An Improved 1-norm SVM for Simultaneous Classification and Variable Selection

Hui Zou
School of Statistics
University of Minnesota
Minneapolis, MN 55455
hzou@stat.umn.edu

Abstract

We propose a novel extension of the 1-norm support vector machine (SVM) for simultaneous feature selection and classification. The new algorithm penalizes the empirical hinge loss by the adaptively weighted 1-norm penalty in which the weights are computed by the 2-norm SVM. Hence the new algorithm is called the hybrid SVM. Simulation and real data examples show that the hybrid SVM not only often improves upon the 1-norm SVM in terms of classification accuracy but also enjoys better feature selection performance.

1 Introduction

We consider the feature selection problem in the support vector machine (SVM) for binary classification. The standard 2-norm SVM (Vapnik 1996) is a widely used classification tool, due to its elegant margin interpretation and highly competitive performance in practice. However, the 2-norm SVM classifier cannot automatically select input features. It is now well known that feature selection is crucial for achieving good classification accuracy if the underlying true model is sparse (Hastie et al. 2001, Friedman et al. 2004). Moreover, in many scientific problems parsimonious models are often preferred, hence feature selection is necessary. Friedman et al. (2004) advocated the *bet-on-sparsity principle* in statistical modeling; namely, procedures that do well in sparse problems should be favored. Fan and Li (2006) gave a comprehensive overview of the importance of feature selection in knowledge discovery.

One way to approach the feature selection problem in classification is to combine a separate feature selection step with the SVM methodology. For example, one could use univariate ranking (Golub et al. 1999) and recursive feature elimination (Guyon et al. 2002) to select a subset of variables and then fit a 2-norm SVM by using the selected subset variables. However, these type of procedures depend on the external feature selection methods. As indicated by the statistical theory developed by Fan and Li (2001, 2006), one could achieve superior performance by doing feature selection and prediction simultaneously. A lot of empirical evidence supports this viewpoint, see Tibshirani (1996), Hastie et al. (2001), Zhu et al. (2003) and Friedman et al. (2004). In the past few years, the 1-norm minimization method for variable selection has attracted a lot of attention. Breiman (1995) invented the non-negative garrote idea which was revisited again recently by Yuan and Lin (2005). Tibshirani (1996) proposed the lasso, a penalized least square method using the 1-norm penalty, for variable selection in linear and generalized linear models. Bradley et al. (1998), Song et al. (2002) and Zhu et al. (2003) considered the 1-norm SVM to accomplish the goal of automatic feature selection in the SVM classifier. The 1-norm SVM penalizes the empirical hinge loss using the lasso (1-norm) penalty. Due to its singularity at the origin (Tibshirani 1996, Fan and Li 2001), the 1-norm penalty is able to shrink some of the coefficients to exact zero. Thus the 1-norm SVM automatically discards irrelevant features by estimating their coefficients by zero. When there are many noise variables, the 1-norm SVM has significant advantages over the 2-norm SVM, for the latter does not select significant variables (Zhu et al. 2003).

In this paper we attempt to further improve upon the 1-norm SVM. As shown in Section 3, the 1-norm SVM often tends to include some noise features in the final model when the underlying model is truly sparse. This phenomenon is closely related to the lack of oracle property of the lasso, as conjectured by Fan and Li (2001) and proven by Zou (2006). Zou (2006) further showed that a modified lasso using the weighted 1-norm penalty could perform as well as if the underlying sparse model were given in advance. Motivated by these empirical and theoretical findings, we propose

a novel extension of the 1-norm SVM by adopting the adaptively weighted 1-norm penalty in the SVM. We construct the adaptive weights using the 2-norm SVM. Thus the new algorithm is called the hybrid SVM. Simulation and real data examples show that the hybrid SVM often outperforms the 1-norm SVM in terms of sparsity and classification accuracy.

The rest of the paper is organized as follows. Section 2 introduces the hybrid SVM methodology. We conduct Monte Carlo simulation to compare the hybrid SVM with the 1-norm and 2-norm SVMs in Section 3. In Section 4 we demonstrate the utility of the hybrid SVM using three benchmark data sets. Section 5 contains some discussion.

2 Methodology

In this section we first briefly discuss the unified representation of the 2-norm and 1-norm SVMs. We then introduce the weighted 1-norm penalty and the hybrid support vector machine.

2.1 Review of the 2-norm and 1-norm SVMs

Let x denote the feature vector. The class labels, y, are coded as $\{1, -1\}$. A classification rule δ is a mapping from x to $\{1, -1\}$ such that a label $\delta(x)$ is assigned to the datum at x. Under the 0-1 loss, the misclassification error of δ is $R(\delta) = P(y \neq \delta(x))$. The smallest classification error is the Bayes error achieved by the Bayes rule $\arg\max_{c\in\{1,-1\}} p(y = c|x)$.

The standard 2-norm SVM finds a hyperplane $(x^T\beta + \beta_0)$ that creates the biggest margin between the training points for class 1 and -1 (Vapnik 1996, Hastie et al. 2001)

$$\begin{aligned} &\max_{\beta,\beta_0} \tfrac{1}{\|\beta\|_2} \\ \text{subject to} \quad &y_i(\beta_0 + x_i^T\beta) \geq 1 - \xi_i, \quad \forall\ i \\ &\xi_i \geq 0, \textstyle\sum \xi_i \leq B, \end{aligned}$$

where ξ_i are slack variables, and B is a pre-specified positive number that controls the overlap between the two classes. The 2-norm SVM has an equivalent *loss + penalty* formulation (Vapnik 1996, Hastie et al. 2001)

$$(\hat{\beta}, \hat{\beta}_0) = \arg\min_{\beta,\beta_0} \sum_{i=1}^{n} \left[1 - y_i(x_i^T\beta + \beta_0)\right]_+ + \lambda\|\beta\|_2^2, \tag{1}$$

where the subscript "+" means the positive part ($z_+ = \max(z, 0)$). The loss function $(1 - t)_+$ is called the hinge loss or SVM loss. Lin (2002) showed that due to the unique property of the hinge loss, the SVM directly approximates the Bayes rule without estimating the conditional class probability.

The 1-norm SVM replaces the 2-norm penalty in (1) with the 1-norm penalty

$$\arg\min_{\beta,\beta_0} \sum_{i=1}^{n} \left[1 - y_i(x_i^T\beta + \beta_0)\right]_+ + \lambda\|\beta\|_1. \tag{2}$$

Note that the 1-norm penalty is not differentiable at zero. This important singularity property ensures that the 1-norm SVM is able to delete many noise features by estimating their coefficients by zero. The 2-norm penalty is differentiable everywhere, thus the 2-norm SVM will use all the input features in classification. When there are many noise variables, the 2-norm SVM suffers severe damage caused by the noise features. Thus the 1-norm SVM is considered a better choice than the 2-norm SVM if the underlying model has a sparse presentation. For more detailed discussion on the advantages of the 1-norm penalty over the 2-norm penalty, the readers are referred to Friedman et al. (2004).

2.2 The hybrid SVM

We are ready to present the technical details of the hybrid support vector machine. Given the n training samples $\{(x_i, y_i)\}_{i=1}^n$, let $\hat{\beta}(\ell_2)$ denote the coefficients in the 2-norm SVM. We define a weight vector as follows

$$w_j = |\hat{\beta}(\ell_2)_j|^{-\gamma} \quad j = 1, \ldots, p \tag{3}$$

where γ is a positive constant. Then the weighted 1-norm penalty is

$$\|\beta\|_{W1} = \sum_{j=1}^{p} w_j|\beta_j|. \tag{4}$$

With such definitions, we propose penalizing the empirical hinge loss by the weighted 1-norm penalty

$$(\hat{\beta}, \hat{\beta}_0) = \arg\min_{\beta,\beta_0} \sum_{i=1}^{n} \left[1 - y_i(x_i^T\beta + \beta_0)\right]_+ + \lambda\|\beta\|_{W1}. \tag{5}$$

The fitted classifier is $\hat{f}(x) = \hat{\beta}_0 + x^T\hat{\beta}$, and the classification rule is $sign(\hat{f}(x))$.

Note that the weighted 1-norm penalty is a data-driven quantity. It is worth noting that if the weights are data-independent, then the weighted 1-norm cannot fix the drawback of the ordinary 1-norm penalty (Zou 2006). The rationale behind the weighted 1-norm penalty is to adaptively penalize each components such that the coefficients of irrelevant variables are shrunken to zero, while reducing the shrinkage bias for the large coefficients of significant variables. Rigorous justification of the usage of the weighted 1-norm penalty is provided in Zou (2006).

2.3 Computation and tuning

The hybrid SVM can be efficiently solved by standard linear programming (LP) software. To derive the LP formulation of the hybrid SVM, we introduce a set of slack variables

$$\xi_i = \left[1 - y_i\left(\sum_{j=1}^{p} x_{ij}^T(\beta_j^+ - \beta_j^-) + \beta_0^+ - \beta_0^-\right)\right]_+ \qquad i = 1, 2, \ldots, n,$$

and we write $\beta_j = \beta_j^+ - \beta_j^-$ where β_j^+ and β_j^- denote the positive and negative parts of β_j, respectively. Then it is not hard to show that the hybrid SVM is equivalent to

$$\arg\min \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{p} |\hat{\beta}(\ell_2)_j|^{-\gamma}(\beta_j^+ - \beta_j^-) \qquad (6)$$

subject to $\forall\ 1 \le i \le n$ and $j = 0, 1, \ldots, p$

$$y_i(\beta_0^+ - \beta_0^- + x_i^T(\beta^+ - \beta^-)) \ge 1 - \xi_i,$$
$$\xi_i \ge 0, \beta_j^+ \ge 0,\ \beta_j^- \ge 0.$$

If $\gamma = 0$, (6) reduces to the 1-norm SVM. Zhu et al. (2003) proposed an efficient algorithm to efficiently solve the 1-norm SVM solution path. Similar algorithm can be used to solve (6) for all λ.

In the hybrid SVM there are two obvious tuning parameters: λ and γ. In principle, there is another tuning parameter, because different regularization parameter in the 2-norm SVM will give different $\hat{\beta}(\ell_2)$, hence different weights for the hybrid SVM. To save computations, we propose the following strategy for selecting the tuning parameters. We first find the best 2-norm SVM classifier based on the training data, then the weights are computed based on the 2-norm SVM classifier. The next step is to apply cross-validation (or validation if an independent validation data set is available) to choose the optimal pair of (λ, γ). In all the numerical examples we choose γ from the set $\{1, 2, 4\}$.

3 Simulation

In this section we conduct simulation experiments to compare the hybrid SVM with the 2-norm and the 1-norm SVMs. We first introduce some notation used in the simulation. We use "C" and "IC" to denote the median number of correctly and incorrectly selected input features, respectively. The term "PPS" stands for the probability of perfect selection (selecting the true model).

3.1 Sparse models : $p < n$

Model 1:Orange data model. In the first simulation example, we considered the "orange data" model in Zhu et al. (2003). We generated 50 observations in each of two classes. The first class ("+") has two independent standard normal inputs x_1, x_2. The second class ("-") also has two standard normal independence inputs, but conditioned on $4.5 \le x_1^2 + x_2^2 \le 8$. To make the classification more difficult, we also included q independent standard normal inputs in the model. let $I(\cdot)$ be the indicator function. The Bayes rule assigns label $1 - 2I(4.5 \le x_1^2 + x_2^2 \le 8)$ to datum $(x_1, x_2, \ldots, x_{2+q})$, thus being independent of the dimension. The Bayes error is 0.0435.

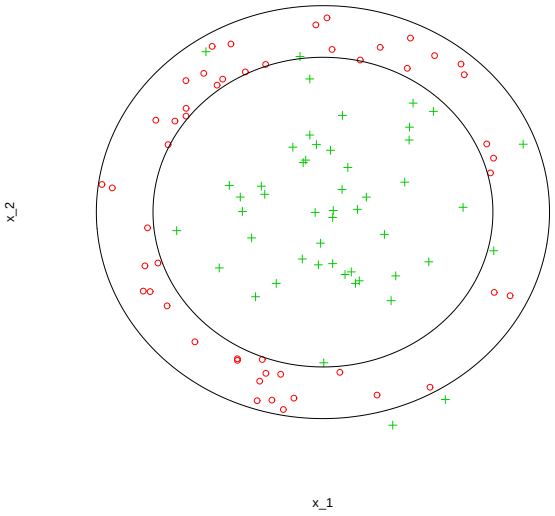

Figure 1: Orange data. The solid black circles are the Bayes decision boundary.

In the original input space, a linear classifier is not sufficient for separating the two classes. We used the enlarged dictionary $D = \{\sqrt{2}x_j, \sqrt{2}x_j x_k, x_j^2, j, k = 1, 2 \ldots, 2+q\}$ to build the SVM classifiers. We also collected an independent validation data set of size 100 for selecting the tuning parameter(s) in each SVM. A test set of 20000 observations was used to evaluate the test error of each SVM. We let q be 2, 4, 6, 8. For each q, the simulation was repeated 500 times.

Table 1 compares the classification accuracy of three SVMs in model 1. Numbers in parentheses are the standard errors. Several observations can be made from Table 1. Both the hybrid SVM and the 1-norm SVM dominates the 2-norm SVM by a good margin. When the number of noise features, q, increases, the classification error of the 2-norm SVM increases

quickly. In contrast, the errors of the hybrid and 1-norm SVMs are much more robust with respect to the value of q. We also see that the hybrid SVM is significantly more accurate than the 1-norm SVM.

q	p	2-norm	1-norm	hybrid
2	14	9.97(0.09)%	8.00(0.04)%	7.27(0.04)%
4	27	12.87(0.11)%	8.21(0.04)%	7.45(0.04)%
6	44	16.17(0.14)%	8.42(0.01)%	7.63(0.05)%
8	65	19.21(0.15)%	8.52(0.06)%	7.65(0.05)%

Table 1: Orange data model: $p < n$, note p is the size of the enlarged dictionary. Compare the misclassification errors of the hybrid SVM, 2-norm and 1-norm SVMs.

Table 2 summarizes the feature selection results by the 1-norm and hybrid SVMs. It indicates that the 1-norm tends to select a few noise features into its final model, but the hybrid SVM has greater tendency to discard all the noise features. The perfect variable selection means that all the true features are selected and all the noise features are eliminated. When q exceeds 12, there are about 100 predictors in the enlarged dictionary. It is very difficult for the 1-norm SVM to exactly identify the ground truth. However, the hybrid SVM consistently have pretty high probabilities of perfect selection.

q	p	1-norm			hybrid		
		C	IC	PPS	C	IC	PPS
2	14	2	3	0.21	2	0	0.72
4	27	2	3	0.19	2	0	0.72
6	44	2	4	0.11	2	0	0.69
8	65	2	5	0.12	2	0	0.71

Table 2: Orange data model, $p < n$. Compare the variable selection results of the hybrid SVM and the 1-norm SVM.

Orange data example clearly shows the benefits of using the weighted 1-norm penalty in the SVM classification. We now consider more sparse models with various correlation structure among predictors.

Model 2. We simulated a training data set consisting of 100 observations from the model $y \sim$ Bernoulli$\{p(x^T\beta + \beta_0)\}$, where $p(u) = \exp(u)/(1 + \exp(u))$. We let $\beta = (3, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 3)$ and $\beta_0 = 0$. The components of x are standard normals, where the correlation between x_i and x_j is ρ. We considered both $\rho = 0.5$ and $\rho = 0$. Note that the Bayes rule is to assign datum $(x_1, \ldots, x_{12})$ to class $2I(x_1 + x_6 + x_{12}) - 1$.

Model 3. We simulated a training data set consisting of 100 observations from the model $y \sim$ Bernoulli$\{p(x^T\beta + \beta_0)\}$, where $p(u) = \exp(u)/(1 + \exp(u))$. We let $\beta = (3, 2, 0, 0, 0, 0, 0, 0, 0)$ and $\beta_0 = 1$. The components of x are standard normal, where the correlation between x_i and x_j is $\rho^{|i-j|}$. We considered both $\rho = 0.5$ and $\rho = 0$. Note that the Bayes rule is to assign datum $(x_1, \ldots, x_9)$ to class $2I(3x_1+2x_2+1)-1$.

Model 2	$\rho = 0.5$ Error %	$\rho = 0$ Error %
2-norm	10.84(0.07)	13.85(0.07)
1-norm	9.99(0.08)	12.85(0.08)
hybrid	9.32(0.08)	11.69(0.06)
Bayes	7.38	10.19
Model 3	$\rho = 0.5$ **Error %**	$\rho = 0$ **Error %**
2-norm	14.36(0.07)	16.54(0.07)
1-norm	13.21(0.07)	15.29(0.07)
hybrid	12.77(0.06)	14.82(0.06)
Bayes	11.89	13.75

Table 3: Simulation model 2 and model 3. Compare classification performance.

Model 2	$\rho = 0.5$			$\rho = 0$		
	C	IC	PPS	C	IC	PPS
1-norm	3	2	0.15	3	2	0.25
hybrid	3	0	0.62	3	0	0.74
Model 3	$\rho = 0.5$			$\rho = 0$		
	C	IC	PPS	C	IC	PPS
1-norm	2	0	0.54	2	2	0.31
hybrid	2	0	0.77	2	0	0.74

Table 4: Simulation model 2 and model 3. Compare feature selection performance.

In both models we collected an independent validation data set of size 100 for selecting the tuning parameter(s) in each SVM. A test set of 20000 observations was used to evaluate the test error of each SVM. The simulation was repeated 500 times. Tables 3 and 4 present the simulation results of model 2 and model 3. In terms of classification accuracy, the 1-norm SVM and the hybrid SVM dominate the 2-norm SVM, and the hybrid SVM significantly outperforms the 1-norm SVM. We observe again that the hybrid SVM has greater tendency to discard all the noise features than the 1-norm SVM. The hybrid SVM consistently has much higher probabilities of perfect selection than the 1-norm SVM.

3.2 Sparse models: $p > n$

In this section we present more simulation experiments to compare the performance of the hybrid SVM and

the 1-norm and 2-norm SVMs in the high-dimensional setting.

Model 4. We used the orange data model again with $q = 12, 16$ and 20. The corresponding number of predictors is $119, 189$ and 275.

q	p	2-norm	1-norm	hybrid
12	119	24.30(0.14)%	8.65(0.05)%	7.66(0.05)%
16	189	27.81(0.14)%	8.61(0.05)%	7.74(0.06)%
20	275	30.40(0.13)%	8.68(0.06)%	7.78(0.07)%

Table 5: Orange data model, $p > n$. Compare the misclassification errors of the hybrid SVM, 2-norm and 1-norm SVMs.

q	p	1-norm			hybrid		
		C	IC	PPS	C	IC	PPS
12	119	2	5	0.064	2	0	0.668
16	189	2	4	0.036	2	0	0.652
20	275	2	6	0.016	2	0	0.566

Table 6: Orange data model $p > n$. Compare the variable selection results of the hybrid SVM and the 1-norm SVM.

Table 5 and Table 6 summarize the simulation results of model 4. We see that the 2-norm SVM suffers from the high-dimensionality, while the 1-norm SVM and the hybrid SVM overcome the curse-of-dimensionality by automatically deleting most of the noise variables. It is also worth emphasizing that even when p greatly exceeds n, the hybrid SVM is still capable of selecting the true model with a high probability. It is also worth mentioning here that the hybrid SVM achieves the remarkable performance only using 50 observations per class. This example clearly demonstrates the advantages of the hybrid SVM over the 1-norm and 2-norm SVMs.

Model 5. We simulated a training data set consisting of 100 observations from the model $y \sim$ Bernoulli$\{p(x^T\beta + \beta_0)\}$ with $\beta = (3, 2, 0, 0, 0, 0, 0, 0, 0)$ and $\beta_0 = 1$. The components of x are standard normal, where the correlation between x_i and x_j is $0.5^{|i-j|}$. We then included 300 independent normal variables as noise features in the predictor set. Thus $p = 309$ and $n = 100$. We repeated the simulation 500 times.

From Table 7 we see that the performance of the 2-norm SVM is severely damaged by the added noise variables, while the 1-norm and hybrid SVMs are much more robust against the noise features. The hybrid SVM does better than the 1-norm SVM in identifying the true model.

Model 5	Error %	C	IC	PPS
2-norm	34.21(0.13)			
1-norm	14.79(0.10)	2	2	0.33
hybrid	14.69(0.13)	2	0	0.50
Bayes	11.89			

Table 7: Results of simulation model 5.

3.3 A dense model

We have seen that in sparse models, the hybrid and 1-norm SVMs dominate the 2-norm SVM in terms of classification accuracy. To have a more complete picture, we need to compare the three SVMs in the case where the true model is dense.

Model 6. We used the same setup in model 3, except that we let $\beta = (3, 3, 3, 3, 3, 3, 3, 3)$ and $\beta_0 = 0$. The components of x are independent standard normal. All eight input features contribute equally in this model.

Bayes	2-norm	1-norm	hybrid
6.36%	8.75(0.05)%	9.44(0.07)%	9.42(0.08)%

Table 8: Simulation model 4. Compare the misclassification errors of the hybrid SVM, 2-norm and 1-norm SVMs.

As can be seen from Table 8, the 2-norm SVM performs the best in this model. This example shows the value of quadratic regularization. On the other hand, the 1-norm SVM and the hybrid SVM only lose a little classification accuracy, and they have tremendous advantages over the 2-norm SVM in the sparse setting. Our simulation shows that in classification problems, the bet on sparsity principle is a good rule to follow.

4 Real Data Examples

The simulation study has demonstrated the promising advantages of the hybrid SVM. We now examine the performance of the hybrid SVM on several real data examples. We considered three benchmark data sets obtained from UCI Machine Learning Repository (Newman & Merz 1998). Note that there are a large number of predictors in these benchmark data sets. `Spam` data contains a training set and a test set. The test set has 1536 observations and test indicators can be downloaded from http://www-stat.stanford.edu/~tibs/ElemStatLearn/. For `ionosphere` and `WDBC` data, we randomly selected 2/3 data for training and the other 1/3 data as the test set. For each SVM, fitting and tuning were done only on the training set, and the classification error was computed using the test set. We repeated the

randomization ten times.

	2-norm Error %	1-norm Error %	hybrid Error %
spam	7.31(0.66)	7.37(0.66)	7.08(0.65)
ionosphere	13.81(0.83)	13.47(0.80)	12.16(0.72)
WDBC	3.26(0.31)	3.37(0.25)	3.29(0.22)

Table 9: Three benchmark data sets: compare classification performance.

	# of predictors	1-norm NSV	hybrid NSV
spam	57	55	41
ionosphere	34	11	8
WDBC	30	11	8

Table 10: Three benchmark data sets: compare feature selection performance. "NSV" is the number of selected variables.

Table 9 compares the three SVMs. In terms of classification error, the hybrid SVM seems to perform slightly better than the 2-norm and 1-norm SVMs.

As can be seen from Table 10, the feature selection results are very interesting. In `spam` data the 1-norm SVM seems to offer little improvement over the 2-norm SVM. The hybrid SVM is able to delete 16 variables, and at the same time, has a smaller classification error than both the 2-norm and 1-norm SVMs. In `WDBC` and `ionosphere` data both the hybrid SVM and the 1-norm SVM greatly reduce the number of used features. The hybrid SVM still uses less features and produces a more accurate classifier than the 1-norm SVM.

5 Discussion

We have proposed the hybrid SVM for simultaneous classification and feature selection. The hybrid SVM is a two stage algorithm. At the first stage, we use the coefficients of a 2-norm SVM classifier to construct the weights in the weighted 1-norm penalty. Then we solve the weighted 1-norm SVM. The 1-norm SVM performs better than the 2-norm SVM when there are many noise features. By using the adaptively weighted 1-norm penalty, the hybrid SVM often more efficiently eliminates the noise features while reducing the shrinkage bias on the significant variables. As a result, the hybrid SVM classifier is often more accurate than the 1-norm SVM. When the underlying model is sparse, the hybrid SVM can identify the exact subset model with a much higher probability than the 1-norm SVM. Thus we regard the hybrid SVM as an improved 1-norm SVM.

In this paper we have used the inverse power function to construct the weights in the weighted 1-norm penalty. As pointed out in Zou (2006), one could compute the weights by using many other functions. For example, let $f(t)$ be a positive continuous function on $(0, \infty)$ such that $\lim_{t\to 0+} \frac{f(t)}{t^s} = \infty$ for some $s > 0$, then we can construct the weights by $w_j = f(|\hat{\beta}(\ell_2)_j|)$. The choice of the weighting function is not critical for large samples. For finite or small samples, we have suggested using cross-validation to choose the weighting function. This strategy worked quite well in our experiments.

References

[1] Breiman, L. (1995), Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

[2] Bradley, P. & Mangasarian, O. (1998), Feature selection via concave minimization and support vector machines, *in* J. Shavlik, ed., ICML'98, Morgan Kaufmann.

[3] D.J. Newman, S. Hettich, C.B. & Merz, C. (1998), UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[4] Fan, J. & Li, R. (2001), Variable Selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**, 1348-1360.

[5] Fan, J. & Li, R. (2006), Statistical challenges with high dimensionality: Feature selection in knowledge discovery, *Proceedings of the Madrid International Congress of Mathematicians 2006*, to appear.

[6] Friedman, J., Hastie, T. Rosset, S., Tibshirani, R. & Zhu, J. (2004), Discussion of boosting papers, *Annals of Statistics* **32**, 102-107.

[7] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J. & Caligiuri, M. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-536.

[8] Guyon, I., Weston, J., Barhill, S. & Vapnik, V. (2002), Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389-422.

[9] Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer Verlag, New York.

[10] Lin, Y. (2002), Support vector machines and the Bayes rule in classification, *Data Mining and Knowledge Discovery* **6**, 259-275.

[11] Song, M., Breneman, C., Bi, J., Sukumar, N., Bennett, K., Cramer, S. & Tugcu, N. (2002), Prediction of protein retention times in anion-exchange chromatography systems using support vector regression, *Journal of Chemical Information and Computer Sciences*, September.

[12] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

[13] Vapnik, V. (1996), *The Nature of Statistical Learning*, Springer Verlag, New York.

[14] Yuan, M. and Lin, Y. (2005), On the Nonnegative Garrote Estimator, Statistics Discussion Paper 2005-25, School of Industrial and Systems Engineering, Georgia Institute of Technology.

[15] Zhu, J., Rosset, S., Hastie, T. & Tibshirani, R. (2003), 1-norm support vector machines, *Neural Information Proceeding Systems 16*.

[16] Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, 101(476) 1418-1429.