Towards Decision-Friendly AUC: Learning Multi-Classifier with AUC$_\mu$

Peifeng Gao [1], **Qianqian Xu** [2, *], **Peisong Wen** [1, 2]
Huiyang Shao [1, 2], **Yuan He** [3], **Qingming Huang** [1, 2, 4, 5,*]
[1] School of Computer Science and Tech., University of Chinese Academy of Sciences
[2] Key Laboratory of Intelligent Information Processing Institute of Computing Technology, Chinese Academy of Sciences
[3] Alibaba Group
[4] BDKM, University of Chinese Academy of Sciences
[5] Peng Cheng Laboratory
{gaopeifeng21, shaohuiyang21}@mails.ucas.ac.cn, {xuqianqian, wenpeisong20z}@ict.ac.cn
heyuan.hy@alibaba-inc.com, qmhuang@ucas.ac.cn

Abstract

Area Under the ROC Curve (AUC) is a widely used ranking metric in imbalanced learning due to its insensitivity to label distributions. As a well-known multiclass extension of AUC, Multiclass AUC (MAUC, a.k.a. M-metric) measures the average AUC of multiple binary classifiers. In this paper, we argue that simply optimizing MAUC is far from enough for imbalanced multi-classification. More precisely, MAUC only focuses on learning scoring functions via ranking optimization, while leaving the decision process unconsidered. Therefore, scoring functions being able to make good decisions might suffer from low performance in terms of MAUC. To overcome this issue, we turn to explore AUC$_\mu$, another multiclass variant of AUC, which further takes the decision process into consideration. Motivated by this fact, we propose a surrogate risk optimization framework to improve model performance from the perspective of AUC$_\mu$. Practically, we propose a two-stage training framework for multi-classification, where at the first stage a scoring function is learned maximizing AUC$_\mu$, and at the second stage we seek for a decision function to improve the F1-metric via our proposed soft F1. Theoretically, we first provide sufficient conditions that optimizing the surrogate losses could lead to the Bayes optimal scoring function. Afterward, we show that the proposed surrogate risk enjoys a generalization bound in order of $\mathcal{O}(1/\sqrt{N})$. Experimental results on four benchmark datasets demonstrate the effectiveness of our proposed method in both AUC$_\mu$ and F1-metric.

Introduction

Over the past few decades, learning with imbalanced data has been attracting researchers' attention in the machine learning community (Japkowicz and Stephen 2002; Cárdenas and Baras 2006; He and Garcia 2009; Li, Chaudhuri, and Tewari 2016; Wang et al. 2021). In some real-world tasks like disease prediction tasks (Hao et al. 2020; Zhou et al. 2020) and rare event detection tasks (Liu et al. 2020a, 2018; Wu et al. 2020), the data distributions are largely skewed, *i.e.*, a few categories occupy the vast majority of observations, whereas the rest account for a small fraction. In this case, commonly used classification metrics like accuracy are not ideal choices to evaluate classifiers since they might ignore the important minority categories. Hence, evaluating the model performance in the case of imbalanced data is critical for imbalanced classification problems.

For binary classification tasks, Area Under the ROC Curve (AUC) has been widely used as one of the standard metrics since it is insensitivity to label distributions. Due to its importance, researchers have made efforts to direct AUC optimization (Rakotomamonjy 2004; Zhang, Saha, and Vishwanathan 2012; Ying, Wen, and Lyu 2016; Liu et al. 2020b). Since several real-world classification tasks involve more than two classes, it is natural to extend AUC to multiclass such that classifiers over imbalanced multiclass datasets can be measured and optimized similarly. More related work is presented in the Appendices.

One simple extension is *Multiclass AUC* (MAUC), which takes the average AUC between classes in a one-vs-one or one-vs-all manner (Hand and Till 2001; Yang et al. 2021a). Despite the success of the MAUC in imbalanced learning, we argue that it is not an ideal metric for imbalanced multi-classification. Specifically, multi-classification can be solved with two stages: *scoring* and *decision process*, where the former is predicting continuous score for each category, and the latter is mapping scores to discrete categories. MAUC focuses on the scoring, but leaves the decision process unconsidered. This might lead to inconsistency of MAUC w.r.t. the same decision results, *i.e.*, even if two scoring functions have same predictions, MAUC might be different, while another multiclass extension AUC$_\mu$(Kleiman and Page 2019) outputs consistent results.

As shown in Fig.1, given two examples $(\boldsymbol{x}_1, 1), (\boldsymbol{x}_2, 2)$ and a scoring function f with two components f^1, f^2, denote $\Delta_1 = f^1(\boldsymbol{x}_1) - f^1(\boldsymbol{x}_2), \Delta_2 = f^2(\boldsymbol{x}_1) - f^2(\boldsymbol{x}_2)$. For any $\tilde{f}$ with $\tilde{f}(\boldsymbol{x}_1) = f(\boldsymbol{x}_1) + b_1, \tilde{f}(\boldsymbol{x}_2) = f(\boldsymbol{x}_2) + b_2$, these two scoring functions output the same predictions, and $\Delta_{1,2}$ might be different with b_1, b_2 changing. However, MAUC measures Δ_1, Δ_2 independently, leading to different evaluation results for f and $\tilde{f}$. Since AUC$_\mu$ measures $\Delta margin = \Delta_1 - \Delta_2$ invariable w.r.t. b_1, b_2, it is able to avoid this issue. To this end, we turn to explore AUC$_\mu$ (Kleiman and Page 2019) which takes both scoring and

*Corresponding author.

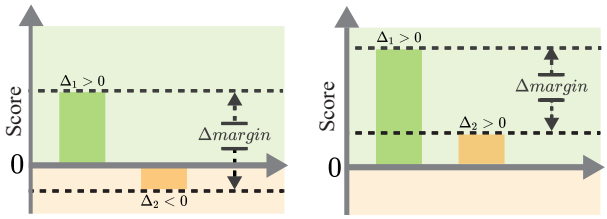


Figure 1: Illustration on inconsistency of MAUC w.r.t. same decisions. Left: f. Right: $\tilde{f}$ with $\tilde{f}(\boldsymbol{x}_1) = f(\boldsymbol{x}_1) + b_1, \tilde{f}(\boldsymbol{x}_2) = f(\boldsymbol{x}_2) + b_2$. We have $\text{MAUC}(f) = 1 \neq \text{MAUC}(\tilde{f}) = 0.5$, $\text{AUC}_\mu(f) = \text{AUC}_\mu(\tilde{f}) = 1$.

decision process into consideration. First of all, on top of Kleiman and Page's work, we further analyze AUC_μ under the imbalanced multi-classification setting. Specifically, we show that AUC_μ could efficiently avoid the imbalanced issue and the decision challenge by considering ranking pairs and sub-classifier pairs simultaneously. Therefore, AUC_μ is more reasonable to measure performance of scoring functions for imbalanced classification problems, and it is necessary to study direct optimization of AUC_μ.

Based on the above considerations, we propose to train a multi-classifier with two-stages: **1)** learn a scoring function with AUC_μ and **2)** search optimal thresholds.

At the first stage, we propose to directly optimize AUC_μ. The main challenge is that AUC_μ is non-differentiable, which makes gradient-based optimization methods infeasible. To overcome this problem, we investigate replacing the non-differentiable 0-1 loss with differentiable surrogate losses. We provide sufficient conditions that the surrogate losses are Fisher consistent with AUC_μ, *i.e.*, optimizing the surrogate risk will lead to Bayes optimal scoring function under the AUC_μ criterion. On top of this, we further propose an empirical surrogate risk of AUC_μ. In this way, the AUC_μ of models can be optimized without acknowledging the data prior. Additionally, we provide generalization error bounds to ensure the expected risk could be optimized.

At the second stage, we target to find a set of thresholds such that the macro F1-metric is maximized. Since the $argmax$ operation is non-differentiable, it is replaced by $softmax$, leading to a differentiable soft F1-metric. By optimizing the thresholds such that the soft F1-metric is minimized, we obtain a better decision function.

In short, the main contribution of this paper is three-fold:

- From the view of decision process, we show that $\text{AUC}_{\boldsymbol{\mu}}$ is a more ideal metric compared to MAUC since both the imbalance issue and the decision process are taken into account.
- We propose a two-stage training framework for imbalanced multi-classification. The key components include: learning a scoring function by optimizing $\text{AUC}_{\boldsymbol{\mu}}$, and searching a decision function with soft F1-metric.
- We propose an empirical surrogate risk minimization framework to directly optimize $\text{AUC}_{\boldsymbol{\mu}}$. Theoretically, we provide consistency analysis and generalization error bounds.

Learning Multi-Classifier with AUC_μ

In this section, we provide the formal definition of MAUC and AUC_μ, and then analyze the properties of both in multi-classification problems.

Preliminary

Notations Denote the sample space as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the feature space with d dimensions and $\mathcal{Y}$ is the label space. Under the context of multi-classification, $\mathcal{Y} = \{1, \ldots, N_C\}$, where N_C is the number of categories. Given two examples $\boldsymbol{z}_1 = (\boldsymbol{x}_1, y_1)$ and $\boldsymbol{z}_2 = (\boldsymbol{x}_2, y_2)$, we denote the proposition that $y_1 = i, y_2 = j$ or $y_1 = j, y_2 = i$ as $\varepsilon^{(i,j)}$. Denote a scoring function for multi-classification as $f \mapsto \mathbb{R}^{N_C}$, where the i-th component $f^i : \mathcal{X} \mapsto \mathbb{R}$ predicts the score of an instance being i-th category, and the hypothesis space of f^i is denoted as $\mathcal{F} = \left\{f^i : \mathcal{X} \mapsto \mathbb{R} \text{ is a measurable function}\right\}$. $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if argument $(\cdot)$ is true and returns 0 otherwise. We define the 0-1 loss function $\tilde{I}(x) = \mathbb{I}(x > 0) + \frac{1}{2}\mathbb{I}(x = 0)$.

Multiclass extensions of AUC The main idea of MAUC is to decompose the multi-class problem into several binary problems in a one-vs-one (ovo) or one-vs-all (ova) manner. As suggested in (Yang et al. 2021a), ovo is more comprehensive, thus we only discuss MAUC decomposed in an ovo manner. Formally, MAUC formulates multiclass AUC metric as the average AUC of each class pair (i, j):

$$\text{MAUC}(f) = \frac{1}{N_C(N_C-1)} \sum_{i}^{N_C} \sum_{j \neq i}^{N_C} \text{AUC}^{(i,j)}(f), \qquad (1)$$

where

$$\text{AUC}^{(i,j)}(f) = \mathop{\mathbb{E}}_{\boldsymbol{z}_1, \boldsymbol{z}_2} \left[\tilde{I}(f^i(\boldsymbol{x}_1) - f^i(\boldsymbol{x}_2)) \middle| \varepsilon^{(i,j)} \right].$$

Intuitively, $\text{AUC}^{(i,j)}(f)$ equals the probability that scores of positive examples are higher than negative ones, thus it is insensitive to the distribution skew.

Similarly, AUC_μ decomposes a multi-class problem in an ovo manner. The main difference from MAUC is the definition of binary AUC:

$$\text{AUC}_\mu(f) = \frac{1}{N_C(N_C-1)} \sum_{i}^{N_C} \sum_{j \neq i}^{N_C} \text{AUC}_\mu^{(i,j)}(f), \qquad (2)$$

where

$$\text{AUC}_\mu^{(i,j)}(f) = \mathop{\mathbb{E}}_{\boldsymbol{z}_1, \boldsymbol{z}_2} \left[\tilde{I}(\Delta margin(i, j, \boldsymbol{x}_1, \boldsymbol{x}_2, f)) \middle| \varepsilon^{(i,j)} \right],$$

and

$$\begin{aligned} &\Delta margin(i, j, \boldsymbol{x}_1, \boldsymbol{x}_2, f) \\ &= \left(f^i(\boldsymbol{x}_1) - f^i(\boldsymbol{x}_2)\right) - \left(f^j(\boldsymbol{x}_1) - f^j(\boldsymbol{x}_2)\right). \end{aligned}$$

More information about how it is defined could be found in (Kleiman and Page 2019).

Multi-classification Given a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a classifier could be viewed as a compound of a scoring function $f : \mathcal{X} \mapsto \mathbb{R}^{N_C}$ and a decision process $g : \mathbb{R}^{N_C} \mapsto \mathcal{Y}$. Typically, g could be implemented as an *argmax* operation:

$$g(f(\boldsymbol{x})) = \arg\max_{i \in [N_C]} f^i(\boldsymbol{x}).$$

AUC_μ in Decision Process

In this subsection, we provide detailed explanations that compared with MAUC, showing that both AUC_μ and MAUC could measure the ability of ranking. However in classification problems, AUC_μ is a more ideal metric considering the decision process. To begin with, we show the necessity of considering the decision process by the following example.

Example 1 (Decision bias of MAUC & AUC_μ)**.** *Given a dataset* $\{(\boldsymbol{x}_1, 1), (\boldsymbol{x}_2, 2)\}$ *and a scoring function* $f(\cdot)$ *with* $f(\boldsymbol{x}_1) = [0.9, 0.1]$, $f(\boldsymbol{x}_2) = [0.8, 0.2]$. *Obviously we have* $\text{MAUC}(f) = \text{AUC}_\mu(f) = 1$. *However, if implementing the decision process* g *as an argmax operation, then the predicted categories are* $g(f(\boldsymbol{x}_1)) = g(f(\boldsymbol{x}_2)) = 1$.

The above phenomenon stems from the fact that AUC only constrains the relative scores for example pairs. Therefore, even if we have an optimal scoring function with $\text{MAUC} = \text{AUC}_\mu(f) = 1$, the classifier might still fail when using the *argmax*. To avoid this issue, it is necessary to introduce thresholds into the decision process as follows.

Definition 1 (Decision process with thresholds)**.** *Given an* N_C*-class scoring function* f *and a set of thresholds* $\boldsymbol{\tau} = \{\tau^i \in \mathbb{R}\}_{i=1}^{N_C}$*, the decision function* g *is defined as*

$$g(f(\boldsymbol{x}); \boldsymbol{\tau}) = \arg\max_{i \in [N_C]} \left(f^i(\boldsymbol{x}) + \tau^i \right).$$

By introducing thresholds into the decision process, the decision bias can be addressed. For example, by setting $\boldsymbol{\tau} = \{0, 0.7\}$ in Example 1, we have $g(f(\boldsymbol{x}_1)) = 1, g(f(\boldsymbol{x}_2)) = 2$, where both $\boldsymbol{x}_1, \boldsymbol{x}_2$ are correctly classified.

Based on the above observation, we argue that the evaluation of scoring functions should take the above defined decision process into consideration. For the sake of the presentation, we propose two concepts to measure scoring functions as follows.

Definition 2 (Equivalent scoring function)**.** *Given two* N_C*-class scoring functions* $f, \tilde{f}$*, and a feature space* $\mathcal{X}$*, if for any thresholds* $\boldsymbol{\tau} \in \mathbb{R}^{N_C}$ *and* $\boldsymbol{x} \in \mathcal{X}$*, the following condition holds:*

$$g(f(\boldsymbol{x}); \boldsymbol{\tau}) = g(\tilde{f}(\boldsymbol{x}); \boldsymbol{\tau}),$$

then it is said that $\tilde{f}$ *is equivalent to* f.

Definition 3 (Optimal scoring function)**.** *Given an* N_C*-class scoring functions* f*, and a sample space* $\mathcal{Z}$*, if there exists a set of thresholds* $\boldsymbol{\tau} \in \mathbb{R}^{N_C}$*, such that the predictions are correct for any* $(\boldsymbol{x}, y) \in \mathcal{Z}$*:*

$$g(f(\boldsymbol{x}); \boldsymbol{\tau}) = y,$$

then it is said that f *is an optimal scoring function.*

Intuitively, if two scoring functions are equivalent, they always lead to same predictions no matter how the decision function is chosen. This motivates us to propose two principles for evaluating scoring functions from the decision perspective:

Claim 1. *An appropriate evaluation metric should:*

- *Output same results for equivalent scoring functions;*
- *Achieve score 1 for an optimal scoring function.*

However, it is easy to give an example that MAUC violates the above principles, while AUC_μ satisfies them.

Example 2 (Failure case of MAUC)**.** *Consider a dataset* $\{(\boldsymbol{x}_1, 1), (\boldsymbol{x}_2, 2), (\boldsymbol{x}_3, 3)\}$ *and two scoring functions* $f, \tilde{f}$ *with*

$$f(\boldsymbol{x}_1) = [0.30, 0.45, 0.25], \tilde{f}(\boldsymbol{x}_1) = [0.10, 0.25, 0.05],$$
$$f(\boldsymbol{x}_2) = [0.20, 0.50, 0.30], \tilde{f}(\boldsymbol{x}_2) = [0.20, 0.50, 0.30],$$
$$f(\boldsymbol{x}_3) = [0.31, 0.40, 0.29], \tilde{f}(\boldsymbol{x}_3) = [0.31, 0.40, 0.29].$$

Obviously, $\tilde{f}$ *is an optimal scoring function (by setting* $\boldsymbol{\tau} = [0.17, 0.01, 0.2]$*), and* $\tilde{f}$ *is equivalent to* f. *According to the definition we have* $\text{MAUC}(f) = 2/3$, $\text{MAUC}(\tilde{f}) = 0.5$. *As opposed to* MAUC*, we have* $\text{AUC}_\mu(f) = \text{AUC}_\mu(\tilde{f}) = 1$.

Besides the above intuitive explanation, we further provide a formal presentation that AUC_μ satisfies our principle in the following propositions. See Appendices for proofs.

Proposition 1. *Given two equivalent scoring function* $f, \tilde{f}$*, we have* $\text{AUC}_\mu(f) = \text{AUC}_\mu(\tilde{f})$.

Proposition 2. *Given an optimal scoring function* f*, we have* $\text{AUC}_\mu(f) = 1$.

In a nutshell, AUC_μ is more consistent with the prediction of a classifier.

Methodology

As analyzed in the last section, we can learn a classifier in a two-stage manner: **1)** train a scoring function f such that $\text{AUC}_\mu(f)$ is maximized; **2)** optimize the thresholds $\boldsymbol{\tau}$, such that the F1-metric of $g \circ f$ is maximized. In this section, we first propose a training framework for maximizing AUC_μ. Then we optimize Soft F1 to derive thresholds $\boldsymbol{\tau}$. By our proposed methods, one could get a classifier that has great ranking and classification performances simultaneously.

Maximization of AUC_μ

In this subsection, we focus on the minimization of expected risk $\text{AUC}_\mu^\downarrow = 1 - \text{AUC}_\mu$:

$$R(f) = \text{AUC}_\mu^\downarrow(f) = \frac{1}{N_C(N_C - 1)} \sum_{i=1}^{N_C} \sum_{j \neq i}^{N_C} \mathop{\mathbb{E}}_{\boldsymbol{z}_1, \boldsymbol{z}_2} \left[\tilde{I}(-\Delta margin(y_1, y_2, \boldsymbol{x}_1, \boldsymbol{x}_2, f)) \middle| \varepsilon^{(i,j)} \right]. \tag{3}$$

Surrogate Risk of $\text{AUC}_\mu^\downarrow$ Since the 0-1 loss $\tilde{I}(\cdot)$ is non-differentiable, gradient-based methods are infeasible to optimize the above objective function. To overcome this problem, we replace $\tilde{I}(\cdot)$ with a differentiable surrogate loss $\ell(\cdot)$, and construct the following surrogate risk:

$$R_\ell(f) = \frac{1}{N_C(N_C-1)} \sum_{i}^{N_C} \sum_{j\neq i}^{N_C} \mathop{\mathbb{E}}_{\boldsymbol{z}_1,\boldsymbol{z}_2} \left[\ell(\Delta margin(i,j,\boldsymbol{x}_1,\boldsymbol{x}_2,f)) \middle| \varepsilon^{(i,j)} \right].$$

However, this brings up another problem: can $R(f)$ be minimized by minimizing $R_\ell(f)$? In search of an answer to this problem, we introduce the concern of Fisher consistency.

Definition 4 (Consistency of $R_\ell(f)$)**.** *Surrogate loss $\ell(\cdot)$ is consistent with $\text{AUC}_\mu^\downarrow$ if for any distributions over sample space and any sequence $\{f_t\}_{t\in\mathbb{N}_+}$, the following condition holds:*

$$R(f_t) \to \inf_{f\in\mathcal{F}^{N_C}} R(f) \ , \ \textit{if} \ R_\ell(f_t) \to \inf_{f\in\mathcal{F}^{N_C}} R_\ell(f).$$

This definition gives conditions that a surrogate loss is helpful to original optimization problem. To investigate what kinds of surrogate loss are consistent, we first focus on the Bayes optimal scoring functions what kinds of scoring function f^* can minimize $R(f)$. Specifically, a scoring function f^* is a Bayes optimal scoring function of $R(f)$ if

$$f^* \in \arg\inf_{f\in\mathcal{F}^{N_C}} R(f).$$

And the following theorem provides sufficient conditions of Bayes optimal scoring functions:

Theorem 1 (**Bayes Optimal Scoring Functions**)**.** *A scoring function f is Bayes optimal if $\forall i \neq j \in \mathcal{Y}, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ s.t. $\eta_i(\boldsymbol{x}_1)\eta_j(\boldsymbol{x}_2) \neq \eta_i(\boldsymbol{x}_2)\eta_j(\boldsymbol{x}_1)$, we have*

$$\left[\eta_i(\boldsymbol{x}_1)\eta_j(\boldsymbol{x}_2) - \eta_j(\boldsymbol{x}_1)\eta_i(\boldsymbol{x}_2)\right] \Delta margin > 0,$$

where

$$\eta_i(x) := \mathop{\mathbb{P}}_{\boldsymbol{z}}(y=i|\boldsymbol{x}),$$
$$\Delta margin := \Delta margin(i,j,\boldsymbol{x}_1,\boldsymbol{x}_2,f).$$

Based on the above theorem, we provide a sufficient condition for consistency of surrogate losses as follows:

Theorem 2 (**Consistent Surrogate Loss of $\text{AUC}_\mu^\downarrow$**)**.** *A surrogate loss ℓ is consistent with $\text{AUC}_\mu^\downarrow$ if it is convex, differentiable at 0 and $\ell^{'}(0) < 0$.*

Remark 1. *Note that the derivation of the above theorem is non-trivial. Previous work (Gao and Zhou 2015; Yang et al. 2021a) provides proofs of binary AUC's and MAUC's consistency by contradiction. However, since AUC_μ involves relative scores between different components of the score function, we find it hard to follow existing techniques. Instead, we propose a new proof. See Appendices for details.*

From the above theorem, we find that most of widely-used surrogate loss functions are consistent with $\text{AUC}_\mu^\downarrow$.

Corollary 1. *These surrogate losses are consistent with $\text{AUC}_\mu^\downarrow$ according to $Thm.2$.*

Exp loss: $\ell(x) = \exp(-ax)$*;*
Square Loss: $\ell(x) = (1-x/a)^2$*;*
Generalized Hinge Loss:

$$\ell_\epsilon(x) = \begin{cases} 1-x & \textit{if } x \leq 1-\epsilon \\ (x-1-\epsilon)^2/4\epsilon & \textit{if } 1-\epsilon \leq x \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

Remark 2. *Note that Hinge Loss $\max(1-x,0)$ equals to Generalized Hinge Loss ℓ_ϵ by taking $\epsilon \to 0$, which means Hinge Loss is an asymptotically consistent surrogate loss.*

Empirical Risk Minimization So far, we have found many proper surrogate losses for surrogate risk minimization. Unfortunately, even with a proper surrogate loss, there is still a challenge to optimize $R_\ell(f)$. Concretely, the minimization of $R_\ell(f)$ is intractable due to the sample distribution is unavailable. This requires us to estimate $R_\ell(f)$ over a finite training set. The following proposition gives an unbiased estimation of $R_\ell(f)$ over a sample set S. The proof could be found in Appendices.

Proposition 3 (**Unbiased Empirical Estimation of $R_\ell(f)$**)**.** *Consider a sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where N is the number of samples. We denote $\{\boldsymbol{x}_j | y_j = i, (\boldsymbol{x}_j, y_j) \in S\}$ as S_i and $|S_i|$ as N_i. The unbiased estimation of $R_\ell(f)$ on S is given by*

$$\hat{R}_{\ell,S}(f) = \frac{1}{N_C(N_C-1)} \sum_{i=1}^{N_C} \sum_{j\neq i}^{N_C} \sum_{\boldsymbol{x}_1\in S_i} \sum_{\boldsymbol{x}_2\in S_j} \frac{\ell(\Delta margin(i,j,\boldsymbol{x}_1,\boldsymbol{x}_2,f))}{N_i N_j}.$$

This allows gradient-based optimization technologies to come into play. By plugging the above empirical risk into the standard optimization framework, we can obtain an approximately optimal scoring function in the sense of AUC_μ.

Decision Thresholds Learning

After learning a scoring function with the above framework, we still need to determine a set of decision thresholds $\boldsymbol{\tau}$ to form a classifier. The decision thresholds are expected to maximize the F1-metric corresponding to the decision results. Formally, given a fixed scoring function f and a training set S, we target to solve the following problem:

$$\max_{\boldsymbol{\tau}\in\mathbb{R}^{N_C}} F(f,\boldsymbol{\tau}) = \frac{1}{N_C} \sum_{i=1}^{N_C} \frac{1}{\left[Prec^i(f,\boldsymbol{\tau})\right]^{-1} + \left[Rec^i(f,\boldsymbol{\tau})\right]^{-1}} + \alpha\|\boldsymbol{\tau}\|_2^2,$$

where α is a hyperparameter to control the norm of $\boldsymbol{\tau}$, and

$$Prec^i(f,\boldsymbol{\tau}) = \frac{\sum_{\boldsymbol{z}\in S} \mathbb{I}\left(g(f(\boldsymbol{x});\boldsymbol{\tau}) = i\right) \mathbb{I}\left(y=i\right)}{\sum_{\boldsymbol{z}\in S} \mathbb{I}\left(g(f(\boldsymbol{x});\boldsymbol{\tau}) = i\right)},$$
$$Rec^i(f,\boldsymbol{\tau}) = \frac{\sum_{\boldsymbol{z}\in S} \mathbb{I}\left(g(f(\boldsymbol{x});\boldsymbol{\tau}) = i\right) \mathbb{I}\left(y=i\right)}{\sum_{\boldsymbol{z}\in S} \mathbb{I}\left(y=i\right)}.$$

However, optimizing the above target involves complicated discrete programming problems. To avoid this, we replace the non-differentiable $argmax$ operation with $softmax$, and the term $\mathbb{I}\left(g(f(\boldsymbol{x});\boldsymbol{\tau})=i\right)$ is softened to the i-th component of the predicted score $f(\boldsymbol{x})+\boldsymbol{\tau}$ after $softmax$:

$$\frac{\exp\left(\lambda\cdot\left(f^i(\boldsymbol{x})+\boldsymbol{\tau}^i\right)\right)}{\sum_j\exp\left(\lambda\cdot\left(f^j(\boldsymbol{x})+\boldsymbol{\tau}^j\right)\right)},$$

where λ is a tunable hyperparameter. In this way, we can optimize $\boldsymbol{\tau}$ with gradient-based methods.

Generalization Analysis

In the previous section, we have proposed a training framework to minimize the empirical risk of AUC_μ. In this section, we explore whether the expected risk could be optimized by minimizing the empirical surrogate risk with sufficient examples. Formally, we target to find out an upper bound of the gap

$$R(f)\leq R_\ell(f)\leq\hat{R}_{\ell.S}(f)+\epsilon(N),$$

such that the upper bound $\epsilon(N)\to 0$ when $N\to\infty$. By choosing surrogate loss $\ell\geq\tilde{l}$, the first inequality is obvious, thus we focus on the second inequality in the rest of this section. All details of proof could be found in Appendices.

Following the standard analysis framework of Probably Approximately Correct (PAC) learning, given fixed labels Y, we denote

$$\sup_f\left[\mathop{\mathbb{E}}_S\left[\hat{R}_{\ell,S}|\mathbf{Y}(S)=Y\right]-\hat{R}_{\ell,S}\right],$$

then we have the following conclusion:

Lemma 1. *Assume $dom(\ell)=[0,B]$, then $\forall f\in\mathcal{F}^{N_C},\delta\in(0,1)$, we have*

$$\mathop{\mathbb{P}}_S\Bigg[R_{\ell,S}(f)\geq\hat{R}_{\ell,S}(f)+\mathop{\mathbb{E}}_{S'}\left[\Phi(S')\middle|\mathbf{Y}(S')=Y\right]+\frac{2B}{N_C}\psi(Y)\sqrt{\frac{\log(1/\delta)}{2N}}\middle|\mathbf{Y}(S)=Y\Bigg]\leq\delta,$$

where $\psi(Y)=\sqrt{\sum_i\frac{N}{N_i}}$, $N_i=\sum_{y\in Y}\mathbb{I}(y=i)$.

The above lemma shows that the generalization error could be upper bounded by bounding

$$\mathop{\mathbb{E}}_{S'}\left[\Phi(S')\middle|\mathbf{Y}(S')=Y\right].$$

To this end, we introduce the concept of Rademacher Complexity (Mohri, Rostamizadeh, and Talwalkar 2018), which is a kind of complexity measure of hypothesis space. Unfortunately, the standard Rademacher Complexity is infeasible here since it required the risk to be expressed as a summation of independent terms, while the risk of AUC_μ is a pair-wise form which couples many examples. To handle this problem, similar to the pair-wise Rademacher Complexity of MAUC (Yang et al. 2021a), we propose the Rademacher Complexity of AUC_μ as follows:

Definition 5 (Rademacher Complexity of $\text{AUC}_\mu^\downarrow$)**.** *The empirical Rademacher Complexity of $\text{AUC}_\mu^\downarrow$ over the dataset S is*

$$\hat{\mathfrak{R}}_S\left[\ell\circ\mathcal{F}^{N_C}\right]=\frac{1}{N_C(N_C-1)}\mathop{\mathbb{E}}_\sigma\left[\sup_{f\in\mathcal{F}_{N_C}}\sum_i^{N_C}\sum_{j\neq i}^{N_C}\sum_{\boldsymbol{x}_m\in S_i}\sum_{\boldsymbol{x}_n\in S_j}T^{i,j,m,n}\right],$$

where

$$T^{i,j,m,n}=\frac{\sigma_i^m+\sigma_j^n}{2}\frac{\ell(\Delta margin(i,j,\boldsymbol{x}_m,\boldsymbol{x}_n,f))}{N_iN_j},$$

$\forall i\in[N_C],\{\sigma_i^m\}_{i,m}$ are i.i.d. Rademacher random variables taking -1 or 1 uniformly. Here $\boldsymbol{x}_m\in S_i$ represents that $\boldsymbol{x}_m$ is the m-th example of category i in dataset S.

With the above definition, $\mathbb{E}_{S'}\left[\Phi(S')\middle|\mathbf{Y}(S')=Y\right]$ can be bounded by the expected Rademacher Complexity of AUC_μ as shown in Lem.2. The proof follows techniques of (Usunier, Amini, and Gallinari 2005; Agarwal et al. 2005; Yang et al. 2021a).

Lemma 2 (**Symmetrization of Rademacher Complexity for $\text{AUC}_\mu^\downarrow$**)**.** *For sake of the presentation, we denote $\mathbf{Y}(S)$ as labels of dataset S. We have the the following condition holds:*

$$\mathop{\mathbb{E}}_S\left[\Phi(S)\middle|\mathbf{Y}(S)=Y\right]\leq 4\mathfrak{R}_Y\left[\ell\circ\mathcal{F}^{N_C}\right],$$

where $\mathfrak{R}_Y\left[\ell\circ\mathcal{F}^{N_C}\right]$ is the conditional expected Rademacher Complexity when the labels Y of dataset S is fixed, i.e.,

$$\mathfrak{R}_Y\left[\ell\circ\mathcal{F}^{N_C}\right]=\mathop{\mathbb{E}}_{S|Y}\hat{\mathfrak{R}}_S\left[\ell\circ\mathcal{F}^{N_C}\right].$$

By combining the Lem.1 and Lem.2, we get a close form to the final result. The last two steps are: 1). bound expected Rademacher Complexity by the empirical form. 2). replace the conditional expectation with the general expectation. The final generalization error bound is given by the following theorem:

Theorem 3 (**Generalization Error Bound of $\text{AUC}_\mu^\downarrow$**)**.** *Assume $dom(\ell)=[0,B]$, then $\forall f\in\mathcal{F}^{N_C},\delta\in(0,1)$, we have*

$$\mathop{\mathbb{P}}_S\Bigg[R_{\ell,S}(f)\geq\hat{R}_{\ell,S}(f)+4\hat{\mathfrak{R}}_S[\ell\circ\mathcal{F}^{N_C}]+\frac{10B}{N_C}\psi(Y)\sqrt{\frac{\log(2/\delta)}{2N}}\Bigg]\leq\delta.$$

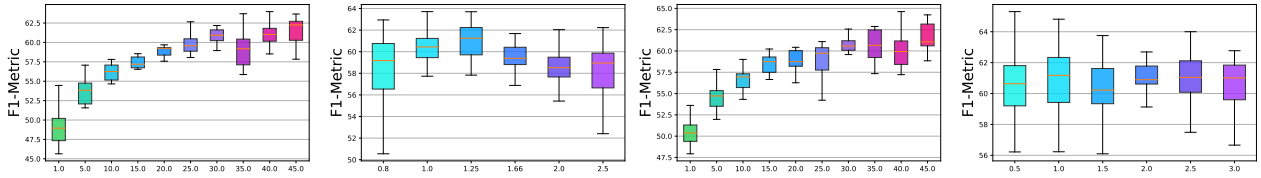
Remark 3. *According to the literature (Yang et al. 2021a; Golowich, Rakhlin, and Shamir 2018; Long and Sedghi 2019), the Rademacher Complexity could be bounded in order of $\mathcal{O}(1/\sqrt{N})$ for a wide range of models including deep neural networks. Therefore, we can draw the conclusion that the following fact holds with high probability:*

$$R_\ell(f)-\hat{R}_{\ell,S}(f)=\mathcal{O}(1/\sqrt{N}).$$

Remark 4. *From the above result, it can be seen that AUC_μ enjoys an imbalance-aware generalization bound like MUAC, showing both AUC_μ and MAUC metric are imbalance-friendly.*

Type	Method	CIFAR-10			CIFAR-100			TinyImageNet	ImageNet		
		50	100	200	50	100	200	100	50	100	200
Baseline	CE	93.90	93.01	92.44	93.20	92.32	91.42	89.54	96.40	95.52	94.97
Imbalanced methods	Focal	94.82	94.90	90.86	93.10	92.43	92.30	88.79	96.26	95.53	94.16
	CBFocal	95.48	94.40	91.16	92.85	92.71	89.88	88.72	96.04	94.90	94.46
	CBCE	94.44	93.99	92.25	92.48	91.94	91.63	89.87	96.68	95.91	94.78
	LDAM	94.49	94.60	92.36	92.38	92.75	91.88	90.68	96.66	95.85	94.58
	TL	94.99	94.45	92.83	93.40	91.15	92.20	89.49	96.69	96.00	94.92
	IHT	95.31	95.13	91.31	93.16	91.52	92.76	89.88	96.71	95.53	95.08
	NM	93.62	92.58	92.55	92.96	92.04	89.75	89.40	96.31	95.99	95.18
AUC-based losses	MAUC + Square	95.15	95.05	90.05	93.28	92.15	92.12	89.71	96.67	95.88	95.08
	MAUC + Exp	94.59	94.15	91.18	93.26	92.41	92.49	89.76	96.51	95.93	95.04
	MAUC + Hinge	93.82	93.96	91.94	93.06	92.11	91.70	89.36	96.42	95.71	95.18
	Ours + Square	95.68	94.54	91.81	**94.51**	**93.54**	**93.53**	**90.90**	**96.75**	95.99	95.13
	Ours + Exp	94.87	94.55	**93.28**	92.96	92.53	92.44	89.12	96.67	**96.02**	**95.30**
	Ours + Hinge	**95.68**	**95.61**	92.20	93.73	92.56	92.41	89.88	96.57	95.79	95.14

Table 1: Performance comparison on AUC_μ. The best results and the best competitors of each type are marked as **bold** and underline.



(a) Square Loss; K. (b) Square Loss; a. (c) Exp Loss; K. (d) Exp Loss; a.

Figure 2: Sensitivity analysis of our method on CIFAR-10 with an imbalance ratio of 50. (a) and (c): sensitivity of the warm-up epochs K. (b) and (d): sensitivity of the hyperparameter a in surrogate losses. More results could be found in Appendices.

Experiments

To demonstrate the effectiveness of our proposed framework, we conduct a series of experiments in four benchmark datasets for imbalanced multi-classification: CIFAR10, CIFAR100 (Krizhevsky 2012), TinyImageNet (Russakovsky et al. 2015) and ImageNet (Deng et al. 2009). These datasets are resampled with imbalance ratios of $50, 100, 200$.

Experimental Settings

Network Architecture For all experiments, we implement the scoring function as ResNet-18 trained from scratch. To normalize the scores into $[0, 1]$, we add a softmax function after the last linear layer of the model.

Competitors We compare popular methods for the imbalanced classification task. We set up standard Cross Entropy loss (**CE**) as a baseline. Other competitors are divided into two technical routes: **1) AUC-based loss functions** including MAUC (Yang et al. 2021a) and ours using square loss, exponential loss and hinge loss respectively; **2) imbalanced methods** including, Focal loss (**Focal**) (Lin et al. 2017), Class-Balanced CE loss (**CB-CE**), Class-Balanced Focal loss (Cui et al. 2019, **CB-Focal**), **LDAM** (Cao et al. 2019). Besides these loss functions, we also compare sampling-based methods including Instance Hardness Threshold sampling (Smith, Martinez, and Giraud-Carrier 2014, **IHT**), Near Miss sampling (Mani and Zhang 2003, **NM**) and Tomek Links sampling (Tomek 1976, **TL**). All sampling-based methods are coupled with the **CE** loss.

Training Strategy Limited by the space, we only present a part of key training settings, and more details are provided in Appendices. We utilize the Adam optimizer (Kingma and Ba 2017) for all methods. The initial learning rates are searched in $[10^{-4}, 10^{-3}]$, and decays by 0.99 per epoch. We keep the models with the highest AUC_μ in the validation set and report the corresponding AUC_μ and F1-metric on the test set. The training epochs are set to 25 for ImageNet and 80 for other datasets.

Experimental Results

In our experiments, evaluation metrics including AUC_μ, MAUC, precision, recall, and F1-metric are reported. Due to the limitation of the space, only test AUC_μ and F1-metric of all methods are shown in Tab.1 and Tab.2 respectively, and more results of other metrics are provided in Appendices. For our method, we report the F1-metric with and without

Type	Method	CIFAR-10			CIFAR-100			TinyImageNet	ImageNet		
		50	100	200	50	100	200	100	50	100	200
Baseline	CE	56.21	49.37	49.14	22.21	25.24	21.04	9.50	11.23	10.28	9.15
Imbalanced methods	Focal	60.39	58.26	50.23	26.92	23.40	16.91	10.72	9.40	7.85	6.41
	CBFocal	61.07	56.92	46.31	23.28	20.87	21.67	9.24	9.48	6.94	6.47
	CBCE	59.38	52.23	46.24	24.89	22.21	16.96	12.88	13.29	11.19	7.85
	LDAM	58.34	54.31	45.91	24.11	19.64	21.07	12.88	10.76	9.35	8.58
	TL	60.49	53.85	48.07	25.16	22.71	22.69	9.33	12.50	11.11	9.22
	IHT	60.25	58.70	44.68	22.83	22.59	19.85	11.61	13.85	10.20	10.12
	NM	47.32	42.31	35.11	23.25	25.86	19.18	9.91	10.54	10.55	8.62
AUC-based losses	MAUC + Square	61.03	57.79	46.01	27.28	22.96	22.40	11.69	14.43	12.51	11.19
	MAUC + Exp	58.32	51.29	48.06	24.92	25.56	22.85	13.94	13.88	12.62	11.02
	MAUC + Hinge	56.15	53.46	43.89	24.91	23.12	23.72	11.66	13.51	11.76	11.30
	Ours + Square	61.12	55.23	45.89	28.18	28.04	23.92	13.04	15.27	12.85	10.69
	Ours + Exp	59.66	52.55	**50.73**	29.42	26.92	21.78	11.71	15.04	12.89	11.96
	Ours + Hinge	60.68	59.42	45.82	29.45	23.34	18.48	14.35	14.94	13.14	11.64
	Ours* + Square	61.24	55.64	46.17	28.26	**28.08**	**23.95**	13.34	**15.39**	12.90	10.73
	Ours* + Exp	60.28	53.05	50.73	29.42	27.26	22.01	12.01	15.15	13.02	**12.07**
	Ours* + Hinge	**61.98**	**59.61**	46.41	**29.53**	23.53	18.64	**14.42**	15.08	**13.26**	11.74

Table 2: Performance comparison on F1-metric.

learning decision thresholds, denoted as **Ours*** and **Ours**, respectively. We have the following observations from Tab.1 and 2: **1)**. Our method shows great superiority in terms of both F1-metric and AUC$_\mu$, which validates the effectiveness of the proposed framework in promoting AUC$_\mu$. **2)**. Compared with MAUC in the perspective of F1, ours outperforms MAUC with a large margin in most cases. This supports our claim that AUC$_\mu$ taking scoring and decision thresholds into consideration simultaneously, thus learning with AUC$_\mu$ could significantly promote the classification performance. **3)**. The AUC-based methods are usually outperforms other competitors. We attribute this to insensitivity of AUC to label distributions and argue AUC-based methods are more suitable for imbalance learning.

Sensitivity Analysis

To study sensitivity of our method to hyperparameters, we conduct a series of empirical studies. Specifically, we study two key hyperparameters: the hyperparameter a controlling surrogate losses and the number of warm-up epochs K. We conduct experiments by grid searching, where K is set in $\{1, 5, 10, 15, 20, 25, 30, 35, 50, 45\}$, a is set in $\{0.8, 1.0, 1.25, 1.66, 2.0, 2.5\}$ for the Square loss and $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. Afterward, we study the effect of a hyperparameter by fixing the other one. The results are shown in Fig.2, where the horizontal and vertical axis represent the hyperparameters, the corresponding F1-Metric, respectively. The length of each box shows the variation of the corresponding parameter.

Effect of a For a of Square Loss, one can observe a clear trend from Fig.2 (b) that the classifier achieves optimal performance when a is around 1 and 1.25. For a of Exp Loss, the trend is not so clear. As Fig.2 (d) shown, when we tune a in $[0.5, 3.0]$, F1-metric do not show any obvious change. The difference comes from the gradient of Square Loss and Exp Loss. The gradient of Square Loss at 0 is very sensitive to a, so choosing the appropriate a is crucial for training. While it does not affect gradient of Exp Loss at 0 that much.

Effect of K According to Fig.2 (a) and (c), the number of warm-up epochs K has a huge impact on training. If K is too small, the classifier cannot be trained well. For both surrogate losses, increasing K from 1 to 30 makes a great improvement in terms of F1-metric. This demonstrates the importance of the warm-up phase for AUC training.

Conclusion

In this work, we argue that MAUC metric only considers the ranking of a classifier, leaving the classification decision process unconsidered. This leads to the phenomenon that a classifier enjoying a high MAUC may have poor classification ability. We turn to study another form of multi-class AUC, AUC$_\mu$. Through simple analysis, we find that AUC$_\mu$ is more consistent with the prediction of a classifier. Inspired by this, we propose an empirical surrogate risk minimization framework to optimize AUC$_\mu$ directly. The main challenge of optimizing AUC$_\mu$ is the 0-1 loss is not differentiable. To overcome this difficulty, we replace 0-1 loss with surrogate loss. Then, we find optimizing surrogate risk of AUC$_\mu$ using some widely-used surrogate losses can lead to Bayes optimal scorer. Moreover, we provide generalization analysis of our training framework. Finally, experiments on four datasets consistently show advantages of our methods.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: 62236008, U21B2038, 61931008, 6212200758 and 61976202, in part by the Fundamental Research Funds for the Central Universities, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000.

References

Agarwal, S.; Graepel, T.; Herbrich, R.; Har-Peled, S.; and Roth, D. 2005. Generalization Bounds for the Area Under the ROC Curve. *Journal of Machine Learning Research*, 6(14): 393–425.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. 32.

Cárdenas, A. A.; and Baras, J. S. 2006. B-ROC Curves for the Assessment of Classifiers over Imbalanced Data Sets. In *AAAI Conference on Artificial Intelligence*, 1581–1584.

Cortes, C.; and Mohri, M. 2003. AUC Optimization vs. Error Rate Minimization. 16.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Ferri, C.; Hernández-Orallo, J.; and Salido, M. A. 2003. Volume under the ROC Surface for Multi-Mlass Problems. In *European Conference on Machine Learning*, 108–120. Springer.

Gao, W.; and Zhou, Z.-H. 2015. On the Consistency of AUC Pairwise Optimization. In *AAAI Conference on Artificial Intelligence*.

Golowich, N.; Rakhlin, A.; and Shamir, O. 2018. Size-independent Sample Complexity of Neural Networks. In *Conference On Learning Theory*, 297–299. PMLR.

Hand, D. J.; and Till, R. J. 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. 45(2): 171–186.

Hao, H.; Fu, H.; Xu, Y.; Yang, J.; Li, F.; Zhang, X.; Liu, J.; and Zhao, Y. 2020. Open-Appositional-Synechial Anterior Chamber Angle Classification in AS-OCT Sequences. In *Medical Image Computing and Computer Assisted Intervention*, 715–724. ISBN 978-3-030-59721-4.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. 21(9): 1263–1284.

Herschtal, A.; and Raskutti, B. 2004. Optimising Area under the ROC Curve using Gradient Descent. In *International Conference on Machine learning*, 49.

Honzík, P.; Kučera, P.; Hynčica, O.; and Jirsík, V. 2009. Novel Method for Evaluation of Multi-Class Area under Receiver Operating Characteristic. In *International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, 1–4. IEEE.

Japkowicz, N.; and Stephen, S. 2002. The Class Imbalance Problem: A Systematic Study. 6(5): 429–449.

Kingma, D. P.; and Ba, J. 2017. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kleiman, R.; and Page, D. 2019. Aucμ: A Performance Metric for Multi-Class Machine Learning Models. In *International Conference on Machine Learning*, 3439–3447. PMLR.

Krizhevsky, A. 2012. Learning Multiple Layers of Features from Tiny Images. *University of Toronto*.

Lane, T. 2000. Extensions of ROC Analysis to Multi-Class Domains. In *ICML-2000 Workshop on Cost-Sensitive Learning, Standford*.

Li, B.; Chaudhuri, S.; and Tewari, A. 2016. Handling Class Imbalance in Link Prediction Using Learning to Rank Techniques. In *AAAI Conference on Artificial Intelligence*, volume 30.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *IEEE/CVF International Conference on Computer Vision*, 2980–2988.

Ling, C. X.; Huang, J.; Zhang, H.; et al. 2003. AUC: A Statistically Consistent and More Discriminating Measure than Accuracy. In *International Joint Conference on Artificial Intelligence*, volume 3, 519–524.

Liu, C.; Zhong, Q.; Ao, X.; Sun, L.; Lin, W.; Feng, J.; He, Q.; and Tang, J. 2020a. Fraud Transactions Detection via Behavior Tree with Local Intention Calibration. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3035–3043.

Liu, M.; Yuan, Z.; Ying, Y.; and Yang, T. 2020b. Stochastic AUC Maximization with Deep Neural Networks. In *International Conference on Learning Representations*.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection – A New Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6536–6545.

Long, P. M.; and Sedghi, H. 2019. Generalization Bounds for Deep Convolutional Neural Networks. In *International Conference on Learning Representations*.

Mani, I.; and Zhang, I. 2003. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*, volume 126, 1–7. International Conference on Machine Learning.

McClish, D. K. 1989. Analyzing a Portion of the ROC Curve. 9(3): 190–195.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of Machine Learning*. MIT press.

Mossman, D. 1999. Three-way Rocs. 19(1): 78–89.

Narasimhan, H.; and Agarwal, S. 2013. A Structural SVM Based Approach for Optimizing Partial AUC. In *International Conference on Machine Learning*, 516–524. PMLR.

Natole, M.; Ying, Y.; and Lyu, S. 2018. Stochastic Proximal Algorithms for AUC Maximization. In *International Conference on Machine Learning*, 3710–3719. PMLR.

Pepe, M. S.; and Thompson, M. L. 2000. Combining Diagnostic Test Results to Increase Accuracy. 1(2): 123–140.

Provost, F.; and Domingos, P. 2000. Well-Trained PETs: Improving Probability Estimation Trees.

Rakotomamonjy, A. 2004. Support Vector Machines and Area under ROC Curve.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet Large Scale Visual Recognition Challenge. 115(3): 211–252.

Smith, M. R.; Martinez, T.; and Giraud-Carrier, C. 2014. An Instance Level Analysis of Data Complexity. 95(2): 225–256.

Tomek, I. 1976. Two Modifications of CNN. SMC-6(11): 769–772.

Usunier, N.; Amini, M.-R.; and Gallinari, P. 2005. A Data-Dependent Generalisation Error Bound for the AUC. In *ICML Workshop on ROC Analysis in Machine Learning*.

Wang, L.; Xu, S.; Wang, X.; and Zhu, Q. 2021. Addressing Class Imbalance in Federated Learning. In *AAAI Conference on Artificial Intelligence*, volume 35, 10165–10173.

Wang, Z.; and Chang, Y.-C. I. 2011. Marker Selection via Maximizing the Partial Area under the ROC Curve of Linear Risk Scores. 12(2): 369–385.

Wu, H.; Hu, Z.; Jia, J.; Bu, Y.; He, X.; and Chua, T.-S. 2020. Mining Unfollow Behavior in Large-scale Online Social Networks via Spatial-Temporal Interaction. In *AAAI Conference on Artificial Intelligence*, volume 34, 254–261.

Yang, B. 2009. The Extension of the Area under the Receiver Operating Characteristic Curve to Multi-Class Problems. In *ISECS International Colloquium on Computing, Communication, Control, and Management*, volume 2, 463–466. IEEE.

Yang, Z.; Xu, Q.; Bao, S.; Cao, X.; and Huang, Q. 2021a. Learning with Multiclass AUC: Theory and Algorithms.

Yang, Z.; Xu, Q.; Bao, S.; He, Y.; Cao, X.; and Huang, Q. 2021b. When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC. In *International Conference on Machine Learning*, 11820–11829. PMLR.

Yang, Z.; Xu, Q.; Bao, S.; He, Y.; Cao, X.; and Huang, Q. 2022. Optimizing Two-way Partial AUC with an End-to-end Framework.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic Online AUC Maximization. 29.

Yuan, Z.; Yan, Y.; Sonka, M.; and Yang, T. 2021. Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. In *IEEE/CVF International Conference on Computer Vision*, 3040–3049.

Zhang, X.; Saha, A.; and Vishwanathan, S. 2012. Smoothing Multivariate Performance Measures. 13(1): 3623–3680.

Zhou, K.; Gao, S.; Cheng, J.; Gu, Z.; Fu, H.; Tu, Z.; Yang, J.; Zhao, Y.; and Liu, J. 2020. Sparse-GAN: Sparsity-constrained Generative Adversarial Network for Anomaly Detection in Retinal OCT Image. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 1227–1231. IEEE.