面向 AUC 优化的高效对抗训练

包世龙[1] 许倩倩[2] 杨智勇[1] 华 聪[1),2)] 韩博宇[1),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 鉴于 ROC 曲线下面积(Area Under the ROC Curve,AUC)对数据分布的不敏感特性,面向 AUC 的对抗训练(AdAUC)近来已成为机器学习领域中抵御长尾分布下对抗攻击的有效范式之一。当前主流方法大多遵循基于平方替代损失的 AUC 对抗训练框架,并将成对比较形式的 AUC 对抗损失重构为一个逐样本的随机鞍点优化问题,克服端到端的计算瓶颈。然而,面向复杂的实际应用场景,基于平方损失设计的 AUC 对抗训练框架恐难以适应多样的下游任务需求。此外,与传统对抗训练范式类似,面向 AUC 的对抗训练方法在提高模型对抗鲁棒性的同时,也会降低模型在正常样本上的 AUC 性能,而目前鲜有针对该问题的有效解决方案。鉴于此,本文对如何构建一般化的高效 AUC 对抗机器学习范式展开系统研究。首先,提出了一种基于标准化分数扰动的通用 AUC 对抗训练框架 (NSAdAUC),在相对温和的条件下,该框架可通过直接扰动模型对样本的预测得分实现对 AUC 指标的攻击,且不依赖于特定的 AUC 替代损失。在此基础上,本文进一步指出鲁棒 AUC 误差可分解为标准 AUC 误差和边界 AUC 误差两项之和,并据此设计了一种基于排序感知对抗正则化的 AUC 对抗训练框架 (RARAdAUC),同时兼顾模型的标准 AUC 和鲁棒 AUC 性能。为验证所提框架的有效性,在 5 个长尾基准数据集上进行了大量实验,结果表明所提 NSAdAUC 和 RARAdAUC 框架在多种对抗攻击下的鲁棒性均优于现有方法,可在平均意义上分别产生 0.94%、5.52%的标准 AUC 和 5.69%、5.41%的鲁棒 AUC 性能提升。

关键词 AUC 优化;对抗训练;对抗鲁棒性;长尾学习;机器学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2025.01551

Efficient Adversarial Training for AUC Optimization

BAO Shi-Long[1] XU Qian-Qian[2] YANG Zhi-Yong[1] HUA Cong[1),2)]

HAN Bo-Yu[1),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, Shenzhen Campus, Sun Yat-sen University, Shenzhen, Guangdong 518107)

Abstract The Area Under the ROC Curve (AUC) is widely recognized as an essential metric for evaluating classification performance, particularly in imbalanced data scenarios, due to its insensitivity to underlying data distribution. Motivated by its promising property, AUC-Oriented Adversarial Training (AT), abbreviated as AdAUC, has recently gained prominence as an effective paradigm for defending against adversarial attacks in real-world long-tail security challenges.

收稿日期:2024-09-18;在线发布日期:2025-03-27。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(62236008,62441232,U21B2038,U23B2051,62122075,62206264,92370102)、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(XDB06801201)、国家资助博士后研究人员计划(GZB20240729)资助。**包世龙**,博士,助理教授,中国计算机学会(CCF)会员,主要研究领域为机器学习与数据挖掘。E-mail:baoshilong@ucas.ac.cn。**许倩倩**(通信作者),博士,研究员,中国计算机学会(CCF)会员,主要研究领域为统计机器学习、多媒体与计算机视觉。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。**华 聪**,博士研究生,主要研究方向为机器学习与多模态学习。**韩博宇**,博士研究生,主要研究方向为计算机视觉与生成模型。**操晓春**,博士,教授,国家杰出青年科学基金入选者,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,IEEE Fellow,国家杰出青年科学基金入选者,中国计算机学会(CCF)会士,主要研究领域为多媒体计算、图像处理、计算机视觉与模式识别。E-mail:qmhuang@ucas.ac.cn。

The core idea of AdAUC is to improve model robustness against adversarial perturbations by optimizing a minimax framework with AUC-inspired AT objectives. To achieve this, existing AdAUC methods typically rely on a squared surrogate loss to approximate and reformulate the pairwise AUC adversarial loss into an instance-wise stochastic saddle point problem (SPP). This transformation alleviates the computational bottlenecks arising from pairwise comparisons in AdAUC. However, despite its advantages, this approach has several limitations. First of all, given that different surrogate optimization methods often lead to varying AUC performances, the current square-base surrogate AdAUC paradigm may lack the flexibility needed to accommodate the diverse robustness requirements of real-world applications. In addition, akin to the traditional AT paradigm, improving adversarial robustness in terms of AUC typically comes at the cost of degraded AUC performance on clean data — An issue commonly referred to as the clean-robustness trade-off in the AT community. Unfortunately, this trade-off between standard AUC and robust AUC remains an open challenge in the current literature, with limited exploration of effective solutions, thereby restricting the practical applicability of AdAUC. To address these issues, this paper systematically investigates a more generalized and efficient AdAUC framework. Specifically, we introduce a novel approach called the Normalized Score-based AdAUC (NSAdAUC), which takes a fundamentally different approach from existing methods. Instead of relying on specific surrogate loss functions, NSAdAUC directly perturbs the model's predicted scores across different samples to optimize adversarial AUC. This direct perturbation strategy allows for a more flexible and effective AT process, free from the constraints of traditional surrogate losses. Taking a step further, we provide a theoretical analysis that decomposes the robust AUC error into two key components: the standard AUC error and the boundary AUC error. This decomposition offers deeper insights into the fundamental trade-offs of AdAUC and serves as a guiding principle for designing more balanced training strategies. Building upon these insights, we propose a Ranking-aware Adversarial Regularization algorithm (RARAdAUC), explicitly designed to balance standard and robust AUC performance. More concretely, RARAdAUC introduces a ranking-based regularization term to mitigate the negative impact of AdAUC on clean data while still enhancing adversarial robustness. Finally, to evaluate the effectiveness of our proposed methods, we conduct extensive experiments on five benchmark datasets with long-tail distributions. The experimental results demonstrate that NSAdAUC and RARAdAUC consistently outperform existing AdAUC approaches. In particular, NSAdAUC achieves an average improvement of 0.94% in standard AUC and 5.69% in robust AUC, while RARAdAUC yields even greater improvements of 5.52% in standard AUC and 5.41% in robust AUC. Our study not only provides new insights into the adversarial robustness of AdAUC but also paves the way for future research into balancing standard and robust performance in adversarial settings.

Keywords AUC optimization; adversarial training; adversarial robustness; long-tail learning; machine learning

1 引 言

ROC 曲线,即受试者操作特征曲线(Receiver Operating Characteristic curve),是机器学习领域一个常用的性能评估曲线,它表示了不同分类阈值下模型预测结果的真阳性率(True Positive Rate,TPR)与假阳性率(False Positive Rate,FPR)之间的关系。ROC 曲线下面积(Area Under the ROC Curve,AUC)度量了在不同阈值下分类器的平均性能。考虑到现实世界中的数据常呈现出长尾分布态势,即仅少部分类别(头部类)在数量上主导地位,其他类

别(尾部类)数据体量相对匮乏，主流的准确率(Accuracy，简称 ACC)等性能评估指标无法准确反映模型对尾部类的预测性能[1-3]。相比之下，AUC 对类别分布不敏感[4]，可较好地聚焦尾部类样本的性能，故更适合作为长尾数据分布场景下的模型性能评估指标。有鉴于此，直接通过优化 AUC 指标构造模型的理念受到了学术界及工业界的广泛关注，已被广泛应用于多种类别分布不平衡场景[5]，如疾病预测[6-7]、异常检测[8]和金融欺诈检测[9]等。

与此同时，随着近年来深度学习、机器学习等人工智能技术的安全隐患逐渐暴露出来—攻击者通过对模型输入施加轻微的、人眼无法察觉的微小扰动便可轻易改变模型的原始预测结果[10]，如何利用 AUC 优化技术提高长尾分布场景下模型的鲁棒性已成为机器学习领域的前沿热点问题之一。目前，对抗训练(Adversarial Training，AT)是提升模型鲁棒性的一个主流范式，其可表述为一个最小最大化(min-max)问题，其内部最大化问题旨在生成能够改变模型预测的对抗样本，随后外部最小化问题基于生成的对抗样本进行模型学习以规避对抗攻击。然而，AUC 优化的成对比较(pairwise)损失函数定义复杂，其中每个正(负)样本和所有负(正)样本相互耦合，因此若以 AUC 指标为攻击目标需同时篡改一对正负样本实现对抗样本生成，难以直接沿用传统 AT 范式中针对逐样本(pointwise)损失设计的对抗学习框架。

为克服该问题，文献[11]首次探索了面向 AUC 指标的对抗训练机制。针对 AUC 损失中正负样本耦合的问题，文献[11]选取平方损失[12]作为 AUC 优化的替代损失函数，并将成对比较形式的 AUC 对抗训练问题重构为一个逐样本的随机鞍点优化问题，实现了正负样本对的有效解耦。针对文献[11]的对抗攻击假设强、对抗样本生成弱以及泛化性能不明确的问题，文献[13]重新审视了基于 AUC 优化的对抗训练方法，提出了一种基于随机方差缩减梯度(Stochastic Variance Reduced Gradient，SVRG)的端到端 AUC 对抗训练算法，并给出了面向 AUC 对抗训练范式的泛化理论分析框架。

尽管面向 AUC 的对抗训练方法取得了一定的进展，但仍存在以下几个问题亟需解决：首先，主流方法仅考虑基于平方替代损失的 AUC 对抗训练问题，无法适用于其他典型的 AUC 替代损失，如指数损失、铰链损失等，可扩展性较差。其次，现有方法可能面临模型优化目标与实际攻防场景不一致的问题。简言之，主流方法通常假设恶意攻击者也将严格遵循正负样本比较对形式(pairwise)生成对抗样本实现对 AUC 指标的攻击，而在实际对抗攻防场景中对抗攻击大多只会应用于模型的单个输入数据，例如以交叉熵损失为基础(针对 ACC 的)进行跨指标对抗攻击，带来额外的安全隐患。此外，以往的研究表明[14]，对抗训练在提高模型对抗鲁棒性同时将不可避免的降低模型在干净样本上的性能，而现有 AUC 的对抗训练研究鲜有对标准 AUC(standard AUC)和鲁棒 AUC(robust AUC)性能间权衡问题的探究。

鉴于此，为实现高效的 AUC 对抗训练，本文首先重新审视了针对 AUC 指标的对抗攻击问题，通过对不同替代损失的单调性分析发现—为降低模型的 AUC 指标，同时操纵一对正负样本对与分别扰动模型对正负样本的预测得分所生成的对抗样本具有相同的全局最优解，其中对于正样本应降低其预测得分而对于负样本应尽可能提高模型的预测得分。基于该良好性质，本文提出基于标准化分数扰动的通用 AUC 对抗训练框架(记为 NSAdAUC)，该方法遵循逐样本形式的对抗样本生成过程，且可适用于任意满足在[−1，1]区间内严格单调递减的 AUC 替代损失，进而实现对不同替代损失对抗训练方法的统一。在此基础上，本文进一步探索了标准 AUC 与鲁棒 AUC 性能间的权衡问题，通过将鲁棒 AUC 误差分解为标准 AUC 误差与边界 AUC 误差之和，提出了一种基于排序感知对抗正则化的 AUC 对抗训练框架(记为 RARAdAUC)。此外，本文还探讨了如何将上述针对二分类问题开发的 AUC 对抗训练框架及其标准 AUC 和鲁棒 AUC 的权衡框架应用于更复杂实际的多分类任务中。最后，在五个常用长尾基准数据集上的实验证明了所提方法的有效性。

2　研究现状

2.1　ROC 曲线和 AUC 优化方法

受试者操作特征曲线(ROC)刻画了所有可能阈值下模型预测的真阳性率(True Positive Rate，TPR)与假阳性率(False Positive Rate，FPR)间的对应关系，其最早被用于分析二战中雷达接收机操作员的行为[15]，并于 1954 年作为信号检测领域的一个基本工具首次出现在学术文献[16]中。由于通过单一 ROC 概率曲线很难直接比较模型间的优劣

性,ROC 曲线下面积(Area Under the ROC curve,AUC)随即被提出,并作为一种响应偏差的分析工具[17]率先被引入到心理物理学的信号检测理论(SDT)中。

与此同时,机器学习领域也掀起了对 ROC 曲线和 AUC 指标的研究热潮[1,18-20]。文献[1]指出,在衡量模型判别能力方面,ROC 和 AUC 是比准确率(Accuracy,ACC)更好的指标。文献[21]指出,最大化 AUC 和最大化 ACC 之间存在着性能的不一致性。文献[22]的进一步分析表明,相比 ACC,AUC 对长尾数据分布更不敏感,通常在长尾数据下能得到更合理的性能。鉴于 AUC 的良好性质,如何开发最大化 AUC 指标的优化算法获得了研究人员们的广泛关注。早期的 AUC 优化研究主要集中在全批次的离线优化上,例如文献[23-24]使用逻辑斯蒂替代损失和传统的梯度下降法来实现 AUC 的最大化。除此之外,文献[25-26]还针对回归问题中的 AUC 优化问题展开了初步探索。

随着大数据时代的到来,AUC 指标的在线优化问题(online learning)和随机优化问题(stochastic optimization)成为 AUC 可扩展性的主要瓶颈之一。为此,文献[27]率先提出一种基于存储采样技术的 AUC 在线优化方法。文献[28]首次将替代损失为平方损失的 AUC 优化问题重构为了一个随机鞍点问题,使 AUC 损失中互相耦合的正负样本对解耦为一系列逐样本损失的和,显著减小了 AUC 优化的计算负担。在此基础上,文献[5,29]提出了具有更快收敛速度的加速版本,并对 AUC 优化在深度神经网络中的应用做出了初步的探索。与此同时,AUC 优化相关领域所取得的有益效果也得到了良好的理论支持。典型地,文献[2,30-32]等推导出 AUC 优化的泛化误差上界。文献[12,33]探索了 AUC 优化问题的一致性分析等。随着技术与理论研究的不断深入完善,AUC 优化已成功应用于深度学习中的多个任务,如局部 AUC 优化[34-35]、多分类 AUC 优化[3]、多任务学习[36]以及联邦学习[37]等,并已成为长尾/不平衡学习等场景下的标准度量之一。

2.2 对抗训练

已有研究表明,通过精心设计的、人眼不易察觉的对抗样本(Adversarial Examples,AEs)可轻易改变深度神经网络(Deep Neural Networks,DNNs)模型的原始预测结果,造成巨大的安全隐患[38-40]。对抗训练目前已成为抵御对抗样本的一种主流范式。

(1)基于均匀分布假设的对抗训练:通过对 DNNs 稳定性的实验观测,文献[41]首次揭示了神经网络模型的脆弱性,并提出利用干净样本和对抗样本共同训练来规避对抗攻击。考虑到神经网络的结构过于复杂,文献[42]提出一种快速梯度符号法(Fast Gradient Sign Method,FGSM),通过一阶线性逼近生成对抗样本,以实现高效的对抗训练。与此同时,文献[43]提出一个通用的最小最大化(min-max)对抗训练框架,其内层为一个最大化问题—通过在给定扰动半径范围内最大化损失函数产生对抗样本。外层为一个最小化问题—利用生成的对抗样本更新模型参数以最小化损失,从而规避潜在对抗攻击。此后,在最小最大化对抗训练框架的基础上,学界逐渐衍生出一系列有效的对抗训练方法和理论研究[44-46]。其中,通过投影梯度下降(Projected Gradient Descent,PGD)迭代生成对抗样本(简称 PGD-K)[10]已成为提高模型对抗鲁棒性的一个主流基准之一[47-50]。为进一步提升对抗训练的性能,文献[51]提出一种双重鲁棒的样本加权对抗训练方法,根据每个对抗样本在训练过程中的重要程度动态调整样本权重,使模型在训练过程中关注更脆弱的样本。文献[52]探讨了对抗训练过程中模型鲁棒性在不同类、样本和任务间的传递关系,并以此设计了一种子集对抗训练方法,通过选择性地扰动易受攻击的部分样本,在保持对抗鲁棒性的同时,显著提高训练效率。此外,文献[14]面向对抗训练框架下模型鲁棒性和准确性之间的权衡(Trade-off)问题展开系统研究,通过将对抗样本的鲁棒误差分解为干净样本的分类误差和对抗样本的边界误差之和,构建首个通用的鲁棒正则化框架 TRADES 平衡模型的准确性与鲁棒性。在此基础上,文献[53]提出一种像素重加权对抗训练方法,通过对每张对抗样本图像中像素的重要性进行动态加权,聚焦对抗扰动对模型决策影响最大的像素,进一步改善模型鲁棒性和准确性的权衡问题。文献[54]提出 CUR(Conserve-Update-Revise)框架,通过保留原始特征、更新对抗特征、修正潜在偏差,调和模型的准确性与鲁棒性。

(2)基于长尾分布假设的对抗训练:尽管上述方法取得了不错的效果,基于最小化错误率准则设计的对抗训练框架假设数据是平衡的,对类别分布较为敏感。考虑到真实世界(尤其是风险敏感应用场景)中的数据大多呈现出长尾分布态势,该类方法极易忽略尾部类性能指标,造成潜在的安全隐患[55]。

鉴于此,部分研究者开始对长尾分布数据下的

对抗训练问题展开研究。文献[56]通过构建尺度不变分类器和两阶段数据重平衡方法提高长尾分布下模型的对抗鲁棒性。文献[57]利用一种重平衡损失指导模型生成更平衡、信息更丰富的对抗样本，并提出一种挖掘尾部类样本的特征间隔正则化方法平衡头部类和尾部类间的特征表示，缓解长尾分布的影响。文献[58]提出利用 Mix-up 数据增强策略提升尾部类样本的对抗鲁棒性。文献[59]提出引入平衡 Softmax 损失提升长尾分布下对抗训练的性能，并进一步探索了一系列有效的图像增强方法缓解对抗训练中易遭遇的过拟合问题，取得了当前最先进的性能。然而，该类方法仍以最大化准确率为目标，本质上是基于交叉熵损失的扩展方法，因而在长尾数据上的鲁棒性提升十分有限。

与此同时，鉴于 AUC 指标的类别分布不敏感特性，面向 AUC 优化的对抗学习方法近来获得了广泛关注，已成为提升长尾分布下模型对抗鲁棒性的有效的范式之一。文献[11]突破了传统最小化错误率的固有对抗训练模式，率先提出建立基于 AUC 优化的对抗训练机制。针对 AUC 对抗样本生成难、计算复杂度高的问题，文献[11]探索了基于平方替代损失[12]的 AUC 对抗训练方法，提出将正负样本对耦合的 AUC 对抗损失等价重构为逐样本求和的形式，进而实现高效的端到端 AUC 对抗训练。在此基础上，文献[13]进一步解决了文献[11]中内层对抗样本生成过程中的全局假设强、对抗攻击弱以及泛化性能不明确的问题，提出了一种基于随机方差缩减梯度(Stochastic Variance Reduced Gradient，SVRG)的端到端 AUC 对抗训练算法，有效提升了长尾分布下模型的对抗鲁棒性，并推导出了面向 AUC 优化的对抗训练框架的泛化误差上界。

3　预备知识

为了便于说明，本节只讨论二分类问题，但本文涉及的方法和结论在多分类场景中仍然适用。令 $\mathcal{X}=\mathcal{X}_+\cup\mathcal{X}_-$ 为所有训练样本的集合，其中 $\mathcal{X}_+=\{x_i^+|x_i^+\in\mathbb{R}^d\}_{i=1}^{n_+}$ 是从分布 $\mathbb{P}$ 独立同分布采样的正样本集合，$\mathcal{X}_-=\{x_j^-|x_j^-\in\mathbb{R}^d\}_{j=1}^{n_-}$ 是从分布 $\mathbb{N}$ 独立同分布采样的负样本集合，n_+、n_- 分别是正负样本的个数且有 $n:=|\mathcal{X}|=n_++n_-$。相应的标签集合记为 $\mathcal{Y}=\{-1,+1\}^n$，$+1$ 表示正样本，-1 表示负样本。记 $f_\theta:\mathcal{X}\to[0,1]$ 为由深度神经网络构建的预测模型，其输出 $f_\theta(x)$ 表示样本 x 被归类为正样本的概率。记指示函数为 $\mathbb{I}(\cdot)$，当且仅当其内部条件为真时返回 1，否则返回 0。此外，令 $\mathbb{B}(x,\epsilon)=\{\hat{x}:\|\hat{x}-x\|_\infty\le\epsilon\}$ 表示最大扰动半径 ϵ 内，关于样本 x 所有可能的对抗样本 $\hat{x}$ 的可行集。

3.1　AUC 优化框架

依据文献[60]，预测模型 f_θ 的 AUC 指标衡量了正样本取得的得分高于负样本得分的概率，即

$$\mathrm{AUC}(f)=\mathbb{E}_{x^+\sim\mathbb{P}}\mathbb{E}_{x^-\sim\mathbb{N}}[\mathbb{I}(f_\theta(x^+)>f_\theta(x^-))],$$

显然，期望最大化 $\mathrm{AUC}(f)$ 指标以获得最优的模型 f_θ。注意到上式等价于：

$$\mathrm{AUC}(f)=1-\mathbb{E}_{x^+\sim\mathbb{P}}\mathbb{E}_{x^-\sim\mathbb{N}}[\mathbb{I}(f_\theta(x^+)\le f_\theta(x^-))],$$

因而在机器学习问题中通常最小化如下问题：

$$\min_\theta\ \mathbb{E}_{x^+\sim\mathbb{P}}\mathbb{E}_{x^-\sim\mathbb{N}}[\mathbb{I}(f_\theta(x^+)\le f_\theta(x^-))] \tag{1}$$

然而，实际问题中的数据分布 $\mathbb{P}$，$\mathbb{N}$ 往往未知，且指示函数 $\mathbb{I}(\cdot)$ 不可微，难以直接进行端到端计算。

鉴于此，现有的 AUC 优化方法[3,37]常最小化如下经验替代风险：

$$\hat{\mathcal{R}}_{\mathrm{AUC}}^{\theta}(\mathcal{X})=\sum_{x^+\in\mathcal{X}_+}\sum_{x^-\in\mathcal{X}_-}\frac{\ell_{\mathrm{surr}}(f_\theta(x^+)-f_\theta(x^-))}{n_+n_-} \tag{2}$$

其中，ℓ_{surr} 为某连续可导的替代损失函数，常用的替代损失包括[1,12]：

(1) 平方损失(Square loss)：

$$\ell_{\mathrm{surr}}(t):=\ell_{\mathrm{square}}(t)=(1-t)^2;$$

(2) 指数损失(Exponential loss)：

$$\ell_{\mathrm{surr}}(t):=\ell_{\exp}(t)=\exp(-t);$$

(3) 铰链损失(Hinge loss)：

$$\ell_{\mathrm{surr}}(t):=\ell_{\mathrm{hinge}}(t)=\max(0,1-t)。$$

3.2　基于最小化错误率准则的对抗训练框架

对抗训练[31,43]一般形式化为一个极大极小优化问题，内层一般为一个最大化问题，旨在产生半径为 ϵ 的球范围内使损失最大的对抗样本；外层为最小化问题，利用生成的对抗样本更新模型参数使扰动损失函数最小，从而使模型适应当前扰动；二者交替进行，通过不断在训练阶段加入对抗样本，提升模型的鲁棒性：

$$\min_\theta\sum_{i=1}^{n}\max_{\hat{x}_i\in\mathbb{B}(x,\epsilon)}\frac{\ell(f_\theta(\hat{x}_i),y_i)}{n} \tag{3}$$

其中，ℓ 为损失函数，常采用交叉熵损失。

为实现对极大极小问题的高效求解，一般通过近似优化器迭代求解内部最大化问题来生成对抗性样本，然后外层最小化根据对抗样本造成的分类误差更新参数 θ。PGD-K[10]是目前一种有效的方法之一，其第 $k\in[K]$ 步的生成过程如下：

$$\hat{x}_i^k=\mathcal{P}_{\mathbb{B}}(\hat{x}_i^{k-1}+\gamma_k\cdot\text{sign}(\nabla_{\hat{x}_i}\ell(f_\theta(\hat{x}_i^{k-1}),y_i))),$$

其中，$\hat{x}_i^0=x_i$，γ_k 为攻击步长，$\mathcal{P}_{\mathbb{B}}$ 表示投影算子，K 为可调的迭代步数。

然而，简单的通过式(3)可能无法同时兼顾自然误差(natural error)和鲁棒误差(robust error)[61]，例如可能在保证模型鲁棒性的同时却显著降低了其在自然样本上的性能。TRADES[14]是当前权衡模型准确率与鲁棒性问题最有竞争力的方法之一：

$$\min_\theta\frac{1}{n}\sum_{i=1}^{n}\{\ell(f_\theta(x_i),y_i)+\max_{\hat{x}_i\in\mathbb{B}(x_i,\epsilon)}\lambda\cdot\mathbb{KL}(f_\theta(x_i),f_\theta(\hat{x}_i))\}\tag{4}$$

其中，$\lambda>0$ 为平衡系数，$\mathbb{KL}$ 表示 KL 散度(Kullback-Leibler divergence)。通过确定一个合理的 λ 值，可较好的实现模型在干净样本上的性能与对抗鲁棒性能间的权衡。

4 方 法

4.1 当前 AUC 对抗训练框架的局限性

如 3.2 节所述，基于最小化错误率准则设计的对抗训练框架，对类别分布极为敏感，难以应对长尾数据中的类别分布变化。由于 AUC 指标对类别分布的不敏感性，已有研究[11,13]对面向 AUC 优化的对抗训练机制展开探索，以期兼顾 AUC 对长尾数据的不敏感性及对抗学习算法对恶意攻击的稳健性。

具体而言，通过对传统对抗训练方法式(3)的简单扩展，基于 AUC 优化的对抗训练方法可形式化为如下极小极大优化问题：

$$\min_\theta\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\max_{\substack{\hat{x}_i^+\in\mathbb{B}(x_i^+,\epsilon),\\ \hat{x}_j^-\in\mathbb{B}(x_j^-,\epsilon)}}\frac{\ell_{\text{surr}}(f_\theta(\hat{x}_i^+)-f_\theta(\hat{x}_j^-))}{n_+n_-}\tag{5}$$

然而，注意到式(5)中的对抗样本生成过程依赖于较强的成对扰动假设，即攻击者将同时篡改一对正负样本以降低模型的 AUC 性能，而在实际测试阶段的对抗攻击普遍仅应用于单个样本，可能造成模型训练与测试阶段表现的不一致。

为克服该问题，文献[11,13]提出了基于逐样本优化的端到端 AUC 对抗训练框架。具体而言，若选取 ℓ_{square} 作为替代损失 ℓ_{surr}，可以证明，基于平方替代损失的 AUC 对抗训练问题式(5)可重构为如下逐样本形式：

$$\min_{\theta,a,b}\max_{\alpha}\sum_{i=1}^{n}\max_{\hat{x}_i\in\mathbb{B}(x_i,\epsilon)}\mathcal{F}(f_\theta,(\hat{x}_i,y_i),a,b,\alpha)+\kappa\text{Reg}\tag{6}$$

其中，$\kappa>0$；Reg 为与对抗样本扰动项无关的正则化项；$\mathcal{F}_i:=\mathcal{F}(f_\theta,(\hat{x}_i,y_i),a,b,\alpha)$ 为逐样本损失函数，定义如下：

$$\mathcal{F}_i=\frac{(f_\theta(\hat{x}_i)-a)^2\mathbb{I}_{y_i=1}}{\hat{p}}+\frac{(f_\theta(\hat{x}_i)-b)^2\mathbb{I}_{y_i=-1}}{1-\hat{p}}+2(\alpha+1)\left(\frac{f_\theta(\hat{x}_i)\mathbb{I}_{y_i=-1}}{1-\hat{p}}-\frac{f_\theta(\hat{x}_i)\mathbb{I}_{y_i=1}}{\hat{p}}\right)-\alpha^2\tag{7}$$

其中，为便于表达，此处将 $\mathbb{I}(y_i=1)$ 和 $\mathbb{I}(y_i=-1)$ 分别简记为 $\mathbb{I}_{y_i=1}$ 和 $\mathbb{I}_{y_i=-1}$，$\hat{p}=n_+/n$，a,b,α 均为额外引入的可学习超参数。

上述框架虽实现了针对 AUC 对抗机器学习算法的端到端优化，但仍存在一定的局限性：

(1) 该重构技巧仅适用于基于平方替代损失的 AUC 优化问题。然而，面对复杂的实际应用场景，平方替代损失并不一定是最优解。

(2) 现有方法对训练与测试阶段可能存在的对抗攻击的不一致性问题考虑不足。式(6)虽遵循逐样本扰动的形式，但其优化形式复杂，且隐含了一个不切实际的假设—攻击者需准确预知额外的 a,b,α 三个参数。

(3)现有 AUC 对抗训练框架也存在类似的标准 AUC 和鲁棒 AUC 间的权衡问题(见实验部分的 5.5.3 节)，而目前却鲜有针对相关问题的研究。

因此，亟需探索通用的高效 AUC 对抗训练方法，以提升模型在长尾分布数据下的实际性能与对抗鲁棒性。

4.2 基于标准化分数扰动的通用 AUC 对抗训练

实现通用 AUC 对抗训练的关键是解决针对 AUC 指标的高效对抗样本生成问题。如 3.1 节所述，AUC 指标本质上衡量了正样本得分高于负样本得分的概率。因此，为实现对模型 AUC 的破坏，对抗攻击本质上是要通过扰动样本尽可能地降低正样本的得分或提高负样本的得分。基于此，通过对式(5)的简单分析不难发现有以下命题成立。

命题 1. 给定分类器 $f_\theta:\mathcal{X}\to[0,1]$，若所选取的替代损失 $\ell_{\text{surr}}(t)$ 在 $t\in[-1,1]$ 严格单调递减，则基于 AUC 优化的对抗训练式(5)等价于如下基于标准化分数扰动(Normalized Score Adversarial Perturbations，NSAd)的双层优化问题：

$$\begin{aligned}&\min_\theta\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\frac{\ell_{\text{surr}}(f_\theta(\hat{x}_i^{+,*})-f_\theta(\hat{x}_j^{-,*}))}{n_+n_-},\\&\text{s.t. }\hat{x}_i^{+,*}=\mathop{\arg\min}_{\hat{x}_i^+\in\mathbb{B}(x_i^+,\epsilon)}f_\theta(\hat{x}_i^+)\\&\qquad\hat{x}_j^{-,*}=\mathop{\arg\max}_{\hat{x}_j^-\in\mathbb{B}(x_j^-,\epsilon)}f_\theta(\hat{x}_j^-)\end{aligned}\tag{8}$$

评论 1. 由 $\ell_{\text{surr}}(t)$ 在 $t\in[-1,1]$ 严格单调递减性质可知有以下等式成立：

$$\max_{\substack{\hat{x}_i^+\in\mathbb{B}(x_i^+,\epsilon),\\ \hat{x}_j^-\in\mathbb{B}(x_j^-,\epsilon)}} \ell_{\text{surr}}(f_\theta(\hat{x}_i^+)-f_\theta(\hat{x}_j^-))=$$

$$\ell_{\text{surr}}\Big(\min_{\substack{\hat{x}_i^+\in\mathbb{B}(x_i^+,\epsilon),\\ \hat{x}_j^-\in\mathbb{B}(x_j^-,\epsilon)}} \{f_\theta(\hat{x}_i^+)-f_\theta(\hat{x}_j^-)\}\Big)=$$

$$\ell_{\text{surr}}\Big(\min_{\hat{x}_i^+\in\mathbb{B}(x_i^+,\epsilon)} f_\theta(\hat{x}_i^+)-\max_{\hat{x}_j^-\in\mathbb{B}(x_j^-,\epsilon)} f_\theta(\hat{x}_j^-)\Big).$$

上式表明在给定扰动半径 ϵ 的条件下，使得 AUC 损失 ℓ_{surr} 最小的一对正负样本与直接扰动模型对样本的预测得分具有相同的全局最优解。

与此同时，不难发现式(8)中的条件约束可进一步统一为如下形式：

$$\hat{x}_i^*=\arg\min_{\hat{x}_i\in\mathbb{B}(x_i,\epsilon)} y_i\cdot f_\theta(\hat{x}_i) \tag{9}$$

其中对于正样本(负样本)有 $y_i=1$($y_i=-1$)。因此该分数扰动过程仅需作用于单个的样本，且未引入任何可学习的复杂参数，克服了正负样本耦合的问题(L2)，从而实现训练与测试阶段攻击方式的统一。

基于命题 1，易证 3.1 节中介绍的平方损失、指数损失、铰链损失等 AUC 优化中常见的替代损失都满足 ℓ_{surr} 在$[-1,1]$区间严格单调递减的约束。因此，本小节提出如下基于标准化分数扰动的通用 AUC 对抗训练框架：

$$\begin{aligned}&\min_\theta \sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\frac{\ell_{\text{surr}}(f_\theta(\hat{x}_i^{+,*})-f_\theta(\hat{x}_j^{-,*}))}{n_+n_-},\\ &\text{s.t. } \hat{x}_i^{+,*}=\arg\min_{\hat{x}_i^+\in\mathbb{B}(x_i^+,\epsilon)} f_\theta(\hat{x}_i^+)\\ &\qquad \hat{x}_j^{-,*}=\arg\max_{\hat{x}_j^-\in\mathbb{B}(x_j^-,\epsilon)} f_\theta(\hat{x}_j^-)\end{aligned} \tag{10}$$

此处替代损失 ℓ_{surr} 则可根据实际应用场景进行选择，克服了(L1)的瓶颈。为实现对式(11)的高效优化，首先通过与式(3)类似的 PGD-K[10] 算法迭代近似求解其内层的基于标准化分数扰动的对抗攻击问题，生成对抗样本；而后采用随机梯度下降法(Stochastic Gradient Descent，SGD)算法优化模型参数 θ，伪代码见算法 1。

算法 1.　通用 AUC 对抗训练框架

输入：模型 $f_\theta:\mathcal{X}\to[0,1]$；训练数据 $\mathcal{X}=\{(x_i,y_i)\}_{i=1}^n$；批量大小 B；SGD 的学习率 α；训练的迭代轮数 T

输出：鲁棒模型 f_θ

过程：

1. 随机初始化模型的参数 θ_0
2. FOR $t=1$ to T DW
3. 　分层采样一批训练数据 S；
4. 　基于标准化分数扰动(算法 2)生成对抗样本；
5. 　基于生成的批量对抗样本 $\hat{S}$，通过 SGD 更新模型参数：

$$\theta_t=\theta_{t-1}-\alpha\nabla_\theta\hat{\mathcal{R}}_{\text{AUC}}^\theta(\hat{S})$$

6. END FOR

算法 2.　基于标准化分数扰动的对抗样本生成

输入：模型 $f_\theta:\mathcal{X}\to[0,1]$；批量数据集合 $S=\{(x_i,y_i)\}_{i=1}^B$；PGD 的迭代步数 K，迭代步长 $\{\gamma_k\}_{k=1}^K$，最大扰动半径 ϵ

输出：对抗扰动后的批量数据集合：$\hat{S}=\{(\hat{x}_i,y_i)\}_{i=1}^B$

过程：

1. FOR 对于每一个样本 $(x_i,y_i)\in S$ DO
2. 　对抗样本的初始化：$\hat{x}_i^0=x_i$
3. 　FOR $k=1$ to K DO
4. 　　IF $y_i=1$ THEN
5. 　　$\hat{x}_i^k=\mathcal{P}_{\mathbb{B}}(\hat{x}_i^{k-1}-\gamma_k\cdot\text{sign}(\nabla_{\hat{x}_i}f_\theta(\hat{x}_i^{k-1})))$
6. 　　END IF
7. 　　IF $y_i=-1$ THEN
8. 　　$\hat{x}_i^k=\mathcal{P}_{\mathbb{B}}(\hat{x}_i^{k-1}+\gamma_k\cdot\text{sign}(\nabla_{\hat{x}_i}f_\theta(\hat{x}_i^{k-1})))$
9. 　　END IF
10. 　END FOR
11. END FOR

讨论：实际上，只要传统对抗训练式(3)中损失函数 $\ell(f_\theta(\hat{x}_i),y_i)$ 关于 $f_\theta(\hat{x}_i)$ 的单调性满足对于正样本 $y_i=+1$ 严格单调递减，负样本 $y_i=-1$ 严格单调递增，其同样可化简为式(9)类似的基于标准化分数扰动的形式。不难发现，常用的二元交叉熵损失(Binary Cross-Entropy，BCE)满足上述条件，因此有：

$$\arg\max_{\hat{x}_i\in\mathbb{B}(x,\epsilon)}\ell_{\text{BCE}}(f_\theta(\hat{x}_i),y_i)\Leftrightarrow\arg\min_{\hat{x}_i\in\mathbb{B}(x,\epsilon)} y_i\cdot f_\theta(\hat{x}_i) \tag{11}$$

通过比较式(8)和(10)可发现，基于标准化分数扰动的 AUC 对抗方法在一定程度上也可实现跨指标的对抗防御，如攻击者可能利用交叉熵损失合成对抗样本(一种面向 ACC 指标的对抗攻击方法，见 3.2 节)破坏模型的 AUC 指标等。

4.3　标准 AUC 与鲁棒 AUC 的权衡方法

受 TRADES 的启发(见 3.2 节)，克服(L3)的一种可能的、简单直接的方法是直接沿用式(4)的正则化框架，而仅将其第一项针对标准 ACC 的优化项替换为标准 AUC 的经验风险：

$$\min_\theta\hat{\mathcal{R}}_{\text{AUC}}^\theta(\mathcal{X})+\frac{\lambda}{n}\sum_{i=1}^n\mathbb{KL}(f_\theta(x_i),f_\theta(\hat{x}_i^*)) \tag{12}$$

其中，$\hat{x}_i^*$ 仍是基于标准化分数扰动式(9)生成的最优对抗样本。

然而，注意到式(12)中的第二项仅仅考虑了逐样本形式的对抗正则化约束(Instance-wise Adversarial Regularization，IAR)，与关注整体排序性能的 AUC 指标性能不一致，难以兼顾标准 AUC 和鲁棒 AUC 间的关系。

具体而言，面向准确率(ACC)权衡问题的

TRADES[14]证明有以下关系成立：

$$\mathcal{E}_{rob}^{ACC}=\mathcal{E}_{nat}^{ACC}+\mathcal{E}_{bdy}^{ACC} \tag{13}$$

其中，

$$\mathcal{E}_{rob}^{ACC}:=\mathbb{E}[\mathbb{I}(y\cdot g(f_\theta(\hat{x}))\leq 0,\exists\hat{x}\in\mathbb{B}(x,\epsilon))],$$

$$\mathcal{E}_{nat}^{ACC}:=\mathbb{E}[\mathbb{I}(y\cdot g(f_\theta(x))\leq 0)],$$

$$\mathcal{E}_{bdy}^{ACC}:=\mathbb{E}[\mathbb{I}(y\cdot g(f_\theta(\hat{x}))\leq 0,\exists\hat{x}\in\mathbb{B}(x,\epsilon)),\ \text{s.t. } y\cdot g(f_\theta(x))>0],$$

其中令

$$g(z):=\mathbb{I}(z\geq 0.5)-\mathbb{I}(z<0.5)。$$

式(13)表明，模型的鲁棒 ACC 性能等于模型在干净样本上的标准 ACC 性能与边界样本的鲁棒 ACC 性能之和，故可基于此设计优化算法兼顾模型的标准和鲁棒 ACC 性能。

显然，由于 $\mathcal{E}_{rob}^{ACC}\neq\mathcal{E}_{nat}^{AUC}+\mathcal{E}_{bdy}^{ACC}$，通过简单的将标准 AUC 代入进 TRADES 的优化框架难以满足实际需求。为解决该问题，本小节通过进一步对鲁棒 AUC 指标的分解问题展开分析，提出一种基于排序感知对抗正则化(Ranking-aware Adversarial Regularization, RAR)AUC 优化框架(简称 RARAdAUC)，通过捕捉样本扰动前后的排序结果之间的差异，实现标准 AUC 和稳健 AUC 间的权衡。

首先，记标准 AUC 误差(standard AUC error)

$$\mathcal{E}_{nat}^{AUC}=\mathbb{E}_{\mathcal{X}}[\hat{\mathcal{R}}_{\mathrm{AUC}}^{\theta}(\mathcal{X})]=\mathbb{E}_{x^+}\mathbb{E}_{x^-}[\mathbb{I}(f_\theta(x^+)\leq f_\theta(x^-))] \tag{14}$$

在此基础上，基于式(5)有如下鲁棒 AUC 误差(robust AUC error)的定义：

$$\mathcal{E}_{rob}^{AUC}=\mathbb{E}_{x^+}\mathbb{E}_{x^-}[\mathbb{I}(f_\theta(\hat{x}^+)\leq f_\theta(\hat{x}^-)),\ \exists\hat{x}^+\in\mathbb{B}(x^+,\epsilon),\hat{x}^-\in\mathbb{B}(x^-,\epsilon)] \tag{15}$$

基于上述定义，易证有以下结论成立。

命题 2. 鲁棒 AUC 误差可以分解为标准 AUC 误差与边界 AUC 误差(boundary AUC error)之和：

$$\mathcal{E}_{rob}^{AUC}=\mathcal{E}_{nat}^{AUC}+\mathcal{E}_{bdy}^{AUC} \tag{16}$$

其中，$\mathcal{E}_{bdy}^{AUC}$ 定义如下：

$$\mathbb{E}_{x^+}\mathbb{E}_{x^-}[\mathbb{I}(f_\theta(x^+)>f_\theta(x^-))\cdot\mathbb{I}(f_\theta(\hat{x}^+)\leq f_\theta(\hat{x}^-)),\ \exists\hat{x}^+\in\mathbb{B}(x^+,\epsilon),\hat{x}^-\in\mathbb{B}(x^-,\epsilon)]。$$

评论 2. 显然，当 $\epsilon=0$ 时有 $\mathcal{E}_{rob}^{AUC}=\mathcal{E}_{nat}^{AUC}$ 成立。因此，若 $\mathbb{I}(f_\theta(x^+)\leq f_\theta(x^-))\neq 0$，此时在 $\epsilon=0$ 处即可扰动成功，故 $\mathbb{I}(f_\theta(\hat{x}^+)\leq f_\theta(\hat{x}^-))\neq 0$ 也成立。若 $\mathbb{I}(f_\theta(x^+)\leq f_\theta(x^-))=0$，则需进一步考虑 $\epsilon>0$ 时的位于分类边界的样本误差，即得 $\mathcal{E}_{bdy}^{AUC}$。

此外，通过将 TRADES[14]中定义的边界 ACC 误差 $\mathcal{E}_{bdy}^{ACC}$ 和命题 2 中定义的边界 AUC 误差 $\mathcal{E}_{bdy}^{AUC}$ 相比较可知，二者最大的区别在于 $\mathcal{E}_{bdy}^{AUC}$ 是基于一对正负样本定义的而不是逐个样本，这与 AUC 指标关注整个列表中正负样本排序关系的特性相吻合。

基于命题 2，本小节提出如下基于排序感知正则化的 AUC 对抗训练方法：

$$\min_\theta\hat{\mathcal{R}}_{\mathrm{AUC}}^{\theta}(\mathcal{X})+\lambda\sum_{x\in\mathcal{X}}\mathbb{D}(\widetilde{f}_\theta(x),\widetilde{f}_\theta(\hat{x}^*)),\ \text{s.t. } \hat{x}^*=\arg\min_{\hat{x}\in\mathbb{B}(x,\epsilon)}y\cdot f_\theta(\hat{x}) \tag{17}$$

其中，$\widetilde{f}_\theta(x)=\dfrac{\exp(f_\theta(x))}{\sum_{x'\in\mathcal{X}}\exp(f_\theta(x'))}$ 衡量了在给定模型对所有样本预测得分的情况下该样本的排名情况[62-63]，$\mathbb{D}(\cdot,\cdot)$为任意的分布度量，如 KL-散度。

最后，式(17)的优化过程与式(11)类似。由于式(17)中的正则项捕捉了干净样本排名与对抗样本排名的差异，因此当 $\hat{\mathcal{R}}_{\mathrm{AUC}}^{\theta}(\mathcal{X})\approx 0$，即任意正负样本对都满足 $f_\theta(x^+)>f_\theta(x^-)$ 且 $\mathbb{D}(\widetilde{f}_\theta(x),\widetilde{f}_\theta(\hat{x}^*))\approx 0$ 时，对抗样本间的排序一致性也将得到满足，即 $\mathbb{I}(f_\theta(\hat{x}^+)>f_\theta(\hat{x}^-))$ 处处成立，进而实现了标准 AUC 与鲁棒 AUC 之间的合理权衡。

伪代码见算法 3，此处为了便于说明，记

$$\hat{\mathcal{R}}_{\mathrm{RAR}}^{\theta}(\mathcal{X},\hat{\mathcal{X}}):=\hat{\mathcal{R}}_{\mathrm{AUC}}^{\theta}(\mathcal{X})+\lambda\sum_{x\in\mathcal{X}}\mathbb{D}(\widetilde{f}_\theta(x),\widetilde{f}_\theta(\hat{x}^*))。$$

算法 3. 标准 AUC 与鲁棒 AUC 的权衡方法

输入：模型 $f_\theta:\mathcal{X}\to[0,1]$；训练数据 $\mathcal{X}=\{(x_i,y_i)\}_{i=1}^n$；批量大小 B；SGD 的学习率 α；训练的迭代轮数 T

输出：鲁棒模型 f_θ

过程：

1. 随机初始化模型的参数 θ_0
2. FOR $t=1$ to T DO
3. 分层采样一批训练数据 S；
4. 基于标准化分数扰动(算法 2)生成对抗样本；
5. 基于生成的批量对抗样本 $\hat{S}$ 计算排序感知正则化 AUC 对抗损失，并通过 SGD 更新模型参数：

$$\theta_t=\theta_{t-1}-\alpha\nabla_\theta\hat{\mathcal{R}}_{\mathrm{RAR}}^{\theta}(S,\hat{S})$$

6. END FOR

4.4 多分类情形下 AUC 的对抗训练问题

前面的章节聚焦于二分类问题，详细介绍了如何开发面向 AUC 的对抗训练框架及其标准 AUC 和鲁棒 AUC 性能间的权衡方法。本节将进一步讨论如何将所提方法应用于实际更普遍的多分类任务。

符号简记：在多分类任务的背景下，记 $\mathcal{Y}=\{y_1,y_2,\cdots,y_C\}$ 为标签空间，其中 $C>2$ 是总的类别数，并设输入空间：

$$\mathcal{X}=\bigcup_{c\in[C]}\mathcal{X}_c,$$

其中，$\mathcal{X}_c=\{x_i\mid x_i\in\mathbb{R}^d\}_{i=1}^{n_c}$ 是属于类别 y_c 的样本集合，d 是样本特征的维度，n_c 是类 c 的样本数量。多分类情况下旨在构建一个参数为 Θ 的分类模型，记为

$f_\Theta=(f_\Theta^{(1)},f_\Theta^{(2)},\cdots,f_\Theta^{(C)}):\mathcal{X}\to[0,1]^C$，其中 $f_\Theta^{(c)}:\mathcal{X}\to[0,1]$，$\forall c\in[C]$表示样本为第 c 个类的概率。

为实现对 f_Θ 的高效学习，现有多分类 AUC (Multi-class AUC，MAUC)优化方法通常遵循“一对多”(One vs All，OVA)的训练范式——基本思想是将多分类 AUC 的性能分解为一组二分类 AUC 的平均值，并已在机器学习社区中取得了巨大的成功[3,9,34]。具体来说，给定一个预测模型 $f_\Theta=(f_\Theta^{(1)},f_\Theta^{(2)},\cdots,f_\Theta^{(C)})$，多类 AUC 问题的目标是最大化以下每类平均 AUC 得分：

$$\mathrm{MAUC}(f)=\frac{1}{C}\sum_{c=1}^{C}\mathrm{AUC}_{c|\neg c}(f^{(c)})。$$

类似于 3.1 节，这里 $\mathrm{AUC}_{c|\neg c}(f^{(c)})$指的是每个二分类器 $f^{(c)}$ 的 AUC 得分，其中正样本是属于第 c 类的样本(即令 $\mathcal{X}_+=\mathcal{X}_c$)，负样本是其他 $y\neq c$ 的样本(即令 $\mathcal{X}_-=\mathcal{X}\setminus\mathcal{X}_c$)。最大化上述目标等价于最小化如下 MAUC 的经验风险：

$$\hat{\mathcal{R}}_{\mathrm{MAUC}}^{\Theta}(\mathcal{X})=\frac{1}{C}\sum_{c=1}^{C}\hat{\mathcal{R}}_{\mathrm{AUC}}^{\Theta^c}(\mathcal{X}_{c|\neg c})\tag{18}$$

其中，$\mathcal{X}_{c|\neg c}=\mathcal{X}_+\cup\mathcal{X}_-$ 需根据不同类 c 划分正负样本，Θ^c 表示 $f^{(c)}$ 的模型参数。

受到上述 MAUC 定义的启发，可以发现，多分类情形下的每个 $f_\Theta^{(c)}$，$\forall c\in[C]$本质上仍可看作一个二元分类器，故仍可采用与命题 1 类似的分数扰动技术攻击每个通道生成对抗样本。

具体而言，由于一个样本(x_i,y_i)，$y_i\in[C]$需同时参与 f_Θ中所有 C 个通道的计算，为实现对多分类 AUC 指标的攻击则需降低该样本在其所属类分类器 $f_\Theta^{(y_i)}$ 处的得分，并提高在其他类分类器处的得分，即

$$\hat{x}_i^*=\arg\min_{\hat{x}_i\in\mathbb{B}(x,\epsilon)}\Big\{f_\Theta^{(y_i)}(\hat{x}_i)-\sum_{c\in[C]\setminus y_i}f_\Theta^{(c)}(\hat{x}_i)\Big\}\tag{19}$$

最后，与式(11)和(17)类似，针对多分类任务中 AUC 的对抗训练问题，仍可通过式(19)生成对抗样本，同时在利用式(18)兼顾多分类模型的 AUC 性能与鲁棒性，伪代码见算法 4。

算法 4.　多分类 AUC 对抗训练框架

输入：模型 $f_\Theta=(f_\Theta^{(1)},f_\Theta^{(2)},\cdots,f_\Theta^{(C)}):\mathcal{X}\to[0,1]^C$；训练数据 $\mathcal{X},\mathcal{Y}$；批量大小 B；SGD 的学习率 α；训练的迭代轮数 T

输出：鲁棒模型 f_Θ

过程：

1. 随机初始化模型的参数Θ_0
2. FOR $t=1$ to T DO
3. 　分层采样一批训练数据 S；
4. 　对 S 中的每个样本采用与算法 2 类似的式(19)生成对抗样本集合 $\hat{S}$
5. 　对于多分类鲁棒 AUC 优化(NSAdAUC)
$$\Theta_t=\Theta_{t-1}-\alpha\nabla_\Theta\hat{\mathcal{R}}_{\mathrm{MAUC}}^{\Theta}(\hat{S})$$
6. 对于多分类鲁棒 AUC 的权衡问题(RARAdAUC)
FOR $c=1$ to C DO
根据是否属于类别 c 划分 S 中的所有样本 $S_{c|\neg c}$
根据是否属于类别 c 划分 $\hat{S}$ 中的所有样本 $\hat{S}_{c|\neg c}$
END FOR
根据划分后的数据，通过 SGD 更新模型参数
$$\Theta_t=\Theta_{t-1}-\alpha\nabla_\Theta\hat{\mathcal{R}}_{\mathrm{MRAR}}^{\Theta}(S,\hat{S}),$$
其中
$$\hat{\mathcal{R}}_{\mathrm{MRAR}}^{\Theta}(S,\hat{S})=\frac{1}{C}\sum_{c=1}^{C}\hat{\mathcal{R}}_{\mathrm{RAR}}^{\Theta^c}(S_{c|\neg c},\hat{S}_{c|\neg c})$$
7. END FOR

讨论：本小节的多分类 MAUC(f) 指标本质上是 3.1 节中传统 AUC(f) 定义的直接扩展[3]，在二分类情况下二者等价。

具体而言，当 $C=2$ 时，通常仅需构建一个单通道分类器 $f_\theta:\mathcal{X}\to[0,1]$即可完成二分类，其输出 $f_\theta(x)$表示样本 x 被分类为正样本的概率(见第 3 节)。若定义一个二通道分类预测器 $f_\Theta=(f_\Theta^{(1)},f_\Theta^{(2)}):\mathcal{X}\to[0,1]^2$，其输出经 Softmax 函数归一化，同时设 $f_\Theta^{(1)}(x)$和 $f_\Theta^{(2)}(x)$分别表示样本 x 为正样本和负样本的概率，此时有

$$\mathrm{MAUC}(f)=\frac{1}{2}\{\mathrm{AUC}_{1|\neg1}(f^{(1)})+\mathrm{AUC}_{2|\neg2}(f^{(2)})\},$$

其中，$\mathrm{AUC}_{1|\neg1}$衡量 $f_\Theta^{(1)}$通道正样本得分高于负样本的概率，而 $\mathrm{AUC}_{2|\neg2}$衡量$f_\Theta^{(2)}$通道负样本得分高于正样本得分的概率。

由于$f_\Theta^{(1)}(x)+f_\Theta^{(2)}(x)=1$，易证 $\mathrm{AUC}_{1|\neg1}(f^{(1)})=\mathrm{AUC}_{2|\neg2}(f^{(2)})$，从而可得

$$\mathrm{MAUC}(f)=\mathrm{AUC}_{1|\neg1}(f^{(1)})=\mathrm{AUC}(f),$$

与 3.1 节的指标等价。

5　实　验

5.1　数据集介绍

本文选取以下四个长尾基准数据集验证所提方法的有效性，数据集的详细统计信息参见表 1.

表 1　数据集描述

数据集	正类 ID	正类样本总数	负类样本总数	不平衡比
CIFAR-10-LT	5～9	891	11515	12.92
CIFAR-100-LT	50～99	946	9901	10.47
Tiny-ImageNet-200-LT	70,81,94,107,111,116,121,133,145,153,164,166	4200	65800	15.67
CheXpert	lung diseases	7635	57496	7.53

(1)二分类CIFAR-10-LT数据集:由文献[11,13]从原始CIFAR-10数据集[64](共10类)中采样得到,其中不同类别的样本量呈指数衰减,且保证最少类的样本数量与最多类的样本量比例为0.01。最后,将前五个类全部看作负类,后五个类视为正类,其正负样本的不平衡比为$\rho=12.92$。

(2)二分类CIFAR-100-LT数据集:采样自CIFAR-100数据集[64](共100类),其不平衡数据的构建策略与二分类CIFAR-10-LT数据集完全相同。这里将前五十个类全部看作负类,后五十个类视为正类,所得二分类CIFAR-100-LT数据集的不平衡比$\rho=10.47$。

(3)二分类Tiny-ImageNet-200-LT数据集:采样自Tiny-ImageNet-200数据集,处理方式和二分类的CIFAR-10-LT数据集保持一致。这里将所有的交通工具类视为正类,其余类记为负类,最终二分类Tiny-ImageNet-200-LT数据集的不平衡比$\rho=15.67$。

(4)二分类CheXpert数据集:采样自ChestX-ray14数据集[65-66]。原始数据集是一个具有五个标签的多标签X光胸片医学图像数据集,每个标签对应一种病症。参照文献[11,13],将至少有一种症状的图像记为正样本,而没有任何症状的图像记为负样本,最终的不平衡比$\rho=7.53$。

5.2 对比方法

为验证所提方法的有效性,将其性能与以下几个当前最先进的对抗训练方法进行比较:

(1)基于交叉熵损失的对抗训练方法(Cross-Entropy based AT,CE-AT)[10]:这是当前以最大化准确率为目标的主流对抗训练范式。

(2)基于鲁棒性与准确率权衡的对抗训练方法(TRADES)[14]:将对抗样本的鲁棒误差分解为干净样本的分类误差和对抗样本的边界误差之和,并以此构建了合理的代理优化损失,使得模型同时兼顾在干净样本上的准确率以及对抗样本上的鲁棒性。

(3)稳健且平衡(Robust and Balanced,RoBal)的对抗训练方法[56]:通过引入尺度不变的余弦分类器和两阶段的数据重平衡方法,提高长尾分布下模型对抗鲁棒性。

(4)基于平衡Softmax损失(Balanced Softmax Loss,BSL)的对抗训练方法[59]:通过引入平衡Softmax损失提升长尾分布下对抗训练的性能,并进一步探索了一系列的图像增强方法缓解其对抗训练中遭遇的过拟合问题,其主要包括了三种有效方法:BSL-AT仅使用平衡的Softmax损失,而不是传统的交叉熵损失(像Robal所使用的),作为最终的对抗训练优化目标;BSL-AT-AuA在BSL-AT的基础上进一步引入Autoaugment(AuA)[67]的数据增强方法;BSL-AT-RA在BSL-AT的基础上进一步引入Randaugment(RA)[68]的数据增强方法。感兴趣读者可参见文献[59]了解详细技术细节。

(5)像素级加权的对抗训练方法[53](Pixel Reweighted AT,PRAT):一种通用的像素重加权技术,通过对每个像素扰动施加不同的权重提高对抗训练的准确性与鲁棒性。此处考虑文献[53]给出的三种变体方法:基于传统交叉熵损失对抗训练[10]的像素重加权(PRAT)、基于TRADES的像素重加权(PRAT-T)以及基于错误率自感知对抗训练[69]权(PRAT-M)。建议感兴趣读者参考文献[53]了解详细技术细节。

(6)双重鲁棒样本加权对抗训练方法[51](Doubly Robust Instance-Reweighted AT,DONE):通过对抗样本加权和干净样本加权实现对抗训练中鲁棒性和准确性之间的平衡。

(7)基于AUC优化的对抗训练方法是提升长尾分布下模型鲁棒性的一种有效的对抗学习新范式。目前的代表工作主要有两个版本,分别记为AdAUC1[11]和AdAUC2[13]。AdAUC1首次将AUC优化引入到对抗训练问题中来,提出将正负样本耦合的、难计算的对抗AUC经验风险重构为一个逐样本的极小极大化问题,从而实现了面向AUC指标的端到端对抗训练,克服了长尾分布对模型的影响。在此基础上,AdAUC2进一步解决了AdAUC1中内层对抗攻击的生成能力较弱且泛化性难评估的问题,同时提出了一个基于随机方差缩减梯度(Stochastic Variance Reduced Gradient,SVRG)的端到端对抗AUC优化算法。

5.3 实验细节

5.3.1 网络结构

对于二分类CIFAR-10-LT和CIFAR-100-LT数据集,选择WideResNet-28[70]作为基础骨干网络。模型输入为$32\times32\times3$的图像,输出为640维的特征向量。对于Tiny-ImageNet-200-LT和CheXpert数据集,利用ResNet-18[71]作为基础骨干网络,其中两个数据集的输入均为$224\times224\times3$的图像,生成1024维的特征向量。最后,在骨干模型的基础上进一步接入一个全连接层以实现分类预测。

5.3.2　训练细节

所有实验的代码均使用 PyTorch[72] 深度学习框架实现，并采用网格搜索来确定最优的超参数值。具体来说，所有数据集上的批次大小(Batch Size)设置为 64，学习率选择范围为 $\{3\times10^{-2},1\times10^{-2},3\times10^{-3},1\times10^{-3},3\times10^{-4},1\times10^{-4}\}$ 且每 20 个 epoch 后学习率将衰减 0.1；对于交叉熵损失、RoBal 等对比方法以及本文提出的方法，使用带有 Nesterov 动量的随机梯度下降(Stochastic Gradient Descent，SGD)优化算法，其中动量(momentum)参数设置为 0.95，权重衰减(weight decay)参数设置为 10^{-4}；对于 AdAUC1 方法，采用原文中提出的随机梯度下降上升优化算法(Stochastic Gradient Descent Ascent，SGDA)，且参数设置均遵循原文中的最优参数；对于 AdAUC2 方法，采用原文中提出的随机方法缩减的梯度(Stochastic Variance Reduced Gradient，SVRG)优化算法，且参数设置均遵循原文中的最优参数。对于本文所提方法，τ 固定为 1，且 $\lambda\in\{1,3,6,8,10\}$。此外，训练阶段的最大对抗扰动范围固定为 $\epsilon=8/255$，攻击步长固定为 2/255，PGD 攻击中的最大 K 设置为 10。最后，鉴于既往 AdAUC1 和 AdAUC2 框架都采用平方损失作为 AUC 对抗训练的替代损失，若不特别说明，本文也默认采取同样的替代损失，以尽可能地保证性能比较过程中的公平性。第 5.6.3 节提供了不同替代损失的消融实验。

5.4　评估方案

本文主要采用以下两种对抗攻击方式评估所有对抗训练方法的 AUC 鲁棒性。

5.4.1　基于标准化分数扰动的对抗攻击

就二分类问题而言，基于标准化分数的对抗攻击方法(Normalized Score-based Adversarial Attack，NSAd)遵循如下方式生成对抗样本：

$$\delta_i^*=\arg\min_{\delta_i\in\mathbb{B}(x_i,\epsilon)} y_i\cdot f_\theta(x_i+\delta_i),$$

其中对于正样本 x_i 有 $y_i=1$，负样本 x_i 有 $y_i=-1$。

对于该问题仍可通过投影梯度下降法(Project Gradient Descent，PGD-K)迭代 K 步近似求解：

$$\hat{x}_i^k=\mathcal{P}_{\mathbb{B}}(\hat{x}_i^{k-1}-y_i\cdot\gamma_k\cdot\mathrm{sign}(\nabla_{\hat{x}_i}f_\theta(\hat{x}_i^{k-1})))$$

5.4.2　跨指标对抗攻击

在此方案中，首先测试了所有方法在受到对抗扰动后 AUC 指标的鲁棒性。实验中使用了两种攻击方法，投影梯度下降攻击法(PGD-K)[10] 和自动攻击法(AutoAttack，AA)[73]。

(1) 基于准确率(交叉熵损失)的攻击(简称 ACCAd)。仍利用 PGD-K 求解，从无扰动的样本开始，沿着最大化交叉熵损失的梯度方向迭代 K 步生成扰动样本，定义如下：

$$x^k=\mathcal{P}_{\mathbb{B}}(x^{k-1}+\gamma_i\cdot\mathrm{sign}(\nabla_x\ell(g_\theta(x^{k-1}),y)))。$$

(2) 自动攻击法(AA)[73]，集成了多种攻击方法，如 Auto-PGD(APGD)、目标版本的 APGD-DLR、FAB[74] 和平方攻击法(Square Attack)[75] 等，算法可自动调整对抗参数实现攻击。

5.5　性能分析

5.5.1　NSAd 攻击下的性能比较

表 2 报告了四个长尾二分类数据集上不同方法在标准化分数攻击下的性能。基于实验结果，可以得出以下结论：

(1) 本文所提方法(即 NSAdAUC 和 RARAdAUC)在大多数情况下都取得了最佳的标准 AUC 和鲁棒 AUC 性能，表明了所提方法的优越性。

(2) 与基于交叉熵损失的(CE-based)对抗训练方法相比，基于 AUC 的对抗训练方法(包括现有的 AUC-based 方法和本文提出的方法)整体上展现了更强的鲁棒性，且在不同数据集上的性能更加稳定。值得注意的是，尽管某些基于交叉熵损失的长尾对抗训练方法(如 Robal 和 BSL)在一定程度上提升了长尾分布下模型的对抗鲁棒性，但其性能提升幅度相对有限。例如，在 Tiny-ImageNet-200-LT 数据集上，整体表现最好的 AUC 方法(RARAdAUC)在 PGD-20 处的性能比 Robal 和 BSL-AT-RA 方法分别提升了 9.33% 和 3.34%。这一结果进一步验证了优化分布不敏感 AUC 指标在应对长尾数据分布下对抗攻击时的显著优势，与本文的研究动机高度一致。

(3) 所提方法显著优于现有 AUC 对抗训练方法(AdAUC1 和 AdAUC2)，这主要归因于本文所提方法具有更简明的形式，避免了求解更为复杂的随机鞍点问题(式(6))，从而在性能上实现了突破。

(4) 在标准 AUC 和鲁棒 AUC 的权衡方面，RARAdAUC 相较于 NSAdAUC 在大多数情况下的性能表现更为合理。尽管在 CIFAR-100-LT 数据集上，RARAdAUC 的鲁棒 AUC 性能略低于 NSAdAUC，但与其他方法(如 TRADES、PART 和 DONE)相比，RARAdAUC 依然展现出极具竞争力的表现。

表 2 在 NSAd 攻击下的标准 AUC 和鲁棒 AUC 性能比较

数据集	类型	方法	Clean	PGD-1	PGD-5	PGD-10	PGD-20
CIFAR-10-LT	CE-based	CE-AT	68.05	63.45	51.83	48.53	47.96
		TRADES	73.52	71.52	66.10	64.53	64.38
		RoBal	75.18	62.81	55.35	52.09	51.94
		BSL-AT	74.29	72.10	65.63	63.90	63.67
		BSL-AT-AuA	73.71	68.78	59.75	57.08	56.69
		BSL-AT-RA	74.21	71.77	64.77	62.05	61.78
		PART	69.14	64.53	56.10	54.53	51.62
		PART-T	74.21	72.47	65.17	64.33	63.86
		PART-M	73.64	71.66	67.18	65.04	64.20
		DONE	74.62	71.83	64.50	62.39	62.18
	AUC-based	AdAUC1	71.23	66.50	56.59	52.30	49.75
		AdAUC2	75.49	72.03	62.89	60.01	59.87
	Ours	NSAdAUC	81.11	79.57	75.19	74.02	73.94
		RARAdAUC	**84.17**	**82.34**	**76.29**	**74.43**	**74.30**
CIFAR-100-LT	CE-based	CE-AT	57.66	51.84	38.41	35.23	34.50
		TRADES	61.54	55.29	41.89	38.43	37.48
		RoBal	61.88	57.03	44.45	40.42	40.03
		BSL-AT	56.72	54.23	40.98	37.45	37.08
		BSL-AT-AuA	57.04	50.37	36.70	33.72	33.15
		BSL-AT-RA	57.90	51.50	37.75	34.11	33.92
		PART	52.56	48.27	37.86	35.33	35.06
		PART-T	61.65	56.52	43.45	39.24	38.83
		PART-M	58.89	54.02	41.93	39.06	38.11
		DONE	54.35	51.18	42.86	40.68	40.50
	AUC-based	AdAUC1	63.13	57.85	45.98	41.79	40.07
		AdAUC2	63.59	60.14	**50.95**	44.81	40.56
	Ours	NSAdAUC	65.18	63.50	49.47	**48.56**	**48.55**
		RARAdAUC	**70.20**	**64.17**	47.26	41.78	41.19
Tiny-ImageNet-200-LT	CE-based	CE-AT	89.72	87.23	78.66	74.59	73.21
		TRADES	89.72	87.65	77.97	74.86	74.57
		RoBal	92.87	88.57	78.64	74.79	74.34
		BSL-AT	91.45	87.40	81.85	79.32	79.02
		BSL-AT-AuA	92.18	88.70	82.53	80.16	80.12
		BSL-AT-RA	92.25	89.02	83.01	80.64	80.33
		PART	81.45	76.17	70.54	66.09	65.66
		PART-T	86.61	82.05	78.46	74.67	74.31
		PART-M	86.24	83.18	78.68	74.57	74.26
		DONE	86.68	84.82	78.45	76.75	76.71
	AUC-based	AdAUC1	93.05	88.62	78.80	73.44	71.35
		AdAUC2	93.73	84.71	78.69	74.78	74.26
	Ours	NSAdAUC	90.19	89.14	85.88	**84.95**	**84.91**
		RARAdAUC	**93.77**	**92.26**	**86.03**	83.83	83.67
CheXpert	CE-based	CE-AT	75.19	75.92	73.29	72.77	72.71
		TRADES	81.96	79.67	72.94	71.10	71.06
		RoBal	75.55	75.06	73.49	72.97	72.77
		BSL-AT	74.87	74.61	73.05	72.80	72.37
		BSL-AT-AuA	77.91	76.59	74.54	74.35	74.03
		BSL-AT-RA	76.51	75.26	73.53	73.41	73.16
		PART	71.88	71.50	68.48	67.91	67.87
		PART-T	77.46	75.04	72.01	70.97	70.78
		PART-M	76.87	74.59	71.41	70.38	70.16
		DONE	73.28	71.24	61.22	58.71	58.43
	AUC-based	AdAUC1	75.67	74.86	73.37	72.92	72.91
		AdAUC2	75.97	75.76	73.58	73.02	72.89
	Ours	NSAdAUC	76.77	75.70	75.33	74.27	74.21
		RARAdAUC	**84.91**	**84.91**	**84.91**	**82.43**	**82.41**

5.5.2　跨指标对抗攻击场景下的性能比较

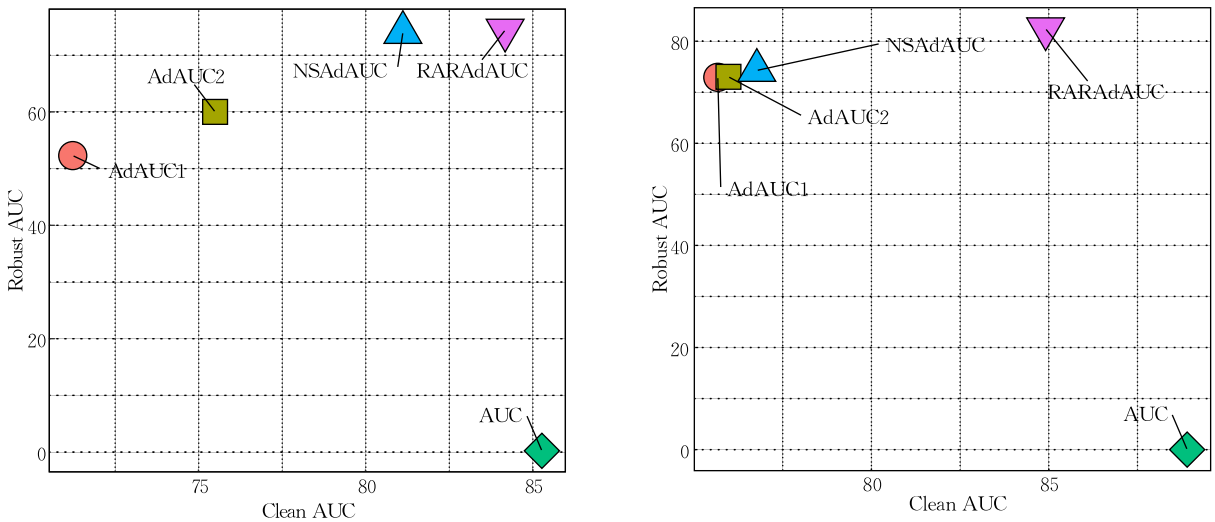
如第 4.1 节所述，在实际对抗攻防场景中，恶意攻击者可能采取各式各样复杂的攻击方式以达到破坏模型的目的。为验证所提方法在该类情况下的鲁棒性，本小节在 CIFAR-10-LT 数据集上模拟了不同方法抵御跨指标攻击的能力（即 ACCAd 和 AA 攻击），实验结果如表 3 所示。结果表明：(1) NSAdAUC 和 RARAdAUC 均展现出超越其他对比方法的卓越性能，验证了 NSAdAUC 和 RARAdAUC 抵御潜在复杂多样对抗攻击的潜力；(2) 在基于 ACCAd 的跨指标对抗攻击场景下，既往面向 AUC 的对抗训练方法，包括 AdAUC1 和 AdAUC2，显著低于本文所提的两种方法，甚至低于部分基于交叉熵损失的方法（如 TRADES BSL-AT 等）。例如，RARAdAUC 在 PGD-20 处的鲁棒性比 AdAUC2 提升了 23.43%。这是由于既往 AUC 对抗训练范式仅考虑成对扰动的对抗攻击方式，存在训练与测试阶段鲁棒性不一致的问题，而本文提出的标准化分数扰动的对抗攻击可较好的克服该问题。

表 3　CIFAR-10-LT 数据集上 ACCAd 和 AA 攻击的标准 AUC 和鲁棒 AUC 性能比较

类型	方法	Clean	PGD-5	PGD-10	PGD-20	AA
CE-based	CE-AT	68.05	48.70	44.17	43.19	43.84
	TRADES	73.52	59.74	56.85	56.64	56.62
	RoBal	75.18	51.61	48.78	48.45	48.45
	BSL-AT	74.29	65.09	64.06	64.01	64.03
	BSL-AT-AuA	73.71	64.58	63.88	63.79	63.84
	BSL-AT-RA	74.21	63.73	62.65	62.52	62.61
	PART	69.14	65.26	64.42	64.35	66.61
	PART-T	74.21	66.41	65.72	65.62	67.98
	PART-M	73.64	66.50	65.78	65.70	66.64
	DONE	74.62	65.65	63.78	63.63	65.35
AUC-based	AdAUC1	71.23	55.91	53.09	52.83	52.91
	AdAUC2	75.49	57.08	53.41	53.02	52.94
Ours	NSAdAUC	81.11	74.58	73.34	73.35	71.32
	RARAdAUC	**84.17**	**77.56**	**76.44**	**76.45**	**74.61**

5.5.3　标准 AUC 和鲁棒 AUC 权衡的验证

为说明 AUC 对抗训练中的权衡问题，图 1 比较了 CIFAR-10-LT 和 CheXpert 数据集上不同 AUC 方法的标准 AUC 和鲁棒 AUC 性能（PGD-10 处）间的关系，其中 AUC 方法表示直接优化式(2)而不考虑对抗训练。显然，仅优化式(2)可取得最好的标准 AUC 性能，而鲁棒 AUC 却趋近于 0。传统的 AUC 对抗训练算法，即 AdAUC1 和 AdAUC2，以及本文所提的 NSAdAUC 方法，虽取得较好的鲁棒 AUC 性能，但标准 AUC 仍然较低。相比之下，RARAdAUC 通过引入排序感知的对抗性正则化，实现了最优的性能权衡，证明了所提方法的有效性。



(a) CIFAR-10-LT数据集　　(b) CheXpert数据集

图 1　CIFAR-10-LT 和 CheXpert 数据集上标准 AUC 和鲁棒 AUC 间的性能比较

5.5.4　多分类问题中的 AUC 鲁棒性

为验证所提方法在多分类问题中的有效性，本小节进一步在多分类 CIFAR-10-LT[64] 数据集上进行了实验。方法的具体细节见 4.4 节，长尾数据集的处理方式与 5.1 节保持一致。表 4 报告了在多分类 CIFAR-10-LT 数据集上不同方法的标准 AUC 性能及在 NSAd 攻击下的鲁棒 AUC 的性能。结果表明，在标准 AUC 和鲁棒 AUC 性能方面，本文所提方法皆显著优于既往方法，包括专为长尾分布对抗攻击场景开发的方法（如 Robal、BSL、AdAUC1 和 AdAUC2）。例如，在 PGD-20 处，所提的 NSAdAUC 和 RARAdAUC 相比于现有最好方法的鲁棒 AUC 性能提升分别达到 6.16%和 4.80%。此外，相比于 NSAdAUC，RARAdAUC 依然可更有效地兼顾多

分类问题中的标准 AUC 与鲁棒 AUC 性能，其在 PGD-20 处的鲁棒 AUC 性能虽比 NSAdAUC 下降了 1.36%，但标准 AUC 性能却提升了 3.11%，这进一步证明了所提方法的有效性。

表 4 多分类 CIFAR-10-LT 数据集上的标准 AUC 和鲁棒 AUC 性能比较

类型	方法	Clean	PGD-1	PGD-5	PGD-10	PGD-20
CE-based	CE-AT	70.45	67.91	59.63	57.70	57.68
	TRADES	79.20	78.75	66.41	63.60	63.27
	RoBal	76.41	74.39	66.98	64.66	64.69
	BSL-AT	64.05	56.13	42.97	41.27	41.22
	BSL-AT-AuA	79.32	72.56	55.19	51.02	50.66
	BSL-AT-RA	77.09	70.39	53.31	49.22	48.83
	PART	68.52	67.17	61.94	59.77	59.40
	PART-T	72.20	70.56	64.16	62.22	61.93
	PART-M	71.08	70.03	63.99	60.38	60.08
	DONE	60.45	58.65	53.55	52.80	52.77
AUC-based	AdAUC1	72.84	71.26	66.61	65.44	65.45
	AdAUC2	75.20	73.46	68.55	67.33	67.36
Ours	NSAdAUC	85.54	82.84	**75.45**	**73.48**	**73.52**
	RARAdAUC	**88.65**	**85.40**	75.08	72.23	72.16

5.5.5 不同长尾分布下的性能比较

为充分证明所提方法在长尾对抗攻击场景下的优势，本小节进一步比较不同不平衡比下各方法针对 NSAd 的鲁棒性。具体而言，本实验基于二分类 CIFAR-10-LT 数据集展开，进一步比较了不平衡比 $\rho=46.82$ 和 $\rho=19.01$ 下的鲁棒性，其他设置与第 5.3.2 节保持一致，结果如表 5 所示。首先，注意到在不同的不平衡场景下，部分方法（如 TRADES 和 PART 等）的鲁棒性甚至低于传统的 CE-AT 方法。一种可能的原因为该类方法仅聚焦于均匀分布对抗训练下的改进，难以适用于长尾分布场景。

其次，传统 AUC 对抗训练方法（AdAUC1 和 AdAUC2）由于优化形式复杂且忽略了标准 AUC 和鲁棒 AUC 权衡问题，致使在不同长尾分布下的性能欠佳。最后，实验结果一致表明，所提方法均显著优于现有的对抗训练方法，证明了所提方法的有效性和通用性。

表 5 CIFAR-10-LT 数据集不同平衡比下标准 AUC 和鲁棒 AUC 性能比较

类型	方法	Clean	PGD-1	PGD-5	PGD-10	PGD-20
不平衡比$\rho=46.82$						
CE-based	CE-AT	70.61	69.69	67.15	66.57	66.56
	TRADES	75.80	70.14	53.89	48.87	48.01
	RoBal	72.29	69.15	59.19	57.04	56.69
	BSL-AT	77.13	72.43	63.29	61.02	60.59
	BSL-AT-AuA	78.29	73.74	61.42	56.81	55.31
	BSL-AT-RA	77.16	73.57	63.98	60.52	59.42
	PART	63.11	58.12	44.88	41.59	41.28
	PART-T	69.34	66.76	59.63	58.31	58.02
	PART-M	68.00	66.19	58.72	56.97	56.64
	DONE	68.71	66.22	59.24	57.26	57.07
AUC-based	AdAUC1	72.34	68.14	56.51	53.18	53.11
	AdAUC2	77.22	74.57	66.22	64.33	63.95
Ours	NSAdAUC	81.76	80.57	77.09	76.15	76.13
	RARAdAUC	**85.16**	**83.48**	**78.29**	**76.92**	**76.87**
不平衡比$\rho=19.01$						
CE-based	CE-AT	68.13	66.81	63.23	62.26	62.16
	TRADES	79.57	75.24	62.26	58.35	57.92
	RoBal	79.14	74.86	62.43	59.64	59.32
	BSL-AT	75.46	72.10	63.56	60.95	60.58
	BSL-AT-AuA	76.73	73.09	61.50	57.97	57.23
	BSL-AT-RA	75.62	71.34	60.95	58.68	58.18
	PART	65.11	61.12	49.88	47.59	47.28
	PART-T	72.35	70.16	63.49	60.79	60.03
	PART-M	71.00	69.78	62.65	59.71	59.43
	DONE	71.36	68.13	58.65	55.88	55.62
AUC-based	AdAUC1	64.52	61.86	55.00	52.82	52.27
	AdAUC2	71.00	67.53	57.01	51.92	51.58
Ours	NSAdAUC	84.44	82.68	77.45	**76.05**	**75.98**
	RARAdAUC	**87.96**	**85.66**	**78.15**	75.49	75.22

5.5.6 均匀分布下的性能比较

尽管本文的主要关注点是长尾分布下的对抗攻防问题，本小节进一步探讨了平衡数据分布下的对抗鲁棒性，以验证所提框架的应用潜力。本小节从原始 CIFAR-10 数据集[64]中选取“猫（cat）”和“鹿（deer）”两类，各 5000 张图片用于训练，另各 1000 张用于测试，并将所提方法与交叉熵损失的对抗训练方法 CE-AT 和 TRADES 进行比较，其他实验配置与第 5.3.2 节保持一致，结果如图 2 所示。

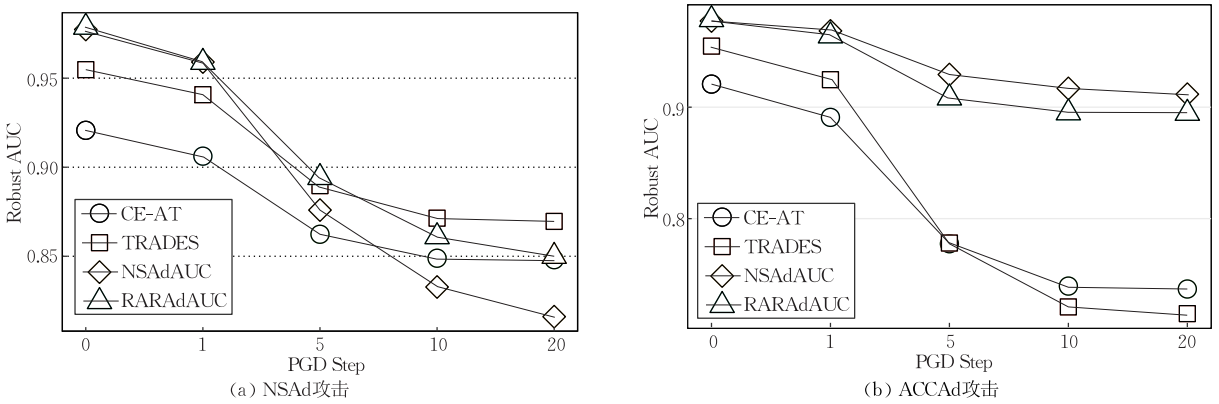


图 2 平衡数据分布下 AUC 对抗训练的性能验证（为了便于描述，PGD step=0 处的性能即为标准 AUC 的性能）

由于 AUC 指标对数据分布变化不敏感，即使在平衡场景下，所提方法仍展现出极具竞争力的性能。具体而言，在 NSAd 攻击下，所提方法在标准 AUC 指标上优于基于交叉熵损失的方法，但在鲁棒 AUC 性能方面有些许的性能下降。而在 ACCAd 攻击下，NSAdAUC 和 RARAdAUC 显著优于其他方法，且性能提升更加明显。例如，在 PGD-20 处的性能，NSAdAUC 比 CE-AT 提升了 15.76%。

综上所述，所提的 AUC 对抗训练框架在平衡数据分布下仍然能够表现出优异的性能。

5.6　定量实验

5.6.1　不同攻击参数的敏感性分析

图 3 展示了所提出的两种方法 NSAdAUC 和 RARAdAUC 在测试阶段不同 NSAd 攻击强度下的鲁棒 AUC 性能。具体而言，该实验在二分类 CIFAR-10-LT 数据集上进行，分别对 AdAUC1、AdAUC2、NSAdAUC、RARAdAUC 方法不同的扰动半径和攻击步长鲁棒性结果进行可视化分析。其中扰动半径和扰动步长参数分别设置为

{4/255, 6/255, 8/255, 10/255}和

{0.5/255, 1/255, 2/255, 4/255}。

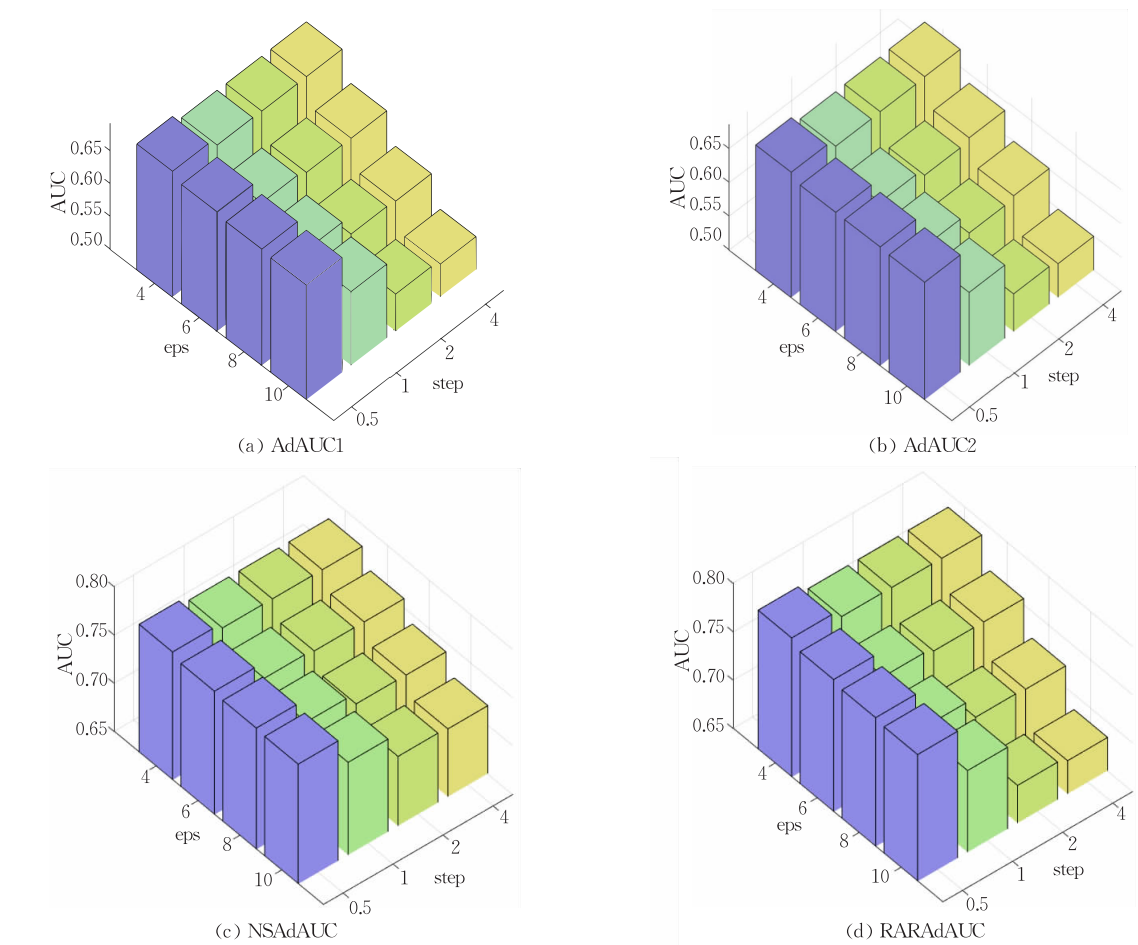


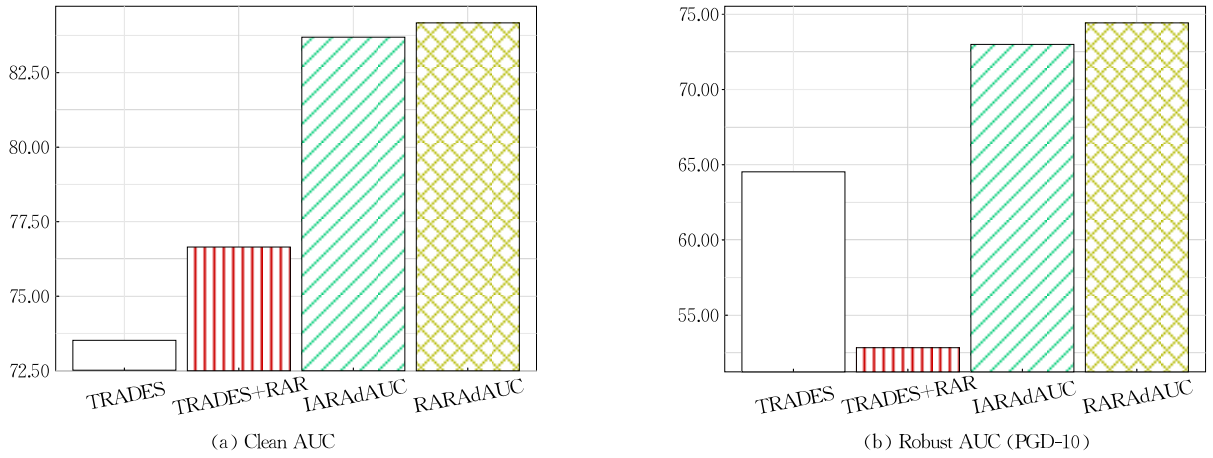
图 3　二分类 CIFAR-10-LT 数据集上不同 NSAd 攻击参数下 AUC 对抗训练的敏感性分析

结果表明，在 NSAd 攻击方式下，所提 NSAdAUC 和 RARAdAUC 方法显示出相似的敏感性趋势，且在攻击步长 2/255 和 4/255 的性能基本类似。与此同时，注意到随着攻击步长和扰动半径的逐步增大，模型的鲁棒 AUC 性能也整体呈现出逐渐降低的趋势。此外，在不同的测试参数下（如不同的扰动半径和攻击步长），本文所提两种方法的性能也显著优于现有的 AUC 对抗训练算法（AdAUC1 和 AdAUC2），这进一步证明了所提方法的有效性。

5.6.2　RARAdAUC 正则化方案的消融分析

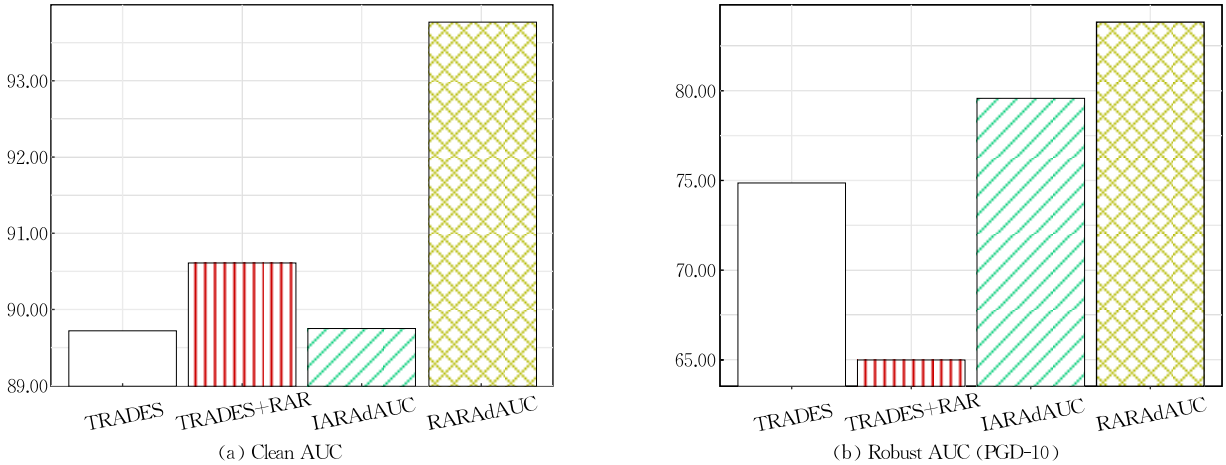
为验证第 4.3 节提出的排序感知正则化方案（RAR）的有效性，本小节进一步将其扩展至以准确率为目标的 TRADES 方案中来。具体而言，将所提的 RAR 策略应用至传统 TRADES 中（简记为 TRADES+RAR），以替换其原始设计的逐样本的对抗正则化（IAR）。与此同时，作为对照，进一步将 IAR 应用至 4.3 节所提的 AUC 权衡问题中来（简记为 IARAdAUC）。实验分别在 CIFAR-10-LT 和

Tiny-ImageNet-200-LT 数据集上进行，结果如图 4 和图 5 所示。首先，由于所提 RAR 策略更多地关注模型预测结果中正负样本的排序关系，故以准确率为优化目标的 TRADES 方法在配备了 RAR 正则化项后（TRADES+RAR），也可以在一定程度上提升在标准 AUC（Clean AUC）上的性能。然而，观察到在鲁棒 AUC 性能方面（Robust AUC），TRADES+RAR 方法表现不佳，这主要是由于 TRADES 与 RAR 的优化目标不一致所致（见第 4.3 节）。与此同时，IARAdAUC 在 clean 和 robust 性能方面均表现不如 RARAdAUC，尤其在 Tiny-ImageNet-200-LT 数据集上，这进一步证明了所提方法的有效性。



(a) Clean AUC　　(b) Robust AUC (PGD-10)

图 4　CIFAR-10-LT 数据集上的消融结果



(a) Clean AUC　　(b) Robust AUC (PGD-10)

图 5　Tiny-ImageNet-200-LT 数据集上的消融结果

5.6.3　不同替代损失的 AUC 对抗训练

如 3.1 节所述，常见的 AUC 替代损失主要有平方损失（square）、指数损失（exp）和铰链损失（hinge）。为此，本小节在 CIFAR-10-LT 数据集上分别对 NSAdAUC 和 RARAdAUC 的可扩展性展开验证。在 NSAd 攻击下的实验结果如表 6 所示。

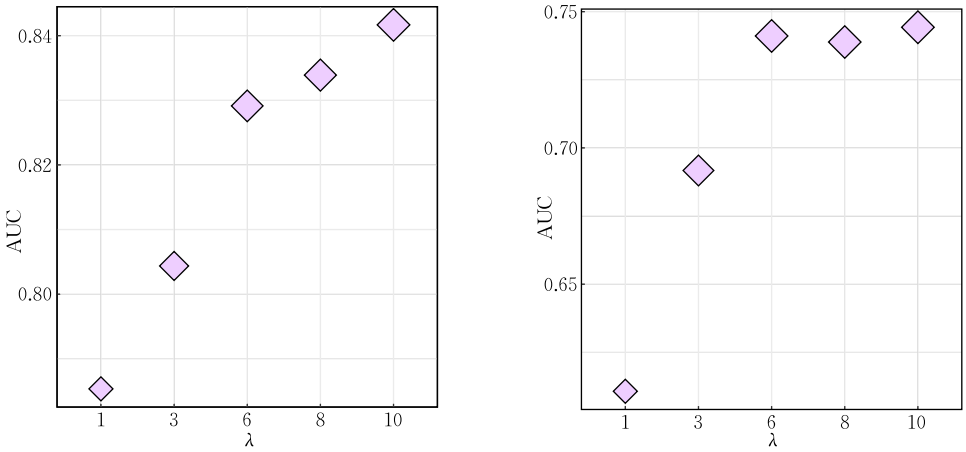
表 6　CIFAR-10-LT 数据集上不同替代损失的标准 AUC 和鲁棒 AUC 性能比较

方法	Clean	PGD-1	PGD-5	PGD-10	PGD-20
NSAdAUC-hinge	80.27	78.29	72.77	71.29	71.19
NSAdAUC-exp	78.92	77.25	72.65	71.37	71.28
NSAdAUC-square	**81.11**	**79.57**	**75.19**	**74.02**	**73.94**
RARAdAUC-hinge	82.56	78.71	66.29	62.82	62.68
RARAdAUC-exp	82.57	78.32	64.58	60.69	60.55
RARAdAUC-square	**84.17**	**82.34**	**76.29**	**74.43**	**74.30**

结果表明，在大多数情况下，所提方法即使在不同替代损失下也均表现竞争力的性能，且优于现有 AdAUC1 和 AdAUC2 方法。此外，总体而言，基于平方替代损失的 AUC 优化方法具有最突出的性能表现，这与现有的研究[31,60]相一致。

5.6.4　平衡系数 λ 的灵敏性分析

图 6 展示了在 CIFAR-10-LT 数据集上所提出 RARAdAUC 方法采用不同平衡系数 $\lambda \in \{1,3,6,8,10\}$ 的性能。结果表明，适当的 λ 值可以显著提高模型的标准 AUC 和鲁棒 AUC 性能，表明了所提的排序感知正则化方案的重要作用。



(a) Clean AUC vs. λ　　(b) Robust AUC (PGD-10) vs. λ

图 6　CIFAR-10-LT 数据集上 λ 的敏感性分析

5.6.5　多分类场景下的训练效率

图 7 展示了在多分类 CIFAR-10-LT 数据集上以下方法单次迭代(epoch)的训练时间开销：(1) CE-AT；(2) TRADES；(3) AdAUC2；(4) NSAdAUC；(5) RARAdAUC。由于 AdAUC1 和 AdAUC2 方法均采用逐样本扰动的形式[11,13]，两者效率接近，故仅考虑 AdAUC2。如预期所示，CE-AT 因其形式简单，实现了最佳的训练效率。与此同时，采用逐样本损失优化的 AdAUC2 也展现出了具有竞争力的训练效率。而 TRADES 则因其对抗样本生成和损失最小化的过程需要同时计算两项损失(见式(4))，导致整体效率较低。相比之下，尽管所提方法需要计算成对比较的 AUC 损失，带来一定的计算负担，但在生成对抗样本时，所提方法通过攻击得分来实现对抗样本生成，无需额外计算复杂的损失函数，从而在训练时间开销上取得了较好的平衡，展现出可接受的训练效率。

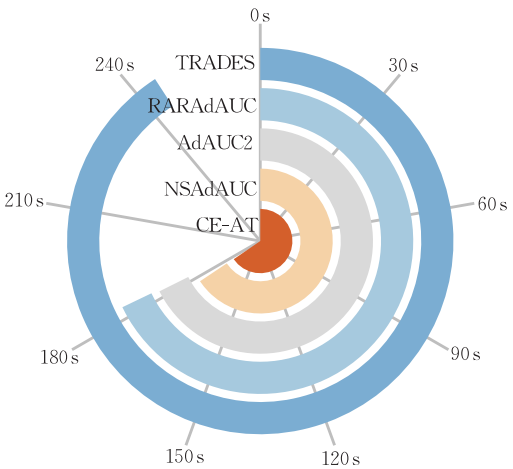


图 7　多分类场景下的训练效率比较，越靠近内圈表示效率越高

6　总结与展望

面向长尾对抗攻防场景，本文重新审视了既往基于平方损失设计的 AUC 对抗训练方法，揭示了其存在的可扩展性差、对抗攻击假设强以及鲁棒性能难权衡等问题。鉴于此，本文首先构建了基于标准化分数扰动的通用 AUC 对抗训练框架 NSAdAUC，通过将不同替代损失下的对抗攻击统一转化为分数攻击，克服可扩展性差、攻击假设强等问题。在此基础上，本文进一步探索了一种基于排序感知对抗正则化的 AUC 对抗训练框架 RARAdAUC，通过显式地优化模型的标准 AUC 性能和决策边界处的正负样本，以兼顾模型的标准 AUC 和鲁棒 AUC 性能。在二分类 CIFAR-10-LT、CIFAR-100-LT、Tiny-ImageNet-200-LT、CheXp-ert 以及多分类 CIFAR-10-LT 数据集上的实验结果表明，本文所提方法在多种攻击方式下均显著优于当前最佳的基线方法，获得了更为先进的模型鲁棒性。最后，一系列的定量实验从多个角度验证了所提方法的可行性及有效性。

尽管本文提出的方法在长尾分布的对抗攻防场景下取得了显著的性能提升，但仍存在一些局限性。首先，AUC 优化的损失函数式(2)需要对所有正负样本对进行计算，在多分类场景中，随着样本数量的增加，对计算资源的需求也随之增加，训练效率可能难以保障。其次，本文的研究仅聚焦于 AUC 优化领域中常见的三种替代损失(包括平方损失、指数损失和铰链损失)，尚未涉及基于一般化损失[76]的 AUC 优化问题，因此在应对复杂多样的下游任务需求时仍显不足。最后，在长尾分布条件下生成的对

抗样本分布可能与干净样本分布存在显著差异,这种分布偏移现象可能对模型的泛化能力产生影响。未来的研究将进一步针对长尾分布下的 AUC 对抗训练问题展开探索,以克服上述潜在的局限性。

参 考 文 献

[1] Ling C X, Huang J, Zhang H. AUC: A statistically consistent and more discriminating measure than accuracy//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2003: 519-526

[2] 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

[3] Yang Z, Xu Q, Bao S, et al. Learning with multiclass AUC: Theory and algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7747-7763

[4] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861-874

[5] Liu M, Yuan Z, Ying Y, Yang T. Stochastic AUC maximization with deep neural networks//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2019: 1-9

[6] Hao H, Fu H, Xu Y, et al. Open-narrow-synechiae anterior chamber angle classification in AS-OCT sequences. arXiv preprint arXiv: abs/2006.05367, 2020

[7] Zhou K, Gao S, Cheng J, et al. Sparse-GAN: Sparsity-constrained generative adversarial network for anomaly detection in retinal OCT image//Proceedings of the IEEE International Symposium on Biomedical Imaging. Iowa City, USA, 2020: 1227-1231

[8] Liu W, Luo W, Lian D, Gao S. Future frame prediction for anomaly detection—A new baseline//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6536-6545

[9] Huang M, Liu Y, Ao X, et al. AUC-oriented graph neural network for fraud detection//Proceedings of the ACM Web Conference. Virtual, Lyon, France, 2022: 1311-1321

[10] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-9

[11] Hou W, Xu Q, Yang Z, et al. AdAUC: End-to-end adversarial AUC optimization against long-tail problems//Proceedings of the International Conference on Machine Learning. Maryland, USA, 2022: 8903-8925

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

[13] Yang Z, Xu Q, Hou W, et al. Revisiting AUC-oriented adversarial training with loss-agnostic perturbations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15494-15511

[14] Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 7472-7482

[15] Wixted J T. The forgotten history of signal detection theory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2020, 46(2): 201-233

[16] Peterson W, Birdsall T, Fox W. The theory of signal detectability. Transactions of the IRE Professional Group on Information Theory, 1954, 4(4): 171-212

[17] Green D M, Swets J A. Signal Detection Theory and Psychophysics. New York, USA: Wiley, 1966

[18] Bowyer K, Kranenburg C, Dougherty S. Edge detector evaluation using empirical ROC curves. Computer Vision and Image Understanding, 2001, 84(1): 77-103

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

[20] Hand D J, Till R J. A simple generalization of the area under the ROC curve for multiple class classification problems. Machine Learning, 2001, 45(2): 171-186

[21] Cortes C, Mohri M. AUC optimization vs. error rate minimization//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2003: 313-320

[22] Carrington AM, Manuel DG, Fieguth PW, et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 329-341

[23] Herschtal A, Raskutti B. Optimising area under the ROC curve using gradient descent//Proceedings of the International Conference on Machine Learning. Banff, Canada, 2004: 385-392

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

[25] Menon A K, Jiang X J, Vembu S, et al. Predicting accurate probabilities with a ranking loss//Proceedings of the International Conference on Machine Learning. Edinburgh, UK, 2012: 703-710

[26] Sculley D. Combined regression and ranking//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 979-988

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

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

[29] Natole M, Ying Y, Lyu S. Stochastic proximal algorithms for AUC maximization//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 3710-3719

[30] Clémençon S, Lugosi G, Vayatis N. Ranking and empirical minimization of U-statistics. The Annals of Statistics, 2008, 36(2): 844-874

[31] Usunier N, Amini M-R, Gallinari P. A data-dependent generalization error bound for the AUC//Proceedings of the International Conference on Machine Learning Workshop on ROC Analysis in Machine Learning. Bonn, Germany, 2005: 1-9

[32] Wu G, Li C, Yin Y. Towards understanding generalization of macro-AUC in multi-label learning//Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 37540-37570

[33] Agarwal S. Surrogate regret bounds for bipartite ranking via strongly proper losses. The Journal of Machine Learning Research, 2014, 15(1): 1653-1674

[34] Yang Z, Xu Q, Bao S, et al. When all we need is a piece of the pie: A generic framework for optimizing two-way partial AUC//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 11820-11829

[35] Yang Z, Xu Q, Bao S, et al. Optimizing two-way partial AUC with an end-to-end framework. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(8): 10228-10246

[36] Jiang W, Li G, Wang Y, et al. Multi-block-single-probe variance reduced estimator for coupled compositional optimization//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 32499-32511

[37] Yuan Z, Guo Z, Xu Y, et al. Federated deep AUC maximization for heterogeneous data with a constant communication complexity//Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021: 12219-12229

[38] Athalye A, Carlini N. On the robustness of the CVPR 2018 white-box adversarial example defenses. arXiv preprint arXiv:1804.03286, 2018

[39] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time//Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference. Prague, Czech Republic, 2013: 387-402

[40] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 274-283

[41] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the International Conference on Learning Representations. Banff, Canada, 2014: 1-9

[42] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-9

[43] Huang R, Xu B, Schuurmans D, Szepesvari C. Learning with a strong adversary. arXiv preprint arXiv:1511.03034, 2015

[44] Bai Y, Gautam T, Sojoudi S. Efficient global optimization of two-layer ReLU networks: Quadratic-time algorithms and adversarial training. SIAM Journal on Mathematics of Data Science, 2023, 5(2): 446-474

[45] Wang Y, Ma X, Bailey J, et al. On the convergence and robustness of adversarial training//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 6586-6595

[46] Bao H, Scott C, Sugiyama M. Calibrated surrogate losses for adversarially robust classification//Proceedings of the Conference on Learning Theory. Virtual, 2020: 408-451

[47] Zhang J, Xu X, Han B, et al. Attacks which do not kill training make adversarial learning stronger//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 11278-11287

[48] Wang J, Zhang H. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks //Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 6628-6637

[49] Wong E, Rice L, Kolter J. Fast is better than free: Revisiting adversarial training//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2019: 1-17

[50] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[51] Sow D, Lin S, Wang Z, Liang Y. Doubly robust instance-reweighted adversarial training//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 1-27

[52] Losch M, Omran M, Stutz D, et al. On adversarial training without perturbing all examples//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 28-49

[53] Zhang J, Liu F, Zhou D, et al. Improving accuracy-robustness trade-off via pixel reweighted adversarial training//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 59382-59402

[54] Gowda S, Zonooz B, Arani E. Conserve-update-revise to cure generalization and robustness trade-off in adversarial training//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 50-71

[55] Li G, Tong W, Yang T. Maximization of average precision for deep learning with adversarial ranking robustness//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023: 15475-15496

[56] Wu T, Liu Z, Huang Q, et al. Adversarial robustness under long-tailed distribution//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8659-8668

[57] Li G, Xu G, Zhang T. Alleviating the effect of data imbalance on adversarial training. arXiv preprint arXiv: 2307. 10205, 2023

[58] Wang W, Shower H, Wan Y, et al. A mix-up strategy to enhance adversarial training with imbalanced data//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham, UK, 2023: 2637-2645

[59] Yue X, Mou N, Wang Q, Zhao L. Revisiting adversarial training under long-tailed distributions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 24492-24501

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

[61] Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A. Robustness may be at odds with accuracy//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-23

[62] Cynthia R. The P-Norm push: A simple convex ranking algorithm that concentrates at the top of the list. Journal of Machine Learning Research, 2009, 10: 2233-2271

[63] Cao Z, Qin T, Liu T Y, et al. Learning to rank: From pairwise approach to listwise approach//Proceedings of the International Conference on Machine Learning. Corvallis, USA, 2007: 129-136

[64] Yann L, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[65] Mahapatra D, Ge Z. Training data independent image registration using generative adversarial networks and domain adaptation. Pattern Recognition, 2020, 100: 107-109

[66] Wang X, Peng Y, Lu L, et al. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2097-2106

[67] Cubuk E D, Zoph B, Mane B, et al. AutoAugment: Learning augmentation strategies from data//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 113-123

[68] Cubuk E D, Zoph B, Shlens J, Le Q V. RandAugment: Practical automated data augmentation with a reduced search space//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 702-703

[69] Wang Y, Zou D, Yi J, et al. Improving adversarial robustness requires revisiting misclassified examples//Proceedings of the International Conference on Learning Representation. Addis Ababa, Ethiopia, 2019: 1-13

[70] Zagoruyko S, Komodakis N. Wide residual networks//Proceedings of the British Machine Vision Conference. York, UK, 2016: 87. 1-87. 12

[71] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[72] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 8024-8035

[73] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 2206-2216

[74] Croce F, Hein M. Minimally distorted adversarial examples with a fast adaptive boundary attack//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 2196-2205

[75] Andriushchenko M, Croce F, Flammarion N, Hein M. Square attack: A query-efficient black-box adversarial attack via random search//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 484-501

[76] Yang Z, Shen W, Ying Y, Yuan X. Stochastic AUC optimization with general loss. Communications on Pure & Applied Analysis, 2020, 19(8): 4191-4212

BAO Shi-Long, Ph. D., assistant professor. His research interests include machine learning and data mining, with a special focus on AUC optimization and collaborative ranking.

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

YANG Zhi-Yong, Ph. D., associate professor. His research interests lie in machine learning and learning theory, with special focus on AUC optimization, meta-learning/multi-task learning, and learning theory for recommender systems.

HUA Cong, Ph. D. candidate. His research interests include machine learning and multi-modal learning.

HAN Bo-Yu, Ph. D. candidate. His research interests include computer vision and generative model.

CAO Xiao-Chun, Ph. D., professor. His main research interests include computer vision and multimedia analysis.

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

Background

With the rapid advancement of AI, security concerns have grown within the machine learning community, particularly the vulnerability that attackers can subtly manipulate inputs to alter a model's predictions. Adversarial Training (AT) has emerged as one of the effective paradigms for enhancing model robustness. However, traditional AT methods are typically developed under the assumption of balanced data distribution, which may be inadequate for real-world applications, where data is often long-tailed.

Given that the tail classes are often more important than head ones, recent studies have introduced the distribution-insensitive AUC metric to the AT community (called AdAUC), yielding promising results. Nevertheless, several challenges remain. Firstly, current methods are limited to squared loss and may not generalize to other AUC surrogates, such as hinge and exponential losses. In fact, the squared surrogate loss may not always be the optimal choice for complex real-world applications, limiting the further development of AdAUC. Additionally, they assume that malicious attackers will adhere to pairwise perturbations of AUC when attacking models, whereas adversarial attacks in real-world scenarios are typically conducted on individual inputs, causing inconsistency between the optimization goal and the practical risk. Furthermore, AT is known to compromise model performance toward clean samples, yet little attention has been paid to how to balance standard AUC with robust AUC in existing studies.

To address these issues, this paper proposes a unified AUC-oriented adversarial training framework. The core idea is to equivalently reformulate adversarial attacks induced by various surrogate AUC losses into normalized score adversarial perturbations. This allows AdAUC to adopt a pointwise adversarial example generation process, and thus can be applied to all mainstream AUC surrogate losses (denoted as NSAdAUC). Taking a step further, this paper starts the first trial to investigate the trade-off between standard AUC and robust AUC performance. By decomposing robust AUC error into the sum of standard AUC error and boundary AUC error, we develop a rank-aware adversarial regularization framework (denoted as RARAdAUC). Additionally, we also explore these methods to more complex multi-class tasks. Finally, comprehensive experiments on five widely used long-tail benchmark datasets consistently demonstrate the effectiveness of our proposed methods.

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, 62441232, U21B2038, U23B2051, 62122075, 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. XDB0680201, and in part by the Postdoctoral Fellowship Program of CPSF under Grant No. GZB20240729.