Gene expression

MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data

Xin Zhou and David P. Tuck*
Department of Pathology, Yale University School of Medicine, New Haven, Connecticut 06510, USA

Received on December 7, 2006; revised on January 24, 2007; accepted on January 26, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Given the thousands of genes and the small number of samples, gene selection has emerged as an important research problem in microarray data analysis. Support Vector Machine—Recursive Feature Elimination (SVM-RFE) is one of a group of recently described algorithms which represent the stat-of-the-art for gene selection. Just like SVM itself, SVM-RFE was originally designed to solve binary gene selection problems. Several groups have extended SVM-RFE to solve multiclass problems using one-versus-all techniques. However, the genes selected from one binary gene selection problem may reduce the classification performance in other binary problems.

Results: In the present study, we propose a family of four extensions to SVM-RFE (called MSVM-RFE) to solve the multiclass gene selection problem, based on different frameworks of multiclass SVMs. By simultaneously considering all classes during the gene selection stages, our proposed extensions identify genes leading to more accurate classification.

Contact: david.tuck@yale.edu

Supplementary information: Supplementary materials, including a detailed review of both binary and multiclass SVMs, and complete experimental results, are available at Bioinformatics online.

1 INTRODUCTION

Microarray technology allows researchers to measure the quantity of mRNA for tens of thousands of genes simultaneously. It has the power to create a comprehensive overview of the gene regulatory network. Studies on DNA microarray data offer opportunities for advancing fundamental biological research and clinical practice (Golub et al., 1999; Ramaswamy et al., 2001; Ross et al., 2000; Staunton et al., 2001; West et al., 2001).

Normally, a microarray data set contains a large number of genes (usually several thousand or more) and a relatively small number of samples or conditions (usually < 100). Among all the genes, many are irrelevant, insignificant or redundant to the discriminant problem under investigation. Hence the identification of informative genes, which have the greatest power for classification, is of fundamental and practical importance to the investigation of specific discriminant problems, such as cancer versus non-cancer (Alon et al., 1999) or different tumor types (Golub et al., 1999; Ramaswamy et al., 2001).

Many gene selection algorithms have been proposed in the context of microarray data analysis over the past few years. However, most of the studies are related to binary (two-class) gene selection problems (Golub et al., 1999; Guyon et al., 2002; Lee et al., 2003; Zhou and Mao, 2005a,b; Zhang et al., 2006, to name a few), and only a few involve multiclass (i.e. more than two classes) gene selection and classification (Chai and Domeniconi, 2004; Li et al., 2004; Ramaswamy et al., 2001; Tibshirani et al., 2002). As the multiclass problem is intrinsically more difficult and presents more challenges, it is worthy of further investigation.

The well-known SVM-RFE (Support Vector Machine—Recursive Feature Elimination) (Guyon et al., 2002) is a simple and efficient algorithm which conducts gene selection in a backward elimination procedure. It has been widely applied in analyzing high-dimensional biological data, such as gene expression data (Frank et al., 2006; Ramaswamy et al., 2001; Zheng et al., 2006), sequence analysis (Das et al., 2006), and protein mass spectra data (Hilario et al., 2006). Just as SVM, SVM-RFE was initially proposed for binary cases. Some researchers (Chai and Domeniconi, 2004; Ramaswamy et al., 2001) employed a simple one-versus-all technique to apply SVM-RFE to solve multiclass problems. However, the genes selected from one binary gene selection problem may reduce the classification performance in other binary problems.

In the present study, we propose four extensions of SVM-RFE to solve the multiclass gene selection problem. By taking all classes simultaneously into consideration during the gene selection stage, our proposed extensions provide genes leading to more accurate classification performance.

The paper is organized as follows. The binary and multiclass SVMs, as well as SVM-RFE, are first briefly reviewed in Section 2. In Section 3, we discuss the existing extension of SVM-RFE for multiclass problems, and propose four new MSVM-RFE algorithms based on different frameworks of multiclass SVMs, respectively, to overcome the shortcoming of the existing extension. The performance of MSVM-RFE algorithms is finally tested with six microarray data sets in Section 4.

*To whom correspondence should be addressed.
2 BACKGROUND
Consider \( n \) training data pairs: \( S = \{ (x_i, y_i) \}, i = 1, \ldots, n \), where \( x_i \in \mathbb{R}^p \) is a feature vector representing the \( i \)-th sample, and \( y_i \) is the class label of \( x_i \). For a binary problem, \( y_i \in \{-1, 1\} \), while for \( k \)-class \( (k > 2) \) problem, \( y_i \in \{1, 2, \ldots, k\} \).

2.1 Support vector machines
Support vector machines (SVMs) (Vapnik, 1998) are powerful classification algorithms that have shown state-of-the-art performance in a variety of biological classification tasks. SVMs are not very sensitive to the curse of dimensionality, and they are well-suited to work with high dimensional data, such as microarray gene expression data (Brown et al., 2000; Furey et al., 2000).

The first generation of SVMs was only designed for binary classification. However, most real-life diagnostic tasks are not binary, and solving multiclass classification problems is much harder than solving binary ones (Mukherjee, 2003; Rifkin et al., 2003). Fortunately, several algorithms have been proposed for extending binary SVMs to multiclass classification. These algorithms are grouped into two types. One is by constructing and combining several binary SVM classifiers. One-versus-all (OVA) and one-versus-one (OVO) SVMs (Kreßel, 1999) are two typical methods in this type. The other type, called ‘all-together’, is to directly solve one optimization problem which takes all classes into consideration (Crammer and Singer, 2001; Weston and Watkins, 1999; Lee et al., 2004). In this section, we will briefly introduce the binary and multiclass SVMs. A detailed review of both binary and multiclass SVMs, including their mathematical formulations, is presented in the Supplemental Material.

2.1.1 Binary SVMs
Intuitively, an SVM searches for a hyperplane with maximal distance between itself and the closest samples from each of two classes (Vapnik, 1998). The decision function of SVMs, just as other linear classifiers, is presented as \( f(x) = w^T x + b \), where \( w = [w_1, w_2, \ldots, w_p]^T \) is the weight vector and \( b \) is a scalar. The mechanism of SVMs is to minimize the following optimization problem:

\[
\min \quad \mathcal{P}(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i, \\
\text{subject to} \quad y_i[w^T x_i + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n, 
\]

where the predefined parameter \( C \) is a trade-off between training accuracy and generalization. Note that the SVM is originally designed for the binary classification problem, hence \( y_i \in \{-1, 1\} \) as we mentioned before. The solution of this optimization problem is given by solving the corresponding dual problem, a quadratic programming (QP) problem with \( n \) variables.

2.1.2 Multiclass SVMs: one-versus-all
This is the simplest and probably the earliest implementation for multiclass SVMs (Bottou et al., 1994). Here, \( k \) binary SVM classifiers are constructed: class 1 (positive) versus all other classes (negative), class 2 versus all other classes, \ldots, class \( k \) versus all other classes. The decision function of the \( r \)-th classifier \( (r = 1, \ldots, k) \), \( f_r(x) = w_r^T x + b_r \), where \( w_r = [w_{r1}, w_{r2}, \ldots, w_{rp}]^T \) is the \( r \)-th weight vector and \( b_r \) is a scalar, can be obtained by solving the following optimization problem,

\[
\min \quad \mathcal{P}(w_r, \xi_r) = \frac{1}{2} \|w_r\|^2 + C \sum_{j=1}^{n} \xi_{rj}, \\
\text{subject to} \quad z^r_i[w_r^T x_i + b_r] \geq 1 - \xi_{rj}, \quad \xi_{rj} \geq 0, \quad j = 1, \ldots, n, 
\]

where \( z^r_i \) is the class label for the \( r \)-th classifier, that is, \( z^r_i = 1 \) if \( y_i = r \) and \(-1 \) otherwise. Note that for multiclass cases \( y_i \in \{1, 2, \ldots, k\} \). When all the \( k \) classifiers are constructed, the combined \( k \) binary classifiers \( [f_1, f_2, \ldots, f_k]^T \) predict the class \( y \) of a (test) sample \( x \) that corresponds to the maximal value of \( k \) binary classifiers, that is,

\[
y = \arg \max_r f_r(x).
\]

Practically we solve the dual problem of (2) whose number of variables is the same as the number of data, so \( k \) QP problems with \( n \) variables are solved in this implementation of multiclass SVMs.

2.1.3 Multiclass SVMs: one-versus-one
Another major method based on multiple binary SVMs is called the OVO method (Kreßel, 1999). This method involves the construction of \( k(k - 1)/2 \) SVM classifiers where each one is trained on data from two classes. In total we need to solve \( k(k - 1)/2 \) QP problems with less than \( n \) variables.

Hsu and Lin (2002) compared several multiclass SVMs, and their experiments indicate that the OVO SVM is more suitable for practical use (with large or medium sample size) concerning both accuracy and running speed. However, some researchers (Statnikov et al., 2005; Yeang et al., 2001) found that, in small sample sized microarray data, OVO SVMs perform worse than OVA SVMs and other ‘all-together’ multiclass SVMs. It is also observed from our experimental studies (refer to Supplemental Material). The reason perhaps is that each binary classifier in the OVO SVM uses only a fraction of total training samples, and the small training set might make these binary classifiers subject to overfitting.

2.1.4 Multiclass SVMs: method by Weston and Watkins (WW)
This is the first ‘all-together’ implementation of multiclass SVMs by solving one single optimization problem (Vapnik 1998; Weston and Watkins, 1999). The idea is similar to the OVA approach. It simultaneously constructs \( k \) binary classifiers where the \( r \)-th function, \( w_r^T x + b_r \), separates the class \( r \) from the others. The formulation is as follows:

\[
\min \quad \mathcal{P}(w, \xi) = \frac{1}{2} \sum_{r=1}^{k} \|w_r\|^2 + C \sum_{j=1}^{n} \sum_{r=1, r \neq j}^{k} \xi_{rj}, \\
\text{subject to} \quad w_r^T x_i + b_r \geq w_{rj}^T x_i + b_j + 2 - \xi_{rj}, \quad \xi_{rj} \geq 0, \quad r = 1, \ldots, k \setminus j.
\]
The decision function is the same as that of OVA SVMs, that is,
\[ y = \arg \max_r (w_r^T x + b_r). \] (4)
The QP problem arising from (3) has \((k - 1)n\) variables.

Given that the computational complexity of the QP optimization algorithm is polynomial to the number of variables, it is computationally much more expensive to solve multiclass classification using ‘all-together’ methods than using methods based on binary SVMs.

2.1.5 Multiclass SVMs: method by Crammer and Singer (CS) This method is similar to WW. It uses fewer slack variables in the constraints of the optimization problem, and there is no bias item \(b\) in the decision function (Crammer and Singer, 2001). Basically, this method solves the following optimization problem:
\[
\begin{align*}
\text{minimize} & \quad P(w, \xi) = \frac{1}{2} \sum_{i=1}^{k} \|w_i\|^2 + C \sum_{i=1}^{n} \xi_i, \\
\text{subject to} & \quad w_i^T x_i - w_j^T x_{ji} \geq 1 - \xi_i, \quad r \neq j, \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots, n.
\end{align*}
\] (5)
The decision function is
\[ y = \arg \max_r (w_r^T x). \]
The corresponding dual problem of (5) is a QP problem with \(kn\) variables. Although the QP problem is larger than that from WW, Crammer and Singer, (2001) decomposed the single problem into multiple QP problems with reduced size, which saves much computation.

2.1.6 Multiclass SVMs: method by Lee, Lin and Wahba (LLW) This method has a promising theoretical property that its solution for the multiclass problem resembles Bayes rule asymptotically. The formulation is as follow:
\[
\begin{align*}
\text{minimize} & \quad P(w, \xi) = \frac{1}{2} \sum_{i=1}^{k} \|w_i\|^2 + C \sum_{i=1}^{n} \sum_{r=1, r \neq j}^{k} \xi_i, \\
\text{subject to} & \quad w_i^T x_i + b_i \leq -\frac{1}{k} \sum_{r=1, r \neq j}^{k} + \xi_i, \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots, n, \quad r \in \{1, \ldots, k\} \setminus j, \\
& \quad \sum_{i=1}^{n} (w_i^T x_i + b_i) = 0, \quad i = 1, \ldots, n.
\end{align*}
\] (6)
The formulation presented here is slightly different from that in Lee et al. (2004). Here we assume that the linear kernel and equal misclassification costs are employed. Please refer to Lee et al. (2004) for the general form. The final decision function is just the same as (4). The single QP dual problem of (6) has \((k - 1)n\) variables.

2.2 Support Vector Machine—Recursive Feature Elimination (SVM-RFE) SVM-RFE (Guyon et al., 2002) conducts feature selection in a sequential backward elimination manner, which starts with all the features and discards one feature at a time. Just like SVM, SVM-RFE was initially proposed for binary problems. The squared coefficients \(w_j^2 (j = 1, \ldots, p)\) of the weight vector \(w\) obtained from binary problem (1) are employed as feature ranking criteria. Intuitively, those features with the largest weights are the most informative. Thus in an iterative procedure of SVM-RFE one trains the SVM classifier, computes the ranking criteria \(w_j^2\) for all features, and discards the feature with the smallest ranking criterion. The procedure is repeated until a small subset of features is obtained.

Actually, the magnitude of \(w_j^2\) just corresponds to the approximate change in the criterion in (1),
\[ J = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i, \] when the \(j\)-th feature is discarded. This is explained by the Optimal Brain Damage (OBD) algorithm (Lecun et al., 1990). The criterion \(J\) can be expanded in Taylor series to second order,
\[ \Delta J(j) = \frac{\partial J}{\partial w_j} \Delta w_j + \frac{\partial^2 J}{\partial^2 w_j} (\Delta w_j)^2 + O((\Delta w_j)^3) \] (7)
At the optimum of \(J\), the first order term can be neglected\(^1\), then Equation (7) becomes \(\Delta J(j) \approx (\Delta w_j)^2\). If we denote by \(J(j)\) the value of \(J\) when the \(j\)-th feature is removed (by setting the corresponding weight to 0), it follows that
\[ J(j) \approx J + w_j^2. \] (8)
Therefore, removing the feature with smallest \(w_j^2\) will cause the least increase of \(J\), which increases the generalization performance at the same time. In other words, SVM-RFE aims to find the gene subset that yields minimal criterion \(J\). This criterion was also employed by Zhou and Mao (2006) in a sequential forward selection.

3 EXTENSIONS OF SVM-RFE FOR MULTICLASS PROBLEMS

3.1 Existing extension: OVA-RFE SVM-RFE (Guyon et al., 2002) is originally designed for binary problems. Some researchers (Chai and Domeniconi, 2004; Ramaswamy et al., 2001; Rifkin et al., 2003) employed RFE for multiclass gene selection by ranking genes for each class separately in the OVA manner.

For \(k\)-class problem (\(k \geq 3\)), \(k\) binary SVM classifiers are constructed using OVA method described in Section 2.1.2. For the \(r\)-th binary classification problem, SVM-RFE is carried out to identify a feature subset \(S_r\) for class \(r\) against all other classes. After \(k\) feature subsets are selected, the final selected subset for the whole multiclass problem is the combination of all the \(k\) subsets. This extension of SVM-RFE is called OVA-RFE in the present work.

\(^1\)Actually, the criterion for OBD is not \(J = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i\), but the Lagrangian function from (1), that is \(L = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i[y_i(w^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i \xi_i\), because \(\frac{\partial L}{\partial w_j} = 0\) and \(\frac{\partial L}{\partial \xi_j} \neq 0\). However, at the optimum of (1) and corresponding dual problem, \(J = L\), moreover, \(\frac{\partial^2 L}{\partial^2 w_j} = 0\) and \(\frac{\partial^2 L}{\partial \xi_j} = 1\), so for simplicity here we use \(J\) instead of \(L\). In the following, we use the similar simplified forms for the extensions of SVM-RFE.
The $k$ subsets $S_r$ ($r = 1, \ldots, k$) are identified independently. According to Section 2.1.2 and 2.2, the subset $S_r$ minimizes the following criterion,

$$J_r = \frac{1}{2} \|w_r\|^2 + C \sum_{i=1}^n \xi_i. \quad (9)$$

However, it cannot guarantee that the final selected feature subset ($S_1 \cup S_2 \cup \ldots \cup S_k$) simultaneously minimizes all the $k$ criteria: $J_1, J_2, \ldots, J_k$. For example, some features selected from one binary problem may reduce the classification performance in other binary problems. As a consequence, the feature subset selected by OVA-RFE might not be an optimal set.

### 3.2 New extensions: MSVM-RFEs

To overcome the limitations of OVA-RFE, we propose MSVM-RFE algorithms for feature selection on multiclass data. As there are two types of multiclass SVMs: methods based on multiple binary SVMs and ‘all-together’ methods, our proposed extensions are derived from these two groups, respectively.

#### 3.2.1 MSVM-RFE based on OVA SVMs

According to our discussion in Section 3.1, the optimal feature subset should simultaneously minimize the $k$ criteria in Equation (9) with $r = 1, \ldots, k$. Therefore, the multiclass feature selection is cast as a multi-objective optimization problem. The simplest way to solve the multi-objective optimization problem is the Weighting Objectives Method (Geoffrion, 1968). Assume that the $k$ classes equally contribute to the classification task. The multi-objective optimization is changed to the following single objective optimization problem by employing equal weights,

$$\text{minimize } J = \frac{1}{k} \sum_{r=1}^k J_r = \frac{1}{k} \sum_{r=1}^k \left( \frac{1}{2} \|w_r\|^2 + C \sum_{i=1}^n \xi_i \right). \quad (10)$$

The OBD algorithm (Lecun et al., 1990) is employed to solve this optimization problem. As in the binary SVM-RFE, the criterion $J$ in (10) is expanded in Taylor series to second order,

$$\Delta J(j) \approx \frac{1}{k} \sum_{r=1}^k \left[ \frac{\partial J}{\partial w_{rj}} \Delta w_{rj} + \frac{\partial^2 J}{\partial^2 w_{rj}} (\Delta w_{rj})^2 \right]. \quad (11)$$

At the optimum of $J$, $\Delta J(j) \approx \frac{1}{k} \sum_{r=1}^k (\Delta w_{rj})^2$. Similar with binary SVM-RFE, it follows that

$$J(j) \approx J + \frac{1}{k} \sum_{r=1}^k (w_{rj})^2. \quad (12)$$

where $J(j)$ is the value of $J$ when the $j$-th feature is eliminated. In consequence, removing the feature with smallest $\sum_{r=1}^k w_{rj}^2$ will cause the least increase of $J$. Therefore, $\sum_{r=1}^k w_{rj}^2$ can be employed as a ranking criterion in the recursive elimination procedure as described in Algorithm 1. Compared with OVA-RFE which independently identifies features from each class, this new extension simultaneously takes all the $k$ classes into consideration.

It is straightforward to apply such techniques on other binary SVMs based methods, such as OVO SVM (Krelle, 1999). However, the extension based on OVO SVMs does not perform as well as other extensions proposed in this paper (see experimental results in the Supplementary Material) on the small sample sized microarray data. It might be the direct impact of the overfitting of OVO classifiers we mentioned in Section 2.1.3. With large sample sizes, we expect the extension based on OVO SVMs to perform as well as or better than other extensions proposed here.

#### 3.2.2 MSVM-RFE based on ‘all-together’ methods

We take WW (Weston and Watkins, 1999, Section 2.1.4) as an example to illustrate our extensions on ‘all-together’ methods. The criterion used in the OBD algorithm (LeCun et al., 1990) is just the cost function in (3), that is,

$$J = \frac{1}{2} \sum_{r=1}^k \|w_r\|^2 + C \sum_{i=1}^n \sum_{r=1}^k \xi_i. \quad (13)$$

Similar with binary SVM-RFE, the criterion $J$ is expanded in Taylor series to second order,

$$\Delta J(j) \approx \sum_{r=1}^k \frac{\partial J}{\partial w_{rj}} \Delta w_{rj} + \sum_{r=1}^k \frac{\partial^2 J}{\partial^2 w_{rj}} (\Delta w_{rj})^2. \quad (14)$$

At the optimum of $J$, it follows that

$$J(j) \approx J + \sum_{r=1}^k w_{rj}^2. \quad (15)$$

So $\sum_{r=1}^k w_{rj}^2$ is still an appropriate ranking criterion for ‘all-together’ SVMs.

The recursive elimination procedure of our proposed extensions is summarized as Algorithm 1. We propose four MSVM-RFE algorithms based on different multiclass SVMs, respectively. They are, for short, MSVM-RFE(OVA) MSVM-RFE(WW), MSVM-RFE(CS) and MSVM-RFE(LLW). To reduce the computational cost, the algorithm can be generalized to discard more than one gene at each step, although discarding several genes may degrade the selection performance.

**Algorithm 1: Multiclass SVM-RFE**

**Input:** $n$ data pairs $\{(x_i, y_i)\}$ $m$ (predefined number of features to be selected) $p$ (number of features in set $S$)

**Output:** the index of $m$ selected features

Initialize $S$ to the full feature set; $S$ is the set of selected features $p = \text{number of features in set } S$; $\text{while } p \neq m \text{ do}$

Train a multiclass SVM with the features in set $S$; $w_r = [w_{r1}, w_{r2}, \ldots, w_{rD}]^T$

Calculate the ranking criteria for set $S$: $c_i = \sum_r w_{rj}^2$; Identify the feature with the smallest ranking criterion; Remove the identified feature from set $S$; $p = \text{number of features in set } S$; $\text{end}$

Return feature set $S$.
4 EXPERIMENTS AND DISCUSSION

We tested our proposed MSVM-RFE algorithms on six microarray data sets. The specifications of the data sets are listed in Table 1. Please refer to the Supplementary Material for more detailed data description and preprocessing.

For assessment of gene selection algorithms, Ambroise and McLachlan (2002) suggested external cross-validation, in which the gene selection and validation are performed on different parts of the sample set, to obtain an unbiased performance estimate. In this work, we employed external 4-fold stratified cross-validation to evaluate the selection performance. The stratified cross-validation (Breiman et al., 1984) is slightly different from regular cross-validation. In k-fold stratified cross-validation, a data set is randomly partitioned into k equally sized folds such that the class distribution in each fold is approximately the same as that in the entire data set. In contrast, regular cross-validation randomly partitions the data set into k-folds without considering class distributions. Kohavi (1995) reported that stratified cross-validation has smaller bias and variance than regular cross-validation. To obtain a more reliable estimate, the external 4-fold cross-validation process was repeated 100 times using different partitions of the data set.

We compared our MSVM-RFE algorithms with two multi-class gene selection algorithms: BSS/WSS and OVA-RFE. BSS/WSS (Ratio of Between-group to Within-groups Sum of Squares) was proposed by Dudoit et al., (2002) to perform gene selection for multi-class problems. BSS/WSS is an individual feature ranking criterion. For a gene j, this ratio is

\[ \frac{BSS(j)}{WSS(j)} = \frac{\sum_{m=1}^{k} \sum_{r=1}^{C} I(y_i = r)(\bar{x}_{ij} - \bar{x}_j)^2}{\sum_{m=1}^{k} \sum_{r=1}^{C} I(y_i = r)(\bar{x}_r - \bar{x}_j)^2}, \]

where \( I() \) is the indicator function, \( \bar{x}_j \) denotes the average expression level of gene j across all samples and \( \bar{x}_r \) denotes the average expression level of gene j across samples belonging to class r. Intuitively, genes with large BSS/WSS ratios (that is, relatively large variation among classes and relatively small variation within classes) are likely relevant to class separation. OVA-RFE is described in Section 3.1.

In Algorithm 1, MSVM-RFE algorithms remove one gene with the smallest ranking criterion in each step. However, the number of genes in microarray data is very large (normally several thousands). If only one gene is removed in each step, the computational cost would be very high. Therefore, in the present work genes corresponding to the ranking criterion in the bottom 10% of the remaining genes were removed to expedite the selection procedure.

To select the relevant genes using MSVM-RFE algorithms, it is important to find the appropriate value of C which is used to construct the multiclass SVMs during the recursive elimination procedure. Generally, the appropriate C should be variant when multiclass SVM classifiers are trained on different gene subsets. However, it might not be easy to tune the parameter C for each SVM classifier, and such tuning would demand substantially increased computation. In the present study, we employed fixed C during the RFE procedure. The parameter was tuned through a simple technique. We assessed the performance of the BSS/WSS algorithm using external cross-validation with a sequence of given values of C (for example, \( C = 10, 1, 0.1, 0.01, 0.001, 0.0001 \), respectively). The optimal value of C is chosen to be the one which gives the minimal external cross-validation error on the selected gene subsets. For the GCM data set, C was set to 0.1 for OVA SVM and corresponding RFE algorithms, such as OVA-RFE and MSVM-RFE (OVA). Accordingly, \( C = 0.1 \) for WW related algorithms, while \( C = 0.01 \) for WW and CS related algorithms.

Another issue is to decide the number of genes to be selected. Generally, finding the optimal number of genes is very difficult. In the present study, we employed the solution in Li et al., (2004). A set of simulations were conducted on the GCM dataset by varying the number of selected genes. When the number of selected genes is >400, the selection performance is not significantly increased, and the variance of external cross-validation error is small and stable. The same results were observed on other datasets. Hence, a maximum of 400 genes were identified in all experiments.

All programs are written in C/C++. We used LIBSVM (Chang and Lin, 2001) to implement OVA SVMs and corresponding gene selection algorithms, including OVA-RFE and MSVM-RFE (OVA). BSVM (Hsu and Lin, 2002) was employed to implement WW2 and CS related algorithms. As there are no available fast implementations of LLW SVMs, we implemented this SVM framework using the OOQP package (Gertz and Wright, 2003) to solve the arising QP problems. The QP problem from LLW SVMs has \( (k-1) \) variables, so the computational cost is very high, especially for large data sets. In this work, we only tested the selection performance when \( k, 2k, \ldots \) genes were selected, respectively, using LLW SVMs on the GCM and 11_Tumors data sets.

The results and genes selected on the GCM data set are presented here for illustration of typical results. (However, all the experimental results are available as Supplementary Material). For the GCM data set, we compared the selection performances of six gene selection algorithms, as shown in Figure 1. As an individual ranking method, the performance of 2

\[ \text{The WW SVM algorithm implemented in BSVM is slightly different from the original algorithm. BSVM solves a modified optimization problem, which adds a term } \sum_{i=1}^{n} b_i^2 \text{ to the objective function (3). By doing so, the corresponding dual problem is simplified, and is easy to be solved using fast decomposition techniques. Considering its fast running speed, we used BSVM for classification and gene selection on large data sets (11_Tumor and GCM datasets), but for other smaller data sets we still used the original WW SVM algorithm implemented by the OOQP package (Gertz and Wright, 2003).} \]

**Table 1.** Description of the data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Samples</th>
<th>Genes</th>
<th>Classes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>198</td>
<td>14122</td>
<td>14</td>
<td>Ramaswamy et al. (2001)</td>
</tr>
<tr>
<td>11_Tumors</td>
<td>174</td>
<td>9700</td>
<td>11</td>
<td>Su et al. (2001)</td>
</tr>
<tr>
<td>NCI Ross</td>
<td>58</td>
<td>5643</td>
<td>8</td>
<td>Ross et al. (2000)</td>
</tr>
<tr>
<td>NCI Stauton</td>
<td>58</td>
<td>3144</td>
<td>8</td>
<td>Stauton et al. (2001)</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>203</td>
<td>5343</td>
<td>5</td>
<td>Bhattacharjee et al. (2001)</td>
</tr>
<tr>
<td>MLL</td>
<td>72</td>
<td>10930</td>
<td>3</td>
<td>Armstrong et al. (2002)</td>
</tr>
</tbody>
</table>
The BSS/WSS algorithm is much worse than those of SVM-RFE extensions. This is because individual ranking methods may select many redundant genes, which add little discriminatory power to the classification problem under investigation, while SVM-RFE extensions that actually evaluate gene subsets instead of individual genes implicitly take into account the effect of irrelevant or redundant genes. The MSVM-RFE algorithms proposed here outperform OVA-RFE as shown in Figure 1. We believe the good performance is due to considering all classes simultaneously during the gene selection procedure. However, the performances of four MSVM-RFE algorithms are not the same. MSVM-RFE(LLW) is not as good as other three algorithms. MSVM-RFE(CS) achieves better performance when only 50–150 genes are selected, and gives similar performance with MSVM-RFE(OVA) and MSVM-RFE(WW) when 200 genes are selected.

The test performances in Figure 1 are based on different classification algorithms, that is, genes selected by BSS/WSS, OVA-RFE or MSVM-RFE(OVA) are evaluated using OVA SVMs, while MSVM-RFEs (CS, WW, LLW) are tested by CS, WW and LLW SVMs, respectively. Therefore, we carried out a fair comparison with all gene selection algorithms assessed by one multiclass SVM. The comparison results are shown in Table 2. The results are similar to those in Figure 1 no matter which multiclass SVM algorithm is employed. If \( \mu_A, \sigma_A, \mu_B \), and \( \sigma_B \) represent means and standard deviations of the error rates from methods \( A \) and \( B \), respectively, we can reach a conclusion that method \( A \) is significantly better than method \( B \) when \( \mu_B \geq \mu_A + \sigma_A \). Based on the rule-of-thumb, our MSVM-RFE algorithms are significantly better than OVA-RFE and BSS/WSS. Furthermore, the performances of MSVM-RFE algorithms are significantly better than that without any gene selection being performed as well.

We combined all the selected genes during cross-validation procedure to identify the most important ones. As we applied our gene selection algorithms 400 times on different subsets of the original data set (4-fold cross-validation repeated 100 times), we actually obtained 400 different selected gene subsets (each contains 400 genes). By computing the frequency of each gene appearing in all the 400 gene subsets, we can identify the important genes which have been most frequently selected for different re-sampling sets. For MSVM-RFE(OVA) on the GCM data set, we finally identified 342 genes with frequency \( \geq 200 \), which are shown in Figure 2. While for MSVM-RFEs (WW, CS, LLW), 351, 353 and 297 markers were identified. A complete list of selected markers is provided with the Supplementary Material.

The MSVM-RFE algorithms proposed here provide improved performance for the challenge of selecting genes which contribute to learning in multiple classes simultaneously. Improving classification performance also affords the opportunity to contribute to biological discovery, however, interpretation of results using these methods requires care since not all genes will be selected based on specificity for a single tumor type. The algorithms proposed here select a mixture of genes which exhibit molecular features specific for the neoplastic process along with features characteristic of a tumor’s tissue of origin, but which may represent normal tissue-specific
processes. However, beyond this, optimal classification performance may also lead to selection of genes which may not be associated primarily with a single tumor or tissue type, but which may contribute to learning differences among various combinations of samples. The potential contribution of gene selection techniques for discovery of biological relationships has been noted by most previous researchers, although their focus has been primarily on the genes which are specific to a single tissue or tumor subtype. Uncovering the true relations may prove to be more complex than in identifying the role of a gene which is highly specific to a single tumor type.

Many of the most tumor or tissue specific-genes which were identified in previous analysis of the data sets which we studied were also selected by our algorithms. For the GCM data set, for instance, we find that a number of the top selected genes are genes which have well established histories as biomarkers for specific tumors, similar to what was noted by Ramaswamy et al. (2001). CEA (carcinoembryonic antigen) is the top ranked gene for colorectal cancer, PSA (prostate-specific antigen) for prostate cancer, CD20 for lymphomas. Perhaps, more interestingly, we have also identified several genes which have more recently been noted to have close biological associations with established biomarkers. Several recent publications have begun to explore roles for these molecules as biomarkers in their own right, and in combination with other markers. For instance, GATA3 was consistently identified as one of the top markers by all the algorithms. It is associated with estrogen receptor status and biological function and is increasingly recognized as a biomarker in its own right, with its own role in breast cancer prognosis (Oh et al., 2006). Galectin-4, also consistently identified as one of the most informative genes, is a lectin for which CEA has been identified as a potential ligand (Ideo et al., 2005).

Again similar to previous studies (Ramaswamy et al., 2001) a number of genes are selected for optimal multiclass classifiers due to their tissue specificity, rather than their association with the neoplastic process, per se. Examples include genes with gastrointestinal tissue origin (intestinal Trefoil factor 3) which are correlated with colorectal tumors, lymphoid specific functions (such as Lymphotoxin beta, and EBI1-ligand chemokine), multiple genes specific to lung tissue (Pulmonary Surfactant-associated protein B precursor, SFTP1, SP-C1). Several genes which are tissue specific to prostate were strongly selected including PSCA (prostate stem cell antigen), semenogelin, prostatic secretory protein 57 and prostate acid phosphatase. In addition to genes known to be involved in the leukemic process including Myc and Myb, whereas others were tissue-specific molecules, or various genes known to be involved in normal hematopoietic differentiation, sometimes associated with specific types of leukocytes, such as leukocyte surface markers CD37, terminal transferase, IL-8 and IL-6. Examples for melanoma include melanoma inhibitory activity (MIA), and differentiation antigen melan-A MLANA; for pancreas, genes directly or indirectly related to pancreatic function such as insulin, glucagon, spasmolytic polypeptide precursor; for CNS, the Zic gene identified by MSVM-RFE(LLW) encodes a member of the ZIC family of C2H2-type zinc finger proteins. Aberrant expression of this gene is seen in medulloblastoma, but perhaps more relevantly it encodes a transcription factor that can bind and transactivate the apolipoprotein E gene, and may serve as an identifier of tissue specificity more than tumor specificity.

Other genes which were consistently selected may characterize the third type of gene which may be less specific to a single tumor type, but serve as key contributors to optimal multiclass classification. It is possible that this group of genes may instead of serving as biomarkers for a specific tumor type may be important areas for further exploration and novel discoveries of molecular elements involved in biological processes relevant to the malignant process across multiple tumor types. Among genes with established functions, Claudin-4, for instance consistently selected as one of the top five genes overall in contributing to multiclass classification, is suspected to be involved in several epithelial based tumor types. Claudin-4 is a member of a family of transmembrane proteins, essential in the formation and maintenance of tight junctions (Hewitt et al., 2006). Several genes were selected which did not appear even in the top 1000 genes for each individual tumor type on the GCM data set (Ramaswamy et al., 2001). For instance, tumor-associated antigen CO-029 is a tetraspanin shown to be associated with integrins (Gesierich et al., 2005) and expressed in several different tumor types. Although it was not specific to a single tumor type, it contributed to sub-classification within multiple groups and therefore contributed to information for

<table>
<thead>
<tr>
<th>Gene selection algorithms</th>
<th>OVA SVM</th>
<th>WW SVM</th>
<th>CS SVM</th>
<th>LLW SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSS/WSS</td>
<td>26.40 ± 1.60</td>
<td>29.90 ± 1.50</td>
<td>25.77 ± 2.13</td>
<td>27.07 ± 1.70</td>
</tr>
<tr>
<td>OVA-RFE</td>
<td>19.50 ± 1.77</td>
<td>19.86 ± 1.81</td>
<td>19.76 ± 1.88</td>
<td>19.47 ± 1.62</td>
</tr>
<tr>
<td>MSVM-RFE(MLA)</td>
<td>16.75 ± 1.48</td>
<td>16.74 ± 1.40</td>
<td>16.76 ± 1.43</td>
<td>16.84 ± 1.45</td>
</tr>
<tr>
<td>MSVM-RFE(CS)</td>
<td>16.82 ± 1.66</td>
<td>16.78 ± 1.66</td>
<td>16.86 ± 1.69</td>
<td>16.82 ± 1.53</td>
</tr>
<tr>
<td>MSVM-RFE(LLW)</td>
<td>16.72 ± 1.54</td>
<td>16.88 ± 1.47</td>
<td>16.89 ± 1.51</td>
<td>16.87 ± 1.51</td>
</tr>
<tr>
<td>Without gene selection</td>
<td>17.58 ± 1.51</td>
<td>17.63 ± 1.61</td>
<td>17.64 ± 1.61</td>
<td>17.54 ± 1.40</td>
</tr>
<tr>
<td></td>
<td>20.64 ± 1.68</td>
<td>21.67 ± 1.70</td>
<td>21.67 ± 1.69</td>
<td>20.10 ± 1.60</td>
</tr>
</tbody>
</table>
the multiclass problem. Therefore, it is possible that further exploration among the list of genes with unknown function which were consistently selected for optimal classification could either be genes with tissue or tumor specificity or could identify genes involved more fundamentally with the neoplastic process.

In this study, we proposed four MSVM-RFE algorithms, which show promising performance compared with other multiclass gene selection algorithms. However, it is difficult to determine the best gene selection approach among the extensions we proposed. There does not seem to exist a clear winner. For example, MSVM-RFE(OVA) achieves the best result on the MLL data sets, while MSVM-RFE(LLW) is best on the NCI Ross and Lung data sets; MSVM-RFE(CS) has superb performance on the GCM, 11_Tumor and NCI Staunton data sets, and MSVM-RFE(WW) is similar to MSVM-RFE(CS) on the 11_Tumor and NCI Staunton data sets. Therefore, which MSVM-RFE is best may depend on the particular data set. So surprisingly, although MSVM-RFE(CS) has the best performance on fully half of the datasets, it performs poorly on the Lung and MLL data sets (refer to Supplementary Material for details). The reason for the poor performance on these data sets is not clear. We conjecture that it may be due to lacking bias item b in the CS SVM classification (Poggio et al., 2002). To identify the factors which influence this requires further investigation.

5 CONCLUSION

In this paper, we proposed four extensions of SVM-RFE (called MSVM-RFE) to solve the multiclass gene selection problem, based on different frameworks for multiclass SVMs. By simultaneously considering all classes during the gene selection stages, our proposed extensions identify genes leading to more accurate classification.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Robert Bjornson and Nicholas Carriero for their help in running codes on the Bulldog Cluster. This work is supported by NIH P30 DK072442. Part of the simulations were run on the Bulldog Cluster, funded by NIH grant RR19895-02, at Yale Center for High Performance Computation in Biolog and Biomedicine.

Conflict of Interest: none declared.

REFERENCES


