学校代码：10284
分 类 号：TP181
密　　级：不涉密
U D C：004.8
学　　号：DG1833006

南京大学

博 士 学 位 论 文

论 文 题 目　弱标记 AUC 优化研究

作 者 姓 名　解铮

专 业 名 称　计算机科学与技术

研 究 方 向　机器学习

导 师 姓 名　黎铭 教授

2023 年 8 月 31 日

答辩委员会主席　　陈松灿 教授

评　　阅　　人　　陈松灿 教授

张敏灵 教授

王　魏 教授

俞　扬 教授

张利军 教授

论文答辩日期　　2023 年 8 月 7 日

研究生签名：

导师签名：

Weakly Labeled AUC Optimization

by

Zheng Xie

Supervised by

Professor Ming Li

A dissertation submitted to

the graduate school of Nanjing University

in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science and Technology

Department of Computer Science and Technology

Nanjing University

August 31, 2023

南京大学学位论文原创性声明

本人郑重声明，所提交的学位论文是本人在导师指导下独立进行科学研究工作所取得的成果。除本论文中已经注明引用的内容外，本论文不包含其他个人或集体已经发表或撰写过的研究成果，也不包含为获得南京大学或其他教育机构的学位证书而使用过的材料。对本文的研究做出重要贡献的个人和集体，均已在论文的致谢部分明确标明。本人郑重申明愿承担本声明的法律责任。

研究生签名： 

日期：2023.8.31

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：弱标记 AUC 优化研究

计算机科学与技术 专业 2018 级博士生姓名：解铮

指导教师（姓名、职称）：黎铭 教授

摘　要

AUC 是机器学习中常用的评价指标，被认为相比准确率能够更好地表征模型阈值无关的整体性能。通过 AUC 优化构建模型，可以避免数据分布不平衡所产生的负面影响，并使模型具有更好的样本识别与排序能力。然而，现有的 AUC 优化方法主要面向强标记的场景，弱标记场景下的 AUC 优化问题鲜有研究关注。为此，本文以标记弱化程度为主线，针对不同程度标记弱化场景下的弱标记 AUC 优化问题开展系统性研究，取得如下创新成果：

1. 针对标记不完全可见的弱标记场景，提出了半监督 AUC 优化方法SAMULT。本文提出了利用无标记数据进行无偏 AUC 风险估计的方法，进而给出了基于风险最小化的半监督 AUC 优化方法。实验表明，该方法无需依赖对类别先验概率的知识，并在标记不完全可见的 AUC 优化任务中取得了良好的效果。

2. 针对标记不完全可见的流式弱标记数据场景，提出了半监督在线 AUC 优化方法 SOLA。本文通过将优化问题重写为随机鞍点问题，解决了半监督在线 AUC 优化中无法基于样本对计算风险的难题，实现了基于单个样本点的随机梯度优化方法。实验表明，该方法在标记不完全可见的流式数据场景 AUC 优化任务中取得了良好的效果，并显著减少了模型更新开销。

3. 针对标记不准确且不完全可见的弱标记场景，提出了弱监督 AUC 优化框架 WSAUC。该框架将不同的弱监督信息转化为统一形式，并基于部分 AUC 优化算法实现了通用的稳健弱监督 AUC 优化方法。实验表明，该框架在标记不准确且不完全可见等多种场景下的 AUC 优化任务中取得了良好的效果。

4. 针对标记不可见的弱标记场景，提出了利用多个无标记集合进行 AUC 优化的方法 U^m-AUC。本文提出仅依赖多个具有不同先验的无标记样本集合构建 AUC 优化模型的方法，并通过将其转化为多标记 AUC 优化问题实现了高效求解。实验表明，该方法在标记不可见的 AUC 优化任务中取得了良好的效果。

关键词：机器学习；AUC；AUC 优化；弱标记学习；半监督学习；弱监督学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Weakly Labeled AUC Optimization

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Zheng Xie

MENTOR: Professor Ming Li

ABSTRACT

AUC is a commonly used evaluation metric in machine learning, which is believed to better characterize a model's threshold-independent overall performance compared with accuracy. By learning models through AUC optimization, negative impacts caused by imbalanced data distribution can be avoided, and the model's sample recognition and ranking ability can be improved. However, existing AUC optimization methods are primarily designed for strongly labeled scenarios, with little research attention given to AUC optimization in weakly labeled scenarios. Therefore, this thesis systematically investigates the weakly labeled AUC optimization problem for different levels of label weakening, and achieves the following innovative results:

1. Proposal of a semi-supervised AUC optimization method SAMULT, for scenarios with incomplete labels: This thesis introduces a method that utilizes unlabeled data to assist in unbiased AUC risk estimation, thereby presenting the semi-supervised AUC optimization method. Experiments show that this method does not rely on prior knowledge of class prior probability, and achieves superior AUC optimization performance under scenarios with incomplete labels.

2. Proposal of a semi-supervised online AUC optimization method SOLA, for scenarios with incomplete labels in streaming data: This thesis reformulates the semi-supervised AUC optimization problem as a stochastic saddle point problem, achieving a stochastic gradient optimization method based on individual data points. This method resolves the difficulties of semi-supervised online AUC optimization, which arise from the inability to compute risks based on sample pairs. Experiments show that this method achieves superior AUC optimization performance under scenarios with incomplete la-

bels in streaming data, and significantly reduces the model update overhead.

3. Proposal of a unified weakly supervised AUC optimization framework WSAUC, for scenarios with inaccurate and incomplete labels: The framework transforms different types of weak supervision information into a unified form and implements a generic and robust weakly supervised AUC optimization method based on partial AUC optimization algorithms. Experiments show that the framework achieves superior AUC optimization performance under various scenarios, including those with inaccurate and incomplete labels.

4. Proposal of an AUC optimization method U^m-AUC, for label-inaccessible scenarios: This thesis proposes a method for learning AUC optimization models with sorely multiple unlabeled sample sets that have different priors, and efficiently solves the problem by transforming it into a multi-labeled AUC optimization problem. Experiments show that this method achieves superior AUC optimization performance under label-inaccessible scenarios.

KEYWORDS: Machine Learning; AUC; AUC Optimization; Weakly Labeled Learning; Semi-Supervised Learning; Weakly Supervised Learning

目　　录

主要符号表

$\boldsymbol{x}$	样本特征
y	样本标记
$\mathcal{X}$	样本集合
$\mathcal{X}_P, \mathcal{X}_N, \mathcal{X}_U$	正样本集合、负样本集合、无标记样本集合
$\mathcal{X}_{\tilde{P}}, \mathcal{X}_{\tilde{N}}$	带噪正标记样本集合、带噪负标记样本集合
$\mathcal{Y}$	标记集合
$f(\boldsymbol{x})$	模型
$\boldsymbol{w}$	模型参数
$\mathbb{I}[\cdot]$	指示函数
$\ell(z)$	损失函数
$\ell_{01}(z)$	0-1 损失函数
R_{PN}	真实 AUC 风险
$\hat{R}_{PN}$	经验 AUC 风险
$p(\boldsymbol{x})$	样本分布
$p_P(\boldsymbol{x}), p_N(\boldsymbol{x})$	正样本分布、负样本分布
$p_{\tilde{P}}(\boldsymbol{x}), p_{\tilde{N}}(\boldsymbol{x})$	带噪正标记样本分布、带噪负标记样本分布
$\{\cdots\}$	集合
$\lvert\{\cdots\}\rvert$	集合中元素的个数
$\lVert\cdot\rVert$	二范数

第一章　绪论

1.1　引言

人工智能是引领未来的战略性技术，正在深刻改变人们的生产生活方式。在2021年颁布的“十四五”规划和2035年远景目标纲要中，人工智能位列事关国家安全和发展全局的基础核心领域之首。其中，机器学习关注如何使计算机系统从数据中自主学习和提取模式以做出预测或进行决策，是人工智能的核心发展领域。国务院印发的《新一代人工智能发展规划》中，多次强调了机器学习相关理论与技术的发展要求。

在机器学习模型的构建和使用中，模型评价指标的选取至关重要。它定义了“模型应该学什么”、反映出“模型学得好不好”。可以认为，如果没有合理的评价指标，机器学习将无法开展。以分类任务为例，主流的模型评价指标包括分类准确率、AUC、F1、查准率、查全率等。在以分类准确率作为模型评价指标时，通常需要假设数据分布较为平衡；否则，分类准确率很可能会给出误导性的结果。例如，当正样本只占样本总量的1%时，一个将所有样本都识别为负类的模型可以轻易取得99%的分类正确率。而AUC（Area Under ROC Curve，ROC曲线下面积）可以衡量模型在不同阈值下对于正负样本的识别能力，其等价于模型将正负样本对正确排序的概率。无论数据分布是否平衡，AUC都能够有效地对模型辨别样本的能力进行评估。因此，在军事探测、搜索引擎、医疗诊断等重要机器学习应用中，AUC均是最主要的评价指标之一[1-4]。

由于AUC的重要意义，如何训练模型使其具有良好的AUC性能表现受到了学术界和工业界的广泛关注。为实现这一目标，“AUC优化”这一学习方法被提出，即在模型求解过程中，通过将AUC指标直接或间接作为优化目标求解，以使模型取得更好的AUC指标。通过这一手段，机器学习模型能够具有更高的排序性能，并能够避免模型在数据分布极端不平衡的情况下训练失败。目前，在该领域下的研究涉及到AUC优化的各个方面，例如批量优化方法[5-6]与在线求

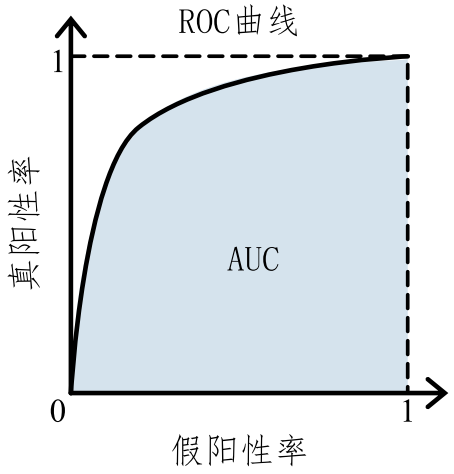


图 1-1 ROC 曲线与 AUC 指标

解方法[7-8]，AUC 优化的理论性质[9-10]，与深度学习的结合[11-12]等。同时，各种 AUC 优化方法也在多种不同的实际问题中展示出良好效果，例如搜索推荐、基因检测、软件分析等[13-17]。

尽管 AUC 优化在机器学习中具有重要地位，但该领域中的大多数研究都集中在解决有完整标注数据的情况下进行 AUC 优化。然而在实际的学习任务中，收集足够的有标注数据数据往往非常昂贵、耗时或困难。因此，有必要研究如何利用弱标记数据构建 AUC 优化模型。但由于 AUC 的数学计算依赖于样本对而非单个样本，这种形式上的区别使得针对分类正确率设计的弱标记学习方法难以直接应用到弱标记场景的 AUC 优化问题上。到目前为止，针对利用弱标记数据求解 AUC 优化问题的研究仍然匮乏。在此背景下，本文围绕弱标记 AUC 优化问题进行研究，能够在标注数据获取困难时构建具有良好 AUC 性能的模型，具有重要现实意义。

1.2 研究现状

1.2.1 AUC 评价指标

“受试者工作特征” 曲线，简称 ROC 曲线，是模型在不同阈值下的真阳性率（true positive rate, TPR）随假阳性率（false positive rate, FPR）变化的曲线，如图 1-1 所示。该方法在 “二战” 期间的 1941 年被提出，用于检测敌机的雷达信

号分析[18]。在此之后，ROC 分析又被广泛用于医疗影响识别、生物识别、灾害预测等领域。上世纪末，Spackman[3]、Provost 等人[4]的工作指出 ROC 曲线对于机器学习模型的评估的有效性，ROC 曲线进一步在机器学习领域受到关注。根据 ROC 曲线的定义，其越接近上方，就意味着模型在产生某个特定假阳性率时对应的真阳性率越高，也就表示模型性能越好。当基于 ROC 曲线对模型进行比较时，若一个模型的 ROC 曲线一直位于另一个模型的上方，则可以认为前者的性能优于后者。当两者的 ROC 曲线发生交叉，则两个模型在不同阈值下互有优劣。一般而言，通常可以通过 ROC 曲线下的面积，也就是 AUC（Area Under ROC Curve）来判断模型的性能。模型的 AUC 越大，ROC 曲线整体越靠上，则代表模型识别正样本的性能越好。

作为模型的评价指标，AUC 的一大优势是其衡量了模型无关于阈值的对样本的排序能力。这使得无论数据分布是否平衡，它总能有效地对模型性能进行衡量；而分类准确率则不然。以二分类问题为例，给定数据集 $\mathcal{D}=\{(\boldsymbol{x},y)_i\}_{i=1}^n \sim p(\boldsymbol{x},y)$，其中样本特征 $\boldsymbol{x}\in\mathcal{R}^d$，其标记 $y\in\{+1,-1\}$。我们希望构建模型 $f:\mathcal{R}^d\to\mathcal{R}$ 将样本特征映射为一个实数。若以较高的分类准确率为目标构建模型，可以以 0 作为阈值对样本 $\boldsymbol{x}$ 进行分类。其在真实分布上的泛化分类准确率可以如下定义：

$$\mathrm{ACC}=\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}\mathbb{I}[\mathrm{sign}(f(\boldsymbol{x}))=y]\,. \tag{1-1}$$

其中 $\mathbb{I}[\cdot]$ 是指示函数，当命题为真时取值为 1，否则为 0。为了最大化模型的泛化分类准确率，主流的机器学习范式包括经验风险最小化（empirical risk minimization，ERM）和结构风险最小化（structural risk minimization，SRM）。以 ERM 为例，模型求解通过最小化其在给定数据集上的经验风险实现：

$$\min_f \quad \frac{1}{n}\sum_{(\boldsymbol{x},y)\sim\mathcal{D}}\ell(yf(x))\,; \tag{1-2}$$

在 SRM 范式下，正则化项也被加入优化目标以降低过拟合的风险。在优化目标中，$\ell(\cdot)$ 是替代损失函数，在不同的模型中有不同的实现。例如对数几率回归对应 $\ell(z)=\log_2(1+\exp(-z))$，软间隔 SVM 对应 $\ell(z)=\max(0,1-z)$。

记正样本分布 $p_P(\boldsymbol{x})=p(x|y=+1)$，负样本分布 $p_N(\boldsymbol{x})=p(x|y=-1)$，正类先验概率 $\pi_p=p(y=+1)$，负类先验概率 $\pi_N=p(y=-1)$。分类准确率可以

被拆分为两个类别上准确率的加权平均：

$$\begin{aligned}\mathrm{ACC} = \pi_P \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_P(\boldsymbol{x})} \mathbb{I}[f(\boldsymbol{x}) > 0] \\ + \pi_N \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_N(\boldsymbol{x})} \mathbb{I}[f(\boldsymbol{x}) < 0]\,.\end{aligned} \tag{1-3}$$

可以看出，当类别分布不均衡，即 $\pi_P \ll \pi_N$ 或 $\pi_P \gg \pi_N$ 时，模型的分类准确率将取决于多数类，而少数类的影响较小。例如，当 $\pi_P = 0.01$，$\pi_N = 0.99$ 时，将所有样本判为负类的模型也可以取得 99% 的分类准确率，而这样的模型并不具有任何识别正样本的能力。此时，通过最小化分类准确率所习得的分类模型就极有可能发生该问题。

AUC 区别于分类准确率，在数据分布不平衡的情况下，也能判断分类模型的好坏。AUC 可以被定义为分类器对样本对排序正确的概率[1]，即一个随机采样的正样本被模型排序在一个随机采样的负样本之前的期望，可以写作：

$$\mathrm{AUC} = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_P(\boldsymbol{x})} \mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_N(\boldsymbol{x}')} [\mathbb{I}[f(\boldsymbol{x}) > f(\boldsymbol{x}')]]\,. \tag{1-4}$$

对于上述将所有样本判为负类的模型，如果其不能正确地对于正样本输出较高的得分，依然会具有较低的 AUC 指标。可以看出，AUC 刻画了分类器无关阈值的样本识别能力，能够更好地表征模型的整体性能。无论数据分布是否平衡，AUC 都能有效地对分类器的性能作出判断。

1.2.2 AUC 优化

为了构建具有较高 AUC 性能的模型，最直接的方式是将 AUC 作为直接或间接的优化目标进行学习[2]。根据上述定义，学习算法可以通过最小化基于样本对的损失函数进行模型求解。最早期在该学习范式上作出探索的工作包括 2003 年提出的 SVM$^{\mathrm{rank}}$[19]，RankBoost[5]等。这些工作基本等价于通过不同的算法最小化数据集上的样本对排序风险，或经验 AUC 风险：

$$\min_f \quad \frac{1}{n_P n_N} \sum_{\boldsymbol{x}\in\mathcal{X}_P} \sum_{\boldsymbol{x}'\in\mathcal{X}_N} [\ell(f(\boldsymbol{x}) - f(\boldsymbol{x}'))]\,, \tag{1-5}$$

其中 $\mathcal{X}_P$ 与 $\mathcal{X}_N$ 分别是正样本与负样本的集合，其大小分别为 n_P 和 n_N。这种学习方法即对应了以 AUC 为导向的 ERM 或 SRM 范式。同期，Cortes 等人[20]在 NIPS 2003 发表的论文中正式提出了 AUC 优化（AUC optimization）这一术语，并从理论上指出了优化分类准确率并不一定导致模型的 AUC 性能提升，这一学习范式逐渐受到更加广泛的关注。2004 年，Herschtal 等人[7]提出基于梯度下降的 AUC 优化算法；2005 年，Joachims[6]提出可以优化 AUC 的 SVM 变体 SVM$^{\text{perf}}$；2007 年，Calders 等人[21]提出了高效的近似 AUC 优化算法。在理论方面，Agarwal 等人[10]针对 AUC 优化问题的泛化误差界进行了研究。Gao 等人[9]对 AUC 优化问题中常用损失函数的渐进一致性作出了分析，给出了 AUC 优化中替代损失函数具有一致性的充分条件，并指出指数替代损失、平方替代损失等损失函数具有 AUC 一致性。

关于 AUC 优化算法与理论的早期工作为后续近 20 年间大量的后续研究打下了良好的基础。在设计高效的 AUC 优化算法方面，一个难点在于 AUC 的损失函数必须基于样本对计算，而难以分解到单个样本计算。不仅在利用梯度下降或随机梯度下降等方法进行求解时，需要面临平方级别的样本对带来的计算开销；针对流式数据在线更新模型时，由于难以保存全量历史样本，也无法利用当前样本与其他所有样本配对计算损失。为了解决这一问题，Zhao 等人[22]提出利用一个较小的样本缓存对历史样本以均等概率保留以实现在线学习场景下的 AUC 优化。Gao 等人[23]通过在线维护样本的协方差矩阵，实现了在使用平方替代损失函数的情况下基于单样本的在线 AUC 优化。此后，Ying 等人[8]通过将 AUC 优化问题转化为鞍点问题，同样在使用平方替代损失的情况下实现了基于单个样本通过随机原始-对偶算法求解。基于该问题转化形式，后续工作进一步提升了算法的计算效率与收敛率。Natole 等人[24]使用非光滑正则化项将在线 AUC 优化的收敛率提升到 $O(\log t/t)$。Lei 等人[25]提出的算法在不使用正则化项或使用强凸正则化项的情况下将收敛率进一步提升至 $O(1/t)$。在与深度模型的结合方面，Liu 等人[11]首次提出了针对 AUC 优化的深度神经网络优化算法与优化器。Yuan 等人[13]指出在利用深度神经网络进行 AUC 优化时，基于间隔的损失相较于平方替代损失更加稳健，优化效果更好，并提出了 PESG 优化算法。Yuan 等人[12]使用 AUC 损失与交叉熵损失相结合的方法缓解了通过深度神经网络直接优化 AUC 损失效果不佳的问题。

此外，一些研究指出部分 AUC（Partial AUC）在特定问题上具有较好的效果，并将一些 AUC 优化算法进行扩展以优化部分 AUC。Narasimhan 等人[26]基于自己选择的方法提出了单向部分 AUC 的优化算法。Yang 等人[27]提出了双向部分 AUC 的定义，并给出了双向部分 AUC 的非参数化估计方法。Yang 等人[28]提出了一种端到端的双向部分 AUC 双层优化方法。Zhu 等人[29]基于分布稳健优化（DRO）损失的形式提出了新的部分 AUC 的优化方法。

基于这些研究成果，AUC 优化算法在各种实际应用任务中取得了成功，例如医疗影像分析[13,30]、基因检测[14]、软件挖掘[15]、语音信号处理与识别[16]、监控视频异常行为检测[17]等。

1.2.3 弱标记 AUC 优化

以上提到的关于 AUC 优化的研究，大都假设能够获取充足的标记数据进行模型学习。然而，在许多现实应用中，完整的标注往往很难获取，人们不得不考虑如何从较弱的标注信息中进行学习。例如，在互联网文本分类任务中，由于互联网产生的数据量巨大，无法对所有数据进行标注，只能标注其中的一小部分。模型需要合理地对无标记数据进行利用，才能有效地进行学习。在医疗影像诊断中，不仅由于专业医生的数量与时间有限，难以对所有影像进行标注，也因为标注难度较大而经常出现标记错误的情况。此时，机器学习算法还需要考虑标记不准确带来的影响。在针对不同地区分析就业率、生育率等宏观指标时，由于隐私及规模问题甚至有可能难以获得个体样本标记，仅能通过地区整体的统计数据进行学习。在这些标记信息逐渐减弱的场景下，以往的 AUC 优化方法难以适用，如何利用弱标记数据进行 AUC 优化成为了亟待研究的问题。

在弱标记 AUC 优化这一方向上，已有一些研究人员开始作出尝试，但目前仍然面临较大的挑战。其原因在于传统利用弱标记进行学习的范式，如半监督学习、正标记-无标记学习等往往需要依赖对于分布的特定假设以利用无标记样本辅助模型学习，而针对样本对的分布提出分布假设十分困难。目前，学术界对于利用弱标记数据构建 AUC 优化模型的探索性研究寥寥无几：SSRankBoost[31]和 SSLROC[32]分别借助 RankBoost[5]和 TSVM[33]的思想，尝试基于临近样本或样本损失的大小为无标记样本产生伪标记以加以利用。PNU-AUC[34]则从正标记-无标记学习的角度对损失函数进行补偿以进行 AUC 风险估计。这些研究工作主要

借鉴针对分类准确率设计的弱标记学习算法的思想，未能对 AUC 优化任务有针对性地设计算法以利用弱标记数据进行学习。当面临标记信息进一步减弱的场景，如标记存在不准确问题，甚至不可见时，目前尚未有研究对如何在这些场景下有效地进行 AUC 优化进行探讨。

1.3 有待研究的问题

如上节中所阐述，AUC 优化作为一类重要的学习方法，可以避免数据分布不平衡所产生的负面影响，并使模型具有更好的样本识别与排序能力，在众多应用领域中已经取得成功。然而，在针对标记信息不同程度弱化的学习场景下如何利用弱标记数据构建 AUC 优化模型这一方面，仍存在许多问题亟待研究。具体而言，有以下几个方面：

1. 在标记不完全可见的场景下，如何有效地进行 AUC 优化模型的构建？不同于面向准确率的学习算法，AUC 风险需要基于样本对定义，难以针对样本对的分布提出假设对无标记数据加以利用。如何针对 AUC 优化问题设计专用的方法仍然有待研究。
2. 如何利用标记不完全可见的流式数据，在线地进行 AUC 优化模型的构建？由于 AUC 损失难以分解到单个样本计算，在无法存储或反复扫描历史数据的流式数据学习场景中，难以利用当前样本与所有历史样本配对计算损失。此外，在数据流式场景下，由于无法保留数据对分布进行估计，也对无标记数据的利用带来了困难。
3. 当标记信息进一步减弱，数据标记不准确且不完全可见的情况下，如何有效地进行 AUC 优化模型的构建？标记信息不准确为模型构建带来了额外的风险，若不对其进行处理，会导致模型性能的下降。当标记信息不准确和不完全可见的问题同时发生，需要综合进行考虑，从微弱的标记信息中进行稳健的 AUC 优化，否则会对模型性能带来极大的负面影响。
4. 在标记信息继续减弱，直至不可见的情况下，能否构建 AUC 优化模型？至少需要何种监督信息才能完成 AUC 优化模型的学习？在缺乏样本标记的情况下，难以建立样本到标记的映射，算法只能通过微弱的监督信息进行学习。在此种情况下设计 AUC 优化算法十分困难。

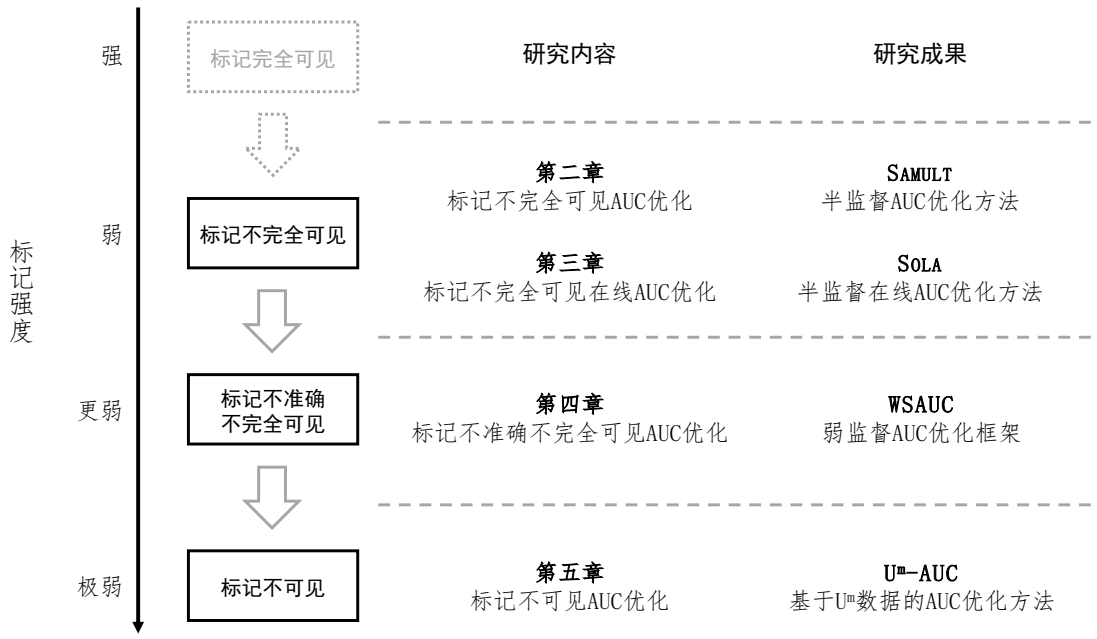


图 1-2 本文研究内容

1.4 本文工作

本文针对如何在不同程度的弱标记场景下构建 AUC 优化模型进行了系统性的研究。针对上一节中提出的四个问题，主要完成四项工作，分别对应于本文的第二章到第五章，如图 1-2 所示。

第二章针对标记不完全可见的场景，提出了半监督 AUC 优化方法 SAMULT。该方法利用 AUC 成对损失的特殊性质，实现了在无需知道先验概率的情况下即可通过无标记数据进行无偏 AUC 风险估计，进而可以在无需猜测无标记样本伪标记的情况下利用无标数据。该方法摆脱了对特定分布假设的依赖，避免了在分布假设不成立时引入的风险，在多个任务上实现了更好的 AUC 优化性能。

第三章针对标记不完全可见的流式数据场景，提出了半监督在线 AUC 优化方法 SOLA。该方法通过将第二章提出的风险最小化问题转化为随机鞍点问题，实现了基于单个样本进行半监督 AUC 损失梯度更新的方法，解决了在线半监督 AUC 优化由于无法基于样本对计算风险而难以实现的困难。该方法首次为面临数据流式产生、分布不平衡且标记不完全可见的学习任务提供了解决方案，并在软件持续构建预测任务中展现出优秀的学习性能和极高的运行效率。

第四章针对标记不准确且不完全可见的场景，提出了弱监督 AUC 优化框架 WSAUC，将从多种不同弱标记数据进行 AUC 优化进行了综合考虑。该框架将

不同的弱监督信息转化为标记混杂的统一形式，并基于一种新型的部分 AUC 优化方法实现了通用的稳健弱监督 AUC 优化方法。该框架首次对多种弱监督信息进行综合研究，实现了利用标记不准确且不完全可见场景下的 AUC 优化，并在多种弱监督场景下取得了良好的 AUC 优化性能。

第五章针对标记不可见的场景，提出了利用多个无标记集合进行 AUC 优化的方法 U^m-AUC。该方法仅依赖二或多个具有不同先验的无标记样本集合及其先验大小次序的知识，将原学习问题转化成为一个多标记场景下的宏平均 AUC（Marco AUC）优化问题，实现了高效求解。借助于 AUC 优化的特殊性，该方法无需依赖对数据集类别先验的知识，比其他针对类似场景的学习方法所需监督信息更少、更符合实际，并取得了良好的学习效果。

最后，**第六章**总结本文汇报的研究工作及贡献，并对未来可能的相关研究方向作出讨论。

第二章　标记不完全可见 AUC 优化

2.1　引言

在许多实际机器学习应用中，收集大量无标记数据相对容易，而为这些数据进行标注却需要非常多的人力和专业知识，难以大量获取。例如，在构建图像分类模型时，可以很容易地收集到互联网上的大量图片，但是为每一张图片进行标注的开销将会难以承担。在进行社交媒体情感分析时，由于用户生成内容数据量大且实时性高，也难以将其全部标注。在这类应用中，机器学习算法需要应对数据标记不完全可见的问题，即只有少量标注数据可用，其余大量的数据没有标注。这是从弱标记数据进行学习的一种典型场景。

为了在标记不完全可见的场景下进行有效学习，通常使用半监督学习方法，以利用无标记数据辅助模型学习。由于标注不完全可见的学习场景的普遍性，半监督学习自提出以来一直以来受到研究者的广泛关注[35-36]，并有许多半监督学习方法已被提出。例如基于生成式模型的半监督学习方法[37-39]、基于低密度间隔的半监督学习方法[33,40-41]、基于图的半监督学习方法[42-44]、基于分歧的方法半监督学习方法等[45-50]。此外，近年来还有许多针对深度学习模型提出的半监督学习方法[51-61]。为了有效利用无标记数据，几乎所有这些方法都会基于一定的分布假设（例如聚类假设，流形假设等）来建立标记数据和无标记数据之间的联系，并通过显式或隐式地估计无标记样本的伪标记来构建学习器。

AUC（ROC 曲线下的面积）是评估学习器性能的广泛使用指标[1]，尤其是在数据分布出现某种不平衡时。在本文第一章中，我们已经概述了 AUC 作为评估指标的重要性。许多研究都详细阐述了如何在有监督场景高效地优化 AUC[7-9,23]，但半监督场景下的 AUC 优化研究较为稀缺。这是因为 AUC 风险基于样本对定义，而针对样本对的分布提出分布假设十分困难。基于各种样本分布假设的半监督方法难以与基于样本对的 AUC 风险计算相结合，也无法直接推广到 AUC 优化任务中使用。

为解决该问题，本章节提出一种利用无标记数据进行 AUC 风险无偏估计的方法。该方法不必根据任何分布假设猜测无标记数据的伪标记，也无需知道先验概率，具有较高的易用性。根据这一风险估计方法，通过将无标记的数据同时视为正例和负例，即可解决半监督 AUC 优化问题。基于这一理论发现，本章节提出了两种新的半监督 AUC 优化方法：一种简单地将无标记数据视为正样本和负样本数据的无偏经验风险最小化方法 SAMULT（Semi-supervised AUC Maximization by treating the UnLabeled data in Two ways），以及将无标记数据随机划分为伪正例和伪负例集来训练基分类器的集成学习方法 SAMPURA（Semi-supervised AUC Maximization by Partitioning Unlabeled data at RAndom）。

本章节提出的方法不必依赖特定分布形态的假设，避免了显式或隐式地基于这些假设进行伪标记估计的方法[31,39]可能出现的偏差导致表现不佳甚至性能下降[62]；也无需早期方法[34]依赖的类别先验等额外知识。实验结果表明，本文提出的方法在性能上优于多个对比方法。此外，该方法还可以通过将所有无标记样本视为负例，解决正标记-无标记数据上的 AUC 优化问题。

2.2 半监督 AUC 优化方法 SAMULT

2.2.1 无偏半监督 AUC 风险估计

为引出半监督 AUC 优化问题，本节首先简述有监督场景下的 AUC 优化问题。在有监督学习中，以二分类问题为例，通常有包含正样本和负样本的有标记数据集用于训练。其中正样本和负样本的集合可以如下表示：

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = +1)\,, \tag{2-1}$$

$$\mathcal{X}_N := \{\boldsymbol{x}'_j\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_N(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = -1)\,. \tag{2-2}$$

为简便起见，本章节中主要讨论线性模型 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$。非线性分类也可以通过对特征进行非线性映射实现。由于 AUC 等价于一个随机采样的正样本被模型排序在一个随机负样本之前的概率[1]，其可以定义为：

$$\mathrm{AUC} = 1 - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))\right]\right]. \tag{2-3}$$

其中，01 损失函数 ℓ_{01} 定义如下：

$$\ell_{01}(z) = \begin{cases} 1, & \text{if } z < 0 \\ 1/2, & \text{if } z = 0 \\ 0, & \text{if } z \geqslant 0 \end{cases} . \tag{2-4}$$

最大化上述 AUC 等价于最小化如下 AUC 风险。为了避免歧义，本文用 PN-AUC 风险代指有监督情况下的 AUC 风险，即通过有标记的正样本（P）与负样本（N）计算的风险：

$$R_{PN} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))\right]\right]. \tag{2-5}$$

在半监督场景下，需要利用无标记样本集合进行训练。虽然无法得知这些无标记样本（U）的标记，但是它们要么是正样本，要么是负样本。因此，无标记样本可以被认为是从正样本分布和负样本分布的混合分布中采样出来的：

$$\mathcal{X}_U := \{\boldsymbol{x}''_k\}_{k=1}^{n_U} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) := \theta_P p_P(\boldsymbol{x}) + \theta_N p_N(\boldsymbol{x}), \tag{2-6}$$

其中 θ_P 和 θ_N 分别是混合分布中正类和负类的先验概率。

基于上述无标记数据分布的定义，本文将证明以下结论：将无标记样本视为负样本，并与正样本进行配对所估计的风险；或将无标记样本视为正样本，与负样本进行配对所估计的风险，在优化时与无偏风险是等价的。定理 2.1 给出了该结论形式化的叙述。

定理 2.1 *通过正样本和视为负样本的无标记样本来估计的 PU-AUC 风险 R_{PU}，以及通过负样本和视为正样本的无标记样本来估计的 UN-AUC 风险 R_{UN}，经线性变换后可等价于有监督 PN-AUC 风险 R_{PN}。两种利用无标记样本的风险 R_{PU} 和 R_{UN} 的定义如下：*

$$R_{PU} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_U}\left[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}''))\right]\right], \tag{2-7}$$

$$R_{UN} = \mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_U}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}''-\boldsymbol{x}'))\right]\right]. \tag{2-8}$$

证明 由于期望的线性性，下式成立：

$$\begin{aligned}
R_{PU} =& \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_U}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}''))]] \\
=& \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\theta_P \mathop{\mathbb{E}}_{\bar{\boldsymbol{x}}\in\mathcal{X}_P}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\bar{\boldsymbol{x}}))] \\
&+ \theta_N \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))]] \\
=& \frac{1}{2}\theta_P + \theta_N \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))]] .
\end{aligned}$$

可以进行第二步推导的原因在于，$\mathcal{X}_P \times \mathcal{X}_P$ 上的一对样本来自于相同分布，将其中任意一个样本排在前面的概率是等同的，其产生的期望风险对称，故第一项等于常数 *1/2*；而第二项等价于 *PN-AUC* 风险乘以系数 θ_N。因此：

$$R_{PU} = \theta_N R_{PN} + \frac{1}{2}\theta_P . \tag{2-9}$$

类似地，可证明下式成立：

$$R_{UN} = \theta_P R_{PN} + \frac{1}{2}\theta_N . \tag{2-10}$$

综上，可得 R_{PU} 和 R_{UN} 与有监督 *AUC* 风险 R_{PN} 经线性变换后等价。由于 θ_P 和 θ_N 分别为正负类的先验概率，满足 $\theta_P > 0$ 和 $\theta_N > 0$。因此当 R_{PU} 或 R_{UN} 取到最小值时，R_{PN} 也同时取到最小值。 □

直观地讲，PU-AUC 风险 R_{PU} 可以被看作是两部分的加权平均值：一部分是有标记的正样本和无标记样本中负样本排序构成的 PN-AUC 风险，另一部分是由有标记的正样本和无标记样本中正样本排序构成的 $\mathcal{X}_P \times \mathcal{X}_P$ 上的 AUC 风险。由于有标记正样本和无标记样本中的正样本来自同一分布，因此一者排在另外一者前面的概率总是 $1/2$。

根据定理 2.1，可以得知优化 R_{PU} 或者 R_{UN} 渐近等价于优化有监督 AUC 风险。因此无标记的数据可以被简单地视为正样本或负样本，即可用于半监督 AUC 优化模型的构建。

由于 $\theta_P + \theta_N = 1$ 恒成立，我们可以进一步将公式 2-9 与公式 2-10 求和得

到下列等式：

$$R_{PU} + R_{UN} - \frac{1}{2} = R_{PN}\,. \tag{2-11}$$

公式 2-11 给出一个有趣的结论：当我们有正例、负例和无标记数据时，可以在不知道类别先验概率 θ_P 和 θ_N 的情况下修正估计量的偏差，实现无偏的 AUC 风险估计。发生这种情况的原因是 AUC 具有对不平衡分布的稳健性；这一结论只在 AUC 优化中成立，而面向分类准确率的半监督学习无法享有这一特性。受到这个理论的启发，本章节提出一种名为 SAMULT 的机器学习方法，并进一步将其扩展为一种集成学习方法 SAMPURA。

2.2.2 无需猜测标记的半监督 AUC 优化方法

上一节中，推导出将 PU-AUC 风险与 PN-AUC 风险相结合可以对有监督 AUC 风险进行估计。根据公式 2-11 ，可以看出即使在不知道无标记数据正负样本的先验概率的情况下，风险估计量的偏差是一个常数，恒为 $1/2$。在这种情况下，风险估计量可以被简单地修正成为无偏估计量。

基于上一节中阐述的结论，本节介绍两种半监督的 AUC 优化方法：SAMULT 及其集成学习扩展版本 SAMPURA。SAMULT 将无标记的数据视为正类和负类数据来计算基于公式 2-11 的风险估计，然后将其与监督 AUC 风险估计量结合，得到一个无偏的 AUC 风险估计量。模型的求解则可以通过最小化经验风险或结构风险实现。此外，由于无标记数据既可以被视为正样本，也可以被视为负样本，我们可以对无标记数据进行多种划分，将它们标记为正类或负类，然后训练多个不同的分类器。基于这个想法，本文进一步将 SAMULT 扩展为一个集成学习版本，即 SAMPURA。它首先通过将无标记的数据随机分成伪正类和伪负类数据集来增加数据，然后通过集成学习得到一个强分类器。

SAMULT 方法 基于公式 2-11 ，本文提出以下使用半监督 AUC 优化中的所有标记和无标记数据的 AUC 风险估计量：

$$\hat{R}_{PNU} = \gamma\hat{R}_{PN} + (1-\gamma)(\hat{R}_{PU} + \hat{R}_{UN} - \frac{1}{2})\,, \tag{2-12}$$

其中 $\gamma \in [0,1]$ 是有监督风险与半监督风险的折中参数，

$$\hat{R}_{PN} = \frac{1}{n_P n_N} \sum_{\boldsymbol{x} \in \mathcal{X}_P} \sum_{\boldsymbol{x}' \in \mathcal{X}_N} \ell(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}')), \tag{2-13}$$

$$\hat{R}_{PU} = \frac{1}{n_P n_U} \sum_{\boldsymbol{x} \in \mathcal{X}_P} \sum_{\boldsymbol{x}'' \in \mathcal{X}_U} \ell(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}'')), \tag{2-14}$$

$$\hat{R}_{UN} = \frac{1}{n_U n_N} \sum_{\boldsymbol{x}'' \in \mathcal{X}_U} \sum_{\boldsymbol{x}' \in \mathcal{X}_N} \ell(\boldsymbol{w}^\top (\boldsymbol{x}'' - \boldsymbol{x}')). \tag{2-15}$$

由于 0-1 损失函数非光滑连续，难以优化，在实践中，本文此处采用平方损失 $\ell(z) = (1-z)^2$ 替代 0-1 损失。目前，已有现有研究证明使用平方替代损失函数在渐进意义下与 AUC 一致[9]。

为了降低过拟合的风险，此处采用结构风险最小化的学习范式，即在优化过程中同时引入 ℓ_2 正则化项：

$$\min_{\boldsymbol{w}} \quad \hat{R}_{PNU}(\boldsymbol{w}) + \lambda ||\boldsymbol{w}||^2, \tag{2-16}$$

其中参数 $\lambda \geqslant 0$ 是正则化项的权重系数。

由于公式 2-12 中的常数项不会改变该优化问题的最优解，在求解时可以忽略。优化问题公式 2-16 的解析解计算方法如下：

$$\begin{aligned} \hat{\boldsymbol{w}} = &(\gamma \boldsymbol{H}_{PN} + (1-\gamma)(\boldsymbol{H}_{PU} + \boldsymbol{H}_{UN}) + \lambda \boldsymbol{I}_d)^{-1} \\ &(\gamma \boldsymbol{h}_{PN} + (1-\gamma)(\boldsymbol{h}_{PU} + \boldsymbol{h}_{UN})) , \end{aligned} \tag{2-17}$$

其中，

$$\boldsymbol{h}_{PN} = \frac{1}{n_P} \boldsymbol{X}_P^\top \mathbf{1}_{n_P} - \frac{1}{n_N} \boldsymbol{X}_N^\top \mathbf{1}_{n_N}, \tag{2-18}$$

$$\boldsymbol{h}_{PU} = \frac{1}{n_P} \boldsymbol{X}_P^\top \mathbf{1}_{n_P} - \frac{1}{n_U} \boldsymbol{X}_U^\top \mathbf{1}_{n_U}, \tag{2-19}$$

$$\boldsymbol{h}_{UN} = \frac{1}{n_U} \boldsymbol{X}_U^\top \mathbf{1}_{n_U} - \frac{1}{n_N} \boldsymbol{X}_N^\top \mathbf{1}_{n_N}, \tag{2-20}$$

$$\begin{aligned} \boldsymbol{H}_{PN} = &\frac{1}{n_P} \boldsymbol{X}_P^\top \boldsymbol{X}_P - \frac{1}{n_P n_N} \boldsymbol{X}_P^\top \mathbf{1}_{n_P} \mathbf{1}_{n_N}^\top \boldsymbol{X}_N \\ &- \frac{1}{n_P n_N} \boldsymbol{X}_N^\top \mathbf{1}_{n_N} \mathbf{1}_{n_P}^\top \boldsymbol{X}_P + \frac{1}{n_N} \boldsymbol{X}_N^\top \boldsymbol{X}_N, \end{aligned} \tag{2-21}$$

算法 1 SAMULT

Input: $\boldsymbol{X}_P, \boldsymbol{X}_N, \boldsymbol{X}_U, \lambda, \gamma$

1: 根据公式 2-18 至公式 2-23 计算 $\boldsymbol{h}_{PN}, \boldsymbol{h}_{PU}, \boldsymbol{h}_{UN}, \boldsymbol{H}_{PN}, \boldsymbol{H}_{PU}, \boldsymbol{H}_{UN}$

2: 根据公式 2-17 计算模型闭式解 $\hat{\boldsymbol{w}}$

Output: 模型参数 $\hat{\boldsymbol{w}}$

$$\begin{aligned}\boldsymbol{H}_{PU} =& \frac{1}{n_P}\boldsymbol{X}_P^\top \boldsymbol{X}_P - \frac{1}{n_P n_U}\boldsymbol{X}_P^\top \mathbf{1}_{n_P}\mathbf{1}_{n_U}^\top \boldsymbol{X}_U \\ & - \frac{1}{n_P n_U}\boldsymbol{X}_U^\top \mathbf{1}_{n_U}\mathbf{1}_{n_P}^\top \boldsymbol{X}_P + \frac{1}{n_U}\boldsymbol{X}_U^\top \boldsymbol{X}_U\,, \end{aligned} \tag{2-22}$$

$$\begin{aligned}\boldsymbol{H}_{UN} =& \frac{1}{n_U}\boldsymbol{X}_U^\top \boldsymbol{X}_U - \frac{1}{n_U n_N}\boldsymbol{X}_U^\top \mathbf{1}_{n_U}\mathbf{1}_{n_N}^\top \boldsymbol{X}_N \\ & - \frac{1}{n_U n_N}\boldsymbol{X}_N^\top \mathbf{1}_{n_N}\mathbf{1}_{n_U}^\top \boldsymbol{X}_U + \frac{1}{n_N}\boldsymbol{X}_N^\top \boldsymbol{X}_N\,, \end{aligned} \tag{2-23}$$

$\boldsymbol{X}_P$，$\boldsymbol{X}_N$ 和 $\boldsymbol{X}_U$ 分别是正、负、无标记样本矩阵，$\mathbf{1}_d$ 是 d 维全 1 向量，$\boldsymbol{I}_d$ 是 d 维单位矩阵。为了避免矩阵求逆的计算复杂度达到 $O(d^3)$，计算时可以采用 Sherman-Morrison 公式来降低计算成本。

值得注意的是，当只有正例和无标记数据可用时，$\hat{R}_{PN}$ 和 $\hat{R}_{UN}$ 均为零，SAMULT 会退化为一种特殊形式，其中只优化 $\hat{R}_{PU}$。忽略参数 λ 和偏差项，SAMULT$^{\text{P+U}}$ 的优化目标等价于：

$$\min_{\boldsymbol{w}} \quad \hat{R}_{PU}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||^2\,. \tag{2-24}$$

该式与将无标记数据视为负例时的有监督 AUC 优化相同。这表明可以通过将无标记数据视为负例来将 PU AUC 优化转化为有监督 AUC 优化来求解。

简而言之，SAMULT 优化的是（无偏）有监督 AUC 风险估计量 $\hat{R}_{PN}$ 和无偏半监督 AUC 风险估计量 $(\hat{R}_{PU} + \hat{R}_{UN} - \frac{1}{2})$ 的加权平均。SAMULT 不需要依赖于为无标记数据产生伪标记，也不需要估计类先验概率以重新加权无标记数据，因此只需几行代码即可实现。算法 1 显示了 SAMULT 的流程。需要注意的是，由于 AUC 仅代表学习器的排名质量，SAMULT 并不确定决策边界。若要生成样本的最终分类，可以采用一些基于排序列表的独立阈值确定策略[63-64]。

SAMPURA 方法 由于最小化 PU 和 UN 样本对的损失有助于学习分类器，一个直观的想法是使用无标记的数据来扩充正负数据。基于该思路，可以将无标记的数据 $\mathcal{X}_U$ 平均分成伪正样本集 $\mathcal{X}_{U^+}$ 和伪负样本集 $\mathcal{X}_{U^-}$，分别扩充原始的正负标

记数据。通过假设 $\mathcal{X}_{P'} = \mathcal{X}_P \cup \mathcal{X}_{U^+}$ 中的样本应该排在 $\mathcal{X}_N$ 中的样本之前，以及 $\mathcal{X}_P$ 中的样本应该排在 $\mathcal{X}_{N'} = \mathcal{X}_N \cup \mathcal{X}_{U^-}$ 中的样本之前，可以定义无偏风险估计量如下：

$$\begin{aligned}\hat{R}_{PNU} =& \hat{R}_{P'N} + \hat{R}_{PN'} - \frac{1}{2} \\ =& \frac{1}{n_{P'} n_N} \sum_{\boldsymbol{x} \in \mathcal{X}_{P'}} \sum_{\boldsymbol{x}' \in \mathcal{X}_N} \ell(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}')) \\ &+ \frac{1}{n_P n_{N'}} \sum_{\boldsymbol{x} \in \mathcal{X}_P} \sum_{\boldsymbol{x}' \in \mathcal{X}_{N'}} \ell(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}')) - \frac{1}{2}.\end{aligned} \tag{2-25}$$

利用无标记数据的不同划分，可以训练多个分类器，并通过集成学习获得一个强分类器，以进一步降低风险估计的方差，使得学习更加稳定。在每个划分中，我们最小化带有 ℓ_2 正则化项的结构风险，以获得基分类器：

$$\min_{\boldsymbol{w}} \quad \hat{R}_{P'N} + \hat{R}_{PN'} + \lambda ||\boldsymbol{w}||^2, \tag{2-26}$$

然后对这些基分类器的权重 $\boldsymbol{w}$ 取平均值来构建最终的集成分类器。

基分类器的解析解可以如下计算：

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{H}_{P'N} + \boldsymbol{H}_{PN'} + \lambda \boldsymbol{I}_d\right)^{-1} \left(\boldsymbol{h}_{P'N} + \boldsymbol{h}_{PN'}\right), \tag{2-27}$$

其中的矩阵定义如下：

$$\boldsymbol{h}_{P'N} = \frac{1}{n_{P'}} \boldsymbol{X}_{P'}^\top \mathbf{1}_{n_{P'}} - \frac{1}{n_N} \boldsymbol{X}_N^\top \mathbf{1}_{n_N}, \tag{2-28}$$

$$\boldsymbol{h}_{PN'} = \frac{1}{n_P} \boldsymbol{X}_P^\top \mathbf{1}_{n_P} - \frac{1}{n_{N'}} \boldsymbol{X}_{N'}^\top \mathbf{1}_{n_{N'}}, \tag{2-29}$$

$$\begin{aligned}\boldsymbol{H}_{P'N} =& \frac{1}{n_{P'}} \boldsymbol{X}_{P'}^\top \boldsymbol{X}_{P'} - \frac{1}{n_{P'} n_N} \boldsymbol{X}_{P'}^\top \mathbf{1}_{n_{P'}} \mathbf{1}_{n_N}^\top \boldsymbol{X}_N \\ &- \frac{1}{n_{P'} n_N} \boldsymbol{X}_N^\top \mathbf{1}_{n_N} \mathbf{1}_{n_{P'}}^\top \boldsymbol{X}_{P'} + \frac{1}{n_N} \boldsymbol{X}_N^\top \boldsymbol{X}_N,\end{aligned} \tag{2-30}$$

$$\begin{aligned}\boldsymbol{H}_{PN'} =& \frac{1}{n_P} \boldsymbol{X}_P^\top \boldsymbol{X}_P - \frac{1}{n_P n_{N'}} \boldsymbol{X}_P^\top \mathbf{1}_{n_P} \mathbf{1}_{n_{N'}}^\top \boldsymbol{X}_N \\ &- \frac{1}{n_P n_{N'}} \boldsymbol{X}_{N'}^\top \mathbf{1}_{n_{N'}} \mathbf{1}_{n_P}^\top \boldsymbol{X}_P + \frac{1}{n_{N'}} \boldsymbol{X}_{N'}^\top \boldsymbol{X}_{N'},\end{aligned} \tag{2-31}$$

$\boldsymbol{X}_{P'}$ 和 $\boldsymbol{X}_{N'}$ 分别是 $\mathcal{X}_{P'}$ 和 $\mathcal{X}_{N'}$ 的样本特征矩阵。

算法 2 SAMPURA

Input: $\boldsymbol{X}_P, \boldsymbol{X}_N, \boldsymbol{X}_U, \lambda, T$

1: **for** $t = 1 \to T$ **do**
2: 将无标记数据 $\boldsymbol{X}_U$ 随机划分成 $\boldsymbol{X}_{U^+}$ 和 $\boldsymbol{X}_{U^-}$
3: 令 $\boldsymbol{X}_{P'} = [\boldsymbol{X}_P^\top \mid \boldsymbol{X}_{U^+}^\top]^\top, \boldsymbol{X}_{N'} = [\boldsymbol{X}_N^\top \mid \boldsymbol{X}_{U^-}^\top]^\top$
4: 根据公式 2-28 至公式 2-31 计算 $\boldsymbol{h}_{P'N}, \boldsymbol{h}_{PN'}\ \boldsymbol{H}_{P'N}, \boldsymbol{H}_{PN'}$
5: 根据公式 2-27 计算模型闭式解 $\hat{\boldsymbol{w}}^{(t)}$
6: **end for**

Output: 集成分类器 $\hat{\boldsymbol{w}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{w}}^{(t)}$

SAMPURA 的流程如算法 2 所示。可以看出，SAMPURA 只在 SAMULT 上进行了简单改动即可实现集成学习的效果。

2.3 理论分析

本节以 SAMULT 方法为例进行理论分析，证明其超额风险界和方差缩减效果。考虑 K 是 $\mathcal{X}^2$ 上的一个核函数，C_w 是一个正实数，令 $\mathcal{F}_K$ 表示函数类：

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

其中 $\|x\|_K = \sqrt{K(x, x)}$。

本节中假设替代损失函数 ℓ 是 L-Lipschitz 连续的，上界为正实数 C_ℓ，并满足不等式 $\ell \geqslant \ell_{01}$。例如，平方损失和指数损失即可满足这些条件。

2.3.1 超额风险

记 $\hat{f}^*_{PNU}$ 为令经验风险 $\hat{R}_{PNU}(f)$ 取得最小值的模型，如下定理可以表明 $\hat{f}^*_{PNU}$ 的风险收敛于函数族 $\mathcal{F}_K$ 中最优模型的风险。

定理 2.2 (超额风险) *设 $\hat{f}^*_{PNU} \in \mathcal{F}_K$ 是令经验风险 $\hat{R}_{PNU}(f)$ 最小的分类器，$f^*_{PN} \in \mathcal{F}_K$ 是令真实风险 $R_{PN}(f)$ 最小的分类器。对于任意 $\delta > 0$，至少以 $1-\delta$ 的概率下式成立：*

$$\begin{aligned} &R_{PN}(\hat{f}^*_{PNU}) - R_{PN}(f^*_{PN}) \\ &\quad \leqslant h(\frac{\delta}{3})\left(\gamma\sqrt{\frac{n_P+n_N}{n_P n_N}} + (1-\gamma)\left(\sqrt{\frac{n_P+n_U}{n_P n_U}} + \sqrt{\frac{n_U+n_N}{n_U n_N}}\right)\right), \end{aligned} \tag{2-32}$$

其中

$$h(\delta)=8\sqrt{2}C_\ell C_w C_x+5\sqrt{2\ln(2/\delta)},$$

n_P, n_N 分别是正、负样本集的大小。

证明 设 $\hat{R}_{PNU}(f)$ 表示 $R_{PNU}(f)$ 的经验估计量。优化 $\hat{R}_{PNU}(f)$ 的超额风险可以表示为

$$\begin{aligned}
&R_{PN}(\hat{f}^*_{PNU})-R_{PN}(f^*_{PN})\\
&\quad=R_{PN}(\hat{f}^*_{PNU})-\hat{R}_{PNU}(\hat{f}^*_{PNU})+\hat{R}_{PNU}(\hat{f}^*_{PNU})\\
&\qquad-\hat{R}_{PNU}(f^*_{PN})+\hat{R}_{PNU}(f^*_{PN})-R_{PN}(f^*_{PN})\\
&\quad\leqslant 2\max_{f\in\mathcal{F}}|\hat{R}_{PNU}(f)-R_{PN}(f)|.
\end{aligned}\tag{2-33}$$

根据公式 *2-11* ，右侧项可以表示为

$$\max_{f\in\mathcal{F}}|\hat{R}_{PNU}(f)-R_{PN}(f)|=\max_{f\in\mathcal{F}}|\hat{R}_{PNU}(f)-R_{PNU}(f)|.\tag{2-34}$$

设 x_i 是第 i 个正样本，x'_j 是第 j 个负样本，根据 *Usunier* 等人[65]中的定理 *6*，对于任意 $\delta>0$，对于任何 $f\in\mathcal{F}_K$，至少有 $1-\delta$ 的概率：

$$\begin{aligned}
\max_{f\in\mathcal{F}}|\hat{R}_{PN}(f)-R_{PN}(f)|\leqslant&\frac{2C_\ell C_w\sqrt{2(n_P+n_N)}}{n_P n_N}\sqrt{\sum_{i,j}\left[\|x_i\|_K^2+\|x'_j\|_K^2-2K(x_i,x'_j)\right]}\\
&+5\sqrt{\frac{n_P+n_N}{2n_P n_N}\ln(2/\delta)}\\
\leqslant&4\sqrt{2}C_\ell C_w C_x\sqrt{\frac{n_P+n_N}{n_P n_N}}+5\sqrt{\frac{(n_P+n_N)}{2n_P n_N}\ln(2/\delta)},
\end{aligned}$$

设 x_i 是第 i 个正样本，x'_j 是第 j 个无标记样本，类似地，有

$$\begin{aligned}
\max_{f\in\mathcal{F}}|\hat{R}_{PU}(f)-R_{PU}(f)|\leqslant&\frac{2C_\ell C_w\sqrt{2(n_P+n_U)}}{n_P n_N}\sqrt{\sum_{i,j}\left[\|x_i\|_K^2+\|x'_j\|_K^2-2K(x_i,x'_j)\right]}\\
&+5\sqrt{\frac{n_P+n_U}{2n_P n_U}\ln(2/\delta)}\\
\leqslant&4\sqrt{2}C_\ell C_w C_x\sqrt{\frac{n_P+n_U}{n_P n_U}}+5\sqrt{\frac{(n_P+n_U)}{2n_P n_U}\ln(2/\delta)},
\end{aligned}$$

设 x_i 是第 i 个无标记样本，x'_j 是第 j 个负样本，类似地，有

$$\begin{aligned}\max_{f\in\mathcal{F}}|\hat{R}_{UN}(f)-R_{UN}(f)| \leqslant & \frac{2C_\ell C_w\sqrt{2(n_P+n_N)}}{n_U n_N}\sqrt{\sum_{i,j}\left[\|x_i\|_K^2+\|x'_j\|_K^2-2K(x_i,x'_j)\right]} \\ & +5\sqrt{\frac{n_U+n_N}{2n_U n_N}\ln\left(2/\delta\right)} \\ \leqslant & 4\sqrt{2}C_\ell C_w C_x\sqrt{\frac{n_U+n_N}{n_U n_N}}+5\sqrt{\frac{(n_U+n_N)}{2n_U n_N}\ln(2/\delta)}.\end{aligned}$$

其中 $C_x=\max\left(\max_i\|x_i\|,\max_j\|x'_j\|\right)$。

简便起见，定义 $h(\delta)=8\sqrt{2}C_\ell C_w C_x+5\sqrt{2\ln\left(2/\delta\right)}$, 对于任意 $\delta>0$，对于任何 $f\in\mathcal{F}_K$，下列不等式至少有 $1-\delta$ 的概率成立：

$$\max_{f\in\mathcal{F}}|\hat{R}_{PN}(f)-R_{PN}(f)| \leqslant \frac{h(\delta)}{2}\sqrt{\frac{n_P+n_N}{n_P n_N}},$$

$$\max_{f\in\mathcal{F}}|\hat{R}_{PU}(f)-R_{PU}(f)| \leqslant \frac{h(\delta)}{2}\sqrt{\frac{n_P+n_U}{n_P n_U}},$$

$$\max_{f\in\mathcal{F}}|\hat{R}_{UN}(f)-R_{UN}(f)| \leqslant \frac{h(\delta)}{2}\sqrt{\frac{n_U+n_N}{n_U n_N}}.$$

简单计算可得对于任意 $\delta'>0$ 以至少 $1-\delta'$ 的概率下式成立：

$$\begin{aligned}&\max_{f\in\mathcal{F}}|\hat{R}_{PNU}(f)-R_{PNU}(f)| \\ \leqslant&\gamma\left(\max_{f\in\mathcal{F}}|\hat{R}_{PN}(f)-R_{PN}(f)|\right) \\ &+(1-\gamma)\left(\max_{f\in\mathcal{F}}|\hat{R}_{PU}(f)-R_{PU}(f)|\right)+(1-\gamma)\left(\max_{f\in\mathcal{F}}|\hat{R}_{UN}(f)-R_{UN}(f)|\right) \\ \leqslant&h(\frac{\delta'}{3})\left(\gamma\sqrt{\frac{n_P+n_N}{n_P n_N}}+(1-\gamma)\sqrt{\frac{n_P+n_U}{n_P n_U}}+(1-\gamma)\sqrt{\frac{n_U+n_N}{n_U n_N}}\right)/2\,.\end{aligned} \tag{2-35}$$

将公式 *2-34* 和不等式 *2-35* 带入不等式 *2-33* 右侧，定理得证。 □

定理 2.2 保证了优化 PNU 损失的超额风险的收敛率为：

$$\mathcal{O}\left(\frac{1}{\sqrt{n}_P}+\frac{1}{\sqrt{n}_N}+\frac{1}{\sqrt{n}_U}\right).$$

2.3.2 方差缩减

前文证明了本章所提出的经验风险估计量是无偏的，并且超额风险有界。下一个问题是，即当 $\gamma < 1$ 时，$\hat{R}_{PNU}(f)$ 的方差是否可以小于 $\hat{R}_{PN}(f)$ 的方差，换言之，是否 $\mathcal{X}_U$ 可以帮助减少估计 R_{PN} 的方差。为回答这个问题，选择任意感兴趣的 f。为简单起见，假设 $n_U \to \infty$，以展示可能取得的方差缩减效果上限。

方差与协方差的定义如下：

$$\sigma_{PN}^2(f) = \mathrm{Var}_{PN}[\ell(f(x_P, x_N))]\,,$$

$$\sigma_{PU}^2(f) = \mathrm{Var}_{PU}[\ell(f(x_P, x_U))]\,,$$

$$\sigma_{UN}^2(f) = \mathrm{Var}_{UN}[\ell(f(x_U, x_N))]\,,$$

$$\tau_{PN,PU}(f) = \mathrm{Cor}_{PN,PU}[\ell(f(x_P, x_N)), \ell(f(x_P, x_U))]\,,$$

$$\tau_{PN,UN}(f) = \mathrm{Cor}_{PN,UN}[\ell(f(x_P, x_N)), \ell(f(x_U, x_N))]\,,$$

$$\tau_{PU,UN}(f) = \mathrm{Cor}_{PU,UN}[\ell(f(x_P, x_U)), \ell(f(x_U, x_N))]\,.$$

基于以上定义，可以证明以下定理。

定理 2.3 *设 $n_U \to \infty$。对于任意固定的 f，经验风险 $\hat{R}_{PNU}(f)$ 的方差的极小值点是*

$$\gamma_{PN} = \arg\min_{\gamma} \mathrm{Var}[\hat{R}_{PNU}(f)] = \frac{\psi_{PNU}}{\psi_{PNU} - \psi_{PN}}\,, \tag{2-36}$$

其中

$$\psi_{PN} = \frac{1}{n_P n_N}\sigma_{PN}^2(f)\,,$$

$$\psi_{PNU} = \frac{1}{n_P}\tau_{PN,PU}(f) + \frac{1}{n_N}\tau_{PN,UN}(f)\,.$$

此外，如果 $\psi_{PNU} > \psi_{PN}$，则对于任意 $\gamma \in (2\gamma_{PN} - 1, 1)$ 满足 $\mathrm{Var}[\hat{R}_{PNU}(f)] \leqslant \mathrm{Var}[\hat{R}_{PN}(f)]$。

证明 经验风险可以表示为

$$\begin{aligned}\hat{R}_{PNU}(f) =&\gamma\hat{R}_{PN}(f) + (1-\gamma)(\hat{R}_{PU}(f) + \hat{R}_{UN}(f) - \frac{1}{2}) \\ =&\frac{\gamma}{n_P n_N}\sum_{i=1}^{n_P}\sum_{j=1}^{n_N}\ell(f(x_i^P, x_j^N)) \\ &+ \frac{\gamma}{n_P n_U}\sum_{i=1}^{n_P}\sum_{j=1}^{n_U}\ell(f(x_i^P, x_j^U)) \\ &+ \frac{\gamma}{n_U n_N}\sum_{i=1}^{n_U}\sum_{j=1}^{n_N}\ell(f(x_i^U, x_j^N)) \\ &+ \frac{1-\gamma}{2}\,.\end{aligned}$$

设 $n_U \to \infty$，则有

$$\begin{aligned}\mathrm{Var}[\hat{R}_{PNU}(f)] =&\gamma^2\frac{\sigma_{PN}^2}{n_P n_N} + 2\gamma(1-\gamma)\frac{\tau_{PN,PU}(f)}{n_{\tilde{P}}} + 2\gamma(1-\gamma)\frac{\tau_{PN,UN}(f)}{n_{\tilde{N}}} \\ =&\gamma^2\psi_{PN} + 2\gamma(1-\gamma)\psi_{PNU}\,,\end{aligned}$$

其中分母为 n_U 的项可以被消掉。

方差最小时，令对 γ 的导数为 0，

$$\begin{aligned}\frac{\mathrm{Var}[\hat{R}_{PNU}(f)]}{\gamma} =&2\gamma\psi_{PN} + (2-2\gamma)\psi_{PNU} \\ =&(2\psi_{PN} - 2\psi_{PNU})\gamma + 2\psi_{PNU} \\ =&0\,.\end{aligned}$$

解上述方程即可得到方差的最小值点。 □

定理 2.3 表明，如果选择适当的 γ，则所提出的风险估计量 $\hat{R}_{PNU}$ 比有监督风险估计量 $\hat{R}_{PN}$ 具有更小的方差，即无标记数据有利于构建模型。

2.4 实验验证

本章节实验使用 20 个通用数据集评估各种方法，其中包括来自 UCI 仓库[66]的 18 个数据集，以及 *ijcnn1* 和 *madelon*[67] 数据集。表 2-1 概述了这些数据集的样本数和特征数。

表 2-1 实验数据集的统计信息

Dataset	# Instance	# Features	Dataset	# Instance	# Features
australian	690	42	*breast*	277	9
breastw	683	9	*clean1*	476	166
colic	188	13	*colic.orig*	205	17
credit-a	653	15	*credit-g*	1,000	20
fourclass	862	2	*german*	1,000	59
haberman	306	14	*heart*	270	9
house	232	16	*ijcnn1*	141,691	22
madelon	2,600	500	*parkinsons*	195	22
phishing	11,055	68	*vehicle*	435	16
vote	232	16	*wdbc*	569	14

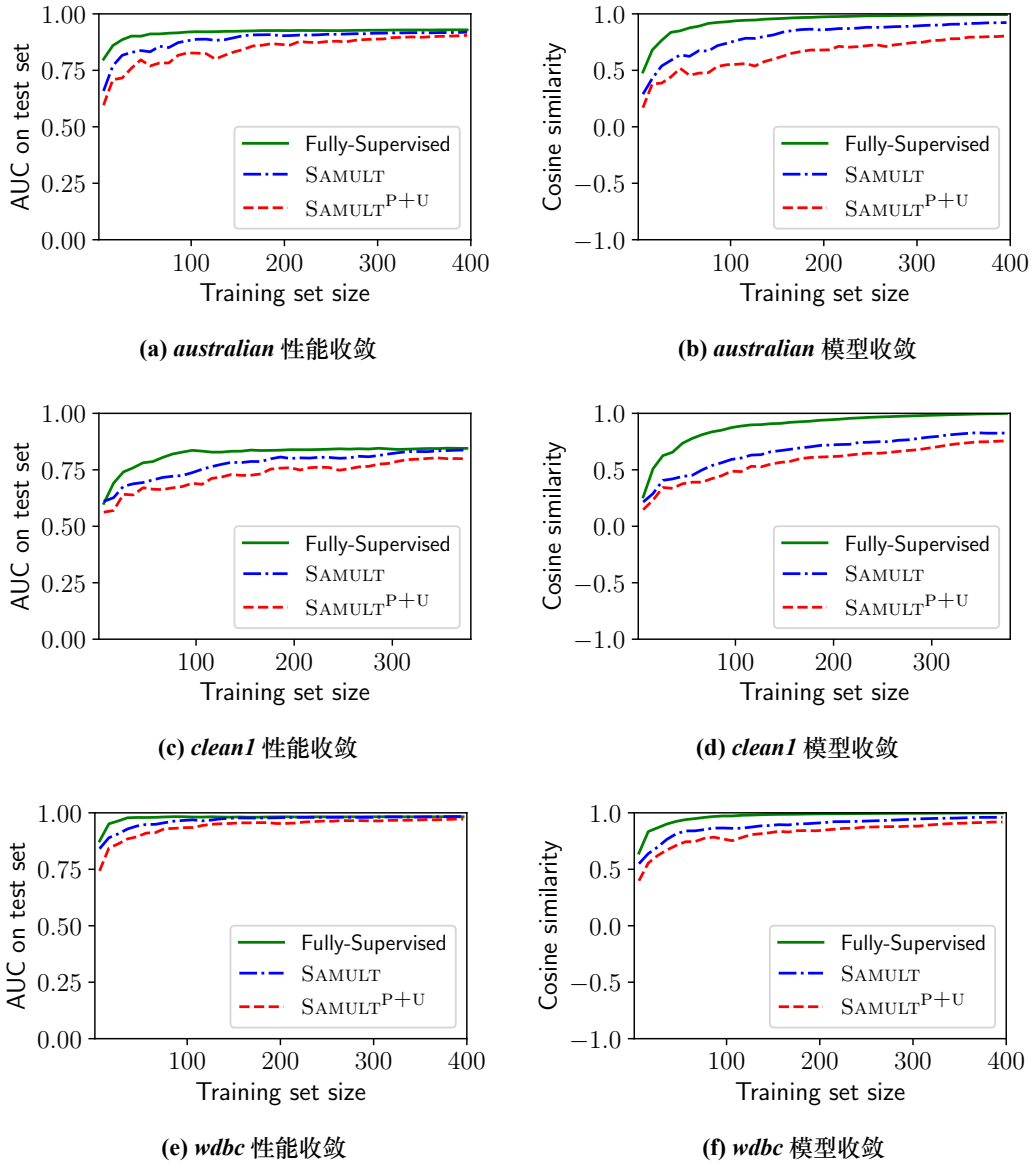
本节首先研究了由 SAMULT（优化 PNU-AUC 风险 $\hat{R}_{PNU}$）和 SAMULT$^{\text{P+U}}$（优化 PU-AUC 风险 $\hat{R}_{PU}$）训练的模型的渐近性质，以展示定理 2.1 的有效性。

其次，本节将汇报提出方法与其他半监督 AUC 优化方法或相近方法的比较结果。所有实验都在随机数据分割的情况下重复进行 50 次，并记录平均 AUC 分数以及标准差。超参数通过 5 折交叉验证的网格搜索进行选择。

当仅有正样本和无标记样本时，SAMULT 退化为一个特殊形式，只需要优化 PU-AUC 风险 $\hat{R}_{PU}$。这在渐近意义下等价于监督 AUC 优化，如定理 2.1 所示。本节针对这种退化情况进行了实证评估。

2.4.1 渐进性质

根据定理 2.1，最小化 PU-AUC 风险或 PNU-AUC 风险的模型会随着更多数据的加入而收敛到监督学习的情况。为了直观地验证这一定理，本小节通过实验考查由不同大小的训练集训练的半监督模型是否能够在 AUC 优化性能和模型相似度（考虑模型的余弦相似度）这两个方面逐渐逼近利用所有标记数据训练的完全监督模型。具体而言，我们选择了三个具有不同特征数量的数据集，逐渐增加训练数据的数量，并通过 SAMULT（优化 PNU 风险 $\hat{R}_{PNU}$）和 SAMULT$^{\text{P+U}}$（优化 PU 风险 $\hat{R}_{PU}$）两种方法构建半监督 AUC 优化模型。这两个模型所使用的训练集中只有 10% 的标记可见，而 SAMULT$^{\text{P+U}}$ 只使用正样本和无标记样本进行训练。然后，这两个方法构建的模型将和利用所有标记构建的基准模型进行比较，其性能可以视为两个半监督分类器的上限。在这组实验中，大约 20% 的数据被



(a) *australian* 性能收敛 (b) *australian* 模型收敛

(c) *clean1* 性能收敛 (d) *clean1* 模型收敛

(e) *wdbc* 性能收敛 (f) *wdbc* 模型收敛

图 2-1 各模型测试 AUC（左）、与基准模型的相似度（$\cos\langle\hat{\boldsymbol{w}}, \hat{\boldsymbol{w}}^*\rangle$）（右）随训练集大小变化曲线

保留作为测试集。实验在 10 次随机数据划分下重复进行并取平均，以降低随机性带来的影响。

图2-1(a)，2-1(c) 和 2-1(e) 显示了从不同大小的训练集中学习的分类器的 AUC 值，图2-1(b)，2-1(d) 和2-1(f) 显示了这些分类器与从所有可用标记数据中学习的最优分类器 $\hat{\boldsymbol{w}}^*$ 的余弦相似度。由图 2-1 可以看出，由 SAMULT 和 SAMULT$^{\text{P+U}}$ 训练的两个半监督分类器在 AUC 分数和余弦相似度方面随着训练数据的增加而收敛到完全监督分类器。由于 SAMULT$^{\text{P+U}}$ 没有利用负样本数据，其收敛速度相较于 SAMULT 略慢。

表 2-2 半监督场景 AUC 优化实验结果

Dataset	Supervised	Log. Reg.	SSRankBoost	PNU-AUC	SAMULT	SAMPURA
australian	.879±.029	.860±.027	.886±.013	**.903±.009**	**.903±.009**	**.903±.009**
breast	.655±.097	.625±.095	.647±.065	**.701±.029**	**.701±.029**	**.704±.026**
breastw	.987±.009	.980±.006	.984±.013	.992±.001	**.996±.001**	**.996±.001**
clean1	.760±.062	.725±.060	.737±.050	.767±.042	.777±.039	**.782±.038**
colic	.829±.112	.818±.074	.721±.062	.858±.013	**.869±.013**	**.870±.013**
colic.orig	.647±.093	.645±.076	.612±.081	.644±.048	**.658±.049**	**.663±.044**
credit-a	.893±.024	.886±.023	.885±.013	**.906±.008**	**.906±.007**	**.906±.008**
credit-g	.719±.043	.709±.030	.665±.027	.748±.018	.748±.018	**.761±.017**
fourclass	.825±.023	.826±.026	.692±.029	**.827±.008**	**.827±.008**	**.828±.006**
german	.683±.057	.672±.048	.709±.025	**.727±.019**	**.727±.019**	**.729±.017**
haberman	.551±.086	.530±.075	**.582±.067**	.547±.051	.551±.045	.556±.043
heart	.857±.065	.842±.060	.823±.042	**.876±.025**	**.876±.025**	**.878±.024**
house	**.975±.038**	.961±.015	**.972±.034**	**.979±.012**	**.979±.012**	**.979±.011**
ijcnn1	**.912±.003**	.901±.004	.902±.002	.904±.009	**.913±.005**	**.915±.004**
madelon	.510±.037	.541±.020	**.571±.023**	.528±.029	.517±.027	.530±.022
parkinsons	.848±.129	.826±.082	.799±.051	**.860±.023**	**.860±.023**	**.863±.011**
phishing	.975±.097	.972±.001	.983±.003	.974±.002	.976±.002	**.985±.002**
vehicle	.932±.038	.922±.022	.912±.039	**.965±.020**	**.965±.020**	**.970±.014**
vote	.965±.038	.951±.015	**.972±.034**	**.979±.012**	**.979±.012**	**.979±.011**
wdbc	.971±.014	.963±.006	.964±.016	**.983±.006**	**.983±.006**	**.983±.005**
#Best/Comp.	2	0	4	11	15	18

2.4.2 方法性能

为了研究本章节提出的半监督 AUC 优化方法 SAMULT 和 SAMPURA 的性能，我们在前面所述的数据集上将 SAMULT 和 SAMPURA 与最为先进的对比方法和基础对比方法进行比较：

- SSRankBoost[31]，一种 boosting 的改进算法，可以在半监督数据上优化基于样本对损失函数，学习二分类排名函数；
- PNU-AUC[34]，一种基于 PU 学习的半监督 AUC 优化方法，通过将半监督问题视作 PU 分类问题和 NU 分类问题的组合进行求解；
- 有监督的 AUC 优化，作为对比方法展示利用无标记数据产生的收益；
- Logistic 回归，作为对比方法以展示显式优化 AUC 的优势。

实验使用随机数据划分重复进行 50 次，并记录平均 AUC 分数和标准差。参数通过 5 倍交叉验证的网格搜索选择。SAMPURA 中的基学习器数量固定为 20。由于 PNU-AUC 还需要估计类先验概率来训练模型，因此在此实验中我们将真实的类先验概率提供给 PNU-AUC。

表 2-2 总结了半监督 AUC 优化的实验结果。对于每个数据集，具有最佳性能的方法以及根据成对 t-检验在显著性水平为 5% 时与最佳方法没有显著差异的方法都用粗体表示。

通过实验结果可以看出，与所有其他方法相比，SAMULT 和 SAMPURA 均实现了最佳性能。在 15 个数据集上，SAMULT 实现了最佳或相当的性能，而在 20 个数据集中，SAMPURA 则在 18 个数据集中实现了最佳或相当的性能，而对比方法 SSRankBoost 和 PNU-AUC 仅分别在 4 个和 11 个数据集中实现了最佳或相当的性能。与有监督 AUC 优化相比，SAMULT 和 SAMPURA 通过借助无标记数据，在大多数数据集上显著提高了性能。

2.4.3 仅有正标记无标记数据的退化场景

当只有正样本和无标记样本时，如前文所述，SAMULT 会退化为一种有监督式的 AUC 优化方法，等价于将无标记数据视为负样本。本小节将展示这种简单方法可以获得比现有的基于正样本和无标记样本的 AUC 优化方法更好的性能。本文将这种退化的方法称为 $\text{SAMULT}^{\text{P+U}}$。

目前，在 PU 学习场景下研究 AUC 优化的工作较少。实验将 $\text{SAMULT}^{\text{P+U}}$ 与两种现有的基于正样本和无标记样本的 AUC 优化方法进行比较：

- PU-RSVM[68]，一种基于排序 SVM 的改进算法，用于利用正-无标记数据构建排序模型；
- PU-AUC[34]，PNU-AUC 的退化版本，是一种通过补偿损失函数实现的基于无偏 AUC 风险估计量的正标记-无标记 AUC 优化方法。

本实验的实验设置与上一个实验相同。由于 PU-AUC 还需要估计类先验概率来训练模型，实验中使用真实值来代替不准确的估计，这会高估 PU-AUC 在实际使用时的性能。

实验结果如表 2-3 所示。对于每个数据集，具有最佳性能的方法以及根据成对 t-检验在显著性水平为 5% 时与最佳方法没有显著差异的方法都用粗体表示。

从结果中可以看出，$\text{SAMULT}^{\text{P+U}}$ 在 20 个数据集中有 17 个取得了最佳表现，而 PU-RSVM 仅在 2 个数据集中取得最佳表现，PU-AUC 在 7 个数据集中取得最佳表现。与 PU-RSVM 相比，在几乎所有数据集上，$\text{SAMULT}^{\text{P+U}}$ 都显示出了很

表 2-3 PU 场景 AUC 优化实验结果

Dataset	PU-RSVM	PU-AUC	SAMULT^{P+U}
australian	.844±.034	.898±.021	**.900±.019**
breast	.615±.104	**.701±.077**	**.701±.077**
breastw	.987±.009	.993±.002	**.996±.002**
clean1	.709±.072	.786±.050	**.796±.050**
colic	.807±.103	**.877±.060**	**.877±.060**
colic.orig	.621±.078	.650±.068	**.670±.061**
credit-a	.876±.028	**.912±.015**	**.912±.015**
credit-g	.688±.047	.755±.028	**.757±.027**
fourclass	.823±.031	**.832±.026**	**.832±.025**
german	.642±.045	.734±.034	**.736±.034**
haberman	**.572±.083**	.561±.081	.555±.080
heart	.835±.073	**.883±.039**	**.883±.040**
house	.945±.029	.980±.012	**.983±.011**
ijcnn1	**.927±.005**	.900±.011	.905±.012
madelon	.470±.015	**.533±.031**	.514±.032
parkinsons	.797±.089	**.870±.033**	**.870±.032**
phishing	.960±.008	.966±.005	**.970±.005**
vehicle	.942±.034	.959±.030	**.966±.025**
vote	.945±.029	.980±.012	**.983±.011**
wdbc	.967±.021	.984±.007	**.985±.007**
#Best/Comp.	2	7	17

大的改进。与 PU-AUC 相比，尽管 SAMULT^{P+U} 在许多数据集上表现几乎相同或略微更好，但 SAMULT^{P+U} 几乎不会表现更差。而且 SAMULT^{P+U} 不依赖于对类先验概率的知识，更符合实际的应用场景。SAMULT^{P+U} 与 PU-AUC 均优化通过正样本和无标记样本所估计的 AUC 风险，这可以解释为何两者性能非常接近。然而，PU-AUC 需要通过已知的类别先验对损失函数进行补偿，即使在先验概率估计准确的情况下也会引入额外的误差，而对先验概率估计不准确时误差将进一步增大。相比之下，SAMULT^{P+U} 无需此步骤，对 AUC 风险的估计更准确，因此能够较为稳定地实现性能的增强。

上述实验结果与分析表明，当只有正类和无标记数据可用时，简单地将无标记数据视为负类就足以学习模型。通过使用 SAMULT^{P+U} 方法进行学习，在正标记-无标记学习场景下，估计类别先验概率和给无标记数据打伪标记的策略则可以被省略。

2.5 本章小结

本章节从理论上证明，在半监督 AUC 优化中，将无标记数据视为正和负数据即可进行无偏的 AUC 风险估计。基于该结论，本章提出了新型的半监督 AUC 优化方法 SAMULT 和 SAMPURA。这两种方法不需要依赖于对分布形态的具体假设，避免了由于假设不成立而带来的方法失效的风险，也无需依赖对分布先验概率的知识。实验结果表明，所提出的方法优于现有的半监督 AUC 优化方法。此外，我们还展示了正标记-无标记 AUC 优化问题也可以通过本章方法的简化版本来解决，该方法简单地将无标记数据视为负数据，不需要任何分布假设或先验知识，也可以取得与现有方法可比或更优的结果。

本章工作已总结成文:

Zheng Xie, Ming Li. "Semi-Supervised AUC Optimization without Guessing Labels of Unlabeled Data." In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2018.

第三章　标记不完全可见在线 AUC 优化

3.1　引言

由于完整的标注信息在许多机器学习任务中难以获得，在应用机器学习算法时，通常不得不设计算法从弱标记数据中进行学习。上一章中，本文针对标记不完全可见的场景提出了半监督 AUC 优化方法，在多种弱监督任务上取得了良好的 AUC 优化性能，且在数据分布不平衡的场景下也具有良好的学习效果。然而，该方法依赖于对于全量数据的存储和批量计算，在大量数据流式产生、对于模型更新效率要求高的场景下难以具体应用。这类任务在大型的工业生产环境中十分常见，如网络产品的异常检测、生产线的运行状态监控识别等。

软件持续集成（Continuous Integration，CI）[69-70]系统中对于软件构建结果的预测任务是上述问题的一个具体例子。在软件开发中，持续集成系统通过频繁的对代码进行集成和测试来及时发现错误和解决冲突。然而，在大多数情况下，成功的构建事件占据了绝大部分，而大型项目的构建往往需要花费数个小时。由于成功的构建不能为开发人员提供任何指导，因此大量的构建工作都是浪费的[71]。在这种情况下，若能提前通过模型预测构建能否通过，则可以将计算资源优先分配给更加可能会失败的代码提交，减少构建开销。这种任务被称为持续构建结果预测[72]或持续缺陷预测[73]。作为一类有代表性的现实应用案例，该任务对于机器学习模型有如下需求：

- **从流式数据中学习**，即大量数据依次产生，难以全部存储，需要模型在线更新。随着软件系统的开发进行，每次提交都会触发一次 CI 事件。预测模型应该能够不断地从一系列 CI 事件中学习，并同时为每个到达的 CI 事件预测构建结果。
- **应对数据标记不完全**，即标注数据只占所有数据中的一小部分。由于 CI 构建的高计算成本和有限资源，构建实际上只会针对一小部分提交进行，以

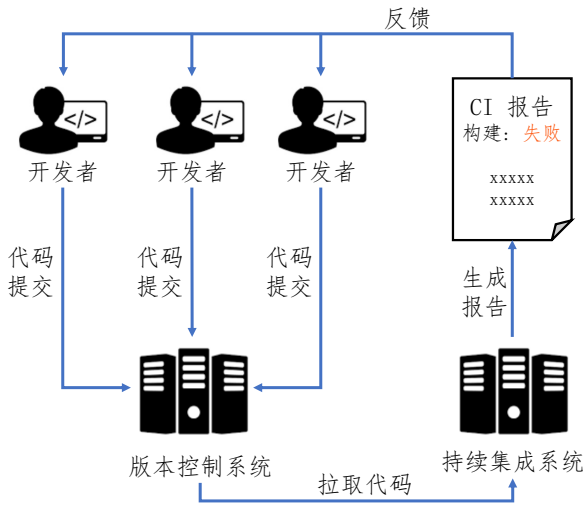


图 3-1 持续集成（CI）流程示意图

获得构建结果标记。学习器必须利用无标记的构建记录进行模型构建。

- **应对类别分布不平衡**，即不同类别数据数量差异较大。在大多数情况下，开发人员会提交被认为是正确的代码进行 CI，这使得构建更容易成功，但是严重的不平衡性使得训练针对分类正确率设计的模型更加困难。
- **需要对样本的排序**，即用户需要模型对样本的重要程度进行排序，以便依次排查。在构建结果预测中，模型需要为每个构建事件提供构建结果的可疑度排序，以便根据 CI 服务器的工作负载筛查不同比例的构建事件。

目前，已有不少研究者对持续构建结果预测问题进行探究。例如 Hassan 等人[72]使用基于决策树的算法和针对性设计的特征解决该任务，并未考虑大型 CI 系统下应用的效率问题。Finlay 等人[74]注意到该任务面临从大量流式数据构建模型的挑战，提出了基于 Hoeffding 树，即一种能够增量更新的决策树来应对该任务，但忽略了数据分布不对称带来的影响。Ni 等人[75]使用级联分类器进一步提高了查找失败构建的准确性，但没有对方法的时效性、应对分布不平衡性的能力作出考量。这些研究工作虽然已经证实持续构建结果预测对于 CI 系统的性能有所帮助，但没有相关方法能够完全符合上述提出的四个任务需求。

实际上，即使不限定在针对该应用场景的研究工作中，也没有现成的机器学习算法能够同时满足这些需求。这是由于从流式数据中构建模型的要求对模型学习带来了较大困难。由于不能存储和反复扫描全量数据，学习算法难以对数据分布进行估计；经典的半监督机器学习策略如基于生成式模型、基于低密度间隔、基于图的半监督学习方法在该场景下均难以应用。学术界目前在半监督

在线学习方向具有代表性的工作有 Goldberg 博士在其攻读博士学位期间提出的在线流形正则（Online Manifold Regularization，OMR）[76]方法和在线主动半监督学习算法（Online Active SemI-Supervised Learning，OASIS）[77]。其中，前者由于仍然需要存储全部（或部分）数据且具有平方级别的时间复杂度，并没有真正意义上解决基于大规模流式数据更新模型的问题；后者则需要依赖专家回答进行主动学习。Liu 等人[78]提出的在线半监督 SVM 由于需要动态维护一个较大的数据缓存的相似度矩阵，也需要较高的时间和空间开销。更不用说这些方法还无法解决数据分布不平衡的问题。因此，若要为软件持续集成系统设计构建结果预测模型，需要有针对性地设计新的学习方法。

根据上面的分析，我们发现若能够将半监督 AUC 优化算法扩展到在线优化场景，则可以较好的满足应用需求。然而，该问题存在两个挑战：第一，由于许多半监督方法依赖对于数据的伪标记进行估计后迭代训练分类器，需要将数据全部存储，使得模型基于单个样本在线更新难以实现。第二，由于 AUC 优化问题的损失函数基于正负样本对定义，同样为基于单个样本计算损失梯度更新模型带来了较大的困难。

在本章节中，我们通过将上一章中所提出的半监督 AUC 优化方法转化成为随机鞍点问题求解，解决了上述两个挑战，进而提出一种新的半监督在线 AUC 优化方法，名为 SOLA（Semi-supervised OnLine AUC optimization）。具体而言，SOLA 基于 AUC 能够进行无偏半监督风险估计的特性，即无标记数据可以在不估计其可能的伪标记的情况下帮助学习，免除了算法需要通过迭代生成伪标记的要求。另外，通过将原经验风险最小化问题转换为可以基于单个样本计算梯度的随机鞍点问题，使得优化可以基于单个样本计算梯度。基于这两点，SOLA 即可无需保留历史数据，基于单个样本实现高效的半监督在线 AUC 优化。SOLA 方法能够满足前面所述的软件持续构建预测任务的四个需求，实验结果表明，该方法能够在该任务上取得优秀的预测效果和极高的模型更新效率。

3.2 半监督在线 AUC 优化方法 SOLA

在介绍 SOLA 方法的具体细节前，先给出学习问题的形式化定义。在软件构建预测任务中，设 $\mathcal{X}_P$ 和 $\mathcal{X}_N$ 分别表示正负样本，即失败和成功的构建事件集

合。由于受到计算资源的限制，并不是所有构建事件都能被执行，因此我们还有一组无标记的样本 $\mathcal{X}_U$。正样本、负样本集合可以被视为分别采样自正样本分布和负样本分布，而无标记样本则采样于样本的真实分布，是正样本分布和负样本分布的混合分布。

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = +1)\,, \tag{3-1}$$

$$\mathcal{X}_N := \{\boldsymbol{x}'_j\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_N(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = -1)\,, \tag{3-2}$$

$$\mathcal{X}_U := \{\boldsymbol{x}''_k\}_{k=1}^{n_U} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) := \theta_P p_P(\boldsymbol{x}) + \theta_N p_N(\boldsymbol{x})\,, \tag{3-3}$$

由于面临的问题具有分布不平衡的问题，分类准确率无法对模型进行准确评估，因此面向分类准确率的模型十分容易失效。在这个任务中，采用 AUC 作为模型的优化目标可以解决这个问题。

与前文定义相同，有标记样本上的 AUC 可以被形式化为：

$$\mathrm{AUC} = 1 - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell_{01}(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}'))\right]\right]. \tag{3-4}$$

因此，AUC 优化问题可以被形式化为一个经验风险最小化问题。在使用 ℓ_2 正则化的情况下，AUC 风险的最小化可以被形式化为：

$$\min_{||\boldsymbol{w}||<R} \quad R_{PN}(\boldsymbol{w})\,, \tag{3-5}$$

其中

$$R_{PN} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\ell(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}')) \tag{3-6}$$

是有监督（正 vs. 负）的 AUC 风险。这里 n_P 和 n_N 分别是正负样本的数量。在实践中，由于 0-1 损失是离散的且难以优化，通常使用平方损失 $\ell(z) = (1-z)^2$ 进行优化。现有工作已经证明平方损失与 AUC 风险一致[9]。AUC 可以被视为一种成对排序损失，这使它自然适用于处理不平衡数据和对样本进行排序。

由于在数据逐步到来时估计数据分布十分困难，因此难以通过打伪标记等常用方法对无标记数据进行利用。我们基于上一章节提出的半监督 AUC 优化的成果来解决该问题，即优化随机抽取的正样本在排名上优于随机抽取的无标记

样本的概率（或无标记样本在排名上优于负样本的概率）等价于无偏的 AUC 优化[79]。有了这个方法，半监督 AUC 优化可以被形式化为以下损失函数的最小化问题以利用无标记的数据。该函数由三个风险估计量组成：

$$\min_{||\boldsymbol{w}||<R} \quad \gamma R_{PN} + (1-\gamma)(R_{PU} + R_{UN})\,, \tag{3-7}$$

其中 $\gamma \in [0,1]$ 是有监督风险和半监督风险的折中系数，

$$R_{PU} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P} \mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_U} \ell(\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{x}''))\,, \tag{3-8}$$

$$R_{UN} = \mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_U} \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N} \ell(\boldsymbol{w}^\top(\boldsymbol{x}''-\boldsymbol{x}')) \tag{3-9}$$

是半监督风险估计量。

直接应用公式 3-7 仍然无法处理流数据，但它为我们提供了一种不需要估计数据分布即可利用无标记数据的方法。将公式 3-7 扩展到在线环境中的另一个障碍是 AUC 的风险是定义在样本对上的，因此不能通过直接使用随机梯度下降来解决该问题。为了克服这个问题，本章节将公式 3-7 重写为一个鞍点问题，并提出了一种算法，可以通过一个极小-极大过程在线更新模型。

定理 3.1 *以下鞍点问题与公式 3-7 等价：*

$$\min_{\substack{||\boldsymbol{w}||<R,\\ a_1,b_1,\\ a_2,b_2,\\ a_3,b_3}} \max_{\alpha_1,\alpha_2,\alpha_3} \gamma f_{PN}(\boldsymbol{w},a_1,b_1,\alpha_1) + (1-\gamma)\big(f_{PU}(\boldsymbol{w},a_2,b_2,\alpha_2) + f_{UN}(\boldsymbol{w},a_3,b_3,\alpha_3)\big)\,, \tag{3-10}$$

其中所涉及到的函数定义如下：

$$f_{PN}(\boldsymbol{w},a,b,\alpha) = \mathop{\mathbb{E}}_{\boldsymbol{x}}[F_{PN}(\boldsymbol{w},a,b,\alpha;\boldsymbol{x})]\,,$$

$$f_{PU}(\boldsymbol{w},a,b,\alpha) = \mathop{\mathbb{E}}_{\boldsymbol{x}}[F_{PU}(\boldsymbol{w},a,b,\alpha;\boldsymbol{x})]\,,$$

$$f_{UN}(\boldsymbol{w},a,b,\alpha) = \mathop{\mathbb{E}}_{\boldsymbol{x}}[F_{UN}(\boldsymbol{w},a,b,\alpha;\boldsymbol{x})]\,,$$

$$
\begin{aligned}
F_{PN}(\boldsymbol{w}, a, b, \alpha; \boldsymbol{x}) = & -\frac{n_P n_N}{n_P + n_N}\alpha^2 \\
& + n_N \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_P]\left((\boldsymbol{w}^\top \boldsymbol{x} - a)^2 - 2(1+\alpha)\boldsymbol{w}^\top \boldsymbol{x}\right) \\
& + n_P \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_N]\left((\boldsymbol{w}^\top \boldsymbol{x} - b)^2 + 2(1+\alpha)\boldsymbol{w}^\top \boldsymbol{x}\right), \\
F_{PU}(\boldsymbol{w}, a, b, \alpha; \boldsymbol{x}) = & -\frac{n_P n_U}{n_P + n_U}\alpha^2 \\
& + n_U \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_P]\left((\boldsymbol{w}^\top \boldsymbol{x} - a)^2 - 2(1+\alpha)\boldsymbol{w}^\top \boldsymbol{x}\right) \\
& + n_P \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_U]\left((\boldsymbol{w}^\top \boldsymbol{x} - b)^2 + 2(1+\alpha)\boldsymbol{w}^\top \boldsymbol{x}\right), \\
F_{UN}(\boldsymbol{w}, a, b, \alpha; \boldsymbol{x}) = & -\frac{n_U n_N}{n_U + n_N}\alpha^2 \\
& + n_N \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_U]\left((\boldsymbol{w}^\top \boldsymbol{x} - a)^2 - 2(1+\alpha)\boldsymbol{w}^\top \boldsymbol{x}\right) \\
& + n_U \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_N]\left((\boldsymbol{w}^\top \boldsymbol{x} - b)^2 + 2(1+\alpha)\boldsymbol{w}^\top \boldsymbol{x}\right).
\end{aligned}
$$

接下来，我们首先证明公式 3-7 和公式 3-10 的等价性，然后给出求解公式 3-10 的在线优化算法。

证明 *AUC* 风险 R_{PN} 可以被写为:

$$
\begin{aligned}
R_{PN} = & \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P} \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N} \ell(\boldsymbol{w}^\top(\boldsymbol{x} - \boldsymbol{x}')) \\
= & \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[(\boldsymbol{w}^\top \boldsymbol{x})^2] + \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[(\boldsymbol{w}^\top \boldsymbol{x}')^2] - 2\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}] \\
& + 2\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'] - 2\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}]\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'] + 1 \\
= & 1 + (\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[(\boldsymbol{w}^\top \boldsymbol{x})^2] - (\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}])^2) + (\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[(\boldsymbol{w}^\top \boldsymbol{x}')^2] - (\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'])^2) \\
& + 2\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'] - 2\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}] + (\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}] - \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'])^2.
\end{aligned}
$$

注意到

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[(\boldsymbol{w}^\top \boldsymbol{x})^2] - (\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}])^2 &= \min_a \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[(\boldsymbol{w}^\top \boldsymbol{x} - a)^2], \\
\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[(\boldsymbol{w}^\top \boldsymbol{x}')^2] - (\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'])^2 &= \min_b \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[(\boldsymbol{w}^\top \boldsymbol{x}' - b)^2],
\end{aligned}
$$

当 $a = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}]$，$b = \mathbb{E}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}']$ 时取得最小值。另外，

$$(\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}] - \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'])^2 = \max_{\alpha} 2\alpha(\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'] - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}]) - \alpha^2\,,$$

当 $\alpha = \mathbb{E}_{\boldsymbol{x}'\in\mathcal{X}_N}[\boldsymbol{w}^\top \boldsymbol{x}'] - \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}_P}[\boldsymbol{w}^\top \boldsymbol{x}]$ 取得最大值。综上可以得出：

$$R_{PN} = 1 + \min_{\substack{||\boldsymbol{w}||<R,\\ a,b}} \max_{\alpha} \mathop{\mathbb{E}}_{\boldsymbol{x}}[F_{PN}(\boldsymbol{w}, a, b, \alpha; \boldsymbol{x})]\,. \tag{3-11}$$

通过类似方法可以证明：

$$R_{PU} = 1 + \min_{\substack{||\boldsymbol{w}||<R,\\ a,b}} \max_{\alpha} \mathop{\mathbb{E}}_{\boldsymbol{x}}[F_{PU}(\boldsymbol{w}, a, b, \alpha; \boldsymbol{x})]\,, \tag{3-12}$$

$$R_{UN} = 1 + \min_{\substack{||\boldsymbol{w}||<R,\\ a,b}} \max_{\alpha} \mathop{\mathbb{E}}_{\boldsymbol{x}}[F_{UN}(\boldsymbol{w}, a, b, \alpha; \boldsymbol{x})]\,, \tag{3-13}$$

□

因此公式 *3-7* 等价于公式 *3-10* 。

由于函数 f_{PN}, f_{PU}, 及 f_{UN} 在原始变量 $(\boldsymbol{w}, a, b)$ 上是凸函数，在对偶变量 α 上是凹函数，因此可以通过在原始变量 $(\boldsymbol{w}, a_1, b_1, a_2, b_2, a_3, b_3)$ 上进行梯度下降，同时在对偶变量 $(\alpha_1, \alpha_2, \alpha_3)$ 上进行梯度上升来解决公式 3-10 。通过使用 F_{PN}，F_{PU}，和 F_{UN} 的梯度作为 f_{PN}，f_{PU} 和 f_{UN} 的梯度的无偏估计进行优化，可以在每个样本上进行计算，更新模型。每个有标记或无标记的样本都会影响 F_{PN}，F_{PU}，和 F_{UN} 三者中的两个函数。因此，在每次迭代中，应计算两个函数的梯度，即在原始变量中进行梯度下降和在对偶变量中进行梯度上升。模型权重 $\boldsymbol{w}$ 的更新规则为：

$$\begin{aligned}\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} &- \eta^{(t)}\gamma\frac{\partial F_{PN}(x^{(t)})}{\partial \boldsymbol{w}^{(t)}} \\ &- \eta^{(t)}(1-\gamma)(\frac{\partial F_{PU}(x^{(t)})}{\partial \boldsymbol{w}^{(t)}} + \frac{\partial F_{UN}(x^{(t)})}{\partial \boldsymbol{w}^{(t)}})\,,\end{aligned} \tag{3-14}$$

在优化 F_{PN} 中，其他优化变量的更新规则为

$$(a_1^{(t+1)}, b_1^{(t+1)}) \leftarrow (a_1^{(t)}, b_1^{(t)}) - \eta^{(t)}\gamma\frac{\partial F_{PN}(x^{(t)})}{\partial (a_1^{(t)}, b_1^{(t)})}\,, \tag{3-15}$$

$$\alpha_1^{(t+1)} \leftarrow \alpha_1^{(t)} + \eta^{(t)}\gamma\frac{\partial F_{PN}}{\partial \alpha_1^{(t)}}\,. \tag{3-16}$$

在优化 F_{PU} 过程中，更新规则为

$$(a_2^{(t+1)}, b_2^{(t+1)}) \leftarrow (a_2^{(t)}, b_2^{(t)}) - \eta^{(t)}(1-\gamma)\frac{\partial F_{PU}(x^{(t)})}{\partial (a_2^{(t)}, b_2^{(t)})}\,, \tag{3-17}$$

$$\alpha_2^{(t+1)} \leftarrow \alpha_2^{(t)} + \eta^{(t)}(1-\gamma)\frac{\partial F_{PU}}{\partial \alpha_2^{(t)}}\,. \tag{3-18}$$

在优化 F_{UN} 过程中，更新规则为

$$(a_3^{(t+1)}, b_3^{(t+1)}) \leftarrow (a_3^{(t)}, b_3^{(t)}) - \eta^{(t)}(1-\gamma)\frac{\partial F_{UN}(x^{(t)})}{\partial (a_3^{(t)}, b_3^{(t)})}\,, \tag{3-19}$$

$$\alpha_3^{(t+1)} \leftarrow \alpha_3^{(t)} + \eta^{(t)}(1-\gamma)\frac{\partial F_{UN}}{\partial \alpha_3^{(t)}}\,. \tag{3-20}$$

算法的流程如算法 3 所示。值得注意的是，即使缺失了三种类型数据（即正样本、负样本和无标记样本）中的其中一种，SOLA 仍然可以正常训练。

3.3 实验验证

3.3.1 实验设置

数据集 为了验证本章节提出的 SOLA 方法的效果，我们在多个真实的持续构建预测任务上验证方法。进行持续构建结果预测需要使用两种数据源：持续集成系统和版本控制系统。前者提供构建结果和相应的提交 ID 的构建信息，而用户可以从后者获取有关项目、代码更改和开发人员的信息。文献[73] 提出了一个用于持续构建结果预测的数据集，其中包含使用 CI 并托管在 GitHub 上的 1265 个开源项目。这些项目使用由三个主流 CI 平台提供的 CI 服务：Jenkins、Travis CI 和 TeamCity。除了具有明确结果标记（即成功或失败）的构建记录外，这三个平台还提供了不明确的构建结标记，包括 Jenkins 上的 “not_build”、“aborted”、“unstable”、Travis CI 上的 “errored” ，TeamCity 上的 “error”、“warning”、“unknown”。带有不明确标记的构建记录在数据集中标记为未知，在学习过程中可以视为无标记的样本。不同项目的数据在构建记录总数、标记记录百分比和成功构建与失败构

算法 3 SOLA

Input: 样本序列 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$

初始化 $t \leftarrow 0, n_P^{(0)} \leftarrow 0, n_N^{(0)} \leftarrow 0, n_U^{(0)} \leftarrow 0$
随机初始化优化变量
for 每个样本 $\boldsymbol{x}^{(t+1)}$ **do**
    if $\boldsymbol{x}^{(t+1)}$ 是正样本 **then**
        $n_P^{(t+1)} \leftarrow n_P^{(t)} + 1$
        根据公式 3-15 和 3-16 更新 (a_1, b_1, α_1)
        根据公式 3-17 和 3-18 更新 (a_2, b_2, α_2)
    end if
    if $\boldsymbol{x}^{(t+1)}$ 是负样本 **then**
        $n_N^{(t+1)} \leftarrow n_N^{(t)} + 1$
        根据公式 3-15 和 3-16 更新 (a_1, b_1, α_1)
        根据公式 3-19 和 3-20 更新 (a_3, b_3, α_3)
    end if
    if $\boldsymbol{x}^{(t+1)}$ 是无标记样本 **then**
        $n_U^{(t+1)} \leftarrow n_U^{(t)} + 1$
        根据公式 3-17 和 3-18 更新 (a_2, b_2, α_2)
        根据公式 3-19 和 3-20 更新 (a_3, b_3, α_3)
    end if
    根据公式 3-14 计算 $\boldsymbol{w}^{(t+1)}$
    if $||\boldsymbol{w}^{(t+1)}|| > R$ **then**
        $\boldsymbol{w}^{(t+1)} \leftarrow R\boldsymbol{w}^{(t+1)}/||\boldsymbol{w}^{(t+1)}||$
    end if
    $t \leftarrow t + 1$
end for

Output: 模型参数 $\boldsymbol{w}^{(t)}$

建比率方面存在显著差异。为了研究方法在不同情况下的性能，本节中，我们选择了 8 个具有不同不平衡比例和大小的项目来评估 SOLA 方法和其他方法。由于必须留出一些标记记录进行评估，因此实验中排除了标记记录少于 200 条的项目。表 3-1 展示了实验中设计到的软件项目的统计数据。实验采用文献[73]中提供的特征以及从提交日志中计算的其他一些特征。提交 ID 和构建时间等与学习无关的特征在实验中被排除。表 3-2 展示了用于构建模型的特征。

对比方法 为了验证同时处理四个挑战的必要性，本章节选择如下的对比方法进行实验，包括针对构建预测任务设计的方法和其他较为适用的方法：

- Hoeffding Tree[74]：基于 Hoeffding Tree 的方法，用于处理软件构建结果的流数据。此方法能够处理流数据，但不能满足另外三个需求。
- OMR[76]：一种半监督学习方法，通过进行在线流形正则化，能够从包含有

表 3-1 项目统计信息

项目名称	成功构建数 (S)	失败构建数 (F)	未知构建数	S:F
deeplearning4j	5	426	422	0.01
capybara	173	185	106	0.94
killbill	841	332	790	2.53
rails	8,048	2,792	805	2.88
codetriage	242	32	15	7.56
oryx	346	40	31	8.65
phony	332	29	24	11.45
stringer	293	9	9	32.56

表 3-2 构建记录特征含义

缩写	特征含义
LBO	最后构建结果 (Last Build Outcome)
NR	修订次数 (Number of Revisions)
NRC	修订代码文件数 (Number of Revised Code Files)
NML	修改行数 (Number of Modified Lines)
NMLC	代码文件中修改行数 (Number of Modified Lines in Code Files)
NDC	不同贡献者数量 (Number of Distinct Committers)
NC	提交次数 (Number of Commits)

标记和无标记的流式数据中学习。它没有考虑数据不平衡，可能会导致性能受到影响。

- SOLAM[8]：一种在线 AUC 优化方法。由于以 AUC 作为优化目标，SOLAM 可以处理数据不平衡，并对构建事件的可疑程度进行排序。然而，它不能利用无标记数据。
- SAMULT[79]：一种批量半监督 AUC 优化方法。虽然它不能处理流数据，在实践中时间效率不佳，但我们在实验中将其用作本文方法的上界参考。

3.3.2 方法性能

本小节通过重复实验评估了各个方法的性能。对于每个软件项目，最后的 100 个有标记样本被用作测试数据，在此之前的样本被用作训练数据。在第 100 个有标记样本之后的无标记样本被忽略。实验使用 AUC 作为性能指标，它反映了方法的排名能力。各个方法的性能如表 3-3 所示。其中，加粗的数字指最优性能或基于成对 t 检验与最优性能可比的性能（显著性水平 5%）。用剑标标注的数据指方法产出的模型将所有数据分类为一类或给出相同打分，不具有实际意义。

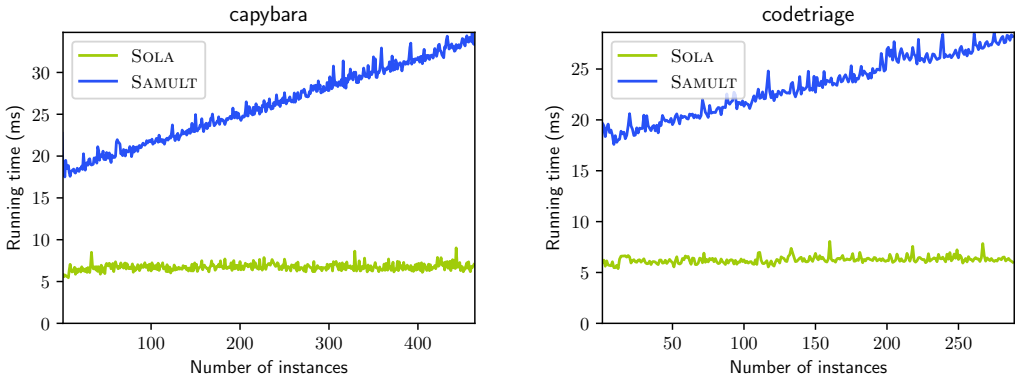
表 3-3 在线半监督 AUC 优化实验结果

Methods	Online Methods				Batch Method
	Hoef. Tree	OMR	SOLAM	SOLA	SAMULT
	—	Semi.	AUC-Opt.	AUC-Opt./Semi.	AUC-Opt./Semi.
deeplearning4j	0.429	0.500†	**0.832**	**0.832**	**0.848**
capybara	0.612	0.604	**0.718**	**0.727**	0.658
killbill	**0.692**	0.686	0.688	0.688	**0.710**
rails	0.531	0.529	**0.611**	**0.616**	**0.627**
codetriage	0.640	0.495	0.648	**0.707**	**0.704**
oryx	0.500†	0.500†	0.665	**0.737**	**0.757**
phony	0.500†	0.500†	0.808	**0.980**	**0.979**
stringer	0.500†	0.500†	0.954	**0.975**	**0.975**
Average	0.551	0.539	0.741	0.783	0.782

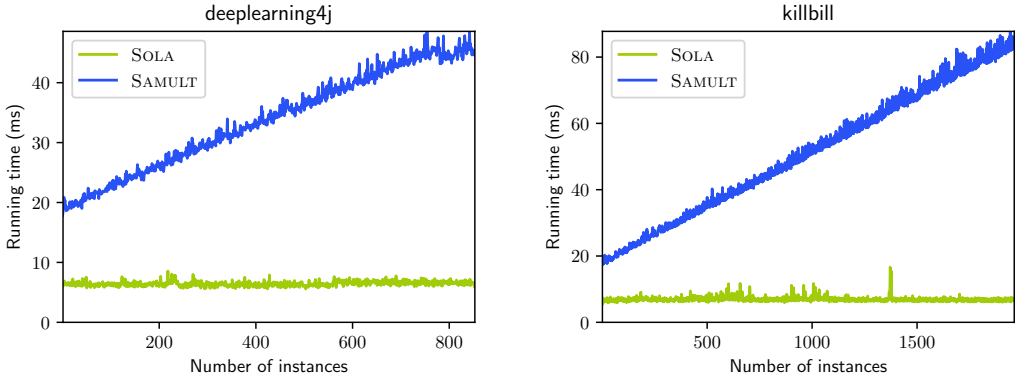
与构建结果预测专用方法比较 SOLA 在大多数项目上都击败了 Hoeffding Tree。值得注意的是，当数据高度不平衡时，例如 *deeplearning4j*、*phony*、*stringer* 等，Hoeffding Tree 无法学习有效的模型，导致 AUC 分数低至 0.5 或更低。当数据集相对平衡时，Hoeffding Tree 可以从足够的正负数据中学习到有效的模型。然而，通过优化 AUC，SOLA 无论数据是否平衡总是可以学习到有效的模型。在平均意义下，SOLA 的 AUC 比 Hoeffding Tree 高出 42.1%。

与忽视数据不平衡性的方法比较 OMR 没有针对不平衡情况优化学习算法，因此当数据高度不平衡时，模型无法进行有效的预测，就像 Hoeffding Tree 一样。只有在如 *capybara* 和 *killbill* 这种数据分布相对平衡的项目中，OMR 才能学习到有效的模型，但性能仍然较差。在平均意义下，SOLA 的 AUC 比 Hoeffding Tree 高出 45.3%。将 OMR 和 Hoeffding Tree 与 SOLA 进行比较，可以得出结论：AUC 优化对于从不平衡数据中学习模型至关重要。

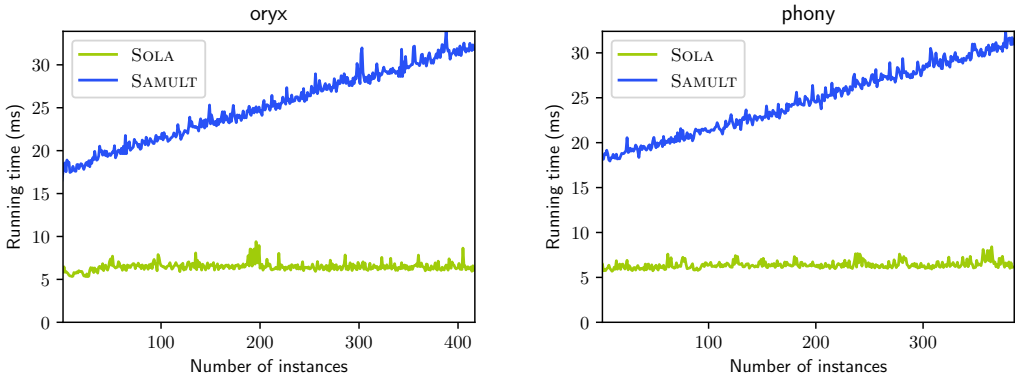
与不使用无标记数据的方法比较 由于 SOLAM 以 AUC 作为优化目标，能够应对数据分布不平衡，因此它可以在所有项目中学习到有效的模型。但是，在大多数情况下，SOLA 由于利用了无标记数据而优于 SOLAM。SOLAM 在 *deeplearning4j*、*killbill* 和 *rails* 中的表现与 SOLA 相当，其原因在于这几个数据集上有足够的有标记数据用于学习模型。但是，通过利用无标记数据辅助学习，SOLA 的性能平均比 SOLAM 提高了 5.7%。



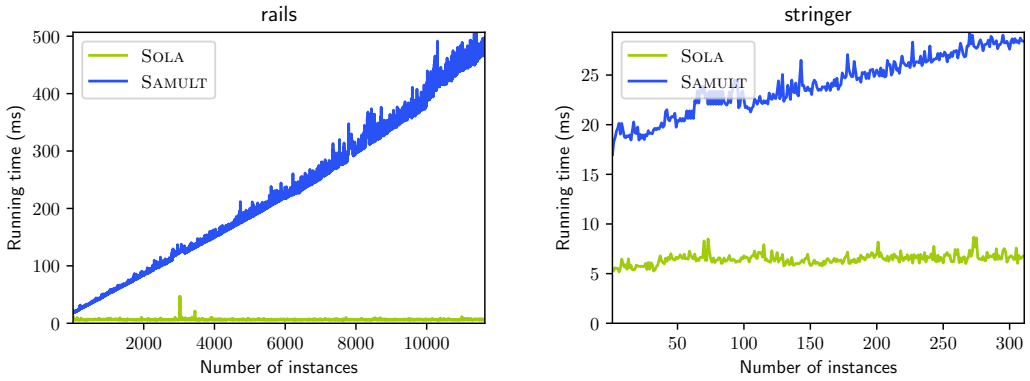
(a) *capybara* (b) *codetriage*



(c) *deeplearning4j* (d) *killbill*



(e) *oryx* (f) *phony*



(g) *rails* (h) *stringer*

图 3-2 各个数据集上 SOLA 与离线算法 SAMULT 模型更新耗时对比

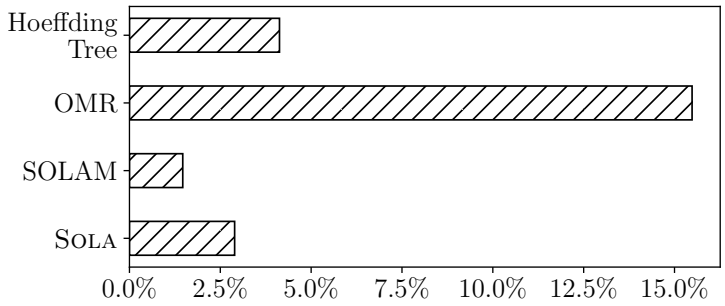


图 3-3 各个在线算法的相对运行时间（以离线算法为基准）

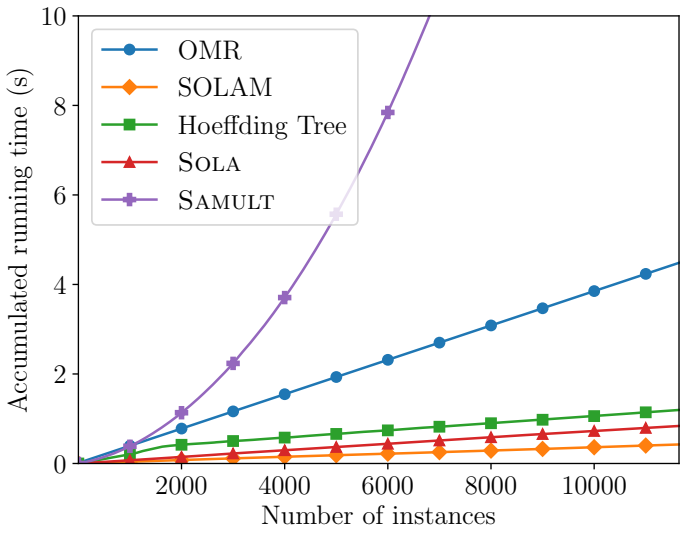


图 3-4 各个算法的累计运行时间

与批量方法比较 SAMULT 是一种离线批量学习方法，可以看作是在线求解方法的上限。如表 3-3 所示，SOLA 的性能与 SAMULT 接近，这意味着 SOLA 通过仅处理每个样本一次就可以从序列数据中有效地学习。然而，下一小节的实验结果将展示出，得益于在线更新的能力，SOLA 的运行时间显著降低。

综合以上实验结果，可以看出，在 CI 构建结果预测中，SOLA 凭借其处理不平衡数据、对样本进行排序和利用无标记数据的能力，优于现有的对比方法。SOLA 还通过在线更新模型节省了大量时间开销。这些结果表明，同时解决四个挑战对 CI 构建结果预测是有益的。

3.3.3 运行时间

为了研究将半监督 AUC 优化方法应用于在线情况下能否减少运行时间，我们评估了 SOLA 和其他对比方法的运行时间。实验模拟每次有新样本到达时更新

模型，并测量模型更新的时间。在实验中，该更新过程均被重复 100 次以消除随机性的影响。图 3-2 展示了 SOLA 与离线算法 SAMULT 更新效率的差异。图 3-3 和图 3-4 分别展示了以规模最大的数据集 *rails* 为例，各个在线算法相对离线算法的运行时间和各个方法的累计运行时间。

在图 3-3 中可以观察到，与批量优化方法 SAMULT 相比，SOLA 在模型更新期间节省了超过 97% 的时间。对比方法中，只有 SOLAM 的时间消耗略好于 SOLA，其原因是 SOLAM 忽略了无标记数据。可以看出，通过采用在线算法，SOLA 大大节省了模型更新的时间。

3.4 本章小结

本章节提出了在线半监督 AUC 优化方法 SOLA，首次将半监督 AUC 优化问题扩展到在线优化的场景。该方法能够很好地处理以持续构建结果预测为代表的一类任务的多个需求，即从流式数据中学习，应对数据标记不完全，应对类别分布不均衡，以及需要对样本排序。实验结果表明，SOLA 方法通过良好地应对上述四个需求，能够取得比其他对比方法更好的效果。

本章工作已总结成文:

Zheng Xie, Ming Li. “Cutting the Software Building Efforts in Continuous Integration by Semi-Supervised Online AUC Optimization.” In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 2018.

第四章　标记不准确不完全可见 AUC 优化

4.1　引言

在诸多完整标记信息难以获取的机器学习任务中，算法不得不通过弱标记信息进行学习。除了最为典型的标记不完全可见的学习场景，还常常会遇到标记 “不准确” 和 “不确切” 的学习场景。例如，由于标注难度较大或标记是从不可靠的来源收集时，部分样本的标记有可能发生错误。在通过众包机制对数据进行标注时，也会出现类似的问题。这一类标记可能出现错误的情况通常被称为噪声标记[80-81]，属于标记不准确的典型场景。另外，对于具有一定结构的数据，标记可能只能在较粗的粒度提供。例如，一张图片中可以含有多个物体，但是对图像类别的标记可能只在图片的级别提供而没有精确到物体；对于细胞功能的标注通常在细胞级别，而这些功能通常是由细胞遗传物质中少量的基因引起的。这种场景通常被称作多示例学习[82]，即标记信息提供在样本包的级别，而一个样本包内对应多个样本。相较于完全、准确、确切的标记信息，这些从不同方面弱化的标记信息为学习带来了不同程度的困难。为了克服这些困难，许多研究者致力于设计算法以不同种类的弱标记信息作为监督进行学习，通常被统称为弱监督学习[83]。

在本文的前两章中，我们提出了半监督 AUC 优化方法和在线优化方法，属于数据标记 “不完全” 中的典型场景。在本章节中，我们进一步考虑数据标记不完全和不准确同时出现时的 AUC 优化问题。这一问题在标注难度高、耗时久的任务上比较常见，例如医疗影像的标注等。为了解决这一问题，需要对不同种类的弱标记带来的影响进行综合考虑。目前，只有少量工作尝试对多种弱监督学习场景进行综合考虑[84-86]：SAFEW[84]关注弱监督学习任务的安全性方面。CEGE[86]提出了一个基于质心估计的弱监督学习通用框架。然而，这些弱监督综合性研究主要面向最优化模型分类准确率。由于 AUC 优化依赖于样本对计算损失，面向分类准确率的弱监督学习方法不能用于优化模型的 AUC 性能。目前学

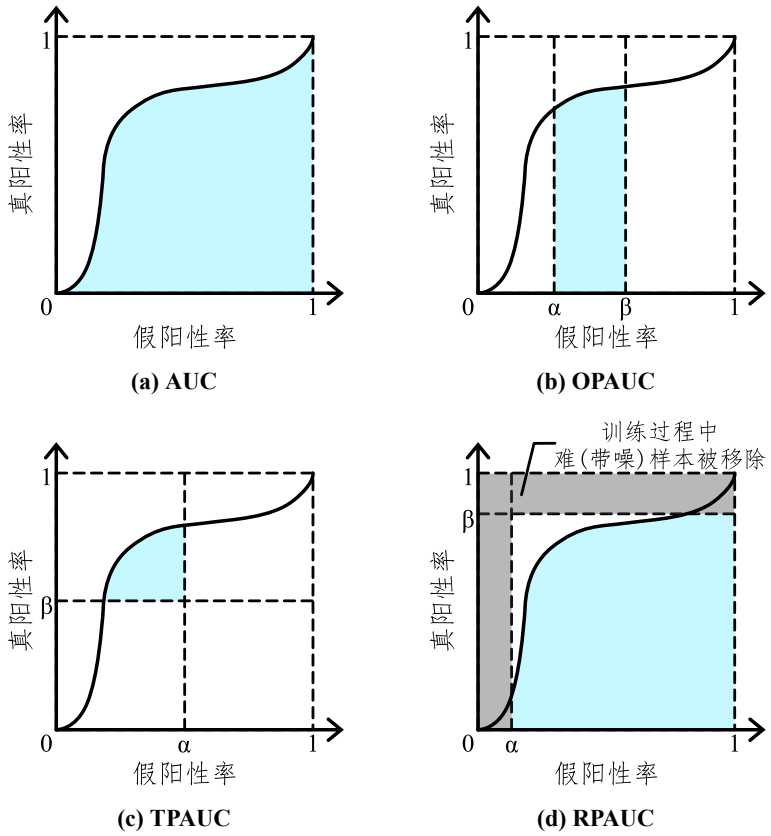


图 4-1 AUC 与各种部分 AUC 变种

术界对于训练 AUC 优化模型时应该如何应对“不准确”、“不确切”标记信息的研究仍然相对匮乏。因此，关于如何利用不同类型的弱标记数据进行 AUC 优化需要针对性研究。

在本章节中，我们提出了 WSAUC（Weakly Supervised AUC Optimization），一个弱监督 AUC 优化问题的统一框架。该框架对多种弱监督场景下的 AUC 优化任务提供了一种统一的解决方案，覆盖了包括标记带噪的学习场景、正标记-无标记学习场景、多示例学习场景和半监督学习场景等。该框架通过适当构造经验风险最小化（ERM）问题，将不同类型的弱监督 AUC 风险与真实的 AUC 风险进行对齐。各种类型的弱监督被转化为数据标记中不同级别的混杂形式，使得在同一个优化框架下解决这些问题成为可能。

此外，WSAUC 框架使用一种新类型的部分 AUC（partial AUC，pAUC）以实现在弱监督的条件下进行稳健的 AUC 优化，本文将其命名为反转部分 AUC（reversed partial AUC，简称 rpAUC），如图 4-1d 所示。本章节将展示最小化经验 rpAUC 风险（或最大化 rpAUC）与在噪声标记学习中常用的稳健训练方法

表 4-1 各种弱监督场景下 AUC 风险与有监督 AUC 风险 R_{PN} 的转化方式

场景	样本集	R	a	b	风险定义
有监督 AUC 优化	$\mathcal{X}_P, \mathcal{X}_N$	R_{PN}	1	0	公式 2-5
标记带噪的 AUC 优化	$\mathcal{X}_{\tilde{P}}, \mathcal{X}_{\tilde{N}}$	$R_{\tilde{P}\tilde{N}}$	$1-\eta_P-\eta_N$	$(\eta_P+\eta_N)/2$	公式 4-4
正标记-无标记 AUC 优化	$\mathcal{X}_P, \mathcal{X}_U$	R_{PU}	π_N	$\pi_P/2$	公式 4-6
多示例 AUC 优化	$\mathcal{X}_{\tilde{P}}, \mathcal{X}_N$	$R_{\tilde{P}N}$	$1-\eta_P$	$\eta_P/2$	公式 4-8
半监督 AUC 优化	$\mathcal{X}_P, \mathcal{X}_U, \mathcal{X}_N$	R_{PNU}	1	0 (修正后)	公式 4-13
标记带噪半监督 AUC 优化	$\mathcal{X}_{\tilde{P}}, \mathcal{X}_U, \mathcal{X}_{\tilde{N}}$	$R_{\tilde{P}\tilde{N}U}$	$1-\eta_P-\eta_N$	$(\eta_P+\eta_N)/2$	公式 4-14
统一风险形式化	$R = aR_{PN} + b$				

如小损失技巧（small loss trick）[87-89]和动态阈值技术[90]等具有类似的机制，证实 rpAUC 可以用作稳健的弱监督 AUC 优化的优化目标。该方法可以与 AUC 优化的 ERM 问题进行结合，并可以通过简单修改现有的 pAUC 优化算法进行简单实现。基于 rpAUC，WSAUC 框架提供了在各种弱监督场景下进行 AUC 优化的统一解决方案，回答了标记信息进一步减弱时应如何构建 AUC 模型的问题。

WSAUC 框架为如何从多种不同种类弱标记数据构建 AUC 优化模型提供了一种综合性的理解。由于将不同种类的弱标记 AUC 优化问题进行了统一，该框架可以直接应用于数据标记不完全可见且不准确的 AUC 优化问题。实验在包括该场景的多种弱监督场景将 WSAUC 框架和不同方法进行了对比，结果显示出 WSAUC 能够在不同的弱监督场景实现稳健的 AUC 优化效果。

4.2 弱监督 AUC 优化框架 WSAUC

本节介绍弱监督 AUC 优化框架 WSAUC。我们首先在第 4.2.1 小节给出弱监督 AUC 优化的统一形式，并在第 4.2.2 小节中展示这种统一形式如何特化到多种弱监督 AUC 优化场景中。这种统一形式可以表达为从两个正负类别混杂的数据集中进行 AUC 优化。通过将不同种类的弱标记信息归纳到同一种形式，标记不准确不完全可见的场景下的 AUC 优化问题可以在该框架下被解决。接下来，第 4.2.3 小节将介绍一种基于部分 AUC 的稳健优化方法，可以减轻统一形式下的弱监督 AUC 优化问题中样本混杂的问题，实现稳健的 AUC 优化效果。在第 4.3 节中，我们将对统一形式的弱监督 AUC 优化问题进行理论分析。表 4-1 中展示了 WSAUC 框架所能覆盖的各种弱监督 AUC 优化场景。

4.2.1 优化问题的形式化统一

本文提出，弱监督 AUC 优化问题可以被统一表示为一种从具有不同类别比例的两个混合样本集中进行 AUC 优化的形式。这种形式类似于无标记-无标记学习[91]，是一种依靠尽可能少的监督信息建立分类模型的学习范式。该统一形式下，设想我们有两个混合样本集 $\mathcal{X}_A$ 和 $\mathcal{X}_B$，其可以被视为分别采样于正负分布不同比例的混合分布：

$$\begin{aligned}\mathcal{X}_A &:= \{\boldsymbol{x}_i\}_{i=1}^{n_A} \overset{\text{i.i.d.}}{\sim} p_A(\boldsymbol{x}) := \theta_A p_P(\boldsymbol{x}) + (1-\theta_A) p_N(\boldsymbol{x})\,;\\ \mathcal{X}_B &:= \{\boldsymbol{x}'_j\}_{j=1}^{n_B} \overset{\text{i.i.d.}}{\sim} p_B(\boldsymbol{x}) := \theta_B p_P(\boldsymbol{x}) + (1-\theta_B) p_N(\boldsymbol{x})\,,\end{aligned}$$

不失一般性，此处假设 $\theta_A > \theta_B$。在分布 p_A 和 p_B 上定义的 AUC 风险为：

$$R_{AB}(f) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_A(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_B(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x},\boldsymbol{x}'))]\right]. \tag{4-1}$$

通过解决经验风险最小化（ERM）问题可以得到模型：

$$\min_f \quad \hat{R}_{AB}(f) = \frac{1}{|\mathcal{X}_A||\mathcal{X}_B|}\sum_{\boldsymbol{x}\in\mathcal{X}_A}\sum_{\boldsymbol{x}'\in\mathcal{X}_B}\ell(f(\boldsymbol{x},\boldsymbol{x}'))\,. \tag{4-2}$$

上述优化问题忽视了样本集合采样于不纯净的混合分布，类似于将两个集合分别视为正负样本集，进行有监督的 AUC 优化（有监督的 AUC 优化形式参见第二章）。然而，可以证明以上由混合分布计算的 AUC 风险通过线性变换可以被转化为真实的 AUC 风险。

定理 4.1 (统一形式化) *在两个混合分布 p_A 和 p_B 上的混合 AUC 风险 R_{AB} 可以被重写为如下统一形式：*

$$R_{AB} = aR_{PN} + b\,, \tag{4-3}$$

其中偏差项 $b = {}^{(1-a)}/_{2}$。基于该公式，真实风险 R_{PN} 可以通过对 R_{AB} 进行线性变换得到。

证明 对于任意 f，可以重写风险如下：

$$
\begin{aligned}
& R_{AB}(f) \\
= & \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_A}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_B}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] \\
= & \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_A}\left[(1-\theta_B)\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]+\theta_B\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_P}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] \\
= & \theta_A(1-\theta_B)\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] \\
& \quad + \theta_A\theta_B\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_P}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] \\
& \quad + (1-\theta_A)(1-\theta_B)\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_N}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] \\
& \quad + (1-\theta_A)\theta_B\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_N}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_P}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] \\
= & \theta_A(1-\theta_B)R_{PN}(f)+\theta_A\theta_B R_{PP}(f) \\
& \quad + (1-\theta_A)(1-\theta_B)R_{NN}(f)+(1-\theta_A)\theta_B R_{NP}(f) \\
= & \theta_A(1-\theta_B)R_{PN}(f)+(1-\theta_A)\theta_B(1-R_{PN}) \\
& \quad + \frac{1}{2}(2\theta_A\theta_B+1-\theta_A-\theta_B) \\
= & (\theta_A-\theta_B)R_{PN}(f)+\frac{1-(\theta_A-\theta_B)}{2}\,.
\end{aligned}
$$

因此原命题得证。 □

推论 4.1 (混合 AUC 风险的一致性) 在上述 *AUC* 优化问题中，假设 f^* 是在两个混合分布 p_A 和 p_B 上最小化混合 *AUC* 风险 R_{AB} 的解，即 $f^*=\operatorname{argmin} R_{AB}$，那么 f^* 也使得真实 *AUC* 风险 R_{PN} 取得最小值，即 R_{AB} 与 R_{PN} 是一致的。

证明 根据定理 *4.1*，对于任意 f 有：

$$
R_{AB}(f)=aR_{PN}(f)+\frac{1-a}{2}\,,
$$

其中 $a>0$。因此，对于任意 f，由于 $a>0$，且 f^* 使 R_{AB} 取得最小值，有下式成立：

$$
R_{PN}(f^*)-R_{PN}(f)=\frac{R_{AB}(f^*)-R_{AB}(f)}{a}\leqslant 0
$$

故，得证 f^* 也使 R_{PN} 取得最小值。 □

本节的剩余部分将给出将这个统一形式化转化为不同的弱监督 AUC 优化问题。将不同的弱监督 AUC 优化任务转化成通用形式求解提供了一种通用的基础解决方案。值得注意的是，在优化分类正确率的时候，需要知道分布的混合比例（即 θ_A 和 θ_B）以纠正估计量的偏差以实现一致性。然而，在 AUC 优化中，即使不知道混合比例，上述风险也可以保持统计一致性。

然而，混合分布的存在仍然会对仅用有限数据的模型学习产生影响，这在公式 4-3 中的系数 a 中有所体现。当样本集中存在高比例的混合时，可以视作标记中出现了大量噪声，直接解决上述优化问题可能不够稳健。在第 4.2.3 小节中，我们通过最小化一种新型的部分 AUC（rpAUC）的经验风险来解决这个问题。

4.2.2 各种弱监督场景的转化方法

标记带噪的 AUC 优化 本节首先讨论标记带噪的 AUC 优化[92]，这是不准确监督场景中最常见的情况，即样本的标记可能会发生错误。这种问题经常会在标注难度较高的任务、或通过众包对数据进行标注等场合[93-101]。

考虑带有标记噪声的二分类任务，每个样本的标记可能以一定的概率被翻转。考虑非对称噪声，即一个正样本可能以概率 η_P 被错误标记为负，同时一个负样本可能以概率 η_N 被错误标记为正。这样的标记带噪的 AUC 优化问题可以通过将噪声比例设置为混合比例，轻松转化为公式 4-1 中定义的问题：

$$\begin{aligned}\mathcal{X}_{\tilde{P}} &:= \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_{\tilde{P}}(\boldsymbol{x}) := (1-\eta_P)p_P(\boldsymbol{x}) + \eta_P p_N(\boldsymbol{x})\,;\\ \mathcal{X}_{\tilde{N}} &:= \{\boldsymbol{x}'_j\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_{\tilde{N}}(\boldsymbol{x}) := \eta_N p_P(\boldsymbol{x}) + (1-\eta_N)p_N(\boldsymbol{x})\,.\end{aligned}$$

标记带噪的学习问题可以看作是两个从混合集合学习或无标记-无标记学习的一种变体。这两个问题表述之间的区别在 Lu 等人[91] 中进行了讨论。具体来说，标记带噪的学习任务假设噪声率 $\eta_P + \eta_N < 0.5$，而类先验概率在有或没有噪声的情况下都保持不变。如果没有这个假设，边缘分布 $p(\boldsymbol{x})$ 可能会发生改变，这就需要在协变量偏移（covariate shift）假设下解决这个问题。

通过在公式 4-1 中简单地替换两个噪声分布 $\mathcal{X}_A$ 和 $\mathcal{X}_B$，就可以得到带噪 AUC 风险：

$$R_{\tilde{P}\tilde{N}}(f) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_{\tilde{P}}(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_{\tilde{N}}(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x},\boldsymbol{x}'))]\right]. \tag{4-4}$$

通过令 $\theta_A = 1 - \eta_P$ 和 $\theta_B = \eta_N$，可以很容易地证明以下推论：

推论 4.2 *优化 $R_{\tilde{P}\tilde{N}}$ 与优化真实分布的 AUC 风险 R_{PN} 一致，且满足：*

$$R_{\tilde{P}\tilde{N}} = \tilde{a} R_{PN} + \frac{1-\tilde{a}}{2},$$
$$\tilde{a} = 1 - \eta_P - \eta_N .$$

这表明带噪 AUC 风险仍然与干净 AUC 风险保持一致。实践中，可以解决以下 ERM 问题来构建模型：

$$\min_f \quad \hat{R}_{\tilde{P}\tilde{N}}(f) = \frac{1}{|\mathcal{X}_{\tilde{P}}||\mathcal{X}_{\tilde{N}}|} \sum_{\boldsymbol{x}\in\mathcal{X}_{\tilde{P}}} \sum_{\boldsymbol{x}'\in\mathcal{X}_{\tilde{N}}} \ell(f(\boldsymbol{x}, \boldsymbol{x}')) . \tag{4-5}$$

正标记-无标记 AUC 优化 接下来讨论仅具有一个类别监督场景：正标记-无标记 AUC 优化[102]。这种场景通常发生在某一类的数据标记比较容易识别，而另一类的数据标记不易识别的情况[103-111]。例如，在欺诈检测中，少部分的交易被投诉为欺诈交易，而未被投诉的交易记录中通常既有正常交易，也有未被发现的欺诈交易。因此，被投诉的交易可以被视为正样本，而其他交易记录可以被视为无标记样本。

假设无标记数据中正类和负类的潜在类先验概率分别为 π_P 和 π_N，在这种情况下，具有正标记的样本可以被视为具有 $\theta_A = 1$ 的纯正样本集合 $\mathcal{X}_P$，而无标记数据则由 $\theta_B = \pi_P$ 的混合集合 $\mathcal{X}_U$ 组成。这两个集合可以被表示为：

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}) ,$$
$$\mathcal{X}_U := \{\boldsymbol{x}'_j\}_{j=1}^{n_U} \overset{\text{i.i.d.}}{\sim} p_U(\boldsymbol{x}) := \pi_P p_P(\boldsymbol{x}) + \pi_N p_N(\boldsymbol{x}) .$$

则正样本和无标记样本集上可以定义 P-U AUC 风险：

$$R_{PU}(f) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_P(\boldsymbol{x})} \left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_U(\boldsymbol{x})} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right] . \tag{4-6}$$

基于该定义，很容易证明如下推论。

推论 4.3 *优化 P-U AUC 风险 R_{PU} 与优化真实的 AUC 风险 R_{PN} 一致，且：*

$$R_{PU} = \pi_N R_{PN} + \frac{\pi_P}{2}.$$

这种形式将正标记-无标记样本的 AUC 优化问题与混合分布的情况建立了联系，即将一个集合视为具有最小的混合率，将另一个集合视为具有最极端的混合比率（与边际分布 $p(\boldsymbol{x})$ 一样混杂）。这种形式表明，通过将无标记数据视为负数据，可以解决具有单侧监督的 AUC 优化问题。在实际应用中，可以通过最小化以下经验风险来解决正标记-无标记的 AUC 优化问题：

$$\min_f \quad \hat{R}_{PU}(f) = \frac{1}{|\mathcal{X}_P||\mathcal{X}_U|} \sum_{\boldsymbol{x}\in\mathcal{X}_P} \sum_{\boldsymbol{x}'\in\mathcal{X}_U} \ell(f(\boldsymbol{x},\boldsymbol{x}')). \tag{4-7}$$

多示例 AUC 优化 多示例 AUC 优化是一种具有不确切监督信息的学习问题。在多示例学习场景中，训练样本被组织成样本包（bag），而标记只在包级别给出。例如，一张图片中可以含有多个物体，但是对图像类别的标记通常只在图片的级别提供而没有精确到物体；对于细胞功能的标注通常在细胞级别，而这些功能通常是由细胞遗传物质中大量基因中的少数引起的。对于多示例学习而言，样本级方法通常假设包的标记取决于是否存在任何正样本，即正样本包中至少存在一个正样本，负样本包中则均为负样本[112-115]。包级方法则假设包的标记由多个概念或样本共同定义，而不取决于单个样本的存在与否[116-119]。在本文中，我们采用了样本级多示例学习的假设。

形式化地讲，在该场景中，用户有一组正例包 $S_P = \{B_i^+\}_{i=1}^{N_P}$ 和一组负例包 $S_N = \{B_i^-\}_{j=1}^{N_N}$。每个正例包 B_i^+ 中至少包含一个正样本，而负例包 B_j^- 则不包含任何正样本。按照现有研究[86]，可以将出现在负例包中的样本视为从纯负分布 p_N 中抽样，而出现在正例包中的样本视为从某个未知比例 $1-\eta_P$ 和 η_P 的正负分布混合分布 $p_{\tilde{P}}$ 中抽样。然后，可以定义在正例包样本和负例包样本上的 AUC 风险：

$$R_{\tilde{P}N}(f) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_{\tilde{P}}(\boldsymbol{x})} \left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_N(\boldsymbol{x})} [\ell_{01}(f(\boldsymbol{x},\boldsymbol{x}'))] \right]. \tag{4-8}$$

基于该多示例 AUC 风险定义，可以证明推论 4.1 在多示例 AUC 优化问题下的推论，陈述如下：

推论 4.4 *优化多示例 AUC 风险 $R_{\tilde{P}N}$ 与优化真实的 AUC 风险 R_{PN} 一致，且满足下列等式:*

$$R_{\tilde{P}N} = (1-\eta_P)R_{PN} + \frac{\eta_P}{2}\,.$$

上述推论为我们提供了一种处理多示例学习中 AUC 优化的方法。实践中，需要首先将样本包的并集构建为样本集 $\mathcal{X}_{\tilde{P}}$ 和 $\mathcal{X}_N$:

$$\mathcal{X}_{\tilde{P}} = \bigcup_{i=1}^{N_P} B_i^+\,, \quad \mathcal{X}_N = \bigcup_{j=1}^{N_N} B_j^-\,, \tag{4-9}$$

然后求解以下 ERM 问题:

$$\min_f \quad \hat{R}_{\tilde{P}N}(f) = \frac{1}{|\mathcal{X}_{\tilde{P}}||\mathcal{X}_N|}\sum_{\boldsymbol{x}\in\mathcal{X}_{\tilde{P}}}\sum_{\boldsymbol{x}'\in\mathcal{X}_N}\ell(f(\boldsymbol{x},\boldsymbol{x}'))\,. \tag{4-10}$$

通过这种方法，模型可以输出样本级别的排序得分。要获得包级别的排序得分，可以简单地计算包中样本得分的最大值。

半监督 AUC 优化 半监督学习是一种常见的标记信息不完全的弱监督学习场景，通常使用有限标记数据和相对较多的无标记数据构建模型。这一类学习方法主要针对标记收集困难，但是数据获取相对容易的学习任务，通过无标记辅助学习，达到比仅仅使用少量有标记数据更好的学习效果。

在半监督场景下，与之前的情况不同，数据可以根据其标记分为三个集合:

$$\begin{aligned}
\mathcal{X}_P &:= \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x})\,,\\
\mathcal{X}_N &:= \{\boldsymbol{x}'_j\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_N(\boldsymbol{x})\,,\\
\mathcal{X}_U &:= \{\boldsymbol{x}''_k\}_{k=1}^{n_U} \overset{\text{i.i.d.}}{\sim} p_U(\boldsymbol{x}) := \pi_P p_P(\boldsymbol{x}) + \pi_N p_N(\boldsymbol{x})\,.
\end{aligned}$$

对于无标记和负样本数据，我们可以对 P-U AUC 风险 (公式 4-6) 进行对称定义，得到 U-N AUC 风险:

$$R_{UN}(f) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_U(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_N(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x},\boldsymbol{x}'))]\right]\,. \tag{4-11}$$

可以证明，通过将 P-U AUC 风险和 U-N AUC 风险相加可以得到一个无偏的风险估计量。即使类先验概率是未知的，风险相加后其偏差总是常数，因此可以被修正[79]。

定理 4.2 *优化 R_{PU} 和 R_{UN} 的和与优化真实的 AUC 风险 R_{PN} 一致，偏差始终为 $\frac{1}{2}$，即：*

$$R_{PU} + R_{UN} - \frac{1}{2} = R_{PN}\,. \tag{4-12}$$

证明 *对于任何 f，有下式成立：*

$$\begin{aligned}
&R_{PU}(f) + R_{UN}(f)\\
=&\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_U}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_U}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right]\\
=&\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\pi_N\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right] + \pi_P\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_P}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right]\\
&+\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_N}\left[\pi_P\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_P}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right] + \pi_N\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_N}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right]\\
=&(\pi_P+\pi_N)R_{PN}(f) + \pi_P R_{PP}(f) + \pi_N R_{NN}(f)\\
=&R_{PN}(f) + \frac{1}{2}\,.
\end{aligned}$$

因此原命题得证。 □

这也表明，在半监督情况下，可以通过减去 $\frac{1}{2}$ 来实现无偏的 AUC 风险估计，而不需要知道类先验。

为了充分利用数据以减少估计方差，可以优化以下风险，而不是直接通过 ERM 最小化公式 4-12：

$$R_{PNU} = \gamma R_{PN} + (1-\gamma)(R_{PU} + R_{UN} - \frac{1}{2})\,, \tag{4-13}$$

其中 γ 是加权系数。

为了计算经验风险 $\hat{R}_{PNU}$，需要对三个数据集对 $\mathcal{X}_P \times \mathcal{X}_N$、$\mathcal{X}_P \times \mathcal{X}_U$ 和 $\mathcal{X}_U \times \mathcal{X}_N$ 中的成对损失进行求和。$(R_{PU} + R_{UN})$ 引起的偏差始终为 1/2，可以通过从经验风险中减去它来进行补偿。经过补偿，经验风险 $\hat{R}_{PNU}$ 就成为了真

实 AUC 风险的无偏估计量。实践中，是否补偿偏差并不影响模型的训练，所以也可以直接忽略偏差项。

标记带噪的半监督 AUC 优化 最后，基于统一形式化，可以进一步考虑可用监督信息既不准确又不完全的情况，即拥有少量的标记带噪的样本和大量的无标记样本的场景。该场景是最为复杂的一种弱监督学习场景。例如，基于缺陷报告的缺陷检测任务属于这种情况[85]。

设 η_P 和 η_N 分别表示正标记和负标记不正确的概率，π_P 和 π_N 表示真实分布的类先验概率，其中 $\pi_P + \pi_N = 1$。值得注意的是，在实际学习中，并不需要知道这些变量的具体取值。在此条件下，样本可以被分类为三个集合：一个带噪的正样本集 $\mathcal{X}_{\tilde{P}} \sim p_{\tilde{P}}$，一个带噪的负样本集 $\mathcal{X}_{\tilde{N}} \sim p_{\tilde{N}}$，和一个无标记的集合 $\mathcal{X}_U \sim p_U$。

与前述情况类似，通过结合带噪的 P-U AUC 风险和 U-N AUC 风险，有偏的风险估计量会产生与标记带噪情况下相同的系数。

推论 4.5 *优化 $R_{\tilde{P}U}$ 和 $R_{U\tilde{N}}$ 的和与优化真实的 AUC 风险 R_{PN} 一致，且满足：*

$$R_{\tilde{P}U} + R_{U\tilde{N}} - \frac{1}{2} = \tilde{a} R_{PN} + \frac{1-\tilde{a}}{2},$$
$$\tilde{a} = 1 - \eta_P - \eta_N .$$

为了降低风险估计量的方差，定义以下风险：

$$R_{\tilde{P}\tilde{N}U} = \gamma R_{\tilde{P}\tilde{N}} + (1-\gamma)(R_{\tilde{P}U} + R_{U\tilde{N}} - \frac{1}{2}) . \tag{4-14}$$

该优化问题的求解方式则与普通半监督 AUC 优化相同。

小节 本小节展示了可以通过最小化一个或多个 AUC 风险项的总和或加权平均来解决各种弱监督 AUC 优化问题。这一类优化问题都可以通过将来自不同集合的样本对合并到基于样本对的 AUC 优化算法来求解。各种不同的弱监督 AUC 优化风险在表 4-1 中作出了总结。本章接下来的内容将为这些问题提供稳健的学习解决方案。在第 4.3 节中，将介绍该框架的理论分析。

4.2.3 基于 rpAUC 的稳健 AUC 优化

在前面的小节中，我们从两个混合的样本集合的 AUC 优化问题的角度，提出了一种弱监督 AUC 优化问题的统一形式。该发现提供了一种在各种弱监督场景下解决 AUC 优化问题的统计一致的方法。然而，对于有限的训练数据，仅通过上述 ERM 问题来训练模型仍然会受到由于数据集混杂带来的负面影响，这种影响反映在超额风险界中的系数中（参见定理 4.3 中的 a 和定理 4.4 中的 $\tilde{a}$）。

为了缓解这个问题，在本节中，将提出一种新的部分 AUC，即 “反转部分 AUC”（rpAUC）。本文将会证明通过在混杂数据集上通过 ERM 实现最大化 rpAUC，WSAUC 可以实现稳健的 AUC 优化。接下来，我们首先给出 rpAUC 的定义，它通过反转双向部分 AUC 的限制得到。然后，本文将展示 rpAUC 优化与 “小损失技巧（small-loss trick）” 之间的等价性，这是一种常用于噪声标记学习的技术[87-89,120]。基于这种等价性，可以得到一种简单的弱监督 AUC 优化解决方案，即通过将训练过程中最小化全 AUC 风险替换为部分 AUC 风险来解决弱监督 AUC 优化问题。

目前，有两种部分 AUC 变体，名为单向部分 AUC（One-way Partial AUC，OPAUC）[26]和双向部分 AUC（Two-way Partial AUC，TPAUC）[27]。TPAUC 考虑 ROC 曲线下 TPR 在 $[\alpha, \beta]$ 区间中的面积，OPAUC 考虑 ROC 曲线下 $\mathrm{FPR} < \alpha$ 且 $\mathrm{TPR} > \beta$ 部分的面积。部分 AUC 并非关注所有样本的排序质量，而只关心特定 FPR 或 TPR 区间的模型排序质量。这使得部分 AUC 能够针对排序列表的某一具体部分进行优化，对于有这类特定要求的应用场景更加适用。例如，对于网页检索任务，往往只关注其头部排序的质量，而对后面的内容关注度较低。

通过对 TPAUC 稍加改动，即翻转其对 TPR 和 FPR 的约束限制，本文提出双向反转部分 AUC（rpAUC），其定义如下。

定义 4.1 *对于模型 f，双向反转部分 AUC（rpAUC）在 FPR 阈值为 α 和 TPR 阈值为 β 时如下定义:*

$$\begin{aligned}
\mathrm{rpAUC}(f;\alpha,\beta) &= 1 - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_A^+(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_B^-(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x},\boldsymbol{x}'))]\right], \\
p_A^+(\boldsymbol{x}) &= p_A(\boldsymbol{x}|f(\boldsymbol{x}) \in [\mathrm{TPR}_f^{-1}(\beta), \infty)), \\
p_B^-(\boldsymbol{x}) &= p_B(\boldsymbol{x}|f(\boldsymbol{x}) \in (-\infty, \mathrm{FPR}_f^{-1}(\alpha)]).
\end{aligned}$$

算法 4 基于 rpAUC 的稳健 WSAUC 优化方法

Input: 纯净或混合的数据集 $\{\mathcal{X}_P[,\mathcal{X}_U][,\mathcal{X}_N]\}$；超参数 α，β

初始化模型 f
for $t = 1 \to T$ **do**
  for $k = 1 \to K$ **do**
    for 对于每一对数据集 $(\mathcal{X}_A, \mathcal{X}_B)$ **do**
      根据比例 α / β 筛除样本
      采样小批量 $B_A \in \mathcal{X}_A^+, B_B \in \mathcal{X}_B^-$
      在小批量上计算 rpAUC 风险 $\hat{R}_{AB}^{(\mathrm{rp})}(f)$
      通过反向传播更新模型 f
    end for
  end for
end for

Output: 模型 f

rpAUC 的示意图如图 4-1d 所示。通过反转 TPR 和 FPR 的约束条件，rpAUC 去掉了 ROC 曲线的最左边和最上面的边缘。根据定义，这相当于消除具有最低排序分数的 β 比例的正样本和具有最高排序分数的 α 比例的负样本。不同于以往两种部分 AUC 是为了适应特定学习任务对样本排序的要求，在训练过程中最大化 rpAUC 实际上等同于在最大化 AUC 时去除了那些产生了最大损失的样本。为了说明这一点，下面证明引起最大损失的样本是那些在 $\mathcal{X}_A$ 中具有最低分数和在 $\mathcal{X}_B$ 中具有最高分数的样本。

样本损失可以定义如下：

$$L(x) = \begin{cases} \frac{1}{|\mathcal{X}_B|} \sum_{\boldsymbol{x}' \in \mathcal{X}_B} \ell(f(\boldsymbol{x}, \boldsymbol{x}')), & \text{if } \boldsymbol{x} \in \mathcal{X}_A, \\ \frac{1}{|\mathcal{X}_A|} \sum_{\boldsymbol{x}' \in \mathcal{X}_A} \ell(f(\boldsymbol{x}', \boldsymbol{x})), & \text{if } \boldsymbol{x} \in \mathcal{X}_B. \end{cases}$$

假设替代损失函数 $l(z)$ 单调非增，则很容易证明 $L(x)$ 与 $f(x)$ 单调变化。因此，可以证明以下命题。

命题 4.1 *假设模型 f 是一个评分函数，则其 rpAUC 风险 $\hat{R}_{AB}^{(\mathrm{rp})}(f;\alpha,\beta)$ 等于在 $\mathcal{X}_A$ 中消除了 $L(x)$ 产生了最大损失的 β 比例样本和在 $\mathcal{X}_B$ 中产生了最大损失的 β 比例样本后的全 AUC 风险 $\hat{R}_{AB}(f)$。*

这个命题揭示了最小化经验 rpAUC 风险等同于在标记带噪的场景中选择干净标记的样本进行学习，后者是标记带噪学习中的一种常用手段。在训练阶段，

最小化以下经验 rpAUC 风险：

$$\hat{R}_{AB}^{(\mathrm{rp})}(f;\alpha,\beta)=\frac{1}{|\mathcal{X}_A^+||\mathcal{X}_B^-|}\sum_{\boldsymbol{x}\in\mathcal{X}_A^+}\sum_{\boldsymbol{x}'\in\mathcal{X}_B^-}\ell(f(\boldsymbol{x},\boldsymbol{x}'))\,, \tag{4-15}$$

其中，$\mathcal{X}_A^+$ 是 $\mathcal{X}_A$ 中具有前 $\lfloor(1-\beta)|\mathcal{X}_A|\rfloor$ 个最高分数的样本的集合，$\mathcal{X}_B^-$ 是 $\mathcal{X}_B$ 中具有最低 $\lfloor(1-\alpha)|\mathcal{X}_B|\rfloor$ 个分数的样本的集合。这可以通过应用任何现有的双向 pAUC 优化算法并反转样本选择来实现，例如文献[26,28-29]中涉及到的优化算法。算法 4 展示了一种 rpAUC 优化过程的实现参考。

需要注意的是，虽然 rpAUC 与 TPAUC 的定义相似，但使用目的完全不同。采用 TPAUC 作为优化目标时，用户只对部分 TPR 和 FPR 区间内的排序性能感兴趣。此时，超参数 α 和 β 根据用户对于模型的要求设置。而用 rpAUC 作为优化目标时，用户仍然对整个区间上的 AUC 指标感兴趣，优化其中的一部分是为了避免弱标记带来的负面影响。此时，超参数 α 和 β 的选择以最大化测试数据上的 AUC 为目的，其最优值跟数据标记的质量和模型拟合能力等因素相关。

4.3 理论分析

本节以标记信息不准确不完全可见的场景为代表，从理论上对所提出的弱监督 AUC 风险进行分析：(1) 证明了 WSAUC 在标记信息不准确不完全可见情况下的超额风险界（excess risk bound），这些界可以应用于上面讨论的所有弱监督 AUC 优化问题，保障了上述方法的泛化性能；(2) 讨论了 WSAUC 在标记信息不准确不完全可见情况下的方差缩减，表明引入无标记数据可以降低方差，实现更准确的风险估计。

考虑 K 是 $\mathcal{X}^2$ 上的一个核函数，C_w 是一个正实数，令 $\mathcal{F}_K$ 表示以下定义在样本空间上的函数族：

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

其中 $\|x\|_K=\sqrt{K(x,x)}$。本节中假设替代损失函数 ℓ 是 L-Lipschitz 连续的，上界为正实数 C_ℓ，并满足不等式 $\ell\geqslant\ell_{01}$。例如，使用平方损失和指数损失作为替代损失函数即可满足这些条件。

4.3.1 超额风险

记最小化混合风险 $\hat{R}_{AB}(f)$ 的模型为 $\hat{f}_{AB}^{*}$，本文给出以下超额风险界，表明 $\hat{f}_{AB}^{*}$ 的风险收敛于函数族 $\mathcal{F}_K$ 中最优模型的风险。

定理 4.3 (统一形式下的超额风险) 设 $\hat{f}_{AB}^{*} \in \mathcal{F}_K$ 是经验风险 $\hat{R}_{AB}(f)$ 的最小值点，$f_{PN}^{*} \in \mathcal{F}_K$ 是真实风险 $R_{PN}(f)$ 的最小值点。对于任意 $\delta > 0$，下式以至少 $1-\delta$ 的概率成立：

$$R_{PN}(\hat{f}_{AB}^{*}) - R_{PN}(f_{PN}^{*}) \leqslant \frac{h(\delta)}{a}\sqrt{\frac{n_A + n_B}{n_A n_B}},$$

其中 $h(\delta) = 8\sqrt{2}C_{\ell}C_w C_x + 5\sqrt{2\ln(2/\delta)}$，$n_A, n_B$ 是采样的混合样本集的大小。

证明 设 $R'_{AB}(f) = \frac{R_{AB} - \frac{1-a}{2}}{a}$ 表示对 R_{AB} 进行线性变换以估计 R_{PN}，$\hat{R}'_{AB}(f)$ 表示它的经验估计量。优化 $\hat{R}_{AB}(f)$ 的超额风险可以表示为

$$\begin{aligned} & R_{PN}(\hat{f}_{AB}^{*}) - R_{PN}(f_{PN}^{*}) \\ & \quad = R_{PN}(\hat{f}_{AB}^{*}) - \hat{R}'_{AB}(\hat{f}_{AB}^{*}) + \hat{R}'_{AB}(\hat{f}_{AB}^{*}) \\ & \qquad - \hat{R}'_{AB}(f_{PN}^{*}) + \hat{R}'_{AB}(f_{PN}^{*}) - R_{PN}(f_{PN}^{*}) \\ & \quad \leqslant 2\max_{f\in\mathcal{F}} |\hat{R}'_{AB}(f) - R_{PN}(f)|. \end{aligned} \tag{4-16}$$

根据定理 *4.1*，右侧项可以表示为

$$\max_{f\in\mathcal{F}} |\hat{R}'_{AB}(f) - R_{PN}(f)| = \max_{f\in\mathcal{F}} |\hat{R}'_{AB}(f) - R'_{AB}(f)|. \tag{4-17}$$

根据 *Usunier* 等人[65]中的定理 *6*，对于任意 $\delta > 0$，对于任何 $f \in \mathcal{F}_K$，至少有 $1-\delta$ 的概率：

$$\begin{aligned} & \max_{f\in\mathcal{F}} |\hat{R}_{AB}(f) - R_{AB}(f)| \\ & \leqslant 4\sqrt{2}C_{\ell}C_w C_x\sqrt{\frac{n_A+n_B}{n_A n_B}} + 5\sqrt{\frac{n_A+n_B}{2n_A n_B}\ln(2/\delta)}, \end{aligned} \tag{4-18}$$

其中 $C_x = \max\left(\max_i \|x_i\|, \max_j \|x'_j\|\right)$。简便起见，定义

$$h(\delta) = 8\sqrt{2}C_\ell C_w C_x + 5\sqrt{2\ln\left(2/\delta\right)}\,.$$

则有

$$\max_{f\in\mathcal{F}} |\hat{R}'_{AB}(f) - R'_{AB}(f)| \leqslant \frac{h(\delta)}{2a}\sqrt{\frac{n_A+n_B}{n_A n_B}}\,. \tag{4-19}$$

将公式 *4-17* 和不等式 *4-19* 应用于不等式 *4-16* 中的右侧项，可证明该定理。 □

定理 4.3 保证了统一形式的超额风险可以被界定，加上置信度项，其量级为

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

记 $\hat{f}^*_{\tilde{P}\tilde{N}U}$ 为令经验风险 $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ 取得最小值的模型，类似地，如下定理可以表明 $\hat{f}^*_{\tilde{P}\tilde{N}U}$ 的风险收敛于函数族 $\mathcal{F}_K$ 中最优模型的风险。

定理 4.4 (标记信息不准确且不完全可见情况的超额风险) 设 $\hat{f}^*_{\tilde{P}\tilde{N}U} \in \mathcal{F}_K$ 是令经验风险 $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ 最小的分类器，$f^*_{PN} \in \mathcal{F}_K$ 是令真实风险 $R_{PN}(f)$ 最小的分类器。对于任意 $\delta > 0$，至少以 $1-\delta$ 的概率下式成立：

$$\begin{aligned}
&R_{PN}(\hat{f}^*_{\tilde{P}\tilde{N}U}) - R_{PN}(f^*_{PN}) \\
&\leqslant \frac{h(\frac{\delta}{3})}{\tilde{a}}\left(\gamma\sqrt{\frac{n_{\tilde{P}}+n_{\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}} + (1-\gamma)\left(\sqrt{\frac{n_{\tilde{P}}+n_U}{n_{\tilde{P}}n_U}} + \sqrt{\frac{n_U+n_{\tilde{N}}}{n_U n_{\tilde{N}}}}\right)\right),
\end{aligned}$$

其中 $h(\delta) = 8\sqrt{2}C_\ell C_w C_x + 5\sqrt{2\ln\left(2/\delta\right)}$，$n_{\tilde{P}}, n_{\tilde{N}}$ 是采样的混合样本集的大小。

证明 设 $R'_{\tilde{P}\tilde{N}U}(f) = \gamma\frac{R_{\tilde{P}\tilde{N}} - \frac{1-\tilde{a}}{2}}{\tilde{a}} + (1-\gamma)\frac{R_{\tilde{P}U} + R_{U\tilde{N}} - \frac{1}{2} - \frac{1-\tilde{a}}{2}}{\tilde{a}}$ 表示对 $R_{\tilde{P}\tilde{N}U}$ 进行线性变换以估计 R_{PN}，$\hat{R}'_{\tilde{P}\tilde{N}U}(f)$ 表示它的经验估计量。与不等式 *4-16* 类似，优化 $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ 的超额风险可以表示为

$$\begin{aligned}
&R_{PN}(\hat{f}^*_{\tilde{P}\tilde{N}U}) - R_{PN}(f^*_{PN}) \\
&\quad= R_{PN}(\hat{f}^*_{\tilde{P}\tilde{N}U}) - \hat{R}'_{\tilde{P}\tilde{N}U}(\hat{f}^*_{\tilde{P}\tilde{N}U}) + \hat{R}'_{\tilde{P}\tilde{N}U}(\hat{f}^*_{\tilde{P}\tilde{N}U}) \\
&\qquad - \hat{R}'_{\tilde{P}\tilde{N}U}(f^*_{PN}) + \hat{R}'_{\tilde{P}\tilde{N}U}(f^*_{PN}) - R_{PN}(f^*_{PN}) \\
&\quad\leqslant 2\max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}\tilde{N}U}(f) - R_{PN}(f)|\,.
\end{aligned} \tag{4-20}$$

根据推论 *4.5*，右侧项可以表示为

$$\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}\tilde{N}U}(f)-R_{PN}(f)|=\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}\tilde{N}U}(f)-R'_{\tilde{P}\tilde{N}U}(f)|\,. \tag{4-21}$$

分别将不等式 *4-19* 中的 $x\in\mathcal{X}_A, x'\in\mathcal{X}_B$ 替换为 $x\in\mathcal{X}_{\tilde{P}}, x'\in\mathcal{X}_{\tilde{N}}$，或 $x\in\mathcal{X}_{\tilde{P}}, x'\in\mathcal{X}_U$，或 $x\in\mathcal{X}_U, x'\in\mathcal{X}_{\tilde{N}}$，则对于任意 $\delta>0$，以至少 $1-\delta$ 的概率，对任意 $f\in\mathcal{F}_K$ 下式成立：

$$\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}\tilde{N}}(f)-R'_{\tilde{P}\tilde{N}}(f)|\leqslant\frac{h(\delta)}{2\tilde{a}}\sqrt{\frac{n_{\tilde{P}}+n_{\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}}\,,$$

$$\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}U}(f)-R'_{\tilde{P}U}(f)|\leqslant\frac{h(\delta)}{2\tilde{a}}\sqrt{\frac{n_{\tilde{P}}+n_{U}}{n_{\tilde{P}}n_{U}}}\,,$$

$$\max_{f\in\mathcal{F}}|\hat{R}'_{U\tilde{N}}(f)-R'_{U\tilde{N}}(f)|\leqslant\frac{h(\delta)}{2\tilde{a}}\sqrt{\frac{n_{U}+n_{\tilde{N}}}{n_{U}n_{\tilde{N}}}}\,.$$

简单计算可得对于任意 $\delta'>0$ 以至少 $1-\delta'$ 的概率下式成立：

$$\begin{aligned}
&\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}\tilde{N}U}(f)-R'_{\tilde{P}\tilde{N}U}(f)|\\
\leqslant\;&\gamma\left(\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}\tilde{N}}(f)-R'_{\tilde{P}\tilde{N}}(f)|\right)\\
&+(1-\gamma)\left(\max_{f\in\mathcal{F}}|\hat{R}'_{\tilde{P}U}(f)-R'_{\tilde{P}U}(f)|\right)\\
&+(1-\gamma)\left(\max_{f\in\mathcal{F}}|\hat{R}'_{U\tilde{N}}(f)-R'_{U\tilde{N}}(f)|\right)\\
\leqslant\;&\frac{h(\frac{\delta'}{3})}{2\tilde{a}}\left(\gamma\sqrt{\frac{n_{\tilde{P}}+n_{\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}}+(1-\gamma)\left(\sqrt{\frac{n_{\tilde{P}}+n_{U}}{n_{\tilde{P}}n_{U}}}+\sqrt{\frac{n_{U}+n_{\tilde{N}}}{n_{U}n_{\tilde{N}}}}\right)\right).
\end{aligned} \tag{4-22}$$

将公式 *4-21* 和不等式 *4-22* 带入不等式 *4-20* 中的右侧项，定理得证。 □

定理 4.4 保证了不准确和不完全情况下的超额风险可以被界定，加上置信度项，其量级为

$$\mathcal{O}\left(\frac{1}{\tilde{a}\sqrt{n_{\tilde{P}}}}+\frac{1}{\tilde{a}\sqrt{n_{\tilde{N}}}}+\frac{1}{\tilde{a}\sqrt{n_{U}}}\right).$$

可以看出，对于 $\gamma=1$，定理 4.4 会退化为定理 4.3 。

4.3.2 方差缩减

前文证明了本章所提出的经验风险估计量是无偏的，并且超额风险收敛。本节接下来研究在标记信息不准确且不完全可见的情况下，即当 $\gamma < 1$ 时，$\hat{R}_{\tilde{P}\tilde{N}U}(f)$ 的方差是否可以小于 $\hat{R}_{\tilde{P}\tilde{N}}(f)$ 的方差，换言之，是否 $\mathcal{X}_U$ 可以帮助减少估计 R_{PN} 的方差。为回答这个问题，选择任意感兴趣的 f。为简单起见，假设 $n_U \to \infty$，以便展示方差缩减效果可能取得的上限。在标记既不精确又不完全可见的情况对应上一节中标记带噪的半监督 AUC 优化。在该场景下，风险估计量的方差与协方差的定义如下：

$$
\begin{aligned}
\sigma^2_{\tilde{P}\tilde{N}}(f) &= \mathrm{Var}_{\tilde{P}\tilde{N}}[\ell(f(x_{\tilde{P}}, x_{\tilde{N}}))]\,, \\
\sigma^2_{\tilde{P}U}(f) &= \mathrm{Var}_{\tilde{P}U}[\ell(f(x_{\tilde{P}}, x_U))]\,, \\
\sigma^2_{U\tilde{N}}(f) &= \mathrm{Var}_{U\tilde{N}}[\ell(f(x_U, x_{\tilde{N}}))]\,, \\
\tau_{\tilde{P}\tilde{N},\tilde{P}U}(f) &= \mathrm{Cor}_{\tilde{P}\tilde{N},\tilde{P}U}[\ell(f(x_{\tilde{P}}, x_{\tilde{N}})), \ell(f(x_{\tilde{P}}, x_U))]\,, \\
\tau_{\tilde{P}\tilde{N},U\tilde{N}}(f) &= \mathrm{Cor}_{\tilde{P}\tilde{N},U\tilde{N}}[\ell(f(x_{\tilde{P}}, x_{\tilde{N}})), \ell(f(x_U, x_{\tilde{N}}))]\,, \\
\tau_{\tilde{P}U,U\tilde{N}}(f) &= \mathrm{Cor}_{\tilde{P}U,U\tilde{N}}[\ell(f(x_{\tilde{P}}, x_U)), \ell(f(x_U, x_{\tilde{N}}))]\,.
\end{aligned}
$$

基于以上定义，可以证明以下定理。

定理 4.5 (标记信息不准确不完全可见情况的方差) *设 $n_U \to \infty$。对于任意固定的 f，经验风险 $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ 的方差的极小值点是*

$$
\gamma_{\tilde{P}\tilde{N}} = \arg\min_{\gamma} \mathrm{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)] = \frac{\psi_{\tilde{P}\tilde{N}U}}{\psi_{\tilde{P}\tilde{N}U} - \psi_{\tilde{P}\tilde{N}}}\,,
$$

其中

$$
\begin{aligned}
\psi_{\tilde{P}\tilde{N}} &= \frac{1}{n_{\tilde{P}} n_{\tilde{N}}} \sigma^2_{\tilde{P}\tilde{N}}(f)\,, \\
\psi_{\tilde{P}\tilde{N}U} &= \frac{1}{n_{\tilde{P}}} \tau_{\tilde{P}\tilde{N},\tilde{P}U}(f) + \frac{1}{n_{\tilde{N}}} \tau_{\tilde{P}\tilde{N},U\tilde{N}}(f)\,.
\end{aligned}
$$

此外，若 $\psi_{\tilde{P}\tilde{N}U} > \psi_{\tilde{P}\tilde{N}}$，则对于任意 $\gamma \in (2\gamma_{\tilde{P}\tilde{N}} - 1, 1)$ 有 $\mathrm{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)] \leqslant \mathrm{Var}[\hat{R}_{\tilde{P}\tilde{N}}(f)]$ 成立，即利用无标记数据的风险估计量的方差更小。

证明 经验风险可以表示为：

$$\begin{aligned}\hat{R}_{\tilde{P}\tilde{N}U}(f) &= \gamma\hat{R}_{\tilde{P}\tilde{N}}(f)+(1-\gamma)(\hat{R}_{\tilde{P}U}(f)+\hat{R}_{U\tilde{N}}(f)-\frac{1}{2}) \\ &= \frac{\gamma}{n_{\tilde{P}}n_{\tilde{N}}}\sum_{i=1}^{n_{\tilde{P}}}\sum_{j=1}^{n_{\tilde{N}}}\ell(f(x_i^{\tilde{P}},x_j^{\tilde{N}})) \\ &\quad+\frac{\gamma}{n_{\tilde{P}}n_U}\sum_{i=1}^{n_{\tilde{P}}}\sum_{j=1}^{n_U}\ell(f(x_i^{\tilde{P}},x_j^U)) \\ &\quad+\frac{\gamma}{n_Un_{\tilde{N}}}\sum_{i=1}^{n_U}\sum_{j=1}^{n_{\tilde{N}}}\ell(f(x_i^U,x_j^{\tilde{N}}))+\frac{1-\gamma}{2}\,.\end{aligned}$$

设 $n_U\to\infty$，则有

$$\begin{aligned}\mathrm{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)] &= \gamma^2\frac{\sigma^2_{\tilde{P}\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}+2\gamma(1-\gamma)\frac{\tau_{\tilde{P}\tilde{N},\tilde{P}U}(f)}{n_{\tilde{P}}} \\ &\quad+2\gamma(1-\gamma)\frac{\tau_{\tilde{P}\tilde{N},U\tilde{N}}(f)}{n_{\tilde{N}}} \\ &= \gamma^2\psi_{\tilde{P}\tilde{N}}+2\gamma(1-\gamma)\psi_{\tilde{P}\tilde{N}U}\,,\end{aligned}$$

其中分母为 n_U 的项可以被消掉。

方差最小时，令对 γ 的导数为 0，

$$\begin{aligned}\frac{\mathrm{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)]}{\gamma} &= 2\gamma\psi_{\tilde{P}\tilde{N}}+(2-2\gamma)\psi_{\tilde{P}\tilde{N}U} \\ &= (2\psi_{\tilde{P}\tilde{N}}-2\psi_{\tilde{P}\tilde{N}U})\gamma+2\psi_{\tilde{P}\tilde{N}U} \\ &= 0\,.\end{aligned}$$

解上述方程即可得到方差的最小值点。 □

定理 4.5 表明，如果选择适当的 γ，则所提出的风险估计量 $\hat{R}_{\tilde{P}\tilde{N}U}$ 比有监督风险估计量 $\hat{R}_{\tilde{P}\tilde{N}}$ 具有更小的方差，即无标记数据有助于建立模型。

4.4 实验验证

本节汇报在常用的基准数据集在上进行的弱监督 AUC 优化问题的实验结果。对于具有不准确、不完全监督的 AUC 优化，实验以图像基准数据集 MNIST、

FashionMNIST、CIFAR10 和 CIFAR100 为基础，为每个任务合成具有不同设置的多个数据集。对于具有不确切监督的 AUC 优化，我们采用了几个在相关文献中常用的多示例学习数据集[1]，包括 Musk1、Musk2、fox、tiger 和 elephant。WSAUC 与多个对比方法在不同的弱监督场景下的比较，证实了该框架是一种在不同弱监督场景下构建 AUC 优化模型的统一解决方案。此外，本节还展示了不同混杂比率下 rpAUC 和 AUC 作为训练目标产生的性能差异，以证实 rpAUC 应对混杂数据的优越性。

4.4.1 多种弱监督场景的学习性能

实验采用的不同弱监督场景下的 AUC 优化方法如下：PNU-AUC[34]及其在正标记-无标记情形下的变体。PNU-AUC 是一种半监督 AUC 优化方法，通过已知先验对标记数据上的风险进行补偿，实现了无偏的 PU 和 NU 估计。有界合页损失（barrier hinge loss），一种对称的损失函数，旨在学习在有损标记下最大化 AUC[121]，以及其他解决标记带噪问题中常用的对称和非对称损失函数。SAMULT[79]及其变体。SAMULT 是本文第二章提出的基于无偏的风险估计的半监督 AUC 优化方法。它的变体也可用于解决正标记-无标记 AUC 优化问题。MI-SVM 和 mi-SVM[113]，两种基于间隔的多示例学习方法。MissSVM[122]采用半监督学习方法，将正包中的样本视为无标记数据。SIL[123]是一种多示例学习方法，从样本包标记中进行学习。多示例学习方法 sbMIL[124]，适用于正包稀疏的情况。对于可以使用深度神经网络作为骨干网络的方法，本文使用相同的网络架构进行比较。网络架构为 [conv(3x3x8), max pooling, conv(3x3x16), max pooling, avg pooling, fc, fc]，激活函数为 ReLU。所有实验重复 10 次，以消除随机性的影响。

标记带噪的 AUC 优化 对于带有噪声标记的 AUC 优化，我们将 WSAUC 与几种带有噪声标记的学习损失/方法进行了比较。正样本和负样本标记的噪声比率分别从 $\{20\%, 30\%, 40\%\}$ 中选取。实验测试了所有不同的正负比率组合，以展示在对称和非对称噪声下的表现。结果如表 4-2 中所示。结果表明，多种对比方法在易于处理的数据集 MNIST 和 FashionMNIST 上的表现比较接近；而在具有挑战性的数据集 CIFAR10 和 CIFAR100 上，WSAUC 取得了相对较大的提升。

1 http://www.uco.es/grupos/kdis/momil/

表 4-2 标记带噪 AUC 优化性能及标准差

Dataset	Pos. noise	20%			30%			40%		
	Neg. noise	20%	30%	40%	20%	30%	40%	20%	30%	40%
MNIST	hinge	**99.6** (0.1)	**99.6** (0.1)	99.4 (0.1)	**99.6** (0.1)	**99.5** (0.1)	99.2 (0.1)	99.4 (0.1)	99.2 (0.1)	98.7 (0.3)
	ramp	99.3 (0.1)	99.2 (0.1)	99.0 (0.1)	99.2 (0.1)	99.0 (0.1)	98.8 (0.1)	99.0 (0.1)	98.8 (0.2)	98.2 (0.3)
	unhinged	88.5 (0.8)	88.5 (0.8)	88.4 (0.8)	88.4 (0.8)	88.3 (0.8)	88.2 (0.7)	88.2 (0.8)	88.1 (0.7)	87.8 (0.6)
	barrier	99.5 (0.1)	99.5 (0.1)	**99.5** (0.1)	99.5 (0.1)	**99.5** (0.1)	**99.5** (0.1)	**99.6** (0.0)	**99.5** (0.0)	**99.5** (0.1)
	WSAUC	**99.6** (0.0)	**99.6** (0.0)	**99.5** (0.1)	**99.6** (0.0)	**99.5** (0.0)	99.4 (0.1)	99.5 (0.1)	99.4 (0.1)	99.1 (0.1)
FMNIST	hinge	99.3 (0.1)	99.2 (0.1)	99.1 (0.1)	99.2 (0.1)	99.2 (0.1)	99.0 (0.1)	99.1 (0.1)	99.0 (0.1)	98.8 (0.1)
	ramp	99.2 (0.1)	99.2 (0.1)	99.1 (0.1)	99.1 (0.1)	99.1 (0.1)	98.9 (0.1)	99.0 (0.1)	98.9 (0.1)	98.6 (0.2)
	unhinged	96.6 (0.7)	96.5 (0.7)	96.4 (0.7)	96.7 (0.7)	96.6 (0.7)	96.6 (0.7)	96.8 (0.6)	96.7 (0.7)	96.7 (0.6)
	barrier	98.9 (0.1)	98.9 (0.1)	98.8 (0.1)	98.9 (0.1)	98.9 (0.1)	98.9 (0.1)	98.8 (0.1)	98.8 (0.1)	98.9 (0.1)
	WSAUC	**99.4** (0.1)	**99.4** (0.1)	**99.3** (0.1)	**99.3** (0.0)	**99.3** (0.1)	**99.2** (0.1)	**99.3** (0.0)	**99.2** (0.1)	**99.1** (0.1)
CIFAR10	hinge	92.8 (0.1)	92.0 (0.3)	91.7 (0.4)	92.1 (0.3)	91.3 (0.2)	90.1 (0.1)	91.5 (0.4)	90.3 (0.3)	88.2 (0.4)
	ramp	87.1 (5.5)	88.5 (3.1)	89.4 (3.2)	89.1 (1.3)	89.2 (2.5)	88.6 (2.4)	87.8 (1.9)	89.2 (1.4)	85.0 (2.2)
	unhinged	73.5 (1.6)	73.3 (3.1)	75.5 (3.5)	72.7 (2.1)	74.2 (5.1)	73.1 (4.7)	71.6 (1.2)	74.4 (4.7)	71.6 (1.1)
	barrier	88.5 (0.7)	88.4 (1.4)	89.0 (1.0)	88.0 (1.4)	88.7 (0.9)	88.4 (0.9)	88.1 (0.8)	88.4 (1.4)	88.2 (0.8)
	WSAUC	**93.6** (0.2)	**92.8** (0.2)	**92.3** (0.2)	**92.9** (0.2)	**92.2** (0.2)	**91.1** (0.3)	**92.2** (0.3)	**90.9** (0.4)	**89.5** (0.2)
CIFAR100	hinge	85.4 (0.5)	83.5 (1.0)	81.7 (1.0)	84.0 (0.8)	82.0 (0.8)	80.1 (1.2)	82.7 (0.6)	**80.4** (0.7)	76.4 (0.6)
	ramp	86.3 (1.4)	82.6 (5.7)	80.7 (3.6)	83.6 (2.6)	**83.7** (1.5)	77.0 (5.6)	80.0 (3.2)	74.6 (5.6)	67.4 (2.6)
	unhinged	63.0 (2.4)	61.5 (0.6)	62.7 (1.5)	64.6 (4.0)	63.4 (3.1)	63.5 (2.0)	62.8 (3.4)	64.9 (2.4)	61.5 (0.9)
	barrier	78.1 (2.0)	78.9 (2.9)	78.5 (2.7)	79.0 (3.7)	79.1 (4.2)	78.8 (3.1)	79.4 (4.0)	78.2 (4.3)	**78.3** (2.5)
	WSAUC	**86.4** (0.4)	**85.3** (0.5)	**83.3** (0.7)	**84.8** (0.5)	82.5 (0.7)	**80.3** (1.0)	**83.1** (0.5)	**80.4** (1.1)	76.5 (1.3)

正标记-无标记 AUC 优化 正标记-无标记任务通常被认为对类先验概率和标记比率很敏感，因此实验中采用不同的先验概率和标记比率进行验证。实验中，正例比率在 $\{20\%, 30\%, 40\%\}$ 中选择，标记比率选择在 $\{5\%, 10\%\}$ 中选择。在该场景下，对比方法包括 PU-AUC 和 SAMULT^{P+U}，它们分别是 PNU-AUC 和 SAMULT 在正标记-无标记场景下的变体。结果如表 4-3 中所示。可以观察到，在各组实验中 WSAUC 性能均有提升，特别是当标记比率相对较低时提升更为明显。这表明，WSAUC 通过优化 rpAUC 带来的稳健性优势在标记数据量较少时更大。

多示例 AUC 优化 对于多示例 AUC 优化，我们将 WSAUC 与下列多标记学习方法进行比较：miSVM、MISVM、MissSVM、SIL 和 sbMIL。实验在多示例学习专用的数据集上进行，即 Musk1、Musk2、fox、tiger 和 elephant。汇报的 AUC 基于样本包的分数进行计算。对于 WSAUC，模型按照样本级别进行训练，取包中最大样本分数作为整个包的分数。结果如表 4-4 中所示。与对比方法相比，WSAUC 取得了相对较大的提升，其中可能的原因有二。第一，多示例学习中正负样本概念存在不对等，更有可能发生严重的分布不平衡现象。对比方法并非针对最大化 AUC 指标设计，应对分布不平衡问题的能力较差，WSAUC 显式地优化 AUC 有利于实现更好的包级别 AUC 性能。第二，多示例学习中正包中往往也只有极少量比例的正样本，因此属于混杂程度较高的 AUC 优化问题。故而利用 rpAUC 的 WSAUC 方法能够更好地应对这一问题。

标记带噪的半监督 AUC 优化 在半监督 AUC 优化任务中，除了展示在 0% 噪声比率下的正常半监督设置中的性能外，我们还评估了在标记数据受到噪声影响的情况下（噪声比率为 20% 或 30%）的性能。据我们所知，不存在现有的 AUC 优化方法能够同时利用无标记和带有噪声标记的数据进行学习。因此，实验采用普通的半监督 AUC 优化方法 PNU-AUC 和 SAMULT 作为对比方法。实验采用的标记比率在 $\{5\%, 10\%\}$ 中选择。结果如表 4-5 中所示。可以观察到，当数据没有噪声时，WSAUC 在所有四个数据集上的表现与对比方法相似。然而，当标记数据既稀少（标记比率为 5%）又有噪声（20% 或 30% 的噪声比率）时，WSAUC 的性能显著优于对比方法。这表明，WSAUC 在处理稀少和不准确的标记数据时特别有用。在其他情况下，WSAUC 的表现与其他先进方法一样好。

表 4-3 正标记-无标记 AUC 优化性能及标准差

Dataset	Pos. ratio	20%		30%		40%	
	Label ratio	5%	10%	5%	10%	5%	10%
MNIST	PU-AUC	86.8 (3.1)	96.8 (0.5)	77.8 (7.4)	94.8 (1.1)	67.2 (0.7)	88.8 (2.9)
	SAMULT^{P+U}	86.7 (3.7)	97.5 (0.4)	74.5 (5.7)	95.3 (1.1)	65.8 (8.4)	89.1 (2.9)
	WSAUC	**93.8** (3.2)	**98.6** (0.1)	**88.5** (3.5)	**98.2** (0.2)	**84.2** (4.6)	**95.7** (1.3)
FMNIST	PU-AUC	90.7 (2.4)	98.5 (0.3)	81.1 (6.1)	98.2 (0.3)	67.6 (10.4)	97.2 (0.6)
	SAMULT^{P+U}	92.4 (2.1)	98.6 (0.3)	83.0 (6.3)	98.4 (0.3)	67.9 (10.6)	97.3 (0.6)
	WSAUC	**98.1** (1.1)	**99.0** (0.1)	**89.0** (1.2)	**98.9** (0.1)	**88.8** (1.7)	**98.6** (0.1)
CIFAR10	PU-AUC	62.6 (7.1)	81.7 (1.2)	57.6 (6.7)	77.2 (3.1)	51.8 (6.3)	70.1 (2.2)
	SAMULT^{P+U}	56.9 (7.1)	81.4 (1.8)	58.9 (7.7)	78.0 (2.2)	51.9 (6.6)	72.5 (3.7)
	WSAUC	**76.6** (7.9)	**86.7** (0.7)	**73.5** (8.7)	**82.7** (1.9)	**65.4** (9.3)	**76.9** (5.1)
CIFAR100	PU-AUC	56.0 (3.6)	68.9 (0.9)	54.4 (3.2)	65.8 (2.1)	52.6 (3.1)	60.9 (2.8)
	SAMULT^{P+U}	54.2 (3.8)	69.6 (1.3)	52.7 (3.5)	67.1 (1.0)	51.1 (3.2)	62.4 (1.8)
	WSAUC	**64.4** (5.4)	**73.7** (1.0)	**62.4** (5.5)	**70.5** (1.1)	**59.6** (5.6)	**66.6** (2.9)

表 4-4 多示例 AUC 优化性能及标准差

Dataset	Musk1	Musk2	fox	tiger	elephant
miSVM	78.8 (10.5)	75.7 (12.2)	52.8 (9.8)	78.6 (9.1)	76.7 (10.1)
MISVM	83.3 (12.0)	83.9 (14.6)	55.5 (10.8)	83.6 (10.3)	88.4 (6.0)
MissSVM	78.9 (10.1)	75.8 (12.5)	50.1 (9.4)	78.2 (8.9)	77.4 (9.2)
SIL	90.7 (9.1)	75.9 (13.4)	59.3 (11.2)	85.9 (10.6)	86.1 (6.4)
sbMIL	74.5 (18.9)	73.9 (16.0)	63.0 (8.4)	78.9 (7.9)	83.4 (7.5)
WSAUC	**96.0** (6.0)	**98.6** (4.5)	**90.4** (5.5)	**95.9** (3.7)	**96.6** (7.1)

表 4-5 （标记带噪）半监督 AUC 优化性能及标准差

Dataset	Noise ratio	0%		20%		30%	
	Label ratio	5%	10%	5%	10%	5%	10%
MNIST	PNU-AUC	95.9 (2.3)	98.5 (0.1)	91.2 (2.7)	98.5 (0.2)	83.2 (4.2)	97.5 (0.5)
	SAMULT	**98.4** (2.8)	**99.3** (0.1)	93.2 (2.5)	**98.9** (0.1)	83.9 (3.9)	98.2 (0.2)
	WSAUC	98.3 (2.7)	**99.3** (0.1)	**96.6** (2.4)	98.8 (0.1)	**95.2** (2.0)	**98.3** (0.1)
FMNIST	PNU-AUC	97.8 (0.8)	98.9 (0.1)	93.5 (1.5)	98.9 (0.1)	89.8 (2.9)	98.7 (0.1)
	SAMULT	**98.4** (0.9)	**99.3** (0.0)	96.1 (1.1)	**99.0** (0.1)	90.3 (2.6)	**98.8** (0.1)
	WSAUC	**98.4** (0.9)	**99.3** (0.1)	**97.7** (0.9)	**99.0** (0.1)	**97.4** (0.9)	**98.8** (0.1)
CIFAR10	PNU-AUC	**78.5** (5.8)	88.9 (0.1)	60.8 (7.8)	85.4 (1.1)	57.9 (8.1)	82.5 (0.9)
	SAMULT	70.3 (7.2)	89.2 (0.4)	65.8 (8.4)	**86.9** (0.8)	60.8 (8.7)	**84.5** (0.6)
	WSAUC	70.9 (7.1)	**89.4** (0.4)	**66.5** (8.0)	86.7 (0.7)	**65.8** (0.7)	83.9 (0.7)
CIFAR100	PNU-AUC	61.0 (4.9)	76.8 (0.5)	56.6 (4.1)	73.8 (0.4)	55.4 (4.3)	69.6 (1.3)
	SAMULT	**64.0** (4.5)	**78.1** (0.7)	60.1 (4.7)	**75.1** (0.3)	57.4 (4.7)	71.2 (1.2)
	WSAUC	**64.0** (4.5)	77.9 (0.6)	**61.4** (5.1)	**75.1** (0.5)	**60.3** (4.0)	**71.7** (1.0)

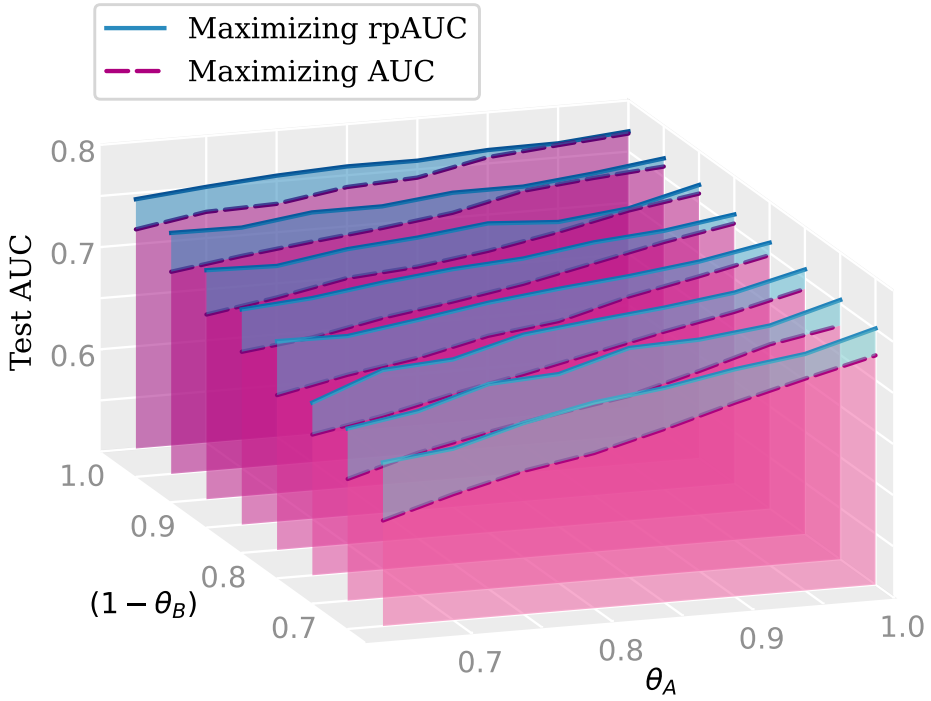


图 4-2 不同混杂比例下 rpAUC 与 AUC 作为训练目标的性能差异

4.4.2 rpAUC 与 AUC 作为优化目标的对比

本章节前文指出，rpAUC 在数据集混杂（或标记带噪）的情况下，是一种比 AUC 更具稳健性的优化目标。为从实验验证这一论断，本文在数据集混杂比例变化的情况下对 rpAUC 和 AUC 作为优化目标进行对比，汇报模型测试 AUC 的差异。实验遵循第 4.2.1 小节中的问题设置，使用不同的 θ_A 和 θ_B 来展示优化 rpAUC 相对于 AUC 的性能增量。这一组实验也可以被视为关于 rpAUC 的消融实验，此时优化完整的 AUC 即为优化 rpAUC 的消融版本。

我们采用 CIFAR100 数据集，并在 $[0.65, 1.0]$ 的范围内，以步长为 0.05 变化类别混合比例 θ_A 和 $(1-\theta_B)$。θ_A 和 $(1-\theta_B)$ 的值越小，相应数据集中的标记错误的比率就越高。实验中采用 5% 的数据进行训练，以研究在数据相对稀少时算法的性能。对于每个类比例组合，实验重复进行 10 次并取平均。结果如图 4-2 所示。其中，蓝色阴影部分表示在训练期间最大化 rpAUC 相较于最大化 AUC，在测试 AUC 上所取得的性能提升。

从图中可以观察到，当数据集干净时，最大化 AUC 或 rpAUC 的性能接近。随着混杂比例的上升，rpAUC 作为训练目标所展示出的优势逐渐扩大。

4.5 本章小结

为了解决标记不准确且不完全可见情况下的而 AUC 优化问题，本章节综合性地研究了不同种类的弱监督 AUC 优化问题，提出了两个主要结果：

1. 指出各种弱监督情况下的 AUC 优化问题可以统一为最小化混杂数据集上的 AUC 风险，并且经验风险最小化与最大化真实 AUC 是一致的。
2. 提出了反向部分 AUC（rpAUC），能够在混杂数据集上实现有效的 AUC 优化，可以作为弱监督 AUC 优化问题的具有稳健性的训练目标。

结合以上结果，本章提出 WSAUC 框架，为多个场景下的弱监督 AUC 优化问题提供了统一的解决方案，并为如何在标记信息进一步减弱，即标记不准确且不完全可见的情况下构建 AUC 优化模型给出了回答。我们希望该框架能够启发相关领域的新研究和应用。

本章工作已总结成文:

Zheng Xie, Yu Liu, Hao-Yuan He, Ming Li, Zhi-Hua Zhou. “Weakly Supervised AUC Optimization: A Unified Partial AUC Approach.” Under review. 2023.

第五章　标记不可见 AUC 优化

5.1　引言

在真实世界的机器学习问题中，为所有数据提供标记信息往往十分困难，因此机器学习方法经常不得不应对从不同程度的弱标记信息学习的问题。在第二章和第三章中，本文讨论了标记不完全可见的 AUC 优化问题。在第四章中，我们进一步综合考虑了利用标记不完全、不准确或不确切的多种弱标记数据中的 AUC 优化问题，并为标记信息进一步弱化到既不完全可见、又不准确的场景提供了构建 AUC 优化模型的学习方法。在这些场景中，学习算法仍然可以观察到一定数量的样本标记，并对其加以利用。

当标记信息进一步减弱，一个更加具有挑战性的场景是从标记完全不可见的数据中构建机器学习模型。为了使得问题可解，至少需要有多个先验概率不同的无标记数据集，并且知道每个数据集的先验概率。这种学习问题被称作从多个无标记数据集学习，或简称为 U^m 学习（learning from multiple unlabeled sets, U^m learning）。在此问题下，一般假定共有 m 个无标记数据集，其中 $m \geqslant 2$。这种情况通常出现在样本可以分为不同组别，并且不同组之间样本被划分为正类的概率各不相同，例如通过不同地区的居民特征预测预测投票率，或通过不同年份的居民特征预测发病率。

U^m 学习中最简单的一种情况，即仅考虑 $m = 2$ 的情况，可以追溯到标记比例学习（learning from label proportions，LLP）问题[125]。针对这一学习问题，Xu 等人[126]和 Krause 等人[127]假设每个聚类对应于一个单独的类，并应用判别聚类方法来构建分类器。Yu 等人[128]提出最小化每个数据集 U$_i$（即经验比例风险）的平均预测概率与类先验之间的距离以进行学习。

由于问题的复杂性，针对 $m \geqslant 3$ 的 U^m 学习算法的研究最早于 2020 年才出现。相关研究包括 Scott 等人[129]将训练在所有未标记集对上的分类器进行集成；Tsai 等人[130]为该问题引入了一致性正则化方法。最近，Lu 等人[131]提出了一种

一致性方法，用于从多个未标记集合中进行分类，这是首个对来自 m 个未标记集合（$m \geqslant 3$）的学习进行分类损失优化的分类器一致性方法。在本章节中，我们首次考虑了从 U^m 数据中学习 AUC 优化模型的问题，该模型最大化了分类器的成对排序能力[1]。这个问题的重要性体现在两个方面：首先，对于很多 U^m 学习场景，模型的排序性能更为重要。例如，可以通过不同年份具有不同发病率的居民特征中筛查出患病风险最高的居民。其次，考虑到具有不同类别先验的多个 U 集合，不平衡问题很可能会对学习过程产生负面影响。因此，采用能够适配数据分布不平衡性的性能度量 AUC 是一个十分自然的选择。

为了实现这一目标，本文提出了 U^m-AUC，首个从 U^m 数据中进行 AUC 优化的机器学习方法。U^m-AUC 将问题转化为多标记 AUC 优化问题，其中多标记学习问题中的每个标记对应于一个二分 AUC 优化子问题。为了克服成对损失计算的平方时间复杂度，U^m-AUC 将问题转化为随机鞍点问题，并通过单样本 AUC 优化算法来解决它。本文的理论分析表明 U^m-AUC 与最优 AUC 优化模型一致，并给出了泛化界。实验证明，U^m-AUC 具有良好的学习性能和对数据分布不平衡的稳健性。此外，本文提出的方法放松了普通 U^m 学习任务中知晓每个无标记数据集先验概率的要求，仅仅需要知道各个无标记集合先验概率大小的相对顺序。这使得本文的方法更加适用于现实场景。通过在多个数据集上的实验，本文验证了 U^m-AUC 在该问题上的有效性。

5.2 基于 U^m 数据的 AUC 优化方法 U^m-AUC

本章节研究在 U^m 设置下的 AUC 优化问题，该问题需要跨多个无标记数据集进行 AUC 优化。在该问题设置下，我们有 $m(m \geqslant 2)$ 个无标记数据集 $\mathrm{U}_1, \cdots, \mathrm{U}_m$，这些数据集具有不同的类别先验概率，其定义如下：

$$\mathrm{U}_i = \{x_{ik}\}_{k=1}^{n_i} \overset{\text{i.i.d.}}{\sim} p_i(x) = \pi_i p_P(x) + (1-\pi_i) p_N(x), \tag{5-1}$$

其中，$p_P(x)$ 和 $p_N(x)$ 分别是正类和负类条件概率分布，π_i 表示第 i 个无标记数据集的类别先验概率。U_i 的大小为 n_i。尽管我们只能获取无标记数据，但我们的目标是构建一个最小化 PN-AUC 风险的分类器（公式 2-5 ）。

为解决此问题，本文提出 U^m-AUC，首个基于 U^m 数据学习的 AUC 优化方

法。区别于现有的 U^m 分类研究需要实现了解类别先验概率[131]，U^m-AUC 只需要基于无标记数据集的类别先验相对顺序的知识，这更加符合实际情况。为了方便起见，不失一般性地，此处假设无标记数据集的类别先验概率按降序排列，即对于 $i<j$，有 $\pi_i \geqslant \pi_j$。另外，假设至少有两个无标记数据集具有不同的先验概率，即 $\pi_1 > \pi_m$；否则，问题将无法解决。

5.2.1 U^m 数据的 AUC 优化一致性分析

为了提供与真实 AUC 一致的解决方案，首先介绍两个无标记数据集的情况，即可以通过具有不同类别先验概率的两个无标记数据集实现一致的 AUC 学习。假设这两个无标记数据集分别是 U$_i$ 和 U$_j$，其中 $\pi_i > \pi_j$。可以最小化以下的 U^2 AUC 风险：

$$R_{ij}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_i(\boldsymbol{x})}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_j(\boldsymbol{x})}[\ell_{01}(f(\boldsymbol{x},\boldsymbol{x}'))]\right] \tag{5-2}$$

学习问题可以通过求解以下的 U^2 AUC ERM 问题来实现：

$$\min_f \quad \hat{R}_{ij}(f) = \frac{1}{n_i n_j}\sum_{\boldsymbol{x}\in \mathrm{U}_i}\sum_{\boldsymbol{x}'\in \mathrm{U}_j}\ell(f(\boldsymbol{x},\boldsymbol{x}'))\,. \tag{5-3}$$

下面的定理表明，U^2 AUC 风险最小化问题与我们需要解决的原始 AUC 优化问题是一致的，即可以通过最小化 U^2 AUC 风险来解决原始的 AUC 优化问题。

定理 5.1 (U^2 AUC 一致性) *假设 f^* 是在两个分布 p_i 和 p_j 上最小化 AUC 风险 R_{ij} 的模型，即 $f^* = \arg\min R_{ij}$，其中 $\pi_i > \pi_j$。则可以推断 f^* 也是令真实 AUC 风险 R_{PN} 的最小化的模型，即 R_{ij} 与 R_{PN} 一致。*

该定理的证明可以参见第四章中定理 4.1 和推论 4.1 及其证明。根据该定理，在仅有混合数据集可用的条件下，通过最小化 U^2 AUC 风险，即可以获得所需的 AUC 优化模型。

利用 U^m 数据，可以通过组合 $m(m-1)/2$ 个具有权重 $z_{ij}>0$ 的 AUC 子问题来构建以下的最小化问题：

$$\min_f \quad R_{U^m}(f) = \sum_{i,j|1\leqslant i<j\leqslant m} z_{ij}R_{ij}(f) \tag{5-4}$$

这对应于 U^m AUC ERM 问题：

$$\min_f \quad \hat{R}_{U^m}(f) = \sum_{i,j|1\leqslant i<j\leqslant m} \sum_{\boldsymbol{x}\in \mathrm{U}_i} \sum_{\boldsymbol{x}'\in \mathrm{U}_j} \frac{z_{ij}\ell(f(\boldsymbol{x},\boldsymbol{x}'))}{n_i n_j}\,. \tag{5-5}$$

同理，此处也可以证明 U^m AUC 风险最小化问题与原始 AUC 优化问题之间的一致性。

定理 5.2 (U^m AUC 一致性) *假设 f^* 是在 m 个分布 $p_1,\cdots,p_m$ 上最小化 AUC 风险 R_{U^m} 的模型，即 $f^* = \arg\min R_{U^m}$，其中对于 $i<j$ 有 $\pi_i > \pi_j$。则可以推断 f^* 也是令真实 AUC 风险 R_{PN} 的最小化的模型，因此 R_{U^m} 与 R_{PN} 一致。*

证明 *根据 R_{U^m} 的定义，有*

$$\begin{aligned} R_{PN}(f^*) - R_{PN}(f) &= \frac{R_{ij}(f^*) - R_{ij}(f)}{a_{ij}} \\ &= \frac{\frac{2R_{U^m}(f^*)-R_{U^m}(f)}{m(m-1)} - \sum_{i,j|1\leqslant i<j\leqslant m} \frac{1-a_{ij}}{2}}{\sum_{i,j|1\leqslant i<j\leqslant m} a_{ij}} \\ &\leqslant 0\,. \end{aligned}$$

因此 f^* 也使得 R_{PN} 取得最小值，命题得证。 □

该定理表明，在仅有不纯的数据集可用的条件下，通过优化 U^m AUC 风险最小化问题，可以获得所需的模型。这表示利用替代损失优化公式 5-5 中所有的 $m(m-1)/2$ 个 AUC 子问题的加权平均损失即为 U^m AUC 风险最小化提供了一种朴素的解决方案。

然而，当数据集数量较多或样本量较多的情况下，这样的解决方案复杂且低效。具体而言，对于 m 个数据集，根据 U^m AUC 风险的定义，总共需要处理 $m(m-1)/2$ 个子问题，当 m 量级较大时，问题将过于复杂。此外，假设样本数为 n，如果使用成对损失来优化每个子问题，那么训练中每个迭代的时间复杂度为 $O(n^2)$。这意味着该方法的时间消耗随着 n 的增长呈二次增长，对于大规模数据集来说，计算上不可行。为了解决上述问题，接下来介绍一种新颖而高效的用于 U^m AUC 风险最小化的训练方法。

5.2.2 简洁高效的学习方法 U^m-AUC

为了简化 U^m AUC 风险最小化的朴素解决方案形式，本节将其转化为一个等价的多标记学习问题，以将子问题的数量减少到 $m-1$。为了降低模型训练的时间成本，本节进一步使用了一种高效的随机优化算法，将时间复杂度从 $O(n^2)$ 降低到 $O(n)$。所提出的方法在图 5-1 中进行了概括。

子问题数量精简 为了减少子问题的数量，此处将 U^m AUC 风险最小化问题转化为一个具有 $m-1$ 个标记的多标记学习问题：令数据集 $\mathrm{U}_1, \cdots, \mathrm{U}_k$ 中的样本在第 k 个位置上标记为 1，数据集 $\mathrm{U}_{k+1}, \cdots, \mathrm{U}_m$ 中的样本在第 k 个位置上标记为 0。

换言之，令标记向量 $\bar{\boldsymbol{y}}^{(k)}$ 为第 k 个无标记数据集 U_k 的标记，其中

$$\bar{\boldsymbol{y}}^{(k)} = [\underbrace{0, 0, \cdots, 0}_{k-1}, \underbrace{1, 1, \cdots, 1}_{m-k}], \tag{5-6}$$

即在标记 $\bar{\boldsymbol{y}}^{(k)}$ 中，前面有 $k-1$ 个负类标记，后面有 $m-k$ 个正类标记。

设 $\boldsymbol{g}(x) = \hat{\boldsymbol{y}}$ 为多标记学习问题的模型输出分数，$\boldsymbol{g}_k(x)$ 为 $\boldsymbol{g}(x)$ 的第 k 维，表示第 k 个子问题的输出。原问题可以转化为最大化多标记宏平均 AUC：

$$\max_{\boldsymbol{g}} \quad \mathrm{AUC}_{\mathrm{macro}}(\boldsymbol{g}) = \frac{1}{m-1} \sum_{k=1,2,\cdots,m} \mathrm{AUC}_k(\boldsymbol{g}_k), \tag{5-7}$$

其中 AUC_k 是第 k 个标记的 AUC：

$$\mathrm{AUC}_k(\boldsymbol{g}_k) = 1 - \sum_{\boldsymbol{x} \in \bigcup_{i \leqslant k} \mathrm{U}_i} \sum_{\boldsymbol{x}' \in \bigcup_{j>k} \mathrm{U}_j} \frac{\ell(\boldsymbol{g}_k(\boldsymbol{x}, \boldsymbol{x}'))}{\sum_{i \leqslant k} n_i \sum_{j>k} n_j}. \tag{5-8}$$

多标记学习问题的第 k 个标记对应一个简单的 AUC 优化子问题：

$$\min_{\boldsymbol{g}_k} \quad \frac{1}{m-1} \sum_{\boldsymbol{x} \in \bigcup_{i \leqslant k} \mathrm{U}_i} \sum_{\boldsymbol{x}' \in \bigcup_{j>k} \mathrm{U}_j} \frac{\ell(\boldsymbol{g}_k(\boldsymbol{x}, \boldsymbol{x}'))}{\sum_{i \leqslant k} n_i \sum_{j>k} n_j}. \tag{5-9}$$

换言之，为了解决多标记学习问题公式 5-7，只需要解决 $m-1$ 个形式为公式 5-9 的子问题。下面解释可以通过解决这个多标记学习问题公式 5-7 来解决 U^m-AUC 问题公式 5-4 的原因。

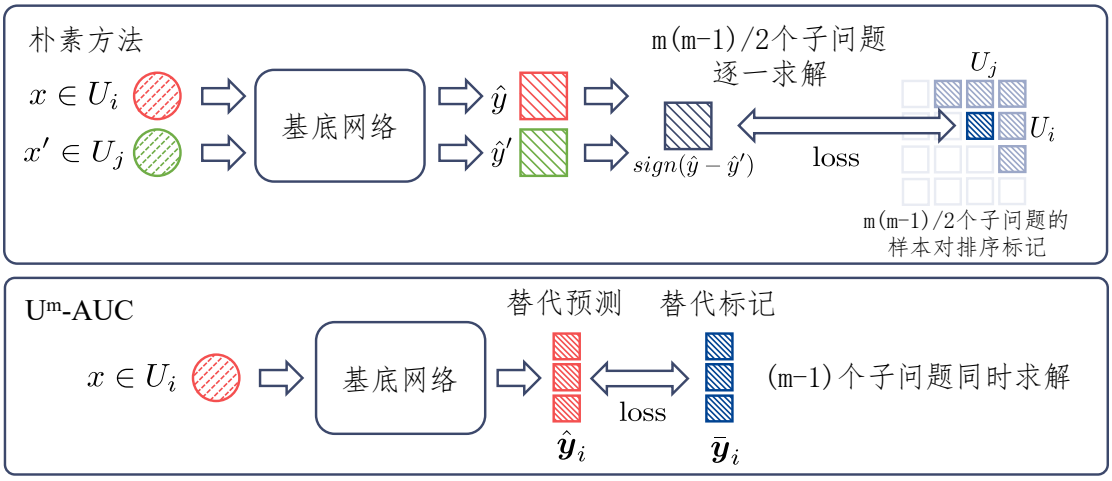


图 5-1 U^m-AUC 方法框架

设 $r_{ijk} = \frac{n_i n_j}{\sum_{i \leqslant k} n_i \sum_{j>k} n_j}$，优化问题公式 5-9 等价于以下形式：

$$\min_{\boldsymbol{g}_k} \quad \frac{1}{m-1} \sum_{i \leqslant k} \sum_{j>k} r_{ijk} \sum_{\boldsymbol{x} \in \mathrm{U}_i} \sum_{\boldsymbol{x}' \in \mathrm{U}_j} \frac{\ell(\boldsymbol{g}_k(\boldsymbol{x}, \boldsymbol{x}'))}{n_i n_j}, \tag{5-10}$$

或可简化为：

$$\min_{\boldsymbol{g}_k} \quad = \frac{1}{m-1} \sum_{i \leqslant k} \sum_{j>k} r_{ijk} \hat{R}_{ij}(\boldsymbol{g}_k), \tag{5-11}$$

这恰好是公式 5-5 中的特殊情况，其中 $z_{ij} = r_{ijk}/(m-1) > 0$。因此，多标记学习问题公式 5-7 的每个子问题都是 U^m AUC 风险最小化问题公式 5-4 。根据定理 5.4，优化这个多标记学习问题等价于解决原始的 AUC 优化问题。基于此，可以将子问题的输出聚合为 $f = \frac{1}{m-1} \sum_{k=1}^{m-1} \boldsymbol{g}_k$ 得到最终模型预测。

总结来说，通过将 U^m AUC 风险最小化问题转化为多标记学习问题，我们只需要优化 $m-1$ 个子问题，而不是像朴素方法中需要求解的平方数量 $(m(m-1)/2)$ 个子问题。

模型高效训练方法 尽管上述方法能将子问题的数量从 $O(m^2)$ 减少到 $O(m)$，但是当在训练数据上优化基于样本对计算的损失时，方法的复杂度过高。每个迭代将花费 $O(n_P \cdot n_N)$ 的时间，其中 n_P 是正样本数，n_N 是负样本数。为了解决这个问题，本文进一步将多个多标记 AUC 优化子问题再次重写为 min-max 问题，使得其可以通过交替优化算法求解。该方法每个迭代只需要 $O(n_P + n_N)$ 的时间，更适用于大规模数据集。

在多标记学习问题公式 5-7 中使用平方替代 AUC 损失，得到如下形式：

$$
\begin{aligned}
&\frac{1}{m-1}\sum_{1\leqslant k<m}\sum_{\boldsymbol{x}\in\bigcup_{i\leqslant k}\mathrm{U}_i}\sum_{\boldsymbol{x}'\in\bigcup_{j>k}\mathrm{U}_j}\frac{(1-f(\boldsymbol{x})+f(\boldsymbol{x}'))^2}{\sum_{i\leqslant k}n_i\sum_{j>k}n_j}\\
&=\frac{1}{m-1}\sum_{1\leqslant k<m}\Bigg(\underbrace{\sum_{\boldsymbol{x}\in\bigcup_{i\leqslant k}\mathrm{U}_i}\frac{(f(\boldsymbol{x})-a_k(f))^2}{\sum_{i\leqslant k}n_i}}_{A_k(f)}+\underbrace{\sum_{\boldsymbol{x}'\in\bigcup_{j>k}\mathrm{U}_j}\frac{(f(\boldsymbol{x}')-b_k(f)))^2}{\sum_{j>k}n_j}}_{B_k(f)}\\
&\qquad\qquad+\underbrace{(1-a_k(f)+b_k(f)))}_{C_k(f)}\Bigg)\\
&=\frac{1}{m-1}\sum_{1\leqslant k<m}\left(A_k(f)+B_k(f)+\max_{\alpha}\{2\alpha(1-a_k(f)+b_k(f))-\alpha^2\}\right),
\end{aligned}
\tag{5-12}
$$

其中

$$
a_k(f)=\frac{1}{\sum_{i\leqslant k}n_i}\sum_{\boldsymbol{x}\in\bigcup_{i\leqslant k}\mathrm{U}_i}f(\boldsymbol{x}),
$$

$$
b_k(f)=\frac{1}{\sum_{j>k}n_j}\sum_{\boldsymbol{x}'\in\bigcup_{j>k}\mathrm{U}_j}f(\boldsymbol{x}').
$$

目标函数公式 5-12 等价于 $m-1$ 个 min-max 问题，可以写作：

$$
\min_{f,a_k,b_k}\max_{\alpha}h(f,a_k,b_k,\alpha):=\mathbb{E}_{\boldsymbol{z}}[H(f,a_k,b_k,\alpha;\boldsymbol{z})],
\tag{5-13}
$$

其中 $\boldsymbol{z}=(\boldsymbol{x},y)$ 是一个随机样本，而

$$
\begin{aligned}
H(f,a_k,b_k,\alpha;\boldsymbol{z})&=(1-p)(f(\boldsymbol{x})-a_k)^2\mathbb{I}[y=1]+p(f(x)-b_k)^2\mathbb{I}[y=-1]\\
&\quad-p(1-p)\alpha^2+2\alpha\left(p(1-p)+pf(\boldsymbol{x})\mathbb{I}[y=-1]-(1-p)f(\boldsymbol{x})\mathbb{I}[y=1]\right),
\end{aligned}
\tag{5-14}
$$

其中 $p=\sum_{i\leqslant k}n_i/(\sum_{i\leqslant k}n_i+\sum_{j>k}n_j)$。

这些 min-max 问题可以通过高效的原始-对偶随机优化方法来解决，使用 PESG[132] 来更新参数，可以基于单样本计算损失。此外，通过将 C_k 替换为 $\max_{\alpha\geqslant 0}\{2\alpha(m-a_k(f)+b_k(f))-\alpha^2\}$，其中引入了一个边界参数 m，可以使得损失函数更加稳健[13]。通过等价问题转换技术和高效优化方法的结合，可以将训练中每个迭代的复杂度从 $O(n^2)$ 降低到 $O(n)$。算法的描述见算法 5 。

算法 5 U^m-AUC

Input: m 个无标记数据集 $U_1, \cdots, U_m$（正类先验概率递减）

for $t = 1, 2, \cdots,$ *num_epochs* **do**
  for $b = 1, 2, \cdots,$ *num_batches* **do**
    初始化模型 g
    从 $\bigcup_{0 \leqslant i \leqslant m} \mathrm{U}_i$ 采样小批量 $\mathcal{B}$
    计算模型输出 $\boldsymbol{g}(\mathcal{B})$
    计算小批量 $\mathcal{B}$ 的多标记损失
    通过 PESG 更新 $\boldsymbol{g}$ 模型参数
  end for
end for
聚合多标记模型 $\boldsymbol{g}$ 得到原问题模型 f

Output: 模型 f

5.3 理论分析

本节给出了 U^2 AUC 和 U^m AUC 的 ERM 问题的超额风险界。设 $\mathcal{X}$ 是特征空间，K 是定义在 $\mathcal{X}^2$ 上的核函数，C_w 是一个严格正实数。令 $\mathcal{F}_K$ 表示以下形式的函数类:

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

其中 $\|x\|_K = \sqrt{K(x, x)}$。假设替代损失函数 ℓ 是 L-Lipschitz 连续的，上界为严格正实数 C_ℓ，并满足不等式 $\ell \geqslant \ell_{01}$。

设 $\hat{f}_{ij}^*$ 是令经验风险 $\hat{R}_{ij}(f)$ 最小化的模型，则可证明其超额风险界，表明 $\hat{f}_{ij}^*$ 的风险收敛于函数族 $\mathcal{F}_K$ 中最优函数的风险。下面的定理给出了形式化的叙述，其证明可以参考定理 4.3 及其证明。

定理 5.3 (U^2 AUC ERM 问题的超额风险) 设 $\hat{f}_{ij}^* \in \mathcal{F}_K$ 是令经验风险 $\hat{R}_{ij}(f)$ 最小化的模型，$f_{PN}^* \in \mathcal{F}_K$ 是令真实风险最小化的模型 $R_{PN}(f)$。对于任意 $\delta > 0$，以至少 $1 - \delta$ 的概率，有下式成立

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

其中 $h(\delta) = 8\sqrt{2} C_\ell C_w C_x + 5\sqrt{2 \ln(2/\delta)}$，$a = \pi_i - \pi_j$，$n_i, n_j$ 分别是 U_i, U_j 中的样本数。

该定理保证了一般情况下的超额风险可以被如下阶数的项所限制。

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

令 $\hat{f}^*_{U^m}$ 表示令经验风险 $\hat{R}_{U^m}(f)$ 最小化的模型，则可证明以下超额风险界，表明 $\hat{f}^*_{U^m}$ 的风险收敛到函数族 $\mathcal{F}_K$ 中最优函数的风险。

定理 5.4 ($\mathbf{U}^m$ AUC ERM 问题的超额风险) 设 $\hat{f}^*_{U^m}\in\mathcal{F}_K$ 是令经验风险 $\hat{R}_{U^m}(f)$ 最小化的模型，$f^*_{PN}\in\mathcal{F}_K$ 是令真实风险 $R_{PN}(f)$ 最小化的模型。对于任意 $\delta>0$ 以至少 $1-\delta$ 的概率，有下式成立

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

其中 $h(\delta)=8\sqrt{2}C_\ell C_w C_x+5\sqrt{2\ln(2/\delta)}$，$s=\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}(\pi_i-\pi_j)$，$n_i,n_j$ 分别是 U_i,U_j 中的样本数。

证明 根据 U^m AUC 的定义，可知：

$$\begin{aligned}R_{U^m}=&\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}R_{ij}\\=&\sum_{i,j|1\leqslant i<j\leqslant m}(z_{ij}((\pi_i-\pi_j)R_{PN}+\frac{1-(\pi_i-\pi_j)}{2}))\\=&\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}(\pi_i-\pi_j)R_{PN}+\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}\frac{1-(\pi_i-\pi_j)}{2}\\=&sR_{PN}+\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}\frac{1-(\pi_i-\pi_j)}{2}\end{aligned}$$

令 $R'_{U^m}(f)=\frac{R_{U^m}-\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}\frac{1-(\pi_i-\pi_j)}{2}}{s}$ 表示对 R_{ij} 进行线性变换以估计 R_{PN}，$\hat{R}'_{U^m}(f)$ 是其经验估计量。优化 $\hat{R}_{U^m}(f)$ 的超额风险可以被写作

$$\begin{aligned}&R_{PN}(\hat{f}^*_{U^m})-R_{PN}(f^*_{PN})\\&=R_{PN}(\hat{f}^*_{U^m})-\hat{R}'_{U^m}(\hat{f}^*_{U^m})+\hat{R}'_{U^m}(\hat{f}^*_{U^m})-\hat{R}'_{U^m}(f^*_{PN})+\hat{R}'_{U^m}(f^*_{PN})-R_{PN}(f^*_{PN})\\&\leqslant 2\max_{f\in\mathcal{F}}|\hat{R}'_{U^m}(f)-R_{PN}(f)|.\end{aligned}\tag{5-15}$$

根据定理 *4.1* 和 U^m *AUC* 的定义，不等式右侧项可以写作

$$\max_{f\in\mathcal{F}}|\hat{R}'_{U^m}(f)-R_{PN}(f)|=\max_{f\in\mathcal{F}}|\hat{R}'_{U^m}(f)-R'_{U^m}(f)|\,. \tag{5-16}$$

根据 *Usunier* 等人[65]中的定理 *6*，对于任意 $\delta>0$，以至少 $1-\delta$ 的概率，对任意 $f\in\mathcal{F}_K$ 有下式成立：

$$\max_{f\in\mathcal{F}}|\hat{R}_{U_iU_j}(f)-R_{U_iU_j}(f)|\leqslant\frac{h(\delta)}{2s}\sqrt{\frac{n_i+n_j}{n_in_j}}. \tag{5-17}$$

因此

$$\begin{aligned}\max_{f\in\mathcal{F}}|\hat{R}'_{U^m}(f)-R'_{U^m}(f)|\leqslant&\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}\max_{f\in\mathcal{F}}|\hat{R}'_{U^m}(f)-R'_{U^m}(f)|\\ \leqslant&\frac{h(\frac{2\delta}{m(m-1)})}{2s}\sum_{i,j|1\leqslant i<j\leqslant m}z_{ij}\sqrt{\frac{n_i+n_j}{n_in_j}}\end{aligned} \tag{5-18}$$

将公式 *5-16* 和公式 *5-18* 带入公式 *5-15* 右侧，定理得证。 □

定理 5.4 保证了超额风险可以被如下阶数的项所限制。

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

很容易看出，当 $m=2$ 且 $z_{12}=1$ 时，定理 5.4 会退化成定理 5.3 。

5.4 实验验证

本节汇报 U^m-AUC 与对比方法进行对比的实验结果。

数据集 实验采用基准数据集 Kuzushiji-MNIST（简称 K-MNIST）[133]、CIFAR-10 和 CIFAR-100[134]来测试 U^m-AUC 的性能，并生成了多个具有不同设置的数据集。这些数据集被转换为二分类数据集，其中对于 K-MNIST，设置奇数和偶数两类，对于 CIFAR 数据集，则设置动物和非动物两类。

实验中，在 $\{10,50\}$ 中选择 m 的取值。若未特殊说明，则每个未标记数据集 U$_i$ 的大小固定为 $n_i=\lceil n_{\text{train}}/m\rceil$。为了模拟不同情况下的数据集分布，实验

将从四种不同的分布中生成类先验 $\{\pi_i\}_{i=1}^m$，并确保类先验不完全相同，以避免出现数学上无法解决的情况。然后，根据公式 5-1 中的定义，从训练集中按照不同正负比例随机采样数据生成每一个无标记集合 U_i 。

模型 对于在 Kuzushiji-MNIST 数据集上的所有实验，使用一个 5 层的多层感知机（MLP）作为骨干模型。对于 CIFAR 数据集的实验，则使用 Resnet32[135]作为骨干模型。方法采用 PESG[132]作为优化器。所有模型训练 150 个 epoch，并报告测试集上的 AUC。

对比方法 实验中，本文将所提出的方法与两个目前最先进的 U^m 分类方法进行了比较：LLP-VAT[130]代表了 EPRM 方法，U^m-SSC[131]代表了 ERM 方法。需要注意的是，先前的方法要求已知类别先验，而在本节的设置中，只能获取未标记数据集的类别先验的相对顺序关系。为了确保性能比较的公平性，此处额外对比了 LLP-VAT 和 U^m-SSC 的较弱版本，分别称为 LLP-VAT* 和 U^m-SSC*，通过使用均匀分布的 $[0,1]$ 之间的值来代替真实先验进行计算。对比方法采用 Adam[136]和交叉熵损失进行优化，遵循原始论文中的标准实现。为了确保公平性，在所有任务中，使用相同的基底模型来实现所有方法。

所有方法的实现均基于 PyTorch[137]，实验在 NVIDIA Tesla V100 GPU 上进行。为了排除随机性的影响，所有实验均以不同的随机种子重复 3 次，并汇报每组实验的均值和标准差。

5.4.1 模型性能

本文在前述三个图像数据集和两个不同的无标记样本集数量上进行了实验。考虑到在实际应用场景中，数据集的类别先验往往不遵循均匀分布，为了更好地模拟实际情况，本文针对每个无标记数据集考虑了四种不同的类别先验分布：$Beta(1,1)$，$Beta(5,1)$，$Beta(5,5)$ 和 $Beta(5,2)$。本节中，分别将这四个分布称为均匀、偏倚、集中和偏倚集中。其具体定义如下所示：

1. $\mathcal{D}_u$（均匀）：类别先验在 $[0,1]$ 上均匀分布；
2. $\mathcal{D}_b$（偏倚）：类别先验偏向一侧，即大多数集合中的正样本多于负样本；

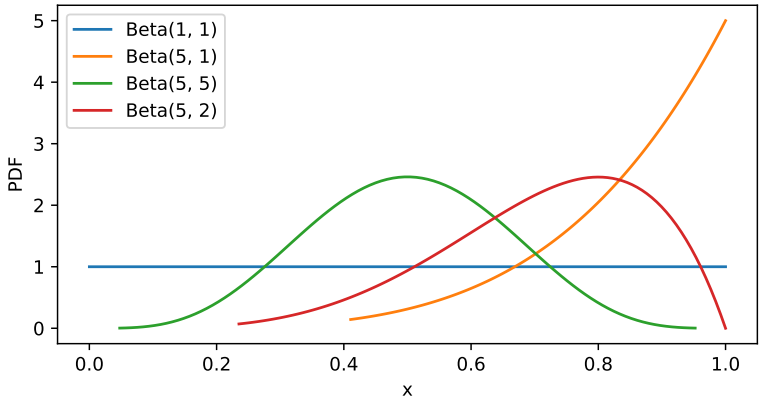


图 5-2 实验中所用 Beta 分布的概率密度函数

3. $\mathcal{D}_c$（集中）：类别先验集中在 0.5 附近，即大多数集合中的正样本和负样本比例接近；
4. $\mathcal{D}_{bc}$（偏倚集中）：类别先验集中在 0.8 附近，即大多数集合中的正样本和负样本比例接近，并且正样本多于负样本。

具体而言，给定无标记数据集数量 m 和每个集合的大小 $\{n_i\}_{i=1}^m$，我们首先从 Beta 分布中采样类别先验 $\{\pi_i\}_{i=1}^m$ 并排序，然后随机从正负样本集合中采样组成无标记数据集。Beta 分布是一个有两个参数 α 和 β 的连续概率分布，其概率密度函数为

$$f(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\qquad 0\leqslant x\leqslant 1\,,$$

其中 $B(\alpha,\beta)$ 是 beta 函数，定义为

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx\,.$$

本文中使用的 Beta 分布的概率密度函数如图 5-2 所示。可以预见，不同数据集的先验差距越小，则分类器构建越困难。

实验汇报了在 $m=10$ 和 $m=50$ 设定下，在不同的数据集和不同的类别先验分布下得到的结果，如表 5-1 和表 5-2 所示。结果表明，本文提出的 U^m-AUC 方法在利用较少信息的情况下也能优于对比方法。可以看出，由于 AUC 优化具有对数据分布不平衡的稳健性，在两种类别先验偏倚和偏倚集中的情况下，U^m-AUC 相较于两种对比方法优势尤其明显。在类别先验均匀和集中的情况下，U^m-AUC 也展现出了较大的性能优势。这可能是因为利用 U^m 数据学习本质上

表 5-1 不同先验分布的测试 AUC（$m = 10$）

Dataset	$\mathcal{D}$	LLP-VAT⋆	LLP-VAT	U^m-SSC⋆	U^m-SSC	U^m-AUC
K-MNIST	$\mathcal{D}_u$	$0.865_{\pm 0.0145}$	$0.896_{\pm 0.0249}$	$0.908_{\pm 0.0073}$	$0.911_{\pm 0.0084}$	$\mathbf{0.938}_{\pm 0.0064}$
	$\mathcal{D}_b$	$0.780_{\pm 0.0225}$	$0.789_{\pm 0.0185}$	$0.833_{\pm 0.0357}$	$0.836_{\pm 0.0521}$	$\mathbf{0.851}_{\pm 0.0616}$
	$\mathcal{D}_c$	$0.853_{\pm 0.0330}$	$0.808_{\pm 0.0131}$	$0.858_{\pm 0.0239}$	$0.856_{\pm 0.0307}$	$\mathbf{0.870}_{\pm 0.0512}$
	$\mathcal{D}_{bc}$	$0.825_{\pm 0.0315}$	$0.798_{\pm 0.0332}$	$0.868_{\pm 0.0255}$	$0.857_{\pm 0.0390}$	$\mathbf{0.896}_{\pm 0.0439}$
CIFAR-10	$\mathcal{D}_u$	$0.856_{\pm 0.0131}$	$0.856_{\pm 0.0066}$	$0.860_{\pm 0.0090}$	$0.859_{\pm 0.0131}$	$\mathbf{0.905}_{\pm 0.0080}$
	$\mathcal{D}_b$	$0.723_{\pm 0.0454}$	$0.737_{\pm 0.0754}$	$0.746_{\pm 0.0614}$	$0.778_{\pm 0.0462}$	$\mathbf{0.866}_{\pm 0.0238}$
	$\mathcal{D}_c$	$0.787_{\pm 0.0172}$	$0.847_{\pm 0.0059}$	$0.792_{\pm 0.0372}$	$0.807_{\pm 0.0209}$	$\mathbf{0.884}_{\pm 0.0046}$
	$\mathcal{D}_{bc}$	$0.769_{\pm 0.0373}$	$0.805_{\pm 0.0231}$	$0.796_{\pm 0.0552}$	$0.812_{\pm 0.0430}$	$\mathbf{0.887}_{\pm 0.0155}$
CIFAR-100	$\mathcal{D}_u$	$0.734_{\pm 0.0092}$	$0.731_{\pm 0.0167}$	$0.747_{\pm 0.0192}$	$0.756_{\pm 0.0115}$	$\mathbf{0.847}_{\pm 0.0121}$
	$\mathcal{D}_b$	$0.630_{\pm 0.0183}$	$0.651_{\pm 0.0210}$	$0.652_{\pm 0.0332}$	$0.667_{\pm 0.0331}$	$\mathbf{0.715}_{\pm 0.0292}$
	$\mathcal{D}_c$	$0.670_{\pm 0.0168}$	$0.707_{\pm 0.0117}$	$0.676_{\pm 0.0363}$	$0.692_{\pm 0.0264}$	$\mathbf{0.757}_{\pm 0.0136}$
	$\mathcal{D}_{bc}$	$0.672_{\pm 0.0359}$	$0.700_{\pm 0.0324}$	$0.683_{\pm 0.0500}$	$0.701_{\pm 0.0415}$	$\mathbf{0.751}_{\pm 0.0641}$

表 5-2 不同先验分布的测试 AUC（$m = 50$）

Dataset	$\mathcal{D}$	LLP-VAT⋆	LLP-VAT	U^m-SSC⋆	U^m-SSC	U^m-AUC
K-MNIST	$\mathcal{D}_u$	$0.896_{\pm 0.0124}$	$0.902_{\pm 0.0102}$	$0.915_{\pm 0.0136}$	$0.915_{\pm 0.0107}$	$\mathbf{0.931}_{\pm 0.0156}$
	$\mathcal{D}_b$	$0.808_{\pm 0.0142}$	$0.787_{\pm 0.0196}$	$0.861_{\pm 0.0102}$	$0.869_{\pm 0.0083}$	$\mathbf{0.883}_{\pm 0.0229}$
	$\mathcal{D}_c$	$0.863_{\pm 0.0206}$	$0.833_{\pm 0.0165}$	$0.855_{\pm 0.0378}$	$0.863_{\pm 0.0417}$	$\mathbf{0.867}_{\pm 0.0125}$
	$\mathcal{D}_{bc}$	$0.860_{\pm 0.0523}$	$0.815_{\pm 0.0052}$	$0.881_{\pm 0.0056}$	$0.885_{\pm 0.0078}$	$\mathbf{0.904}_{\pm 0.0012}$
CIFAR-10	$\mathcal{D}_u$	$0.852_{\pm 0.0079}$	$0.857_{\pm 0.0073}$	$0.853_{\pm 0.0030}$	$0.854_{\pm 0.0492}$	$\mathbf{0.889}_{\pm 0.0083}$
	$\mathcal{D}_b$	$0.757_{\pm 0.0250}$	$0.742_{\pm 0.0847}$	$0.794_{\pm 0.0278}$	$0.806_{\pm 0.0204}$	$\mathbf{0.861}_{\pm 0.0097}$
	$\mathcal{D}_c$	$0.790_{\pm 0.0132}$	$0.852_{\pm 0.0038}$	$0.807_{\pm 0.0101}$	$0.808_{\pm 0.0062}$	$\mathbf{0.861}_{\pm 0.0138}$
	$\mathcal{D}_{bc}$	$0.804_{\pm 0.0056}$	$0.830_{\pm 0.0235}$	$0.826_{\pm 0.0059}$	$0.832_{\pm 0.0052}$	$\mathbf{0.873}_{\pm 0.0074}$
CIFAR-100	$\mathcal{D}_u$	$0.739_{\pm 0.0036}$	$0.738_{\pm 0.0084}$	$0.742_{\pm 0.0647}$	$0.744_{\pm 0.0084}$	$\mathbf{0.844}_{\pm 0.0042}$
	$\mathcal{D}_b$	$0.669_{\pm 0.0199}$	$0.673_{\pm 0.0363}$	$0.686_{\pm 0.0103}$	$0.696_{\pm 0.0068}$	$\mathbf{0.756}_{\pm 0.0281}$
	$\mathcal{D}_c$	$0.689_{\pm 0.0075}$	$0.724_{\pm 0.0065}$	$0.700_{\pm 0.0018}$	$0.703_{\pm 0.0097}$	$\mathbf{0.790}_{\pm 0.0085}$
	$\mathcal{D}_{bc}$	$0.699_{\pm 0.0065}$	$0.718_{\pm 0.0082}$	$0.714_{\pm 0.0009}$	$0.717_{\pm 0.0024}$	$\mathbf{0.812}_{\pm 0.0163}$

表 5-3 不同不平衡设置下的测试 AUC

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

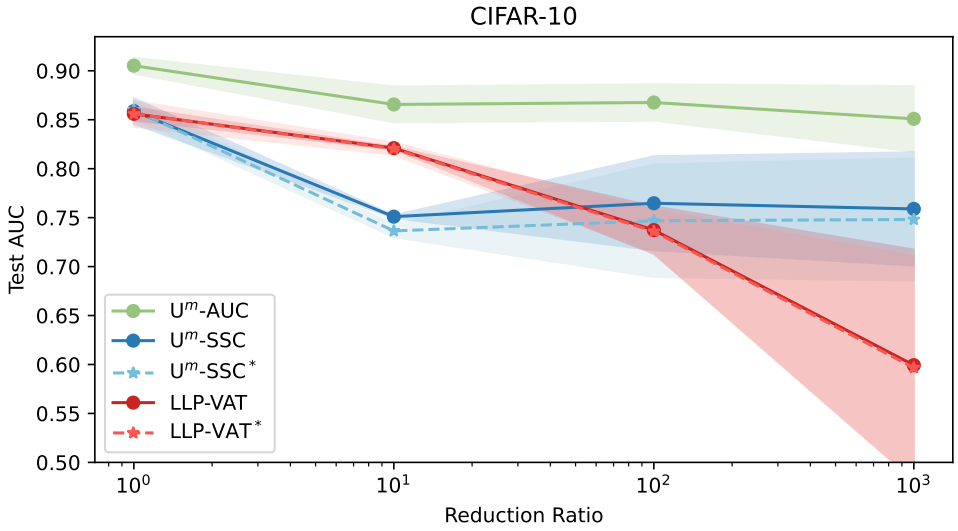


图 5-3 各方法在 CIFAR-10 数据集上不同缩减比例（数据不平衡性）下的测试 AUC

包含非常多的子分类问题，而正负类别分布不平衡的情况总会在一部分子分类问题中发生。

5.4.2 对数据集大小不平衡的稳健性

除了正负类先验分布不平衡的情况，在 U^m 的设置中，经常也会遇到每个数据集大小不平衡的情况。当遇到该问题时，常常会导致偏向较大数据集的模型，同时在少数类上表现不佳。

为了评估所提出的方法对不平衡数据集的稳健性，本文进行了各种不平衡设置的实验。具体而言，本小节按照相关文献[131]的方法，采用以下两种方式生成不平衡数据集：

1. 大小缩减：随机选择一半（$\lceil m/2 \rceil$ 个）数据集，并将它们的大小改变为 $\lceil \tau \cdot (n_{\text{train}}/m) \rceil$，其中 τ 是缩减比率。
2. 随机：从范围 $[0, n_{\text{train}}]$ 中随机采样数据集大小 n_i，使得 $\sum_{i=1}^{m} n_i = n_{\text{train}}$。

表 5-3 中呈现了 U^m-AUC 在不同不平衡数据集上的测试 AUC。结果表明，本文方法受数据集不平衡问题的影响相对较小。随着缩减比率的减小，测试性能缓慢下降，测试性能的方差也缓慢增加。

此外，我们在连续的不平衡程度设置下测试了各个方法。具体而言，首先生成类先验 $\mathcal{D}_u$ 并选择具有最小 π_i 的 $\lceil m/2 \rceil$ 个数据集 U$_i$，然后将它们的大小更

改为 $\lceil n_{\text{train}}/(r \cdot m) \rceil$，其中 r 是缩减比例。在 CIFAR-10 数据集上，使用 $m = 10$ 进行的实验结果如图 5-3 所示。可以看出，U^m-AUC 相较于对比方法 U^m-SSC 和 LLP-VAT，在不平衡设置下具有更小的测试性能下降，测试性能方差增加幅度较小，对数据不平衡的适应性更好。U^m-SSC 在出现较小缩减比例时性能就有明显下降；而 LLP-VAT 在数据不平衡时，性能下降十分剧烈。

5.5 本章小结

在本章节中，我们研究了从多个未标记数据集中构建 AUC 优化模型的问题。在该场景下，标记信息进一步减弱，数据标记完全不可见。为了解决这个问题，本章提出了 U^m-AUC，它不仅具有理论保证，还能够高效快速地求解。相较于其他 U^m 学习方法，U^m-AUC 不需要知道每一个无标记数据集的具体先验概率，仅仅需要知道其相对大小顺序，这更加符合现实场景。实验证明，U^m-AUC 相比于其他对比方法表现出了更好的性能和稳健性。

本章工作已总结成文:

Zheng Xie, Yu Liu, Ming Li. “AUC Optimization from Multiple Unlabeled Datasets.” Under review. 2023.

第六章　结束语

本论文中的研究内容主要涉及国家自然科学基金创新研究群体项目“面向开放动态环境的机器学习”（61921006），国家自然科学基金面上项目“面向开放动态环境的软件自适应学习研究”（62076121），国家重点研发计划课题“智能无人集群系统全局规划及协同行为管控”（2017YFB1001903）等项目。

AUC 作为机器学习任务中常用的评价指标，具有不受数据分布不平衡影响、能够衡量模型排序性能等优势。为构建具有良好 AUC 指标性能的机器学习模型，AUC 优化这一学习范式被提出，并受到了广泛关注。然而，针对如何在弱标记的场景下构建 AUC 优化模型这一问题，目前相关研究较少。

为解决这一问题，本文针对如何在不同程度的弱标记场景下构建 AUC 优化模型进行了系统性的研究。具体而言，包含以下四个方面：

第二章针对标记不完全可见的场景，提出了半监督 AUC 优化方法 SAMULT。该方法利用 AUC 成对损失的特殊性质，实现了在无需知道先验概率的情况下即可通过无标记数据进行无偏 AUC 风险估计，进而可以在无需猜测无标记样本伪标记的情况下利用无标数据。该方法摆脱了对特定分布假设的依赖，避免了在分布假设不成立时引入的风险，在多个任务上实现了更好的 AUC 优化性能。

第三章针对标记不完全可见的流式数据场景，提出了半监督在线 AUC 优化方法 SOLA。该方法通过将第二章提出的风险最小化问题转化为随机鞍点问题，实现了基于单个样本进行半监督 AUC 损失梯度更新的方法，解决了在线半监督 AUC 优化由于无法基于样本对计算风险而难以实现的困难。该方法首次为面临数据流式产生、分布不平衡且标记不完全可见的学习任务提供了解决方案，并在软件持续构建预测任务中展现出优秀的学习性能和极高的运行效率。

第四章针对标记不准确且不完全可见的场景，提出了弱监督 AUC 优化框架 WSAUC，将从多种不同弱标记数据进行 AUC 优化进行了综合考虑。该框架将不同的弱监督信息转化为标记混杂的统一形式，并基于一种新型的部分 AUC 优

化方法实现了通用的稳健弱监督 AUC 优化方法。该框架首次对多种弱监督信息进行综合研究，实现了利用标记不准确且不完全可见场景下的 AUC 优化，并在多种弱监督场景下取得了良好的 AUC 优化性能。

第五章针对标记不可见的场景，提出了利用多个无标记集合进行 AUC 优化的方法 U^m-AUC。该方法仅依赖二或多个具有不同先验的无标记样本集合及其先验大小次序的知识，将原学习问题转化成为一个多标记场景下的宏平均 AUC（Marco AUC）优化问题，实现了高效求解。借助于 AUC 优化的特殊性，该方法无需依赖对数据集类别先验的知识，比其他针对类似场景的学习方法所需监督信息更少、更符合实际，并取得了良好的学习效果。

对于弱标记场景下的 AUC 优化问题，未来还有以下问题有待进一步解决：

如何在开放、动态的机器学习环境下构建 AUC 优化模型？在现实机器学习任务中，经常会面临环境变化的问题，例如类别变化、特征变化、数据分布变化等。AUC 指标能够良好应对数据分布不平衡的场景；在本文中，已经对如何在弱标记场景下利用 AUC 指标的特性设计方法进行了探讨。当环境发生变化时，变化前的给出的标记不能完全代表变化后的环境，可以被视作一种较弱的标记信息。AUC 优化问题在弱标记学习场景体现出的特性能否在动态变化的场景中被合理利用、如何针对动态变化的场景提出新的 AUC 优化方法值得探讨。

本文针对弱标记 AUC 问题的理论能否应用到其他排序指标的最优化学习任务中？在机器学习任务中，还有许多其他常用的模型评价指标，如 F1、AUPRC、AP（平均精确度）、NDCG 等。这些模型评价指标与 AUC 在形式上具有相似性。能否、如何将本文针对 AUC 优化提出的理论及方法推广到其他指标的弱标记优化任务上将是一个有趣的研究方向。

参考文献

[1] James A. Hanley, Barbara J. McNeil. “The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve.” *Radiology*, 1982, 143(1): 29-36.

[2] Tianbao Yang, Yiming Ying. “AUC Maximization in the Era of Big Data and AI: A Survey.” *ACM Computing Surveys*, 2022, 55(8): 1-37.

[3] Kent A. Spackman. “Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning.” In: *Proceedings of the 6th International Workshop on Machine Learning*. 1989: 160-163.

[4] Foster J. Provost, Tom Fawcett, Ron Kohavi. “The Case against Accuracy Estimation for Comparing Induction Algorithms.” In: *Proceedings of the 15th International Conference on Machine Learning*. 1998: 445-453.

[5] Yoav Freund, Raj Iyer, Robert E. Schapire, Yoram Singer. “An Efficient Boosting Algorithm for Combining Preferences.” *Journal of Machine Learning Research*, 2003, 4: 933-969.

[6] Thorsten Joachims. “A Support Vector Method for Multivariate Performance Measures.” In: *Proceedings of the 22nd International Conference on Machine Learning*. 2005: 377-384.

[7] Alan Herschtal, Bhavani Raskutti. “Optimising Area Under the ROC Curve Using Gradient Descent.” In: *Proceedings of the 21st International Conference on Machine Learning*. 2004: 49-56.

[8] Yiming Ying, Longyin Wen, Siwei Lyu. “Stochastic Online AUC Maximization.” In: *Advances in Neural Information Processing Systems 29*. 2016: 451-459.

[9] Wei Gao, Zhi-Hua Zhou. “On the Consistency of AUC Pairwise Optimization.” In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. 2015: 939-945.

[10] Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, Dan Roth. “Generalization Bounds for the Area Under the ROC Curve.” *Journal of Machine Learning Research*, 2005, 6: 393-425.

[11] Mingrui Liu, Zhuoning Yuan, Yiming Ying, Tianbao Yang. “Stochastic AUC Maximization with Deep Neural Networks.” In: *Proceedings of the International Conference on Learning Representations*. 2020.

[12] Zhuoning Yuan, Zhishuai Guo, Nitesh Chawla, Tianbao Yang. “Compositional Training for End-to-End Deep AUC Maximization.” In: *Proceedings of the International Conference on Learning Representations*. 2022.

[13] Zhuoning Yuan, Yan Yan, Milan Sonka, Tianbao Yang. “Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification.” In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*. 2021: 3040-3049.

[14] Zhenqiu Liu, Terry Hyslop. “Partial AUC for Differentiated Gene Detection.” In: *Proceedings of the 2010 IEEE International Conference on BioInformatics and BioEngineering*. 2010: 310-311.

[15] Zheng Xie, Ming Li. “Cutting the Software Building Efforts in Continuous Integration by Semi-Supervised Online AUC Optimization.” In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 2018: 2875-2881.

[16] Zhongxin Bai, Xiao-Lei Zhang, Jingdong Chen. “Partial AUC Optimization Based Deep Speaker Embeddings with Class-Center Learning for Text-Independent Speaker Verification.” In: *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2020: 6819-6823.

[17] Asghar Feizi. “Hierarchical Detection of Abnormal Behaviors in Video Surveillance through Modeling Normal Behaviors Based on AUC Maximization.” *Soft Computing*, 2020, 24(14): 10401-10413.

[18] David M. Green, John A. Swets. Signal Detection Theory and Psychophysics. New York: John Wiley, 1966.

[19] Thorsten Joachims. “Optimizing Search Engines using Clickthrough Data.” In: *ACM Conference on Knowledge Discovery and Data Mining*. 2002.

[20] Corinna Cortes, Mehryar Mohri. “AUC Optimization Vs. Error Rate Minimization.” In: *Advances in Neural Information Processing Systems 16*. 2003: 313-320.

[21] Toon Calders, Szymon Jaroszewicz. “Efficient AUC Optimization for Classification.” In: *Knowledge Discovery in Databases: PKDD 2007*. 2007: 42-53.

[22] Peilin Zhao, Steven C. H. Hoi, Rong Jin, Tianbao Yang. “Online AUC Maximization.” In: *Proceedings of the 28th International Conference on Machine Learning*. 2011: 233-240.

[23] Wei Gao, Rong Jin, Shenghuo Zhu, Zhi-Hua Zhou. “One-Pass AUC Optimization.” In: *Proceedings of the 30th International Conference on Machine Learning*. 2013: 906-914.

[24] Michael Natole, Yiming Ying, Siwei Lyu. “Stochastic Proximal Algorithms for AUC Maximization.” In: Proceedings of Machine Learning Research: *Proceedings of the 35th International Conference on Machine Learning*: vol. 80. 2018: 3710-3719.

[25] Yunwen Lei, Yiming Ying. “Stochastic Proximal AUC Maximization.” *Journal of Machine Learning Research*, 2021, 22(61): 2832-2876.

[26] Harikrishna Narasimhan, Shivani Agarwal. “SVM$^{tight}_{pAUC}$: A New Support Vector Method for Optimizing Partial AUC Based on a Tight Convex Upper Bound.” In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2013: 167-175.

[27] Hanfang Yang, Kun Lu, Xiang Lyu, Feifang Hu. “Two-Way Partial AUC and Its Properties.” *Statistical Methods in Medical Research*, 2019, 28(1): 184-195.

[28] Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, Qingming Huang. “When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC.” In: *Proceedings of the 38th International Conference on Machine Learning*. 2021: 11820-11829.

[29] Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, Tianbao Yang. “When AUC meets DRO: Optimizing Partial AUC for Deep Learning with Non-Convex Convergence Guarantee.” In: *Proceedings of the 39th International Conference on Machine Learning*. 2022: 27548-27573.

[30] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, et al. “Big Self-Supervised Models Advance Medical Image Classification.” In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*. 2021: 3458-3468.

[31] Massih Reza Amini, Tuong Vinh Truong, Cyril Goutte. “A Boosting Algorithm for Learning Bipartite Ranking Functions with Partially Labeled Data.” In: *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2008: 99-106.

[32] Shijun Wang, Diana Li, Nicholas Petrick, Berkman Sahiner, Marius George Linguraru, Ronald M. Summers. “Optimizing Area under the ROC Curve Using Semi-Supervised Learning.” *Pattern Recognition*, 2015, 48(1): 276-287.

[33] Thorsten Joachims. “Transductive Inference for Text Classification Using Support Vector Machines.” In: *Proceedings of the 16th International Conference on Machine Learning*. 1999: 200-209.

[34] Tomoya Sakai, Gang Niu, Masashi Sugiyama. “Semi-Supervised AUC Optimization Based on Positive-Unlabeled Learning.” *Machine Learning*, 2018, 107: 767-794.

[35] Olivier Chapelle, Bernhard Schlkopf, Alexander Zien. Semi-Supervised Learning. The MIT Press, 2006.

[36] Xiaojin Zhu, Andrew B. Goldberg, Ronald Brachman, Thomas Dietterich. Introduction to Semi-Supervised Learning. Morgan, 2009.

[37] Behzad M. Shahshahani, David A. Landgrebe. “The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon.” *IEEE Transactions on Geoscience and Remote Sensing*, 1994, 32(5): 1087-1095.

[38] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, Tom Mitchell. “Text Classification from Labeled and Unlabeled Documents Using EM.” *Machine Learning*, 2000, 39(2): 103-134.

[39] Akinori Fujino, Naonori Ueda. “A Semi-Supervised AUC Optimization Method with Generative Models.” In: *Proceedings of the IEEE 16th International Conference on Data Mining*. IEEE, 2016: 883-888.

[40] Olivier Chapelle, Mingmin Chi, Alexander Zien. “A Continuation Method for Semi-Supervised SVMs.” In: *Proceedings of the 23rd International Conference on Machine Learning*. 2006: 185-192.

[41] Yu-Feng Li, James T. Kwok, Zhi-Hua Zhou. “Cost-sensitive Semi-Supervised Support Vector Machine.” In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. AAAI Press, 2010: 500-505.

[42] Avrim Blum, Shuchi Chawla. “Learning from Labeled and Unlabeled Data Using Graph Mincuts.” In: *Proceedings of the 18th International Conference on Machine Learning*. 2001: 19-26.

[43] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty. “Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions.” In: *Proceedings of the 20th International Conference on Machine Learning*. 2003: 912-919.

[44] Fei Wang, Changshui Zhang. “Label Propagation Through Linear Neighborhoods.” *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(1): 55-67.

[45] Avrim Blum, Tom Mitchell. “Combining Labeled and Unlabeled Data with Co-

training." In: *Proceedings of the 11th Annual Conference on Computational Learning Theory*. ACM, 1998: 92-100.

[46] Sally A. Goldman, Yan Zhou. "Enhancing Supervised Learning with Unlabeled Data." In: *Proceedings of the 27th International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2000: 327-334.

[47] Zhi-Hua Zhou, Ming Li. "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers." *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(11): 1529-1541.

[48] Ming Li, Zhi-Hua Zhou. "Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples." *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 2007, 37(6): 1088-1098.

[49] Ming Li, Hang Li, Zhi-Hua Zhou. "Semi-Supervised Document Retrieval." *Information Processing and Management*, 2009, 45(3): 341-355.

[50] Zhi-Hua Zhou, Ming Li. "Semi-Supervised Learning by Disagreement." *Knowledge and Information Systems*, 2010, 24(3): 415-439.

[51] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, Tapani Raiko. "Semi-supervised Learning with Ladder Networks." In: *Advances in Neural Information Processing Systems 28*. 2015: 3546-3554.

[52] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz. "mixup: Beyond Empirical Risk Minimization." In: *Proceedings of the International Conference on Learning Representations*. 2018.

[53] Yves Grandvalet, Yoshua Bengio. "Semi-supervised Learning by Entropy Minimization." In: *Advances in Neural Information Processing Systems 17*. 2005: 529-536.

[54] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, Max Welling. "Semi-supervised Learning with Deep Generative Models." In: *Advances in Neural Information Processing Systems 27*. 2014: 3581-3589.

[55] Abhishek Kumar, Prasanna Sattigeri, Tom Fletcher. “Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference.” In: *Advances in Neural Information Processing Systems 30*. 2017: 5534-5544.

[56] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, Gabriele Monfardini. “The Graph Neural Network Model.” *IEEE Transactions on Neural Networks*, 2008, 20(1): 61-80.

[57] Thomas N. Kipf, Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks.” In: *Proceedings of the International Conference on Learning Representations*. 2017.

[58] Yujia Li, Daniel Tarlow, Marc Brockschmidt, Richard Zemel. “Gated Graph Sequence Neural Networks.” In: *Proceedings of the International Conference on Learning Representations*. 2016.

[59] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, Shin Ishii. “Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41: 1979-1993.

[60] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, Colin A Raffel. “MixMatch: A Holistic Approach to Semi-Supervised Learning.” In: *Advances in Neural Information Processing Systems 32*. 2019: 5049-5059.

[61] Kihyuk Sohn, David Berthelot, Nicholas Carlini, et al. “FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence.” In: *Advances in Neural Information Processing Systems 33*. 2020: 596-608.

[62] Fábio Gagliardi Cozman, Ira Cohen. “Unlabeled Data Can Degrade Classification Performance of Generative Classifiers.” In: *Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference*. AAAI Press, 2002: 327-331.

[63] Avi Arampatzis, Jaap Kamps, Stephen Robertson. “Where to Stop Reading a

Ranked List?: Threshold Optimization Using Truncated Score Distributions." In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2009: 524-531.

[64] Zachary C. Lipton, Charles Elkan, Balakrishnan Naryanaswamy. "Optimal Thresholding of Classifiers to Maximize F1 Measure." In: *Proceedings of the 2014 European Conference on Machine Learning and Knowledge Discovery in Databases*. 2014: 225-239.

[65] Nicolas Usunier, Massih-Reza Amini, Patrick Gallinari. "A Data-dependent Generalisation Error Bound for the AUC." In: *ICML'05 Workshop on ROC Analysis in Machine Learning*. 2005.

[66] Moshe Lichman. UCI Machine Learning Repository. University of California, Irvine, School of Information. 2013. http://archive.ics.uci.edu/ml.

[67] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, Gideon Dror. "Result Analysis of the NIPS 2003 Feature Selection Challenge." In: *Advances in Neural Information Processing Systems 17*. The MIT Press, 2005: 545-552.

[68] Sundararajan Sellamanickam, Priyanka Garg, Sathiya Keerthi Selvaraj. "A Pairwise Ranking Based Approach to Learning with Positive and Unlabeled Examples." In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 2011: 663-672.

[69] Grady Booch. Object Oriented Design with Applications. Redwood City, California: Benjamin-Cummings Publishing Co., Inc., 1991.

[70] Paul Duvall, Stephen M. Matyas, Andrew Glover. Continuous Integration: Improving Software Quality and Reducing Risk. Addison-Wesley Professional, 2007.

[71] Michael Hilton, Timothy Tunnell, Kai Huang, Darko Marinov, Danny Dig. "Usage, Costs, and Benefits of Continuous Integration in Open-source Projects." In: *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*. New York, New York, 2016: 426-437.

[72] Ahmed E. Hassan, Ken Zhang. “Using Decision Trees to Predict the Certification Result of a Build.” In: *Proceedings of the 21st IEEE/ACM International Conference on Automated Software Engineering*. 2006: 189-198.

[73] Lech Madeyski, Marcin Kawalerowicz. “Continuous Defect Prediction: The Idea and a Related Dataset.” In: *Proceedings of the 14th International Conference on Mining Software Repositories*. Piscataway, New Jersey, 2017: 515-518.

[74] Jacqui Finlay, Russel Pears, Andy M. Connor. “Data Stream Mining for Predicting Software Build Outcomes Using Source Code Metrics.” *Information and Software Technology*, 2014, 56(2): 183-198.

[75] Ansong Ni, Ming Li. “Cost-effective Build Outcome Prediction Using Cascaded Classifiers.” In: *Proceedings of the 14th International Conference on Mining Software Repositories*. Piscataway, New Jersey, 2017: 455-458.

[76] Andrew B. Goldberg, Ming Li, Xiaojin Zhu. “Online Manifold Regularization: A New Learning Setting and Empirical Study.” In: *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg, 2008: 393-407.

[77] Andrew Goldberg, Xiaojin Zhu, Alex Furger, Jun-Ming Xu. “OASIS: Online Active Semi-Supervised Learning.” In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence*: 1. AAAI Press, 2011: 362-367.

[78] Ying Liu, Zhen Xu, Chunguang Li. “Online Semi-Supervised Support Vector Machine.” *Information Sciences*, 2018, 439-440: 125-141.

[79] Zheng Xie, Ming Li. “Semi-Supervised AUC Optimization Without Guessing Labels of Unlabeled Data.” In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI Press, 2018: 4310-4317.

[80] Bo Han, Quanming Yao, Tongliang Liu, et al. “A Survey of Label-noise Representation Learning: Past, Present and Future.” *arXiv preprint arXiv:2011.04406*, 2021.

[81] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee. “Learning From Noisy Labels With Deep Neural Networks: A Survey.” *IEEE Trans-*

actions on Neural Networks and Learning Systems, 2022: 1-19.

[82] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, Ghyslain Gagnon. “Multiple Instance Learning: A Survey of Problem Characteristics and Applications.” *Pattern Recognition*, 2018, 77: 329-353.

[83] Zhi-Hua Zhou. “A Brief Introduction to Weakly Supervised Learning.” *National Science Review*, 2017, 5(1): 44-53.

[84] Yu-Feng Li, Lan-Zhe Guo, Zhi-Hua Zhou. “Towards Safe Weakly Supervised Learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43: 334-346.

[85] Zhen-Yu Zhang, Peng Zhao, Yuan Jiang, Zhi-Hua Zhou. “Learning From Incomplete and Inaccurate Supervision.” *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(12): 5854-5868.

[86] Chen Gong, Jian Yang, Jane You, Masashi Sugiyama. “Centroid Estimation with Guaranteed Efficiency: A General Framework for Weakly Supervised Learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 44(6): 2841-2855.

[87] Tongliang Liu, Dacheng Tao. “Classification with Noisy Labels by Importance Reweighting.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38: 447-461.

[88] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, Li Fei-Fei. “MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels.” In: *Proceedings of the 35th International Conference on Machine Learning*. 2018: 2304-2313.

[89] Mengye Ren, Wenyuan Zeng, Bin Yang, Raquel Urtasun. “Learning to Reweight Examples for Robust Deep Learning.” In: *Proceedings of the 35th International Conference on Machine Learning*. 2018: 4334-4343.

[90] Hao Yang, Youzhi Jin, Ziyin Li, et al. “Learning from Noisy Labels via Dynamic Loss Thresholding.” *arXiv preprint arXiv:2104.02570*, 2021.

[91] Nan Lu, Gang Niu, Aditya K. Menon, Masashi Sugiyama. “On the Minimal Supervision for Training Any Binary Classifier from Only Unlabeled Data.” In: *Proceedings of the International Conference on Learning Representations*. 2019.

[92] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, Ambuj Tewari. “Learning with Noisy Labels.” In: *Advances in Neural Information Processing Systems 26*. 2013: 1196-1204.

[93] Zhilu Zhang, Mert Sabuncu. “Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels.” In: *Advances in Neural Information Processing Systems 31*. 2018: 8792-8802.

[94] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, James Bailey. “Symmetric Cross Entropy for Robust Learning With Noisy Labels.” In: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*. 2019: 322-330.

[95] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, Xiaogang Wang. “Learning From Massive Noisy Labeled Data for Image Classification.” In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. 2015: 2691-2699.

[96] Jacob Goldberger, Ehud Ben-Reuven. “Training Deep Neural-Networks Using a Noise Adaptation Layer.” In: *Proceedings of the International Conference on Learning Representations*. 2017.

[97] Bo Han, Jiangchao Yao, Gang Niu, et al. “Masking: A New Perspective of Noisy Supervision.” In: *Advances in Neural Information Processing Systems 31*. 2018: 5841-5851.

[98] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, Nathan Silberman. “Learning From Noisy Labels by Regularized Estimation of Annotator Confusion.” In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: 11244-11253.

[99] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, Sanjiv Kumar. “Does Label Smoothing Mitigate Label Noise?” In: *Proceedings of the 37th International Conference on Machine Learning*. 2020: 6448-6458.

[100] Aditya Krishna Menon, Brendan Van Rooyen, Cheng Soon Ong, Robert C. Williamson. “Learning from Corrupted Binary Labels via Class-Probability Estimation.” In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015: 125-134.

[101] Aritra Ghosh, Himanshu Kumar, P. S. Sastry. “Robust Loss Functions under Label Noise for Deep Neural Networks.” In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. AAAI Press, 2017: 1919-1925.

[102] Jessa Bekker, Jesse Davis. “Learning from Positive and Unlabeled Data: A Survey.” *Machine Learning*, 2020, 109(4): 719-760.

[103] Bing Liu, Wee Sun Lee, Philip S. Yu, Xiaoli Li. “Partially Supervised Classification of Text Documents.” In: *Proceedings of the 19th International Conference on Machine Learning*. 2002: 387-394.

[104] Xiaoli Li, Bing Liu. “Learning to Classify Texts Using Positive and Unlabeled Data.” In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. 2003: 587-592.

[105] Wee Sun Lee, Bing Liu. “Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression.” In: *Proceedings of the 20th International Conference on Machine Learning*. 2003: 448-455.

[106] Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, John R. Leathwick. “Presence-Only Data and the EM Algorithm.” *Biometrics*, 2008, 65(2): 554-563.

[107] Charles Elkan, Keith Noto. “Learning Classifiers from Only Positive and Unlabeled Data.” In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008: 213-220.

[108] Marthinus C du Plessis, Gang Niu, Masashi Sugiyama. “Analysis of Learning

from Positive and Unlabeled Data." In: *Advances in Neural Information Processing Systems 27*. 2014: 719-760.

[109] Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, Masashi Sugiyama. "Positive-Unlabeled Learning with Non-Negative Risk Estimator." In: *Advances in Neural Information Processing Systems 30*. 2017: 1674-1684.

[110] Chen Gong, Hong Shi, Tongliang Liu, Chuang Zhang, Jian Yang, Dacheng Tao. "Loss Decomposition and Centroid Estimation for Positive and Unlabeled Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43: 918-932.

[111] Marthinus C du Plessis, Gang Niu, Masashi Sugiyama. "Convex Formulation for Learning from Positive and Unlabeled Data." In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015: 1386-1394.

[112] Thomas G. Dietterich, Richard H. Lathrop, Tomás Lozano-Pérez. "Solving the Multiple Instance Problem with Axis-Parallel Rectangles." *Artificial Intelligence*, 1997, 89(1): 31-71.

[113] Stuart Andrews, Ioannis Tsochantaridis, Thomas Hofmann. "Support Vector Machines for Multiple-Instance Learning." In: *Advances in Neural Information Processing Systems 15*. 2002: 577-584.

[114] Marc-André Carbonneau, Eric Granger, Alexandre J. Raymond, Ghyslain Gagnon. "Robust Multiple-Instance Learning Ensembles Using Random Subspace Instance Selection." *Pattern Recognition*, 2016, 58: 83-99.

[115] Yanshan Xiao, Bo Liu, Zhifeng Hao. "A Sphere-Description-Based Approach for Multiple-Instance Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(2): 242-257.

[116] Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, Alex J. Smola. "Multi-Instance Kernels." In: *Proceedings of the 19th International Conference on Machine Learning*. 2002: 179-186.

[117] Xiu-Shen Wei, Jianxin Wu, Zhi-Hua Zhou. "Scalable Algorithms for Multi-

Instance Learning." *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28(4): 975-987.

[118] Maximilian Ilse, Jakub Tomczak, Max Welling. "Attention-based Deep Multiple Instance Learning." In: *Proceedings of the 35th International Conference on Machine Learning*: vol. 80. 2018: 2127-2136.

[119] Xinggang Wang, Yongluan Yan, Peng Tang, Wenyu Liu, Xiaojie Guo. "Bag Similarity Network for Deep Multi-Instance Learning." *Information Sciences*, 2019, 504: 578-588.

[120] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, et al. "A Closer Look at Memorization in Deep Networks." In: *Proceedings of the 34th International Conference on Machine Learning*. 2017: 233-242.

[121] Nontawat Charoenphakdee, Jongyeong Lee, Masashi Sugiyama. "On Symmetric Losses for Learning from Corrupted Labels." In: *Proceedings of the 36th International Conference on Machine Learning*. 2019: 961-970.

[122] Zhi-Hua Zhou, Jun-Ming Xu. "On the Relation between Multi-Instance Learning and Semi-Supervised Learning." In: *Proceedings of the 24th International Conference on Machine Learning*. 2007: 1167-1174.

[123] Soumya Ray, Mark Craven. "Supervised versus Multiple Instance Learning: An Empirical Comparison." In: *Proceedings of the 22nd International Conference on Machine Learning*. 2005: 697-704.

[124] Razvan C. Bunescu, Raymond J. Mooney. "Multiple Instance Learning for Sparse Positive Bags." In: *Proceedings of the 24th International Conference on Machine Learning*. 2007: 105-112.

[125] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, Quoc V Le. "Estimating Labels from Label Proportions." In: *Proceedings of the 25th International Conference on Machine Learning*. 2008: 776-783.

[126] Linli Xu, James Neufeld, Bryce Larson, Dale Schuurmans. "Maximum Margin Clustering." In: *Advances in Neural Information Processing Systems 17*. 2004:

1537-1544.

[127] Andreas Krause, Pietro Perona, Ryan Gomes. “Discriminative Clustering by Regularized Information Maximization.” In: *Advances in Neural Information Processing Systems 23*. 2010: 775-783.

[128] Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, Shih-Fu Chang. “On Learning from Label Proportions.” *arXiv preprint arXiv:1402.5902*, 2014.

[129] Clayton Scott, Jianxin Zhang. “Learning from Label Proportions: A Mutual Contamination Framework.” In: *Advances in Neural Information Processing Systems 35*. 2020: 22256-22267.

[130] Kuen-Han Tsai, Hsuan-Tien Lin. “Learning from Label Proportions with Consistency Regularization.” In: *Proceedings of the 12th Asian Conference on Machine Learning*: vol. 129. 2020: 513-528.

[131] Nan Lu, Shida Lei, Gang Niu, Issei Sato, Masashi Sugiyama. “Binary Classification from Multiple Unlabeled Datasets via Surrogate Set Classification.” In: *Proceedings of the 38th International Conference on Machine Learning*: vol. 139. 2021: 7134-7144.

[132] Zhishuai Guo, Yan Yan, Zhuoning Yuan, Tianbao Yang. “Fast Objective & Duality Gap Convergence for Non-Convex Strongly-Concave Min-Max Problems with PL Condition.” *Journal of Machine Learning Research*, 2023, 24(148): 1-63.

[133] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, David Ha. “Deep Learning for Classical Japanese Literature.” *arXiv preprint arXiv:1812.01718*, 2018.

[134] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

[135] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. “Deep Residual Learning for Image Recognition.” In: *Proceedings of the 2016 IEEE Conference on*

Computer Vision and Pattern Recognition. 2016: 770-778.

[136] Diederik P Kingma, Jimmy Ba. “Adam: A Method for Stochastic Optimization.” In: *Proceedings of the International Conference on Learning Representations*. 2015.

[137] Adam Paszke, Sam Gross, Francisco Massa, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In: *Advances in Neural Information Processing Systems 32*. 2019: 8024-8035.

攻读博士学位期间研究成果

攻读博士学位期间发表的学术论文

[1] **Zheng Xie**, Ming Li. “Semi-Supervised AUC Optimization without Guessing Labels of Unlabeled Data.” In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2018. (*CCF-A 类会议*)

[2] **Zheng Xie**, Ming Li. “Cutting the Software Building Efforts in Continuous Integration by Semi-Supervised Online AUC Optimization.” In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 2018. (*CCF-A 类会议*)

[3] **Zheng Xie**, Hui Sun, Ming Li. “Semi-Supervised Learning with Support Isolation by Small-Paced Self-Training.” In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. 2023. (*CCF-A 类会议*)

[4] Hui Sun, **Zheng Xie**, Xin-Ye Li, Ming Li. “Cooperative and Adversarial Learning: Co-Enhancing Discriminability and Transferability in Domain Adaptation.” In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. 2023. (*CCF-A 类会议*)

[5] **解铮**, 黎铭. “基于代价敏感间隔分布优化的软件缺陷定位.” 软件学报, 2017, 28(11). (*CCF-A 类中文期刊*)

攻读博士学位期间在投的学术论文

[6] **Zheng Xie**, Yu Liu, Hao-Yuan He, Ming Li, Zhi-Hua Zhou. “Weakly Supervised AUC Optimization: A Unified Partial AUC Approach.” *Under review*, 2023.

[7] **Zheng Xie**, Yu Liu, Ming Li. “AUC Optimization from Multiple Unlabeled Datasets.” *Under review*, 2023.

[8] Hao-Yuan He, Yu Liu, Ren-Biao Liu, **Zheng Xie**, Ming Li. “Probabilistic Instance Dependent Label Refinement for Noisy Label Learning.” *Under review*, 2023.

[9] Ren-Biao Liu, Chao-Zhi Zhang, Jiang-Tian Xue, **Zheng Xie**, Ming Li. “Towards Mitigating Noisy Message Propagation in Graph-based Semi-Supervised Learning.” *Under review*, 2023.

[10] Ren-Biao Liu, Jiang-Tian Xue, Zi-Yu Mao, Zhi-Cun Lv, **Zheng Xie**, Ming Li. “GAGN: Generative Augmented Graph Network for Graph based Semi-Supervised Learning.” *Under review*, 2023.

[11] Xin-Ye Li, Jiang-Tian Xue, **Zheng Xie**, Ming Li. “Think Outside the Code: Brainstorming Boosts Large Language Models in Code Generation.” *Under review*, 2023.

[12] Yi-Fan Ma, Yali Du, **Zheng Xie**, Ming Li. “Learning Unified Semantic Space via Triplet Comparison for Cross-Project Bug Localization.” *Under review*, 2023.

攻读博士学位期间参与的科研课题

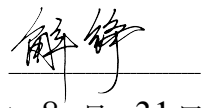
1. 国家自然科学基金面上项目 “基于机器学习的小样本软件缺陷检测技术的研究” （61272217）
2. 国家自然科学基金优秀青年科学基金项目 “半监督学习及其在软件缺陷检测中的应用” （61422304）
3. 国家重点研发计划课题 “智能无人集群系统全局规划及协同行为管控” （2017YFB1001903）
4. 国家自然科学基金创新研究群体项目 “面向开放动态环境的机器学习” （61921006）
5. 国家自然科学基金面上项目 “面向开放动态环境的软件自适应学习研究” （62076121）

攻读博士学位期间获得的主要奖励与荣誉

1. OPPO 首届人工智能 6G 大赛三等奖，2022 年
2. 南京大学优秀博士生提升计划，2021 年
3. 南京大学人工智能奖学金，2018 年
4. AAAI 学生奖学金，2018 年
5. 南京大学海航奖学金，2017 年

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名： 

2023 年 8 月 31 日

论文题名	弱标记 AUC 优化研究				
研究生学号	DG1833006	所在院系	计算机科学与技术系	学位年度	2023
论文级别	□ 学术学位硕士 □ 专业学位硕士 ☑ 学术学位博士 □ 专业学位博士				
作者 Email	xiez@lamda.nju.edu.cn				
导师姓名	黎铭				

论文涉密情况：

☑ 不保密

□ 保密，保密期（_____年____月____日至_____年____月____日）