Semi-Supervised Learning for Maximizing the Partial AUC

Tomoharu Iwata, Akinori Fujino, Naonori Ueda
NTT Communication Science Laboratories, Kyoto, Japan
{tomoharu.iwata.gy, akinori.fujino.yh, naonori.ueda.fr}@hco.ntt.co.jp

Abstract

The partial area under a receiver operating characteristic curve (pAUC) is a performance measurement for binary classification problems that summarizes the true positive rate with the specific range of the false positive rate. Obtaining classifiers that achieve high pAUC is important in a wide variety of applications, such as cancer screening and spam filtering. Although many methods have been proposed for maximizing the pAUC, existing methods require many labeled data for training. In this paper, we propose a semi-supervised learning method for maximizing the pAUC, which trains a classifier with a small amount of labeled data and a large amount of unlabeled data. To exploit the unlabeled data, we derive two approximations of the pAUC: the first is calculated from positive and unlabeled data, and the second is calculated from negative and unlabeled data. A classifier is trained by maximizing the weighted sum of the two approximations of the pAUC and the pAUC that is calculated from positive and negative data. With experiments using various datasets, we demonstrate that the proposed method achieves higher test pAUCs than existing methods.

1 Introduction

The area under a receiver operating characteristic (ROC) curve (AUC) is widely used for performance measurement with binary classification problems (Bradley 1997; Huang and Ling 2005). The ROC curve is the true positive rate (TPR) as a function of the false positive rate (FPR). By maximizing the AUC, we can obtain a classifier that achieves high average TPRs over all FPR values from zero to one (Brefeld and Scheffer 2005; Wang and Tang 2009; Ding et al. 2015; Ying, Wen, and Lyu 2016; Han and Zhao 2010; Zhou, Lai, and Yen 2009; Zhao et al. 2011).

In many applications, we would like to achieve a high TPR with a specific FPR range. For example, in cancer screening applications, maintaining a low FPR is important if we are to eliminate unnecessary and costly biopsies (Baker and Pinsky 2001). In a spam detection system, we can accept only a low FPR if we are to prevent legitimate emails from being identified as spam. In such applications, a partial AUC (pAUC) is more appropriate than an AUC. The pAUC is the partial area under the ROC curve with a specific FPR range as shown in Figure 1(a). Many methods for maximizing the pAUC have been proposed (Komori and Eguchi 2010; Ricamato and Tortorella 2011; Narasimhan and Agarwal 2013a; 2013b; Ueda and Fujino 2018; Wang and Chang 2010). However, existing pAUC maximization methods require many labeled data for training, which are expensive to prepare.

In this paper, we propose a semi-supervised learning method for maximizing the pAUC to achieve a high pAUC with a small amount of labeled data and a large amount of unlabeled data. Unlabeled data are usually easier to prepare than labeled data. To exploit unlabeled data, we define the unlabeled positive rate (UPR), which is the probability that the decision function score of unlabeled data is higher than a threshold. We then derive two approximations of the pAUC from the partial area under the curves of the UPR and the TPR or FPR: the first is calculated from positive and unlabeled data, and the second is calculated from negative and unlabeled data.

A classifier is trained by maximizing the weighted sum of the two approximations of the pAUC and the pAUC that is calculated from positive and negative data. For classifiers, the proposed method can use any differentiable functions, such as logistic regression and neural networks. Although several semi-supervised methods for AUC maximization have been proposed (Fujino and Ueda 2016; Sakai, Niu, and Sugiyama 2018; Kiryo et al. 2017), they are inapplicable to pAUC maximization. The main contributions of this paper are as follows:

1. We derive two approximations of the pAUC that are calculated using unlabeled data (Section 4).
2. We propose a semi-supervised learning method for maximizing the pAUC based on the approximations of the pAUC (Section 5). Our work is the first attempt for semi-supervised pAUC maximization to our knowledge.
3. We empirically demonstrate that the proposed method performs better than existing supervised and semi-supervised methods using various datasets for anomaly detection (Section 6).

2 Preliminaries

Let $\mathbf{x} \in \mathbb{R}^D$ be a D-dimensional feature vector, and $y \in \{\pm 1\}$ be a binary class label. Decision function $s : \mathbb{R}^D \to \mathbb{R}$ outputs a score for classification, where the score is used to estimate the class label by, $\hat{y} = \mathrm{sign}(s(\mathbf{x}) - h)$, where h is a threshold and $\mathrm{sign}(\cdot)$ is the sign function; $\mathrm{sign}(A) = +1$ if $A \geq 0$ and $\mathrm{sign}(A) = -1$ otherwise. The true positive rate of threshold h is defined by the probability that the score of positive data is higher than threshold h, $\mathrm{TPR}(h) = \int_{-\infty}^{\infty} f_{\mathrm{P}}(s) I(s > h) ds$, where $f_{\mathrm{P}}(s)$ is the score distribution of positive data, and $I(\cdot)$ is the indicator function; $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. Similarly, the false positive rate of threshold h is defined by the probability that the score of negative data is higher than threshold h, $\mathrm{FPR}(h) = \int_{-\infty}^{\infty} f_{\mathrm{N}}(s) I(s > h) ds$, where $f_{\mathrm{N}}(s)$ is the score distribution of negative data. The AUC is the area under the curve of $\mathrm{TPR}(h)$ against $\mathrm{FPR}(h)$ with varying threshold h, and it is calculated by

$$\begin{aligned}\mathrm{AUC} &= \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(u)) du \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\mathrm{P}}(s) f_{\mathrm{N}}(s') I(s > s') ds ds', \qquad (1)\end{aligned}$$

where $\mathrm{FPR}^{-1}(u) = \inf\{h \in \mathbb{R} | \mathrm{FPR}(h) \leq u\}$ (Cortes and Mohri 2004). The AUC is the probability that scores sampled from the positive distribution are higher than those from the negative distribution.

The partial AUC (pAUC) between α and β, where $0 \leq \alpha < \beta \leq 1$, is the normalized partial area under the curve of TPR against FPR where the FPR is between α and β as follows,

$$\begin{aligned}&\mathrm{pAUC}(\alpha, \beta) = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} \mathrm{TPR}(\mathrm{FPR}^{-1}(u)) du \\ &= \frac{1}{\beta - \alpha} \int_{\mathrm{FPR}^{-1}(\beta)}^{\mathrm{FPR}^{-1}(\alpha)} \int_{-\infty}^{\infty} f_{\mathrm{P}}(s) f_{\mathrm{N}}(s') I(s > s') ds ds'. \\ &= \mathbb{E}[I(s(\mathbf{x}^{\mathrm{P}}) > s(\mathbf{x}^{\mathrm{N}}))], \qquad (2)\end{aligned}$$

where $\mathbb{E}[\cdot]$ represents the expectation, $\mathbf{x}^{\mathrm{P}}$ is a sample from the positive data distribution, and $\mathbf{x}^{\mathrm{N}}$ is a sample from the negative data distribution with the FPR between α and β. Figure 1(b) illustrates how the pAUC is calculated using positive and negative data.

Given a set of positive samples $\mathcal{P} = \{\mathbf{x}_m^{\mathrm{P}}\}_{m=1}^{M_{\mathrm{P}}}$ and a set of negative samples $\mathcal{N} = \{\mathbf{x}_m^{\mathrm{N}}\}_{m=1}^{M_{\mathrm{N}}}$, an empirical pAUC is calculated by

$$\begin{aligned}&\widehat{\mathrm{pAUC}}(\alpha, \beta) = \frac{1}{(\beta - \alpha) M_{\mathrm{P}} M_{\mathrm{N}}} \\ &\times \sum_{\mathbf{x}_m^{\mathrm{P}} \in \mathcal{P}} \Big[(j_\alpha - \alpha M_{\mathrm{N}}) I(s(\mathbf{x}_m^{\mathrm{P}}) > s(\mathbf{x}_{(j_\alpha)}^{\mathrm{N}})) \\ &+ \sum_{j = j_\alpha + 1}^{j_\beta} I(s(\mathbf{x}_m^{\mathrm{P}}) > s(\mathbf{x}_{(j)}^{\mathrm{N}})) \\ &+ (\beta M_{\mathrm{N}} - j_\beta) I(s(\mathbf{x}_m^{\mathrm{P}}) > s(\mathbf{x}_{(j_\beta + 1)}^{\mathrm{N}})) \Big], \qquad (3)\end{aligned}$$

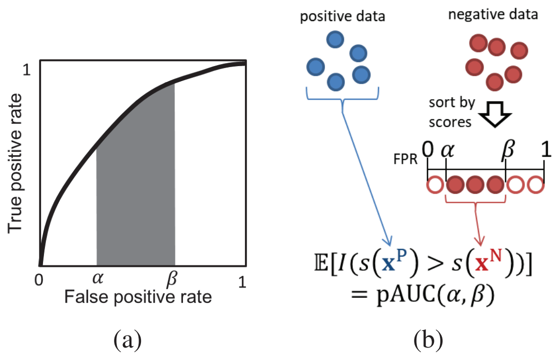

True positive rate
1
0
α
β
1
False positive rate
positive data
negative data
sort by
scores
FPR
0
α
β
1
$\mathbb{E}[I(s(\mathbf{x}^{\mathrm{P}}) > s(\mathbf{x}^{\mathrm{N}}))]$
$= \mathrm{pAUC}(\alpha, \beta)$
(a)
(b)

Figure 1: (a) $\mathrm{pAUC}(\alpha, \beta)$: Partial area under the curve of the true positive rate against the false positive rate between α and β. (b) $\mathrm{pAUC}(\alpha, \beta)$ is calculated by the probability that the score of a positive sample is higher than that of a negative sample that lies between α and β when sorted by its score.

where $j_\alpha = \lceil \alpha M_{\mathrm{N}} \rceil$, $j_\beta = \lfloor \beta M_{\mathrm{N}} \rfloor$, and $\mathbf{x}_{(j)}^{\mathrm{N}}$ denotes the negative sample in $\mathcal{N}$ ranked in the jth position among negatives in a descending order of scores $s(\mathbf{x})$ (Dodd and Pepe 2003; Narasimhan and Agarwal 2013a). The empirical pAUC is the empirical probability that positive samples have higher scores than negative samples that are ranked between j_α and j_β. The computational complexity of calculating Eq.(3) is $O(M_{\mathrm{N}} \log M_{\mathrm{N}} + (\beta - \alpha) M_{\mathrm{P}} M_{\mathrm{N}})$, where the first term is for sorting M_{N} negative samples, and the second term is for comparing M_{P} positive samples and $(\beta - \alpha) M_{\mathrm{N}}$ negative samples in the second term.

3 Problem formulation

Assume that we are given a set of positive samples $\mathcal{P} = \{\mathbf{x}_m^{\mathrm{P}}\}_{m=1}^{M_{\mathrm{P}}}$, a set of negative samples $\mathcal{N} = \{\mathbf{x}_m^{\mathrm{N}}\}_{m=1}^{M_{\mathrm{N}}}$, and a set of unlabeled samples $\mathcal{U} = \{\mathbf{x}_m^{\mathrm{U}}\}_{m=1}^{M_{\mathrm{U}}}$. We would like to obtain a decision function that has a high pAUC with given α and β for unseen samples.

We assume that the positive ratio of unlabeled samples θ_{P} is known. When labeled and unlabeled data are generated from the same distribution $p(\mathbf{x}, y)$, θ_{P} can be easily estimated by using the empirical positive probability with the labeled data. When labeled and unlabeled data are generated from different distributions, θ_{P} can be estimated by using methods described in (Saerens, Latinne, and Decaestecker 2002; Iyer, Nath, and Sarawagi 2014; Du Plessis and Sugiyama 2014).

4 Partial AUC with unlabeled data

In this section, we derive two approximations of the pAUC that are calculated using unlabeled data. Intuitively speaking, we sort unlabeled data by their scores, consider unlabeled data with specific ranges as positive or negative, and approximate the pAUC using the estimated positive or negative data. Figure 2 shows an overview of the calculation of the two approximated pAUCs that use positive, negative and unlabeled data, which is explained in detail in this sec-

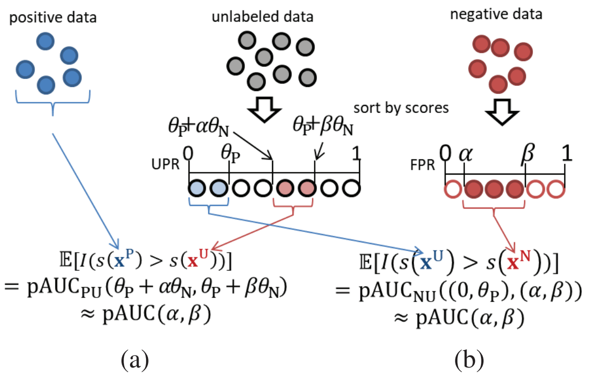

positive data
unlabeled data
negative data
sort by scores
$\theta_P+\alpha\theta_N$
$\theta_P+\beta\theta_N$
UPR
0
θ_P
1
FPR
0
α
β
1
$\mathbb{E}[I(s(\mathbf{x}^P) > s(\mathbf{x}^U))]$
$= \mathrm{pAUC}_{PU}(\theta_P + \alpha\theta_N, \theta_P + \beta\theta_N)$
$\approx \mathrm{pAUC}(\alpha, \beta)$
$\mathbb{E}[I(s(\mathbf{x}^U) > s(\mathbf{x}^N))]$
$= \mathrm{pAUC}_{NU}((0, \theta_P), (\alpha, \beta))$
$\approx \mathrm{pAUC}(\alpha, \beta)$
(a)
(b)

Figure 2: (a) Approximated pAUC with positive and unlabeled data is calculated by the probability that the score of a positive sample is higher than that of an unlabeled sample between $\theta_P + \alpha\theta_N$ and $\theta_P + \beta\theta_N$ when sorted by its score, where θ_P is the positive probability of unlabeled data, $\theta_N = 1 - \theta_P$ is the negative probability of unlabeled data. (b) Approximated pAUC with negative and unlabeled data is calculated by the probability that the score of an unlabeled sample between 0 and θ_P is higher than that of a negative sample between α and β.

tion as follows: We first define the unlabeled positive rate (UPR) and the classifier with the threshold using the UPR in Section 4.1. Then, in Section 4.2, we derive an approximated pAUC that is calculated using positive and unlabeled data by considering the partial area under the curve of the TPR against the UPR. Similarly, in Section 4.3, we derive another approximated pAUC that is calculated using negative and unlabeled data by considering the partial area under the curve of the UPR against the FPR.

4.1 Unlabeled positive rate

We newly define unlabeled positive rate of threshold h, $\mathrm{UPR}(h)$, as the probability that the score of unlabeled data is higher than threshold h in a similar way to the TRP and FPR as follows,

$$\mathrm{UPR}(h) = \int_{-\infty}^{\infty} f_U(s)I(s > h)ds, \tag{4}$$

where $f_U(s) = \theta_P f_P(s) + \theta_N f_N(s)$ is the score distribution of unlabeled data, θ_P is the positive probability of unlabeled data, $\theta_N = 1 - \theta_P$ is the negative probability of unlabeled data, and $0 \leq \theta_P, \theta_N \leq 1$.

Given score s, we estimate the class label of unlabeled data with threshold $h = \mathrm{UPR}^{-1}(\theta_P)$ as follows,

$$\hat{y}(s) = \begin{cases} +1 & s > \mathrm{UPR}^{-1}(\theta_P), \\ -1 & \text{otherwise}, \end{cases} \tag{5}$$

where $\mathrm{UPR}^{-1}(\theta) = \inf\{h \in \mathbb{R} | \mathrm{UPR}(h) \leq \theta\}$, and $\mathrm{UPR}^{-1}(\theta_P)$ is the threshold when the UPR is θ_P. With this threshold, the probability that unlabeled data are classified as positive becomes θ_P from the definition of the UPR in Eq.(4) as follows, $\int_{-\infty}^{\infty} f_U(s)I(\hat{y}(s) = 1)ds = \int_{-\infty}^{\infty} f_U(s)I(s > \mathrm{UPR}^{-1}(\theta_P))ds = \theta_P$. When the classifier in Eq.(5) is assumed, unlabeled data with scores higher than the threshold, $s > \mathrm{UPR}^{-1}(\theta_P)$, are classified as positive, and unlabeled data with scores lower than the threshold $s \leq \mathrm{UPR}^{-1}(\theta_P)$, are classified as negative. Therefore, with (5), we obtain estimates of the positive and negative score distributions, $\hat{f}_P(s)$ and $\hat{f}_N(s)$, by the split of the score distribution of the unlabeled data $f_U(s)$ at $s = \mathrm{UPR}^{-1}(\theta_P)$ as follows,

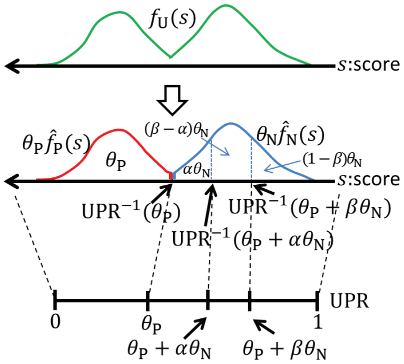

$f_U(s)$
s:score
$\theta_P\hat{f}_P(s)$
θ_P
$(\beta-\alpha)\theta_N$
$\alpha\theta_N$
$\theta_N\hat{f}_N(s)$
$(1-\beta)\theta_N$
s:score
$\mathrm{UPR}^{-1}(\theta_P)$
$\mathrm{UPR}^{-1}(\theta_P + \beta\theta_N)$
$\mathrm{UPR}^{-1}(\theta_P + \alpha\theta_N)$
UPR
0
θ_P
1
$\theta_P + \alpha\theta_N$
$\theta_P + \beta\theta_N$

Figure 3: Top: Unlabeled score distribution $f_U(s)$. Middle: By splitting the unlabeled score distribution at threshold $h = \mathrm{UPR}^{-1}(\theta_P)$, we obtain the estimated positive score distribution, $\hat{f}_P(s)$ (red), and the estimated negative score distribution, $\hat{f}_N(s)$ (blue). Bottom: Unlabeled positive rate.

$$f_U(s) \approx \begin{cases} \theta_P \hat{f}_P(s) & \text{if } s > \mathrm{UPR}^{-1}(\theta_P), \\ \theta_N \hat{f}_N(s) & \text{otherwise}, \end{cases} \tag{6}$$

which is illustrated in the top and middle of Figure 3.

Since the area of the unlabeled score distribution $f_U(s)$ between $\mathrm{UPR}^{-1}(\theta_P)$ and $\mathrm{UPR}^{-1}(\theta_P + \alpha\theta_N)$ is $\alpha\theta_N$ as in the middle of Figure 3, the area of the estimated negative score distribution $\hat{f}_N(s) \approx \frac{f_U(s)}{\theta_N}$ between $\mathrm{UPR}^{-1}(\theta_P)$ and $\mathrm{UPR}^{-1}(\theta_P + \alpha\theta_N)$ is α. In addition, there are no estimated negative data $\hat{f}_N(s) = 0$ if $s > \mathrm{UPR}^{-1}(\theta_P)$ from Eq.(6). Therefore, the score when the UPR is $\theta_P + \alpha\theta_N$ is the same as the score when the estimated FPR is α as follows,

$$\mathrm{UPR}^{-1}(\theta_P + \alpha\theta_N) = \widehat{\mathrm{FPR}}^{-1}(\alpha), \tag{7}$$

where $\widehat{\mathrm{FPR}}$ is the false positive rate with an estimated negative score distribution $\hat{f}_N(s)$. Similarly, the score when the UPR is $\theta_P + \beta\theta_N$ is the same as the score when the FPR is β as follows,

$$\mathrm{UPR}^{-1}(\theta_P + \beta\theta_N) = \widehat{\mathrm{FPR}}^{-1}(\beta). \tag{8}$$

The relationship between the estimated negative score distribution and the UPR is illustrated in the middle and bottom of Figure 3. With Eqs.(7) and (8), the unlabeled data between $\theta_P + \alpha\theta_N$ and $\theta_P + \beta\theta_N$ are assumed to be negative data between α and β.

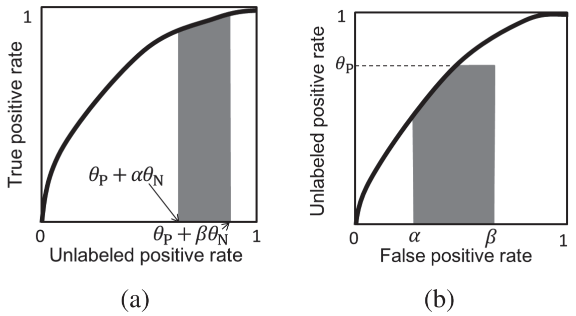


Figure 4: (a) $\mathrm{pAUC}_{\mathrm{PU}}(\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}, \theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}})$: Partial area under the curve of the true positive rate against the unlabeled positive rate between $\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}$ and $\theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}}$. (b) $\mathrm{pAUC}_{\mathrm{NU}}((0, \theta_{\mathrm{P}}), (\alpha, \beta))$: Partial area under the curve of the unlabeled positive rate against the false positive rate between α and β of the false positive rate and between 0 and θ_{P} of the unlabeled positive rate.

4.2 Partial AUC with positive and unlabeled data

We define $\mathrm{pAUC}_{\mathrm{PU}}(\alpha', \beta')$, which is the normalized partial area under the curve of $\mathrm{TPR}(h)$ against $\mathrm{UPR}(h)$ with varying threshold h where $\mathrm{UPR}(h)$ is between α' and β', as follows,

$$\mathrm{pAUC}_{\mathrm{PU}}(\alpha', \beta') = \frac{1}{\beta' - \alpha'} \times \int_{\mathrm{UPR}^{-1}(\beta')}^{\mathrm{UPR}^{-1}(\alpha')} \int_{-\infty}^{\infty} f_{\mathrm{P}}(s) f_{\mathrm{U}}(s') I(s > s') ds ds', \quad (9)$$

which is calculated with positive and unlabeled distributions. Then, the partial area between $\alpha' = \theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}$ and $\beta' = \theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}}$ becomes an approximation of the pAUC as follows,

$$\begin{aligned} &\mathrm{pAUC}_{\mathrm{PU}}(\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}, \theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}}) = \frac{1}{(\beta - \alpha)\theta_{\mathrm{N}}} \\ &\times \int_{\mathrm{UPR}^{-1}(\theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}})}^{\mathrm{UPR}^{-1}(\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}})} \int_{-\infty}^{\infty} f_{\mathrm{P}}(s) f_{\mathrm{U}}(s') I(s > s') ds ds' \\ &\approx \int_{\widehat{\mathrm{FPR}}^{-1}(\beta)}^{\widehat{\mathrm{FPR}}^{-1}(\alpha)} \int_{-\infty}^{\infty} f_{\mathrm{P}}(s) \theta_{\mathrm{N}} \hat{f}_{\mathrm{N}}(s') I(s > s') ds ds' \\ &\times \frac{1}{(\beta - \alpha)\theta_{\mathrm{N}}} \approx \mathrm{pAUC}(\alpha, \beta), \end{aligned} \quad (10)$$

where we used Eqs.(7), (8) and $f_{\mathrm{U}}(s) \approx \theta_{\mathrm{N}} \hat{f}_{\mathrm{N}}(s)$ if $s \leq \mathrm{UPR}^{-1}(\theta_{\mathrm{P}})$ in Eq.(6). An example of $\mathrm{pAUC}_{\mathrm{PU}}(\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}, \theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}})$ is shown in Figure 4(a). $\mathrm{pAUC}_{\mathrm{PU}}(\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}, \theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}})$ is the probability that scores sampled from positive data are higher than those of unlabeled data ranked between the $\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}$ and $\theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}}$ ratios with a descending order of scores.

4.3 Partial AUC with negative and unlabeled data

We define $\mathrm{pAUC}_{\mathrm{NU}}((\gamma, \eta), (\alpha', \beta'))$, which is the normalized partial area under the curve of $\mathrm{UPR}(h)$ against $\mathrm{FPR}(h)$ for different values of h where $\mathrm{FPR}(h)$ is between α' and β' and $\mathrm{UPR}(h)$ is between γ and η, as follows,

$$\begin{aligned} &\mathrm{pAUC}_{\mathrm{NU}}((\gamma, \eta), (\alpha', \beta')) = \frac{1}{(\beta' - \alpha')(\eta - \gamma)} \\ &\times \int_{\mathrm{FPR}^{-1}(\beta')}^{\mathrm{FPR}^{-1}(\alpha')} \int_{\mathrm{UPR}^{-1}(\eta)}^{\mathrm{UPR}^{-1}(\gamma)} f_{\mathrm{U}}(s) f_{\mathrm{N}}(s') I(s > s') ds ds', \end{aligned} \quad (11)$$

which is calculated with negative and unlabeled distributions. Then, the partial area between $(\gamma, \eta) = (0, \theta_{\mathrm{P}})$ and $(\alpha', \beta') = (\alpha, \beta)$ becomes an approximation of the pAUC as follows,

$$\begin{aligned} &\mathrm{pAUC}_{\mathrm{NU}}((0, \theta_{\mathrm{P}}), (\alpha, \beta)) = \frac{1}{\theta_{\mathrm{P}}(\beta - \alpha)} \\ &\times \int_{\mathrm{FPR}^{-1}(\beta)}^{\mathrm{FPR}^{-1}(\alpha)} \int_{\mathrm{UPR}^{-1}(\theta_{\mathrm{P}})}^{\infty} f_{\mathrm{U}}(s) f_{\mathrm{N}}(s') I(s > s') ds ds' \\ &\approx \int_{\mathrm{FPR}^{-1}(\beta)}^{\mathrm{FPR}^{-1}(\alpha)} \int_{\mathrm{UPR}^{-1}(\theta_{\mathrm{P}})}^{\infty} \theta_{\mathrm{P}} \hat{f}_{\mathrm{P}}(s) f_{\mathrm{N}}(s') I(s > s') ds ds' \\ &\times \frac{1}{\theta_{\mathrm{P}}(\beta - \alpha)} \approx \mathrm{pAUC}(\alpha, \beta), \end{aligned} \quad (12)$$

where we used $f_{\mathrm{U}}(s) = \theta_{\mathrm{P}} \hat{f}_{\mathrm{P}}(s)$ if $s > \mathrm{UPR}^{-1}(\theta_{\mathrm{P}})$ in Eq.(6), and the start of the integration interval of variable s, $\mathrm{UPR}^{-1}(\theta_{\mathrm{P}})$, can be $-\infty$ since $\hat{f}_{\mathrm{P}}(s) = 0$ if $s \leq \mathrm{UPR}^{-1}(\theta_{\mathrm{P}})$. An example of $\mathrm{pAUC}_{\mathrm{NU}}((0, \theta_{\mathrm{P}}), (\alpha, \beta))$ is shown in Figure 4(b). $\mathrm{pAUC}_{\mathrm{NU}}((0, \theta_{\mathrm{P}}), (\alpha, \beta))$ is the probability that scores of unlabeled data ranked between the 0 and θ_{P} ratios are higher than those from negative data ranked between the α and β ratios.

The pAUC can be also approximated using only unlabeled data by assuming that unlabeled samples between 0 and θ_{P} are positive and those between $\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}$ and $\theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}}$ are negative when sorted by their scores. However, this approximated pAUC is not useful for training a classifier since their labels are estimated by the classifier to be trained. On the other hand, the approximated pAUCs with positive/negative and unlabeled data used in the proposed method are based on true positive/negative data.

5 Partial AUC maximization using unlabeled data

The empirical estimate of the partial area under the curve of the TPR against the UPR in Eq.(10) is given by

$$\begin{aligned} &\widehat{\mathrm{pAUC}}_{\mathrm{PU}}(\theta_{\mathrm{P}} + \alpha\theta_{\mathrm{N}}, \theta_{\mathrm{P}} + \beta\theta_{\mathrm{N}}) = \frac{1}{(\beta - \alpha)\theta_{\mathrm{N}} M_{\mathrm{P}} M_{\mathrm{U}}} \\ &\times \sum_{\mathbf{x}_m^{\mathrm{P}} \in \mathcal{P}} \Big[(k_{\bar{\alpha}} - \bar{\alpha} M_{\mathrm{U}}) I(s(\mathbf{x}_m^{\mathrm{P}}) > s(\mathbf{x}_{(k_{\bar{\alpha}})}^{\mathrm{U}})) \\ &+ \sum_{k = k_{\bar{\alpha}} + 1}^{k_{\bar{\beta}}} I(s(\mathbf{x}_m^{\mathrm{P}}) > s(\mathbf{x}_{(k)}^{\mathrm{U}})) \\ &+ (\bar{\beta} M_{\mathrm{U}} - k_{\bar{\beta}}) I(s(\mathbf{x}_m^{\mathrm{P}}) > s(\mathbf{x}_{(k_{\bar{\beta}} + 1)}^{\mathrm{U}})) \Big], \end{aligned} \quad (13)$$

where $\bar{\alpha} = \theta_{\rm P} + \alpha\theta_{\rm N}$, $\bar{\beta} = \theta_{\rm P} + \beta\theta_{\rm N}$, $k_{\bar{\alpha}} = \lceil \bar{\alpha} M_{\rm U} \rceil$, $k_{\bar{\beta}} = \lfloor \bar{\beta} M_{\rm U} \rfloor$, and $\mathbf{x}^{\rm U}_{(k)}$ denotes the unlabeled sample ranked in the kth position among unlabeled samples $\mathcal{U}$ in a descending order of scores. $\widehat{\rm pAUC}_{\rm PU}(\theta_{\rm P} + \alpha\theta_{\rm N}, \theta_{\rm P} + \beta\theta_{\rm N})$ is calculated from positive and unlabeled data. The computational complexity of calculating Eq.(13) is $O(M_{\rm U} \log M_{\rm U} + (\beta - \alpha)\theta_{\rm N} M_{\rm P} M_{\rm U})$, where the first term is for sorting $M_{\rm U}$ unlabeled samples, and the second term is for comparing $M_{\rm P}$ positive samples and $(\beta - \alpha)\theta_{\rm N} M_{\rm U}$ unlabeled samples in the second term.

The empirical estimate of the partial area under the curve of the UPR against the FPR in Eq.(12) is given by

$$
\begin{aligned}
&\widehat{\rm pAUC}_{\rm NU}((0, \theta_{\rm P}), (\alpha, \beta)) = \frac{1}{(\beta - \alpha)\theta_{\rm P} M_{\rm U} M_{\rm N}} \\
&\times \Bigg[(j_\alpha - \alpha M_{\rm N}) \sum_{k=1}^{k_{\theta_{\rm P}}} I(s(\mathbf{x}^{\rm U}_{(k)}) > s(\mathbf{x}^{\rm N}_{(j_\alpha)})) \\
&+ \sum_{k=1}^{k_{\theta_{\rm P}}} \sum_{j=j_\alpha+1}^{j_\beta} I(s(\mathbf{x}^{\rm U}_{(k)}) > s(\mathbf{x}^{\rm N}_{(j)})) \\
&+ (\beta M_{\rm N} - j_\beta) \sum_{k=1}^{k_{\theta_{\rm P}}} I(s(\mathbf{x}^{\rm U}_{(k)}) > s(\mathbf{x}^{\rm N}_{(j_\beta+1)})) \\
&+ (\theta_{\rm P} M_{\rm U} - k_{\theta_{\rm P}}) \sum_{j=j_\alpha+1}^{j_\beta} I(s(\mathbf{x}^{\rm U}_{(k_{\theta_{\rm P}}+1)}) > s(\mathbf{x}^{\rm N}_{(j)})) \\
&+ (\theta_{\rm P} M_{\rm U} - k_{\theta_{\rm P}})(\beta M_{\rm N} - j_\beta) I(s(\mathbf{x}^{\rm U}_{(k_{\theta_{\rm P}}+1)}) > s(\mathbf{x}^{\rm N}_{(j_\beta+1)})) \Bigg],
\end{aligned} \tag{14}
$$

where $k_{\theta_{\rm P}} = \lfloor \theta_{\rm P} M_{\rm U} \rfloor$. $\widehat{\rm pAUC}_{\rm NU}((0, \theta_{\rm P}), (\alpha, \beta))$ is calculated from negative and unlabeled data. Figure 5 illustrates terms in Eq.(14). The computational complexity of calculating Eq.(14) is $O(M_{\rm U} \log M_{\rm U} + M_{\rm N} \log M_{\rm N} + (\beta - \alpha)\theta_{\rm N} M_{\rm U} M_{\rm N})$, where the first and second terms are for sorting $M_{\rm U}$ unlabeled and $M_{\rm N}$ negative samples, respectively, and the third term is for comparing $\theta_{\rm P} M_{\rm U}$ unlabeled samples and $(\beta - \alpha) M_{\rm N}$ negative samples at the second term.

We learn the decision function by maximizing the empirical pAUCs. For the decision function, we can use any differentiable functions, such as logistic regression models and neural networks. To make the empirical pAUCs differentiable, we used sigmoid function $\sigma(A) = \frac{1}{1+\exp(-A)}$ instead of indicator function $I(A)$ in empirical pAUCs, which is often used for a smooth approximation of the indicator function. For example, the second term in Eq.(3) is approximated by $I(s(\mathbf{x}^{\rm P}_m) > s(\mathbf{x}^{\rm N})) \approx \sigma(s(\mathbf{x}^{\rm P}_m) - s(\mathbf{x}^{\rm N}))$. Let $\widetilde{\rm pAUC}(\alpha, \beta)$ indicate the approximation of $\widehat{\rm pAUC}(\alpha, \beta)$ smoothed by the sigmoid function. Our objective function to be maximized is

$$
\begin{aligned}
L = &\lambda_1 \widetilde{\rm pAUC}(\alpha, \beta) + \lambda_2 \widetilde{\rm pAUC}_{\rm PU}(\theta_{\rm P} + \alpha\theta_{\rm N}, \theta_{\rm P} + \beta\theta_{\rm N}) \\
&+ \lambda_3 \widetilde{\rm pAUC}_{\rm NU}((0, \theta_{\rm P}), (\alpha, \beta)),
\end{aligned} \tag{15}
$$

where the first term is the smoothed empirical pAUC with positive and negative data in Eq.(3), the second term is the smoothed empirical pAUC with positive and unlabeled data in Eq.(13), the third term is the smoothed empirical pAUC with negative and unlabeled data in Eq.(14), and $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters, $\lambda_1, \lambda_2, \lambda_3 \geq 0$, $\sum_{i=1}^{3} \lambda_i = 1$. The hyperparameters can be tuned with validation data. When $\lambda_1 = 1, \lambda_2 = \lambda_3 = 0$, the proposed method corresponds to a supervised pAUC maximization method. We can robustly approximate the partial AUC by using unlabeled data as well as labeled data. The proposed method is applicable even when only positive (negative) and unlabeled data are available since the second (third) term requires only positive (negative) and unlabeled data. In this sense, the proposed method is related to methods for learning from positive and unlabeled data (Du Plessis, Niu, and Sugiyama 2015; Lee and Liu 2003; Li and Liu 2003; Elkan and Noto 2008) although they are not for pAUC maximization.

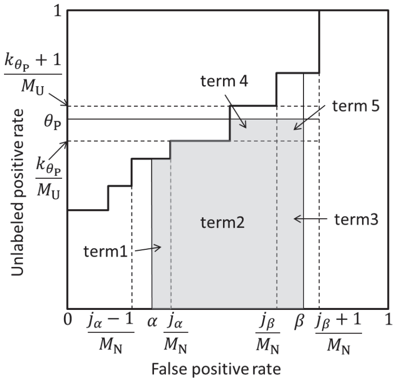

1
$\frac{k_{\theta_{\rm P}}+1}{M_{\rm U}}$
$\theta_{\rm P}$
$\frac{k_{\theta_{\rm P}}}{M_{\rm U}}$
Unlabeled positive rate
term 4
term 5
term2
term3
term1
0
$\frac{j_\alpha - 1}{M_{\rm N}}$
α
$\frac{j_\alpha}{M_{\rm N}}$
$\frac{j_\beta}{M_{\rm N}}$
β
$\frac{j_\beta + 1}{M_{\rm N}}$
1
False positive rate

Figure 5: The empirical partial area under the curve of the unlabeled positive rate against the false positive rate, $\widehat{\rm pAUC}_{\rm NU}((0, \theta_{\rm P}), (\alpha, \beta))$, in Eq.(14).

6 Experiments

6.1 Data

We evaluated the effectiveness of the proposed method by using the following nine datasets for anomaly detection (Campos et al. 2016) [1]: Annthyroid, Cardiotocography, InternetAds, KDDCup99, PageBlocks, Pima, SpamBase, Waveform and Wilt, where the feature vector dimensionalities were 21, 21, 1555, 79, 10, 8, 57, 21 and 5, respectively. For each dataset, we used 50 labeled and 300 unlabeled samples for training, 50 labeled samples for validation, and the remaining samples for testing, where the positive ratio was set at 0.1. For each dataset, we randomly sampled 30 training, validation and test data sets, and calculated the average pAUC over the 30 sets.

6.2 Comparing methods

We compared the proposed semi-supervised learning method for maximizing pAUC with the following seven methods: CE, MA, MPA, ST, SS, SSR, pSS and pSSR.

[1] The datasets were obtained from http://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/.

Table 1: Test pAUCs with (a) $\alpha = 0, \beta = 0.1$, (b) $\alpha = 0, \beta = 0.3$ and (c) $\alpha = 0.1, \beta = 0.2$. Values in bold typeface are not statistically different (at 5% level) from the best performing method in each row according to a paired t-test. The bottom row shows the average test pAUC over all datasets, and values in bold typeface indicate the method that achieved the best.

(a) $\text{pAUC}(0.0, 0.1)$

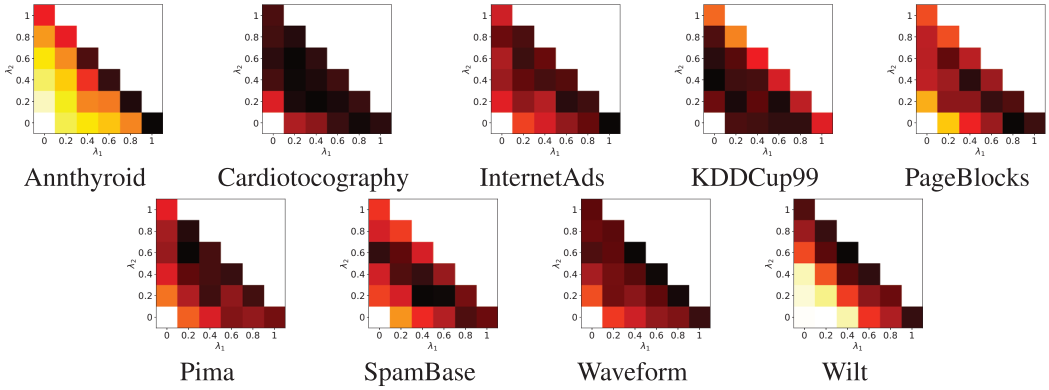
	CE	MA	MPA	ST	SS	SSR	pSS	pSSR	Ours
Annthyroid	0.227	0.236	0.384	0.357	**0.399**	**0.422**	0.258	**0.457**	**0.388**
Cardiotocography	0.464	0.473	0.493	0.494	0.420	0.450	0.467	0.393	**0.527**
InternetAds	0.540	0.570	0.565	**0.632**	0.496	0.464	0.527	0.446	0.580
KDDCup99	**0.880**	**0.868**	**0.874**	**0.890**	0.837	0.832	0.867	0.802	**0.884**
PageBlocks	0.528	0.518	**0.593**	**0.596**	**0.599**	**0.599**	0.553	0.568	**0.598**
Pima	0.057	0.118	**0.188**	**0.189**	**0.179**	0.130	0.127	0.118	**0.206**
SpamBase	0.408	0.438	**0.461**	**0.469**	0.422	0.393	0.435	0.416	**0.484**
Waveform	**0.270**	0.253	**0.288**	**0.283**	0.268	**0.281**	**0.305**	0.226	**0.306**
Wilt	0.100	0.195	0.594	0.549	**0.648**	0.403	0.260	**0.703**	**0.681**
Average	0.386	0.408	0.493	0.496	0.474	0.442	0.422	0.459	**0.517**

(b) $\text{pAUC}(0.0, 0.3)$

	CE	MA	MPA	ST	SS	SSR	pSS	pSSR	Ours
Annthyroid	0.442	0.436	**0.517**	0.494	**0.516**	0.445	0.428	**0.506**	**0.503**
Cardiotocography	0.680	0.705	0.698	0.701	0.661	0.665	0.686	0.637	**0.725**
InternetAds	0.664	0.697	0.695	**0.723**	0.629	0.631	0.621	0.590	0.672
KDDCup99	0.949	0.941	0.944	**0.956**	0.929	0.914	0.943	0.904	**0.961**
PageBlocks	0.679	0.677	0.717	0.724	**0.746**	**0.744**	**0.729**	**0.753**	**0.727**
Pima	0.255	0.324	**0.387**	**0.383**	**0.384**	**0.364**	0.327	**0.346**	**0.355**
SpamBase	**0.698**	**0.690**	**0.691**	**0.691**	0.663	0.627	0.662	0.617	**0.687**
Waveform	**0.624**	**0.619**	0.598	**0.628**	0.571	0.548	**0.595**	0.500	**0.609**
Wilt	0.326	0.440	0.813	0.780	0.803	0.687	0.539	0.790	**0.845**
Average	0.591	0.614	0.673	0.675	0.656	0.625	0.614	0.627	**0.676**

(c) $\text{pAUC}(0.1, 0.2)$

	CE	MA	MPA	ST	SS	SSR	pSS	pSSR	Ours
Annthyroid	0.480	0.469	**0.526**	**0.512**	**0.537**	0.459	0.454	0.456	**0.510**
Cardiotocography	**0.729**	**0.750**	**0.752**	**0.760**	0.697	0.685	**0.746**	0.601	**0.761**
InternetAds	0.697	**0.734**	**0.729**	**0.734**	0.611	0.637	0.663	0.558	**0.724**
KDDCup99	0.982	0.977	0.982	**0.986**	0.967	0.956	0.973	0.963	**0.988**
PageBlocks	0.713	0.718	0.751	0.740	**0.784**	**0.782**	**0.776**	0.708	**0.763**
Pima	0.294	0.353	0.388	0.404	**0.425**	0.404	0.376	0.337	**0.447**
SpamBase	**0.764**	**0.760**	**0.775**	**0.774**	0.713	0.688	0.727	0.623	**0.768**
Waveform	**0.708**	**0.695**	0.626	0.634	0.536	0.594	**0.683**	0.522	0.654
Wilt	0.341	0.462	0.700	0.691	**0.854**	0.714	0.567	**0.858**	**0.865**
Average	0.634	0.658	0.692	0.693	0.681	0.658	0.663	0.625	**0.720**



Annthyroid Cardiotocography InternetAds KDDCup99 PageBlocks

Pima SpamBase Waveform Wilt

Figure 6: Test pAUCs with $\alpha = 0, \beta = 0.1$ obtained by the proposed method with different hyperparameters λ_1, λ_2, $\lambda_3 = 1 - \lambda_1 - \lambda_2$. X-axis is hyperparameter λ_1, y-axis is λ_2, and a darker color indicates a better test pAUC.

The CE, MA and MPA methods are supervised learning methods. The CE method learns neural network parameters by minimizing the cross entropy loss. The MA and MPA methods learn parameters by maximizing the smoothed empirical AUC and pAUC, respectively.

The ST method is a self-training based semi-supervised learning method for maximizing the pAUC. With the ST method, at each step, the neural network is (re)trained by maximizing the smoothed empirical pAUC using labeled data, and then some unlabeled samples are added to the labeled data by assigning labels using the neural network, where unlabeled samples with high positive/negative probabilities are selected to be added.

The SS and SSR methods are semi-supervised learning methods for maximizing the AUC, where the SS method does not need to estimate the positive ratio (Xie and Li 2018), and the SSR method is based on positive-unlabeled learning (Sakai, Niu, and Sugiyama 2018). The pSS and pSSR methods are semi-supervised learning methods for maximizing the pAUCs, where the pAUC is calculated by $\mathrm{pAUC}(\alpha, \beta) = \mathrm{AUC} - \mathrm{pAUC}(0, \alpha) - \mathrm{pAUC}(\beta, 1)$. In particular, the pSS and pSSR methods calculate the AUC in the same way as the SS and SSR methods using labeled and unlabeled samples, respectively, and derive the $\mathrm{pAUC}(\alpha, \beta)$ by subtracting $\mathrm{pAUC}(0, \alpha)$ and $\mathrm{pAUC}(\beta, 1)$, which are estimated using labeled samples, from the AUC. With the SSR and pSSR methods, we used non-negative estimators (Kiryo et al. 2017).

For decision functions $s(\mathbf{x})$ with all the methods including our proposed method, we used the same neural network architecture, which was a three-layer feed-forward neural network with 32 hidden units and rectified linear units (ReLU) for the activation functions. We optimized the neural network parameters using ADAM (Kingma and Ba 2015) with a learning rate of 0.1 and a batch size of 1,024. The empirical AUC and pAUC were calculated using samples in each batch for training. The weight decay parameter was set at 10^{-3}. The hyperparameters of the proposed, SS, SSR, pSS and pSSR methods were selected from 0,0.2,0.4,0.6,0.8 and 1 using the validation pAUC. With the ST method, the number of unlabeled samples to be labeled for each step is tuned from $\{5, 10, 15, 20, 25\}$ using the validation pAUC, where the number of epochs for each retraining step was 100. The validation pAUC was also used for early stopping with all methods, where the maximum number of training epochs was 3,000. We implemented all the methods based on PyTorch (Paszke et al. 2017).

6.3 Results

Table 1 shows test pAUCs with three different pairs of α and β. The proposed method achieved the performance that was not statistically different from the best performing method in most cases. The test pAUC with MPA was higher than that with the other supervised methods, i.e., CE and MA, since MPA directly optimized the pAUC. MA was better than CE because the AUC is more closely related to the pAUC than the cross entropy loss. On average, the performance of ST was slightly better than MPA, but it was worse than the proposed method especially when (α, β) were $(0, 0.1)$ and $(0.1, 0.2)$. ST decisively assigns partial unlabeled data with a high confidence score to positive or negative, which changes the negative score distribution and makes it difficult to identify negative samples where the false positive rate is between α and β. Besides, ST treats assigned unlabeled data and labeled data equally. On the other hand, the proposed method provisionally assigns all unlabeled data to positive or negative, and treats assigned unlabeled data and labeled data unequally. These differences resulted in the better performance of the proposed method than ST. By using unlabeled data, the semi-supervised methods that maximize the AUC, i.e., SS and SSR, often performed better than the supervised MA method. However, they performed worse than the proposed method since they maximize the AUC but not the pAUC. The semi-supervised methods that maximize the pAUC, i.e., pSS and pSSR, sometimes performed worse than MPA, SS and pSSR. This would be because the pAUC estimated by pSS and pSSR, which subtracted the estimated pAUCs from the estimated AUC, was not accurate. On the other hand, the proposed method achieved a robust pAUC estimation by sorting unlabeled data with their scores. The average computational time for training with the proposed method on the InternetAds dataset, which had the largest feature vector dimensionality and took the longest time among the nine datasets, was 29.3 seconds on computers with 2.60GHz CPUs.

Figure 6 shows the test pAUC with $\alpha = 0, \beta = 0.1$ when using the proposed method with different hyperparameters λ_1, λ_2, $\lambda_3 = 1 - \lambda_1 - \lambda_2$. The best hyperparameter setting differed across datasets. The proposed method achieved good performance by selecting the hyperparameters using validation data.

7 Conclusion

In this paper, we derived two approximations of the partial AUC that are calculated using unlabeled data, and proposed a semi-supervised learning method for pAUC maximization that trains a classifier robustly by maximizing the two approximated partial AUCs as well as the partial AUC using labeled data. We confirmed experimentally that our proposed method performed better than existing methods. For future work, we would like to evaluate the proposed method with different types of applications, such as anomaly detection (Yamanaka et al. 2019), and different types of classifiers, such as tree-based methods (Levatić et al. 2017; 2018).

References

Baker, S. G., and Pinsky, P. F. 2001. A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. *Journal of the American Statistical Association* 96(454):421–428.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159.

Brefeld, U., and Scheffer, T. 2005. AUC maximizing support vector learning. In *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*.

Campos, G. O.; Zimek, A.; Sander, J.; Campello, R. J.; Micenková, B.; Schubert, E.; Assent, I.; and Houle, M. E. 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* 30(4):891–927.

Cortes, C., and Mohri, M. 2004. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems*, 313–320.

Ding, Y.; Zhao, P.; Hoi, S. C.; and Ong, Y.-S. 2015. An adaptive gradient method for online AUC maximization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Dodd, L. E., and Pepe, M. S. 2003. Partial AUC estimation and regression. *Biometrics* 59(3):614–623.

Du Plessis, M. C., and Sugiyama, M. 2014. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks* 50:110–119.

Du Plessis, M.; Niu, G.; and Sugiyama, M. 2015. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, 1386–1394.

Elkan, C., and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220. ACM.

Fujino, A., and Ueda, N. 2016. A semi-supervised AUC optimization method with generative models. In *16th International Conference on Data Mining*, 883–888. IEEE.

Han, G., and Zhao, C. 2010. AUC maximization linear classifier based on active learning and its application. *Neurocomputing* 73(7-9):1272–1280.

Huang, J., and Ling, C. X. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 17(3):299–310.

Iyer, A.; Nath, S.; and Sarawagi, S. 2014. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, 530–538.

Kingma, D. P., and Ba, J. 2015. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kiryo, R.; Niu, G.; du Plessis, M. C.; and Sugiyama, M. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, 1675–1685.

Komori, O., and Eguchi, S. 2010. A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics* 11(1):314.

Lee, W. S., and Liu, B. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *International Conference on Machine Learning*, volume 3, 448–455.

Levatić, J.; Ceci, M.; Kocev, D.; and Džeroski, S. 2017. Self-training for multi-target regression with tree ensembles. *Knowledge-Based Systems* 123:41–60.

Levatić, J.; Kocev, D.; Ceci, M.; and Džeroski, S. 2018. Semi-supervised trees for multi-target regression. *Information Sciences* 450:109–127.

Li, X., and Liu, B. 2003. Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, 587–592.

Narasimhan, H., and Agarwal, S. 2013a. A structural SVM based approach for optimizing partial AUC. In *International Conference on Machine Learning*, 516–524.

Narasimhan, H., and Agarwal, S. 2013b. SVM pAUC tight: a new support vector method for optimizing partial AUC based on a tight convex upper bound. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 167–175. ACM.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Ricamato, M. T., and Tortorella, F. 2011. Partial AUC maximization in a linear combination of dichotomizers. *Pattern Recognition* 44(10-11):2669–2677.

Saerens, M.; Latinne, P.; and Decaestecker, C. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation* 14(1):21–41.

Sakai, T.; Niu, G.; and Sugiyama, M. 2018. Semi-supervised AUC optimization based on positive-unlabeled learning. *Machine Learning* 107(4):767–794.

Ueda, N., and Fujino, A. 2018. Partial AUC maximization via nonlinear scoring functions. *arXiv preprint arXiv:1806.04838*.

Wang, Z., and Chang, Y.-C. I. 2010. Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* 12(2):369–385.

Wang, R., and Tang, K. 2009. Feature selection for maximizing the area under the ROC curve. In *International Conference on Data Mining Workshops*, 400–405. IEEE.

Xie, Z., and Li, M. 2018. Semi-supervised AUC optimization without guessing labels of unlabeled data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yamanaka, Y.; Iwata, T.; Takahashi, H.; Yamada, M.; and Kanai, S. 2019. Autoencoding binary classifiers for supervised anomaly detection. In *PRICAI*, 647–659.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, 451–459.

Zhao, P.; Hoi, S. C.; Jin, R.; and YANG, T. 2011. Online AUC maximization. In *International Conference on Machine Learning*.

Zhou, L.; Lai, K. K.; and Yen, J. 2009. Credit scoring models with AUC maximization based on weighted SVM. *International Journal of Information Technology & Decision Making* 8(04):677–696.