doi:10.12068/j.issn.1005－3026.2025.20240153

基于属性补全的异质图表示学习算法

陈东明，刘嘉明，梁春美，王冬琦
（东北大学 软件学院，辽宁 沈阳 110819）

摘 要：在异质图数据收集中，由于隐私保护政策或版权限制，节点属性缺失现象普遍存在．针对属性不完备和属性完全缺失两种情况，提出了一种基于属性补全的异质图表示学习算法（HGAC）．对于属性不完备的节点，通过构建属性空间的邻接矩阵并执行图卷积来获取缺失的属性；将属性视为抽象节点，在元路径的引导下，对学习节点和属性进行拓扑嵌入，利用拓扑嵌入间的相似性来补全完全缺失的属性．在3个真实数据集上进行实验，结果表明，该算法有效提升了下游任务的性能，并具有较强的泛化能力．

关 键 词：图表示学习；异质图；属性缺失；属性补全；元路径
中图分类号：TP 391 **文献标志码**：A **文章编号**：1005－3026（2025）09－0025－09

Heterogeneous Graph Representation Learning Algorithm Based on Attribute Completion

CHEN Dong-ming，*LIU Jia-ming*，*LIANG Chun-mei*，*WANG Dong-qi*
（Software College，Northeastern University，Shenyang 110819，China. Corresponding author：LIU Jia-ming，E-mail：2271401@stu.neu.edu.cn）

Abstract：In the process of collecting heterogeneous graph data，node attributes are often missing due to privacy protection policies or copyright constraints. Regarding both incomplete attributes and completely missing attributes，a heterogeneous graph representation learning algorithm based on attribute completion（HGAC）was proposed. For nodes with incomplete attributes，the missing attributes were obtained by constructing an adjacency matrix in the attribute space and performing graph convolution. Subsequently，the attributes were regarded as abstract nodes，and under the guidance of meta-paths，the topological embeddings of both nodes and attributes were learned. The similarity among the topological embeddings were then used to complete completely missing attributes. Experiments conducted on three real datasets demonstrate that the proposed algorithm effectively enhances the performance of downstream tasks and possesses strong generalization capability.

Key words：graph representation learning；heterogeneous graph；attribute missing；attribute completion；meta-path

在现实世界中，复杂的系统通常可以建模为图数据结构的形式，根据图中节点和连边的类型，可以将图分为同质图和异质图[1]．同质图中仅有一种类型的节点和连边，其信息传递是在同种类型的节点之间进行的；异质图在结构和语义上更为复杂，可以有多种类型的节点和连边，允许不同类型的节点有不同的属性，从而能够完整、自然地对现实世界的网络数据进行建模[2]．

在图中，节点间复杂的结构信息能够反映节点之间的连接关系，属性信息作为对节点特性的补充，能够丰富图结构信息，使得图表示学习能够更全面地理解图数据．现存的多数基于图神经网络的异质图表示学习方法也被解释为由图结构引导的平滑邻居节点属性，因此这些模型对节

收稿日期：2024－07－29
基金项目：辽宁省重点研发计划项目（2024JH2/102400072）；辽宁省应用基础研究计划项目（2023JH2/101300185）；中央高校基本科研业务费专项资金资助项目（2024GFZD03）．
作者简介：陈东明（1971—），男，安徽怀宁人，东北大学教授，博士生导师．

点属性非常敏感.为了学习到良好的节点表示,需要保证输入到模型中的图具有完整的节点属性.然而,在现实数据收集过程中,由于隐私、成本等原因,这一条件往往是无法成立的.例如,在学术网络中,由于版权限制或数据库收录不全,某些论文的详细信息可能无法获取;在社交网络中,出于隐私保护的原因,用户选择不公开自身信息等,这些情况都会导致图中节点属性不完备或完全缺失.对于以上问题,已有研究者采用零填充和均值填充来插补缺失的属性,但这些方法可能会引入许多不相关的信息,从而导致语义偏差.另外,也有许多学者设计出复杂的属性补全模型[3],并将属性补全与图表示学习相结合,取得了显著的效果,但这些模型只适用于属性不完备或属性完全缺失中的一种情况,当图中同时存在这两种情况时,上述模型的应用将会受限.

基于以上讨论,本文提出一种基于属性补全的异质图表示学习算法(HGAC),该算法可以兼顾图中节点属性不完备和完全缺失的情况,充分利用可信赖的观测值(即属性完备的节点及图的拓扑结构信息)来补全未知属性.对于属性完全缺失的目标节点,即使其邻居属性不完备,也可以通过本算法获得高质量的节点表示.在3个真实数据集上进行的节点分类、节点聚类等实验验证结果表明,所提出的算法具有明显优势.

1 相关工作

图中节点属性缺失是数据挖掘中普遍存在的问题,早期的图神经网络(GNN)通常使用零填充和均值填充来插补缺失的属性,如图卷积网络(graph convolutional network,GCN)[4]、异质图注意力网络(heterogeneous attention network,HAN)[5]、元路径聚合异质图神经网络(meta-path aggregated graph neural network,MAGNN)[6]等.其中GCN[4]主要依赖于节点的完整属性信息来进行特征聚合,对缺失属性的处理能力较弱,并且没有考虑异质图中不同节点类型之间的差异以及缺失属性对图卷积过程的影响.HAN[5]主要关注节点的邻居和元路径信息,没有专门的机制来处理节点属性的缺失.MAGNN[6]对属性缺失的处理依赖于简单的平均填充,容易引入噪声,降低模型的鲁棒性和准确性.

随着深度学习的发展,已有学者提出了几种方法来专门处理节点属性缺失的问题.图属性补全嵌入(graph attribute completion embedding,GRAPE)[7]创新性地将属性视为节点,构建了节点和属性之间的关系,利用GNN同时完成了属性补全和标签预测的任务.图卷积网络矩阵分解(graph convolutional network matrix factorization,GCNMF)[8]考虑到传统策略将属性输入和图表示学习分离会降低性能,将属性补全和图表示学习的处理集成在同一个GNN架构中,利用高斯混合模型(GMM)表示缺失属性,将GMM转换为GNN层,并通过端到端的学习过程训练整个系统.结构-属性对齐转换器(structure-attribute alignment transformer,SAT)[9]提出了一个共享潜在空间的假设,认为缺失的属性可以通过拓扑表示来补充.然后使用双编码器分别学习属性和拓扑的表示,并通过对抗性学习对齐成对的潜在表示.迭代拓扑细化(iterative topology refinement,ITR)[10]首先采用属性缺失样本的结构嵌入作为初始化嵌入,再根据亲和结构聚合属性观测样本的可靠嵌入,对初始值进行细化.

上述方法均适用于同质图,异质图的复杂性和多样化的网络结构使得属性补全成为HINs的一大挑战.异质图神经网络-属性补全(heterogeneous graph neural network via attribute completion,HGNN-AC)[3]首先将属性补全应用到异质图中,它为异质图构建了专门的属性补全模型,通过浅层算法获得节点的结构嵌入;之后利用这种结构信息来引导属性信息的补全;最终结合其他GNN模型得到最终的嵌入.但是该方法主要依赖于一阶邻居进行属性聚合,忽略了高阶邻居的潜在贡献,同时未能充分考虑节点间的属性相似性,可能在缺失属性较多时影响其性能.属性补全异质嵌入网络(attribute completion heterogeneous embedding network,AC-HEN)[11]不仅利用了图的结构信息,同时也在特征空间进行邻居聚合,获得多视图嵌入,设计出一个异质图的通用属性补全框架.但其在处理高阶邻居聚合时,仅依赖简单的加权机制,未能充分利用异质图中不同类型节点之间的复杂关系.异质图对比属性补全(heterogeneous graph contrastive attribute completion,HGCA)[12]在设计属性补全方法时考虑了图中节点标签不足的情况,采用对比学习策略来统一无监督异质框架下的属性补全和图表示学习,取得了具有说服力的实验结果.然而,该方法更多依赖于对比学习策略,未能有效利用部分标签信息来进一步提升属性补全的精度.

自动属性补全(automatic attribute completion, AutoAC)[13]将神经架构搜索(neural architecture search, NAS)引入属性补全任务中,提出了一个表达性补全操作的搜索空间和一个可微属性补全框架.异质残差图注意力网络(heterogeneous residual graph attention network, HetReGAT)[14]首先学习节点的拓扑信息以生成节点嵌入,然后利用这些嵌入计算节点间的权重来补全缺失的属性,并设计了异质残差图注意力网络来学习节点的完整信息.但这些方法对全局结构信息的利用不足,尤其对高阶邻居和不同类型节点间的复杂关系没有充分考虑.

为了解决上述问题,本文提出了HGAC算法,通过在属性空间构建邻接矩阵,并聚合相似节点的属性来补全不完备节点的属性,从而有效减少不完备节点对目标节点的负面影响.为进一步提升属性补全的准确性,HGAC结合图的拓扑结构与元路径,设计了带属性的随机游走算法,不仅捕捉高阶语义关系,还融合高阶邻居信息,使补全属性包含丰富的全局结构信息.针对图中同时存在属性不完备和完全缺失的情况,HGAC创新性地结合属性空间和结构信息进行双重聚合,采用多头注意力机制动态调整邻居的贡献,从而保证了在复杂异质图中的鲁棒性和优异表现.

2　问题定义

给定一个图 $G=(V,E,T,R,X)$,其中 V 代表图中节点的集合,E 代表图中连边的集合,T 代表图中节点类型的集合,R 代表图中连边类型的集合,X 代表节点的属性集合.对于节点来说,有一个节点类型的映射函数 $\phi: V\to T$,将每个节点映射到一种节点类型;同样地,对于连边来说,有一个连边类型的映射函数 $\psi: E\to R$,将每条连边映射到一种连边类型.当 $|T|+|R|>2$ 时,称图 G 为异质图.

定义 1　属性缺失的图:给定一个图 $G=(V,E,T,R)$ 以及对应的邻接矩阵 $\boldsymbol{A}$ 和节点属性集合 X,存在节点集合 $V^{\mathrm{M}}\subset V$,其中的节点属性完全缺失,对应的属性集合表示为 X^{M},同时,存在节点集合 $V^{\mathrm{I}}\subset V$,其中的节点属性不完备,对应的属性集合表示为 X^{I},并以掩码矩阵 $\boldsymbol{I}$ 来记录 V^{I} 中的属性缺失情况.当 $\boldsymbol{I}[i][j]=1$ 时,代表节点 i 的第 j 个属性缺失;反之,当 $\boldsymbol{I}[i][j]=0$ 时,代表节点 i 的第 j 个属性存在.将除上述两种节点之外的节点称为属性完整的节点,用 V^{O} 表示,对应的属性集合为 X^{O}.对于图 G,若满足 $V=V^{\mathrm{O}}\cup V^{\mathrm{I}}\cup V^{\mathrm{M}}$,$X=X^{\mathrm{O}}\cup X^{\mathrm{I}}\cup X^{\mathrm{M}}$,则称图 G 为属性缺失的图.为了便于后续论述,将属性不完备的节点简称为不完备节点,将属性完全缺失的节点简称为缺失节点,将属性完整的节点简称为完整节点.

定义 2　属性补全:对于一个包含属性缺失或属性不完备节点的图,属性补全旨在利用图中已知的属性和图的拓扑结构等可用信息,完成对节点缺失属性的补全,使补全后的属性能够输入到后续的图神经网络中,从而获得有利于下游任务的节点表示.

定义 3　异质图表示学习:给定一个异质图 $G=(V,E,T,R,X)$,异质图表示学习旨在通过某个映射函数,利用图的结构及节点属性等信息,将异质图中的每个节点 $v\in V$ 映射到一个低维空间,学习得到一个低维稠密向量 $\boldsymbol{h}_v=F_\theta(G,v)$,该向量维度远小于图中节点的总数.

3　算法设计与实现

图 1 为本文提出的HGAC算法的整体框架,由不完备节点的属性补全和缺失节点的属性补全两部分组成.

1) 不完备节点的属性补全.由于图中同时存在属性不完备节点和属性完全缺失节点,因此在对目标节点进行属性补全时,很容易受到不完备节点的影响,尤其是当不完备节点与目标节点直接相连时,其缺失的属性会错误地传递给目标节点,导致补全结果产生偏差.因此,在对目标节点进行属性补全之前,需要对图中不完备节点进行处理.为了充分利用已观测到的属性信息,本文算法将属性完整的节点作为源节点来预测不完备节点缺失的属性.通过构造属性空间的邻接矩阵,并利用该邻接矩阵进行属性聚合来完成对不完备节点的属性补全.

步骤 1　由于图中包含大量具有完整属性的节点,无法使用所有节点作为源节点,因此需要设计一种策略从中选出最适合的源节点.本文借鉴 K 最近邻(KNN)算法的思想,对于每个属性不完备的节点 $v\in V^{\mathrm{I}}$(即属性部分缺失的节点),通过计算其与属性完整节点 $u\in V^{\mathrm{O}}$ 的相似度,选取最为相

似的一组完整节点作为源节点.相似度的计算公式为

$$s_{v,u}=\cos\left(\boldsymbol{x}_v,\boldsymbol{x}_u\right)=\frac{\boldsymbol{x}_v\cdot\boldsymbol{x}_u}{\left|\boldsymbol{x}_v\right|\left|\boldsymbol{x}_u\right|}. \tag{1}$$

其中:$\boldsymbol{x}_v$代表不完备节点v的现有属性向量;$\boldsymbol{x}_u$代表完整节点u的属性向量.

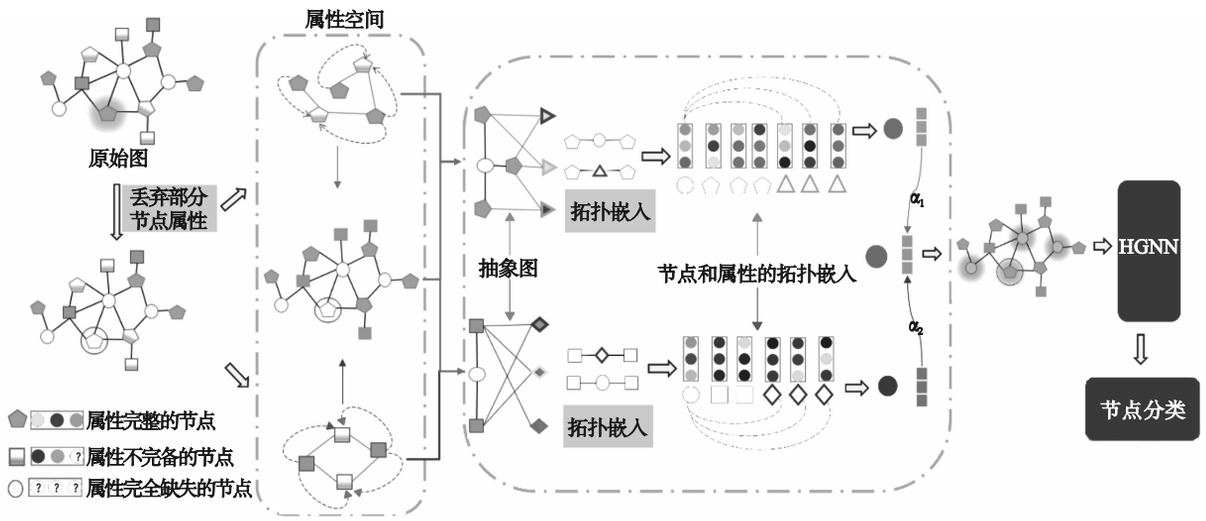

图1 HGAC整体框架

Fig. 1 Overall architecture of HGAC

在计算了不完备节点与完整节点的相似度后,为每个不完备节点选择前K个完整节点作为源节点,并为这些节点建立连接,从而构建属性空间的邻接矩阵$\boldsymbol{A}_{\mathrm{a}}$.在该邻接矩阵中,不完备节点与其所选源节点之间的值$\boldsymbol{A}_{\mathrm{a}}[v][u]$如式(2)所示.

$$\boldsymbol{A}_{\mathrm{a}}[v][u]=\begin{cases}1, & u\in S_v;\\ 0, & \text{其他}.\end{cases} \tag{2}$$

其中,S_v是不完备节点v的源节点集合.

以本文使用的3个数据集为例,在ACM数据集中,不完备的作者节点通过计算其与完整论文节点之间的余弦相似度,选取相似度最高的K个论文节点作为源节点,构建属性空间的邻接矩阵;在IMDB数据集中,不完备的导演和演员节点通过与其对应的完整电影节点进行相似度计算,选取最相似的电影节点作为源节点;在DBLP数据集中,不完备的术语和作者节点则通过与完整的论文节点进行相似度计算,选取最相似的论文节点作为源节点.通过这些构建后的属性邻接矩阵,不完备节点与源节点之间建立连接,再通过属性聚合过程可以有效补全缺失的属性,从而提升节点嵌入表示的质量.

步骤2 利用GCN[4]执行属性信息聚合,得到节点v补全后的属性:

$$\boldsymbol{x}_v^{\mathrm{IC}}=\mathrm{GCN}\left(\boldsymbol{A}_{\mathrm{a}},\boldsymbol{X}\right)=\sigma\left(\boldsymbol{D}_{\mathrm{a}}^{-\frac{1}{2}}\boldsymbol{A}_{\mathrm{a}}\boldsymbol{D}_{\mathrm{a}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}\right). \tag{3}$$

其中:$\boldsymbol{D}_{\mathrm{a}}$为$\boldsymbol{A}_{\mathrm{a}}$的度矩阵;$\boldsymbol{X}$为节点的属性矩阵;$\boldsymbol{W}$为可训练的权重矩阵;$\boldsymbol{x}_v^{\mathrm{IC}}$即为节点$v$补全后的属性.值得注意的是,由于不完备节点缺失部分属性,所以在使用GCN进行聚合时,没有将节点自身的属性信息纳入聚合范围,即$\boldsymbol{A}_{\mathrm{a}}$没有添加自环.在后续计算中,将不完备节点的属性设置为

$$\boldsymbol{x}_{vj}=\begin{cases}\boldsymbol{x}_v, & \boldsymbol{I}[v][j]=1;\\ \boldsymbol{x}_v^{\mathrm{IC}}, & \boldsymbol{I}[v][j]=0.\end{cases} \tag{4}$$

即对于节点现有的属性,使用原始值;而对于缺失的属性,使用补全后的值.

2) 缺失节点的属性补全.在得到不完备节点的属性后,可以根据其属性值来构建抽象图,完成对缺失节点的属性补全.

步骤1 在对缺失节点属性进行补全的过程中,受GRAPE[7]启发,本文算法将属性也视为一种节点,根据属性值添加原始节点和属性节点间的连边,并将属性值作为连边的权重,构成带有属性节点的抽象图.但与GRAPE不同的是,本文的研究对象为异质图,其本身就含有丰富的节点和语义信息,无法直接使用GRAPE中改造的GraphSAGE来学习节点嵌入和属性值.由于在上一步已经将不完备属性中的缺失值进行了补全,因此原始节点与属性节点之间的连边是准确的.基于这一观察,考虑采用带有属性的随机游走方式,捕获原始节点和属性节点以及原始图中各节点间的高阶关系.

步骤2 在异质图中,元路径是一种捕获节点间高阶关系的工具.Metapath2vec[15]通过基于

元路径的随机游走学习节点嵌入,以捕获异质图中丰富的语义关系,并为下游任务提供低维、稠密且语义丰富的节点表示.因此,本文在Metapath2vec的基础上,通过将属性值加入到游走概率中,来获取原始节点与属性节点的拓扑嵌入.在执行随机游走之前,首先根据节点类型的不同将图划分为多个带有属性节点的抽象子图$\{g_1,g_2,\cdots,g_n\}$,之后按照设计的元路径,在子图上执行带有属性的随机游走,给定元路径$P=t_1\xrightarrow{R_1}t_2\xrightarrow{R_2}\cdots\xrightarrow{R_{l-1}}t_l$,由节点$v_i$游走到下一个节点的概率为

$$p\left(v_{i+1}|v_i,P\right)=\begin{cases}\dfrac{w_{i,i+1}}{\left|N_{t_{i+1}}\left(v_i\right)\right|}, & A_{i,i+1}=1,\ \varphi\left(v_{i+1}\right)=t_{i+1};\\ 0, & \text{其他}.\end{cases}\tag{5}$$

式中:$A_{i,i+1}$表示相邻节点间的连边数;$N_{t_{i+1}}\left(v_i\right)$为节点$v_i$的$t_{i+1}$类型的邻居,即只有下一个节点满足元路径的规则$\varphi$时,才有可能被访问;$w_{i,i+1}$为节点$v_i$和$v_{i+1}$间的权重,在原始节点与属性节点间,该值为节点的属性值,在两个原始节点间,该值为1.在获取到游走序列后,将序列输入到异质Skip-Gram中,通过最大化随机游走捕获的局部相邻节点结构的概率来学习原始节点和属性节点的拓扑嵌入.目标函数可以表述为

$$\arg\max\sum_{v\in V}\sum_{t\in T}\sum_{u\in N_{t(v)}}\lg p\left(u|v;\theta\right),\tag{6}$$

$$p\left(u|v;\theta\right)=\frac{\exp\left(\boldsymbol{z}_u\cdot\boldsymbol{z}_v^{\mathrm{T}}\right)}{\sum_{o\in V}\exp\left(\boldsymbol{z}_o\cdot\boldsymbol{z}_v^{\mathrm{T}}\right)}.\tag{7}$$

式中:θ为模型参数;$\boldsymbol{z}$为节点的表示向量.

步骤3　由于在随机游走过程中结合了原始节点和属性节点,并且在游走概率中加入了两种节点间的属性值,所以获得的嵌入包含了原始节点和这些属性之间的相似关系,则可以将节点嵌入和属性嵌入之间的相似值作为节点的属性值.

对于节点$v\in V^{\mathrm{M}}$在子图g上的第j个属性值,其计算公式为

$$\boldsymbol{x}_j^{\mathrm{MC}}=\boldsymbol{z}_v\boldsymbol{W}_1\left(\boldsymbol{z}_j^{\mathrm{attr}}\right)^{\mathrm{T}}.\tag{8}$$

给定所有节点和属性的嵌入$\boldsymbol{Z}_{\mathrm{node}}$和$\boldsymbol{Z}_{\mathrm{attr}}$,补全后的属性可以表示为

$$X_g^{\mathrm{MC}}=\boldsymbol{Z}_{\mathrm{node}}\boldsymbol{W}_1\boldsymbol{Z}_{\mathrm{attr}}^{\mathrm{T}}.\tag{9}$$

不同类型的节点对目标节点有不同的贡献,因此需要一个聚合机制将补全得到的属性进行聚合.由于注意力机制能够自动学习并赋予不同类型节点以不同的权重,本文使用注意力机制聚合不同类型子图下的属性值.

为学习每个子图的权重,首先使用多层感知机(MLP)对属性进行非线性变换,通过注意力向量$\boldsymbol{q}$衡量多种属性间的相似性,得到各子图的重要性系数:

$$w_g=\frac{1}{\left|V^{\mathrm{M}}\right|}\sum_{v\in V^{\mathrm{M}}}\boldsymbol{q}^{\mathrm{T}}\cdot\tanh\left(\boldsymbol{W}_2\cdot X_g^{\mathrm{MC}}+\boldsymbol{b}\right).\tag{10}$$

式中$\boldsymbol{b}$为偏置向量.之后使用Softmax函数对重要性系数进行归一化,得到每个子图的权重:

$$\alpha_g=\frac{\exp\left(w_g\right)}{\sum_{g=1}^{\mathrm{n}}\exp\left(w_g\right)}.\tag{11}$$

最后,使用头数为H的多头注意力机制对不同子图的属性进行加权求和,得到补全后的属性:

$$X^{\mathrm{MC}}=\mathrm{mean}\left(\sum_{h}^{H}\sum_{g=1}^{n}\alpha_g^{(h)}\cdot X_g^{\mathrm{MC}}\right).\tag{12}$$

经过上述两阶段的补全后,图G的属性集可以表示为

$$X^{\mathrm{new}}=\left\{\boldsymbol{x}_u^{\mathrm{o}},\boldsymbol{x}_v,\boldsymbol{x}_w^{\mathrm{MC}}|\forall u\in V^{\mathrm{O}},\forall v\in V^{\mathrm{I}},\forall w\in V^{\mathrm{M}}\right\}.\tag{13}$$

因此,可以将完整的属性集X^{new}和图拓扑结构共同输入异质图神经网络中,得到节点的最终嵌入$\boldsymbol{Z}$:

$$\boldsymbol{Z}=\mathrm{HGNN}\left(G,X^{\mathrm{new}}\right).\tag{14}$$

式中,HGNN可以为任何一个异质图神经网络模型,在本实验中采用MAGNN.

为约束节点的属性补全过程,确保补全后的属性可以提高异质图神经网络模型的性能,本文首先删除了部分完整节点的所有属性,并在属性补全过程中重建这些属性.通过计算被删除属性和重建属性之间的补全损失L_{MC},使属性补全过程具有指导性和可学习性:

$$V_{\mathrm{drop}}^{\mathrm{O}}=\alpha V^{\mathrm{O}},\tag{15}$$

$$L_{\mathrm{MC}}=\frac{1}{\left|V_{\mathrm{drop}}^{\mathrm{O}}\right|}\sum_{v\in V_{\mathrm{drop}}^{\mathrm{O}}}\sqrt{\left(\boldsymbol{x}_v^{\mathrm{O}}-\boldsymbol{x}_v^{\mathrm{MC}}\right)^2}.\tag{16}$$

式中,α为属性丢弃比例.对于下游任务,如节点分类,使用交叉熵损失函数优化新算法中的参数:

$$L_{\mathrm{prediction}}=-\sum_{l\in Y_l}Y_{v_l}\cdot\ln\left(C\cdot\boldsymbol{z}_{v_l}\right).\tag{17}$$

其中:Y_{v_l}和$\boldsymbol{z}_{v_l}$分别是节点v_l的标签和表示向量;C是分类器参数.

结合两阶段属性补全和下游任务,算法整体的损失函数为

$$L=\lambda L_{\mathrm{MC}}+L_{\mathrm{prediction}}.\tag{18}$$

其中,λ为平衡这两部分损失的系数.

4 实验与分析

4.1 实验数据集及设置

本文使用在异质图神经网络研究中被广泛应用的ACM,IMDB和DBLP 3个真实数据集进行实验评估,并将实验结果与近年来影响广泛的方法进行对比分析.表1为在3个数据集上的评估信息.

本文选择近几年具有代表性的图表示学习算法作为对比基线以验证所提算法的有效性.

表1 数据集统计信息

Table 1 Dataset statistics

数据集	评价指标	Metapath2vec	GCN	HAN	MAGNN	AC-HEN	HGNN-AC	HGAC
ACM	NMI	21.22	51.40	61.37	63.17	64.82	64.54	**65.22**
	ARI	21.00	53.01	64.39	67.41	68.84	68.26	**69.75**
IMDB	NMI	0.89	7.46	10.62	10.39	11.94	13.02	**13.29**
	ARI	0.22	7.69	10.01	11.11	12.17	13.43	**13.87**
DBLP	NMI	74.23	73.45	77.49	79.64	79.51	79.47	**80.69**
	ARI	78.11	77.50	82.95	82.80	84.51	84.66	**84.95**

注:加粗表示各项指标中的最优结果,下同.

Metapath2vec[15]:一种浅层的异质图嵌入算法,通过基于元路径的随机游走并结合Skip-Gram来生成嵌入.本文在所有可能的元路径上进行了测试,并记录最佳结果.

GCN[4]:一个适用于同质图的图卷积网络,可以同时利用图的结构信息和节点属性信息.本文测试了其在基于不同元路径生成的同质子图上的效果,并记录最佳结果.

HAN[5]:一个使用注意力机制的异质图神经网络,分别通过节点级和语义级注意力来聚合不同邻居和不同元路径的信息.

MAGNN[6]:一个异质图神经网络模型,综合考虑了节点的内容特征、元路径内部信息以及多个元路径间的关系.

HGNN-AC[3]:该模型在拓扑结构的引导下,通过注意力机制完成对一阶邻居节点属性的聚合,实现对属性完全缺失节点的补全.

AC-HEN[11]:该模型分别使用属性聚合和结构聚合来获得属性视图和结构视图的嵌入,以弱监督学习的方式完成对不完备节点属性的补全.

对于所有基线算法中使用的数据集,均保留其在原论文中的设置,即对于GCN,HAN和MAGNN,保留数据集中所有节点的属性;对于AC-HEN,在保留数据集中所有节点属性的同时,选择ACM和DBLP数据集中20%的作者节点,随机丢弃这些节点30%的属性,以及选择IMDB数据集中20%的电影节点,同样随机丢弃这些节点30%的属性.对于HGNN-AC和本文所提算法HGAC,仅保留ACM和DBLP数据集中的论文节点以及IMDB数据集中的电影节点的原始属性,同时从每个数据集保留属性的节点中选择20%,随机丢弃这些节点30%的属性.

对于基线算法中的参数,设置均为其在原论文中的值.对于本文提出的HGAC,将Metapath2vec中随机游走的窗口大小设为5、步长设为100、嵌入维数设为64、K设为4、注意力头数设为8、注意力向量维度设为256.使用Adam作为优化器,学习率为0.001、权重衰减设为0.001.将epoch参数设为200、batch_size设为128、patience设为100、平衡参数λ_1和λ_2均设为0.4、属性丢失比例α设为0.3.

4.2 实验结果与分析

4.2.1 节点分类

节点分类是一种有效评估节点表示质量的方式,本文分别对ACM,IMDB和DBLP数据集中的论文、电影、作者节点进行了多分类实验.在训练前,将数据集中有标签的节点按10%,10%,80%的比例划分为训练集、验证集和测试集,之后将训练集输入到本文提出的模型中,通过损失函数不断优化模型,并利用验证集调整模型中的超参数以提高节点表示的质量.最后,将测试集的数据分别以20%,40%,60%和80%的训练比例输入线性支持向量机(support vector machine,SVM)分类器中.对于分类结果,本文采用Micro-F1和Macro-F1进行评估,结果如表2所示.

本文提出的HGAC算法在ACM,IMDB和DBLP数据集上均表现出了最佳效果.这表明HGAC算法在不同训练集比例下都能取得最佳的分类性能,尤其是在训练集占比仅为20%的情况下,其Macro-F1和Micro-F1值与训练集占比

80%时的结果相差不大,这体现了HGAC算法在小样本场景下依然具有良好的泛化能力.相比之下,其他传统算法如Metapath2vec,GCN和HAN等,在属性不完备或完全缺失时性能下降明显,特别是在IMDB数据集上的表现较为有限.这些算法通常依赖节点的完整属性信息,而在实际数据中,由于隐私、版权等问题,节点属性往往不完备或缺失,导致模型难以学习到有效的节点表示.

表2　节点分类实验结果对比

Table 2　Comparison of experimental results of node classification　%

数据集	评价指标	训练比例	Metapath2vec	GCN	HAN	MAGNN	AC-HEN	HGNN-AC	HGAC
ACM	Macro-F1	20	69.43	72.31	90.01	88.01	90.09	90.87	**91.49**
		40	70.21	73.83	90.82	89.42	90.55	90.97	**91.84**
		60	70.57	72.59	91.51	90.39	91.22	91.13	**92. 21**
		80	71.74	72.23	91.71	90.79	91.39	91.52	**92.77**
	Micro-F1	20	71.13	75.27	89.89	88.08	90.43	90.75	**91.63**
		40	72.10	76.21	90.73	89.48	90.79	91.02	**92.16**
		60	72.85	76.10	91.37	90.42	91.23	91.61	**92.45**
		80	73.24	75.91	91.56	90.80	91.41	92.15	**93.05**
IMDB	Macro-F1	20	46.42	44.75	57.21	58.11	58.67	58.63	**59.53**
		40	47.70	45.26	57.54	59.39	59.11	59.39	**59.77**
		60	48.25	47.70	57.93	59.97	60.17	60.02	**60.28**
		80	48.73	48.25	58.16	60.02	60.34	60.57	**60.84**
	Micro-F1	20	48.08	47.44	57.24	58.16	58.66	58.72	**59.69**
		40	49.55	47.62	57.48	59.46	58.99	59.42	**60.11**
		60	50.06	48.49	57.68	60.05	59.87	59.84	**60.71**
		80	50.68	48.73	57.97	60.15	60.03	60.32	**61.12**
DBLP	Macro-F1	20	90.12	90.06	91.44	92.23	92.29	92.13	**93.18**
		40	90.74	90.37	91.62	92.87	92.79	92.74	**93.73**
		60	91.32	90.57	92.01	93.02	93.15	93.27	**94.05**
		80	91.68	90.74	92.15	93.18	93.22	93.37	**94.47**
	Micro-F1	20	91.25	90.53	91.89	92.35	92.43	92.58	**92.94**
		40	91.79	90.83	92.32	93.05	92.96	93.11	**93.56**
		60	92.33	91.01	92.59	93.38	93.28	93.52	**93.98**
		80	92.70	91.15	92.68	93.47	93.39	93.58	**94.39**

4.2.2　节点聚类

本文使用K均值聚类(K-means clustering,K-means)算法对3个数据集进行了节点聚类实验.与节点分类类似,首先利用异质图表示学习算法学习到节点的表示向量,然后将该表示向量输入到K-means算法中.将K-means算法中的簇数设置为数据集中聚类节点的类别数,最后以归一化互信息(NMI)和调整兰德指数(ARI)作为评价指标来衡量聚类效果,最终结果如表3所示.结果表明,HGAC同样优于其他对比算法.与未加入属性补全的算法相比,HGAC通过属性补全得到的节点表示向量更能准确地捕捉节点之间的相似性和差异性,从而在NMI和ARI等评价指标上均取得了最高分数.

传统算法如GCN,HAN和MAGNN,由于未考虑属性缺失问题,导致其在聚类任务中的表现明显落后.虽然AC-HEN和HGNN-AC等算法也引入了属性补全机制,但它们要么仅对一阶邻居进行属性补全,要么只针对某一类节点的属性不完备情况,这使得其对复杂网络的适应能力较弱.而HGAC通过结合高阶拓扑结构信息和元路径,能够更全面地补全不同类型节点的属性,因此在聚类任务中展现了更强的鲁棒性和竞争力.

4.2.3　参数分析

通过在ACM数据集上进行节点分类实验,研究关键实验参数对算法性能的影响,包括平衡系数λ、属性丢弃比例α、隐藏嵌入维度d、特征空间中最近邻节点数k.其中将测试集数据以40%的训练比例输入到线性支持向量机分类器中.同

时为保证实验的公平性,除测试变量外,其他参数均保持不变,结果如图2所示.

表3 节点聚类实验结果对比

Table 3 Comparison of experimental results of node clustering %

数据集	评价指标	Metapath2vec	GCN	HAN	MAGNN	AC-HEN	HGNN-AC	HGAC
ACM	NMI	21.22	51.40	61.37	63.17	64.82	64.54	**65.22**
	ARI	21.00	53.01	64.39	67.41	68.84	68.26	**69.75**
IMDB	NMI	0.89	7.46	10.62	10.39	11.94	13.02	**13.29**
	ARI	0.22	7.69	10.01	11.11	12.17	13.43	**13.87**
DBLP	NMI	74.23	73.45	77.49	79.64	79.51	79.47	**80.69**
	ARI	78.11	77.50	82.95	82.80	84.51	84.66	**84.95**

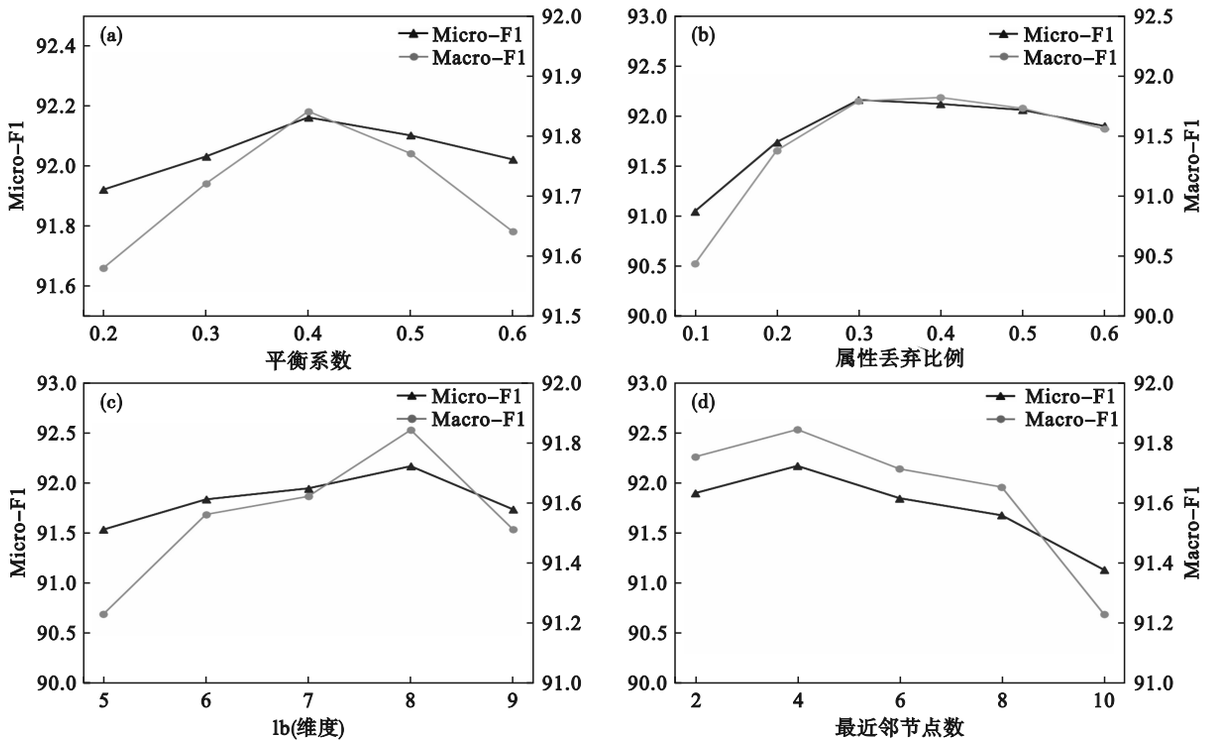

图2 不同参数对节点分类的影响

Fig. 2 Impact of different parameters on node classification

(a)—平衡系数;(b)—属性丢弃比例;(c)—维度;(d)—最近邻节点数.

从图2可以看出,平衡系数λ的最佳值为0.4,随着该系数的增大,算法性能呈现先上升后下降的趋势.这是由于当系数过小时,会削弱补全损失在属性补全过程中的指导作用;而当系数过大时,则会削弱下游任务损失的作用.

对于属性丢弃比例α,其最佳取值为0.3,随着属性丢弃比例的增大,算法性能呈现先上升后缓慢下降的趋势,这表明HGAC需要一个合适的属性丢弃比例.丢弃过多属性将导致用于属性补全的节点过少,从而使得补全结果不准确,而丢弃过少属性将导致用于计算损失的节点过少,无法有效优化算法.

随着嵌入维度的增加,模型的性能呈现先上升后下降的趋势.当维度较小时,模型无法充分捕捉节点的所有特征信息,导致性能不佳;但当维度过大时,冗余信息可能会干扰模型性能,尤其在512维度处性能明显下降.最佳的嵌入维度出现在256附近,此时模型的分类性能达到峰值.

随着最近邻节点数的增加,模型性能呈现先上升后下降的趋势.当$k=4$时,模型分类性能达到最佳,此时能够有效利用相似节点进行属性补全.过大的k值会引入较多噪声信息,导致性能下降.

4.2.4 时空复杂度分析

HGAC算法分为3步:第一步为不完备节点的属性补全,计算相似度并进行聚合,其时间复杂度为$O(N_{\mathrm{I}}\cdot N_{\mathrm{O}}\cdot d+k\cdot N_{\mathrm{I}}\cdot d^2)$,其中$N_I$是不完备节点数,$N_{\mathrm{O}}$是完整节点数,$k$是聚合的近邻数,$d$是节点特征维度;第二步为不完备节点的属性补

全,基于元路径的随机游走和嵌入计算,其时间复杂度为$O(W\cdot L\cdot d_{\mathrm{m}})$,其中$W$是随机游走次数,$L$为游走步长,$d_{\mathrm{m}}$是节点的平均度数;第三步为多头注意力聚合,时间复杂度为$O(H\cdot N\cdot d^2)$,其中H为注意力头数,N为节点总数,d为嵌入维度.空间复杂度主要考虑属性邻接矩阵和嵌入向量的存储开销,为$O(N_{\mathrm{I}}\cdot k+N\cdot d)$,加上GCN和多头注意力机制的参数,总复杂度为$O(L\cdot d^2+H\cdot d^2)$.

HGAC算法虽然较为复杂,但通过优化属性补全和多头注意力机制,能够更好地处理不完备节点和完全缺失节点的情况.其时间复杂度与AC-HEN相近,但在处理属性缺失方面更具优势,在处理该类问题时表现出更强的表达能力和鲁棒性.

5　结　　语

本文提出了一种新颖的基于属性补全的异质图表示学习算法(HGAC).该算法通过构建属性空间的邻接矩阵及利用原始节点与属性节点的拓扑嵌入,有效应对节点属性不完备及完全缺失的挑战,确保即便在邻居节点属性不完备的情况下,目标节点也能获得高质量的表示.本文算法不仅在处理现有模型敏感的节点属性问题时显示出明显优势,还通过利用可靠的观测值和图的拓扑结构信息来补全未知属性,进一步增强了其应用范围和效果.在3个真实数据集上进行的节点分类与聚类实验结果均验证了该算法的显著优越性,表现出较强的泛化能力.HGAC有望在更广泛的实际应用场景中发挥更大的潜力,并为异质图表示学习领域的研究提供新的思路与方向.

尽管HGAC算法在节点属性补全和表示学习任务中展现了显著优势,但仍存在进一步优化和拓展的空间.未来的研究可以针对大规模图数据,探索更加高效的图神经网络架构,以降低计算复杂度和时间成本,从而应对实际应用中的规模化挑战.此外,融合多模态数据(如文本、图像等非结构化信息)将有助于增强模型的表达能力和适用性.与此同时,在动态异质图建模中的应用也是未来值得深入探索的重要方向.

参考文献:

[1] Chen F, Wang Y C, Wang B, et al. Graph representation learning: a survey[J]. *APSIPA Transactions on Signal and Information Processing*, 2020, 9: e15.

[2] 周丽华, 王家龙, 王丽珍, 等. 异质信息网络表征学习综述[J]. 计算机学报, 2022, 45(1): 160-189.
(Zhou Li-hua, Wang Jia-long, Wang Li-zhen, et al. Heterogeneous information network representation learning: a survey [J]. *Chinese Journal of Computers*, 2022, 45(1): 160-189.)

[3] Jin D, Huo C Y, Liang C D, et al. Heterogeneous graph neural network via attribute completion[C]//Proceedings of the Web Conference 2021. Ljubljana, 2021: 391-400.

[4] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]//International Conference on Learning Representations. Toulon, 2017: 1-14.

[5] Wang X, Ji H Y, Shi C, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference. San Francisco, 2019: 2022-2032.

[6] Fu X, Zhang J, Meng Z, et al. MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding[C]//Proceedings of the Web Conference 2020. Taipei, 2020: 2331-2341.

[7] You J, Ma X, Ding Y, et al. Handling missing data with graph representation learning [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 19075-19087.

[8] Taguchi H, Liu X, Murata T. Graph convolutional networks for graphs containing missing features[J]. *Future Generation Computer Systems*, 2021, 117: 155-168.

[9] Chen X, Chen S, Yao J, et al. Learning on attribute-missing graphs[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 44(2): 740-757.

[10] Tu W X, Zhou S H, Liu X W, et al. Initializing then refining: a simple graph attribute imputation network[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Vienna, 2022: 3494-3500.

[11] Wang K, Yu Y, Huang C, et al. Heterogeneous graph neural network for attribute completion [J]. *Knowledge-Based Systems*, 2022, 251: 109171.

[12] He D X, Liang C D, Huo C Y, et al. Analyzing heterogeneous networks with missing attributes by unsupervised contrastive learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(4): 4438-4450.

[13] Zhu G, Zhu Z, Wang W, et al. Autoac: towards automated attribute completion for heterogeneous graph neural network [C]//2023 IEEE 39th International Conference on Data Engineering (ICDE). Anaheim, 2023: 2808-2821.

[14] Li C, Yan Y Y, Fu J H, et al. HetReGAT-FC: heterogeneous residual graph attention network via feature completion[J]. *Information Sciences*, 2023, 632: 424-438.

[15] Dong Y X, Chawla N V, Swami A. Metapath2vec: scalable representation learning for heterogeneous networks [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, 2017: 135-144.