

基于学者属性与多元路径信息的科研合作预测

马健兵¹, 张同强¹, 黄博²

(1. 成都信息工程大学 计算机学院, 四川 成都 610225; 2. 电子科技大学 四川省人民医院门诊部, 四川 成都 610072)

摘要: 学术合作拓展研究视野, 推动了科学研究的多样性与创新性。近年来, 异构学术网络中基于图表示学习的融合方法展现出良好效果, 但现有合作预测研究往往只利用少量元路径信息, 缺乏对其他类别信息的引入, 导致预测结果在相关类别场景下无法达到最优。引入整体类别信息可全面捕捉不同的合作动机。为处理整体类别信息, 提出一种新颖的层次化元路径融合图神经网络模型 HMFG。该模型从学者个体属性和异构结构信息两方面建模, 通过轻量级均值聚合获取邻居特征, 采用多层特征投影统一元路径维度, 同时结合多头注意力与门控机制完成高质量的语义融合。在 WOS 与 SCHOLAT 数据集上开展实验, 结果表明, HMFG 在两个数据集上均表现出稳定且优异的预测效果, 在基于文本特征构建的 WOS 数据中, ACC、F1 和 AUC 分别实现了 2.29%、2.04% 和 1.08% 的最小性能提升。

关键词: 科研合作预测; 异构学术网络; 图表示学习; 链接预测

DOI: 10.11907/rjdk.251859

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)008-0015-08

扫描二维码阅读全文:



Scientific Collaboration Prediction Based on Researcher Attributes and Multiple Meta-Path Information

MA Jianbing¹, ZHANG Tongqiang¹, HUANG Bo²

(1. School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China; 2. Department of Out-patient, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 610072, China)

Abstract: Academic collaboration broadens research horizons and promotes diversity and innovation in scientific research. In recent years, fusion methods based on graph representation learning in heterogeneous academic networks have demonstrated good performance. However, existing collaborative prediction studies often utilize only a small amount of meta-path information and lack the introduction of other types of information, resulting in suboptimal prediction results in related category scenarios. Introducing overall category information can comprehensively capture different collaboration motivations. To handle overall category information, a novel hierarchical meta-path fusion graph neural network model, HMFG, is proposed. This model models from both scholar individual attributes and heterogeneous structural information, obtains neighbor features through lightweight mean aggregation, unifies meta-path dimensions using multi-layer feature projection, and combines multi-head attention and gating mechanisms to achieve high-quality semantic fusion. Experiments conducted on the WOS and SCHOLAT datasets show that HMFG exhibits stable and excellent predictive performance on both datasets. In the WOS dataset constructed based on text features, ACC, F1, and AUC achieve minimum performance improvements of 2.29%, 2.04%, and 1.08%, respectively.

Key Words: academic collaboration prediction; heterogeneous academic network; graph representation learning; link prediction

0 引言

近年来, 科研合作已成为学术界研究的一种热门趋势。大量研究表明, 科研合作不仅能增加研究成果数量, 而且能提升研究工作质量^[1]。此外, 学术合作有助于推动

科学研究的多元化和创新性, 更好地解决学术问题^[2]。合作发表的学术论文被引用率更高, 也更有利于开放科学发展。科研合作预测在学术界也被称为学术合作推荐或预测, 相关研究主要从学者个体特征或学术网络关系展开。前者侧重于利用学者的属性信息刻画研究兴趣, 并基于相似性为目标学者找到潜在合作者^[3]; 后者则依托网络中的

收稿日期: 2025-12-30

基金项目: 四川省科技计划项目(2023YFS0192)

作者简介: 马健兵(1981-), 男, 博士, 成都信息工程大学计算机学院教授、博士生导师, 研究方向为人工智能知识表示、文本信息处理、智慧气象; 张同强(2000-), 男, 成都信息工程大学计算机学院硕士研究生, 研究方向为知识挖掘与推荐系统; 黄博(1986-), 女, 博士, 电子科技大学四川省人民医院门诊部副教授, 研究方向为慢病防治。本文通讯作者: 张同强。

关系结构,通过网络指标分析、图表示学习等手段建模学者间的关联^[4]。然而,仅依赖历史合作关系信息往往难以充分反映合作形成机制,研究主题相似性、兴趣重叠度以及学科互补性等因素同样会影响合作概率^[5]。如何有效利用多元信息提升合作预测效果,成为该领域的重要研究方向。

1 相关工作

科研合作预测方法主要分为基于内容、基于结构、基于混合编码的方法3类。早期方法仅采用内容或结构信息对潜在合作者进行预测。例如,Rodrigues等^[6]利用TF-IDF表征学者兴趣,通过计算学者之间的兴趣相似度进行科研合作预测,验证了文本特征对合作关系预测的有效性,但难以充分挖掘文本中的深层语义信息。Lu等^[7]采用BERT模型对学者研究兴趣进行语义建模,并应用于科研合作预测任务,相较传统文本表示方法获得了更好的预测效果,但未充分考虑学术网络的结构关系。熊回香等^[8]采用Word2Vec提取论文文本特征,并结合共被引相似度开展科研合作预测,提高了预测准确率,但未考虑学术网络的结构关系。余传明等^[9]采用LINE与Node2Vec相结合的方法学习网络结构表示,融合局部与全局结构信息进行科研合作预测,有效提升了预测性能,但忽略了节点属性信息对合作关系的影响。Yang等^[10]采用异构随机游走方法挖掘多路径语义信息,并应用于科研合作预测,验证了异构网络建模的有效性,但对多元路径语义信息的融合仍不够充分。随着研究的不断深入,科研合作预测逐渐向内容信息与结构信息相融合的方向发展。例如,Tu等^[11]采用卷积神经网络提取学者文本特征,并结合注意力机制建模邻居节点的重要性,通过文本信息与网络结构信息的联合学习进行科研合作预测,取得了较好的预测效果,但其主要针对同构网络,难以充分利用异构网络中的多源语义信息。Yuan等^[12]采用预训练模型获取学者属性特征作为节点初始表示,并结合图神经网络进行科研合作预测,实现了节点属性与网络结构信息的融合,但对异构网络中复杂语义关系的建模能力仍有待提升。以上方法虽然提升了预测效果,但多聚焦于同构图,对异构网络中多源语义与节点属性的联合建模仍关注不足^[13]。

在提升学者合作预测效果的研究中,如何挖掘异构学术网络中的潜在合作动机至关重要。这些动机通常对应网络中的不同元路径,反映了语义层面的合作意图。以往工作多聚焦于少量元路径及其对合作预测的影响。例如,Hu等^[14]在跨领域合作研究中定义了RR、RTR、RTTR三条元路径,分别刻画“重复合作”“主题相同”“跨领域互补”3种动机。Li等^[15]构建了包含作者(A)、论文(P)、组织(O)、出版物(V)4类节点的异构网络,仅采用AOA、APA、APV-PA三种元路径,通过向量相似度计算输出最终合作者推荐结果。Li等^[16]提出的HCSP模型基于LSTM与多头注意

力提取兴趣特征,并融合APA、APPA、APVPA三条元路径进行邻居信息聚合。上述研究均仅考虑少量元路径,未能充分利用异构图的多源信息。为此,本文在网络中引入整体类别信息,并结合学者个人属性,提出一种新颖的层次化元路径融合图神经网络模型HMFG(Hierarchical Meta-path Fusion Graph Neural Network),主要贡献点为:①从学者个人属性和社交网络结构双视角丰富特征,有效缓解了数据稀疏性问题;②构建多信息异构图学术网络,基于元路径分析揭示不同潜在合作动机对科研合作预测结果的影响;③提出一种新颖的层次化元路径融合方法,充分利用多源信息优势来提升预测性能。

2 HMFG模型构建

如图1所示,HMFG模型主要包括以下5个模块:①学术异构图构建;②学者属性嵌入;③基于元路径的邻居属性融合;④语义特征投影;⑤基于Transformer的层次化元路径融合。模型融合学者属性与网络结构信息,实现科研合作关系的有效预测,并为异构图多路径语义建模提供新思路。

2.1 学术异构图构建

为全面刻画科研合作中的多元关系,本文构建了一个学术异构图 $G=(V,E)$,异构图学术网络节点类型包括作者(A)、期刊(J)、关键字(K)以及学术机构(I),边的关系则由作者合作、作者论文发表期刊、作者所属机构、作者涉及关键词主题、关键词共现等学术活动中自然产生的联系构成。基于上述多维结构,本文进一步设计了多条元路径以捕捉潜在的合作动因,其中A-A-A表示作者通过共同合作者间接连接,反映协同网络的传播性;A-K-A表示作者在研究主题上的相似性;A-I-A和A-J-A分别体现了来自相同机构或在相同期刊上发表经历可能带来的合作意愿;A-K-K-A则用于挖掘不同作者在关键词共现基础上的跨学科兴趣交集。通过这一建模方式,网络不仅保留了丰富的语义结构,而且为后续的代表学习与合作预测提供了坚实基础。

2.2 学者属性嵌入

已有研究表明,嵌入学者个人属性能有效提高合作关系的准确性,还能降低网络结构数据的稀疏性^[17]。具体而言,学者的属性信息主要来源于其历史发表论文的标题内容,标题被视作其研究主题的直接体现。每个论文标题 t_i 作为预训练BERT模型的输入,取最后一层[CLS]向量作为特征表示。表示为:

$$H_i = \text{BERT}(t_i)[[\text{CLS}]] \quad (1)$$

式中: $H_i \in R^{768}$ 表示第 i 个标题的语义特征向量。由于每一位作者都至少有3篇以上的论文标题,为综合反映作者的研究主题,对于作者 j ,收集所有标题的特征向量集合 $\{H_1, H_2, \dots, H_N\}$,采用平均池化的策略生成作者特征向量。

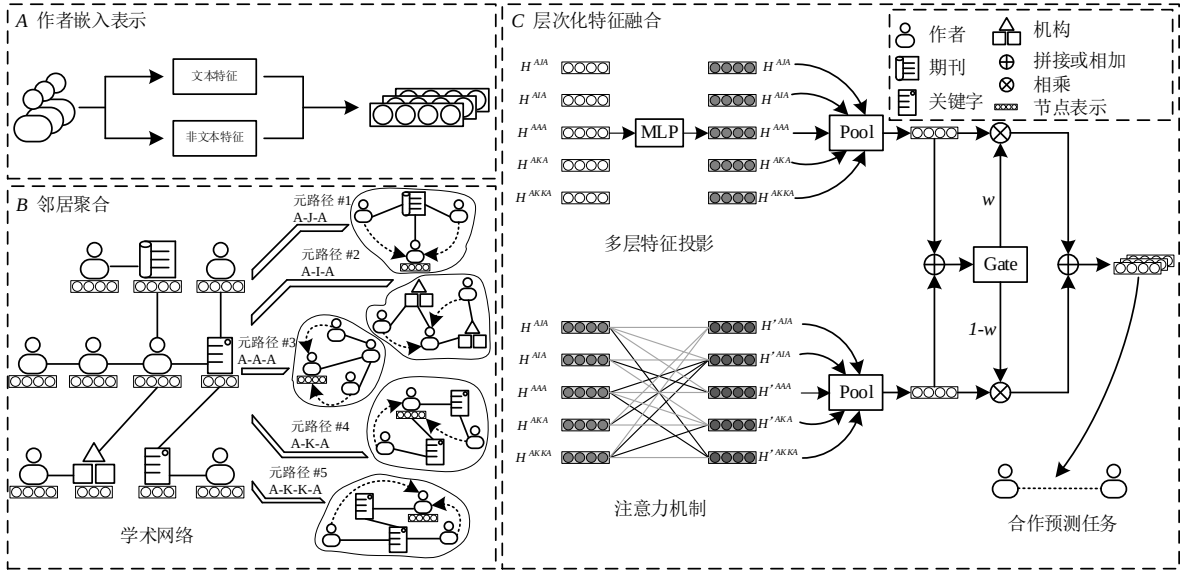


Fig. 1 Framework of HMFG model

图1 HMFG模型框架

表示为:

$$V_j = \frac{1}{N} \sum_{i=1}^N H_i \quad (2)$$

式中: $V_j \in R^{768}$ 表示作者 j 的综合特征向量; N 表示作者 j 过去发表论文的数量。

此外,学者属性不仅包括研究兴趣等文本信息,还包括学术背景和学术影响力等非文本信息。具体来说,学术背景包括学者的历史合作次数、科研基金次数以及发表的论文数量;学术影响力则通过学者的被引用次数、近期论文的被使用次数以及 H-index 等指标进行衡量。这些非文本信息与研究兴趣特征向量相结合可构成更加丰富的学者属性特征向量,表达形式为:

$$A_j = [N_j^{hc}, N_j^{qj}, N_j^{pub}, N_j^{cit}, N_j^{ru}, H_j] \quad (3)$$

$$V_j^{full} = [V_j || A_j] \quad (4)$$

式中: N_j^{hc} 表示历史合作次数; N_j^{qj} 表示获得科研基金项目次数; N_j^{pub} 表示发表论文数量; N_j^{cit} 表示累计被引用次数; N_j^{ru} 表示学者近 180 天论文的使用/访问次数; H_j 表示学者 j 的 H-index。上述非文本信息均可从 WoS 数据库中获取; V_j^{full} 为融合文本与非文本信息的学者完整属性表示,用于评估模型在更加复杂且贴近真实学术环境下的预测性能。

2.3 基于元路径的邻居属性融合

已有研究表明,简单的均值聚合可获得与传统加权求和相近的效果,且在稳定性和计算效率上更具优势^[18]。受此启发,本文定义元路径类型实例的集合为 $S^p = \{p1, p2, p3, \dots, pk\}$, pk 表示某元路径类型的所有实例。对于目标节点 i ,其在该元路径下的邻居特征聚合公式为:

$$x_i = \frac{1}{|N(p^k)|} \sum_{j \in p^k} x_j \quad (5)$$

式中: $|N(p^k)|$ 表示 p^k 元路径类型下与节点 i 相连的邻居集合, x_i 为节点 i 经过该路径聚合后的表示; x_j 为邻居节

点特征向量。

为提升长元路径的计算效率与扩展性,采用基于归一化邻接矩阵的矩阵乘法表示特征传播。记 $S^X = \{X^1, X^2, \dots, X^V\}$, 其中 $X^V \in R^{n_v \times d}$ 表示节点类型特征矩阵, n_v 为节点数, d 为特征维度。边类型集合为 $S^e = \{e_{c1, c2}, e_{c1, c3}, \dots, e_{ci, cj}\}$, $e_{ci, cj}$ 表示异构图中不同边的关系。元路径语义特征传播表示为:

$$P^i = \tilde{A}e_{c1, c2} \tilde{A}e_{c2, c3} \dots \tilde{A}e_{ci, cj} \tilde{A}e_{cj, ck} X^{ck} \quad (6)$$

式中: $\tilde{A}e_{ci, cj} = D^{-1}e_{ci, cj}Ae_{ci, cj}$, $Ae_{ci, cj}$ 为邻接矩阵, D 为度矩阵。

2.4 语义特征投影

为统一一节点特征空间,本研究引入一种由多层堆叠结构构成的特征映射机制。每一层投影模块对元路径特征进行独立线性映射后首先引入 LayerNormlayer 对特征进行归一化处理;其次,通过 PReLU 激活函数增加非线性表达能力;最后,施加 Dropout 以提升模型的泛化能力。通过堆叠多个单层模块可使模型学习更加全面的信息。该过程可形式化表达为:

$$X^{(l+1)} = \text{Dropout}(\text{PReLU}(\text{LayerNorm}(X^{(l)}W^{(l)} + b^{(l)}))) \quad (7)$$

式中: $W^{(l)}$ 和 $b^{(l)}$ 为第 l 层的可学习权重与偏置参数。不同元路径与不同层的参数彼此独立学习,通过堆叠多层该类模块,可获得最终的统一语义表示。公式为:

$$X_{out} = f_L \circ f_{L-1} \circ \dots \circ f_1(X_{in}) \quad (8)$$

式中: f_L 表示第 L 层的投影操作; X_{in} 和 X_{out} 分别表示语义特征投影的初始输入与最终输出结果。多层结构的引入有助于捕捉复杂的语义变化和深层语义关联,为后续的语义融合模块提供统一且高质量的特征表达。

2.5 基于 Transformer 的层次化元路径融合

为更有效地挖掘多元路径之间的语义互补性与交互

关系,本文提出一种基于Transformer架构的层次化元路径特征融合方法,以提升合作预测性能。具体来说,该方法由3个部分组成:一是多头自注意力机制,用于建模元路径之间潜在的依赖关系;二是双重特征池化策略,用于抑制信息冗余并增强全局表示能力;三是引入门控融合网络,用于动态捕捉复杂的语义关系,并进行层次化语义特征融合。设元路径语义特征矩阵集合为 $P = \{P_1, P_2, \dots, P_N\} \in R^{N \times d}$ 。对于每个头 $h \in \{1, \dots, H\}$,分别计算查询、键和值向量。公式为:

$$Q_{P_i}^{(h)} = W_h^Q P_i, K_{P_i}^{(h)} = W_h^K P_i, V_{P_i}^{(h)} = W_h^V P_i, P_i \in P \quad (9)$$

基于缩放点积注意力机制计算路径键的注意力权重。公式为:

$$\alpha^h(P_i, P_j) = \frac{\exp(Q_{P_i}^{(h)} \cdot K_{P_j}^{(h)T} / \sqrt{d/H})}{\sum_{P_i \in P} \exp(Q_{P_i}^{(h)} \cdot K_{P_i}^{(h)T} / \sqrt{d/H})} \quad (10)$$

据此得到第 h 个头的输出。表示为:

$$Attn_h = \sum_{P_i \in P} \alpha^h(P_i, P_j) V_{P_i}^{(h)} \quad (11)$$

将所有注意力头的输出拼接后映射至融合空间。表示为:

$$MultiHead = Concat(Attn_1, \dots, Attn_h) \cdot W^o \quad (12)$$

式中: W_h^Q, W_h^K, W_h^V, W^o 为可学习的参数矩阵, $W^o \in R^{d \times d}$; $\sqrt{d/H}$ 为缩放因子。

通过多头注意力得到不同元路径特征后,采用双重特征池化生成两种全局特征表示。公式为:

$$AttnAvg = \frac{1}{M} \sum_{m=1}^M AttnOut^{(m)} \quad (13)$$

$$FeatAvg = \frac{1}{M} \sum_{m=1}^M X^{(m)} \quad (14)$$

式中: M 表示元路径数量。得到注意力全局特征与原始全局特征后,对双重特征进行拼接并输入门控网络来计算融合权重,据此生成最后的语义融合向量。表示为:

$$W_g = \sigma(MLP([AttnAvg \oplus FeatAvg])) \in [0, 1]^B \quad (15)$$

$$Fused = W_g \odot AttnAvg + (1 - W_g) \odot FeatAvg \quad (16)$$

式中: B 表示批次维度大小; σ 表示Sigmoid函数。通过上述层次化融合过程可获得包含多元语义信息的节点嵌入表示,最终用于作者之间的链路预测二分类任务。

3 实验方法与结果分析

3.1 实验环境与参数设置

本实验CPU为AMD Ryzen Threadripper 3960X 24-Core Processor,内存128 GB, GPU为NVIDIA GeForce RTX 4090(24 GB显存)。模型通过Python 3.11实现,基于PyTorch与PyTorch Geometric(PyG)框架完成训练与测试。在参数设置方面,经过多次实验,模型参数设置如下:节点特征维度统一映射至128维;Dropout设置为0.5,模型采用Adam优化器进行训练,学习率设置为0.000 45,损失函

数采用BCEWithLogitsLoss。训练过程中最大迭代轮数设置为5 000,并采用Early Stopping策略,避免模型过拟合。

在模型优化方面,采用带Logit稳定机制的二元交叉熵损失函数。公式为:

$$L = \frac{1}{|\mathcal{E}|} \sum_{(u,v) \in \mathcal{E}} \max(\hat{y}_{uv}, 0) - y_{uv} \hat{y}_{uv} + \log(1 + \exp(-|\hat{y}_{uv}|)) \quad (17)$$

式中: $y_{uv} \in \{0, 1\}$ 表示学者之间真实的合作关系; $\mathcal{E}$ 为合作关系边的集合; $\hat{y}_{uv}$ 表示学者之间预测合作的概率。

采用内积作为解码函数,计算作者 u 与 v 之间的合作概率。计算公式为:

$$P(u, v) = \sigma(Z_u^T Z_v) \quad (18)$$

3.2 实验数据

本研究使用的数据来源于Web of Science Core Collection。通过设置检索条件WC=Information Science AND CU=PEOPLES R CHINA,共获得初始文献记录179 379条。在此基础上,限定文献类型为Article,研究领域为Computer Science Information Systems,时间范围为2009–2023年,出版机构为IEEE、Elsevier、Springer Nature、MDPI与Wiley-Hindawi后,得到106 626篇文献。剔除缺失关键信息的文献后,最终保留105 622篇有效文献。为解决作者重名问题,参考文献[16]方法,结合机构与ORCID信息消歧,并保留发表论文不少于3篇的作者。最终得到作者(A)13 623名、论文(P)48 085篇、期刊(J)94种、机构(I)2 279个、关键词(K)经聚类归并为88 032个语义类别。据此构建含4类节点(A、J、I、K)与5类边(A-A、A-J、A-I、A-K、K-K)的异构学术图,并对边去除重复与自环。同时,引入国内知名学者网(SCHOLAT)公开数据集集中的SCHOLAT Multiplex Network。该数据集包含丰富的学者社交关系、团队属性及多维度动态学术交互信息,能全面评估模型在真实学术社交场景下的泛化能力与鲁棒性。为适应异构图模型的输入需求,对SCHOLAT Multiplex Network进行预处理,最终得到作者(A)、团队(T)、课程(C)3类节点,以及作者和作者、作者和团队、作者和课程三类边。数据集统计信息见表1。

Table 1 Dataset statistics

表1 数据集统计信息

数据集	总节点数量	节点类型	总边数	边类型	作者节点数
WOS	104 008	4	1 125 429	5	13 623
SCHOLAT	16 838	3	415 707	3	2 302

本文研究任务为异构学术网络中的科研合作预测,因此仅将作者—作者(A-A)合作边作为目标关系进行数据集划分,作者—论文、作者—关键词、作者—机构等其他类型边均保留在异构图中,用于节点表示学习,不参与划分。按照8:1:1的比例将数据集划分为训练集、验证集和测试集。其中,正样本为网络中真实存在合作关系的作者对;负样本为网络中不存在合作关系的作者对,通过随机负采样方式构建,实验中正负样本比例设置为1:1。

3.3 评价指标

采用 AUC (Area Under Curve)、F1 值和准确率 (Accuracy, ACC) 3 种指标评估模型的链路预测性能。其中, AUC 用于衡量模型对正负样本的区分能力。计算公式为:

$$AUC = \frac{n^+ + 0.5n^+}{N} \quad (19)$$

式中: n^+ 表示正样本得分高于负样本的对数; n^+ 表示得分相等的对数; N 表示总样本数。

F1 值用于评估模型降低误报和漏报的综合能力。计算公式为:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (20)$$

式中: Precision 为精确率; Recall 为回收率。

ACC 为模型准确率。计算公式为:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

式中: TP 表示真实存在合作关系且被模型正确预测为存在合作关系的作者对数量; TN 表示真实不存在合作关系且被模型正确预测为不存在合作关系的作者对数量; FP 表示真实不存在合作关系但被模型错误预测为存在合作关系的作者对数量; FN 表示真实存在合作关系但被模型错误预测为不存在合作关系的作者对数量。

3.4 基线模型

从随机游走方法、无元路径方法、基于元路径的方法 3 个类别中选择 9 个具有代表性的基线模型与本文方法进行比较, 分别为: ① DeepWalk^[19]。其对每个节点执行多次随机游走, 生成多条路径序列, 最终结合不同路径信息学习节点表示, 用于下游任务; ② Metapath2vec^[20]。其基于 DeepWalk 思想, 引入元路径指导随机游走过程, 使模型能够进行语义一致的路径采样, 提升节点表示的语义表达能力; ③ HetSANN^[21]。其基于类型特定的注意力机制和多层 GAT 结构设计, 针对不同关系类型选择不同的注意力函数计算, 并通过多层 GAT 拓展感受野, 聚合多跳节点信息; ④ HGT^[22]。其采用基于 Transformer 的类型感知注意力机制, 无需显式定义元路径信息, 根据节点类型和边类型捕捉复杂语义, 完成节点特征更新; ⑤ DiffGraph^[23]。其通过潜在异构图的前向与反向扩散实现去噪与跨类型语义转移, 将辅助异构数据映射到目标语义空间, 提升节点表示质量; ⑥ HAN^[24]。其将注意力机制引入异构图神经网络, 并设计分层结构, 从邻居节点注意力和语义注意力进行融合节点特征, 取得了较好效果; ⑦ SeHGNN^[25]。其在预处理阶段采用均值聚合的方式提取元路径语义特征, 并将元路径映射为统一维度, 采用基于 Transformer 的注意力机制融合不同语义信息; ⑧ MECCH^[26]。其定义了新的图形结构 Metapath 上下文, 采用图池化函数作为 Metapath 上下文编码器, 并通过一维卷积核融合不同元路径的语义特征; ⑨ NoiseHGNN^[27]。其通过节点特征相似性构建高阶图, 并基于该合成图提取目标图, 最大化两者的结构一致

性, 增强节点表示能力。

3.5 实验结果与分析

3.5.1 比较实验

各模型在 WOS 和 SCHOLAT 数据集上的表现见表 2。综合 ACC、F1 值和 AUC 3 项指标来看, 本文模型展现出优越性能, 与对照模型相比分别实现 2.29%、2.04%、1.08% 的最小提升。SeHGNN 与 MECCH 亦表现稳定, 具备较强的竞争力。整体来看, 基于元路径的方法普遍优于不依赖元路径的方法, 而后者又优于传统的随机游走方法。需要指出的是, 基于元路径的方法通常依赖预定义的元路径集合, 对任务理解和领域知识要求较高。在 SCHOLAT 数据集上的实验结果表明, 在包含更复杂社会关系与组织结构的真实学术社交场景中, 本文模型在 AUC、F1 值和 ACC 等评价指标上均取得最优性能, 证实了其在真实学术社交网络中的泛化能力与鲁棒性。

Table 2 Performance comparison of each models on different datasets

表 2 各模型在不同数据集上的性能比较

模型	WOS			SCHOLAT		
	AUC	F1 值	ACC	AUC	F1 值	ACC
Deepwalk	0.762 5	0.711 8	0.699 5	0.778 9	0.723 6	0.707 6
Metapath2vec	0.806 5	0.741 5	0.727 5	0.818 3	0.757 3	0.734 7
HGT	0.842 0	0.771 3	0.753 0	0.854 4	0.780 6	0.762 9
HetSANN	0.909 4	0.790 0	0.737 8	0.896 5	0.770 0	0.726 6
DiffGraph	0.903 0	0.797 7	0.752 4	0.908 3	0.778 1	0.723 9
HAN	0.814 9	0.789 0	0.794 7	0.826 3	0.756 5	0.784 3
SEHGNN	0.910 3	0.837 5	0.819 0	0.891 4	0.822 8	0.815 7
MECCH	0.912 5	0.837 0	0.822 0	0.903 8	0.822 0	0.800 8
NoiseHGNN	0.920 2	0.824 3	0.800 2	0.869 1	0.784 4	0.751 7
HMFG	0.931 0	0.857 9	0.844 9	0.913 9	0.837 6	0.827 2

仅基于 BERT 提取的论文标题文本特征维度过于单一, 难以刻画复杂多变的现实需求, 因此本文引入学术背景与学术影响力等非文本属性丰富作者特征, 实验结果如表 3 所示。可以看出, 在 WOS 数据集上, 属性强化使多数模型的性能指标有所提升, 本文模型仍保持最优性能, 验证了其对深层合作动机的理解能力以及在复杂属性条件下的有效性。

为进一步验证模型在真实场景下的识别精度与抗噪能力, 本研究将正负样本比例调整为 1:10, 并在 WOS 数据集上开展实验, 结果如表 4 所示。在非平衡采样条件下, 由于正负样本比例严重失衡, F1 值和 ACC 指标不再适用, 仅以 AUC 衡量模型性能。实验结果表明, 多数模型的 AUC 与基线相比波动轻微, 显示出对样本分布变化的稳健性, 其中 HMFG 仍保持最高 AUC 的值 (0.913 9), 整体性能优势明显, 展示了其在海量候选合作对中的抗噪能力与真实识别精度。

3.5.2 元路径信息融合分析

为验证多信息融合策略的有效性, 评估各模型在融合不同数量元路径时的整体表现。对相同数量元路径组合下的各项指标取平均, 以衡量该数量级下模型的整体性

Table 3 Comparison of model performance after incorporating non-textual attributes

表3 加入非文本属性后各模型性能比较

模型	WOS		
	AUC	F1 值	ACC
Deepwalk	0.771 2	0.727 5	0.711 0
Metapath2vec	0.822 1	0.763 0	0.739 1
HGT	0.858 1	0.776 1	0.755 1
HetSANN	0.893 8	0.786 5	0.735 7
DiffGraph	0.934 1	0.795 3	0.747 6
HAN	0.844 9	0.797 7	0.805 3
SEHGNN	0.913 3	0.839 1	0.821 1
MECCH	0.915 9	0.841 1	0.825 1
NoiseHGNN	0.923 4	0.825 0	0.803 1
HMFG	0.934 8	0.861 4	0.850 9

能。3条元路径组合共有10种,4条元路径组合共5种,5条元路径组合仅1种,相关实验结果如图2所示。从平均

Table 4 Results of imbalanced sampling experiment

表4 非平衡样本实验结果

模型	非平衡样本(WOS)	
	AUC	差值
Deepwalk	0.761 0	0.001 5
Metapath2vec	0.790 1	0.016 4
HGT	0.828 9	0.013 1
HetSANN	0.873 7	0.035 7
DiffGraph	0.881 6	0.021 4
HAN	0.789 5	0.025 4
SEHGNN	0.892 1	0.018 2
MECCH	0.884 1	0.028 4
NoiseHGNN	0.891 3	0.028 9
HMFG	0.913 3	0.017 7

性能变化趋势来看,随着元路径特征的增多,模型各指标呈现上升趋势,说明融合更多元路径信息有助于提升模型的整体表现。

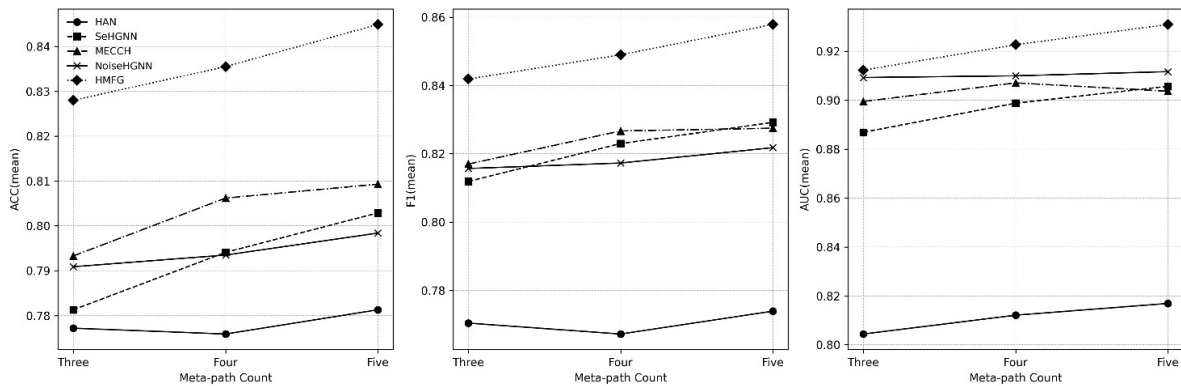


Fig. 2 The average performance of each model under different numbers of meta-paths

图2 不同元路径数量下各模型的平均性能

各模型在不同元路径数量下的最佳性能如图3所示。可以看出,各模型在不同数量路径下的最佳性能与其平均性能之间可能存在一定差异。例如,在4条元路径组合下,Se-

HGNN与MECCH模型的平均ACC和F1值与其各自的最佳性能结果相反。相比之下,本文模型的最佳性能与平均性能指标变化趋势保持一致,展现出良好的稳定性。

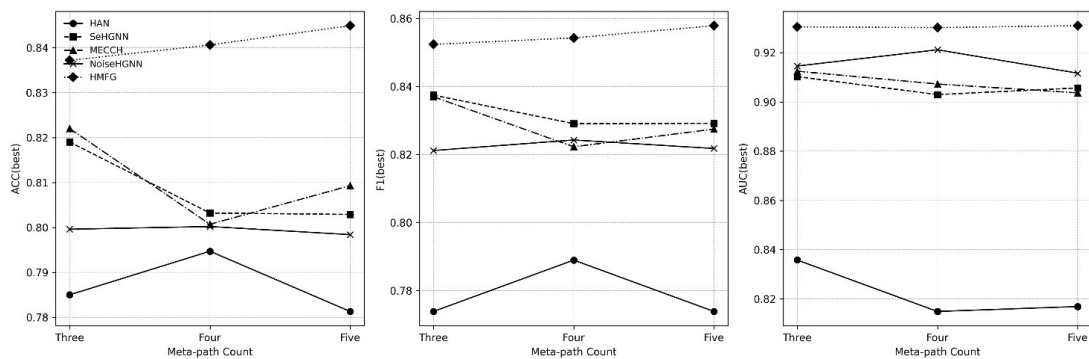


Fig. 3 The optimal performance of each model under different numbers of meta-paths

图3 各模型在不同元路径数量下的最佳性能

为深入分析元路径融合策略的有效性,表5展示了模型在核心路径(AJA、AIA、AKA、AKKA)及其他路径下的表现,结果表明其不同路径下均能保持稳定性能,且整体信息路径普遍较其子路径组合更具优势。核心组合下的子路径组合表现稳定,ACC维持在0.830 9~0.837 2的较高

水平。值得注意的是,部分包含AAA路径的组合同样具有竞争力,甚至在部分指标上优于核心组合。例如,AAA\AKA\AJA的ACC为0.835 4,高于AKA\AIA\AKKA的0.830 9,说明AAA元路径特征具有重要的语义补充作用。

Table 5 Prediction performance of core path and other path combinations for HMFG

表 5 HMFG 核心路径及其他路径组合的预测性能			
元路径	ACC	F1 值	AUC
AAA\AJA\AKA\AKKA	0.835 7	0.849 1	0.923 8
AAA\AJA\AIA\AKA	0.835 7	0.848 5	0.922 0
AAA\AJA\AIA\AKKA	0.834 2	0.847 7	0.920 1
AAA\AIA\AKA\AKKA	0.831 2	0.845 0	0.917 1
AKA\AJA\AIA	0.837 2	0.852 4	0.930 5
AKA\AJA\AKKA	0.834 5	0.848 3	0.925 1
AJA\AIA\AKKA	0.836 9	0.847 5	0.912 3
AKA\AIA\AKKA	0.830 9	0.845 5	0.924 1
AAA\AKA\AJA	0.835 4	0.847 1	0.914 8

各模型性能最佳的元路径组合如表6所示。可以看出,不同模型对元路径信息的敏感性和融合策略存在差异,从侧面印证了在复杂异构网络中,不同类型元路径所承载的语义信息均不可或缺。尽管各模型在3条元路径组合下表现最优的路径集各不相同,但这些组合通常是在其4条路径融合中表现最佳路径组合的子集。此外,个别少路径组合的表现优于多路径组合,表明部分元路径之间存在语义干扰,信息融合不当可能抑制有效特征表达。相比之下,本文模型在不同数量元路径融合情境下均保持稳定优势,且在引入更多信息的情况下取得最优结果,反映出模型具备较强的融合能力与路径鲁棒性。

Table 6 Best meta-path combinations for model performance

表 6 各模型性能最佳的元路径组合			
模型	3 条元路径组合	4 条元路径组合	5 条元路径组合
HAN	AAA\AKA\AKKA	AAA\AJA\AKA\AKKA	AAA\AJA\AIA\AKA\AKKA
NoiseHGNN	AAA\AKA\AKKA	AAA\AJA\AKA\AKKA	AAA\AJA\AIA\AKA\AKKA
SEHGNN	AJA\AIA\AKKA	AAA\AJA\AIA\AKKA	AAA\AJA\AIA\AKA\AKKA
MECCH	AAA\AIA\AKKA	AAA\AJA\AIA\AKKA	AAA\AJA\AIA\AKA\AKKA
HMFG	AKA\AJA\AIA	AJA\AIA\AKA\AKKA	AAA\AJA\AIA\AKA\AKKA

3.5.3 可解释性分析

为验证层次化融合的有效性,对 HMFG 模型的门控权重进行可视化分析,结果如图4所示。可以看出,大部分作者的权重集中在0.6~0.9之间,说明注意力特征占主导且信息量更丰富;所有作者均采用两种特征的混合,没有大量作者处于 $W=0$ 或 $W=1$ 的极端情况,表明原始特征仍能提供有效基线信息。不同作者的融合权重存在差异,体现了层次化特征融合的灵活性。总体来看, HMFG 并非单纯依赖注意力机制,而是通过自适应融合在注意力特征与原始特征之间找到最佳平衡,增强了表达能力与预测稳定性。

为进一步分析不同元路径在科研合作预测中的贡献,基于 HMFG 多头注意力机制学习到的元路径权重进行全局聚合,得到各类路径的整体重要性分布,具体如图5所示。可以看出,不同元路径在模型决策中的贡献存在显著差异。其中,AAA 路径权重最高,表明基于共同合作者的结构邻近性在合作预测中起主导作用;AIA 权重较低,说明机构共属关系区分能力有限。高阶组合路径,如 AKKA 权重进一步下降,表明过长路径易引入冗余信息,对合作

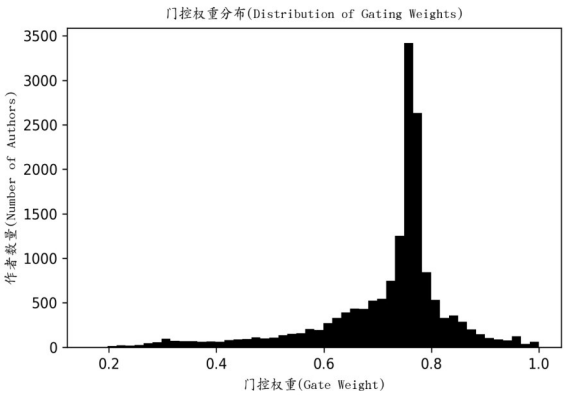


Fig. 4 Gate weight distribution

图 4 门控权重分布

建模增益有限。总体而言,中国信息科学领域的科研合作网络主要依赖于作者间的结构关联以及研究兴趣的相似性,而非机构归属因素。

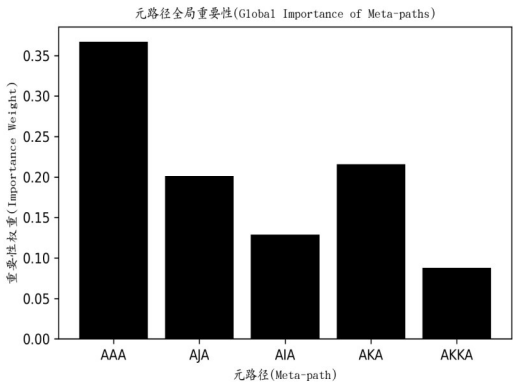


Fig. 5 Global importance of meta-paths

图 5 元路径全局重要性

3.5.4 消融实验

为验证语义融合模块各组成部分的有效性,设计3种变体进行消融实验,其中 Variant#1 移除了注意机制模块和门控网络模块,仅使用语义投影后原始的元路径特征; Variant#2 保留多头注意力融合特征,移除原始全局特征和门控网络模块; Variant#3 保留池化后的两个全局特征,去掉门控网络模块,对两个全局特征进行等权平均融合。如表7所示, Variant#1 性能最低,说明原始语义特征难以充分表达多路径间复杂关系; Variant#2 和 Variant#3 实验结果均显示多头注意力机制对模型性能有显著促进作用,且 Variant#3 整体表现优于 Variant#2,表明在注意力机制的基础上引入双通道信息融合是有效的。完整模型的3项指标超越所有变体,且训练过程更加稳定,表明门控网络在信息选择与降噪过程中发挥着关键作用。

Table 7 Results of ablation study

表 7 消融实验结果			
消融变体	ACC	F1 值	AUC
Variant#1	0.813 2	0.837 4	0.920 0
Variant#2	0.837 8	0.849 7	0.928 8
Variant#3	0.839 4	0.850 0	0.918 9
HMFG	0.844 9	0.857 9	0.931 0

4 结语

本文针对科研合作预测中合作动机复杂、关系信息多样等问题构建了多信息异构图学术网络,并提出一种基于学者属性与多信息元路径融合的科研合作预测模型 HM-FG。具体而言,该模型利用 BERT 提取学者历史论文标题的兴趣特征,采用元路径引导的邻居聚合与多层语义投影,并在融合阶段引入多头注意力、双通道池化与门控机制,有效提升了模型的语义融合能力与表达力。在 WOS 与 SCHOLAT 两类数据集上进行了系统验证,证实 HMFG 模型在不同元路径组合下均保持稳定且优异的预测性能,显著优于主流基线模型,并在引入整体类别信息时取得最佳结果,展现出良好的语义融合效果与路径鲁棒性。相关实验结果揭示了合作动机结构,如期刊、机构与研究兴趣互补等多层语义特征对预测的贡献。然而,本研究仍存在拓展空间。例如,虽然预定义多类型元路径在实验中取得了较好效果,但元路径构建仍依赖一定的领域先验知识,缺乏自动化与自适应能力。未来将探索自适应元路径建模策略,如路径重要性学习或自动化搜索机制,以降低人工先验依赖。此外,本文主要基于静态学术网络开展研究,研究者兴趣与合作关系具有明显的时间依赖性,未来可引入时间感知的异构图或动态图神经网络,以提升预测结果的前瞻性和实用价值。

参考文献:

- [1] ABRAMO G, D'ANGELO C A, DI COSTA F. Research collaboration and productivity: is there correlation?[J]. Higher Education, 2009, 57(2): 155-171.
- [2] ABRAMO G, D'ANGELO C A, DI COSTA F. Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications[J]. Journal of the American Society for Information Science and Technology, 2012, 63(11): 2206-2222.
- [3] HU Z, ZHOU J, WEI W, et al. Predicting cross-domain collaboration using multi-task learning[J]. Expert Systems with Applications, 2024, 255: 124570.
- [4] SUGIYAMA K, KUMAR T, KAN M Y, et al. Identifying citing sentences in research papers using supervised learning[C]//2010 International Conference on Information Retrieval & Knowledge Management, 2010: 67-72.
- [5] JEH G, WIDOM J. Simrank: a measure of structural-context similarity[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002: 538-543.
- [6] RODRIGUES M W, BRANDAO W C, ZÁRATE L E. Recommending scientific collaboration from researchgate[C]//2018 7th Brazilian Conference on Intelligent Systems (BRACIS), 2018: 336-341.
- [7] LU Y, YUAN M, LIU J, et al. Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion[J]. Scientometrics, 2023, 128(2): 1367-1393.
- [8] XIONG H X, LI X M, DU J. Scholar recommendation based on academic keywords and co-citation[J]. Journal of Intelligence, 2021, 40(7): 725-733.
熊回香, 李晓敏, 杜瑾. 基于学术关键词与共被引的学者推荐研究[J]. 情报学报, 2021, 40(7): 725-733.
- [9] YU C M, LIN A C, ZHONG Y C, et al. Research on scientific collaboration recommendation based on network representation learning[J]. Journal of Intelligence, 2019, 38(5): 500-511.
余传明, 林奥琛, 钟韵辞, 等. 基于网络表示学习的科研合作推荐研究[J]. 情报学报, 2019, 38(5): 500-511.
- [10] YANG C, LIU T, CHEN X, et al. HNRWalker: recommending academic collaborators with dynamic transition probabilities in heterogeneous networks[J]. Scientometrics, 2020, 123(1): 429-449.
- [11] TU C, LIU H, LIU Z, et al. Cane: context-aware network embedding for relation modeling[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1722-1731.
- [12] YUAN C, HE Y, LIN R, et al. Graph embedding for scholar recommendation in academic social networks[J]. Frontiers in Physics, 2021, 9: 768006.
- [13] SONG X, ZHANG Y, PAN R, et al. Link prediction for statistical collaboration networks incorporating institutes and research interests[J]. IEEE Access, 2022, 10: 104954-104965.
- [14] HU Z, ZHOU J, WEI W, et al. Predicting cross-domain collaboration using multi-task learning[J]. Expert Systems with Applications, 2024, 255: 124570.
- [15] LI H, HU Y. Metapath and attribute-based academic collaborator recommendation in heterogeneous academic networks[J]. Scientometrics, 2024, 129(7): 4295-4315.
- [16] LI X, WANG M, LIU X. Predicting collaborative relationship among scholars by integrating scholars' content-based and structure-based features[J]. Scientometrics, 2024, 129(6): 3225-3244.
- [17] JIN B, ZHANG Y, ZHU Q, et al. Heterformer: Transformer-based deep node representation learning on heterogeneous text-rich networks[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023: 1020-1031.
- [18] ZHANG W, YIN Z, SHENG Z, et al. Graph attention multi-layer perceptron[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022: 4560-4570.
- [19] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014: 701-710.
- [20] DONG Y, 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, 2017: 135-144.
- [21] HONG H, GUO H, LIN Y, et al. An attention-based graph neural network for heterogeneous structural learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 4132-4139.
- [22] HU Z, DONG Y, WANG K, et al. Heterogeneous graph transformer[C]//Proceedings of the Web Conference, 2020: 2704-2710.
- [23] LI Z, XIA L, HUA H, et al. DiffGraph: heterogeneous graph diffusion model[C]//Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, 2025: 40-49.
- [24] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference, 2019: 2022-2032.
- [25] YANG X, YAN M, PAN S, et al. Simple and efficient heterogeneous graph neural network[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 10816-10824.
- [26] FU X, KING I. MECCH: metapath context convolution-based heterogeneous graph neural networks[J]. Neural Networks, 2024, 170: 266-275.
- [27] XIONG Z, XIE C, DUAN H, et al. NoiseHGNN: synthesized similarity graph-based neural network for noised heterogeneous graph representation learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 21725-21733.

(责任编辑:尹晨茹)