类别条件噪声下的半监督 AUC 优化理论与算法

姜阳邦彦[1] 许倩倩[2] 杨智勇[1] 郝前秀[2] 操晓春[3] 黄庆明[1),2)]

[1)](中国科学院大学计算机科学与技术学院 北京 101408)

[2)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[3)](中山大学网络空间安全学院 广东 深圳 518107)

摘 要 现有半监督 AUC 优化方法通常假设数据标注是准确的。然而在许多实际应用中,研究者往往会同时面临标注量不足和不准确的问题。为此,该文首次尝试在不完整和不准确的数据标注情况下优化 AUC 指标。具体而言,通过分析,对称替代损失在某些情况下可以在半监督问题中具有噪声鲁棒性。在此基础上,该文构建了一个鲁棒半监督 AUC 优化框架,其导出的经验风险无需估计噪声率。此外,通过紧致泛化上界的分析表明,当模型基于足够大的训练数据集进行学习时,其在未见数据上能够很好地泛化。随后,使用 Barrier hinge 损失对该框架进行实例化。为加快训练过程,进一步开发了一种加速算法,将损失和梯度估计的复杂度从 $O(n^2)$ 降低至 $O(n\log n)$,在实验中可获得高达 200 倍的加速。最后,通过在 15 个基准数据集上进行实验验证,证明了所提方法的有效性。

关键词 半监督学习;AUC 优化;标签噪声;二分类问题;不平衡数据学习

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2025.00136

Semi-Supervised AUC Optimization with Class-Conditional Noises: Theory and Algorithms

JIANG Yang-Bang-Yan[1] XU Qian-Qian[2] YANG Zhi-Yong[1]

HAO Qian-Xiu[2] CAO Xiao-Chun[3] HUANG Qing-Ming[1),2)]

[1)](School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408)

[2)](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

[3)](School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen, Guangdong 518107)

Abstract The Area Under ROC Curve (AUC) is one of the most popular evaluation metrics for classification performance, which measures the probability of a positive sample to be scored higher than a negative one. Compared to the accuracy-based metrics, AUC is insensitive to label distributions and misclassification costs, making it a better choice under a long-tail scenario. Accordingly, there is a large amount of work trying to directly optimize AUC, especially under incomplete supervision due to the costly labeling process. Despite the great progress on learning from incomplete data, existing semi-supervised AUC optimization methods usually assume that the labeled data is clean and trustworthy. However, in many practical applications, we must simultaneously face insufficiency of data and inaccuracy of annotations. Motivated by this fact, we present the

收稿日期:2024-02-26;在线发布日期:2024-09-14。本课题得到新一代人工智能国家科技重大专项(2018AAA0102000)、国家自然科学基金项目(62236008,U21B2038,U23B2051,61931008,62122075,62406305,62476068,62471013,62206264,92370102)、中国科学院青年促进会优秀会员项目、中国科学院战略性先导科技专项(XDB0680000)、中国科学院计算技术研究所创新课题(E000000)、中国博士后科学基金项目(2023M743441)、国家资助博士后研究人员计划(GZB20230732)资助。**姜阳邦彦**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为机器学习与计算机视觉。E-mail:jiangyangbangyan@ucas.ac.cn。**许倩倩**(通信作者),博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为统计机器学习及其在多媒体领域的应用。E-mail:xuqianqian@ict.ac.cn。**杨智勇**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为机器学习理论与方法。**郝前秀**,硕士,工程师,主要研究方向为机器学习与多媒体分析。**操晓春**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为计算机视觉、多媒体分析。**黄庆明**(通信作者),博士,讲席教授,中国计算机学会(CCF)会士,主要研究领域为多媒体计算、计算机视觉、模式识别。E-mail:qmhuang@ucas.ac.cn。

first trial on optimizing AUC under the context of incomplete and inaccurate data annotations. The challenges of this task lie in that there are (1) neither a subset of clean labeled instances to propagate confidential labels for the unlabeled data, (2) nor sufficient labeled samples to learn the patterns of noise-free distribution. To address these challenges, we present a reformulation of the AUC score when the data suffers from class-conditional noise. With such a reformulation, the challenging label propagation process in (1) could be dropped. More importantly, one can derive a much more simplified expression of true AUC where the impact of the noise distribution is removed, eliminating the problem in (2). Specifically, we show that the symmetric surrogate losses are noise-robust in the semi-supervised setting under certain scenarios. With the help of symmetric loss functions, AUC surrogate risks developed by noisy positive/negative, noisy positive/unlabeled, and noisy negative/unlabeled data all could recover the ideal AUC surrogate risks. On top of this result, we construct a robust semi-supervised AUC optimization framework along with the induced empirical risks does not need to estimate the noise rates. Taking a step further, a tight algorithm-dependent uniform generalization upper bound of the excess risk is presented to show that a model learned from a sufficiently large given training dataset could generalize well to further unseen data. Here we develop an algorithm-dependent hypothesis class out of the entire hypothesis class, which concentrates on the expected hypothesis outputted by a given randomized learning algorithm. Compared to the earlier-known results, we show that the upper bound of excess risk could converge at a much faster rate. Practically, we propose an instantiation of our framework with the Barrier hinge loss. To speed up the training process, an acceleration algorithm is developed for this loss to reduce the complexity of loss and gradient evaluation from $O(n^2)$ to $O(n\log n)$, which leads to up to 200x speedup in the experiments. The key idea is that by properly reformulation, the optimization objective functions can be divided into several parts where each part can be computed recursively with dynamic programming. Finally, we conduct extensive experiments on 15 benchmark datasets including ablation study, sensitivity analysis and complexity analysis. The empirical results speak to the effectiveness of our proposed method.

Keywords　semi-supervised learning; AUC optimization; label noise; binary classification; imbalanced data learning

1　引　言

ROC曲线下夹面积(Area Under the ROC Curve, AUC)是分类问题中最常见的性能评估指标之一，它衡量了正样本得分高于负样本的概率。相比基于准确率的指标，AUC 对标签分布和误分类代价不敏感，在长尾场景下更具优势。该特性对于存在类别不平衡情况的数据(即某一类样本量占据主导地位)或任务(即不同类别之间的误分类代价有所差异)尤其有效。

由于 AUC 在处理不平衡数据方面的固有优势，学界提出了大量工作以直接优化 AUC 指标，并在理论分析[1-3]和优化效率[4-7]等方面取得了显著的成功。近年来，针对仅有少量数据标注而大部分数据未标注的不完整监督下的 AUC 优化工作，出现了蓬勃发展的趋势[8-11]。该类半监督 AUC 优化的需求源于一些场景下标注获取极高的难度和成本。例如，在疾病检测中，即使对于医生来说，标注过程的耗时也较长；在生物实验中，获取样本标注甚至可能需要花费数天的时间。

尽管现有的半监督 AUC 优化方法在不完整标注数据学习上取得了巨大进展，但该类方法均假设有标注数据干净、可信，而在实践中往往并非如此。例如，标注者的不专业或者不专心将导致人工标注数据不准确。并且在生物实验中，由于设备限制或其他复杂原因，获得的标签也并非总是正确的。

鉴于此，本文旨在在不完整监督和噪声标签场

景下进行 AUC 优化。其主要挑战在于:(1)缺少干净的标注子集以向无标注数据传播可靠的标签信息;(2)缺乏足量有标注样本以学习无噪声分布下的模式。为了应对这些挑战,本文提出了一种在数据中存在类别条件噪声时的 AUC 得分重构方法,以避免显式地进行标签传播和无噪声分布的估计。通过该重构,可以避免(1)中的标签传播过程。更重要的是,可以导出真实 AUC 更简化的表达形式,消除了噪声分布的影响。该发现说明,可以使用基于经验分布的有偏 AUC 得分来恢复无噪声分布下的 AUC,从而解决(2)中的问题。基于此,本文提出了一个基于对称损失和辅助 AUC 得分的半监督 AUC 优化目标函数。

为了探究该目标函数的泛化能力,本文进一步给出了其额外风险的上界,其中引入了一个算法依赖的局部假设类(Local hypothesis class),它集中于由给定随机学习算法输出的期望假设,使得上界更为紧致。理论结果表明,当无标注样本的数量足够大于有标注样本时,额外风险(Excess risk)的上界可以达到 $O(1/n^{+}+1/n^{-}+1/n^{\sharp})$,与早期已知的 $O(1/\sqrt{n^{+}}+1/\sqrt{n^{-}}+1/\sqrt{n^{\sharp}})$ 相比要紧得多。其中,n^{+}、n^{-} 和 $n^{\sharp}$ 分别表示正例、负例和无标注样本的数量。通过这一更紧致的上界,经验 AUC 风险可以以更快速率收敛到期望 AUC 风险。

此外,由于经验 AUC 风险需要进行逐对求和,损失和梯度估计的复杂度是数据集规模的平方级别,难以应用于大规模数据集。为了解决该问题,本文进一步为所采用的对称损失开发了一个加速算法。其关键思想在于,通过适当的重构,将目标函数分解为几个部分,各部分均可通过动态规划进行递推计算。所提出的加速算法成功将算法复杂度从 $O(n^2)$ 降至 $O(n\log n)$,其中 n 为数据集规模。

总体而言,本文的主要贡献可总结为如下三个方面:

(1)面向不完整和不准确的数据,提出了一种早期的 AUC 优化方法尝试。在理论上证明了对称损失函数对于类别条件标签噪声的鲁棒性。在此基础上,提出了一个基于对称损失的鲁棒半监督 AUC 优化框架,无需估计噪声比例。

(2)在传统 AUC 学习理论基础之上,推导出一种算法相关的额外风险统一上界。在一定条件下,成功将上界的阶数从 $O(1/\sqrt{n^{+}}+1/\sqrt{n^{-}}+1/\sqrt{n^{\sharp}})$ 降至 $O(1/n^{+}+1/n^{-}+1/n^{\sharp})$。

(3)进一步开发了一种基于 Barrier hinge 损失的 AUC 优化加速算法,将计算复杂度从 $O(n^2)$ 降至 $O(n\log n)$。

此外,在 15 个真实数据集上的实验结果展现了所提方法的优越性,模拟数据上的实验证明了所提的加速算法能够获得显著的加速比。

2 相关工作

本节将简要回顾与标签噪声学习和半监督 AUC 优化相关的研究工作。

2.1 标签噪声学习

鉴于现实场景中标签噪声的普遍存在,学界已有大量研究关注如何从噪声标签中进行学习[12-13],主要可归纳为以下四类:(1)基于样本选择的方法[14-22],旨在从噪声数据中选择具有可信标注的样本进行训练;(2)基于损失调整的方法,通过估计标签转移矩阵来校正损失[23-25],或对每个样本进行重要性加权[26-27]以减轻噪声样本的影响;(3)基于模型结构的方法[28-32],试图通过精心设计的结构来建模标签转移概率;(4)基于鲁棒替代损失函数的方法。由于鲁棒损失函数无需利用干净样本训练分类器,也无需利用大量数据标注来建模复杂噪声,因此本文主要关注鲁棒替代损失函数。接下来对这方面的研究进行简要介绍。最初,有研究发现在 0-1 损失下的风险最小化对均匀噪声(Uniform noise)具有鲁棒性[33]。通过反例,该研究还证明了标准的凸损失函数如 Hinge、对数和指数损失,对均匀噪声并不具有鲁棒性[34]。随后,文献[35]提出了 Unhinged 损失,该损失是分类校准的、凸的,并且对对称标签噪声具有鲁棒性。近期,文献[36]提出了一种 Barrier hinge 损失,该损失只在特定区域内满足对称条件。本文在一定程度上受到了对称损失鲁棒性的启发。不同之处在于,上述方法仅在有监督场景下研究了对称损失的鲁棒性,而本研究在理论上证明了对称损失在半监督场景下也具有鲁棒性。

2.2 半监督 AUC 优化

半监督 AUC 优化旨在实现对类别不平衡和标注不完整二分类数据的学习。尽管该场景较为常见,但鲜少有对应的研究。最初,文献[8]假设相似的有、无标注样本应该具有相似的标签,并以此为各无标注样本分配伪标签。然后,在 RankBoost 框架[37]中分别对有、无标注样本进行排序误差最小化。随后,文献[38]在基于间隔最大化理论的 AUC 优化框架

中，将无标注数据纳入了优化约束，以在优化过程中猜测无标注数据的标签。此后，文献[9]利用生成模型同时学习有、无标注样本的分布，并基于无标注样本的分布进行了半监督 AUC 优化。

然而，上述方法都依赖于特定的分布假设，而当分布假设与实际情况不符时将导致标签估计不准确，使得分类器性能较差。为了解决该问题，文献[10]提出了一种无偏的半监督 AUC 优化方法，不依赖于特定的分布假设。然而，文献[11]指出，该方法仍然需要准确估计类别先验概率以对无标注数据重新加权，而当有标注数据极少时，该过程较为困难。因此，该研究在理论上证明了可以将无标注数据同时视为正样本和负样本，无需任何分布假设或类别先验估计即可无偏估计 AUC 风险。随后的工作[39]将其进一步扩展为半监督在线 AUC 优化方法。

尽管取得了一定成功，上述半监督 AUC 优化方法仍假设数据标注可信，而在实践中往往并非如此。鉴于此，本文对不完整、不准确且可能不平衡数据集上的 AUC 优化进行了初步尝试。本文的研究遵循文献[10-11]的框架，与选取的模型无关。此外，本文在理论上和实验上证明，尽管文献[11]具有无需估计类别先验的优势，但由于同时使用了正例无标注学习(Positive unlabeled learning)和负例无标注学习(Negative unlabeled learning)，仍可能性能较差。

2.3　学习理论

(1) AUC 优化。传统的学习理论基于训练样本相互独立且同分布的假设。然而，在包括 AUC 学习在内的二分排序问题中，训练样本之间存在相互依赖关系。为了解决该问题，学界进行了许多努力，打破了独立同分布的假设。首次对排序问题泛化性质的研究可以追溯到 RankBoost[37]，它从一致收敛的角度推导出了 VC 维的泛化误差上界。之后，文献[40]基于扩展的 Rademacher 复杂度推导出了一种一致收敛上界。该文献引入了一组新的组合参数，将相互依赖的样本集划分为一系列独立子集。在这些独立子集中，样本满足独立同分布的假设，因此可以将传统的技术(例如 McDiarmid 定理[41])扩展到独立子集上。然而，该方法只关注整个假设类中最坏情况的假设，而不考虑学习算法本身。由于模型通常是经过良好训练的，分析这样一个最坏情况下的假设意义较小。在该工作的基础上，本研究为半监督 AUC 优化提供了一个更紧致的局部额外风险上界，其关键思想是基于算法相关的假设子类构建和局部 Rademacher 复杂度的技术[42]。

(2) 算法稳定性。算法稳定性(Algorithmic stability)的概念至少可以追溯到文献[43-45]，这些工作指出，如果对算法进行微小扰动，其输出只会产生微小影响，则该算法是稳定的。接着，一些开创性的工作[46-47]给出了构建基于一致稳定性(Uniform stability)概念的算法相关泛化误差界的通用框架。在此基础上，学界取得了面向不同算法的算法相关泛化性质的研究成果[48-54]。其中，文献[54]与本研究密切相关，该研究发展了一种算法相关的假设子类，并将传统的 Rademacher 复杂度扩展到算法相关的假设子类中。然而，该工作立足于损失函数是独立同分布项的有限和这一假设，无法直接适用于 AUC 优化问题。

3　预备知识

在传统有监督二分类问题中，假设有干净的正样本集 $\mathcal{X}^+$ 和干净的负样本集 $\mathcal{X}^-$，分别独立同分布抽取自条件密度分布 $\mathcal{D}_P=\Pr(x\mid y=1)$ 和 $\mathcal{D}_N=\Pr(x\mid y=-1)$。记 $g:\mathbb{R}^d\mapsto\mathbb{R}$ 为样本得分函数，并根据得分的符号进行分类：$\hat{y}=\operatorname{sign}(g(x))$。AUC 优化分类器旨在通过最大化 AUC 指标来训练分类器 g。由于 AUC 等价于任意正样本比负样本排序靠前的概率[55]，有

$$AUC(g)=1-\underbrace{\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\ell_{0\text{-}1}(g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-))}_{\text{AUC风险}} \tag{1}$$

其中，$\mathbb{I}[A]$ 为指示函数，当 A 成立时等于 1，否则等于 0；$\ell_{0\text{-}1}(x)=\frac{1}{2}-\frac{1}{2}\operatorname{sign}(x)$ 是 0-损失函数。不难看出，最大化 AUC 相当于最小化式(1)中的第二项，即 AUC 风险。由于 0-1 损失不可微，较难优化，因而通常将其替换为一个可微的替代损失 $\ell(\cdot)$ 作为优化目标。为简单起见，记 $f(\boldsymbol{x},\boldsymbol{x}')=g(\boldsymbol{x})-g(\boldsymbol{x}')$。则可得到 AUC 替代风险 $R(f)=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-))$。实际上，由于真实正负类分布情况未知，所以通过最小化经验风险 $\hat{R}(f)$ 间接最小化期望风险：

$$\hat{R}(f)=\frac{1}{n^+n^-}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sum_{\boldsymbol{x}^-\in\mathcal{X}^-}\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-)) \tag{2}$$

其中，$n^+=|\mathcal{X}^+|$ 和 $n^-=|\mathcal{X}^-|$ 分别为正、负样本数。

4　模型设计

4.1　问题定义

本文考虑了一种更具挑战性但也至关重要的场

景，其中大量数据未标注，且已标注数据中存在不准确的标注。假设含噪声正样本集$\widetilde{\mathcal{X}}^+$、含噪声负样本集$\widetilde{\mathcal{X}}^-$和无标注样本集$\mathcal{X}^\sharp$抽取自以下分布：

$$\widetilde{\mathcal{X}}^+=\{\boldsymbol{x}_i^+\}_{i=1}^{n^+}\overset{i.i.d.}{\sim}\pi\cdot\Pr(\boldsymbol{x}|y=1)+(1-\pi)\cdot\Pr(\boldsymbol{x}|y=-1),$$

$$\widetilde{\mathcal{X}}^-=\{\boldsymbol{x}_j^-\}_{j=1}^{n^-}\overset{i.i.d.}{\sim}\pi'\cdot\Pr(\boldsymbol{x}|y=1)+(1-\pi')\cdot\Pr(\boldsymbol{x}|y=-1),$$

$$\mathcal{X}^\sharp=\{\boldsymbol{x}_k^\sharp\}_{k=1}^{n^\sharp}\overset{i.i.d.}{\sim}\theta^+\cdot\Pr(\boldsymbol{x}|y=1)+\theta^-\cdot\Pr(\boldsymbol{x}|y=-1),$$

其中，n^+、n^-和$n^\sharp$分别表示正、负和无标注样本的数量。假设数据中存在类别条件噪声[56-57]，即$1-\pi=\Pr(y=-1|\tilde{y}=1)$，$\pi'=\Pr(y=1|\tilde{y}=-1)$，$\theta^+=\Pr(y=1)$，$\theta^-=\Pr(y=-1)$。显然，$\theta^++\theta^-=1$。通常情况下，噪声率不会很高，因此可假设$\pi'<\theta^+<\pi$。其中，$\pi'<\pi$保证了观测到的标签大多数都是正确的；当$\pi'=\pi$时，各类别均有一半的标签噪声，含噪声正负样本的数据分布相同，学习难以进行；而当$\pi'>\pi$时，可以通过交换含噪声正负样本的数据分布转化为$\pi'<\pi$[56-57]。而为了保证数据可学习，进一步有$\pi'<\theta^+<\pi$。事实上，现有真实噪声数据集的噪声率通常也不高于40%[58]。

4.2 风险形式

为书写便利，将$\widetilde{\mathcal{X}}^+$、$\widetilde{\mathcal{X}}^-$和$\mathcal{X}^\sharp$的分布分别表示为$\mathcal{D}_{\tilde{\mathrm{P}}}$、$\mathcal{D}_{\tilde{\mathrm{N}}}$和$\mathcal{D}_U$。为利用无标注数据，首先引入以下三种存在标签噪声的有偏AUC替代风险：

$$\begin{aligned}\widetilde{R}_{+-}(f)&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_{\tilde{\mathrm{P}}}}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_{\tilde{\mathrm{N}}}}\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-)),\\ \widetilde{R}_{+\sharp}(f)&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_{\tilde{\mathrm{P}}}}\mathbb{E}_{\boldsymbol{x}^\sharp\sim\mathcal{D}_U}\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^\sharp)),\\ \widetilde{R}_{\sharp-}(f)&=\mathbb{E}_{\boldsymbol{x}^\sharp\sim\mathcal{D}_U}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_{\tilde{\mathrm{N}}}}\ell(f(\boldsymbol{x}^\sharp,\boldsymbol{x}^-))\end{aligned}\tag{3}$$

命题1. 令ℓ为对称损失，使得$\ell(z)+\ell(-z)=\mathrm{const}$。则有

$$\begin{aligned}\widetilde{R}_{+\sharp}(f)&=(\pi-\theta^+)R(f)+\frac{\theta^++1-\pi}{2}\mathrm{const},\\ \widetilde{R}_{\sharp-}(f)&=(\theta^+-\pi')R(f)+\frac{\pi'+\theta^-}{2}\mathrm{const},\\ \widetilde{R}_{+-}(f)&=(\pi-\pi')R(f)+\frac{1-\pi+\pi'}{2}\mathrm{const}\end{aligned}\tag{4}$$

证明详见附录A。

该命题展示了对称损失可支持半监督AUC优化框架处理标注不完整且不准确的数据。具体而言，通过损失函数的对称性质，可得出含噪声正/负样本和无标注样本上的有偏AUC替代风险$\widetilde{R}_{+\sharp}/\widetilde{R}_{\sharp-}$与干净正、负样本上的AUC替代风险$R$的线性关系。由此可知，二者的优化过程等价。即使无法直接计算干净样本上的风险，也可以通过最小化含噪声样本上的有偏风险来间接优化干净样本上的风险。同样地，对于含噪声正、负样本上的有偏AUC替代风险$\widetilde{R}_{+-}$，其与R的线性关系也成立。这种等价关系也消除了对噪声比例π和π'以及类别先验θ^+和θ^-的依赖，使得模型在无需估计上述比例的情况下也可进行学习。

4.3 优化目标

根据上述结果，可引出基于对称损失的鲁棒半监督AUC优化框架。直观地说，由于这三种风险都与理想的AUC替代风险成线性关系，它们都可以直接用于优化分类器/得分函数。一种自然的方式是构建以下组合风险：

$$\widetilde{R}_{P\text{-and-}N}(f)\triangleq\alpha_1\widetilde{R}_{\pm}(f)+\alpha_2\widetilde{R}_{+\sharp}(f)+\alpha_3\widetilde{R}_{\sharp-}(f)\tag{5}$$

其中，α_1、α_2、$\alpha_3>0$是满足$\alpha_1+\alpha_2+\alpha_3=1$的均衡参数。受文献[59]的启发，还可提出以下组合风险形式：

$$\widetilde{R}_{P\text{-or-}N}(f)\triangleq\begin{cases}\widetilde{R}_{PU}^{\alpha}(f), & \alpha\geq0\\ \widetilde{R}_{UN}^{-\alpha}(f), & \alpha<0\end{cases}\tag{6}$$

其中，$\alpha\in[-1,1]$为均衡参数，且有

$$\begin{aligned}\widetilde{R}_{PU}^{\alpha}(f)&\triangleq(1-\alpha)\widetilde{R}_{\pm}(f)+\alpha\widetilde{R}_{+\sharp}(f),\\ \widetilde{R}_{UN}^{\alpha}(f)&\triangleq(1-\alpha)\widetilde{R}_{\pm}(f)+\alpha\widetilde{R}_{\sharp-}(f)\end{aligned}\tag{7}$$

显然，$\widetilde{R}_{PU}^{\alpha}$由$\widetilde{R}_{+-}$和$\widetilde{R}_{+\sharp}$组合得到，而$\widetilde{R}_{UN}^{\alpha}$由$\widetilde{R}_{+-}$和$\widetilde{R}_{\sharp-}$组合得到。由第5.3节中的讨论可知，$\widetilde{R}_{P\text{-or-}N}$得到的额外风险的阶数比$\widetilde{R}_{P\text{-and-}N}$更低。此外，第7节中的实验结果也表明，$\widetilde{R}_{P\text{-or-}N}$通常表现更好。因此，本文采用$\widetilde{R}_{P\text{-or-}N}$而非$\widetilde{R}_{P\text{-and-}N}$。

由于实际数据分布未知，转而优化其对应的经验风险：

$$\begin{aligned}\mathcal{L}_{P\text{-or-}N}=&\frac{1-|\alpha|}{n^+n^-}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sum_{\boldsymbol{x}^-\in\mathcal{X}^-}\ell(g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-))+\\ &\frac{\mathbb{I}[\alpha>0]|\alpha|}{n^+n^\sharp}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sum_{\boldsymbol{x}^\sharp\in\mathcal{X}^\sharp}\ell(g(\boldsymbol{x}^+)-g(\boldsymbol{x}^\sharp))+\\ &\frac{\mathbb{I}[\alpha<0]|\alpha|}{n^\sharp n^-}\sum_{\boldsymbol{x}^\sharp\in\mathcal{X}^\sharp}\sum_{\boldsymbol{x}^-\in\mathcal{X}^-}\ell(g(\boldsymbol{x}^\sharp)-g(\boldsymbol{x}^-))+\\ &\gamma\|\Theta\|_2^2\end{aligned}\tag{8}$$

其中，Θ是模型参数，γ是正则化项的均衡参数。

4.4 对称替代损失

上述风险实例化的关键步骤在于寻找一个对称的替代损失。常用的对称替代损失包括Sigmoid损失和Ramp损失。然而，这些损失均为非凸函数，可能导致优化问题落入局部最小值。而另一方面，又没有任何下界为凸函数的损失函数（如SVM中使用的Hinge损失或AdaBoost中使用的指数损失）满足对称条件。鉴于此，本文采用了近期提出的Barrier hinge损失[36]作为一种折中选择：

$$\ell_{\text{barrier}}(z)=\begin{cases}-b(z+r)+r, & z\leq\dfrac{b\cdot r}{1-b}\\ r-z, & \dfrac{b\cdot r}{1-b}<z<r\\ b(z-r), & z\geq r\end{cases}\quad(9)$$

其中，$b>1$，$r>0$。显然该损失为凸函数，且在$[-r,r]$是对称的。此外，当 $b\gg1$ 时，随着输入值偏离对称范围，函数值会急剧增加。因而，该函数可通过较大损失值产生的较大梯度迫使输入值回到对称范围，以满足对噪声鲁棒性的要求。后文的实验将展示随着训练的进行，得分差异终将落在对称范围内。

表 1 中展示了其他常见替代损失函数及其凸和对称性质。当损失函数具有对称性质时，才能保证命题 1 成立，即优化含噪声半监督 AUC 风险等价于优化干净数据上的 AUC 风险，使得所提目标函数具有噪声鲁棒性。另一方面，凸函数相较于非凸函数具有更好的优化性质。在这些替代损失函数中，Barrier hinge 损失同时满足了凸性和局部对称性，且相较于同样凸且局部对称的 Hinge 损失，在非对称范围内的函数值急剧增加，有利于模型通过学习将损失函数的输出控制在对称范围，保证模型鲁棒性，因而具有明显优势。

表 1　替代损失及其相关性质

替代损失	$l(z)$	凸性质	对称性质
Sigmoid	$1/(1+\exp(z))$	非凸	全局对称
Ramp	$\max(0,\min(1,0.5-0.5z))$	非凸	全局对称
Squared	$(1-z)^2$	凸	不对称
Hinge	$\max(0,1-z)$	凸	局部对称(对称范围$[-1,1]$)
Savage	$1/(1+\exp(2z))^2$	非凸	近似对称($z\to\infty$ 时接近对称)
Barrier hinge	$\max(-b(r+z)+r,\max(b(z-r),r-z))$	凸	局部对称(对称范围$[-r,r]$)且非对称范围内函数值急增加

5　理论分析

观察式(6)中提出的风险，根据 α 的取值，$\widetilde{R}^{\alpha}_{PU}(f)$ 和 $\widetilde{R}^{\alpha}_{UN}(f)$ 其中之一将被优化。由于对二者的分析较为类似，因此在本节中以 $\widetilde{R}^{\alpha}_{PU}(f)$ 为例进行展示，其优化目标如下：

$$\widetilde{R}^{\alpha}_{PU}(f)=(1-\alpha)\widetilde{R}_{\pm}(f)+\alpha\widetilde{R}_{+\sharp}(f)。$$

事实上，由于数据分布未知，需通过有限样本上的经验风险 $\widetilde{R}^{\alpha}_{PU}(f,S)$ 对 $\widetilde{R}^{\alpha}_{PU}(f)$ 进行估计。那么，如何使从有限样本中学习得到的分类器 f 较好地泛化，使得额外风险(Excess risk)$\widetilde{R}^{\alpha}_{PU}(f)-\widetilde{R}^{\alpha}_{PU}(f,S)$ 尽可能小？

标准的额外风险分析方式[41]具有如下局限性。首先，大多数方法假设训练数据独立同分布。显然，该假设无法直接适用于 AUC 优化，因为 AUC 优化问题中训练样本是相互依赖的。其次，它们只考虑整个假设类(Hypothesis class)的最坏情况，而不考虑采用的学习算法本身。然而，由于模型通常被假定为经过了良好训练，这种最坏情况的假设几乎不会作为学习算法的输出出现，由此阻碍了对额外风险上界的进一步改进。

本文则通过考虑学习算法稳定性来为(半监督) AUC 优化提供更紧的额外风险上界。首次证明了在某些情况下，AUC 优化框架的额外风险可以从先前已知的$O(1/\sqrt{n^+}+1/\sqrt{n^-}+1/\sqrt{n^{\sharp}})$降为$O(1/n^++1/n^-+1/n^{\sharp})$。

证明思路如下。为了分析相互依赖的数据，借鉴文献[60-61]的思想，将相互依赖的变量集合分解为一系列独立的变量子集，并引入分数 Rademacher 复杂度(Fractional Rademacher complexity)。为了改进最坏情况假设框架，通过学习算法的一致参数稳定性(Uniform argument stability)，找到一个假设子类使得随机学习算法的期望输出以较高概率属于其中。然后，引入该假设子类的分数 Rademacher 复杂度，由于其依赖于算法，相较于传统算法无关的分数 Rademacher 复杂度更紧。之后，借助局部 Rademacher 复杂度技术[42,54]，推导出变形额外风险(Deformed excess risk)的一致上界，并将其应用于 SGD 学习算法。

由于篇幅限制，正文仅简要介绍主要结果，请参考附录 B～D 以获取详细推导。

5.1　变形额外风险的一致上界

设 f_S 为学习算法在数据集 S 上的输出。本文仅考虑线性假设类 $\mathcal{F}$，即 $f(\widetilde{X}_i)$ 可建模为一个线性函数 $\boldsymbol{w}^{\mathrm{T}}\widetilde{X}_i$，其中 $\boldsymbol{w}$ 为模型参数。因此可以表示为

$$f(\widetilde{X}_i)=\langle f,\widetilde{X}_i\rangle=\boldsymbol{w}^{\mathrm{T}}\widetilde{X}_i,\ \forall f\in\mathcal{F}。$$

令 $\widetilde{R}^{\alpha}_{PU}(f)$ 的经验版本为 $\widetilde{R}^{\alpha}_{PU}(f,S)$。类似地，将 $\widetilde{R}_{+-}(f)$ 和 $\widetilde{R}_{+\sharp}(f)$ 的经验版本定义为 $\widetilde{R}_{+-}(f,S)$ 和 $\widetilde{R}_{+\sharp}(f,S)$。变形额外风险定义为 $\widetilde{R}^{\alpha}_{PU}(f)-\dfrac{a}{a-1}\widetilde{R}^{\alpha}_{PU}(f,S)$，其中 $a>1$ 为常数。由于

$$\sup_{f\in\mathcal{B}_r}\left[\widetilde{R}^{a}_{PU}(f)-\frac{a}{a-1}\widetilde{R}^{a}_{PU}(f,S)\right]\leq$$
$$(1-\alpha)\sup_{f\in\mathcal{B}_r}\left[\widetilde{R}_{\pm}(f)-\frac{a}{a-1}\widetilde{R}_{\pm}(f,S)\right]+$$
$$\alpha\sup_{f\in\mathcal{B}_r}\left[\widetilde{R}_{+-}(f)-\frac{a}{a-1}\widetilde{R}_{+-}(f,S)\right] \tag{10}$$

可以分别计算不等式右侧两个上确界项的上界。由于这两项的证明思路相同，下面仅展示第一个上确界项的计算方式。有以下定理。

定理 1. 非正式版主要结果。假设损失函数 L-Lipschitz连续。输入特征的范数$\|\widetilde{X}_i\|_2$不大于 B_f。若学习算法$\mathcal{A}$是一致参数稳定的，其在数据集 S 上学到的分类器记为 f_S，令 $a>1$，则对任意 $\delta>0$，以下不等式以较高概率成立：

$$\widetilde{R}_{+-}(f_S)-\frac{a}{a-1}\widetilde{R}_{+-}(f_S,S)\leq O\left(\frac{\beta(n)}{\sqrt{\min(n^+,n^-)}}\right),$$

其中，$\beta(n)$取决于所采用算法的稳定性以及训练数据集大小。

5.2 随机梯度下降算法的一致上界

本节将上节中给出的结论应用于随机梯度下降(Stochastic Gradient Descent，SGD)算法。假设对于第 t 步迭代，随机选出一个正样本 i_+、负样本 i_- 和无标注样本 $i_\sharp$ 作为样本三元组，其中正样本 i_+ 和负样本 i_- 用于更新$\widetilde{R}_{+-}$，而正样本i_+和无标注样本 $i_\sharp$ 用于更新$\widetilde{R}_{+\sharp}$。记G_{f,η_t}为第 t 步迭代的 SGD 更新规则：

$$\boldsymbol{\theta}_t-\eta_t\mathrm{big}[(1-\alpha)\cdot\nabla_{\boldsymbol{\theta}_t}\ell(f,(Z_{i_+},Z_{i_-}))+$$
$$\alpha\cdot\nabla_{\boldsymbol{\theta}_t}\ell(f,(Z_{i_+},Z_{i_\sharp}))],$$

其中，η_t为第 t 步的学习率。

则可以给出如下采用 SGD 算法的鲁棒半监督 AUC 优化问题的更紧上界。其证明详见附录 D。

定理 2. SGD 算法的一致上界。假设损失函数是凸、L-Lipschitz 连续、s-平滑的，大小不超过 B_ℓ，输入特征的范数$\|\widetilde{X}_i\|_2$不大于B_f。记 f_T为T 步 SGD 的输出，其学习率满足 $\eta_t<2/s$。令 $a>1$，则对任意 $\delta>0$，以下不等式以至少 $1-2\delta$ 的概率成立：

$$\widetilde{R}^{a}_{PU}(f_T)-\frac{a}{a-1}\widetilde{R}^{a}_{PU}(f_T,S)\leq$$
$$\underbrace{\frac{(6a+8)B_\ell\log(1/\delta)}{3}\left(\frac{1-\alpha}{\min(n^+,n^-)}+\frac{\alpha}{\min(n^+,n^\sharp)}\right)}_{\text{置信度项}}+$$
$$\underbrace{8LB_f\sqrt{2B_f\log\frac{2}{\gamma}}\left(\frac{(1-\alpha)\cdot\beta(n)}{\sqrt{\min(n^+,n^-)}}+\frac{\alpha\cdot\beta(n)}{\sqrt{\min(n^+,n^\sharp)}}\right)}_{\text{复杂度项}},$$

且有 $\beta(n)\leq 2LB_f\sqrt{\frac{1}{n^+}+\frac{1}{n^-}+\frac{1}{n^\sharp}}$。

显然，由于

$$\frac{1-\alpha}{\min(n^+,n^-)}+\frac{\alpha}{\min(n^+,n^\sharp)}=O\left(\frac{1}{n^+}+\frac{1}{n^-}+\frac{1}{n^\sharp}\right),$$

且

$$\frac{(1-\alpha)\cdot\beta(n)}{\sqrt{\min(n^+,n^-)}}+\frac{\alpha\cdot\beta(n)}{\sqrt{\min(n^+,n^\sharp)}}=O\left(\frac{1}{n^+}+\frac{1}{n^-}+\frac{1}{n^\sharp}\right),$$

可以看出置信度项和复杂度项的阶数都已经降为 $O(1/n^++1/n^-+1/n^\sharp)$。因此，本定理给出了更紧的泛化上界。另一方面，上述分析中的假设条件均为现有优化理论分析中的常规假设[62]，较容易满足：(1) 输入特征范数有界。可以通过特征归一化实现；(2) 损失函数是凸的。Barrier hinge、Squared、Hinge 等损失函数均满足凸性；(3) 损失值有界。对于线性模型而言，当输入有界、参数大小有界(可由权重衰减实现)的情况下，损失函数输出通常也有界；(4) 损失函数 L-Lipschitz 连续。当损失输入、输出值有界时，上述损失函数均满足 Lipschitz 连续性定义(即次梯度有上界)，且 Barrier hinge 损失的 Lipschitz 常数 L 为超参数 b；(5) 损失函数 s-平滑。线性分段损失如 Barrier hinge 与 Hinge 在每个线性区间上均是平滑的，整体而言几乎处处平滑(smooth almost everywhere)，对理论分析过程并无影响。

5.3 $\widetilde{R}_{P\text{-or-}N}(f)$与$\widetilde{R}_{P\text{-and-}N}(f)$的对比

记式(6)中$\widetilde{R}_{P\text{-or-}N}(f)$的经验版本为$\widetilde{R}_{P\text{-or-}N}(f,S)$。可以得到使用随机梯度下降算法时的变形额外风险上界：

$$\widetilde{R}_{P\text{-or-}N}(f_S)-\frac{a}{a-1}\widetilde{R}_{P\text{-or-}N}(f_S,S)$$
$$\overset{(a)}{\leq}\xi(\delta)\left(\frac{1-\alpha}{\min(n^+,n^-)}+\frac{\alpha}{\min(n^+,n^\sharp)}\right)+$$
$$\zeta(n,\gamma)\left(\frac{1-\alpha}{\sqrt{\min(n^+,n^-)}}+\frac{\alpha}{\sqrt{\min(n^+,n^\sharp)}}\right),\alpha>0$$

或

$$\overset{(b)}{\leq}\xi(\delta)\left(\frac{1-\alpha}{\min(n^+,n^-)}+\frac{\alpha}{\min(n^-,n^\sharp)}\right)+$$
$$\zeta(n,\gamma)\left(\frac{1-\alpha}{\sqrt{\min(n^+,n^-)}}+\frac{\alpha}{\sqrt{\min(n^-,n^\sharp)}}\right),\alpha<0$$

其中

$$\xi(\delta)=\frac{(6a+8)B_\ell\log(1/\delta)}{3},$$

$$\zeta(n,\gamma)=16L^2B_f^2\sqrt{2B_f\log(2/\gamma)}\sqrt{\frac{1}{n^+}+\frac{1}{n^-}+\frac{1}{n^\sharp}}。$$

类似地，对于式(5)中 $\widetilde{R}_{P\text{-and-}N}(f)$ 的经验版本$\widetilde{R}_{P\text{-and-}N}(f,S)$有

$$\widetilde{R}_{P\text{-and-}N}(f_S)-\frac{a}{a-1}\widetilde{R}_{P\text{-and-}N}(f_S,S)\leq$$

$$\xi(\delta)\left(\frac{\alpha_1}{\min(n^+,n^-)}+\frac{\alpha_2}{\min(n^+,n^\sharp)}+\frac{\alpha_3}{\min(n^-,n^\sharp)}\right)+$$

$$\zeta(n,\gamma)\left(\frac{\alpha_1}{\sqrt{\min(n^+,n^-)}}+\frac{\alpha_2}{\sqrt{\min(n^+,n^\sharp)}}+\frac{\alpha_3}{\sqrt{\min(n^-,n^\sharp)}}\right)。$$

对于 $\widetilde{R}_{P\text{-or-}N}(f)$ 和 $\widetilde{R}_{P\text{-and-}N}(f)$ 的泛化误差而言，由于无标注样本的数量通常远大于有标注样本，因而仅考虑 $n^\sharp \gg n^+$ 且 $n^\sharp \gg n^-$ 的情况。此外，对于 P-and-N 风险，假设 $\alpha_1+\alpha_2+\alpha_3=1$。可以观察到：

(1) 当 $n^+>n^-$ 时，(a) 的阶更小；

(2) 当 $n^->n^+$ 时，(b) 的阶更小。

总而言之，额外风险的最小阶可从 P-or-N 风险之一中得到。因而相较于朴素的 P-and-N 风险，P-or-N 风险更好。

6　加速算法

本节从应用的角度出发，为 Barrier hinge 损失函数的损失和梯度估计设计了一种加速算法。不失一般性，考虑如下经验替代风险的一般形式：

$$\hat{R}^\ell=\frac{1}{n^+n^-}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sum_{\boldsymbol{x}^-\in\mathcal{X}^-}\ell(g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-)) \quad (11)$$

可以看到，替代损失的成对形式导致了 $O(n^2)$ 的计算复杂度。因此，为了降低复杂度，希望将正、负样本解耦。由于 Barrier hinge 损失是分段线性的，可以用其替换 ℓ，并依据线性区间将式(11)重写为

$$\hat{R}^{\ell_{\text{barrier}}}=$$

$$\underbrace{\frac{1}{n^+n^-}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sum_{\boldsymbol{x}^-\in\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}^+)}r-b\cdot r-b[g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-)]}_{\text{(A)}}+$$

$$\underbrace{\frac{1}{n^+n^-}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}}_{\text{(B)}}+$$

$$\underbrace{\frac{1}{n^+n^-}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sum_{\boldsymbol{x}^-\in\mathcal{Q}^{\mathrm{R}}(\boldsymbol{x}^+)}b[g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-)]-b\cdot r}_{\text{(C)}} \quad (12)$$

其中

$$\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}^+)=\{\boldsymbol{x}^-\mid g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-)\leq\frac{b\cdot r}{1-b};\boldsymbol{x}^-\in\mathcal{X}^-\},$$

$$\mathcal{Q}^{\mathrm{C}}(\boldsymbol{x}^+)=\left\{\boldsymbol{x}^-\,\middle|\,\frac{b\cdot r}{1-b}<g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-)<r;\boldsymbol{x}^-\in\mathcal{X}^-\right\},$$

$$\mathcal{Q}^{\mathrm{R}}(\boldsymbol{x}^+)=\{\boldsymbol{x}^-\mid g(\boldsymbol{x}^+)-g(\boldsymbol{x}^-)\geq r;\boldsymbol{x}^-\in\mathcal{X}^-\},$$

分别为与正样本 $\boldsymbol{x}^+$ 的得分差值落入 Barrier hinge 对应线性区间中的负样本集合。

6.1　项(A)的计算

式(12)的第一项可写为

$$\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\sigma^{\mathrm{L}}(\boldsymbol{x}^+)[r-b\cdot r-b\cdot g(\boldsymbol{x}^+)]+\varpi^{\mathrm{L}}(\boldsymbol{x}^+) \quad (13)$$

其中

$$\sigma^{\mathrm{L}}(\boldsymbol{x}^+)=\sum_{\boldsymbol{x}^-\in\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}^+)}\frac{1}{n^+n^-},$$

$$\varpi^{\mathrm{L}}(\boldsymbol{x}^+)=\sum_{\boldsymbol{x}^-\in\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}^+)}\frac{b}{n^+n^-}g(\boldsymbol{x}^-)。$$

注意到若 $\sigma^{\mathrm{L}}(\boldsymbol{x}^+)$ 和 $\omega^{\mathrm{L}}(\boldsymbol{x}^+)$ 已知，则式(13)的复杂度仅为 $O(n^+)$。因此，需高效计算 $\sigma^{\mathrm{L}}(\boldsymbol{x}^+)$ 和 $\varpi^{\mathrm{L}}(\boldsymbol{x}^+)$，以实现项(A)的整体高效计算。

(1) 样本排序。首先根据 $g(\cdot)$ 将正、负两类样本各自按降序排序，分别表示为 $\boldsymbol{x}_0^{+\downarrow},\boldsymbol{x}_1^{+\downarrow},\cdots,\boldsymbol{x}_{n^+-1}^{+\downarrow}$ 和 $\boldsymbol{x}_0^{-\downarrow},\boldsymbol{x}_1^{-\downarrow},\cdots,\boldsymbol{x}_{n^--1}^{-\downarrow}$。换言之，则有

$$g(\boldsymbol{x}_0^{+\downarrow})\geq g(\boldsymbol{x}_1^{+\downarrow})\geq\cdots\geq g(\boldsymbol{x}_{n^+-1}^{+\downarrow}),$$

$$g(\boldsymbol{x}_0^{-\downarrow})\geq g(\boldsymbol{x}_1^{-\downarrow})\geq\cdots\geq g(\boldsymbol{x}_{n^--1}^{-\downarrow}),$$

其中，排序操作的复杂度为 $O(n\log n)$。

(2) 递推计算。为简单起见，将 $\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}_k^{+\downarrow})$、$\mathcal{Q}^{\mathrm{C}}(\boldsymbol{x}_k^{+\downarrow})$ 和 $\mathcal{Q}^{\mathrm{R}}(\boldsymbol{x}_k^{+\downarrow})$ 分别记为 $\mathcal{Q}_k^{\mathrm{L}}$、$\mathcal{Q}_k^{\mathrm{C}}$ 和 $\mathcal{Q}_k^{\mathrm{R}}$。高效计算 σ^{L} 和 ϖ^{L} 的关键在于：

$$\mathcal{Q}_0^{\mathrm{L}}\subseteq\mathcal{Q}_1^{\mathrm{L}}\subseteq\cdots\subseteq\mathcal{Q}_{n^+-1}^{\mathrm{L}},$$

因此，有

$$\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}_{k+1}^{+\downarrow})=\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}_k^{+\downarrow})\cup\{\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}_{k+1}^{+\downarrow})\setminus\mathcal{Q}^{\mathrm{L}}(\boldsymbol{x}_k^{+\downarrow})\}。$$

则通过对所有排序后的正/负样本进行一次复杂度为 $O(n)$ 的高效动态规划，可以从 $\mathcal{Q}_0^{\mathrm{L}}$ 开始递推计算出所有 $\mathcal{Q}_k^{\mathrm{L}}$，$k=0,1,\cdots,n^+-1$。计算 $\mathcal{Q}_k^{\mathrm{L}}$ 的动态规划过程见算法 1。

算法 1.　项(A)中的索引计算。

输入：$\boldsymbol{x}_0^{+\downarrow},\boldsymbol{x}_1^{+\downarrow},\cdots,\boldsymbol{x}_{n^+-1}^{+\downarrow},\boldsymbol{x}_0^{-\downarrow},\boldsymbol{x}_1^{-\downarrow},\cdots,\boldsymbol{x}_{n^--1}^{-\downarrow},b,r$

输出：$\boldsymbol{I}^{\mathrm{L}}$

```
p ← 0.                                          ▷ 正样本指针
q ← 0.                                          ▷ 负样本指针
I^L_{0:(n^+-1)} ← −1                            ▷ Q^L_p = {x_k^{−↓}}_{k=0}^{I^L_p}
while p < n^+, q < n^− do
    if g(x_p^{+↓}) − g(x_q^{−↓}) < (b·r)/(1−b) then
        q ← q + 1.
    else
        I^L_p ← q − 1.                          ▷ I^L_p = −1 即 Q^L_p = ∅
        p ← p + 1.
    end if
end while
if p < n^+ then                                 ▷ 即 q = n^−
    I^L_{p:(n^+-1)} ← q − 1.
end if
```

随后，σ^{L}和ϖ^{L}可由如下递推计算得到：

$$\sigma^{\mathrm{L}}(\boldsymbol{x}_{k+1}^{+\downarrow})=\sigma^{\mathrm{L}}(\boldsymbol{x}_k^{+\downarrow})+\sum_{x^-\in\mathcal{Q}_{k+1}^{\mathrm{L}}\setminus\mathcal{Q}_k^{\mathrm{L}}}\frac{1}{n^+n^-},$$

$$\varpi^{\mathrm{L}}(\boldsymbol{x}_{k+1}^{+\downarrow})=\varpi^{\mathrm{L}}(\boldsymbol{x}_k^{+\downarrow})+\sum_{x^-\in\mathcal{Q}_{k+1}^{\mathrm{L}}\setminus\mathcal{Q}_k^{\mathrm{L}}}\frac{b}{n^+n^-}g(\boldsymbol{x}_k^{-\downarrow}) \quad (14)$$

显然，上述递推计算的时间复杂度为$O(n)$。

(3)求和。最后，计算各正样本上的损失值并进行求和即可完成式(13)的计算，该步复杂度为$O(n)$。

在梯度估计时，式(12)中项(A)关于参数Θ的梯度可写为

$$\sum_{x^+\in\mathcal{X}^+}\sigma^{\mathrm{L}}(\boldsymbol{x}^+)[-b\nabla_\Theta g(\boldsymbol{x}^+)]+\Xi^{\mathrm{L}}(\boldsymbol{x}^+) \quad (15)$$

其中

$$\Xi^{\mathrm{L}}(\boldsymbol{x}^+)=\sum_{x^-\in\mathcal{Q}^{\mathrm{L}}(x^+)}\frac{b}{n^+n^-}\nabla_\Theta g(\boldsymbol{x}^-) \quad (16)$$

类似地，$\Xi^{\mathrm{L}}(\boldsymbol{x}_k^+)$可由如下递推计算得到：

$$\Xi^{\mathrm{L}}(\boldsymbol{x}_{k+1}^{+\downarrow})=\Xi^{\mathrm{L}}(\boldsymbol{x}_k^{+\downarrow})+\sum_{x^-\in\mathcal{Q}_{k+1}^{\mathrm{L}}\setminus\mathcal{Q}_k^{\mathrm{L}}}\frac{b\cdot\nabla_\Theta g(\boldsymbol{x}^{-\downarrow})}{n^+n^-} \quad (17)$$

项(A)的详细损失和梯度计算过程见算法 2。

算法 2. 项(A)中的损失与梯度计算。

输入：$\boldsymbol{x}_0^{+\downarrow},\boldsymbol{x}_1^{+\downarrow},\cdots,\boldsymbol{x}_{n^+-1}^{+\downarrow}$及$\boldsymbol{x}_0^{-\downarrow},\boldsymbol{x}_1^{-\downarrow},\cdots,\boldsymbol{x}_{n^--1}^{-\downarrow},b,r,\boldsymbol{I}^{\mathrm{L}}$

输出：loss，grad

$t_0\leftarrow 0$，$p\leftarrow 0$.

$\sigma_{\mathrm{buff}}\leftarrow 0$，$\varpi_{\mathrm{buff}}\leftarrow 0$，$\Xi_{\mathrm{buff}}\leftarrow 0$.

$\boldsymbol{\sigma}_{0:(n^+-1)}\leftarrow 0$，$\boldsymbol{\varpi}_{0:(n^+-1)}\leftarrow 0$，$\boldsymbol{\Xi}_{0:(n^+-1)}\leftarrow 0$

while $p<n^+$ do

if $I_p^L\neq t_0-1$ then

$\sigma_{\mathrm{buff}}\leftarrow\sigma_{\mathrm{buff}}+\sum_{k=t_0}^{I_p^L}\frac{1}{n^+n^-}$.

$\varpi_{\mathrm{buff}}\leftarrow\varpi_{\mathrm{buff}}+\sum_{k=t_0}^{I_p^L}\frac{b}{n^+n^-}\cdot g(\boldsymbol{x}_k^{-\downarrow})$.

$\Xi_{\mathrm{buff}}\leftarrow\Xi_{\mathrm{buff}}+\sum_{k=t_0}^{I_p^L}\frac{b}{n^+n^-}\cdot\nabla_\Theta g(\boldsymbol{x}_k^{-\downarrow})$.

$t_0\leftarrow I_p^L+1$.

end if

$\sigma_p\leftarrow\sigma_{\mathrm{buff}}$，$\varpi_p\leftarrow\varpi_{\mathrm{buff}}$，$\Xi_p\leftarrow\Xi_{\mathrm{buff}}$.

$p\leftarrow p+1$.

end while

$\boldsymbol{g}^+\leftarrow[g(\boldsymbol{x}_0^{+\downarrow}),g(\boldsymbol{x}_1^{+\downarrow}),\cdots,g(\boldsymbol{x}_{n^+-1}^{+\downarrow})]^{\mathrm{T}}$

$\nabla_\Theta\boldsymbol{g}^+\leftarrow[\nabla_\Theta g(\boldsymbol{x}_0^{+\downarrow}),\nabla_\Theta g(\boldsymbol{x}_1^{+\downarrow}),\cdots,\nabla_\Theta g(\boldsymbol{x}_{n^+-1}^{+\downarrow})]^{\mathrm{T}}$

loss$\leftarrow\boldsymbol{\sigma}^{\mathrm{T}}(r-b\cdot r-b\cdot\boldsymbol{g}^+)+\mathbf{1}^{\mathrm{T}}\boldsymbol{\varpi}$

grad$\leftarrow -b\cdot\boldsymbol{\sigma}^{\mathrm{T}}\nabla_\Theta\boldsymbol{g}^++\mathbf{1}^{\mathrm{T}}\Xi$

6.2 项(C)的计算

同样地，由于

$$\mathcal{Q}_{n^+-1}^{\mathrm{R}}\subseteq\mathcal{Q}_{n^+-2}^{\mathrm{R}}\subseteq\cdots\subseteq\mathcal{Q}_0^{\mathrm{R}},$$

$\mathcal{Q}^{\mathrm{R}}(\boldsymbol{x}_k^{+\downarrow})$，$k=0,1,\cdots,n^+-1$也可通过单次数据集遍历来高效计算。因而，项(C)及其梯度也可以通过与上一小节类似的递推方式来计算。但有所不同的是，需从得分最小样本$x_{n^+-1}^{+\downarrow}$到最大样本$x_0^{+\downarrow}$的方向遍历。

6.3 项(B)的计算

式(12)的第二项，即项(B)，可写为

$$\sum_{x^+\in\mathcal{X}^+}\sigma^{\mathrm{C}}(\boldsymbol{x}^+)[r-g(\boldsymbol{x}^+)]+\varpi^{\mathrm{C}}(\boldsymbol{x}^+),$$

其中

$$\sigma^{\mathrm{C}}(\boldsymbol{x}^+)=\sum_{x^-\in\mathcal{Q}^{\mathrm{C}}(x^+)}\frac{1}{n^+n^-},$$

$$\varpi^{\mathrm{C}}(\boldsymbol{x}^+)=\sum_{x^-\in\mathcal{Q}^{\mathrm{C}}(x^+)}\frac{g(\boldsymbol{x}^-)}{n^+n^-}。$$

由于$\mathcal{Q}_k^{\mathrm{C}}$既非$\mathcal{Q}_{k-1}^{\mathrm{C}}$的子集，也非$\mathcal{Q}_{k+1}^{\mathrm{C}}$的子集，项(B)及其梯度的计算相对复杂。尽管有$\mathcal{Q}_k^{\mathrm{C}}=\mathcal{X}^-\setminus\mathcal{Q}_k^{\mathrm{L}}\setminus\mathcal{Q}_k^{\mathrm{R}}$，但其无法直接转化为递推计算规则。幸而，通过对$\mathcal{Q}_k^{\mathrm{L}}$和$\mathcal{Q}_k^{\mathrm{R}}$进一步分解，可得如下$\mathcal{Q}_k^{\mathrm{C}}$的递推关系：

$$\mathcal{Q}_{k+1}^{\mathrm{C}}=\mathcal{Q}_k^{\mathrm{C}}\setminus[\mathcal{Q}_{k+1}^{\mathrm{L}}\setminus\mathcal{Q}_k^{\mathrm{L}}]\cup[\mathcal{Q}_k^{\mathrm{R}}\setminus\mathcal{Q}_{k+1}^{\mathrm{R}}],$$

因而表明式(12)中项(B)的递推计算规则如下：

$$\sigma^{\mathrm{C}}(\boldsymbol{x}_{k+1}^{+\downarrow})=\sigma^{\mathrm{C}}(\boldsymbol{x}_k^{+\downarrow})-\sum_{x^-\in\mathcal{Q}_{k+1}^{\mathrm{L}}\setminus\mathcal{Q}_k^{\mathrm{L}}}\frac{1}{n^+n^-}+\sum_{x^-\in\mathcal{Q}_k^{\mathrm{R}}\setminus\mathcal{Q}_{k+1}^{\mathrm{R}}}\frac{1}{n^+n^-}。$$

$\varpi^{\mathrm{C}}(\boldsymbol{x}_k^{+\downarrow})$与$\Xi^{\mathrm{C}}(\boldsymbol{x}_k^+)$的计算与$\sigma^{\mathrm{C}}(\boldsymbol{x}_k^{+\downarrow})$类似，因而在此省略。具体算法流程见算法 3。

算法 3. 项(B)中的损失与梯度计算。

输入：$\{\boldsymbol{x}_k^{+\downarrow}\}_{k=1}^{n^+-1},\{\boldsymbol{x}_k^{-\downarrow}\}_{k=1}^{n^--1},b,r,\boldsymbol{I}^{\mathrm{L}},\boldsymbol{I}^{\mathrm{R}}$

输出：loss，grad

$t_0\leftarrow I_0^L+1$，$t_1\leftarrow I_0^R$，$p\leftarrow 0$.

$\sigma_{\mathrm{buff}}\leftarrow\sum_{k=t_0}^{t_1-1}\frac{1}{n^+n^-}$.

$\varpi_{\mathrm{buff}}\leftarrow\sum_{k=t_0}^{t_1-1}\frac{1}{n^+n^-}g(\boldsymbol{x}_k^{-\downarrow})$.

$\Xi_{\mathrm{buff}}\leftarrow\sum_{k=t_0}^{t_1-1}\frac{1}{n^+n^-}\nabla_\Theta g(\boldsymbol{x}_k^{-\downarrow})$

$\boldsymbol{\sigma}_{0:(n^+-1)}\leftarrow 0$，$\boldsymbol{\varpi}_{0:(n^+-1)}\leftarrow 0$，$\boldsymbol{\Xi}_{0:(n^+-1)}\leftarrow 0$

while $p<n^+$ do

if $I_p^L\neq t_0-1$ then

$\sigma_{\mathrm{buff}}\leftarrow\sigma_{\mathrm{buff}}+\sum_{k=t_0}^{I_p^L}\frac{1}{n^+n^-}$.

$\varpi_{\mathrm{buff}}\leftarrow\varpi_{\mathrm{buff}}+\sum_{k=t_0}^{I_p^L}\frac{1}{n^+n^-}\cdot g(\boldsymbol{x}_k^{-\downarrow})$.

$\Xi_{\mathrm{buff}}\leftarrow\Xi_{\mathrm{buff}}+\sum_{k=t_0}^{I_p^L}\frac{1}{n^+n^-}\cdot\nabla_\Theta g(\boldsymbol{x}_k^{-\downarrow})$.

$t_0\leftarrow I_p^L+1$.

end if

if $I_p^R\neq t_1$ then

$\sigma_{\mathrm{buff}}\leftarrow\sigma_{\mathrm{buff}}+\sum_{k=t_1}^{I_p^R-1}\frac{1}{n^+n^-}$.

$\varpi_{\mathrm{buff}}\leftarrow\varpi_{\mathrm{buff}}+\sum_{k=t_1}^{I_p^R-1}\frac{1}{n^+n^-}\cdot g(\boldsymbol{x}_k^{-\downarrow})$.

16. $\Xi_{\text{buff}} \leftarrow \Xi_{\text{buff}} + \sum_{k=t_1}^{I_p^R-1} \frac{1}{n^+ n^-} \cdot \nabla_\Theta g(\boldsymbol{x}_k^{-\downarrow})$.

17. $t_1 \leftarrow I_p^R$.

18. end if

19. $\sigma_p \leftarrow \sigma_{\text{buff}}$, $\varpi_p \leftarrow \varpi_{\text{buff}}$, $\Xi_p \leftarrow \Xi_{\text{buff}}$.

20. $p \leftarrow p+1$.

21. end while

22. $\boldsymbol{g}^+ \leftarrow [g(\boldsymbol{x}_0^{+\downarrow}), g(\boldsymbol{x}_1^{+\downarrow}), \cdots, g(\boldsymbol{x}_{n^+-1}^{+\downarrow})]^{\mathrm{T}}$

23. $\nabla_\Theta \boldsymbol{g}^+ \leftarrow [\nabla_\Theta g(\boldsymbol{x}_0^{+\downarrow}), \nabla_\Theta g(\boldsymbol{x}_1^{+\downarrow}), \cdots, \nabla_\Theta g(\boldsymbol{x}_{n^+-1}^{+\downarrow})]^{\mathrm{T}}$

24. $\text{loss} \leftarrow \boldsymbol{\sigma}^{\mathrm{T}}(r-\boldsymbol{g}^+) + \mathbf{1}^{\mathrm{T}}\boldsymbol{\varpi}$

25. $\text{grad} \leftarrow -\boldsymbol{\sigma}^{\mathrm{T}} \nabla_\Theta \boldsymbol{g}^+ + \mathbf{1}^{\mathrm{T}}\boldsymbol{\Xi}$

显然，上述三项计算过程中的最高时间复杂度来自预处理排序操作，为 $O(n\log n)$。尽管单一线性区间中的排序、递推、求和三个步骤无法并行，但不同线性区间(即项(A)/(B)/(C))的递推(或求和)计算均可以在一定程度上并行，以降低复杂度中的常数系数。由此，所提出的加速算法将损失和梯度计算的复杂度从 $O(n^2)$ 降为 $O(n\log n)$，可极大程度提高大规模数据集上的训练速度。在第 7 节的实验中也展示出了显著的加速比。此外，值得注意的是，所提出算法也可以轻松适配其他具有不超过 3 个线性区间的分段线性可导替代损失函数，如 Ramp 损失。

7 实　验

7.1 数据集

为了验证所提方法的有效性，在来自 Keel[①]、UCI[②]、Kaggle 的 15 个常用二分类数据集上进行了实验。这些数据集包括了类别平衡和不平衡的数据，并涵盖了金融、医学、生物信息等不同的应用场景。表 2 给出了数据集的详细统计信息。

表 2　数据集统计信息

数据集	来源	样本量	θ^+/θ^-	特征维度
australian	Keel	690	0.8016	9
banana	Keel	5300	0.8126	3
credit-g	UCI	1000	0.4286	25
heart	Keel	270	1.2500	14
ionosphere	Keel	351	0.5600	34
monk-2	Keel	432	1.1176	7
phoneme	Keel	5404	0.4154	6
saheart	Keel	462	0.5298	9
segment0	Keel	2308	0.1662	20
spambase	Keel	4597	0.6506	58
titanic	Keel	2201	0.4772	4
wdbc	Keel	569	1.6840	31
wine	Keel	1599	1.1492	12
yeast1	Keel	1484	0.4066	9
Surgical	Kaggle	14635	2.9661	25

7.2 实验细节

实验中得分函数简单采用了基本的线性分类模型 $y=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$，其中 $\boldsymbol{w}\in\mathbb{R}^d$，$d$ 为各数据集输入特征 $\boldsymbol{x}$ 的维度。Barrier hinge 损失函数的超参数固定为 $b=200$ 和 $r=5$。将正则项超参数 γ 的值固定为 0.5，并以 0.1 的间隔在 $[-0.9, 0.9]$ 内搜索 α 的值。对于所有数据集，均采用小批量梯度下降法，其初始学习率设置为 0.2，每 5 轮迭代以 0.999 的比例衰减。模型参数 $\boldsymbol{w}$ 使用正态分布 $\mathcal{N}(0,1)$ 进行初始化。所有实验均在配 Intel(R) Xeon(R) Silver 4110 CPU、Ubuntu 16.04.6 操作系统的服务器上进行，各代码在 Python 3.7 环境下使用 NumPy 1.16.2 实现，其中损失和梯度计算部分使用 Cython 0.29.6 实现。

数据集构建方面，通过分层抽样，将正负样本以 70∶15∶15 的比例划分为训练、验证、测试集。同时，移除训练集中 85% 的数据标签，并通过正负标签的翻转生成含噪声的半监督训练数据。由于控制带噪标签生成过程的参数 π, π' 受 θ^+, θ^- 限制，无法直接将相同的 π, π' 应用于所有数据集。因此，通过 $\rho\in[0, 0.5)$ 和 $k\in(0,1)$ 两个参数间接控制噪声比例。具体而言，设置 $1-\pi=\mathbb{I}[\theta^- > \rho]\cdot\rho + (1-\mathbb{I}[\theta^- > \rho])\cdot k\cdot\theta^-$ 和 $\pi'=\mathbb{I}[\theta^+ > \rho]\cdot\rho + (1-\mathbb{I}[\theta^+ > \rho])\cdot k\cdot\theta^+$，其中 $\mathbb{I}[\cdot]$ 为指示函数。除非特别说明，实验结果均在 $k=0.9$ 和 $\rho=0.42$ 场景下得到。不同 ρ 值下的结果也在后续实验结果中给出。

为提升实验可靠性，针对每个数据集独立生成了 15 组训练验证/测试集。基于这 15 个验证集的平均结果进行选择最佳超参数，并记录 15 个测试集上的平均结果。

7.3 对比方法

为了验证所提出算法的有效性，采用了如下两类方法进行对比。

(1) 朴素半监督 AUC 优化方法：

① SSRB[8]。首先根据各无标签样本最相似的有标签数据为其分配伪标签，然后基于 RankBoost 模型[38]在有、无标签数据上分别进行 AUC 优化。

② PNUAUC[10]。自适应将所有无标签数据视为正或负样本，并进行基于平方损失的 AUC 优化。由于该方法需要给定 θ^+ 的值，在实验中通过搜索以找到其最佳数值。

③ Samult[11]。在 PNUAUC 基础上进行了简

① https://sci2s.ugr.es/keel/datasets.php
② http://archive.ics.uci.edu/ml/datasets

化，同时将无标签数据视为正样本和负样本，无需估计 θ^+，同样采用平方替代损失。

(2) 基于对称替代损失的所提方法变体：

① sigAUC。采用 Sigmoid 损失作为 AUC 替代损失函数，该损失函数对称但非凸。

② rampAUC。采用 Ramp 损失作为替代损失函数，该损失函数同样对称但非凸。

③ savAUC。采用 Savage 损失作为替代损失函数，该函数在 z 趋近∞时接近对称。

④ hinAUC。采用 Hinge 损失作为替代损失函数，该函数在[−1,1]内满足对称条件。

⑤ Ours。本文所提方法，采用 Barrier hinge 损失并优化 $\mathcal{L}^{\ell}_{P\text{-or-}N}$。

7.4 结果分析

所有数据集上的实验结果见表 3。可以看出，所提方法在绝大多数数据集上都表现出色，始终优于其他对比方法。在所有对比方法中，SSRB 方法在大部分数据集上表现不佳。其原因在于该方法依赖有标签样本为无标签样本分配伪标签。因此，当有标签样本不可信时，其性能将大幅下降。而对比同样基于风险变换的 PNUAUC 和 Samult 方法，所提方法通过对标签噪声的建模获得了显著性能提升。此外，所提方法相较于其他变体的优势如下：

(1) 与其他局部近似对称损失函数(如 Hinge 和 Savage)相比，Barrier hinge 损失表现更好。其原因在于当 $b\gg1$ 时，该损失将对超出对称范围的得分差值进行严厉的惩罚。因此，随着训练进行，得分差值将逐渐落入其对称区域，使得损失/梯度评估最终均在对称范围内进行。该特性使得 Barrier hinge 损失与对称损失函数的区别不大。

表 3 15 个数据集上的性能对比(15 次实验的平均 AUC 以及标准差)

	Ours	SSRB	PNUAUC	Samult	hinAUC	savAUC	sigAUC	rampAUC
australian	**.6531**(.0617)	.5529(.0681)*	.6002(.0839)*	.5900(.1063)*	.5671(.1183)*	.6076(.1077)	.6303(.0820)	.6352(.0860)
banana	**.5688**(.0366)	.5309(.0711)	.5084(.0355)*	.5046(.0346)*	.5346(.0420)*	.5181(.0425)*	.5302(.0404)*	.5165(.0432)*
credit-g	**.6072**(.0623)	.5130(.0213)*	.5457(.0924)*	.5476(.0916)*	.5727(.0867)	.5797(.0776)	.5379(.1042)*	.5766(.0764)
heart	**.7134**(.1366)	.5919(.0846)*	.6288(.1881)*	.6291(.1724)*	.6245(.1832)*	.6425(.2017)*	.5090(.2375)*	.6818(.1156)*
ionosphere	.6813(.1126)	.5473(.0533)*	.6483(.1320)	.6542(.1260)	.6762(.1555)	.6728(.1333)	.7317(.1009)	.6913(.1189)
monk-2	**.6960**(.1377)	.5206(.0755)*	.6308(.1432)	.6345(.1471)	.6689(.1528)	.6170(.1699)*	.6708(.1184)	.6879(.1189)
yeast1	**.7196**(.0774)	.5354(.0714)*	.6841(.1093)*	.6785(.1115)*	.6503(.1614)*	.6766(.1275)*	.6973(.0928)*	.6311(.1842)
segment0	**.6823**(.1134)	.5833(.1182)*	.6299(.0703)*	.6228(.0610)*	.6125(.0748)*	.6102(.0555)*	.6222(.0699)*	.6383(.0869)
wine	**.5872**(.0963)	.5276(.0459)*	.5282(.1300)*	.5267(.1202)*	.5624(.1103)	.5823(.1111)	.5655(.1220)	.5826(.0969)
wdbc	**.9533**(.0303)	.6066(.1219)*	.9222(.0615)	.9378(.0270)*	.8059(.3045)*	.8476(.2743)	.8883(.2170)	.8810(.2298)
phoneme	**.7844**(.0194)	.5428(.0507)*	.7503(.0454)*	.7467(.0485)*	.7662(.0346)*	.7717(.0349)	.7721(.0319)*	.7672(.0367)*
saheart	**.6253**(.0917)	.5064(.0477)*	.5985(.0812)*	.5999(.0840)*	.5928(.1203)	.5867(.1111)*	.5804(.1427)	.5632(.1182)
Surgical	**.6095**(.0365)	.5508(.0553)*	.5771(.0513)*	.5901(.0531)*	.5917(.0502)	.5864(.0907)	.5782(.0592)*	.5783(.0836)
spambase	**.6647**(.0696)	.5220(.0320)*	.6222(.0748)*	.6367(.0834)	.6241(.1090)*	.6325(.0906)	.6409(.1069)	.6381(.0642)*
titanic	.7053(.0429)	.5532(.0526)*	.6400(.1439)	.6390(.1426)	.7116(.0225)	.6408(.1514)	.6211(.1602)	.6506(.1399)
win/tie/lose	—	14/1/0	11/4/0	11/4/0	8/7/0	6/9/0	7/8/0	4/11/0

注：最优结果用粗体表示，次好结果用下划线表示(*表示所提方法在相应数据集上显著优于对比方法(在 95% 显著水平下进行配对 t 检验))。

为了证明这一点，可以观察加速算法所使用的 $n^+\sigma^C$ 值的变化情况，其中 $n^+\sigma^C_i$ 恰好是落在第 i 个正样本的对称范围内的负样本比例。即 $n^+\sigma^C_i$ 越大，就会有更多的正负样本对被推入对称范围。其变化如图 1 所示。可以观察到，Barrier hinge 损失逐渐将得分差推入对称范围内。

(2) 与对称损失函数(Sigmoid 和 Ramp 损失)相比，Barrier hinge 损失的结果也相对更好。这可能是因为全局对称的损失函数是非凸的，可能导致局部最优解，而 Barrier hinge 损失是凸的，存在全局最优解。然而在所提方法的变体中，对称损失的性能大多数时候优于 Hinge 和 Savage 等近似/局部对称损失。这也再次侧面印证了输入得分差值可能经常落入 Hinge 和 Savage 等的非对称区域，此时若损失函

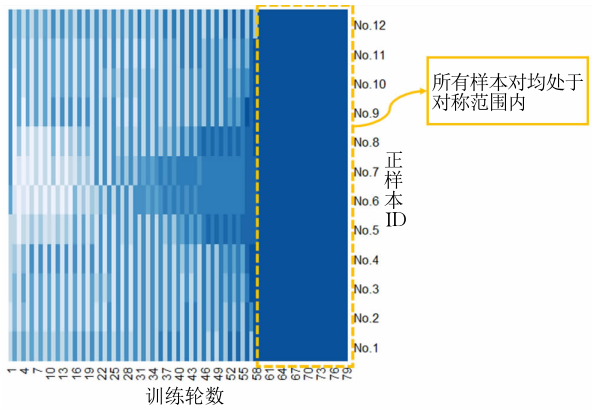
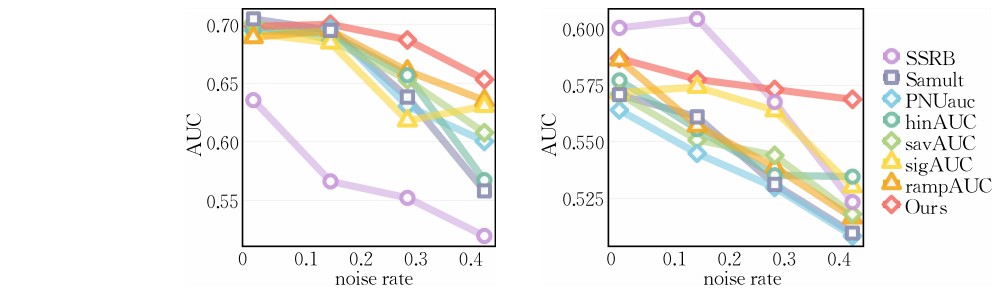


图 1 训练过程中 $n^+\boldsymbol{\sigma}^C\in[0,1]$ 的变化(蓝色越深，值越大。$n^+\sigma^C_i$ 代表落在第 i 个正样本的 Barrier hinge 损失对称范围内的负样本比例。图示表明，随着训练的进行，几乎所有得分差值都进入了对称范围)

数无法尽快将得分差值推入对称区域，将明显降低模型的鲁棒性。

7.5　不同噪声比例下的性能

各方法在不同噪声比例下（即 $\rho=0.0$、0.14、0.28、0.42 时）的性能见图 2(a)和(b)。可以发现，对比方法中，SSRB 方法在不同数据集上的性能波动较大，且受噪声影响最大，其他方法在噪声比例大的时候性能均有明显下降。而随着噪声增多，所提方法逐渐展示出相较于对比方法更大的优势，其性能更加稳定，说明了其对标签噪声的鲁棒性。



(a) australian　　(b) banana

图 2　不同噪声比例下的结果（随着噪声比例增大，所提方法的性能优势更加明显）

7.6　*P*-or-*N* 与 *P*-and-*N* 风险的对比

由于所提出的三个风险可以恢复理想的 AUC 风险，进一步研究了其不同组合的影响，结果如图 3 所示。从图中可以观察到，在大多数情况下，*P*-or-*N* 优于其他组合。一方面，这表明无标注数据的确对使用所提出的 $\widetilde{R}^{\ell}_{+\#}$ 或 $\widetilde{R}^{\ell}_{\#-}$ 风险进行优化的分类器的学习有较大帮助。另一方面，*P*-and-*N* 有时比仅使用有标注数据的模型表现更差（如 saheart、monk-2 和 titanic 数据集），表明将无标注数据同时视为正样本和负样本并非总是有效。然而，所提方法始终能够获得更好的结果。

7.7　敏感度分析

b 和 r 的影响：b 和 r 是 Barrier hinge 损失中的超参数，其敏感度分析结果分别见图 4(a)和(b)。图中表明，大幅改变 b 和 r 的值后模型 AUC 变化不大，因此这两者对所提方法的性能影响均较小。

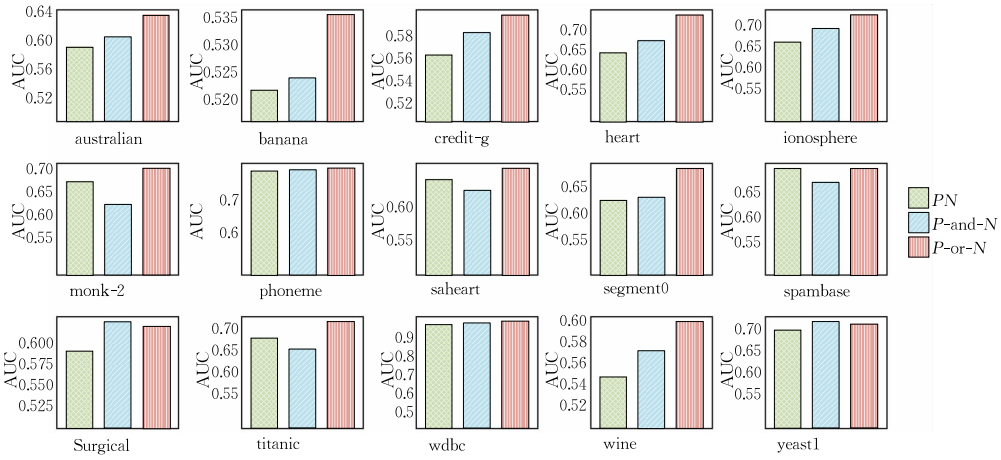
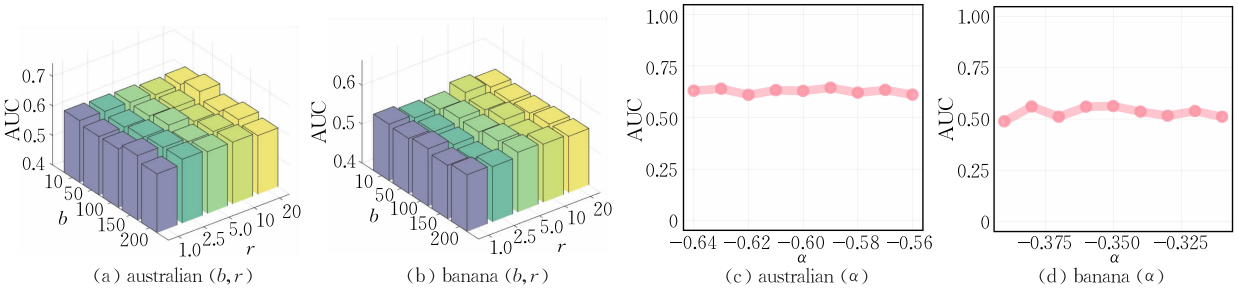


图 3　对三种风险不同组合的研究（*PN* 代表仅优化 $\widetilde{R}_{+-}$ 风险。*P*-and-*N* 和 *P*-or-*N*(Ours)分别代表优化 $\widetilde{R}_{P\text{-and-}N}$ 和 $\widetilde{R}_{P\text{-or-}N}$。结果显示，所提方法大部分时候优于 *P*-and-*N* 组合）



(a) australian (b,r)　(b) banana (b,r)　(c) australian (α)　(d) banana (α)

图 4　超参数 b、r 和 α 的敏感度分析（Barrier hinge 损失对 $b\geq 1$ 和 r 并不敏感，α 的影响也不大）

α 的影响:该超参数用于均衡正负样本上的经验误差和正(或负)样本、无标注样本上的经验误差项。显然,α 是一个较为重要的超参数,因此仅在最优结果的超参数值附近进行了小范围的敏感性分析,结果见图 4(c)和(d)。从中可以看出,所提方法在一定范围内均可保持较高的性能,因而在实际应用中无需对 α 的取值进行精细的搜索即可获得良好的性能。

7.8 计算复杂度分析

前文提到所提加速算法可将损失和梯度估计的计算复杂度从 $O(n^2)$ 降为 $O(n\log n)$。为验证该结论,在 $n^+:n^-=1:1$ 的情况下,记录采用加速算法和未采用加速算法的模型在不同样本量 n 下的训练时间,其加速比见图 5。可以看出,加速比与样本大小成正比,与理论分析一致,且在数据量较大时,加速比可达 200 倍,进一步说明了所提算法的有效性。在真实数据集上,PNUAUC 和 Samult 未经加速,而所提方法及其变体(Ours、sigAUC、hinAUC 等)均使用了加速算法,其训练时间对比也基本符合图 5 中的结果。因此,所提算法可以显著加速大规模数据集上的训练。

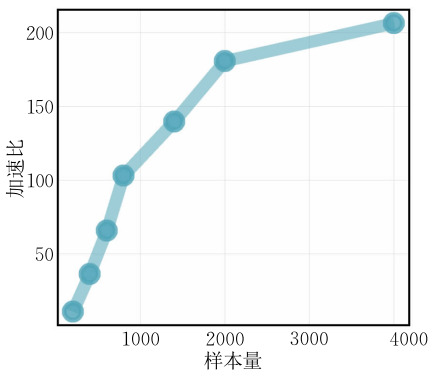


图 5 所提加速算法的加速比

8 结 论

现有的半监督 AUC 优化方法均假设有标注数据干净、可信,难以适用于实际场景。鉴于此,本文首次尝试解决了在不准确和不完整标注下的 AUC 优化问题。首先,理论上展示了在对称损失函数的帮助下,由含噪声的正样本/负样本、含噪声的正样本/无标注样本和含噪声的负样本/无标注样本所得到的 AUC 替代风险均能恢复理想 AUC 替代风险。基于此,提出了一个鲁棒的半监督 AUC 优化框架,无需噪声比例和类别先验估计即可间接对理想 AUC 替代风险进行优化。对所提框架的泛化性能分析表明,所提框架学习到的分类器可较好地泛化到未见的数据。为了实际应用,进一步提出了一种加速算法,将损失和梯度评估的复杂度从 $O(n^2)$ 降至 $O(n\log n)$。15 个真实数据集上的实验结果展现了所提方法的优越性。

另一方面,本文遵循了风险最小化相关标签噪声学习方法的假设,采用了常见的类别条件噪声模型,无法覆盖实际场景中的所有噪声特性。实例相关噪声以及更为复杂的混合噪声等情况将作为未来的拓展研究方向之一,以提高模型的普适性。

参 考 文 献

[1] Narasimhan H, Agarwal S. SVMpAUCtight: A new support vector method for optimizing partial AUC based on a tight convex upper bound//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 167-175

[2] Gao W, Jin R, Zhu S, et al. One-pass AUC optimization//Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 906-914

[3] Gao W, Zhou Z. On the consistency of AUC pairwise optimization//Proceedings of the International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 939-945

[4] Calders T, Jaroszewicz S. Efficient AUC optimization for classification//Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery. Berlin, Germany, 2007: 42-53

[5] Zhao P, Hoi S, Jin R, et al. Online AUC maximization//Proceedings of the International Conference on Machine Learning. Bellevue, USA, 2011: 233-240

[6] Ying Y, Wen L, Lyu S. Stochastic online AUC maximization//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 451-459

[7] Liu M, Yuan Z, Ying Y, et al. Stochastic AUC maximization with deep neural networks//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-9

[8] Amini M R, Truong T V, Goutte C. A boosting algorithm for learning bipartite ranking functions with partially labeled data//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 99-106

[9] Fujino A, Ueda N. A semi-supervised AUC optimization method with generative models//Proceedings of the 2016 IEEE 16th International Conference on Data Mining. Barcelona, Spain, 2016: 883-888

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

[11] Xie Z, Li M. Semi-supervised AUC optimization without guessing labels of unlabeled data//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 4310-4317

[12] Xia X, Liu T, Han B, et al. Part-dependent label noise: Towards instance-dependent label noise//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 7597-7610

[13] Cheng J, Liu T, Ramamohanarao K, et al. Learning with bounded instance and label-dependent label noise//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 1789-1799

[14] Wheway V. Using boosting to detect noisy data//Proceedings of the PRICAI 2000 Workshop. Melbourne, Australia, 2000: 123-132

[15] Malach E, Shalev-Shwartz S. Decoupling "when to update" from "how to update"//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 960-970

[16] Song H, Kim M, Lee J G. SELFIE: Refurbishing unclean samples for robust deep learning//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 5907-5915

[17] Lyu Y, Tsang I W. Curriculum loss: Robust learning and generalization against label corruption//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 10-18

[18] Jiang L, Zhou Z, Leung T, et al. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 2309-2318

[19] Han B, Yao Q, Yu X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 8536-8546

[20] Yu X, Han B, Yao J, et al. How does disagreement help generalization against label corruption?//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 7164-7173

[21] Wang Y, Liu W, Ma X, et al. Iterative learning with open-set noisy labels//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8688-8696

[22] Shen Y, Sanghavi S. Learning with bad training data via iterative trimmed loss minimization//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 5739-5748

[23] Patrini G, Rozza A, Menon A K, et al. Making deep neural networks robust to label noise: A loss correction approach//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2233-2241

[24] Hendrycks D, Mazeika M, Wilson D, et al. Using trusted data to train deep networks on labels corrupted by severe noise//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 10477-10486

[25] Arazo E, Ortego D, Albert P, et al. Unsupervised label noise modeling and loss correction//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 312-321

[26] Wang R, Liu T, Tao D. Multiclass learning with partially corrupted labels. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2568-2580

[27] Liu T, Tao D. Classification with noisy labels by importance reweighting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(3): 447-461

[28] Xiao T, Xia T, Yang Y, et al. Learning from massive noisy labeled data for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2691-2699

[29] Goldberger J, Ben-Reuven E. Training deep neural-networks using a noise adaptation layer//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017: 1-9

[30] Bekker A J, Goldberger J. Training deep neural-networks based on unreliable labels//Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 2682-2686

[31] Han B, Yao J, Niu G, et al. Masking: A new perspective of noisy supervision//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 5841-5851

[32] Yao J, Wang J, Tsang I W, et al. Deep learning from noisy image labels with quality embedding. IEEE Transactions on Image Processing, 2019, 28(4): 1909-1922

[33] Manwani N, Sastry P. Noise tolerance under risk minimization. IEEE Transactions on Cybernetics, 2013, 43(3): 1146-1151

[34] Ghosh A, Manwani N, Sastry P. Making risk minimization tolerant to label noise. Neurocomputing, 2015, 160: 93-107

[35] van Rooyen B, Menon A K, Williamson R C. Learning with symmetric label noise: The importance of being unhinged//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2015: 10-18

[36] Charoenphakdee N, Lee J, Sugiyama M. On symmetric losses for learning from corrupted labels//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 961-970

[37] Freund Y, Iyer R D, Schapire R E, et al. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003, 4(11): 933-969

[38] Wang S, Li D, Petrick N, et al. Optimizing area under the ROC curve using semi-supervised learning. Pattern Recognition, 2015, 48(1): 276-287

[39] Xie Z, Li M. Cutting the software building efforts in continuous integration by semi-supervised online AUC optimization//Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 2875-2881

[40] Agarwal S, Graepel T, Herbrich R, et al. Generalization bounds for the area under the ROC curve. Journal of Machine Learning Research, 2005, 6(4): 393-425

[41] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. Second Edition. USA: MIT Press, 2018

[42] Bartlett P L, Bousquet O, Mendelson S. Local Rademacher complexities. The Annals of Statistics, 2005, 33(4): 1497-1537

[43] Rogers W, Wagner T. A finite sample distribution-free performance bound for local discrimination rules. The Annals of Statistics, 1978, 6(3): 506-514

[44] Devroye L, Wagner T J. Distribution-free inequalities for the deleted and holdout error estimates. IEEE Transactions on Information Theory, 1979, 25(2): 202-207

[45] Devroye L, Wagner T. Distribution-free performance bounds for potential function rules. IEEE Transactions on Information Theory, 1979, 25(5): 601-604

[46] Bousquet O, Elisseeff A. Stability and generalization. Journal of Machine Learning Research, 2002, 2(Mar): 499-526

[47] Kutin S, Niyogi P. Almost-everywhere algorithmic stability and generalization error//Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence. Edmonton, Canada, 2002: 275-282

[48] Zhang T. Leave-one-out bounds for kernel methods. Neural Computing, 2003, 15(6): 1397-1437

[49] Liu T, Tao D, Song M, et al. Algorithm-dependent generalization bounds for multi-task learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(2): 227-241

[50] Xu C, Liu T, Tao D, et al. Local Rademacher complexity for multi-label learning. IEEE Transactions on Image Processing, 2016, 25(3): 1495-1507

[51] Pensia A, Jog V, Loh P. Generalization error bounds for noisy, iterative algorithms//Proceedings of the 2018 IEEE International Symposium on Information Theory. Vail, USA, 2018: 546-550

[52] Mou W, Wang L, Zhai X, et al. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints//Proceedings of the Conference on Learning Theory. Stockholm, Sweden, 2018: 605-638

[53] Chen Y, Jin C, Yu B. Stability and convergence trade-off of iterative optimization algorithms. arXiv preprint arXiv:1804.01619, 2018

[54] Liu T, Lugosi G, Neu G, et al. Algorithmic stability and hypothesis complexity//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2159-2167

[55] Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, 143(1): 29-36

[56] Scott C, Blanchard G, Handy G. Classification with asymmetric label noise: Consistency and maximal denoising//Proceedings of the Conference on Learning Theory. NJ, USA, 2013: 489-511

[57] Menon A, van Rooyen B, Ong C S, et al. Learning from corrupted binary labels via class-probability estimation//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 125-134

[58] Wei J, Zhu Z, Cheng H, et al. Learning with noisy labels revisited: A study using real-world human annotations//Proceedings of the International Conference on Learning Representations. Virtual Event, 2022: 1-9

[59] Sakai T, Plessis M C, Niu G, et al. Semi-supervised classification based on classification from positive and unlabeled data//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2998-3006

[60] Amini M R, Usunier N. Learning with Partially Labeled and Interdependent Data. New York, USA: Springer, 2015

[61] Usunier N, Amini M, Gallinari P. Generalization error bounds for classifiers trained with interdependent data//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2005: 1369-1376

[62] Hardt M, Recht B, Singer Y. Train faster, generalize better: Stability of stochastic gradient descent//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1225-1234

[63] Janson S. Large deviations for sums of partly dependent random variables. Random Structures & Algorithms, 2004, 24(3): 234-248

附录 A. 命题 1 的证明。

对于 $\widetilde{R}_{+\sharp}(f)$,有如下引理:

引理 1. 令 $\gamma(\boldsymbol{x},\boldsymbol{x}')=\ell(f(\boldsymbol{x},\boldsymbol{x}'))+\ell(f(\boldsymbol{x}',\boldsymbol{x}))$. 则 $\widetilde{R}_{+\sharp}(f)$可写为

$$\widetilde{R}_{+\sharp}=(\pi-\theta^+)R(f)+\underbrace{(1-\pi)\theta^+\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\gamma(\boldsymbol{x}^+,\boldsymbol{x}^-)]}_{\text{额外项}}+\underbrace{\frac{\pi\theta^+}{2}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}[\gamma(\boldsymbol{x}'^+,\boldsymbol{x}^+)]}_{\text{额外项}}+\underbrace{\frac{(1-\pi)\theta^-}{2}\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\gamma(\boldsymbol{x}'^-,\boldsymbol{x}^-)]}_{\text{额外项}}.$$

上述引理表明,所提出的基于含噪声正样本和无标注样本的 AUC 替代风险等价于理想的 AUC 替代风险,再加上一些与优化过程无关的多余项。幸运的是,通过对称替代损失函数 ℓ,可以得到一个有效的解决方案。

推论 1. 令 ℓ 为对称损失,即 $\ell(z)+\ell(-z)=\text{const}$. 则 $\widetilde{R}_{+\sharp}(f)$可简化为

$$\widetilde{R}_{+\sharp}(f)=(\pi-\theta^+)R(f)+\frac{\theta^++1-\pi}{2}\text{const} \tag{18}$$

证明. 将含噪声的正例和无标注数据分布代入风险 $\widetilde{R}_{+\sharp}$中,有

$$\widetilde{R}^{\ell}_{+\sharp}=\pi\theta^+\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}'^+,\boldsymbol{x}^+))]+(1-\pi)\theta^+\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^-,\boldsymbol{x}^+))]+\pi\theta^-\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-))]+$$

$$(1-\pi)\theta^- \mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}'^-,\boldsymbol{x}^-))] \quad (19)$$

设 $\gamma^\ell(\boldsymbol{x},\boldsymbol{x}')=\ell(f(\boldsymbol{x},\boldsymbol{x}'))+\ell(f(\boldsymbol{x}',\boldsymbol{x}))$，并记

$$\begin{aligned}
A&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))]\\
B&=\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^-,\boldsymbol{x}^+))]\\
C&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-))]=R^\ell(f)\\
D&=\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}'^-,\boldsymbol{x}^-))]\\
\gamma^\ell&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-))+\ell(f(\boldsymbol{x}^-,\boldsymbol{x}^+))]\\
&=B+C
\end{aligned} \quad (20)$$

首先需证明：

$$\begin{aligned}
A&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))]\\
&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}\left[\frac{\gamma^\ell(\boldsymbol{x}^+,\boldsymbol{x}^{++})}{2}\right]
\end{aligned} \quad (21)$$

为证明该式，有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))]\\
&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\mathbb{I}_{\boldsymbol{x}^+=\boldsymbol{x}'^+}\ell(0)+\mathbb{I}_{\boldsymbol{x}^+\neq\boldsymbol{x}'^+}\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))]\\
&=0+\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[1\times\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))]\\
&=\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}\left[\frac{\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))+\ell(f(\boldsymbol{x}'^+,\boldsymbol{x}^+))}{2}\right]
\end{aligned} \quad (22)$$

类似地，有

$$D=\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}\left[\frac{\gamma^\ell(\boldsymbol{x}^-,\boldsymbol{x}'^-)}{2}\right] \quad (23)$$

将式(20)、(21)和式(23)代入式(19)，有

$$\begin{aligned}
\widetilde{R}^\ell_{+\sharp}&=\pi\theta^+A+(1-\pi)\theta^+B+\pi\theta^-C+(1-\pi)\theta^-D\\
&=\pi\theta^+A+(1-\pi)\theta^+(\gamma^\ell-C)+\pi\theta^-C+(1-\pi)\theta^-D\\
&=\pi\theta^+A+(1-\pi)\theta^+\gamma^\ell+(\pi-\theta^++\pi)C+\\
&\quad(1-\pi)\theta^-D\\
&=(\pi-\theta^+)R^\ell(f)+\\
&\quad(1-\pi)\theta^+\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\gamma^\ell(\boldsymbol{x}^+,\boldsymbol{x}^-)]+\\
&\quad\frac{\pi\theta^+}{2}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}[\gamma^\ell(\boldsymbol{x}'^+,\boldsymbol{x}^+)]+\\
&\quad\frac{(1-\pi)\theta^-}{2}\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\gamma^\ell(\boldsymbol{x}'^-,\boldsymbol{x}^-)]
\end{aligned} \quad (24)$$

证毕.

类似$\widetilde{R}_{+\sharp}$，可以得到 $\widetilde{R}_{\sharp-}$ 对应的引理与推论。

引理 2.　设 $\gamma(\boldsymbol{x},\boldsymbol{x}')=\ell(f(\boldsymbol{x},\boldsymbol{x}'))+\ell(f(\boldsymbol{x}',\boldsymbol{x}))$。则 $\widetilde{R}_{\sharp-}(f)$可写为

$$\begin{aligned}
\widetilde{R}_{\sharp-}&=(\theta^+-\pi')R(f)+\\
&\quad\underbrace{\pi'\theta^-\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\gamma(\boldsymbol{x}^+,\boldsymbol{x}^-)]}_{\text{额外项}}+\\
&\quad\underbrace{\frac{\pi'\theta^+}{2}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\gamma(\boldsymbol{x}^+,\boldsymbol{x}'^+)]}_{\text{额外项}}+\\
&\quad\underbrace{\frac{(1-\pi')\theta^-}{2}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}[\gamma(\boldsymbol{x}^-,\boldsymbol{x}'^-)]}_{\text{额外项}}.
\end{aligned}$$

证明.　该引理的证明与引理 1 的证明相似。

$$\begin{aligned}
\widetilde{R}^\ell_{\sharp-}&=\pi'\theta^+\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}'^+))]+\\
&\quad(1-\pi')\theta^+\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}^+,\boldsymbol{x}^-))]+\\
&\quad\pi'\theta^-\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}[\ell(f(\boldsymbol{x}^-,\boldsymbol{x}^+))]+\\
&\quad(1-\pi')\theta^-\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}[\ell(f(\boldsymbol{x}^-,\boldsymbol{x}'^-))]\\
&=\pi'\theta^+A+(1-\pi')\theta^+C+\pi'\theta^-B+(1-\pi')\theta^-D\\
&=\pi'\theta^+A+(1-\pi')\theta^+C+\pi'\theta^-(\gamma^\ell-C)+(1-\pi')\theta^-D\\
&=\pi'\theta^+A+(\theta^+-\pi')C+\pi'\theta^-\gamma^\ell+(1-\pi')\theta^-D\\
&=(\theta^+-\pi')R^\ell(f)+\pi'\theta^-\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}[\gamma^\ell(\boldsymbol{x}^+,\boldsymbol{x}^-)]+\\
&\quad\frac{\pi'\theta^+}{2}\mathbb{E}_{\boldsymbol{x}^+\sim\mathcal{D}_P}\mathbb{E}_{\boldsymbol{x}'^+\sim\mathcal{D}_P}[\gamma^\ell(\boldsymbol{x}^+,\boldsymbol{x}'^+)]+\\
&\quad\frac{(1-\pi')\theta^-}{2}\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{D}_N}\mathbb{E}_{\boldsymbol{x}'^-\sim\mathcal{D}_N}[\gamma^\ell(\boldsymbol{x}^-,\boldsymbol{x}'^-)]
\end{aligned} \quad (25)$$

其中，A、B、C、D 和 γ^ℓ 的定义与推论 1 证明中的相同。　证毕.

类似地，有

推论 2.　令 ℓ 为对称损失，即 $\ell(z)+\ell(-z)=\text{const}$. 则 $\widetilde{R}_{\sharp-}(f)$可简化为

$$\widetilde{R}_{\sharp-}(f)=(\theta^+-\pi')R(f)+\frac{\pi'+\theta^-}{2}\text{const} \quad (26)$$

$\widetilde{R}_{+-}$ 与 R 间的线性关系证明见文献[36]，结合其与推论 1、2，可以得到命题 1。

附录 B.　相互依赖数据的泛化误差上界。

本节中将引入分数 Rademacher 复杂度的概念，为此，需要将成对的数据集转化为逐实例的数据集。因此，首先重新陈述问题设置，所使用的符号表示见表 4。

表 4　理论分析中使用的符号表示

符号	描述
S	原始数据集
S^i	与 S 仅在第 i 个样本上不同的数据集
n	数据集 S 的样本量
X	数据集 S 的一个样本特征
T	数据集变换
$\widetilde{S}_1,\widetilde{S}_2$	变换后的数据集
$\widetilde{N}_1,\widetilde{N}_2$	$\widetilde{N}_i=\lvert\widetilde{S}_i\rvert, i=1,2$
N	数据集 $\widetilde{S}$ 的样本量
$\widetilde{X}$	数据集 $\widetilde{S}$ 的一个样本特征
$\widetilde{Y}$	数据集 $\widetilde{S}$ 的一个样本标签
$\widetilde{Z}$	$\widetilde{Z}=(\widetilde{X},\widetilde{Y})$
$\mathcal{V}_k$	$\widetilde{S}$ 的分数覆盖的第 k 个子集
ω_k	顶点子集$\mathcal{V}_k$的权重
$\lvert\mathcal{V}_k\rvert$	$\mathcal{V}_k$的样本量
$\chi(\widetilde{S})$	$\widetilde{S}$ 的分数染色数
$\mathcal{A}$	学习算法
f_S	算法$\mathcal{A}$得到的分类器
$\mathcal{F}$	假设类
B_r	以$\mathbb{E}_{A,S}f_S$为中心的$\mathcal{F}$的子集
σ	Rademacher 随机变量
δ,γ	置信度参数
$R(\ell,S)$	S 上使用损失函数 ℓ 计算得到的经验损失
$\mathfrak{R}$	Rademacher 复杂度
$\lVert X\rVert_2$	随机变量 X 的范数
B_f	$\lVert X\rVert_2$的上界
B_ℓ	损失函数 ℓ 的上界
L	损失函数 ℓ 的 Lipschitz 常数
η_t	SGD 第 t 步的学习率
$\alpha_+(n)$	正样本的稳定性参数
$\alpha_-(n)$	负样本的稳定性参数
$\alpha_\sharp(n)$	无标注样本的稳定性参数

给定包含 n 个样本的原始数据集 $S\in\mathcal{X}\times\mathcal{Y}$，其中正、负、无标注样本的数量分别为 n^+、n^- 和 $n^\sharp$。记$\mathcal{D}^n$为S的分布。令 i_+、i_- 和 $i_\sharp$ 分别为某一正、负、无标注样本的编号，X_i 为第 i

个样本的特征。假设存在数据集变换 $T:S\mapsto\{\widetilde{S}_1,\widetilde{S}_2\}\in\mathcal{X}^2\times\mathcal{Y}$，从原始数据集中创建了两个新的训练数据集

$$\widetilde{S}_1=\{(X_{i_+},X_{i_-},+1)\mid i_+,i_-\in[n]\}$$

和

$$\widetilde{S}_2=\{(X_{i_+},X_{i_\neq},+1)\mid i_+,i_\neq\in[n]\}$$

令 $\widetilde{X}_i$ 表示新数据集上的第 i 个样本对，其中 $\widetilde{S}_1$ 上的形式为 (X_{i_+},X_{i_-})，$\widetilde{S}_2$ 上的形式为 $(X_{i_+},X_{i_\neq})$。记 $\widetilde{Y}_i$ 为 $\widetilde{X}_i$ 的标签，且 $\widetilde{Z}_i=(\widetilde{X}_i,\widetilde{Y}_i)$。并且，令 $N_1=|\widetilde{S}_1|$，$N_2=|\widetilde{S}_2|$。容易看出，$N_1=n^+n^-$，$N_2=n^+n^\neq$。

给定学习算法 $\mathcal{A}:\{\widetilde{S}_1,\widetilde{S}_2\}\mapsto\mathcal{F}$，其目标在于从新的训练集 $\widetilde{S}$ 中学习一个泛化性好的假设 $f\in\mathcal{F}:\widetilde{X}\mapsto\widetilde{Y}$。通过损失函数 $\ell(f,\widetilde{Z})$ 度量 f 在样本 $\widetilde{Z}$ 上的预测质量，有

$$\widetilde{R}^{\alpha}_{PU}(f)=\mathbb{E}_S\left[\frac{1-\alpha}{n^+n^-}\sum_{\widetilde{Z}\in\widetilde{S}_1}\ell(f,\widetilde{Z})+\frac{\alpha}{n^+n^\neq}\sum_{\widetilde{Z}\in\widetilde{S}_2}\ell(f,\widetilde{Z})\right]。$$

由于 $\widetilde{S}_1$ 和 $\widetilde{S}_2$ 相互依赖，$\widetilde{R}^{\alpha}_{PU}(f)$ 并非独立变量之和。例如，两个训练样本 $\widetilde{X}_i=(X_{i_+},X_{i_-})$ 和 $\widetilde{X}_j=(X_{j_+},X_{j_-})$ 可能共享相同成分，如 $i_+=j_+$。

接下来，引入分析这种相互依赖数据所需的概念。在不失一般性的情况下，当上下文清晰时，将 $\widetilde{S}$ 表示为 $\widetilde{S}_1$ 或 $\widetilde{S}_2$、N 表示为 N_1 或 N_2。

考虑到训练集 $\widetilde{S}$ 是相互依赖的，可由此导出一个依赖图 $\mathcal{G}=(\mathcal{V},\boldsymbol{A})$，其中 $\mathcal{V}=\{1,2,\cdots,N\}$ 是 $\widetilde{S}$ 的样本索引集合，邻接矩阵 $\boldsymbol{A}$ 表示样本间的相互依赖关系。即，若样本 $\widetilde{Z}_i$ 和 $\widetilde{Z}_j$ 共享某一成分，则顶点 i 和 j 相邻，反之不相邻。给定依赖图后，有以下概念[61]。

定义 1. 分数覆盖(Fractional cover)。图的分数覆盖由顶点子集序列 $\{\mathcal{V}_1,\mathcal{V}_2,\cdots,\mathcal{V}_K\}$ 以及相应的权重序列 $\{\omega_1,\omega_2,\cdots,\omega_K\}$ 组成，其中：

(1) 对于每个 k，$\mathcal{V}_k$ 都是一组非相邻独立的顶点，满足 $\bigcup_{k=1}^{K}\mathcal{V}_k=\mathcal{V}$。

(2) $\forall i\in\mathcal{V},\sum_{k=1}^{K}\omega_k\mathbb{I}[i\in\mathcal{V}_k]\geq 1$，其中 $\omega_k>0$。

如果对于所有顶点，上述不等式均变成等式，则该分数覆盖称为适当的(proper)。适当分数覆盖的存在性由文献[63]保证。在本文的剩余部分，将仅考虑适当分数覆盖。

定义 2. 分数染色数(Fractional chromatic number)。图 $\mathcal{G}$ 的分数染色数 $\chi(\mathcal{G})$ 定义为 $\sum_{k=1}^{K}\omega_k$ 的最小值。

由于图 $\mathcal{G}$ 与数据集 $\widetilde{S}$ 相关，也可将分数染色数 $\chi(\mathcal{G})$ 表示为 $\chi(\widetilde{S})$。

若有适当分数覆盖 $\{(\omega_k,\mathcal{V}_k)\}_{k=1}^{K}$ 以及定义在顶点上的随机变量 Z_i，则有如下 Janson 分解成立：

$$\sum_{i=1}^{N}Z_i=\sum_{i=1}^{N}\sum_{k=1}^{K}\omega_k\Psi[i\in\mathcal{V}_k]Z_i=\sum_{k=1}^{K}\omega_k\sum_{i\in\mathcal{V}_k}Z_i \tag{27}$$

上式表明，相互依赖变量之和可被重构为独立变量之和的加权和。注意，由于 $\mathcal{V}_k$ 仅包含非相邻顶点，其内部求和是在独立变量之间进行的，标准分析可以适用。因此，据此可将许多传统的独立数据集上的分析工具扩展到相互依赖的数据集上(更多细节参见文献[60])。该分解将贯穿于后文的理论分析中。此外，记 $|\mathcal{V}_k|$ 为 $\mathcal{V}_k$ 的样本数量，根据该分解，有

$$\sum_{k=1}^{K}\omega_k|\mathcal{V}_k|=N。$$

接下来，引入分数 Rademacher 复杂度，它将标准的 Rademacher 复杂度扩展到相互依赖的数据集上。

定义 3. 分数 Rademacher 复杂度。令 $\{(\omega_k,\mathcal{V}_k)\}_{k=1}^{K}$ 为数据集 $\widetilde{S}$ 的适当分数覆盖。分数 Rademacher 复杂度定义如下：

$$\mathfrak{R}_n^{\widetilde{S}}(\mathcal{F})=\sum_{k=1}^{K}\frac{\omega_k}{N}\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i\in\mathcal{V}_k}\sigma_i f(\widetilde{X}_i)\right] \tag{28}$$

其中，$\boldsymbol{\sigma}\in\{-1,+1\}^N$ 为满足 $\mathbb{P}(\sigma_i=-1)=\mathbb{P}(\sigma_i=1)=1/2$ 的独立 Rademacher 变量。

因此，经验分数 Rademacher 复杂度的定义如下：

$$\mathfrak{R}_n^{\widetilde{S}}(F,S)=\sum_{k=1}^{K}\frac{\omega_k}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i\in\mathcal{V}_k}\sigma_i f(\widetilde{X}_i)\right] \tag{29}$$

关于 $\chi(\widetilde{S})$ 的具体值，根据文献[60]，有

$$(\widetilde{S}_1)=\max(n^+,n^-) \tag{30}$$

类似地，

$$|\chi(\widetilde{S}_2)=\max(n^+,n^\neq) \tag{31}$$

基于上述概念，根据之前的研究[10]，可以得到 $\widetilde{R}^{\alpha}_{PU}(f)$ 的上界由其经验误差加上复杂度与置信度项组成。

定理 3. 假设损失函数 ℓ 的上界为 $B\ell$。则对任意 $\delta>0$，以下不等式对任意 $f\in\mathcal{F}$ 以至少 $1-\delta$ 的概率成立：

$$\begin{aligned}&R^{\alpha}_{PU}(f)-R^{\alpha}_{PU}(f,S)\\&\leq\underbrace{(1-\alpha)\mathfrak{R}_n^{\widetilde{S}_1}(\ell\circ\mathcal{F})+\alpha\mathfrak{R}_n^{\widetilde{S}_2}(\ell\circ\mathcal{F})}_{\text{复杂度项}}+\\&\underbrace{(1-\alpha)B_\ell\sqrt{\frac{\chi(\widetilde{S}_1)\log 2/\delta}{2N_1}}+\alpha B_\ell\sqrt{\frac{\chi(\widetilde{S}_2)\log 2/\delta}{2N_2}}}_{\text{置信度项}}\end{aligned} \tag{32}$$

注意到复杂度项和置信度项的阶均为

$$O((1-\alpha)/\sqrt{\min(n^+,n^-)}+\alpha/\sqrt{\min(n^+,n^\neq)}),$$

接下来需要进一步使得这两项更紧。

附录 C. 构建局部假设类。

在定理 3 中，对于任意 $f\in\mathcal{F}$ 均存在一致上界。因此，$\mathcal{F}$ 中的最坏情况假设限制了对于更紧上界的计算。为解决该问题，需要找到一个局部假设类 $\mathcal{B}_r$，使得“期望假设”以较高概率居于其中。此处的期望假设是由随机学习算法 $\mathcal{A}$ 输出的假设(即模型)的期望，记为 $\mathbb{E}_{\mathcal{A}}f_S$。其随机性可能来自模型参数的初始化，或诸如 SGD 等算法的小批次采样操作。为了找到包含期望假设的局部假设类 $\mathcal{B}_r$，需要利用学习算法的一致参数稳定性[54]。接下来详细阐述相应的概念。

通常情况下，学习算法的稳定性反映了训练数据微小变化对学习算法输出的影响。设 f_S 为学习算法在数据集 S 上的输出。本文仅考虑线性假设类 $\mathcal{F}$，即 $f(\widetilde{X}_i)$ 可建模为一个线性函数 $\boldsymbol{w}^{\mathrm{T}}\widetilde{X}_i$，其中 $\boldsymbol{w}$ 为模型参数。因而可以将其表示为

$$f(\widetilde{X}_i)=\langle f,\widetilde{X}_i\rangle=\boldsymbol{w}^{\mathrm{T}}\widetilde{X}_i,\ \forall f\in\mathcal{F}。$$

在半监督 AUC 优化问题中，一致参数稳定性定义如下。

定义 4. 一致参数稳定性(Uniform Argument Stability)。

记 $S^i=\{Z_1,Z_2,\cdots,Z_{i-1},Z_i',Z_{i+1},\cdots,Z_n\}$ 为与 S 仅在第 i 个样本处不同的数据集。若存在 $\alpha_+(n)$、$\alpha_-(n)$ 和 $\alpha_\sharp(n)$ 使得任意 $S\sim\mathcal{D}^n$、$f\in\mathcal{F}$、正样本 i_+、负样本 i_- 和无标注样本 $i_\sharp$ 均满足以下不等式，则学习算法 $\mathcal{A}$ 是一致参数稳定的：

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_+}}\|]\leq\alpha_+(n),\\&\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_-}}\|]\leq\alpha_-(n),\\&\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_\sharp}}\|]\leq\alpha_\sharp(n)\end{aligned}\tag{33}$$

其中，$\alpha_+(n),\alpha_-(n),\alpha_\sharp(n)>0$，且

$$\|f_S-f_{S^i}\|=\sup_{\widetilde{X}\in\widetilde{\mathcal{X}},\|\widetilde{X}\|_2\leq1}|\langle f_S,\widetilde{X}\rangle-\langle f_{S^i},\widetilde{X}\rangle|\tag{34}$$

一致参数稳定性表明，训练数据上较小变化仅会导致所学习到的分类器 f_S 的微小变化。实际上，这样的算法将输出一个集中在其期望值 $\mathbb{E}_{\mathcal{A},S}f_S$ 附近的期望假设 $\mathbb{E}_{\mathcal{A}}f_S$，如以下引理所示。

引理 3.　额外假设特征空间中的任意特征 $\widetilde{X}$ 均满足 $\|\widetilde{X}\|_2\leq B_f$。若学习算法 $\mathcal{A}$ 一致参数稳定，且上界为 $\alpha_+(n)$、$\alpha_-(n)$ 和 $\alpha_\sharp(n)$，则对任意置信度参数 $\gamma>0$，有

$$\mathbb{P}(\|\mathbb{E}_{\mathcal{A}}f_S-\mathbb{E}_{\mathcal{A},S}f_S\|\leq\beta(n)\sqrt{2B_f\log(2/\gamma)})\geq1-\gamma\tag{35}$$

其中 $\beta(n)=\sqrt{n^+\alpha_+^2(n)+n^-\alpha_-^2(n)+n^\sharp\alpha_\sharp^2(n)}$。

证明.　引入鞅差序列 $D_t=\mathbb{E}(\mathbb{E}_{\mathcal{A}}f_S\,|\,Z_1,Z_2,\cdots,Z_t)-\mathbb{E}(\mathbb{E}_{\mathcal{A}}f_S\,|\,Z_1,Z_2,\cdots,Z_{t-1})$，可以得到以下关系：

$$\mathbb{E}_{\mathcal{A}}f_S-\mathbb{E}_{\mathcal{A},S}f_S=\sum_{t=1}^{n}D_t,$$

则有

$$\begin{aligned}&\sum_{t=1}^{n}\|D_t\|_\infty^2\\&=\sum_{t=1}^{n}\|\mathbb{E}(\mathbb{E}_{\mathcal{A}}f_S\,|\,Z_1,Z_2,\cdots,Z_t)-\mathbb{E}(\mathbb{E}_{\mathcal{A}}f_S\,|\,Z_1,Z_2,\cdots,Z_{t-1})\|_\infty^2\\&=\sum_{t=1}^{n}\|\mathbb{E}(\mathbb{E}_{\mathcal{A}}f_S-\mathbb{E}_{\mathcal{A}}f_{S^t}\,|\,Z_1,Z_2,\cdots,Z_t)\|_\infty^2\\&\leq\sum_{t=1}^{n}(\mathbb{E}(\|\mathbb{E}_{\mathcal{A}}f_S-\mathbb{E}_{\mathcal{A}}f_{S^t}\|_\infty\,|\,Z_1,Z_2,\cdots,Z_t))^2\\&\leq\sum_{t=1}^{n}(\mathbb{E}(\mathbb{E}_{\mathcal{A}}\|f_S-f_{S^t}\|_\infty\,|\,Z_1,Z_2,\cdots,Z_t))^2\\&\leq B_f(n^+\alpha_+^2(n)+n^-\alpha_-^2(n)+n^\sharp\alpha_\sharp^2(n))\end{aligned}\tag{36}$$

令 $n^+\alpha_+^2(n)+n^-\alpha_-^2(n)+n^\sharp\alpha_\sharp^2(n)=\beta^2(n)$，可得

$$\beta(n)\triangleq\sqrt{n^+\alpha_+^2(n)+n^-\alpha_-^2(n)+n^\sharp\alpha_\sharp^2(n)}\tag{37}$$

因此，根据文献[54]的命题 1，有

$$\mathbb{P}(\|\mathbb{E}_{\mathcal{A}}f_S-\mathbb{E}_{\mathcal{A},S}f_S\|\leq\beta(n)\sqrt{2B_f\log(2/\gamma)})\geq1-\gamma。$$

证毕.

引理 3 表明 $\mathbb{E}_{\mathcal{A}}f_S$ 以高概率集中在其期望值 $\mathbb{E}_{\mathcal{A},S}f_S$ 附近。基于此，可构建如下依赖于算法的局部假设类。

定义 5.　算法假设(Algorithmic Hypothesis)。使用置信度参数 $\gamma>0$，记 $r(n;\gamma)=\beta(n)\sqrt{2B_f\log(2/\gamma)}$。对于一致参数稳定的学习算法，其算法假设定义为

$$\mathcal{B}_r=\{f\in\mathcal{F}\,|\,\|f-\mathbb{E}_{\mathcal{A},S}f_S\|\leq r(n;\gamma)\}\tag{38}$$

值得注意的是，根据引理 3，$\mathbb{E}_{\mathcal{A}}f_S\in\mathcal{B}_r$ 至少以 $1-\gamma$ 的概率成立。$\mathcal{F}$、$\mathcal{B}_r$、最坏情况假设和期望假设 $\mathbb{E}_{\mathcal{A}}f_S$ 间的关系如图 6 所示。基于此，只需关注 $\mathcal{B}_r$ 而非整个假设类 $\mathcal{F}$。$\mathcal{B}_r$ 的分数 Rademacher 复杂度定义如下。

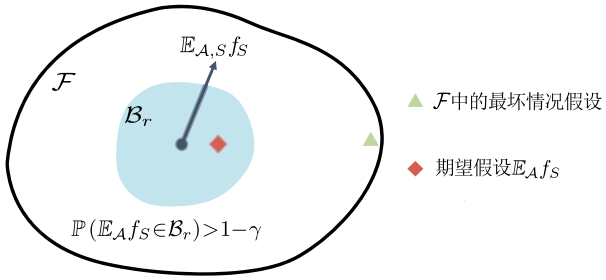


图 6　假设类 $\mathcal{F}$ 与集中假设类(Concentrated hypothesis class) $\mathcal{B}_r$ 的关系以及最坏情况假设与期望假设的区别

定义 6.　算法假设的分数 Rademacher 复杂度。令 $\{(\omega_k,\mathcal{V}_k)\}_{k=1}^K$ 为数据集 $\widetilde{S}$ 的适当分数覆盖。假设学习算法是一致参数稳定的，其算法假设记为 $\mathcal{B}_r$。则 $\mathcal{B}_r$ 的分数 Rademacher 复杂度定义为

$$\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r)=\sum_{k=1}^{K}\frac{\omega_k}{N}\mathbb{E}_S\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f\in\mathcal{B}_r}\sum_{i\in\mathcal{V}_k}\sigma_i f(\widetilde{X}_i)\Big]\tag{39}$$

最后，给出 $\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r)$ 的一个上界，作为后续结果的基础。

定理 4.　令 $\{(\omega_k,\mathcal{V}_k)\}_{k=1}^K$ 为数据集 $\widetilde{S}$ 的适当分数覆盖。假设模型为线性，其特征空间为满足 $\|\widetilde{X}_i\|_2\leq B_f$ 的希尔伯特空间，其中 $B_f>0$ 为常数，且学习算法是一致参数稳定的。记 $\mathcal{B}_r$ 为算法假设，则其分数 Rademacher 复杂度满足

$$\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r)\leq B_f\beta(n)\sqrt{2B_f\chi(\widetilde{S}_1)\log(2/\gamma)}\tag{40}$$

证明.　根据定义 X 和式(27)中的 Janson 分解，有

$$\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r)=\sum_{k=1}^{K}\frac{\omega_k|\mathcal{V}_k|}{N}\mathbb{E}_S\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\sigma_i\langle f,\widetilde{X}_i\rangle\Big]\tag{41}$$

定义

$$\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r,\mathcal{V}_k)\triangleq\mathbb{E}_S\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\sigma_i\langle f,\widetilde{X}_i\rangle\Big],$$

有

$$\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r)=\sum_{k=1}^{K}\frac{\omega_k|\mathcal{V}_k|}{N}\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r,\mathcal{V}_k)。$$

然后，为了求 $\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r)$ 的上界，仅需计算 $\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r,\mathcal{V}_k)$。类似于文献[54]中定理 1 的证明，有

$$\begin{aligned}&\mathfrak{R}_n^{\widetilde{S}}(\mathcal{B}_r,\mathcal{V}_k)\\&=\mathbb{E}\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\sigma_i f(\widetilde{X}_i)\\&=\mathbb{E}\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}(\sigma_i\langle f,\widetilde{X}_i\rangle-\sigma_i\langle\mathbb{E}_{\mathcal{A},S}f_S,\widetilde{X}_i\rangle+\\&\quad\sigma_i\langle\mathbb{E}_{\mathcal{A},S}f_S,\widetilde{X}_i\rangle)\\&\leq\mathbb{E}\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\sigma_i(\langle f,\widetilde{X}_i\rangle-\langle\mathbb{E}_{\mathcal{A},S}f_S,\widetilde{X}_i\rangle)\\&=\mathbb{E}\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\sigma_i\langle f-\mathbb{E}_{\mathcal{A},S}f_S,\widetilde{X}_i\rangle\\&\leq\mathbb{E}\sup_{f\in\mathcal{B}_r}\frac{1}{|\mathcal{V}_k|}\|f-\mathbb{E}_{\mathcal{A},S}f_S\|\cdot\Big\|\sum_{i\in\mathcal{V}_k}\sigma_i\widetilde{X}_i\Big\|_2\\&\leq\frac{r(n;\gamma)}{|\mathcal{V}_k|}\mathbb{E}\Big\|\sum_{i\in\mathcal{V}_k}\sigma_i\widetilde{X}_i\Big\|_2\end{aligned}$$

$$\overset{(c)}{\leq}\frac{r(n;\gamma)}{|\mathcal{V}_k|}\Big(\sum_{i\in\mathcal{V}_k}\|\widetilde{X}_i\|_2^2\Big)^{1/2}$$

$$\leq B_f\beta(n)\sqrt{2B_f\log(2/\gamma)}\Big(\frac{1}{|\mathcal{V}_k|}\Big)^{1/2} \tag{42}$$

其中，由于希尔伯特空间是带有 ℓ_2 范数的巴拿赫空间，式(42)中不等号(c)成立(详见文献[54])。将式(42)代回式(41)，可得

$$\mathfrak{R}_n^{\mathcal{S}}(\mathcal{B}_r)\leq B_f\beta(n)\sqrt{2B_f\log(2/\gamma)}\sum_{k=1}^{K}\frac{\omega_k|\mathcal{V}_k|}{N}\Big(\frac{1}{|\mathcal{V}_k|}\Big)^{1/2}$$

$$\overset{(d)}{\leq}B_f\beta(n)\sqrt{2B_f\log(2/\gamma)}\left(\frac{\sum_{k=1}^{K}\omega_k}{N}\right)^{1/2}$$

$$=B_f\beta(n)\sqrt{\frac{2B_f\chi(\widetilde{S}_1)\log(2/\gamma)}{N}}$$

其中，根据 Jensen 不等式，上式中的不等号(d)成立。 证毕.

附录 D. 变形额外风险的一致上界(定理1的证明)。

结合算法假设的分数 Rademacher 复杂度以及变形额外风险，可以获得更紧的一致上界。

根据正文第 5.1 节，式(10)右侧两个上确界项的上界证明思路相同，在此仅给出第一个上确界项的计算方式。定理 1 的正式版本如下。

定理 1. 正式版。假设损失函数 L-Lipschitz 连续。输入特征的范数$\|\widetilde{X}_i\|_2$不大于B_f。若学习算法$\mathcal{A}$是一致参数稳定的，令 $a>1$，则对任意 $\delta>0$，以下不等式以至少 $1-2\delta$ 的概率成立：

$$\widetilde{R}_{+-}(f_S)-\frac{a}{a-1}\widetilde{R}_{+-}(f_S,S)\leq\frac{(6a+8)\chi(\widetilde{S}_1)B_\ell\log(1/\delta)}{3N_1}+8LB_f\beta(n)\sqrt{\frac{2B_f\chi(\widetilde{S}_1)\log(2/\delta)}{N_1}}。$$

证明. 通过式(27)的分解，有

$$\widetilde{R}_{+-}(f_S)-\frac{a}{a-1}\widetilde{R}_{+-}(f_S,S)$$

$$=\mathbb{E}_S\left[\frac{1}{N_1}\sum_{k=1}^{K}\omega_k\sum_{i\in\mathcal{V}_k}\ell(f_S,\widetilde{Z}_i)\right]-\frac{a}{a-1}\frac{1}{N_1}\sum_{k=1}^{K}\omega_k\sum_{i\in\mathcal{V}_k}\ell(f_S,\widetilde{Z}_i)$$

$$=\sum_{k=1}^{K}\frac{\omega_k|\mathcal{V}_k|}{N_1}\Big(\mathbb{E}_S\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\ell(f_S,\widetilde{Z}_i)-\frac{a}{a-1}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\ell(f_S,\widetilde{Z}_i)\Big)。$$

类似文献[54]，以下不等式至少以 $1-2\delta$ 的概率成立：

$$\sup_{f\in\mathcal{B}_r}\Big(\mathbb{E}_S\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\ell(f,\widetilde{Z}_i)-\frac{a}{a-1}\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\ell(f,\widetilde{Z}_i)\Big)$$

$$\leq\frac{(6a+8)B_\ell\log(1/\delta)}{3|\mathcal{V}_k|}+8\mathfrak{R}_n^{\mathcal{S}}(\ell\circ\mathcal{B}_r,\mathcal{V}_k)。$$

因而，有

$$\widetilde{R}_{+-}(f_S)-\frac{a}{a-1}\widetilde{R}_{+-}(f_S,S)$$

$$\leq\sum_{k=1}^{K}\frac{\omega_k|\mathcal{V}_k|}{N_1}\Big(\frac{(6a+8)B_\ell\log(1/\delta)}{3|\mathcal{V}_k|}+8\mathfrak{R}_n^{\mathcal{S}}(\ell\circ\mathcal{B}_r,\mathcal{V}_k)\Big)$$

$$=\frac{(6a+8)\chi(\widetilde{S}_1)B_\ell\log(1/\delta)}{3N_1}+8\mathfrak{R}_n^{\mathcal{S}}(\ell\circ\mathcal{B}_r) \tag{43}$$

假设损失函数 L-Lipschitz 连续。根据 Talagrand 压缩引理[41]，算法假设 $\ell\circ\mathcal{B}_r$的分数 Rademacher 复杂度为

$$\mathfrak{R}_n^{\mathcal{S}}(\ell\circ\mathcal{B}_r)\leq L\cdot\mathfrak{R}_n^{\mathcal{S}}(\mathcal{B}_r)\leq LB_f\beta(n)\sqrt{\frac{2B_f\chi(\widetilde{S}_1)\log(2/\gamma)}{N_1}}。$$

最后，将$\mathfrak{R}_n^{\mathcal{S}}(\ell\circ\mathcal{B}_r)$代回式(43)即可证明该定理。 证毕.

附录 E. 随机梯度方法上的应用(定理 2 的证明)。

本节将上节中给出的结论应用于随机梯度下降(SGD)算法。首先证明 SGD 是一致参数稳定的，并同时给出 $\beta(n)$ 的具体上界。然后可以得到SGD算法的变形额外风险上界。

为便于分析，采用常用的 s-平滑假设。

定义 7. s-平滑(s-smoothness)。若可微损失函数$\ell(f,\cdot)$对所有$f\in\mathcal{F}$均满足以下条件，则称其为 s-平滑的：

$$\|\nabla_f\ell(f,\cdot)-\nabla_{f'}\ell(f',\cdot)\|\leq s\|f-f'\| \tag{44}$$

接下来给出 SGD 算法的 $\alpha_+(n)$、$\alpha_-(n)$和 $\alpha_\sharp(n)$上界。

引理 4. 假设损失函数 $\ell(f,\cdot)$是 s-平滑的、凸的且 L-Lipschitz 连续的。若 SGD 算法运行了 T 步，其学习率满足 $\eta_t\leq 2/s$，则 SGD 满足一致参数稳定性，且有

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_+}}\|]\leq\alpha_+(n)\leq\frac{2LB_f}{n^+}\sum_{t=1}^{T}\eta_t,$$

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_-}}\|]\leq\alpha_-(n)\leq\frac{2LB_f}{n^-}\sum_{t=1}^{T}\eta_t,$$

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_\sharp}}\|]\leq\alpha_\sharp(n)\leq\frac{2LB_f}{n^\sharp}\sum_{t=1}^{T}\eta_t。$$

证明. 记 $G_1,\cdots,G_T$和$G'_1,\cdots,G'_T$分别为在数据集 S 和 S^{i_+}上运行 T 步 SGD 所产生的梯度更新，$\boldsymbol{\theta}_T$ 和 $\boldsymbol{\theta}'_T$分别为对应输出模型的参数。根据$f(\cdot;\widetilde{Z})$的 Lipschitz 条件，有

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_+}}\|]\leq\mathbb{E}_{\mathcal{A}}[\delta_T] \tag{45}$$

其中 $\delta_T=\|\boldsymbol{\theta}_T-\boldsymbol{\theta}'_T\|_2$。注意到在第 t 步，S 和 S^{i_+} 所选样本有

$$\frac{n^-n^\sharp}{n^+n^-n^\sharp}=\frac{1}{n^+}$$

的概率不同，而通过 SGD 选择相同样本的概率为 $1-1/n^+$。依照文献[62]的证明思路，可以得出对第 t 步有以下结论成立：

$$\mathbb{E}_{\mathcal{A}}[\delta_{t+1}]\leq\frac{1}{n^+}(\mathbb{E}_{\mathcal{A}}[\delta_t]+2\eta_tLB_f)+\Big(1-\frac{1}{n^+}\Big)\mathbb{E}_{\mathcal{A}}[\delta_t] \tag{46}$$

递归求和所有 T 步，有

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_+}}\|]\leq\mathbb{E}_{\mathcal{A}}[\delta_T]\leq\frac{2LB_f}{n^+}\sum_{t=1}^{T}\eta_t \tag{47}$$

类似地，对 S 和 S^{i_-}，有

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^i}\|]\leq\frac{2LB_f}{n^-}\sum_{t=1}^{T}\eta_t \tag{48}$$

对 S 和 $S^{i_\sharp}$，有

$$\mathbb{E}_{\mathcal{A}}[\|f_S-f_{S^{i_\sharp}}\|]\leq\frac{2LB_f}{n^\sharp}\sum_{t=1}^{T}\eta_t \tag{49}$$

证毕.

根据 $\alpha_+(n)$、$\alpha_-(n)$和 $\alpha_\sharp(n)$的上界以及式(37)中 $\beta(n)$的定义，可以得到 $\beta(n)$的上界如下：

$$\beta(n)\leq 2LB_f\sqrt{\frac{1}{n^+}+\frac{1}{n^-}+\frac{1}{n^\sharp}}。$$

结合上式和定理 1，可以得到定理 2。

JIANG Yang-Bang-Yan, Ph. D., lecturer. Her research interests include machine learning and computer vision.

XU Qian-Qian, Ph. D., professor. Her research interests include statistical machine learning and its applications in multimedia.

YANG Zhi-Yong, Ph. D., associate professor. His research interests include theoretical and algorithmic aspects of machine learning.

HAO Qian-Xiu, M. S., engineer. Her research interests include machine learning and multimedia analysis.

CAO Xiao-Chun, Ph. D., professor. His research interests include computer vision and AI security.

HUANG Qing-Ming, Ph. D., chair professor. His research interests include multimedia computing, computer vision and pattern recognition.

Background

The Area Under ROC Curve (AUC) is one of the most popular evaluation metrics for classification performance which is insensitive to label distributions and misclassification costs. Seeing the inherent advantages of AUC in dealing with imbalanced data, there is a large amount of work trying to directly optimize AUC, especially under incomplete supervision where only a limited number of data are labeled.

Existing semi-supervised AUC optimization methods usually assume that the labeled data is accurate. However, in many applications, we must simultaneously face insufficiency of data and inaccuracy of annotations.

Motivated by this fact, we present the first trial on optimizing AUC under the context of incomplete and inaccurate data annotations. More specifically, we show that the symmetric surrogate losses are noise-robust in the semi-supervised setting under certain scenarios. On top of this result, we construct a robust semi-supervised AUC optimization framework along with the induced empirical risks does not need to estimate the noise rates. Taking a step further, a tight bound of the excess risk is presented to show that a model learned from a sufficiently large given training dataset could generalize well to further unseen data. Practically, we propose an instantiation of our framework with the Barrier hinge loss. To speed up the training process, an acceleration algorithm is developed to reduce the complexity of loss and gradient evaluation from $O(n^2)$ to $O(n\log n)$, which leads to up to 200x speedup in the experiments.

This work was supported in part by the National Key R&D Program of China under Grant No. 2018AAA0102000, in part by the National Natural Science Foundation of China: Nos. 62236008, U21B2038, U23B2051, 61931008, 62122075, 62406305, 62476068, 62471013, 62206264, and 92370102, in part by the Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680000, in part by the Innovation Funding of ICT, CAS under Grant No. E000000, in part by the China Postdoctoral Science Foundation (CPSF) under Grant No. 2023M743441, and in part by the Postdoctoral Fellowship Program of CPSF under Grant No. GZB20230732.