计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(4): 863−875, 2025

基于随机游走路径的自监督图拓扑不平衡学习

秦者云[1] 卢宪凯[1] 袭肖明[2] 任春晓[3] 聂秀山[2] 尹义龙[1]
[1](山东大学软件学院 济南 250101)
[2](山东建筑大学计算机科学与技术学院 济南 250101)
[3](山东省国际人才交流服务中心 济南 250101)
(zheyunqin@gmail.com)

Self-Supervised Graph Topology-Imbalance Learning Based on Random Walk Paths

Qin Zheyun[1], Lu Xiankai[1], Xi Xiaoming[2], Ren Chunxiao[3], Nie Xiushan[2], and Yin Yilong[1]
[1](*School of Software*, *Shandong University*, *Jinan* 250101)
[2](*School of Computer Science and Technology*, *Shandong Jianzhu University*, *Jinan* 250101)
[3](*Shandong International Talent Exchange Service Center*, *Jinan* 250101)

Abstract The problem of topological imbalance in graphs, arising from the non-uniform and asymmetric distribution of nodes in the topological space, significantly hampers the performance of graph neural networks. Current research predominantly focuses on labeled nodes, with relatively less attention given to unlabeled nodes. To address this challenge, we propose a self-supervised learning method based on random walk paths aimed at tackling the issues posed by topological imbalance, including the constraints imposed by homogeneity assumptions, topological distance decay, and annotation attenuation. Our method introduces the concept of multi-hop paths within the subgraph neighborhood, aiming to comprehensively capture relationships and local features among nodes. Firstly, through a strategy of aggregating between paths, we can learn both homogeneous and heterogeneous features within multi-hop paths, thereby not only preserving the nodes' original attributes but also maintaining their initial structural connections in the random walk sequences. Additionally, by combining a strategy of aggregating subgraph samples based on multiple paths with structured contrastive loss, we maximize the intrinsic features of local subgraphs for the same node, enhancing the expressive power of graph representations. Experimental results validate the effectiveness and generalization performance of our method across various imbalanced scenarios. This research provides a novel approach and perspective for addressing topological imbalance issues.

Key words graph topology-imbalance; self-supervised graph representation learning; random walk path; homogeneity assumption; topological distance decay

摘 要 图拓扑不平衡问题是由于节点在拓扑空间中的不均匀和不对称分布,对图神经网络性能产生了严重的负面影响.当前的研究主要侧重于标记节点,而对无标记节点的关注较少.为应对这一挑战,提出了一种基于随机游走路径的自监督学习方法,旨在解决拓扑不平衡问题带来的同质性假设限制、拓扑距离衰减以及注释衰减等难题.所提方法引入了多跳路径的子图邻域概念,以更全面地捕捉节点之间的关系和局部特征.首先,通过路径间聚合策略学习多跳路径中的同质和异质特征,不仅保留了节点的原始属性,而且维护了它们在随机游走序列中的初始结构连接.此外,结合了基于多条路径的子图采样和子图生

收稿日期: 2023-10-11; **修回日期**: 2024-10-31
基金项目: 国家自然科学基金重点项目(U23A20389)
This work was supported by the Key Program of the National Natural Science Foundation of China (U23A20389).
通信作者: 尹义龙(ylyin@sdu.edu.cn)

成策略以及结构化的对比损失,最大化了同一节点局部子图的内在特征,从而增强了图表示的表达能力.经过实验验证,所提方法在多种不平衡场景下都表现出了出色的有效性和泛化性能.这一研究为解决图拓扑不平衡问题提供了新的方法和视角.

关键词 图拓扑不平衡;自监督图表示学习;随机行走路径;同质性假设;拓扑距离衰减

中图法分类号 TP181

DOI: 10.7544/issn1000-1239.202330813 **CSTR:** 32373.14.issn1000-1239.202330813

近年来,图数据在各个领域的广泛应用已经引起了广泛的关注和研究[1-2].图数据的独特之处在于其能够以节点和边的方式捕捉实体之间复杂的关系和结构.然而,这种复杂性也给传统的数据处理方法带来了挑战.为了更好地利用图数据中的信息,图表示学习已经成为分析图结构数据的重要策略.该领域旨在将图中的节点映射到低维稠密嵌入中来保留图的属性和结构特征,以便在后续的分析任务中使用.作为图表示学习的重要分支,图神经网络(graph neural networks, GNNs)[3]通过逐层的消息传递和聚合来捕捉节点的上下文信息,从而生成节点的嵌入表示.这种方法的优势在于它能够保留图的局部结构和全局属性,有助于更好地捕捉节点之间的关系.然而,由于图固有的不对称性和不均匀性,GNNs在学习图时面临着一个特殊的挑战,即拓扑不平衡问题[4].

拓扑不平衡问题是由于图节点之间复杂的连接关系导致的不平衡性,这表现为节点在拓扑空间中位置的不对称和不均匀分布.这种不平衡性影响了传统的图表示学习方法,使得它们难以准确地捕捉节点之间的关系.尽管近年来图上的不平衡学习引起了广泛的研究兴趣,但大多数研究侧重于解决类不平衡问题,即每个类别的标记节点数量不平衡.拓扑不平衡问题是最近提出并受到广泛关注的一个问题,它与节点之间的连接和拓扑结构有关.ReNode[4]从标签传播的角度理解了拓扑不平衡问题,并提出了一种样本重加权方法.然而,ReNode基于同质性假设将节点拓扑边界作为决策边界,在现实世界中并不适用.PASTEL[5]则从监督信息分布不足和过度挤压2个新视角探索了拓扑不平衡问题,并通过位置编码和类内冲突来增强类内归纳偏差.HyperIMBA[6]使用双曲几何嵌入来衡量标记节点层次,重新加权监督信息的权重.然而,这些方法主要集中于标记节点的拓扑不平衡问题,而未充分探讨无标记节点的拓扑不均衡问题.

本文将自监督图拓扑不平衡学习(self-supervised graph topology-imbalance learning)定义为一个特定于图的不平衡学习问题,专注于图数据中拓扑不平衡导致的学习到的自监督表示质量较差的问题.相对于已有的基于标记节点的拓扑不平衡研究,本文面临3个挑战:1)同构图假设限制.现有方法隐含地假设输入图是同构的,这限制了模型对图的表达能力.如图1所示,同跳邻居节点(v_2,v_4)和不同跳邻居节点(v_2,v_3)对中心节点v_1的影响不同.2)拓扑距离衰减.节点的影响力随着拓扑距离的增加而衰减,导致远离中心节点的邻域节点无法带来有用的信息.如图1所示,v_1的4跳邻居v_3带给v_1的有效信息远小于1跳邻居v_2和v_4.3)注释限制.自监督学习需要从数据本身挖掘监督信号,可利用的监督信息天然弱于标记节点.

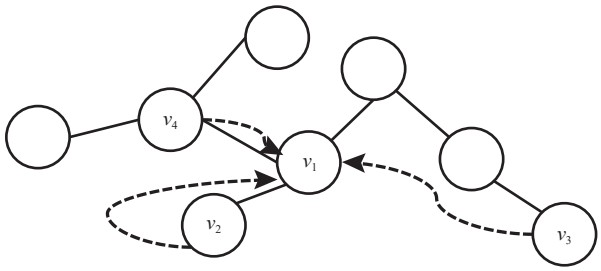


Fig. 1 Challenges in self-supervised topological-unbalance problems

图1 自监督拓扑不平衡问题中所面对的挑战

为了克服自监督图学习中的拓扑不平衡限制,本文提出了一种基于随机游走路径的自监督图拓扑不平衡学习框架.首先,本文利用随机游走获取全局结构嵌入信息,通过基于注意力机制的聚合多跳路径以更精确地捕捉节点的高阶邻域信息.随机游走形成的多跳路径不仅保留了这多个节点的原始属性,还维护了它们在随机游走序列中的初始结构连接[7].这种基于路径的邻域扩展方式相对于传统方法更能够有效地表达目标节点的同质和异质的邻域分布特征,从而突破同质性假设的限制.其次,为了获取较远距离节点的有效信息,本文提出了一种基于聚合多跳路径的子图采样策略,专注于挖掘局部拓扑关系作为监督信号,关注更细粒度的局部结构信息,以

减轻因节点之间距离过远而导致的信息不足问题，学习更具有判别力的局部拓扑特征. 最后，为了学习到更强的自监督信号，本文聚焦于图的天然属性，即拓扑结构. 本文采用针对节点匹配的沃瑟斯坦距离（Wasserstein distance, WD）和针对边匹配的格罗莫夫-沃瑟斯坦距离（Gromov-Wasserstein distance, GWD）[8]来构建结构对比损失.

本文的核心贡献在于提出了一种基于随机游走的子图采样方法，以应对图拓扑不平衡的挑战. 该方法具有 3 个特质：1）异质性. 有效地捕获中心节点周围的局部结构信息，包括同质性和异质性结构信息. 这一特性突破了同构图假设和拓扑衰减的限制，能够更好地理解和分析具有复杂结构的图数据. 2）灵活性. 通过可学习的参数来灵活控制随机游走的行为，使其更加适应不同的图结构和任务需求. 这种灵活性使得我们的方法能够适应各种不同类型的图数据，从而具有更广泛的适用性和实用性. 3）随机性. 引入了一定程度的随机性，从而能够在一定程度上避免陷入局部最优解，增加了算法的鲁棒性和泛化性能.

在接下来的章节中，本文将详细介绍本文提出的基于随机游走路径的自监督图拓扑不平衡学习框架，并通过在多个公开数据集上的实验证明其有效性. 本文的主要贡献包括：

1）在自监督领域扩展了图拓扑不平衡问题，着重研究了由拓扑不平衡引起的低质量图表示学习问题.

2）基于聚合多跳路径获取全局高阶邻域信息，以更好地建模节点对之间的同质性和异质性关系；并且基于聚合多跳路径提取子图，学习稀疏且紧凑的子图局部拓扑特征，增强远距离节点的贡献；最大化结构对比损失来增强类内归纳偏差，基于图天然的拓扑属性增强自监督信号的质量.

3）实验结果证明了本文方法在各种不平衡场景下的有效性和泛化性.

1 相关工作

本节主要介绍了图表示学习、图对比自监督学习、图不均衡学习以及图结构学习的研究方法和进展.

1.1 图表示学习

近年来，图神经网络在图结构数据的表示学习方面取得了显著的研究进展. 这些网络通过将图的拓扑结构和节点的特征相结合，实现了对节点表示的学习. 通常，传统的图神经网络包括 2 个主要步骤：邻域节点的信息聚合和特征变换. 首先，通过聚合相邻节点的表示信息以及自身节点的表示信息来更新每个节点的特征表示. 接下来，通过共享的线性变换将每个节点的特征映射到一个新的特征空间中. 例如，ChebyNet[9]引入了快速局部卷积滤波器，从而有效地减少了计算成本；基于注意力机制的模型[10-11]通过考虑相邻节点之间的交互关系来计算权重，从而取代了传统的基于节点度的权重计算. 然而，这些方法通常需要完整的图数据作为输入，这在处理大规模图数据时效率可能较低. 为了应对上述问题，GraphSAGE[12]提出了一种基于采样的方法以提高大规模图数据的处理效率. 该方法通过对中心节点的邻域节点进行小批量采样，然后聚合这些采样节点的特征来更新中心节点的表示，进而迭代地更新整个图的节点表示.RAW-GNN[7]基于路径设计了多种组合策略，同时学习图的同质性与异质性信息. 然而，现有的方法主要集中在有监督学习框架下依赖大量标记样本，而获取这些标记样本通常需要耗费大量的人力和时间.

需要注意的是，前述研究大多基于同质图的假设[13-14]，即假设连接的节点具有相似的特征. 这一同质性假设限制了模型对异构图的适应能力. 在实际应用中，许多图数据都呈现出异构性，即连接的节点可能属于不同的类别或具有不同的特征. 例如，蛋白质之间的相互作用有不同类型，如物理相互作用和代谢通路. 然而，传统的同质图方法难以有效地捕捉这些不同类型的相互作用，限制了对蛋白质关系的准确建模. 同样，社交网络中的人际连接关系也被视为异构图，因为不同类型的连接表示不同的社交关系，如朋友、同事和家庭关系. 传统的同质图方法难以区分和利用这些不同类型的连接，从而无法准确地捕捉社交网络中的多样性和复杂性.

为了解决这些问题，一些方法基于元路径的概念设计了特定于异构图的图表示学习框架. 例如，Metapath2vec[15]生成由单个元路径引导的随机漫步，然后将其馈送到 skip-gram 模型以生成节点嵌入. HIN2vec[16]进行多种预测训练任务来学习异构图的节点和元路径表示. HERec[17]基于元路径邻居将异构图转换为同构图，并应用深度随机游走模型来学习目标类型的节点嵌入. HAN[18]以类似的方式将异构图转换为多个基于元路径的同构图，但使用图注意力网络架构来聚合来自邻居的信息，并利用注意力机制来组合各种元路径. 然而，这些方法仍然存在监督

学习框架的限制,依赖大量标记样本来获得高质量的监督信号.相比之下,本文方法基于聚合多跳路径来学习高阶全局嵌入和局部子图拓扑结构,以提高图表示的质量,并且能够更好地处理异构图的复杂性.

1.2 图对比自监督学习

图对比自监督学习已经成为图表示学习领域的主要研究方向,其核心思想是通过鼓励保持标签不变分布之间的一致性以及对其他分布之间的一致性来学习图表示.这个范式的成功在很大程度上依赖于图增强和负样本采样策略.图增强方法通过合理的图数据变换扩展了数据集并提升了模型的泛化能力.一些方法主要关注对图中的节点或边进行操作,例如,修改节点特征或调整邻接矩阵.此类方法包括对节点特征进行修改[19-21],以及对节点进行静态表示学习[22].另一些方法则专注于添加或删除图中的边来调整图的结构[23-24].在强调均匀性的背景下,图对比方法通常采用负样本采样策略来生成用于对比学习的负样本.图对比方法可以借鉴自然语言处理领域的一些技巧,例如引入负样本对来构建对比任务以及采用负样本队列或批量负样本的方式.这2种策略都旨在为模型提供足够的负样本,以使其更好地学习图表示.GSC[8]采用插值的方式生成与采样子图结构相似的有意义的负样本,并使用基于最优传输的结构对比损失优化嵌入空间.然而其仅仅通过宽度优先搜索的方式采集中心节点的近距离邻居,没有考虑到拓扑距离衰减的问题,不能充分的应对图拓扑不平衡的挑战.受GSC启发,本文关注到子图结构带来的局部拓扑信息对图表示的质量有着重要的影响.因此,提出了一种基于路径邻域的负样本采样策略,该策略关注包含了远距离节点信息的局部拓扑信息,并使用采样子图构造结构对比损失来提高图表示的质量.这种策略有助于解决图拓扑不平衡的挑战,从而更好地捕捉图数据的特征.

1.3 图不均衡学习

迄今为止,图不平衡学习的研究主要集中在处理类不平衡问题,其中每个类别的标记节点数量存在明显的不平衡[5,25].这种情况下,模型通常对多数类别的推理表现优于少数类别[26].为了解决类不平衡问题,已有的方法可以大致分为2类:数据级重采样和算法级重加权.数据级重采样方法通常在数据选择阶段重新采样[27]或增加数据[28]以平衡每个类的数据数量.这种方法的目标是通过改变数据分布来解决不平衡问题.算法级重加权方法[29]在训练阶段根据数据的数量对不同数据样本进行权重调整,以便更多地关注少数类别的样本.这种方法的重点是在训练过程中调整样本的权重,以平衡不同类别之间的贡献.然而,直接将这种方法应用于图数据时,通常忽略了图特有的拓扑属性.ReNode[4]针对图拓扑不平衡问题,采用了类似于重加权方法的思路,但是仍然遵循了同质性假设,从而限制了模型的表达能力.PASTEL[5]提供了一种新的视角来解决图拓扑不平衡问题,它通过标签重分布来缓解拓扑不平衡问题,以最大化类内标签的影响.HyperIMBA[6]利用双曲几何嵌入来衡量标记节点的层次结构.随后,根据节点的层次结构调整标签信息传播,改变目标边际,从而解决层次不平衡带来的问题.然而,文献[4–6]方法都是基于标记节点的监督信息不足而设计的,没有充分考虑无标记节点的情况.相比之下,本文将图拓扑不平衡问题引入了自监督图表示学习领域,通过考虑子图邻域信息的方式,突破了图拓扑不均衡问题对自监督图表示学习的限制.这一拓展为解决图数据中的不平衡性提供了新的视角,特别关注了图的拓扑结构不均衡问题,而不仅仅是标记节点的问题.这种方法有望更好地捕捉图数据的复杂性和不均衡性,从而提高图表示学习的质量.

1.4 图结构学习

图结构学习[30]旨在通过学习优化的图结构来提高图表示学习的质量.目前,研究主要集中在提升图神经网络(GNN)模型的鲁棒性.然而,也有一些研究工作致力于探索结构学习对于改进图表示的潜力.例如,GCN-LPA[31]赋予同一类别的2个节点之间的边不同的权重,以获得更优的表示结果.SDRF[32]则进一步考虑了Ricci曲率,并根据其对边进行了重新排列.在研究节点之间的关系建模时,使用了多种测量方法[33-34],包括节点特征、节点度、节点编码和边属性.此外,子图结构在图表示学习中也备受关注,因为它被证明能够更好地捕捉节点之间的关系.通过挖掘子图的局部特征和结构信息,可以更准确地描述节点之间的联系.在这项研究中,本文提出了一种子图结构学习策略,旨在应对图拓扑不平衡的问题.本文利用宽度优先搜索和深度优先搜索方法,更有效地捕获节点之间的底层联系,增强图表示学习的性能.

2 自监督图拓扑不平衡学习方法

本节将介绍本文提出的自监督框架.首先,在2.1节中提出了基本符号的定义.接着,在2.2节中介

绍了通过随机游走聚合的多跳路径来获取包含同质和异构信息的全局高阶图嵌入. 然后, 在 2.3 节中描述了如何从聚合的多跳路径中提取局部子图, 并自适应生成对比子图以构建正负样本. 然后用沃瑟斯坦结构对比损失和格罗莫夫-沃瑟斯坦结构对比损失分别优化子图的节点和边, 以提高图表示质量.

2.1 基本符号

本文将一个包含N个节点的无向无权网络定义为$G=(V,E)$, 其中包括节点集合$V=\{v_n|n\in\mathbb{N}\}$和边集合$E=\{e_{ij}\}\subseteq V\times V$. 在实际操作中, 边集合由邻接矩阵$\boldsymbol{A}=(a_{ij})\in\mathbb{R}^{n\times n}$表示, 其中$a_{ij}=1$表示节点$v_i$和$v_j$之间有边相连, 否则$a_{ij}=0$.

接下来, 本文定义一条基于随机游走的长度为K的路径P_i^r, 它由节点v_i作为中心节点和漫步者的出发节点, 通过随机游走$K-1$个节点构成, 即$P_i^r=\{v_{p_k}^r|k\in K\}$. 对于中心节点$v_i$, 不同的漫步者会生成不同的路径. 在这里, r表示随机游走策略.

2.2 基于聚合多跳路径的图嵌入

为了有效地捕捉目标节点的同质和异构邻域分布特征, 本文引入了基于注意力机制的路径聚合方法, 将同质性和异质性信息聚合到一个统一的嵌入空间中. 具体而言, 本文执行近距离邻域和远距离邻域 2 种随机游走策略, 分别采样中心节点的同质性邻域分布和异质性邻域分布. 每个采样策略都会生成多条路径, 以更准确地捕获邻域的分布情况(详见 2.2.1 节). 接下来, 本文执行路径间信息聚合, 分别针对从近距离节点和远距离节点策略中采样得到的路径, 利用注意力机制学习不同路径的重要性, 将不同路径特有的信息聚合在一起, 分别形成特定于同质性和异质性的全局嵌入信息. 最后执行同/异质路径聚合, 以获得全局嵌入表示(详见 2.2.2 节).

2.2.1 随机游走路径采样

为了更全面地捕获中心节点的邻域信息, 本文使用近距离策略和远距离策略 2 种随机游走策略来采样多跳路径. 近距离策略以较小的跳数采样邻域信息, 主要捕获同质性信息; 远距离策略则采样更远的邻域信息, 以捕获异质性信息. 为了最大限度地保留节点的网络邻域分布和学习更丰富表示, 本文采用一个有偏随机游走过程[35], 可以有效地探索不同的邻域.

考虑一个已经遍历了边$e_{i-1,i}$, 现在在节点i上的随机游走, 并将要访问下一个节点$i+1$. 本文将非归一化的转移概率定义为

$$P_{pq}=\begin{cases}\dfrac{1}{p}, \text{ if } d_{i-1,i+1}=0,\\ 1, \text{ if } d_{i-1,i+1}=1,\\ \dfrac{1}{q}, \text{ if } d_{i-1,i+1}=2,\\ 0, \text{其他},\end{cases} \tag{1}$$

其中$d_{i-1,i+1}$表示节点v_{i-1}和v_{i+1}之间的最短路径长度, $d_{i-1,i+1}=1$表示v_{i+1}是v_{i-1}的 1 跳邻居. 当本文设置$p<1$且$q>1$时, 随机游走的倾向是类似于广度优先搜索, 它会收集目标节点的相邻节点的信息, 并且有助于提取同质性信息, 如图 2 所示. 而当$p>1$且$q<1$时, 随机游走的倾向是类似于深度优先搜索, 它更倾向于按照顺序到达距离源节点越来越远的节点, 从而有助于提取异质性信息.

2.2.2 路径间信息聚合

在随机游走策略r的指导下, 多条采样路径可以有效地探索中心节点的拓扑结构. 在多跳路径采样之后, 本文需要将不同路径的信息进行聚合, 以获取

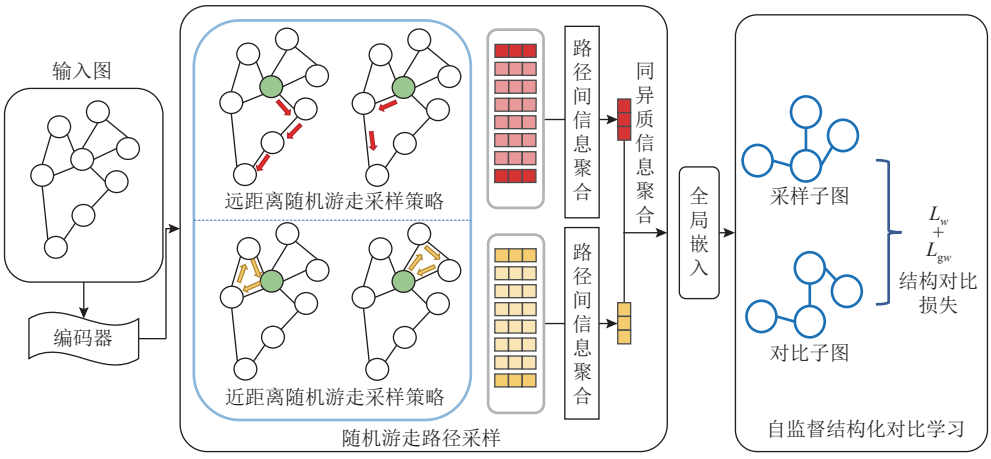


Fig. 2 Illustration of our method

图 2 本文方法的示意图

目标节点的全局邻域信息. 本文遵循路径同分布的假设[7]: 在给定某种采样策略的情况下, 生成的路径遵循相应的分布. 由于所有路径都是从同一个分布中采样的, 因此将这些路径进行求和可以近似得到相应的分布. 而由于远距离节点根据距离不同, 对中心节点贡献的衰减程度不同, 因此不能简单地通过节点嵌入求和的方式来近似分布. 为此, 本文引入了注意力机制, 用于学习中心节点v_i的P_i^r中不同路径的重要性权重, 并将路径信息组合起来(为了简化表示, 省略上标r).

$$s_p = GRU\left(\left\{g_{p_k} | k \in K\right\}\right), \tag{2}$$

$$\boldsymbol{W}_{\text{path}} = softmax\left(\theta\left(\boldsymbol{W}_{\text{node}}^{\text{T}} s_p\right)\right), \tag{3}$$

其中g_{p_k}是节点v_{p_k}使用编码器获取的编码特征, 然后使用序列编码器 GRU(gate recurrent unit)[36] 获取每个路径编码特征s_p, $\boldsymbol{W}_{\text{node}}$是可学习的注意力参数, $\boldsymbol{W}_{\text{path}}$表示归一化后的每个路径的权重. $\theta(\cdot)$是 LeakyReLU 激活函数, *softmax* 函数用于归一化权重.

为了进一步稳定学习过程, 本文采用了标准的多头注意力聚合全局嵌入, 然后将它们的嵌入连接在一起, 形式化表示为:

$$e_i^h = \boldsymbol{W}_{\text{head}}\left(\sum_{p \in P_i} \boldsymbol{W}_{\text{path}} s_p\right), \tag{4}$$

$$e_i = cat\left(e_i^h\right), h \in \{1,2,\ldots,H\}, \tag{5}$$

其中$\boldsymbol{W}_{\text{head}}$是可学习的多头注意力权重, $cat(\cdot)$表示特征拼接, e_i是中心节点v_i的随机游走策略r的全局嵌入. 最后, 为了聚合中心节点v_i的同/异质信息, 本文采用串联的方式将近距离策略($r=1$)和远距离策略($r=2$)得到的全局嵌入进行拼接, 从而得到中心节点v_i的全局嵌入e_i, 用于后续的图嵌入学习.

$$e_i = cat\left(e_i^1, e_i^2\right). \tag{6}$$

2.3 自监督结构化对比学习

本文需要基于图的整体结构嵌入生成正负样本对, 其中正样本对应于相似的节点, 而负样本对应于不相似的节点. 为了实现这一目标, 本文首先基于聚合多跳路径进行子图采样, 然后根据目标节点的邻域结构自动生成对比子图. 为了准确描述子图之间的几何差异, 本文引入了沃瑟斯坦结构化距离和格罗莫夫-沃瑟斯坦距离作为度量对比子图相似性的方式, 以表示构建样本之间的对比损失.

2.3.1 基于聚合多跳路径的子图采样

具体来说, 本文首先基于聚合多跳路径进行子图采样, 如图 3 与算法 1 所示, 力求关注更远距离的节点. 本文随机选择一个节点v_i作为中心节点. 从中心节点v_i开始, 按照多跳路径进行子图采样. 这个路径的长度由参数K, 即采样路径的最大长度控制. 首先, 采样 1 阶邻域节点, 这是与中心节点v_i直接相连的节点. 如果采样路径的长度仍未达到目标长度K, 则采用远距离随机游走策略继续采样. 这时使用了式(1)中定义的参数$p>1$和$q<1$, 它们控制了游走的随机性和偏向性以获取更远距离的节点信息. 继续采样, 直到采样路径的长度达到了目标长度K. 最终得到了一个采样子图, 该子图包含了中心节点v_i以及与其距离不超过K的节点和边. 这个采样子图被记为P_i^{o}, 其中o表示用于子图采样的随机游走策略.

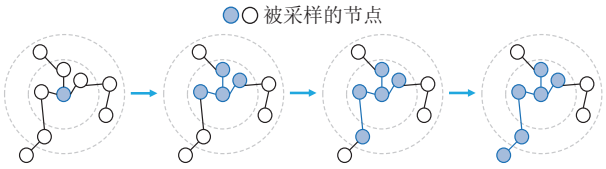


Fig. 3 Illustration of sub-graph sampling process

图 3 子图采样过程示意图

算法 1. 子图采样伪代码.

输入: 图G, 中心节点v_i, 采样路径长度K, 参数p, 参数q;

输出: 采样子图P_i^{o}.

① 初始化: 采样路径长度$len_path=0$, 采样子图P_i^{o}包含中心节点v_i;
② while $len_path < K$
③ if $len_path = 1$
④ 随机选择 1 阶邻域节点v_j与v_i相连;
⑤ 将边(v_i, v_j)加入采样子图P_i^{o};
⑥ $len_path = +1$;
⑦ 将v_j设为新的中心节点v_i;
⑧ else
⑨ 根据概率参数p和q进行随机游走, 选择下一个节点v_k;
⑩ 将边(v_i, v_k)加入采样子图P_i^{o};
⑪ $len_path = +1$;
⑫ 将v_k设为新的中心节点v_i;
⑬ end if
⑭ end while
⑮ return 采样子图P_i^{o}.

完成采样子图后, 需要为子图构造合适的对比视图(contrastive view), 这是基于对比的图神经网络自监督学习方法的核心. 大多数现有的对比图神经网络学习方法采用预定义的对比视图, 例如随机丢

弃、遮挡部分节点特征或者随机丢弃节点之间的边. 此类方法通常无法适用于新的输入或保持原图数据的结构(因为图数据的异质性无法很好地迁移这些预定义的任务). 并且不同于图像的增广方法那样不一定能保证底层语义的稳定性[37]. 为了保持对比视图的拓扑结构与采样子图相似, 构造的对比视图与采样子图拥有相同的节点数量和边连接关系, 并通过采样路径中节点之间的嵌入关系来生成对比视图节点的嵌入. 节点 $v_j \in P_i^{\mathrm{o}}$ 和节点 $v_l \in P_i^{\mathrm{o}}$ 之间的嵌入关系通过聚合节点相似性实现:

$$sim\left(v_j, v_l\right) = \theta\left(\boldsymbol{W}_c[\boldsymbol{W}_v v_j \| \boldsymbol{W}_v v_l]\right), \tag{7}$$

其中 $\boldsymbol{W}_c$ 和 $\boldsymbol{W}_v$ 是可学习的权重矩阵.

与式(2)相似, 归一化权重后, 通过聚合的方式为对比视图中的每个节点生成新的嵌入

$$W_l = softmax\left(sim\left(v_j, v_l\right)\right), \tag{8}$$

$$\widehat{v_j} = \sum_{l=1}^{P_i^{\mathrm{o}}} W_l v_l. \tag{9}$$

由于不同的对比视图需要引入充足的方差才能保证模型有良好的性能[37]. 本文采用较低的概率来进行边缘扰动, 在引入方差的同时保证对比视图的拓扑关系. 通过这些步骤, 本文得到了包含节点特征和边的对比视图, 也成为对比子图, 记作 $\hat{P}_i^{\mathrm{o}}$.

2.3.2 基于最优传输的结构对比损失

在基于聚合多跳路径捕获的子图和构造的对比视图中, 包含了局部拓扑节点分布, 因此使用基于拓扑信息的结构对比损失能够有效地挖掘图的局部信息. 本文使用沃瑟斯坦对比损失来约束正负样本的节点相似性, 使用格罗莫夫-沃瑟斯坦对比损失来约束正负样本的边相似度. 通过构造基于节点和边的结构对比损失增强监督信号, 缓解拓扑距离衰减带来的负面影响.

沃瑟斯坦对比损失遵循一般的 InfoNEC 范式, 用于衡量正样本对和负样本对之间的相似性:

$$L_{\mathrm{w}} = \frac{-1}{N(M+1)} \sum_{i=1}^{N} \left[\log S_1 + \sum_{j=1}^{M} \log S_2 \right], \tag{10}$$

其中 $S_1 = \exp\left(-\dfrac{D_{\mathrm{w}}(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}})}{\tau}\right)$, $S_2 = 1 - \exp\left(-\dfrac{D_{\mathrm{w}}(P_i^{\mathrm{o}}, P_-^{\mathrm{o}})}{\tau}\right)$, N 是采样的子图数量, M 是每个子图的负样本数量, τ 是温度参数. 同一个训练批次中, 对于 P_i^{o}, 对比子图 $\hat{P}_i^{\mathrm{o}}$ 作为正样本, 其他采样子图可作为负样本 P_-^{o}. $D_{\mathrm{w}}(\cdot,\cdot)$ 表示样本节点之间的沃瑟斯坦距离函数. 通过使用基于沃瑟斯坦距离的对比损失, 可以最大化正子图之间的节点相似性, 最小化负子图之间的节点相似性.

沃瑟斯坦距离用于衡量子图的节点之间的相似性. 以 P_i^{o}, $\hat{P}_i^{\mathrm{o}}$ 为例, 沃瑟斯坦距离可以表示为:

$$c\left(e_i, \hat{e}_j\right) = \exp\left(-\frac{\delta\left(e_i, \hat{e}_j\right)}{\tau}\right), \tag{11}$$

$$D_{\mathrm{w}}\left(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}}\right) = \min_{\boldsymbol{T}} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} c\left(e_i, \hat{e}_j\right), \tag{12}$$

其中 $\boldsymbol{T} \in \pi(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}})$, $\pi(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}})$ 表示 2 个子图节点之间的所有联合分布. $c(e_i, \hat{e}_j)$ 表示子图 P_i^{o} 中节点 v_i 与子图 $\hat{P}_i^{\mathrm{o}}$ 中节点 v_j 之间的传输成本, e_i 和 $\hat{e}_j$ 表示节点特征, τ 是温度参数, $\delta(\cdot)$ 表示节点特征之间的余弦相似性. 矩阵 $\boldsymbol{T}$ 表示传输计划, 其中 T_{ij} 表示从 v_i 到 v_j 转移的质量. 可以通过应用 Sinkhorn-Knopp 算法与熵正则化器[38-39]来获得最小化的 $\boldsymbol{T}$:

$$D_{\mathrm{w}}\left(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}}\right) = \min_{\boldsymbol{T}} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} c\left(e_i, \hat{e}_j\right) + \kappa h(\boldsymbol{T}), \tag{13}$$

$$h(\boldsymbol{T}) = \sum_{ij} T_{ij} \log T_{ij}, \tag{14}$$

其中 $h(\boldsymbol{T})$ 作为额外的正则化项, 减少节点传输的随机性, 避免大多数节点传输到 1 个节点. κ 是控制传输平滑度的参数, $\kappa > 1$. 本文使用一个较小的 κ 值来确保将 P_i^{o} 的每个节点传输给 $\hat{P}_i^{\mathrm{o}}$ 的一个节点.

与沃瑟斯坦对比损失相似, 格罗莫夫-沃瑟斯坦对比损失定义为:

$$L_{\mathrm{gw}} = \frac{-1}{N(M+1)} \sum_{i=1}^{N} \left[\log S_3 + \sum_{j=1}^{M} \log S_4 \right], \tag{15}$$

其中 $S_3 = \exp\left(-\dfrac{D_{\mathrm{gw}}(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}})}{\tau}\right)$, $S_4 = 1 - \exp\left(-\dfrac{D_{\mathrm{gw}}(P_i^{\mathrm{o}}, P_-^{\mathrm{o}})}{\tau}\right)$, $D_{\mathrm{gw}}(\cdot,\cdot)$ 表示样本边之间的格罗莫夫-沃瑟斯坦距离函数. 基于格罗莫夫-沃瑟斯坦距离的对比损失可以最大化正子图边缘之间的相似性和最小化负子图边缘之间的相似性, 从而捕获子图之间的几何差异.

格罗莫夫-沃瑟斯坦距离用于计算每个子图内节点对之间的距离, 以测量子图之间的边缘相似性. 对于 2 个子图 $P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}}$, 格罗莫夫-沃瑟斯坦距离定义为:

$$\hat{c}\left(h_i, \hat{h}_j, h_{i'}, \hat{h}_{j'}\right) = \left\| c(h_i, h_{i'}) - c\left(\hat{h}_j, \hat{h}_{j'}\right) \right\|_2, \tag{16}$$

$$D_{\mathrm{gw}}\left(P_i^{\mathrm{o}}, \hat{P}_i^{\mathrm{o}}\right) = \min_{\boldsymbol{T}} \sum_{i,i',j,j'} T_{ij} T_{i'j'} \hat{c}\left(h_i, \hat{h}_j, h_{i'}, \hat{h}_{j'}\right), \tag{17}$$

其中 $\hat{c}(\cdot)$ 是用于衡量 2 个子图之间的边缘差异的成本函数.

最后, 本文得到了最终的损失函数:

$$L = \alpha L_{\mathrm{w}} + (1-\alpha) L_{\mathrm{gw}}, \tag{18}$$

其中 α 表示用于控制不同损失函数重要性的超参数.

3 实 验

在本节中, 本文进行了广泛的实验, 以自监督方式评估本文方法在直推式和归纳式节点分类任务上的性能. 将本文方法与其他基准方法进行了比较, 包括无监督方法和有监督方法. 此外, 本文进行了消融研究来验证本文方法的有效性.

3.1 具体实现

为了评估本文方法的有效性, 在 5 个基准数据集上进行了 3 个实验: 1)文献主题分类. 在 Cora, Citeseer, Pubmed 引文网络数据集[40]上进行了文献主题分类实验. 在这些数据集中, 图由节点和边组成, 其中节点对应于文章, 边对应于引用. 每个节点都有 1 个词袋表示和 1 个类标签. 2)蛋白质角色分类. 使用 PPI 网络[41]进行了蛋白质角色分类实验. 这个数据集包含多个图, 其中 20 个图用于训练, 2 个图用于验证, 2 个图用于测试. 每个节点具有 50 个特征, 包括定位基因集、基序基因集、免疫特征以及代表基因本体的 121 个标签. 3)社交网络社区结构预测. 进行了社交网络中红人帖子的社区结构预测实验. 在 Reddit 数据集[42]中, 节点对应于帖子, 如果同一用户在 2 个帖子上都有推荐, 则存在边. 节点特性包括帖子标题、内容、连接以及其他指标(如帖子分数和评论数量). 根据归纳设置, 本文使用前 20 天的帖子进行训练, 其余的帖子用于测试. 数据集的统计信息如表 1 所示.

Table 1 Statistics of the Datasets

表 1 数据集的统计信息

统计信息	Cora	Citeseer	Pubmed	PPI	Reddit
节点数	2 708	3 327	19 717	56 944	232 965
边数	5 429	4 732	44 338	818 716	11 606 919
特征维度	1 433	3 703	500	50	602
类别数	7	6	3	121	41
用于训练的节点数	140	120	60	44 906	153 757
用于验证的节点数	500	500	500	6 514	23 297
用于测试的节点数	1 000	1 000	1 000	5 524	55 911

由于不同数据集具有不同的属性, 本文在 3 种实验环境下采用了不同的编码器. 实验使用了 PyTorch 和 torch_geometric 库来实现, 实验在 4 个 NVIDIA RTX 3090 GPU 上进行. 本文方法以自监督的方式学习节点表示, 然后在下游任务中使用学习到的表示进行节点分类. 在对比损失中, 为了减少计算负担, 本文设置 M=2. 对于结构化损失函数, 设置超参数 $\alpha = 0.5$. 为此, 本文训练和测试了一个简单的线性分类器(例如逻辑回归). 本文模型的训练是通过最小化对比学习损失函数来进行的. 使用了 Adam 优化器, 初始学习率为 0.000 1(在 Pubmed 数据集上为 0.000 5). 节点表示的维度设置为 1 024(对于 PPI 数据集为 256, 使用 4 个注意力头). 为了公平比较, 本文对所有方法都使用了相同的 10 个随机分割, 并使用准确率作为评估指标. 对于多图的归纳式学习, 本文在每个可用的图上训练编码器, 并使用学到的编码器来生成新图的节点表示. 对于大型图的归纳学习, 本文通过采样节点的小批量来训练编码器.

为了公平比较, 本文遵循了一些先前的无监督相关工作设置, 所有结果均带有标准差. 如表 1 所示, Cora, Citeseer, Pubmed, Reddit 这 4 个数据集的节点类别数较少, 因此采用准确率, 而不是 $F1$ 分数作为评价指标. 而 PPI 数据集具有 121 个节点数, 并且由于样本量较大, 样本数量相对均衡, 为了直接考虑每个样本的预测结果, 文本选择使用 Micro-$F1$ 分数而不是 Weighted-$F1$ 分数.

数据集的不同属性需要使用不同的编码器来提取节点嵌入. 为了公平比较, 本文与文献 [8] 的方法保持一致. 节点分类任务可以根据其设置分为直推式和归纳式. 在直推式设置中, 本文可以在训练期间访问整个图以及图中所有节点的特征. 而在归纳式设置中, 本文的目标是生成训练期间未使用过的节点的表示. 这些未见过的节点可能是演化图中的新节点(例如社交网络中的新用户), 本文将这种情况称为单图归纳式学习. 它们也可以来自以前未见过的图, 这有助于在具有相似特征形式的多个图之间进行泛化, 本文将这种情况称为多图归纳式学习.

1)针对小图的直推式学习. 对于直推式学习任务(如 Cora, Citeseer, Pubmed 数据集), 本文使用图卷积网络 GCN 作为编码器. 编码器的架构如下所示:

$$GCN_i(\boldsymbol{X},\boldsymbol{A}) = \sigma\left(\hat{\boldsymbol{D}}^{\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}_i\right), \tag{19}$$

$$f(\boldsymbol{X},\boldsymbol{A}) = GCN_2(GCN_1(\boldsymbol{X},\boldsymbol{A}),\boldsymbol{A}), \tag{20}$$

其中 $\boldsymbol{X}$ 表示图中节点的特征矩阵, $\hat{\boldsymbol{A}}$ 表示具有自环的邻接矩阵 $\hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, $\boldsymbol{I}$ 是单位矩阵, $\hat{\boldsymbol{D}}$ 是关于 $\hat{\boldsymbol{A}}$ 的度矩阵, σ 是非线性激活函数(如 ReLU, PReLU), $\boldsymbol{W}_i$ 是可学习的参数矩阵.

2)针对大图的归纳式学习. 由于 Reddit 数据集规模较大, 本文不再使用依赖于固定已知邻接矩阵的图卷积方式来对图进行编码. 相反, 本文采用了 GraphSAGE[12]作为编码器, 并引入了残差连接. 其定

义为：

$$MP_i(\boldsymbol{X},\boldsymbol{A})=\sigma([\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{A}}\boldsymbol{X}\|\boldsymbol{X}]\boldsymbol{W}_i),\tag{21}$$

$$f(\boldsymbol{X},\boldsymbol{A})=MP_3(MP_2(MP_1(\boldsymbol{X},\boldsymbol{A}),\boldsymbol{A}),\boldsymbol{A}),\tag{22}$$

其中$\hat{\boldsymbol{D}}^{-1}$为均值池化传播操作，‖表示特征拼接.

与文献 [3] 不同的是，在中心节点进行邻域采样之后，本文对每个中心节点的子图进行采样. 然后，将子图中包含的所有节点视为新的中心节点，并为每个新中心节点采样一个新的邻域，以获取中心节点的表示. 具体来说，本文在第 1, 2, 3 层分别采样了 10, 10, 25 个节点.

3)针对多图的归纳式学习. 对于 PPI 数据集，本文采用了图注意力网络作为编码器，并同样引入了残差连接. 其定义为：

$$\widehat{MP}_i(\boldsymbol{X},\boldsymbol{A})=\sigma\left(\left[\boldsymbol{X}\boldsymbol{W}_i';\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_i\right]\right),\tag{23}$$

$$f(\boldsymbol{X},\boldsymbol{A})=\widehat{MP}_3\left(\widehat{MP}_2\left(\widehat{MP}_1(\boldsymbol{X},\boldsymbol{A}),\boldsymbol{A}\right),\boldsymbol{A}\right),\tag{24}$$

其中$\boldsymbol{W}_i'$和$\boldsymbol{W}_i$都是可学习的参数矩阵.

3.2 比较结果

对比结果如表 2[8,12,39-55] 所示，其他模型的结果均来源于之前的公开文献.

首先，本文进行了与现有监督方法的比较，其中包括一些经典的基于 GCN 的模型，例如 GCN, GAT, GraphSAGE 等. 在 Cora, Citeseer, Pubmed 数据集上，GCN 和 GAT 表现出了相对不错的性能，但本文的自监督模型显著优于它们，不仅在准确率方面表现更出色，而且在 Micro-*F*1 分数上也更高. 对于 PPI 数据集，GCN 和 GAT 的性能相对较低，而本文模型取得了显著的改进，这表明本文模型在处理大规模生物网络数据时具有显著的优势. 在 Reddit 数据集上，GCN 和 GraphSAGE 表现相对较差，而本文模型再次取得了卓越的结果. 与经典的半监督方法相比，本文模型在大多数情况下都表现出了明显的优势. 具体而言，本文模型明显优于 SGC 方法，并且在 Cora 数据集上取得了与 PPNP 相当的效果，在 Citeseer 数据集上稍逊于 PPNP. 值得一提的是，本文模型不仅适用于直推式的节点分类任务，还可以推广到归纳式节点分类任务上，表现出更好的鲁棒性和泛化性. 接着，将本文模型与最新的自监督图表示学习方法进行了比较. 从实验结果可以看出，本文模型在各个数据集上都表现出色，取得了最高的准确率和 Micro-*F*1 分数. 以 Cora 数据集为例，本文模型的准确率达到了 84.6%，而最好的竞争对手 GSC 的准确率为 84.5%. 这表明本文模型在自监督图表示学习领域处于领先地位. 综合来看，本文的实验结果表明，本文提出的自监督图表示学习模型在多个任务和数据集上都取得了显著的性能提升.

Table 2 Performance of Each Model Tested on All Datasets

表 2 测试的各个模型在全部数据集上的性能 %

方法类型	模型	直推式			归纳式	
		Cora	Citeseer	Pubmed	PPI	Reddit
监督方法	GCN	81.4±0.6	70.3±0.7	76.8±0.6	51.5±0.6	93.3±0.1
	GAT	83.0±0.7	72.5±0.7	79.0±0.3	97.3±0.2	
	GraphSAGE	79.2±1.5	71.2±0.5	73.1±1.4	51.3±3.2	92.1±1.1
半监督方法	SGC	81.0±0.0	71.9±0.1	78.9±0.0		
	PPNP	85.3±0.2	75.8±0.3			
无监督方法	Raw features	56.6±0.4	57.8±0.2	69.1±0.2	42.5±0.3	58.5±0.1
	DeepWalk	67.2	43.2	65.3	52.9	32.4
	GAE	71.5±0.4	65.8±0.4	72.1±0.5		
	GraphSAGE	75.2±1.5	59.4±0.9	70.1±1.4	46.5±0.7	90.8±1.1
	GRACE	81.9±0.4	71.2±0.5	80.6±0.4	69.7±0.2	94.7±0.0
	DGI	82.3±0.6	71.8±0.7	76.8±0.6	63.8±0.2	94.0±0.1
	BGRL	82.7±0.6	71.1±0.8	79.6±0.5	73.6±0.2	94.2±0.0
	GMI	83.0±0.3	73.0±0.3	79.9±0.2	65.0±0.0	95.0±0.0
	GraphCL	83.6±0.5	72.5±0.7	79.8±0.5	65.9±0.6	95.1±0.1
	Subg-Con	83.5±0.5	73.2±0.2	81.0±0.1	66.9±0.2	95.2±0.0
	InfoGCL	83.5±0.3	73.5±0.4	79.1±0.2		
	CCA-SSG	84.0±0.4	73.1±0.3	81.0±0.4	73.3±0.2	95.1±0.0
	GraphMAE	84.2±0.4	73.4±0.4	81.1±0.4	74.5±0.3	96.0±0.1
	IDGL	84.5±0.3	74.1±0.2			
	GSC	84.6±0.1	73.7±0.1	82.1±0.2	69.1±0.3	95.3±0.1
	本文模型	85.3±0.2	74.1±0.4	82.7±0.2	75.4±0.2	96.5±0.0

注：Cora，Citeseer，Pubmed，Reddit 数据集采用准确率作为评价指标，PPI 数据集采用 Micro-*F*1 分数作为评价指标. “±”前表示平均值，“±”后表示标准差.

3.3 消融实验

在本节中，本文进行了广泛的消融实验，以探究本文方法的各个组成组件对模型性能的影响，针对 3 个核心要素进行了消融研究，包括基于随机游走路径的图嵌入策略、基于随机游走路径的子图生成策略和基于最优传输的结构对比损失，如表 3 所示.

3.3.1 核心组件

首先，本文建立了一个基线模型，该模型不使用基于随机游走路径的图嵌入策略，而是直接使用编码器生成节点嵌入，通过随机采样的方式生成采样子图，并通过节点扰动构造正负样本；然后使用 InfoNCE 对比损失函数来优化模型. 基线模型在 Cora, Pubmed, PPI 数据集上的性能分别为 82.3%, 81.2% 的准确率和 72.9% 的 Micro-*F*1 得分.

Table 3 Ablation Study Results on Core Components

表 3 关于核心组件的消融实验结果 %

核心组件			数据集		
图嵌入	子图生成	结构对比损失	Cora	Pubmed	PPI
			82.3	81.2	72.9
√			83.0	81.7	73.4
√	√		84.5	82.1	74.2
√	√	√	85.3	82.7	75.4

注：Cora，Pubmed 数据集采用准确率作为评价指标，PPI 数据集采用 Micro-*F*1 分数作为评价指标.

接着，逐步引入核心要素，评估它们对模型性能的影响. 首先，添加了基于随机游走路径的图嵌入策略，这进一步提高了模型在所有 3 个数据集上的性能，特别是在 Cora 上，准确率分别提高到了 83.0 个百分点，在 PPI 上，Micro-*F*1 提高 73.4 个百分点. 然后，引入了基于随机游走路径的子图生成策略，这进一步提高了性能，使模型在 Cora, Pubmed, PPI 上的性能分别增长了 1.5 个百分点，1.0 个百分点，0.8 个百分点. 这表明子图生成策略的引入在提高模型性能方面起到了关键作用. 最后，引入了基于最优传输的结构对比损失，进一步提高了模型性能，使其在 Cora, Pubmed 上的准确率分别达到了 85.3%, 82.7%，在 PPI 上的 Micro-*F*1 达到了 75.4%. 这表明结构对比损失对于模型的性能提升具有重要意义.

综上所述，消融实验结果清晰地表明了本文方法中各个核心要素的重要性. 基于随机游走路径的图嵌入策略、基于随机游走路径的子图生成策略和基于最优传输的结构对比损失相互协同工作，使得本文方法在节点分类任务上取得了卓越的性能. 这些结果不仅证明了本文方法的有效性，还为本文方法在实际应用的潜力提供了强有力的支持.

3.3.2 图嵌入策略

为了评估基于随机游走路径的图嵌入策略(详见 2.2 节)中基于注意力的路径间聚合策略和同/异质聚合策略对模型性能的影响，进行了一系列实验，如表 4 所示，其中设置了一个图嵌入的基线模型，该模型使用可学习的线性层来替代聚合策略，并且保持与组合策略得到的特征维度相同. 以下是实验的主要观察和分析：1)路径间聚合与同/异质聚合的比较. 在所有数据集上观察到，采用路径间聚合策略的模型性能略高于采用同/异质聚合策略的模型性能. 这表明，路径间聚合策略能够更好地捕捉节点之间的局部关系，从而提高了嵌入质量. 2)聚合策略与线性层的比较. 与线性层相比，采用聚合策略在所有数据集上都表现出更好的性能. 这表明，引入基于注意力的聚合策略能够有效地提高图嵌入模型的表达能力，从而增强了对图结构的建模能力. 总的来说，实验结果强调了基于注意力的策略内和策略间组合器在图嵌入任务中的重要性. 这些聚合策略不仅提高了模型性能，还有望在其他图分析任务中产生显著的改进.

Table 4 Ablation Study Results on Graph Embedding Strategies

表 4 图嵌入策略的消融实验结果 %

策略	Cora	Pubmed	PPI
线性层+线性层	84.6	82.1	74.8
路径间聚合+线性层	84.9	82.3	75.2
线性层+同/异质聚合	85.0	82.5	75.2
路径间聚合+同/异质聚合	85.3	82.7	75.4

注：Cora，Pubmed 数据集采用准确率作为评价指标，PPI 数据集采用 Micro-*F*1 分数作为评价指标.

3.3.3 子图采样与正/负样本生成策略

在实验中，本文比较了不同的子图抽样方法，具体是重要性评分和随机漫步，并将实验结果列于表 5 中. 为了保证公平比较，本文对相同大小的子图进行了采样. 以下是从实验结果中得出的主要观察和分析：1)基于多跳路径采样的子图采样方法效果最佳. 与其他采样方法相比，基于多跳路径采样的子图采样方法在所有数据集上都取得了最佳性能. 这证实了多跳路径采样子图能够更全面地捕获局部信息，有助于更好地区分不同的子图. 造成这一优势的部分原因是，多跳路径采样能够更好地保留图的结构特征，从而增强了图对比学习的性能. 2)基于聚合的子图生成方法效果显著. 本文还比较了基于随机扰动和基于聚合的子图生成方法. 实验结果表明，基于聚合的方法能够有效地捕获图的固有局部结构，从而提高了图对比学习的性能. 这一结果强调了对于子图生成、图的结构和拓扑特性的考虑是非常重要的.

Table 5 Ablation Study Results on Subgraph Sampling and Positive/Negative Sample Generation

表 5 子图采样与正/负样本生成的消融实验结果 %

策略	Cora	Pubmed	PPI
重要性评分+随机扰动	85.4	81.3	74.8
随机游走+随机扰动	84.8	82.0	74.8
多跳路径采样+随机扰动	85.2	82.5	75.3
重要性评分+聚合生成	84.9	81.5	75.0
随机游走+聚合生成	85.1	82.2	75.3
多跳路径采样+聚合生成	85.3	82.7	75.4

注：Cora，Pubmed 数据集采用准确率作为评价指标，PPI 数据集采用 Micro-*F*1 分数作为评价指标.

4 结论与展望

本文将图拓扑不平衡问题引入了自监督图表示学习领域，重点研究了未标记节点的不平衡学习问题. 提出了一个自监督图拓扑不平衡学习框架，该框架通过引入多跳路径的子图邻域概念，充分挖掘了子图的局部结构信息，捕获了同质和异质的邻域信息. 这个框架充分地利用了子图结构信息，采用基于聚合多条路径的策略来采样子图和生成正/负样本，并构造结构对比损失来计算子图之间的相似度，进而有效地捕获图的内在局部结构，从而很好地表示了图之间的差异. 通过在多个基准数据集上进行广泛的实验，本文证明了所提出的自监督图拓扑不平衡学习框架相对于有监督和无监督的图表示学习方法来说具有更好的性能.

作者贡献声明：秦者云提出了算法思路并进行了实验；卢宪凯和裘肖明改进了方案并修改论文；任春晓和聂秀山改进方案；尹义龙提出指导意见并修改论文.

参 考 文 献

[1] Kang Zhao, Liu Liang, Han Meng. Semi-supervised classification based on transformed learning[J]. Journal of Computer Research and Development, 2023, 60(1): 103−111 (in Chinese)
(康昭，刘亮，韩蒙. 基于转换学习的半监督分类[J]. 计算机研究与发展，2023，60(1)：103−111)

[2] Hamilton W L, Ying R, Leskovec J. Representation learning on graphs: Methods and applications[J]. arXiv preprint, arXiv: 1709.05584, 2017

[3] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint, arXiv: 1609.02907, 2016

[4] Chen Deli, Lin Yankai, Zhao Guangxiang, et al. Topology-imbalance learning for semi-supervised node classification[C]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 29885−29897

[5] Sun Qingyun, Li Jianxin, Yuan Haonan, et al. Position-aware structure learning for graph topology-imbalance by relieving under-reaching and over-squashing[C]//Proc of the 31st ACM Int Conf on Information & Knowledge Management. New York: ACM, 2022: 1848−1857

[6] Fu Xingcheng, Wei Yuecen, Sun Qingyun, et al. Hyperbolic geometric graph representation learning for hierarchy-imbalance node classification[C]//Proc of the ACM Web Conf. New York: ACM, 2023: 460−468

[7] Jin Di, Wang Rui, Ge Meng, et al. RAW-GNN: Random walk aggregation based graph neural network[J]. arXiv preprint, arXiv: 2206.13953, 2022

[8] Han Yuehui, Hui Le, Jiang Haobo, et al. Generative subgraph contrast for self-supervised graph representation learning[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 91−107

[9] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[C/OL]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2016[2024-09-29]. https://proceedings.neurips.cc/paper_files/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html

[10] Veličković P, Cucurull G, Casanova, et al. Graph attention networks[J] arXiv preprint, arXiv: 1710.10903, 2017

[11] Yun S, Jeong M, Kim R, et al. Graph transformer networks[C/OL]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2019[2024-09-29]. https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html

[12] Hamilton W, Ying Zhitao, Leskovec J. Inductive representation learning on large graphs[C/OL]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2017[2024-09-29]. https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033-da9c6fb5ba83c7a7ebea9-Abstract.html

[13] Fu Xinyu, Zhang Jiani, Meng Ziqiao, et al. Ma-GNN: Metapath aggregated graph neural network for heterogeneous graph embedding[C]//Proc of the Web Conf. New York: ACM, 2020: 2331−2341

[14] Xie Xiaojie, Liang Ying, Wang Zisen, et al. Heterogeneous network node classification method based on graph convolution[J]. Journal of Computer Research and Development, 2022, 59(7): 1470−1485 (in Chinese)
(谢小杰，梁英，王梓森，等. 基于图卷积的异质网络节点分类方法[J]. 计算机研究与发展，2022，59(7)：1470−1485)

[15] Dong Yuxiao, Chawla N V, Swami A. Metapath2vec: Scalable representation learning for heterogeneous networks[C]//Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 135−144

[16] Fu Taoyang, Lee W C, Lei Zhen. HIN2vec: Explore meta-paths in heterogeneous information networks for representation learning [C]//Proc of the ACM on Conf on Information and Knowledge Management. New York: ACM, 2017: 1797−1806

[17] Shi Chuan, Hu Binbin, Zhao W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357−370

[18] Wang Xiao, Ji Houye, Shi Chuan, et al. Heterogeneous graph attention network[C]//Proc of the World Wide Web Conf. New York: ACM, 2019: 2022−2032

[19] Wang Yiwei, Wang Wei, Liang Yuxuan, et al. Mixup for node and graph classification[C]//Proc of the Web Conf. New York: ACM, 2021: 3663−3674

[20] Feng Wenzheng, Zhang Jie, Dong Yuxiao, et al. Graph random neural networks for semi-supervised learning on graphs[C]//Proc of the

Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 22092−22103

[21] You Yuning, Chen Tianlong, Sui Yongduo, et al. Graph contrastive learning with augmentations[C]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 5812−5823

[22] Zhu Yun, Guo Jianhao, Wu Fei, et al. Rosa: A robust self-aligned framework for node-node graph contrastive learning[J]. arXiv preprint, arXiv: 2204.13846, 2022

[23] Gasteiger J, Weißenberger S, Günnemann S. Diffusion improves graph learning[C/OL]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2019[2024-09-29]. https://proceedings.neurips.cc/paper/2019/hash/23c894276a2c5a16470e6a31f4618d73-Abstract.html

[24] Zhao Tong, Liu Yozen, Neves L, et al. Data augmentation for graph neural networks[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 11015−11023

[25] He Haibo, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263−1284

[26] Guo Haixiang, Li Yijing, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220−239

[27] Xu Xiaolong, Chen Wen, Sun Yanfei. Over-sampling algorithm for imbalanced data classification[J]. Journal of Systems Engineering and Electronics, 2019, 30(6): 1182−1191

[28] Park J, Song J, Yang E. Graphens: Neighbor-aware ego network synthesis for class-imbalanced node classification[C/OL]//Proc of the 10th Int Conf on Learning Representations. Washington: ICLR, 2021[2024-09-29]. https://openreview.net/forum?id=MXEl7i-iru

[29] Cao Kaidi, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss[C/OL]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2019[2024-09-29]. https://proceedings.neurips.cc/paper/2019/hash/6214-61af90cadfdaf0e8d4cc25129f91-Abstract.html

[30] Zhu Yanqiao, Xu Weizhi, Zhang Jinghao, et al. Deep graph structure learning for robust representations: A survey[J]. arXiv preprint, arXiv: 2103.03036, 2021

[31] Wang Hongwei, Leskovec J. Combining graph convolutional neural networks and label propagation[J]. ACM Transactions on Information Systems, 2021, 40(4): 1−27

[32] Topping J, Di Giovanni F, Chamberlain B P, et al. Understanding over-squashing and bottlenecks on graphs via curvature[J]. arXiv preprint, arXiv: 2111.14522, 2021

[33] Zhao Jianan, Wang Xiao, Shi Chuan, et al. Heterogeneous graph structure learning for graph neural networks[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 4697−4705

[34] Zhao Jianan, Wang Xiao, Shi Chuan, et al. Robust graph representation learning via neural sparsification[C]//Proc of the 37th Int Conf on Machine Learning. New York: ICML, 2020: 11458−11468

[35] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks[C]//Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855−864

[36] Dey R, Salem F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//Proc of the 60th IEEE Int Midwest Symp on Circuits and Systems. Piscataway, NJ: IEEE, 2017: 1597−1600

[37] Yin Yihang, Wang Qingzhong, Huang Siyu, et al. Autogcl: Automated graph contrastive learning via learnable view generators[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 8892−8900

[38] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport[C/OL]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2013[2024-09-29]. https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html

[39] Chizat L, Roussillon P, Léger F, et al. Faster Wasserstein distance estimation with the Sinkhorn divergence[C]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 2257−2269

[40] Sen P, Namata G, Bilgic M, et al. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93−93

[41] Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks[J]. Bioinformatics, 2017, 33(14): i190−i198

[42] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701−710

[43] Kipf T N, Welling M. Variational graph auto-encoders[J]. arXiv preprint, arXiv: 1611.07308, 2016

[44] Zhu Yanqiao, Xu Yichen, Yu Feng, et al. Deep graph contrastive representation learning[J]. arXiv preprint, arXiv: 2006.04131, 2020

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

[46] Thakoor S, Tallec C, Azar M G, et al. Large-scale representation learning on graphs via bootstrapping[J]. arXiv preprint, arXiv: 2102.06514, 2021

[47] Peng Zhen, Huang Wenbing, Luo Minnan, et al. Graph representation learning via graphical mutual information maximization[C]//Proc of the Web Conf. New York: ACM, 2020: 259−270

[48] Hafidi H, Ghogho M, Ciblat P, et al. Graphcl: Contrastive self-supervised learning of graph representations[J]. arXiv preprint, arXiv: 2007.08025, 2020

[49] Jiao Yizhu, Xiong Yun, Zhang Jiawei, et al. Sub-graph contrast for scalable self-supervised graph representation learning[C]//Proc of the 20th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2020: 222−231

[50] Xu Dongkuan, Wei Cheng, Luo Dongsheng, et al. Infogcl: Information-aware graph contrastive learning[C]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 30414−30425

[51] Zhang Hengrui, Wu Qitian, Yan Junchi, et al. From canonical correlation analysis to self-supervised graph neural networks[C]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 76−89

[52] Hou Zhenyu, Liu Xiao, Cen Yukuo, et al. Graphmae: Self-supervised masked graph autoencoders[C]//Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 594−604

[53] Yu Chen, Wu Lingfei, Zaki M. Iterative deep graph learning for graph neural networks: Better and robust node embeddings[C]//Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 19314−19326

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

[55] Gasteiger J, Bojchevski A, Günnemann S. Predict then propagate: Graph neural networks meet personalized pagerank[J]. arXiv preprint, arXiv: 1810.05997, 2018



Qin Zheyun, born in 1997. PhD candidate. His main research interest includes machine learning.

秦者云, 1997 年生. 博士研究生. 主要研究方向为机器学习.



Lu Xiankai, born in 1990. PhD, professor, master supervisor. His main research interests include machine learning and computer version.

卢宪凯, 1990 年生. 博士, 研究员, 硕士生导师. 主要研究方向为机器学习、计算机视觉.



Xi Xiaoming, born in 1987. PhD, professor, master supervisor. His main research interests include pattern recognition, computer vision, and medical image processing.

袭肖明, 1987 年生. 博士, 教授, 硕士生导师. 主要研究方向为模式识别、计算机视觉、医学图像处理.



Ren Chunxiao, born in 1978. PhD, professor. His main research interests include pattern recognition, and machine learning.

任春晓, 1978 年生. 博士, 研究员. 主要研究方向为模式识别、机器学习.



Nie Xiushan, born in 1981. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include artificial intelligence and intelligent media analysis, and computer application technology.

聂秀山, 1981 年生. 博士, 教授, 博士生导师, CCF 高级会员. 主要研究方向为人工智能与智能媒体分析、计算机应用技术.



Yin Yilong, born in 1972. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interest includes machine learning.

尹义龙, 1972 年生. 博士, 教授, 博士生导师. CCF 杰出会员. 主要研究方向为机器学习.