论文

0. Abstract

```
本篇论文实现的模型是：CPTWSVM
目的是：estimate the conditional probability function
方法（策略）是：implement the empirical risk minimization on training data
借助的算法是：block decomposition algorithm
用途是：二分类 / 多分类
```

In this paper, we *estimate the conditional probability function* by presenting a new twin SVM model (CPTWSVM) in *binary and multiclass classification problems*. The motivation of CPTWSVM is to *implement the empirical risk minimization on training data*, which is hard to realize in traditional twin SVMs.

> *本文通过在二分类和多分类问题中提出一种新的双支持向量机模型（CPTWSVM）来估计条件概率函数。CPTWSVM 的动机是在训练数据上实现经验风险最小化，这在传统的双支持向量机中难以实现。*

In each subproblem of CPTWSVM, it measures the empirical risk and outputs the corresponding probability estimate of each class, which eliminates the problems of inconsistent measurement in **twin SVMs**.

> *在 CPTWSVM 的每个子问题中，它度量经验风险并输出每个类别的相应概率估计，这消除了双支持向量机中度量不一致的问题。*

Though an additional **discriminant objective function** is introduced, the optimization problem size of each subproblem is smaller than conditional probability SVM, and is solved by **block decomposition algorithm** efficiently. In addition, we extend CPTWSVM to multiclass classification by *estimating the conditional probability of each class*, and maintaining the above properties.

> *虽然引入了一个额外的判别目标函数，但每个子问题的优化问题规模小于条件概率支持向量机，并且可以通过块分解算法高效求解。此外，我们通过估计每个类别的条件概率将CPTWSVM扩展到多分类，并保持上述特性。*

Numerical experiments on benchmark and real application datasets demonstrate that CPTWSVM outputs the estimate of *probability and the data projection* well, resulting in better generalization ability than some leading TWSVMs communities, in terms of binary and multiclass classification.

> *在基准和实际应用数据集上的数值实验表明，CPTWSVM 能很好地输出概率估计和数据投影，在二分类和多分类方面，比一些领先的双支持向量机社区具有更好的泛化能力。*

1. Introduce

```
TWSVM的问题：
1. 训练过程中不考虑样本是否被正确分类
2. 无法度量经验风险，需要考虑3种一致性问题，并且这种一致性问题本身的意义尚不明确
3. 本身不能给出概率，可解释性差

本文的贡献：
1. 提出了CPTWSVM，为每个类别估计条件概率函数，从而可以轻松度量经验风险。
2. CPTWSVM可以计算训练样本的条件概率，并且不用考虑一致性问题。
3. CPTWSVM的优化问题具有比PSVM更小的规模，且可以通过块算法很好地解决。
4. CPTWSVM可以在保留上述性质的基础上，（轻松地）扩展到多分类问题中。
```

paragraphs1

Twin support vector machine (TWSVM) [1,2] and their variants, are well known as the extension of support vector machines (SVM) [3,4] and **proximal support vector machine via generalized eigenvalues (GEPSVM)** [5]. Compared to SVM, the main characteristic of TWSVM is that it divides a whole model of SVM into several sub-models. It has the idea of breaking up the whole into parts and turning a complex problem into several simple sub-problems.

> *双支持向量机（TWSVM）[1,2] 及其变体，众所周知是支持向量机（SVM）[3,4] 和基于广义特征值的近端支持向量机（GEPSVM）[5] 的扩展。与支持向量机相比，TWSVM 的主要特点是将支持向量机的整个模型划分为几个子模型。它具有化整为零、将复杂问题转化为若干简单子问题的思想。*

From this point of view, TWSVM is more suitable for solving complex and large-scale problems. Therefore, TWSVM has been extended to multiclass [7,8], regression [9–11], clustering [12,13], semi-supervised learning [14,15], and many other machine learning problems [16,17], et al.

> *从这个角度来看，TWSVM 更适合解决复杂和大规模问题。因此，TWSVM 已被扩展到多分类 [7,8]、回归 [9,10,11]、聚类 [12,13]、半监督学习 [14,15] 以及许多其他机器学习问题 [16,17] 等。*

paragraphs2

In TWSVM, the sub-problems are solved paralleled. There are many problems that are worth studying on *how to measure the different results of sub-problems and optimally combine them*. Specifically, how to measure the empirical risk or accurately measure the loss of misclassification of training samples is a critical problem.

> *在 TWSVM 中，子问题是并行求解的（指的是彼此互不影响）。关于如何度量子问题的不同结果并最优地组合它们，存在许多值得研究的问题。具体来说，如何度量经验风险或准确度量训练样本的误分类损失是一个关键问题。*

In training procedure, TWSVM [1,18] constructs two hyperplanes separately, one of them is close to one of classes and far away from the other class of the training samples. In testing procedure, a new sample is classified as which hyperplane is closer to it. *This shows that in training procedure, we do not measure whether the samples are correctly classified*.

> *在训练过程中，TWSVM [1,18] 分别构造两个超平面，其中一个超平面靠近一个类别的训练样本，而远离另一个类别的训练样本。在测试过程中，一个新的样本被分类到距离它更近的超平面所属的类别。这表明在训练过程中，我们并不度量样本是否被正确分类。*

Therefore, *we cannot measure empirical risk in TWSVM accurately*, but it is easy to measure in SVM. In contrast, there are many studies on the loss of TWSVM, Mehrkanoon et al. [19] defined the *within-class and between-classes loss for each sub-model, and studied the least square within-class loss compared to least square, hinge and pinball between-classes losses for TWSVM*.

> *因此，我们无法在 TWSVM 中准确度量经验风险，但在 SVM 中却很容易度量。相比之下，关于 TWSVM 损失的研究有很多，Mehrkanoon 等人 [19] 为每个子模型定义了类内损失和类间损失，并研究了 TWSVM 中类内最小二乘损失与类间最小二乘损失、铰链损失和 pinball 损失的比较。*

Shao et al. [20] further proposed linear loss for each sub-model in TWSVM, which made the sub-problems of TWSVM easier to optimize. And many convex or non-convex losses TWSVMs have been presented, such as [21– 25].

> *邵等人 [20] 进一步在 TWSVM 中为每个子模型提出了线性损失，这使得 TWSVM 的子问题更容易优化。并且提出了许多凸或非凸损失的 TWSVM，例如 [21,22,23,24,25]。然而，它们都不能度量 TWSVM 中的误分类损失。*

However, *all of them can not measure the loss of misclassification in TWSVM. Thus, how to measure the loss of misclassification in TWSVM properly is still an open question*.

> *因此，如何恰当地度量 TWSVM 中的误分类损失仍然是一个悬而未决的问题。*

paragraphs3

The consistency of measurement in TWSVM is another problem that is worth studying. We need to *measure the distance(similarity) or loss* in each sub-model and in prediction, and there are *three types of consistency of measurement that need to be considered in TWSVM*.

> *TWSVM 中度量的一致性是需要研究的另一个问题。我们需要在每个子模型和预测中度量距离（相似性）或损失，在 TWSVM 中需要考虑三种类型的一致性度量。*

The first one is the *consistency of the distances between within-class and between-classes for each sub-model*. In TWSVM [16,30], it measures the within-class loss by least squares loss and the between-classes loss by hinge loss, and the above different losses studies are concerned with the consistency of the distances between within-class and between-classes for each sub-model.

> *第一种是每个子模型中类内和类间距离的一致性。在 TWSVM [16,30] 中，它通过最小二乘损失度量类内损失，通过铰链损失度量类间损失，而上述不同损失的研究关注的是每个子模型中类内和类间距离的一致性。*

The second one is *the consistency of the distances during the training and prediction procedure*. For this type of consistency, Shao et al. [26,27] pointed out that GEPSVM is consistent while TWSVMs are not consistent, and some training and predicting consistencies TWSVMs had been presented [28–30], and experiments show that this type of consistency will improve the generalization ability of TWSVM.

> *第二种是在训练和预测过程中距离的一致性。对于这种类型的一致性，邵等人 [26, 27] 指出 GEFSVM 是一致的，而 TWSVM 是不一致的，并且已经提出了一些具有训练和预测一致性的 TWSVM [28, 29, 30]，实验表明这种一致性将提高 TWSVM 的泛化能力。*

The third one is the *consistency of the distances between sub-models*. While for the third type of consistency, there is little study.

> *第三种是子模型之间距离的一致性。而对于第三种一致性，研究很少。*

In fact, up to now, *it is not clear whether these three kinds of consistency/inconsistency are beneficial and how these measurements are beneficial to TWSVM*. These inconsistencies bring some *inconvenience to the downstream tasks of classification*. For example, TWSVM can not give the probability output [31–33], in order to give the probability of TWSVM, [31] designed a two-stage algorithm based on its output. While for knowledge transfer, TWSVM [34,35] needs to transfer every output of each subproblems.

> *事实上，到目前为止，尚不清楚这三种一致性/不一致性是否有益，以及这些度量如何有益于 TWSVM。这些不一致性给分类的下游任务带来了一些不便。例如，TWSVM 不能给出概率输出 [31, 32, 33]，为了给出 TWSVM 的概率，[31] 基于其输出设计了一个两阶段算法。而对于知识迁移，TWSVM [34, 35] 需要迁移每个子问题的每一个输出。*

paragraphs4

Recently, Vapnik and Izmailov [36] had *reinforced SVM by $p(y|x) = \langle w, x \rangle + b$ for binary classification and obtained a more reasonable SVM model (called PSVM for short) from a statistical point of view*. And this conditional probability estimation is more convenient for statistical calculation and knowledge transfer [37,38]. This inspired us to study TWSVM from the perspective of probability estimation.

> *最近，Vapnik 和 Izmailov [36] 通过 $p(y \mid x) = \langle w, x \rangle + b$ 对二分类问题强化了支持向量机，并从统计角度获得了一个更合理的支持向量机模型（简称 PSVM）。这种条件概率估计更方便进行统计计算和知识迁移 [37, 38]。这启发了我们从概率估计的角度研究 TWSVM。*

In this paper, we focus on *minimizing empirical risk and output the probability estimation in TWSVM*. To realize the empirical risk minimization, we set the $p(y_k|x) = \langle w_k, x \rangle + b_k$ and estimate it from the training data for each class, where $k = 1, 2, \ldots, K$ means the category. Similar to TWSVM, we estimate $p(y_k|x)$ separately.

> *在本文中，我们专注于在 TWSVM 中最小化经验风险并输出概率估计。为了实现经验风险最小化，我们设定 $p(y_k|x) = \langle w_k, x \rangle + b_k$ 并从训练数据中为每个类别估计它，其中 $k = 1, 2, \ldots, K$ 表示类别。类似于 TWSVM，我们分别估计 $p(y_k \mid x)$。*

However, *by using the probability estimation, the empirical risk is easy to calculate and we don't need to consider the problem of measurement consistency*. In addition, we add a **discriminant objective function** in each sub-model, and the optimization problem size of each sub-problem is smaller than PSVM. By introducing the block decomposition algorithm, the subproblems are solved iteratively by solving a series of small problems.

> *然而，通过使用概率估计，经验风险很容易计算，我们不需要考虑度量一致性的问题。此外，我们在每个子模型中添加了一个判别目标函数，并且每个子问题的优化问题规模小于 PSVM。通过引入块分解算法，子问题通过求解一系列小问题迭代求解。*

Experimental results show that the proposed model outputs the probability estimations and data projections well, and it has better generalization ability than traditional TWSVM on most datasets. To sum up, the *main contributions of the paper are*:

> *实验结果表明，所提出的模型能很好地输出概率估计和数据投影，并且在大多数数据集上比传统的 TWSVM 具有更好的泛化能力。总结来说，本文的主要贡献是：*

i) Conditional probability twin SVM(CPTWSVM) is proposed, *the conditional probability function is estimated for each class and the empirical risk can measure by CPTWSVM easily.*

> *i) 提出了条件概率双支持向量机（CPTWSVM），为每个类别估计条件概率函数，并且 CPTWSVM 可以轻松度量经验风险。*

ii) CPTWSVM not only *can measure the misclassification of the training samples for each class, but also returns the discriminant projections and conditional probability estimations of each class*. In addition, there is no need to consider the problem of measurement consistency in CPTWSVM.

> *ii) CPTWSVM 不仅可以度量每个类别的训练样本的误分类情况，而且返回每个类别的判别投影和条件概率估计。此外，在 CPTWSVM 中不需要考虑度量一致性的问题。*

iii) The sub-problem of CPTWSVM is a *PSVMtype quadratic programming problem*. It has a *small optimization problem size than PSVM, and could be solved by block decomposition algorithm*.

> *iii) CPTWSVM 的子问题是一个 PSVM 型的二次规划问题。它的优化问题规模比 PSVM 小，并且可以通过块分解算法求解。*

iv) CPTWSVM *can be easily extended to multiclass classification, and maintain the above properties*.

> *iv) CPTWSVM 可以轻松扩展到多分类，并保持上述特性。*

Comparing with some leading SVMs and TWSVMs on benchmark datasets, extensive numerical experiments demonstrate that our proposed method achieves better performance including higher prediction accuracy and better interpretative output.

> *在基准数据集上与一些领先的支持向量机和双支持向量机进行比较，大量的数值实验表明，我们提出的方法取得了更好的性能，包括更高的预测准确率和更好的可解释输出。*

paragraphs5

The remainder of this paper is organized as follows: Section 2 gives a brief overview of relative SVM and TWSVM. Section 3 presents our proposed CPTWSVM and related theoretical analysis. Section 4 discusses the results of experiments conducted. Finally, concluding remarks are given in Section 5.

> *本文的其余部分组织如下：第 2 节简要概述了相关的支持向量机和双支持向量机。第 3 节提出了我们建议的 CPTWSVM 及相关的理论分析。第 4 节讨论了进行的实验结果。最后，第 5 节给出了结论性评论。*

2. Relative Work

Consider a binary classification problem in the n-dimensional real space R^n. The set of training samples is represented by $T=\{(x_i,y_i)|i=1,\dots m\}$, where $x_i\in R_n$ is the input and $y_i\in\{+1,-1\}$ is the corresponding output. Denote I is the set of indices i such that $I=\{1,2,\dots,m\}$, I^1 and I^2 represent the sets of indices of positive and negative training samples, and $I=I^1\cup I^2$. Besides, for multiclass classification problem, $y_i\in\{1,2,\dots,K\}$. Denote I^k represents the indices set of the class k for $k=1,\dots,K$, and $I=I^1\cup\dots\cup I^k$. We also let $z=\varphi(x)$ and $\varphi(x)$ is a mapping function.

> *考虑 n 维实空间 R^n 中的二分类问题。训练样本集表示为 $T=\{(x_i,y_i)|i=1,\dots m\}$ ，其中 $x_i\in R_n$ 是输入，$y_i\in\{+1,-1\}$ 是对应的输出。记 I 是索引i的集合，使得$I=\{1,2,\dots,m\}$， I^1 和 I^2 分别表示正类和负类训练样本的索引集合，且$I=I^1\cup I^2$ 。此外，对于多分类问题，$y_i\in\{1,2,\dots,K\}$。记 I^k 表示类别 k 的索引集合，其中$k=1,\dots,K$，且$I=I_1\cup\dots\cup I_k$。我们还令$z=\varphi(x)$，其中 $\varphi(x)$ 是一个映射函数。*

2.1. Review of SVMs

```
介绍了SVM的原问题和PSVM问题的优化形式及其对偶的优化形式
```

Paragraphs1

Traditional SVM constructs a separating hyperplane $\langle w,z\rangle+b=0$ by solving

$$\begin{gathered} min_{w,b,\xi_i}\frac{1}{2}||w||^2+C\sum_{i\in I}\xi_i \\ s.t.\ y_i(\langle w,z_i\rangle+b)\geq 1-\xi_i,\xi_i\geq 0,i\in I \end{gathered} \qquad (1)$$

where C is a positive penalty parameter. Minimize the regularization term $\frac{1}{2}||w||^2$ is equivalent to maximize of the margin between two parallel supporting hyperplanes $\langle w,z\rangle+b=1$ and $\langle w,z\rangle+b=-1$, and minimize the slack variable ξ is equivalent to minimizing the margin error, which is also called *the loss or the empirical risk of SVM*.

> *传统的支持向量机通过求解(1)来构造一个分离超平面 $\langle w,z\rangle+b=0$, 其中 C 是一个正惩罚参数。最小化正则化项 $\frac{1}{2}||w||^2$, 等价于最大化两个平行支撑超平面 $\langle w,z\rangle+b=1$ 和 $\langle w,z\rangle+b=-1$之间的间隔，最小化松弛变量 ξ 等价于最小化间隔误差，这也称为支持向量机的损失或经验风险。*

Traditional [47], we solve its dual problem

$$\begin{gathered} min_\alpha\frac{1}{2}\sum_{i,j\in I}\langle z_i,z_j\rangle\alpha_i\alpha_j-\sum_j\alpha_j \\ s.t.\sum_{i\in I}y_i\alpha_i=0,0\leq\alpha_i\leq C,i\in I \end{gathered} \qquad (2)$$

to obtain the optimal solution w and b of the primal problem (1). And the support vectors are the training samples $\{(x_i, y_i)|i \in I\}$ corresponding to the non-zero $\alpha_i, i \in I$ in duality problem (2), which determines hyperplane. Besides, *for a new sample x^*, when $\langle w, \varphi(x^*)\rangle + b > 0$, it has been classified into positive class $y = +1$, otherwise, it is classified into negative class*.

> *传统上 [47]，我们求解其对偶问题(2), 以获得原始问题(1)的最优解 w 和 b。支持向量是对应对偶问题 (2) 中非零 $\alpha_i, i \in I$ 的训练样本 $\{\langle x_i, y_i\rangle \mid i \in I\}$，它们决定了超平面。此外，对于一个新样本 x^*，当 $\langle w, \varphi(x^*)\rangle + b > 0$ 时，它被分类为正类 $y = +1$，否则，它被分类为负类。*

To obtain a probability of belonging to the classes, the standard implementations of SVM typically use Platt's model-driven approximation [39] for this probability, or some accurate data-driven estimates [40].

> *为了获得属于各类别的概率，支持向量机的标准实现通常使用 Platt 的模型驱动近似 [39] 来计算该概率，或者使用一些精确的数据驱动估计 [40]。*

Paragraphs2

Recently, *to obtain the probability function directly*, Vapnik and Izmailov [36] *consider $p(y|x) = \langle w, z\rangle + b$ from a statistical point of view*, and take values of $\langle w, z\rangle + b$ between 0 and 1: $0 \leq \langle w, z_i\rangle + b \leq 1, i \in I$, for all training samples.

> *最近，为了直接获得概率函数，Vapnik 和 Izmailov [36] 从统计角度考虑 $p(y \mid x) = \langle w, z\rangle + b$，并令" $\langle w, z\rangle + b$ 的值在 0 和 1 之间"对所有训练样本成立。*

Then, a probability estimation SVM (PSVM) is constructed by solving

$$
\begin{array}{c}
min_{w,b,\xi} \frac{1}{2}||w||^2 + \frac{C}{\varepsilon}\sum_{i \in I} \xi_i, \\
s.t. \;\; y_i(\langle w, z_i\rangle + b - 0.5) \geq 0.5 \times \varepsilon - \xi_i \quad \xi_i \geq 0, i \in I \\
0 \leq \langle w, z_i\rangle + b \leq 1, i \in I
\end{array}
\tag{3}
$$

where $\varepsilon > 0$ is a small value. The geometric explanation of this problem is clear, that is, the maximum margin classifier under the probability function estimation.

> *然后，通过求解以下问题构建一个概率估计 SVM（PSVM）。其中ε是一个小量，这个问题的几何解释是清晰的，即在概率函数估计下的最大间隔分类器。*

The primal problem (3) is also solved by its dual form as

$$
\begin{array}{c}
min_{\alpha,\beta,\gamma} \frac{1}{2} \sum_{i,j \in I} (y_i\alpha_i + \beta_i - \gamma_i)^\top (y_j\alpha_j + \beta_j - \gamma_j)\langle z_i, z_j\rangle - \sum_{i \in I}(0.5 \times \alpha_i(y_i + \varepsilon) - \gamma_i) \\
s.t. \sum_{i \in I}(y_i\alpha_i + \beta_i - \gamma_i) = 0, 0 \leq \alpha_i \leq \frac{C}{\varepsilon}, \beta_i \geq 0, \gamma_i \geq 0, i \in I
\end{array}
\tag{4}
$$

By solving (4), we obtain the optimal solution w and b. Similarly, the support vectors are the training samples $\{(x_i, y_i)|i \in I\}$ corresponding to the non-zero α_i, β_i and γ_i, $i \in I$ in duality problem (4). As for new samples x^*, when $\langle w, \varphi(x^*)\rangle + b > 0.5$, it has been classified into positive class $y = +1$, otherwise, it is classified into negative class.

> *通过求解 (4)，我们获得最优解 w 和 b。类似地，支持向量是对应对偶问题 (4) 中非零 α_i, β_i 和 $\gamma_i, i \in I$ 的训练样本 $\{\langle x_i, y_i\rangle \mid i \in I\}$。对于新样本 x^*，当 $\langle w, \varphi(x^*)\rangle + b > 0.5$ 时，它被分类为正类 $y = +1$，否则被分类为负类。*

Paragraphs3

Compared to SVM, PSVM has the following characteristics: i) For training sample $z_i, i \in I$, the value of $\langle w, z_i\rangle + b$ is in $[0, 1]$, and $\langle w, z_i\rangle + b$ is the conditional probability estimation of training samples. That is to say, PSVM could output the probability directly. ii) Due to $\langle w, z_i\rangle + b$ is bounded, its kernel matrix is also bounded, i.e., $||K||_F \leq \delta$, where $|| \cdot ||_F$ is Frobenius norm and δ is a positive bounded number. From the statistic learning theory [41,42], we can see that the conditional probability estimation SVM is more reasonable than traditional SVM.

> *与 SVM 相比，PSVM 具有以下特点：i) 对于训练样本 z_i，$i \in I$，$\langle w, z_i\rangle + b$ 的值在 $[0, 1]$ 内，且 $\langle w, z_i\rangle + b$ 是训练样本的条件概率估计。也就是说，PSVM 可以直接输出概率。ii) 由于 $\langle w, z_i\rangle + b$ 是有界的，其核矩阵也是有界的，即 $\| K \|_F \leq \delta$，其中 $\| \cdot \| \ F$ 是 Frobenius 范数，δ是一个正有界数。从统计学习理论 [41,42] 可以看出，条件概率估计 SVM 比传统 SVM 更合理。*

2.2 Review of TWSVMs

Paragraphs1

The twin support vector machine (TWSVM) [1,2] seeks a pair of nonparallel hyperplanes defined as $f_1(x) = \langle w_1, z\rangle + b_1 = 0$ and $f_2(x) = \langle w_2, z\rangle + b_2 = 0$ in ϕ space, such that each hyperplane is closer to the data samples of one class and far from the data samples of the other class to some extent. A new data sample is assigned to Class $+1$ or -1 depending upon its proximity to the two nonparallel hyperplanes.

> *双支持向量机（TWSVM）[1,2] 寻找一对在 ϕ 空间中定义为 $f_1(x) = \langle w_1, z\rangle + b_1 = 0$ 和 $f_2(x) = \langle w_2, z\rangle + b_2 = 0$ 的非平行超平面，使得每个超平面在某种程度上更接近一个类别的数据样本，而远离另一个类别的数据样本。一个新的数据样本根据其与两个非平行超平面的接近程度被分配到 +1 类或 −1 类。*

In order to find the hyperplanes, the solutions to the following primal problems are required

\begin{gather}min_{w_1,b_1,\xi} \frac{1}{2}(||w_1||^2 + b_1^2) + \frac{C_1}{2} \sum_{i∈I^1} (〈w_1, z_i〉 + b_1)^2 + C_2 \sum

> *为了找到这些超平面，需要求解以下原始问题的解*

where C_1, C_2, C_3 and C_4 are positive parameters, and $||\cdot||$ signifies the 2-norm. Their geometric meaning is clear: for example, the second objective function(目标函数) in (5) makes Class +1 proximate to(与…相邻近) the hyperplane $\langle w_1, z\rangle + b_1 = 0$, while the constraints make Class -1 bounded in the hyperplane $\langle w_1, z\rangle + b_1 = -1$, while the third objective function is to minimize the loss of the constraint. The first objective function is the regular term.

> *其中 C_1、C_2、C_3 和 C_4 是正参数，$||\cdot||$ 表示 2-范数。它们的几何意义是清晰的：例如，(5) 中的第二个目标函数使 $+1$ 类邻近于超平面 $\langle w_1, z\rangle + b_1 = 0$，而约束条件使 -1 类有界于超平面 $\langle w_1, z\rangle + b_1 = -1$，第三个目标函数则是最小化约束的损失。第一个目标函数是正则项。*

Paragraphs2

Let $\tilde{z} = [z; 1]$, $\mathbb{I}_{i,j} = 1$ if $i = j$ and $\mathbb{I}_{i,j} = 0$ otherwise. The dual form of the (5) and (6) are

$$min_\alpha \frac{1}{2} \sum_{i,j\in I^2} \tilde{z}_i(C_1 \sum_{i,j\in I^1} (\langle \tilde{z}_i, \tilde{z}_j\rangle + \mathbb{I}_{i,j}))^{-1} \tilde{z}_j \alpha_i \alpha_j - \sum_{i\in I^2} \alpha_i \quad (7)$$
$$s.t. 0 \le \alpha_i \le C_2, i \in I^2,$$

and

$$min_\beta \frac{1}{2} \sum_{i,j\in I^1} \tilde{z}_i(C_3 \sum_{i,j\in I^2} (\langle \tilde{z}_i, \tilde{z}_j\rangle + \mathbb{I}_{i,j}))^{-1} \tilde{z}_j \beta_i \beta_j - \sum_{i\in I^1} \beta_i \quad (8)$$
$$s.t. 0 \le \beta_i \le C_4, i \in I^1.$$

> *令 $\tilde{z} = [z; 1]$，$\mathbb{I}_{i,j} = 1$如果 $i = j$，否则 $\mathbb{I}_{i,j} = 0$。则，(5) 和 (6) 的对偶形式为：(7) 和 (8)。*

Once the solutions of problems (7) and (8) are obtained, (w_1, b_1) and (w_2, b_2) could be calculate by Karush-Kuhn-Tucker (KKT) conditions, a new sample $x^* \in R^n$ is assigned to class $y(y \in \{+1, -1\})$, depending on which of the two hyperplanes is closer, i.e.,

$$Class\ y = argmin_{k=1,2} \frac{|\langle w_k, \varphi(x^*)\rangle + b_k|}{||w_k||} \quad (9)$$

where $|\cdot|$ is the absolute value. In TWSVM [1], the support vectors of the positive hyperplane is defined as the negative samples corresponding to $0 < \alpha_i < C_2, (i \in I^2)$ and the support vectors of the negative hyperplane is defined as the positive samples corresponding to $0 < \beta_i < C_2, (i \in I^1)$.

> *一旦获得问题 (7) 和 (8) 的解，(w_1, b_1) 和 (w_2, b_2) 可以通过 Karush-Kuhn-Tucker (KKT) 条件计算得到。一个新样本 $x^* \in R^n$ 根据与哪个超平面更近被分配到类别* y&nbsp;(y \in \{+1,−1 \}*，即：(9)。其中 $|\cdot|$ 是绝对值。在 TWSVM [1] 中，正超平面的支持向量定义为对应于*0<α_i<C_2&nbsp;(i∈I^2)*的负类样本，而负超平面的支持向量定义为对应于* 0<β_i<C_4&nbsp;(i∈I^1) *的正类样本。*

Paragraphs3

Compared to SVM, TWSVM has the following characteristics: i) For each class, *it constructs a hyperplane $f_k = \langle w_k, z\rangle + b_k = 0$, and all hyperplanes are constructed separately*. Each size of *the optimization sub-problem in TWSVM is smaller than that of SVM*. And at the same time, we expect the sub-problem in TWSVM is simpler than the whole one in SVM. ii) The f_k is obtained by measuring the within-class similarity and between-class dissimilarity, and the new sample is assigned by the class to which fk is close. That is to say, *the measurement of similarity/dissimilarity in each model and decision may be different*, and *it is hard to measure the loss of misclassification of the training samples directly*.

> *与SVM相比，TWSVM具有以下特点：*

```
i) 对于每个类别，它分别构造一个超平面 $f_k = \langle w_k, z \rangle + b_k = 0$。TWSVM中每个优化子问题的规模小于SVM的优化问题规模。同时，我们期望TWSVM中的子问题比SVM中的整体问题更简单。
ii) $f_k$ 是通过衡量类内相似性和类间相异性获得的，新样本根据其最接近的$f_k$所属的类别进行划分。也就是说，每个模型以及最终决策中所用的相似性/相异性度量标准可能不一致，因此难以直接衡量训练样本的误分类损失。
```

From the above characteristics, we can see that TWSVM divides the classification into sub-models, it is more flexible in constructing the small-size sub-models. However, *the inconsistency of the similarity/dissimilarity in TWSVM lets it hard to estimate the misclassification rate of the training samples and probability output*.

> *从以上特点可以看出，TWSVM将分类问题分解为多个子模型，在构建小规模子模型方面更具灵活性。然而，TWSVM中相似性/相异性度量的不一致性，导致其难以估计训练样本的误分类率并输出概率。*

3. Conditional probability twin support vector machines

3.1 CPTWSVM for binary problem

Paragraphs1

Firstly, consider the binary classification problem, we estimate the conditional probability of each class as $p(y=1|x) = f_1(z) = \langle w_1, z\rangle + b_1$ for positive class and $p(y=-1|x) = f_2(z) = \langle w_2, z\rangle + b_2$ for negative class. Specifically, we estimate them from the training set separately as

$$\begin{aligned} & min_{w_1,b_1,\xi}\frac{1}{2}(||w_1||^2 + b_1^2) + C_1\sum_{i\in I^1}\xi_i - C_2\sum_{i\in I} y_i(\langle w_1, z_i\rangle + b_1) \\ & s.t.\ \langle w_1, z_i\rangle + b_1 \geq 0.5 - \xi_i, \xi_i \geq 0, i \in I^1 \\ & 0 \leq \langle w_1, z_i\rangle + b_1 \leq 1, i \in I \end{aligned} \tag{10}$$

and

$$\begin{aligned} & min_{w_2,b_2,\eta}\frac{1}{2}(||w_2||^2 + b_2^2) + C_3\sum_{i\in I^2}\eta_i + C_4\sum_{i\in I} y_i(\langle w_2, z_i\rangle + b_2) \\ & s.t.\ \langle w_2, z_i\rangle + b_2 \geq 0.5 - \eta_i, \eta_i \geq 0, i \in I^2 \\ & 0 \leq \langle w_2, z_i\rangle + b_2 \leq 1, i \in I \end{aligned} \tag{11}$$

where C_1, C_2, C_3 and C_4 are positive parameters.

> *首先，考虑二分类问题，我们估计每个类别的条件概率：对于正类，$p(y=1 \mid x) = f_1(z) = \langle w_1, z\rangle + b_1$；对于负类，$p(y=-1 \mid x) = f_2(z) = \langle w_2, z\rangle + b_2$。具体来说，我们分别从训练集中估计它们，如(10)，（11）所示。其中 C_1、C_2、C_3 和 C_4 是正参数。*

Their geometric meaning are clear: for (10), its last constraints make conditional probability $f_1(z_i)$ is in $[0,1]$ for training samples, and if the value of sample in class $+1$ is smaller than 0.5, than it is misclassified, and we give it a penalty ξ in the first constraint.

> *它们的几何意义很明确：对于 (10)，其最后一个约束条件使得训练样本的条件概率 $f_1(z_i)$ 在 $[0,1]$ 区间内，并且如果正类中样本的值小于 0.5，则它被错误分类，我们在第一个约束条件中给予它一个惩罚 ξ。*

The first term of objective function is the regularized term, *the second term minimizes the misclassified positive samples*, and *the third term maximizes the conditional probability of positive class and minimizes the conditional probability of negative class together*, also called the discriminant term. For (11), it has a similar geometric meaning. For the above models, we can see that *all the measures about samples are in probability space, and if positive samples are misclassified, then it will be punishment in (10) and if negative samples are misclassified, then it will be punishment in (11)*.

> *目标函数的第一项是正则化项，第二项最小化被错误分类的正样本数量，第三项共同最大化正类的条件概率并最小化负类的条件概率，也称为判别项。对于 (11)，它具有类似的几何意义。对于上述模型，我们可以看到所有关于样本的度量都是在概率空间中进行，如果正样本被错误分类，则会在 (10) 中受到惩罚；如果负样本被错误分类，则会在 (11) 中受到惩罚。*

Paragraphs2

The above problems (10) and (11) can be solved by their dual forms. Let $\tilde{w}_k = [w_k; b_k]^T$, $k = 1, 2$ and $\tilde{z}_i = [z_i; 1]^T$, $i \in I$. The Lagrange function of the problem (10) is given by

$$
\begin{aligned}
&L(\tilde{w}_1, \xi, \alpha, \beta, \gamma, \delta) \\
&= \frac{1}{2}\tilde{w}_1^\top \tilde{w}_1 + C_1 \sum_{i\in I^1}\xi_i - C_2 \sum_{i\in I} y_i\langle \tilde{w}_1, \tilde{z}_i\rangle - \sum_{i\in I^1}\alpha_i(\langle\tilde{w}_1,\tilde{z}_i\rangle - 0.5 + \xi_i) \\
&- \sum_{i\in I}\beta_i\langle\tilde{w}_1,\tilde{z}_i\rangle + \sum_{i\in I}\gamma_i(\langle\tilde{w}_1,\tilde{z}_i\rangle - 1) - \sum_{i\in I^1}\delta_i\xi_i
\end{aligned} \quad (12)
$$

> 上述问题 (10) 和 (11) 可以通过它们的对偶形式求解。令 $\tilde{w}_k = [w_k; b_k]^T$ $k = 1, 2$ 且 $\tilde{z}_i = [z_i; 1]^T$, $i \in I$。问题 (10) 的拉格朗日函数由（12）给出。

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{m_1})^T$, $\beta = (\beta_1, \beta_2, \cdots, \beta_m)^T$, $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_m)^T$ and $\delta = (\delta_1, \delta_2, \cdots, \delta_{m_1})^T$ are the Lagrangian multiplier vectors. The KKT conditions for the problem (10) are given by

$$
\begin{gathered}
\nabla_{\tilde{w}_1} L = \tilde{w}_1 - \sum_{i\in I^1}\alpha_i\tilde{z}_i - \sum_{i\in I}\beta_i\tilde{z}_i + \sum_{i\in I}\gamma_i\tilde{z}_i - C_2\sum_{i\in I} y_i\tilde{z}_i = 0 \quad (13) \\
\nabla_{\xi_i} L = C_1 - \alpha_i - \delta_i = 0, i \in I^1 \quad (14) \\
\langle\tilde{w}_1, \tilde{z}_i\rangle + \xi_i \geq 0.5, \xi_i \geq 0, i \in I^1 \quad (15) \\
0 \leq \langle w_1, z_i\rangle + b_1 \leq 1, i \in I \quad (16) \\
\alpha_i(\langle\tilde{w}_1, \tilde{z}_i\rangle - \xi_i + 0.5) = 0, \alpha_i \geq 0, i \in I^1 \quad (17) \\
\beta_i\langle\tilde{w}_1, \tilde{z}_i\rangle = 0, \beta_i \geq 0, i \in I \quad (18) \\
\gamma_i(\langle\tilde{w}_1, \tilde{z}_i\rangle - 1) = 0, \gamma_i \geq 0, i \in I \quad (19) \\
\delta_i\xi_i = 0, i \in I^1 \quad (20)
\end{gathered}
$$

> 其中$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{m_1})^T$，$\beta = (\beta_1, \beta_2, \cdots, \beta_m)^T$, $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_m)^T$ 和 $\delta = (\delta_1, \delta_2, \cdots, \delta_{m_1})^T$ 是拉格朗日乘子向量。问题 (10) 的 KKT 条件由(13)到(20)给出。

Take the above KKT conditions into (12), then we obtain the dual QPP for (10) as

$$
\begin{gathered}
min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i,i'\in I^1, j,j'\in I}\langle\alpha_i\tilde{z}_i + \beta_j\tilde{z}_j - \gamma_j\tilde{z}_j, \alpha_{i'}\tilde{z}_{i'} + \beta_{j'}\tilde{z}_{j'} - \gamma_{j'}\tilde{z}_{j'}\rangle \\
-\frac{1}{2}\sum_{i\in I^1}\alpha_i + C_2\sum_{i\in I^1, j\in I} y_j\alpha_i\langle\tilde{z}_j, \tilde{z}_i\rangle + C_2\sum_{j,j'\in I} y_j(\beta_j - \gamma_j)\langle\tilde{z}_j, \tilde{z}_{j'}\rangle + \sum_{j\in I}\gamma_j \quad (21) \\
s.t. 0 \leq \alpha_i \leq C_1, i \in I^1, \beta_j \geq 0, \gamma_j \geq 0, j \in I.
\end{gathered}
$$

Paragraphs3

The above problem only has the operation of the inner product for samples, and the kernel methods could be used directly similar to that of SVM. Once the solutions to (α, β, γ) are obtained, then

$$
\begin{gathered}
w_1 = \sum_{i\in I^1}\alpha_i z_i + \sum_{i\in I}(\beta_i - \gamma_i + C_2 y_i)z_i \quad (22) \\
b_1 = \sum_{i\in I^1}\alpha_i + \sum_{i\in I}(\beta_i - \gamma_i + C_2 y_i). \, (23)
\end{gathered}
$$

In a similar way, we obtain the dual of the problem (11) as

$$
\begin{gathered}
min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i,i'\in I^2, j,j'\in I}\langle\alpha_i\tilde{z}_i + \beta_j\tilde{z}_j - \gamma_j\tilde{z}_j, \alpha_{i'}\tilde{z}_{i'} + \beta_{j'}\tilde{z}_{j'} - \gamma_{j'}\tilde{z}_{j'}\rangle \\
-\frac{1}{2}\sum_{i\in I^2}\alpha_i + C_4\sum_{i\in I^2, j\in I} y_j\alpha_i\langle\tilde{z}_j, \tilde{z}_i\rangle + C_4\sum_{j,j'\in I} y_j(\beta_j - \gamma_j)\langle\tilde{z}_j, \tilde{z}_{j'}\rangle + \sum_{j\in I}\gamma_j \quad (24) \\
s.t. 0 \leq \alpha_i \leq C_3, i \in I^2, \beta_j \geq 0, \gamma_j \geq 0, j \in I.
\end{gathered}
$$

And when the solutions to (α, β, γ) in (24) are obtained, we have

$$
\begin{gathered}
w_2 = \sum_{i\in I^2}\alpha_i z_i + \sum_{i\in I}(\beta_i - \gamma_i + C_4 y_i)z_i \quad (25) \\
b_2 = \sum_{i\in I^2}\alpha_i + \sum_{i\in I}(\beta_i - \gamma_i + C_4 y_i). \, (26)
\end{gathered}
$$

> 将上述 KKT 条件代入 (12)，我们得到 (10) 的对偶二次规划问题如（21）所示，上述问题只涉及样本的内积运算，可以直接使用类似 SVM 的核方法。一旦获得 (α, β, γ) 的解，则有（22）（23）所示公式。类似地，我们得到问题 (11) 的对偶形式（24）（25）（26）.

Paragraphs4

A new sample $x^* \in R^n$ is assigned to class $y(y = +1, -1)$, depending on which of the two function is larger, i.e.,
$Class\ y = argmax_{k=1,2}\langle w_k, \varphi(x^*)\rangle + b_k$ (27).

> 一个新的样本 $x^* \in R^n$ 被分配到类别 $y(y = +1, -1)$，取决于两个函数中哪个值更大，如（27）所示

Paragraphs5

For CPTWSVM, we point out that for each training sample z_i, $\langle w_k, z_i\rangle + b_k \in [0,1]$, but may be $\langle w_1, z_i\rangle + b_1 + \langle w_2, z_i\rangle + b_2 \neq 1$. This is because each of the (21) and (24) only estimates the conditional probability of the samples belonging to each class, and the probability of not belonging to this class is not estimated.

> *对于 CPTWSVM，我们指出对于每个训练样本 z_i，有$\langle w_k, z_i\rangle + b_k \in [0,1]$，但可能$\langle w_1, z_i\rangle + b_1 + \langle w_2, z_i\rangle + b_2 \neq 1$。这是因为每个问题 (21) 和 (24) 仅估计样本属于每个类别的条件概率，而未估计不属于该类别的概率。*

In decision, we choose the one with the largest value as its probability estimate, this does not affect decision-making and consistency of (27). In addition, for testing sample x^*, $\langle w_k, \varphi(x^*)\rangle + b_k$ may be doesn't belongs to $[0,1]$, $k = 1,2$, which is similar to that of PSVM [36]. We show that only need to specify the values greater than 1 and less than 0 as 1 and 0, then we can estimate the probability of the testing sample, the experimental results show that it is efficient and effective.

> *在决策时，我们选择具有最大值的那个作为其概率估计，这并不影响决策和 (27) 的一致性。此外，对于测试样本 x^*，$\langle w_k, \varphi(x^*)\rangle + b_k$ 可能不属于 $[0,1]$ 区间，这与 PSVM 类似。我们指出，只需将大于 1 和小于 0 的值分别指定为 1 和 0，就可以估计测试样本的概率，实验结果表明这是有效且高效的。*

In addition, CPTWSVM can also directly define the corresponding support vectors through dual variables the support vectors of (21) is defined as the training samples $\{(z_i, y_i)|i \in I\}$ corresponding to $\{0 < \alpha_i \leq C_1, \beta_j > 0, \gamma_j > 0 | i \in I^1, j \in I\}$, and the support vectors of (24) is similar to the above definition.

> *此外，CPTWSVM 也可以通过对偶变量直接定义相应的支持向量：问题 (21) 的支持向量定义为对应于 $\{0 < \alpha_i \leq C_1, \beta_j > 0, \gamma_j > 0 | i \in I^1, j \in I\}$ 的训练样本 $\{(z_i, y_i)|i \in I\}$，问题 (24) 的支持向量定义与上述类似。*

3.2. Block iterative algorithm to problems (21) and (24)

Paragraphs1

We consider the solution of problem (21), where the solution of problem (24) can be obtained in the same way. Problem (21) is a QPP that can be solved by some standard QPP solvers. However, compared with the dual problem (2) of SVM, the size of problem (21) is double plus the size of the positive samples. Too many variables put more burdens on storage and computing, so, we introduce a block iterative algorithm to reduce the computer memory requirement.

> *我们考虑问题 (21) 的求解，问题 (24) 的求解方法相同。问题 (21) 是一个二次规划问题，可以通过一些标准的 QPP 求解器求解。然而，与 SVM 的对偶问题 (2) 相比，问题 (21) 的规模是正类样本数量的两倍多。过多的变量给存储和计算带来了更多负担，因此，我们引入一种块迭代算法来减少计算机内存需求。*

In iterations, we split the variables into three parts, α, β and γ，

$$\alpha^{(t)} = argmin_{0\leq\alpha\leq C_1}\frac{1}{2}\sum_{i,i\prime\in I^1}\alpha_i\langle\tilde{z}_i, \tilde{z}_{i\prime}\rangle\alpha_{i\prime} +$$

$$\sum_{i\in I^1, j\in I}\alpha_i\langle\tilde{z}_i, \tilde{z}_j\rangle(\beta_j^{(t-1)} - \gamma_j^{(t-1)}) + C_2\sum_{i\in I^1, j\in I}y_i\alpha_i\langle\tilde{z}_i, \tilde{z}_j\rangle - \frac{1}{2}\sum_{i\in I^1}\alpha_i \quad (28)$$

$$\beta^{(t)} = argmin_{\beta\geq 0}\frac{1}{2}\sum_{j,j\prime\in I}\beta_j\langle\tilde{z}_j, \tilde{z}_{j\prime}\rangle\beta_{j\prime} +$$

$$\sum_{i\in I^1, j\in I}\beta_j\langle\tilde{z}_i, \tilde{z}_j\rangle(\alpha_i^{(t)} - \gamma_j^{(t-1)}) + C_2\sum_{j,j'\in I}y_j\beta_j\langle\tilde{z}_j, \tilde{z}_{j'}\rangle \quad (29)$$

$$\gamma^{(t)} = argmin_{\gamma\geq 0}\frac{1}{2}\sum_{j,j\prime\in I}\gamma_j\langle\tilde{z}_j, \tilde{z}_{j\prime}\rangle\gamma_{j\prime} -$$

$$\sum_{i\in I^1, j\in I}\gamma_j\langle\tilde{z}_i, \tilde{z}_j\rangle(\alpha_i^{(t)} + \beta_j^{(t)}) - C_2\sum_{j,j'\in I}y_j\gamma_j\langle\tilde{z}_j, \tilde{z}_{j'}\rangle + \sum_{j\in I}\gamma_i \quad (30)$$

> *在迭代中，我们将变量分为三部分：α、β 和 γ，*

where $\{(x_i, y_i)|i \in I\}, I = I^1 \cup I^2$, $z = \varphi(x)$, $\tilde{z}_i = [z_i; 1]^T$, and $C_1, C_2 > 0$. In each iteration, we fix two of them, solve the other one, and the stop condition is

$$\frac{\sqrt{\sum_{i\in I^1}(\alpha_i^{(t)} - \alpha_i^{(t-1)})^2 + \sum_{j\in I}(\beta_j^{(t)} - \beta_j^{(t-1)})^2 + \sum_{j\in I}(\gamma_i^{(t)} - \gamma_i^{(t-1)})^2}}{\sum_{i\in I^1}(\alpha_i^{(t-1)})^2 + \sum_{j\in I}(\beta_j^{(t-1)})^2 + \sum_{j\in I}(\gamma_j^{(t-1)})^2} < tol \quad (31)$$

here tol is a small tolerance. Due to the subproblems about α, β and γ are also QPPs, they could be solved efficiently by some standard QPP solvers or some fast solvers, such as the dual coordinate descent (DCD) algorithm [44–46]. Algorithm 1 exhibits the detailed procedure of solving algorithm.

> 其中 $\{\langle x_i, y_i\rangle \mid i \in I\}$，$I = I^1 \cup I^2$，$z = \varphi(x)$，$\tilde{z}_i = [z_i; 1]^T$，且 $C_1, C_2 > 0$。在每次迭代中，我们固定其中两个变量，求解另一个变量，停止条件为（31）。这里 tol 是一个小的容差。由于关于 α、β 和 γ 的子问题也是 QPPs，它们可以通过一些标准的 QPP 求解器或一些快速求解器（例如对偶坐标下降法）高效求解。Algorithm 1 展示了求解的详细过程。

3.3. CPTWSVM for multiclass problem

Paragraphs1

Consider K classes multiclass classification problem, we estimate the conditional probability function $p(y = k|x) = f_k(z) = \langle w_k, z\rangle + b_k$ for each class, where $z = \varphi(x)$ and $k \in \{1, \cdots, K\}$. Then, we estimate them from the training set as

$$\begin{aligned} min_{w_k, b_1, \xi^{(k)}} \frac{1}{2}(||w_k||^2 + b_k^2) + C_1 \sum_{i \in I^k} \xi_i^{(k)} - C_2 \sum_{i \in I} y_i(\langle w_k, z_i\rangle + b_k) \\ s.t. \langle w_k, z_i\rangle + b_k \geq 0.5 - \xi_i^{(k)}, \xi_i^{(k)} \geq 0, i \in I^k \\ 0 \leq \langle w_k, z_i\rangle + b_k \leq 1, i \in I \end{aligned} \tag{32}$$

> 考虑一个K类多分类问题，我们为每个类别估计条件概率函数 $p(y = k|x) = f_k(z) = \langle w_k, z\rangle + b_k$，其中 $z = \varphi(x)$ 且 $k \in 1, \cdots, K$ 。然后，我们从训练集中估计它们，如(32)所示

where C_1 and C_2 are positive parameters. The geometric meaning of (32) is similar to that of binary models, its last constraints make conditional probability $f_k(z_i)$ is in $[0, 1]$ for training samples, and *if the value of sample in class k is smaller than 0.5, then it is misclassified, and we give it a penalty ξ in the first constraint*. The first term of objective function is the regularized term, the second term minimizes the misclassified positive samples, and the third term is the linear discriminant term, which maximums the conditional probability of the class k and minimizes the conditional probability of the other classes together.

> 其中 C_1 和 C_2 是正参数。公式（32）的几何意义与二分类模型类似：其最后一个约束确保训练样本的条件概率 $f_k(z_i)$ 位于 $[0, 1]$ 区间内；如果类别 k 中样本的函数值小于 0.5，则视为误分类，并在第一个约束中通过 ξ 给予惩罚。目标函数的第一项为正则化项，第二项最小化误分类的正样本数量，第三项为线性判别项，其共同最大化类别 的条件概率并最小化其他类别的条件概率。

Paragraphs2

Let $\tilde{w}_k$ = $[w_k; b_k]^T$, $k = 1, \ldots, K$ and $\tilde{z}_i = [z_i; 1]^T$, $i \in I$. The primal problem of CPTWSVM for multiclass problem (32) can be solved by the dual form as

$$\begin{aligned} min_{\alpha, \beta, \gamma} \frac{1}{2} \sum_{i \in I^k, j \in I} \langle \alpha_i \tilde{z}_i + \beta_j \tilde{z}_j - \gamma_j \tilde{z}_j, \alpha_i \tilde{z}_i + \beta_j \tilde{z}_j - \gamma_j \tilde{z}_j\rangle - \frac{1}{2} \sum_{i \in I^k} \alpha_i + \\ C_2 \sum_{i \in I^k, j \in I}^{m_k} y_j \alpha_i \langle \tilde{z}_j, \tilde{z}_i\rangle + C_2 \sum_{j, j' \in I} y_j(\beta_j - \gamma_j)\langle \tilde{z}_j, \tilde{z}_{j\prime}\rangle + \sum_{j \in I} \gamma_j \\ s.t. 0 \leq \alpha_i \leq C_1, i \in I^k, \beta_j \geq 0, \gamma_j \geq 0, j \in I \end{aligned} \tag{33}$$

> 令 $\tilde{w}_k = [w_k; b_k]^T$ （$k = 1, \ldots, K$）和 $\tilde{z}_i = [z_i; 1]^T$ （$i \in I$）。多分类问题CPTWSVM的原始问题(32)可通过其对偶形式求解，如公式（33）所示：

The above (33) is similar to that of (21), *its solution (α, β, γ) could be obtained in the same way*, and $f_k(x)$ is calculated as

$$w_k = \sum_{i \in I^k} \alpha_i z_i + \sum_{i \in I} (\beta_i - \gamma_i + C_2 y_i) z_i, (34)$$

$$b_k = \sum_{i \in I^k} \alpha_i + \sum_{i \in I} (\beta_i - \gamma_i + C_2 y_i). (35)$$

> 上述公式(33)与公式(21)类似，其解 (α, β, γ) 可通过相同方式获得，且 $f_k(x)$ 的计算如 (34) 和 (35)所示

Similarly, the support vectors of (33) is defined as the training samples $\{(z_i, y_i)|i \in I\}$ corresponding to $\{0 < \alpha_i^{(k)} \leq C_j, \beta_j^{(k)} > 0, \gamma_j^{(k)} > 0|i \in I^k, j \in I\}$. A new sample $x^* \in R^n$ is assigned to class $k(k = 1, 2, \ldots, K)$, depending on which of the function is larger, i.e.,

$$Class\ y = argmax_{k=1,\ldots,K, i \in I^k} \langle w_k, \varphi(x^*)\rangle + b_k \ \ (36)$$

> 类似地，公式(33)的支持向量定义为对应于 $\{0 < \alpha_i^{(k)} \leq C_j, \beta_j^{(k)} > 0, \gamma_j^{(k)} > 0 \mid i \in I^k, j \in I\}$ 的训练样本 $(z_i, y_i) \mid i \in I$。一个新样本 $x^* \in \mathbb{R}^n$ 被分配到类别 k（$k = 1, 2, \ldots, K$），取决于哪个函数值最大，即公式(36)

3.4 Discussions

Paragraphs1

Now, we discuss the main characteristics of CPTWSVM. Here, we show some properties mainly compared to TWSVM and SVM. *Compared to TWSVM*:

> *现在，我们讨论 CPTWSVM 的主要特性。这里，我们主要展示与 TWSVM 和 SVM 相比的一些属性。与 TWSVM 相比：*

i) For each class, *it constructs an $f_k(z)$ to estimate $p(y=k|x)$, and all $f_k(z)$ are constructed separately*, while TWSVM construct hyperplanes. For each subproblem, CPTWSVM *measures the loss of misclassification rate of whether it belongs to this class k*, while TWSVM cannot measure the loss of misclassification.

> *i) 对于每个类别，它构建一个 $f_k(z)$ 来估计 $p(y=k \mid x)$，并且所有 $f_k(z)$ 是分开构建的，而 TWSVM 构建超平面。对于每个子问题，CPTWSVM 度量样本是否属于此类 k 的错误分类损失，而 TWSVM 无法度量错误分类损失。*

ii) In CPTWSVM, *all of the computation of z_i is based on the inner product*, so the model *can be extended to kernel method in RKHS directly*. While, TWSVM *needs the representation theory to extend to kernel, and the dual problems of TWSVM need to compute the inverse of data matrix*.

> *ii) 在 CPTWSVM 中，所有关于 z_i 的计算都基于内积，因此该模型可以直接扩展到 RKHS 中的核方法。而 TWSVM 需要表示理论来扩展到核方法，并且 TWSVM 的对偶问题需要计算数据矩阵的逆。*

iii) For each class, *$f_k(z) \approx p(y=k|x)$ is estimated by measuring the probability*, and the new sample is assigned to the class with the biggest probability $f_k(z)$, *the measurement of loss for each class and decision is consistent*. While in TWSVM, *the distances of the within-class and between-class to the hyperplanes are not the same*.

> *iii) 对于每个类别，$f_k(z) \approx p(y=k \mid x)$ 是通过度量概率来估计的，新样本被分配到具有最大概率 $f_k(z)$ 的类别，每个类别的损失度量和决策是一致的。而在 TWSVM 中，类内和类间到超平面的距离度量方式不同。*

iv) The output of CPTWSVM is bounded, and at the same time, its kernel matrix is bounded, which is more matching the statistical learning theory. While the output of the TWSVM is unbounded. Due to the bounded constraints, the optimization problem size of CPTWSVM is larger than that of TWSVM, but the *inverse of **quadratic term matrixes(二次项)** need be solved in TWSVM and its computational complexity is very high*.

> *iv) CPTWSVM 的输出是有界的，同时其核矩阵也是有界的，这更符合统计学习理论。而 TWSVM 的输出是无界的。由于有界约束，CPTWSVM 的优化问题规模比 TWSVM 大，但 TWSVM 需要求解二次项矩阵的逆，其计算复杂度非常高。*

Paragraphs2

Compared to SVM: i) Similar to that of $p(y=1|x)$ is introduced in PSVM compared to SVM, the $p(y=k|x)$ is introduced in CPTWSVM compared to TWSVM. CPTWSVM maintains the properties of support vectors and only calculates the inner product of samples in the duality of SVM. However, it estimates $p(y=k|x)$ separately.

> *与 SVM 相比：i) 类似于 PSVM 相比 SVM 引入了 $p(y=1 \mid x)$，CPTWSVM 相比 TWSVM 引入了 $p(y=k \mid x)$。CPTWSVM 保持了支持向量的特性，并且在 SVM 的对偶形式中只计算样本的内积。然而，它是分别估计 $p(y=k \mid x)$ 的。*

ii) Compared to PSVM, the max $\sum_{i\in I} y_i(\langle w, z_i\rangle + b)$ is introduced in CPTWSVM, it maximizes the within-class similarity and minimizes the between-class dissimilarity. In fact, it is easy to extend to PSVM as

$$\begin{gathered}
min_{w,b,\xi}\frac{1}{2}||w||^2 + \frac{C_1}{\varepsilon}\sum_{i\in I}\xi_i - C_2\sum_{i\in I} y_i(\langle w, z_i\rangle + b) \\
s.t.\ \langle w, z_i\rangle + b \geq 0.5\varepsilon - \xi_i,\ \xi_i \geq 0, i \in I^1 \\
\langle w, zi\rangle + b \leq 0.5\varepsilon + \xi_i,\ \xi_i \geq 0, i \in I^2 \\
0 \leq \langle w, z_i\rangle + b \leq 1, i \in I
\end{gathered} \tag{37}$$

where C_1 and C_2 are positive parameters.

> *ii) 与 PSVM 相比，CPTWSVM 引入了 $max \sum_{i\in I} y_i(\langle w, z_i\rangle + b)$，它最大化类内相似性并最小化类间相异性。实际上，可以很容易地将其扩展为 PSVM，如(37)所示，其中 C_1 和 C_2 是正参数。*

To show *the effectiveness of the within-class similarity and between-class dissimilarity term*, we conduct the compared experiments on this model (called CPSVM) too. The dual form of CPSVM is

$$
\begin{aligned}
&min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i,j\in I}(\alpha_i y_i+\beta_i-\gamma_i)^\top(\alpha_j y_j+\beta_j-\gamma_j)\langle z_i,z_j\rangle\\
&+C_2\sum_{i,j\in I}(y_i y_j\alpha_j+y_i\beta_j-y_i\gamma_j)\langle z_i,z_j\rangle+\sum_{i\in I}(\gamma_i-0.5\times\alpha_i(y_i+\varepsilon))\\
&s.t.\sum_{i\in I}(\alpha_i y_i+\beta_i-\gamma_i)=-C_2\sum_{i\in I}y_i,i\in I,\\
&0\leq\alpha_i\leq\frac{C_1}{\varepsilon},\beta_i\geq 0,\gamma_i\geq 0,i\in I.
\end{aligned}
\tag{38}
$$

> *为了展示类内相似性和类间相异性项的有效性，我们也对这个模型（称为 CPSVM）进行了对比实验。CPSVM 的对偶形式是*

From (38), we can see that *the dual problem size of CPSVM is the same as PSVM, and each sub-problem of CPTWSVM is smaller than CPSVM*. For the multiclass classification problem, the *one-vs-all strategy multiclass CPSVM is similar to that of multiclass CPTWSVM*, but with more variables.

> *从 (38) 可以看出，CPSVM 的对偶问题规模与 PSVM 相同，而 CPTWSVM 的每个子问题规模小于 CPSVM。对于多分类问题，使用一对多策略的多分类 CPSVM 与多分类 CPTWSVM 类似，但变量更多。*

4. Experiments

In this section, we focused on comparison of the proposed conditional probability models (CPTWSVM and CPSVM) and the state-of-the-art SVM and TWSVM classifiers (SVM [47], PSVM [36], TWSVM [2], LSTSVM [6]). All methods are implemented by using Matlab R2020a on a desktop with 8 Intel® CoreTM i7-7700K processors (4.20 GHz)and 32GB RAM.

> *在本节中，我们重点比较所提出的条件概率模型（CPTWSVM 和 CPSVM）与最先进的 SVM 和 TWSVM 分类器（SVM [47]、PSVM [36]、TWSVM [2]、LSTSVM [6]）。所有方法均在配备 8 个 Intel® Core™ i7-7700K 处理器（4.20 GHz）和 32GB RAM 的台式机上，使用 Matlab R2020a 实现。*

4.1 Experiment setup

4.1.1. Parameter setting

The parameters of all 6 classifiers (with linear kernel) are displayed in Table 1. The trade-off parameters C, C_1, C_2, C_3, C_4 are selected from $\{2^{-8},\ldots,2^{8}\}$, and for PSVM and CPSVM, the parameter ε is tuned from $\{2^{-8},\ldots,2^{0}\}$. In order to utilize kernel trick [48], the radius basic function (rbf) kernel $K(x,x\prime)=exp(-\mu||x-x\prime||^2)$ is used, and the kernel parameter μ is obtained by searching in the range $\{2^{-6},\ldots,2^{6}\}$. Besides, we set $C_1=C_3, C_2=C_4$ in TWSVM and CPTWSVM for the sake of *reducing training time*.

> *所有 6 个分类器（使用线性核）的参数如表 1 所示。权衡参数 C, C_1, C_2, C_3, C_4 从 $\{2^{-8},\ldots,2^{8}\}$ 中选取，对于 PSVM 和 CPSVM，参数 ε 从 $\{2^{-8},\ldots,2^{0}\}$ 中调整。为了利用核技巧 [48]，使用径向基函数（rbf）核 $K(x,x\prime)=exp(-\mu||x-x\prime||^2)$，核参数 μ 通过在 $\{2^{-6},\ldots,2^{6}\}$ 范围内搜索获得。此外，为了减少训练时间，我们在 TWSVM 和 CPTWSVM 中设置 $C_1=C_3, C_2=C_4$。*

4.1.2. Evaluation metrics

In order to *evaluate the performance of classifiers*, the *confusion matrix and several evaluation criteria* are introduced. Confusion matrix is shown in Table 2. TP, TN, FP, FN are the number of correctly classified positive samples, negative samples that are correctly predicted, negative samples that are misclassified as positive samples, positive samples that are misclassified as negative samples, respectively.

- Acc: Accuracy (Acc) represents the proportion of correctly classified samples, which is defined as $Acc=100\%\times\frac{(TP+TN)}{(TP+FP+TN+FN)}$. The optimal parameters are obtained by conducting the standard 10 fold cross validation [43] according to the Acc, and the reported testing Acc is the 10 times average Acc under the optimal parameters.
- SVs: SVs is the number of support vectors, the corresponding definition for support vectors is described above.
- TPR & FPR: The true positive rate (TPR), also called hit rate or recall, is the probability of the actual positive samples that are tested positive. Likewise, the false positive rate (FPR), also called false alarm rate, is the probability of actual negative samples that are tested positive. Their definitions are as follows $TPR=\frac{TP}{TP+FN}, FPR=\frac{FP}{FP+TN}$.
- ROC & AUC: The receiver operating characteristics (ROC) [49,50] graph *depicts relative tradeoffs between benefits (TP) and costs (FP), in which TPR is the vertical axis and FPR is the horizontal axis*. The AUC value is the area under the ROC curve. The ROC and AUC have the advantage of visualizing and comparing the classifier performance without regard to class distributions or error costs.

> *为了评估分类器的性能，引入了混淆矩阵和若干评估标准。**混淆矩阵** 如表 2 所示。TP、TN、FP、FN 分别表示正确分类的正样本数、正确预测的负样本数、被误分类为正样本的负样本数、被误分类为负样本的正样本数。*

> - *Acc：准确率（Acc）表示正确分类样本的比例，其定义为$Acc = 100\% \times \frac{(TP+TN)}{(TP+FP+TN+FN)}$。通过执行标准的10折交叉验证 [43] 并根据 Acc 确定最优参数，报告的测试 Acc 是在最优参数下 10 次平均的 Acc。*
> - *SVs：SVs 是支持向量的数量，支持向量的相应定义如上所述。*
> - *TPR 和 FPR：真正例率（TPR），也称为命中率或召回率，是实际正样本被检测为阳性的概率。同样，假正例率（FPR），也称为误报率，是实际负样本被检测为阳性的概率。它们的定义如下：*

```
$TPR = \frac{TP}{TP + FN}$, $FPR = \frac{FP}{FP + TN}$
```

> *-ROC 和 AUC：接收者操作特征（ROC）图 [49, 50] 描述了收益（TP）和成本（FP）之间的相对权衡，其中 TPR 是纵轴，FPR 是横轴。AUC 值是 ROC 曲线下的面积。ROC 和 AUC 的优点在于可以在不考虑类别分布或错误成本的情况下可视化和比较分类器性能。*

4.2. Experimental results on the benchmark datasets

Paragraphs1

Table 3 lists the basic description of 25 benchmark datasets, where m, n and k are the number of samples, features and categories, respectively. The classification results for benchmark datasets are summarized in Tables 4 and 5, the best Acc is shown in boldface.

> *表 3 列出了 25 个基准数据集的基本描述，其中 m、n 和 k 分别是样本数、特征数和类别数。基准数据集的分类结果总结在表 4 和表 5 中，最佳 Acc 以粗体显示。*

Paragraphs2

Linear results: From Table 4, it is easy to see that the Acc of linear CPTWSVM is slightly better than that of linear SVM, PSVM, CPSVM, TWSVM and LSTSVM on the majority of datasets, and CPTWSVM obviously gains higher Accs than LSTSVM on 22 out of 25 datasets. It is noteworthy that the Acc of CPSVM is slightly higher than that of PSVM on the whole, indicating that adding the discriminant term $\sum_{i \in I} y_i(\langle w, z_i \rangle + b)$ on the basis of PSVM can effectively improve the classification performance. Meanwhile, the Acc of CPTWSVM is slightly higher than CPSVM, which indicates that TWSVM paradigm plays an important role in classification.

> ***线性结果**：从表 4 可以很容易地看出，在大多数数据集上，线性 CPTWSVM 的 Acc 略优于线性 SVM、PSVM、CPSVM、TWSVM 和 LSTSVM，并且 CPTWSVM 在 25 个数据集中的 22 个上明显比 LSTSVM 获得更高的 Acc。值得注意的是，CPSVM 的 Acc 总体上略高于 PSVM，表明在 PSVM 的基础上添加判别项 $\sum_{i \in I} y_i(\langle w, z_i \rangle + b)$ 可以有效提高分类性能。同时，CPTWSVM 的 Acc 略高于 CPSVM，这表明 TWSVM 范式在分类中起着重要作用。*

Paragraphs3

Table 4 also compares SVs for all classifiers. It can be seen that LSTSVM always has the most SVs, this is because LSTSVM uses the least square loss function, which is not sparse. While TWSVM has more SVs than CPTWSVM, this is because TWSVM also has a least square loss function, but CPTWSVM only has L1 sparse loss.

> *表 4 还比较了所有分类器的 SVs。可以看出，LSTSVM 总是拥有最多的 SVs，这是因为 LSTSVM 使用最小二乘损失函数，该函数不具有稀疏性。而 TWSVM 的 SVs 比 CPTWSVM 多，这是因为 TWSVM 也有一个最小二乘损失函数，但 CPTWSVM 只有 L_1 稀疏损失。*

Paragraphs4

Further, it can be found from Table 4 that *the training time of CPTWSVM is lower than CPSVM but higher than TWSVM on most cases*, this is because *the scale of the QPP in CPTWSVM is smaller than that of CPSVM but higher than that of TWSVM*.

> *进一步，从表 4 中可以发现，在大多数情况下，CPTWSVM 的训练时间低于 CPSVM 但高于 TWSVM，这是因为 CPTWSVM 中的 QPP 规模小于 CPSVM 但高于 TWSVM。*

Nonlinear results: Table 5 displays the results of nonlinear kernel for all classifiers on the 25 benchmark datasets. It is obvious that the Acc of nonlinear CPTWSVM is slightly better than that of other nonlinear classifiers. Similarly, CPTWSVM gains higher Accs than LSTSVM on 21 out of 25 datasets, and therefore confirms the conclusion above further.

> ***非线性结果**：表 5 显示了所有分类器在 25 个基准数据集上的非线性核结果。很明显，非线性 CPTWSVM 的 Acc 略优于其他非线性分类器。类似地，CPTWSVM 在 25 个数据集中的 21 个上获得了比 LSTSVM 更高的 Acc，这进一步证实了上述结论。*

Paragraphs5

Non-parametric statistical analysis: Fig. 1 presents the critical difference (CD) diagrams of the Friedman test with Nemenyi posthoc test [52] at significance level $\alpha = 10\%$. Let R_j be the average rank of $j_{th}(k = 6)$ classifiers on $N = 25$ datasets, the Friedman statistic $\chi_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}] = 27.71$ is larger than the critical value $\chi_\alpha^2(k-1) = 9.24$, and thus the null hypothesis is rejected, and the Nemenyi post-hoc test follows. The performance of pairwise classifiers will be significantly different if their average ranks differ by at least critical difference value $CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} = 1.37$, where $q_\alpha = 2.589$ for Nemenyi test.

> ***非线性统计分析：** 图 1 展示了在显著性水平 $\alpha = 10\%$ 下，带有Nemenyi事后检验 [52] 的Friedman检验的临界差异（CD）图。设 R_j 为第 j 个分类器（$k = 6$）在 $N = 25$ 个数据集上的平均排名，Friedman统计量 $\chi_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}] = 27.71$ 大于临界值 $\chi_\alpha^2(k-1) = 9.24$，因此拒绝零假设，并进行 Nemenyi 事后检验。如果两个分类器的平均排名差异至少为临界差异值 $CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} = 1.37$，则它们的性能被认为存在显著差异，其中对于Nemenyi检验，$q_\alpha = 2.589$。*

From Fig. 1, we can see that CPTWSVM is ranked first on average and significantly outperforms other classifiers. Note that *those connected classifiers are considered to have little differences between them*.

> *从图 1 可以看出，CPTWSVM 平均排名第一，并且显著优于其他分类器。注意，那些用线连接起来的分类器被认为它们之间差异不大。*

Additionally, *the contrast estimation based on medians [53] is applied over 25 datasets to conduct statistical analysis*. The estimations of linear and nonlinear cases are summarized in Tables 6 and 7, we can clearly see that *CPTWSVM always achieves positive difference values with respect to the benchmark methods*, which validates the statistical significance of CPTWSVM against other classifiers.

> *此外，基于中位数的对比估计 [53] 被应用于 25 个数据集进行统计分析。线性和非线性情况下的估计总结在表 6 和表 7 中，我们可以清楚地看到，CPTWSVM 相对于基准方法总是获得正的差异值，这验证了 CPTWSVM 相对于其他分类器的统计显著性。*

4.3. Projection results for TWSVM, LSTSVM and CPTWSVM

Paragraphs1

To illustrate *the consistency of measurement and classification results of the classifiers on test samples*, Figs. 2 and 3 show two-dimensional and three-dimensional scatter plots of the test samples for *six benchmark datasets* by TWSVM, LSTSVM and CPTWSVM.

> *为了说明分类器在测试样本上度量的一致性和分类结果，图 2 和图 3 分别展示了通过 TWSVM、LSTSVM 和 CPTWSVM 对六个基准数据集的测试样本进行二维和三维散点图绘制。*

Paragraphs2

The plots of TWSVM and LSTSVM are obtained by plotting points with coordinates (distance from hyperplane 1, distance from hyperplane -1), i.e., for a test sample x_i, the distances from hyperplane 1 and -1 are $\frac{|\langle w_1,\varphi(x_i)\rangle+b_1|}{\|w_1\|}$ and $\frac{|\langle w_2,\varphi(x_i)\rangle+b_2|}{\|w_2\|}$, respectively. The plots of CPTWSVM are obtained by plotting test samples with the outputting probability of hyperplane 1 and hyperplane -1, i.e., the probability of hyperplane 1 and -1 are $p(y = 1 \mid x_i) = \langle w_1, \varphi(x_i)\rangle + b_1$ and $p(y = -1 \mid x_i) = \langle w_2, \varphi(x_i)\rangle + b_2$, respectively.

> *TWSVM 和 LSTSVM 的图是通过绘制坐标为（到超平面 1 的距离，到超平面 -1 的距离）的点获得的，即对于一个测试样本 x_i，到超平面 1 和 -1 的距离分别是 $\frac{|\langle w_1,\varphi(x_i)\rangle+b_1|}{\|w_1\|}$ 和 $\frac{|\langle w_2,\varphi(x_i)\rangle+b_2|}{\|w_2\|}$。CPTWSVM 的图是通过绘制测试样本的超平面 1 和超平面 -1 的输出概率获得的，即超平面 1 和 -1 的概率分别是 $p(y = 1 \mid x_i) = \langle w_1, \varphi(x_i)\rangle + b_1$ 和 $p(y = -1 \mid x_i) = \langle w_2, \varphi(x_i)\rangle + b_2$。*

Hence, *the clusters of points indicate how well the classification criterion is able to discriminate between the two classes*. The separating line has also been plotted in Fig. 2, positive samples are plotted as red“+” and negative samples are plotted as blue“×”. *In addition, the red dotted line and blue dotted line in CPTWSVM represent the cumulative distribution curve of positive and negative samples respectively*.

> *因此，点的簇状分布表明分类标准区分两个类别的能力如何。分隔线也在图 2 中绘出，正样本用红色"+"绘制，负样本用蓝色"×"绘制。此外，CPTWSVM 中的红色虚线和蓝色虚线分别代表正样本和负样本的累积分布曲线。*

Paragraphs3

From Fig. 2, it can be seen that, for these binary datasets, *CPTWSVM outputs the projection in the consistency measurement, since CPTWSVM outputs the probability values of two classes, cumulative distribution curves can be easily obtained*. *While the projection distances of the two classes are different, and the decision boundary is asymmetric* in TWSVM and LSTSVM.

> *从图 2 可以看出，对于这些二分类数据集，CPTWSVM 在一致性度量下输出投影，因为 CPTWSVM 输出两个类别的概率值，可以轻松获得累积分布曲线。而在 TWSVM 和 LSTSVM 中，两个类别的投影距离不同，且决策边界不对称。*

Besides, compared with TWSVM and LSTSVM, the *separating line of CPTWSVM lies on the diagonal of 0–1 box, and the output is probability value, which is not affected by the consistency of measurement*, while *TWSVM and LSTSVM are more susceptible to consistency*. In this case, *the two classes are clearly separated in CPTWSVM, while the projections of the two classes are less distinct in TWSVM and LSTSVM*.

> *此外，与 TWSVM 和 LSTSVM 相比，CPTWSVM 的分隔线位于 0-1 盒的对角线上，并且输出是概率值，不受度量一致性的影响，而 TWSVM 和 LSTSVM 更容易受到一致性的影响。在这种情况下，两个类别在 CPTWSVM 中清晰分离，而在 TWSVM 和 LSTSVM 中，两个类别的投影区分度较低。*

Paragraphs4

Figure 3 shows the projection of three-dimensional test samples for Iris and Seeds datasets by TWSVM, LSTSVM and CPTWSVM. Since the one-vs-rest is used, the test samples of each class can output the corresponding decision value, i.e., for a test sample x_i , the distance from the hyperplane $k(k=1,2,3)$ is $\frac{|\langle w_k,\varphi(x_i)\rangle+b_k|}{\|w_k\|}$ and the probability of hyperplane $k(k=1,2,3)$ is $p(y=k|x_i)=\langle w_k,\varphi(x_i)\rangle+b_k$. From Fig. 3, we can see that CPTWSVM, as above, obtains better separation than TWSVM and LSTSVM.

> *图 3 显示了通过 TWSVM、LSTSVM 和 CPTWSVM 对 Iris 和 Seeds 数据集的测试样本进行三维投影。由于使用一对多策略，每个类别的测试样本可以输出相应的决策值，即对于一个测试样本 x_i，到超平面 $k(k=1,2,3)$ 的距离是 $\frac{|\langle w_k,\varphi(x_i)\rangle+b_k|}{\|w_k\|}$，而到超平面 $k(k=1,2,3)$的概率是$p(y=k\mid x_i)=\langle w_k,\varphi(x_i)\rangle+b_k$。从图 3 可以看出，与上述类似，CPTWSVM 获得了比 TWSVM 和 LSTSVM 更好的分离效果。*

4.4. ROC for TWSVM, LSTSVM and CPTWSVM

Figure 4 shows the ROC curves and corresponding AUC values of the above six benchmark datasets by TWSVM, LSTSVM, and CPTWSVM. Note that the ROC curve and AUC of multiclass are complicated, we calculate the TPR and FPR of the each category [49,51], and then averaging TPRs and FPRs of all categories to obtain the macro-average ROC curve and AUC.

> *图 4 显示了通过 TWSVM、LSTSVM 和 CPTWSVM 对上述六个基准数据集的 ROC 曲线和相应的 AUC 值。注意多类别的 ROC 曲线和 AUC 计算较为复杂，我们计算每个类别的 TPR 和 FPR [49,51]，然后对所有类别的 TPR 和 FPR 取平均以获得宏平均 ROC 曲线和 AUC。*

As can be seen from the Fig. 4, ROC curves of CPTWSVM (red line) are generally above that of TWSVM(blue line) and LSTSVM(green line) regardless of binary or multiclass, and AUC values of CPTWSVM are significantly higher than TWSVM and LSTSVM, and therefore above conclusions can also be confirmed.

> *从图 4 可以看出，无论是二分类还是多分类，CPTWSVM（红线）的 ROC 曲线通常位于 TWSVM（蓝线）和 LSTSVM（绿线）之上，并且 CPTWSVM 的 AUC 值显著高于 TWSVM 和 LSTSVM，因此上述结论也得到了证实。*

4.5. The influence of discriminant term on CPSVM and CPTWSVM

Paragraphs1

To explore the influence of the discriminant term $\sum_{i\in I} y_i(\langle w,z_i\rangle+b)$ on CPSVM (37) and CPTWSVM (10), we presented the parameter curve of C_2 on fixing optimal of other parameters, and C_2 is selected from $\{0,2^{-8},\cdots,2^8\}$. Here, $C_2=0$ means without the discriminant term, which is used as a basic to compare whether the linear discriminant term has an impact on the classifier.

> *为了探究判别项 $\sum_{i\in I} y_i(\langle w,z_i\rangle+b)$ 对 CPSVM (37) 和 CPTWSVM (10) 的影响，我们在固定其他参数最优的情况下展示了 C_2 的参数曲线，C_2 从 $\{0,2^{-8},\cdots,2^8\}$ 中选取。这里，$C_2=0$ 表示没有判别项，用作比较线性判别项是否对分类器有影响的基础。*

Paragraphs2

Figure 5 shows the parameter curves of the discriminant term for the above six benckmark datasets, where the abscissa is the range of C2 and the ordinate is the accuracy(%) of classifier, the red and blue lines represent CPTWSVM and CPSVM, respectively. It can be seen from Fig. 5 that the discriminant term is of great meaning to both CPSVM and CPTWSVM.

> *图 5 显示了上述六个基准数据集判别项的参数曲线，其中横坐标是 C_2 的范围，纵坐标是分类器的准确率（%），红线和蓝线分别代表 CPTWSVM 和 CPSVM。从图 5 可以看出，判别项对 CPSVM 和 CPTWSVM 都非常有意义。*

Specifically, the optimal results of CPSVM and CPTWSVM are not available when $C_2 = 0$, indicating that the addition of linear discriminant term is useful. In addition, CPTWSVM generally have the lower Accs at $C_2 = 0$ for six benckmark datasets, which further indicates that the discriminant term has a significant influence on CPTWSVM.

> *具体来说，当 $C_2 = 0$ 时，CPSVM 和 CPTWSVM 都无法获得最优结果，表明添加线性判别项是有用的。此外，对于六个基准数据集，CPTWSVM 在 $C_2 = 0$ 时通常具有较低的 Acc，这进一步表明判别项对 CPTWSVM 有显著影响。*

4.6. Experiments on real-world cases

Paragraphs1

In this section, we present 4 real-world cases in Table 8 and Fig. 6, of which Heart and Cradio are disease detection data with sample sizes of 1025 and 70000, Year1 and Taiwan are credit evaluation data with sample sizes of 7027 and 30000, respectively. It can be seen from Table 8 that the proposed CPTWSVM obtains better performance on 3 of 4 real cases.

> *在本节中，我们在表 8 和图 6 中展示了 4 个真实世界案例，其中 Heart 和 Cradio 是疾病检测数据，样本量分别为 1025 和 70000，Year1 和 Taiwan 是信用评估数据，样本量分别为 7027 和 30000。从表 8 可以看出，所提出的 CPTWSVM 在 4 个真实案例中的 3 个上获得了更好的性能。*

Besides, as shown in Fig. 6, the ROC curve of CPTWSVM is higher than other classifiers in Heart, Cradio and Year1 datasets, which further shows that combining TWSVM paradigm with discriminant term and probability output not only obtains better classification performance, but also have better interpretation.

> *此外，如图 6 所示，在 Heart、Cradio 和 Year1 数据集中，CPTWSVM 的 ROC 曲线高于其他分类器，这进一步表明将 TWSVM 范式与判别项和概率输出相结合不仅获得了更好的分类性能，而且具有更好的可解释性。*

Paragraphs2

In addition, in order to visualize the relationship between accuracy improvements and computing costs, Fig. 7 shows the histograms of accuracy and training time on 4 case datasets. Note that the training time of the histograms is taken as $log(1 + Time(s))$ for the sake of positive values.

> *此外，为了可视化准确率提升与计算成本之间的关系，图 7 显示了 4 个案例数据集上准确率和训练时间的直方图。注意，直方图中的训练时间取为 $\log(1 + Time(s))$ 以保证为正值。*

As shown in Fig. 7, the training time of CPTWSVM is higher than that of TWSVM on 2 of 4 datasets, this is because the constraints of probability output increase the scale of solution, but the accuracy improves significantly, which confirms the above conclusions further.

> *如图 7 所示，在 4 个数据集中的 2 个上，CPTWSVM 的训练时间高于 TWSVM，这是因为概率输出的约束增加了求解规模，但准确率显著提高，这进一步证实了上述结论。*

5. Conclusion

In this paper, a *conditional probability twin SVM model (CPTWSVM)* for *binary and multiclass classification* is presented. CPTWSVM *changes the measurement from distance to probability, overcoming some inconsistency problems in TWSVM*.

> *本文提出了一种用于二分类和多分类的条件概率双支持向量机模型（CPTWSVM）。CPTWSVM 将度量从距离改为概率，克服了 TWSVM 中的一些不一致性问题。*

Further, *the probabilistic output* makes the CPTWSVM has better *interpretability and robustness*. The *subproblems of CPTWSVM* are *QPPs and could be solved by block decomposition algorithm efficiently*.

> *此外，概率输出使 CPTWSVM 具有更好的可解释性和鲁棒性。CPTWSVM 的子问题是二次规划问题（QPP），可以通过块分解算法高效求解。*

Numerical experiments and real applications demonstrate that CPTWSVM outputs the probability estimation and data projection well, resulting in better generalization and interpretability ability than TWSVMs. Our future research focuses on *how to improve the computational efficiency of CPTWSVM and apply probability to the downstream tasks of classification*. For example, the **probability-type knowledge transfer in TWSVM** is very interesting.

- probability-type knowledge transfer in TWSVM

> *数值实验和实际应用表明，CPTWSVM 能很好地输出概率估计和数据投影， 具有比 TWSVMs 更好的泛化能力和可解释性。我们未来的研究重点是如何提高 CPTWSVM 的计算效率并将概率应用于分类的下游任务。例如，TWSVM 中的概率型知识迁移非*

> 常有趣。

笔记 Note

> 这里汇总一下行文思路

算法介绍的思路（取自2~3）

- 经典SVM（软间隔）
  - 原问题：

$$\begin{gathered} min_{w,b,\xi_i}\frac{1}{2}||w||^2+C\sum_{i\in I}\xi_i \\ s.t.\ y_i(\langle w,z_i\rangle+b)\geq 1-\xi_i,\xi_i\geq 0,i\in I \end{gathered} \tag{1}$$

    - 分类正确时$y_i(\langle w,z_i\rangle+b)$为正，反之为负（$y_i$要么 + 1 要么 - 1 不会影响值的大小）
    - 支持向量：ξ_i不为零的向量为支持向量
    - 对新样本 x^*，当 $\langle w,\varphi(x^*)\rangle+b>0$ 时，它被分类为正类。否则，它被分类为负类。
  - 对偶问题：

$$\begin{gathered} min_{\alpha}\frac{1}{2}\sum_{i,j\in I}\langle z_i,z_j\rangle\alpha_i\alpha_j-\sum_j\alpha_j \\ s.t.\sum_{i\in I}y_i\alpha_i=0,0\leq\alpha_i\leq C,i\in I \end{gathered} \tag{2}$$

- PSVM的改进
  - 原问题：

$$\begin{gathered} min_{w,b,\xi}\frac{1}{2}||w||^2+\frac{C}{\varepsilon}\sum_{i\in I}\xi_i, \\ s.t.\ \ y_i(\langle w,z_i\rangle+b-0.5)\geq 0.5\times\varepsilon-\xi_i\ \ \ \xi_i\geq 0,i\in I \\ 0\leq\langle w,z_i\rangle+b\leq 1,i\in I \end{gathered} \tag{3}$$

  - 在PSVM中，根据约束$\langle w,\varphi(x^*)\rangle+b>0.5$是$0\sim 1$之间的有界值
  - 支持向量：ξ_i不为零的向量为支持向量
  - ε是一个小量, 有两个作用: 1. 缩放误分类惩罚 2. 控制分类间隔的宽度
  - 对新样本 x^*，当 $\langle w,\varphi(x^*)\rangle+b>0.5$ 时，它被分类为正类，否则被分类为负类。

```
- 对偶问题：
```

$$\begin{gathered} min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i,j\in I}(y_i\alpha_i+\beta_i-\gamma_i)^\top(y_j\alpha_j+\beta_j-\gamma_j)\langle z_i,z_j\rangle-\sum_{i\in I}(0.5\times\alpha_i(y_i+\varepsilon)-\gamma_i) \\ s.t.\sum_{i\in I}(y_i\alpha_i+\beta_i-\gamma_i)=0,0\leq\alpha_i\leq\frac{C}{\varepsilon},\beta_i\geq 0,\gamma_i\geq 0,i\in I \end{gathered} \tag{4}$$

```
与 SVM 相比，PSVM 具有以下特点：
i) 对于训练样本 $z_i，i∈I$，$⟨w,z_i⟩+b$ 的值在 $[0,1]$ 内，且 $⟨w,z_i⟩+b$ 是训练样本的条件概率估计。也就是说，PSVM 可以直接输出概率。
ii) 由于 $⟨w,z_i⟩+b$ 是有界的，其核矩阵也是有界的，即 $｜｜K｜｜_F≤δ$，其中$//·//F$ 是 Frobenius 范数，$\delta$是一个正有界数。

- 给定一个数据集 ${z_1, z_2, ..., z_n}$ 和一个核函数 $k(·,·)$，**核矩阵 $K$** 是一个 $n×n$ 的对称矩阵，其中每个元素 $K_{ij}$ 定义为：$K_{i,j}=k(z_i,z_j)=⟨\varphi(z_i),\varphi(z_j)⟩$。 $\varphi(·)$ 是将原始数据映射到高维特征空间的函数。

- **Frobenius范数**是矩阵的范数，可以看作是**向量欧几里得范数在矩阵上的推广**. 对于一个 $m×n$ 的矩阵 $A$，其Frobenius范数定义为：$｜｜A｜｜_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n ｜a_{ij}｜^2}$
```

- 经典TWSVM
  - 原问题：$$\begin{gather}min{w_1,b_1,\xi} \frac{1}{2}(||w_1||^2 + b_1^2) + \frac{C_1}{2} \sum{i∈I^1} (⟨w1, z_i⟩ + b_1)^2 + C_2 \sum{i∈I^2} ξ_i \

s.t. - (⟨w1, z_i⟩ + b_1) ≥ 1 - \xi_i, \xi_i ≥ 0, i ∈ I^2 \end{gather} \ \ (5)

\begin{gather}min{w2,b_2,\eta} \frac{1}{2}(||w_2||^2 + b_2^2) + \frac{C_3}{2} \sum{i∈I^2} (⟨w2, z_i⟩ + b_2)^2 + C_4 \sum{i∈I^1} \eta_i \

s.t. - (⟨w2, z_i⟩ + b_2) ≥ 1 - \eta_i, \eta_i ≥ 0, i ∈ I^1 \end{gather} \ \ (6)

- 第二个目标函数使 $+1$ 类邻近于超平面 $⟨w_1,z⟩+b_1=0$ - 第三个目标函数则是最小化约束的损失 - 第一个目标函数是正则项 - 约束

\begin{gather}Class \ y = argmin{k=1,2} \frac{ | ⟨w_k,\varphi(x^*)⟩ + b_k |}{||w_k||} \end{gather} \ \ (9)$$

```
- 令 $\tilde{z} =[z;1]$, $\mathbb{I}_{i,j}=1$如果 $i=j$, 否则 $\mathbb{I}_{i,j}= 0$。则其对偶形式为:
$$\begin{gather}min_{α} \frac{1}{2} \sum_{i,j∈I^2} \tilde{z}_i (C_1 \sum_{i,j∈I^1}( ⟨\tilde{z}_i, \tilde{z}_j⟩
+ \mathbb{I}_{i,j}))^{-1}\tilde{z}_j α_i α_j - \sum_{i∈I^2} α_i \\ s.t. 0 ≤ α_i ≤ C_2, i ∈ I^2, \end{gather} \ \
(7) $$ $$\begin{gather} min_β \frac{1}{2} \sum_{i,j∈I^1} \tilde{z}_i ( C_3 \sum_{i,j∈I^2}( ⟨\tilde{z}_i,
\tilde{z}_j⟩ + \mathbb{I}_{i,j}))^{-1} \tilde{z}_j β_i β_j - \sum_{i∈I^1} β_i \\ s.t. 0 ≤ β_i ≤ C_4, i ∈ I^1.
\end{gather} \ \ (8) $$

与SVM相比，TWSVM具有以下特点:
i) 对于每个类别，它分别构造一个超平面 $f_k = \langle w_k, z \rangle + b_k = 0$。TWSVM中每个优化子问题的规模小于SVM的优化
问题规模。同时，我们期望TWSVM中的子问题比SVM中的整体问题更简单。
ii) $f_k$ 是通过衡量类内相似性和类间相异性获得的，新样本根据其最接近的$f_k$所属的类别进行划分。也就是说，每个模型以及最终决
策中所用的相似性/相异性度量标准可能不一致，因此难以直接衡量训练样本的误分类损失。
```

- 二分类CPTWSVM
  - 原问题：

$$\begin{gathered}min_{w_1,b_1,\xi}\frac{1}{2}(||w_1||^2+b_1^2)+C_1\sum_{i\in I^1}\xi_i-C_2\sum_{i\in I}y_i(\langle w_1,z_i\rangle+b_1)\\ s.t.\,\langle w_1,z_i\rangle+b_1\geq 0.5-\xi_i,\xi_i\geq 0,i\in I^1\\ 0\leq\langle w_1,z_i\rangle+b_1\leq 1,i\in I\end{gathered}\qquad(10)$$

$$\begin{gathered}min_{w_2,b_2,\eta}\frac{1}{2}(||w_2||^2+b_2^2)+C_3\sum_{i\in I^2}\eta_i+C_4\sum_{i\in I}y_i(\langle w_2,z_i\rangle+b_2)\\ s.t.\,\langle w_2,z_i\rangle+b_2\geq 0.5-\eta_i,\eta_i\geq 0,i\in I^2\\ 0\leq\langle w_2,z_i\rangle+b_2\leq 1,i\in I\end{gathered}\qquad(11)$$

    - 目标函数的第一项是正则化项
    - 第二项最小化被错误分类的正样本数量（同样也涉及第一个约束）
    - 第三项共同最大化正类的条件概率并最小化负类的条件概率，也称为判别项
    - 最后一个约束保证了$\langle w_1,z\rangle+b_1$和$\langle w_2,z\rangle+b_2$属于$0\sim 1$之间
    - 分界面1的支持向量：满足$\{0<\alpha_i\leq C_1,\beta_j>0,\gamma_j>0|i\in I^1,j\in I\}$ 的训练样本i
    - 一个新样本 $x^*\in R^n$ 的分类取决于哪个类别的概率更大

$$Class\ y=argmax_{k=1,2}\langle w_k,\varphi(x^*)\rangle+b_k\quad(27)$$

  - 对偶问题：

$$\begin{gathered}min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i,i\prime\in I^1,j,j\prime\in I}\langle\alpha_i\tilde{z}_i+\beta_j\tilde{z}_j-\gamma_j\tilde{z}_j,\alpha_{i\prime}\tilde{z}_{i\prime}+\beta_{j\prime}\tilde{z}_{j\prime}-\gamma_{j\prime}\tilde{z}_{j\prime}\rangle\\ -\frac{1}{2}\sum_{i\in I^1}\alpha_i+C_2\sum_{i\in I^1,j\in I}y_j\alpha_i\langle\tilde{z}_j,\tilde{z}_i\rangle+C_2\sum_{j,j\prime\in I}y_j(\beta_j-\gamma_j)\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle+\sum_{j\in I}\gamma_j\\ s.t.\,0\leq\alpha_i\leq C_1,i\in I^1,\beta_j\geq 0,\gamma_j\geq 0,j\in I.\end{gathered}\qquad(21)$$

$$\begin{gathered}min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i,i\prime\in I^2,j,j\prime\in I}\langle\alpha_i\tilde{z}_i+\beta_j\tilde{z}_j-\gamma_j\tilde{z}_j,\alpha_{i\prime}\tilde{z}_{i\prime}+\beta_{j\prime}\tilde{z}_{j\prime}-\gamma_{j\prime}\tilde{z}_{j\prime}\rangle\\ -\frac{1}{2}\sum_{i\in I^2}\alpha_i+C_4\sum_{i\in I^2,j\in I}y_j\alpha_i\langle\tilde{z}_j,\tilde{z}_i\rangle+C_4\sum_{j,j\prime\in I}y_j(\beta_j-\gamma_j)\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle+\sum_{j\in I}\gamma_j\\ s.t.\,0\leq\alpha_i\leq C_3,i\in I^2,\beta_j\geq 0,\gamma_j\geq 0,j\in I.\end{gathered}\qquad(24)$$

  且根据上述参数可得：

$$\begin{gathered}w_1=\sum_{i\in I^1}\alpha_i z_i+\sum_{i\in I}(\beta_i-\gamma_i+C_2y_i)z_i\quad(22)\\ b_1=\sum_{i\in I^1}\alpha_i+\sum_{i\in I}(\beta_i-\gamma_i+C_2y_i).\,(23)\end{gathered}$$

$$\begin{gathered} w_2=\sum_{i\in I^2}\alpha_i z_i+\sum_{i\in I}(\beta_i-\gamma_i+C_4 y_i)z_i \;\; (25) \\ b_2=\sum_{i\in I^2}\alpha_i+\sum_{i\in I}(\beta_i-\gamma_i+C_4 y_i). \, (26) \end{gathered}$$

- CPTWSVM有一种称作块算法的求解方法，如下所示：
  Algorithm1: Block iterative algorithm for solving (21).
  Input: Training set $\{(x_i,y_i)|i\in I\}, I=I_1\cup I_2\cup\ldots,$ parameters C_1 and C_2
  Initialize: $\beta_j^{(0)}=0$ and $\gamma_j^{(0)}=0$ for $j\in I$; $\alpha_j^{(0)}=0$ for $j\in I^k$
  For: $t=1,\ldots,T_{max}$ do
  (1) Update $\{\alpha_i^{(t)}|i\in I^k\}$ by solving (28)
  (2) Update $\{\beta_j^{(t)}|j\in I\}$ by solving (29)
  (3) Update $\{\gamma_j^{(t)}|j\in I\}$ by solving (30)
  (4) Until stop condition (31);
  End;
  Output: $\{\alpha_i=\alpha_i^{(t)},\beta_j=\beta_j^{(t)},\gamma_j=\gamma_j^{(t)}|i\in I^k,j\in I\}$
  其中（28）（29）（30）如下所示：

$$\begin{gathered} \alpha^{k(t)}=argmin_{0\le\alpha_i^k\le C_1}\frac{1}{2}\sum_{i,i\prime\in I^k}\alpha_i^k\langle\tilde{z}_i,\tilde{z}_{i\prime}\rangle\alpha_{i\prime}^k+ \\ \sum_{i\in I^k,j\in I}\alpha_i^k\langle\tilde{z}_i,\tilde{z}_j\rangle(\beta_j^{k(t-1)}-\gamma_j^{k(t-1)})+C_2\sum_{i\in I^k,j\in I}y_j\alpha_i^k\langle\tilde{z}_j,\tilde{z}_i\rangle-\frac{1}{2}\sum_{i\in I^k}\alpha_i^k \end{gathered} \quad (28)$$

$$\begin{gathered} \beta^{k(t)}=argmin_{\beta_j^k\ge 0}\frac{1}{2}\sum_{j,j\prime\in I}\beta_j^k\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle\beta_{j\prime}^k+ \\ \sum_{i\in I^k,j\in I}\beta_j^k\langle\tilde{z}_i,\tilde{z}_j\rangle(\alpha_i^{k(t)}-\gamma_j^{k(t-1)})+C_2\sum_{j,j\prime\in I}y_j\beta_j^k\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle \end{gathered} \quad (29)$$

$$\begin{gathered} \gamma^{k(t)}=argmin_{\gamma_j^k\ge 0}\frac{1}{2}\sum_{j,j\prime\in I}\gamma_j^k\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle\gamma_{j\prime}^k- \\ \sum_{i\in I^k,j\in I}\gamma_j^k\langle\tilde{z}_i,\tilde{z}_j\rangle(\alpha_i^{k(t)}+\beta_j^{k(t)})-C_2\sum_{j,j\prime\in I}y_j\gamma_j^k\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle+\sum_{j\in I}\gamma_j^k \end{gathered} \quad (30)$$

$$\frac{\sqrt{\sum_{k=1}^K\left[\sum_{i\in I^k}(\alpha_i^{k(t)}-\alpha_i^{k(t-1)})^2+\sum_{j\in I}(\beta_j^{k(t)}-\beta_j^{k(t-1)})^2+\sum_{j\in I}(\gamma_j^{k(t)}-\gamma_j^{k(t-1)})^2\right]}}{\sqrt{\sum_{k=1}^K\left[\sum_{i\in I^k}(\alpha_i^{k(t-1)})^2+\sum_{j\in I}(\beta_j^{k(t-1)})^2+\sum_{j\in I}(\gamma_j^{k(t-1)})^2\right]}}<tol \quad (31)$$

- 多分类CPTWSVM
  - 原问题：

$$\begin{gathered} min_{w_k,b_1,\xi^{(k)}}\frac{1}{2}(||w_k||^2+b_k^2)+C_1\sum_{i\in I^k}\xi_i^{(k)}-C_2\sum_{i\in I}y_i(\langle w_k,z_i\rangle+b_k) \\ s.t.\langle w_k,z_i\rangle+b_k\ge 0.5-\xi_i^{(k)},\xi_i^{(k)}\ge 0,i\in I^k \\ 0\le\langle w_k,z_i\rangle+b_k\le 1,i\in I \end{gathered} \quad (32)$$

    - 无论是每个项的含义，支持向量，新样本分类都与二分类没有差异
    - 只有一点，在多分类中，$y_i\in\{+1,-1\}$ 是k类时为1，不是k类时为-1
  - 对偶问题：

$$\begin{gathered} min_{\alpha,\beta,\gamma}\frac{1}{2}\sum_{i\in I^k,j\in I}\langle\alpha_i\tilde{z}_i+\beta_j\tilde{z}_j-\gamma_j\tilde{z}_j,\alpha_i\tilde{z}_i+\beta_j\tilde{z}_j-\gamma_j\tilde{z}_j\rangle-\frac{1}{2}\sum_{i\in I^k}\alpha_i+ \\ C_2\sum_{i\in I^k,j\in I}^{m_k}y_j\alpha_i\langle\tilde{z}_j,\tilde{z}_i\rangle+C_2\sum_{j,j'\in I}y_j(\beta_j-\gamma_j)\langle\tilde{z}_j,\tilde{z}_{j\prime}\rangle+\sum_{j\in I}\gamma_j \\ s.t.0\le\alpha_i\le C_1,i\in I^k,\beta_j\ge 0,\gamma_j\ge 0,j\in I \end{gathered} \quad (33)$$

  与TWSVM相比CPTWSVM有以下特点：
  i) 对于每个类别，它构建一个 $f_k(z)$ 来估计 $p(y=k\mid x)$，并且所有 $f_k(z)$ 是分开构建的，而 TWSVM 构建超平面。对于每个子问题，CPTWSVM 度量“样本是否属于k类的错误分类损失”，而 TWSVM 无法度量错误分类损失。
  ii) 在 CPTWSVM 中，所有关于 z_i 的计算都基于内积，因此该模型可以直接扩展到 RKHS 中的核方法。而TWSVM需要表示理论来扩展到核方法，并且TWSVM 的对偶问题需要计算数据矩阵的逆。
  iii) 对于每个类别，$f_k(z)\approx p(y=k\mid x)$ 是通过度量概率来估计的，新样本被分配到具有最大概率 $f_k(z)$ 的类别，每个类别的损失度量和决策是一致的。而在 TWSVM 中，类内和类间到超平面的距离度量方式不同。

iv) CPTWSVM 的输出是有界的，同时其核矩阵也是有界的，这更符合统计学习理论。而 TWSVM 的输出是无界的。由于有界约束，CPTWSVM 的优化问题规模比 TWSVM 大，但 TWSVM 需要求解二次项矩阵的逆，其计算复杂度非常高。

与 SVM 相比CPTWSVM有以下特点：
i) 类似于 PSVM 相比 SVM 引入了 $p(y=1 \mid x)$，CPTWSVM 相比 TWSVM 引入了 $p(y=k \mid x)$。CPTWSVM 保持了支持向量的特性，并且在 SVM 的对偶形式中只计算样本的内积。然而，它是分别估计 $p(y=k \mid x)$ 的。
ii) 与 PSVM 相比，CPTWSVM 引入了 $max \sum_{i \in I} y_i(\langle w, z_i \rangle + b)$，它最大化类内相似性并最小化类间相异性。实际上，可以很容易地将其扩展为 PSVM。

- CPSVM：
  - 原问题：

$$\begin{gathered} min_{w,b,\xi} \frac{1}{2}||w||^2 + \frac{C_1}{\varepsilon}\sum_{i \in I}\xi_i - C_2 \sum_{i \in I} y_i(\langle w, z_i \rangle + b) \\ s.t. \langle w, z_i \rangle + b \geq 0.5\varepsilon - \xi_i,\ \xi_i \geq 0, i \in I^1 \\ \langle w, zi \rangle + b \leq 0.5\varepsilon + \xi_i,\ \xi_i \geq 0, i \in I^2 \\ 0 \leq \langle w, z_i \rangle + b \leq 1, i \in I \end{gathered} \tag{37}$$

  - 对偶问题：

$$\begin{gathered} min_{\alpha,\beta,\gamma} \frac{1}{2}\sum_{i,j \in I}(\alpha_i y_i + \beta_i - \gamma_i)^\top(\alpha_j y_j + \beta_j - \gamma_j)\langle z_i, z_j \rangle \\ +C_2 \sum_{i,j \in I}(y_i y_j \alpha_j + y_i \beta_j - y_i \gamma_j)\langle z_i, z_j \rangle + \sum_{i \in I}(\gamma_i - 0.5 \times \alpha_i(y_i + \varepsilon)) \\ s.t. \sum_{i \in I}(\alpha_i y_i + \beta_i - \gamma_i) = -C_2 \sum_{i \in I} y_i, i \in I, \\ 0 \leq \alpha_i \leq \frac{C_1}{\varepsilon}, \beta_i \geq 0, \gamma_i \geq 0, i \in I. \end{gathered} \tag{38}$$

  CPSVM 的对偶问题规模与 PSVM 相同，而 CPTWSVM 的每个子问题规模小于 CPSVM。对于多分类问题，使用一对多策略的多分类 CPSVM 与多分类 CPTWSVM 类似，但变量更多。

实验设计的思路（取自 4）

- 参与对比的模型：SVM | PSVM | TWSVM | LSTSVM | CPSVM | CPTWSVM
- 参数配置（Parameter setting）：
  1. The trade-off parameters$C, C_1, C_2, C_3, C_4 : \{2^{-8}, \ldots, 2^8\}$
  2. For PSVM and CPSVM, ε is tuned from $\{2^{-8}, \ldots, 2^0\}$
  3. set $C_1 = C_3, C_2 = C_4$ in TWSVM and CPTWSVM
  4. rbf kernel $K(x, x\prime) = exp(-\mu||x - x\prime||^2)$, μ is in $\{2^{-6}, \ldots, 2^6\}$ | *(linear表示不用kernel)*
- 评估度量（Evaluation metrics ）：
  1. Acc: $Acc = 100\% \times \frac{(TP+TN)}{(TP+FP+TN+FN)}$
  2. SVs: 支持向量的数量
  3. TPR & FPR: $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$, 用于绘制ROC和AUC
  4. ROC 和 AUC: TPR和FPR分别为横轴和纵轴
  5. Friedman test with Nemenyi posthoc test: 非线性统计分析(容我详述)
  6. The contrast estimation based on medians: 基于中位数的对比估计(容我详述)

带有Nemenyi事后检验的Friedman检验: Friedman检验是一种非参数的方差分析，它不要求数据服从正态分布，非常适合用于比较多个算法在多个数据集上的性能。
Step1. 假设有 k 个算法，有 N 个数据集，在每个数据集上运行所有算法，并得到一个性能指标。
Step2. 对于每个算法 j，计算它在所有 N 个数据集上的排名的平均值 R_j。
Step3. 计算**Friedman统计量**，$\chi_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}]$
Step4. 将计算出的 $\chi^2{}_F$ 与卡方分布在显著性水平 α 下的临界值 $\chi^2{}_\alpha(k-1)$ 进行比较 (α由我们定)
Step5. 如果 $\chi^2{}_F > \chi^2{}_\alpha(k-1)$，则拒绝**零假设 (H₀)**(所有算法的性能都是相同的，没有显著差异)
Step6. 计算临界差异: $CD = q_\alpha * \sqrt{\frac{k(k+1)}{6N}}$, 其中$q_\alpha$需要查表获得
Step7. 如果 $|R_i - R_j| > CD$，则认为算法 i 和 j 的性能存在显著差异
本论文设显著性水平$\alpha = 10\%$，进而$\chi_F^2 = 27.71$, $\chi_\alpha^2(k-1) = 9.24$，$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} = 1.37$

基于中位数的对比估计: 使用“中位数”而不是“均值”来衡量*典型的性能差异*，因为中位数对异常值不敏感，更加稳健。
Step1. 假设有 k 个算法，有 N 个数据集，在每个数据集上运行所有算法，并得到一个性能指标。

Step2. 对任意算法i,j，计算二者在N个数据集上的差值：差值 = 算法i指标 - 算法j指标
Step3. 对任意算法i,j，我们得到一个差值列表，取出其中的中位数$mid_{i,j}$
Step4. mid_i,j就表示模型i,j的性能差异, 进而可以绘制相应二维表格, 横轴/纵轴均为"算法"

- 数据集选取：Table3中有25个数据集，其中Aus，Echo，Horse和WDBC有两个维度的特征；Iris和Seeds有三个维度的特征。其中样本数、特征数和类别数分别记为 m、n 和 k。
- 评估方法：
  - 选参：Standard 10 fold cross validation according to the Acc（10折交叉）
  - 算Acc：The 10 times average Acc under the optimal parameters（最优参数时的均值）
- 可视化部分：
  - Table4/Table5：平均准确率 (标准差) | SV数 | 训练时间
    结论：1. 线性 CPTWSVM 的 Acc 略优于线性 SVM、PSVM、CPSVM、TWSVM 和 LSTSVM，且CPTWSVM 在 25 个数据集中的 22 个上明显比 LSTSVM 获得更高的 Acc；2. CPSVM 的 Acc 总体上略高于 PSVM，表明在 PSVM 的基础上添加判别项 $\sum_{i\in I} y_i(\langle w, z_i\rangle + b)$ 可以有效提高分类性能。3. CPTWSVM 的 Acc 略高于 CPSVM，这表明 TWSVM 范式在分类中起着重要作用。4. LSTSVM 总是拥有最多的 SVs，这是因为 LSTSVM 使用最小二乘损失函数，该函数不具有稀疏性。5. TWSVM 的 SVs 比 CPTWSVM 多，这是因为 TWSVM 也有一个最小二乘损失函数，但 CPTWSVM 只有 L_1 稀疏损失。6. 在大多数情况下，CPTWSVM 的训练时间低于 CPSVM 但高于 TWSVM，这是因为 CPTWSVM 中的 QPP 规模小于 CPSVM 但高于 TWSVM。7. 非线性 CPTWSVM 的 Acc 略优于其他非线性分类器。8. 类似地，非线性CPTWSVM 在 25 个数据集中的 21 个上获得了比非线性 LSTSVM 更高的 Acc，这进一步证实了上述结论。
  - Fig1：非线性统计分析(计算方法如前所述)
    结论：CPTWSVM 平均排名第一，并且显著优于其他分类器。（那些用线连接起来的分类器被认为它们之间差异不大。）
  - Table6/Table7：基于中位数的对比估计(计算方法如前所述)
    结论：CPTWSVM 相对于其他基准方法总是获得正的差异值，这验证了 CPTWSVM 相对于其他分类器的统计显著性。
  - Fig2：二维特征可视化分析(绘制方法见下方)
    结论：1. 对于这些二分类数据集，CPTWSVM 在一致性度量下输出投影，因为 CPTWSVM 输出两个类别的概率值，可以轻松获得累积分布曲线。而在 TWSVM 和 LSTSVM 中，两个类别的投影距离不同，且决策边界不对称。2. 与 TWSVM 和 LSTSVM 相比，CPTWSVM 的分隔线位于 0-1 盒的对角线上，并且输出是概率值，不受度量一致性的影响，而 TWSVM 和 LSTSVM 更容易受到一致性的影响。3. 两个类别在 CPTWSVM 中清晰分离，而在 TWSVM 和 LSTSVM 中，两个类别的投影区分度较低。
  - Fig3：三维特征可视化分析(没啥计算方法)
    结论：从Fig3可以看出，与上述类似，CPTWSVM 获得了比 TWSVM 和 LSTSVM 更好的分离效果。
  - Fig4：六个数据集的ROC曲线与AUC值分析(计算方法见下)
    结论：从Fig4可以看出，无论是二分类还是多分类，CPTWSVM（红线）的 ROC 曲线通常位于 TWSVM（蓝线）和 LSTSVM（绿线）之上，并且 CPTWSVM 的 AUC 值显著高于 TWSVM 和 LSTSVM，因此上述结论也得到了证实。
  - Fig5：CPSVM 与 CPTWSVM 中 C_2/C_4 取值对模型的影响分析(绘制方法见下)
    结论：从Fig5 可以看出，判别项对 CPSVM 和 CPTWSVM 都非常有意义。具体来说，当 $C_2/C_4 = 0$ 时，CPSVM 和 CPTWSVM 都无法获得最优结果，表明添加线性判别项是有用的。此外，对于六个基准数据集，CPTWSVM 在 $C_2/C_4 = 0$ 时通常具有较低的 Acc，这进一步表明判别项对 CPTWSVM 有显著影响。
  - Fig6：6种模型在真实数据集上的ROC曲线与AUC值分析(对每个模型绘制ROC和AUC即可)
    结论：如Fig6 所示，在 Heart、Cradio 和 Year1 数据集中，CPTWSVM 的 ROC 曲线高于其他分类器，这进一步表明将 TWSVM 范式与判别项和概率输出相结合不仅获得了更好的分类性能，而且具有更好的可解释性。
  - Fig7：6种模型在真实数据集上准确率和训练时间的直方图(时间取为 $log(1+t)$以保证正值)
    结论：如图Fig7 所示，在 4 个数据集中的 2 个上，CPTWSVM 的训练时间高于 TWSVM，这是因为概率输出的约束增加了求解规模，但准确率显著提高。

Fig2的绘制：

1. TWSVM 和 LSTSVM 的图是通过绘制坐标为（到超平面 1 的距离，到超平面 -1 的距离）的点获得的，即对于一个测试样本 x_i，到超平面 1 和 -1 的距离分别是 $\frac{|\langle w_1,\varphi(x_i)\rangle+b_1|}{\|w_1\|}$ 和 $\frac{|\langle w_2,\varphi(x_i)\rangle+b_2|}{\|w_2\|}$。
2. CPTWSVM 的图是通过绘制测试样本的超平面 1 和超平面 -1 的输出概率获得的，即超平面 1 和 -1 的概率分别是 $p(y=1 \mid x_i) = \langle w_1, \varphi(x_i)\rangle + b_1$和 $p(y=-1 \mid x_i) = \langle w_2, \varphi(x_i)\rangle + b_2$。
3. 除此之外，CPTWSVM 中的红色虚线和蓝色虚线分别代表正样本和负样本的**累积分布曲线**（**累计分布曲线**是**累计分布函数**的图形化表示。其含义是：对于一个随机变量，有多少比例的数据点"落在"某个特定值之下。如果两条曲线**分离得很开**，几乎没有重叠，说明模型能非常清晰地区分两类样本。）

Fig4的绘制：
4. 我们计算每个类别的 TPR 和 FPR
5. 然后对所有类别的 TPR 和 FPR 取平均以获得宏平均 ROC 曲线和 AUC。

Fig5的绘制：

6. 横坐标是 C_2 的范围，纵坐标是分类器的准确率（%），红线和蓝线分别代表 CPTWSVM 和 CPSVM。

7. 固定其他参数最优的情况下令 C_2/C_4 从 $\{0, 2^{-8}, \cdots, 2^8\}$ 中选取