Stable Gene Selection from Microarray Data via Sample Weighting

Lei Yu, Yue Han, and Michael E. Berens

Abstract—Feature selection from gene expression microarray data is a widely used technique for selecting candidate genes in various cancer studies. Besides predictive ability of the selected genes, an important aspect in evaluating a selection method is the stability of the selected genes. Experts instinctively have high confidence in the result of a selection method that selects similar sets of genes under some variations to the samples. However, a common problem of existing feature selection methods for gene expression data is that the selected genes by the same method often vary significantly with sample variations. In this work, we propose a general framework of sample weighting to improve the stability of feature selection methods under sample variations. The framework first weights each sample in a given training set according to its influence to the estimation of feature relevance, and then provides the weighted training set to a feature selection method. We also develop an efficient margin-based sample weighting algorithm under this framework. Experiments on a set of microarray data sets show that the proposed algorithm significantly improves the stability of representative feature selection algorithms such as SVM-RFE and ReliefF, without sacrificing their classification performance. Moreover, the proposed algorithm also leads to more stable gene signatures than the state-of-the-art ensemble method, particularly for small signature sizes.

Index Terms—Feature selection, gene selection, stability, classification, gene expression microarray.

1 Introduction

The identification and validation of molecular biomarkers for cancer diagnosis, prognosis, and therapeutic targets is an important problem in cancer genomics. Due to the time-consuming, costly, and labor-intensive nature of clinical and biological validation experiments, it is crucial to select a list of high-potential biomarker candidates for validation [27]. Gene expression microarray data [13] are widely used for identifying candidate genes in various cancer studies. From a machine learning viewpoint, the selection of candidate genes in this context can be regarded as a problem of feature selection from high-dimensional labeled data, where the aim is to find a small subset of features (genes) that best explain the difference between samples of distinct phenotypes.

Many feature selection methods have been adopted for gene selection from microarray data, and have shown good classification performance of the selected genes [23], [31], [36], [39]. However, a common problem with existing gene selection methods is that the selected genes by the same method often vary significantly with some variations of the samples in the same data set [7], [10], [20]. To make the matters worse, different methods or different parameter settings of the same method may also result in largely different subsets of genes for the same set of samples. The instability of the resulting gene signatures raises serious doubts about the reliability of the selected genes as biomarker candidates and hinders biologists from deciding candidates for subsequent validations.

The stability issue of feature selection has recently become a topic of strong interest in both the machine learning and the bioinformatics communities. Several studies have developed stability measures and assessed the stability of existing feature selection methods [3], [7], [19], [20], [21]. Others have employed ensemble (Ens.) techniques to improve the stability of feature selection results, including: Bayesian model averaging [22], [37], aggregating the results of a collection of feature ranking methods [9], [35], and aggregating the results of the same feature selection method from bootstrapped subsets of samples [1], [7], [8].

The stability of feature selection is a complicated issue, inviting an abundance of approaches to improve the stability of feature selection results. Several major factors affect the stability of feature selection results: the mechanisms of feature selection methods, the underlying data distribution, and the sample size [11], [25]. It is important to note that stability of feature selection methods should not be investigated alone, but always together with the predictive performance of the selected genes. Biologists will not be interested in a strategy (e.g., arbitrarily selecting the same set of genes regardless of the input samples) that yields very stable gene signatures but bad predictive models.

In this work, we focus on the stability of feature selection methods under sample variations. We propose a sample weighting (SW) framework to improve the stability of feature selection methods. The framework is motivated by importance sampling, one of the commonly used variance reduction techniques [30]. The main idea of this framework

- *L. Yu and Y. Han are with the Department of Computer Science, State University of New York, Binghamton, NY 13902-6000. E-mail: {lyu, yhan1}@binghamton.edu.*
- *M.E. Berens is with the Cancer and Cell Biology Division, Translational Genomics Research Institute, 445 N. Fifth Street, Phoenix, AZ 85004. E-mail: mberens@tgen.org.*

Manuscript received 17 May 2010; revised 25 Oct. 2010; accepted 25 Dec. 2010; published online 7 Mar. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2010-05-0124. Digital Object Identifier no. 10.1109/TCBB.2011.44.

is to first weight each sample in a training set according to its influence to the estimation of feature relevance, and then provide the weighted training set to a feature selection method. Intuitively, different samples in a training set could have different influence on the feature selection result according to their views (or local profiles) of the relevance of each feature. If a sample shows a noticeably distinct local profile from the other samples, its absence or presence in the training data will substantially affect the feature selection result. In order to improve stability, samples with outlying local profiles need to be weighted differently from the rest of the samples. To this end, we propose a margin-based sample weighting algorithm which assigns a weight to each sample according to the outlying degree of its local profile of feature relevance compared with other samples. The local profile of feature relevance at a given sample is measured based on the hypothesis margin of the sample.

Our experiments on a set of public microarray data sets show that the proposed sample weighting framework significantly improves the stability of SVM-RFE [15] and ReliefF [29] while maintaining their classification performance. The sample weighting framework is also compared with the bagging ensemble framework [1] based on SVM-RFE and ReliefF. Experimental results show that the improvement to the stability of both algorithms by sample weighting is generally more significant than the improvement by bagging ensemble, particularly for small signature sizes (a few tens of genes).

The rest of the paper is organized as follows: Section 2 proposes a margin-based sample weighting algorithm under the general framework of sample weighting. Section 3 describes experimental setup. Section 4 presents and discusses experimental results. Section 5 provides concluding remarks and points out some future research directions.

2 Margin-Based Sample Weighting

In a recent study [16], Han and Yu proposed a theoretical framework about stable feature selection which defines the stability of feature selection from a sample variance perspective and shows that the stability of feature selection under training data variations can be improved by variance reduction techniques. The sample weighting framework proposed in this study is motivated by importance sampling, one of the commonly used variance reduction techniques [30]. The theory of importance sampling suggests that in order to reduce the variance of a Monte Carlo estimator (e.g., the estimate of feature relevance by a feature weighting algorithm based on a training set), instead of performing i.i.d. sampling, we should increase the number of samples taken from regions which contribute more to the quantity of interest and decrease the number of samples taken from other regions. When given only the empirical distribution in a training set, although we cannot redo the sampling process, we can simulate the effect of importance sampling by increasing the weights of samples taken from more important regions and decreasing the weights of those from other regions. Therefore, the problem of variance reduction for feature selection boils down to finding an empirical solution to estimating the importance of samples with respect to feature evaluation and weighting samples accordingly. Section 2.1 provides some preliminaries. Section 2.2 presents the main ideas of the proposed sample weighting framework. Section 2.3 provides the technical details of the margin-based sample weighting algorithm developed under this framework.

2.1 Preliminaries

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denote a training set of n labeled samples, where $\mathbf{x}_i$ is a sample vector in the feature space $\Re^d$ defined by d features $X_1, \ldots, X_d$, and y is the value of the class variable Y. For gene selection, a gene expression microarray data set, consisting of the expression levels of d genes across n samples labeled by experimental conditions, can be represented as a training set for feature selection, with each gene represented by a feature.

Margins play an important role in modern machine learning research, and have been used both for theoretical generalization bounds and as guidelines for algorithm design. They measure the confidence of a classifier with respect to its decisions [5]. As described in [6], there are two natural ways of defining the margin of a sample with respect to a classifier. *Sample margin* measures the distance between a sample and the decision boundary of a classifier. Support Vector Machine (SVM) [5], for example, uses this type of margin; it finds the separating hyperplane with the largest sample margin for support vectors. An alternative definition, *hypothesis margin*, measures the distance between the hypothesis of a sample and the closest hypothesis that assigns alternative label to the sample. Hypothesis margin requires a distance measure between hypotheses (classifiers). For example, AdaBoost [12] uses this type of margin with the L_1-norm as the distance measure between hypotheses. Feature selection methods developed under the large margin principles such as SVM-RFE [15] and ReliefF [29] evaluate the relevance of features according to their respective contributions to the margins.

For 1-Nearest Neighbor (1NN) classifier, authors of [6] proved that 1) the hypothesis margin lower bounds the sample margin; and 2) the hypothesis margin of a sample $\mathbf{x}$ with respect to a training set D can be computed by the following formula:

$$\theta_D(\mathbf{x}) = \frac{1}{2}\left(\|\mathbf{x} - \mathbf{x}^M\| - \|\mathbf{x} - \mathbf{x}^H\|\right),$$

where $\mathbf{x}^H$ and $\mathbf{x}^M$ represent the nearest samples (called Hit and Miss) to $\mathbf{x}$ in D with the same and opposite class labels, respectively. Since hypothesis margin is easy to compute and large hypothesis margin ensures large sample margin, we focus on hypothesis margin in this paper.

2.2 Margin Vector Feature Space

In our framework of sample weighting for stable gene selection, we employ the concept of hypothesis margin in a novel way. By decomposing the margin of a sample along each dimension, the sample in the original feature space can be represented by a new vector (called *margin vector*) in the *margin vector feature space* defined as follows:

Definition 1. *Let $\mathbf{x} = (x_1, \ldots, x_d)$ be a sample in the original feature space $\Re^d$, and $\mathbf{x}^H$ and $\mathbf{x}^M$ represent the nearest samples to $\mathbf{x}$ with the same and opposite class labels, respectively. For each $\mathbf{x} \in \Re^d$, $\mathbf{x}$ can be mapped to $\mathbf{x}' = (x'_1, \ldots, x'_d)$ in a new feature space $\Re'^d$ according to*

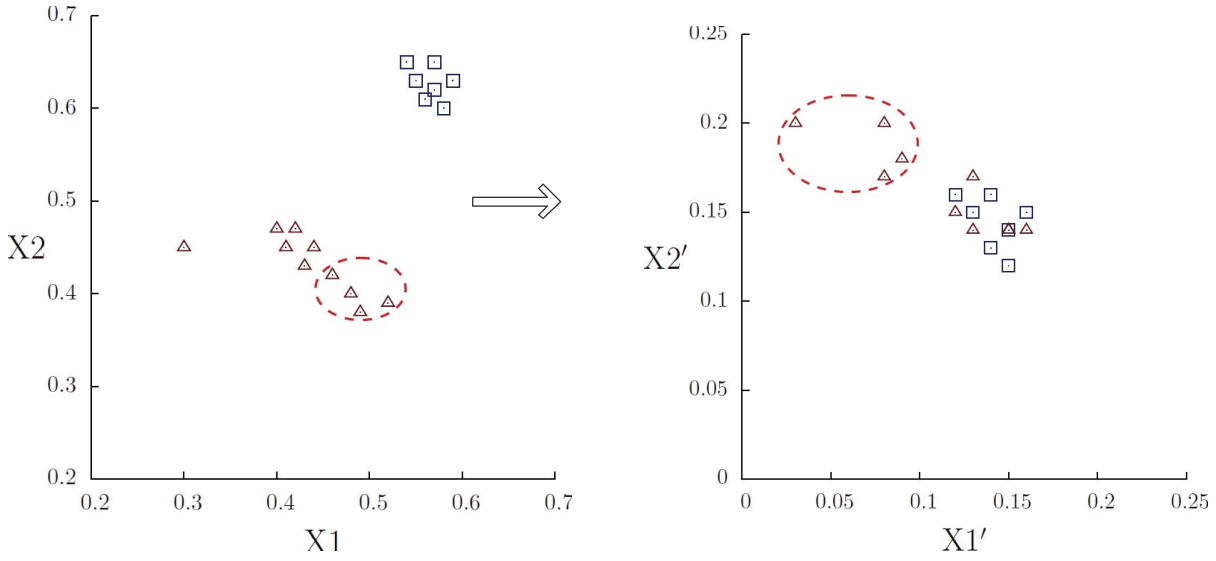


Fig. 1. An illustrative example for Margin Vector Feature Space. Each data point in the original feature space (left) is projected to the margin vector feature space (right) according to its hypothesis margin in the original feature space. The class labels of data points are distinguished by triangles and squares.

$$x'_j = |x_j - x_j^M| - |x_j - x_j^H|, \quad (1)$$

where x'_j is the jth coordinate of $\mathbf{x}'$ in the new feature space $\Re'^d$, and x_j, x_j^M, or x_j^H is the jth coordinate of $\mathbf{x}$, $\mathbf{x}^H$, or $\mathbf{x}^M$ in the original feature space $\Re^d$, respectively. Vector $\mathbf{x}'$ is called the margin vector of $\mathbf{x}$, and $\Re'^d$ is called the margin vector feature space.

In essence, $\mathbf{x}'$ captures the local profile of feature relevance for all features at $\mathbf{x}$. The larger the value of x'_j, the more feature X_j contributes to the margin of sample $\mathbf{x}$. Thus, the margin vector feature space captures local feature relevance profiles (margin vectors) for all samples in the original feature space.

Fig. 1 illustrates the idea of margin vector feature space through a 2D example. Each labeled data point (triangle or square) is a sample with two features. Each sample in the original feature space (left) is projected into the margin vector feature space (right) according to (1). We can clearly see that samples labeled with triangles exhibit largely different outlying degrees in the two feature spaces. Specifically, those in the dashed ovals are evenly distributed within the proximity to the rest of the triangles (except the outlier on the leftmost) in the original feature space, but are clearly separated from the majority of the samples in the margin vector feature space. The outlier triangle in the original space becomes part of the majority group in the margin vector feature space. To decide the overall relevance of features $X1$ versus $X2$, one intuitive idea is to take the average over all margin vectors, as adopted by the well-known ReliefF algorithm [29]. However, since the samples in the dashed oval exhibit largely distinct margin vectors from the rest of the samples, the presence or absence of these samples in the training set will affect the global decision on which feature is more relevant.

From the illustrative example, we can see that the margin vector feature space captures the distance among samples with respect to their margin vectors (instead of feature values in the original space), and enables the detection of samples that largely deviate from others in this respect. By identifying and reducing the emphasis on these outlying samples, more stable results can be produced from a feature selection method. In the next section, we will further discuss how to exploit such discrepancy to weight samples in order to alleviate the affect of training data variations on feature selection results.

2.3 Algorithm

The previous definition on margin vector feature space only considers one nearest neighbor from each class. To reduce the affect of noise or outliers in the training set on the transformed feature space, multiple nearest neighbors from each class can be used to compute the margin vector of a sample. In this work, we consider all neighbors from each class for a given sample. Equation (1) can then be extended to

$$x'_j = \sum_{l=1}^{m} |x_j - x_j^{M_l}| - \sum_{l=1}^{h} |x_j - x_j^{H_l}|, \quad (2)$$

where $x_j^{H_l}$ or $x_j^{M_l}$ denotes the jth component of the lth neighbor to $\mathbf{x}$ with the same or opposite class label, respectively. m or h represents the total number of Misses or Hits ($m + h$ equals the total number of samples in the training set excluding $\mathbf{x}$).

Once the margin vector feature space is generated, the next task is to exploit the discrepancy of margin vectors in this space to weight samples in the original space. To quantitatively evaluate the outlying degree of each margin vector $\mathbf{x}'$, we measure the average distance of $\mathbf{x}'$ to all other margin vectors; greater average distance indicates higher outlying degree. As illustrated in Fig. 1, the global decision of feature relevance is more sensitive to samples that largely deviate from the rest of the samples in the margin vector feature space than to samples that have low outlying degrees. To improve the stability of a feature selection method under training data variations, we assign lower weights to samples with higher outlying degrees. This decision is consistent with the intuition behind importance

TABLE 1
Summary of Microarray Data Sets

Name	# Features	# Samples	Source
Colon	2000	62	[2]
Leukemia	7129	72	[13]
Prostate	6034	102	[32]
Lung	12533	181	[14]

sampling introduced earlier. Specifically, the weight for a sample $\mathbf{x}$ in the original feature space is given by

$$W(\mathbf{x}) = \frac{1/\overline{dist}(\mathbf{x}')}{\sum_{i=1}^{n} 1/\overline{dist}(\mathbf{x}'_i)}, \quad (3)$$

where

$$\overline{dist}(\mathbf{x}') = \frac{1}{n-1} \sum_{i=1, \mathbf{x}'_i \neq \mathbf{x}'}^{n-1} dist(\mathbf{x}', \mathbf{x}'_i).$$

Algorithm 1 outlines the key steps of the margin-based sample weighting algorithm. Both feature space transformation and sample weighting involve distance computation along all features for all pairs of samples: the former in the original feature space, and the latter in the margin vector feature space. Since these computations dominate the time complexity of the algorithm, the overall time complexity of the algorithm is $O(n^2 * d)$, where n is the sample size and d is the number of features (genes). Therefore, the algorithm is very efficient for microarray data with small sample size (i.e., $n \ll d$).

Algorithm 1. Margin Based Sample Weighting
Input: data $D = \{\mathbf{x}_i\}_{i=1}^{n}$
Output: weight vector $\mathbf{w}$ for all samples in D
// Feature Space Transformation
for $i = 1$ **to** n **do**
  for $j = 1$ **to** d **do**
    For $\mathbf{x}_i$, compute $x'_{i,j}$ according to Eq. (2)
  end for
end for
// Sample Weighting
Calculate and store pair-wise distances among all margin vectors $\mathbf{x}'_i$
for $i = 1$ **to** n **do**
  For $\mathbf{x}_i$, compute its weight according to Eq. (3)
end for

3 Experimental Setup

Before we present experimental results in the next section, we describe microarray data sets used, methods in comparison, evaluation measures, and experimental procedures.

3.1 Microarray Data

We experimented with four frequently studied public gene expression microarray data sets summarized in Table 1. The Colon cancer data set [2] has been frequently used in previous studies in gene selection and classification. It consists of the gene expression profiles of 2,000 genes for 62 tissue samples among which 40 are colon cancer tissues and 22 are normal tissues. The Leukemia data set [13] is another widely used benchmark data set. It consists of gene expression profiles of two classes of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). The data set consists of 7,129 genes and 72 samples (47 ALL and 25 AML). The Prostate data set [32] consists of gene expression profiles of 6,034 genes for 52 prostate tumor samples and 50 normal samples. The Lung cancer data set [14] consists of gene expression profiles of 12,533 genes for 181 lung tissue samples among which 31 are of malignant pleural mesothelioma (MPM) and 150 are of adenocarcinoma (ADCA).

3.2 Methods in Comparison

We choose SVM-RFE and ReliefF as the baseline algorithms for experimental study. We evaluate the effectiveness of the proposed sample weighting framework for these two algorithms. Furthermore, we compare the sample weighting framework with a recently proposed ensemble framework using SVM-RFE and ReliefF as the base algorithms.

3.2.1 Baseline Algorithms: SVM-RFE and ReliefF

Although much simpler feature selection methods are available [24], SVM-RFE is chosen as a baseline because it is known to provide state-of-the-art classification performance and widely used in microarray data. Recently, the original algorithm has been improved by several studies such as bootstrapped SVM-RFE [8], two-stage SVM-RFE [33], and SVM-RFE combined with MRMR filter [26]. SVM-RFE is intrinsically a multivariate feature selection method in the sense that it considers feature interaction while evaluating the relevance of features.

The main process of SVM-RFE is to recursively eliminate features of low weights, using SVM to determine feature weights. Starting from the full set of features, at each iteration, the algorithm trains a linear SVM classifier based on the remaining set of features, ranks features according to the squared values of feature weights in the optimal hyperplane, and eliminates one or more features with the lowest weights. This recursive feature elimination (RFE) process stops until all features have been removed or a desired number of features is reached. Our implementation of SVM-RFE is based on Weka's [34] implementation of soft-margin SVM using linear kernel and default C parameter. As suggested by the authors of SVM-RFE, 10 percent of the remaining features are eliminated at each iteration to speed up the RFE process.

ReliefF [29] is chosen as another representative algorithm for margin-based feature selection. It is a simple and efficient feature weighting algorithm which considers all features together in evaluating the relevance of features. The main idea of ReliefF is to weight features according to how well their values distinguish between samples that are similar to each other. Specifically, for a two-class problem, the weight for each feature X_j is determined as follows:

$$W(X_j) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{l=1}^{K} \left(\left| x_{i,j} - x_{i,j}^{M_l} \right| - \left| x_{i,j} - x_{i,j}^{H_l} \right| \right), \quad (4)$$

where $x_{i,j}$, $x_{i,j}^{M_l}$, or $x_{i,j}^{H_l}$ denotes the jth component of sample $\mathbf{x}_i$, its lth closest Miss $\mathbf{x}_i^{M_l}$, or its lth closest Hit $\mathbf{x}_i^{H_l}$,

respectively. n is the total number of samples, and K is the number of Hits or Misses considered for each sample. We used Weka's implementation of ReliefF with the default setting $K = 10$.

ReliefF appears similar to our proposed sample weighting algorithm since both algorithms involve distance calculation between a sample and its Hits or Misses along each feature for all samples. However, the two algorithms are intrinsically different. ReliefF is a feature weighting algorithm; it produces *feature* weights according to (4) which does not explicitly construct the margin vector for each sample but takes an average of the margins over all samples. Our sample weighting algorithm produces *sample* weights by explicitly projecting each sample to its margin vector in the margin vector feature space (based on (2)) and a successive sample weighting procedure in the margin vector feature space. Our sample weighting algorithm can be used as a preprocessing step for any feature selection algorithms which can be extended to incorporate sample weights.

3.2.2 Sample Weighting SVM-RFE and Sample Weighting ReliefF

Given a training set, SVM-RFE and ReliefF select features based on the original training set where every sample is equally weighted. They can be extended to work on a weighted training set produced by the proposed sample weighting algorithm. We refer to this version of SVM-RFE or ReliefF as sample weighting SVM-RFE or sample weighting ReliefF, respectively. We next explain how sample weights are incorporated into each algorithm.

For SVM-RFE, feature weights are determined based on the final chosen hyperplane of soft-margin SVM which is decided by the trade-off between maximizing the margin and minimizing the training error [5]. With a sample weight $w_i > 0$ assigned to each sample, the original objective function of soft-margin SVM is extended as follows:

$$\min \ \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{n} w_i \xi_i, \tag{5}$$

where the first component is the opposite of the margin, and ξ_i (value of the slack variable) and C in the second component, respectively, capture the error of each sample caused by the hyperplane and the error penalty. For samples with $\xi_i > 0$, increased or decreased sample weight influences the error term (and hence the chosen hyperplane) by amplifying or reducing the effect of ξ_i. When all samples have equal weight, (5) becomes the original objective function of soft-margin SVM.

For ReliefF, feature weights are determined based on the weighting function in (4). With a sample weight $w_i > 0$ assigned to each sample, the original weighting function is extended as follows:

$$w(X_j) = \sum_{i=1}^{n} w_i \sum_{l=1}^{K} \left(w_i^{M_l} \left| x_{i,j} - x_{i,j}^{M_l} \right| - w_i^{H_l} \left| x_{i,j} - x_{i,j}^{H_l} \right| \right), \tag{6}$$

where w_i, $w_i^{M_l}$, or $w_i^{H_l}$ denotes the weight of sample $\mathbf{x}_i$, its lth closest Miss $\mathbf{x}_i^{M_l}$, or its lth closest Hit $\mathbf{x}_i^{H_l}$, respectively. Intuitively, samples with higher weights will have bigger influence on deciding the feature weights, and vice versa. Equation (6) becomes the original weighting function (4) when all samples have equal weight $1/n$, and all Hits and Misses have equal weight $1/K$.

3.2.3 Ensemble SVM-RFE and Ensemble ReliefF

As mentioned in the Introduction, several ensemble methods have been developed to improve the stability of feature selection methods. We choose a most recent bagging ensemble framework [1] to compare with our proposed sample weighting framework. Given a training set, the bagging ensemble framework first generates a number of bootstrapped training sets (by random sampling with replacement), and then repeatedly applies a base feature selection method (e.g., SVM-RFE or ReliefF) on each of the newly created training sets to generate a number of feature rankings. To aggregate the different rankings into a final consensus ranking, the complete linear aggregation scheme used in [1] decides the consensus ranking by summing the ranks of a feature decided based on all bootstrapped training sets. We refer to the ensemble version of SVM-RFE or ReliefF as ensemble SVM-RFE or ensemble ReliefF, respectively. In our implementation, we use 40 bootstrapped training sets to construct each ensemble, as suggested by Abeel et al. [1].

3.3 Evaluation Measures

3.3.1 Stability Measures

Following [1] and [20], we take a similarity-based approach where the stability of a feature selection method is measured by the average over all pairwise similarity comparisons among all feature subsets (gene signatures) obtained by the same method from different subsamplings of a data set. Let $\mathcal{D} = \{D_i\}_{i=1}^{q}$ be a set of subsamplings of a data set of the same size, and $\mathbf{r}_i$ be the feature subset selected by a feature selection method $\mathcal{F}$ on the subsampling D_i. The stability of $\mathcal{F}$ over $\mathcal{D}$ is given by

$$\overline{S}_{\mathcal{D},\mathcal{F}} = \frac{2\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} S(\mathbf{r}_i, \mathbf{r}_j)}{q(q-1)}, \tag{7}$$

where $S(\mathbf{r}_i, \mathbf{r}_j)$ represents a similarity measure between subsets $\mathbf{r}_i$ and $\mathbf{r}_j$.

The stability of a feature selection method depends on the specific choice of the similarity measure $S(\mathbf{r}_i, \mathbf{r}_j)$. Simple measures such as the percentage of overlap or Jaccard index can be applied as in [20]. These measures tend to produce higher values for larger subsets due to the increased bias of selecting overlapping features by chance. To correct this bias, Kuncheva suggested the use of the Kuncheva index [21], defined as follows:

$$S(\mathbf{r}_i, \mathbf{r}_j) = \frac{r - (k^2/d)}{k - (k^2/d)}, \tag{8}$$

where d denotes the total number of features in a data set, $k = |\mathbf{r}_i| = |\mathbf{r}_j|$ denotes the size of the selected subsets, and $r = |\mathbf{r}_i \cap \mathbf{r}_j|$ is the number of common features in both subsets. The Kuncheva index takes values in $[-1, 1]$, with larger value indicating larger number of common features in both subsets. The k^2/d term in the index corrects a bias due to the chance of selecting common features between

two randomly chosen subsets. An index close to zero reflects that the overlap between two subsets is mostly due to chance.

The Kuncheva index only considers overlapping genes between two gene subsets, without taking into account nonoverlapping but highly correlated genes which may correspond to coordinated molecular changes. To address this issue, Zhang et al. proposed a measure called percentage of overlapping genes-related, POGR, defined as follows [38]:

$$POGR(\mathbf{r}_i, \mathbf{r}_j) = \frac{r + O_{i,j}}{k_i}, \tag{9}$$

where $k_i = |\mathbf{r}_i|$ denotes the size of the gene subset $\mathbf{r}_i$, $r = |\mathbf{r}_i \cap \mathbf{r}_j|$ denotes the number of overlapping genes, and $O_{i,j}$ denotes the number of genes in $\mathbf{r}_i$ which are not shared but significantly positively correlated with at least one gene in $\mathbf{r}_j$. To normalize the bias effect of subset size, nPOGR, the normalized POGR, is defined as

$$nPOGR(\mathbf{r}_i, \mathbf{r}_j) = \frac{r + O_{i,j} - E(r) - E(O_{i,j})}{k_i - E(r) - E(O_{i,j})}, \tag{10}$$

where $E(r)$ is the expected number of overlapping genes, and $E(O_{i,j})$ is the expected number of genes in $\mathbf{r}_i$ which are not shared but significantly positively correlated with at least one gene in $\mathbf{r}_j$, for two gene subsets (with sizes $|\mathbf{r}_i|$ and $|\mathbf{r}_j|$) randomly extracted from a given data set. The term $E(r)$ or $E(O_{i,j})$, respectively, corrects the bias due to the chance of selecting common genes or selecting significantly correlated genes between two randomly chosen gene subsets. Both definitions of POGR and nPOGR are nonsymmetric because it is possible that $|\mathbf{r}_i| \neq |\mathbf{r}_j|$ and/or $O_{i,j} \neq O_{j,i}$.

In our stability study, since we are interested in pairwise similarity between a number of gene subsets of equal size, we extend the original nPOGR measure into a symmetric measure by combining $nPOGR(\mathbf{r}_i, \mathbf{r}_j)$ and $nPOGR(\mathbf{r}_j, \mathbf{r}_i)$ as follows:

$$nPOGR(\mathbf{r}_i, \mathbf{r}_j) = \frac{r + O - E(r) - E(O)}{k - E(r) - E(O)}, \tag{11}$$

where $O = (O_{i,j} + O_{j,i})/2$, and $E(O)$ is the expected number of genes in one gene subset which are not shared but significantly positively correlated with at least one gene in the other subset, for any pair of gene subsets (with the same size) randomly extracted from a given data set. Note that this measure becomes the Kuncheva index if the two terms O and $E(O)$ about significantly correlated genes are removed. According to Zhang et al. [38], $E(O)$ is estimated based on 10,000 randomly generated pairs of gene subsets. Significantly correlated genes are determined based on Pearson correlation with 0.1 percent FDR control.

3.3.2 Classification Performance Measure

Since the data sets used in this study contain imbalanced class distributions (in particular, the lung cancer data set), we adopt a commonly used measure in this context, the area under the receiver operating characteristic (ROC) curve (denoted as AUC), to compare the classification performance of different methods. AUC is a function of two class-specific measures: sensitivity and specificity, defined as the proportions of correctly classified samples in the positive and the negative classes, respectively.

3.4 Experimental Procedures

For each data set used in the study, the entire data set was randomly split into the training set and the test set, with 2/3 of all the samples of each class in the training set, and the rest in the test set. The conventional, ensemble, and sample weighting versions of a baseline algorithm (SVM-RFE or ReliefF) were applied on the training set to select subsets of genes at various sizes (signature sizes as in Fig. 2 and Table 2). For each selected subset, both a linear SVM classifier (with default C parameter in Weka) and a K-nearest neighbor classifier (K = 1) were trained based on the selected genes and the training set, and then tested on the corresponding test set. For each data set, the above procedures were repeated 100 times. The stability of a selection method was measured over the 100 subsamplings of the data set according to (7). The classification performance of the method was measured by the average AUC over the 100 random training/test splits.

4 Results

4.1 Stability Performance

Fig. 2 reports the stability performance of the conventional, ensemble, and sample weighting versions of the SVM-RFE and ReliefF algorithms based on the Kuncheva index and nPOGR measures for the four microarray data sets used in our study. The result discussion below starts with a comparison among the three versions of SVM-RFE based on the Kuncheva index, and then expands to the alternative nPOGR measure and ReliefF algorithm.

From the four subgraphs in column (A), we can observe the following three major trends. First, the stability of SVM-RFE is very low for all of the four data sets. For example, at signature size 100 for the Colon data, the pairwise similarity between two gene signatures, on average (over the 100 generated gene signatures), is only 0.3 by the Kuncheva index, indicating on average roughly 30 percent overlap between any pair of gene signatures. Second, sample weighting consistently improves the stability of the gene signatures selected by SVM-RFE for all data sets at various signature sizes. In particular, the improvement becomes more significant as the signature size gets smaller. This is important because biologists usually focus on a few tens of most relevant genes to identify biomarker candidates for validation. Third, sample weighting is in general more effective than ensemble at improving the stability of SVM-RFE, especially at small signature sizes.

We now examine the stability performance of the three versions of SVM-RFE based on the nPOGR measure which takes into account significantly correlated genes in addition to overlapping genes between two gene signatures. From the four subgraphs in column (B), we can observe the same three trends discussed above for the three versions of SVM-RFE, since the stability curves measured by the nPOGR (in column (B)) are in general very close to those measured by the Kuncheva index (in column (A)) for the same algorithm and the same data set. Interestingly, considering significantly correlated genes does not improve the stability performance at our experimental settings. A close examination of the calculation of the nPOGR formula in (11) showed that usually a number of significantly correlated genes were

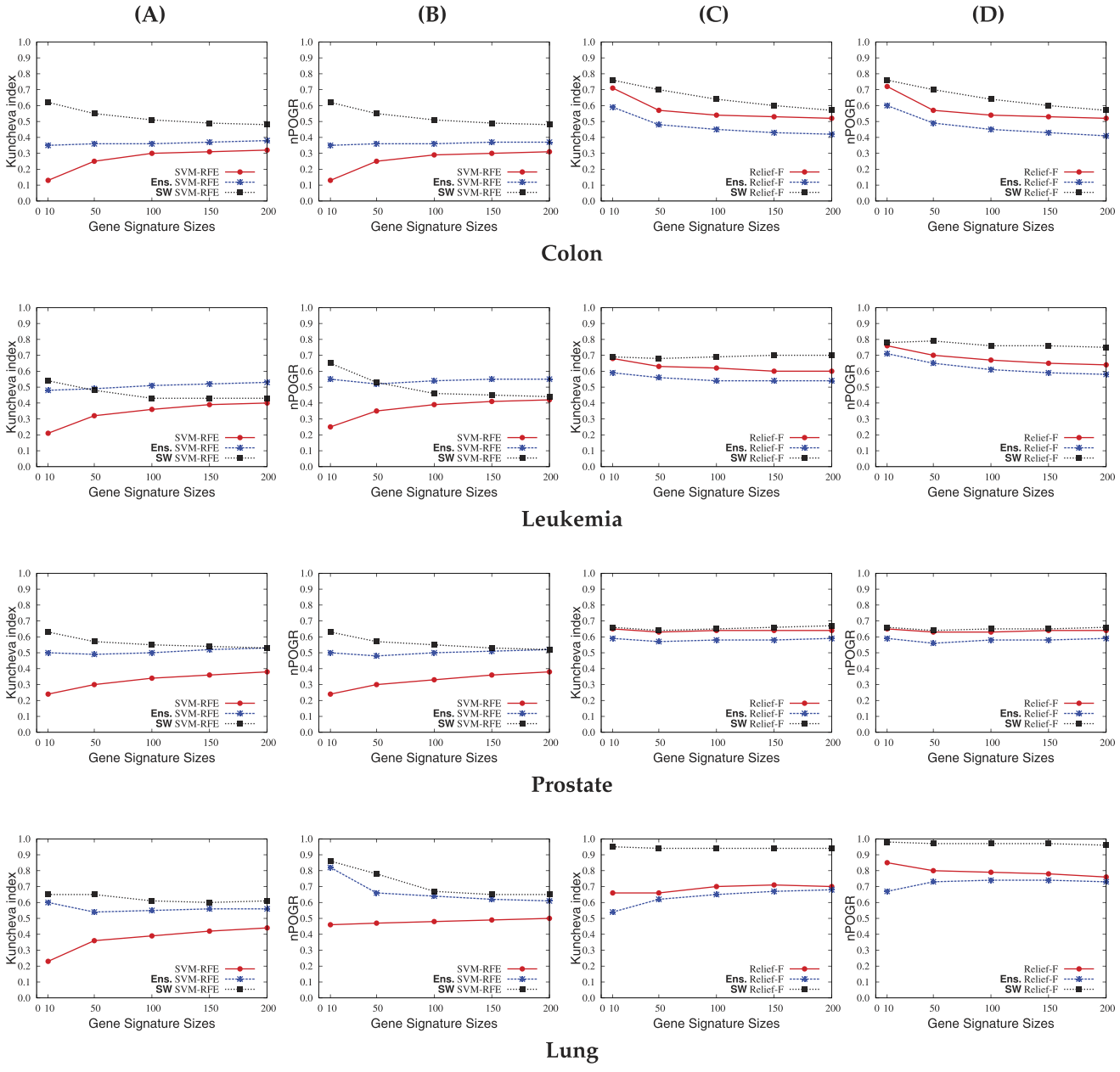


Fig. 2. Stability of the conventional, Ensemble, and Sample Weighting versions of SVM-RFE and ReliefF at increasing gene signature sizes. Subgraphs in column (A) based on the Kuncheva index and SVM-RFE; column (B) based on the nPOGR and SVM-RFE; column (C) based on the Kuncheva index and ReliefF; and column (D) based on the nPOGR and ReliefF.

identified in each case (i.e., term $O < 0$), but this number was often offset by the term $E(O)$, the expected number of significantly correlated genes from random selection. A few exceptions happened at small signature sizes for the Lung data, which led to noticeable increases of stability based on the nPOGR measure.

The above observations verify the effectiveness of sample weighting at improving the stability of SVM-RFE. We now examine its effectiveness for ReliefF. Subgraphs in columns (C) and (D) report the stability of the three versions of ReliefF based on the Kuncheva index and nPOGR measures, respectively. We can observe that the stability of ReliefF is consistently higher than SVM-RFE under either measure. Although ReliefF shows relatively more stable results, sample weighting still consistently improves its stability for all data sets except the Prostate data where ReliefF and SW ReliefF perform nearly the same. It is worth mentioning that for the Lung data, SW ReliefF exhibits almost perfect stability based on either measure. In contrast to sample weighting, the ensemble method exhibits a negative effect on the stability of ReliefF.

Overall, results from Fig. 2 verify that the proposed sample weighting framework is an effective approach to improving the stability of representative feature selection algorithms such as SVM-RFE and ReliefF. Compared with bagging-based ensemble, sample weighing is in general more effective. Although the stability appearance of a feature selection algorithm depends on the choice of stability measures, results from Fig. 2 suggest that considering significantly correlated genes in a stability measure does not dismiss the instability problem for existing feature selection algorithms under training data variations.

TABLE 2
Classification Performance Measured by the AUC (Average Value ± Standard Deviation) of the Linear SVM for the Conventional, Ensemble, and Sample Weighting Versions of SVM-RFE and ReliefF at Increasing Gene Signature Sizes

Data	Selection Method	Gene Signature Size				
		10	50	100	150	200
Colon	SVM-RFE	76.4±9.5	77.5±8.2	79.2±8.7	79.4±8.5	80.1±8.7
	Ens. SVM-RFE	80.3±7.9	79.4±9.0	78.6±8.3	78.6±9.1	79.4±8.7
	SW SVM-RFE	79.5±9.1	81.2±8.4	78.4±10.0	76.2±10.0	76.2±9.5
	ReliefF	78.8±8.8	80.1±8.8	78.5±8.7	77.5±8.9	76.1±8.5
	Ens. ReliefF	78.9±8.9	80.2±9.9	79.1±9.4	77.3±9.6	76.1±9.0
	SW ReliefF	78.3±8.2	79.6±9.4	78.1±9.4	76.4±10.0	75.4±10.0
Leukemia	SVM-RFE	92.8±5.8	96.3±3.8	96.9±3.3	96.8±3.5	97.0±3.4
	Ens. SVM-RFE	92.9±5.4	96.4±3.9	97.2±3.1	97.0±3.4	96.7±3.5
	SW SVM-RFE	91.2±5.6	96.2±3.9	96.4±3.3	96.5±3.4	96.8±3.5
	ReliefF	91.5±5.3	95.2±4.7	95.9±4.1	96.1±3.9	96.4±3.4
	Ens. ReliefF	91.3±5.5	94.7±4.3	95.7±4.0	96.3±3.7	96.2±3.8
	SW ReliefF	91.2±5.6	94.5±5.2	95.7±4.7	95.2±4.9	95.3±5.0
Prostate	SVM-RFE	89.8±5.1	91.3±4.1	92.1±3.8	92.1±4.3	92.2±3.9
	Ens. SVM-RFE	92.9±4.1	92.0±4.5	92.0±4.6	92.6±4.0	92.7±4.3
	SW SVM-RFE	93.4±3.6	91.3±4.5	90.0±4.8	90.7±4.9	91.2±4.7
	ReliefF	93.3±3.8	93.0±4.1	91.4±4.4	91.4±4.2	91.7±4.2
	Ens. ReliefF	93.4±3.5	92.4±4.0	91.4±4.1	91.0±4.4	91.9±4.2
	SW ReliefF	93.3±3.8	92.7±3.8	91.4±4.1	91.3±4.7	91.4±4.1
Lung	SVM-RFE	95.8±4.3	96.8±3.1	96.9±3.1	96.8±3.1	96.8±3.1
	Ens. SVM-RFE	96.3±3.5	96.9±3.2	96.9±3.1	97.0±3.1	96.9±3.1
	SW SVM-RFE	94.7±4.7	96.9±3.1	96.9±3.1	97.3±3.1	97.2±3.1
	ReliefF	96.2±4.2	96.7±3.2	96.7±3.3	97.0±3.3	97.4±3.0
	Ens. ReliefF	97.0±3.0	97.0±3.1	97.1±3.1	97.2±3.2	97.5±3.0
	SW ReliefF	96.8±4.7	96.7±4.0	98.6±2.4	98.4±2.4	98.8±2.2

4.2 Classification Performance

Table 2 reports the classification performance (by the AUC) of the linear SVM classifier for the conventional, ensemble, and sample weighting versions of SVM-RFE and ReliefF at increasing gene signature sizes for the four microarray data sets. We can observe that the three versions of SVM-RFE (or ReliefF) in general perform very similar under each signature size. Although there are some marginal differences among the three versions in terms of the average AUC values at places, the differences are not statistically significant, given the large standard deviation values caused by the small sample size of the test sets. These observations, together with those from the stability graphs in Fig. 2, suggest that different feature selection algorithms can lead to similarly good classification performance, while their stability can largely vary. Moreover, the increased stability to the baseline algorithms (SVM-RFE or ReliefF) resulted from the sample weighting process is not at the price of classification performance. Results based on 1NN classifier are very similar to those reported in Table 2, and thus excluded for conciseness of presentation.

4.3 Consensus Gene Signatures

We further demonstrate the effect of improved stability on constructing consensus gene signatures, using SVM-RFE as an example. Specifically, we compare the three versions of SVM-RFE by examining the selection frequency of each gene across the 100 random training/test splits of a given data set for each selection method. Given a data set, a selection method, and a gene signature size (e.g., 50), some genes are more consistently represented across the 100 generated gene signatures than others. A consensus gene signature can be constructed by extracting those genes frequently selected over many samplings. From a majority voting perspective, a gene is retained in the consensus gene signature (hence called a consensus gene), if it is represented in more than 50 percent of all gene signatures generated by the same method. The threshold 50 percent may be increased to shorten the consensus signatures while increasing the confidence on the consensus genes.

Fig. 3 shows the selection frequency curves of the three versions of SVM-RFE for the Colon data. Each curve shows the selection frequencies of all features selected by a corresponding version at signature size 50 (features occurring in none of the 100 signatures not shown). As a reference, the frequency curve for perfect stability (i.e., the same 50 genes appearing in all of the 100 signatures) is also shown. The area under the curve of each selection method

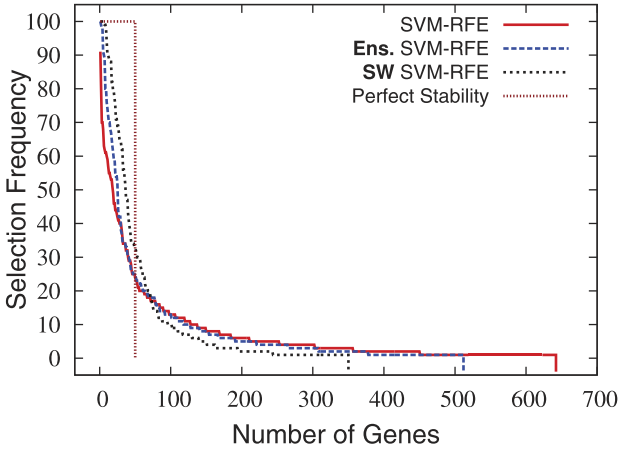


Fig. 3. Selection frequency plots of the conventional, Ensemble, and Sample Weighting versions of SVM-RFE for the Colon data. Each plot shows how many genes occur in at least how many of the 100 gene signatures of size 50 selected by each method. The area under the curve of each method equals the area under the perfect stability curve (100×50). The more overlap between the two areas, the more stable the method is.

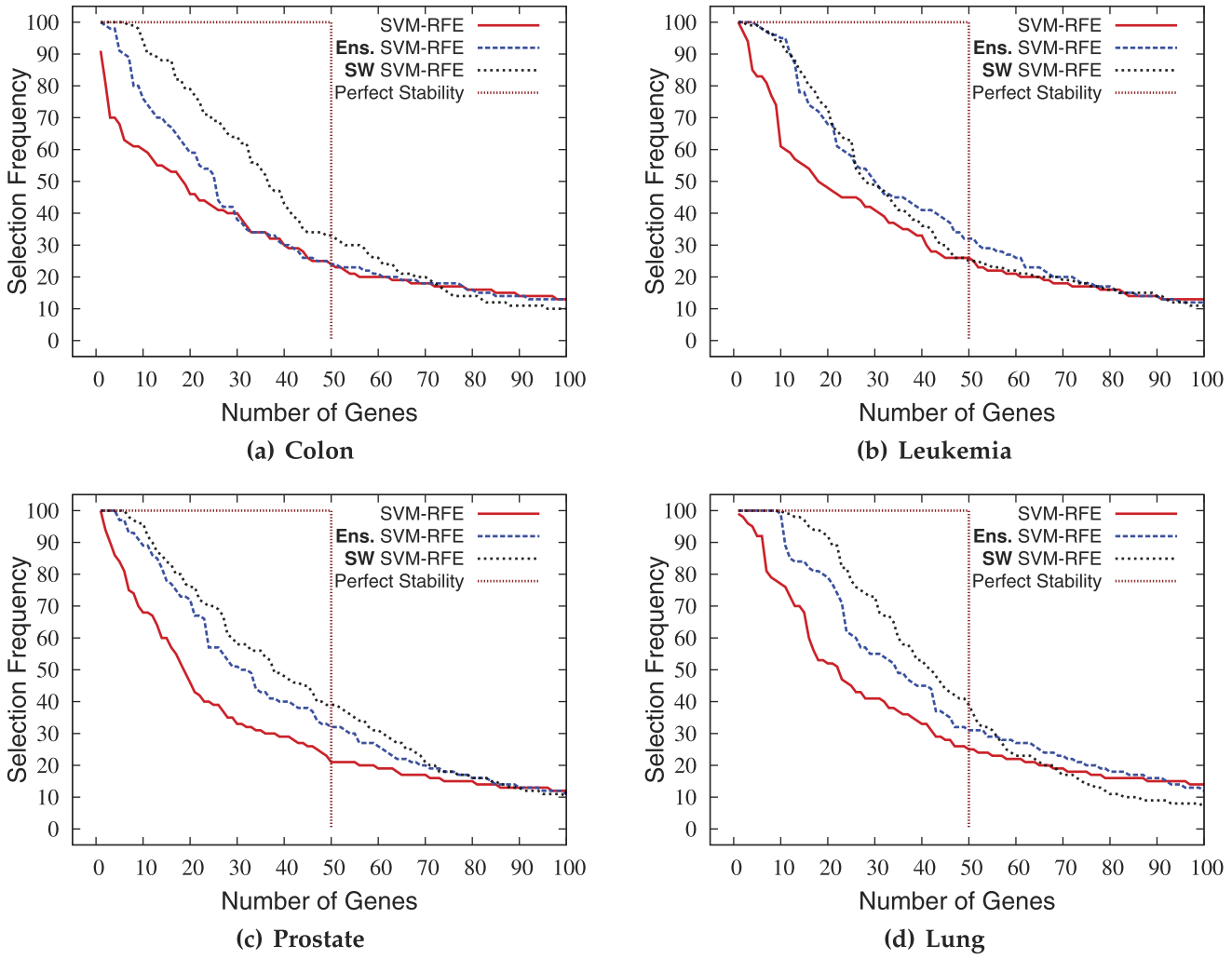


Fig. 4. Selection frequency plots of the conventional, Ensemble, and Sample Weighting versions of SVM-RFE for four data sets. Each plot shows how many genes occur in at least how many of the 100 gene signatures of size 50 selected by each version. Only the top 100 most frequently selected genes are included. **(a) Colon** provides a "zoomed in" view of Fig. 3.

equals the area under the perfect stability curve (i.e., the size of the 100×50 gene signature matrix). The more overlap between the area under the curve of a selection method and the area under the perfect stability curve, the more stable the selection method is. Comparing the curves of the three versions, we can observe that the sample weighting version is more stable than the other two versions; its curve is closer to the perfect stability curve (on the left side of the perfect stability curve) and has a much shorter tail (on the right size of the perfect stability curve).

Fig. 4a provides a "zoomed in" view of the frequency plots in Fig. 3 by focusing on the top 100 most frequently selected genes for the Colon data. Clearly, the sample weighting version consistently selects more consensus genes than the other two versions at various frequency threshold levels from 50 to 100 percent. For example, at the 50 percent threshold level, the conventional, ensemble, and sample weighting versions, respectively, select 18, 25, and 36 consensus genes. If the threshold level is increased to 85 percent, the consensus gene signature sizes for the three versions will shrink to 1, 7, and 16, respectively. These numbers are also reported in Table 3.

We performed the same analysis as above for all the four data sets used in this study. For the sake of conciseness of presentation, figures showing the full view of the frequency plots (as Fig. 3) for the Leukemia, Prostate, and Lung data

TABLE 3
The Numbers of Genes above Certain Selection Frequencies across 100 Gene Signatures of Size 50 Selected by the Conventional, Ensemble, and Sample Weighting Versions of SVM-RFE

Data	Selection Method	Frequency Intervals		
		[1,100]	(50,100]	(85,100]
Colon	SVM-RFE	642	18	1
	Ens. SVM-RFE	512	25	7
	SW SVM-RFE	**350**	**36**	**16**
Leukemia	SVM-RFE	688	18	4
	Ens. SVM-RFE	**397**	**30**	13
	SW SVM-RFE	469	28	**14**
Prostate	SVM-RFE	722	18	4
	Ens. SVM-RFE	371	32	13
	SW SVM-RFE	**262**	**37**	**15**
Lung	SVM-RFE	558	22	6
	Ens. SVM-RFE	308	34	12
	SW SVM-RFE	**246**	**42**	**22**

are not included. The "zoomed in" view of the frequency plots for each data set is provided in Fig. 4. Furthermore, Table 3 precisely reports the total numbers of genes that are selected in at least one of the 100 generated gene signatures as well as the numbers of consensus genes with frequency over 50 and 85 for each version on each data set.

From Fig. 4 and Table 3, we can observe that for all the four data sets used in our study, sample weighting significantly improves the stability of SVM-RFE. Comparing sample weighting with ensemble, the former clearly outperforms the latter for the Colon, Prostate, and Lung data. For the Leukemia data, the two methods perform very similar, with ensemble being slightly better. Such observations are consistent with those from Fig. 2 in Section 4.1, where the stability performance is measured with respect to pairwise gene signature similarity at various signature sizes. Fig. 4 and Table 3 in this section provide a more detailed view about the stability of the three versions of SVM-RFE at signature size 50 for each data set as shown in column (A) of Fig. 2. The analysis on consensus gene signatures also demonstrates the potential impact of the stability improvement by sample weighing. Sample weighting enables SVM-RFE to consistently select much fewer genes of low frequency and produce much bigger consensus gene signatures. Compared to unstable gene signatures, such consensus gene signatures may lead to higher confidence of biologists in selecting gene candidates for further examination and validation.

5 Conclusions

This paper studies the stability of feature selection from gene expression microarray data sets. The first contribution of this paper is a general framework of sample weighting to improve the stability of existing feature selection methods. The framework weights each sample in a training set according to its influence to the estimation of feature relevance, and then provides the weighted training set to a feature selection method. Various concrete sample weighting algorithms can be developed under this framework. The second contribution of this paper is the margin-based sample weighting algorithm developed under the general framework. The algorithm assigns a weight to each sample according to the outlying degree of its local profile of feature relevance (margin vector) compared with other samples. Our empirical study based on gene expression data sets has shown that the margin-based sample weighting algorithm is effective at improving the stability of representative SVM-RFE and ReliefF algorithms without sacrificing their predictive performance. The results suggest that the general framework of sample weighting is a promising approach to improving the stability of feature selection methods for gene selection. It is worth noting that the framework is not limited to feature selection from gene expression microarray data. It can be applied as a preprocessing step to feature selection from other types of high-throughput data such as protein mass spectrometry (MS) [28] and single nucleotide polymorphism (SNP) microarrays [4].

In the future work, we plan to develop additional sample weighting algorithms under the general framework and investigate their effectiveness on different feature selection methods. Since the sample weighting framework is not limited to work with a particular selection method, it is reasonable to expect that this framework could improve the stability of other selection methods as well. To apply the sample weighting framework to other selection methods, these selection methods need to be extended to take weighted samples as input and consider sample weights in feature evaluation.

Our empirical study also compared the sample weighting framework with the bagging ensemble framework. Although the former is in general more effective at improving the stability of the SVM-RFE method for the data sets used in this paper, it is not always the case. It would be interesting to study the effect of the two frameworks combined together on the stability of a feature selection method. It is worth noting that the sample weighting framework is computationally more efficient than the ensemble framework. The former applies an efficient sample weighting algorithm to a training set only once before feature selection, while the latter has to apply a base selection method to a training set a number of times.

As mentioned in the Introduction, stability of feature selection is an important and complicated issue; therefore, it would be worthwhile to develop additional frameworks to improve the stability of feature selection methods. Recent results show that incorporation of prior knowledge about feature relevance into the feature selection process [18] or feature selection based on several related data sets through transfer learning [17] is worth investigating.

Acknowledgments

The authors would like to thank the editor and anonymous reviewers for their helpful comments. This work was partially supported by the US National Science Foundation (NSF Grant No. 0855204).

References

[1] T. Abeel, T. Helleputte, Y.V. 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] 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, pp. 6745-6750, 1999.

[3] A.L. Boulesteix and M. Slawski, "Stability and Aggregation of Ranked Gene Lists," *Briefings in Bioinformatics,* vol. 10, no. 5, pp. 556-568, 2009.

[4] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley, and E.S. Lander, "Characterization of Single-Nucleotide Polymorphisms in Coding Regions of Human Genes," *Nature Genetics,* vol. 22, pp. 231-238, 1999.

[5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[6] K. Crammer, R. Gilad-Bachrach, and A. Navot, "Margin Analysis of the LVQ Algorithm," *Proc. 17th Conf. Neural Information Processing Systems,* pp. 462-469, 2002.

[7] C.A. Davis, F. Gerick, V. Hintermair, C.C. Friedel, K. Fundel, R. Küffner, and R. Zimmer, "Reliable Gene Signatures for Microarray Classification: Assessment of Stability and Performance," *Bioinformatics,* vol. 22, pp. 2356-2363, 2006.

[8] K.B. Duan, J.C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data," *IEEE Trans. NanoBioscience,* vol. 4, no. 3, pp. 228-234, Sept. 2005.

[9] J. Dutkowski and A. Gambin, "On Consensus Biomarker Selection," *BMC Bioinformatics,* vol. 8(Suppl 5):S5, 2007, doi:10.1186/1471-2105-8-S5-S5.

[10] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome Signature Genes in Breast Cancer: Is There a Unique Set?" *Bioinformatics,* vol. 21, pp. 171-178, 2005.

[11] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of Samples Are Needed to Generate a Robust Gene List for Predicting Outcome in Cancer," *Proc. Nat'l Academy of Sciences USA,* vol. 103, no. 15, pp. 5923-5928, 2006.

[12] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Computer Systems and Science,* vol. 55, no. 1, pp. 119-139, 1997.

[13] 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," *Science,* vol. 286, pp. 531-537, 1999.

[14] G.J. Gordon, R.V. Jensen, L. Hsiaoand, 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 Research,* vol. 62, pp. 4963-4967, 2002.

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

[16] Y. Han and L. Yu, "A Variance Reduction Framework for Stable Feature Selection," *Proc. 10th IEEE Int'l Conf. Data Mining,* pp. 206-215, 2010.

[17] T. Helleputte and P. Dupont, "Feature Selection by Transfer Learning with Linear Regularized Models," *Proc. 19th European Conf. Machine Learning (ECML '09),* pp. 533-547, 2009.

[18] T. Helleputte and P. Dupont, "Partially Supervised Feature Selection with Regularized Linear Models," *Proc. 26th Int'l Conf. Machine Learning,* pp. 409-416, 2009.

[19] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, "Algebraic Stability Indicators for Ranked Lists in Molecular Profiling," *Bioinformatics,* vol. 24, no. 2, pp. 258-264, 2008.

[20] A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces," *Knowledge and Information Systems,* vol. 12, pp. 95-116, 2007.

[21] L. Kuncheva, "A Stability Index for Feature Selection," *Proc. 25th Int'l Multi-Conf.: Artificial Intelligence and Applications,* pp. 390-395, 2007.

[22] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, and B.K. Mallick, "Gene Selection: A Bayesian Variable Selection Approach," *Bioinformatics,* vol. 19, no. 1, pp. 90-97, 2003.

[23] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," *Bioinformatics,* vol. 20, pp. 2429-2437, 2004.

[24] 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,* vol. 13, pp. 51-60, 2002.

[25] S. Loscalzo, L. Yu, and C. Ding, "Consensus Group Based Stable Feature Selection," *Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09),* pp. 567-576, http://portal.acm.org/citation.cfm?id=1557019.1557084, 2009.

[26] P.A. Mundra and J.C. Rajapakse, "SVM-RFE with MRMR Filter for Gene Selection," *IEEE Trans. NanoBioscience,* vol. 9, no. 1, pp. 31-37, Mar. 2010.

[27] M.S. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, M. Winget, and Y. Yasui, "Phases of Biomarker Development for Early Detection of Cancer," *J. Nat'l Cancer Inst.,* vol. 93, pp. 1054-1060, 2001.

[28] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, and L.A. Liotta, "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer," *Lancet,* vol. 359, pp. 572-577, 2002.

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

[30] B.Y. Rubinstein, *Simulation and the Monte Carlo Method.* John Wiley & Sons, 1981.

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

[32] 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. 2, pp. 203-209, 2002.

[33] Y. 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 2007.

[34] I.H. Witten and E. Frank, *Data Mining - Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers, 2005.

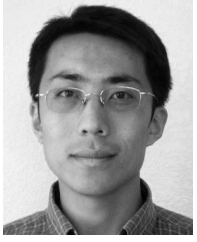
[35] Y.H. Yang, Y. Xiao, and M.R. Segal, "Identifying Differentially Expressed Genes from Microarray Experiments via Statistic Synthesis," *Bioinformatics,* vol. 21, no. 7, pp. 1084-1093, 2005.

[36] J. Ye, J. Chen, R. Janardan, and S. Kumar, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 1, no. 4, pp. 181-190, Oct.-Dec. 2004.

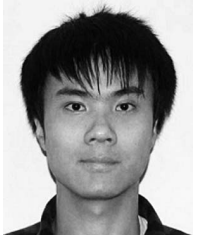
[37] K.Y. Yeung, R.E. Bumgarner, and A.E. Raftery, "Bayesian Model Averaging: Development of an Improved Multi-Class, Gene Selection and Classification Tool for Microarray Data," *Bioinformatics,* vol. 21, no. 10, pp. 2394-2402, 2005.

[38] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang, C. Wang, and Z. Guo, "Evaluating Reproducibility of Differential Expression Discoveries in Microarray Studies by Considering Correlated Molecular Changes," *Bioinformatics,* vol. 25, no. 13, pp. 1662-1668, 2009.

[39] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature Selection for Gene Expression Using Model-Based Entropy," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 7, no. 1, pp. 25-36, Jan.-Mar. 2010.



Lei Yu received the bachelor's degree in computer engineering from Dalian University of Technology, China, in 1999, and the PhD degree in computer science from Arizona State University in 2005. He is currently an associate professor at the Department of Computer Science at Binghamton University, State University of New York. His research interests include data mining, machine learning, and bioinformatics. He is a member of the IEEE.



Yue Han received the bachelor's degree in computer engineering from Beijing Jiaotong University, China, in 2007. He is currently working toward the PhD degree in computer science at Binghamton University, State University of New York. His research interests include data mining, machine learning, and bioinformatics.

Michael E. Berens received the bachelor's degree in zoology from Arizona State University in 1976 and the PhD degree in biology from the University of Arizona in 1982. He conducted postdoc research on Experimental Oncology at the University of Zurich, Switzerland, during 1982-1984. He is currently a senior research investigator and the director of the Cancer and Cell Biology Division at the Translational Genomics Research Institute in Phoenix, Arizona.