计算机科学
COMPUTER SCIENCE

基于协同网络与度量学习的标签噪声鲁棒联邦学习方法

吴飞, 张家宾, 岳晓凡, 季一木, 荆晓远

引用本文

吴飞, 张家宾, 岳晓凡, 季一木, 荆晓远. 基于协同网络与度量学习的标签噪声鲁棒联邦学习方法[J]. 计算机科学, 2024, 51(10): 391-398.

WU Fei, ZHANG Jiabin, YUE Xiaofan, JI Yimu, JING Xiaoyuan. Collaborative Network and Metric Learning Based Label Noise Robust Federated LearningMethod [J]. Computer Science, 2024, 51(10): 391-398.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900050

基于协同网络与度量学习的标签噪声鲁棒联邦学习方法

吴 飞[1] 张家宾[1] 岳晓凡[1] 季一木[2] 荆晓远[3]
1 南京邮电大学自动化学院、人工智能学院 南京 210003
2 南京邮电大学计算机学院 南京 210003
3 武汉大学计算机学院 武汉 430072

摘 要 针对联邦学习中标签噪声问题的研究较少,目前的主流方法是,服务器端引入基准数据集对客户端的模型进行评估,对客户端的聚合权重、特征类中心进行控制等,但大多数方法区分噪声客户端/噪声样本的能力尚有提升空间。文中提出了一种基于协同网络与度量学习的标签噪声鲁棒联邦学习方法。该方法由以下3部分组成:1)客户端互评分机制:客户端为彼此模型评分,构建评分矩阵,进一步将其转化为邻接矩阵,以区分干净/噪声客户端。2)协同网络模块:通过构建两个协同对等的联邦网络模型,使用简森-香农散度为协同网络彼此的训练区分干净样本与噪声样本。3)联邦-协同网络三元组损失:为噪声样本设计损失函数,约束同一噪声样本协同网络的输出特征。在CIFAR-10和CIFAR-100两个公开数据集上进行实验验证,结果表明所提方法在准确性上具有优势。

关键词:鲁棒联邦学习;标签噪声;协同网络;度量学习
中图分类号 TP391

Collaborative Network and Metric Learning Based Label Noise Robust Federated Learning Method

WU Fei[1],ZHANG Jiabin[1],YUE Xiaofan[1],JI Yimu[2] and JING Xiaoyuan[3]
1 College of Automation,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
2 School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
3 School of Computer Science,Wuhan University,Wuhan 430072,China

Abstract Currently,there is limited research on the problem of label noise in federated learning. The main approaches involve introducing a benchmark dataset on the server side to evaluate the client's model,controlling the aggregation weights and feature class centers of the clients. However,most methods still have room for improvement in distinguishing noisy clients or noisy samples. This paper proposes a label-noise robust federated learning method based on co-networks and metric learning. The method consists of the following three parts:1) Client mutual evaluation mechanism. Clients score each other's models,construct a rating matrix,and further transform it into an adjacency matrix to differentiate clean/noisy clients. 2) Collaborative network module. By constructing two collaborative equivalent federated network models,the Jensen-Shannon divergence is used to distinguish clean samples from noisy samples for the training of collaborative networks. 3) Federated-collaborative network triplet loss. A loss function is designed to constrain the output features of the collaborative networks for the same noisy samples. Experimental verification is conducted on the publicly available datasets CIFAR-10 and CIFAR-100,and the results demonstrate the superiority of the proposed method in accuracy.

Keywords Robust federated learning,Label noise,Collaborative network,Metric learning

到稿日期:2023-09-11 返修日期:2024-02-06
基金项目:国家自然科学基金(62076139);之江实验室开放课题(2021KF0AB05);未来网络科研基金项目(FNSRFP-2021-YB-15);南京邮电大学1311人才计划
This work was supported by the National Natural Science Foundation of China(62076139),Open Research Project of Zhejiang Lab(2021KF0AB05),Future Network Scientific Research Fund Project(FNSRFP-2021-YB-15) and 1311 Talent Program of Nanjing University of Posts and Telecommunications.
通信作者:吴飞(wufei_8888@126.com)

1 引言

联邦学习是一种大规模协作深度学习技术，客户端可以在保持本地数据隐私的条件下联合训练深度学习模型[1]。然而在现实的联邦学习应用中，由于不同客户端的能力、偏见、硬件可靠性等存在差异，不同客户端之间标注人员标注的角度不一致，其标注结果也不完全一致[2]。一些客户端可能会具有干净的、质量高的数据，但是另一些客户端可能会存在不同程度的标签噪声数据，导致标签质量突出问题[3]。在联邦学习系统中，标签质量差异化的存在会大大降低联邦学习模型的准确性和鲁棒性[4]。

现有应对标签噪声的联邦学习方法有引入基准数据集[4]、约束类特征中心[5]、评估客户端噪声水平的多阶段联邦学习[6]等方法。引入基准数据集将不可避免地带来潜在的数据偏见，约束类特征中心不适用于标签类数较多的场景，更准确地区分噪声与干净客户端需要进行进一步的研究。

针对以上问题，本文提出了基于协同网络与度量学习的标签噪声鲁棒联邦学习方法(Co-Network and Metric Learning Based Label-noise Robust Federated Learning，Fed-CNML)。本文的主要贡献为：

1)提出了客户端互评分机制(Mutual Evaluation Mechanism for Clients，MEMC)，用于区分干净/噪声客户端。本地客户端训练的模型由其他客户端进行评分，评分依据是测试精度。获得评价矩阵后，通过设定的阈值进一步获得邻接矩阵，进而依据客户端的邻居数量区分噪声客户端与干净客户端。

2)提出了联邦-协同机制(Federal-Collaborative Mechanism，FCM)，使用干净客户端获得两个对等的联邦-协同网络模型 A 和 B，联邦-协同网络模型(A/B)通过计算预测概率与原始标签的散度(Jensen-Shannon Divergence，简森-香农散度，JSD[7])为客户端上本地-协同网络模型(B/A)的训练区分干净样本与噪声样本。

3)设计了联邦-协同网络三元组损失(Federal-Collaborative Network Triplet Loss，FCNTL)。为了减少含有噪声客户端的模型参与联邦聚合对联邦模型产生的负面影响，本文引入度量学习，用于本地模型与全局模型、协同网络模型之间的对齐。一方面增大两个协同网络对同一噪声样本的特征距离，另一方面约束本地模型与联邦模型之间的特征距离。

4)在两个广泛使用的数据集 CIFAR-10 和 CIFAR100[8]上通过实验验证了 Fed-CNML 方法的有效性。

2 相关工作

近年来，在传统机器学习(非联邦学习)中常用的应对标签噪声的代表方法有：

1)优化损失函数。优化损失函数是解决标签噪声问题最常用的方法[9-12]，这些损失函数是专为标签噪声设计的，对异常值的敏感性较低，有助于减少噪声标签的影响。例如，Tanaka 等[9]提出了一种联合更新网络参数和标签的方法 Joint-Opt。他们发现高学习率可以抑制深度神经网络的记忆能力，并阻止它完全拟合标签。因此提出假设：干净的标签损失函数值应较小，噪声标签数据损失函数值应较大。根据这一假设，调整学习率以防止网络对错误伪标签过拟合。设计两个损失项，一个用于防止开始时所有样本被划分到同一个类中，另一个用于防止模型陷入局部最优解。

2)协同网络。协同网络是通过同时训练两个网络，使得它们相互合作，从而解决标签噪声问题[13-14]。Han 等[13]提出的协同指导(Co-teaching)同时训练两个深度神经网络，让它们互相合作。每个网络前馈所有样本的数据，并选择一些可能是干净标签的样本；两个网络相互通信，选择应该使用哪些数据进行训练；每个网络使用协同网络认为的干净样本训练。Li 等[14]提出了基于半监督的 DivideMix 方法，也同时训练两个网络，让每个网络在样本损失分布上拟合高斯混合模型(GMM 模型)；然后利用 GMM 模型，将训练数据分为干净数据和噪声数据。由于噪声数据原始标签不可信，因此其可以作为无标签数据使用。Li 等还设计了 Co-Divide 模块使标签与未标签数据分别在两个相同的模型上训练，让这两个网络彼此互斥，这样能够让网络过滤掉不同类型的错误以及防止确认性偏差。

3)样本分类。样本分类是常用的处理标签噪声的方法[15-18]。例如 Gu 等[15]将自步学习思想融入噪声标签学习中，通过设计的样本评估算法对数据集中得到的样本进行分组(分为简单容易判断标签干净的样本、复杂不容易判断标签的样本等分组)。优先用简单的干净的数据集开始训练分类器，然后逐渐增加复杂度更高的噪声数据到训练集中，不断提高训练集的多样性，以提升分类器的性能。Smart 等[16]通过自监督获得预分类器，筛选可信的干净样本，通过干净样本引导其他噪声样本进行半监督学习。Patel 等[17]提出了一种自适应样本选择策略，该策略仅依赖于给定小批量的批量统计选择样本进行训练。Xu 等[18]提出了自适应近邻聚类的标签噪声过滤算法 AdNN，该算法把标签噪声检测问题转化成离群点检测问题，然后根据相对密度去除离群点中的非噪声样本得到噪声备选集，最后通过噪声因子对噪声备选集中的离群点进行噪声识别和过滤。

4)利用图像和标签之间的关系。Iscen 等[19]提出邻居一致性学习，利用特征空间中样本之间的相似性，鼓励每个样本的预测与其邻域相似。

上述方法在建模的过程中均需要访问全部的样本数据，因此这些方法都不能直接应用于联邦学习场景中。现有联邦学习中标签噪声鲁棒的代表方法有：

1)引入基准数据。Chen 等[4]提出了一种引入基准数据集的方法，在服务器上维护一组基准样本，通过计算联邦模型在本地数据集上的性能与客户端本地模型在基准数据集中的性能之间的互交叉熵量化客户端本地数据的可信度。对可信度执行信用加权编排，以基于联邦模型中客户端可信度值调整客户端权重。但是，引入一个外部基准数据集同样具有挑战性，这将不可避免地带来潜在的数据偏见。如何选择合适的基准数据集、如何确定基准模型的适应程度等问题，需要进一步的研究和探索。

2)约束类特征中心。Yang 等[5]提出了一种通过类特征中心应对联邦学习噪声问题的方法 RoFL。这些特征中心是每个客户端上本地数据的类中心特征，由服务器在每轮联邦

学习中对齐得到全局类特征中心。这个全局类特征中心约束客户端的类特征中心，使客户端保持一致的决策边界。该方法存在一定局限，不适用于标签类较多的场景。

3）设计方法评估客户端噪声水平的多阶段联邦学习。Xu 等[6]提出了一种多阶段标签噪声鲁棒联邦学习方法 Fed-Corr，利用在所有客户端上独立测量的模型预测子空间的维数，识别有噪声的客户端；然后根据样本损失，识别有噪声客户端上的错误标签，提出了一种基于估计的局部噪声水平的自适应局部近端正则化项（LID 分数[20]）来估计客户端的噪声水平。根据 4.3 节的实验可知，LID 分数并不能有效地反映出客户端整体的噪声水平。

3　Fed-CNML 方法

图 1 给出了基于协同网络与度量学习的标签噪声鲁棒联邦学习方法（Fed-CNML）的总体框架。网络架构分为两个模块：一是评价矩阵计算模块，其使用客户端互评分机制区分干净与噪声客户端；二是协同网络-联邦学习模块，其使用协同网络为另一个网络的训练区分干净样本与噪声样本。

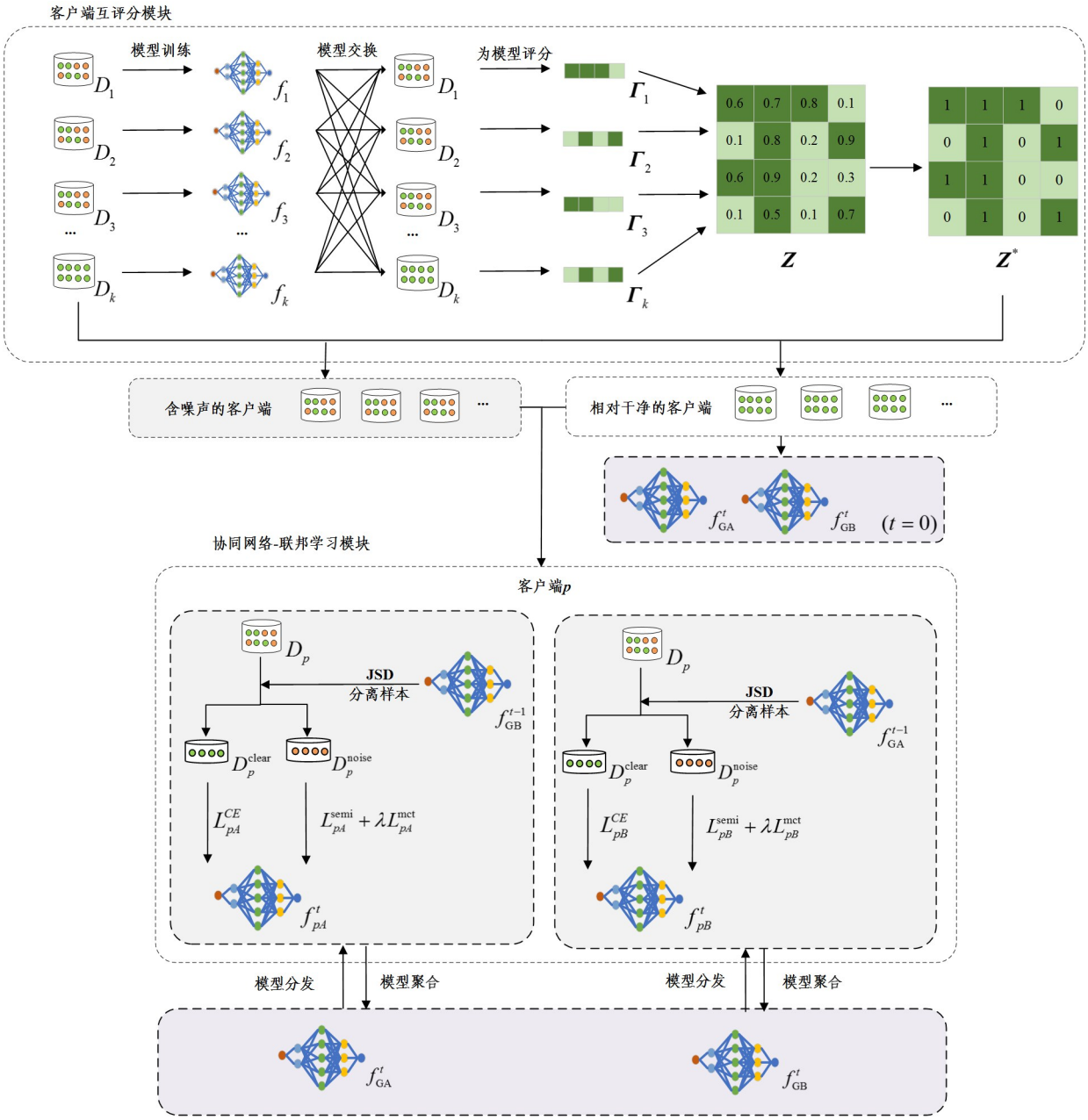


图 1　基于协同网络与度量学习的标签噪声鲁棒联邦学习方法总体框架

Fig. 1　Overall framework of Fed-CNML

首先，客户端利用自身数据为其他客户端本地训练的模型进行评分，构建评价矩阵，再进一步构建邻接矩阵以估计客户端噪声水平，区分干净的或者含有噪声的客户端。在联邦学习中，优先使用干净客户端训练获得两个协同模型，一个网络为另一个网络的训练筛选噪声客户端中的干净样本。为了利用噪声数据并降低噪声数据对联邦模型的影响，设计了联邦-协同网络三元组损失。

3.1　客户端互评分机制

在联邦学习中，如何准确判断客户端的噪声情况十分重要。在预训练阶段筛选出干净的客户端有利于防止噪声客户

端在前期参与训练，影响模型收敛。基于此，本文提出了一种不依赖额外的基准数据集的客户端互评分机制(MEMC)，以区分干净与噪声客户端。

在联邦学习场景中，客户端的原始数据保留在客户端当中，通过上传模型参数与梯度合作训练一个共享的全局模型[21]。Li 等提出在联邦学习中应用委员会机制来解决恶意客户端攻击的问题[21]，Che 等提出一种新的无服务器联邦学习框架，即基于委员会机制的联邦学习[22]。为了选取委员会客户端，这些方法将客户端模型参数或者梯度信息发送给其他客户端进行评价，所选取的委员会客户端可以获取其他所有客户端的模型参数、梯度等信息。客户端发送梯度信息给其他客户端，符合联邦学习的基本原则。

设在联邦学习中有 K 个客户端，第 p 个客户端拥有本地数据 $D_p=\{(x_p^i, y_p^i)\}(i=1,\cdots,N_p)$，其中每个客户端有 $N_p=|D_p|$ 个数据样本。数据样本 (x_p^i, y_p^i) 由图像 x_p^i 和原始标签 $y^i{}_p$ 组成，标签 $y_p^i=[y_{p1}^i, y_{p2}^i, \cdots, y_{pC}^i]^{\mathrm{T}}$。

在所有的 K 个客户端中构建独立的本地模型 f_p，$(p=1,\cdots,K)$，其参数为 θ_p。使用最小化交叉熵(Cross-Entropy, CE)损失来更新 f_p：

$$L_p^{\mathrm{CE}}=-\frac{1}{N_p}\sum_{i=1}^{N_p}(y_p^i)^{\mathrm{T}}\log\hat{y}_p^i \tag{1}$$

其中，$\hat{y}_p^i=\mathrm{softmax}(f_p(x_p^i))$ 是模型 f_p 对于 x_p^i 预测的概率分数。

在每个客户端都经过一定轮次的训练后，客户端将自己的本地模型的参数都发送给服务端，服务端将所有客户端模型参数汇总后，发送给各个客户端。客户端 p 得到其他客户端的模型参数 $\theta_q(q=1,\cdots,K)$ 后，使用样本 D_p 作为测试集在 $q(q=1,\cdots,K)$，客户端上获得测试精度 φ^{pq}，进一步可以得到一个维度为 K 的评价向量 $\boldsymbol{\Gamma}_p=\begin{pmatrix}\varphi^{p1}\\ \vdots\\ \varphi^{pK}\end{pmatrix}$。若要求模型参数梯度等信息不发送给其他客户端，则本文提出的互评分机制中也可采用黑盒评价策略。因为客户端不需要对模型参数进行汇总，所以各个客户端可以将自身的模型与参数封装为黑盒模型，其他客户使用本地数据对其进行黑盒测试。黑盒模型不直接泄漏模型参数信息与结构信息，有助于保护模型的具体细节。

然后将各个客户评分列向量 $\boldsymbol{\Gamma}_p(p=1,\cdots,K)$ 聚合，形成客户端评价矩阵：

$$\boldsymbol{Z}=(\boldsymbol{\Gamma}_1\ \cdots\ \boldsymbol{\Gamma}_K)=\begin{pmatrix}\varphi^{11} & \cdots & \varphi^{K1}\\ \vdots & \ddots & \vdots\\ \varphi^{1K} & \cdots & \varphi^{KK}\end{pmatrix} \tag{2}$$

通常，干净客户端的模型相对于噪声客户端训练时收敛速度快，容易取得良好的模型。在噪声客户端中，样本的标签存在多样性、随机性噪声，因此模型收敛速度慢，且模型因受到噪声影响而准确性较低，表现为模型只在自己客户端的训练集上测试结果良好，而在干净的数据集或者含有不同噪声的其他客户端的训练数据上准确性表现差。干净的客户端作为测试集在其他干净客户的模型上进行测试能取得相对较好的测试精度，在噪声客户端的模型上的测试精度较差。因此，评价矩阵可以反映出客户端数据与其他客户端模型之间相互“匹配”的程度。干净的客户端之间“匹配度”较高，噪声客户端与其他客户端的“匹配度”较低，噪声越严重的客户端之间的“匹配度”越低。通过可信度阈值 μ_1 对评价矩阵二值化，认为测试精度大于阈值的“匹配度”具有邻居关系，可以得到邻接矩阵 $\boldsymbol{Z}^*$：

$$\boldsymbol{Z}_{pq}^*=\begin{cases}1, & \varphi^{pq}\geqslant\mu_1\\ 0, & \varphi^{pq}<\mu_1\end{cases} \tag{3}$$

$$\mu_1=\max(\varphi^{pq})-(\max(\varphi^{pq})-\mathrm{avg}(\varphi^{pq}))/\tau_1$$

其中，$(p\neq q; p,q=1,\cdots,K)$，$\tau_1$ 是滤波系数，$\max(\varphi^{pq})$ 是 φ^{pq} 的最大值，$\mathrm{avg}(\varphi^{pq})$ 是 φ^{pq} 的平均值，该设置可以实现自适应调整阈值。

客户端 p 的邻居数量一定程度上可以反映客户端 p 与其他客户端的相似度，η_p 是邻接矩阵 $\boldsymbol{Z}^*$ 的行和，也即客户端邻居数量。

$$\eta_p=\sum_{j=1}^{K}\boldsymbol{Z}_{jp}^* \tag{4}$$

其中，μ_2 为设定的邻居数量阈值，和客户端总体数量相关。$\eta_p>\mu_2(p=1,\cdots,K)$ 的客户端被认为是干净客户端构成集合 S_{clear}，其他为噪声客户端 S_{noise}。

3.2 联邦-协同学习机制

本文方法使用预测概率与原始标签的散度(Jensen-Shannon Divergence, JSD)来区分干净样本与噪声样本。客户端使用联邦模型为每个样本计算一个 JSD，并将样本 JSD 发送给服务器，由服务器根据所有样本的 JSD 信息计算区分干净样本与噪声样本的阈值。但是单一网络为自己划分训练样本容易出现既当裁判又当选手的情形，样本选择错误会造成误差累计。因此本文设计了联邦-协同机制(FCM)，一个协同模型为另一个协同模型的训练选择样本。这两个模型分别独立地训练出两个不同的模型(这两个模型是分别随机初始化的)。

首先，第 p 个客户端构建独立的本地模型 f_{pA}^t 与 $f_{pB}^t(p=1,\cdots,K)$，其参数分别为 θ_{pA}^t 与 θ_{pB}^t，t 表示在第 t 轮联邦学习。本地训练数据 $D_p=\{(x_p^i, y_p^i)\}$。使用客户端样本数量作为联邦学习的聚合权重：

$$\begin{aligned}\theta_{\mathrm{GA}}^t&=\sum_{p\in S}\frac{N_p}{\sum_{i\in S}N_i}(\theta_{pA}^t)\\ \theta_{\mathrm{GB}}^t&=\sum_{p\in S}\frac{N_p}{\sum_{i\in S}N_i}(\theta_{pB}^t)\end{aligned} \tag{5}$$

其中，S 表示参与训练的客户端集合，θ_{GA}^t 和 θ_{GB}^t 表示第 t 轮中聚合后的联邦模型 f_{GA}^t 与 f_{GB}^t 的参数。客户端每轮聚合得到全局参数 θ_{GA}^t 和 θ_{GB}^t 后，使用全局参数更新本地参数 θ_{pA}^{t+1} 和 θ_{pB}^{t+1}。

在含有噪声的客户端参与训练之前，客户端互评分机制筛选出的干净客户端 S_c 使用交叉熵损失进行一定轮次的预热训练，经过预热训练后得到两个初始的联邦全局模型 f_{GA}^t 与 f_{GB}^t，$t=0$。

经过干净客户端的预热训练之后，噪声客户端也将加入训练，因此需要区分干净与噪声样本。干净客户端只是相对干净，用来预热网络模型。区分干净与噪声样本时，干净客户端同样参与区分。干净客户端的样本因为参与了预热训练，

所以在被划分时会更容易被划分为干净样本。干净客户端中的极少数噪声样本因为数量较少，所以对模型影响较小。

样本 x_p^i 的原始标签是 y_p^i，在两个协同网络上的预测概率分别是：

$$\begin{aligned}\hat{y}_{\mathrm{GA}}^i&=\mathrm{softmax}(f_{\mathrm{GA}}^t(x_p^i))\\ \hat{y}_{\mathrm{GB}}^i&=\mathrm{softmax}(f_{\mathrm{GB}}^t(x_p^i))\end{aligned}\tag{6}$$

设 J_{pA}^i 和 J_{pB}^i 是原始标签 y_p^i 和预测概率 $\hat{y}_{pA}^i$ 与 $\hat{y}_{pB}^i$ 之间的 JS 散度：

$$\begin{aligned}J_{pA}^i&=JSD(y_p^i,\hat{y}_{pB}^i)\\ J_{pB}^i&=JSD(y_p^i,\hat{y}_{pA}^i)\end{aligned}\tag{7}$$

其中，$JSD(\cdot)$ 是 Jensen-Shannon Divergence 函数。J_{pB}^i 由 f_{GA}^t 计算的 $\hat{y}_{pA}^i$ 得到，为 f_{GB}^t 的训练区分干净与噪声样本。同理，J_{pA}^i 由 f_{GB}^t 计算的 $\hat{y}_{pB}^i$ 得到，为 f_{GA}^t 的训练区分干净与噪声样本。

每个客户端上传样本的 JSD 到服务器，服务器分别对 J_{pA}^i 和 J_{pB}^i 进行排序，计算区分噪声样本与干净样本的阈值 J_{cutoff}^B 与 J_{cutoff}^A：

$$\begin{aligned}J_{\mathrm{cutoff}}^A&=\begin{cases}J_{\mathrm{avg}}^A-(J_{\mathrm{avg}}^A-J_{\mathrm{min}}^A)/\tau_2, & \text{if } J_{\mathrm{avg}}^A\geqslant J_u\\ J_{\mathrm{avg}}^A, & \text{other}\end{cases}\\ J_{\mathrm{cutoff}}^B&=\begin{cases}J_{\mathrm{avg}}^B-(J_{\mathrm{avg}}^B-J_{\mathrm{min}}^B)/\tau_2, & \text{if } J_{\mathrm{avg}}^B\geqslant J_u\\ J_{\mathrm{avg}}^B, & \text{other}\end{cases}\end{aligned}\tag{8}$$

其中，J_{avg}^A 和 J_{avg}^B、J_{min}^A 和 J_{min}^B 分别是 J_{pA}^i 和 J_{pB}^i（其中 $0\leqslant p<K$，$0\leqslant i<N_p$）的平均值和最小值；τ_2 是滤波系数；J_u 是样本选择的阈值，作用是当 J_{avg}^A 和 J_{avg}^B 大于阈值 J_u 时，适当越小阈值 J_{cutoff}^B 和 J_{cutoff}^A 以避免过于保守选择干净样本，从而实现自适应的阈值调整。

服务器将样本筛选阈值 J_{cutoff}^A 和 J_{cutoff}^B 发送至客户端。客户端 p 根据 J_{cutoff}^A 和 J_{cutoff}^B 分别为 f_{pA}^{t+1} 和 f_{pB}^{t+1} 的训练划分干净样本集合 D_{pA}^{clear} 和 D_{pB}^{clear} 以及噪声样本集合 D_{pA}^{noise} 和 D_{pB}^{noise}。

3.3　联邦-协同网络三元组损失设计

通过联邦协同机制，将所有客户端中的样本划分为干净样本集合 D_{pA}^{clear} 和 D_{pB}^{clear} 以及噪声样本集合 D_{pA}^{noise} 和 D_{pB}^{noise}。干净集合 D_{pA}^{clear} 和 D_{pB}^{clear} 的训练使用最小化交叉熵(CE)损失。

$$\begin{aligned}L_{pA}^{\mathrm{CE}}&=-\frac{1}{|D_{pA}^{\mathrm{clear}}|}\sum_{i\in D_{pA}^+}(y_p^i)^{\mathrm{T}}\log\hat{y}_{pA}^i\\ L_{pB}^{\mathrm{CE}}&=-\frac{1}{|D_{pB}^{\mathrm{clear}}|}\sum_{i\in D_{pB}^+}(y_p^i)^{\mathrm{T}}\log\hat{y}_{pB}^i\end{aligned}\tag{9}$$

其中，$\hat{y}_{pA}^i=\mathrm{softmax}(f_{pA}^t(x_p^i))$ 是对应于 x_p^i 在模型 f_{pA}^t 上预测的概率分数，$\hat{y}_{pB}^i=\mathrm{softmax}(f_{pB}^t(x_p^i))$ 是对应于 x_p^i 在模型 f_{pB}^t 上预测的概率分数。

对于噪声样本，通常采用伪标签的方式重新赋予标签进行训练，但是伪标签若不正确，将会引入新的误差。考虑到联邦学习中的全局模型是使用多方数据训练出来的，一般来说，全局模型的准确性明显优于单个本地模型的准确性，因此本方法通过约束本地模型学习到的特征表示与全局模型学习到的特征表示之间的距离，限制学习错误标签带偏模型的影响。

两个协同模型使用了同样的伪标签计算损失，如果伪标签错误，误差将累计，样本上错误的伪标签将不容易再矫正。若噪声样本在两个协同模型上输出特征出现差异，则可以避免自我训练中的确认偏差，过滤不同类型的错误，进一步降低两个模型学习到错误伪标签噪声样本带来的累积误差风险。因此本文设计联邦-协同网络三元组损失(FCNTL)，增大两个协同网络模型学习到的特征表示之间的距离，使协同网络尽可能学习到噪声样本不同的特征。

因此，对于噪声集合 D_{pA}^{noise} 和 D_{pB}^{noise} 中的样本，使用两个协同网络的综合预测经过锐化处理[23]后的 α_p^i 作为伪标签。计算交叉熵损失为 L_{pA}^{semi} 和 L_{pB}^{semi}。

$$\begin{aligned}L_{pA}^{\mathrm{semi}}&=-\frac{1}{|D_{pA}^{\mathrm{noise}}|}\sum_{i\in D_{p_A}^{\mathrm{noise}}}(\alpha_p^i)^{\mathrm{T}}\log\hat{y}_{pA}^i\\ L_{pB}^{\mathrm{semi}}&=-\frac{1}{|D_{pB}^{\mathrm{noise}}|}\sum_{i\in D_{pB}^{\mathrm{noise}}}(\alpha_p^i)^{\mathrm{T}}\log\hat{y}_{pB}^i\end{aligned}\tag{10}$$

为了降低 D_{pA}^{noise} 和 D_{pB}^{noise} 中的样本因伪标签不正确而对全局模型产生的影响，本文引入度量损失限制样本在本地模型上特征与全局模型上的特征的距离。以模型 f_{pA}^t 为例，对于一个样本 x_p^i，它在本地模型 f_{pA}^t 和全局模型 f_{GA}^t 上的特征距离被定义为：

$$\begin{aligned}&u(Y_{pA}^t,Y_{\mathrm{GA}}^{t-1})=\|Y_{pA}^t-Y_{\mathrm{GA}}^{t-1}\|_2^2\\ &Y_{pA}^t=f_{pA}^t(x_p^i)\\ &Y_{\mathrm{GA}}^{t-1}=f_{\mathrm{GA}}^{t-1}(x_p^i)\end{aligned}\tag{11}$$

其中，Y_{pA}^t 为样本 x_p^i 经过本地模型 f_{pA}^t 的非线性特征表示，Y_{GA}^{t-1} 为样本 x_p^i 在上一轮的全局模型 f_{GA}^{t-1} 的非线性特征表示。

为了增大噪声样本在两个协同网络上的特征距离，本文增大两个网络之间噪声样本特征的差异。在本地模型 f_{pA}^t 和 $t-1$ 轮全局模型 f_{pB}^{t-1} 的特征距离被定义为：

$$\begin{aligned}&u(Y_{pA}^t,Y_{\mathrm{GB}}^{t-1})=\|Y_{pA}^t-Y_{\mathrm{GB}}^{t-1}\|_2^2\\ &Y_{\mathrm{GB}}^{t-1}=f_{\mathrm{GB}}^{t-1}(x_p^i)\end{aligned}\tag{12}$$

其中，Y_{GB}^{t-1} 为样本 x_p^i 在上一轮全局模型 f_{pB}^{t-1} 上的非线性特征表示。

将度量学习应用于本地模型 f_{pA}^t 与全局模型 f_{GA}^{t-1} 和 f_{GB}^{t-1} 的对齐，定义联邦-协同网络三元组损失(FCNTL)，函数定义如下：

$$L_{pA}^{\mathrm{mct}}=-\frac{1}{N_p}v(u(Y_{pA}^t,Y_{\mathrm{GA}}^{t-1})-u(Y_{pA}^t,Y_{\mathrm{GB}}^{t-1}))\tag{13}$$

其中，$v(x)=\max(0,x)$ 是一个铰链损失函数。对距离 $u(Y_{pA}^t,Y_{\mathrm{GA}}^{t-1})$ 和 $u(Y_{pA}^t,Y_{\mathrm{GB}}^{t-1})$ 施加了约束，期望样本在本地模型 f_{pA}^t 与全局模型 f_{GA}^t 之间的特征距离小，从而限制学习噪声样本时本地模型更新的漂移幅度；期望本地模型 f_{pA}^t 到全局模型 f_{GB}^t 之间的特征距离大，从而增大全局模型 f_{GA}^t 与全局模型 f_{GB}^t 在噪声样本的特征差异。

交叉熵损失 L_{pA}^{semi} 和 L_{pB}^{semi} 与联邦-协同网络三元组损失 L_{pA}^{mct} 和 L_{pB}^{mct} 相结合，得到噪声数据 D_{pA}^{noise} 与 D_{pB}^{noise} 的总损失函数为：

$$\begin{aligned}L_{pA}^{\mathrm{tol}-}&=L_{pA}^{\mathrm{semi}}+\lambda L_{pA}^{\mathrm{mct}}\\ L_{pB}^{\mathrm{tol}-}&=L_{pB}^{\mathrm{semi}}+\lambda L_{pB}^{\mathrm{mct}}\end{aligned}\tag{14}$$

其中，λ 为模型对比三元组损失系数。

综上所述，干净样本与噪声样本分别采用不同的损失计算方式，总损失为：

$$L_{pA}^{\mathrm{tol}}=L_{pA}^{\mathrm{CE}}+L_{pA}^{\mathrm{tol}-}$$
$$L_{pB}^{\mathrm{tol}}=L_{pB}^{\mathrm{CE}}+L_{pB}^{\mathrm{tol}-} \tag{15}$$

联邦学习结束后的最终模型是 f_{GA}^{t} 与 f_{GB}^{t}。在使用模型对测试集图片分类时，采用 A 模型与 B 模型结合的方式，A 模型与 B 模型各自的预测概率如式(6)所示，测试样本 x^i 的最终预测概率 $\hat{y}^i$ 为 $\hat{y}_{\mathrm{GA}}^{i}$ 和 $\hat{y}_{\mathrm{GB}}^{i}$ 的加权和：

$$y^i=\frac{\hat{y}_{\mathrm{GA}}^{i}+\hat{y}_{\mathrm{GB}}^{i}}{2} \tag{16}$$

4 实验及结果分析

4.1 数据集介绍

本文使用了两个广泛使用的公开数据集 CIFAR-10 和 CIFAR-100 进行了实验。CIFAR-10 和 CIFAR-100 是经典的图像分类数据集，分别包含 10 个、100 个类别的 60 000 张 32 * 32 像素的 RGB 彩色图像。其中 50 000 张用于训练，10 000 张用于测试。本文在多个噪声水平下进行了实验，以证明 Fed-CNML 对本地标签质量偏差具有鲁棒性。数据集的详细信息如表 1 所列。

表 1 CIFAR-10 和 CIFAR-100 数据集的统计数据

Table 1 Statistics of CIFAR-10 and CIFAR-100 datasets

数据集	CIFAR-10	CIFAR-100
训练集大小	50 000	50 000
测试集大小	10 000	10 000
图像尺寸	32 * 32 * 3	32 * 32 * 3
标签类数	10	100
客户端数	100	50

4.2 实验设置

使用 ResNet18[24] 作为基础网络模型，使用随机梯度下降(SGD)优化器执行优化，设置如下：初始学习率(LR)为 0.02，权重衰减为 5×10^{-4}，动量值为 0.9，批量大小为 10。每个网络训练 350 个 epoch，同时每 120 个 epoch 线性衰减学习率(lr-decay)0.1。CIFAR-10 和 CIFAR-100 分别采用了 25、50 epoch 的预热期以挑选干净客户端。本实验使用 CIFAR10-Policy 进行数据扩充。

本实验使用与 Fed-Corr 方法[6] 一致的噪声添加策略。含噪客户端比例 ρ 是含噪客户端占所有客户端的比例，客户端噪声率 κ_p 是 p 客户端训练集中噪声样本占该客户端所有样本的比例。噪声客户端中，噪声样本占该客户端样本的比例 κ_p 是 0～1 之间的随机数，添加噪声类型是随机标签噪声。干净客户端中标签均为正确的，$\kappa_p=0$。

4.3 客户端互评分机制有效性评估

本节对比了 Fed-Corr 方法[6] 中基于 LID 分数判断客户端噪声情况的方法。通过图示的形式展示了本文所提出的客户端互评分机制在不同噪声比例下筛选干净客户端的表现。

以 CIFAR-10 数据集为例进行展示。图 2 左图是在噪声率设置为 $\rho=0.4$ 时，本文提出的客户端选择机制(MEMC)客户端邻居数量与噪声水平关系图，图 2 右图是 Fed-Corr 中 LID 分数与噪声水平的关系图。其中每一个点代表一个客户端，纵坐标表示客户端的噪声率，横坐标分别表示客户端邻居的数量以及客户端的 LID 分数。

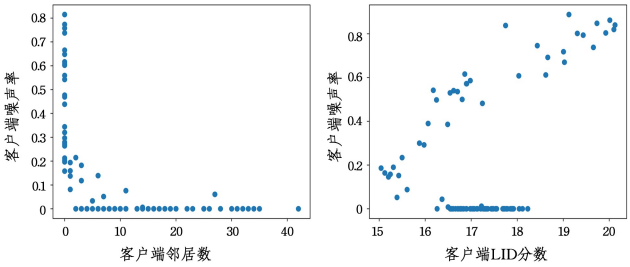
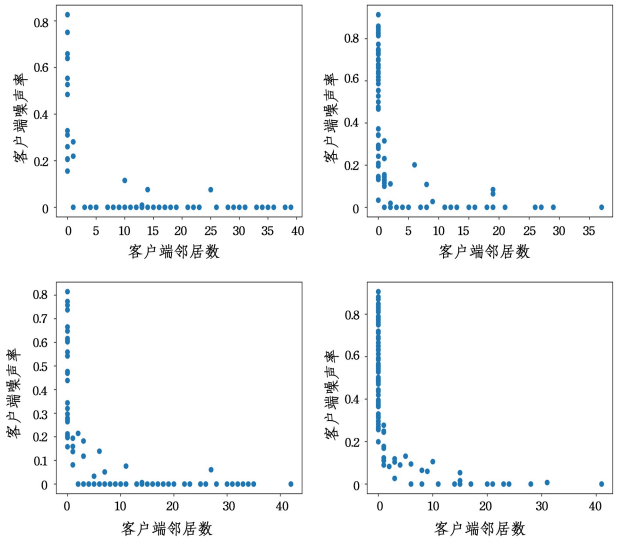


图 2 $\rho=0.4$ 时，客户端邻居数、客户端 LID 分数与客户端噪声率的对比关系

Fig. 2 When $\rho=0.4$, number of client neighbors or client LID score vs. noise rate

如图 2 左图所示，噪声客户端与干净客户端在邻居数量上具有明显差异。如图 2 右图所示，不同噪声率的客户端的 LID 分数分布分散，依据客户端 LID 分数难以有效区分干净客户端与噪声客户端。本文提出的客户端选择机制(MEMC)根据客户端邻居数量能够更好地区分噪声客户端与干净客户端。

为了调研客户端互评分机制在不同比例噪声下的性能，本实验对多组不同比例($\rho=0.2, 0.4, 0.6, 0.8$(*100%))的噪声设置进行了对比实验。客户端邻居数量与噪声比例的关系如图 3 所示。图中每一个点代表一个客户端，纵坐标表示客户端的噪声率，横坐标表示客户端邻居的数量。



注：从左上至右下含噪客户端比例分别是 $\rho=0.2, 0.4, 0.6, 0.8$。

图 3 不同含噪客户端比例下客户端邻居数与噪声率关系图

Fig. 3 Relationship between the number of client neighbors and noise rate with different noisy-client proportions

客户端邻居数与相对干净客户端的绝对数量有关。因此，本文对过滤客户端时邻居数量阈值 μ_2 进行了实验。在 100 个客户端的 CIFAR-10 数据集的实验中，在不同噪声客户端比例下，设置不同阈值 μ_2 来研究对实验产生的影响。实验结果如表 2 所列，在阈值 $\mu_2=1$ 和 5 的实验结果高于 0，10，15 等情况。阈值过大或者过小均会对实验结果产生影响，阈值设置为 0 表示所有客户端均参与初始模型的训练。

表 2　CIFAR-10 数据集、100 个客户端情况下，不同 μ_2 对结果的影响

Table 2　Influence of different values of μ_2 on the results of MEMC on CIFAR-10 dataset with 100 clients

噪声客户端比例	$\mu_2=0$	$\mu_2=1$	$\mu_2=5$	$\mu_2=10$	$\mu_2=15$
$\rho=0.8$	93.02	**93.53**	93.51	92.45	89.11
$\rho=0.6$	92.80	94.01	**94.30**	92.24	91.90
$\rho=0.4$	92.44	93.30	**93.49**	92.99	92.85
$\rho=0.2$	92.00	94.43	**94.49**	94.14	92.65

阈值过小，噪声客户端没有被筛选出来，大量噪声样本参与初始模型的训练。因为初始模型受到影响，所以接下来使用 JSD 进行样本区分，噪声样本更容易被分进干净样本内。

阈值过大，客户端筛选越严格，训练初始模型所用的样本数量就越少，初始模型的泛化能力随着客户端数量的减少而降低。因此，接下来使用 JSD 进行样本区分，干净样本则不容易被区分出来。

阈值过大或者过小最终都会对实验结果产生一定影响。经过实验，阈值设置为所有客户端数量的 1%～5% 比较合适。在 CIFAR10 以及 CIFAR100 数据集上进行的实验中，客户端数分别是 100 和 50，阈值 μ_2 分别取 5 和 2。

4.4　对比方法与实验结果分析

为了评估本文方法 Fed-CNML 的性能，将其与以下两类最先进的方法进行比较：1）针对传统数据集中场景（非联邦学习场景）的标签噪声方法（JointOpt[7] 和 DivideMix[12]），作为参考上限；2）针对联邦学习场景的标签噪声方法（RoFL[5] 和 FedCorr[6]）。作为参考，本文统一采取与它们实验相同的实验设置，对比方法的结果使用 FedCorr 论文中报告的结果或使用其公开的代码在配置 A6000 显卡的服务器上运算得出。

Fed-CNML 方法在 $\rho=0.0, 0.2, 0.4, 0.6, 0.8, 1.0$（$*100\%$）等不同噪声水平设置下，在 CIFAR-10 数据集上的测试精度平均值（5 次实验）和标准差如表 3 所列，最优结果用粗体表示。在集中学习场景中，使用与联邦学习设置中完全相同的噪声处理数据集。此外，集中学习场景中的精度也可以视为联邦学习中精度的上限。在绝大多数噪声设置中，本文所提出的 Fed-CNML 方法在精度上取得了显著提升。

表 3　CIFAR-10 数据集上测试准确度的平均值（5 次试验）和标准差

Table 3　Average (5 trials) and standard deviation of test accuracies on CIFAR-10

设置	对比方法	精度（%）±标准差					
		$\rho=0.0$	$\rho=0.2$	$\rho=0.4$	$\rho=0.6$	$\rho=0.8$	$\rho=1.0$
集中场景	JointOpt	93.73±0.21	92.70±0.10	92.29±0.37	91.26±0.46	89.18±0.29	84.65±0.05
	DivideMix	95.64±0.05	95.12±0.11	96.39±0.09	96.07±0.06	94.21±0.27	94.50±0.10
联邦场景	RoFL	88.33±0.07	88.29±0.20	88.25±0.33	87.77±0.83	87.08±0.65	86.40±0.47
	FedCorr	93.82±0.41	94.29±0.32	**94.01±0.22**	92.93±0.25	91.52±0.50	89.67±0.13
	Fed-CNML	**95.24±0.04**	**94.49±0.24**	93.85±0.36	**94.30±0.41**	**93.51±0.20**	**93.96±0.33**

在取不同 ρ 的 6 组实验设置下，本方法与集中场景的标签噪声鲁棒学习方法 DivideMix 之间的平均精度差距为 1.31%，平均精度超过目前标签噪声联邦学习中最好的 FedCorr 方法 1.82%。

由于 RoFL 和 JointOpt 并没有报告在 CIFAR-100 数据集上的精度，因此本方法在 CIFAR-100 数据集上与 DivideMix 和 FedCorr 进行了对比，结果如表 4 所列，最优结果用粗体表示。在所有实验设置中，本文提出的 Fed-CNML 方法在精度上取得了显著提升，本文方法与集中场景的标签噪声鲁棒学习方法 DivideMix 之间的平均精度差距为 4.87%，平均精度超过目前标签噪声联邦学习中最好的 FedCorr 方法 5.99%。

表 4　CIFAR-100 数据集上测试准确度的平均值（5 次试验）和标准差

Table 4　Average (5 trials) and standard deviation of test accuracies on CIFAR-100

设置	对比方法	精度（%）±标准差					
		$\rho=0.0$	$\rho=0.2$	$\rho=0.4$	$\rho=0.6$	$\rho=0.8$	$\rho=1.0$
集中场景	DivideMix	76.21±0.22	76.51±0.25	77.30±0.32	76.64±0.60	75.83±0.19	74.60±0.38
联邦场景	FedCorr	72.56±2.07	72.10±1.40	71.05±1.24	70.56±2.21	58.60±4.54	57.89±6.10
	Fed-CNML	**76.05±0.08**	**74.76±0.21**	**73.66±0.26**	**73.15±0.34**	**70.02±0.20**	**65.07±0.48**

4.5　消融实验

本节评估了 Fed-CNML 中组件的有效性，将没有使用客户端互评分机制的版本称为 Fed-C1，将没有使用联邦-协同机制的版本称为 Fed-C2，将没有使用联邦-协同网络三元组损失的版本称为 Fed-C3。将这 3 种版本与完整的 Fed-CNML 方法在 CIFAR10 数据集上进行了实验。对比实验结果如表 5 所列。

由表 5 的实验结果可以看出，Fed-CNML-1，Fed-CNML-2 以及 Fed-CNML-3 在 4 种实验设置中的表现都明显不如完整版本的 Fed-CNML。其中联邦-协同机制对精度影响最大；其次是联邦-协同网络三元组损失，对精度提升有一定作用。本文设计的 Fed-CNML 通过对客户端进行筛选、协同网络联邦学习方法、联邦-协同度量学习损失，提升了标签噪声的学习效果，对解决联邦学习中的标签噪声问题效果显著。

表 5　CIFAR-10 的消融实验结果

Table 5　Ablation experiment results on CIFAR-10

消融方法	$\rho=0.2$	$\rho=0.4$	$\rho=0.6$	$\rho=0.8$	$\rho=1.0$
Fed-C1	92.00	92.44	92.80	93.02	92.18
Fed-C2	87.97	86.88	82.41	77.44	78.15
Fed-C3	92.31	90.26	89.45	89.50	89.81
Fed-CNML	**94.49**	**93.49**	**94.30**	**93.51**	**93.96**

结束语　本文提出了一种基于协同网络和度量学习的标签噪声鲁棒联邦学习方法。该方法由 3 个部分组成。1）客户

端相互评价机制:客户端对彼此的模型进行评分,构造评分矩阵,并进一步将其转化为邻接矩阵,以区分干净/噪声的客户端。2)协同网络模块:通过构建两个协作等效联邦网络模型,利用JSD进行协作网络训练,以区分干净样本和噪声样本。3)联邦协作网络三元组损失:针对噪声样本设计损失函数,以约束协同网络对于同一噪声样本的输出特征。本研究在两个广泛应用的数据集上进行了综合实验,证明了本文方法在准确性上具有优势。

参 考 文 献

[1] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[2] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Artificial Intelligence and Statistics. 2017: 1273-1282.

[3] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[4] CHEN Y, YANG X, QIN X, et al. Dealing with label quality disparity in federated learning[M]//Federated Learning: Privacy and Incentive. 2020: 108-121.

[5] YANG S, PARK H, BYUN J, et al. Robust federated learning with noisy labels[J]. IEEE Intelligent Systems, 2022, 37(2): 35-43.

[6] XU J, CHEN Z, QUEK T Q S, et al. Fedcorr: Multi-stage federated learning for label noise correction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10184-10193.

[7] FUGLEDE B, TOPSOE F. Jensen-Shannon divergence and Hilbert space embedding[C]// International Symposium on Information Theory. 2004: 31-36.

[8] KRIZHEVSKY A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009.

[9] TANAKA D, IKAMI D, YAMASAKI T, et al. Joint optimization framework for learning with noisy labels[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5552-5560.

[10] WANG Y, MA X, CHEN Z, et al. Symmetric cross entropy for robust learning with noisy labels[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 322-330.

[11] THULASIDASAN S, BHATTACHARYA T, BILMES J, et al. Combating label noise in deep learning using abstention[C]. International Conference on Machine Learning. 2019: 6234-6243.

[12] GHOSH A, KUMAR H, SASTRY P S. Robust loss functions under label noise for deep neural networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017: 1919-1925.

[13] HAN B, YAO Q, YU X, et al. Robust training of deep neural networks with extremely noisy labels[J] arXiv: 1804. 06872, 2020.

[14] LI J, SOCHER R, HOI S C H. Dividemix: Learning with noisy labels as semi-supervised learning[C]// International Conference on Learning Representations. 2020.

[15] GU N, FAN M, MENG D. Robust semi-supervised classification for noisy labels based on self-paced learning[J]. IEEE Signal Processing Letters, 2016, 23(12): 1806-1810.

[16] SMART B, CARNEIRO G. Bootstrapping the Relationship Between Images and Their Clean and Noisy Labels[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 5344-5354.

[17] PATEL D, SASTRY P S. Adaptive sample selection for robust learning under label noise[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 3932-3942.

[18] XU M L, JIANG G X, WANG W J. Label noise filtering framework based on outlier detection[J]. Computer Science, 2024, 51(2): 87-99.

[19] ISCEN A, VALMADRE J, ARNAB A, et al. Learning with neighbor consistency for noisy labels[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4672-4681.

[20] HOULE M E. Dimensionality, discriminability, density and distance distributions[C]// International Conference on Data Mining Workshops. 2013: 468-473.

[21] LI Y, CHEN C, LIU N, et al. A blockchain-based decentralized federated learning framework with committee consensus[J]. IEEE Network, 2020, 35(1): 234-241.

[22] CHE C, LI X, CHEN C, et al. A decentralized federated learning framework via committee mechanism with convergence guarantee[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(12): 4783-4800.

[23] YOU Y, CHEN T, SUI Y, et al. Graph contrastive learning with augmentations[C]// Advances in Neural Information Processing Systems. 2020: 5812-5823.

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

WU Fei, born in 1989, Ph.D, professor. His main research interests include pattern recognition and machine learning.

(责任编辑:何杨)