



研究与开发

基于深度学习的多样化复杂网络影响力节点识别

马玉磊, 郭莎莎

(新乡学院计算机与信息工程学院, 河南 新乡 453000)

摘要: 为提高多样化复杂网络中影响力节点识别的准确性和鲁棒性, 提出一种基于深度学习的多样化复杂网络影响力节点识别方法。首先, 采用多个中心性指标从不同方面评估节点在网络拓扑结构中的重要性, 通过可学习权重向量自适应地决定不同复杂网络中各指标的权重; 接着, 提出一种支持不同特征维度的 Transformer 框架; 最后, 利用 Transformer 模型对不同距离的邻居信息进行分级聚合, 以提取邻域的上下文信息。在多种复杂网络数据集上完成了验证实验, 结果表明, 所提方法在不同规模、不同类型的复杂网络上均取得了较好的影响力节点识别性能, 有效提高了影响力节点识别的准确性和鲁棒性。

关键词: 复杂网络; 深度学习; 自注意力机制; 中心性度量; 影响力节点

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2025107

Influential nodes recognition of diverse complex network based on deep learning

MA Yulei, GUO Shasha

Department of Computer and Information Engineering, Xinxiang University, Xinxiang 453000, China

Abstract: To improve the accuracy and robustness of influential node recognition in diverse complex networks, a deep learning-based recognition method for influential nodes in diverse complex networks was proposed. Firstly, multiple centrality indexes were utilized to evaluate the importance of network topology from different perspectives, the weight of each index in different complex networks was decided adaptively through the learnable weight vector. Secondly, a new Transformer framework that could handle features of different dimensions was proposed. Finally, the Transformer model was employed to realize hierarchical aggregation of the neighbor information in different distances, so as to extract the contextual information of the neighborhood. Validation experiments were carried on multiple complex network datasets, the results showed that the proposed method achieved a good recognition performance of influential nodes for the complex networks of different scales and different categories, effectively improving

收稿日期: 2024-11-05; 修回日期: 2025-04-07

通信作者: 马玉磊, mayulei8219@163.com

基金项目: 河南省科技厅重点研发与推广专项(科技攻关)(No.212102210405); 新乡学院教育教学改革研究与实践项目(No.31)

Foundation Items: The Key Research & Development and Promotion Special Project (Science and Technology Key Technology Research) of Henan Provincial Department of Science and Technology (No.212102210405), Xinxiang University's Teaching Reform Research and Practice Project (No.31)

the accuracy and robustness of influential node recognition.

Key words: complex network, deep learning, self-attention mechanism, centrality measurement, influential node

0 引言

复杂网络可用于表示真实世界的一些复杂系统,如社交网络^[1]、蛋白质互作网络^[2]以及国际贸易网络等。复杂网络的节点间存在错综复杂的关系,其中一部分节点对网络结构和网络功能具有更大的影响力,因此,在复杂网络中发现与识别这些具有影响力的节点既具有理论意义,又具有实践价值^[3]。

节点中心性^[4]是一种判断节点影响力或重要性的常用指标,它主要从网络拓扑结构方面评估节点的重要性。节点中心性指标可分为3种类型,即局部中心性、全局中心性和半局部中心性。局部中心性不考虑网络拓扑的总体分布,得到的影响力值往往低于真实值,如度中心性^[5]和亲密度中心性^[6]。全局中心性的度量结果优于局部中心性,但其计算复杂度高,无法应用于大规模复杂网络,如特征向量中心性^[7]。半局部中心性结合直接邻居与有限跳数间接邻居来评估节点影响力,使准确性和计算效率能够达到较好的平衡,如介数中心性^[8]和半局部中心性^[9]。针对节点中心性指标的局限性,文献[10]基于信息熵和K-shell方法提出一种新的节点影响力度量方法,该方法根据节点周围的拓扑信息测量节点的传播能力,以避免K-shell法的“富人俱乐部现象”。文献[11]提出一种基于交叉熵的节点重要性度量方法,该方法将节点邻域拓扑信息与中心节点信息融合,利用交叉熵来量化节点信息的差异。文献[10]和文献[11]分别利用信息熵和交叉熵量化节点的特征信息,再结合节点邻域拓扑结构评估节点的重要性,由于熵无法充分利用节点的相关信息,因此,此类方法难以充分聚合邻域节点的信息。

随着人工智能技术的发展,研究人员将深度神经网络用于复杂网络节点的信息提取与聚合,

其中图神经网络(graph neural network, GNN)和图卷积网络(graph convolutional network, GCN)的性能较为突出。总体而言, GNN 和 GCN 利用节点间连接逐层进行信息传递与聚合,在聚合过程中对不同邻居节点的贡献加以区分,可有效捕获网络的拓扑结构信息和关系信息。文献[12]提出了融合项目影响力的 GNN,文献[13]利用 GNN 提取节点影响力和节点间拓扑结构信息,文献[14]提出了基于多层 GCN 的影响力节点识别方法,文献[15]提出了基于正则等价相似性 GCN 的影响力节点识别方法。文献[12-15]将 GNN 和 GCN 应用于复杂网络影响力识别问题,取得了良好的效果,但 GNN 和 GCN 仍存在不足之处:二者均要求节点特征的维度固定且统一,因此无法适用于多样化的复杂网络;在经过多轮迭代更新后,二者均会出现节点特征过平滑现象,因此所能结合的邻域范围有限。

综上,复杂网络中节点影响力受多种因素影响,通过单一度量指标不足以准确评估不同网络的节点影响力,且 GNN 和 GCN 存在节点特征维度固定和过平滑现象,难以处理多样化的复杂网络。为提高多样化复杂网络中影响力节点识别的准确性和鲁棒性,本文提出一种基于改进 Transformer 模型的多样化复杂网络影响力节点识别方法。本文的主要贡献可总结为3点。

(1) 采用多个中心性指标从不同方面评估节点在网络拓扑结构中的中心性,通过可学习权重自适应地决定不同网络指标的重要性。

(2) 提出一种改进的 Transformer (improved Transformer, ImTf) 模型来聚合邻域节点的特征信息, ImTf 模型支持不同的节点特征维度。

(3) ImTf 模型对不同距离的邻居信息进行分级聚合,在聚合过程中利用 Transformer 编码器提



取目标节点的邻域上下文信息, 并且 ImTf 模型的位置嵌入机制和多头注意力机制可避免 GNN 和 GCN 的过平滑问题。

本文所提方法集成网络拓扑结构信息、节点特征信息和邻域上下文信息, 以期提高多样化复杂网络影响力节点识别的准确性和鲁棒性。

1 系统框架设计

将复杂网络表示为一个无向无权图 $G=(V,E)$, 其中 V 和 E 分别为 G 的节点集和连接集, V 可表示为 $\{v_1, v_2, \dots, v_N\}$, 其中 N 为节点总数, E 可表示为 $\{e_{u,v}, \forall u \in V, \forall v \in V\}$, 其中 $e_{u,v}$ 表示节点 u 和 v 之间的连接, G 的边总数为 M 。若节点 u 与 v 之间存在连接, 那么 $e_{u,v}$ 等于 1, 否则等于 0。假设节点 u 的一跳邻居列表为 $\Gamma(u)$ 。从一个给定复杂网络可提取出两个矩阵, 即邻接矩阵 A 和节点特征矩阵 B 。 A 包含网络拓扑结构信息, A 的行向量 A_i 表示节点 v_i 与其他节点间是否存在连接, 元素 A_{ij} 表示节点 v_i 与 v_j 之间是否存在连接。 B 包含各节点的特征信息, 其行向量 B_i 表示节点 v_i 的特征向量。

影响力节点识别步骤如图 1 所示。由图 1 可知, 影响力节点识别主要包含以下 3 个步骤。

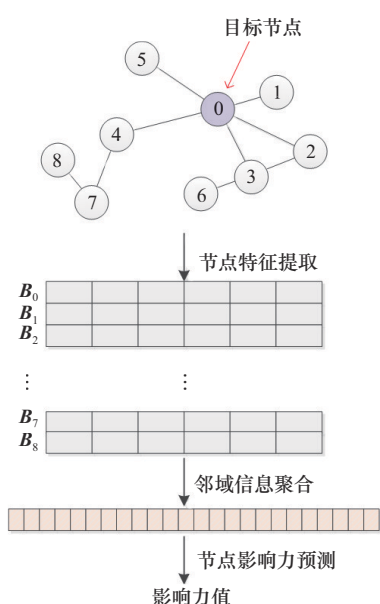


图1 影响力节点识别步骤

步骤1 将目标节点的索引设为0, 按广度优先搜索, 并建立其邻域节点的索引。

步骤2 提取目标节点和邻域所有节点的特征, 对邻域信息进行分级聚合。

步骤3 输出目标节点的影响力值。

2 影响力节点识别

2.1 节点特征提取

2.1.1 节点影响力度量

局部中心性、半局部中心性和全局中心性从不同角度对复杂网络的节点重要性进行评估, 由于全局中心性的计算复杂度随着网络规模扩大而大幅升高, 难以应用于大规模复杂网络, 因此本文仅使用局部中心性与半局部中心性度量节点的重要性。

局部中心性主要关注于节点的一跳邻居, 如 DC。DC 的计算式为:

$$DC(u) = \sum_{v \in V} e_{u,v} \quad (1)$$

半局部中心性在网络局部信息和全局信息之间取得平衡, 如 BC 和 SC。BC 评估了目标节点对网络传播的重要性, 其计算式为:

$$BC(u) = \sum_{v \neq w \neq u \in V} \frac{\delta_{v,w}(u)}{\delta_{v,w}} \quad (2)$$

其中, $\delta_{v,w}$ 为节点 v 与节点 w 之间最短路径的跳数, $\delta_{v,w}(u)$ 表示经过 u 的路径跳数。

SC 评估了该节点对其一跳邻居和二跳邻居的重要性, 其计算式为:

$$SC(u) = \sum_{v \in \Gamma(u)} \sum_{w \in \Gamma(v)} DC(w) \quad (3)$$

其中, $DC(w)$ 为节点 w 的度。

中心性度量仅考虑了节点与其邻居的连接情况, 并未考虑其邻域的节点密度, 而邻域节点密度和邻域相似性也对节点影响力具有关键作用。度数相同的节点在密集邻域和稀疏邻域内的影响力存在差异, 因此采用邻域密集度 (neighbor-

hood density degree, NDD) 来评估节点的邻域密度, 其计算式为:

$$\text{NDD}(v_i) = \frac{|\Gamma(v_i)|}{\sum_{v_j \in \Gamma(v_i)} |\Gamma(v_j)|} \quad (4)$$

其中, $\Gamma(v_i)$ 为 v_i 的所有一跳距离邻居节点。

共同邻居数 (common neighbor, CN) 用于评估节点与其邻居节点的共同邻居数, 其计算式为:

$$\text{CN}(v_i) = \sum_{v_j \in \Gamma(v_i)} |\Gamma(v_i) \cap \Gamma(v_j)| \quad (5)$$

其中, $|\Gamma(v_i) \cap \Gamma(v_j)|$ 表示节点 v_i 与其邻居节点 v_j 的共同邻居数。

2.1.2 特征向量建立

由于 DC、BC、SC、NDD 和 CN 的取值范围不同, 因此采用最小-最大归一化 (min-max normalization) 将每种度量的原始数据转换到 [0,1] 范围内。假设 $\mathbf{B}_i = [B_{i,1}, B_{i,2}, \dots, B_{i,Z}]$ 为节点 v_i 的归一化影响力特征向量, $B_{i,z}$ 表示节点 v_i 的第 z 个影响力指标。本文的影响力特征向量 $\mathbf{B}_i$ 依次为 DC(u)、BC(u)、SC(u)、NDD(G_u) 和 CN(G_u), 分别可通过式 (1) ~ 式 (5) 计算得到。

2.2 邻域信息聚合

邻域信息聚合流程如图2所示, 从网络邻接矩阵 $\mathbf{A}$ 出发, 先按广度优先搜索目标节点邻域内的节点集, 再按照节点与目标节点的距离将邻域节点排列成一个有序序列, 其中目标节点位于序列首位。邻域信息聚合过程主要包含3个模块, 即分集嵌入模块、Transformer编码器和分级注意力聚合模块。

分集嵌入模块从节点特征矩阵 $\mathbf{B}$ 提取目标节点邻域内所有节点的特征向量, p_i 为目标节点邻域节点序列生成分级的位置 Token, 每个节点的位置 Token 与特征向量拼接成位置嵌入向量。

Transformer 编码器对位置嵌入向量进行特征提取, 所提取特征中包含目标节点与邻域节点间的相关性。每个 Transformer 编码器由若干个结构相同的 Transformer 编码块级联而成, 多头自注意力 (multi-head self-attention, MSA) 机制是 Transformer 的核心模块, MSA 并行捕获位置嵌入向量中 Token 间的前后相关性, 由此可捕获目标节点与邻域内各节点间的相关性。

分级注意力聚合模块为不同距离的邻居节点赋予不同的权重。Transformer 编码器将邻域内所

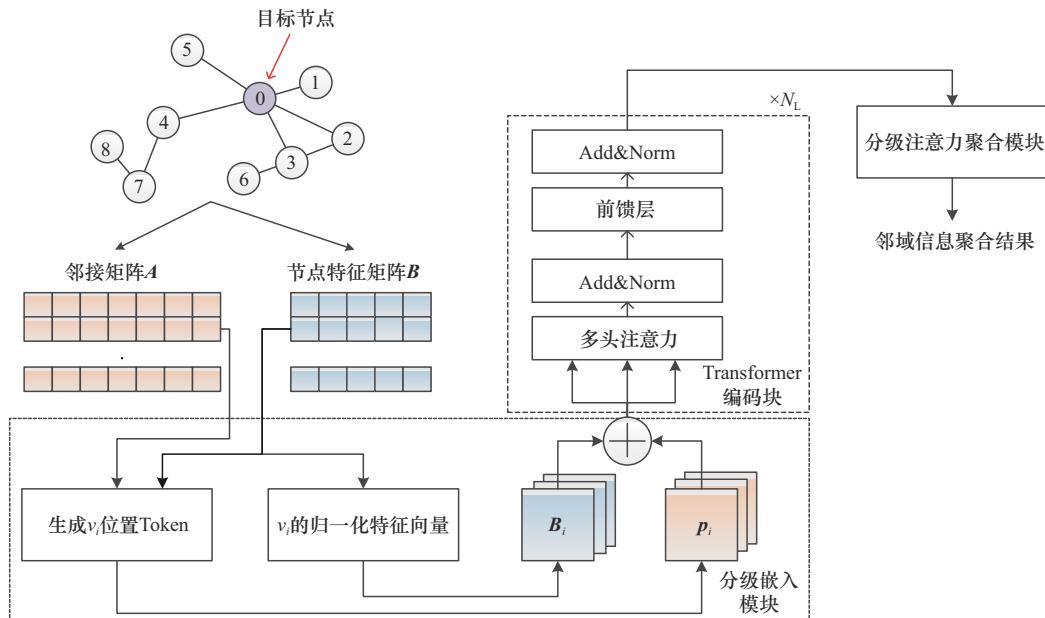


图2 邻域信息聚合流程



有节点视为具有同等重要性,然而在影响力节点识别问题中,近距离邻居节点对识别目标节点影响力的重要性高于远距离邻居节点,因此,将Transformer编码器输出的特征序列传入分级注意力模块,为近距离邻居节点的信息赋予高权重,为远距离邻居节点的信息赋予低权重。

2.2.1 分级嵌入模块

分级嵌入模块的流程如图3所示。Transformer嵌入包含输入嵌入和位置嵌入两个步骤,其中,输入嵌入负责将输入数据转换成Transformer要求的向量形式;位置嵌入负责为向量增加位置Token,Transformer编码器通过位置Token捕获向量中元素间的前后相关性,从而捕获目标节点与邻域内各节点间的相关性。

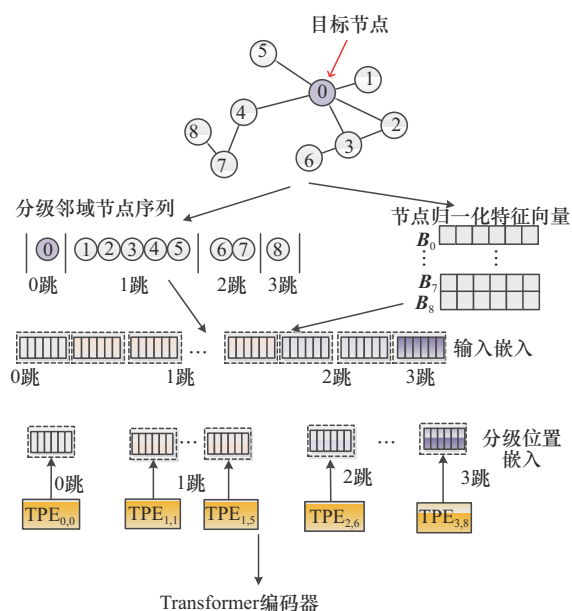


图3 分级嵌入模块的流程

分集嵌入模块的具体处理过程如下。

(1) 输入嵌入。按广度优先搜索目标节点邻域内的节点集,其中目标节点位于序列首位,并标注为第0级,之后为一跳距离的邻居节点集,并分别标注为第1级,以此类推,得到一个分级的邻域节点序列。假设邻域节点数为 N_{nb} ,在邻域节点序列中的每个节点位置插入该节点的特征

向量,得到一个长度为 $Z \times N_{nb}$ 的邻域节点特征序列 $\mathbf{X}$ 。

(2) 分级位置嵌入。原Transformer模型提供的位置嵌入方法为:

$$\begin{cases} PE^{(p, 2i)} = \sin\left(\frac{p}{10000^{2i/d}}\right) \\ PE^{(p, 2i+1)} = \cos\left(\frac{p}{10000^{2i/d}}\right) \end{cases} \quad (6)$$

其中, p 为当前位置的索引, d 为位置嵌入向量的长度, i 为位置嵌入向量的维度索引。式(6)中上下两个表达式分别为偶数位置和奇数位置的位置嵌入。

式(6)用于对输入序列 $\mathbf{X}$ 中每个节点的位置进行唯一编码,Transformer编码器基于这种位置编码能够理解 $\mathbf{X}$ 中节点的前后顺序。如前文的图3所示,第1跳的邻居数为5,第2跳的邻居数为2,可见不同距离的邻居节点数量可能不同。此外,邻域信息聚合任务需要对不同距离的邻居节点信息进行不同的处理,因此本文设计了分级位置嵌入机制,为不同距离的邻居节点分配不同的级数。分级位置嵌入方法为:

$$\begin{cases} TPE^{(p, 2i)} = \sin\left(\frac{p + h_p \times 100000}{10000^{2i/d}}\right) \\ TPE^{(p, 2i+1)} = \cos\left(\frac{p + h_p \times 100000}{10000^{2i/d}}\right) \end{cases} \quad (7)$$

其中, h_p 为位置 p 处节点与目标节点间的跳数。式(7)中上下两个表达式分别为偶数位置和奇数位置的分级位置嵌入。

通过式(7)在位置嵌入向量中增加节点与目标节点的跳数信息,从而得到分级的位置嵌入向量。在下文中,分级注意力模块从Transformer编码向量中解码得到跳数信息,并为不同距离的邻居节点信息赋予不同的权重。

2.2.2 Transformer 编码器

Transformer是一种处理序列的有效模型,通过MSA机制实现对长序列依赖性的捕捉。考虑

此过程的运算量大, Transformer 将多注意力头并行运算来提高模型的运行效率, 从而在长序列特征提取的质量和运算速度间取得了较好的平衡。本文使用 Transformer 编码器提取节点的邻域上下文信息。Transformer 编码器结构如图 4 所示, 每个编码器由 N_L 个结构相同的编码块级联而成, 每个编码块的处理流程依次为多头注意力、残差连接和层标准化 (Add&Norm)、前馈层、残差连接和层标准化。编码块的核心是利用 MSA 机制捕获输入向量中各位置之间的相关性, 它由多个单头注意力模块并行构成, 因此训练效率更高。由图 4 可知, 每个编码块包含两个残差连接, 这两个残差连接可防止神经网络深度过深引起的过拟合问题。

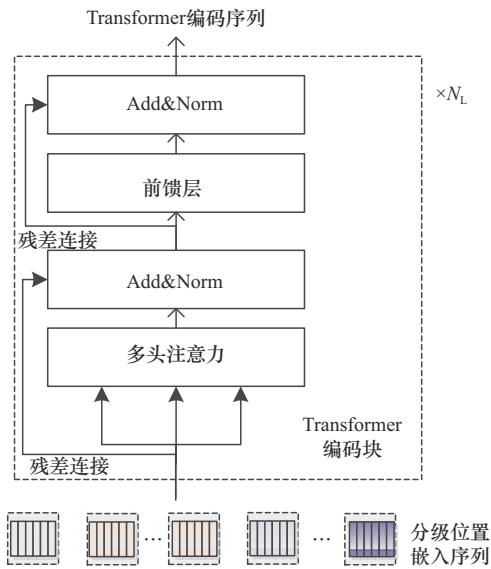


图4 Transformer 编码器结构

MSA 由多个自注意力头并行而成。假设输入的分级位置嵌入向量为 X_{TPE} , 其维度为 $N \times d$, N 为 X_{TPE} 包含的节点数, d 为每个节点的特征维度。每个自注意力 h 头将 X_{TPE} 映射到 3 个不同的子空间, 可表示为:

$$\begin{cases} Q = X_{TPE} W^Q \\ K = X_{TPE} W^K \\ V = X_{TPE} W^V \end{cases} \quad (8)$$

其中, W^Q 、 W^K 和 W^V 均为可学习映射矩阵, W^Q 和 W^K 的维度均为 $d \times d_K$, d_K 为子空间 Q 和子空间 K 的维度, W^V 的维度为 $d \times d_V$, d_V 为子空间 V 的维度。

接着, 采用 softmax 函数从子空间 Q 、 K 和 V 评估输入节点间的相关性, 计算方法为:

$$h = SA(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_K}} \right) V \quad (9)$$

将所有自注意力头的输出拼接成 MSA 的结果, 可表示为:

$$\begin{aligned} X_{MH} &= \text{Concat}(h_1, h_2, \dots, h_R) = \\ &\text{Concat} \left(SA(QW_1^Q, KW_1^K, VW_1^V), \dots, \right. \\ &\left. SA(QW_R^Q, KW_R^K, VW_R^V) \right) W^O \end{aligned} \quad (10)$$

其中, $\text{Concat}(\cdot)$ 为向量拼接操作, R 为注意力头数, W^O 为 MSA 的可学习输出权重矩阵。

X_{TPE} 与 X_{MH} 先进行残差连接, 再进行层标准化处理, 此过程可表示为:

$$X_{LN} = \text{LN}(X_{MH} + X_{TPE}) \quad (11)$$

其中, $\text{LN}(\cdot)$ 为层标准化处理。

X_{LN} 的输出传入前馈层进行处理, 并与 X_{LN} 的输入进行残差连接和层标准化处理, 得到最终的编码结果为:

$$X^{\text{Tran}} = \text{LN}(\text{FFN}(X_{LN}) + X_{LN}) \quad (12)$$

其中, $\text{FFN}(\cdot)$ 为前馈层。

Transformer 编码器利用 MSA 机制提取不同节点间的相关性, 其输出向量包含丰富的邻域上下文信息。

2.2.3 分级注意力聚合模块

分级注意力聚合模块的流程如图 5 所示。首先采用 softmax 函数计算邻域每一跳邻居节点的重要性, 将 softmax 函数结果作为该跳邻居的注意力系数, 即距离越近, 分配的注意力系数越高。假设目标节点 v_i 的邻域节点最大跳数为 J , 那么不同跳数邻居节点的注意力系数为:



$$\alpha_{i,j} = \frac{\exp\left(\sigma\left[\delta_j \tanh\left(W_j X_{i,j}^{\text{Tran}}\right)\right]\right)}{\sum_{j'=0}^J \exp\left(\sigma\left[\delta_{j'} \tanh\left(W_{j'} X_{i,j'}^{\text{Tran}}\right)\right]\right)} \quad (13)$$

其中, j 为该节点与目标节点间的跳数, $\tanh(\cdot)$ 为双曲正切函数激活函数, δ_j 和 W_j 为第 j 跳的可学习矩阵, 在网络训练的反向传播中根据当前损失更新 δ_j 和 W_j 的权重。

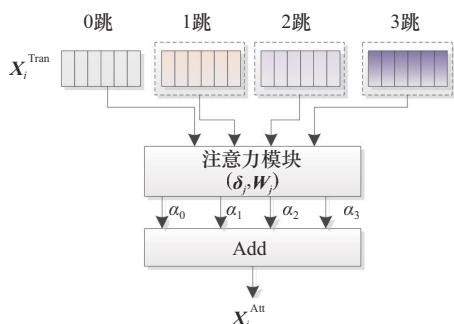


图5 分级注意力聚合模块的流程

接着, 对目标节点 v_i 的邻域信息进行分级注意力聚合, 可表示为:

$$X_i^{\text{Att}} = \sum_{j=0}^J \alpha_{i,j} X_{i,j}^{\text{Tran}} \quad (14)$$

2.3 节点影响力预测

本文通过全连接层将复杂网络节点影响力预测问题建模为一个回归问题, 全连接层由两层网络组成, 每层的节点数均等于邻域节点数量 N_{nb} 。采用易感-感染-康复 (susceptible-infected-recovered, SIR) 模型得到网络节点的正定影响力值, 在训练过程中将 ImTf 模型预测的影响力值与正定影响力值进行比较。考虑均方误差 (mean square error, MSE) 收敛速度快且能放大偏离点的差异, 因此采用 MSE 作为损失函数, 可表示为:

$$\ell_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2 \quad (15)$$

其中, y_i 为正定影响力值, y'_i 为 ImTf 模型预测的影响力值。

3 实验结果与分析

3.1 实验数据集

本文采用 4 个不同类型的复杂网络作为 Benchmark 数据集, 以评估所提方法的性能。实验数据集的基本属性见表 1。表 1 展示了 4 种复杂网络的基本属性, 其中, Email 数据集^[16]来自罗维拉-威尔吉利大学的邮件交流关系, Jazz 数据集^[17]来自爵士乐手间的合作关系, Router 数据集^[18]来自互联网流量数据, USAir 数据集^[19]来自美国航空运输网。

表 1 实验数据集的基本属性

属性	Email 数据集	Jazz 数据集	Router 数据集	USAir 数据集
边数	5 451	742	625 8	212 6
节点数	1 133	198	5022	332
平均度	9.622	7.495	2.492	12.807
最大度	71	100	106	139
流行阈值	0.056 5	0.026 6	0.078 6	0.023 1

3.2 正定影响力计算方法

使用 SIR 传播模型^[20]在 4 个复杂网络上模拟传染病在人群中的传播过程, SIR 模型将节点分为 3 种状态: 易感者 (susceptible)、感染者 (infected)、康复者 (recovered)。假设 S 、 I 和 R 分别表示易感者、感染者和康复者的人数。在 $t=0$ 时, 仅有一个感染节点 v_i , 其他均为易感者。在之后的每个时刻, 一个感染者可能以一定的概率感染其易感者邻居节点, 此概率称为传播率 β , 感染者也可能以一定的概率复原, 此概率称为复原率, 本文的复原率设为 1。重复上述传播过程, 直至网络中不存在感染者节点, 最终计算康复者占总人数的比例 r , 将 r 作为节点 v_i 的影响力值。当传播率 β 在流行阈值 β_{th} 附近时, r 会随着传播率的升高而大幅升高, β_{th} 的计算式为^[20]:

$$\beta_{\text{th}} = \frac{k}{(\underline{k})^2 - \underline{k}} \quad (16)$$

其中, k 为网络的平均节点度数。

3.3 性能评价标准

采用 Kendall tau 系数 (Kendall tau coefficient, KtC) 作为性能评价标准, 基于 KtC 度量 SIR 模型正定影响力排序序列和本文所提方法预测排序序列间的一致性。假设两个排序序列分别为 $Y' = (y'_1, y'_2, \dots, y'_N)$ 和 $Y = (y_1, y_2, \dots, y_N)$, KtC 的计算式为:

$$\tau = \frac{2(c_+ - c_-)}{(N+1)N} \quad (17)$$

其中, c_+ 为一致对的数量, c_- 为不一致对的数量。KtC 值的范围为 $[-1, 1]$, 值越高, 表示两个顺序序列的一致性越高。

3.4 实验环境与网络训练

在一台服务器上进行影响力识别实验, 处理器为 Intel Core i9-12900K, 内存为 128 GB, 显卡为 NVIDIA RTX A6000。编程环境为 Python 3.7.6, 深度学习框架为 PyTorch 1.7.0。使用 Huggingface 库建立本文 ImTf 模型的 Transformer 编码器, MSA 的注意力头数为 10。

由于在真实网络模拟 SIR 传播的成本过高, 采用 Barabási - Albert 模型^[21]生成 500 个小规模网络, 在小规模网络上进行 SIR 模拟。考虑 SIR 传播的随机性, 每个网络独立模拟 100 次, 计算每个节点的平均影响力值作为真实影响力值。从 500 个小规模网络中随机选取 450 个作为训练集, 剩余 50 个作为验证集。使用 Adam 优化器在小规模网络上训练 ImTf 模型, 以决定 ImTf 模型的超参数和可学习参数。Adam 优化器的学习率为 0.001, 最大 epoch 数为 100。将批大小分别设为 32、64、96、128、256, 观察批大小对模型的影响。不同批大小下 ImTf 模型平均训练损失和平均验证损失如图 6 所示。由图 6 可知, 当批大小为 32 和 64 时, ImTf 模型收敛的训练损失和验证损失均较高; 当批大小为 128 时, 训练损失与验证损失均较低, 且训练损失与验证损失相差较小, 这说明

此时的神经网络泛化性能更好。经实验比较后, 批大小设为 128。

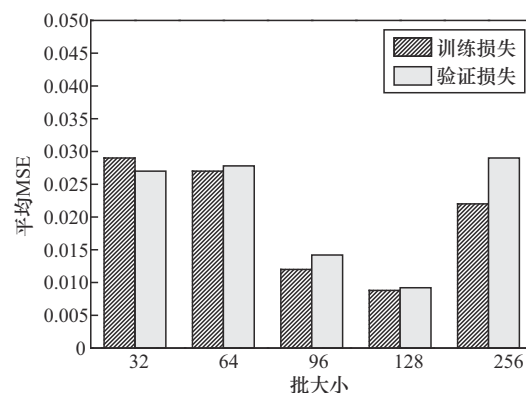


图6 不同批大小下 ImTf 模型平均训练损失和平均验证损失

3.5 参数实验

为评估不同邻域节点数量 N_{nb} 和 Transformer 编码块数量 N_L 对 ImTf 模型性能的影响, 对这两个参数分别进行参数实验。首先, 将 N_L 固定为 2, 传播率 β 固定为 $1.6 \times \beta_{th}$, N_{nb} 分别设为 16、24、32、40、48 和 56, 观察不同邻域节点数量对影响力预测性能的影响。考虑 SIR 模型具有随机性, 在每个参数条件下对网络独立地完成 1 000 次 SIR 传播, 统计每个节点的平均影响力值作为正定影响力值。基于 KtC 度量 SIR 模型正定影响力排序序列和 ImTf 模型预测排序序列间的一致性, 不同邻域节点数量的影响力预测性能如图 7 所示。由图 7 可知, 当邻域节点数较少时, ImTf 模型聚集的邻域信息较少, 所提取特征的判别性较弱。随着邻域节点数的增加, ImTf 模型的节点影响力识别准确率呈现升高趋势。然而, 当邻域节点数较多时, 不同节点间邻域重叠率升高, 此时节点聚集的邻域特征信息逐渐趋同, 导致所提取特征的判别性降低。经实验比较, ImTf 模型在 Email 网络、Jazz 网络、Router 网络和 USAir 网络上的最佳邻域节点数分别为 32、40、24 和 40。在后续实验中, 各复杂网络的参数 N_{nb} 均取最佳值。

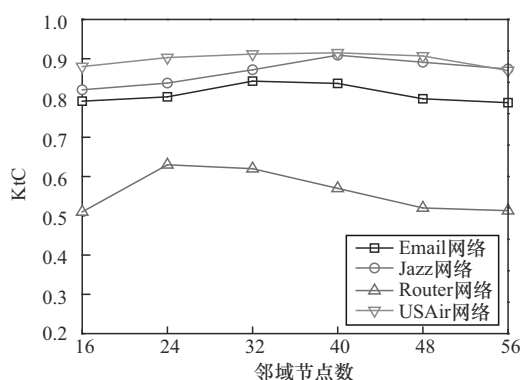


图7 不同邻域节点数量的影响力预测性能

将各复杂网络的参数 N_{nb} 均设为最佳值, 传播率 β 固定为 $1.6 \times \beta_{th}$, N_L 分别设为 1、2、3、4 和 5, 观察不同 Transformer 编码块数对影响力预测性能的影响。在每个参数条件下对网络独立地完成 1 000 次 SIR 传播, 统计每个节点的平均影响力值作为正定影响力值。基于 KtC 度量 SIR 模型正定影响力排序序列和 ImTf 模型预测排序序列间的一致性, 不同编码块数的影响力预测性能如图 8 所示。由图 8 可知, 随着编码块数的增加, ImTf 模型的节点影响力识别准确率呈现升高趋势, 其中 Email 网络和 USAir 网络的最佳编码块数为 3~5, Jazz 网络的最佳编码块数为 4 和 5。考虑模型的复杂度, 将 Email 网络和 USAir 网络的最佳编码块数设为 3, Jazz 网络的最佳编码块数设为 4, Router 网络的最佳编码块数设为 2。在后续实验中, 各复杂网络的参数 N_L 均取最佳值。

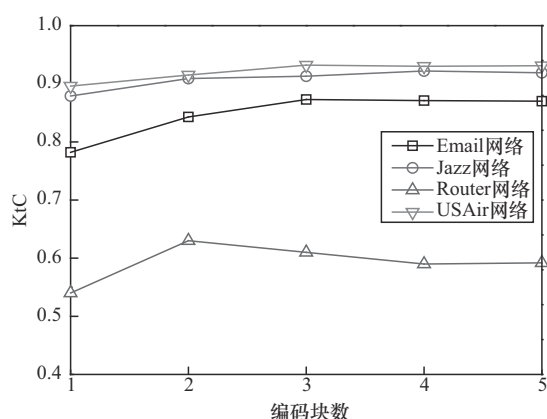


图8 不同编码块数的影响力预测性能

3.6 对比实验

为解决 GNN 和 GCN 处理影响力节点识别问题的局限性, 实验选取引言部分所提及的 OlapGN 方法^[14]、REGCN 方法^[15]作为对比方法。此外, 选择 CGNN 方法^[22]与 HIKS 方法^[23]作为对比方法, 其中 CGNN 方法结合了卷积神经网络和 GNN 技术, HIKS 方法结合了边权重学习与 K-shell 方法。将传播率 β 设为 $1 \times \beta_{th}$ 、 $1.2 \times \beta_{th}$ 、 $1.4 \times \beta_{th}$ 、 $1.6 \times \beta_{th}$ 、 $1.8 \times \beta_{th}$ 和 $2 \times \beta_{th}$, 评估各方法在不同传播率下的影响力节点识别性能。在每个参数条件下对网络独立地完成 1 000 次 SIR 传播, 统计每个节点的平均影响力值作为正定影响力值。基于 KtC 度量 SIR 模型正定影响力排序序列和各方法预测排序序列间的一致性, 复杂网络影响力节点识别的对比实验结果如图 9 所示。由图 9 (a) 可知, 本文所提方法和 REGCN 方法的影响力识别性能均随网络传播率的升高而升高。经实验比较, 在各个网络传播率下, OlapGN 方法和 CGNN 方法的影响力识别性能均低于 REGCN 方法和本文所提方法。总体而言, 所提 ImTf 模型在 Email 网络上的影响力节点识别性能大幅优于基于 GNN 和 GCN 的方法, 且优于 HIKS 方法。由图 9 (b) 可知, OlapGN 方法和 REGCN 方法在 Jazz 网络上的影响力节点识别性能明显优于 CGNN 方法, 但依然明显低于 ImTf 模型。由图 9 (c) 可知, OlapGN 方法、REGCN 方法和 ImTf 在 Router 网络上的影响力节点识别性能受传播率的影响较大, 而 CGNN 方法的性能较稳定, 经过比较, ImTf 在 Router 网络上的影响力节点识别性能展现出明显的优势。由图 9 (d) 可知, 4 种方法在 USAir 网络上的影响力节点识别性能受传播率的影响较小, 相较于其他 3 种对比方法, ImTf 方法优势十分明显。

综上, 基于 GNN 和 GCN 的复杂网络影响力节点识别方法经过多轮迭代更新后, 会出现节点特征过平滑现象, 因此, 所能结合的邻域范围有限, 且在不同网络上的性能差异较大。ImTf 模型

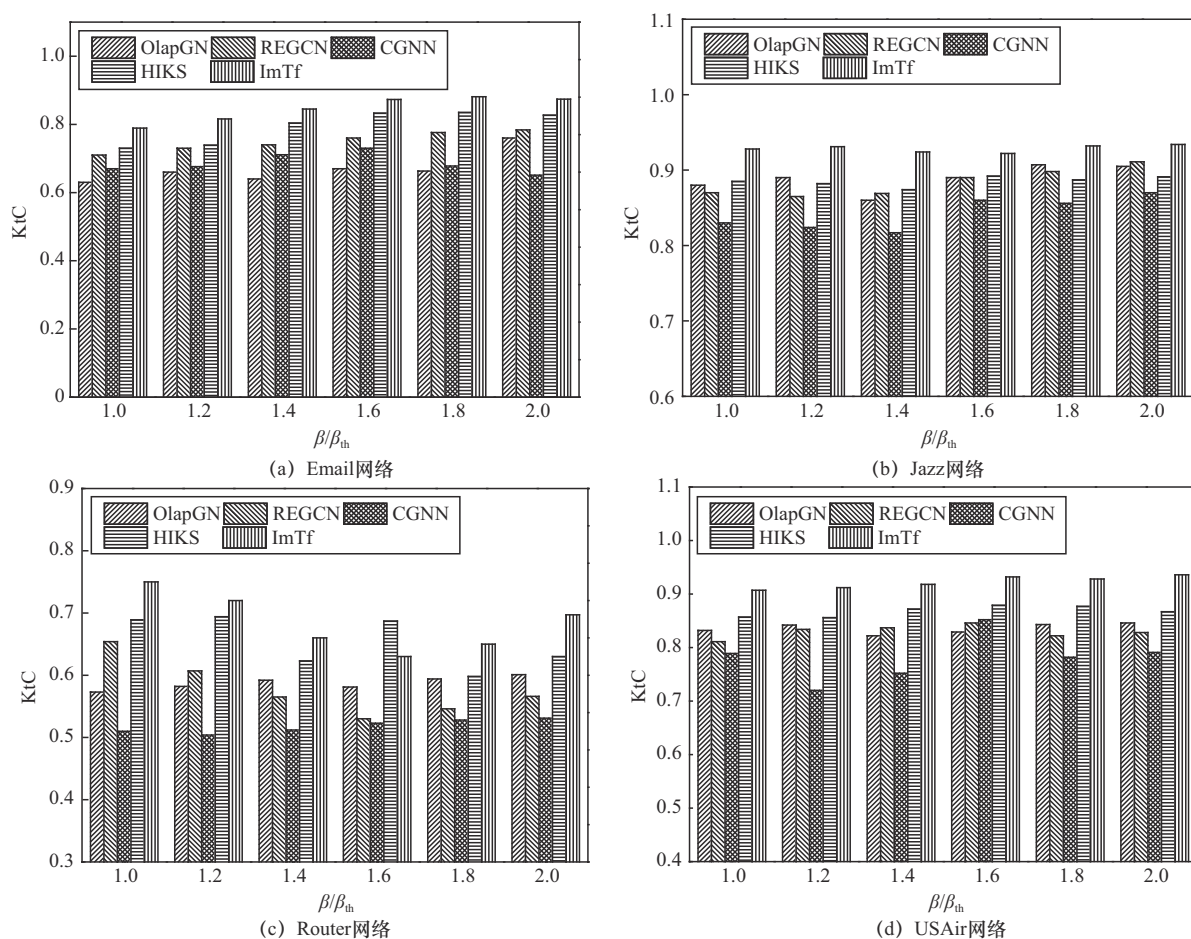


图9 复杂网络影响力节点识别的对比实验结果

的位置嵌入机制和MSA机制可避免GNN和GCN的过平滑问题,因此,本文所提方法能更有效地聚合节点邻域内的节点信息,通过集成网络拓扑结构信息、节点特征信息和邻域上下文信息,有效提高了不同条件下复杂网络影响力节点识别的准确性和鲁棒性。

4 结束语

本文提出了一种基于深度学习的多样化复杂网络影响力节点识别方法,以提高多样化复杂网络中影响力节点识别的准确性和稳定性。利用Transformer对不同距离邻居信息进行分级聚合,在聚合过程中分析了目标节点对邻域内节点的依赖性。最终,通过集成网络拓扑结构信息、节点特征信息和邻域上下文信息,有效提高了不同条

件下复杂网络影响力节点识别的准确性和鲁棒性。

当邻域节点较少时,ImTf模型所提取特征的判别性较弱;当邻域节点较多时,不同节点间邻域重叠率较高,此时节点聚集的邻域特征信息逐渐趋同,导致所提取特征的判别性降低。因此,邻域节点数是决定ImTf模型性能的关键因素。本文通过参数实验确定了不同网络的最佳邻域节点数和编码块数,但尚不确定这些参数是否具有通用性。未来将进一步研究这些参数的自适应选择方法,如引入超参数优化技术或强化学习,根据每个复杂网络的特点自适应地决定合适的邻域节点数和编码块数,以提高模型在未知网络中的适应性。



参考文献:

- [1] 曾祥宇, 龙海霞, 杨旭华. 基于马尔可夫相似性增强和网络嵌入的社区发现[J]. 计算机科学, 2023, 50(4): 56-62.
ZENG X Y, LONG H X, YANG X H. Community detection based on Markov similarity enhancement and network embedding[J]. Computer Science, 2023, 50(4): 56-62.
- [2] CHEN X, ZHANG M, YIN S, et al. Construction of protein-protein interaction networks on capsicum against sepsis by regulating of molecular mechanisms: bioinformatics approaches[J]. Journal of Biomedical Nanotechnology, 2023, 19(6): 1062-1069.
- [3] 汪亭亭, 梁宗文, 张若曦. 基于信息熵与迭代因子的复杂网络节点重要性评价方法[J]. 物理学报, 2023, 72(4): 337-347.
WANG T T, LIANG Z W, ZHANG R X. Importance evaluation method of complex network nodes based on information entropy and iteration factor[J]. Acta Physica Sinica, 2023, 72(4): 337-347.
- [4] 许星舟. 基于节点中心性和标签传播算法的社区检测[J]. 计算机应用与软件, 2024, 41(3): 290-296, 344.
XU X Z. Community detection based on node centrality and label propagation algorithm[J]. Computer Applications and Software, 2024, 41(3): 290-296, 344.
- [5] WANG Y Z, WANG R, WANG Y, et al. The alterations of brain network degree centrality in patients with neovascular glaucoma: a resting-state fMRI study[J]. Neurological Sciences, 2023, 44(8): 2915-2922.
- [6] MA J L, WANG S Y. A closeness centrality reconnection strategy to suppress the traffic-driven epidemic spreading[J]. International Journal of Modern Physics C, 2024, 35(5): 2450053.
- [7] 刘宽, 刘海员, 张雷. 一种基于特征向量中心性的基因调控网络结构推测算法[J]. 南开大学学报(自然科学版), 2023, 56(4): 40-44.
LIU K, LIU H Y, ZHANG L. An algorithm for gene regulatory network structure inference based on eigenvector centrality[J]. Acta Scientiarum Naturalium Universitatis Nankaiensis, 2023, 56(4): 40-44.
- [8] 李静, 路庆昌, 徐鹏程, 等. 考虑站点耦合关系的大型城市轨道交通网络级联失效分析[J]. 浙江大学学报(工学版), 2024, 58(9): 1945-1955.
LI J, LU Q C, XU P C, et al. Cascading failure analysis of large-scale urban rail transit network considering station coupling relationship[J]. Journal of Zhejiang University (Engineering Science), 2024, 58(9): 1945-1955.
- [9] WANG X H, YANG Q, LIU M Z, et al. Comprehensive influence of topological location and neighbor information on identifying influential nodes in complex networks[J]. PLoS One, 2021, 16(5): e0251208.
- [10] 原慧琳, 冯宠. 基于 k-shell 熵的影响力节点的排序与识别[J]. 计算机科学, 2022, 49(S2): 226-230.
YUAN H L, FENG C. Ranking and identification of influence nodes based on k-shell entropy[J]. Computer Science, 2022, 49(S2): 226-230.
- [11] 龚志豪, 蒋沅, 代冀阳, 等. 基于交叉熵的节点重要性排序算法[J]. 电子科技大学学报, 2023, 52(6): 944-953.
GONG Z H, JIANG Y, DAI J Y, et al. Node importance ranking algorithm based on cross entropy[J]. Journal of University of Electronic Science and Technology of China, 2023, 52(6): 944-953.
- [12] 孙轩宇, 史艳翠. 融合项目影响力的图神经网络会话推荐模型[J]. 计算机应用, 2023, 43(12): 3689-3696.
SUN X Y, SHI Y C. Session-based recommendation model by graph neural network fused with item influence[J]. Journal of Computer Applications, 2023, 43(12): 3689-3696.
- [13] 臧玕珣, 徐鑫航. 基于网络嵌入的农产品销售推荐系统[J]. 计算机技术与发展, 2022, 32(10): 209-214.
ZANG J X, XU X H. Recommendation system for agricultural products marketing channels based on network embedding[J]. Computer Technology and Development, 2022, 32(10): 209-214.
- [14] RASHID Y, BHAT J I. OlapGN: a multi-layered graph convolution network-based model for locating influential nodes in graph networks[J]. Knowledge-Based Systems, 2024, 283: 111163.
- [15] WU Y H, HU Y M, YIN S Y, et al. A graph convolutional network model based on regular equivalence for identifying influential nodes in complex networks[J]. Knowledge-Based Systems, 2024, 301: 112235.
- [16] GUIMERÀ R, DANON L, DÍAZ-GUILERAA, et al. Self-similar community structure in a network of human interactions[J]. Physical Review E, 2003, 68(6): 065103.
- [17] GLEISER P M, DANON L. Community structure in Jazz[J]. Advances in Complex Systems, 2003, 6(4): 565-573.
- [18] SPRING N, MAHAJAN R, WETHERALL D, et al. Measuring ISP topologies with rocketfuel[J]. IEEE/ACM Transactions on Networking, 2004, 12(1): 2-16.
- [19] 袁榕, 宋玉蓉, 孟繁荣. 一种基于加权网络拓扑权重的链路预测方法[J]. 计算机科学, 2020, 47(5): 265-270.
YUAN R, SONG Y R, MENG F R. Link prediction method based on weighted network topology weight[J]. Computer Science, 2020, 47(5): 265-270.
- [20] PASTOR-SATORRAS R, CASTELLANOC, VANMIEGHEM P, et al. Epidemic processes in complex networks[J]. Reviews of

Modern Physics, 2015, 87(3): 925-979.

- [21] BARABÁSI A L, BONABEAU E. Scale-free networks[J]. Scientific American, 2003, 288(5): 60-69.
- [22] ZHANG M, WANG X J, JIN L, et al. A new approach for evaluating node importance in complex networks via deep learning methods[J]. Neurocomputing, 2022, 497: 13-27.
- [23] CHAKRAVARTHY T S, SELVARAJ L. HIKS: a K-shell-weighted hybrid approach method for detecting influential nodes in complex networks using possible edge weights[J]. International Journal of Communication Systems, 2024, 37(7): e5722.

[作者简介]



马玉磊 (1982-), 男, 新乡学院计算机与信息工程学院副教授, 主要研究方向为计算机网络与网络分析。

郭莎莎 (1993-), 女, 新乡学院计算机与信息工程学院助教, 主要研究方向为图像处理。