第 59 卷　第 5 期
2023 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 59, No. 5
Sept., 2023

DOI:10.13232/j.cnki.jnju.2023.05.010

基于图自监督对比学习的社交媒体谣言检测

乔禹涵[1,2],贾彩燕[1,2*]

(1.北京交通大学计算机与信息技术学院,北京,100044;
2.交通数据分析与挖掘北京市重点实验室,北京交通大学,北京,100044)

摘　要:网络社交媒体的快速发展提供了便捷的信息获取方式,但也滋生了谣言和虚假新闻,现有的谣言检测模型在有标注数据充足时能有效解决分类问题,然而谣言可用的标注数据有限,各种针对谣言特点精心设计的模型倾向于过拟合,同时,现有模型的鲁棒性不足,谣言传播者恶意破坏谣言传播结构会使模型出现分类错误.针对以上问题,采用自监督的图对比学习方法,对原始谣言传播图进行不同方式的数据增强来模拟对原图的扰动,建立自监督对比学习任务,使图编码器捕获谣言更趋本质的特征,缓解了过拟合,提高了模型的鲁棒性与泛化性能.在来源于主流社交媒体平台的三个公开数据集Twitter15,Twitter16和PHEME上进行了对比实验,实验结果显示,提出的模型的准确率比基准模型分别提高3.4%,1.8%和1.2%,证实了图自监督对比学习方法在谣言检测任务上的有效性.

关键词:谣言检测,自监督学习,对比学习,图表示学习

中图分类号:TP389.1　　**文献标志码:**A

Rumor detection on social media based on graph contrastive self-supervised learning

Qiao Yuhan[1,2], Jia Caiyan[1,2*]

(1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China;
2. Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, 100044, China)

Abstract: The rapid development of social media provides a convenient way to obtain information, meanwhile it helps the spread of rumors. Generally, with enough labeled data, existing rumor detection models can effectively solve rumor classification problems. However, due to limited labeled data of rumors, previous methods carefully designed for the characteristics of rumors tend to over-fit. Besides, existing rumor detection models are not robust enough. To solve the above problems, the graph contrastive self-supervised learning approach is adopted. A contrastive loss is defined to make graph encoders capture more essential and intrinsic features of rumors, alleviating the over-fitting and improving the robustness and generalization of the model. Experiments on three public datasets Twitter15, Twitter16 and PHEME has enhanced the accuracy of 3.4%, 1.8% and 1.2% respectively compared with the baseline, confirming the effectiveness of the proposed method.

Key words: rumor detection, self-supervised learning, contrastive learning, graph representation learning

国内外社交媒体平台已成为大众获取信息的主要渠道,然而,便捷的信息获取方式也为虚假信息的传播提供了有利条件.谣言的传播会损害社会安定及公众利益,因此高效准确地进行谣言检

基金项目:中央高校基本科研业务费(2019JBZ110)
收稿日期:2023－07－17
*通讯联系人,E-mail:cyjia@bjtu.edu.cn

测至关重要. 社会心理学文献[1]将谣言定义为一个广泛传播的未经证实或故意捏造的事件,谣言检测的目标是对未经证实事件的真假进行判断. 谣言检测的相关研究已从传统的基于特征工程的方法演变为深度学习方法. 考虑谣言传播的拓扑结构,近年来基于谣言传播结构的检测方法不断出现. Ma et al[2]首次利用谣言的传播结构信息,使用递归神经网络来捕获谣言传播的结构特征. Bian et al[3]在此基础上开创性地将谣言检测建模为图的分类问题,首次将图神经网络(Graph Neural Networks,GNN)应用于谣言检测,借助图神经网络强大的图表示学习能力来捕获谣言传播图的全局特征. 随后,结合谣言传播结构的基于图表示学习的各种谣言检测方法开始涌现.

通常,在有标注数据充足的情况下,深度学习模型能有效地解决分类问题,各种针对谣言特点精心设计的检测模型也取得了良好的效果. 但由于对谣言的标注耗时耗力,有标注谣言数据难以大量获得,现实中的有标注谣言数据极为有限,常用的公开数据集(Twitter15,Twitter16,PHEME)样本数量较少,针对谣言特点精心设计的方法存在过拟合风险. 同时,现有模型的鲁棒性不足,如图 1 所示,谣言传播者恶意破坏谣言传播结构,容易使模型分类出现错误.

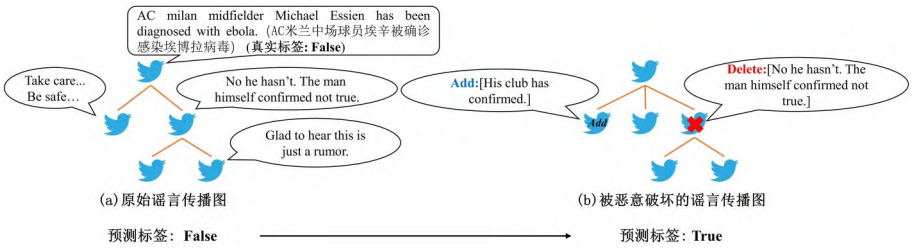

图 1 破坏谣言传播结构致使检测结果发生错误的实例

Fig. 1 An instance of rumor detection model making mistakes caused by perturbing the rumor propagation structures

自监督对比学习方法不利用额外标注信息,通过将数据分别与正例样本和负例样本在特征空间进行对比来得到更本质的特征表示,但目前在谣言检测领域对其的应用依旧匮乏. 本文将谣言检测视为图结构数据的分类问题,建立图自监督对比学习的辅助任务. 结合谣言特点提出三种图的扰动方式,将两个经过数据增强(可视为噪声扰动)的谣言传播图输入图编码器得到高层图表示,再通过判断两个扰动图是否来自同一原始图来建立自监督对比损失,将有监督任务和自监督对比任务联合训练,使图编码器捕获谣言更趋向本质的特征,缓解过拟合的负面影响,提高模型的泛化性能与鲁棒性.

1 相关工作

1.1 谣言检测相关工作 现有的谣言检测方法分三种:(1)基于特征工程的传统方法;(2)深度学习方法;(3)基于谣言传播结构的方法. 早期的谣言检测研究[4−6]根据谣言帖子的文本内容、用户资料、传播模式等来设计人工特征,这类基于特征工程的方法费时费力,提取的特征针对性强,泛化能力差. 近年来基于深度学习的检测方法不断涌现,如 Ma et al[7]和 Yu et al[8]分别采用循环神经网络(Recurrent Neural Networks,RNN)和卷积神经网络(Convolutional Neural Networks,CNN),从谣言帖子的时间序列中学习谣言的特征表示,Liu and Wu[9]同时利用 RNN 和 CNN 根据时间序列提取用户特征. 然而,这些方法忽略了谣言传播的拓扑结构. 为了利用谣言的传播结构信息,Ma et al[2]基于谣言双向传播树,建立递归神经网络,同时从帖子文本内容和传播结构两方面学习谣言特征表示. Khoo et al[10]利用 Transformer[11]架构建模帖子长距离之间的联系,并在其中融入传播树的结构信息. Bian et al[3]利用谣言传播图结构,设计了双向图卷积神经网络,借助图卷积网络强大的图表示学习能力来获取谣言全局结构特征. Wei et al[12]提出谣言传播的不确定性,对图卷积网络中的邻接矩阵进行动态更新. Lin et al[13]将谣言传播图作为无向图,采用层次化的注意力机制网络,充分利用了源帖子的信息.

1.2　图自监督对比学习相关工作　自监督学习的相关研究可分为对比式模型和生成式模型. 对比学习是一种对比式模型，首先兴起于视觉领域. Chen et al[14]的SimCLR利用对比学习提高视觉表示的质量. He et al[15]的Momentum Contrast方法利用Memory Bank存储负样本，大大增加了负样本的数量，缓解了显存不足的问题. Hjelm et al[16]提出Deep Infomax(DIM)来最大化一张图片的局部和全局上下文的互信息. 随后，对比学习开始在图结构数据上被大量应用. Veličković et al[17]提出Deep Graph Infomax (DGI)，将DIM方法拓展应用到图数据，最大化图级表示与节点表示的互信息. Hassani and Khasahmadi[18]通过建立多视角对比来最大化不同视图的互信息. Zhu et al[19]通过节点之间的对比来构建对比学习的正负样本. You et al[20]利用数据增强后的图级表示构建对比损失. 自监督对比学习任务的建立，使图编码器能捕获图更本质的高层特征.

在谣言检测领域，使用图自监督学习方法的研究还极其有限. Zhang et al[21]利用神经主题模型W-LDA，以Wasserstein自编码器获取谣言传播路径中对事件不敏感的主题模式，并以此重构谣言回复路径的词频. He et al[22]对数据增强后的帖子节点表示和原谣言图表示进行互信息最大化，使用预训练后微调的方法得到了更鲁棒的谣言表示. 然而，谣言传播图中的帖子节点较多，计算对比损失需要较大的计算量，使对比学习不高效. Sun et al[23]使用有监督的对比学习方法，利用谣言的类别标签信息，使同类样本的图表示在对比空间拉近，不同类样本的图表示远离，提高了谣言图特征表示的质量，并利用对抗学习提高了模型的鲁棒性，然而因其依赖标签信息，仍存在过拟合的风险. 为了减少对标签信息的依赖，缓解过拟合问题并提高模型的泛化能力，本文使用自监督的图对比学习方法，同时，为了进一步使对比学习更加高效，减少计算量，采用图级表示的实例之间的对比学习，并采用联合训练的方式，将自监督对比损失作为有监督分类损失的正则项，缓解了有标注数据匮乏造成的过拟合问题，提升了模型的泛化性能与鲁棒性.

2　问题描述

将谣言定义为一组谣言事件(Rumor Events)的集合$C=\{C_1, C_2, \cdots, C_n\}$，$C_i$表示其中第$i$个谣言事件，$n$表示所有谣言事件的数量. $C_i=\{r_i, x_1^i, x_2^i, \cdots, x_{m-1}^i, G_i\}$，$r_i$表示第$i$个谣言的源帖子(Source Post)，$x_j^i$表示第$j$个回复帖子，$m$表示第$i$个谣言所有帖子的数量. 虽然所有回复帖子以序列顺序排列，但基于帖子之间的回复关系使整个谣言事件可以建立为一个带有传播关系的谣言传播图. 用$G_i=(V_i, E_i)$表示第i个事件的谣言传播图，V_i表示以源帖子r_i为根节点的所有帖子节点的集合，E_i表示所有边的集合. 如果x_2^i是对x_1^i的回复帖子，则存在一个直接的连边$x_1^i \rightarrow x_2^i$. $\boldsymbol{X}\in\mathbb{R}^{m\times d}$，$\boldsymbol{A}\in\{0,1\}^{m\times m}$分别表示谣言传播图的特征矩阵和邻接矩阵.

谣言检测任务的目标是学习一个分类器$f: C_i \rightarrow Y_i$，Y_i是谣言的类别标签. 常用数据集将谣言分为四类：Non-Rumor(非谣言)，False-Rumor(验证为假的谣言)，True-Rumor(验证为真的谣言)，Unverified-Rumor(未经验证的谣言).

3　基于图自监督对比学习的谣言检测方法RD-GCSL

3.1　RD-GCSL谣言检测模型　提出一个通用的谣言图自监督对比学习检测框架RD-GCSL(Rumor Detection with Graph Contrastive Self-Supervised Learning)，如图2所示，该框架由五个模块组成.(1)数据增强模块：扰动原始谣言传播图的结构，生成两个新的谣言传播图；(2)图编码器模块：基于GNN模型的图编码器对谣言传播图进行节点特征聚合与更新，获取谣言图级别的特征表示；(3)投影头：基于前馈神经网络的映射层，将图的特征表示映射到对比空间；(4)对比损失：利用数据增强后得到的图级表示构建正负样本对，建立自监督对比损失；(5)谣言分类器：将图级别表示输入全连接层，预测谣言类别标签.

3.2　数据增强　数据增强的目的是在不改变数据原始语义标签的条件下，对原数据进行一定程度的变换，生成新的可用数据. 谣言的传播结构

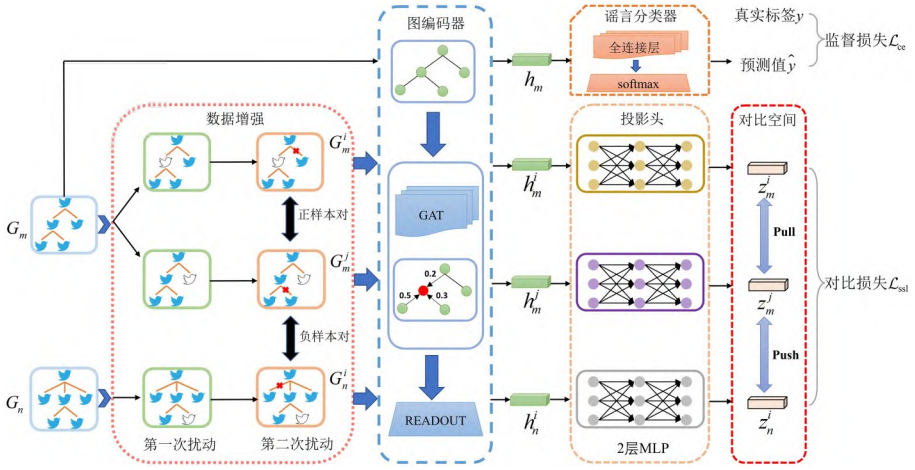

图 2 RD-GCSL 谣言检测模型图

Fig. 2 The architecture of RD-GCSL rumor detection model

通常具有不确定性[12],谣言制造者经常蓄意为虚假的事件发布支持的帖子或移除反对的帖子,此外,谣言传播图自身也包含一部分噪声信息. 为了使谣言检测模型具有更强的鲁棒性与泛化性能,对谣言事件的原始传播图 G 进行两次扰动,生成两个新的扰动图 $\hat{G}_i, \hat{G}_j$. 在之前图表示学习的相关工作[20]中,提出的基于图数据的各种数据增强策略在图分类任务中已被证明简单有效. 本文结合谣言传播的具体特点,设计了三种图级数据增强策略:移除边(Edge Removing,ER)、移除节点(Node Dropping,ND)、掩盖节点特征(Feature Masking,FM),如图 3 所示.

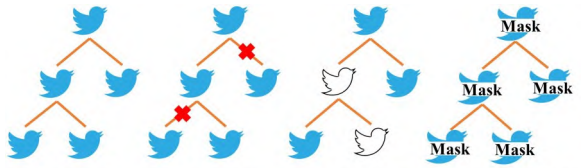

图 3 不同的图数据增强策略

Fig. 3 Various graph augmentation strategies

第一种策略是移除边. 社交网络中,谣言传播图的结构通常具有不确定性,回复帖子与被回复帖子不一定有直接的关联. 例如,一些社交网络用户没有遵循严格的回复关系,而是将回复帖子放置于谣言传播图的任意节点. 为了建模此种情况,使用随机丢弃谣言传播图连边的策略,具体方法:对邻接矩阵为 $\boldsymbol{A}$,特征矩阵为 $\boldsymbol{X}$ 的谣言传播图 $G=(V,E)$,以概率 r 对原始边的集合随机采样并丢弃.

第二种策略是移除节点. 实际的谣言传播过程中某些谣言制造者或恶意传播者蓄意为虚假信息回复支持帖子,或将提供证据戳穿虚假信息的回复帖子删除,以逃避谣言检测. 此外,社交网络中的用户也可随时将其回复的帖子删除,造成回复信息的缺失. 为了建模以上现象,提高谣言检测模型的鲁棒性,使用随机丢弃谣言传播图节点的策略,具体方法:以概率 r 对原始节点的集合随机采样,移除采样得到的节点和其对应的连边.

第三种策略是掩盖节点特征. 社交媒体平台的便利性使用户回复的文本信息不需要具有高度的规范性,常包含一定噪声或歧义,例如拼写错误、特殊字符、俚语等,造成原始的语义信息具有一定噪声或偏置. 为了建模此种现象,使用节点特征掩盖的策略,具体方法:以概率 r 对节点特征矩阵 $\boldsymbol{X}$ 的 d 个维度随机采样,将特征矩阵 $\boldsymbol{X}$ 中对应采样到的维度置 0.

数据增强是对比学习最关键的模块,样本对生成的策略会直接影响对比学习的质量. 对原始数据做的扰动过少会使对比学习任务过于简单,图编码器无法捕获谣言图的本质特征. 对原始数据做的扰动过多,可能造成有效信息丢失过多(详细验证见 4.3.2). 为了使对比学习的过程更加高效,每次对原始图的扰动都使用两种不同的数据增强方法的组合连续扰动.

3.3　图编码器　图编码器的作用是对输入图编码来获取图级别的特征表示，但本文提出的图自监督对比学习方法不依赖特定的图编码器. 考虑到谣言传播树的特点，对于一则谣言帖子，其所有回复帖子的重要程度并不相同. 图注意力网络(Graph Attention Networks，GAT)[24]在对待邻居节点(回复帖子)时，对邻居节点指派不同级别的权重进行聚合，而图卷积网络(Graph Convolutional Networks，GCN)[25]将所有邻居帖子节点同等对待. 因此，为了提高帖子表示的质量，减少噪声信息的权重，使用L层的GAT作为图编码器. $H^{(l)}=\left[h_r^{(l)}, h_{x_2}^{(l)}, \cdots, h_{x_m}^{(l)}\right]^T$代表帖子节点在第$l$层的隐层表示，其中$H^{(0)}=X$. 注意力系数的计算如下：

$$\alpha_{i,j}^{(l)}=\frac{\exp\left(\phi\left(a^T\left[W^{(l)}h_{x_i}^{(l)}\,\|\,W^{(l)}h_{x_j}^{(l)}\right]\right)\right)}{\sum_{j\in\mathcal{N}_i}\exp\left(\phi\left(a^T\left[W^{(l)}h_{x_i}^{(l)}\,\|\,W^{(l)}h_{x_j}^{(l)}\right]\right)\right)} \tag{1}$$

其中，$\alpha_{i,j}^{(l)}$代表帖子x_j对帖子x_i的重要性，a和$W^{(l)}$代表权重参数，$\|$代表拼接操作，$\mathcal{N}_i$代表x_i自身及其一阶邻居，ϕ代表激活函数(如LeakyReLU).

节点的聚合更新如下：

$$h_{x_i}^{(l+1)}=\mathrm{ReLU}\left(\sum_{j\in\mathcal{N}_i}\alpha_{i,j}^{(l)}W^{(l)}h_{x_j}^{(l)}\right) \tag{2}$$

对网络最后一层节点进行平均池化，获得整个图的全局表示：

$$h=meanpooling\left(H^{(L)}\right) \tag{3}$$

分别将无向的谣言事件原图G_m和两个扰动图$\hat{G}_m^i$，$\hat{G}_m^j$输入共享参数的图注意力网络，得到对应的图级表示，分别为$h_m\in\mathbb{R}^{d_1}$，$h_m^i\in\mathbb{R}^{d_1}$和$h_m^j\in\mathbb{R}^{d_1}$.

3.4　投影头　一个非线性变换$g(\cdot)$由两层感知机组成. 将图编码器输出的两个扰动图的图级表示h_m^i和h_m^j投影到隐空间得到$z_m^i\in\mathbb{R}^{d_2}$和$z_m^j\in\mathbb{R}^{d_2}$，进行对比损失的计算：

$$z_m^i=g\left(h_m^i\right), z_m^j=g\left(h_m^j\right) \tag{4}$$

3.5　对比损失　每轮训练中，每个minibatch中的N个图经过数据增强生成$2N$个扰动图，选取一个扰动图的表示z_m^i作为锚节点，与其来自同一个原图的扰动图的特征表示z_m^j为正样本，除此之外的$2N-2$个扰动图的特征都视为负样本. 通过最大化正样本的一致性(最小化负样本的一致性)，建立自监督对比学习损失：

$$\mathcal{L}_{\mathrm{ssl}}=-\lg\frac{\exp\left(z_m^i\cdot z_m^j/\tau\right)}{\exp\left(z_m^j\cdot z_m^j/\tau\right)+\sum_{\mathrm{neg}}\exp\left(z_m^j\cdot z_{\mathrm{neg}}/\tau\right)} \tag{5}$$

其中，τ表示温度系数，z_{neg}表示随机采样的负样本.

3.6　谣言分类器　将谣言原始图的图级表示h_m输入全连接层和一个softmax层：

$$\hat{y}=\mathrm{softmax}\left(W_c h_m+b_c\right) \tag{6}$$

其中，$\hat{y}\in\mathbb{R}^{1\times C}$是预测的谣言各类别的概率分布，$C$表示谣言类别的数量，$W_c$和$b_c$是可学习的参数矩阵.

利用数据真实标签信息，计算预测值和真实分布的交叉熵，得到有监督分类损失：

$$\mathcal{L}_{\mathrm{ce}}=-\sum_{i=1}^{n}y_i\lg\left(\hat{y}_i\right) \tag{7}$$

有监督分类损失和自监督对比学习损失相加作为总损失：

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{ce}}+\lambda\mathcal{L}_{\mathrm{ssl}} \tag{8}$$

其中，λ表示自监督损失的权重超参数.

4　实验分析

4.1　实验设置

4.1.1　数据集　使用来源于主流社交媒体平台的三个公开数据集Twitter15[26]，Twitter16[26]和PHEME[27]进行实验，每则谣言事件的标签都通过谣言揭穿网站(如snopes.com，Emergent.info等)来标定. 所有数据集包含四种类型的标签：Non-Rumor(非谣言)，False-Rumor(经验证真实值为假的谣言)，True-Rumor(经验证真实值为真的谣言)，Unverified-Rumor(未经验证的谣言). Twitter15，Twitter16两个数据集中谣言各类别的数量相对均衡，然而，现实中虚假谣言的数量远少于真实事件的数量，因此，实验另外选取了类别数量不平衡的数据集PHEME进行补充. 表1列出了所有数据集的详细统计信息.

4.1.2　评价指标和参数设置　与RvNN[2]，Bi-GCN[3]等方法的实验设置一致，所有数据集按照4∶1的比例划分为训练集和测试集，采用5折交叉验证，以不同的随机种子运行10次并汇报平均值. 采用与其他研究者相同的评价指标：准确率(Accuracy)和$F1$. 参数设置：谣言传播图初始节

表 1 数据集的统计信息

Table 1 Statistics of datasets

数据集	Twitter15	Twitter16	PHEME
谣言事件	1490	818	6425
非谣言(NR)	374	205	4023
验证为假的谣言(FR)	370	205	638
未验证的谣言(UR)	374	203	698
验证为真的谣言(TR)	372	205	1067
用户数	276663	173487	48843
帖子数	331612	204820	197852

点的文本特征采用5000维的TF-IDF特征,图神经网络中每个节点的隐层特征维度为64,图注意力网络的层数为2,dropout参数为0.5,batch size为256(Twitter16为128),学习率为0.0005,两次数据扰动的比率$r=\{0.1,0.2,0.3,0.4,0.5\}$,通过网格搜索选取最佳组合,自监督损失项权重$\lambda=1$,对比损失中温度系数$\tau=0.2$,采用Adam优化器更新参数.每次训练迭代200个epoches,验证集的*loss*在10个epoches之内不再下降时采取早停机制.

4.2 与主流模型的对比实验

4.2.1 对比模型

(1)RvNN[2]:是基于GRU单元和树结构递归神经网络的谣言检测方法.

(2)BiGCN[3]:是基于GCN的模型,利用谣言传播的有向图,分自上而下和自下而上两部分提取谣言的高层特征.

(3)UDGAT:是本文使用的图编码器,使用GAT并将谣言传播图作为无向图,其与BiGCN模型相比,大量减少了模型参数.

(4)ClaHi-GAT[13]:是基于GAT的模型,采用层次化的注意力机制来充分利用源帖子的信息.

(5)RDEA[22]:是基于GCN的对比学习方法,将帖子节点表示和原谣言图表示互信息最大化,使用预训练后微调的方法得到了更鲁棒的谣言表示.

(6)SRD-PSID[28]:是多视角的对比学习方法,利用两个编码器将传播路径与源帖文本编码得到的两个表示作为两个不同视角进行对比.

(7)RD-GCSL模型:是本文提出的自监督图对比学习谣言检测方法,以UDGAT作为图编码器,对数据增强的两个谣言图进行图级别的对比,建立自监督辅助任务,与有监督分类任务联合训练.

4.2.2 实验结果与分析

表2~4展示了各谣言检测模型在Twitter15,Twitter16和PHEME三个数据集上的性能,表中黑体字表示最优的性能.由表可见,在基准模型中,RvNN和BiGCN等深度学习模型通过捕获谣言的文本和结构信息,学习到了高层级的谣言特征,提升了谣言检测的效果.本文方法在之前研究的基础上,建立了新的自监督对比学习任务,使图编码器编码得到的谣言图表示具有谣言更本质的特征,缓解了因有标注数据少造成的过拟合问题,提高了模型的泛化性能与鲁棒性.提出的模型RD-GCSL在Twitter15,Twitter16和PHEME数据集上分别达到88.0%,88.9%,85.6%的准确率,与未使用对比学习的基模型UDGAT相比,分别提升3.4%,1.8%,1.2%,验证了自监督对比学习方法的有效性.

表 2 Twitter15数据集上的实验结果

Table 2 Experimental results on Twitter15 dataset

模型	准确率	*F*1			
		NR	FR	TR	UR
RvNN	0.723±0.8%	0.682	0.758	0.821	0.654
BiGCN	0.843±0.4%	0.788	0.860	0.895	0.808
UDGAT	0.846±0.2%	0.792	0.849	0.906	0.829
ClaHi-GAT	0.859±0.4%	0.831	0.864	0.901	0.834
RDEA	0.855±0.6%	0.831	0.857	0.903	0.816
RD-GCSL	**0.880±0.3%**	**0.851**	**0.886**	**0.926**	**0.852**

表 3 Twitter16数据集上的实验结果

Table 3 Experimental results on Twitter16 dataset

模型	准确率	*F*1			
		NR	FR	TR	UR
RvNN	0.737±0.9%	0.662	0.743	0.835	0.708
BiGCN	0.858±0.5%	0.767	0.854	0.925	0.867
UDGAT	0.871±0.3%	0.794	0.876	0.927	0.870
ClaHi-GAT	0.882±0.4%	0.827	**0.887**	0.936	0.874
RDEA	0.880±0.5%	0.823	0.878	0.937	0.875
RD-GCSL	**0.889±0.3%**	**0.833**	0.882	**0.949**	**0.886**

表4　PHEME数据集上的实验结果

Table 4　Experimental results on PHEME dataset

模型	准确率	*F*1			
		NR	FR	TR	UR
BiGCN	0.847±0.2%	0.910	0.634	0.655	0.500
UDGAT	0.844±0.2%	0.902	0.658	0.833	0.485
ClaHi-GAT	0.846±0.1%	0.896	0.670	0.623	0.515
SRD-PSID	0.838±0.3%	0.905	**0.774**	**0.734**	**0.604**
RD-GCSL	**0.856±0.1%**	**0.915**	0.669	0.607	0.530

为了进一步说明自监督对比学习方法能缓解标注数据不足带来的过拟合影响，仅使用少量样本(10%，20%，50%)进行训练．表5展示了少量样本训练的实验结果，表中“Δ”代表准确率的增益．由表可见，在有标注的训练数据有限时，提出的自监督对比学习模型RD-GCSL在所有数据集上的准确率和基准模型UDGAT相比，仍有明显提升，进一步验证了自监督对比学习方法的有效性．

表5　不同训练数据规模下的实验结果

Table 5　Experimental results with various scales of labeled training data

数据集	模型	10%		20%		50%		80%	
		准确率	Δ	准确率	Δ	准确率	Δ	准确率	Δ
Twitter15	UDGAT	0.608	—	0.684	—	0.769	—	0.846	—
	RD-GCSL	0.626	↑1.8%	0.702	↑1.8%	0.803	↑3.4%	0.880	↑3.4%
Twitter16	UDGAT	0.594	—	0.723	—	0.820	—	0.871	—
	RD-GCSL	0.626	↑3.2%	0.743	↑2.0%	0.838	↑1.8%	0.889	↑1.8%
PHEME	UDGAT	0.738	—	0.766	—	0.797	—	0.844	—
	RD-GCSL	0.745	↑0.7%	0.776	↑1.0%	0.807	↑1.0%	0.856	↑1.2%

4.3　消融实验

4.3.1　谣言图编码器模块的影响　本文提出的RD-GCSL不依赖特定的谣言图编码器，能作为一个通用的框架来提高现有谣言检测模型的效果．为了验证其对不同的谣言图编码器普遍有效，使用三种谣言图编码器UDGAT，BiGCN，ClaHi-GAT，结合本文的图自监督对比学习方法进行实验．用-GCSL代表提出的自监督对比学习的模型，表中“Δ”代表准确率的增益．

表6给出了三种不同的谣言图编码器结合提出的对比学习方法后在所有数据集上的准确率．由表可见，谣言图编码器结合提出的对比学习方法，使其性能获得了提升，证明本文提出的对比学习方法作为一个通用的框架，可以提升已有的谣言检测模型的效果．

4.3.2　数据增强模块的影响　数据增强作为对比学习最关键的模块，其生成的样本对将直接影响对比学习的质量．根据三种不同的图扰动方法，可以构建样本对多种扰动方式的组合．此外，数据扰动的比例r也将决定对比学习的质量．为了探究不同数据增强方法对自监督对比学习效果的影响，进行以下实验．

表6　对比学习结合不同图编码器的实验结果

Table 6　Experimental results of contrastive learning by various graph encoders

模型	Twitter15		Twitter16		PHEME	
	准确率	Δ	准确率	Δ	准确率	Δ
UDGAT	0.846	—	0.871	—	0.844	—
UDGAT-GCSL	0.880	↑3.4%	0.889	↑1.8%	0.856	↑1.2%
BiGCN	0.843	—	0.858	—	0.847	—
BiGCN-GCSL	0.881	↑3.8%	0.888	↑3.0%	0.850	↑0.3%
ClaHi-GAT	0.859	—	0.882	—	0.846	—
ClaHi-GAT-GCSL	0.872	↑1.3%	0.892	↑1.0%	0.852	↑0.6%

4.3.2.1 不同数据增强策略的影响 分别对原始图进行单种方法扰动(移除边(ER)、移除节点(ND)、掩盖节点属性(FM))、两种不同方法组合连续扰动、三种不同方法组合连续扰动生成扰动图. 每种方法的扰动比例从 $r=\{0.1,0.2,0.3,0.4,0.5\}$ 中选取最优参数.

表 7 展示了不同数据增强策略的影响,表中黑体字表示性能最优. 由表可见,不同的增强方法在不同的数据集上的效果不同,移除边略好于其他两种策略,采用两种不同方法连续扰动的策略效果略好于单种方法扰动和三种方法连续扰动的策略. 由此可以推断,对比学习样本对的生成不应过于简单,因为这会降低对比学习的质量,但也不应过于复杂,因为对原图进行过多扰动会造成有效信息的丢失.

表 7 数据增强策略的影响

Table 7 Experimental results with various data augmentation strategies

数据增强策略	准确率		
	Twitter15	Twitter16	PHEME
ND	0.871	0.883	0.851
ER	0.873	0.888	0.852
FM	0.869	0.887	0.855
ND+FM	**0.880**	0.888	**0.856**
ND+ER	0.875	**0.889**	0.853
ER+FM	0.873	0.887	0.855
ND+ER+FM	0.866	0.885	0.855

4.3.2.2 不同数据增强比例 r 的影响 为了研究扰动比例对图对比学习效果的影响,采用三种方法连续扰动的策略(ND+ER+FM),以不同的扰动比例 $\{0.1,0.2,\cdots,0.8,0.9\}$ 进行实验,实验结果如图 4 所示. 由图可见,扰动比例分别为 0.3,0.5,0.5 时,模型在 Twitter15,Twitter16,PHEME 三个数据集上表现最好. 随着扰动比例的增大,模型分类的准确率明显降低,说明对原图做过多的扰动会引入过多的噪声,丢失原图的有效信息,也说明建立更困难的对比学习任务不一定会提升对比学习的效果.

4.3.3 投影头模块的影响 为了验证模型中投影头模块的作用,进行了消融实验,实验结果如表 8 所示,表中 w/o PH(without projection head)代表去掉投影头模块的模型. 由表可见,不使用投影头时,对比学习模型的表现明显下降,在 Twitter15,Twitter16,PHEME 数据集上的准确率分别下降 3.1%,2.3%,3.9%. 说明由图编码器得到的图特征表示样本对要经过投影头的非线性变换,在变换后的隐空间中计算对比损失才能确保对比学习的质量,证明了投影头模块的重要性.

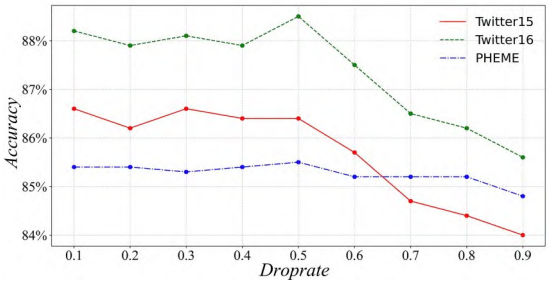

图 4 不同扰动比例的影响

Fig. 4 Effect of various perturbation ratios

表 8 投影头对模型的影响

Table 8 Effect of projection head

数据集	模型	准确率	$F1$			
			NR	FR	TR	UR
Twitter15	UDGAT-GCSL	0.880	0.848	0.880	0.923	0.833
	w/o PH	0.849	0.822	0.841	0.901	0.817
Twitter16	UDGAT-GCSL	0.889	0.833	0.882	0.949	0.886
	w/o PH	0.866	0.799	0.858	0.945	0.846
PHEME	UDGAT-GCSL	0.856	0.915	0.662	0.639	0.510
	w/o PH	0.817	0.899	0.555	0.565	0.430

4.3.4 泛化性能验证实验 为了验证提出的图自监督对比学习模型在鲁棒性、泛化性能上的提升以及对过拟合问题的缓解效果,设计了如下的实验. 对原始测试集中的谣言传播图进行两种不同类型的数据增强,将所得扰动图的类标签设置为其所对应原图的谣言类别标签. 表 9 展示了没有使用图自监督对比学习的基模型 UDGAT 和本文模型 RD-GCSL 在新构建的测试集上的效果,并与没有进行数据增强的原始数据集上的效果进行比较,表中"Δ"代表准确率的增益. 由表可见,对原始测试集进行扰动之后,所有模型的分类准确率均有所下降. 但本文模型 RD-GCSL 在扰动测试集上下降的精度明显小于没有使用自监督

表9　泛化性能的验证实验

Table 9　Experiment of generalization performance

数据集	模型	准确率		
		原始测试集	扰动测试集	Δ
Twitter15	UDGAT	0.846	0.823	↓2.3%
	RD-GCSL	0.880	0.871	↓0.9%
Twitter16	UDGAT	0.871	0.835	↓3.6%
	RD-GCSL	0.889	0.878	↓1.1%
PHEME	UDGAT	0.844	0.830	↓1.4%
	RD-GCSL	0.856	0.849	↓0.7%

方法的基模型UDGAT,证明RD-GCSL得益于自监督对比学习任务的构建,展示了较好的鲁棒性与泛化性能,缓解了过拟合问题.

5　结论

针对目前谣言有标注数据有限,现有的谣言检测模型存在过拟合与鲁棒性不足的问题,提出一种新的基于图自监督对比学习的谣言检测方法.建立图自监督对比学习任务,和有监督分类任务联合学习,使图编码器能捕获谣言更本质的图结构特征,缓解了有标注数据匮乏造成的过拟合问题,提升了模型的泛化性能与鲁棒性.在Twitter15,Twitter16和PHEME三个公开数据集上进行的实验中,本文提出的方法在使用全部有标注数据和仅使用部分有标注数据的条件下,均比基准方法取得了更高的准确率和$F1$,验证了本文方法在谣言检测问题上的有效性.通过消融实验,探究了图编码器模块、数据增强模块和投影头模块对模型的影响,并验证了提出的自监督对比学习方法不依赖于特定的谣言图编码器,能作为一个通用框架提高现有谣言检测模型的性能.

参考文献

[1] DiFonzo N, Bordia P. Rumor, gossip and urban legends. Diogenes, 2007, 54(1): 19－35.

[2] Ma J, Gao W, Wong K F. Rumor detection on twitter with tree - structured recursive neural networks// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Volume 1. Long Papers. Melbourne, Australia: Association for Computational Linguistics, 2018: 1980－1989.

[3] Bian T, Xiao X, Xu T Y, et al. Rumor detection on social media with bi-directional graph convolutional networks. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 549－556.

[4] Castillo C, Mendoza M, Poblete B. Information credibility on twitter// Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India: Association for Computing Machinery, 2011: 675－684.

[5] Yang F, Liu Y, Yu X H, et al. Automatic detection of rumor on Sina Weibo// Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. Beijing, China: Association for Computing Machinery, 2012, Article No. 13.

[6] Liu X M, Nourbakhsh A, Li Q Z, et al. Real-time rumor debunking on twitter// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia: Association for Computing Machinery, 2015: 1867－1870.

[7] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks// Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, NY, USA: AAAI Press, 2016: 3818－3824.

[8] Yu F, Liu Q, Wu S, et al. A convolutional approach for misinformation identification// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017: 3901－3907.

[9] Liu Y, Wu Y F B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks// Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, LA, USA: AAAI Press, 2018: 354－361.

[10] Khoo L M S, Chieu H L, Qian Z, et al. Interpretable rumor detection in microblogs by attending to user interactions. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8783－8790.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need// Proceedings of the 31st International

Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017: 6000－6010.

[12] Wei L W, Hu D, Zhou W, et al. Towards propagation uncertainty: Edge - enhanced Bayesian graph convolutional networks for rumor detection// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Volume 1. Long Papers. Bangkok, Thailand: Association for Computational Linguistics, 2021: 3845－3854.

[13] Lin H Z, Ma J, Cheng M F, et al. Rumor detection on twitter with claim - guided hierarchical graph attention networks//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 10035－10047.

[14] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: JMLR.org, 2020: 1597－1607.

[15] He K M, Fan H Q, Wu Y X, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 9726－9735.

[16] Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization. 2019, arXiv: 1808.06670.

[17] Veličković P, Fedus W, Hamilton W L, et al. Deep graph infomax. 2018, arXiv:1809.10341.

[18] Hassani K, Khasahmadi A H. Contrastive multi-view representation learning on graphs. 2020, arXiv: 2006.05582.

[19] Zhu Y Q, Xu Y C, Yu F, et al. Deep graph contrastive representation learning. 2020, arXiv: 2006.04131.

[20] You Y N, Chen T L, Sui Y D, et al. Graph contrastive learning with augmentations// Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020: 5812－5823.

[21] Zhang P F, Ran H Y, Jia C Y, et al. A lightweight propagation path aggregating network with neural topic model for rumor detection. Neurocomputing, 2021(458): 468－477.

[22] He Z Y, Li C, Zhou F, et al. Rumor detection on social media with event augmentations//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event, Canada: Association for Computing Machinery, 2021: 2020－2024.

[23] Sun T N, Qian Z, Dong S J, et al. Rumor detection on social media with graph adversarial contrastive learning//Proceedings of the ACM Web Conference 2022. Lyon, France: Association for Computing Machinery, 2022: 2789－2797.

[24] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2018, arXiv:1710.10903.

[25] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2017, arXiv: 1609.02907.

[26] Ma J, Gao W, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Volume 1. Long Papers. Vancouver, Canada: Association for Computational Linguistics, 2017: 708－717.

[27] Zubiaga A, Liakata M, Procter R, et al. Analysing how people orient to and spread rumours in social media by looking at conversational threads. 2016, arXiv:1511.07487.

[28] Gao Y, Wang X, He X N, et al. Rumor detection with self - supervised learning on texts and social graph. 2022, arXiv:2204.08838.

（责任编辑 杨可盛）