AUC Optimization with a Reject Option

Song-Qing Shen, Bin-Bin Yang, Wei Gao*
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, 210023, China
{shensq, yangbb, gaow}@lamda.nju.edu.cn

Abstract

Making an erroneous decision may cause serious results in diverse mission-critical tasks such as medical diagnosis and bioinformatics. Previous work focuses on classification with a reject option, i.e., abstain rather than classify an instance of low confidence. Most mission-critical tasks are always accompanied with class imbalance and cost sensitivity, where AUC has been shown a preferable measure than accuracy in classification. In this work, we propose the framework of AUC optimization with a reject option, and the basic idea is to withhold the decision of ranking a pair of positive and negative instances with a lower cost, rather than mis-ranking. We obtain the Bayes optimal solution for ranking, and learn the reject function and score function for ranking, simultaneously. An online algorithm has been developed for AUC optimization with a reject option, by considering the convex relaxation and plug-in rule. We verify, both theoretically and empirically, the effectiveness of the proposed algorithm.

Introduction

Making an erroneous decision may cause serious results in many mission-critical tasks such as medical diagnostic (Hamid et al. 2017), biometric verification (Golfarelli, Maio, and Malton 1997), myoelectric pattern-recognition control (Robertson, Englehart, and Scheme 2018), and so on. For example, in an electronic nose system, researchers may be totally exposed to toxic or dangerous chemicals when the system makes an incorrect prediction (Hatami and Chira 2013), and for a mobile robot, an imprecise localization may bring about large deviation of navigation route (Marinho et al. 2018). The mission-critical studies have received much attention as machine learning techniques go to more real applications (Hatami and Chira 2013; Hamid et al. 2017; Robertson, Englehart, and Scheme 2018).

A classical solution for mission-critical tasks is to consider the classification with a reject option, also known as selective classification (Golfarelli, Maio, and Malton 1997; Bounsiar, Grall, and Beauseroy 2006; Bartlett and Wegkamp 2008; Grandvalet et al. 2009; Hamid et al. 2017; Robertson, Englehart, and Scheme 2018; Marinho et al. 2018). The basic idea is to abstain an uncertain instance with a lower cost so as to avoid mis-classification, and the cost is often more acceptable in practice. Relevant research could be traced back to Chow's rule (Chow 1970), where the Bayes decision is provided to tradeoff the optimal error versus reject rate. Most previous studies concern the reduction of classification error, which is equivalent to the improvement of predictive accuracy of classifiers.

Many mission-critical tasks are always accompanied with class imbalance and cost sensitivity, where the number of one class may overwhelm the others, and the minor class receives primary interest with higher cost. For example, in medical diagnosis, the number of healthy persons is much larger than that of patients, whereas the cost of erroneously diagnosing a patient as healthy may be much higher than that of diagnosing a healthy person as a patient. The Area Under the ROC Curve (or short for AUC) is preferable to accuracy as an evaluation measure in diverse tasks such as class-imbalance learning and cost-sensitive learning (Cortes and Mohri 2004; Gao et al. 2016; Liu et al. 2018), and various algorithm have been developed for AUC optimization (Freund et al. 2003; Zhao et al. 2011; Gao et al. 2016; Liu et al. 2018). However, it remains open for mission-critical tasks to take AUC into consideration.

This work tries to study AUC as an evaluation measure for mission-critical tasks, and the main contributions are summarized as follows:

- We propose the framework of AUC optimization with a reject option for mission-critical tasks, and present the Bayes optimal solution. The optimal score function for ranking and the reject function are also provided based on the conditional probability of data distribution.
- From the Bayes optimal solution, we develop an online algorithm for AUC optimization with a reject option by using plug-in rule and convex surrogate loss, and learn the score function and the reject function simultaneously. Our algorithm is guaranteed theoretically with regret bounds.
- We finally present extensive empirical studies to verify the effectiveness of the proposed algorithm by comparing with state-of-the-art algorithm on AUC optimization and classification with a reject option.

*This research is supported by NSFC(61921006, 61876078).

Related Work

Classification with a reject option has been a classical framework to deal with mission-critical tasks. The basic idea is to provide an option to reject an instance with a lower cost, rather than mis-classification. Chow (1957; 1970) provided the Bayes optimal decision from the trade-off between error rate and rejection.

Various SVMs-style variants have been developed for classification with a reject option (Fumera and Roli 2002; Bounsiar, Grall, and Beauseroy 2006; Bartlett and Wegkamp 2008; Grandvalet et al. 2009). Cortes, DeSalvo, and Mohri (2016) introduced a general framework of classification with a reject option. Recent years have witnessed the reject option in deep neural networks (Geifman and El-Yaniv 2017). Tortorella (2005) and Pietraszek (2005) studied an optimal rejection rule based on the ROC curve.

The studies on AUC could date back to the 1970's in signal detection theory (Egan 1975). AUC has been an important performance measure for ranking (Metz 1978; Ferri, Hernández-Orallo, and Flach 2011). Herschtal and Raskutti (2004) proposed the RankOpt algorithm based on gradient descent method, and some variants of traditional algorithms have been developed for AUC optimization, such as boosting (Freund et al. 2003; Rudin and Schapire 2009) and SVM (Brefeld and Scheffer 2005; Joachims 2005).

Zhao et al. (2011) proposed an online AUC optimization algorithm based on reservoir sampling. Gao et al. (2016) proposed an online AUC optimization algorithm using the consistent square loss. Ying, Wen, and Lyu (2016) further built on a saddle point formulation for a consistent square loss and proposed a stochastic algorithm with convergence analysis. Based on this saddle point formulation, Natole, Ying, and Lyu (2018) and Liu et al. (2018) proposed new stochastic algorithms with tighter regret bounds.

The rest of the paper is organized as follows. We begin with the some preliminaries, and propose the framework of AUC optimization with a reject option. We then develop the AUCRO algorithm with theoretical guarantees. We finally conduct extensive experiments and conclude this work.

Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} = \{-1, +1\}$ be the instance and label space, respectively, and assume that $\mathcal{D}$ is an underlying (unknown) distribution over joint space $\mathcal{X} \times \mathcal{Y}$. Denote by

$$\eta(\mathbf{x}) = \Pr[y = +1|\mathbf{x}]$$

the conditional probability of positive instances according to distribution $\mathcal{D}$. We can also express distribution $\mathcal{D}$ with triplet $(\mathcal{D}^+, \mathcal{D}^-, p)$, where $\mathcal{D}^+(\mathbf{x}) = \Pr[\mathbf{x}|y = +1]$, $\mathcal{D}^-(\mathbf{x}) = \Pr[\mathbf{x}|y = -1]$ and $p = \Pr[y = +1]$.

Notice that the distribution $\mathcal{D}$ is unknown in practice, and what we can observe is a training sample S_n of size n, i.e.,

$$S_n = \{(\mathbf{x}_1^+, +1), (\mathbf{x}_2^+, +1), \cdots, (\mathbf{x}_{n_+}^+, +1), \\ (\mathbf{x}_1^-, -1), (\mathbf{x}_2^-, -1), \cdots, (\mathbf{x}_{n_-}^-, -1)\},$$

where each element is drawn independently and identically (i.i.d) from distribution $\mathcal{D}$. Here, we denote by n_+ and n_- the cardinality of positive and negative instances in training sample S_n, respectively, and $n = n_- + n_+$.

Let $f\colon \mathcal{X} \to \mathbb{R}$ be a score function. Given a sample S_n, the AUC w.r.t. function f is defined as

$$\mathrm{AUC}(f, S_n) = 1 - \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \ell(f, \mathbf{x}_i^+, \mathbf{x}_j^-),$$

where the ranking loss $\ell(f, \mathbf{x}_i^+, \mathbf{x}_j^-)$ is given by

$$\ell(f, \mathbf{x}_i^+, \mathbf{x}_j^-) = \mathbb{I}[f(\mathbf{x}_i^+) < f(\mathbf{x}_j^-)] + \frac{1}{2}\mathbb{I}[f(\mathbf{x}_i^+) = f(\mathbf{x}_j^-)].$$

Here, $\mathbb{I}[\cdot]$ is the indicator function which returns 1 if the argument is true and 0 otherwise. Essentially, AUC is equivalent to the Wilcoxon-Mann-Whitney statistic (Yan et al. 2003).

AUC Optimization with a Reject Option

For AUC optimization with a reject option, we introduce an augmented function $r\colon \mathbb{R} \times \mathbb{R} \to \{0, 1\}$, and make a reject option by setting

$$r(f(\mathbf{x}_i^+), f(\mathbf{x}_j^-)) = 0.$$

In such case, there is no definitive ranking between $\mathbf{x}_i^+$ and $\mathbf{x}_j^-$. We further introduce the loss function for AUC with a reject option as follows:

$$\begin{aligned} L(r, f, \mathbf{x}_i^+, \mathbf{x}_j^-) &= \ell(f, \mathbf{x}_i^+, \mathbf{x}_j^-) r(f(\mathbf{x}_i^+), f(\mathbf{x}_j^-)) \\ &\quad + d(1 - r(f(\mathbf{x}_i^+), f(\mathbf{x}_j^-))) \\ &= \big(\ell(f, \mathbf{x}_i^+, \mathbf{x}_j^-) - d\big) r(f(\mathbf{x}_i^+), f(\mathbf{x}_j^-)) + d \end{aligned}$$

where d is the cost of a reject option. For convenience, we introduce a constant

$$\kappa = d/(1 - d). \tag{1}$$

Our goal is to optimize the empirical risk over the training data S_n defined below:

$$R(f, S_n) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} L(r, f, \mathbf{x}_i^+, \mathbf{x}_j^-).$$

Based on the Bayes optimal solution from Theorem 1 (to be shown in the next section), we consider the reject function as follows:

$$r\big(f(\mathbf{x}_i^+), f(\mathbf{x}_j^-)\big) = \mathbb{I}\left[\frac{f(\mathbf{x}_j^-)}{f(\mathbf{x}_i^+)} > \frac{1}{\kappa}\right] + \mathbb{I}\left[\frac{f(\mathbf{x}_j^-)}{f(\mathbf{x}_i^+)} < \kappa\right],$$

and the score function f is given by

$$f(\mathbf{x}) = 1 - 1/\eta(\mathbf{x}).$$

We restrict $d \in (0, 1/2]$, which is similar to the scenario in classification with a reject option (Herbei and Wegkamp 2006; Bartlett and Wegkamp 2008), and we reject pairwise instances for smaller d. The problem can be degenerated to the traditional AUC optimization without any reject option as for $d = 1/2$.

By plugging the reject function $r(\cdot,\cdot)$ into loss function $L(r, f, \mathbf{x}_i^+, \mathbf{x}_j^-)$, we have

$$L(r, f, \mathbf{x}_i^+, \mathbf{x}_j^-) = (1-d)\mathbb{I}\left[\frac{f(\mathbf{x}_j^-)}{f(\mathbf{x}_i^+)} < \kappa\right] + d\mathbb{I}\left[\frac{f(\mathbf{x}_j^-)}{f(\mathbf{x}_i^+)} < \frac{1}{\kappa}\right].$$

We can not directly optimize the loss function L since the class-conditional probability $\eta(\mathbf{x})$ is unknown. For this problem, we consider plug-in rules (Herbei and Wegkamp 2006; Devroye, Györfi, and Lugosi 2013), where an empirical conditional probability $\widehat{\eta}(\mathbf{x})$ is estimated to approximate the true conditional probability $\eta(\mathbf{x})$.

In this work, we focus on logistic regression for plug-in rule to estimate $\eta(\mathbf{x})$ as in the work of (Herbei and Wegkamp 2006), that is,

$$\widehat{\eta}(\mathbf{x}) = \frac{e^{\mathbf{w}^\top \mathbf{x}+b}}{1+e^{\mathbf{w}^\top \mathbf{x}+b}}.$$

Substituting into the score function f, we have

$$f(\mathbf{x}) = -e^{-(\mathbf{w}^\top \mathbf{x}+b)}. \tag{2}$$

This follows that

$$R(f, S_n) = \sum_{i=1}^{n^+}\sum_{j=1}^{n^-} \frac{d\mathbb{I}\left[\ln\kappa + \mathbf{w}^\top(\mathbf{x}_i^+ - \mathbf{x}_j^-) \le 0\right]}{n^+n^-} + \sum_{i=1}^{n^+}\sum_{j=1}^{n^-} \frac{(1-d)\mathbb{I}\left[-\ln\kappa + \mathbf{w}^\top(\mathbf{x}_i^+ - \mathbf{x}_j^-) \le 0\right]}{n^+n^-}. \tag{3}$$

For non-convex loss function in Eqn. (3), we exploit the surrogate loss functions as follows:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n^+}\sum_{j=1}^{n^-} \frac{1}{2n^+n^-}\Big\{d\psi\left(\ln\kappa + \mathbf{w}^\top(\mathbf{x}_i^+ - \mathbf{x}_j^-)\right) + (1-d)\psi\left(-\ln\kappa + \mathbf{w}^\top(\mathbf{x}_i^+ - \mathbf{x}_j^-)\right)\Big\} + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{4}$$

where λ is a regularization parameter and ψ is a convex surrogate loss function.

We adopt the square loss in Eqn. (4) as in the works of (Gao and Zhou 2015; Gao et al. 2016), i.e.,

$$\mathcal{L}(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n^+}\sum_{j=1}^{n^-} \frac{d\left(1-\ln\kappa - \mathbf{w}^\top(\mathbf{x}_i^+ - \mathbf{x}_j^-)\right)^2}{2n^+n^-} + \sum_{i=1}^{n^+}\sum_{j=1}^{n^-} \frac{(1-d)\left(1+\ln\kappa - \mathbf{w}^\top(\mathbf{x}_i^+ - \mathbf{x}_j^-)\right)^2}{2n^+n^-}.$$

Motivated from (Gao et al. 2016), we rewrite the loss function $\mathcal{L}(\mathbf{w})$ as a sum of losses for each individual training instance $\sum_{t=1}^{T}\mathcal{L}_t(\mathbf{w})$, where $\mathcal{L}_t(\mathbf{w})$ is equal to

$$\frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{t-1} \frac{\mathbb{I}\left[y_i \ne y_t\right] d\left(1-\ln\kappa - y_t\mathbf{w}^\top(\mathbf{x}_t - \mathbf{x}_i)\right)^2}{2\left|\{i\in[t-1] : y_iy_t = -1\}\right|} + \sum_{i=1}^{t-1} \frac{\mathbb{I}\left[y_i \ne y_t\right](1-d)\left(1+\ln\kappa - y_t\mathbf{w}^\top(\mathbf{x}_t - \mathbf{x}_i)\right)^2}{2\left|\{i\in[t-1] : y_iy_t = -1\}\right|}.$$

Algorithm 1 The AUCRO Algorithm

Input:

Training instances $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$, the regularization parameter $\lambda > 0$ and stepsizes $\{\eta_t\}_{t=1}^T$.

Initialization:

Set $T_0^+ = T_0^- = 0$, $\mathbf{c}_0^+ = \mathbf{c}_0^- = \mathbf{0}$, $\mathbf{w}_0 = \mathbf{0}$ and $\Gamma_0^+ = \Gamma_0^- = [\mathbf{0}]_{m\times m}$.

for $t = 1, 2, \ldots, T$ **do**
  Receive a training instance $(\mathbf{x}_t, y_t)$;
  if $y_t = 1$ **then**
    $T_t^+ = T_{t-1}^+ + 1$ and $T_t^- = T_{t-1}^-$;
    $\mathbf{c}_t^+ = \mathbf{c}_{t-1}^+ + \frac{1}{T_t^+}(\mathbf{x}_t - \mathbf{c}_{t-1}^+)$ and $\mathbf{c}_t^- = \mathbf{c}_{t-1}^-$;
    Update Γ_t^+ using Eqn. (8);
    Calculate the gradient $\nabla\mathcal{L}_t(\mathbf{w}_{t-1})$;
  else
    $T_t^- = T_{t-1}^- + 1$ and $T_t^+ = T_{t-1}^+$;
    $\mathbf{c}_t^- = \mathbf{c}_{t-1}^- + \frac{1}{T_t^-}(\mathbf{x}_t - \mathbf{c}_{t-1}^-)$ and $\mathbf{c}_t^+ = \mathbf{c}_{t-1}^+$;
    Update Γ_t^- using Eqn. (9);
    Calculate the gradient $\nabla\mathcal{L}_t(\mathbf{w}_{t-1})$;
  end if
  $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t\nabla\mathcal{L}_t(\mathbf{w}_{t-1})$
end for

Notice that this is an unbiased estimation to $\mathcal{L}(\mathbf{w})$. We denote by T_t^+ and T_t^- the number of positive and negative instances in sequence S_t, respectively. We also set $\mathcal{L}_t(\mathbf{w}) = 0$ for the case $T_t^+T_t^- = 0$. For $y_t = 1$, we have the gradient

$$\nabla\mathcal{L}_t(\mathbf{w}) = \lambda\mathbf{w} + \mathbf{x}_t\mathbf{x}_t^\top\mathbf{w} - \alpha\mathbf{x}_t + \sum_{i:y_i=-1}\left(\alpha\mathbf{x}_i + (\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{x}_i\mathbf{x}_t^\top - \mathbf{x}_t\mathbf{x}_i^\top)\mathbf{w}\right)/T_t^-, \tag{5}$$

where $\alpha = 1 - \ln\kappa(2d-1)$. We further denote by

$$\mathbf{c}_t^- = \sum_{i:y_i=-1}\mathbf{x}_i/T_t^-,$$
$$\Gamma_t^- = \sum_{i:y_i=-1}\left(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{c}_t^-\left[\mathbf{c}_t^-\right]^\top\right)/T_t^-,$$

the mean and covariance matrix of negative class, respectively. Then, Eqn. (5) can be simplified as

$$\nabla\mathcal{L}_t(\mathbf{w}) = \lambda\mathbf{w} + \alpha(\mathbf{c}_t^- - \mathbf{x}_t) + (\mathbf{c}_t^- - \mathbf{x}_t)(\mathbf{c}_t^- - \mathbf{x}_t)^\top\mathbf{w} + \Gamma_t^-\mathbf{w}. \tag{6}$$

Similarly, we calculate the gradient for $y_t = -1$ as follows:

$$\nabla\mathcal{L}_t(\mathbf{w}) = \lambda\mathbf{w} - \alpha(\mathbf{c}_t^+ - \mathbf{x}_t) + (\mathbf{c}_t^+ - \mathbf{x}_t)(\mathbf{c}_t^+ - \mathbf{x}_t)^\top\mathbf{w} + \Gamma_t^+\mathbf{w}, \tag{7}$$

where

$$\mathbf{c}_t^+ = \sum_{i:y_i=1}\mathbf{x}_i/T_t^+,$$
$$\Gamma_t^+ = \sum_{i:y_i=1}\left(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{c}_t^+\left[\mathbf{c}_t^+\right]^\top\right)/T_t^+.$$

Algorithm 1 presents the detailed description of AUC optimization with a Reject Option (AUCRO). Specifically, we update Γ_t^+ (Line 6) or Γ_t^- (Line 11) for each iteration, respectively, as follows:

$$\Gamma_t^+ = \Gamma_{t-1}^+ + \mathbf{c}_{t-1}^+ \left[\mathbf{c}_{t-1}^+\right]^\top - \mathbf{c}_t^+ \left[\mathbf{c}_t^+\right]^\top + \left(\mathbf{x}_t\mathbf{x}_t^\top - \Gamma_{t-1}^+ - \mathbf{c}_{t-1}^+ \left[\mathbf{c}_{t-1}^+\right]^\top\right)/T_t^+, \quad (8)$$

$$\Gamma_t^- = \Gamma_{t-1}^- + \mathbf{c}_{t-1}^- \left[\mathbf{c}_{t-1}^-\right]^\top - \mathbf{c}_t^- \left[\mathbf{c}_t^-\right]^\top + \left(\mathbf{x}_t\mathbf{x}_t^\top - \Gamma_{t-1}^- - \mathbf{c}_{t-1}^- \left[\mathbf{c}_{t-1}^-\right]^\top\right)/T_t^-. \quad (9)$$

We finally use the stochastic gradient descent to update the classifier as follows:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \nabla \mathcal{L}_t(\mathbf{w}_{t-1}),$$

where η_t is the stepsize for each t-th iteration.

Main Theoretical Analysis

This section presents theoretical analysis. We first provide the Bayes optimal solution for optimal ranking, as well as the reject function. Define the expected risk with respect to distribution $\mathcal{D}$ as follows:

$$R(f, \mathcal{D}) = E_{\mathbf{x}^+\sim\mathcal{D}^+, \mathbf{x}^-\sim\mathcal{D}^-} L(r, f, \mathbf{x}^+, \mathbf{x}^-).$$

Theorem 1. *Given distribution $\mathcal{D}$, the Bayes optimal function f^* is defined by*

$$f^*(\mathbf{x}) = 1 - 1/\eta(\mathbf{x}),$$

that is, we have $R(f, \mathcal{D}) \geq R(f^*, \mathcal{D})$ for any score function $f\colon \mathcal{X} \to \mathbb{R}$. Here, we assume $\eta(\mathbf{x}) \in (0, 1)$ without loss of generality. We also have the corresponding reject function, for each positive instance $\mathbf{x}^+$ and negative one $\mathbf{x}^-$,

$$r\big(f(\mathbf{x}_i^+), f(\mathbf{x}_j^-)\big) = \mathbb{I}\left[\frac{f(\mathbf{x}_j^-)}{f(\mathbf{x}_i^+)} > \frac{1}{\kappa}\right] + \mathbb{I}\left[\frac{f(\mathbf{x}_j^-)}{f(\mathbf{x}_i^+)} < \kappa\right]$$

where the κ is defined by Eqn. (1)*.*

Proof. Recall that our optimization objective is given by

$$\arg\min_f E_{\mathbf{x}_1\sim\mathcal{D}^+, \mathbf{x}_2\sim\mathcal{D}^-} L(r, f, \mathbf{x}_1, \mathbf{x}_2).$$

Without changing the optimal Bayes solution, the optimization objective can be transformed as

$$\arg\min_f \underset{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2)\sim\mathcal{D}^2}{E} \big[\mathbb{I}[y_1 > y_2] L(r, f, \mathbf{x}_1, \mathbf{x}_2) + \mathbb{I}[y_2 > y_1] L(r, f, \mathbf{x}_2, \mathbf{x}_1)\big].$$

It is easy to observe that the reject function r is symmetric, and we have

$$r\big(f(\mathbf{x}_1), f(\mathbf{x}_2)\big) = r\big(f(\mathbf{x}_2), f(\mathbf{x}_1)\big) = 0$$

if there is no definitive ranking between $\mathbf{x}_1$ and $\mathbf{x}_2$. We also denote by

$$r = r\big(f(\mathbf{x}_1), f(\mathbf{x}_2)\big) = r\big(f(\mathbf{x}_2), f(\mathbf{x}_1)\big).$$

Table 1: Benchmark datasets

dataset	#instance	#feature	dataset	#instance	#feature
diabetes	768	8	a9a	32,561	123
fourclass	862	2	w8a	49,749	300
german	1,000	24	connect-4	67,557	126
splice	3,175	60	acoustic	78,823	50
letter	20,000	16	covtype	581,012	54

This follows that

$$\arg\min_f \underset{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2)\sim\mathcal{D}^2}{E} \big[d(1-r)(a+b-2ab) + ra(1-b)\ell(f, \mathbf{x}_1, \mathbf{x}_2) + rb(1-a)\ell(f, \mathbf{x}_2, \mathbf{x}_1)\big]$$

where $a = \eta(\mathbf{x}_1)$ and $b = \eta(\mathbf{x}_2)$. Minimize the above formula, and it is easy to obtain

$$r\big(f(\mathbf{x}_1), f(\mathbf{x}_2)\big) = r\big(f(\mathbf{x}_2), f(\mathbf{x}_1)\big) = 0$$

when $\kappa \leq (1-1/b)/(1-1/a) \leq 1/\kappa$. It is easy to find the Bayes optimal solution f^*

$$f^*(\mathbf{x}) = 1 - \frac{1}{\eta(\mathbf{x})}$$

and the reject function r is given by

$$r\big(f(\mathbf{x}_1), f(\mathbf{x}_2)\big) = \mathbb{I}\left[\frac{f(\mathbf{x}_2)}{f(\mathbf{x}_1)} > \frac{1}{\kappa}\right] + \mathbb{I}\left[\frac{f(\mathbf{x}_2)}{f(\mathbf{x}_1)} < \kappa\right].$$

We also have the Bayes risk

$$R^* = \underset{\mathbf{x}^+\sim\mathcal{D}^+, \mathbf{x}^-\sim\mathcal{D}^-}{E} \big[\min\{\eta(\mathbf{x}^+), \eta(\mathbf{x}^-), d\eta(\mathbf{x}^+) + d\eta(\mathbf{x}^-) + (1-2d)\eta(\mathbf{x}^+)\eta(\mathbf{x}^-)\} - \eta(\mathbf{x}^+)\eta(\mathbf{x}^-)\big].$$

This completes the proof. □

Remark: $f^*(\mathbf{x}) = 1 - 1/\eta(\mathbf{x})$ is not the only Bayes optimal solution, whereas it is helpful for our later analysis. Notice that the reject option takes effect only for $d \in (0, 1/2]$, that is why we restrict the range of d in the previous section. Without loss of generality, we assume $\eta(\mathbf{x}) \in (0, 1)$ for any $\mathbf{x} \in \mathcal{X}$ in the above proof.

In the rest of this work, we present the convergence analysis of the proposed AUCRO algorithm. We first introduce a lemma for smooth functions from (Nesterov 2003)[Theorem 2.1.5] as follows:

Lemma 2. *Given a linear function space $\mathcal{W} \subseteq \mathbb{R}^m$, if a function $g\colon \mathcal{W} \to \mathbb{R}$ is β-smooth, then, for every $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, we have*

$$g(\mathbf{w}) - g(\mathbf{w}') \leq \langle \nabla g(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{1}{2}\beta\|\mathbf{w} - \mathbf{w}'\|^2.$$

Define $\mathbf{w}_* = \arg\min_{\mathbf{w}} \sum_t \mathcal{L}_t(\mathbf{w})$. The following theorem guarantees the convergence of Algorithm 1.

Theorem 3. *For $\|\mathbf{x}_t\| \leq 1$ ($t \in [T]$), $\|\mathbf{w}_*\| \leq D$ and $TL^* = \sum_{t=1}^T \mathcal{L}_t(\mathbf{w}_*)$, we have*

$$\sum\nolimits_t \mathcal{L}_t(\mathbf{w}_{t-1}) - \sum\nolimits_t \mathcal{L}_t(\mathbf{w}_*) \leq 2\theta D^2 + D\sqrt{2\theta TL^*}$$

where $\theta = 4 + \lambda$ and $\eta_t = 1/(\theta + \sqrt{\theta^2 + 2\theta TL^*/D^2})$.

Table 2: Comparison of the testing AUC values (mean±std.) on benchmark datasets

	Compared methods				AUCRO				
dataset	FSAUC	OPAUC	OAM_{seq}	OAM_{gd}	d=0.5	d=0.48	d=0.46	d=0.44	d=0.42
diabetes	.8143±.0235	.8335±.0319	.8035±.0243	.8305±.0232	.8318±.0232	.8549±.0232	.8768±.0233	.8960±.0231	.9129±.0220
fourclass	.8318±.0311	.8279±.0239	.8295±.0303	.8317±.0303	.8303±.0269	.8534±.0264	.8769±.0254	.9004±.0224	.9248±.0197
german	.7791±.0392	.7911±.0281	.7762±.0356	.7547±.0302	.7896±.0332	.8146±.0346	.8385±.0352	.8613±.0350	.8826±.0343
splice	.9088±.0087	.9240±.0084	.8965±.0130	.8905±.0145	.9243±.0078	.9435±.0072	.9591±.0063	.9714±.0053	.9810±.0044
letter	.8106±.0064	.8116±.0046	.7455±.0128	.7826±.0146	.8108±.0040	.8404±.0040	.8687±.0039	.8956±.0039	.9204±.0038
a9a	.8977±.0037	.9002±.0034	.8637±.0044	.8507±.0114	.8994±.0034	.9206±.0033	.9390±.0031	.9548±.0028	.9680±.0024
w8a	.9482±.0058	.9657±.0046	.8858±.0133	.9336±.0145	.9703±.0038	.9816±.0029	.9884±.0021	.9928±.0013	.9954±.0009
connect-4	.8589±.0029	.8585±.0026	.7351±.0124	.7986±.0135	.8578±.0027	.8831±.0027	.9058±.0026	.9259±.0024	.9435±.0022
acoustic	.8017±.0038	.8018±.0026	.7765±.0046	.7661±.0121	.8010±.0030	.8370±.0033	.8636±.0033	.8844±.0030	.9006±.0031
covtype	.8230±.0014	.8235±.0009	.6770±.0231	.7459±.0272	.8232±.0014	.8494±.0014	.8734±.0014	.8953±.0013	.9155±.0012

Proof. Since we have defined $\mathcal{L}_t(\mathbf{w}) = 0$ for $T_t^+ T_t^- = 0$ above, this easy case will not be analyzed. Here we only consider the general case $T_t^+ T_t^- \neq 0$ in our proof. Recall aforementioned $\mathcal{L}_t(\mathbf{w})$ and it is easy to calculate the gradient

$$\nabla \mathcal{L}_t(\mathbf{w}) = \lambda \mathbf{w} - \frac{\sum_{i=1}^{t-1} \mathbb{I}\left[y_i \neq y_t\right]}{|\{i \in [t-1] : y_i y_t = -1\}|} \times \Big\{ d\big(1 - \ln \kappa - y_t \mathbf{w}^\top (\mathbf{x}_t - \mathbf{x}_i)\big) y_t (\mathbf{x}_t - \mathbf{x}_i) + (1-d)\big(1 + \ln \kappa - y_t \mathbf{w}^\top (\mathbf{x}_t - \mathbf{x}_i)\big) y_t (\mathbf{x}_t - \mathbf{x}_i) \Big\}.$$

For $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ and $\|\mathbf{x}_t\| \leq 1$,we have

$$\|\nabla \mathcal{L}_t(\mathbf{w}') - \nabla \mathcal{L}_t(\mathbf{w}')\| \leq \theta \|\mathbf{w}' - \mathbf{w}\|,$$

which implies that $\mathcal{L}_t$ is θ-smooth. Denote by

$$\mathbf{w}_{t*} = \arg\min_{\mathbf{w}} \mathcal{L}_t(\mathbf{w}),$$

and this gives $\nabla \mathcal{L}_t(\mathbf{w}_{t*}) = 0$ from the convex and differentiable loss $\mathcal{L}_t$. Based on Lemma 2, we have

$$0 \leq \mathcal{L}_t(\mathbf{w}_{t*}) \leq \min_c \Big[\mathcal{L}_t\big(\mathbf{w}_{t-1} - c \nabla \mathcal{L}_t(\mathbf{w}_{t-1})\big)\Big] \leq \min_c \Big[\mathcal{L}_t(\mathbf{w}_{t-1}) - c\|\nabla \mathcal{L}_t(\mathbf{w}_{t-1})\|^2 + \frac{1}{2}\theta c^2 \|\nabla \mathcal{L}_t(\mathbf{w}_{t-1})\|^2\Big] = \mathcal{L}_t(\mathbf{w}_{t-1}) - \frac{1}{2\theta}\|\nabla \mathcal{L}_t(\mathbf{w}_{t-1})\|^2. \quad (10)$$

From the convexity of function $\mathcal{L}_{t-1}$, we have

$$\mathcal{L}_t(\mathbf{w}_{t-1}) - \mathcal{L}_t(\mathbf{w}_*) \leq \langle \nabla \mathcal{L}_t(\mathbf{w}_{t-1}), \mathbf{w}_{t-1} - \mathbf{w}_* \rangle. \quad (11)$$

Also, we have

$$\|\mathbf{w}_t - \mathbf{w}_*\|^2 = \|\mathbf{w}_{t-1} - \eta_t \nabla \mathcal{L}_t(\mathbf{w}_{t-1}) - \mathbf{w}_*\|^2 = \|\mathbf{w}_{t-1} - \mathbf{w}_*\|^2 - 2\eta_t \langle \nabla \mathcal{L}_t(\mathbf{w}_{t-1}), \mathbf{w}_{t-1} - \mathbf{w}_* \rangle + \eta_t^2 \|\nabla \mathcal{L}_t(\mathbf{w}_{t-1})\|^2. \quad (12)$$

By using Eqns. (10) and (11) in Eqn. (12), we have

$$(1 - \theta\eta_t)\mathcal{L}_t(\mathbf{w}_{t-1}) - \mathcal{L}_{t-1}(\mathbf{w}_*) \leq \frac{1}{2\eta_t}\|\mathbf{w}_{t-1} - \mathbf{w}_*\|^2 - \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}_*\|^2.$$

Summing over $t = 1, \cdots, T$ and rearranging, we obtain

$$\sum_{t=1}^{T}(1 - \theta\eta_t)\mathcal{L}_t(\mathbf{w}_{t-1}) - \sum_{t=1}^{T}\mathcal{L}_{t-1}(\mathbf{w}_*) \leq \frac{1}{2\eta_1}\|\mathbf{w}_0 - \mathbf{w}_*\|^2 - \frac{1}{2\eta_T}\|\mathbf{w}_T - \mathbf{w}_*\|^2 + \sum_{t=1}^{T-1}\Big(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\Big)\|\mathbf{w}_t - \mathbf{w}_*\|.$$

By setting $\eta_t = \eta$, we have

$$\frac{1}{2\eta_1}\|\mathbf{w}_0 - \mathbf{w}_*\|^2 - \frac{1}{2\eta_T}\|\mathbf{w}_T - \mathbf{w}_*\|^2 \leq \frac{1}{2\eta}\|\mathbf{w}_*\| \leq \frac{D^2}{2\eta}.$$

From $\mathbf{w}_0 = \mathbf{0}$ and $\|\mathbf{w}_*\| \leq D$, we finally get

$$\sum_{t=1}^{T}\mathcal{L}_t(\mathbf{w}_{t-1}) - \sum_{t=1}^{T}\mathcal{L}_t(\mathbf{w}_*) \leq \frac{1}{1 - \theta\eta}\left(\frac{D^2}{2\eta} + \theta\eta T L^*\right).$$

By setting

$$\eta = 1/(\theta + \sqrt{\theta^2 + 2\theta T L^*/D^2})$$

and simple derivations, the theorem holds as desired. □

Remark: This theorem presents an $O(1/T)$ convergence rate for the AUCRO algorithm if the distribution is separable, i.e., $L^* = 0$, and an $O(1/\sqrt{T})$ convergence rate for general case. Besides, the convergence rate can be improved to $O(\log T/T)$ by exploring the strongly convexity of $\mathcal{L}_t(\mathbf{w}_t)$. Here we consider the standard and suboptimal regret analysis on exploiting the smoothness of $\mathcal{L}_t(\mathbf{w}_t)$ for simplicity.

Experiments

We evaluate the performance of our method on ten benchmark datasets, as summarized in Table 1[1]. We compare our method with state-of-the-art AUC optimization algorithms without a reject option. We then compare our method with conventional rejection algorithms that reject instances of low confidence. We finally analyze the parameter influence.

The features have been scaled to $[-1, 1]$ for all datasets. Multi-class datasets have been transformed into binary ones by randomly partitioning classes into two groups, where each group contains almost the same number of classes.

[1]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/

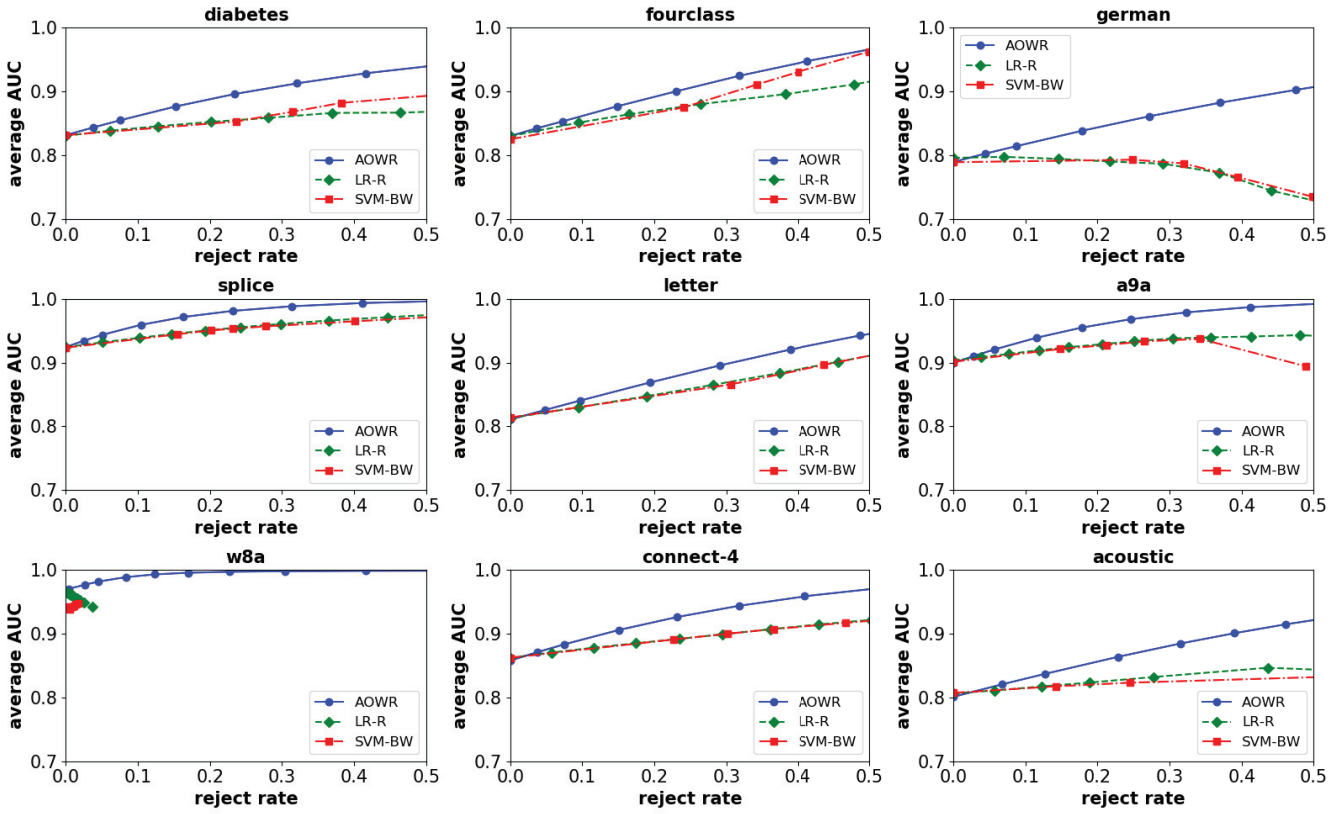

Figure 1: The curves of reject-rate vs AUC on benchmark datasets

Comparisons with AUC Approches

We compare our method with four state-of-the-art algorithms on AUC optimization as follows:

- **FSAUC**: An AUC optimization algorithm which builds on a saddle point formulation for the consistent square loss (Liu et al. 2018);
- **OPAUC**: An online AUC optimization algorithm which employ the consistent square loss. (Gao et al. 2016);
- **OAM$_{\text{seq}}$**: An online AUC optimization algorithm based on reservoir sampling which updates the model w.r.t. the sequence of pairwise instances (Zhao et al. 2011);
- **OAM$_{\text{gd}}$**: An online AUC optimization algorithm based on reservoir sampling which updates model with gradient descent approach (Zhao et al. 2011).

Two trials of 5-fold cross-validation is executed on training sets to decide the learning rate $\eta_t \in 2^{[-12:\,10]}$ and the regularized parameter $\lambda \in 2^{[-10:\,2]}$ for our algorithm. For FSAUC, we tune the initial stepsize $\eta_1 \in 2^{[-10:\,10]}$ and the parameter $R \in 10^{[-1:\,5]}$, as recommended in (Liu et al. 2018). For OPAUC, stepsize η_t is decided within the range $2^{[-12:\,10]}$ and the regularization parameter λ is decided within the range $2^{[-10:\,2]}$ as recommended in (Gao et al. 2016). For OAM$_{\text{seq}}$ and OAM$_{\text{gd}}$, the buffer sizes are fixed to be 100 and the penalty parameter C is decided within $2^{[-10:\,10]}$ as recommended in (Zhao et al. 2011).

The performance of compared methods are evaluated by five trials of 5-fold cross-validation, where AUC values are obtained by averaging over these 25 runs, as summarized in Table 2. Firstly, when reject cost $d = 0.5$, that is to say when our method does not reject any pairwise instances, our method achieve better AUC than OAM$_{\text{seq}}$ and OAM$_{\text{gd}}$ and the performance is comparable to FSAUC and OPAUC. Secondly, when reject cost $d < 0.5$, that is to say when our method begin to reject some uncertain pairwise instances, the AUC achieved goes up as reject cost d goes down. On the other hand, FSAUC, OPAUC, OAM$_{\text{seq}}$ and OAM$_{\text{gd}}$ have no ability to reject uncertain pairwise instances.

Comparisons with Rejection Approaches

Conventional algorithms on classification with a reject option abstain uncertain instances without prediction. Since our method is the first to reject making a rank, to show the effectiveness of our method, we make comparisons with rejection approaches **LR-R** and **SVM-BW** in enhancing AUC:

- **LR-R**: A rejection algorithm which is embedded with plug-in rule (logistic regression rule) to decide whether to reject an instance or not (Herbei and Wegkamp 2006);
- **SVM-BW**: An SVM-based algorithm using a generalized hinge loss which allow the classifier to reject instances according to the distance to the classification hyperplane (Bartlett and Wegkamp 2008).

We evaluate the performance of AUCRO and conventional rejection methods on 'reject-rate v.s. AUC' curve. This is inspired by (Nadeem, Zucker, and Hanczar 2009) that proposed 'reject-rate v.s. accuracy' curve. In the comparisons of conventional rejection methods, the method getting higher accuracy under the same reject rate is better. More specifically, the more area under the curve, the better the method is. This way of comparison is similar to AUC comparison as AUC is the area under the 'ROC-AUC curve'. So here we compare the area under 'reject-rate v.s. AUC' curve to show the effectiveness of our method.

For AUCRO, parameters are tuned in the same way as aforementioned. For LR-R, two trials of 5-fold cross-validation is executed on training sets to decide the regularized parameter $\lambda \in 2^{[-10:\,10]}$. For SVM-BW, we tune the regularization parameter $\lambda \in 2^{[-10:\,10]}$ and choose the linear kernel as kernel function.

The performance of compared methods are evaluated by five trials of 5-fold cross-validation, where AUC values and corresponding reject rates are obtained by averaging over these 25 runs. The 'reject-rate v.s. AUC' curves on benchmark datasets are drawn in Figure 1. It can be observed that when the same percentage of instances are rejected, our method achieves higher AUC and higher growth rate. When the reject rate is zero, though AUCRO is an online algorithm, it gets comparable AUC to the other two batch algorithms. Note that AUCs of the dataset w8a never grow up and change slowly in LR-R and SVM-BW. We think that w8a is a severely class-imbalanced dataset where instances in majority class is dozens of times more than in minority class such that conventional univariate rejection algorithms cannot improve AUC with rejecting instances. However, our method can process such severely imbalanced datasets well.

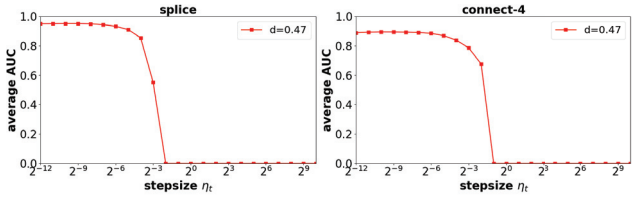

Figure 2: Influence of stepsize η_t

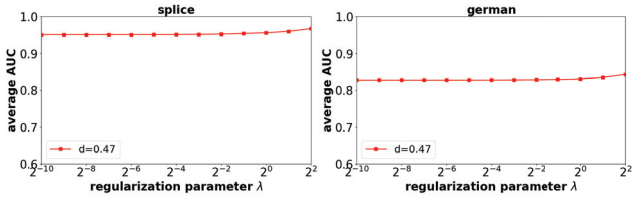

Figure 3: Influence of regularization parameter λ

Parameter Analysis

We analyze the influence of parameters in this section. Since different reject cost d lead to curves of similar shape in the analysis of η_t and λ, we only show the case $d = 0.47$ in the figure for simplicity. And due to page limit, we only present the results of several datasets for the study of each parameter, but the trends are similar on other datasets.

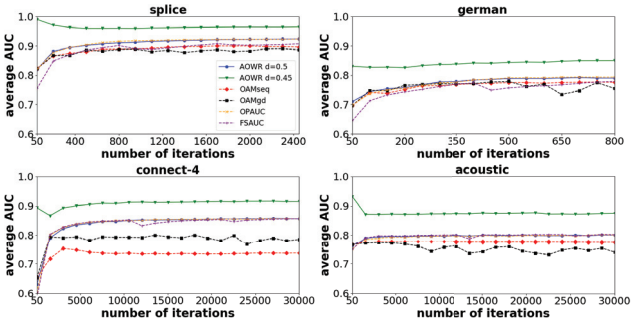

Figure 4: Influence of iterations

Figure 2 shows that there is a relatively big range $\eta_t \in [2^{-12}, 2^{-4}]$ where AUCRO achieves good results with $d = 0.47$. When η_t is set to values larger than 2^{-4}, the performance drops off dramatically. Figure 3 shows that AUCRO is relatively not sensitive to the regularization parameter λ. It can be seen that when λ is set to a big value, AUCRO may get slightly higher AUC. That's because larger λ prefers a smaller $\|\mathbf{w}\|_2$, accompanied with similar scores and higher reject rate. Figure 4 shows the influence of the number of iterations for AUCRO, FSAUC, OPAUC, OAM_{seq} and OAM_{gd}. When $d = 0.5$, AUCRO converges faster and more smoothly than FSAUC, OAM_{seq} and OAM_{gd}, and gets a comparable convergence rate to OPAUC. When $d < 0.5$, AUCRO converges even faster and there lies the difference that AUCRO achieves high AUC before the convergence due to the zero initialization of $\mathbf{w}$ which leads to high reject rates in the initial iterations.

Conclusion

This work introduces the framework for AUC optimization with a reject option. We present the Bayes rule for optimal ranking and interpret the design of the reject function. We then present an online algorithm for AUC optimization with a reject option based on plug-in rule and convex relaxation of surrogate loss. We verify the effectiveness of the proposed algorithm empirically and theoretically. An interesting work in the future is to consider different reject option into AUC optimization with tighter regret bound, and we could also apply reject option in the deep neural networks for ranking.

References

Bartlett, P. L., and Wegkamp, M. H. 2008. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* 9:1823–1840.

Bounsiar, A.; Grall, E.; and Beauseroy, P. 2006. A kernel based rejection method for supervised classification. *International Journal of Computational Intelligence* 3(4):312–321.

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

Chow, C. K. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers* EC-6(4):247–254.

Chow, C. K. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16(1):41–46.

Cortes, C., and Mohri, M. 2004. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems 17*, 313–320.

Cortes, C.; DeSalvo, G.; and Mohri, M. 2016. Learning with rejection. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory*, 67–82.

Devroye, L.; Györfi, L.; and Lugosi, G. 2013. *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Egan, J. P. 1975. *Signal Detection Theory and ROC-analysis*. New York: Academic press.

Ferri, C.; Hernández-Orallo, J.; and Flach, P. A. 2011. A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning*, 657–664.

Freund, Y.; Iyer, R.; Schapire, R. E.; and Singer, Y. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* 4:933–969.

Fumera, G., and Roli, F. 2002. Support vector machines with embedded reject option. In *International Workshop on Support Vector Machines*, 68–82. Springer.

Gao, W., and Zhou, Z.-H. 2015. On the consistency of AUC pairwise optimization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 939–945.

Gao, W.; Wang, L.; Jin, R.; Zhu, S.; and Zhou, Z.-H. 2016. One-pass AUC optimization. *Artificial Intelligence* 236:1–29.

Geifman, Y., and El-Yaniv, R. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, 4878–4887.

Golfarelli, M.; Maio, D.; and Malton, D. 1997. On the error-reject trade-off in biometric verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7):786–796.

Grandvalet, Y.; Rakotomamonjy, A.; Keshet, J.; and Canu, S. 2009. Support vector machines with a reject option. In *Advances in Neural Information Processing Systems 22*, 537–544.

Hamid, K.; Asif, A.; Abbasi, W.; Sabih, D.; et al. 2017. Machine learning with abstention for automated liver disease diagnosis. In *Proceedings of the 15th International Conference on Frontiers of Information Technology*, 356–361.

Hatami, N., and Chira, C. 2013. Classifiers with a reject option for early time-series classification. In *Proceedings of 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning*, 9–16.

Herbei, R., and Wegkamp, M. H. 2006. Classification with reject option. *Canadian Journal of Statistics* 34(4):709–721.

Herschtal, A., and Raskutti, B. 2004. Optimising area under the ROC curve using gradient descent. In *Proceedings of the 21st International Conference on Machine Learning*, 49.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, 377–384.

Liu, M.; Zhang, X.; Chen, Z.; Wang, X.; and Yang, T. 2018. Fast stochastic AUC maximization with o (1/n)-convergence rate. In *Proceedings of the 35th International Conference on Machine Learning*, 3195–3203.

Marinho, L. B.; Filho, P. P. R.; Almeida, J. S.; Souza, J. W. M.; Junior, A. H. S.; and de Albuquerque, V. H. C. 2018. A novel mobile robot localization approach based on classification with rejection option using computer vision. *Computers & Electrical Engineering* 68:26 – 43.

Metz, C. E. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 8(4):283–298.

Nadeem, M. S. A.; Zucker, J.-D.; and Hanczar, B. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, 65–81.

Natole, Jr., M.; Ying, Y.; and Lyu, S. 2018. Stochastic proximal algorithms for AUC maximization. In *Proceedings of the 35th International Conference on Machine Learning*, 3710–3719.

Nesterov, Y. 2003. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.

Pietraszek, T. 2005. Optimizing abstaining classifiers using ROC analysis. In *Proceedings of the 22nd International conference on Machine Learning*, 665–672.

Robertson, J. W.; Englehart, K. B.; and Scheme, E. J. 2018. Rejection of systemic and operator errors in a real-time myoelectric control task. In *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5640–5643. IEEE.

Rudin, C., and Schapire, R. E. 2009. Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research* 10:2193–2232.

Tortorella, F. 2005. A ROC-based reject rule for dichotomizers. *Pattern Recognition Letters* 26(2):167–180.

Yan, L.; Dodier, R. H.; Mozer, M.; and Wolniewicz, R. H. 2003. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the 20th International Conference on Machine Learning*, 848–855.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems 29*, 451–459.

Zhao, P.; Hoi, S. C.; Jin, R.; and Yang, T. 2011. Online AUC maximization. In *Proceedings of the 28th International Conference on Machine Learning*, 233–240.