计算机研究与发展
*Journal of Computer Research and Development*
ISSN 1000-1239,CN 11-1777/TP

# 《计算机研究与发展》网络首发论文

题目：基于图卷积的异质网络节点分类方法
作者：谢小杰，梁英，王梓森，刘政君
收稿日期：2021-02-05
网络首发日期：2021-10-18
引用格式：谢小杰，梁英，王梓森，刘政君. 基于图卷积的异质网络节点分类方法[J/OL]. 计算机研究与发展.
https://kns.cnki.net/kcms/detail/11.1777.TP.20211015.1818.002.html

计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20210124

# 基于图卷积的异质网络节点分类方法

谢小杰[1,2,3] 梁英[1,3] 王梓森[1,2,3] 刘政君[1,2,3]

[1]（中国科学院计算技术研究所泛在计算系统研究中心 北京 100190）

[2]（中国科学院大学计算机科学与技术学院 北京 100049）

[3]（移动计算与新型终端北京市重点实验室（中国科学院计算技术研究所） 北京 100190）

（mailbox_of_xxj@126.com）

# Heterogeneous Network Node Classification Method Based on Graph Convolution

Xie Xiaojie[1,2,3], Liang Ying[1,3], Wang Zisen[1,2,3], and Liu Zhengjun[1,2,3]

[1] （*Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190）

[2] （*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 100049）

[3]（*Beijing Key Laboratory of Mobile Computing and New Devices*（*Institute of Computing Technology, Chinese Academy of Sciences*）, *Beijing* 100190）

**Abstract** Graph neural networks can effectively learn network semantic information and have achieved good performance on node classification tasks, but still facing challenges: how to make the best of rich heterogeneous semantic information and comprehensive structural information to make node classification more accurate. To resolve the above challenges, based on the graph convolution operation,HNCF(heterogeneous network node classification framework) is proposed to solve the node classification task in heterogeneous networks, including two steps of heterogeneous network reduction and graph convolution node classification. Firstly, through the designed heterogeneous network reduction rules, HNCF simplifies a heterogeneous network into a semantic homogeneous network and retains semantic information of the heterogeneous network through relation representations between nodes, reducing the complexity of network structure modeling. Then, based on the message passing framework, a graph convolution node classification method is designed to learn network structure information on the semantic homogeneous network, such as neighbor weights without 1-sum constraint, to discover the differences of relations and neighbor semantic extraction. Finally, heterogeneous node representations are generated and used to classify nodes to identify node category labels. Experiments on three public node classification datasets show that HNCF can make the best of heterogeneous semantic information and effectively learn network structure information such as reasonable neighbor weights to improve the performance of heterogeneous network node classification.

**Key words** heterogeneous network; graph neural network; node classification; semantic relation; neighbor weight

**摘要** 图神经网络能够有效学习网络语义信息，在节点分类任务上取得了良好的效果．但仍面临挑战：如何充分利用异质网络丰富语义信息和全面结构信息使节点分类更精准．针对上述问题，提出了一种基于图卷积的异质网络节点分类框架(heterogeneous node classification framework, HNCF)，包括异质网络约简和图卷积节点分类，解决异质网络节点分类问题．通过设计转换规则约简异质网络，将异质网络化简为语义化同质网络，利用节点间的关系表示保留异质网络多语义信息，降低网络结构建模复杂度；基于消息传递框架设计图卷积节

**收稿日期**：2021-02-05；**修回日期**：2021-08-30
**基金项目**：国家重点研发计划项目 (2018YFB1004700)
This work was supported by the the National Key Research and Development Program of China(2018YFB1004700).
**通信作者**：梁英（liangy@ict.ac.cn）

点分类方法，在语义化同质网络上学习无 1-sum 约束的邻居权重等网络结构信息，深入挖掘关系语义特征，发现不同连接关系和邻居语义提取的差异性，生成节点的异质语义表示用于节点分类，识别节点类别标签．在 3 个公开的节点分类数据集上进行了实验，结果表明 HNCF 能够充分利用异质网络多种语义信息，有效学习邻居节点权重等网络结构信息，提升节点分类效果．

**关键词** 异质网络；图神经网络；节点分类；语义关系；邻居权重

**中图法分类号** TP391

近年来，图神经网络受到了研究者的广泛关注．通过端到端的方式对不规则网络结构（非欧氏空间）进行分析和挖掘，图神经网络能够有效揭示网络中的节点特征、关系特征和网络结构等隐含语义信息，生成上下文相关的节点表示，在知识图谱、社交网络、推荐系统等方面发挥了重要的作用[1]．

节点分类是图神经网络最重要的下游任务之一，目的是利用已标注节点学习网络语义信息，生成节点特征表示，识别未标注节点的类别[2]．例如，在引文网络中，节点表示文献，文献引用构建了节点间的连接关系，通过学习文献的文本和引用关系特征，可以生成文献的特征表示，预测文献的研究主题[3]．

图神经网络节点分类主要分为同质网络节点分类和异质网络节点分类．同质网络仅包含一种节点类型，且节点间的连接关系单一，如引文网络[3]中文献间的引用关系，协作网络[4]中学者间的合作关系和社交网络[5]中用户间的关注关系等．而现实世界中的数据关系复杂多样，构成的网络更多是异质网络，它包含多种类型的节点和关系，不同类型的节点以及节点间的关系揭示了不同的语义信息．例如，在学术网络[6]中，除了文献间的引用关系外，还包括学者与文献间的写作关系、文献与期刊间的发表关系等．

图神经网络能够利用网络中丰富的语义信息和全面的结构信息进行精准的节点分类，但是现有研究在节点分类上仍然面临着如下挑战：

1）如何充分利用异质网络中丰富的语义信息，提升节点分类的效果．同质网络仅包含一种节点类型和关系类型，语义信息单一，在建模复杂数据关系上存在明显的局限性．异质网络由多种类型的节点和关系组成，语义信息丰富，能够有效描述多样化的数据特征．充分利用异质网络的多语义信息开展研究，可以适应更加复杂的数据场景．

2）如何充分利用网络中全面的结构信息，学习合理的节点特征表示．异质网络结构复杂，仅使用对称元路径[6]揭示网络中同类型节点间的内在联系，不能充分发现邻居节点的相关性．需要引入邻居节点权重，结合节点特征、关系特征、邻域特征等全面的网络结构信息进行分析，区分不同邻居节点的重要性，生成更合理的节点表示，提升节点分类效果．

为应对上述挑战，本文提出了一种基于图卷积的异质网络节点分类框架(heterogeneous node classification framework, HNCF)，用于解决异质网络上的节点分类问题．首先，约简异质网络，将异质网络转换为语义化同质网络，简化网络结构并保留异质网络的多语义信息；然后，基于图神经网络中的消息传递框架，设计图卷积节点分类方法，在语义化同质网络上学习合理的邻居权重，生成节点异质语义表示，并利用节点异质语义表示识别节点的类别标签．同时，分别在 3 个公开数据集上，对本文设计的异质网络约简方法和图卷积节点分类方法进行了实验对比和分析，结果表明本文所提方法有效提升了节点分类效果．

本文的主要贡献包括 3 个方面：

1）提出了一种基于图卷积的异质节点分类框架 HNCF，通过将异质网络约简为语义化同质网络，并利用消息传递框架实现图卷积节点分类，有效地解决了异质网络上的节点分类问题．

2）提出了一种异质网络约简方法，结合对称元路径和关系表示池化函数，设计转换规则，生成语义化同质网络，简化了异质网络结构，保留了丰富语义信息，便于适应消息传递框架．

3）提出了一种图卷积节点分类方法，基于消息传递框架，计算无 1-sum 约束的邻居节点权重，加权聚合邻域特征，生成节点语义表示，充分利用了语义化同质网络的结构信息，进一步提升了分类效果．

## 1 相关工作

图神经网络能够充分利用网络中丰富的语义信息和全面的网络结构信息，以端到端的方式学习任务相关的节点表示，解决节点分类等问题．根据网络类型的不同，可将图神经网络节点分类的方法分为同质图神经网络方法和异质图神经网络方法．

在同质图神经网络方法中，Tomas 等人[3]提出了一种图卷积网络(graph convolution network, GCN)，通过编码一阶邻域结构学习节点表示，综合利用了网络结构和节点特征信息，提升了节点分类效果．由于

GCN 无法适应大规模动态网络，Hamilton 等人[7]提出了一种图神经网络框架 GraphSAGE（graph sample and aggregate），对中心节点的高阶邻域进行采样，结合不同的池化方式聚合邻域特征，提升了分类性能．Velickovic 等人[8]认为不同邻居节点具有不同的重要性，提出了图注意力网络(graph attention network, GAT)学习邻居节点权重，通过多头注意力对邻域特征进行加权聚合，生成了更加合理的节点表示，获得了比 GraphSAGE 更好的分类效果．为了学习同质网络中潜在的关系特征，Wang 等人[9]提出了一种邻边卷积方法，通过节点特征表示的差异性体现关系特征，将关系包含的语义信息融入到了节点表示学习中．Gilmer 等人[10]提出了一种通用的消息传递框架(message passing neural network, MPNN)，对现有的图卷积网络进行了泛化，通过消息传递规则和聚合函数学习节点表示，有效建模了节点特征、关系特征和邻域结构等网络信息．虽然同质图神经网络节点分类方法能够充分学习网络结构特征，但是仅能建模单一语义关系，无法直接运用在异质网络中．因此，现有研究更关注于异质图神经网络方法，以充分利用网络语义信息，适应复杂的数据场景．

异质图神经网络方法重点关注网络结构和语义关系建模．异质网络结构由不同类型的节点和关系组成，表达了复杂数据的信息交互，融入异质结构特征有利于提升下游任务效果．在网络结构建模方法中，为了充分利用邻域结构信息，异质图卷积（heterogeneous graph neural network，HetGNN）方法[11]提取了属性、文本、图片等异质邻居节点特征，结合注意力机制[12]生成包含丰富特征的节点表示．与 HetGNN 方法不同，Hu 等人[13]提出了一种异质网络 Transformer 框架，通过时序编码和图采样对大型动态异质网络进行预训练，提取异质网络结构特征，提升节点分类等任务的效果．Yun 等人[14]从结构生成的角度出发，提出了一种端到端的异质网络生成框架 GTN(graph transformer networks)，通过邻接矩阵学习软选择机制，构建新的网络结构用于节点表示学习．异质网络语义信息主要通过关系类型和关系组合表达，建模多语义关系能够有效揭示丰富的语义信息．在语义关系建模方法中，为了减少异质网络建模的复杂性，HAN(heterogeneous graph attention network)方法[15]通过对称元路径将异质网络分解为多个表达不同语义关系的同质网络，并利用 GAT 聚合邻域特征，综合考虑了不同邻居节点的重要性．陈亦琦等人[16]提出了一种复合关系图卷积网络(composite relation graph convolution network, CRGCN)，通过建模属性网络中用户-属性等基本关系和用户-属性-用户等复合关系，学习编码关系特征的节点表示．Vashishth 等人[17]提出了一种多关系图神经网络框架 (composition-based multi-relational graph convolutional networks, CompGCN)，根据不同关系类型和邻居类型设置不同的组合算子，通过聚合邻域特征学习节点特征表示．异质网络结构复杂多样，直接建模难度较高，与网络结构建模方法相比，语义关系建模方法可以简化异质网络结构，降低语义和结构信息提取的复杂度．在语义关系建模方法中，虽然现有方法利用对称元路径简化异质网络结构，保留了丰富语义特征，但忽略了网络结构中关系语义信息的深度挖掘，无法充分发现不同连接关系和邻域节点语义提取的差异性，导致难以准确捕捉网络中的语义和结构特征，影响节点分类效果．

本文提出的方法充分利用了异质网络语义和结构信息，有效解决了异质网络节点分类问题．该方法可以深入挖掘语义关系特征，通过保留异质语义信息的关系表示向量和无 1-sum 约束的邻居节点权重充分发现不同连接关系和邻居语义提取的差异性，提升节点分类效果．

## 2 基于图卷积的异质网络节点分类框架 HNCF

重点介绍信息网络、异质网络、对称元路径、单节点多关系异质网络、语义化同质网络等概念及定义，并对异质网络节点分类框架 HNCF 进行具体描述．

### 2.1 概念定义

**定义 1.** 信息网络. 信息网络被定义为无向图 $G=(V, E, A, P, \phi, \varphi)$，其中，$V$ 表示节点集合，$E$ 表示关系集合，$A$ 表示节点类型集合，$P$ 表示关系类型集合，$\phi: V \to A$ 为节点类型映射函数，$\varphi: E \to P$ 为关系类型映射函数．

异质网络是具有多种节点和关系类型的信息网络，多样化的节点和关系共同描述了异质网络中的语义信息，见定义 2．

**定义 2.** 异质网络[15]. 给定信息网络 $G_h=(V_h, E_h, A_h, P_h, \phi_h, \varphi_h)$，如果$|A_h|+|P_h|>2$，则称 $G_h$ 为异质信息网络，简称为异质网络．

**例 1.** 图 1(a)所示的异质网络由多种类型的节点和关系构成，节点集合 $V_h=\{z_1, s_1, s_2, s_3, u_1, u_2\}$，关系集合 $E_h=\{<z_1, s_1>, <z_1, s_2>, <z_1, s_3>, <s_1, s_2>, <s_2, s_3>, <s_1, u_1>, <s_2, u_1>, <s_2, u_2>, <s_3, u_2>\}$，节点类型集合 $A_h=\{Z, S, U\}$，关系类型集合 $P_h=\{ZS, SS, SU\}$，满足异质网络$|A_h|+|P_h|>2$ 的约束．其中：节点映射 $\phi_h(z_1)=Z$，$\phi_h(s_1)=\phi_h(s_2)=\phi_h(s_3)=S$，$\phi_h(u_1)=\phi_h(u_2)=U$；关系映射函数$\varphi_h(<z_1, s_1>)=\varphi_h(<z_1, s_2>)=\varphi_h(<z_1,$

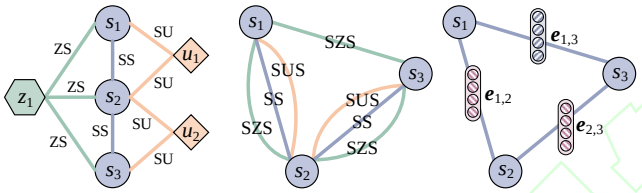
$s_3$>) = ZS，$\varphi_h$(<$s_1$, $s_2$>) = $\varphi_h$(<$s_2$, $s_3$>) = SS，$\varphi_h$(<$s_1$, $u_1$>) = $\varphi_h$(<$s_2$, $u_1$>) = $\varphi_h$(<$s_2$, $u_2$>) = $\varphi_h$(<$s_3$, $u_2$>) = SU.

元路径描述了异质网络上不同类型节点间的组合关系，而对称元路径是首尾对称的元路径，表达了同类型节点间的语义关系，详见定义 3 和定义 4.

**定义** 3. 元路径[15]. 异质网络 $G_h$ 上长度为$\delta(\delta>0)$的元路径 $f$ 被定义为：$A_1 \circ P_1 \circ \ldots \circ A_i \circ P_i \circ \ldots \circ P_l \circ A_{\delta+1}$ (简写为 $A_1 \ldots A_i \ldots A_{\delta+1}$)，描述了节点类型 $A_1$ 和 $A_{\delta+1}$ 之间的组合关系 $P_1 \to P_2 \to \ldots \to P_\delta$.

**定义** 4. 对称元路径. 给定异质网络 $G_h$ 上的元路径 $f = A_1 \circ P_1 \circ \ldots \circ A_i \circ P_i \circ \ldots \circ P_l \circ A_{\delta+1}$，如果对$\forall i \in [1, \delta+1]$，都有 $A_i = A_{\delta+2-i}$，则称元路径 $f$ 为对称元路径.

**例** 2. 在图 1(a)所示的异质网络中，ZSU 是 1 条长度为 2 的元路径，描述了节点类型 Z,S,U 之间的组合关系，SUS 是 1 条长度为 2 的对称元路径，$s_1$—$u_1$—$s_2$ 是对称元路径 SUS 的实例，表达了 S 类型节点 $s_1$，$s_2$ 通过与 $u_1$ 节点相连构建的语义关系.




(a) 异质网络 (b) 单节点多关系异质网络 (c) 语义化同质网络

Fig 1. Examples of different types of graph structures

图 1 不同类型的网络结构示例

单节点多关系异质网络由一种节点类型和多种关系类型构成，是异质网络的一个特例. 该网络仅包含单一类型的节点，且节点之间具有多种类型的关系，详见定义 5.

**定义** 5. 单节点多关系异质网络. 给定信息网络 $G_r = (V_r, E_r, A_r, P_r, \phi_r, \varphi_r)$，如果$|A_r| = 1$ 且$|P_r| > 1$，则称 $G_r$ 为单节点多关系异质网络. 此时，$E_r = \{<v_i, v_j, t> \mid v_i, v_j \in V_r \wedge t \in P_r\}$，即 $v_i$ 与 $v_j$ 间的任意关系均对应一种关系类型.

**例** 3. 如图 1(b)所示，单节点多关系异质网络由一种节点类型和多种关系类型构成，节点集合 $V_r = \{s_1, s_2, s_3\}$，关系集合 $E_r$ = {<$s_1$, $s_2$, SZS>, <$s_1$, $s_2$, SUS>, <$s_1$, $s_2$, SS>, <$s_2$, $s_3$, SZS>, <$s_2$, $s_3$, SUS>, <$s_2$, $s_3$, SS>, <$s_1$, $s_3$, SZS>}，节点类型集合 $A_r$ = {S}，关系类型集合 $P_r$ = {SZS, SS, SUS}，满足单节点多关系异质网络$|A_r| = 1$ 和$|P_r| > 1$ 的约束. 其中，节点映射函数$\phi_r(s_1) = \phi_r(s_2) = \phi_r(s_3)$ = S，关系映射函数$\varphi_r$(<$s_1$, $s_2$>) = $\varphi_r$(<$s_2$, $s_3$>) = {SZS, SS, SUS}，$\varphi_r$(<$s_1$, $s_3$>) = {SUS}.

与异质网络不同的是，语义化同质网络仅包含单一的节点类型和关系类型，同类型节点间的关系表示描述了关系包含的语义信息.

**定义** 6. 语义化同质网络. 给定信息网络 $G_g = (V_g, E_g, A_g, P_g, \phi_g, \varphi_g)$，如果$|A_g| = 1$ 且$|P_g| = 1$，则称 $G_g$ 为语义化同质网络，简称为同质网络. 此时，$E_g = \{<v_i, v_j, \boldsymbol{e}_{i,j}> \mid v_i, v_j \in V_g \wedge \boldsymbol{e}_{i,j} \in \mathbb{R}^m\}$，$P_g = \{t_g\}$，$v_i$ 与 $v_j$ 间仅具有唯一关系类型 $t_g$，且对应 $m$ 维异质关系表示 $\boldsymbol{e}_{i,j}$，表达了关系的异质语义信息.

**例** 4. 如图 1(c)所示，语义化同质网络由一种类型的节点和关系构成，节点集合 $V_g = \{s_1, s_2, s_3\}$，关系集合 $E_g$ = {<$s_1$, $s_2$, $\boldsymbol{e}_{1,2}$>, <$s_2$, $s_2$, $\boldsymbol{e}_{2,3}$>, <$s_1$, $s_3$, $\boldsymbol{e}_{1,3}$>}，节点类型集合 $A_g$ = {S}，关系类型集合 $P_g = \{t_g\}$，满足语义化同质网络$|A_g| = 1$ 和$|P_g| = 1$ 的约束. 其中，节点映射函数$\phi_g(s_1) = \phi_g(s_2) = \phi_g(s_3)$ = S，关系映射函数$\varphi_g$(<$s_1$, $s_2$>) = $\varphi_g$(<$s_2$, $s_3$>) = $\varphi_g$(<$s_1$, $s_3$>) = $\{t_g\}$.

为了方便叙述，表 1 给出了 HNCF 框架中用到的符号，并解释了符号代表的含义.

**Table 1 Symbol Explanation Table**

**表 1 符号对照表**

| 符号 | 含义 |
|---|---|
| $A_c$ | 待分类的节点类型 |
| $v_i \overset{f}{\leftrightarrow} v_j$ | 节点 $v_i$ 和 $v_j$ 通过对称元路径 $f$ 相连 |
| $M$ | 对称元路径集合 |
| $\omega$ | 对称元路径类型映射函数 |
| $T$ | 关系类型嵌入表 |
| $\boldsymbol{x}_i, \boldsymbol{x}_j$ | 中心节点 $v_i$ 和邻居节点 $v_j$ 的初始特征表示 |
| $\boldsymbol{e}_t$ | 关系类型 $t$ 的关系表示 |
| $\boldsymbol{e}_{i,j}$ | 中心节点 $v_i$ 和邻居节点 $v_j$ 的异质关系表示 |
| $\alpha_j$ | 邻居节点 $v_j$ 的权重 |
| $\boldsymbol{W}_\cdot, \boldsymbol{b}_\cdot$ | 全连接层权重和偏量，$\cdot$ 表示下标 |
| $b_{\varphi_t(v_i, v_j)}$ | 关系组合$\varphi_t(v_i, v_j)$的全连接层偏量 |
| $N_i$ | 中心节点 $v_i$ 的邻居节点集合 |
| $\boldsymbol{w}_j$ | 邻居节点 $v_j$ 的加权特征表示 |
| $\boldsymbol{h}_i$ | 中心节点 $v_i$ 的邻域上下文表示 |
| $\boldsymbol{x}_i'$ | 中心节点 $v_i$ 的更新节点表示 |
| $\boldsymbol{P}_i$ | 中心节点 $v_i$ 的类别概率分布 |

## 2.2 框架描述

为了对异质网络上特定类型的节点进行分类，本文提出一种基于图卷积的异质网络节点分类框架 HNCF，学习异质网络上的语义关系和网络结构信息，提升节点类别标签识别的精准性.

异质网络的结构复杂、语义丰富，为了合理地学习异质网络中的语义特征，通常将异质网络转换成同质网络，然后利用图神经网络学习节点表示，但在转换同质网络的过程中，缺乏关系特征建模，难以有效利用语义关系信息. 因此，HNCF 首先通过异质网络约简，将异质网络转换成语义化同质网络，简化网络

结构并充分保留异质网络的多语义信息.

DGL(deep graph library)[18]，PyG(PyTorch geometric)[19]等主流图神经网络计算库使用的消息传递框架是一种图卷积网络泛化框架，支持节点特征、关系特征和邻域结构等特征信息建模[10]. 为了融入异质语义特征、充分利用关系特征和邻居权重等网络结构信息，HNCF 基于消息传递框架设计图卷积节点分类方法，学习合理的邻居节点权重，进一步生成节点异质语义表示，提升节点类别标签识别精度.

HNCF 主要包括 2 个阶段，分别是异质网络约简和图卷积节点分类，如图 2 所示.

1）异质网络约简.给定待学习表示的节点类型，将异质网络简化为语义化同质网络，保留多种语义信息，便于适应消息传递框架. 首先，利用对称元路径，将异质网络转换成单节点多关系异质网络，通过节点间不同的关系类型保存异质网络中的多种语义信息. 然后，对节点间的多重关系进行语义融合，将单节点多关系异质网络进一步转换成语义化同质网络，利用节点间的关系表示向量保存融合后的异质语义信息.

2）图卷积节点分类.基于消息传递框架和语义化同质网络，设计图卷积网络节点分类方法，融入异质语义信息和网络结构信息. 首先，结合关系表示向量，学习邻居权重，区分邻居节点重要性，并对邻域特征进行加权聚合，生成节点异质语义表示. 然后，利用全连接层将节点异质语义表示投影到类别空间，计算节点的类别概率分布，并根据类别概率分布确定节点的类别标签.

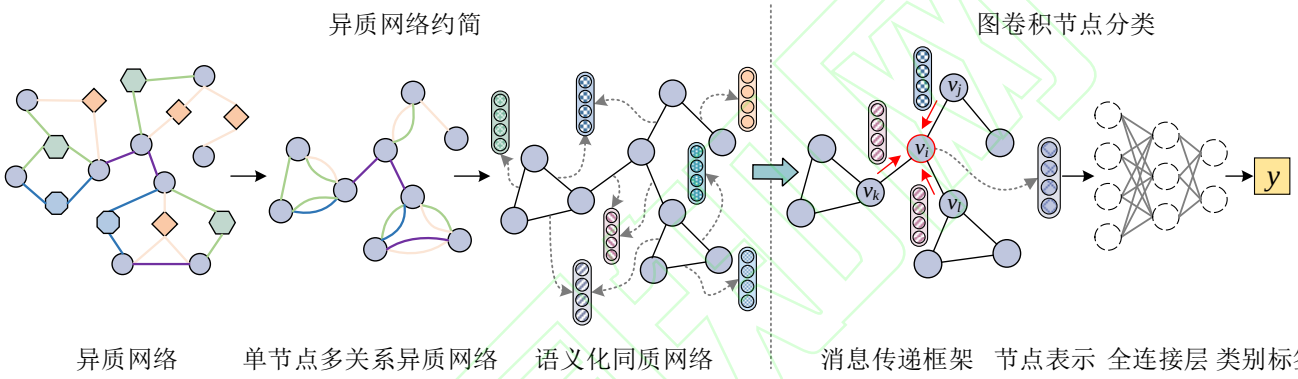



Fig. 2 Heterogeneous node classification framework HNCF

图 2 基于图卷积的异质网络节点分类框架 HNCF

## 3 异质网络约简方法

本节详述 HNCF 框架的异质网络约简方法，设计转换规则，将异质网络简化为语义化同质网络，利用节点间的关系表示保存异质网络中的多语义信息.

### 3.1 异质网络-单节点多关系异质网络转换

首先将异质网络 $G_h$ 转换为单节点多关系异质网络 $G_r$，受文献[15]启发，本文设计了转化规则进行网络结构初步化简，通过不同的对称元路径构建同类型节点间的多语义关系.

对于给定的特定节点类型 $A_c$，令 $M$ 为异质网络 $G_h$ 上的对称元路径集合，即 $M=\{f=A_1A_2\ldots A_{\delta+1} \mid A_1=A_c \wedge \forall i \in [1, l+1], A_i=A_{\delta+2-i}\}$. 下面给出从异质网络 $G_h=(V_h, E_h, A_h, P_h, \phi_h, \varphi_h)$到单节点多关系异质网络 $G_r=(V_r, E_r, A_r, P_r, \phi_r, \varphi_r)$的网络结构转换规则.

**规则** 1. 节点转换规则. 给定节点类型 $A_c$，$\forall v_i \in V_h$，若$\varphi_h(v_i)=A_c \Rightarrow$ 构建节点集合 $V_r=\{v_i \mid v_i \in V_h \wedge \varphi_h(v_i)=A_c\}$.

**规则** 2. 关系转换规则. 记$v_i \overset{f}{\leftrightarrow} v_j$为 $v_i$ 和 $v_j$ 之间在异质网络 $G_h$ 上存在对称元路径 $f$ 连接. $\forall v_i, v_j \in V_r$ 且 $v_i \neq v_j$，若$\exists f \in M$，使得$v_i \overset{f}{\leftrightarrow} v_j \Rightarrow$ 构建关系集合 $E_r=\{<v_i, v_j, t> \mid v_i \overset{f}{\leftrightarrow} v_j \wedge t=\omega(f), \omega: M \to P_r\}$.

**规则** 3. 类型转换规则. 构建节点类型集合 $A_r=\{A_c\}$；关系类型集合 $P_r=\{t \mid t=\omega(f) \wedge f \in M\}$.

**规则** 4. 映射转换规则. 令节点类型映射函数$\phi_r=\phi_h$；关系类型映射函数$\varphi_r: E_r \to P_r$，$\varphi_r(v_i, v_j)=\{t \mid <v_i, v_j, t> \in E_r \wedge v_i, v_j \in V_r \wedge t \in P_r\}$.

规则 1 转换后的节点集合 $V_r$ 仅保留给定类型 $A_c$ 的节点，且 $V_r \subset V_h$.

规则 2 构建转换后的关系集合 $E_r$，如果 $V_r$ 中的节点 $v_i$ 和 $v_j$ 在异质网络 $G_h$ 上通过对称元路径 $f$ 进行连接，则 $v_i$ 和 $v_j$ 之间在单节点多关系异质网络 $G_r$ 中具有连接关系，对应的关系类型 $t$ 可以保留化简前节点间的语义信息，$v_i$ 和 $v_j$ 之间的连接关系可以有多种，且关系类型也不一定相同. $\omega$为对称元路径集合 $M$ 到关系类型集合 $P_r$ 的映射函数，构建了对称元路径 $f$ 与关系类型 $t$ 之间的一一映射关系.

规则 3 转换后的节点类型集合 $A_r$ 只有一种类型 $A_c$，且 $A_r \subset A_h$；规则 4 中节点类型映射函数保持不变，

即$\phi_r$与$\phi_h$相同. 与转换前的关系类型不同，规则 3 和规则 4 将生成新的关系类型，在异质网络 $G_h$ 中节点类型为 $A_c$ 的节点之间，如果存在不同的对称元路径，则分别对应单节点多关系异质网络 $G_r$ 中不同的关系类型.

利用对称元路径进行网络结构转换，可以保留特定的节点类型，并构建节点间的多重语义关系，不仅降低了网络结构复杂度，而且保留了节点间的异质语义信息.

**例** 5. 如图 1(a)和 1(b)所示，设 $A_c$ = S，$M$ = {SZS, SUS, SS}，根据异质网络-单节点多关系异质网络转换规则，可知 $V_r$ = {$s_1$, $s_2$, $s_3$}，$E_r$ = {<$s_1$, $s_2$, SZS>, <$s_1$, $s_2$, SUS>, <$s_1$, $s_2$, SS>, <$s_2$, $s_3$, SZS>, <$s_2$, $s_3$, SUS>, <$s_2$, $s_3$, SS>, <$s_1$, $s_3$, SZS>}，$A_r$ = {S}，$P_r$ = {SZS, SUS, SS}. 其中，$\omega$分别将元路径映射为 SZS，SUS，SS 类型.

### 3. 2 单节点多关系异质网络-同质网络转换

由 3.1 节的网络结构转化规则，得到单节点多关系异质网络 $G_r$. 本节进一步讨论简化网络结构的方法，把 $G_r$ 转换为语义化同质网络 $G_g$，并将异质语义融入到 $G_g$ 的关系表示中.

本文首先利用 embedding 方法[20]和 $G_r$ 的关系类型集合 $P_r$ 构建关系类型嵌入表 $T$，对于每种关系类型 $t \in P_r$，在 $T$ 中设置对应的向量表示 $\boldsymbol{e}_t \in \mathbb{R}^m$，通过向量形式表达关系类型的语义信息，$m$ 为关系类型表示的维度.

给定关系类型嵌入表 $T$，通过规则 5~8 所示的关系语义转换规则将单节点多关系异质网络 $G_r$ = ($V_r$, $E_r$, $A_r$, $P_r$, $\phi_r$, $\varphi_r$)转换为语义化同质网络 $G_g$ = ($V_g$, $E_g$, $A_g$, $P_g$, $\phi_g$, $\varphi_g$).

**规则** 5. 节点转换规则. 构建节点集合 $V_g = V_r$.

**规则** 6. 关系转换规则. 构建关系集合 $E_g$ = {<$v_i$, $v_j$, $\boldsymbol{e}_{i,j}$> | $\boldsymbol{e}_{i,j} = \Theta_{t\in\varphi_r(v_i, v_j)}\boldsymbol{e}_t \wedge \boldsymbol{e}_t \in T$}.

**规则** 7. 类型转换规则. 构建节点类型集合 $A_g = A_r$，关系类型集合 $P_g$ = {$t_g$}.

**规则** 8. 映射转换规则. 设置节点类型映射函数 $\phi_g = \phi_r$，构建关系类型映射函数 $\varphi_g : E_g \to P_g$，$\varphi_g(v_i, v_j)$ = {$t_g$ | <$v_i$, $v_j$, $\boldsymbol{e}_{i,j}$> $\in E_g \wedge t_g \in P_g$}.

由于转换前后节点未发生变化，规则 5 转换后的节点集合保持不变，保留了单节点多关系异质网络中 $G_r$ 中的所有节点.

规则 6 转换后的关系类型集合 $E_g$ 中，$v_i$ 和 $v_j$ 间的多种关系类型$\varphi_r(v_i, v_j)$被 max，add，mean 等池化聚合函数$\Theta$转换为单一关系类型 $t_g$，并将融合后的多种关系语义信息保存到异质关系表示 $\boldsymbol{e}_{i,j}$ 中.

因为语义化同质网络 $G_g$ 中只包含一种节点类型和关系类型，规则 7 转换后节点类型集合 $A_g$ 仅包含一种节点类型 $A_c$，$A_g = A_r$. 同时，规则 8 会生成新的关系类型 $t_g$，转换为语义化同质网络 $G_g$ 后，节点间的关系类型均相同.

通过关系语义转换规则，节点 $v_i$ 和节点 $v_j$ 间的多语义关系被转换成了具有异质关系表示 $\boldsymbol{e}_{i,j}$ 的单一关系连接，不仅融合了多种异质语义信息，同时进一步将单节点多关系异质网络简化为了语义化同质网络，便于适应消息传递框架.

**例** 6. 如图 1(b)和 1(c)所示，通过为关系类型 SZS，SUS，SS 设置向量表示 $\boldsymbol{e}_{SZS}$，$\boldsymbol{e}_{SUS}$，$\boldsymbol{e}_{SS}$，并根据单节点多关系异质网络-同质网络转换规则，可得 $V_g$ = {$s_1$, $s_2$, $s_3$}，$E_g$ = {<$s_1$, $s_2$, $\boldsymbol{e}_{1,2}$>, <$s_2$, $s_2$, $\boldsymbol{e}_{2,3}$>, <$s_1$, $s_3$, $\boldsymbol{e}_{1,3}$>}，$A_g$ = {S}，$P_g$ = {$t_g$}.

## 4 图卷积节点分类方法

本节讨论 HNCF 框架的图卷积节点分类方法. 基于消息传递框架，设计消息传递过程，学习语义化同质网络上的邻居节点权重、语义关系、邻域上下文等网络结构特征，生成更新节点表示，并输入单层全连接层，计算节点的类别概率分布，识别类别标签.

### 4. 1 消息传递过程

消息传递框架是一种图卷积泛化框架，能够灵活建模节点特征、关系特征和邻域特征，主要通过邻域特征聚合和节点表示更新实现图卷积操作[10]，分别如式(1)、式(2)所示：

$$\boldsymbol{h}_i = \sum_{j\in N_i} F(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{e}_{ij}), \tag{1}$$

$$\boldsymbol{x}_i' = Q(\boldsymbol{h}_i, \boldsymbol{x}_i), \tag{2}$$

其中，$N_i$ 表示节点 $v_i$ 的邻居节点集合，$\boldsymbol{x}_i$ , $\boldsymbol{x}_j$ 为节点 $v_i$ 和 $v_j$ 的特征表示，$F$，$Q$ 分别为消息传递函数和节点更新函数，$\boldsymbol{h}_i$ 表示节点 $v_i$ 的邻域上下文表示，$\boldsymbol{x}_i'$ 表示 $v_i$ 更新后的节点表示.

通过消息传递函数 $F$，消息传递框架获取邻域节点 $v_j$ 特征，聚合生成邻域上下文表示 $\boldsymbol{h}_i$，并利用节点更新函数 $Q$ 对 $v_i$ 的特征表示进行更新，得到图卷积后的更新节点表示 $\boldsymbol{x}_i'$.

本文利用消息传递框架，学习语义化同质网络 $G_g$ 中的节点表示，用于识别节点类别标签. 给定节点初始特征表示集合 $X$ = {$\boldsymbol{x}_i \in \mathbb{R}^n | v_i \in V_r$}，可从节点的文本、属性等内容中提取，或利用 embedding 方法[20]学习节点的初始特征表示.

图 3 展示了由中心节点 $v_i$ 及其邻居节点 $v_j$,$v_k$,$v_l$ 构成的语义化同质网络上的消息传递过程. 其中，$\boldsymbol{x}_i$,$\boldsymbol{x}_j$,$\boldsymbol{x}_k$,$\boldsymbol{x}_l$ 分别为节点 $v_i$,$v_j$,$v_k$,$v_l$ 的初始特征表示；$\boldsymbol{e}_{i,j}$,$\boldsymbol{e}_{i,k}$,$\boldsymbol{e}_{i,l}$ 分别为中心节点 $v_i$ 与邻居节点 $v_j$,$v_k$,$v_l$ 之间的异质关系表示. 具体步骤如下：

1）邻居权重计算．以邻居节点 $v_j$ 为例，根据 $v_i$ 和 $v_j$ 的初始特征表示之差 $\boldsymbol{x}_i \ominus \boldsymbol{x}_j$，以及异质关系表示 $\boldsymbol{e}_{i,j}$，计算邻居节点 $v_j$ 的邻居权重 $\alpha_j$.

2）加权邻域聚合．利用非线性变换提取邻居节点初始特征表示和异质关系表示中的特征，并通过邻居权重加权聚合，获取中心节点 $v_i$ 的邻域上下文表示 $\boldsymbol{h}_i$.

3）节点表示更新．利用邻域上下文表示 $\boldsymbol{h}_i$ 对中心节点 $v_i$ 的特征表示进行更新，得到包含异质语义特征的更新节点表示 $\boldsymbol{x}_i'$.

4.1.1 邻居权重计算

为了区分不同邻居节点的重要性程度，减少不相关邻居节点引入的噪声，本文借鉴注意力机制[12]和关系卷积[9]的思想，设计了具体的邻居权重计算方法.利用中心节点 $v_i$ 和邻居 $v_j$ 的特征表示之差 $\boldsymbol{x}_i \ominus \boldsymbol{x}_j$ 和异质关系表示 $\boldsymbol{e}_{i,j}$，计算邻居节点 $v_j$ 的邻居权重 $\alpha_j$，确定邻居节点的重要性，具体计算方法如式(3)：

$$\alpha_j = \sigma((\boldsymbol{W}_{\text{wgt}} \cdot \Psi(\boldsymbol{x}_i \ominus \boldsymbol{x}_j, \boldsymbol{e}_{i,j})) + b_{\varphi_r(v_i, v_j)}). \quad (3)$$

式(3)中，$\alpha_j \in (0, 1)$为邻居节点 $v_j$ 相对于 $v_i$ 的权重，$\sigma$为 sigmoid 函数，$\boldsymbol{W}_{\text{wgt}} \in \mathbb{R}^{1\times(n+m)}$ 是全连接层的权重，$\cdot$ 为矩阵乘法运算符，$\ominus$ 为向量减法运算符，$+$ 表示标量加法运算符，$\Psi$为向量拼接函数.

同时，式(3)为每种关系类型组合学习特定的偏量 $b_{\varphi_r(v_i, v_j)} \in \mathbb{R}^1$，以尽可能的区分中心节点 $v_i$ 和不同邻居节点 $v_j$ 之间的语义关系．其中，关系类型组合偏量的个数为 $2^{|T|}-1$，$|T|$表示关系类型的数目．设关系类型嵌入表 $T$ 中包含 2 种关系类型 $t_1$ 和 $t_2$，则可能的关系类型组合为$\{\{t_1\}, \{t_2\}, \{t_1, t_2\}\}$.

与同质网络不同，异质网络约简后构建的邻居节点通常数量不一，相关性差异较大，需要区分不同邻居节点的重要性，以生成合理的节点表示．GAT[8]等图注意力机制在计算邻居权重时存在 1-sum 约束，需要利用 softmax 函数对邻居权重进行归一化，使所有权重之和为 1．然而，1-sum 约束对邻居节点数量敏感，使得邻居权重的确定产生偏向性，邻居节点数量较少时会使权重偏大，数量较多时会使权重偏小，导致邻居权重学习相对不合理．例如，当邻居节点数为 100 时，如果邻居节点均非常相关，1-sum 约束会使得邻居权重约为 0.01，过小的邻居权重会导致难以提取单个邻居节点的特征．同时，当邻居节点数为 10，如果邻居节点均不相关，1-sum 约束仍会使得邻居权重约为 0.1，为不相关邻居节点设置了较大的权重，从而引入了噪声．

因此，本文去除了邻居节点权重的 1-sum 约束，通过 sigmoid 函数将权重限定在(0, 1)内，权重之和不为 1．同时，邻居节点权重的大小由初始特征表示之差 $\boldsymbol{x}_i \ominus \boldsymbol{x}_j$、异质关系表示 $\boldsymbol{e}_{i,j}$、全连接层权重 $\boldsymbol{W}_{\text{wgt}}$ 和组合类型偏量 $b_{\varphi_r(v_i, v_j)}$ 共同确定，不需要使用 softmax 进行归一化，不受邻居节点数量的干扰，缓解了因邻居节点数量变化引起的偏向性问题．节点特征表示之差 $\boldsymbol{x}_i \ominus \boldsymbol{x}_j$、异质关系表示 $\boldsymbol{e}_{i,j}$ 和组合类型偏量 $b_{\varphi_r(v_i, v_j)}$ 分别体现了初始特征分布和语义关系的不同，充分考虑了中心节点和邻居节点连接关系的差异性．

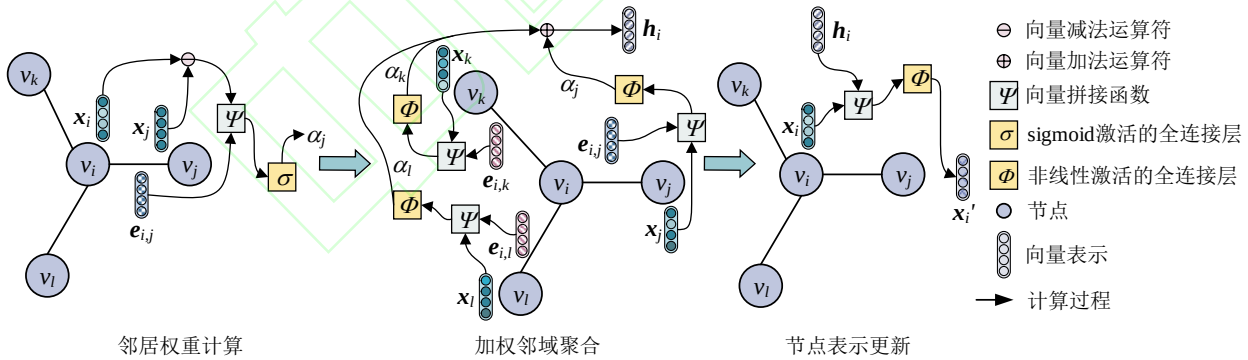



Fig. 3 The message passing process

图 3 消息传递过程

4.1.2 加权邻域聚合

在计算邻居权重之后，首先会对邻居节点 $v_j$ 的节点特征 $\boldsymbol{x}_j$ 和异质关系表示 $\boldsymbol{e}_{i,j}$ 进行非线性变换，然后利用邻居节点权重 $\alpha_j$ 计算加权特征表示 $\boldsymbol{w}_j$，并提取 $v_i$ 的邻域上下文表示 $\boldsymbol{h}_i$，如式(4)和式(5)：

$$\boldsymbol{w}_j = \alpha_j \times \Phi((\boldsymbol{W}_{\text{ngb}} \cdot \Psi(\boldsymbol{x}_j, \boldsymbol{e}_{i,j})) \oplus \boldsymbol{b}_{\text{ngb}}), \quad (4)$$

$$\boldsymbol{h}_i = \frac{1}{|N_i|} \sum_{j\in N_i} \boldsymbol{w}_j. \quad (5)$$

式(4)和式(5)中，$N_i$ 表示中心节点 $v_i$ 的邻居节点集合，$\Phi$是非线性激活函数，$\boldsymbol{W}_{\text{ngb}} \in \mathbb{R}^{o\times(n+m)}$和 $\boldsymbol{b}_{\text{ngb}} \in \mathbb{R}^o$ 分别是全连接层的权重和偏量，$o$ 表示邻域上下文表示的维度，$\cdot$为矩阵乘法运算符，$\times$表示标量乘法运算符，$\oplus$ 表示向量加法运算符，$\Psi$为向量拼接函数.

加权邻域聚合综合考虑了邻居节点的特征和权重，以及中心节点 $v_i$ 与邻居节点 $v_j$ 之间的异质关系表

示 $\boldsymbol{e}_{i,j}$ 等多种网络结构信息，充分挖掘了连接关系和邻居节点语义提取的差异性，获得了更加合理的邻域特征，减少了邻域结构中的噪声.

此外，类似于 GCN[3]，在进行加权邻域聚合时，本文使用度归一化平衡因不同节点度引起的尺度不一致问题，同时避免在邻居节点数目较多时和权重较大时可能产生的梯度爆炸问题.

本文将式(1)分为了邻居权重计算和加权邻域聚合 2 个步骤. 首先，结合异质语义表示 $\boldsymbol{e}_{i,j}$ 学习邻居节点权重 $\alpha_j$，充分利用关系表示中的异质语义信息. 然后，利用全连接层和非线性转换函数实现了消息传递函数 $F$，根据不同邻居节点的重要性提取加权特征表示 $\boldsymbol{w}_j$. 最后，聚合邻域特征，生成中心节点 $v_i$ 的邻域上下文表示 $\boldsymbol{h}_i$.

4.1.3 节点表示更新

加权邻域聚合后，根据邻域上下文表示 $\boldsymbol{h}_i$ 对中心节点 $v_i$ 的节点表示进行更新，得到包含异质语义信息的更新节点表示 $\boldsymbol{x}_i'$，如式(6)：

$$\boldsymbol{x}_i' = \Phi((\boldsymbol{W}_{\text{upt}} \cdot \Psi(\boldsymbol{x}_i, \boldsymbol{h}_i)) \oplus \boldsymbol{b}_{\text{upt}}). \tag{6}$$

式(6)中，$\Phi$是非线性激活函数，$\boldsymbol{W}_{\text{upt}} \in \mathbb{R}^{d\times(n+o)}$和 $\boldsymbol{b}_{\text{upt}} \in \mathbb{R}^d$ 表示全连接层的权重和偏量，$d$ 表示更新节点表示维度，• 为矩阵乘法运算符，⊕ 表示向量加法运算符，$\Psi$为向量拼接函数.

式(6)利用全连接层和非线性转换函数实现了式(2)中的节点更新函数 $Q$. 与中心节点 $v_i$ 的初始特征表示相比，更新后的节点表示 $\boldsymbol{x}_i'$不仅考虑了 $v_i$ 的初始特征，同时也参考了网络邻域中不同邻居节点的特征表示、异质语义关系和重要性，综合融入了语义化同质网络中的多种网络结构特征.

## 4. 2 节点类别识别

将 $\boldsymbol{x}_i'$投影到类别空间，经过 softmax 函数激活后，得到中心节点 $v_i$ 的类别概率分布 $\boldsymbol{P}_i$，如式(7)：

$$\boldsymbol{P}_i = \text{softmax}(\boldsymbol{W}_{\text{cls}} \cdot \boldsymbol{x}_i'). \tag{7}$$

式(7)中，$\boldsymbol{W}_{\text{cls}} \in \mathbb{R}^{d\times K}$ 表示全连接层的权重，$K$ 表示节点类别标签数，• 为矩阵乘法运算符.

对于类别标签向量 $Y = (y_1, y_2, \ldots, y_K)$，根据类别概率分布 $\boldsymbol{P}_i = (p_1, p_2, \ldots, p_K)$，确定概率最大的分量对应的类别标签 $y$ 作为中心节点 $v_i$ 最终的分类结果.

基于异质图神经网络节点分类框架 HNCF 设计异质网络节点分类算法，参见算法 1. 输入为异质网络 $G_{\text{h}}$、对称元路径集合 $M$、节点类型 $A_{\text{c}}$、节点特征表示集合 $X$，输出类别标签集合 $L$.

具体地，算法 1 的第①②行表示初始化类别标签集合 $L$，并利用第 3 节异质网络约简方法，将异质网络 $G_{\text{h}}$ 转换成同质网络 $G_{\text{g}}$. 第③~⑭行表示迭代地计算每个节点 $v_i \in V_{\text{r}}$ 的类别标签 $y$ 的过程. 第⑮行返回算法的计算结果.

其中，第④行表示利用函数 *GetNeighbors* 和关系集合 $E_{\text{g}}$ 获取 $v_i$ 的邻居节点集合. 第⑤~⑧行表示计算邻居节点 $v_j$ 的权重$\alpha_j$和加权特征表示 $\boldsymbol{w}_j$.

第⑨行表示根据邻居的加权特征表示 $\boldsymbol{w}_j$，计算中心节点 $v_i$ 的邻域上下文表示 $\boldsymbol{h}_i$. 第⑩行表示利用 $\boldsymbol{h}_i$ 进行节点表示更新，生成 $v_i$ 的异质语义表示 $\boldsymbol{x}_i'$.

第⑪~⑭行表示根据异质语义表示 $\boldsymbol{x}_i'$，计算类别概率分布 $\boldsymbol{P}_i$，并利用函数 *GetNodeLabel* 获取类别标签 $y_i$，加入到类别标签集合 $L$ 中.

假设$|N|$表示 $V_{\text{r}}$ 中节点的平均邻居数，根据 HNCF 的计算步骤，时间复杂度可表示为$O(m\times|V_{\text{r}}|\times|M|\times|N| + o\times(m+n)\times|V_{\text{r}}|\times|N| + d\times(n+o)\times|V_{\text{r}}|)$，$|V_{\text{r}}|$，$|M|$分别表示节点集合 $V_{\text{r}}$ 和对称元路径集合 $M$ 的大小. 当网络规模较大时，可以重写为$O(|V_{\text{r}}|\times|N|)$，即时间复杂度主要受 $G_{\text{r}}$ 的节点数和平均邻居数影响.

HNCF 的空间复杂度可表示为 $O(n\times|V_{\text{r}}| + |E_{\text{r}}| + m\times|M| + o\times(n+m) + d\times(n+o))$，当网络规模较大时，空间复杂度可以重写为 $O(n\times|V_{\text{r}}| + |E_{\text{r}}|)$，即空间复杂度主要受 $G_{\text{r}}$ 中节点的初始特征表示维度、节点数和关系数影响.

**算法** 1. 异质网络节点分类算法.

输入：异质网络 $G_{\text{h}} = (V_{\text{h}}, E_{\text{h}}, A_{\text{h}}, P_{\text{h}}, \phi_{\text{h}}, \varphi_{\text{h}})$，对称元路径集合 $M$，待分类的节点类型 $A_{\text{c}}$，节点初始特征表示集合 $X$；

输出：节点类别标签集合 $L$.

① $L = \varnothing$； /*初始化类别标签集合*/

/*将异质网络约简为语义化同质网络*/

② $G_{\text{g}} = Reduce(G_{\text{h}}, M, A_{\text{c}})$;

③ for $v_i$ in $V_{\text{g}}$ do

/*迭代处理每个中心节点*/

/*获取 $v_i$ 的邻居节点集合 $N_i$ */

④ $N_i = GetNeighbors(v_i, E_{\text{g}})$;

⑤ for $v_j$ in $N_i$ do

/*迭代处理每个邻居节点*/

/*计算邻居节点 $v_j$ 的权重系数$\alpha_j$ */

⑥ $\alpha_j = \sigma((\boldsymbol{W}_{\text{wgt}} \cdot \Psi(\boldsymbol{x}_i \ominus \boldsymbol{x}_j, \boldsymbol{e}_{i,j})) + b_{\varphi_{\text{r}}(v_i, v_j)})$;

/*计算邻居节点 $v_j$ 的加权特征表示 $\boldsymbol{w}_j$ */

⑦ $\boldsymbol{w}_j = \alpha_j \times \Phi((\boldsymbol{W}_{\text{ngb}} \cdot \Psi(\boldsymbol{x}_j, \boldsymbol{e}_{i,j})) \oplus \boldsymbol{b}_{\text{ngb}})$;

⑧ end for /*终止内层循环*/

/*计算中心节点 $v_i$ 的邻域上下文表示 $\boldsymbol{h}_i$ */

⑨ $\boldsymbol{h}_i = 1 / N_i \cdot \sum_{j\in N_i} \boldsymbol{w}_j$;

/*根据邻域上下文表示更新中心节点表示*/

⑩ $\boldsymbol{x}_i' = \Phi((\boldsymbol{W}_{\text{upt}} \cdot \Psi(\boldsymbol{x}_i, \boldsymbol{h}_i)) \oplus \boldsymbol{b}_{\text{upt}})$;

⑪ $\boldsymbol{P}_i = \text{softmax}(\boldsymbol{W}_{\text{cls}} \cdot \boldsymbol{x}_i')$;

/*计算类别概率分布*/

⑫ $y_i = GetNodeLabel(\boldsymbol{P}_i)$;

/*获取类别标签*/

⑬ $L = L \cup \{y_i\}$;

/*将 $v_i$ 的类别标签加入 $L$ 中*/

⑭ end for /*终止外层循环*/
⑮ return *L*. /*返回输出结果*/

## 5 实验及效果评估

### 5.1 实验数据

为了评估所提方法 HNCF 的有效性，本文使用文献[15]公开发布的 3 个异质网络节点分类数据集：ACM,DBLP,IMDB 进行了实验. 数据集的具体说明如表 2 所示. 对于 DBLP 数据集，由于元路径 APTPA 包含的语义信息较少，本文仅使用元路径 APA 和 APCPA 进行实验. 同时，对于 IMDB 数据集，因为只能获得未经处理的原始数据，本文根据文献[15]的方法进行处理，生成实验数据.

**Table 2 Statistics of Node Classification Datasets**

**表 2 节点分类数据集的统计信息**

| 数据集 | 节点数 | 特征数 | 对称元路径 | 类别数 |
|---|---|---|---|---|
| ACM | 3025 | 1870 | PSP, PAP | 3 |
| DBLP | 4057 | 334 | APA, APCPA | 4 |
| IMDB | 3081 | 1226 | MAM, MDM | 3 |

本文按4∶2∶4的比例将每个数据集划分为训练集、验证集和测试集. 其中，训练集用于训练模型，验证集用于验证模型训练效果，测试集用于评估模型性能. 采用 macro-F1[21]和 balanced accuracy[22]作为评价指标，用于评估节点分类的综合性能和类别不平衡情况下的学习效果.

### 5.2 对比方法

为了验证 HNCF 的有效性，实验选取了 5 种同质和异质图神经网络进行对比：

1）GCN[3]. 同质图神经网络，对邻居节点的特征表示进行均值池化学习节点特征表示.

2）GraphSAGE[7]. 同质图神经网络，对邻域进行采样，并聚合采样邻域的特征. 本实验仅使用 1-hop 邻居，并采用 mean 作为聚合运算符.

3）GAT[8]. 同质图神经网络，在一阶局部子图上执行注意力机制，并为中心节点邻域中的不同邻居指定不同的权重.

4）HAN[15]. 基于 GAT 的异质图神经网络，通过节点级和语义级的注意力的方式，聚合邻居节点的特征来生成节点表示.

5）CompGCN[17]. 异质图神经网络，通过关系类型组合算子对多关系网络进行建模，并结合节点和关系特征表示聚合邻域的语义特征.

HNCF 是本文所提的基于图卷积的异质网络节点分类方法的具体实现框架. 同时，为了评估 1-sum 约束对节点表示学习的影响，本文使用 $\text{HNCF}_{\text{+1-sum}}$ 表示去除度归一化并添加邻居权重 1-sum 约束后的方法. 为了评估异质网络多语义信息的有效性，使用 $\text{HNCF}_{\text{-mult}}$ 表示仅使用单条元路径约简，执行 HNCF 所得到的最优结果. 此外，本文使用 $\text{HNCF}_{\text{-wght}}$ 表示不对邻居节点进行区分、将邻居权重均设为 1 的方法，用于评估邻居权重学习的重要性.

实验基于 PyTorch 和 PyG，实现了 HNCF 和其他对比方法，并利用 Cross Entropy[21]和 Adam[23]优化模型参数. 为了进行公平比较，实验为所有中心节点添加自环以保留中心节点特征信息，并统一设置更新节点表示维度 $d$ 为 64，训练批量大小为 16.

与文献[15]相同，对于 GCN,GraphSAGE,GAT, $\text{HNCF}_{\text{-mult}}$ 方法，本实验对所有元路径的性能进行测试，并展示最优结果. 对于 CompGCN 和 HNCF 方法，在 ACM 数据集上，由于通过 PSP 查找到的邻居过多造成网络退化，本实验对基于 PSP 的邻居节点进行采样，使其邻居数与 PAP 相同.

同时，实验测试了学习率为 0.0001,0.0005,0.001,0.005,0.01 的情况，并使用容忍度为 20 轮的提前终止策略[21]防止模型过拟合. 由于模型参数的随机初始化和提前终止会对最终结果造成一定影响，本文采用每个方法在每个数据集上运行 10 次，取计算结果平均值.

### 5.3 实验结果

表 3 和表 4 分别展示了不同图神经网络方法在 ACM,DBLP,IMDB 数据集上的 macro-F1 和 balanced accuracy 值，可以得出如下几点结论：

首先，HNCF 在 3 个节点分类数据集上取得了最佳性能，验证了本文所提方法的有效性. 具体地，在 IMDB 数据集上，HNCF 的性能大大优于对比方法，分别在 macro-F1 和 balanced accuracy 上提高了 2.42%~5.68%和 2.9%~5.78%. 同时，对于 ACM 和 DBLP 数据集，HNCF 在 maro-F1 上分别提高了 0.7%~1.98%和 0.83%~2.99%，在 balanced accuracy 上分别提升了 0.87%~2.23%和 0.88%~2.27%.

**Table 3 Performance Comparison on macro-F1 / %**

**表 3 不同方法在 macro-F1/%上的对比结果**

| 方法 | 数据集 | | |
|---|---|---|---|
| | ACM | DBLP | IMDB |
| GCN | 92.35 | 89.56 | 65.12 |
| GraphSAGE | 93.00 | 90.29 | 64.29 |
| GAT | 93.05 | 90.60 | 65.08 |
| HAN | 93.63 | 91.72 | 67.55 |
| CompGCN | 92.78 | 91.45 | 66.62 |
| **$\text{HNCF}_{\text{+1-sum}}$** | 92.11 | 90.79 | 67.65 |
| **$\text{HNCF}_{\text{-mult}}$** | 93.91 | 92.42 | 64.73 |

| | | | |
|---|---|---|---|
| **HNCF$_{-wght}$** | 94.08 | 92.08 | 68.93 |
| **HNCF** | **94.33** | **92.55** | **69.97** |

其次，异质图神经网络方法的性能比同质图神经网络方法更好（CompGCN 在 ACM 数据集上的 macro-F1 性能除外），表明异质图神经网络具有更强大的语义表达能力，能够学习更加丰富的异质语义特征，提升节点分类效果．特别地，HNCF 的 maroc-F1 值在 ACM 数据集上比同质图神经网络高 1.28%~1.98%，在 DBLP 数据集上高 1.95%~2.99%，在 IMDB 数据集上高 4.58%~5.68%．同时，HNCF 的 balanced accuracy 值在 ACM 数据集上的提升幅度为 1.68%~2.23%，在 DBLP 数据集上为 1.75%~2.27%，在 IMDB 数据集上为 4.98%~5.78%．此外，分析实验数据发现，同质图神经网络方法 GCN，GraphSAGE，GAT，HNCF$_{-mult}$ 在 ACM，DBLP，IMDB 数据集上的最优对称元路径均相同，分别为 PAP，APCPA，MAM，说明对称元路径的选取是影响节点分类效果的主要因素，但是各方法的实验结果存在较大差异，说明不同方法也会对分类效果造成一定影响．

第三，HNCF 的结果优于 HNCF$_{+1-sum}$ 方法，表明去除 1-sum 约束有利于提升 HNCF 的节点分类效果．由实验结果可知，添加 1-sum 约束后，HNCF 在 ACM，DBLP，IMDB 上的 macro-F1 值分别下降了 2.22%，1.76%，2.32%，balanced accuracy 值分别下降了 2.13%，1.5%，1.58%．同时，实验分析发现，去除 1-sum 约束有利于增大邻居权重和节点表示每个特征维度的方差，从而增加节点的区分能力．

最后，HNCF 的性能要优于 HNCF$_{-mult}$ 和 HNCF$_{-wght}$，表明 HNCF 可以通过异质语义融合和邻居权重学习获得更好的节点表示．仅使用一条对称元路径后，HNCF$_{-mult}$ 无法获取异质语义信息，导致在 ACM，DBLP，IMDB 数据集上的 macro-F1 值分别下降了 0.42%，0.13%，5.24%，在 ACM 和 IMDB 数据集上的 balanced accuracy 值分别下降了 0.38%和 5.36%．从 HNCF$_{-wght}$ 的 macro-F1 值可以看出，学习邻居节点权重对 ACM 和 DBLP 数据集有轻微影响，分别增加了 0.25%和 0.47%．同时，不学习邻居节点权重会对 IMDB 数据集产生较大影响，macro-F1 降低了 1.04%，balanced accuracy 降低了 1.35%．

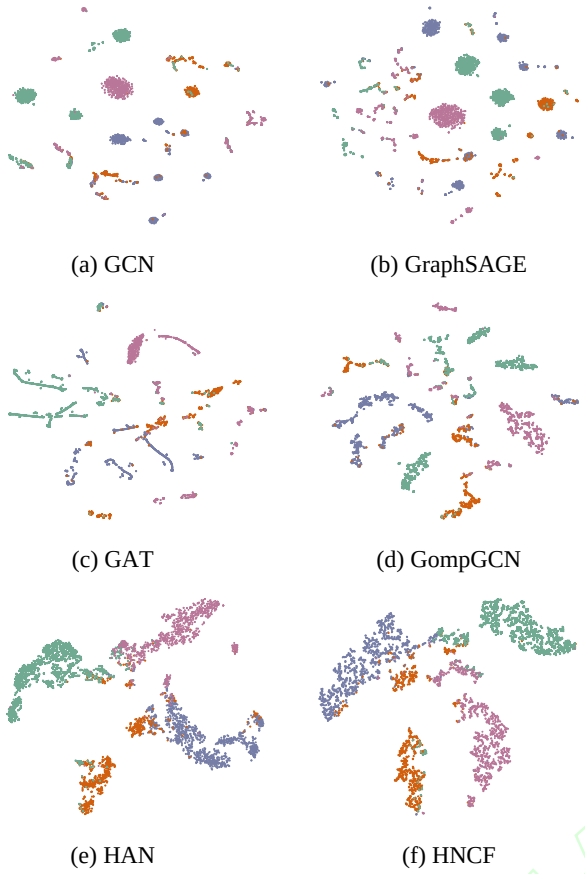
**Table 4 Performance Comparison on balanced accuracy / %**

**表 4 不同方法在 balanced accuracy/%上的对比结果**

| 方法 | 数据集 | | |
|---|---|---|---|
| | ACM | DBLP | IMDB |
| GCN | 92.07 | 90.11 | 63.76 |
| GraphSAGE | 92.62 | 90.63 | 63.88 |
| GAT | 92.52 | 90.49 | 64.56 |
| HAN | 93.43 | 91.50 | 66.64 |
| CompGCN | 92.63 | 91.03 | 65.99 |
| **HNCF$_{+1-sum}$** | 92.17 | 90.88 | 67.96 |
| **HNCF$_{-mult}$** | 93.92 | **92.49** | 64.18 |
| **HNCF$_{-wght}$** | 93.99 | 91.99 | 68.19 |
| **HNCF** | **94.30** | 92.38 | **69.54** |

为了直观地展示 HNCF 所学节点表示的有效性，本实验利用 t-SNE[24]方法对不同方法在 DBLP 数据集上学习的节点表示进行降维，得到 2 维可视化结果如图 4 所示．注意，由于 t-SNE 方法在降维时选择的坐标轴不同，可视化结果中节点类别的相对位置发生了改变，并不影响结果的分析．

由图 4 可知，与同质图神经网络方法相比，异质图神经网络方法可以学习更好的节点表示．GCN,GraphSAGE,GAT 仅考虑异质网络中的单一语义，会使得节点的每个类别都包含多个呈团状或条纹状的簇，导致决策边界不平滑，从而干扰节点类别的预测．在异质图神经网络中，相比于 HAN,HNCF，CompGCN 未考虑邻居节点的重要性，不能有效区分邻居类型，虽然可以将相同类别的节点聚在一起，但是仍然存在相对分散的情况，从而导致错误的识别结果．HAN 和 HNCF 均可以将节点聚成了 4 个几乎完全分离的、在空间中分布良好的簇，因为相同类型的节点集中在特定区域，决策边界更加平滑，从而使得分类结果更好．但与 HAN 不同，HNCF 消除了邻居权重的 1-sum 约束，并利用了多种语义关系特征，移除了部分邻域噪声，减少了节点类别的重叠．



(a) GCN (b) GraphSAGE

(c) GAT (d) GompGCN

(e) HAN (f) HNCF

Fig. 4 Visualization on the DBLP dataset

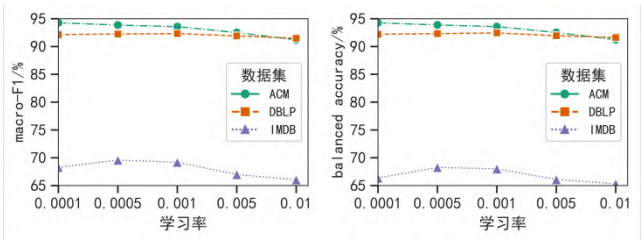
图 4 DBLP 数据集上的节点表示可视化

### 5.4 参数分析

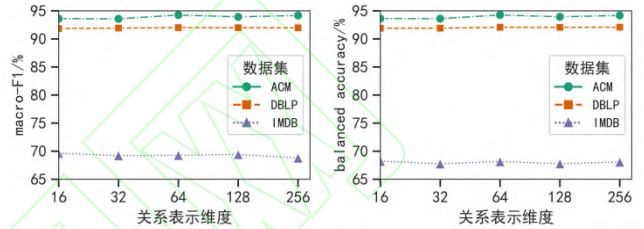
为了评估模型参数的影响，本文对学习率 $r_{\text{learn}}$，关系表示维度 $m$，邻域上下文表示维度 $o$ 和更新节点表示维度 $d$ 进行了参数分析，结果如图 5 所示.

首先，学习率对 HNCF 有着较大的影响，大学习率会导致 HNCF 的性能显著下降. 在学习率较小时，增加学习率会略微降低 ACM 数据集上的性能、略微提升 DBLP 和 IMDB 数据集上的性能，而随着学习率的继续增大，HNCF 在 3 个数据集上的 macro-F1 值和 balanced accuracy 值均出现明显的下降趋势，说明学习率会影响 HNCF 的训练过程和学习效果，学习率较大时会造成 HNCF 训练效果变差，且不同数据的最佳学习率存在差异. 因此，需要仔细调整学习率，以获得最优的节点分类结果.

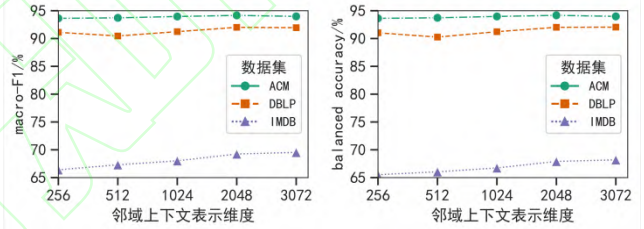
其次，关系表示维度 $m$ 对 HNCF 的影响较小. 随着关系表示维度的增大，HNCF 在 DBLP 数据上的性能未发生较大变化，在 IMDB 数据集上的性能会产生细微幅度的波动，而在 ACM 数据集上的 macro-F1 值和 balanced accuracy 值会略微增加，而后基本保持不变. 由表 2 可知，3 个数据集均包含 2 条对称元路径，在异质网络约简过程中，均会产生 2 种关系类型，并学习对应的关系向量表示. 在关系表示维度 $m$ 较小时，增加 $m$ 有利于区分 2 种不同的关系类型，从而略微提升 HNCF 在 ACM 数据集上的分类效果. 但随着 $m$ 的增加，对关系类型的区分能力基本保持不变，使得 HNCF 对 $m$ 变得不敏感，导致性能变化较小. 因此，在保证足以区分不同关系类型的情况下，关系表示维度 $m$ 的变化对 HNCF 的结果影响较小.



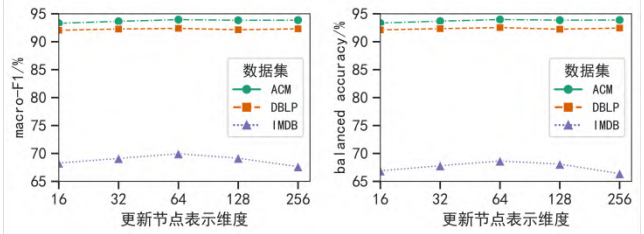

(a) 学习率敏感性分析




(b) 关系表示维度敏感性分析




(c) 邻域上下文表示维度敏感性分析




(d) 更新节点表示维度敏感性分析

Fig. 5 Parameter sensitivity study

图 5 参数敏感性分析

第三，增加邻域上下文表示维度 $o$ 可以改善 HNCF 的节点分类效果. 随着 $o$ 的增大，ACM,DBLP,IMDB 数据集的 macro-F1 值和 balanced accuracy 值总体上均获得了提升. HNCF 在提取邻域上下文表示时，需要加权聚合邻居节点特征，增加 $o$ 的大小可以提升 HNCF 的模型容量，从而学习更多的邻域特征，增强最终的节点分类效果. 因此，可以将邻域上下文表示维度设置为相对较大的值，以确保 HNCF 达到较好的识别效果.

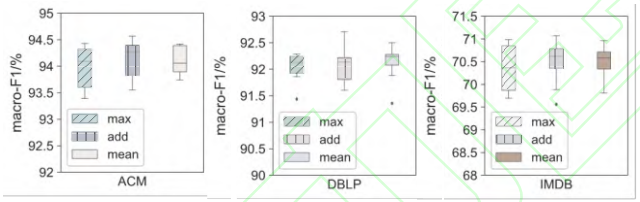
最后，更新节点表示维度 $d$ 会对 HNCF 造成一定影响. $d$ 的增大对 DBLP 数据集的影响不大，但是会略微提升 HNCF 在 ACM 数据集上的性能，并造成

IMDB 数据集上性能的大幅变化. 邻域上下文表示和节点初始特征表示是节点表示更新的主要依据，$d$ 的变化会导致 HNCF 提取不同的特征，进而影响更新节点表示的质量. 因此，需要仔细调整更新节点表示维度，维度过大或过小均会影响节点分类效果.

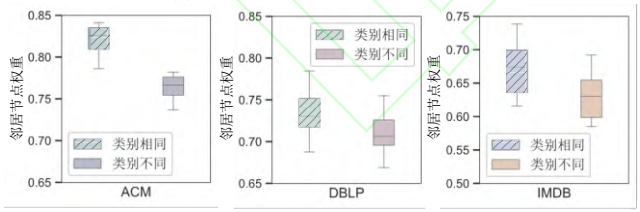
### 5.5 实例分析

为了进一步分析关系表示池化运算符的影响和邻居节点权重学习的效果，本文分别在 ACM,DBLP,IMDB 数据集上运行 10 次 HNCF 方法，并对关系表示池化运算符 max,add,mean 的 macro-F1 值以及按“邻居节点是否与中心节点具有相同的类别”分类的邻居节点权重进行了数据统计，详见图 6.

如图 6(a)所示，不同的关系表示池化运算符对 HNCF 的分类效果有着较大的影响. 在 ACM 数据集上，max 运算符的表现最差，4 分位间距分布在较大的区间内，性能相对不稳定；而与 mean 运算符相比，add 运算符取得了更好的均值和最大值，但也会产生更差的最小值. 在 DBLP 数据集上，add 运算符虽然能取得最好的最大值，但是 4 分位间距分布更大，稳定性不如 max 和 mean 运算符. 在 IMDB 数据集上，add 运算符和 mean 运算符的结果相差不大，均比 max 运算符表现更好. 注意，本文在训练过程中使用了提前终止防止过拟合，且模型参数是随机初始化的，所以在训练不充分的情况下会产生较差的意外结果，导致 max,add,mean 的统计结果中出现离群值.




(a) 关系表示池化运算符统计




(b) 邻居节点权重统计

Fig. 6 Statistics of edge pooling operators and neighbor weights

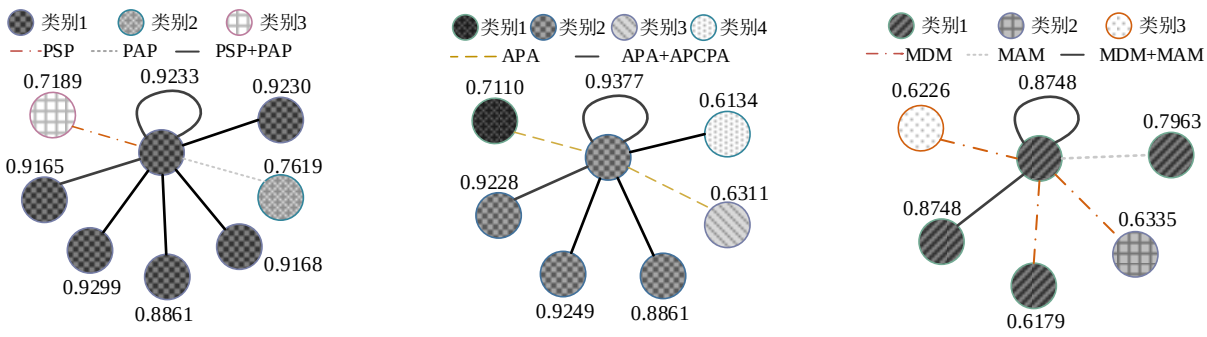
图 6 关系表示池化运算符和邻居节点权重的统计信息

max 运算符倾向于提取最重要的语义关系特征，弱化了相邻节点间不同关系类型组合的差异性，更适用于以某一关系类型为主的单节点多关系异质网络. 而 add 运算符和 mean 运算符综合考虑了关系类型组合中所有的关系类型，在关系类型信息表达能力相差不明显的情况下，性能相对更加鲁棒. 其中，mean 运算符会对所有关系类型表示取平均，移除了关系类型数的影响，而 add 运算符通过直接加和的数值累计保留了关系类型数特征，有利于提升方法的最优性能，但也更易受到关系类型数的影响，多个关系类型表示加和可能会使池化后的异质关系表示在向量空间中发生较大偏移，进而产生较差的节点分类结果.

因此，不同的异质语义融合方式会造成 HNCF 分类性能上的差异，应该仔细选择关系表示池化运算符，以使 HNCF 达到最好的分类效果.

从图 6(b)中可以看出，在 3 个数据集上，相同类别邻居节点的最大值、中位数和最小值均大于不同类别邻居节点，说明 HNCF 可以为相同类别的邻居节点学习较高的权重，从而区分不同类型邻居节点的重要性程度，有利于减轻不重要的邻居节点引入的噪声，使得生成的节点表示更加合理.

为了进一步说明邻居节点权重的学习情况，我们分别在 ACM,DBLP,IMDB 数据集中提取了以节点 676、节点 1054 和节点 2867 为中心节点的单节点多关系异质网络. 如图 7 所示，不同图例的节点表示不同的类别，不同图例的线表示中心节点与邻居节点之间的不同连接关系，不同数字表示不同节点对应的邻居节点权重.




(a) ACM 数据集编号 676 节点权重实例 (b) DBLP 数据集编号 1054 节点权重实例 (c) IMDB 数据集编号 2867 节点权重实例

Fig. 7 Examples of neighbor weight learning

图 7 邻居权重学习实例

从图 7 看出，具有相同类别的邻居节点通常能够学习更高的权重，而与中心节点类别不同的邻居节点权重更低，与图 6(b)的数据相符合．节点的初始特征表示包含了先验的类别信息，HNCF 在学习邻居权重时利用初始特征表示之差$\boldsymbol{x}_i \ominus \boldsymbol{x}_j$作为特征，引入了节点初始特征表示的差异性，相同类别节点间的初始特征表示相对更小，从而提升了 HNCF 的节点区分能力，使邻居权重学习更加合理．

此外，即使邻居节点与中心节点的类别相同，不同类型的连接关系也会导致不同的邻居权重．因为 HNCF 在学习邻居权重时考虑了节点间的异质关系表示 $\boldsymbol{e}_{i,j}$，表达了中心节点与邻居节点间的语义关系特征，使得不同连接关系在异质网络约简后的异质关系表示不同，进而影响邻居节点权重的确定．

因此，HNCF 可以通过学习合理的邻居节点权重，确定邻居节点的重要性，生成更加合理的节点表示．

## 6 总结

本文提出了一种基于图卷积的异质网络节点分类框架 HNCF，利用转换规则将异质网络约简为语义化同质网络，并基于消息传递框架设计图卷积节点分类方法，解决了异质网络上的节点分类问题．不仅简化了异质网络结构，保留了多种语义信息，而且充分利用了网络结构信息，提升了节点分类性能．在 3 个公开的异质网络节点分类数据集上验证了本文所提方法的有效性．未来的研究将扩展 HNCF 到链接预测、节点聚类等下游任务中，以适应更多的异质网络分析与挖掘场景．

## 参 考 文 献


[1] Wu Zonghan, Pan Shirui, Chen Fengwen, et al. A comprehensive survey on graph neural networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24

[2] Zhang Daokun, Yin Jie, Zhu Xingquan. Network representation learning: A survey [J]. IEEE Transactions on Big Data, 2020, 1(6): 3-28

[3] Tomas N K, Max W. Semi-supervised classification with graph convolutional networks [C/OL]//Proc of the 5th Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2017[2020-10-05]. https://arxiv.org/pdf/1609.02907.pdf

[4] Liu Zheng, Xie Xing, Chen Lei. Context-aware academic collaborator recommendation [C]//Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1870–1879

[5] Xu Keyulu, Hu Weihua, Leskovec J, et al. How powerful are graph neural networks? [C/OL]//Proc of the 7th Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2019[2020-10-05]. https://arxiv.org/pdf/1810.00826.pdf

[6] Dong Yuxiao, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks [C]//Proc of the 23rd SIGKDD ACM Int Conf on Knowledge Discovery and Data Mining. New York:ACM, 2017: 135–144

[7] Hamilton W, Ying R, Leskovec J. Inductive representation learning on large graphs [C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 1025–1035

[8] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks [C/OL]//Proc of the 6th Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2018[2020-10-05]. https://arxiv.org/pdf/1710.10903.pdf

[9] Wang Yue, Sun Yongbin, Liu Ziwei, et al. Dynamic graph CNN for learning on point clouds [J]. ACM Transactions on Graphics, 2018, 38(5): 1–12

[10] Gilmer J, Schoenholz S S, Riley P F, et al. Neural message passing for quantum chemistry [C]//Proc of the 34th Int Conf on Machine Learning. New York:ACM, 2017: 1263–1272

[11] Zhang Chuxu, Song Dongjin, Huang Chao, et al. Heterogeneous graph neural network [C]//Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2019: 793–803

[12] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C/OL]//Proc of the 3rd Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2015[2020-10-05]. https://arxiv.org/pdf/1409.0473.pdf

[13] Hu Ziniu, Dong Yuxiao, Wang Kuansan, et al. Heterogeneous graph transformer [C]//Proc of the 20th World Wide Web Conf. New York: ACM, 2020: 2704-2710

[14] Yun Songjun, Jeong M, Kim R, et al. Graph transformer networks [C]//Proc of the 33rd Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2019: 11960-11970

[15] Wang Xiao, Ji Houye, Shi Chuan, et al. Heterogeneous graph attention network [C]//Proc of the 19th World Wide Web Conf. New York: ACM, 2019: 2022–2032

[16] Chen Yiqi, Qian Tieyun, Li Wanli, et al. Exploiting composite relation graph convolution for attributed network embedding[J]. Journal of Computer Research and Development, 2020, 57(8): 1674-1682 (in Chinese)

(陈亦琦, 钱铁云, 李万理, 等. 基于复合关系图卷积的属性网络嵌入方法 [J]. 计算机研究与发展, 2020, 57(8): 1674-1682)

[17] Vashishth S, Sanyal S, Nitin V, et al. Composition-based multi-relational graph convolutional networks [C/OL]//Proc of the 8th Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2020[2021-01-05]. https://arxiv.org/pdf/1911.03082v1.pdf

[18] Wang Minjie, Zheng Da, Ye Zihao, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks [EB/OL]. (2019-04-03)[2021-01-05]. https://arxiv.org/abs/1909.01315

[19] Fey M, Lenssen J E. Fast graph representation learning with PyTorch Geometric[C/OL]//Proc of the 7th Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2019[2021-01-05]. https://arxiv.org/pdf/1903.02428.pdf

[20] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representation of words and phrases and their compositionality [C]//Proc of the 27th Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 3111–3119

[21] Ian G, Bengio Y, Courville A. Deep Learning [M]. Cambridge, MA: MIT Press, 2015: 115-189

[22] Brodersen K H, Ong C S, Stephan K E, et al. The balanced accuracy and its posterior distribution [C]//Proc of the 20th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3121–3124

[23] Kingma D P, Ba J. Adam: A method for stochastic optimization [C/OL]//Proc of the 3rd Int Conf on Learning Representations. Piscateway, NJ: IEEE, 2015[2021-01-05]. https://arxiv.org/pdf/1412.6980.pdf

[24] Maaten L, Hinton G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9: 2579–2605(只有卷)

Xie Xiaojie, born in 1997. MS candidate. His main research interests include information retrieval and machine learning.

谢小杰，1997 年生. 硕士研究生. 主要研究方向为信息检索、机器学习.

Liang Ying ， born in 1962. Senior engineer, senior member of CCF. Her main research interests include data mining, big data process, ubiquitous computing etc..

梁英,1962 年生.高级工程师，CCF 高级会员. 主要研究方向为数据挖掘、大数据处理、普适计算等.

Wang Zisen, born in 1998. MS candidate. His main research interests include information retrieval and machine learning.

王梓森,1998 年生.硕士研究生.主要研究方向为信息检索、机器学习.

Liu Zhengjun, born in 1997. MS candidate. His main research interests include information retrieval and machine learning.

刘政君，1997 年生.硕士研究生.主要研究方向为信息检索、机器学习.

本文校对负责人：王梓森

手机：13127075627

E-mail：wzs_0120@163.com