Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 8, No. 1, pp. 3–14, January 2012

SDP RELAXATION FOR SEMI-SUPERVISED SUPPORT VECTOR MACHINE*

Y.Q. Bai, Y. Chen and B.L. Niu

This paper is to remember Professor Paul Tseng with great sorrow.

Abstract: Semi-Supervised Support Vector Machine (S^3VM) is based on applying the margin maximization principle to both labeled and unlabeled sets. The formulation of S^3VM leads to a mixed integer nonlinear optimization problem. In this paper we first consider a semidefinite programming (SDP) relaxation to the mixed integer nonlinear optimization problem associated with S^3VM. To reduce the size of the SDP relaxation formulation, we further modify the SDP problem by decomposing the semidefinite positive matrix into a sequence of small-size matrices. Finally, we apply the modified SDP relaxation to two artificial and five real-world classification problems under a common experimental setting. The numerical examples show that the modified SDP relaxation is effective. In particular, the relative error of the modified SDP relaxation is within 3% for protein classification test problems.

Key words: *semi-supervised support vector machines, semidefinite programming, mixed integer nonlinear programming*

Mathematics Subject Classification: *90C22, 90C11, 90C90*

1 Introduction

Semi-Supervised Support Vector Machine (S^3VM) has received considerable attention recently for its wide applications in machine learning, text classification, protein classification as well as other aspects. S^3VM is based on applying the margin maximization principle to both labeled set (the training set) and unlabeled set (the working set). The idea of S^3VM is as follows. Given a training set of labeled data and a working set of unlabeled data, the objective of S^3VM is to assign class labels to the working set such that the "best" support vector machine (SVM) is constructed. The approach was first proposed by Vapnik and Sterin [5, 7] and was referred to as Transductive SVM. Since the Transductive SVM derives an inductive rule for the entire input space which is extended by the training set, this approach was also referred to as S^3VM by Chapelle et al. in [5]. Vapnik in the book [19] suggested that S^3VM should deliver better results than the traditional SVM.

The formulation of S^3VM is associated with a mixed integer nonlinear programming (MINLP) problem. There is a wide spectrum of techniques for solving the MINLP problem associated with S^3VM. The review paper presented by Chapelle et al. [5] summarized several approaches, including the local combinatorial search discussed by Joachims [12], the continuation techniques presented by Chapelle et al. [4], the branch-and-bound algorithms

*This research is supported by the grant from National Natural Science Foundation of China (No. 11071158) and the Key Disciplines of Shanghai Municipality (No. S30104).

ISSN 1348-9151

by Bennett and Demiriz [2] and the semidefinite programming by Bie and Cristianini [8]. Among all existed approaches the semidefinite programming (SDP) is an important one because it constructs a tight convex relaxation which can be solved through semidefinite programming, see, e.g., [3, 10, 11, 13, 14, 20]. Bie and Cristianini in [7] provided an SDP relaxation for the MINLP problem associated with S^{3}VM for the problem of binary classification. Their relaxation model has a large scale, say, $O((l+(n-l)^2)^2(l+(n-l)^{2.5})$, where l stands for the number of the labeled examples and n for the number of both the labeled and unlabeled examples. Obviously, their approach is very expensive and may cause computational difficulties for practical applications.

In this paper we first consider a semidefinite programming (SDP) relaxation to the MINLP problem associated with S^{3}VM. To reduce the size of the SDP relaxation formulation, we further modify the SDP problem by decomposing the semidefinite positive matrix into a sequence of small-size matrices. Finally, we apply the modified SDP relaxation to two artificial and five real-world classification problems under a common experimental setting. The numerical examples show that the modified SDP relaxation is effective. In particular, the relative error of the modified SDP relaxation is within 3% for protein classification test problems.

The paper is organized as follows. In Section 2 we briefly recall the standard SVM and S^{3}VM for binary classification problems. In Section 3 we first reformulate the MINLP problem associated with S^{3}VM into an equivalent continuous quadratic programming (QP) problem, and then we present an SDP relaxation for the QP problem. Section 4 discusses the modification of the SDP relaxation by decomposing the large-size semidefinite positive matrix into a sequence of small-size matrices. The numerical performance of the modified SDP relaxation to two artificial and five real-world classification problems are shown in Section 5. Finally, some concluding remarks are given in Section 6.

Throughout the paper, R^n denotes the n-dimensional Euclidean space and $R^{n\times m}$ the $n\times m$-dimensional matrix space. S^n stands for the set of $n\times n$-dimensional symmetric matrices, S^n_+ for the cone of n-dimensional positive semidefinite matrices. $\mathbf{0}_{n\times m}$ denotes the $n\times m$-dimensional all zeros matrix, $\mathbf{e}$ is the n-dimensional all ones vector and I_n the $n\times n$-dimensional identity matrix. For $A, B\in S^n$, $A_{i,j}$ denotes the (i,j) entry of A and $A_{\{i_1,\ldots,i_k\}}$ the principal sub-matrix indexed by $\{i_1,\ldots,i_k\}$. $A\succeq\mathbf{0}$ means that A is positive semidefinite and $A\bullet B$ the inner product of matrices A and B.

2 SVM and S^{3}VM

In this section, we briefly review the standard SVM and S^{3}VM for binary classification problems. For multi-class classification problems, the most popular approach is to reduce the single multi-class problem into a series of binary classification problems. We first restrict our attention to linear classifiers. Here our introduction of SVM and S^{3}VM are based on [5, 6, 7].

Given a training set consisting of l labeled data $\{(\mathbf{x}_i,y_i)\}_{i=1}^l$ and a working set consisting of $(n-l)$ unlabeled data $\{\mathbf{x}_i\}_{i=l+1}^n$, where $y_i\in\{-1,+1\}$ for $i=1,\ldots,l$ and $\mathbf{x}_i\in R^d$ for $i=1,\ldots,n$.

If the data are linearly separable, then there exists a vector $\mathbf{w}\in R^d$ and a scalar $b\in R$ such that a separating hyperplane $\mathbf{w}^T\mathbf{x}+b=0$ is constructed by using the training set. Then the separating hyperplane is used to predicate the labels of data in the working set by letting $y_i=1$ if $\mathbf{w}^T\mathbf{x}_i+b>0$ and $y_i=-1$ if $\mathbf{w}^T\mathbf{x}_i+b<0$ for $i=l+1,\ldots,n$.

Figure 1 shows two examples of separating hyperplane. The figure in the left part illustrates the case of many possible separating hyperplane for two data sets.

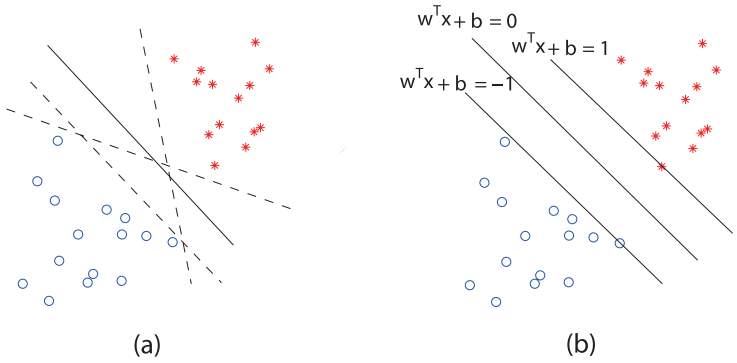

Figure 1: Separating hyperplane for maximizing the margin.

The statistical learning theory suggests that the hyperplane which maximizes the margin (maximizes the distance between it and the nearest data point of each class) shall give rise to the best estimation. Namely it should work the best on new data [19].

If there is a separating hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$, then there are a nonzero vector $\mathbf{w}$ and a scalar b such that

$$\mathbf{w}^T\mathbf{x}_i + b \geq 1 \text{ if } y_i = 1 \text{ and } \mathbf{w}^T\mathbf{x}_i + b \leq -1 \text{ if } y_i = -1 \text{ for } i = 1, \ldots, l \tag{2.1}$$

or equivalently

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \text{ for } i = 1, \ldots, l. \tag{2.2}$$

Therefore, it is equivalent to maximize the distance between the two parallel planes $\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$ (see the right part of Figure 1). The Euclidean distance between these two planes is $\dfrac{2}{\|\mathbf{w}\|_2}$. Hence, the hard-margin SVM can be formulated as the following convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \ i = 1, \ldots, l. \end{aligned} \tag{2.3}$$

The dual problem of (2.3) is the standard hard-margin SVM:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \ & \mathbf{e}^T\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^T \mathrm{diag}(\mathbf{y}) K \mathrm{diag}(\mathbf{y})\boldsymbol{\alpha} \\ \text{s.t.} \ & \mathbf{y}^T\boldsymbol{\alpha} = 0, \ \boldsymbol{\alpha} \geq \mathbf{0}, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^T$ is the Lagrange multiplier or the dual variables of (2.3) and $K_{i,j} = \mathbf{x}_i^T\mathbf{x}_j$.

If the data are not linear separable, then the nonnegative variables, $\xi_i \geq 0, i = 1, \ldots, l$, are introduced and the constraints (2.3) are modified as

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \text{ for } i = 1, \ldots, l, \tag{2.5}$$

where $\xi_i, i = 1, ..., l$ are introduced to measure the misclassification errors. Naturally, ξ_i are added to the objective function in (2.5) to control the misclassification errors, and the problem (2.3) is modified into the soft-margin SVM:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{l}\xi_i \\ \text{s.t.} & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ i = 1, \ldots, l, \\ & \xi_i \geq 0,\ i = 1, \ldots, l, \end{array} \tag{2.6}$$

where $C > 0$ is a fixed penalty parameter and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_l)^T$. The dual of this optimization problem is the standard soft-margin SVM:

$$(\mathcal{P}_{\text{SVM}}) \qquad \begin{array}{ll} \max\limits_{\boldsymbol{\alpha}} & \mathbf{e}^T\boldsymbol{\alpha} - \dfrac{1}{2}\boldsymbol{\alpha}^T\text{diag}(\mathbf{y})K\text{diag}(\mathbf{y})\boldsymbol{\alpha} \\ \text{s.t.} & \mathbf{y}^T\boldsymbol{\alpha} = 0,\ \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{e}. \end{array} \tag{2.7}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^T$ is the Lagrange multiplier or the dual variable of (2.6) and $K_{i,j} = \mathbf{x}_i^T\mathbf{x}_j$.

Since many practical classification problems are too complicated to use the linear classifiers, the nonlinear classifiers are needed to classify the practical problems. An approach called kernel technique is introduced to convert nonlinear classifiers into linear classifiers in a higher dimensional space in terms of a map. Let $\phi : R^d \to R^h (h \geq d)$ denote such a map and $K_{i,j} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. It is easy to observe that ϕ always appears in linear classifiers in the inner product form. Thus a kernel function $k(\cdot, \cdot)$ can be constructed such that $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. Many kernel functions have been introduced in the literatures. The following two are the typical kernel functions:

$$\begin{aligned} \text{linear kernel:}\ & k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j, \\ \text{Gaussian kernel:}\ & k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2). \end{aligned}$$

The Mercer Theorem states that a symmetric function $k(\mathbf{x}_i, \mathbf{x}_j)$ on a finite input space is a kernel function if and only if the matrix $K = \big(k(\mathbf{x}_i, \mathbf{x}_j)\big)_{i,j=1}^{n}$ is positive semidefinite [6].

Unlike SVM, the idea of S^{3}VM is to apply the maximum margin principle to both training and working sets. Let $\gamma(y)$ denote the optimal value of $(\mathcal{P}_{\text{SVM}})$ and $\mathbf{y}^u = (y_{l+1}, \ldots, y_n)^T$ the unknown labels for the data in the working set. Following the papers [5, 7], the formulation of S^{3}VM is as follows:

$$\begin{array}{ll} \min\limits_{\mathbf{y}^u} & \gamma(\mathbf{y}) \\ \text{s.t.} & y_i \in \{-1, +1\},\ i = l + 1, \ldots, n, \end{array} \tag{2.8}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$. Without loss of generality, the intercept b is assumed to be 0, which can be easily realized since the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ can be converted into $\mathbf{w}^{*T}\mathbf{x} = 0$ by letting $\mathbf{w}^* = (\mathbf{w}^T, b)^T$. Thus the constraint $\mathbf{y}^T\boldsymbol{\alpha} = 0$ will disappear in the problem $(\mathcal{P}_{\text{SVM}})$. Substituting $\gamma(\mathbf{y})$ with its dual problem, problem (2.8) can be rewritten as the following optimization problem

$$(\mathcal{P}_{\text{S}^3\text{VM}}) \qquad \begin{array}{ll} \min\limits_{\boldsymbol{\mu},\boldsymbol{\delta},\mathbf{y}^u} & \dfrac{1}{2}(\mathbf{e} + \boldsymbol{\mu} - \boldsymbol{\delta})^T\text{diag}(\mathbf{y})K^{-1}\text{diag}(\mathbf{y})(\mathbf{e} + \boldsymbol{\mu} - \boldsymbol{\delta}) + C\mathbf{e}^T\boldsymbol{\delta} \\ \text{s.t.} & \boldsymbol{\mu} \geq \mathbf{0},\ \boldsymbol{\delta} \geq \mathbf{0}, \\ & y_i \in \{-1, +1\},\ i = l + 1, \ldots, n, \end{array} \tag{2.9}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ are the dual variables of $(\mathcal{P}_{\mathrm{S^3VM}})$ and K is the kernel matrix with the pseudo-inverse K^{-1}.

The objective function of $(\mathcal{P}_{\mathrm{S^3VM}})$ is an inhomogeneous polynomial of degree 4 so that it is a nonconvex and nonlinear function. Problem $(\mathcal{P}_{\mathrm{S^3VM}})$ contains both continuous variables $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ and discrete variable $\mathbf{y}^u$.

3 SDP Relaxation to S^{3}VM

In this section we first introduce two novel transformations such that the discrete variable $\mathbf{y}^u$ is implied in terms of two continuous variables. We aim to rewrite the problem $(\mathcal{P}_{\mathrm{S^3VM}})$ into an equivalent continuous quadratic programming (QP) problem. Then we further relax the QP problem into an SDP problem.

We intend to modify the problem $(\mathcal{P}_{\mathrm{S^3VM}})$ by using the following transformations:

$$\boldsymbol{\omega} = \mathrm{diag}\,(\mathbf{y})(\mathbf{e} + \boldsymbol{\mu}),\ \boldsymbol{\nu} = \mathrm{diag}\,(\mathbf{y})\boldsymbol{\delta}. \tag{3.1}$$

These two transformations aim to imply the discrete variable $\mathbf{y}^u$ in the problem $(\mathcal{P}_{\mathrm{S^3VM}})$ in terms of two continuous variables $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$.

Given $\boldsymbol{\mu} \geq \mathbf{0}$ and $\boldsymbol{\delta} \geq \mathbf{0}$, the equations in (3.1) are essentially equivalent to

$$\omega_i y_i \geq 0,\ i = 1, \ldots, l \ \text{ and } \ \omega_i^2 \geq 1,\ \nu_i \omega_i \geq 0,\ \nu_i^2 = \delta_i^2,\ i = 1, \ldots, n. \tag{3.2}$$

Thus $(\mathcal{P}_{\mathrm{S^3VM}})$ can be rewritten into the following equivalent continuous optimization problem:

$$\begin{array}{ll} \min_{\boldsymbol{\omega},\boldsymbol{\nu},\boldsymbol{\delta},t} & t \\ \text{s.t.} & \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\nu})^T K^{-1}(\boldsymbol{\omega} - \boldsymbol{\nu}) + C\mathbf{e}^T\boldsymbol{\delta} \leq t, \\ & \omega_i^2 \geq 1,\ \nu_i\omega_i \geq 0,\ \nu_i^2 = \delta_i^2,\ i = 1, \ldots, n, \\ & \omega_i y_i \geq 0,\ i = 1, \ldots, l, \\ & \boldsymbol{\delta} \geq \mathbf{0}. \end{array} \tag{3.3}$$

Let $\mathbf{z} = (\boldsymbol{\omega}^T, \boldsymbol{\nu}^T, \boldsymbol{\delta}^T, t)^T$ be a $(3n+1)$-dimensional vector. Problem (3.3) can be rewritten as the following QP problem.

$$\begin{array}{lll} (\mathcal{P}_{\mathrm{QP}}) & \min_{\mathbf{z}} & \mathbf{c}^T\mathbf{z} \\ & \text{s.t.} & \mathbf{z}^T G\mathbf{z} + \mathbf{g}^T\mathbf{z} \leq 0, \\ & & \mathbf{z}^T A_i\mathbf{z} + 1 \leq 0,\ \mathbf{z}^T B_i\mathbf{z} \leq 0,\ \mathbf{z}^T D_i\mathbf{z} = 0,\ i = 1, \ldots, n, \\ & & H\mathbf{z} \leq \mathbf{0}, \end{array} \tag{3.4}$$

where $\mathbf{c} = (\mathbf{0}_{1\times 3n}, 1)^T \in R^{3n+1}$ and $\mathbf{g} = (\mathbf{0}_{1\times 2n}, C\mathbf{e}^T, -1)^T \in R^{3n+1}$, and G and H are defined as follows, respectively.

$$\begin{aligned} G &= \begin{pmatrix} \frac{1}{2}P^T K^{-1} P & \mathbf{0}_{(3n+1)\times 1} \\ \mathbf{0}_{1\times(3n+1)} & 0 \end{pmatrix} \in S^{3n+1}, \\ H &= \begin{pmatrix} -\mathrm{diag}\,(\bar{\mathbf{y}}) & \mathbf{0}_{l\times(2n-l)} & \mathbf{0}_{n\times n} & \mathbf{0}_{l\times 1} \\ \mathbf{0}_{n\times l} & \mathbf{0}_{n\times(2n-l)} & -I_n & \mathbf{0}_{n\times 1} \end{pmatrix} \in R^{(l+n)\times(3n+1)}, \end{aligned}$$

where $P = (I_n, -I_n, \mathbf{0}_{n\times n})$, $\bar{\mathbf{y}} = (y_1, \ldots, y_l)^T$. For $i = 1, \ldots, n$, $A_i \in S^{3n+1}$ is an all zeros matrix except for the ith diagonal which is equal to -1, $B_i \in S^{3n+1}$ is an all zeros matrix except for the two elements $(i, n+i)$ and $(n+i, i)$ which are equal to -1, and $D_i \in S^{3n+1}$ is an all zeros matrix except for the two elements $(n+i, n+i)$ and $(2n+i, 2n+i)$ which are equal to 1 and -1, respectively.

Therefore, problem $(\mathcal{P}_{\mathrm{S^3VM}})$ is converted into a QP problem $(\mathcal{P}_{\mathrm{QP}})$. Obviously, the objective function is a convex function since it is linear. However, except for the first quadratic constraint, the other constraints are not convex. The idea of the relaxation is to introduce into some convex form.

Let $X = \begin{pmatrix} 1 & \mathbf{z}^T \\ \mathbf{z} & Z \end{pmatrix}$. It is easy to verify that $Z = \mathbf{z}\mathbf{z}^T$ is equivalent to $X \succeq 0$ and $\mathrm{rank}\,(X) = 1$. We then present an SDP relaxation for $(\mathcal{P}_{\mathrm{QP}})$ below.

The problem $(\mathcal{P}_{\mathrm{QP}})$ is equivalent to

$$\begin{aligned} \min_{\mathbf{z},Z} \ & \mathbf{c}^T\mathbf{z} \\ \text{s.t.} \ & G \bullet Z + \mathbf{g}^T\mathbf{z} \le 0, \\ & A_i \bullet Z + 1 \le 0,\ B_i \bullet Z \le 0,\ D_i \bullet Z = 0,\ i = 1, \ldots, n, \\ & H\mathbf{z} \le 0, \\ & X = \begin{pmatrix} 1 & \mathbf{z}^T \\ \mathbf{z} & Z \end{pmatrix} \in S_+^{3n+2} \text{ and } \mathrm{rank}(X) = 1. \end{aligned} \tag{3.5}$$

Dropping the constraint rank(X)=1, we then obtain the following SDP relaxation:

$$(\mathcal{P}_{\mathrm{SDP}}) \qquad \begin{aligned} \min_{\mathbf{z},Z} \ & \mathbf{c}^T\mathbf{z} \\ \text{s.t.} \ & G \bullet Z + \mathbf{g}^T\mathbf{z} \le 0, \\ & A_i \bullet Z + 1 \le 0,\ B_i \bullet Z \le 0,\ D_i \bullet Z = 0,\ i = 1, \ldots, n, \\ & H\mathbf{z} \le 0, \\ & X \succeq 0. \end{aligned} \tag{3.6}$$

If the problem $(\mathcal{P}_{\mathrm{SDP}})$ is solved and the optimal solution $\bar{\mathbf{z}}$ is obtained, we can use the signs of $\bar{z}_i$ for $i = l+1, \ldots, n$ to label y_i for the unlabeled data $\mathbf{x}_i$ for $i = l+1, \ldots, n$ according to the definitions of $\boldsymbol{\omega}$ and $\bar{\mathbf{z}}$.

Note that problem $(\mathcal{P}_{\mathrm{SDP}})$ is a large-size problem with $(9n^2+15n+4)/2$ variables. It also has mixed constraints, including a convex quadratic constraint, $(4n + l)$ linear constraints and a positive semidefinite positive matrix constraint. The semidefinite positive matrix X is a large size matrix. This leads to the issue of how to decompose the semidefinite positive matrix X.

4 Modification of SDP Relaxation

We have observed that problem $(\mathcal{P}_{\mathrm{SDP}})$ is a large-size problem since the vector $\mathbf{z}$ is a $(3n+1)$-dimensional vector and the matrix X is a $(3n+1) \times (3n+1)$-dimensional matrix. Since $\mathbf{z} = (\boldsymbol{\omega}^T, \boldsymbol{\nu}^T, \boldsymbol{\delta}^T, t)^T$ and the relations among $\boldsymbol{\omega}$, $\boldsymbol{\nu}$ and $\boldsymbol{\delta}$ depend on the system of inequalities and equation of $\omega_i^2 \ge 1,\ \nu_i\omega_i \ge 0,\ \nu_i^2 = \delta_i^2,\ i = 1, \ldots, n$, respectively. In fact these are the constraints of problem (3.3). Note that there is no relation among ω_i, ν_i and δ_i if their indexes are not equal.

Motivated by this observation, we shall attempt to reduce the size of problem $(\mathcal{P}_{\mathrm{SDP}})$ by decomposing X.

Instead of requiring $X \succeq 0$ in $(\mathcal{P}_{\mathrm{SDP}})$, we decompose it into a set of small-size positive semidefinite principal sub-matrices.

Let

$$X_i = \begin{pmatrix} 1 & \mathbf{z}_i^T \\ \mathbf{z}_i & Z_i \end{pmatrix} \succeq 0,\ i = 1, \ldots, n, \tag{4.1}$$

where $\mathbf{z}_i = (z_i, z_{n+i}, z_{2n+i})^T$ and $Z_i = Z_{\{i,n+i,2n+i\}}$. Constraints $A_i \bullet Z + 1 \leq 0,\ B_i \bullet Z \leq 0,\ D_i \bullet Z = 0,\ i = 1, \ldots, n$ are then equivalent to

$$\widehat{A}_i \bullet Z_i + 1 \leq 0,\ \widehat{B}_i \bullet Z_i \leq 0,\ \widehat{D}_i \bullet Z_i = 0,\ i = 1, \ldots, n \tag{4.2}$$

where $\widehat{A}_i = A_{i\{i,n+i,2n+i\}}$, $\widehat{B}_i = B_{i\{i,n+i,2n+i\}}$ and $\widehat{D}_i = D_{i\{i,n+i,2n+i\}}$. This leads to the following relaxed problem:

$$\begin{aligned} (\mathcal{P}_{\mathrm{FRSDP}}) \qquad & \min_{\mathbf{z}, Z_i} \ \mathbf{c}^T \mathbf{z} \\ & \text{s.t.} \ \ \mathbf{z}^T G \mathbf{z} + \mathbf{g}^T \mathbf{z} \leq 0, \\ & \qquad \widehat{A}_i \bullet Z_i + 1 \leq 0,\ \widehat{B}_i \bullet Z_i \leq 0,\ \widehat{D}_i \bullet Z_i = 0,\ i = 1, \ldots, n, \\ & \qquad H\mathbf{z} \leq 0, \\ & \qquad X_i \succeq 0,\ i = 1, \ldots, n. \end{aligned} \tag{4.3}$$

Note that the number of variables of problem $(\mathcal{P}_{\mathrm{FRSDP}})$ is reduced to $(9n+1)$. Obviously, it is much less than that of problem $(\mathcal{P}_{\mathrm{SDP}})$, which has $(9n^2+15n+4)/2$ variables. Moreover, we replace the constraint $\mathbf{z}^T G \mathbf{z} + \mathbf{g}^T \mathbf{z} \leq 0$ by a second-order cone constraint due to a computational consideration.

The following proposition states that the optimal solution provided by problem $(\mathcal{P}_{\mathrm{FRSDP}})$ is a lower bound of problem $(\mathcal{P}_{\mathrm{SDP}})$.

Proposition 4.1. *Let $\mathcal{F}^{SDP}$ and $\mathcal{F}^{FRSDP}$ denote the feasible regions of $(\mathcal{P}_{SDP})$ and $(\mathcal{P}_{FRSDP})$, respectively. One has*

$$\mathcal{F}^{SDP} \subseteq \mathcal{F}^{FRSDP}.$$

It is straightforward to prove Proposition 4.1 and we omit the proof here. The following theorem establishes the relation between $(\mathcal{P}_{\mathrm{FRSDP}})$ and $(\mathcal{P}_{\mathrm{QP}})$.

Theorem 4.2. *Let $\mathbf{z}^*$ and $X_i^*,\ i = 1, \ldots, n$ be a solution of $(\mathcal{P}_{FRSDP})$. If rank$(X_i^*) = 1$ for $i = 1, \ldots, n$, then $\mathbf{z}^*$ is a solution of $(\mathcal{P}_{QP})$.*

Proof. Suppose that $\mathrm{rank}(X_i^*) = 1$ for $i = 1, \ldots, n$. To prove $\mathbf{z}^*$ is a solution of $(\mathcal{P}_{\mathrm{QP}})$, we need to verify that $\mathbf{z}^*$ satisfies the constraints of $(\mathcal{P}_{\mathrm{QP}})$. In other words, $\mathbf{z}^*$ must satisfy $\mathbf{z}^T A_i \mathbf{z} + 1 \leq 0,\ \mathbf{z}^T B_i \mathbf{z} \leq 0,\ \mathbf{z}^T D_i \mathbf{z} = 0,\ i = 1, \ldots, n$, respectively. By using the definition of X_i, $\mathrm{rank}(X_i^*) = 1$ implies that $Z_i^* - \mathbf{z}_i^* \mathbf{z}_i^{*T} = 0$. Thus we have

$$Z_{i,i}^* = z_i^{*2},\ Z_{n+i,i}^* = z_{n+i}^* z_i^*,\ Z_{n+i,n+i}^* = z_{n+i}^{*2},\ Z_{2n+i,2n+i}^* = z_{2n+i}^{*2}. \tag{4.4}$$

Combining (4.4) and the definitions of A_i, B_i, D_i, $\widehat{A}_i$, $\widehat{B}_i$ and $\widehat{D}_i$, it follows that

$$\begin{aligned} \mathbf{z}^{*T} A_i \mathbf{z}^* + 1 &= z_i^{*2} + 1 = Z_{i,i}^* + 1 = \widehat{A}_i \bullet Z_i^* + 1 \leq 0, \\ \mathbf{z}^{*T} B_i \mathbf{z}^* &= 2 z_{n+i}^* z_i^* = 2 Z_{n+i,i}^* = \widehat{B}_i \bullet Z_i^* \leq 0, \end{aligned}$$

and

$$\mathbf{z}^{*T} D_i \mathbf{z}^* = z_{n+i}^{*2} - z_{2n+i}^{*2} = Z_{n+i,n+i}^* - Z_{2n+i,2n+i}^* = \widehat{D}_i \bullet Z_i^* = 0.$$

This proves the theorem. □

Table 1: Basic properties of the six data sets.

Data set	Classes	Dimension	Points	Comment
g241c	2	241	1500	artificial
g241d	2	241	1500	artificial
USPS	2	241	1500	imbalanced
COIL2	6	241	1500	
BCI	2	117	400	
Text	2	11,960	1500	sparse discrete

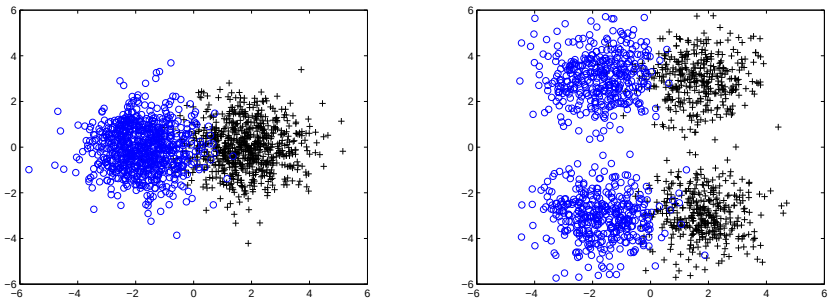

Figure 2: Two-dimensional examples of g241c (left) and g241d (right). Black plus signs, class +1; blue circles, class −1.

5 Numerical Examples

In this section, we apply the relaxation ($\mathcal{P}_{\text{FRSDP}}$) to two artificial data sets, g241c and g241d, and five real-world data sets, USPS, COIL2, BCI, Text and SCOP40mini, respectively. The real-world data sets are derived from the fields of the hand written digit recognition, the image recognition, the brain-computer interface, the text classification and the protein classification. The computational results are obtained by solving the relaxation ($\mathcal{P}_{\text{FRSDP}}$) with SeDuMi 1.3 of Matlab 7.6 (R2008a) on a workstation under Red Hat Linux 5.1.

Six data sets, g241c, g241d, USPS, COIL2, BCI and Text, are obtained from http://www.kyb.tuebingen.mpg.de/ssl-book/ [4]. Their basic properties are shown in Table 1.

The data sets g241c and g241d are generated from two and four normal distributions respectively. Two-dimensional examples of g241c and g241d are shown in Figure 2. The data set USPS is derived from the famous USPS set of handwritten digits. The digits "2" and "5" are assigned to the class +1, and all the others form class −1. The data set COIL2 includes 1500 images of 24 different objects taken from different angles. Each class consists of 12 objects. The data set BCI is derived from the research toward the development of a brain-computer interface (BCI). Each data of BCI is resulted from the subject's imagination of moments, the left hand (class −1) or the right hand (class +1). The data set Text is the 5 comp.* groups from the Newsgroups data set and the goal is to classify the ibm category versus the rest.

Table 2: Average test errors (in %, in front of /) and average CPU-times (in second, behind /) on g241c, g241d, USPS, COIL2, BCI and Text.

		g241c	g241d	USPS	COIL2	BCI	Text
FRSDP	L	47.48/110	46.28/110	45.74/145	48.47/101	31.44/27	41.90/1894
	G	30.95/1370	30.33/1436	13.51/1383	19.40/1262	31.33/37	32.85/1298
SVM	G	35.81	35.74	15.76	20.1	43.28	47.46

Table 3: Average test errors (in %, in front of /) and average CPU-times (in second, behind /) on SCOP40mini.

		BLAST	SW	NW	LA	PRIDE
FRSDP	G	3.01/2534	3.00/2462	3.00/2329	3.00/2543	3.53/2356
SVM	G	2.84	2.49	2.46	2.25	2.89

The data set SCOP40mini is derived from the data set SCOP40_Minidatabase (Accession Number: PCB00019) downloaded from the Protein Classification Benchmark Collection [18]. The protein classification is on hierarchial levels: the first two levels, family and superfamily, describe near and far evolutionary relationships; the third, fold, describes geometrical relationships. The goal of SCOP40_Minidatabase is to classify protein domain sequences and structures into superfamilies, based on families. The members in the same superfamily form class $+1$, and the rest members form class -1. In the data set SCOP40_Minidatabase, there are 55 families. The data set SCOP40mini is generated by discarding the families with less than 10 members. Thus, there are 32 families in SCOP40mini.

For g241c, g241d, USPS, COIL2, BCI and Text, each data set has been equipped with 12 subsets of 100 labeled data [4]. Thus there are 12 classification tasks for each data set. Here we use both linear kernel and Gaussian kernel. The parameter σ is chosen as the median of the pairwise distances and the parameter C is fixed to 3500 for linear kernel and 1.5 for Gaussian kernel. For SCOP40mini, the members outside the family but within the same superfamily and partial members outside the superfamily are selected as labeled data [18]. Since protein sequences are alphabetic sequences with variant lengths, they should be represented as fixed-length vectors of real numbers so that they can be used as inputs for classification algorithms. Many approaches have been developed to represent protein sequences, such as BLAST [1], Smith-Waterman (SW) [17], Needleman-Wunsch (NW) [15], Local Alignment Kernel (LA) [16], PRIDE [9] and so on. Here we use the approaches, BLAST, SW, NW, LA and PRIDE, to show their effect on our method. We only use the Gaussian kernel for this data set. The parameter σ is also chosen as the median of the pairwise distances and C is fixed to 3500.

For the purpose of comparison, we present the results of the SVM equipped with Gaussian kernel. The parameter σ is chosen as the median of the pairwise distances. The results are obtained by using Spider 1.71.

Table 2 reports the results of ($\mathcal{P}_{\mathrm{FRSDP}}$) and SVM on g241c, g241d, USPS, COIL2, BCI and Text. The results are the average test errors (in %) and average CPU-times (in second) of the 12 tasks. L denotes the linear kernel and G denotes the Gaussian kernel. We see from Table 2 that ($\mathcal{P}_{\mathrm{FRSDP}}$) equipped with Gaussian kernel performs better than that equipped with linear kernel on these data sets, though the former needs more time than the latter

except for the case of Text. There is an interesting phenomenon that in the case of Text, ($\mathcal{P}_{\rm FRSDP}$) equipped with Gaussian kernel not only performs better but also needs less time than that equipped with linear kernel. We also see from Table 2 that the performance of ($\mathcal{P}_{\rm FRSDP}$) equipped with Gaussian kernel is evidently better than that of SVM. The average test errors of ($\mathcal{P}_{\rm FRSDP}$) are approximate 80% of that of SVM except in the case of COIL2. Especially in the case of Text, the average error of ($\mathcal{P}_{\rm FRSDP}$) is only about 69.2% of that of SVM. The results show that our method is effective on these data sets. Furthermore, its performance on Text confirms that S^3VMs are particularly well suited for text classification and several other (typically high-dimensional) learning problems [4].

Table 3 shows the results of ($\mathcal{P}_{\rm FRSDP}$) and SVM on the data set SCOP40mini. The results are the average test errors (in %) and average CPU-times (in second) of the 32 tasks. We see that although ($\mathcal{P}_{\rm FRSDP}$) performs a little worse than SVM, its average test errors are only about 3%. We also see that the approaches, BLAST, SW, NW, LA and PRIDE, have approximately same average test error and average CPU-time. The results on SCOP40mini indicate that our method is well suited for dealing with the protein classification problems.

6 Conclusions and Remarks

We have presented two SDP relaxations for the MINLP problem ($\mathcal{P}_{\rm S^3VM}$) associated with S^3VM. By decomposing the semidefinite positive matrix into a sequence of small-size matrices, our second SDP relaxation ($\mathcal{P}_{\rm FRSDP}$) has the reasonable size compared with the first SDP relaxation. Furthermore, we have applied the modified SDP relaxation ($\mathcal{P}_{\rm FRSDP}$) to two artificial and five real-world classification problems which are derived from the five fields of the hand-written digit recognition, the image recognition, the brain-computer interface, the text classification and the protein classification. The numerical examples indicated that ($\mathcal{P}_{\rm FRSDP}$) is effective. In particular, the relative error of ($\mathcal{P}_{\rm FRSDP}$) is within 3% for protein classification test problems.

There are two directions for future research. The first one is to establish a second-order cone relaxation for S^3VM because it is well-known that the other possibility is to further modify Problem (P_{S^3VM}) to suit other needs. The other direction is to modify problem ($\mathcal{P}_{\rm S^3VM}$) such that it is able to deal with various cases. For example, we can add a balance constraint $-\lambda \leq \sum_{i=l+1}^{n} y_i \leq \lambda$ to the problem, where λ is a given constant [21]. The balance constraint can prevent the unlabeled data from being classified into the same class.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* 215 (1990) 403–410.

[2] K.P. Bennett and A. Demiriz, Semi-supervised support vector machines, *Adv. Neural Inf. Process. Syst.* 11 (1999) 368–374.

[3] P. Biswas, and Y. Ye, Semidefinite programming for ad hoc wireless sensor network localization, in *Proceedings of the 3rd international symposium on Information processing in sensor networks*, ACM, Berkeley, California, USA, 2004, pp. 46–54.

[4] O. Chapelle, B. Schölkopf and A. Zien, Semi-supervised Learning, MIT Press, Cambridge, MA, 2006.

[5] O. Chapelle, V. Sindhwani. and S.S. Keerthi, Optimization techniques for semi-supervised support vector machines, *J. Mach. Learn. Res.* 9 (2008) 203–233.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, 2000.

[7] T. De Bie and N. Cristianini, Convex methods for transduction, *Adv. Neural Inf. Process. Syst.* 16 (2003) 73–80.

[8] T. De Bie and N. Cristianini, Semi-supervised learning using semidefinite programming, in *Semi-supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien (eds.),MIT Press, Cambridge, MA, 2006.

[9] Z. Gáspári, K. Vlahovicek and S. Pongor, Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm, *Bioinformatics* 21 (2005) 3322–3323.

[10] M.X. Goemans, Semidefinite programming in combinatorial optimization, *Math. Program.* 79 (1997) 143–161.

[11] C. Helmberg, Semidefinite Programming for Combinatorial Optimization, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2000.

[12] T. Joachims, Transductive inference for text classification using support vector machines, in *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, 1999, pp. 200–209.

[13] M. Kojima and L. Tunçel, Cones of matrices and successive convex relaxations of nonconvex sets, *SIAM J. Optim.* 10 (2000) 750–778.

[14] J. Lasserre, An explicit exact SDP relaxation for nonlinear 0-1 programs, in *Lecture Notes in Computer Science*, K. Aardal and M.H. Gerards (eds.), Vol. 2081, Springer, New York, 2001, pp. 293–303.

[15] S.B. Needleman. and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.* 48 (1970) 443–453.

[16] H. Saigo, J.P. Vert, N. Ueda and T. Akutsu, Protein homology detection using string alignment kernels, *Bioinformatics* 20 (2004) 1682–1689.

[17] T.F. Smith and M.S. Waterman, Identification of common molecular subsequences, *J. Mol. Biol.* 147 (1981) 195–197.

[18] P. Sonego,M. Pacurar, S. Dhir, A. Kertesz-Farkas, A. Kocsor, Z. Gaspari, J.A.M. Leunissen and S. Pongor, A Protein Classification Benchmark collection for machine learning, *Nucleic Acids Res.* 35 (2007) D232–D236.

[19] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[20] Z. Wang,S. Zheng,S. Boyd and Y. Ye, Further relaxations of the SDP approach to sensor network localization, *SIAM J. Optim.* 19 (2008) 655–673.

[21] L. Xu, J. Neufeld, B. Larson and D. Schuurmans,Maximum margin clustering, *Adv. Neural Inf. Process. Syst.* 17 (2005) 1537–1544.

Manuscript received 31 January 2010
revised 15 September 2010, 7 February 2011
accepted for publication 13 April 2011

Y.Q. Bai
Department of Mathematics, Shanghai University
Shanghai, 200444, China
E-mail address: yqbai@shu.edu.cn

Y. Chen
Department of Shanghai University
E-mail address: Chenyi-fire@163.com

B.L. Niu
Department of Shanghai University
E-mail address: Amynew0202@163.com