The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Semi-Supervised AUC Optimization without Guessing Labels of Unlabeled Data

Zheng Xie, Ming Li
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center for Novel Software Technology and Industrialization
Nanjing 210023, China
Email: {xiez, lim}@lamda.nju.edu.cn

Abstract

Semi-supervised learning, which aims to construct learners that automatically exploit the large amount of unlabeled data in addition to the limited labeled data, has been widely applied in many real-world applications. AUC is a well-known performance measure for a learner, and directly optimizing AUC may result in a better prediction performance. Thus, semi-supervised AUC optimization has drawn much attention. Existing semi-supervised AUC optimization methods exploit unlabeled data by explicitly or implicitly estimating the possible labels of the unlabeled data based on various distributional assumptions. However, these assumptions may be violated in many real-world applications, and estimating labels based on the violated assumption may lead to poor performance. In this paper, we argue that, in semi-supervised AUC optimization, it is unnecessary to guess the possible labels of the unlabeled data or prior probability based on any distributional assumptions. We analytically show that the AUC risk can be estimated unbiasedly by simply treating the unlabeled data as both positive and negative. Based on this finding, two semi-supervised AUC optimization methods named SAMULT and SAMPURA are proposed. Experimental results indicate that the proposed methods outperform the existing methods.

Introduction

In many real-world applications, collecting a large amount of unlabeled data is relatively easy, while obtaining the labels for the collected data is rather expensive since much human effort and expertise is required for labeling. Semi-supervised learning (Chapelle, Schlkopf, and Zien 2006; Zhu et al. 2009), aiming to construct learners that automatically exploit the large amount of unlabeled data in addition to the limited labeled data in the purpose of improving the learning performance, has drawn significant attention. Many semi-supervised learning methods have been proposed. To effectively exploit the unlabeled data, almost all of these methods elaborate to link the labeled data and the unlabeled data based on certain distributional assumption (Chapelle, Schlkopf, and Zien 2006), such as the cluster assumption, the manifold assumption, etc., and construct the learner by explicitly or implicitly estimating the labels of unlabeled instances.

AUC (area under ROC curve) (Hanley and McNeil 1982), which measures the probability of a randomly drawn positive instance being ranked before a randomly drawn negative instance, is a widely-used performance measure for a learner, especially when the data distribution exhibits certain imbalance. Directly optimizing the AUC during the learning procedure usually lead to a better prediction performance. Many studies are elaborated to show that AUC can be effectively and efficiently optimized (Herschtal and Raskutti 2004; Gao and Zhou 2015; Gao et al. 2013; Ying, Wen, and Lyu 2016).

Some efforts have been devoted to AUC optimization in semi-supervised settings (Amini, Truong, and Goutte 2008; Fujino and Ueda 2016). These methods rely on the aforementioned distributional assumptions. However, those assumptions may be violated in real-world applications and the learner may be biased when explicitly or implicitly estimating the labels based on the violated assumptions, resulting in poor performance or even performance degradation (Cozman and Cohen 2002). Sakai, Niu, and Sugiyama (2017) proposed an unbiased semi-supervised AUC optimization method based on positive-unlabeled learning. However, such a method requires an accurate estimation of the prior probability to reweight the unlabeled data, which is usually difficult especially when the number of labeled data is extremely small.

In this paper, we argue that in semi-supervised AUC optimization problem, it is unnecessary to guess the possible labels of the unlabeled data or prior probability based on any distributional assumptions. We theoretically show that the AUC risk can be estimated unbiasedly by treating the unlabeled data as both positive and negative without any distributional assumptions.

Such a theoretical finding enables us to address semi-supervised AUC optimization problem by simply treating the unlabeled data as both positive and negative data, *without designing specific strategies* to identify the possible labels of the unlabeled data. Based on this theoretical finding, we propose two novel semi-supervised AUC optimization methods: SAMULT, a straightforward unbiased method by regarding the unlabeled data as both positive and negative data, and SAMPURA, an ensemble method by random partitioning the unlabeled data into pseudo-positive and pseudo-negative sets to train base classifiers. The experimental re-

sults based on linear models indicate that the proposed methods outperform the competing methods, and such a result can be easily generalized to non-linear model case. In addition, our finding can also facilitate AUC optimization using only positive and unlabeled data, by simply treating all unlabeled instances as the negatives.

The remainder of this paper is organized as follows. We first introduce preliminaries. Then, we present our theoretical finding, describe the two proposed semi-supervised AUC optimization methods crafted based on the finding and report the experimental results. Finally, we discuss on the related works and conclude the paper.

Preliminary

In supervised learning, the sets of positive and negative data can be denoted as:

$$\mathcal{X}_{\mathrm{P}} := \{\boldsymbol{x}_i\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{P}}(\boldsymbol{x}) := p(\boldsymbol{x} \mid y=+1)\,, \text{ and}$$
$$\mathcal{X}_{\mathrm{N}} := \{\boldsymbol{x}'_j\}_{j=1}^{n_{\mathrm{N}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{N}}(\boldsymbol{x}) := p(\boldsymbol{x} \mid y=-1)\,.$$

For simplicity, we assume a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ in this paper. The non-linear case can be realized by applying a non-linear feature mapping over the input space.

Since AUC is equivalent to the probability of a randomly drawn positive instance being ranked before a randomly drawn negative instance (Hanley and McNeil 1982), it can be formulated as:

$$\mathrm{AUC} = 1 - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))]\big]\,. \tag{1}$$

Maximizing AUC is equivalent to minimizing the following AUC risk. To avoid confusion, we denote the supervised AUC risk as PN-AUC risk:

$$R_{\mathrm{PN}} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))]\big]\,. \tag{2}$$

In semi-supervised setting, the set of unlabeled data is available. The unlabeled instances are considered to be drawn from a mixture of the positive distribution and negative distribution:

$$\mathcal{X}_{\mathrm{U}} := \{\boldsymbol{x}''_k\}_{k=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) = \theta_P p_{\mathrm{P}}(\boldsymbol{x}) + \theta_N p_{\mathrm{N}}(\boldsymbol{x})\,, \tag{3}$$

where θ_P and θ_N are the prior probabilities of the positive and negative class.

The existing semi-supervised AUC optimization methods (Amini, Truong, and Goutte 2008; Fujino and Ueda 2016; Sakai, Niu, and Sugiyama 2017) rely on distributional assumptions or prior knowledge to infer the labels of unlabeled data or to reweight them. If the assumptions hold or the prior knowledge is reliable, these methods may achieve a reasonably good performance. However, when the assumptions are violated or the the prior knowledge is not available, these methods may perform poorly.

We argue that in semi-supervised AUC optimization, it is unnecessary to estimate the labels of unlabeled data. We theoretically and empirically show that we can achieve unbiased semi-supervised AUC optimization without distributional assumptions or prior knowledge about the distribution or class prior probabilities.

Unbiased Estimation without Guessing Label

The general idea of this work is to show that optimizing the risk estimated with the positive data and the unlabeled data treated as negative (or the negative data and the unlabeled data treated as positive) is equivalent to optimizing an unbiased risk. Theorem 1 gives a formal statement.

Theorem 1. *PU-AUC risk R_{PU} which is estimated by positive and unlabeled data treated as negative data, and UN-AUC risk R_{NU} which is estimated by negative and unlabeled data treated as positive data, are equivalent to the supervised PN-AUC risk R_{PN} with a linear transformation, where R_{PU} and R_{NU} are defined as:*

$$R_{\mathrm{PU}} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}''))]\big]\,, \tag{4}$$

$$R_{\mathrm{NU}} = \mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}''-\boldsymbol{x}'))]\big]\,. \tag{5}$$

Proof. Due to the linearity of the expectation, we have:

$$\begin{aligned}
R_{\mathrm{PU}} &= \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}''))]\big] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}[\theta_P \mathop{\mathbb{E}}_{\bar{\boldsymbol{x}}\in\mathcal{X}_{\mathrm{P}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\bar{\boldsymbol{x}}))] \\
&\qquad + \theta_N \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))]] \\
&= \frac{1}{2}\theta_P + \theta_N \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))]\big]\,.
\end{aligned}$$

It is noteworthy that the expected risk of pairs over $\mathcal{X}_{\mathrm{P}} \times \mathcal{X}_{\mathrm{P}}$ is symmetric, so it is equal to a constant. Thus

$$R_{\mathrm{PU}} = \theta_N R_{\mathrm{PN}} + \frac{1}{2}\theta_P\,. \tag{6}$$

Similarly, we also have:

$$R_{\mathrm{NU}} = \theta_P R_{\mathrm{PN}} + \frac{1}{2}\theta_N\,. \tag{7}$$

Above all, we prove that R_{PU} and R_{NU} is equivalent to the supervised AUC risk R_{PN} with a linear transformation. □

Intuitively, the PU-AUC risk R_{PU} can be regarded as a weighted average of two parts: one is the PN-AUC risk on the labeled positive data and unlabeled negative data, and the other is AUC risk over $\mathcal{X}_{\mathrm{P}} \times \mathcal{X}_{\mathrm{P}}$ yielded by positive labeled data and positive unlabeled data. Since the labeled positive data and unlabeled positive data are drawn from the same distribution, the probability of ranking a labeled positive instance after an unlabeled positive instance is $1/2$.

Theorem 1 suggests that optimizing R_{PU} or R_{NU} is equivalent to optimizing the supervised AUC risk asymptotically. Thus, the unlabeled data can be simply treated as the positives or the negatives, and it is unnecessary to estimate the labels of unlabeled data based on certain assumption.

Furthermore, with the fact that $\theta_P + \theta_N = 1$, summing up Eq. (6) and Eq. (7) results in:

$$R_{\mathrm{PU}} + R_{\mathrm{NU}} - \frac{1}{2} = R_{\mathrm{PN}}\,. \tag{8}$$

Eq. (8) suggests that when we have positive, negative and unlabeled data, we can conduct unbiased AUC risk estimation without knowing the class prior probabilities θ_P

and θ_N. This allows us to design new unbiased AUC risk estimators that help to utilize the unlabeled data. Inspired by this fact, we propose a method named SAMULT (Semi-supervised AUC Maximization by treating the UnLabeled data in Two ways), and further extend it to an ensemble method SAMPURA (Semi-supervised AUC Maximization by Partitioning Unlabeled data at RAndom). Further, when only positive and unlabeled data are available, R_{PN} and R_{NU} become zero and only R_{PU} is used to learn the model. In this case, SAMULT degenerates to a special form similar to the supervised AUC optimization. The technical details of the two methods are described in the following section.

Proposed Methods

Based on our theoretical finding, we propose two semi-supervised AUC optimization methods: SAMULT and SAMPURA. SAMULT treats the unlabeled data as both positive and negative data to compute the risk estimator based on Eq. (8), and then combines it with a supervised AUC risk estimator to get an unbiased AUC risk estimator. Moreover, since we do not need to estimate the labels from the whole dataset, we can bootstrap-sample multiple data set from the unlabeled data, label them as either positive or negative and trains multiple classifiers. Thus, we further extend SAMULT to an ensemble version, namely SAMPURA, which firstly trains multiple base classifiers by randomly partitioning the unlabeled data into pseudo-positive and pseudo-negative datasets to augment the labeled data, and then aggregates them to generate a strong classifier.

SAMULT

Based on Eq. (8), we propose the following AUC risk estimator that utilizes all the labeled and unlabeled data in semi-supervised AUC optimization:

$$\widehat{R}_{\mathrm{PNU}} = \gamma\widehat{R}_{\mathrm{PN}} + (1-\gamma)(\widehat{R}_{\mathrm{PU}} + \widehat{R}_{\mathrm{NU}} - \frac{1}{2}), \quad (9)$$

where $\gamma \in [0,1]$ is the trade-off parameter, and

$$\widehat{R}_{\mathrm{PN}} = \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}}} \sum_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}} \sum_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}} \ell(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}')), \quad (10)$$

$$\widehat{R}_{\mathrm{PU}} = \frac{1}{n_{\mathrm{P}} n_{\mathrm{U}}} \sum_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}} \sum_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}} \ell(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'')), \quad (11)$$

$$\widehat{R}_{\mathrm{NU}} = \frac{1}{n_{\mathrm{U}} n_{\mathrm{N}}} \sum_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}} \sum_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}} \ell(\boldsymbol{w}^\top(\boldsymbol{x}''-\boldsymbol{x}')). \quad (12)$$

Since the 0-1 loss is discrete and difficult to optimize, in practice, we replace the 0-1 loss with square loss $\ell(z) = (1-z)^2$, which has been proven to be consistent with AUC asymptotically (Gao and Zhou 2015).

To reduce the risk of overfitting, we optimize the risk estimator with an ℓ_2-regularizer to learn the classifier:

$$\min_{\boldsymbol{w}} \quad \widehat{R}_{\mathrm{PNU}}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||^2, \quad (13)$$

where the parameter $\lambda \geq 0$ is a trade-off between the risk and the regularizer.

In practice, we omit the constant in Eq. (9) which leads to the same minimizer of Eq. (9). The analytical solution of the optimization problem (Eq. (13)) can be computed as:

$$\begin{aligned}\widehat{\boldsymbol{w}} = &(\gamma\boldsymbol{H}_{\mathrm{PN}} + (1-\gamma)(\boldsymbol{H}_{\mathrm{PU}} + \boldsymbol{H}_{\mathrm{NU}}) + \lambda\boldsymbol{I}_d)^{-1} \\ &(\gamma\boldsymbol{h}_{\mathrm{PN}} + (1-\gamma)(\boldsymbol{h}_{\mathrm{PU}} + \boldsymbol{h}_{\mathrm{NU}})),\end{aligned} \quad (14)$$

where

$$\begin{aligned}
\boldsymbol{h}_{\mathrm{PN}} =& \frac{1}{n_{\mathrm{P}}}\boldsymbol{X}_{\mathrm{P}}^\top\mathbf{1}_{n_{\mathrm{P}}} - \frac{1}{n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{N}}^\top\mathbf{1}_{n_{\mathrm{N}}}, \\
\boldsymbol{h}_{\mathrm{PU}} =& \frac{1}{n_{\mathrm{P}}}\boldsymbol{X}_{\mathrm{P}}^\top\mathbf{1}_{n_{\mathrm{P}}} - \frac{1}{n_{\mathrm{U}}}\boldsymbol{X}_{\mathrm{U}}^\top\mathbf{1}_{n_{\mathrm{U}}}, \\
\boldsymbol{h}_{\mathrm{NU}} =& \frac{1}{n_{\mathrm{U}}}\boldsymbol{X}_{\mathrm{U}}^\top\mathbf{1}_{n_{\mathrm{U}}} - \frac{1}{n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{N}}^\top\mathbf{1}_{n_{\mathrm{N}}}, \\
\boldsymbol{H}_{\mathrm{PN}} =& \frac{1}{n_{\mathrm{P}}}\boldsymbol{X}_{\mathrm{P}}^\top\boldsymbol{X}_{\mathrm{P}} - \frac{1}{n_{\mathrm{P}}n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{P}}^\top\mathbf{1}_{n_{\mathrm{P}}}\mathbf{1}_{n_{\mathrm{N}}}^\top\boldsymbol{X}_{\mathrm{N}} \\
&- \frac{1}{n_{\mathrm{P}}n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{N}}^\top\mathbf{1}_{n_{\mathrm{N}}}\mathbf{1}_{n_{\mathrm{P}}}^\top\boldsymbol{X}_{\mathrm{P}} + \frac{1}{n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{N}}^\top\boldsymbol{X}_{\mathrm{N}}, \\
\boldsymbol{H}_{\mathrm{PU}} =& \frac{1}{n_{\mathrm{P}}}\boldsymbol{X}_{\mathrm{P}}^\top\boldsymbol{X}_{\mathrm{P}} - \frac{1}{n_{\mathrm{P}}n_{\mathrm{U}}}\boldsymbol{X}_{\mathrm{P}}^\top\mathbf{1}_{n_{\mathrm{P}}}\mathbf{1}_{n_{\mathrm{U}}}^\top\boldsymbol{X}_{\mathrm{U}} \\
&- \frac{1}{n_{\mathrm{P}}n_{\mathrm{U}}}\boldsymbol{X}_{\mathrm{U}}^\top\mathbf{1}_{n_{\mathrm{U}}}\mathbf{1}_{n_{\mathrm{P}}}^\top\boldsymbol{X}_{\mathrm{P}} + \frac{1}{n_{\mathrm{U}}}\boldsymbol{X}_{\mathrm{U}}^\top\boldsymbol{X}_{\mathrm{U}}, \\
\boldsymbol{H}_{\mathrm{NU}} =& \frac{1}{n_{\mathrm{U}}}\boldsymbol{X}_{\mathrm{U}}^\top\boldsymbol{X}_{\mathrm{U}} - \frac{1}{n_{\mathrm{U}}n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{U}}^\top\mathbf{1}_{n_{\mathrm{U}}}\mathbf{1}_{n_{\mathrm{N}}}^\top\boldsymbol{X}_{\mathrm{N}} \\
&- \frac{1}{n_{\mathrm{U}}n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{N}}^\top\mathbf{1}_{n_{\mathrm{N}}}\mathbf{1}_{n_{\mathrm{U}}}^\top\boldsymbol{X}_{\mathrm{U}} + \frac{1}{n_{\mathrm{N}}}\boldsymbol{X}_{\mathrm{N}}^\top\boldsymbol{X}_{\mathrm{N}},
\end{aligned}$$

$\boldsymbol{X}_{\mathrm{P}}$, $\boldsymbol{X}_{\mathrm{N}}$, and$\boldsymbol{X}_{\mathrm{U}}$ are positive, negative and unlabeled instance matrices, respectively, $\mathbf{1}_d$ is d-dimensional all-one vector, and $\boldsymbol{I}_d$ is the d-dimensional identity matrix. To avoid the $O(d^3)$ computation of matrix inverse, we adopt Sherman-Morrison formula to bring down the computational cost in practice.

When only the positive and unlabeled data are available, $\widehat{R}_{\mathrm{PN}}$ and $\widehat{R}_{\mathrm{NU}}$ are zero, and SAMULT degenerates to a special form where only $\widehat{R}_{\mathrm{PU}}$ is optimized. Omitting the trading-off parameter λ and the constant term, the optimization objective of SAMULT$^{\mathrm{P+U}}$ becomes:

$$\min_{\boldsymbol{w}} \quad \widehat{R}_{\mathrm{PU}}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||^2. \quad (15)$$

This formulation is identical to the supervised AUC optimization by treating the unlabeled data as negative. This indicates that the positive-unlabeled AUC optimization problem can be solved as a supervised AUC optimization problem by treating the unlabeled data as negative data.

In summary, SAMULT optimizes a weighted average of an unbiased supervised AUC risk estimator $\widehat{R}_{\mathrm{PN}}$ and an unbiased semi-supervised AUC risk $(\widehat{R}_{\mathrm{PU}} + \widehat{R}_{\mathrm{NU}} - \frac{1}{2})$. SAMULT does not require carefully designed strategies to identify the labels of the unlabeled data, or any estimation of the class prior probabilities to reweight the unlabeled data, so it is easy to implement with just a few lines of code. Algorithm 1 shows the procedure of SAMULT. It is noteworthy that since AUC optimization aims to optimize the ranking quality of the learner, SAMULT does

not determine the decision boundary. To generate the final classification of instances, the stand-alone thresholding strategies (e.g., (Arampatzis, Kamps, and Robertson 2009; Lipton, Elkan, and Naryanaswamy 2014)) based on the ranking list generated by the AUC optimization can be employed.

Algorithm 1 SAMULT

1: **Input:** $\boldsymbol{X}_{\mathrm{P}}, \boldsymbol{X}_{\mathrm{N}}, \boldsymbol{X}_{\mathrm{U}}, \lambda, \gamma$
2: Compute $\boldsymbol{h}_{\mathrm{PN}}, \boldsymbol{h}_{\mathrm{PU}}, \boldsymbol{h}_{\mathrm{NU}}, \boldsymbol{H}_{\mathrm{PN}}, \boldsymbol{H}_{\mathrm{PU}}, \boldsymbol{H}_{\mathrm{NU}}$
3: Compute closed form solution of $\widehat{\boldsymbol{w}}$ (Eq. (14))
4: **Output:** $\widehat{\boldsymbol{w}}$

SAMPURA

Since minimizing the losses over both PU and UN instance pairs helps to learn the classifier, a straightforward idea is to use the unlabeled data to augment the positive and negative data. We equally divide the unlabeled data $\mathcal{X}_{\mathrm{U}}$ into pseudo-positive data $\mathcal{X}_{\mathrm{U}+}$ and pseudo-negative data $\mathcal{X}_{\mathrm{U}-}$ to augment the original positively and negatively labeled data, respectively. By assuming that the instances in $\mathcal{X}_{\mathrm{P}'} = \mathcal{X}_{\mathrm{P}} \cup \mathcal{X}_{\mathrm{U}+}$ should be ranked before those in $\mathcal{X}_{\mathrm{N}}$, and the instances in $\mathcal{X}_{\mathrm{P}}$ should be ranked before those in $\mathcal{X}_{\mathrm{N}'} = \mathcal{X}_{\mathrm{N}} \cup \mathcal{X}_{\mathrm{U}-}$, the unbiased risk estimator can be defined as:

$$\begin{aligned}\widehat{R}_{\mathrm{PNU}} =& \frac{1}{3}(\widehat{R}_{\mathrm{P}'\mathrm{N}} + \widehat{R}_{\mathrm{PN}'} - \frac{1}{2}) \\ =& \frac{1}{3}\Big(\frac{1}{n_{\mathrm{P}'} n_{\mathrm{N}}} \sum_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}'}} \sum_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}} \ell(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}')) \\ &+ \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}'}} \sum_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}} \sum_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}'}} \ell(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}')) - \frac{1}{2}\Big) .\end{aligned} \tag{16}$$

By generating different partition of the unlabeled data, we can train multiple classifiers to obtain an ensemble, which may potentially reduce the variance. In each partition, we minimize an equivalent empirical risk with an ℓ_2-regularizer to obtain the base classifier:

$$\min_{\boldsymbol{w}} \quad \widehat{R}_{\mathrm{P}'\mathrm{N}} + \widehat{R}_{\mathrm{PN}'} + \lambda||\boldsymbol{w}||^2 , \tag{17}$$

and then take the average of those $\boldsymbol{w}$s to construct the final ensemble.

The analytical solution of the base classifiers can be computed by

$$\widehat{\boldsymbol{w}} = \left(\boldsymbol{H}_{\mathrm{P}'\mathrm{N}} + \boldsymbol{H}_{\mathrm{PN}'} + \lambda \boldsymbol{I}_d\right)^{-1} \left(\boldsymbol{h}_{\mathrm{P}'\mathrm{N}} + \boldsymbol{h}_{\mathrm{PN}'}\right) , \tag{18}$$

where

$$\begin{aligned}\boldsymbol{h}_{\mathrm{P}'\mathrm{N}} =& \frac{1}{n_{\mathrm{P}'}} \boldsymbol{X}_{\mathrm{P}'}^\top \mathbf{1}_{n_{\mathrm{P}'}} - \frac{1}{n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^\top \mathbf{1}_{n_{\mathrm{N}}} , \\ \boldsymbol{h}_{\mathrm{PN}'} =& \frac{1}{n_{\mathrm{P}}} \boldsymbol{X}_{\mathrm{P}}^\top \mathbf{1}_{n_{\mathrm{P}}} - \frac{1}{n_{\mathrm{N}'}} \boldsymbol{X}_{\mathrm{N}'}^\top \mathbf{1}_{n_{\mathrm{N}'}} , \\ \boldsymbol{H}_{\mathrm{P}'\mathrm{N}} =& \frac{1}{n_{\mathrm{P}'}} \boldsymbol{X}_{\mathrm{P}'}^\top \boldsymbol{X}_{\mathrm{P}'} - \frac{1}{n_{\mathrm{P}'} n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{P}'}^\top \mathbf{1}_{n_{\mathrm{P}'}} \mathbf{1}_{n_{\mathrm{N}}}^\top \boldsymbol{X}_{\mathrm{N}} \\ &- \frac{1}{n_{\mathrm{P}'} n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^\top \mathbf{1}_{n_{\mathrm{N}}} \mathbf{1}_{n_{\mathrm{P}'}}^\top \boldsymbol{X}_{\mathrm{P}'} + \frac{1}{n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^\top \boldsymbol{X}_{\mathrm{N}} , \\ \boldsymbol{H}_{\mathrm{PN}'} =& \frac{1}{n_{\mathrm{P}}} \boldsymbol{X}_{\mathrm{P}}^\top \boldsymbol{X}_{\mathrm{P}} - \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}'}} \boldsymbol{X}_{\mathrm{P}}^\top \mathbf{1}_{n_{\mathrm{P}}} \mathbf{1}_{n_{\mathrm{N}'}}^\top \boldsymbol{X}_{\mathrm{N}} \\ &- \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}'}} \boldsymbol{X}_{\mathrm{N}'}^\top \mathbf{1}_{n_{\mathrm{N}'}} \mathbf{1}_{n_{\mathrm{P}}}^\top \boldsymbol{X}_{\mathrm{P}} + \frac{1}{n_{\mathrm{N}'}} \boldsymbol{X}_{\mathrm{N}'}^\top \boldsymbol{X}_{\mathrm{N}'} ,\end{aligned}$$

and $\boldsymbol{X}_{\mathrm{P}'}$ and $\boldsymbol{X}_{\mathrm{N}'}$ are the instance matrices of $\mathcal{X}_{\mathrm{P}'}$ and $\mathcal{X}_{\mathrm{N}'}$, respectively.

The procedure of SAMPURA is shown in Algorithm 2.

Algorithm 2 SAMPURA

1: **Input:** $\boldsymbol{X}_{\mathrm{P}}, \boldsymbol{X}_{\mathrm{N}}, \boldsymbol{X}_{\mathrm{U}}, \lambda, T$
2: **for** $t = 1 \to T$ **do**
3: Random partition $\boldsymbol{X}_{\mathrm{U}}$ into $\boldsymbol{X}_{\mathrm{U}+}$ and $\boldsymbol{X}_{\mathrm{U}-}$ equally
4: Compute $\boldsymbol{h}_{\mathrm{P}'\mathrm{N}}, \boldsymbol{h}_{\mathrm{PN}'}, \boldsymbol{H}_{\mathrm{P}'\mathrm{N}}, \boldsymbol{H}_{\mathrm{PN}'}$
5: $\boldsymbol{X}_{\mathrm{P}'} = [\boldsymbol{X}_{\mathrm{P}}^\top \mid \boldsymbol{X}_{\mathrm{U}+}^\top]^\top, \boldsymbol{X}_{\mathrm{N}'} = [\boldsymbol{X}_{\mathrm{N}}^\top \mid \boldsymbol{X}_{\mathrm{U}-}^\top]^\top$
6: Compute $\boldsymbol{h}_{\mathrm{P}'\mathrm{N}}, \boldsymbol{h}_{\mathrm{PN}'}\ \boldsymbol{H}_{\mathrm{P}'\mathrm{N}}, \boldsymbol{H}_{\mathrm{PN}'}$
7: Compute closed form solution of $\widehat{\boldsymbol{w}}^{(t)}$ (Eq. (18))
8: **end for**
9: **Output:** $\widehat{\boldsymbol{w}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{w}}^{(t)}$

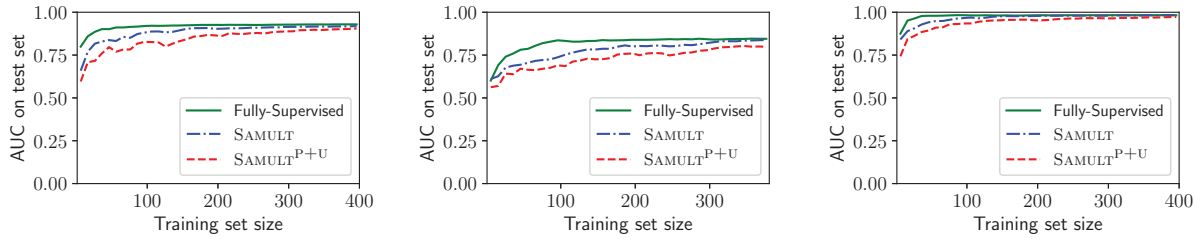
Experiment

We evaluate our methods on 20 widely-used datasets, including 18 datasets from UCI repository (Lichman 2013), and the *ijcnn1* and the *madelon* (Guyon et al. 2005). Table 1 summarizes the statistics of these datasets.

Table 1: Statistics of the datasets.

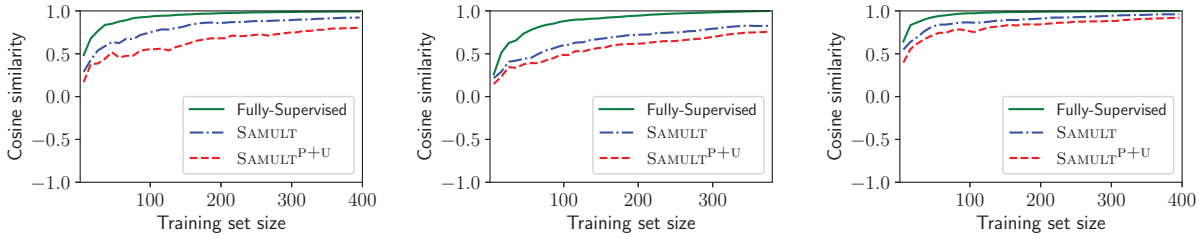
Dataset	# Instance	# Features
australian	690	42
breast	277	9
breastw	683	9
clean1	476	166
colic	188	13
colic.orig	205	17
credit-a	653	15
credit-g	1,000	20
fourclass	862	2
german	1,000	59
haberman	306	14
heart	270	9
house	232	16
ijcnn1	141,691	22
madelon	2600	500
parkinsons	195	22
phishing	11,055	68
vehicle	435	16
vote	232	16
wdbc	569	14

We firstly examine the asymptotical property of the model trained by SAMULT (optimizing PNU-AUC risk $\widehat{R}_{\mathrm{PNU}}$) and SAMULT$^{\mathrm{P+U}}$ (optimizing PU-AUC risk $\widehat{R}_{\mathrm{PU}}$), to show the validity of Theorem 1.



(a) *australian* (b) *clean1* (c) *wdbc*

Figure 1: Figure a, b, and c show the AUC (on test set) vs. training set size curves of classifiers trained by fully-supervised, SAMULT, and SAMULT$^{\text{P+U}}$. For fully-supervised case, all data in training set is labeled. For SAMULT and SAMULT$^{\text{P+U}}$, 10% data in training set is labeled. SAMULT$^{\text{P+U}}$ uses only positive and unlabeled data. Roughly 20% data is held out as the test set.



(a) *australian* (b) *clean1* (c) *wdbc*

Figure 2: Figure a, b, and c shows the cosine similarity ($\cos\langle \widehat{\boldsymbol{w}}, \widehat{\boldsymbol{w}}^* \rangle$) vs. training set size curves of classifiers trained by fully-supervised, SAMULT, and SAMULT$^{\text{P+U}}$, where $\widehat{\boldsymbol{w}}^*$ is learned from all available data with label, and $\widehat{\boldsymbol{w}}$ is learned from training sets of different sizes. For fully-supervised, all data in training set is labeled. For SAMULT and SAMULT$^{\text{P+U}}$, 10% data in training set is labeled. SAMULT$^{\text{P+U}}$ uses only positive and unlabeled data. Roughly 20% data is held out as the test set.

Secondly, we compare our methods to the state-of-the-art semi-supervised AUC optimization methods. All experiments are repeated 50 times with random data partition, and the average AUC scores as well as the standard deviations are recorded. The parameters are chosen by grid search through a 5-fold cross validation.

When only positive and unlabeled data are available, SAMULT degenerates to a special form that only the PU-AUC risk $\widehat{R}_{\text{PU}}$ needs to be optimized, which is equivalent to supervised AUC optimization asymptotically as shown in Theorem 1. We conduct empirical evaluation with respect to this degenerated case.

Asymptotical Property

According to Theorem 1, the model that minimizes the PU-AUC risk or the PNU-AUC risk converges to the supervised case as more data are available. We choose three datasets with different number of features to illustrate the performance of the classifiers learned from training sets of different sizes by SAMULT, which optimizes the PNU risk $\widehat{R}_{\text{PNU}}$, and by SAMULT$^{\text{P+U}}$, which optimizes the PU risk $\widehat{R}_{\text{PU}}$, to illustrate the validity of Theorem 1. For each experiment, 10% of the data in the training set is labeled for SAMULT and SAMULT$^{\text{P+U}}$ to train the classifiers, and SAMULT$^{\text{P+U}}$ uses only positive and unlabeled data. We also trained a fully-supervised classifier (i.e., the golden rule) using all data in training set labeled for comparison, whose performance could be considered as an upper bound of the two semi-supervised classifiers. Roughly 20% of the data is held out as the test set. The experiments are repeated 10 times with random data shuffle and partition.

Figure 1a, 1b, and 1c show the AUC of the classifiers learned from the training sets of different sizes, and Figure 2a, 2b, and 2c show the cosine similarity of the classifiers with the optimal classifier $\widehat{\boldsymbol{w}}^*$ that learned from all available data with label. It can be seen that two semi-supervised classifiers trained by SAMULT and SAMULT$^{\text{P+U}}$ converge to the fully-supervised classifier as the increasing of the training data, in terms of AUC scores and cosine similarities. The classifier trained by SAMULT converges faster than the one by SAMULT$^{\text{P+U}}$, since the latter does not utilize the negative data.

Performance of the Compared Methods

We report the performance of our proposed semi-supervised AUC optimization methods SAMULT and SAMPURA. We compare SAMULT and SAMPURA with two state-of-the-art semi-supervised AUC optimization methods: SSRankBoost (Amini, Truong, and Goutte 2008), a boosting based algorithm for learning a bipartite ranking function with partially labeled data, and PNU-AUC (Sakai, Niu, and Sugiyama 2017), which is a semi-supervised AUC optimiza-

Table 2: Experimental results (mean±std of the AUC) on semi-supervised datasets over 50 repetition. The boldfaces denote the best or comparable methods in terms of the AUC, according to the pairwise t-test at the significance level 5%. The numbers of best or comparable case of each method are shown in the last row.

Dataset	Supervised	Log. Reg.	SSRankBoost	PNU-AUC	SAMULT	SAMPURA
australian	.879±.029	.860±.027	.886±.013	**.903±.009**	**.903±.009**	**.903±.009**
breast	.655±.097	.625±.095	.647±.065	**.701±.029**	**.701±.029**	**.704±.026**
breastw	.987±.009	.980±.006	.984±.013	.992±.001	**.996±.001**	**.996±.001**
clean1	.760±.062	.725±.060	.737±.050	.767±.042	.777±.039	**.782±.038**
colic	.829±.112	.818±.074	.721±.062	.858±.013	**.869±.013**	**.870±.013**
colic.orig	.647±.093	.645±.076	.612±.081	.644±.048	**.658±.049**	**.663±.044**
credit-a	.893±.024	.886±.023	.885±.013	**.906±.008**	**.906±.007**	**.906±.008**
credit-g	.719±.043	.709±.030	.665±.027	.748±.018	.748±.018	**.761±.017**
fourclass	.825±.023	.826±.026	.692±.029	**.827±.008**	**.827±.008**	**.828±.006**
german	.683±.057	.672±.048	.709±.025	**.727±.019**	**.727±.019**	**.729±.017**
haberman	.551±.086	.530±.075	**.582±.067**	.547±.051	.551±.045	.556±.043
heart	.857±.065	.842±.060	.823±.042	**.876±.025**	**.876±.025**	**.878±.024**
house	**.975±.038**	.961±.015	**.972±.034**	**.979±.012**	**.979±.012**	**.979±.011**
ijcnn1	**.912±.003**	.901±.004	.902±.002	.904±.009	**.913±.005**	**.915±.004**
madelon	.510±.037	.541±.020	**.571±.023**	.528±.029	.517±.027	.530±.022
parkinsons	.848±.129	.826±.082	.799±.051	**.860±.023**	**.860±.023**	**.863±.011**
phishing	.975±.097	.972±.001	.983±.003	.974±.002	.976±.002	**.985±.002**
vehicle	.932±.038	.922±.022	.912±.039	**.965±.020**	**.965±.020**	**.970±.014**
vote	.965±.038	.951±.015	**.972±.034**	**.979±.012**	**.979±.012**	**.979±.011**
wdbc	.971±.014	.963±.006	.964±.016	**.983±.006**	**.983±.006**	**.983±.005**
# Best/Comp.	2	0	4	11	15	18

tion method based on positive-unlabeled learning. Two simple baseline methods, supervised AUC optimization and logistic regression, are also compared, to show the benefit of utilizing unlabeled data and explicitly optimizing AUC.

All experiments are repeated 50 times with random data partition, and the average AUC scores as well as the standard deviations are recorded. The parameters are chosen by grid search through a 5-fold cross validation. The number of base learners in SAMPURA is fixed on 20. Since PNU-AUC also requires an estimation of the class prior probabilities to train the model, we feed the ground-truth class prior probabilities to PNU-AUC in this experiment.

Table 2 summarizes the experimental results on semi-supervised AUC optimization. For each dataset, the method with the best performance as well as the methods which are comparable to the best according to a pairwise t-test at the significance level 5%, are boldfaced.

It can be observed that SAMULT and SAMPURA achieve best performance compared to all the competing methods. SAMULT achieves best or comparable performance on 15 datasets and SAMPURA achieves on 18 out of 20 datasets, while SSRankBoost and PNU-AUC only achieves best or comparable on 4 and 11 datasets, respectively. Compared to supervised AUC optimization, SAMULT and SAMPURA improve the performance significantly on most of the datasets. The two proposed methods are easy to implement and use, since they do not require designing strategies to guess the possible labels of the unlabeled data or prior knowledge of the data.

Degeneration to AUC Optimization for Positive and Unlabeled Data

While only the positive and unlabeled data are available, SAMULT degenerates to a supervised AUC optimization style method by treating the unlabeled data as negative, as previously mentioned. In this subsection, we show that this simple method can obtain better performance than existing positive-unlabeled AUC optimization methods. We refer to this degenerated method as SAMULT$^{\mathrm{P+U}}$.

We compare SAMULT$^{\mathrm{P+U}}$ with two existing AUC optimization methods based on only positive and unlabeled data: PU-RSVM (Sellamanickam, Garg, and Selvaraj 2011), a ranking SVM based method for positive and unlabeled data, and PU-AUC (Sakai, Niu, and Sugiyama 2017), a positive-unlabeled AUC optimization method by optimizing an unbiased AUC risk estimator relies only on positive and unlabeled data.

We use the same experimental setting as the previous experiment. Since PU-AUC also requires an estimation of the class prior probabilities to train the model, we feed it with the ground-truth class prior probabilities instead of its estimation, which is expected to further push-up the performance of PU-AUC.

The experimental results are shown in Table 3. The best or comparable methods on each dataset is boldfaced, according to paired t-test at the significance level 5%.

It is shown that SAMULT$^{\mathrm{P+U}}$ achieves best performance on 17 out of 20 datasets, while PU-RSVM only achieves

Table 3: Experimental results (mean±std of the AUC) on positive-unlabeled datasets over 50 repetition. The boldfaces denote the best or comparable methods in terms of the AUC, according to the paired t-test at the significance level 5%. The numbers of best or comparable case of each method are shown in the last row.

Dataset	PU-RSVM	PU-AUC	SAMULT^{P+U}
australian	.844±.034	.898±.021	**.900±.019**
breast	.615±.104	**.701±.077**	**.701±.077**
breastw	.987±.009	.993±.002	**.996±.002**
clean1	.709±.072	.786±.050	**.796±.050**
colic	.807±.103	**.877±.060**	**.877±.060**
colic.orig	.621±.078	.650±.068	**.670±.061**
credit-a	.876±.028	**.912±.015**	**.912±.015**
credit-g	.688±.047	.755±.028	**.757±.027**
fourclass	.823±.031	**.832±.026**	**.832±.025**
german	.642±.045	.734±.034	**.736±.034**
haberman	**.572±.083**	.561±.081	.555±.080
heart	.835±.073	**.883±.039**	**.883±.040**
house	.945±.029	.980±.012	**.983±.011**
ijcnn1	**.927±.005**	.900±.011	.905±.012
madelon	.470±.015	**.533±.031**	.514±.032
parkinsons	.797±.089	**.870±.033**	**.870±.032**
phishing	.960±.008	.966±.005	**.970±.005**
vehicle	.942±.034	.959±.030	**.966±.025**
vote	.945±.029	.980±.012	**.983±.011**
wdbc	.967±.021	.984±.007	**.985±.007**
#Best/Com.	2	7	17

the best performance on 2 and PU-AUC on 7 datasets, respectively. Compared with PU-RSVM, SAMULT^{P+U} shows great improvement on almost all of the datasets. Compared with PU-SVM, although SAMULT^{P+U} performs almost the same or slightly better on many datasets, SAMULT^{P+U} hardly shows a worse performance. SAMULT^{P+U} does not bother to estimate the class prior probabilities while PU-AUC needs this extra step.

Such experimental results suggest that when optimizing AUC with only the positive and unlabeled data available, simply treating the unlabeled data as negative is enough to learn the model. The estimation of the class prior probabilities and the strategies to identify the possible labels of the unlabeled data could be unnecessary.

Related Work

Semi-Supervised Learning

Semi-supervised learning is highly demanded in many real-world applications, since it is often easy to gather plenty of unlabeled data but collecting labeled data could be expensive. Most of the semi-supervised learning algorithms can be roughly divided into four categories. Generative-model based methods assume that the data is generated by a latent distribution, and estimate the distribution with EM procedure (Shahshahani and Landgrebe 1994; Nigam et al. 2000; Fujino and Ueda 2016). Low density separation based methods try to find a separation hyperplane through a low-density area to separate the labeled and unlabeled data (Joachims 1999; Chapelle, Chi, and Zien 2006; Li, Kwok, and Zhou 2010). Graph-based methods estimate the possible labels of the unlabeled data by label propagation on the graph built from the data (Blum and Chawla 2001; Wang and Zhang 2008). Disagreement-based methods, such as co-training, exploit the disagreements among multiple learners during the learning process (Blum and Mitchell 1998; Goldman and Zhou 2000; Zhou and Li 2005; Li and Zhou 2007; Li, Li, and Zhou 2009).

AUC Optimization

AUC is a widely-used performance measure for classifiers, especially for problems that are highly imbalanced. Optimizing AUC is a common method to learn classifiers that rank the positive data before the negative data. Owing to the non-convexity and discontinuousness, many surrogate losses are proposed. Gao and Zhou (2015) studied the consistency of those surrogate losses theoretically. For online AUC optimization, Gao et al. (2013) proposed a method that maintains a covariance matrix to optimize AUC online, and Ying, Wen, and Lyu (2016) formalized the online AUC optimization problem as a stochastic saddle point problem to solve it with stochastic gradient based algorithm, which overcomes the challenge that AUC should be computed over instance pairs and thus storing all the data is needed.

For semi-supervised AUC optimization, Amini, Truong, and Goutte (2008) extended the RankBoost algorithm to semi-supervised case to learn a ranking function. Fujino and Ueda (2016) use a generative model based algorithm to utilize unlabeled data to optimize AUC. Sakai, Niu, and Sugiyama (2017) proposed a semi-supervised AUC optimization method by reweighting the unlabeled data, from a positive-unlabeled learning perspective.

Conclusion

Many semi-supervised methods rely on strong distributional assumptions or prior knowledge to guess the possible labels of the unlabeled data. In this paper, we argue that, in semi-supervised AUC optimization, it is unnecessary to design sophisticated strategies to estimate the possible labels of the unlabeled data or the class prior probabilities. We theoretically show that treating unlabeled data as both positive and negative data leads to an unbiased AUC risk estimation asymptotically in semi-supervised AUC optimization, based on which two semi-supervised AUC optimization methods, namely SAMULT and SAMPURA, is proposed. Experimental results indicate that the proposed method outperform the state-of-the-art semi-supervised AUC optimization methods. Additionally, we show that the positive-unlabeled AUC optimization problem can also be addressed by a degenerated version of our method that simply treats the unlabeled data as negative, without any distributional assumption or prior knowledge.

The current work is based on linear models on binary classification problem, extending this method to non-linear

and multi-class cases and conducting extensive evaluation w.r.t more performance measures will be done in future. Moreover, solving the proposed optimization problem using stochastic gradient descent may further scale up the proposed methods to extremely large data sets, which would be another interesting future work.

Acknowledgments

This research was supported by National Key Research and Development Program (2017YFB1001903) and NSFC (61422304, 61272217). The authors would like to thank Dr. Yu-Feng Li for his valuable comments on this work, and Mr. Hao Zhang for proof-reading this paper.

References

Amini, M. R.; Truong, T. V.; and Goutte, C. 2008. A Boosting Algorithm for Learning Bipartite Ranking Functions with Partially Labeled Data. In *Proceedings of the Thirty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 99–106. ACM.

Arampatzis, A.; Kamps, J.; and Robertson, S. 2009. Where to Stop Reading a Ranked List?: Threshold Optimization Using Truncated Score Distributions. In *Proceedings of the Thirty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval*, 524–531. ACM.

Blum, A., and Chawla, S. 2001. Learning from Labeled and Unlabeled Data Using Graph Mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 19–26. Morgan Kaufmann Publishers Inc.

Blum, A., and Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 92–100. ACM.

Chapelle, O.; Chi, M.; and Zien, A. 2006. A Continuation Method for Semi-Supervised SVMs. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, 185–192. ACM.

Chapelle, O.; Schlkopf, B.; and Zien, A. 2006. *Semi-Supervised Learning*. The MIT Press.

Cozman, F. G., and Cohen, I. 2002. Unlabeled Data Can Degrade Classification Performance of Generative Classifiers. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, 327–331. AAAI Press.

Fujino, A., and Ueda, N. 2016. A Semi-Supervised AUC Optimization Method with Generative Models. In *IEEE Sixteenth International Conference on Data Mining*, 883–888. IEEE.

Gao, W., and Zhou, Z.-H. 2015. On the Consistency of AUC Pairwise Optimization. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 939–945. AAAI Press.

Gao, W.; Jin, R.; Zhu, S.; and Zhou, Z.-H. 2013. One-Pass AUC Optimization. In *Proceedings of the Thirtieth International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 906–914. PMLR.

Goldman, S. A., and Zhou, Y. 2000. Enhancing Supervised Learning with Unlabeled Data. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 327–334. Morgan Kaufmann Publishers Inc.

Guyon, I.; Gunn, S.; Ben-Hur, A.; and Dror, G. 2005. Result Analysis of the NIPS 2003 Feature Selection Challenge. In *Advances in Neural Information Processing Systems 17*. The MIT Press. 545–552.

Hanley, J. A., and McNeil, B. J. 1982. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143(1):29–36.

Herschtal, A., and Raskutti, B. 2004. Optimising Area Under the ROC Curve Using Gradient Descent. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 49–56. ACM.

Joachims, T. 1999. Transductive Inference for Text Classification Using Support Vector Machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 200–209. Morgan Kaufmann Publishers Inc.

Li, M., and Zhou, Z.-H. 2007. Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* 37(6):1088–1098.

Li, Y.-F.; Kwok, J. T.; and Zhou, Z.-H. 2010. Cost-sensitive Semi-Supervised Support Vector Machine. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 500–505. AAAI Press.

Li, M.; Li, H.; and Zhou, Z.-H. 2009. Semi-Supervised Document Retrieval. *Information Processing and Management* 45(3):341–355.

Lichman, M. 2013. UCI Machine Learning Repository.

Lipton, Z. C.; Elkan, C.; and Naryanaswamy, B. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. In *Proceedings of the 2014 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, 225–239. Berlin, Heidelberg: Springer-Verlag.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning* 39(2-3):103–134.

Sakai, T.; Niu, G.; and Sugiyama, M. 2017. Semi-Supervised AUC Optimization Based on Positive-Unlabeled Learning. *Machine Learning*.

Sellamanickam, S.; Garg, P.; and Selvaraj, S. K. 2011. A Pairwise Ranking Based Approach to Learning with Positive and Unlabeled Examples. In *Proceedings of the Twentieth ACM International Conference on Information and Knowledge Management*, 663–672. ACM.

Shahshahani, B. M., and Landgrebe, D. A. 1994. The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon. *IEEE Transactions on Geoscience and Remote Sensing* 32(5):1087–1095.

Wang, F., and Zhang, C. 2008. Label Propagation Through Linear Neighborhoods. *IEEE Transactions on Knowledge and Data Engineering* 20(1):55–67.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic Online AUC Maximization. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 451–459.

Zhou, Z.-H., and Li, M. 2005. Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. *IEEE Transactions on Knowledge and Data Engineering* 17(11):1529–1541.

Zhu, X.; Goldberg, A. B.; Brachman, R.; and Dietterich, T. 2009. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers.