Gene selection using support vector machines with nonconvex penalty

Hao Helen Zhang¹,², Jeongyoun Ahn², Xiaodong Lin³, and Cheolwoo Park⁴

¹ Department of Statistics, North Carolina State University, Raleigh, NC 27695,
² Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599,
³ Department of Mathematical Sciences, University of Cincinnati, OH 45221,
⁴ Department of Statistics, University of Georgia, Athens, GA 30602, USA.

ABSTRACT

Motivation: With the development of DNA microarray technology, scientists can now measure the expression levels of thousands of genes simultaneously in one single experiment. One current difficulty in interpreting microarray data comes from their innate nature of “high dimensional low sample size.” Therefore, robust and accurate gene selection methods are required to identify differentially expressed group of genes across different samples, e.g., between cancerous and normal cells. Successful gene selection will help to classify different cancer types, lead to a better understanding of genetic signatures in cancers, and improve treatment strategies. Although gene selection and cancer classification are two closely related problems, most existing approaches handle them separately by selecting genes prior to classification. We provide a unified procedure for simultaneous gene selection and cancer classification, achieving high accuracy in both aspects.

Results: In this paper we develop a novel type of regularization in support vector machines (SVMs) to identify important genes for cancer classification. A special nonconvex penalty, called the smoothly clipped absolute deviation penalty, is imposed on the hinge loss function in the SVM. By systematically thresholding small estimates to zeros, the new procedure eliminates redundant genes automatically and yields a compact and accurate classifier. A successive quadratic algorithm is proposed to convert the non-differentiable and nonconvex optimization problem into easily solved linear equation systems. The method is applied to two real data sets and has produced very promising results.

Availability: MATLAB codes are available upon request from the authors.

Contact: hzhang@stat.ncsu.edu

Supplementary information:
http://www4.stat.ncsu.edu/hzhang/pub.html

1 INTRODUCTION

We consider the problem of gene selection for cancer classification using microarray gene expression data. The objective of gene selection is two-fold: to provide a better understanding of the underlying biological system that generates data, and to improve the prediction performance of classifiers. Effective gene selection often leads to a compact classifier with better accuracy and interpretability (Kitter, 1986).

Gene selection is treated as a variable selection problem in statistics and a dimension reduction problem in machine learning. Many greedy algorithms have been developed in the literature. Gene-ranking methods are particularly popular, which select genes according to some pre-determined ranking criteria. There are two main types of ranking criteria: correlation coefficients (Golub et al., 1999; Furey et al., 2000; Pavlidis et al., 2001) and hypothesis testing statistics. Two-sample t-test methods include parametric tests (Devore and Peck, 1997; Thomas et al., 2001; Pan, 2002) and nonparametric tests (Troyanskaya et al., 2002; He, 2004). Although being useful in practice, all these methods select important genes based on individual gene information thus fail to take into account mutual information among genes. Dimension reduction techniques project the full data onto the first few principal directions then conduct classification in the low dimensional subspace. West (2003) proposed the idea of “meta-genomes” which are linear combinations of the original genes. One disadvantage of projection methods is that none of the original genes can be discarded since each principal component generally involves all the genes.

Support vector machines (Boser et al., 1992; Vapnik, 1995; Cristianini and Shawe-Taylor, 1999) have demonstrated superior performances in classifying high dimensional and low sample size data. However, the standard SVM can suffer from the presence of redundant variables (Hastie et al., 2001; Guyon et al., 2002), since its decision rule utilizes all the variables without discrimination. Several methods have...
been proposed for variable selection in the SVM (Furey et al., 2000; Rakotomamonjy, 2003; Grandvalet and Canu, 2002; Mukherjee et al., 2000; Chapelle et al., 2002; Weston et al., 2000). Guyon et al. (2002) developed the recursive feature elimination (RFE) algorithm which successively eliminates features by training a sequence of SVM classifiers. Bradley and Mangasarian (1998) suggested the $L_2$ SVM which imposes the absolute value penalty on the directional vector of the separating plane.

Different from all the methods above, we formulate the SVM as a regularization problem with a novel form of the penalty. The optimization problem consists of two parts: the first part which separates two classes by maximizing the margin between them and classifying structures where a linear separation is not plausible, the second part is the Bayes rule sign function. Often we have a misclassification error. For classifying the data with complicated structures where a linear separation is not plausible, the classification problem is to learn a discrimination function $g(x) = \text{Prob}(Y = +1|X = x)$. With the 0-1 loss function, the optimal rule minimizing the expected loss $EL[f(X, Y)]$ is the Bayes rule $\text{sign}[g(x) - 1/2]$. If $f(x) = 0$, the point $x$ is randomly classified as $+1$ or $-1$. The function $\text{sign}(t)$ takes value 1 if $t > 0$ and takes value $-1$ if $t < 0$.

Support Vector Machines

Support vector machine (SVM) is a large margin classifier which separates two classes by maximizing the margin between them. For non-separable data, the soft-margin SVM uses the slack variable to control an upper bound of the misclassification error. For classifying the data with complicated structures where a linear separation is not plausible, the SVM is applicable to any biological data of high dimensional low sample size. Its performances on one microarray dataset and one metabolism dataset are illustrated in the paper.

2 METHODS

Given a training set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$ is the input vector and $y_i \in \{+1, -1\}$ indicates its class label, the classification problem is to learn a discrimination rule $f: \mathbb{R}^d \to \{+1, -1\}$ so that we can assign a class label to any new subject observed in the future. For microarray gene expression data, $x_i$ represents the expression levels of $d$ genes of the $i$th sample tissue and $y_i$ is "normal" or "cancerous"; often we have $d \gg n$. In the statistical framework, we assume $(x_i, y_i)$’s are independent realizations of the random pair $(X, Y)$ which follows a joint distribution $P(X, Y)$. Define $g(x) = \text{Prob}(Y = +1|X = x)$. With the 0-1 loss

$$L[f(x), y] = \begin{cases} 1 & \text{if } yf(x) < 0 \\ 0 & \text{if } yf(x) > 0, \end{cases}$$

the optimal rule minimizing the expected loss $EL[f(X, Y)]$ is the Bayes rule $\text{sign}[g(x) - 1/2]$. If $f(x) = 0$, the point $x$ is randomly classified as $+1$ or $-1$. The function $\text{sign}(t)$ takes value 1 if $t > 0$ and takes value $-1$ if $t < 0$.

Fig. 1. Hard-thresholding penalty $L_0$; soft-thresholding penalties $L_{0.5}, L_1,$ and $L_2$ penalty.

The SVM finds $f(x) = b + w \cdot h(x)$ by minimizing

$$\frac{1}{n} \sum_{i=1}^n [1 - y_i(b + w \cdot h(x_i))]_+ + \lambda\|w\|^2; \quad (1)$$

where $b$ is constant, $w$ is the directional vector, and $D = \{h_1(x), \ldots, h_q(x)\}$ is a dictionary of basis functions. The parameter $\lambda$ controls the trade-off between minimizing the loss function and maximizing the margin. The hinge loss $[1 - yf(x)]_+$ is a convex upper bound for the 0-1 loss. One nice property of the SVM is that its solution only depends on a small subset of the training set called “support vectors”. The standard SVM can not select important variables, since all the input variables are used for constructing the classifier. For variable selection purposes, the following thresholding penalty functions can be used to replace the $L_2$ penalty $\|w\|^2$:

$$L_0(w) = \sum_{j=1}^q I(w_j \neq 0),$$
$$L_1(w) = \sum_{j=1}^q |w_j|,$$
$$L_2(w) = \sum_{j=1}^q |w_j|^\gamma, \quad 0 < \gamma < 1.$$
for signal denoising, where the former leaves the magnitudes of coefficients unchanged if they are larger than a given threshold, while the latter shrinks them to zero by the threshold value. It is known that the $L_\gamma$ function is a soft-thresholding penalty only if $\gamma \leq 1$ (Bradley and Mangasarian, 1998). This explains why the standard SVM corresponding to $\gamma = 2$ does not select variables. In Figure 1, the first three penalty functions are all non-differentiable at the origin, which is a necessary condition for a penalty function to produce sparse solutions (Fan and Li, 2001).

In the statistics literature, the $L_1$ penalty is also known as the LASSO (Tibshirani, 1996) and widely used for variable selection in linear regression models. The $L_1$ SVM was proposed by Bradley and Mangasarian (1998). Recently Zhu et al. (2003) studied its solution properties and suggested an algorithm to find the whole solution path over a range of tuning parameters. Fung and Mangasarian (2004) developed a fast Newton algorithm to solve its dual problem. Other methods incorporating the model parsimony include some Bayesian methods (Lee et al., 2003; Bae and Mallick, 2004).

The SCAD SVM

Though the $L_1$ penalty gives sparse solutions, the estimates can be biased for large coefficients since larger penalties are imposed on larger coefficients. In linear regression models, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty which overcomes the biasness problem of the $L_1$ penalty. They showed that the SCAD penalty produces sparse solutions by thresholding small estimates to zero, provides nearly unbiased estimates for large coefficients, and gives a model continuous in data. In this paper, we propose the SCAD SVM to conduct variable selection in the context of classification, and study its performances on high dimensional low sample size data.

The SCAD function, as plotted in Figure 2, is symmetric, nonconvex, and singular at the origin. Though having the same form as the $L_1$ penalty at the neighborhood of zero, the SCAD applies a constant penalty for large coefficients while the $L_1$ penalty increases linearly as the coefficient increases. It is this distinct feature that guards the SCAD penalty against producing biases for estimating large coefficients. Mathematically, the SCAD penalty has the expression

$$p_\alpha(|w|) = \begin{cases} 
\lambda |w| & \text{if } |w| \leq \lambda, \\
(\lambda |w|^2 - 2\alpha \lambda |w| + \alpha^2 \lambda^2)^{1/(\alpha - 1)} & \text{if } \lambda < |w| \leq \alpha \lambda, \\
(\alpha + 1) \lambda^2 |w|^{\alpha - 1} & \text{if } |w| > \alpha \lambda,
\end{cases} \quad (2)$$

where $\alpha > 2$ and $\lambda > 0$ are tuning parameters. In Figure 2, we have $\alpha = 3$ and $\lambda = 0.4$. The function in (2) is a quadratic spline function with two knots at $\lambda$ and $\alpha \lambda$. Except being singular at the origin, the function $p_\alpha(w)$ has a continuous first-order derivative. We propose the SCAD SVM as

$$\min_w \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(b + w \cdot h(x_i))]_+ + \epsilon \sum_{j=1}^{q} p_\alpha(|w_j|). \quad (3)$$

The objective function in (3) consists of the hinge loss part and the SCAD penalty on $w$. The parameter $\lambda$ balances the trade-off between data fitting and model parsimony. If $\lambda$ is too small, the procedure tends to overfit the training data and gives a classifier with little sparsity; if $\lambda$ is too large, the produced classifier can be very sparse but have a poor discriminating power. To tune $\lambda$ properly, we generate a tuning set for the simulated data and use cross validation for the real data. Fan and Li (2001) showed that the Bayes risks are not sensitive to the choice of $\alpha$, and $\alpha = 3.7$ is a good choice for various problems. We also use $\alpha = 3.7$ in our examples.

Interestingly, when $d \gg n$, linear classifiers often give better performances than nonlinear ones in many applications (Hastie et al., 2001), even though nonlinear methods are known to be more flexible. This fact is related to the asymptotic results in Hall et al. (2005): when $d \gg n$, under mild assumptions for data distribution, the pairwise distances between any two points are approximately identical to each other so the data points form an $n$-simplex. Linear classifiers then become natural choices to discriminate two simples. Since we focus on classifying high dimensional low sample size data, only linear SVMs are considered in this paper. In other words, we use the input vector $x$ as basis functions, i.e., $h(x) = x$ and $q = d$.

3 ALGORITHM

The standard SVM and $L_1$ SVM are often solved using quadratic programming and linear programming methods. However, many standard optimization packages fail to solve (3), because the hinge loss function is not differentiable at zero and the SCAD penalty is not convex in $w$. In this section, we propose an iterative algorithm to solve the SCAD SVM efficiently, and show that only a series of linear equation systems need to be solved.

Successive quadratic algorithm (SQA) is a generalization of Newton’s method for unconstrained optimization in that it finds a step away from the current point by minimizing a quadratic approximation of the problem. Numerous optimization
packages, including NPSOL, NLPQL, OPSYC, OPTIMA, and MATLAB, are found on this approach (More and Wright, 1993). We propose using the SQA to solve the SCAD SVM.

Denote the objective function in (3) by \(A(b, w)\). For each \(i\), we have \(y_i^2 = 1\) and
\[
[1 - y_i(b + w \cdot x_i)]^+ = \frac{1 - y_i(b + w \cdot x_i) + |y_i - (b + w \cdot x_i)|}{2}.
\]
Assume an initial value \((b_0, w_0)\) is given. Consider the local quadratic approximation for the second term in (4):
\[
[y_i - (b + w \cdot x_i)] \approx \frac{1}{2} \frac{y_i - (b + w \cdot x_i)}{y_i - (b_0 + w_0 \cdot x_i)} + \frac{1}{2} |y_i - (b_0 + w_0 \cdot x_i)|.
\]
For the SCAD penalty \(p_\lambda(w_j)\), we use the following quadratic approximation
\[
p_\lambda(w_j) \approx p_\lambda(w_{j0}) + p_\lambda^\prime(w_{j0}) \frac{(w_{j0})^2 - w_j^2}{2w_{j0}}.
\]
It is easy to check that both approximating functions have the same gradient as the original functions at the current point \((b_0, w_0)\). Thus minimizing the local quadratic approximation assures the convergence of the algorithm towards the correct descending direction of the original function. The quadratic form of the entire objective \(A(b, w)\) is given as
\[
A(b, w) \approx \frac{1}{2} - \frac{1}{2n} \sum_{i=1}^{n} y_i(b + w \cdot x_i)
\]
\[
+ \frac{1}{4n} \sum_{i=1}^{n} |y_i - (b_0 + w_0 \cdot x_i)|
\]
\[
+ \frac{1}{4n} \sum_{i=1}^{n} |y_i - (b_0 + w_0 \cdot x_i)|^2
\]
\[
+ \sum_{j=1}^{d} \left[ p_\lambda(w_{j0}) + \frac{p_\lambda^\prime(w_{j0})}{2w_{j0}} (w_j^2 - w_{j0}^2) \right].
\]
Removing the terms which do not involve \((b, w)\), we get
\[
\tilde{A}(b, w) = -\sum_{i=1}^{n} \frac{y_i(b + w \cdot x_i)}{2n} + \sum_{j=1}^{d} \frac{p_\lambda(w_{j0})}{2w_{j0}} w_j^2
\]
\[
- \frac{1}{2n} \sum_{i=1}^{n} y_i(b + w \cdot x_i)
\]
\[
+ \frac{1}{4n} \sum_{i=1}^{n} |y_i - (b_0 + w_0 \cdot x_i)|^2
\]
\[
+ \sum_{j=1}^{d} \frac{(b + w \cdot x_j)^2}{|y_i - (b_0 + w_0 \cdot x_j)|}.
\]
Define the matrix \(X = [1, x_1, \ldots, x_d]\), where 1 is the vector of 1’s with length \(n\) and \(x_j\) is the \(j\)th input vector. Define \(y = [y_1, \ldots, y_n]^T\), \(w = [w_1, \ldots, w_d]^T\), and \(e = [e_1, \ldots, e_n]^T\) with \(e_i = y_i - (b_0 + w_0 \cdot x_i)\). Define \(r = [y_1/|e_1|, \ldots, y_n/|e_n|]^T\), \(D_1 = \frac{1}{2n} \text{diag}[1/|e_1|, \ldots, 1/|e_n|]\), \(P = \frac{1}{2n} (y + r)^T X\), and \(D_2 = \text{diag}[0, p_\lambda(w_{j0})/|w_{j0}|, \ldots, p_\lambda(w_{j0})/|w_{j0}|] \cdot [w_{j0}]\). Then the approximate optimization problem becomes
\[
\min_{b, w} \tilde{A}(b, w) = \frac{1}{2} (b^T w) - P(b) - P(b).
\]
Since (5) is quadratic in \((b, w)\), solving (5) is equivalent to solving the following linear equation system
\[
Q(\hat{b}) = P.
\]
We propose the following algorithm to solve the SCAD SVM by solving a series of linear equation systems iteratively:

step 1: Set \(k = 1\) and specify the initial value \((b^{(1)}, w^{(1)})\).
step 2: Let \((b_0, w_0) = (b^{(k)}, w^{(k)})\). Minimize \(A(b, w)\) by solving (6). The solution is denoted as \((b^{(k+1)}, w^{(k+1)})\).
step 3: Let \(k = k + 1\). Go to step 2 until convergence.

If some \(w^{(k)}_j\) is very close to zero, say, smaller than a certain threshold, then the \(j\)th variable is regarded as redundant. Following Fan and Li (2001), we remove the \(j\)th column from the matrix \(X\) and adjust the coefficient matrix in (6) correspondingly. The iteration then continues as a reduced optimization problem. The algorithm stops when there is no change in \((b^{(k)}, w^{(k)})\). The \(j\)th variable is regarded as unimportant if \(|w_j| < \epsilon\), where \(\epsilon\) is a pre-selected small positive thresholding value. Based on our experiences, the solutions from the standard SVM provide good starting values. In all of our examples, this algorithm converges quickly.

4 SIMULATION

In the high dimensional low sample size setting, we simulate a data set which contains many redundant variables. Four methods are compared: standard SVM, \(L_1\) SVM, SCAD SVM, and another classifier called distance weighted discrimination (DWD). The DWD is a large-margin classifier developed using the second-order cone programming by Marron et al. (2004), and it does not suffer from the data piling problem as the standard SVM. We use the OSU SVM package (www.ece.osu.edu/~majoi/osu-svm) to implement the SVM and the algorithm of Fung and Mangasarian (2004) to implement the \(L_1\) SVM. A tuning set with the same size as the training set is used to choose the optimal \(\lambda\). Each classifier is evaluated on a test set of size 500. The thresholding value we used for removing variables is 0.001.

This example is a modification of the example used in Weston et al. (2000). There are \(d = 200\) inputs and only the first two are relevant. The probability of \(Y = +1\) or \(-1\) is equal. The inputs \(X_1\) and \(X_2\) are drawn from a mixture of Normal distributions: with probability 0.7, we have \(X_1 = Y N(3, 1)\) and \(X_2 = N(0, 1)\); with probability 0.3, we have \(X_1 = N(0, 1)\) and \(X_2 = Y N(3, 1)\). The inputs \(X_j, j = 1, 2, \ldots, d\).
$3, \ldots, 200$ are independently generated from $N(0, 20)$. The Bayes rule is slightly nonlinear around the origin but can be approximated well by linear functions. We consider various settings for the sample size: $n = 20, 30, 40, 50, 70, 100$. In each setting, we run thirty replicates and plot the average test errors in Figure 3. As $n$ increases, the test errors of the SCAD SVM and $L_1$ SVM decrease prominently compared to those of the DWD and the standard SVM. This suggests that variable selection is important when too many redundant variables are present. Furthermore, the SCAD SVM consistently outperforms the $L_1$ SVM in all the settings.

In Figure 4, for one particular simulated data of size $n = 100$ we plot the classification boundaries given by the Bayes rule and four learning methods. Note the Bayes rule is available only for simulated data. We use “o” for the points from the $+1$ class and “x” for those from the $-1$ class. The line symbols are: the DWD (dashed), the SVM (dotted), the $L_1$ SVM (dash-dotted), the SCAD SVM (thick solid), and the Bayes rule (thin solid). Since only $X_1$ and $X_2$ are truly relevant to the classification boundary, all the classifiers have been projected from the 200-dimensional input space to the first two-dimensional subspace. Figure 4 shows that the SCAD SVM classifier is the closest to the Bayes rule among all the classifiers. This explains why the SCAD SVM has the smallest test error rate as shown in Figure 3.

### Table 1. Number of variables selected by various methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>n=20</th>
<th>n=30</th>
<th>n=40</th>
<th>n=50</th>
<th>n=70</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>DWD</td>
<td>99.53</td>
<td>99.83</td>
<td>102.13</td>
<td>100.23</td>
<td>98.53</td>
<td>98.97</td>
</tr>
<tr>
<td></td>
<td>(0.83)</td>
<td>(0.73)</td>
<td>(0.76)</td>
<td>(0.67)</td>
<td>(0.70)</td>
<td>(0.61)</td>
</tr>
<tr>
<td>SVM</td>
<td>64.07</td>
<td>72.67</td>
<td>81.67</td>
<td>78.93</td>
<td>82.80</td>
<td>87.10</td>
</tr>
<tr>
<td></td>
<td>(1.47)</td>
<td>(1.48)</td>
<td>(1.45)</td>
<td>(0.96)</td>
<td>(1.10)</td>
<td>(0.99)</td>
</tr>
<tr>
<td>$L_1$ SVM</td>
<td>15.37</td>
<td>9.10</td>
<td>12.37</td>
<td>11.17</td>
<td>12.47</td>
<td>14.03</td>
</tr>
<tr>
<td></td>
<td>(2.39)</td>
<td>(0.82)</td>
<td>(1.37)</td>
<td>(0.57)</td>
<td>(0.62)</td>
<td>(0.85)</td>
</tr>
<tr>
<td>SCAD SVM</td>
<td>8.00</td>
<td>7.90</td>
<td>7.53</td>
<td>6.47</td>
<td>4.73</td>
<td>5.27</td>
</tr>
<tr>
<td></td>
<td>(0.62)</td>
<td>(0.58)</td>
<td>(0.66)</td>
<td>(0.70)</td>
<td>(0.28)</td>
<td>(0.52)</td>
</tr>
</tbody>
</table>

Table 2. Frequency of selecting correct variables in thirty runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>n=20</th>
<th>n=30</th>
<th>n=40</th>
<th>n=50</th>
<th>n=70</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$ SVM</td>
<td>1</td>
<td>4</td>
<td>13</td>
<td>22</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>SCAD SVM</td>
<td>0</td>
<td>7</td>
<td>19</td>
<td>25</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

Fig. 3. Average test error rates plotted against the sample size.

Fig. 4. Bayes rule and the classification boundaries given by four methods (projected onto the first two dimensions) in the simulated example.
5 REAL DATA

In practice, gene-ranking methods are widely used to select genes which are highly differentiated between two types of tissues prior to training. Ranking criteria are often based on the t-statistic (Pan, 2002) or correlation coefficients. For each gene \( x_j \), the mean \( \mu_j^+ \) (resp. \( \mu_j^- \)) and standard deviation \( \sigma_j^+ \) (resp. \( \sigma_j^- \)) using only the tissues labeled +1 (resp. −1) are calculated. Define \( w_j = (\mu_j^+ - \mu_j^-)/ (\sigma_j^+ + \sigma_j^-) \). Golub et al. (1999) selected \( p \) genes with the largest positive \( w_j \)'s and \( p \) genes with the largest negative \( w_j \)'s. Furey et al. (2000) used \( |w_j| \) to select top \( p \) genes. Pavlidis et al. (2001) suggested a Fisher-discriminant type correlation coefficient. We compare our method with two ranking methods: t-test and Furey et al. (2000). Each ranking criteria is first applied to select the top 50 and 100 genes, then the standard SVM is fitted. There are two problems with ranking methods: (1) one has to specify the number of selected genes \( p \) in advance and often subjectively; (2) the selection is individual-based and hence ignores correlation among genes.

UNC Breast Cancer Dataset

Three public microarray gene expression data sets are used in this section. They are from Perou et al. (2000), van’t Veer et al. (2002), Sotiriou et al. (2003), respectively, and for convenience we use “Stanford”, “Rosetta”, and “Singapore” to refer them. Originally the three sets have 5,974 genes and 104 patients, 24,187 genes and 97 patients, and 7,650 genes and 99 patients, respectively. In Hu et al. (2005), the three data sets are imputed for missing values, combined, and then corrected to adjust the bias since they are from three different batches. The DWD is used for the batch adjustment process; see Benito et al. (2004) for a detailed description of the systematic bias adjustment for microarray data using the DWD. As for the gene identifier for the combined data set, UniGene is used since it is most convenient to map the identifiers from each data set to UniGene identifier (Build 161). In case of multiple occurrences of a UCID, the median value is used.

The combined data set has 2,924 genes and totally 300 patients. Our primary interest is to select important genes and use them to classify the tissues into different types of breast cancer. We use the source information to separate the whole data into three folds naturally. We train each classifier on two folds and test it on the remaining one. For example, the SCAD SVM is first trained on Rosetta and Singapore data, then tested on Stanford data. We refer this as “Stanford” learning. Then we repeat this procedure for the other two learnings, referred as the “Rosetta” learning and the “Singapore” learning. To choose the tuning parameter \( \lambda \), we use ten-fold cross validation within the training set.

Table 3 shows the cross validation error in each learning and the average error rate for five methods. In the Stanford learning, the SCAD SVM has the lowest error rate 0.115. In the Rosetta learning, the SCAD SVM and the SVM are equally best. In the Singapore learning, the SVM is best and the SCAD SVM is slightly worse and the second best. Two ranking methods give the same result (hence only t-test reported), and they are worse than the SVM which uses all the genes. It can be explained by that the ranking methods select individual genes separately and ignore their correlations. For a fair comparison, the DWD is not included here because it was used for pre-processing the combined data. Overall speaking, the SCAD SVM gives the lowest error rate among all.

<table>
<thead>
<tr>
<th></th>
<th>Stanford</th>
<th>Rosetta</th>
<th>Singapore</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-test (p=50)</td>
<td>.202</td>
<td>.217</td>
<td>.192</td>
<td>.203</td>
</tr>
<tr>
<td>t-test (p=100)</td>
<td>.192</td>
<td>.206</td>
<td>.111</td>
<td>.170</td>
</tr>
<tr>
<td>SVM</td>
<td>.154</td>
<td>.175</td>
<td>.051</td>
<td>.127</td>
</tr>
<tr>
<td>( L_1 ) SVM</td>
<td>.125</td>
<td>.216</td>
<td>.081</td>
<td>.141</td>
</tr>
<tr>
<td>SCAD SVM</td>
<td>.115</td>
<td>.175</td>
<td>.061</td>
<td>.117</td>
</tr>
</tbody>
</table>

Table 4 gives the number of genes selected in each learning by the SCAD SVM and the \( L_1 \) SVM. The \( L_1 \) SVM selects 59 ~ 72 genes in each learning, where the SCAD SVM only selects 15 ~ 31 genes for each case. Note that the misclassification rates of the SCAD SVM shown in Table 2 are only based on the selected 15 ~ 31 genes, which shows the very strong gene selection power of the method. Also note that the gene selection results of both methods are consistent in the sense that they both select the smallest number of genes for prediction in the Stanford learning, and both select the largest number of genes in the Singapore learning.

<table>
<thead>
<tr>
<th></th>
<th>Stanford</th>
<th>Rosetta</th>
<th>Singapore</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L_1 ) SVM</td>
<td>59</td>
<td>63</td>
<td>72</td>
<td>65</td>
</tr>
<tr>
<td>SCAD SVM</td>
<td>15</td>
<td>19</td>
<td>31</td>
<td>22</td>
</tr>
</tbody>
</table>

Figure 5 lists the UniGene identifiers of all the genes that are selected at least three times by either the SCAD SVM or the \( L_1 \) SVM. In the second and third columns are the frequencies of each gene selected in three learnings by, respectively, the SCAD SVM and the \( L_1 \) SVM. The sum of these two columns is in the fourth column, and the fifth column lists the number of times that each gene selected by the t-test. The sixth column shows whether or not the selected gene is in the list of “intrinsic” genes selected by Perou et al. (2000). The last column displays the corresponding descriptive names of UniGene identifiers. We can see that the top gene Hs.169946
is selected by all the methods in each learning and also classified as an intrinsic gene. Hs.79136 and Hs.80420, both intrinsic genes, are selected three times by the SCAD SVM and intrinsic and also selected by t-test. However, there are 9 out of total 27 selected genes which are neither intrinsic nor selected by t-test. This suggests that one should consider the multivariate gene selection approaches, such as the SCAD SVM and the L1 SVM, rather than individual gene-by-gene methods such as t-test procedures.

Metabolism Dataset

Metabolic datasets contain the quantitative measurements of all small molecule metabolites in biological samples. Some biological studies show that most of the metabolites are not informative in predicting disease or non-disease outcomes (Stitt and Fernie, 2003). Consequently, hybrid methods that incorporate variable selection with classification techniques can be very effective in analyzing datasets of this sort. Our metabolism data set is provided by Metabolon Inc. and we are actually one of the first research groups to analyze it.

There are metabolic profiles of 63 subjects: 32 healthy subjects and 31 subjects diagnosed with a certain disease. Within the patient group, 9 subjects are taking medication and 22 are not. For each sample, its metabolic profile contains the intensity levels of 317 compounds (metabolites). Table 5 shows the average leave-one-out cross validation error and the number of metabolites selected by each method. The SCAD SVM gives the smallest cross validation error 0.143. Moreover, the SCAD SVM selects 18 important metabolites out of 317 and the L1 SVM selects 32 metabolites. Hence the SCAD SVM achieves the highest classification accuracy using the fewest number of metabolites. This result has great implications on metabolic studies, since one main issue from biological aspects is to identify which metabolites are more relevant to the occurrence of the disease.

6 DISCUSSION

For high dimensional low sample size data, redundant input variables can affect the performances of classifiers. How to combine variable selection and classification in a unified framework has become an imminent problem. In this paper, we propose a new regularization technique for simultaneous classification and variable selection in the SVM. Compared with other methods, our nonconvex penalty function achieves more compactness and better accuracy, showing great potential for gene selection in cancer classification problems. The non-convexity of the penalty function introduces greater difficulties in optimization. To address this problem, we have developed an iterative procedure based on the successive quadratic algorithm to implement the SCAD SVM efficiently. In both simulated and real data analysis, we found that this algorithm converges quickly. Overall, the SCAD SVM gives competitive results in terms of both variable selection and classification.

7 ACKNOWLEDGMENT

This research was supported by Statistical and Applied Mathematical Sciences Institute (SAMS) data mining and machine learning program during 2003-2004. Hao Helen Zhang’s research was partially supported by NSF grant DMS-0405913. The authors are grateful to Dr. C. M. Perou and his colleagues for providing breast cancer microarray data sets.
REFERENCES


