SVM-RFE With MRMR Filter for Gene Selection

Piyushkumar A. Mundra and Jagath C. Rajapakse*, *Senior Member, IEEE*

***Abstract*—We enhance the support vector machine recursive feature elimination (SVM-RFE) method for gene selection by incorporating a minimum-redundancy maximum-relevancy (MRMR) filter. The relevancy of a set of genes are measured by the mutual information among genes and class labels, and the redundancy is given by the mutual information among the genes. The method improved identification of cancer tissues from benign tissues on several benchmark datasets, as it takes into account the redundancy among the genes during their selection. The method selected a less number of genes compared to MRMR or SVM-RFE on most datasets. Gene ontology analyses revealed that the method selected genes that are relevant for distinguishing cancerous samples and have similar functional properties. The method provides a framework for combining filter methods and wrapper methods of gene selection, as illustrated with MRMR and SVM-RFE methods.**

***Index Terms*—Cancer classification, gene redundancy, gene relevancy, mutual information, support vector machine recursive feature elimination (SVM-RFE) .**

I. Introduction

THE ADVENT of DNA microarray technology has enabled simultaneous measurements of expressions of thousands of genes. However, due to high cost of experiments, sample sizes of gene expression measurements remain in hundreds, compared to tens of thousands of genes involved. As there are only a few samples (observations) compared to the genes (features), the extraction of useful information from microarray data is hindered by the curse of input dimensionality as well as by the computational instabilities. Therefore, selection of relevant genes remains a challenge in the analysis of gene expression data [1]. In this paper, we address the problem of distinguishing cancer samples from benign samples by collecting gene expressions by microarrays.

The genes or input features to a classifier can be broadly categorized into two types: relevant or redundant. The relevancy of a gene is measured with respect to the output class labels and relates to the importance of the gene for the classification task [2]. It is usually measured by the mutual information or correlation between gene expressions and class labels. Highly correlated genes tend to deteriorate the performance and become redundant for classification [3]. Mutual information or correlation among genes is often used to measure the redundancy of a set of genes. Usually, optimal classification accuracy is achieved by a set of maximally relevant and minimally redundant genes.

Gene selection has recently attracted many scientists in functional genomics and numerous algorithms have, therefore, resulted [4]–[13]. The aim of gene selection is to select a small subset of genes from a larger pool, rendering not only a good performance of classification, but also biologically meaningful insights. Gene selection methods are classified into two types: filter methods and wrapper methods [14]. Filter methods evaluate a gene subset by looking at the intrinsic characteristics of data with respect to class labels [4], while wrapper methods evaluate the goodness of a gene subset by the accuracy of its learning or classification method. Wrapper methods of gene selection are embedded in the classification process—so better in principle—but are more complex and could be computationally expensive.

Several algorithms have been developed to maximize the relevancy of a gene subset while minimizing the redundancy among the genes [5]–[9], [13], for example, MRMR criterion based on mutual information [6], [7]. Though a good gene subset should contain genes that are highly relevant and nonredundant, weakly relevant (but nonredundant) genes help the correlation-based feature selection algorithms [2], [3] and a tradeoff between relevancy and redundancy of genes may be useful for classification [15]. The MRMR method does not allow a tradeoff between relevancy and redundancy of genes. Greedy algorithms and simulated annealing have been attempted to determine the optimal tradeoff between the relevancy and the redundancy of a set of genes [5], [16]. In another study, the relevancy–redundancy criterion was attempted in two stages [9]: using Wilcoxon test or *F*-test, the relevant gene set was obtained from original microarray dataset, and subsequently, redundant genes were removed from the selected gene set by controlling the upper bound of Bayes error. Ooi *et al.* in [8] studied the tradeoff between relevancy and redundancy in multiclass gene selection problem by introducing a data-dependent tuning parameter called differential degree of prioritization. Recently, ReliefF and MRMR algorithms were combined in a two-stage strategy for large-scale gene selection [13]. In the first stage, a small subset of genes was selected using ReliefF, and then, MRMR method was applied to select nonredundant genes into the subset. All the aforementioned methods are filter approaches and do not incorporate classifier operation into the gene selection process.

Wrapper approaches usually achieve higher classification accuracies by embedding classifier characteristics into the gene selection process [10], [11], [17]–[25]. Least-square (LS) SVM-based leave-one-out sequential forward-selection algorithm was proposed for gene selection [11], [23]. Recursive cluster-elimination-based approach has been introduced to rank gene clusters in classification [10]. Using the principles of bagging and random gene selection for tree building, a random forest-based approach was proposed to measure the importance of a gene in classification [18]. Wahde and Szallasi

Manuscript received July 14, 2008. First published October 30, 2009; current version published March 24, 2010. *Asterisk indicates corresponding author.*

P. A. Mundra is with the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, 637553 Singapore.

*J. C. Rajapakse is with the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, 637553 Singapore, and also with the Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139-4307 USA (e-mail: asjagath@ntu.edu.sg).

Digital Object Identifier 10.1109/TNB.2009.2035284

1536-1241/$26.00 © 2009 IEEE

reviewed evolutionary-algorithms-based wrapper methods, in which, gene selection is achieved by optimizing a selection criteria by using genetic operations [17]. Though the aforementioned methods seem to outperform filter methods, each has its own pros and cons, and most suffer from high computational cost and instabilities.

Support vector machine recursive feature elimination (SVM-RFE) approach for gene selection [21] has recently attracted many researchers. SVM-RFE is a multivariate gene ranking method that uses SVM weights as the ranking criterion of genes. Earlier, we introduced multiple SVM-RFE (MSVM-RFE), where the SVM was trained on multiple subsets of data and the genes were ranked using statistical analysis of gene weights in multiple runs, and demonstrated its applications in selecting genes in microarray data [22] SVM-RFE was also applied to identify peaks in mass spectrometry data [26]. The performance of SVM-RFE becomes unstable at some values of the gene filter-out factor, i.e., the number of genes eliminated in each iteration. To overcome this, two-stage SVM-RFE algorithm was proposed [24], where initial gene subset was selected using several MSVM-RFE models with different gene filter-out factors, and in the second stage, genes were selected by eliminating one gene at each iteration. A fuzzy granular SVM-RFE algorithm was recently proposed by incorporating statistical learning, fuzzy clustering, and granular computing to select highly informative genes [25].

In wrapper methods, the classifier characteristics such as SVM weights in SVM-RFE provide a criterion to rank genes based on their relevancy, but they do not account for the redundancy among the genes [27]. Our aim is to combine classifier characteristics with a filter criterion that could minimize the redundancy among selected genes, resulting in a selection of a small subset of genes and improved classification accuracy. In this paper, we propose an approach that incorporate mutual-information-based MRMR filter in SVM-RFE to minimize the redundancy among the selected genes. As seen later, our approach, referred to as *SVM-RFE with MRMR filter*, improved the accuracy of classification and yielded smaller gene sets on several benchmark cancer gene expression datasets. Experiments show that our method outperforms MRMR and SVM-RFE methods, as well as other popular methods on most datasets, and selects genes that are biologically relevant in discriminating cancer samples and have properties belonging to the same pathway.

The manuscript is organized as follows. Section II describes MRMR and SVM-RFE methods, and gives a detailed description of the proposed algorithm: SVM-RFE with MRMR Filter. The numerical experiments on four gene expression datasets are demonstrated in Section III. The performance of our algorithm and comparison with earlier approaches are presented. Section IV concludes the manuscript with a discussion.

II. Methods

A. Minimum Redundancy–Maximum Relevancy (MRMR)

The MRMR method aims at selecting maximally relevant and minimally redundant set of genes for discriminating tissue classes. In this paper, we use mutual-information-based MRMR criterion to find a maximally relevant and minimally redundant set of genes.

Let $D = \{x_{i,k}\}_{n \times K}$ denote a microarray gene expression data matrix, where $x_{i,k}$ is the expression of gene i in sample k, n denotes the number of genes measured by the microarray, and K denotes the number of samples. Let $x_k = (x_{1,k}, x_{2,k}, \ldots, x_{n,k})$ denote the kth sample of gene expressions and $x_{i\cdot} = (x_{i,1}, x_{i,2}, \ldots, x_{i,K})$ denote the gene expressions of ith gene across samples. Let $G = \{1, 2, \ldots, n\}$ be the indexed set representing the genes. In this paper, we address two class classification of tissue samples into cancer or benign tissues. Let the target class label of sample k be $y_k = \ell \in \{+1, -1\}$, taking values $+1$ or -1 for benign or cancerous tissues, respectively. The mutual information between class labels ℓ and gene i will quantify the relevancy of gene i for the classification. The relevancy R_S of genes in a subset $S \subset G$ is given by

$$R_S = \frac{1}{|S|} \sum_{\ell} \sum_{i \in S} I(\ell, i) \tag{1}$$

where $I(\ell, i) = \sum_{x_{i\cdot}} p(\ell, x_{i\cdot}) \log \frac{p(\ell, x_{i\cdot})}{p(\ell) p(x_{i\cdot})}$ is the mutual information between class labels ℓ and gene i, where the summation is taken over the space of gene expression values. The redundancy of a gene subset is determined by the mutual information among the genes. The redundancy of gene i with the other genes in the subset S is given by

$$Q_{S,i} = \frac{1}{|S|^2} \sum_{i' \in S, i' \neq i} I(i, i'). \tag{2}$$

In MRMR method, gene ranking is performed by optimizing the ratio of the relevancy of a gene to the redundancy of the genes in the set. The maximally relevant and minimally redundant gene i^* in the set S is given by

$$i^* = \arg\max_{i \in S} \frac{R_S}{Q_{S,i}}. \tag{3}$$

The relevancy and redundancy measures of genes can be combined in many ways, but the quotient in (3) has been found to select highly relevant genes with least redundancy [6]. The effectiveness of this ratio is reflected by its consistently good performance shown on various expression datasets than by other criterion, such as the difference between relevancy and redundancy [6], [7]. After selecting the top-ranked genes, the subsequent genes are selected by forward selection, maximizing the criterion given in (3).

B. Support Vector Machine Recursive Feature Elimination (SVM-RFE)

SVM-RFE was introduced by Guyon *et. al.,* for ranking genes from gene expression data for cancer classification [21]. It is now being widely used for gene selection and several improvements have been recently suggested [22], [24], [25]. SVM-RFE, starting with all the genes, removes the gene that is least significant for classification recursively in a backward elimination manner. The ranking score is given by the components of the weight vector w of the SVM as follows:

$$w = \sum_k \alpha_k y_k x_k \tag{4}$$

where $y_k \in \ell$ is the class label of the sample x_k and the summation is taken over all the training samples. α_k is the Lagrange multipliers involved in maximizing the margin of separation of the classes [28]. If w_i denotes the component weight connecting to the gene i, ${w_i}^2$ gives a measure the ranking of the gene i based on its effect on the margin of separation upon removal [21], [22]. For computational efficiency, more than one gene can be removed at a single step, though it may have negative effect on selection of genes when the set of genes is small [21].

C. SVM-RFE with MRMR Filter

The MRMR filter, when used alone, may not yield optimal accuracy because the classifier performs independently and is not involved in the selection of genes. On the other hand, SVM-RFE does not take into account the redundancy among genes. Our aim is to improve the gene selection in SVM-RFE by introducing an MRMR filter to minimize the redundancy among relevant genes. As seen later, this improves the classification accuracy by compromising relevancy and redundancy of genes relating to cancer.

In our approach of SVM-RFE with MRMR filter, the genes are ranked by a convex combination of the relevancy given by SVM weights and the MRMR criterion. For ith gene, the ranking measure r_i is given by

$$r_i = \beta\,|w_i| + (1-\beta)\frac{R_{S,i}}{Q_{S,i}} \tag{5}$$

where the parameter $\beta \in [0,1]$ determines the tradeoff between SVM ranking and MRMR ranking, and the relevancy $R_{S,i}$ of gene i in the set S on classification is given by

$$R_{S,i} = \frac{1}{|S|}\sum_{\ell} I(\ell, i) \qquad \forall i \in S. \tag{6}$$

To facilitate the backward selection, we use gene-wise MRMR criterion for ranking in the present method. Also, we use $|w_i|$ as the gene relevancy measure from SVM to better compromise with redundancy of genes. Algorithm 1 illustrates an iteration of SVM-RFE with MRMR filter method of ranking genes: the least important gene at a time is identified after ranking the genes in the subset $S \subset G$. In each iteration, one (or more) of least significant genes are removed and the remaining subset will go through the removal process iteratively, until the removal of any more genes does not improve the classifier performance.

D. Gene Ontology (GO)-Based Gene Similarity

GO (www.geneontology.org) provides a structured and controlled vocabulary to annotate gene and gene products of an organism. The functions of genes are distributed in terms of three hierarchies (or taxonomies): molecular function (MF), biological process (BP), and cellular compartment (CC). The interrelationships among the terms are represented in a directed acyclic graphs (DAGs), where each node of the graph represents a term and an edge represents *is-a* or *part-of* relation between two terms. GO helps to evaluate functional similarities among a set of genes [22], [29], as there exists a direct relationship between semantic similarity of gene pairs and their structural- and sequence-based similarity [30], [31]. The semantic similarity represents the biological similarity among a set of genes. In order to find biological relevancy and redundancy of genes selected by our method, we investigate into GO-based biological semantics.

Algorithm 1 : SVM-RFE with MRMR Filter for ranking genes

```
begin
Set $\beta$
Given set of genes, $S \subset G$
Ranked set of genes, R = { }
repeat
  Train linear SVM with gene set $S$
  Calculate the weight of each gene $w_i$
  for each gene $i \in S$ do
        Compute $R_{S,i}$ and $Q_{S,i}$
        Compute $r_i$
  end for
  Select the gene with smallest ranking score, $i^* = \arg\min\{r_i\}$
  Update $R = R \cup \{i^*\}$ ; $S = S \setminus \{i^*\}$ ;
until all genes are ranked
end : output $R$
```

The concept of semantic similarity among genes is derived from how GO terms describing the gene function relate to one another. We use Lin's semantic similarity measure that estimates similarity on the basis of parent commonality of two query terms and incorporates the information content of the query terms [32]. For each term t, let $p(t)$ be the probability of finding a child of t in the annotation database. If t_i and t_j are two query terms and $a\,(t_i, t_j)$ represents the set of parent terms shared by both t_i and t_j, then the similarity is given by

$$\rho\,(t_i, t_j) = \frac{2 \times \max_{t \in a(t_i,t_j)} \log\,(p\,(t))}{\log\,(p\,(t_i)) + \log\,(p\,(t_j))}. \tag{7}$$

To compute semantic similarity between two genes i and j, let T_i and T_j be the sets of terms annotating the two genes, respectively. The semantic similarity between the two genes is defined as the average interset similarity of terms in sets T_i and T_j. The biological similarity of the genes in the set S is given by

$$\rho\,(S) = \frac{2}{|S|\,(|S|-1)} \sum_{i,j \in S, i \neq j} \frac{1}{|T_i|\,|T_j|} \sum_{t_i \in T_i, t_j \in T_j} \rho\,(t_i, t_j)\,. \tag{8}$$

III. Experiments and Results

A. Data

We evaluated the performance of the proposed method on four microarray gene expression cancer datasets, namely, colon [33], leukemia [4], hepato [34], and prostate [35]. These datasets have

TABLE I
NUMBER OF GENES AND SAMPLES IN GENE EXPRESSION DATASETS

Dataset	*Genes*	**Training**	**Testing**
Colon	2000	40	22
Leukemia	7129	38	34
Hepato	7129	33	27
Prostate	12600	102	34

been widely used to benchmark gene selection algorithms. In colon cancer dataset, no separate testing set is available, and therefore, we randomly divided the original dataset into separate training and testing datasets. The numbers of samples and genes in the datasets are given in Table I.

B. Preprocessing

Datasets were normalized to zero mean and unit variance, based on gene expressions of a particular sample. The gene expression values were directly used as input features to SVM-RFE, but were discretized in order to compute mutual information to evaluate MRMR values. The discretization of expression values were done to represent over, no, and underexpression of the gene. The discretization $\tilde{x}$ of the gene expression x was obtained as follows:

$$\tilde{x} = \begin{cases} +2, & \text{if } x > \mu + \sigma/2 \\ -2, & \text{if } x < \mu - \sigma/2 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where the discretized value $\tilde{x} = +2, 0,$ or -2 represents an overexpression, no expression, and an underexpression of the gene, respectively. μ and σ denote the mean and standard deviation of the gene expressions.

C. Parameter Estimation

In each iteration of Algorithm 1, we trained the linear SVM model with the selected subset S. The linear SVM sensitivity parameter η was estimated from the set of $\{2^{-20}, 2^{-19}, \ldots, 2^{0}, \ldots, 2^{15}\}$, giving the maximum Matthew's correlation coefficient (MCC[1]) on tenfold cross validation on training data. MCC was chosen because of the small sample sizes and the imbalances of labels in most datasets. The value of η was used with linear SVM model to select genes in SVM-RFE and SVM-RFE with MRMR filter methods. The β value for the SVM-RFE with filter method was then determined empirically from the set $\{0.2, 0.4, 0.5, 0.6, 0.8\}$ based on the best tenfold cross-validation performance.

D. Performance Evaluation and Implementation

Starting with all the genes measured, we used Algorithm 1 to gradually remove genes to select the most relevant and minimally redundant subset of genes. To increase the speed of gene selection with both SVM-RFE and SVM-RFE with MRMR filter methods, we eliminated 100 genes in one iteration, when the number of genes in the gene subset was equal to or greater than 10 000, 10 genes if the number of genes was less than 10 000, but greater than 1000, and one gene at a time if the set contained less than 1000 genes.

After parameter estimation with tenfold cross validation, genes were selected using optimal parameters on training data. Using the selected genes, the performance of the methods were evaluated on test samples. Because of the small number of test samples and imbalances of training and testing sets, the test errors were evaluated on bootstrap samples: training and testing datasets were merged and all samples were then partitioned for training and testing sets for 100 times. The performance measures such as accuracy, sensitivity, and specificity were averaged over 100 trials. The genes were selected from the gene subset giving the least-average test error. MCC was also used to evaluate the performances on test data. We performed pairwise *t*-test to determine whether there is a statistical significance of the differences of performances of the present method over the other methods. In addition to MRMR and SVM-RFE methods, we compared our method with several other widely used methods: Bayes-error-based filter with k-nearest neighbor (KNN) and SVM [9], *t*-test with Fisher discriminant analysis [36], and LS-bound with SVM [23].

We used MATLAB toolbox from the original authors for experiments with MRMR method [6]. The classification accuracy of the gene subset was evaluated using SVM. We also compared our results with the SVM-RFE method. For gene selection and testing with SVM, we used LIBSVM—version 2.84 software [37]. For implementation of LS-bound SVM method, the original source codes (available in [23, supplementary material]) were utilized. For GO analysis, we used *GOSim* [38] package developed in *R* to compute the Lin's similarity measures. Gene Entrez Ids were obtained from National Center for Biotechnology Information (NCBI) database. The genes without Human Genome Organization (HUGO) symbols were excluded from GO analysis.

E. Results

Table II gives a comparison of classification performances of MRMR, SVM-RFE, and SVM-RFE with filter methods on four datasets. As seen, the performance of SVM-RFE with MRMR filter was significantly better in most of the performance measures on leukemia, hepato, and prostate datasets. SVM with MRMR filter showed significant improvement in the sensitivity against SVM-RFE method on colon data. The numbers of genes selected by SVM-RFE with filter were lower than those selected by MRMR or SVM-RFE methods except that for prostate cancer, the present method selected one more gene than MRMR. Table III shows a comparison of performances of the present method with other existing methods for gene selection. As seen, SVM-RFE with MRMR filter gave higher classification accuracies on Colon, Leukemia and Prostate cancer datasets, which were statistically significant. In hepato dataset, there was no statistical difference between the accuracies between Bayes followed by SVM and the present method. However, the numbers of genes selected by the present method were relatively higher compared to the other methods.

[1] MCC $= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

TABLE II
PERFORMANCE OF MRMR, SVM-RFE, AND SVM-RFE WITH MRMR FILTER AND STATISTICAL SIGNIFICANCE OF DIFFERENCES OF PERFORMANCE OF SVM-RFE WITH FILTER

Dataset	**Measure**	**MRMR**	**SVM-RFE**	**SVM-RFE with MRMR**	**Significance of Difference**	
					with MRMR	with SVM-RFE
Colon	# Genes	90	90	78		
	Accuracy	91.00 ± 5.85	91.00 ± 5.17	91.68 ± 5.13	...	...
	Sensitivity	87.38 ± 11.72	86.75 ± 10.18	89.50 ± 9.18	...	$p < 0.05$
	Specificity	93.07 ± 6.05	93.43 ± 5.53	92.93 ± 5.79	...	...
	MCC	0.81 ± 0.13	0.81 ± 0.11	0.83 ± 0.11	...	...
Leukemia	# Genes	91	47	37		
	Accuracy	97.18 ± 2.86	97.88 ± 2.07	98.35 ± 1.93	$p < 0.01$	$p < 0.05$
	Sensitivity	93.36 ± 6.98	95.00 ± 5.13	96.36 ± 4.37	$p < 0.001$	$p < 0.05$
	Specificity	99.85 ± 0.86	99.90 ± 0.70	99.75 ± 1.31	...	...
	MCC	0.94 ± 0.06	0.96 ± 0.04	0.97 ± 0.04	$p < 0.01$	$p < 0.05$
Hepato	# Genes	63	79	11		
	Accuracy	84.78 ± 5.38	89.19 ± 5.88	88.15 ± 4.59	$p < 0.001$	...
	Sensitivity	73.00 ± 19.96	80.25 ± 12.20	88.12 ± 10.26	$p < 0.001$	$p < 0.001$
	Specificity	89.74 ± 6.91	92.95 ± 6.95	88.16 ± 5.72	...	$p < 0.001$
	MCC	0.64 ± 0.13	0.75 ± 0.13	0.74 ± 0.10	$p < 0.001$	...
Prostate	# Genes	9	85	10		
	Accuracy	96.47 ± 3.07	96.24 ± 3.37	98.29 ± 2.30	$p < 0.001$	$p < 0.001$
	Sensitivity	96.08 ± 3.68	95.88 ± 4.08	98.16 ± 2.69	$p < 0.001$	$p < 0.001$
	Specificity	97.56 ± 5.6	97.22 ± 5.56	98.67 ± 3.96	...	$p < 0.05$
	MCC	0.92 ± 0.07	0.91 ± 0.08	0.96 ± 0.05	$p < 0.001$	$p < 0.001$

TABLE III
COMPARISON OF GENES SELECTED, ACCURACIES, AND SIGNIFICANCE OF DIFFERENCE OF PERFORMANCES WITH OTHER METHODS

Dataset	**Measure**	**Bayes + KNN**	**Bayes + SVM**	**t-test + FDA**	**LS-Bound + SVM**	**SVM-RFE with MRMR**
Colon	# Genes	10	10	**3**	18	78
	Accuracy	88.23	86.27	82.68	85.23	**91.68**
	Significance	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	...
Leukemia	# Genes	7	8	**2**	5	37
	Accuracy	95.71	97.12	90.68	94.74	**98.35**
	Significance	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	...
Hepato	# Genes	9	10	**3**	4	11
	Accuracy	86.00	**88.96**	73.00	84.74	88.12
	Significance	$p < 0.01$	...	$p < 0.001$	$p < 0.001$	...
Prostate	# Genes	16	**3**	11	4	10
	Accuracy	97.29	95.41	87.29	96.06	**98.29**
	Significance	$p < 0.01$	$p < 0.001$	$p < 0.001$	$p < 0.001$	...

TABLE IV
RELEVANCY AND GO ANALYSIS OF SELECTED GENE SUBSETS

Dataset	**Method**	**Selected Genes**	**Relevancy**	**BP**	**MF**	**CC**
Colon	MRMR	90	0.27	0.40	0.38	0.55
	SVM-RFE	90	0.21	0.41	0.38	0.56
	SVM-RFE with MRMR	78	**0.28**	0.41	0.38	0.56
Leukemia	MRMR	91	0.39	0.42	0.37	0.59
	SVM-RFE	47	0.38	0.42	0.38	0.60
	SVM-RFE with MRMR	37	**0.40**	0.42	0.38	0.61
Hepato	MRMR	63	**0.28**	0.40	0.36	0.56
	SVM-RFE	79	0.20	0.41	0.35	0.58
	SVM-RFE with MRMR	11	**0.28**	0.48	0.35	0.63
Prostate	MRMR	9	0.28	0.41	0.32	0.57
	SVM-RFE	85	0.22	0.41	0.37	0.57
	SVM-RFE with MRMR	10	**0.30**	0.50	0.36	0.61

Table IV shows the mutual-information-based relevancies of selected gene subsets by different methods on four datasets. For all four datasets, the present method selected comparable or slightly more relevant genes compared to other methods, but the subtle differences were statistically insignificant. To evaluate biological similarity among genes, we performed GO-based similarity analysis with Lin's similarity measure on the selected gene sets. Table IV gives GO-based similarities for best gene subsets selected by different methods. There were no statistically significant differences among GO measures of genes selected by different methods. This indicates that the present method selected functionally similar genes to those selected by other methods. Except for the MRMR-selected genes on Prostate data, the present method always selected a smaller subset of genes, yet functionally similar and coherent to those sets selected by other methods.

IV. Discussion

The redundancy among genes expressions in microarray data hinders the identification of cancerous tissues from benign tissues. In order to remove the redundancy or collinearity among genes, we introduced MRMR filter in the ranking of genes by the SVM-RFE method. The efficacy of embedding of MRMR filter in SVM-RFE was evidenced by improved classification performance on benchmark datasets. It enhanced the gene selection based on SVM weights. On most datasets tested, the proposed approach outperformed other methods in the classification because it was able to reduce the redundancy among the selected genes. Furthermore, the present method selected a less number of genes compared to MRMR and SVM-RFE methods on most datasets. We also observed improvements of standard deviations of performance measures (accuracy, sensitivity, specificity, and MCC) over bootstrap samples by the proposed method in the experiments, indicating more stability in gene selection than the other methods.

The proposed algorithm is computationally more expensive than SVM-RFE or MRMR methods. By removing a set of genes instead of one gene at a time, the present algorithm can be expedited. The use of discretized data for computation of mutual information could lead to a loss of information, but might improve the robustness to noise. But other methods which discretize the data optimally into a number of levels could further improve the present method. Parameter β and the sensitivity η of the SVM were empirically determined. Investigation into other methods that could estimate both parameters together may yield better estimates of parameters.

It is not advisable to determine relevant and redundant genes independent of the classifier, for example, using only MRMR criterion. On the other hand, a simple filter might improve the selection of genes if it understands how the filtered genes are used by the classifier. We proposed to incorporate MRMR criterion into the ranking scheme of SVM-RFE, so that the filter takes care of the redundancies among genes. This work can be extended to investigate into other filter criteria. Similarly important is to investigate the enhancement of introducing filters, such as MRMR, into the recent development of SVM-RFE, such as MSVM-RFE, two-stage SVM-RFE, and fuzzy-granular SVM-RFE. However, the scope of this paper is to demonstrate how the incorporation of redundancy, such as by MRMR filter, into a classifier-based gene selection method could improve the process of gene selection.

Functional similarity exists among the genes involved in a specific BP, but usually hinders computational methods of tissue classification. The GO analysis shows that functional similarities among the genes in terms of biological pathways, MFs, and colocalization of the top-ranked genes by the present method were comparable to those selected by MRMR or SVM-RFE methods. However, SVM-RFE with MRMR filter selects a fewer number of genes and attempts to eliminate redundant genes some of which may have biological functions important to cancer. Therefore, once a good classification is achieved, the original gene set need to be scanned to find genes that are biologically similar to those selected. Ideally, one could integrate the biological measures of relevancy and redundancy, using information from GO, into the gene selection process, so that the results are directed by biological underpinnings. However, this is not feasible at this time, as many genes measured by microarrays do not have entries in the GO database.

Acknowledgment

The authors would like to thank researchers who made the datasets and source codes used in this research publicly available. Special thanks goes to Dr. K. Duan for his inputs in the early part of this research.

References

[1] A. Blum and A. Langley, “Selection of relevant features and examples in machine learning,” *Artif. Intell.*, vol. 97, pp. 245–271, 1997.

[2] R. Kohavi and G. John, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, pp. 273–324, 1997.

[3] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[4] 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.

[5] X. Liu, A. Krishnan, and A. Mondry, “An entropy-based gene selection method for cancer classification using microarray data,,” *BMC Bioinformatics*, vol. 6, p. 76, 2005.

[6] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *J. Bioinformatics Comput. Biol.*, vol. 3, pp. 185–205, 2005.

[7] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[8] C. Ooi, M. Chetty, and S. Teng, “Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data,” *BMC Bioinformatics*, vol. 7, pp. 320–339, 2006.

[9] J. Zhang and H. Deng, “Gene selection for classification of microarray data based on Bayes error,” *BMC Bioinformatics*, vol. 8, p. 370, 2007.

[10] M. Yousef, S. Jung, L. Showe, and M. Showe, “Recursive cluster elimination (rce) for classification and feature selection from gene expression data,” *BMC Bioinformatics*, vol. 8, p. 144, 2007.

[11] E. Tang, P. Suganthan, and X. Yao, “Gene selection algorithms for microarray data based on least square support vector machine,” *BMC Bioinformatics*, vol. 7, p. 95, 2006.

[12] S. Baker and S. Kramer, “Identifying genes that contributes most to good classification in microarray,” *BMC Bioinformatics*, vol. 7, p. 407, 2006.

[13] Y. Zhang, C. Ding, and T. Li, “Gene selection algorithm by combining reliefF and mRMR,” *BMC Genomics*, vol. 9, no. 2, p. S27, 2008.

[14] I. Inza, P. Larranaga, R. Blanco, and A. Cerrolaza, “Filter versus wrapper gene selection approaches in DNA microarray domains,” *Artif. Intell. Med.*, vol. 31, pp. 91–103, 2004.

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

[16] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[17] M. Wahde and Z. Szallasi, “A survey of methods for classification of gene-expression data using evolutionary algorithm,” *Expert Rev. Molecular Diagnostics*, vol. 6, no. 1, pp. 101–110, 2006.

[18] R. Diaz-Uriarte and S. Andres, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, p. 3, 2006.

[19] R. Ruiz, J. Riquelme, and J. Aguilar-Ruiz, “Incremental wrapper-based gene selection from microarray data for cancer classification,” *Pattern Recog.*, vol. 39, pp. 2383–2392, 2006.

[20] A. Rakotomamonjy, “Variable selection using SVM based criteria,” *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, 2003.

[21] I. Guyon, J. Weston, S. Barhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, vol. 46, pp. 389–422, 2002.

[22] D. Kai-Bo, J. Rajapakse, H. Wang, and F. Azuaje, “Multiple SVM-RFE for gene selection in cancer classification with expression data,” *IEEE Trans. Nanobiosci.*, vol. 4, no. 3, pp. 228–234, Sep. 2005.

[23] X. Zhou and K. Mao, “LS bound based gene selection for DNA microarray data,” *Bioinformatics*, vol. 21, no. 8, pp. 1559–1564, 2005.

[24] Y. Tang, Y.-Q. Zhang, and Z. Huang, “Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis,” *IEEE Trans. Comput. Biol. Bioinformatics*, vol. 4, no. 3, pp. 365–381, Jul.–Sep. 2007.

[25] Y. Tang, Y.-Q. Zhang, Z. Huang, X. Hu, and Y. Zhao, “Recursive fuzzy granulation for gene subset extraction and cancer classification,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 6, pp. 723–730, Nov. 2008.

[26] J. Rajapakse, D. Kai-Bo, and W. Yeo, “Proteomic cancer classification with mass spectrometry data,” *Amer. J. Pharmacogenom.*, vol. 5, pp. 281–292, 2005.

[27] Z.-X. Xie, Q.-H. Hu, and D.-R. Yu, “Improved feature selection algorithm based on SVM and correlation,” in *Int. Symp. Neural Netw.,* (LNBI Series 3971), J. Wang, Z. Yi, J. M. Zurada, B.-L. Lu, and Y. Hujun, Eds., 2006, pp. 1373–1380.

[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[29] H. Wang, F. Azuaze, O. Bodenreider, and J. Dopazo, “Gene expression correlation and gene ontology based similarity: An assessment of quantitative relationships,” in *Proc. IEEE Symp. Comput. Intell. Bioinformatics Comput. Biol.*, 2004, pp. 25–31.

[30] P. Lord, R. Stevens, A. Brass, and A. Goble, “Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation,” *Bioinformatics*, vol. 19, pp. 1275–1283, 2003.

[31] J. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. Mato, L. Martinez-Cruz, F. Corrales, and A. Rubio, “Correlation between gene expression and go semantic similarity,” *IEEE Trans. Comput. Biol. Bioinformatics*, vol. 2, no. 4, pp. 330–338, Oct.–Dec. 2005.

[32] D. Lin, “An information-theoretic definition of similarity,” in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 296–304.

[33] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” *Proc. Nat. Acad. Sci. USA*, vol. 96, pp. 6745–6750, 1999.

[34] N. Iizuka, M. Oka, H. Yamada Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, K. Hamada, H. Nakayama, K. Ishitsuka, T. Miyamoto, A. Hirabayashi, S. Uchimura, and Y. Hamamoto, “Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection,” *Lancet*, vol. 361, pp. 923–929, 2003.

[35] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D’Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers, “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, vol. 1, pp. 203–209, 2002.

[36] C. Lai, M. Reinders, L. van’t Veer, and L. Wessels, “A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets,” *BMC Bioinformatics*, vol. 7, p. 235, 2006.

[37] C. Chang and C. Lin. (2001). Libsvm: A library for support vector machines [Online]. Available: www.csie.ntu.edu.tw/ cjlin/libsvm.

[38] H. Frohlich, N. Speer, A. Poustka, and T. Beiβbarth, “Gosim—An R-package for computation of information theoretic go similarities between terms and gene products,” *BMC Bioinformatics*, vol. 8, p. 166, 2007.

Piyushkumar A. Mundra received the B.E. degree in chemical enginnering from Nirma Institute of Technology, Ahmedabad, India, and the M.Tech. degree in bioprocess technology from the Institute of Chemical Technology, previously known as University Institute of Chemical Technology, Mumbai, India. He is currently working toward the Ph.D. degree at the School of Computer Engineering, Nanyang Technological University, Singapore.

For more than a year, he was a Research Project Assistant in the National Chemical Laboratory, Pune, India. His current research interests include feature selection, classification algorithms, genetic algorithms, and gene regulatory networks.

Jagath C. Rajapakse (S’90–M’91–SM’00) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Buffalo, Buffalo, NY.

He was a Visiting Scientist at the Max-Planck-Institute of Cognitive and Brain Sciences, Leipzig, Germany and a Visiting Fellow at the National Institute of Mental Health, Bethesda, MD. He is currently a Professor of computer engineering and the Director of the BioInformatics Research Centre, Nanyang Technological University, Singapore. He is also a Visiting Professor in the Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge. His research interests include neuroinformatics, bioinformatics, modeling brain connectivity through functional brain imaging, and building pathways from gene and protein expressions obtained by microarrays, and high-content microscopic imaging. He has authored or coauthored more than 225 peer-reviewed international conference and journal papers, and book chapters. Prof. Rajapakse was an Associate Editor of the IEEE Transactions on Medical Imaging, the IEEE Transactions on Medical Imaging, and the IEEE Transactions on Neural Networks, and in editorial boards of several other journals.