Received August 23, 2017, accepted September 25, 2017, date of publication October 11, 2017,
date of current version November 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2761125

L2P-Norm Distance Twin Support Vector Machine

XU MA, QIAOLIN YE, AND HE YAN
College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
Laboratory for Internet of Things and Mobile Internet Technology of Jiangsu Province, Huaiyin Institute of Technology, Nanjing 223003, China
Jangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing 210094, China

Corresponding author: Qiaolin Ye (yqlcom@njfu.edu.cn)

The work was supported in part by the National Science Foundation of China under Grant 61401214, Grant 61603184, Grant 61773210, and Grant 31670554, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20171453 and Grant BK20161527, in part by the Jiangsu Key Laboratory for Internet of Things and Mobile Internet Technology, and in part by the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety.

ABSTRACT A twin support vector machine (TWSVM) is an effective classifier, especially for binary data, which is defined by squared l_2-norm distance in the objective function. Since squared l_2-norm distance is susceptible to outliers, it is desirable to develop a revised TWSVM. In this paper, a new robust TWSVM via $l_{2,p}$-norm formulations was proposed, because it suppress the influence of outliers better than l_1-norm or squared l_2-norm minimizations. However, the resulted objective is challenging to solve, because it is non-smooth and non-convex. As an important work, we systematically derive an efficient iterative algorithm to minimize the pth order of l_2-norm distances. Theoretical support shows that the iterative algorithm is effective in the resolution to improve TWSVM via $l_{2,p}$-norm instead of squared l_2-norm distances. A large number of experiments show that $l_{2,p}$-norm distances twin support vector machine can treat the noise data effectively and has a better accuracy.

INDEX TERMS TWSVM, L2P-norm, robustness.

I. INTRODUCTION

In many applications in data mining and pattern recognition, support vector machine (SVM) [1]–[3] has been a vital method for classification over the past decade. It has been successfully applied to a broad range of fields [4]–[10]. For the standard SVM, the main idea is to get an optimal separating hyper plane to maximize the margin between the two types of data sets [3], [11], [12]. An advantage of SVM is regulating the trade-off between structural complexity and empirical risk.

However, SVM may not constantly satisfy real world applications. Due to the need to solve the quadratic programming problems (QPPs) [13], the computational complexity would be a problem. Also, SVM is not acceptable while dealing with some special datasets, such as binary datasets [14]–[17], imbalanced datasets [18]–[21], SVM is not acceptable. As a result, SVM derived a set of variants.

In 2001, G. Fungand et al. proposed an algorithm PSVM [22], two parallel planes are pushed apart as far as possible to classify points. Instead of figuring out a quadratic or a linear equation, PSVM only need to solve a single system of linear equations. The formulation of PSVM makes the solution of SVM fast and effectual. In 2006, O. L. Mangasarian and E. W. Wild proposed a nonparallel plane classifier via generalized eigenvalue (GEPSVM) [23]. The novel approach for classification problems cuts down the requisite that the bounding generated by SVMs is parallel in the input space.

Different from PSVM and GEPSVM, a new nonparallel plane classifier Twin Support Vector Machine (TWSVM) was proposed by Jayadeva in 2007 [16]. It solves a pair of quadratic programming problems. Each of the two quadratic programming problems has the expression of a typical SVM, except that not all data points are used in the constraints of either problem at the same time.

Nevertheless, the above related works are based on squared l_2-norm distance metric, which is sensitive to outliers. To develop a robust method, l_1-norm distance metric has been introduced in many papers [24]–[29]. The l_1-norm is more robust and sparse than l_2-norm and it is the optimal convex approximation of the l_0-norm. l_1-norm is more suitable to optimize than l_0-norm because l_0-norm is an NP-hard optimization problem [30], [31]. Particularly, Newton Method for Linear Programming Twin Support Vector Machines (NLPTSVM) [38]–[40] is one such extension. NLPTSVM uses l_1-norm to improve the robustness and obtains the solution by Newton–Armijo algorithm. Instead of solving a pair of quadratic programming problems, NLPTSVM solves just two linear equations. Consequently, this algorithm is not only robust, but also simple and fast.

Extensive studies have shown that l_1-norm minimizations and not-squared l_2-norm ($l_{2,p}-$norm, $0 < p \leq 2$) minimizations can provide robustness for the objectives. The $l_{2,p}-$norm can better tolerate the biases caused by the outlying data, especially the outliers are far from the normal data. Also, not only l_1-norm, $l_{2,p}-$norm also has a better sparsity than squared l_2-norm [34]. This will allow the algorithm to have fewer support vector points. Thus, many researches have improved the various models through $l_{2,p}-$norm distances [33], [34]. Inspired by the above, we are immersed in the problem of robust TWSVM on data set with outlier data samples in this paper. In classical TWSVM, it is willing to minimize the distance with the squared l_2-norm distance. As we know, squared l_2-norm distance can expand the error distance of samples.

Based on this recognition, we proposed a new TWSVM objective based on $l_{2,p}-$norm distance, termed as pTWSVM, as a robust SVM classifier. The new method solves a pair of quadratic programming problems, both of which are formulated using the $l_{2,p}-$norm. It is interesting from a number of perspectives as follows:

1) The resulted objective is based on the $p-$ th order of the l_2-norm distance, which is more comprehensive than conventional TWSVM. The conventional TWSVM is a special case of this new method when $p = 2$.

2) Compared with TWSVM, our new method is more robust against outlier data samples. It provides a robust alternative to TWSVM.

3) The resulted objective of the proposed pTWSVM is difficult for us to solve, because the formulation is non-smooth and non-convex. To solve this problem, we present an efficient and simple iterative algorithm, which is proved to find a local optimal solution.

4) Extensive empirical evaluations demonstrate the new method outperforms related state-of-art methods on various data sets.

The remaining content of this paper is organized as follows.

In Section II, we briefly introduce the related work. In Section III, we dwell on our theoretical work for the new method in detail, including the improvement and related proof. Section IV is about the extension of nonlinear kernel. The experimental results are presented in Section V. Finally, Section VI summarizes this paper.

II. RELATED WORK

In this paper, the vectors are all column vectors. A row vector will be defined by transposing a column vector via a prime superscript T. Suppose there are m data points belongs to two classes. Let the positive class patterns to be denoted by a set of m1 row vectors A_i $(i = 1, 2 \ldots m1)$ in the n-dimensional real space R^n. Also, the negative class patterns are denoted by row vectors B_i $(i = 1, 2 \ldots m2)$. The n denotes the dimension of the data. For any matrix M, the $i-th$ row, $j-th$ column are denoted by m^i, m_j respectively. The squared l_2-norm of this matrix is defined as :

$$\|M\|_2 = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} \left\| m^i \right\|_2^1$$

Therefore, the widely used squared l_2-norm can be defined as:

$$\|M\|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} m_{ij}^2 = \sum_{i=1}^{n} \left\| m^i \right\|_2^2$$

The l_2-norm can be generalized to the $p-$th order l_2-norm ($l_{2,p}-$norm)

$$\|\mathrm{M}\|_2^{\mathrm{p}} = \sum_{i=1}^{n} \sum_{j=1}^{m} m_{ij}^p = \sum_{i=1}^{n} \left\| m^i \right\|_2^p$$

However, when $0 < p < 1$, the $l_{2,p}-$norm is not a valid matrix norm because it does not satisfy the triangle inequality for norms. Still, we call it norms for convenience in this paper.

In addition, $e1$ and $e2$ are vectors of ones of appropriate dimension. I denotes the identity matrix of arbitrary dimension.

A. GEPSVM

The proximal support vector machine via generalized eigenvalues(GEPSVM) [23] is a great classifier for binary classification problems, which is based on the squared $l2-$norm distance metric. The goal of GEPSVM classifier is to obtain two nonparallel planes

$$\mathrm{xw}^1 + \mathrm{b}^1 = 0 \text{ and } \mathrm{xw}^2 + \mathrm{b}^2 = 0. \tag{1}$$

so as to minimize the Euclidean distance of the planes from two type data points respectively. In order to minimize the Euclidean distance for each plane, we need to solve the following optimization problem:

$$\min_{\mathrm{w1,b1}\neq 0} \frac{\left\|\mathrm{Aw}^1 + \mathrm{e}^1\mathrm{b1}\right\|^2}{\left\|\mathrm{Bw}^1 + \mathrm{e}^2\mathrm{b1}\right\|^2} \tag{2}$$

where $\mathrm{w}^1, \mathrm{b1} \neq 0$ and the $\|\cdot\|$ denotes the l_2-norm. To get a stable solution and to avoid singular, we always introduce a regularization term. This formulation can be converted to Rayleigh Quotient problem as follows:

$$\min_{z\neq 0} \frac{\mathrm{z^T Gz}}{\mathrm{z^T Hz}} \tag{3}$$

where G and H are symmetric matrices in real space $\mathrm{R}^{(n+1)\times(n+1)}$ and z presents the classification plane. The G, H and z are defined as:

$$\begin{aligned} G &= [Ae]^T \times [Ae] \\ H &= [Be]^T \times [Be] \\ z &= [w1, b1]^T \end{aligned} \tag{4}$$

The solution of formulation (3) is obtained by solving the generalized eigenvalue problem via the properties of Rayleigh Quotient [23], [35].

$$\mathrm{Gz} = \mu \mathrm{Hz}, \quad \mathrm{z} \neq 0. \tag{5}$$

It's easy to get the minimum objective value of formulation (2) when z is the value of the eigenvector corresponding to the smallest eigenvalue μ. Therefore, we can obtain the plane $\mathrm{xw}^1 + \mathrm{b}^1 = 0$, which is close to patterns of positive class and far away from patterns of negative class. And vice versa, we can get another one by the same method.

B. TWSVM

Suppose we have data points of n-dimensional belonging to two classes represented by matrices A and B respectively. TWSVM [16] devotes to obtaining two nonparallel hyper planes, each plane is as close as possible to one type points and as far as possible to the rest.

The result of TWSVM can be obtained by solving the following pairs of quadratic programming problems such that:

$$\begin{aligned} &\min_{\mathrm{w}^1,\mathrm{b}^1,\mathrm{q}} \ \frac{1}{2}\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|^2 + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ &\text{subjectto} \ -\left(\mathrm{Bw}^1 + \mathrm{e}_2\mathrm{b}^1\right) + \mathrm{q} \geq \mathrm{e}_2, \quad \mathrm{q} \geq 0 \end{aligned} \tag{6}$$

$$\begin{aligned} &\min_{\mathrm{w}^2,\mathrm{b}^2,\mathrm{q}} \ \frac{1}{2}\left\|\mathrm{Bw}^2 + \mathrm{e}_2\mathrm{b}^2\right\|^2 + \mathrm{c}^2\mathrm{e}_1^{\mathrm{T}}\mathrm{q} \\ &\text{subjectto} \ -\left(\mathrm{Aw}^2 + \mathrm{e}_1\mathrm{b}^2\right) + \mathrm{q} \geq \mathrm{e}_1, \quad \mathrm{q} \geq 0 \end{aligned} \tag{7}$$

where $\mathrm{c}^1, \mathrm{c}^2 > 0$ are parameters. The two nonparallel planes are obtained by $\mathrm{w}^1, \mathrm{b}^1, \mathrm{w}^2, \mathrm{b}^2$:

$$\mathrm{X}^{\mathrm{T}}\mathrm{w}^1 + \mathrm{b}^1 = 0 \quad \text{and} \ \mathrm{X}^{\mathrm{T}}\mathrm{w}^2 + \mathrm{b}^2 = 0 \tag{8}$$

We can classify a new data point by comparing its geometrical margin to the two planes respectively.

III. P-ORDER TWIN SUPPORT VECTOR MACHINE

A. OPTIMIZATION ALGORITHM TO THE PROPOSED METHOD

It clearly demonstrates that the squared l_2-norm distance in the formulation of TWSVM. However, the squared l_2-norm distance may be not satisfied for the problem. The result we obtained could be affected by the outliers pronouncedly. That is, $p - th$ order l_2-norm is a proficient method to replace squared l_2-norm distance. If we can find appropriate value of p, the algorithm will emphasize normal data points and repress outlier data points best. Assuming the squared distance is a benchmark, if $\mathrm{p} < 2$, the distance between data points will be shortened and the influence of outlier data samples will be alleviated. The paper holds the notion that the best value of p is determined by the percentage of outlier data points. In fact, smaller p value means sparser representation, and using $\mathrm{l}_{2,\mathrm{p}}-$norm can find sparse solutions than the widely used squared l_2-norm [33]. The experimental results in [42] have demonstrated that $\mathrm{l}_{2,\mathrm{p}}-$norm does obtain sparser solutions than squared l_2-norm and $\mathrm{l}_{2,1}-$norm.

Thus, the exact formulation of $p - th$ order l_2-norm TWSVM is

$$\begin{aligned} &\min_{\mathrm{w}^1,\mathrm{b}^1,\mathrm{q}} \ \frac{1}{2}\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|^{\mathrm{p}} + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ &\text{subject to} \ -\left(\mathrm{Bw}^1 + \mathrm{e}_2\mathrm{b}^1\right) + \mathrm{q} \geq \mathrm{e}_2, \quad \mathrm{q} \geq 0 \end{aligned} \tag{9}$$

$$\begin{aligned} &\min_{\mathrm{w}^2,\mathrm{b}^2,\mathrm{q}} \ \frac{1}{2}\left\|\mathrm{Bw}^2 + \mathrm{e}_2\mathrm{b}^2\right\|^{\mathrm{p}} + \mathrm{c}^2\mathrm{e}_1^{\mathrm{T}}\mathrm{q} \\ &\text{subject to} \ -\left(\mathrm{Aw}^2 + \mathrm{e}_1\mathrm{b}^2\right) + \mathrm{q} \geq \mathrm{e}_1, \quad \mathrm{q} \geq 0 \end{aligned} \tag{10}$$

The Lagrange corresponding to formulation (9) is given by:

$$\begin{aligned} \mathcal{L}\left(\mathrm{w}^1, \mathrm{b}^1, \mathrm{q}, \alpha, \beta\right) = \frac{1}{2}\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}} + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ + \alpha^{\mathrm{T}}\left[\mathrm{Bw}^1 + \mathrm{e}_2\mathrm{b}^1 - \mathrm{q} + \mathrm{e}_2\right] - \beta^{\mathrm{T}}\mathrm{q} \end{aligned} \tag{11}$$

where α and β are the vectors of Lagrange multipliers.

Obviously, the formulation (11) involves $\mathrm{l}_{2,\mathrm{p}}-$norm regularization. Hence it is hard to derive the solution directly. To address this issue, we make a good idea, which is splitting the distance $\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}}$ to squared and $(p-2)th$ order:

$$\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}} = \left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}-2}\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^2 \tag{12}$$

We use $\mathcal{S}$ to denote the $(p-2)$ *th* order term such that

$$\mathcal{S} = \left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}-2}. \tag{13}$$

The corresponding Lagrange function can be written as:

$$\begin{aligned} \mathcal{L}\left(\mathrm{w}^1, \mathrm{b}^1, \mathrm{q}, \alpha, \beta\right) = \frac{1}{2}\mathcal{S}\left\|\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^2 + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ + \alpha^{\mathrm{T}}\left[\mathrm{Bw}^1 + \mathrm{e}_2\mathrm{b}^1 - \mathrm{q} + \mathrm{e}_2\right] - \beta^{\mathrm{T}}\mathrm{q} \end{aligned} \tag{14}$$

The Karush-Kuhn-Tucker (K.K.T) [13] necessary and sufficient optimality conditions for the problem are given by

$$\frac{\partial \mathcal{L}}{\partial \mathrm{w}^1} = \mathcal{S}\mathrm{A}^{\mathrm{T}}\left(\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right) + \mathrm{B}^{\mathrm{T}}\alpha = 0 \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial \mathrm{b}^1} = \mathcal{S}\mathrm{e}_1^{\mathrm{T}}\left(\mathrm{Aw}^1 + \mathrm{e}_1\mathrm{b}^1\right) + \mathrm{e}_2^{\mathrm{T}}\alpha = 0 \tag{16}$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{q}} = \mathrm{c}^1\mathrm{e}_2 - \alpha - \beta = 0 \tag{17}$$

$$\frac{\partial \mathrm{L}}{\partial \alpha} = -\left(\mathrm{Bw}^1 + \mathrm{e}_2\mathrm{b}^1\right) + \mathrm{q} - \mathrm{e}_2 \geq 0 \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \mathrm{q} \geq 0 \tag{19}$$

$$\alpha^{\mathrm{T}}\left(\mathrm{Bw}^1 + \mathrm{e}_2\mathrm{b}^1 - \mathrm{q} + \mathrm{e}_2\right) = 0 \tag{20}$$

$$\beta^{\mathrm{T}}\mathrm{q} = 0 \tag{21}$$

$$\alpha \geq 0, \quad \beta \geq 0 \tag{22}$$

According to $\alpha \geq 0$, $\beta \geq 0$, $c_1 e_2 - \alpha - \beta = 0$, we have

$$0 \leq \alpha \leq c^1 \tag{23}$$

We define

$$H = [Ae_1], \quad G = [Be_2], \quad u = [w^1, b^1]^T \tag{24}$$

With these notations, (13) may be rewritten as

$$\mathcal{S} = \|Hu\|_2^{p-2}. \tag{25}$$

Combining (15) and (16) leads to

$$\mathcal{S}\left[A^T \; e_1^T\right][A \; e_1]\left[w^1, b^1\right]^T + \left[B^T \; e_2^T\right]\alpha = 0. \tag{26}$$

For convenience, we can rewrite (26) as following:

$$\mathcal{S}H^T Hu + G^T\alpha = 0, \tag{27}$$

i.e.,

$$u = -\left(\frac{1}{\mathcal{S}}H^T H\right)^{-1} G\alpha. \tag{28}$$

Although $\mathcal{S}H^T H$ is always positive semi-definite, it is possible that it may not be well conditioned in some situations. So the problem can be regularized by introducing a regularization term as follows:

$$u = -\left(\frac{1}{\mathcal{S}}H^T H + \varepsilon I\right)^{-1} G\alpha, \tag{29}$$

where $\varepsilon > 0$ and I is an identity matrix of appropriate dimensions.

Using the Lagrange function and the K.K.T. conditions above, we obtain the Wolfe dual of $l_{2,p}-$norm TWSVM as follows:

$$\begin{aligned} \min_{\alpha} \; & \frac{1}{2}\alpha^T G\left(\frac{1}{\mathcal{S}}H^T H\right)^{-1} G^T\alpha - e_2^T\alpha \\ & 0 \leq \alpha \leq c^1 \end{aligned} \tag{30}$$

Similarly, another dual is obtained as:

$$\begin{aligned} \min_{\gamma} \; & \frac{1}{2}\gamma^T P\left(\frac{1}{\mathcal{S}}Q^T Q\right)^{-1} P^T\gamma - e_1^T\gamma \\ & 0 \leq \gamma \leq c^2 \end{aligned} \tag{31}$$

We can obtain the optimal u via an iterative algorithm. $\mathcal{S}$ is calculated with the current calculated α. The iteration is started with an initialized u and repeated until the objective converges. The u and $\mathcal{S}$ are re-changed adaptively during each iteration.

Anther one is similarly like the process above.

Before performing the experiments, we first analyze the difference between our new algorithm and TWSVM and NLPTSVM. It is clear to see that when $p = 2$, pTWSVM is equivalent to TWSVM. That is, TWSVM is actually a special case of pTWSVM. When $p = 1$, pTWSVM is TWSVM based on l_1-norm. However, it is not the same as Newton Method for Linear Programming Twin Support Vector Machines (NLPTSVM) because of the objectives.

Algorithm 1 The Algorithm to Solve the Problem

Input : Training data $A \in R^{m1\times n}$, $B \in R^{m2\times n}$, parameter p, $c1$, $c2$.
Give out $H \in R^{m1\times(n+1)}$ $G \in R^{m2\times(n+1)}$ $I \in R^{(n+1)\times(n+1)}$.
Initialize $u \in R^{(n+1)\times 1}$.
Until objective converges, do
1. Calculate $\mathcal{S} = \|Hu^T\|_2^{p-2}$;
2. Calculate α via dual function;
3. Update u, add regularization term if necessary;
End
Output $u \in R^{(n+1)\times 1}$.

NLPTSVM introduces the regularization terms in the objective to alleviate over-fitting problem. Besides, although both pTWSVM and NLPTSVM are iterative algorithms, their processes are different. Different from NLPTSVM, whose solution is obtained by solving a pair of dual exterior penalty problems as unconstrained minimization problems using Newton–Armijo algorithm, pTWSVM solves a pair of quadratic programming problems.

B. CONVERGENCE ANALYSIS

To prove the convergence of the new algorithm, we need the following useful lemma which has been proved in [34]:

Lemma 1: For any nonzero vectors u, v, when $0 < p \leq 2$, the following inequality holds:

$$\|u\|_2^p - \frac{p}{2}\|v\|_2^{p-2}\|u\|_2^2 \leq \|v\|_2^p - \frac{p}{2}\|v\|_2^{p-2}\|u\|_2^2 \tag{32}$$

Theorem 1: This algorithm can monotonically decrease the objective of the problem (9) in each iteration and make the objective function value converge to a local optimum.

Proof: Recall the formulation of our new method

$$\begin{aligned} J(u) = \min_{w^1,b^1,q} \; & \frac{1}{2}\|Hu\|^p + c^1 e_2^T q \\ \text{subject to } & -(Gu) + q \geq e_2, \quad q \geq 0. \end{aligned} \tag{33}$$

Formulation (33) is equivalent to formulation (9) and let J represent the objective value. Suppose $\tilde{u}$ is the solution of the $(t+1)$th iteration of the algorithm:

$$\begin{aligned} \tilde{u} &= \arg\min_{w^1,b^1,q} \frac{1}{2}\|Hu\|^p + c^1 e_2^T q \\ &= \arg\min_{w^1,b^1,q} \frac{1}{2}\mathcal{S}\|Hu\|^2 + c^1 e_2^T q. \end{aligned}$$

Note that $\mathcal{S} = \|Hu\|_2^{p-2}$, so we have

$$\begin{aligned} & \frac{1}{2}\|Hu\|_2^{p-2}\|H\tilde{u}\|^2 + c^1 e_2^T q \leq \frac{1}{2}\|Hu\|_2^{p-2}\|Hu\|^2 + c^1 e_2^T q \\ \Longrightarrow \; & \frac{1}{2}\|Hu\|_2^{p-2}\|H\tilde{u}\|^2 \leq \frac{1}{2}\|Hu\|_2^{p-2}\|Hu\|^2 \\ \Longrightarrow \; & \frac{p}{2}\|Hu\|_2^{p-2}\|H\tilde{u}\|^2 \leq \frac{p}{2}\|Hu\|_2^{p-2}\|Hu\|^2 \end{aligned} \tag{34}$$

According to Lemma 1, we have

$$\|H\tilde{u}\|_2^p - \frac{p}{2}\|Hu\|_2^{p-2}\|H\tilde{u}\|_2^2 \leq \|Hu\|_2^p - \frac{p}{2}\|Hu\|_2^{p-2}\|u\|_2^2. \tag{35}$$

Combining (34) and (35) leads to

$$\begin{aligned} &\|\mathrm{H}\tilde{\mathrm{u}}\|_2^{\mathrm{p}} \le \|\mathrm{Hu}\|_2^{\mathrm{p}} \\ \Longrightarrow\ &\frac{1}{2}\|\mathrm{H}\tilde{\mathrm{u}}\|_2^{\mathrm{p}} + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \le \frac{1}{2}\|\mathrm{Hu}\|_2^{\mathrm{p}} + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ \Longrightarrow\ &\mathrm{J}(\tilde{\mathrm{u}}) \le \mathrm{J}(\mathrm{u}). \end{aligned} \tag{36}$$

Thus, in each iteration, the algorithm decrements the objective function monotonically until the algorithm converges.

C. TIME COMPLEXITY

To optimize the objective function of pTWSVM, the most time consuming operation is to solve the pair of quadratic programming problems. In general, each QPP of the PTWSVM focuses on only about half of the data compared to traditional support vector machines. This is the same as conventional TWSVM. Thus, the complexity of each iteration is about $(m/2)^3$. The next section has experimentally demonstrated that the pTWSVM only needs to iterate three or four times to converge. Therefore, the full pTWSVM complexity will not be more than $4\times\left(2\times(m/2)^3\right) = m^3$. It is comparable to the traditional SVM.

IV. THE NONLINEAR KERNEL CLASSIFIER

In order to extend our method to nonlinear classifiers, we modify the new algorithm by the kernel method [36], [37]. We consider the kernel-generated surfaces for TWSVM instead of planes as follows:

$$\mathrm{K}\left(\mathrm{x}^{\mathrm{T}}, \mathrm{C}^{\mathrm{T}}\right)\mathrm{u}^1 + \mathrm{b}^1 = 0, \text{ and } \mathrm{K}\left(\mathrm{x}^{\mathrm{T}}, \mathrm{C}^{\mathrm{T}}\right)\mathrm{u}^2 + \mathrm{b}^2 = 0 \tag{37}$$

where $\mathrm{C}^{\mathrm{T}} = [\mathrm{A}\ \mathrm{B}]^{\mathrm{T}}$ and K is an appropriately chosen kernel. Note that if the K is a linear kernel like $\mathrm{K}\left(\mathrm{x}^{\mathrm{T}}, \mathrm{C}^{\mathrm{T}}\right) = \mathrm{x}^{\mathrm{T}}\mathrm{C}$, it will degenerate into an ordinary plane.

We construct an optimization objective KPTWSVM as follows:

$$\begin{aligned} &\min_{\mathrm{w}^1,\mathrm{b}^1,\mathrm{q}} \ \frac{1}{2}\left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|^{\mathrm{p}} + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ &\text{subject to } -\left(\mathrm{K}\left(\mathrm{B},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_2\mathrm{b}^1\right) + \mathrm{q} \ge \mathrm{e}_2, \quad \mathrm{q} \ge 0 \end{aligned} \tag{38}$$

where $\mathrm{c}^1 > 0$ is a parameter. Next, we define a Lagrange function $\mathcal{L}$ corresponding to the above:

$$\begin{aligned} \mathcal{L}\left(\mathrm{w}^1, \mathrm{b}^1, \mathrm{q}, \alpha, \beta\right) = {} & \frac{1}{2}\left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}} + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ & + \alpha^{\mathrm{T}}\left[\mathrm{K}\left(\mathrm{B},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_2\mathrm{b}^1 - \mathrm{q} + \mathrm{e}_2\right] \\ & - \beta^{\mathrm{T}}\mathrm{q} \end{aligned} \tag{39}$$

To solve the problem, we split the distance into two parts such that:

$$\begin{aligned} &\left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}} \\ &= \left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}-2} \cdot \left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^2 \end{aligned} \tag{40}$$

In formulation (40), $\left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^{\mathrm{p}-2}$ can be represented by $\mathcal{S}$. The Lagrange function is updated as follows:

$$\begin{aligned} \mathcal{L}\left(\mathrm{w}^1, \mathrm{b}^1, \mathrm{q}, \alpha, \beta\right) = {} & \frac{1}{2}\mathcal{S}\left\|\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right\|_2^2 + \mathrm{c}^1\mathrm{e}_2^{\mathrm{T}}\mathrm{q} \\ & + \alpha\left[\mathrm{K}\left(\mathrm{B},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_2\mathrm{b}^1 - \mathrm{q} + \mathrm{e}_2\right] \\ & - \beta\mathrm{q} \end{aligned} \tag{41}$$

We obtain the following K.K.T. conditions for KPTWSVM as follows:

$$\frac{\partial\mathcal{L}}{\partial\mathrm{w}^1} = \mathcal{S}\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)^{\mathrm{T}}\left[\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right] + \mathrm{K}(\mathrm{B},\mathrm{C}^{\mathrm{T}})^{\mathrm{T}}\alpha = 0 \tag{42}$$

$$\frac{\partial\mathcal{L}}{\partial\mathrm{b}^1} = \mathcal{S}\mathrm{e}_1^{\mathrm{T}}\left[\mathrm{K}(\mathrm{A},\mathrm{C}^{\mathrm{T}})\mathrm{w}^1 + \mathrm{e}_1\mathrm{b}^1\right] + \mathrm{e}_2^{\mathrm{T}}\alpha = 0 \tag{43}$$

$$\frac{\partial\mathcal{L}}{\partial\mathrm{q}} = \mathrm{c}^1\mathrm{e}_2 - \alpha - \beta = 0 \tag{44}$$

$$\frac{\partial\mathcal{L}}{\partial\alpha} = -\left[\mathrm{K}\left(\mathrm{B},\mathrm{C}^{\mathrm{T}}\right)\mathrm{w}^1 + \mathrm{e}_2\mathrm{b}^1\right] + q - \mathrm{e}_2 \ge 0 \tag{45}$$

$$\frac{\partial\mathcal{L}}{\partial\beta} = \mathrm{q} \ge 0 \tag{46}$$

$$\alpha^{\mathrm{T}}\left[\mathrm{K}(\mathrm{B},\mathrm{C}^{\mathrm{T}})\mathrm{w}^1 + \mathrm{e}_2\mathrm{b}^1 - \mathrm{q} + \mathrm{e}_2\right] = 0 \tag{47}$$

$$\beta^{\mathrm{T}}\mathrm{q} = 0 \tag{48}$$

$$\alpha \ge 0, \quad \beta \ge 0 \tag{49}$$

Combining (42) and (43), we arrive at

$$\mathcal{S}\left[\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)^{\mathrm{T}}\mathrm{e}_1^{\mathrm{T}}\right]\left[\mathrm{K}\left(\mathrm{A},\mathrm{C}^{\mathrm{T}}\right)\mathrm{e}_1\right]\left[\mathrm{w}^1, \mathrm{b}^1\right]^{\mathrm{T}} + \left[\mathrm{K}(\mathrm{B},\mathrm{C}^{\mathrm{T}})^{\mathrm{T}}\mathrm{e}_2^{\mathrm{T}}\right]\alpha = 0 \tag{50}$$

Let

$$\mathrm{E} = \left[\mathrm{K}(\mathrm{A},\mathrm{C}^{\mathrm{T}})^{\mathrm{T}}\mathrm{e}_1^{\mathrm{T}}\right], \quad \mathrm{R} = \left[\mathrm{K}(\mathrm{B},\mathrm{C}^{\mathrm{T}})^{\mathrm{T}}\mathrm{e}_2^{\mathrm{T}}\right] \tag{51}$$

and the augmented vector $\mathrm{u} = \left[\mathrm{w}^1, \mathrm{b}^1\right]^{\mathrm{T}}$. Then the formulation can be rewritten as:

$$\mathcal{S}\mathrm{E}^{\mathrm{T}}\mathrm{Eu} + \mathrm{R}^{\mathrm{T}}\alpha = 0 \tag{52}$$

i.e.

$$\mathrm{u} = -\frac{1}{\mathcal{S}}(\mathrm{E}^{\mathrm{T}}\mathrm{E})^{-1}\mathrm{R}^{\mathrm{T}}\alpha \tag{53}$$

The Wolfe dual of KPTWSVM is given by

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\cdot\frac{1}{\mathcal{S}}\alpha^{\mathrm{T}}\mathrm{R}\left(\mathrm{H}^{\mathrm{T}}\mathrm{H}\right)^{-1}\mathrm{R}^{\mathrm{T}}\alpha + \mathrm{e}_2^{\mathrm{T}}\alpha \\ &0 \le \alpha \le \mathrm{c}^1 \end{aligned} \tag{54}$$

In a similar manner, another KPTWSVM kernel-generated surface can be obtained by solving a new dual function.

Once the two KPTWSVM problems are solved, a new data point can be classified in a manner similar to the linear case.

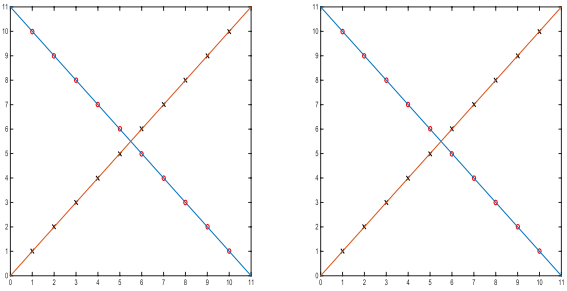
In the actual experiments, if the number of patterns is large, then the rectangular kernel technique can be used to reduce the dimensionality of KPTWSVM. In the linear case, a regularization term is always being useful.

V. EXPERIMENTAL RESULTS

In this section, experiments are conducted to evaluate the performance of our new method. We first compare the conventional TWSVM and pTWSVM on the artificial data set. Then we compare the pTWSVM with four widely used classifiers on several diverse public data sets. We also study the effect of the change in the value of p on the experimental results. And following this, experiments on robustness are displayed. Finally, analysis of iteration is reported.

A. BINARY DATA

In this subsection, a toy experiment is presented to show the difference between the traditional TWSVM and our new method. A simple data set was constructed by several points distributed on $y = x$ and $y = -x + 10$ respectively. The two classes of points are strictly binary data. In a two-dimensional Cartesian coordinate system, each of them need two lines perpendicular. The data set is strictly distributed on the two lines and has no noise. Although pTWSVM is committed to improving the robustness of TWSVM, it should have the same accuracy as TWSVM in the case of no noise. Moreover, since there is no noise, the algorithm only need to iterate once to achieve the final convergence results. Recovered images of Fig.1 show the classification surfaces of TWSVM and pTWSVM, respectively. Also, the binary dataset has been displayed as points in the images.



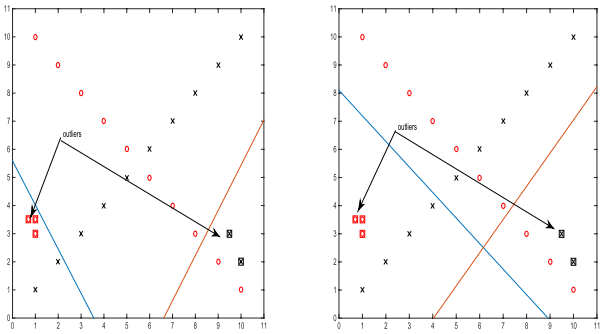
(TWSVM) (pTWSVM)

FIGURE 1. Binary data experiments pictures.

It is suggested that the two algorithms have a good classification effect on binary data sets and the classification surfaces are almost the same in Fig.1, and the result is in line with our expectations conjecture.

In order to introduce the outliers, we simulate some of the data points that shifted their original distribution, and mark these points with boxes. Afterwards, the same experiments will be carried out again to observe the difference between the obtained classification. The new data set and the classification surfaces of two methods are displayed in Fig.2.

From the Fig.2 we can find that the classification surfaces of TWSVM and pTWSVM are similar in terms of structure, and pTWSVM provides a better classification. This proves that pTWSVM is much less susceptible to outliers than TWSVM, and has good robustness.



(TWSVM) (pTWSVM)

FIGURE 2. Binary data with noise experiments pictures.

TABLE 1. Data sets descriptions.

Data set	Number	Dimension
heart	270	13
australian	690	14
pima	768	8
monk1	561	6
sonar	208	60
spect	267	44
cancer	683	9
ionodata	351	34
haberman	306	3
monk3	554	6
wpbc	194	33
bupa	345	6
checkdata	297	13

B. COMPARISON OF ACCURACY

In this subsection, several diverse public data sets are collected to compare the performance of different classification algorithms. The descriptions of the datasets are given in table 1. All the data sets are selected from the UCI Repository [41].

For fairness, a linear kernel is used in every comparison algorithm. We compare our algorithms with some of the widely used algorithms, including original TWSVM, SVM, GEPSVM, L1GEPSVM [28] and the latest NLPTSVM. To check the statistical significance of the new method, we perform the paired t-tests comparing these methods to our new method. The significant difference from pTWSVM based on p-value < 0.05. The p-value<0.05 indicates a great difference between two classification accuracy values appears. The 10-fold method was used to obtain the best parameters for each algorithm and the p value for pTWSVM. And all the parameters except p value are obtained over the range $\left(2^i | i = -7, -6, \ldots + 7\right)$. We present the average accuracy, time and the standard deviation in TABLE 2. The best performance on different datasets is shown in bold.

Form table2 we can find that pTWSVM performs best on the vast majority of data sets compared to several other algorithms. Comparing pTWSVM with TWSVM alone, we know that pTWSVM is always more accurate than TWSVM classification, although it is not on a very individual data set, with only a difference of less than 0.1%, which can be ignored. This situation can be explained that TWSVM is a special case of pTWSVM. When the parameter p of the

TABLE 2. Test set accuracy with a linear kernel (average ± standard deviation).

	pTWSVM	L1GEP	TWSVM	SVM	GEPSVM	NLPTSVM
	Accuracy Time(s) p-value	Accuracy Time(s) p-value	Accuracy Time(s) p-value	Accuracy Time(s) p-value	Accuracy Time(s) p-value	Accuracy Time(s) p-value
heart	**0.8444±2.7716** 0.1623 —	0.7889±5.5679 0.0132 0.0720	0.8296±3.9545 0.0071 0.5675	0.8259±3.0089 0.9383 0.4698	0.7963±4.3823 0.7859 0.1113	0.6741± 1.4815 0.0427 5.97e-5
australian	0.8493±2.5269 1.2176 —	0.6794±4.9658 0.0211 8.62e-6	0.8464±4.0059 0.1180 0.8613	**0.8551±1.6524** 8.1210 0.6857	0.6609±4.6603 1.0614 1.65e-6	0.5768± 3.0607 0.8684 2.55e-7
pima	**0.7657±3.8292** 1.1706 —	0.7552±4.0560 0.0137 0.5455	0.7539±2.3068 0.0412 0.5268	0.7539±3.4394 1.8497 0.5713	0.7487±4.2436 0.9329 0.3572	0.7408± 4.2634 0.9378 0.3558
monk1	0.7007±7.0755 0.3543 —	**0.7986±3.9823** 0.0125 5.06e-7	0.7058±3.1856 0.0934 0.7047	0.5545±9.2960 0.1614 5.10e-7	0.7665±2.2915 0.8432 0.0515	0.6649± 4.5580 0.0777 0.1641
sonar	0.6825±10.0408 0.3965 —	0.7117±4.8810 0.0158 0.0816	0.6870±5.5519 0.0079 0.8062	**0.7408±3.5758** 1.5948 0.0293	0.7262±9.5241 4.2953 0.0184	0.7211± 6.1541 0.0257 0.2800
spect	**0.7941±1.5024** 0.1442 —	0.5882±4.8803 0.0187 2.02e-6	0.7936±5.4955 0.0062 0.9740	0.7157±4.4007 1.5253 0.0041	0.7827±5.0937 2.7397 0.6591	0.7939± 5.4955 0.0253 0.9951
cancer	0.9663±1.2882 1.4262 —	0.9196±7.1479 0.0159 0.0033	0.9664±1.6358 0.0925 0.9934	**0.9722±1.1695** 0.2452 0.6237	0.9561±2.2693 1.0705 0.4251	0.9547± 1.8037 0.3123 0.3608
ionodata	**0.9089±1.9005** 0.2017 —	0.8261±4.4920 0.0140 3.19e-4	0.8575±5.6590 0.0094 0.0121	0.8604±3.1790 1.4361 0.0204	0.7976±4.4062 2.1234 6.43e-4	0.8689± 5.6814 0.3446 0.2272
haberman	0.6320±19.5167 0.1335 —	**0.7518±4.782**1 0.0123 1.80e-4	0.7352±5.1755 0.0079 0.0014	0.6403±21.1010 0.2823 0.6761	0.7485±5.0298 0.7074 6.57e-4	0.7319± 5.2829 0.0204 0.0105
monk3	0.8284±5.9756 0.6786 —	**0.8700±2.1738** 0.0142 0.0908	0.7816±2.7899 0.0361 0.0240	0.4801±3.5137 0.1020 8.39e-9	0.7978±3.6369 0.8342 0.0619	0.7780± 3.3755 0.5351 0.0323
wpbc	**0.7891±5.8439** 0.1236 —	0.7267±7.3870 0.0131 0.0042	0.7626±7.0603 0.0060 0.1583	0.7316±6.7910 1.8354 0.0298	0.7629±6.5623 1.4888 0.1706	0.7626± 7.5985 0.0634 0.5226
bupa	**0.6986±3.3553** 0.2546 —	0.5449±5.1526 0.0118 3.99e-5	0.6725±4.7274 0.0091 0.1921	0.6609±6.2571 0.8766 0.0232	0.5391±4.0372 0.7477 4.03e-6	0.6290± 7.2522 0.0975 0.1008
checkdata	0.5360±4.8724 1.3981 —	**0.5720±5.8429** 0.0197 0.0134	0.5080±4.7603 0.0785 0.0703	0.5180±4.7476 0.6881 0.1096	0.5220±5.7845 0.9978 0.4931	0.5180± 3.9064 0.5537 0.4245

pTWSVM is fixed to 2, the pTWSVM is transformed into TWSVM. Results from our toy experiment show that when $p = 2$, the hyper-planes obtained by pTWSVM are the same as that of TWSVM, and only loop once. In theory, when p is not fixed in 2, the pTWSVM provides more parameter selections to optimize the algorithm. In addition, from the table2, we can find the standard deviation of the new method is always smaller than the standard deviation of other methods for most data sets. This implies that our proposed new method has better robustness and our algorithm has higher stability, which is in line with our expectations.

Clearly, the many corresponding p-values in Table2 are less than 0.05, i.e., the accuracies of pTWSVM are obviously greater than those of the other classifiers on most datasets. For example, the p-values of the tests comparing pTWSVM with NLPTSVM on ionodata and monk3 data sets are 0.0121 and 0.0240 respectively, leading us to conclude that pTWSVM is significantly better than TWSVM on the two data sets. This also appears in SVM. Besides, we observe that in some data sets, pTWSVM does not have the highest accuracy. For example, the accuracy of L1GEP on australian and cancer. However, the p-values of the t-tests comparing pTWSVM with them on these data sets are, respectively, 0.6857 and 0.6237, which leads us to conclude that there is not a great difference between them in statistically significant. The p-values of t-tests also prove that there are significant differences between pTWSVM and NLPTSVM on four data sets.

When concerning the computational cost shown in Table 2, it is to be noted that NLPTSVM is always faster than pTWSVM. This can be explained from their formulations. Although they are both iterative algorithms based on

traditional TWSVM, but pTWSVM solves the quadratic programming problems (QPPs), NLPTSVM solves the linear programming problems (LPPs).

The experimental results indicate that pTWSVM is not only effective, but also can be a better choice for most data sets.

C. STUDY THE p VALUE OF THE NEW PROPOSED METHOD

The new method arises a problem of value of p. Considering the objective function, we hold the notion that the p value is under the influence of outliers. In order to get a higher accuracy, the greater proportion of noise, the smaller value of p, and vice versa. Formula (9) perspicuously indicates that p value directly affects the result of the objective. Splitting the formulation into two parts: the functional margin of outlier data points and the functional margin of normal data points. The role of p value is to emphasize the proportion of the two parts. In summary, we hold the notion that the parameter p value can directly affect experiment accuracy.

We experiment on australian, sonar, spect and several other benchmark data sets as examples. In order to measure the effect of p on accuracy, we set the remaining parameters to a specific value $c1 = c2 = 1$. Then we record the accuracy of the different p values. We vary p of the proposed objective in the range of 0.1 to 2 to study its impacts to the classification performance. Through the experimental data, we simulate the corresponding correct rate curve. All the records are presented in Fig.3.

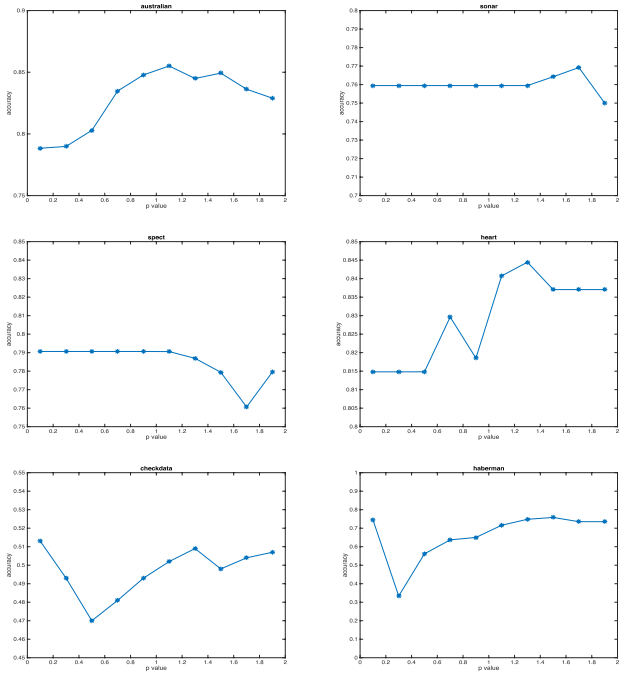


FIGURE 3. Accuracy with different p value.

Fig. 3 shows that the determination of p is strongly related to the specific dataset. We can find two conclusions in the figure 3: one is, when the p value is too small, the classification accuracy is not very stable; another is, when the p value is between 1.0 to 1.2, pTWSVM always have a very good performance. Above mentioned can be explained from three aspects. Firstly, when p is small, the value of δ could be so extremely big that the value of the objective function is not accurate. Next, the regularization is set to $1e-7$, it may have an effect on the calculation results for singularity problems. Lastly, the data distribution and numerical size of the data set can affect the calculation process. However, When the p value is a little larger, these problems will be greatly alleviated and the classification performance will rise and stabilize.

In order to have a better accuracy, we have adopted a strategy that the most appropriate p is selected from $\{0.1, 0.2 \ldots 2.0\}$ through 10-fold cross validation.

D. CONVERGENCE STUDY OF SOLUTION ALGORITHM

As the proposed algorithm is an iterative algorithm, an important issue is the convergence property of our new method. In the previous chapter, we have rigorously proved its convergence in theory, and now we study its convergence from the experiment empirically. We experiment on several data sets and the p value is fixed. The objective values of our proposed algorithm on the four data sets in each iteration are plotted in the Fig. 4.

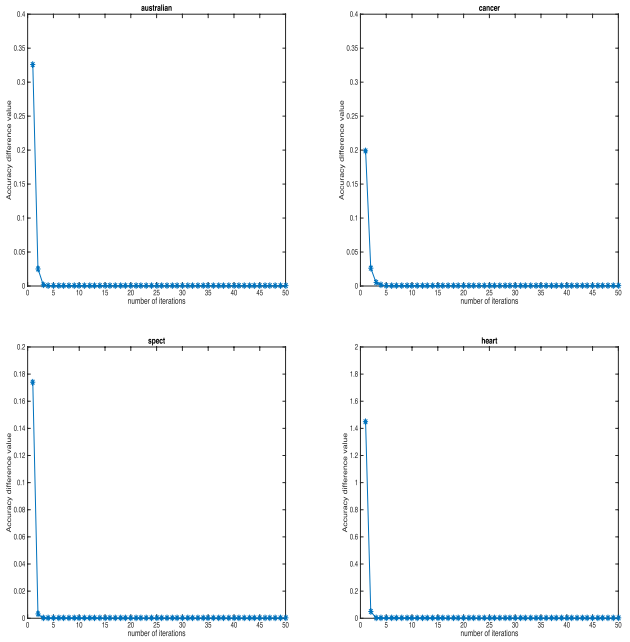


FIGURE 4. Number of iterations vs. the objective value difference.

Fig.4 shows that the objective values of our new proposed algorithm keep to decrease along with the iterative processes. Moreover, the algorithm typically converges to the asymptote within five times for each data set, which means that the algorithm is computationally and temporally feasible. Upon these experimental results, we set a stopping threshold of 10^{-5} in our experiments, which is sufficient to achieve satisfactory results in terms of convergence.

TABLE 3. Test set accuracy with 20% noise (average ± standard deviation).

	pTWSVM	L1GEP	TWSVM	SVM	GEPSVM	NLPTSVM
	Accuracy p-value	Accuracy p-value	Accuracy p-value	Accuracy p-value	Accuracy p-value	Accuracy p-value
heart	**0.7037±8.8425** –	0.6778±6.6872 0.2522	0.6852±1.1712 0.4363	0.7037±4.5361 0.9993	0.6556±5.9259 0.0820	0.6630±2.1596 0.0922
australian	**0.6812±2.9701** –	0.6261±7.2551 0.0037	0.6580±4.7716 0.1955	0.5928±3.1553 3.39e-4	0.6101 ±5.5263 0.0068	0.5754±3.1950 7.08e-4
pima	**0.7527±3.2041** –	0.7279±3.1829 0.2118	0.7473±5.0706 0.7443	0.7435±3.3122 0.5015	0.7240±2.7397 0.2411	0.7174±5.5374 0.0258
monk1	0.6863±4.1689 –	0.8002±4.2124 0.0010	0.6542±2.9356 0.1227	0.5437±6.6010 1.47e-5	**0.8003±2.8472** 3.41e-5	0.6649±4.5580 0.2543
sonar	**0.7502±8.0145** –	0.7024±9.1912 0.0232	0.6820±8.6080 0.0127	0.7498± 6.9030 0.9782	0.7355±2.1783 0.3558	0.7259±6.5317 0.1043
spect	0.7679±4.3534 –	0.5579±5.2548 1.68e-6	0.7901±5.0184 0.2039	0.7230±5.4347 0.0331	0.7718±3.6723 0.7666	**0.7939±5.4955** 0.1662
cancer	0.9605±1.9395 –	0.9590±0.5932 0.9173	0.9649±1.5532 0.7347	**0.9678±0.9945** 0.5161	0.9532±1.5048 0.5975	0.9561±1.9514 0.7156
ionodata	**0.9088±2.8048** –	0.8120±4.4297 1.26e-4	0.8632±4.6761 0.0593	0.8719±2.3386 0.0948	0.8120±3.7534 3.03e-5	0.8718±5.1995 0.0874
haberman	0.7451±4.6970 –	0.7418±4.0654 0.8491	0.7254±5.0644 0.3094	0.7420±2.6346 0.8794	**0.7548±4.6652** 0.5593	0.7287±5.1614 0.3124
monk3	**0.8683±5.0077** –	0.8411±3.6339 0.0813	0.7960±1.9319 0.0010	0.7038±14.7974 1.75e-6	0.7961±2.9466 2.86e-4	0.7780±3.3755 2.29e-4
wpbc	**0.7935±7.1711** –	0.6804±7.5218 1.74e-5	0.7470±5.2652 0.0181	0.6080±3.7282 5.20e-7	0.7675±5.7140 0.0560	0.7626±7.5985 0.0166
bupa	**0.6812±5.1034** –	0.6174±3.7345 0.0065	0.6493±10.5867 0.1574	0.6464±4.3575 0.1724	0.5188±4.8847 1.24e-5	0.6319±6.8347 0.0771
checkdata	0.5340±4.5100 –	**0.5770±4.6968** 0.0920	0.5100±2.7203 0.2867	0.5070±1.9900 0.1932	0.5300±1.9494 0.8108	0.5174±3.8419 0.4727

E. ROBUSTNESS AGAINST NOISE SAMPLES

Since the main advantage of the new proposed pTWSVM algorithm dedicated to process noisy samples, we will focus on the processing of the data sets with outliers in the following experiments.

To emulate the outlier data samples, given the input data set $X = [x_1, \ldots, x_n] \in R^{m\times n}$, we corrupt it by a noise matrix $\tilde{X} \in R^{d\times n}$ whose elements are i.i.d. standard Gaussian variables. Then we carry out the same learning and clustering procedures on $X + \sigma\tilde{X}$ as those on the original data, where $\delta = \text{nf}\frac{\|X\|_F}{\|\tilde{X}\|_F}$ and nf is the given noise factor. In all our experiments, we set nf $= 0.1$. We compare our new method against other methods as before and report the classification results in Table3.

As is shown in Table3, in the case of adding the same noise, the new proposed pTWSVM demonstrates its strong robustness. pTWSVM exhibits the highest classification accuracy on different data sets. Comparing with the classification results when no noise is added, it is also suggested that the pTWSVM classification accuracy is reduced in each algorithm. Also, we note that in these five data sets, which pTWSVM does not have the best accuracies, the corresponding p-values are 3.41e-5, 0.1662, 0.5161, 0.5593, 0.0920 respectively. The five p-values have only one less than 0.05, which means that the other four do not show any significant difference in statistical significance.

The differences in the accuracy of each algorithm will be obtained by contrasting with the performances of the original data and contaminated data. To get a deep association, we take a different η value in experiments. The following pictures summarizes the performance of different algorithms on some benchmark datasets with different values of η.

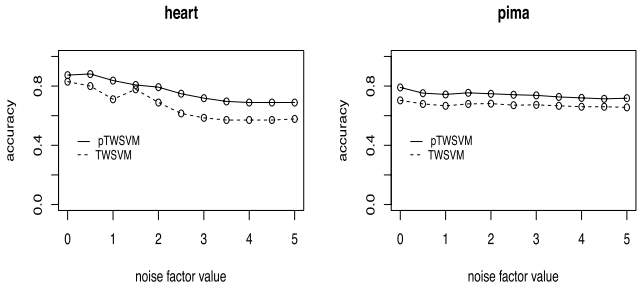


FIGURE 5. Accuracy with different noise factor value.

From the Fig.5, we can get the following points:

Firstly, the proposed pTWSVM method is consistently better than the TWSVM method on the experimental data sets, which demonstrate that the proposed method is able to effectively improve the clustering accuracy on noisy data with outlier data samples. This also shows that the new pTWSVM method in the practical application can achieve better results.

Secondly, no matter what the noise factor value is, the accuracy of pTWSVM is always higher than that of TWSVM. Although the improvements by the pTWSVM method over the comparative methods on the original benchmark data sets without noise are mediocre as shown in Table2, the improvements by our new method of the

contaminated data with outlier data samples are considerably larger. For example, on the heart data set with outliers, the average pTWSVM accuracy of different η value is 0.7481, and TWSVM accuracy is 0.6633. So our proposed method improves the clustering accuracy over the TWSVM method by $12.78\% = (0.7481 - 0.6633)/0.6633$. In contrast, the improvement of clustering accuracy on the same data set under the noiseless condition is about $4.47\% = (0.8667 - 0.8296)/0.8296$. The same situation can be seen on all the other experimental data sets, which show that the proposed method has better capability to cluster on contaminated data.

Finally, the pictures show that the change in accuracy of pTWSVM is flat and does not change much, which clearly indicates that the new proposed pTWSVM method is faster and easier to stabilize than original TWSVM method. The feature confirms pTWSVM method's robustness against outlier data samples.

VI. CONCLUSIONS

We have proposed a robust TWSVM based on the $l_{2,p}-$norm distance, which formulated a non-smooth and non-convex minimization problem. Comparing to the squared l_2-norm distance, the $l_{2,p}-$norm TWSVM has better accuracy and it is very robust against outlying data samples. The new method takes much more challenging optimization problem than that in the traditional TWSVM. To solve the problem, we introduced an efficient iterative algorithm and provided the rigorous theoretical analysis of the convergence of our algorithm.

There are still several directions to investigate in the future. First, the problem of dealing with the singularity. It is addressed by regularization in our study. Second, during each iteration, if the p value is too small, such as 0.1,0.2, then the $\mathcal{S}$ value will become very large. This would lead to not accurate. Finally, deciding the values of parameters is still an open problem, whereas, has not been solved in many algorithms as well.

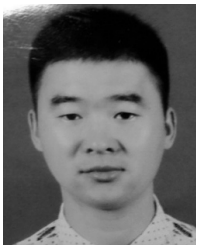
REFERENCES

[1] M. M. Adankon and M. Cheriet, "Support vector machine," *Comput. Sci.*, vol. 1, no. 4, pp. 1–28, 2002.
[2] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995, pp. 988–999.
[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[4] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, Jan. 2002.
[5] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, p. 1155, 2007.
[6] N. Chen, W. Lu, J. Yang, and G. Li, *Support Vector Machine in Chemistry*. Singapore: World Scientific, 2004.
[7] P. Shih and C. Liu, "Face detection using discriminating feature analysis and support vector machine in video," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2006, pp. 407–410.
[8] J. H. Min and Y.-C. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert Syst. Appl.*, vol. 28, no. 4, pp. 603–614, May 2005.
[9] B. Schoelkopf, K. Tsuda, and J.-P. Vert, Eds., *Support Vector Machine Applications in Computational Biology*. Cambridge, MA, USA: MIT Press, 2004.
[10] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucl. Acids Res.*, vol. 36, no. 9, pp. 3025–3030, 2008.
[11] P. S. Bradley and O. L. Mangasarian, "Massive data discrimination via linear support vector machines," *Optim. Methods Softw.*, vol. 13, no. 1, pp. 1–10, 2000.
[12] R. A. Lordo, "Learning from data: Concepts, theory, and methods," *Technometrics*, vol. 43, no. 1, pp. 105–106, 2008.
[13] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, "Nonlinear programming: Theory and algorithms," *J. Oper. Res. Soc.*, vol. 45, no. 7, pp. 846–846, 1979.
[14] M. A. Kumar and M. Gopal, "A comparison study on multiple binary-class SVM methods for unilabel text categorization," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1437–1444, Aug. 2010.
[15] J. Jayadeva, R. Khemchandani, and S. Chandra, "TWSVM for unsupervised and semi-supervised learning," in *Twin Support Vector Machines*. Springer, 2017, pp. 125–152.
[16] J. Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 905–910, May 2007.
[17] W. U. Enying and L. Jia, "Research on multiclass classification algorithm based on binary tree SVM," *J. Chongqing Normal Univ.*, vol. 33, pp. 102–106, 2016.
[18] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *Explorations Newslett.*, vol. 6, no. 1, pp. 1–6, 2004.
[19] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach," *Acm SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 30–39, 2004.
[20] N. Japkowicz, "Learning from imbalanced data sets," *Ai Mag.*, 2000.
[21] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.
[22] G. Fung and O. L. Mangasarian, "Proximal support vector machine classifiers," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 77–86.
[23] O. L. Mangasarian and E. W. Wild, "Multisurface proximal support vector machine classification via generalized eigenvalues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 69–74, Jan. 2006.
[24] Y. Dodge, *Statistical Data Analysis Based on the L1-Norm and Related Methods*, vol. 6. Amsterdam, The Netherlands: North Holland, 1987, pp. 280–282.
[25] A. Eriksson and A. van den Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L1 norm," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 771–778.
[26] S. Gao, Q. Ye, and N. Ye, "1-norm least squares twin support vector machines," *Neurocomputing*, vol. 74, no. 17, pp. 3590–3597, Oct. 2011.
[27] J. J. Liang and W. U. De, "Sparse least square support vector machine with L1 norm," *Comput. Eng. Des.*, 2014.
[28] H. Yan, Q. Ye, T. Zhang, D. Yu, X. Yuan, Y. Xu, and L. Fu, "Least squares twin bounded support vector machines based on L1-norm distance metric for classification," *Pattern Recognit.*, vol. 74, pp. 434–447, 2017.
[29] R. Yan, Q. Ye, L. Zhang, N. Ye, and X. Shu, "A feature selection method for projection twin support vector machine," *Neural Process. Lett.*, vol. 3, pp. 1–18, 2017.
[30] H. Firouzi, M. Farivar, M. Babaie-Zadeh, and C. Jutten, "Approximate sparse decomposition based on smoothed ℓ^0-norm," *Mathematics*, 2012.
[31] M. Thiao, T. P. Dinh, and H. A. L. Thi, "DC programming approach for a class of nonconvex programs involving l0 norm," *Commun. Comput. Inf. Sci.*, vol. 14, no. 2, pp. 348–357, 2008.
[32] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.
[33] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 796–808, Apr. 2016.
[34] H. Wang, F. Nie, and H. Huang, "Learning robust locality preserving projection via p-order minimization," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3059–3065.
[35] B. N. Parlett, *The Symmetric Eigenvalue Problem*. Philadelphia, PA, USA: SIAM, 1980, pp. 1–22.
[36] B. SchÃűlkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999, pp. 1095–1097.

[37] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 323–330.
[38] M. Tanveer, "Robust and sparse linear programming twin support vector machines," *Cognit. Comput.*, vol. 7, no. 1, pp. 137–149, Feb. 2015.
[39] M. Tanveer, "Smoothing technique on linear programming twin support vector machines," *Int. J. Mach. Learn. Comput.*, vol. 3, no. 2, pp. 240–244, 2013.
[40] M. Tanveer and K. Shubham, "A regularization on Lagrangian twin support vector regression," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 3, pp. 807–821, Jun. 2017.
[41] C. L. Blake and C. J. Merz. (1998). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA, USA. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
[42] L. Wang, S. Chen, and Y. Wang, "A unified algorithm for mixed $\ell_{2,p}$-minimizations and its application in feature selection," *Comput. Optim. Appl.*, vol. 58, no. 2, pp. 409–421, Jun. 2014.

XU MA was born in 1993. He is currently pursuing the master's degree with the College of Information Science and Technology, Nanjing Forestry University, Jiangsu, China. His main research interests include pattern recognition, machine learning, data mining, and image processing.

QIAOLIN YE received the B.S. degree in computer science from the Nanjing Institute of Technology, Nanjing, China, in 2007, the M.S. degree in computer science and technology from Nanjing Forestry University, Nanjing, in 2009, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Jiangsu, China, in 2013. He is currently an Associate Professor with the Computer Science Department, Nanjing Forestry University, Nanjing, China. He has authored over 50 scientific papers. Some of them are published in the IEEE TNNLS, the IEEE TIFS, and the IEEE TCSVT. His research interests include machine learning, data mining, and pattern recognition.



HE YAN was born in 1988. He is currently pursuing the master's degree with the College of Information Science and Technology, Nanjing Forestry University, Nanjing, China. His main research interests include pattern recognition, machine learning, and data mining.