International Journal of Pure and Applied Mathematics

Volume 35 No. 2 2007, 251-258

OPTIMAL CONDITIONS AND NUMERICAL EXPERIMENTS IN PARAMETER SELECTION PROBLEMS OF SVM

Yu-Lin Dong[1 §], Yun Wang[2], Zun-Quan Xia[3]

[1,2,3]Department of Applied Mathematics
Dalian University of Technology
Dalian, 116024, P.R. CHINA
[1]e-mail: dyl@student.dlut.edu.cn
[2]e-mail: wangyun_3412@163.com
[3]e-mail: zqxiazhh@dlut.edu.cn

Abstract: Support vector machine (SVM) is a widely used tool for classification. In this paper, we present a new SVM model to calculate the optimal value of cost parameter C for particular problems of linearity non-separability of data. The new SVM model is formulated in the form of one of MPEC problems with an integer objective function. A lower bound, positive number, C_0 is required to provide for avoiding choosing a candidate set of C. Numerical experiments show that this model for choice of C is suitable for solving SVM problems.

AMS Subject Classification: 90C90
Key Words: data mining, support vector machine, cost parameter, nonlinear programming, MPEC problem

1. Introduction

Consider a support vector machine (SVM) classifier for the binary classification setting. Given a set of training data $T = \{x_1, x_2, \ldots, x_m\} \in R^n$ along with labels $\{y_1, y_2, \ldots, y_m\} \in \{1, -1\}$, we aim to find a linear function of the form

Received: December 26, 2006

© 2007, Academic Publications Ltd.

[§]Correspondence author

$f(x) = w^T x + b$, where $w \in R^n$ and $b \in R$, such that a new data x is assigned to a label +1 if $f(x) > 0$, and a label -1 otherwise. The SVM classifier is determined by w and b which can be obtained by solving the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \ i = 1, \cdots, m, \quad (1.1)$$

where $\xi_i \geq 0, 1 \leq i \leq m$ are the slack variables to allow some classification errors and C is the so-called cost parameter to control the balance between the 'margin' and classification error.

The performance of a SVM classifier depends on the selection of C, see [10]. The value of C is usually pre-defined, or determined by a tuning procedure, see [4]. Recently, Schittkowski [9] has proposed a two-level approach to choose optimal SVM parameters for support vector machines with L_2-norm hinge loss function.

In this paper, we consider an efficient two-level approach for optimizing the cost parameter in the standard SVM model (1.1). At the first level, SVM classifiers are constructed based on some training data. Due to the L_1-norm measurement of loss, we propose a new nonlinear programming technique to minimize the classification error by choosing a cost parameter C at the second level. In order to tackle the nonsmoothness in our model, we approximate the objective function by a smoothing function. More important, numerical experimental results show that the cost parameter C given by our approach has significantly improved the accurancy of SVM classifiers, comparing to a pre-assigned C in conventional method.

This paper is organized as follows. In Section 2, we state our SVM formulation. In Section 3, we establish the existence of a solution to the formulation. In Section 4, we report results of numerical experiments. In Section 5, we conclude the paper with some remarks.

2. Mathematical Model

We choose two subsets, denoted by A and B, from a training data set T. Set A is used for constructing a SVM classifier and set B for evaluating the classifier. The goal is to find a SVM classifier such that the classification error based on set B is minimized. Note that the accuracy of the classifier depends on the cost parameter. Here, we treat it as a variable instead of a parameter.

At the first level, the cost parameter C is fixed. Define $\mathcal{A} = \{i \,|\, x_i \in A\}$ and $\mathcal{B} = \{i \,|\, x_i \in B\}$. The SVM solutions based on set A determine a set of classifiers, which is defined by

$$\mathcal{S}(C) = \arg\min_{w,b,\xi} \left\{ \frac{1}{2}\|w\|_2^2 + C\sum_{i\in\mathcal{A}} \xi_i \,\middle|\, y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i \in \mathcal{A} \right\}.$$

At the second level, an optimal C is chosen for minimizing the classification error based on the set B. In this paper, the classification error is measured by the number of data points being wrongly classified. If $x_k, k \in \mathcal{B}$, is correctly classified, then $-y_k(w^T x_k + b) < 0$. If nonnegative variables z_k are introduced, the problem can be summarized as follows.

$$\begin{array}{rl} \min\limits_{w,b,\xi,z,C} & \sum\limits_{k\in\mathcal{B}} (z_k)_* \\ \text{s.t.} & (w,b) \in \mathcal{S}(C), \\ & z_k \geq -y_k(w^T x_k + b), \quad k \in \mathcal{B}, \\ & z_k \geq 0, \quad k \in \mathcal{B}, \\ & C \geq C_0, \end{array} \tag{2.1}$$

where $(z_k)_*$ is a step function, $(z_k)_* = 1$, if $z_k > 0$, 0, otherwise.

Since $(z_k)_*$ in (2.1) is not differentiable, gradient-based nonlinear programming techniques cannot be used for solving problem (2.1). Instead of solving (2.1) directly, we approximate function $(z_k)_*$ by a differentiable function introduced in [6] as $t(x,\beta) := 1 - e^{-\beta x}$, where $\beta > 0,\ x \geq 0$. Now, problem (2.1) can be approximately solved by

$$\min_{w,b,\xi,z,C,\alpha,\mu} \sum_{k\in\mathcal{B}} -e^{-\beta z_k} \text{ s.t. constraints in (2.1).} \tag{2.2}$$

3. Optimality Conditions

The model (2.2) is also called mathematical program with equilibrium constraints (MPECs), in which the essential constraint $w \in \mathcal{S}(C)$ is defined by a parametric quadratic programming.

A general MPEC problem can be stated below, see [3]:

$$\min_y \{F(x,y) \,|\, x \in \Psi(y), y \in Y\}, \tag{3.1}$$

where, $F : R^n \times R^m \to R,\ Y \subseteq R^m$, is a closed set, $\Psi(y)$ is defined as follows:

$$\Psi(y) = \operatorname{Arg}\min_x \{f(x,y) \,|\, g(x,y) \leq 0,\ h(x,y) = 0\}, \tag{3.2}$$

where, $f : R^n \times R^m \to R,\ g : R^n \times R^m \to R^p,\ h : R^n \times R^m \to R^q$, $g(x,y) = (g_1(x,y), \cdots, g_p(x,y))^T,\ h(x,y) = (h_1(x,y), \cdots, h_q(x,y))^T$ are sufficiently smooth (vector valued functions).

The following assumptions are used in this paper:

(A1) The set $\{(x,y) \in R^n \times R^m \,|\, g(x,y) \leq 0,\ h(x,y) = 0\}$ is non-empty and compact.

(A2) (MFCQ) We say that (MFCQ) is satisfied at (x^0, y^0) if there exists a direction $d \in R^n$ satisfying

$$\begin{array}{ll} \bigtriangledown_x g_i(x^0,y^0)d < 0, & \text{for each } i \in I(x^0,y^0) := \{j \,|\, g_j(x^0,y^0) = 0\}, \\ \bigtriangledown_x h_j(x^0,y^0)d = 0, & \text{for each } j = 1, \cdots, q, \end{array}$$

and the gradients $\{\bigtriangledown_x h_j(x^0,y^0) \,|\, j = 1, \cdots, q\}$ are linearly independent.

Now we give the optimality conditions of problem (2.2).

Proposition 3.1. *Let (x^0, y^0) be a local optimal solution of problem (2.2). Let (A1) and (A2) be satisfied for the lower level problem at all points $x \in \Psi(y^0)$. Then, there exist $k_0 \geq 0, r^j \geq 0, \mu_i \geq 0, \lambda_i \geq 0, \varphi_k \geq 0, \psi_k \geq 0, \sigma \geq 0$, such that*

$$k_0 + \sum_{j \in \mathcal{A}} r^j + \sum_{i \in \mathcal{A}} (\lambda_i + \mu_i) + \sum_{k \in \mathcal{B}} (\varphi_k + \psi_k) + \sigma = 1,$$

and there exist $\xi^j \in S(C)$, $j = 1, \cdots, m$, such that

$$\begin{array}{l} rC - \mu_i - \lambda_i = 0, \quad \forall i \in \mathcal{A}\,, \\ k_0 \beta z_k^* \exp(-\beta z_k^*) - \varphi_k - \psi_k = 0, \quad \forall k \in \mathcal{B}\,, \\ rw^0 - \sum\limits_i \mu_i x_i y_i - \sum\limits_k \varphi_k x_k y_k = 0, \quad \forall i \in \mathcal{A}, \forall k \in \mathcal{B}\,, \\ \sum\limits_{i \in \mathcal{A}} \mu_i y_i + \sum\limits_{k \in \mathcal{B}} \varphi_k y_k = 0, \quad \forall i \in \mathcal{A}, \forall k \in \mathcal{B}\,, \\ \sum\limits_{j \in \mathcal{A}} (\sum\limits_{i \in \mathcal{A}})\xi_i - \sum\limits_{i \in \mathcal{A}} \xi^j) = \sigma, \quad \forall i \in \mathcal{A}, \\ r = \sum\limits_{i \in \mathcal{A}} r^i. \end{array}$$

Proof. Let $q = (w, b, \xi, z, C)$,

$$\begin{array}{l} v(C) = \min\limits_{w,b,\xi} \{\frac{1}{2} w^T w + C \sum\limits_{i \in \mathcal{A}} \xi_i \,|\, y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i \in \mathcal{A}\}, \\ L(q) = \frac{1}{2} w^T w + C \sum\limits_{i \in \mathcal{A}} \xi_i + \sum\limits_{i \in \mathcal{A}} l_i(-y_i(w^T x_i + b) + 1 - \xi_i) - \sum\limits_{i \in \mathcal{A}} m_i \xi_i. \end{array}$$

We have

$$\begin{aligned}\inf_{(w,b,\xi)\in\mathcal{S}(C)}\inf_{l,m} D_{z,C}L(w,b,\xi,z^*,C^*)(d,h)&\\ \le V'_-(z^*,C^*,d,h) \le V'_+(z^*,C^*,d,h)&\\ \le \inf_{(w,b,\xi)\in\mathcal{S}(C)}\sup_{(l,m)} D_{z,C}L(w,b,\xi,z^*,C^*)(d,h)\,,&\end{aligned}$$

i.e.

$$V'(z^*,C^*,d,h)=\inf_{(w,b,\xi)\in\mathcal{S}(C)}\sum_{i\in\mathcal{A}}\xi_i d=\inf_{\xi\in\mathcal{S}(C)}\sum_{i\in\mathcal{A}}\xi_i d\,.$$

Problem (2.1) is equivalent to

$$\begin{array}{ll}\min\limits_{q} & \sum\limits_{k\in\mathcal{B}}\exp(-\beta z_k)\\ \text{s.t.} & \frac{1}{2}w^Tw+C\sum\limits_{i\in\mathcal{A}}\xi_i-V(z,C)\le 0,\\ & y_i(w^Tx_i+b)\ge 1-\xi_i,\\ & \xi_i\ge 0,\\ & z_k+y_k(w^Tx_k+b)\ge 0,\quad k\in\mathcal{B},\\ & z_k\ge 0,\quad k\in\mathcal{B},\\ & C\ge C_0.\end{array}\tag{3.3}$$

Let

$$\begin{aligned}\varphi(q)=\max\{&-\sum_{k\in B}\exp(-\beta z_k)+\sum_{k\in B}\exp(-\beta z_k^*),1-\xi_i-y_i(w^Tx_i+b),\\ &-\xi_i,-z_k-y_k(w^Tx_k+b),\ -z_k,\alpha_0-\alpha,\frac{1}{2}w^Tw+C\sum_{i\in\mathcal{A}}\xi_i-V(z,C)\}.\end{aligned}$$

Obviously, $\varphi\ge 0$, and the optimal solution of (1.1) $q^*\in\arg\min\varphi$. Thus, the original problem transformed into $\min\varphi(q)$, satisfy $\varphi'(q^*)(S)\ge 0$, $\forall S\in R^{n+1+|A|+|B|+1}$ at local optimal point. Let $f(q)=-\sum\limits_{k\in B}\exp(-\beta z_k)$, $f_1(q)=1-\xi_i-y_i(w^Tx_i+b)$, $f_2(q)=-\xi_i$, $f_3(q)=-z_k-y_k(w^Tx_k+b)$, $f_4(q)=-z_k$, $f_5(q)=\alpha_0-\alpha$, $f_6(q)=\frac{1}{2}w^Tw+C\sum\limits_{i\in\mathcal{A}}\xi_i-V(z,C)$. Then we have

$$\begin{aligned}0\le\max\{&Df(q^*)S,Df_1(q^*)S,Df_2(q^*)S,Df_3(q^*)S,\\ &Df_4(q^*)S,Df_5(q^*)S,Df_6(q^*)S\}\le\sigma(S,\Phi),\end{aligned}$$

where,

$$\Phi=\operatorname{conv}\{Df(q^*),Df_1(q^*),Df_2(q^*),Df_3(q^*),Df_4(q^*),Df_5(q^*),Df_6(q^*)\}.$$

Datasets size	C	TTRC(%)	TTEC(%)	C	TTRC(%)	TTEC(%)
WPBC(24m)	1.0	82.87	80.09	0.7	82.87	80.75
WPBC(60m)	1.0	75.83	63.64	4.232395	76.06	66.37
Ionosphere	1.0	92.66	86.02	0.128388	89.84	87.73
Cleveland	1.0	85.48	83.18	0.133076	85.78	84.53
Wine	1.0	98.04	96.92	0.107698	98.29	97.69
BUPA Liver	1.0	69.79	68.12	13.121832	71.30	70.15
Tic-tac-toe	1.0	66.24	65.45	2.161537	72.29	70.69
Sonar	1.0	83.71	77.43	2.014522	85.31	80.29

Table 1: Numerical results. TTRC: Tenfold cross validation training correctness. TTEC: Tenfold cross validation testing correctness

By $\sigma(S, \Phi) \geq 0$, we get $0 \in \Phi$, and by Carathéodory's Theorem [8], we get the result. □

4. Numerical Experiments

In this section, we modify a method proposed in [5] to solve problem (2.2). We demonstrate now the effectiveness of this approach by comparing it numerically with the model that C takes the default value 1.0. All our experiments are run on the Intel(R) AT/AT Compatible with CPU 3.0G and 2GM RAM. We ran all tests on eight publicly available datasets: the Wisconsin Prognostic Breast Cancer Database and six datasets, Ionosphere, Cleveland Heart Problem, Wine, Bupa Liver, Tic-tac-toe and Sonar from the Irvine Machine Learning Database Repository. For wine data set, we select class 1 and class 2 for our numerical experiment. We randomly extract 10% points from the training set as a surrogate of the testing set, and the rest as the training set. We assign 0.1 to C_0 and use our model to get the optimal cost parameter C. We performed tenfold cross-validation on each dataset and use tenfold training correctness and tenfold testing correctness to evaluate how well the cost parameter C generalizes to future data. We randomly divide the every data set into ten fold data sets firstly, and then do numerical experiment with C taking the default value 1.0 and use the value got by our model respectively. We summarize all these results in Table 1.

It is obvious that our testing correctness is higher for any datasets and our training correctness is higher for any one among the eight datasets with the exception of the third dataset listed in Table 1. For example, for Tic-tac-toe dataset, its testing correctness is 70.69 when C takes 2.161537, and its testing correctness is 65.45 when C takes the default value 1.0, which is 5 percent higher than the one obtained when C takes 1.0. At the same time, there are three datasets whose testing correctness are over 2.0 percent higher than the ones calculated when C takes the default value.

5. Final Remarks

In this paper we have proposed an approach to predetermine optimal cost parameter of a support vector machine. Numerical experiments have shown that this method is efficient for support vector machine. Future work includes the improvement of our algorithm as well as choosing the kernel parameters using this method in nonlinear support vector machine, which seems to be more efficient and promising.

References

[1] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, *Knowledge Discovery and Data Mining*, **2** (1998), 21-167.

[2] N. Cristianini, J. Shwve-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge (2000).

[3] S. Dempe, *Foundations of Bilevel Programming*, Kluwer Academic Publishers, The Netherlands (2002).

[4] Y.J. Lee, O.L. Mangasarian, SSVM: A smooth support vector machine, *Computational Optimization and Applications*, **20** (2001), 5-22.

[5] G.H. Lin, M. Fukushima, New relaxation method for mathematical programs with complementarity constraints, *Journal of Optimization Theory and Applications*, **118** (2003), 81-116.

[6] O.L. Mangasarian, Machine learning via polyhedral concave minimization, *Mathematical Programming Technical Report 95-20*, November 1995; In: *Applied Mathematics and Parallel Computing – Festschrift forKlaus Ritter* (Ed-s: H. Fischer, B. Riedmueller, S. Schaeffler) Physica-Verlag, Germany (1996), 175-188.

[7] J. Nocedal, S.J. Wright, *Numerical Optimization*, Springer, New York (1999).

[8] R.T. Rockefeller, *Convex Analysis*, Princeton University Press, Princeton (1970).

[9] K. Schittkowski, Optimal parameter selection in support vector machines, *Journal of Industrial and Management Optimization*, **1** (2005), 465-476.

[10] H. Trevor, T. Saharon, T. Robert, Z. Ji, The entire regularization path for the support vector machine, *Journal of Machine Learning Research*, **5** (2004), 1391-1415.

[11] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-verlag, New York (1996).