计算机科学
COMPUTER SCIENCE

基于异构网络表征学习的作者学术行为预测

黄丽，朱焱，李春平

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900078

基于异构网络表征学习的作者学术行为预测

黄 丽[1] 朱 焱[1] 李春平[2]
1 西南交通大学计算机与人工智能学院 成都 611756
2 清华大学软件学院 北京 100091
(793275643@qq.com)

摘 要 作者学术行为预测旨在从异构学术网络中挖掘作者的行为关系,以促进科研合作,产出高水平、高质量的研究成果。现有的节点表示方法大多未考虑节点的语义特征、内容特征、全局结构等,难以有效学习网络中节点的低维特性。为有效融合节点的多维特征和全局结构,提出了一种集成BiLSTM、注意力机制和聚类算法的异构网络表示学习方法HNEMA,以提高学术网络中作者的学术行为预测效果。HNEMA首先基于BiLSTM和注意力机制融合节点的多维特征,聚合同一元路径下或不同元路径下相同类型的邻居,随后聚合待表征节点的所有邻居的多维特征。基于此,采用聚类算法捕获节点的全局特征,从而全面有效地学习节点的低维特性。在全面特征学习的基础上,应用逻辑回归分类器预测作者的学术行为。在3个公开数据集上的验证实验结果表明,相比其他方法,HNEMA在AUC和F1指标上都有一定程度的提升。

关键词:异构网络;网络表征学习;链接预测;元路径
中图法分类号 TP183

Author's Academic Behavior Prediction Based on Heterogeneous Network Representation Learning

HUANG Li[1], ZHU Yan[1] and LI Chun-ping[2]
1 School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
2 School of Software, Tsinghua University, Beijing 100091, China

Abstract The author's academic behavior prediction aims to mine the behavioral relationships of authors from heterogeneous academic networks to promote scientific research cooperation and produce high-level and high-quality research results. Most of the existing methods of node representation learning do not consider the semantic feature, content feature, global structure of the node, etc. It is difficult to effectively learn the low-dimensional characteristics of the node in the network. In order to effectively integrate the multi-dimensional features and global structure of nodes, a heterogeneous network representation learning method (HNEMA) that integrates BiLSTM, attention mechanism and clustering algorithm is proposed to improve the predictive effect of author's academic behavior. HNEMA first integrates the multi-dimensional features of nodes based on BiLSTM and attention mechanism, aggregates the same type of neighbors on the same meta-path or different meta-paths, and then aggregates the multi-dimensional features of all neighbors of the node to be characterized. Based on this, a clustering algorithm is used to capture the global features of the node, so as to comprehensively and effectively learn the low-dimensional characteristics of the node. On the basis of comprehensive feature learning, logistic regression classifier is used to predict author's academic behavior. Validation experiments on three public datasets show that HNEMA has a certain degree of improvement in AUC and F1 indicators compared to other methods.

Keywords Heterogeneous network, Network representation learning, Link prediction, Meta-path

1 引言

学术社交网络蕴含着丰富的信息,越来越多的学者据此搜索相关文献、了解研究趋势、关注水平突出的研究机构或个人、开展学术交流合作等。根据网络结构,社交网络可分为同构网络和异构网络,同构网络仅包含一种类型的节点且节点间只存在一种类型的关系,例如图1中两个红色圈分别表示学术网络中的合作者网络和引用网络,它们都只包含一种作者类型和论文类型的节点,只有一种合作类型和引用类型的关系;而异构网络包含丰富的结构和语义信息,除了包含

到稿日期:2021-09-10 返修日期:2022-01-25
基金项目:四川省科技计划项目(2019YFSY0032)
This work was supported by the Sichuan Province Science and Technology Project(2019YFSY0032).
通信作者:朱焱(yzhu@swjtu.edu.cn)

多种类型的节点外，还考虑了不同类型节点之间的相互影响。图 1 描绘了一个异构学术网络，包含作者、文章、论文发表载体 3 种类型的节点，以及节点之间的多种关系，如发表、合作、引用等。异构学术网络包含丰富的节点及节点之间的关联关系，有利于提高节点间多种关系预测的准确率。预测异构学术网络中作者的学术行为，旨在挖掘作者之间潜在的、未来可能出现的关系（链接），如可能的合作或文章引用。有效的学术行为关系预测能够帮助学者快速、准确地找到学术资源，从而提高学者的科学创新和知识发现行动效率，实现科研效率高效化。

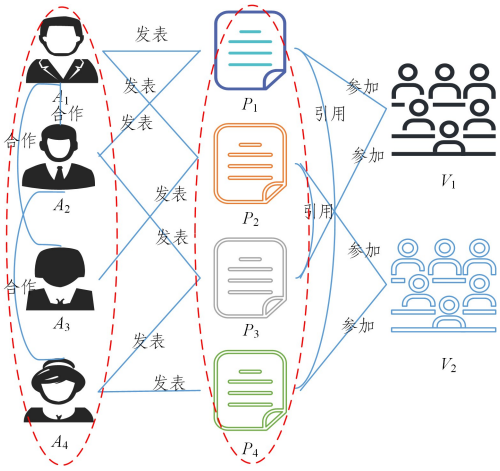


图 1　异构学术网络（电子版为彩图）

Fig. 1　Heterogeneous academic network

常用的传统链接预测方法有两种：1）基于相似性指标的方法，此类方法对实时性要求较高，对精度要求较低，不考虑边和节点属性的问题；2）基于矩阵运算和基于概率模型[1]的方法，这类方法计算量较大且时间复杂度较高。近年来涌现了大量节点表示学习方法，该类方法常通过捕获网络结构特性、节点属性等信息来完成网络中节点表示的学习，即用低维稠密的向量表示网络中的任意节点[2]，可用于链接预测分析、节点分类等任务。该类方法容易实现多维特征的聚合，在海量或稀疏的网络中具有很大的优势，有助于从多维异构数据角度提升预测准确率，很适合异构网络的链接预测。

异构网络具有节点异质性、节点属性多维性、节点间关系多样性的特点。而现有节点表示学习方法不能有效地表示和融合节点的多维特征，且大多数未考虑节点的全局结构。如 Dong 等于 2017 年提出了 Metapath2vec[3]，该方法通过元路径的随机游走来构建每个目标节点的异构邻域，然后使用 Skip-Gram 模型来处理获取到的领域，从而学习目标节点的表示。但该方法只能基于一种元路径来完成节点的表示，且未考虑节点的内容特征。Wang 等于 2019 年提出了一种异构图神经网络分层注意力机制的方法 HAN[4]，该方法涉及到节点和语义级注意力，虽然使用注意力融合了多种元路径特征，但忽略了元路径中间节点的信息。Zhang 等提出了创新的表示学习方法 HetGNN[5]，该方法能够同时捕捉结构和内容的异构性，但节点的结构特征表示仅使用 Deepwalk[6] 完成，未区分节点的类型，同时未考虑全局结构。

综上所述，现有异构网络表示学习方法存在以下局限性：1）未考虑节点的内容特征，如 Node2vec[7]，HeRec[8]，HHNE[9]，HGT[10]，MetaGATE[11]，Meta-GNN[12] 等；2）仅学习一条元路径信息，丢失了网络中其他类型元路径的信息，如 Metapath2vec 和 HHNE 等；3）考虑了多条元路径，但在提取特征时，只选择与初始节点相同类型的节点，忽略了中间不同类型的节点，如 HeRec，HAN，HAHE[13] 等；4）只考虑了节点的局部特征，未考虑节点的全局特征，如 HetGNN，MAGNN[14]，HetGAN[15]，RHINE[16]。

针对以上问题，受 Metapath2vec 与 HetGNN 的启发，本文提出了一种改进的表示学习方法（Heterogeneous Network Embedding Based on Meta-path and Attribute，HNEMA）。HNEMA 结合 BiLSTM 和注意力机制，融合多类元路径下节点的结构特征、内容特征、社区特征和节点的邻居信息，并利用聚类算法学习节点全局结构。与其他模型相比，HNEMA 更关注节点的多维特征，包括邻居信息，特别加强了学习高阶邻居信息的能力，从而使目标节点的特征提取更为全面合理。

2　相关概念及定义

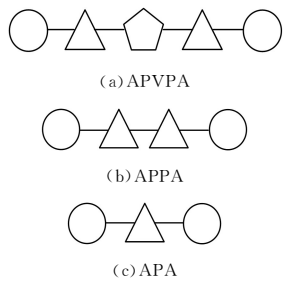
定义 1（异构网络）　给定网络 $G=(V,E)$，V 和 E 分别为节点集和连边集。存在节点类型的映射函数 $\phi:V\to A$ 和连边类型的映射函数 $\varphi:E\to R$。A 和 R 表示预定义的节点类型和连边类型，其中 $|A|+|R|>2$，即网络中节点类型和节点之间关系的类型包含两种及以上，称该网络为异构网络。

如图 1 所示，异构网络包含作者、文章、论文发表载体 3 种类型的节点及合作、引用、发表等多种类型的关系。

定义 2（网络表示学习）　给定异构网络 $G=(V,E)$，网络表示学习旨在通过一系列学习训练找到一种映射 $f(v_i)\to y_i$，$\forall i\in\{1,2,\cdots,n\}$，其中 $|y_i|\ll|V|$，即用低维稠密的向量来表示网络中的任意节点。

定义 3（元路径）　一条路径被定义为 $A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots\xrightarrow{R_t}A_{t+1}$，缩写为 $A_1A_2\cdots A_{t+1}$，其中节点 A_1 与 A_{t+1} 之间的复合关系被描述为 $R=R_1\circ R_2\circ\cdots\circ R_t$。元路径用于捕获异构网络中基于特定范式的结构和语义信息。

在学术网络中常用的几种元路径结构如图 2 所示。



（a）APVPA

（b）APPA

（c）APA

图 2　元路径

Fig. 2　Meta-path

图 2 中，APVPA 表示两个作者（A）在同一个期刊或者会议（V）发表文章（P），APPA 表示两个作者发表的文章存在引用关系，APA 表示两个作者合作发表了同一篇文章。

3 多维特征学习表征的 HNEMA 方法

HNEMA 方法基于神经网络融合节点在多条元路径下的结构特征、内容特征和社区特征，利用 BiLSTM(Bi-directional Long Short-Term Memory)和注意力机制聚合邻居信息，使用聚类算法保留节点的全局性，以完成节点的表示学习。此方法可细分为 5 个部分，即节点采样、节点特征的聚合、同一元路径下相同类型邻居节点的聚合、不同元路径下相同类型邻居节点的聚合和不同类型邻居节点的总聚合，如图 3 所示。

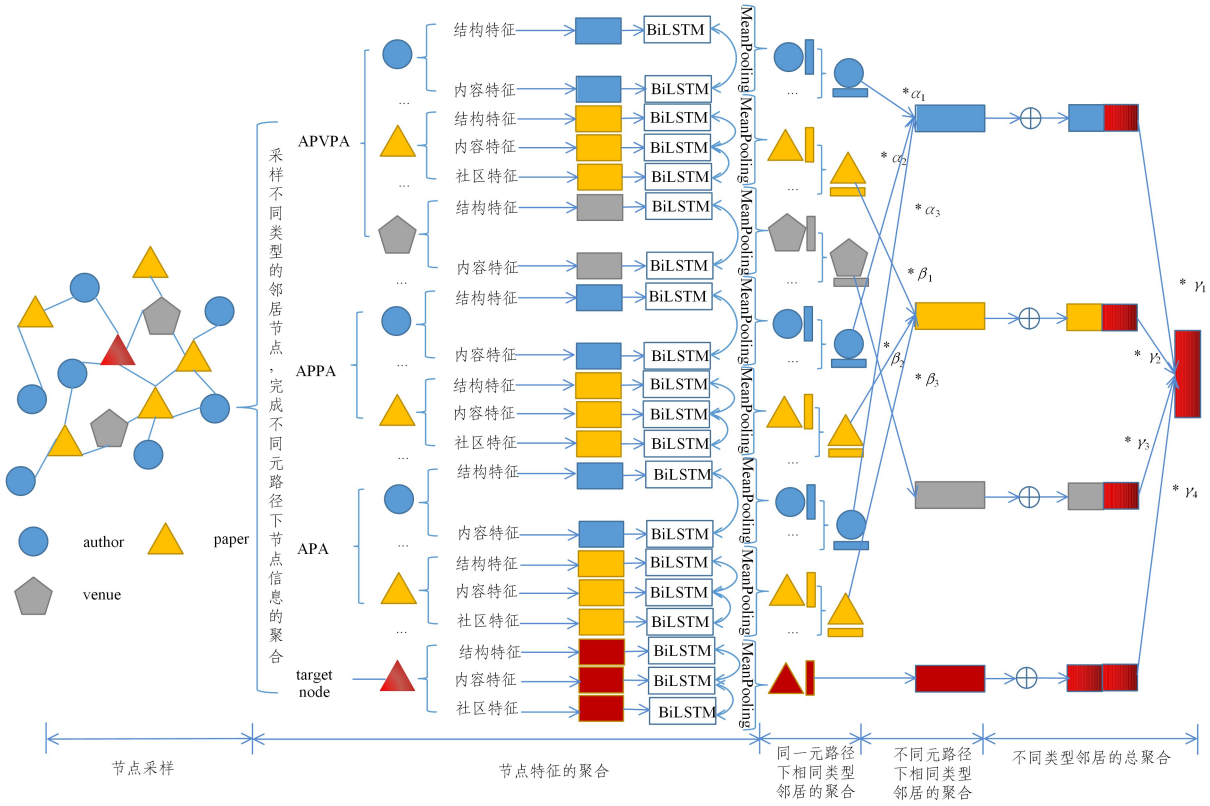


图 3 HNEMA 方法模型图

Fig. 3 Model diagram of HNEMA

3.1 节点采样

大多数网络表示方法的节点采样直接从一阶邻居中获得，但在异构网络中，仅通过一阶邻居无法捕获到所有类型的邻居。如图 1 所示，作者类型的节点无法与论文发表载体类型的节点直接相连。其次，若采样的邻居是弱相关的节点，则会影响待表征目标节点的表示。在异构网络中，不同节点的邻居节点类型和数量可能不同。为了获得更全面的异构邻居信息，本文的节点采样采取基于随机游走和元路径相结合的采样方式。

以某节点 v_i 为待表征的目标节点，采样步骤为：1)从 v_i 开始进行固定路径长度的随机游走，然后采样数量固定但类型不同的邻居节点，同时为了增加采样到的邻居节点与目标节点的相关性，游走过程中设置了重启随机游走，每次游走都有 50%的概率返回到目标节点重新游走；2)针对每一个节点，按照给定的元路径模式，对基于元路径的邻居节点进行采样；3)将前两步采样到的邻居拼接作为目标节点 v_i 的邻居序列 L。从 L 中计算每个邻居节点出现的次数，按照出现次数多到少的顺序排列，并筛选前 k 个强相关且为不同类型的邻居节点作为目标节点 v_i 的邻居集。

此采样过程能覆盖目标节点的所有邻居节点类型，不仅采样了一二阶邻居，还增强了高阶邻居、语义相关邻居对目标节点表征的能力。

3.2 节点特征的聚合

本文提出的 HNEMA 方法对文章类型的节点提取结构特征、内容特征和社区特征，对作者类型和论文发表载体类型的节点提取结构特征和内容特性。受 HetGNN 和 GraphSAGE[17]方法的启发，BiLSTM 能够捕获不同特征之间的交互信息，更好地聚合节点特征，因此基于 BiLSTM 和平均池化层完成节点的多维特征聚合。节点的特征聚合函数如式(1)所示：

$$f_1(v)=\frac{\sum_{i\in C_v}[\overrightarrow{LSTM}\{\mathcal{F}C_{\theta_x}(\boldsymbol{x}_i)\}\oplus\overleftarrow{LSTM}\{\mathcal{F}C_{\theta_x}(\boldsymbol{x}_i)\}]}{|C_v|} \tag{1}$$

其中，v 表示目标节点，$f_1(v)$为节点的特征聚合表示；$\mathcal{F}C_{\theta_x}$是一个特征映射函数，指参数为 θ_x 的全连接神经网络；C_v 为特征集合；$\boldsymbol{x}_i$ 为节点的第 i 个特征表示；$\oplus$表示拼接。

(1)结构特征。采用异构网络的经典表示学习方法 Metapath2vec，目标节点按照给定的元路径模式进行随机游走，保存基于元路径的邻居节点信息，再通过 Skip-Gram 完成目标节点的表示学习。例如，给定元路径实例 $V_1 \xrightarrow{R_1}$

$V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \cdots \xrightarrow{R_{l-1}} V_l$，目标节点与其基于元路径的邻居节点的转移概率为：

$$p(v^{i+1} \mid v_t^i, P) = \begin{cases} \dfrac{1}{|N_{t+1}(v_t^i)|}, & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1} = t+1) \\ 0, & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1} \neq t+1) \\ 0, & (v^{i+1}, v_t^i) \notin E \end{cases} \tag{2}$$

其中，$v_t^i \in V_t$ 表示第 i 步的节点且类型为 t，$N_{t+1}(v_t^i)$ 表示节点 v_t^i 的邻居节点类型为 $t+1$ 的节点集，$(v^{i+1}, v_t^i) \in E$ 表示第 i 和 $i+1$ 节点之间有链接；$\phi(v^{i+1} = t+1)$ 表示节点 v^{i+1} 的类型符合定义的下一个节点类型。本文通过 Metapath2vec 方法完成多条元路径下节点的结构特征表示学习，如 APVPA，APPA，APA。

(2)内容特征。不同类型的节点拥有不同的内容特征。本文作者类型节点的内容特征包含作者发表文章的关键字、摘要信息；文章类型节点的内容特征包含文章的摘要、标题和关键字信息；论文发表载体类型节点的内容特征包含在某会议或期刊上发表的文章摘要信息，内容特征部分采用 Doc2vec[18] 方法完成节点的表示。

(3)社区特征。据调查，目前在异构网络表示学习中使用社区特征的工作相对较少，因为社区结构比一般“结构特征”(如 3.2 节中的(1))更为高阶。学习高阶结构的邻居信息能丰富节点表示，有利于后续的学术行为预测。学术网络是一个异构网络，但其中包含同构网络，如引用网络等。在引用网络中，同一主题的论文形成社区的可能性更大，社区特征进一步描述了文章类型节点之间的相互依赖关系。基于此，在引用网络中采用 GEMSEC(Graph Embedding with Self Clustering)[19] 方法捕获节点的社区特征。GEMSEC 方法的目标函数如式(3)所示，其包含 Skip-Gram 目标函数和聚类目标函数，保证采样序列中在同一个窗口内的节点具有相似的向量表示，同时最小化目标节点和其簇中心的距离，使得同一社区节点的表示近似。

$$O = \min_{f,u}\Big(\sum_{v \in V}\Big[\ln\Big(\sum_{u \in V}\exp(f(v) \cdot f(u))\Big) - \sum_{n_i \in N_s(v)} f(n_i) \cdot f(v)\Big]\Big) + \gamma \cdot \sum_{v \in V} \min_{c \in C} \| f(v) - \boldsymbol{u}_c \|_2^2 \tag{3}$$

其中，f 是节点的表示向量；$f(u)$ 是聚类中心的表示向量；$f(v)$ 表示目标节点的表示向量；$f(n_i)$ 表示与目标节点 v 在同一个移动窗口内的节点的表示向量；$\boldsymbol{u}_c$ 表示第 c 个聚簇中心的表示向量；C 表示聚类中心集；γ 是一个超参数，代表聚类代价的权重系数。

3.3　邻居节点的聚合

(1)同一元路径下相同类型邻居节点的聚合。因 BiLSTM 能挖掘各个邻居节点之间特征的深层交互信息，且参数较少，结构简单，能够简单有效地聚合相同类型邻居节点的信息，故基于 BiLSTM 和平均池化层完成同一元路径下相同类型邻居节点的聚合，如式(4)所示：

$$f_2(v) = \frac{\sum_{i \in M_v}[\overrightarrow{LSTM}\{\mathcal{F}C_{\theta_y}(\boldsymbol{y}_i)\} \oplus \overleftarrow{LSTM}\{\mathcal{F}C_{\theta_y}(\boldsymbol{y}_i)\}]}{|M_v|} \tag{4}$$

其中，$f_2(v)$ 表示同一元路径下相同类型邻居节点的聚合表示；$\mathcal{F}C_{\theta_y}$ 是一个映射函数，表示参数为 θ_y 的全连接神经网络；M_v 表示同一元路径下相同类型邻居节点的集合；$\boldsymbol{y}_i$ 表示同一元路径下相同类型的第 i 个邻居节点的表示向量；$\oplus$ 表示拼接。

(2)不同元路径下相同类型邻居节点的聚合。不同的元路径对目标节点的表示具有不同的重要性，为了更好地聚合 3 种元路径学习的节点的语义特征，本文采用注意力机制为不同元路径赋予不同的权重，完成来自不同元路径下相同类型邻居信息的聚合。此方法在 MetaGATE 和 MAGNN 中都有采用。聚合函数如式(5)所示：

$$\boldsymbol{\varepsilon}_v = \sum_{\boldsymbol{f}_i \in \mathcal{F}(v)} \alpha^{v,i} \boldsymbol{f}_i \tag{5}$$

其中，$\alpha^{v,i}$ 表示相同类型邻居节点在第 i 条元路径下的重要性；v 表示目标节点；$\boldsymbol{f}_i$ 指相同类型邻居节点在第 i 条元路径下的表示；$\mathcal{F}(v)$ 表示多条元路径下邻居节点的表示集合。

(3)不同类型邻居节点的总聚合。因不同类型的邻居节点对目标节点 v 的贡献度不同，故也采用注意力机制为不同类型的邻居节点赋予不同的权重。目标节点的表示如式(6)所示：

$$\boldsymbol{z}_v = \sum_{\boldsymbol{\varepsilon}_i \in E(v)} \beta^{v,i} \boldsymbol{\varepsilon}_i \tag{6}$$

其中，$\beta^{v,i}$ 表示第 i 种类型邻居节点的重要性；$\boldsymbol{\varepsilon}_i$ 表示第 i 种类型邻居节点的表示向量；$E(v)$ 表示不同类型邻居节点的表示集合；$\boldsymbol{z}_v$ 表示目标节点的表示向量。

3.4　目标函数

常见的表示学习优化目标为最小化目标节点和正样本之间的距离，最大化目标节点和负样本之间的距离。本文方法在常见的表示学习目标函数中还加入了 k-means 聚类目标，其中每个簇代表着不同的研究领域。k-means 聚类的优化目标是最小化目标节点与其最近簇中心点的距离，增强簇的内聚度，保留节点的全局结构，使相近邻域的节点具有相似的表示向量。

目标节点与正负样本之间的距离如式(7)所示：

$$O_1 = \sum_{v \in V} \min \| f(v) - f(v_c) \|_2^2 + \sum_{v \in V} \max \| f(v) - f(v_{c'}) \|_2^2 \tag{7}$$

其中，$f(v)$ 表示目标节点的表示向量；v_c 表示正样本；$v_{c'}$ 表示负样本。

目标节点与其最近簇中心之间的距离如式(8)所示：

$$O_2 = \sum_{v \in V} \min_{0 < q \leqslant Q} \| f(v) - \boldsymbol{\mu}_q \|_2^2 \tag{8}$$

其中，$\boldsymbol{\mu}_q$ 表示簇中心的表示向量；Q 表示聚类中心集。

式(9)是总目标函数；式(10)是 γ 值的计算方式；式(9)和式(10)中 γ 是聚类权重系数。当 γ 的初始值较大时，节点的学习向其最近簇中心靠近，聚类的节点可能无法包含与目标节点相关的邻居节点，不能准确学习节点的信息。当 γ 的初始值较小时，初期主要通过学习邻居节点信息完成目标节点的表示。随着训练次数增加，γ 值增加，同时逐渐增大聚类系数权重，使相邻节点更准确地分配到同一个簇中，以更好地学习节点的表示。

$$O=O_1+\gamma * O_2 \tag{9}$$

$$\gamma^{(n+1)}=\gamma^{(n)}\,10^{\frac{-\log_{10}^{\gamma^{(n)}}}{n+1}} \tag{10}$$

3.5 HNEMA 算法

HNEMA 算法的具体描述如算法 1 所示。

算法 1 HNEMA 算法

输入：异构网络 G=(V,E)，节点的结构特征、内容特征、社区特征，嵌入维度 d，训练轮次 e，聚簇数 k，聚类权重系数 γ

输出：节点表示向量矩阵 **Z**

1. 对目标节点进行随机游走和元路径的采样，得到邻居序列 N_{total}
2. 从 N_{total} 中选择强相关的不同类型的邻居节点：
3. $A=\{A_m(v_i) \mid N_{v_i}\in V_{author}, m=10\}$，
   $P=\{P_n(v_i) \mid N_{v_i}\in V_{paper}, n=10\}$，
   $C=\{C_t(v_i) \mid N_{v_i}\in V_{venue}, t=3\}$，$S=A+P+C$
4. for j=1,2,3,…,e：
5. for 目标节点 v_i in V：
6. 对于目标节点 v_i 的邻居集合 S：
7. 聚合不同元路径下节点的不同特征：
$$\mathbf{f}_{v_i}=\frac{\sum[\overrightarrow{\mathrm{LSTM}}\{\mathcal{F}C_{\theta_x}(\mathbf{x}_j)\}\oplus\overleftarrow{\mathrm{LSTM}}\{\mathcal{F}C_{\theta_x}(\mathbf{x}_j)\}]}{|C_{v_i}|}, j\in C_{v_i}$$
8. 聚合同一元路径下相同类型的邻居信息：
$$\mathbf{F}_{v_i}=\frac{\sum[\overrightarrow{\mathrm{LSTM}}\{\mathcal{F}C_{\theta_l}(\mathbf{f}_l)\}\oplus\overleftarrow{\mathrm{LSTM}}\{\mathcal{F}C_{\theta_l}(\mathbf{f}_l)\}]}{|M_{v_i}|}, l\in M_{v_i}$$
9. 聚合不同元路径下相同类型的邻居节点信息：$\mathbf{E}_{v_i}=\sum\lambda_{v_i}^{p}\mathbf{F}_p$，$v_i\in V, 1\leqslant p\leqslant 3$
10. 聚合不同类型的邻居节点信息：$\mathbf{z}_{v_i}=\sum\omega_{\xi_i}E_r$，$v_i\in V, 1\leqslant r\leqslant 3$
11. 得到目标节点 v_i 的向量表示 $\mathbf{z}_{v_i}$
12. 前向传播，计算交叉熵损失和聚类损失 loss
13. 最小化损失，计算相关参数，更新梯度，反向传播，更新参数
14. End for
15. End for
16. Return **Z**

4 实验

4.1 数据集描述

实验采用 Aminer 平台上公开的学术网络数据集 Citation network V1，V2 以及 ACM，经数据预处理后，数据集包含的相关信息如表 1 所列。其中，Author 表示作者数，Paper 表示文章数，以此类推。A-P 表示作者和文章的关系数，以此类推。

表 1 数据集统计信息

Table 1 Dataset statistics information

	V1	V2	ACM
Author	28646	352068	485899
Paper	21044	315866	302395
Venue	18	296	333
A-P	69311	762997	957568
P-P	46931	59337	60462
P-V	21044	315866	302395
Time	2006－2015	1996－2005	2012－2015

4.2 基准方法

为了验证 HNEMA 方法的有效性，将其与以下几种网络表示学习方法进行比较。

(1)Deepwalk[6]：采用随机游走方法和 Skip-gram 模型完成节点表示学习的方法。

(2)Metapath2vec[3]：采用指定的一种元路径模式的游走方法和 Skip-gram 完成节点表示学习的方法。

(3)HHNE[9]：在双曲空间进行异构网络表示学习的方法。

(4)SHNE[20]：一种联合优化图结构紧密性和文本语义相关性，学习在文本关联的异构图中的节点表示方法。

(5)HetGNN[5]：一种基于强相关邻居节点进行优先采样，并结合节点的结构特征和内容特征的节点表示方法。

4.3 实验设置及评价标准

为了验证 HNEMA 方法能否有效表征异构学术网络中节点的特征，设计了 4 个实验，在本文学习的节点特征上进行学术行为预测和论文发表载体推荐。

从基于历史信息预测学术行为的角度出发，本文按照时间参数划分数据集。若时间小于或等于 t，则为训练集；若时间大于 t，则为测试集。t 值的设定与训练集和测试集的数量比例有关，文中采用的训练集与测试集的比例约为 7∶3。在实验过程中，默认学习率为 0.001，采用 Adam 优化器，向量维度设置为 128，epoch 设置为 70，聚类权重系数初始化设置为 0.001，每个节点的邻居个数和聚类个数 k 为超参数。基准方法参数参考原文，选择最优的参数设置。

为了评估方法的有效性，实验中使用的评价指标为 F1 值和 AUC，定义如式(11)、式(12)所示：

$$F\text{-}Measure=\frac{1+\alpha^2}{\alpha^2} * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

$$AUC=\frac{\sum_{i\in positiveClass} Rank_i - M(M+1)/2}{M\times N} \tag{12}$$

其中，M 和 N 分别表示正、负样本数，$Rank_i$ 表示第 i 条正样本置信度排序序号。

4.4 实验设计与分析

实验 1 作者的学术行为预测

本实验在表 1 所列的数据集上完成作者与作者的合作关系预测(A-P-A)，作者与文章的引用关系预测(A-P-P)，作者与论文发表载体的参与关系预测(A-P-V)。本文采用 HNEMA 表征异构学术网络，对比方法采用各自的表征方法，基于学习提取后的特征，应用逻辑回归分类器进行作者的学术行为预测实验。表 2－表 4 列出了 3 类学术行为的预测实验结果。

表 2 作者与作者的合作关系预测实验结果

Table 2 Cooperation prediction experiment results

Algorithm	V1(2011)		V2(2002)		ACM(2014)	
	AUC	*F1*	*AUC*	*F1*	*AUC*	*F1*
Deepwalk	0.794	0.272	0.766	0.488	0.821	0.471
Metapath2vec	0.770	0.679	0.745	0.653	0.727	0.649
HHNE	0.698	0.503	0.674	0.498	0.692	0.524
SHNE	0.812	0.642	0.695	0.606	0.687	0.579
HetGNN	0.799	0.675	**0.793**	0.634	0.818	0.671
HNEMA	**0.852**	**0.750**	0.784	**0.666**	**0.831**	**0.679**

表 3　作者与文章的引用关系预测实验结果

Table 3　Citation prediction experiment results

Algorithm	V1(2011)		V2(2002)		ACM(2014)	
	AUC	*F1*	*AUC*	*F1*	*AUC*	*F1*
Deepwalk	0.724	0.433	0.791	0.614	0.834	0.640
Metapath2vec	0.624	0.684	0.742	0.687	0.771	0.701
HHNE	0.707	0.573	0.688	0.587	0.739	0.651
SHNE	0.799	0.691	0.687	0.635	0.775	0.693
HetGNN	0.829	0.740	0.803	0.691	0.841	0.744
HNEMA	**0.857**	**0.774**	**0.804**	**0.692**	**0.868**	**0.778**

表 4　作者与论文发表载体的参与关系预测实验结果

Table 4　Participant prediction experiment results

Algorithm	V1(2011)		V2(2002)		ACM(2014)	
	AUC	*F1*	*AUC*	*F1*	*AUC*	*F1*
Deepwalk	0.643	0.345	0.821	0.628	0.902	0.801
Metapath2vec	0.627	0.524	0.894	0.725	0.887	0.741
HHNE	0.608	0.387	0.618	0.494	0.664	0.570
SHNE	0.722	0.496	0.749	0.642	0.872	0.757
HetGNN	0.786	0.559	0.826	0.706	0.889	0.801
HNEMA	**0.817**	**0.623**	**0.900**	**0.773**	**0.938**	**0.827**

3 类预测实验结果表明，基于 HNEMA 的异构网络特征表征方法能显著提升预测性能。例如，在作者与作者的合作关系预测上，AUC 增长了 1%～15.4%，F1 值增长了 0.8%～47.8%。

实验 2　论文发表载体的个性化推荐实验

基于 HNEMA 异构网络表征学习，本文完成了论文发表载体的个性化推荐实验，其中 R 表示 Recall，P 表示 Precision，实验结果如表 5 所列。

表 5　V1 的论文发表载体推荐实验结果

Table 5　Venue recommended experimental results in V1

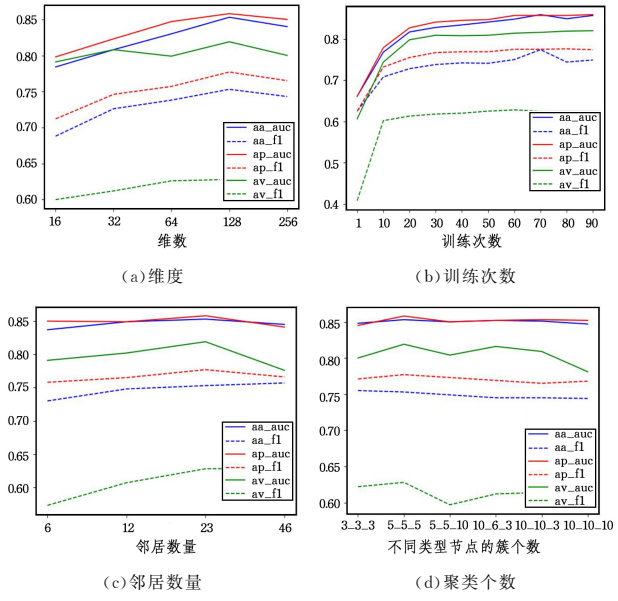
Algorithm	*R*@3	*R*@5	*P*@3	*P*@5	*F1*@3	*F1*@5
Deepwalk	0.267	0.431	0.122	0.117	0.167	0.184
Node2vec	0.428	0.545	0.186	0.145	0.259	0.229
MP2ec	0.206	0.311	0.095	0.088	0.131	0.137
HHNE	0.293	0.456	0.133	0.123	0.183	0.194
SHNE	0.421	0.578	0.189	0.157	0.262	0.247
HetGNN	0.648	0.750	0.282	0.201	0.393	0.317
HNEMA	**0.659**	**0.771**	**0.287**	**0.207**	**0.400**	**0.323**

由表 5 可知，本文算法的召回率、精准率、F1 均优于其他方法，其中 F1 值提高了 0.6%～26.9%。

HNEMA 方法能有效提升行为预测和个性化推荐性能，这归因于它的 3 个优势：1）通过 BiLSTM 能够有效地融合节点的多维特征，包括高阶社区特征；2）通过注意力机制赋予不同元路径以不同的权重，较大程度地增强了节点与邻居节点的结构和语义信息的学习能力；3）综合表征节点的局部结构和全局结构，能较全面地学习网络结构信息。

实验 3　参数合理设置实验

超参数分析实验基于 V1 数据集完成。由图 4(a)可知，当维数为 16～128 时，AUC 和 F1 所衡量的预测性能都有上升。随着维数继续增加，性能趋于平缓或下降，因此本文选取最佳维度为 128。图 4(b)显示，随着训练轮次的增加，预测性能呈上升趋势，当训练轮次为 70 时，训练效果基本稳定。由图 4(c)可知，当邻居数量小于 20 时，预测性能呈上升趋势；当大于 30 时，因加入了一些弱相关的邻居干扰节点的学习，预测效果有所下降，故邻居数量应设置为 20～30。图 4(d)显示，3 种类型的节点聚类个数分别为 5，即划分为 5 个研究邻域时，预测性能最好。在不同的数据集上，维度、训练次数、邻居数量、聚类个数均可能不同，因为不同的数据集具有不同的数据规模和网络结构。



(a)维度　(b)训练次数　(c)邻居数量　(d)聚类个数

图 4　超参数分析实验

Fig. 4　Hyperparameters analysis experiments

结束语　本文提出的 HNEMA 方法旨在综合、全面地学习节点的多维特征、邻居信息和全局结构，有效表征异构网络。基于 HNEMA，本文实现了作者的 3 种学术行为预测。研究和实验结果表明，HNEMA 方法具有有效融合节点信息、多方面学习节点的结构和语义特征、捕获高阶邻居信息的优势。

虽然元路径提供了很强的结构和语义信息，但是需要人为设定，未来工作将研究元路径的自动生成。此外，异构网络表征中的时空效率也是将来的研究重点之一。

参 考 文 献

[1] KUMAR A, SINGH S S, SINGH K, et al. Link prediction techniques, applications, and performance: A survey[J/OL]. Physica A: Statistical Mechanics and its Applications, 2020, 553. https://doi.org/10.1016/j.physa.

[2] SHI C, SUN Y. Research progress of heterogeneous network representation learning[J]. Communications of the CCF, 2018, 14(3): 16-20.

[3] 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.

[4] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network[C]// The World Wide Web Conference. 2019: 2022-2032.

[5] ZHANG C, SONG D, HUANG C, et al. Heterogeneous graph neural network [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mi-

ning. 2019: 793-803.
[6] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations [C]// Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press. 2014: 701-710.
[7] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press. 2016: 855-864.
[8] WANG X, SHI C, HU B, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357-370.
[9] ZHANG Y, SHI C. Hyperbolic heterogeneous information network embedding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 5337-5344.
[10] HU Z, DONG Y, WANG K. Heterogeneous graph transformer [C]// Proceedings of The Web Conference. 2020: 2704-2710.
[11] CAO M, MA X, XU M, et al. Heterogeneous information network embedding with meta-path based on graph attention networks [C]// International Conference on Artificial Neural Networks. 2019: 622-634.
[12] SANKAR A, ZHANG X, CHANG K C. Meta-gnn: metagraph neural network for semi-supervised learning in attributed heterogeneous information networks[C]// Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2019: 137-144.
[13] ZHOU S, BU J, WANG X, et al. Hahe: Hierarchical attentive heterogeneous information network embedding[J]. arXiv: 1902.01475, 2019.
[14] 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: 2331-2341.
[15] HU B, FANG Y, SHI C. Adversarial learning on heterogeneous information networks[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 120-129.
[16] LU Y, SHI C, HU L, et al. Relation structure-aware heterogeneous information network embedding [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 4456-4463.
[17] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 1025-1035.
[18] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]// International conference on machine learning. PMLR, 2014: 1188-1196.
[19] ROZEMBERCZKI B, DAVIES R, SARKAR R, et al. Gemsec: Graph embedding with self clustering [C]// Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2019: 65-72.
[20] ZHANG C, SWAMI A, CHAWLA N V. Shne: Representation learning for semantic-associated heterogeneous networks [C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019: 690-698.

HUANG Li, born in 1996, postgraduate. Her main research interests include representation learning, data mining and link prediction.

ZHU Yan, born in 1965, Ph. D, professor, Ph. D supervisor, is a member of China Computer Federation. Her main research interests include data mining, Web anomaly and intelligent analysis.

(责任编辑: 何杨)