

Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data

Kai-Bo Duan, Jagath C. Rajapakse*, *Senior Member, IEEE*, Haiying Wang, *Member, IEEE*, and Francisco Azuaje, *Senior Member, IEEE*

Abstract—This paper proposes a new feature selection method that uses a backward elimination procedure similar to that implemented in support vector machine recursive feature elimination (SVM-RFE). Unlike the SVM-RFE method, at each step, the proposed approach computes the feature ranking score from a statistical analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data. We tested the proposed method on four gene expression datasets for cancer classification. The results show that the proposed feature selection method selects better gene subsets than the original SVM-RFE and improves the classification accuracy. A Gene Ontology-based similarity assessment indicates that the selected subsets are functionally diverse, further validating our gene selection method. This investigation also suggests that, for gene expression-based cancer classification, average test error from multiple partitions of training and test sets can be recommended as a reference of performance quality.

Index Terms—Cancer classification, feature selection, gene expression, gene ontology, semantic similarity, support vector machine recursive feature elimination (SVM-RFE).

I. INTRODUCTION

RECENTLY there has been an explosion in using gene expression phenotype for identification and classification in a variety of diagnostic areas. However, nowadays, for most of gene expression data for cancer classification, the number of training samples is still very small compared to the large number of genes involved in the experiments. When the number of genes is significantly greater than the number of samples, it is possible to find both random and biologically relevant correlations of gene behavior with the sample categories. To protect against spurious results, identifying a smallest possible but most informative subset of genes is the goal of *gene selection*. This is an important machine learning problem, which is referred to as feature selection [1], [2]. In addition, a small subset of genes is

also desirable in developing gene expression-based diagnostic tools. However, the small number of training samples and a large number of genes make gene selection a more relevant and challenging problem in gene expression-based cancer classification.

Traditional statistical methods for clustering and classification have been extensively used for gene selection [3]–[6]. Support vector machines (SVMs) [7], [8] also have been widely used for solving classification problems. In [9], linear SVMs are used in a backward elimination procedure for gene selection, and the selection procedure is therein referred to as SVM recursive feature elimination (SVM-RFE). Compared with other feature selection methods, SVM-RFE is a scalable, efficient wrappers method. Further information about other feature selection methods can be found in [10].

In this paper, we present a new gene selection method that uses a backward elimination procedure similar to that of SVM-RFE [9], but which, at each step, computes the ranking scores from a statistical analysis of the weight vectors of multiple linear SVMs trained on subsamples of the original training data by resampling. Correspondingly, a new feature ranking criterion is proposed. This method is tested on cancer classification tasks based on gene expression data.

The remainder of the paper is organized as follows: in Section II, we review the SVM-RFE method; in Section III, we review cross-validation (CV) performance estimation and describe the new gene selection method, which will be referred to as MSVM-RFE; in Section IV, MSVM-RFE and SVM-RFE are compared and evaluated with numerical experiments on four gene expression data for cancer classification; in Section V, we analyze the experiment results and conclude the paper.

II. SVM-RFE FEATURE SELECTION

SVM-RFE feature selection method was proposed in [9] to conduct gene selection for cancer classification. Nested subsets of features are selected in a sequential backward elimination manner, which starts with all the feature variables and removes one feature variable at a time. At each step, the coefficients of the weight vector \mathbf{w} of a linear SVM are used to compute the feature ranking score. The feature—say, the i th feature—with the smallest ranking score $c_i = (w_i)^2$ is eliminated, where w_i represents the corresponding component in the weight vector \mathbf{w} .

Using $c_i = (w_i)^2$ as the ranking criterion corresponds to removing the feature whose removal changes the objective function the least. This objective function is chosen to be $J = (1/2)\|\mathbf{w}\|^2$ in SVM-RFE. This is explained by the Optimal Brain Damage (OBD) algorithm [11], which approximates the change in objective function caused by removing

Manuscript received March 22, 2005; revised June 14, 2005. The work of J. C. Rajapakse is supported by Nanyang Technological University (NTU), the Ministry of Education (MOE), the Science and Engineering Research Council (SERC) and Bio-Medical Research Council (BMRC) of the Agency of Science and Technology for Research (A*Star), and National Grid Office (NGO), Singapore and Singapore-MIT Alliance (Computation and Systems Biology program). Asterisk indicates corresponding author.

K.-B. Duan is with the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: askbduan@ntu.edu.sg).

*J. C. Rajapakse is with the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (e-mail: asjagath@ntu.edu.sg).

H. Wang is with the School of Computing and Mathematics, University of Ulster, Jordanstown, UK (e-mail: hy.wang@ulster.ac.uk).

F. Azuaje is with the Computer Science Research Institute, University of Ulster, Jordanstown, UK (e-mail: fj.azuaje@ieee.org).

Digital Object Identifier 10.1109/TNB.2005.853657

a given feature by expanding the objective function in Taylor series to second order

$$\Delta J(i) = \frac{\partial J}{\partial w_i} \Delta w_i + \frac{\partial^2 J}{\partial w_i^2} (\Delta w_i)^2. \quad (1)$$

At the optimum of J , the first-order term can be neglected and with $J = (1/2)\|\mathbf{w}\|^2$, (1) becomes

$$\Delta J(i) = (\Delta w_i)^2. \quad (2)$$

$\Delta w_i = w_i$ corresponds to removing the i th feature.

Another explanation for using $(w_i)^2$ as the ranking criterion is from the sensitivity analysis of the objective function $J = (1/2)\|\mathbf{w}\|^2$ with respect to a variable. To compute the gradient, a virtual scaling factor ν is introduced into the kernel function [12] and $k(\mathbf{x}_i, \mathbf{x}_j)$ becomes $k(\nu \cdot \mathbf{x}_i, \nu \cdot \mathbf{x}_j)$. For a linear SVM (with a linear kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$), using the fact that $\nu_k = 1$, the sensitivity can be computed as

$$\begin{aligned} \frac{\partial J}{\partial \nu_k} &= \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial \nu_k} \\ &= \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j (2\nu_k x_k^2) \\ &= w_k^2. \end{aligned}$$

The recursive elimination procedure of SVM-RFE is implemented as follows:

- 1) Start: ranked feature set $R = []$; selected feature subset $S = [1, \dots, d]$;
- 2) Repeat until all features are ranked:
 - a) Train a linear SVM with features in set S as input variables;
 - b) Compute the weight vector;
 - c) Compute the ranking scores for features in set S : $c_i = (w_i)^2$;
 - d) Find the feature with the smallest ranking score: $e = \arg \min_i c_i$;
 - e) Update: $R = [e, R]$, $S = S - [e]$;
- 3) Output: Ranked feature list R .

Due to computing efficiency reasons, the algorithm can be generalized to remove more than one feature per step [9]. However, the removal of several features at a time may degrade the performance of the feature selection method.

III. MSVM-RFE

In this section, we will review the CV techniques and describe the new feature selection method, MSVM-RFE.

A. CV

CV [13] is basically a method for estimating predictive generalization error based on resampling. The resulting estimate of generalization error is often used for model selection by choosing the model that has the smallest estimated generalization error.

CV has many variants; in k -fold CV, the data samples are randomly split into k mutually exclusive subsets (or folds) of approximately equal size. A learning model is trained k times,

each time leaving out one of the subsets from training, but using only the omitted subset to compute whatever error criterion. The average error from k times of training and testing gives an estimate of the generalization error. A variant of this algorithm is known as ‘‘leave-one-out’’ (LOO) CV, in which k equals the sample size and one sample is left out in each training-test experiment. To reduce variability of CV estimate, k -fold CV may be run multiple times and an average of the estimates can be computed.

B. MSVM-RFE

Many feature selection methods are sensitive to small perturbations of the experimental conditions, as discussed in [10]. If the data have redundant variables, different subsets of variables with identical predictive power may be obtained according to initial conditions of the algorithm and the removal or addition of a few variables or training samples. One way to stabilize a feature selection method is to repeat the selection procedure on several subsamples from bootstrap resampling of training data. Bootstrap sampling samples the data uniformly with replacement, which results in a subsample of the original data. The final feature subset can be the union of variables selected by various repetitions on bootstrap subsamples, or can be decided by analyzing the behavior of variables being selected in variant bootstrap repetitions of the selection procedure.

The bootstrap stabilization idea can be applied to SVM-RFE. However, instead of applying this idea on SVM-RFE as a whole, we may apply it on each step of the recursive procedure of SVM-RFE. At each step of the SVM-RFE procedure, we train multiple linear SVMs on subsamples of training data and compute the feature ranking scores from statistical analysis of $(w_i)^2$, which is the feature ranking score of SVM-RFE. The subsamples can be obtained by bootstrap resampling of the training data, or by other resampling methods, such as k -fold CV.

Suppose that we have t linear SVMs trained on different subsamples of original training data. Let \mathbf{w}_j be the weight vector of the j th linear SVMs and w_{ji} be the corresponding weight value associated with the i th feature; let $v_{ji} = (w_{ji})^2$. We can compute the feature ranking score with the following criterion:

$$c_i = \frac{\bar{v}_i}{\sigma_{v_i}} \quad (3)$$

where \bar{v}_i and σ_{v_i} are mean and standard deviation of variable v_i

$$\bar{v}_i = \frac{1}{t} \sum_{j=1}^t v_{ji} \quad (4)$$

$$\sigma_{v_i} = \sqrt{\frac{\sum_{j=1}^t (v_{ji} - \bar{v}_i)^2}{t-1}}. \quad (5)$$

Before computing the ranking score for each feature, it is important to normalize the weight vectors

$$\mathbf{w}_j = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}. \quad (6)$$

We refer to this new feature selection method as MSVM-RFE, where MSVM stands for multiple SVMs.

Thus, the recursive procedure of MSVM-RFE can be described as follows:

- 1) Start: ranked feature set $R = []$; selected subset $S = [1, \dots, d]$;
- 2) Repeat until all features are ranked:
 - a) Train t linear SVMs on subsamples of the original training data, with features in set S as input variables;
 - b) Compute and normalize the weight vectors;
 - c) Compute the ranking scores c_i for features in S using (3);
 - d) Find the feature with the smallest ranking score: $e = \arg \min_i c_i$;
 - e) Update: $R = [e, R]$, $S = S - [e]$;
- 3) Output: Ranked feature list R .

As in SVM-RFE, we can eliminate several feature variables per step for computation efficiency at the risk of possible performance degradation. Also note that in order to use SVM-RFE and MSVM-RFE, it is important to normalize the values of each feature variable across the samples. The proposed MSVM-RFE is computationally more expensive than SVM-RFE. However, as feature selection is a prestep for building a good classifier, it is worthwhile to go through a computationally more expensive way if a better feature subset can be selected.

In our numerical study, k -fold CV will be chosen as the resampling method. Multiple linear SVMs will be obtained from m multiple runs of k -fold CV. The results reported in this paper were obtained with $m = 20$ and $k = 5$. Thus, at each step of MSVM-RFE, altogether we will have $t = 100$ linear SVMs. Theoretically, within the allowance of computation cost, m should be as large as possible. Multiple runs of fivefold or tenfold CV are often used for estimation of generalization error.

Although the recursive procedure of SVM-RFE and MSVM-RFE produces nested feature subsets, it does not tell us which is the best subset. If multiple runs of k -fold CV give a good estimate of the generalization error, we may be able to decide the best feature subset on the basis of this estimate within the recursive procedure of MSVM-RFE itself. This is also the reason why we choose k -fold CV as the resampling method for MSVM-RFE.

IV. NUMERICAL EXPERIMENTS

A. Experiment Settings

We evaluate SVM-RFE and MSVM-RFE on four gene expression datasets: Breast Cancer (Breast), Colon Tumor (Colon), ALL-AML Leukemia (Leukemia), and Lung Cancer (Lung). We obtained the datasets from [14]. We use the same training and test data reported in the previous studies, without changing the sample sizes, so that our results can be objectively compared with earlier methods. Table I summarizes some basic information about the datasets, including the number of training and test samples. More detailed information about the datasets can be found from [14] and the references therein.

SVMs have two very important hyperparameters, the kernel function and the regularization parameter C . For good performance, these parameters must be chosen carefully [15]. In our

TABLE I
SIZES OF TRAINING AND TEST SETS, NUMBER OF FEATURES, OF THE FOUR GENE EXPRESSION DATASETS

Dataset	#Training	#Testing	#Total Genes
Breast	78	19	24481
Colon	42	20	2000
Leukemia	38	34	7129
Lung	32	149	12533

study, linear SVMs are used, and we have only the C parameter to tune. The C values for these linear SVMs, either in performance validation or in the recursive feature selection procedures of SVM-RFE and MSVM-RFE, on all datasets, are chosen from the finite set $\{2^{-20}, \dots, 2^0, \dots, 2^{15}\}$ on the basis of classification performance estimate by tenfold CV. In MSVM-RFE, for each recursive step, the 100 linear SVMs from 20 runs of fivefold CV use the same C value that is determined beforehand by a tenfold CV search procedure as just described.

To speed up the feature selection in the numerical experiments on both SVM-RFE and MSVM-RFE, we eliminate r ($r \geq 1$) features each time when the number of features n in feature subset S is large. We choose $r = 100$ if $n > 10000$, $r = 10$ if $1000 < n \leq 10000$, and $r = 1$ if $n \leq 1000$.

The goodness of a gene subset is usually assessed by the performance of a classifier built on the training set with only features in the subset as input variables. As the features selected matters more than the classification algorithm used [9], different feature selection methods are compared with the same classification algorithm. Linear algorithms are commonly used in cancer classification with gene expression data. In this study, linear SVM is chosen as the classification algorithm.

Validation error on test set is usually used to assess the performance of a classifier. However, the training and test sets of our gene expression data are small and the test error may not be reliable due to possible "unfortunate" partition of the training and test sets. Thus, instead of reporting such a test error from one partition of training and test sets, we do as follows: we merge the training set and test set and then partition the total samples again into a training set and a test set randomly by stratified sampling, for 100 times; for each partition, we train a classifier on the training set and then test it on the corresponding test set; from this 100 trials, we compute the averages of test error, sensitivity, and specificity.

In our study, on each dataset, the gene selection procedure is carried out solely on the training set. The goodness of a feature subset is examined by the performance of a linear SVM classifier trained with features in the subset as input variables. For each method, we test feature subsets with size ranging from 1 to 200. In Figs. 1–4, we plot the average test errors of linear SVM classifiers on gene subsets selected respectively by SVM-RFE and MSVM-RFE, on the four datasets. We take the feature subset with the least average test error as the best feature subset and report the performance of the best feature subset of each gene selection method in Table II.

SVMs are capable of dealing with a large number of input variables with a slight increase in computation complexity. To see if feature selection improves the performance of SVMs, we

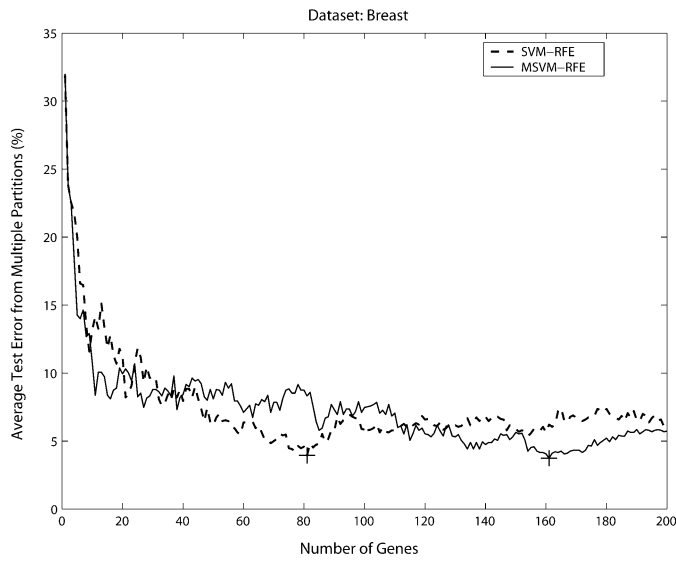


Fig. 1. Performance of feature subsets selected by SVM-RFE and MSVM-RFE, on Breast dataset. The performance of the best feature subset selected by each methods is denoted by “+.”

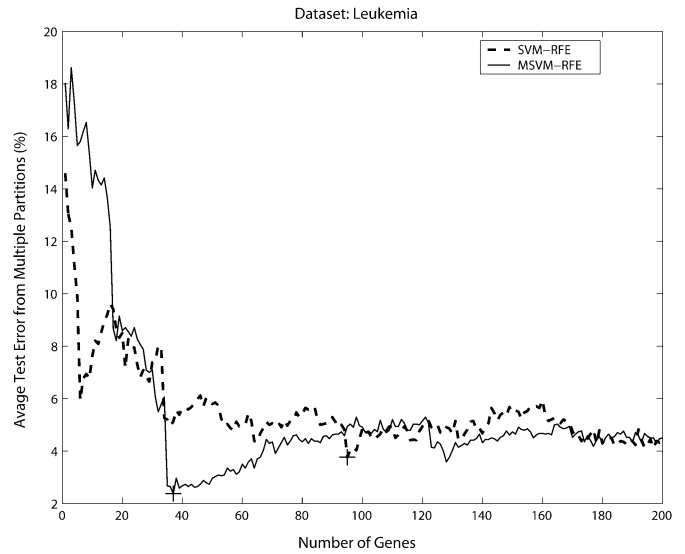


Fig. 3. Performance of feature subsets selected by SVM-RFE and MSVM-RFE, on Leukemia dataset. The performance of the best feature subset selected by each method is denoted by “+.”

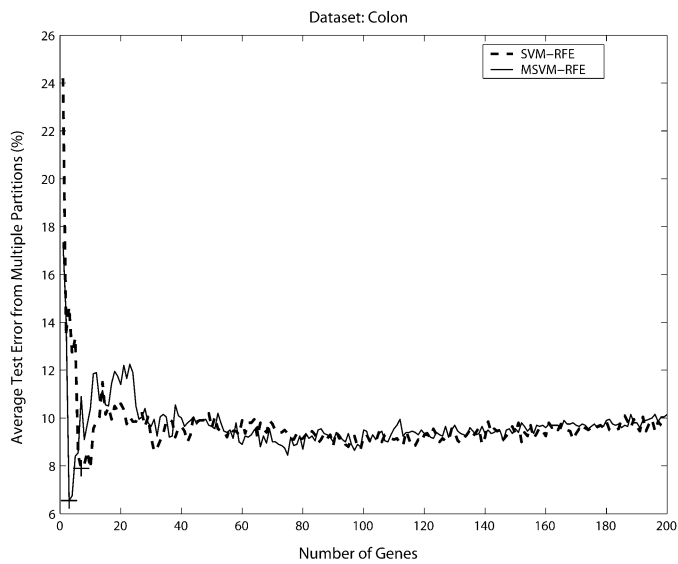


Fig. 2. Performance of feature subsets selected by SVM-RFE and MSVM-RFE, on Colon dataset. The performance of the best feature subset selected by each method is denoted by “+.”

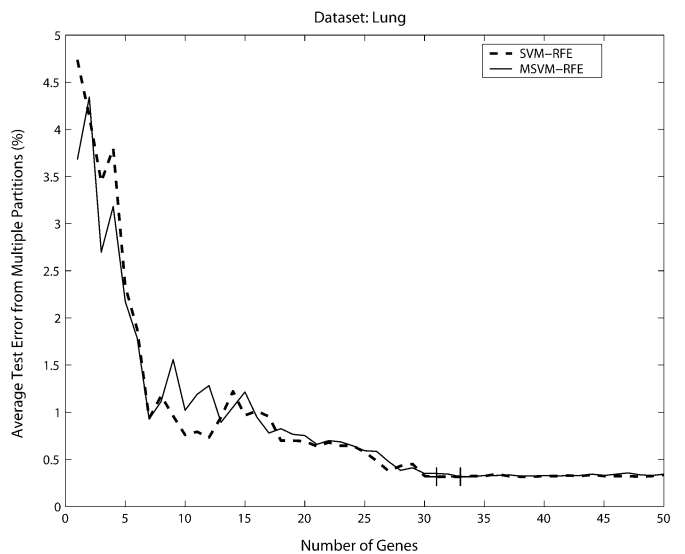


Fig. 4. Performance of feature subsets selected by SVM-RFE and MSVM-RFE, on Lung dataset. The performance of the best feature subset selected by each method is denoted by “+.”

also train and test linear SVMs with full number of genes as input variables, on the same 100 random partitions of training and test sets. The performance of SVM with full number of genes on four datasets are reported in Table II, together with the performance of SVM with gene feature selection by SVM-RFE and MSVM-RFE.

From Figs. 1–4 and Table II we can see that, on Breast, Colon, and Leukemia datasets, the best feature subsets selected by MSVM-RFE give better classification accuracy than the best feature subsets selected by SVM-RFE. On Lung dataset, the best feature subsets of the both methods achieve same perfect classification accuracy. At the best feature subsets of SVM-RFE and MSVM-RFE, good sensitivity and specificity are also attained.

The experiment results in Table II also show that, even for SVMs, which are capable of handling a large number of input

variables, the classification performance is improved significantly with gene selection either by SVM-RFE or MSVM-RFE on all the datasets.

B. Gene Ontology (GO)-Based Similarity Analysis of Selected Genes

The GO [16] is one of the most important knowledge resources within the bioinformatics community, which provides a structured and controlled vocabulary to annotate genes and gene products. It comprises three hierarchies, sometimes referred to as taxonomies or “aspects,” that respectively hold terms describing the molecular function (MF), biological process (BP), and cellular component (CC). The vocabularies (one for each ontology) and their relationships are represented in the form of directed acyclic graphs (DAGs), where the terms are represented

TABLE II
PERFORMANCE OF SVMs WITHOUT AND WITH FEATURE SELECTION BY SVM-RFE AND MSVM-RFE. VALUES OF TEST ERROR, SENSITIVITY, AND SPECIFICITY ARE ALL IN PERCENTAGES

Dataset	Measurement	SVM	SVM-RFE	MSVM-RFE
Breast	# Genes	Full (24481)	81	161
	Test Error	35.26±9.71	3.95±4.57	3.74±4.32
	Sensitivity	58.75±13.42	94.58±6.94	94.83±6.68
	Specificity	75.00±17.27	98.57±4.31	98.71±4.11
Colon	# Genes	Full (2000)	7	3
	Test Error	18.30±6.86	7.90±4.78	6.55±4.48
	Sensitivity	74.00±16.67	83.71±11.84	83.57±12.08
	Specificity	85.85±7.63	96.62±5.82	98.77±3.04
Leukemia	# Genes	Full (12582)	95	37
	Test Error	12.91±6.37	3.76±4.35	2.38±4.32
	Sensitivity	99.40±1.92	100.00±0.00	100.00±0.00
	Specificity	69.50±14.64	90.86±10.55	94.21±10.49
Lung	# Genes	Full (12533)	31	33
	Test Error	0.48±0.78	0.32±0.34	0.32±0.34
	Sensitivity	96.80±7.37	96.87±3.34	96.87±3.34
	Specificity	99.83±0.38	100.00±0.00	100.00±0.00

as nodes and connected by relationships represented as edges. In such a hierarchy each term may represent a “child node” of one or more “parent nodes” and the child-to-parent relationship is of two types: “is-a” and “part-of.” By providing a framework to store different repositories using the same standard vocabulary, the GO may facilitate information querying across the databases and several applications for supporting predictive data analysis in functional genomics.

A recent study confirmed that, in general, high and low GO-based similarity values are associated with high and low expression correlation values, respectively [17]. Gene selection algorithms typically aim to identify genes that are less correlated to each other so as to reduce the redundancy and maximize the information content. Moreover, functional diversity may be an indicator of the quality of the set of genes selected. Thus, relationships between GO-based similarity and gene expression correlation may offer a new approach to assessing the relevance of a set of genes selected. Such an assessment may be implemented by analyzing the pairs of GO-based similarities calculated for a selected set of genes.

Several information content-based measures have been proposed [18]–[20], which can be applied to GO-based similarity estimation. In this paper, we report results using the similarity measure proposed by Lin [20], which has shown to be an effective indicator of GO-based functional similarity [17]. Let O be the set of terms in GO; for each term $o \in O$, let $p(o)$ be the probability of finding a child of o in the taxonomy; let $U(o_i, o_j)$ be the set of parent terms shared by terms o_i and o_j . Lin’s similarity measure is defined as

$$\rho(o_i, o_j) = \frac{2 \times \max_{o \in U(o_i, o_j)} [\log(p(o))]}{\log(p(o_i)) + \log(p(o_j))}. \quad (7)$$

The value of Lin’s similarity measure varies between zero and one. The measure described in (7) estimates the similarity between a pair of GO terms. For a set G of n GO terms, we can

TABLE III
AVERAGE SIMILARITY AND DISTRIBUTION OF BEST GENE SUBSETS SELECTED BY SVM-RFE AND MSVM-RFE, ON BREAST DATASET. A TOTAL OF 76 HUGO SYMBOLS (2850 GENE PAIRS) AND 144 HUGO SYMBOLS (10299 GENE PAIRS) ARE ANALYZED, RESPECTIVELY, FOR SVM-RFE AND MSVM-RFE

Hierarchy	Method	Average Similarity	Similarity Distribution				
			0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
MF	SVM-RFE	0.0112	2787	39	19	3	2
	MSVM-RFE	0.0062	10181	70	29	9	7
BP	SVM-RFE	0.0177	2760	69	15	4	2
	MSVM-RFE	0.0094	10134	128	24	5	5
CC	SVM-RFE	0.0171	2779	35	12	6	18
	MSVM-RFE	0.0087	10161	64	37	13	21

TABLE IV
AVERAGE SIMILARITY AND DISTRIBUTION OF BEST GENE SUBSETS SELECTED BY SVM-RFE AND MSVM-RFE, ON COLON DATASET. A TOTAL OF 5 HUGO SYMBOLS (10 GENE PAIRS) AND 2 HUGO SYMBOLS (1 GENE PAIRS) ARE ANALYZED, RESPECTIVELY, FOR SVM-RFE AND MSVM-RFE

Hierarchy	Method	Average Similarity	Similarity Distribution				
			0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
MF	SVM-RFE	0.0074	10	0	0	0	0
	MSVM-RFE	0.0000	1	0	0	0	0
BP	SVM-RFE	0.0259	10	0	0	0	0
	MSVM-RFE	0.0000	1	0	0	0	0
CC	SVM-RFE	0.1631	7	2	0	0	1
	MSVM-RFE	0.0000	1	0	0	0	0

measure the similarity of terms within this set by, for instance, averaging the similarities of the $n(n-1)/2$ unique pairs of terms

$$\rho(G) = \frac{2}{n(n-1)} \sum_{o_i, o_j \in G, o_i \neq o_j} \rho(o_i, o_j) \quad (8)$$

where $\rho(o_i, o_j)$ is computed using (7). The similarity of a pair of genes may be then computed as the average similarity between terms originating from the two genes as described in [17].

Based on the February 2004 GO Annotation@EBI (GOA) release, we studied the functional similarity exhibited by pairs of genes within the subsets selected by SVM-RFE and MSVM-RFE, on the three datasets (except for Lung dataset, where we cannot find the gene names from the data source). Genes having no Human Genome Organization¹ (HUGO) symbol² were excluded from this analysis. The average similarity measured by (8) and the similarity distribution over all the gene pairs are reported in Tables III–V. GO-based similarity values were obtained using annotations from the three GO hierarchies: MF, BP, and CC.

As seen from Tables III–V, the average GO-based similarities of all selected gene subsets on all the datasets are low in all the three aspects of GO; the distribution analyses also show that the similarities of most gene pairs in the gene sets are located to the very lower end; on all the datasets and in the three aspects of GO, average similarities of genes selected by MSVM-RFE are even lower than those of genes selected by SVM-RFE.

¹[Online.] Available: <http://www.hugo-international.org/>

²Refer to the HUGO Gene Nomenclature Committee Web site for HUGO gene nomenclature. [Online.] Available: <http://www.gene.ucl.ac.uk/nomenclature/index.html>

TABLE V

AVERAGE SIMILARITY AND DISTRIBUTION OF BEST GENE SUBSETS SELECTED BY SVM-RFE AND MSVM-RFE, ON LEUKEMIA DATASET. A TOTAL OF 92 HUGO SYMBOLS (4186 GENE PAIRS) AND 37 HUGO SYMBOLS (666 GENE PAIRS) ARE ANALYZED, RESPECTIVELY, FOR SVM-RFE AND MSVM-RFE

Hierarchy	Method	Average Similarity	Similarity Distribution				
			0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
MF	SVM-RFE	0.0145	4086	69	22	5	4
	MSVM-RFE	0.0115	658	2	5	0	1
BP	SVM-RFE	0.0209	4066	97	19	2	2
	MSVM-RFE	0.0197	649	13	3	0	1
CC	SVM-RFE	0.0324	3946	141	57	16	26
	MSVM-RFE	0.0130	649	14	1	1	1

V. DISCUSSION AND CONCLUSION

As seen in Table II and Figs. 1–4, the performance of MSVM-RFE is better than or comparable to SVM-RFE. It is also noted from Table II that, with feature selection either by SVM-RFE or MSVM-RFE, SVM shows better performance on all datasets. This observation further testifies the usefulness and importance of the gene selection for cancer classification with expression data.

Although the multiple linear SVMs for MSVM-RFE in our numerical study are from multiple runs of k -fold CV, we believe multiple SVMs trained on subsamples from bootstrap or other resampling methods will also work well under the same framework of MSVM-RFE.

From Table II, we observe that the standard deviations of performance measures (test error, sensitivity, and specificity) over the 100 times of training and testing are usually large. This implies that the variability of results from one single test is large, and such test results thus are not fair performance references due to possible “unfortunate” partition. When the data size is small, the chance of an unfortunate partition is high. For gene expression data, mean-standard-deviation format of performance measurement from multiple partitions of training and test sets can be a fairer reference of performance quality.

The results from the GO-based similarity analysis in Tables III–V show low similarities among the selected gene set. The similarity of most pairs of genes are also low, which is an indicator of functional diversity. Moreover, as low GO-based similarity typically implies low gene expression correlation, the low average similarity shown may also be associated with a low degree of (annotation- and expression-based) redundancy among the selected genes. The low similarity values among the genes selected by MSVM-RFE may in a way explain the better performance of MSVM-RFE over SVM-RFE. Such a GO-based similarity assessment analysis provides us with an alternative way to assess the relevance of the gene subsets derived from a gene selection procedure.

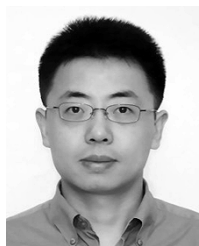
As we mentioned earlier, the use of CV instead of bootstrap sampling as the resampling method explores the possibility to determine the best feature subset within the recursive procedure of MSVM-RFE on the basis of the generalization error estimate from multiple runs of k -fold CV. This generalization error estimate could have a very poor correlation with average test error from the 100 times of training and testing, which is taken by us

as the “true” error estimate. The poor generalization error estimation of multiple runs of k -fold CV is due to the small sizes of training sets of the gene expression datasets used in our study.

We conclude that: 1) the proposed MSVM-RFE method can select better gene subsets than SVM-RFE and improve the cancer classification accuracy; 2) gene selection also improves the performance of SVMs and is a necessary step for cancer classification with gene expression data; and 3) GO-based similarity values of pairs of genes belonging to subsets selected by MSVM-RFE are significantly low, which may be seen as an indicator of functional diversity (or redundancy reduction). This further supports our hypothesis that MSVM-RFE is a powerful and meaningful approach to gene selection in cancer classification. In addition, the mean-standard-deviation format of performance report from multiple partitions of training and test sets is recommended as a more accurate reference of performance for cancer classification with gene expression data when the number of available data samples is small.

REFERENCES

- [1] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.
- [2] A. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.
- [3] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, “Molecular classification of cancer: Class discovery and class prediction by gene expression,” *Science*, vol. 286, pp. 531–537, 1999.
- [4] A. Alizadeh *et al.*, “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, pp. 503–511, 2000.
- [5] M. Eisen, P. Spellman, P. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” in *Proc. Nat. Acad. Sci. USA*, vol. 95, 1998, pp. 14 863–14 868.
- [6] I. Hedenfalk *et al.*, “Gene expression profiles in hereditary breast cancer,” *New Eng. J. Med.*, vol. 344, pp. 539–558, 2001.
- [7] B. Boser, I. Guyon, and V. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proc. 5th Annu. Workshop Computational Learning Theory*, 1992, pp. 114–152.
- [8] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
- [9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.
- [10] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, no. 3, pp. 1157–1182, Mar. 2003.
- [11] Y. LeCun, J. Denker, S. Solla, R. Howard, and L. D. Jackel, “Optimal brain damage,” in *Advances in Neural Information Processing Systems II*, D. S. Touretzky, Ed. Mateo, CA: Morgan Kaufmann, 1990.
- [12] A. Rakotomamonjy, “Variable selection using SVM-based criteria,” *J. Mach. Learn. Res. (Special Issue on Variable Selection)*, vol. 3, pp. 1357–1370, 2003.
- [13] M. Stone, “Cross-validated choice and assessment of statistical predictions,” *J. Roy. Stat. Soc. B*, vol. 36, no. 1, pp. 111–147, 1974.
- [14] J. Li and H. Liu. (2002) Kent Ridge Bio-Medical Data Set Repository. [Online]. Available: <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>
- [15] K.-B. Duan, S. Keerthi, and A. Poo, “Evaluation of simple performance measures for tuning SVM hyperparameters,” *Neurocomputing*, vol. 51, pp. 41–59, Apr. 2003.
- [16] Gene Ontology Consortium, “Creating the gene ontology resource: Design and implementation,” in *Genome Res.*, 2001, vol. 11, pp. 1245–1233.
- [17] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, “Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships,” in *Proc. 2004 IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology*, pp. 25–31.
- [18] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” in *Proc. 14th Int. Joint Conf. Artificial Intelligence*, 1995, pp. 448–453.
- [19] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” in *Proc. Int. Conf. Research in Computational Linguistics*, 1998, pp. 19–33.
- [20] D. Lin, “An information-theoretic definition of similarity,” in *Proc. 15th Int. Conf. Machine Learning*, 1998, pp. 296–304.



Kai-Bo Duan received the B.Eng. degree in power engineering and the M.Eng. degree in mechanical engineering from Nanjing University of Aeronautics and Astronautics, China, in 1996 and 1999, respectively, and the Ph.D. degree in mechanical engineering from National University of Singapore in 2004 for his research work on SVMs and kernel classification methods.

He joined the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, in 2003 and is currently a

Research Fellow working on machine learning methods to computational biology. His research interests include machine learning, data mining, and computational biology.



Jagath C. Rajapakse (S'90–M'91–SM'00) received the B.Sc. (Eng.) degree with First-Class Honors in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 1985 and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the State University of New York, Buffalo, in 1989 and 1993, respectively.

He was a Visiting Fellow at the National Institute of Mental Health, Bethesda, MD, and a Visiting Scientist at the Max-Planck-Institute of Cognitive Neuroscience, Leipzig, Germany. He is currently an As-

sociate Professor in the School of Computer Engineering (SCE) and the Deputy Director of the BioInformatics Research Centre (BIRC) at Nanyang Technological University (NTU), Singapore. He has authored more than 160 research publications in refereed journals, books, and conference proceedings, in the fields of brain imaging, computational biology, and machine learning. His papers are among the highly cited papers in these fields. The mission of his research is to investigate the neural correlates and the genetic mechanisms of human brain function and disease by using neuroimaging and bioinformatics approaches, leading to new therapies and drugs for brain disease. His current research projects are supported by grants from NTU, the Ministry of Education (MOE), Science and Engineering Research Council (SERC) and Bio-Medical Research Council (BMRC) of the Agency of Science and Technology for Research (A*Star), and National Grid Office (NGO), Singapore and Singapore-MIT Alliance (Computation and Systems Biology program).



Haiying Wang (S'03–M'05) received the B.Eng. and M.Sc. degrees in optical electronics engineering from Zhejiang University, Hangzhou, China, in 1987 and 1989, respectively, and the Ph.D. degree in artificial intelligence in biomedicine from the University of Ulster, Jordanstown, U.K., in 2004.

He was a Senior Engineer in applied electronics at the Fujian Electronic Technology Institute, Fuzhou, China, before he joined the University of Ulster in 2000. He is currently a Postdoctoral Research Fellow within the School of Computing and Mathematics,

University of Ulster. He has published several publications in journals, books and conference proceedings in electronic systems and biomedical informatics. His research interests include knowledge engineering, data mining, machine learning, and their applications in medical informatics and bioinformatics.

Francisco Azuaje (M'96–SM'05) received the B.Sc. degree in electronic engineering from Simon Bolivar University, Caracas, Venezuela, in 1995. He was a student of the Master in Policy and Management of Technological Innovation at Central University of Venezuela in 1996 and received the Ph.D. on artificial intelligence and medical informatics from the University of Ulster, Jordanstown, U.K., in 2000.

Before joining the University of Ulster as a Reader in 2002, he was a Lecturer at the Department of Computer Science of Trinity College, Dublin, Ireland. He has published several refereed publications in journals, books and conference proceedings relating to the areas of bioinformatics, artificial intelligence, and medical informatics. He is an Editorial Board Member of *BioMedical Engineering OnLine*, *Cancer Informatics*, and the *Online Journal of Bioinformatics*. He has coedited two books in the areas of bioinformatics and systems biology.

Dr. Azuaje is an Editorial Board Member of the IEEE TRANSACTIONS ON NANOBIOSCIENCE. He administrates the *IEEE Forum on Bioinformatics and Systems Biology*.