Statistica Sinica Preprint No: SS-2020-0287	
Title	An iterative algorithm to learn from positive and unlabeled examples
Manuscript ID	SS-2020-0287
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0287
Complete List of Authors	Xin Liu, Qingle Zheng, Xiaotong Shen and Shaoli Wang
Corresponding Author	Shaoli Wang
E-mail	swang@shufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

Statistica Sinica

An iterative algorithm to learn from positive and unlabeled examples

Xin Liu[1], Qingle Zheng[1], Xiaotong Shen[2] and Shaoli Wang[1]

[1] *Shanghai University of Finance and Economics*

[2] *University of Minnesota*

Abstract: In semi-supervised learning, a training sample comprises both labeled and unlabeled instances from each class under consideration. In practice, an important, yet challenging issue is the detection of novel classes that may be absent from the training sample. Here, we focus on the binary situation in which labeled instances come from the positive class, and unlabeled instances come from both classes. In particular, we propose a semi-supervised large-margin classifier to learn the negative (novel) class based on pseudo-data generated iteratively using an estimated model. Numerically, we employ an efficient algorithm to implement the proposed method using the hinge loss and ψ-loss functions. Theoretically, we derive a learning theory for the new classifier in order to quantify the misclassification error. Finally, a numerical analysis demonstrates that the proposed method compares favorably with its competitors on simulated examples, and is highly competitive on benchmark examples.

Key words and phrases: Biased SVM, Iterative algorithm, Large-margins, PU learning.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

1 Introduction

In semi-supervised learning, a large amount of labeled and unlabeled data are observed together in order to enhance the predictive accuracy of a classifier (Vapnik, 1998; Chapelle and Zien, 2005; Wang and Shen, 2007; Wang, Shen, and Pan, 2009). For most existing methods, instances from all classes are required. Therefore, these methods cannot detect a novel class if it is absent from the training sample. This sort of problem arises in many applications, such as text classification (Liu et al., 2002; Denis, Gilleron, and Tommasi, 2002), where relevant documents are retrieved without labor-intensively labeling irrelevant documents, and disease gene prediction (Calvo et al., 2007), where disease genes are identified in the presence of positive instances, but not negative ones. In this study, we consider the situation in which labeled instances come from one (positive) class, and unlabeled instances come from both classes. By minimizing the generalization error, we construct a semi-supervised learner capable of detecting the novel class. In fact, any classification can be cast into the novel-class-detection framework with labeled instances from only one class and a large number of unlabeled instances from both classes.

We now briefly review the pertinent literature. In terms of text classification, variants of one-class support vector machines (SVMs) have been proposed to estimate the support of positive data without using unlabeled samples (Tax and Duin, 1999; Manevitz and Yousef, 2001; Schölkopf et al., 2001; Pierre Geurts, 2011). The naive Bayes approach

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

has been applied to the positive and unlabeled classification problem. Here, examples include the positive naive Bayes approach (Denis, Gilleron, and Tommasi, 2002) and the positive tree-augmented naive Bayes approach (Calvo, Larrañaga and Lozano, 2007). However, either they perform poorly when a large number of unlabeled instances are discarded (Liu et al., 2003), or the computation cost becomes high, with limited improvement. Two-step algorithms have also been developed to solve the problem. The first step extracts a fraction of the reliable negative instances from the unlabeled sample, and then the second step trains classifiers based on the positive and reliable negative instances. These two steps are repeated iteratively until no reliable negative instances can be identified in the unlabeled sample. Examples of such algorithms include spy-EM (Liu et al., 2002), positive example-based learning (Yu, Han, and Chang, 2002), and the SVM with a Rocchio extraction (Li and Liu, 2003). Note that a scheme maximizing the number of negative classified instances among unlabeled samples, while classifying positive samples correctly, leads to good overall performance (Liu et al., 2002). Moreover, by adjusting the misclassification costs of the two classes due to asymmetry, weighted methods are obtained. Here, examples include the weighted logistic regression (Lee and Liu, 2003), biased SVM (BSVM) (Liu et al., 2003), and re-weighting method (Elkan and Noto, 2008). Liu et al. (2003) demonstrate experimentally that the BSVM outperforms various two-step algorithms. Recently, bagging tactics have been employed, yielding comparative performance (Mordelet and Vert, 2014). Global and local learning

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

from positive and unlabeled examples adapts the intrinsic geometric information in the training data set. A biased least square SVM (BLSSVM) has also been proposed (Ke et al., 2018). The learning theory on the risk estimator for positive and unlabeled instances is partially established and examined in, for example, Kiryo et al. (2017), Natarajan et al. (2018), and Tanielian and Vasile (2019).

To detect the negative (novel) class, we propose a semi-supervised large-margin classifier that combines the benefits of large margins and the BSVM method (Liu et al., 2003), and iteratively generates pseudo-samples for training. The proposed classifier incorporates the predicted values of unlabeled instances appropriately, and then iteratively trains a biased model based on the pseudo-training samples, with original labeled instances remaining unchanged at each iteration step. Additionally, the proposed method adjusts the weights adaptively to tackle the imbalance issue, if there is any, yielding a more accurate classification. This iterative scheme usually leads to an improvement at each iteration, thereby outperforming its counterparts without a weight adjustment. To implement the proposed large-margin classifier using the hinge loss and ψ-loss functions, we employ an inexact alternating direction method of multipliers (IADMM) algorithm (Wang et al., 2013), which decouples variables for efficient computation.

Our numerical analysis indicates that the newly proposed method compares favorably with the state-of-the-art BSVM and bagging SVM (BASVM) in terms of the generalization error (Mordelet and Vert, 2014). More importantly, the proposed method

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

achieves nearly the performance of the classifiers with complete data, indicating that the re-weighting scheme does lead to an overall improvement. Theoretically, we establish a novel learning theory for the ψ-loss, providing insight into the connection between the performance of the proposed method and the sample size, tuning parameter, and loss function in semi-supervised learning. In particular, the theory confirms the simulation results.

The rest of paper is organized as follows. Section 2 presents a general weighted large-margin classification model and the proposed method. Section 3 develops an algorithm based on the IADMM for implementation. Section 4 introduces a new tuning criterion with only positive labeled data and unlabeled data. In Section 5, the proposed method is compared against its strong competitors on two simulated examples and two benchmark examples. In Section 6, we investigate the theoretical properties of the proposed method. Section 7 discusses the proposed method and the underlying problem. All technical proofs are deferred to the appendix.

2 Methodology

2.1 Weighted Large-Margin Classification

Given a training sample $(\mathbf{x}_i, y_i)_{i=1}^n$ with $y_i \in \{1, -1\}$, for $1 \le i \le n$, the objective function of the weighted large-margin classification (Osuna, Freund, and Girosi, 1997)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

is

$$\min_{f\in\mathcal{F}} \quad C_+ \sum_{y_i=1} L(y_i f(\mathbf{x}_i)) + C_- \sum_{y_j=-1} L(y_j f(\mathbf{x}_j)) + J(f), \tag{2.1}$$

where $\mathcal{F}$ is the candidate set of decision functions, $L(\cdot)$ is the margin loss function of the functional margin $z = yf(\mathbf{x})$, $J(\cdot)$ is a regularization term that controls the complexity of the decision function f, and C_+ and C_- are nonnegative tuning parameters controlling the trade-off between the fits for the positive and negative classes, respectively, and the complexity of the decision function. A margin loss $L(z)$ is called a large margin if it is decreasing in the variable z; that is, a large margin loss penalizes small margins, pushing correctly specified instances away from the classification boundary. Given a decision function f, the corresponding classification rule is $\text{sign}(f(\mathbf{x}))$. For linear classification problems, $\mathcal{F} = \{f(\mathbf{x}) = b_0 + \mathbf{b}^T\mathbf{x} \equiv (1, \mathbf{x}^T)\bar{\mathbf{b}}\}$, where $\bar{\mathbf{b}} = (b_0, \mathbf{b}^T)^T$, and the commonly used regularizer is $J(f) = \|\mathbf{b}\|^2/2$, the reciprocal of the geometric margin. For nonlinear classification, $\mathcal{F} = \{f(\mathbf{x}) = b_0 + \sum_{i=1}^n b_i K(\mathbf{x}, \mathbf{x}_i)\}$ and $J(f) = \sum_{1\le i,j\le n} b_i K(\mathbf{x}_i, \mathbf{x}_j) b_j/2$, where $K(\cdot,\cdot)$ is a reproducing kernel, see Gu (2000) and Wahba (1990) for the reproducing kernel Hilbert spaces. Moreover, different large-margin loss functions lead to different learning machines. In this study, we consider a linear classification with the hinge loss $L(z) = (1-z)_+$ (Cortes and Vapnik, 1995) and the ψ-loss $\psi(z) = \min(1, (1-z)_+)$ (Shen et al., 2003). The hinge loss is the most commonly used loss function in classification problems, owing to its good performance and convexity. However, the hinge loss is not robust to outliers, because of unboundedness. Hence, a bounded loss function, ψ-loss, is

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

also used as an alternative. The numerical analysis in Section 5 shows that our proposed method with ψ-loss outperforms that with the hinge loss. Our proposed method can also adapt to other loss functions as well.

2.2 Proposed Method

In light of the preceding discussion, we propose the following cost function based on (2.1):

$$S(f, \mathbf{y}) = C\Big(\frac{1}{n_+}\sum_{y_i=1} L(y_i f(\mathbf{x}_i)) + \frac{1}{n_-}\sum_{y_j=-1} L(y_j f(\mathbf{x}_j))\Big) + J(f), \tag{2.2}$$

where n_+ and n_- are the numbers of instances of positive and negative classes, respectively, in the training sample. This weighting scheme assigns a large weight to the small class and a small weight to the large class, which mitigates the imbalance and misclassification. Note that the tuning parameter C can be rescaled to one by introducing another tuning parameter λ into $J(f)$, controlling the level of the penalty.

The motivation for our proposed approach comes from model (2.1). The BSVM (Liu et al. (2003)) fits (2.1) based on a pseudo-training sample consisting of the original positive instances and unlabeled observations treated as pseudo-negative instances. Obviously, such a scheme is biased owing to mislabeling of unlabeled data. However, some correctly labeled negative instances, together with the original positive instances, are useful for estimating the decision boundary using (2.2). In addition, incorrectly labeled positive instances have little impact on the decision boundary, given the missing-at-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

random assumption (Assumption A1 in Section 6). As a result, the classifier sign($\hat{f}^{(1)}$) based on (2.2) yields a better decision boundary than that of the classifier sign($\hat{f}^{(0)}$), which labels all unlabeled instances as negative. Furthermore, the subsequent refitting by the classifier sign($\hat{f}^{(2)}$) trained based on the original positives and the predicted labels of unlabeled data given by classifier sign($\hat{f}^{(1)}$) leads to a more accurate classification. This is confirmed by Theorem 3. This iterative train-and-refit procedure continues until a certain termination criterion is met when no further improvement is possible.

For the following analysis, we denote the observations $(\mathbf{x}_i, y_i)_{i=1}^{n_l}$ in the training set as the labeled data, where $y_i = 1$, for $1 \le i \le n_l$, and $(\mathbf{x}_j)_{j=n_l+1}^{n}$ as the unlabeled data. We summarize the iteration scheme below.

Algorithm 1

For $k = 0, 1, \ldots,$

Step 1 (Initialization): Train $\hat{f}^{(0)}$ using $\mathbf{x}_i$ and $y_i = I(1 \le i \le n_l) - I(n_l + 1 \le i \le n)$, for $i = 1, \ldots, n$. Specify a precision $\varepsilon > 0$, and set up the initial pseudo-training sample using the initial classifier sign($\hat{f}^{(0)}$): $y_j^0 = \text{sign}(\hat{f}^{(0)}(\mathbf{x}_j))$, for $n_l + 1 \le j \le n$, and $y_i^0 = y_i = 1$, for $1 \le i \le n_l$.

Step 2 (Iteration): Given the pseudo-sample $(\mathbf{x}_i, y_i^k)_{i=1}^{n}$, compute the classifier $\hat{f}^{(k+1)}$ by minimizing $S(f, \mathbf{y}^k)$, where $\mathbf{y}^k = (y_1^k, \cdots, y_n^k)^T$. Reclassify the data as $y_i^{k+1} = y_i$, for $1 \le i \le n_l$, and $y_j^{k+1} = \text{sign}(\hat{f}^{(k+1)}(\mathbf{x}_j))$, for $n_l + 1 \le j \le n$.

Step 3 (Termination): If $S(\hat{f}^{(k+1)}, \mathbf{y}^{k+1}) > S(\hat{f}^{(k+1)}, \mathbf{y}^k)$, terminate; otherwise, re-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

peat steps 2 and 3 until $|S(\hat{f}^{(k+1)}, \mathbf{y}^{k+1}) - S(\hat{f}^{(k)}, \mathbf{y}^k)| \leq \varepsilon |S(\hat{f}^{(k)}, \mathbf{y}^k)|$. The final classifier $\hat{f}_C$ is $\hat{f}^{(K)}$, where K is the number of iterations.

Note that in Algorithm 1, the minimization of $S(f, \mathbf{y})$ with the hinge loss in Step 2 appears to be a special case of the minimization problem with the ψ-loss introduced in Section 3. This iterative scheme bears the properties described in Theorems 1 and 2 below.

Theorem 1. *(Monotonicity) $S(\hat{f}^{(k)}, \mathbf{y}^k)$ is a decreasing function in k. Hence, the iterative algorithm converges as $k \to \infty$. That is, for any given precision $\varepsilon > 0$, the algorithm terminates in a finite number of steps.*

Theorem 2. *Suppose that $P(\sum_{Y_i^k=1} \mathbf{X}_i / n_+^k \neq \sum_{Y_j^k=-1} \mathbf{X}_j / n_-^k) > 0$; for the ψ-loss function, suppose further that an additional condition $P(\sum_{Y_i^k=1} \mathbf{X}_i / n_+^k \neq 0, \sum_{Y_j^k=-1} \mathbf{X}_j / n_-^k \neq 0) > 0$ holds. Then, $P(\hat{\mathbf{b}}^{k+1} \neq 0) > 0$, for any constant $C > 0$.*

Theorem 2 claims that as long as the covariates' sample mean vector of the positive class is not equal to that of the negative class, and both are away from the zero vector in the kth iteration, the coefficient vector is estimated as nonzero with a positive probability in the $(k+1)$th iteration, such that the decision function $f(\mathbf{x}) = b_0 + \mathbf{b}^T\mathbf{x}$ can be identified. Furthermore, the negative class that is absent from the training data set is recovered with a positive probability.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

3 Nonconvex Minimization, Difference Convex Programming, and the IADMM

Often, when the hinge loss is used with $J(f) = \|\mathbf{b}\|^2/2$, the objective function (2.2) is convex. However, when the hinge loss is replaced by the ψ-loss, the objective function becomes nonconvex. In what follows, we develop an efficient algorithm for the nonconvex minimization. The objective function (2.2) with the ψ-loss becomes

$$\min_{\bar{\mathbf{b}}} \quad \frac{1}{2}\|\mathbf{b}\|^2 + \sum_{i=1}^{n} C_{y_i}\psi(y_i f(\mathbf{x}_i)), \tag{3.1}$$

where $\bar{\mathbf{x}}_i = (1, \mathbf{x}_i^T)^T, \bar{\mathbf{b}} = (b_0, \mathbf{b}^T)^T$, $f(\mathbf{x}_i) = \bar{\mathbf{x}}_i^T\bar{\mathbf{b}}$, and $\psi(z) = \min((1-z)_+, 1)$.

To solve the above minimization, we employ a difference convex algorithm (An and Tao, 1997) and the IADMM (Wang et al., 2013). First, we decompose the loss function $\psi = \psi_1 + \psi_2$, where $\psi_1(z) = (1-z)_+$, which is the hinge loss, and $\psi_2(z) = z\mathbf{1}(z < 0)$, and replace ψ_2 with its majorization. Specifically, given the m-step solution $\bar{\mathbf{b}}^m$, we substitute $\langle \nabla\psi_2(\bar{\mathbf{b}}^m), \bar{\mathbf{b}}\rangle$ for $\psi_2(\bar{\mathbf{b}})$ after ignoring the constant term. Next, in the $(m+1)$-step, we solve the following sub-problem:

$$\min_{\bar{\mathbf{b}}} \quad \frac{1}{2}\|\mathbf{b}\|^2 + \sum_{i=1}^{n} C_{y_i}\Big((1 - y_i f(\mathbf{x}_i))_+ + y_i f(\mathbf{x}_i)\mathbf{1}(y_i f^m(\mathbf{x}_i) < 0)\Big), \tag{3.2}$$

where $\mathbf{1}(\cdot)$ is the indicator function. After introducing the slack variables ξ_i and η_i, (3.2)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

becomes

$$\min_{\bar{\mathbf{b}},\boldsymbol{\xi},\boldsymbol{\eta}} \quad \frac{1}{2}\|\mathbf{b}\|^2+\sum_{i=1}^n C_{y_i}\Big(\xi_i+y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}\mathbf{1}(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^m<0)\Big), \quad \text{subject to}$$
$$1-y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}=\xi_i-\eta_i, \quad \xi_i\geq 0,\eta_i\geq 0,\ i=1,\ldots,n. \tag{3.3}$$

The corresponding augmented Lagrangian of (3.3) $L(\bar{\mathbf{b}},\boldsymbol{\xi},\boldsymbol{\eta},\mathbf{u})$ is

$$\frac{1}{2}\|\mathbf{b}\|^2+\sum_{i=1}^n C_{y_i}\Big(\xi_i+y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}\mathbf{1}(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^m<0)\Big)+\rho\sum_{i=1}^n(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}-1+\xi_i-\eta_i+u_i)^2,$$

where $\mathbf{u}=(u_i)_{i=1}^n$ denotes the vectorized Lagrangian multipliers. Given $\bar{\mathbf{b}}^t,\boldsymbol{\xi}^t,\boldsymbol{\eta}^t$, and $\mathbf{u}^t$, we solve the following sub-problems iteratively using the alternating direction method of multipliers (ADMM, Boyd et al. (2011)):

$$\bar{\mathbf{b}}^{t+1}=\operatorname*{argmin}_{\bar{\mathbf{b}}}\frac{1}{2}\|\mathbf{b}\|^2+\sum_{i=1}^n C_{y_i}y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}\mathbf{1}(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^m<0)$$
$$+\frac{\rho}{2}\sum_{i=1}^n(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}-1+\xi_i^t-\eta_i^t+u_i^t)^2, \tag{3.4}$$

$$(\xi_i^{t+1},\eta_i^{t+1})=\operatorname*{argmin}_{\xi_i\geq 0,\eta_i\geq 0}\sum_{i=1}^n C_{y_i}\xi_i+\frac{\rho}{2}\sum_{i=1}^n(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^{t+1}-1+\xi_i-\eta_i+u_i^t)^2, \tag{3.5}$$

$$u_i^{t+1}=u_i^t+y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^{t+1}-1+\xi_i^{t+1}-\eta_i^{t+1}. \tag{3.6}$$

The whole iteration procedure completes using a certain termination rule, specified below. Specifically, to solve (3.4), we employ the IADMM, which updates (3.4) by linearizing its last two terms and adding a proximal term $\|\bar{\mathbf{b}}-\bar{\mathbf{b}}^t\|_2^2$. This yields

$$\bar{\mathbf{b}}^{t+1}=\operatorname*{argmin}_{\bar{\mathbf{b}}}\frac{1}{2}\|\mathbf{b}\|^2+\frac{\zeta}{2}\|\bar{\mathbf{b}}-\bar{\mathbf{b}}^t\|^2+\rho\bar{\mathbf{b}}^T\bar{\mathbf{v}}^t, \tag{3.7}$$

where $\zeta>0$ is a prespecified constant, and $\bar{\mathbf{v}}^t=(v_0,\mathbf{v}^T)^T=\sum_{i=1}^n(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}-1+\xi_i-\eta_i+$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

$u_i - C_{y_i}\mathbf{1}(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^m < 0)/\rho)y_i\bar{\mathbf{x}}_i$. The analytic solution of (3.7) is

$$b_0^{t+1} = b_0^t - \frac{\rho}{\zeta}v_0^t, \quad \mathbf{b}^{t+1} = \frac{\zeta\mathbf{b}^t - \rho\mathbf{v}^t}{1+\zeta}. \tag{3.8}$$

Similarly, problem (3.5) has the following closed-form solution:

$$\xi_i^{t+1} = \max(-y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^{t+1} + 1 - u_i^t - \frac{C_{y_i}}{\rho}, 0), \quad \eta_i^{t+1} = \max(y_i\bar{\mathbf{x}}_i^T\bar{\mathbf{b}}^{t+1} - 1 + u_i^t, 0). \tag{3.9}$$

To give a stopping rule, let $A = (y_1\bar{\mathbf{x}}_1, \cdots, y_n\bar{\mathbf{x}}_n)^T$, and define

$$\mathbf{r}^{t+1} = A\bar{\mathbf{b}}^{t+1} - 1 + \boldsymbol{\xi}^{t+1} - \boldsymbol{\eta}^{t+1}, \quad \mathbf{s}^{t+1} = \rho A^T(\boldsymbol{\xi}^{t+1} - \boldsymbol{\eta}^{t+1} - \boldsymbol{\xi}^t + \boldsymbol{\eta}^t),$$

$$\epsilon_{\text{pri}} = \sqrt{n}\epsilon + \epsilon\max\{\|A\bar{\mathbf{b}}^{t+1}\|_2, \|\boldsymbol{\xi}^{t+1} - \boldsymbol{\eta}^{t+1}\|_2, 1\}, \quad \epsilon_{\text{dual}} = \sqrt{p}\epsilon + \epsilon\rho\|A^T\mathbf{u}^{t+1}\|_2,$$

where $\epsilon > 0$ is the tolerance. The iteration for (3.2) terminates when $\|\mathbf{r}^{t+1}\|_2 < \epsilon_{\text{pri}}$ and $\|\mathbf{s}^{t+1}\|_2 < \epsilon_{\text{dual}}$, or it reaches the maximum number of iterations. The computation strategy for solving (3.1) is summarized in the next algorithm.

Algorithm 2

Step 1 (Initialization): Specify $\bar{\mathbf{b}}^0, \boldsymbol{\xi}^0, \boldsymbol{\eta}^0, \mathbf{u}^0, \rho$, and ζ.

Step 2 (IADMM iteration): Given $\bar{\mathbf{b}}^m$, solve (3.2) to yield $\bar{\mathbf{b}}^{m+1}$ using the IADMM iteration by updating (3.6), (3.8), and (3.9) iteratively until $\|\mathbf{r}^{t+1}\|_2 < \epsilon_{\text{pri}}$ and $\|\mathbf{s}^{t+1}\|_2 < \epsilon_{\text{dual}}$, or it reaches the maximum number of iterations M_{ADMM}.

Step 3 (DCA iteration): Repeat Step 2 until $\|\bar{\mathbf{b}}^m - \bar{\mathbf{b}}^{m+1}\|/\|\bar{\mathbf{b}}^m\| < \varepsilon$ or it reaches the maximum number of iterations M_{DCA}.

With the hinge loss function, the minimization of $S(f, \boldsymbol{y})$ can be solved using the preceding algorithm without the ψ_2 part in Step 2, followed by Step 3. The solution to

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

(2.2) with the hinge loss can serve as the initial value for the algorithm with the ψ-loss. Importantly, an iterative improvement of the ψ-learning solution is often seen over the corresponding SVM solution. In terms of convergence, Algorithm 2 converges rapidly, owing to the finite-step termination property of the DC algorithm and the IADMM.

4 Tuning Without Negative Instances

In classification, tuning parameters are usually selected using cross-validation by minimizing the classification error over a tuning set of data with complete label information. However, in our problem, negative instances are unavailable for the tuning set, which makes the cross-validation scheme infeasible. To overcome this difficulty, Lee and Liu (2003) propose the criterion $r^2/\Pr(\text{sign}(f(X)) = 1)$, which is proportional to the square of the geometric mean of the precision and the recall of retrieving the positive class. This criterion tries to mimic the behavior of an F-score, the harmonic mean of the precision and the recall. However, when a classifier's performance is evaluated using the classification error, this criterion may not be relevant, because it has no direct relationship with the error. Consequently, to target the classification error, we propose a new criterion for selecting the tuning parameters, as follows. Note that the classification error $\text{Err}(f) = \Pr(\text{sign}(\text{f(X)}) \neq Y) = 1 - \Pr(\text{sign}(f(X)) = -1, Y = -1) - \Pr(\text{sign}(f(X)) = 1, Y = 1)$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

can be rewritten as

$$\Pr(\mathrm{sign}(f(X)) = 1) + 2\Pr(Y = 1)\Pr(\mathrm{sign}(f(X)) = -1|Y = 1) - \Pr(Y = 1).$$

Therefore, because $\Pr(Y = 1)$ at the population level does not contain the tuning parameter, minimizing the classification error with respect to this parameter is equivalent to minimizing

$$\begin{aligned}
&\Pr(\mathrm{sign}(f(X)) = 1) + 2\Pr(Y = 1)\Pr(\mathrm{sign}(f(X)) = -1|Y = 1) \\
&\quad = \big(w\Pr(\mathrm{sign}(f(X)) = 1) + (1 - w)\Pr(\mathrm{sign}(f(X)) = -1|Y = 1)\big) * \big(1 + 2\Pr(Y = 1)\big) \\
&\quad \propto \mathrm{Err}^*(f),
\end{aligned}$$

where $w = 1/\big(1 + 2\Pr(Y = 1)\big)$, and

$$\mathrm{Err}^*(f) = \big(w\Pr(\mathrm{sign}(f(X)) = 1) + (1 - w)\Pr(\mathrm{sign}(\mathrm{f(X)}) = -1|Y = 1)\big). \tag{4.1}$$

It is clear that $\Pr(\mathrm{sign}(f(X)) = -1|Y = 1)$ decreases as $\Pr(\mathrm{sign}(f(X) = 1))$ increases, and vice versa. Thus, by estimating $\Pr(\mathrm{sign}(f(X)) = 1)$ and $\Pr(\mathrm{sign}(f(X)) = -1|Y = 1)$ using a tuning sample that contains instances with the positive class, the tuning parameter can be selected by minimizing the proposed criterion $\mathrm{Err}^*(f)$ in (4.1) empirically, provided that we have knowledge of $\Pr(Y = 1)$ and w. In real applications, the value of $\Pr(Y = 1)$ may either come from prior information, such as the prevalence of a disease in the whole population, or be estimated empirically using the percentage of positively labeled instances in the training set. However, the latter approach tends to underestimate the probability, because positive instances in the unlabeled data are

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

treated as unlabeled instances. Our simulation shows that this criterion performs well for tuning.

5 Numerical Examples

This section compares the proposed method with two strong competitors using simulations: the BSVM (Liu et al., 2003) and the BASVM (Mordelet and Vert, 2014). We denote the ψ-learning version of tbe BSVM as BPSI, and denote our iterative methods with the hinge loss and the ψ-loss as ISVM and IPSI, respectively. All methods are computed using R 3.5.0.

For the simulations, the test error (the classification error on the test set), averaged over 100 independent replications, is used to evaluate the performance of a method. We define the amount of improvement of an iterative classifier over its biased counterpart in terms of the Bayesian regret:

$$\frac{(T(biased) - T(Bayes)) - (T(iterative) - T(Bayes))}{T(biased) - T(Bayes)}, \tag{5.1}$$

where $T(\cdot)$ and $T(Bayes)$ represent the test error of a method and the Bayes error, respectively. For real examples, because the Bayes rule is unknown, we define the amount of improvement as

$$\frac{T(biased) - T(iterative)}{T(biased)}, \tag{5.2}$$

which may underestimate the amount of improvement compared to (5.1).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

5.1 Simulated and Real-Data Examples

Two simulated and two real-data examples are examined, in which unlabeled instances are generated by dropping the labels of some instances. Examples 1 and 2 are simulated following the set up of Wang and Shen (2007), where the two Bayes errors are 0.1587 and 0.089, respectively. The two real examples, HEART and SPAM, are available in the UCI Machine Learning Repository (Lichman, 2013). Here, HEART focuses on heart disease classification, based on 13 numeric-valued clinical attributes, and SPAM discriminates spam from normal e-mails based on 57 frequency attributes.

To generate the one-class situation, in two real examples, each class is treated as a novel/negative class once, with the other treated as a positive class. Two cases with different sizes of positively labeled and unlabeled samples are considered. In the first case, the data are split randomly into three parts, with five positively labeled and 95 unlabeled instances for training, and 100 labeled instances for tuning; the remaining 800 instances in Examples 1 and 2 and the 97 in HEART are used for testing. In the second case, the data are divided randomly into three parts, with 10 positively labeled instances and 90 unlabeled instances for training, and 100 labeled instances for tuning; again, the remaining 800 in Examples 1 and 2 and the 97 in HEART are used for testing. For SPAM, the sizes of the training and tuning samples increase to 200, and the remaining 4201 instances are used for testing. Note that all 100 instances in the tuning set for the two cases are considered **labeled**, which allows us to select the tuning parameters of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

different methods using a usual criterion, such as the generalization error on the tuning set.

For tuning, the generalization error, defined as $GE(f) = P(Y \neq \text{sign}(f(X)))$, is minimized with respect to the tuning parameters over a set of grid points within the tuning domain. More specifically, for the BSVM and BPSI, there are two tuning parameters, C_+ and C_-; for the BASVM, there are four tuning parameters, C_+, C_-, the size of the bootstrap samples K, and the number of bootstraps T; for the BLSSVM, there are four tuning parameters, C_+, C_-, a radial basis function kernel parameter σ, and a parameter λ in the regularization term for local discrepancies in the labels. For our iterative methods ISVM and IPSI, there is only one parameter C.

The search set of C and C_- is $\{10^{-4+j/10}; j = 0, \ldots, 80\}$, and that of $w = C_-/(C_+ + C_-)$ is $\{0.01, \ldots, 0.15\}$. For the BASVM, to reduce the computational cost, we tune the parameter C and the other parameters using the default setting of Mordelet and Vert (2014); that is, $w = n_+/(n_+ + n_-)$, the size of the bootstrap samples $K = n_l$, and the number of bootstraps $T = 35$ if $K \leq 20$; otherwise, $T = 11$. For σ and λ in the BLSSVM, both vary in the set $\{2^j; j = -6, -5, \ldots, 6\}$, as suggested in the setting of Ke et al. (2018).

For testing, a classification model with estimated tuning parameters is evaluated over a test set. The averaged test error based on 100 replications is reported in Table 1.

Table 1 about here

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

As indicated in Table 1, ISVM and IPSI outperform their counterparts BSVM and BPSI in all cases. In particular, in the simulated examples, the amounts of improvement of ISVM and IPSI over BSVM and BPSI range from 1.43% to 34.91%, respectively. In the real examples, the amounts of improvement of the iterative method over its biased counterpart range from 7.35% to 23.46%. This shows that an iterative improvement does occur with the proposed method over its biased counterpart. Compared with the BSVM, the BASVM performs relatively poorly in most cases, indicating that the suggested criterion does not work well in our examples. Note that the improvements of our proposed method over the BSVM in cases 1 and 2 for Example 2 in Tables 1 and 2 are both significant, considering 500 repetitions at a 5% significance level. To ensure a fair comparison with other data sets, we still use 100 repetitions. The proposed method with the ψ-loss, BPSI, performs better than its SVM counterpart, BSVM, in most cases, primarily because of the difference in the loss functions.

5.2 Performance with the Proposed Tuning Criterion

When the tuning data set contains only unlabeled data, the generalization error is not applicable directly, as described above. Therefore, this section examines the performance of the four methods using the tuning criterion proposed in (4.1) in Section 4, **in the absence of labeled instances from a novel class**. Specifically, the data are divided randomly into three parts in case 1, with five labeled positive instances and 95 unlabeled

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

instances for training, five labeled positive instances and 95 unlabeled instances for tuning, and the remaining instances used for testing in Examples 1 and 2 and HEART. In case 2, the data are divided randomly into three parts, with 10 labeled positive instances and 90 unlabeled instances for training, 10 labeled positive instances and 90 unlabeled instances for tuning, and the remaining instances used for testing in Examples 1 and 2 and HEART. For SPAM, the sizes of the training and tuning samples are doubled, and the remaining 4201 instances are used for testing in both cases. For the proposed tuning criterion in (4.1), w is specified by its definition, where $\Pr(\text{sign}(f(X) = 1)$ is replaced by 0.5, owing to the prior information that the generated data are balanced. Then, the tuning criterion is minimized over the tuning set, and the tuning parameters with the smallest criterion value are selected. Finally, we test the fitted model using the selected tuning parameters over the testing set. The averaged test errors based on 100 replications are reported in Table 2. We also set $\Pr(\text{sign}(f(X) = 1)$ as the sample proportion of the labeled class, finding that the performance of the classifiers was similar. The result is omitted to conserve space.

As suggested by Table 2, the ISVM and IPSI outperform the BSVM and BPSI in all cases. The amounts of improvement range from 7.36% to 46.12%. Compared with Table 1, the performance of the biased methods deteriorates after tuning. Interestingly, although the BASVM underperforms against the BSVM in Table 1, it outperforms the BSVM after tuning. One possible explanation is that a higher tuning error is anticipated

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

because the BASVM involves more tuning parameters than those of the other methods. Overall, a comparison of Tables 1 and 2 shows that the tuning criterion performs well in terms of selecting the tuning parameters, leading to good accuracy of classification.

Table 2 about here

6 Statistical Learning Theory

6.1 Theory

In binary classification, the Bayes classifier is defined as $\bar{f}_B = \text{sign}(P(Y = 1|X = x) - 1/2)$, which is a global minimizer of the generalization error $GE(f) = P(Y \neq \text{sign}(f(X)))$. Let $\text{sign}(\hat{f}_C)$ be the corresponding classifier defined by the ψ-loss in Algorithm 1. In what follows, we establish an error bound in terms of the Bayesian regret $e(\hat{f}_C, \bar{f}_B) = GE(\hat{f}_C) - GE(\bar{f}_B) \geq 0$, which is the difference between the generalization errors of our classifier and the Bayes rule. In particular, we establish a probability error bound for $e(\hat{f}_C, \bar{f}_B)$ as a function of the complexity of the candidate decision function set $\mathcal{F}$, the sample size of the labeled data n_l, the sample size of the unlabeled data n_u, the tuning parameter $\lambda = (nC)^{-1}$, the error of the initial classifier $\delta_n^{(0)}$, the sample proportion of negative instances r_n, and the maximum iteration step K. Moreover,we also show that, in the absence of labeled negative instances, the proposed method is still able to recover the performance of supervised ψ-learning based on complete data in terms of the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

rate of convergence under certain assumptions. Let $\mathbf{Z} = (\mathbf{X}, Y)$, $V(f, \mathbf{Z}) = \psi(Yf(\mathbf{X}))$ and $e_V(f, \bar{f}_B) = E(V(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z}))$, the Bayesian regret under the loss $V(f, \mathbf{Z})$, which is $\psi(Yf(\mathbf{X}))$. Furthermore, we assume the following conditions hold.

Assumption A1: (Distribution) Let $P(\mathbf{x}, y)$ denote the joint distribution of $(\mathbf{X}, Y)$. Then, $(\mathbf{x}_i)_{i=1}^{n_l}$ are drawn independently from the conditional distribution $P_{\mathbf{X}|Y=1}(\mathbf{x}, y)$, and $(\mathbf{x}_i)_{i=n_l+1}^{n}$ are drawn independently from the marginal distribution $P_{\mathbf{X}}(\mathbf{x}, y)$.

Assumption A2: (Approximation) For a positive sequence $\eta_n \to 0$ as $n \to \infty$, there exists $f^* \in \mathcal{F}$, such that $e_V(f^*, \bar{f}_B) \leq \eta_n$.

Assumption A3: (Smoothness) There exist positive constants α, β, ζ, and a_i, for $i = 0, 1, 2$, such that for any sufficiently small $\delta > 0$,

$$\sup_{\{f \in \mathcal{F}: e_V(f, \bar{f}_B) \leq \delta\}} e(f, \bar{f}_B) \leq a_0 \delta^{\alpha}, \tag{6.1}$$

$$\sup_{\{f \in \mathcal{F}: e_V(f, \bar{f}_B) \leq \delta\}} \|\text{sign}(f) - \text{sign}(\bar{f}_B)\|_1 \leq a_1 \delta^{\beta}, \tag{6.2}$$

$$\sup_{\{f \in \mathcal{F}: e_V(f, \bar{f}_B) \leq \delta\}} \text{Var}(V(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z})) \leq a_2 \delta^{\zeta}. \tag{6.3}$$

Remark. Assumption A2 is also used by Shen et al. (2003), and it ensures that the Bayes rule $\bar{f}_B$ can be well approximated by decision functions in $\mathcal{F}$. Assumption A3 measures the local behavior of $e(f, \bar{f}_B)$, $\|\text{sign}(f) - \text{sign}(\bar{f}_B)\|_1$, and $\text{Var}(V(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z}))$ within a neighborhood of $\bar{f}_B$. A similar assumption is used in Wang, Shen, and Pan (2009).

To describe Assumption A4, we introduce the L_2-metric entropy with bracketing

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

for the function class $\mathcal{F}$. Given any $\varepsilon > 0$, $\{(f_i^l, f_i^u)\}_{i=1}^I$ satisfying $\|f_i^l - f_i^u\|_2 \leq \varepsilon$, for $i = 1, \ldots, I$, is called an ε-bracketing function set of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists i such that $f_i^l \leq f \leq f_i^u$. Then, the L_2-metric entropy with bracketing for the function class $\mathcal{F}$ is defined as the smallest $\log(I)$, and is denoted by $H_B(\varepsilon, \mathcal{F})$. Using the above notation, Assumption A4 is formally given in the following.

Assumption A4: (Complexity) For some constants $a_i > 0$, for $i = 3, 4, 5$, and $\varepsilon_n > 0$,

$$\sup_{k \geq 2} \phi(\varepsilon_n, k) \leq a_5 n^{1/2}, \tag{6.4}$$

where $\phi(\varepsilon, k) = \int_{a_4 N}^{a_3^{1/2} N^{\min(1,\zeta)/2}} H_B^{1/2}(u, \mathcal{F}(k)) du/N$, $\mathcal{F}(k) = \{V(f, \mathbf{z}) - V(f^*, \mathbf{z}) : f \in \mathcal{F}, J(f) \leq k\}$, $N = N(\varepsilon, \lambda, k) = \min(\varepsilon^2 + \lambda(k/2 - 1)J^*, 1)$, and $J^* = \max(1, J(f^*))$.

Refer to Shen et al. (2003) for more details on Assumption 4. Combining the technical assumptions from A1 to A4, the following results are established.

Theorem 3. *Under Assumptions A1–A4 and* $\delta_n^2 = \min(\max(\varepsilon_n^2, 4\eta_n), 1) \geq 4\lambda J^*$, *there exist some positive constants* a_6 *and* a_7, *such that*

$$\begin{aligned}
&P\Big(e(\hat{f}_C, \bar{f}_B) \geq a_0 \max(\delta_n^{2\alpha}, (\rho_n(\delta_n^{(0)})^2)^{\alpha \max(1, B^K)}\Big) \\
\leq &P\Big(e_V(\hat{f}^{(0)}, \bar{f}_B) \geq \rho_n(\delta_n^{(0)})^2\Big) + 24K \; exp(-a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)}) + \\
&24K \; exp\Big(-a_7 n_u\big(r_n - a_1 \rho_n^{\beta}(\rho_n(\delta_n^{(0)})^2)^{\beta \min(1, B^K)}\big)(\lambda J^*)^{2-\min(1,\zeta)}\Big) + K\rho_n^{-\beta},
\end{aligned}$$

where $B = \frac{2\beta\zeta}{1+\max(0,1-\beta)}$, *$K$ is the finite number of iterations of Algorithm 1 at termination,* $\rho_n > 0$ *is a real number, and* r_n *denotes the sample proportion of truly negative instances.*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

Theorem 3 establishes a finite-sample probability bound for $e(\hat{f}_C, \bar{f}_B)$. The parameter B measures the level of difficulty of the underlying problem, with smaller B indicating more difficulty. Note that B is proportional to β and ζ in Assumption A3. As $n_l, n_u \to \infty$, we obtain the convergence rate of the IPSI, which is determined by the error rate of the corresponding supervised ψ-learning with complete data, error rate of the initial classifier, and maximum iteration steps K.

Corollary 1. *Under the assumptions of Theorem 3, as $n_l, n_u \to \infty$,*

$$|e(\hat{f}_C, \bar{f}_B)| = O_p\Big(\max\big(\delta_n^{2\alpha}, (\rho_n(\delta_n^{(0)})^2)^{\alpha \max(1, B^K)}\big)\Big) \text{ and}$$
$$E|e(\hat{f}_C, \bar{f}_B)| = O\Big(\max\big(\delta_n^{2\alpha}, (\rho_n(\delta_n^{(0)})^2)^{\alpha \max(1, B^K)}\big)\Big),$$

provided that the initial classifier satisfying $P\big(e_V(\hat{f}^{(0)}, \bar{f}_B) \geq \rho_n(\delta_n^{(0)})^2\big) \to 0$, with $\rho_n \to \infty$ and $\rho_n(\delta_n^{(0)})^2 \to 0$, $a_1\rho_n^{\beta}(\rho_n(\delta_n^{(0)})^2)^{\beta \min(1, B^K)} < r_n$, and the tuning parameter λ is selected such that $n_l(\lambda J^*)^{2-\min(1,\zeta)}$ and $n_u\big(r_n - a_1\rho_n^{\beta}(\rho_n(\delta_n^{(0)})^2)^{\beta \min(1, B^K)}\big)(\lambda J^*)^{2-\min(1,\zeta)}$ are bounded away from zero.

The parameter B describes two cases. When $B > 1$, the IPSI reaches the convergence rate of its supervised counterpart with complete data (Shen et al. (2003)). However, this is not guaranteed when $B \leq 1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

6.2 A Theoretical Example

We apply Theorem 3 to a specific learning example to obtain an error rate for the proposed method IPSI in terms of the Bayesian regret. Consider a linear classification problem in which the unlabeled data $\mathbf{X} = (X_1, X_2)^T$ form a sample from a marginal density $q(x) = \frac{1}{2}(1+\theta_1)|x|^{\theta_1}$, for $-1 \leq x \leq 1$, with $\theta_1 > 0$. Given $\mathbf{x} = (x_1, x_2)^T$, the conditional distribution of the positive label is $P(Y=1|\mathbf{x}) = \frac{1}{2}\text{sign}(x_1)|x_1|^{\theta_2} + \frac{1}{2}$ with $\theta_2 > 0$, where the parameters θ_1 and θ_2 describe the shape of the marginal density near the origin and the shape of the conditional class probability around 0.5, respectively. The labeled data are a random sample from $P(\mathbf{x}|Y=1)$. Note that $f_B = x_1$.

Assumption A1 is easily satisfied. We now verify Assumptions A2-A4. For simplicity, we restrict $\mathcal{F}$ to $\mathcal{F}_1 = \{f(x) = (1, x_1)\mathbf{w} : \mathbf{w} \in \mathcal{R}^2\}$ because X_1 and X_2 are independent. For assumption A2, let $f^* = nf_B$. Then, we have $e_V(f^*, \bar{f}_B) \leq P(|nf_B(X_1)| \leq 1) \leq \frac{1+\theta_1}{n} = \eta_n$. Because $e_V(f, \bar{f}_B) \geq e(f, \bar{f}_B)$, (6.1) in Assumption A3 holds for $\alpha = 1$. Direct calculations yield that there exist constants $c_1, c_2 > 0$ such that for $f \in \mathcal{F}_1$, $e_V(f, \bar{f}_B) \geq e(f, \bar{f}_B) = c_1(-\frac{d_0}{1+d_1})^{1+\theta_1+\theta_2}$ and $E|\text{sign}(f) - \text{sign}(\bar{f}_B)| = c_2(-\frac{d_0}{1+d_1})^{1+\theta_1}$, with $w_f = w_{f_B} + (d_0, d_1)^T$, which implies that $\beta = \frac{1+\theta_1}{1+\theta_1+\theta_2}$ in (6.2). To check (6.3), by the triangle inequality, $\text{Var}(V(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z})) \leq E|V(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z})| \leq \Delta_1 + \Delta_2$, where $\Delta_1 = E|l(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z})| \leq E|\text{sign}(f) - \text{sign}(\bar{f}_B)| \leq c_3 e_V(f, \bar{f}_B)^{\frac{1+\theta_1}{1+\theta_1+\theta_2}}, \Delta_2 = E(V(f, \mathbf{Z}) - l(f, \mathbf{Z})) = E(V(f, \mathbf{Z}) - V(\bar{f}_B, \mathbf{Z})) + E(l(\bar{f}_B, \mathbf{Z}) - l(f, \mathbf{Z})) \leq 2e_V(f, \bar{f}_B)$, and c_3 is a constant. Hence, (6.3) holds with $\zeta = \frac{1+\theta_1}{1+\theta_1+\theta_2}$. For (6.4), let $\phi_1(\varepsilon, k) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

$a_3(\log(1/N^{1/2}))^{1/2}/N^{1/2}$. By Lemma 6 of Wang and Shen (2007), solving (6.4) yields $\varepsilon_n = (\log n/n)^{1/2}$ when $C/J^* \sim \delta_n^{-2}n^{-1} \sim (\log n)^{-1}$. Therefore, $B = \frac{2(1+\theta_1)^2}{(1+\theta_1+2\theta_2)(1+\theta_1+\theta_2)}$. Applying Theorem 3 yields $E|e(\hat{f}_C, \bar{f}_B)| = O(\max(n^{-1}\log n, (\rho_n(\delta_n^{(0)})^2)^{\max(1,B^K)}))$. When $B > 1$ or, equivalently, $1+\theta_1 > \frac{3+\sqrt{17}}{2}\theta_2$, the rate is $O(n^{-1}\log n)$ for sufficiently large K, and is $O(\rho_n(\delta_n^{(0)})^2)$ otherwise.

It is clear that our proposed method achieves a fast rate $n^{-1}\log n$ when θ_1 is larger than θ_2, indicating that the marginal density $q(x)$ is low around the origin. This is in accordance with the low density separation condition of Chapelle and Zien (2005) for semi-supervised learning.

7 Discussion

This study develops a large-margin semi-supervised classifier for detecting a novel class with labeled instances from only one class. In particular, the proposed method achieves higher prediction accuracy. The numerical analysis illustrates that our method is highly competitive against the state-of-the-art BSVM and BASVM. The theoretical results show that it can recover the performance of its supervised counterpart with complete data. Note that the proposed method involves only one tuning parameter, as opposed to the two tuning parameters for the BSVM, reducing the cost of tuning numerically. Finally, a generalization of the proposed method to multiclass learning may require

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

further investigation.

Acknowledgements Liu's research was supported by the Fundamental Research Funds for the Central Universities. Zheng's research was supported by the Graduate Innovation Foundation of Shanghai University of Finance and Economics, grant CXJJ-2014-461. Shen's research was partially supported by US National Science Foundation DMS-1712564 and DMS-1952539, and Wang's research was partially supported by NSFC grant 11371235.

Appendix

A. Proofs

Proof of Theorem 1: Note that $S(\hat{f}^{(k+1)}, \mathbf{y}^{k+1}) \leq S(\hat{f}^{(k+1)}, \mathbf{y}^k)$ and $\hat{f}^{(k+1)}$ minimizes the objective $S(f, \mathbf{y}^k)$. Then $S(\hat{f}^{(k+1)}, \mathbf{y}^{k+1}) \leq S(\hat{f}^{(k)}, \mathbf{y}^k)$. That is, $S(\hat{f}^{(k)}, \mathbf{y}^k)$ is decreasing in k. Therefore, Algorithm 1 converges as $k \to \infty$ and terminates finitely for any given precision ε. This completes the proof.

Proof of Theorem 2: Let $\hat{b}_0^{k+1} = \text{argmin}_{b_0} S((b_0, \mathbf{0}_p); \mathbf{Y}^k)$, then it suffices to show that $P(\partial S((\hat{b}_0^{k+1}, \mathbf{0}_p))/\partial \mathbf{b} \neq \mathbf{0}_p) > 0$. It is easy to see that $\hat{b}_0^{k+1}$ can be any constant in $[-1, 1]$. Furthermore, $\partial S((\hat{b}_0^{k+1}, \mathbf{0}_p))/\partial \mathbf{b} = \sum_{Y_i^k=1} \partial L(\hat{b}_0^{k+1})\mathbf{X}_i/n_+^k - \sum_{Y_j^k=-1} \partial L(-\hat{b}_0^{k+1})\mathbf{X}_j/n_-^k$, where ∂ represents the partial sub-gradient. For the hinge loss $L(z) = (1-z)_+$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

$\partial S((\hat{b}_0^{k+1}, \mathbf{0}_p))/\partial \mathbf{b} \neq \mathbf{0}_p$ is equivalent to $\sum_{Y_i^k=1} \mathbf{X}_i/n_+^k \neq \sum_{Y_j^k=-1} \mathbf{X}_j/n_-^k$. For the ψ-loss, we need $\sum_{Y_i^k=1} \mathbf{X}_i/n_+^k \neq 0$ and $\sum_{Y_j^k=-1} \mathbf{X}_j/n_-^k \neq 0$ additionally. Therefore, under the conditions of Theorem 2, $P(\hat{\mathbf{b}}^{k+1} \neq \mathbf{0}_p) > 0$.

Proof of Theorem 3: Firstly, we bound the probability of the ratio of incorrectly classified unlabeled instances using $\text{sign}(\hat{f}^{(k)})$ by the tail probability of $e_V(\hat{f}^{(k)}, \bar{f}_B)$. Denote by $D_f = \{\text{sign}(\hat{f}^{(k)}(\mathbf{X}_j)) \neq \text{sign}(\bar{f}_B(\mathbf{X}_j)), n_l + 1 \leq j \leq n\}$ the set of incorrectly classified instances and $n_f = \#D_f$. By Markov's inequality, the fact that $E(\frac{n_f}{n}) = \frac{n_u}{n} E\|\text{sign}(\hat{f}^{(k)}) - \text{sign}(\bar{f}_B)\|_1$, and (6.2), we obtain

$$\begin{aligned} P\Big(\frac{n_f}{n} \geq a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta\Big) \leq & P\Big(\|\text{sign}(\hat{f}^{(k)}) - \text{sign}(\bar{f}_B)\|_1 \geq a_1(\rho_n(\delta_n^{(k)})^2)^\beta\Big) \\ & + P\Big(\frac{n_f}{n} \geq \rho_n^\beta \|\text{sign}(\hat{f}^{(k)}) - \text{sign}(\bar{f}_B)\|_1\Big) \\ \leq & P\Big(e_V(\hat{f}^{(k)}, \bar{f}_B) \geq \rho_n(\delta_n^{(k)})^2\Big) + \rho_n^{-\beta}. \qquad \text{(A.1)} \end{aligned}$$

Then we will establish the connection between $P\Big(e_V(\hat{f}^{(k+1)}, \bar{f}_B) \geq \rho_n(\delta_n^{(k+1)})^2\Big)$ and $P\Big(e_V(\hat{f}^{(k)}, \bar{f}_B) \geq \rho_n(\delta_n^{(k)})^2\Big)$, where $\rho_n(\delta_n^{(k+1)})^2 = (\rho_n(\delta_n^{(k)})^2)^B$ and $B = \frac{2\beta\zeta}{1+\max(0,1-\beta)}$. For simplicity, let $\delta_k^2 = \rho_n(\delta_n^{(k)})^2$. Moreover, $\mathbf{Z}_j = (\mathbf{X}_j, Y_j)$ with $Y_j = \text{sign}(\hat{f}^{(k)}(\mathbf{X}_j)), n_l + 1 \leq j \leq n$. Define a scaled empirical process $E_{n_+^k}(V(f^*, \mathbf{Z}) - V(f, \mathbf{Z})) = \frac{1}{n_+^k} \sum_{Y_i=1} \big(V(f^*, \mathbf{Z}_i) - V(f, \mathbf{Z}_i) - E(V(f^*, \mathbf{Z}_i) - V(f, \mathbf{Z}_i))\big)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

By the definition of $\hat{f}^{(k)}$ and (A.1), we have

$$\begin{aligned}
&P\Big(e_V(\hat{f}^{(k+1)}, \bar{f}_B) \geq \rho_n(\delta_n^{(k+1)})^2\Big)\\
\leq &P\Big(\frac{n_f}{n} \geq a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta\Big) + P^*\Big(\sup_{N_k} \frac{1}{n_+^k}\sum_{Y_i=1}\big(V(f^*, \mathbf{Z}_i) - V(f, \mathbf{Z}_i)\big) +\\
&\frac{1}{n_-^k}\sum_{Y_j=-1}\big(V(f^*, \mathbf{Z}_j) - V(f, \mathbf{Z}_j)\big) + \lambda(J(f^*) - J(f)) \geq 0, \frac{n_f}{n} \leq a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta\Big)\\
&\leq P\Big(e_V(\hat{f}^{(k)}, \bar{f}_B) \geq \rho_n(\delta_n^{(k)})^2\Big) + \rho_n^{-\beta} + I_1 + I_2, \qquad \text{(A.2)}
\end{aligned}$$

where $N_k = \{f \in \mathcal{F} : e_V(f, \bar{f}_B) \geq \delta_{k+1}^2\}$, $I_1 = P^*\big(\sup_{N_k} \frac{1}{n_+^k}\sum_{Y_i=1}(\tilde{V}(f^*, \mathbf{Z}_i) - \tilde{V}(f, \mathbf{Z}_i)) \geq 0, \frac{n_f}{n} \leq a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta\big)$, $I_2 = P^*\big(\sup_{N_k} \frac{1}{n_-^k}\sum_{Y_j=-1}(V(f^*, \mathbf{Z}_j) - V(f, \mathbf{Z}_j)) \geq 0, \frac{n_f}{n} \leq a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta\big)$, and $\tilde{V}(f, \mathbf{Z}) = V(f, \mathbf{Z}) + \lambda J(f)$.

To bound I_1, we partition N_k into a sequence of sets $A_{s,t}$ with $A_{s,t} = \{f \in \mathcal{F} : 2^{s-1}\delta_{k+1}^2 \leq e_V(f, \bar{f}_B) < 2^s\delta_{k+1}^2, 2^{t-1}J^* \leq J(f) < 2^tJ^*\}$ and $A_{s,0} = \{f \in \mathcal{F} : 2^{s-1}\delta_{k+1}^2 \leq e_V(f, \bar{f}_B) < 2^s\delta_{k+1}^2, J(f) < J^*\}; s, t = 1, 2, \ldots$ Thus it suffices to bound I_1 and I_2 separately over each $A_{s,t}$. To bound I_1, we need to bound the first and second moments of $\tilde{V}(f, \mathbf{Z}) - \tilde{V}(f^*, \mathbf{Z})|Y = 1$ over each $A_{s,t}$. Without loss of generality, assume that $e_{V|Y}(f, \bar{f}_B) \geq c_1 e_V(f, \bar{f}_B)$, $\delta_k^2 \geq \delta_n^2$, $J(f^*) \geq 1$, and thereby $J^* = \max(J(f^*), 1) = J(f^*)$.

For the first moment, since $\delta_{k+1}^2 \geq 4\lambda J(f^*)$, we obtain

$$\inf_{A_{s,t}} E(\tilde{V}(f, \mathbf{Z}) - \tilde{V}(f^*, \mathbf{Z})|Y = 1) \geq (c_1 2^{s-1} - 1/4)\delta_{k+1}^2 + \lambda(2^{t-1} - 1)J(f^*) = M(s, t),$$

$$\inf_{A_{s,0}} E(\tilde{V}(f, \mathbf{Z}) - \tilde{V}(f^*, \mathbf{Z})|Y = 1) \geq (c_1 2^{s-1} - 1/2)\delta_{k+1}^2 = M(s, 0),$$

where $s, t = 1, 2, \ldots$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

For the second moment, note that $\text{Var}(V(f,\mathbf{Z}) - V(f^*,\mathbf{Z})) \leq 2(\text{Var}(V(f,\mathbf{Z}) - V(\bar{f}_B,\mathbf{Z})) + \text{Var}(V(f^*,\mathbf{Z}) - V(\bar{f}_B,\mathbf{Z})))$. By Assumption A3,

$$\sup_{A_{s,t}} \text{Var}(\tilde{V}(f,\mathbf{Z}) - \tilde{V}(f^*,\mathbf{Z})|Y=1) \leq \sup_{A_{s,t}} \frac{\text{Var}(V(f,\mathbf{Z}) - V(f^*,\mathbf{Z}))}{1-r} \leq \frac{4a_2}{1-r} M(s,t)^{\zeta} = \nu(s,t)^2,$$

where r is the population proportion of truly negative instances and $s = 1, 2, \cdots, t = 0, 1, \ldots$

Note that $I_1 \leq I_3 + I_4$, where $I_3 = \sum_{s,t=1}^{\infty} P^*\big(\sup_{A_{s,t}} E_{n_+^k}(V(f^*,\mathbf{Z}) - V(f,\mathbf{Z})) \geq M(s,t)\big)$ and $I_4 = \sum_{s=1}^{\infty} P^*\big(\sup_{A_{s,0}} E_{n_+^k}(V(f^*,\mathbf{Z}) - V(f,\mathbf{Z})) \geq M(s,t)\big)$. By Assumption A4, a direct application of the Theorem 3 of Shen and Wong (1994) with $M = \sqrt{n_+^k} M(s,t), \nu = \nu(s,t)^2, \varepsilon = 1/2, T = 2$ leads to that

$$\begin{aligned} I_3 \leq & \sum_{s,t=1}^{\infty} 3 \exp\Big(- \frac{(1-\varepsilon) n_+^k M(s,t)^2}{2(4\nu(s,t)^2 + 2M(s,t)/3)} \Big) \\ \leq & \sum_{s,t=1}^{\infty} 3 \exp\big(- a_6 n_l M(s,t)^{2-\min(1,\zeta)} \big) \\ \leq & \sum_{s,t=1}^{\infty} 3 \exp\Big(- a_6 n_l \big((c_1 2^{s-1} - 1/4)\delta_{k+1}^2 + \lambda(2^{t-1} - 1) J(f^*) \big)^{2-\min(1,\zeta)} \Big) \\ \leq & 3 \exp\big(- a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)} \big) / \big(1 - \exp(-a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)}) \big)^2, \end{aligned}$$

where $a_6 > 0$ is a constant.

Similarly, $I_4 \leq 3 \exp\big(-a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)}\big) / \big(1 - \exp(-a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)})\big)^2$. Therefore, by combining the bounds of I_3 and I_4, we have that

$$I_1 \leq 6 \exp\big(- a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)} \big) / \big(1 - \exp(-a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)}) \big)^2.$$

For simplicity, assume $\exp\big(- a_6 n_l (\lambda J^*)^{2-\min(1,\zeta)} \big) \leq 1/2$. Hence $I_1 \leq 24 \exp\big(-$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

$a_6 n_l(\lambda J^*)^{2-\min(1,\zeta)}\big)$. Similarly, $I_2 \le 24\exp\big(-a_7 n_u(r_n - a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta)(\lambda J^*)^{2-\min(1,\zeta)}\big)$, where r_n is the sample proportion of truly negative instances.

By substituting the upper bounds of I_1 and I_2 into (A.2), $P\Big(e_V(\hat f^{(k+1)}, \bar f_B) \ge \rho_n(\delta_n^{(k+1)})^2\Big) \le P\Big(e_V(\hat f^{(k)}, \bar f_B) \ge \rho_n(\delta_n^{(k)})^2\Big) + \rho_n^{-\beta} + 24\exp(-a_6 n_l(\lambda J^*)^{2-\min(1,\zeta)}) + 24\exp(-a_7 n_u(r_n - a_1(\rho_n^2(\delta_n^{(k)})^2)^\beta)(\lambda J^*)^{2-\min(1,\zeta)})$. Iterating this inequality yields that

$$
\begin{aligned}
&P\Big(e_V(\hat f^{(K)}, \bar f_B) \ge (\rho_n(\delta_n^{(0)})^2)^{\max(1,B^K)}\Big)\\
\le& P\Big(e_V(\hat f^{(0)}, \bar f_B) \ge \rho_n(\delta_n^{(0)})^2\Big) + 24K\exp\big(-a_6 n_l(\lambda J^*)^{2-\min(1,\zeta)}\big)+\\
&24K\exp\big(-a_7 n_u(r_n - a_1\rho_n^\beta(\rho_n(\delta_n^{(0)})^2)^{\beta\min(1,B^K)})(\lambda J^*)^{2-\min(1,\zeta)}\big) + K\rho_n^{-\beta}.
\end{aligned}
$$

Then Theorem 3 follows from Assumption A3 and $\delta_k^2 \ge \max(\varepsilon_n^2, 4\eta_n) = \delta_n^2$ for any k.

Proof of Corollary 1: It follows from Theorem 3 immediately.

References

An, L. and Tao, P. (1997). Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *J. Glob. Optim.* **11**, 253-285.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein. J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**, 1-122.

Calvo, B., Larrañaga, P. and Lozano, J. A. (2007). Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recogn. Lett.* **28**, 2375-2384.

Calvo, B., López-Bigas, N., Furney, S. J., Larrañaga, P. and Lozano., J. A. (2007). A partially supervised

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

classification approach to dominant and recessive human disease gene prediction. *Comput. Meth. Prog. Bio.* **85**, 229-237.

Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. *AISTATS*, 57-64.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273-297.

Denis, F., Gilleron, R. and Tommasi, M. (2002). Text classification from positive and unlabeled examples. *Proceedings of the Ninth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 1927-1934.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213-220.

Geurts, P.(2011). Learning from positive and unlabeled examples by enforcing statistical significance. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 305-314.

Gu, C. (2000). Multidimension smoothing with splines. *Smoothing and Regression: Approaches, Computation and Application*, 329-354.

Ke, T., Jing, L., Lv, H., Zhang, L. and Hu, Y. (2018). Global and local learning from positive and unlabeled examples. *Appl. Intell.* **48**, 2373-2392.

Kiryo, R., Niu, G., Du Plessis, M. C. and Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. *Adv. Neural. Inf. Process. Syst.*, 1675-1685.

Lee, W. S. and Liu, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression. *ICML* **3**, 448-455.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

Li, X. and Liu, B. (2003). Learning to classify texts using positive and unlabeled data. *IJCAI* **3**, 587-592.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Liu, B., Dai, Y., Li, X., Lee, W. S. and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. *ICDM*, 179-186.

Liu, B., Lee, W. S., Yu, P. S. and Li, X. (2002). Partially supervised classification of text documents. *ICML* **2**, 387-394.

Manevitz, L. M. and Yousef, M. (2001). One-class SVMs for document classification. *J. Mach. Learn. Res.*, **2**, 139-154.

Mordelet, F. and Vert, J. P. (2014). A bagging svm to learn from positive and unlabeled examples. *Pattern Recogn. Lett.* **37**, 201-209.

Natarajan, N., Dhillon, I, Ravikumar, P. and Tewari, A. (2018). Cost-sensitive learning with noisy labels. *J. Mach. Learn. Res.* **18**, 1-33.

Osuna, E., Freund, R. and Girosi, F. (1997). Support vector machines: Training and applications. AI Memo 1602, Massachusetts Institute of Technology.

Tanielian, U. and Vasile, F.(2019). Relaxed softmax for PU learning. *Proceedings of the thirteenth ACM Conference on Recommender Systems*, 119-127.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.* **13**, 1443-1471.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

Shen, X., Tseng, G. C., Zhang, X. and Wong, W. H. (2003). On ψ-learning. *J. Am. Stat. Assoc.* **98**, 724-734.

Shen X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Stat.* **22**, 580-615.

Tax, D. M. J. and Duin, R. P. W. (1999). Support vector domain description. *Pattern Recogn. Lett.* **20**, 1191-1199.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Wahba, G. (1990). Spline models for observational data. *Series in Applied Mathematics*, Vol. 59. SIAM, Philadelphia.

Wang, H., Banerjee, A., Hsieh, C. J., Ravikumar, P. K. and Dhillon, I. S. (2013). Large scale distributed sparse precision estimation. *Adv. Neural. Inf. Process. Syst.*, 584-592.

Wang J. and Shen, X. (2007). Large margin semi-supervised learning. *J. Mach. Learn. Res.* **8**, 1867-1891.

Wang, J., Shen, X. and Pan, W. (2009). On efficient large margin semi-supervised learning: Method and theory. *J. Mach. Learn. Res.* **10**, 719-742.

Yu, H., Han, J. and Chang, K. C. C. (2002). PEBL: positive example based learning for web page classification using SVM. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 239-248.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, 200433, P.R. China
E-mail: liu.xin@mail.shufe.edu.cn

School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, 200433, P.R. China
E-mail: zqlehome@gmail.com

School of Statistics
University of Minnesota
Minneapolis, MN 55347, USA
E-mail: xshen@stat.umn.edu

School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, 200433, P.R. China
E-mail: swang@shufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

Table 1: Averaged test errors tuned using the generalization error based on the tuning sample with all labels known, as well as the corresponding standard errors (in parentheses), over 100 independent replications. In Case 1, $n_u = 19n_l$, $n_l = 5$ in Eg. 1, Eg. 2, and HEART, $n_l = 10$ in SPAM. In Case 2, $n_u = 9n_l$, $n_l = 10$ in Eg. 1, Eg. 2, and HEART, $n_l = 20$ in SPAM. The amount of improvement is defined in (5.1) and (5.2).

Data	Example 1	Example 2	HEART	HEART	SPAM	SPAM
(n, dim)	$(1000, 2)$	$(1000, 2)$	$(297, 13)$	$(297, 13)$	$(4601, 57)$	$(4601, 57)$
Novelty	-1	-1	absent	present	no	yes
			Case 1			
BASVM	.2237(.0072)	.1914(.0074)	.2545(.0084)	.2807(.0076)	.1762(.0048)	.2629(.0054)
BSVM	.1974(.0053)	.1543(.0056)	.2544(.0077)	.2642(.0076)	.1904(.0047)	.2391(.0051)
BLSSVM	.1913(.0051)	.1519(.0052)	.2395(.0071)	.2477(.0077)	.1881(.0042)	.2287(.0052)
ISVM	.1871(.0047)	.1488(.0072)	.2053(.0069)	.2044(.0063)	.1512(.0045)	.2055(.0077)
Improv.	24.10%	7.86%	16.19%	20.51%	18.83%	12.61%
BPSI	.1958(.0042)	.1507(.0064)	.2175(.0073)	.2189(.0064)	.1669(.0045)	.1850(.0051)
IPSI	.1879(.0047)	.1474(.0072)	.1949(.0078)	.2028(.0077)	.1331(.0028)	.1529(.0044)
Improv.	21.31%	5.33%	10.38%	7.35%	20.25%	17.38%
			Case 2			
BASVM	.1921(.0039)	.1497(.0048)	.2161(.0047)	.2505(.0056)	.1345(.0017)	.2178(.0041)
BSVM	.1812(.0030)	.1275(.0028)	.2172(.0049)	.2267(.0056)	.1517(.0022)	.1904(.0041)
BLSSVM	.1803(.0030)	.1276(.0029)	.2037(.0046)	.2102(.0053)	.1466(.0023)	.1755(.0042)
ISVM	.1742(.0023)	.1269(.0033)	.1863(.0041)	.1819(.0038)	.1289(.0015)	.1387(.0022)
Improv.	28.62%	1.43%	12.18%	17.24%	14.36%	23.46%
BPSI	.1834(.0031)	.1327(.0030)	.2093(.0045)	.1990(.0045)	.1465(.0021)	.1489(.0026)
IPSI	.1748(.0024)	.1277(.0033)	.1816(.0039)	.1810(.0037)	.1290(.0015)	.1376(.0021)
Improv.	34.91%	11.39%	13.2%	9.02%	11.94%	7.58%

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0287

Table 2: Averaged test errors tuned using our criterion in Section 4 based on the tuning sample with labeled positive instances, and unlabeled instances, as well as the corresponding standard errors (in parentheses), over 100 independent replications. In Case 1, $n_u = 19n_l$, $n_l = 5$ in Eg. 1, Eg. 2, and HEART, $n_l = 10$ in SPAM. In Case 2, $n_u = 9n_l$, $n_l = 10$ in Eg. 1, Eg. 2, and HEART, $n_l = 20$ in SPAM. The amount of improvement is defined in (5.1) and (5.2).

Data	Example 1	Example 2	HEART	HEART	SPAM	SPAM
(n, dim)	$(1000, 2)$	$(1000, 2)$	$(297, 13)$	$(297, 13)$	$(4601, 57)$	$(4601, 57)$
Novelty	-1	-1	absent	present	no	yes
			Case 1			
BASVM	.2163(.0065)	.2034(.0072)	.2762(.0078)	.2919(.0082)	.1762(.0043)	.2696(.0052)
BSVM	.2362(.0071)	.2123(.0085)	.3007(.0091)	.3178(.0089)	.2158(.0061)	.3117(.0090)
BLSSVM	.2213(.0068)	.2011(.0076)	.2812(.0086)	.2912(.0086)	.1962(.0058)	.2888(.0083)
ISVM	.1916(.0057)	.1712(.0080)	.2251(.0088)	.2481(.0083)	.1574(.0048)	.2390(.0083)
Improv.	46.12%	27.13%	20.02%	18.54%	25.78%	24.12%
BPSI	.2041(.0055)	.1712(.0075)	.2538(.0086)	.2419(.0080)	.1736(.0049)	.2254(.0070)
IPSI	.1818(.0055)	.1627(.0082)	.2201(.0082)	.2383(.0081)	.1377(.0030)	.1693(.0059)
Improv.	27.22%	7.36%	15.13%	2.99%	22.84%	24.71%
			Case 2			
BASVM	.1941(.0041)	.1614(.0049)	.2285(.0055)	.2613(.0065)	.1389(.0024)	.2202(.0045)
BSVM	.2001(.0044)	.1489(.0042)	.2476(.0062)	.2696(.0076)	.1702(.0036)	.2621(.0081)
BLSSVM	.1912(.0044)	.1453(.0041)	.2372(.0058)	.2402(.0071)	.1588(.0040)	.2284(.0076)
ISVM	.1752(.0026)	.1321(.0035)	.2009(.0049)	.1963(.0045)	.1281(.0015)	.1497(.0041)
Improv.	40.24%	23.06%	15.14%	24.24%	21.98%	36.24%
BPSI	.1891(.0030)	.1351(.0037)	.2202(.0047)	.2100(.0060)	.1512(.0025)	.1586(.0040)
IPSI	.1722(.0023)	.1287(.0032)	.1988(.0051)	.1989(.0050)	.1265(.0014)	.1413(.0031)
Improv.	40.62%	13.29%	9.80%	7.03%	15.75%	9.02%