Asia-Pacific Journal of Operational Research
Vol. 33, No. 6 (2016) 1650046 (22 pages)

DOI: 10.1142/S0217595916500469

Soft Quadratic Surface Support Vector Machine for Binary Classification

Jian Luo

School of Management Science and Engineering
Dongbei University of Finance and Economics
Dalian 116025, P. R. China
luojian546@hotmail.com

Shu-Cherng Fang

Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University, Raleigh, NC 27695-7906, USA
fang@ncsu.edu

Zhibin Deng*

School of Economics and Management
University of Chinese Academy of Sciences
Beijing 100190, P. R. China
zhibindeng@ucas.ac.cn

Xiaoling Guo

School of Science, China University of Mining and Technology
Beijing 100083, P. R. China
guoxiaoling@ucas.ac.cn

Received 29 September 2015
Revised 16 May 2016
Accepted 1 August 2016
Published 28 November 2016

In this paper, a kernel-free soft quadratic surface support vector machine model is proposed for binary classification directly using a quadratic function for separation. Properties (including the solvability, uniqueness and support vector representation of the optimal solution) of the proposed model are derived. Results of computational experiments on some artificial and real-world classifying data sets indicate that the proposed soft quadratic surface support vector machine model may outperform Dagher's quadratic model and other soft support vector machine models with a Quadratic or Gaussian kernel in terms of the classification accuracy and robustness.

Keywords: Data mining; support vector machine; binary classification; quadratic optimization; kernel-free SVM.

*Corresponding author.

1. Introduction

Binary classification is an important task of information extraction from data. As a commonly used effective classification technique in machine learning, the support vector machine (SVM) has been adopted for texture classification (Kim *et al.*, 2012), customer churn prediction (Chen *et al.*, 2012) and financial prediction (Cao *et al.*, 2011).

For binary classification, a training data set of n records $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$ is given, where $\mathbf{x}^i = [x_1^i, x_2^i, \ldots, x_m^i]^T \in \mathbb{R}^m$ indicates the position of the ith training point in the m-dimensional space, and the label of $y^i = +1$ or -1 indicates that point $\mathbf{x}^i$ belongs to Class 1 or Class 2, respectively. As an optimization-based binary classification technique, the SVM model was first proposed by Cortes and Vapnik (1995). The basic concept of the SVM model is to find the parameter vector $(\mathbf{u}, d) \in \mathbb{R}^m \times \mathbb{R}^1$ of a hyperplane

$$f(\mathbf{x}) := \mathbf{u}^T\mathbf{x} + d = 0 \tag{1}$$

that separates the n training points $\{\mathbf{x}^1, \ldots, \mathbf{x}^i, \ldots, \mathbf{x}^n\}$ into the two classes as distinctly as possible.

In real-world binary classification applications, the given training data set is often contaminated by outliers and noise. It is quite possible that no hyperplane can really separate all points apart such that each point belongs to the right class. To address this issue, Vapnik (1998) developed a soft SVM model using a continuous measure of misclassification error. However, the soft SVM model still does not work well when the training data set is only separable by a nonlinear surface in the m-dimensional space. To overcome this difficulty, an indirect approach is to map each training point $\mathbf{x}^i \in \mathbb{R}^m$ into a corresponding point $\phi(\mathbf{x}^i)$ in a higher dimensional space $\mathbb{R}^l$ using a nonlinear kernel function $\phi(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^l$, where $l \geq m$, followed by using an SVM to separate $\{\phi(\mathbf{x}^i)\}$ apart in the $\mathbb{R}^l$ space. Vapnik (1998) proposed a soft SVM model with a kernel to seek a hyperplane which separates all mapped training points into two related classes as distinctly as possible.

However, for a given data set, there is no universal rule to automatically choose a most suitable kernel for usage. Moreover, the performance of a soft SVM model with a kernel depends heavily on the selected parameter set embedded in the kernel function (Schölkopf and Smola, 2002). Appropriate parameters are chosen with intuition to produce a minimal cross-validated misclassification rate. Unfortunately, some SVMs with kernels need to compute the inverse of a perturbed kernel matrix in solving its dual problem when the kernel matrix becomes singular (Cristianini and Shawe-Taylor, 2000), or to decompose the kernel matrix for usage in the primal problem. These two techniques usually require extra computational effort to produce an approximated solution.

The objective of this paper is to develop a kernel-free nonlinear SVM model following the logic of soft SVM models. To the best of our knowledge, from optimization point of view, Dagher (2008) proposed a kernel-free quadratic SVM (QSVM) model and tested it against the soft SVM models with either a Quadratic or

Gaussian kernel on some public data sets. However, there is no theoretical analysis of QSVM model and in all Dagher's computational experiments the cross-terms in the objective function of the model was omitted to avoid the computational difficulty. In this paper, following the logic of linear SVMs, to form a soft quadratic surface SVM (SQSSVM) model, we directly introduce the quadratic surface into the soft linear SVMs. The proposed SQSSVM model can handle those difficult cases with a large amount of outliers and noise, which can not be classified well by Dagher's model and other linear SVMs with kernels. We will derive some theoretical properties (such as solvability, uniqueness and support vector representation (Schölkopf and Smola, 2002) of the optimal solution) of the proposed SQSSVM model and conduct computational experiments on randomly generated and public data sets to show that the new model indeed may outperform Dagher's model and soft SVM models with a Quadratic or Gaussian kernel.

The rest of the paper is arranged as following: Section 2 provides a review of the linear SVM and Dagher's QSVM models. The SQSSVM model is proposed in Sec. 3. Some theoretical properties of the SQSSVM model are derived from the optimization point of view in Sec. 4. The SQSSVM model is tested on artificial and five public benchmark classifying data sets with results reported in Secs. 5 and 6. Some concluding remarks are given in Sec. 7.

In this paper, $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^m$ the m-dimensional Euclidean space, $\mathbb{R}^{m\times m}$ the space of all $m \times m$ matrices, and $\|\mathbf{x}\|_2$ means the ℓ_2-norm of vector $\mathbf{x}$.

2. Review of SVM and Dagher's QSVM Models

In this section, we review some basic ideas of linear SVM models for binary classification. Dagher's QSVM models are also reviewed.

Definition 1 (Deng *et al.*, 2013). Consider a training data set $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$, where $\mathbf{x}^i \in \mathbb{R}^m, y^i \in \{+1, -1\}, i = 1, \ldots, n$. If there exists $(\mathbf{u}, d) \in \mathbb{R}^m \times \mathbb{R}^1$ such that

$$y^i(\mathbf{u}^T\mathbf{x}^i + d) \geq 1 \tag{2}$$

for all $i = 1, \ldots, n$, then we say the training data set is linearly separable.

Given a training point $\mathbf{x}^i \in \mathbb{R}^m$, its class label $y^i \in \{+1, -1\}$ and a linear function $f(\mathbf{x}) = \mathbf{u}^T\mathbf{x} + d$, where $(\mathbf{u}, d) \in \mathbb{R}^m \times \mathbb{R}^1$, the following four definitions are then given.

Definition 2 (Dagher, 2008). $\hat{\beta}_i := y^i f(\mathbf{x}^i)$ is called the functional margin at point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = 0$.

Definition 3. The vector $\nabla f(\mathbf{x}^i)$ $(= \mathbf{u})$ is called the gradient direction at point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = f(\mathbf{x}^i)$. If $y^i = +1$ (or -1), the negative (or positive) gradient direction $-\mathbf{u}$ (or $\mathbf{u}$) is called the related gradient direction at point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = f(\mathbf{x}^i)$.

Definition 4 (Dagher, 2008). The related gradient direction at point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = f(\mathbf{x}^i)$ intercepts the hyperplane $f(\mathbf{x}) = 0$ at a point $\mathbf{x}^{\mathbf{B}}$. The length of segment $\overline{\mathbf{x}^i\mathbf{x}^{\mathbf{B}}}$, denoted as β_i, is called the geometrical margin at point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = 0$.

Definition 5. The related gradient direction at point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = f(\mathbf{x}^i)$ intercepts the hyperplane $f(\mathbf{x}) = +1$ (or -1) at a point $\mathbf{x}^{\mathbf{I}}$. The length of segment $\overline{\mathbf{x}^{\mathbf{I}}\mathbf{x}^{\mathbf{B}}}$, denoted as $\bar{\beta}_i$, is called the relative geometrical margin at the point $\mathbf{x}^i$ with respect to $f(\mathbf{x}) = 0$.

Figure 1 illustrates $\mathbf{x}^i, \mathbf{x}^{\mathbf{I}}, \mathbf{x}^{\mathbf{B}}, \hat{\beta}_i, \beta_i$ and $\bar{\beta}_i$ for $m = 2$. In this figure, the red line is the separating line of the two classes. Moreover, the relationship between the functional and geometrical margin at a training point $\mathbf{x}^i$ is $\beta_i = \frac{\hat{\beta}_i}{\|\mathbf{u}\|_2}$ (Dagher, 2008). Also, expression (2) implies that each training point has a no-less-than 1 functional margin. Then, at point $\mathbf{x}^i$, $\bar{\beta}_i = \|\mathbf{x}^{\mathbf{B}} - \mathbf{x}^{\mathbf{I}}\|_2 = \frac{1}{\|\mathbf{u}\|_2}$. Thus, in this situation, for all training points $\mathbf{x}^i, i = 1, \ldots, n$, $\bar{\beta}_1 = \bar{\beta}_2 = \cdots = \bar{\beta}_n$. However, this phenomenon does not happen during the building of the QSSVM models in Sec. 3. The objective of the SVM model can be restated as "to maximize the sum of the relative geometrical margins at all training points with respect to $f(\mathbf{x}) = 0$ subject to the condition that each training point has a no-less-than 1 functional margin". (See Fig. 1, where the distance between the two blue lines is maximized subject to the condition that no training point exists between the two blue lines.)

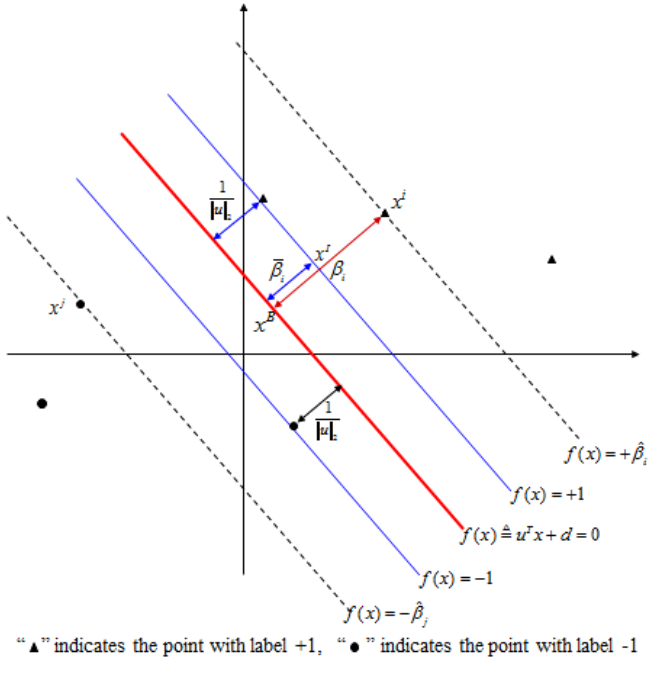

Fig. 1. Building a linear SVM.

Hence, for a linearly separable training data set $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)\}$, the following optimization problem was proposed (Boser *et al.*, 1992):

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{u}\|_2^2 \\ s.t. \quad & y^i(\mathbf{u}^T\mathbf{x}^i + d) \geq 1, \quad i = 1, 2, \ldots, n, \\ & (\mathbf{u}, d) \in \mathbb{R}^m \times \mathbb{R}^1. \end{aligned} \tag{SVM}$$

However, in general, the training data set is not linearly separable but separable by a nonlinear surface (Drucker *et al.*, 1999). Then, we have the following SVM model with a kernel for nonlinear separation (Cortes and Vapnik, 1995):

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{v}\|_2^2 \\ s.t. \quad & y^i(\mathbf{v}^T\phi(\mathbf{x}^i) + d) \geq 1, \quad i = 1, 2, \ldots, n, \\ & (\mathbf{v}, d) \in \mathbb{R}^l \times \mathbb{R}^1, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{KSVM}$$

where $\phi(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^l$, with $m \leq l$, is a nonlinear kernel function. Also, with respect to $\phi(\mathbf{x})$, a kernel for two training points $\mathbf{x}^i$ and $\mathbf{x}^j$ is defined as $K(\mathbf{x}^i, \mathbf{x}^j) = \phi(\mathbf{x}^i)^T\phi(\mathbf{x}^j)$ (Vapnik, 2000). Two well-known kernels are Gaussian kernel $K(\mathbf{x}^i, \mathbf{x}^j) = \exp(-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|_2^2}{2\sigma^2})$ and Quadratic kernel $K(\mathbf{x}^i, \mathbf{x}^j) = (a + (\mathbf{x}^i)^T\mathbf{x}^j)^2$ (Schölkopf and Smola, 2002). Moreover, to handle the training data set with outliers and noise, Vapnik (1998) proposed a soft SVM model with a kernel, by adding a slack variable $\xi_i \geq 0$ for each constraint in (SVM) and a number $\hat{\eta} > 0$ as the penalty value for each ξ_i appeared in the objective function:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{v}\|_2^2 + \hat{\eta}\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y^i(\mathbf{v}^T\phi(\mathbf{x}^i) + d) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n, \\ & (\mathbf{v}, d) \in \mathbb{R}^l \times \mathbb{R}^1, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{SKSVM}$$

Lin (2001) provided an optimization point of view of soft SVMs with kernels. And some variants of soft SVMs with different kernels can be referred to Chen *et al.* (2012), Liu and Yuan (2011), Martin-Barragan *et al.* (2007).

In Dagher (2008), the parameter set $(W, \mathbf{b}, c)$ of a quadratic surface

$$g(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T W\mathbf{x} + \mathbf{b}^T\mathbf{x} + c = 0, \tag{3}$$

where

$$W = W^T = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{12} & w_{22} & \cdots & w_{2m} \\ \vdots & & & \\ w_{1m} & w_{2m} & \cdots & w_{mm} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m, \quad \text{and} \quad c \in \mathbb{R},$$

was found directly to separate the n training points into two classes, without using any kernel function. Dagher (2008) first proposed the following QSVM model:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}(W\mathbf{x}^i)^T W\mathbf{x}^i + 2\sum_{i=1}^{n}\mathbf{b}^T W\mathbf{x}^i + n\mathbf{b}^T\mathbf{b} \\ s.t. \quad & y^i\left(\frac{1}{2}(\mathbf{x}^i)^T W\mathbf{x}^i + \mathbf{b}^T\mathbf{x}^i + c\right) \geq 1, \quad i = 1,\ldots,n, \\ & W = W^T \in \mathbb{R}^{m\times m}, \quad (\mathbf{b}, c) \in \mathbb{R}^m \times \mathbb{R}^1. \end{aligned} \tag{QSVM1}$$

Moreover, to avoid the computational difficulty, Dagher proposed the following modified QSVM model by omitting the cross-terms in the objective function:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}(W\mathbf{x}^i)^T W\mathbf{x}^i + n\mathbf{b}^T\mathbf{b} \\ s.t. \quad & y^i\left(\frac{1}{2}(\mathbf{x}^i)^T W\mathbf{x}^i + \mathbf{b}^T\mathbf{x}^i + c\right) \geq 1, \quad i = 1,\ldots,n, \\ & W = W^T \in \mathbb{R}^{m\times m}, \quad (\mathbf{b}, c) \in \mathbb{R}^m \times \mathbb{R}^1. \end{aligned} \tag{QSVM2}$$

and then used this model for all the numerical tests and experiments. In Sec. 3, following the logic of linear SVM models, we will introduce the quadratic surface directly into soft SVM model to propose a kernel-free SQSSVM model.

3. Quadratic Surface Support Vector Machine Models

In this section, a quadratic surface is used to separate the training data set into two classes instead of a hyperplane, to develop a SQSSVM model.

The proposed quadratic surface SVM (QSSVM) intends to find the parameter set $(W, \mathbf{b}, c)$ of a quadratic surface $g(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T W\mathbf{x} + \mathbf{b}^T\mathbf{x} + c = 0$ that separates the n training points $\{\mathbf{x}^1,\ldots,\mathbf{x}^i,\ldots,\mathbf{x}^n\}$ into two classes, with a maximum separation.

Definition 6. Consider a training data set $\{(\mathbf{x}^1, y^1),\ldots,(\mathbf{x}^i, y^i),\ldots,(\mathbf{x}^n, y^n)\}$, where $\mathbf{x}^i \in \mathbb{R}^m$, $y^i \in \{+1, -1\}, i = 1,\ldots,n$. If there exist $W = W^T \in \mathbb{R}^{m\times m}$, $(\mathbf{b}, c) \in \mathbb{R}^m \times \mathbb{R}^1$ such that

$$y^i\left(\frac{1}{2}(\mathbf{x}^i)^T W\mathbf{x}^i + \mathbf{u}^T\mathbf{x}^i + c\right) \geq 1, \tag{4}$$

for all $i = 1,\ldots,n$, then we say the training data set is quadratically separable.

Given a training point $\mathbf{x}^i \in \mathbb{R}^m$, its class label $y^i \in \{+1, -1\}$ and a quadratic function $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T W\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, where $W = W^T \in \mathbb{R}^{m\times m}$ and $(\mathbf{b}, c) \in \mathbb{R}^m \times \mathbb{R}^1$,

the following four definitions are then given:

Definition 7 (Dagher, 2008). $\hat{\gamma}_i := y^i g(\mathbf{x}^i)$ is called the functional margin at point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = 0$.

Definition 8. The vector $\nabla g(\mathbf{x}^i)$ $(= W\mathbf{x}^i + \mathbf{b})$ is called the gradient direction at point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = g(\mathbf{x}^i)$. If $y^i = +1$ (or -1), the negative (or positive) gradient direction $-\nabla g(\mathbf{x}^i)$ (or $\nabla g(\mathbf{x}^i)$) is called the related gradient direction at point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = g(\mathbf{x}^i)$.

Definition 9 (Dagher, 2008). The related gradient direction at point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = g(\mathbf{x}^i)$ intercepts the quadratic surface $g(\mathbf{x}) = 0$ at a point $\mathbf{x}^{\mathbf{B}}$. The length of the segment $\overline{\mathbf{x}^i\mathbf{x}^{\mathbf{B}}}$, denoted as γ_i, is called the geometrical margin at point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = 0$.

Definition 10. The related gradient direction at point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = g(\mathbf{x}^i)$ intercepts the surface $g(\mathbf{x}) = +1$ (or -1) at a point $\mathbf{x}^{\mathbf{I}}$. The length of the segment $\overline{\mathbf{x}^{\mathbf{I}}\mathbf{x}^{\mathbf{B}}}$, denoted as $\bar{\gamma}_i$, is called the relative geometrical margin at the point $\mathbf{x}^i$ with respect to $g(\mathbf{x}) = 0$.

Figure 2 illustrates the $\mathbf{x}^i$, $\mathbf{x}^{\mathbf{I}}$, $\mathbf{x}^{\mathbf{B}}$, $\hat{\gamma}_i$, γ_i and $\bar{\gamma}_i$ for $m = 2$, where $g(\mathbf{x}) = 0$ is the red separating quadratic curve. Expression (4) implies that $\hat{\gamma}_i = y^i g(\mathbf{x}^i) \geq 1$, $i = 1, \ldots, n$, which indicates that each training point has a no-less-than 1 functional margin.

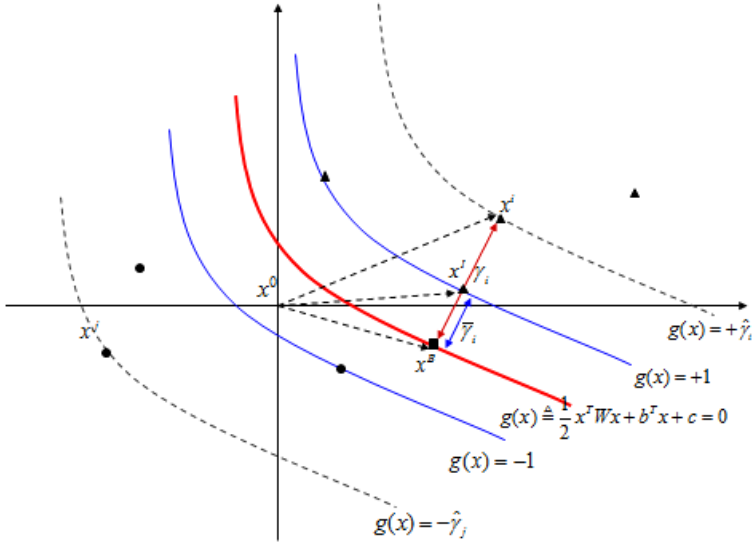

Fig. 2. Building a quadratic surface SVM.

Moreover, the relative geometrical margin $\bar{\gamma}_i$ at the point $\mathbf{x}^i$ can be approximated as follows. Let $\mathbf{x^0}$ be the origin of $\mathbb{R}^m$. We may see from Fig. 2 that $\overrightarrow{\mathbf{x^0x^B}} = \overrightarrow{\mathbf{x^0x^I}} + \overrightarrow{\mathbf{x^Ix^B}}$ and $\overrightarrow{\mathbf{x^Ix^B}} = -\bar{\gamma}_i \frac{\nabla g(\mathbf{x}^i)}{\|\nabla g(\mathbf{x}^i)\|_2}$. Thus, $\mathbf{x^B} = \mathbf{x^I} - \bar{\gamma}_i \frac{\nabla g(\mathbf{x}^i)}{\|\nabla g(\mathbf{x}^i)\|_2}$. Taylor's expansion says that $g(\mathbf{x^B}) \approx g(\mathbf{x^I}) + \nabla g(\mathbf{x^I})^T(\mathbf{x^B} - \mathbf{x^I})$. Noting that $g(\mathbf{x^B}) = 0$ and $g(\mathbf{x^I}) = 1$, then we have

$$0 \approx 1 + \nabla g(\mathbf{x^I})^T(\mathbf{x^B} - \mathbf{x^I}) = 1 + \nabla g(\mathbf{x^I})^T \left(-\bar{\gamma}_i \frac{\nabla g(\mathbf{x}^i)}{\|\nabla g(\mathbf{x}^i)\|_2}\right),$$

which infers that $\bar{\gamma}_i \approx \frac{\|\nabla g(\mathbf{x}^i)\|_2}{\nabla g(\mathbf{x^I})^T \nabla g(\mathbf{x}^i)}$. Similarly,

$$g(\mathbf{x^I}) \approx g(\mathbf{x}^i) + \nabla g(\mathbf{x}^i)^T(\mathbf{x^I} - \mathbf{x}^i)$$

$$g(\mathbf{x}^i) \approx g(\mathbf{x^I}) + \nabla g(\mathbf{x^I})^T(\mathbf{x}^i - \mathbf{x^I})$$

and $\mathbf{x^I} - \mathbf{x}^i = -\frac{\gamma_i - \bar{\gamma}_i}{\|\nabla g(\mathbf{x}^i)\|_2} \nabla g(\mathbf{x}^i)$, which is inferred by $\overrightarrow{\mathbf{x^0x^I}} - \overrightarrow{\mathbf{x^0x}^i} = \overrightarrow{\mathbf{x}^i\mathbf{x^I}}$. Hence $\nabla g(\mathbf{x^I})^T \nabla g(\mathbf{x}^i) \approx \nabla g(\mathbf{x}^i)^T \nabla g(\mathbf{x}^i)$. Consequently, at point $\mathbf{x}^i$, $\bar{\gamma}_i = \|\mathbf{x^B} - \mathbf{x^I}\|_2 \approx \frac{\|\nabla g(\mathbf{x}^i)\|_2}{\nabla g(\mathbf{x^I})^T \nabla g(\mathbf{x}^i)} \approx \frac{1}{\|\nabla g(\mathbf{x}^i)\|_2} = \frac{1}{\|W\mathbf{x}^i + \mathbf{b}\|_2}$. Note that, in general, $\bar{\gamma}_i \neq \bar{\gamma}_j$ for $\mathbf{x}^i \neq \mathbf{x}^j$. This situation is different from that in the SVM model.

The objective of the QSSVM can be stated as "to maximize the sum of the approximated relative geometrical margins at all training points with respect to $g(\mathbf{x}) = 0$ subject to the condition that each training point has a no-less-than 1 functional margin". (See Fig. 2, where the distance between the two blue curves is maximized subject to the condition that no training point exists between the two blue curves.)

Thus, for a quadratically separable training data set $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)\}$, we consider the following quadratic surface SVM model:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \|W\mathbf{x}^i + \mathbf{b}\|_2^2 \\ s.t. \quad & y^i \left(\frac{1}{2}(\mathbf{x}^i)^T W \mathbf{x}^i + \mathbf{b}^T \mathbf{x}^i + c\right) \geq 1, \quad i = 1, \ldots, n, \qquad \text{(QSSVM)} \\ & W = W^T \in \mathbb{R}^{m \times m}, \quad (\mathbf{b}, c) \in \mathbb{R}^m \times \mathbb{R}^1. \end{aligned}$$

However, if the training data set is not quadratically separable, for a separating quadratic surface $g(\mathbf{x}) = 0$, one of the following two situations would occur for some training points:

Situation 1 : for point $\mathbf{x}^i, y^i = -1$, but $\frac{1}{2}(\mathbf{x}^i)^T W \mathbf{x}^i + \mathbf{b}^T \mathbf{x}^i + c > -1$,

Situation 2 : for point $\mathbf{x}^j, y^j = +1$, but $\frac{1}{2}(\mathbf{x}^j)^T W \mathbf{x}^j + \mathbf{b}^T \mathbf{x}^j + c < +1$.

These points are referred to as the outliers of the data set with respect to $g(\mathbf{x}) = 0$ (Brooks, 2011). In this case, the proposed model (QSSVM) would become infeasible, since no quadratic surface can separate all training points into their

corresponding classes correctly. To take care of this situation, similar to the development of the soft SVM model, we add a slack variable $\xi_i \geq 0$ for each constraint in (QSSVM) and a number $\hat{\eta} > 0$ as the penalty value for each ξ_i in the objective function. Then we propose the following SQSSVM model:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \|W\mathbf{x}^i + \mathbf{b}\|_2^2 + \hat{\eta} \sum_{i=1}^{n} \xi_i \\ s.t. \quad & y^i \left(\frac{1}{2} (\mathbf{x}^i)^T W \mathbf{x}^i + \mathbf{b}^T \mathbf{x}^i + c \right) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \\ & \xi_i \geq 0, \quad i = 1, \ldots, n, \\ & W = W^T \in \mathbb{R}^{m \times m}, \quad (\mathbf{b}, c) \in \mathbb{R}^m \times \mathbb{R}^1. \end{aligned} \tag{SQSSVM}$$

The main advantages of the model (SQSSVM) over the model (QSSVM) are reflected in the generalization ability and robustness to outliers.

Note that in models (QSSVM) and (SQSSVM), the matrix W is symmetric. To simplify these two models, we may convert each of them into an equivalent form as follows. First, let $\mathfrak{W}$ be the vector formed by taking the $\frac{m^2+m}{2}$ elements of the upper elements of the upper triangle part of the matrix W, i.e.,

$$\mathfrak{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} & w_{22} & \cdots & w_{2m} & \cdots & w_{mm} \end{bmatrix}^T \in \mathbb{R}^{\frac{m^2+m}{2}}. \tag{5}$$

Then, construct an $m \times (\frac{m^2+m}{2})$ matrix M^i for the training point $\mathbf{x}^i = [x_1^i, x_2^i, \ldots, x_m^i]^T \in \mathbb{R}^m$ as follows. For the j-th row of M^i in $\mathbb{R}^{\frac{m^2+m}{2}}, j = 1, \ldots, m$, check the elements of $\mathfrak{W}$ one by one. If the pth element of $\mathfrak{W}$ is w_{jk} or w_{kj} for some $k = 1, 2, \ldots, m$, then assign the pth element of the jth row of M^i to be x_k^i. Otherwise, assign it to be 0.

Also let matrix $H^i = [M^i, I] \in \mathbb{R}^{m \times (\frac{m^2+m}{2} + m)}, i = 1, \ldots, n$, where I is the m-dimensional identity matrix. Then, define the vector of variables $\mathbf{z} = \begin{bmatrix} \mathfrak{W} \\ \mathbf{b} \end{bmatrix} \in \mathbb{R}^{\frac{m^2+3m}{2}}$ and the vector $\mathbf{s}^i = [\frac{1}{2}x_1^i x_1^i, \ldots, x_1^i x_m^i, \frac{1}{2}x_2^i x_2^i, \ldots, x_2^i x_m^i, \ldots, \frac{1}{2}x_{m-1}^i x_{m-1}^i, x_{m-1}^i x_m^i, \frac{1}{2}x_m^i x_m^i, x_1^i, x_2^i, \ldots, x_m^i] \in \mathbb{R}^{\frac{(m+1)m}{2}+m}$. So the objective of the model (QSSVM) becomes $\sum_{i=1}^n \|W\mathbf{x}^i + \mathbf{b}\|_2^2 = \sum_{i=1}^n \|H^i\mathbf{z}\|_2^2 = \sum_{i=1}^n (H^i\mathbf{z})^T(H^i\mathbf{z}) = \sum_{i=1}^n \mathbf{z}^T(H^i)^T H^i\mathbf{z} = \mathbf{z}^T(\sum_{i=1}^n (H^i)^T H^i)\mathbf{z}$. Let $G = \sum_{i=1}^n (H^i)^T H^i \in \mathbb{R}^{(\frac{m^2+3m}{2}) \times (\frac{m^2+3m}{2})}$, then model (QSSVM) becomes

$$\begin{aligned} \min \quad & \mathbf{z}^T G \mathbf{z} \\ s.t. \quad & y^i((\mathbf{s}^i)^T\mathbf{z} + c) \geq 1, \quad i = 1, \ldots, n, \\ & (\mathbf{z}, c) \in \mathbb{R}^{\frac{m^2+3m}{2}} \times \mathbb{R}^1. \end{aligned} \tag{QSSVM$'$}$$

Similarly, the model (SQSSVM) can be reformulated as

$$\begin{aligned} \min \quad & \mathbf{z}^T G\mathbf{z} + \hat{\eta}\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y^i((\mathbf{s}^i)^T\mathbf{z} + c) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \qquad \text{(SQSSVM}')\\ & (\mathbf{z}, c) \in \mathbb{R}^{\frac{m^2+3m}{2}} \times \mathbb{R}^1, \quad \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

Note that G is positive semidefinite since $\mathbf{z}^T G\mathbf{z} = \sum_{i=1}^{n} \|H^i\mathbf{z}\|_2^2 \geq 0$ for any $\mathbf{z} \in \mathbb{R}^{\frac{m^2+3m}{2}}$. Consequently, both of models (QSSVM$'$) and (SQSSVM$'$) are linearly constrained convex quadratic optimization problems, which can be solved efficiently (Fang and Puthenpura, 1993).

4. Theoretical Properties of the SQSSVM Model

In this section, we study some theoretical properties of the model (SQSSVM$'$). The solvability of the model (SQSSVM$'$) is first studied as follows.

Theorem 1. *For any given training data set* $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$ *and* $\hat{\eta} > 0$, *there exists an optimal solution to the model* (*SQSSVM*$'$) *with a finite objective value.*

If the training data set is quadratically separable, it is easy to verify that the model (QSSVM$'$) has at least one optimal solution using a similar proof of Theorem 1. Then, the next result states the relationship between the optimal solutions of models (QSSVM$'$) and (SQSSVM$'$).

Theorem 2. *For any given* $\hat{\eta} > 0$, *let* $(\mathbf{z}^{\hat{\eta}}, c^{\hat{\eta}}, \boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}})$ *be an optimal solution of model* (*SQSSVM*$'$) *and assume that the sequence* $\{(\mathbf{z}^{\hat{\eta}}, c^{\hat{\eta}}, \boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}})\}$ *converges to* $(\mathbf{z}^*, c^*, \boldsymbol{\xi}^*)$ *as* $\hat{\eta} \to \infty$. *If the training data set is quadratically separable, then* $\boldsymbol{\xi}^* = \mathbf{0}$ (*where* $\mathbf{0} = (0, \ldots, 0)^T \in \mathbb{R}^n$) *and* $(\mathbf{z}^*, c^*)$ *is an optimal solution of model* (*QSSVM*$'$).

Let $\mathcal{F}^* = \{(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathbb{R}^{\frac{m^2+3m}{2}} \times \mathbb{R}^1 \times \mathbb{R}^n | (\mathbf{z}, c, \boldsymbol{\xi})$ is an optimal solution to the model (SQSSVM$'$)$\}$. Then $\mathcal{F}^* \neq \emptyset$ by Theorem 1. Moreover, we have the next three results.

Theorem 3. *For any given training data set* $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$ *and* $\hat{\eta} > 0$, *if* G *is positive definite, then the optimal solution of model* (*SQSSVM*$'$) *is unique with respect to the variable* $\mathbf{z}$.

Thus, for any given training data set, if G is positive definite, the main characteristics of the separating quadratic surface are uniquely determined by the optimal solution of model (SQSSVM$'$) with respect to the variable $\mathbf{z}$.

Theorem 4. *For any given training data set* $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$ *and* $\hat{\eta} > 0$, *if* G *is positive definite, then there exist constants* $\underline{c}$ *and* $\overline{c}$ *such that* $\underline{c} \leq c \leq \overline{c}$, *for any* $(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathcal{F}^*$.

Theorem 5. *If the training data set is quadratically separable and G is positive definite, then for any given sufficiently large $\hat{\eta} > 0$, the optimal solution of model (SQSSVM′) is unique with respect to the variable c.*

From Theorems 3 and 5, we know that if the training data set is quadratically separable and G is positive definite, then, for any sufficiently large $\hat{\eta} > 0$, the model (SQSSVM′) generates a unique separating quadratic surface. Generally speaking, for any given training data set with G being positive definite, we may solve the model (SQSSVM′) with a sufficiently large $\hat{\eta} > 0$ to generate a separating quadratic surface for binary classification.

Note that if the matrix G in model (SQSSVM′) is only positive semidefinite, we can always append a perturbation such that the matrix $G + \epsilon I$ ($\epsilon > 0$, I is the identity matrix) becomes positive definite. Then, consider the following perturbed model:

$$\begin{aligned} \min \quad & \mathbf{z}^T(G + \epsilon I)\mathbf{z} + \hat{\eta}\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y^i((\mathbf{s}^i)^T\mathbf{z} + c) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n, \\ & \xi_i \geq 0, \quad i = 1, \ldots, n, \qquad \text{(SQSSVM′-}\epsilon\text{)} \\ & (\mathbf{z}, c) \in \mathbb{R}^{\frac{m^2+3m+2}{2}}. \end{aligned}$$

Similar to the proof of Theorem 1, we can verify that the model (SQSSVM′-ϵ) has at least one optimal solution. Let $(\mathbf{z}^\epsilon, c^\epsilon, \boldsymbol{\xi}^\epsilon)$ be an optimal solution of model (SQSSVM′-ϵ), then the model (SQSSVM′) and its perturbed model (SQSSVM′-ϵ) are related by the next two results.

Lemma 6. *For any given training data set $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$ and $\hat{\eta} > 0$, if the optimal value of model (SQSSVM′) is v and the optimal value of model (SQSSVM′-ϵ) is v_ϵ, for a given $\epsilon > 0$, then $v_\epsilon \to v$ as $\epsilon \to 0$.*

Remark 1. For any given $\hat{\eta} > 0$ and $0 < \epsilon_1 < \epsilon_2$, we have $v_{\epsilon_1} \leq (\mathbf{z}^{\epsilon_2})^T G\mathbf{z}^{\epsilon_2} + \epsilon_1(\mathbf{z}^{\epsilon_2})^T(\mathbf{z}^{\epsilon_2}) + \hat{\eta}\sum_{i=1}^{n}\xi_i^{\epsilon_2} < (\mathbf{z}^{\epsilon_2})^T G\mathbf{z}^{\epsilon_2} + \epsilon_2(\mathbf{z}^{\epsilon_2})^T(\mathbf{z}^{\epsilon_2}) + \hat{\eta}\sum_{i=1}^{n}\xi_i^{\epsilon_2} = v_{\epsilon_2}$. Hence the sequence $\{v_\epsilon\}$ monotonically decreases to v as $\epsilon \searrow 0$.

Theorem 7. *For any given training data set $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^n, y^n)\}$ and $\hat{\eta} > 0$, if the sequence $\{(\mathbf{z}^\epsilon, c^\epsilon, \boldsymbol{\xi}^\epsilon)\}$ converges to $(\mathbf{z}^0, c^0, \boldsymbol{\xi}^0)$ as $\epsilon \to 0$, then $(\mathbf{z}^0, c^0, \boldsymbol{\xi}^0) \in \mathcal{F}^*$ and $(\mathbf{z}^0)^T\mathbf{z}^0 \leq \mathbf{z}^T\mathbf{z}$, for any $(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathcal{F}^*$.*

In conclusion, for a training data set with G being positive semidefinite only, we may solve the perturbed model (SQSSVM′-ϵ) with a sufficiently small $\epsilon > 0$ to generate a separating quadratic surface for binary classification. Hence, without loss of generality, G is supposed to be positive definite in the SQSSVM model.

Furthermore, the dual problem of (SQSSVM′) can be formulated as:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{4} \left(\sum_{i=1}^{n} \alpha_i y^i \mathbf{s}^i \right)^T G^{-1} \left(\sum_{i=1}^{n} \alpha_i y^i \mathbf{s}^i \right) \\ s.t. \quad & \sum_{i=1}^{n} \alpha_i y^i = 0, \qquad \text{(DSQSSVM)} \\ & 0 \le \alpha_i \le \hat{\eta}, \quad i = 1, \ldots, n. \end{aligned}$$

Problems (SQSSVM′) and (DSQSSVM) are both linearly constrained convex quadratic optimization problems, no duality gap exists. The optimality conditions (KKT conditions) for problems (SQSSVM′) and (DSQSSVM) are:

$$y^i((\mathbf{s}^i)^T \mathbf{z} + c) \ge 1 - \xi_i, \quad \xi_i \ge 0, \quad i = 1, \ldots, n,$$

$$\sum_{i=1}^{n} \alpha_i y^i = 0, \quad 0 \le \alpha_i \le \hat{\eta}, \quad i = 1, \ldots, n,$$

$$\alpha_i(y^i((\mathbf{s}^i)^T \mathbf{z} + c) - 1 + \xi_i) = 0, \quad \xi_i(\hat{\eta} - \alpha_i) = 0, \quad i = 1, \ldots, n.$$

We can solve the above optimality conditions to get the optimal solutions $(\mathbf{z}^*, c^*, \boldsymbol{\xi}^*)$ and $(\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$ of problems (SQSSVM′) and (DSQSSVM), respectively. Then we have the following results by Lagrangian dual theory:

$$\mathbf{z}^* = \sum_{i=1}^{n} \alpha_i^* \left(\frac{1}{2} y^i G^{-1} \mathbf{s}^i \right).$$

- If $\alpha_i^* = 0$; then $\xi_i^* = 0$ and $y^i((\mathbf{s}^i)^T \mathbf{z}^* + c^*) \ge 1$, which indicate that point $\mathbf{x}^i$ is inside the scope of one class.
- If $0 < \alpha_i^* < \hat{\eta}$; then $\xi_i^* = 0$ and $y^i((\mathbf{s}^i)^T \mathbf{z}^* + c^*) = 1$, which indicate that point $\mathbf{x}^i$ is a support vector (Schölkopf and Smola, 2002) on the boundary.
- If $\alpha_i^* = \hat{\eta}$; then $\xi_i^* \ge 0$ and $y^i((\mathbf{s}^i)^T \mathbf{z}^* + c^*) \le 1$, which indicate that point $\mathbf{x}^i$ may be an outlier.

Thus, we can solve the problem (DSQSSVM) to obtain its optimal solution $(\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$ and then get parameters $(\mathbf{z}^*, c^*)$ of the separating quadratic surface by expression (5) and $c^* = 1/y^{sv} - (\mathbf{s}^{sv})^T \mathbf{z}^*$ (for any support vector). The SQSSVM model is related to the support vector theory and the parameters $(\mathbf{z}^*, c^*)$ of obtained classifier is a linear combination of the opportune evaluations of the training data set $\{(\mathbf{x}^i, y^i), i = 1, \ldots, n\}$.

5. Computational Experiments on Effectiveness and Efficiency

In this section, we test the classification accuracy and efficiency of the SQSSVM model on three-dimensional (3D) artificial and real-world classifying data sets. We also test Dagher's modified QSVM model, our QSSVM model and soft SVM models

with a Quadratic or Gaussian kernel for comparisons. All computational experiments in this paper are performed using MATLAB (R2013a) software on a personal laptop equipped with Intel Core i5 2.40 GHz CPU, 2.5 GB usable RAM and Microsoft Windows 7 Professional. The SQSSVM, QSSVM and Dagher's modified QSVM models are implemented using the interior point algorithm in the module "quadprog" of MATLAB, while the soft SVM models with a Quadratic or Gaussian kernel are implemented using the sequential minimal optimization algorithm (Chang and Lin, 2011) in the MATLAB code. For all models, we use the grid method to find the best parameter $\hat{\eta}$ and a (i.e., the parameter in Quadratic kernel): $\log_2 \hat{\eta}, \log_2 a \in \{-4, -5, \ldots, 9, 10\}$, the parameter σ of Gaussian kernel is set to be the median of the between-class pairwise Euclidean distances of training points as in Liu and Yuan (2011). Note that, the recorded CPU time in this paper doesn't include the time of tuning parameters in the model.

First, we generate 90 different 3D quadratic surfaces in three types, i.e., 30 quadratic surfaces for each type. Since a 3D quadratic surface $\frac{1}{2}\mathbf{x}^T W\mathbf{x}+\mathbf{b}^T\mathbf{x}+c=0$ can be characterized by the eigenvalues of the matrix W, the first, second and third type has one, two and three positive eigenvalues, respectively. For each 3D quadratic surface, 200 points (labeled as Class 1) are generated on one side and another 200 points (labeled as Class 2) are generated on the other side. In this way, a total of 90 3D artificial classifying data sets are generated. One example of the second type artificial data set is shown in Fig. 3, where the red and blue points are labeled as Class 1 and Class 2, respectively.

Next, for each 3D artificial classifying data set, we randomly pick $k\%$ of the 400 points ($\frac{k}{2}\%$ from Class 1 and $\frac{k}{2}\%$ from Class 2) as the training data set. The soft SVM models with a Quadratic or Gaussian kernel, Dagher's modified QSVM model, QSSVM model and SQSSVM model are trained, respectively, using the training data set to generate parameters in corresponding classifiers. Then these classifiers are used to classify the remaining $400(1-k\%)$ points in the 3D artificial data set and misclassification rates are calculated. In our experiments, like most

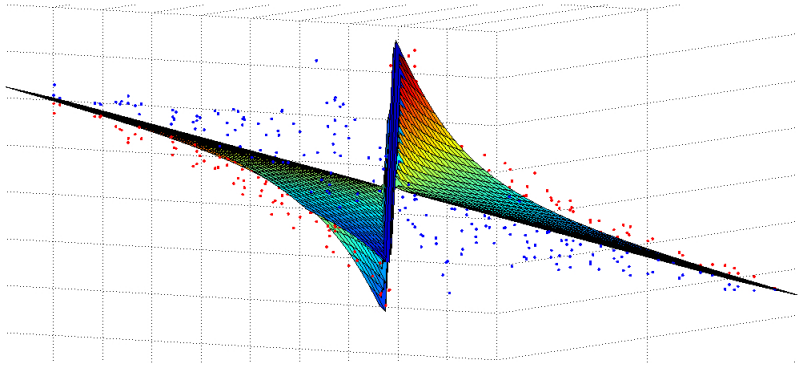

Fig. 3. A second type artificial data set.

common practices reported in (Dagher, 2008; Liu and Yuan, 2011), k is chosen to be 10, 20 and 40. To be statistically meaningful, for each given k, we conduct tests with selected k% of the 400 points for all 90 3D artificial classifying data sets. For each model, the mean, standard deviation (std), minimum (min) and maximum (max) of misclassification rates (MR) and average CPU time of the 90 data sets are reported in Table 2. A smaller misclassification rate indicates that the corresponding model performs better for binary classification. The soft SVM models with Quadratic and Gaussian kernel are denoted by "SVM_GausKer" and "SVM_QuadKer", respectively, in all tables of this paper.

To further test the classification accuracy and efficiency of the SQSSVM model, five public benchmark databases, often used in (Brooks, 2011; Dagher, 2008; Liu and Yuan, 2011) from the most popular data sets in UCI repository (Bache and Lichman, 2013), are used for additional experiments. For binary classification, if one database contains more than two classes, we only choose two classes of this database for test. The descriptions of these real-world data sets for binary classifications are listed in Table 1.

For each of these real-world Data sets, we randomly select k% as the training data set, the remaining points are included as the testing data set. The SQSSVM model is first trained using the training data set to generate the parameters of the classifying quadratic surface. Then the quadratic surface is used to classify the testing data set and the misclassification rate is calculated. Like before, k is chosen to be 10, 20 and 40. To be statistically meaningful, for each given k, we repeat the test with randomly selected k% points for 100 times. The mean and standard deviation of the calculated misclassification rates and average CPU time of the 100 experiments are reported in Tables 2 and 3. For fair comparisons, we carried out similar experiments for the other four models using the same training and testing data set, and recorded the results in Tables 2 and 3. Furthermore, for the Skin data set of large size, k is set to be 0.25%, 0.5% and 1% to show the effectiveness of the models and to avoid memory overflows.

From Tables 2 and 3, we have the following observations: For most experiments on the tested data sets, the mean and standard deviation of misclassification rates produced by the SQSSVM model are smaller than those produced by the other models; there is not much difference among the CPU time of all models; the

Table 1. Descriptions of real-world data sets.

Data set	Class 1		Class 2	
	Name	# of points	Name	# of points
Iris	Versicolour	50	Virginica	50
Car Evaluation	Unacc	1210	Acc	384
Wisconsin Breast Cancer	Benign	458	Malignant	241
Seeds	Kama	70	Rosa	70
Skin	Skin	194198	Non-skin	50859

Table 2. Artificial, iris and car evaluation data test.

k%	Model type	Artificial data		Iris data		Car evaluation data	
		MR(%) mean/std	CPU(s)	MR(%) mean/std	CPU(s)	MR(%) mean/std	CPU(s)
10%	SVM_GausKer	49.21/2.27	0.03	14.07/7.00	0.02	14.10/1.63	0.14
	SVM_QuadKer	49.84/2.56	0.13	16.21/8.22	0.04	13.72/4.62	0.38
	QSVM2	15.81/14.65	0.11	11.12/5.56	0.02	11.35/2.17	0.18
	QSSVM	14.62/14.09	0.12	10.87/5.41	0.02	9.99/1.89	0.19
	SQSSVM	13.53/11.82	0.13	9.56/4.88	0.02	8.60/1.56	0.19
20%	SVM_GausKer	47.24/2.16	0.05	8.82/4.28	0.03	9.13/0.99	0.21
	SVM_QuadKer	48.29/2.47	0.21	11.39/5.55	0.05	8.46/5.64	0.57
	QSVM2	11.55/13.73	0.17	7.93/4.32	0.02	8.88/2.93	0.29
	QSSVM	10.30/13.55	0.19	7.74/4.18	0.02	7.54/2.86	0.30
	SQSSVM	8.14/8.12	0.22	7.21/3.49	0.02	6.38/1.09	0.32
40%	SVM_GausKer	45.61/2.02	0.07	6.85/2.68	0.08	6.91/8.97	0.48
	SVM_QuadKer	47.35/2.21	0.30	9.15/4.51	0.09	6.81/8.94	1.35
	QSVM2	9.97/15.63	0.28	8.41/5.51	0.05	15.94/5.18	0.71
	QSSVM	8.87/15.59	0.34	8.15/5.39	0.06	15.48/5.08	0.72
	SQSSVM	5.19/6.88	0.41	5.33/3.22	0.06	4.61/0.95	0.73

Table 3. Wisconsin breast cancer, seeds and skin data test.

Model type	WBC data			Seeds data			Skin data		
	k%	MR(%) mean/std	CPU(s)	k%	MR(%) mean/std	CPU(s)	k%	MR(%) mean/std	CPU(s)
SVM_GausKer	10%	5.80/0.95	0.05	10%	11.21/5.32	2.76	0.25%	1.34/0.23	1.61
SVM_QuadKer		7.93/2.23	0.06		12.97/5.99	3.99		1.81/0.40	0.76
QSVM2		6.36/2.00	0.16		10.56/3.61	3.01		0.73/0.54	3.10
QSSVM		5.51/1.71	0.16		10.33/3.46	3.05		0.71/0.50	3.21
SQSSVM		4.63/1.18	0.17		9.56/1.97	3.11		0.47/0.21	3.38
SVM_GausKer	20%	4.81/0.88	0.07	20%	9.16/2.61	3.29	0.5%	1.02/0.06	2.06
SVM_QuadKer		6.59/1.86	0.08		9.86/5.11	5.12		1.26/0.13	1.07
QSVM2		6.52/1.81	0.25		8.32/2.51	7.41		0.67/1.84	3.32
QSSVM		5.26/1.37	0.26		8.22/2.45	7.48		0.64/1.78	3.49
SQSSVM		3.74/0.64	0.27		7.92/2.39	7.56		0.29/0.09	3.77
SVM_GausKer	40%	3.61/0.59	0.11	40%	7.89/1.69	3.60	1%	0.90/0.04	2.91
SVM_QuadKer		5.27/0.91	0.17		7.97/4.09	5.91		1.01/0.08	1.65
QSVM2		6.86/1.49	0.44		7.06/2.14	9.19		0.56/0.60	3.93
QSSVM		4.40/0.97	0.45		7.01/2.13	9.28		0.52/0.53	4.12
SQSSVM		3.14/0.30	0.47		6.81/1.93	9.41		0.21/0.03	4.41

QSSVM model produces more accurate classification than Dagher's modified QSVM model.

6. Computational Experiments on Robustness

For real world applications, available data sets usually contain outliers (Brooks, 2011), (Brooks, 2011), i.e., the points mislabeled on the wrong side. Hence, in this

the effectiveness and robustness of the SQSSVM model using the artificial training data with outliers.

For each 3D artificial classifying data set, we randomly pick $k\%$ of all 400 points as the training data set like before, the remaining points are included as the testing data set. Then $p\%$ of all points in this training data set are randomly selected and deliberately mislabeled as the outliers. Also, like most common practices reported in Brooks (2011), p is chosen to be 5, 10 and 15. Afterwards, for each given p, all five models are trained respectively using the training data set with $p\%$ outliers to generate the parameters in the corresponding classifiers. Then, we used the classifiers to classify the corresponding testing data set and calculated the misclassification rates. Like before, k is chosen to be 10, 20 and 40. For each given k and p, the tests are repeated with selected $k\%$ of the 400 points for all 90 3D artificial training data sets with $p\%$ outliers. Finally, for p being 5, 10 and 15, the computational results are reported in Table 4. To be compared, the computational results in Table 2 are also included in Table 4 as p being 0.

From Table 4, we have the following observations: For all experiments on the artificial data sets, the mean of the misclassification rates produced by the SQSSVM model is much smaller than that produced by Dagher's modified QSVM model, the QSSVM model and soft SVM models with a Quadratic or Gaussian kernel; as the size of the training data set or the percentage of outliers increases, the superiority of the SQSSVM model in classification accuracy becomes more evident, and the performance of Dagher's modified QSVM model gets much worse.

7. Conclusions and Discussions

In this paper, we have proposed an SQSSVM model for data classification directly using a quadratic function for separation. We have not only studied the theoretical

Table 4. Artificial training data set with $p\%$ outliers.

$k\%$	Model type	Mean misclassification rate (%)				CPU time (s)			
		$p=0$	$p=5$	$p=10$	$p=15$	$p=0$	$p=5$	$p=10$	$p=15$
10%	SVM_GausKer	49.21	47.83	47.86	48.45	0.03	0.03	0.04	0.03
	SVM_QuadKer	49.84	48.36	48.77	48.81	0.13	0.13	0.14	0.15
	QSVM2	15.81	36.56	39.67	43.76	0.11	0.11	0.12	0.13
	QSSVM	14.62	35.97	39.32	43.63	0.12	0.12	0.13	0.13
	SQSSVM	13.53	18.61	20.10	23.98	0.13	0.13	0.12	0.14
20%	SVM_GausKer	47.24	46.98	47.56	47.80	0.05	0.05	0.06	0.05
	SVM_QuadKer	48.29	47.85	48.61	48.87	0.21	0.21	0.22	0.22
	QSVM2	11.55	41.40	42.88	44.70	0.17	0.17	0.18	0.18
	QSSVM	10.30	40.82	42.53	44.56	0.19	0.19	0.20	0.19
	SQSSVM	8.14	11.56	14.47	15.29	0.22	0.22	0.23	0.21
40%	SVM_GausKer	45.61	45.25	46.61	46.03	0.07	0.07	0.08	0.08
	SVM_QuadKer	47.35	46.23	47.27	47.23	0.30	0.30	0.31	0.31
	QSVM2	9.97	45.06	45.70	47.93	0.28	0.28	0.28	0.29
	QSSVM	8.87	44.52	45.41	47.80	0.34	0.34	0.34	0.35
	SQSSVM	5.19	8.85	11.89	12.45	0.41	0.41	0.42	0.42

properties of the proposed model, but also conducted extensive computational experiments to validate its superior performance. Our major findings are summarized as below.

- Unlike soft SVM models with kernels, the proposed SQSSVM model does not need to use any kernel function or to tune the parameters in the kernel function for binary classification, to save much effort and time.
- For the tested artificial data, the SQSSVM model yields much more accurate classification than Dagher's modified model, the QSSVM model and soft SVM models with a Quadratic or Gaussian kernel. As the number of outliers in the training data set increases, the superiority in the classification accuracy of the SQSSVM model becomes even more evident. Also the SQSSVM model is shown to be much more robust than the QSSVM and Dagher's modified QSVM model.
- For all numerical experiments, the QSSVM model outperforms Dagher's modified QSVM model in terms of classification accuracy.
- The SQSSVM model successfully handles five real-world benchmark classifying data sets much more accurately than other models.

We are interested in studying the statistical properties of the classifier produced by the SQSSVM model such as its relationships with Vapnik–Chervonenkis bounds and reproducing kernel Hilbert space theories. We are also interested in applying the SQSSVM model in breast cancer diagnosis (Martin-Barragan *et al.*, 2007) and planar segmentation for urban terrain data (Luo *et al.*, 2013).

Appendix: Proof of Theorems 1–7

The proof of Theorem 1 can be derived as follows.

Proof. Take arbitrary $(\tilde{\mathbf{z}}, \tilde{c}) \in \mathbb{R}^{\frac{m^2+3m+2}{2}}$ and let $\tilde{\xi}_i = \max\{0, 1-y^i((\mathbf{s}^i)^T\tilde{\mathbf{z}}+\tilde{c})\}, i = 1, \ldots, n$. It is easy to see that $(\tilde{\mathbf{z}}, \tilde{c}, \tilde{\boldsymbol{\xi}})$ is feasible to the model (SQSSVM$'$). Note that the objective function is continuous and the feasible domain is a closed convex set defined by linear inequalities. Moreover, for any $\mathbf{z} \in \mathbb{R}^{\frac{m^2+3m}{2}}$ and $\xi_i \geq 0, i = 1, \ldots, n$, $\mathbf{z}^T G\mathbf{z} + \hat{\eta}\sum_{i=1}^n \xi_i = \sum_{i=1}^n (\|H^i\mathbf{z}\|_2^2 + \hat{\eta}\xi_i) \geq 0$, which indicates that the objective value is bounded below by 0 over the feasible domain. Hence there exists an optimal solution with a finite objective value. □

The proof of Theorem 2 can be derived as follows.

Proof. When the training data set is quadratically separable, it is not difficult to see that there exists a feasible solution $(\hat{\mathbf{z}}, \hat{c}, \mathbf{0})$ to model (SQSSVM$'$) with a given $\hat{\eta} > 0$. We first prove that $\sum_{i=1}^n \xi_i^{\hat{\eta}} \to 0$ as $\hat{\eta} \to \infty$ by contradiction. Suppose that there exists a given $\delta > 0$ such that for any $\hat{\eta} \geq \hat{\eta}^* := \frac{\hat{\mathbf{z}}^T G\hat{\mathbf{z}}+1}{\delta} > 0$, we have $|\sum_{i=1}^n \xi_i^{\hat{\eta}} - 0| = \sum_{i=1}^n \xi_i^{\hat{\eta}} \geq \delta$. Then, for the optimal solution $(\mathbf{z}^{\hat{\eta}}, c^{\hat{\eta}}, \boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}})$ of model (SQSSVM$'$) with any given $\hat{\eta} \geq \hat{\eta}^*$, we have ${\mathbf{z}^{\hat{\eta}}}^T G\mathbf{z}^{\hat{\eta}} + \hat{\eta}\sum_{i=1}^n \xi_i^{\hat{\eta}} \geq 0 + \hat{\eta}^*\delta =$

$0 + \hat{\mathbf{z}}^T G\hat{\mathbf{z}} + 1 > \hat{\mathbf{z}}^T G\hat{\mathbf{z}} + 0$ since G is positive semidefinite. Therefore, for any given $\hat{\eta} \geq \hat{\eta}^*$, $(\mathbf{z}^{\hat{\eta}}, c^{\hat{\eta}}, \boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}})$ can not be an optimal solution because $(\hat{\mathbf{z}}, \hat{c}, \mathbf{0})$ is feasible to the model (SQSSVM$'$). This contradiction leads to that $\sum_{i=1}^{n} \xi_i^{\hat{\eta}} \to 0$ as $\hat{\eta} \to \infty$. Consequently, $\boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}} \to \boldsymbol{\xi}^{*} = \mathbf{0}$ as $\hat{\eta} \to \infty$.

Next, we prove that $(\mathbf{z}^*, c^*)$ is an optimal solution to model (QSSVM$'$). Since $(\mathbf{z}^{\hat{\eta}}, c^{\hat{\eta}}, \boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}})$ is feasible to the model (SQSSVM$'$) for all $\hat{\eta} > 0$ and the linear constraints are in a closed form, we have $\{(\mathbf{z}^{\hat{\eta}}, c^{\hat{\eta}}, \boldsymbol{\xi}^{\hat{\boldsymbol{\eta}}})\}$ converges to $(\mathbf{z}^*, c^*, \mathbf{0})$ as $\hat{\eta} \to \infty$, and $y^i((\mathbf{s}^i)^T\mathbf{z}^* + c^*) \geq 1, i = 1, \ldots, n$. Hence $(\mathbf{z}^*, c^*)$ is feasible to model (QSSVM$'$). Moreover, let $(\bar{\mathbf{z}}, \bar{c})$ be an optimal solution to model (QSSVM$'$). Then $(\bar{\mathbf{z}}, \bar{c}, \mathbf{0})$ is feasible to model (SQSSVM$'$). Consequently, we have ${\mathbf{z}^{\hat{\eta}}}^T G\mathbf{z}^{\hat{\eta}} + \hat{\eta}\sum_{i=1}^{n} \xi_i^{\hat{\eta}} \leq \bar{\mathbf{z}}^T G\bar{\mathbf{z}} + 0$. Let $\hat{\eta} \to \infty$ and assume that $0 * \infty = 0$ without loss of generality, then we have ${\mathbf{z}^*}^T G\mathbf{z}^* \leq \bar{\mathbf{z}}^T G\bar{\mathbf{z}}$. Therefore, $(\mathbf{z}^*, c^*)$ is an optimal solution to model (QSSVM$'$). □

The proof of Theorem 3 can be derived as follows.

Proof. Assume that $(\hat{\mathbf{z}}, \hat{c}, \hat{\boldsymbol{\xi}}) \in \mathcal{F}^*$, $(\bar{\mathbf{z}}, \bar{c}, \bar{\boldsymbol{\xi}}) \in \mathcal{F}^*$ and $\hat{\mathbf{z}} \neq \bar{\mathbf{z}}$. For any $0 < \delta < 1$, $(\tilde{\mathbf{z}}, \tilde{c}, \tilde{\boldsymbol{\xi}}) := \delta(\hat{\mathbf{z}}, \hat{c}, \hat{\boldsymbol{\xi}}) + (1-\delta)(\bar{\mathbf{z}}, \bar{c}, \bar{\boldsymbol{\xi}})$ is feasible to model (SQSSVM$'$) due to the convexity of the feasible domain. Therefore,

$$\tilde{\mathbf{z}}^T G\tilde{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n} \tilde{\xi}_i \geq \hat{\mathbf{z}}^T G\hat{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n} \hat{\xi}_i,$$

$$\tilde{\mathbf{z}}^T G\tilde{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n} \tilde{\xi}_i \geq \bar{\mathbf{z}}^T G\bar{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n} \bar{\xi}_i.$$

Multiplying the first inequality by δ and the second by $(1-\delta)$, we have $\tilde{\mathbf{z}}^T G\tilde{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n} \tilde{\xi}_i \geq \delta\hat{\mathbf{z}}^T G\hat{\mathbf{z}} + (1-\delta)\bar{\mathbf{z}}^T G\bar{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n}(\delta\hat{\xi}_i + (1-\delta)\bar{\xi}_i)$. Equivalently, $[\delta\hat{\mathbf{z}} + (1-\delta)\bar{\mathbf{z}}]^T G[\delta\hat{\mathbf{z}} + (1-\delta)\bar{\mathbf{z}}] \geq \delta\hat{\mathbf{z}}^T G\hat{\mathbf{z}} + (1-\delta)\bar{\mathbf{z}}^T G\bar{\mathbf{z}}$, and $\delta(1-\delta)(\hat{\mathbf{z}} - \bar{\mathbf{z}})^T G(\hat{\mathbf{z}} - \bar{\mathbf{z}}) \leq 0$. When G is positive definite, we have $\hat{\mathbf{z}} - \bar{\mathbf{z}} = 0$, which contradicts to the assumption that $\hat{\mathbf{z}} \neq \bar{\mathbf{z}}$. □

The proof of Theorem 4 can be derived as follows.

Proof. Let $(\hat{\mathbf{z}}, \hat{c}, \hat{\boldsymbol{\xi}}) \in \mathcal{F}^*$. When G is positive definite, by Theorem 3, $\hat{\mathbf{z}}$ is uniquely determined and, for each $(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathcal{F}^*$, we have $\mathbf{z} = \hat{\mathbf{z}}$ and $\mathbf{z}^T G\mathbf{z} + \hat{\eta}\sum_{i=1}^{n} \xi_i = \hat{\mathbf{z}}^T G\hat{\mathbf{z}} + \hat{\eta}\sum_{i=1}^{n} \hat{\xi}_i$. Consequently, $\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \hat{\xi}_i := \bar{\delta}$, which is uniquely determined. Since $\xi_i \geq 0$ for any i, we have $\xi_i \leq \sum_{i=1}^{n} \xi_i = \bar{\delta}$. Therefore, $c \leq \xi_j - 1 - (\mathbf{s}^j)^T\hat{\mathbf{z}} \leq \bar{\delta} - 1 - (\mathbf{s}^j)^T\hat{\mathbf{z}}$ for $j \in \{j : y^j = -1\}$, and $c \geq 1 - \xi_j - (\mathbf{s}^j)^T\hat{\mathbf{z}} \geq 1 - \bar{\delta} - (\mathbf{s}^j)^T\hat{\mathbf{z}}$ for $j \in \{j : y^j = +1\}$. Let $\overline{c} = \min_{\{j:y^j=-1\}}\{\bar{\delta} - 1 - (\mathbf{s}^j)^T\hat{\mathbf{z}}\}$, $\underline{c} = \max_{\{j:y^j=+1\}}\{1 - \bar{\delta} - (\mathbf{s}^j)^T\hat{\mathbf{z}}\}$, then we have $\underline{c} \leq c \leq \overline{c}$. □

The proof of Theorem 5 can be derived as follows.

Proof. When the training data set is quadratically separable, by a similar proof of Theorem 2, for the model (SQSSVM$'$) with any given sufficiently large $\hat{\eta} > 0$ and

$(\breve{\mathbf{z}}, \breve{c}, \breve{\boldsymbol{\xi}}) \in \mathcal{F}^*$, we know $\breve{\boldsymbol{\xi}} = \mathbf{0}$. Hence $(\breve{\mathbf{z}}, \breve{c}, \mathbf{0})$ is feasible to the model (SQSSVM$'$), which indicates that $y^i((\mathbf{s}^i)^T\breve{\mathbf{z}} + \breve{c}) \geq 1, \forall i$. We first prove that there exists a $j \in \{j : y^j = +1\}$ such that $y^j((\mathbf{s}^j)^T\breve{\mathbf{z}} + \breve{c}) = 1$ as follows by contradiction.

Assume this conclusion is wrong, i.e., then we have

$$(\mathbf{s}^j)^T\breve{\mathbf{z}} + \breve{c} > 1, \quad \text{for } j \in \{j : y^j = +1\}, \tag{B1}$$

$$(\mathbf{s}^j)^T\breve{\mathbf{z}} + \breve{c} \leq -1, \quad \text{for } j \in \{j : y^j = -1\}. \tag{B2}$$

Let $\tilde{\mathbf{z}} = \delta\breve{\mathbf{z}}$ and $\tilde{c} = \delta(\breve{c} + 1) - 1$, for some $\delta \in (0, 1)$. Then expression (B2) is equivalent to

$$(\mathbf{s}^j)^T\tilde{\mathbf{z}} + \tilde{c} \leq -1, \quad \text{for } j \in \{j : y^j = -1\}. \tag{B3}$$

Moreover, for $j \in \{j : y^j = +1\}$, from expression (B1), we have

$$\lim_{\delta \to 1^-} [(\mathbf{s}^j)^T\tilde{\mathbf{z}} + \tilde{c}] = \lim_{\delta \to 1^-} [\delta(\mathbf{s}^j)^T\breve{\mathbf{z}} + \delta(\breve{c} + 1) - 1] = (\mathbf{s}^j)^T\breve{\mathbf{z}} + \breve{c} > 1.$$

Hence, there exists a $\delta \in (0, 1)$ such that

$$(\mathbf{s}^j)^T\tilde{\mathbf{z}} + \tilde{c} > 1, \quad \text{for } j \in \{j : y^j = +1\}. \tag{B4}$$

The expressions (B3) and (B4) indicate that $(\tilde{\mathbf{z}}, \tilde{c}, \mathbf{0})$ is feasible to model (SQSSVM$'$) and the corresponding objective value is $\tilde{\mathbf{z}}^T G\tilde{\mathbf{z}} + 0 = \delta^2\breve{\mathbf{z}}^T G\breve{\mathbf{z}} < \breve{\mathbf{z}}^T G\breve{\mathbf{z}} + 0$, which indicates that $(\breve{\mathbf{z}}, \breve{c}, \mathbf{0})$ is not an optimal solution. This contradiction infers that there exists a $j \in \{j : y^j = +1\}$ such that $y^j((\mathbf{s}^j)^T\breve{\mathbf{z}} + \breve{c}) = 1$.

Suppose that the model (SQSSVM$'$) has another optimal solution $(\hat{\mathbf{z}}, \hat{c}, \hat{\boldsymbol{\xi}})$. As before, we have $\hat{\boldsymbol{\xi}} = \mathbf{0}$. When G is positive definite, we know $\breve{\mathbf{z}} = \hat{\mathbf{z}}$ from Theorem 3. Rewrite the two optimal solutions as $(\breve{\mathbf{z}}, \breve{c}, \mathbf{0})$ and $(\breve{\mathbf{z}}, \hat{c}, \mathbf{0})$, respectively. From the above arguments, we know there exist j and $\bar{j} \in \{j : y^j = +1\}$ such that

$$(\mathbf{s}^{\bar{j}})^T\breve{\mathbf{z}} + \breve{c} = 1, \quad (\mathbf{s}^j)^T\breve{\mathbf{z}} + \breve{c} \geq 1; \quad (\mathbf{s}^j)^T\breve{\mathbf{z}} + \hat{c} = 1, \quad (\mathbf{s}^{\bar{j}})^T\breve{\mathbf{z}} + \hat{c} \geq 1.$$

Therefore, we have $\breve{c} \geq \hat{c}$ and $\breve{c} \leq \hat{c}$ using the above expressions. In other words, we have $\breve{c} = \hat{c}$. □

The proof of Lemma 6 can be derived as follows.

Proof. Let $(\tilde{\mathbf{z}}, \tilde{c}, \tilde{\boldsymbol{\xi}}) \in \mathcal{F}^*$. If $\|\tilde{\mathbf{z}}\| \neq 0$, for $(\mathbf{z}^\epsilon, c^\epsilon, \boldsymbol{\xi}^\epsilon)$ and any $\delta > 0$, there exists $\epsilon_0 = \frac{\delta}{(\tilde{\mathbf{z}}^T)(\tilde{\mathbf{z}})}$ such that when $0 < \epsilon < \epsilon_0$, $v \leq \mathbf{z}^{\epsilon T} G\mathbf{z}^\epsilon + \hat{\eta}\sum_{i=1}^n \xi_i^\epsilon \leq v_\epsilon \leq \tilde{\mathbf{z}}^T(G + \epsilon I)\tilde{\mathbf{z}} + \hat{\eta}\sum_{i=1}^n \tilde{\xi}_i = v + \epsilon(\tilde{\mathbf{z}})^T(\tilde{\mathbf{z}}) < v + \delta$. That is, $|v_\epsilon - v| < \delta$. If $\|\tilde{\mathbf{z}}\| = 0$, by the expression $v \leq \mathbf{z}^{\epsilon T} G\mathbf{z}^\epsilon + \hat{\eta}\sum_{i=1}^n \xi_i^\epsilon \leq v_\epsilon \leq \tilde{\mathbf{z}}^T(G + \epsilon I)\tilde{\mathbf{z}} + \hat{\eta}\sum_{i=1}^n \tilde{\xi}_i = v + \epsilon(\tilde{\mathbf{z}})^T(\tilde{\mathbf{z}}) = v$. we have that $v = v_\epsilon$. Therefore, $v_\epsilon \to v$ as $\epsilon \to 0$. □

The proof of Theorem 7 can be derived as follows.

Proof. When $\{(\mathbf{z}^\epsilon, c^\epsilon, \boldsymbol{\xi}^\epsilon)\} \to (\mathbf{z}^0, c^0, \boldsymbol{\xi}^0)$ as $\epsilon \to 0$, obviously $(\mathbf{z}^0, c^0, \boldsymbol{\xi}^0)$ is feasible to model (SQSSVM$'$). By Lemma 6, we have $v_\epsilon \to v$ as $\epsilon \to 0$. Hence we know

$(\mathbf{z}^0, c^0, \boldsymbol{\xi}^0) \in \mathcal{F}^*$. For any $\epsilon > 0$ and any $(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathcal{F}^*$, note that $(\mathbf{z}, c, \boldsymbol{\xi})$ is feasible to problem (SQSSVM′-ϵ). Therefore,

$$(\mathbf{z}^\epsilon)^T(G + \epsilon I)\mathbf{z}^\epsilon + \hat{\eta}\sum_{i=1}^n \xi_i^\epsilon = (\mathbf{z}^\epsilon)^T G\mathbf{z}^\epsilon + \epsilon(\mathbf{z}^\epsilon)^T\mathbf{z}^\epsilon + \hat{\eta}\sum_{i=1}^n \xi_i^\epsilon$$

$$\leq (\mathbf{z})^T(G + \epsilon I)\mathbf{z} + \hat{\eta}\sum_{i=1}^n \xi_i = (\mathbf{z})^T G\mathbf{z} + \epsilon(\mathbf{z})^T\mathbf{z} + \hat{\eta}\sum_{i=1}^n \xi_i.$$

Since $(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathcal{F}^*$, we have $(\mathbf{z})^T G\mathbf{z} + \hat{\eta}\sum_{i=1}^n \xi_i \leq (\mathbf{z}^\epsilon)^T G\mathbf{z}^\epsilon + \hat{\eta}\sum_{i=1}^n \xi_i^\epsilon$. Consequently, $(\mathbf{z}^\epsilon)^T\mathbf{z}^\epsilon \leq (\mathbf{z})^T\mathbf{z}$ for any $\epsilon > 0$. Hence, as $\epsilon \to 0$, $(\mathbf{z}^0)^T\mathbf{z}^0 \leq (\mathbf{z})^T\mathbf{z}$ for any $(\mathbf{z}, c, \boldsymbol{\xi}) \in \mathcal{F}^*$. □

Acknowledgment

Fang's research has been supported by United States Army Research Office Grant No. W911NF-15-1-0223; Deng's research has been supported by National Natural Science Foundation of China Grant No. 11501543 and UCAS Scientific Research Foundation for Young Faculty Grant No. Y551037Y00.

References

Bache, K and M Lichman (2013). UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Boser, B, I Guyon and VN Vapnik (1992). A training algorithm for optimal margin classifiers, in: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Brooks, JP (2011). Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59, 467–479.

Cao, Y, GY Wan and FQ Wang (2011). Predicting financial distress of Chinese listed companies using rough set theory and support vector machine. *Asia-Pacific Journal of Operational Research*, 28, 95–109.

Chang, CC and CJ Lin (2013). LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chen, Z, Z Fan and M Sun (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, 223, 461–472.

Cortes, C and VN Vapnik (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Cristianini, N and J Shawe-Taylor (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1st edn. Cambridge: Cambridge University Press.

Dagher, I (2008). Quadratic kernel-free non-linear support vector machine. *Journal of Global Optimization*, 41, 15–30.

Deng, N, Y Tian and C Zhang (2013). *Support Vector Machines*: *Optimization Based Theory*, *Algorithms*, *and Extensions*. Boca Raton, FL: CRC Press.

Drucker, H, D Wu and VN Vapnik (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10, 1048–1054.

Fang, SC and S Puthenpura (1993). *Linear Optimization and Extensions*: *Theory and Algorithms.* Englewood Cliffs, NJ: Prentice Hall.

Kim, KI, K Jung, SH Park and HJ Kim (2012). Support vector machines for texture classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 1542–1550.

Lin, CJ (2001). Formulations of support vector machines: a note from an optimization point of view. *Neural Computation*, 13, 307–317.

Liu, Y and M Yuan (2011). Reinforced multicategory support vector machines. *Journal of Computational and Graphical Statistics*, 20, 901–919.

Luo, J, Z Deng, D Bulatov, JE Lavery and SC Fang (2013). Comparison of an ℓ_1-regression-based and a RANSAC-based planar segmentation procedure for urban terrain data with many outliers. in *Proceedings of SPIE 8892*, *Image and Signal Processing for Remote Sensing XIX*, 2013, doi:10.1117/12.2028627.

Martin-Barragan, B, R Lillo and J Romo (2014). Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232, 146–155.

Schölkopf, B and AJ Smola (2002). *Learning with Kernels.* Cambridge, MA: MIT Press.

Vapnik, VN (1998). *Statistical Learning Theory.* New York, NY: Wiley-Interscience Publication.

Vapnik, VN (2000). *The Nature of Statistical Learning Theory.* New York, NY: Springer-Verlag.

Biography

Jian Luo is an Assistant Professor in the School of Management Science and Engineering at Dongbei University of Finance and Economics. He received his Bachelor degree and Master degree from Wuhan University in China in 2007 and 2009, and his PhD degree from North Carolina State University in USA in 2014. He was a research assistant in the US Army Research Office in 2012–2014. He is the author or coauthor of papers published in Journal of the Operational Research Society, Journal of Industrial and Management Optimization, Soft Computing, among others. His research interests include machine learning techniques with applications, data mining, pattern analysis, fuzzy and nonlinear optimization, and geometric design.

Shu-Cherng Fang is the Walter Clark Chair and Alumni Distinguished Graduate Professor in the Department of Industrial and Systems Engineering and Graduate Program in Operations Research at NC State University. His key research interest is on optimization theory and algorithms with real life applications such as intelligent human–machine decision support systems, terrain data representation, logistics and supply chain management, telecommunications and bio-informatics. He received his Bachelor degree from the National Tsing Hua University in Taiwan and PhD degree from Northwestern University in USA. Before he joined NC State University, Dr. Fang had been working at AT&T Engineering Research Center, AT&T Bell Laboratories, and AT&T Corporate Headquarters. He is also associated with many universities in Australia, China, Hong Kong and Taiwan. He has published a few journal articles and books. He is the Founding Editor-in-Chief

of Fuzzy Optimization and Decision Making. His new book, coauthored by Prof. Wenxun Xing of Tsinghua University, on "Linear Conic Optimization" has just been published by the Science Press in August, 2013.

Zhibin Deng is an Assistant Professor in the School of Economics & Management at the University of Chinese Academy of Sciences. He received his Bachelor degree and Master degree from Tsinghua University in China in 2007 and 2009 respectively, and PhD degree from North Carolina State University in USA in 2013. He was a research assistant in the US Army Research Office in 2011–2013. He received Edwards P. Fitts Fellowship for his excellent academic record when admitted to North Carolina State University. He is the author or coauthor of papers published in European Journal of Operational Research, Journal of Global Optimization, and Journal of Industrial and Management Optimization. His research interests include quadratic optimization, linear conic optimization and mixed-integer optimization.

Xiaoling Guo received her PhD degree in 2014, from Department of Mathematical Sciences, Tingshua University, Beijing, China. She did a two-year postdoctoral research in School of Engineering at University of Chinese Academy of Sciences. She is currently a lecturer in Department of Mathematics in China University of Mining and Technology, Beijing. She is the author or coauthor of papers published in Journal of Systems Sciences and Systems Engineering, Journal of Industrial and Management Optimization, Optimization, Journal of the Operations Research Society of China, among others. Her research interests include quadratic programming, linear conic programming, algorithm designing, project management and emergency management.