面向半监督归纳式学习的自训练增强图模型

杨瀚轩[1),2)] 余昭昕[2),1)] 李子乾[3)] 徐会芳[4)] 孔庆超[2),1)]

[1)](中国科学院大学人工智能学院 北京 100049)

[2)](中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190)

[3)](国家电网有限公司客户服务中心 天津 300304)

[4)](中国电力科学研究院有限公司 北京 100192)

摘 要 图表示学习是图数据分析的一个基础研究问题,在多种应用领域中均具有重要的研究价值。不同于一般的直推式学习,归纳式图表示学习要求对训练过程中不可见的未知节点进行推理和分类,因此具有更大的研究挑战。现有归纳式学习方法主要采用建立在全监督学习下的图神经网络,这些方法依赖于大量带标注的数据进行训练,因而在面对可见结构中节点标注稀缺的半监督归纳式学习问题时可能存在模型过拟合问题。本文首次提出半监督归纳式图表示学习问题,并建立了一种自训练增强的归纳式图(Self-Training Augmented Inductive Graph,STAIG)模型,该模型由一个使用图神经网络学习节点向量表示的编码器和一个通过重构节点标签和属性特征训练模型的解码器组成。针对半监督归纳式图学习问题,所提出的模型采用自训练增强方法,并在编码器中提出一种基于随机游走的节点掩码方法提高预测未知节点的泛化性。在此基础上,为了进一步应对标注稀缺问题,该模型使用解码器生成节点伪标签来增强标注信息,并通过置信度过滤机制提高伪标签的可靠性。基于基准归纳式学习图数据集的实验验证了本文提出的STAIG模型在半监督节点分类任务上取得了优于对比方法的结果,且在标注数据比例低于10%的弱监督学习设置下具有显著优势。

关键词 归纳式图表示学习;半监督节点分类;变分图自编码;自训练增强

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.02263

Self-Training Augmented Graph Model for Semi-Supervised Inductive Learning

YANG Han-Xuan[1),2)] YU Zhao-Xin[2),1)] LI Zi-Qian[3)] XU Hui-Fang[4)] KONG Qing-Chao[2),1)]

[1)](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049)

[2)](State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

[3)](State Grid Corporation of China Customer Service Center, Tianjin 300304)

[4)](China Electric Power Research Institute, Beijing 100192)

Abstract Graph representation learning is a fundamental research issue in graph data analysis with significant research value across various application domains. Unlike traditional transductive learning, inductive graph representation learning requires inferring and classifying unknown nodes that are invisible during the training procedure, making it a more challenging research issue. Existing inductive learning methods primarily use graph neural networks (GNNs) under fully supervised learning. These methods rely on large amounts of annotated data for training, and thus

收稿日期:2024-12-26;在线发布日期:2025-07-11。本课题得到国家电网有限公司电力大型语言模型关键技术研究及在客服中的示范应用科技项目(No.5700-202353595A-3-2-ZN)资助。**杨瀚轩**,博士,主要研究领域为图表示学习、变分自编码。E-mail:yanghanxuan2020@ia.ac.cn。**余昭昕**,博士,主要研究领域为深度学习、情感认知建模。**李子乾**,硕士,高级工程师,主要研究领域为人工智能技术研究和应用。**徐会芳**,硕士,高级工程师,主要研究领域为电力领域知识图谱、图机器学习。**孔庆超**(通信作者),博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为社会计算、大语言模型。E-mail:qingchao.kong@ia.ac.cn。

may experience model overfitting when addressing semi-supervised inductive learning problems with sparse node labels in the visible structure. This paper first proposes the semi-supervised inductive graph representation learning problem and establishes the Self-Training Augmented Inductive Graph (STAIG) model. This model consists of an encoder that learns node vector representations using a graph neural network and a decoder that trains the model by reconstructing node labels and attribute features. To address the semi-supervised inductive graph learning problem, the proposed model employs the self-training augmentation technique, and proposes a node masking method based on random walks in the encoder to improve the generalizability of predicting unknown nodes. Furthermore, to address the issue of label scarcity, the model employs the decoder to generate pseudo-node labels to enhance label information, with a confidence filtering mechanism to improve the reliability of the pseudo-labels. Experiments based on benchmark inductive learning graph datasets validate that the proposed STAIG model achieves superior results for semi-supervised node classification compared to the comparative methods and demonstrates significant advantages under weakly supervised learning settings with less than 10% labeled data.

Keywords inductive graph representation learning; semi-supervised node classification; variational graph auto-encoder; self-training augmentation

1 引 言

图表示学习是图数据分析的一个基础研究问题，旨在将图数据的拓扑结构以及节点的属性特征和类别标签等信息映射为低维向量表示（或嵌入），用于处理节点分类等图数据下游任务。一般来说，根据任务设置的不同，图表示学习问题可以分为直推式学习和归纳式学习。传统的直推式学习问题假设所学习的图拓扑结构完全已知且固定不变，所有节点的结构信息和属性特征在训练过程中均可见。与之不同，归纳式学习问题假设部分图结构未知，在模型训练过程中，只有部分已知节点的结构信息可见，而所有待推理（分类）的未知节点，其结构信息和属性特征均不可见。这类图表示学习问题在许多现实世界的图数据中都具有十分重要的研究和应用价值，例如演化网络[1-3]（evolving network）和跨图网络[4-5]（cross-graph network）等。

解决归纳式学习问题要求模型对结构信息不完全或可变的图数据进行处理，因此需要具有更强的未知结构推理和泛化能力[6]。现有归纳式图表示学习方法主要基于图神经网络（Graph Neural Network，GNN）[6-8]，这些方法基于图拓扑结构对节点邻域信息进行聚合来得到包含节点的多步邻域信息的聚合节点表示，从而通过挖掘图数据的局部邻域信息来提升模型对未知节点的推理能力。然而，现有归纳式学习方法均采用全监督学习设置，即要求可见图结构中的所有节点都具有标注信息以用于训练模型。该问题严重限制了这些方法的实际应用价值，因为在许多现实场景中，由于数据丢失或标注成本过高等原因，经常存在大量无标注节点，仅有少部分节点具有标注信息。现有方法难以基于已知节点的结构和少量标注信息训练模型，因而无法对未知节点类别进行有效推理。

在标注数据稀缺的场景下，本文提出半监督归纳式图表示学习问题。如表1所示，与假设训练数据的拓扑结构信息不完全但标注信息完全的传统全监督归纳式学习问题不同，半监督归纳式图表示学习要求训练数据的拓扑结构信息和节点标注信息均不完全。现有方法处理半监督归纳式图学习问题主要有以下技术挑战：(1)由于部分图结构不可见导致拓扑结构信息不完全，现有半监督节点分类方法[7-17]无法获取待推理的未知节点结构信息，因此需要基于可见图结构的分布信息训练模型，并采用随机掩码等方式提升模型处理包含部分未知节点的不完整图结构的泛化能力。(2)由于训练数据的可见节点标注稀缺导致标注信息不完全，现有归纳式图学习方法[6,18-22]可能对少量的有标注节点发生过拟合，导致对未知节点的类别推理能力下降，因此需要更加充分地挖掘和利用可见图结构中无标注数据的结构信息，并通过生成伪标签等方式增强节点的标注信息。

表1　图表示学习任务设置比较

	结构信息	标注信息
半监督学习	完全	不完全
归纳式学习	不完全	完全
半监督归纳式学习	不完全	不完全

针对半监督归纳式图表示学习问题，本文进一步提出自训练增强的归纳式图(Self-Training Augmented Inductive Graph，STAIG)模型，包括一个图卷积网络(graph convolutional network，GCN)编码器和一个标签重构解码器。该模型采用一种自训练增强技术，为了适应归纳式学习问题中部分节点结构不可见的问题，在编码器中采用随机游走方法对部分节点进行掩码，从而适应处理包含未知节点的可变图结构，改进模型推理未知节点的泛化性。在此基础上，为了缓解半监督学习设置下的标注数据稀缺问题，该模型使用解码器生成节点伪标签来增强标注信息，并采用一种置信度过滤机制对所生成伪标签进行筛选，最后通过重构节点标签及属性特征训练模型。基于基准归纳式图数据集的实验验证了所提出模型与当前最佳节点分类方法对比的有效性。

本文的主要研究贡献总结如下：(1)本文首次提出半监督归纳式图表示学习问题，并基于编码器-解码器框架建立了一种自训练增强的归纳式图模型STAIG，通过重构节点标签和属性特征端到端地解决半监督节点分类任务；(2)为了适应归纳式学习场景下部分结构不可见的问题，该模型采用一种基于随机游走的节点掩码方法，提升处理包含未知节点的可变图结构的适应性和泛化性；(3)针对半监督学习场景下的标注数据稀缺问题，该模型通过迭代生成节点伪标签来增强数据的标注信息，并采用一种置信度过滤机制提高伪标签的可靠性；(4)基于归纳式图数据集的实验结果表明，本文提出的STAIG模型在节点分类任务上均取得了良好的性能，且在弱监督学习场景下较对比方法具有显著优势。

本文的组织结构如下：第2节介绍相关工作；第3节给出半监督归纳式图表示学习的问题定义；第4节详细介绍本文提出的STAIG模型方法；第5节对所提出方法的有效性进行实验验证和结果分析；第6节对本文内容进行总结。

2　相关工作

本节简要介绍与半监督节点分类和归纳式图表示学习相关的代表性研究工作。

2.1　半监督节点分类

半监督节点分类任务旨在利用拓扑结构和稀缺标签信息预测图节点的类别。代表性方法主要采用GNN框架，这些方法基于图拓扑结构对节点邻域信息进行聚合来学习节点的向量表示，并通过构建多层网络结构多次聚合邻居节点特征，得到包含节点的多步邻域信息的聚合节点表示。Kipf等[7]首先提出采用GNN方法解决半监督节点分类问题并建立GCN方法，该方法使用归一化的图拉普拉斯矩阵对邻居节点进行均值池化，从而聚合节点的邻域信息。图注意力网络(graph attention network，GAT)[8]采用注意力机制来执行邻域聚合，使用一个子注意力网络来计算每个邻居节点的注意力权重，并利用注意力机制参数化每个邻居节点的权重。Zhu等[9]针对节点类别分布不均衡问题提出BNE(balanced neighbor exploration)方法，通过一种邻居探索算法来依据节点间的拓扑关系构建数据类别分布均衡的训练集。Liu等[10]针对结构、特征和标注不完全的弱信息场景提出D^{2}PT(dual-channel diffused propagation then transformation)方法，通过使用一种基于图扩散机制的对偶模型框架来缓解数据不足和孤立节点问题。此外，也有工作针对异质图[11]以及超图结构[12,23]等特殊任务场景展开研究。与这些工作不同，本文主要聚焦一般的同质图半监督节点分类问题。

针对节点分类任务，Zhang等[13]提出图注意力多层感知器(graph attention multi-layer perceptron，GAMLP)，该模型采用一种标签传播方法，即将节点标签转化为独热编码，并将其与节点属性特征相结合构成模型的输入特征，再使用GNN对节点的标签和属性特征进行邻域聚合，得到包含标签信息的GNN节点表示。为了应对半监督学习下节点标注稀缺的问题，有工作进一步提出使用一个预训练教师模型来生成节点的伪标签，用以增强数据的标签信息[14-15]，这些方法需要训练一个额外的教师模型，因而训练过程较为复杂，模型存储成本较高。还有工作基于Transformer预训练模型提出无监督图表示学习方法，通过采用节点随机掩码和非对称编码器-解码器结构降低模型的内存消耗[24]。最近的工作中，Yang等[16]提出MGCN(mixed graph contrastive network)，该方法基于线性插值得到多视图节点插值表示和标签，并通过最小化不同视图之间同一节点的插值表示来缓解表示坍塌问题。Peng等[17]基

于图对比学习框架提出LGGCL(label-guided graph contrastive learning)方法,该方法利用节点伪标签划分对比学习正负样本,并分别基于属性特征聚类对齐对正样本进行筛选,基于伪标签概率分布对负样本进行重加权。此外,还有一些工作采用变分图自编码(variational graph auto-encoder,VGAE)[25]框架来无监督地学习节点表示,然后使用这些节点表示训练一个线性分类器用于节点分类。例如GraphMAE[26]对节点的属性特征进行随机掩码作为模型的输入特征,然后通过重构原始属性特征来训练模型。MaskGAE[27]对图中的某些边或路进行随机掩码,以减轻模型对图数据的邻近结构特征过拟合的问题。然而,上述工作均主要研究直推式图表示学习问题,即要求训练图数据与待推理的测试图数据的结构特征完全一致,但对于更具有挑战性和应用价值的归纳式图学习问题仍缺少半监督节点分类任务相关研究。

2.2 归纳式图表示学习

归纳式图表示学习问题要求对在模型训练过程中不可见的未知节点进行分类预测,因此需要模型具有更强的结构泛化能力。相关工作中,Hamilton等[6]首先提出归纳式图表示学习问题,即要求用于训练的图结构与待推理的图结构不一致。对于节点分类任务,归纳式学习问题需要模型基于训练数据中的已知节点进行优化,并对未知节点的向量表示进行泛化推理。针对此问题,Hamilton等提出GraphSAGE方法,该方法采用一种基于图拓扑结构的节点采样技术,对每个节点随机选取固定数量的邻居节点样本进行邻域聚合,并建立多层网络结构聚合多步邻域信息,通过学习节点的邻域复合表示提升模型推理未知节点结构的泛化性。此后,一些工作针对归纳式学习问题展开进一步研究,例如Xu等[18]针对动态图数据的归纳式学习问题,使用自注意力机制来学习节点的动态邻域信息,并使用函数化的时间表示向量作为位置编码。Zeng等[19]提出GraphSAINT,通过对子图而非节点或边进行采样进一步提升归纳式学习的效率和准确性。Wang等[20]提出使用基于因果的随机游走来学习动态图数据的三角闭包结构特征,从而提升模型处理动态图的归纳式节点分类任务性能。

在最近的研究工作中,Gao等[21]提出一种归纳式图压缩方法,该方法通过节点多对一映射的方式将一个大规模图压缩为一个节点数较少的模拟图,然后对模拟图进行归纳式推理,以降低处理大规模图数据的计算和存储负担。Yao等[22]针对动态演化图数据归纳式学习问题,提出将GraphSAGE方法引入循环神经网络框架,递归地学习和推理动态图数据的结构表示。然而,上述归纳式图表示学习方法均建立在可见图结构中所有节点均具有标注信息的全监督学习条件下,而没有考虑仅有少量已知节点具有标注的半监督归纳式学习问题,因而在处理半监督学习问题时,现有方法可能由于训练数据的标注信息稀缺而发生过拟合。基于此,本文首次提出半监督归纳图表示学习问题,并建立一种生成式图模型,采用节点随机掩码方法生成伪标签来增强数据的标注信息。

3 问题定义

本节给出半监督归纳式图表示学习问题的正式定义。

给定一个包含N个节点的图结构$\mathcal{G}$,$\mathcal{V}$表示所有节点的集合,$\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)'\in\mathbb{R}^{N\times D_0}$表示节点属性特征,$\boldsymbol{Y}=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N)'\in(0,1)^{N\times D}$表示节点标签,其中$D_0$和$D$分别为输入特征维度和标签类别数。假设图$\mathcal{G}$中有一部分节点在训练和推理过程中均可见,称为可见节点,其余节点仅在推理过程中可见,称为未知节点。分别使用$\mathcal{V}_{\text{vis}}$和$\mathcal{V}_{\text{inv}}$表示可见节点和未知节点集合,且$\mathcal{V}_{\text{vis}}\cup\mathcal{V}_{\text{inv}}=\mathcal{V}$,$\mathcal{G}_{\text{vis}}$表示仅包含所有可见节点的子图结构,则归纳式图表示学习的目标是基于$\mathcal{G}_{\text{vis}}$以及可见节点的属性特征$\boldsymbol{x}_i$和标签$\boldsymbol{y}_i(i\in\mathcal{V}_{\text{vis}})$训练一个模型$\mathcal{M}$,再使用$\mathcal{M}$来推理未知节点的标签$\boldsymbol{y}_j(j\in\mathcal{V}_{\text{vis}})$。

上述归纳式图学习问题假设可见节点中所有节点均有标注。在此基础上,半监督归纳式图学习要求在可见节点$\mathcal{V}_{\text{vis}}$中,只有部分节点有标注。记$\mathcal{V}_{\text{lab}}\subseteq\mathcal{V}_{\text{vis}}$为有标注节点集,则半监督归纳式图表示学习的目标是基于$\mathcal{G}_{\text{vis}}$、可见节点属性特征$\boldsymbol{x}_i(i\in\mathcal{V}_{\text{vis}})$以及节点标签$\boldsymbol{y}_{i'}(i'\in\mathcal{V}_{\text{lab}})$训练一个模型$\mathcal{M}$,再使用$\mathcal{M}$来推理未知节点的标签$\boldsymbol{y}_j(j\in\mathcal{V}_{\text{vis}})$。

4 自训练增强的归纳式图模型

本节对所提出的自训练增强的归纳式图(STAIG)模型展开详细介绍。该模型采用生成式变分自编码[28](variational auto-encoder,VAE)框架,由一个使用GCN算子来学习节点表示的编码器

和一个通过重构节点标签和节点属性特征来训练模型的解码器组成。此外，为了应对半监督学习下的标注稀缺问题，本节提出一种自训练增强方法，通过使用模型迭代生成节点伪标签来增强训练数据的标注信息。

STAIG模型的整体框架如图1所示。模型训练过程分为两个阶段，对于第t次迭代，当$t\leqslant T_0$时训练处于预热(warm-up)阶段，此阶段仅使用节点的真实标签$\boldsymbol{Y}$来训练模型，当$t>T_0$时训练进入自训练阶段，此阶段使用由上一轮迭代后得到的模型$\mathcal{M}^{(t-1)}$生成的伪标签$\tilde{\boldsymbol{Y}}$来增强真实标签。然后，GCN编码器将$\boldsymbol{Y}$或$\tilde{\boldsymbol{Y}}$与节点属性特征$\boldsymbol{X}$拼接作为输入特征，使用GCN层学习变分正态后验的均值μ和标准差σ，并通过采样得到节点表示$\boldsymbol{Z}$。最后，标签重构解码器使用$\boldsymbol{Z}$通过前馈神经网络(feedforward neural network，FNN)重构节点标签$\hat{\boldsymbol{Y}}$和属性特征$\hat{\boldsymbol{X}}$，并分别计算真实标签重构损失$\mathcal{L}_{\text{lab}}$、伪标签重构损失$\mathcal{L}_{\text{pseu}}$、属性特征重构损失$\mathcal{L}_{\text{feat}}$和一个KL(Kullback-Leibler)散度以优化模型。

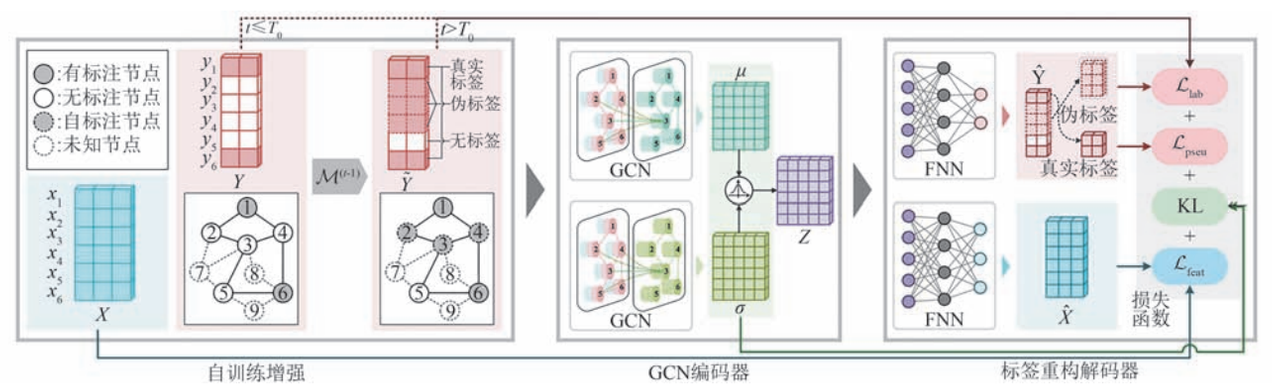

图1　STAIG模型框架

4.1　自训练增强方法

为了增强半监督归纳式图表示学习中的数据标注，本模型基于自训练方法，使用上轮迭代训练得到的模型来生成节点伪标签，这些伪标签经过置信过滤后用于增强训练数据的标注信息。

本自训练增强方法整体流程如图2所示。该方法以包含无标注节点的原始图邻接矩阵为输入，采用随机游走对图节点进行K轮掩码处理(图中$K=3$)，进而将所得到的多个掩码图邻接矩阵输入模型$\mathcal{M}^{(t-1)}$来生成节点标签$\hat{\boldsymbol{Y}}_k^{(t)}$，$k=1,\cdots,K$，并对基于不同掩码图生成的节点标签取均值得到伪标签$\bar{\boldsymbol{Y}}^{(t)}$，最后基于一种置信度过滤机制对该伪标签进行筛选，同时结合有标注节点的真实标签得到增强标签$\tilde{\boldsymbol{Y}}^{(t)}$。

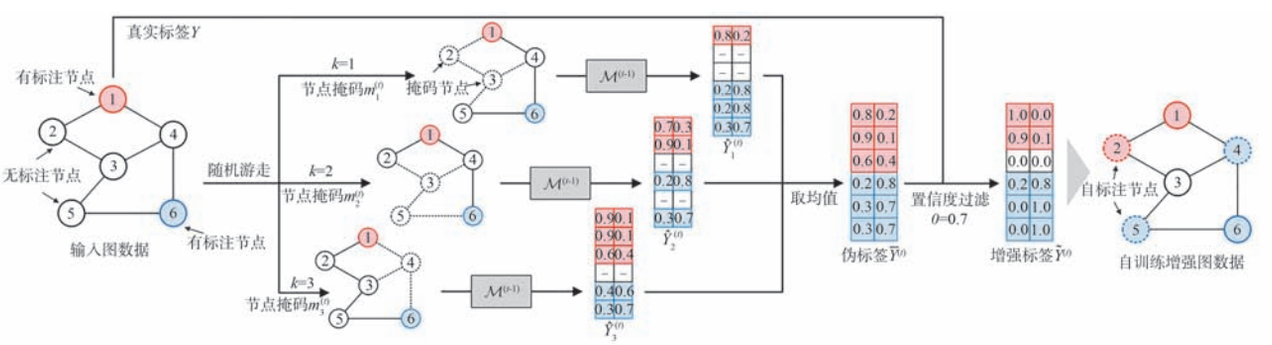

图2　自训练增强方法示意图

4.1.1　基于节点掩码的伪标签增强

在自训练阶段，本方法首先使用模型对所有无标注可见节点的类别标签重复进行多次预测，然后取多次预测结果的均值作为无标注节点的伪标签。在每次预测中，采用一种节点随机掩码方法将输入图中的部分无标注节点掩码，使得输入图中包含部分结构不可见的未知节点。

针对图掩码问题，现有基于VAE的图表示学习

方法已有一定研究，例如对节点输入特征掩码[25,29-31]或对边掩码[26,32]等。与这些工作相比，本方法主要关注解决归纳式学习问题，因此选择将每个节点的拓扑结构和属性特征完全掩码（而非仅掩码属性特征），从而将掩码节点作为训练数据中的未知节点，以提升模型的未知节点推理和泛化能力。

具体来说，本方法基于随机游走对节点进行完全掩码。图随机游走是指从一个随机根节点出发，每次选择一个当前节点的邻居节点作为下一个节点，并在一定步数内不断重复此过程。考虑到绝大多数现实世界中的归纳式学习图数据集都是具有动态演化特征，即已知节点与未知节点之间具有出现时间先后的关系，因此在结构上通常符合某种随机游走特征。依据采样策略不同，图随机游走采样一般包括深度优先采样（depth-first sampling，DFS）和广度优先采样（bread-first sampling，BFS）策略，其中DFS是指与当前节点相比，新采样节点与根节点的最短步长距离更长，与之相反，BFS则是指新采样节点与根节点的最短步长距离更短。此外，若新采样节点与根节点的最短步长距离不变，则为均衡采样策略。图随机游走技术在现有工作[26,33-36]中已有一定研究，然而这些方法均仅适用于直推式图表示学习问题[6]，且未考虑标注信息稀缺的半监督学习情形。本方法将随机游走技术引入基于GNN的生成式变分图自编码框架，并结合多种采样策略对图结构进行掩码以生成节点伪标签，从而缓解归纳式学习场景下的标注稀缺问题。

首先，本方法使用模型 $\mathcal{M}^{(t-1)}$ 基于节点随机掩码方法生成 K 次节点标签 $\hat{\boldsymbol{Y}}_k^{(t)}\in(0,1)^{N\times C}$，$k=1,\cdots,K$。为了适应归纳式学习问题中的可变图结构，每次随机将一些节点掩码为未知节点并得到掩码后的邻接矩阵 $\tilde{\boldsymbol{A}}_k^{(t)}$。

节点掩码 $\boldsymbol{m}_k^{(t)}=(m_{1,k}^{(t)},\cdots,m_{N,k}^{(t)})'\in\{0,1\}^N$ 通过随机游走采样得到。令 C 表示游走总步长，节点 i 为第 c 步（$c=2,\cdots,C-1$）掩码节点，即 $m_{i,k}^{(t)}=1$，则第 $c+1$ 步节点 j 的掩码概率为

$$P\left(m_{j,k}^{(t)}=1|m_{i,k}^{(t)}=1\right)=\begin{cases}\dfrac{\pi_{\mathrm{ij}}}{\sum\limits_{h\in\mathbb{N}_i}\pi_{\mathrm{ih}}}, & j\in\mathbb{N}_i,\\ 0, & \text{otherwise},\end{cases}\tag{1}$$

其中，$\mathbb{N}_i$ 表示节点 i 的邻居节点集，$\pi_{ij}>0$ 表示由节点 i 移动到邻居节点 j 的偏好。参考Grover和Leskovec[34]，令节点 r 为第 c-1 步掩码节点，则移动偏好 π_{ij} 定义为

$$\pi_{ij}=\begin{cases}\delta, & d(r,j)=0,\\ 1, & d(r,j)=1,\\ \gamma, & d(r,j)=2\end{cases}\tag{2}$$

其中，$d(r,j)$ 为节点 r 与 j 之间的步长距离，$d(r,j)=0$ 表示 r 与 j 为同一节点，δ 和 γ 分别表示BFS和DFS策略偏好超参数。

如图3所示，从当前节点 i 出发，若移动至节点 r，则 $d(r,j)=0$，新采样节点 j 返回至已采样节点 r，即表示BFS策略；若移动至节点 j_1，则 $d(r,j)=1$，新采样节点 j 较当前节点 i 与 r 之间的步长距离不变，即表示均衡采样策略；若移动至节点 j_2 或 j_3，则 $d(r,j)=2$，新采样节点 j 较当前节点 i 与 r 之间的步长距离增加，即表示DFS策略。因此，当 $\delta>\gamma$ 时，模型更倾向于采用BFS策略；反之，模型更倾向于采用DFS策略。

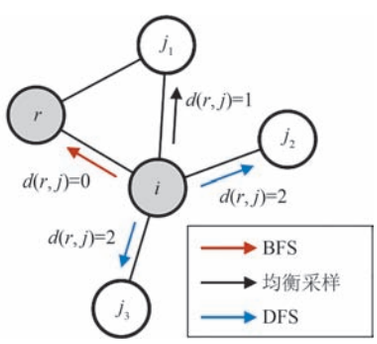

图3 基于随机游走的节点掩码

4.1.2 伪标签置信度过滤机制

为了保证伪标签具有较高的置信度，本方法对模型以不同节点掩码图 $\tilde{\boldsymbol{A}}_k^{(t)}$ 为输入预测得到的节点标签 $\hat{\boldsymbol{Y}}_k^{(t)}$ 取平均值，作为无标注节点的伪标签 $\bar{\boldsymbol{Y}}^{(t)}=(\bar{\boldsymbol{y}}_1^{(t)},\cdots,\bar{\boldsymbol{y}}_N^{(t)})'$，即

$$\bar{\boldsymbol{Y}}^{(t)}=\frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{Y}}_k^{(t)}.\tag{3}$$

然后，本方法通过设置一个阈值 θ 来过滤掉置信度较低的伪标签，即只有当 K 次生成的节点标签的均值高于此阈值时，模型才认为所生成的伪标签是有效的。因此，最终用于模型训练的增强节点标签 $\tilde{\boldsymbol{Y}}^{(t)}=(\tilde{\boldsymbol{y}}_1^{(t)},\cdots,\tilde{\boldsymbol{y}}_N^{(t)})'$ 为

$$\tilde{\boldsymbol{y}}_i^{(t)}=\begin{cases}\boldsymbol{y}_i, & i\in\mathcal{V}_{\mathrm{lab}},\\ \bar{\boldsymbol{y}}_i^{(t)}, & i\in\mathcal{V}\backslash\mathcal{V}_{\mathrm{lab}}\text{ and }\max\left(\bar{\boldsymbol{y}}_i^{(t)}\right)>\theta,\\ \boldsymbol{0}, & \text{otherwise},\end{cases}\tag{4}$$

其中，$\mathcal{V}$ 和 $\mathcal{V}_{\mathrm{lab}}$ 分别表示全部节点集和有标注节点集，$\max\left(\bar{\boldsymbol{y}}_i^{(t)}\right)$ 表示向量 $\bar{\boldsymbol{y}}_i^{(t)}$ 中元素的最大值。

本自训练增强方法具体执行步骤如算法1所示。该算法以图邻接矩阵$\boldsymbol{A}$、节点属性特征$\boldsymbol{X}$和独热编码的节点标签特征$\boldsymbol{Y}$作为输入，并输出使用伪标签增强的节点标签$\tilde{\boldsymbol{Y}}^{(t)}$。首先通过随机游走得到节点掩码$\boldsymbol{m}_k^{(t)}$（第1行），并对$\boldsymbol{A}$进行随机掩码得到掩码处理后的邻接矩阵$\tilde{\boldsymbol{A}}_k^{(t)}$（第3行），然后将$\tilde{\boldsymbol{A}}_k^{(t)}$、$\boldsymbol{X}$和$\boldsymbol{Y}$输入上轮迭代后的STAIG模型$\mathcal{M}(t-1)$以得到重构节点标签$\hat{\boldsymbol{Y}}_k^{(t)}$（第4行）。重复$K$次上述过程，并对多次生成的重构标签取平均得到伪标签$\bar{\boldsymbol{Y}}^{(t)}$（第6行）。最后，基于置信度过滤机制选择置信度较高的伪标签并得到最终的增强节点标签$\tilde{\boldsymbol{Y}}^{(t)}$（第7行）。

算法1.　自训练增强方法

输入：图邻接矩阵$\boldsymbol{A}$；节点属性特征$\boldsymbol{X}$；节点标签特征$\boldsymbol{Y}$

输出：第t次迭代的增强节点标签特征$\tilde{\boldsymbol{Y}}^{(t)}$

1．　FOR $k=1,\cdots,K$ DO

2．　　$\boldsymbol{m}_k^{(t)}=\mathrm{RandomWalk}(\boldsymbol{A})$；

3．　　$\tilde{\boldsymbol{A}}_k^{(t)}=\mathrm{Mask}(\boldsymbol{A},\boldsymbol{m}_k^{(t)})$；

4．　　$\hat{\boldsymbol{Y}}_k^{(t)}=\mathcal{M}^{(t-1)}(\tilde{\boldsymbol{A}}_k^{(t)},\boldsymbol{X},\boldsymbol{Y})$；

5．　ENDFOR

6．　$\bar{\boldsymbol{Y}}^{(t)}=\mathrm{Average}(\hat{\boldsymbol{Y}}_k^{(t)})$；

7．　$\tilde{\boldsymbol{Y}}^{(t)}=\mathrm{ConfidenceFiltering}(\bar{\boldsymbol{Y}}^{(t)})$；

下面本节分别对所提出的STAIG模型结构展开详细介绍，包括一个基于GCN的编码器和一个生成式标签重构解码器。

4.2　GNN编码器

编码器使用GNN执行邻域聚合以学习节点向量表示，例如GCN[7]、GraphSAGE[6]、GAT[8]等。本节以其中较为经典且参数量较少的GCN算子为例，使用归一化图拉普拉斯矩阵对邻居节点特征进行聚合。

给定一个包含N个节点的图邻接矩阵$\boldsymbol{A}\in\{0,1\}^{N\times N}$，则第$l$层（$l=1,\cdots,L$）的GCN隐藏表示$\boldsymbol{H}^{(l)}=(\boldsymbol{h}_1^{(l)},\cdots,\boldsymbol{h}_N^{(l)})'$为

$$\boldsymbol{h}_i^{(l)}=\mathrm{GCN}^{(l)}\left(\boldsymbol{A},\ \boldsymbol{h}_i^{(l-1)}\right),\tag{5}$$

其中，输入特征$\boldsymbol{H}^{(0)}$由节点属性特征$\boldsymbol{X}$（如果可用）与节点标签特征$\boldsymbol{Y}$（无标注节点和未知节点编码为零向量）拼接得到。此外，为了缓解半监督学习中的标注稀缺性问题，本方法使用基于节点掩码的自训练增强方法，通过迭代生成伪标签来增强数据标注。因此，输入特征可以表示为

$$\boldsymbol{H}^{(0)}=\left[\boldsymbol{X}\,\middle\|\,\tilde{\boldsymbol{Y}}\right],\tag{6}$$

其中，$\tilde{\boldsymbol{Y}}$是自训练增强后的节点标签特征。

然后利用GCN节点表示作为变分参数，通过MC采样生成正态隐变量$\boldsymbol{Z}=(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N)'$作为节点表示。对于$i=1,\cdots,N$，节点表示的生成方式为

$$\boldsymbol{z}_i\sim\mathrm{Normal}\left(\boldsymbol{\mu}_i,\mathrm{diag}\left(\boldsymbol{\sigma}_i^2\right)\right),\tag{7}$$

其中，均值$\boldsymbol{\mu}_i$和标准差$\boldsymbol{\sigma}_i$参数由GCN的输出层（$l=L$）得到。参考VGAE[24]，本方法采用重参数化技巧对参数进行梯度优化。

4.3　标签重构解码器

为了利用标签信息监督训练模型，本方法提出标签重构解码器，使用FNN分别对节点标签和节点属性特征进行重构，即

$$\hat{\boldsymbol{Y}}=\mathrm{softmax}\left(\mathrm{FNN}_y(\boldsymbol{Z})\right),\tag{8}$$

$$\hat{\boldsymbol{X}}=\mathrm{FNN}_x(\boldsymbol{Z}),\tag{9}$$

本模型的损失函数定义为重构损失以及节点表示的变分后验与先验分布之间的KL散度之和。重构损失包含重构节点标签和属性特征，对于有标注节点的标签重构，本模型采用交叉熵（cross entropy，CE）损失函数，对于无标注节点的伪标签重构以及所有节点的属性特征重构，本模型采用均方误差（mean square error，MSE）损失函数，即

$$\mathcal{L}_{\mathrm{lab}}=-\sum_{i\in\mathcal{V}_{\mathrm{lab}}}\boldsymbol{y}_i\cdot\log\left(\boldsymbol{y}_i\right),\tag{10}$$

$$\mathcal{L}_{\mathrm{pseu}}=\sum\nolimits_{i\in\mathcal{V}\setminus\mathcal{V}_{\mathrm{lab}}}\left\|\bar{\boldsymbol{y}}_i-\hat{\boldsymbol{y}}_i\right\|^2\cdot I_{[\max(\bar{\boldsymbol{y}}_i)>\theta]},\tag{11}$$

$$\mathcal{L}_{\mathrm{feat}}=\sum_{i\in\mathcal{V}}\left\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\right\|^2,\tag{12}$$

其中，$I_{[\max(\bar{\boldsymbol{y}}_i)>\theta]}$为一个示性函数，表示对伪标签$\bar{\boldsymbol{y}}_i$进行置信度过滤，即仅使用最大元素值高于阈值$\theta$的$\bar{\boldsymbol{y}}_i$来计算伪标签重构损失。

最后，完整的损失函数为

$$\mathcal{L}=\mathcal{L}_{\mathrm{lab}}+\lambda_{\mathrm{pseu}}\mathcal{L}_{\mathrm{pseu}}+\lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}}+\lambda_{\mathrm{kl}}\sum_{i\in\mathcal{V}}\mathrm{KL}\left[q(\boldsymbol{z}_i)\,\|\,p(\boldsymbol{z}_i)\right],\tag{13}$$

其中，$\mathrm{KL}\left[q(\boldsymbol{z}_i)\,\|\,p(\boldsymbol{z}_i)\right]$为节点表示$\boldsymbol{z}_i$的变分后验分布$q(\boldsymbol{z}_i)$和标准正态先验分布$p(\boldsymbol{z}_i)$之间的KL散度，$\lambda_{\mathrm{pseu}}\mathcal{L}_{\mathrm{pseu}}$、$\lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}}$和$\lambda_{\mathrm{kl}}$为超参数，分别用于调节伪标签重构损失$\mathcal{L}_{\mathrm{pseu}}$、属性特征重构损失$\mathcal{L}_{\mathrm{feat}}$和KL散度的权重。为了充分利用数据的真实标签优化模型，本方法设置各调节参数取值小于1，即有标注节点重构损失$\mathcal{L}_{\mathrm{lab}}$的相对权重最大。

模型训练步骤如算法2所示。该算法以图邻接矩阵A、节点属性特征$\boldsymbol{X}$和独热编码的节点标签特征$\boldsymbol{Y}$作为输入，并输出节点的重构标签$\hat{\boldsymbol{Y}}$。首先，依据所处训练阶段判断是否采用自训练增强方法来增强节点标签（第2-6行），并将节点属性特征$\boldsymbol{X}$与（自

训练增强后的）节点标签特征 $\tilde{\boldsymbol{Y}}$ 拼接作为编码器输入（第7行），进而使用GCN编码器学习正态分布的变分后验参数并通过MC采样生成正态隐变量 $\boldsymbol{Z}$（第8-13行）。然后，使用解码器通过FNN重构节点标签 $\hat{\boldsymbol{Y}}$ 和属性特征 $\hat{\boldsymbol{X}}$（第14、15行），并计算重构损失和KL散度（第16-19行）。最后，使用随机梯度下降算法优化权值参数（第20行，其中 $\omega^{(t)}$ 表示第 t 次迭代后的权值参数集合，κ 表示学习率）。重复上述过程，直至模型收敛。

算法2. STAIG模型训练

输入：图邻接矩阵A；节点属性特征 $\boldsymbol{X}$；节点标签特征 $\boldsymbol{Y}$

输出：重构节点标签特征 $\hat{\boldsymbol{Y}}$

1. FOR $t=1,\cdots,T$ DO
2. IF $t\leqslant T_0$ THEN
3. $\tilde{\boldsymbol{Y}}=\boldsymbol{Y}$;
4. ELSE
5. $\tilde{\boldsymbol{Y}}=\text{Augmentation}(\boldsymbol{A},\boldsymbol{X},\boldsymbol{Y})$;
6. ENDIF
7. $\boldsymbol{H}^{(0)}=[\boldsymbol{X}\|\tilde{\boldsymbol{Y}}]$;
8. FOR $l=1,\cdots,L-1$ DO
9. $\boldsymbol{H}^{(l)}=\text{GCN}_h^{(l-1)}(\boldsymbol{A},\boldsymbol{H}^{(l-1)})$;
10. ENDFOR
11. $\boldsymbol{\mu}=\text{GCN}_{\mu}^{(L)}(\boldsymbol{H}^{(L-1)})$;
12. $\boldsymbol{\sigma}=\text{GCN}_{\sigma}^{(L)}(\boldsymbol{H}^{(L-1)})$;
13. $\boldsymbol{Z}=\text{NormalSampling}(\boldsymbol{\mu},\boldsymbol{\sigma})$;
14. $\hat{\boldsymbol{Y}}=\text{softmax}(\text{FNN}_y(\boldsymbol{Z}))$;
15. $\hat{\boldsymbol{X}}=\text{FNN}_x(\boldsymbol{Z})$;
16. $\mathcal{L}_{\text{lab}}=\text{CE}(\boldsymbol{Y},\hat{\boldsymbol{Y}})$;
17. $\mathcal{L}_{\text{pseu}}=\text{MSE}(\tilde{\boldsymbol{Y}},\hat{\boldsymbol{Y}})$;
18. $\mathcal{L}_{\text{feat}}=\text{MSE}(\boldsymbol{X},\hat{\boldsymbol{X}})$;
19. $\mathcal{L}=\mathcal{L}_{\text{lab}}+\lambda_{\text{pseu}}\mathcal{L}_{\text{pseu}}+\lambda_{\text{feat}}\mathcal{L}_{\text{feat}}+\lambda_{\text{kl}}\text{KL}[q(\boldsymbol{Z})\|p(\boldsymbol{Z})]$;
20. $\omega^{(t)}\leftarrow\omega^{(t-1)}-\kappa\nabla_{\omega}\mathcal{L}$;
21. ENDFOR

5 实验与分析

为了验证本文提出的方法在归纳式图表示学习上的有效性，本节在多个基准图数据集上进行有监督和半监督归纳式节点分类实验，并进一步进行参数敏感性实验。

5.1 数据集

本实验考虑四个基准真实世界归纳式图学习数据集，包括一个跨图网络，即Flickr[19]，以及三个演化网络，即Reddit[6]、Elliptic[3]和ogbn-arxiv[37]。这些数据集已在归纳式图表示学习和节点分类相关研究中广泛使用，且节点数量规模较大，因而可以用于更好地验证所提出方法的有效性和可扩展性等。各数据集的主要描述性统计信息如表2所示，其详细介绍如下：

（1）Flickr是由一个大型图片分享网站采集的由800个自我中心网络（ego network）构成的跨图网络，每张图片作为一个节点，如果两张图片之间存在一些共同的元信息（例如位置和标签等），则形成一条边。这些图片根据其标签不同进行分类。节点属性特征为每张图片的词袋特征向量。该数据集以每个自我中心网络为单位进行分割，并分别随机选择50%、25%、25%的自我中心网络作为训练集、测试集和验证集。

（2）Reddit数据集是由一个新闻评论网站在2014年9月期间的全部发帖构成的演化网络，其中每个节点代表一篇帖子，如果两个帖子被同一用户评论过，则两个节点之间形成一条边。节点标签表示帖子所属社区。节点属性特征为帖子标题和评论的词向量。该数据集按发帖时间对节点进行分割，其中前20天的帖子用作训练集，后10天的帖子中70%用作测试集，30%用作验证集。

（3）Elliptic是一个约两周内比特币交易的演化网络，其中节点表示交易，边表示比特币流。每个节点被标记为“合法”或“非法”两类。节点属性特征由除交易时间以外的其他交易信息构成。该数据集按交易时间对节点进行分割，其中第一周的交易用作训练集，第二周的交易中70%用作测试集，30%用作验证集。

（4）ogbn-arxiv是一个1971～2020年发表的计算机科学领域论文引用演化网络，其中每个节点表示一篇论文，边表示论文间的引用关系（本实验假设所有边为无向边）。所有节点依据不同研究方向分为40个类别。节点属性特征为每篇论文题目和摘要的平均词向量。该数据集按论文发表时间对节点进行分割，其中2017年及以前的用作训练集，2018的用作验证集，2019年及以后的用作测试集。

为了比较模型在不同弱监督或半监督学习条件下的归纳式图学习性能，本实验考虑了两种不同的

表2 归纳式节点分类数据集描述性统计信息

	可见节点数	未知节点数	可见边数	未知边数	特征维度	类别数
Flickr	44,625	44,625	218,140	681,616	500	7
Reddit	153,932	79,033	10,753,238	12,460,600	602	41
Elliptic	22,722	13,152	19,500	17,124	165	2
ogbn-arxiv	90,941	78,402	738,066	1,577,532	128	40

训练集标注比例。具体来说，对于每个数据集，本实验分别在训练集中随机选取1%和10%的节点作为有标注数据，训练集中所有其他节点均作为无标注数据。依照归纳式学习的标准设置，训练过程中测试集和验证集的所有节点（及其属性特征和相关的边）均不可见。

5.2　对比方法

本实验将所提出的STAIG模型与目前最佳节点分类方法进行了比较，包括基于GNN的半监督或弱监督节点分类方法，即GCN[7]、BNE[9]、D^{2}PT[10]、MGCN[16]和LGGCL[17]，以及基于GNN的归纳式节点分类方法，即GraphSAGE[6]、GAT[8]、GraphSAINT[19]和GAMLP[13]，以及基于VAE的方法，即VGAE[24]、GraphMAE[25]和MaskGAE[26]。

基于GNN的半监督或弱监督节点分类方法详细介绍如下：

（1）GCN基于图拉普拉斯算子执行邻域聚合来学习节点的向量表示，是最经典的半监督图表示学习方法之一。

（2）BNE针对节点类别分布不均衡问题，通过一种邻居探索算法来依据节点间的拓扑关系构建数据类别分布均衡的训练集。

（3）D^{2}PT针对结构、特征和标注不完全的弱信息场景，使用一种基于图扩散机制的对偶模型框架来缓解数据不足和孤立节点问题。

（4）MGCN基于线性插值得到多视图节点插值表示和标签，并通过最小化不同视图之间同一节点的插值表示来缓解表示坍塌问题。

（5）LGGCL利用节点伪标签划分对比学习正负样本，并分别基于属性特征聚类对齐对正样本进行筛选，基于伪标签概率分布对负样本进行重加权。

基于GNN的归纳式节点分类方法详细介绍如下：

（1）GraphSAGE是首个针对归纳式学习问题设计的一种基于图拓扑进行邻居节点采样和聚合的图表示学习方法。

（2）GAT利用注意力机制来学习节点的邻域聚合表示，并通过使用一个额外的神经网络来学习注意力权重。

（3）GraphSAINT通过对子图而非节点或边进行采样来提升归纳式图表示学习的效率和性能。

（4）GAMLP利用一种标签传播方法来强化模型对于数据标注信息的利用，该方法是目前在节点分类任务上准确率最高的方法之一。

基于VAE的对比方法详细介绍如下：

（1）VGAE首次将VAE框架应用于图表示学习，该方法使用GCN编码器生成正态隐变量作为节点表示，并通过重构邻接矩阵来无监督地训练模型。

（2）GraphMAE对节点的属性特征进行随机掩码作为模型的输入特征，然后通过重构原始属性特征来训练模型。

（3）MaskGAE对图中的某些边或路径进行随机掩码，以减轻模型对图数据的邻近结构特征过拟合的问题。

由于基于VAE的方法无法以端到端训练的方式利用标签信息。依照这类方法的标准实验设置[25-26]，本实验首先以无监督学习方式训练这些模型，然后使用所学习的节点表示训练一个逻辑回归模型来进行节点分类。

5.3　超参数设置

对于本文所提出的STAIG模型，本实验中编码器使用2个GCN层，每层的特征维度为512，解码器使用3个全连接层，每层的特征维度依然为512。伪标签重构损失权重λ_{pseu}和属性特征重构损失权重λ_{feat}均设置为0.1，KL散度权重采用预热方法动态调整，即随着训练迭代次数增加，λ_{kl}取值从0.001逐渐增大至1。

在训练过程中，模型首先训练1个时期（epoch）作为预热阶段，然后进入增强阶段并使用所提出的自训练增强方法进行标注数据增强，其中超参数设置为生成次数$K\in\{1,2\}$，游走步长$C=5$，置信度阈值$\theta=0.9$，随机游走起点通过伯努利采样从所有可见节点中以概率0.2随机选取，采样策略参数δ和γ通过网格搜索得到。学习率设置为$\kappa\in\{0.001,0.005\}$。对比方法的GNN层数及各层特征维度设置为与STAIG模型编码器相同，其他模型超参数使用其开源代码中的默认值。所有模型均基于NVIDIA A100 40 GB GPU设备使用Adam优化器[38]进行梯度优化，并使用早停（early-stopping）训练策略以减轻过拟合问题，迭代次数少于1000个epoch。

5.4　节点分类实验结果和分析

表3展示了各数据集上的半监督节点分类实验结果。实验结果验证了本文所提出的STAIG模型在半监督学习设置下具有至少与对比方法相当的性能，特别在弱监督学习（1%标注数据比例）设置下，STAIG模型显著优于所有对比方法。

表3 半监督归纳式节点分类准确率(%)实验结果

	Flickr		Reddit		Elliptic		ogbn-arxiv	
	1%	10%	1%	10%	1%	10%	1%	10%
GCN	42.8±0.3	47.3±0.2	92.1±0.0	94.0±0.0	69.4±0.3	85.4±0.3	61.5±0.5	66.1±0.1
BNE	45.6±0.2	48.1±0.1	91.2±0.1	93.8±0.1	88.9±0.1	89.6±0.2	61.2±0.2	66.0±0.3
D^2PT	45.5±0.9	48.2±0.7	79.2±0.4	80.1±0.6	88.2±2.9	89.5±0.1	23.4±1.3	59.6±0.6
MGCN	47.2±0.2	8.4±0.1	92.3±0.1	93.9±0.0	88.4±0.6	86.5±0.6	60.6±0.1	66.0±0.1
LGGCL	45.8±0.4	48.1±0.3	91.4±0.2	93.2±0.1	89.0±0.6	89.2±0.5	60.4±0.1	65.7±0.1
GraphSAGE	34.3±0.5	41.1±0.4	88.9±0.1	93.9±0.0	58.8±0.4	86.7±0.3	58.0±0.7	63.4±0.2
GAT	34.3±0.5	41.1±0.4	91.6±0.3	93.7±0.1	70.1±0.9	86.9±0.5	60.9±0.5	65.9±0.3
GraphSAINT	41.4±1.1	47.6±1.3	66.0±0.6	91.6±0.7	77.9±1.7	89.8±0.4	59.0±0.1	61.7±0.5
GAMLP	32.5±2.5	41.0±1.4	84.6±1.4	94.3±0.9	85.5±5.4	90.1±0.3	56.8±0.3	66.3±0.1
VGAE	39.6±0.6	42.5±0.8	64.2±2.3	73.0±1.2	62.5±0.6	82.2±0.7	49.4±2.5	57.2±2.7
GraphMAE	42.5±0.6	43.1±0.6	91.8±0.1	93.2±0.4	84.8±0.6	86.0±0.8	57.6±0.1	61.2±0.3
MaskGAE	42.7±0.3	45.5±1.2	88.1±0.4	93.5±0.0	85.6±2.1	86.8±1.4	59.7±0.2	63.6±0.2
STAIG	**48.3±0.1**	**48.8±0.5**	**93.8±0.1**	**94.8±0.0**	**90.1±0.2**	**90.7±0.2**	**63.6±0.5**	**67.3±0.3**

注:加粗表示最优结果,下同。

具体而言,随着标注比例从10%下降到1%,大多数对比方法的性能显著下降。例如,GAMLP在Flickr、Reddit和Elliptic数据集上的分类准确率分别下降了约8.5%、9.7%和4.6%。相比之下,本文提出的STAIG模型在这三个数据集上的分类准确率分别仅下降了1.3%、1.0%和0.6%。这些结果表明,STAIG在处理不同类型数据集时,均展现出了强大的泛化能力,特别是在标注数据稀缺的情况下,依然能够保持优异的性能。

在弱监督学习条件下,有限的标注数据无法支撑对大量的模型参数进行优化,因此对比方法在标注数据稀缺的情况下易表现出显著性能下降。另一方面,STAIG模型可以通过本文介绍的自训练增强方法有效缓解标注稀缺带来的参数优化困难,同时减轻模型可能对少量标注数据产生过拟合的问题。具体来说,首先,STAIG利用了自训练增强方法,通过迭代地为未标注数据分配伪标签,在训练过程中增加了有效的监督信息。这种方法在弱监督环境下尤其有效,因为它能够利用大量未标注的数据,从而提升模型的泛化能力和稳健性。其次,STAIG模型采用随机游走对节点进行掩码,该掩码方法能够在标注数据稀缺的情况下更加充分地挖掘图结构信息,提高了模型对结构信息和标注信息均不完全的半监督归纳式学习场景的适应性。

为了更全面地展示各模型在不同标注数据比例下的性能变化情况,本节对STAIG模型与几种最有代表性的对比方法进行了详细对比。这些对比方法包括半监督节点分类方法BNE和D^2PT,归纳式学习方法GraphSAINT和GAMLP,以及基于VAE的方法MaskGAE。实验考虑了1%、5%、10%、20%、50%以及100%等多种训练数据标注比例,结果如图4所示。实验结果直观地表明,在不同的标注比例下,STAIG模型的性能变化整体较对比方法更加平稳,且随着标注比例的降低,STAIG模型的优势不断增强。

具体来说,当标注数据比例极低时,STAIG模型表现显著优于其他方法,特别是在处理Reddit动态演化网络中的新增节点时,STAIG的自训练增强方法

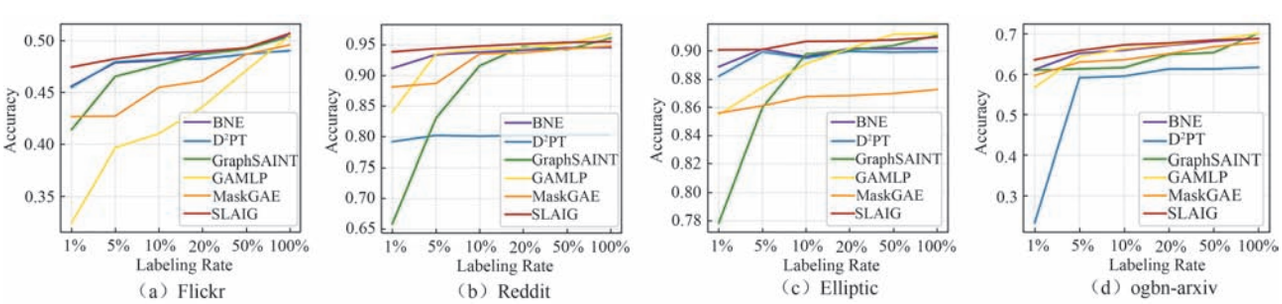

图4 不同标注比例下节点分类实验结果

显现出其独特优势。BNE和D^2PT等半监督方法虽然在部分情况下表现良好，但总体上仍不如STAIG稳健。随着标注数据比例增加，各方法的性能都有所提升，但STAIG的优势依然明显，特别是在处理不连通小规模子图时，其自训练增强方法和高阶节点邻域信息利用效果更为突出。在标注数据比例较高的情况下，STAIG依然保持领先地位。尽管部分方法如GraphSAINT和GAMLP在标注比例50%以上时也表现较好，但在综合性能上仍逊于STAIG。

此外，STAIG在所有标注数据比例下都表现出较高的稳健性，尤其在标注数据稀缺的情况下，其性能优势更加明显。这表明STAIG能够有效利用少量标注数据，通过自训练增强和高阶节点邻域信息的利用，提升模型的泛化能力。在各归纳式学习图数据集上，STAIG均具有显著性能优势。这主要是因为本模型通过自训练增强方法生成具有高置信度的节点伪标签，使其在较低标注数据比例的弱监督学习设置下保持对未知节点的分类推理能力。尽管部分对比方法如GraphSAINT和GAMLP在高标注数据比例下表现较好，但在标注数据稀缺的情况下，其性能显著下降。这表明这些方法在处理弱监督学习任务时存在一定的局限性，无法像STAIG一样有效应对标注数据不足的问题。综上所述，STAIG在不同标注数据比例下都表现出优越的性能和稳健性，特别是在标注数据稀缺的情况下，其优势更加显著。

5.5　消融实验结果和分析

本节还对STAIG模型中各组件的有效性进行了消融实验验证。消融实验设计了以下五种变体：

（1）w/o walk表示不使用随机游走，而采用伯努利采样对节点进行随机掩码（掩码概率为0.2）以生成伪标签。

（2）w/o mask表示使用未作掩码处理的图生成伪标签。

（3）w/o filter表示不对所生成的伪标签进行置信度过滤，直接用作增强节点标签。

（4）w/o pseudo表示仅使用真实标签作为输入标签特征并对其进行重构。

（5）w/o feature表示去除特征重构损失。

实验结果如表4所示。结果显示，完整的STAIG模型在所有数据集和标注比例下均表现最佳。具体来说，对于自训练增强方法的掩码策略，通过对比STAIG与w/o walk变体结果可以看出，与完全随机掩码相比，基于随机游走对节点进行掩码能够帮助模型更好地适应真实归纳式学习图数据中的未知节点分布，从而有效提升模型性能。进一步对比w/o walk与w/o mask变体，可以看出完全不使用节点掩码方法生成伪标签会导致模型性能进一步下降，这表明掩码处理能够帮助模型更有效地利用图的结构信息，提高生成伪标签的质量，从而提升分类性能。此外，对于伪标签置信度过滤机制，对比STAIG与w/o filter变体结果可以看出，通过筛选具有高置信度的伪标签作为最终的增强标签可以避免模型在自训练阶段被噪声标签干扰而发生负优化，保证自训练增强方法的收敛性，从而提高模型的节点分类性能。

表4　半监督归纳式节点分类准确率（%）消融实验结果

	Flickr		Reddit		Elliptic		ogbn-arxiv	
	1%	10%	1%	10%	1%	10%	1%	10%
STAIG	**48.3±0.1**	**48.8±0.5**	**93.8±0.1**	**94.8±0.0**	**90.1±0.2**	**90.7±0.2**	**63.6±0.5**	**67.3±0.3**
w/o walk	46.7±1.1	48.3±0.8	93.1±0.1	94.6±0.1	90.0±0.1	90.3±0.3	62.3±0.9	66.4±0.2
w/o mask	45.7±1.0	47.3±0.5	92.6±0.1	94.1±0.1	88.3±0.2	89.4±0.7	59.3±0.6	62.5±0.5
w/o filter	46.0±1.2	47.5±0.8	92.2±0.3	93.1±0.3	87.4±0.6	88.6±0.9	55.3±1.4	60.9±1.0
w/o pseudo	45.3±0.0	47.2±0.2	92.2±0.2	93.5±0.5	88.2±0.1	89.0±0.6	59.1±0.7	62.6±0.5
w/o feature	47.2±0.2	48.0±0.4	93.3±0.1	93.6±0.1	86.0±3.7	88.1±0.0	62.2±0.4	65.5±0.5

对于模型损失函数中的伪标签重构损失，对比STAIG与w/o pseudo变体，可以看出重构具有高置信度的节点伪标签能够有效提升模型性能，这表明伪标签在弱监督学习设置下能够显著缓解标注数据稀缺的问题，提升模型的泛化能力。对于属性特征重构损失，对比STAIG与w/o feature变体结果可以看出，特征重构损失对模型性能的提升具有重要作用，这表明最小化特征重构损失能够帮助模型更好地学习节点的属性信息，从而提升分类性能。

5.6　参数敏感性分析

本实验还对所提出的STAIG模型中的5个重要超参数进行了敏感性分析，包括生成次数K、游走

步长C、置信度阈值θ、伪标签重构损失权重λ_{pseu}和属性特征重构损失权重λ_{feat}。

生成次数K敏感性分析实验结果如图5所示，其中各曲线表示数据集的不同标注比例，阴影部分表示基于3次独立重复试验的95%置信区间。结果表明，当$K\leqslant 2$时，STAIG模型表现更优。随着K取值增大，模型性能整体趋于平稳，但在某些数据集（如Elliptic）上呈现一定下降趋势，这可能是由于过大的K值会导致通过置信度筛选的有效伪标签数量太少，从而使得所提出的自训练增强方法未能充分发挥作用。

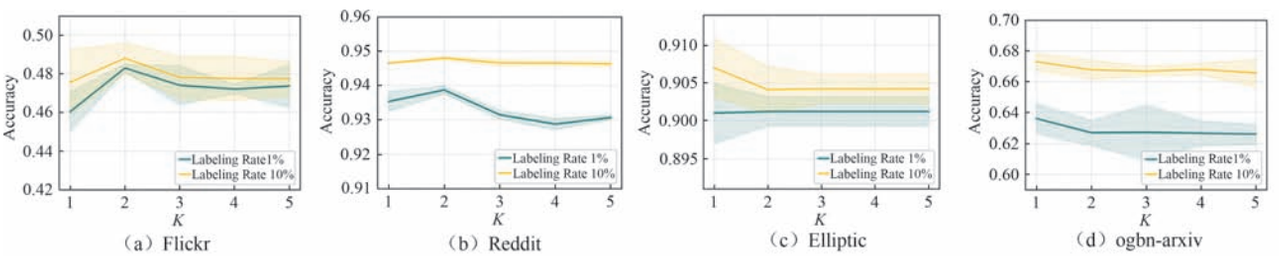

图5 生成次数K参数敏感性分析结果

游走步长C敏感性分析实验结果如图6所示。结果表明，当C取值接近5时，模型性能达到峰值。太大的C值会导致掩码节点过多，从而降低模型生成的伪标签的置信度。而太小的C值则会导致掩码节点过少，从而影响模型推理未知图结构的泛化性。注意当$C=1$时，模型退化为基于伯努利采样随机掩码的变体。因此，选择适中的C值（例如$C=5$）可以在提高伪标签置信度和保持模型泛化性之间取得平衡，从而提升模型的总体性能。

置信度阈值θ敏感性分析实验结果如图7所示。结果表明，随着θ趋近于1，分类准确率整体呈上升趋势。这验证了较高的伪标签置信度对于提升模型性能的必要性。高置信度阈值能够确保模型仅使用高质量的伪标签进行训练，减少噪声标签的影响，从而提高分类准确率。然而，过高的阈值也可能导致伪标签数量不足，无法充分利用未标注数据的信息。因此，在实际应用中，需要根据具体情况选择适当的置信度阈值，以在保证伪标签质量和数量之间取得平衡。

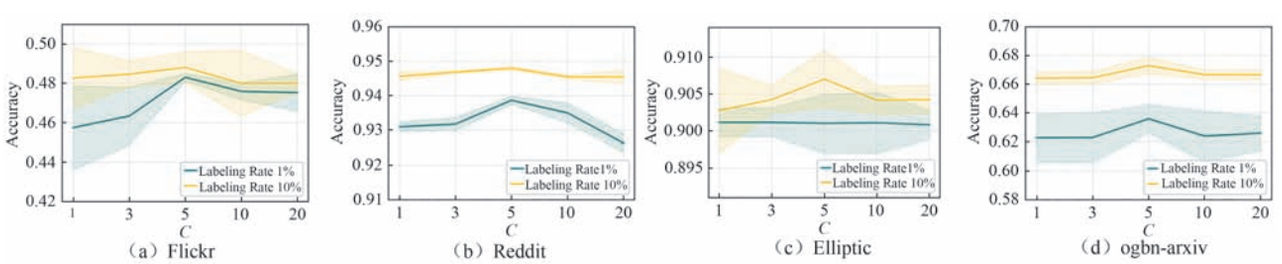

图6 游走步长C参数敏感性分析结果

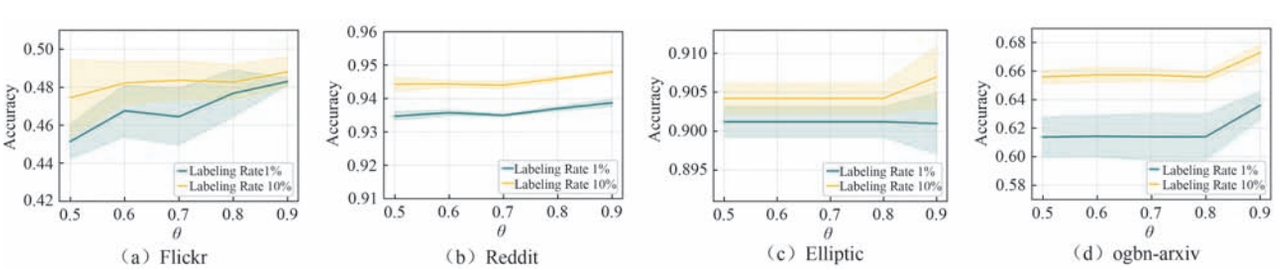

图7 置信度阈值θ参数敏感性分析结果

伪标签重构损失权重λ_{pseu}敏感性分析实验结果如图8所示。结果表明，当λ_{pseu}取值较小时模型分类性能更优，且在大多数数据集上随着该参数取值增大，模型性能变化较为稳定。而在Elliptic数据集上，当λ_{pseu}值大于0.5时，模型性能有明显下降，这表明过大的λ_{pseu}值可能导致模型对所生成伪标签中的噪声信息过拟合。因此，选择较小的λ_{pseu}值更有利于模型在优化过程中充分利用数据中的真实标签。

属性特征重构损失权重λ_{feat}敏感性分析实验结果如图9所示。结果与λ_{pseu}参数实验结果相似，即当

λ_{feat}取值较小时模型分类性能更优。其中，在Elliptic数据集上，当λ_{feat}值大于1时，模型在1%标注比例下的分类性能有较明显的下降，这表明过大的λ_{feat}值可能导致模型在优化过程中过分重视节点属性特征。因此，应选择较小的λ_{feat}值，避免模型对通常更加重要的节点类别标签欠拟合。

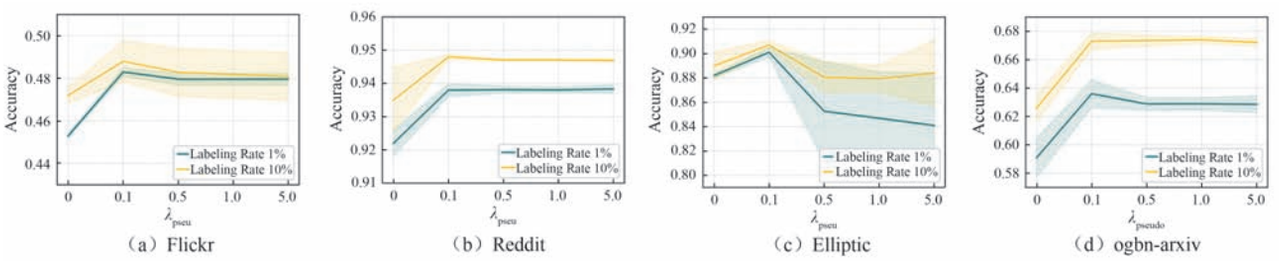

图8　伪标签重构损失权重λ_{pseu}参数敏感性分析结果

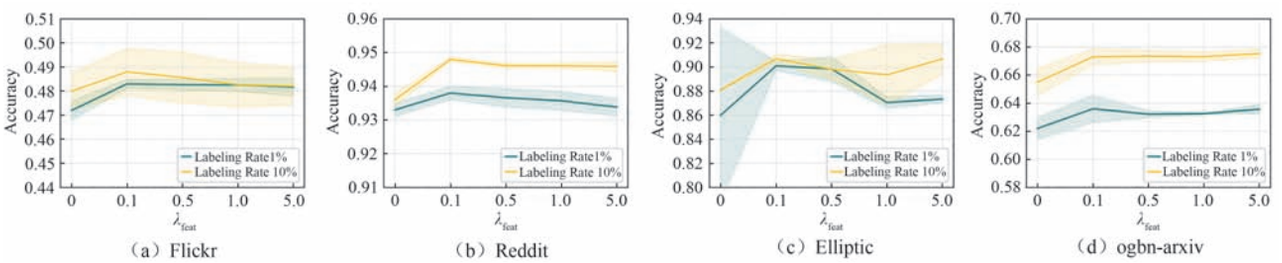

图9　属性特征重构损失权重λ_{feat}参数敏感性分析结果

6　总　结

本文针对半监督归纳式图表示学习问题提出了生成式STAIG模型。该模型由一个通过执行邻域聚合学习节点表示的GCN编码器和一个通过最小化节点标签和属性特征重构损失进行模型训练的标签重构解码器组成。为了基于VGAE框架利用节点的标注信息，该模型将节点标签编码为独热标签特征，然后与节点属性特征拼接构成模型的输入特征，再对输入的标签和属性特征（而非邻接矩阵）进行重构。此外，为了减轻半监督学习设置下节点标注的稀缺性并提高模型处理归纳式学习问题的泛化性，该模型提出一种自训练增强方法，通过对节点进行随机掩码来使用模型自身生成伪节点标签，用以增强数据的标注信息。最后，实验验证了本文提出的STAIG模型在半监督节点分类任务的有效性，在弱监督学习场景下具有显著优势。

在本文基础上，未来计划将本研究内容与图预训练及其他数据稀疏应用领域相结合。例如，在图预训练领域，一些最新研究采用基于“预训练-提示-微调”的范式来提升图预训练模型在多种下游任务上的性能表现[39-40]，但目前仍缺少针对归纳式图表示学习问题的图预训练模型研究，因此未来工作可以着重建立和改进在归纳式学习场景下的图预训练以及图提示方法。此外，针对一些数据资源稀疏的应用场景，如具有隐藏用户节点的社交网络谣言检测[41]以及稀疏关系预测[42]等任务，未来工作可以通过建立图模型挖掘社交网络用户的结构信息，并采用自训练增强方法缓解标注数据稀缺问题。

参　考　文　献

[1] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA, 2005: 177-187

[2] Paranjape A, Benson A R, Leskovec J. Motifs in temporal networks//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. Cambridge, UK, 2017: 601-610

[3] Weber M, Domeniconi G, Chen J, et al. Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics//Proceedings of the KDD 2019 Workshop on Anomaly Detection in Finance. Anchorage, USA, 2019

[4] Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 2005, 102(43): 15545-15550

[5] Rozemberczki B, Davies R, Sarkar R, et al. GEMSEC: Graph

embedding with self clustering//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Vancouver, Canada, 2019: 65-72

[6] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs//Advances in Neural Information Processing Systems: Vol. 30. Long Beach, USA, 2017

[7] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the 2017 International Conference on Learning Representations. Toulon, France, 2017

[8] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks//Proceedings of the 2018 International Conference on Learning Representations. Vancouver, Canada, 2018

[9] Zhu Z, Xing H, Xu Y. Balanced neighbor exploration for semi-supervised node classification on imbalanced graph data. Information Sciences, 2023, 631: 31-44

[10] Liu Y, Ding K, Wang J, et al. Learning strong graph neural networks with weak information//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA, 2023: 1559 - 1571

[11] Wu Y, Wang Y, Wang X, et al. Motif-Based Hypergraph Convolution Network for Semi-Supervised Node Classification on Heterogeneous Graph. Chinese Journal of Computers. 2021, 44(11): 2248-2260 (in Chinese)
(吴越, 王英, 王鑫等. 基于超图卷积的异质网络半监督节点分类. 计算机学报, 2021, 44(11): 2248-2260)

[12] Sun X, Yin H, Liu B, et al. Heterogeneous hypergraph embedding for graph classification//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. Virtual, Israel, 2021: 725-733

[13] Zhang W, Yin Z, Sheng Z, et al. Graph attention multi-layer perceptron//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA, 2022: 4560 - 4570

[14] Zhang C, He Y, Cen Y, et al. SCR: Training graph neural networks with consistency regularization. 2022. arXiv: 2112.04319

[15] Sun C, Hu J, Gu H, et al. Scalable and adaptive graph neural networks with self-label-enhanced training. Pattern Recognition, 2025, 160: 111210

[16] Yang X, Wang Y, Liu Y, et al. Mixed graph contrastive network for semi-supervised node classification. ACM Transactions on Knowledge Discovery from Data, 2024, 18(7): 1-19

[17] Peng M, Juan X, Li Z. Label-guided graph contrastive learning for semi-supervised node classification. Expert Systems with Applications, 2024, 239: 122385

[18] Xu D, Ruan C, Korpeoglu E, et al. Inductive representation learning on temporal graphs//Proceedings of the 2020 International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020:20-30

[19] Zeng H, Zhou H, Srivastava A, et al. GraphSAINT: Graph sampling based inductive learning method//Proceedings of the 2020 International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[20] Wang Y, Chang Y Y, Liu Y, et al. Inductive representation learning in temporal networks via causal anonymous walks//Proceedings of the 2021 International Conference on Learning Representations. Virtual, Austria, 2021

[21] Gao X, Chen T, Zang Y, et al. Graph condensation for inductive node representation learning//Proceedings of the IEEE 40th International Conference on Data Engineering. Utrecht, The Netherlands, 2024: 3056-3069

[22] Yao H Y, Zhang C Y, Yao Z L, et al. A recurrent graph neural network for inductive representation learning on dynamic graphs. Pattern Recognition, 2024, 154: 110577

[23] Sun X, Cheng H, Liu B, et al. Self-supervised hypergraph representation learning for sociological analysis. IEEE Transactions on Knowledge and Data Engineering, 2023, 35 (11): 11860-11871

[24] Zhang S, Chen H, Yang H, et al. Graph masked autoencoders with transformers. 2022, arXiv: 2202.08391

[25] Kipf T N, Welling M. Variational graph auto-encoders//Proceedings of the NeurIPS 2016 Workshop on Bayesian Deep Learning. Barcelona, Spain, 2016

[26] Hou Z, Liu X, Cen Y, et al. GraphMAE: Self-supervised masked graph autoencoders//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA, 2022: 594 - 604

[27] Li J, Wu R, Sun W, et al. What's behind the mask: Understanding masked graph modeling for graph autoencoders//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA, 2023: 1268 - 1279

[28] Kingma D P, Welling M. Auto-encoding variational Bayes//Proceedings of the 2014 International Conference on Learning Representations. Banff, Canada, 2014

[29] Tu W, Liao Q, Zhou S, et al. RARE: Robust masked graph autoencoder. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(10): 5340-5353

[30] Zheng Y, Jia C. ProtoMGAE: Prototype-aware masked graph auto-encoder for graph representation learning. ACM Transactions on Knowledge Discovery from Data, 2024, 18 (6): 137

[31] Yuan X, Zhang C, Tian Y, et al. Mitigating severe robustness degradation on graphs//Proceedings of the 2024 International Conference on Learning Representations. Vienna, Austria, 2024

[32] Tan Q, Liu N, Huang X, et al. MGAE: Masked autoencoders for self-supervised learning on graphs. 2022. arXiv: 2201.02534

[33] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2014: 701-710

[34] Grover A, Leskovec J. node2vec: Scalable feature learning for networks//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 855-864

[35] Nikolentzos G, Vazirgiannis M. Random walk graph neural networks//Advances in Neural Information Processing

Systems: Vol. 33. Virtual, 2020: 16211-16222

[36] Chen D, Schulz T H, Borgwardt K. Learning long range dependencies on graphs via random walks. arXiv: 2406.03386, 2024

[37] Hu W, Fey M, Zitnik M, et al. Open graph benchmark: Datasets for machine learning on graphs//Advances in Neural Information Processing Systems. Virtual, 2020, 33: 22118-22133

[38] Kingma D P, Ba J. Adam: A method for stochastic optimization//Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA, 2015

[39] Sun M, Zhou K, He X, et al. GPPT: Graph pre-training and prompt tuning to generalize graph neural networks// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA, 2022: 1717-1727

[40] Sun X, Cheng H, Li J, et al. All in one: Multi-task prompting for graph neural networks//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA, 2023: 2120-2131

[41] Liu B, Sun X, Ni Z, et al. Co-Detection of crowdturfing microblogs and spammers in online social networks. World Wide Web, 2020, 23: 573-607

[42] Guan Y, Sun X, Sun Y. Sparse relation prediction based on hypergraph neural networks in online social networks. World Wide Web, 2023, 26(1): 7-31

YANG Han-Xuan, Ph. D. His research interests include graph representation learning and variational auto-encoders.

YU Zhao-Xin, Ph. D. candidate. His research interests include deep learning and emotion cognitive modeling.

LI Zi-Qian, M. S., senior engineer. His research interests include artificial intelligence technology research and application.

XU Hui-Fang, M. S., senior engineer. Her research interests include knowledge graphs in the electric power sector and graph machine learning.

KONG Qing-Chao, Ph. D., associate professor. His research interests include social computing and large language models.

Background

Graph representation learning aims to learn low-dimensional embeddings of graph nodes and has become a critical problem with a number of downstream applications. For many real-world graph data, such as the evolving networks and cross-graph networks, many graph nodes and their related edges are unseen during the training process, a. k. a., the inductive learning problem, which requires graph models to adapt to variable graph structures and infer representations of the unseen nodes. Furthermore, in practice, due to the data incompleteness or expensive annotating cost, nodes of the visible structure can be scarcely labeled. This brings about the problem, which is even more challenging since not only the proximity information of the unseen structure is unavailable, but also the label scarcity problem often leads to the over-fitting issue.

Although the semi-supervised graph learning tasks, represented by node classification, have been well studied by extensive previous work, these methods are trained under the transductive learning setting, which assumes the graph structure to be static with all nodes visible during both the training and inference processes. For inductive learning, existing graph models are typically based on GNNs. These methods have shown good generalizability for predicting the unseen nodes, but they rely on plenty of annotated nodes under the supervised learning setting. In summary, existing inductive graph models tend to be over-fitted for (weakly) semi-supervised learning due to the lack of annotated nodes, and thus the semi-supervised inductive learning problem still remains to be investigated.

In this paper, we focus on the semi-supervised inductive graph representation learning problem and propose the STAIG model. Our model includes a GCN encoder and a novel label reconstruction decoder. To deal with the label scarcity problem under semi-supervised learning, the encoder takes node labels as one-hot input features, which have been augmented with pseudo-node labels generated by the model itself. In addition, to adapt to the variable graph structure between the training and inference processes, we randomly mask some nodes and reconstruct the labels of masked nodes in the decoder, so as to boost the model generalizability for inferring the representations of unseen nodes. Experimental results based on inductive learning graph datasets verify the effectiveness of our model for semi-supervised node classification.

This work was supported by the Science and Technology Program of the Headquarters of State Grid Corporation of China, Research on Key Technologies of Electric Power Large Language Model and Its Demonstration Application in Intelligent Customer Service, under Grant No. 5700-202353595A-3-2-ZN.