基于联邦学习的主动半监督短文本分类方法

孔德焱[1],冀振燕[2*],杨燕燕[1],刘 洋[1],刘吉强[2]
(1. 北京交通大学软件学院,北京 100044;
2. 北京交通大学网络空间安全学院智能交通数据安全与隐私保护北京市重点实验室,北京 100044)

摘 要: 短文本分类应用广泛,是当前的研究热点,但受到短文本标注数据稀缺和数据隐私保护不便集中训练的影响,分类效果不佳. 针对上述问题,我们提出了基于联邦学习的主动半监督异质图注意力网络模型(Active Semi-Supervised Learning empowered Heterogeneous Graph ATtention network model based on Federated learning, Fed-ASSL-HGAT),通过设计新颖的主动半监督学习(Active Semi-Supervised Learning, ASSL)框架生成高质量标注样本赋能异质图注意力网络(Heterogeneous Graph ATttention network model, HGAT),引入联邦学习对部署在不同节点的模型进行联合训练以满足数据隐私保护需求. 所提出的ASSL框架通过将主动学习的多类别标注转化成二元类别标注,可大大降低标注难度;设计基于信息增益的选择策略筛选软、硬标签,以防止信息损失;通过半监督学习选择高准确率、高稳定性的正负样本打伪标签以确保标注质量. 实验结果表明,所提出的ASSL-HGAT(S)在AGNews、Snippets、TagMyNews数据集上相比HGAT基线模型 F_1 值分别提升2.45%、8.11%、7.46%. 融合联邦学习所进一步提出的Fed-ASSL-HGAT模型可在不泄漏隐私数据的情况下满足性能要求.

关键词: 异质图神经网络;主动学习;半监督学习;联邦学习
基金项目: 国家自然科学基金(No.52175493, No.51935002)
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3517-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20230703

An Active Semi-Supervised Short Text Classification Method Based on Federated Learning

KONG De-yan[1], JI Zhen-yan[2*], YANG Yan-yan[1], LIU Yang[1], LIU Ji-qiang[2]
(1. *School of Software Engineering, Beijing Jiaotong University, Beijing* 100044, *China*;
2. *Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, School of Cyberspace Science and Techonology, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: Short-text classification is broadly used and is a current hot research spot. However, the performance of short-text classification is hampered by the sca1rcity of annotated data for short texts and the challenges of centralized training for private data. To address these issues, we propose Fed-ASSL-HGAT (Active Semi-Supervised Heterogeneous Graph ATtention network model based on Federated learning), an active semi-supervised heterogeneous graph attention network model based on federated learning. This model utilizes the innovative active semi-supervised learning (ASSL) framework to generate high-quality labeled samples for empowering the heterogeneous graph attention network (HGAT) model. Additionally, federated learning is introduced to facilitate the joint training of the models deployed on different nodes, thereby satisfying the requirements of data privacy protection. The proposed ASSL framework significantly reduces the annotation difficulty by transforming the multi-class annotation task into a binary classification task. To mitigate information loss, we employ a selection strategy based on information gain to filter soft and hard labels. Semi-supervised learning is employed to select positive and negative samples with high accuracy and stability for pseudo-labeling, thereby ensuring the labeling quality. Experimental results demonstrate that the proposed ASSL-HGAT (Active Semi-supervised Learning Empowered Heterogeneous Graph Attention Network) model achieves improvements of 2.45%, 8.11%, and 7.46% in F_1 scores comparing with the HGAT baseline model on the AGNews, Snippets, and TagMyNews datasets, respectively. By incorporating the federat-

收稿日期:2023-07-26;修回日期:2024-03-04;责任编辑:王天慧
*通讯作者:冀振燕

ed learning, the Fed-ASSL-HGAT model can meet the performance requirements without scarifying data privacy.

Key words: heterogeneous graph neural network; active learning; semi-supervised learning; federated learning
Foundation Item(s): National Natural Science Foundation of China (No.52175493, No.51935002)

1 引言

短文本分类算法已广泛应用于各种场景,如新闻分类[1]、情感分类、标签推荐、软件需求分类[2]等任务中,模型在有大规模、集中式训练数据可用时可达到领先性能. 如,Yang等[3]提出的异质图注意力网络(Heterogeneous Graph ATtention network model, HGAT)在短文本基准数据上取得了先进性能. 然而,一些实际场景中数据难以共享,分类类别较多,标注数据稀缺,而通过人工标注样本又十分耗时. 因此,针对上述问题,研究人员开始探索如何结合主动学习和半监督学习、联邦学习等方法,以实现在不泄漏隐私数据的情况下满足性能要求.

主动学习(Active Learning, AL)是一种高效的打标签方法,通过选择和标记少量含有丰富信息的未标记样本来构建高性能分类器. 半监督学习(Semi-Supervised Learning, SSL)可充分利用无标签数据来学习更多的特征,对经过有标签数据训练的模型进行增强. 联邦学习方法[4]旨在使多个用户的客户端能够联合训练模型,并保持与其他用户或中心服务器的数据隔离.

基于上述启发,本文提出了基于联邦学习的主动半监督异质图注意力网络模型(Active Semi-Supervised Learning empowered Heterogeneous Graph ATtention network model based on Federated learning, Fed-ASSL-HGAT),将主动学习和半监督学习有机结合,有效解决标签稀缺对HGAT模型性能的挑战;并通过将主动学习问题转化为二元判断问题,简化了多类别分类任务的标注难度,减轻了专家标注的负担;通过引入基于信息增益的主动学习策略增强了标签监督;通过引入联邦学习联合训练不同节点的模型解决数据隐私问题.

2 相关工作

2.1 基于伪标签的半监督学习方法

基于伪标签的半监督学习方法旨在利用已训练模型选择具有高置信度预测的无标签样本,并将生成的伪标签样本添加到训练集中,以提高模型的分类性能. 未标记样本通常会提供额外的信息,帮助模型更好地捕捉数据分布,从而增强其泛化能力. Iscen等[5]提出了一种基于流形假设的可转换标签传播方法,通过最近邻图生成伪标签. Haase-Schütz等[6]对未标记的数据集进行分区,并在每个分区上重新训练初始化的神经网络,之后使用先前训练过的网络来过滤用于训练新网络的标签. Berthelot等[7]提出了一种基于标签平滑技术和数据增强的半监督学习方法MixMatch,有效地利用未标记数据信息,提高了模型性能和泛化能力. Zhou等[8]利用自监督学习的方式来增强半监督学习的能力,使其在不需要额外标签的情况下,可以更好地利用未标记的数据.

基于伪标签的传统半监督学习技术是跨领域的,但表现不佳,因为校准不良的模型会产生通用、大量错误的伪标签,导致噪声训练. 为此,Rizve等[9]提出了一种具有竞争力的降噪策略,通过不确定性机制和负样本学习来减少噪声训练. 因此,本文使用基于不确定性感知的伪标签选择框架来对异质图注意力网络进行赋能.

2.2 主动学习

主动学习是一种被广泛研究的解决标签瓶颈的方法,其最核心的部分是设计查询策略. 其查询策略分为基于信息量、代表性以及两者混合的查询策略[10]. Zhang等[11]提出了一种新型主动半监督学习方法对医学图像进行分类,该方法通过对抗扰动和密度感知熵来寻找最具有代表性和信息量样本进行标注. Zhang等[12]提出了一种基于图神经网络的主动学习方法,采用软标签选择策略来减少对专家的依赖. 本文使用基于信息增益的主动学习方法来降低标注成本.

2.3 联邦学习

联邦学习是一种分布式的机器学习框架,允许中央协调服务器对分布在多个客户端上的本地非独立同分布训练数据进行分布式训练来学习共享模型,同时确保客户端对数据的控制权,有助于打破数据孤岛的限制. 联邦学习自2016年谷歌提出后便引起学术界和工业界的强烈关注. Konečný等[13]首次提出了FedAvg方法来聚合客户端模型,在此基础上Hsu等[14]通过FedAvgM方法引入动量来积累模型更新. 联邦学习还从更实际的角度进行了研究,如Jeong等[15]关注了联邦学习中数据标签的不足,并提出了半监督解决方案. Ahn等[16]讨论了联邦主动学习的范式. 基于上述启发,本文将主动半监督学习赋能异构图注意力网络(Active Semi-Supervised Learning empowered Heterogeneous Graph ATtention network, ASSL-HGAT)和联邦学习相融合,同时解决了标签样本不足和数据安全的问题.

3 基于联邦学习的主动半监督短文本分类方法

针对高质量标注样本稀缺、标注难度大问题,本文

首先提出了基于主动半监督学习的异质信息网络(如图1所示),其主要包含三个模块:异质信息网络(Heterogeneous Information Network,HIN)、异质图注意力网络HGAT和主动半监督学习(Active Semi-Supervised Learning,ASSL)策略.HIN负责融合多模块特征;HGAT负责对HIN进行编码,同时使用双层注意力机制来减少特征融合时产生的噪声;ASSL负责从无标签样本集合中选择出高质量的样本进行标注,以丰富模型的训练数据.其中,基于信息增益的AL负责选择出最有价值的样本让专家进行判断标注,基于SSL的伪标签选择器负责选择出高准确率且高稳定性的正负样本进行打伪标签,最后将这些样本和训练集混合得到新的混合样本集合,结合负样本学习不断迭代训练和优化ASSL-HGAT模型.

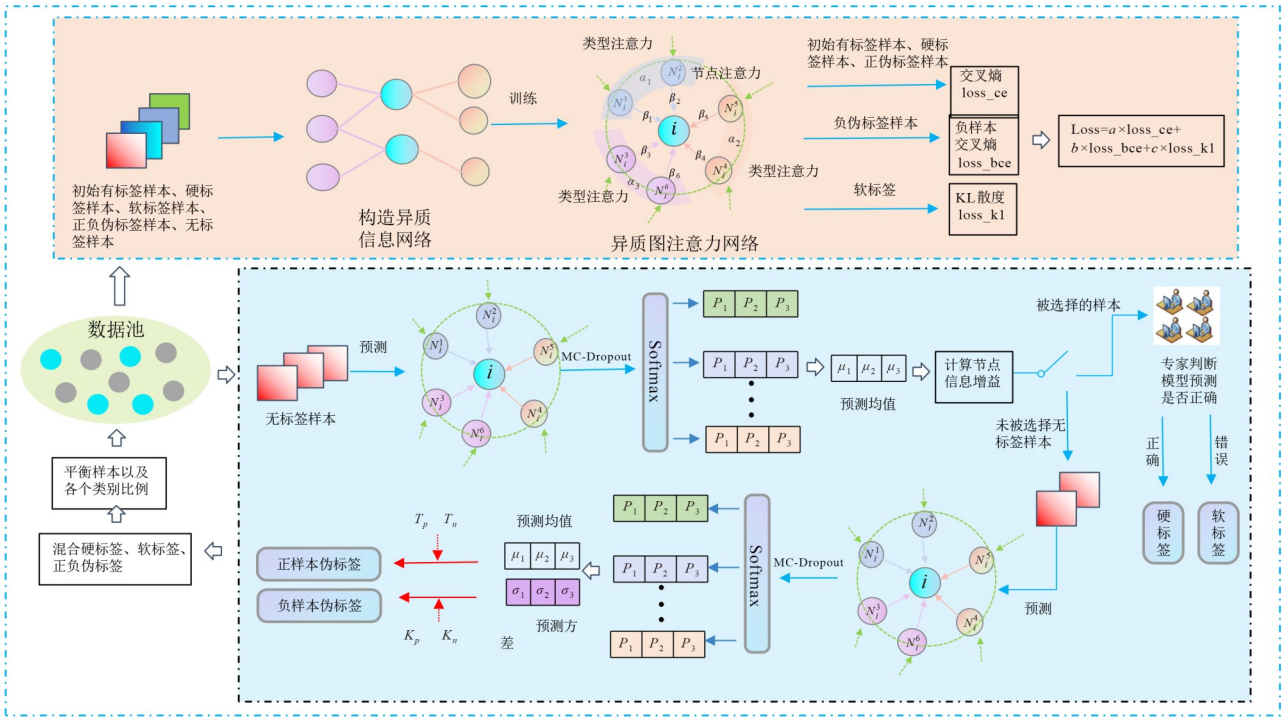


图1　基于主动半监督学习的短文本分类方法

本文的ASSL框架是将基于信息增益的主动学习与基于伪标签的半监督学习有机融合,旨在高效地利用有限的标注数据和大量的无标注数据训练模型.

3.1　异质信息网络

异质信息网络被应用于构造知识图谱增强的异质图结构,如图2所示.此时的节点集合是由文本节点、主题节点和实体节点构成,其中,$D=\{d_1,d_2,\cdots,d_m\}$表示文本节点集合,$T=\{t_1,t_2,\cdots,t_k\}$表示主题节点集合,$E=\{e_1,e_2,\cdots,e_n\}$表示在知识图谱中的实体节点集合.知识图增强的异质图结构可以有效地将文本的主题分布与知识库中的实体嵌入相结合,可提高分类性能,缓解短文本的语义稀疏问题.

主题信息节点表示使用LDA无监督聚类算法挖掘可能潜在的主题.本文使用开源的bert-as-service服务来生成基于BERT的短文本向量嵌入和语义信息特征.

为了解决短文本的语义稀疏问题,可利用来自外部知识源的附加信息作为先验知识.本模型通过实体链接工具TAGME[3]链接到知识图谱,从文本中识别实体并链接到维基百科中对应实体的描述信息.

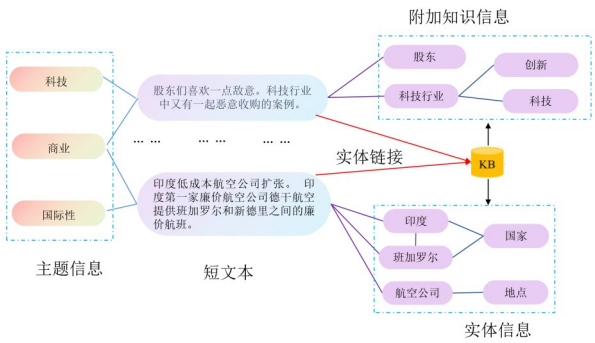


图2　异质信息网络结构

3.2　异质图注意力网络

异质图卷积的目标是聚合短文本周围的节点主题信息和实体信息.对于每一个实体,使用知识图谱中实体的描述文本或者概念信息作为其嵌入向量表示.对于图$G=(N,R)$,定义领接矩阵为$\boldsymbol{A}$,则加入自连接的邻接矩阵为$\boldsymbol{A}'=\boldsymbol{A}+\boldsymbol{I}$.定义$\boldsymbol{M}$为度矩阵,其中,

$\boldsymbol{M}_{ii}=\sum_j \boldsymbol{A}'_{ij}$. 考虑到不同的节点类型，将多种节点的特征空间组合起来生成一个新的特征空间 $T=\{\tau_1,\tau_2,\tau_3\}$. 第 l 层的传播公式如下所示：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\sum_{\tau\in T}\widetilde{\boldsymbol{A}}_\tau\times\boldsymbol{H}_\tau^{(l)}\times\boldsymbol{W}_\tau^{(l)}\right) \tag{1}$$

其中，$\widetilde{\boldsymbol{A}}_\tau\in\mathbb{R}^{|N|\times|N_\tau|}$ 是 $\tilde{\boldsymbol{A}}=\boldsymbol{M}^{-1/2}\boldsymbol{A}'\boldsymbol{M}^{-1/2}$ 的一个子矩阵. $\boldsymbol{H}^{(l+1)}$ 是通过聚合 $\boldsymbol{H}^{(l)}$ 附近节点特征生成的.

接着使用双层注意力机制来减少特征融合时产生的噪声. 给定一个节点 v，类型级别的注意力分数是根据当前节点嵌入表示 h_v 和类型嵌入 h_τ 计算的，这里的 $h_\tau=\sum_{v'}\widetilde{\boldsymbol{A}}_{vv'}h_{v'}$，其中，$h_{v'}$ 表示节点 v 附近类型为 τ 的特征. 类型级别注意力计算公式如下：

$$a_\tau=\sigma\left(\boldsymbol{\mu}_\tau^T\left[h_v\right]\left[h_\tau\right]\right) \tag{2}$$

其中，$\boldsymbol{\mu}_\tau^T$ 是类型注意力向量，$\sigma(\cdot)$ 表示 ReLU 激活函数. 类型注意力的权重归一化公式如下：

$$a_\tau=\frac{\exp(a_\tau)}{\sum_{\tau'\in T}\exp(a_{\tau'})} \tag{3}$$

根据当前节点 h_v、其附近节点 $h_{v'}(v'\in N_v)$ 和类型级注意力 $\alpha_{\tau'}$，可以计算节点级注意力分数，计算公式如下：

$$b_{vv'}=\sigma\left(v^T\times\alpha_{\tau'}\left[h_v\|h_{v'}\right]\right) \tag{4}$$

节点注意力的权重归一化计算公式如下：

$$\beta_{vv'}=\frac{\exp(b_{vv'})}{\sum_{i\in N_v}\exp(b_{vi})} \tag{5}$$

最后，在异质图卷积中加入类型级别和节点级别的双层注意力，公式如下：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\sum_{\tau\in T}\boldsymbol{B}_\tau\times\boldsymbol{H}_\tau^{(l)}\times\boldsymbol{W}_\tau^{(l)}\right) \tag{6}$$

其中，$\boldsymbol{B}_\tau$ 表示注意力矩阵，第 v 行和 v' 列的元素对应式(5)中 $\beta_{vv'}$.

3.3 基于信息增益的主动学习方法

传统的主动学习方法假定专家可以始终提供硬标签. 但在某些领域或类别繁多的情况下，专家也很难给出确切的类标签. 例如，“ogbn-papers100M”任务旨在通过引用网络将 arXiv 论文的标签推断到 172 个主题领域中，但机器学习专家无法明确标记金融学科领域的查询实例，因为不属于其专业知识范围. 针对这一问题，作者提出基于信息增益的主动学习方法，其核心思想是将主动学习转化为二元分类任务，这意味着领域专家只需判断模型预测标签正确与否，大大降低了标注难度.

图 3 对比了基于二元标记策略的主动学习方法与传统的主动学习方式. 对于任意给定节点，传统的主动学习方法会提出“它属于哪个类别?”的问题，专家需给出确切的类别标签. 基于二元策略方法会提出“样本是否属于第一个类别?”，这种问题更容易得到回答.

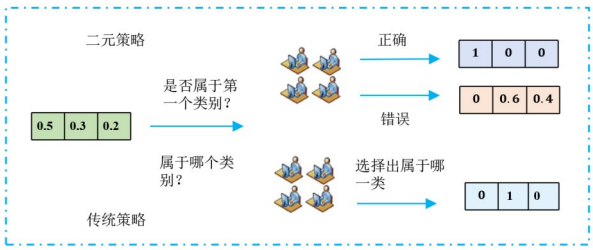


图 3 二元标记策略

与先前让专家直接给出确切类别的标注方法不同，使用二元标记策略方法专家只需要判断模型预测结果 $\hat{y}$ 是否正确. 对于预测标签正确的样本，采用交叉熵计算损失，并将其作为硬标签，对于预测标签错误的样本，采用 KL(Kullback-Leibler)散度计算损失，并将其作为软标签. 这样，无论是正确还是错误的预测，都能提供信息增益，从而加强标签监督. 通过专家的判断，样本标签的不确定性将降低.

信息增益是指熵减少的期望，即样本 v_i 的信息增益 $\mathrm{IG}(v_i)=\mathrm{E}\left[H(\hat{y}_i)-H(\dot{y}_i)\right]$，其中，$H$ 表示熵函数，$H(\hat{y}_i)$ 表示预测样本的原始熵，$H(\dot{y}_i)$ 表示专家判断后预测样本的熵，信息增益函数定义如下：

$$\begin{aligned}\mathrm{IG}(v_i)=&P(v_i+)\left(H(\hat{y}_i)-H(\dot{y}_i,v_i+)\right)\\&+P(v_i-)\left(H(\hat{y}_i)-H(\dot{y}_i,v_i-)\right)\end{aligned} \tag{7}$$

其中，$P(v_i+)$ 表示样本 v_i 预测标签是正确的概率，$P(v_i-)$ 表示样本 v_i 预测标签是错误的概率；在样本预测标签正确的情况下，专家判断后其熵会下降到 0，即 $H(\dot{y}_i,v_i+)=0$，故信息增益函数简化为

$$\mathrm{IG}(v_i)=H(\hat{y}_i)+P(v_i-)H(\dot{y}_i,v_i-) \tag{8}$$

其中，$H(\dot{y}_i,v_i+)$、$H(\dot{y}_i,v_i-)$ 分别为专家判断样本预测标签是正确或错误的熵.

3.4 基于不确定性感知的伪标签选择器

传统的半监督伪标签选择方法由于网络校准差，导致伪标签样本不正确、训练混乱、适应性差，因此效果不佳. 研究表明[9]，选择不确定性值低的样本可以显著减少校准不良的后果，提高通用性. 基于此，本文使用 MC-Dropout[17] 策略，在模型预测阶段多次使用 dropout 对同一输入(即同一无标签数据)进行多次预测，计算预测值的平均值和方差，以获取各无标签数据的不确定性值. 定义 $D_L=\left\{\left(\boldsymbol{x}^{(i)},y^{(i)}\right)\right\}\left(i=1,2,\cdots,N_L\right)$ 表示有 N_L 个样本的有标签样本集合；定义 $D_U=$

$\{(\boldsymbol{x}^{(i)})\}(i=1,2,\cdots,N_U)$表示有$N_U$个样本的无标签样本集合. 其中,向量$\boldsymbol{x}^{(i)}$是文本输入,$y^{(i)}=\{y_1^{(i)},y_2^{(i)},\cdots,y_c^{(i)}\}$是其所对应的标签. 首先,使用有标签样本集合$D_L$对模型进行监督训练;接着使用训练好的模型$f_\theta$对无标签样本集合$D_U$进行多次预测,并计算预测结果的均值和方差;利用均值和方差同时作为置信度,对无标签样本集合D_U打伪标签,并选择出正负样本伪标签$\tilde{y}^{(i)}$;混合所选择的伪标签数据和所述有标签数据集,获得混合样本集合$\tilde{D}=\{(\boldsymbol{x}^{(i)},\tilde{y}^{(i)})\}(i=1,2,\cdots,N_U+N_L)$.

定义$\boldsymbol{g}^{(i)}=[g_1^{(i)},g_2^{(i)},\cdots,g_c^{(i)}]\in\{0,1\}^C$表示一个二进制向量,即第$i$个样本中每一个类别是否被选为伪标签. 如果$g_1^{(i)}=1$表示$\tilde{y}^{(i)}$被选择为伪标签;反之,如果$g_1^{(i)}=0$表示$\tilde{y}^{(i)}$未选择为伪标签. 本文使用设正样本置信阈值$T_p$和负样本置信度阈值$T_n$来筛选出正负伪标签. $\boldsymbol{g}^{(i)}$由以下规则获得:

$$\begin{aligned}g_c^{(i)}=&1\left[\mu(p_c^{(i)})\leqslant K_p\right]1\left[p_c^{(i)}\geqslant T_p\right]\\&+1\left[\mu(p_c^{(i)})\leqslant K_n\right]1\left[p_c^{(i)}\leqslant T_n\right]\end{aligned}\tag{9}$$

其中,$p_c^{(i)}$表示类别c是样本标签的概率均值,$\mu(p)$表示概率p的不确定性,K_p和K_n表示选择正负样本的不确定性阈值即方差阈值.

3.5　融合联邦学习

为保护数据隐私并遵守数据安全法规,我们进一步提出了基于联邦学习的主动半监督短文本分类方法Fed-ASSL-HGAT,图4展示了Fed-ASSL-HGAT的联邦学习框架,该框架由一个服务器和多个客户端组成.

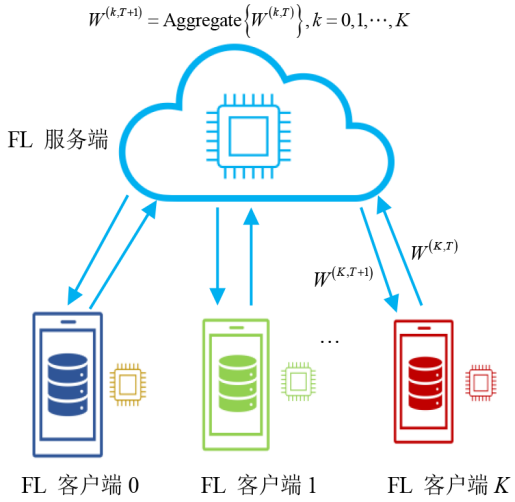


图4　联邦学习FedAvg框架

其中,服务器负责聚合来自通信正常客户端的模型参数,并生成全局共享模型参数;客户端主要负责从服务端下载全局贡献模型参数并更新本地模型参数,接着采用ASSL策略对模型进一步进行训练,训练完成后将各自的模型参数上传到服务器进行聚合. 具体地,服务端存储共享模型,客户端存储本地模型和本地私有数据,服务端对各模型参数聚合的公式如下:

$$W^{(k,T+1)}=\sum_{k=0}^{K}\frac{n_k}{n}W^{(k,T+1)}\tag{10}$$

其中,k表示客户端的数量,T表示迭代的轮次,n表示训练总样本数,n_k表示第k个客户端的训练样本数量.

3.6　损失函数

模型输出softmax层的计算公式如下:

$$Z=\text{softmax}\left(\boldsymbol{H}^{(L)}\right)\tag{11}$$

其中,输入是短文本的嵌入向量$\boldsymbol{H}^{(L)}$. 在模型训练过程中,训练数据的交叉熵损失计算公式如下:

$$L_{\text{CE}}=\text{CrossEntropy}\left(Y_p,Z\right)+\text{KullbackLeibler}(Y_s,Z)\tag{12}$$

其中,Y_p表示硬标签的标签概率分布,Y_s表示软标签的标签概率分布. 负样本交叉熵的损失计算如下:

$$L_{\text{NCE}}\left(\tilde{y}^{(i)},\hat{y}^{(i)}\right)=-\frac{1}{s^{(i)}}\sum_{c=1}^{C}g_c^{(i)}\left(1-\tilde{y}^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)\tag{13}$$

其中,$s^{(i)}$表示为第i个样本选择的伪标签的数量,$\hat{y}^{(i)}=f_\theta(x^{(i)})$表示模型$f_\theta$的概率输出. 最终的损失函数如下:

$$L=L_{\text{CE}}+L_{\text{NCE}}\tag{14}$$

4　实验分析

4.1　实验数据集

实验数据集采用3个开源的、具有代表性的半监督短文本分类数据集:

(1)AGNews数据集主要用于新闻文本分类,包含世界、体育、商业和科技四类;

(2)Snippets是从网络搜索引擎中检索到的文本片段;

(3)TagMyNews是一个常用的文本分类数据集,包含来自BBC新闻网站的新闻文章.

数据集信息如表1所示. 对于各数据集中的每个类别随机选择100个标注样本,其中,一半用于训练,另一半用于验证. 剩余样本一部分作为测试集,一部分作为未标注数据集.

4.2　实验评价指标及基准方法

本文采用分类任务最常用的精确度评价指标:准确率(ACCuracy,ACC)和F_1分数(F_1-score,F_1).

ASSL-HGAT的主干模型来自于HGAT[3],文献[3]通过实验对比了HGAT和其他半监督短文本分类方法,如PTE[18]、GAT[19]、HAN[20]、TextGCN[21]等,验证了HGAT的优越性能. 因此,所提出模型以HGAT为基础进行改进并以其作为基线模型.

HAN:基于预定义的元路径将HIN转化为多个同

表1 实验中使用的数据集

数据集	类别数量	训练集	验证集	测试集	无标签样本集合	文本字符长度(最大/最小/均值)
AGNews	4	200	200	2 640	3 000	770/75/195
Snippets	8	400	400	3 200	6 000	297/19/230
TagMyNews	7	350	350	4 550	5 250	105/15/80

质子网络,然后应用图注意力网络.

HGAT:使用短文本和附加信息构造HIN,然后应用基于双层注意机制的图神经网络对HIN进行嵌入.

4.3 实验环境和模型超参数

本文所有实验均在相同软硬件环境下进行,采用深度学习开源框架Pytorch进行模型搭建,编程语言采用Python3.6,硬件配置为NVIDIA Tesla P100-PCIE-12 GB、Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20 GHz、CUDA 10.2. 所提出模型与HGAT使用相同超参数,BERT模型中句子最大序列长度设置为128,隐藏层维度为512,学习率为0.005,丢弃率设置为0.8,推理阶段前向传播次数设置为10,正负样本置信度阈值分别为0.7、0.1,正负样本不确定性阈值为0.05.

4.4 实验结果与分析

4.4.1 ASSL-HGAT性能评估

为了验证ASSL-HGAT的性能,我们设计实验将其与其变体模型、基线HGAT模型进行了对比.

表2展示了ASSL-HGAT及其变体模型在AGNews、Snippets和TagMyNews测试集上的准确率和F_1分数. UPS(Uncertainty-aware Pseudo-label Selection)表示基于不确定性感知的伪标签选择框架. C代表类别数,ALL代表混合未标注样本和标注样本集合. 由于AL和SSL在每次迭代选择的样本都是不同的,因此带有AL或SSL模型的实验结果是有偏差的. 为了消除由随机性带来的影响并证明所提出带有AL或SSL方法的鲁棒性,对每组弱监督学习的实验进行了6轮重复实验,并计算平均值和标准偏差. 为了获得模型的性能上限,将所有训练数据和带真实标签的未标记数据集混合在HGAT(No.1)和HGAT+BERT(No.2)模型上进行有监督学习实验.

整体来看,所提出模型及其变体在3个数据集上的性能均优于HGAT基线,证明了主动半监督学习策略在半监督短文本分类中的实用性.

尽管HGAT与其他半监督模型相比具有明显优势[3],但仅靠图传播无法充分利用未标记数据,模型HGAT+UPS使用不确定性感知来提高正负伪标签的准确率,并采用负样本学习来纠正模型. 表中No.4和No.5显示,HGAT+UPS模型与HGAT模型相比F_1值在AGNews、Snippets、TagMyNews上分别实现了1.36%、2.82%和4.10%的性能提升. UPS框架可以提高模型性能的主要原因有两点:(1)未标记样本通常会提供额外的信息,帮助模型更好地捕捉数据分布,从而增强其泛化能力.(2)通过不确定性机制和负样本学习,模型可以从未标记样本数据集中选择出大量高度稳定且准确的正负样本用于标注,防止标注错误的噪声样本影响模型性能,提高模型的鲁棒性. 对比表中的No.4和No.6,引入主动学习策略的HGAT+AL模型达到了第二优性能,与No.4模型相比F_1值在AGNews、Snippets、TagMyNews上分别提升了2.22%、7.69%和5.69%.

表2 模型精度对比 单位:%

No	Text Embedding	Model	Labeled samples	Metrics	AGNews	Snippets	TagMyNews
1	Word2Vec	HGAT	ALL	ACC↑	85.71	94.18	65.85
				F_1↑	85.76	94.45	66.02
2	BERT	HGAT	ALL	ACC↑	89.39	92.09	79.16
				F_1↑	89.32	91.71	79.05
3	Word2Vec	HGAT	50C	ACC↑	78.23±0.71	78.46±1.01	42.54±0.35
				F_1↑	77.89±0.83	77.40±1.09	42.25±0.39
4	BERT	HGAT	50C	ACC↑	86.08±0.17	82.76±0.59	68.57±0.09
				F_1↑	85.99±0.18	82.04±0.49	68.42±0.19
5	BERT	HGAT+UPS	50C	ACC↑	87.44±0.38	85.82±0.67	72.54±0.16
				F_1↑	87.35±0.39	84.86±0.67	72.52±0.17
6	BERT	HGAT+AL	50C	ACC↑	<u>88.32±0.55</u>	<u>89.80±0.52</u>	<u>74.31±0.54</u>
				F_1↑	<u>88.21±0.55</u>	<u>89.73±0.81</u>	<u>74.11±0.49</u>
7	BERT	ASSL-HGAT	50C	ACC↑	**88.53**±0.31	**90.26**±0.64	**75.94**±0.45
				F_1↑	**88.44**±0.30	**90.15**±0.56	**75.88**±0.46

注:加粗数据表示最优的结果,下划线表示次优结果.

实验证明所提出的基于信息增益的主动学习策略能够在降低专家标注难度的同时显著提升模型的性能(尤其是在分类类别较多的场景下,专家标注难度显著降低).

我们提出的ASSL-HGAT模型分类性能最优,在AGNews、Snippets、TagMyNews数据集上F_1值分别达到88.44%、90.15%、75.88%. ASSL-HGAT模型在AGNews上超过No.1有监督方法的上限性能2.68%,在TagMyNews上超过No.1有监督方法的上限性能9.86%. 表中No.5~No.7的结果显示,主动学习和半监督学习可有效地协同工作,形成闭环学习,提高短文本分类的性能.

4.4.2 基于信息增益的主动学习方法的性能评估

为了进一步验证所提出基于信息增益的主动学习策略的优越性,我们在各种主动学习策略下对HGAT进行了实验,实验中所有模型均采用BERT预训练模型生成文本嵌入向量表示. 设置标签预算为每类50个标签样本,以展示端到端的准确性.

表3展示了AL-HGAT模型在不同主动学习选择策略下的精度. 相比于选择标注,将主动学习转化成为二元分类任务之后,各个主动学习策略的精确度均有提升,尤其是多分类数据集上性能提高显著. 例如,基于信息增益的主动学习选择策略在判断标注方式下相比于选择标注在AGNews、Snippets、TagMyNews数据集上F_1值分别提高了0.8%、5.16%、3.87%. 这表明相同预算情况下,判断标注方式不仅能降低专家标注样本的难度,同时也能保证模型性能的稳定提升.

表3 AL-HGAT模型在不同主动学习选择策略下的精度对比　　单位:%

标注方式	标注数量	主动学习策略	AGNews/4分类		Snippets/8分类		TagMyNews/7分类	
			ACC↑	F_1↑	ACC↑	F_1↑	ACC↑	F_1↑
选择标注	ALL	有监督	89.39	89.32	92.09	91.71	79.16	79.05
	50C	有监督	86.38	86.27	81.03	80.50	68.03	67.96
	10C ~ 50C	随机	86.33±0.64	86.24±0.65	82.06±2.67	81.41±2.52	69.14±0.79	68.97±0.77
		最小置信度	86.72±0.35	86.60±0.38	84.11±1.11	83.61±0.92	71.40±0.23	71.35±0.18
		密度熵	87.20±0.76	87.12±0.75	84.65±0.89	84.26±0.33	70.89±0.64	70.36±0.82
		核心集	87.19±0.17	87.10±0.18	84.22±0.51	83.99±0.52	71.58±0.37	71.57±0.48
		信息增益	87.61±0.52	87.41±0.45	85.14±0.42	84.57±0.37	70.34±0.74	70.24±0.67
判断标注	10C ~ 50C	随机	87.01±0.49	86.90±0.46	85.83±1.50	85.10±1.59	72.75±0.72	72.59±0.53
		最小置信度	88.05±0.41	87.94±0.41	88.53±0.57	88.19±0.59	72.80±0.90	72.78±0.96
		密度熵	87.59±0.43	87.51±0.44	87.56±0.75	86.69±0.70	72.95±0.74	73.00±0.73
		核心集	87.97±0.10	87.86±0.11	86.42±0.46	86.44±1.07	73.37±0.62	73.18±0.66
		信息增益	**88.32**±0.55	**88.21**±0.55	**89.80**±0.52	**89.73**±0.81	**74.31**±0.54	**74.11**±0.49

注:加粗数据表示最优的结果,下划线表示次优结果.

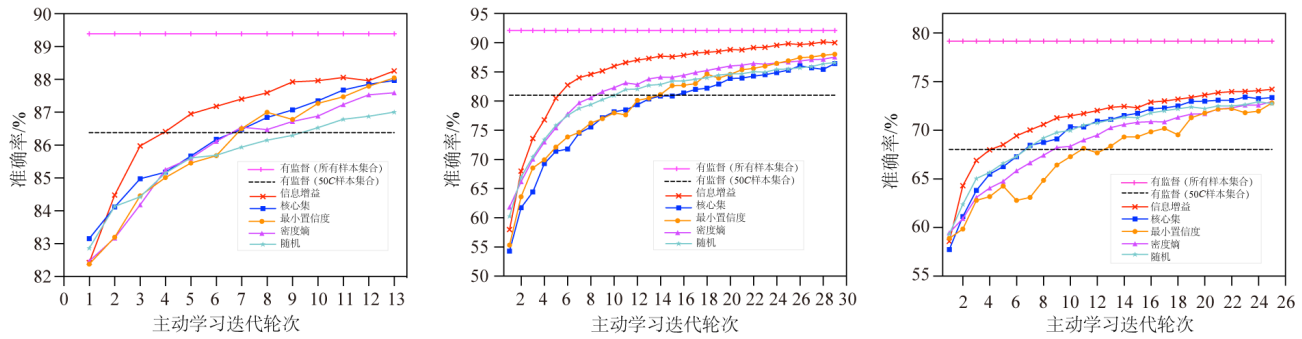
与训练样本标注数量是50C的有监督学习比较,除了随机选择策略外,其他主动学习选择策略精确度均有提升,说明标注预算相同情况下,通过主动学习选择的样本质量更高,可提高模型性能. 其中,在判断标注场景下,基于信息增益的主动学习选择策略的模型性能提升最大,在AGNews、Snippets、TagMyNews数据集上F_1值分别提高了1.94%、9.23%、6.15%,证明信息增益选择策略在判断标注方式下对模型性能的提升有显著优势,也特别适合判断标注的场景. 图5分别展示了判断场景下不同的主动学习选择策略AL-HGAT模型在三个数据集上的准确率趋势图. 整体来看,判断场景下对于三个数据集,基于信息增益的主动学习选择策略在每一轮迭代均处于性能领先,证明了基于信息增益的AL-HGAT模型在标注场景具有性能优势和强大的泛化能力.

4.4.3 基于联邦学习和主动半监督学习的短文本分类方法的性能评估

为了评估提出的Fed-ASSL-HGAT的性能,我们对数据集进行了重新划分并构建非独立同分布数据集. 具体地,首先将测试集、训练集和无标注样本进行混合得到V_{mix},将验证集作为测试结果,使用狄雷克雷对V_{mix}进行划分得到一个客户端的样本集合$V_{\text{mix},i}$,设len_i表示$V_{\text{mix},i}$样本集合的大小,对于包含主动学习的模型训练集的初始样本数为$0.05\times\text{len}_i$,主动学习阶段每次每个客户端增加$0.05\times\text{len}_i$个样本数量,直到增加到$0.3\times\text{len}_i$终止迭代.

本文基于狄雷克雷分布[22](q~Dir(αp))将现有数据集划分到不同的客户端上,使其私有数据满足非独立同分布. 在这里,p表示在N个类别上的先验类分布,α是用来控制生成不同的程度的非独立同分布数据,α越大表示客户端之间类分布越相似,α越小表示客户端之间类分布差距越大. 对于联邦学习的实验,所有数据集都设置10个客户端,α设置为0.3、0.5和1.

实验结果如表4所示,其中,Fed-HGAT表示HGAT模型与FedAvg算法融合后的模型,Fed-ASSL-HGAT(S)



(a) AGNews (b) Snippets (c) TagMyNews

图5 判断场景下不同的主动学习选择策略模型趋势图

表示在选择标注方式下IG-ASSL-HGAT与FedAvg算法融合后的模型，Fed-ASSL-HGAT(J)表示在判断标注方式下IG-ASSL-HGAT与FedAvg算法融合后的模型．通过模型在非独立同分布数据集上的精度比较可看出，随着浓度参数α的不断变大，模型的训练效果也在不断提升，并在$\alpha=1$的时候各个模型的效果达到了较为满意的性能AGNews的实验结果显示，Fed-HGAT在$\alpha=1$的非独立同分布数据上ACC达到87.84%，与有监督(ALL)仅相差1.52%，F_1值达到87.74%，与有监督(ALL)相差少了1.58%，即在保护数据隐私的情况下，仅使用30%的训练样本也能达到令人满意的效果．

表4 模型在非独立同分布数据集上精度比较 单位：%

模型	浓度参数的α	AGNews/4分类		Snippets/8分类		TagMyNews/7分类	
		ACC↑	F_1↑	ACC↑	F_1↑	ACC↑	F_1↑
HGAT	有监督(ALL)	89.39	89.32	92.09	91.71	79.16	79.05
Fed-HGAT	0.3	68.78	60.11	85.91	80.88	66.24	64.72
Fed-ASSL-HGAT (S)		66.51	58.66	83.09	77.78	53.56	49.59
Fed-ASSL-HGAT (J)		61.28	61.08	88.56	87.81	64.08	61.03
Fed-HGAT	0.5	85.18	84.95	89.06	88.17	69.51	68.21
Fed-ASSL-HGAT (S)		69.72	61.42	86.41	84.21	54.63	54.09
Fed-ASSL-HGAT (J)		74.84	71.93	90.71	89.89	69.94	69.51
Fed-HGAT	1	**87.84**	**87.74**	89.78	88.9	74.39	74.89
Fed-ASSL-HGAT (S)		80.15	79.40	87.06	86.07	64.54	63.95
Fed-ASSL-HGAT (J)		84.24	84.24	**93.78**	**93.77**	**77.09**	**77.36**

注：加粗数据表示最优的结果．

Snippets的实验结果显示，Fed-ASSL-HGAT(J)在$\alpha=1$的非独立同分布数据上F_1值达到93.77%，比有监督(ALL)提高了2.06%，表明Fed-ASSL-HGAT(J)在Snippets数据集的性能优越性．TagMyNews的实验结果显示，Fed-ASSL-HGAT(J)在$\alpha=1$的非独立同分布数据上F_1值达到77.36%，与有监督(ALL)相差了1.69%．对比Fed-ASSL-HGAT(S)和Fed-ASSL-HGAT(J)两组模型实验结果，当$\alpha=0.5$和$\alpha=1$时，Fed-ASSL-HGAT(J)模型性能一直领先，表明融合使用判断标注方式的基于信息增益的主动半监督学习方法和联邦学习算法效果更佳．

Fed-ASSL-HGAT模型通过融合联邦学习和主动半监督学习异质图注意力网络实现了标记样本少量情况下对文本的准确分类，同时在保护数据隐私情况下充分挖掘数据价值．此外，将集中式训练模式转换为联邦式训练模式，打破了单个显卡的大小限制，可在多张显卡中进行联邦式合作训练．其中，半监督学习改善了联邦学习中客户端标记数据稀缺的问题，主动学习可以提供高质量的有标注数据，三者结合显著提高了模型分类的准确度和泛化能力．

5 结论

本文针对高质量标注样本稀缺、标注难度大问题，首先提出了基于主动半监督学习的异质图注意力网络ASSL-HGAT，使用一个新颖的ASSL框架对HGAT进行赋能．将基于信息增益的主动学习与基于伪标签的半监督学习进行平滑融合，提高了训练数据的质量和模型的泛化能力，实现了在高质量标注样本稀缺的情况下，准确对常见的短文本进行分类，其中，针对多类别分类任务中样本标注困难问题，将主动学习转化成一个二元判断问题，通过二元标记策略大大降低了专家

标注数据的难度. 融合联邦学习的Fed-ASSL-HGAT模型可在不泄漏隐私数据的情况下满足性能要求.

参考文献

[1] 张昱, 刘开峰, 张全新, 等. 基于组合-卷积神经网络的中文新闻文本分类[J]. 电子学报, 2021, 49(6): 1059-1067.
ZHANG Y, LIU K F, ZHANG Q X, et al. A combined-convolutional neural network for Chinese news text classification[J]. Acta Electronica Sinica, 2021, 49(6): 1059-1067. (in Chinese)

[2] 李雪莹, 王田路, 梁鹏, 等. 基于系统模型的用户评论中非功能需求的自动分类[J]. 电子学报, 2022, 50(9): 2079-2089.
LI X Y, WANG T L, LIANG P, et al. Automatic classification of non-functional requirements in App user reviews based on system model[J]. Acta Electronica Sinica, 2022, 50(9): 2079-2089. (in Chinese)

[3] YANG T C, HU L M, SHI C, et al. HGAT: Heterogeneous graph attention networks for semi-supervised short text classification[J]. ACM Transactions on Information Systems, 39(3): 32.

[4] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[EB/OL]. (2016-02-17) [2023-07-01]. http://arxiv.org/abs/1602.05629.

[5] ISCEN A, TOLIAS G, AVRITHIS Y, et al. Label propagation for deep semi-supervised learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 5065-5074.

[6] HAASE-SCHÜTZ C, STAL R, HERTLEIN H, et al. Iterative label improvement: Robust training by confidence based filtering and dataset partitioning[C]//2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2021: 9483-9490.

[7] BERTHELOT D, CARLINI N, Goodfellow I, et al. Mixmatch: A holistic approach to semi-supervised learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2019: 5049-5059.

[8] ZHOU T Y, WANG S J, BILMES J A. Time-consistent self-supervision for semi-supervised learning[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 11523-11533.

[9] RIZVE M N, DUARTE K, RAWAT Y S, et al. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning[EB/OL]. (2021-01-15)[2023-07-01]. http://arxiv.org/abs/2101.06329.

[10] 李延超, 肖甫, 陈志, 等. 自适应主动半监督学习方法[J]. 软件学报, 2020, 31(12): 3808-3822.
LI Y C, XIAO F, CHEN Z, et al. Adaptive active learning for semi-supervised learning[J]. Journal of Software, 2020, 31(12): 3808-3822. (in Chinese)

[11] ZHANG W Q, ZHU L, HALLINAN J, et al. BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 20634-20644.

[12] ZHANG W T, WANG Y X, YOU Z B, et al. Information gain propagation: A new way to graph active learning with soft labels[EB/OL]. (2022-03-02)[2023-07-01]. http://arxiv.org/abs/2203.01093.

[13] KONEČNÝ J, MCMAHAN H B, RAMAGE D, et al. Federated optimization: Distributed machine learning for on-device intelligence[EB/OL]. (2016-10-08) [2023-07-01]. http://arxiv.org/abs/1610.02527.

[14] HSU T M H, QI H, BROWN M. Federated visual classification with real-world data distribution[C]//European Conference on Computer Vision. Cham: Springer, 2020: 76-92.

[15] JEONG W, YOON J, YANG E, et al. Federated semi-supervised learning with inter-client consistency & disjoint learning[EB/OL]. (2020-06-22) [2023-07-01]. http://arxiv.org/abs/2006.12097.

[16] AHN J H, KIM K, KOH J, et al. Federated active learning (F-AL): An efficient annotation strategy for federated learning[EB/OL]. (2022-02-01) [2023-07-01]. http://arxiv.org/abs/2202.00195.

[17] GAL Y, GHAHRAMANI Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. New York: ACM, 2016: 1050-1059.

[18] TANG J, QU M, MEI Q Z. PTE: Predictive text embedding through large-scale heterogeneous text networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1165-1174.

[19] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. (2017-10-30) [2023-07-01]. http://arxiv.org/abs/1710.10903.
[20] WANG X, JI H Y, SHI C, et al. Heterogeneous graph attention network[C]//WWW'19: The World Wide Web Conference. New York: ACM, 2019: 2022-2032.
[21] YAO L, MAO C S, LUO Y. Graph convolutional networks for text classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 7370-7377.
[22] HSU T M H, QI H, BROWN M. Measuring the effects of non-identical data distribution for federated visual classification[EB/OL]. (2019-09-13) [2023-07-01]. https://arxiv.org/abs/1909.06335.

作者简介

孔德焱 男,1999年生于安徽省阜阳市. 现为北京交通大学软件学院硕士研究生. 主要研究方向为自然语言处理、图神经网络、弱监督学习.
E-mail: 1402039615@qq.com

冀振燕 女,1972年生于河南省许昌市. 现为北京交通大学软件学院教授. 主要研究方向为人工智能与大数据、智能交通领域软件工程. 中国电子学会会员编号:E190022451S.
E-mail: zhyji@bjtu.edu.cn

杨燕燕 女,1986年3月出生于河南省郑州市. 现为北京交通大学软件学院副教授. 主要研究方向为机器学习、不确定性人工智能.
E-mail: yangyy@bjtu.edu.cn

刘 洋 男,2000年5月生于辽宁省沈阳市. 现为北京交通大学软件学院硕士研究生. 主要研究方向为自然语言处理.
E-mail: liuyang052630@163.com

刘吉强 男,1973年3月生于山东省海阳市. 现为北京交通大学软件学院教授. 主要研究方向为隐私保护、物联网安全、区块链、可信计算等.
E-mail: jqliu@bjtu.edu.cn