doi:10.3772/j.issn.1002-0470.2024.04.002

基于样本动态权重的课程式半监督学习方法①

朱 徽②[*,**] 胡 斌[*] 宋怡宁[***] 赵晓芳[*,****]

([*]中国科学院计算技术研究所 北京 100190)

([**]中国科学院大学 北京 100049)

([***]中央军委国防动员部信息中心 北京 100034)

([****]中科苏州智能计算技术研究院 苏州 215028)

摘 要 本文针对半监督场景中极度匮乏的监督信号导致的标签传播困难、模型训练严重受噪声干扰等问题展开研究。伪标签化带来的噪声和低数据利用率导致的确认偏差,会随着自训练过程造成错误累积,进而形成不可逆偏差,损害性能。本文提出基于样本动态权重的课程式半监督学习方法,旨在通过非离散的课程设计,鼓励模型由简单至困难地利用样本,逐步构建分类面,进而缓解伪标签化过程中的噪声产生,增强模型泛化能力。从类内角度,提供弱监督信号的高置信度伪标签被混合用于构建特征原型,估计样本的学习难度。从类间角度,标签嵌入被用于评估类间语义相关度,课程式地减弱训练前期对语义相关类别间的辨别。在通用的半监督学习基准数据集上进行了广泛的实验和分析,证明了方法的有效性。

关键词 半监督学习;特征表示向量;课程学习;特征原型;语义相关度

深度学习的成功在很大程度上依赖于大规模神经网络[1-2]的支持,而这些神经网络又往往由大量有标签数据所驱动。然而,获取完全标注的数据集是耗时的、昂贵的,且过度依赖专家经验。相比有标签数据,充足的无标签数据广泛存在且容易收集。由此,使用少量有标签数据结合大量无标签数据完成模型训练的半监督学习方法受到广泛关注。随着半监督领域研究的发展,传统监督学习算法中“泛化性能不足”、“模型容易过拟合”、“严重依赖数据标注质量”等问题在一定程度上得到有效改善。

现有的主流半监督学习算法主要采用 2 类机制:一致性正则法和伪标签化法。其中,基于一致性正则的半监督学习方法[3-6]激励模型对于同一输入图像进行不同程度扰动后,反馈出一致的预测结果;基于伪标签化的半监督学习方法[7-10]利用现阶段模型产生高置信度的伪标签,并将其进一步用于指导后续的模型训练。一些最新的半监督学习方法融合 2 种范式[11-12],引入更复杂的数据增强方法[3,13]和学习策略[14-15],实现了对于更强扰动模型仍表现出空间一致性,提高了伪标签生成的可靠性,进而取得了显著效果。

然而,半监督场景中监督信号匮乏,特别是当目标任务覆盖更多类别数和更少有标签样本的情况下,已有方法无差别地利用样本为全部类别间构建分类面难度极大,出现的确认偏差、欠拟合和噪声干扰问题将影响伪标签生成的准确率,进而损害模型性能。同时,无标签数据中包含离群点、非典型样本以及噪声。在半监督学习初期,对于这些困难样本的伪标签生成容易出现错误,误导模型训练方向,进而随着自训练过程造成更严重的错误累积和不可逆

① 国家重点研发计划(2021YFF0703800)资助项目。

② 男,1995 年生,博士生;研究方向:计算机视觉,机器学习,人工智能;联系人,E-mail:zhuhui@ict.ac.cn。

(收稿日期:2023-02-08)

偏差。

现有的半监督方法同等对待各样本与类别,样本在训练过程中各个阶段所占据的重要程度没有被深入思考和探究,而各类别间样本的潜在相似和差异信息也往往被忽略。例如,在使用猫这一类别的无标签图像时,图片类型可能包括照片、卡通、素描等,其学习难度存在较大差异;而在区分哈士奇、狼和椅子多个类别样本时,前两者具有很大的相似性,而与后者关联性不强。由此,有效评估样本在训练过程中不同阶段的重要性,计算类别间潜在相似度,有利于更合理地辅助分类面的构建,特别是在训练的早期阶段,可以大幅提高伪标签生成的可靠性,缓解噪声干扰以及错误累积问题。

本文提出了一种基于样本动态权重的课程式半监督学习方法,对于样本的类内和类间权重进行动态调整,并通过一个非离散的课程设计,鼓励模型由简单至困难地利用样本,逐步构建分类面,进而增强模型的泛化能力。具体而言,从样本类内角度,本工作通过评估无标签样本的学习难度,课程式地动态调整类内各样本对损失的贡献。从样本类间角度,本工作通过引入标签嵌入评估类别间的语义相关度,并设计课程式学习算法,逐步有侧重地构建类间分类面。本文在 3 个通用的半监督学习基准数据集(CIFAR-10、CIFAR-100 和 STL-10)上进行了一系列的验证实验、消融实验和分析性实验,以证明所提出方法的有效性。

1　相关工作

1.1　半监督学习

在基于深度学习的人工智能产业中,数据规模往往成为算法研发的关键要素。拥有更多可利用的数据,就可以驱动容量更大、更复杂的模型,或将同等规模的模型训练到更优的效果,进而增强算法核心竞争力。由此,节省专家标注成本且同时利用更多易采集数据的半监督学习受到了广泛关注。

半监督学习算法依据其机制主要分为 2 类:一致性正则法和伪标签化法(自训练)。作为基于一致性正则法的半监督算法原型,Pseudo-Ensembles[16]直接应用了高斯噪声和 Dropout 噪声。随后,一系列方法,包括 Π-Model[17]、Mean Teacher[6]、VAT[5]以及 MixMatch[4],分别通过利用随机正则化和随机扰动,使用旧模型检查点,利用模型参数值的指数移动平均值(exponential moving average,EMA)以及引入更复杂的 Mixup[18]正则化,实现了半监督算法性能的大幅度提升。此外,S^4L[19]通过引入自监督形式的损失进一步激励空间一致性,FeatMatch[20]提出了一种可学习的基于特征的细化和增强模块。而基于伪标签化法的半监督方法范式最先由文献[21]提出,随后 TSSDL[10]通过引入基于 k 近邻(k-nearest neighbors,kNN)结果密度的置信水平,以克服不可靠的标签估计可能导致噪声训练的问题。Label Propagation[8]采用了一种基于流形假设的推导式标签传播方法来进行预测。为解决伪标签化过程中的噪声干扰问题,PENCIL[22]通过更新标签分布来纠正噪声标签,R2-D2[23]使用标签概率分布作为伪标签,UPS[9]提出了一种不确定性感知伪标签选择框架以最小化较差网络校准的影响。

近期的半监督学习方法进一步结合了一致性正则法和伪标签化法的思想。UDA[12]和 ReMixMatch[3]依据决策边界通过高密度区域的低可能性假设,通过锐化人工标签从而鼓励模型产生高置信度预测。FixMatch[11]采用了基于置信度阈值的伪标签生成方法,并抛弃了诸多冗余组件来简化半监督学习框架。为进一步提高算法性能,更多学习策略被融入半监督算法,例如,课程学习[7,14]、协同训练[15]等。此外,从预定义的集合中随机选择若干变换操作的复杂数据增强方法的引入,例如,CTAugment[3]、RandAugment[13]等,也在半监督学习领域取得了显著成效。

1.2　课程学习

课程学习的本质是激励模型在学习困难样本之前预先学习简单样本[24-25]。近期提出的半监督学习方法尝试将课程学习策略融入算法[7,14,26],从而尽可能更有效地利用无标签样本。准确地识别困难样本,逐步调整每个样本的参与度是课程学习成功的关键。本文设计了一个非离散课程,以鼓励模型在持续的半监督训练过程中,由简单至复杂地逐步构建分类面,缓解错误产生,增强模型泛化能力。

2　本文方法

本节重点介绍基于样本动态权重的课程式半监督学习方法。首先简要介绍半监督任务中关键的问题描述，然后从类内和类间 2 个角度分别介绍课程式学习算法细节，最后描述将 2 种课程联合用于模型训练的方法总述。

本方法的整体流程如图 1 所示，具体步骤可以概述如下。

（1）将强数据增强方法（A）应用于无标签样本，将弱数据增强方法（α）应用于全部的有标签样本和无标签样本，得到模型对于全部输入样本的对应特征（图中特征提取网络 f_Θ 的输出）和预测概率分布（图中分类器 p_m 的输出）。

（2）利用标签嵌入矩阵计算类间语义相关度 h（详见 2.4 节），并基于此调整伪标签化过程中置信度评估函数（详见 2.5 节，式(8)）。

（3）启动伪标签化过程，将保留伪标签的无标签样本与有标签样本混合后，计算特征原型（详见 2.2 节），并基于此计算无标签样本的类内权重 w（详见 2.3 节）。

（4）使用基于类内样本动态权重的课程式学习方法（算法 1）和基于类间样本动态权重的课程式学习方法（算法 2），利用弱数据增强后有标签图像 $\alpha(\boldsymbol{B}_l)$ 的预测概率分布和标签计算有监督损失 L_s，利用强数据增强后无标签图像 $A(\boldsymbol{B}_u)$ 的预测概率分布和生成的伪标签计算无监督损失 L_u，进而计算模型训练的整体损失 L。

（5）将整体损失函数作为优化目标，使用优化器（例如，随机梯度下降（stochastic gradient descent，SGD））进行模型训练。

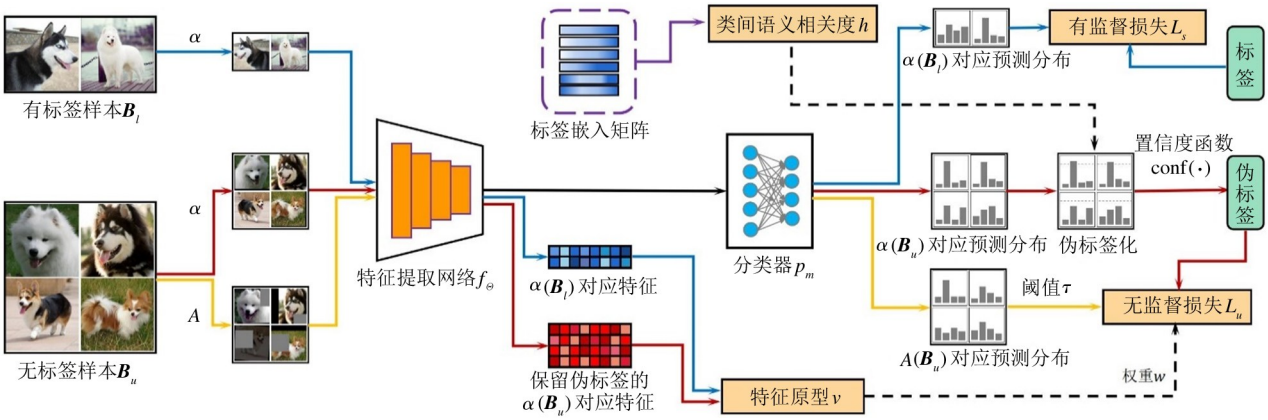


图 1　本文方法的整体流程图

2.1　问题描述

半监督学习方法的目标通常是基于一个完整的训练数据集 $\boldsymbol{D}=\{\boldsymbol{D}_L,\boldsymbol{D}_U\}$，训练一个由可训练参数 Θ 构成的模型 f_Θ。该训练集包含一个由 N_L 个有标签样本构成的有标签数据集 $\boldsymbol{D}_L$ 和一个由 N_U 个无标签样本构成的无标签数据集 $\boldsymbol{D}_U$，其中 $N_L \ll N_U$。对于有标签数据集 $\boldsymbol{D}_L=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{N_L}$，$\boldsymbol{x}_i$ 表示第 i 个样本，$\boldsymbol{y}_i=[y_i^1,y_i^2,\cdots,y_i^K]\in\{0,1\}^K$ 是对应的 K 个类别的独热编码真实标签。其中 $y_i^k=1$ 表示标签中第 k 个类别存在。对于无标签数据集 $\boldsymbol{D}_U=\{\boldsymbol{u}_l\}_{l=1}^{N_U}$，每个样本不包含任何标签信息。

2.2　基于特征原型的类内样本学习难度估计

在训练数据集中，属于同一类别的样本差异性往往很大。以图像数据为例，其图片类型可能多样，单个样本的质量也不尽相同。因此，数据集中各样本的学习难度通常具有显著差异。不对样本学习难度加以区分而盲目使用大量无标签样本，容易产生大量噪声。特别是在监督信号极度匮乏的半监督学习场景中，往往会导致更为严重的确认偏差和错误累积现象。尽可能准确地评估样本的学习难度，有利于设计出更合理的学习算法，进而提升数据利用率，增强算法的核心竞争力。但值得注意的是，如果分割出特定的有标签验证集进行样本评估是耗时

的,这种方法在标注信息极其珍贵的半监督学习场景中,更加行不通。

为了低成本地估计样本的学习难度,本文依据能提供监督信号的样本的特征向量构建特征原型,并基于此衡量样本与特征原型的相似度,进而估计类内样本的学习难度。尽管在半监督学习场景中,有标签样本数量稀少,准确的监督信号极度匮乏,但高置信度的伪标签可以提供充足的弱监督信号。因此,混合置信度较高的保留伪标签的无标签样本,有利于更准确地构建各类别的特征原型,并在一定程度上抵抗离群点等不确定性干扰甚至噪声。

具有可学习参数 Θ 的模型在概念上可以划分为 2 部分:一个将输入映射到一个特征描述的特征提取网络 $f_\Theta: X \to \mathbb{R}^M$ 和一个应用于 f_Θ 之后作为分类器的全连接层(含 softmax 层) $p_{\mathrm{m}}: \mathbb{R}^M \to \mathbb{R}^K$。给定一小批次数量为 B 的有标签数据 $\boldsymbol{B}_l$ 和一小批次数量为 μB 的无标签数据 $\boldsymbol{B}_\mu$,其中 μ 是一个标量超参数用于表示相对权重。在半监督学习基本范式中,伪标签的生成由置信度评估函数所约束。给定一个随机的无标签样本 $\boldsymbol{u}_\xi$,$\alpha(\boldsymbol{u}_\xi)$ 表示其弱增强版本,将其作为模型输入能够得到弱增强分支对应的预测概率分布 $\boldsymbol{q}_\xi = p_{\mathrm{m}}(y \mid f_\Theta(\alpha(\boldsymbol{u}_\xi)))$,当且仅当 $\max(\boldsymbol{q}_\xi) \geqslant \gamma$ 时,该无标签样本将保留伪标签 $\hat{\boldsymbol{q}}_\xi = \operatorname{argmax}(\boldsymbol{q}_\xi)$,并参与后续无监督损失计算,其中 γ 为伪标签化置信度阈值。

使用 $\boldsymbol{B}_l^k \subset \boldsymbol{B}_l$ 和 $\boldsymbol{B}_\mu^k \subset \boldsymbol{B}_\mu$ 表示属于第 k 个类别的有标签样本和保留伪标签的无标签样本的子集。给定一个随机样本 $\boldsymbol{x}_\xi$,$f_\Theta(\alpha(\boldsymbol{x}_\xi))$ 表示其弱增强版本 $\alpha(\boldsymbol{x}_\xi)$ 对应的特征向量,通过计算属于第 k 个类别的样本的平均向量得到特征原型 $\boldsymbol{v}^k \in \mathbb{R}^M$,这一过程可以用公式表示为

$$\boldsymbol{v}^k = \frac{1}{|\boldsymbol{B}_l^k| + |\boldsymbol{B}_\mu^k|} \sum_{\boldsymbol{x}_t \in \boldsymbol{B}_l^k \cup \boldsymbol{B}_u^k} f_\Theta(\alpha(\boldsymbol{x}_t)) \tag{1}$$

其中,$\boldsymbol{x}_t$ 表示第 t 个样本,$|\cdot|$ 表示基数,即该批次内样本的数量。令 d_ξ^k 表示无标签样本 $\boldsymbol{u}_\xi^k$ 到特征原型 $\boldsymbol{v}^k$ 的欧氏距离,$\boldsymbol{u}_\xi^k$ 与 $\boldsymbol{v}^k$ 之间的相似度 s_ξ 计算方法如下:

$$s_\xi = \frac{1}{d_\xi^k} = \frac{1}{\| f_\Theta(\alpha(\boldsymbol{u}_\xi)) - \boldsymbol{v}^k \|_2} \tag{2}$$

这一相似度则能够近似表示样本的学习难度。也就是说,样本的特征向量距离特征原型越近,表示该样本特征典型,相对更具有代表性,模型容易学习;反之,样本的特征向量距离特征原型越远,表示该样本在特征空间内处于类别边缘,甚至以离群点形式存在,模型学习往往更加困难。

如图 2 所示,无标签样本 $\boldsymbol{u}_2$、$\boldsymbol{u}_3$ 相较 $\boldsymbol{u}_1$ 距离特征原型 $\boldsymbol{v}_2$ 更远,其学习难度更大,有较高可能性为离群点或噪声样本。利用本方法所构建的特征原型可以较好地评估各个类别内样本的学习难度,辅助设计更合理的学习算法,进而更有效地利用无标签样本。随着半监督训练过程的推进,模型准确率提升,特征向量将更加准确,生成伪标签的数量及其可信度稳步提升,特征原型的构建以及相似度的评估也更加可靠。同时,由于有标签样本的参与,本方法具有较强的准确性和鲁棒性,可以很好地处理不确定样本和噪声。对于不确定的样本,本方法可以帮助评估伪标签置信度及可靠性,降低伪标签生成过程中错误的产生。对于噪声样本,可以较好地控制其在损失计算中的参与度,降低噪声在早期阶段对半监督训练过程的干扰。

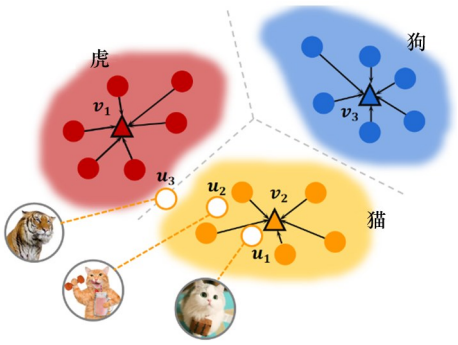


图 2　基于特征原型估计无标签样本学习难度

2.3　基于类内样本动态权重的课程式学习方法

在半监督学习过程中,监督信号匮乏,若不对样本学习难度加以区分,前期盲目学习困难样本,往往会产生大量噪声伪标签,并随着自训练过程导致错误累积和放大,进而误导模型训练方向。反之,如果能够从容易学习的样本出发,辐射其他无标签样本,由易到难地进行标签传播,引导模型学习逐步深入,能够有效缓解噪声干扰、确认偏差和错误累积问题。

由此,考虑对学习难度较低的无标签样本在训练前期赋予较高的权重,而对学习困难的不可靠样本在训练前期赋予较低的权重,并基于此实现课程式地推进半监督学习进程。利用计算得到的无标签样本与特征原型间的相似度 $\boldsymbol{s}=[s_1, s_2, \cdots]$,为无标签样本分配动态权重。对于一个随机的无标签样本 $\boldsymbol{u}_\xi$,其在损失中所占初始权重可以表示为

$$w_\xi = \left(\frac{s_\xi}{\max(\boldsymbol{s})}\right)^{\Omega} \tag{3}$$

其中,Ω 是一个“温度”超参数,用于进一步调节权重大小。接下来,将无标签样本对应的动态权重应用于无监督损失计算中。具体地,依据训练轮次,将课程式地改变无标签样本权重在无监督损失中的参与度。同样地,对 $\boldsymbol{u}_\xi$ 进行弱增强后,将所得到的图像 $\alpha(\boldsymbol{u}_\xi)$ 送入模型,得到对应的预测概率分布 $\boldsymbol{q}_\xi = p_{\mathrm{m}}(y \mid f_\Theta(\alpha(\boldsymbol{u}_\xi)))$,再通过计算这一概率分布中概率最大的值所属的类别(即 argmax 函数)得到伪标签 $\widehat{\boldsymbol{q}_\xi} = \mathrm{argmax}(\boldsymbol{q}_\xi)$。令 $A(\boldsymbol{u}_\xi)$ 表示无标签样本 $\boldsymbol{u}_\xi$ 的强增强版本,将其作为模型输入能够得到强增强分支对应的预测概率分布 $\boldsymbol{Q}_\xi = p_{\mathrm{m}}(y \mid f_\Theta(A(\boldsymbol{u}_\xi)))$。依据半监督学习中通用的一致性损失,即交叉熵损失 $H(\cdot)$,计算伪标签 $\widehat{\boldsymbol{q}_\xi}$ 与强增强版本对应概率分布 $\boldsymbol{Q}_\xi$ 之间的交叉熵,即 $H(\widehat{\boldsymbol{q}_\xi}, \boldsymbol{Q}_\xi)$。由此,样本 $\boldsymbol{u}_\xi$ 对应的无监督损失可以表示为

$$l(\boldsymbol{u}_\xi) = w_\xi^{\delta} H(\widehat{\boldsymbol{q}_\xi}, \boldsymbol{Q}_\xi) \tag{4}$$

其中,$\delta = 1 - \dfrac{t}{T}$ 表示用于调节权重 w_ξ 进行课程式衰减变化的超参数,t 是当前的训练轮次,T 是总训练轮次。值得注意的是,基于置信度评估函数,未保留伪标签的无标签样本因其伪标签可靠性不足,不会实际参与到损失计算中来。

基于类内样本动态权重的课程式学习方法的伪代码见算法 1。

其中第 1 行为使用有标签数据计算有监督损失,第 2 ~ 7 行为根据伪标签置信度选择保留伪标签的无标签样本,第 8 行为计算各类别特征原型,第 9 行为计算无标签样本动态权重,第 10 ~ 11 行为基于无标签样本动态权重计算无监督损失,第 12 行为计算用于模型训练的整体损失。

这一基于样本动态权重的课程式学习方法能够有效降低伪标签化过程中的噪声产生,进一步避免错误伪标签随着自训练过程不断累积和传播,有利于提高无标签数据的利用率,提升模型训练得到的最终性能。值得注意的是,性能更优的中间模型,能够产生更加可靠的样本特征向量和高置信度伪标签,这对上节所描述的类内样本学习难度估计具有积极的促进作用。由此,二者之间(基于特征原型的样本类内学习难度估计与基于类内样本动态权重的课程式学习)在效果上能够相互促进,并随着半监督自训练过程的推进而持续地相互增益。

算法 1　基于类内样本动态权重的课程式学习方法

输入:一批次数量为 B 的有标签数据 $\boldsymbol{B}_l$,一批次数量为 μB 的无标签数据 $\boldsymbol{B}_\mu$,数据增强方法 $\{\alpha, A\}$,伪标签置信度阈值 γ,当前训练轮次 t,总训练轮次 T

输出:整体损失 L

1: 计算有监督损失 $L_s = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i)\in \boldsymbol{B}_l} H(\boldsymbol{y}_i, p_{\mathrm{m}}(y \mid f_\Theta(\alpha(\boldsymbol{x}_i))))$

2: **for** $i = 1, \cdots, \mu B$ **do**

3: 　计算预测概率分布 $\boldsymbol{q}_i = p_{\mathrm{m}}(y \mid f_\Theta(\alpha(u_i)))$

4: 　**if** $\max(\boldsymbol{q}_i) \geqslant \gamma$ **then**

5: 　　将 $\boldsymbol{u}_i$ 加入保留伪标签的样本集合 $\boldsymbol{B}_{pl}$

6: 　**end if**

7: **end for**

8: 依据集合 $\boldsymbol{B}_l$ 与 $\boldsymbol{B}_{pl}$ 中样本的特征向量计算全部类别的特征原型 $\boldsymbol{v} = \{\boldsymbol{v}^1, \boldsymbol{v}^2, \cdots, \boldsymbol{v}^K\}$ ▷ 式(1)

9: 计算相似度 $\boldsymbol{s}$ 与权重 $\boldsymbol{w}$ ▷式(2) ~ (3)

10: 计算超参数 $\delta = 1 - t/T$

11: 计算无监督损失 $L_u = \sum_{\boldsymbol{u}_j\in \boldsymbol{B}_u} \mathbb{I}_{[\max(\boldsymbol{q}_j)\geqslant\gamma]} w_j^{\delta} H(\widehat{\boldsymbol{q}_j}, p_{\mathrm{m}}(y \mid f_\Theta(A(u_j))))$ ▷ 式(4)

12: 计算整体损失 $L = L_s + \lambda_u L_u$

2.4　基于标签嵌入的类间语义相关度计算方法

word2vec[27] 与 Global Vector (GloVe)[28],作为 2 种通过学习类别标签的词嵌入来获取语义信息的通用方法,其性能和效果是相近的。二者都基于单词共现进行统计并在大型文本语料库中进行预训练。本工作使用 GloVe 方法在特定语料库中预训练的模型,进行语义信息的获取。该方法是一种基于计数的无监督算法,通过在欧几里得空间中学习词嵌入,并通过统计方法反映语义关系。

类别间语义相关度示意图如图 3 所示。对数据集中的所有类别标签分别进行处理,得到表示语义信息的标签嵌入矩阵,记作 $\boldsymbol{W}=[\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_K]\in R^{K\times Z}$。对于任一标签类别 ξ,通过计算该类别与其他所有类别词嵌入向量间的余弦相似度,得到该类别与其他各个类别的语义相关度。该随机类别 ξ 与某一特定类别 p 的语义相关度可以表示为

$$h_{\xi,p}=\mathrm{Cos\ Sim}(\boldsymbol{W}_\xi,\boldsymbol{W}_p)=\frac{\boldsymbol{W}_\xi\boldsymbol{W}_p^{\mathrm{T}}}{\|\boldsymbol{W}_\xi\|\ \|\boldsymbol{W}_p\|} \tag{5}$$

其中,Cos Sim(·)表示 2 个同维度向量间的余弦相似度。由此,该随机类别 ξ 与其他所有类别语义相关度分布可以表示为 $\boldsymbol{h}_\xi=[h_{\xi,1},h_{\xi,2},\cdots,h_{\xi,K}]$。

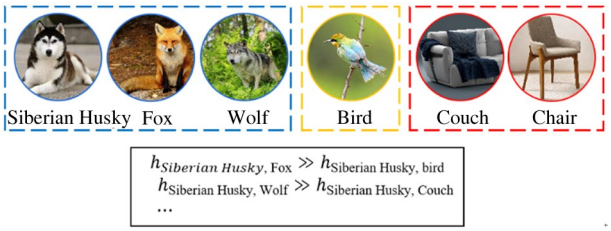


图 3　类别间语义相关度示意

为了在一定程度上削弱语义无关类别的噪声干扰,进一步降低语义相关度较低类别在分布中所占权重,根据低密度假设,使用锐化操作降低该分布中类别间的熵。锐化函数可以表示为

$$h'_{\xi,i}=\mathrm{sharpen}(\boldsymbol{h}_\xi,\Gamma)_i=\frac{h_{\xi,i}^{\Gamma}}{\sum_{r=1}^{K}h_{\xi,r}^{\Gamma}} \tag{6}$$

其中,$h'_{\xi,i}$ 表示锐化后的类别 ξ 与类别 i 之间的语义相似度,Γ 是一个“温度”超参数,用于调节锐化力度。当 $\Gamma\to1.0$,锐化操作力度减弱。锐化后,应根据最大语义相关度的值(即与本类别间的相似度),将分布 $\boldsymbol{h}'_\xi=[h'_{\xi,1},h'_{\xi,2},\cdots,h'_{\xi,K}]$ 内所有值进行放大回调,由于锐化操作前 $\max(\boldsymbol{h}_\xi)=1.0$,回调操作可以表示为

$$\widetilde{\boldsymbol{h}}_\xi=\frac{\boldsymbol{h}'_\xi}{\max(\boldsymbol{h}'_\xi)} \tag{7}$$

其中,$\max(\boldsymbol{h}'_\xi)$ 表示锐化操作后语义相关度的最大值(与自身类别的相似度)。语义相关度分布 $\widetilde{\boldsymbol{h}}_\xi$ 作为基础,将进一步用于半监督算法中的置信度函数的设计和类间样本动态权重的课程式调整(见下文)。

2.5　基于类间样本动态权重的课程式学习方法

上文提及,同时对全部类别构建分类面容易造成欠拟合问题,甚至出现误判。对于监督信号更加匮乏的半监督场景,直接对全部类别进行辨别更加艰难。因此,对于无标签样本,本文希望在训练的早期阶段减少对于语义相关类别内部的细粒度辨别,即优先构建语义相关度较低的类间分类面。与基于类内样本动态权重的课程式学习相似,同样为类间样本设计了一个非离散的课程,用于动态修改语义相关类别在置信度计算中的参与度。

给定随机无标签样本 $\boldsymbol{u}_\xi$,将其弱增强版本 $\alpha(\boldsymbol{u}_\xi)$ 作为输入,得到预测类分布 $\boldsymbol{q}_\xi=p_{\mathrm{m}}(y\mid f_\Theta(\alpha(\boldsymbol{u}_\xi)))$。在伪标签化过程中,对于每个无标签的样本,不再仅根据其最大概率所属预测类别 $\widehat{q_\xi}$ 生成伪标签,而是同时将其他类别的预测概率依据其语义相关度,以一定比例引入置信度评估函数 conf(·)中。基于类间样本语义相关度计算结果 $\widetilde{\boldsymbol{h}}_\xi$,半监督学习算法中的置信度评估函数可以修改为

$$\mathrm{conf}(\boldsymbol{u}_\xi)=\max(\boldsymbol{q}_\xi)+\delta\sum_{p=1}^{K}\mathbb{I}_{[p\neq\widehat{q_\xi}]}\ \widetilde{h}_{\widehat{q_\xi},p}\ \boldsymbol{q}_{\xi p} \tag{8}$$

其中,与上文相同,$\delta=1-\frac{t}{T}$ 表示用于调节语义相关类别进行课程式衰减变化的超参数,t 是当前的训练轮次,T 是总训练轮次,δ 逐渐从 1 衰减至 0。也就是说,随着训练迭代次数的增加,其他类别的语义相关度对确定保留伪标签及其是否包含在损失函数中的参与程度逐渐下降。

同时,对于无标签样本在基于一致性正则的半监督学习范式中的强增强分支预测分布 $\boldsymbol{Q}_\xi=p_{\mathrm{m}}(y\mid f_\Theta(A(\boldsymbol{u}_\xi)))$,同样以依据各个类别的语义相关度计算概率之和的形式动态调整类间样本对损失贡献的权重。如果概率之和足够大,表明该样本与语义相关度较高的类别区分困难,本文将在无监督损失函数中课程式地降低该样本的权重。也就是说,在半监督训练的早期阶段,不会对于这些语义相关度较高类别间区分不清的样本过多关注。因此,模型倾向于优先学习对于语义相关度较低类别之间没有足够信心的无标签样本。具体地,计算每个保留伪

标签的无标签样本的损失过程可以表示为

$$l'(\boldsymbol{u}_\xi)=\begin{cases} l(\boldsymbol{u}_\xi) & \sum_{k=1}^{K}\widetilde{h}_{\widehat{q_\xi},k}\,\boldsymbol{Q}_{\xi_k}\leqslant\tau \\ \left(1-\dfrac{\delta}{\lambda}\right)l(\boldsymbol{u}_\xi) & \text{其他}\end{cases} \tag{9}$$

其中,若使用基于类内样本动态权重的课程式学习方法,则基础损失 $l(\boldsymbol{u}_\xi)=w_\xi^\delta H(\widehat{\boldsymbol{q}_\xi},\boldsymbol{Q}_\xi)$,否则,基础损失应为标准交叉熵损失,即 $l(\boldsymbol{u}_\xi)=H(\widehat{\boldsymbol{q}_\xi},\boldsymbol{Q}_\xi)$。$\delta$ 仍然用于调整样本在损失中所对应权重进行课程式变化,λ 是用于进一步调节权重变化幅度的因子,而 τ 是一个表示阈值的超参数。

以 3 个类别[dog, wolf, plane]为例,给定一个 dog 类别的样本,阈值 $\tau=0.7$,dog 类别与其他类别间的语义相关度为[1.0, 0.9, 0.1]。若样本 A 的强增强版本对应预测概率分布为[0.5, 0.4, 0.1],则其依据语义相关度计算概率和为 $1.0\times0.5+0.9\times0.4+0.1\times0.1=0.87>0.7$。而样本 B 的强增强版本对应预测概率分布为[0.5, 0.1, 0.4],则其依据语义相关度计算概率和为 $1.0\times0.5+0.9\times0.1+0.1\times0.4=0.63<0.7$。由此,样本 A 属于语义相关度较高类别间分辨不清(dog 与 wolf),将对其在损失中所占的权重进行动态削减,而样本 B 属于语义相关度较低类别间分辨不清(dog 与 plane),将优先对其进行学习。

基于类间样本动态权重的课程式学习方法的伪代码见算法 2。

算法 2　基于类间样本动态权重的课程式学习方法

输入:一批次数量为 B 的有标签数据 $\boldsymbol{B}_l$,一批次数量为 μB 的无标签数据 $\boldsymbol{B}_u$,数据增强方法 $\{\alpha, A\}$,伪标签置信度阈值 η,标签嵌入矩阵 $\boldsymbol{W}$,当前训练轮次 t,总训练轮次 T,超参数 λ,阈值 τ

输出:整体损失 L

1:　计算有监督损失 $L_s=\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\boldsymbol{B}_l}H(\boldsymbol{y}_i,p_{\mathrm{m}}(y\mid f_\Theta(\alpha(\boldsymbol{x}_i))))$

2:　**for** $i=1,\cdots,K$ **do**

3:　　依据标签嵌入矩阵 $\boldsymbol{W}$ 计算类别 i 与其他类别间的初始语义相关度 $\boldsymbol{h}_i$　▷式(5)

4:　　对 $\boldsymbol{h}_i$ 锐化和回调,得到 $\widetilde{\boldsymbol{h}}_i$　▷式(6)~(7)

5:　**end for**

6:　将无监督损失 L_u 初始值置为 0

7:　**for** $i=1,\cdots,\mu B$ **do**

8:　　计算伪标签 $\widehat{\boldsymbol{q}_i}=\arg\max(\boldsymbol{q}_i)$

9:　　基于 $\widetilde{h}_{\widehat{q_i}}$ 计算置信度 $\mathrm{conf}(\boldsymbol{u}_i)$　▷式(4)

10:　　计算预测概率分布 $\boldsymbol{Q}_i=p_{\mathrm{m}}(y\mid f_\Theta(A(\boldsymbol{u}_i)))$

11:　　计算超参数 $\delta=1-t/T$

12:　　**if** $\mathbb{I}_{[\mathrm{conf}(\boldsymbol{u}_i)\geqslant\eta]}$ **then**

13:　　　**if** $\sum_{k=1}^{K}\widetilde{\boldsymbol{h}}_{\widehat{q_i},k}\,\boldsymbol{Q}_{i_k}\leqslant\tau$ **then**

14:　　　　$L_u\leftarrow L_u+H(\widehat{\boldsymbol{q}_i},\boldsymbol{Q}_i)$　▷式(9)

15:　　　**else**

16:　　　　$L_u\leftarrow L_u+\left(1-\dfrac{\delta}{\lambda}\right)H(\widehat{\boldsymbol{q}_i},\boldsymbol{Q}_i)$　▷式(9)

17:　　　**end if**

18:　　**end if**

19:　**end for**

20:　计算整体损失 $L=L_s+\lambda_u L_u$

其中第 1 行为使用有标签数据计算有监督损失,第 2~5 行为基于标签嵌入矩阵计算类间语义相关度,第 6~19 行为基于修改后的置信度评估函数和语义相关度计算无监督损失,第 20 行为计算用于模型训练的整体损失。

2.6　课程式半监督学习方法总述

半监督学习的总体损失函数由有监督损失和无监督损失两部分组成,可以表示为

$$L=L_s+\lambda_u L_u \tag{10}$$

其中,λ_u 表示两部分损失间比例。由于本方法主要面向无标签数据,因此,有监督损失部分为使用有标签数据及其真实标签的常规交叉熵损失,即:

$$L_s=\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\boldsymbol{B}_l}H(\boldsymbol{y}_i,p_{\mathrm{m}}(y\mid f_\Theta(\alpha(\boldsymbol{x}_i)))) \tag{11}$$

对于无监督损失,联合使用类内和类间课程学习方法,结合置信度评估函数和每个无标签样本的损失,可将无监督损失函数表示为

$$L_u=\sum_{\boldsymbol{u}_j\in B_u}\mathbb{I}_{[\mathrm{conf}(\boldsymbol{u}_j)\geqslant\eta]}\,l'(\boldsymbol{u}_j) \tag{12}$$

其中,$l'(\boldsymbol{u}_j)$ 包含类内样本动态权重项 w_j^δ。使用优化器,将这一总体损失函数 L 作为优化目标,进行半监督学习模型训练。

3　实验与分析

本节将重点介绍实验数据集与评价指标、实验设置、实验结果与分析等内容。

3.1　实验数据集与评价指标

对于半监督学习任务,常用的基准数据集包括CIFAR-10[29]、CIFAR-100[29]和STL-10[30]。

CIFAR-10数据集包含50 000张训练图像和10 000张尺寸为32像素×32像素的测试图像。该数据集共覆盖10个不同类别(飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车),每个类别包含5 000个训练图像。

与CIFAR-10数据集类似,CIFAR-100数据集包括50 000张训练图像和10 000张尺寸为32像素×32像素的测试图像。它由100个更详细的类别组成,每个类别包含500个训练图像。

STL-10数据集是专门为开发无监督和半监督学习而设计的,它的灵感来自于CIFAR-10数据集,并从ImageNet数据集[31]中的有标签样本内提取。该数据集包含5 000个有标签图像,100 000个无标签图像和8 000个测试图像,每个图像尺寸为96像素×96像素。有标签的集合包含10个不同的类别(飞机、鸟、汽车、猫、鹿、狗、马、猴子、船、卡车)。无标签的图像是从相似但更广泛的分布中提取获得的,覆盖了熊、兔子、火车、公共汽车等类别。因此,对于半监督学习方法来说,它是一个更具挑战性的基准数据集。

对于半监督学习任务,对所有数据集依照标准评估协议[4,11]分割并构造半监督任务评价体系。具体地,丢弃全部样本标签从而构造无标签数据集,同时保留一小部分类别平衡的标签及其对应图像构造有标签数据集。对于半监督算法性能的评价,本工作相关实验主要依据其在测试集上的分类准确率(%)或分类错误率(%)进行评估。其中,准确率$=\frac{\text{正确分类的样本数}}{\text{样本总数}}$,错误率$=\frac{\text{错误分类的样本数}}{\text{样本总数}}$。本文图表中所报告的结果是基于3或5组不同初始数据所测得的实验结果的平均值和标准差。

3.2　实验设置

本工作代码实现基于Pytorch[32]框架,并在NVIDIA Tesla V100和A100 GPU上进行实验。为了进行公平的半监督学习性能对比,依照文献[3,11],使用相同的实验设置和通用的超参数设定。对于CIFAR-10、CIFAR-100和STL-10数据集,依照文献[11],分别使用相同的主干网络WRN-28-2(1.5 M)、WRN-28-8(23.4 M)和WRN-37-2(5.9 M)。

对于训练策略和参数设定,使用Nesterov动量为0.90的随机梯度下降(SGD)作为优化器来训练主干网络2^{20}个迭代(1 024个轮次)。初始学习率设置为0.03,批处理大小为64。对于学习率计划,使用余弦学习率衰减SGDR[33],它将学习率调整为$\eta_t=\eta_0\cos\left(\frac{7\pi t}{16T}\right)$,其中$\eta_0$是初始学习率,$t$是当前训练轮次,$T$是总训练轮数。除将CIFAR-100数据集的权重衰减设置为1×10^{-3}外,其余数据集的权重衰减设置为5×10^{-4}。实验中使用了指数移动平均EMA[6],衰减参数设置为0.999。依照文献[11],置信度阈值γ设置为0.950,有标签数据与无标签数据的比例μ为7.000,有监督与无监督损失比例λ_u为1.000。数据增强类型与参数与文献[11]相同,即弱增强采用水平翻转和随机裁剪,强增强使用RandAugment[13]与Cutout[34]。

对于GloVe方法,利用在Common Crawl数据集中由840 B个字符训练的模型版本进行词嵌入向量的提取,每个词嵌入的维数为300。值得注意的是,当一个类标签包含一个以上的单词时,例如"aquarium fish",其对应的词嵌入是通过平均所有单词的词向量得到的。

对于本方法引入的超参数,η设置为与γ相同,即0.95,λ设置为2.00,τ设置为0.70,"温度"超参数Γ和Ω分别设置为2.00和0.01。

3.3　实验结果

3.3.1　与最先进方法的性能比较

与最先进方法的性能对比,是基于同样的主干网络、训练策略和实验设置,使用不同算法将模型训练至收敛时,对比不同算法对应模型的最终性能和表现,以展示算法应用的真实效果。表1、表2和表3分别展示了在3种通用基准数据集(CIFAR-10、CI-

FAR-100 和 STL-10)上,本文方法与全监督训练基线以及最先进半监督方法的性能比较。对于半监督图像分类任务,本文实验在基准数据集上使用了不同的初始标签数量设定,以展示对于不同数量有标签数据的半监督任务,算法进行标签传播和无标签数据利用的性能表现差异。

表 1　在 CIFAR-10 数据集上与最先进方法的性能比较(带†方法的主干网络是 CNN-13,其余方法主干网络是 WRN-28;带‡结果基于代码库复现得到,其余结果来自原文公布)

方法	CIFAR-10(错误率/%)		
	40 标签	250 标签	4 000 标签
仅监督训练	64.01 ± 0.76	39.12 ± 0.77	12.74 ± 0.29
MixMatch[4]	47.54 ± 11.50	11.05 ± 0.86	6.42 ± 0.10
UDA[12]	29.05 ± 5.93	8.82 ± 1.08	4.88 ± 0.18
ReMixMatch[3]	19.10 ± 9.64	5.44 ± 0.05	4.72 ± 0.13
FixMatch[11]	13.81 ± 3.37	5.07 ± 0.65	4.26 ± 0.05
Curriculum Labeling[7]	-	-	5.09 ± 0.18
FlexMatch[14]	5.93 ± 0.41‡	5.23 ± 0.32‡	4.31 ± 0.07‡
SemCo[15]	-	5.12 ± 0.27	3.80 ± 0.08
UPS[9]	-	-	6.39 ± 0.02†
CoMatch[35]	6.91 ± 1.39	4.91 ± 0.33	-
Dash[36]	13.22 ± 3.75	4.56 ± 0.13	4.08 ± 0.06
FixMatch + Sel + Reg[37]	8.48 ± 2.81	5.57 ± 0.65	-
LESS[38]	6.80 ± 2.10	4.90 ± 0.80	-
CADR[39]	5.59	5.65	4.41
本文方法	**5.48 ± 1.12**	**4.42 ± 0.52**	**3.94 ± 0.10**

表 2　在 CIFAR-100 数据集上与最先进方法的性能比较(带†方法的主干网络是 CNN-13,其余方法主干网络是 WRN-28)

方法	CIFAR-100(错误率/%)		
	400 标签	2 500 标签	10 000 标签
仅监督训练	79.47 ± 0.18	52.88 ± 0.51	32.55 ± 0.21
MixMatch[4]	67.61 ± 1.32	39.94 ± 0.37	28.31 ± 0.33
UDA[12]	59.28 ± 0.88	33.13 ± 0.22	24.50 ± 0.25
ReMixMatch[3]	44.28 ± 2.06	27.43 ± 0.31	23.03 ± 0.56
FixMatch[11]	48.85 ± 1.75	28.29 ± 0.11	22.60 ± 0.12
SemCo[15]	-	31.93 ± 0.01	24.45 ± 0.12
UPS[9]	-	-	32.00 ± 0.49†
Dash[36]	44.76 ± 0.96	27.18 ± 0.21	21.97 ± 0.14
FixMatch + Sel + Reg[37]	51.99 ± 1.72	34.79 ± 0.49	-
LESS[38]	48.70 ± 2.40	-	-
CADR[39]	47.10	29.39	23.07
本文方法	**40.56 ± 0.92**	**25.82 ± 0.22**	**20.96 ± 0.12**

如表 1 所示,在 CIFAR-10 数据集上,本方法对于不同的初始标签数量(每类别 4、25、400 个标签)设定都产生了非常有竞争力的结果。这一显著的性能提升可以归功于本方法合理分配样本动态权重并

基于此实现了更合理的半监督学习思路和更有效的无标签数据利用能力。如表 2 所示,在 CIFAR-100 数据集上,本方法对于不同的初始标签数量设定(每类别 4、25、100 个标签)所产生的结果都达到了最先进的水平。值得注意的是,CIFAR-100 数据集包含更多的类别数量,由此,本方法有效组织了众多类别的分类面构建顺序,从而能够产生更显著的性能提升。如表 3 所示,在 STL-10 数据集上,本文方法的性能同样优于竞争对手。由此,可以证明本方法对于更大和更复杂的数据集同样能够表现出稳定且出色的效果。此外,本文方法在较少初始标签数量设定的实验中效果提升更加显著,这归功于本方法对于类内和类间无标签样本进行了更合理的利用,有效降低了噪声干扰,提高了更困难标签传播场景的准确率。

表 3　在 STL-10 数据集上与最先进方法的性能比较

方法	STL-10(错误率/%)
	1 000 标签
仅监督训练	20.66 ± 0.83
Π-Model[17]	26.23 ± 0.82
Mean Teacher[6]	21.43 ± 2.39
MixMatch[4]	10.18 ± 1.46
UDA[12]	7.66 ± 0.56
FixMatch[11]	7.98 ± 1.50
Dash[36]	7.26 ± 0.40
本文方法	**6.16 ± 0.84**

3.3.2　消融实验结果

消融实验能够定量反映本方法添加每个模块后对算法效果的影响,由此直观地展示本方法中每个模块对性能提升的贡献。

在 CIFAR-10 数据集上,使用不同的初始标签数量(每类别 4、400 个标签)设定,进行了消融实验分析。消融实验结果见表 4,其中,“全监督基线”表示仅使用有标签数据训练,“半监督基线”表示使用半监督基本框架,利用有标签和无标签数据训练,“课程 1”表示引入基于类内样本动态权重的课程式学习方法,“课程 2”表示引入基于类间样本动态权重的课程式学习方法,“课程 1 + 课程 2”表示联合使用 2 种课程。诚然,半监督方法的最大性能提升源自有效利用无标签数据。但是,本方法中所提出的 2 个课程式学习方法(即基于类内样本动态权重的课程式学习与基于类间样本动态权重的课程式学习)均能够在基线的基础上,进一步提升性能。值得注意的是,更少初始标签数量(每类别 4 个标签)设定的实验更具有挑战性,因此有效改进所带来的效果增益通常更显著。联合使用 2 种课程式学习方法,相较使用单一课程式学习,性能提升更加显著。这主要是由于本工作中所提出的 2 个课程式方法之间相互正交,分别作用于样本的类内和类间层面。由此,基于课程式学习的改进,由简单至困难地选择重点学习样本、逐步构建分类面,能够有效地提升半监督学习方法的性能。

表 4　消融实验结果(错误率/%)

方法	40 标签	4 000 标签
全监督基线	64.01 ± 0.76	12.74 ± 0.29
半监督基线	13.81 ± 3.37	4.26 ± 0.05
课程 1	8.18 ± 1.50	4.10 ± 0.11
课程 2	6.93 ± 0.96	4.06 ± 0.08
课程 1 + 课程 2	5.48 ± 1.12	3.94 ± 0.10

3.3.3　分析性实验结果

除主实验与消融实验外,本文补充了更多分析性实验以便更全面、深入地展示本方法的整体效果,主要包括训练过程中准确率及损失变化曲线、课程作用时间段分析、伪标签化过程改进效果分析和使用 t-SNE[40] 进行特征可视化。

对于训练过程中准确率及损失变化曲线分析,在 CIFAR-10 数据集上使用40 初始标签数量设定进行了实验测定。如图 4 所示,本文方法相较基线半监督方法具有显著的性能提升。具体而言,相较基线方法,本方法对应的测试错误率和损失值下降更快,收敛时的准确率明显提升,过拟合现象也在一定程度上得到有效缓解。

对于本方法中课程作用时间段的分析,重点粗粒度地探究在 2 种课程式学习算法中,课程式权重调节作用时间段(前期训练阶段、前中期训练阶段以及全部训练阶段)对于算法最终性能的影响。将课程作用的总训练轮次 T 分别设置为200(前期

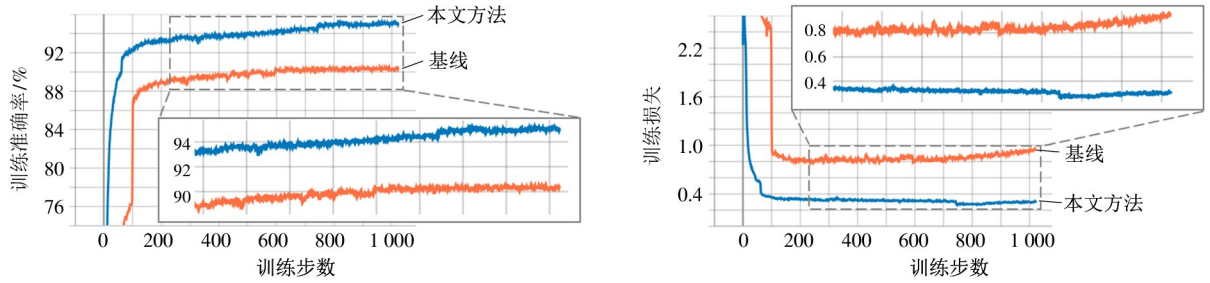


图 4　在 CIFAR-10 数据集上 40 标签数量设定的训练过程中，本文方法与基线的准确率及损失变化曲线对比

阶段）、500（前中期阶段）和 1 000（全部阶段），并在 CIFAR-10 数据集上使用 40 初始标签数量设定进行了实验测定。实验结果和效果对比如图 5 所示。可以观察到，2 种课程作用的时间段对算法效果没有非常显著的影响，尤其是基于类间样本动态权重的课程式学习方法（课程 2）。也就是说，本方法对课程作用阶段并不敏感。对于算法的具体应用而言，只要保证在模型训练的早期阶段使用一段时间进行简单样本和分类面的基础性学习就可以发挥显著作用。

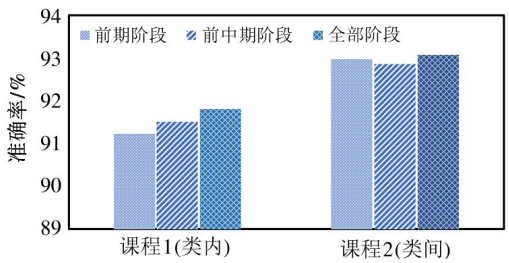
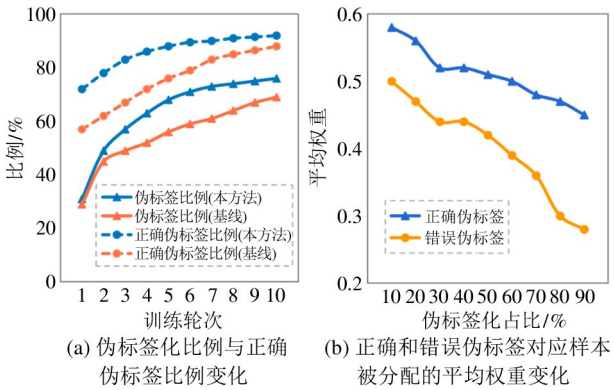


图 5　课程学习作用时间段对算法效果的影响

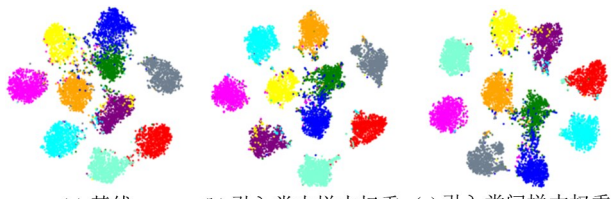
对于伪标签化过程改进效果分析如图 6 所示，由图可知，本文方法相较基线能够快速产生大量伪标签指导半监督模型训练；同时，全部伪标签中正确的伪标签比例相较基线也大幅提升。此外，随着伪标签化占比逐步增加，噪声样本被成功地分配了更低的权重。这一分析结果可以证明，基于样本动态权重的课程式半监督学习方法，在很大程度上改善了半监督中关键的机制——伪标签化过程，这得益于本文方法由简单至困难地学习样本、构建分类面，减少了噪声产生和错误累积。由于伪标签化的效率和准确率大幅提升，本方法对于半监督学习的性能有着极大程度的增益。



(a) 伪标签化比例与正确伪标签比例变化　(b) 正确和错误伪标签对应样本被分配的平均权重变化

图 6　本方法与基线关于伪标签化过程的对比

对于使用 t-SNE[40] 对学习到的特征描述进行可视化，结果如图 7 所示。可以直观地注意到，引入基于类内样本权重的课程式学习后学到的特征描述由于缓解了离群点和不确定样本引起的噪声训练而呈现出更好的类内紧凑性，而引入基于类间样本权重的课程式学习后学到的特征描述具有更好的类间可分离性，这可以归功于更合理、更有效的分类面构建顺序。因此，利用样本动态权重设计课程式学习方法，有利于获得更出色的模型分类结果。



(a) 基线　(b) 引入类内样本权重　(c) 引入类间样本权重

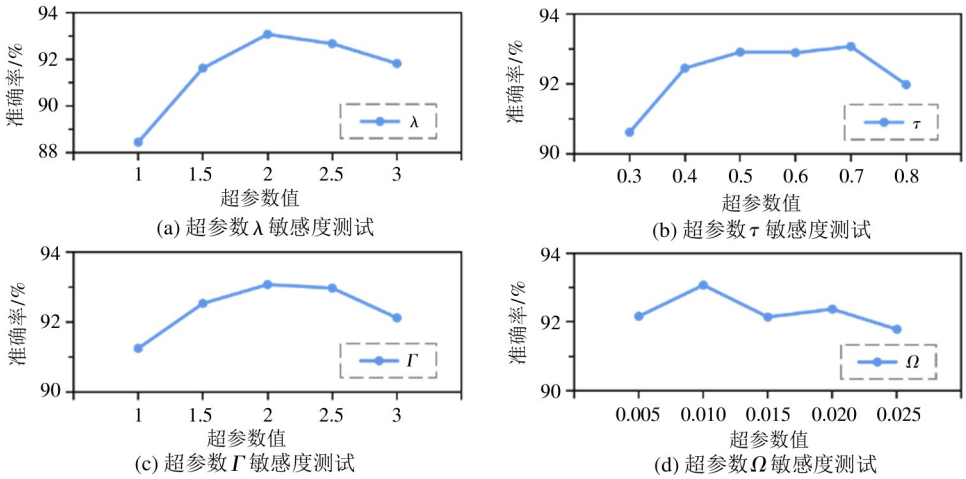
图 7　在 CIFAR-10 数据集上使用 40 标签设定，利用 t-SNE 进行特征可视化

3.3.4　超参数分析

对于本文方法引入的超参数，在 CIFAR-10 数据集上使用 40 标签设定进行了敏感度测试。具体地，测试的超参数包括：$\lambda \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$，

$\tau \in \{0.3,0.4,0.5,0.6,0.7,0.8\}$, $\Gamma \in \{1.0,1.5,2.0,2.5,3.0\}$, $\Omega \in \{0.005,0.010,0.015,0.020,0.025\}$。

结果如图 8 所示,由图可知,本文方法在一定程度上对超参数的具体取值并不敏感,这得益于课程学习的机制使得学习过程平滑,许多关键超参数具有一定的自我调节空间,这也可以从侧面证明本文方法的鲁棒性。



(a) 超参数 λ 敏感度测试　(b) 超参数 τ 敏感度测试

(c) 超参数 Γ 敏感度测试　(d) 超参数 Ω 敏感度测试

图 8　超参数分析,使用 CIFAR-10 数据集进行实验测定

4　结　论

本文介绍了基于样本动态权重的课程式半监督学习方法,旨在鼓励模型由简单至困难地利用样本,逐步构建分类面,进而缓解伪标签化过程中噪声的产生,增强模型的泛化能力。为区分类内样本学习难度,本文方法通过混合标签和高置信度伪标签信息构建特征原型,并基于此计算样本动态权重,进而课程式地调整样本对损失的贡献。为评估类别间语义相似度,本文引入标签嵌入方法 GloVe,并基于此调整置信度函数,课程式地调整样本所占权重,逐步构建分类面。在 3 个半监督学习基准数据集上进行的一系列实验和分析表明,本文方法能够有效改善半监督学习中的焦点问题,在性能方面达到了先进的水平,且在实际应用也具有重要意义。本文工作中使用的类内样本学习难度和类间语义相关度的评估方法并不局限于此,未来工作中,可以尝试探索更多可行性的替代和优化方法,例如,构建语义层级结构等。

参考文献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:770-778.

[2] TAN M, LE Q. EfficientNet: rethinking model scaling for convolutional neural networks[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA:ICML, 2019:6105-6114.

[3] BERTHELOT D, CARLINI N, CUBUK E. D, et al. ReMixMatch: semi-supervised learning with distribution matching and augmentation anchoring[C]//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020:1-10.

[4] BERTHELOT D, CARLINI N, GOODFELLOW I J, et al. MixMatch: a holistic approach to semi-supervised learning[C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS, 2019:5050-5060.

[5] MIYATO T, MAEDA S, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,41(8):1979-1993.

[6] TARVAINEN A, VALPOLA H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning result[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, USA: NIPS, 2017:1195-1204.

[7] CASCANTE-BONILLA P, TAN F, QI Y, et al. Curriculum labeling: revisiting pseudo-labeling for semi-super-

vised learning[C]//The 35th AAAI Conference on Artificial Intelligence. Online: AAAI Press, 2021: 6912-6920.
[8] ISCEN A, TOLIAS G, AVRITHIS Y, et al. Label propagation for deep semi-supervised learning[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:5070-5079.
[9] RIZVE M. N, DUARTE K, RAWAT Y S, et al. Indefense of pseudo-labeling: an uncertainty-aware pseudo-label selection framework for semi-supervised learning[C]//The 9th International Conference on Learning Representations. Online: ICLR, 2021:1-19.
[10] SHI W, GONG Y, DING C, et al. Transductive semi-supervised deep learning using min-max features[C]// Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: ECCV, 2018:311-327.
[11] SOHN K, BERTHELOT D, CARLINI N, et al. FixMatch: simplifying semi-supervised learning with consistency and confidence[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. Online: NIPS, 2020:596-608.
[12] XIE Q, DAI Z, HOVY E H, et al. Unsupervised data augmentation for consistency training[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. Online: NIPS, 2020:6256-6268.
[13] CUBUK E D, ZOPH B, SHLENS J, et al. Randaugment: practical automated data augmentation with a reduced search space[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE, 2020:3008-3017.
[14] ZHANG B, WANG Y, HOU W, et al. FlexMatch:boosting semi-supervised learning with curriculum pseudo labeling[C]//Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021. Online: NIPS, 2021:18408-18419.
[15] NASSAR I, HERATH S, ABBASNEJAD E, et al. All labels are not created equal: enhancing semi-supervision via label grouping and co-training[C]//IEEE Conference on Computer Vision and Pattern Recognition. Online: IEEE, 2021:7241-7250.
[16] BACHMAN P, ALSHARIF O, PRECUP D. Learning with pseudo-ensembles[C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. Montreal, Canada: NIPS, 2014:3365-3373.
[17] LAINE S, AILA T. Temporal ensembling for semi-supervised learning[C]//The 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017: 1-13.
[18] ZHANG H, CISSÉ M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[C]//The 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018:1-13.
[19] BEYER L, ZHAI X, OLIVER A, et al. S^4L: self-supervised semi-supervised learning[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:1476-1485.
[20] KUO C, MA C, HUANG J, et al. Featmatch: feature-based augmentation for semi-supervised learning[C]// The 16th European Conference on Computer Vision. Glasgow, UK: ECCV, 2020:479-495.
[21] LEE D H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks[C]//ICML 2013 Workshop: Challenges in Representation Learning. Atlanta, USA: ICML, 2013:1-6.
[22] YI K, WU J. Probabilistic end-to-end noise correction for learning with noisy labels[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:7017-7025.
[23] WANG G, WU J. Repetitivere prediction deep decipher for semi-supervised learning[C]//The 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020:6170-6177.
[24] WANG X, CHEN Y, ZHU W. A survey on curriculum learning[J]. IEEE Transactionson Pattern Analysis and Machine Intelligence, 2022,44(9):4555-4576.
[25] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ICML, 2009:41-48.
[26] YU Q, IKAMI D, IRIE G, et al. Multi-task curriculum framework for open-set semi-supervised learning[C]// The 16th European Conference on Computer Vision. Glasgow, UK:ECCV, 2020:438-454.
[27] MIKOLOV T, CHEN K, CORRADO GREG, et al. Efficient estimation of word representations in vector space [C]//The 1st International Conference on Learning Representations (Workshop Track Proceedings). Scottsdale, USA:ICLR, 2013:1-12.
[28] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014:1532-1543.
[29] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images [R]. Toronto: University of Toronto, 2009.
[30] COATES A, NG A Y, LEE H. An analysis of single-layer networks in unsupervised feature learning[C]//The 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: AISTATS, 2011: 215-223.
[31] RUSSAKOVSKY O, DENG J, SU H, et al. Image net

large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015,115(3):211-252.
[32] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in Pytorch[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 Workshop. Long Beach, USA: NIPS, 2017:1-4.
[33] LOSHCHILOV I, HUTTER F. SGDR: stochastic gradient descent with warm restarts[C]// The 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017:1-16.
[34] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout [EB/OL]. (2017-11-29) [2023-02-07]. https://arxiv.org/pdf/1708.04552v2.pdf.
[35] LI J, XIONG C, HOI S C H. Comatch: semi-supervised learning with contrastive graph regularization [C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021:9455-9464.
[36] XU Y, SHANG L, YE J, et al. Dash: semi-supervised learning with dynamic thresholding [C]// Proceedings of the 38th International Conference on Machine Learning. Online: ICML, 2021:11525-11536.
[37] KIM N R, LEE J H. Propagation regularizer for semi-supervised learning with extremely scarce labeled samples [C]// Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022:14381-14390.
[38] LUCAS T, WEINZAEPFEL P, ROGEZ G. Barely-supervised learning: semi-supervised learning with very few labeled images[C]// The 35th AAAI Conference on Artificial Intelligence. Online: AAAI Press, 2022:1881-1889.
[39] HU X, NIU Y, MIAO C, et al. On non-random missing labels in semi-supervised learning [C]// The 10th International Conference on Learning Representations. Online: ICLR, 2022:1-12.
[40] MAATEN L, HINTON G. Visualizing data using T-SNE [J]. Journal of Machine Learning Research, 2008, 9: 2579-2605

Curriculum paradigm based on the dynamic weights of samples for semi-supervised learning

ZHU Hui[*,**], HU Bin[*], SONG Yining[***], ZHAO Xiaofang[*,****]
([*] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)
([**] University of Chinese Academy of Sciences, Beijing 100049)
([***] Information Center of National Defense Mobilization Department of Central Military Commission, Beijing 100034)
([****] Institute of Intelligent Computing Technology, Suzhou, Chinese Academy of Sciences, Suzhou 215028)

Abstract

This work studies the difficulty of label propagation and serious noise interference in model training, which are due to the extreme lack of supervision signals in semi-supervised learning scenarios. Noise from pseudo-labeling and confirmation bias caused by low data utilization will lead to error accumulation along with the self-training process, thus forming irreversible deviation and damaging the performance. In this paper, a curriculum paradigm based on the dynamic weights of samples for semi-supervised learning is proposed, aiming at encouraging the model to utilize samples from easy to hard and gradually construct hyperplanes based on the non-discrete curriculum, so as to alleviate the generation of noise in the pseudo-labeling process and enhance the generalization ability of the model. Specifically, from the intra-class perspective, prototypes of features are constructed by mixing pseudo-labels with high confidence, which can provide weak supervision signals. Then, the learning difficulties of samples are estimated. From the inter-class perspective, label embedding is used to evaluate the semantic relevancy between categories, and the discrimination between semantically related categories are weaken in the early stage of training. Comprehensive experiments and analyses are conducted on commonly-used semi-supervised learning benchmark datasets to demonstrate the effectiveness of this method.

Key words: semi-supervised learning, feature representation vector, curriculum learning, prototype of features, semantic relevancy