计算机科学
COMPUTER SCIENCE

基于无监督集群级的科技论文异质图节点表示学习方法

宋杰，梁美玉，薛哲，杜军平，寇菲菲

引用本文

宋杰，梁美玉，薛哲，杜军平，寇菲菲. 基于无监督集群级的科技论文异质图节点表示学习方法[J]. 计算机科学, 2022, 49(9): 64-69.

SONG Jie, LIANG Mei-yu, XUE Zhe, DU Jun-ping, KOU Fei-fei. Scientific Paper Heterogeneous Graph Node Representation Learning Method Based onUnsupervised Clustering Level[J]. Computer Science, 2022, 49(9): 64-69.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500196

基于无监督集群级的科技论文异质图节点表示学习方法

宋 杰 梁美玉 薛 哲 杜军平 寇菲菲
北京邮电大学计算机学院(国家示范性软件学院)智能通信软件与多媒体北京市重点实验室 北京 100876
(songs@bupt.edu.cn)

摘 要 科技论文数据的知识表征是一个有待解决的问题,而如何学习科技论文异质网络中论文节点的表示是解决这一问题的核心。文中提出了一种基于无监督集群级的科技论文异质图节点表示学习方法(Unsupervised Cluster-level Scientific Paper Heterogeneous Graph Node Representation Learning Method,UCHL),以获取科技论文异质图中节点(作者、机构与论文等)的表示。基于科技论文异质图表示对整个异质图进行链接预测,获取节点之间边的关系,即论文与论文之间的关联关系。实验结果表明,在真实的科技论文数据集上,所提方法在多项评测指标上都取得了更优的性能。

关键词:科技论文;异质图网络;图表示学习;链接预测;无监督学习
中图法分类号 TP391

Scientific Paper Heterogeneous Graph Node Representation Learning Method Based on Unsupervised Clustering Level

SONG Jie,LIANG Mei-yu,XUE Zhe,DU Jun-ping and KOU Fei-fei
Beijing Key Laboratory of Intelligent Communication Software and Multimedia,School of Computer Science(National Pilot Software Engineering School),Beijing University of Posts and Telecommunications,Beijing 100876,China

Abstract Knowledge representation of scientific paper data is a problem to be solved,and how to learn the representation of paper nodes in scientific paper heterogeneous network is the core to solve this problem. This paper proposes an unsupervised cluster-level scientific paper heterogeneous graph node representation learning method(UCHL),aiming at obtaining the representation of nodes (authors,institutions,papers,etc.) in the heterogeneous graph of scientific papers. Based on the heterogeneous graph representation,this paper performs link prediction on the entire heterogeneous graph and obtains the relationship between the edges of the nodes,that is,the relationship between paper and paper. Experiments results show that the proposed method achieves excellent performance on multiple evaluation metrics on real scientific paper datasets.

Keywords Scientific paper,Heterogeneous graph network,Graph representation learning,Link prediction,Unsupervised learning

1 引言

异质图节点间存在多种类型的边(关系),同时每一条边所具有的不同属性也会导致节点间的远近亲疏。目前异质图处理的难点在于,一方面要处理图的结构信息,另一方面要关注每个节点所具有的属性。传统的机器学习方法专注于单个节点的特征,而忽略了结构信息。图神经网络通过递归邻域聚合策略学习节点的新特征向量,许多有监督的图神经网络模型被提出,用于进行图数据表示学习。然而,科技论文异质图中的标记数据并不总是可用的,这些算法不适用于科技论文异质图的无监督学习。

在异质图研究中,元路径已被广泛用于表示具有不同语义的复合关系,UCHL 利用元路径的结构对异构图中的连接语义进行建模,基于不同的元路径,将异构图分解为特定语义的同构图,应用图卷积神经网络来捕获具有特定语义的节点的局部表示,并基于注意力机制聚合不同语义的节点表示,通过最大化局部与全局、局部与集群簇中心的互信息,且在不依赖于任何监督标签指导信息的情况下,学习嵌入了图级语义结构信息的表示。本文的主要贡献如下:

(1)提出了一种基于无监督集群级的科技论文异质图节点表示学习方法 UCHL,通过学习科技论文异质网络中论文数据的节点表示来获取科技论文数据的知识表征。

(2)基于互信息理论以及集群簇中心学习科技论文异质图节点的低维嵌入空间表示,同时保留了图的语义结构信息以及节点特征信息。

到稿日期:2022-05-20 返修日期:2022-07-05
基金项目:国家重点研发计划(2018YFB1402600);国家自然科学基金(61877006,61802028,62002027)
This work was supported by the National Key R & D Program of China(2018YFB1402600) and National Natural Science Foundation of China(61877006,61802028,62002027).
通信作者:梁美玉(meiyu1210@bupt.edu.cn)

2　相关工作

Deepwalk[1]是使用游走的图嵌入的技术之一，通过从一个节点移动到另一个节点来实现图的遍历。Node2vec[2]是最早尝试从图形结构化数据中学习的深度学习技术之一，它考虑广度优先搜索与深度优先搜索过程。graph2vec[3]本质上是学习嵌入图的一组由用户指定边数的子图，并将其传递给神经网络进行分类。SDNE[4]尝试从两个不同的指标中学习，即一阶接近度与二阶接近度。LINE[5]明确定义了两个函数，一个用于一阶接近，另一个用于二阶接近。HARP[6-7]通过更好的权重初始化方式来改进解决方案，避免局部最优，并使用图粗化技术将相关节点聚合为“超级节点”，这本质上是一个图预处理步骤，可简化图以加快训练速度。

GCN[8]是一个多层图卷积神经网络，它将邻接矩阵 $\boldsymbol{A}$ 和特征矩阵 $\boldsymbol{X}$ 编码为嵌入 $\boldsymbol{H}$。GATs[9]在 GCN(Graph Convolutional Networks)的基础上不使用拉普拉斯矩阵而使用注意力系数，将顶点特征之间的相关性更好地融入了模型。随后，各种 GCN 的变体被提出，用于解决学习图嵌入的问题。GraphSAGE[10]对图中每个顶点的邻居顶点进行采样；GCN-LPA[11]分析了图卷积网络(GCN)和标签传播算法(LPA)之间的理论关系；CNMPGNN[12-13]是基于共同邻居的主题，其对结构模式进行了泛化和丰富；ACM-GCN[14]考虑到图结构和输入特征对 GNN(Graph Neural Networks)的影响，提出了自适应通道混合(ACM)框架，自适应地利用每个 GNN 层中的聚合、多样性和身份渠道来解决有害的异嗜性。

Metapath2vec[15]是基于元路径的异质图嵌入方法，其只能处理特定的一个元路径。ESim[16-17]可以利用多个元路径，但无法了解元路径的重要性。HAN[18]学习邻居的重要性和基于注意力机制的多个元路径。MAGNN[19]考虑了元路径中的中间节点，聚合元路径内和元路径间的信息。HetGNN[20-21]采用随机游走策略对邻居进行采样，使用专门的 Bi-LSTM 来整合异构节点特征和相邻节点。HetSANN[22]通过类型感知注意层来学习不同类型的相邻节点以及相关边。基于 Transformer[23]的架构，HGT[24]学习不同节点的特征以及与特定类型参数的关系。DGI[25]介绍了如何将 DIM[26-27]的最大化互信息思想应用到图领域中。而在 DGI 的基础上，HDGI[28]是第一个在异质图表示学习中应用最大化互信息的无监督方法。GIC[29]是利用集群级节点信息进行无监督图表示学习的方法，也是 DGI 工作的拓展，然而它只能应用在同质图中。

3　基于无监督集群级的科技论文异质图节点表示学习方法

为了学习科技论文数据的知识表征，本文提出了一种基于无监督集群级的科技论文异质图节点表示学习方法(UCHL)，用于获取科技论文异质图中节点的表示，该方法的整体框架如图 1 所示。

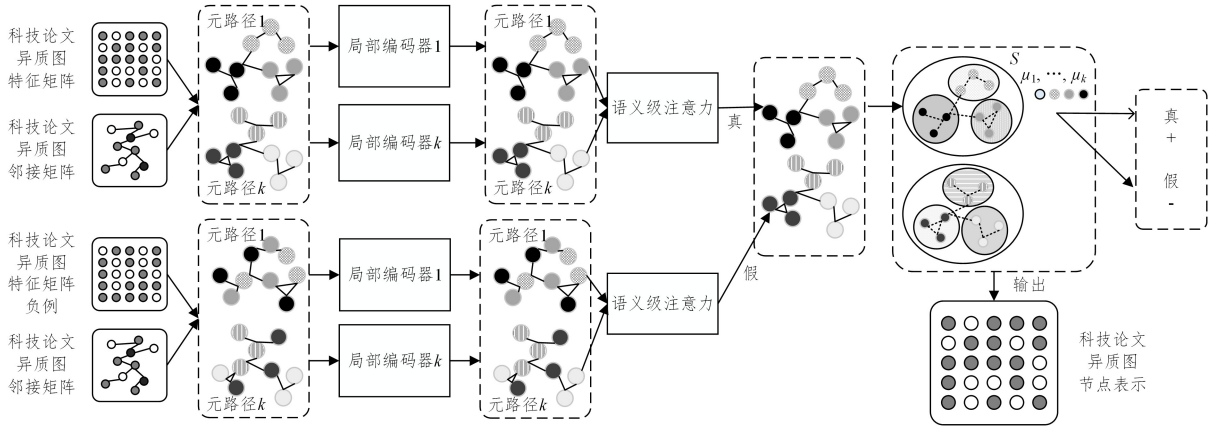


图 1　基于无监督集群级的科技论文异质图节点表示学习方法的框架

Fig. 1　Framework of UCHL

3.1　问题定义

科技论文异质图表示为 $G=(V,E)$，由一个对象集 V 和连接集 E 组成。异质图还与节点类型映射函数 $\varphi: V\to A$ 和连接类型映射函数 $\omega: E\to B$ 相关联。A 和 B 表示预定义对象类型和连接类型的集合，其中 $|A|+|B|>2$。基于科技论文异质图 G 和节点特征集 $\boldsymbol{X}$，UCHL 的目标是学习一个 d 维节点表示 $H\in R^{|V|\times d}$，并且 H 包含了图结构信息以及节点特征信息。本文只关注科技论文节点 V 的表示学习。

在科技论文异质图中，两个相邻节点之间可能被不同类型的边连接，定义这些边为元路径。因此元路径集合定义为 $\{\varphi_1,\varphi_2,\cdots,\varphi_k\}$，其中 φ_i 表示第 i 种元路径类型，例如在科技论文异质图中，同一个作者的两篇论文之间的边关系与同一个学科的两篇论文之间的边关系就是不同的元路径。对于元路径 φ_i，如果节点 $v_m\in V$ 和 $v_n\in V$ 之间的元路径关系存在，那么 v_m 和 v_n 是基于 φ_i 的邻接节点。这样的信息就可以用元路径邻接矩阵 $\boldsymbol{A}^i\in R^{|V|\times|V|}$ 来表示，若 $A^i_{mn}=A^i_{nm}=1$，则代表第 m 个节点与第 n 个节点是连接的，否则代表不连接。

3.2　UCHL 整体架构

UCHL 整体结构图如图 1 所示。针对所有特定的元路径节点进行表示学习，计算出每个元路径下的邻接矩阵 $\boldsymbol{A}^i$，并构建相应的负样本数据，将它们同时输入局部编码器进行编码。元路径局部编码器是分层结构，分别根据每个基于元路径的邻接矩阵学习单个节点表示；然后通过语义级注意力对其进行聚合，得到基于元路径的局部表示编码器的输出

节点表示 H；最终获得全局图表示 $s\in R^{1\times d}$ 以及集群簇 μ_c。

3.3 科技论文异质图的负样本构建

利用负样本构建出原始图中不存在的节点，在科技论文异质图 G 中，从基于元路径的邻接矩阵集合中获得了丰富而复杂的结构信息。负样本生成的过程如式(1)所示：

$$(\widetilde{\boldsymbol{X}},\{\boldsymbol{A}^1,\cdots,\boldsymbol{A}^k\})=C(\boldsymbol{X},\{\boldsymbol{A}^1,\cdots,\boldsymbol{A}^k\}) \tag{1}$$

其中，C 表示随机打乱函数。保持所有基于元路径的邻接矩阵不变，改变节点的索引以破坏它们之间的节点级连接。整个图的结构没有变化，但是每个节点对应的初始特征发生了变化。

3.4 科技论文异质图的层级局部编码器

基于元路径的层级局部编码器有两层结构。本文首先从每个基于元路径的邻接矩阵 $\boldsymbol{A}^i(i=1,2,\cdots,k)$ 中分别导出一个节点表示，然后通过注意力机制聚合所有元路径下的节点表示。采用图注意力神经网络作为一个层次局部编码器，推导出一个包含初始节点特征 $\boldsymbol{X}$ 和邻接矩阵 $\boldsymbol{A}^i(i=1,2,\cdots,k)$ 的节点表示 H^i，学习到的节点隐藏表示如式(2)所示：

$$\vec{h}_p^i=\Big\Vert_{k=1}^{K}\sigma\Big(\sum_{q\in N_p^i}\alpha_{pq}^i W\vec{x}_j\Big) \tag{2}$$

其中，$\Vert$ 是连接操作，W 是共享的特定元路径线性权重矩阵变换，N_p^i 是节点 p 基于 φ_i 的邻居节点集，α_{pq}^i 是基于 φ_i 的 p 与 q 两连接节点的注意力权重，K 是多头注意力机制中的头数。

经过层级局部编码器的学习后，可以获得基于不同语义的元路径节点表示集$\{H^1,H^2,\cdots,H^k\}$。基于每个元路径的节点学习的表示仅包含异质图中的语义特定信息，为了聚合节点的表示，添加了一个语义注意层来学习每个元路径应该分配的权重$\{S^1,S^2,\cdots,S^k\}$，语义注意力层学习的注意力权重由节点是否属于原始图的二元交叉熵损失来引导。

为了使基于不同元路径的表示具有可比性，需要使用线性变换来转换每个节点的表示，由共享权重矩阵 $\boldsymbol{W}_{sem}$ 和共享偏置向量 $\boldsymbol{b}$ 来参数化。基于不同元路径的表示的重要性将通过共享注意力向量 $\boldsymbol{q}$ 来衡量。元路径 φ_i 的重要性可以通过式(3)来计算：

$$e^i=\frac{1}{N}\sum_{j=1}^{N}\tanh(\boldsymbol{q}^{\mathrm{T}}\cdot[\boldsymbol{W}_{sem}\cdot h_j^i+\boldsymbol{b}]) \tag{3}$$

根据元路径的重要性，本文将使用 $softmax$ 函数对其进行归一化，如式(4)所示：

$$S^i=softmax(e^i)=\frac{\exp(e^i)}{\sum_{j=1}^{k}\exp(e^j)} \tag{4}$$

不同元路径的权重被用作系数进行线性组合，从而计算最终图全局表示 H，如式(5)所示：

$$H=\sum_{i=1}^{k}S^i\cdot H^i \tag{5}$$

通过层级局部编码器以及语义级注意力来获取真实的图节点表示 H 与假的图节点表示 $\widetilde{H}$ 之后，全局图表示 $s\in R^{1\times d}$ 是通过式(6)对所有节点表示进行计算得到的。

$$s=\sigma\Big(\frac{1}{N}\sum_{i=1}^{N}h_i\Big) \tag{6}$$

其中，σ 代表逻辑 sigmoid 函数，N 为节点个数，h_i 表示每个节点的表示。集群簇 μ_r 通过对细粒度表示进行聚类，然后计算集群簇内所有节点的平均值获得。$\mu_r(r=1,2,\cdots,R)$ 是由 K-means 聚类的可微分版本的层获取的。在最终优化时，μ_r 的更新可以通过式(7)和式(8)迭代进行。

$$\mu_r=\frac{\sum_i c_{ir}h_i}{\sum_i c_{ir}} \tag{7}$$

$$c_{ir}=\frac{\exp(-\beta\mathrm{sim}(h_i,\mu_k))}{\sum_j\exp(-\beta\mathrm{sim}(h_i,\mu_j))},j=1,\cdots,R \tag{8}$$

其中，$\mathrm{sim}(*,*)$表示两个实例之间的相似性函数；β 是超参数，它趋向于无穷大，为每个集群簇分配给出了一个二进制值。

为了估计和最大化互信息，根据每个节点 i 所属的集群簇计算 z_i，它代表每个节点的相应集群表示，然后最大化每个节点的 h_i 和 z_i 之间的相互信息。为了计算每个节点 i 的 z_i，本文应用了节点 i 所属集群的表示的加权平均值，如式(9)所示：

$$z_i=\sigma\Big(\sum_{r=1}^{R}c_{ir}\mu_r\Big) \tag{9}$$

其中，c_{ir} 与式(8)中的含义相同，代表节点 i 被分配到集群 r 的程度，是一个软分配值；σ 代表逻辑 sigmoid 函数。

3.5 科技论文异质图的互信息代理

本文定义了一个判别器，将其作为通过为正例分配比负例更高的分数来估计互信息的代理器，通过将真实图中节点表示 h_i 与节点集群表示 z_i 配对来获得正例，通过虚假图中节点表示 $\widetilde{h}_i$ 与节点集群表示 z_i 配对来获得负例。判别器函数 D 是估计节点表示和图全局表示之间的相互信息的代理，使用双线性评分函数，如式(10)所示：

$$D(h_i,z_i)=\sigma(h_i^T z_i) \tag{10}$$

其中，σ 代表逻辑 sigmoid 非线性函数。

判别器是一个标准的二元交叉熵损失函数，其目的是最大化联合分布(正例)中的样本的预期对数比和边际分布(负例)的乘积。正例是真实输入图 G 的 s 与 h_i 的配对，但负例是图 G 的 s 与 $\widetilde{h}_i$ 的配对，并且考虑到集群簇的信息，因此判别器 D 的损失函数由两部分组成，其中一部分 L_g 为图全局表示与节点表示之间的交叉熵损失，而 L_c 为集群表示与节点表示之间的交叉熵损失，分别如式(11)和式(12)所示：

$$L_g=\frac{1}{2N}\Big(\sum_{i=1}^{N}E_{(X,A)}[\log D(h_i,s)]+\sum_{i=1}^{N}E_{(\widetilde{X},\widetilde{A})}[\log(1-D(\widetilde{h}_i,s))]\Big) \tag{11}$$

$$L_c=\frac{1}{2N}\Big(\sum_{i=1}^{N}E_{(X,A)}[\log D(h_i,z_i)]+\sum_{i=1}^{N}E_{(\widetilde{X},\widetilde{A})}[\log(1-D(\widetilde{h}_i,z_i))]\Big) \tag{12}$$

UCHL 总的损失函数如式(13)所示：

$$L=\theta L_g+(1-\theta)L_c \tag{13}$$

其中，$\theta\in[0,1]$是一个超参数。本文通过不断优化式(13)中的损失，来获取一个最优解，以学习到科技论文异质图的节点表示。算法的流程如算法 1 所示。

算法 1 UCHL 算法

输入：科技论文异质图特征矩阵 **X**，科技论文异质图邻接矩阵 **A**

输出：科技论文异质图节点表示 H

1. 构建相应的负样本数据；

2. 通过层级局部编码器学习各元路径下的节点表示；
3. 通过注意力机制聚合各元路径下的节点表示；
4. 通过语义注意层来学习每个元路径应该分配的权重 $S^1, S^2, \cdots, S^k$；
5. 通过科技论文异质图的互信息代理来学习最终的节点表示。

4　实验结果与分析

4.1　数据集

为了验证提出的基于无监督集群级的科技论文异质图节点表示学习方法(UCHL)的性能，在 3 个真实的公开数据集上与其他相关算法进行了对比实验。

(1)DBLP 数据集：包含 14 328 篇论文(P)、4 057 位作者(A)、20 个会议(C)、8789 个术语(T)。作者分为 4 个领域：数据库、数据挖掘、机器学习、信息检索。本文使用元路径集{APA，APCPA，APTPA}来进行实验。

(2) ACM 数据集：提取发表在 KDD，SIGMOD，SIGCOMM，MobiCOMM 和 VLDB 上的论文，并将论文分为 3 类(数据库、无线通信、数据挖掘)。本文构建了一个异质图，包括 3 025 篇论文 (P)、5835 位作者 (A) 和 56 种主题 (S)。使用元路径集 {PAP，PSP} 来进行实验。

(3)AMINER 数据集：包含 4 472 位作者(A)、7 623 篇论文(P)、101 个会议(C)和 10 个论文分类(L)。本文使用元路径集 {APA，APCPA，APTPA} 进行实验。

数据集的相关细节介绍如表 1 所列。

表 1　异质图数据集详细信息

Table 1　Heterogeneous graph dataset details

Dataset	Relation	A	B	A-B	Metapath
DBLP	P-A	14328	4057	19645	APA
	P-C	14328	20	14328	APCPA
	P-T	14327	8789	88420	APTPA
ACM	P-A	3025	5835	9744	PAP
	P-S	3025	56	3025	PSP
AMINER	P-A	7623	4472	15213	APA
	P-C	7623	101	4158	APCPA
	P-L	7623	10	7623	APLPA

4.2　对比方法

本文将提出的 UCHL 与以下几种无监督表示学习方法进行比较。

Raw Feature：直接采用节点文本特征。

DeepWalk[1]：为同构图设计的基于随机游走的网络。

Metapath2vec[15]：基于元路径，但只能处理特定的一个元路径。

ESim[16]：从多个元路径中捕获语义信息。

同时，本文也对比了以下几种有监督的表示学习方法。

GCN[8]：用于同构图中节点分类的半监督方法。

GAT[9]：基于注意力机制的有监督表示学习方法。

HAN[18]：采用节点级注意力和语义级注意力捕获来自所有元路径的信息。

在实验中，一些边隐藏在输入图中，目标是基于计算的节点表示来预测这些边的存在，节点 i 与节点 j 之间边的概率由 $\sigma(h_i^{\mathrm{T}}h_j)$ 给出，其中 σ 为逻辑 sigmoid 函数。设置 5% 的正样本和负样本的边作为验证集，10% 的正样本和负样本的边作为测试集，学习的表示特征维度 $d=16$。测试了 ROC 曲线下的面积 AUC 分数，它等同于随机选择的边比随机选择的负边排名更高的可能性以及平均精度 AP 分数，即准确率-召回率曲线下的面积，这里准确率由 TP/(TP+FP) 计算得出，召回率由 TP/(TP+FN) 计算得出。其中，TP 表示正例被预测为正例数，FP 表示负例被预测为正例数，FN 表示正例被预测为负例数。

4.3　UCHL 的实验结果与分析

本节对提出的 UCHL 与其他方法进行了性能对比以及参数分析等。

在 DBLP 数据集上的对比实验结果如表 2 所列，由于提出的 UCHL 方法充分考虑了不同元路径对整体结果的影响，因此其在 AUC 和 AP 指标上比所有对比方法都取得了更好的性能。与性能指标排名第二的方法 HAN 相比，UCHL 在 *AUC* 和 *AP* 上分别提升了 2.59% 和 3.58%。而与直接使用原始特征相比，UCHL 在 AUC 和 AP 指标上都有超过 20% 的性能提升。

表 2　UCHL 与其他对比方法在 DBLP 上的性能比较

Table 2　Performance comparison of UCHL and other comparative methods on DBLP

(单位：%)

	AUC	*AP*
Raw Feature	70.24	69.73
DeepWalk	75.53	74.69
Metapath2vec	79.65	79.12
ESim	87.45	86.91
GCN	77.56	76.14
GAT	84.34	83.28
HAN	90.67	89.53
UCHL	93.26	93.11

表 3 列出了在 ACM 数据集上 UCHL 与其他对比方法的实验结果，可以观察到 UCHL 依然取得了最好的表现，高出原始特征或 DeepWalk 等传统方法约 20%。与对比方法中效果最好的有监督方法 HAN 和无监督方法 ESim 相比，UCHL 的 *AUC* 和 *AP* 分别提升了 5.22%，4.11% 和 4.58%，5.56%。相比其他数据集，UCHL 在 ACM 数据集上的提升幅度最大，其原因可能是 ACM 数据集规模最小。

表 3　UCHL 与其他对比方法在 ACM 上的性能比较

Table 3　Performance comparison of UCHL and other comparative methods on ACM

(单位：%)

	AUC	*AP*
Raw Feature	69.98	68.22
DeepWalk	72.82	71.97
Metapath2vec	76.27	73.71
ESim	88.37	86.13
GCN	78.02	77.36
GAT	84.68	82.87
HAN	87.26	87.11
UCHL	92.48	91.69

AMINER 数据集上的实验结果如表 4 所列。可以看出，提出的 UCHL 在 *AUC* 和 *AP* 上的性能依然优于所有对比方法，其中在原始特征的基础上，UCHL 在 *AUC* 和 *AP* 上分别达到了近 17% 和近 20% 的提升，而与次优的方法 HAN

相比，UCHL 在 AUC 上提升了 2.29%，在 AP 上提升了 1.07%。UCHL 结合了集群的簇关系并加入了互信息关系约束，在所有方法中取得了最佳性能。

表 4 UCHL 与其他对比方法在 AMINER 上的性能比较

Table 4 Performance comparison of UCHL and other comparative methods on AMINER

（单位：%）

	AUC	*AP*
Raw Feature	66.16	63.91
DeepWalk	68.94	67.39
Metapath2vec	71.39	68.97
ESim	83.68	81.61
GCN	74.33	73.31
GAT	79.62	79.17
HAN	83.14	82.90
UCHL	85.43	83.97

UCHL 提供了一个超参数 θ 来调整两部分的损失对模型整体效果的影响。为了验证式(14)中超参数 θ 对 UCHL 的影响，在 3 个数据集上分别进行了参数实验，实验结果如图 2 和图 3 所示。在[0.1,0.9]中搜索 θ，步长设置为 0.1，进行 9 次超参数 θ 不同的参数实验，记录实验在 3 个数据集上的结果，并且观察 UCHL 的 AUC 和 AP 的变化情况。

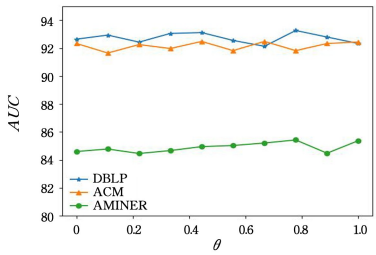


图 2 UCHL 在 3 个数据集上对不同 θ 的 AUC 值

Fig. 2 AUC values of UCHL on three datasets with different θ

从图 2 可以观察到，随着 θ 在 0.1 到 0.9 之间变化，UCHL 在 DBLP，ACM 和 AMINER 这 3 个数据集上 AUC 性能几乎保持不变，也就是说超参数的变化对所提方法最终影响不大。

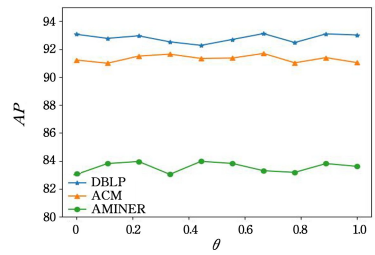
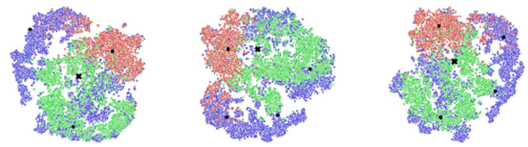


图 3 UCHL 在 3 个数据集上对不同 θ 的 AP 值

Fig. 3 AP values of UCHL on three datasets with different θ

图 3 的结果表明，UCHL 在 DBLP，ACM 和 AMINER 这 3 个数据集上 AP 性能随着 θ 波动而保持稳定，最优性能一般出现在 θ 处于[0.4～0.7]区间时。整体来看，UCHL 对参数并不敏感。为了验证集群簇数 R 对 UCHL 的影响，以 ACM 数据集为例，我们将学习到的科技论文表示利用 TSNE 降维绘制了 2D 空间图节点表示，由于 ACM 数据集中论文被分成了 3 类，因此在绘制 TSNE 降维图时将类别数设置为 3，取 R 分别为 3，4，5，绘制结果如图 4 所示。



(a)$R=3$ (b)$R=4$ (c)$R=5$

图 4 不同集群簇数 R 对 UCHL 学习的节点表示质量的影响

Fig. 4 Effects of different cluster numbers R on the quality of node representation learned by UCHL

使用轮廓分数 SIL 来评估结果。对于 ACM 数据集，图 4(a)是数据 $R=3$ 时 UCHL 学习到的节点表示绘制而成的，图 4(b)和图 4(c)分别是在 $R=4$ 和 $R=5$ 的条件下得到的。可以看出，UCHL 在 ACM 数据集上表现良好，可以根据节点表示来识别集群。R 的变化也不会引起节点表示质量的大幅度波动，这也可以从 SIL 分数的变化中得到印证，图 4 中 UCHL 在 $R=3$，$R=4$ 和 $R=5$ 这 3 种情况下的 SIL 分数分别为0.1185，0.1114，0.1243。这说明在互信息和集群簇中心的双重约束下，UCHL 可以快速聚集语义向量空间中相似的节点表示，并且同一簇中的节点也靠近该类的簇中心。

结束语 本文分析了互信息最大化以及集群簇中心在科技论文异质图表示学习中的影响，并提出了一种无监督集群级科技论文异质图节点表示学习方法(UCHL)，实现了对科技论文的深度语义表示学习。UCHL 利用元路径的结构对异构图中的关联语义进行建模，基于不同的元路径，将异构图分解为特定语义的同构图，并基于图卷积神经网络来捕获具有特定语义的节点的局部表示。UCHL 基于互信息最大化以及集群簇聚类技术，利用注意力机制聚合不同语义的节点表示，通过最大化局部与全局、局部与集群簇中心的互信息，在无需任何监督指导的情况下更好地学习包含图级结构信息的科技论文异质图节点表示。通过在 3 个科技论文数据集上对提出的 UCHL 进行了验证，实验结果表明，UCHL 学习到的科技论文异质图节点表示在科技论文异质图的链接预测任务中取得了最好的性能。

参 考 文 献

[1] 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.

[2] 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. 2016: 855-864.

[3] NARAYANAN A, CHANDRAMOHAN M, VENKATESAN R, et al. graph2vec: Learning distributed representations of graphs[J]. arXiv: 1707.05005, 2017.

[4] WANG D, CUI P, ZHU W. Structural deep network embedding [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1225-1234.

[5] TANG J, QU M, WANG M, et al. Line: Large-scale information

network embedding[C]// Proceedings of the 24th International Conference on World Wide Web. 2015:1067-1077.

[6] CHEN H,PEROZZI B,HU Y,et al. Harp: Hierarchical representation learning for networks[J]. arXiv:1706.07845,2017.

[7] MENG D Y,JIA Y M,DU J P,et al. Data-driven control for relative degree systems via iterative learning[J]. IEEE Transactions on Neural Networks,2011,22(12):2213-2225.

[8] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609.02907,2016.

[9] VELIČKOVIĆ P,CUCURULL G,CASANOVA A,et al. Graph attention networks[J]. arXiv:1710.10903,2017.

[10] 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.

[11] WANG H,LESKOVEC J. Unifying graph convolutional neural networks and label propagation[J]. arXiv:2002.06755,2020.

[12] ZHANG F,BU T M. CN-Motifs Perceptive Graph Neural Networks[J]. IEEE Access,2021,9:151285-151293.

[13] FANG Y K,DENG W H,DU J P,et al. Identity-aware CycleGAN for face photo-sketch synthesis and recognition[J]. Pattern Recognition,2020,102:1-36.

[14] LUAN S,HUA C,LU Q,et al. Is Heterophily A Real Nightmare For Graph Neural Networks To Do Node Classification?[J]. arXiv:2109.05641,2021.

[15] DONG Y,CHAWLA N V,SWAMI A. metapath2vec: Scalable representation learning for heterogeneousnetworks[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017:135-144.

[16] SHANG J,QU M,LIU J,et al. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks[J]. arXiv:1610.09769,2016.

[17] XUE Z,DU J P,DU D W,et al. Deep low-rank subspace ensemble for multi-view clustering[J]. Information Sciences,2019,482:210-227.

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

[19] 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.

[20] 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 Mining. 2019:793-803.

[21] HU W M,GAO J,LI B,et al. Anomaly detection using local kernel density estimation and context-based regression[J]. IEEE Transactions on Knowledge and Data Engineering,2018,32(2):218-233.

[22] 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. 2020,34(4):4132-4139.

[23] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. arXiv:1706.03762,2017.

[24] HU Z,DONG Y,WANG K,et al. Heterogeneous graph transformer[C]// Proceedings of the Web Conference. 2020:2704-2710.

[25] VELICKOVIC P,FEDUS W,HAMILTON W L,et al. Deep Graph Infomax[J]. ICLR (Poster),2019,2(3):4.

[26] HJELM R D,FEDOROV A,LAVOIE-MARCHILDON S,et al. Learning deep representations by mutual information estimation and maximization[J]. arXiv:1808.06670,2018.

[27] LI W L,JIA Y M,DU J P. Variance-constrained state estimation for nonlinearly coupled complex networks[J]. IEEE Transactions on Cybernetics,2017,48(2):818-824.

[28] REN Y,LIU B,HUANG C,et al. Heterogeneous deep graph infomax[J]. arXiv:1911.08538,2019.

[29] MAVROMATIS C,KARYPIS G. Graph InfoClust: Leveraging cluster-level node information for unsupervised graph representation learning[J]. arXiv:2009.06946,2020.

SONG Jie, born in 1997, master. His main research interests include data mining, information retrieval and machine learning.

LIANG Mei-yu, born in 1985, associate professor, Ph.D. Her main research interests include artificial intelligence, data mining, multimedia information processing and computer vision.

（责任编辑：何杨）