中国科学: 信息科学 2024 年 第 54 卷 第 1 期: 144–158
SCIENTIA SINICA Informationis
论文

面向标签噪声的联合训练框架

魏琦[1], 孙皓亮[1*], 马玉玲[2], 尹义龙[1*]

1. 山东大学软件学院, 济南 250101
2. 山东建筑大学计算机科学与技术学院, 济南 250101
* 通信作者. E-mail: haolsun@sdu.edu.cn, ylyin@sdu.edu.cn

收稿日期: 2022–10–13; 修回日期: 2023–03–23; 接受日期: 2023–04–06; 网络出版日期: 2024–01–15

国家自然科学基金 (批准号: 62106129, 62176139, 62177031)、山东省自然科学基金 (批准号: ZR2021QF053, ZR2021ZD15) 和中国博士后科学基金 (批准号: 2021TQ0195, 2021M701984) 资助项目

摘要 当前面向标签噪声的鲁棒性学习通常依赖样本选择和标签修正两种策略, 但是这两类方法均存在缺陷. 基于样本选择的方法忽略了被过滤掉的样本中的有效信息, 进而降低了模型的性能. 基于标签修正的方法常使用自标签技术而引起模型的错误积累问题. 对此, 本文提出了一个集成样本选择、标签修正的联合训练框架. 针对样本选择模块, 本文设计了一种新的选择标准, 通过在线选择的方法对所挑选的样本集合进行更新. 相较于现有选择标准, 本文提出的标准可保留更多边界样本, 提升了模型对决策边界的学习性能, 增强了模型的泛化性能. 针对标签修正模块, 本文提出了一种联合标签修正策略. 相比于传统的自标签修正技术, 该模块通过联合特征空间视角, 对噪声样本进行多视角的标签修正, 解决了传统自标签技术的错误累积问题. 此外, 本文引入对比学习正则化项, 提升了标签修正效果和模型表征学习能力. 本文方法在 4 个测试基准上取得了当前最好分类效果, 验证了所提训练框架的有效性.

关键词 标签噪声学习, 样本选择, 标签修正, 对比学习

1 引言

标签噪声学习 (learning with noisy labels) [1~5] 作为机器学习研究中典型的弱监督学习问题, 一直以来备受关注. 当前深度学习中模型的良好泛化性能依赖于大规模且标注精确的训练数据 [6]. 但是在实际应用中, 精细地标注数据将带来显著的资源消耗. 此外, 在某些特定的专业领域 (如医疗 [7]、地理卫星图像分析 [8]), 标注的质量还受到人为主观因素的影响. 因此, 这些因素导致了训练数据中不可避免地引入噪声标签. 针对训练数据中带有噪声标签的任务, 标签噪声学习的目标是降低噪声标签对训练过程的影响, 从而保证模型的泛化性.

引用格式: 魏琦, 孙皓亮, 马玉玲, 等. 面向标签噪声的联合训练框架. 中国科学: 信息科学, 2024, 54: 144–158, doi: 10.1360/SSI-2022-0395
Wei Q, Sun H L, Ma Y L, et al. A joint training framework for learning with noisy labels (in Chinese). Sci Sin Inform, 2024, 54: 144–158, doi: 10.1360/SSI-2022-0395

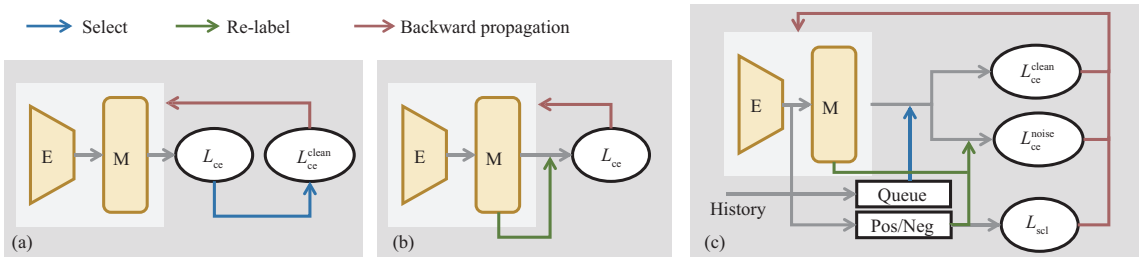

图 1 (网络版彩图) 样本选择 (a)、标签修正 (b) 和本文训练框架 (c) 对比示意图. E, M 分别表示解码器和分类层, $L_{\rm scl}$ 表示监督对比损失

Figure 1 (Color online) Compared sample selection (a) and label correct (b) with our framework (c). E, M denote the encoder module and the classifier layer, respectively. $L_{\rm scl}$ denotes the supervised contrastive learning loss

神经网络的记忆效应对标签噪声学习具有重要指导作用. 该效应描述了一种现象, 即深度模型在训练前期倾向于记忆干净样本而忽略噪声标签的影响, 从而实现泛化性能的快速提升 [3]. 基于此观察, 对于标签噪声学习的方法研究主要包括两类: (1) 基于样本选择的方法 [3,4,9~13], 其核心思想是通过一个预先设定的选择标准 (如小损失准则), 从噪声训练数据集中选出一个可靠的干净子集, 在该子集上进行训练, 可以避免模型泛化性能受到大量噪声标签的影响; (2) 基于标签修正的方法 [14~18], 利用模型自身判别性对样本做出预测, 并将其作为样本的修正伪标签, 使用修正后的伪标签来监督模型训练. 以上两种策略在标签噪声学习任务上取得了较好性能.

但是, 上述两种策略均存在一定的不足. 基于样本选择的方法未能利用噪声标签样本所包含的潜在判别信息, 丢弃所有噪声样本可能会导致模型泛化性能降低, 尤其在噪声比例较大的极端情况下, 选出的小部分数据的分布与真实数据的分布存在较大偏差, 导致算法的性能退化. 而基于标签修正的策略依赖于模型预测输出的自训练过程, 容易引起错误累积 [3], 即被修正错误的标签用于随后的模型训练, 会加深模型对错误标签的记忆进而降低修正标签的质量.

针对上述两类方法中存在的问题, 本文提出了一种结合样本选择和标签修正策略的联合训练框架, 并定义了两种新的模块. 在样本选择模块中, 本文引入信息熵来衡量样本预测结果的不确定性. 利用额外的记忆序列存储每个样本在不同训练时刻的输出结果, 通过信息熵衡量单个样本在邻近轮次的预测结果的确定性程度, 挑选出确定性高的样本作为干净样本. 此外, 本文提出了一种新的惩罚项, 使得噪声样本的预测熵值提升, 进而增强噪声样本和干净样本之间的区分性. 相较于传统的小损失准则, 基于信息熵的选择标准能充分探索到困难样本, 促进模型对决策边界的学习. 在标签修正模块中, 本框架联合多个特征空间视角, 利用类原型技术为噪声样本提供修正标签. 本文基于度量的修正技术, 通过平衡两个视角的修正标签, 可有效提高标签修正模块中修正结果的可信度, 缓解不可靠修正标签导致的错误累积. 此外, 本文设计了基于对比学习的正则化项, 可进一步提高特征视角的修正标签质量和模型表征能力. 通过充分的对比、消融实验与分析, 证明了本文提出框架的有效性和先进性.

本文主要框架图及对比结果如图 1 所示. 本文的主要贡献点可以总结如下.

- 构建了一个标签噪声鲁棒的训练框架, 集成了样本选择和标签修正两个模块. 在面向噪声标签的情况下, 实现了模型的鲁棒性学习.
- 设计了基于信息熵的样本选择标准, 该准则在历史预测结果上衡量所有样本的信息熵, 并通过惩罚项进一步增大了噪声样本的熵值, 可有效地从训练数据中过滤噪声样本.
- 提出了多视角的标签修正策略, 通过平衡基于度量和基于模型预测输出的权重, 缓解自标签技术中的错误累积问题. 通过引入对比学习增强特征提取器学习性能, 能进一步提升模型表征能力和修

正标签准确率.

- 本文提出的训练框架在 4 个带有噪声标签的测试基准上均取得了最好结果.

2 相关工作

目前, 对标签噪声学习的方法研究主要集中在样本选择和标签修正. 本节将从这两方面对当前相关方法展开讨论.

2.1 基于样本选择方法的研究进展

基于样本选择方法 [3, 4, 9~12] 的核心思想是从含有标签噪声的数据集中选出一个干净的子集用于模型训练, 其关键点是如何设计出更加高效的选择准则. 早期的工作通常使用小损失准则进行样本选择 [3, 4, 9~11]. 小损失准则的基本假设认为在噪声比例符合特定条件时, 模型在训练的早期阶段会优先拟合干净标签样本, 因此模型对干净样本产生的损失值较小而对噪声标签样本产生较大的损失值. 当前对小损失准则的改进主要集中在两方面. (1) 基于阈值的方法, 该类型的代表工作有 Co-teaching [3] 和 JoCoR [19]. 在 Co-teaching 中, 作者首先通过交叉验证估计训练集中的噪声比例, 后利用该比例设定损失阈值. 高于损失阈值的样本被认为是噪声样本, 反之, 则为干净样本. 此外, 作者借助双分支网络, 将其中一个网络筛选出的干净样本供给另一个网络训练, 从而避免样本选择误差所导致的错误积累. 不同于 Co-teaching, JoCoR 使用双分支网络的输出计算一致性正则化损失, 并加入到了样本的训练损失中. 该一致性损失约束, 使得噪声标签样本往往产生较大的损失值, 从而增强了干净样本和噪声标签样本之间的区分性. (2) 基于分布检测的方法, 其代表工作有 DivideMix [4] 和 Beta 混合模型 [20]. 这两项工作均引入了动态小损失技巧, 采用预先设定的双峰分布拟合所有训练数据的损失值, 从而实现对干净样本和噪声样本的划分. 两者的区别是分别采用了混合高斯和混合 Beta 分布, 后者指出混合 Beta 分布在挑选干净样本上比混合高斯分布更加通用和有效. 相比于基于阈值的方法, 该类方法的优势在于不需要预先估计训练数据的噪声比例, 避免复杂的交叉验证.

近期, 通过模型输入波动的策略进行样本选择引起了广泛的关注 [13, 15, 21]. 这些工作认为当模型拟合标签噪声时, 噪声通常会破坏干净数据中编码的信号 [21]. 因而模型对噪声样本的预测往往变得不自信, 在序列的预测过程中, 模型输出结果通常会出现波动. 例如, SFT [13] 通过一个额外的动量队列存储模型历史输出, 并在序列化的预测结果中检测波动发生, 从而筛除噪声样本.

2.2 基于标签修正方法的研究进展

早期的标签修正工作往往借助了额外的推断步骤或网络结构, 例如, 构建知识图谱 [22] 和图网络 [23]. 近期有关标签修正的相关工作可以分为两大类. (1) 基于噪声转移矩阵的方法 [24~26]. 这类工作构建标签转移矩阵可以估计训练数据标签从干净转为噪声的概率分布, 并将该转移矩阵加入目标函数中, 降低标签噪声对模型训练的影响. 代表性的工作有 Dual-T [26], T-Revision [25]. 但这类工作往往需要借助一个额外的干净验证集 (锚点数据) 进行估计. 最新研究集中在脱离锚点数据对转移矩阵进行无偏估计 [27]. 此外, 大多估计噪声转移矩阵的方法依赖对数据分布的假设, 文献 [28] 在没有对数据分布进行额外假设的情况下, 提供了噪声估计的上界. (2) 利用网络输出进行标签修正的自训练方法. 代表性工作 ProSelfLC [16] 采用渐进式策略, 利用模型的输出对部分样本修正, 并逐渐增大修正标签的可信度. 此外, 元学习 [17, 18, 29] 也被引入到标签修正策略中来. 通过 bi-level 的训练策略, 将元网络嵌入到分类网络的训练中, 实现了端到端的标签修正. 例如, MLSC [18] 建立了一个元标签修正器, 利用

上一阶段的分类网络输出和自身标签作为网络输入, 输出了修正标签并且监督分类网络学习.

相比于现阶段的方法, 本文提出的框架从 3 个方面缓解了标签噪声对模型的影响: (1) 本文提出了基于信息熵的选择准则去筛选出干净样本. 相比于传统的小损失准则, 该准则能选出边界样本, 有利于模型学习决策边界; (2) 本文设计了一个联合标签修正方法, 可同时从特征空间和模型预测的角度对噪声样本修正, 缓解了当前标签修正方法中存在的错误积累问题; (3) 本文将基于对比学习的表示学习方法集成到训练中, 进一步提高了标签修正的准确率.

3 方法

第 3.1 小节概述标签噪声学习的问题设置以及整体框架. 第 3.2 小节介绍了基于信息熵的选择标准, 第 3.3 小节介绍了基于对比学习的联合标签修正机制. 第 3.4 小节通过两种噪声鲁棒策略的结合, 实现了更高效的噪声标签学习.

3.1 问题设置

假设有 N 个标签样本 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$, 且 D 中含有未知个数的噪声标签. 在本文所研究的多分类问题中, 标签 $y_i\in\{1,\dots,K\}$, K 表示类别数量. 在常规分类任务中, 对于给定模型, 已知其参数为 θ. 对于训练数据中每个小批量 $\mathcal{M}$, 目标函数为 $L=\frac{1}{|\mathcal{M}|}\sum_{\boldsymbol{x}\in\mathcal{M}}\mathcal{L}(\boldsymbol{x},y;\theta^t)$, 其中 $\mathcal{L}$ 表示损失函数. 在面向标签噪声学习的任务中, 本文改进了上述目标函数, 使模型在处于噪声环境的训练时能更加鲁棒.

3.2 基于信息熵的样本选择

本小节设计的基于熵的样本选择准则, 实现了在包含标签的训练集合 $\mathcal{M}$ 中对干净子集 $\mathcal{C}$ 的高效选择. 在该准则中, 本文假设模型对噪声样本的预测往往呈现出波动的状态, 即当前轮次的预测结果与随后的预测结果不相同 [3,4], 此时, 该噪声样本的预测序列元素集合 $P(\boldsymbol{x})$ 的熵值较大. 而干净样本的预测结果相对稳定, 其预测序列集合熵值较小. 基于该假设, 本文引入信息熵函数 entropy$(\cdot)$ 来衡量样本预测结果的确定性程度. 将模型预测结果稳定性高的样本认定为干净样本. 但随着训练的进行, 噪声样本的输出也会趋于稳定, 达到低熵状态. 为了降低上述情况的影响, 本文设计了一种惩罚项 penalty$(\cdot)$ 以排除高置信噪声样本的干扰. 本文定义的选择准则如下.

定义1 对于样本 $\boldsymbol{x}$, 给定超参数 ϵ, 当其预测结果的确定性 $H(\boldsymbol{x})\leqslant\epsilon$ 时, 则为干净样本. 给定模型对样本 $\boldsymbol{x}$ 的预测序列 $P(\boldsymbol{x})$, 确定性函数表示为

$$H(\boldsymbol{x})=\text{entropy}(P(\boldsymbol{x}))+\text{penalty}(P(\boldsymbol{x})). \tag{1}$$

下文分别阐述预测序列 $P(\boldsymbol{x})$、信息熵函数 entropy$(\cdot)$ 以及惩罚项 penalty$(\cdot)$ 的定义.

对于样本 $\boldsymbol{x}$, 本文引入一个记忆模块 $P(\boldsymbol{x})$ 保存模型在多个时刻的输出预测. 在实际应用中, 记忆模块 $P(\boldsymbol{x})$ 可表示成序列的形式, 通过 "先进先出" 的更新策略确保最新的预测结果用于样本选择. 首先, 将模型对样本 $\boldsymbol{x}$ 给出的预测分布表示为 $f_\theta(\boldsymbol{x})$, 其中 $y^t=\arg\max(f_\theta^t(\boldsymbol{x}))$ 和 $p^t=f_\theta^t(\boldsymbol{x})$ 分别表示在第 t 轮中, 模型将样本 $\boldsymbol{x}$ 以概率 p^t 预测成类型 y^t. 对于样本 $\boldsymbol{x}$, 记忆模块 $P(\boldsymbol{x})$ 保存模型在多个时刻的输出预测, 可表示成 $P(\boldsymbol{x})=\{[y^t=y]\cdot p^t\}_{t=1}^T$, 其中 $[\cdot]\in\{0,1\}$, T 表示记忆模块大小.

对于第 i 个样本 $\boldsymbol{x}_i$, 通过其记忆模块 $P(\boldsymbol{x}_i)$ 计算信息熵 entropy$(\cdot)$ 衡量预测不确定性. 其具体计算形式如下:

$$\text{entropy}(P(\boldsymbol{x}_i))=-\sum_{t=1}^{T}P_t(\boldsymbol{x}_i)\log P_t(\boldsymbol{x}_i). \tag{2}$$

此外, 用以排除高置信噪声样本的干扰的惩罚项 penalty(·) 具体是

$$\text{penalty}(P(\boldsymbol{x}_i)) = -\mu \cdot \sum_{t=1}^{T} \mathbb{1}(P_t(\boldsymbol{x}_i) = 0), \tag{3}$$

其中, $\mathbb{1}(\cdot)$ 为指示函数, μ 是惩罚因子超参.

根据定义 1, 可从所有训练数据中挑选出干净样本并构成干净子集 $\mathcal{C}$, 即通过一个超参数 ϵ, 从所有样本对应的集合 $I = \{H(\boldsymbol{x}_i)\}_{i=1}^{N}$ 中选出确定性值较小的一部分. 为了缩小对参数 ϵ 的搜索空间大小, 引入归一化函数对所有样本确定性集合进行处理 $I' = \text{Normalize}(I)$, 故 $H'(\boldsymbol{x}_i) \in [0,1]$, 且 $\epsilon \in [0,1]$. 该基于带惩罚项的信息熵选择准则选出的干净子集可表示为

$$\mathcal{C} = \{(\boldsymbol{x}_i, y_i) | H'(\boldsymbol{x}_i) \leqslant \epsilon\}_{i=1}^{N}. \tag{4}$$

集合 $\mathcal{C}$ 作为干净训练集, 用于后续模型训练.

3.3 基于对比学习的标签修正机制

为了解决标签修正方法中常见的错误累积问题, 即模型使用其自身预测分布 $f(\boldsymbol{x})$ 或者预测类别 $\arg\max f(\boldsymbol{x})$ 作为其修正标签用于自身训练, 其偶然产生的误差会不断累积. 本文从度量学习的角度, 通过度量样本的特征向量与类原型向量距离来对噪声样本进行标签修正. 同时调整基于度量的和基于模型预测输出的修正标签的权重, 实现了更准确的标签修正结果. 直观上, 混合两个不同视角产生的修正标签对错误的修正标签更加鲁棒, 从而缓解了错误标签引起的误差积累. 此外, 本文引入监督对比学习来提升特征空间的判别性, 以此进一步增加修正标签的准确率, 并通过设计重排序策略, 避免监督对比学习受到标签噪声影响.

标签修正策略. 通过构建类原型, 度量噪声样本特征向量与类原型向量的距离来进行标签修正. 对于噪声子集 $\mathcal{N}$ 的每个样本 $(\boldsymbol{x}, y)$, 其基于特征空间的修正标签可以表示为 $y^{\text{m}} = [p_1^{\text{m}}, \ldots, p_K^{\text{m}}]$. 其中, p_k^{m} 可以表示为

$$p_k^{\text{m}} = \frac{\exp(-F_{\text{dis}}(\boldsymbol{z}, \boldsymbol{c}_k))}{\sum_{j=1}^{K} \exp(-F_{\text{dis}}(\boldsymbol{z}, \boldsymbol{c}_j))}, \quad \sum_{k=1}^{K} p_k^{\text{m}} = 1, \tag{5}$$

其中, $\boldsymbol{z}$ 和 $\boldsymbol{c}_k$ 分别表示样本的特征向量和第 k 类样本的类原型向量 $\boldsymbol{c}_k = \frac{1}{|D_k|} \sum_{(\boldsymbol{x}_i, y_i) \in D_k} \boldsymbol{z}_i$, F_{dis} 表示度量函数. 随后, 通过一个平衡系数 λ, 调节基于度量和模型预测输出的两个修正标签 y^{m} 和 $y^{\text{c}} = f_\theta(\boldsymbol{x})$ 的权重. 对于噪声样本 $(\boldsymbol{x}, y) \in \mathcal{N}$, 其修正标签可表示为

$$y^{\text{refurb}} = \lambda \cdot y^{\text{m}} + (1 - \lambda) \cdot y^{\text{c}}. \tag{6}$$

此外, λ 被设计成与训练轮次相关的退火系数, 即 $\lambda = (1 - T/T_{\max}) \times 0.8$, 其中, T 和 $T_{\max}$ 表示当前训练轮次和最大训练轮次. 在训练前期对基于特征空间的修正标签给予更多权重, 随着模型性能提升逐渐增大基于模型视角的修正标签的权重.

监督对比学习. 现阶段工作表明, 特征空间的判别性能可以通过对比学习显著增强 [30]. 因此, 本文引入了有监督对比学习来提升特征空间的判别性, 从而增强联合标签修正策略的性能. 具体实现方法是拉近同类样本的特征向量在特征空间中的距离, 扩大异类样本的特征向量的距离. 有监督的对比学习损失可以被写成

$$\mathcal{L}_{\text{SCL}}(\boldsymbol{x}_i, y_i, \boldsymbol{z}_i) = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{p \in \text{Pos}} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau) + \sum_{n \in \text{Neg}} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n / \tau)}, \tag{7}$$

其中 $\boldsymbol{z}_i$ 表示样本 $\boldsymbol{x}_i$ 的特征向量, τ 表示放缩系数, Pos 和 Neg 分别表示当前样本的同类样本集合和异类样本集合. 在实现过程中, Pos 和 Neg 来自所挑选出的干净子集 $\mathcal{C}$. 由于训练前期模型性能不足, 集合 $\mathcal{C}$ 可能带有噪声样本, 导致监督对比学习受到噪声标签影响.

因此, 本文设计了一个排序策略, 称为重排序. 该策略对样本可靠度进行排序, 选择可信度较高的样本来进行对比学习, 以此避免噪声信号. 具体地, 通过构建类原型, 度量样本特征向量与其类原型距离, 并以此进行排序. 由于靠近类原型的样本更能代表该类的特征属性, 属于干净样本的可靠度更大. 对于第 k 类样本, 借助类原型和预先定义的度量函数 $F_{\mathrm{dis}}(a,b)$, 可以选出距离类原型最近的一部分样本. 该集合可表示为

$$D_k^* = \{(\boldsymbol{x}_i, y_i) | F_{\mathrm{dis}}(\boldsymbol{z}_i, \boldsymbol{c}_k) \leqslant \beta\}, \tag{8}$$

其中, β 是一个超参数. 为了使 β 独立于预先设置的度量函数, 将其表示为一个比例阈值. 通过重排序策略所挑选出用于对比学习的样本集合可表示为 $\mathcal{C}^* = \{D_1^* \cap \cdots \cap D_K^*\}$. 最终, 对于样本 $(\boldsymbol{x}_i, y_i)$, 式 (7) 中的 Pos 和 Neg 分别可以写成 $\mathrm{Pos} = D_{k=y_i}^*$, $\mathrm{Neg} = D_{k\neq y_i}^*$.

3.4 模型训练

本文学习算法采用批量学习的策略. 对于训练数据中每个小批量 $\mathcal{M} \in D^N$, 通过基于信息熵的样本选择准则可分为干净样本集合 $\mathcal{C}$ 和噪声样本 $\mathcal{N}$, 其中 $\mathcal{N} = \mathcal{M} - \mathcal{C}$. 对于每个样本 $(\boldsymbol{x}, y) \in \mathcal{N}$, 本文通过基于对比学习的联合标签修正框架得到其修正标签 y^{refurb}. 综上, 模型训练样本集为 $\{(\boldsymbol{x}, y) \in \mathcal{C}\} \cup \{(\boldsymbol{x}, y^{\mathrm{refurb}}) \in \mathcal{N}\}$. 本文框架的目标函数可表示为

$$L = \frac{1}{|\mathcal{M}|}\left[\sum_{\boldsymbol{x}\in\mathcal{C}} \mathcal{L}(\boldsymbol{x}, y) + \sum_{\boldsymbol{x}\in\mathcal{N}} \mathcal{L}(\boldsymbol{x}, y^{\mathrm{refurb}}) + \alpha \sum_{\boldsymbol{x}\in\mathcal{C}^*} \mathcal{L}_{\mathrm{SCL}}(\boldsymbol{x}, y)\right], \tag{9}$$

其中, $\mathcal{L}_{\mathrm{SCL}}$ 是由式 (7) 计算所得的对比损失, $\mathcal{C}^*$ 表示通过重排序策略从 $\mathcal{C}$ 中选出用于对比学习的集合, α 是超参平衡系数. 具体训练流程可参见算法 1.

4 实验

本节对所提出的框架进行验证和分析, 并在 4 个公开数据集上, 与现有的标签噪声学习方法进行对比. 此外, 本节通过充分的消融实验, 验证本文方法中各模块的具体作用和有效性.

实验数据. 为了模拟真实世界中的噪声数据集, 常通过人为的方式将常用数据集的干净标签转换成错误标签. 通过一个噪声转移矩阵 $\boldsymbol{T}$, 将干净样本标签 y 以概率转移 p 为噪声标签 $\hat{y}$. 根据噪声类型不同, 噪声标签生成方式可分为如图 2 所示的 3 种噪声类型.

- 随机噪声类型 (symmetric label noise, 简记为 Sym.). 样本标签随机变为其他任意一类标签.
- 类别依赖噪声类型 (pair label noise, 简记为 Pair). 样本标签固定转为某一类标签.
- 实例依赖噪声类型 (instance-dependent label noise, 简记为 Inst.). 不同于前两种噪声生成方法, 实例依赖噪声倾向于将决策边界的样本标签替换成噪声标签, 该类型噪声对模型鲁棒性训练最具有挑战.

本文所测试的基准数据集包含 CIFAR-10, CIFAR-100, Clothing-1M 和 Food-101N. 其中对于 CIFAR-10 和 CIFAR-100, 使用上述噪声转移矩阵, 生成不同噪声类型和噪声比例的训练数据. Clothing-1M 和 Food-101N 属于开放场景下的噪声数据集, 能有效测评鲁棒性算法在真实噪声环境下的性能表现. 其噪声比例分别为 39% 和 20%.

算法 1 学习算法

Require: 训练集 $\mathcal{D}^N$, 记忆模块大小 T, 熵阈值 ϵ, 排序比例 β, 距离函数 F_{dis}, 平衡系数 λ, α.
Ensure: 最优模型参数 θ^*.

$\theta^0, \text{MB} \leftarrow \text{WarmUp}(\mathcal{D}_N)$; //初始化模型参数 θ 和记忆模块 MB.
while $e <$ MaxEpoch **do**
  $\{H_i\}_{i=1}^N \leftarrow \text{MB}$; //计算每个样本预测结果的不确信度 ▷ 式 (1).
  $\mathcal{C}, \mathcal{N} \leftarrow (\mathcal{D}_{\text{train}}^N, \text{Normalize}(\{H_i\}_{i=1}^N))$; //划分训练数据 ▷ 式 (4).
  $\{c_k\}_{i=1}^K \leftarrow \mathcal{C}$; //构建类原型.
  $\mathcal{C}^* \leftarrow (\mathcal{C}, \{c_k\}_{i=1}^K, F_{\text{dis}})$; //根据类原型进一步划分子集 ▷ 式 (8).
  for iter $\in \{1, \ldots, \text{num\_iters}\}$ **do**
    Draw two mini-batch $\{(\boldsymbol{x}_l, y_l)\}_{l=1}^B, \{(\boldsymbol{x}_u)\}_{u=1}^B$ from $\mathcal{C}, \mathcal{N}$, respectively;
    for $u \in \{1, \ldots, B\}$ **do**
      $y_u^{\text{m}} \leftarrow (\boldsymbol{x}_u, \{c_k\}_{i=1}^K)$; //基于类原型进行标签修正 ▷ 式 (5).
      $y_u^{\text{refurb}} \leftarrow \lambda \cdot y_u^{\text{m}} + (1-\lambda) \cdot y_u^{\text{c}}$;
    end for
    for $(\boldsymbol{x}_b, y_b) \in \{(\boldsymbol{x}_l, y_l)_{l=1}^B \cap (\boldsymbol{x}_u, y_u^{\text{refurb}})_{u=1}^B\}$ **do**
      RandomSampler(Pos, Neg) from $\mathcal{C}^*$;
      $L_{\text{SCL}}^b \leftarrow (\boldsymbol{x}_b, y_b, \text{Pos}, \text{Neg})$; //对比损失 ▷ 式 (7).
      $L^b \leftarrow L_{\text{CE}}^b + \alpha \cdot L_{\text{SCL}}^b$;
    end for
    $\theta^e \leftarrow \text{SGD}(\frac{1}{2B} \sum_{b=1}^{2B} L^b; \theta^e)$;
  end for
  Evaluate the training set $\mathcal{D}_{\text{train}}^N$ and update MB;
end while
return θ^*.

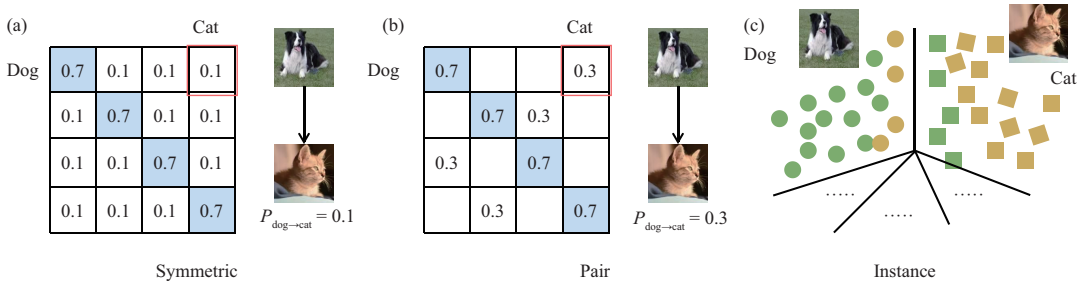

图 2 (网络版彩图) 噪声生成示例

Figure 2 (Color online) Examples of generation of noisy labels. (a) Symmetric; (b) pair; (c) instance

神经网络设置. 参考 DivideMix [4], ELR [11], 本文分别采用了 ResNet-18 和 ResNet-34 [31] 作为 CIFAR-10 和 CIFAR-100 训练的骨干网络. 在实验中, 采用动量为 0.9, 权重衰减为 5E−4 的 SGD 作为模型训练的优化器. 学习率初始化为 0.02, 小批量大小设置为 64. 训练轮数设置为 200 轮, 学习率在第 150 轮衰减为初始的 0.1 倍. 为了给模型预热和初始化选择阶段中的记忆模块, 本文在正式训练前引入了一个预热阶段. 对 CIFAR-10 和 CIFAR-100, 分别训练 10 轮和 30 轮以预热模型. 在数据处理部分, 本文仅使用了随机裁剪和水平翻转两种数据增强策略. 除此之外, 在与基于半监督的鲁棒性学习方法进行比较时, 本文使用了带有 PreAct 模块的 ResNet-18 模型, 其余训练细节保持不变.

对于真实世界中的噪声数据集 Clothing-1M 和 Food-101N, 本文使用了在 ImageNet 上预训练的 ResNet-50 作为骨干网络. 在实验中, 采用动量为 0.9, 权重衰减为 1E−3 的 SGD 作为模型训练的优

表 1 CIFAR-10 和 CIFAR-100 数据集上未使用半监督技术方法的分类准确率 (%) 结果对比

Table 1 Comparison with the state-of-the-art without SSL on CIFAR-10 and CIFAR-100 (%) [a)]

Dataset	Method	Symmetric		Pair	Instance	
		20%	50%	40%	20%	40%
CIFAR-10 (ResNet-18)	Co-teaching [3]	87.16±0.1	72.80±0.4	74.51±1.1	86.54±0.1	79.98±0.3
	Peer Loss [32]	88.69±0.3	79.52±0.5	76.08±0.6	88.19±0.5	81.53±0.7
	CORES* [21]	89.97±0.1	81.47±0.5	78.82±0.7	89.67±0.3	82.99±0.5
	CDR [33]	89.72±0.3	82.64±0.8	78.47±0.5	90.41±0.3	83.07±1.3
	Me-Momentum [10]	91.44±0.3	87.17±0.4	81.94±0.3	90.86±0.2	86.66±0.9
	PES [34]	92.38±0.4	87.45±0.3	89.52±0.9	92.69±0.4	89.73±0.5
	Base	93.21±0.3	90.49±0.4	90.55±0.2	92.87±0.2	90.59±0.3
	Ours	**93.37±0.2**	**92.02±0.1**	**92.14±0.3**	**93.02±0.3**	**92.05±0.2**
CIFAR-100 (ResNet-34)	Co-teaching [3]	59.28±0.4	41.37±0.1	39.22±0.6	57.24±0.6	45.69±0.9
	Peer Loss [32]	64.87±0.3	48.55±0.6	40.97±0.2	63.82±0.3	47.91±0.5
	CORES* [21]	65.01±0.2	50.07±0.4	43.66±0.5	64.86±0.5	49.62±0.7
	CDR [33]	66.52±0.2	55.30±0.9	46.27±0.9	67.33±0.6	55.94±0.5
	Me-Momentum [10]	68.03±0.5	60.42±0.6	52.88±0.5	68.11±0.5	58.58±1.2
	PES [34]	68.89±0.4	58.90±2.7	59.08±1.1	70.49±0.7	65.68±1.4
	Base	72.07±0.3	67.04±0.2	65.50±0.3	73.37±0.2	71.07±0.3
	Ours	**72.21±0.2**	**69.70±0.4**	**68.87±0.5**	**72.69±0.3**	**71.24±0.4**

a) The bold and the underline represent the best and the second best testing accuracy, respectively. Base denotes the entropy-based sample selection criterion.

化器. 学习率初始化为 0.002, 小批量大小设置为 32. 对 Clothing-1M 数据集, 训练轮数为 30, 学习率在 25 轮时衰减为初始的 0.1 倍. 对 Food-101N 数据集, 训练轮数为 50, 学习率在 40 轮时衰减为初始的 0.1 倍. 实验中, 图片尺寸初设为 256×256, 并且使用随机水平翻转和随机裁剪的方式将其裁剪为 224×224. 预训练轮数均为 3 轮.

超参数设置. 本文提出的方法部分一共包含 3 个超参数: (1) 基于熵的选择准则中记忆模块的大小 T; (2) 基于熵的选择准则中熵的阈值 ϵ; (3) 联合标签修正方法中的比例阈值 β. 对于所有实验, 超参数统一设置为 $T = 5, \epsilon = 0.2, \beta = 40\%$.

4.1 实验结果

基线方法. 由于半监督方法能有效对抗噪声标签, 现阶段方法常使用半监督技术而取得显著的性能提升 [4, 9, 11, 12]. 因此, 引入了 MixMatch 去提升本文框架的性能. 本小节方法比较分为两部分. (1) 未使用半监督的方法: Co-teaching [3], Peer Loss [32], CDR [33], CORES* [21], Me-Momentum [10], PES [34]. (2) 基于半监督的方法: SELF [12], M-correction [9], DivideMix [4], ELR+ [11].

表 1 报告了当前未使用半监督技术的标签噪声学习方法在多种标签噪声的 CIFAR-100 和 CIFAR-100 数据集上的结果. 其中 Base 表示本文提出的基于熵的选择标准. 从表中结果可以看出, 即使只使用本文框架中的样本选择部分, 也可以胜过当前的最优方法. 尤其是当噪声比例提高时, 本文方法的优势更加明显. 比如在 Sym. 50% 噪声情况下, 仅仅是基于熵的样本选择方法在 CIFAR-10 数据集上提高了 3.04%, 在 CIFAR-100 上提高了 6.62%. 此外本文提出的联合标签修正也是有效的. 在任何一

表 2 CIFAR-10 和 CIFAR-100 数据集上使用半监督技术比较结果 (%)

Table 2 Comparison with the state-of-the-art with SSL on CIFAR-10 and CIFAR-100 (%)[a)]

Method/noise	CIFAR-10			CIFAR-100		
	Sym. 20%	Sym. 50%	Sym. 80%	Sym. 20%	Sym. 50%	Sym. 80%
SELF [12]	94.1±0.2	90.5±0.6	77.8±0.4	72.9±0.1	60.5±0.4	43.7±0.8
M-correction [9]	94.0	92.0	86.8	73.9	66.1	48.2
DivideMix [4]	95.2	94.2	**93.0**	75.2	72.2	**57.9**
ELR+ [11]	94.9±0.2	93.6±0.1	90.4±0.2	75.5±0.2	71.0±0.2	50.4±0.8
Ours (semi)	**95.6±0.2**	**95.1±0.2**	91.7±0.5	**75.6±0.1**	**72.9±0.3**	55.7±0.6
	Inst. 20%	Inst. 40%	Pair 40%	Inst. 20%	Inst. 40%	Pair 40%
SELF [12]	94.9±0.2	90.2±0.6	91.3±0.9	73.9±0.2	68.3±0.3	64.4±0.8
DivideMix [4]	95.5±0.1	94.5±0.2	93.4±0.5	75.2±0.2	70.9±0.1	71.0±0.7
ELR+ [11]	94.9±0.1	94.3±0.2	93.6±0.5	**75.8±0.1**	**74.3±0.3**	72.3±0.9
Ours (semi)	**95.6±0.1**	**95.1±0.1**	**95.0±0.4**	75.7±0.2	73.4±0.4	**72.9±0.5**

a) The bold and the underline represent the best and the second best testing accuracy, respectively.

表 3 Clothing-1M 和 Food-101N 数据集上各方法结果比较 (%)

Table 3 Comparison with the state-of-the-art on Clothing-1M and Food-101N (%)[a)]

Clothing-1M (pre-trained ResNet-50)								
CE	Co-teaching [3]	DMI [35]	T-Revision [25]	JNPL [36]	DivideMix [4]	ELR+ [11]	Ours	Ours (semi)
69.21	58.68	72.46	74.18	74.15	74.76	74.81	74.71	**74.96**
Food-101N (pre-trained ResNet-50)								
CE	Co-teaching [3]	DMI [35]	T-Revision [25]	GCE [37]	S2E [38]	PLC [39]	Ours	Ours (semi)
84.03	83.73	85.52	85.97	84.96	84.97	85.28	86.37	**87.22**

a) The best performance is highlighted by bold.

种噪声情况下, 该模块对整体框架性能均有提升, 同时提升幅度也随着噪声比例升高而变大. 上述结果表明了本文所提出的训练框架在面向噪声标签的鲁棒性学习中, 表现出了较大的优势, 以上结果表明基于熵的选择标准在筛选噪声样本上具有巨大潜力.

表 2 报告了当前使用半监督技术的标签噪声学习方法在多种标签噪声的 CIFAR-10 和 CIFAR-100 数据集上的结果. 在当前所提出框架的基础上, 本文引入了半监督学习中经典的 MixMatch 方法加强模型的抗噪效果. 可以看出, 在各种噪声类型和噪声比例情况下, 引入 MixMatch 均带来了较大的性能提升. 同时, 相比于当前的最优方法, 本文方法在绝大多数情况下, 都取得了一定的提升. 即使在 CIFAR-10 数据集 Sym. 80% 比例的噪声情况下, 依然取得了 91.7% 的测试准确率.

表 3 [35~39] 报告了标签噪声学习方法在真实世界中噪声数据集上的结果对比. 在数据集 Clothing-1M 上, 参考之前的文献 DivideMix, 从训练数据中随机采样出 1000 个小批量去确保每类数据中的噪声标签处于平衡状态. 本文框架在两个真实数据集上取得了最好性能. 其中, 在数据集 Food-101N 上, 本文框架相比之前的最优方法 T-Revision 提高了 1.25%.

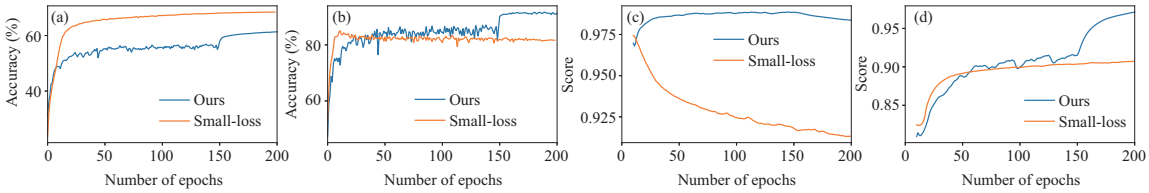

图 3 (网络版彩图) 本文样本选择准则与小损失技巧在 CIFAR-10 Sym. 50% 的训练曲线对比

Figure 3 (Color online) Comparison results with the small-loss criterion on CIFAR-10 Sym. 50%. (a) Train accuracy; (b) test accuracy; (c) precision score; (d) recall score

表 4 各样本选择方法在 CIFAR-10 (左) 和 CIFAR-100 (右) 上的 F_1-scores 结果对比

Table 4 Comparison results of F_1-scores on CIFAR-10 (left) and CIFAR-100 (right)[a)]

Criterion	Method	Sym. 60%	Pair 40%	Inst. 40%	Sym. 60%	Pair 40%	Inst. 40%
Small-loss	Co-teaching [3]	0.841	0.673	0.651	0.740	0.598	0.577
	JoCoR [19]	0.892	**0.704**	0.669	0.782	**0.617**	**0.591**
	CORES [21]	**0.929**	0.491	**0.874**	**0.923**	0.277	0.097
Other	CL [40]	0.810	–	0.629	0.790	–	0.500
Entropy	Ours	**0.961**	**0.973**	**0.984**	**0.896**	**0.849**	**0.901**

a) "–" denotes the missing value in original paper.

4.2 各模块有效性分析

基于熵的选择准则. 为了验证本文所提出的选择标准优于当前诸多基于小损失的选择标准, 本文从两个视角 "查准率" 和 "查全率" 对当前选择标准进行了分析, 即良好的选择标准既要保证选出子集的干净程度, 也要尽可能多从噪声训练集中选出干净样本. 为了公平比较, 实验中只使用了本文框架中的样本选择部分与先前基于小损失准则的方法, 同时分别引入准确率 Precision, 召回率 Recall 和 F_1-scores 作为衡量标准.

对比结果如图 3 所示. 在 40% 的噪声比例情况下, 通过小损失准则训练的模型在含有噪声标签的训练集上展现出严重的过拟合现象 (即准确率超过 60%), 导致了其在测试阶段取得了较低的泛化性. 图 3(a) 和 (b) 中结果表明本文所提出的选择标准在对抗噪声标签上的优越性. 图 3(c) 和 (d) 刻画了样本选择结果的准确率和召回率, 本文基于熵值的选择标准明显优于小损失标准, 达到了更高的准确率和召回率. 同时可以看出, 随着训练的进行, 小损失标准选出的干净样本集合中噪声比例逐步提升. 小损失标准对难样本的筛选不够有效, 导致接近 10% 的边界样本被判定为噪声样本而丢弃. 此外, 表 4 [40] 展示了几种典型改进的小损失标准在干净样本筛选过程中的 F_1-scores. 相比现有 3 种改进的小损失标准, 本文基于信息熵的选择准则在大多数噪声情况下都展示出了最优的选择结果.

(分析) 与小损失准则对比. 该样本选择准则在不同训练阶段的样本选择结果如图 4 所示. 相对于小损失准则中总体损失值服从于双峰分布, 在本文提出的准则中, 训练样本预测置信度所产生的熵更加服从于 0-1 分布. 这种分布迫使样本只有两个确定性选择 (噪声样本或者干净样本), 而不是双峰分布中以一定概率属于某一个峰. 直观上, 小损失准则中的不确定选择更加容易包容噪声. 其次, 在本文的选择准则中, 阈值的设置是一个独立于训练数据中噪声比例的超参数, 因此不需要像小损失准则中使用复杂的交叉验证预先来确定选择阈值. 在后续的实验部分, 本文在 4 个测试基准的多种噪声比例上, 采用相同的熵阈值 $\epsilon = 0.2$ 均取得了有竞争力的结果.

(分析) 如何有效区分难样本和边界样本. 从预测结果的熵的角度来看, 难样本和边界样本展现出

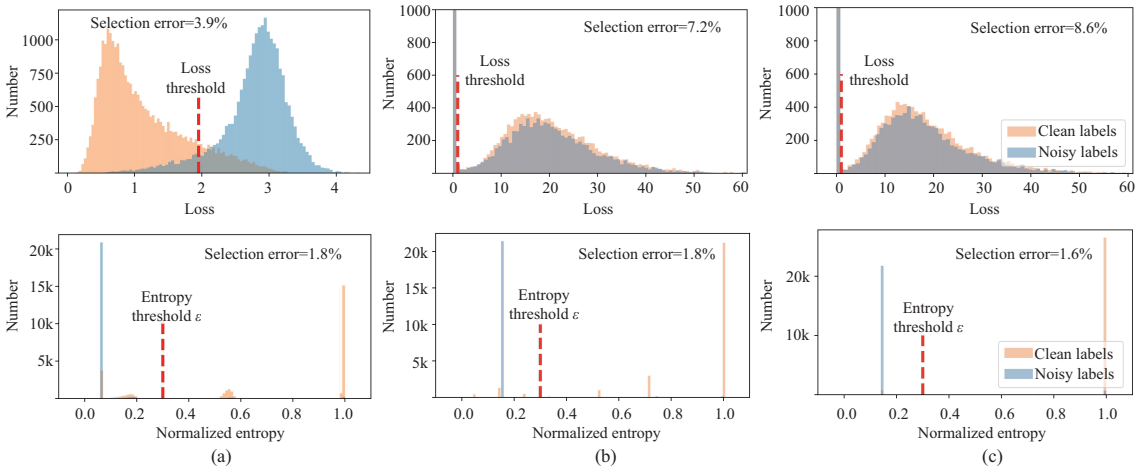

图 4 (网络版彩图) 本文选择标准与小损失准则对比, 本文准则在 CIFAR-10 Sym. 50% 噪声下实现了更低的选择误差

Figure 4 (Color online) Compared with the small-loss criterion, the proposed entropy-based criterion attains the lower selection error under CIFAR-10 50% symmetric label noise. (a) Epoch 10; (b) Epoch 100; (c) Epoch 200

相似的结果. 但是随着模型训练以及其性能的提升, 模型对于难样本展现出更强的学习能力, 而对噪声样本依然保持不稳定的预测结果. 此时, 本文选择标准中的惩罚项给予了历史预测结果不稳定的样本一个惩罚值. 因此, 噪声样本的信息熵降低, 而难样本的信息熵不会受到影响. 最终, 通过一个熵阈值, 本文方法能在选择过程中保留难样本, 而过滤掉边界样本. 由于难样本无法通过可视化的手段显示出来, 本文尝试从样本选择结果的 Recall 准则来展示本文选择标准的优越性. 从图 3(c) 和 (d) 中可以观察到, 本文选择标准达到了 98% 以上的召回率, 即 98% 的干净样本中被本文选择标准选出, 此结果可以证明绝大多数难样本能被本文准则有效选出.

基于对比学习的标签矫正模块. 在标签矫正模块中, 生成标签的质量同时依赖于基于度量和基于模型输出的标签. 通过比较传统的自标签技术, 本文所提出的标签校正模块在修正噪声标签上更具有竞争力. 相关对比结果如图 5 所示. 图 5(a) 记录了各种方法所生成修正标签的准确率. 相比较单一地使用模型输出作为预测标签, 本文所提出的混合标签修正技巧取得了更好的修正效果. 此外, 考虑到预训练技术可以有效提高模型初期的特征表现性能, 本文使用了现有的自监督表征学习技术, 例如 SimCLR, 对模型特征提取器进行了预训练. 从图中结果可以看到, 添加预训练平均提升了 2% 的标签修正准确率. 图 5(b) 中刻画了添加排序环节后的测试准确率和训练准确率. 从图中学习曲线可知, 排序环节有效抑制了模型在噪声训练数据上的过度拟合, 有效提升了分类准确率.

本文所提出的模块主要包含三部分, 基于熵的样本选择准则, 联合标签修正策略和监督对比损失. 表 5 展示了消融实验的结果. 相比于直观的交叉熵损失, 本文所提出的框架分别在 CIFAR-10 40% 和 80% 随机噪声中带来了 13.5% 和 36.5% 的提升. 其中, 值得一提的是仅依靠样本选择模块, 模型在 CIFAR-10 (Sym. 40%) 的测试正确率可达到 90.72%. 此外, 联合标签修正在噪声比例增大时, 带来的模型性能提升也是逐渐增大的. 因此, 本文框架各部分的作用得到了有效验证.

4.3 超参数选择

本文所提出的框架中包含 3 个超参数, 分别是历史模块大小 T, 熵阈值 ϵ 和重排序比例 β.

历史模块大小 T 和熵阈值 ϵ. 通过网格搜索策略, 在空间 $T \in \{3,5,7,9\}$ 和 $\epsilon \in \{0.2,0.4,0.6,0.8\}$

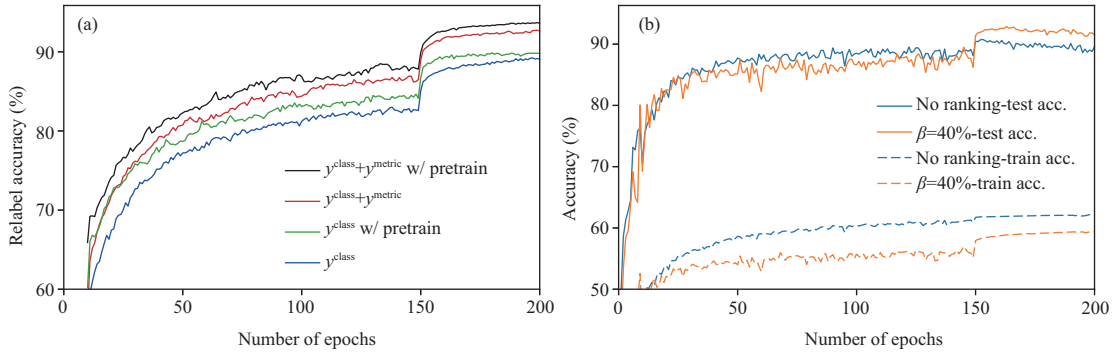

图 5 (网络版彩图) 在 **CIFAR-10 Inst. 50%** 上, (a) 验证了混合标签修正技术的有效性, (b) 验证了标签修正中重排序的有效性

Figure 5 (Color online) On CIFAR-10 with 50% instance label noise, the quality of pseudo-labels (a) verifies the effectiveness of the proposed hybrid label correction function; training curves (b) verifies the effectiveness of the reranking phase

表 5 在 **CIFAR-10 & 100** 上, 通过消融实验对框架中各模块有效性分析

Table 5 Ablation study results of test accuracy (%) on CIFAR-10 & 100[a)]

Components			CIFAR-10		CIFAR-100	
Entropy-S	Joint-LC	SCL	Sym. 40%	Sym. 80%	Sym. 40%	Sym. 80%
Cross entropy			77.20	36.44	47.41	10.96
✓	–	–	90.72	72.90	66.97	31.67
✓	–	✓	91.34	73.87	67.74	31.99
✓	✓	–	92.31	75.06	69.69	33.01
✓	✓	✓	92.73	76.14	70.30	33.38

a) Entropy-S, Joint-LC, and SCL denote entropy-based selection, joint label correction, and supervised contrastive learning, respectively.

表 6 超参数 T 和 ϵ 网格搜索的分类准确率 (%) 对比

Table 6 Comparison results of test accuracy (%) of hyper-parameters T and ϵ via grid search[a)]

T	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$
3	91.527	89.070	82.182	81.731
5	**92.337**	87.633	83.890	77.507
7	91.820	87.367	84.278	80.082
9	91.582	86.998	84.451	77.324

a) The best performance is highlighted in bold.

中, 对两个超参数进行了搜索. 在 CIFAR-10 Sym. 50% 的噪声设置下, 对比结果如表 6 所示. 从挑选样本的角度来看, 在 $T = 5, \epsilon = 0.2$ 的组合下取得了最好结果. 同时, 样本选择效果对熵阈值 ϵ 表示较为敏感. 在 ϵ 固定时, 最终选择结果并不会较大程度地受到历史模块大小 T 的影响. 因此, 在本文所有实验中, 均采用了 $T = 5, \epsilon = 0.2$ 的超参数组合.

排序比例阈值 β. 本文从测试准确率、样本选择召回率和准确率 3 个角度对超参数进行了评估, 在 CIFAR-10 Sym. 50% 的噪声条件下, 对比结果如图 6 所示. 在三项评价标准下, $\beta = 40\%$ 均取得了最好结果. 这反映了可信赖样本数量对类原型的构建是比较重要的. 过多选入样本容易引入噪声数据,

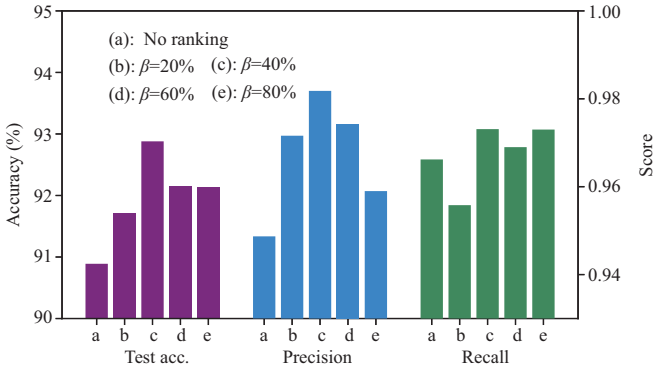

图 6 (网络版彩图) 超参数 β 的敏感性分析

Figure 6 (Color online) Sensitivity analysis of hyper-parameter β

而样本量过少容易导致类分布的偏移. 在本文所有实验中, 均使用了 $\beta = 40\%$.

5 结论

本文面向标签噪声学习设计了一个集成样本选择和标签修正的联合训练框架. 在样本选择中, 本文提出了一个基于信息熵的选择标准. 该标准通过信息熵来衡量模型对样本预测序列的确定性程度, 并挑选预测序列稳定的样本作为干净样本. 该准则还带有一个惩罚项, 以此避免高置信噪声样本被选中. 在标签修正中, 本文在特征空间中构建了类原型, 通过度量样本特征向量到类中心距离并结合预测结果进行标签修正. 同时, 本文引入监督对比学习来提升特征空间的判别性, 提升修正标签的准确率. 通过一个重排序策略避免对比学习框架受到噪声信号的干扰. 在 4 个测试基准上的丰富实验表明本文框架在面向标签噪声学习任务上具有较强鲁棒性.

参考文献

1 Angluin D, Laird P. Learning from noisy examples. Machine Learn, 1988, 2: 343–370
2 Frenay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst, 2014, 25: 845–869
3 Han B, Yao Q, Yu X, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Proceedings of Conference on Neural Information Processing Systems, 2018
4 Li J, Socher R, Hoi S C. DivideMix: learning with noisy labels as semi-supervised learning. In: Proceedings of International Conference on Learning Representations, 2020
5 Wei Q, Feng L, Sun H, et al. Fine-grained classification with noisy labels. 2023. ArXiv:2303.02404
6 Zhou Z H. Open-environment machine learning. Natl Sci Rev, 2022, 9: nwac123
7 Karimi D, Dou H, Warfield S K, et al. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal, 2020, 65: 101759
8 Blum A, Kalai A, Wasserman H. Noise-tolerant learning, the parity problem, and the statistical query model. J ACM, 2003, 50: 506–519
9 Arazo E, Ortego D, Albert P, et al. Unsupervised label noise modeling and loss correction. In: Proceedings of International Conference on Machine Learning, 2019
10 Bai Y, Liu T. Me-Momentum: extracting hard confident examples from noisily labeled data. In: Proceedings of International Conference on Computer Vision, 2021
11 Liu S, Niles-Weed J, Razavian N, et al. Early-learning regularization prevents memorization of noisy labels. In: Proceedings of Conference on Neural Information Processing Systems, 2020

12 Nguyen D T, Mummadi C K, Ngo T P N, et al. SELF: learning to filter noisy labels with self-ensembling. In: Proceedings of International Conference on Learning Representations, 2020
13 Wei Q, Sun H, Lu X, et al. Self-Filtering: a noise-aware sample selection for label noise with confidence penalization. In: Proceedings of European Conference on Computer Vision, 2022
14 Tanaka D, Ikami D, Yamasaki T, et al. Joint optimization framework for learning with noisy labels. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018
15 Song H, Kim M, Lee J G. SELFIE: refurbishing unclean samples for robust deep learning. In: Proceedings of International Conference on Machine Learning, 2019
16 Wang X, Hua Y, Kodirov E, et al. ProSelfLC: progressive self label correction for training robust deep neural networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021
17 Wu Y, Shu J, Xie Q, et al. Learning to purify noisy labels via meta soft label corrector. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021
18 Zheng G, Awadallah A H, Dumais S. Meta label correction for noisy label learning. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021
19 Wei H, Feng L, Chen X, et al. Combating noisy labels by agreement: a joint training method with co-regularization. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020
20 Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training. In: Proceedings of Conference on Neural Information Processing Systems, 2020
21 Cheng H, Zhu Z, Li X, et al. Learning with instance-dependent label noise: a sample sieve approach. In: Proceedings of International Conference on Learning Representations, 2021
22 Li Y, Yang J, Song Y, et al. Learning from noisy labels with distillation. In: Proceedings of International Conference on Computer Vision, 2017
23 Xiao T, Xia T, Yang Y, et al. Learning from massive noisy labeled data for image classification. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2015
24 Hendrycks D, Mazeika M, Wilson D, et al. Using trusted data to train deep networks on labels corrupted by severe noise. In: Proceedings of Conference on Neural Information Processing Systems, 2018
25 Xia X, Liu T, Wang N, et al. Are anchor points really indispensable in label-noise learning? In: Proceedings of Conference on Neural Information Processing Systems, 2019
26 Yao Y, Liu T, Han B, et al. Dual T: reducing estimation error for transition matrix in label-noise learning. In: Proceedings of Conference on Neural Information Processing Systems, 2020
27 Li X, Liu T, Han B, et al. Provably end-to-end label-noise learning without anchor points. In: Proceedings of International Conference on Machine Learning, 2021
28 Gao W, Zhang T, Yang B B, et al. On the noise estimation statistics. Artif Intelligence, 2021, 293: 103451
29 Sun H, Guo C, Wei Q, et al. Learning to rectify for robust learning with noisy labels. Pattern Recognition, 2022, 124: 108467
30 Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. In: Proceedings of Conference on Neural Information Processing Systems, 2020
31 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2016
32 Liu Y, Guo H. Peer loss functions: learning from noisy labels without knowing noise rates. In: Proceedings of International Conference on Machine Learning, 2020
33 Xia X, Liu T, Han B, et al. Robust early-learning: hindering the memorization of noisy labels. In: Proceedings of International Conference on Learning Representations, 2020
34 Bai Y, Yang E, Han B, et al. Understanding and improving early stopping for learning with noisy labels. In: Proceedings of Conference on Neural Information Processing Systems, 2021
35 Xu Y, Cao P, Kong Y, et al. L_DMI: an information-theoretic noise-robust loss function. In: Proceedings of Conference on Neural Information Processing Systems, 2019
36 Kim Y, Yun J, Shon H, et al. Joint negative and positive learning for noisy labels. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021
37 Zhang Z, Sabuncu M R. Generalized cross entropy loss for training deep neural networks with noisy labels.

In: Proceedings of Conference on Neural Information Processing Systems, 2018
38 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, 2020
39 Zhang Y, Zheng S, Wu P, et al. Learning with feature-dependent label noise: a progressive approach. In: Proceedings of International Conference on Learning Representations, 2021
40 Northcutt C, Jiang L, Chuang I. Confident learning: estimating uncertainty in dataset labels. J Artif Intell Res, 2021, 70: 1373–1411

A joint training framework for learning with noisy labels

Qi WEI[1], Haoliang SUN[1*], Yuling MA[2] & Yilong YIN[1*]

1. *School of Software, Shandong University, Jinan* 250101*, China*;
2. *School of Computer Science and Technology, Shandong Jianzhu University, Jinan* 250101*, China*
* Corresponding author. E-mail: haolsun@sdu.edu.cn, ylyin@sdu.edu.cn

Abstract Sample selection and label correction are two effective strategies for learning with noisy labels (LNL). Both of these strategies have limitations for LNL tasks. Sample selection strategies usually ignore discriminative information in discarded samples, thereby degrading the performance of the algorithm. Label correction strategies commonly leverage self-labeling techniques, resulting in notorious error accumulation. In this paper, we propose a new learning framework that combines sample selection with label correction. Specifically, a novel selection criterion is designed to update the selected set online. Compared with the existing criterion, our proposal retains more boundary samples for the decision and can improve the generalization ability of the learning algorithm. For the label correction phase, we designed a joint label correction function. Compared with the conventional self-label strategy, a multiview label correction is proposed to combine multiple feature space views, which can alleviate the effect of error accumulation. In addition, we propose a contrastive learning regularization term to enhance the learning of modified label quality and model representation from the feature perspective. Our framework achieves the new state of the art on four challenging benchmarks, demonstrating its great effectiveness in LNL.

Keywords learning with noisy labels, sample selection, label correction, contrastive learning