AUC Optimization from Multiple Unlabeled Datasets

Zheng Xie[*], **Yu Liu**[*], **Ming Li**[†]

National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{xiez, liuy, lim}@lamda.nju.edu.cn

Abstract

Weakly supervised learning aims to make machine learning more powerful when the perfect supervision is unavailable, and has attracted much attention from researchers. Among the various scenarios of weak supervision, one of the most challenging cases is learning from multiple unlabeled (U) datasets with only a little knowledge of the class priors, or U^m learning for short. In this paper, we study the problem of building an AUC (area under ROC curve) optimal model from multiple unlabeled datasets, which maximizes the pairwise ranking ability of the classifier. We propose U^m-AUC, an AUC optimization approach that converts the U^m data into a multi-label AUC optimization problem, and can be trained efficiently. We show that the proposed U^m-AUC is effective theoretically and empirically.

1 Introduction

Since obtaining perfect supervision is usually challenging in real-world machine learning problems, the machine learning approaches often have to deal with inaccurate, incomplete, or inexact supervisions, collectively referred to as weak supervision (Zhou 2017). To achieve this, many researchers have devoted into the area of *weakly supervised learning*, such as semi-supervised learning (Zhu et al. 2009), positive-unlabeled learning (Bekker and Davis 2020), noisy label learning (Han et al. 2021), etc.

Among multiple scenarios of weakly supervised learning, one of the most challenging scenarios is to learn classifiers from m unlabeled (U) datasets with different class priors, i.e., the proportions of positive instances in the sets. Such a learning task is usually referred to as U^m learning. This scenario usually occur when the instances can be categorized into different groups, and the probability of an instance to be positive varies across the groups, e.g., for predicting voting rates or morbidity rates. Prior studies include Scott and Zhang (2020), which ensembles the classifiers trained on all pairs of the unlabeled sets; Tsai and Lin (2020), which introduces consistency regularization for the problem. Recently, Lu et al. (2021) proposed a consistent approach for classification from multiple unlabeled sets, which is the first classifier-consistent approach for learning from m unlabeled sets ($m > 2$) that optimizes a classification loss.

In this paper, we further consider the problem of learning an AUC (area under ROC curve) optimization model from the U^m data, which maximizes the pairwise ranking ability of the classifier (Hanley and McNeil 1982). The importance of this problem lie in two folds: First, we note that for certain scenarios, the ranking performance of the model is more important. E.g., ranking items with coarse-grind rank labels. Second, given multiple U sets with different class priors, the imbalance issue is very likely to affect the learning process. Thus, taking an imbalance-aware performance measure, i.e., AUC, is naturally appropriate for the problem.

To achieve this goal, we introduce U^m-AUC, a novel AUC optimization approach from U^m data. U^m-AUC solves the problem as a multi-label AUC optimization problem, as each label of the multi-label learning problem corresponds to a pseudo binary AUC optimization sub-problem. To overcome the quadratic time complexity of the pairwise loss computation, we convert the problem and solve it through a pointwise AUC optimization algorithm. Our theoretical analysis shows that U^m-AUC is consistent with the optimal AUC optimization model, and provides the generalization bound. Experiments show that our approach outperforms the state-of-the-art methods and has superior robustness.

Our main contributions are highlighted as follows:

- To the best of our knowledge, we present the first algorithm for optimizing AUC in U^m scenarios. It is also the first one to address the U^m problem without relying on exact class priors. Importantly, our algorithm is simple yet efficient.
- Furthermore, we conduct a comprehensive theoretical analysis of the proposed methodology, demonstrating its validity and assessing its excess risk.
- We perform experiments on various settings using multiple benchmark datasets, and the results demonstrate that our proposed method consistently outperforms the state-of-the-art methods and performs robustly under different imbalance settings.

The reminder of our paper is organized as follows. We first introduce preliminary in section 2. Then, we introduce the U^m-AUC approach and conducts the theoretical analysis in section 3, while section 4 shows the experimental results.

[*]These authors contributed equally.
[†]Ming Li is the corresponding author.

Finally, section 5 introduces the related works, and section 6 concludes the paper.

2 Preliminary

In the fully supervised AUC optimization, we are given a dataset sampled from a specific distribution

$$\mathcal{X}_L := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)\,. \tag{1}$$

For convenience, we refer to positive and negative data as samples from two particular distributions:

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = +1)\,, \text{ and}$$
$$\mathcal{X}_N := \{\boldsymbol{x}'_j\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_N(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = -1)\,,$$

there we have $\mathcal{X}_L = \mathcal{X}_P \cup \mathcal{X}_N$.

Let $f : \mathcal{X} \to \mathcal{R}$ be a scoring function. It is expected that positive instances will have a higher score than negative ones. For a threshold value t, we define the true positive rate $\mathrm{TPR}_f(t) = \Pr(f(x) \geq t | y = 1)$ and the false positive rate $\mathrm{FPR}_f(t) = \Pr(f(x) \geq t | y = 0)$. The AUC is defined as the area under the ROC curve:

$$\mathrm{AUC} = \int_{-\infty}^{+\infty} \mathrm{TPR}_f(\mathrm{FPR}_f^{-1}(t))dt. \tag{2}$$

Previous study (Hanley and McNeil 1982) introduced that the AUC equals the probability that a randomized positive instance scores higher than a randomized negative instance score, so that the AUC of the model f can be formulated as:

$$\mathrm{AUC} = 1 - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_P(\boldsymbol{x})}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_N(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x}) - f(\boldsymbol{x}'))]\big]\,. \tag{3}$$

Here $\ell_{01}(z) = \mathbb{I}[z < 0]$. Without creating ambiguity, we will denote $f(\boldsymbol{x}, \boldsymbol{x}')$ as $f(\boldsymbol{x}) - f(\boldsymbol{x}')$ for clarity.

The maximization of the AUC is equivalent to the minimization of the following AUC risk. Since the true AUC risk measures the error rate of ranking positive instances over negative instances, we refer to the true AUC risk as the PN-AUC risk to avoid confusion:

$$R_{PN}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_P(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_N(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))]\right]\,. \tag{4}$$

With a finite sample, we typically solve the following empirical risk minimization (ERM) problem:

$$\min_f \quad \hat{R}_{PN}(f) = \frac{1}{|\mathcal{X}_P||\mathcal{X}_N|} \sum_{\boldsymbol{x} \in \mathcal{X}_P} \sum_{\boldsymbol{x}' \in \mathcal{X}_N} \ell(f(\boldsymbol{x}, \boldsymbol{x}'))\,. \tag{5}$$

3 U^m-AUC: The Method

In this paper, we study AUC optimization under U^m setting, which involves optimizing AUC across multiple unlabeled datasets. Suppose we are given $m (m \geq 2)$ unlabeled datasets $\mathrm{U}_1, \ldots, \mathrm{U}_m$ with different class prior probabilities, as defined by the following equation:

$$\mathrm{U}_i = \{x_{ik}\}_{k=1}^{n_i} \overset{\text{i.i.d.}}{\sim} p_i(x) = \pi_i p_P(x) + (1 - \pi_i) p_N(x)\,, \tag{6}$$

where $p_P(x)$ and $p_N(x)$ are the positive and negative class-conditional probability distributions, respectively, and π_i denotes the class prior of the i-th unlabeled set. The size of U_i is n_i. Although we only have access to unlabeled data, our objective is to build a classifier that minimizes the PN-AUC risk eq. (4).

To achieve this goal, we propose U^m-AUC, a novel AUC optimization approach that learns from U^m data. Unlike the previous studies on U^m classification who require the knowledge of the class prior (Lu et al. 2021), U^m-AUC replaces this requirement by only knowing knowledge of the relative order of the unlabeled sets based on their class priors, which is more realistic.

We next introduce the U^m-AUC approach. For convenience, and without loss of generality, we assume that the class priors of the unlabeled sets are in descending order, i.e., $\pi_i \geq \pi_j$ for $i < j$. Additionally, we assume that at least two unlabeled sets have different priors, i.e., $\pi_1 > \pi_m$; otherwise, the problem is unsolvable.

Consistent AUC Learning from U^m Data

To provide a solution that is consistent with the true AUC, we first introduce the two sets case, i.e., one can achieve consistent AUC learning through two unlabeled sets with different class priors. Such a result is discussed in previous studies (Charoenphakdee, Lee, and Sugiyama 2019). Suppose the two unlabeled sets are U_i and U_j with $\pi_i > \pi_j$. We can minimize the following U^2 AUC risk:

$$R_{ij}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_i(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_j(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))]\right] \tag{7}$$

by solve the following U^2 AUC ERM problem:

$$\min_f \quad \hat{R}_{ij}(f) = \frac{1}{n_i n_j} \sum_{\boldsymbol{x} \in \mathrm{U}_i} \sum_{\boldsymbol{x}' \in \mathrm{U}_j} \ell(f(\boldsymbol{x}, \boldsymbol{x}'))\,. \tag{8}$$

The following theorem shows that the U^2 AUC risk minimization problem is consistent with the original AUC optimization problem we need to solve, i.e., we can solve the original AUC optimization problem by minimizing the U^2 AUC risk.

Theorem 1 (U^2 AUC consistency)**.** *Suppose f^* is a minimizer of the AUC risk R_{ij} over two distributions p_i and p_j where $\pi_i > \pi_j$, i.e., $f^* = \arg\min R_{ij}$. Then, it follows that f^* is also a minimizer of the true AUC risk R_{PN} , and thus, R_{ij} is consistent with R_{PN}.*

Therefore, by minimizing the U^2 AUC risk under the condition that only impure data sets are available, we can obtain the desired model.

With the U^m data, we can construct the following minimization problem by composing $m(m-1)/2$ AUC subproblems with weights $z_{ij} > 0$:

$$\min_f \quad R_{U^m}(f) = \sum_{i,j|1 \leq i < j \leq m} z_{ij} R_{ij}(f) \tag{9}$$

which corresponds to the U^m AUC ERM problem

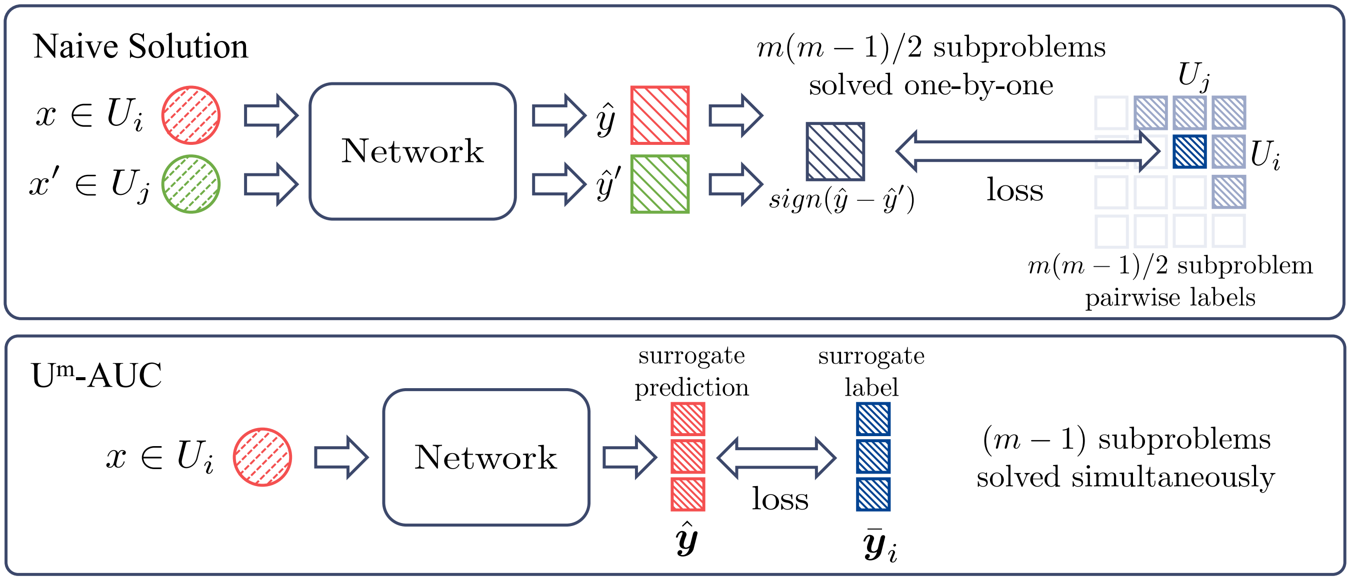

Figure 1: Framework Demonstration. The U^m-AUC employs an efficient stochastic optimization algorithm, eliminating the need for pairwise loss and reducing the time complexity to $O(n)$. Additionally, it simplifies the naive solution by transforming it into a multi-label learning problem, reducing the number of sub-problems and resulting in a more concise problem formulation.

$$\min_f \quad \hat{R}_{U^m}(f) = \sum_{i,j|1\leq i<j\leq m} \sum_{\boldsymbol{x}\in \mathrm{U}_i} \sum_{\boldsymbol{x}'\in \mathrm{U}_j} \frac{z_{ij}\ell(f(\boldsymbol{x},\boldsymbol{x}'))}{n_i n_j} . \tag{10}$$

As well, we can theoretically demonstrate the consistency between U^m AUC risk minimization problem and the original AUC optimization problem.

Theorem 2 (U^m AUC consistency)**.** *Suppose f^* is a minimizer of the AUC risk R_{U^m} over m distributions $p_1, \cdots, p_m$ where $\pi_i > \pi_j$ for $i < j$, i.e., $f^* = \arg\min R_{U^m}$. Then, it follows that f^* is also a minimizer of the true AUC risk R_{PN}, and thus, R_{U^m} is consistent with R_{PN}.*

Similarly, the desired model can be obtained by optimizing the U^m AUC risk minimization problem under the condition that only impure data sets are available. This means that the ERM in eq. (10) provides a naive solution for U^m AUC risk minimization by solving all the $m(m-1)/2$ AUC sub-problems using a pairwise surrogate loss based on the definition of the AUC score (Gao and Zhou 2015), and minimizing the loss over all instance pairs that belong to different U sets.

However, such a solution can be complex and inefficient, especially when dealing with a large number of datasets and a huge amount of data in each dataset. For instance, with m datasets, we need to handle $m(m-1)/2$ sub-problems according to the definition of U^m AUC risk, which can be complex when m grows. Furthermore, assuming that the number of samples is n, if a pairwise loss is used to optimize each sub-problem, the time complexity of each epoch in training is $O(n^2)$. This means that the time consumption of the method grows quadratically with n, making it computationally infeasible for large-scale datasets. To address the aforementioned issues, we propose a novel and efficient training algorithm for U^m AUC risk minimization.

U^m-AUC: Simple and Efficient Learning for U^m AUC Risk Minimization

To simplify the form of naive solution to U^m AUC risk minimization, we transform it into an equivalent multi-label learning problem, reducing the number of sub-problems to $m-1$. To decrease the time cost of training the model, we use an efficient stochastic optimization algorithm, reducing the time complexity from $O(n^2)$ to $O(n)$. The proposed approach is demonstrated in the fig. 1.

Reduction in the Number of Sub-problems To reduce the number of sub-problems, we transform the U^m AUC risk minimization problem into a multi-label learning problem with $m-1$ labels. We let the samples in datasets $\mathrm{U}_1, \cdots, \mathrm{U}_k$ have label 1 at the k-th position and let the samples in datasets $\mathrm{U}_{k+1}, \cdots, \mathrm{U}_m$ have label 0 at the k-th position.

Specifically, we assign the surrogate label $\bar{\boldsymbol{y}}^{(k)}$ be the label of the k-th unlabeled set U_k, where

$$\bar{\boldsymbol{y}}^{(k)} = [\underbrace{0, 0, \ldots, 0}_{k-1}, \underbrace{1, 1, \ldots, 1}_{m-k}] \tag{11}$$

has $k-1$ negative labels in the front and $m-k$ positive labels in the rear.

Let $\boldsymbol{g}(x) = \hat{\boldsymbol{y}}$ be the model output score for the multi-label learning problem, and $\boldsymbol{g}_k(x)$ be the k-th dimension of $\boldsymbol{g}(x)$, which denote the output of k-th sub-problem, the multi-label learning problem can be formalized as:

$$\max_{\boldsymbol{g}} \quad \mathrm{AUC}_{\mathrm{macro}}(\boldsymbol{g}) = \frac{1}{m-1} \sum_{k=1,2,\cdots,m} \mathrm{AUC}_k(\boldsymbol{g}_k) , \tag{12}$$

and AUC_k is AUC on the k-th label, which has

$$\text{AUC}_k(\boldsymbol{g}_k) = 1 - \sum_{\boldsymbol{x}\in\bigcup_{i\le k}\text{U}_i} \sum_{\boldsymbol{x}'\in\bigcup_{j>k}\text{U}_j} \frac{\ell(\boldsymbol{g}_k(\boldsymbol{x},\boldsymbol{x}'))}{\sum_{i\le k} n_i \sum_{j>k} n_j}. \tag{13}$$

For the k-th label, the sub-problem of the multi-label learning problem is a simple AUC optimization problem:

$$\min_{\boldsymbol{g}_k} \quad \frac{1}{m-1} \sum_{\boldsymbol{x}\in\bigcup_{i\le k}\text{U}_i} \sum_{\boldsymbol{x}'\in\bigcup_{j>k}\text{U}_j} \frac{\ell(\boldsymbol{g}_k(\boldsymbol{x},\boldsymbol{x}'))}{\sum_{i\le k} n_i \sum_{j>k} n_j}. \tag{14}$$

That is, in order to solve the multi-label learning problem described in eq. (12), we can optimize $m-1$ sub-problems of the form described in eq. (14) only. The following explanation outlines why the U^m-AUC problem eq. (9) can be solved by solving this multi-label learning problem eq. (12).

Let $r_{ijk} = (n_i n_j)/(\sum_{i\le k} n_i \sum_{j>k} n_j)$, the optimization problem eq. (14) is equivalent to

$$\min_{\boldsymbol{g}_k} \quad \frac{1}{m-1} \sum_{i\le k}\sum_{j>k} r_{ijk} \sum_{\boldsymbol{x}\in\text{U}_i} \sum_{\boldsymbol{x}'\in\text{U}_j} \frac{\ell(\boldsymbol{g}_k(\boldsymbol{x},\boldsymbol{x}'))}{n_i n_j}, \tag{15}$$

or simplified as:

$$\min_{\boldsymbol{g}_k} \quad \frac{1}{m-1} \sum_{i\le k}\sum_{j>k} r_{ijk} \hat{R}_{ij}(\boldsymbol{g}_k). \tag{16}$$

This is exactly the special case of eq. (10)'s where $z_{ij} = r_{ijk}/(m-1) > 0$. Therefore, each sub-problem of multi-label learning problem eq. (12) is an ERM problem for the U^m AUC risk minimization problem eq. (9). According to theorem 2, optimizing this multi-label learning problem is equivalent to solving the original AUC optimization problem. Thus, we aggregate the output of sub-problem as $f = \frac{1}{m-1}\sum_{k=1}^{m-1} \boldsymbol{g}_k$.

In summary, by transforming the U^m AUC risk minimization problem into a multi-label learning problem, we only need to optimize $m-1$, rather than $m(m-1)/2$ sub-problems as in the naive approach.

Efficient Training of the Model Although we have reduced the number of the sub-problems from $O(m^2)$ to $O(m)$, this approach may not be practical for large datasets when optimizing the pairwise loss on training data, since the pairwise method suffers from severe scalability issue, as each epoch will take $O(n_P \cdot n_N)$ time with n_P positive and n_N negative samples. This issue has been discussed and efficient methods for AUC optimization have been proposed in several previous works like (Yuan et al. 2021; Liu et al. 2020). These method will take only $O(n_P + n_N)$ time each epoch, making them more suitable for large-scale datasets.

Consider using the square surrogate AUC loss in the multi-label learning problem eq. (12), with the derivation process shown in Appendix, we get

$$\begin{aligned}
&\frac{1}{m-1}\sum_{1\le k<m}\sum_{\boldsymbol{x}\in\bigcup_{i\le k}\text{U}_i}\sum_{\boldsymbol{x}'\in\bigcup_{j>k}\text{U}_j}\frac{(1-f(\boldsymbol{x})+f(\boldsymbol{x}'))^2}{\sum_{i\le k} n_i \sum_{j>k} n_j}\\
&= \frac{1}{m-1}\sum_{1\le k<m}\Big(\underbrace{\sum_{\boldsymbol{x}\in\bigcup_{i\le k}\text{U}_i}\frac{(f(\boldsymbol{x})-a_k(f))^2}{\sum_{i\le k} n_i}}_{A_k(f)}\\
&\quad + \underbrace{\sum_{\boldsymbol{x}'\in\bigcup_{j>k}\text{U}_j}\frac{(f(\boldsymbol{x}')-b_k(f)))^2}{\sum_{j>k} n_j}}_{B_k(f)} + \underbrace{(1-a_k(f)+b_k(f))^2}_{C_k(f)}\Big)\\
&= \frac{1}{m-1}\sum_{1\le k<m}(A_k(f)+B_k(f)\\
&\qquad + \max_{\alpha_k}\{2\alpha_k(1-a_k(f)+b_k(f))-\alpha_k^2\})
\end{aligned} \tag{17}$$

where $a_k(f) = \sum_{\boldsymbol{x}\in\bigcup_{i\le k}\text{U}_i} f(\boldsymbol{x})/\sum_{i\le k} n_i$, and $b_k(f) = \sum_{\boldsymbol{x}'\in\bigcup_{j>k}\text{U}_j} f(\boldsymbol{x}')/\sum_{j>k} n_j$.

Following previous work (Ying, Wen, and Lyu 2016), the objective eq. (17) is equivalent to $(m-1)$ min-max problems

$$\min_{f,a_k,b_k}\max_{\alpha_k} h(f,a_k,b_k,\alpha_k) := \mathbb{E}_{\boldsymbol{z}}[H(f,a_k,b_k,\alpha_k;\boldsymbol{z})], \tag{18}$$

where $\boldsymbol{z} = (\boldsymbol{x}, y)$ is a random sample, and

$$\begin{aligned}
H(f,a_k,b_k,\alpha_k;\boldsymbol{z}) =&(1-p)(f(\boldsymbol{x})-a_k)^2\mathbb{I}_{[y=1]}\\
&+ p(f(x)-b_k)^2\mathbb{I}_{[y=-1]}\\
&- p(1-p)\alpha_k^2\\
&+ 2\alpha_k(p(1-p)+pf(\boldsymbol{x})\mathbb{I}_{[y=-1]}\\
&\qquad - (1-p)f(\boldsymbol{x})\mathbb{I}_{[y=1]}),
\end{aligned} \tag{19}$$

with $p = \sum_{i\le k} n_i/(\sum_{i\le k} n_i + \sum_{j>k} n_j)$.

Besides, we replace the C_k with $\max_{\alpha_k\ge 0}\{2\alpha_k(m - a_k(f) + b_k(f)) - \alpha_k^2\}$, which has a margin parameter m to make the loss function more robust (Yuan et al. 2021).

The min-max problems can be efficiently solved using primal-dual stochastic optimization techniques, eliminating the need for explicit construction of positive-negative pairs. In our implementation, we leverage the Polyak-Łojasiewicz (PL) conditions (Guo et al. 2020) of the objective functions in the min-max problems eq. (18), and update the parameters accordingly to solve the multi-label learning problem.

Through a combination of equivalence problem conversion techniques and efficient optimization methods, the complexity of each epoch in training can be reduced from $O(n^2)$ to $O(n)$. The algorithm is described in algorithm 1.

Theoretical Analysis

In this subsection, we provide a theoretical analysis of the approach described above. Specifically, we prove excess risk bounds for the ERM problem of U^2 AUC and U^m AUC.

Let $\mathcal{X}$ be the feature space, K be a kernel over $\mathcal{X}^2$, and C_w be a strictly positive real number. Let $\mathcal{F}_K$ be a class of

Algorithm 1 U^m-AUC

Input: Model $\boldsymbol{g}$, m sets of unlabeled data with class priors in descending order $U_1, \cdots, U_m$, training epochs *num_epochs*, number of batches *num_batches*.

for $t = 1, 2, \ldots,$ *num_epochs* **do**
  for $b = 1, 2, \ldots,$ *num_batches* **do**
    Fetch mini-batch $\mathcal{B}$ from $\bigcup_{0 \leq i \leq m} \text{U}_i$
    Forward $\mathcal{B}$ and get $\boldsymbol{g}(\mathcal{B})$
    Compute multi-label loss of the mini-batch $\mathcal{B}$
    Update the parameters of $\boldsymbol{g}$
  end for
end for
Aggregate the $\boldsymbol{g}$ by $f = \frac{1}{m-1} \sum_{k=1}^{m-1} \boldsymbol{g}_k$

Output: f

functions defined as:

$$\mathcal{F}_K = \{ f_w : \mathcal{X} \to R, f_w(x) = K(w, x) | \|w\|_k \leq C_w \},$$

where $\|x\|_K = \sqrt{K(x, x)}$. We also assume that the surrogate loss ℓ is L-Lipschitz continuous, bounded by a strictly positive real number C_ℓ, and satisfies inequality $\ell \geq \ell_{01}$.

Let $\hat{f}_{ij}^*$ be the minimizer of empirical risk $\hat{R}_{ij}(f)$, we introduce the following excess risk bound, showing that the risk of $\hat{f}_{ij}^*$ converges to risk of the optimal function in the function family $\mathcal{F}_K$.

Theorem 3 (Excess Risk for U^2 AUC ERM problem)**.** *Assume that $\hat{f}_{ij}^* \in \mathcal{F}_K$ is the minimizer of empirical risk $\hat{R}_{ij}(f)$, $f_{PN}^* \in \mathcal{F}_K$ is the minimizer of true risk $R_{PN}(f)$. For any $\delta > 0$, with the probability at least $1 - \delta$, we have*

$$R_{PN}(\hat{f}_{ij}^*) - R_{PN}(f_{PN}^*) \leq \frac{h(\delta)}{a} \sqrt{\frac{n_i + n_j}{n_i n_j}},$$

where $h(\delta) = 8\sqrt{2} C_\ell C_w C_x + 5\sqrt{2 \ln(2/\delta)}$, $a = \pi_i - \pi_j$, and n_i, n_j is the size of dataset U_i, U_j.

Theorem 3 guarantees that the excess risk of general case can be bounded plus the term of order

$$\mathcal{O}\left(\frac{1}{a\sqrt{n_i}} + \frac{1}{a\sqrt{n_j}} \right).$$

Let $\hat{f}_{U^m}^*$ be the minimizer of empirical risk $\hat{R}_{U^m}(f)$, we introduce the following excess risk bound, showing that the risk of $\hat{f}_{U^m}^*$ converges to risk of the optimal function in the function family $\mathcal{F}_K$.

Theorem 4 (Excess Risk for U^m AUC ERM problem)**.** *Assume that $\hat{f}_{U^m}^* \in \mathcal{F}_K$ is the minimizer of empirical risk $\hat{R}_{U^m}(f)$, $f_{PN}^* \in \mathcal{F}_K$ is the minimizer of true risk $R_{PN}(f)$. For any $\delta > 0$, with the probability at least $1 - \delta$, we have*

$$R_{PN}(\hat{f}_{U^m}^*) - R_{PN}(f_{PN}^*) \leq \frac{h(\frac{2\delta}{m(m-1)})}{s} \sum_{i,j|1 \leq i < j \leq m} z_{ij} \sqrt{\frac{n_i + n_j}{n_i n_j}},$$

where $h(\delta) = 8\sqrt{2} C_\ell C_w C_x + 5\sqrt{2 \ln(2/\delta)}$, and $s = \sum_{i,j|1 \leq i < j \leq m} z_{ij}(\pi_i - \pi_j)$, n_i, n_j is the size of unlabeled dataset U_i, U_j.

Theorem 4 guarantees that the excess risk of general case can be bounded plus the term of order

$$\mathcal{O}\left(\frac{1}{s} \sum_{i,j|1 \leq i < j \leq m} z_{ij} \sqrt{\frac{n_i + n_j}{n_i n_j}} \right).$$

It is evident that theorem 4 degenerates into theorem 3 when $m = 2$ and $z_{12} = 1$.

4 Experiments

In this section, we report the experimental results of the proposed U^m-AUC, compared to state-of-the-art U^m classification approaches.

Datasets We tested the performance of U^m-AUC using the benchmark datasets Kuzushiji-MNIST (K-MNIST for short) (Clanuwat et al. 2018), CIFAR-10, and CIFAR-100 (Krizhevsky, Hinton et al. 2009) with synthesizing multiple datasets with different settings. We transformed these datasets into binary classification datasets, where we classified odd vs. even class IDs for K-MNIST and animals vs. non-animals for CIFAR datasets. In the experiments, we choose $m \in \{10, 50\}$, and the size of each unlabeled data set U_i is fixed to $n_i = \lceil n_{\text{train}}/m \rceil$, unless otherwise specified. To simulate the distribution of the dataset in different cases, we will generate the class priors $\{\pi_i\}_{i=1}^m$ from four different distributions. We then randomly sampled data from the training set into U_i using the definition in eq. (6).

Models We use a 5-layer MLP (multi-layer perceptron) as our model for all experiments on the K-MNIST dataset, and use the Resnet32 (He et al. 2016) for experiments on the CIFAR datasets. We train all models for 150 epochs, and report the AUC on the test set at the final epoch.

Baselines In our experiments, we compared our method with state-of-the-art U^m classification methods: LLP-VAT (Tsai and Lin 2020) on behalf of EPRM methods, and U^m-SSC (Lu et al. 2021) on behalf of ERM methods. Note that the previous methods require exact class priors, while in our setting, we can only obtain relative order relations for the class priors of the unlabeled dataset. To ensure fairness in performance comparisons, we created weaker versions of LLP-VAT and U^m-SSC, called LLP-VAT* and U^m-SSC*, respectively, by giving them priors obtained by dividing $[0, 1]$ uniformly instead of using the true priors. We used Adam (Kingma and Ba 2014) and cross-entropy loss for their optimization, following the standard implementation in the original paper. To ensure fairness, we used the same model to implement all methods in all tasks.

Our implementation is based on PyTorch (Paszke et al. 2019), and experiments are conducted on an NVIDIA Tesla V100 GPU. To ensure the robustness of the results, all experiments are repeated 3 times with different random seed, and we report the mean values with standard deviations. For more details about the experiment, please refer to Appendix.

Dataset	$\mathcal{D}$	LLP-VAT*	LLP-VAT	U^m-SSC*	U^m-SSC	U^m-AUC
K-MNIST ($m = 10$)	$\mathcal{D}_u$	$0.865_{\pm 0.0145}$	$0.896_{\pm 0.0249}$	$0.908_{\pm 0.0073}$	$0.911_{\pm 0.0084}$	$\mathbf{0.938}_{\pm 0.0064}$
	$\mathcal{D}_b$	$0.780_{\pm 0.0225}$	$0.789_{\pm 0.0185}$	$0.833_{\pm 0.0357}$	$0.836_{\pm 0.0521}$	$\mathbf{0.851}_{\pm 0.0616}$
	$\mathcal{D}_c$	$0.853_{\pm 0.0330}$	$0.808_{\pm 0.0131}$	$0.858_{\pm 0.0239}$	$0.856_{\pm 0.0307}$	$\mathbf{0.870}_{\pm 0.0512}$
	$\mathcal{D}_{bc}$	$0.825_{\pm 0.0315}$	$0.798_{\pm 0.0332}$	$0.868_{\pm 0.0255}$	$0.857_{\pm 0.0390}$	$\mathbf{0.896}_{\pm 0.0439}$
CIFAR-10 ($m = 10$)	$\mathcal{D}_u$	$0.856_{\pm 0.0131}$	$0.856_{\pm 0.0066}$	$0.860_{\pm 0.0090}$	$0.859_{\pm 0.0131}$	$\mathbf{0.905}_{\pm 0.0080}$
	$\mathcal{D}_b$	$0.723_{\pm 0.0454}$	$0.737_{\pm 0.0754}$	$0.746_{\pm 0.0614}$	$0.778_{\pm 0.0462}$	$\mathbf{0.866}_{\pm 0.0238}$
	$\mathcal{D}_c$	$0.787_{\pm 0.0172}$	$0.847_{\pm 0.0059}$	$0.792_{\pm 0.0372}$	$0.807_{\pm 0.0209}$	$\mathbf{0.884}_{\pm 0.0046}$
	$\mathcal{D}_{bc}$	$0.769_{\pm 0.0373}$	$0.805_{\pm 0.0231}$	$0.796_{\pm 0.0552}$	$0.812_{\pm 0.0430}$	$\mathbf{0.887}_{\pm 0.0155}$
CIFAR-100 ($m = 10$)	$\mathcal{D}_u$	$0.734_{\pm 0.0092}$	$0.731_{\pm 0.0167}$	$0.747_{\pm 0.0192}$	$0.756_{\pm 0.0115}$	$\mathbf{0.847}_{\pm 0.0121}$
	$\mathcal{D}_b$	$0.630_{\pm 0.0183}$	$0.651_{\pm 0.0210}$	$0.652_{\pm 0.0332}$	$0.667_{\pm 0.0331}$	$\mathbf{0.715}_{\pm 0.0292}$
	$\mathcal{D}_c$	$0.670_{\pm 0.0168}$	$0.707_{\pm 0.0117}$	$0.676_{\pm 0.0363}$	$0.692_{\pm 0.0264}$	$\mathbf{0.757}_{\pm 0.0136}$
	$\mathcal{D}_{bc}$	$0.672_{\pm 0.0359}$	$0.700_{\pm 0.0324}$	$0.683_{\pm 0.0500}$	$0.701_{\pm 0.0415}$	$\mathbf{0.751}_{\pm 0.0641}$
K-MNIST ($m = 50$)	$\mathcal{D}_u$	$0.896_{\pm 0.0124}$	$0.902_{\pm 0.0102}$	$0.915_{\pm 0.0136}$	$0.915_{\pm 0.0107}$	$\mathbf{0.931}_{\pm 0.0156}$
	$\mathcal{D}_b$	$0.808_{\pm 0.0142}$	$0.787_{\pm 0.0196}$	$0.861_{\pm 0.0102}$	$0.869_{\pm 0.0083}$	$\mathbf{0.883}_{\pm 0.0229}$
	$\mathcal{D}_c$	$\mathbf{0.863}_{\pm 0.0206}$	$0.833_{\pm 0.0165}$	$0.855_{\pm 0.0378}$	$\mathbf{0.863}_{\pm 0.0417}$	$\mathbf{0.867}_{\pm 0.0125}$
	$\mathcal{D}_{bc}$	$0.860_{\pm 0.0523}$	$0.815_{\pm 0.0052}$	$0.881_{\pm 0.0056}$	$0.885_{\pm 0.0078}$	$\mathbf{0.904}_{\pm 0.0012}$
CIFAR-10 ($m = 50$)	$\mathcal{D}_u$	$0.852_{\pm 0.0079}$	$0.857_{\pm 0.0073}$	$0.853_{\pm 0.0030}$	$0.854_{\pm 0.0492}$	$\mathbf{0.889}_{\pm 0.0083}$
	$\mathcal{D}_b$	$0.757_{\pm 0.0250}$	$0.742_{\pm 0.0847}$	$0.794_{\pm 0.0278}$	$0.806_{\pm 0.0204}$	$\mathbf{0.861}_{\pm 0.0097}$
	$\mathcal{D}_c$	$0.790_{\pm 0.0132}$	$0.852_{\pm 0.0038}$	$0.807_{\pm 0.0101}$	$0.808_{\pm 0.0062}$	$\mathbf{0.861}_{\pm 0.0138}$
	$\mathcal{D}_{bc}$	$0.804_{\pm 0.0056}$	$0.830_{\pm 0.0235}$	$0.826_{\pm 0.0059}$	$0.832_{\pm 0.0052}$	$\mathbf{0.873}_{\pm 0.0074}$
CIFAR-100 ($m = 50$)	$\mathcal{D}_u$	$0.739_{\pm 0.0036}$	$0.738_{\pm 0.0084}$	$0.742_{\pm 0.0647}$	$0.744_{\pm 0.0084}$	$\mathbf{0.844}_{\pm 0.0042}$
	$\mathcal{D}_b$	$0.669_{\pm 0.0199}$	$0.673_{\pm 0.0363}$	$0.686_{\pm 0.0103}$	$0.696_{\pm 0.0068}$	$\mathbf{0.756}_{\pm 0.0281}$
	$\mathcal{D}_c$	$0.689_{\pm 0.0075}$	$0.724_{\pm 0.0065}$	$0.700_{\pm 0.0018}$	$0.703_{\pm 0.0097}$	$\mathbf{0.790}_{\pm 0.0085}$
	$\mathcal{D}_{bc}$	$0.699_{\pm 0.0065}$	$0.718_{\pm 0.0082}$	$0.714_{\pm 0.0009}$	$0.717_{\pm 0.0024}$	$\mathbf{0.812}_{\pm 0.0163}$

Table 1: Test AUC (mean±std) on benchmark datasets, under different prior distributions. Bolded font represents the best or comparable methods in the row, according to the pairwise t-test at the significance level 5%.

Comparison with Baseline Methods

To compare our approach with the baseline methods, we conducted experiments on the three image datasets and two different numbers of bags, as described above. In real-world scenarios, the class priors of datasets often do not follow a uniform distribution. To better simulate real-world situations, we considered four different class prior distributions for each image dataset and for each number of bags: $Beta(1, 1)$, $Beta(5, 1)$, $Beta(5, 5)$, and $Beta(5, 2)$. We refer to these four distributions as uniform, biased, concentrated, and biased concentrated, respectively. These four distributions represent four distinct cases as follows:

1. $\mathcal{D}_u$ (Uniform): the class priors are sampled from uniform distribution on $[0, 1]$;
2. $\mathcal{D}_b$ (Biased): the class priors are sampled from the distribution concentrated on one side, i.e., most sets have more positive samples than negative samples;
3. $\mathcal{D}_c$ (Concentrated): the class priors are sampled from the distribution concentrated around 0.5, i.e., most sets have close proportions of positive and negative samples;
4. $\mathcal{D}_{bc}$ (Biased Concentrated): the class priors are sampled from the distribution concentrated around 0.8, i.e., most sets have close proportions of positive and negative samples, and positive samples more than negative samples.

Our experiments encompass a broader range of settings compared to previous work. Specifically, the setting explored in (Lu et al. 2021) is just our $\mathcal{D}_u$ (Uniform) setting.

For $m = 10$ and $m = 50$, the results obtained from different datasets and varied distributions of class priors are reported in table 1. The results demonstrates that our proposed method, U^m-AUC, outperforms the baselines, even when a smaller amount of information is utilized.

Robustness to Imbalanced Datasets

One of the most prevalent challenges in classification tasks is handling class-imbalanced datasets. In the U^m setting, we also encounter imbalanced datasets. If the size of each dataset is imbalanced, it can result in biased models that prioritize the larger datasets.

To assess the robustness of our method against imbalanced datasets, we conducted tests using various settings. Specifically, we generated imbalanced datasets in two ways following the approach proposed in (Lu et al. 2021):

1. Size Reduction: With reduce ratio τ, randomly select $\lceil m/2 \rceil$ datasets, and change their size to $\lceil \tau \cdot (n_{\text{train}}/m) \rceil$.
2. Random: Randomly sample dataset size n_i from range $[0, n_{\text{train}}]$, such that $\sum_{i=1}^{m} n_i = n_{\text{train}}$.

The test AUC of U^m-AUC on with different imbalance datasets is presented in table 2. It indicates that our method

Dataset	m	$\tau = 0.8$	$\tau = 0.6$	$\tau = 0.4$	$\tau = 0.2$	Random
K-MNIST	10	$0.936_{\pm 0.0042}$	$0.934_{\pm 0.0095}$	$0.926_{\pm 0.0038}$	$0.928_{\pm 0.0046}$	$0.928_{\pm 0.0196}$
	50	$0.938_{\pm 0.0106}$	$0.932_{\pm 0.0047}$	$0.937_{\pm 0.0097}$	$0.928_{\pm 0.0042}$	$0.941_{\pm 0.0186}$
CIFAR-10	10	$0.907_{\pm 0.0087}$	$0.901_{\pm 0.0053}$	$0.901_{\pm 0.0039}$	$0.895_{\pm 0.0026}$	$0.904_{\pm 0.0123}$
	50	$0.900_{\pm 0.0022}$	$0.895_{\pm 0.0080}$	$0.893_{\pm 0.0023}$	$0.890_{\pm 0.0147}$	$0.902_{\pm 0.0048}$
CIFAR-100	10	$0.842_{\pm 0.0098}$	$0.835_{\pm 0.0036}$	$0.827_{\pm 0.0228}$	$0.817_{\pm 0.0243}$	$0.803_{\pm 0.0366}$
	50	$0.795_{\pm 0.0090}$	$0.805_{\pm 0.0067}$	$0.785_{\pm 0.0210}$	$0.777_{\pm 0.0213}$	$0.811_{\pm 0.0125}$

Table 2: Test AUC on benchmark datasets with different imbalanced setting.

is reasonably robust to the imbalance settings, as it exhibits a slow decline in test performance and a slow increase in test performance variance as the reduction ratio decreases.

5 Related Works

U^m Classification U^m classification involves learning a classifier from m unlabeled datasets, where we have limited information about each dataset, typically the class priors of each dataset. It appears in scenarios where individual labels are unavailable for privacy, technical, or ethical reasons, such as patient data collected from different areas (Croft et al. 2018) and voter data collected in different years (Newman 2003). The U^m classification setting is a case of *weak supervised learning* (Zhou 2017), and can be traced back to a classical problem of *learning with label proportions* (LLP) (Quadrianto et al. 2008). There are three main categories of previous approaches to solving the U^m classification problem. The first category is clustering-based approaches. Xu et al. (2004) and Krause, Perona, and Gomes (2010) assumed that each cluster corresponds to a single class and applied discriminative clustering methods to solve the problem. The second category is empirical proportion risk minimization-based approaches. Yu et al. (2014) aimed to minimize the distance between the average predicted probabilities and the class priors for each dataset U$_i$, which is known as empirical proportion risk. Meanwhile, Tsai and Lin (2020) introduced consistency regularization to the problem. The third category is empirical risk minimization-based approaches. Scott and Zhang (2020) extended the U^2 classification problem by ensembling classifiers trained on all pairs of unlabeled sets, while Lu et al. (2021) tackled the problem through a surrogate set classification task. However, all of the previous works on U^m classification have required knowledge of the class priors and are unable to address situations where the priors are inaccurate or only the relative order of the unlabeled sets' class priors is known.

AUC Optimization AUC is a commonly used performance measure that holds significant value in a variety of scenarios, and the use of AUC is particularly advantageous when dealing with imbalanced data or when the model's ranking capability is of utmost importance. While the goal is to train models with better AUC, studies (Cortes and Mohri 2003) have shown that algorithms that maximize model accuracy do not necessarily maximize the AUC score. Accordingly, numerous studies have been dedicated to directly optimizing the AUC for decades (Yang and Ying 2022). To enable efficient optimization of AUC, Gao et al. (2013) proposed an AUC optimization algorithm using a covariance matrix, while Ying, Wen, and Lyu (2016) optimized the AUC optimize problem as a stochastic saddle point problem with stochastic gradient-based methods. For AUC optimization with deep neural models, Liu et al. (2020) introduced deep AUC maximization based on a non-convex min-max problem, and some AUC optimization frameworks are proposed (Yang 2022; Yuan et al. 2023). Recently, there are also studies of partial-AUC and multi-class AUC optimization (Yang et al. 2021a,b; Zhu et al. 2022; Yao, Lin, and Yang 2022). In addition to the algorithms, significant work has been conducted on the theoretical aspects. For example, Gao and Zhou (2015) investigated the consistency of commonly used surrogate losses for AUC, while Agarwal et al. (2005) and Usunier, Amini, and Gallinari (2005) studied the generalization bounds of AUC optimization models. The research on AUC optimization has led to the development of numerous real-world applications, such as software build outcome prediction (Xie and Li 2018a), medical image classification (Yuan et al. 2021), and molecular property prediction (Wang et al. 2022). Most recently, there has been a growing body of research on weakly supervised AUC optimization. For example, Sakai, Niu, and Sugiyama (2018) and Xie and Li (2018b) studied semi-supervised AUC optimization, Charoenphakdee, Lee, and Sugiyama (2019) studied the properties of AUC optimization under label noise, and Xie et al. (2023) offered a universal solution for AUC optimization in various weakly supervised scenarios. However, to the best of our knowledge, there has been no investigation into AUC optimization for U^m classification to date.

6 Conclusion

In this work, we investigate the challenge of constructing AUC optimization models from multiple unlabeled datasets. To address this problem, we propose U^m-AUC, a novel AUC optimization method with both simplicity and efficiency. U^m-AUC is the first solution for AUC optimization under the U^m learning scenario and provides a solution for U^m learning without exact knowledge of the class priors. Furthermore, theoretical analysis demonstrates the validity of U^m-AUC, while empirical evaluation demonstrates that U^m-AUC exhibits superiority and robustness compared to the state-of-the-art alternatives.

Acknowledgements

This research was supported by NSFC (62076121, 619210 06), and Key Program of Hubei (JD2023008).

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: 393–425.

Bekker, J.; and Davis, J. 2020. Learning from Positive and Unlabeled Data: A Survey. *Machine Learning*, 109(4): 719–760.

Charoenphakdee, N.; Lee, J.; and Sugiyama, M. 2019. On Symmetric Losses for Learning from Corrupted Labels. In *Proceedings of 36th International Conference on Machine Learning*, 961–970.

Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; and Ha, D. 2018. Deep Learning for Classical Japanese Literature. arXiv:1812.01718.

Cortes, C.; and Mohri, M. 2003. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems 16*.

Croft, J. B.; Wheaton, A. G.; Liu, Y.; Xu, F.; Lu, H.; Matthews, K. A.; Cunningham, T. J.; Wang, Y.; and Holt, J. B. 2018. Urban-rural county and state differences in chronic obstructive pulmonary disease—United States, 2015. *Morbidity and Mortality Weekly Report*, 67(7): 205.

Gao, W.; Jin, R.; Zhu, S.; and Zhou, Z.-H. 2013. One-Pass AUC Optimization. In *Proceedings of 30th International Conference on Machine Learning*, 906–914.

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

Guo, Z.; Yan, Y.; Yuan, Z.; and Yang, T. 2020. Fast objective & duality gap convergence for nonconvex-strongly-concave min-max problems. arXiv:2006.06889.

Han, B.; Yao, Q.; Liu, T.; Niu, G.; Tsang, I. W.; Kwok, J. T.; and Sugiyama, M. 2021. A Survey of Label-noise Representation Learning: Past, Present and Future. arXiv:2011.04406.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Krause, A.; Perona, P.; and Gomes, R. 2010. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems 23*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

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

Lu, N.; Lei, S.; Niu, G.; Sato, I.; and Sugiyama, M. 2021. Binary Classification from Multiple Unlabeled Datasets via Surrogate Set Classification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 7134–7144.

Newman, B. 2003. Integrity and presidential approval, 1980–2000. *Public Opinion Quarterly*, 67(3): 335–367.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035.

Quadrianto, N.; Smola, A. J.; Caetano, T. S.; and Le, Q. V. 2008. Estimating labels from label proportions. In *Proceedings of the 25th International Conference on Machine Learning*, 776–783.

Sakai, T.; Niu, G.; and Sugiyama, M. 2018. Semi-Supervised AUC Optimization Based on Positive-Unlabeled Learning. *Machine Learning*, 107: 767–794.

Scott, C.; and Zhang, J. 2020. Learning from Label Proportions: A Mutual Contamination Framework. In *Advances in Neural Information Processing Systems 35*, 22256–22267.

Tsai, K.-H.; and Lin, H.-T. 2020. Learning from Label Proportions with Consistency Regularization. In *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129, 513–528.

Usunier, N.; Amini, M.-R.; and Gallinari, P. 2005. A Data-dependent Generalisation Error Bound for the AUC. In *ICML'05 Workshop ROC Analysis in Machine Learning*.

Wang, Z.; Liu, M.; Luo, Y.; Xu, Z.; Xie, Y.; Wang, L.; Cai, L.; Qi, Q.; Yuan, Z.; Yang, T.; et al. 2022. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. *Bioinformatics*, 38(9): 2579–2586.

Xie, Z.; and Li, M. 2018a. Cutting the Software Building Efforts in Continuous Integration by Semi-Supervised Online AUC Optimization. In *Proceedings of 27th International Joint Conference on Artificial Intelligence*, 2875–2881.

Xie, Z.; and Li, M. 2018b. Semi-Supervised AUC Optimization Without Guessing Labels of Unlabeled Data. In *Proceedings of 32nd AAAI Conference on Artificial Intelligence*, 4310–4317.

Xie, Z.; Liu, Y.; He, H.-Y.; Li, M.; and Zhou, Z.-H. 2023. Weakly Supervised AUC Optimization: A Unified Partial AUC Approach. arXiv:2305.14258.

Xu, L.; Neufeld, J.; Larson, B.; and Schuurmans, D. 2004. Maximum margin clustering. In *Advances in Neural Information Processing Systems 17*.

Yang, T. 2022. Algorithmic Foundations of Empirical X-Risk Minimization. arXiv:2206.00439.

Yang, T.; and Ying, Y. 2022. AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys*, 55(8): 1–37.

Yang, Z.; Xu, Q.; Bao, S.; Cao, X.; and Huang, Q. 2021a. Learning with multiclass AUC: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7747–7763.

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.

Yao, Y.; Lin, Q.; and Yang, T. 2022. Large-scale optimization of partial auc in a range of false positive rates. *Advances in Neural Information Processing Systems*, 35: 31239–31253.

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

Yu, F. X.; Choromanski, K.; Kumar, S.; Jebara, T.; and Chang, S.-F. 2014. On learning from label proportions. arXiv:1402.5902.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3040–3049.

Yuan, Z.; Zhu, D.; Qiu, Z.-H.; Li, G.; Wang, X.; and Yang, T. 2023. LibAUC: A Deep Learning Library for X-Risk Optimization. In *29th SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53.

Zhu, D.; Li, G.; Wang, B.; Wu, X.; and Yang, T. 2022. When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In *International Conference on Machine Learning*, 27548–27573. PMLR.

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