出的一种基于元路径的图随机遍历技术. 对于给定的异质信息网络 $G=(V,E)$ 和元路径 $P=A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots A_t\xrightarrow{R_t}A_{t+1}\cdots\xrightarrow{R_l}A_{l+1}$, 随机游走的起始节点为 A_1 类型节点, 第 $i+1$ 个游走节点的选择概率如式 (2) 所示. 其中 v_t^i 表示 A_t 类型的节点, $N_{t+1}\left(v_t^i\right)$ 表示节点 v_t^i 的邻域中 A_{t+1} 类型的节点数量. 第 $i+1$ 个游走节点应从节点 v_t^i 的所有 A_{t+1} 类型的邻居节点中随机选择. 基于节点选择概率, 随机游走将在元路径的指导下游走出包含元路径语义信息以及异质信息网络中结构信息的节点序列.

$$s\left(v^{i+1}\left|v_t^i,P\right.\right)=\begin{cases}\dfrac{1}{\left|N_{t+1}\left(v_t^i\right)\right|}, & \left(v^{i+1},v_t^i\right)\in E,\Phi\left(v^{i+1}\right)=A_{t+1}\\ 0, & \left(v^{i+1},v_t^i\right)\in E,\Phi\left(v^{i+1}\right)\neq A_{t+1}\\ 0, & \left(v^{i+1},v_t^i\right)\notin E\end{cases} \tag{2}$$

Skip-gram 模型是 Mikolov T 等[4]提出的用于自然语言处理中学习大型数据集中单词的连续向量表征的神经网络模型. Skip-gram 模型具有三层网络结构, 分别为输入层、隐藏层和输出层, 并提出了式 (3) 所示的损失函数[20]. 其中, C 为上下文中单词数量, V 为语料库中单词数量, w_I 表示输入的单词, $w_{O,i}$ 表示第 i 个输出的上下文单词, j_c^* 为输出层输出的第 c 个上下文单词在语料库中的真实索引, u 表示单词从隐藏层到输出层过程中的计算分数. 该模型输入为由文本中句子构成的语料库, 通过最小化损失函数, 学习语料库中单词的低维表征.

$$\begin{aligned}E&=-\log p(w_{O,1},w_{O,2},\cdots,w_{O,C}\,|w_I)\\&=-\log\prod_{c=1}^{C}\frac{\exp(u_{c,j_c^*})}{\sum_{j'=1}^{V}\exp(u_{j'})}\\&=-\sum_{c=1}^{C}u_{j_c^*}+C\cdot\log\sum_{j'=1}^{V}\exp(u_{j'})\end{aligned} \tag{3}$$

目前, DeepWalk[3]、Node2Vec[5]、Metapath2Vec[15] 等研究发现将信息网络中节点信息映射为自然语言可应用 Skip-gram 模型学习信息网络中节点的低维表征. 基于元路径的随机游走技术可提取包含元路径语义信息、网络结构信息的节点序列, 从而将异质信息网络中的节点信息映射为自然语言, 进而可结合 Skip-gram 模型学习异质信息网络中节点的低维表征.

如图 2 中阶段 3 所示, 首先应用基于元路径的随机游走技术获取异质信息网络中的节点序列. 对任意元路径 $p_i\in P$ 获取其相应的节点序列集 c_{p_i} 并构建语料库集合 $C=\{c_{p_1},c_{p_2},\cdots,c_{p_n}\}$.

对语料库集合中任意一个节点序列集 c_{p_i} 应用 Skip-gram 模型学习异质信息网络的低维表征 M_{p_i}. 此时, 任意元路径 p_i 都有唯一的低维表征 M_{p_i} 与之对应. 各个低维表征构成了基于不同元路径的低维表征集合 $M=\{M_{p_1},M_{p_2},\cdots,M_{p_n}\}$.

2.4 阶段 4: 基于元路径权重对低维表征进行融合

此阶段基于元路径权重集合 $W=\{w_{p1},w_{p2},\cdots,w_{p_n}\}$ 对低维表征集合 $M=\{M_{p_1},M_{p_2},\cdots,M_{p_n}\}$ 进行加权融合. 对于任意的低维表征 M_{p_i} 均基于相应的元路径 p_i, 所以低维表征 M_{p_i} 中仅包含元路径 p_i 所表示的语义信息, 导致基于单一元路径的低维表征中缺失其它元路径表示的语义信息. 而元路径因语义信息的不同对表征学习的重要程度不同, 从而具有不同的权重. 所以对基于不同元路径的低维表征进行加权融合可得到融合不同元路径语义信息的低维表征, 从而提高低维表征质量. 因此, 本文提出了式 (4) 所示的低维表征融合公式, 并基于该公式实现了基于元路径权重的低维表征融合算法.

$$M_W=\sum_{i=1}^{n}w_{p_i}\times M_{p_i} \tag{4}$$

如算法 1 所示, 该算法的输入为元路径权重集合、低维表征集合以及低维表征维度, 然后依次对低维表征中 d 个特征分量进行加权融合, 得到融合不同元路径语义信息的低维表征 M_W. 低维表征 M_W 不仅包含不同元路径的语义信息, 而且还包含丰富的网络结构信息. 以上特点使得基于融合元路径权重的低维表征在低维空间中具有良好的表示、推理能力, 并且可有效应用于数据挖掘任务.

算法 1. 基于元路径权重的低维表征融合算法

输入: 元路径权重集合 $W=\{w_{p_1},w_{p_2},\cdots,w_{p_n}\}$, 低维表征集合 $M=\{M_{p_1},M_{p_2},\cdots,M_{p_n}\}$, 维度 d

输出: 融合元路径权重的低维表征 M_W

1. **for** i=0,1,2,$\cdots$,$d-1$ **do**
2. $M_W[i]=w_{p_1}\times M_{p_1}[i]+w_{p_2}\times M_{p_2}[i]+\cdots+w_{p_n}\times M_{p_n}[i]$
3. **end for**

3 实验结果与分析

为证明本文提出的基于融合元路径权重的异质网

络表征学习方法的正确性以及在数据挖掘任务中的有效性, 本文对实际数据集进行了节点分类对比试验.

3.1 实验数据集

实验所用数据集为 AMIner[15,21]数据集, 该数据集为典型的异质学术文献信息网络. 如表 1 所示, 该数据集中包含作者、文章、会议 3 种节点类型, 共计 4891 819 个数据节点, 其中 246 678 个带标签的作者节点被分为 8 个类别, 分别为 Computing Systems, Theoretical Computer Science, Computer Networks & Wireless Communication, Computer Graphics, Human Computer Interaction, Computational Linguistics, Computer Vision & Pattern Recognition, Databases & Information Systems.

表 1 AMIner 数据集中的节点

节点类型	节点数量
作者	1693 531 (246 678 带标签)
文章	3194 405
会议	3883
合计	4891 819

如表 2 所示, AMiner 数据集中共包含 12 518 144 个边, 其中表示文章与作者之间撰写与被撰写关系的边共 9323 739 个, 表示文章与会议之间发表与被发表关系的边共 3194 405 个.

表 2 AMIner 数据集中的边

边类型	边数量
文章-作者 (作者-文章)	9323 739
文章-会议 (会议-文章)	3194 405
合计	12 518 144

此外, 本文在 AMIner 数据集的基础上构建数据规模较小的子数据集 AMIner-Small, 用于验证本文提出的基于融合元路径权重的异质网络表征学习方法对不同数据规模的异质信息网络的表征学习能力. 如表 3 所示, AMIner-Small 数据集中数据规模远远小于 AMiner 数据集.

表 3 AMIner-Small 数据集中的节点

节点类型	节点数量
作者	1290(675 带标签)
文章	500
会议	10
合计	1800

3.2 评价指标

在分类实验中, 数据的低维表征质量对实验结果具有重要影响, 因此本文通过实验结果评价低维表征质量, 进而分析异质网络表征学习方法的正确性、有效性.

本文采用分类精确率 (Precision)、召回率 (Recall)、Micro-F1 分数、Macro-F1 分数评价分类实验结果, 从而评价不同异质网络表征学习方法的正确性、在数据挖掘任务中的有效性.

分类精确率为预测为正类的样本中实际为正类的样本比例. 召回率表示预测为正类的样本数占全部正类样本数的比例. F1 分数 (Micro-F1 分数、Macro-F1 分数) 表示精确度和召回率的加权平均值. 以上 4 个评价指标值越高表示分类实验越精确, 相应的低维表征质量越高、异质网络表征学习方法越正确、有效.

3.3 节点分类实验

3.3.1 AMIner-Small 数据集的节点分类实验

采用 HIN2Vec[17]异质网络表征框架作为对比实验方法. 不同于之前基于 Skip-gram 模型的表征方法, HIN2Vec 核心是一个神经网路模型, 并且将元路径视为节点间的不同类型关系, 然后通过捕获节点间不同类型关系学习节点的低维表征.

首先在 AMIner-Small 数据集的基础上构建元路径集合并学习各个元路径的权重. 权重学习实验重复十次, 结果如表 4 所示, 其中 APA 的权重均值为 0.01, APAPA 的权重均值为 0.02, APCPA 的权重均值为 0.97. 根据元路径权重学习结果发现在 AMIner-Small 数据集中元路径 APCPA 表示的语义信息对异质网络表征学习的重要程度远高于 APA、APAPA 表示的语义信息, 而 APA、APAPA 表示的语义信息对异质网络表征学习的重要程度则十分接近.

表 4 元路径及其权重

元路径	权重范围	权重均值
APA	0.005 ~ 0.013	0.01
APAPA	0.01 ~ 0.03	0.02
APCPA	0.96 ~ 0.98	0.97
合计	—	1

在元路径集合及权重的基础上分别应用本文提出的基于融合元路径权重的异质网络表征学习方法以及 HIN2Vec 框架学习 AMIner-Small 数据集中节点的低维表征. 然后将带标签的 675 个作者节点的低维表征作为特征向量训练和测试 SVM 分类器. 分类实验中

将 675 个低维表征按 70%/30% 比例随机分为训练数据集与测试数据集, 分类结果是取 10 次实验结果的均值. 具体实验结果如表 5 所示, 其中 FMPW 表示本文提出的基于融合元路径权重的异质网络表征学习方法.

表 5 AMIner-Small 数据集中作者节点分类结果

方法	精确率	召回率	Micro-F1	Macro-F1
HIN2Vec(APA)	0.4784	0.2786	0.4560	0.2685
HIN2Vec(APAPA)	0.4233	0.2103	0.4093	0.2000
HIN2Vec(APCPA)	0.5814	0.4399	0.5567	0.4257
HIN2Vec ({APA/APAPA/APCPA})	0.6012	0.4230	0.5911	0.4171
FMPW	**0.6400**	**0.5078**	**0.6404**	**0.4947**

根据实验结果发现本文提出的基于融合元路径权重的异质网络表征学习方法在分类精确率、召回率、Micro-F1 分数、Macro-F1 分数上均明显高于 HIN2Vec 方法.该结果表明基于融合元路径权重的异质网络表征学习方法对小规模异质网络具有良好的表征学习能力, 证明了该方法的正确性、有效性.

3.3.2 AMIner 数据集的节点分类实验

由于 AMIner 数据集中数据规模远大于 AMiner-Small 数据集, 导致 HIN2Vec 不能处理 AMiner 数据集, 所以本文采用 Metapath2Vec[15]异质网络表征方法作为对比实验方法. Metapath2Vec 应用基于单条元路径的随机游走获取异质网络中的结构信息并结合 Skip-gram 模型需学习异质网络的低维表征.

此部分实验中, 实验步骤与 AMIner-Small 数据集中分类的实验步骤一致, 首先提取元路径 APA、APAPA、APCPA 构成元路径集合并学习其权重, 然后分别采用本文提出的基于融合元路径权重的异质网络表征学习方法和 Metapath2Vec 方法学习 AMIner 数据集中节点的低维表征.

元路径权重学习的实验结果与 AMIner-Small 数据集中的元路径权重学习结果一致, 即 APA 的权重均值为 0.01, APAPA 的权重均值为 0.02, APCPA 的权重均值为 0.97. 该结果表示在 AMIner 数据集中 APCPA 表示的语义信息对异质网络表征学习的影响程度最大.

本文在全部节点的低维表征中随机挑选 47 108 个带标签的作者的低维表征作为 SVM 分类器的特征向量, 其中训练集比例为 10%~90%, 其余节点为测试集. 实验重复十次并取平均值, 结果如图 3 所示, 其中 FMPW 表示本文提出的基于融合元路径权重的异质网络表征学习方法.

根据实验结果可知, 随着训练集比例的提高, 分类结果越加精确. 而且本文提出的基于融合元路径权重的异质网络表征学习方法的分类精确率、召回率、Micro-F1 分数、Macro-F1 分数中均明显高于基于元路径 APA 和基于元路径 APAPA 的 Metapath2Vec 方法, 但是仅率高于基于 APCPA 的 Metapath2Vec 方法. 造成以上结果的原因在于, 元路径 APCPA 的权重为 0.97, 导致融合不同元路径的低维表征中 APCPA 对应的低维表征占主要比例. 该结果从侧面验证了元路径权重学习结果的正确性. 此外, 基于图 3 所示的实验结果发现基于不同元路径的 Metapath2Vec 方法学习的低维表征质量差别大, 导致应用 Metapath2Vec 方法学习异质网络的低维表征时结果具有不确定性. 而本文提出的基于融合元路径权重的异质网络表征学习方法可得出最优结果, 从而有效解决上述问题.

3.4 实验分析

综合以上实验结果可知, 基于融合元路径权重的异质网络表征学习方法可应用于不同数据规模的异质网络, 并且在不同数据规模的异质网络中分类实验结果优于基准方法 HIN2Vec 和 Metapath2Vec. 因此本文提出的基于融合元路径权重的异质网络表征学习方法对不同数据规模的异质网络具有良好的表征学习能力, 可学习得到高质量的低维表征, 可有效应用于数据挖掘任务, 并且优于基于单条元路径的异质网络表征学习方法.

4 结论

本文提出基于融合元路径权重的异质网络表征学习方法, 通过元路径权重学习表明元路对异质网络表征学习的重要程度, 并以此为基础对基于不同元路径的低维表征进行加权融合, 得到融合不同元路径语义信息的异质网络表征. 该方法解决了基于单条元路径的异质网络表征学习方法不能包含其它元路径语义信息而导致的低维表征中缺失结构信息、语义信息的问题. 同时通过对比试验证明本文提出的基于融合元路径权重的异质网络表征学习方法在不同数据规模的异质网络中具有良好的表征学习能力, 并且可有效应用于数据挖掘任务. 在未来的工作中, 将对如何提高大规模异质网络的表征学习效率进行深入研究.

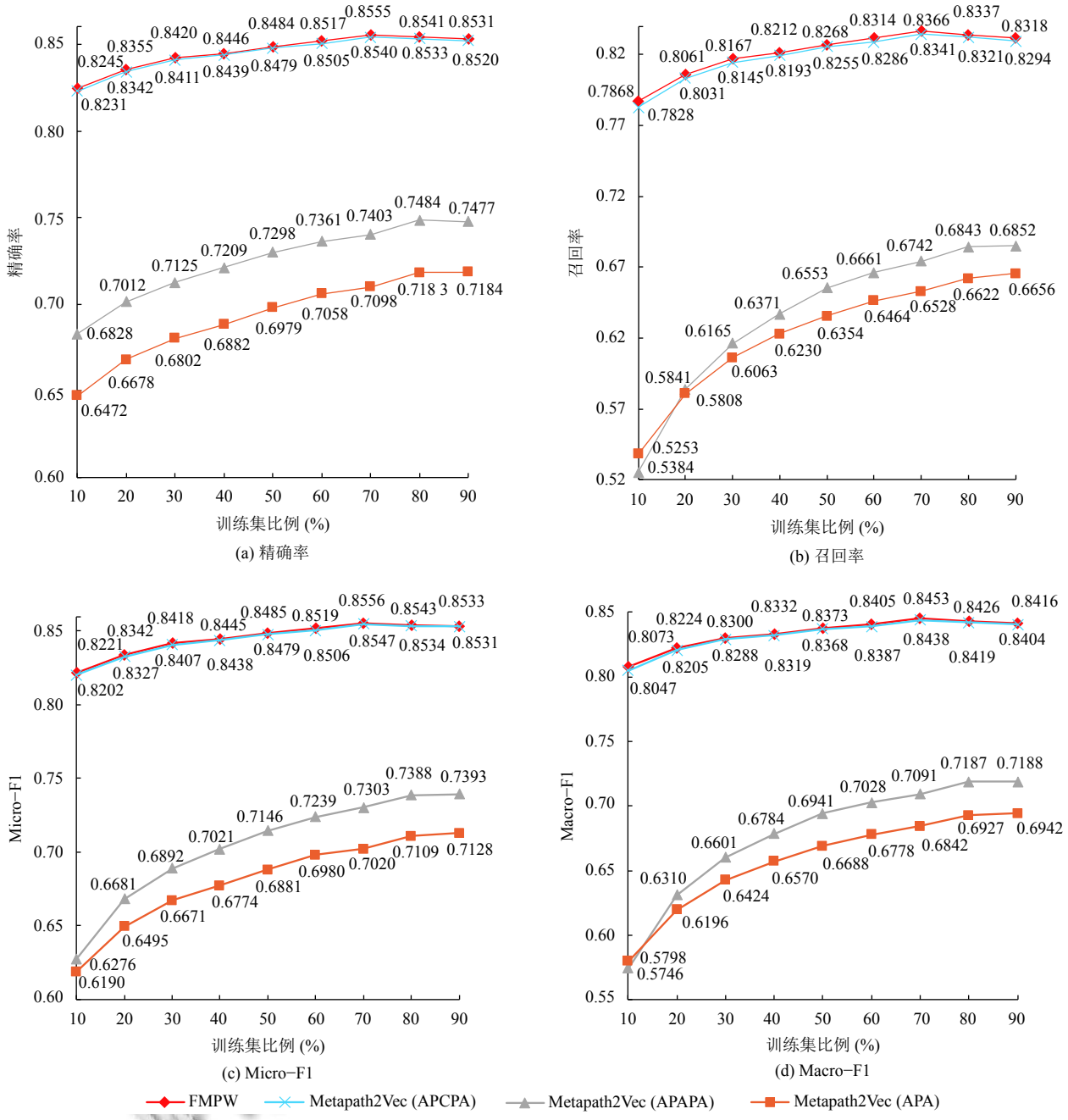

图 3 AMIner 数据集中作者节点分类结果

参考文献

1 Cai HY, Zheng VW, Chang KCC. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616–1637. [doi: 10.1109/TKDE.2018.2807452]

2 蒋宗礼, 张津丽, 杜永萍, 等. 基于堆叠降噪自编码器的异质网络的层次构建与节点分类. 北京工业大学学报, 2018, 44(9): 1217–1226. [doi: 10.11936/bjutxb2017040032]

3 Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2014. 701–710.

4 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013.

5 Grover A, Leskovec J. Node2Vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 855–864.

6 Tang J, Qu M, Wang MZ, *et al*. LINE: Large-scale information network embedding. Proceedings of the 24th International Conference on World Wide Web. Florence, Italy. 2015. 1067–1077.

7 Yang C, Liu ZY, Zhao DL, *et al*. Network representation learning with rich text information. Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina. 2015. 2111–2117.

8 Cao SS, Lu W, Xu QK. GraRep: Learning graph representations with global structural information. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia. 2015. 891–900.

9 Tu CC, Zhang WC, Liu ZY, *et al*. Max-margin deepwalk: Discriminative learning of network representation. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, NY, USA. 2016. 3889–3895.

10 Shi C, Li YT, Zhang JW, *et al*. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17–37. [doi: 10.1109/TKDE.2016.2598561]

11 石川, 孙怡舟. 异质网络表征学习的研究进展. 中国计算机学会通讯, 2018, 14(3): 16–21.

12 Sun YZ, Han JW, Yan XF, *et al*. PathSim: Meta path-based top-K similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment, 2011, 4(11): 992–1003.

13 Sun YZ, Han JW. Mining heterogeneous information networks: Principles and methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery, 2012, 3(2): 1–159. [doi: 10.2200/S00433ED1V01Y201207DMK005]

14 Zhang JL, Jiang ZL, Li T. CHIN: Classification with META-PATH in heterogeneous information networks. Proceedings of the 1st International Conference on Applied Informatics. Bogotá, Colombia. 2018. 63–74.

15 Dong YX, Chawla NV, Swami A. Metapath2Vec: Scalable representation learning for heterogeneous networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, NS, Canada. 2017. 135–144.

16 石川, 孙怡舟, 菲利普·俞. 异质信息网络的研究现状和未来发展. 中国计算机学会通讯, 2017, 13(11): 35–40.

17 Fu TY, Lee WC, Lei Z. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore. 2017. 1797–1806.

18 Gupta M, Kumar P, Bhasker B. HeteClass: A meta-path based framework for transductive classification of objects in heterogeneous information networks. Expert Systems with Applications, 2017, 68: 106–122. [doi: 10.1016/j.eswa.2016.10.013]

19 Gupta M, Kumar P, Bhasker B. A new relevance measure for heterogeneous networks. Proceedings of the 17th International Conference on Big Data Analytics and Knowledge Discovery. Valencia, Spain. 2015. 165–177.

20 Rong X. Word2Vec parameter learning explained. arXiv preprint arXiv: 1411.2738, 2014.

21 Tang J, Zhang J, Yao LM, *et al*. ArnetMiner: Extraction and mining of academic social networks. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, USA. 2008. 990–998.