中国科学: 信息科学 2024 年 第 54 卷 第 3 期: 603–622
SCIENTIA SINICA Informationis
论文

基于混合比例估计的标签噪声学习方法

郑庆华[1,3], 曹书植[1,3], 阮建飞[1,3*], 赵锐[1,3], 董博[2,4*]

1. 西安交通大学计算机科学与技术学院, 西安 710049
2. 西安交通大学继续教育学院, 西安 710049
3. 智能网络与网络安全教育部重点实验室, 西安 710049
4. 陕西省天地网技术重点实验室, 西安 710049
* 通信作者. E-mail: jianfei.ruan@hotmail.com, dong.bo@xjtu.edu.cn

收稿日期: 2023–05–06; 接受日期: 2023–08–16; 网络出版日期: 2024–03–07

科技创新 2030 —“新一代人工智能” 重大项目 (批准号: 2020AAA0108800)、国家自然科学基金 (批准号: 62037001, 61721002, 62002282)、教育部创新团队项目 (批准号: IRT_17R86)、西安交通大学本科教学改革研究项目 (批准号: 20JX04Y) 和西安交大 – 税友集团税务大数据协同创新项目资助

摘要 近年来, 人工智能蓬勃发展, 伴随着计算硬件算力的提升, 深度学习已成为了人工智能算法的新范式. 然而深度学习依赖大量精确标注的数据, 在现实的多类别分类场景中, 受限于标注成本和隐私数据保护等因素, 大量精准标注的数据往往难以获得. 近些年, 移动众包和网络爬虫这类经济廉价的数据收集方法被广泛采用, 但他们不可避免地引入了错误标注, 即标签噪声. 鉴于深度神经网络强大的数据拟合能力, 标签噪声的存在将造成算法的过拟合, 严重制约了深度学习方法的泛化能力. 针对标签噪声问题, 现有研究大多显式或隐式地依赖锚点 (明确属于某一类别的样本), 然而在现实场景中锚点难以获取, 这使得现有解决方案不再适用. 为解决这一问题, 本文创造性地将多类别标签噪声学习问题转化为混合比例估计 (mixture proportion estimation, MPE) 问题, 构建了一种不依赖锚点的满足统计一致性的学习算法. 本文的主要贡献包括: (1) 对现有的仅适用于二组成物 MPE 场景的 R-MPE (regrouping-MPE) 方法进行推广, 提出了多组成物场景下不依赖不可约假设的 MPE 方法 MR-MPE (multi-component oriented R-MPE); (2) 理论上证明了多类别分类场景下标签噪声学习算法锚点假设和 MPE 问题不可约假设的等价性, 并基于所提出的 MR-MPE 方法构建了不依赖锚点的满足统计一致性的算法. 本文在合成噪声数据集和真实噪声数据集上分别与现有算法进行了对比实验, 结果显示本文所提算法在多个数据集上均展现出了最优的性能; 同时, 在移除锚点的情况下, 本文对算法的鲁棒性进行了测试, 验证了所提算法不依赖锚点的特性.

关键词 混合比例估计, 多类别分类, 标签噪声学习, 锚点, 不可约假设, 统计一致性

引用格式: 郑庆华, 曹书植, 阮建飞, 等. 基于混合比例估计的标签噪声学习方法. 中国科学: 信息科学, 2024, 54: 603–622, doi: 10.1360/SSI-2023-0126
Zheng Q H, Cao S Z, Ruan J F, et al. Label-noise learning via mixture proportion estimation (in Chinese). Sci Sin Inform, 2024, 54: 603–622, doi: 10.1360/SSI-2023-0126

1 引言

伴随着数字化产业的蓬勃发展和 “智能 +” 时代的加速到来, 人工智能技术方兴未艾. 在摩尔 (Moore) 定律 [1] 的指引下, 集成电路工艺日新月异, 计算机计算能力大幅提升. 这使得数据驱动的计算智能 —— 深度学习 —— 在大规模工程化应用方面取得了长足进步, 被广泛地应用于计算机视觉 [2]、自然语言处理 [3]、推荐系统 [4,5] 等领域.

深度学习作为数据驱动的方法, 其性能依赖大量的精准标注数据. 但在现实的多类别分类场景中, 受限于人工标注成本和隐私数据保护等因素, 大量的精确标注数据难以获取. 近些年, 移动众包 [6~8] 和网络爬虫技术 [9] 这类经济廉价的数据收集方法被广泛地使用, 但它们不可避免地引入了错误标注, 即标签噪声 [10,11]. 而基于多层神经网络的深度学习方法, 从 MLP [12] 到 LeNet [13] 到 VGG [14] 到 GoogleNet [15] 到 ResNet [16] 到 ViT [17,18], 神经网络规模日益庞大, 对数据的拟合能力愈发增强. 标签噪声的存在将造成神经网络的过拟合, 严重制约了深度学习方法的泛化能力 [10].

为了解决多类别分类场景下的标签噪声学习问题, 现有方法主要从启发式算法 [19~24] 和统计一致性算法 [25~31] 两个角度展开研究. 前者一般通过选取可信样本或者纠正样本标签的方法来减少标签噪声带来的副作用, 然而其并不能在理论上保持统计一致性 (与真实标签数据监督下构建的分类器具有一致的分类风险), 这使得这一类算法缺乏理论保障. 后者虽然实现了理论上的统计一致性, 但其中绝大部分的算法依赖锚点 (明确属于某一类别的样本) [32,33], 然而在锚点难以获取的现实场景中, 此类方案不再适用. 目前一些方法开始尝试在不依赖锚点的情况下探索统计一致性算法, 但这些方法却引入了额外的假设, 例如, 对数据的稀疏性提出了要求 [34]. 因此, 在多类别分类场景下, 如何仅基于标签噪声数据, 在保证统计一致性的前提下, 构建不依赖锚点和其他假设的学习算法, 已经成为了一个亟待解决的问题.

针对以上问题, 本文从混合比例估计 (mixture proportion estimation, MPE) 问题的角度出发, 将多类别分类场景下的标签噪声学习问题转化为了 MPE 问题, 进而构建了不依赖锚点假设的统计一致性算法, 本文的主要贡献包括以下 3 个方面. (1) 对现有的仅适用于二组成物场景的 MPE 方法 R-MPE (regrouping-MPE) 进行推广, 提出了多组成物场景下不依赖不可约假设的 MPE 方法 MR-MPE (multi-component oriented R-MPE). (2) 理论上证明了多类别分类场景下标签噪声学习算法锚点假设和 MPE 问题不可约假设的等价性, 并基于所提出的 MR-MPE 方法构建了不依赖锚点的统计一致性的标签噪声学习方法. (3) 本文分别在合成噪声数据集 CIFAR-10 和 CIFAR-100 以及真实噪声数据 CIFAR-10N [35], CIFAR-100N [35] 和 Yorúbá [36] 上与现有主流方法进行了对比实验, 本文所提方法在多个数据集上均展现出了最优的性能, 其中本文所提算法在 CIFAR-10 (20% 噪声率)、CIFAR-100 (20% 和 50% 噪声率)、CIFAR-10N、CIFAR-100N 和 Yorúbá 数据集上均取得了最优的分类精度; 同时, 本文通过人工移除近似锚点的方式, 验证了本文所提方法在锚点缺失情况下相比现有方法具有更好的鲁棒性.

本文剩余章节组织如下: 第 2 节总结了标签噪声学习的相关工作; 第 3 节首先回顾了 MPE 问题的定义和相关工作, 之后阐述了不可约假设的概念并分析了该假设不成立情况下, MPE 方法的误差, 最终, 介绍了不依赖不可约假设的 MPE 方法 R-MPE; 第 4 节则定义了多组成物场景下的 MPE 问题, 并提出了不依赖不可约假设的多组成物场景下的 MPE 方法 MR-MPE; 第 5 节将标签噪声学习问题进一步转化为多组成物场景下的 MPE 问题, 并提出了基于 MR-MPE 的标签噪声学习方法; 第 6 节则对本文所提方法在合成噪声数据集和真实噪声数据集上进行了实验测试, 进一步验证了本文提出方法的优越性; 最终, 第 7 节对本文进行了总结与展望.

2 相关工作

针对标签噪声问题, 目前主流的算法可分为启发式算法与统计一致性算法, 本节将对这两类算法进行具体的介绍.

2.1 启发式算法

启发式算法通常从筛选可信样本或纠正样本标签的角度入手, 最大程度地消除样本的标签噪声信息. 根据启发式算法的具体实施策略, 现有启发式算法可大体上分为: 可信样本选取和样本标签纠正.

可信样本选取策略. 筛选可信的样本 (无噪声标签样本) 对神经网络进行训练, 尽可能地减少标签噪声样本带来的副作用. 代表性地, Jiang 等 [37] 提出了 MentorNet 学习算法, 该方法从 Curriculum Learning (课程学习) [38] 的角度出发, 通过构建一个额外的网络筛选可信的样本来指导神经网络的训练, 但该方法不可避免地引入了神经网络的选择偏差 [39]. 针对这一问题, 为了减少单一网络筛选样本带来的选择偏差, Malach 等 [19] 提出了 Decoupling 方法, 该方法通过同时构建两个神经网络, 在全部样本中筛选预测不一致的样本作为训练样本来对两个神经网络同时进行训练. Han 等 [20] 提出的 Co-teaching 方法则进一步地提升了 Decoupling 方法, 具体地, 该方法通过同时训练两个神经网络进行相互教学, 对相同的输入数据, 两个网络同时进行前向传播计算, 并各自筛选相同比例的损失较小的样本作为可信样本, 进而基于对方筛选的可信样本进行反向传播训练, 从而避免了各自选择偏向带来的误差. 然而上述方法均将损失较小的样本作为可信样本, 但这样的筛选标准使得位于决策边界的样本被丢弃, 导致这一类方法缺乏泛化能力. 针对这一问题, Wei 等 [40] 提出了 SFT 方法, 该方法基于样本历史预测中类别的波动情况筛选可信样本, 从而避免了将损失较小的样本作为可信样本所带来的选择偏差.

样本标签纠正策略. 从纠正噪声标签的角度出发, 减少噪声标签数据带来的干扰. 其中, Tanaka 等 [24] 将样本的标签视为一个随机变量, 与神经网络的参数一起进行迭代, 在训练过程中纠正样本的标签, 并将标签纠正后的数据集作为监督数据指导神经网络的训练; 另一种方法则从主动学习 [41] 的角度出发, 挑选分类困难的样本进行人工重标注, 对样本标签进行纠正, 从而提升分类器的性能. 上述两种策略均在一定程度上减少了标签噪声数据带来的副作用, 然而, 它们并不能在理论上保持统计一致性, 即不能保证基于标签噪声数据构建的分类器与以正确标注数据作为监督构建的分类器具有相同的分类风险.

2.2 统计一致性算法

针对启发式算法存在的理论保障欠缺的问题, 学术界开始研究统计一致性算法. 统计一致性算法一般通过修改传统损失函数的方式来保证算法的统计一致性. 其中, 转移矩阵的精准估计对损失函数的修正至关重要. 在一个 c 分类的多分类的任务中, 转移矩阵 T 是一个 $c\times c$ 的方阵, 其第 i 行, 第 j 列的元素 T_{ij} 表示给定真实标签 $Y=i$ 情况下噪声标签 $\bar{Y}=j$ 的条件概率, 即有 $T_{ij}=P(\bar{Y}=j|Y=i)$. 根据转移矩阵的估计是否利用锚点 (如定义 1 所示), 统计一致性算法可以分为基于锚点的统计一致性算法和不基于锚点的统计一致性算法两类.

定义1 (锚点) 在一个具有 c ($c\geqslant 2$) 类的多分类任务中, 若存在一个样本 x^i, 其属于 $Y=i$, $i\in\{1,2,\ldots,c\}$ 类的后验概率为 1 或趋近于 1, 即有 $P(Y=i|X=x^i)\to 1$, 则称样本 x^i 为锚点.

基于锚点的统计一致性算法. 基于锚点的统计一致性算法利用锚点来初始化转移矩阵, 进而构建统计一致性算法. 假定样本 X 的噪声标签后验概率为 $P(\bar{Y}|X)=[P(\bar{Y}=1|X),\ldots,P(\bar{Y}=c|X)]^{\mathrm{T}}$,

真实标签后验概率为 $P(Y|X) = [P(Y = 1|X), \ldots, P(Y = c|X)]^{\mathrm{T}}$, 则两者之间满足关系: $P(\bar{Y}|X) = T^{\mathrm{T}}P(Y|X)$. 由于样本对应的噪声后验概率 $P(\bar{Y}|X)$ 可通过噪声数据估计得到, 在转移矩阵 T 确定的情况下, 真实标签后验概率 $P(Y|X)$ 可以直接推理得到. 因此, 确定转移矩阵 T 是标签噪声学习的关键, 当下的基于锚点的统计一致性算法均显式或隐式地利用锚点来初始化转移矩阵, 具体而言, 给定锚点 x^i, 结合其性质 $P(Y = i|X = x^i) \to 1, P(Y \neq i|X = x^i) \to 0$ 可以得到如下的关系:

$$P\left(\bar{Y} = j\middle|x^i\right) = \sum_{k=1}^{c} P(\bar{Y} = j\,|Y = k)\, P\left(Y = k\middle|x^i\right) = P(\hat{Y} = j|Y = i) = T_{ij}. \tag{1}$$

给定明确属于 $Y = i$ 的锚点 x^i, 通过估计其噪声标签后验概率 $P(\bar{Y} = j|X = x^i), j \in \{1, 2, \ldots, c\}$, 可以唯一地确定转移矩阵 T 的第 i 行元素. 因此, 在每一类别对应锚点均已知的情况下, 转移矩阵 T 可以被唯一地确定. 现有标签噪声研究均显式或隐式地依赖锚点, 具体而言, Yu 等 [31] 探索了完全噪声情况下的学习方法, 并假设明确属于每一个类别的一组真实标签样本 (锚点) 已知, 这显式地利用了锚点. Liu 等 [25] 和 Patrini 等 [30] 尽管不直接要求得到一组额外的真实标签样本 (锚点), 但是这一类方法均使用噪声类后验概率 $P(\bar{Y}|X = x) \to 1$ 的样本作为近似的锚点, 这一类方法则隐式地依赖于锚点. 尽管上述算法在理论上保证了统计一致性, 即该类算法学习得到的分类器将趋于真实标签样本监督下的最优贝叶斯分类器, 然而它们显式或隐式地依赖锚点. 在实际问题中, 锚点往往难以获取, 这是由于当下的大规模的数据往往来自于爬虫或众包收集, 由于缺乏专家标注, 这一类经济廉价的数据收集方式所获取的样本标签质量较差, 由于数据规模庞大, 直接获取一组真实标签样本十分困难. 同时, 在大规模数据集中, 样本类别数目极多, 受限于标注成本, 针对每一个类别均进行专家标注, 获取明确属于这一类的样本 (锚点) 并不现实. 例如在 ImageNet 数据集之中, 仅样本类别就超过了 2 万个, 即使每一个类别只获取一个锚点, 也将对数据收集工作带来极大挑战. 而在特定的场景下, 锚点样本可能并不存在 [32, 33], 例如在一些文本分类的场景中, 一些文本的语义是极其模糊的, 对于特定的文本类别而言, 难以定义明确属于该类别的锚点样本, 这使得锚点可能并不存在. 因此, 锚点假设过于强烈 [42, 43], 在现实的场景之中难以满足.

不基于锚点的统计一致性算法. 为了解决锚点依赖算法的局限性, 学术界开始研究不基于锚点的统计一致性算法. 具体而言, Xia 等 [32] 提出了 T-Revision 方法, 区别于现有方法直接利用已知锚点初始化转移矩阵的策略, 该方法将分类任务划分为两个阶段: 第一阶段, 筛选噪声标签后验概率较大的样本点作为近似锚点, 用于初始化转移矩阵; 第二阶段, 通过加入松弛变量迭代地修正第一阶段得到的转移矩阵, 从而得到估计更为准确的转移矩阵. T-Revision 方法尽管不直接使用锚点来估计转移矩阵, 却间接地使用了近似锚点, 若原始的样本中实际上不存在锚点, 则此处利用近似锚点估计得到的初始的转移矩阵会产生较大的误差, 因此该方法并不能真正地解决统计一致性算法存在的锚点依赖问题. Li 等 [34] 则提出了 VolMinNet 的方法, 该方法从几何的角度出发, 将标签噪声问题转化为了一个最小化单纯形面积的最优化问题. 然而单纯形的优化往往依赖于有效的噪声类后验概率估计, 而在现实场景之中, 噪声类后验概率容易被过拟合, 这影响了该方法的性能. 为了解决这一问题, 在 VolMinNet 方法的基础上, Cheng 等 [44] 提出了 CCR 方法, 该方法将单纯形从噪声类后验空间投影至真实类后验空间避免了噪声类后验概率难以估计的问题, 通过最大化真实类后验概率对应的单纯形面积并结合正则化约束来学习转移矩阵. 以上的 VolMinNet 和 CCR 方法虽不直接依赖锚点, 但是却依赖数据稀疏性的先验知识 —— Sufficiently Scattered 假设 [34], 而数据是否满足该假设在计算上难以检验. 因此, VolMinNet 和 CCR 方法均没有在根本上解决锚点依赖的问题. 针对上述问题, Zhu 等 [33] 提出了 HOC 方法, 该方法从可聚类的条件出发, 使用样本表征近邻间的三阶一致性关系来估计转移矩阵. 然而, 聚

类方法的引入带来了新的误差, 也未能真正解决锚点依赖的问题. 因此, 如何构建不基于锚点的统计一致性算法仍是一个亟待解决的问题.

3 混合比例估计 (MPE) 问题

本节介绍了 MPE 问题的定义, 回顾了关于 MPE 问题的相关工作, 分析了 MPE 问题中不可约假设与标签噪声学习问题中锚点假设的关系, 总结了传统 MPE 方法存在的误差, 并介绍了不依赖不可约假设的 MPE 方法 R-MPE.

3.1 MPE 问题的定义

MPE 问题 [23, 26] 是一个统计推理问题: 假设一个混合物分布 F 可以表示为两个组成物分布 H 和 G 的凸组合, MPE 问题的目标是确定各组成物所对应的比例系数, 该问题可形式化表示为

$$F = (1-\kappa)G + \kappa H, \tag{2}$$

其中, F, G 和 H 是定义在希尔伯特空间 [45] $\mathcal{X}$ 上的分布函数, 这里的 κ 是组成物 H 所对应的比例系数. F 表示混合物的分布, G 和 H 表示两个组成物的分布. 假设 X_F 和 X_H 分别表示从混合物分布 F 和组成物分布 H 中采样得到的样本集, 且在该问题中是已知的, MPE 问题的目标是在没有任何关于组成物分布 G 观测数据的情况下, 求解各组成物所对应的比例系数.

3.2 MPE 问题的相关工作

在 MPE 问题中, 由于仅有从混合物分布 F 和组成物分布 H 中采样的样本集 X_F 和 X_H, 若不对未知组成物分布 G 做任何假设, 该问题是不可解的 [46]. 举例来说, 如果组成物分布 G 可表达为 $G = (1-\beta)K + \beta H$, 其中 $\beta \in (0,1]$, K 是一个定义在希尔伯特空间 $\mathcal{X}$ 上的分布, 则混合物分布 F 可进一步地表达为 $F = (1-\kappa)G + \kappa H = (1-\kappa)(1-\beta)K + ((1-\kappa)\beta + \kappa)H$. 这表明, 如果对分布 G 不做任何约束, κ 和 $(1-\kappa)\beta + \kappa$ 均是合理的比例系数. 因此, 在不增加额外假设的情况之下, κ 不具有唯一解. 为了唯一地确定 κ, 相关的研究 [42, 47, 48] 均对组成物分布 G 的潜在形式做出了假设. 目前绝大多数的 MPE 研究均假设组成物分布 G 满足不可约假设 (见定义 2) [47].

定义2 (不可约假设) 如果不存在这样的一种分解满足关系 $G = (1-\beta)K + \beta H$, 其中 G, K, H 均是定义在希尔伯特空间 $\mathcal{X}$ 的概率分布, 且 $0 < \beta \leqslant 1$, 则称分布 G 对分布 H 是不可约的.

在组成物分布 G 满足不可约假设的情况下, MPE 问题的对应解 κ 是唯一的. 假设 $f_F(x)$, $f_G(x)$ 和 $f_H(x)$ 分别代表分布 F, G 和 H 对应的密度函数, 则对于 $\forall x \in \mathcal{X}$, 有

$$f_F(x) = (1-\kappa) f_G(x) + \kappa f_H(x). \tag{3}$$

根据定义 2, 若组成物分布 G 对分布 H 是不可约的, 则形如 $f_G(x) = (1-\beta) f_K(x) + \beta f_H(x), \beta \in (0,1]$ 的分解是不存在的. 也就是说, 在不可约假设成立的情况下, 对于 $\forall \beta \in (0,1]$ 都存在 $x \subseteq \mathcal{X}$, 使得 $S(x) = f_G(x) - \beta f_H(x) < 0$ (确保 $f_K(x)$ 不是有效的概率密度函数). 考虑以上存在性条件成立的情况, 当 $f_H(x) = 0$ 时, $S(x) = f_G(x) \geqslant 0$ 恒成立, 这时候不存在 $x \subseteq \mathcal{X}$ 使得 $S(x) < 0$. 故只需要考虑 $f_H(x) > 0$ 的情况, 这时候, $\exists x \subseteq \mathcal{X}, f_H(x) > 0$, 使得 $S(x) < 0$, 即存在 $f_G(x) < \beta f_H(x)$. 该条件可以进一步地表达为对于 $\forall \beta \in (0,1]$, 应当存在 $x \subseteq \mathcal{X}, f_H(x) > 0$, 使得 $\beta > f_G(x)/f_H(x)$, 则上述条件可

以进一步表达为如下的形式:

$$\beta > \frac{f_G(x)}{f_H(x)} \geqslant \inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_G(x)}{f_H(x)}. \tag{4}$$

由于 $f_G(x) \geqslant 0$ 且有 $f_H(x) > 0$, 可以得到如下的关系:

$$\gamma = \inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_G(x)}{f_H(x)} \geqslant 0. \tag{5}$$

假设 $\gamma > 0$, 则有 $\beta > \gamma > 0$, 这与 $\beta \in (0, 1]$ 的任意性要求相违背, 因此假设错误. 所以在不可约假设成立的情况下, $\gamma = 0$, 即有如下的关系:

$$\gamma = \inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_G(x)}{f_H(x)} = 0, \tag{6}$$

在式 (6) 中, $f_G(x)/f_H(x)$ 的下界取值于希尔伯特空间上 $f_H(x) > 0$ 的部分, 即有 $x \subseteq \mathcal{X}, f_H(x) > 0$. 将式 (3) 左右两边同时除以 $f_H(x)$ 并取下界, 则有

$$\inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_F(x)}{f_H(x)} = (1-\kappa) \inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_G(x)}{f_H(x)} + \kappa, \tag{7}$$

结合式 (6), 可以得到

$$\kappa = \kappa(F|H) = \inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_F(x)}{f_H(x)}, \tag{8}$$

其中, $\kappa(F|H)$ 为 $f_F(x)/f_H(x)$ 的下界, 代表了组成物分布 H 占混合物分布 F 的最大比例. 式 (8) 表明, 在分布 G 对分布 H 满足不可约假设的情况下, MPE 问题对应解 κ 将趋于 $\kappa(F|H)$, 在理论上 MPE 问题的解将被唯一地确定. 通过求解 $\kappa(F|H)$, 可以估计得到 MPE 问题的比例系数 κ.

现有的 MPE 研究均依赖不可约假设. 具体地, Blanchard 等 [47] 最早在半监督的异常值检测任务中提出了 MPE 问题, 并且给出了不可约假设的定义, 证明了在不可约假设满足的情况下, κ 是唯一的. 之后 Scott [49] 在 Blanchard 等的工作基础上进行了改进, 构建了类条件概率密度估计器, 并通过 ROC 曲线来求解 κ. 近年来, 更多的新的 MPE 方法被提出, 例如 Yu 等 [42] 从极大似然估计的角度出发求解 κ, Ramaswamy 等 [48] 则从核均值嵌入的角度出发, 将原始的问题转化为了一个凸优化问题来间接地求解 κ. 尽管上述 MPE 方法的具体实现方式不同, 然而其本质上均显式或隐式地通过求解式 (8) 来估计 κ, 因此传统的 MPE 方法可以总结为算法 1 的形式.

Algorithm 1 Traditional MPE

Input: F: the mixture distribution; H: the component distribution; C: a set of all possible latent distributions.
Output: $\kappa(F|H)$: the maximum mixture proportion of H in F.

```
κ(F|H) ⇐ 0;
for κ = 0 to 1 do
    M ⇐ (F − κH)/(1 − κ);
    if M ∈ C then
        κ(F|H) ⇐ κ;
    end if
end for
```

3.3 不可约假设和锚点假设的关联性分析

根据式 (6), 如果组成物分布 G 对分布 H 是不可约的, 则一定存在一个样本 $x' \in \mathcal{X}$, 满足 $f_G(x') \to 0$, $f_H(x') > 0$. 以二分类的场景为例, 假设 Y 表示样本所属的类别, 分布 G 和 H 分别表示正样本 $(Y=+1)$ 对应的分布和负样本 $(Y=-1)$ 对应的分布, 由于 $f_G(x)$ 表示分布 G 对应的密度函数, 则 $f_G(x)$ 表示正样本分布对应的密度函数, 即有 $f_G(x) = P(x|Y=+1)$. 根据不可约假设, 由于存在一个样本 $x' \in \mathcal{X}$ 满足 $f_G(x') \to 0$, 则一定有 $P(x'|Y=+1) \to 0$. 根据贝叶斯公式, 可以得到如下的关系:

$$P(x'|Y=+1) = \frac{P(Y=+1|x')\,P(x')}{P(Y=+1)} \to 0. \tag{9}$$

通过式 (9) 可以进一步得到 $P(Y=+1|x') \to 0$, 在二分类任务中, 任意的样本 x' 均满足条件: $P(Y=+1|x') + P(Y=-1|x') = 1$, 因此, 样本 x' 满足 $P(Y=-1|x') \to 1$, 这表明样本 x' 是一个锚点 (见定义 1). 以上的讨论表明, MPE 问题中的不可约假设本质上与标签噪声学习问题中的锚点假设是等价的. 如果可以在不依赖不可约假设的情况下求解 MPE 问题, 则可以在此基础上构建不借助锚点的标签噪声学习方法.

3.4 误差分析

根据 3.2 小节的讨论, 现有的 MPE 方法均依赖不可约假设. 然而, 由于没有任何关于组成物分布 G 的观测数据, 式 (6) 中不可约假设成立与否无法验证. 当不可约假设不成立时, 组成物分布 G 和 H 对应的密度函数满足如式 (10) 所示关系:

$$\gamma = \inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_G(x)}{f_H(x)} > 0, \tag{10}$$

这时, 通过求解式 (8) 估计 κ 会引入估计误差, 如式 (11) 所示:

$$\kappa(F|H) = \kappa + (1-\kappa)\inf_{x\subseteq\mathcal{X}, f_H(x)>0} \frac{f_G(x)}{f_H(x)} = \kappa + (1-\kappa)\gamma. \tag{11}$$

式 (11) 说明, 在不可约假设不成立时, 传统 MPE 方法会引入 $(1-\kappa)\gamma$ 的估计误差. 而当 γ 较大时, 估计误差较大, 这严重地制约了传统 MPE 方法的性能. 因此, 如何构建不依赖不可约 (锚点) 假设的 MPE 方法, 是改进传统 MPE 方法的主要方向.

3.5 R-MPE

当不可约假设不成立时, 直接使用传统 MPE 方法估计 κ 将会引入较大的估计误差. 针对这一问题, Yao 等 [46] 提出了一种不借助额外假设的 MPE 方法 R-MPE. R-MPE 方法首先重构一个全新的 MPE 问题 —— 新的组成物分布 H^r 和 G^r 天然地满足不可约假设 —— 进而可以基于传统的 MPE 方法求解新问题的对应解 κ^r. R-MPE 同时在理论上保证了全新问题的对应解 κ^r 趋于原始 MPE 问题的对应解 κ, 从而间接地求解了原始 MPE 问题. R-MPE 算法总结如算法 2 所示.

算法 2 中, 超参数 P_r 表示用于重构组成物的样本复制比例系数. 从理论上而言, 超参数 P_r 越小, R-MPE 方法产生的估计误差越小. 然而在现实问题中, 当 P_r 过分小时, R-MPE 方法对现有的 MPE 方法的改进效果并不明显. 根据相关研究 [46], 当超参数 P_r 选取为 10% 时, R-MPE 可以展现出最好的效果, 最大限度缩小了估计的误差. 因此, 在本文的后续实验中, 将固定超参数 P_r 为 10%.

Algorithm 2 R-MPE

Input: X_F: positive sample i.i.d. drawn from the distribution F; X_H: negative sample i.i.d. drawn from the distribution H; P_r: the percentage of the sample needed to copy from X_F to X_H.

output: κ^r: the mixture proportion.

1: Train a binary classifier h with X_F and X_H;

2: Assign each example $x \in X_F$ with a class posterior probability $P(Y=1|X=x)$ predicted by trained classifier h on the samples X_F and X_H;

3: Obtain X_{H^r} by copying $P_r \times |X_F|$ examples with the small probability from F to X_H;

4: Estimate κ^r by employing Algorithm 1 with inputs X_F and X_{H^r}.

4 多组成物场景的 MPE 方法 MR-MPE

4.1 多组成物场景的 MPE 问题

以上所讨论的 MPE 问题仅限于二组成物的场景, 由于标签噪声问题本质上是一个 MPE 问题[46], 并且在现实场景中, 多类别分类的场景更加普遍. 因此, 将 MPE 问题进一步推广至多组成物的场景下十分必要. 本小节定义了多组成物场景下的 MPE 问题: 假设一个混合物分布 F 是 c ($c>2$) 个组成物分布的凸组合, 该问题可形式化表示为

$$F=\sum_{n=1}^{c-1}\kappa_n H_n+\kappa G. \tag{12}$$

假设已知从混合物分布 F 中采样的样本集 X_F 和分别从其中 $c-1$ 个组成物分布 H_n ($n\in\{1,2,\ldots,c-1\}$) 中采样的样本集 X_{H_n} ($n\in\{1,2,\ldots,c-1\}$), 而没有另一个组成物分布 G 的观测数据, 多组成物场景下 MPE 问题的目标是求解一系列比例系数 κ, κ_n ($n\in\{1,2,\ldots,c-1\}$). 多组成物场景下的 MPE 问题可以进一步地表示为

$$\begin{aligned} F&=\sum_{n=1}^{c-1}\kappa_n H_n+\kappa G\\ &=\kappa_i H_i+\sum_{n=1,n\neq i}^{c-1}\kappa_n H_n+\kappa G\\ &=\kappa_i H_i+(1-\kappa_i)\,M_i,\quad i\in\{1,2,\ldots,c-1\}, \end{aligned} \tag{13}$$

其中, M_i 满足

$$M_i=\frac{\kappa G+\sum_{n=1,n\neq i}^{c-1}\kappa_n H_n}{1-\kappa_i},\quad i\in\{1,2,\ldots,c-1\}. \tag{14}$$

式 (13) 和 (14) 表明, 通过将多组成物 MPE 问题的组成物进行线性组合, 一个 c ($c>2$) 组成物的 MPE 问题可以拆分为 $c-1$ 个二组成物的 MPE 问题. 也就是说, 若需要求解一个 c 组成物的 MPE 问题所对应的全部的混合比例系数, 则只需依次求解 $c-1$ 个二组成物的 MPE 问题.

4.2 不依赖不可约假设的多组成物 MPE 方法

根据 4.1 小节的结论可知, 如果要不依赖不可约假设地求解多组成物场景下的 MPE 问题, 可以将其分解为一系列二组成物的子问题, 进而基于 R-MPE 方法对各子问题求解. 因此, 可以推广得到不依赖不可约假设的多组成物 MPE 方法 MR-MPE. 在 MR-MPE 中, 首先根据式 (13) 将 c 组成物的

MPE 问题转化为 $c-1$ 个二组成物的 MPE 子问题, 并基于算法 2 求解子问题对应的比例系数, 进而得到组成物 MPE 问题对应的比例系数, 其具体流程总结如算法 3 所示.

Algorithm 3 MR-MPE

Input: c ($c>2$): the number of components; X_F: the sample i.i.d. drawn from the distribution F; X_{H_n} ($n\in\{1,2,\ldots,c-1\}$): the sample i.i.d. drawn from the distributions H_n ($n\in\{1,2,\ldots,c-1\}$), respectively; P_r: the percentage of the sample needed to copy.

Output: κ, κ_n ($n\in\{1,2,\ldots,c-1\}$): the mixture proportion in multi-component MPE problem.

1: $i\Leftarrow 1$;
2: **while** $i<c$ **do**
3: According to (13), rewrite the multi-component MPE problem as the form of $F=\kappa_i H_i+(1-\kappa_i)M_i$;
4: Take the samples X_F and X_{H_i} i.i.d. drawn from the distributions F and H_i, respectively, as two inputs in Algorithm 2;
5: Estimate the mixture proportion κ_i using Algorithm 2;
6: $i\Leftarrow i+1$;
7: **end while**
8: $\kappa\Leftarrow 1-\sum_{n=1}^{c-1}\kappa_n$.

5 基于 MR-MPE 的标签噪声学习方法

本节首先将多类别分类的标签噪声学习问题中的反向转移矩阵估计问题转化为多组成物的 MPE 问题, 并基于本文 4.2 小节提出的 MR-MPE 对反向转移矩阵进行估计, 进而构造不依赖锚点的满足统计一致性的标签噪声学习方法.

5.1 基于 MR-MPE 的反向转移矩阵估计方法

在标签噪声学习问题中, 反向转移矩阵 Q 的精确估计对于构建统计一致的标签噪声学习方法至关重要. 区别于转移矩阵 T, 反向转移矩阵 Q 表示了给定样本噪声标签情况下真实标签的条件概率 (见定义 3), 从而建立了真实标签后验概率与噪声标签后验概率之间的联系.

定义3 (反向转移矩阵) 在一个共有 c ($c>2$) 类别的多类别分类问题之中, 假设样本所对应的真实标签和噪声标签的随机变量分别为 Y 和 $\bar{Y}$, 则在该问题之中, 反向转移矩阵 Q 是一个 $c\times c$ 的方阵. 其中, 第 i 行, 第 j 列的元素 Q_{ij} 表示噪声标签 $\bar{Y}=j$ 情况下真实标签为 $Y=i$ 的条件概率, 即有 $Q_{ij}=P(Y=i|\bar{Y}=j)$.

根据本文 3.3 小节的分析可知, MPE 问题中的不可约假设和标签噪声学习问题中的锚点假设实际上是等价的. 若可在不依赖不可约假设的情况下求解 MPE 问题, 则可以构建不依赖锚点的标签噪声学习方法. 而在更为常见的多分类标签噪声学习问题之中, 其对应的反向转移矩阵估计问题与多组成物场景下的 MPE 问题联系紧密. 假设多类别分类的标签噪声学习问题一共具有 c ($c>2$) 个标签类别, 令样本对应的真实标签和噪声标签分别为 Y 和 $\bar{Y}$, 假设样本 X 对应的噪声标签 $\bar{Y}$ 仅仅依赖于真实标签 Y 而独立于样本 X [31], 则噪声标签类和真实标签类分别对应的样本密度函数 $P(X|\bar{Y})$ 和 $P(X|Y)$ 之间的关系如式 (15) 所示:

$$P(X|\bar{Y}=j)=\sum_{i=1}^{c}P(X|Y=i)P(Y=i|\bar{Y}=j). \tag{15}$$

记 $\bar{P}_i = P(X|\bar{Y}=i)$, $P_i = P(X|Y=i)$, 则式 (15) 可以表示为矩阵的形式:

$$\begin{bmatrix} \bar{P}_1 \\ \bar{P}_2 \\ \vdots \\ \bar{P}_c \end{bmatrix} = Q^{\mathrm{T}} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_c \end{bmatrix}. \tag{16}$$

为了将式 (16) 转化为多组成物 MPE 问题的形式, 可将其进行进一步的分解, 假定式 (16) 可分解为如下的形式:

$$\begin{bmatrix} \bar{P}_1 \\ \bar{P}_2 \\ \vdots \\ \bar{P}_c \end{bmatrix} = H' \begin{bmatrix} \bar{P}_1 \\ \bar{P}_2 \\ \vdots \\ \bar{P}_c \end{bmatrix} + G' \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_c \end{bmatrix}, \tag{17}$$

其中, 假定矩阵 H' 是一个 $c \times c$ 的方阵, 并且其对角线元素为 0, 而矩阵 G' 是一个 $c \times c$ 的对角方阵. 当矩阵 H' 和矩阵 G' 满足以上的形式要求时, 式 (17) 可以表达为

$$\bar{P}_i = \sum_{j=1, j\neq i}^{c} H'_{ij}\bar{P}_j + G'_{ii}P_i,\ i \in \{1, 2, \ldots, c\}, \tag{18}$$

这与多组成物场景下的 MPE 问题形式完全一致. 因此, 如果以上的分解是存在的, 则式 (16) 可以转化为一系列多组成物的 MPE 问题.

以下将证明, 当矩阵 Q^{T} 可逆时, 形如式 (17) 的分解是存在且唯一的.

证明 假定式 (17) 的分解是存在的, 则式 (17) 可进一步地表示为

$$\begin{bmatrix} \bar{P}_1 \\ \bar{P}_2 \\ \vdots \\ \bar{P}_c \end{bmatrix} = (I - H')^{-1} G' \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_c \end{bmatrix}, \tag{19}$$

其中, I 表示单位矩阵. 对比式 (16) 和 (19), 矩阵 Q^{T} 应当满足如式 (20) 所示关系:

$$Q^{\mathrm{T}} = (I - H')^{-1} G'. \tag{20}$$

式 (20) 可进一步地表示为

$$Q^{\mathrm{T}^{-1}} = G'^{-1}(I - H'). \tag{21}$$

由于矩阵 H' 是对角元素为 0 的 $c \times c$ 的方阵, 而矩阵 G' 是一个 $c \times c$ 的对角方阵. 假定矩阵 H' 和 G' 的第 i 行 j 列元素分别为 H'_{ij} 和 G'_{ij}, 记矩阵 Q^{T} 的逆矩阵为 Q^I, 其第 i 行 j 列元素为 Q^I_{ij}, 则其满足关系:

$$Q^I_{ii} = \frac{1}{G'_{ii}},\ i \in \{1, 2, \ldots, c\}, \tag{22}$$

$$Q^I_{ij} = \frac{-H'_{ij}}{G'_{ii}},\ i, j \in \{1, 2, \ldots, c\},\ i \neq j. \tag{23}$$

当矩阵 Q^{T} 可逆时, 按照式 (22) 和 (23) 可分解得到矩阵 H' 和 G', 因此式 (17) 的分解是存在的, 以下将证明式 (17) 分解的唯一性, 假定式 (17) 存在另一种分解满足

$$\begin{bmatrix}\bar{P}_1\\ \bar{P}_2\\ \vdots\\ \bar{P}_c\end{bmatrix} = A\begin{bmatrix}\bar{P}_1\\ \bar{P}_2\\ \vdots\\ \bar{P}_c\end{bmatrix} + B\begin{bmatrix}P_1\\ P_2\\ \vdots\\ P_c\end{bmatrix}, \tag{24}$$

其中, 矩阵 $A = H'$ 和矩阵 $B = G'$ 的关系不可同时成立. 根据式 (21) 可得到关系 $(G'^{-1} - B^{-1}) = G'^{-1}H' - B^{-1}A$, 由于矩阵 G' 和矩阵 B 均是对角矩阵而矩阵 H' 和矩阵 A 是对角线元素为 0 的矩阵, 因此满足关系 $G'^{-1}H' = B^{-1}A = 0$. 这表明 $G'^{-1} - B^{-1} = 0$, 即有 $B = G'$, 根据式 (21), 可以进一步得到关系 $A = H'$, 这与前提假设相矛盾. 因此, 如式 (17) 所示的矩阵分解是存在并且唯一的.

根据定义 3 中矩阵 Q 的物理含义, 可以得到如下的关系:

$$\sum_{i=1}^{c} Q_{ij} = 1. \tag{25}$$

上述公式表明矩阵 Q 是列和为 1 的矩阵, 因此矩阵 Q 的转置矩阵 Q^{T} 是行和为 1 的矩阵, 故而矩阵 Q^{T} 的逆矩阵 Q^{I} 也是行和为 1 的矩阵, 于是可以得到如下的约束关系:

$$G'_{ii} + \sum_{j=1}^{c} H'_{ij} = 1,\ i \in \{1, 2, \ldots, c\}. \tag{26}$$

因此, 式 (18) 所述的 c 个等式等价于 c 个多组成物场景下的 MPE 问题. 其中, $\bar{P}_i$ $(i \in \{1, 2, \ldots, c\})$ 表示噪声类条件概率分布, 其可以通过标签噪声样本估计求解, 因此在该问题中是已知的. 根据 5.1 小节的讨论, c 个 c 组成物的 MPE 问题等价于 $c \times (c-1)$ 个二组成物场景下的 MPE 问题. 式 (18) 进一步表明, 矩阵 H' 的非对角元素 H'_{ij} $(i \neq j)$ 和矩阵 G' 的对角元素 G'_{ii} 正是每一个多组成物场景下 MPE 问题待求解的比例系数. 利用第 4 节提出的 MR-MPE 方法 (如算法 3 所示) 可以依次求解矩阵 H' 的各个元素, 进而根据式 (26) 求解矩阵 G', 并且最终根据式 (20) 的方法估计得到反向转移矩阵 Q. 综上所述, 可以得到基于 MR-MPE 的反向转移矩阵估计方法, 该方法具体流程总结如算法 4 所示.

Algorithm 4 MR-MPE-based inverse transition matrix estimation method

Input: c: the number of components; X_i $(i \in \{1, 2, \ldots, c\})$: the sample i.i.d. drawn from the noise label class $\bar{Y} = i$ $(i \in \{1, 2, \ldots, c\})$; P_r: the percentage of the sample needed to copy.

Output: Q: inverse transition matrix.

1: $i \Leftarrow 1$;
2: **while** $i \leqslant c$ **do**
3: Take the sample X_i i.i.d. drawn from the component distribution $\bar{P}_i$ as input X_F in Algorithm 3;
4: Take samples X_j $(j \in \{1, 2, \ldots, c\}, j \neq i)$ i.i.d. drawn from the component distribution $\bar{P}_j$ $(j \in \{1, 2, \ldots, c\}, j \neq i)$ as inputs X_{H_k} in Algorithm 3 that sampled from distributions H_k $(k \in \{1, 2, \ldots, c-1\})$, respectively;
5: Estimate the H'_{ij} and G'_{ii} by employing Algorithm 3;
6: $i \Leftarrow i + 1$;
7: **end while**
8: $Q \Leftarrow [(I - H')^{-1} G']^{\mathrm{T}}$.

5.2 不依赖锚点的标签噪声学习方法

由于反向转移矩阵表示噪声标签到真实标签的条件转换概率, 在反向转移矩阵已知的情况下, 反向转移矩阵可以将噪声标签后验概率预测转化为真实标签后验概率预测. 由于仅有标签噪声样本, 若采用标签噪声样本作为监督信息来指导分类网络的训练, 可以得到噪声标签后验概率的预测. 假定待分类的目标类别个数为 c ($c > 2$), 则对于任意的样本 X, 其噪声标签后验概率 $P(\bar{Y}|X) = [P(\bar{Y} = 1|X), \ldots, P(\bar{Y} = c|X)]^{\mathrm{T}} = [\bar{P}'_1, \ldots, \bar{P}'_c]^{\mathrm{T}}$ 与该样本真实标签后验概率 $P(Y|X) = [P(Y = 1|X), \ldots, P(Y = c|X)]^{\mathrm{T}} = [P'_1, \ldots, P'_c]^{\mathrm{T}}$ 具有转换关系, 如式 (27) 所示:

$$\begin{bmatrix} P'_1 \\ P'_2 \\ \vdots \\ P'_c \end{bmatrix} = \begin{bmatrix} P(Y=1|\bar{Y}=1) & \cdots & P(Y=1|\bar{Y}=c) \\ \vdots & & \vdots \\ P(Y=c|\bar{Y}=1) & \cdots & P(Y=c|\bar{Y}=c) \end{bmatrix} \begin{bmatrix} \bar{P}'_1 \\ \bar{P}'_2 \\ \vdots \\ \bar{P}'_c \end{bmatrix}, \tag{27}$$

其矩阵形式可表示为

$$P(Y|X) = QP(\bar{Y}|X), \tag{28}$$

其中, 反向转移矩阵 Q 满足关系 $Q_{ij} = P(Y = i|\bar{Y} = j)$ ($i, j \in \{1, 2, \ldots, c\}$). 式 (28) 表明, 反向转移矩阵 Q 可以将噪声标签后验概率预测转换为真实标签后验概率预测. 由于 $P(\bar{Y}|X)$ 可以通过标签噪声数据计算, 在反向转移矩阵 Q 估计良好的情况下, 真实标签后验概率 $P(Y|X)$ 可以精准估计, 从而预测样本的真实标签. 基于 MR-MPE 方法的标签噪声学习方法主要可以分为两个阶段: (1) 噪声标签预测阶段: 将标签噪声数据作为监督信息训练分类网络, 用于预测样本的噪声标签 $P(\bar{Y}|X)$; (2) 真实标签预测阶段: 首先利用噪声标签数据估计反向转移矩阵, 其次, 固定第一阶段的网络参数并将反向转移矩阵 (由算法 4 得到) 作为原始网络 softmax 之后的线性转移层, 在线性转移层后输出真实标签预测 $P(Y|X)$. MR-MPE 的整体框架流程如图 1 所示.

6 实验结果与分析

本节将对 MR-MPE 进行实验测试. 具体地, 本节分别在合成噪声数据集和真实噪声数据上进行对比实验来验证本文方法的优越性; 同时, 本节对比验证了本文方法在锚点不存在情况下的鲁棒性.

6.1 实验设置

6.1.1 数据集

本节将分别在合成噪声数据集 CIFAR-10 和 CIFAR-100 以及真实噪声数据集 CIFAR-10N, CIFAR-100N 和 Yorúbá 上进行实验测试. CIFAR-10 数据集是一个用于识别普适物体的小型图像数据集, 一共包含 10 个类别的 RGB 彩色图片: 飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车. 该数据集包含了 50000 张训练样本和 10000 张测试样本, 图片均是 3 通道, 尺寸为 32×32. CIFAR-100 数据集是由同样规格的彩色图像组成的, 但该数据集共有 100 类, 每个类有 500 张图片作为训练样本, 100 张图片作为测试样本. 本节利用标准数据集 CIFAR-10 和 CIFAR-100 来人工生成噪声标签, 用于模拟真实的噪声标签数据. 具体地, 本节分别采用对称的均匀噪声生成方式和非对称的配对噪声生成方式来合成样本对应的噪声标签. 在均匀噪声的场景中, 每一类样本将按照 $1 - \eta$ (η 为噪声率) 的概率保持

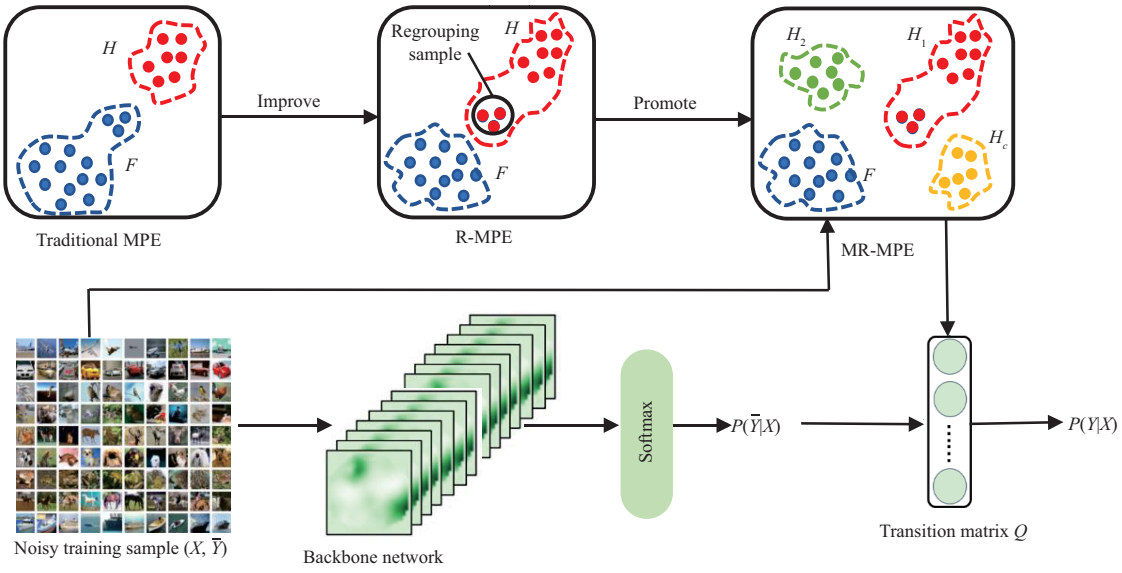

图 1 (网络版彩图) MR-MPE 算法框架图

Figure 1 (Color online) Overall framework of MR-MPE

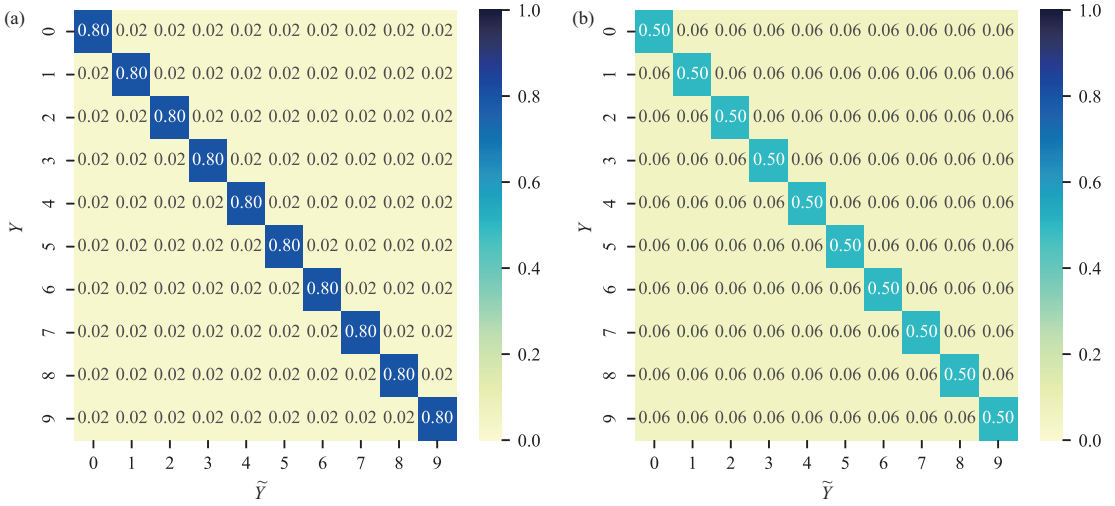

图 2 (网络版彩图) (a) CIFAR-10 20% 和 (b) CIFAR-10 50% 均匀噪声转移矩阵

Figure 2 (Color online) Transition matrix on CIFAR-10 dataset with 20% (a) and 50% (b) symmetric noise

其真实标签, 并按照 $(1-\eta)/(c-1)$ (c 为总类别数) 的概率均匀地转移为其他类的噪声标签. 而在配对噪声的场景中, 每一类样本仍将以 $1-\eta$ 的概率保持其真实标签, 并以 η 的概率转移为某一特定类别的噪声标签. 本节分别对 CIFAR-10 和 CIFAR-100 数据集按照 20% 和 50% 的噪声率来生成均匀噪声, 按照 20% 和 40% 的噪声率来生成配对噪声. 以 CIFAR-10 数据集为例, 当噪声率为 20% 和 50% 的均匀噪声时, 对应的转移矩阵的情况分别如图 2(a) 和 (b) 所示, 当噪声率为 20% 和 40% 的配对噪声时, 对应的转移矩阵的情况分别如图 3(a) 和 (b) 所示, 其中纵轴为真实标签, 而横轴为噪声标签, 图中数值表示真实标签到噪声标签的转移概率.

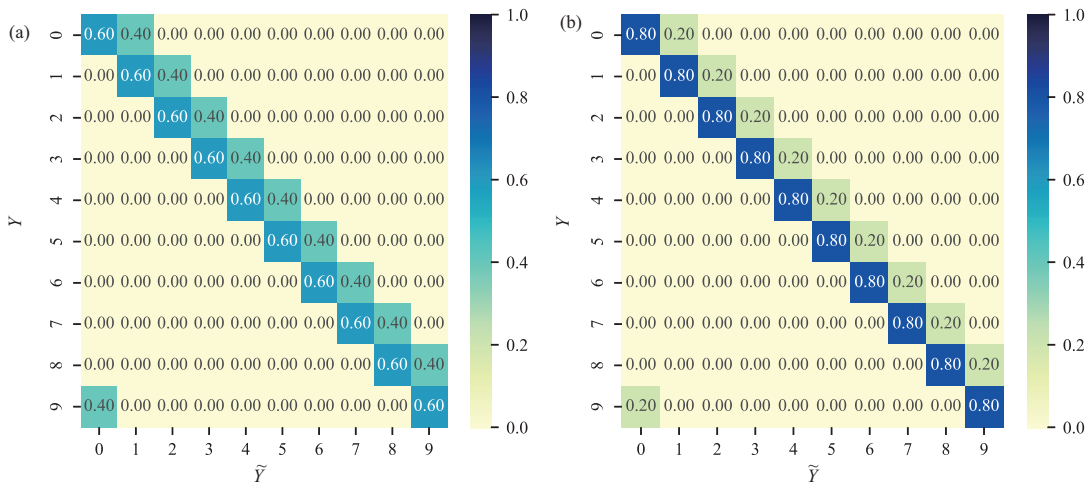

图 3 (网络版彩图) (a) CIFAR-10 20% 和 (b) CIFAR-10 40% 配对噪声转移矩阵

Figure 3 (Color online) Transition matrix on CIFAR-10 dataset with 20% (a) and 40% (b) pair-flipped noise

真实噪声数据集 CIFAR-10N 和 CIFAR-100N 分别是 CIFAR-10 和 CIFAR-100 数据集的人工标注版本, 二者均由加利福尼亚大学圣克鲁斯分校 (University of California, Santa Cruz, UCSC) 在亚马逊的 Amazon Mechanical Turk 众包平台收集 [35], 其样本分别与 CIFAR-10 和 CIFAR-100 数据集完全一致. 不同地, 二者对应的标签均来自人工标注, 其中不可避免地含有噪声标签. 本实验将合成噪声数据集 CIFAR-10 和 CIFAR-100 以及真实噪声数据集 CIFAR-10N 和 CIFAR-100N 对应的 50000 张训练样本按照 9:1 的比例划分为训练集和验证集, 将 10000 张测试样本作为测试集, 用于后续实验评估. 同时, 为了验证本文算法在不同类型数据集上均有良好效果, 本文选取文本数据集 Yorúbá 进行进一步实验. Yorúbá 是第三大非洲土著语言, Yorúbá 数据集包含了 1340 个训练样本、189 个验证样本和 379 个测试样本, 该数据集由 BBC 新闻收集, 其样本为 Yorúbá 语言的文本段, 标签为文本对应的主题. Yorúbá 数据标签一共包含了 7 个类别, 其训练样本和验证样本标签是含有噪声的而其测试样本标签则是人工标注的. 以上 3 个数据集反映了现实场景中的标签噪声的分布情况, 可用于检验标签学习算法在实际场景中的实用性.

6.1.2 对比方法

本节的实验将与现有的主流的标签噪声学习算法进行对比. 具体地, 在合成噪声数据集和真实噪声数据集上, 本节分别选取启发式算法: JoCoR [22], Co-teaching [20], Co-teaching+ [21] 和统计一致性算法: T-Revision [32], Forward [30], L-DMI [50], CCR [44] 作为对比实验方法. 同时, 为了验证本文提出算法不依赖锚点的特性, 本节将在移除锚点的情况下测试算法的性能. 本节选取不直接依赖锚点的标签噪声学习方法: VolMinNet [34], T-Revision [32], HOC [33] 作为移除锚点实验中的对比方法, 与本文提出的 MR-MPE 方法进行比较, 用于验证本文提出算法的优越性和鲁棒性.

6.1.3 实现细节

本节在 CIFAR-10 和 CIFAR-10N 数据集上使用 ResNet-18 作为 backbone 网络, 在 CIFAR-100 和 CIFAR-100N 数据集上使用 ResNet-34 作为 backbone 网络, 并且全部实验使用 Adam [51] 优化器. 在 Yorúbá 数据集上, 本节使用 BERT [52] 文本预训练网络作为 backbone 网络, 并使用 AdamW 优化器.

表 1 不同算法在均匀噪声数据集上分类精度 (%)

Table 1 Classification accuracy (%) of different algorithms on symmetric noisy datasets[a)]

Method	CIFAR-10		CIFAR-100	
	Sym-20%	Sym-50%	Sym-20%	Sym-50%
T-Revision [32]	86.55 ± 0.47	80.35 ± 1.55	59.67 ± 0.54	49.14 ± 0.45
Forward [30]	88.17 ± 0.78	80.34 ± 0.43	60.24 ± 0.85	38.68 ± 1.14
JoCoR [22]	84.36 ± 0.61	79.51 ± 0.16	52.60 ± 0.93	43.15 ± 0.35
Co-teaching [20]	88.99 ± 0.25	83.13 ± 0.55	44.71 ± 2.26	32.69 ± 0.61
Co-teaching+ [21]	87.87 ± 0.65	83.43 ± 1.89	47.30 ± 0.96	30.47 ± 1.01
L-DMI [50]	87.34 ± 0.58	79.54 ± 0.53	55.99 ± 0.68	34.53 ± 1.63
CCR [44]	89.85 ± 0.52	84.03 ± 0.21	$\mathbf{67.15 \pm 0.42}$	51.83 ± 0.21
MR-MPE	$\mathbf{90.25 \pm 0.26}$	$\mathbf{84.12 \pm 0.27}$	62.32 ± 0.26	$\mathbf{51.98 \pm 0.23}$

a) The superior results are emphasized in bold.

在 CIFAR-10, CIFAR-10N 数据集上, 本节设置初始学习率为 3.85E−4, 权重衰减系数为 4E−5, 批处理样本个数为 128. 同时设置算法训练次数 epoch 为 100, 并且在第 20, 35, 70 和 85 个 epoch 结束时下调学习率至当下学习率的 1/100. 在 CIFAR-100 和 CIFAR-100N 数据集上, 本节设置初始学习率为 4E−4, 权重衰减系数为 4E−5, 批处理样本个数为 128, 算法训练次数 epoch 为 80, 在训练过程中, 分别在第 25 和 35 个 epoch 结束时下调学习率至当下学习率的 1/10. 而在文本分类的 Yorúbá 数据集上, 本节设置初始学习率为 3E−5, 批处理样本个数为 32, 算法训练次数 epoch 为 15.

6.2 分类精度评价

6.2.1 合成噪声数据集上的实验结果

本小节分别在均匀噪声和配对噪声的条件下进行对比实验, 并且根据 6.1.3 小节中的参数设置进行训练, 在 5 次重复实验的情况下, 本文所提 MR-MPE 算法与各个对比方法在均匀噪声条件下和配对噪声条件下所对应的平均最佳分类精度和标准差分别总结如表 1 和 2 所示. 通过对比实验可知, 相比现有的标签噪声学习方法, 在均匀的噪声条件下, 本文提出的 MR-MPE 在 CIFAR-10 和 CIFAR-100 (50% 噪声率) 的数据集上均表现出最优越的性能. 而在 CIFAR-100 (20% 噪声率) 的数据集上, MR-MPE 则取得了近似最优的分类结果. 而在配对噪声的条件下, 本文提出的 MR-MPE 在 CIFAR-10 和 CIFAR-100 (20% 噪声率) 的数据集上均表现出最优越的性能. 而在 CIFAR-100 (40% 噪声率) 的数据集上, MR-MPE 则取得了近似最优的分类结果. MR-MPR 在 CIFAR-100 噪声数据集上性能有所下降, 这是由于该数据集的分类类别较多, 待估计的转移矩阵足有 10000 个参数, 需要解决的 MPE 问题个数大幅上升造成了估计误差的积累从而导致了转移矩阵估计精度的下降, 这在一定程度上影响了 MR-MPE 方法的性能; 而表现较好的 CCR 方法由于引入了额外的正则化对转移矩阵的形式进行约束, 从而在 CIFAR-100 噪声数据集上仍保持了较好的性能.

6.2.2 真实噪声数据集上的实验结果

为进一步验证本文提出的 MR-MPE 方法的优越性, 本小节在真实的标签噪声数据集 CIFAR-10N, CIFAR-100N 和 Yorúbá 上进行对比实验来测试 MR-MPE 在真实噪声情况下的性能. 在 5 次重复实验的基础上, 各算法在真实噪声数据集上对应的平均分类精度与标准差如表 3 所示. 实验结果表明, 在 3 个真实噪声数据集上, 本文提出的 MR-MPE 算法均取得了最优的分类精度. 真实噪声数据集上的对

表 2　不同算法在配对噪声数据集上分类精度 (%)

Table 2　Classification accuracy (%) of different algorithms on pair-flipped noisy datasets[a)]

Method	CIFAR-10		CIFAR-100	
	Pair-20%	Pair-40%	Pair-20%	Pair-40%
T-Revision [32]	89.41 ± 0.68	84.71 ± 1.28	58.68 ± 1.07	51.13 ± 0.57
Forward [30]	88.49 ± 1.12	85.31 ± 0.75	58.91 ± 1.31	48.12 ± 0.87
JoCoR [22]	84.69 ± 0.43	65.69 ± 0.39	53.39 ± 0.57	35.25 ± 0.79
Co-teaching [20]	85.09 ± 0.91	81.39 ± 0.71	42.16 ± 0.32	31.63 ± 0.76
Co-teaching+ [21]	89.09 ± 0.61	84.49 ± 0.46	47.49 ± 0.41	35.58 ± 0.88
L-DMI [50]	88.65 ± 0.83	86.21 ± 1.13	52.42 ± 1.73	42.31 ± 0.93
CCR [44]	90.08 ± 0.41	87.18 ± 0.15	66.11 ± 0.31	**64.27 ± 0.24**
MR-MPE	**90.57 ± 0.18**	**87.43 ± 0.31**	**66.43 ± 0.39**	60.09 ± 0.89

a) The superior results are emphasized in bold.

表 3　不同算法在真实噪声数据集上分类精度 (%)

Table 3　Classification accuracy (%) of different algorithms on real-world noisy datasets[a)]

Method	CIFAR-10N	CIFAR-100N	Yorúbá
T-Revision [32]	85.44 ± 0.52	50.84 ± 0.77	60.23 ± 0.63
Forward [30]	88.35 ± 0.21	57.02 ± 0.14	66.18 ± 0.81
JoCoR [22]	84.71 ± 0.39	45.41 ± 0.19	59.15 ± 0.41
Co-teaching [20]	88.43 ± 0.07	50.32 ± 0.79	61.29 ± 1.47
Co-teaching+ [21]	88.53 ± 0.09	52.33 ± 0.49	61.98 ± 1.22
L-DMI [50]	88.34 ± 0.17	53.45 ± 0.26	63.76 ± 0.42
VolMinNet [34]	87.68 ± 0.21	54.41 ± 0.67	67.01 ± 0.51
HOC [33]	86.27 ± 1.53	55.42 ± 0.83	66.36 ± 0.73
MR-MPE	**88.61 ± 0.04**	**57.26 ± 0.15**	**67.98 ± 0.37**

a) The superior results are emphasized in bold.

比实验进一步地证明了本文提出的方法在现实中的实用性和泛化能力.

6.2.3　移除锚点的鲁棒性实验

相比现有的标签噪声学习方法, 本文提出的 MR-MPE 方法的主要优势在于不依赖锚点. 目前, 不直接依赖锚点的统计一致性算法主要包括了 VolMinNet [34], T-Revision [32], HOC [33], 而这些方法却直接或间接地使用了近似锚点或额外假设, 并没有真正地解决锚点依赖的问题.

为验证在锚点不存在情况下本文提出方法的优越性, 本小节将在移除锚点的情况下对算法的鲁棒性进行测试. 具体地, 本小节首先利用真实标签样本作为监督信息来训练分类网络, 其次对样本的真实标签进行预测, 并根据设定的阈值剔除训练集和验证集中最大预测后验概率大于阈值的样本点 (近似锚点). 本文分别在噪声率为 20% 的 CIFAR-10 合成噪声数据集和 CIFAR-10N 真实噪声数据集上进行移除锚点的实验测试. 本文设置阈值为 1, 0.95, 0.9, 0.8, 0.7, 0.6 和 0.5, 随着阈值的下降, 移除的锚点或近似锚点越多. 测试 MR-MPE 和对比方法在不同阈值下的准确率, 对应的实验结果如图 4(a) 和 (b) 所示.

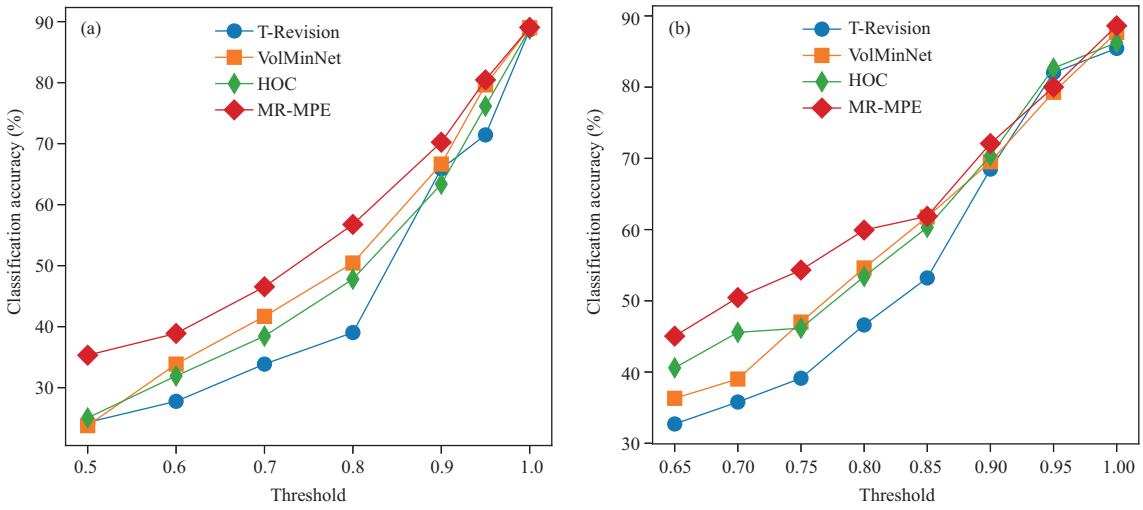

图 4 (网络版彩图) 不同阈值条件下不同算法在 CIFAR-10 20% 噪声率 (a) 和 CIFAR-10N 数据集 (b) 上分类精度

Figure 4 (Color online) Classification accuracy of different algorithms with different thresholds on CIFAR-10 dataset with 20% noise rate (a) and CIFAR-10N dataset (b)

根据上述实验结果可知, 本文提出的方法在移除锚点的情况下整体相比对比方法具有更好的性能表现. 具体地, 相比于现有的不依赖锚点的标签噪声学习算法, 随着移除锚点比例的上升, 本文提出的方法精度下降更少, 受锚点的影响较小. 这表明了, 本文提出的方法在锚点较少或者不存在的情况下, 依然维持了更好的性能, 从而验证了本文提出的算法不依赖锚点的特性, 进一步地显示了 MR-MPE 算法的鲁棒性.

7 结论与展望

本文对现有标签噪声学习方法存在的锚点依赖问题进行了改进, 将多类别分类的标签噪声学习问题转化为了多组成物的 MPE 问题, 并利用本文所提的 MR-MPE 方法估计反向转移矩阵, 进而构造了不依赖锚点的统计一致性算法. 同时, 本文在多个数据集上进行了对比试验, 验证了 MR-MPE 方法的优越性和鲁棒性. 然而, 本文所讨论的多类别分类的标签噪声学习问题仍然限于噪声率较小的情况 (噪声率 $< 50\%$), 同时, 本文所考虑的均匀的噪声生成方式所产生的噪声分布仍是较为简单的噪声分布情况, 而在现实场景中, 实际噪声的分布情况更加复杂. 未来的工作将进一步探索更极端噪声率情况下的鲁棒算法, 研究更复杂的噪声分布特征, 进一步地提升 MR-MPE 方法的实用性.

参考文献

1 Mack C A. Fifty years of Moore's law. IEEE Trans Semicond Manufact, 2011, 24: 202–207
2 Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: a brief review. Comput Intelligence Neurosci, 2018, 2018: 1–13
3 Ranjan N, Mundada K, Phaltane K, et al. A survey on techniques in NLP. Int J Comput Appl, 2016, 134: 6–9
4 Qin C, Zhu H S, Zhuang F Z, et al. A survey on knowledge graph-based recommender systems. Sci Sin Inform, 2020, 50: 937–956 [秦川, 祝恒书, 庄福振, 等. 基于知识图谱的推荐系统研究综述. 中国科学: 信息科学, 2020, 50: 937–956]
5 Lu J, Wu D, Mao M, et al. Recommender system application developments: a survey. Decision Support Syst, 2015, 74: 12–32

6 Chen C J, Jiang L, Lei N, et al. An interactive feature selection method based on learning-from-crowds. Sci Sin Inform, 2020, 50: 794–812 [陈长建, 姜流, 雷娜, 等. 基于众包学习的交互式特征选择方法. 中国科学: 信息科学, 2020, 50: 794–812]

7 Tong Y, Zhou Z, Zeng Y, et al. Spatial crowdsourcing: a survey. VLDB J, 2020, 29: 217–250

8 Behrend T S, Sharek D J, Meade A W, et al. The viability of crowdsourcing for survey research. Behav Res, 2011, 43: 800–813

9 Razzaq A, Yang X. Digital finance and green growth in China: appraising inclusive digital finance using web crawler technology and big data. Tech Forecasting Soc Change, 2023, 188: 122262

10 Zhang C, Bengio S, Hardt M, et al. Understanding deep learning (still) requires rethinking generalization. Commun ACM, 2021, 64: 107–115

11 Gao W, Zhang T, Yang B B, et al. On the noise estimation statistics. Artif Intelligence, 2021, 293: 103451

12 Karlik B, Olgac A V. Performance analysis of various activation functions in generalized MLP architectures of neural networks. Int J Artif Intell Expert Syst, 2011, 1: 111–122

13 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324

14 Sengupta A, Ye Y, Wang R, et al. Going deeper in spiking neural networks: VGG and residual architectures. Front Neurosci, 2019, 13: 95

15 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 1–9

16 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778

17 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020. ArXiv:2010.11929

18 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 10012–10022

19 Malach E, Shalev-Shwartz S. Decoupling "when to update" from "how to update". In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), 2017. 961–971

20 Han B, Yao Q, Yu X, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018. 8536–8546

21 Yu X, Han B, Yao J, et al. How does disagreement help generalization against label corruption? In: Proceedings of International Conference on Machine Learning (ICML), 2019. 7164–7173

22 Wei H, Feng L, Chen X, et al. Combating noisy labels by agreement: a joint training method with co-regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 13726–13735

23 Yao Q, Yang H, Han B, et al. Searching to exploit memorization effect in learning with noisy labels. In: Proceedings of International Conference on Machine Learning (ICML), 2020. 10789–10798

24 Tanaka D, Ikami D, Yamasaki T, et al. Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 5552–5560

25 Liu T, Tao D. Classification with noisy labels by importance reweighting. IEEE Trans Pattern Anal Mach Intell, 2015, 38: 447–461

26 Northcutt C G, Wu T, Chuang I L. Learning with confident examples: rank pruning for robust classification with noisy labels. 2017. ArXiv:1705.01936

27 Natarajan N, Dhillon I S, Ravikumar P, et al. Learning with noisy labels. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS), 2013. 1196–1204

28 Zhang Z, Sabuncu M R. Generalized cross-entropy loss for training deep neural networks with noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018. 8792–8802

29 Thekumparampil K K, Khetan A, Lin Z, et al. Robustness of conditional GANs to noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018. 10292–10303

30 Patrini G, Rozza A, Krishna Menon A, et al. Making deep neural networks robust to label noise: a loss correction approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1944–1952
31 Yu X, Liu T, Gong M, et al. Learning with biased complementary labels. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 68–83
32 Xia X, Liu T, Wang N, et al. Are anchor points really indispensable in label-noise learning? In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), 2019. 6838–6849
33 Zhu Z, Song Y, Liu Y. Clusterability as an alternative to anchor points when learning with noisy labels. In: Proceedings of International Conference on Machine Learning (ICML), 2021. 12912–12923
34 Li X, Liu T, Han B, et al. Provably end-to-end label-noise learning without anchor point. In: Proceedings of International Conference on Machine Learning (ICML), 2021. 6403–6413
35 Wei J, Zhu Z, Cheng H, et al. Learning with noisy labels revisited: a study using real-world human annotations. 2021. ArXiv:2110.12088
36 Zhu D, Hedderich M A, Zhai F, et al. Is BERT robust to label noise? A study on learning with noisy labels in text classification. In: Proceedings of the 3rd Workshop on Insights from Negative Results in NLP, 2022. 62–67
37 Jiang L, Zhou Z, Leung T, et al. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: Proceedings of International Conference on Machine Learning (ICML), 2018. 2304–2313
38 Bengio Y, Louradour J, Collobert R, et al. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML) 2009. 41–48
39 Tsotsos J K, Luo J. Probing the effect of selection bias on generalization: a thought experiment. 2021. ArXiv:2105.09934
40 Wei Q, Sun H, Lu X, et al. Self-filtering: a noise-aware sample selection for label noise with confidence penalization. In: Proceedings of the European Conference on Computer Vision (ECCV), 2022. 516–532
41 Kremer J, Sha F, Igel C. Robust active label correction. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2018. 308–316
42 Yu X, Liu T, Gong M, et al. An efficient and provable approach for mixture proportion estimation using linear independence assumption. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 4480–4489
43 Vandermeulen R A, Scott C D. An operator theoretic approach to nonparametric mixture models. Ann Statist, 2019, 47: 2704–2733
44 Cheng D, Ning Y, Wang N, et al. Class-dependent label-noise learning with cycle-consistency regularization. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2022. 35: 11104–11116
45 Yang Z C, Liu F, Gorshkov A V, et al. Hilbert-space fragmentation from strict confinement. Phys Rev Lett, 2020, 124: 207602
46 Yao Y, Liu T, Han B, et al. Towards mixture proportion estimation without irreducibility. 2020. ArXiv:2002.03673
47 Blanchard G, Lee G, Scott C. Semi-supervised novelty detection. J Machine Learning Res, 2010, 11: 2973–3009
48 Ramaswamy H, Scott C, Tewari A. Mixture proportion estimation via kernel embeddings of distributions. In: Proceedings of International Conference on Machine Learning (ICML), 2016. 2052–2060
49 Scott C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), 2015. 838–846
50 Xu Y, Cao P, Kong Y, et al. $\mathcal{L}_{\mathrm{DMI}}$: a novel information-theoretic loss function for training deep nets robust to label noise. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), 2019. 6225–6236
51 Kingma D P, Ba J. Adam: a method for stochastic optimization. 2014. ArXiv:1412.6980
52 Kenton J D M W C, Toutanova L K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186

Label-noise learning via mixture proportion estimation

Qinghua ZHENG[1,3], Shuzhi CAO[1,3], Jianfei RUAN[1,3*], Rui ZHAO[1,3] & Bo DONG[2,4*]

1. *School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049*, China*;
2. *School of Continuing Education, Xi'an Jiaotong University, Xi'an* 710049*, China*;
3. *Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an* 710049*, China*;
4. *Shaanxi Province Key Lab of Satellite and Terrestrial Network Technology Research and Development, Xi'an* 710049*, China*
* Corresponding author. E-mail: jianfei.ruan@hotmail.com, dong.bo@xjtu.edu.cn

Abstract With the rise of artificial intelligence in recent years, along with the improvement of hardware computing power, deep learning has emerged as the new paradigm for artificial intelligence algorithms. In realistic multi-class classification scenarios, deep learning relies heavily on the availability of massive manually labeled data; the limitations of labeling costs and privacy protections, however, often make it difficult to obtain adequate amounts of appropriately labeled data for deep learning. Recently, crowdsourcing and web crawling have provided an easy way to collect large amounts of labeled data, but they are limited by the inevitable introduction of label noise. As deep neural networks have a high capacity to fit noisy labels, it is challenging to train deep networks robustly with noisy labels. For robust learning, existing works commonly rely explicitly or implicitly on a given set of anchor points, i.e., instances that almost certainly belong to the true classes. Unfortunately, anchor points are difficult to obtain in practice, which makes these works fragile. To address this problem, in this paper, we build an anchor-free statistically consistent algorithm in the presence of label noise by creatively transforming the multi-class label-noise learning problem into a mixture proportion estimation (MPE) problem. This paper makes the following contributions: (i) we for the first time generalize the existing Regrouping-MPE (R-MPE) method that is only suitable for two-component scenarios, and propose a multi-component oriented R-MPE (MR-MPE) method without relying on the common irreducible assumption; and (ii) from a theoretical perspective, we demonstrate that the anchor point hypothesis for label-noise learning is equivalent to the irreducible hypothesis for MPE problems in the context of multi-class classification. Therefore, an anchor-free statistically consistent label-noise learning algorithm is subsequently constructed based on the proposed MR-MPE method. In this paper, comparative experiments with existing algorithms are conducted on both synthetic noisy datasets and real-world noisy datasets. The results demonstrate that the proposed algorithm performs most effectively on multiple datasets. Additionally, the robustness of the proposed algorithm is verified when anchor points are removed.

Keywords mixture proportion estimation, multi-class classification, label-noise learning, anchor point, irreducible assumption, statistical consistency