结构-特征协同防御的图神经网络

韩继辉[1], 石玉鹏[1], 黄子奇[2], 张安琳[3], 黄道颖[1]

(1.郑州轻工业大学 计算机科学与技术学院,河南 郑州 450001;2.北方信息控制研究院集团有限公司,江苏 南京 211153;3.郑州轻工业大学 工程训练中心,河南 郑州 450001)

摘　要:为解决图神经网络在复杂扰动环境下的节点表征退化问题,本文提出一种结构-特征协同防御的图神经网络 SFCoRobustGNN。在结构层面引入稀疏注意力机制,融合结构先验以动态抑制异常边;在特征层面结合通道门控机制与非线性特征混合模块(FeatureMixPro),增强模型对特征扰动的适应能力;通过对抗训练与多目标优化策略,实现双路径协同防御。在 Cora、Citeseer 等多个基准数据集上的实验表明:面对不同强度的结构扰动(5%~40%)与特征攻击($\varepsilon=0.01\sim0.10$),所提方法优于主流基线方法,节点分类准确率明显提升。在 ogbn-products 大规模数据集上,即使面对 20%扰动率的 MetaAttack 攻击,仍能保持 71.82%的准确率,展现了良好的扩展性。消融实验验证了各模块的有效性及协同效应。所提方法有效抑制了复杂扰动下的性能衰减,并展现出良好的泛化性。

关键词:图神经网络; 鲁棒性;结构扰动;特征扰动;稀疏注意力

中图分类号:TP18;TN929.5　　**文献标志码**:A　　**doi**:10.13705/j.issn.1671-6833.2026.04.010

图神经网络(graph neural networks,GNN)凭借其对图结构数据强大的表征能力,已被广泛应用于社交网络分析、推荐系统、生物信息学等多个领域[1]。然而,在实际应用场景中,图数据往往同时受到结构扰动(如边的添加、删除或伪造)和特征扰动(如节点属性污染与对抗噪声)的双重影响,严重削弱了模型的泛化能力与预测可靠性[2]。已有研究表明,GNN 对图结构扰动表现出高度敏感性,即便仅对图结构进行少量恶意篡改,也会导致节点分类性能显著下降。因此,如何提升 GNN 在复杂扰动环境下的鲁棒性,已成为图表示学习研究中亟待解决的关键问题[3]。

针对上述挑战,现有研究方法可归纳为以下 3 类。第一类为结构优化方法,该类方法主要从图拓扑结构入手,通过剪枝、边权重调整或结构重构等技术抑制异常连接、恢复图结构语义。例如,Zhang 等[4]提出的 GNNGuard 通过计算节点间特征相似度对邻接边进行动态加权,以减轻噪声传播;Chen 等[5]提出了一种对比学习框架,为无监督生成鲁棒节点表示提供了通用范式,其核心思想可适配于图结构重构任务;Jin 等[6]提出的 ProGNN 将邻接矩阵参数转化为可学习变量,并施加稀疏性与平滑性约束,实现了图结构与节点特征的联合学习。尽管如此,此类方法大多依赖预处理或静态图优化策略,难以对训练过程中出现的动态扰动做出及时响应;此外,这些方法往往更关注拓扑结构的修正,而对特征空间中存在的对抗性扰动缺乏显式建模机制。第二类为表示增强方法,该方法侧重于通过改进特征学习过程获得对扰动具有不变性的节点表示,主要策略包括多视图建模、扰动模拟和表示一致性对齐[7]。例如,Veličković 等[8]通过最大化局部节点表示与全局图摘要间的互信息,实现无监督鲁棒表示学习;Kamhoua 等[9]通过随机移除边和掩蔽特征生成多个增强视图,并利用对比学习促进不同视图间表示的一致性;Zhu 等[10]提出的 GraphCL 与 Thakoor 等[11]提出的 BGRL 进一步扩展对比学习,证实这些方法提升鲁棒性的潜力。此类方法虽不需要显式修改图结构,具有较强的通用性,但在面对针对特

收稿日期:2026-01-28;**修订日期**:2026-02-28

基金项目:河南省科技攻关项目(252102211089;232102210064)

作者简介:韩继辉(1987—),男,河南周口人,郑州轻工业大学副教授,博士,主要从事复杂网络、图深度学习方向研究,E-mail:hanjihui@zzuli.edu.cn。

通信作者:黄道颖(1967—),男,河南信阳人,郑州轻工业大学教授,博士,主要从事分布式计算、智能算法方向研究,E-mail:dyhuang@zzuli.edu.cn。

征空间的高强度对抗攻击时,其防御效能仍显不足;同时,现有方法大多未能显式区分结构扰动与特征扰动,缺乏对复合攻击的有效应对机制。第三类为鲁棒训练方法,该类方法通过设计具有抗干扰能力的优化目标、正则化项或训练策略,以提升模型整体的鲁棒性。例如,Liu 等[12]引入节点级平滑度调节机制,在促进有效信息传播的同时抑制噪声扩散。此外,一些研究尝试将对抗训练与随机平滑技术引入图学习领域,从理论层面为模型鲁棒性提供保障。这类方法通常支持端到端训练,并具有较强的适应性,但也常有训练过程不稳定、计算开销大、超参数敏感等问题,使其难以适用于超大规模图数据或对实时性要求较高的实际场景。

现有鲁棒图学习方法多针对单一类型扰动设计,缺乏对结构与特征扰动的联合建模与协同防御机制,在现实场景下面对拓扑与属性双重攻击时,其稳定性和泛化能力明显不足。具体表现为结构优化方法依赖固定先验,适应性有限;特征噪声建模能力较弱,尤其对非线性扰动缺乏有效处理;部分方法计算复杂度高,难以扩展至大规模图数据。针对上述局限,本文提出结构-特征协同防御的图神经网络 SFCoRobustGNN,该框架在结构层面融合稀疏注意力与拓扑约束,动态抑制噪声边;在特征层面引入通道级门控调制机制与非线性特征混合模块(FeatureMixPro),增强对特征扰动的适应性;训练中结合对抗生成与多目标优化,实现双路径协同防御。

1 结构-特征协同防御的图神经网络

本文提出的 SFCoRobustGNN 核心在于从结构性与特征性两个互补维度共同构建鲁棒性,模型框架如图 1 所示。该框架在 STABLE[13]结构优化思路的基础上,引入以下 3 项关键设计。

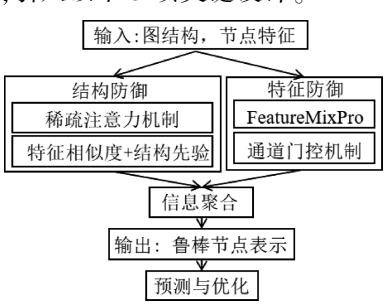

图 1 SFCoRobustGNN 框架图

Figure 1 SFCoRobustGNN framework diagram

1.1 稀疏注意力机制

传统的图注意力机制(GAT)仅依赖节点特征计算邻居权重,在遭受特征污染或伪造攻击时可能被误导[14]。为提升鲁棒性,本文在 GAT 的基础上设计了稀疏注意力机制,同时引入特征相似度和结构先验约束[15]。

结构先验 ϕ_{ij} 的选择是动态稀疏注意力机制的核心之一,需与图数据的语义特性相匹配。本文重点探讨 3 种广泛使用的先验度量如下。

(1)共邻居数(common neighbors,CN):

$$\phi_{ij}^{\mathrm{CN}} = |N(i) \cap N(j)|。 \tag{1}$$

该指标假设共享邻居越多的节点间关联越强,适用于连接关系清晰、且共同邻居信息具有判别性的网络(如社交网络 Polblogs)。

(2)Jaccard 系数(Jaccard coefficient,JC):

$$\phi_{ij}^{\mathrm{JC}} = |N(i) \cap N(j)| / |N(i) \cup N(j)|。 \tag{2}$$

该系数通过归一化缓解了节点度偏差,更能捕捉节点间的结构相似性,适用于节点属性相似性至关重要的网络(如引文网络 Cora、Citeseer)。

(3)Adamic-Adar 指数(Adamic-Adar,AA):

$$\phi_{ij}^{\mathrm{AA}} = \sum_{z \in N(i) \cap N(j)} 1/\log |N(z)|。 \tag{3}$$

该指数为稀少共同邻居赋予更高权重,适用于识别异质网络中具有稀缺性价值的隐含关系(如知识图谱)。

给定节点 i 与 j 的特征向量 $\boldsymbol{h}_i, \boldsymbol{h}_j \in \mathbf{R}^d$,未归一化注意力打分定义为

$$e_{ij} = \mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i \| \boldsymbol{W}\boldsymbol{h}_j]) - \lambda(1 - \mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)) - \mu(1 - \phi_{ij})。 \tag{4}$$

式中:$\boldsymbol{W}$ 为可学习的线性变换矩阵;$\boldsymbol{a}$ 为注意力权重向量;$\|$ 表示向量拼接操作;$\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)$ 表示特征相似度(如余弦相似度);ϕ_{ij} 表示结构先验;λ、μ 均为超参数,用于控制对低相似度和低先验可信度边的惩罚。

归一化后得到注意力系数:

$$\alpha_{ij} = \exp(e_{ij}) \Big/ \sum_{k \in N(i)} \exp(e_{ik})。 \tag{5}$$

节点更新则为加权聚合形式:

$$\boldsymbol{h}'_i = \sigma\Big(\sum_{j \in N(i)} \alpha_{ij} \boldsymbol{W}\boldsymbol{h}_j\Big)。 \tag{6}$$

式中:$\sigma(\cdot)$ 为 Sigmoid 激活函数。

该机制在聚合邻域信息时不仅考虑特征相似性,还结合结构先验进行双重约束。当边的特征相似度低或结构先验不足时,其权重将被动态压制。该设计克服了“仅依赖特征相似度”的弱点,使得模型在特征污染或伪造攻击下依然能够保持稳定。与静态边剪枝方法相比,该稀疏注意力具备在线自适应调节能力,可在动态扰动条件下持续发挥防御作用。

1.2 通道门控机制

为减少节点特征中低质量或噪声成分的影响,

引入通道级门控调制机制,对节点表示的每一维度重要性进行动态建模,从而增强关键特征并抑制干扰信息。对于每个节点的中间表示向量 $\boldsymbol{z}_i \in \mathbf{R}^d$,本文使用一个共享的门控网络生成通道注意力向量 $\boldsymbol{g}_i \in \mathbf{R}^d$,形式如下所示:

$$\boldsymbol{g}_i = \sigma(\boldsymbol{U}_g \boldsymbol{z}_i + \boldsymbol{b}_g)。 \tag{7}$$

式中:$\boldsymbol{U}_g$ 与 $\boldsymbol{b}_g$ 分别为可学习的投影矩阵和偏置向量,保证门控权重在[0,1]。最终输出通过门控方式进行通道加权融合:

$$\tilde{\boldsymbol{z}}_i = \boldsymbol{g}_i \odot \boldsymbol{z}_i。 \tag{8}$$

式中:$\odot$ 表示逐元素乘法。

该机制本质上对节点的每个通道维度进行软选择,从而抑制低质量或高噪声的特征维度,仅保留对分类任务具有判别性的表示信息。与传统 Dropout 不同,门控机制是一种数据驱动的自适应调制策略,能够针对通道维度的重要性动态建模,从而在特征扰动或遮蔽情况下抑制噪声传播。

1.3 特征混合增强策略

为建模复杂的特征扰动,本文设计了增强型特征混合策略 FeatureMixPro,该策略通过非线性混合机制与类型条件混合相结合,并采用自适应调度训练策略,共同提升模型鲁棒性。

1.3.1 非线性混合与类型条件机制

给定一批次内的节点对特征 $(\boldsymbol{h}_i, \boldsymbol{h}_j)$ 及其对应的节点类型 (t_i, t_j)(对于同构图,所有节点类型相同),其混合表示生成如下:

$$\tilde{\boldsymbol{h}} = f_\phi(\gamma \boldsymbol{W}_{t_i} \boldsymbol{h}_i + (1-\gamma)\boldsymbol{W}_{t_j} \boldsymbol{h}_j)。 \tag{9}$$

式中:f_ϕ 为一个简单的非线性前馈网络,负责将线性混合后的特征映射到更具判别性的非线性空间;$\boldsymbol{W}_t$ 为可学习的类型特定投影矩阵,当 $t_i = t_j$ 时,使用同一矩阵,当 $t_i \neq t_j$ 时,使用不同的矩阵,以此保障不同类型节点在混合前被投影到可比的语义空间,避免“苹果与橘子”式的无效混合。

所设计的非线性投影网络 $\boldsymbol{f}_\phi$ 的核心作用在于学习一个扰动不变的表示空间,以模拟真实世界中的复杂非线性扰动。以分子图为例,一个“羟基(—OH)”与一个“甲基($—CH_3$)”的原始特征向量在输入空间中可能相距甚远。简单的线性插值会生成大量无化学意义的中间特征。而 f_ϕ 通过其非线性变换能力,能够学习到这两种官能团在特定上下文环境下的高阶语义表示,即将它们均映射到“可置换烷基”这一共享的语义概念附近,从而有效模拟“官能团置换”这一复杂扰动,增强模型鲁棒性。其数学本质是学习一个非线性映射 $f_\phi: \mathbf{R}^d \to \mathbf{R}^k$,使得

$$\| f_\phi(\boldsymbol{h}_{OH}) - f_\phi(\boldsymbol{h}_{CH_3}) \| < \| \boldsymbol{h}_{OH} - \boldsymbol{h}_{CH_3} \|。 \tag{10}$$

1.3.2 自适应调度训练策略

为平衡鲁棒性与收敛稳定性,混合强度 β 随训练周期衰减。具体地,在第 e 轮:

$$\beta_e = \beta_{initial} \cdot \exp(-e/T_{decay})。 \tag{11}$$

这使得训练早期积极进行特征增强,而后期逐渐减弱,让模型专注于清洁数据的拟合。

1.4 损失函数

模型整体通过交叉熵损失进行训练:

$$\mathscr{L}_{CE} = -\sum_{i \in V_L} \boldsymbol{y}_i^{\mathrm{T}} \log \hat{\boldsymbol{y}}_i。 \tag{12}$$

式中:V_L 表示有标签节点集合。

通过稀疏注意力机制、通道门控机制和 FeatureMixPro 的协同作用,SFCoRobustGNN 在结构层与特征层实现了噪声抑制与特征增强。

2 实验验证

本文通过系统实验,综合评估 SFCoRobustGNN 在多种扰动类型与攻击强度下的鲁棒性表现,验证其在图智能环境中应对结构与特征扰动的综合能力。

2.1 实验设置

2.1.1 数据集

如表 1 所示,本文在 5 个经典图分类数据集上进行评估,确保全面比较各种扰动情况下的鲁棒性表现。其中,对于 Polblogs 数据集,由于缺乏原生特征,本文遵循文献[13]的标准,使用单位矩阵作为输入特征。

表 1 图分类数据集

Table 1 Graph classification datasets

数据集	节点数	边数	类别数	特征维度
Cora[16]	2 485	5 069	7	1 433
Citeseer[17]	2 110	3 668	6	3 703
Polblogs[18]	1 222	16 714	2	
PubMed[19]	19 717	44 338	3	500
ogbn-products[20]	2 449 029	61 859 140	47	100

2.1.2 攻击方法

为全面评估模型在不同扰动类型下的鲁棒性,本文设计了结构扰动攻击与特征扰动攻击两类攻击场景。

(1)结构扰动攻击。此类攻击旨在篡改图的结构信息(邻接矩阵)。本文采用 3 种典型方法:Meta-Attack[21] 基于元学习优化的强攻击方法,能够生成有效的攻击边;DICE[22] 随机断开图中的边并建立跨类别的连接,模拟结构破坏;Random 完全随机地

增加和删除边,模拟无结构规律的攻击。

(2)特征扰动攻击。此类攻击旨在污染节点的特征信息(特征矩阵)。本文采用两种经典的对抗性特征攻击方法。PGD[23]在特征空间计算梯度,生成最大化模型分类损失的对抗性扰动,并投影到预设的扰动幅度约束 ε 内;FGSM[24]为一种快速生成对抗样本的方法,沿损失函数梯度的符号方向添加扰动。对于 PGD 和 FGSM 攻击,本文设置特征扰动的最大扰动幅度约束 $\varepsilon \in \{0.01, 0.05, 0.10\}$,以模拟不同强度的对抗性特征噪声。

2.1.3 对比方法

本文选择涵盖了结构修复、表示增强与训练优化 3 大类策略的方法作为性能对比基线方法:GCN[25]、GNNGuard[4]、RGCN[26]、ProGNN[6]、STABLE[13]。所有方法均在相同的数据划分与攻击扰动下进行评估,实验结果为 10 次独立运行的节点分类准确率平均值及标准差。

2.1.4 实验环境与参数设置

本文所有实验均在统一的高性能计算平台上完成,以确保结果的可比性与可复现性。硬件使用 2 个 Intel Xeon Platinum 8358P CPU 与 4 张 NVIDIA GeForce RTX 3090 GPU,软件环境基于 Ubuntu 20.04.6、PyTorch 1.13.1,并集成 DGL 0.9.1 与 PyG 2.3.0 图计算库,所有任务运行于 Docker 容器中。

模型隐藏层维度设为 256,采用 4 头稀疏注意力机制(每头 64 维,温度系数 $\tau = 0.5$)。训练使用 AdamW 优化器(学习率 0.001,权重衰减 10^{-4}),结合余弦退火调度与梯度裁剪(阈值 2.0),批量大小为 128,最大训练轮数为 500,并启用早停策略(耐心值 50)。鲁棒性相关参数包括:结构先验权重($\lambda_1 = 0.7, \lambda_2 = 0.3$)、特征混合系数(初始值 0.8,衰减率 0.95)、通道丢弃率为 0.2 以及扰动幅度 $\varepsilon = 0.01$。

2.2 实验结果

2.2.1 结构扰动攻击

针对 MetaAttack 攻击下的鲁棒性评估实验,在 Cora、Citeseer、Polblogs、PubMed 和 ogbn-products 这 5 个标准图数据集上,通过设置 0%~20%的边扰动比例作为扰动率,系统对比了 GCN、GNNGuard、RGCN、ProGNN、STABLE 及本文提出的 SFCoRobustGNN 模型的防御性能,如图 2 所示。

从图 2 可以看出,随着攻击强度增加,模型性能基本呈下降趋势,而 SFCoRobustGNN 在多数情况下保持最优或接近最优的准确率,在 Cora 的 20%攻击率下准确率达 79.70%,显著高于 GCN(60.31%);在 Citeseer 的 15% 和 20% 攻击率下分别达到 73.44%和 72.10%;在 Polblogs 的 20%攻击率下准确率为 89.94%,STABLE(88.46%)表现接近;在 PubMed 上始终保持最高或次高准确率;在 ogbn-products 的 20%攻击率下准确率为 71.82%,显著优于 GCN(54.12%) 和 GNNGuard(51.79%)。实验结果表明,SFCoRobustGNN 通过稀疏注意力机制动态抑制异常边,结合多目标优化与对抗训练策略,能够有效抵御不同类型图结构扰动。在不同规模和图特性的数据集上,该框架均表现出稳定且优越的防御性能,验证了其结构-特征协同防御机制的有效性。

本文采用 DICE 攻击方法对上述模型进行测试。实验同样在 5 个数据集上进行,攻击率设置为 20%~40%,结果如图 3 所示。

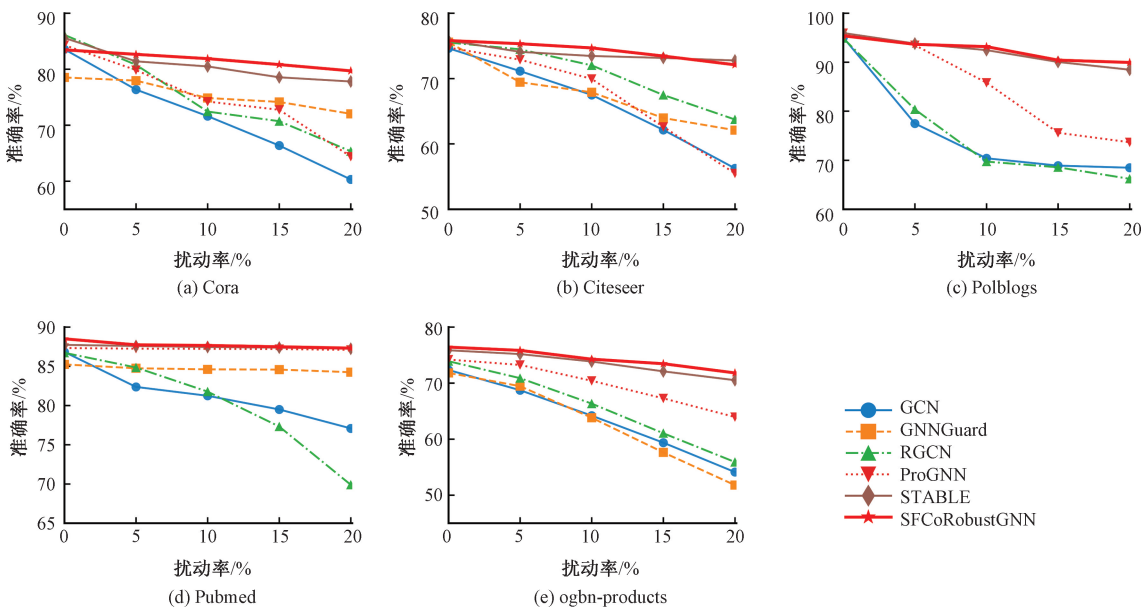

图 2 MetaAttack 攻击多数据集对比

Figure 2 Comparison of MetaAttack attacks on multiple datasets

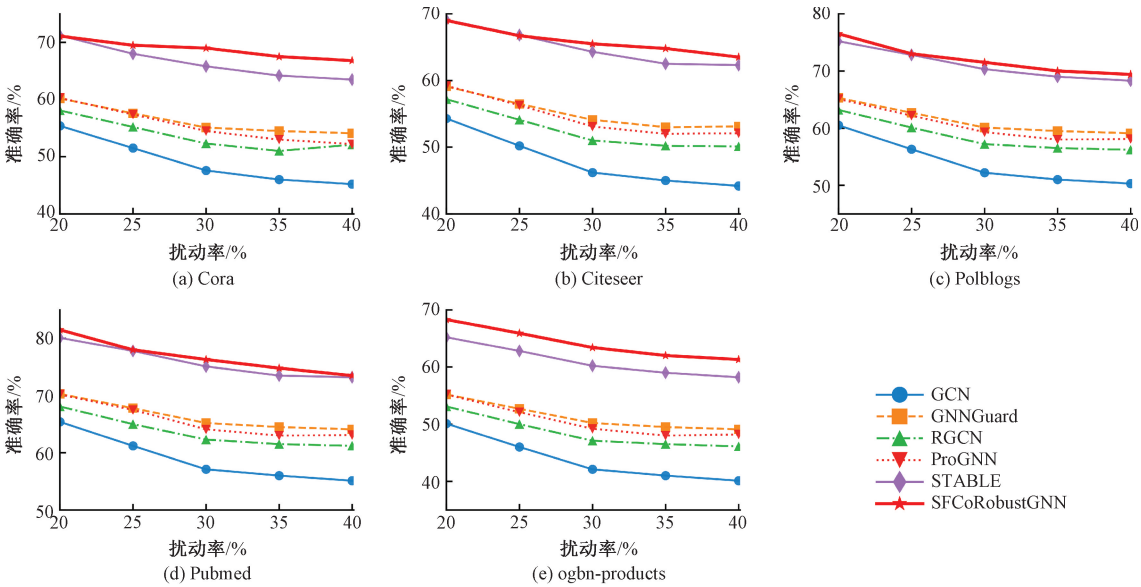

图 3　DICE 攻击多数据集对比

Figure 3　Comparison of DICE attacks on multiple datasets

从图 3 可以看出,随着攻击率的增加,模型的性能基本呈下降趋势。在 Cora 上,SFCoRobustGNN 与 STABLE 在 40%攻击率的准确率分别保持在 66.8%和 63.5%,显著优于 GCN(45.2%)及其他基线方法;在 Citeseer 上,SFCoRobustGNN 多数情况下优于或持平 STABLE,在 35%和 40%攻击率下分别达到 64.8%和 63.4%;在 Polblogs 上,SFCoRobustGNN 表现最佳,20%和 40%攻击率下准确率分别为 76.5%和 69.4%;在 PubMed 上,SFCoRobustGNN 准确率高于或接近 STABLE;在 ogbn-products 上,SFCoRobustGNN 优势最为明显,40%攻击率下准确率为 61.3%,高于 STABLE(58.2%)及其他基线方法。实验证明,面对非针对性攻击,SFCoRobustGNN 能有效缓解随机结构扰动,表现出优越且稳定的鲁棒性能。

为评估模型在非针对性攻击下的鲁棒性,本文进一步采用随机攻击策略,实验同样在 5 个数据集上进行,攻击率为 20%~40%,结果如图 4 所示。从图 4 可以看出,随着攻击率的增加,模型性能基本都呈下降趋势,但 SFCoRobustGNN 在多数情况下仍保持最优或接近最优,在 Cora 上的准确率在 20%和 40%攻击率下分别为 77.2%和 69.1%,优于 GCN 等基线方法;在 Citeseer 上,30%攻击率下准确率达 72.4%,高于 STABLE 的 69.2%;在 Polblogs 上表现

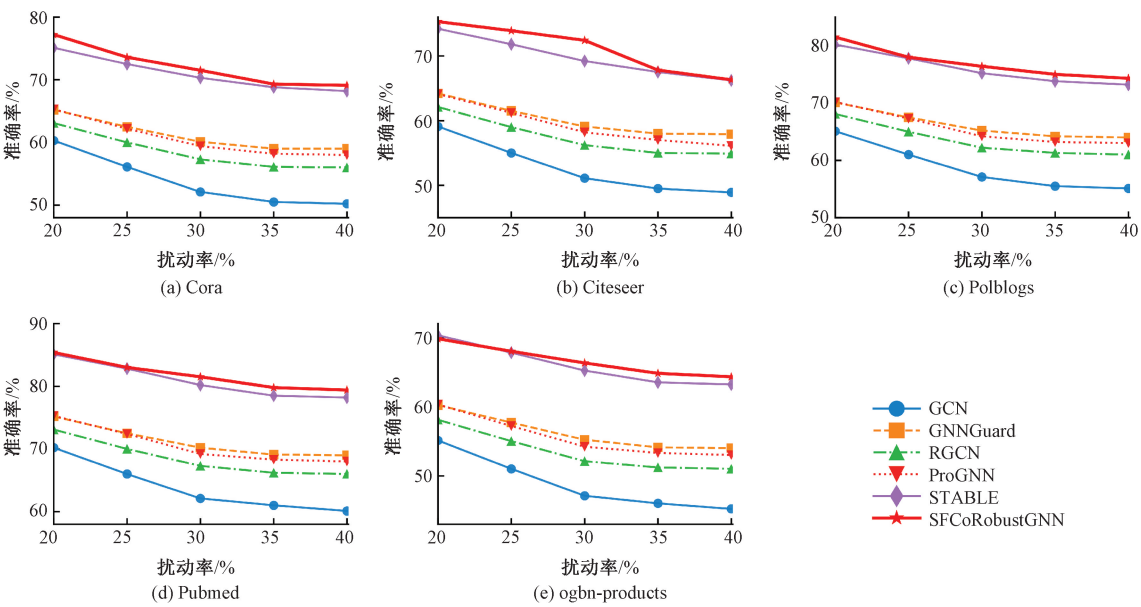

图 4　Random 攻击多数据集对比

Figure 4　Comparison of Random attacks on multiple datasets

最佳,20%和40%攻击率下分别为81.5%和74.3%;在PubMed上性能最佳;在ogbn-products上,高攻击率(35%~40%)下准确率分别为64.8%和64.3%,优于其他方法。实验结果表明,SFCoRobustGNN在面对随机结构扰动时具有较强的抗干扰能力,稀疏注意力机制能够动态适应无规律攻击模式,结合多目标优化训练,验证了该框架在复杂扰动环境下的稳定性。

2.2.2 特征扰动攻击

本文采用PGD攻击方法评估模型在连续特征攻击下的鲁棒性,结果如表2所示。实验结果显示,随着ε的增大,所有模型性能均有所下降,而SFCoRobustGNN在多数情况下仍保持最优或接近最优的准确率,在Cora上,$\varepsilon=0.01$时准确率为80.15%,显著优于GCN等基线方法,$\varepsilon=0.10$时性能与STABLE相当且标准差更低;Citeseer上,SFCoRobustGNN在各级攻击下均取得最高准确率,且方差最低;在PubMed上,其表现与STABLE相近,并保持最低方差;在ogbn-products上,中高攻击强度($\varepsilon=0.05,0.10$)下其准确率(72.62%、68.56%)优于传统GCN及部分基线方法。结果表明,SFCoRobustGNN在面对PGD攻击时不仅能维持较高分类性能,还表现出更稳定的预测行为,验证了其在特征扰动下的强鲁棒性。

本文采用FGSM攻击方法评估模型在快速对抗攻击下的鲁棒性,结果如表2所示。实验表明,随着ε的增大,所有模型性能均下降,而SFCoRobustGNN在多数情况下仍保持最优或接近最优,在Cora上,$\varepsilon=0.01$时的准确率为81.64%,显著优于GCN等基线方法;$\varepsilon=0.10$时为74.92%,优于所有对比方法;在Citeseer上,$\varepsilon=0.05$时SFCoRobustGNN的准确率最高(76.84%),且方差始终较低。在PubMed上的表现与STABLE相近,同时标准差更低,稳定性更优;在ogbn-products上,高强度攻击下准确率为70.85%,优于其他方法。结果表明,SFCoRobustGNN在FGSM攻击下具有更高的分类准确率和更稳定的预测性能,展现出良好的对抗鲁棒性。

2.3 结构先验选择对比分析

为系统评估不同结构先验对图神经网络鲁棒性

表2 PGD和FGSM攻击下各模型节点分类准确率

Table 2 Node classification accuracy of each model under PGD and FGSM attack

攻击方法	数据集	ε	准确率/%					
			GCN	GNNGuard	RGCN	ProGNN	STABLE	SFCoRobustGNN
PGD	Cora	0.01	71.32	75.68	73.45	76.82	79.41	**80.15**
		0.05	63.17	70.24	68.93	72.56	75.83	**76.30**
		0.10	54.26	64.37	62.18	67.42	**70.15**	69.42
	Citeseer	0.01	69.84	73.20	71.67	74.93	77.62	**78.73**
		0.05	61.53	68.42	66.84	70.27	73.85	**74.16**
		0.10	52.67	62.18	60.32	65.74	69.43	**70.28**
	PubMed	0.01	80.15	83.27	82.06	84.62	**86.93**	86.24
		0.05	74.38	79.16	77.84	80.75	83.42	**84.73**
		0.10	67.92	74.35	72.61	76.83	79.67	**79.95**
	ogbn-products	0.01	65.73	69.42	68.15	71.86	**74.28**	73.85
		0.05	60.24	65.37	63.82	67.45	70.63	**72.62**
		0.10	53.16	60.83	58.94	63.27	67.83	**68.56**
FGSM	Cora	0.01	73.85	77.12	75.64	78.93	80.25	**81.64**
		0.05	67.42	72.86	72.25	74.67	**77.43**	77.27
		0.10	59.73	68.54	68.82	70.93	73.86	**74.92**
	Citeseer	0.01	72.46	75.83	74.27	77.62	**79.54**	79.35
		0.05	65.38	70.27	68.93	72.85	75.67	**76.84**
		0.10	57.62	66.73	64.85	69.42	**72.18**	72.13
	PubMed	0.01	82.67	85.34	84.16	86.72	**88.45**	88.26
		0.05	77.83	81.75	80.42	83.64	85.27	**85.73**
		0.10	71.45	77.62	75.83	79.86	81.94	**82.42**
	ogbn-products	0.01	68.92	72.45	71.18	74.63	76.85	**77.24**
		0.05	63.75	68.37	66.92	70.84	73.42	**73.63**
		0.10	56.83	63.27	61.45	66.18	69.75	**70.85**

的影响,本文在 Cora(引文网络)、Polblogs(社交网络)及 ogbn-products(大规模异质商品网络)3 个具有代表性的数据集上,在 20%攻击率的 MetaAttack 攻击下,对比分析了共邻居数(CN)、Jaccard 系数(JC)及 Adamic-Adar 指数(AA)这 3 种先验机制的防御性能。结果如表 3 所示,可知结构先验的性能与图数据的语义特性密切相关。在引文网络 Cora 中,Jaccard 系数凭借其归一化优势有效抑制节点度偏差,取得最高准确率 80.2%;在社交网络 Polblogs 中,共邻居数通过直接捕捉显式强关联,表现出最佳鲁棒性 89.5%;而在异质图 ogbn-products 上,Adamic-Adar 指数通过挖掘稀少但高价值共现关系,达到最高准确率 72.1%。

表 3　不同结构先验在不同类型数据集上的性能对比

Table 3　Performance comparison of structure priors on different types of dataset

数据集	先验类型	准确率/%	标准差
Cora	CN	78.1	±1.2
	JC	**80.2**	±0.8
	AA	79.6	±0.9
Polblogs	CN	**89.5**	±0.7
	JC	87.1	±1.1
	AA	88.3	±1.0
ogbn-products	CN	70.8	±0.7
	JC	71.5	±0.6
	AA	**72.1**	±0.5

基于上述结果,本文提出如下先验选择准则为图结构先验的选择提供实证依据与系统指导:属性相似性驱动的同质图(如引文网络)宜优先选用 JC;连接驱动型网络(如社交网络)推荐 CN;而异质大规模图(如电商网络)可优先尝试 AA。

2.4　特征分布对比实验

为验证 FeatureMixPro 模块在对抗性攻击下的特征净化效能,本文在 Cora 数据集上进行量化分析。在 PGD 攻击($\varepsilon=0.10$)条件下,对比了线性混合与 FeatureMixPro 非线性变换的特征空间质量,结果如表 4 所示。结果表明,FeatureMixPro 通过非线性变换根本性改善了特征空间结构。轮廓系数从负值转为正值,证明特征从不可聚类状态转变为可有效区分;平均类间距激增 2 784.0%,显示不同类别被显著分离;虽然平均类内距增加 161.4%,但其增幅远低于平均类间距,净效应体现为特征判别性的实质性提升。实验结果有力证实了 FeatureMixPro 模块能够有效净化特征噪声,重构出具有高判别性的鲁棒特征表示,为模型在对抗攻击下的稳定性能提供了理论支撑。

表 4　PGD 攻击下特征空间指标定量对比

Table 4　Feature space quantity comparison in PGD attack

评价指标	指标定量	
	线性混合	FeatureMixPro
轮廓系数	−0.032 3	**0.019 8**
平均类内距	2.769 1	**7.237 1**
平均类间距	0.148 4	**4.279 9**

2.5　消融实验

为验证 SFCoRobustGNN 中各模块的有效性及其协同作用,本文在 Cora 数据集上进行 MetaAttack 结构扰动攻击的消融实验,结果如表 5 所示。由表 5 可知,所提出的各模块均对模型鲁棒性具有重要贡献,且模块间存在显著的协同增强效应。

表 5　消融实验结果

Table 5　Ablation experiment results

模型变体	不同扰动率下的准确率/%			
	5%	10%	15%	20%
完整模型	**82.65**	**81.88**	**80.82**	**79.70**
无结构先验	78.35	73.61	69.87	67.20
无特征混合	82.54	79.45	76.83	73.80
无门控机制	81.01	78.92	75.16	70.50
无结构先验+无特征混合	75.28	70.15	65.43	62.10

在不同的扰动率下,完整模型 SFCoRobustGNN 都可以保持最高的节点分类准确率。在扰动率为 20%时,去除结构先验约束模块后,准确率下降了 12.50 百分点,说明结构先验对维持图语义完整性具有关键作用;移除特征混合模块导致准确率下降 5.90 百分点,表明该模块有效增强了节点表示对扰动的适应性;而禁用门控机制导致准确率下降 9.2 百分点,验证了该机制在重要邻居筛方面的有效性。当同时去除结构先验和特征混合模块时,模型性能急剧下降至 62.10%,准确率损失高达 17.60 百分点,该损失显著高于单一模块单独消融的损失,充分证明各组件并非独立起作用,而是存在强烈的协同效应,共同构成了模型的鲁棒性基础。消融实验结果表明,SFCoRobustGNN 中各模块均是不可或缺的,且整体设计在有效抵御图结构攻击方面表现出高度的协同性与稳定性。

3　结论

本文提出了一种面向结构与特征扰动协同攻击的鲁棒图神经网络 SFCoRobustGNN。该框架通过三重机制实现协同防御:结构上采用融合特征相似度与结构先验的稀疏注意力机制,抑制异常连接与噪

声传播;特征层面引入通道级门控调制机制与增强型非线性混合模块 FeatureMixPro,实现特征污染的自适应净化与语义增强;训练中融合对抗生成与多目标优化,提升模型在复杂扰动下的泛化能力。

在 Cora、Citeseer 等多个图数据集上的实验表明,本方法在结构攻击与特征攻击下均表现最优。在不同强度的结构与特征扰动下,模型性能显著优于基线方法,展现出优异的鲁棒性。消融实验进一步验证了结构先验与特征混合等核心模块各自的重要贡献及其协同效应。本文在理论上为图鲁棒学习提供了双重约束与类型感知混合的新思路,在应用上具备端到端、高效、易集成的优点,适用于社交网络、推荐系统和生物医学等高安全场景。

尽管表现出色,本方法仍存在局限性:在高稀疏图中结构先验信息不足可能导致约束失效;特征维度方差极不平衡时门控机制可能失准;当前框架未涵盖异构图与动态图场景。未来工作将探索图补全技术、特征选择策略,并扩展框架以支持异构与动态图数据,进一步提升算法的普适性与鲁棒性。

参考文献:

[1] Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2847-2856.

[2] Zhu Yanqiao, Xu Yichen, Yu Feng, et al. Deep graph contrastive representation learning[PP/OL]. V2. arXiv (2020-07-13) [2025-09-30]. https://arxiv.org/abs/2006.04131.

[3] Battaglia P W, Hamrick J B, Bapst V, et al. Relational inductive biases, deep learning, and graph networks[PP/OL]. V3. arXiv(2018-10-17)[2025-09-30]. https://arxiv.org/abs/1806.01261.

[4] Zhang Xiang, Zitnik M. GNNGuard: defending graph neural networks against adversarial attacks[PP/OL]. V3. arXiv(2020-10-28)[2025-09-30]. https://arxiv.org/abs/2006.08149.

[5] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[PP/OL]. V3. arXiv(2020-07-01)[2025-09-30]. https://arxiv.org/abs/2002.05709.

[6] Jin Wei, Ma Yao, Liu Xiaorui, et al. Graph structure learning for robust graph neural networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2020: 66-74.

[7] Li Gege, Ye Zhonglin, Cao Shujuan, et al. An unsupervised link prediction algorithm based on an approximate graph neural network framework[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(6): 75-82. [李格格, 冶忠林, 曹淑娟, 等. 一种近似图神经网络框架的无监督链路预测算法[J]. 郑州大学学报(工学版), 2024, 45(6): 75-82.]

[8] Veličković P, Fedus W, Hamilton W L, et al. Deep graph infomax[PP/OL]. V2. arXiv (2018-12-21) [2025-09-31]. https://arxiv.org/abs/1809.10341.

[9] Fanseu Kamhoua B, Zhang Lin, Ma Kaili, et al. GRACE: A general graph convolution framework for attributed graph clustering[J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(3): 1-31.

[10] Zhu Yanqiao, Xu Yichen, Yu Feng, et al. Graph contrastive learning with adaptive augmentation[C]//Proceedings of the Web Conference 2021. New York: ACM, 2021: 2069-2080.

[11] Thakoor S, Tallec C, Azar M G, et al. Large-scale representation learning on graphs via bootstrapping[PP/OL]. V3. arXiv (2023-02-20)[2025-09-30]. https://doi.org/10.48550/arXiv.2102.06514.

[12] Liu Xiaorui, Jin Wei, Ma Yao, et al. Elastic graph neural networks[PP/OL]. V1. arXiv (2021-07-05)[2025-09-30]. https://arxiv.org/abs/2107.06996.

[13] Li Kuan, Liu Yang, Ao Xiang, et al. Reliable representations make a stronger defender: unsupervised structure refinement for robust GNN[C]//Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 925-935.

[14] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[EB/OL]. V3. arXiv (2018-02-04) [2025-09-30]. https://arxiv.org/abs/1710.10903.

[15] Egressy B, Von Niederhäusern L, Blanuša J, et al. Provably powerful graph neural networks for directed multigraphs[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(10): 11838-11846.

[16] Ahmed N, Rossi R A, Zhou Rong. Cora[DS/OL]. [2025-10-10]. http://networkrepository.com/cora.php.

[17] Ahmed N, Rossi R A, Zhou Rong. Citeseer[DS/OL]. [2025-10-10]. http://networkrepository.com/citeseer.php.

[18] Adamic L, Glance N. Polblogs[DS/OL]. [2025-10-10]. http://networkrepository.com/polblogs.php.

[19] Namata G M, London B, Getoor L, et al. PubMed diabetes dataset[DS/OL]. [2025-10-10]. https://linqs.org/datasets/#pubmed-diabetes.

[20] HU Weihua, Fey M, Zitnik M, et al. Dataset ogbn-prod-

ucts[DS/OL]. Open Graph Benchmark, 2020. https://ogb. stanford. edu/docs/nodeprop/#ogbn-products.

[21] Zügner D, Borchert O, Akbarnejad A, et al. Adversarial attacks on graph neural networks: perturbations and their patterns[J]. ACM Transactions on Knowledge Discovery from Data, 2020, 14(5): 1-31.

[22] Waniek M, Michalak T P, Wooldridge M J, et al. Hiding individuals and communities in a social network[J]. Nature Human Behaviour, 2018, 2(2): 139-147.

[23] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[PP/OL]. V4. arXiv (2019-09-04)[2025-10-10]. https://doi.org/10.48550/arXiv.1706.06083.

[24] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [PP/OL]. V3. arXiv (2015-03-20)[2025-09-30]. https://arxiv.org/abs/1412.6572.

[25] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [PP/OL]. V4. arXiv (2017-02-22)[2025-09-30]. https://arxiv.org/abs/1609.02907.

[26] Zhu Dingyuan, Zhang Ziwei, Cui Peng, et al. Robust graph convolutional networks against adversarial attacks [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 1399-1407.

A Graph Neural Network for Structure-Feature Collaborative Defense

HAN Jihui[1], SHI Yupeng[1], HUANG Ziqi[2], ZHANG Anlin[3], HUANG Daoying[1]

(1. College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450001, China; 2. North Information Control Research Academy Group Co., Ltd., Nanjing 211153, China; 3. Engineering Training Center, Zhengzhou University of Light Industry, Zhengzhou 450001, China)

Abstract: To address the degradation of node representations in graph neural networks under complex perturbation environments, a structure-feature collaborative defense graph neural network named SFCoRobustGNN was proposed. Structurally, a sparse attention mechanism that integrated structure priors to dynamically suppress anomalous edges was introduce. Feature-wise, a channel gating mechanism was combined with a nonlinear feature mixing module (FeatureMixPro) to enhance the model's adaptability to feature perturbations. A collaborative dual-pathway defense was achieved through adversarial training and a multi-objective optimization strategy. Experiments on multiple benchmark datasets, including Cora and Citeseer, demonstrated that the proposed method outperformed most mainstream baseline metods under various intensities of structure perturbations (5%-40%) and feature attacks (ε=0.01-0.10), showing significant improvement in node classification accuracy. On the large-scale ogbn-products dataset, it maintained an accuracy of 71.82% even under a 20% MetaAttack structure perturbation, demonstrating its strong scalability. Ablation studies validated the effectiveness and synergistic effects of each module. The proposed method effectively mitigated performance degradation under complex perturbations and exhibited excellent generalization.

Keywords: graph neural networks; robustness; structure perturbation; feature perturbation; sparse attention