

HeSPS-GNN: 基于子图级别图神经网络的链接预测方法

傅继彬, 白惠元, 赵培元

(河南财经政法大学 计算机与信息工程学院, 河南 郑州 450046)

摘要: 链接预测是图数据分析中的关键任务, 子图级图神经网络方法在该任务中表现出较强的结构建模能力, 但现有方法在子图抽取阶段普遍存在子图规模不一致的问题, 导致在排序池化层中产生节点裁剪或零向量填充, 从而引发信息损失, 尤其在稀疏图场景下性能下降明显。针对上述问题, 提出一种预设子图节点数量的层次采样方法 HeSPS。该方法以目标节点对为中心, 采用由近及远的分层扩展策略进行邻居采样, 在保证子图节点数量基本一致的同时尽可能保留原图结构信息。基于 HeSPS 构建子图级链接预测模型 HeSPS-GNN, 并在多个公开基准数据集上进行实验验证。结果表明, 与 SEAL 和 SPG4LP 等主流子图方法相比, HeSPS-GNN 在 AUC 指标上取得更优性能, 尤其是在 Power 和 Router 等稀疏图数据集上, 其 AUC 最高提升达 4.26%。该方法能有效降低排序池化带来的信息损失, 并提升稀疏图场景下链接预测的稳定性与准确性。

关键词: 链接预测; 图神经网络; 闭包图; 简单路径图; 排序池化层

DOI: 10.11907/rjdlk.251772

扫描二维码阅读全文:



中图分类号: TP181

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-08

HeSPS-GNN: Subgraph-level Graph Neural Network Approach for Link Prediction

FU Jibin, BAI Huiyuan, ZHAO Peiyuan

(School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou 450046, China)

Abstract: Link prediction is a key task in graph data analysis. Subgraph-level graph neural network methods demonstrate strong structural modeling capabilities in this task. However, existing methods generally suffer from inconsistent subgraph sizes during the subgraph extraction stage, leading to node pruning or zero vector padding in the ranking pooling layer, which results in information loss, especially in sparse graph scenarios where performance degradation is significant. To address the aforementioned issues, we propose a hierarchical sampling method called HeSPS that presets the number of subgraph nodes. This method centers around target node pairs and adopts a hierarchical expansion strategy from near to far for neighbor sampling, ensuring that the number of subgraph nodes remains basically consistent while preserving as much structural information from the original graph as possible. Based on HeSPS, we construct a subgraph-level link prediction model called HeSPS-GNN and conduct experimental validation on multiple public benchmark datasets. The results show that compared to mainstream subgraph methods such as SEAL and SPG4LP, HeSPS-GNN achieves superior performance in terms of the AUC metric, especially on sparse graph datasets such as Power and Router, where the AUC improvement is up to 4.26%. This method effectively reduces the information loss caused by ranking pooling and enhances the stability and accuracy of link prediction in sparse graph scenarios.

Key Words: link prediction; graph neural networks; enclosing subgraph; simple path graph; sortpooling layer

收稿日期: 2025-11-19

作者简介: 傅继彬(1975-), 男, 博士, CCF 会员, 河南财经政法大学计算机与信息工程学院副教授, 研究方向为人工智能、深度学习; 白惠元(2001-), 男, CCF 会员, 河南财经政法大学计算机与信息工程学院硕士研究生, 研究方向为人工智能、深度学习; 赵培元(2002-), 男, CCF 会员, 河南财经政法大学计算机与信息工程学院硕士研究生, 研究方向为差分隐私、人工智能。本文通讯作者: 白惠元。

0 引言

现实世界中的大量复杂关系数据可以抽象为图结构,如社交网络、交通网络、生物分子网络和信息网络等。然而,这些图往往是不完整的,由于观测限制、数据噪声或关系的动态演化,图中的边并不能完全刻画节点之间的真实关联。链接预测任务旨在根据已有的图结构信息来预测节点之间潜在或缺失的连接,在推荐系统、关系补全、异常检测和生物信息分析等领域具有重要应用价值^[1-4]。

随着图神经网络(Graph Neural Network, GNN)的发展,基于深度学习的链接预测方法逐渐成为研究热点。该类方法通过对图结构和节点属性进行联合建模,在预测精度和泛化能力方面相较传统启发式方法和嵌入方法取得明显进展。尤其是近年来提出的子图级链接预测方法,其通过构建以目标节点对为中心的局部子图,并将链接预测问题转化为子图分类任务,显著增强了模型对局部结构模式的表达能力。

然而,现有子图级方法在子图抽取和建模过程中仍面临一些挑战,特别是在稀疏图场景下,子图规模不稳定、结构信息利用不足等问题容易导致模型性能下降。因此,研究更加稳定、有效的子图构建与建模策略,对于提升子图级链接预测方法的整体性能具有重要意义。

1 相关工作

链接预测作为图学习领域中的重要研究方向,已形成较为系统的研究体系。相关方法大体可分为启发式方法、图嵌入方法、基于图神经网络的节点级方法以及近年来兴起的子图级方法。早期的链接预测研究主要基于启发式指标,通过刻画节点之间的局部结构相似性进行预测。例如,Common Neighbors^[5]、Jaccard Coefficient^[6]、Adamic-Adar^[7]以及 Preferential Attachment^[8]等方法分别从共同邻居数量、邻居集合重合度和节点度分布等角度对潜在链接进行评分。该类方法计算简单、可解释性强,但仅利用局部结构信息,难以刻画复杂的高阶结构模式,在结构复杂或稀疏图中预测能力受限。

随着表示学习的发展,研究者们提出多种基于图嵌入的链接预测方法。例如,DeepWalk^[9]、node2vec^[10]和 LINE^[11]等方法通过随机游走或边建模策略,将节点映射到低维向量空间,并利用向量相似度进行链接预测。该类方法在一定程度上提升了预测性能和模型可扩展性,但其本质仍是面向节点的表示学习,未显式建模待预测节点对之间的局部结构关系,且在结构差异较大的场景中泛化能力有限。

近年来,GNN的提出为链接预测提供了新的建模范

式。例如,GCN^[12]、GraphSAGE^[13]和 GAT^[14]等方法通过消息传递机制聚合邻居信息,学习节点的高阶表示,并通过节点嵌入的组合或内积实现链接预测。该类方法能够端到端地融合节点属性与图结构信息,但通常在节点级别进行建模,对节点对之间的结构交互刻画不足,难以捕捉局部子结构模式。为弥补节点级方法在结构建模方面的不足,子图级链接预测方法逐渐受到关注。例如,Zhang等^[15]提出的 SEAL 方法为每一条候选边构建目标节点对的 h-hop 闭包子图,并结合双半径节点标签和 DGCNN(Dynamic Graph CNN)模型,将链接预测问题转化为子图分类任务,在多个基准数据集上取得了较好效果。然而,该方法在实际应用中仍存在一定局限性:一方面,闭包阶数 h 的增大会导致子图规模迅速膨胀,带来较高的计算开销;另一方面,不同候选边对应的子图规模差异较大,在排序池化层中不可避免地产生节点裁剪或零向量填充,造成结构信息损失,尤其在稀疏图场景下影响模型性能^[16]。后续研究从不同角度对 SEAL 框架进行了拓展。例如,文献[17]提出的 GraIL 通过子图推理机制提升了模型在归纳场景下的泛化能力。文献[18]提出的 NBFNet 从路径推理角度对节点对关系进行建模,加强了长路径依赖的表达能力。该类方法在一定程度上提升了子图级链接预测的性能,但并未从子图规模一致性的角度系统解决排序池化阶段的信息损失问题。另一类研究以目标节点对之间的路径为中心构建子图。例如,文献[19]提出的 SPG4LP 方法通过枚举目标节点对之间的简单路径构建路径子图,在强调语义路径信息的同时有效控制了子图规模,在部分长距离链接预测任务中的表现优于封闭子图方法。然而,该方法通常会裁剪目标节点周围的邻域结构,忽略了一跳邻居等对信息传递至关重要的节点,从而导致局部结构信息利用不足。此外,路径子图的规模同样存在不稳定问题,在排序池化层中仍可能引入信息损失。为提升子图抽取的效率和可扩展性,部分研究尝试通过随机采样或简化建模策略改进子图级链接预测方法。例如,文献[20]提出的 ScaLed 采用随机游走方式抽取子图以缓解计算开销。文献[21]提出的 S3GRL 通过简化子图聚合过程并引入预计算扩散算子,提高了模型在大规模图上的效率。然而,这些方法主要关注计算复杂度和扩展性问题,对排序池化阶段子图规模不一致所带来的信息损失关注较少。

综上所述,现有链接预测方法在结构建模能力和预测性能方面取得了显著进展,但在子图级方法中仍普遍存在子图规模不稳定、排序池化信息损失以及稀疏节点建模能力不足等问题。针对上述不足,本文从子图采样策略入手,提出一种预设定子图节点数量的层次采样方法 HeSPS(Hierarchical Sampling with Preset Size),并在此基础上构建子图级链接预测模型 HeSPS-GNN,以提升模型在不同图结构,尤其是稀疏图场景下的预测性能。

2 预备知识与问题概述

2.1 闭包图与简单路径图

对于给定无向图 $G=(V,E)$, 其中 V 表示图中所有节点的集合, E 表示已观察到的边的集合。以下给出闭包图和简单路径图的定义。

定义一(h-hop 闭包图): 对于一个无向图 $G=(V,E)$, 给定两个节点 $x, y \in V$, 其 h -hop 闭包图记作 S_{xy} , 其由满足 $\{ild(i, x) \leq h \text{ 或 } d(i, y) \leq h\}$ 条件的所有节点诱导而成。

h -hop 闭包图以节点周围跳数为采样标准, 包含目标节点对的邻居节点及其链接信息。图 1 为闭包图示例, 其中红色节点为目标节点, 蓝色节点为目标节点采样的 h 跳周围节点。

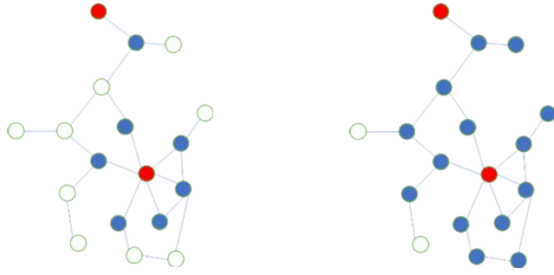


Fig. 1 h-hop enclosing subgraph

图 1 h-hop 闭包图

h -hop 闭包图包含目标节点对周围的一些重要结构, 包括共同邻居、路径、环等。将该子图输入图神经网络可弥补纯节点嵌入方法, 如 GCN、GraphSage 等, 忽视节点对结构关系的不足。然而, h -hop 闭包图也存在一些缺陷: 首先, 每一个候选边都需要提取一个子图, 随着 h 的增大, 子图规模会急剧增加, 一般情况下当 h 取 3 时, 子图规模趋近于全图规模, 而大规模图或海量候选边场景下的计算成本过高; 其次, h -hop 闭包图是基于跳数采样节点且跳数不易过大, 具有很强的局部性, 限制了其感知范围, 无法利用远距离结构关系, 对跨社区或长路径链接预测能力较弱; 最后, 当目标节点中有稀疏节点时, 基于跳数的采样方式大大限制了目标节点的采样能力。

定义二(k-path 简单路径图): 给定无向图 $G=(V,E)$, 整数 k , 对于图中任意两个节点 $x, y \in V$, x, y 之间所有长度 $\leq k$ 的路径集合称为在整数 k 限制下的简单路径图, 记作 $SPG_{x,y}^k$ 。k-path 简单路径子图以目标节点对之间的路径长度为采样标准, 包含目标节点对之间路径中的节点和链接信息。图 2 为简单路径子图示例。

与 h -hop 闭包图相比, k -path 简单路径图更强调路径性而非邻域性, 善于捕捉目标节点对之间的连接性和结构依赖。受限于采样方式, 简单路径子图往往更加稀疏, 有效控制了子图规模, 而且善于长距离目标节点对的链接预测, 但亦存在一定缺陷: 简单路径子图仅包含路径上的节点, 忽略了目标节点周围的邻居节点。在 GNN 中, 目标节

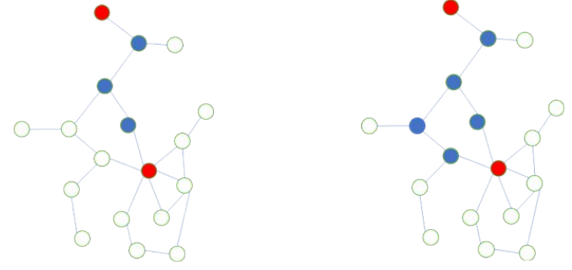


Fig. 2 k-path simple path subgraph

图 2 k-path 简单路径图

点周围被忽略的邻居节点将无法进行信息传递, 可能导致关键信息丢失, 从而影响节点嵌入表示, 降低链接预测准确率。

2.2 GNN

2.2.1 GNN 的表示

GNN 采用消息传递机制, 在每一轮迭代更新中, GNN 通过从邻居节点聚合的信息来更新每个节点的嵌入表示。第 k 轮的更新过程可形式化表示为:

$$\{ild(i, x) \leq h \text{ 或 } d(i, y) \leq h\} \quad (1)$$

$$h_u^{(k+1)} = \phi^{(k)}(h_u^{(k)}, m_{N(u)}^{(k)}) \quad (2)$$

式中: $h_u^{(k)}$ 表示节点 u 在第 k 层的节点嵌入表示; $m_{N(u)}^{(k)}$ 表示来自邻居的聚合消息; ϕ 为具有可学习参数的更新函数, 通常采用多层感知机实现。通常情况下, 最终得到的节点嵌入会被用作该节点的代表向量^[22]。

2.2.2 DGCNN

DGCNN 是一种用于图分类任务的 GNN 架构, 由基于传播机制的图卷积层和排序池化层组成。对于一个封闭子图, 其邻接矩阵为 $A \in \{0, 1\}^{n \times n}$, 节点信息矩阵为 $X \in R^{n \times c}$, DGCNN 使用的图卷积层用公式表示为 $Z = f(\tilde{D}^{-1} \tilde{A} X W)$, 其中 $\tilde{A} = A + I$, $\tilde{A}$ 为加上自环的邻接矩阵, $\tilde{D}$ 为 $\tilde{A}$ 的度矩阵, 定义为 $\tilde{D}_{i,i} = \sum_j \tilde{A}_{i,j}$, W 为可训练的图卷积权重矩阵, f 为非线性激活函数, Z 为新的节点状态矩阵。公式的含义是初始节点特征先经过线性变换, 再通过归一化传播矩阵将信息传递给邻居节点, 则卷积后的第 i 行 Z_i 包含自身节点特征及其一阶邻居特征。DGCNN 堆叠多个上述图卷积层, 并将每层的节点表示拼接作为最终节点表示, 从而提取多阶节点特征。DGCNN 中的排序池化层根据最后一层图卷积输出对节点表示进行排序, 实现节点顺序的同构不变性。随后采用 max-k pooling 操作将排序后的节点序列统一为固定长度, 以便在该序列上训练传统的 1-D CNN, 进行图的二分类任务。

2.3 问题概述

图的链接预测任务可被转化为子图的二分类任务, 因此子图的抽取方法尤为关键。不论是 h -hop 闭包图还是 k -path 简单路径图, 均存在各自的缺陷。更关键的是, 这两种方法采样的子图进入 DGCNN 模型后, 排序池化层会对其进行截断或填充, 不论是 h -hop 闭包图还是 k -path 简单路径图, 一张图采样出的样本子图节点数量差异很大。

当子图过于稀疏时,排序池化层将大量填充零向量,在模型训练时会影响模型效用;当子图规模过于庞大时,又将截断大量节点,其中可能包含重要节点,造成信息损失。为解决上述两种采样方式存在的缺陷,本文提出一种预设子图节点数量的层次子图采样策略,将采样后的子图输入 GNN 模型,以提高模型效用。

3 HeSPS-GNN 模型构建

HeSPS-GNN 模型包含正负样本生成、子图提取、子图信息矩阵构建和 GNN 构建 4 个步骤。框架如图 3 所示。

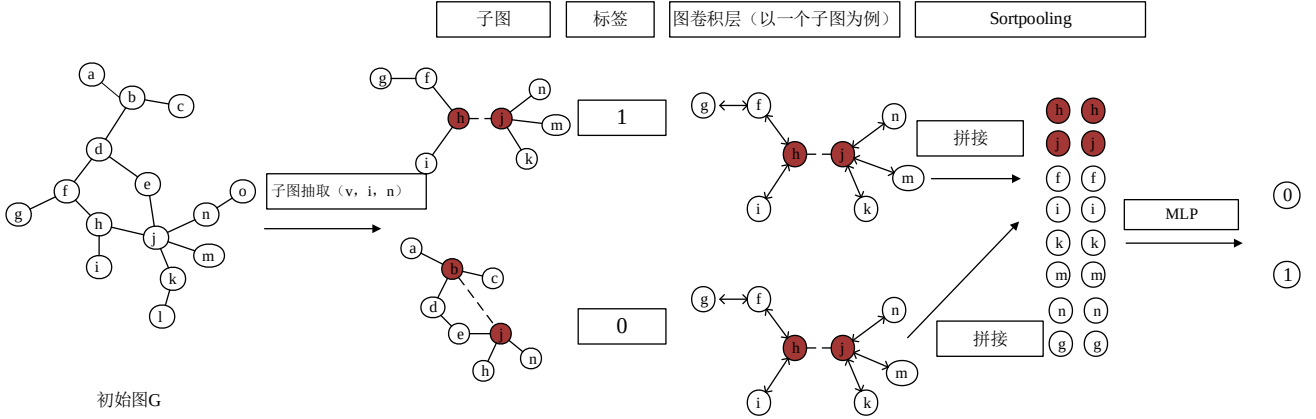


Fig. 3 Framework of HeSPS-GNN model

图3 HeSPS-GNN模型框架

3.1 正负样本生成与子图提取

子图级别的链接预测任务实际上是将节点之间的链接是否存在的问题转化为子图的正负样本二分类问题,因此在子图抽取前应当先采集正负样本。本研究采用算法 1 实现正负样本的采样。

算法 1: 正负样本采样

输入:原始图 $G=(V,E)$,负采样率 λ ,子图提取函数 g_n ;

输出:目标链接的子图 S_{vi} ,标签 y_{vi} ;

1. $NEG \leftarrow \phi$
2. FOR $v \in V$ DO
3. $\bar{N}_1(v) \leftarrow \text{sample}(V \setminus \lambda |N_1(v)|)$
4. $NEG \leftarrow \bar{N}_1(v)$
5. FOR $i \in N_1(v) \cup \bar{N}_1(v)$ DO
6. $S_{vi} \leftarrow g_n(G, (v, i), n)$
7. $y_{vi} = \begin{cases} 1 & \text{if } i \in N_1(v) \\ 0 & \text{otherwise} \end{cases}$
8. RETURN $\{S_{vi}, y_{vi} | (v, i) \in E \cup NEG\}$

遍历图中节点集合 V 的每一个节点 v ,被遍历的节点为目标节点对的第一个节点。采样 λ 乘该节点一跳邻居数量的非邻居节点作为目标节点对的第二个节点,这样采样出来的节点对样本为负样本,其他节点对为正样本。以目标节点对为中心,采用 g_n 函数构建子图,该函数用算法 2 描述。若目标节点对为正样本,则子图标签为 1;若目标节点对为负样本,则子图标签为 0。

算法 2: 预设节点数量的层次子图抽取方法

输入:原始图 $G=(V,E)$,目标节点对 (v,i) ,每个目标节点采样节点的个数 n

输出:以 (v,i) 为中心构建的子图 S_{vi}

1. $V_{\text{sub}} \leftarrow \{v, i\}$

2. $D[v] \leftarrow 0; D[i] \leftarrow 0$

3. FOR each 中心节点 IN $\{v, i\}$ DO

4. 当前节点集 $\leftarrow \{\text{中心节点}\}$

5. 已访问节点 $\leftarrow \{v, i\}$

6. 剩余采样数 $\leftarrow n$

7. $h \leftarrow 0$

8. WHILE 剩余采样数 > 0 AND 当前节点集 $\neq \phi$ DO

9. $h = h + 1$

10. 候选邻居集 $\leftarrow \phi$

11. FOR node IN 当前节点集 DO

12. 候选邻居集 $\leftarrow$ 候选邻居集 $\cup \text{Neighbor}(G, \text{node})$

13. 候选邻居集 $\leftarrow$ 候选邻居集 $\setminus$ 已访问节点

14. IF |候选邻居集| = 0 THEN

15. BREAK

16. IF |候选邻居集| $>$ 剩余采样数 THEN

17. 候选邻居集 $\leftarrow \text{Random_sample}(\text{剩余采样数})$

18. $V_{\text{sub}} \leftarrow V_{\text{sub}} \cup \text{候选邻居集}$

19. 已访问节点 $\leftarrow$ 已访问节点 $\cup$ 候选邻居集

20. FOR node IN 候选邻居集 DO

21. $D[\text{node}] \leftarrow h$

22. 剩余节点数 $\leftarrow$ 剩余采样数 $- |$ 候选邻居集 $|$

23. 当前节点集 $\leftarrow$ 候选邻居集

24. $S_{vi} \leftarrow \text{Subgraph}(G, V_{\text{sub}})$

25. IF $(v, i) \in E$ THEN

26. Remove_Edge($S_{vi}, (v, i)$)

27. RETURN S_{vi}

算法 2 从目标节点对的两个端点分别以 BFS 方式扩展

邻居节点,直到每个端点采样够 n 个邻居节点为止,这样能尽可能保证子图节点个数的一致性,从而在后续 DC-GNN 的排序池化层中减少信息损失。同时,当目标节点为稀疏节点时,这样的采样策略能摆脱 h-hop 闭包图受跳数限制的局限性,在 n 个采样邻居的条件下不断挖掘稀疏节点在图中的结构信息,避免简单路径子图中大量删去目标节点周围邻居节点的情况。

3.2 子图信息矩阵构建

完成目标链接的子图提取后,需要对每一个子图进行节点信息矩阵构建,并将其作为 GNN 的输入。节点信息由节点自身属性信息矩阵、图嵌入表示矩阵和结构化标签矩阵 3 种矩阵拼接而成。对于有节点自身属性的数据集,特征矩阵即为自身属性,如果一些数据集没有自身属性,则用单位矩阵替代。图嵌入表示用 node2vec 学习,由于会学习到链接信息,在 node2vec 之前需要将负采样链接也加入到图中。结构化标签采用双半径节点标签 DRNL (Dual Radius Node Label),设某一个节点 u 到目标节点 v 和 i 的最短距离分别为 $d_v(u)$ 和 $d_i(u)$,则 DRNL 标签的计算公式为:

$$DRNL(u) = 1 + \min(d_v(u) + d_i(u)) + \left\lfloor \frac{d_v(u) + d_i(u)}{2} \right\rfloor + \mod(d_v(u) + d_i(u), 2) - 1 \quad (3)$$

该编码方式能有效区分不同结构角色的节点,如目标节点、共邻节点、两跳邻居等,从而增强 GNN 对局部结构的感知能力。

3.3 GNN 构建

完成目标链接的子图提取和子图信息矩阵构建后,采

用 DGCNN 对每个子图进行表示学习并判断目标链接是否存在。如图 4 所示,DGCNN 由多层图卷积和一层排序池化模块组成。图卷积层对输入的子图进行多层图卷积操作,逐层聚合邻居信息,生成节点的高阶表示。设输入的节点特征矩阵为 X ,邻居矩阵为 A ,则图卷积操作的公式为 $Z = f(\tilde{D}^{-1}\tilde{A}XW)$ 。由于子图大小的不确定性和后续全连接层要求输入的一致性,图卷积层后需设置一个排序池化层 (Sortpooling),该层按照每个节点表示的最后一个维度的值进行排序,保留前 k 个节点。

本研究的采样机制基本适配于排序池化层,每个目标节点采样 n 个节点,则目标链接理想状态下会采样出 $2n$ 个节点,将 k 值设置为 $2n$ 将大大避免排序池化层填充的大量零向量或删除大量节点,其中大概率包含重要节点。既然指定了每个目标节点按照层次遍历采样 n 个节点,为什么还需要排序池化层? 这是由于在实际采样中,有些数据集是非连通图,某些目标节点无法采样够 n 个节点。此外,当两个目标节点分别先采样出节点个数为 n 的子图,然后合并成一个子图时,有些节点重复存在,需要进行去重操作,最终采样出的子图个数仍然有可能小于 $2n$ 个,因此排序池化层仍需要存在。本研究的采样机制大大减少了排序池化操作带来的信息损失,提升了预测精度。排序池化操作后,所有子图的图表示维度保持一致。子图的特征矩阵被输入到一维卷积神经网络进行局部特征提取,随后展平进入全连接层,最终输出一个二分类的概率分布,从而预测链接的存在性。

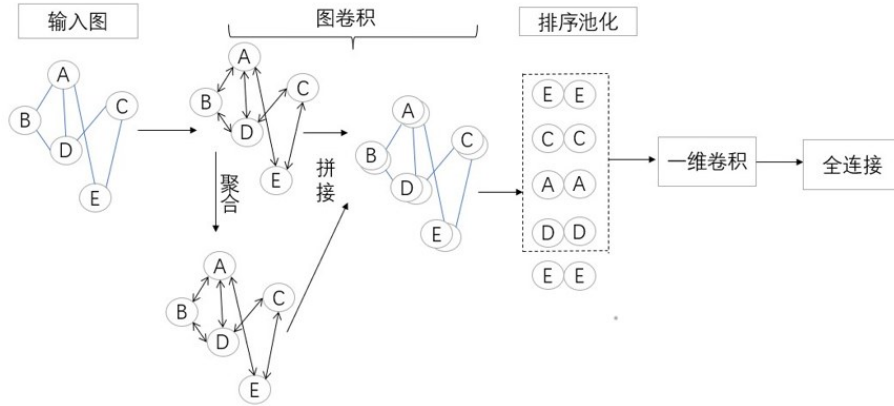


Fig. 4 Framework of DGCNN model

图4 DGCNN模型框架

4 实验方法与结果分析

4.1 实验环境与参数设置

实验在运行 Ubuntu 20.04 的工作站上进行,硬件配置为 16 核 Intel® Xeon® Platinum 8481C 处理器和 1 块显存为 24GB 的 NVIDIA RTX4090D 显卡,软件环境为 Python 3.8、PyTorch 1.11.0 和 CUDA 11.3。

负采样时 λ 设置为 1,即正样本数量和负样本数量相

同。node2vec 的图嵌入用 128 维向量表示。图卷积层有 4 层,每一层的维度为 $[32, 32, 32, 1]$,每一层均使用 tanh 激活函数。排序池化层中 k 的值为目标链接采样的所有子图中个数最多的子图节点数,即 $2n$ 。1D 卷积层有两层,通道数分别为 16 和 32,激活函数采用 ReLU。

4.2 数据集

采用 USAir、Power、Router、NS、PB、Celegans、Yeast 7 个公开网络数据集进行实验,来源于 Network Repository、KONECT^[23] 及 SNAP^[24] 数据集平台,具体信息见表 1。训练

集、验证集和测试集的比例为 85:5:10。当图中节点边的数量约等于或略大于图中节点数量时,判定该图为稀疏图。

Table 1 Dataset information

表 1 数据集信息

数据集	数据描述	节点	边	平均度	是否稀疏
USAir	美国航空交通网络	332	2 126	12.81	否
Power	美国电力输配网络	4 941	6 594	2.67	是
Router	互联网路由拓扑网络	5 022	6 258	2.49	是
NS	网络科学合作网络	1 589	2 742	3.45	否
PB	政治博客链接网络	1 223	16 714	27.36	否
Celegans	秀丽隐杆线虫神经连接网络	297	2 148	14.46	否
Yeast	酵母蛋白质互作网络	2 375	11 693	9.85	否

4.3 基线方法与评价指标

选择以下 3 类基线方法与本文方法进行比较:①启发式方法。包括 CN(Common Neighbors)、PA(Preferential Attachment)、AA(Adamic-Adar);②嵌入式方法。包括 node2vec、LINE;③基于 GNN 子图级别的链接预测方法。包括 SEAL、SPG4LP。

采用 AUC(Area Under ROC Curve)作为链接预测方法的评价标准。计算公式为:

$$AUC = \frac{\sum_{i \in P} \sum_{j \in N} I(score_i > score_j)}{|P| \times |N|} \quad (4)$$

式中: P 表示正样本集合; N 表示负样本集合; I 在条件成立取 1,否则为 0。

4.4 实验结果分析

4.4.1 与基线方法的比较

不同方法的 AUC 如表 2 所示。总体来说,基于 GNN 子图级别的链接预测方法的性能优于启发式方法和嵌入式方法。具体来说,除了在 Celegans 数据集上,本文方法的 AUC 均优于基线方法,尤其是在 Power 和 Router 这样的稀疏图场景下,本文方法的优势更加明显。在 Celegans 数据集上,本文方法的 AUC 略低于 SPG4LP,这可能是由于该数据集节点规模较小且连接较为致密,固定规模子图在扩展过程中引入较多结构相似的邻居节点,降低了子图之间的判别性。SPG4LP 更强调节点对之间的路径结构,因而在该数据集上能够更有效地刻画连接关系。

4.4.2 子图节点个数的影响

子图节点个数是影响实验结果的重要参数。在 6 个

Table 2 AUC of different methods

表 2 不同方法的 AUC

方法	USAir	Power	Router	NS	PB	Celegans	Yeast
CN	93.11	58.80	56.21	94.48	92.04	85.13	90.02
PA	88.84	44.79	47.58	69.15	90.20	74.79	81.83
AA	95.04	58.90	56.43	95.09	92.36	86.95	88.89
node2vec	91.05	75.22	66.69	92.52	85.25	84.12	94.05
LINE	85.56	56.41	68.14	81.60	78.81	70.20	87.98
SEAL	96.30	90.15	95.84	98.52	94.22	88.28	96.81
SPG4LP	96.24	86.53	91.05	97.18	95.47	90.59	96.89
HeSPS-GNN	97.19	94.41	97.21	99.25	95.86	88.12	97.80

数据集上(Celegans 数据集上的性能波动不大,不再列举),本文方法在不同子图节点个数 n 下的 AUC 如图 5 所示。可以看出,当 n 取值较小时,AUC 指标基本不会过低;当 n 取值较大时,AUC 指标不会断崖式下滑,说明本文方法可选的参数范围较大,且表现相对稳定。值得注意的是,当场景为稀疏图时,采样时选取合适的 n 值会大大提升模型性能。

4.4.3 不同子图采样策略在排序池化层的信息损失

基于 GNN 子图级别的链接预测方法,如 SEAL、SPG4LP 的预测效果受 DGCNN 中排序池化层的影响,因为不论是 h -hop 闭包图还是 k -path 简单路径图,其采样出的子图规模均具有较大差异,导致子图的特征矩阵在排序池化层会进行大量裁剪或填充,裁剪会造成信息损失,而大量填充零向量则对后续模型计算不利。为验证信息损失情况,本文在 USAir 和 Power 两个数据集上进行比较实验,在相同排序池化层 k 值下比较 h -hop 闭包图、 k -path 简单路径图和本文提出的 HeSPS 3 种策略采样出的所有子图在排序池化层中平均填充零向量的个数和裁剪率。二者的计算公式分别为:

$$\text{平均填充零向量的个数} = \frac{\text{排序池化层中填充零向量的总数}}{\text{子图总数}} \quad (5)$$

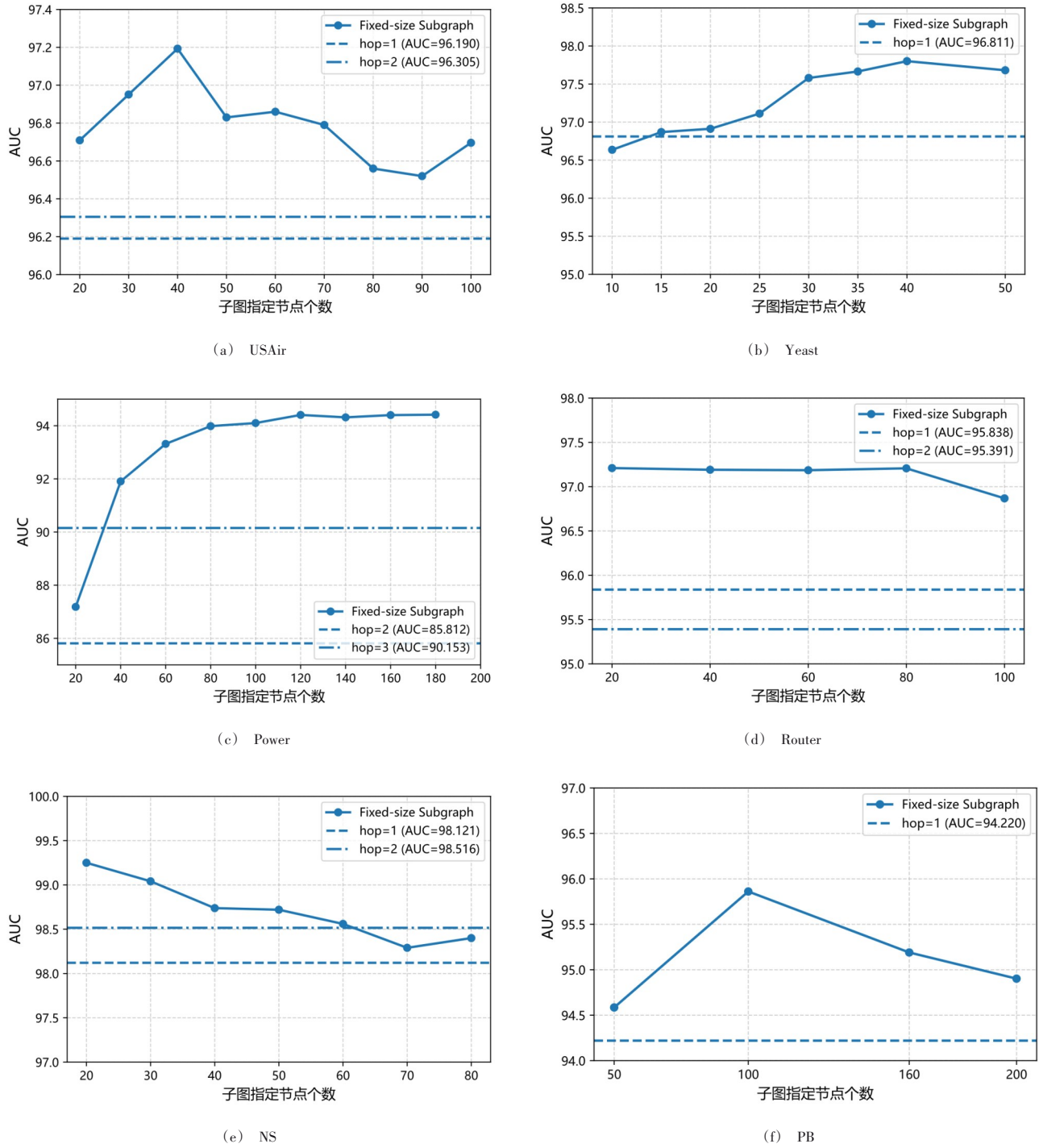
$$\text{裁剪率} = \frac{\text{裁剪后所有子图剩余边总数}}{\text{裁剪前所有子图的边总数}} \quad (6)$$

不同采样策略在 USAir 数据集上的实验结果如表 3 所示。可以看出,HeSPS 采样出的子图平均填充零向量个数显著少于 h -hop 闭包图和 k -path 简单路径图,且由于 HeSPS 的 k 值设定为所有子图的最大节点个数,不存在裁剪问题。SEAL 采用的 h -hop 闭包图采样策略在 k 值较小时,填充零向量个数不多但裁剪率很高;当 k 值较大时,裁剪率有所下降但平均填充零向量的个数大大增加。SPG4LP 采用的简单路径图采样策略得到的子图规模一般较小,填充零向量个数普遍较多,裁剪率亦为零。此外,HeSPS 采样策略在不同 k 值下的 AUC 均高于 h -hop 闭包图和 k -path 简单路径图采样策略。

不同采样策略在 Power 数据集上的实验结果如表 4 所示。该数据集为稀疏图, h -hop 闭包图采样策略在目标节点为稀疏节点时采样的子图规模往往很小,需要填充大量零向量,影响模型效果。相较而言,HeSPS 控制了子图规模的相对一致,显著减少了零向量填充个数,同样不存在裁剪问题,在稀疏图场景下显著提高了 AUC 指标。在该数据集上,简单路径图采样策略在路径长度设置较长的情况下采样节点数仍然较少,因此设置的 k 值也较小,即使如此,该采样策略的 AUC 指标仍然低于 85,因此不在表 4 中列举其数据。

5 结语

本文针对现有基于子图建模的链接预测方法存在的

Fig. 5 AUC of the proposed method under different numbers of subgraph nodes n 图5 本文方法在不同子图节点个数 n 下的 AUC

子图结构不一致、信息损失严重、稀疏节点建模能力弱以及结构保留性差等问题,提出一种预设子图节点数量的层次采样子图策略 HeSPS。该策略通过设定子图节点数上限,并采用由近及远的分层采样方式,有效控制了子图规模的一致性,适配排序池化层的处理机制,增强了稀疏节点的结构信息表达能力,同时避免了传统闭包采样中的子图膨胀问题和路径采样中的边结构丢失问题。在多个

真实图数据集上的实验结果表明,HeSPS在保持计算效率的同时显著提升了链接预测性能,尤其在稀疏图场景中表现出优越的 AUC 指标增益,验证了其有效性与泛化能力。未来工作可以从以下几个方向进一步拓展:一是结合节点属性信息与边语义信息增强子图表达能力;二是探索自适应的子图采样机制以适应多样化图结构;三是在大规模图和动态图场景中进一步验证该方法的可扩展性与实用性;

Table 3 Experimental results of different sampling strategies on the USAir dataset

表3 不同采样策略在USAir数据集上的实验结果

k	采样策略	平均填充零向量个数	裁剪率/%	AUC
20	1-hop	3.69	73.71	93.763
	6-path	6.72	0	94.689
	HeSPS	1.52	0	96.709
30	1-hop	8.03	51.23	93.985
	6-path	16.71	0	93.498
	HeSPS	2.85	0	96.951
40	1-hop	13.01	30.79	94.544
	6-path	16.71	0	93.498
	HeSPS	4.39	0	97.193
50	1-hop	18.52	17.94	95.058
	6-path	36.72	0	93.280
	HeSPS	6.19	0	96.830
60	1-hop	24.09	10.33	95.492
	6-path	46.71	0	92.116
	HeSPS	8.19	0	96.860

Table 4 Experimental results of different sampling strategies on the Power dataset

表4 不同采样策略在Power数据集上的实验结果

k	子图	平均填充零向量个数	裁剪率/%	AUC
60	3-hop	27.42	4.59	90.153
	HeSPS	11.65	0	93.316
80	3-hop	46.13	1.21	90.621
	HeSPS	16.44	0	93.371
100	3-hop	65.77	0.31	90.610
	HeSPS	21.49	0	93.985
120	3-hop	85.68	0.09	90.580
	HeSPS	26.72	0	94.101
140	3-hop	105.65	0.02	90.612
	HeSPS	32.12	0	94.406

四是分析链接预测任务在GNN模型中存在的隐私问题并研究隐私保护方法。

参考文献:

- [1] WU J, XIE H, JIANG H W. A survey of graph neural network recommender systems[J]. Computer Science and Exploration, 2022, 16(10): 2249-2263.
吴静, 谢辉, 姜火文. 图神经网络推荐系统综述[J]. 计算机科学与探索, 2022, 16(10): 2249-2263.
- [2] XIE Z W. Research on knowledge graph completion methods based on graph neural networks[D]. Wuhan: Wuhan University, 2023.
谢志文. 基于图神经网络的知识图谱补全方法研究[D]. 武汉: 武汉大学, 2023.
- [3] CHEN A Q, CHEN R, KUANG Z F, et al. Imbalanced fraud detection based on graph neural networks[J]. Computer Engineering, 2023, 49(11): 150-159.
陈安琪, 陈睿, 卞祝芳, 等. 基于图神经网络的不平衡欺诈检测研究[J]. 计算机工程, 2023, 49(11): 150-159.
- [4] FOUT A, BYRD J, SHARIAT B, et al. Protein interface prediction using graph convolutional networks [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6530-6539.
- [5] YAO L, WANG L, PAN L, et al. Link prediction based on common-neighbors for dynamic social network [J]. Procedia Computer Science, 2016, 83: 82-89.
- [6] NIWATTANAKUL S, SINGTHONGCHAI J, NAENUDORN E, et al. Using of Jaccard coefficient for keywords similarity [C]//Proceedings of the International Multiconference of Engineers and Computer Scientists, 2013: 380-384.
- [7] ADAMIC L A, ADAR E. Friends and neighbors on the Web [J]. Social Networks, 2003, 25(3): 211-230.
- [8] NEWMAN M E J. Clustering and preferential attachment in growing networks [J]. Physical Review E, 2001, 64(2): 025102.
- [9] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014: 701-710.
- [10] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 855-864.
- [11] TANG J, QU M, WANG M, et al. Line: large-scale information network embedding [C]//Proceedings of the 24th International Conference on World Wide Web, 2015: 1067-1077.
- [12] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [DB/OL]. <https://arxiv.org/abs/1609.02907>.
- [13] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 1025-1035.
- [14] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks [DB/OL]. <https://arxiv.org/abs/1710.10903>.
- [15] ZHANG M, CHEN Y. Link prediction based on graph neural networks [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 5171-5181.
- [16] ZHANG M, CHEN Y. Weisfeiler-lehman neural machine for link prediction [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 575-583.
- [17] TERU K, DENIS E, HAMILTON W. Inductive relation prediction by sub-graph reasoning [C]//International Conference on Machine Learning, 2020: 9448-9457.
- [18] ZHU Z, ZHANG Z, XHONNEUX L P, et al. Neural bellman-ford networks: a general graph neural network framework for link prediction [J]. Advances in Neural Information Processing Systems, 2021, 34: 29476-29490.
- [19] LI Z R, ZHENG W G. Link prediction based on simple path graph [J/OL]. Computer Engineering, 2025-06-26. <https://doi.org/10.19678/j.issn.1000-3428.0070041>.
李志仁, 郑卫国. 基于简单路径图的链接预测 [J/OL]. 计算机工程, 2025-06-26. <https://doi.org/10.19678/j.issn.1000-3428.0070041>.
- [20] LOUIS P, JACOB S A, SALEHI-ABARI A. Simplifying subgraph representation learning for scalable link prediction [DB/OL]. <https://arxiv.org/abs/2301.12562>.
- [21] LOUIS P, JACOB S A, SALEHI-ABARI A. Sampling enclosing subgraphs for link prediction [C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022: 4269-4273.
- [22] XIANG Z, WANG T, WANG D. Preserving node-level privacy in graph neural networks [C]//2024 IEEE Symposium on Security and Privacy, 2024: 4714-4732.
- [23] KUNEGIS J. KONECT: the Koblenz network collection [C]//Proceedings of the 22nd International World Wide Web Conference Companion, 2013: 1343-1350.
- [24] LESKOVEC J, KREVL A. SNAP datasets: Stanford large network dataset collection [EB/OL]. <https://snap.stanford.edu/data/>.

(责任编辑: 尹晨茹)