迭代式的深度 PU 学习与类别先验估计框架

赵昀睿[1] 许倩倩[2] 姜阳邦彦[3),4)] 黄庆明[1),2),5),6)]

[1)](中国科学院大学计算机科学与技术学院 北京 101408)

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

[3)](中国科学院信息工程研究所信息安全国家重点实验室 北京 100093)

[4)](中国科学院大学网络空间安全学院 北京 100049)

[5)](中国科学院大学大数据挖掘与知识管理重点实验室 北京 101408)

[6)](鹏城实验室 广东 深圳 518055)

摘 要 近年来,深度学习在诸多任务上展现了优异的性能,其一般基于海量数据并采用有监督的学习方式,依赖于完整的数据标签信息.然而在现实应用场景中,收集大量标签往往成本高昂.因此,如何利用未经充分标注的数据进行学习成为了当下的主要挑战.二分类问题中的从正例和无标签(Positive-Unlabeled,PU)样本数据进行学习,简称 PU 学习,即为其一.当前主流的 PU 学习算法需要准确无误的类别先验知识,但实际上类别先验通常难以获得,需要估计.已有的类别先验估计算法则主要面向传统的机器学习分类器进行设计,无法直接运用在大规模数据集上,因而不利于发挥深度学习在大规模数据集上的优势.为克服以上问题,本文提出了一个基于无监督混合模型的迭代式深度 PU 学习与类别先验估计框架.它利用了深度神经网络对正例和负例给出的预测分数具有不同的分布这一特性,使用双高斯成分的混合模型近似拟合预测分数的混合分布.其中,各个高斯分量分别代表了正类和负类的条件概率分布,混合权重系数代表了类别先验.结合半监督学习中的平均教师和温度锐化技术,所提框架在类别先验未知以及数据缺失负例监督的条件下,估计类别先验的同时进行 PU 数据上的深度学习,二者相互促进.在基准数据集 MNIST、Fashion-MNIST、CIFAR-10 和实际应用数据集 Alzheimer 上的实验结果验证了所提框架的有效性,准确率分别为 94.66%、95.16%、89.98%和 73.20%,该结果不仅超越了现有基于类别先验估计的 PU 学习算法,更可与基于真实类别先验的最前沿算法相媲美.

关键词 PU 学习;类别先验估计;半监督学习;弱监督学习;深度学习

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02667

An Iterative Framework for Deep PU Learning and Class Prior Estimation

ZHAO Yun-Rui[1)] XU Qian-Qian[2)] JIANG Yang-Bang-Yan[3),4)] HUANG Qing-Ming[1),2),5),6)]

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

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

[3)](*State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

[4)](*School of Cyber Security, University of Chinese Academy of Sciences, Beijing* 100049)

[5)](*Key Laboratory of Big Data Mining and Knowledge Management (BDKM), University of Chinese Academy of Sciences, Beijing* 101408)

[6)](*Peng Cheng Laboratory, Shenzhen, Guangdong* 518055)

Abstract Deep learning models have achieved superior performance with a large amount of data. Such success usually depends on complete label information in a fully-supervised training style. However, collecting all the labels can be very expensive. Consequently, weakly supervised learning,

收稿日期:2022-02-09;在线发布日期:2022-08-27.本课题研究由科技创新 2030-"新一代人工智能"重大项目(2018AAA0102000)、国家自然科学基金项目(U21B2038,61931008,6212200758,61976202)、中央高校基本科研业务费专项资金、中科院青促会会员项目以及中国科学院战略性先导科技专项(XDB28000000)资助.**赵昀睿**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究方向为弱监督学习.E-mail:zhaoyunrui20@mails.ucas.ac.cn.**许倩倩**(通信作者),博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为统计机器学习及其在多媒体领域的应用.E-mail:xuqianqian@ict.ac.cn.**姜阳邦彦**,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为机器学习与计算机视觉.**黄庆明**(通信作者),博士,讲席教授,中国计算机学会(CCF)会士,主要研究领域为多媒体计算、图像处理、计算机视觉和模式识别.E-mail:qmhuang@ucas.ac.cn.

which aims at learning with incomplete, inexact, or inaccurate supervision, has attracted the machine learning community in the past decade. One of these real-world applications could be Positive-Unlabeled (PU) learning, where we have to train binary classifiers from a few positive examples with much more unlabeled data. Current state-of-the-art PU methods rely on the ground—truth class prior (i.e., the proportion of the positive samples to the unlabeled data), which is hard to obtain and needs to be estimated in practice. While previous studies of class prior estimation mainly focused on traditional machine learning models, few could deal with a relatively large-scale dataset or be applied to deep learning scenarios. To solve such problems, we propose an iterative framework for deep PU learning and class prior estimation utilizing an unsupervised mixture model in the paper. Specifically, positive and negative classes are supposed to have distinct predicted score distributions intuitively. We investigate and demonstrate our intuition and approximate the score distributions by a Gaussian Mixture Model with two components. Each component represents a relevant class-conditional distribution, and the positive weight could be approximate to the ground truth class prior. We further incorporate techniques such as mean teacher and temperature sharpening from semi-supervised learning to stabilize the whole process and boost performance. Our proposed framework could estimate the class prior and learn from PU data simultaneously, achieving well-matched performance with other PU competitors based on the ground-truth prior. Experiments on three benchmark datasets (i.e., MNIST, Fashion-MNIST, and CIFAR-10) and one practical application (i.e., Alzheimer) validate the effectiveness of our framework. Our algorithm reaches the accuracy of 94.66%, 95.16%, 89.98%, and 73.20%, respectively, with limited labels and unknown class prior.

Keywords positive-unlabeled learning; class prior estimation; semi-supervised learning; weakly supervised learning; deep learning

1 引 言

近年来,随着互联网与信息技术的发展,人类进入了大数据时代.机器学习,特别是基于海量数据的深度学习,受到了广泛关注并取得了突破性进展,在诸多任务上的表现[1-3]达到甚至超越了人类水平.然而,深度学习算法通常采用有监督的训练方式,其取得的优异性能依赖于海量标注数据,尤其离不开完整类别标签信息的指导.当数据中的标签数目较少时,深度监督学习算法容易产生过拟合现象,难以发掘隐含在数据背后的规律,模型的泛化性能因之大打折扣.另一方面,对于许多实际应用场景而言,数据标注代价高昂[4].在人力和物力有限的条件下,人们仅能获取少量数据的标签.鉴于此,对标注数据依赖更少的学习方式成为当下热门,半监督学习即为其中之一.

半监督学习最早由 Merz 等人[5]提出并应用在分类问题上,其基于少量的标注数据,旨在挖掘无标签数据中蕴含的信息,用以辅助提升模型的性能.尽管半监督学习在一定程度上减轻了对标注数据的依赖,但是一般的半监督学习算法要求标注数据中同时包含各个类别,限制了其普适性.对于部分二分类的实际应用问题,除了数据标注任务本身较为困难之外,还存在负类标签难以获取的问题.此时,常规的半监督学习算法将因缺失负例监督而失效.例如在阿尔兹海默症(Alzheimer's Disease,AD)的诊断问题[6]中,病理特征明显的患者容易被临床确诊,可以看作有标签的正例;而早期 AD 病患与正常人无殊,难以被临床确诊.也就是说,未被确诊的人仍旧存在患病可能,属于无标签样本.类似情况还存在于恶意 URL 检测[7]、虚假评论检测[8]、冷冻电镜的粒子拾取[9]等任务上.由此可见,在二分类半监督学习中,进一步考虑仅利用正例和无标签样本进行学习(Positive-Unlabeled Learning,PU),简称为 PU 学习,具有重要的研究价值,近年来受到了机器学习社区的广泛关注[10].

如图 1 所示,相较于一般的二分类半监督学习,PU 学习的标注数据缺失负例信息,因而更具挑战性.PU 学习算法根据其对无标签数据的处理方式,

可分为两大流派，分别是两阶段式的 PU 学习和基于风险重构的 PU 学习. 其中，两阶段式的 PU 学习符合思维惯性，该类方法[11-13]首先找出可靠的负例，从而将 PU 问题转换为常规的二分类半监督学习问题后，再运用(半)监督学习算法. 然而，此类方法依赖于不同类别数据分布的严格可分性和较为理想的标签分配机制. 因此，基于风险重构的 PU 学习[6,14-16]成为主流方法，其将二分类监督学习的风险重构为关于正例和无标签样本的风险，从而无需事先在无标签数据中识别出可靠的负例. nnPU[16]成功将基于风险重构的 PU 学习扩展到了深度场景，Self-PU[6]则结合自步(Self-Paced)学习的训练策略和深度模型的自校准(Self-Calibration)与自蒸馏(Self-Distillation)技术，在 PU 学习的基准数据集上取得了最先进的分类性能.

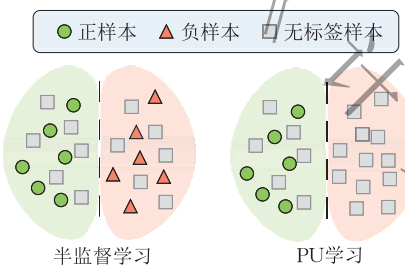


图 1　二分类半监督学习与 PU 学习的联系与区别

尽管基于风险重构的 PU 学习取得了巨大成功，但该类方法严重依赖于真实的类别先验知识，否则将无法构造 PU 学习的无偏风险估计器. 对于多数实际问题而言，类别先验往往未知，这大大降低了此类 PU 学习算法的实用价值. 那么，可否事先随意设置一个类别先验值，再应用此类算法呢?答案是否定的. 实验表明(如图 2 所示)，不准确的类别先验预设值将会严重损害算法性能. 基于风险重构的 PU 学习与类别先验的依赖关系，驱动了类别先验估计的相关研究[17-20]. 但这些方法大多仅能在小规模数据集和传统机器学习方法上达到较优水平，无法直接应用在基于海量数据的深度学习方法上，需要额外进行主成分分析(Principal Components Analysis，PCA)和随机采样等操作[19].

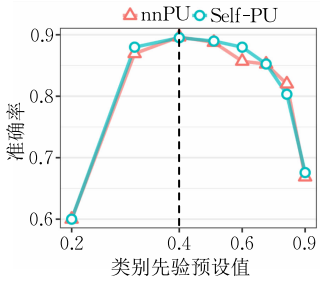


图 2　基于风险重构的 PU 算法 nnPU[16]和 Self-PU[6]在 CIFAR-10 上的准确率随类别先验预设值的变化曲线(黑色虚线为类别先验真实值(0.4). 总体来看，类别先验预设值越偏离真实值，算法的准确率越差)

在二分类深度神经网络过拟合前，其预测分数的分布关于正、负类别存在差异，它们的混合分布呈现双峰状[22]. 基于上述观察，本文提出了一种迭代式的深度 PU 学习与类别先验估计框架，其由类别先验估计模块、PU 学习模块和迭代过程稳定模块构成. 其中，类别先验估计模块基于双成分的高斯混合模型(Gaussian Mixture Model，GMM)，对深度模型的预测分数进行无监督式建模，以估计基于风险重构的 PU 学习所必需的类别先验，且不受 PU 数据缺失负类标签的限制. 此外，运用 EM 算法求解 GMM 所需的时间复杂度、空间复杂度均为 $O(N)$，占用的计算资源少. 因此，所提方法可广泛应用于各种规模的数据集，进而可以有效应对深度场景下 PU 数据的类别先验未知问题.

综上所述，本文的主要贡献总结如下：

(1)本文提出了一种迭代式的深度 PU 学习与类别先验估计框架. 该框架利用了深度 PU 学习与类别先验估计的相互促进关系，能够有效应对深度场景下 PU 数据的类别先验未知问题.

(2)所提框架的核心思想在于使用 GMM 对深度神经网络输出的预测分数建模，进而估计类别先验. 在深度神经网络其对训练数据过拟合之前，正、负类的预测分数条件分布可分别近似为两条钟形曲线，其混合分布可用 GMM 拟合.

(3)在所提框架中，结合使用半监督学习常用的平均教师(Mean Teacher)[23]和温度锐化(Temperature Sharpening)技术[24]，将二者作为迭代过程稳定模块，以平稳类别先验的估计过程，进而提升模型的分类性能.

(4)本文在三个基准数据集 MNIST、F-MNIST 和 CIFAR-10 以及实际应用数据集 Alzheimer 上进行了详细的对比实验、消融实验和超参数敏感度分析，验证了所提框架的有效性及其设计的合理性.

2　相关工作

2.1　PU 学习

PU 学习方法主要可以分为两大流派，分别是基于风险重构的 PU 学习和两阶段式的 PU 学习. 当前主流方法基于风险重构的思想，在优化过程中

构造了 PU 数据关于二分类监督学习的风险无偏估计器，其基本形式如下：

$$R=\pi_P R_P^+ + R_U^- - \pi_P R_P^- \tag{1}$$

其中，$\pi_P=\Pr(y=1)$ 表示（正）类别先验，简称类别先验，R_P^+ 表示将正类样本预测成正类的期望风险，同理 R_P^- 表示正类样本预测成负类的期望风险，R_U^- 则表示将无标签样本预测成负类的期望风险. Du Plessis 等人于 2014 年首先提出了 PU 学习的无偏风险估计（unbiased PU Risk Estimator，uPU）[14]，并进一步指出训练过程中使用满足特定条件的代理损失函数（Surrogate Loss Function）可以有效降低计算成本[15]. 自此以后，机器学习社区涌现出许多致力于增强或改进 uPU 的工作. 深度神经网络以其强大的拟合能力著称，那么直接优化 PU 学习的无偏经验风险可能导致 R_N^- 的无偏估计值小于 0. 为了防止深度神经网络过度优化将无标签样本预测成负类的风险，即 R_U^-，Kiryo 等人在 uPU 中关于 R_N^- 的无偏估计器上增加了非负约束，由此提出了 PU 学习的非负风险估计器（non-negative PU Risk Estimator，nnPU）[16]. 此外，Self-PU[6] 在 nnPU 的基础上引入了自监督学习（Self-Supervision Learning）任务，通过导师网络（Mentor Net）[25] 和平均教师（Mean Teacher）[23] 的架构，实现深度模型的自校准和自蒸馏，并结合自步学习的训练策略，在 PU 学习的基准数据集上取得了最优性能. PUSB[26] 提出了次序不变假设（Invariance of Order），即样本被标注的概率与该样本是正例的概率正相关. 并基于该假设，从理论上证明了当 uPU 使用对数损失函数时，其贝叶斯最优解满足预测分数与正类条件概率正相关. aPU[27] 进一步拓展到训练数据与测试数据的正类条件分布不一致的情况. ImbPU[28] 基于风险重构的思想构造了等价于对少数类进行过采样的优化目标，以应对 PU 数据的类别不平衡问题.

PU 学习的另一大流派是两阶段式的 PU 学习，这类方法符合人们的惯性思维，首先识别出一定数目的可靠负例后，便可在第二阶段运用（半）监督学习算法. 由此可见，两阶段式 PU 学习的研究重心和主要区别在于可靠负例的识别机制. 文献[29]从被标注正例中随机选取一定数目的间谍（Spy），将之加入无标签样本集合中；这样一来，可靠的负例即为后验概率小于任一间谍的无标签样本. 文献[30]基于图进行标签传播，利用高斯核构建邻接矩阵，欧式空间上特征相似的样本将被划分至同一类别. 文献[31]使用朴素贝叶斯对样本被标注的条件概率进行建模，再依据上述条件概率熵的大小，筛选出可靠的正例或负例. 文献[32]使用 k-means 进行聚类，可靠负例即来自距离被标注正例最远的样本簇. RP[33] 通过计算各个样本的模型预测置信度得分，以筛选出高置信度的样本，它们的类别即为模型的预测结果. 文献[34]从生成学习（Generative Learning）的角度，在生成对抗网络（Generative Adversarial Network，GAN）的框架下联合使用了正例生成器和负例生成器，以及正例判别器、无标签样本判别器和负例鉴别器. 其中，无标签样本判别器用以保证生成数据的合理性. KLDCE[35] 首先将 PU 学习问题转化为标签噪声问题，然后通过对有噪声的负例集合进行质心估计来消减标签噪声的副作用. 基于实例依赖（Instance Dependent）的 PU 学习[36] 将类别标签视为由样本输入特征确定的隐变量，并借助 EM 的算法框架学习分类器. 两阶段式的 PU 学习依赖于所选样本的数量及其可靠性，因此要求不同类别的数据分布严格可分和较为理想的标签分配机制.

2.2 类别先验估计

绝大多数的 PU 学习方法依赖于类别先验知识，即 π_P. 然而，在实际应用中，类别先验往往未知. 为此，针对 PU 数据的类别先验估计成为重要的研究课题. 其中，PE[17] 基于部分匹配准则，最小化无标签样本分布和被标注正例分布之间的 Pearson 距离. 其中，类别先验即为最优匹配时有标签正例分布的标度因子. KM1 和 KM2[18] 提出了一种基于核嵌入的类别先验估计方法，将正类分布从总体分布中分离出来. 类别先验即为最小负例比例低于给定阈值的最大值. TIcE[19] 指出当完全随机选择（Selected Completely At Random，SCAR）假设成立时，给定任一数据子域，在该子域中样本被标记的频率是正类样本标签频率的下确界. 通过决策树归纳，该方法旨在找到令样本被标记频率最大的子域，以逼近真实的标签频率，进而估计类别先验. CDMM[20] 基于逻辑回归，选取其参数对数似然的上界凹函数，并通过循环坐标下降（Cyclic Coordinate Descent，CD）和 MM 优化算法（Minimization-Majorization，MM）近似求解上述上界凹函数，从而估计正类标签频度，进而估计类别先验.

已有的类别先验估计算法[17-20] 往往采用非深度的传统机器学习算法，尽管在小规模数据集上成效卓著，但无法直接应用于深度学习场景下的大规模数据集. 此类算法的核心思想是混合比例估计，即无标签样本的概率分布是正类样本分布和负类样本

分布的线性组合，其中的混合系数表示类别先验. 当不可约(Irreducibility)假设[21]成立时，正类样本的支撑集(Support)不能包含于负类样本的支撑集中，那么类别先验可以由混合概率密度与正类样本概率密度比值的最小值估计得到. 由此可见，基于混合比例估计的类别先验估计算法具备较为可靠的理论保障，并且在低维、小规模数据上得到了广泛验证. 然而，由于“维数灾难”和计算资源有限等问题，高维、大规模数据上的概率密度估计十分困难，若是简单地对数据进行降维/采样处理，则可能会破坏不可约假设，最终致使算法失效. 考虑到深度模型强大的特征提取能力以及模型训练与类别先验估计之间存在相互促进的关系，所提框架在模型输出的一维预测分数上进行类别先验估计，有效降低了计算复杂度. 另外，尽管预测分数浓缩了输入样本的类别信息，但预测分数的分布难以满足不可约假设. 为此，所提框架使用高斯混合模型对预测分数建模，而高斯混合模型可以有效应对不可约假设不成立的情况. 最后，根据半监督学习的平滑假设和聚类假设，所提框架进一步引入了平均教师和温度锐化技术，有效改善了方法的有效性和稳定性.

2.3　半监督学习

半监督学习算法常利用平滑假设、聚类假设和流形假设三大基本假设[37-38]来建立无标签数据的学习器. 近期的深度半监督学习工作主要从平滑假设和聚类假设着手. 平滑假设要求数据稠密区域内距离相近的样本具有相似的类别标签. 基于平滑假设的半监督学习方法在优化目标中添加无监督分量对无标签数据进行正则化. 这类方法简单有效，通过对数据或模型参数进行扰动，利用无标签数据引导网络生成信息更丰富的抽象表示. Rasmus 等人[39]首先提出梯形网络，它将前馈网络作为自动编码器(Auto-Encoder)的编码器，添加解码器，加入重构损失；伪标签集成[40]、Π-Model[41]等对神经网络模型本身进行扰动，计算不同网络输出之间的一致性损失. 平均教师[23]对模型的历史参数进行时序上的集成，取得了明显的性能提升. 聚类假设要求分类决策边界附近的数据分布稀疏. 根据聚类假设，模型在无标签数据上的输出熵应该尽可能小. 半监督深度学习的先进方法 MixMatch[42]、UDA[24]和 ReMixMatch[43]均使用了温度锐化技术，间接降低了模型的输出熵. 流形假设指高维数据可以被映射至的低维流形中，那么局部邻域内的样本应当具有相似的类别标签[44]. 现有的深度半监督模型甚至可在部分基准数据集上用较少的标注数据达到与使用完整标注数据训练相近的性能[45].

除了半监督学习之外，主流的无标签数据处理方法还包括主动学习(Active Learning)、迁移学习(Transfer Learning)等. 其中，主动学习严重依赖于专家知识，需要通过专家与模型之间的频繁交互额外进行人工标注；迁移学习则依赖于迁移前后任务或数据分布之间的相关性，且要求标注数据同时包含正负例. 相较于上述两者，PU 学习作为二分类半监督学习的特例，无需额外的人工标注，不依赖于其它任务或模型，在诸如 AD 诊断问题等负类标签难以获取的特定应用场景下更具潜力.

3　方　法

3.1　问题定义

PU 学习是二分类问题中的一类特殊情况，因此，首先对二分类问题进行回顾. 在二分类问题中，输入空间是一个 d 维实数向量子空间，即 $\mathcal{X}\subseteq\mathbb{R}^d$；标签空间 $\mathcal{Y}=\{0,1\}$，其中 0 表示负类，1 表示正类. 设 $p(\boldsymbol{x},y)$ 为真实数据分布 $(\mathcal{X},\mathcal{Y})$ 的联合概率密度，$p(\boldsymbol{x})$ 为关于某一样本其输入向量 $\boldsymbol{x}$ 的边缘分布，$y\in\mathcal{Y}$ 为其类别标签. 那么，二分类问题中的正例集合和负例集合可以分别表示为

$$\boldsymbol{X}_P=\{\boldsymbol{x}\}^{n_P}\sim p_P(\boldsymbol{x}) \tag{2}$$

$$\boldsymbol{X}_N=\{\boldsymbol{x}\}^{n_N}\sim p_N(\boldsymbol{x}) \tag{3}$$

其中，n_P 和 n_N 分别为正例和负例集合的基数，即对应集合的样本个数；$p_P(\boldsymbol{x})=\Pr(\boldsymbol{x}|y=1)$ 和 $p_N(\boldsymbol{x})=\Pr(\boldsymbol{x}|y=0)$ 分别表示正类和负类的条件概率. 基于上述定义，全体样本的输入向量集合 $\boldsymbol{X}=\boldsymbol{X}_P\cup\boldsymbol{X}_N$ 可以表示为

$$\boldsymbol{X}=\{\boldsymbol{x}\}^{n}\sim p(\boldsymbol{x}) \tag{4}$$

$$p(\boldsymbol{x})=\pi_P\cdot p_P(\boldsymbol{x})+\pi_N\cdot p_N(\boldsymbol{x}) \tag{5}$$

其中，样本总数 $n=n_P+n_N$；$\pi_P=\Pr(y=1)$ 表示(正)类别先验，$\pi_N=1-\pi_P$ 表示负类别先验.

给定深度神经网络结构 f，二分类监督学习的目标是学习模型的参数 $\boldsymbol{\Theta}$，使得分类器 $f(\cdot;\boldsymbol{\Theta}):\mathbb{R}^d\to\mathbb{R}$ 在真实数据分布 $(\mathcal{X},\mathcal{Y})$ 上取得最小的预测误差期望，即

$$\boldsymbol{\Theta}^*=\arg\min_{\boldsymbol{\Theta}}R_{PN},$$
$$R_{PN}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\ell_{0-1}(f(\boldsymbol{x};\boldsymbol{\Theta}),y)\right] \tag{6}$$

其中，$\ell_{0-1}(\hat{y},y)=\dfrac{1-sign(\hat{y}\cdot y-1/2)}{2}$ 表示 0-1 损失函数. 然而，0-1 损失函数不可导，致使 R_{PN} 难以

优化.为此,通过引入代理损失函数 ℓ,近似替代式(6)中的优化目标:

$$R_{PN}^{\ell}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),y)] \tag{7}$$

相较于二分类监督学习,PU 学习的数据集 $\boldsymbol{X}_{\mathrm{PU}}=\boldsymbol{X}_L\cup\boldsymbol{X}_U$ 仅由少量正例和大量的无标签数据构成,其中 $\boldsymbol{X}_L$ 表示有标签的正例集合,$\boldsymbol{X}_U$ 表示无标签样本集合.基于 SCAR 假设[10],即给定样本所属类别 y,该样本被标注的概率与其输入特征 $\boldsymbol{x}$ 无关,以及单一训练集(Single-Training-Set)设置[10],被标注的正样本服从真实的正类条件分布,而全体样本服从真实的数据分布,即:

$$\boldsymbol{X}_L=\{\boldsymbol{x}\}^{n_L}\sim p_P(\boldsymbol{x}) \tag{8}$$

$$\boldsymbol{X}_{\mathrm{PU}}=\{\boldsymbol{x}\}^{n}\sim p(\boldsymbol{x}) \tag{9}$$

3.2 流程框架

本文提出了一种迭代式的深度 PU 学习与类别先验估计框架,以克服深度场景下类别先验未知问题.所提框架由三大模块组成,分别是基于非负 PU 损失的 PU 学习模块、基于 GMM 的类别先验估计模块以及基于平均教师与温度锐化的迭代过程稳定模块,其流程架构如图 3 所示.

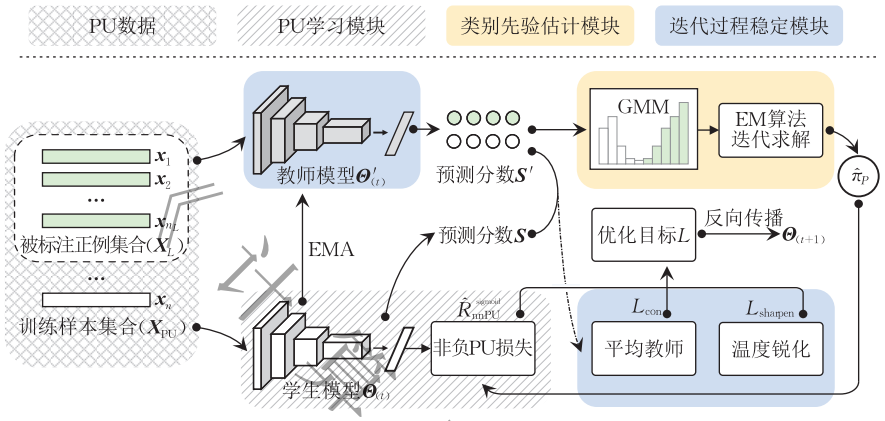


图 3 迭代式的深度 PU 学习与类别先验估计框架

首先,PU 数据的全体样本 $\boldsymbol{X}_{\mathrm{PU}}$ 被同时输入至两个网络结构相同而参数不同的学生模型 $f(\cdot;\boldsymbol{\Theta})$ 和教师模型 $f(\cdot;\boldsymbol{\Theta}')$ 中,教师模型的参数 $\boldsymbol{\Theta}'$ 由学生模型的历史参数进行时序上的指数窗口滑动平均(Exponential Moving Average,EMA)运算得到;由此得到学生模型和教师模型各自对 $\boldsymbol{X}_{\mathrm{PU}}$ 的预测分数 $\boldsymbol{S}=f(\boldsymbol{X}_{\mathrm{PU}};\boldsymbol{\Theta})$ 和 $\boldsymbol{S}'=f(\boldsymbol{X}_{\mathrm{PU}};\boldsymbol{\Theta}')$;之后,类别先验估计模块对 $\boldsymbol{S}'$ 使用 GMM 进行无监督式建模,并通过 EM 算法迭代多步直至收敛,以求解 GMM 的参数,由此得到类别先验估计值 $\hat{\pi}_P$;PU 学习模块再借助 $\hat{\pi}_P$ 计算非负 PU 损失 $\hat{R}_{\mathrm{nnPU}}^{\ell_{\mathrm{sigmoid}}}$;最后结合迭代过程稳定模块中教师模型和学生模型预测分数之间的一致性损失 L_{con},以及温度锐化损失 L_{sharpen},进而得到了框架的总体优化目标 L,并使用梯度反向传播算法更新 $\boldsymbol{\Theta}$.上述步骤迭代反复执行,直至 L 收敛或达到预设迭代次数.接下来将依次对所提框架的各个模块进行详细介绍.

3.3 PU 学习模块

PU 学习模块负责从训练数据中学习其潜在的分类规则.它包含一个基于深度神经网络 f 的分类器,并定义了 f 在 PU 数据上进行训练的优化目标.式(7)给出了二分类监督学习的优化目标 R_{PN}^{ℓ},其可重构为不同类别期望损失的线性组合:

$$R_{PN}^{\ell}=\pi_P R_P^{\ell,+}+\pi_N R_N^{\ell,-} \tag{10}$$

其中,$R_P^{\ell,+}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_P(\boldsymbol{x})}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),1)]$ 表示在正例集合中将样本预测为正类的期望损失,$R_N^{\ell,-}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_N(\boldsymbol{x})}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),0)]$ 表示在负例集合中将样本预测为负类的期望损失.在 SCAR 假设下,式(8)成立,也就是说,$\boldsymbol{X}_L$ 与 $\boldsymbol{X}_P$ 具有相同的类条件分布 $p_P(\boldsymbol{x})$.因此,根据经验风险最小化原则(Empirical Risk Minimization,ERM),$\hat{R}_L^{\ell,+}=\mathop{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{X}_L}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),1)]$ 是关于 $R_P^{\ell,+}$ 的无偏估计.根据式(9),$\boldsymbol{X}_{\mathrm{PU}}$ 与 $\boldsymbol{X}$ 同分布,于是以下等式关系成立:

$$\begin{aligned}R_{\mathrm{PU}}^{\ell,-}&=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),-1)]\\&=\pi_P R_P^{\ell,-}+\pi_N R_N^{\ell,-}\end{aligned} \tag{11}$$

由此可见,$\pi_N R_N^{\ell,-}=R_{\mathrm{PU}}^{\ell,-}-\pi_P R_P^{\ell,-}$ 可由 $\hat{R}_{\mathrm{PU}}^{\ell,-}=\mathop{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{X}_{\mathrm{PU}}}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),0)]$ 和 $\hat{R}_L^{\ell,-}=\mathop{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{X}_L}[\ell(f(\boldsymbol{x};\boldsymbol{\Theta}),0)]$ 间接估计[14].综上所述,可以得到 R_{PN}^{ℓ} 关于 PU 数据的无偏估计,即:

$$\hat{R}_{\mathrm{PU}}^{\ell}=\pi_P\hat{R}_L^{\ell,+}+\hat{R}_{\mathrm{PU}}^{\ell,-}-\pi_P\hat{R}_L^{\ell,-} \tag{12}$$

若代理损失函数 $\ell(\cdot,\cdot)$ 满足 $\ell(\cdot,1)+\ell(\cdot,0)=1$,那么 $\hat{R}_L^{\ell,-}=1-\hat{R}_L^{\ell,+}$ 成立,将之代入式(12),有:

$$\hat{R}_{\mathrm{PU}}^{\ell}=\pi_P\hat{R}_L^{\ell,+}+\hat{R}_{\mathrm{PU}}^{\ell,-}-\pi_P+\pi_P\hat{R}_L^{\ell,+} \tag{13}$$

其中，$\hat{R}_{\mathrm{PU}}^{\ell}$ 仅需计算 $\hat{R}_{L}^{\ell,+}$ 和 $\hat{R}_{\mathrm{PU}}^{\ell,-}$，相较于式(12)不需要计算 $\hat{R}_{L}^{\ell,-}$，因而有效降低了计算复杂度[15].

由于深度神经网络具有强大的拟合能力，在实际训练过程中，模型过度优化 $\hat{R}_{\mathrm{PU}}^{\ell,-}$，以至于关于 $R_{N}^{\ell,-}$ 的无偏估计小于 0[16]，与真实情况下 $R_{N}^{\ell,-}\geqslant 0$ 相悖. 为此，需要向式(13)引入非负约束：

$$\hat{R}_{\mathrm{nnPU}}^{\ell}=\pi_P\hat{R}_{L}^{\ell,+}+\max(0,\hat{R}_{\mathrm{PU}}^{\ell,-}+\pi_P\hat{R}_{L}^{\ell,+}-\pi_P) \tag{14}$$

所提框架的代理损失函数选取 PU 学习中常用的 sigmoid 损失，即：

$$\ell_{\mathrm{sigmoid}}(\hat{y},y)=1-\frac{1}{1+\exp[\hat{y}\cdot(2y-1)]} \tag{15}$$

显然，ℓ_{sigmoid} 满足 $\ell_{\mathrm{sigmoid}}(\cdot,1)+\ell_{\mathrm{sigmoid}}(\cdot,0)=1$. 因此，所提框架中 PU 学习模块的优化目标为

$$\hat{R}_{\mathrm{nnPU}}^{\ell_{\mathrm{sigmoid}}}=\pi_P\hat{R}_{L}^{\ell_{\mathrm{sigmoid}},+}+\max(0,\hat{R}_{\mathrm{PU}}^{\ell_{\mathrm{sigmoid}},-}+\pi_P\hat{R}_{L}^{\ell_{\mathrm{sigmoid}},+}-\pi_P) \tag{16}$$

由上式可知，类别先验 π_P 是构造关于 PU 数据无偏风险估计器的关键. 然而，实际应用场景的 π_P 往往未知，这将导致 PU 学习模块无法正常运作. 为此，本文设计了类别先验估计模块.

3.4　类别先验估计模块

类别先验估计模块利用不同类别分布的可分性和深度神经网络的抽象表征能力，通过由两个高斯成分的 GMM 对全体样本的预测分数进行无监督式地建模. 对于 GMM 中均值较大的高斯成分，它的权重即为 PU 学习模块必需的类别先验. 其中，样本 $\boldsymbol{x}$ 关于模型 $\boldsymbol{\Theta}$ 的预测分数是对条件概率 $\Pr(y=1|\boldsymbol{x})$ 的建模，其定义如下：

$$s=\mathrm{sigmoid}[f(\boldsymbol{x};\boldsymbol{\Theta})]\in[0,1] \tag{17}$$

混合模型是一类无监督建模方法[46-48]，其中，GMM[48]的应用最为广泛. 使用双高斯成分的 GMM 对预测分数建模的依据基于以下观察. 在深度神经网络发生过拟合现象之前，正类的预测分数与负类的预测分数呈现出不同的分布. 如图 4 所示，正类的预测分数集中分布在分值较高的区间，而负类的预测分数集中分布在分值较低的区间，两者分别构成了两条中间高、两端低的钟形曲线；预测分数的混合分布由上述两条钟形曲线加权叠加而成. 由此可见，不同类别的预测分数分布可以近似为不同均值和方差的高斯分布；预测分数的混合分布可近似为上述高斯分布的线性组合. 因此，双高斯成分的 GMM 对预测分数 $\boldsymbol{S}$ 的建模如下：

$$\begin{aligned}\boldsymbol{S}=\{s\}^n&\overset{i.i.d}{\sim}\pi_P\Pr(s|y=1)+\pi_N\Pr(s|y=0)\\&=\pi_P\mathcal{N}_P(s)+\pi_N\mathcal{N}_N(s)\end{aligned} \tag{18}$$

$$\mathcal{N}_P(s)=\mathcal{N}(s|\mu_P,\sigma_P),\mathcal{N}_N(s)=\mathcal{N}(s|\mu_N,\sigma_N) \tag{19}$$

其中，类别标签 y 是隐变量；π_P 既是类别先验，同时也是 GMM 中 $\mathcal{N}_P(s)$ 的混合系数；同理，$\pi_N=1-\pi_P$ 既是负类别先验，同时也是 $\mathcal{N}_N(s)$ 的混合系数；$\mathcal{N}_P(\cdot)$，$\mathcal{N}_N(\cdot)$ 分别表示正类和负类预测分数各自所服从的高斯分布，$\mathcal{N}(s|\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right)$，$\mu,\sigma$ 分别表示其均值和标准差；因为正类的预测分数整体上应当大于负类，所以有 $\mu_N<\mu_P$.

接下来需要对 GMM 中的类别先验 π_P 进行求解. 考虑到 GMM 中包含隐变量 y，选择最大期望算法(Expectation Maximization，EM)[49]进行求解.

3.4.1　EM 算法

EM 算法[49]主要通过期望(Expectation，E)步和最大化(Maximization，M)步反复迭代直至似然函数收敛至局部最优解. 令双高斯成分的所有参数 $\boldsymbol{\Phi}=(\mu_P,\sigma_P,\mu_N,\sigma_N)$，根据极大似然优化准则，首先为其构造对数似然函数：

$$\mathcal{L}(\boldsymbol{S}|\boldsymbol{\Phi},\pi_P,\pi_N)=\sum_{s\in\boldsymbol{S}}\log[\pi_P\mathcal{N}_P(s)+\pi_N\mathcal{N}_N(s)] \tag{20}$$

直接对上式进行优化过于复杂. EM 算法则间接优化 $\mathcal{L}(\boldsymbol{S}|\boldsymbol{\Phi})$ 的一阶近似，有效降低了计算复杂度. 令 $\boldsymbol{\Phi}^{(t_{\mathrm{in}})}=[\mu_P^{(t_{\mathrm{in}})},\sigma_P^{(t_{\mathrm{in}})},\mu_N^{(t_{\mathrm{in}})},\sigma_N^{(t_{\mathrm{in}})}]$，$\pi_P^{(t_{\mathrm{in}})},\pi_N^{(t_{\mathrm{in}})}$ 表示 EM 算法经过 t_{in} 次迭代后 GMM 的参数，$\mathcal{N}_P^{(t_{\mathrm{in}})},\mathcal{N}_N^{(t_{\mathrm{in}})}$ 表示 t_{in} 次迭代后的高斯成分，EM 算法在 $t_{\mathrm{in}}+1$ 次迭代中首先执行 E 步，即计算给定 s 的类条件概率：

$$\begin{aligned}Q_P^{(t_{\mathrm{in}})}(s)&=\Pr(y=1|s;\boldsymbol{\Phi}^{(t_{\mathrm{in}})})\\&=\frac{\pi_P^{(t_{\mathrm{in}})}\mathcal{N}_P^{(t_{\mathrm{in}})}(s)}{\pi_P^{(t_{\mathrm{in}})}\mathcal{N}_P^{(t_{\mathrm{in}})}(s)+(1-\pi_N^{(t_{\mathrm{in}})})\mathcal{N}_N^{(t_{\mathrm{in}})}(s)}\end{aligned} \tag{21}$$

$$Q_N^{(t_{\mathrm{in}})}(s)=1-Q_P^{(t_{\mathrm{in}})}(s) \tag{22}$$

进而得到关于 $\mathcal{L}(\boldsymbol{S}|\boldsymbol{\Phi},\pi_P,\pi_N)$ 的一阶近似：

$$\begin{aligned}\mathcal{L}^{(t_{\mathrm{in}})}(\boldsymbol{S}|\boldsymbol{\Phi})=\sum_{s\in\boldsymbol{S}}[&Q_P^{(t_{\mathrm{in}})}(s)\log(\mathcal{N}_P^{(t_{\mathrm{in}})}(s))+\\&Q_N^{(t_{\mathrm{in}})}(s)\log(\mathcal{N}_N^{(t_{\mathrm{in}})}(s))]\end{aligned} \tag{23}$$

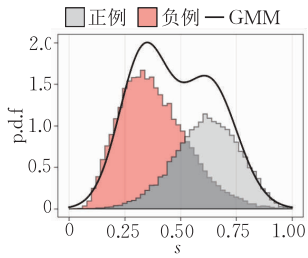


图 4　CIFAR-10 上 GMM 拟合预测分数混合分布实例图

接下来，EM 算法通过 M 步最大化 $\mathcal{L}^{(t_{\text{in}})}(\boldsymbol{S}|\boldsymbol{\Phi})$，以得到第 $t_{\text{in}}+1$ 次迭代后的参数：

$$\boldsymbol{\Phi}^{(t_{\text{in}}+1)}=\arg\max_{\boldsymbol{\Phi}}\mathcal{L}^{(t_{\text{in}})}(\boldsymbol{S}|\boldsymbol{\Phi}) \tag{24}$$

$$\pi_P^{(t_{\text{in}}+1)}=\mathop{\mathbb{E}}_{s\in\boldsymbol{S}}[Q_P^{(t_{\text{in}})}] \tag{25}$$

$$\pi_N^{(t_{\text{in}}+1)}=1-\pi_P^{(t_{\text{in}}+1)} \tag{26}$$

其中，式(24)关于 $\boldsymbol{\Phi}$ 的参数更新方程可转换为以下形式[50]：

$$\begin{aligned}
\mu_P^{(t_{\text{in}}+1)}&=\frac{1}{\pi_P^{(t_{\text{in}}+1)}}\mathop{\mathbb{E}}_{s\in\boldsymbol{S}}[sQ_P^{(t_{\text{in}})}],\\
\sigma_P^{(t_{\text{in}}+1)}&=\frac{1}{\pi_P^{(t_{\text{in}}+1)}}\mathop{\mathbb{E}}_{s\in\boldsymbol{S}}[(s-\mu_P^{(t_{\text{in}}+1)})^2Q_P^{(t_{\text{in}})}],\\
\mu_N^{(t_{\text{in}}+1)}&=\frac{1}{\pi_N^{(t_{\text{in}}+1)}}\mathop{\mathbb{E}}_{s\in\boldsymbol{S}}[sQ_N^{(t_{\text{in}})}],\\
\sigma_N^{(t_{\text{in}}+1)}&=\frac{1}{\pi_N^{(t_{\text{in}}+1)}}\mathop{\mathbb{E}}_{s\in\boldsymbol{S}}[(s-\mu_N^{(t_{\text{in}}+1)})^2Q_N^{(t_{\text{in}})}]
\end{aligned} \tag{27}$$

EM 算法求解 GMM 的伪代码如算法 1 所描述，E 步和 M 步反复迭代执行，收敛后可得类别先验 π_P 的估计值 $\hat{\pi}_P$.

算法 1. 类别先验估计.

输入：

$\boldsymbol{S}$：PU 数据的预测分数，$\boldsymbol{S}=[s_1;s_2;\cdots;s_n]$

$\boldsymbol{\Phi}^{(0)}$：双高斯成分的初始参数

$\pi_P^{(0)},\pi_N^{(0)}$：GMM 的初始混合系数

输出：

$\hat{\pi}_P$：类别先验估计值

过程：

#初始化

$t_{\text{in}}=0$

REPEAT

#E 步

根据式(21)和(22)计算 $Q_P^{(t_{\text{in}})}(s),Q_N^{(t_{\text{in}})}(s)$.

#M 步

根据式(27)计算 $\boldsymbol{\Phi}^{(t_{\text{in}}+1)}$.

根据式(25)和(26)计算 $\pi_P^{(t_{\text{in}}+1)},\pi_N^{(t_{\text{in}}+1)}$.

$t_{\text{in}}=t_{\text{in}}+1$

UNTIL(收敛)

$\hat{\pi}_P=\pi_P^{(t_{\text{in}})}$

GMM 拟合预测分数的效果如图 4 所示，其概率密度函数(Probability Density Function，p.d.f.)大致符合正类和负类预测分数各自所服从分布的叠加. 尽管 PU 学习模块和类别先验估计模块的协同工作基本达成了本文目标，即在类别先验未知条件下进行深度 PU 学习，但它无法确保其迭代过程的稳定性，$\hat{\pi}_P$ 可能随着迭代轮次增长而发生震荡，进而影响模型最终的分类性能. 为此，本文设计了迭代过程稳定模块，以平稳所提框架的迭代过程.

3.5 迭代过程稳定模块

迭代过程稳定模块基于平滑假设和聚类假设，通过平均教师[23]的架构对模型历史参数进行集成，以提升深度模型的抗扰动能力；另外结合使用温度锐化技术，加强深度模型对可信样本的预测置信度，图 4 中不同类别的钟形曲线因之更可区分，从而平稳了类别先验估计的迭代过程.

3.5.1 平均教师

平均教师[23]由深度神经网络结构相同的学生模型和教师模型组成. 在所提框架中，学生模型为 PU 学习模块中的深度神经网络，其参数为 $\boldsymbol{\Theta}$，由 PU 学习模块训练得到；教师模型的参数 $\boldsymbol{\Theta}'$ 不直接在 PU 数据上学习，而是由学生模型的历史参数经过时序上的 EMA 运算得到，即：

$$\boldsymbol{\Theta}'_{(t+1)}=\alpha\boldsymbol{\Theta}'_{(t)}+(1-\alpha)\boldsymbol{\Theta}_{(t+1)} \tag{28}$$

其中，$\boldsymbol{\Theta}_{(t+1)}$，$\boldsymbol{\Theta}'_{(t+1)}$ 表示所提框架第 $t+1$ 次迭代时学生模型和教师模型的参数；$\boldsymbol{\Theta}'_{(t)}$ 则为第 t 次迭代时模型的参数；α 是 EMA 的平滑系数，其取值在[0,1]之间.

给定样本 $\boldsymbol{x}$，教师模型的输出 $f(\boldsymbol{x};\boldsymbol{\Theta}')$ 可以看作对学生模型输出 $f(\boldsymbol{x};\boldsymbol{\Theta})$ 的轻微扰动. 根据平滑假设，若模型的输出高度可信，那么 $f(\boldsymbol{x};\boldsymbol{\Theta}')$ 和 $f(\boldsymbol{x};\boldsymbol{\Theta})$ 应当具有一致性. 因此，平均教师将教师模型的输出作为监督信息，指导学生模型的训练. 首先将样本 $\boldsymbol{x}$ 关于学生模型 $\boldsymbol{\Theta}$ 的输出置信度 c 定义为由学生模型给出的最大类条件概率，即：

$$c=\max(s,1-s) \tag{29}$$

结合基于置信度的掩码技术[24]，并将置信度阈值设为 τ，那么学生模型和教师模型的输出一致性损失可定义为

$$L_{\text{con}}=\frac{1}{n}\sum_{\boldsymbol{x}\in\boldsymbol{X}_{\text{PU}}}\mathbb{1}(c\geqslant\tau)\left\|f(\boldsymbol{x};\boldsymbol{\Theta})-f(\boldsymbol{x};\boldsymbol{\Theta}')\right\|_2^2 \tag{30}$$

其中，$\mathbb{1}(\cdot)$是指示函数，当满足条件$(\cdot)$时函数值取 1，反之取 0.

3.5.2 温度锐化

温度锐化通过调整预测分数的温度(Temperature，T)，降低预测类别条件分布的信息熵. 其理论依据是聚类假设，即分类决策面不应穿过数据密集的区域. 因此，样本 $\boldsymbol{x}$ 经过模型 $f(\cdot;\boldsymbol{\Theta})$ 的预测类别条件分布具有较低的信息熵. 给定预测分数 s，温度锐化使用锐化函数来降低上述信息熵：

$$Sharpen(s,T)=\frac{s^{1/T}}{s^{1/T}+(1-s)^{1/T}} \tag{31}$$

结合平均教师中提及的基于置信度的掩码技术，只对高置信度的预测分数进行温度锐化，那么温度锐化的优化目标可写作：

$$L_{\text{sharpen}}=\frac{1}{n}\sum_{x\in \boldsymbol{X}_{\text{PU}}}\mathbb{1}(c\geq\tau)\left\|s-Sharpen(s,T)\right\|_2^2 \quad (32)$$

3.6　优化目标

综上所述，所提框架由 PU 学习模块、类别先验估计模块和迭代过程稳定模块组成，其单次迭代步骤中的整体优化目标为

$$L=\hat{R}_{\text{nnPU}}^{\ell_{\text{sigmoid}}}+\lambda_1 L_{\text{con}}+\lambda_2 L_{\text{sharpen}} \quad (33)$$

其中，λ_1，λ_2 为超参数. 所提框架反复迭代执行深度 PU 学习和类别先验估计，直至 L 收敛或迭代次数达到预设值，其算法伪代码如算法 2 所描述.

算法 2.　迭代式的深度 PU 学习与类别先验估计.

输入：

$\boldsymbol{X}_{\text{PU}}$：输入特征矩阵，$\boldsymbol{X}_{\text{PU}}=[\boldsymbol{x}_1;\boldsymbol{x}_2;\cdots;\boldsymbol{x}_n]$.

$\boldsymbol{X}_L$：被标注正例的输入特征矩阵，$\boldsymbol{X}_L=[\boldsymbol{x}_1;\boldsymbol{x}_2;\cdots;\boldsymbol{x}_{n_L}]$.

α：式(28)中的平滑系数.

τ：式(30)和(32)中的置信度阈值.

λ_1，λ_2：优化目标(33)中一致性损失和温度锐化的超参数.

$f(\cdot;\cdot)$：所选深度神经网络结构对应的函数映射.

$\boldsymbol{\Theta}_{(0)}$，$\boldsymbol{\Theta}'_{(0)}$：学生模型和教师模型的初始参数.

输出：

$\hat{\pi}_P$：正类先验估计值.

$\boldsymbol{\Theta}$，$\boldsymbol{\Theta}'$：算法结束时学生模型和教师模型的参数.

过程：

$t=0$

REPEAT

$\boldsymbol{S},\boldsymbol{S}'=\text{sigmoid}[f(\boldsymbol{X}_{\text{PU}};\boldsymbol{\Theta}_{(t)})],\text{sigmoid}[f(\boldsymbol{X}_{\text{PU}};\boldsymbol{\Theta}'_{(t)})]$

对 $\boldsymbol{S}'$ 运用算法 1，得到 $\hat{\pi}_P$.

根据式(16)计算 $\hat{R}_{\text{nnPU}}^{\ell_{\text{sigmoid}}}$.

根据式(30)计算 L_{con}.

根据式(32)计算 L_{sharpen}.

$L=\hat{R}_{\text{nnPU}}^{\ell_{\text{sigmoid}}}+\lambda_1 L_{\text{con}}+\lambda_2 L_{\text{sharpen}}$

对 L 运用梯度反向传播算法，得到 $\boldsymbol{\Theta}_{(t+1)}$.

$\boldsymbol{\Theta}'_{(t+1)}=\alpha\boldsymbol{\Theta}'_{(t)}+(1-\alpha)\boldsymbol{\Theta}_{(t+1)}$

$t=t+1$

UNTIL(收敛或迭代次数达到预设值)

$\boldsymbol{\Theta},\boldsymbol{\Theta}'=\boldsymbol{\Theta}_{(t)},\boldsymbol{\Theta}'_{(t)}$

3.7　设计选项讨论

在对所提框架进行设计的过程中，存在多个设计选项，分别是：(1) 在算法流程方面，是迭代式地进行深度 PU 学习和类别先验估计，抑或是基于单次类别先验估计的深度 PU 学习；(2) 类别先验的估计方式，是使用 GMM 拟合预测分数的混合分布，从而间接估计类别先验，抑或直接统计伪标签的正类频率；(3) 迭代过程稳定模块的必要性. 接下来对上述设计选项进行探讨.

3.7.1　迭代式 vs 单次估计

为解决类别先验未知问题，以往工作[17-20]通常首先进行类别先验估计，进而基于类别先验估计值进行 PU 学习. 然而，单次估计的随机性较强. 况且训练初期深度模型的学习不够充分，基于预测分数进行单次估计的质量更加难以得到保障. 此外，在深度神经网络过拟合前，深度模型的分类性能的提升一般意味着其提取的高层级特征更具代表性. 上述增益无法为单次估计所利用. 因此，所提框架选用迭代式地进行深度 PU 学习和类别先验估计的算法流程，即随着深度模型的训练，预测分数愈发可靠，从中估计的类别先验将更加准确；更准确的估计值进一步促使模型分类性能的提升. 深度 PU 学习和类别先验估计二者相互促进.

3.7.2　GMM vs 伪标签频率

关于如何利用预测分数估计类别先验，其最为简单直接的做法是统计伪标签中的正类频率：

$$\hat{\pi}_P=\frac{1}{n}\sum_{x\in \boldsymbol{X}_{\text{PU}}}\mathbb{1}(s\geq 0.5) \quad (34)$$

尽管基于伪标签的估计方法符合常理，且易于实现，但是它对预测分数的分布要求较为严苛. 预测分数不仅需要具有较好的类别可分性，且其分类面需位于 $s=0.5$ 附近. 然而，上述要求往往难以满足. 考虑当预测分数的分值整体偏大或偏小时，伪标签的正类频率将大幅高于或低于真实的类别先验. 图 5 展示了

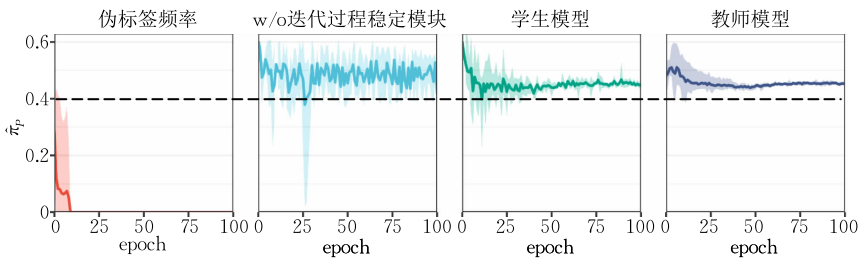


图 5　CIFAR-10 上不同方法(伪标签、GMM、本文方法)的 $\hat{\pi}_P$ 随训练 epoch 数的变化曲线(其中，彩色实线为重复 5 次实验的均值；阴影部分由 5 次实验的最小值和最大值决定. 黑色虚线表示 $\pi_P=0.4$)

CIFAR-10 上基于伪标签频率预测的 $\hat{\pi}_P$ 随着训练轮次的变化趋势. 可以发现,训练初期偏低的 $\hat{\pi}_P$ 导致了模型对正类的拟合不足,于是其输出的预测分数整体偏低,进而造成了更低的 $\hat{\pi}_P$,由此形成了确认偏差(Confirmation Bias)[51],即模型过拟合于训练初期不准确的预测,模型的预测分数和 $\hat{\pi}_P$ 逐渐向 0 靠近. 此外,消融实验结果分析(见 4.5 节)进一步说明了基于伪标签的 $\hat{\pi}_P$ 不可行. 相较于伪标签频率法,基于 GMM 的类别先验估计方法通过将预测分数的分布建模为双高斯成分的混合分布,于是无论预测分数整体偏大或偏小,只要其分布大致符合两条钟形曲线加权叠加的特征,就能较为准确地估计类别先验,从而应对预测分数分布整体偏移问题. 图 5 和表 4 表明,迭代式的 GMM 基本达成本文目标,并取得了良好的分类性能.

3.7.3 迭代过程稳定模块的必要性

实验发现,尽管迭代式的 GMM 基本能够估计类别先验,并且取得了良好的分类性能,但是其类别先验估计过程不够稳定. 如图 5 左二所示,$\hat{\pi}_P$ 随着训练轮次的变化而震荡,5 次实验的差异较大. 为此,本文特别设计了迭代过程稳定模块,通过引入平均教师[23]和温度锐化[24]以平稳迭代过程. 如图 5 右一、右二所示,迭代过程稳定模块的效果明显,$\hat{\pi}_P$ 的震荡现象得到有效缓解,5 次实验的标准差明显缩小,这说明了迭代过程稳定模块使得所提框架更加稳定可靠. 另外,平均教师中学生模型与教师模型的输出一致性为模型训练提供了训练标签以外的自监督信息,模型最终的分类性能因之得到了进一步提升(见表 4).

4 实 验

4.1 数据集与主干网络

本文的实验设置整体上与论文[6]保持一致,在三个基准数据集(手写数字识别数据集 MNIST[52]、服饰图片分类数据集 F-MNIST[53]、普适物体识别数据集 CIFAR-10[54])和一个实际应用数据集(阿尔茨海默病神经成像数据集 Alzheimer①)上进行了实验. 各个数据集的基本信息和训练时使用的主干网络如表 1 所示,接下来对各个数据集进行详细介绍.

表 1 数据集与主干网络概述

数据集	训练集大小	测试集大小	输入尺寸	π_P	标签个数	正类/负类	主干网络
MNIST	60000	10000	28×28	0.49	1000	奇数/偶数	多层感知机(6 层)
MNIST^2	60000	10000	28×28	0.48	1000	0～4/5～9	多层感知机(6 层)
F-MNIST	60000	10000	28×28	0.40	500	上装/非上装	多层感知机(6 层)
F-MNIST^2	60000	10000	28×28	0.30	500	1,4,7/其它	多层感知机(6 层)
CIFAR-10	50000	10000	3×32×32	0.40	1000	交通工具/动物	卷积神经网络(13 层)
CIFAR-10^2	10000	2000	3×32×32	0.50	2000	猫/狗	卷积神经网络(13 层)
Alzheimer	5121	1279	3×224×224	0.50	769	痴呆/健康	Resnet-18

4.1.1 MNIST

MNIST[52]共有 70000 个手写阿拉伯数字图像,其中训练集和测试集分别包含 60000 和 10000 张图片. 每个图片的输入尺寸为 28×28. 将奇数 1、3、5、7、9 当作正类,偶数 0、2、4、6、8 作为负类. 经统计,类别先验 π_P 为 0.49. 同时,为符合 PU 学习的实验设定,本文随机选择了训练集中的 1000 个正样本,保留其标签,并将剩余未被选择的样本作为无标签数据. MNIST 数据集使用的主干网络是一个 6 层的多层感知机,其隐层由 4 个无偏置的线性层构成,每一个线性层都紧接着批归一化[55]操作和 ReLU[56]激活函数. MNIST^2 将小于 5 的数当作正类,反之则为负类. 经统计,类别先验 π_P 为 0.48.

4.1.2 F-MNIST

F-MNIST[53]数据集是服饰图片分类任务下的 MNIST 数据集,输入图片尺寸,训练、测试样本数及类别数的设定同经典的 MNIST 相同. F-MNIST 将上装类:0(T 恤)、2(套衫)、4(外套)、6(汗衫)视作正类,将非上装类:1(裤子)、3(裙子)、5(凉鞋)、7(运动鞋)、8(包)、9(踝靴)视作负类,则 π_P 为 0.40. 为符合 PU 学习的实验设定,本文随机选择了训练集中的 500 个正样本,保留其标签,并将剩余未被选择的样本作为无标签数据. F-MNIST 使用的主干网络与 MNIST 相同. 类似地,F-MNIST^2 将 1、4、7 视作正类,π_P 为 0.30.

4.1.3 CIFAR-10

CIFAR10[54]是一个常用于目标识别的图像基

① Alzheimer's Dataset. https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images

准数据集，由 60000 个 3×32×32 彩色图像组成，包含 10 类对象，其中，每个类有 6000 个图像. CIFAR-10 将交通工具类：0(飞机)、1(汽车)、8(船)和 9(卡车)视作正类，将动物类：2(鸟类)、3(猫)、4(鹿)、5(狗)、6(蛙)、7(马)视作负类. 可知，π_P 为 0.40. 为符合 PU 实验设定，本文随机选择了训练集中的 1000 个正样本，保留其标签，并将剩余未被选择的样本作为无标签数据. CIFAR10 采用的主干网络为全卷积深度神经网络[57]. CIFAR10[2] 则将 3(猫)视作正类，5(狗)视作负类，π_P 为 0.50，标签个数为 2000.

4.1.4　Alzheimer

Alzheimer 数据集是关于阿尔兹海默症的多分类数据集，总共包含了 4 类共 5760 张脑部磁共振成像(Magnetic Resonance Imaging，MRI)，训练集和测试集的大小分别为 5121 和 639. 如图 6 所示，AD 有 4 个病程发展阶段，包括非痴呆(Non-Demented)、轻微痴呆(VeryMild Demented)、轻度痴呆(Mild Demented)和中度痴呆(Moderate Demented). 其中，非痴呆与轻微痴呆的 MRI 相近. 值得注意的是，正常衰老与 AD 之间存在一种过度状态，称作轻度认知功能障碍(Mild Cognitive Impairment，MCI). 该状态下的 MRI 与常人无异，然而部分 MCI 被临床诊断为 AD[6]. 因此，训练集中的中度痴呆和轻度痴呆样本均可被临床诊断，视作正样本. 而轻微痴呆的 MRI 与健康人相差无几，共同组成了无标签数据. 本文采用 ImageNet 预训练后的残差网络 Resnet-18[58] 作为 Alzheimer 数据集的主干网络，其输入图像尺寸被统一裁剪为 3×224×224.

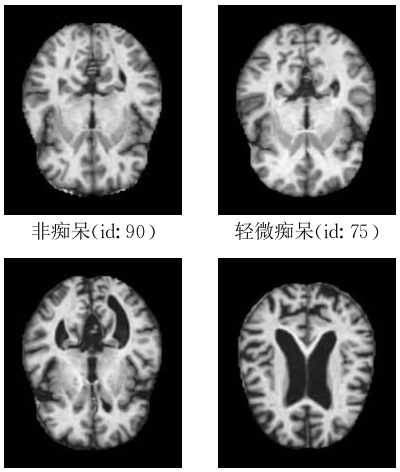


图 6　AD 的 4 个病程发展阶段对应的 MRI(非痴呆的 MRI(左上)与轻微痴呆的 MRI(右上)十分相似)

4.2　评价指标

二分类问题可根据样本真实标签的正负以及模型预测的正确与否，划分为 4 种基本情况，分别是样本被：(1) 正确地预测成正类(TruePositive，TP)；(2) 正确地预测成负类(TrueNegative，TN)；(3) 错误地预测成正类(False Positive，FP)和(4) 错误地预测成负类(False Negative，FN). 基于上述 4 种情况定义以下 4 类评价指标，分别是准确率(Accuracy，Acc)、精确率(Precision)、召回率(Recall)和 $F1$-score(f_1).

(1) 准确率，反映算法预测的总体正确程度，即全部预测中，预测正确的数目占比.

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{35}$$

(2) 精确率，即预测结果为正类中，预测正确的数目占比.

$$precision=\frac{TP}{TP+FP} \tag{36}$$

(3) 召回率，即真实标签为正类中，预测正确的数目占比.

$$recall=\frac{TP}{TP+FN} \tag{37}$$

(4) $F1$-score，精确率和召回率的调和均值，用于综合评估精确率和召回率，是 PU 学习的重要评价指标.

$$f_1=\frac{2\times precision\times recall}{precision+recall} \tag{38}$$

4.3　实现细节

所提方法基于深度学习框架 Pytorch 实现. 所有实验均在GeForce RTX 3090上进行. 为确保GMM的初次拟合结果，首先对所提框架的学生模型进行 5 个轮次的热身训练(warm up)，以使得初始预测分数的类别条件分布基本可分. 式(33)中超参数 λ_1 在 MNIST、F-MNIST、CIFAR-10 和 Alzheimer 的取值分别为 0.05、0.05、0.3 和 0.3；而超参数 λ_2 采取余弦退火的策略，即随着训练轮次的增加，其值逐渐从 0 增至 1. 式(28)中的平滑系数 α 设为 0.999，式(32)中的温度 T 均设为−10，置信度阈值 τ 均设为 0.8. 基准数据集的初始步长均设置为 5e-4，Alzheimer 的初始步长则为 1e-4；在训练过程中使用 Adam 优化器[59] 和余弦退火策略器 Cosine-AnnealingLR[60] 对步长进行调整；其中，Adam 的权重衰减系数为 5e-3. 基准数据集的训练批大小(Batch Size)均设置为 256，而 Alzheimer 的训练批大小为 128. 为了保证实验的准确性和可复现性，程序的每次运行都会指定一个整数作为 NumPy、Pytorch、random 标准库、CUDA 等模块的随机种子. 所有实验在随机种子 0～4 下重

复 5 次进行,并记录教师模型或学生模型在验证集上准确率最高时对应各项指标的均值和标准差.

4.4 对比实验

4.4.1 对比方法

(1) supervised,即有监督学习方法,将所有无标签样本当作负例.由于 PU 学习中无标签样本数目往往远大于被标注的正例数目,即 $n_U \gg n_L$,深度模型将过拟合于负例.因此,对无标签数据进行下采样,即从集合 $\boldsymbol{X}-\boldsymbol{X}_L$ 中无放回地随机抽取 n_L 个样本当作负例,与已有的 n_L 个正例构成二分类监督学习训练集.显然,上述操作可能会将无标签数据中的正例当作负例,从而引入了标签噪声(Label Noise).

(2) uPE 是 PE[17] 与 uPU[15] 的结合.PE[17] 通过实现正类条件密度与混合密度之间的部分匹配估计类别先验.其中,Pearson 散度用于衡量上述密度间的差异.

(3) nnPE 是 PE 与 nnPU[16] 的结合.

(4) uKM1 是 KM1[18] 与 uPU 的结合.KM1 和 KM2[18] 基于再生核希尔伯特核空间.在该空间下,正类条件分布是混合分布核的重新加权.通过最小化正类条件分布的核嵌入距离,找到最佳的权重和类别先验估计值.

(5) nnKM1 是 KM1 与 nnPU 的结合.

(6) uKM2 是 KM2 与 uPU 的结合.

(7) nnKM2 是 KM2 与 nnPU 的结合.

(8) uTIcE 是 TIcE[19] 与 uPU 的结合.TIcE[19] 指出任一数据子域的正类标签频率是真实正类标签频率的下界,并通过自顶向下的决策树,搜索使得正类标签频率最大的数据子域,从而间接逼近真实的正类标签频率,进而估计类别先验.

(9) nnTIcE 是 TIcE 与 nnPU 的结合.

(10) uCDMM 是 CDMM[20] 与 uPU 的结合.CDMM[20] 结合循环坐标下降和 MM 优化框架,优化逻辑回归的参数似然函数,估计正类标签频度,进而估计类别先验.

(11) nnCDMM 是 TIcE 与 nnPU 的结合.

(12) Ours,即本文所提框架.

4.4.2 实验结果及分析

表 2 展示了不同算法在各个数据集上的对比实验结果.总体来看,所提方法在多数评价指标上取得了最优性能.

具体而言,Ours 在 MNIST 上取得了最优的准确率、召回率和 $F1$-score,相较于次优结果分别平均提升了 8.33%、3.54% 和 7.84%;在 MNIST^2 上取得了最优的准确率和 $F1$-score,较次优结果分别平均提升了 0.64% 和 0.42%.在 FMNIST 上取得了最优的准确率、召回率和 $F1$-score,较次优结果分别平均提升了 1.77%、2.79% 和 2.27%;在 FMNIST^2 上取得了最优的准确率、召回率和 $F1$-score,较次优结果平均分别提升 0.52%、2.21% 和 1.16%.在 CIFAR10 上取得了最好的准确率、召回率和 $F1$-score,较次优结果分别平均提升了 3.40%、4.40% 和 4.55%;在 CIFAR-10^2 上取得了最优的准确率和 $F1$-score,较次优结果分别平均提升了 7.51% 和 3.39%.由于 Alzheimer 的主干网络是 ImageNet 预训练后的模型,这减轻了类别先验估计值不准确的负面影响.Ours 的优势尽管有所削减,但仍在主要评价指标准确率和 $F1$-score 上取得了最优性能,较次优结果分别平均提升了 1.25% 和 0.86%.此外,Alzheimer 上的实验结果表明所提框架有助于 AD 的早期诊断,具有实用价值.Ours 与 supervised 在 Alzheimer 上针对病患各个病程的检出率如图 7 所示,鉴于 Alzheimer 训练集中所有的中度痴呆或轻度痴呆均为有标签的正例,而轻微痴呆缺失标签,Ours 对轻微痴呆的检出率相较 supervised 具有 40.40% 的显著提升,对中度痴呆也有近 17.87% 的提升.此外,由于测试集中属于轻度痴呆的样本数仅有 12 个,Ours 和 supervised 都对轻度痴呆达到了 100% 的检出率.考虑到轻微痴呆和正常人的 MRI 相近,现实条件下轻微痴呆难以被专家诊断,Ours 相较于 supervised 在轻微痴呆检出率上的显著提升,说明了所提方法有助于 AD 病患尽早诊断和治疗,进而利于保障患者的生命健康.

Ours 的先进性很大程度上可以归因于其类别先验估计准确,且稳定可靠.以 MNIST、FMNIST、CIFAR-10 和 Alzheimer 为例,不同类别先验估计算法在各个数据集上的对比实验结果如图 8 所示,观察可知,Ours 的类别先验估计均值在 CIFAR-10 和 Alzheimer 上最接近真实的类别先验,其在 MNIST、F-MNIST 上的 $\hat{\pi}_P$ 与 π_P 也仅平均相差 0.01 和 0.03.此外,Ours 随机 5 次实验的 $\hat{\pi}_P$ 标准差较小,最高不超过 0.02.由此可见,Ours 兼具较小的类别先验估计误差和标准差,因而具有更好且更稳定的分类性能.相反的是,其它类别先验估计算法无法兼顾类别先验估计的准确性和稳定性.例如 KM1 在 MNIST 上的估计均值最接近真实值,但其 5 次实验的差异

表 2　对比实验结果(单元格记录了均值(标准差),最优性能以加粗黑体标出,次优性能以倾斜黑体标出.根据配对 t 检验的结果,分别用√(×)表示 Ours 明显优于(更差)的相应方法,置信水平为 95%)　　(单位:%)

数据集	评价指标	supervised	uPE	nnPE	uKM1	nnKM1	uKM2	nnKM2	uTIcE	nnTIcE	uCDMM	nnCDM-M	Ours
MNIST	*ACC*	85.89√ (0.43)	86.07√ (0.75)	***86.33*** √ (***0.35***)	81.88√ (8.48)	85.44√ (1.59)	64.27√ (18.54)	64.33√ (18.61)	77.74 (14.37)	79.92 (11.69)	66.89√ (12.19)	57.74√ (15.62)	**94.66** (**0.37**)
	precision	89.35√ (1.69)	85.83√ (1.15)	83.48√ (1.22)	86.50√ (2.05)	83.70√ (3.02)	95.09 (6.78)	94.97 (7.09)	94.74 (1.68)	94.29 (1.28)	84.87 (10.78)	**97.13** (**6.42**)	***95.41*** (***1.05***)
	recall	81.09√ (2.47)	85.93√ (1.71)	***90.14*** √ (***2.27***)	74.73√ (19.87)	87.77√ (4.97)	31.97√ (43.84)	32.28√ (44.22)	58.53 (31.66)	63.24 (25.84)	43.39√ (30.89)	17.06√ (38.09)	**93.68** (**0.94**)
	F-score	84.98√ (0.70)	85.87√ (0.81)	***86.66*** √ (***0.52***)	78.93√ (13.81)	85.56√ (1.81)	33.43√ (45.79)	33.55√ (45.94)	67.18 (29.97)	73.06 (19.33)	50.30√ (31.48)	17.12√ (38.18)	**94.53** (**0.37**)
MNIST^2	*ACC*	81.84√ (0.40)	91.22√ (0.33)	***92.70*** √ (***0.32***)	74.11 (23.29)	75.23 (24.30)	87.80√ (2.11)	89.67√ (1.88)	62.83√ (15.50)	66.09 (17.91)	68.55√ (11.87)	71.23√ (12.51)	**93.34** (**0.75**)
	precision	90.91√ (1.04)	92.89√ (1.94)	91.92√ (1.12)	***96.30*** (***3.46***)	95.51 (4.21)	92.35√ (2.12)	93.31√ (1.44)	96.19 (4.45)	**97.04** (**2.82**)	94.69 (2.17)	94.38 (2.50)	95.25 (1.51)
	recall	71.86√ (1.30)	89.87 (2.35)	**94.09**× (**1.04**)	53.14 (48.55)	56.38 (51.48)	83.26√ (5.41)	86.15√ (4.79)	29.38√ (31.62)	36.01√ (37.10)	40.80√ (23.82)	47.20√ (26.81)	***91.64*** (***1.03***)
	F-score	80.26√ (0.57)	91.31√ (0.41)	***92.98*** √ (***0.28***)	54.66 (49.90)	55.93 (51.06)	87.45√ (2.58)	89.50√ (2.24)	37.36√ (37.32)	42.87√ (41.12)	53.82√ (24.32)	58.79√ (25.89)	**93.40** (**0.72**)
F-MNIST	*ACC*	90.87√ (5.09)	***93.39*** √ (***0.11***)	93.32√ (0.15)	86.87 (15.02)	87.39 (13.72)	86.43 (14.78)	86.48 (14.80)	62.53√ (5.48)	66.42√ (8.65)	88.33 (11.52)	86.58 (14.87)	**95.16** (**0.30**)
	precision	91.06√ (0.39)	90.99√ (0.57)	90.29√ (1.21)	93.72 (3.68)	93.18 (3.26)	92.46 (4.38)	92.80 (4.21)	**98.98**× (**2.28**)	***95.50*** (***6.47***)	92.37 (2.43)	92.51 (4.39)	92.09 (0.52)
	recall	85.58√ (1.31)	92.65√ (0.60)	***93.38*** √ (***1.83***)	73.47 (41.10)	75.10 (37.98)	73.78 (41.28)	73.49 (32.28)	6.67√ (14.49)	17.59√ (23.38)	77.91 (32.31)	74.13 (41.44)	**96.17** (**0.90**)
	F-score	88.23√ (7.30)	***91.81*** √ (***0.13***)	91.79√ (0.30)	73.58 (41.13)	76.23 (35.07)	73.10 (40.87)	73.13 (33.55)	10.00√ (21.53)	24.64√ (29.12)	80.19 (26.17)	73.30 (40.98)	**94.08** (**0.38**)
F-MNIST^2	*ACC*	86.55√ (1.01)	89.59√ (0.40)	89.66√ (0.47)	85.77√ (2.59)	88.70√ (1.43)	86.55√ (2.36)	88.98√ (0.70)	73.62√ (4.51)	73.96√ (6.21)	89.28√ (0.36)	***89.95*** √ (***0.22***)	**90.47** (**0.34**)
	precision	79.64√ (2.66)	84.28 (2.15)	83.11 (2.23)	86.41× (3.03)	86.66× (1.79)	83.72 (2.34)	86.34× (1.93)	***97.48***× (***2.74***)	**97.66**× (**3.22**)	85.80 (4.35)	84.07 (1.27)	83.87 (1.81)
	recall	74.24√ (2.23)	80.43√ (3.20)	***82.36*** √ (***1.72***)	62.81√ (12.35)	73.67√ (4.75)	68.49√ (9.18)	75.31√ (4.58)	12.50√ (15.48)	14.03√ (21.96)	77.51√ (4.73)	82.10 (2.24)	**84.57** (**1.28**)
	F-score	76.81√ (1.65)	82.24√ (0.97)	82.70√ (0.49)	72.06√ (7.43)	79.58√ (3.09)	75.08√ (5.77)	80.34√ (2.02)	19.48√ (23.54)	19.79√ (28.83)	81.25√ (0.83)	***83.04*** √ (***0.65***)	**84.20** (**0.30**)
CIFAR-10	*ACC*	84.70√ (0.52)	82.86√ (2.87)	83.78√ (0.82)	75.40 (19.84)	84.92√ (0.95)	84.31√ (4.92)	***86.58*** √ (***2.57***)	81.84 (12.37)	82.57 (12.64)	74.81 (13.69)	79.27 (10.86)	**89.98** (**0.83**)
	precision	83.64√ (2.28)	79.42√ (6.00)	80.45√ (2.26)	73.01 (18.71)	80.60√ (2.64)	82.26√ (3.44)	84.04√ (2.26)	88.82 (6.41)	**90.54** (**5.32**)	***90.33*** (***8.99***)	87.29 (7.27)	87.36 (0.82)
	recall	76.98√ (4.21)	77.97√ (5.07)	78.66√ (1.33)	***83.23*** (***9.51***)	82.24√ (2.46)	77.31 (12.96)	82.01 (6.09)	65.19 (36.79)	65.18 (36.50)	45.78 (42.03)	59.44 (33.44)	**87.63** (**1.79**)
	F-score	80.06√ (1.42)	78.47√ (3.15)	79.51√ (0.69)	75.49 (10.35)	81.35√ (0.98)	79.31 (8.32)	***82.94*** √ (***3.71***)	66.90 (37.52)	67.74 (37.88)	47.92 (43.86)	63.07 (35.33)	**87.49** (**1.13**)
CIFAR-10^2	*ACC*	62.47√ (1.01)	63.87√ (8.50)	55.71√ (3.82)	65.74 (9.05)	66.46 (10.66)	53.28√ (7.33)	50.00√ (0.00)	65.22√ (3.11)	53.86√ (3.85)	64.10 (8.27)	***66.30*** (***9.35***)	**73.81** (**1.50**)
	precision	61.46√ (1.31)	60.41√ (7.13)	53.29√ (2.29)	78.19 (12.35)	76.91 (13.15)	***95.11***× (***10.94***)	**100.00**× (**0.00**)	60.94√ (3.55)	52.21√ (2.27)	76.47 (13.29)	76.55 (13.29)	71.96 (0.97)
	recall	67.36√ (7.35)	89.06× (7.36)	***96.20***× (***3.23***)	50.34 (28.50)	54.72 (33.66)	9.70√ (21.69)	0.00√ (0.00)	86.90 (5.26)	**97.16**× (**3.28**)	48.20 (27.48)	55.80 (31.58)	78.00 (3.04)
	F-score	64.07√ (3.04)	71.43√ (2.96)	68.52√ (1.25)	53.93 (30.29)	55.30 (32.21)	11.81√ (26.42)	0.00√ (0.00)	***71.45*** √ (***0.89***)	67.85√ (1.12)	51.94 (29.29)	56.10 (31.47)	**74.84** (**1.79**)
Alzheimer	*ACC*	61.47√ (2.22)	69.74√ (1.68)	70.74√ (1.57)	70.45√ (0.40)	***71.95*** √ (***0.39***)	69.85√ (0.98)	71.62√ (0.74)	70.27√ (0.70)	71.62√ (0.68)	63.77√ (3.49)	65.86√ (1.55)	**73.20** (**0.55**)
	precision	78.04√ (14.08)	72.43 (2.13)	70.05√ (1.77)	71.02 (2.22)	70.56√ (1.67)	71.96 (2.14)	70.30√ (0.80)	71.58 (2.22)	70.44√ (0.93)	***75.66*** (***6.65***)	**75.86** (**5.76**)	72.25 (0.73)
	recall	38.09√ (17.97)	63.98√ (6.46)	72.90 (9.07)	69.48 (6.17)	**75.52** (**4.47**)	65.41√ (6.56)	74.84 (2.21)	67.61√ (5.66)	74.49 (2.80)	43.00√ (15.97)	47.76√ (8.82)	***75.30*** (***2.45***)
	F-score	47.57√ (12.96)	67.73√ (3.41)	71.13 (3.67)	70.04√ (2.15)	***72.86*** (***1.29***)	68.31√ (2.76)	72.48√ (1.00)	69.36√ (2.11)	72.38 (1.10)	52.68√ (12.45)	57.91√ (5.30)	**73.72** (**0.95**)

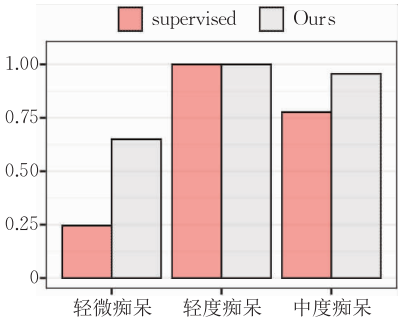


图 7 Ours 与 supervised 在 Alzheimer 上患者各个病程(轻微痴呆、轻度痴呆和中度痴呆)的检出率

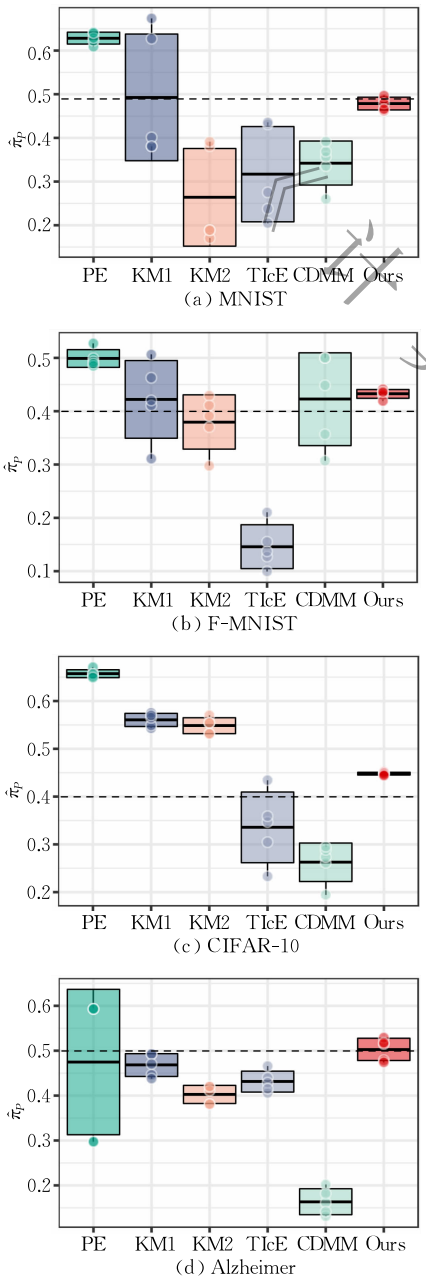


图 8 不同类别先验估计方法在各个数据集的均值-标准差箱型图(黑色虚线表示各个数据集的 π_P)

过大,因此 uKM1 和 nnKM1 的平均分类性能欠佳,且具有较大的标准差;PE 在 MNIST 上的类别先验估计最稳定,但其 $\hat{\pi}_P$ 过于偏离 π_P,因此 uPE 和 nnPE 的平均分类性能欠佳.

4.4.3 已知 π_P 的对比实验结果与分析

本文致力于解决 PU 学习中深度场景下类别先验未知问题.实验发现,即使所提框架不利用 π_P,也能达到与已知 π_P 的 PU 算法相媲美的分类性能,甚至在部分结果上略有超越.表 3 展示了 MNIST、FMNIST、CIFAR-10 和 Alzheimer 数据集上 Ours 与基于真实类别先验 PU 算法的对比实验结果.其中,uPU 是 PU 学习的无偏风险估计器;nnPU 是 uPU 针对深度模型的扩展;Self-PU 在 nnPU 的基础上引入自步学习训练策略和深度模型的自校准与自蒸馏,代表了 PU 学习的最先进水平.Ours 整体性能最佳,在 MNIST 上的准确率、精确率和 $F1$-score 上分别平均超出次优结果 0.80%、2.17%和

表 3 Ours 与已知 π_P 的 *PU* 算法对比实验结果(根据配对 t 检验的结果,分别用√(×)表示 Ours 明显优于(更差)的相应方法,置信水平为 95%) (单位:%)

数据集	评价指标	uPU	nnPU	Self-PU	Ours
MNIST	*ACC*	91.57√ (1.27)	93.26√ (0.38)	***93.86***√ (***0.30***)	**94.66** (**0.37**)
	precision	92.50√ (1.25)	***93.24***√ (***0.84***)	93.12√ (0.89)	**95.41** (**1.05**)
	recall	90.24√ (2.90)	93.08√ (1.28)	**94.54**× (**0.99**)	***93.68*** (***0.94***)
	F-score	91.33 (1.44)	93.15 (0.42)	***93.82*** (***0.31***)	**94.53** (**0.37**)
F-MNIST	*ACC*	93.66√ (0.31)	94.39√ (0.42)	***94.75***√ (***0.25***)	**95.16** (**0.30**)
	precision	***92.78*** (***1.91***)	**92.88** (**0.94**)	91.73√ (0.80)	92.09 (0.52)
	recall	91.33√ (2.06)	93.13√ (0.84)	***95.50***√ (***0.61***)	**96.17** (**0.90**)
	F-score	92.02√ (0.40)	93.00√ (0.52)	***93.57***√ (***0.28***)	**94.08** (**0.38**)
CIFAR-10	*ACC*	88.56√ (0.52)	89.07√ (0.38)	***89.28***√ (***0.72***)	**89.98** (**0.83**)
	precision	***87.35*** (***2.57***)	86.35 (1.68)	86.16√ (0.78)	**87.36** (**0.82**)
	recall	83.66√ (2.98)	86.39 (2.06)	***87.21*** (***2.35***)	**87.63** (**1.79**)
	F-score	85.39√ (0.72)	86.34√ (0.52)	***86.67***√ (***1.06***)	**87.49** (**1.13**)
Alzheimer	*ACC*	71.09√ (0.70)	72.12√ (0.46)	***72.59***√ (***0.64***)	**73.20** (**0.55**)
	precision	***71.55*** (***2.55***)	69.05√ (1.79)	69.83√ (1.03)	**72.25** (**0.73**)
	recall	70.49 (6.10)	**80.44** (**4.18**)	***79.56*** (***3.36***)	75.30 (2.45)
	F-score	70.81√ (1.93)	***74.22*** (***0.79***)	**74.34** (**1.13**)	73.72 (0.95)

0.71%；在 F-MNIST 上的准确率、召回率和 $F1$-score 超出次优结果 0.41%、0.67% 和 0.51%；在 CIFAR-10 上的所有指标超出次优结果 0.60%、0.01%、0.42% 和 0.82%；在 Alzheimer 上的准确率和精确率超出次优结果 0.61% 和 0.70%. 接下来的消融实验结果进一步揭示了所提框架基于类别先验估计，但优于基于真实类别先验 PU 算法的原因.

4.5　消融实验分析

为评估所提框架中各个设计的有效性与合理性，以及探究对比实验结果中 Ours 性能优异的原因，分别对 GMM、平均教师（L_{con}）和温度锐化（L_{sharpen}）在基准数据集上进行了消融实验，其结果如表 4 所示. 其中，无 GMM 的类别先验估计为迭代式地统计伪标签的类别频率. 总体而言，GMM 用于估计类别先验，为深度 PU 学习奠定了基础；平均教师大幅提升了分类性能，温度锐化则进一步巩固并提升了分类性能；Ours 有效结合了三者的优势，性能最佳. 下面将结合实验结果对各个设计的作用进行分析.

(1) GMM. 如表 4 所示，无 GMM 的方法在各个基准数据集上的评价指标均显著落后于 GMM（序号 0 vs 1、2 vs 4、3 vs 5、6 vs 8）. 以准确率为例，1 比 0 在 MNIST、F-MNIST 和 CIFAR-10 上分别平均提升了 1.76%、3.85% 和 22.72%；4 比 2 在各个基准数据集上分别提升了 3.17%、0.08% 和 16.82%；5 比 3 分别提升了 1.58%、3.46% 和 20.81%；8 比 6 分别提升了 3.42%、1.18% 和 17.15%. 值得一提的是，1 仅仅基于 GMM，便取得了接近基于真实类别先验的 PU 算法性能. 另外，无 GMM 的标准差远高于 GMM 的标准差. 以准确率为例，MNIST、F-MNIST 和 CIFAR-10 上无 GMM 的标准差最大可分别达到 6.29%、7.15% 和 12.06%，而 GMM 的标准差最高不超过 2.12%. 无 GMM 部分标准差较大的原因在于其类比先验估计过程出现了如图 5 左二所示的情况，即训练初期偏低的 $\hat{\pi}_P$ 导致了模型对正类的拟合不足，于是其输出的预测分数整体偏低，进而造成了更低的 $\hat{\pi}_P$，最终导致算法失效. 由此可见，GMM 的表现更加稳固可靠. 综上所述，GMM 能够更好地估计类别先验，为所提框架的分类性能带来了大幅提升.

表 4　基准数据集上的消融实验结果（✓表示包含对应设计）　　（单位：%）

数据集	序号	GMM	平均教师	伪标签	温度锐化	acc	precision	recall	f-score
MNIST	0					91.84(1.22)	94.02(1.53)	89.17(3.38)	91.48(1.42)
	1	✓				93.60(0.53)	92.98(0.73)	**94.13(1.35)**	93.54(0.58)
	2		✓			90.83(6.05)	94.51(0.46)	86.37(12.78)	89.84(7.85)
	3				✓	92.05(1.16)	94.18(1.53)	89.44(3.26)	91.71(1.38)
	4	✓	✓			94.00(0.46)	94.13(0.42)	93.67(1.26)	93.89(0.51)
	5	✓			✓	93.63(0.48)	94.29(0.71)	92.68(0.72)	93.48(0.49)
	6		✓		✓	91.24(6.29)	94.76(0.75)	86.99(13.18)	90.27(8.11)
	7	✓	✓	✓		**94.72(0.15)**	***95.31(1.05)***	***93.92(1.12)***	**94.60(0.16)**
	8(Ours)	✓	✓		✓	***94.66(0.37)***	**95.41(1.05)**	93.68(0.94)	***94.53(0.37)***
F-MNIST	0					90.86(6.90)	**94.36(1.49)**	82.30(19.20)	86.68(12.52)
	1	✓				94.71(0.35)	91.61(1.43)	95.57(1.71)	93.53(0.43)
	2		✓			94.89(0.24)	92.22(0.79)	95.26(1.07)	93.71(0.31)
	3				✓	91.42(7.15)	93.13(2.22)	85.22(20.65)	87.56(12.95)
	4	✓	✓			94.97(0.33)	92.17(0.90)	95.58(1.21)	93.83(0.42)
	5	✓			✓	94.88(0.36)	92.08(0.44)	95.41(0.82)	93.71(0.46)
	6		✓		✓	93.98(0.83)	***93.22(1.38)***	91.64(2.80)	92.39(1.16)
	7	✓	✓	✓		***95.05(0.36)***	92.22(0.96)	***95.73(1.01)***	***93.93(0.43)***
	8(Ours)	✓	✓		✓	**95.16(0.30)**	92.09(0.52)	**96.17(0.90)**	**94.08(0.38)**
CIFAR-10	0					66.64(10.34)	32.54(44.64)	21.05(31.74)	24.95(35.84)
	1	✓				89.36(0.55)	***87.19(1.59)***	86.12(2.33)	86.62(0.80)
	2		✓			72.79(12.06)	79.64(8.82)	39.07(33.48)	47.24(31.22)
	3				✓	67.15(11.40)	33.16(45.56)	21.86(33.36)	25.68(37.20)
	4	✓	✓			89.61(0.65)	86.89(1.01)	87.18(0.89)	87.03(0.80)
	5	✓			✓	87.96(2.12)	82.89(2.41)	***88.11(3.82)***	85.4(2.70)
	6		✓		✓	72.73(11.95)	79.73(8.91)	38.78(32.96)	47.13(31.04)
	7	✓	✓	✓		***89.95(0.54)***	86.86(0.64)	**88.23(1.72)**	**87.53(0.79)**
	8(Ours)	✓	✓		✓	**89.98(0.83)**	**87.36(0.82)**	87.63(1.79)	***87.49(1.13)***

(2)平均教师.仍以准确率为例,仅仅对比是否使用平均教师,2 比 0 在 F-MNIST 和 CIFAR-10 上分别提升了 4.03%和 6.15%;4 比 1 在各个基准数据集上分别提升了 0.40%、0.26%和 0.25%;6 比 3 在 F-MNIST 和 CIFAR-10 上分别提升了 2.56%和 5.58%;7 比 5 在各个数据集上分别提升了 1.03%、0.28%和 2.02%.由此可见,平均教师基于平滑假设为模型训练引入了额外的监督信息,进一步提升了分类性能.此外,8 比 5 的准确率标准差分别降低了 0.11%、0.06%和 1.29%,这说明了平均教师有助于所提框架的稳定性.

(3)温度锐化.以准确率为例,仅对比是否使用温度锐化,3 比 0 在所有基准数据集分别提升了 0.21%、0.56%和 0.51%;5 比 1 在 MNIST 和 F-MNIST 上分别提升了 0.03%和 0.17%;6 比 2 在 MNIST 上提升了 0.41%;8 比 4 在各个数据集上分别提升了 0.66%、0.19%和 0.37.由此可见,温度锐化进一步提升了分类性能.

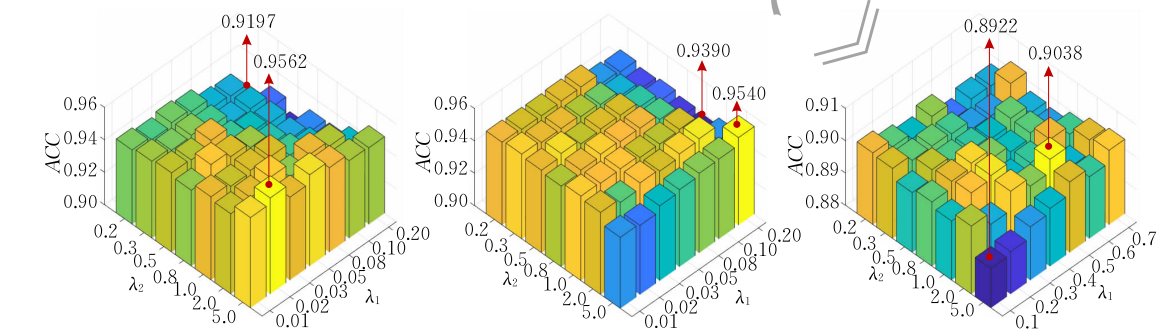
目前,半监督学习的流行做法是对无标签数据打上伪标签.根据聚类假设,伪标签和温度锐化的作用都是使得模型在无标签数据上的输出信息熵应该尽可能小,那么它们的实际效果应当相似.消融实验通过 7 vs 8 对二者进行了比较.实验结果显示,温度锐化与伪标签的各项指标十分接近.

4.6 超参数敏感性分析

所提框架的超参数包括平均教师中式(28)里的平滑系数 α、温度锐化中式(32)里的温度 T、置信度阈值 τ、优化目标中式(33)里的 λ_1 和 λ_2.鉴于 α、T 和 τ 的设置策略已被广泛研究[24,42-43],接下来重点对所提框架的主要参数 λ_1 和 λ_2 展开分析.

λ_1 和 λ_2 分别控制平均教师(L_{con})和温度锐化(L_{sharpen})的重要程度.图 9 展示了基准数据集上的超参数敏感度分析结果,可以发现所提方法对超参数的微小扰动不敏感,具有良好的稳定性.具体而言,在 MNIST 上,λ_1 和 λ_2 的取值区间分别为[0.01,0.2]和[0.2,5],准确率均值为 94.21%,标准差仅为 0.85%;在 F-MNIST 上 λ_1 和 λ_2 的取值区间分别为[0.01,0.2]和[0.2,5],准确率均值为 94.89%,标准差仅为 0.37%;在 CIFAR-10 上 λ_1 和 λ_2 的取值区间分别为[0.1,0.7]和[0.2,5],准确率均值为 89.95%,标准差仅为 0.26%.此外,合理地设置 λ_1 和 λ_2 有利于所提方法分类性能的进一步提升.例如,MNIST 在 $\lambda_1=0.01$ 和 $\lambda_2=5$ 处取得了最佳性能,对应准确率为 95.91%;F-MNIST 在 $\lambda_1=0.2$ 和 $\lambda_2=5$ 处取得了最佳性能,准确率为 95.40%.CIFAR-10 在 $\lambda_1=0.5$ 和 $\lambda_2=2$ 处取得了最佳性能,准确率为 90.45%.

所提方法关于 λ_1 和 λ_2 的稳定性一定程度上可归因于基于置信度的掩码技术.在置信度掩码的作用下,只有高度可信的样本参与了 L_{con} 和 L_{sharpen} 的计算,大多数有害的监督信息被滤除.因此,λ_1 和 λ_2 的微小扰动不会对算法性能造成严重的负面影响.



(a) MNIST (b) F-MNIST (c) CIFAR-10

图 9 基准数据集上的超参数敏感度分析

4.7 主干网络的影响

在此前的对比实验中,不同的数据集采用了不同的神经网络结构.本节探究了不同主干网络对实验结论的影响,结果如表 5 所示.可以发现,所提方法在相同的主干网络设置下均取得了最优的准确率和 $F1$-score,进一步验证了所提方法的有效性.总体来看,在相同的 PU 算法设置下,模型复杂度越高,对应的分类性能通常更好.然而,对于所提框架而言,Resnet-18 的分类性能与 CNN-13 接近.以准确率和 $F1$-score 为例,Resnet-18 较 CNN-13 分别平均仅提升 0.09%和 0.04%.这说明 CNN-13 足以胜任 CIFAR-10 的普适物体分类任务.

表 5　CIFAR-10 上不同的主干网络对算法性能的影响　　（单位：%）

主干网络	评价指标	supervised	uPE	nnPE	uKM1	nnKM1	uKM2	nnKM2	uTIcE	nnTIcE	uCDMM	nnCDMM	Ours
MLP-6	*ACC*	78.69	75.90	76.35	76.31	79.72	77.14	***80.11***	74.47	75.63	70.07	74.22	**81.81**
		(1.08)	(1.14)	(0.45)	(1.46)	(0.75)	(1.05)	(***0.74***)	(8.12)	(8.76)	(6.66)	(7.97)	(**0.12**)
	precision	71.69	67.33	64.84	68.90	71.55	71.28	72.39	81.15	82.16	**86.48**	***82.61***	78.44
		(2.58)	(1.95)	(0.50)	(3.57)	(1.71)	(3.30)	(1.58)	(10.74)	(10.17)	(**8.44**)	(***9.94***)	(0.95)
	recall	77.63	77.53	**89.31**	75.11	***82.02***	72.41	81.38	52.46	55.05	32.44	49.38	75.22
		(3.68)	(4.57)	(**0.64**)	(4.24)	(***2.04***)	(4.43)	(1.76)	(29.45)	(31.12)	(22.67)	(27.87)	(1.39)
	*F*1	74.45	71.98	75.13	71.72	76.40	***71.68***	76.60	56.41	58.26	42.17	55.12	**76.78**
		(1.04)	(1.63)	(0.42)	(1.08)	(0.55)	(***0.97***)	(0.62)	(31.56)	(32.62)	(26.62)	(30.86)	(**0.30**)
CNN-13	*ACC*	84.70	82.86	83.78	75.40	84.92	84.31	***86.58***	81.84	82.57	74.81	79.27	**89.98**
		(0.52)	(2.87)	(0.82)	(19.84)	(0.95)	(4.92)	(***2.57***)	(12.37)	(12.64)	(13.69)	(10.86)	(**0.83**)
	precision	83.64	79.42	80.45	73.01	80.60	82.26	84.04	88.82	**90.54**	***90.33***	87.29	87.36
		(2.28)	(6.00)	(2.26)	(18.71)	(2.64)	(3.44)	(2.26)	(6.41)	(**5.32**)	(***8.99***)	(7.27)	(0.82)
	recall	76.98	77.97	78.66	***83.23***	82.24	77.31	82.01	65.19	65.18	45.78	59.44	**87.63**
		(4.21)	(5.07)	(1.33)	(***9.51***)	(2.46)	(12.96)	(6.09)	(36.79)	(36.50)	(42.03)	(33.44)	(**1.79**)
	F-score	80.06	78.47	79.51	75.49	81.35	79.31	***82.94***	66.90	67.74	47.92	63.07	**87.49**
		(1.42)	(3.15)	(0.69)	(10.35)	(0.98)	(8.32)	(***3.71***)	(37.52)	(37.88)	(43.86)	(35.33)	(**1.13**)
Resnet-18	*ACC*	85.35	89.36	86.05	88.89	88.10	***89.07***	88.25	81.04	82.43	76.02	79.11	**90.07**
		(0.47)	(0.52)	(0.73)	(0.27)	(0.54)	(***0.33***)	(0.36)	(12.29)	(12.68)	(10.03)	(11.56)	(**0.44**)
	precision	84.68	86.98	80.05	85.70	83.52	86.42	83.36	**90.90**	90.12	***90.81***	90.35	87.93
		(0.11)	(1.18)	(3.17)	(2.10)	(3.36)	(1.74)	(1.41)	(**5.14**)	(5.53)	(***5.76***)	(5.93)	(0.96)
	recall	77.39	86.34	***87.24***	86.84	**87.88**	86.32	88.34	60.38	65.31	46.54	55.86	87.17
		(1.59)	(1.07)	(***6.73***)	(2.78)	(**4.08**)	(2.17)	(2.66)	(35.36)	(36.96)	(30.04)	(34.54)	(1.69)
	*F*1	80.86	86.65	83.27	86.21	85.51	***86.33***	85.74	64.86	67.53	55.00	61.31	**87.53**
		(0.81)	(0.62)	(1.64)	(0.46)	(0.55)	(***0.45***)	(0.65)	(36.84)	(37.87)	(32.62)	(35.59)	(**0.65**)

5　总　结

本文提出了一种迭代式的深度 PU 学习与类别先验估计框架. 该框架能准确且稳定地估计类别先验，并利用所得先验估计值训练深度模型，因此无需已知数据的真实先验分布，从而更利于解决实际应用场景下的 PU 学习问题. 本文所提出的迭代框架包括以下核心步骤：(1) 将网络的预测分数建模为 GMM，从而估计类别先验；(2) 基于类别先验的估计值，计算非负 PU 风险；(3) 结合平均教师和温度锐化技术，提高模型性能及稳定性. 本文通过消融实验验证了上述核心设计的合理性，并在多个数据集上进行了大量对比实验，验证了该框架的先进性. 其中，在 Alzheimer 上的实验结果显示该框架能应用于基于 MRI 的 AD 自动识别任务，且效果优于现有方法，进而表明所提框架兼具科学价值和实用价值.

尽管本文为解决大规模 PU 数据上的类别先验估计问题做出了早期探索与尝试，但所提框架仍然存在一些不足之处. 所提框架的主要局限性在于其学习过程较长，超参数个数较多. 为得到更好的分类性能，选取最优的超参数较为繁琐. 此外，所提框架中的深度模型输出预测分数是一个黑盒过程，其可解释性在理论方面相对欠缺. 后续将考虑在算法的简洁性、理论性与实用性方面做出改进.

参　考　文　献

[1] Hestness J, Narang S, Ardalani N, et al. Deep learning scaling is predictable, empirically. arXiv preprint arXiv: 1712.00409, 2017

[2] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683, 2019

[3] Mahajan D, Girshick R, Ramanathan V, et al. Exploring the limits of weakly supervised pretraining//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 181-196

[4] Zhou Z H. A brief introduction to weakly supervised learning. National Science Review, 2018, 5(1): 44-53

[5] Merz C J, Clair D C S, Bond W E. Semi-supervised adaptive resonance theory (smart2)//Proceedings of the International Joint Conference on Neural Networks. Baltimore, USA, 1992, 3: 851-856

[6] Chen X, Chen W, Chen T, et al. Self-PU: Self boosted and calibrated positive-unlabeled training//Proceedings of the International Conference on Machine Learning. Virtual Event, 2020: 1510-1519

[7] Zhang Y L, Li L, Zhou J, et al. POSTER: A PU learning based system for potential malicious URL detection//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA, 2017: 2599-2601

[8] Li H, Chen Z, Liu B, et al. Spotting fake reviews via collective positive-unlabeled learning//Proceedings of the International

Conference on Data Mining. Shenzhen, China, 2014: 899-904

[9] Bepler T, Morin A, Rapp M, et al. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. Nature Methods, 2019, 16(11): 1153-1160

[10] Bekker J, Davis J. Learning from positive and unlabeled data: A survey. Machine Learning, 2020, 109(4): 719-760

[11] He F, Liu T, Webb G I, et al. Instance-dependent pu learning by Bayesian optimal relabeling. arXiv preprint arXiv:1808.02180, 2018

[12] Ienco D, Pensa R G. Positive and unlabeled learning in categorical data. Neurocomputing, 2016, 196: 113-124

[13] Liu L, Peng T. Clustering-based method for positive and unlabeled text categorization enhanced by improved TFIDF. Journal of Information Science and Engineering, 2014, 30(5): 1463-1481

[14] Du Plessis M C, Niu G, Sugiyama M. Analysis of learning from positive and unlabeled data//Advances in Neural Information Processing Systems, Montréal Canada, 2014, 27: 703-711

[15] Du Plessis M, Niu G, Sugiyama M. Convex formulation for learning from positive and unlabeled data//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 1386-1394

[16] Kiryo R, Niu G, Du Plessis M C, et al. Positive-unlabeled learning with non-negative risk estimator//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 1674-1684

[17] Du Plessis M C, Sugiyama M. Class prior estimation from positive and unlabeled data. IEICE Transactions on Information and Systems, 2014, 97(5): 1358-1362

[18] Ramaswamy H, Scott C, Tewari A. Mixture proportion estimation via kernel embeddings of distributions//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 2052-2060

[19] Bekker J, Davis J. Estimating the class prior in positive and unlabeled data through decision tree induction//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2712-2719

[20] Łazęcka M, Mielniczuk J, Teisseyre P. Estimating the class prior for positive and unlabelled data via logistic regression//Advances in Data Analysis and Classification, 2021, 15(4): 1039-1068

[21] Blanchard G, Lee G, Scott C. Semi-supervised novelty detection. Journal of Machine Learning Research, 2010, 11: 2973-3009

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

[23] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv: 1703.01780, 2017

[24] Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training//Advances in Neural Information Processing Systems. Virtual Event, 2020, 33: 6256-6268

[25] Jiang L, Zhou Z, Leung T, et al. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 2304-2313

[26] Kato M, Teshima T, Honda J. Learning from positive and unlabeled data with a selection bias//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[27] Hammoudeh Z, Lowd D. Learning from positive and unlabeled data with arbitrary positive shift//Advances in Neural Information Processing Systems. Virtual Event, 2020, 33: 13088-13099

[28] Su G, Chen W, Xu M. Positive-unlabeled learning from imbalanced data//Proceedings of the International Joint Conference on Artificial Intelligence. Virtual Event, 2021: 2995-3001

[29] Liu B, Lee W S, Yu P S, et al. Partially supervised classification of text documents//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2002, 2(485): 387-394

[30] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency//Advances in Neural Information Processing Systems. Montréal Canada, 2004: 321-328

[31] Li X L, Liu B, Ng S K. Learning to identify unexpected instances in the test set//Proceedings of the International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 2802-2807

[32] Du Plessis M C, Sugiyama M. Semi-supervised learning of class balance under class-prior change by distribution matching. Neural Networks, 2014, 50: 110-119

[33] Northcutt C G, Wu T, Chuang I L. Learning with confident examples: Rank pruning for robust classification with noisy labels. arXiv preprint arXiv:1705.01936, 2017

[34] Hsieh Y G, Niu G, Sugiyama M. Classification from positive, unlabeled and biased negative data//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 2820-2829

[35] Gong C, Shi H, Liu T, et al. Loss decomposition and centroid estimation for positive and unlabeled learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(3): 918-932

[36] Gong C, Wang Q, Liu T, et al. Instance-dependent positive and unlabeled learning with labeling bias estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4163-4177

[37] Zhu X, Goldberg A B. Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1-130

[38] Zhu X J. Semi-supervised learning literature survey. University of Wisconsin-Madison, Madison, USA: Technical Reports: TR1530, 2005

[39] Rasmus A, Berglund M, Honkala M, et al. Semi-supervised learning with ladder networks//Advances in Neural Information Processing Systems. Montréal, Canada, 2015, 28: 3546-3554

[40] Bachman P, Alsharif O, Precup D. Learning with pseudo-ensembles//Advances in Neural Information Processing Systems. Montréal, Canada, 2014, 27: 3365-3373

[41] Laine S, Aila T. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242, 2016

[42] Berthelot D, Carlini N, Goodfellow I, et al. MixMatch: A holistic approach to semi-supervised learning//Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, USA, 2019: 5049-5059

[43] Berthelot D, Carlini N, Cubuk E D, et al. ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[44] Niyogi P. Manifold regularization and semi-supervised learning: Some theoretical analyses. Journal of Machine Learning Research, 2013, 14: 1229-1250

[45] Sohn K, Berthelot D, Carlini N, et al. FixMatch: Simplifying semi-supervised learning with consistency and confidence//Advances in Neural Information Processing Systems. Virtual Event, 2020, 33: 596-608

[46] Stauffer C, Grimson W E L. Adaptive background mixture models for real-time tracking//Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. Fort Collins, USA, 1999: 246-252

[47] Ma Z, Leijon A. Bayesian estimation of beta mixture models with variational inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(11): 2160-2173

[48] Permuter H, Francos J, Jermyn I. A study of Gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recognition, 2006, 39(4): 695-706

[49] McLachlan G J, Krishnan T. The EM Algorithm and Extensions. Hoboken, USA: John Wiley & Sons, 2007

[50] Shental N, Bar-Hillel A, Hertz T, et al. Computing Gaussian mixture models with EM using equivalence constraints//Advances in Neural Information Processing Systems. Vancouver, Canada, 2004, 16(8): 465-472

[51] Arazo E, Ortego D, Albert P, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning//Proceedings of the International Joint Conference on Neural Networks. Virtual Event, 2020: 1-8

[52] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[53] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017

[54] Krizhevsky A, Hinton G. Learning Multiple Layers of Features from Tiny Images[M. S. dissertation]. University of Toronto, Toronto, Canada, 2009

[55] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 448-456

[56] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines//Proceedings of the International Conference on International Conference on Machine Learning. Haifa, Israel, 2010: 807-814

[57] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net. arXiv preprint arXiv: 1412.6806, 2014

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

[59] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[60] Loshchilov I, Hutter F. Fixing weight decay regularization in Adam. arXiv preprint arXiv:1711.05101, 2017

ZHAO Yun-Rui, M. S. candidate. His research interest is weakly supervised learning.

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

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

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

Background

With the advent of big data, deep neural networks have attracted extensive attention and their performance has reached or even surpassed the human level in various tasks. Specifically, such great success usually relies on full supervision by a large amount of labelled data. However, it is hard to obtain intact label information in many real-world applications, even those with only binary options. For example, observed interactions between users and items in recommendation systems are labelled positives. Since many unconcerned factors like lack of exposure or other coincidences could account for missing interactions, we cannot view all the unobserved interactions as negatives. Similar scenarios include Alzheimer's disease recognition, malicious URL detection, and particle picking in cryo-electron micrographs, where we only have access to a few labelled positives with plenty of unlabeled data. Such a great demand motivates us to learn from positive and unlabeled data, also known as PU learning.

While PU learning based on unbiased risk estimators has achieved the state-of-the-art performance on several benchmarks, it relies on the knowledge of class prior, which might be unknown in reality. To this end, there emerge some class prior estimation algorithms designed for PU data. However, these algorithms are found difficult in handling large-scale data.

In this paper, we propose an iterative framework for deep PU learning and class prior estimation utilizing an unsupervised mixture model. Specifically, positive and negative classes have distinct predicted score distributions intuitively. We investigate and demonstrate our intuition and approximate the score distributions by a Gaussian Mixture Model with two components each representing the relevant class-conditional distribution. Accordingly, class prior is estimated through the weights of the Gaussian Mixture Model. With mean teacher and temperature sharpening incorporated, our proposed framework could estimate the class prior and learn from PU data simultaneously, achieving competent performance with other PU competitors based on the ground-truth class prior.

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