DOI:10.19652/j.cnki.femt.2305427

多层次结构与半监督学习的谣言检测研究*

张岩珂[1,2] 但志平[1,2] 董方敏[1,2] 高 准[1,2] 张洪志[1,2]
(1.三峡大学水电工程智能视觉监测湖北省重点实验室 宜昌 443002;
2.三峡大学计算机与信息学院 宜昌 443002)

摘 要: 当前谣言检测工作主要基于监督学习,需要人为标记数据而导致检测具有滞后性。为了充分利用大量的未标记数据,及时检测社交网络中的虚假谣言。提出了一种基于多层次结构与半监督学习谣言检测模型(multi-level semi spuervised graph convolutional neural network,MSGCN)。该模型构建了一种多层次检测模块,基于图卷积网络对有限的标记样本进行训练以提取多层次传播结构特征、扩散结构特征和全局结构特征。其次,引入随机模型扰动集成无标签数据的动态输出进行一致性预测,提出互补伪标签法来获取高质量伪标签数据,并将其加入标记数据扩充样本。最后在有监督交叉熵损失和无监督一致性损失约束下提高模型质量。在公开的 Twitter15、Twitter16 和 Weibo 数据集上的实验结果表明,所提出模型在 30% 标记样本下准确率达到 88.3%、90.1%和 95.5%,在少量的标记样本下便可达到优异的成绩。

关键词: 谣言检测;半监督;层次结构;伪标签
中图分类号: TP183 **文献标识码:** A **国家标准学科分类代码:** 520.604

Research on rumor detection based on multilevel structure and semi supervised learning

Zhang Yanke[1,2] Dan Zhiping[1,2] Dong Fangmin[1,2] Gao Zhun[1,2] Zhang Hongzhi[1,2]
(1. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China; 2. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China)

Abstract: Social media generates a large amount of information, only a small portion of which can be labeled by professionals as true or false rumors. To make full use of the vast amount of unlabeled data and detect false rumors in a timely manner, proposes a model called MSGCN based on multi-level structure and semi supervised learning. This model constructs a multi-level detection module based on graph convolutional neural network to train limited labeled samples to extract multi-level propagation structure features, diffusion structure features, and global structure features. By perturbing the random model and integrating the dynamic output of unlabeled data for consistent prediction, the complementary pseudo label method is used to label the high confidence unlabeled data calculated by the model and add it to the training set to expand the sample. Under supervised cross-entropy loss and unsupervised consistency loss constraints, the model shows excellent performance. The experimental results on public Twitter15, Twitter16, and Weibo datasets show that the proposed model achieves accuracy of 88.3%, 90.1% and 95.5% under 30% labeled samples, can achieve excellent performance with a small number of labeled samples.

Keywords: rumor detection; semi-supervised; multilevel structure; pseudo label

0 引 言

网络社交媒体(如微博、Twitter 等)由于其信息的实时共享性和全球连通性,已成为多数人依赖的社交媒体服务平台。然而,也正是由于这些特性,这类网络社交媒体上充斥着大量的谣言。谣言包括故意制造或可证实的虚

收稿日期: 2023-08-14
* **基金项目:** NSFC-新疆联合基金(U1703261)项目资助

假信息，以误导读者，其动机是追求个人或组织利润[1]。谣言会给社会带来巨大的负面影响，严重的甚至会危害到人们生命财产安全。阻止谣言传播的重要方法之一就是及时发现谣言。谣言检测就是通过分析提取信息的内容和传播模式等相关信息来判断该信息的真实性[2]。

国内外许多学者都针对谣言检测领域进行了深入的研究。早期的谣言检测主要利用机器学习，人工提取特征来进行检测[3]。近年来随着深度学习的发展，神经网络可以挖掘到更多的细节，研究者们开始关注深层次的特征信息[4-8]。强子珊等[9]通过构建多模态的异质图来处理文本、图片和用户之间的特征，在小型话题社区谣言检测中获得了较高的准确率。冯茹嘉[10]等利用预训练模型解决了一词多义等问题，显著提升了检测精度。Ma 等[11]提出利用递归神经网络(RNN)来捕捉潜在的时空语义特征，并通过添加复杂的递归单元和隐藏层来有效的提高谣言检测的性能，在 Twitter 数据集上取得较好效果。李奥等[12]提出一种生成对抗网络模型，通过对抗网络生成器和判别器的相互促进作用，强化谣言文本的学习。Bian 等[13]通过自顶向下和自下向上的节点更新来获得谣言的传播和扩散特征，并提出了一种双向图卷积的谣言检测模型，在微博和 Twitter 数据集上取得良好成绩，同时证实了有向的传播结构有助于提升谣言的检测效果。谢欣彤等[14]提出基于传播用户代表性特征学习方法，从多种用户行为进行分析筛选后构造其传播路径，提升了检测效率。

这些监督方法虽然在不同的数据集上获得了优异的成绩，但都需要依靠大量的标记数据进行支撑，并且多数依赖人工标注，存在工作量巨大且一定程度的检测滞后性，同时谣言事件在社交媒体上的传播速度非常快，因此在实际情况中只有非常有限的标记数据可用于谣言检测，致使有监督模型在应用中通常会因标记样本不足而表现出较差的性能。此外，难以保证大数据标注的一致性[15]。随着数据量的增加，标注的不一致现象会越来越严重。

因此，使用未标记的数据来进行谣言检测、更加有效的利用少量的标记样本成为一种实用性较高的解决方案，具有重要的理论研究意义和广泛的社会应用价值。

针对以上问题，本文提出一种基于多层次结构与半监督学习谣言检测模型(multi-level semi spuervised graph convolutional neural network，MSGCN)的方法。该模型构建了事件的传播、扩散和全局结构特征以学习事件的多层次结构以更好地利用谣言的有向特征和全局关联特征。同时利用有限的标记数据引导模型的训练方向，使用大量的无标签数据增强模型的泛化性能。针对无标签数据可能引入过多的噪声问题，通过集成无标签数据的动态输出并进行一致性训练来有效降低噪声的影响，并提出互补伪标签生成高质量伪标签扩充样本。

本文提出一种多层次结构与半监督学习的谣言检测方法。结合事件的多层次结构特征，同时使用标记数据和未标记数据联合训练模型，提升了检测效能，在少量标记数据下也可有效进行谣言检测。为了增强无标记数据利用效率，引入集成无标签输出的一致性训练，聚焦大量无标签数据的有效利用来增强模型的性能。摒弃了传统采用单一阈值生成伪标签的方法，提出互补标签生成伪标签的方法以减少噪声，提高伪标签的质量。

1 半监督谣言检测方法

本文提出一种基于传播结构与半监督学习的谣言检测模型 MSGCN。模型总体框架如图 1 所示，该框架主要由 4 部分构成，分别为输入表示层、检测模块、输出模块和伪标签生成模块。

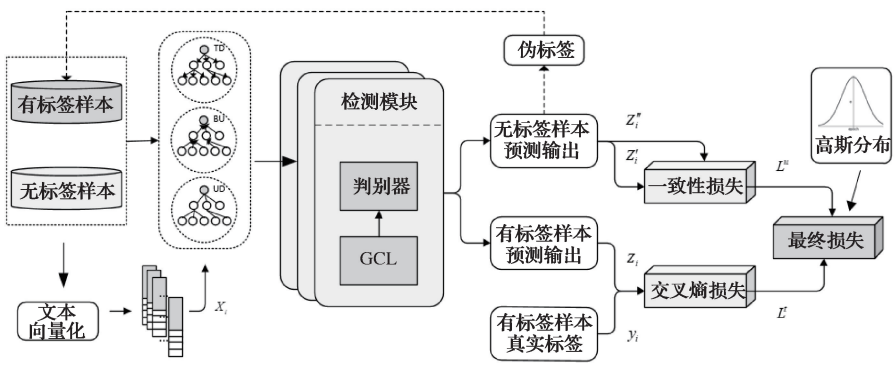

图 1 MSGCN 半监督网络模型架构

1.1 任务定义

定义 1 给定一组少量的标记数据集 C，记为 $C=\{(c_1,y_1),(c_2,y_2),\cdots,(c_n,y_n)\}$，其中 c_i 是第 i 个事件，y_i 是第 i 个事件的标签，n 为事件的个数。其中每个事件 c_i 具体地可表示为 $c_i=\{m_i,n_1^i,n_2^i,\cdots,n_{i-1}^i,G_i\}$，其中 m_i 表示事件原帖，n_j^i 表示第 j 个转发帖，G_i 为事件帖子的传播图。

定义 2 记一组大量的无标记数据 X 记为 $X=\{x_1,x_2,\cdots,x_m\}$，其中 m 为无标记数量。令 $f_\theta(\boldsymbol{Z}\mid x)$ 表示输入为 x 的样本，经过 θ 判别模型预测输出 $\boldsymbol{Z}$。

1.2 输入表示层

图卷积网络已被证实可以有效获取异构图信息[16-17]，由于网络社交媒体谣言信息的传播为一种异构图结构[18]，所以本文利用图卷积网络来获取其结构特征。

本文将文本类型的源推文与响应推文转换为 TF-IDF 值，并获取到 BiGCN 模型中的传播特征和扩散特征。首先将提取到的文本特征向量嵌入到节点中，并将事件 c_i 的文本特征向量使用 $\boldsymbol{X}_i=[\boldsymbol{x}_0^i,\boldsymbol{x}_1^i,\cdots,\boldsymbol{x}_{n_{i-1}}^i]\in\boldsymbol{R}^{N_i\times\boldsymbol{D}_i}$ 表示，$\boldsymbol{D}_i$ 为特征维度。如果节点之间存在转发或评论关系，则认为两个节点之间存在连接，建立边关系。最后将父节点指向子节点方向的图表示为谣言的传播图，将子节点指向父节点的图表示为谣言的扩散图。同时为了兼顾谣言的全局结构特征，引入无向图来获取谣言传播的全局结构特征。

对于图 $G_i=\{G_i^{TD},G_i^{BU},G_i^{UD}\}=\{E_i,V_i\}$，构建其事件 c_i 的邻接矩阵 $\boldsymbol{A}_i\in\boldsymbol{R}^{N_i\times N_i}$，$N_i$ 为节点数目，$E_i=\{c_i^{st}\mid s,t=0,1,\cdots,n_{i-1}\}$ 为事件 c_i 中所有转发节点与评论节点之间边的集合。$V_i=\{m_i,n_1^i,n_2^i,\cdots,n_{i-1}^i\}$ 为事件之间节点的集合。G_i^{TD}、G_i^{BU}、G_i^{UD} 分别表示为传播图、扩散图和无向图，其邻接矩阵为 $\boldsymbol{A}_i^{TD}=\boldsymbol{A}_i,\boldsymbol{A}_i^{BU}=\boldsymbol{A}_i^{\mathrm{T}},\boldsymbol{A}_i^{UD}=\boldsymbol{A}_i+\boldsymbol{A}_i^{\mathrm{T}}$，邻接矩阵分别包含了根节点和其子节点间所有的传播信息、扩散信息和全局结构信息。

1.3 检测模块

本文使用 3 个图卷积层捕捉 3 种层次图的空间特征，每个图卷积层都是相互独立的，并且各包含两层 GCN 模块。如图 2 所示，经过 3 个独立的图卷积层后获得了传播和扩散的双向信息和全局特征，并输出为一个向量矩阵 $\boldsymbol{H}\in\boldsymbol{R}^{N\times c}$，$N$ 是节点数，c 是分类数。用非线性函数表示为：

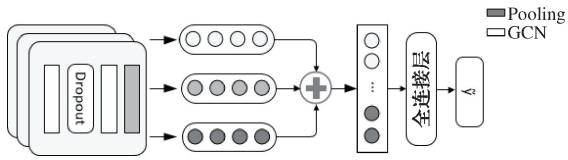

图 2 多层次检测模块

$$\boldsymbol{H}^{(d+1)}=\sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^d\boldsymbol{W}^d)=\sigma(\hat{\boldsymbol{A}}\boldsymbol{H}^d\boldsymbol{W}^d)\tag{1}$$

式中：d 是网络层数；$\sigma(x)$ 为激活函数；$\boldsymbol{W}$ 为参数矩阵。$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$ 表示邻接矩阵 $\boldsymbol{A}'$ 进行归一化后的结果矩阵，$\tilde{\boldsymbol{D}}$ 为矩阵 $\hat{\boldsymbol{A}}$ 的对角矩阵。同时为了充分获取特征表示，防止层级间特征流失，进行特征增强处理，将每层隐藏特征向量与前一层进行拼接，如下式所示：

$$\tilde{\boldsymbol{H}}_m^d=\sigma(\hat{\boldsymbol{A}}_i^m\tilde{\boldsymbol{H}}_m^{(d-1)}\boldsymbol{w}_m^{(d-1)})\tag{2}$$

$$\boldsymbol{H}_m^d=concat(\tilde{\boldsymbol{H}}_m^d,\tilde{\boldsymbol{H}}_m^{(d-1)})\tag{3}$$

式中：d 是网络层数；$m\in\{TD,BU,UD\}$ 表示 3 种不同的层次图；$\tilde{\boldsymbol{H}}_m^d$ 为隐藏特征矩阵，如 $\tilde{\boldsymbol{H}}_{TD}^1$ 为 TD 图中第 1 层输出的隐藏特征矩阵。$\boldsymbol{w}_m^d\in\boldsymbol{R}^{k\times v_m}$ 是隐藏权重矩阵。$\sigma(x)$ 为非线性激活函数 ReLU。Dropout[19]通过无偏差地在网络中加入噪音而用于 GCL 层中避免过拟现象。由于可用于训练的有标记样本在训练前期不够充足，使用一层 Dropout 防止有标记数据的过拟合，同时也可在无标记样本中加入一定扰动，通过 DropEdge 来随机去除有向图中的边来获得同一样本不同的无标记输入，从而增强模型的泛化性。紧接着将矩阵 $\tilde{\boldsymbol{H}}_m^d$ 进行均值池化以获取所有节点特征信息。均值池化公式如下：

$$\boldsymbol{S}_m=mean\,pooling(\tilde{\boldsymbol{H}}_m^{\mathrm{d}})\tag{4}$$

然后将 3 个图卷积层池化后的特征进行拼接以获得其多层次特征信息。

$$\boldsymbol{S}=concat(\boldsymbol{S}_{TD},\boldsymbol{S}_{BU},\boldsymbol{S}_{UD})\tag{5}$$

拼接后的信息通过全连接层进行降维。最后为了方便表示预测结果，本文使用 Softmax 函数进行归一化，得到归一化后的结果为事件不同类别标签的向量概率。表达式如下：

$$\tilde{y}=Soft\max(FC(\boldsymbol{S}))\tag{6}$$

全连接层 $FC(\boldsymbol{S})=\boldsymbol{W}_i\boldsymbol{S}+b_i$，其中，$\boldsymbol{W}_i$ 和 b_i 为可学习的权重矩阵和偏置项。

1.4 损失函数

对于有标记样本，将检测模块的预测标签与样本的真实标签的交叉熵作为有监督交叉熵损失来评估标记样本。

$$L_i^t=-\frac{1}{|B|}\sum_{i\in B\cap C}\log\tilde{y}_i[y_i]\tag{7}$$

式中：B 是学习过程中的小批量；y_i 表示谣言事件 c_i 对应的真实标签；$\tilde{y}_i$ 表示对于谣言事件 c_i 模型预测的输出。

对于大量的无标记数据 X 进行两次检测，由于模型中使用了 Dropout，会使模型在训练时具有不同的连接权重。所以对无标签数据，要提高在模型扰动情况下对两次输入的预测结果一致性表示，最小化两个分布之间的度量，集成无标记数据的输出并进行一致性约束，无监督一致性损失为：

$$L_i^u=\frac{1}{Y|B|}\sum_{i\in B}\|\boldsymbol{Z}'_i-\boldsymbol{Z}''_i\|^2\tag{8}$$

式中：Y 为标签所有的值的集合；$\boldsymbol{Z}'_i$ 为无标记数据 c_i 在模型中的第 1 次训练后的预测输出；$\boldsymbol{Z}''_i$ 为该无标记数据第 2 次训练后的预测输出。

一致性训练通过使无标记数据的两次预测保持一致，更接近无标记数据的准确标签。降低噪声对模型的影响，对大量无标签数据进行充分利用。

总体损失函数 L 为有监督交叉熵损失 L_i^t 与无监督一致性损失 L_i^u 的加权：

$$L=L_i^t+w(t)L_i^u\tag{9}$$

式中：$w(t)$ 是随着训练时间 t 的增加缓慢沿高斯分布提升的权重函数。

$$w(t)=\exp\left(-5\left(1-\frac{t}{T}\right)^2\right)\tag{10}$$

最小化损失函数会逐渐将标记数据信息传播到未标记数据上。在训练初期，标记数据贡献较大，在训练后期，无标记数据将比标记数据贡献更多。

1.5 互补伪标签

传统的伪标签大多是严格的 one-hot 标签，即将置信度最大的类别作为标签。这通常会导致模型产生了错误的高置信伪标签而在训练中引入噪声，使模型性能受限于标签质量。为了降低错误伪标签的影响，本文提出互补伪标签的方法来获取高质量的伪标签。

互补伪标签由正负标签组成。具体来说，将无标记数据的两次输出分别定义正标签和负标签，其中正标签定义为：

$$\hat{y}_i = (\boldsymbol{Z}'_i \geqslant \lambda)_j \tag{11}$$

负标签定义为：

$$\check{y}_i = (\boldsymbol{Z}''_i \leqslant \gamma)_j \tag{12}$$

式中：λ 为正向阈值；γ 为负向阈值；$j \in Y$ 为各个类别标签。在二分类问题时，[0.1,0.9]相较于[0.4,0.6]更能推出正确的标签，但在模型训练初期，出现后者的可能性更高且多分类任务更容易出现相差不大的预测值，此时传统的伪标签会引入大量的噪声。在互补伪标签中，正伪标签用来筛选出部分可能的类别项，而当各分类预测类别相近时，负标签可以滤去不正确的类别。相较于传统的伪标签，互补伪标签可以同时监督多个类别并且有效减少错误标签带来的偏差问题。通过互补伪标签得到最终标签为：

$$\bar{y} = \operatorname{argmax}\left(\frac{\| \hat{y}_i \|_j}{\| \hat{y}_i \|_j + \| \boldsymbol{Z}''_i - \check{y}_i \|_j}\right) \tag{13}$$

式中：$\| \hat{y}_i \|_j$ 表示各个类别的模长；$\bar{y}$ 为最终要加入到标记数据集中的伪标签。通过一方面去除不正确类别，一方面增强正确的类别项进而获取高质量的伪标签。

2 实验与分析

2.1 数据集和预处理

本文选择的是公开的社交媒体谣言数据集 Twitter15、Twitter16 与 Weibo 作为实验数据集。其中，Liu 等[20]创建的 Twitter15 数据集共包含 1 490 个事件，Ma 等[21]创建的 Twitter16 数据集共包含 818 个事件。Ma 等[11]创建的 Weibo 数据集包含 2 351 个谣言事件和 2 313 个非谣言事件。在这些数据集中，节点指的是用户，边表示转发或回复关系，特征是根据 TF-IDF 值所提取的前 5 000 个字段。Twitter15 与 Twitter16 数据集包含 4 个标签，非谣言(NR)、虚假谣言(FR)、真实谣言(TR)和未经证实的谣言(UR)。数据集统计数据如表 1 所示。

由于数据集为监督数据集，本文去除了部分数据集中数据的标签，将其分为无标签组和有标签组，具体划分数据如表 2 所示。

表 1 数据集事件信息统计

参数	Weibo	Twitter15	Twitter16
# posts	3 805 656	331 612	204 820
# users	2 746 818	276 663	173 487
# events	4 664	1 490	818
# true rumors	2 351	374	205
# false rumors	2 313	370	205
# unverified rumors	0	374	203
# non-rumors	0	372	205

表 2 数据集划分

		Weibo	Twitter15	Twitter16
标记数据 10%	有标签数据	466	149	82
	无标签数据	4 198	1 341	736
标记数据 20%	有标签数据	933	298	164
	无标签数据	3 731	1 192	654
标记数据 30%	有标签数据	1 398	447	246
	无标签数据	3 266	1 043	572

2.2 评价指标与基线方法

为了评估模型性能，本文使用准确率(accuracy)、F_1 值与召回率(recall)作为评价指标与其他算法进行对比。准确率与 F_1 值的计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

式中：TP、TN、FP、FN 的具体含义如表 3 所示。

表 3 混淆矩阵

	实际正类	实际负类
预测正类	TP	FP
预测负类	FN	TN

本文选取了在谣言检测领域中取得过突出效果的模型作为比较的基线方法，并在 3 种数据集上进行对比。选用的基线方法如下。

1)PPC[22]，一种在时间序列上联合使用 RNN 和 CNN 模型的谣言检测方法，通过谣言传播路径上的用户特征得到谣言表示。

2)RvNN[11]，一种使用树状递归神经网络的谣言检测模型，通过聚合文本信息和结构信息得到谣言表示特征。

3)BiGCN[13]，一种基于图卷积网络的谣言检测方法。在构造无向树的基础上，利用 GCN 对树结构进行自顶向下和自底向上的卷积，获取谣言的传播和扩散特征。

4)PLAN[23]，一种基于 Transformer 注意力机制的谣言检测方法，通过将谣言传播的过程构建成时间序列，并

利用 Transformer 注意力机制学习谣言的表示。

5)EBGCN[24]，一种探究谣言传播不确定性的研究，通过采用贝叶斯方法自适应地重新考虑潜在关系的可靠性。

6)PPTK[25]，一种基于微博信息传播树的路径树核谣言检测研究，通过将用户的影响力、情感反馈和帖子内容等特征嵌入节点中来计算路径相似度。

由于 PPTK 方法没有公开实现代码，本文使用其原文中的实验结果，PPTK 方法仅对比了 Weibo 数据集。

2.3 实验设置

所有实验采用 python3.7 实现，使用 pytorch1.7.1 版本深度学习框架，CUDA 版本 11.0。硬件环境为 NVIDIA Geforce RTX 3060 显卡。

所有实验中每个节点的隐藏特征向量的维度为 64。初始特征向量维度为 5 000。最大迭代次数为 200 次，激活函数为 ReLU。伪标签负向置信度阈值 $\gamma=0.05$，正向置信度阈值在二分类问题时为 $\lambda=0.85$，在多分类问题时为 $\lambda=0.45$。采用 Adam 算法为优化函数，学习率 $\eta=0.000\ 5$，$Dropout=0.5$。用于训练的标记数据和未标记数据同验证数据的比例为 4∶1。

2.4 实验结果与分析

表 4～6 所示为本文模型 MSGCN 和其他基线方法在 Twitter15、Twitter16 和 Weibo 数据集上的谣言检测效果。由于 Weibo 数据集是二分类类型，本文给出真谣言事件和假谣言事件的准确率、召回率和 F_1 值；对于 Twitter15 和 Twitter16 数据集包含的 4 种不同形式的谣言类别，采用 F_1 值全面评测模型性能。其中每个评价指标的最优值用加粗字体表示。

表 4 Twitter15 数据集对比结果

方法	ACC	NR	FR	TR	UR
		F_1	F_1	F_1	F_1
RvNN	0.723	0.682	0.758	0.821	0.654
PPC	0.842	0.811	0.875	0.818	0.790
BiGCN	0.852	0.821	0.857	0.908	0.812
PLAN	0.845	0.823	0.858	0.895	0.802
EBGCN	0.876	0.854	**0.865**	0.918	0.818
MSGCN(10%)	0.843	0.797	0.793	0.869	0.790
MSGCN(20%)	0.871	0.848	0.826	0.891	0.809
MSGCN(30%)	**0.883**	**0.857**	0.849	**0.920**	**0.825**

由表 4～6 可知，在 3 个数据集上，MSGCN 在不同标记数据下的准确率总体优于其他谣言检测方法，其中在 Twitter15 和 Twitter16 数据集上使用 30%的标记数据准确率达到 88.3%和 90.1%。在微博数据集上准确率达到 95.5%，这表明本文方法使用少量的标记数据仍可以获得可接受的性能。RvNN 模型认为传播图中所有节点贡献度相同，对传播节点敏感，但忽略了文本信息特征导致效果不理想。PPC 模型由于联合使用了 RNN 和 CNN 结构

表 5 Twitter16 数据集对比结果

方法	ACC	NR	FR	TR	UR
		F_1	F_1	F_1	F_1
RvNN	0.737	0.622	0.743	0.835	0.708
PPC	0.863	0.820	0.898	0.843	0.837
BiGCN	0.886	0.830	0.861	0.933	0.871
PLAN	0.874	0.853	0.839	0.917	0.880
EBGCN	0.895	0.864	0.881	**0.944**	0.874
MSGCN(10%)	0.889	0.847	0.885	0.890	0.835
MSGCN(20%)	0.896	0.864	0.844	0.933	0.862
MSGCN(30%)	**0.901**	**0.881**	**0.898**	0.923	**0.885**

表 6 Weibo 数据集对比结果

方法	分类	ACC	F_1	Recall
RvNN	F	0.908	0.905	0.897
	T		0.911	0.918
PPC	F	0.921	0.923	0.950
	T		0.918	0.889
BiGCN	F	0.940	0.947	0.920
	T		0.933	0.943
PLAN	F	0.943	0.943	0.948
	T		0.942	0.937
EBGCN	F	0.941	0.943	0.946
	T		0.932	0.945
PPTK	F	0.935	0.933	0.910
	T		0.937	0.961
MSGCN(10%)	F	0.937	0.940	0.935
	T		0.945	0.951
MSGCN(20%)	F	0.950	0.954	0.938
	T		0.948	0.955
MSGCN(30%)	F	0.955	0.967	0.953
	T		0.951	0.960

表现出比 RvNN 模型更好的检测效果。但 PPC 模型仅使用线性结构进行特征提取，忽略了传播过程中的异构图信息。同样的，PLAN 模型由于使用 Transformer 结构提取特征并且关注到帖子之间的隐性特征从而获得了进一步的提升，但也忽略了对有向信息的利用。BiGCN 与 EBGCN 模型引入了图卷积网络并研究谣言传播特征，但未考虑全局结构信息。由于 Weibo 数据中包含信息复杂且繁多，PPTK 在考虑大量特征的同时也引入更多的噪声，导致其准确率有所下降。当有标记数据量增大时，效果提升变缓，例如在 Twitter15 数据集上，MSGCN 在 20%标记数据情况下准确率比 10%标记数据提高 2.8%，在 30%标记数据情况下准确率比 20%数据提高 1.2%。

上述比对都是同大量有标记数据集的方法进行的，且已经表现出更好的性能，为了进一步验证方法的有效性，改变上述基线方法标记数量的大小再进行比对，结果如表 7 所示。

表7 不同标记数据对比结果

数据集	算法	10%	20%	30%
Twitter15	RvNN	0.336	0.552	0.539
	PPC	0.613	0.711	0.747
	BiGCN	0.628	0.709	0.759
	PLAN	0.633	0.731	0.775
	MSGCN	**0.843**	**0.871**	**0.883**
Twitter16	RvNN	0.361	0.556	0.588
	PPC	0.577	0.621	0.689
	BiGCN	0.634	0.711	0.776
	PLAN	0.673	0.742	0.787
	MSGCN	**0.889**	**0.896**	**0.901**

由表7可以看出，在只有少量有标记数据训练模型的情况下，MSGCN的准确率远高于基线方法，这表明MSGCN可以在谣言传播的早期快速识别谣言，且不需要耗费巨量的人力进行样本标注，从而避免谣言检测的滞后性所造成的损失。

图3、4所示分别为不同标记数据下MSGCN在不同数据集中准确率和损失值的变化曲线。由图3可以直观看出MSGCN可以在短时间内利用少量的标记样本取得较高的准确率，例如在Twitter16和Weibo数据集中，30%标记数据下MSGCN在20次Epoch训练下准确率便接近80%，在Twitter15数据集中也只需30次Epoch便可接近80%的准确率。这表明MSGCN可以有效的对谣言传播事件进行快速谣言检测，快速谣言检测指在谣言传播的早期阶段进行检测谣言，可以有效地避免谣言传播所造成的危害。同时由图4可以直观看出，聚合了交叉熵损失和一致性损失的总体损失收敛速度较快，在较快的时间内便可达到较低的损失率，这也表明MSGCN模型具有良好的健壮性。

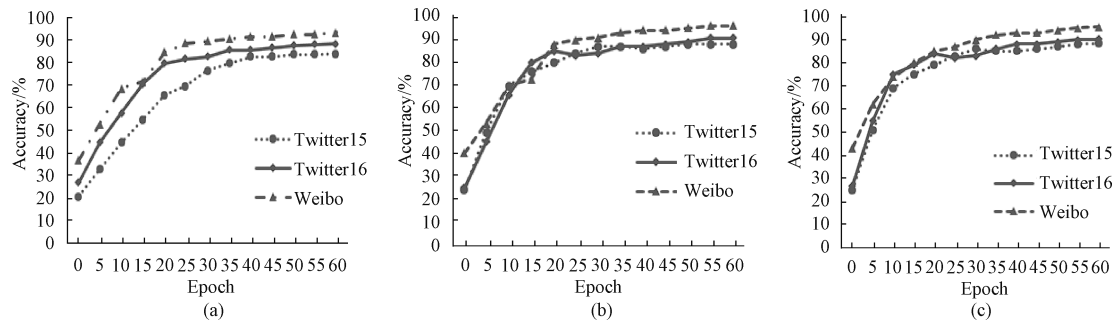

图3 (a)10%标记数据准确率变化；(b)20%标记数据准确率变化；(c)30%标记数据准确率变化

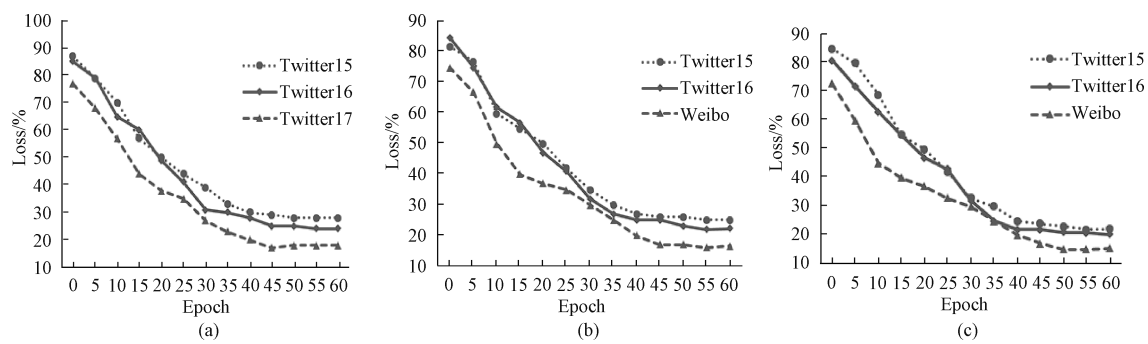

图4 (a)10%标记数据损失值变化；(b)20%标记数据损失值变化；(c)30%标记数据损失值变化

2.5 参数分析

1)无标记数据对模型的影响

为了验证无标记数据对MSGCN模型的准确率影响，选用Twitter15和Twitter16两个数据集进行实验，令每个类别有标记数据数量为20，每个类别无标记数据数量分别为50、100、150观察模型随着无标记数据数量的改变的准确率变化，结果如图5、6所示。

由图5、6可知，当有标记数据初始值为20时，随着无标记数据的增加，MSGCN的准确率和各分类类别的F_1值有不同程度的增长，在Twitter15数据集上每种类别分别增加50个无标记数据，准确率分别增加6%和4%，在Twitter16数据集上每种类别分别增加50无标记数据，准确率分别增加6%和2%。实验进一步验证了无标记数据对模型训练的有效性。

2)互补伪标签对模型的影响

为了验证互补伪标签对MSCGN模型的影响，分别在10%、20%和30%标记数据下对比互补伪标签和传统伪标签。实验结果如图7～9所示。

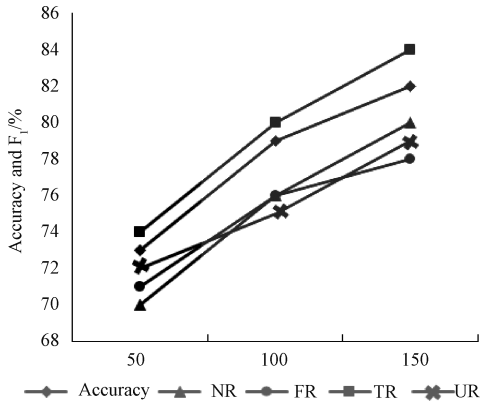

图 5 Twitter15 无标记数据影响

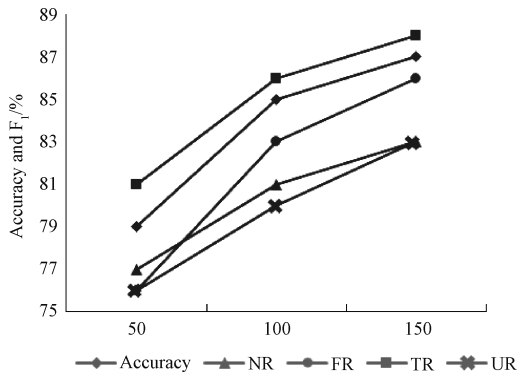

图 6 Twitter16 无标记数据影响

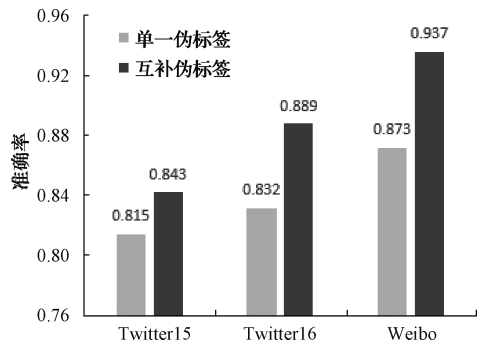

图 7 10%标记样本下不同伪标签的影响

由图 7～9 可知，在不同的标记数据和不同的数据集下，使用互补伪标签的准确率均高于传统单一伪标签。表明互补伪标签比单一伪标签可以生成质量更高的伪标签，减少模型的噪声，从而提高模型的检测效率。同时可以看出，在标记数据越少时，互补伪标签比单一伪标签的表现越好，表明互补伪标签可以好更利用大量的无标记数据，有效减少模型学习不均衡等问题。

2.6 消融实验

为了验证 MSGCN 方法中每个模块的有效性。以 Twitter15 数据集为例，从每个类别中分别选取 20%的标记样本，进行消融实验分析。从 MSGCN 中分别移除伪标

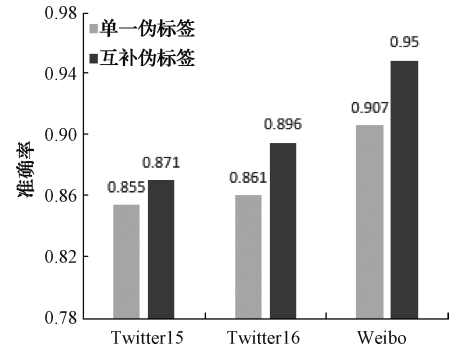

图 8 20%标记样本下不同伪标签的影响

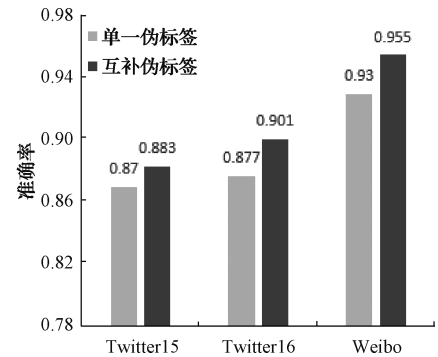

图 9 30%标记样本下不同伪标签的影响

签生成(PL)、去除检测模块中 Dropou 层，移除一致性训练损失(CT_Loss)和 DropEdge 来测试不同部分对实验结果的影响。消融实验结果如图 10 所示。

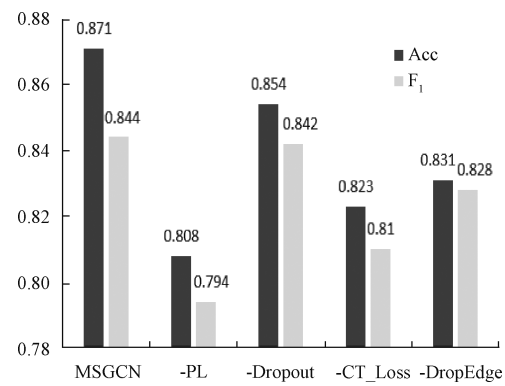

图 10 消融实验

由图 10 可以看出，移除不同的模块后，准确率和 F_1 值都会有不同程度的下降，移除伪标签生成、去除检测模块中 Dropout 层、移除一致性训练损失、去除 DropEdge 后准确率分别下降 6.3%、2.7%、4.8%和 4.0%。移除伪标签生成准确率下降较大，表明通过正负伪标签方法，产生了较多的有效样本。一致性训练对无标签样本具有较好的利用率，并且更接近无标记数据的准确标签，DropEdge 对样本进行扰动显著提升了模型的泛化能力。推文的评论与转发等有向信息可以为谣言事件的监测提供指向性

信息，Dropout 层可以使模型更好地学习谣言的多层次结构特征，增强模型的鲁棒性。

3 结 论

由于网络谣言传播迅速，及时遏制谣言传播至关重要，然而很少有数据样本可以在短时间内被标记使用，使监督算法存在一定滞后性。因此本文提出了一种多层次结构和半监督学习的谣言检测模型，利用一致性损失和互补伪标签充分利用大量无标记数据提升模型的泛化能力，使用多层次结构在图卷积网络中有效提取自顶向下和自底向上两个传播方向的特征和全局结构特征。在 3 个公开的数据集上的实验表明，MSGCN 可以利用少量有标记数据便可获得满意的准确率，利用大量无标记数据后同其他方法相比具有显著的提升。在未来的工作中，将研究数据不均衡问题和深入探究谣言的传播规律。

参 考 文 献

[1] SHU K, SLIVA A, WANG S, et al. Fake news detection on social media: A data mining perspective[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36.

[2] POTTHAST M, KIESEL J, REINARTZ K, et al. A stylometric inquiry into hyperpartisan and fake news[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 1: 231-240.

[3] 高玉君，梁刚，蒋方婷，等. 社会网络谣言检测综述[J]. 电子学报，2020，48(7)：1421-1435.

[4] 张少钦，杜圣东，张晓博，等. 融合多模态信息的社交网络谣言检测方法[J]. 计算机科学，2021，48(5)：117-123.

[5] 陈志毅，隋杰. 基于 DeepFM 和卷积神经网络的集成式多模态谣言检测方法[J]. 计算机科学，2022，49(1)：101-107.

[6] 葛晓义，张明书，魏彬，等. 基于双重情感感知的可解释谣言检测[J]. 中文信息学报，2022，36(9)：129-138.

[7] 赵志杰，张艳艳，毛翔宇. 基于改进 Adam 优化算法的中文短文本分类方法[J]. 电子测量技术，2022，45(23)：132-138.

[8] 梁欣怡，行鸿彦，侯天浩. 基于自监督特征增强的 CNN-BiLSTM 网络入侵检测方法[J]. 电子测量与仪器学报，2022，36(10)：65-73.

[9] 强子珊，顾益军. 基于多模态异质图的社交媒体谣言检测模型[J]. 数据分析与知识发现，2023，7(11)：68-78.

[10] 冯茹嘉，张海军，潘伟民. 基于预训练语言模型的早期微博谣言检测[J]. 计算机与数字工程，2023，51(5)：1075-1080，1184.

[11] MA J, GAO W, WONG K F. Rumor detection on twitter with tree-structured recursive neural networks[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1980-1989.

[12] 李奥，但志平，董方敏，等. 基于改进生成对抗网络的谣言检测方法[J]. 中文信息学报，2020，34(9)：78-88.

[13] BIAN T, XIAO X, XU T Y, et al. Rumor detection on social media with bi-directional graph convolutional networks [C]. Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 549-556.

[14] 谢欣彤，胡悦阳，刘譞哲，等. 传播用户代表性特征学习的谣言检测方法[J]. 计算机科学与探索，2022，16(6)：1334-1342.

[15] 盛晓辉，沈海龙. 基于数据增强和相似伪标签的半监督文本分类算法[J]. 计算机应用研究，2023，40(4)：1019-1023，1051.

[16] 李云红，魏妮娜，张晓丹. 基于多方向 Gabor 滤波器的图像分割[J]. 国外电子测量技术，2017，36(3)：20-23.

[17] 孙晶晶，静大海. 基于神经网络复杂背景下车牌识别系统的研究[J]. 国外电子测量技术，2017，36(8)：22-25.

[18] 郭秋实，李晨曦，刘金硕. 引入知识表示的图卷积网络谣言检测方法[J]. 计算机应用研究，2022，39(7)：2032-2036.

[19] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(6): 1929-1958.

[20] LIU X, NOURBAKHSH A, LI Q, et al. Real-time rumor debunking on twitter[C]. Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015: 1867-1870.

[21] MA J, GAO W, WONG K F. Detect rumors in microblog posts using propagation structure via kernel learning[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 708-717.

[22] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21: 1-67.

[23] KHOO L, CHIEU H L, QIAN Z, et al. Interpretable rumor detection in microblogs by attending to user interactions [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[24] WEI L, HU D, ZHOU W, et al. Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection[C]. ACL-IJCNLP, 2021.

[25] 徐建民,孙朋,吴树芳.传播路径树核学习的微博谣言检测方法[J].计算机科学,2022,49(6):342-349.

作 者 简 介

张岩珂,硕士研究生,主要研究方向为自然语言处理。
E-mail:ctguzyk@163.com

但志平(通信作者),博士,教授,主要研究方向为自然语言处理和计算机视觉。
E-mail:zpdan@ctgu.edu.cn

董方敏,博士,教授,博士生导师,主要研究方向为计算机图形图像处理和智能信息处理。