion fusion is to integrate multiple feature selection criteria according to the feature rankings. In the fusion process, a basis criterion first produces a feature ranking, where each feature has a *position* (or *order*) value; and then, a ranking combination algorithm is applied onto all feature rankings to generate the final consensus ranking. The ranking-based multicriterion fusion procedure is illustrated in Fig. 5.

Compared to the score-based fusion in Fig. 4, the ranking-based fusion requires a basis criterion to produce a feature ranking rather than a feature score vector.

Ranking combination (usually termed as rank aggregation) is a common problem in many fields and it has been extensively studied [12], [20], [25], [29]. Among the existing aggregation methods, *Borda count*, which is originally a voting method based on rankings is simple yet effective rank aggregation method [41], [43]. Suppose, there are m voters (i.e., basis criteria in this study) and a fix set of p candidates (i.e., features). In Borda count, each voter i first

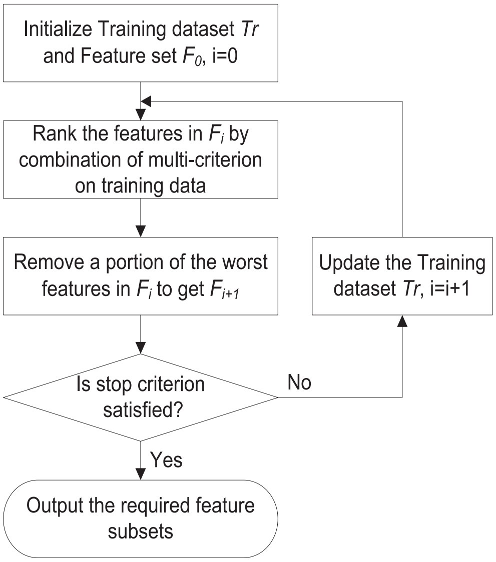

Initialize Training dataset Tr and Feature set F0, i=0
Rank the features in Fi by combination of multi-criterion on training data
Remove a portion of the worst features in Fi to get Fi+1
Update the Training dataset Tr, i=i+1
Is stop criterion satisfied?
No
Yes
Output the required feature subsets

Fig. 6. The procedure of MCF-RFE.

gives points to all candidates to generate a point vector $\mathbf{v}_i$ as follows: the top ranked candidate is given p points, the second ranked candidate is given $p-1$ points, and so on. The final points of candidates are the sum of points from the m voters

$$\mathbf{v} = \sum_{i=1}^{m} \mathbf{v}_i \tag{14}$$

and the aggregated ranking is obtained by descendingly sorting the final points in $\mathbf{v}$.

4 Multicriterion Fusion Based Recursive Feature Elimination (MCF-RFE) Algorithm

Section 3 presents a robust feature evaluation and ranking method based on multicriterion fusion. If the purpose of feature selection is to improve classification, the above feature ranking method may not necessary be a good choice. It is well acknowledged that a collection of the best features does not necessarily produce the best feature subset. In order to obtain a feature subset to produce good classification results, the multicriterion fusion-based feature evaluating method must be combined with a search strategy.

Recursive feature elimination (RFE) is a frequently used search strategy in feature selection, see, for example, [17]. The RFE search procedure can be briefly summarize as follows:

RFE Procedure:

1. Given the full feature set F_0, set $i = 0$.
2. Evaluate the merit of each feature in the feature set F_i.
3. Remove the least important feature(s) from F_i to obtain feature set F_{i+1}.
4. Set $i = i + 1$ and goes to Step 2 until a stopping criterion is satisfied.

The RFE algorithm generates a group of nested feature subsets, i.e., $F_0 \supset F_1 \supset F_2 \ldots$. The original RFE eliminates one feature at each iteration and could be computational intensive if it is applied to high-dimensional data such as gene-expression data. For computational efficiency, a variant of the RFE is to eliminate a portion of the features at each iteration (e.g., 50 percent). Based on the RFE strategy and SVM, Guyon et al. [17] proposed a feature subset selection algorithm SVM-RFE (Support Vector Machine Recursive Feature Elimination), where the merit of a feature is evaluated in terms of its corresponding coefficient in the SVM classifier. SVM-RFE produces feature subsets leading to good classification performance and is often used as a benchmark algorithm [14], [39]. Despite its popularity, SVM-RFE lacks robustness. Motivated by strengths of multicriterion fusion, we formulate a robust feature selection algorithm by combining the multicriterion fusion-based feature evaluation and the RFE search strategy. We name the new algorithm as MCF-RFE (Multicriterion Fusion-based Recursive Feature Elimination) whose procedure is described in Fig. 6.

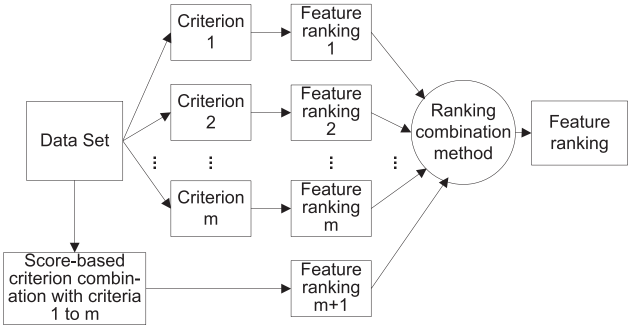


Fig. 7. Score and ranking-based multicriterion fusion.

In multiple-criteria fusion, both score-based and ranking-based fusion methods are used: a score-based fusion method is first used to generate a feature ranking, which is then added to the m feature rankings produced by individual basis criteria. After that, the $m+1$ feature rankings are aggregated by a combination to generate the final feature ranking. The fusion procedure is illustrated in Fig. 7.

5 Experiments and Results

5.1 Data Sets

In the experimental study, extensive experiments were conducted on the following five gene-expression data sets:

Colon Data [2]: This type of data were first described by Alon et al., in 1999. The data set contains expression levels of 2,000 genes for 62 samples including 22 normal samples, and 40 colon cancer samples. The task is to distinguish between *normal* and *tumor* samples. The original data can be downloaded from http://microarray.princeton.edu/oncology/.

Leukemia Data [15]: Introduced by Golub et al., in 1999, this data set contains expression levels of 7,129 genes for 47 ALL (Acute lymphoblastic leukemia) leukemia patients and 25 AML (Acute myelogenous leukemia) leukemia patients. The original data can be downloaded at: http://www-genome. wi.mit.edu/cgi-bin/cancer/datasets.cgi.

Prostate Data [37]: This data set contains expression level of 12,600 genes for 136 samples including 77 prostate tumors, and 59 normal samples. The data set was first described by Singh et al., and the original data are available at: http://www-genome.wi.mit.edu/cgi- bin/cancer/datasets.cgi.

CNS Data [31]: The goal of this study is the molecular investigation of treatment effectiveness for embryonal CNS

TABLE 2
Data Sets Characteristics

Name	# Class 1	# Class 2	# Features	*SDR*
Colon	22	40	2000	3.1%
Leukemia	25	47	7129	1.01%
Prostate	59	77	12600	1.08%
CNS	21	39	7129	0.84%
DLBCL	19	58	7129	1.08%

(Central Nervous System) tumors. The task is to distinguish between *failed* and *succeed* treatment outcomes. There are 60 patients with 7,129 genes in this data set, where 21 patients are survivors and 39 patients are failures. The original data are available at: http://www-genome.wi.mit.edu/mpr/CNS/.

DLBCL Data [35]: This set of data contains 58 DLBCL (diffuse large b-cell lymphoma) samples and 19 FL (Follicular Lymphoma) samples with 7,129 genes. The data are available at: http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi.

The five data sets are briefly summarized in Table 2, where SDR denotes the sample-to-dimension ratio.

5.2 Experimental Setup

In the experiment, linear SVM was employed as the pattern classifier. The linear SVM was implemented using the LIBSVM [6] toolbox of version 2.88 and parameter C was empirically set to 0.01 for the first two data sets and 0.1 for the last three data sets. Due to small sample size of the data sets, classification error was estimated using *.632* bootstrap [8] with 300 repeats. Considering the class imbalance in the gene-expression data sets, AUC (area under the ROC curve) [4] was also employed in the evaluation and comparative study of the feature selection algorithms.

Besides assessment in classification performance, feature selections algorithms were also evaluated in terms of their capacity to deal with training data perturbation based on the robustness measure (3).

For a pattern classification system without a feature selection component, the variance of classification error estimation reflects the sensitivity of a classifier to variations in training and testing samples. For a pattern classification system that includes a feature selection component, although the variance of classification error estimation is the combined effects of both the feature selection algorithm and the pattern classifier, variance still can be used as an indication of the robustness of the feature selection algorithm. This is because a feature selection algorithm sensitive to training data variation usually produces feature subsets leading to large variance in the classification error estimation. For this reason, the standard deviation of the classification error estimation for each feature selection algorithm was investigated in the experiment.

In addition to linear SVM, KNN ($K = 3$) was also used in the experiment. It was found that KNN classification results could obviously support the conclusions made from linear SVM classification results. Due to page limitation, we only present the classification results by linear SVM in this paper. For classification results by KNN, readers can refer to the supplementary files on the website: http://www3.ntu.edu.sg/home2006/yang0159/RobustFS.htm.

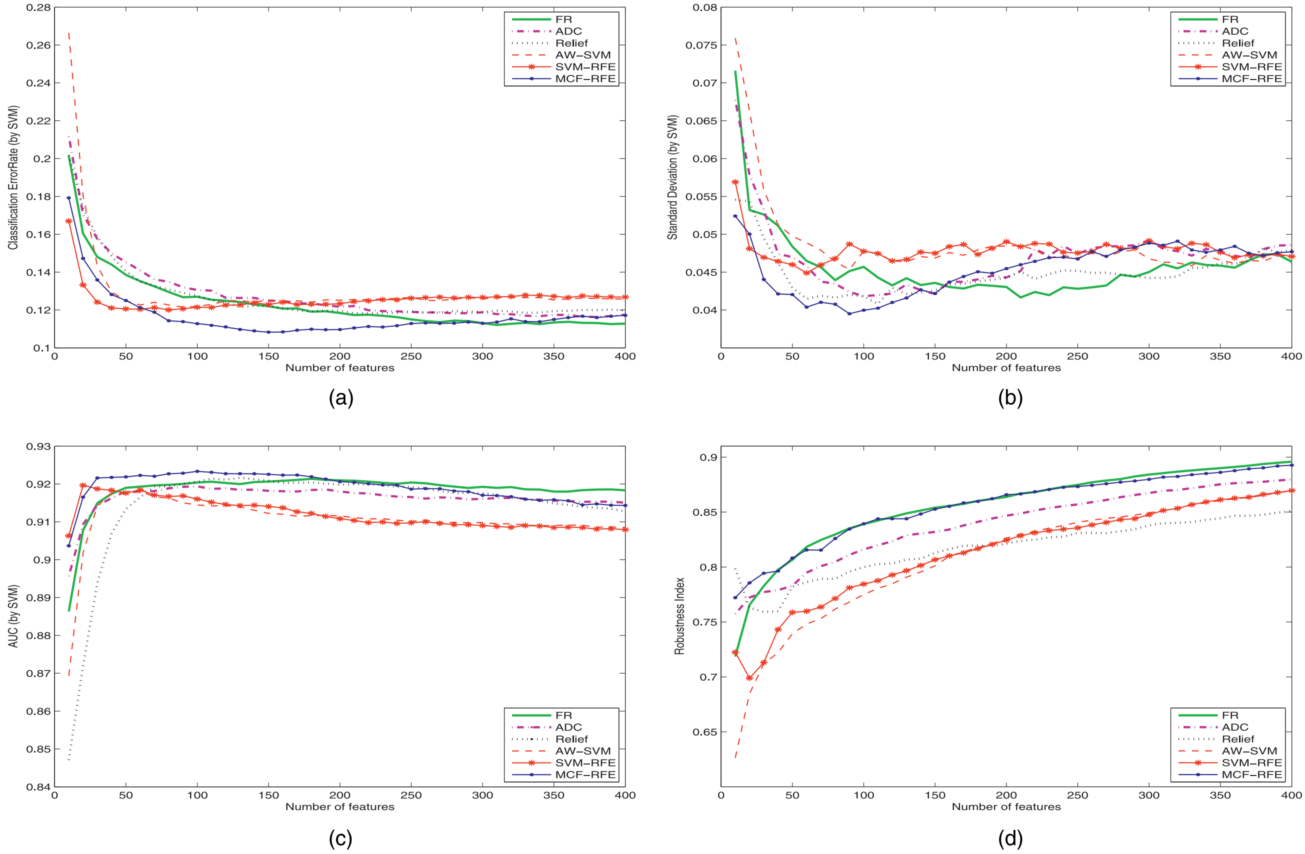
5.3 Comparative Study of MCF-RFE with Basis Criteria and SVM-RFE

Figs. 8, 9, 10, 11, and 12 show the classification results including estimated classification error, standard deviation of classification error estimation and AUC values, as well as the feature stability of six feature selection methods including Fisher's ratio (FR), ADC, Relief, AW-SVM, SVM-RFE, and MCF-RFE, on the five data sets.

We first compare MCF-RFE with the four ranking-based basis criteria: Fisher's ratio (FR), ADC, Relief, and AW-SVM. From the estimated classification error on the five data sets (Figs. 8a, 9a, 10a, 11a, and 12a), we can observe that MCF-RFE outperforms the basis criteria on all five data sets. For example, on DLBCL data (Fig. 12a), MCF-RFE produces the least classification error of 2.6 percent with only 120 features, while the best basis criterion for DLBCL data, Fisher's ratio, produces an error of 3.8 percent with the same number of features, and achieves its least classification error of 2.8 percent with 290 features. The AUC results (see Figs. 8c, 9c, 10c, 11c, and 12c) are in line with those of the classification error except on CNS data (Fig. 11c) where MCF-RFE is slightly interior to AW-SVM when the number of selected features is less than 130. The comparison of the standard deviation of error estimation (Figs. 8b, 9b, 10b, 11b, and 12b) also proves the effectiveness of MCF-RFE. Take again the results on DLBCL data as an example (see Fig. 12b), the standard deviation is 0.024 for MCF-RFE and 0.031 for Fisher's ratio when 120 features are selected. As for feature stability, we find that MCF-RFE does not perform the best but it produces a compromised result of the four basis criteria and the difference between MCF-RFE, and the stablest basis criterion is not much.

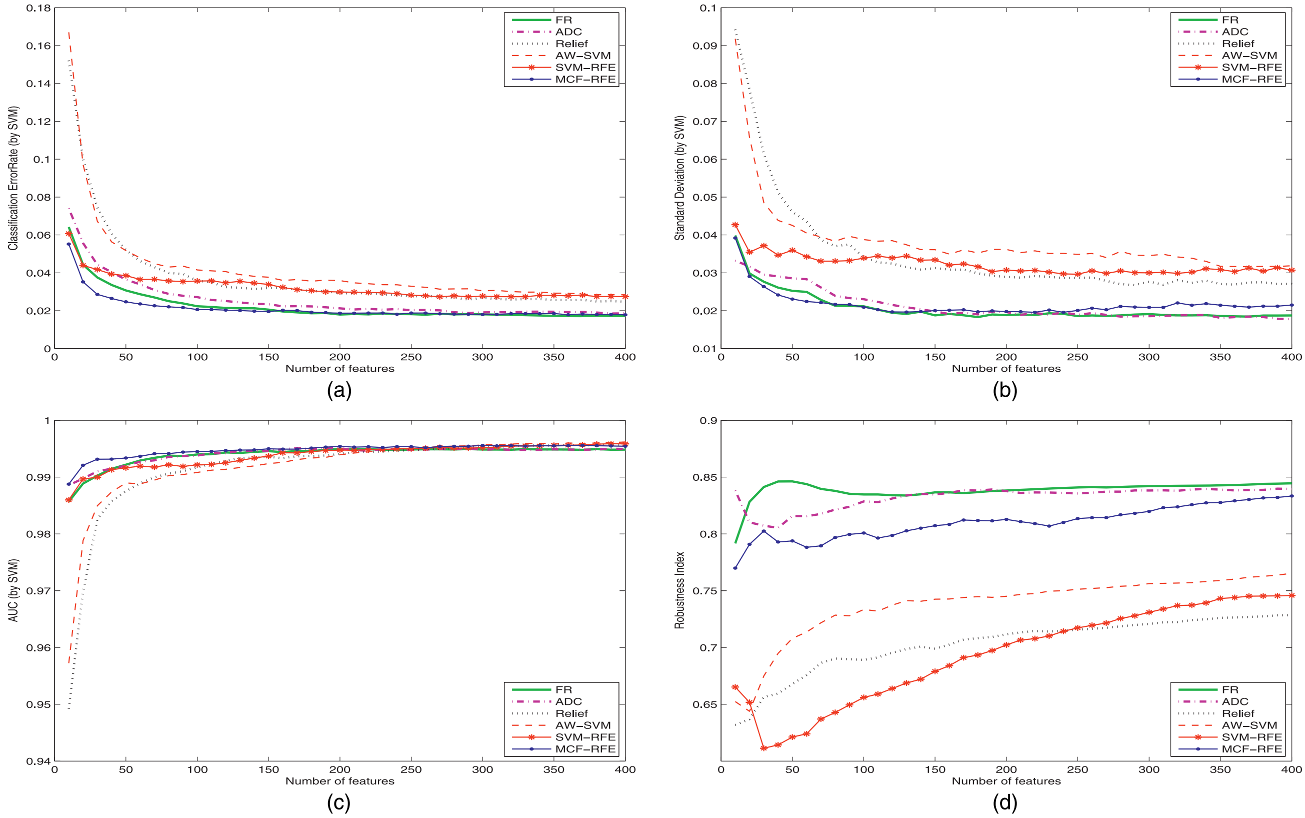
In evaluating a feature selection algorithm, both classification performance and feature stability should be considered. But classification performance should be the first consideration and feature stability should be the secondary because a stable but classification-ineffective selection result does not make any sense. Based on the above considerations, we think the MCF-RFE outperforms the basis criteria because it produces feature subsets with better classification performance and reasonably good stability.

When it comes to the performance comparison between MCF-RFE and SVM-RFE, we observe that MCF-RFE achieves substantial improvements over SVM-RFE in the feature stability performance on all five data sets (refer to Figs. 8d, 9d, 10d, 11d, and 12d). MCF-RFE also produces better classification performance than SVM-RFE except the AUC on CNS data. Here, we still take the results on DLBCL data (see Fig. 12) as an example. With a subset of 120 features, the respective values of estimated classification error, standard deviation of error estimation, AUC and feature stability of MCF-RFE versus SVM-RFE are 2.6 percent versus 3.6 percent, 0.024 versus 0.030, 0.996 versus 0.992 and 80.7 percent versus 67.2 percent. The good results of MCF-RFE again prove the strengths of multiple-criteria fusion.



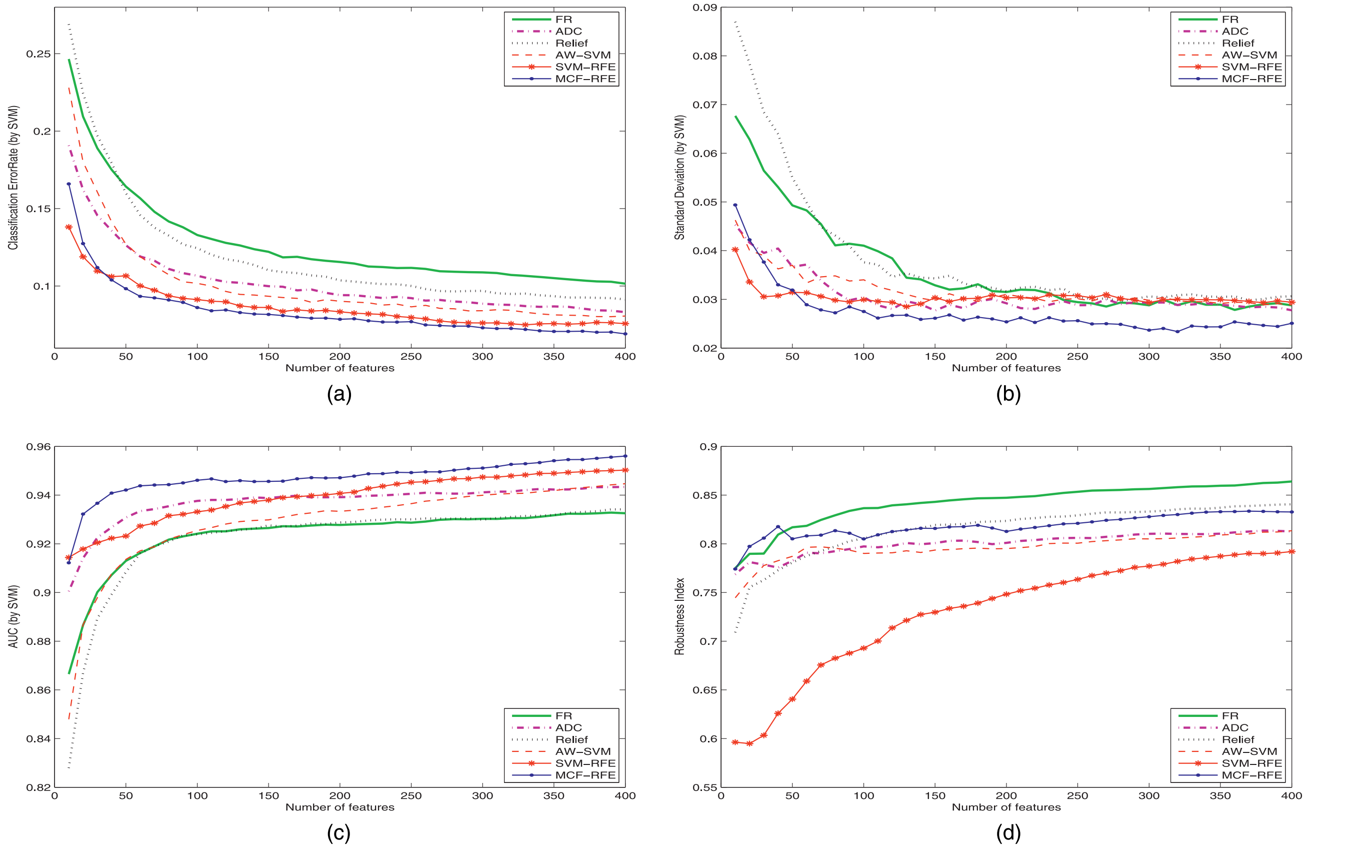
(a) (b) (c) (d)

Fig. 8. Performance comparisons on colon data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.



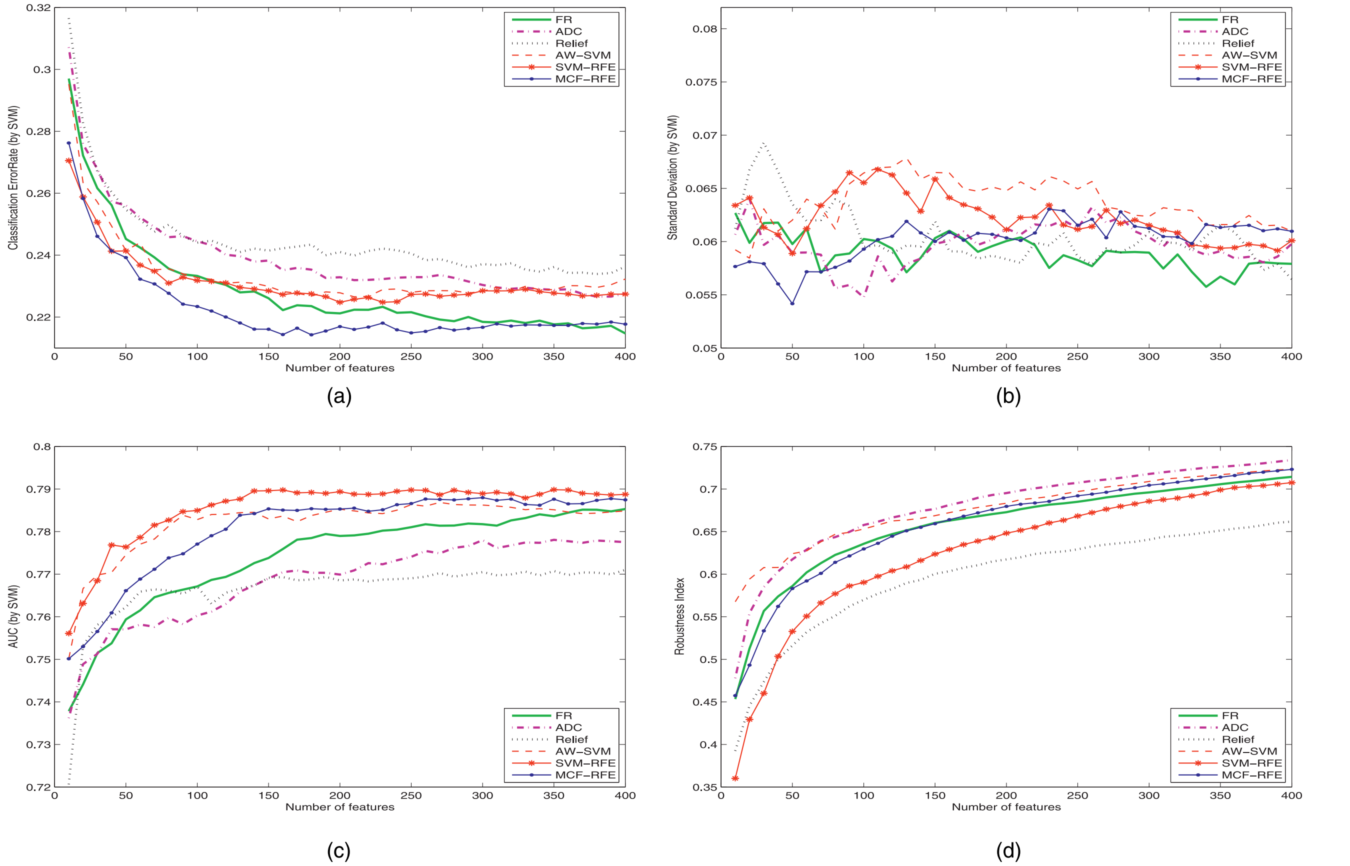
(a) (b) (c) (d)

Fig. 9. Performance comparisons on leukemia data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.



(a) (b) (c) (d)

Fig. 10. Performance comparisons on prostate data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.



(a) (b) (c) (d)

Fig. 11. Performance comparisons on CNS data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.

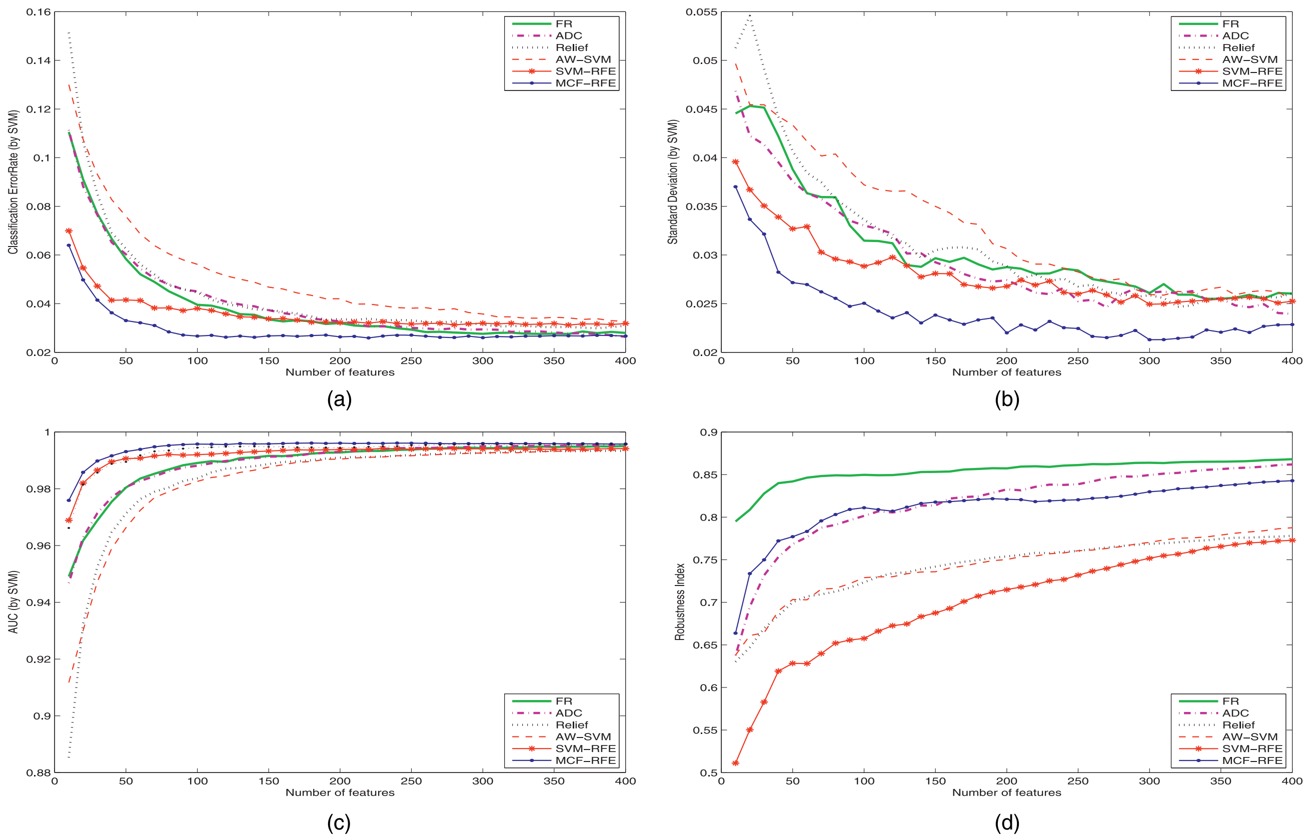


Fig. 12. Performance comparisons on DLBCL data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.

5.4 Comparative Study of MCF-RFE with Bagging-Based Ensemble Technique (BBET)

This part conducts a comparative study between MCF-RFE and the bagging-based ensemble feature selection technique (BBET) proposed by Saeys et al. [34]. BBET is based on instance perturbation and can be applied to any of ranking-based feature selection algorithms. In this study, the procedure of BBET is as follows: first, a number of bags (i.e., subsamples) are generated from the training samples using resampling with replacement technique and a feature ranking algorithm is performed on each of the bags to produce separate feature rankings; then all the feature rankings are combined to form a final feature ranking using an aggregation method. In the experiment, BBET was applied to each of the four basis criteria with 40 bags (which is the same as in [34]) and *Borda count* was used as the linear aggregation method. Due to page limitation, only the experimental results on Colon, Prostate, and CNS data sets are presented in Figs. 13, 14, and 15, and the results on Leukemia and DLBCL data sets can be found on the same website as that for KNN classification results.

We first compare the feature stability performances because BBET is originally designed to improve robustness of feature selection algorithms. From the stability results in Figs. 13d, 14d, and 15d, and on website, it can be observed that MCF-RFE performs better than or comparable to the best basis criterion with BBET. For the classification performance including estimated classification error, standard deviation of error estimation and AUC, MCF-RFE performs better with exceptions of AUC and standard deviation on CNS data.

After comparing carefully, the results of the four basis criteria with BBET with those without BBET, we find that BBET may not always be beneficial, which is different from the conclusion in [34] that BBET generally provides more robust results. For example, in most of the cases, *AW-SVM* benefits from BBET while *Fisher's ratio* does not. We think one possible reason is the small sample size of gene-expression data. In addition, due to the resampling with replacement used in *.632* bootstrap in our experiments, only about $63.2\% \times 63.2\% = 39.9\%$ of the original samples were retained for construction of each feature selector in the ensemble, while in [34] $90\% \times 63.2\% = 56.9\%$ of the original samples were used to construct each feature selector.

6 Conclusion

In this paper, we have analyzed and discussed multicriterion fusion for feature selection on high-dimensional and small-sized data. Motivated by the strengths of fusion of multiple criteria and the recursive feature elimination (RFE) search strategy, we have proposed a feature subset selection algorithm-MCF-RFE. Extensive experiment study of MCF-RFE with Fisher's ratio, Relief, ADC (asymmetric dependency coefficient), AW-SVM (absolute weight of SVM) and the commonly used benchmark algorithm SVM-RFE based on three performance indices including classification error, standard deviation of error estimation and feature stability has been conducted, and the results show that MCF-RFE outperforms in classification performance with reasonably good stability. An comparative study

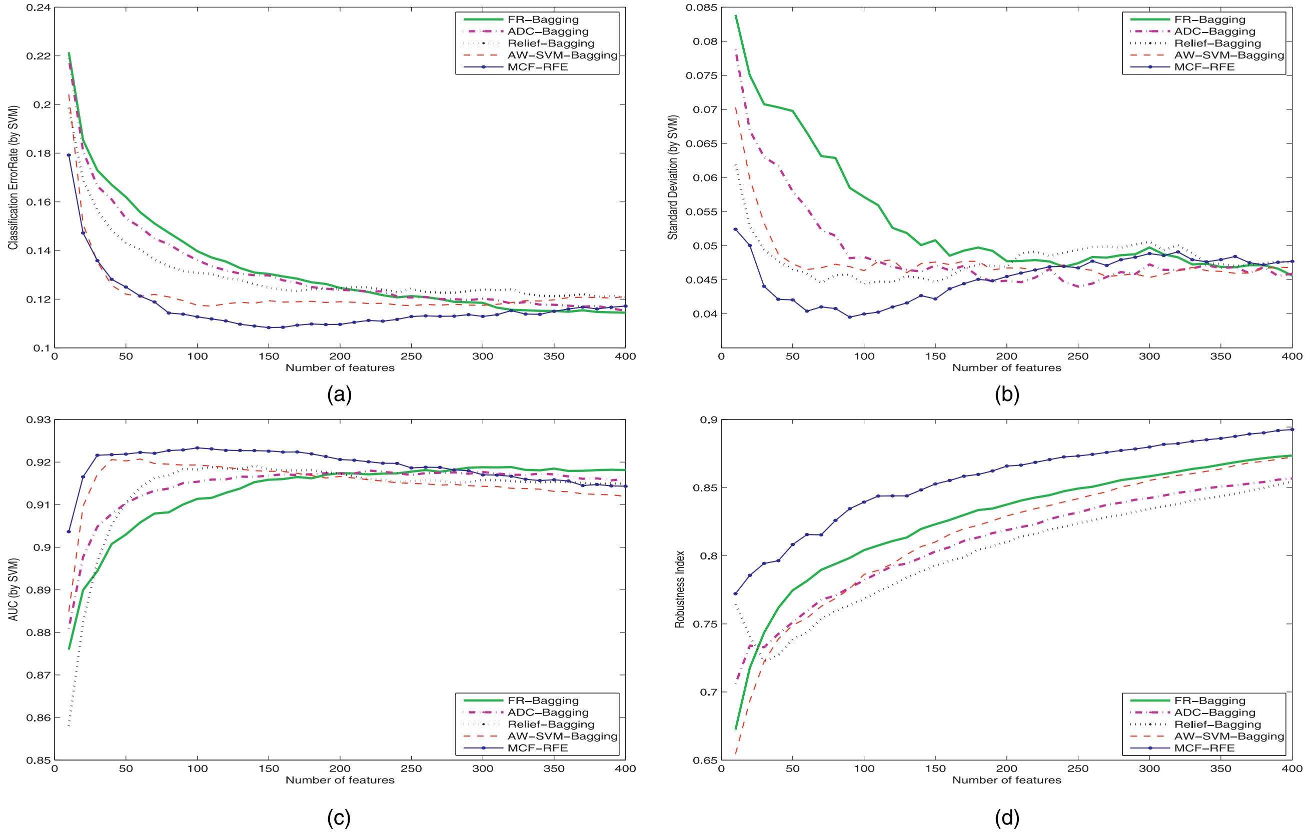


Fig. 13. MCF-RFE and bagging-based ensemble comparisons on colon data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.

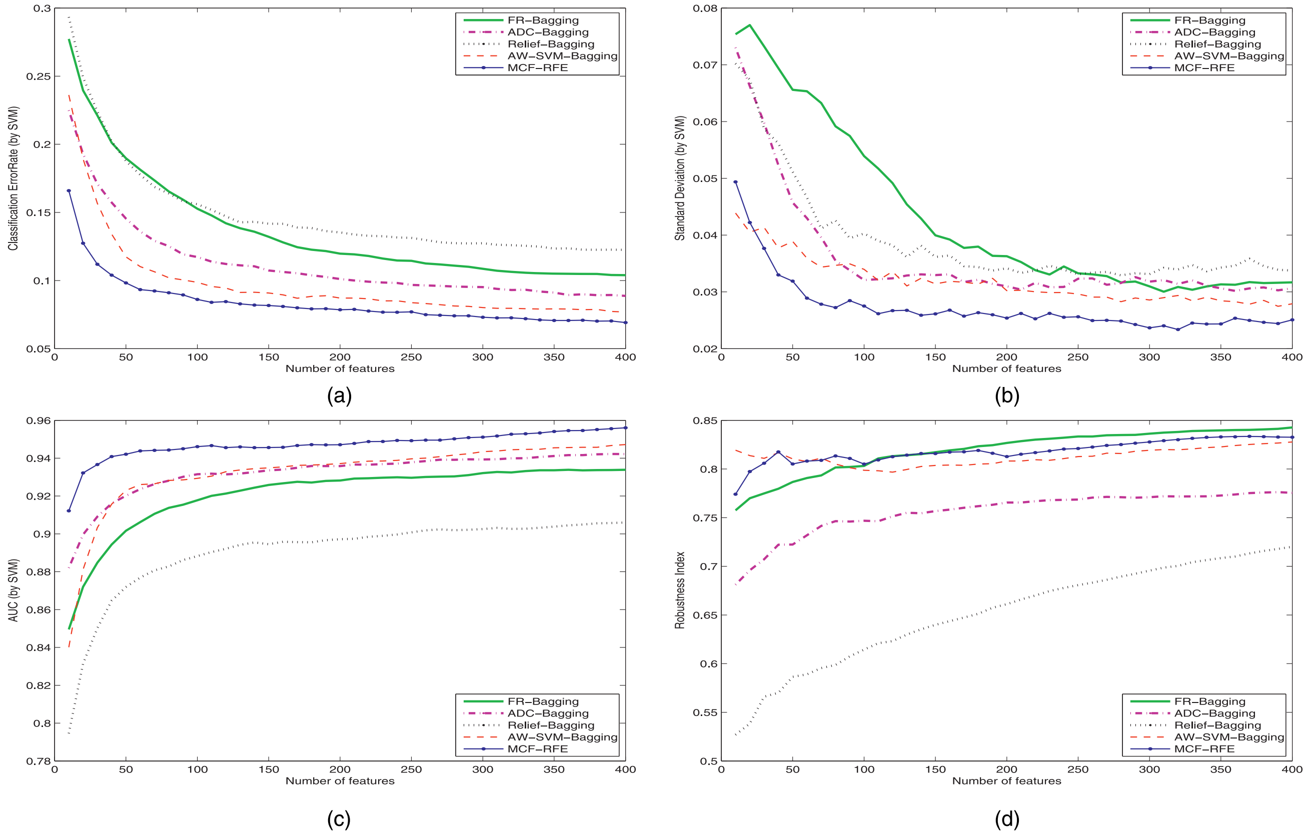
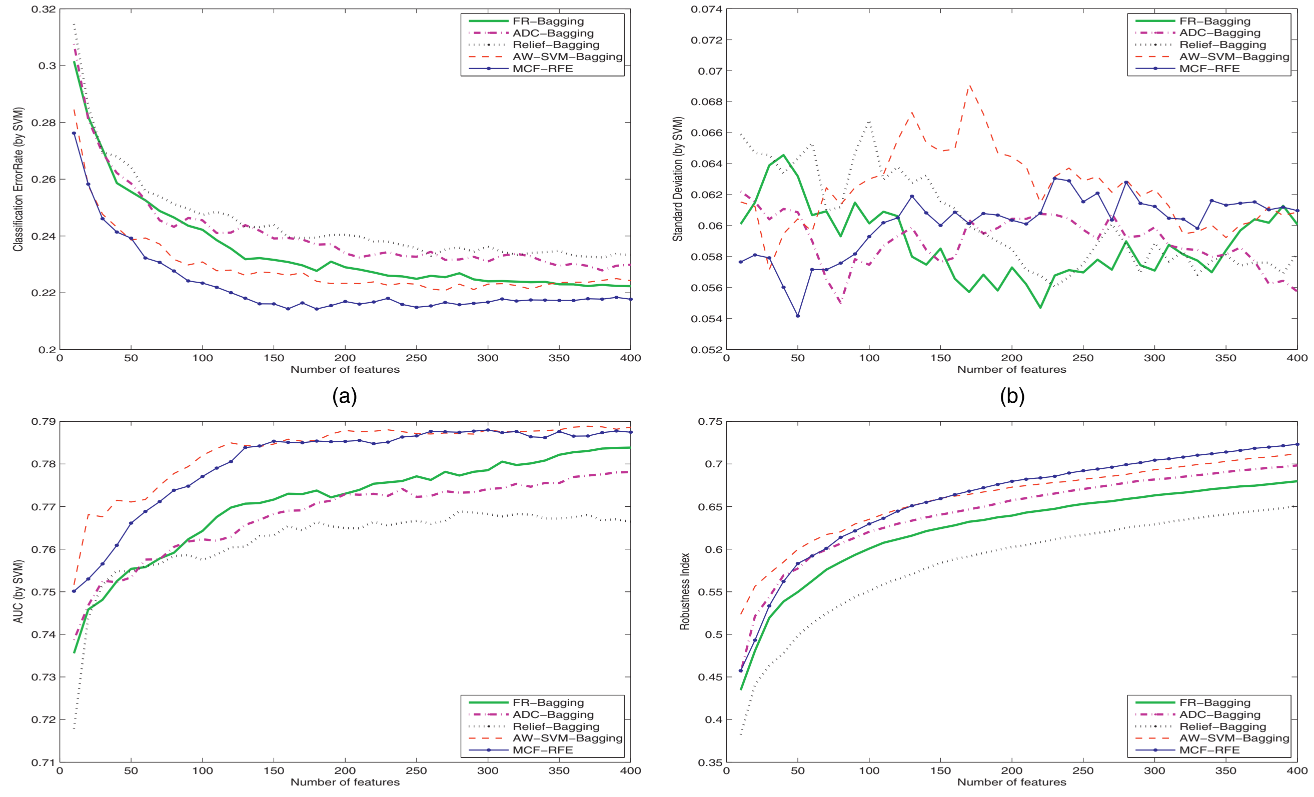


Fig. 14. MCF-RFE and bagging-based ensemble comparisons on prostate data. (a) Classification error. (b) Standard deviation of error estimation.



(a) (b)

(c) (d)

Fig. 15. MCF-RFE and bagging-based ensemble comparisons on CNS data. (a) Classification error. (b) Standard deviation of error estimation. (c) AUC. (d) Feature stability.

between our proposed algorithm and the bagging-based ensemble technique (BBET) has also proved the strengths of multicriterion fusion.

Acknowledgments

The authors would like to thank the reviewers for their detailed and valuable comments and suggestions which have helped to improve the manuscript in many ways.

References

[1] F.K. Ahmad, N.M. Norwawi, S. Deris, and N.H. Othman, "A Review of Feature Selection Techniques via Gene Expression Profiles," *Proc. Int'l Symp. Information Technology (ITSim '08)*, pp. 1-7, 2008.

[2] U. Alon, N. Barkai, D.A. Notterman, K. Gishdagger, S. Ybarradagger, D. Mackdagger, 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'l Academy of Sciences USA*, vol. 96, no. 12, pp. 6745-6750, June 1999.

[3] G. Bontempi, "A Blocking Strategy to Improve Gene Selection for Classification of Gene Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 4, no. 2, pp. 293-300, Apr. 2007.

[4] A.P. Bradley, "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[5] U. Braga-Neto and E.R. Dougherty, "Is Cross-Validation Valid for Small-Sample Microarray Classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374-380, 2004.

[6] C.C. Chang and C.J. Lin, "LIBSVM : A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[7] S.M. Chao and E.R. Dougherty, "The Peaking Phenomenon in the Presence of Feature-Selection," *Pattern Recoginition Letters*, vol. 29, no. 11, pp. 1667-1674, 2008.

[8] M.R. Chernick, *Bootstrap Methods: A Guide for Practitioners and Researchers*, second ed., John Wiley & Sons, 2007.

[9] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, Sept. 1995.

[10] T.G. Dietterich, "Machine Learning Research: Four Current Directions," *Artificial Intelligence Magazine*, vol. 18, no. 4, pp. 97-136, 1997.

[11] T.G. Dietterich, "Ensemble Methods in Machine Learning," *Proc. First Int'l Workshop Multiple Classifier Systems (MCS '00)*, vol. 1857, pp. 1-15, 2000.

[12] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank Aggregation Revisited," *Proc. Int'l World Wide Web Conf.*, May 2001.

[13] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics*, vol. 16, no. 10, pp. 906-914, Oct. 2000.

[14] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, "Semisupervised Learning for Molecular Profiling," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 110-118, Oct. 2005.

[15] T. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, Oct. 1999.

[16] I. Guyon, http://www.clopinet.com/isabelle/Projects/NIPS2001/, 2009.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, nos. 1-3, pp. 389-422, Jan. 2002.

[18] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research*, vol. 3, pp. 1157-1182, Mar. 2003.

[19] G. Gulgezen, Z. Cataltepe, and L. Yu, "Stable and Accurate Feature Selection," *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases: Part I (ECML PKDD '09)*, vol. 5781, pp. 455-468, 2009.

[20] D.F. Hsu and T. Isak, "Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval," *Information Retrieval*, vol. 8, no. 3, pp. 449-480, Jan. 2005.

[21] J. Hua, Z.X. Xiong, J. Lowey, E. Suh, and E.R. Dougherty, "Optimal Number of Features as a Function of Sample Size for Various Classification Rules," *Bioinformatics,* vol. 21, no. 8, pp. 1509-1515, 2005.

[22] D. Huang and T.W.S. Chow, "Effective Gene Selection Method with Small Sample Sets Using Gradient-Based and Point Injection Techniques," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 3, pp. 467-475, July-Sept. 2007.

[23] A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces," *Knowledge and Information Systems,* vol. 12, no. 1, pp. 95-116, May 2007.

[24] E. Kim and J. Ko, "Dynamic Classifier Integration Method," *Proc. Int'l Workshop Multiple Classifier Systems (MCS '05),* pp. 97-107, 2005.

[25] A. Klementiev, D. Roth, and K. Small, "An Unsupervised Learning Algorithm for Rank Aggregation," *Proc. European Conf. Machine Learning (ECML '07),* pp. 616-623, 2007.

[26] R. Kohavi and G.H. John, "Wrapper for Feature Subset Selection," *Artificial Intelligence,* vol. 97, nos. 1-2, pp. 273-324, Dec. 1997.

[27] P. Krizek, J. Kittler, and V. Hlavac, "Improving Stability of Feature Selection Methods," *Proc. 12th Int'l Conf. Computer Analysis of Images and Patterns (CAIP'07)* , vol. 4673, pp. 929-936, 2007.

[28] Y. Leung and Y. Hung, "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 7, no. 1, pp. 108-117, Jan.-Mar. 2010.

[29] Y.T. Liu, T.Y. Liu, T. Qin, Z.M. Ma, and H. Li, "Supervised Rank Aggregation," *Proc. Int'l Conf. World Wide Web,* pp. 481-490, 2007.

[30] S. Loscalzo, L. Yu, and C. Ding, "Consensus Group Stable Feature Selection," *Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD '09),* pp. 567-576, 2009.

[31] S.L. Pomeroy et al., "Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression," *Nature,* vol. 415, pp. 265-271, 2002.

[32] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning,* vol. 53, pp. 23-69, 2003.

[33] Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007.

[34] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," *Proc. European Conf. Machine Learning and Knowledge Discovery,* pp. 313-25, 2008.

[35] M.A. Shipp et al., "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning," *Nature Medicine,* vol. 8, no. 1, pp. 68-74, Jan. 2002.

[36] D.V. Shridhar, E.B. Bartlett, and R.C. Seagrave, "Information Theoretic Subset Selection for Neural Network Models," *Computers and Chemical Eng.,* vol. 22, nos. 4-5, pp. 613-626, 1998.

[37] D. Singh et al., "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell,* vol. 1, no. 2, pp. 203-209, 2002.

[38] P. Somol and J. Novovicova, "Evaluating the Stability of Feature Selectors that Optimize Feature Subset Cardinality," *Proc. Int'l Workshop Structural, Syntactic, and Statistical Pattern Recognition,* pp. 956-966, 2008.

[39] Y.C. Tang, Y.Q. Zhang, and Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 3, pp. 365-381, July-Sept. 2007.

[40] K. Tumer and J. Ghosh, *Theoretical Foundations of Linear and Order Statistics Combiners for Neural Pattern Classifiers,* Technical Report 95-02-98, The Computer and Vision Research Center, Univ. of Texas, 1998.

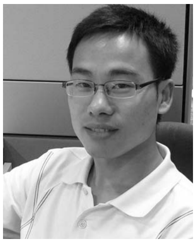
[41] M. van Erp and L. Schomaker, "Variants of the Borda Count Method for Combining Ranked Classifier Hypotheses," *Proc. Int'l Workshop Frontiers in Handwriting Recognition,* pp. 443-452, 2000.

[42] V. Vapnik, *Statistical Learning Theory.* John Wiley & Sons, 1998.

[43] Wikipedia. http://en.wikipedia.org/wiki/Borda_count, 2009.

[44] L. Yu, C. Ding, and S. Loscalzo, "Stable Feature Selection via Dense Feature Groups," *Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD '08),* pp. 803-811, 2008.

[45] X. Zhou and K.Z. Mao, "The Ties Problem Resulting from Counting-Based Error Estimators and Its Impact on Gene Selection Algorithms," *Bioinformatics,* vol. 22, no. 20, pp. 2507-2515, 2006.



Feng Yang received both the bachelor's and master's degrees from Xi'an Jiaotong University, China. He is currently pursuing the PhD degree at the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. He is currently working on robustness of gene selection for microarray gene-expression data. His research interests include dimension reduction, feature selection of very high-dimensional and small sample size data, and classification of gene-expression data.

K.Z. Mao received the BEng, MEng, and PhD degrees from Jinan University, Northeastern University, and Sheffield University, in 1989, 1992, and 1998, respectively. He was a research associate at the University of Sheffield from April 1998 to September 1998, a research fellow at the Centre for Signal Processing (now part of I2R of A-Star) from September 1998 to May 2001. He joined the School of Electrical and Electronic Engineering, Nanyang Technological University as an assistant professor in June 2001, where, currently, he is an associate professor. His research interests include machine learning, computational intelligence, biomedical image analysis, and bioinformatics.