Robust Twin Bounded Support Vector Classifier with Manifold Regularization

Junhong Zhang, Zhihui Lai, Heng Kong, Linlin Shen

***Abstract*—Support vector machine (SVM), as a supervised learning method, has different kinds of varieties with significant performance. In recent years, more researches focused on nonparallel SVM, where twin support vector machine (TWSVM) is the typical one. In order to reduce the influence of outliers, more robust distance measurements are considered in these methods, but the discriminability of the models is neglected. In this paper, we propose robust manifold twin bounded support vector machine (RMTBSVM), which considers both robustness and discriminability. Specifically, a novel norm, i.e., capped L_1-norm, is used as the distance metric for robustness, and a robust manifold regularization is added to further improve the robustness and classification performance. In addition, we also use kernel method to extend the proposed RMTBSVM for nonlinear classification. We introduce the optimization problems of the proposed model. Subsequently, effective algorithms for both linear and nonlinear cases are proposed and proved to be convergent. Moreover, the experiments are conducted to verify the effectiveness of our model. Compared with other methods under the SVM framework, the proposed RMTBSVM shows better classification accuracy and robustness.**

***Index Terms*—twin support vector machine, robust capped L_1-norm, manifold regularization, kernel method.**

I. Introduction

SUPPORT vector machine (SVM) [1] is one of the most commonly used tools for classification tasks in machine learning. It is generally applied in image classification [2], action recognition [3], energy resources prediction [4], and so on. The essential points of support vector classification (SVC) are the maximum margin principle, dual theory, and kernel tricks [5], [6]. The traditional SVC expects to solve an optimization problem that minimizes structural risk based on hinge loss and L_2-norm regularization. Based on dual theory, the principle of SVC finally leads to a very large quadratic programming problem (QPP). To solve the QPP more efficiently, many algorithms were proposed, such as sequential minimal optimization (SMO) [7] and successive over-relaxation (SOR) [8]. However, these algorithms might be still time-consuming in large-scale classification tasks. Because of the large QPP in conventional SVM, many effective improved methods of SVM were proposed in early studies. Suykens et al proposed least square support vector machine (LSSVM), which uses equality constraints instead of inequalities in SVM [9], and the computation burden is decreased. Mangasarian et al proposed Lagrangian support vector machine based on implicit Lagrangian formulation, which can efficiently work in large datasets [10].

Nonparallel support vector machine (NPSVM) is a new genre of SVM [6]. Compared with classical SVM, NPSVM uses nonparallel hyperplanes to classify data. Twin support vector machine (TWSVM) [11] is a special case of NPSVM. The main idea of TWSVM is to find two hyperplanes for two classes, and each hyperplane is closer to the corresponding class and far away from the other class. Therefore, TWSVM is formulated with two optimization problems for two hyperplanes respectively, which leads to a pair of QPPs. Compared with the large QPP in SVM, these two QPPs in TWSVM are both small and easy to solve. Thus TWSVM can be trained more efficiently than SVM. Additionally, TWSVM can effectively learn more complicated distribution than SVM, for instance, the "cross planes" dataset [12]. Therefore, TWSVM is widely applied in the classification problems. There are many extensive methods based on TWSVM. Shao et al presented twin bounded support vector machine (TBSVM) implementing thestructural risk minimization principle with L_2-norm regularization [12]. Kumar et al proposed the least square version of TWSVM, which can be solved with linear equation efficiently instead of QPPs [13].

However, most of methods mentioned above use squared L_2-norm as metric and are sensitive to outliers, since the impact of errors will be amplified by the square operator. How to improve the robustness of discriminative methods (e.g., SVM, regression) is still an important and open topic in machine learning [14]. To improve the robustness, many discriminative learning methods based on more robust metric (e.g. L_1-norm, $L_{2,1}$-norm) were proposed, such as L_1-fisher discriminant analysis (L_1-LDA), robust feature selection (RFS), and robust discriminant regression (RDR) [15]–[17]. The robust metric is also widely studied in recent researches in the TWSVM framework. Xu et al presented pin-TWSVM which uses robust pinball loss in the objective functions [18]. Yan et al presented L_1-TWSVM and its least square version L_1-LSTBSVM, which are both based on L_1-norm distance metric [19], [20]. Recently, the robustness of TWSVM model

This work was supported in part by the Natural Science Foundation of China under Grant 61976145, Grant 61802267 and Grant 61732011, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grants JCYJ20180305124834854 and JCYJ20190813100801664.

J. Zhang, Z. Lai and L. Shen are Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China. (email: apple_zjh@163.com, lai_zhi_hui@163.com, llshen@szu.edu.cn)

Heng Kong is with the Department of Breast and Thyroid Surgery, BaoAn Central Hospital of Shenzhen, BaoAn district, Shenzhen, Guangdong Province, China. (email: generaldoc@126.com)

were further improved by introducing capped L_1-norm metric, such as CTWSVM, FRTBSVM, R-CTWSVM+ [21]–[23], and robust correntropy-based metric, such as RCTSVM, ARTSVM [24], [25]. However, these methods have not further discussed the case of nonlinear classification so far.

Manifold learning is also an important topic in machine learning in the past 20 years. Preserving the manifold structure can improve the discriminative property of data [26]. Therefore, many improved methods of SVM are developed based on the technique of manifold learning framework. Local coding method is applied in SVM to derives a locally linear classifer in the early studies [27]. Recently, manifold regularization has been a popular technique widely applied in SVM models to preserve manifold structure, which can effectively implement semi-supervised learning [28], [29]. In this framework, Laplacian TWSVM (Lap-TWSVM) was developed for semi-supervised learning under the manifold regularization framework [30], [31].

In this paper, we take advantages of robust metric and manifold regularization techniques to propose a novel model with significant robustness and discriminative performance. We improve Lap-TWSVM and CTWSVM [21], [31] and propose a comprehensive method, namely Robust Manifold Twin Bounded Support Vector Machine (RMTBSVM). It greatly enhances the robustness and classification performance of TBSVM. The main contributions of this paper can be concluded as below:

- We propose RMTBSVM, a novel nonparallel SVM model with capped L_1-norm loss and manifold regularization in L_1-norm. The different robust metrics are systematically integrated into the model to take full advantage of their function for SVM classifier's design. Therefore, the robustness of SVM has been greatly improved. As such, the proposed RMTBSVM is not only robust to outliers but has stronger discriminative ability.
- We present the optimization problems of the proposed model and derive effective algorithms to solve the problems. The convergence of the algorithm is proved.
- We further extend our model to nonlinear conditions. Kernel method is used to implement nonlinear classification. The theoretical analysis is also presented in this paper.
- Experiments are carried out to verify the effectiveness of RMTBSVM. The results suggest that RMTBSVM is more robust and performs better in classification tasks, which is verified by statistical significance in this paper.

The rest of this paper is organized as follows: In section II, we discuss the related works and give the motivation of our method. In section III, we propose our model, and design an effective algorithm to solve the optimization problem. Further analysis of our method is presented in section IV. In section V, we evaluate the performance of the RMTBSVM by performing a set of experiments. Finally, a conclusion is made in section VI.

II. Preliminary

In this section, we review some related works, including TWSVM, TBSVM, manifold regularization and capped L_1-norm loss.

A. TWSVM and TBSVM

Let us consider a binary classification problem. The training sample matrix is $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where column vector $\mathbf{x}_i \in \mathbb{R}^d$ is the sample point. The label of i-th sample is $y_i \in \{-1, 1\}$. We use matrix $\mathbf{A} \in \mathbb{R}^{d\times n_A}$ to denote the samples of positive class (i.e. $y_i = 1$), and use $\mathbf{B} \in \mathbb{R}^{d\times n_B}$ to denote the samples of negative class (i.e. $y_i = -1$), where $n_A + n_B = n$. We define

$$\mathbf{A} = [\mathbf{x}_1^+, \mathbf{x}_2^+, \cdots, \mathbf{x}_{n_A}^+],\ \mathbf{B} = [\mathbf{x}_1^-, \mathbf{x}_2^-, \cdots, \mathbf{x}_{n_B}^-],$$

where $\mathbf{x}_i^+$ (or $\mathbf{x}_i^-$) represents the i-th sample of positive (or negative) class. For convenience, we suppose in $\mathbf{X}$ the samples are grouped by the labels, that is, $\mathbf{X} = [\mathbf{A}, \mathbf{B}]$.

Twin support vector machine (TWSVM) [11] defines two nonparallel hyperplanes for classification task:

$$f_1(\mathbf{x}) = \mathbf{w}_1^T\mathbf{x} + b_1 = 0, \quad f_2(\mathbf{x}) = \mathbf{w}_2^T\mathbf{x} + b_2 = 0,$$

where $\mathbf{w}_1 \in \mathbb{R}^d, \mathbf{w}_2 \in \mathbb{R}^d, b_1 \in \mathbb{R}$ and $b_2 \in \mathbb{R}$. TWSVM assumes that positive samples are closed to the hyperplane $f_1(\mathbf{x}) = 0$ but far away from $f_2(\mathbf{x}) = 0$, and vice versa. Therefore, the hyperplanes can be obtained by solving following pair of optimization problems:

$$\begin{aligned} &\min_{\mathbf{w}_1,b_1,\xi} \frac{1}{2}\|\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1\|_2^2 + c_1\mathbf{e}_2^T\xi, \\ &\text{s.t. } -(\mathbf{B}^T\mathbf{w}_2 + b_1\mathbf{e}_2) + \xi \geq \mathbf{e}_2, \xi \geq 0, \end{aligned} \tag{1}$$

$$\begin{aligned} &\min_{\mathbf{w}_2,b_2,\eta} \frac{1}{2}\|\mathbf{B}^T\mathbf{w}_2 + b_2\mathbf{e}_2\|_2^2 + c_2^T\mathbf{e}_1\eta, \\ &\text{s.t. } (\mathbf{A}^T\mathbf{w}_1 + b_2\mathbf{e}_1) - \eta \geq \mathbf{e}_1, \eta \geq 0, \end{aligned} \tag{2}$$

where $\mathbf{e}_1 \in \mathbb{R}^{n_A}, \mathbf{e}_2 \in \mathbb{R}^{n_B}$ are vectors of ones, and $\xi \in \mathbb{R}^{n_B}, \eta \in \mathbb{R}^{n_A}$ denote slack variables generated from hinge loss. Generally, to implement structural risk minimization principle and avoid over-fitting issues, L_2-regularization term is added to (1) and (2) to obtain an improved version of TWSVM, namely twin bounded support vector machine (TBSVM) [12]:

$$\begin{aligned} &\min_{\mathbf{w}_1,b_1,\xi} \frac{1}{2}\|\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1\|_2^2 + c_1\mathbf{e}_2^T\xi + \frac{r_1}{2}\left(\|\mathbf{w}_1\|_2^2 + b_1^2\right), \\ &\text{s.t. } -(\mathbf{B}^T\mathbf{w}_2 + b_1\mathbf{e}_2) + \xi \geq \mathbf{e}_2, \xi \geq 0, \end{aligned} \tag{3}$$

$$\begin{aligned} &\min_{\mathbf{w}_2,b_2,\eta} \frac{1}{2}\|\mathbf{B}^T\mathbf{w}_2 + b_2\mathbf{e}_2\|_2^2 + c_2\mathbf{e}_1^T\eta + \frac{r_2}{2}\left(\|\mathbf{w}_2\|_2^2 + b_2^2\right), \\ &\text{s.t. } (\mathbf{A}^T\mathbf{w}_1 + b_2\mathbf{e}_1) - \eta \geq \mathbf{e}_1, \eta \geq 0. \end{aligned} \tag{4}$$

Based on dual theory, we can obtain the dual problem of (3) and (4):

$$\max_{\mathbf{0}\leq\alpha\leq c_1\mathbf{e}_2} \mathbf{e}_2^T\alpha - \frac{1}{2}\alpha^T\tilde{\mathbf{B}}^T(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T + r_1\mathbf{I})^{-1}\tilde{\mathbf{B}}\alpha, \tag{5}$$

$$\max_{\mathbf{0}\leq\gamma\leq c_2\mathbf{e}_1} \mathbf{e}_1^T\gamma - \frac{1}{2}\gamma^T\tilde{\mathbf{A}}^T(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T + r_2\mathbf{I})^{-1}\tilde{\mathbf{A}}\gamma, \tag{6}$$

where $\tilde{\mathbf{A}} = [\mathbf{A}^T, \mathbf{e}_1]^T, \tilde{\mathbf{B}} = [\mathbf{B}^T, \mathbf{e}_2]^T$ and $\mathbf{0}, \mathbf{I}$ denote zero-vector and identity matrix both in appropriate dimension. [12] points out that (5) and (6) has same formulation of QPP:

$$\max_{\theta}\ \mathbf{e}^T\theta - \frac{1}{2}\theta^T\mathbf{Q}\theta, \quad \text{s.t. } \mathbf{0} \leq \theta \leq c\mathbf{e}, \tag{7}$$

which can be efficiently solved with successive over-relaxation

algorithm (SOR) [8]. If α, γ in (5) and (6) are obtained, we can easily compute $\mathbf{w}_1, b_1, \mathbf{w}_2, b_2$.

Remark 1. TWSVM (or TBSVM) uses kernel method to implement nonlinear classification. The corresponding optimization problems are similar with the linear case. For details, see [12].

Finally, for a new data $\mathbf{x}$, TWSVM and TBSVM uses the same decision function to determine its label:

$$h(\mathbf{x}) = \operatorname{sgn}\left(\frac{|\mathbf{w}_2^T\mathbf{x} + b_2|}{\|\mathbf{w}_2\|} - \frac{|\mathbf{w}_1^T\mathbf{x} + b_1|}{\|\mathbf{w}_1\|}\right). \tag{8}$$

The multi-class TWSVM (or TBSVM) classifier can be naturally derived with one-versus-rest strategy (OVR), which is OVR-TWSVM [32]. Consider a dataset with k classes. We denote $\mathbf{X}_i$ as the sample matrix of i-th class, and we define:

$$\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_k], \tag{9}$$

$$\mathbf{X}_{-i} = [\mathbf{X}_1, \cdots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \cdots, \mathbf{X}_k]. \tag{10}$$

OVR-TWSVM generates k hyperplanes for classification, and each hyperplane corresponds to one class. The hyperplane of i-th class is obtained by solving the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{w}_i, b_i, \xi_i} \frac{1}{2}\|\mathbf{X}_i^T\mathbf{w}_i + b_i\mathbf{e}_i\|_2^2 + c_i\mathbf{e}_{-i}^T\xi_i,\\ &\text{s.t. } -(\mathbf{X}_{-i}^T\mathbf{w}_i + b_i\mathbf{e}_{-i}) + \xi_i \geq \mathbf{e}_{-i}, \xi_i \geq 0.\end{aligned} \tag{11}$$

where $(\mathbf{w}_i, b_i)$ indicates the hyperplane of i-th class, c_i is the penalty parameter, ξ_i denotes the slack variables, and $\mathbf{e}_i, \mathbf{e}_{-i}$ are vectors of ones of proper dimensions. Problem (11) is directly extended from binary-class TWSVM (1), and it takes the i-th class as the positive class and the rest classes as the negative class. Therefore, we need to solve k QPPs to obtain an OVR-TWSVM classifier. The geometric meaning of OVR-TWSVM is that the i-th hyperplane should be close to the data of i-th class and far away from the others. Hence the corresponding decision function is

$$h(\mathbf{x}) = \min_{i=1,2,\cdots,k}\left(\frac{|\mathbf{w}_i^T\mathbf{x} + b_i|}{\|\mathbf{w}_i\|}\right). \tag{12}$$

OVR-TWSVM is intuitive and easy to implement. The OVR strategy can be also applied in the improved methods of TWSVM easily, including the method we propose in this paper.

B. Manifold regularization

Manifold regularization is a general technique commonly used in semi-supervised learning [33]. Suppose we have dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and let matrix $\mathbf{F} \in \mathbb{R}^{n\times n}$ denote the adjacency graph for training samples. The standard optimization framework of manifold regularization can be written as

$$\begin{aligned}\mathbf{f}^* = \arg\min_{\mathbf{f}\in\mathcal{H}} \sum_{i=1}^n \mathcal{L}(\mathbf{x}_i, y_i, \mathbf{f}) + \lambda_I \sum_{i,j=1}^n (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 W_{ij}\\ + \lambda_H\|\mathbf{f}\|_{\mathcal{H}}^2,\end{aligned} \tag{13}$$

where λ_I, λ_H are regularization parameters. The second part of (13) is the manifold regularizer, with which the model can exploit and preserve locally geometric structure of data. This term can be written as $\mathbf{f}^T\mathbf{L}\mathbf{f}$, where $\mathbf{L}$ is the Laplacian matrix of $\mathbf{W}$.

For example, [31] proposed a semi-supervised model based on TBSVM, namely Laplacian TWSVM (Lap-TWSVM). Its optimization problems can be formulated as

$$\begin{aligned}\min_{\mathbf{w}_1, b_1, \xi} \frac{1}{2}\|\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1\|_2^2 + c_1\mathbf{e}_2^T\xi + \frac{r_1}{2}\left(\|\mathbf{w}_1\|_2^2 + b_1^2\right)\\ + \sum_{i,j=1}^n (\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\mathbf{x}_j)^2 F_{ij},\\ \text{s.t. } -(\mathbf{B}^T\mathbf{w}_2 + b_1\mathbf{e}_2) + \xi \geq \mathbf{e}_2, \xi \geq 0,\end{aligned} \tag{14}$$

$$\begin{aligned}\min_{\mathbf{w}_2, b_2, \eta} \frac{1}{2}\|\mathbf{B}^T\mathbf{w}_2 + b_2\mathbf{e}_2\|_2^2 + c_2\mathbf{e}_1^T\eta + \frac{r_2}{2}\left(\|\mathbf{w}_2\|_2^2 + b_2^2\right)\\ + \sum_{i,j=1}^n (\mathbf{w}_2^T\mathbf{x}_i - \mathbf{w}_2^T\mathbf{x}_j)^2 F_{ij},\\ \text{s.t. } (\mathbf{A}^T\mathbf{w}_1 + b_2\mathbf{e}_1) - \eta \geq \mathbf{e}_1, \eta \geq 0.\end{aligned} \tag{15}$$

Compared with (3) and (4), these new optimization problems have a new manifold regularization term. The problems can be naturally converted into the quadratic programming framework (7) and solved efficiently [31].

C. Capped L_1-norm and CTWSVM

In general, we always use different norms of vector to measure loss or regularize the learning model. L_2-norm is commonly used in many models, which is defined as

$$\|\mathbf{x}\|_2 = (\mathbf{x}^T\mathbf{x})^{1/2}. \tag{16}$$

We usually square this term to simplify computation, i.e. $\|\mathbf{x}\|_2^2$ is used. However, L_2-norm is sensitive since the square term amplifies the impact of the outliers [15], [21]. In contrast, L_1-norm is more robust since it is defined as the sum of the absolute value of the elements in vector:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|. \tag{17}$$

Based on L_1-norm, we then introduce an operator derived from L_1-norm, called capped L_1-norm. It is defined with a positive parameter ϵ:

$$\|\mathbf{x}\|_{1,\epsilon} = \sum_{i=1}^n \min\left(|x_i|, \epsilon\right), \quad \epsilon > 0. \tag{18}$$

Remark 2. It should be noted that $\|\cdot\|_{1,\epsilon}$ operator is not absolutely homogeneous, i.e. $\|\rho\mathbf{x}\|_{1,\epsilon} \not\equiv |\rho|\|\mathbf{x}\|_{1,\epsilon}$. Therefore, it cannot be defined as a norm. Here we use this norm-like notation just for convenience.

The capped norm is considered more robust to outliers [34]. Here we briefly compare the robustness of the loss based on L_2, L_1 and capped L_1-norm. Fig. 1 shows the curve of different losses. We can see that for the large u_i, which might be seen as an outlier, capped L_1-norm limits the impact of the error to ϵ. Hence capped L_1-norm can significantly reduce the influence of noises or outliers. In recent researches, more robust losses are used to improve the robustness of

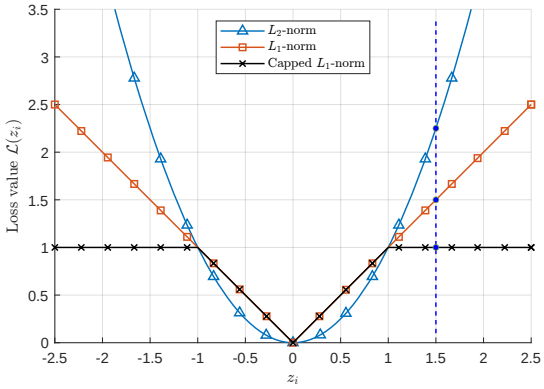

Fig. 1. Comparison of L_2, L_1, and capped L_1-loss ($\epsilon = 1$). The x-axis denotes the input of loss function. For outlier point, say $u_i = 1.5$ (the vertical line), we can see that capped L_1-loss $>$ L_1-loss $>$ L_2-loss. Therefore, L_2-loss is the most sensitive, the L_1-norm is second, and capped L_1-norm is the most robust to outliers.

TWSVM [19], [21], [23]. Capped twin support vector machine (CTWSVM) is a typical one which uses the following capped L_1-loss:

$$\mathcal{L}_1(\mathbf{x}_i, y_i, \mathbf{f}) = \begin{cases} \min\left(\left|\mathbf{w}_1^T\mathbf{x}_i + b\right|, \epsilon_1\right), & y_i > 0; \\ \min[(1 + b_1 + \mathbf{w}_1^T\mathbf{x}_i)_+, \epsilon_2], & y_i < 0; \end{cases},$$

$$\mathcal{L}_2(\mathbf{x}_i, y_i, \mathbf{f}) = \begin{cases} \min\left(\left|\mathbf{w}_2^T\mathbf{x}_i + b\right|, \epsilon_3\right), & y_i > 0; \\ \min[(1 - b_1 - \mathbf{w}_2^T\mathbf{x}_i)_+, \epsilon_4], & y_i < 0. \end{cases}$$

Therefore, the optimization problems of CTWSVM are

$$\begin{aligned} &\min_{\mathbf{w}_1, b_1, \xi} \|\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1\|_{1,\epsilon_1} + c_1\|\xi\|_{1,\epsilon_2}, \\ &\text{s.t. } -(\mathbf{B}^T\mathbf{w}_1 + b_1\mathbf{e}_2) + \xi \geq \mathbf{e}_2,\ \xi \geq \mathbf{0}, \qquad (19) \\ &\min_{\mathbf{w}_2, b_2, \eta} \|\mathbf{B}^T\mathbf{w}_1 + b_2\mathbf{e}_1\|_{1,\epsilon_3} + c_2\|\eta\|_{1,\epsilon_4}, \\ &\text{s.t. } (\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_2) - \eta \geq \mathbf{e}_1,\ \eta \geq \mathbf{0}. \qquad (20) \end{aligned}$$

III. Robust Manifold Twin Bounded Support Vector Machine

In this section, we first introduce the motivation of the proposed model, and present the optimization problems of RMTBSVM. Then we analyze the optimization problems and present the algorithms for problems based on the analysis.

A. The motivation of RMTBSVM

We mention Lap-TWSVM and CTWSVM above, which perform well in semi-supervised learning and robust learning, respectively. However, a drawback of Lap-TWSVM is the model is sensitive to outliers since it uses L_2-loss, and CTWSVM cannot exploit locally manifold structures. Therefore, it is worth trying to take advantages of both frameworks. That is, using a more robust loss with manifold regularization based on TBSVM not only enhances the robustness to noise, but enables the proposed model to exploit the manifold structure for robust classification.

In order to improve the classification performance, we consider introducing supervised information to manifold regularization. Thus, a special regularized term with graph preservation is designed in the proposed model via using L_1-norm as metric to improve the robustness, which is presented as

$$\sum_{i,j=1}^{n} \left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right| F_{ij}. \qquad (21)$$

In CTWSVM, the capped L_1-norm is applied for misclassification loss for robustness. Therefore, the loss is at most ϵ for these points [21]. However, it restricts the penalties for misclassification to a small value, which probably limits the discriminability of the model. Since $\|\xi\|_1 = \mathbf{e}_2^T\xi \geq \|\xi\|_{1,\epsilon}$, from the intuition, minimizing L_1-loss derives a larger between-class scatter after projection than capped L_1-loss, so according to Fisher discriminant criterion, it is supposed to improve the performance of classification. Hence in our model, L_1-loss is applied to balance robustness and discriminability. On the other hand, it also reduces the complexity of the model.

Furthermore, [19], [21], [22] does not consider the case of nonlinear classification. However, most of the datasets have complicated nonlinear distribution, and as a result, the performance of the linear model is degraded. Therefore, to adapt the model to nonlinear cases and improve the performance, we propose an effective approach for nonlinear classification based on kernel theory to further extend the proposed model.

B. Linear classification

We first consider the linear cases of RMTBSVM. The loss term of TBSVM is substituted by capped L_1-loss and L_1-hinge loss, and more robust L_1-norm manifold regularizer is added. Therefore, we present the optimization problems of RMTBSVM:

$$\begin{aligned} \min_{\mathbf{w}_1, b_1, \xi} &\|\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1\|_{1,\epsilon} + c_1\mathbf{e}_2^T\xi + \frac{r_1}{2}(\|\mathbf{w}_1\|_2^2 + b_1^2) \\ &+ \frac{\mu_1}{2}\sum_{i,j=1}^{n} |\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\mathbf{x}_j| F_{ij}, \\ &\text{s.t. } -(\mathbf{B}^T\mathbf{w}_1 + b_1\mathbf{e}_2) + \xi \geq \mathbf{e}_2,\ \xi \geq 0, \qquad (22) \end{aligned}$$

$$\begin{aligned} \min_{\mathbf{w}_2, b_2, \eta} &\|\mathbf{B}^T\mathbf{w}_2 + b_2\mathbf{e}_2\|_{1,\epsilon} + c_2\mathbf{e}_1^T\eta + \frac{r_2}{2}(\|\mathbf{w}_2\|_2^2 + b_2^2) \\ &+ \frac{\mu_2}{2}\sum_{i,j=1}^{n} |\mathbf{w}_2^T\mathbf{x}_i - \mathbf{w}_2^T\mathbf{x}_j| F_{ij}, \\ &\text{s.t. } (\mathbf{A}^T\mathbf{w}_2 + b_2\mathbf{e}_1) - \eta \geq \mathbf{e}_1,\ \eta \geq 0. \qquad (23) \end{aligned}$$

The optimization problems (22) and (23) are complicated. Here we simplify them in two steps. First, we use augment vectors to represent $\mathbf{w}_1, b_1$ and $\mathbf{w}_2, b_2$. Thus we define

$$\mathbf{v}_1 = \begin{bmatrix} \mathbf{w}_1 \\ b_1 \end{bmatrix},\ \mathbf{v}_2 = \begin{bmatrix} \mathbf{w}_2 \\ b_2 \end{bmatrix},$$
$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ \mathbf{e}_1^T \end{bmatrix},\ \tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{B} \\ \mathbf{e}_2^T \end{bmatrix},\ \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \mathbf{e}^T \end{bmatrix}.$$

where $\mathbf{e} \in \mathbb{R}^n$ is the vector of ones. Second, the optimization problems are concave and hard to solve due to the capped norm term. An elegant way is to substitute the objective function with a convex one, such that minimizing the new objective leads to the solution of the original problem. Thus inspired by

[35], we reformulate (22) and (23) in the proposed model as the following convex optimization problems:

$$\min_{\mathbf{v}_1,\xi} \ \frac{1}{2}\mathbf{v}_1^T\tilde{\mathbf{A}}\mathbf{D}_1\tilde{\mathbf{A}}^T\mathbf{v}_1 + c_1\mathbf{e}_2^T\xi + \frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{v}_1^T\tilde{\mathbf{x}}_i - \mathbf{v}_1^T\tilde{\mathbf{x}}_j|F_{ij} + \frac{r_1}{2}\|\mathbf{v}_1\|_2^2, \ \text{s.t.} \ -\tilde{\mathbf{B}}^T\mathbf{v}_1 + \xi \geq \mathbf{e}_2, \ \xi \geq \mathbf{0}, \tag{24}$$

$$\min_{\mathbf{v}_2,\eta} \ \frac{1}{2}\mathbf{v}_2^T\tilde{\mathbf{B}}\mathbf{D}_2\tilde{\mathbf{B}}^T\mathbf{v}_2 + c_2\mathbf{e}_1^T\eta + \frac{\mu_2}{2}\sum_{i,j=1}^{n}|\mathbf{v}_2^T\tilde{\mathbf{x}}_i - \mathbf{v}_2^T\tilde{\mathbf{x}}_j|F_{ij} + \frac{r_2}{2}\|\mathbf{v}_2\|_2^2, \ \text{s.t.} \ \tilde{\mathbf{A}}^T\mathbf{v}_2 - \eta \geq \mathbf{e}_1, \ \eta \geq \mathbf{0}. \tag{25}$$

where $\mathbf{D}_1, \mathbf{D}_2$ are diagonal matrices, and the diagonal entries are

$$(D_1)_{ii} = \frac{\mathcal{I}(\mathbf{v}_1^T\tilde{\mathbf{x}}_i^+ \leq \epsilon)}{|\mathbf{v}_1^T\tilde{\mathbf{x}}_i^+|}, \ i = 1, 2, \cdots, n_A, \tag{26}$$

$$(D_2)_{jj} = \frac{\mathcal{I}(\mathbf{v}_2^T\tilde{\mathbf{x}}_j^- \leq \epsilon)}{|\mathbf{v}_2^T\tilde{\mathbf{x}}_j^-|}, \ j = 1, 2, \cdots, n_B, \tag{27}$$

where $\mathcal{I}(\cdot)$ denotes the indicative function.

Remark 3. For the reason why the proposed minimization problems (24) and (25) leads to the solution of original problems, see section IV-A.

Since the optimization of (25) is similar with (24), we only analyze the optimization problem (24). The manifold regularization term in (24) can be formulated as:

$$\sum_{i,j=1}^{n}|\mathbf{v}_1^T\tilde{\mathbf{x}}_i - \mathbf{v}_1^T\tilde{\mathbf{x}}_j|F_{ij} = \mathbf{v}_1^T\tilde{\mathbf{X}}(\mathbf{D}_G - \mathbf{G})\tilde{\mathbf{X}}^T\mathbf{v}_1 = \mathbf{v}_1^T\tilde{\mathbf{X}}\mathbf{L}_G\tilde{\mathbf{X}}^T\mathbf{v}_1,$$

where $\mathbf{L}_G = \mathbf{D}_G - \mathbf{G}$, and $\mathbf{G}, \mathbf{D}_G$ are defined as

$$G_{ij} = \frac{F_{ij}}{2|\mathbf{v}_1^T\tilde{\mathbf{x}}_i - \mathbf{v}_1^T\tilde{\mathbf{x}}_j|}, \tag{28}$$

$$\mathbf{D}_G = \text{diag}\left(\sum_{j=1}^{n}G_{1j}, \sum_{j=1}^{n}G_{2j}, \cdots, \sum_{j=1}^{n}G_{nj}\right). \tag{29}$$

Therefore, the Lagrangian function of (24) is:

$$L_1(\Theta_1) = \frac{1}{2}\mathbf{v}_1^T\tilde{\mathbf{A}}\mathbf{D}_1\tilde{\mathbf{A}}^T\mathbf{v}_1 + \frac{r_1}{2}\mathbf{v}_1^T\mathbf{v}_1 + \frac{\mu_1}{2}\mathbf{v}_1^T\mathbf{X}\mathbf{L}_G\mathbf{X}^T\mathbf{v}_1 + c_1\mathbf{e}_2^T\xi + \alpha^T(\mathbf{e}_2 + \tilde{\mathbf{B}}\mathbf{v}_1 - \xi) - \beta^T\xi, \tag{30}$$

where $\Theta_1 = \{\mathbf{v}_1, \xi, \alpha, \beta\}$ with α, β as Lagrangian multipliers, and $\alpha, \beta \geq \mathbf{0}$. Take the partial derivative w.r.t $\mathbf{v}_1$ and ξ to be zero, we have

$$\frac{\partial L_1}{\partial \mathbf{v}_1} = \tilde{\mathbf{A}}\mathbf{D}_1\tilde{\mathbf{A}}^T\mathbf{v}_1 + r_1\mathbf{v}_1 + \mu_1\mathbf{X}\mathbf{L}_G\mathbf{X}^T\mathbf{v}_1 + \alpha^T\tilde{\mathbf{B}}^T\mathbf{v}_1 = \mathbf{0}$$
$$\Rightarrow \mathbf{v}_1 = -(\tilde{\mathbf{A}}\mathbf{D}_1\tilde{\mathbf{A}}^T + \mu_1\tilde{\mathbf{X}}\mathbf{L}_G\tilde{\mathbf{X}}^T + r_1\mathbf{I})^{-1}\tilde{\mathbf{B}}\alpha, \tag{31}$$

$$\frac{\partial L_1}{\partial \xi} = c_1\mathbf{e}_2 - \alpha - \beta = \mathbf{0}$$
$$\Rightarrow \beta = c_1\mathbf{e}_2 - \alpha, \tag{32}$$

where $\mathbf{I}$ is identity matrix with appropriate size. Since $\beta \geq \mathbf{0}$, we have

$$\mathbf{0} \leq \alpha \leq c_1\mathbf{e}_2. \tag{33}$$

Algorithm 1 Training linear RMTBSVM

Input: Data matrices $\mathbf{A}, \mathbf{B}$, parameters $c_1, c_2, r_1, r_2, \mu_1, \mu_2$, steps limitation T.
Output: Augment vectors $\mathbf{v}_1, \mathbf{v}_2$.

```
Compute matrices Ã, B̃, X̃.
Initialize v1, v2 with the solution of TBSVM.
for t = 1 to T do
    Compute D1, G, DG with (26), (28), (29). Compute D2, H, DH with (27), (36), (37).
    Compute LG = DG − G and LH = DH − H.
    Compute α, γ by solving (34) and (35) with SOR.
    Compute v1, v2 with (31), (38).
    if v1 and v2 are changeless then
        Break;
    end if
end for
return v1, v2.
```

From (30), (31) and (32), we can obtain the dual of the primal optimization problem:

$$\max_{\alpha} \ \mathbf{e}_2^T\alpha - \frac{1}{2}\alpha^T\tilde{\mathbf{B}}^T(\tilde{\mathbf{A}}\mathbf{D}_1\tilde{\mathbf{A}}^T + \mu_1\tilde{\mathbf{X}}\mathbf{L}_G\tilde{\mathbf{X}}^T + r_1\mathbf{I})^{-1}\tilde{\mathbf{B}}\alpha, \ \text{s.t.} \ \mathbf{0} \leq \alpha \leq c_1\mathbf{e}_2. \tag{34}$$

By solving (34), the argument vector $\mathbf{v}_1$ can be computed with (31). That is, $\mathbf{w}_1$ and b_1 are obtained.

Similarly, we can obtain the dual of (25)

$$\max_{\gamma} \ \mathbf{e}_1^T\gamma - \frac{1}{2}\gamma^T\tilde{\mathbf{A}}^T(\tilde{\mathbf{B}}\mathbf{D}_2\tilde{\mathbf{B}}^T + \mu_2\tilde{\mathbf{X}}\mathbf{L}_H\tilde{\mathbf{X}}^T + r_2\mathbf{I})^{-1}\tilde{\mathbf{A}}\gamma, \ \text{s.t.} \ \mathbf{0} \leq \gamma \leq c_2\mathbf{e}_1, \tag{35}$$

where γ is the Lagrangian multiplier, and $\mathbf{L}_H = \mathbf{D}_H - \mathbf{H}$, with $\mathbf{D}_H, \mathbf{H}$ defined as

$$H_{ij} = \frac{F_{ij}}{2|\mathbf{v}_2^T\tilde{\mathbf{x}}_i - \mathbf{v}_2^T\tilde{\mathbf{x}}_j|}, \tag{36}$$

$$\mathbf{D}_H = \text{diag}\left(\sum_{j=1}^{n}H_{1j}, \sum_{j=1}^{n}H_{2j}, \cdots, \sum_{j=1}^{n}H_{nj}\right). \tag{37}$$

With the solution of (35), $\mathbf{v}_2$ is given by

$$\mathbf{v}_2 = (\tilde{\mathbf{B}}\mathbf{D}_2\tilde{\mathbf{B}}^T + \mu_1\tilde{\mathbf{X}}\mathbf{L}_H\tilde{\mathbf{X}}^T + r_2\mathbf{I})^{-1}\tilde{\mathbf{A}}\gamma. \tag{38}$$

It should be noticed that $\{\mathbf{D}_1, \mathbf{L}_G\}$ and $\{\mathbf{D}_2, \mathbf{L}_H\}$ are both dependent to $\mathbf{v}_1$ and $\mathbf{v}_2$, which means only one step computation cannot obtain the optimal solution of the primal optimization problem (24) and (25). Therefore, we propose an iterative algorithm to compute $\mathbf{v}_1, \mathbf{v}_2$. More details are shown in algorithm 1.

With $\mathbf{v}_1, \mathbf{v}_2$, we can obtain $\mathbf{w}_1, b_1$ and $\mathbf{w}_2, b_2$. Then we use the decision function (8) to classify new points.

C. Nonlinear classification with kernel method

Similar to the linear cases, the optimization problems of nonlinear RMTBSVM are formulated as

$$\min_{\mathbf{w}_1,b_1,\xi} \ \|\mathbf{A}_\Phi^T\mathbf{w}_1 + b_1\mathbf{e}_1\|_{1,\epsilon} + \frac{r_1}{2}(\|\mathbf{w}_1\|_{\mathcal{H}}^2 + b_1^2) + c_1\mathbf{e}_2^T\xi$$

$$+\frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{w}_1^T\phi(\mathbf{x}_i)-\mathbf{w}_1^T\phi(\mathbf{x}_j)|F_{ij},$$
$$\text{s.t. } -(\mathbf{B}_\Phi^T\mathbf{w}_1+b_1\mathbf{e}_2)+\xi\geq\mathbf{e}_2, \tag{39}$$

$$\min_{\mathbf{w}_2,b_2,\eta}\|\mathbf{B}_\Phi^T\mathbf{w}_2+b_2\mathbf{e}_2\|_{1,\epsilon}+\frac{r_2}{2}(\|\mathbf{w}_2\|_{\mathcal{H}}^2+b_2^2)+c_2\mathbf{e}_1^T\eta$$
$$+\frac{\mu_2}{2}\sum_{i,j=1}^{n}|\mathbf{w}_2^T\phi(\mathbf{x}_i)-\mathbf{w}_2^T\phi(\mathbf{x}_j)|F_{ij},$$
$$\text{s.t. } (\mathbf{A}_\Phi^T\mathbf{w}_2+b_2\mathbf{e}_1)-\eta\geq\mathbf{e}_1, \tag{40}$$

where

$$\mathbf{A}_\Phi=[\phi(\mathbf{x}_1^+),\phi(\mathbf{x}_2^+),\cdots,\phi(\mathbf{x}_{n_A}^+)],$$
$$\mathbf{B}_\Phi=[\phi(\mathbf{x}_1^-),\phi(\mathbf{x}_2^-),\cdots,\phi(\mathbf{x}_{n_B}^-)],$$
$$\mathbf{X}_\Phi=[\mathbf{A}_\Phi,\mathbf{B}_\Phi],$$

and $\phi(\cdot)$ is nonlinear feature mapping, $\|\cdot\|_{\mathcal{H}}$ denotes the norm defined by the inner product in the corresponding Hilbert feature space $\mathcal{H}$. In general, we cannot explicitly express the feature mapping $\phi(\cdot)$, but the kernel function of $\phi(\cdot)$ is known, and can be defined as

$$\mathcal{K}:\mathbb{R}^n\times\mathbb{R}^n\mapsto\mathbb{R},\quad \mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j).$$

We then reformulate the primal optimization problems as

$$\min_{\mathbf{w}_1,b_1,\xi}\frac{1}{2}(\mathbf{A}_\Phi^T\mathbf{w}_1+b_1\mathbf{e}_1)^T\mathbf{D}_1^\Phi(\mathbf{A}_\Phi^T\mathbf{w}_1+b_1\mathbf{e}_1)+c_1\mathbf{e}_2^T\xi$$
$$+\frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{w}_1^T\phi(\mathbf{x}_i)-\mathbf{w}_1^T\phi(\mathbf{x}_j)|F_{ij}+\frac{r_1}{2}(\|\mathbf{w}_1\|_{\mathcal{H}}^2+b_1^2),$$
$$\text{s.t. } -(\mathbf{B}_\Phi^T\mathbf{w}_1+b_1e_2)+\xi\geq\mathbf{e}_2, \tag{41}$$

$$\min_{\mathbf{w}_2,b_2,\eta}\frac{1}{2}(\mathbf{B}_\Phi^T\mathbf{w}_2+b_2\mathbf{e}_2)^T\mathbf{D}_2^\Phi(\mathbf{B}_\Phi^T\mathbf{w}_2+b_2\mathbf{e}_2)+c_2\mathbf{e}_1^T\eta$$
$$+\frac{\mu_2}{2}\sum_{i,j=1}^{n}|\mathbf{w}_2^T\phi(\mathbf{x}_i)-\mathbf{w}_2^T\phi(\mathbf{x}_j)|F_{ij}+\frac{r_2}{2}(\|\mathbf{w}_2\|_{\mathcal{H}}^2+b_2^2),$$
$$\text{s.t. } (\mathbf{A}_\Phi^T\mathbf{w}_2+b_2\mathbf{e}_1)-\eta\geq\mathbf{e}_1, \tag{42}$$

where $\mathbf{D}_1^\Phi$ and $\mathbf{D}_2^\Phi$ are defined in Table I. We suppose that the optimal $\mathbf{w}_1,\mathbf{w}_2$ can be formulated as the linear combination of training samples, i.e.

$$\mathbf{w}_1^*=\sum_{i=1}^{n}\mathbf{p}_i\phi(\mathbf{x}_i)=\mathbf{X}_\Phi\mathbf{p},\mathbf{w}_2^*=\sum_{i=1}^{n}\mathbf{q}_i\phi(\mathbf{x}_i)=\mathbf{X}_\Phi\mathbf{q} \tag{43}$$

where $\mathbf{p},\mathbf{q}$ are linear representation coefficient. This assumption greatly simplifies the problem. (we will justify why it is valid in section IV-A). We further rewritten the optimization problem (41) and (42) as follows:

$$\min_{\mathbf{p},b_1,\xi}(\mathbf{A}_\Phi^T\mathbf{X}_\Phi\mathbf{p}+b_1\mathbf{e}_1)^T\mathbf{D}_1^\Phi(\mathbf{A}_\Phi^T\mathbf{X}_\Phi\mathbf{p}+b_1\mathbf{e}_1)+c_1\mathbf{e}_2^T\xi$$
$$+\frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{p}^T\mathbf{K}(:,\mathbf{x}_i)-\mathbf{p}^T\mathbf{K}(:,\mathbf{x}_j)|F_{ij}+\frac{r_1}{2}(\mathbf{p}^T\mathbf{K}\mathbf{p}+b_1^2),$$
$$\text{s.t. } -(\mathbf{K}_B^T\mathbf{p}+b_1\mathbf{e}_2)+\xi\geq\mathbf{e}_2,$$

$$\min_{\mathbf{q},b_2,\eta}(\mathbf{B}_\Phi^T\mathbf{X}_\Phi q+b_2\mathbf{e}_2)^T\mathbf{D}_2^\Phi(\mathbf{B}_\Phi^T\mathbf{X}_\Phi\mathbf{q}+b_2\mathbf{e}_2)+c_2e_1^T\xi$$
$$+\frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{q}^T\mathbf{K}(:,\mathbf{x}_i)-\mathbf{q}^T\mathbf{K}(:,\mathbf{x}_j)|F_{ij}+\frac{r_2}{2}(\mathbf{q}^T\mathbf{K}\mathbf{q}+b_2^2),$$
$$\text{s.t. } (\mathbf{K}_A^T\mathbf{q}+b_2\mathbf{e}_1)-\eta\geq\mathbf{e}_1,$$

where $\mathbf{K}(:,\mathbf{x}_i)$ denotes i-th column of kernel matrix $\mathbf{K}$. Though $\mathbf{A}_\Phi,\mathbf{B}_\Phi,\mathbf{X}_\Phi$ are dependent to $\phi(\cdot)$, the kernel matrices $\mathbf{K}_A=\mathbf{X}_\Phi^T\mathbf{A}_\Phi$, $\mathbf{K}_B=\mathbf{X}_\Phi^T\mathbf{B}_\Phi$ and $\mathbf{K}=\mathbf{X}_\Phi^T\mathbf{X}_\Phi$ can be computed with kernel functions. Therefore we only need to use kernel matrices to reformulate the optimization problems:

$$\min_{\mathbf{u}_1,\xi}\frac{1}{2}\mathbf{u}_1^T\tilde{\mathbf{K}}_A\mathbf{D}_1^\Phi\tilde{\mathbf{K}}_A\mathbf{u}_1+\frac{r_1}{2}\mathbf{u}_1^T\mathbf{K}_p\mathbf{u}_1+c_1\mathbf{e}_2^T\xi$$
$$+\frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{u}^T\tilde{\mathbf{K}}(:,\mathbf{x}_i)-\mathbf{u}^T\tilde{\mathbf{K}}(:,\mathbf{x}_j)|F_{ij},$$
$$\text{s.t. } -\tilde{\mathbf{K}}_B^T\mathbf{u}_1+\xi\geq\mathbf{e}_2, \tag{44}$$

$$\min_{\mathbf{u}_2,\eta}\frac{1}{2}\mathbf{u}_2^T\tilde{\mathbf{K}}_B\mathbf{D}_2^\Phi\tilde{\mathbf{K}}_B\mathbf{u}_2+\frac{r_2}{2}\mathbf{u}_2^T\mathbf{K}_p\mathbf{u}_2+c_2\mathbf{e}_1^T\eta$$
$$+\frac{\mu_1}{2}\sum_{i,j=1}^{n}|\mathbf{u}_2^T\tilde{\mathbf{K}}(:,\mathbf{x}_i)-\mathbf{u}_2^T\tilde{\mathbf{K}}(:,\mathbf{x}_j)|F_{ij},$$
$$\text{s.t. } (\mathbf{K}_A^T\mathbf{q}+b_2\mathbf{e}_1)-\eta\geq\mathbf{e}_1, \tag{45}$$

and other notations can be find in Table I. Similar with linear case, we can obtain

$$\mathbf{u}_1=-(\tilde{\mathbf{K}}_A\mathbf{D}_1^\Phi\tilde{\mathbf{K}}_A^T+\mu_1\tilde{\mathbf{K}}\mathbf{L}_G^\Phi\tilde{\mathbf{K}}+r_1\mathbf{K}_p)^{-1}\tilde{\mathbf{K}}_B\alpha, \tag{46}$$
$$\mathbf{u}_2=(\tilde{\mathbf{K}}_B\mathbf{D}_2^\Phi\tilde{\mathbf{K}}_B^T+\mu_2\tilde{\mathbf{K}}\mathbf{L}_H^\Phi\tilde{\mathbf{K}}+r_2\mathbf{K}_p)^{-1}\tilde{\mathbf{K}}_A\gamma, \tag{47}$$

where α and γ are Lagrangian multipliers. The dual problems of (44) and (45) are

$$\max_{\alpha}-\frac{1}{2}\alpha^T\tilde{\mathbf{K}}_A^T(\tilde{\mathbf{K}}_A\mathbf{D}_1^\Phi\tilde{\mathbf{K}}_A^T+\mu_2\tilde{\mathbf{K}}\mathbf{L}_G^\Phi\tilde{\mathbf{K}}+r_1\mathbf{K}_p)^{-1}\tilde{\mathbf{K}}_B\alpha$$
$$+\mathbf{e}_2^T\alpha,\quad \text{s.t. } \mathbf{0}\leq\gamma\leq c_1\mathbf{e}_2, \tag{48}$$

$$\max_{\gamma}-\frac{1}{2}\gamma^T\tilde{\mathbf{K}}_A^T(\tilde{\mathbf{K}}_B\mathbf{D}_2^\Phi\tilde{\mathbf{K}}_B^T+\mu_2\tilde{\mathbf{K}}\mathbf{L}_H^\Phi\tilde{\mathbf{K}}+r_2\mathbf{K}_p)^{-1}\tilde{\mathbf{K}}_A\gamma$$
$$+\mathbf{e}_1^T\gamma,\quad \text{s.t. } \mathbf{0}\leq\gamma\leq c_2\mathbf{e}_1. \tag{49}$$

It is the same for nonlinear case that we should use iterative algorithm to compute $\mathbf{u}_1,\mathbf{u}_2$. The details of our proposed algorithm are listed in algorithm 2.

Finally, similar to (8), we define the decision function of nonlinear RMTBSVM as

$$h(\mathbf{x})=\text{sgn}\left(\frac{|\mathbf{q}^T\mathbf{X}_\Phi^T\mathbf{x}+b_2|}{\sqrt{\mathbf{q}^T\mathbf{K}\mathbf{q}}}-\frac{|\mathbf{p}^T\mathbf{X}_\Phi^T\mathbf{x}+b_1|}{\sqrt{\mathbf{p}^T\mathbf{K}\mathbf{p}}}\right), \tag{50}$$

where $\mathbf{X}_\Phi^T\mathbf{x}$ can be computed with kernel function.

IV. Algorithm analysis

In this section, we first give the convergence analysis for the proposed method for linear and nonlinear cases, respectively. And then, we also reveal more discriminative properties of RMTBSVM based on manifold regularization view.

A. Convergence analysis

In this section, we prove the convergence of the proposed algorithms in linear and nonlinear cases.

TABLE I
NOTATION TABLE FOR NONLINEAR RMTBSVM

Description	Definition of vectors or matrices
Augment vectors or matrices	$\mathbf{u}_1 = \begin{bmatrix}\mathbf{p}\\ b_1\end{bmatrix},\ \mathbf{u}_2 = \begin{bmatrix}\mathbf{q}\\ b_2\end{bmatrix},\ \tilde{\mathbf{K}}_A = \begin{bmatrix}\mathbf{K}_A\\ \mathbf{e}_1^T\end{bmatrix},\ \tilde{\mathbf{K}}_B = \begin{bmatrix}\mathbf{K}_B\\ \mathbf{e}_2^T\end{bmatrix},\ \tilde{\mathbf{K}} = \begin{bmatrix}\mathbf{K}\\ \mathbf{e}^T\end{bmatrix},\ \mathbf{K}_p = \begin{bmatrix}\mathbf{K} & \mathbf{0}\\ \mathbf{0}^T & 1\end{bmatrix}.$
Diagonal matrices $\mathbf{D}_1^\Phi, \mathbf{D}_2^\Phi$	$\mathbf{D}_1^\Phi = \operatorname{diag}\left\{\frac{\mathcal{I}(\mathbf{u}_1^T\tilde{\mathbf{K}}(:,\mathbf{x}_i^+)\le\epsilon)}{\lvert\mathbf{u}_1^T\tilde{\mathbf{K}}(:,\mathbf{x}_i^+)\rvert}\right\}_{i=1}^{n_A},\ \mathbf{D}_2^\Phi = \operatorname{diag}\left\{\frac{\mathcal{I}(\mathbf{u}_2^T\tilde{\mathbf{K}}(:,\mathbf{x}_i^-)\le\epsilon)}{\lvert\mathbf{u}_2^T\tilde{\mathbf{K}}(:,\mathbf{x}_i^-)\rvert}\right\}_{i=1}^{n_B}.$
Graph matrices $\mathbf{G}^\Phi, \mathbf{H}^\Phi$	$G_{ij}^\Phi = \frac{F_{ij}}{2\lvert\mathbf{u}_1^T\tilde{\mathbf{K}}(:,\mathbf{x}_i) - \mathbf{u}_1^T\tilde{\mathbf{K}}(:,\mathbf{x}_j)\rvert},\ H_{ij}^\Phi = \frac{F_{ij}}{2\lvert\mathbf{u}_2^T\tilde{\mathbf{K}}(:,\mathbf{x}_i) - \mathbf{u}_2^T\tilde{\mathbf{K}}(:,\mathbf{x}_j)\rvert}.$
Laplacian matrices $\mathbf{L}_G^\Phi, \mathbf{L}_H^\Phi$	$\mathbf{D}_G^\Phi = \operatorname{diag}\left\{\sum_{j=1}^n G_{ij}^\Phi\right\}_{i=1}^n, \mathbf{D}_H^\Phi = \operatorname{diag}\left\{\sum_{j=1}^n H_{ij}^\Phi\right\}_{i=1}^n,\ \begin{matrix}\mathbf{L}_G^\Phi = \mathbf{D}_G^\Phi - \mathbf{G}^\Phi\\ \mathbf{L}_H^\Phi = \mathbf{D}_H^\Phi - \mathbf{H}^\Phi\end{matrix}.$

Algorithm 2 Training nonlinear RMTBSVM

Input: Data matrices $\mathbf{A}, \mathbf{B}$, kernel function $\mathcal{K}$, parameters $c_1, c_2, r_1, r_2, \mu_1, \mu_2$, steps limitation T.
Output: Augment vectors u_1, u_2.

```
Compute matrices K̃_A, K̃_B, K̃, K_p with Table I.
Initialize u_1, u_2 with the solution of kernel TBSVM.
for t = 1 to T do
    Compute D_1^Φ, D_2^Φ, L_G^Φ, L_H^Φ based on Table I.
    Compute α, γ by solving (48) and (49) with SOR.
    Compute u_1, u_2 with (46), (47).
    if u_1 and u_2 are changeless then
        Break;
    end if
end for
return u_1, u_2.
```

Recall that in section III-B, we reformulate the objective function in (22) to obtain an easy-to-solved optimization problem (24). To explain the relationship between (22) and (24), we propose the following theorem.

Theorem 1. In each iteration, optimizing (24) is equivalent to minimizing an upper bound of the objective function in (22).

Proof: Inspired by [35], we first define two functions:

$$\mathbf{h}: \mathbb{R}^d \mapsto \mathbb{R}_+^d, \quad h(\mathbf{x}) = \left[|x_1|, |x_2|, \cdots, |x_d|\right]^T, \tag{51}$$

$$g: \mathbb{R}_+^d \mapsto \mathbb{R}_+, \quad g(\mathbf{x}) = \sum_{i=1}^n \min\{x_i, \epsilon\}. \tag{52}$$

Then we can rewrite the primal problem (22) as

$$\begin{aligned}&\min_{\mathbf{w}_1, b_1} g(\mathbf{h}(\mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1)) + \mathcal{R}_K + \mathcal{R}_M\\ =&\min_{\mathbf{w}_1, b_1} g(\mathbf{h}(\mathbf{z})) + \mathcal{R},\end{aligned} \tag{53}$$

where $\mathbf{z} = \mathbf{z}(\mathbf{w}_1, b_1) = \mathbf{A}^T\mathbf{w}_1 + b_1\mathbf{e}_1$, and $\mathcal{R}$ is defined as

$$\mathcal{R} = \mathcal{R}(\mathbf{w}_1, b_1, \mathbf{G}) = \mathcal{R}_K(\mathbf{w}_1, b_1) + \mathcal{R}_M(\mathbf{w}_1, b_1, \mathbf{G}), \tag{54}$$

$$\mathcal{R}_K(\mathbf{w}_1, b_1) = \sum_{i=1}^{n_B}(1 - b_1 - \mathbf{w}_1^T\mathbf{x}_i^-)_+ + \frac{r_1}{2}(\|\mathbf{w}_1\|_2^2 + b_1^2), \tag{55}$$

$$\mathcal{R}_M(\mathbf{w}_1, b_1, \mathbf{G}) = \frac{\mu_1}{2}\sum_{i,j=1}^n (\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\mathbf{x}_j)^2 G_{ij}, \tag{56}$$

In t-th iteration, since $g(\cdot)$ is a concave function, the inequality (57) holds based on the definition of sub-gradient:

$$g(\mathbf{h}(\mathbf{z})) \le g(\mathbf{h}(\mathbf{z}^{(t)})) + \left\langle \Omega^{(t)}, \mathbf{h}(\mathbf{z}) - \mathbf{h}(\mathbf{z}^{(t)})\right\rangle, \tag{57}$$

where $\Omega^{(t)}$ is a sub-gradient of $g(\mathbf{u})$ at $\mathbf{u} = \mathbf{h}(\mathbf{z}^{(t)})$. We can find one of sub-gradients of $g(\mathbf{u})$ at $\mathbf{u} = \mathbf{h}(\mathbf{z}^{(t)})$

$$\Omega^{(t)} = \frac{1}{2}\left[\mathcal{I}(z_1^{(t)} \le \epsilon), \mathcal{I}(z_2^{(t)} \le \epsilon), \cdots, \mathcal{I}(z_{n_A}^{(t)} \le \epsilon)\right]^T. \tag{58}$$

By adding $\mathcal{R}$ on both sides in (57) we have

$$\begin{aligned}&g(\mathbf{h}(\mathbf{z})) + \mathcal{R}\\ \le\ & g(\mathbf{h}(\mathbf{z}^{(t)})) + \left\langle \Omega^{(t)}, \mathbf{h}(\mathbf{z}) - \mathbf{h}(\mathbf{z}^{(t)})\right\rangle + \mathcal{R},\end{aligned} \tag{59}$$

where $\langle\cdot,\cdot\rangle$ denotes inner product operator. Then we minimize the right side of (59) with respect to $\mathbf{w}_1, b_1$. Since $\Omega^{(t)}$ and $\mathbf{h}(\mathbf{z}^{(t)})$ are both seen as constants, we have

$$\begin{aligned}&\min_{\mathbf{w}_1, b_1} g(\mathbf{h}(\mathbf{z}^{(t)})) + \left\langle \Omega^{(t)}, \mathbf{h}(\mathbf{z}) - \mathbf{h}(\mathbf{z}^{(t)})\right\rangle + \mathcal{R}\\ \Leftrightarrow &\min_{\mathbf{w}_1, b_1} (\Omega^{(t)})^T\mathbf{h}(\mathbf{z}) + \mathcal{R}.\end{aligned} \tag{60}$$

We can easily find that (60) is equivalent to the optimization problem (24). Hence we say optimizing (24) is equivalent to minimizing an upper bound of the original problem. ■

We can understand Theorem 1 as follows: The primal ob-

jective function in (22) is non-convex and is hard to minimize directly, but Theorem 1 ensures the proposed algorithm minimizing an upper bound of the original objective in each step. Based on Theorem 1, we can further prove the convergence of algorithm 1. We begin with the following lemma [16]:

Lemma 1. For all positive real number a, b, the following inequality holds

$$\sqrt{a} - \frac{a}{2\sqrt{b}} \leq \sqrt{b} - \frac{b}{2\sqrt{b}}. \tag{61}$$

With Lemma 1, we have the following theorem.

Theorem 2. In each iteration, algorithm 1 monotonically decreases the value of objective function (22). Therefore, the iterative series will converge to a local optimum.

Proof: For convenience, we denote the objective function of (22) in t-th iteration as $\mathcal{G}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{G}^{(t)})$. According to the proof of Theorem 1, we have

$$\mathcal{G}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{G}^{(t)}) = g(\mathbf{h}(\mathbf{z}^{(t)})) + \mathcal{R}^{(t)}, \tag{62}$$

where $\mathbf{z}^{(t)} = \mathbf{A}^T\mathbf{w}_1^{(t)} + b_1^{(t)}\mathbf{e}_1$ and $\mathcal{R}^{(t)} = \mathcal{R}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{G}^{(t)})$ is given by (54). Therefore, the conclusion we will prove is $\mathcal{G}(\mathbf{w}_1^{(t+1)}, b_1^{(t+1)}, \mathbf{G}^{(t+1)}) \leq \mathcal{G}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{G}^{(t)})$. On the other hand, we denote the objective function of (24) as $\mathcal{F}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{D}_1^{(t)}, \mathbf{G}^{(t)})$. Similarly we have

$$\mathcal{F}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{D}_1^{(t)}, \mathbf{G}^{(t)}) = (\Omega^{(t)})^T\mathbf{h}(\mathbf{z}^{(t)}) + \mathcal{R}^{(t)}. \tag{63}$$

In algorithm 1, we first solve the dual problem (34) and use (31) to obtain $\mathbf{v}_1^{(t+1)}$, which further minimizes the objective function. Therefore, we have

$$\mathcal{F}(\mathbf{w}_1^{(t+1)}, b_1^{(t+1)}, \mathbf{D}_1^{(t)}, \mathbf{G}^{(t)}) \leq \mathcal{F}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{D}_1^{(t)}, \mathbf{G}^{(t)}). \tag{64}$$

For simplification, we define $\mathcal{R}_L$ as:

$$\mathcal{R}_L(\mathbf{w}_1, b_1; \Omega^{(t)}) = (\Omega^{(t)})^T\mathbf{h}(\mathbf{z}) + \mathcal{R}_K,$$

where $\Omega^{(t)}$ is seen as a known constant with respective to $\mathcal{R}_L$. Then we can rewritten the above inequality (64) as

$$\begin{aligned}
&\mathcal{R}_L^{(t+1)} + \frac{\mu_1}{2}\sum_{i,j=1}^{n}\frac{|(\mathbf{w}_1^{(t+1)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t+1)})^T\mathbf{x}_j|^2}{2|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|}F_{ij} \\
&\leq \mathcal{R}_L^{(t)} + \frac{\mu_1}{2}\sum_{i,j=1}^{n}\frac{|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|^2}{2|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|}F_{ij},
\end{aligned} \tag{65}$$

where

$$\begin{aligned}
\mathcal{R}_L^{(t)} &= \mathcal{R}_L(\mathbf{w}_1^{(t)}, b_1^{(t)}; \Omega^{(t)}), \\
\mathcal{R}_L^{(t+1)} &= \mathcal{R}_L(\mathbf{w}_1^{(t+1)}, b_1^{(t+1)}; \Omega^{(t)}).
\end{aligned}$$

We can further write the inequality (65) as

$$\begin{aligned}
&\mathcal{R}_L^{(t+1)} + \frac{\mu_1}{2}\sum_{i,j=1}^{n}\frac{(|(\mathbf{w}_1^{(t+1)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t+1)})^T\mathbf{x}_j|F_{ij}^2}{2|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij}} \\
&\leq \mathcal{R}_L^{(t)} + \frac{\mu_1}{2}\sum_{i,j=1}^{n}\frac{(|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij})^2}{2|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij}}.
\end{aligned} \tag{66}$$

In Lemma 1, if we set

$$\begin{aligned}
a &= \left|\left((\mathbf{w}_1^{(t+1)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t+1)})^T\mathbf{x}_j\right)F_{ij}\right|^2, \\
b &= \left|\left((\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j\right)F_{ij}\right|^2,
\end{aligned}$$

then we can easily obtain following inequality

$$\begin{aligned}
&\sum_{i,j=1}^{n}|(\mathbf{w}_1^{(t+1)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t+1)})^T\mathbf{x}_j|F_{ij} \\
&- \sum_{i,j=1}^{n}\frac{|(\mathbf{w}_1^{(t+1)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t+1)})^T\mathbf{x}_j|^2F_{ij}^2}{2|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij}} \\
&\leq \sum_{i,j=1}^{n}|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij} \\
&- \sum_{i,j=1}^{n}\frac{|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|^2F_{ij}^2}{2|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij}}.
\end{aligned} \tag{67}$$

Combining two inequalities (66) and (67), we can obtain

$$\begin{aligned}
&\mathcal{R}_L^{(t+1)} + \frac{\mu_1}{2}\sum_{i,j=1}^{n}|(\mathbf{w}_1^{(t+1)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t+1)})^T\mathbf{x}_j|F_{ij} \\
\leq\ &\mathcal{R}_L^{(t)} + \frac{\mu_1}{2}\sum_{i,j=1}^{n}|(\mathbf{w}_1^{(t)})^T\mathbf{x}_i - (\mathbf{w}_1^{(t)})^T\mathbf{x}_j|F_{ij},
\end{aligned}$$

which is same as $\mathcal{R}_L^{(t+1)} + \mathcal{R}_M^{(t+1)} \leq \mathcal{R}_L^{(t)} + \mathcal{R}_M^{(t)}$, and it is equivalent to

$$(\Omega^{(t)})^T\mathbf{h}(\mathbf{z}^{(t+1)}) + \mathcal{R}^{(t+1)} \leq (\Omega^{(t)})^T\mathbf{h}(\mathbf{z}^{(t)}) + \mathcal{R}^{(t)}. \tag{68}$$

Therefore, we have the following inequality

$$\begin{aligned}
&g(\mathbf{h}(\mathbf{z}^{(t+1)})) + \mathcal{R}^{(t+1)} \\
\leq\ &g(\mathbf{h}(\mathbf{z}^{(t)})) + \left\langle\Omega^{(t)}, \mathbf{h}(\mathbf{z}^{(t+1)}) - \mathbf{h}(\mathbf{z}^{(t)})\right\rangle + \mathcal{R}^{(t+1)} \\
\leq\ &g(\mathbf{h}(\mathbf{z}^{(t)})) + \left\langle\Omega^{(t)}, \mathbf{h}(\mathbf{z}^{(t)}) - \mathbf{h}(\mathbf{z}^{(t)})\right\rangle + \mathcal{R}^{(t)} \\
=\ &g(\mathbf{h}(\mathbf{z}^{(t)})) + \mathcal{R}^{(t)}.
\end{aligned}$$

That is, $\mathcal{G}(\mathbf{w}_1^{(t+1)}, b_1^{(t+1)}, \mathbf{G}^{(t+1)}) \leq \mathcal{G}(\mathbf{w}_1^{(t)}, b_1^{(t)}, \mathbf{G}^{(t)})$. Therefore, algorithm 1 monotonically decreases objective function in (22). ■

In the nonlinear case, the kernel function is used and the training samples are mapped to a Hilbert feature spaces. The Representer Theorem proposed in [33] gives the form of the solution to the optimization problems under manifold regularization framework, which is shown as Lemma 2.

Lemma 2 (Representer Theorem [33], [36])**.** For the optimization problem (13), the optimal $\mathbf{f}^*$ admits an expansion:

$$\mathbf{f}^* = \sum_{i=1}^{n}\alpha_i\mathcal{K}(\cdot, \mathbf{x}_i), \tag{69}$$

where $\mathcal{K}(\cdot,\cdot)$ is the kernel function.

With this lemma, we have the following theorem ensuring the convergence of algorithm 2.

Theorem 3. In each iteration, algorithm 2 monotonically

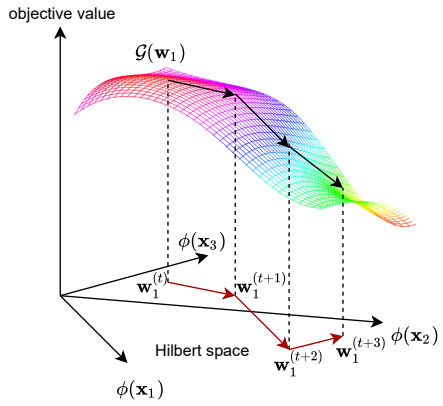

Fig. 2. A visualization of the iterative process in nonlinear RMTBSVM. For convenience, the bias b is ignored in this figure. In t-th iteration, the optimal $\mathbf{w}_1^{(t_*)}$ should be in the span of $\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_n)$. Since $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t_*)}$, the objective will monotonically decrease.

decreases the value of objective function (44). Therefore, the iterative series will converge to a local optimum.

Proof: In t-th iteration, we rewritten optimization problem (41):

$$\begin{aligned}&\min_{\mathbf{w}_1,b_1,\xi} (\mathbf{A}_\Phi^T\mathbf{w}_1 + b_1\mathbf{e}_1)^T\mathbf{D}_1^\Phi(\mathbf{A}_\Phi^T\mathbf{w}_1 + b_1\mathbf{e}_1)\\ &+ c_1\sum_{i=1}^{n_B}(1 + b_1 + \mathbf{w}_1^T\phi(\mathbf{x}_i^-))_+ + \frac{r_1}{2}(\|\mathbf{w}_1\|_{\mathcal{H}}^2 + b_1^2)\\ &+ \frac{\mu_1}{2}\sum_{i,j=1}^{n}\left[\mathbf{w}_1^T\phi(\mathbf{x}_i) - \mathbf{w}_1^T\phi(\mathbf{x}_j)\right]^2 G_{ij}^{(t)}.\end{aligned}$$

This optimization problem is in the form of (13). Thus according to Lemma 2, in t-th iteration, the optimal solution $\mathbf{w}_1^{(t_*)}$ is formulated as

$$\mathbf{w}_1^{(t_*)} = \sum_{i=1}^{n} p_i\phi(\mathbf{x}_i).$$

Therefore, in algorithm 2, the derived $\mathbf{w}^{(t+1)}$ satisfies $\mathbf{w}_1^{(t+1)} = \mathbf{w}_1^{(t_*)}$. That is, $\mathbf{w}^{(t+1)}$ minimizes the objective function in (41) in each iteration, which is same as (64). Hence the remained proof is similar to that of Theorem 2 and it is omitted for avoiding repetition. ∎

From Theorem 3, in each iteration, the Representer Theorem gives a form of optimal $\mathbf{w}$, which is the linear combination of $\phi(\mathbf{x}_i)$. Based on this, the optimal $\mathbf{w}^{(t_*)}$ can be obtained by iteration. This is the reason why we should formulate $\mathbf{w}^*$ as (43). We illustrate the iteration process in Fig. 2. For simplification, we ignore the bias term b_1.

B. Multi-class RMTBSVM

The RMTBSVM are discussed as binary classifier above. Here, we extend our proposed RMTBSVM with OVR strategy for multi-class classification. We use the same notations as section II-A. The i-th optimization problem of k-class RMTBSVM is formulated as

$$\begin{aligned}&\min_{\mathbf{w}_i,b_i,\xi_i} \frac{1}{2}\|\mathbf{X}_i^T\mathbf{w}_i + b_i\mathbf{e}_i\|_2^2 + c_i\mathbf{e}_{-i}^T\xi_i + \frac{r_i}{2}(\|\mathbf{w}_i\|_2^2 + b_i^2),\\ &+ \frac{\mu_i}{2}\sum_{r,s=1}^{n}|\mathbf{w}_i^T\mathbf{x}_r - \mathbf{w}_i^T\mathbf{x}_s|F_{rs},\\ &\text{s.t. } -(\mathbf{X}_{-i}^T\mathbf{w}_i + b_i\mathbf{e}_{-i}) + \xi_i \geq \mathbf{e}_{-i}, \xi_i \geq 0. \qquad (70)\end{aligned}$$

where r_i, μ_i are regularization parameters for i-th class, This problem can be easily solved by algorithm 1.

C. The discriminant of RMTBSVM

In this section, we study the manifold regularization term in our method. In general, the geometric neighbor information is used in the construction of graph matrix $\mathbf{F}$. In our work, we add supervised information in $\mathbf{F}$ to enhance the discriminative ability of model, which is defined as

$$F_{ij} = \begin{cases} 1/n_A, & y_i = y_j = 1;\\ 1/n_B, & y_i = y_j = -1;\\ 0, & \text{Otherwise.}\end{cases} \qquad (71)$$

We consider the manifold regularization with L_2-norm metric. According to [37], we have

$$\begin{aligned}&\sum_{i,j=1}^{n}(\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\mathbf{x}_j)^2F_{ij}\\ &= \frac{1}{n_A}\sum_{i,j=1}^{n_A}[\mathbf{w}_1^T(\mathbf{x}_i^+ - \mathbf{x}_j^+)]^2 + \frac{1}{n_B}\sum_{i,j=1}^{n_B}[\mathbf{w}_1^T(\mathbf{x}_i^- - \mathbf{x}_j^-)]^2\\ &= \mathbf{w}_1^T(\mathbf{I} - \frac{1}{n_A}\mathbf{e}_1\mathbf{e}_1^T)\mathbf{w}_1 + \mathbf{w}_1^T(\mathbf{I} - \frac{1}{n_B}\mathbf{e}_2\mathbf{e}_2^T)\mathbf{w}_1\\ &= \mathbf{w}_1^T(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{w}_1,\end{aligned}$$

where $\mathbf{S}_1, \mathbf{S}_2$ denote within-class scatter of positive and negative class, respectively. According to fisher discriminant principle, minimizing this term derives a more discriminative model. Since it minimizes the sum of within-class scatter, the samples in the same class will be mapped to be close to each other. Similarly, in L_1-norm case, the manifold regularization term becomes

$$\frac{1}{n_A}\sum_{i,j=1}^{n_A}\left|\mathbf{w}_1^T(\mathbf{x}_i^+ - \mathbf{x}_j^+)\right| + \frac{1}{n_B}\sum_{i,j=1}^{n_B}\left|\mathbf{w}_2^T(\mathbf{x}_i^- - \mathbf{x}_j^-)\right|,$$

which leads the model to minimize within-class scatter based on L_1-norm metric. Therefore, our method can better discriminate the data from different classes with strong robustness. This is another reason why the proposed method is more robust than other methods.

V. Experiments

In this section, we carried the experiments on both binary-class and multi-class datasets to test the effectiveness of our model. We compare our method with other methods in classification accuracy and robustness. The experiments were run on PC (CPU: Intel Core i7, 2.30GHz; RAM: 8.00GB; OS: 64-bit Windows10).

TABLE II
THE PROFILE OF DATASETS USED IN THE EXPERIMENTS. IN THE EXPERIMENTS, THE RATIO OF THE NUMBER OF TRAINING SAMPLE VERSUS THE TEST SAMPE AS 6:4.

Dataset	#feature	#sample	#class
Heart	13	270	2
Sonar	60	208	2
Australian	14	690	2
Breast	10	683	2
German	24	1000	2
Adult	119	2000	2
Mushroom	112	4000	2
BreastMNIST	28×28	702	2
PneumoniaMNIST	28×28	2000	2
DNA	180	2000	3
Pendigits	16	2000	10
USPS	16×16	2000	10
DermaMNIST	28×28×3	5000	7

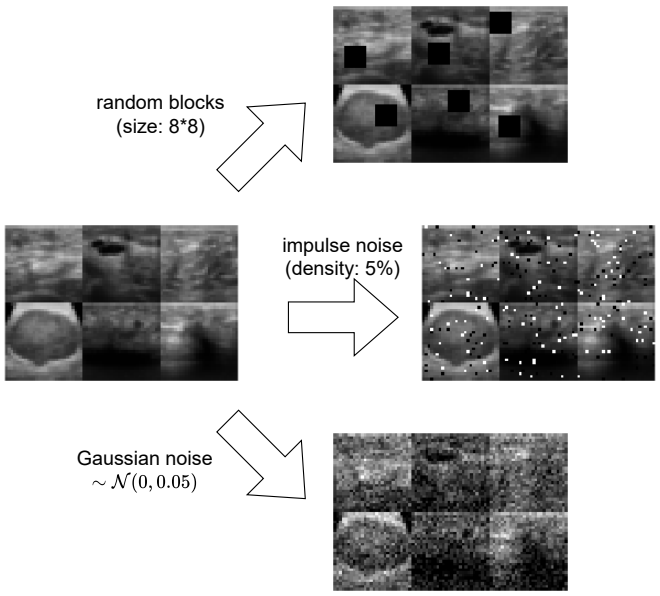

Fig. 3. The original images and images corrupted by different noises

A. Experimental setup

In the experiments, data preprocessing was first carried out. The continuous features were scaled to $[-1, 1]$ interval, and the discrete features were expanded and encoded with one-hot coding. We perform experiments on medical image datasets [38], including BreastMNIST [39], PneumoniaMNIST [40] and DermaMNIST datasets [41]. The first two are binary-class datasets and the third is a multi-class dataset. The images of these datasets were preprocessed as 28×28 pixels [38]. The datasets we used in the experiments are shown in Table II. We split each dataset into training set and testing set with the proportion of samples as 6:4.

In the experiments, we tune the hyperparameters for the best performance. Since RMTBSVM has many hyperparameters, for simplification, we set the parameters as $c_i = c$, $r_i = r$ and $\mu_i = \mu$ $(i = 1, 2, \cdots, k)$, where k is the number of categories. The optimal c is selected from $\{2^{-5}, 2^{-4}, \cdots, 2^5\}$. r is chosen from $\{10^{-3}, 10^{-2}, \cdots, 10^3\}$, and μ is selected from $\{0\} \cup \{10^i : i = -7, -6, \cdots, 0\}$. The kernel method was used for nonlinear classification. In the experiments, we mainly use RBF kernel:

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right),$$

and γ is selected from $\{2^{-5}, \cdots, 2^5\}$. The optimal range of the threshold parameter ϵ usually varies with different datasets. To determine it, we first set $\epsilon = \infty$ and compute $Z = |\mathbf{w}^T \phi(\mathbf{x}) + b|$ with all training samples, and estimate the mean value $\mathbb{E}(Z)$ and standard deviation $\sigma(Z)$. Then the optimal ϵ is searched in the interval $[\mathbb{E}(Z), \mathbb{E}(Z)+3\sigma(Z)]$, and we use $3\sigma(Z)/10$ as step size to adjust ϵ. To search the optimal combination of parameters, we perform grid searching with three-fold cross-validation. The parameters corresponding to the highest average accuracy on the validation set are selected as the best parameters.

In the experiments, The proposed RMTBSVM was compared with other five baseline SVM methods, including SVM implemented by LIBSVM [42], LSSVM [43], TBSVM [12], L_1-TWSVM [19] and CTWSVM [21]. We also make comparison with deep learning methods. Three different deep neural networks are considered, i.e., convolutional neural network (CNN) [44], residual neural network (ResNet) [45], and vision transformer (ViT) [46]. The loss functions of the deep neural networks are set as cross entropy loss. In addition, we also study the performance of RMTBSVM using the deep learning features. The features are extracted with the deep neural networks and used as the input of RMTBSVM.

B. General experimental results

For each dataset, we randomly select the training data and testing data for 10 times. Then we record the training time and compute the average classification accuracy as well as the standard deviation of each method. The results are shown in Table III. The highest accuracy of each dataset is highlighted by boldface. Then, we perform paired t-test comparing RMTBSVM with other methods, and compute the p-value to check the statistical significance. The null hypothesis is that the testing accuracy of RMTBSVM has no difference compared to other methods. We report the results of significance tests in Table IV. The null hypothesis is rejected if $p < 0.05$, which means that the two methods have significantly different performances.

From Table III, we can observe that our proposed RMTBSVM achieves the best accuracies among all methods on 10 out of 13 datasets. Our method obtains significantly higher accuracies on Sonar, BreastMNIST, and DermaMNIST datasets than the second-best method. The baseline method, SVM, achieves higher accuracy than RMTBSVM on German dataset, but obtain significantly poor performance on Sonar, Pendigits and DermaMNIST. The reason may be the traditional SVM cannot infer the complex intrinsic distribution of these datasets. In general, the nonparallel SVM methods (i.e., TBSVM, L_1-TWSVM, CTWSVM, and RMTBSVN) perform

TABLE III
THE AVERAGE TRAINING TIME (S), CLASSIFICATION ACCURACY (%), AND THE CORRESPONDING STANDARD DEVIATION ON DIFFERENT DATASETS.

DataSets	SVM	LSSVM	TBSVM	L_1-TWSVM	CTWSVM	RMTBSVM
Heart	82.78±2.66 0.0008	81.85±2.24 0.0015	82.22±3.01 0.0060	81.76±2.72 0.0565	82.13±2.75 0.0514	**83.89**±2.07 0.0123
Mushroom	99.93±0.05 0.6201	99.93±0.05 0.9078	99.92±0.04 2.8160	99.97±0.57 7.1499	99.98±0.57 13.7819	**100.00**±0.00 4.0142
German	**74.28**±1.52 0.0299	73.33±1.65 0.0466	71.70±1.64 0.1651	71.40±1.58 2.342	71.53±1.74 2.8643	74.20±1.01 0.3765
Sonar	80.24±4.70 0.0032	80.12±4.48 0.0021	82.65±4.83 0.0086	82.17±3.52 0.0148	82.05±4.15 0.0205	**85.90**±3.05 0.0248
Australian	85.65±2.30 0.0049	84.60±1.75 0.0146	86.05±1.90 0.0492	85.65±2.04 0.6903	**86.52**±1.83 0.9249	86.49±2.03 0.1146
Breast	96.30±0.66 0.0017	96.26±0.73 0.0102	96.52±0.62 0.0311	96.04±0.83 0.3371	**96.59**±0.73 0.4578	96.56±0.57 0.0284
Adult	80.89±1.24 0.0802	80.45±1.05 0.2427	80.33±0.65 0.5661	80.06±1.13 1.6855	81.11±1.06 3.7605	**81.36**±1.22 4.9880
BreastMNIST	76.44±2.47 0.1247	75.52±1.42 0.0067	78.64±1.91 0.0337	75.41±1.55 0.1613	78.72±2.32 0.3355	**81.78**±1.99 0.255
PneumoniaMNIST	94.38±0.99 0.7389	92.65±1.40 0.0953	93.70±0.75 0.4702	95.16±0.72 2.7011	95.53±0.42 2.6490	**95.89**±0.56 5.0186
Pendigits	88.43±1.84 0.0203	90.13±1.39 0.0979	93.78±2.05 0.5159	94.73±0.83 9.5677	94.93±0.90 10.0287	**95.30**±1.03 8.3801
DNA	94.15±0.68 0.5556	94.90±0.44 0.5635	90.19±1.29 1.5880	94.88±0.64 3.2264	94.86±0.64 3.4162	**95.13**±0.74 4.6764
DermaMNIST	67.53±0.97 6.7839	67.53±0.97 0.2768	67.53±0.97 2.3347	69.93±0.79 4.0855	69.80±0.99 4.4527	**70.58**±1.00 6.1879
USPS	95.28±0.66 0.2179	95.06±0.72 0.6141	95.94±0.73 8.6429	94.80±0.64 12.3894	95.15±0.51 13.2304	**96.50**±0.51 18.3854

TABLE IV
THE RESULTANT P-VALUE OF PARIED T-TEST. "*" INDICATES THAT $p < 0.05$ FOR ALL METHODS ON THIS DATASET.

Datasets	SVM	LSSVM	TBSVM	L_1-TWSVM	CTWSVM
Heart	0.111373	0.046202	0.016328	0.033793	0.004439
Mushroom	0.003241	0.003241	0.000959	0.081126	0.222868
German	0.036029	0.004471	0.282486	0.005975	0.004898
Sonar*	0.000695	0.000049	0.000084	0.002604	0.000375
Australian	0.001165	0.000162	0.193422	0.000023	0.890531
Breast	0.343436	0.258547	0.879343	0.060694	0.872288
Adult	0.039209	0.002816	0.023989	0.002109	0.334350
BreastMNIST*	0.000751	0.000031	0.000375	0.000017	0.004202
PneumoniaMNIST*	0.000142	0.000048	0.000006	0.013186	0.038407
DNA	0.000008	0.158919	4.78e-08	0.031948	0.068701
Pendigits	0.000013	0.000012	0.011162	0.136480	0.240969
DermaMNIST*	0.000001	0.000013	0.002162	0.001168	0.000007
USPS*	0.000083	0.000008	0.001120	0.000041	0.000164

TABLE V
THE AVERAGE CLASSIFICATION ACCURACY (%) AND STANDARD DEVIATION OF RMTBSVM BASED ON DEEP LEARNING

DataSets	CNN	ResNet	ViT	CNN+ours	ResNet+ours	ViT+ours	RMTBSVM
DermaMNIST	68.79±1.04	65.43±1.32	69.48±0.84	67.53±0.97	**74.78**±0.95	72.52±0.69	71.03±1.12
USPS	97.56±0.82	94.63±0.66	93.67±0.37	97.68±0.64	93.46±0.79	**98.57**±0.51	96.59±0.51

TABLE VI
THE AVERAGE TRAINING TIME (S), CLASSIFICATION ACCURACY (%), AND THE CORRESPONDING STANDARD DEVIATION ON NOISED DATASETS ($\tau = 0.2$).

DataSets	SVM	LSSVM	TBSVM	L_1-TWSVM	CTWSVM	RMTBSVM
Heart	81.29±2.18 0.0013	80.46±2.82 0.0032	82.41±1.94 0.0083	81.85±3.13 0.0486	81.48±3.31 0.0520	**83.80**±2.16 0.0152
Mushroom	99.75±0.17 0.1042	99.85±0.05 1.1493	**99.93**±0.05 3.2108	99.91±0.07 9.9200	99.91±0.06 18.2734	99.92±0.05 3.9047
German	71.20±1.39 0.0233	70.73±1.49 0.0346	72.13±1.75 0.1120	70.25±2.13 0.2196	72.53±2.15 2.8643	**73.20**±1.23 0.2087
Sonar	74.22±3.66 0.0024	72.17±6.27 0.0038	78.31±5.02 0.0093	78.67±4.44 0.0115	79.88±3.96 0.0112	**84.58**±4.87 0.0217
Australian	84.64±2.63 0.0052	84.96±1.56 0.0141	84.78±2.03 0.0416	81.59±1.67 0.2095	82.79±1.57 0.6369	**85.11**±1.76 0.0815
Breast	96.48±0.81 0.0033	96.37±0.95 0.0108	96.41±0.92 0.0265	95.53±1.05 0.2847	96.30±0.78 0.3867	**96.59**±0.97 0.0262
Adult	79.90±1.31 0.1744	79.94±1.44 0.0655	80.23±0.73 0.4997	78.54±1.07 3.5769	80.11±0.96 2.7945	**81.13**±1.10 4.7451
BreastMNIST	72.78±1.36 0.1532	73.09±1.26 0.0063	75.59±1.37 0.0319	74.84±1.91 0.1165	76.55±1.57 0.1064	**80.46**±1.79 0.2715
PneumoniaMNIST	94.35±0.81 1.1315	86.41±2.42 0.1362	90.18±1.76 0.4997	94.40±0.84 1.0214	93.86±0.94 1.3912	**94.50**±0.69 4.2485
Pendigits	85.55±1.93 0.0157	86.73±1.98 0.0719	89.45±1.27 0.2660	89.73±1.18 6.8596	**90.25**±1.53 8.6570	89.68±0.74 6.4048
DNA	92.81±0.74 0.2860	93.25±0.72 0.0952	93.06±0.85 0.5309	91.51±0.85 1.4957	93.11±0.99 1.4749	**93.50**±1.07 2.1751
DermaMNIST	67.53±0.97 5.7182	68.15±1.23 0.2997	69.69±0.84 2.6915	69.63±1.11 5.5031	69.91±1.15 5.3726	**71.03**±1.12 8.8269
USPS	92.10±0.59 0.2179	95.11±0.65 0.2579	95.33±0.73 1.8666	95.34±0.54 4.7147	94.71±0.66 4.9551	**95.76**±0.63 9.0862

better than SVM on these datasets, but RMTBSVM still obtain the best classification accuracy. CTWSVM performs slightly better than RMTBSVM on Australian and Breast datasets, However, notice that the p-values of CTWSVM on these two datasets are 0.8905 and 0.8723, which suggests that the accuracy of RMTBSVM and CTWSVM has no difference on Australian and Breas datasets statistically. Besides, we can see that most entries of Table IV are less than 0.05, and $p < 0.05$ always holds for all methods on Sonar, Adult, PneumoniaMNIST, USPS, and DermaMNIST datasets. Therefore, RMTBSVM significantly outperforms other SVM methods on these datasets.

In terms of time consumption, The proposed RMTBSVM costs similar time with CTWSVM, L1-TWSVM in most cases. It turns out that L_1-TWSVM, CTWSVM, and the proposed RMTBSVM cost more time than the conventional SVM. The main reason is that these methods use alternatively iterative strategies to compute the optimal solution.

C. Robustness analysis

To study the robustness of RMTBSVM, we introduced Gaussian noise into data and check the performance of different methods. Denote $\mathbf{X}_{\mathcal{N}}$ as the noised data matrix, which is defined as

$$\mathbf{X}_{\mathcal{N}} = \mathbf{X} + \tau \frac{\|\mathbf{X}\|_F}{\|\mathbf{M}\|_F}\mathbf{M},$$

where $\tau \in [0, 1]$ is the noise factor, and matrix $\mathbf{M}$ is a random noise matrix with $m_{ij} \sim \mathcal{N}(0, 1)$ [47]. In the experiments, $\tau \in \{0, 0.2, 0.4, 0.6\}$.

We first set $\tau = 0.2$ and record the performance of methods on noised data, The corresponding results are shown in Table VI. We can observe that the accuracies of most methods degrade due to noises. However, RMTBSVM still obtain the best accuracies on 11 out of 13 datasets. The accuracy of SVM, LSSVM and TBSVM obviously degrades on many datasets. By contrast, the performance of L_1-TWSVM, CTWSVM, and RMTBSVM are relatively stable and better than SVM, LSSVM, and TBSVM. This phenomenon demonstrates that the L_1-norm metric based methods are more robust against noise, and this also validates the robustness of our method.

We conducted another experiment that study the influence of the different noise factors. Fig. 4(a) and Fig. 4(b) illustrates the classification accuracies of six SVM methods with different noise factor τ. We can see that the proposed RMTBSVM always possesses the best accuracy regardless of the presence of noises. The performance of LSSVM and TBSVM suffers significant degradation as τ increases. By contrast,

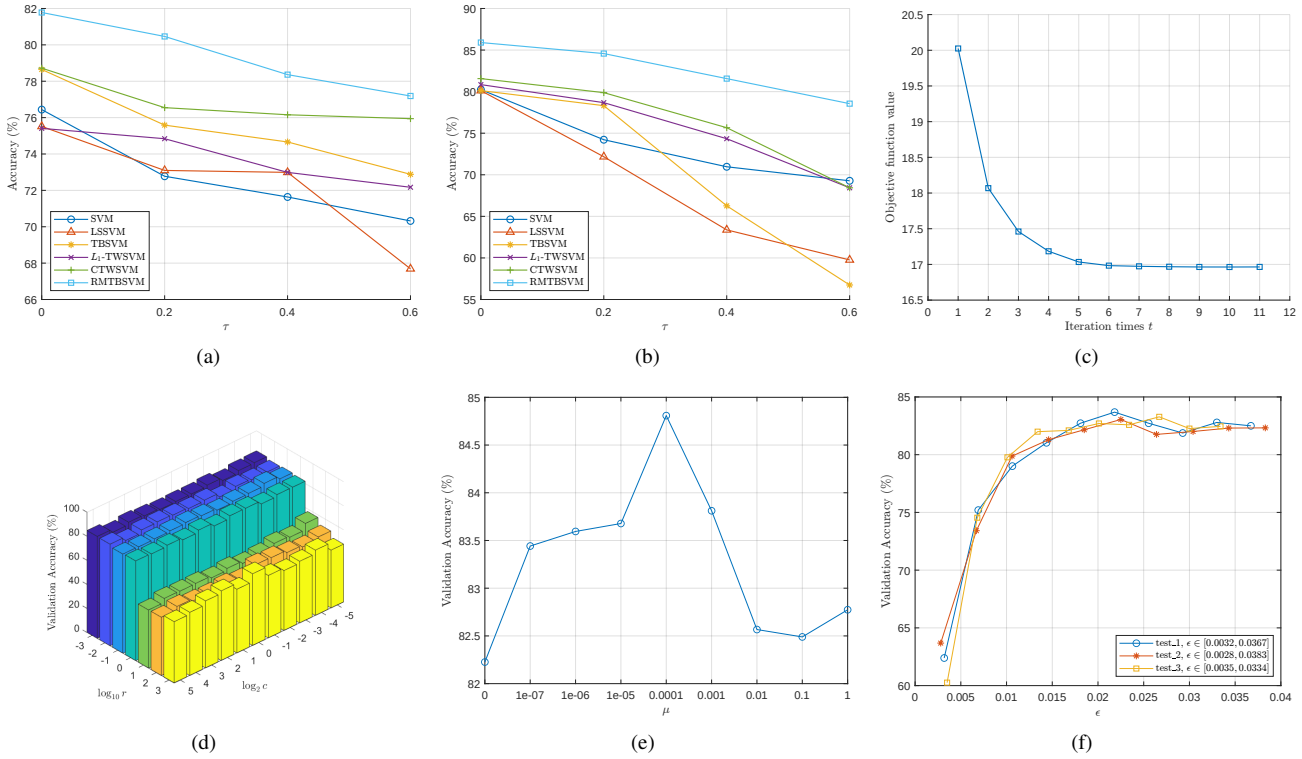

Fig. 4. (a), (b): the classification accuracy of each method in BreastMNIST dataset and Sonar dataset, respectively. (c): the convergence of RMTBSVM. (d): the classification accuracy v.s. parameters of c, r on Sonar dataset. (e): accuracy v.s. parameter μ on Sonar dataset. (f): accuracy v.s. parameter ϵ on Sonar dataset with different training/testing sets.

TABLE VII
THE AVERAGE CLASSIFICATION ACCURACY (%) AND STANDARD DEVIATION ON IMAGE DATASETS WITH NOISES

DataSets (noise type)	SVM	LSSVM	TBSVM	L_1-TWSVM	CTWSVM	RMTBSVM
BreastMNIST (Gaussian)	72.84±1.49	66.76±1.62	73.99±1.10	76.12±2.26	76.87±1.97	**78.90**±2.00
BreastMNIST (block)	71.71±1.98	64.59±3.46	72.38±1.53	73.56±1.01	75.59±1.29	**77.51**±1.89
BreastMNIST (impluse)	69.29±2.36	64.23±2.74	72.46±2.68	75.55±2.29	75.80±2.16	**78.01**±1.47
PneumoniaMNIST (Gaussian)	94.14±0.90	88.34±1.84	91.94±1.24	94.80±0.55	94.09±0.70	**94.66**±0.53
PneumoniaMNIST (block)	92.51±1.12	85.01±1.70	81.26±2.02	93.20±0.69	91.59±0.94	**94.11**±0.65
PneumoniaMNIST (impulse)	93.95±1.03	83.69±2.39	89.64±1.55	94.40±0.79	93.75±0.95	**94.54**±0.59

RMTBSVM, CTWSVM and L_1-TWSVM are more stable. However, CTWSVM and L_1-TWSVM are still not as good as our method. This experiment suggests once again that our proposed RMTBSVM possesses strong robustness and better classification performance.

We performed an extra experiment of robustness analysis on image data. We used different types of noise to corrupt the image data. The instances of the corrupted images from BreastMNIST are visualized in Fig. 3. For each dataset, the images are corrupted by random blocks with the size of 8×8, impulse noises with the density of 5%, and Gaussian noises of $\mathcal{N}(0, 0.05)$. The results of the experiments are presented in Table VII. We can observe that the performance of SVM, LSSVM is severely affected by noises, where the random blocks and impulse noises have larger impact. However, L_1-TWSVM, CTWSVM and RMTBSVM still retain high accuracy in spite of the presence of different noises. Obviously, our proposed RMTBSVM still achieves the highest accuracy on these datasets. Therefore, the experimental results further suggest that RMTBSVM has better insensitivity to different types of noise.

D. Convergence analysis

We analyze the convergence of RMTBSVM theoretically in section IV-A. To validate it, we conducted the experiment on Heart dataset. Fig. 4(c) illustrates the objective value in each iteration step. The figure demonstrates the objective value monotonically decreases in each iteration, which is consistent with our previous analysis. In addition, the algorithm converges in few steps (always less than 10 steps), which indicates that the proposed algorithms converge in limited iteration steps.

E. Parameters study

We report the accuracy of our method versus the hyperparameters. Fig. 4(d) illustrates the grid search result of param-

eters c and r on Sonar dataset. It is clear that the parameter r obviously influences the performance of RMTBSVM. The performance is degraded significantly when r is large. It turns out that the ideal value of r usually satisfies $r \leq 1$. By contrast, the accuracy is relatively robust to c in the searching interval. In general, the performance is best when $c \in [2^{-4}, 1]$ and $r \in [10^{-3}, 1]$.

Fig. 4(e) shows the validation accuracy versus the manifold regularization parameter μ. One can find that r influences the performance to a certain extent. The accuracy is relatively low when μ is too large or too small. Especially, when $\mu = 0$, the validation accuracy is lowest as shown in Fig. 4(e). That is, the performance is worse if we eliminate the manifold regularization term. This phenomenon suggests that the manifold regularization in RMTBSVM enhances the discriminative ability and further improves performance. In the experiments, we find that the potentially optimal interval of μ is $[10^{-7}, 10^{-3}]$ approximately.

Finally, ϵ is also a key hyperparameter in RMTBSVM. Since the optimal range of ϵ is dependent on the training data, we used a different way to determine the optimal ϵ (see section V-A). Fig. 4(f) illustrates the accuracy versus ϵ on different training/testing sets. Each curve represents one partitioning of training and testing dataset. It is clear that the curves share a similar pattern. The accuracy is low when ϵ is small. As ϵ increases, the accuracy grows up to a certain extent, but declines slightly when ϵ achieves the largest value.= The best values of ϵ corresponding to the three tests are 0.0218, 0.0225, and 0.0267, respectively, which have no significant difference. Therefore, the selection of ϵ is insensitive regarding the training/testing sets.

F. Experiment with deep learning

In this section, we conducted a brief experiment on DermaMNIST and USPS datasets to study the joint learning of deep neural networks and RMTBSVM. Since the proposed RMTBSVM is not a feature extraction method, it is an interesting problem that whether RMTBSVM performs better than the feature extraction methods. Therefore, we considered comparing RMTBSVM with deep learning methods. We also used the output features (i.e., the input of the last fully connected layer in the classification neural networks) as the input of RMTBSVM to check the performance. The experimental results are shown in Table V, where "CNN+ours" denotes RMTBSVM using features extracted by CNN, and the others are similar. We find some interesting phenomena in the experiments.

First, RMTBSVM (using the original features) obtains better performance than deep neural networks on DermaMNIST dataset. We observe that RMTBSVM achieves the accuracy of 71.03%, while the best result of deep learning method is 69.48%, obtained by ViT. This suggests the proposed RMTBSVM can perform better classification accuracy than deep learning methods, which further validate the effectiveness of the proposed method.

Another observation is that RMTBSVM improves the performance of deep learning methods when the deep neural network features are used. The accuracy of ResNet is 65.43% on DermaMNIST dataset, which is poor compared with other methods. However, RMTBSVM obtains the accuracy of 74.48% when using the features from ResNet, and achieves the best result on DermaMNIST dataset. The proposed RMTBSVM fails to achieve better accuracy than CNN on USPS dataset, but RMTBSVM still obtains best accuracy on USPS using the features extracted from ViT. These phenomena demonstrate that RMTBSVM can further improve the performance of the neural networks via the features extracted by the networks. This also verifies the good classification performance of RMTBSVM.

VI. Conclusion

In this paper, we propose a novel robust model under support vector machine framework, namely robust manifold twin bounded support vector machine (RMTBSVM). Capped L_1-norm is used as the robust distance metric to reduce the impact of outliers. The manifold regularization, which imposes the model to exploit the geometric structure of data, is integrated into the proposed model to enhance the discriminability. Besides, since the existing methods such as L_1-TWSVM and CTWSVM have not discussed nonlinear classification, we further generalize our model for nonlinear classification by kernel method. The algorithms for both linear and nonlinear cases are presented, and the convergence of the algorithms is then analyzed. The experimental results show that the proposed RMTBSVM outperforms the other types of SVM in not only classification accuracy but also the robustness against noises.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[3] J. Liu, J. Yang, Y. Zhang, and X. He, "Action recognition by multiple features and hyper-sphere multi-class svm," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3744–3747.

[4] A. Zendehboudi, M. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: A review," *Journal of Cleaner Production*, vol. 199, pp. 272–285, 2018.

[5] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

[6] Y. Tian, Z. Qi, X. Ju, Y. Shi, and X. Liu, "Nonparallel support vector machines for pattern classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1067–1079, 2014.

[7] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft, Tech. Rep. MSR-TR-98-14, April 1998.

[8] O. L. Mangasarian and D. R. Musicant, "Successive overrelaxation for support vector machines," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1032–1037, 1999.

[9] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[10] O. L. Mangasarian and D. R. Musicant, "Lagrangian support vector machines," *Journal of Machine Learning Research*, vol. 1, no. Mar, pp. 161–177, 2001.

[11] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[12] Y.-H. Shao, C.-H. Zhang, X.-B. Wang, and N.-Y. Deng, "Improvements on twin support vector machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 962–968, 2011.

[13] M. A. Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7535–7543, 2009.
[14] D. Huang, R. Cabral, and F. D. l. Torre, "Robust regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 363–375, 2016.
[15] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with l1-norm," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 828–842, 2014.
[16] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1813–1821, 2010.
[17] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, "Robust discriminant regression for feature extraction," *IEEE Transactions on Cybernetics*, vol. 48, no. 8, pp. 2472–2484, 2018.
[18] Y. Xu, Z. Yang, and X. Pan, "A novel twin support-vector machine with pinball loss," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 359–370, 2017.
[19] H. Yan, Q.-L. Ye, and D.-J. Yu, "Efficient and robust twsvm classification via a minimum l1-norm distance metric criterion," *Machine Learning*, vol. 108, no. 6, pp. 993–1018, 2019.
[20] H. Yan, Q. Ye, T. Zhang, D.-J. Yu, X. Yuan, Y. Xu, and L. Fu, "Least squares twin bounded support vector machines based on l1-norm distance metric for classification," *Pattern Recognition*, vol. 74, pp. 434–447, 2018.
[21] C. Wang, Q. Ye, P. Luo, N. Ye, and L. Fu, "Robust capped l1-norm twin support vector machine," *Neural Networks*, vol. 114, pp. 47–59, 2019.
[22] J. Ma, L. Yang, and Q. Sun, "Capped l1-norm distance metric-based fast robust twin bounded support vector machine," *Neurocomputing*, vol. 412, pp. 295–311, 2020.
[23] Y. Li, H. Sun, W. Yan, and Q. Cui, "R-CTSVM+: Robust capped L1-norm twin support vector machine with privileged information," *Information Sciences*, vol. 574, pp. 12–32, 2021.
[24] C. Yuan, L. Yang, and P. Sun, "Correntropy-based metric for robust twin support vector machine," *Information Sciences*, vol. 545, pp. 82–101, 2021.
[25] J. Ma, L. Yang, and Q. Sun, "Adaptive robust learning framework for twin support vector machine classification," *Knowledge-Based Systems*, vol. 211, p. 106536, 2021.
[26] L. Rossi, A. Torsello, and E. R. Hancock, "Unfolding kernel embeddings of graphs: Enhancing class separation through manifold learning," *Pattern Recognition*, vol. 48, no. 11, pp. 3357–3370, 2015.
[27] L. Ladicky and P. H. Torr, "Locally linear support vector machines," in *ICML*, 2011.
[28] S. Sun and X. Xie, "Semisupervised support vector machines with tangent space intrinsic manifold regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 9, pp. 1827–1839, 2016.
[29] X. Xie and S. Sun, "General multi-view semi-supervised least squares support vector machines with multi-manifold regularization," *Information Fusion*, vol. 62, pp. 63–72, 2020.
[30] R. Khemchandani, S. Chandra *et al.*, "TWSVM for unsupervised and semi-supervised learning," in *Twin Support Vector Machines*. Springer, 2017, pp. 125–152.
[31] Z. Qi, Y. Tian, and Y. Shi, "Laplacian twin support vector machine for semi-supervised classification," *Neural Networks*, vol. 35, pp. 46–53, 2012.
[32] J. Xie, K. Hone, W. Xie, X. Gao, Y. Shi, and X. Liu, "Extending twin support vector machine classifier for multi-category classification problems," *Intelligent Data Analysis*, vol. 17, no. 4, pp. 649–664, 2013.
[33] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, no. Nov, pp. 2399–2434, 2006.
[34] L. Zhang, M. Luo, Z. Li, F. Nie, H. Zhang, J. Liu, and Q. Zheng, "Large-scale robust semisupervised classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 907–917, 2019.
[35] P. Gong, J. Ye, and C. Zhang, "Multi-stage multi-task feature learning," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2979–3010, 2013.
[36] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 709–722, 2013.
[37] Xiaofei He, Shuicheng Yan, Yuxiao Hu, P. Niyogi, and Hong-Jiang Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.
[38] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," *arXiv e-prints*, pp. arXiv–2010, 2020.
[39] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.
[40] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
[41] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[43] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. P. Vandewalle, *Least squares support vector machines*. World Scientific, 2002.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[47] H. Wang, F. Nie, and H. Huang, "Robust distance metric learning via simultaneous L_1-norm minimization and maximization," in *International Conference on Machine Learning*, 2014, pp. 1836–1844.

Junhong Zhang is now pursuing the B.S degree in Shenzhen University, Shenzhen. He is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060. His research interests include machine learning and pattern recognition.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, M.S. degree from Jinan University, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a Research Associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. He has published over 150 scientific articles. Now he is an associate editor of International Journal of Machine Learning and Cybernetics. For more information including all papers and related codes, the readers are referred to the website (http://www.scholat.com/laizhihui).

Heng Kong received the M.D. and B.S. degree from Chongqing Medical University, M.S. degree from Guangzhou Medical University, and PhD. degree from Southern Medical University, China, in 2000, 2005 and 2008, respectively. She works as a visiting scholar in Cancer Center of Georgia Reagent University at Augusta in USA in 2014-2016. She is a professor and director in department of thyroid and breast surgery, BaoAn Central Hospital of Shenzhen (the fifth affiliated Hospital of Shenzhen University), Guangdong province. She is also doing basic and clinic research associated breast and thyroid cancer. Her research interests include gene therapy, immunotherapy, early diagnosis and prognosis analysis of breast cancer, and tumor image processing and recognition using machine learning and artificial intelligent methods.

Linlin Shen received the B.Sc. degree from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 2005. He was a Research Fellow with Medical School, University of Nottingham, researching brain image processing of magnetic resonance imaging. He is currently a Professor and a Director with the Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His current research interests include Gabor wavelets, face/palmprint recognition, medical image processing, and hyperspectral image classification.