

伪标签半监督通信辐射源个体识别方法

吕昊远, 俞璐, 陈璞

(陆军工程大学通信工程学院, 南京 210007)

摘要: 针对通信辐射源个体识别技术中有标签信号样本不足导致个体识别准确率较低的问题, 提出了基于伪标签半监督深度学习的辐射源个体识别方法, 该方法利用加权平均思想改进了伪标签的赋值方式, 有效增强了伪标签的质量, 提升了网络模型的鲁棒性; 介绍了如何基于伪标签思想设计半监督深度学习的方法, 并运用熵正则化算法的概念从理论方面解释了伪标签的有效性; 实验设计了适合于信号样本的卷积神经网络, 采取不同数目的有标签样本与无标签样本组建的训练集方案, 得到了改进的伪标签半监督方法在测试集的识别准确率, 结果表明, 该方法较全监督方法和改进前的伪标签半监督方法有着更好的识别效果和更强的优越性。

关键词: 辐射源个体识别; 伪标签半监督学习; 加权平均; 熵正则化; 卷积神经网络

Pseudo Label Semi-supervised Communication Emitter Identification Method

LÜ Haoyuan, YU Lu, CHEN Pu

(College of Communication Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: Aiming at the problem of low recognition accuracy caused by insufficient labeled signal samples in communication emitter identification technology, a semi-supervised deep learning method for emitter identification based on pseudo label is proposed. This method uses the idea of weighted average to improve the assignment method of pseudo label, effectively enhances the quality of pseudo label, and improves the robustness of network model. This paper introduces how to design a semi-supervised deep learning method based on the idea of pseudo label, and explains the effectiveness of pseudo label theoretically by using the concept of entropy regularization algorithm. Convolutional neural network is designed for signal samples. Taking different number of labeled samples and unlabeled samples as training set, the recognition accuracy of the improved pseudo label semi-supervised method in the test set is obtained. The results show that the method has better recognition effect and stronger superiority than the full supervised method and the original pseudo label semi-supervised method.

Keywords: emitter identification; pseudo label semi-supervised learning; weighted average; entropy regularization; convolutional neural network

0 引言

在现代战争日趋复杂的电磁环境空间中, 电子对抗发挥着举足轻重的作用。作为重要的非合作识别手段, 电子侦察是电子对抗的基础和前提^[1]。电子侦察手段将获取通信辐射源信号数据的幅度、脉宽、载频等参数经过处理, 得到通信辐射源目标的信息, 从而做到对于信息制取权的实时掌握^[2]。

随着技术的不断发展, 新型多功能辐射源信号参数呈现出多变性, 信号样式也变得越来越复杂, 不同模式、个体间的信号参数交错, 对于辐射源目标的识别, 传统电子侦察识别方法就愈加困难。为了能够准确地识别、干扰、打击敌方目标, 首要环节就是做到辐射源个体区分, 由此提出辐射源个体识别 (SEI, specific emitter identification) 技术^[3]。“个体特征”就是由通信设备内部的不同制造工艺

而导致所发射信号的细微差异。辐射源个体识别技术通过测量手段获取信号的外部特征, 提取出反映目标身份的信息, 确定发射所得信号的特定辐射源个体^[4]。

作为人工智能领域近十年来最受关注的技术之一, 深度学习通过神经网络将底层的特征映射到高层, 并且通过高层将特征抽象出来, 从而发现数据的分布式特征表示^[5]。深度学习的快速发展为通信辐射源个体识别提供了新的处理思路和研究方法, 深度学习直接从输入的信号数据中提取个体特征, 跨越人工设计特征阶段^[6], 应用在辐射源个体识别问题中可以节省大量科研成本。

深度网络训练需要庞大的数据集, 但在实际环境中获取到具有标签信息的信号样本数量非常有限, “小样本”问题严重制约了通信辐射源个体识别技术的应用与发展^[7], 在这种情况下, 深度学习方法中引入半监督的思想, 充分利用大量无标签信号样本内在的结构信息提取通信辐射源

收稿日期: 2021-03-18; 修回日期: 2021-05-10。

基金项目: 国家自然科学基金(61702543)。

作者简介: 吕昊远(1997-), 男, 山西晋中人, 硕士研究生, 主要从事深度学习、辐射源个体识别方向的研究。

俞璐(1973-), 女, 吉林长春人, 博士, 副教授, 主要从事多媒体信息处理、模式识别、图像处理方向的研究。

引用格式: 吕昊远, 俞璐, 陈璞. 伪标签半监督通信辐射源个体识别方法[J]. 计算机测量与控制, 2021, 29(7): 229-234.

个体本征特征,从半监督学习的角度提高通信辐射源个体识别技术的性能,这在信号侦察对抗和无线电监视管控等领域具有重大的研究意义和应用价值^[8]。

先前的研究工作中,基于端到端半监督深度学习的通信辐射源个体识别技术研究较少。黄健航^[9]提出的通信辐射源个体识别方法用到了半监督矩形网络^[10],苟嫣^[11]提出的迁移极限学习机采用基于实例的迁移学习方法。半监督学习方面,Lee^[12]提出的基于伪标签的半监督深度学习方法将伪标签训练机制加入到深度卷积神经网络中,在MNI-ST图像数据集中达到较高的识别准确率。

本文在伪标签半监督方法的基础上,提出一种基于伪标签半监督深度学习的通信辐射源个体识别方法,并在生成伪标签中使用加权平均的思想进行改进,在5台同型号USRP辐射源采集的数据集上验证算法的鲁棒性。并与全监督方法和改进前的伪标签半监督方法进行比对,实验结果显示,改进后的伪标签半监督算法在少量有标签信号样本条件下能够达到更好的识别性能。

1 伪标签半监督

在基于伪标签的半监督学习方法中,深度网络以有监督的方式同时训练有标签和无标签数据。对于无标签数据,使用网络的预测输出作为其伪标签,加入到网络中再次训练,充分挖掘出无标签样本中与有标签样本完全相同的个体信息,最终在测试集上取得较好的识别效果^[13]。

本章分别介绍半监督学习的思想和基于伪标签的半监督深度学习方法,并引入熵正则化的概念,从理论方面说明伪标签方法的有效性。

1.1 半监督学习

相比于稀缺的有标签样本,无标签样本数量众多且相对容易获取。为了能够有效利用宝贵的有标签样本,也为了使大量且相对便宜的无标签样本不至于浪费而发挥出最大效用,研究学者们提出了半监督学习技术^[14]。

通过近些年的不断探索发展,半监督学习已经成为机器学习领域的研究热点。它的基本思想是使用大量无标签样本辅助少量有标签样本进行学习,达到提高学习效果的目的^[15]。如图1(a)所示,有4个分属于两个不同类别的有标签样本点,当仅考虑这4个有标签样本点时,只能假设出线性决策边界,但如果考虑到数量较多的无标签样本点(白色图形)时,如图1(b)所示,就可以绘制出正确的非线性决策边界而确定真正的样本分布。

因此,在半监督学习中,结合使用稀少的有标签和海量无标签样本,得到更精确的决策边界来提高模型的推广性能^[16]。

1.2 伪标签方法

伪标签就是对于无标签样本数据,通过网络预测结果,赋予其标签值。赋标签值时只选择对每个未标记样本具有最大预测概率的类别。

$$y'_i = \begin{cases} 1 & \text{if } i = \operatorname{argmax}_f f_i(x) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

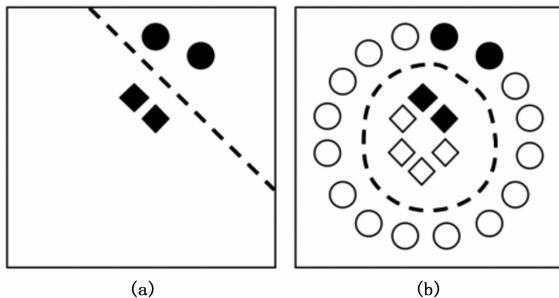


图1 半监督学习示意图

再以有监督的方式同时使用有标签和无标签数据对网络进行训练。对于未标记的数据,其伪标签值要随着网络权重值的变化而不断更新^[17]。由于伪标签的可信度以及有标签数据和无标签数据的数量规模的差距,两部分损失之间的平衡对于网络训练的性能至关重要,因此整体损失函数为:

$$Loss = \frac{1}{n} \sum_{m=1}^n \sum_{i=1}^C L(y_i^m, f_i^m) + \alpha(t) \frac{1}{n'} \sum_{m=1}^{n'} \sum_{i=1}^C L(y_i'^m, f_i'^m) \quad (2)$$

C 是类别数, n 是有标签样本数目, n' 是无标签样本数目, L 是交叉熵损失函数, f 是神经网络对于有标签样本的输出预测, y 是对应的标签, f' 是神经网络对于无标签样本的输出预测, y' 是其对应的伪标签。 $\alpha(t)$ 是平衡两部分损失权重的系数函数, t 是当前的迭代轮次。

在训练过程中, $\alpha(t)$ 决定着无标签数据损失部分在整体损失中所占比重的大小。在训练初始阶段,因为网络输出预测不够准确,所以赋予的伪标签值准确性较低,此时如果 $\alpha(t)$ 值过大,增大无标签数据损失权重会导致训练性能退化,但 $\alpha(t)$ 值太小就不能充分利用无标签数据的好处,对于网络性能提升有限。设置 $\alpha(t)$ 为退火过程,初始值为0,再随着训练迭代次数的增加而缓慢增长并最终固定,网络的预测能力也会随之增强^[18]。

$$\alpha(t) = \begin{cases} 0 & t < T_1 \\ \frac{t-T_1}{T_2-T_1} \alpha_f & T_1 \leq t \leq T_2 \\ \alpha_f & T_2 \leq t \end{cases} \quad (3)$$

T_1 为加入无标签损失时的起始轮次值, T_2 为权重固定的轮次值, α_f 为最终无标签损失部分的权重固定值,3个值的设置需要在实验中结合具体应用场景设置。这个退火过程的设置可以使得优化中避免较差的局部最小值,使得无标签数据的伪标签尽可能地类似于真实标签。

1.3 熵正则化

理想的深度网络模型中,不同的辐射源信号在特征空间分类边界上密度较小,而来自同一个辐射源的信号特征距离较近、密度较大并且具有连续性。熵正则化从这个思想出发,在最大后验估计框架下充分利用无标签信号样本^[19]。熵正则化通过最小化无标签数据类别的条件熵来支持类之间的低密度分离,其表示形式如下:

$$H(y | x') = -\frac{1}{n'} \sum_{m=1}^{n'} \sum_{i=1}^C P(y_i^m = 1 | x'^m) \log P(y_i^m = 1 | x'^m) \quad (4)$$

n' 是无标签信号样本数目, C 是辐射源的类别数目, y_i^m 是第 m 个无标签信号样本的假设标签, x'^m 是第 m 个无标签信号样本的输入, 最大化后验概率分布为:

$$C(\theta, \lambda) = \sum_{m=1}^n \log P(y^m | x^m; \theta) - \lambda H(y | x'; \theta) \quad (5)$$

n 是有标签信号样本的数量, x^m 是第 m 个有标签信号样本, λ 是权重系数平衡两部分。熵正则化的意义在于让模型对于无标签信号样本的识别概率尽可能地集中在某一类, 减小决策边界附近的信号密度。通过最大化有标签信号数据的条件对数似然和最小化无标签样本的熵, 使得深度网络的泛化能力增强。通过比较式 (5) 和式 (2) 可以看得, 前文所提的伪标签方法相当于熵正则化, 式 (5) 中的前后两项分别对应式 (2) 中的前后两项, 权重系数 λ 对应于 α 。

2 深度学习方案设计

2.1 算法过程

基于伪标签半监督深度学习的通信辐射源个体识别方法训练流程如图 2 所示。

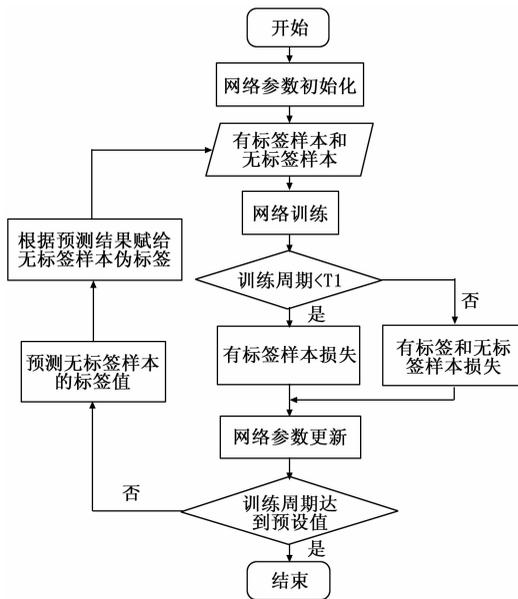


图 2 训练流程

首先初始化深度网络的参数, 少量有标签和大量无标签信号样本构成算法的输入数据, 接着开始训练样本数据, 由于在训练初始阶段模型的预测准确率不高, 伪标签的可信度非常低, 所以当训练周期数小于 T_1 时, 只计算有标签数据部分的损失值。当模型训练到一定阶段 (周期数大于 T_1) 时, 网络的预测能力增强, 这时伪标签信息相对可信, 结合有标签和无标签信号样本两部分损失, 构建整体损失函数, 并根据式 (3), 在接下来的训练中, 动态调整两部分损失的权重系数。

伪标签原算法在一轮迭代中, 将网络的输出预测值强

锐化作为样本的伪标签, 然后在同一轮中计算输出预测和伪标签的损失值, 这对于本实验的辐射源信号数据而言, 在训练初期伪标签的优劣条件下都存在问题 (具体如表 1 所示)。本文引入加权平均的思想, 提出改进的轮次标签法。

如图 3 所示, x 是无标签信号样本输入, 将 x 在上一个训练轮次中的预测结果 \bar{y} 与本轮次的预测结果 \tilde{y} 进行加权平均计算, 得到一个新的预测值 \hat{y} 作为其伪标签。

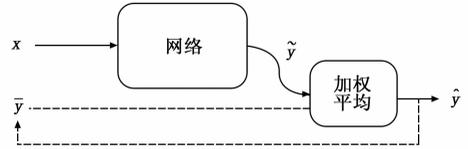


图 3 轮次标签

轮次标签法在训练初期, 有效增强了伪标签的质量, 提升了网络模型的鲁棒性, 具体的算法改进前后, 在预测正确和错误两种情形下的比较如表 1 所示。

表 1 算法改进前后比较

	预测正确	预测错误
原方法	无标签样本损失过小, 无法发挥作用	伪标签产生错误的指导, 严重影响模型训练
轮次标签	增大无标签样本损失值, 强化作用	平滑错误预测, 增强模型对错误预测的容忍

在训练过程每个周期内, 网络对于无标签信号样本都要预测标签值, 将加权平均结果作为其伪标签, 然后就可以作为有标签信号样本, 加入到训练集中重复训练, 直至训练周期达到预设值结束。

2.2 网络结构

根据输入信号样本的形式, 以及对网络的需求关系, 设计出如图 4 所示的 CNN 模型, 除输入输出层外, 中间的特征提取过程包括两个卷积层和两个全连接层。具体的信号样本划分, 构建数据集会在下一章中说明。

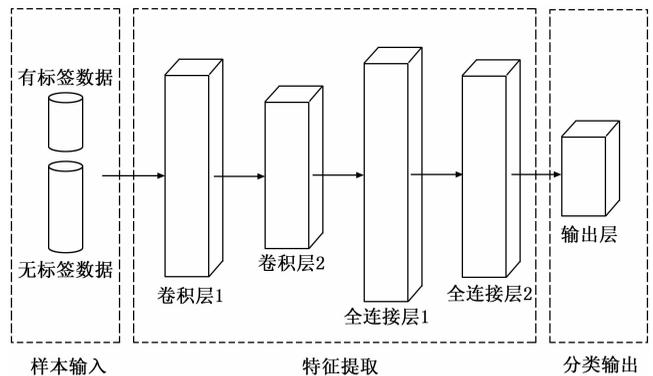


图 4 CNN 结构图

网络的输入样本维数为 2×128 , 两个卷积层的核数量分别为 32, 16, 大小分别是 (1, 3), (2, 3)。第一个全连接层的神经元个数为 128, 第二个全连接层神经元个数为

5, 对应 5 个辐射源个体的 5 分类问题, 最后通过 softmax 层输出预测值。除最后一层使用 softmax 激活函数, 其余层使用 ReLU 激活函数。并且在每层后会连接至 dropout 层进行正则化, 参数设置为 0.1, 实验选用交叉熵损失函数和 Adam 优化器。

3 实验与分析

3.1 实验数据准备

为了评估基于伪标签半监督深度学习算法对于通信辐射源个体识别问题的可行性与有效性, 在实验室条件下, 采用 6 台 USRP N210 设备作为信号样本采集的辐射源, 其中 5 台作为发射端, 固定 1 台为接收端, 在采取的辐射源信号上实验验证。

发送端和接收端程序采用 LabVIEW 软件实现, USRP 通过网络电缆与计算机相连^[20]。所采集的数据是相互正交的 IQ 两路载波信号。采集数据过程中设置频率带宽为 1 GHz, 采样频率为 1 MHz。

对采集到的通信辐射源信号进行预处理制作成数据样本集, 去除采样帧初始阶段帧与帧切换时产生的不规则样本点, 并且进行功率归一化比例变换, 再设置 5 类标签值, 将辐射源个体类别进行标号。

将所采集到的信号样本按照一定比例划分为训练集和测试集, 并且选取训练集中的少量样本作为有标签训练集, 和大量样本作为无标签训练集。设置有标签信号样本数为 1 000 和 2 000。有标签信号样本与无标签信号样本的比例大小为 (1: 3)。测试集的样本数为 500。

3.2 实验结果分析

本文使用深度学习框架 PyTorch 实现对深度神经网络的构建以及算法的训练与测试。计算机系统环境为 Windows10-64-bit, 开发环境为 anaconda+pytorch+pycharm, 软硬件环境配置如表 2 所示。

表 2 软硬件环境

硬件	CPU	Intel(R) Core(TM) i7-10870H
	内存	DDR4 8G * 2
	显卡	Nvidia RTX 2070 Max-Q
软件	Anaconda 版本	Anaconda3 2020.02
	Python 版本	python 3.7.9
	Pytorch 版本	torch 1.7.1

半监督训练过程中, 设置 200 次的迭代次数, 根据多次实验效果, 设置 T_1 的值为 30, T_2 的值为 180, α_f 的值为 0.2。

图 5~8 分别绘制在不同有标签信号样本数量条件下, 训练过程中训练样本的损失函数和识别准确率随着迭代次数增加的变化曲线。识别准确率包括全监督训练和半监督训练, 半监督训练中又分为有标签信号样本和无标签信号样本两部分, 其中无标签样本的原有标签值只用在每次迭代后计算样本的识别准确率, 不投入深度网络指导训练。

可以看出, 当迭代次数小于 T_1 时, 两种训练方式的损

失和准确率的变化趋势一致, 当迭代次数为 T_1 时, 半监督训练过程中由于加入了无标签信号样本这部分损失值, 整体的训练损失值产生波动, 在 T_2 之后逐渐趋于稳定。与全监督训练比较, 因为半监督训练中使用了更多的无标签信号样本, 所以损失值收敛的速度更慢波动更大, 且随着无标签信号样本的数量增多, 这一现象会更加明显。半监督训练过程, 相比较于有标签信号样本, 无标签信号样本的识别准确度较低, 这是因为有标签信号样本在训练过程中使用了真实标签值, 而无标签信号样本而无标签信号样本在训练过程中使用的伪标签值与真实标签值还存在着一些差异。

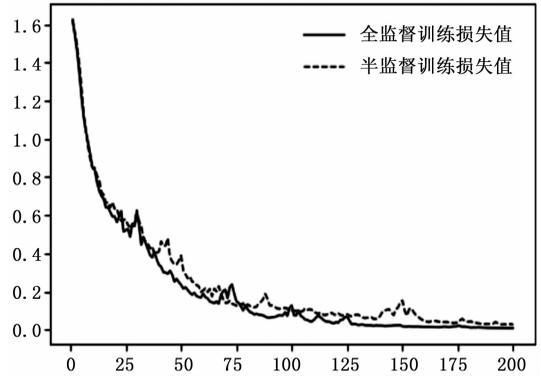


图 5 1 000 个有标签样本的训练过程中的损失值

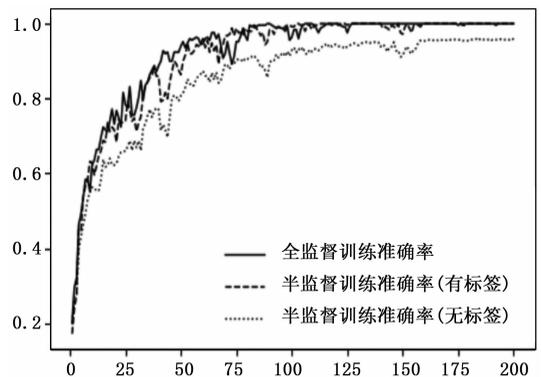


图 6 1 000 个有标签样本的训练过程中的准确率

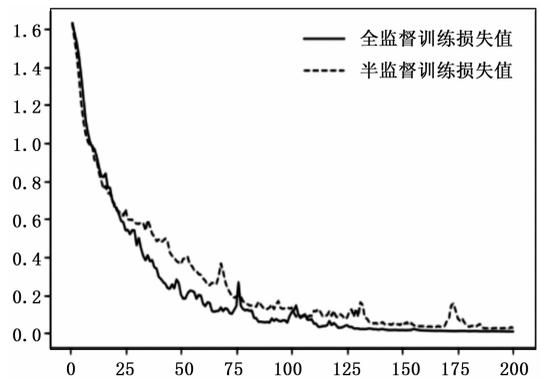


图 7 2 000 个有标签样本的训练过程中的损失值

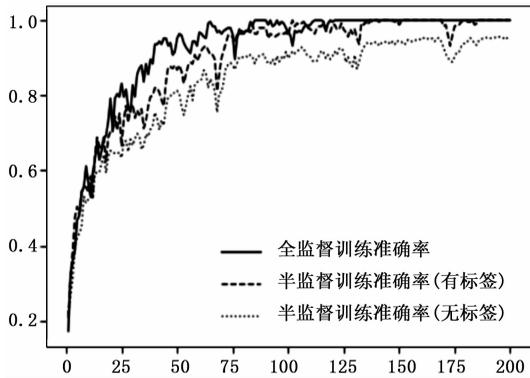


图 8 2 000 个有标签样本的训练过程中的准确率

图 9 和图 10 将两种条件下测试集的结果用混淆矩阵展示, 颜色表示识别程度, 颜色越深表示识别为此类的概率越大。混淆矩阵对角线上的数据表示被正确分类的准确率, 其余表示被错误分类, 由此可见, 改进的伪标签半监督方法在测试集上分别达到了 88.92% 和 94.14% 的识别准确率, 第二类辐射源的识别准确率最高, 相比之下, 第一类辐射源与第三类辐射源的差别较小, 在识别过程中有一定的识别难度。

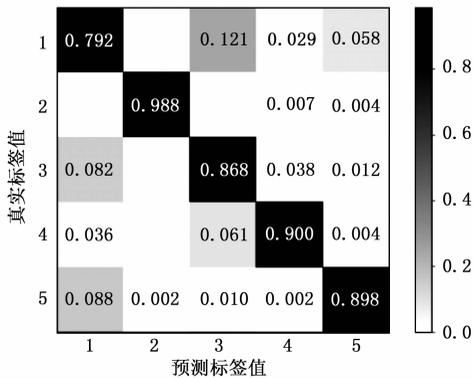


图 9 有标签样本数为 1 000 的混淆矩阵

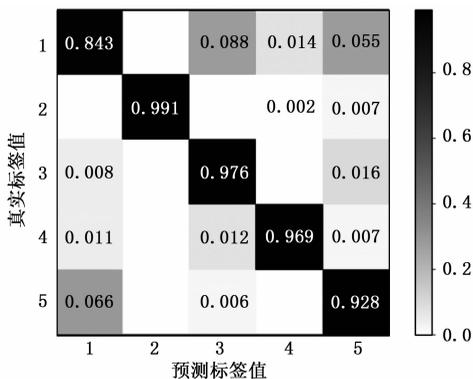


图 10 有标签样本数为 2 000 的混淆矩阵

设置有标签和无标签信号样本数比例, 并与两种有标签样本数构建不同的组合方式, 如表 3 和表 4 所示, 通过 100 次蒙特卡洛实验最终得出不同设置条件下全监督方法, 原伪标签半监督方法和改进后的伪标签半监督方法的识别

准确率, 其中带 * 的为改进后的伪标签半监督方法。

表 3 1 000 标签数时的识别准确率

训练方法	有标签与无标签的信号样本数比例			
	1:1	1:2	1:3	1:4
全监督/%	85.69			
伪标签半监督/%	86.19	87.02	87.82	85.70
伪标签半监督*/%	86.75	87.53	88.92	85.86

表 4 2 000 标签数时的识别准确率

训练方法	有标签与无标签的信号样本数比例			
	1:1	1:2	1:3	1:4
全监督/%	92.28			
伪标签半监督/%	92.51	92.97	93.83	92.44
伪标签半监督*/%	92.98	93.39	94.14	92.85

全监督训练过程中只用到有标签样本, 所以改变有标签训练样本所占其值不变。全监督方法受标签数目影响较大, 当标签数目较少时, 测试集的识别准确率较低, 加入无标签样本的半监督在标签数目较少的情况下, 取得了较高的识别准确率, 相比于原伪标签方法, 改进后的伪标签方法识别准确率得到提升。

一定范围内当无标签样本数所占比例增大时, 半监督学习的识别准确率会增大, 实验中发现有标签样本和无标签样本数目为 (1:3) 时效果最好, 但如果继续增大无标签数目, 模型的识别性能反而会下降, 且结果具有波动性, 这是因为过多的无标签样本会使生成错误伪标签的概率增大, 训练过程中就会受到错误的伪标签影响使得识别准确率降低。

当标签数目较多时, 全监督和半监督方法的识别准确率相差较小, 这也说明了, 只有在标签信号样本数目较少的情况下, 伪标签半监督学习在训练过程中才更能体现出避免过拟合增强鲁棒的特性。

4 结束语

本文提出了一种基于伪标签半监督深度学习的通信辐射源个体识别方法, 介绍了伪标签半监督深度学习方法的提出背景、理论基础和算法过程, 并且根据加权平均的思想加入了轮次标签法。对比分析改进的伪标签半监督方法与全监督方法和原有的伪标签半监督方法在 5 台 USRP 辐射源数据集上的训练过程和分类结果, 当有标签信号样本数为 1 000 和 2 000, 所占样本总数的四分之一时, 分别得到 88.92% 和 94.14% 的识别准确率, 证明了在少标签样本条件下的辐射源个体识别问题上有着很好的应用效果。在后续的工作中, 将对于伪标签质量的提高, 深度神经网络改进等方面继续深入研究探索。

参考文献:

[1] 刘 博, 辐射源个体识别技术的发展现状及应用建议 [J]. 电子信息对抗技术, 2019, 34 (4): 40-43.

