J. Nonlinear Var. Anal. 8 (2024), No. 6, pp. 953-971
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.8.2024.6.07

SPARSE LEAST SQUARES K-SVCR MULTI-CLASS CLASSIFICATION

HOSSEIN MOOSAEI

Department of Informatics, Faculty of Science,
Jan Evangelista Purkyně University, Ústí nad Labem, Czech Republic

Abstract. This paper introduces a novel model, the sparse least squares K-class support vector classification-regression with adaptive ℓ_p-norm (PLSTKSVC), to tackle challenges in multi-class classification. Leveraging a "1-versus-1-versus-rest" structure, PLSTKSVC dynamically adjusts the parameter p based on the data, enabling an adaptive learning framework. By incorporating cardinality-constrained optimization, the model seamlessly integrates feature selection and classification. Although the ℓ_p-norm is non-convex for $0 < p < 1$, PLSTKSVC efficiently addresses the associated optimization via linear systems of equations. PLSTKSVC offers several advantages, including simultaneous feature selection and classification, robust theoretical foundations, algorithmic efficiency, and strong empirical validation. The model's theoretical contributions include lower bounds on non-zero solution entries and upper bounds on the optimal solution norm. Experimental results on multi-class classification datasets highlight the superior performance of PLSTKSVC, establishing it as a significant advancement in machine learning.

Keywords. Feature selection; Least squares K-class support vector classification-regression; Multi-class classification; Support vector machines; Sparse optimization; ℓ_p-norm.

1. INTRODUCTION

Support vector machines (SVM), initially introduced for binary classification in [8, 11], have found widespread applications in various domains. It has been successfully utilized in face recognition [14], heart disease detection [3], energy prediction [1, 36, 39], Raman spectroscopy [15], and biomedicine [35], among others. The core concept of SVM involves identifying the maximum margin between two hyperplanes, addressing a constrained convex quadratic programming problem (QPP). Over the past decades, SVM, along with other binary classification algorithms, has seen numerous variants, extensions, and applications [4, 12, 13, 19, 28, 30, 31, 32, 33, 38].

Variants and extensions of binary classification algorithms often limit their direct applicability to practical scenarios, where multi-class classification is frequently encountered [37]. Two commonly employed strategies come into play when addressing multi-class classification within the SVM framework. The first strategy, known as "1-versus-1," involves constructing $\frac{k(k-1)}{2}$ binary classifiers [21]. However, this approach may yield suboptimal results as it overlooks some training samples in the training process of each classifier. The second strategy, "1-versus-rest,"

*Corresponding author.
E-mail address: hmoosaei@gmail.com; hossein.moosaei@ujep.cz (H. Moosaei).
Received 30 May 2024; Accepted 12 September 2024; Published online 25 October 2024.

constructs K binary classifiers, with each classifier encompassing all the training samples of a specific class [17]. While this strategy ensures that every class is represented in the training of at least one classifier, it introduces the potential challenge of class imbalance.

Angulo et al. [2] introduced an innovative approach known as K-class support vector classification-regression (K-SVCR) for addressing K-class classification problems. Utilizing a "1-versus-1-versus-rest" structure with ternary outputs $\{-1,0,+1\}$ and constructing $\frac{k(k-1)}{2}$ classifiers, each trained with the entire dataset, K-SVCR effectively mitigates information loss and class imbalance risks. As a result, K-SVCR surpasses SVM methods in multi-class classification scenarios. Moosaei and Hladík [27, 29] later introduced a modified version of K-SVCR, termed the least squares K-class support vector classification-regression machine (LSK-SVCR). In this adaptation, they converted inequality constraints into equality constraints and opted for a 2-norm instead of a 1-norm to minimize slack variables in the primal problem of K-SVCR. This intelligent adjustment leads to an efficient algorithm characterized by robust generalization performance.

Feature selection is crucial to data mining and machine learning, especially in high-dimensional applications. This pivotal step has gained substantial attention for reducing data dimensionality, improving classification accuracy by utilizing only relevant data, and accelerating the learning process. There are two widely recognized approaches to feature selection. The first approach involves a two-stage process, where significant features are initially chosen from the original feature set. In the second stage, classification algorithms are applied to the refined datasets [10]. Contrastingly, the second approach seamlessly integrates feature selection with classification. Here, the solution obtained during the classification phase exhibits the desired sparse characteristic [24].

The problem of cardinality-constrained optimization has attracted considerable attention across diverse fields, including optimization, machine learning, computational finance, and operations management [6, 25]. Despite its practical relevance, solving optimization problems with cardinality constraints poses challenges due to their non-convex and discontinuous nature. The optimization problem constrained by cardinality is formally represented as follows:

$$\begin{aligned} &\min_x f(x) \\ &\text{subject to } \|x\|_0 < k, \\ &x \in X. \end{aligned} \tag{1.1}$$

In this context, f signifies the objective function, where $k > 0$ is a designated positive integer, and $X \subset \mathbb{R}^n$ is a subset that could encompass additional constraints on x. It is crucial to emphasize that the stipulation $k < n$ is indispensable to uphold the cardinality condition. The cardinality-constrained problem aims to tackle the intricate task of minimizing two factors in sparse optimization problems, specifically, both f and $\|x\|_0$. Alternatively, an adjusted version of the problem is introduced, amalgamating the two criteria:

$$\begin{aligned} &\min_x f(x) + \rho\|x\|_0 \\ &\text{subject to } x \in X, \end{aligned} \tag{1.2}$$

where $\rho > 0$ serves as a regularization parameter. It is important to recognize that the ℓ_0-norm is an integer-valued, non-convex, and non-smooth function. Optimization problems involving $\|x\|_0$ are known to be NP-hard [34], making problem (1.2) computationally challenging. Considering the connection between $\|x\|_0$ and $\|x\|_p$ with $0 < p < 1$ [9], an alternative problem is

introduced:

$$\min_x f(x) + \rho \|x\|_p \text{ s.t. } x \in X. \tag{1.3}$$

This formulation explores the approximation of $\|x\|_0$ by $\|x\|_p$ with $0 < p < 1$, providing a potential avenue for addressing the computational challenges associated with cardinality-constrained optimization problems.

This study introduces an innovative optimization-based approach that seamlessly integrates feature selection and classification. This is achieved by leveraging the sparse optimization problem (1.3) and proposing a novel model called the sparse least squares K-class support vector classification-regression machine with adaptive ℓ_p-norm, referred to as PLSTKSVC. Our model incorporates an adaptive learning technique with ℓ_p-norm, where $0 < p < 1$. Notably, the parameter p is dynamically selected by the model based on the available data. Despite the non-convex and non-smooth nature of ℓ_p-norm for $0 < p < 1$, we address these challenges through techniques that approximate these terms with their convex and smooth counterparts, enhancing the overall approximations of the underlying problems. The solution approach involves solving systems of linear equations, rendering our method simple and efficient.

Key advantages of our proposed approach include:

- **Simultaneous feature selection and classification:** Our novel approach minimizes the p-norms of hyperplane weights in the objective function, enabling concurrent feature selection and classification simultaneously.
- **Theoretical Foundations:** We present fundamental theorems and properties of the proposed model, encompassing lower bounds on non-zero entries in solutions and upper bounds on the norm of the optimal solution.
- **Algorithmic Efficiency:** The introduced algorithm for finding sparse solutions to non-smooth and non-convex optimization problems accelerates the training process by solving a series of linear systems of equations, contributing to computational efficiency.
- **Empirical Validation:** Experiments on multi-class classification datasets validate the effectiveness of our proposed method in both feature selection and classification tasks.

In summary, PLSTKSVC emerges as a powerful tool, demonstrating outstanding classification performance while utilizing fewer features compared to its counterparts.

The remainder of this paper is structured as follows: Section 2 provides a succinct overview of K-support vector classification regression (K-SVCR) and least squares K-SVCR. Section 3 details our sparse LSK-SVCR method. Section 4 presents our solving strategy. Section 5 discusses the nonlinear case. Section 6 covers the decision rule. Experimental results demonstrating the efficiency of our proposed algorithm on UCI datasets are presented in Section 7. The paper concludes with final remarks in Section 8.

Notations. Let $a = [a_i]$ symbolize a vector in $\mathbb{R}^n$. If f denotes a real-valued function defined on the n-dimensional real space $\mathbb{R}^n$, the gradient of f concerning x is represented by $\frac{\partial f}{\partial x}$, which manifests as a column vector in $\mathbb{R}^n$. The transpose of a matrix A is denoted as A^T. For two vectors x and y existing in the n-dimensional real space, their scalar product is denoted as $x^T y$. The 2-norm of a vector $x \in \mathbb{R}^n$ is denoted as $\|x\|$. The column vector consisting of ones with arbitrary dimension is denoted by e. In the context of matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times l}$, the kernel $k(A,B)$ is a versatile function mapping $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times l}$ to $\mathbb{R}^{m\times l}$. Specifically, for column vectors x and y in $\mathbb{R}^n$ and matrix $A \in \mathbb{R}^{m\times n}$, $k(x^T, y)$ results in a real number, $k(x^T, A^T)$ yields a

row vector in $\mathbb{R}^m$, and $k(A,A^T)$ generates an $m \times m$ matrix. The identity matrix of size $n \times n$ is denoted as I_n, and the matrix operation $[A;B]$ merges matrices A and B as illustrated below:

$$[A;\, B] = \begin{bmatrix} A \\ B \end{bmatrix}.$$

2. Background

This section serves as a foundation, introducing two critical concepts: K-support vector classification regression and least squares K-SVCR. These fundamental elements are essential for understanding the upcoming development and application of our approach.

2.1. **K-support vector classification-regression.** The K-SVCR approach, first introduced by Angulo et al. [2], offers a novel perspective on multi-class classification with ternary outputs $\{-1,0,+1\}$. It presents a support vector classification-regression machine specifically designed for K-class classification. In the decomposition phase, K-SVCR evaluates all training data within a "1-versus-1-versus-rest" framework. This evaluation is performed using a hybrid support vector machine (SVM) that integrates both classification and regression techniques. A visual representation of the K-SVCR method is provided in Figure 1.

The K-SVCR formulation can be expressed as a convex quadratic programming problem of the following form:

$$\begin{aligned} \min_{w,b,\zeta_1,\zeta_2,\phi,\phi^*} \quad & \frac{1}{2}\|w\|^2 + c_1(e_1^T\zeta_1 + e_2^T\zeta_2) + c_2 e_3^T(\phi+\phi^*) \\ \text{subject to} \quad & Aw + e_1 b \geq e_1 - \zeta_1, \\ & -(Bw + e_2 b) \geq e_2 - \zeta_2, \\ & -\delta e_3 - \phi^* \leq Cw + e_3 b \leq \delta e_3 + \phi, \\ & \zeta_1, \zeta_2, \phi, \phi^* \geq 0, \end{aligned} \tag{2.1}$$

where c_1 and c_2 denote the regularization parameters, ζ_1, ζ_2, ϕ, and ϕ^* stand for positive slack variables, and e_1, e_2, and e_3 represent vectors of ones with appropriate dimensions. To ensure no overlap, it is essential for the positive parameter δ to be strictly less than 1.

The dual formulation of (2.1) can be written as:

$$\begin{aligned} \max_{\gamma} \quad & q^T\gamma - \frac{1}{2}\gamma^T H\gamma, \\ \text{subject to} \quad & 0 \leq \gamma \leq F, \end{aligned} \tag{2.2}$$

where $Q = \begin{bmatrix} A^T & -B^T & C^T & -C^T \end{bmatrix}$, $H = Q^T Q$, $q = \begin{bmatrix} e_1; & e_2; & -\delta e_3; & -\delta e_3 \end{bmatrix}$, and $F = \begin{bmatrix} c_1 e_1; & c_1 e_2; & c_2 e_3; & c_2 e_3 \end{bmatrix}$. Through the solution of the above quadratic optimization problem with box constraints, we can derive the separating hyperplane given by $f(x) = w^T x + b$.

2.2. **Least squares K-SVCR.** Moosaei and Hladík introduced the least squares version of the K-SVCR method and named LSK-SVCR in [27]. This algorithm evaluates the training points in a structure "1-versus-1-versus-rest" with ternary outputs $\{-1,0,+1\}$. We will shortly discuss the linear case, and the results can be extended for the nonlinear case as did in [27].

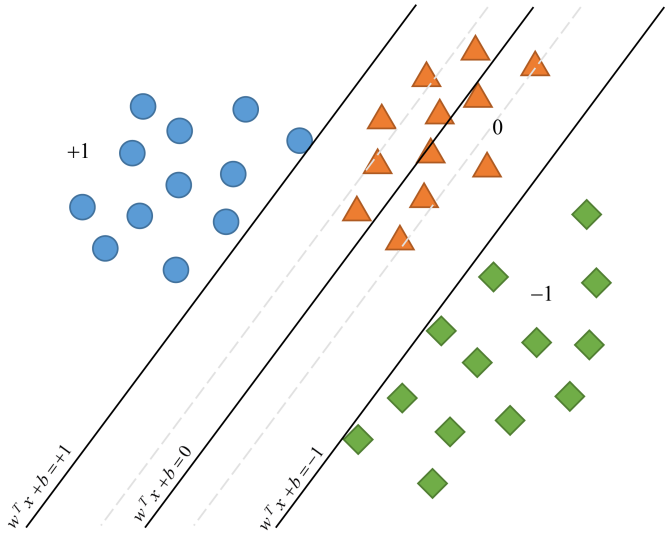

FIGURE 1. The depiction of K-SVCR's geometric properties [27].

The LSK-SVCR model can be presented as follows:

$$\begin{aligned} \min_{w,b,\zeta_1,\zeta_2,\phi,\phi^*} \quad & \frac{1}{2}\|w\|^2 + c_1\left(\|\zeta_1\|^2 + \|\zeta_2\|^2\right) + c_2\|\phi\|^2 + c_3\|\phi^*\|^2 \\ \text{subject to} \quad & e_1 - (Aw + e_1 b) = \zeta_1, \\ & e_2 + (Bw + e_2 b) = \zeta_2, \\ & Cw + e_3 b - \delta e_3 = \phi^*, \\ & -Cw - e_3 b - \delta e_3 = \phi. \end{aligned} \tag{2.3}$$

Here, ζ_1, ζ_2, ϕ, and ϕ^* denote positive slack variables, while c_1, c_2, and c_3 represent penalty parameters. Additionally, the positive parameter δ is constrained to be less than 1 to prevent overlap.

Upon substituting the constraints into the objective function, the solution to this problem can be obtained by solving a linear system of equations, depicted as follows:

$$\begin{aligned} \begin{bmatrix} w \\ b \end{bmatrix} = & -\left[c_1 E^T E + c_1 F^T F + c_2 G^T G + c_3 G^T G\right]^{-1} \\ & (-c_1 E^T e_1 + c_1 F^T e_2 + c_2 \delta G^T e_3 - c_3 \delta G^T e_3), \end{aligned} \tag{2.4}$$

where $E = [A\ e_1]$, $F = [B\ e_2]$, and $G = [C\ e_3]$.

Algorithm 1 outlines the classification process utilizing Linear LSK-SVCR.

3. SPARSE LSK-SVCR

In this section, we explore sparse LSK-SVCR, a novel approach that integrates the principles of sparsity with LSK-SVCR. By combining these elements, the methodology aims to improve the efficiency and interpretability of classification tasks. We denote this introduced method as p-norm LSK-SVCR (PLSK-SVCR).

Algorithm 1 Linear LSK-SVCR

Input: A, B, and C represent datasets with class labels $+1$, -1, and 0 respectively

1: Construct augmented datasets E, F, and G by appending a column of ones to A, B, and C respectively.
2: Choose penalty parameters c_1, c_2, c_3, and a threshold parameter δ.
3: Compute the hyperplane parameters (w,b) using the method described in (2.4).
4: Classify data points into classes $+1$, -1, or 0 based on the decision function (6.1).

Initially, we enhance the primal problem of K-SVCR (2.1) to (3.1), wherein we modify the objective function to incorporate the square of the 2-norm of slack variables ζ_1, ζ_2, ϕ, and ϕ^* instead of their 1-norm. Additionally, we refine the regularization term from $\frac{1}{2}\|w\|^2$ to a more generalized form $\frac{c}{2}(\|w\|^2+b^2)$. Consequently, the ensuing minimization problem is presented.

$$\begin{aligned}\min_{w,b,\zeta_1,\zeta_2,\phi,\phi^*}\quad & \frac{c}{2}(\|w\|^2+b^2)+\frac{c_1}{2}(\|\zeta_1\|^2+\|\zeta_2\|^2)+\frac{c_2}{2}\|\phi\|^2+\frac{c_3}{2}\|\phi^*\|^2 \\ \text{subject to}\quad & e_1-(Aw+e_1b)=\zeta_1,\\ & e_2+(Bw+e_2b)=\zeta_2,\\ & Cw+e_3b-\delta e_3=\phi^*,\\ & -Cw-e_3b-\delta e_3=\phi,\end{aligned} \tag{3.1}$$

where $c>0$, $c_1>0$, and $c_2>0$ represent the regularization parameters, and ζ_1, ζ_2, ϕ, and ϕ^* denote positive slack variables. We aim to discover the most sparse solution to the problem (3.1), which entails minimizing the number of non-zero components and the objective functions. To achieve this objective, we replace $\|w\|^2$ with $\|w\|_0$ in the aforementioned problem, leading to the following challenging problem:

$$\begin{aligned}\min_{w,b,\zeta_1,\zeta_2,\phi,\phi^*}\quad & \frac{c_1}{2}(\|\zeta_1\|^2+\|\zeta_2\|^2)+\frac{c_2}{2}\|\phi\|^2+\frac{c_3}{2}\|\phi^*\|^2+\frac{c}{2}(\|w\|_0+b^2) \\ \text{subject to}\quad & e_1-(Aw+e_1b)=\zeta_1,\\ & e_2+(Bw+e_2b)=\zeta_2,\\ & Cw+e_3b-\delta e_3=\phi^*,\\ & -Cw-e_3b-\delta e_3=\phi.\end{aligned} \tag{3.2}$$

Hence, by substituting the constraints into the objective function, problem (3.2) is transformed into the subsequent unconstrained optimization problem:

$$\begin{aligned}\min_{w,b}\ & \frac{c_1}{2}\|e_1-Aw-e_1b\|^2+\frac{c_1}{2}\|e_2+Bw+e_2b\|^2\\ & +\frac{c_2}{2}\|-Cw-e_3b-\delta e_3\|^2+\frac{c_3}{2}\|Cw+e_3b-\delta e_3\|^2+\frac{c}{2}(\|w\|_0+b^2).\end{aligned} \tag{3.3}$$

The l_0 norm $\|x\|_0$ is inherently related to the l_p norm $\|x\|_p$ with $0<p<1$, as demonstrated by the following relationship [9]:

$$\|x\|_0=\lim_{p\to 0}\|x\|_p^p=\lim_{p\to 0}\sum_{i=1}^{n}|x_i|^p. \tag{3.4}$$

Additionally, as shown in [28], it is established that within broad assumptions, the $\|x\|_p$ formulation provides a tight approximation of $\|x\|_0$. As the parameter p approaches zero, the problems converge, resulting in equivalence in terms of both optimal value and optimal solutions.

Hence, within the objective functions of both problems (3.2) and (3.3), the expression $\|w\|_0$ can be estimated by the l_p norm, denoted as $\|w\|_p^p$ where $0 < p < 1$. This approximation leads us to address the following problem instead of tackling (3.3):

$$\begin{aligned}\min_{w,b}\ \Phi(w,b) = \min_{w,b}\ & \frac{c_1}{2}\|e_1 - Aw - e_1 b\|^2 + \frac{c_1}{2}\|e_2 + Bw + e_2 b\|^2 \\ & + \frac{c_2}{2}\|-Cw - e_3 b - \delta e_3\|^2 + \frac{c_3}{2}\|Cw + e_3 b - \delta e_3\|^2 + \frac{c}{2}(\|w\|_p^p + b^2). \end{aligned} \tag{3.5}$$

Following the methods outlined in [26], we aim to derive lower bounds on the absolute values of the non-zero components in the optimal solution. Specifically, we seek to define these lower and upper bounds so that any component of the optimal solution within these bounds is necessarily zero. Drawing inspiration from the work in [26], we formulate and prove the following theorem.

Theorem 3.1. *Let (w^*, b^*) denote an optimal solution to problem (3.5), and let (w_0, b_0) represent an arbitrary point. For any $i \in \{1, 2, \ldots, n\}$, if w_i^* lies in the interval $(-I, I)$, then we have $w_i^* = 0$, where*

$$I = \left[\frac{\frac{c}{2} p}{(c_1\|A\| + c_1\|B\| + (c_2 + c_3)\|C\|)\sqrt{\Phi(w_0, b_0)}} \right]^{\frac{1}{1-p}}.$$

Proof. Let k be the number of non-zero elements in w^*, i.e., $k = \|w^*\|_0$. Assume, without loss of generality, that the optimal solution has the form $w^* = (w_1^*, w_2^*, \ldots, w_k^*, 0, \ldots, 0)^T$. Let $Z^* = (w_1^*, w_2^*, \ldots, w_k^*)^T$ be nonzero components of the optimal solution. The optimization problem (3.5) can be written as:

$$\begin{aligned}\min_{Z,b}\ F(Z,b) = & \frac{c_1}{2}\|(e_1 - (\tilde{A}Z + e_1 b))\|^2 + \frac{c_1}{2}\|(e_2 + \tilde{B}Z + e_2 b)\|^2 \\ & + \frac{c_2}{2}\|(-\delta e_3 - \tilde{C}Z - e_3 b)\|^2 + \frac{c_3}{2}\|(-\delta e_3 + \tilde{C}Z + e_3 b)\|^2 + \frac{c}{2}(\|Z\|_p^p + b_1^2). \end{aligned} \tag{3.6}$$

If (Z^*, b^*) constitutes a local optimal solution for (3.6), it must adhere to the first-order necessary condition. Given that all components of Z^* are non-zero, the function $F(Z,b)$ attains differentiability at (Z^*, b^*), necessitating that the first derivative with respect to Z vanishes at (Z^*, b^*). This condition is expressed as:

$$\begin{aligned} 0 = & -c_1\tilde{A}^T(e_1 - (\tilde{A}Z^* + e_1 b^*)) + c_1\tilde{B}^T(e_2 + \tilde{B}Z^* + e_2 b^*) \\ & - c_2\tilde{C}^T(-\delta e_3 - \tilde{C}Z^* - e_3 b^*) + c_3\tilde{C}^T(-\delta e_3 + \tilde{C}Z^* + e_3 b^*) + \frac{c}{2}(\nabla_Z^* \|Z^*\|_p^p). \end{aligned} \tag{3.7}$$

By utilizing the knowledge that $\nabla_Z \|Z^*\|_p^p = p\,\mathrm{diag}(|Z^*|)^{p-1}\,\mathrm{sgn}(Z^*)$, we can rewrite (3.7) as:

$$\begin{aligned} \frac{c}{2} p(\mathrm{diag}(|Z^*|)^{p-1}\,\mathrm{sgn}(Z^*)) = & c_1\tilde{A}^T(e_1 - (\tilde{A}Z^* + e_1 b^*)) - c_1\tilde{B}^T(e_2 + \tilde{B}Z^* + e_2 b^*) \\ & + c_2\tilde{C}^T(-\delta e_3 - \tilde{C}Z^* - e_3 b^*) - c_3\tilde{C}^T(-\delta e_3 + \tilde{C}Z^* + e_3 b^*). \end{aligned}$$

Thus we have

$$\begin{aligned}\frac{c}{2}p\|(\operatorname{diag}(|Z^*|)^{p-1}\operatorname{sgn}(Z^*))\| \\ = \|c_1\tilde{A}^T(e_1-(\tilde{A}Z^*+e_1b^*))-c_1\tilde{B}^T(e_2+\tilde{B}Z^*+e_2b^*) \\ +c_2\tilde{C}^T(-\delta e_3-\tilde{C}Z^*-e_3b^*)-c_3\tilde{C}^T(-\delta e_3+\tilde{C}Z^*+e_3b^*)\|.\end{aligned} \tag{3.8}$$

It follows that

$$\begin{aligned}&\frac{c}{2}p\|\operatorname{diag}(|Z^*|)^{p-1}\| \\ &\leqslant c_1\|\tilde{A}^T\|\cdot\|(e_1-(\tilde{A}Z^*+e_1b^*))\|+c_1\|\tilde{B}^T\|\cdot\|(e_2+\tilde{B}Z^*+e_2b^*)\| \\ &\quad +c_2\|\tilde{C}^T\|\cdot\|(-\delta e-\tilde{C}Z^*-eb^*)\|+c_2\|\tilde{C}^T\|\cdot\|(-\delta e_3+\tilde{C}Z^*+e_3b^*)\| \\ &\leqslant (c_1\|A^T\|+c_1\|B^T\|+(c_2+c_3)\|C^T\|)\sqrt{F(w_0,b_0)}.\end{aligned} \tag{3.9}$$

Then we obtain

$$\begin{aligned}\frac{c}{2}p\min_{1\leqslant i\leqslant k}|Z_i^*|^{p-1} &\leqslant \frac{c}{2}p\|\operatorname{diag}(|Z^*|)^{p-1}\| \\ &\leqslant (c_1\|A\|+c_1\|B\|+(c_2+c_3)\|C\|)\sqrt{F(w_0,b_0)}.\end{aligned}$$

Finally, we can conclude that:

$$|Z_i^*|\geqslant\min_{1\leqslant i\leqslant k}|Z_i^*|\geqslant\left(\frac{\frac{c}{2}p}{(c_1\|A\|+c_1\|B\|+(c_2+c_3)\|C\|)\sqrt{F(w_0,b_0)}}\right)^{\frac{1}{1-p}}. \tag{3.10}$$

Hence, for any local optimal solution (w^*,b^*) to problem (3.5), if $w_i^*\in(-I,I)$, it follows that $w_i^*=0$, for $i=1,\ldots,n$. This conclusion completes the proof. □

The theorem above has demonstrated a connection between the selection of penalty parameters, the parameter p, and the sparsity of the solution.

Naturally, 0 represents a lower bound, and n (the number of features) serves as an upper bound for $\|w^*\|_0$. However, in [26], an upper bound for the number of nonzero components of the optimal solution is proposed. This indicates at least how many features can be removed, as illustrated in the following theorem.

Corollary 3.1. *Let (w^*,b^*) represent an optimal solution of problem (3.5), and (w_0,b_0) be any arbitrary point. Then, it follows that:*

$$\|w^*\|_0\leqslant\min\left\{n,\frac{2\Phi(w_0,b_0)}{cI^p}\right\}.$$

Now, we introduce a new upper bound for $\|w^*\|_0$ in the following theorem.

Theorem 3.2. *Suppose that $\tilde{A}\in\mathbb{R}^{m_1\times n}$, $\tilde{B}\in\mathbb{R}^{m_2\times n}$, and $\tilde{C}\in\mathbb{R}^{m_3\times n}$ are matrices defined in Theorem 3.1. Then, let $m=m_1+m_2+m_3$. It follows that $\|w^*\|_0\leq m$.*

Proof. In the proof of Theorem 3.1, we assumed that (Z^*,b^*) is a local optimal solution of the problem (3.6). Consequently, the second-order necessary conditions are satisfied at (Z^*,b^*), indicating that the Hessian matrix is positive semi-definite. Specifically, the Hessian matrix is given by

$$\nabla_Z^2F(Z^*,b^*)=c_1\tilde{A}^T\tilde{A}+c_1\tilde{B}^T\tilde{B}+(c_2+c_3)\tilde{C}^T\tilde{C}+\frac{c}{2}p(p-1)\operatorname{diag}(|Z^*|^{p-2}), \tag{3.11}$$

which is positive semi-definite. Additionally, we know

$$\frac{c}{2}p(p-1)\,\mathrm{diag}(|Z^*|^{p-2}) \text{ is negative.}$$

Thus $c_1\tilde{A}^T\tilde{A}+c_1\tilde{B}^T\tilde{B}+(c_2+c_3)\tilde{C}^T\tilde{C}$ must be positive definite for any chosen $c_1,c_2,c_3>0$. By setting $c_1=1$ and $c_2=c_3=\frac{1}{2}$, we conclude that $\tilde{A}^T\tilde{A}+\tilde{B}^T\tilde{B}+\tilde{C}^T\tilde{C}$ is positive definite. Consequently, the matrix $[\tilde{A};\tilde{B};\tilde{C}]$ is invertible, implying that its columns are linearly independent. This leads to the conclusion that the number of its columns cannot exceed the number of its rows, i.e., $\|w^*\|_0 \leq m$. □

4. Solving Strategy

We begin by demonstrating the attainability of the solution to the optimization problem defined by (3.5) within PLSK-SVCR.

Theorem 4.1. *For any chosen value of p where $0<p<1$, the optimal solution to the optimization problem (3.5) can be achieved.*

Proof. Given $\Phi(w,b) \geqslant \frac{c}{2}(\|w\|_p^p+b^2)$, the objective function is bounded below. Moreover, as $\|w\|\to\infty$, $\Phi(w,b)\to\infty$. Hence, the objective function is both continuous and coercive, ensuring the existence of an optimal solution (see [5]). □

Despite the existence of the optimal solution to problem (3.5), its attainment poses a formidable challenge, primarily due to the presence of the non-smooth and non-convex term $\|w\|_p^p$ in the objective function. To address the challenge presented by non-differentiability, we implement the approach of replacing the non-smooth term with a smooth counterpart. Thus, drawing inspiration from [22, 28], we undertake the approximation of $\sum_{i=1}^n(|[w]_i|+\varepsilon_0)^p$, where $\varepsilon_0>0$ is a very small number. This value will gradually approach zero, enabling us to achieve an approximation of $\|\cdot\|_p^p$. As a consequence, the problem defined by (3.5) is approximated by the subsequent smooth formulation

$$\begin{aligned}\min_{w,b}\ &\frac{c_1}{2}\|e_1-Aw-e_1b\|^2+\frac{c_1}{2}\|e_2+Bw+e_2b\|^2+\frac{c_2}{2}\|-Cw-e_3b-\delta e_3\|^2\\&+\frac{c_3}{2}\|Cw+e_3b-\delta e_3\|^2+\frac{c}{2}\left(\sum_{i=1}^n(|[w]_i|+\varepsilon_0)^p+b^2\right).\end{aligned}\tag{4.1}$$

In problem (4.1), although the objective function is differentiable, its non-convexity arises from the inclusion of the term $\sum_{i=1}^n(|[w]_i|+\varepsilon_0)^p$ for $0<p<1$. To address this concern, this non-convex term is substituted with convex term $\|\beta\otimes w\|_2^2$, where β is a variable that can be adjusted to achieve the desired approximation.

Hence, we can formulate the subsequent optimization problem, which exhibits both smoothness and convexity

$$\begin{aligned}\min_{w,b}\ &\frac{c_1}{2}\|e_1-Aw-e_1b\|^2+\frac{c_1}{2}\|e_2+Bw+e_2b\|^2+\frac{c_2}{2}\|-Cw-e_3b-\delta e_3\|^2\\&+\frac{c_3}{2}\|Cw+e_3b-\delta e_3\|^2+\frac{c}{2}(\|\beta\otimes w\|^2+b^2).\end{aligned}\tag{4.2}$$

Within the context of the previously mentioned optimization problem, the computation of β plays a crucial role. To enhance the accuracy of approximations, we intend to introduce an iterative procedure aimed at achieving refined forms like (4.2) for addressing problem (4.1).

To proceed, we start by initializing $\beta^{(0)}$ as $(\beta_1^{(0)},\dots,\beta_n^{(0)})^T$. This allows us to obtain solution $(w^{(0)},b^{(0)})$ for the problem described in (4.2) using the initial value $\beta=\beta^{(0)}$.

In iteration k, let us assume that we have estimated (w,b) as $(w^{(k)},b^{(k)})$. Now, consider obtaining the solution $(w^{(k+1)},b^{(k+1)})$ for the following optimization problem:

$$\begin{aligned}\min_{w,b}\ &\frac{c_1}{2}\|e_1-Aw-e_1b\|^2+\frac{c_1}{2}\|e_2+Bw+e_2b\|^2+\frac{c_2}{2}\|-Cw-e_3b-\delta e_3\|^2\\&+\frac{c_3}{2}\|Cw+e_3b-\delta e_3\|^2+\frac{c}{2}\big(\|\beta^{(k+1)}\otimes w\|^2+b^2\big),\end{aligned}\tag{4.3}$$

where $\beta^{(k+1)}=(\beta_1^{(k+1)},\dots,\beta_n^{(k+1)})^T$ is the weight vector.

To determine $\beta^{(k+1)}$ in the aforementioned problem, we assume that the problems (4.1) and (4.3) have the same steepest descent direction at the current points $(w^{(k)},b^{(k)})$. Then, $\beta^{(k+1)}$ must satisfy the following equation:

$$p|w_i^{(k)}+\varepsilon_0|^{p-1}\operatorname{sgn}(w_i^{(k)})=2(\beta_i^{(k+1)})^2w_i^{(k)}.\tag{4.4}$$

Consequently, we arrive at viable choices for $\beta^{(k+1)}$. To avoid division by zero, we incorporate a very small value ε_0 into the denominator, resulting in the following expression:

$$\beta_i^{(k+1)}=\sqrt{\frac{p|w_i^{(k)}+\varepsilon_0|^{p-1}}{2|w_i^{(k)}|+\varepsilon_0}}.\tag{4.5}$$

The Lagrangian function for problem (4.3) is defined as follows:

$$\begin{aligned}L(w,b)=&\frac{c_1}{2}\|e_1-Aw-e_1b\|^2+\frac{c_1}{2}\|e_2+Bw+e_2b\|^2+\frac{c_2}{2}\|-Cw-e_3b-\delta e_3\|^2\\&+\frac{c_3}{2}\|Cw+e_3b-\delta e_3\|^2+\frac{c}{2}(\|\beta' w\|^2+b^2),\end{aligned}$$

where $\beta'=\operatorname{diag}(\beta_1^{(k+1)},\dots,\beta_n^{(k+1)})$. The necessary and sufficient Karush-Kuhn-Tucker (KKT) conditions are outlined by:

$$\begin{aligned}\frac{\partial L}{\partial w}=&\,c_1(-A^T)(e_1-Aw-e_1b)+c_1B^T(e_2+Bw+e_2b)\\&+c_2(-C^T)(-Cw-e_3b-\delta e_3)+c_3C^T(Cw+e_3b-\delta e_3)+c\beta'^2w=0,\\\frac{\partial L}{\partial b}=&\,c_1(-e_1^T)(e_1-Aw-e_1b)+c_1e_2^T(e_2+Bw+e_2b)\\&+c_2(-e_3^T)(-Cw-e_3b-\delta e_3)+c_3e_3^T(Cw+e_3b-\delta e_3)+cb=0.\end{aligned}$$

The aforementioned KKT conditions for problem (4.3) result in the solutions of the linear system of equations:

$$\begin{aligned}\begin{bmatrix}w\\b\end{bmatrix}=&-\left[c_1E^TE+c_1F^TF+c_2G^TG+c_3G^TG+cD_1\right]^{-1}\\&(-c_1E^Te_1+c_1F^Te_2+c_2\delta G^Te_3-c_3\delta G^Te_3),\end{aligned}\tag{4.6}$$

where $E=[A\ e_1]$, $F=[B\ e_2]$, and $G=[C\ e_3]$, and $D_1=\operatorname{diag}\big((\beta_1^{(k+1)})^2,\dots,(\beta_n^{(k+1)})^2,1\big)$. This leads us to an estimation of the solution for problem (3.5). The utilization of Theorem 3.1 becomes instrumental in selecting relevant features and identifying the non-zero components

Algorithm 2 Linear l_p-norm Least Squares K-SVCR (PLSK-SVCR)

Input: Give matrices $A \in \mathbb{R}^{m_1 \times n}$ of class $+1$, $B \in \mathbb{R}^{m_2 \times n}$ of class -1, and $C \in \mathbb{R}^{m_3 \times n}$ of class 0; Form matrices $E = [A\ e_1]$, $F = [B\ e_2]$, and $G = [C\ e_3]$; Select appropriate parameters c_1, c_2, c_3, c, parameters $p, \varepsilon \in (0,1)$, and a very small positive number $\varepsilon_0 > 0$.

Output:

- The optimal solution for problem (3.5).
- The sets of chosen feature indices, as derived from Theorem 3.1 :

$$F' = \{i : |w_i^*| > I_i\}, \quad i = 1, \ldots, n.$$

- To categorize a new point x_i, apply decision function (6.1) and make the decision based on this rule; here, each w_i consists of components from F', and the components of x_i correspond to those in the feature set F' associated with w_i.

The process:

1: Construct the optimization problem (3.5). Initiate $\beta^{(0)}$ randomly and set $k = 1$. Compute the solution $(w^{(k)}, b^{(k)})$ using (4.6), then update $\beta^{(k+1)}$ as per (4.5).

2: The stopping criteria: If $\|\beta^{(k+1)} - \beta^{(k)}\| < \varepsilon_0$, terminate and provide the optimal solution $(w^*, b^*) = (w^{(k)}, b^{(k)})$ for problem (3.5). If not satisfied, update $k = k+1$ and return to step 1.

within the solutions. Building upon the explanations provided above, we proceed to introduce an algorithm, referred to as Algorithm 2, that serves to demonstrate the implementation of our proposed method, PLSK-SVCR, effectively. This algorithm outlines the steps and processes involved in applying the PLSK-SVCR approach in a clear and systematic manner.

5. NONLINEAR

In real-world situations, using a linear kernel might not suffice for effectively separating many classification tasks. To address non-linear problems, samples are often transformed into a higher-dimensional feature space. In this subsection, we extend the linear PLSK-SVCR approach to handle non-linear cases. Our goal is to identify the following kernel surface:

$$k(x^T, D^T)w + b = 0,$$

where $k(\cdot,\cdot)$ is a kernel function and $D = [A; B; C]$.

The optimization problem (3.5) can be reformulated into its nonlinear primal form as follows:

$$\begin{aligned} \min_{w,b} \ \Psi(w,b) = {} & \frac{1}{2} c_1 \|e_1 - k(A, D^T)w - e_1 b\| + c_1 \|e_2 + k(B, D^T)w + e_2 b\| \\ & + c_2 \| - k(C, D^T)w - e_3 b - \delta e_3\|^2 + c_3 \|k(C, D^T)w + e_3 b - \delta e_3\| + \frac{c}{2}(\|w\|_p^p + b^2). \end{aligned} \quad (5.1)$$

All preceding theorems can be adjusted to accommodate nonlinear scenarios. Here, we highlight one of them without providing proof, as the proof methodology is akin to its employed in the linear case. The remaining theorems presented in the linear context can be similarly tailored to suit the nonlinear setting. The following theorem provides insight into determining the non-zero components of any optimal solution of the nonlinear K-SVCR. Its proof closely resembles that of Theorem 3.1, hence it is omitted here.

Algorithm 3 Nonlinear l_p-norm Least Squares K-SVCR (PLSK-SVCR)

Input: Give matrices $A \in \mathbb{R}^{m_1 \times n}$ of class $+1$, $B \in \mathbb{R}^{m_2 \times n}$ of class -1, $C \in \mathbb{R}^{m_3 \times n}$ of class 0, and and $D = [A; B; C]$. Choose a kernel function K. Form matrices $M = [k(A, D^T)\ e_1] \in R^{m_1 \times (m+1)}$, $N = [k(B, D^T)\ e_2] \in R^{m_2 \times (m+1)}$, $P = [k(C, D^T)\ e_3] \in R^{m_3 \times (m+1)}$. Select appropriate parameters c_1, c_2, c_3, c, parameters $p, \varepsilon \in (0,1)$, a very small positive number $\varepsilon_0 > 0$, and also the parameter of the kernel γ.

Output:

- The optimal solution for problem (5.1).
- The sets of chosen feature indices, as derived from Theorem 5.1.
- To categorize a new point x_i, apply decision function (6.2) and make the decision based on this rule.

The process:

1: Construct the optimization problem (5.1). Initiate $\beta^{(0)}$ randomly and set $k = 1$. Compute the solution $(w^{(k)}, b^{(k)})$ using (5.2), then update $\beta^{(k+1)}$ as per (4.5).

2: The stopping criteria: If $\|\beta^{(k+1)} - \beta^{(k)}\| < \varepsilon_0$, terminate and provide the optimal solution $(w^*, b^*) = (w^{(k)}, b^{(k)})$ for problem (5.1). If not satisfied, update $k = k+1$ and return to step 1.

Theorem 5.1. *Let (w^*, b^*) denote an optimal solution of problem (5.1), and (w_0, b_0) represent an arbitrary point. For any $i \in \{1, 2, \ldots, n\}$, if $w_i^* \in (-L, L)$, then $w_i^* = 0$, where*

$$L = \left[\frac{\frac{c}{2} p}{(2c_1 \|k(A, D^T)\| + 2c_1 \|k(B, D^T)\| + (c_2 + c_3)\|k(C, D^T)\|) \sqrt{\Psi(w_0, b_0)}} \right]^{\frac{1}{1-p}}.$$

Problem (5.1) can be approached iteratively by employing a methodology analogous to that used in the linear case. This iterative process is particularly effective when dealing with convex optimization problems, as it allows us to gradually refine the solution by leveraging key principles from the linear scenario.

In this context, we adopt an approach where, just like in the linear case, the solution to this convex optimization problem is obtained by solving a sequence of subproblems. Each subproblem incrementally improves the objective function by updating the variables based on current estimates, ensuring that the solution converges towards the optimal point.

The iterative scheme involves solving the following system of equations at each step:

$$\begin{bmatrix} w \\ b \end{bmatrix} = -\left[c_1 M^T M + c_1 N^T N + c_2 P^T P + c_3 P^T P + c D_1\right]^{-1} (-c_1 M^T e_1 + c_1 N^T e_2 + c_2 \delta P^T e_3 - c_3 \delta P^T e_3), \tag{5.2}$$

where $M = [k(A, D^T)\ e_1] \in R^{m_1 \times (m+1)}$, $N = [k(B, D^T)\ e_2] \in R^{m_2 \times (m+1)}$, $P = [k(C, D^T)\ e_3] \in R^{m_3 \times (m+1)}$, $D_1 = \operatorname{diag}\left((\beta_1^{(k+1)})^2, \ldots, (\beta_m^{(k+1)})^2, 1\right)$, $D = [A; B; C]$ and $m = m_1 + m_2 + m_3$.

A detailed description of our nonlinear PLSK-SVCR method with integrated feature selection is provided in Algorithm 3.

6. Decision Rule

Multi-class classification techniques assess all training points using the "1-versus-1-versus-rest" framework with ternary outputs $\{-1,0,+1\}$. When assessing a new testing point x_i, its class label is determined using the following decision functions:

For linear K-SVCR, LSK-SVCR, and PLSK-SVCR:

$$f(x_i) = \begin{cases} +1, & x_i^T w + b \geq \delta, \\ -1, & x_i^T w + b \leq -\delta, \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

For nonlinear K-SVCR, LSK-SVCR, and PLSK-SVCR:

$$f(x_i) = \begin{cases} +1, & k(x_i^T, D^T)w + b \geq \delta, \\ -1, & k(x_i^T, D^T)w + b \leq -\delta, \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

In a classification problem with K classes, the ""1-versus-1-versus-rest" strategy involves creating $\frac{K(K-1)}{2}$ classifiers.To determine the class label for a test sample x_i, each classifier contributes a vote to its predicted class. The final classification is assigned to the class that receives the most votes, following the commonly used max-voting rule.

7. Numerical Experiments

To assess the effectiveness of our proposed method, we conducted experiments using PLSK-SVCR on various benchmark datasets. We then compared these results with those obtained using LSK-SVCR. All experiments were conducted using Matlab 2019b on a PC equipped with an Intel(R) CORE(TM) i7-7700HQ CPU@2.80GHz and 16 GB of RAM.

Additionally, we applied a 5-fold cross-validation technique to comprehensively evaluate the performance of the algorithms in terms of accuracy and number of features.

In the 5-fold cross-validation procedure, the dataset is randomly divided into five approximately equal-sized subsets, with one serving as the test set and the remaining four as the training set. This process is repeated five times, and the average accuracy from these testing runs is used as the classification performance measure. It is important to note that accuracy is defined as the ratio of correct predictions to the total number of predictions and is typically presented as a percentage by multiplying the calculated ratio by 100.

7.1. **Optimal parameter selection.**

The accuracy of classification hinges on the precise selection of parameters. The impact of parameters on the classification accuracy of both K-SVCR and LSK-SVCR has been discussed in the previous work [27]. In the context of the wine dataset, the influence of the parameter p on accuracy is depicted in Fig.2. Similarly, Fig.3 and Fig.4 illustrate the effects of p on accuracy and the count of features for the DNA dataset, respectively. As illustrated by Fig.3, the accuracy remains unaffected by variations in p, while Fig. 4 highlights its impact on the number of features. This signifies that lower feature counts can yield the same level of accuracy. These graphical representations underscore the profound reliance on accuracy and feature count on parameter choices. Therefore, the meticulous selection of parameters emerges as a pivotal determinant in the classifiers' performance and feature selection outcomes. In other words, the classification performance of these algorithms is closely linked

to how we choose parameters. To simplify this, we utilized the grid search method to find the best parameter values [16, 19]. In our experiments, for finding the best values, we chose c_1, c_2, c_3, and c from the range $\{2^i | i = -8, -7, \ldots, 7, 8\}$. Additionally, parameter δ came from the set $\{0.1, 0.3, \ldots, 0.9\}$, and p for PLSK-SVCR was selected from $\{0.1, 0.3, \ldots, 0.9\}$.

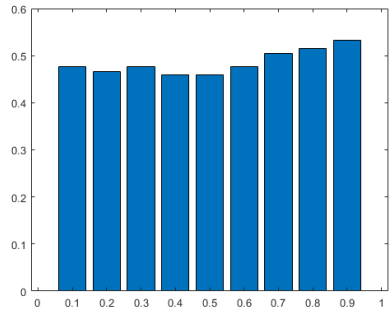

FIGURE 2. Effect of parameter p on accuracy for Wine data ($c = c_1 = c_2 = c_3 = 1000, \delta = 0.1$).

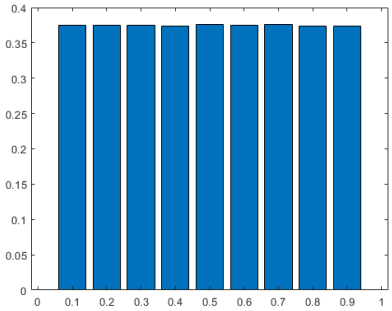

FIGURE 3. Effect of parameter p on accuracy for Dna data ($c = c_1 = c_2 = c_3 = 1000, \delta = 0.1$).

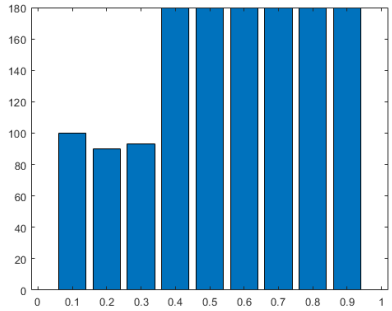

FIGURE 4. Impact of parameter p on the number of features for DNA data ($c = c_1 = c_2 = c_3 = 1000, \delta = 0.1$).

7.2. **Results comparisons and discussion for UCI datases.** In this section, we assess the performance of PLSK-SVCR in comparison to LSK-SVCR. We evaluated our methodology using six datasets. The datasets SRBCT, genRNASeq, MLL, CLL-SUB-111, TOX-171, and Lung are gene microarray data. A brief summary of the datasets is provided below.

The SRBCT and MLL can be downloaded from https://csse.szu.edu.cn/staff/zhuzx/datasets.html and genRNASeq from the UCI Machine Learning Repository [23].

The CLL-SUB-111, TOX-171, and Lung datasets can be downloaded from https://jundongl.github.io/scikit-feature/datasets.html.

The Small Round Blue Cell Tumors (SRBCT) microarray dataset comprises a total of 83 samples distributed across four distinct classes: the Ewing family of tumors (EWS), Burkitt lymphoma (BL), neuroblastoma (NB), and rhabdomyosarcoma (RMS) [20]. Each sample within this dataset contains precisely 2,308 gene expression values. Specifically, among the 83 samples, there are 29 samples classified as EWS, 11 samples as BL, 18 samples as NB, and 25 samples as RMS.

The genRNASeq dataset, which investigates gene expression in cancer using RNA-Seq technology, encompasses a diverse range of gene expressions sourced from patients diagnosed with various tumor types, including BRCA, KIRC, COAD, LUAD, and PRAD. With a comprehensive total of 801 samples, this dataset provides an extensive overview of gene expression patterns across different cancer types. It includes a total of 20,531 genes, making it a crucial resource for delving into the molecular intricacies of these cancers and offering promise for the identification of potential biomarkers or therapeutic targets.

The MLL dataset comprises three classes of leukemia: lymphoblastic leukemia (ALL), myeloid lymphoid leukemia (MLL), and acute myeloid leukemia (AML). It consists of a total of 72 samples, distributed as follows: 24 samples for ALL, 20 samples for MLL, and 28 samples for AML. The dataset contains 12,582 features [18].

The CLL-SUB-111 dataset comprises gene expression data obtained from high-density oligonucleotide arrays. It includes genetically and clinically distinct subgroups of B cell chronic lymphocytic leukemia (B-CLL), making it an invaluable resource for research. With 11,340 features and 111 samples, this dataset enables the exploration of molecular profiles and clinical variations within B-CLL.

The TOX-171 is a toxicology dataset that integrates clinical chemistry and expression data from the livers of rats, sampled 48 hours after exposure to three types of toxicants: alpha-naphthyl-isothiocyanate, dimethylnitrosamine, and N-methylformamide. Each compound corresponds to a specific toxicity mechanism, while the fourth class comprises samples from untreated controls. The dataset consists of continuous data, encompassing 171 samples and 5,748 features. It serves as a comprehensive resource for toxicology research, offering insights into toxicological phenomena and patterns through its structured representation into four distinct classes.

The dataset referred to as LUNG, cited in [7], is composed of a total of 203 samples categorized across four lung cancer types: adenocarcinomas, squamous cell lung carcinomas, pulmonary carcinoids, small-cell lung carcinomas, along with normal lung tissues. These classes comprise 139, 21, 20, 6, and 17 samples, respectively. Within each sample, data pertaining to 12,600 genes is recorded. In order to refine the dataset, genes exhibiting standard deviations

https://jundongl.github.io/scikit-feature/datasets.html

TABLE 1. Dataset description.

Data set	# Samples	# Features	# Classes
SRBCT	83	2308	4
genRNASeq	801	20531	5
MLL	72	12584	3
CLL-SUB-111	111	11340	3
TOX-171	171	5748	4
Lung	203	3312	4

TABLE 2. Classification accuracy of LSK-SVCR, and PLSK-SVCR.

Dataset	LSK-SVCR	PLSK-SVCR
Size	Acc (%)±Std Feature	Acc (%) ±Std Feature
SRBCT	92.65 ± 6.8 2308	**100.00** ± 00.00 **460**±2.65
genRNASeq	98.21 ± 0.09 20531	**99.5** ± 0.08 **53**±2
MLL	**97.54** ± **1.20** 12584	93.28 ± 6.19 **53.2**±2.49
CLL-SUB-111	51.9 ± 8.3 11340	**55.85** ± 14.15 **92**±4.06
TOX-171	90.11 ± 7.17 5748	**96.47** ± 3.01 **137.2**±5.36
Lung	93.91 ± 2.91 3312	**95.55** ± 3.29 **294.8**±2.77

lower than 50 expression units were excluded. Consequently, the refined dataset contains 203 samples and 3,312 genes, as documented in [7]. Table 1 presents a concise summary of the datasets.

To analyze the results presented in Table 2, we assessed the classification accuracy of two methods, LSK-SVCR and PLSK-SVCR, across various datasets. Our examination reveals insights into the performance of these methods under different dataset conditions.

In the SRBCT dataset, LSK-SVCR achieved an accuracy of 92.65% with a standard deviation of 6.8%. Surprisingly, PLSK-SVCR outperformed LSK-SVCR with a perfect accuracy of 100.00% and an impressively low standard deviation of 0.00%. This significant improvement suggests the potential of PLSK-SVCR for future selection, especially considering its lower feature count compared to LSK-SVCR.

Moving to the genRNASeq dataset, both methods demonstrated strong performance. LSK-SVCR achieved a high accuracy of 98.21% with a minimal standard deviation of 0.09%, while PLSK-SVCR also performed well, achieving an accuracy of 99.5% with a similarly low standard deviation of 0.08%. These results indicate the effectiveness of both methods in handling

datasets with high dimensionality and complex patterns, but we note that our proposed PLSK-SVCR method achieved a slightly better accuracy while utilizing lower features.

However, the MLL dataset showcased a slightly different trend. Here, LSK-SVCR marginally outperformed PLSK-SVCR, achieving an accuracy of 97.54% with a standard deviation of 1.20%. In contrast, PLSK-SVCR achieved an accuracy of 93.28% with a higher standard deviation of 6.19%. This suggests that while PLSK-SVCR may not consistently outperform LSK-SVCR across all datasets, its performance remains competitive and merits consideration for feature selection.

In the remaining datasets, namely CLL-SUB-111 and TOX-171, PLSK-SVCR showed significant improvements compared to LSK-SVCR. For example, in the CLL-SUB-111 dataset, LSK-SVCR achieved an accuracy of 51.9%, whereas PLSK-SVCR improved upon this accuracy significantly, achieving 55.85%. Notably, PLSK-SVCR achieved this higher accuracy with lower features due to its feature selection mechanism. Similarly, in the TOX-171 dataset, PLSK-SVCR achieved a higher accuracy of 96.47% compared to LSK-SVCR's 90.11%, with lower standard deviation (3.01% versus 7.17%).

Lastly, in the Lung dataset, both methods performed well, with LSK-SVCR achieving an accuracy of 93.91%, while PLSK-SVCR slightly improved upon this with an accuracy of 95.55%. Notably, PLSK-SVCR achieved this using significantly fewer features.

In summary, while PLSK-SVCR shows promising potential for enhancing the accuracy and performance of LSK-SVCR through feature selection, further research is required to fully explore its effectiveness and applicability across diverse datasets.

8. Conclusion and Feature Works

In summary, this paper introduces PLSTKSVC, a novel model designed to address challenges in multi-class classification. By dynamically selecting the parameter p based on available data and seamlessly integrating feature selection and classification through cardinality-constrained optimization, PLSTKSVC offers an adaptive learning approach. Despite the non-convex nature of the ℓ_p-norm, PLSTKSVC efficiently solves optimization problems using a linear system of equations.

PLSTKSVC's key advantages include simultaneous feature selection and classification, established theoretical foundations, algorithmic efficiency, and empirical validation. Theorems regarding lower bounds of non-zero entries in solutions and upper bounds for the norm of the optimal solution enhance its theoretical understanding. Experimental results on multi-class classification datasets demonstrate PLSTKSVC's superior performance, making it a valuable contribution to machine learning.

Looking forward, there are promising avenues for future research, particularly concerning features:

- Exploring methods to improve the interpretability of feature selections made by PLSK-SVCR, providing deeper insights into the relevance of chosen features in various contexts.
- Evaluating the performance of PLSK-SVCR on imbalanced datasets and exploring potential adjustments or enhancements to ensure robust performance in such scenarios.

In conclusion, PLSK-SVCR has demonstrated significant strengths in feature-efficient multi-class classification. Future endeavors can build upon these findings to refine its capabilities and

explore innovative directions to improve feature selection and adaptability in diverse machine learning applications.

REFERENCES

[1] A. Ahmad, M. Hassan, M. Abdullah, H. Rahman, F. Hussin, H. Abdullah, R. Saidur, A review on applications of ann and svm for building electrical energy consumption forecasting, Renew. Sustain. Energy Rev. 33 (2014), 102-109.

[2] C. Angulo, X. Parra, A. Catala, K-SVCR. A support vector machine for multi-class classification, Neurocomputing 55 (2003), pp. 57-77.

[3] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A. Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm, Comput. Methods Programs Biomed. 141 (2017), 19-26.

[4] F. Bazikar, S. Ketabchi, H. Moosaei, Dc programming and dca for parametric-margin ν-support vector machine, Appl. Intell. 50 (2020), 1763-1774.

[5] A. Beck, Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB, SIAM, 2014.

[6] D. Bertsimas, R. Shioda, Algorithm for cardinality-constrained quadratic optimization, Comput. Optim. Appl. 43 (2009), 1-22.

[7] A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. 98 (2001), 13790-13795.

[8] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, In: Haussler, D. (ed.) Proceedings of the fifth annual workshop on Computational learning theory, pp. 144–152, ACM, New York, 1992.

[9] A. Bruckstein, D. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (2009), 34-81.

[10] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014), 16-28.

[11] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995), 273-297.

[12] S. Ding, S. Shi, W. Jia, Research on fingerprint classification based on twin support vector machine, IET Image Process. 14 (2019), 231-235.

[13] S. Ding, N. Zhang, X. Zhang, F. Wu, Twin support vector machine: theory, algorithm and applications, Neural Comput. Appl. 28 (2017), 3119-3130.

[14] O. Déniz, M. Castrillon, M. Hernández, Face recognition using independent component analysis and support vector machines, Pattern Recognit. Lett. 24 (2003), 2153-2157.

[15] M.B. Fenn, P. Xanthopoulos, G. Pyrgiotakis, S.R. Grobmyer, P.M. Pardalos, L.L. Hench, Raman spectroscopy for clinical oncology, Adv. Opt. Technol. 2011 (2011), 213783.

[16] C.W. Hsu, C.C. Chang, C.J. Lin, A practical guide to support vector classification, https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (2003).

[17] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2002), 415-425.

[18] S. Kar, K. Sharma, M. Maitra, Gene selection from microarray gene expression data for classification of cancer subgroups employing pso and adaptive k-nearest neighborhood technique, Expert Syst. Appl. 42 (2015), 612-627.

[19] S. Ketabchi, H. Moosaei, M. Razzaghi, P. Pardalos, An improvement on parametric ν-support vector algorithm for classification, Ann. Oper. Res. 276 (2019), 155-168.

[20] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (2001), 673-679.

[21] U. Kressel, Pairwise classification and support vector machines. In: Scholkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 255-268. MIT Press (1998)

[22] M.J. Lai, J. Wang, An unconstrained l_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems, SIAM J. Optim. 21 (2011), 82-101.
[23] M. Lichman, Uci machine learning repository (2013), http://archive.ics.uci.edu/ml
[24] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Inf. Sci. 181 (2011), 115-128.
[25] A. Miller, Subset Selection in Regression. CRC Press, Boca Raton, FL, 2nd edn., 2002.
[26] H. Moosaei, M. Hladík, Bounds for sparse solutions of k-svcr multi-class classification model, In: International Conference on Learning and Intelligent Optimization. pp. 136-144, Springer, 2022.
[27] H. Moosaei, M. Hladík, Least squares approach to k-svcr multi-class classification with its applications, Ann. Math. Artif. Intell. 90 (2022), 873-892.
[28] H. Moosaei, M. Hladík, Sparse solution of least-squares twin multi-class support vector machine using l_0 and l_p-norm for classification and feature selection." Neural Networks 166 (2023), 471-486.
[29] H. Moosaei, M. Hladík, Least squares k-svcr multi-class classification. In: Kotsireas, I., Pardalos, P. (eds.) Learning and Intelligent Optimization, 14th International Conference, LION 14, Athens, Greece, May 24-28, 2020, Revised Selected Papers, pp. 117–127. Springer, Cham (2020)
[30] H. Moosaei, S. Ketabchi, M. Razzaghi, M. Tanveer, Generalized twin support vector machines, Neural Process. Lett. 53 (2021), 1545–1564.
[31] H. Moosaei, F. Bazikar, M. Hladík, Multi-task twin support vector machine with universum data, Engineering Applications of Artificial Intelligence 132 (2024), 107951.
[32] H. Moosaei, F. Bazikar, S. Ketabchi, M. Hladík, Universum parametric-margin ν-support vector machine for classification using the difference of convex functions algorithm, Appl. Intelligence 52 (2022), 2634-2654.
[33] H. Moosaei, A. Mousavi, M. Hladík, Z. Gao, Sparse l1-norm quadratic surface support vector machine with universum data, Soft Computing 27 (2023), 5567-5586.
[34] B. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 24 (1995), 227-234.
[35] P. Pardalos, V. Boginski, A. Vazacopoulos, Data Mining in Biomedicine, Springer Optimization and Its Applications, vol. 7 (2007)
[36] M. Shao, X. Wang, Z. Bu, X. Chen, Y. Wang, Prediction of energy consumption in hotel buildings via support vector machines, Sustain. Cities Soc. 57 (2020), 102128.
[37] L. Tang, Y. Tian, P. Pardalos, A novel perspective on multiclass classification: regular simplex support vector machine, Info. Sci. 480 (2019), 324-338.
[38] T. Trafalis, H. Ince, Support vector machine for regression and applications to financial forecasting. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000, Neural Computing: New Challenges and Perspectives for the New Millennium. vol. 6, pp. 348–353 (2000)
[39] H. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (2012), 3586-3592.