中国科学: 信息科学 2024 年 第 54 卷 第 3 期: 623–637
SCIENTIA SINICA Informationis
论文

稳健选择伪标注的混合式半监督学习

郭兰哲, 李宇峰*

计算机软件新技术国家重点实验室 (南京大学), 南京 210023
* 通信作者. E-mail: liyf@nju.edu.cn

收稿日期: 2022–11–04; 修回日期: 2023–03–24; 接受日期: 2023–04–06; 网络出版日期: 2024–03–12

国家自然科学基金 (批准号: 62176118, 61921006) 和中国人工智能学会 – 华为 MindSpore 学术奖励基金资助项目

摘要 半监督学习旨在数据标注缺乏的情形下利用无标注数据提升学习性能, 是重要的机器学习范式. 尽管不少研究报道表明半监督学习取得了优异的性能表现, 然而其在面临诸多实践任务时仍存在伪标注质量判断困难、超参数选择敏感、理论指导缺乏等瓶颈. 针对上述挑战, 本文提出一种稳健选择伪标注的混合式半监督学习方法, 通过综合利用模型预测结果之间的分歧自适应地判断伪标注质量, 无需预设超参数, 显著提升了半监督学习的稳健性. 本文在理论上证明了新方法的错误率随训练轮数的增加而显著下降. 实验验证了本文方法较主流技术取得了明显的性能提升, 例如, 相较于在 CIFAR-10 数据集中表现最优的半监督学习技术 FixMatch, 新方法的分类错误率下降了 11% 以上, 在更具挑战的 STL-10 数据集中分类错误率下降了 18.8%.

关键词 机器学习, 深度学习, 半监督学习, 伪标注, 稳健性

1 引言

机器学习, 尤其是深度学习, 近年来在多种监督学习任务中取得了巨大的成功 [1,2], 这些成功的一个重要前提条件是存在大量有标注数据用于模型训练. 然而在很多现实应用中, 获取数据标注往往是非常困难的, 需要耗费大量的人力、物力和财力 [3,4]. 例如, 在医学影像分析任务中, 标注影像中的病灶需要医学专家的领域知识 [5]; 在基因识别任务中, 标注基因的类别需要昂贵的专业设备 [6]; 在图像分类任务中, 获取大量图片的类别标注需要聘请大量数据标注人员 [7]. 相比之下, 在许多现实任务中, 无标注数据的获取往往是非常容易的. 因此, 亟需研究能够利用无标注数据提升学习性能的机器学习技术.

半监督学习 (semi-supervised learning, SSL) 是一种能够有效利用无标注数据提升学习性能, 降低机器学习对大规模数据标注需求的强大学习范式. 半监督学习的发展经历了统计半监督学习和深度半监督学习两个重要时期. 统计半监督学习时期的代表方法包括生成式半监督学习、半监督支持向

引用格式: 郭兰哲, 李宇峰. 稳健选择伪标注的混合式半监督学习. 中国科学: 信息科学, 2024, 54: 623–637, doi: 10.1360/SSI-2022-0421
Guo L Z, Li Y F. Robust pseudo-label selection for holistic semi-supervised learning (in Chinese). Sci Sin Inform, 2024, 54: 623–637, doi: 10.1360/SSI-2022-0421

量机、图半监督学习和基于分歧的半监督学习 [2,3]. 深度半监督学习时期的代表方法包括熵最小化方法, 鼓励模型在无标注数据上输出熵值较小 (即置信度较高) 的预测结果, 从而使所得决策边界处于数据密度较低的区域 [8,9]; 一致性正则方法, 鼓励模型在相似的样本上产生相似的预测结果 [10~15]. 近年来, 有大量研究人员指出, 相比于只采用一种无标注数据利用策略的半监督学习方法, 同时考虑多种学习策略的混合式半监督学习方法往往可以取得更好的效果 [16], 代表性方法例如 MixMatch [16], ReMixMatch [17], FixMatch [18] 等, 通过综合考虑熵最小化和一致性正则方法, 在多种半监督学习基准数据集, 如 CIFAR-10 [19], SVHN [20] 中可以取得更优的性能.

尽管大量研究报道混合式半监督学习取得了优异的性能表现, 然而现有方法仍然面临伪标注质量判断困难、超参数选择敏感、理论指导缺乏等瓶颈, 影响其在更多现实任务中的应用. 现有混合式半监督学习的基本做法: 一方面利用模型预测结果为无标注样本生成标注, 即伪标注 (pseudo-label), 并通过选择置信度较高的伪标注实现熵最小化; 另一方面对无标注样本进行增广, 最小化模型在原始样本和增广样本之间预测结果的不一致性, 从而实现一致性正则. 此类方法的一个关键问题是如何为无标注数据赋予正确的伪标注. 针对该问题, 目前主要采用一些基于超参数的方法, 例如, MixMatch 方法利用锐化 (sharpen) 函数降低模型预测结果的熵值, 锐化函数的温度 (temperature) 超参数需要用户预先设定; FixMatch 方法利用置信度阈值选择置信度较高的伪标注, 阈值超参数需要用户预先指定. 此类做法存在严重的性能瓶颈: 一方面超参数的设计需要一定的经验和领域知识, 不合适的超参数会严重影响模型学习性能; 另一方面超参数需要训练开始前预先指定, 然而模型的学习状态是随着训练轮数不断变化的, 伪标注的选择也要自适应调整, 例如, 在训练初始阶段, 模型预测结果的置信度普遍偏低, 随着训练过程的持续, 模型预测结果的置信度会逐渐增加, 基于用户预先设定的置信度阈值来选择伪标注显然是不合理的. 这些瓶颈导致半监督学习的稳健性难以保证, 如何稳健自适应地选择伪标注已经成为了提升半监督学习性能的关键问题 [21~23].

为了解决上述问题, 本文提出了一种稳健选择伪标注的新型混合式半监督学习方法 TriMatch, 该方法同时训练多个学习器, 通过综合利用多学习器预测结果之间的分歧判断伪标注的质量, 实现了不依赖超参数, 从数据中自适应学习如何选择伪标注的效果. 具体而言, 在训练过程中, TriMatch 方法采用弱增广、强增广的数据增广策略为无标注数据生成 3 个视图 (即 3 个数据增广版本), 然后基于每个视图各自训练学习器, 如果其中两个学习器在无标注样本的弱增广版本上产生的预测结果一致, 则为该样本的强增广版本赋予伪标注, 并将伪标注样本提供给第 3 个学习器作为新增的有标注样本以监督学习的方式进行模型更新. 理论分析证明了 TriMatch 方法学习过程中模型的分类误差会随模型训练轮数的增加逐渐下降, 实验验证了 TriMatch 方法在 4 种半监督学习基准数据集 CIFAR-10, CIFAR100, SVHN 和 STL-10 以及不同数量的标注数据下均取得了最优的性能. 相比现有最先进的半监督学习方法 FixMatch, 在 CIFAR-10 数据集中, TriMatch 方法分类错误率下降 11.8%, 在更具挑战的 STL-10 数据集中, TriMatch 方法分类错误率下降 18.8%.

本文后续组织结构如下: 第 2 节介绍半监督学习相关研究工作; 第 3 节介绍现有混合式半监督学习方法以及本文提出的稳健选择伪标注的新型混合式半监督学习方法; 第 4 节通过理论分析证明所提方法的有效性; 第 5 节进行实验分析论证所提方法的先进性; 最后, 第 6 节对本文进行总结并展望未来研究工作.

2 相关工作

本节介绍与本文研究相关的半监督学习工作, 包括基于分歧的方法、熵最小化方法、一致性正则

方法、混合式半监督学习方法和稳健半监督学习, 更多半监督学习研究可以参考文献 [2,3,21].

基于分歧的方法采用多个学习器同时进行训练, 学习器之间的 "分歧" 对无标注数据的利用至关重要. 协同训练 (co-training) [24] 是此类方法的代表, 协同训练针对多视图 (multi-view) 数据设计, 假设数据拥有两个充分且条件独立的视图, 在此情形下, 首先在每个视图上基于有标注数据分别训练一个学习器, 然后让每个学习器挑选自己置信度最高的无标注样本赋予伪标注, 并将伪标注样本提供给另一个学习器作为新增的有标注样本用于训练更新, 这个过程不断迭代进行, 直到两个分类器都不再发生变化, 或达到预先设定的轮数. 值得一提的是, 即使数据不具有多个视图, 只要多个学习器之间具有显著的分歧便可通过互相提供伪标注样本的方式来提升泛化性能 [25], 例如, 通过不同的学习算法、不同的数据采样、不同的参数设置等都可以产生学习器之间的分歧. 三体训练 (tri-training) [26,27] 也是此类方法的重要代表, 其同时训练 3 个学习器, 通过 "少数服从多数" 的准则来产生伪标注样本, 并将学习器进行集成. 该方法将半监督学习与集成学习 (ensemble learning) 这两个长期独立的发展领域联系起来, 证明了将学习器集成起来更有助于提升半监督学习性能 [28]. 本文工作与基于分歧的半监督学习主要区别在于本文工作是一种新型混合式半监督学习方法, 在考虑分歧的同时, 综合考虑了熵最小化、一致性正则策略, 从而实现更优的性能提升.

熵最小化方法基于低密度分隔假设 (low-density separation assumption), 认为模型的决策边界应该穿过数据低密度的区域. 为了实现该目标, Grandvalet 和 Bengio [8] 提出熵最小化正则项, 显式地优化模型 f 在无标注样本 $\boldsymbol{x}$ 上的预测结果 $f(\boldsymbol{x})$ 的熵值, 并从理论上分析熵最小化有助于产生位于低密度区域的决策边界; Lee [9] 进一步提出 Pseudo-Labeling 方法, 在模型训练过程中将预测置信度较高的预测结果作为伪标注赋予无标注样本, 并将该无标注样本及其伪标注添加到有标注数据集中, 以监督学习的方式进行模型训练, 由于置信度较高的预测结果具有更低的熵值, 因此该方法可以隐式地实现熵最小化的目标.

一致性正则方法基于平滑假设 (smooth assumption), 认为相似的样本应当具有相似的标注. 为了实现该目标, 将数据增广策略应用至无标注样本, 鼓励模型在原始无标注样本的不同增广版本上产生一致的预测结果. 形式化而言, 对于无标注样本 $\boldsymbol{x}$, 一致性正则最小化损失函数: $\|f(\mathrm{Augment}(\boldsymbol{x})) - f(\mathrm{Augment}'(\boldsymbol{x}))\|_2^2$, 其中 $\mathrm{Augment}(\boldsymbol{x})$ 和 $\mathrm{Augment}'(\boldsymbol{x})$ 表示对样本 $\boldsymbol{x}$ 进行不同的数据增广, 例如, 对于图像数据, 常用的数据增广策略包括旋转、裁剪、随机翻转等. Π-Model 方法 [11] 直接将上式与标注数据中的交叉熵损失相结合作为模型的损失函数进行优化; Temporal Ensembling 方法 [12] 引入集成学习的策略, 最小化模型在当前无标注样本上的预测结果和多轮预测集成结果之间的均方误差; Mean Teacher 方法 [13] 进一步将优化目标中其中一项的模型输出用基于指数移动平均值 (exponential moving average, EMA) 策略得到的集成模型输出替代; VAT 方法 [14] 将数据增广部分替换为对抗的数据增广, 即寻找能够最大化输出分布变化的特征增广; UDA 方法 [15] 指出对于不同的数据类型, 如图像、文本、语音等, 应结合先验知识采用领域相关的增广策略以进一步提升半监督学习性能.

混合式半监督学习方法认为相比于单一的熵最小化方法和一致性正则方法, 综合考虑两种方法更有助于性能提升. 代表方法如下: MixMatch [16], 首先对无标注样本做多次数据增广, 得到模型在多个增广样本上的平均预测结果, 并引入锐化函数产生熵最小化的猜测标注 (guessing label), 然后利用 MixUp [29] 策略混合标注数据和无标注数据, 并最小化模型在无标注样本上的预测结果与猜测标注之间的均方误差; ReMixMatch [17] 在 MixMatch 的基础上提出标注分布对齐和强增广、弱增广的数据增广策略, 进一步提升了 MixMatch 方法的性能; FixMatch [18] 利用置信度阈值选择模型在弱增广样本上产生的置信度高的预测结果作为伪标注, 并将强增广样本和对应的伪标注作为标注数据以监督学习的方式进行训练. 大量研究指出, 混合式半监督学习方法在多种半监督学习基准任务中取得了目前最

先进的性能表现. 然而, 如上文所述, 现有混合式半监督学习存在伪标注质量判断困难、超参数选择敏感、理论指导缺乏等瓶颈, 限制其在真实任务中的广泛应用.

稳健半监督学习研究如何保障利用无标注数据进行学习的模型性能稳健性. 现有稳健半监督学习研究主要从数据质量、模型构建、评价指标 3 个方面展开 [30]. 例如, 在数据质量层面, Oliver 等 [21] 通过对多种半监督学习算法进行评估, 指出无标注数据分布偏移会严重影响半监督学习稳健性; Guo 等 [31] 针对无标注数据包含未见类的问题进行研究, 提升半监督学习在类别增量场景下的稳健性. 在模型构建层面, Li 等 [5] 提出基于模型集成的稳健半监督模型构建方法, 通过最大化最坏情况下的性能增益避免半监督学习性能退化; Wei 等 [32] 提出稳健模型选择方法, 降低半监督模型选择不确定性造成的风险. 在评价指标层面, Li 等 [33] 通过最大化半监督学习模型在多种评价指标下的综合性能, 提升半监督学习应对不同评价指标的稳健性. 本文工作主要关注模型构建层面, 研究如何提升半监督学习模型训练过程中对伪标注选择的稳健性.

3 本文工作

本节介绍半监督学习的问题设定、相关的符号及其含义, 现有混合式半监督学习方法以及本文提出的新型混合式半监督学习方法 TriMatch.

3.1 问题设定

首先对半监督学习的基本概念做简单介绍. 在半监督学习任务中, 训练数据由两部分组成, 包含 N 个有标注样本的有标注数据集 $\mathcal{D}^l = \{(\boldsymbol{x}_1^l, \boldsymbol{y}_1^l), \ldots, (\boldsymbol{x}_N^l, \boldsymbol{y}_N^l)\}$, 以及包含 M 个无标注样本的无标注数据集 $\mathcal{D}^u = \{\boldsymbol{x}_1^u, \ldots, \boldsymbol{x}_M^u\}$, 通常来说, 无标注数据的数量远大于有标注数据, 即 $M \gg N$. $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$, $\boldsymbol{y} \in \mathcal{Y} = \{0,1\}^K$, 其中 $\mathcal{X}$ 表示样本的特征空间, $\mathcal{Y}$ 表示样本的类别空间, d 表示特征空间的维度, K 表示类别空间的维度. 半监督学习的目标是通过拟合训练数据 $\mathcal{D}_l$ 和 $\mathcal{D}_u$ 学习从特征空间到类别空间的映射模型 $f(\boldsymbol{x};\theta) : \{\mathcal{X};\Theta\} \to \mathcal{Y}$, 其中 $\theta \in \Theta$ 表示模型 f 的参数, 使其能够在未见的测试数据上取得良好的泛化性能.

半监督学习的损失函数 $\mathcal{L}$ 通常由两部分组成: 监督损失 $\mathcal{L}_s$ 和无监督损失 $\mathcal{L}_u$, 即 $\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u$, 其中 $\lambda_u > 0$ 为两个损失之间的平衡超参数. 监督损失通常来说可以直接计算模型在有标注数据上的预测结果与样本真实标注之间的交叉熵损失 (cross-entropy loss), 如下所示:

$$\begin{aligned} \mathcal{L}_s &= \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{y}_i, f(\boldsymbol{y}|\boldsymbol{x}_i;\theta)) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} -y_{i,k}\log f(\boldsymbol{y}=k|\boldsymbol{x};\theta), \end{aligned} \tag{1}$$

其中 $f(\boldsymbol{y}|\boldsymbol{x};\theta) \in [0,1]^K$ 表示模型 f 对于输入样本 $\boldsymbol{x}$ 的预测概率, $H(\cdot,\cdot)$ 表示交叉熵函数.

不同的无监督损失 $\mathcal{L}_u$ 的构造方式将产生不同的半监督学习方法. 接下来将以 FixMatch 为例介绍现有混合式半监督学习方法如何构造无监督损失函数 $\mathcal{L}_u$.

3.2 现有混合式半监督学习

FixMatch 作为代表性的混合式半监督学习方法, 由于其简单的步骤和优异的性能, 取得了大量的关注. 研究指出 FixMatch 在多个半监督学习基准数据集, 如 CIFAR, SVHN 中取得了目前最先进的

性能 [18], 例如, 在图像分类任务数据集 CIFAR-10 中, FixMatch 只利用 10% 的有标注数据便可以取得与利用所有数据标注的全监督学习相近的性能.

FixMatch 首先对无标注样本采用弱增广和强增广策略进行数据增广, 然后利用模型在弱增广样本上的预测结果为对应的强增广样本赋予伪标注, 并且当且仅当伪标注的预测置信度高于某个用户预设的置信度阈值时, 该伪标注才会被选择. 在得到伪标注之后, FixMatch 以监督学习的方式优化模型在强增广样本上的预测结果和伪标注之间的交叉熵损失.

具体而言, 在每个训练轮次, 给定一批 B 个有标注样本 $\{(\boldsymbol{x}_b^l, \boldsymbol{y}_b^l) : b \in (1, \ldots, B)\}$ 和 μB 个无标注样本 $\{\boldsymbol{x}_b^u : b \in (1, \ldots, \mu B)\}$, 其中 μ 表示有标注数据和无标注数据之间批样本数量的比例. 对于无标注样本 $\boldsymbol{x}_b^u$, 首先对样本进行弱增广, 并得到模型在对应弱增广样本上的预测概率:

$$\boldsymbol{q}_b = f(\boldsymbol{y}|\alpha(\boldsymbol{x}_b^u); \theta), \tag{2}$$

其中 $\alpha(\cdot)$ 表示弱增广操作.

根据模型预测结果 $\boldsymbol{q}_b$, 选择预测概率最大的类别作为该样本的伪标注, 如下所示:

$$\hat{\boldsymbol{y}}_b^u = \arg\max(\boldsymbol{q}_b), \tag{3}$$

其中 $\arg\max$ 操作可以得到一个独热 (one-hot) 概率分布.

然后, 对该样本进行强增广, 并得到模型在强增广样本上的预测概率:

$$\boldsymbol{q}_b^s = f(\boldsymbol{y}|\mathcal{A}(\boldsymbol{x}_b^u); \theta), \tag{4}$$

其中 $\mathcal{A}(\cdot)$ 表示强增广操作.

通过计算模型在强增广样本上预测概率与对应伪标注 $\hat{\boldsymbol{y}}_b^u$ 之间的交叉熵可以得到模型在该样本上的损失为

$$H(\hat{\boldsymbol{y}}_b^u, f(\boldsymbol{y}|\mathcal{A}(\boldsymbol{x}_b^u); \theta)). \tag{5}$$

对于伪标注选择的问题, FixMatch 仅保留模型在弱增广样本上预测置信度高于某个预设阈值 τ 的伪标注, 例如 $\tau = 0.95$, 所以, FixMatch 的无监督损失函数可以写作

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{I}(\max(\boldsymbol{q}_b) \geqslant \tau) H(\hat{\boldsymbol{y}}_b^u, f(\boldsymbol{y}|\mathcal{A}(\boldsymbol{x}_b^u); \theta)), \tag{6}$$

其中 $\mathbb{I}(\cdot)$ 为指示函数, 成立则取值为 1, 反之则为 0.

从上述步骤可以看出, 现有混合式半监督学习方法一方面通过选择置信度较高的模型预测结果作为无标注样本的伪标注, 实现熵最小化; 另一方面鼓励模型在样本的多个增广版本之间输出一致的预测结果, 实现一致性正则. 但上述过程依赖超参数 τ 的选择, 若超参数选择不当, 则会影响伪标注的质量, 从而导致半监督学习性能不稳健.

3.3 稳健选择伪标注的混合式半监督学习

尽管研究报道指出上述以 FixMatch 为代表的混合式半监督学习方法通过同时考虑熵最小化和一致性正则, 在多种基准数据集上取得了优异的性能, 然而, 现有混合式半监督学习方法仍然存在严重局限: (1) 现有方法根据预先设定的某个固定超参数阈值来选择伪标注, 没有考虑到模型的预测置信度分布会随训练过程的变化而变化, 导致无法在模型训练的各个阶段都能选择到最优的伪标注; (2) 现有方

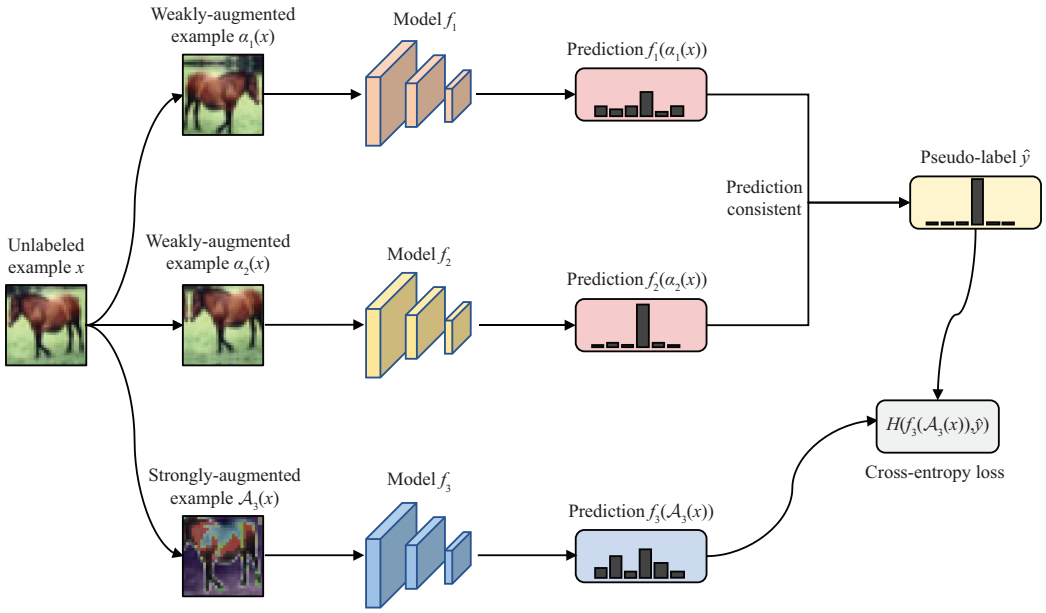

图 1 (网络版彩图) TriMatch 算法框架: TriMatch 采用 3 个模型 $\{f_1, f_2, f_3\}$ 进行训练, 在每次迭代中, 对于训练样本 x, 将两个弱增广的样本 $\alpha_1(x)$ 和 $\alpha_2(x)$ 输入给其中两个模型 (例如 f_1, f_2), 如果两个模型输出的最大预测概率属于同一个类, 则将该预测结果转化为该样本的伪标注, 输入给第 3 个模型 (例如 f_3), 并最小化该伪标注和模型在强增广样本 $\mathcal{A}(x)$ 上的预测结果之间的交叉熵损失

Figure 1 (Color online) Diagram of TriMatch. TriMatch adopts three models. At each iteration, two weakly-augmented training examples (the top two) are fed into two models to obtain predictions. When the two models assign the maximum probability to the same class, the prediction is converted to a one-hot pseudo-label. Then, we compute the model's prediction for a strong augmentation of the same image (bottom). The model is trained to make its prediction on the strongly-augmented version match the pseudo-label via a cross-entropy loss

法依赖用户指定的超参数, 而超参数的设计需要一定的领域和经验知识, 不合适的超参数选择将会导致模型性能显著下降. 这些缺陷导致现有半监督学习方法缺乏稳健性, 影响了半监督学习的广泛应用. 为了解决上述难题, 本文在综合考虑熵最小化和一致性正则的基础上, 进一步考虑基于分歧的无标注数据利用策略, 提出了稳健选择伪标注的新型混合式半监督学习方法 TriMatch. TriMatch 在训练过程中引入多个模型, 通过综合利用多个模型预测结果之间的分歧判断伪标注的质量, 突破现有方法伪标注选择需要设置超参数的局限, 同时可以自适应地在模型训练的各个阶段选择合适的伪标注, 解决现有方法未考虑模型预测分布变化的问题, 提升伪标注选择的稳健性.

具体而言, TriMatch 同时采用 3 个模型 $\{f_1, f_2, f_3\}$ 进行训练, 在每一轮迭代过程中, 给定一批 B 个有标注样本 $\{(\boldsymbol{x}_b^l, \boldsymbol{y}_b^l) : b \in (1, \ldots, B)\}$ 和 μB 个无标注样本 $\{\boldsymbol{x}_b^u : b \in (1, \ldots, \mu B)\}$, 对于每个有标注样本 $\boldsymbol{x}_b^l$, 首先进行一次弱增广操作, 获得每一个模型 $f_i (i \in \{1, 2, 3\})$ 在对应弱增广样本 $\alpha(\boldsymbol{x}_b^l)$ 上的预测概率, 并计算模型预测概率 $f_i(\boldsymbol{y}|\alpha(\boldsymbol{x}_b^l); \theta_i)$ 和样本真实标注 $\boldsymbol{y}_b^l$ 之间的交叉熵损失作为该模型的监督损失 $\mathcal{L}_{s,i}$:

$$\mathcal{L}_{s,i} = \frac{1}{B} \sum_{b=1}^{B} H(\boldsymbol{y}_b^l, f_i(\boldsymbol{y}|\alpha(\boldsymbol{x}_b^l); \theta_i)), \quad i \in \{1, 2, 3\}. \tag{7}$$

对于每个无标注样本 x_u^b, 为了增加模型之间的分歧, 采用 3 种不同的弱增广和强增广操作, 然后

算法 1 TriMatch 方法伪代码

Require: 有标注数据 $\mathcal{X}=\{(\boldsymbol{x}_b^l,\boldsymbol{y}_b^l):b\in(1,\ldots,B)\}$, 无标注数据 $\mathcal{U}=\{\boldsymbol{x}_b^u:b\in(1,\ldots,\mu B\}$, 无监督损失权重 λ_u;

for $i=1$ **to** 3 **do**

  $\mathcal{L}_{s,i}=\frac{1}{B}\sum_{b=1}^{B}H(\boldsymbol{y}_b^l,f_i(\boldsymbol{y}|\alpha(\boldsymbol{x}_b^l);\theta_i))$; // 计算模型在有标注数据上的交叉熵损失;

end for

for $b=1$ **to** μB **do**

  for $i=1$ **to** 3 **do**

    $\boldsymbol{q}_{b,i}=f_i(\boldsymbol{y}|\alpha_i(\boldsymbol{x}_b^u);\theta_i)$; // 第 i 个模型在第 b 个弱增广无标注样本上的预测概率;

    $\hat{\boldsymbol{y}}_{b,i}=\arg\max(\boldsymbol{q}_{b,i}),i\in\{1,2,3\}$; //根据模型预测概率得到独热编码的伪标注;

  end for

  for $i=1$ **to** 3 **do**

    $\text{mask}_{b,i}=\mathbb{I}(\hat{\boldsymbol{y}}_{b,j}=\hat{\boldsymbol{y}}_{b,k}),\{j,k\}=\{1,2,3\}\setminus i$; // 判断其余两个模型在弱增广样本上的预测标注是否一致;

  end for

end for

for $i=1$ **to** 3 **do**

  $\mathcal{L}_{u,i}=\frac{1}{\mu B}\sum_{b=1}^{\mu B}\text{mask}_{b,i}H(\hat{\boldsymbol{y}}_{b,j},f_i(\boldsymbol{y}|\mathcal{A}_i(\boldsymbol{x}_b^u);\theta_i)),\ j\in\{1,2,3\}\setminus i$; // 其余两个模型预测结果一致时, 选择该伪标注作为强增广样本的真实标注并计算交叉熵损失;

end for

计算总体损失 $\sum_{i=1}^{3}\mathcal{L}_{s,i}+\lambda_u\mathcal{L}_{u,i}$ 并更新模型.

获得第 i 个模型在第 i 个弱增广样本 $\alpha_i(\boldsymbol{x}_b^u)$ 的上预测概率 $\boldsymbol{q}_{b,i}$, 如下所示:

$$\boldsymbol{q}_{b,i}=f_i(\boldsymbol{y}|\alpha_i(\boldsymbol{x}_b^u);\theta_i). \tag{8}$$

然后采用预测概率最大的类别作为第 i 个模型在样本 x_u^b 上产生的伪标注, 即

$$\hat{\boldsymbol{y}}_{b,i}=\arg\max(\boldsymbol{q}_{b,i}). \tag{9}$$

对于第 i 个模型, 如果其余两个模型在无标注样本 $\boldsymbol{x}_b^u$ 上产生的伪标注一致, 则选择该伪标注:

$$\text{mask}_{b,i}=\mathbb{I}(\hat{\boldsymbol{y}}_{b,j}=\hat{\boldsymbol{y}}_{b,k}),\quad \{j,k\}=\{1,2,3\}\setminus i. \tag{10}$$

然后计算该模型在对应强增广样本 $\mathcal{A}_i(\boldsymbol{x}_b^u)$ 上的预测结果和伪标注之间的交叉熵损失, 最终得到的无监督损失函数如下所示:

$$\mathcal{L}_{u,i}=\frac{1}{\mu B}\sum_{b=1}^{\mu B}\text{mask}_{b,i}H(\hat{\boldsymbol{y}}_{b,j},f_i(\boldsymbol{y}|\mathcal{A}_i(\boldsymbol{x}_b^u);\theta_i)),\quad j\in\{1,2,3\}\setminus i. \tag{11}$$

模型训练的总体损失为每个模型的监督损失和无监督损失之和, 即 $\mathcal{L}=\sum_{i=1}^{3}\mathcal{L}_{s,i}+\lambda_u\mathcal{L}_{u,i}$. 在模型预测过程中, 采用 3 个模型预测结果的加权平均结果作为最终概率以产生预测标注.

图 1 展示了 TriMatch 方法中无监督损失的计算过程, 算法 1 总结了 TriMatch 方法的伪代码. 从上述步骤可以看出, TriMatch 方法引入多个模型, 通过考虑模型预测结果之间的分歧, 有效避免了超参数的设置, 自适应地考虑了模型在训练过程中的状态变化.

4 理论分析

本节对 TriMatch 方法进行理论分析, 证明模型错误率将随训练轮数的增加逐渐下降.

在模型训练过程中, TriMatch 根据模型预测结果为无标注样本赋予伪标注并以监督学习的方式进行优化, 在此过程中不可避免地会加入错误的伪标注, 即模型训练数据中包含噪声样本, 根据概率近似正确 (probably approximately correct, PAC) 学习理论, 当训练数据中包含噪声样本时, 有如下结果 [34].

引理1 令 $\mathcal{F}$ 表示模型假设空间, η 表示训练数据中噪声样本的比例 ($\eta < 0.5$), 如果训练样本数量 m 满足

$$m \geqslant \frac{2}{\epsilon^2(1-2\eta)^2} \ln\left(\frac{2|\mathcal{F}|}{\delta}\right), \tag{12}$$

则对于从训练数据中学习得到的模型 f 和假设空间中最优的模型 f^*, 我们有

$$\Pr[d(f, f^*) \geqslant \epsilon] \leqslant \delta, \tag{13}$$

其中 $d(f, f^*)$ 表示模型 f 相比最优模型 f^* 错误率, $\epsilon > 0$, $\delta < 1$.

引理 1 给出了当训练数据中包含标注噪声样本时, 训练模型所需的样本数量. 受文献 [26] 启发, 基于引理 1 可以得到如下结论.

定理1 对于任意模型 f_i, 令 L^t 表示其余两个模型 f_j 和 f_k 预测一致的无标注样本数量, e^t 表示模型 f_j 和 f_k 同时预测错误的概率, 如果 $L^t \geqslant L^{t-1}$ 且 $e^t L^t \leqslant e^{t-1} L^{t-1}$, 那么模型 f_i 的错误率将随训练轮数的增加而下降, 即

$$\epsilon^t \leqslant \epsilon^{t-1}. \tag{14}$$

证明 在模型训练的第 t 轮, 模型 f_i 的训练样本总数量为 m^t, 其中包括 N 个标注样本和另外两个模型 f_j 和 f_k 预测结果一致的无标注样本, 即 $m^t = N + L^t$. 模型 f_j 和 f_k 预测结果一致并且预测结果错误的样本数量为 $e^t L^t$, 则在当前轮次中模型 f_i 的训练数据中所包含的噪声样本比例为

$$\eta^t = \frac{e^t L^t}{N + L^t}. \tag{15}$$

根据引理 1 可知, 若要使模型 f_i 的错误率 ϵ 逐渐减小, 则 $m^t(1-2\eta^t)^2$ 需逐渐增加.

基于式 (15), 可知

$$m^t(1-2\eta^t)^2 = (N + L^t)\left(1 - \frac{2e^t L^t}{N + L^t}\right)^2. \tag{16}$$

由式 (16) 可知, 若要使 $m_t(1-2\eta^t)^2$ 逐渐增加, 需满足

$$(N + L^t)\left(1 - \frac{2e^t L^t}{N + L^t}\right)^2 \geqslant (N + L^{t-1})\left(1 - \frac{2e^{t-1} L^{t-1}}{N + L^{t-1}}\right)^2. \tag{17}$$

当以下两个条件满足时, 式 (17) 显然成立:

$$L^t \geqslant L^{t-1}, \tag{18}$$

$$e^t L^t \leqslant e^{t-1} L^{t-1}. \tag{19}$$

定理 1 表明, 如果在模型训练过程中每一轮选择的伪标注数量逐渐增加, 同时选择错误的伪标注数量逐渐减少, 即 $L^t \geqslant L^{t-1}$, $e^t L^t \leqslant e^{t-1} L^{t-1}$, 则模型的错误率将逐渐减小. 我们通过实验证明这两个条件在实践中均是容易满足的, 具体而言, 在半监督学习基准数据集 CIFAR-10 上进行实验, 每一轮选择的伪标注总数量 L^t 和选择错误的伪标注总数量 $e^t L^t$ 随训练轮数的变化曲线如图 2 所示, 从结果中可以看出选择的伪标注总数量随迭代轮数增加不断增加, 选择错误的伪标注数量随迭代轮数增加不断减少, 这符合定理要求的假设条件, 为本文方法提供了理论保障.

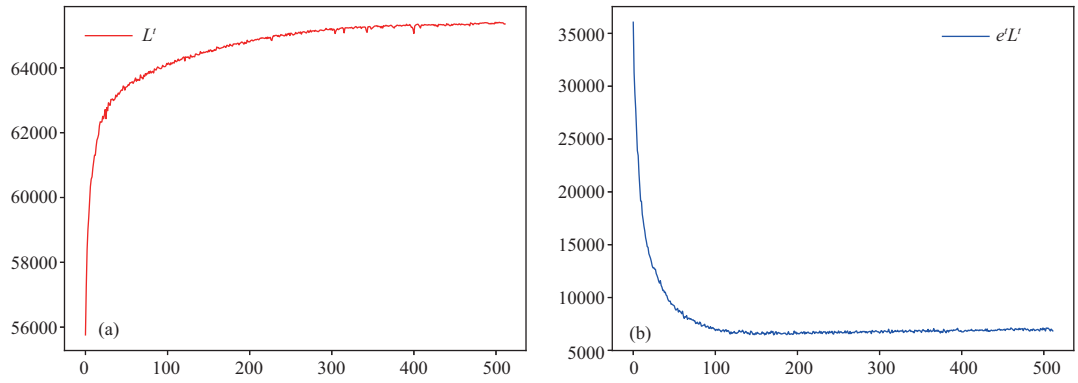

图 2 (网络版彩图) 每一轮选择的 (a) 伪标注数量和 (b) 选择错误的伪标注数量随模型训练轮数的变化情况

Figure 2 (Color online) The number of selected pseudo-labels and wrongly selected pseudo-labels through training iterations

5 实验验证

本节将通过多个基准数据集上的实验验证所提新型半监督学习方法的有效性.

5.1 数据集

选取多个半监督学习的基准数据集, 包括 CIFAR-10, CIFAR-100 [19], SVHN [20] 和 STL-10 [35], 用以评估各种半监督学习方法的性能表现. CIFAR 数据集包括 50000 个训练样本和 10000 个测试样本, 每个样本为 32×32 像素的图像, 其中 CIFAR-10 数据集具有 10 个类别, 每个类别包含 6000 个样本, CIFAR-100 数据集具有 100 个类别, 每个类别包含 600 个样本. SVHN 数据集是一个数字识别数据集, 源自谷歌的街景门牌号码, 包括 73257 个训练样本和 26032 个测试样本, 类别总数为 10. STL-10 数据集包含 5000 个标注样本和 100000 个无标注样本, 类别总数为 10, 值得注意的是, STL-10 的无标注数据是类别不平衡的, 其中各类别的样本比例未知.

参考文献 [18], 采取 10 种不同的半监督学习设置进行实验, 对于 CIFAR-10 数据集, 在每个类别随机采样 4, 25, 400 个样本作为有标注数据, 其余作为无标注数据; 对于 CIFAR-100 和 SVHN 数据集, 在每个类别随机采样 4, 25, 100 个样本作为有标注数据, 其余作为无标注数据; 对于 STL-10 数据集, 采用原始数据集对标注数据和无标注数据的默认划分.

5.2 对比方法

为了证明本文所提方法的先进性, 选取以下代表性的深度半监督学习方法作为对比方法:

• Pseudo-Labeling [9]: Pseudo-Labeling 是代表性的基于熵最小化的方法, 其在模型的训练过程中利用模型预测结果为无标注样本赋予伪标注, 并将伪标注样本添加至标注数据集中辅助模型训练. 由于伪标注选择的过程考虑了预测置信度, 所以降低了模型预测结果的熵值 [8].

• Π-Model [11]: Π-Model 是代表性的基于一致性正则的方法, 其对无标注样本做随机数据增广, 然后最小化模型在原始样本和数据增广后的样本上输出的预测概率之间的均方误差.

• Mean Teacher [13]: 在 Π-Model 方法的基础上, Mean Teacher 进一步对模型参数进行集成, 通过引入 EMA 操作, 得到多轮训练产生的模型参数的集成, 然后计算当前轮的模型和集成模型在无标注数据上输出的预测概率之间的均方误差作为无监督损失.

• MixMatch [16]: MixMatch 是首个被提出的混合式半监督学习方法, 其首先对无标注样本进行 k 次随机数据增广, 得到模型在 K 个样本输出的预测结果的平均值, 然后通过调整预测概率分布的温度来锐化得到熵值较低的预测结果, 并优化该低熵值结果与模型预测结果之间的均方误差, 此外 MixMatch 还引入 MixUp [30] 策略混合标注和无标注数据进行训练以进一步提升性能. MixMatch 同时考虑了熵最小化和一致性正则, 相比只考虑单一策略的半监督学习方法取得了更优的性能表现.

• ReMixMatch [17]: 在 MixMatch 方法的基础上, ReMixMatch 进一步引入分布对齐 (distribution alignment) 和增广锚定 (augmentation anchoring) 技术, 使无标注数据的预测分布与标注数据分布相匹配, 同时引入弱增广和强增广的数据增广策略, 最小化模型在弱增广无标注样本和强增广无标注样本预测结果之间的均方误差.

• FixMatch [18]: FixMatch 简化了 MixMatch 和 ReMixMatch 方法, 首先基于模型对弱增广样本的预测结果生成独热伪标注, 同时仅当模型预测结果超过某个预先设定的置信度阈值时, 才将该伪标注赋予对应的强增广样本进行模型训练. 目前, FixMatch 在多种半监督学习的基准数据集中取得了最优的性能表现.

5.3 参数设置

我们采用宽残差网络 (wide resnet) [36] 作为分类器网络, 因为该网络架构是目前深度半监督学习方法最常用的模型结构 [21]. 对于 CIFAR-10 和 SVHN 数据集, 网络深度设置为 28, 宽度设置为 2; 对于 CIFAR-100 数据集, 网络深度设置为 28, 宽度设置为 8; 对于 STL-10 数据集, 网络深度设置为 37, 宽度设置为 2. 对于所有的数据集, 我们采用动量随机梯度下降 (stochastic gradient descent with momentum) 算法优化模型参数, 其初始学习率 (learning rate) 设置为 0.03, 动量参数设置为 0.9, 并采用余弦学习率衰减 (cosine learning rate decay) 策略来控制学习率在训练过程中的变化. 对于 CIFAR-10, SVHN 和 STL-10 数据集, 权重衰减 (weight decay) 参数设置为 5×10^{-4}, 对于 CIFAR-100 数据集, 权重衰减参数设置为 1×10^{-3}. 模型总训练轮数为 2^{20}, 每轮随机采样的标注数据规模为 64, 无标注数据为 7×64. 对于弱增广和强增广策略, 参照文献 [18] 中的设置, 其中强增广部分分别采用 RandAugment [37] 和 CTAugment [16] 实现. 参照以往的半监督学习方法, 本文采用了 EMA 策略, 参数设置为 0.999. 实验采用 Mindspore 架构. 对于所有的方法, 随机运行 5 次实验, 并汇报分类误差的均值和标准差.

5.4 实验结果

本小节汇报了所有对比方法在多个数据集上的分类误差, 结果如表 1 和 2 所示. 从实验结果可以看出, 现有的同时考虑熵最小化和一致性正则的半监督学习方法, 如 MixMatch, ReMixMatch 和 FixMatch, 相比单一的熵最小化方法 Pseudo-Labeling 和一致性正则方法 Π-Model 和 Mean Teacher, 可以取得更优的性能表现, 这验证了混合式半监督学习方法的有效性. 而本文方法通过多个模型预测结果之间的分歧判断伪标注质量, 相比现有方法在所有数据集上都取得了更优的性能. 例如, 在 CIFAR-10 数据集中, 当每个类别标注样本为 400 时, 本文方法相比 FixMatch, 分类错误率下降 11.8%, 相比 MixMatch 和 ReMixMatch 分别下降 45.3% 和 25.6%; 在更具挑战的 STL-10 数据集中, 本文方法相比 FixMatch, 分类错误率下降 18.8%, 相比 MixMatch 和 ReMixMatch 分别下降 32.2% 和 14.8%. 这是因为本文方法在熵最小化和一致性正则的基础上, 进一步考虑了基于分歧的无标注数据利用策略, 这些结果充分证明了本文方法的先进性, 也说明了同时考虑 3 种无标注数据利用策略相比现有只考虑两种策略的混合式半监督学习方法可以取得更优的性能, 进一步启发半监督学习算法设计.

表 1 所有对比方法在 CIFAR-10 和 CIFAR-100 上不同标注数据量下的分类误差 (均值 ± 标准差)

Table 1 Classification error (mean ± std) comparison on CIFAR-10 and CIFAR-100 datasets with varying number of labels[a)]

	CIFAR-10				CIFAR-100	
	250 labels	500 labels	1000 labels	4000 labels	2500 labels	10000 labels
Pseudo-Labeling	49.92±0.93	41.23±0.58	30.84±1.18	16.13±0.12	57.30±0.43	36.13±0.42
Π-Model	54.32±1.89	42.09±0.23	31.58±0.84	15.23±0.42	57.28±0.47	37.82±0.11
Mean Teacher	32.32±1.97	41.87±2.79	18.03±0.58	10.51±0.36	54.82±0.64	35.49±0.37
MixMatch	11.13±0.92	9.87±0.34	7.91±0.36	6.36±0.17	40.03±0.42	29.14±0.32
ReMixMatch	6.23±0.54	6.08±0.37	5.82±0.26	4.96±0.09	28.52±0.41	24.13±0.48
FixMatch (CTA)	5.29±0.54	4.96±0.14	4.71±0.20	4.58±0.18	28.98±0.34	23.89±0.12
FixMatch (RA)	5.13±0.26	4.82±0.11	4.60±0.05	4.32±0.05	28.16±0.26	23.43±0.18
Ours (CTA)	5.10±0.45	4.74±0.17	4.51±0.23	4.03±0.16	26.81±0.29	22.41±0.22
Ours (RA)	**4.96±0.31**	**4.62±0.25**	**4.33±0.17**	**3.81±0.10**	**26.72±0.28**	**21.31±0.20**

a) The best results are in bold.

表 2 所有对比方法在 SVHN 和 STL-10 上不同标注数据量下的分类误差 (均值 ± 标准差)

Table 2 Classification error (mean ± std) comparison on SVHN and STL-10 datasets with varying number of labels[a)]

	SVHN				STL-10
	250 labels	500 labels	1000 labels	4000 labels	1000 labels
Pseudo-Labeling	22.67±0.83	18.23±0.37	11.02±0.79	7.32±0.10	26.93±0.79
Π-Model	18.56±0.37	11.47±0.28	8.94±0.23	7.20±0.11	26.15±0.84
Mean Teacher	7.32±1.51	4.93±0.56	3.95±0.34	3.01±0.13	22.42±1.76
MixMatch	4.03±0.34	3.78±0.30	3.69±0.31	3.31±0.12	10.20±0.51
ReMixMatch	3.56±0.48	3.21±0.33	2.94±0.25	2.69±0.11	8.43±0.46
FixMatch (CTA)	2.67±0.58	2.50±0.11	2.38±0.15	2.34±0.11	8.62±0.58
FixMatch (RA)	2.53±0.36	2.41±0.08	2.39±0.11	2.28±0.14	8.52±0.91
Ours (CTA)	2.35±0.27	2.23±0.05	2.28±0.10	2.15±0.06	7.14±0.32
Ours (RA)	**2.26±0.12**	**2.20±0.09**	**2.11±0.07**	**2.03±0.02**	**6.92±0.48**

a) The best results are in bold.

5.5 分析与探讨

本小节通过两个实验设置对所提方法的稳健性和通用性进行更深入的分析和探讨.

稳健性分析. 为了分析为何本文所提新方法在伪标注选择问题中更具稳健性, 本文比较了 TriMatch 和 FixMatch 方法在训练过程中选择的伪标注的情况, 包括每一轮选择的伪标注样本的总数, 选择正确的伪标注样本数量和选择错误的伪标注样本数量, 结果如图 3 所示. 可以看出, 在模型训练初期, 相比 FixMatch 方法, 本文所提 TriMatch 方法倾向于选择更多的无标注样本, 这是因为 FixMatch 采用一个固定的高置信度阈值来选择伪标注, 而在训练初期, 模型预测置信度往往比较低, 所以预测置信度超过该阈值的无标注样本数量比较少, 导致大量预测正确的伪标注没有被选择, 这说明采用固定阈值的方法进行伪标注选择没有考虑到训练过程中模型预测置信度的变化, 导致大量正确的伪标注被遗漏. 此外, 相比 FixMatch, TriMatch 方法在训练过程中能够选择数量更多的正确伪标注, 同时选择错误的伪标注数量并没有显著增加, 这进一步说明了利用多个模型预测结果之间的分歧进行伪标注选择

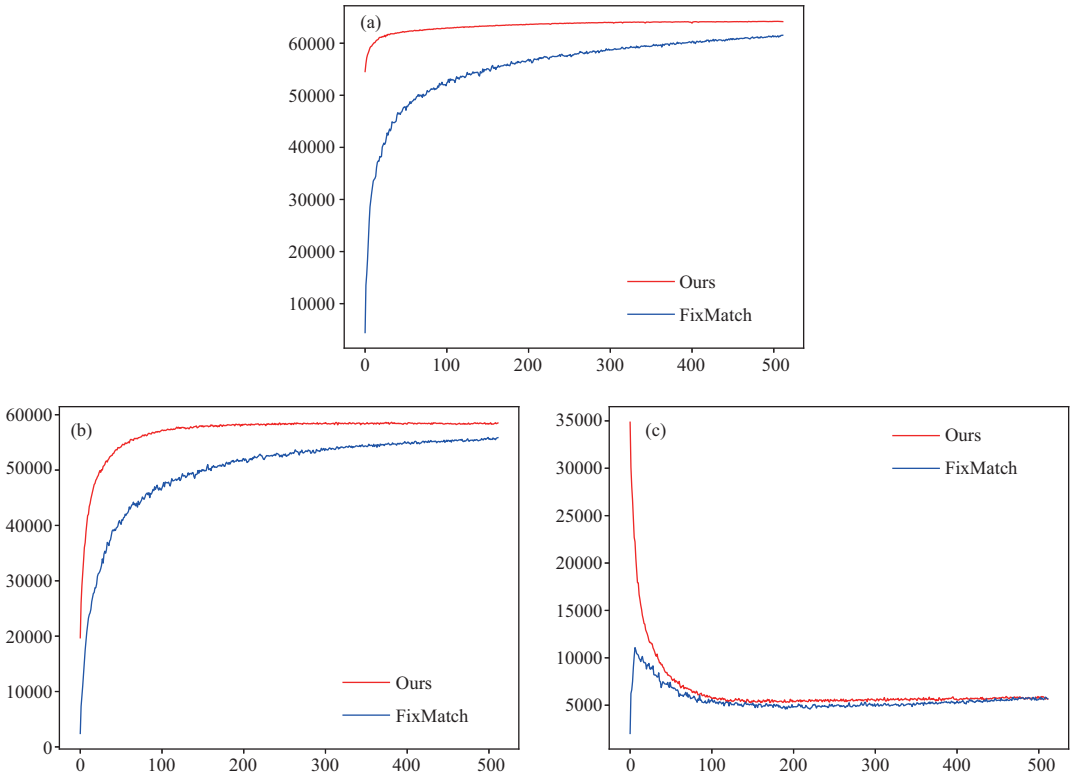

图 3 (网络版彩图) 在模型训练过程中每轮选择的 (a) 伪标注总数量、(b) 选择正确的伪标注数量和 (c) 选择错误的伪标注数量

Figure 3 (Color online) The number of (a) selected pseudo-labels, (b) correct pseudo-labels, and (c) wrong pseudo-labels during the training procedure

相比于利用固定置信度阈值进行伪标注选择性能更加稳健.

通用性分析. 为了进一步说明利用模型之间的分歧来选择伪标注选择这一策略的通用性, 本文还将该策略与 Pseudo-Labeling 方法进行结合. 原始 Pseudo-Labeling 方法利用模型的预测置信度选择伪标注, 在原始 Pseudo-Labeling 方法训练过程中, 引入多个模型, 利用多个模型预测结果之间的分歧辅助伪标注选择, 所得实验结果如表 3 所示. 从结果可以看出, 在 4 个半监督学习基准数据集中, 通过基于分歧的策略选择伪标注, 相比原始 Pseudo-Labeling 方法取得了一致有效的性能提升, 这进一步说明了基于多模型间的分歧进行伪标注选择是一种通用且有效的伪标注选择策略.

6 总结与展望

半监督学习是机器学习领域的基础问题之一, 然而现有半监督学习存在伪标注质量判断困难、超参数选择敏感、理论指导缺乏等瓶颈问题. 针对该问题, 本文创新性地提出一种稳健选择伪标注的新型混合式半监督学习方法 TriMatch, 该方法通过同时训练多个模型, 利用模型预测结果之间的分歧自适应地选择伪标注, 无需预设超参数, 显著缓解了半监督学习中伪标注选择困难这一核心问题. 理论分析证明了本文方法的分类错误率随训练轮数的增加逐步下降, 在多个半监督学习基准数据集上的实验验证了相比现有半监督学习方法, 本文方法取得了更先进的性能表现.

本文方法在静态封闭环境下已经取得了良好的性能, 下一步工作我们考虑将本文方法应用于动态

表 3 本文方法和 Pseudo-Labeling (PL) 方法结合在多个数据集上的分类误差

Table 3 Comparison of classification error for Pseudo-Labeling and our proposal with Pseudo-Labeling[a)]

	CIFAR-10				CIFAR-100	
	250 labels	500 labels	1000 labels	4000 labels	2500 labels	10000 labels
PL	49.92±0.93	41.23±0.58	30.84±1.18	16.13±0.12	57.30±0.43	36.13±0.42
Ours-PL	**46.58±0.87**	**38.23±0.41**	**27.45±0.76**	**15.15±0.31**	**54.38±0.53**	**34.79±0.28**
	SVHN				STL-10	
	250 labels	500 labels	1000 labels	4000 labels	1000 labels	
PL	22.67±0.83	18.23±0.37	11.02±0.79	7.32±0.10	26.93±0.79	
Ours-PL	**19.48±0.74**	**16.86±0.45**	**10.03±0.57**	**6.51±0.32**	**24.87±1.04**	

a) The best results are in bold.

开放环境 [7,38,39] 中, 考虑数据的分布、类别、属性、评价指标都有可能发生变化的情况, 以进一步提升半监督学习稳健应用的范畴.

参考文献

1 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444

2 Zhou Z H. Machine Learning. Beijing: Tsinghua University Press, 2016 [周志华. 机器学习. 北京: 清华大学出版社, 2016]

3 Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning. Cambridge: MIT Press, 2006

4 Zhou Z H. A brief introduction to weakly supervised learning. Natl Sci Rev, 2018, 5: 44–53

5 Li Y F, Guo L Z, Zhou Z H. Towards safe weakly supervised learning. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 334–346

6 Guo L Z, Zhou Z, Shao J J, et al. Learning from imbalanced and incomplete supervision with its application to ride-sharing liability judgment. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Singapore, 2021. 487–495

7 Guo L Z, Zhang Z Y, Jiang Y, et al. Safe deep semi-supervised learning for unseen-class unlabeled data. In: Proceedings of the 37th International Conference on Machine Learning, Vienna, 2020. 3897–3906

8 Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Proceedings of the 17th International Conference on Neural Information Processing Systems, Vancouver, 2004. 529–536

9 Lee D H. Pseudo-Label: the simple and efficient semi-supervised learning method for deep neural networks. In: Proceedings of the 30th International Conference on International Conference on Machine Learning Workshop, Atlanta, 2013

10 Rasmus A, Berglund M, Honkala M, et al. Semi-supervised learning with ladder networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, 2015. 3546–3554

11 Sajjadi M, Javanmardi M, Tasdizen T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, 2016. 1163–1171

12 Laine S, Aila T. Temporal ensembling for semi-supervised learning. In: Proceedings of the 5th International Conference on Learning Representations, Toulon, 2017

13 Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, 2017. 1195–1204

14 Miyato T, Maeda S, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 1979–1993

15 Xie Q Z, Dai Z H, Hovy E, et al. Unsupervised data augmentation for consistency training. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, 2020. 6256–6268

16 Berthelot D, Carlini N, Goodfellow I, et al. MixMatch: a holistic approach to semi-supervised learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, 2019. 5050–5060
17 Berthelot D, Carlini N, Cubuk E D, et al. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring. In: Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, 2020
18 Sohn K, Berthelot D, Carlini N, et al. FixMatch: simplifying semi-supervised learning with consistency and confidence. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, 2020. 596–608
19 Krizhevsky A, Hinton G. Learning Multiple Layers of Features From Tiny Images. Technical Report, 2009
20 Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Granada, 2011
21 Oliver A, Odena A, Raffel C, et al. Realistic evaluation of deep semi-supervised learning algorithms. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, 2018. 3239–3250
22 Guo L Z, Li Y F. Class-imbalanced semi-supervised learning with adaptive thresholding. In: Proceedings of the 39th International Conference on Machine Learning, Baltimore, 2022. 8082–8094
23 Xu Y, Shang L, Ye J X, et al. Dash: semi-supervised learning with dynamic thresholding. In: Proceedings of the 38th International Conference on Machine Learning, Vienna, 2021. 11525–11536
24 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual conference on Computational learning theory, Madison, 1998. 92–100
25 Zhou Z H. Disagreement-based semi-supervised learning. Acta Automatica Sin, 2013, 39: 1871–1878 [周志华. 基于分歧的半监督学习. 自动化学报, 2013, 39: 1871–1878]
26 Zhou Z H, Li Ming. Tri-training: exploiting unlabeled data using three classifiers. IEEE Trans Knowl Data Eng, 2005, 17: 1529–1541
27 Chen D D, Wang W, Gao W, et al. Tri-net for semi-supervised deep learning. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, 2018. 2014–2020
28 Zhou Z H. When semi-supervised learning meets ensemble learning. In: Proceedings of the 8th International Workshop on Multiple Classifier Systems, Reykjavik, 2009. 529–538
29 Zhang H Y, Cissé M, Dauphin Y N, et al. MixUp: beyond empirical risk minimization. In: Proceedings of the 6th International Conference on Learning Representations, Vancouver, 2018
30 Li Y F, Liang D M. Safe semi-supervised learning: a brief introduction. Front Comput Sci, 2019, 13: 669–676
31 Guo L Z, Zhang Y G, Wu Z F, et al. Robust semi-supervised learning when not all classes have labels. In: Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, 2022
32 Wei T, Wang H, Tu W W, et al. Robust model selection for positive and unlabeled learning with constraints. Sci China Inf Sci, 2022, 65: 212101
33 Li Y F, Kwok J T, Zhou Z H. Towards safe semi-supervised learning for multivariate performance measures. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 1816–1822
34 Angluin D, Laird P. Learning from noisy examples. Machine Learn, 1988, 2: 343–370
35 Coates A, Ng A Y, Lee H. An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, 2011. 215–223
36 Zagoruyko S, Komodakis N. Wide residual networks. In: Proceedings of the British Machine Vision Conference, 2016
37 Cubuk E D, Zoph B, Shlens J, et al. RandAugment: practical automated data augmentation with a reduced search space. In: Proceedings of the 34th Conference on Neural Information Processing Systems, Vancouver, 2020. 18613–18624
38 Guo L Z, Zhou Z, Li Y F. Record: resource constrained semi-supervised learning under distribution shift. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Virtual Event, 2020. 1636–1644
39 Zhou Z H. Open-environment machine learning. Natl Sci Rev, 2022, 9: nwac123

Robust pseudo-label selection for holistic semi-supervised learning

Lanzhe GUO & Yufeng LI*

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023*, China*
* Corresponding author. E-mail: liyf@nju.edu.cn

Abstract Semi-supervised learning (SSL) is a powerful paradigm for leveraging unlabeled data to mitigate the reliance on large labeled datasets. Although it has been reported that SSL methods achieve significant performance on multiple benchmark datasets, they still have critical limitations when applied to real-world tasks, such as being difficult to determine the quality of pseudo-labels, being sensitive to hyper-parameter choices, lacking theoretical guarantee. To address these issues, we propose a new holistic SSL approach with robust pseudo-label selection. Specifically, our proposal selects pseudo-labels adaptively based on the disagreement of model predictions without pre-defined hyper-parameters. Theoretically, we prove that the classification error decreases with the training iterations. Experimentally, we achieve state-of-the-art performance by a large margin across various datasets. For example, compared with the SOTA SSL algorithm FixMatch, we reduce the error by 11.8% on the CIFAR-10 dataset and 18.8% on the more difficult STL-10 dataset.

Keywords machine learning, deep learning, semi-supervised learning, pseudo-label, robust