The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 351–357

# An MPEC Model for Selecting Optimal Parameter in Support Vector Machines

Yu-Lin Dong[1,*] Zun-Quan Xia[2,†]
Ming-Zheng Wang[2,3,‡]

[1] College of Information Science and Technology,
Shandong University of Science and Technology, Qingdao 266510, China
[2] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
[3] Management School, Dalian University of Technology, Dalian 116024, China

**Abstract** In this paper, we present a new MPEC model for calculating the optimal value of cost parameter $C$ for particular problems of linear non-separability of data. The objective function of the new model is an integer lower semi-continuous one. Smoothing technique is employed for solving this model, and the relationship between the MPEC model and its associated smoothing problem is given. It is proved that one of the global solution of the smoothing problem is also a solution of the MPEC problem. Numerical experiments show that this model is more efficient for choosing the parameter $C$.

**Keywords** Support vector machine; cost parameter; MPEC problem; nonsmooth optimization.

## 1 Introduction

Consider a support vector machine (SVM) [2][3] [4][10] 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 decision function of the form $f(x) = w^T x + b$, 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:

$$\begin{array}{ll} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|_2^2 + C\sum\limits_{i=1}^{m}\xi_i \\ \text{s.t.} & y_i(w^T x_i + b) \geq 1 - \xi_i,\ i = 1, \cdots, m, \\ & \xi_i \geq 0, \qquad i = 1, \cdots, m, \end{array} \tag{1}$$

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.

*Email:dyulin@sina.com
†Email:zqxiazhh@dlut.edu.cn
‡Email:wangmzhm@163.com

The value of $C$ is usually pre-defined, or determined by a tuning procedure [6]. In the latter case, a candidate set $V$ of $C$ and a tuning set are needed. A typical candidate set $V$ is composed of finite positive real numbers in an ascending order. A tuning set can be selected from the training set. For each $C$ in the candidate set $V$, a SVM classifier is constructed and the correctness of the classifier is computed. The selected $C$ is the one maximizing classification correctness for the tuning set. However, It is not easy to determine a candidate set $V$. There is no criteria to choose a proper set $V$. If the size of set $V$ is too large, the tuning procedure may take too much time. If the size of set $V$ is too small, a proper $C$ may not be found.

Recently, Schittkowski [9] has proposed a two-level approach to choose optimal SVM parameters. Different from the standard SVM model (1), the SVM model discussed in [9] uses $L_2$-norm measuring the hinge loss of misclassification. Due to the measurement of loss in $L_2$-norm, a gradient-based optimization method can be used to search an optimal cost parameter. However, Studies [3] show that measuring loss with $L_1$-norm gives smaller classification error and is particularly suitable for some types of data. In this case, derivatives of the functions do not exist, and hence, gradient-based techniques cannot be applied.

In this paper, we consider an efficient bilevel approach for optimizing the cost parameter in the standard SVM model (1). It is formulated in the form of one of bilevel programming problems with an integer objective function. In order to tackle the nonsmoothness of the objective function, we approximate the objective function by a smoothing function, and we proved that an exact solution of nonsmooth model can be obtained from a solution of the smoothed model for a finite value of the smoothing parameter.

This paper is organized as follows. In Section 2, an MPEC model for choosing the cost parameter is presented. In Section 3, a concave approximation is chosen for solving the MPEC model, and the relationship between the solutions of approximation problem and MPEC model is given. In Section 4, results on numerical experiments are reported.

## 2 MPEC Model

We choose two subsets, denoted by $A$ and $B$, from the 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. Here, we treat the cost parameter $C$ as a variable instead of a parameter.

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

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

At the upper level, in the objective function, the optimal $C$ is need to be 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 $x_k, k \in \mathscr{B}$, 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,

$$\begin{array}{ll} \min\limits_{w,b,\xi,z,C} & \sum\limits_{k\in\mathscr{B}} (z_k)_* \\ \text{s. t.} & (w,b) \in \mathscr{S}(C), \\ & z_k \geq -y_k(w^T x_k + b), \ \ k \in \mathscr{B}, \\ & z_k \geq 0, \ \ k \in \mathscr{B}, \\ & C \geq C_0, \end{array} \tag{2}$$

where $(z_k)_*$ is a step function,

$$(z_k)_* = \begin{cases} 1 & z_k > 0; \\ 0 & \text{otherwise.} \end{cases}$$

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

In order to solve (2), we consider the KKT conditions of (1). The Lagrangian function of (1) is defined as

$$L_p \equiv \frac{1}{2}\|w\|^2 + C\sum_{i\in\mathscr{A}} \xi_i - \sum_{i\in\mathscr{A}} \alpha_i\{y_i(w^T x_i + b) - 1 + \xi_i\} - \sum_{i\in\mathscr{A}} \mu_i\xi_i, \tag{3}$$

where the $\alpha_i, \ i \in \mathscr{A}$, are the Lagrange multipliers introduced to the inequality constraints $y_i(w^T x_i + b) \geq 1 - \xi_i, \ i = 1, \cdots, m$, and $\mu_i$ are the Lagrange multipliers for $\xi_i \geq 0, \ i = 1, \cdots, m$. The KKT conditions for the primal problem (1) are

$$\begin{array}{ll} \frac{\partial L_p}{\partial w} : \ w - \sum\limits_{i\in\mathscr{A}} \alpha_i y_i x_i = 0, & \\ \frac{\partial L_p}{\partial b} : \ \sum\limits_{i\in\mathscr{A}} y_i \alpha_i = 0, & \\ \frac{\partial L_p}{\partial \xi_i} : \alpha_i + \mu_i = C, & \forall i \in \mathscr{A}, \\ \mu_i \xi_i = 0, & \forall i \in \mathscr{A}, \\ y_i(w^T x_i + b) - 1 + \xi_i \geq 0, & \forall i \in \mathscr{A}, \\ \alpha_i\{y_i(w^T x_i + b) - 1 + \xi_i\} = 0, & \forall i \in \mathscr{A}, \\ \xi_i \geq 0, \ \ \alpha_i \geq 0, \ \ \mu_i \geq 0, & \forall i \in \mathscr{A}. \end{array} \tag{4}$$

Note that the constraints of problem (1) are linear. The intersection of the sets of feasible directions with the sets of descent directions coincides with the intersection of the sets of feasible directions for linearized constraints with the sets of descent directions. In other words, the regularity condition holds [5]. Hence, the KKT conditions must be satisfied at the optimal solutions to (1). Also, (1) is convex, thus the KKT conditions are sufficient. In summary, we have

**Proposition 1:** The KKT conditions for problem (1) are necessary and sufficient for $w, b, \xi$ to be solution of (1). □

This proposition implies that (1) can be solved by finding solutions to those KKT conditions. Thus, we can rewrite (2) as

$$\begin{array}{lll}
\min\limits_{w,b,\xi,z,C,\alpha,\mu} & \sum\limits_{k\in\mathscr{B}} (z_k)_* & \\
\text{s. t.} & w - \sum\limits_{i\in\mathscr{A}} \alpha_i y_i x_i, & = 0, \\
& \sum\limits_{i\in\mathscr{A}} y_i \alpha_i & = 0, \\
& \alpha_i + \mu_i & = C, \ i \in \mathscr{A}, \\
& \alpha_i \{y_i(w^T x_i + b) - 1 + \xi_i\} & = 0, \ i \in \mathscr{A}, \\
& \mu_i \xi_i & = 0, \ i \in \mathscr{A}, \\
& y_i(w^T x_i + b) - 1 + \xi_i & \geq 0, \ i \in \mathscr{A}, \\
& \xi_i \geq 0, \quad \alpha_i \geq 0, \quad \mu_i & \geq 0, \ i \in \mathscr{A}, \\
& z_k + y_k(w^T x_k + b) & \geq 0, \ k \in \mathscr{B}, \\
& z_k & \geq 0, \ k \in \mathscr{B}, \\
& C & \geq C_0.
\end{array} \tag{5}$$

# 3 Concave Approximation

Note that $(z_k)_*$ in (5) is not differentiable, gradient-based nonlinear programming techniques can not be used for solving problem (5). Instead of solving (5) directly, we approximate function $(z_k)_*$ by a differentiable function $t(x,\beta) := 1 - e^{-\beta x}$, $\beta > 0$, $x \geq 0$, see [1][7]. Now, (5) can be approximately solved by

$$\min_{w,b,\xi,z,C,\alpha,\mu} \left\{ \sum_{k\in\mathscr{B}} (1 - e^{-\beta z_k}) \mid \text{ constraints in (5)} \right\}. \tag{6}$$

We have the following proposition.

**Proposition 2:** Solutions to the problem (6) exist.

**proof** Firstly, we show that the feasible regions of (6) are not empty. Since for any $C \geq C_0$, (1) is a quadratic programming, the feasible region of (1) is an closed nonempty set (e.g. take $b = 0$, $w_i = 0$, $\xi_i = 1$, $i = 1, \cdots, m$ ), the minimal solutions to problem (1) can always be obtained. Thus, For any $C \geq C_0$, $\mathscr{S}(C)$ is not empty, and the solution set to the KKT conditions is not empty. Obviously, as long as $z_k$ is large enough, problems (6) have feasible solutions. Secondly, we notice that the feasible regions of (6) are closed. Then, we can derive the conclusion from the fact that the objective functions of problems (6) are concave and bounded below. □

Let $p = (z, w, b, \xi, C)$, $\gamma = (\alpha, \mu, 0)$, and by using some standard transformation, the constraint region of (6) can be written as a more general form

$$\Omega = \{s := (p,\gamma) \mid Mp \leq d, N\gamma \leq d', \gamma \geq 0, Dp + E\gamma \leq 0, \gamma^T(Mp - d) = 0\}, \tag{7}$$

where $M, N, D$ and $E$ are coefficients transformed from the constraints of (5). Note that $z_k \geq 0, k \in \mathscr{B}$, thus the nonsmooth problem can be stated as follows

$$\min_{s\in\Omega} h^T |s|_*, \tag{8}$$

and the smooth problem (6) can be transformed into the following concave minimization problem

$$\min_{s\in\Omega} h^T(\mathbf{1}-e^{-\beta|s|}), \tag{9}$$

where $h=(1,\cdots,1,0,\cdots,0)^T\in R^k$, which the first $|\mathscr{B}|$ components are 1, where $|\mathscr{B}|$ is the dimension of variable $z$, $\mathbf{1}=(1,\cdots,1)^T$, $|s|_*=(|s_1|_*,\cdots,|s_k|_*)^T$, $e^{-\beta|s|}=(e^{-\beta|s_1|},\cdots,e^{-\beta|s_k|})^T$.

We have the following theorem.

**Theorem 1**: Let $\Omega$ be defined by (7) that contains no straight lines going to infinity in both directions, and let $h\geq 0$. Then for a sufficiently large positive but finite value $\beta_0$ of $\beta$, the smooth problem (9) has a global solution that also solves the original nonsmooth problem (8).

**Proof:** Note first that the following obvious relations

$$h^T|s|_*\geq h^T(\mathbf{1}-e^{-\beta|s|}),\ \forall s\in R^k. \tag{10}$$

and

$$\lim_{\beta_i\to\infty} h^T(\mathbf{1}-e^{-\beta_i|s|})=h^T|s|_*,\ \forall s\in R^k. \tag{11}$$

hold. Define

$$\begin{aligned} F=&\ \{s:=(p,\gamma)\mid \gamma^T(Mp-d)=0\},\\ \Omega_1=&\{s:=(p,\gamma)\mid Mp\leq d,\gamma\geq 0,Dp+E\gamma\leq 0,N\gamma\leq d'\}. \end{aligned}$$

One has

$$\begin{aligned} F &= \{s\mid \gamma_1(M_1^Tp-d_1)=0,\cdots,\gamma_l(M_l^Tp-d_l)=0\}\\ &= \textstyle\bigcap_{i=1}^l\{s\mid \gamma_i(M_i^Tp-d_i)=0\}\\ &= \textstyle\bigcap_{i=1}^l\{\{s\mid \gamma_i=0,M_i^Tp-d_i\leq 0\} \text{ or } \{s\mid \gamma_i\geq 0,M_i^Tp-d_i=0\}\}, \end{aligned}$$

where $\gamma=(\gamma_1,\cdots,\gamma_l)\in R^l$. Let

$$\begin{aligned} F_1 &= \{s\mid \gamma_1=0,M_1^Tp-d_1\leq 0,\ \cdots,\ \gamma_l=0,M_l^Tp-d_l\leq 0\},\\ F_2 &= \{s\mid \gamma_1\geq 0,M_1^Tp-d_1=0,\ \cdots,\ \gamma_l=0,M_l^Tp-d_l\leq 0\},\\ &\cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\\ F_{2^l} &= \{s\mid \gamma_1\geq 0,M_1^Tp-d_1=0,\ \cdots,\ \gamma_l\geq 0,M_l^Tp-d_l=0\}. \end{aligned}$$

Then, we have

$$\Omega=\Omega_1\cap F=\Omega_1\cap(\cup_{i=1}^{2^l}F_i)=\cup_{i=1}^{2^l}(\Omega_1\cap F_i),$$

where $\Omega_1\cap F_i$, $i=1,\cdots,2^l$, are polyhedral sets. Thus, $\Omega$ is a union of finite polyhedral sets. The smoothing problem (9) is equivalent to the following concave minimization problem

$$\min_{(s,u)\in T} h^T(\mathbf{1}-e^{-\beta u}), \tag{12}$$

where $T := \{(s,u) \mid s \in \Omega, -u \le s \le u\} = \bigcup_{i=1}^{2^l} \{(s,u) \mid s \in \Omega_1 \cap F_i, -u \le s \le u\}$.

Since the objective function of this problem is concave in $(s,u)$ on $R^{2k}$ and is bounded below on $T$, if follows by [8] that it has a vertex $(s(\beta),u(\beta))$ of one of the polyhedral $\{(s,u) \mid s \in \Omega_1 \cap F_i, -u \le s \le u\}$ as a global solution for each $\beta > 0$. Since $T$ has a finite number of such vertices, one vertex, say $(\bar{s},\bar{u})$, will repeatedly solve (12) for some sequence $\{\beta_0,\beta_1,\cdots\} \uparrow \infty$. Hence for $\beta_i \ge \beta_0$,

$$\begin{aligned} h^T(\mathbf{1}-e^{-\beta_i\bar{u}}) &= h^T(\mathbf{1}-e^{-\beta_i u(\beta_i)}) \\ &= \min_{(s,u)\in T} h^T(\mathbf{1}-e^{-\beta_i u}) \\ &= \min_{s\in\Omega} h^T(\mathbf{1}-e^{-\beta_i|s|}) \\ &\le \inf_{s\in\Omega} h^T|s|_*, \end{aligned}$$

where the last inequality follows from (10). Letting $i \to \infty$ if follows by (11) that

$$h^T|\bar{s}|_* \le h^T|\bar{u}| = \lim_{i\to\infty} h^T(\mathbf{1}-e^{-\beta_i\bar{u}}) \le \inf_{s\in\Omega} h^T|s|_*$$

Since $\bar{s} \in \Omega$, it follows that $\bar{s}$ solves (8), i.e., $\bar{s}$ is the solution of (5). □

# 4 Numerical experiments

We demonstrate now the effectiveness of this approach by comparing it numerically with the model that $C$ takes the default value 1.0. All experiments are run on the Intel(R) AT/AT Compatible with CPU 3.0G and 2GM RAM. We ran all tests on six publicly available datasets: the Wisconsin Prognostic Breast Cancer Database and six datasets, Ionosphere, Cleveland Heart Problem, Wine and Tic-tac-toe from the Irvine Machine Learning Database Repository. For wine data set, class 1 and class 2 are selected for numerical experiment. We randomly extract 10% points from the training set as set $B$, and the rest as the set $A$. 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. Experimental results are summarized in Table 1.

# References

[1] Bradley P. S., Mangasarian O. L., Rosen J. B.: Parsimonious least norm approximation. Computational Optimization and Applications, 11(1)(1998) 5-21

[2] Burges, C.: A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2(2)(1998) 1-47

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

[4] Deng N. Y., Tian Y. J.: A new approach for data mining: Support Vector Machine. Scientific Press. Beijing, P. R. China, 2004 (in chinese).

Table 1: Numerical results

| Datasets size mxn | C | Tenfold Training Correctness, % | Tenfold Testing Correctness, % |
|---|---|---|---|
| WPBC(24 months) | 1.0 | 82.87 | 80.09 |
| 155x32 | 0.7 | 82.87 | 80.75 |
| WPBC(60 months) | 1.0 | 75.83 | 63.64 |
| 110x32 | 4.333454 | 76.06 | 66.36 |
| Ionosphere | 1.0 | 92.66 | 86.02 |
| 351x34 | 0.128388 | 89.84 | 87.73 |
| Cleveland | 1.0 | 85.48 | 83.18 |
| 297x13 | 0.133076 | 85.78 | 84.53 |
| Wine | 1.0 | 98.04 | 96.92 |
| 130x13 | 0.107698 | 98.29 | 97.69 |
| Tic-tac-toe | 1.0 | 66.24 | 65.45 |
| 958x9 | 2.161537 | 72.29 | 70.69 |

[5] Fletcher, R.: Practical Methods of Optimization, John Wiley and Sons, Inc., 2nd edition. (1987)

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

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

[8] Rockafellar, R. T. Convex Analysis. Princeton University Press, Princeton, New Jersey, 1970

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

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