计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330506
61(7): 1629−1642, 2024

基于信息瓶颈理论的鲁棒少标签虚假信息检测

王吉宏[1] 赵书庆[1] 罗敏楠[1] 刘 欢[1] 赵 翔[2] 郑庆华[1]
[1](西安交通大学计算机科学与技术学院 西安 710049)
[2](国防科技大学大数据与决策实验室 长沙 410003)
(wang1946456505@stu.xjtu.edu.cn)

Robust Few-Label Misinformation Detection Based on Information Bottleneck Theory

Wang Jihong[1], Zhao Shuqing[1], Luo Minnan[1], Liu Huan[1], Zhao Xiang[2], and Zheng Qinghua[1]
[1] (*School of Computer Science and Technology*, *Xi'an Jiaotong University*, *Xi'an* 710049)
[2] (*Laboratory for Big Data and Decision*, *National University of Defense Technology*, *Changsha* 410003)

Abstract Misinformation detection is crucial for the social stability. Researches show that there are substantial distinctions between misinformation and real information in terms of information content and propagation structure. Consequently, recent researchers mainly focus on improving the accuracy of misinformation detection by jointly considering the information content and propagation structure. However, these methods can be infeasible in practice since they highly rely on manual label information. The manual labels can be expensive since they require extensive comparison with official reports and other evidence. Moreover, the spreaders of misinformation can adversarially manipulate the information content and propagation structure by controlling reviews and other methods. Such behaviors may exacerbate the challenges of misinformation detection. To address these problems, we propose a robust few-label misinformation detection method based on information bottleneck theory. Specifically, to mitigate the dependence on labeled data, we propose to integrate the unlabeled sample information by employing the mutual information maximization technique. Furthermore, to improve the robustness of our method against the adversarial manipulation of misinformation spreaders, we employ the adversarial training strategy to simulate the behaviors of the spreaders and propose to learn robust representations based on the information bottleneck theory. The learned representations can effectively embed the essential information in the misinformation while discarding the adversarial information involved by the spreaders. Empirical evaluations validate the effectiveness of the proposed approach, demonstrating superior performance compared with benchmark methods in terms of few-label detection and robustness.

Key words misinformation detection; graph neural network; mutual information; graph representation learning; robust representation learning; few-label learning

收稿日期: 2023-06-15; **修回日期**: 2023-12-04

基金项目: 国家重点研发计划项目(2022YFB3102600); 国家自然科学基金项目(62192781, 62272374, 62202367, 62250009, 62137002, 61937001); 国家自然科学基金创新研究群体(61721002); 教育部创新研究团队(IRT_17R86); 中国工程科学技术知识中心项目及中国工程院项目; 王宽诚教育基金项目

This work was supported by the National Key Research and Development Program of China (2022YFB3102600), the National Natural Science Foundation of China (62192781, 62272374, 62202367, 62250009, 62137002, 61937001), the Innovative Research Group of the National Natural Science Foundation of China (61721002), the Innovation Research Team of Ministry of Education (IRT_17R86), the Project of China Knowledge Center for Engineering Science and Technology and Project of Chinese Academy of Engineering, the Project of K. C. Wong Education.

通信作者: 罗敏楠(minnluo@xjtu.edu.cn)

摘 要 虚假信息检测对于维护网络舆情安全具有重要意义.研究表明,虚假信息在信息内容和传播结构上较真实信息具有显著不同.为此,近年来研究致力于挖掘信息内容和信息传播结构,提升虚假信息检测的精准性.然而,现实场景中虚假信息的标注往往需要大量地与官方报道等比照分析,代价较为昂贵,现有方法对标注信息的过分依赖限制了其实际应用.此外,虚假信息传播者可通过在评论区控评等手段恶意操纵虚假信息的传播,增加了虚假信息检测的难度.为此,基于信息瓶颈理论提出一种鲁棒少标签虚假信息检测方法,通过互信息最大化技术融合无标注样本信息,克服虚假信息检测对标签的过分依赖问题;并通过对抗训练的策略模拟虚假信息传播者的恶意操纵行为,基于信息瓶颈理论学习鲁棒的虚假信息表征,在高质量表征虚假信息的同时消除恶意操纵行为的影响.实验表明,该方法在少标签识别和鲁棒性2个方面均取得了优于基准方法的效果.

关键词 虚假信息检测;图神经网络;互信息;图表示学习;鲁棒表示学习;少标签学习

中图法分类号 TP391

随着移动互联网技术的蓬勃发展,微信、微博、推特等社交媒体成为网络用户获取信息的重要媒介.社交媒体实时性、开放性和便捷性的特点加速了网络信息的传播.与此同时,由于其低门槛、弱审核的特点,社交媒体成为了滋养虚假有害信息的温床.肆意泛滥传播的虚假信息不仅会损害官方媒体的权威性和公信力,更可能导致潜在的经济、政治危害[1-3].为此,针对虚假信息进行检测识别,是净化互联网环境、保障公共网络信息安全的重要手段.

虚假信息检测近年来受到研究者的普遍关注.研究表明,人为杜撰的虚假信息在信息内容和传播结构上较真实信息具有显著的差异性[4-6].一方面,虚假信息往往呈现出更加强烈的主观性和情感煽动性[4];另一方面虚假信息的传播往往更加广泛,且传播过程中存在更加剧烈的争论探讨[6].针对虚假信息的以上特点,近年来研究者提出基于深度神经网络的虚假信息检测方法[7-13],以捕捉虚假信息在信息内容和传播途径上的特征分布.基于信息内容的方法[7-9]往往采用循环神经网络(recurrent neural network, RNN)或卷积神经网络(convolutional neural network, CNN)表征信息文本和图片特征.然而此类方法主要聚焦于虚假信息内容本身,忽略了其在社交网络中的传播拓扑结构.为此,近年来研究者[10-11, 13]引入图神经网络(graph neural network, GNN)表征分析虚假信息的传播拓扑,以更加精准地进行虚假信息的检测和识别.

然而,现有大多数工作主要集中于虚假信息检测的精准性,忽视了实际场景中虚假信息的标签稀缺性和恶意操纵性特点.具体而言,实际场景中虚假信息的标注需要与大量的官方报道、政府公告等进行比照分析,人工标注代价较为昂贵.现有虚假信息检测方法往往采用有监督学习策略,对标签信息具有较高的依赖性,因而限制了其在实际场景中的应用.此外,虚假信息在传播过程中具有明显的恶意操纵性.虚假信息传播者往往通过在相关评论区进行控评,删除反对性评论以及增加诱导性评论,误导社交用户,提升其可信性和说服力[12].此类恶意操纵行为为虚假信息的信息内容和传播拓扑添加了人为扰动,增加了检测的难度.

考虑到现有方法的以上缺陷,本文提出基于信息瓶颈理论[14-15]的鲁棒少标签虚假信息检测方法.虚假信息检测可抽象为图分类问题,将虚假信息的信息内容和传播拓扑分别视为属性图的节点特征和拓扑结构,通过图神经网络在少标签场景下学习鲁棒高效的图表征,并基于图表征引入少量标注信息进行虚假信息检测.具体而言,为克服现有方法对标签信息的高度依赖性问题,可通过互信息最大化(mutual information maximization, Infomax)技术,同时利用无标注样本信息与少量标注样本,学习高质量虚假信息表征,实现少标签场景下的虚假信息检测;为克服虚假信息传播过程中存在的恶意操纵性问题,可通过对抗训练的方式模拟虚假信息传播者的恶意操纵行为,并基于信息瓶颈理论消除对抗性信息,学习鲁棒的无监督虚假信息表征.

本文的主要贡献包括3个方面:

1)提出同时利用无监督样本信息与少量标签信息,在少标签场景下实现高效的虚假信息检测,克服了虚假信息标注代价高的难题;

2)基于信息瓶颈理论提出了鲁棒无监督虚假信息表示学习方法,克服了虚假信息传播过程中存在的恶意操纵性问题;

3)在真实数据集上对所提出的方法进行验证.

实验表明，所提方法在少标签识别和鲁棒性方面均优于现有基准方法.

1 相关工作

本节针对现有虚假信息检测工作和信息瓶颈理论进行简要介绍. 根据检测方法所采用的信息类型，现有虚假信息检测工作大致可分为基于信息内容的方法和基于图的方法.

1.1 基于信息内容的方法

基于信息内容的方法通过挖掘信息内容本身训练虚假信息分类器，针对虚假信息进行分类. 早期方法采用人工提取特征的策略[16-19]. 例如，Castillo 等人[20] 采用人工提取的语言特征基于决策树模型进行虚假信息的检测. Yang 等人[18] 通过分析信息内容和用户的统计学信息，使用基于径向基函数(radial basis function, RBF)核的支持向量机(support vector machine, SVM)检测虚假信息. 此类方法依赖于从信息内容和用户信息中人工提取特征，需要大量的领域知识和人工标注，并且依赖于特定的数据集，因此在实际应用中具有较大的局限性.

随着深度学习的发展，研究者提出利用深度神经网络自主挖掘信息内容[21-23]，进而进行虚假信息检测. 此类方法无需人工提取特征，因而具备更强的泛化性能. 例如 Ma 等人[8] 利用长短期记忆(long short term memory, LSTM)网络针对虚假信息文本进行挖掘，提取隐藏特征. RvNN[24] 基于递归神经网络(recursive neural networks)挖掘虚假信息. Yu 等人[25] 采用 CNN 提取序列特征，形成关键特征之间的高层交互. Ma 等人[26] 采用对抗生成网络(generative adversarial network, GAN)以捕捉文本中存在的低频非平凡模式，增强模型的鲁棒性. 此外，现有一些工作除直接分析虚假信息文本外，还采用了用户立场、用户可信度以及多模态信息提升虚假信息检测的性能[27-29]. 此类方法虽然利用深度神经网络取得了较好的虚假信息检测效果，并且无需昂贵的人工特征，但其忽略了虚假信息在社交网络中传播形成的拓扑结构，无法捕捉社交拓扑信息.

1.2 基于图的方法

除却信息内容本身，虚假信息的传播拓扑对于虚假信息检测也有重要作用. 因此，基于图的方法近年来受到研究者的广泛关注[10, 12-13, 24]. 例如，Ma 等人[24] 通过构建自底向上和由上向下的树状神经网络捕捉虚假信息的传播拓扑，在推特上进行虚假信息检测；Bian 等人[10] 提出一种双向图神经网络，用于发掘虚假信息自底向上和由上向下的传播过程；Yuan 等人[13] 通过异质信息网络，联合挖掘社交网络中的用户、信息和评论组成的异质图，实现融合局部和全局特征的虚假信息检测；Li 等人[30] 基于 Transformer 模型融合社交网络中的异质信息，实现高效的虚假信息检测；InfOP[31] 融合了用户信息与虚假信息文本，通过图结构学习推断信息传播过程中的隐式关联关系，迭代地针对信息传播结构和虚假信息分类器进行训练.

现有大多工作主要集中于虚假信息检测的精准性，忽视了实际场景中虚假信息标注代价高且容易被虚假信息传播者恶意操纵的特点. 目前仅有 Sun 等人[12] 在研究中关注了虚假信息检测的鲁棒性. 该研究提出 GACL 方法，通过基于标签信息的对比学习策略挖掘虚假信息特征，并引入对抗性特征变换(adversarial feature transformation, AFT)模块提升检测的鲁棒性. 然而其所采用的对比学习策略依赖于标注信息，在少标签场景下性能有所欠缺. 此外，GACL 引入的 AFT 模块仅在隐层网络中添加扰动，其分布较真实场景下的恶意操纵行为有所不同，此鲁棒性策略在真实场景中可能面临分布差异性(distribution shift)问题.

1.3 信息瓶颈理论

信息瓶颈理论近年来受到研究者的广泛关注，被广泛应用于计算机视觉[32-33]、自然语言处理[34-35]、图学习[36-37] 等各个领域. 例如，在计算机视觉领域，Ma 等人[32] 采用信息瓶颈理论实现从图像到音频的转化，他们提出的多模态信息瓶颈方法克服了传统生成模型对跨模态数据对的依赖. Wang 等人[33] 通过信息瓶颈理论提出一种可解释的多任务信息瓶颈网络，通过最大化潜在表示和类标签之间的互信息，同时最小化潜在表示和输入共享的信息，实现可解释的乳腺图像诊断. 在自然语言处理领域，Zhang 等人[34] 通过信息瓶颈理论消除文本中的任务无关信息，提升文本任务的鲁棒性. Mahabadi 等人[35] 基于信息瓶颈理论实现对大规模语言模型的微调，消除其在特定任务上的无关信息，在低资源文本任务上取得了优越的性能. 在图学习领域，Sun 等人[37] 提出基于信息瓶颈理论的图结构学习算法，学习图中缺失的隐式结构信息，提升了多种图神经网络的性能. Wang 等人[36] 提出基于信息瓶颈理论的图神经网络事后解释方法，为图神经的表征提供归因依据，剖析其表征逻辑. 尽管信息瓶颈理论在深度学习领域受到了广泛的关注，其在虚假信息检测领域的应用仍缺乏探索.

2 符号定义及相关理论

在本节中，首先针对所用符号进行简要说明，而后针对虚假信息检测任务进行形式化定义，最后针对信息瓶颈理论进行简要介绍.

2.1 符号说明

使用大写字母表示随机变量(如Y等)，对应的花体字母表示随机变量的取值空间(如$\mathcal{Y}$等)，对应的小写字母表示随机变量的实例(如y等). 加粗大写字母表示矩阵，如$\boldsymbol{A}$，$\boldsymbol{X}$.$p(Y)$表示随机变量Y的概率分布. $p(Y=y)$表示随机变量Y在值y上的概率密度，简写为$p(y)$.

2.2 问题定义

为更加精准地对虚假信息进行识别检测，本文针对虚假信息的信息内容与传播拓扑进行分析，实现少标签场景下的鲁棒虚假信息检测. 具体而言，虚假信息的信息内容及其拓扑可表示为一个属性图$G=\{\mathcal{V},\mathcal{E},\mathcal{X}\}$，其中$\mathcal{V}$表示图$G$的节点集合，即社交网络用户，$\mathcal{E}$表示社交网络用户在社交网络中的交互行为，如评论、转发等，$\mathcal{X}$表示对应用户节点所发表的信息内容，如评论文本、博文内容等. 本文通过预训练模型BERT[38]将此类信息内容进行编码，表征为结构化低维表征. 此属性图可形式化为矩阵形式$G=\{\boldsymbol{A},\boldsymbol{X}\}$，其中$\boldsymbol{A}\in\mathbb{R}^{n\times n}$表示图的邻接矩阵，$\boldsymbol{X}\in\mathbb{R}^{n\times d}$表示图的节点特征矩阵. n，d分别表示节点数量与节点特征维度. 当前虚假信息检测任务通常将网络信息分为4类[39]，$\mathcal{Y}\in\{\mathrm{F},\mathrm{T},\mathrm{U},\mathrm{N}\}$，F,T,U,N分别表示证实为假的传闻、证实为真的传闻、未经证实的传闻以及非传闻信息. 虚假信息检测旨在学习一个分类器$f:\mathcal{G}\to\mathcal{Y}$，将虚假信息及其传播拓扑$G\in\mathcal{G}$归类为以上4种信息类别.

鲁棒少标签虚假信息检测旨在信息标签获取代价高，且信息传播过程中存在人为恶意操纵的场景下进行高效的虚假信息检测. 形式化地，给定虚假信息数据集$\mathcal{D}=\{\mathcal{D}_{\mathrm{L}},\mathcal{D}_{\mathrm{U}}\}$，其中包含少量的标注样本$\mathcal{D}_{\mathrm{L}}$和大量的无标注样本$\mathcal{D}_{\mathrm{U}}$，即$|\mathcal{D}_{\mathrm{L}}|\ll|\mathcal{D}_{\mathrm{U}}|$. 虚假信息$G\in\mathcal{D}$的信息内容(节点特征)和传播拓扑(拓扑结构)可能受到传播者的恶意操纵，$G'=\mathcal{M}(G)$，其中$\mathcal{M}$表示恶意操纵行为. 鲁棒少标签虚假信息检测指基于少标签数据集$\mathcal{D}$学习鲁棒分类器$f:\mathcal{M}(\mathcal{D})\to\mathcal{Y}$，当信息存在恶意操纵时，仍能完成虚假信息的识别.

2.3 信息瓶颈理论

给定输入X和对应的标签Y，信息瓶颈理论[14-15, 40]旨在学习鲁棒压缩表征Z，使得表征Z在尽可能保存标签Y信息的同时消除X中的标签无关信息. 形式化地，假设存在马尔可夫链$Z\to X\to Y$，即在给定X的条件下，Z相对于Y条件独立. 信息瓶颈(information bottleneck，IB)理论通过求解以下优化问题学习鲁棒表征Z：

$$\max_{Z}\mathcal{L}_{\mathrm{IB}}=I(Z;Y)-\beta I(X;Z)\propto I(Z;Y)-\beta' I(Z;X|Y),$$

其中互信息(mutual information，MI)用于衡量2个随机变量之间的相关性$I(X;Z)=\int_x\int_z p(x,z)\ln\dfrac{p(x,z)}{p(x)p(z)}\mathrm{d}x\mathrm{d}z$. $I(Z;Y)$表示表征Z所蕴含的标签相关信息；$I(Z;X|Y)$表示表征Z中蕴含的标签无关信息[40]；超参数β和β'用于权衡表征的信息量和压缩量.

3 鲁棒少标签虚假信息检测

本节首先介绍如何同时利用无标注样本信息和少量标注样本在少标签场景下学习高质量虚假信息表征；然后阐述基于信息瓶颈理论的鲁棒无监督虚假信息表示学习方法；最后通过联合优化少标签目标函数和鲁棒性目标函数，实现少标签场景下的鲁棒虚假信息检测.

3.1 少标签虚假信息检测

虚假信息的标注往往需要对比官方报道、分析多方信息，其成本较为昂贵，限制了虚假信息检测算法的实际应用. 为此，本文考虑少标签场景下的虚假信息检测问题. 为缓和对标注信息的依赖，可行的办法是通过互信息最大化[41-42](mutual information maximization，Infomax)技术充分挖掘无标注样本信息，并通过引入少量标注样本引导表征学习.

互信息最大化技术被广泛应用于无监督图表示学习领域[43-44]，其通过最大化图表征与图之间的互信息，在低维表示空间学习与高维图最相关的表征向量，在无标签引导的场景下挖掘图的关键特征. 本文提出通过互信息最大化技术，并引入少量标签信息，在少标签场景下学习高质量虚假信息表征. 具体而言，给定虚假信息数据集$\mathcal{D}=\{\mathcal{D}_{\mathrm{L}},\mathcal{D}_{\mathrm{U}}\}$，其中包含少量标注样本$\mathcal{D}_{\mathrm{L}}$和大量的无标注样本$\mathcal{D}_{\mathrm{U}}$，本文通过最大化表征$Z$和虚假信息图$G$之间的互信息，学习虚假信息图编码器$p_\theta(Z|G)$，表征现实场景中广泛存在的无标注虚假信息数据：

$$\max_{Z\sim p_\theta(Z|G)} \mathcal{L}_{\mathrm{U}}(\mathcal{D};\theta)=I(Z;G)\ , \tag{1}$$

其中虚假信息属性图$G\sim\mathcal{D}$采样自数据集$\mathcal{D}$, θ表示图编码器对应的可训练参数. 为同时表征虚假信息的信息内容和传播拓扑, 本文采用2层图卷积神经网络[45](graph convolutional network, GCN)作为编码器:

$$\boldsymbol{H}=\boldsymbol{A}^{*}\sigma(\boldsymbol{A}^{*}\boldsymbol{X}\boldsymbol{W}_1)\boldsymbol{W}_2,$$

其中$\boldsymbol{A}^{*}$表示图G的对称拉普拉斯矩阵[45], $\boldsymbol{X}$表示图G的节点特征矩阵, $\boldsymbol{H}$表示GCN生成的表征. $\theta=\{\boldsymbol{W}_1,\boldsymbol{W}_2\}$表示图卷积神经网络的可训练参数, σ表示ReLu激活函数. 上式以无监督的方式学习虚假信息的低维表征, 为引入少量标注信息指导表征的生成, 本文引入有监督项$I(Z;Y)$:

$$\max_{Z\sim p_\theta(Z|G)} \mathcal{L}_{\mathrm{FL}}(\mathcal{D}_{\mathrm{L}};\mathcal{D};\theta)=I(Z;G)+\gamma I(Z;Y)\ , \tag{2}$$

其中标签信息$Y\sim\mathcal{D}_{\mathrm{L}}$采样自标注数据集$\mathcal{D}_{\mathrm{L}}$, 超参数$\gamma$用于权衡有监督项$I(Z;Y)$与无监督项$I(Z;G)$的比重. 通过求解式(2)优化问题, 可以充分利用现实场景中普遍存在的无标注信息和少量代价较为高昂的有标注虚假信息, 实现在少标签场景下的虚假信息检测.

3.2 鲁棒虚假信息检测

在3.1节中, 本文提出基于互信息最大化技术的少标签虚假信息检测. 然而, 研究表明, 通过互信息最大化或对比学习等方式获得的图表征往往存在鲁棒性问题[46]. 具体而言, 现实场景中虚假信息的传播者可能通过评论区控评等策略躲避虚假信息检测[12]. 换言之, 社交网络中虚假信息对应的属性图G可能被虚假信息传播者恶意操纵, 导致其拓扑结构和节点特征信息被改变, 从而误导虚假信息检测算法的预测结果. 本文中, 将未经虚假信息传播者恶意操纵的属性图G称为原始图, 将存在恶意操纵行为的属性图G'称为对抗图. 恶意操纵行为可形式化为$G'=\mathcal{M}(G)$. 相比原始图G, 对抗图G'经过虚假信息传播者的恶意操纵(如删评等操纵行为), 其模式更加隐蔽, 因而更加难以检测.

为消除虚假信息传播者恶意操纵行为的影响, 本文提出鲁棒虚假信息检测, 研究属性图中存在对抗性扰动信息的场景下, 如何进行高效的虚假信息检测. 形式化地, 给定对抗图G', 本文旨在学习鲁棒虚假信息表征$Z'\sim p_\theta(Z'|G')$, 使得Z'在消除对抗图G'中对抗信息的同时, 保留原始图G中的信息. 为此, 本文借鉴信息瓶颈理论[14-15, 40]的思想, 提出鲁棒信息瓶颈(robust information bottleneck, RIB)理论权衡表征Z'的信息量与压缩量, 在保留原始图关键信息的同时, 压缩对抗图中引入的对抗信息:

$$\begin{aligned}&\max_{Z'\sim p_\theta(Z'|G')} \mathcal{L}_{\mathrm{RIB}}(\mathcal{D};\theta)=I(G;Z')-\beta I(Z';G'|G),\\ &\text{s.t. } G'=M(G),\end{aligned} \tag{3}$$

其中$I(G;Z')$被称为信息项, 其衡量Z'中保留的原始图信息; $I(Z';G'|G)$被称为对抗项, 其衡量Z'中保留的对抗信息. 对抗信息由恶意操纵行为$\mathcal{M}$引入, 仅存在于G'. 超参数β用于平衡信息项和对抗项的权重. 虚假信息原始图$G\sim\mathcal{D}$采样自数据集$\mathcal{D}$, 对抗图G'经由虚假信息传播者的恶意操纵$\mathcal{M}$生成. 式(3)通过同时最大化信息项和最小化对抗项, 实现在保留原始图G中信息的同时, 消除对抗图G'中的对抗信息.

为模拟现实场景中虚假信息传播者对原始图G的恶意操纵行为$\mathcal{M}$, 本文采用对抗训练的策略生成对抗图G'. 具体而言, 本文通过最小化目标函数$\mathcal{L}_{\mathrm{RIB}}$得到对抗图$G'$, 使得$G'$能够污染表征$Z'$, 从而影响最终检测结果:

$$\begin{aligned}&\max_{Z'\sim p_\theta(Z'|G')} \mathcal{L}_{\mathrm{RIB}}(\mathcal{D};\theta)=I(G;Z')-\beta I(Z';G'|G),\\ &\text{s.t. } G'=\arg\min_{G'}\mathcal{L}_{\mathrm{RIB}}(\mathcal{D};\theta),\ d(G',G)\leqslant\epsilon,\end{aligned} \tag{4}$$

其中$d(G',G)$表示原始图G与对抗图G'之间的距离, ϵ用于限制恶意操纵行为对原始图G添加的扰动量. 在实验中, 本文采用l_1范式实例化距离度量函数$d(G',G)$, 分别限制图中边的扰动量和节点特征的扰动量.

3.3 鲁棒少标签虚假信息检测

通过联合式(2)与式(4), 可实现鲁棒少标签虚假信息检测. 形式化地, 定义鲁棒少标签目标函数为

$$\begin{aligned}L(\mathcal{D}_{\mathrm{L}};\mathcal{D};\theta)=&(1-\alpha)\mathcal{L}_{\mathrm{FL}}(\mathcal{D}_{\mathrm{L}};\mathcal{D};\theta)+\alpha\mathcal{L}_{\mathrm{RIB}}(\mathcal{D};\theta)=\\ &(1-\alpha)I(Z;G)+\gamma' I(Z;Y)+\\ &\alpha I(G;Z')-\beta' I(Z';G'|G),\end{aligned} \tag{5}$$

其中超参数α用于权衡少标签目标函数$\mathcal{L}_{\mathrm{FL}}(\mathcal{D}_L;\mathcal{D};\theta)$与鲁棒性目标函数$\mathcal{L}_{\mathrm{RIB}}(\mathcal{D};\theta)$之间的权重. 2个目标函数内部的超参数$\gamma$和$\beta$经过加权后可重新表示为$\gamma'=\gamma(1-\alpha)$, $\beta'=\beta\alpha$. 通过针对$\mathcal{L}(\mathcal{D}_{\mathrm{L}};\mathcal{D};\theta)$进行对抗训练, 可实现鲁棒少标签的虚假信息检测:

$$\begin{aligned}&\max_{Z'\sim p_\theta(Z'|G'),Z\sim p_\theta(Z|G)} \mathcal{L}(\mathcal{D}_{\mathrm{L}};\mathcal{D};\theta)\\ &\text{s.t. } G'=\arg\min_{G'}\mathcal{L}_{\mathrm{RIB}}(\mathcal{D};\theta),\ d(G',G)\leqslant\epsilon.\end{aligned} \tag{6}$$

上述优化问题为2层优化问题, 在内层优化中, 需要搜索最优对抗图G', 使其能够最小化$\mathcal{L}_{\mathrm{RIB}}$. 该问题在文献中被称为图对抗攻击问题[47-48]. 图对抗攻击可同时作用于图拓扑结构和图节点特征. 然而, 由于图拓扑结构的离散性特点, 针对图拓扑结构的对抗攻击算法往往具有较高的时间复杂度. 为降低对抗图的求解成本, 本文针对图拓扑结构和图节点特征

选择不同的对抗扰动策略. 对于离散的图拓扑结构，本文采用随机扰动策略，用以快速改变图结构. 具体而言，本文以p的概率随机删除图G中的边结构，以生成对抗图G'的拓扑结构. 对于图的节点特征，本文采用投影梯度下降[49](projected gradient descent, PGD)算法求解最优扰动图，使其能够最小化$\mathcal{L}_{\mathrm{RIB}}$.

3.4 优化求解

式(6)中的目标函数$\mathcal{L}(\mathcal{D}_{\mathrm{L}};\mathcal{D};\theta)$涉及4个互信息优化项$I(Z:G)$, $I(Z;Y)$, $I(G;Z')$, $I(Z';G'|G)$. 现实场景中，互信息的计算过程往往涉及针对高维数据先验分布的积分，由于先验分布的不可知和高维数据积分的极高复杂度，互信息的优化一直以来是机器学习领域的一大难题[50-51]，下文将重点阐述本文如何针对以上4项互信息进行优化.

1)最大化$I(Z;G)$和$I(G;Z')$. 为了优化$I(Z:G)$和$I(G;Z')$，本文选取被广泛采用的詹森-香农散度(Jensen-Shannon divergence, JSD)互信息估计器[52]：

$$\hat{I}(Z;G)=E_{p(G,Z)}[\ln D(Z;G)]+E_{p(G),p(Z)}[\ln(1-D(Z;G))].$$

此式通过对比学习范式估计图及其表征之间的互信息. $I(G;Z')$可通过类似方法进行估计. 判别器D用于判断表征Z与图G是否采样自联合分布$p(G,Z)$. 判别器D采用简化图卷积网络[53](simplifying graph convolutional network)与双线性网络(bilinear network)架构：

$$D(Z;G)=Biliner(SGC(G);Z),$$

其中$SGC(G)=\boldsymbol{A}^{*K}\boldsymbol{X}\boldsymbol{W}$，$\boldsymbol{A}^{*}$表示图$G$的对称拉普拉斯矩阵[45]，$\boldsymbol{X}$表示图$G$的节点特征矩阵，$\boldsymbol{W}$表示简化图卷积网络的可训练参数. 简化图卷积网络用于学习图G的图摘要，将图G映射至低维向量空间. 双线性网络层用于判断G的图摘要和表征Z是否采样自联合分布$p(G,Z)$.

2)最大化$I(Z;Y)$. 为优化$I(Z;Y)$，本文通过变分推断的方法[54]推导出$I(Z;Y)$的下确界：

$$I(Z;Y)=\int p(Y,Z)\ln p(Y|Z)\mathrm{d}Y\mathrm{d}Z+H(Y),$$

其中Y的熵$H(Y)$仅与标签Y自身相关，因此可忽略不计. 由于先验分布$p(Y|Z)$往往不可知，本文选取其变分近似$q_\varphi(Y|Z)$进行估计：

$$\begin{aligned}I(Z;Y)=&\int p(Y,Z)\ln q_\varphi(Y|Z)\mathrm{d}Y\mathrm{d}Z+\\&KL\left[p(Y|Z)|q_\varphi(Y|Z)\right]+H(Y)\geqslant\\&\int p(Y,Z)\ln q_\varphi(Y|Z)\mathrm{d}Y\mathrm{d}Z\approx\\&\frac{1}{N}\sum_{i=1}^{N}\ln q_\varphi(y_i|z_i).\end{aligned}\tag{7}$$

式(7)中，N表示批训练中采样的批样本数量. 式(7)变分推导表明，$I(Z;Y)$可通过基于分类器$q_\varphi(Y|Z)$的分类损失进行优化. 在实验中，选取逻辑回归分类器作为变分近似$q_\varphi(Y|Z)$，φ为逻辑回归的可训练参数，此分类器可用于最终的虚假信息检测. 由于Z同时基于有标注样本和无标注样本进行学习，缓解了标签依赖问题. 同时基于信息瓶颈理论的对抗训练策略提升了Z的鲁棒性，使得基于Z的虚假信息检测效果更加鲁棒.

3)最小化$I(Z';G'|G)$. 根据条件互信息的定义，$I(Z';G'|G)$可形式化为

$$I(Z';G'|G)=E_{p(G,G',Z')}\left[\ln\frac{p(G)\,p(G,G',Z')}{p(G',G)\,p(Z',G)}\right].$$

在本文中，表征Z'基于对抗图G'生成，G'由原始图G经过对抗攻击生成，由此可认为$p(G,G',Z')=p_\theta(Z'|G')\,p(G'|G)\,p(G)$. 互信息$I(Z';G'|G)$可重写为

$$\begin{aligned}I(Z';G'|G)=&E_{p(G,G')}E_{p_\theta(Z'|G')}\left[\ln\frac{p_\theta(Z'|G')}{p(Z'|G)}\right]=\\&E_{p(G,G')}E_{p_\theta(Z'|G')}\left[\ln\frac{p_\theta(Z'|G')\,p_\theta(Z|G)}{p_\theta(Z|G)\,p(Z'|G)}\right]=\\&KL\left[p_\theta(Z'|G')|p_\theta(Z|G)\right]-\\&KL\left[p_\theta(Z|G)|p(Z'|G)\right]\leqslant\\&KL\left[p_\theta(Z'|G')|p_\theta(Z|G)\right],\end{aligned}\tag{8}$$

其中$p_\theta(Z'|G')$和$p_\theta(Z|G)$分别表示对抗图G'和原始图G经过编码器编码的过程. 在实验中，本文假设2项表征的条件分布均服从高斯分布，形式化地：$p_\theta(Z'|G')\sim\mathcal{N}(\mu_\theta(G'),\sigma_\theta(G'))$, $p_\theta(Z|G)\sim\mathcal{N}(\mu_\theta(G),\sigma_\theta(G))$. 其中$\mu_\theta(\cdot)$和$\sigma_\theta(\cdot)$由GCN实现，$\theta$表示其可训练参数. 式(8)表明，通过最小化对抗图表征和原始图表征的Kullback-Leibler散度，可最小化$I(Z';G'|G)$，消除G'中对抗性信息的影响.

图1展示了本文所提方法的整体框架，该方法主要分为4步：首先，本文将虚假信息的信息内容经过预训练模型BERT编码，构造虚假信息属性图；然后，基于图对抗攻击算法向原始图G添加对抗扰动，生成对抗图G'；其次，通过GCN编码器学习虚假信息表征，并通过优化以上互信息项实现鲁棒虚假信息表示学习；最后，基于鲁棒图表征融入少量标注样本信息，实现少标签场景下的鲁棒虚假信息检测，将输入信息分为4类，即$\mathcal{Y}\in\{\mathrm{F,T,U,N}\}$，F,T,U,N分别表示证实为假的传闻、证实为真的传闻、未经证实的传闻以及非传闻信息. 整体算法流程可总结为算法1.

算法1. 鲁棒少标签虚假信息检测.

输入：虚假信息数据集$\mathcal{D}=\{\mathcal{D}_{\mathrm{U}},\mathcal{D}_{\mathrm{L}}\}$；

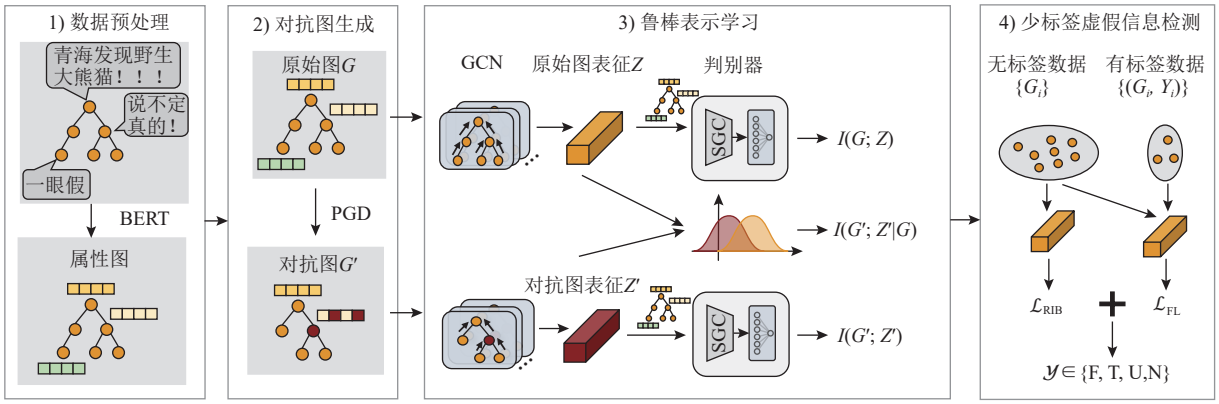


Fig. 1 The framework of our proposed method

图 1 本文所提方法的框架

输出：虚假信息检测 Y.

① 数据预处理. 将数据集中的信息文本内容经过 BERT 编码为结构化低维表征，构建为属性图 $G=\{\boldsymbol{A},\boldsymbol{X}\}$.

② 对抗图生成. 通过随机扰动策略改变属性图 G 的拓扑结构，通过 PGD 算法改变属性图 G 的节点特征，生成对抗图 $G'=\{\boldsymbol{A}',\boldsymbol{X}'\}$.

③ 表示学习. 通过 3.3 节所提出的互信息优化方法，对式(6)进行求解，以学习少标签场景下的鲁棒虚假信息表征 Z.

④ 虚假信息识别. 对于给定属性图 G，通过学得的鲁棒少标签虚假信息表征 Z 和式(7)中的虚假信息分类器 $q_\varphi(Y|Z)$，将待判别虚假信息归类为 4 类，即 $\mathcal{Y}\in\{\mathrm{F},\mathrm{T},\mathrm{U},\mathrm{N}\}$.

4 实　　验

本节通过在真实数据集上的大量实验验证本文所提的虚假信息检测方法在少标签识别和鲁棒性方面的优越性. 具体而言，本节致力于回答 3 个问题：

问题 1. 少标签场景下，本文所提方法较基准方法在识别精准性方面表现如何.

问题 2. 当虚假信息传播过程中存在人为恶意扰动时，本文所提方法较基准方法在鲁棒性方面表现如何.

问题 3. 本文所提出的少标签目标函数和鲁棒性目标函数如何影响虚假信息检测的效果.

4.1 数据集

为评估所提虚假信息检测方法的效果，本文基于公开真实数据集 Twitter15 和 Twitter16[39] 进行充分的实验验证. 2 个数据集均收集于美国最有影响力的社交网站推特，其中每条虚假信息内容由源推文和关联推文 2 部分组成，源推文包含 4 种标签类别：证实为假的传闻(F)、证实为真的传闻(T)、未经认证的传闻(U)以及非传闻信息(N)，适用于 4 分类任务. 关联推文指与源推文存在转发、评论等关联关系的推文. 根据源推文与其关联推文之间的传播关系，可以为每个源推文构建对应的属性图结构. 4 个数据集的统计信息见表 1.

Table 1 Statistics of the Datasets

表 1 数据集统计信息

统计项目	Twitter15	Twitter16
源推文	1 490	818
用户	276 663	173 487
关联推文	331 612	204 820
证实为假的传闻	370	205
证实为真的传闻	372	205
未经证实的传闻	374	203
非传闻	374	205

4.2 基准方法

本文将 5 种基准方法与本文方法进行比较：

1)BERT[38]. 是一种基于 Transformer 架构的预训练语言模型，可以用于对谣言的文本信息进行编码，并在此基础上训练逻辑回归分类器进行虚假信息检测.

2)RvNN[24]. 是一种基于树状结构的谣言检测分类器，可以通过自上而下和自下而上的树结构来传播信息，从而获得虚假信息的表征.

3)UDGCN[10]. 使用图卷积神经网络来进行谣言检测，并且使用根节点增强策略提高检测性能.

4)BiGCN[10]. 综合考虑了虚假信息的传播与扩散特性，使用自上而下的有向图卷积神经网络来建模虚假信息从源推文开始的传播过程，使用自下而上的有向图卷积神经网络将分散的评论信息聚合到源推文.

5)GACL[12]. 基于 GCN 通过基于标签的对比学习策略学习挖掘虚假信息特征,并引入对抗性特征变换(adversarial feature transformation, AFT)模块,在学习全局拓扑表征的同时消除噪声和对抗性信息的影响.

6)本文方法. 通过融合无标注样本信息和少量标注样本,基于信息瓶颈理论在少标签场景下实现鲁棒的虚假信息检测.

4.3 实验细节

在数据预处理阶段,本文通过 BERT[38] 模型将数据集中的每条推文信息进行编码,表征为 768 维向量,并将其作为该推文对应的节点特征. 通过将推文作为图中节点,将推文之间的关联关系作为拓扑结构,虚假信息可构建为一个属性图. 在训练阶段,本文遵循与现有工作[10, 12]一致的数据集划分策略,随机将数据集划分为 5 份,进行 5 折交叉验证. 本文选取 2 层[45]GCN 作为编码器,用以表征虚假信息属性图,其隐藏层和图表征维度均设置为 64. 基于虚假信息表征,本文采用多层感知机(multilayer perceptron, MLP)进行虚假信息的分类. 本文选择 Adam 优化器[55] 针对神经网络进行优化,并将学习率设置为 0.001. 在测试阶段,本文选择准确率(accuracy)和 $F1$ 分数评估虚假信息检测的效果.

4.4 少标签虚假信息检测

现实场景中,虚假信息的标注往往需要大量的人工参与,标注代价较为高昂. 为回答问题 1,本文通过实验探究本文方法在少标签场景下的虚假信息检测效果. 首先,本文在 Twitter15 数据集上为每个类别(U, N, T, F)分别随机选取 100 个带有标签的样本,在 Twitter16 数据集上为每个类别选取 50 个带有标签的样本,探究本文方法和基准方法在少标签场景下的虚假信息检测效果. 表 2 显示了各方法在 Twitter15 和 Twitter16 的虚假信息检测准确率和 $F1$ 分数. 可以观测出,少标签场景下相比现有基准方法,本文方法在准确率指标和 $F1$ 分数上总体取得了更加优越的表现. 比如,在 Twitter15 上,本文方法达到 81.02% 的准确率,而表现最好的基准方法 GACL 为 78.66%;在 Twitter16 上,本文方法达到 82.31% 的准确率,次优方法 GACL 为 80.77%. 本文方法在少标签场景下的优越性能归功于本文所采用的少标签训练策略,通过互信息最大化技术挖掘无标注样本信息,同时引入少量标签信息指导表征生成,实现在少标签场景下的高效虚假信息检测.

Table 2 Detection Results of Each Method on Twitter15 and Twitter16 Datasets

表 2 各方法在 Twitter15 和 Twitter16 数据集上的检测效果 %

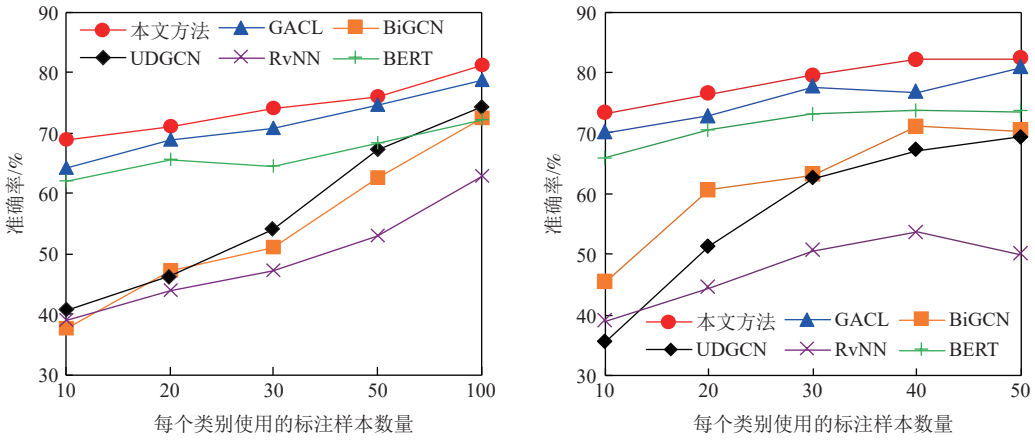
方法	Twitter15					Twitter16				
	准确率	$F1$(U)	$F1$(N)	$F1$(T)	$F1$(F)	准确率	$F1$(U)	$F1$(N)	$F1$(T)	$F1$(F)
BERT[38]	72.13	70.67	90.22	66.22	58.18	73.54	78.85	80.70	79.29	48.58
RvNN[24]	62.86	55.61	66.88	63.93	63.07	50.03	40.33	43.71	59.93	55.29
BiGCN[10]	72.44	68.49	73.54	**79.22**	69.37	70.31	60.54	68.07	82.41	68.42
UDGCN[10]	74.17	72.58	70.94	81.55	71.45	69.38	62.05	59.84	84.37	68.83
GACL[12]	78.66	74.24	91.78	77.01	69.44	80.77	84.37	81.86	88.60	66.33
本文方法	**81.02**	**77.99**	**92.08**	78.15	**73.82**	**82.31**	**84.87**	**82.26**	**90.32**	**69.07**

注:Twitter15 上每类别使用 100 个标注样本,Twitter16 上每类别使用 50 个标注样本. 黑体数值表示最优值.

为进一步探究不同标签量下本文方法的性能变化,我们在 Twitter15 和 Twitter16 上分别随机采样不同数量的标注样本,并观测其检测准确率的变化,结果如图 2 所示. 可以观察出,本文方法在少标签场景下较基准方法展现出更为优越的性能,且当标签越少时,较有监督方法如 RvNN, BiGCN 和 UDGCN 等性能提升明显. 此外,基于对比学习的 GACL 和基于预训练模型的 BERT 等在一定程度上利用了无标注样本信息,展现出仅次于本文方法的效果. 此现象表明通过挖掘无标注样本潜在特征分布,能够有效提升虚假信息检测的效果,降低其对标注信息的依赖. 然而由于 BERT 缺乏拓扑表征能力,GACL 的对比学习策略依赖于标签信息,二者的效果较本文方法仍有欠缺.

4.5 虚假信息检测的鲁棒性

现实场景中,虚假信息的传播过程中可能存在恶意操纵行为,虚假信息传播者通过评论区控评等手段扰动虚假信息及其传播拓扑,增加了虚假信息检测的难度. 为回答问题 2,验证本文方法在人为恶意扰动场景下的检测鲁棒性,本文通过向数据集中添加不同规模的随机噪声,观测检测准确率的变化. 具体而言,本文通过在测试集中以一定比率随机丢



(a) Twitter15 (b) Twitter16

Fig. 2 Detection accuracy on Twitter15 and Twitter16 datasets with different amounts of labeled examples

图 2 Twitter15 和 Twitter16 数据集上使用不同数量标注样本的检测准确率

弃边和掩除节点的方式扰动虚假信息属性图、模拟虚假信息传播者的恶意控评(如删评)等手段，评估虚假信息检测方法的鲁棒性. 结果如图 3 所示. 可以观测到，随着扰动量的增加，虚假信息检测的准确率呈现一定的下降趋势，其中，非鲁棒性检测方法 BiGCN 与 UDGCN 下降幅度明显，受扰动影响较大. 本文方法与 GACL 采用对抗训练的策略提升模型鲁棒性，受扰动影响相对较小，呈现更强的鲁棒性. 此外，由于 GACL 采用的对抗性特征变换模块仅在隐层网络中引入扰动信息，其分布较真实扰动存在差异. 与之相反，本文通过对抗攻击算法直接在针对输入数据添加最优扰动，相比 GACL 可以实现更强的鲁棒性.

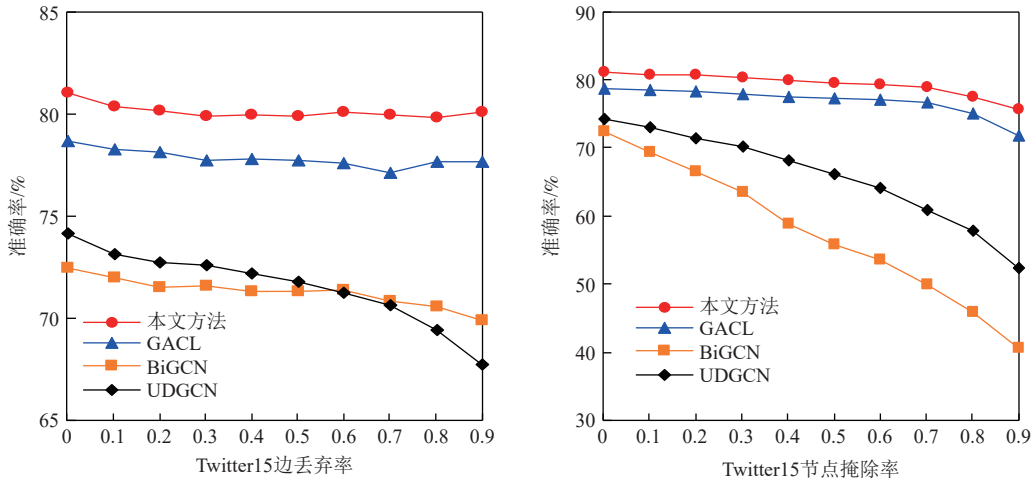
4.6 消融分析

为回答问题 3，分析本文所提出的少标签目标函数和鲁棒性目标函数如何影响最终虚假信息检测的效果，本文方法进行了消融分析，研究本文方法在消去不同模块后的性能变化. 具体而言，本文考虑 2 种变体：

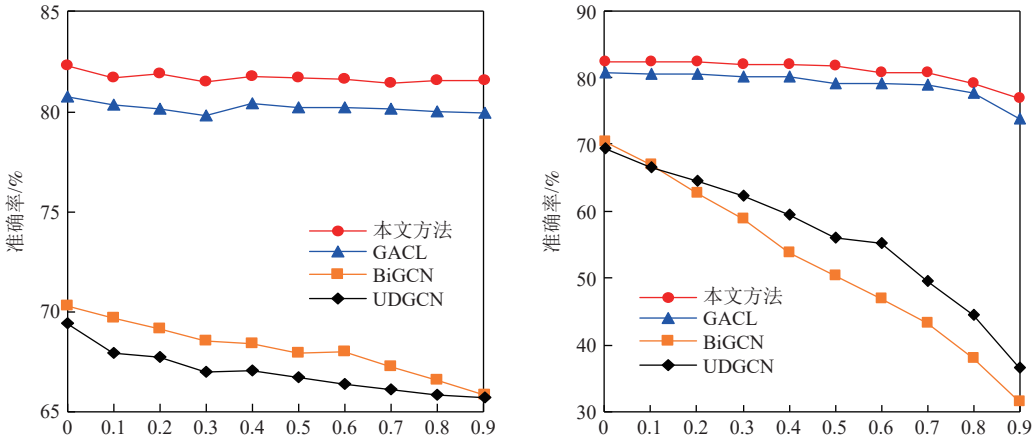
1)本文方法(w/o_ad). 仅保留少标签目标函数，不考虑鲁棒性目标函数.

2)本文方法(w/o_mi). 仅保留有监督项，不考虑基于互信息最大化的无监督目标函数，以及不考虑鲁棒性目标函数.

表 3 给出了本文方法及其 2 种变体在数据集 Twitter15 和 Twitter16 上不同标签量下的检测准确率. 以每类别仅使用 10 个标注样本时为例，在 Twitter15 和 Twitter16 上，本文方法(w/o_ad)的准确率相比原始本文方法分别下降了 4.88 个百分点和 1.69 个百分点. 这是由于不考虑鲁棒性目标函数的情况下，部分存在人为恶意扰动的虚假信息更难被模型捕捉，导致检测性能下降. 此外，在 Twitter15 和 Twitter16 上，本文方法(w/o_mi)的准确率相比原始方法分别下降了 5.67 个百分点和 3.38 个百分点，这是由于失去基



(a) 在Twitter15上丢弃边 (b) 在Twitter15上删除节点



(c) 在Twitter16上丢弃边 (d) 在Twitter16上删除节点

Fig. 3 Detection accuracy with varying perturbation rate

图 3 不同扰动量下的检测准确率

Table 3 Detection Accuracy of Our Method and Its Variants on Twitter15 and Twitter16 Datasets

表 3 本文方法与其变体在 Twitter15 和 Twitter16 数据集上的检测准确率 %

方法	Twitter15 的每个类别使用标注样本数量					Twitter16 的每个类别使用标注样本数量				
	10	20	30	50	100	10	20	30	40	50
本文方法	**68.74**	**71.1**	**73.94**	**75.91**	**81.02**	**73.23**	**76.46**	**79.54**	**82.15**	**82.31**
本文方法（w/o_ad）	63.86	70.24	73.15	75.2	80.55	71.54	76.00	78.46	81.23	80.92
本文方法（w/o_mi）	63.07	70.94	70.31	74.17	79.84	69.85	76.31	78.15	80.46	81.54

注：黑体数值表示最优值.

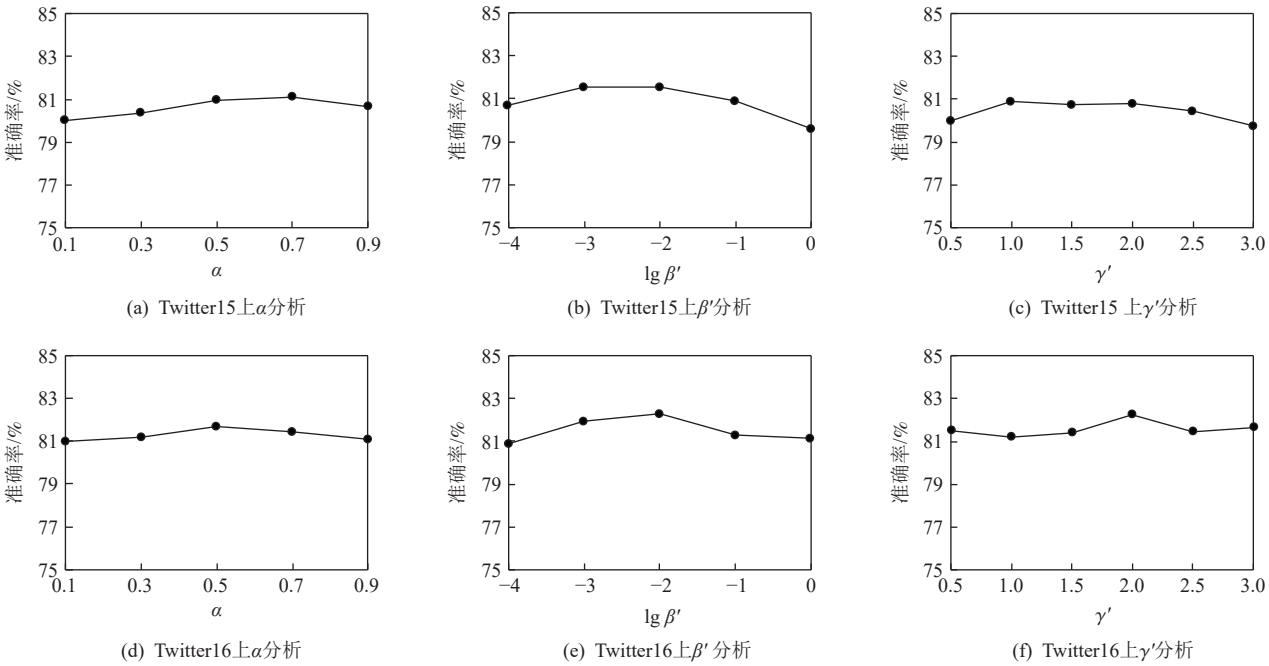
于互信息最大化的无监督项后，模型完全依赖极少的标注样本进行学习，缺乏从大量未标注样本中挖掘信息的能力.

4.7 参数分析

为了探索超参数的不同取值对模型性能的影响，本文研究了式(5)中不同的α，β'，γ'取值下模型的检测准确率，结果如图 4 所示. 由图 4 可以得出：1)超参数α用于权衡少标签目标函数与鲁棒性目标函数的比重，较小的α表明模型更加注重检测的鲁棒性，较大的α表明模型更加注重少标签下的性能. 本文选取取值空间$\alpha\in\{0.1,0.3,0.5,0.7,0.9\}$进行分析. 结果表明，当$\alpha$值过大时，模型在训练阶段不能学习对抗性的输入数据，不足以确保模型的鲁棒，导致测试集中部分数据更容易逃脱检测. 而当α值过小时，则会忽视少标签目标函数的重要性，使得模型不能从无标注样本中获取足够的信息. 2)超参数β'用于权衡信息瓶颈理论中信息项与对抗项的比重，较小的β'会更多关注信息项，较大的β'会更多关注对抗项. 本文选取取值空间$\beta'\in\{0.000\,1,0.001,0.01,0.1,1\}$进行分析. 结果表明，较小的$\beta'$会导致表征缺乏屏蔽对抗性信息的能力，影响检测效果；较大的β'会导致表征缺乏表征关键信息的能力，无法有效识别虚假信息. 3)超参数γ'用于权衡有监督项与无监督项的比重，较小的γ'会更多关注无监督信息，较大的γ'会更多关注有监督信息. 本文选取取值空间$\gamma'\in\{0.5,1.0,1.5,2.0,2.5,3.0\}$进行分析. 结果表明过大或过小的$\gamma'$会仅关注有监督信息或无监督信息，无法充分利用数据集中的信息，影响最终检测结果.

5 结 论

本文针对虚假信息检测的标签稀缺性和恶意扰动性问题进行了研究，提出一种基于信息瓶颈理论的鲁棒少标签虚假信息检测方法. 本文通过互信息最大化技术融合无标注样本信息，克服虚假信息检测对标签的过分依赖问题；并通过对抗训练的策略模拟虚假信息传播者的恶意操纵行为，基于信息瓶颈理论学习鲁棒的虚假信息表征，在高质量表征虚假信息的同时消除恶意操纵行为的影响. 实验表明，本文方法在少标签识别和鲁棒性 2 个方面均取得了



(a) Twitter15上α分析 (b) Twitter15上β'分析 (c) Twitter15 上γ'分析

(d) Twitter16上α分析 (e) Twitter16上β' 分析 (f) Twitter16上γ'分析

Fig. 4 Detection accuracy with varying hyperparameters

图 4 不同超参数下的检测准确率

优于基准方法的效果.

本文方法基于独立同分布假设，然而实际场景存在测试环境分布偏移(out of distribution, OOD)的问题. 模型受限于所采集的数据，往往无法泛化至不同语种、不同社交平台以及不同信息主题的虚假信息检测. 如何进行跨语种、跨平台、跨主题的虚假信息检测是未来值得研究的重点方向. 此外，本文鲁棒性评估方法主要针对社交平台中常见的随机噪声数据，当平台中存在对抗攻击时如何进行鲁棒性评估尚待研究.

作者贡献声明：王吉宏提出了算法思路和实验方案并撰写论文；赵书庆负责完成实验并撰写论文，与王吉宏同等贡献；罗敏楠针对论文提出指导意见并修改论文；刘欢针对论文提出了修改意见；赵翔针对论文进行了修改和指导；郑庆华对论文所研究的问题提出了指导性意见.

参 考 文 献

[1] Amrita B, Shu Kai, Gao Min, et al. Disinformation in the online information ecosystem: Detection, mitigation and challenges[J]. Journal of Computer Research and Development, 2021, 58(7): 1353−1365 (in Chinese)
(Amrita B, 舒凯, 高旻, 等. 网络信息生态系统中的虚假信息: 检测、缓解与挑战[J]. 计算机研究与发展, 2021, 58(7): 1353−1365)

[2] Qi Peng, Cao Juan, Sheng Qiang. Semantics-enhanced multi-modal fake news detection[J]. Journal of Computer Research and Development, 2021, 58(7): 1456−1465 (in Chinese)
(亓鹏, 曹娟, 盛强. 语义增强的多模态虚假新闻检测[J]. 计算机研究与发展, 2021, 58(7): 1456−1465)

[3] Xu Mingda, Zhang Zike, Xu Xiaoke. Research on spreading mechanism of false information in social networks by motif degree[J]. Journal of Computer Research and Development, 2021, 58(7): 1425−1435 (in Chinese)
(徐铭达, 张子柯, 许小可. 基于模体度的社交网络虚假信息传播机制研究[J]. 计算机研究与发展, 2021, 58(7): 1425−1435)

[4] Shu Kai, Sliva A, Wang Suhang, et al. Fake news detection on social media: A data mining perspective [C]//Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2017, 19(1): 22−36

[5] Shu Kai, Mahudeswaran D, Wang Suhang, et al. Hierarchical propagation networks for fake news detection: Investigation and exploitation[C]//Proc of the 14th Int AAAI Conf on Web and Social Media. Palo Alto, CA: AAAI, 2020: 626−637

[6] Zhou Xinyi, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities[J]. ACM Computing Surveys, 2020, 53(5): 1−40

[7] Lu Yiju, Li Chengte. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 505−514

[8] Ma Jing, Gao Wei, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proc of the 25th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 3818–3824

[9] Wu Lianwei, Rao Yuan, Zhao Yongqiang, et al. DTCA: Decision tree-based co-attention networks for explainable claim verification[C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 1024−1035

[10] Bian Tian, Xiao Xi, Xu Tingyang, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 549−556

[11] Nguyen V, Sugiyama K, Nakov P, et al. Fang: Leveraging social context for fake news detection using graph representation[C]//Proc of the 29th ACM Int Conf on Information & Knowledge Management. New York: ACM, 2020: 1165−1174

[12] Sun Tiening, Qian Zhong, Dong Sujun, et al. Rumor detection on social media with graph adversarial contrastive learning[C]//Proc of the 34th ACM Web Conf. New York: ACM, 2022: 2789−2797

[13] Yuan Chunyuan, Ma Qianwen, Zhou Wei, et al. Jointly embedding the local and global relations of heterogeneous graph for rumor detection[C]//Proc of the 19th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2019: 796−805

[14] Tishby N, Pereira F C, Bialek W. The information bottleneck method[J]. arXiv preprint, arXiv: physics/0004057, 2000

[15] Alemi A A, Fischer I, Dillon Joshua V, et al. Deep variational information bottleneck[C/OL]//Proc of the 5th Int Conf on Learning Representations. New York: OpenReview.net, 2017[2023-11-14]. https://openreview.net/pdf?id=HyxQzBceg

[16] Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media[C]//Proc of the 13th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2013: 1103−1108

[17] Wu Ke, Yang Song, Zhu K Q. False rumors detection on sina weibo by propagation structures[C]//Proc of the 31st IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2015: 651−662

[18] Yang Fan, Liu Yang, Yu Xiaohui, et al. Automatic detection of rumor on sina weibo[C/OL]//Proc of the 18th ACM SIGKDD Workshop on Mining Data Semantics. New York: ACM, 2012[2023-11-14].https://dl.acm.org/doi/10.1145/2350190.2350203

[19] Zhao Zhe, Resnick P, Mei Qiaozhu. Enquiring minds: Early detection of rumors in social media from enquiry posts[C]//Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 1395−1405

[20] Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]//Proc of the 20th Int Conf on World Wide Web. Berlin: Springer, 2011: 675−684

[21] Qi Peng, Cao Juan, Yang Tianyun, et al. Exploiting multi-domain visual information for fake news detection[C]//Proc of the 19th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2019: 518−527

[22] Schwarz S, Theóphilo A, Rocha A. EMET: Embeddings from multilingual-encoder transformer for fake news detection[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 2777−2781

[23] Udandarao V, Maiti A, Srivatsav D, et al. Cobra: Contrastive bi-modal representation algorithm [J]. arXiv preprint, arXiv: 2005.03687, 2020

[24] Ma Jing, Gao Wei, Wong K. Rumor detection on Twitter with tree-structured recursive neural networks[C]//Proc of the 56th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2018: 1980−1989

[25] Yu Feng, Liu Qiang, Wu Shu, et al. A convolutional approach for misinformation identification[C]//Proc of the 26th Int Joint Conf on Artificial Intelligence. Palo Alto, CA : AAAI, 2017: 3901−3907

[26] Ma Jing, Gao Wei, Wong K. Detect rumors on Twitter by promoting information campaigns with generative adversarial learning[C]//Proc of the 31st World Wide Web Conf. New York: ACM, 2019: 3049−3055

[27] Wei Penghui, Xu Nan, Mao Wenji. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2019: 4786−4797

[28] Jin Zhiwei, Cao Juan, Guo Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 795−816

[29] Li Quanzhi, Zhang Qiong, Si Luo. Eventai at Semeval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information[C]//Proc of the 13th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2019: 855−859

[30] Li Tianle, Sun Yushi, Hsu S, et al. Fake news detection with heterogeneous Transformer [J]. arXiv preprint, arXiv: 2205.03100, 2020

[31] Mehta N, Pacheco Maria L, Goldwasser D. Tackling fake news detection by continually improving social context representations using graph neural networks[C]//Proc of the 60th Annual Meeting of the ACL Stroudsburg. Stroudsburg, PA: ACL, 2022: 1363−1380

[32] Ma Shuang, Mcduff D, Song Y. Unpaired image-to-speech synthesis with multimodal information bottleneck[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 7597−7606

[33] Wang Junxia, Zheng Yuanjie, Ma Jun, et al. Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation [J/OL]. Medical Image Analysis, 2023[2023-11-14].https://www.sciencedirect.com/science/article/abs/pii/S1361841522003152

[34] Zhang Cenyuan, Zhou Xiang, Wan Yixin, et al. Improving the adversarial robustness of nlp models by information bottleneck[J]. arXiv preprint, arXiv: 2206.05511, 2022

[35] Mahabadi K, Belinkov Y, Henderson J. Variational information bottleneck for effective low-resource fine-tuning[C/OL]//Proc of the 9th Int Conf on Learning Representations. New York: OpenReview.net, 2021[2023-11-14].https://openreview.net/forum?id= kvhzKz-_DMF

[36] Wang Jihong, Luo Minnan, Li Jundong, et al. Empower post-hoc graph explanations with information bottleneck: A pre-training and fine-tuning perspective[C]//Proc of the 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2023:

2349–2360

[37] Sun Qingyun, Li Jianxin, Peng Hao, et al. Graph structure learning with variational information bottleneck[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 4165−4174

[38] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the 2019 Conf of the North American Chapter of the ACL: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[39] Ma Jing, Gao Wei, Wong K. Detect rumors in microblog posts using propagation structure via kernel learning[C]//Proc of the 55th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2017: 708−717

[40] Federici M, Dutta A, Forré P, et al. Learning robust representations via multi-view information bottleneck[C/OL]//Proc of the 8th Int Conf on Learning Representations. New York: OpenReview. net, 2020[2023-11-14].https://openreview.net/pdf?id=B1xwcyHFDr

[41] Tschannen M, Djolonga J, Rubenstein P K, et al. On mutual information maximization for representation learning[C/OL]//Proc of the 8th Int Conf on Learning Representations. New York: OpenReview. net, 2020[2023-11-14].https://openreview.net/pdf?id=rk xoh24FPH

[42] Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization [C/OL]//Proc of the 7th Int Conf on Learning Representations. New York: OpenReview. net, 2019[2023-11-14].https://openreview.net/forum?id=Bklr3j0cKX

[43] Veličković P, Fedus W, Hamilton William L, et al. Deep graph infomax[C/OL]//Proc of the 7th Int Conf on Learning Representations. New York: OpenReview. net, 2019[2023-11-14].https://openreview.net/pdf?id=rklz9iAcKQ

[44] Peng Zhen, Huang Wenbing, Luo Minnan, et al. Graph representation learning via graphical mutual information maximization[C]//Proc of the 32nd Web Conf. Berlin: Springer, 2020: 259−270

[45] Kipf T, Welling M. Semi-supervised classification with graph convolutional networks[C/OL]//Proc of the 5th Int Conf on Learning Representations. New York: OpenReview. net, 2017[2023-11-14]. https://openreview.net/forum?id=SJU4ayYgl

[46] In Y, Yoon K, Park C. Similarity preserving adversarial graph contrastive learning[C]//Proc of the 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2023: 867−878

[47] Sun Yiwei, Wang Suhang, Tang Xianfeng, et al. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach[C]//Proc of the 32nd Web Conf. Berlin: Springer, 2020: 673−683

[48] Xu Kaidi, Chen Hongge, Liu Sijia, et al. Topology attack and defense for graph neural networks: An optimization perspective[C]//Proc of the 28th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2019: 3961−3967

[49] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C/OL]//Proc of the 6th Int Conf on Learning Representations. New York: OpenReview. net, 2018[2023-11-14].https://openreview.net/forum?id=rJzIBfZAb

[50] Belghazi Mohamed I, Baratin A, Rajeshwar S, et al. Mutual information neural estimation[C]//Proc of the 35th Int Conf on Machine Learning. New York: PMLR, 2018: 531−540

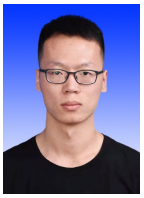
[51] Poole B, Ozair S, Van Den Oord A, et al. On variational bounds of mutual information[C]//Proc of the 36th Int Conf on Machine Learning. New York: PMLR, 2019: 5171−5180

[52] Nowozin S, Cseke B, Tomioka R. F-GAN: Training generative neural samplers using variational divergence minimization[C]//Proc of the 30th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2016: 271−279

[53] Wu Felix, Souza A, Zhang Tianyi, et al. Simplifying graph convolutional networks[C]//Proc of the 36th Int Conf on Machine Learning. New York: PMLR: 2019: 6861−6871

[54] Kingma P, Welling M. Auto-encoding variational Bayes[C/OL]//Proc of the 2nd Int Conf on Learning Representations. New York: OpenReview. net, 2013[2023-11-14]. https://openreview.net/forum?id = 33X9fd2-9FyZd

[55] Kingma P, Jimmy B. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980, 2014



Wang Jihong, born in 1997. PhD candidate. His main research interests include graph learning and trusty machine learning.

王吉宏, 1997 年生. 博士研究生. 主要研究方向为图学习、可信机器学习.



Zhao Shuqing, born in 2001. Master candidate. His main research interests include graph learning and explainable deep learning.

赵书庆, 2001 年生. 硕士研究生. 主要研究方向为图学习、可解释深度学习.



Luo Minnan, born in 1984. Professor, PhD supervisor. Her main research interests include machine learning, graph learning, and cross-media retrieval.

罗敏楠, 1984 年生. 教授, 博士生导师. 主要研究方向为机器学习、图学习、跨媒体数据挖掘.



Liu Huan, born in 1990. PhD, assistant professor. His main research interests include machine learning, computer vision, and public opinion analysis.

刘　欢, 1990 年生. 博士, 助理教授. 主要研究方向为机器学习、计算机视觉、网络舆情分析.



Zhao Xiang, born in 1986. PhD, professor. His main research interests include graph data management and mining, and intelligent analytics.

赵 翔, 1986 年生. 博士, 教授. 主要研究方向为图数据管理与挖掘、智能分析.



Zheng Qinghua, born in 1969. Professor, PhD supervisor. His main research interests include theory and technology of intelligent learning environment, and network public opinion and harmful information monitoring.

郑庆华, 1969 年生. 教授, 博士生导师. 主要研究方向为智能学习环境理论与技术、网络舆情及有害信息监控.