

## Article

# A Multiclass Nonparallel Parametric-Margin Support Vector Machine

Shu-Wang Du <sup>1</sup>, Ming-Chuan Zhang <sup>2</sup>, Pei Chen <sup>2</sup>, Hui-Feng Sun <sup>2</sup>, Wei-Jie Chen <sup>1,2,\*</sup>  and Yuan-Hai Shao <sup>3</sup> 

<sup>1</sup> Zhijiang College, Zhejiang University of Technology, Shaoxing 312030, China; dusw@zjut.edu.cn

<sup>2</sup> College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310024, China; 2112012210@zjut.edu.cn (M.-C.Z.); 2112112292@zjut.edu.cn (P.C.); 2112112244@zjut.edu.cn (H.-F.S.)

<sup>3</sup> Management School, Hainan University, Haikou 570228, China; shaoyuanhai@hainanu.edu.cn

\* Correspondence: wjchen@zjut.edu.cn

**Abstract:** The twin parametric-margin support vector machine (TPMSVM) is an excellent kernel-based nonparallel classifier. However, TPMSVM was originally designed for binary classification, which is unsuitable for real-world multiclass applications. Therefore, this paper extends TPMSVM for multiclass classification and proposes a novel  $K$  multiclass nonparallel parametric-margin support vector machine (MNP-KSVC). Specifically, our MNP-KSVC enjoys the following characteristics. (1) Under the “one-versus-one-versus-rest” multiclass framework, MNP-KSVC encodes the complicated multiclass learning task into a series of subproblems with the ternary output  $\{-1, 0, +1\}$ . In contrast to the “one-versus-one” or “one-versus-rest” strategy, each subproblem not only focuses on separating the two selected class instances but also considers the side information of the remaining class instances. (2) MNP-KSVC aims to find a pair of nonparallel parametric-margin hyperplanes for each subproblem. As a result, these hyperplanes are closer to their corresponding class and at least one distance away from the other class. At the same time, they attempt to bound the remaining class instances into an insensitive region. (3) MNP-KSVC utilizes a hybrid classification and regression loss joined with the regularization to formulate its optimization model. Then, the optimal solutions are derived from the corresponding dual problems. Finally, we conduct numerical experiments to compare the proposed method with four state-of-the-art multiclass models: Multi-SVM, MBSVM, MTPMSVM, and Twin-KSVC. Experimental results demonstrate the feasibility and effectiveness of MNP-KSVC in terms of multiclass accuracy and learning time.

**Keywords:** multiclass classification; twin support vector machine; nonparallel hyperplane; hybrid classification and regression; MNP-KSVC



**Citation:** Du, S.-W.; Zhang, M.-C.; Chen, P.; Sun, H.-F.; Chen, W.-J.; Shao, Y.-H. A Multiclass Nonparallel Parametric-Margin Support Vector Machine. *Information* **2021**, *12*, 515. <https://doi.org/10.3390/info12120515>

Academic Editor: Ognjen Arandjelović

Received: 15 November 2021

Accepted: 8 December 2021

Published: 10 December 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Data mining has become an essential tool for integrating information technology and industrialization due to the growing size of available databases [1]. One of the main applications in data mining is supervised classification. This aims to assign a label to an unseen instance that is as correct as possible from a given set of classes based on the learning model. Recently, support vector machine (SVM) [2,3] has been a preeminent maximum-margin learning paradigm for data classification. The basic idea of SVM is to find an optimal decision boundary via maximizing the margin between two parallel support hyperplanes. Compared with the neural network, SVM has the following attractive features [3]: (1) the structural risk minimization principle is implemented in SVM to control the upper bound of the generalization error, leading to an excellent generalization ability; (2) the global optimum can be achieved by optimizing a quadratic programming problem (QPP). Furthermore, kernel techniques enable SVM to handle complicated nonlinear learning tasks effectively. During recent decades, SVM has been successfully applied in a wide variety of fields ranging from scene classification [4], fault diagnosis [5,6],

EEG classification [7,8], pathological diagnosis [9,10], and bioinformatics [8] to power applications [11].

One limitation in the classical SVM is the strictly parallel requirements for support vector hyperplanes. Namely, parallel hyperplanes are challenging to use to capture a data structure with a cross-plane distribution [12,13], such as “XOR” problems. To alleviate the above issue, the nonparallel SVM models have been proposed in the literature [13–17] during the past years. This approach relaxes the parallel requirement in SVM and seeks nonparallel hyperplanes for different classes. The pioneering work is the generalized eigenvalue proximal SVM (GEPsVM) proposed by Mangasarian and Wild [12], which attempts to find a pair of nonparallel hyperplanes via solving eigenvalue problems (EPs). Subsequently, Jayadeva et al. [13] proposed a novel QPP-type nonparallel model for classification, named TWSVM. The idea of TWSVM is to generate two nonparallel hyperplanes such that each hyperplane is closer to one of the two classes and is at least one apart from the other class. Compared with the classical SVM, the nonparallel SVM models (GEPsVM and TWSVM) have lower computational complexity and better generalization ability. Therefore, in the last few years, they have been studied extensively and developed rapidly, including a least squares version of TWSVM (LSTsVM) [16], structural risk minimization version of TWSVM (TBSVM) [17],  $\nu$ -PTSVM [18], nonparallel SVM (NPSVM) [19,20], nonparallel projection SVM (NPrSVM) [21], and so on [21–28].

The above nonparallel models were mainly proposed for binary classification problems. However, most real-world applications [29–32] are related to multiclass classifications such as disease diagnosis, fault detection, image recognition, and text categorization. Therefore, many researchers are interested in extending SVM models from binary to multiclass classification. Generally, the decomposition procedure has been considered to be an effective way to achieve multiclass extensions. Yang et al. [33] proposed a multiple birth SVM (MBSVM) for multiclass classification based on the “one-versus-rest” strategy, which is the first multiclass extension of the nonparallel SVM model. Angulo et al. [34] proposed an “one-versus-one-versus-rest” multiclass framework. In contrast to the “one-versus-one” strategy, it constructs  $k(k-1)/2$  binary classifiers with all data points, which can avoid the risk of information loss and class distortion problems. Following this framework, Xu et al. [35] proposed a multiclass extension of TWSVM, termed Twin-KSVC. Results show that Twin-KSVC has a better generalization ability than MBSVM in most cases. Nasiri et al. [36] formulated Twin-KSVC in the least-squared sense to boost learning efficiency and further presented the LST-KSVC model. Lima et al. [32] proposed an improvement on LST-KSVC (ILST-KSVC) with regularization to implement the structural risk minimization principle.

As a successful extension of SVM, the twin parametric-margin support vector machine (TPMSVM) [15] was proposed to pursue a pair of parametric-margin nonparallel hyperplanes. Unlike GEPsVM and TWSVM, each hyperplane in TPMSVM aims to be closer to its class and far away from the other class. The parametric-margin mechanism enables TPMSVM to be suitable for many cases and results in better generalization performance. However, TPMSVM can only deal with binary classification learning tasks. The above motivates us to propose a novel  $K$  multiclass nonparallel parametric-margin support vector machine, termed MNP-KSVC. The proposed MNP-KSVC is endowed with the following attractive advantages:

- The proposed MNP-KSVC encodes the  $K$  multiclass learning task into a series of “one-versus-one-versus-rest” subproblems with all the training instances. Then, it encodes the outputs of subproblems with the ternary output  $\{-1, 0, +1\}$ , which helps to deal with imbalanced cases.
- For each subproblem, MNP-KSVC aims to find a pair of nonparallel parametric-margin to separate the two selected classes together with the remaining classes. Unlike TPMSVM, each parametric-margin hyperplane is closer to its class and at least at a distance away from the other class, meanwhile mapping the remaining instances into a region.

- To implement the empirical risks, MNP-KSVC considers the hybrid classification and regression loss. The Hinge loss is utilized to penalize the errors of the focused two class instances and the  $\varepsilon$ -intensive loss for the remaining class instances.
- Extensive numerical experiments are performed on several multiclass UCI benchmark datasets, and their results are compared with four models (Multi-SVM, MBSVM, MTPMSVM, and Twin-KSVC). The comparative results indicate the effectiveness and feasibility of the proposed MNP-KSVC for multiclass classification.

The remainder of this paper is organized as follows. Section 2 briefly introduces notations and related works. Section 3 proposes our MNP-KSVC. The model optimization is also discussed in Section 3. The nonlinear version of MNP-KSVC is extended in Section 4. Experimental results are described in Sections 5 and 6 presents a discussion and future work.

## 2. Preliminaries

In this section, we first describe the notations used throughout the paper. Then, we briefly revisit the nonparallel classifier PTSVM and its variants.

### 2.1. Notations

In this paper, scalars are denoted by lowercase italic letters, vectors by lowercase bold face letters, and matrices by uppercase letters. All vectors are column vectors unless transformed to row vectors by a prime superscript  $(\cdot)'$ . A vector of zeros of arbitrary dimensions is represented by  $\mathbf{0}$ . In addition, we denote  $\mathbf{e}$  as a vector of ones and  $\mathbf{I}$  as an identity matrix of arbitrary dimensions. Moreover, let  $\|\cdot\|$  stand for the  $L_2$ -norm.

### 2.2. TPMSVM

The TPMSVM [15] is originally proposed for binary classification with a heteroscedastic noise structure. It attempts to seek two nonparallel parametric-margin hyperplanes via the following QPPs:

$$\begin{aligned} \min_{w_1, b_1} \quad & \frac{1}{2} \|w_1\|^2 + v_1 \sum_{i \in \mathcal{I}_+} \xi_{1i} + c_1 \sum_{j \in \mathcal{I}_-} \eta_{1j}, \\ \text{s.t.} \quad & w_1' x_i + b_1 \geq -\xi_{1i}, \quad \xi_{1i} \geq 0, \\ & w_1' x_j + b_1 = \eta_{1j}, \end{aligned} \quad (1)$$

and

$$\begin{aligned} \min_{w_2, b_2} \quad & \frac{1}{2} \|w_2\|^2 + v_2 \sum_{i \in \mathcal{I}_-} \xi_{2i} - c_2 \sum_{j \in \mathcal{I}_+} \eta_{2j}, \\ \text{s.t.} \quad & w_2' x_i + b_2 \leq \xi_{2i}, \quad \xi_{2i} \geq 0, \\ & w_2' x_j + b_2 = \eta_{2j}, \end{aligned} \quad (2)$$

where  $v_1, v_2, c_1, c_2$  are positive parameters, and the decision hyperplane is half of the sum of the two hyperplanes. To obtain the solutions of problems (1) and (2), one needs to resort to the dual problems

$$\begin{aligned} \min_{\alpha_1} \quad & \frac{1}{2} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_+} \alpha_{1i}^2 x_i' x_j - v_1 \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} \alpha_{1i} x_i' x_j, \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I}_+} \alpha_{1i} = v_1, \\ & 0 \leq \alpha_{1i} \leq c_1 \end{aligned} \quad (3)$$

and

$$\begin{aligned} \min_{\alpha_2} \quad & \frac{1}{2} \sum_{i \in \mathcal{I}_-} \sum_{j \in \mathcal{I}_-} \alpha_{2i}^2 x_i' x_j - \nu_2 \sum_{i \in \mathcal{I}_-} \sum_{j \in \mathcal{I}_+} \alpha_{2i} x_i' x_j, \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I}_-} \alpha_{2i} = \nu_2, \\ & 0 \leq \alpha_{2i} \leq c_2. \end{aligned} \quad (4)$$

Then, the solutions  $(w_1, b_1)$  and  $(w_2, b_2)$  of dual problems (3) and (4) can be calculated according to the Karush–Kuhn–Tucker (KKT) conditions [3],

$$w_1 = \sum_{i \in \mathcal{I}_+} \alpha_{1i} x_i - \nu_1 \sum_{j \in \mathcal{I}_-} x_j \text{ and } w_2 = - \sum_{i \in \mathcal{I}_-} \alpha_{2i} x_i + \nu_2 \sum_{j \in \mathcal{I}_+} x_j, \quad (5)$$

$$b_1 = - \frac{1}{|\mathcal{I}_{SV_1}|} \sum_{i \in \mathcal{I}_{SV_1}} w_1' x_i \text{ and } b_2 = - \frac{1}{|\mathcal{I}_{SV_2}|} \sum_{i \in \mathcal{I}_{SV_2}} w_2' x_i, \quad (6)$$

where  $\mathcal{I}_{SV_1}$  and  $\mathcal{I}_{SV_2}$  are indices of the support vector set.

Note that TPMSVM can capture more complex heteroscedastic error structures via parametric-margin hyperplanes compared with TWSVM [13]. However, the variable  $b$  in TPMSVM is not strictly convex, leading to a lack of a unique solution. Moreover, the optimization problems (1) and (2) are only designed for binary classification tasks, and thus are unsuitable for many real-world multiclass learning applications.

### 3. The Proposed MNP-KSVC

#### 3.1. Model Formulation

To address the above issues in TPMSVM, this subsection proposes a novel  $K$  multiclass nonparallel parametric-margin support vector machine, termed MNP-KSVC. Inspired by the “hybrid classification and regression” learning paradigm [35], MNP-KSVC decomposes the complicated  $K$  multiclass learning task into a series of “one-versus-one-versus-rest” subproblems. Each subproblem focuses on the separation of the two selected classes together with the remaining classes. Here, we utilize  $\{-1, 1\}$  to represent the label of the two selected classes and 0 to label the rest. Namely, the subproblem is encoded with the ternary output  $\{-1, 0, 1\}$ . The main idea of MNP-KSVC is to find a pair of nonparallel parametric-margin hyperplanes for each subproblem,

$$f_1(x) = w_1' x + b_1 = 0 \text{ and } f_2(x) = w_2' x + b_2 = 0, \quad (7)$$

such that each one is approximate to the corresponding class while as far as possible from the other class on one side. Moreover, the remaining classes are restricted in a region between these hyperplanes.

Formally, we use  $\mathcal{I}_k$  to express the set of indices for instances belonging to the  $k$  label in the subproblem, where  $k$  is in  $\{-1, 0, 1\}$ . Inspired by the TPMSVM [15], our MNP-KSVC considers the following two loss functions for the above ternary output learning problem:

$$\begin{aligned} \mathcal{R}_1^{emp} = & \nu_1 \sum_{j \in \mathcal{I}_-} (w_1' x_j + b_1) + c_1 \sum_{i \in \mathcal{I}_+} \max(0, 1 - (w_1' x_i + b_1)) \\ & + c_2 \sum_{l \in \mathcal{I}_0} \max(0, \varepsilon - 1 + (w_1' x_l + b_1)) \end{aligned} \quad (8)$$

and

$$\begin{aligned} \mathcal{R}_2^{emp} = & -\nu_2 \sum_{i \in \mathcal{I}_-} (w_2' x_i + b_2) + c_3 \sum_{j \in \mathcal{I}_+} \max(0, 1 + (w_2' x_j + b_2)) \\ & + c_4 \sum_{l \in \mathcal{I}_0} \max(0, \varepsilon - 1 - (w_2' x_l + b_2)), \end{aligned} \quad (9)$$

where  $\nu_1, \nu_2, c_1, c_2, c_3, c_4 > 0$  are penalty parameters and  $\varepsilon \in (0, 1]$  is a margin parameter. Introducing the regularization term  $\|w\|_2^2 + b^2$  yields the primal problems of MNP-KSVC:

$$\begin{aligned} \min_{w_1, \xi_1, \eta_1} \quad & \frac{1}{2}(\|w_1\|^2 + b_1^2) + \nu_1 \sum_{j \in \mathcal{I}_-} (w_1' x_j + b_1) + c_1 \sum_{i \in \mathcal{I}_+} \xi_{1i} + c_3 \sum_{l \in \mathcal{I}_0} \eta_{1l} \\ \text{s.t.} \quad & (w_1' x_i + b_1) \geq 1 - \xi_{1i}, \quad \xi_{1i} \geq 0, \\ & -(w_1' x_l + b_1) \geq \varepsilon - 1 - \eta_{1l}, \quad \eta_{1l} \geq 0 \end{aligned} \quad (10)$$

and

$$\begin{aligned} \min_{w_2, \xi_2, \eta_2} \quad & \frac{1}{2}(\|w_2\|^2 + b_2^2) - \nu_2 \sum_{i \in \mathcal{I}_+} (w_2' x_i + b_2) + c_2 \sum_{j \in \mathcal{I}_-} \xi_{2j} + c_4 \sum_{l \in \mathcal{I}_0} \eta_{2l} \\ \text{s.t.} \quad & -(w_2' x_j + b_2) \geq 1 - \xi_{2j}, \quad \xi_{2j} \geq 0, \\ & (w_2' x_l + b_2) \geq \varepsilon - 1 - \eta_{2l}, \quad \eta_{2l} \geq 0, \end{aligned} \quad (11)$$

where  $\xi_1, \eta_1, \xi_2, \eta_2$  are non-negative slack vectors.

To deliver the mechanism of MNP-KSVC, we now give the following analysis and geometrical explanation for problem (10):

- The first term is the  $L_2$ -norm of  $w_1$  and  $b_1$ . Minimizing this is done with the aim of regulating the model complexity of MNP-KSVC and avoiding over-fitting. Furthermore, this regularization term makes the QPPs strictly convex, leading to a unique solution.
- The second term is the sum of the projection value of the  $-1$  labeled instances  $x_{j \in \mathcal{I}_-}$  on  $f_1(x_j)$ . Optimizing this term leads instances  $x_{j \in \mathcal{I}_-}$  to be as far as possible from the  $+1$  labeled parametric-margin hyperplane  $f_1(x_j)$ .
- The third term with the first constraint requires the projection values of the  $+1$  labeled instances  $x_{i \in \mathcal{I}_+}$  on hyperplane  $f_1(x_j)$  to be not less than 1. Otherwise, a slack vector  $\xi_{1i}$  is introduced to measure its error when the constraint is violated.
- The last term with the second constraint aims for the projection values of the remaining 0 labeled instances  $x_{l \in \mathcal{I}_0}$  on hyperplane  $f_1(x_j)$  to be not more than  $1 - \varepsilon$ . Otherwise, a slack variable  $\eta_{1l}$  is utilized to measure the corresponding error. Optimizing this is done with the aim of keeping instances  $x_{l \in \mathcal{I}_0}$  at least  $\varepsilon$  distance from the  $+1$  labeled instances. Moreover,  $\varepsilon$  controls the margin between “+” and “0” labeled instances.

The geometrical explanation for problem (11) is similar. Let  $u_1 = [w_1 \ b_1]$ ,  $u_2 = [w_2 \ b_2]$ ,  $\tilde{x} = [x \ 1]$ . For the sake of simplicity, denote  $A = \{\tilde{x}\}_{i \in \mathcal{I}_+}$ ,  $B = \{\tilde{x}\}_{j \in \mathcal{I}_-}$  and  $C = \{\tilde{x}\}_{l \in \mathcal{I}_0}$  as the instances belonging to  $+1$ ,  $-1$ , and  $0$  labels, respectively. Then, the matrix formulations of problems (10) and (11) can be expressed as

$$\begin{aligned} \min_{u_1, \xi_1, \eta_1} \quad & \frac{1}{2}\|u_1\|^2 + \nu_1 e_-' B u_1 + c_1 e_+' \xi_1 + c_3 e_0' \eta_1 \\ \text{s.t.} \quad & A u_1 \geq e_+ - \xi_1, \quad \xi_1 \geq 0, \\ & -C u_1 \geq (\varepsilon - 1)e_0 - \eta_1, \quad \eta_1 \geq 0 \end{aligned} \quad (12)$$

and

$$\begin{aligned} \min_{u_2, \xi_2, \eta_2} \quad & \frac{1}{2}\|u_2\|^2 - \nu_2 e_+' A u_2 + c_2 e_-' \xi_2 + c_4 e_0' \eta_2 \\ \text{s.t.} \quad & -B u_2 \geq e_- - \xi_2, \quad \xi_2 \geq 0, \\ & C u_2 \geq (\varepsilon - 1)e_0 - \eta_2, \quad \eta_2 \geq 0. \end{aligned} \quad (13)$$

In what follows, we discuss the solutions of problems (12) and (13).

### 3.2. Model Optimization

To obtain solutions to problems (12) and (13), we first derive their dual problems by Theorem 1.

**Theorem 1.** *Optimization problems*

$$\begin{aligned} \min_{\alpha_1, \beta_1} \quad & \frac{1}{2} [\alpha'_1 \quad \beta'_1] \begin{bmatrix} AA' & -AC' \\ -CA' & CC' \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \\ & - (v_1 e'_- [BA' - BC'] + [e'_+ (\varepsilon - 1) e'_0]) \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{0} \leq \alpha_1 \leq c_1 e_+, \quad \mathbf{0} \leq \beta_1 \leq c_3 e_0 \end{aligned} \quad (14)$$

and

$$\begin{aligned} \min_{\alpha_2, \beta_2} \quad & \frac{1}{2} [\alpha'_2 \quad \beta'_2] \begin{bmatrix} BB' & -BC' \\ -CB' & CC' \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix} \\ & - (v_2 e'_+ [AB' - AC'] + [e'_- (\varepsilon - 1) e'_0]) \begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{0} \leq \alpha_2 \leq c_3 e_-, \quad \mathbf{0} \leq \beta_2 \leq c_4 e_0 \end{aligned} \quad (15)$$

are the dual problems of (12) and (13), respectively.

**Proof of Theorem 1.** Taking problem (12) as an example, firstly, we construct its Lagrangian function as

$$\begin{aligned} \mathcal{L}(\Xi_1) = \quad & \frac{1}{2} \|u_1\|^2 + v_1 e'_- B u_1 + c_1 e'_+ \xi_1 + c_3 e'_0 \eta_1 \\ & - \alpha'_1 (A u_1 + \xi_1 - e_+) - \beta'_1 (-C u_1 + \eta_1 - (\varepsilon - 1) e_0) \\ & - \varphi'_1 \xi_1 - \gamma_1 \eta_1, \end{aligned} \quad (16)$$

where  $\alpha_1, \beta_1, \varphi_1, \gamma_1$  are the non-negative Lagrange multipliers to constraints of problem (12), and  $\Xi_1 = \{u_1, \xi_1, \eta_1, \alpha_1, \beta_1, \varphi_1, \gamma_1\}$ . According to KKT conditions [3,37], the Lagrangian function (16) has to be maximized with its dual variables  $\alpha_1, \beta_1, \varphi_1, \gamma_1$  while being minimized with its primal variables  $u_1, \xi_1, \eta_1$ . Differentiate  $\mathcal{L}(\Xi_1)$  with respect to  $u_1, \eta_1, \xi_1$ ; then, optimal conditions of problem (12) are obtained by

$$\nabla \mathcal{L}_{u_1} = u_1 + v_1 B' e_- - A' \alpha_1 + C' \beta_1 = \mathbf{0}, \quad (17)$$

$$\nabla \mathcal{L}_{\xi_1} = c_1 e_+ - \alpha_1 - \varphi_1 = \mathbf{0}, \quad (18)$$

$$\nabla \mathcal{L}_{\eta_1} = c_3 e_0 - \beta_1 - \gamma_1 = \mathbf{0}, \quad (19)$$

$$\alpha'_1 (A u_1 + \xi_1 - e_+) = 0, \quad (20)$$

$$\beta'_1 (-C u_1 + \eta_1 - (\varepsilon - 1) e_0) = 0, \quad (21)$$

$$\varphi'_1 \xi_1 = 0, \quad (22)$$

$$\gamma_1 \eta_1 = 0. \quad (23)$$

From (17), we have

$$u_1 = A' \alpha_1 - C' \beta_1 - v_1 B' e_-. \quad (24)$$

Since  $\varphi_1, \gamma_1 \geq 0$ , from (18) and (19), we derive

$$\mathbf{0} \leq \alpha_1 \leq c_1 e_+ \quad \text{and} \quad \mathbf{0} \leq \beta_1 \leq c_3 e_0. \quad (25)$$

Finally, substituting (24) into the Lagrangian function (16) and using KKT conditions (17)–(23), the dual problem of (12) can be formulated as

$$\begin{aligned} \min_{\alpha_1, \beta_1} \quad & \frac{1}{2} [\alpha'_1 \ \beta'_1] \begin{bmatrix} AA' & -AC' \\ -CA' & CC' \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \\ & - (\nu_1 e'_- [BA' \ -BC'] + [e'_+ \ (\varepsilon - 1)e'_0]) \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{0} \leq \alpha_1 \leq c_1 e_+, \ \mathbf{0} \leq \beta_1 \leq c_3 e_0. \end{aligned} \quad (26)$$

Similarly, we can derive the dual problem of (13) as problem (15).  $\square$

For ease of notation, define  $q_1 = [\alpha_1; \beta_1]$ ,  $N_1 = [A; -C]$ ,  $h_1 = [e_+; (\varepsilon - 1)e_0]$ , and  $e_{q1} = [c_1 e_+; c_3 e_0]$ . Then, problem (14) can be succinctly reformulated as

$$\begin{aligned} \min_{q_1} \quad & \frac{1}{2} q'_1 H_1 q_1 - d_1 q_1 \\ \text{s.t.} \quad & \mathbf{0} \leq q_1 \leq e_{q1} \end{aligned} \quad (27)$$

where

$$H_1 = N_1 N'_1 = \begin{bmatrix} AA' & -AC' \\ -CA' & CC' \end{bmatrix} \quad (28)$$

and

$$d_1 = \nu_1 e'_- B N'_1 + h'_1 = \nu_1 e'_- [BA' \ -BC'] + [e'_+ \ (\varepsilon - 1)e'_0] \quad (29)$$

Similarly, define  $q_2 = [\alpha_2; \beta_2]$ ,  $N_2 = [-B; C]$ ,  $h_2 = [e_-; (\varepsilon - 1)e_0]$ , and  $e_{q2} = [c_2 e_-; c_4 e_0]$  for problem (15). Then, it can be reformulated as

$$\begin{aligned} \min_{q_2} \quad & \frac{1}{2} q'_2 H_2 q_2 - d_2 q_2 \\ \text{s.t.} \quad & \mathbf{0} \leq q_2 \leq e_{q2} \end{aligned} \quad (30)$$

where

$$H_2 = N_2 N'_2 = \begin{bmatrix} BB' & -BC' \\ -CB' & CC' \end{bmatrix} \quad (31)$$

and

$$d_2 = \nu_2 e'_+ A N'_2 + h'_2 = \nu_2 e'_+ [AB' \ -AC'] + [e'_- \ (\varepsilon - 1)e'_0] \quad (32)$$

After solving dual problems (27) and (30) by the standard QPP solver, we can obtain the solutions to primal problems (12) and (13) by Proposition 1 according to KKT conditions without proof.

**Proposition 1.** Suppose that  $q_1 = [\alpha_1; \beta_1]$  and  $q_2 = [\alpha_2; \beta_2]$  are solutions to dual problems (27) and (30), respectively. Then, solutions  $u_1$  and  $u_2$  to primal problems (12) and (13) can be formulated by

$$u_1 = N'_1 q_1 - \nu_1 B' e_- = A' \alpha_1 - C' \beta_1 - \nu_1 B' e_- \quad (33)$$

and

$$u_2 = -N'_2 q_2 + \nu_2 A' e_+ = -B' \alpha_2 + C' \beta_2 + \nu_2 A' e_+ \quad (34)$$

### 3.3. Decision Rule

As mentioned in Section 3.1, our MNP-KSVC decomposes the multiclass learning task into a series of subproblems with the “one-versus-one-versus-rest” strategy. Specifically, we construct  $K(K - 1)/2$  MNP-KSVC classifiers for  $K$ -class classification. For each  $(k_1, k_2)$ -pair classifier, we relabel the dataset with ternary outputs  $\{-1, 0, +1\}$  according to the two selected and the remaining class instances. Namely, we label “+1”, “−1”, and “0” to instances belonging to the  $k_i$  class,  $k_j$  class, and all the remaining classes, respectively. Then, we train it on the new relabeled dataset by solving problems (12) and (13).

As for the decision, we predict the label of an unseen instance  $x$  on the voting strategy from the ensemble results of  $K(K - 1)/2$  MNP-KSVC classifiers. Namely, we determine the vote via each classifier according to the regions in which  $x$  is located. Taking the  $(k_1, k_2)$ -pair classifier as



an example, firstly, we reformulate the parametric-margin hyperplanes (7) w.r.t.  $u = [w; b]$  and  $\tilde{x} = [x \ 1]$  as

$$f_1(x) = w'_1 x + b_1 = u'_1 \tilde{x} \text{ and } f_2(x) = w'_2 x + b_2 = u'_2 \tilde{x}, \quad (35)$$

If  $x$  is located above the “+” hyperplane  $f_1(x) > 1 - \varepsilon$ —i.e., satisfying the condition  $u'_1 \tilde{x} > 1 - \varepsilon$ —we vote it to the  $k_i$  class. On the other hand, if  $x$  is located below the “-” hyperplane  $f_2(x) < -1 + \varepsilon$ , we vote it to the  $k_j$  class. Otherwise, it belongs to the remaining classes. In summary, the decision function for the  $(k_1, k_2)$ -pair classifier can be expressed as

$$g(x) = \begin{cases} +1 & f_1(x) > 1 - \varepsilon, \\ -1 & f_2(x) < -1 + \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (36)$$

Finally, the given unseen instance  $x$  is assigned to the class label that gets the most votes. In summary, the whole procedure of MNP-KSVC is established in Algorithm 1 with the Figure 1.

---

**Algorithm 1** The procedure of MNP-KSVC

---

- 1: Input dataset  $\mathcal{T} = \{(x_i, y_i) | 1 \leq i \leq m\}$ , where  $x_i \in \mathbb{R}^n$  and  $y_i = \{1, \dots, K\}$ .
- 2: Choose parameters  $\nu_1, \nu_2, c_1, c_2, c_3, c_4 > 0$  and  $\varepsilon \in (0, 1]$ .
- Training Procedure:**
- 3: **for**  $k_i$  in range( $1, \dots, K$ ) **do**
- 4:   **for**  $k_j$  in range( $k_i + 1, \dots, K$ ) **do**
- 5:     Relabel “+1”, “-1”, and “0” to instances belonging to  $k_i, k_j$ , and the rest classes.
- 6:     Construct  $A, B$ , and  $C$  for problem (12) and (13) of  $(k_1, k_2)$ -pair classifier.
- 7:     Solve the corresponding dual problem (27) and (30) by QPP solver. Then, get their solution  $q_1 = [\alpha_1; \beta_1]$  and  $q_2 = [\alpha_2; \beta_2]$ .
- 8:     Build the auxiliary functions for classifier w.r.t. Proposition 1

$$k_1\text{-class} : f_1(x) = w'_1 x + b_1 = q'_1 N_1 x - \nu_1 e'_- Bx \quad (37)$$

and

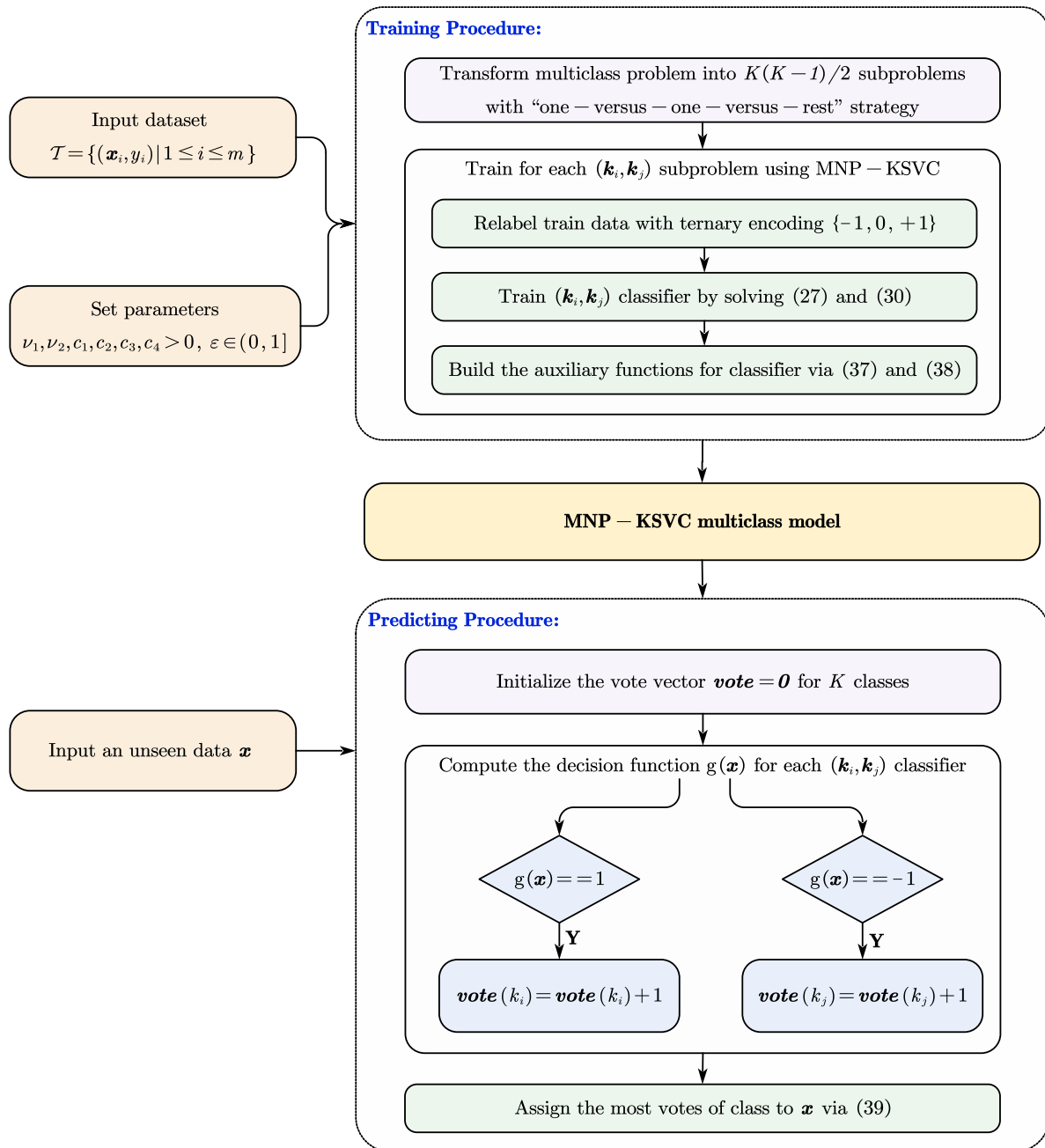
$$k_2\text{-class} : f_2(x) = w'_2 x + b_2 = -q'_2 N_2 x + \nu_2 e'_+ Ax \quad (38)$$

- 9:   **end for**
- 10: **end for**
- Predicting Procedure:**
- 11: For an unseen instance  $x$ , assign it to class  $y$  via the following vote strategy.
- 12: Initialize the vote vector  $vote = \mathbf{0}$  for  $K$  classes.
- 13: **for**  $k_i$  in range( $1, \dots, K$ ) **do**
- 14:   **for**  $k_j$  in range( $k_i + 1, \dots, K$ ) **do**
- 15:     Compute the decision function  $g(x)$  for the  $(k_1, k_2)$ -pair classifier w.r.t. (36).
- 16:     **if**  $g(x) == 1$  **then**
- 17:       Update  $vote(k_i) + = 1$
- 18:     **else if**  $g(x) == -1$  **then**
- 19:       Update  $vote(k_j) + = 1$
- 20:     **end if**
- 21:   **end for**
- 22: **end for**
- 23: Finally, assign the most votes of class to  $x$  via

$$\text{label}(x) \leftarrow \arg \max_{k=\{1, \dots, K\}} vote(k) \quad (39)$$


---





**Figure 1.** The flowchart for the training and predicting procedures of the proposed MNP-KSVC model.

#### 4. Model Extension to the Nonlinear Case

In practice, the linear classifier is sometimes not suitable for many real-world nonlinear learning tasks [18,21,30]. One of the effective solutions is to map linearly non-separable instances into the feature space. Thus, in this section, we focus on the nonlinear extension of MNP-KSVC.

To construct our nonlinear MNP-KSVC, we consider the feature mapping  $\mathbf{x}^\varphi = \varphi(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathcal{H}$  (RKHS, Reproducing Kernel Hilbert Space) with the kernel trick [3].

Define  $A^\phi = \{\phi(\tilde{x}_i)\}_{i \in \mathcal{I}_+}$ ,  $B^\phi = \{\phi(\tilde{x}_j)\}_{j \in \mathcal{I}_-}$ , and  $C^\phi = \{\phi(\tilde{x}_l)\}_{l \in \mathcal{I}_0}$ . Then, the nonlinear MNP-KSVC model optimizes the following two primal QPPs:

$$\begin{aligned} \min_{u_1, \xi_1, \eta_1} \quad & \frac{1}{2} \|u_1\|^2 + v_1 e'_- B_\phi u_1 + c_1 e'_+ \xi_1 + c_3 e'_0 \eta_1 \\ \text{s.t.} \quad & A_\phi u_1 \geq e_+ - \xi_1, \quad \xi_1 \geq 0, \\ & -C_\phi u_1 \geq (\varepsilon - 1)e_0 - \eta_1, \quad \eta_1 \geq 0 \end{aligned} \quad (40)$$

and

$$\begin{aligned} \min_{u_2, \xi_2, \eta_2} \quad & \frac{1}{2} \|u_2\|^2 - v_2 e'_+ A_\phi u_2 + c_2 e'_- \xi_2 + c_4 e'_0 \eta_2 \\ \text{s.t.} \quad & -B_\phi u_2 \geq e_- - \xi_2, \quad \xi_2 \geq 0, \\ & C_\phi u_2 \geq (\varepsilon - 1)e_0 - \eta_2, \quad \eta_2 \geq 0. \end{aligned} \quad (41)$$

where  $\xi_1, \eta_1, \xi_2, \eta_2$  are non-negative slack vectors. Because the formulations of nonlinear problems (40) and (41) are similar to the linear cases (12) and (13), we can obtain their solutions in a similar manner.

In what follows, we define the kernel operation for MNP-KSVC.

**Definition 1.** Suppose that  $K(\cdot, \cdot)$  is an appropriate kernel function; then, the kernel operation in matrix form is defined as

$$K(A, B) = \langle A_\phi, B_\phi \rangle, \quad (42)$$

whose  $ij$ -th element can be computed by

$$K(A, B)_{ij} = \phi(x_i)' \phi(x_j) = K(x_i, x_j). \quad (43)$$

Then, we can derive the dual problems of (40) and (41) by Theorem 2.

**Theorem 2.** Optimization problems

$$\begin{aligned} \min_{\alpha_1, \beta_1} \quad & \frac{1}{2} [\alpha'_1 \quad \beta'_1] \begin{bmatrix} K(A, A) & -K(A, C) \\ -K(C, A) & K(C, C) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \\ & - (v_1 e'_- [K(B, A) - K(B, C)] + [e'_+ (\varepsilon - 1)e'_0]) \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \\ \text{s.t.} \quad & 0 \leq \alpha_1 \leq c_1 e_+, \quad 0 \leq \beta_1 \leq c_3 e_0 \end{aligned} \quad (44)$$

and

$$\begin{aligned} \min_{\alpha_2, \beta_2} \quad & \frac{1}{2} [\alpha'_2 \quad \beta'_2] \begin{bmatrix} K(B, B) & -K(B, C) \\ -K(C, B) & K(C, C) \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix} \\ & - (v_2 e'_+ [K(A, B) - K(A, C)] + [e'_- (\varepsilon - 1)e'_0]) \begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix} \\ \text{s.t.} \quad & 0 \leq \alpha_2 \leq c_3 e_-, \quad 0 \leq \beta_2 \leq c_4 e_0 \end{aligned} \quad (45)$$

are the dual problems of (40) and (41), respectively.

**Proof of Theorem 2.** By introducing non-negative Lagrange multipliers  $\alpha_1, \beta_1, \varphi_1$ , and  $\gamma_1$  to constraints of problem (40), its Lagrangian function is built as

$$\begin{aligned} \mathcal{L}(\Xi_1) = \quad & \frac{1}{2} \|u_1\|^2 + v_1 e'_- B_\phi u_1 + c_1 e'_+ \xi_1 + c_3 e'_0 \eta_1 \\ & - \alpha'_1 (A_\phi u_1 + \xi_1 - e_+) - \beta'_1 (-C_\phi u_1 + \eta_1 - (\varepsilon - 1)e_0) \\ & - \varphi'_1 \xi_1 - \gamma_1 \eta_1, \end{aligned} \quad (46)$$

where  $\Xi_1 = \{u_1, \xi_1, \eta_1, \alpha_1, \beta_1, \varphi_1, \gamma_1\}$ . Differentiate  $\mathcal{L}(\Xi_1)$  with respect to  $u_1, \eta_1, \xi_1$ ; then, optimal conditions of problem (40) are obtained by

$$\nabla \mathcal{L}_{u_1} = u_1 + v_1 B'_\phi e_- - A'_\phi \alpha_1 + C'_\phi \beta_1 = 0, \quad (47)$$

$$\nabla \mathcal{L}_{\xi_1} = c_1 e_+ - \alpha_1 - \varphi_1 = 0, \quad (48)$$

$$\nabla \mathcal{L}_{\eta_1} = c_3 e_0 - \beta_1 - \gamma_1 = 0, \quad (49)$$

$$\alpha'_1 (A_\phi u_1 + \xi_1 - e_+) = 0, \quad (50)$$

$$\beta'_1 (-C_\phi u_1 + \eta_1 - (\varepsilon - 1)e_0) = 0, \quad (51)$$

$$\varphi'_1 \xi_1 = 0, \quad (52)$$

$$\gamma_1 \eta_1 = 0. \quad (53)$$

From (47), we have

$$u_1 = A'_\phi \alpha_1 - C'_\phi \beta_1 - v_1 B'_\phi e_-. \quad (54)$$

Since  $\varphi_1, \gamma_1 \geq 0$ , from (48) and (49), we derive

$$0 \leq \alpha_1 \leq c_1 e_+ \text{ and } 0 \leq \beta_1 \leq c_3 e_0. \quad (55)$$

Finally, substituting (54) into the Lagrangian function (16) and using KKT conditions (47)–(53), the dual problem of (40) can be formulated as (44). Similarly, we can derive the dual problem of (41) as problem (45).  $\square$

The procedure of the nonlinear MNP-KSVC is similar to that of the linear one, but with the following minor modifications in Algorithm 1:

- In contrast to some existing nonparallel SVMs, we do not need to consider the extra kernel-generated technique since only inner products appear in dual problems (14) and (15) of the linear MNP-KSVC. These dual formulations enable MNP-KSVC to behave consistently in linear and nonlinear cases. Thus, taking appropriate kernel functions instead of inner products in the Hessian matrix of dual problem (14) and (15), i.e.,

$$H_{\phi 1} = \begin{bmatrix} K(A, A) & -K(A, C) \\ -K(C, A) & K(C, C) \end{bmatrix} \text{ and } H_{\phi 2} = \begin{bmatrix} K(B, B) & -K(B, C) \\ -K(C, B) & K(C, C) \end{bmatrix}, \quad (56)$$

we can obtain the dual formation of the nonlinear MNP-KSVC in (44) and (45).

- Once we obtain the solution  $q_1 = [\alpha_1; \beta_1]$  and  $q_2 = [\alpha_2; \beta_2]$  to problems (44) and (45) respectively, the corresponding primal solutions  $u_1$  and  $u_2$  in feature space can be formulated by

$$u_1 = A'_\phi \alpha_1 - C'_\phi \beta_1 - v_1 B'_\phi e_- \quad (57)$$

and

$$u_2 = -B'_\phi \alpha_2 + C'_\phi \beta_2 + v_2 A'_\phi e_+ \quad (58)$$

- For an unseen instance  $x$ , construct the following decision functions for  $(i, j)$ -pair nonlinear MNP-KSVC classifier as

$$g_\phi(x) = \begin{cases} +1 & f_{\phi 1}(\tilde{x}) > 1 - \varepsilon, \\ -1 & f_{\phi 2}(\tilde{x}) < -1 + \varepsilon, \\ 0 & \text{otherwise,} \end{cases} \quad (59)$$

where  $\tilde{x} = [x \ 1]$ , the auxiliary functions  $f_{\phi 1}(x)$  and  $f_{\phi 2}(x)$  in feature space can be expressed as

$$f_{\phi 1}(\tilde{x}) = K(u_1, \tilde{x}) = \alpha'_1 K(A, \tilde{x}) - \beta'_1 K(C, \tilde{x}) - \nu_1 e'_- K(B, \tilde{x}) \quad (60)$$

and

$$f_{\phi 2}(\tilde{x}) = K(u_2, \tilde{x}) = -\alpha'_2 K(B, \tilde{x}) + \beta'_2 K(C, \tilde{x}) + \nu_2 e'_+ K(A, \tilde{x}) \quad (61)$$

## 5. Numerical Experiments

### 5.1. Experimental Setting

To demonstrate the validity of MNP-KSVC, we perform extensive experiments on several benchmark datasets that are commonly used for testing machine learning algorithms. In experiments, we focus on comparing MNP-KSVC and four state-of-the-art multiclass models—Multi-SVM, MBSVM, MTPMSVM, and Twin-KSVC—detailed as follows:

- Multi-SVM [38]: The idea is similar to the “one-versus-all” SVM [3]. However, it generates  $K$  binary SVM classifiers by solving the one large dual QPP. That is, the  $k$ -th classifier is trained with the  $k$ -th class instances encoded with positive labels and the remaining class instances with negative labels. Then, the label of an unseen instance is assigned by the “voting” scheme. The penalty parameter for each classifier in Multi-SVM is  $c$ .
- MBSVM [33]: It is the multiclass extension of the binary TWSVM, which is based on the “one-versus-all” strategy. MBSVM aims to find  $K$  nonparallel hyperplanes by solving  $K$  QPPs simultaneously. Specifically, the  $k$ -th class instances are as far away as the  $k$ -th hyperplane while the remaining instances are proximal to the  $k$ -th hyperplane. An unseen instance is assigned to the label depending on to which of the  $K$  hyperplanes it lies farthest away. The penalty parameter for each classifier in MBSVM is  $c$ .
- MTPMSVM: Inspired by MBSVM [33], we use the “one-versus-all” strategy to implement the multiclass version of TPMSVM [15] as a baseline. In contrast to MBSVM, it aims to find  $K$  parametric-margin hyperplanes, such that each hyperplane is closer to its corresponding class instances and as far away from the remaining class instances. The penalty parameters for each classifier in MTPMSVM are  $(\nu, c)$ .
- Twin-KSVC [35]: It is another novel multiclass extension of TWSVM. Twin-KSVC evaluates all the training instances in a “one-versus-one-versus-rest” structure with the ternary output  $\{-1, 0, +1\}$ . It aims to find a pair of nonparallel hyperplanes for each of two kinds of samples selected from  $K$  classes. The remaining class instances are mapped into a region within these two nonparallel hyperplanes. The penalty parameters for each classifier in Twin-KSVC are  $(c_1, c_2, c_3, c_4, \varepsilon)$ .

All methods are implemented by MATLAB on a PC with an i7 Intel Core processor with 32 GB RAM. The quadratic programming problems (QPPs) of all the classifiers are solved by the “quadprog” function in MATLAB. Now, we describe the setting of our experiments:

- Similar to [35,38], we use multiclass accuracy to measure each classifier, defined as

$$\text{accuracy} = \frac{1}{N} \sum_{k=1}^K \sum_{x \in \mathcal{I}_k} I(\hat{g}(x) = k) \quad (62)$$

where  $N$  is the scalar of the dataset,  $K$  is the total classes,  $\hat{g}(x)$  is the prediction of the classifier, and  $I(\cdot)$  is an indicator function, which returns 1 if the class matches and 0 otherwise. Moreover, we adopt training time to represent the learning efficiency.

- To reduce the complexity of parameter selection for multiclass classifiers, we use the same parameter setting for each learning subproblem. Specifically, we set the same for all classifiers  $c$  in MBSVM and MTPMSVM,  $\nu, c$  in MTPMSVM,  $c_1 = c_2, c_3 = c_4$

in Twin-KSVC, and  $v_1 = v_2, c_1 = c_2, c_3 = c_4$  in MNP-KSVC. For the nonlinear case, the RBF kernel  $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\gamma})$  is considered, where  $\gamma > 0$  is the kernel parameter.

- It is usually unknown beforehand which parameters are optimal for classifiers at hand. Thus, we employ the 10-fold cross-validation technique [3] for parameter selection. In detail, each dataset is randomly partitioned into 10 subsets with similar sizes and distributions. Then, the union of 9 subsets is used as the training set, while the remaining 1 is used as the testing set. Furthermore, we apply the grid-based approach [3] to obtain the optimal parameters of each classifier. Namely, the penalty parameters  $c, c_1, c_2, \nu, \nu_1$  and the kernel parameter  $\gamma$  are selected from  $\{2^i | i = -6, \dots, 6\}$ , while the margin parameter  $\varepsilon$  is chosen from  $\{i | i = 0.1, 0.2, \dots, 0.9\}$ . Once selected, we return them to learn the final decision function.

## 5.2. Result Comparison and Discussion

For comparison, we consider 10 real-world multiclass datasets from the UCI machine learning repository (the UCI datasets are available at <http://archive.ics.uci.edu/ml> (accessed on 10 September 2021)), whose statistics are summarized in Table 1. These datasets represent a wide range of domains (include phytology, bioinformatics, pathology, and so on), sizes (from 178 to 2175), features (from 4 to 34), and classes (from 3 to 10). All datasets are normalized before training such that features are transformed into  $[-1, 1]$ . Moreover, we carry out experiments as follows. Firstly, each dataset is divided into 2 subsets: 70% for training and 30% for testing. Then, we train classifiers with 10-fold cross-validation executions. Finally, we predict the testing set with the fine-tuning classifiers. Each experiment is repeated 10 times.

**Table 1.** Statistics for benchmark datasets used in experiments. # denotes the corresponding quantity.

Datasets	#Instances	#Training	#Testing	#Attributes	#Class
Balance	625	438	187	4	3
Ecoli	327	229	98	7	5
Iris	150	105	45	4	3
Glass	214	150	64	13	6
Wine	178	125	53	13	3
Thyroid	215	150	65	5	3
Dermatology	358	251	107	34	6
Shuttle	2175	1522	653	9	5
Contraceptive	1473	1031	442	9	3
Pen Based	1100	770	330	16	10

Table 2 and 3 contain a summary of learning results for the proposed MNP-KSVC model with other compared methods using linear and nonlinear kernels, respectively. The results on 10 benchmark datasets include the mean and standard of the testing multiclass accuracy (%), whose best performance is highlighted in bold. The comparison results reveal the following:

- MNP-KSVC yields better performance than other classifiers in terms of accuracy on almost all datasets. This confirms the efficacy of the proposed MNP-KSVC on the multiclass learning tasks.
- Nonparallel based models (MBSVM, MTPMSVM, Twin-KSVC, and MNP-KSVC) outperform the traditional SVM model. The reason is that SVM simply utilizes the parallel hyperplanes to learn the decision function, leading to less capability to capture the underlying multiclass distributions.
- MNP-KSVC has a better generalization ability than MBSVM and Twin-KSVC in most cases. For instance, MNP-KSVC obtains a higher accuracy (84.59%) than MBSVM (80.82%) and Twin-KSVC (80.12%) on the Ecoli dataset in the linear case. Similar results can be obtained from the other datasets. Since MBSVM and Twin-KSVC simply

implement empirical risk minimization, they are easy to overfit. The regularization terms are considered in our MNP-KSVC, which regulates the model complexity to avoid overfitting.

- MTPMSVM is another multiclass extension, which is based on the “one-versus-rest” strategy. With the help of the “hybrid classification and regression” learning paradigm, our MNP-KSVC can learn more multiclass discriminate information.
- Furthermore, we count the number of Superior/Inferior (W/L) instances to the compared classifier on all datasets for both linear and nonlinear cases, listed at the bottom of Tables 2 and 3. The results indicate that MNP-KSVC achieves the best results against others in terms of both W/L and average accuracy.

**Table 2.** Performance comparison on benchmark datasets for linear classifiers, in terms of **mean  $\pm$  std** of the testing multiclass accuracy (%). The Win/Loss (W/L) denotes the number of datasets for which MNP-KSVC is superior/inferior to the compared classifiers. Ave. Acc and rank denote each classifier’s average accuracy and rank over all datasets.

Datasets	Multi-SVM	MBSVM	MTPMSVM	Twin-KSVC	MNP-KSVC
Balance	79.91 $\pm$ 4.86	86.14 $\pm$ 4.63	87.43 $\pm$ 3.56	86.64 $\pm$ 1.92	<b>88.33 <math>\pm</math> 2.03</b>
Ecoli	72.32 $\pm$ 7.03	73.62 $\pm$ 4.95	73.77 $\pm$ 4.33	74.62 $\pm$ 3.86	<b>75.91 <math>\pm</math> 2.52</b>
Iris	93.24 $\pm$ 2.66	92.96 $\pm$ 2.24	93.21 $\pm$ 2.53	92.69 $\pm$ 3.24	<b>94.13 <math>\pm</math> 2.49</b>
Glass	69.18 $\pm$ 9.85	<b>72.89 <math>\pm</math> 7.28</b>	68.68 $\pm$ 6.79	71.65 $\pm$ 5.96	71.38 $\pm$ 5.38
Wine	93.24 $\pm$ 3.02	95.28 $\pm$ 1.46	<b>98.19 <math>\pm</math> 1.61</b>	97.23 $\pm$ 1.12	97.14 $\pm$ 1.27
Thyroid	90.24 $\pm$ 2.53	93.74 $\pm$ 1.58	92.92 $\pm$ 1.43	96.97 $\pm$ 1.08	<b>97.52 <math>\pm</math> 1.54</b>
Dermatology	81.82 $\pm$ 3.79	86.67 $\pm$ 1.69	84.46 $\pm$ 3.29	<b>90.37 <math>\pm</math> 2.32</b>	89.06 $\pm$ 3.16
Shuttle	71.58 $\pm$ 4.78	<b>84.04 <math>\pm</math> 2.92</b>	77.17 $\pm$ 3.3	78.76 $\pm$ 4.81	83.16 $\pm$ 1.92
Contraceptive	38.53 $\pm$ 3.76	43.95 $\pm$ 2.51	<b>44.65 <math>\pm</math> 2.93</b>	42.25 $\pm$ 2.45	44.22 $\pm$ 2.03
Pen Based	79.59 $\pm$ 3.75	85.94 $\pm$ 1.26	81.94 $\pm$ 2.04	83.21 $\pm$ 2.77	<b>86.78 <math>\pm</math> 1.37</b>
Ave. Acc	76.97	81.52	80.24	81.43	82.76
W/L	10/0	8/2	8/2	9/1	/
Ave. rank	4.6	2.9	3.1	2.67	1.7

To provide more statistical evidence [39–41], we have further performed the non-parametric Friedman test to check whether there are significant differences between MNP-KSVC and other compared classifiers. The bottom lines of Tables 2 and 3 list the average ranks of each classifier calculated by the Friedman test according to multiclass accuracy. Results show that MNP-KSVC has the first rank in linear and nonlinear cases, followed by NPSVM and RPTSVM successively. Now, we calculate the  $\mathcal{X}_F^2$  value for the Friedman test as

$$\mathcal{X}_F^2 = \frac{12 \times N}{k(k+1)} \left[ \sum_{i=1}^k r_i^2 - \frac{k(k+1)^2}{4} \right] \quad (63)$$

where  $N$  is the number of datasets,  $k$  is the number of classifiers, and  $r_i$  is the average rank on  $N$  datasets for the  $i$ -th model. For the linear case, we compute term  $\sum_{i=1}^k r_i^2$  in Table 2 as

$$\begin{aligned} \sum_{i=1}^k r_i^2 &= 4.6^2 + 2.9^2 + 3.1^2 + 2.7^2 + 1.7^2 \\ &\approx 49.1989 \end{aligned} \quad (64)$$

**Table 3.** Performance comparison on benchmark datasets for nonlinear classifiers, in terms of the **mean  $\pm$  std** of the testing multiclass accuracy (%). The Win/Loss (W/L) denotes the number of datasets for which MNP-KSVC is superior/inferior to the compared classifiers. Ave. Acc and rank denote each classifier’s average accuracy and rank over all datasets.

Datasets	Multi-SVM	MBSVM	MTPMSVM	Twin-KSVC	MNP-KSVC
Balance	79.94 $\pm$ 5.57	87.13 $\pm$ 4.57	89.92 $\pm$ 3.27	90.17 $\pm$ 4.08	<b>91.41 <math>\pm</math> 3.42</b>
Ecoli	79.81 $\pm$ 5.19	80.82 $\pm$ 4.25	<b>84.74 <math>\pm</math> 3.49</b>	80.12 $\pm$ 4.49	84.59 $\pm$ 3.83
Iris	90.26 $\pm$ 2.65	96.89 $\pm$ 1.83	97.39 $\pm$ 2.14	94.36 $\pm$ 2.03	<b>98.04 <math>\pm</math> 1.47</b>
Glass	58.66 $\pm$ 4.76	52.64 $\pm$ 4.08	62.73 $\pm$ 3.95	56.01 $\pm$ 4.26	<b>64.12 <math>\pm</math> 2.98</b>
Wine	94.36 $\pm$ 2.12	94.31 $\pm$ 1.87	<b>98.38 <math>\pm</math> 2.62</b>	97.26 $\pm$ 1.61	98.04 $\pm$ 1.45
Thyroid	91.82 $\pm$ 1.75	93.27 $\pm$ 1.95	95.34 $\pm$ 0.89	94.14 $\pm$ 1.05	<b>95.63 <math>\pm</math> 0.84</b>
Pen Based	86.53 $\pm$ 3.93	<b>89.51 <math>\pm</math> 3.29</b>	85.06 $\pm$ 3.68	86.12 $\pm$ 2.47	88.78 $\pm$ 2.68
Dermatology	84.43 $\pm$ 4.29	83.82 $\pm$ 3.86	84.51 $\pm$ 2.64	83.33 $\pm$ 3.29	<b>85.26 <math>\pm</math> 3.12</b>
Shuttle	74.36 $\pm$ 3.39	83.74 $\pm$ 2.73	87.06 $\pm$ 2.89	86.91 $\pm$ 2.38	<b>89.37 <math>\pm</math> 1.73</b>
Contraceptive	42.09 $\pm$ 4.95	44.28 $\pm$ 4.02	45.93 $\pm$ 3.85	<b>47.51 <math>\pm</math> 3.57</b>	47.47 $\pm$ 3.68
Ave. Acc	78.22	80.64	83.11	81.59	84.27
W/L	10/0	9/1	8/2	9/1	/
Ave. rank	4.3	3.69	2.3	3.3	1.4

Then, substituting  $k = 5$ ,  $N = 10$  and (64) into (63), we have

$$\chi_F^2 = \frac{12 \times 10}{5(5+1)} \left[ 49.1989 - \frac{5(5+1)^2}{4} \right] \approx 16.7956 \quad (65)$$

Based on the above Friedman statistic  $\chi_F^2 = 16.7956$ , we calculate the  $F$ -distribution statistic  $\mathcal{F}_F$  with  $(k - 1, (k - 1)(N - 1)) = (4, 36)$  degrees of freedom as

$$\mathcal{F}_F = \frac{(N - 1) \times \chi_F^2}{N(k - 1) - \chi_F^2} = \frac{9 \times 16.7956}{10 \times 4 - 16.7956} \approx 6.5143 \quad (66)$$

Moreover, we compute the  $p$ -value, which rejects the null hypothesis at the level of significance  $\alpha = 0.05$ . Similarly, we calculate the statistic for the nonlinear case, as summarized in Table 4. The results reject the null hypothesis for both linear and nonlinear cases and reveal the existence of significant differences in the performances of classifiers.

**Table 4.** Results of Friedman test on learning results.

	Statistic $\chi_F^2$	$p$ -Value	Hypothesis
Linear	6.5143	$2.9503 \times 10^{-4}$	reject
Nonlinear	10.2004	$1.2267 \times 10^{-5}$	reject

Furthermore, we record the average learning time of each classifier for the above UCI datasets experiments, as shown in Figures 2 and 3. The results show that our MNP-KSVC is faster than Multi-SVM and Twin-KSVC, while slightly slower than MBSVM and MTPMSVM for linear and nonlinear cases. Multi-SVM performs the slowest of all classifiers because Multi-SVM needs to solve larger problems than the nonparallel-based classifiers. Moreover, the Hessian matrix of dual QPPs in MNP-KSVC avoids the time-costly matrix inversion, leading to greater effectiveness than Twin-KSVC. Overall, the above results confirm the feasibility of Twin-KSVC.



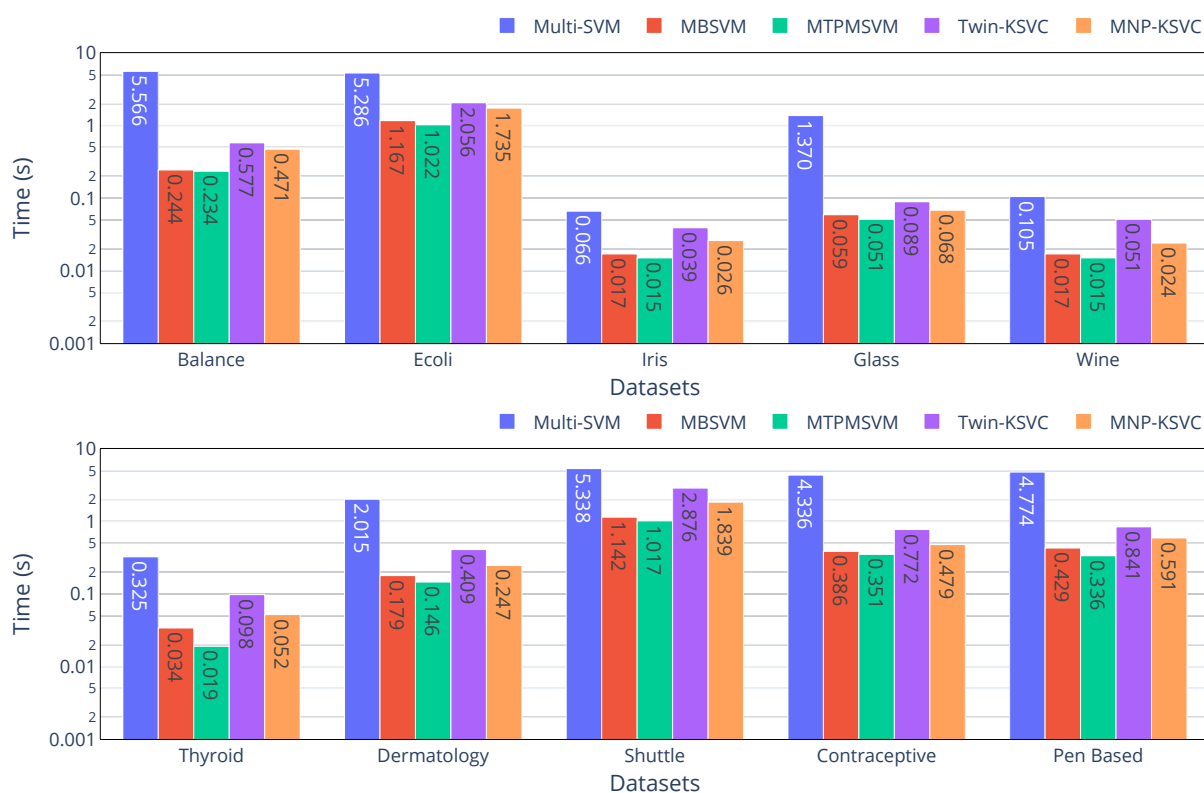


Figure 2. The learning times on benchmark datasets for linear classifiers.

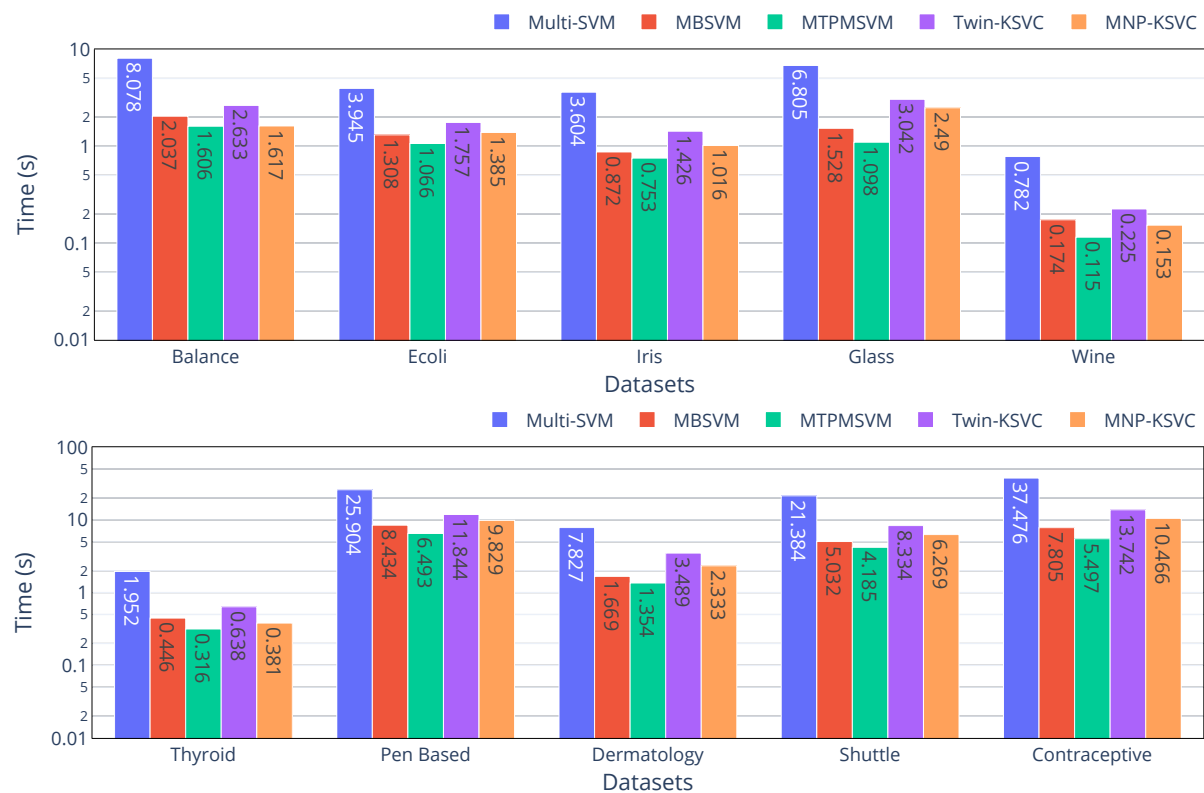


Figure 3. The learning times on benchmark datasets for nonlinear classifiers.

## 6. Discussion and Future Work

This paper proposes a novel  $K$  multiclass nonparallel parametric-margin support vector machine, termed MNP-KSVC. Specifically, our MNP-KSVC has the following attractive merits:

- For the  $K$ -class learning task, our MNP-KSVC first transforms the complicated multiclass problem into  $K(K-1)/2$  subproblems via a “one-versus-one-versus-rest” strategy. Each subproblem focuses on separating the two selected classes and the rest of the classes. That is, we utilize  $\{-1, 1\}$  to represent the label of the two selected classes and 0 to label the rest. Unlike the “one-versus-all” strategy used in Multi-SVM, MBSVM, and MTPMSVM, this encoding strategy can alleviate the imbalanced issues that sometimes occur in multiclass learning [32,35].
- For each subproblem, our MNP-KSVC aims to learn a pair of nonparallel parametric-margin hyperplanes (36) with the ternary encoding  $\{-1, 0, +1\}$ . These parametric-margin hyperplanes are closer to their corresponding class and at least one distance from the other class. Meanwhile, they restrict the rest of the instances into an insensitive region. A hybrid classification and regression loss joined with the regularization is further utilized to formulate the optimization problems (10) and (11) of MNP-KSVC.
- Moreover, the nonlinear extension is also presented to deal with the nonlinear multiclass learning tasks. In contrast to MBSVM [33] and Twin-KSVC [35], the linear and nonlinear models in MNP-KSVC are consistent. Applying the linear kernel in the nonlinear problems (44) and (45) results in the same formulations as the original linear problems (14) and (15).
- Extensive experiments on various datasets demonstrate the effectiveness of the proposed MNP-KSVC compared with Multi-SVM, MBSVM, MTPMSVM, and Twin-KSVC.

There are several interesting directions to research in the future, such as extensions to semi-supervised learning [26,42], multi-label learning [22], and privilege-information learning [43].

**Author Contributions:** Conceptualization, S.-W.D. and W.-J.C.; Funding acquisition, W.-J.C. and Y.-H.S.; Investigation, M.-C.Z. and P.C.; Methodology, S.-W.D., M.-C.Z. and Y.-H.S.; Project administration, S.-W.D. and W.-J.C.; Supervision, W.-J.C. and Y.-H.S.; Validation, M.-C.Z. and H.-F.S.; Visualization, P.C. and H.-F.S.; Writing—original draft, S.-W.D., P.C. and H.-F.S.; Writing—review and editing, S.-W.D. and W.-J.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, Grant number: 61603338, 11871183, 61866010, and 11426202; the Natural Science Foundation of Zhejiang Province of China, Grant number: LY21F030013; the Natural Science Foundation of Hainan Province of China, Grant number: 120RC449; the Scientific Research Foundation of Hainan University, Grant number: kyqd(sk)1804.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing not applicable. No new data were created or analyzed in this study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

QPP	Quadratic Problem Programming
KKT	Karush–Kuhn–Tucker
SVM	Support Vector Machine
TWSVM	Twin Support Vector Machine
Multi-SVM	Multiclass Support Vector Machine
MBSVM	Multiple Birth Support Vector Machine
MTPMSVM	Multiple Twin Parametric Margin Support Vector Machine
Twin-KSVC	Multiclass Twin Support Vector Classifier
MNP-KSVC	Multiclass Nonparallel Parametric-Margin Support Vector Classifier

## References

1. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: San Francisco, CA, USA, 2012.
2. Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
3. Deng, N.; Tian, Y.; Zhang, C. *Support Vector Machines: Theory, Algorithms and Extensions*; CRC Press: Philadelphia, PA, USA, 2013.
4. Sitaula, C.; Aryal, S.; Xiang, Y.; Basnet, A.; Lu, X. Content and context features for scene image representation. *Knowl.-Based Syst.* **2021**, *232*, 107470. [[CrossRef](#)]
5. Ma, S.; Cheng, B.; Shang, Z.; Liu, G. Scattering transform and LSPTSVM based fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.* **2018**, *104*, 155–170. [[CrossRef](#)]
6. Liu, T.; Yan, D.; Wang, R.; Yan, N.; Chen, G. Identification of Fake Stereo Audio Using SVM and CNN. *Information* **2021**, *12*, 263. [[CrossRef](#)]
7. You, S.D. Classification of Relaxation and Concentration Mental States with EEG. *Information* **2021**, *12*, 187. [[CrossRef](#)]
8. Kang, J.; Han, X.; Song, J.; Niu, Z.; Li, X. The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data. *Comput. Biol. Med.* **2020**, *120*, 103722. [[CrossRef](#)]
9. Lazcano, R.; Salvador, R.; Marrero-Martin, M.; Leporati, F.; Juarez, E.; Callico, G.M.; Sanz, C.; Madronal, D.; Florimbi, G.; Sancho, J.; et al. Parallel Implementations Assessment of a Spatial-Spectral Classifier for Hyperspectral Clinical Applications. *IEEE Access* **2019**, *7*, 152316–152333. [[CrossRef](#)]
10. Fabelo, H.; Ortega, S.; Szolna, A.; Bulters, D.; Pineiro, J.F.; Kabwama, S.; J-O'Shanahan, A.; Bulstrode, H.; Bisshopp, S.; Kiran, B.R.; et al. In-Vivo Hyperspectral Human Brain Image Database for Brain Cancer Detection. *IEEE Access* **2019**, *7*, 39098–39116. [[CrossRef](#)]
11. Roy, S.D.; Debbarma, S. A novel OC-SVM based ensemble learning framework for attack detection in AGC loop of power systems. *Electr. Power Syst. Res.* **2022**, *202*, 107625. [[CrossRef](#)]
12. Mangasarian, O.L.; Wild, E.W. Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 69–74. [[CrossRef](#)] [[PubMed](#)]
13. Jayadeva.; Khemchandani, R.; Chandra, S. Twin Support Vector Machines for Pattern Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 905–910. [[CrossRef](#)] [[PubMed](#)]
14. Ding, S.; Hua, X. An overview on nonparallel hyperplane support vector machine algorithms. *Neural Comput. Appl.* **2014**, *25*, 975–982. [[CrossRef](#)]
15. Peng, X. TPMSVM: A novel twin parametric-margin support vector machine for pattern recognition. *Pattern Recogn.* **2011**, *44*, 2678–2692. [[CrossRef](#)]
16. Arun Kumar, M.; Gopal, M. Least squares twin support vector machines for pattern classification. *Expert Syst. Appl.* **2009**, *36*, 7535–7543. [[CrossRef](#)]
17. Shao, Y.; Zhang, C.; Wang, X.; Deng, N. Improvements on Twin Support Vector Machines. *IEEE Trans. Neural Netw.* **2011**, *22*, 962–968. [[CrossRef](#)]
18. Chen, W.J.; Shao, Y.H.; Li, C.N.; Liu, M.Z.; Wang, Z.; Deng, N.Y.  $\nu$ -projection twin support vector machine for pattern classification. *Neurocomputing* **2020**, *376*, 10–24. [[CrossRef](#)]
19. Tian, Y.; Qi, Z.; Ju, X.; Shi, Y.; Liu, X. Nonparallel Support Vector Machines for Pattern Classification. *IEEE Trans. Cybern.* **2014**, *44*, 1067–1079. [[CrossRef](#)]
20. Tian, Y.; Ping, Y. Large-scale linear nonparallel support vector machine solver. *Neural Netw.* **2014**, *50*, 166–174. [[CrossRef](#)] [[PubMed](#)]
21. Chen, W.; Shao, Y.; Li, C.; Wang, Y.; Liu, M.; Wang, Z. NPrSVM: Nonparallel sparse projection support vector machine with efficient algorithm. *Appl. Soft Comput.* **2020**, *90*, 106142. [[CrossRef](#)]
22. Chen, W.; Shao, Y.; Li, C.; Deng, N. MLTSVM: A novel twin support vector machine to multi-label learning. *Pattern Recogn.* **2016**, *52*, 61–74. [[CrossRef](#)]
23. Bai, L.; Shao, Y.H.; Wang, Z.; Chen, W.J.; Deng, N.Y. Multiple Flat Projections for Cross-Manifold Clustering. *IEEE Trans. Cybern.* **2021**, 1–15. Available online: <https://ieeexplore.ieee.org/document/9343292> (accessed on 20 October 2021).

24. Hou, Q.; Liu, L.; Zhen, L.; Jing, L. A novel projection nonparallel support vector machine for pattern classification. *Eng. Appl. Artif. Intel.* **2018**, *75*, 64–75. [[CrossRef](#)]
25. Liu, L.; Chu, M.; Gong, R.; Zhang, L. An Improved Nonparallel Support Vector Machine. *IEEE Trans. Neur. Net. Lear.* **2021**, *32*, 5129–5143. [[CrossRef](#)]
26. Chen, W.; Shao, Y.; Deng, N.; Feng, Z. Laplacian least squares twin support vector machine for semi-supervised classification. *Neurocomputing* **2014**, *145*, 465–476. [[CrossRef](#)]
27. Shao, Y.; Chen, W.; Zhang, J.; Wang, Z.; Deng, N. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recogn.* **2014**, *47*, 3158–3167. [[CrossRef](#)]
28. Chen, W.; Shao, Y.; Ning, H. Laplacian smooth twin support vector machine for semi-supervised classification. *Int. J. Mach. Learn. Cyber.* **2014**, *5*, 459–468. [[CrossRef](#)]
29. Gao, Z.; Fang, S.C.; Gao, X.; Luo, J.; Medhin, N. A novel kernel-free least squares twin support vector machine for fast and accurate multi-class classification. *Knowl.-Based Syst.* **2021**, *226*, 107123. [[CrossRef](#)]
30. Ding, S.; Zhao, X.; Zhang, J.; Zhang, X.; Xue, Y. A review on multi-class TWSVM. *Artif. Intell. Rev.* **2017**, *52*, 775–801. [[CrossRef](#)]
31. Qiang, W.; Zhang, J.; Zhen, L.; Jing, L. Robust weighted linear loss twin multi-class support vector regression for large-scale classification. *Signal Process.* **2020**, *170*, 107449. [[CrossRef](#)]
32. de Lima, M.D.; Costa, N.L.; Barbosa, R. Improvements on least squares twin multi-class classification support vector machine. *Neurocomputing* **2018**, *313*, 196–205. [[CrossRef](#)]
33. Yang, Z.; Shao, Y.; Zhang, X. Multiple birth support vector machine for multi-class classification. *Neural Comput. Appl.* **2013**, *22*, 153–161. [[CrossRef](#)]
34. Angulo, C.; Parra, X.; Català, A. K-SVCR. A support vector machine for multi-class classification. *Neurocomputing* **2003**, *55*, 57–77. [[CrossRef](#)]
35. Xu, Y.; Guo, R.; Wang, L. A Twin Multi-Class Classification Support Vector Machine. *Cogn. Comput.* **2013**, *5*, 580–588. [[CrossRef](#)]
36. Nasiri, J.A.; Moghadam Charkari, N.; Jalili, S. Least squares twin multi-class classification support vector machine. *Pattern Recogn.* **2015**, *48*, 984–992. [[CrossRef](#)]
37. Mangasarian, O.L. *Nonlinear Programming*; SIAM Press: Philadelphia, PA, USA, 1993.
38. Tomar, D.; Agarwal, S. A comparison on multi-class classification methods based on least squares twin support vector machine. *Knowl.-Based Syst.* **2015**, *81*, 131–147. [[CrossRef](#)]
39. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *200*, 675–701. [[CrossRef](#)]
40. Yu, Z.; Wang, Z.; You, J.; Zhang, J.; Liu, J.; Wong, H.S.; Han, G. A New Kind of Nonparametric Test for Statistical Comparison of Multiple Classifiers Over Multiple Datasets. *IEEE Trans. Cybern.* **2017**, *47*, 4418–4431. [[CrossRef](#)] [[PubMed](#)]
41. Hatamlou, A. Black hole: A new heuristic optimization approach for data clustering. *Inform. Sci.* **2013**, *222*, 175–184. [[CrossRef](#)]
42. Chen, W.; Shao, Y.; Xu, D.; Fu, Y. Manifold proximal support vector machine for semi-supervised classification. *Appl. Intell.* **2014**, *40*, 623–638. [[CrossRef](#)]
43. Li, Y.; Sun, H.; Yan, W.; Cui, Q. R-CTSVM+: Robust capped L1-norm twin support vector machine with privileged information. *Inform. Sci.* **2021**, *574*, 12–32. [[CrossRef](#)]