计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600085

基于融合元路径图卷积的异质网络表示学习

蒋宗礼　李苗苗　张津丽
北京工业大学信息学部　北京 100124
(jiangzl@bjut.edu.cn)

摘　要　近年来,网络表示学习(Network Representation Learning,NRL)作为一种在低维空间中表示节点来分析异质信息网络(Heterogeneous Information Networks,HIN)的有效方法受到越来越多的关注。基于随机游走的方法是目前网络表示学习常用的方法,然而这些方法大多基于浅层神经网络,难以捕获异质网络结构信息。图卷积神经网络(Gragh Convolutional Network,GCN)是一种流行的能对图进行深度学习的方法,能够更好地利用网络拓扑结构,但目前的GCN设计针对的是同质信息网络,忽略了网络中丰富的语义信息。为了有效地挖掘异质信息网络中的语义信息和高度非线性的网络结构信息,进而提高网络表示的效果,文中提出了一种基于融合元路径的图卷积异质网络表示学习算法(MG2vec)。该算法首先通过基于元路径的关联度量方法来获取异质信息网络中丰富的语义信息;然后采用图卷积神经网络进行深度学习,捕捉节点和邻居节点的特征,弥补浅层模型捕捉网络结构信息能力不足的缺陷,从而实现将丰富的语义信息和结构信息更好地融入低维的节点表示中。在数据集DBLP和IMDB上分别进行实验,相比DeepWalk,node2vec和Metapath2vec算法,所提MG2vec算法在多标签分类任务上的分类精确率更高且性能更优,精确率和Macro-F1值分别达到了94.49%和94.16%,且与DeepWalk相比分别最高提升了26.05%和28.73%。实验结果证明,MG2vec算法的性能优于经典的网络表示学习算法,具有更好的异质信息网络表示效果。

关键词:网络表示学习;异质信息网络;元路径;语义信息;网络结构信息;图卷积网络
中图法分类号　TP183

Graph Convolution of Fusion Meta-path Based Heterogeneous Network Representation Learning

JIANG Zong-li,LI Miao-miao and ZHANG Jin-li
Department of Information Technology,Beijing University of Technology,Beijing 100124,China

Abstract　In recent years,network representation learning has received more and more attention as an effective method for analyzing heterogeneous information networks by representing nodes in a low-dimensional space. Random walk based methods are currently popular methods to learn network embedding,however,most of these methods are based on shallow neural networks,which make it difficult to capture heterogeneous network structure information. The graph convolutional network (GCN) is a popular method for deep learning of graphs,which is known to be capable of better exploitation of network topology,but current design of GCN is intended for homogenous networks,ignoring the rich semantic information in the network. In order to effectively mine the semantic information and highly nonlinear network structure information in heterogeneous information networks,this paper proposes a heterogeneous network representation learning algorithm based on graph convolution of fusion meta-path (MG2vec)to improve the effect of network representation. Firstly,the algorithm obtains rich semantic information in heterogeneous information networks through relevance measurement based on meta-paths. Then the graph convolution network is used for deep learning to capture the characteristics of nodes and neighbor nodes,to make up for the deficiency of shallow model in capturing the information of the network structure,so as to better integrate rich semantic information and structural information into the low-dimensional node representation. Experiments are carried out on DBLP and IMDB,compared with DeepWalk,node2vec and Metapath2vec classical algorithms,the proposed MG2vec algorithm has higher classification accuracy and better performance in multi-label classification tasks,the precision and Macro-F1 value can be respectively up to 94.49% and 94.16%,and the both of values are up to 26.05% and 28.73% higher respectively than DeepWalk. The experimental results show that the performance of MG2vec algorithm is better than that of classical network representation learning algorithms,and MG2vec has better heterogeneous information network representation effect.

Keywords　Network representation learning,Heterogeneous information network,Meta-path,Semantics information,Network structure information,Graph convolutional networks

1　引言

信息网络,如社交与通信网络、论文引用网络、科技文献信息网络,在现实世界中无处不在。对信息网络数据的分析已经受到各行各业的普遍关注,其中,网络分析的一个关键问题是研究如何合理地表示网络中的特征信息。随着机器学习

到稿日期:2019-06-17　返修日期:2019-10-16　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:李苗苗(867743373@qq.com)

技术的发展，学习网络中的节点特征成为了一项新兴研究任务[1]。网络表示学习可以将网络数据表示成一种高效合理的向量形式，进而可将得到的向量表示运用到节点分类[2]、网络可视化等常见的应用任务中，这对解决信息网络背景下的各种实际应用问题具有重要意义[3]。

目前大多数传统网络表示学习主要关注由同一种类型的节点和边组成的同质信息网络，异质信息网络的相关研究较少。相比同质信息网络，异质信息网络中不同类型的节点和边包含了更加复杂的结构信息以及丰富的语义信息[4]，如何对异质信息网络中的节点或者边进行有效地表示学习是一个困难且有挑战性的问题。Dong 等[5]提出 Metapath2vec 方法来解决异质网络表示学习带来的挑战，但该方法采用的 skip-gram 模型是浅层模型，不能很好地捕获高度非线性网络结构，从而产生非最优网络表示结果。

近年来，图卷积神经网络(GCN)[6-7]在网络学习中的应用越来越多，它是一种非常强大的在图上进行机器学习的神经网络框架，即使是随机初始化的两层 GCN 也可以生成网络中节点的有用特征表示。GCN 适用于任意拓扑结构的图，在学习过程中能同时融合特征信息和结构信息，但该模型只能对同质网络进行深度学习，无法直接应用于异质网络，因此不能捕获异质网络的结构信息。

针对上述问题，本文提出了一种融合元路径的 GCN 异质网络表示学习模型 MG2vec。该模型旨在获取异质信息网络更多的结构信息和语义信息，通过图卷积神经网络融合元路径来实现异质网络表示学习，进而将学习到的向量表示更好地应用到网络分析任务中。首先，利用源对象和目标对象类型相同的元路径在异质信息网络上进行关联度量，通过计算所有目标类型对象之间的关联性构建相应的同质网络；然后，利用图卷积网络对计算出的同质网络进行深度学习，捕捉节点与邻居节点的特征，弥补浅层模型捕捉到的网络结构信息不足的缺陷，实现语义信息和结构信息的融合表示学习；最后，通过多标签分类任务来测试本文方法的有效性。

2 相关研究

网络表示学习[8]旨在将网络信息嵌入一个低维的潜在空间中，进而将学到的低维稠密特征向量作为机器学习算法的输入，从而实现网络分析任务。

Perozzi 等[9]受 Word2vec 神经网络语言模型[10-11]的启发，提出了基于随机游走的 DeepWalk 算法，将随机游走对网络结构采样得到的节点序列作为 skip-gram 模型的输入，从而得到网络表征，并使用随机梯度下降方法优化参数。随后，Tang 等[12]提出的 LINE 算法同时保留了网络的一阶和二阶信息，较好地保留了网络的局部结构和全局结构，从而获得了良好的表示效果。Grover 等[13]则是在 DeepWalk 的基础上通过改进随机游走的过程提出 node2vec 算法。该算通过引入参数 p 和 q 均衡游走过程中的广度优先搜索和深度优先搜索，通过最大化地保留网络邻近节点的相似度来提高节点的表示性能。Kipf 等[7]提出了一种直接作用于网络结构上的卷积神经网络(Convolutional Neural Networks，CNN)算法。该算法通过傅里叶变换将图的拉普拉斯矩阵的特征向量矩阵转化到空间上，然后进行卷积操作，利用基于边的标签传播规则对模型进行训练，从而实现半监督网络表示学习。深度神经网络对网络复杂的非线性关系具有强大的表示能力[14]，因此该算法能充分保留网络的结构信息，并可快速有效地将其应用到节点分类任务中。

上述网络表示学习算法都针对的是同质信息网络，忽略了网络中丰富的语义信息。而异质信息网络分布更广泛，能更好、更完整有效地反映现实世界。Dong 等[5]受 node2vec 算法的启发，利用元路径来指导随机游走，然后将改进的随机游走策略与 skip-gram 模型结合，建立了 Metapath2vec 算法。该算法训练出的节点表示向量同时包含了网络的结构信息和网络中的语义信息，提高了网络表征的效果，但该算法基于浅层神经网络，不能很好地保留网络结构信息。

3 基于融合元路径的 GCN 异质网络表示学习模型

3.1 问题定义

定义 1(异质信息网络，HIN[15]) 设网络 $G=(V,E)$，其中，V 是所有实体节点的集合，E 是所有关系边的集合。节点类型映射函数 $\varphi: V\rightarrow A$。边类型映射函数 $\psi: E\rightarrow R$。$\forall v\in V$，$\varphi(v)\in A$，对 $\forall e\in E$，$\psi(e)\in R$。当 $|A|>1$ 或 $|R|>1$ 时，网络被称为异质信息网络。

图 1 是一个典型的异质信息网络，包含 author，paper 和 conference 3 类节点，不同节点之间存在不同类型的关系。

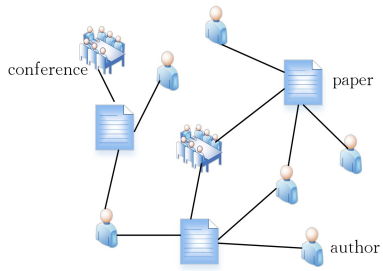

图 1 异质信息网络

Fig. 1 Heterogeneous information network

定义 2(网络表示学习，NRL) 设信息网络 $G=(V,E)$，对 $\forall v\in V$，学习低维向量表示 $r_v\in R^d$，r_v 是一个稠密的低维实数向量，其中 d 远小于 $|V|$。

定义 3(网络模式，Network Schema[16]) 给定一个异质信息网络 $G=(V,E)$，映射函数 $\varphi: V\rightarrow A$ 和 $\psi: E\rightarrow R$，其中 A 表示顶点类型，R 表示边类型，则其网络模式为 $T_G=(A,R)$。

图 2 给出了图 1 所示网络的网络模式，描述了网络包含的类型和相应的关系。

图 2 网络模式

Fig. 2 Network schema

定义 4(元路径，Meta-path[17]) 元路径 $P=(A_1A_2\cdots A_{l+1})$ 是网络模式 $T_G=(A,R)$ 上的链接两类对象的一条路径：$A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots A_t\xrightarrow{R_t}A_{t+1}\cdots\xrightarrow{R_{l-1}}A_l$。其中 $R=R_1\circ R_2\circ\cdots\circ R_{l-1}$，$\circ$ 代表对象类型之间的一种复合关系。

3.2 融合元路径的 GCN 模型

3.2.1 基于元路径的关联度量

Gupta 等[18]提出了一种基于元路径的新的关联度量方

法，它可以度量异质信息网络中相同类型对象或不同类型对象之间的关联程度。给定一条元路径 $P=(A_1A_2\cdots A_{l+1})$，A_1 和 A_{l+1} 的类型不同，则首先根据式(1)计算只有 A_1 和 A_{l+1} 两种类型的对象的异质网络的二分表示 M；然后根据式(2)计算源对象 $a_{1i}\in A_1$ 和目标对象 $b_{(l+1)j}\in A_{l+1}$ 之间的关联性。

$$M=\boldsymbol{W}_{A_1A_2}\times\boldsymbol{W}_{A_2A_3}\times\cdots\times\boldsymbol{W}_{A_lA_{l+1}} \tag{1}$$

其中，$\boldsymbol{W}_{A_iA_j}$ 是 A_i 和 A_j 的邻接矩阵。

$$Rel(a_i,b_j|P)=\frac{w_{a_ib_j}\left(\frac{1}{deg(a_i)}+\frac{1}{deg(b_j)}\right)}{\frac{1}{deg(a_i)}\sum_j w_{a_ib_j}+\frac{1}{deg(b_j)}\sum_i w_{a_ib_j}} \tag{2}$$

其中，$w_{a_ib_j}$ 是二分表示 $M(a_i,b_j)$ 的值，表示链接对象 a_i 和 b_j 的路径数量；$deg(a_i)$ 和 $deg(b_j)$ 分别是对象 a_i 和 b_j 在二分表示中的度。

因此计算元路径的源对象和目标对象的类型不同时，两者之间的关联性 $DPRel(a_i,b_j|P)$ 用式(3)计算。

$$DPRel(a_i,b_j|P)=Rel(a_i,b_j|P) \tag{3}$$

由于需要构建基于元路径的目标对象所对应的同质网络，因此还要计算相同类型的对象之间的关联性，即元路径的源对象和目标对象的类型相同。例如，在 DBLP 数据集中选取元路径 APCPA，然后基于这条元路径计算 author 的关联性，其中 A 表示 author，P 表示 paper，C 表示 conference，C 是元路径中间对象的类型。本文将元路径 APCPA 划分成两个等长子路径 $P_L=APC$ 和 $P_R=CPA$，C 为中间对象集合，根据子路径 P_L 计算源对象 $a_i\in A$ 和所有 $b_k\in C$ 的关联性 $Rel(a_i,b_k|P)$。同样地，根据子路径 P_R 的反向路径 P_R^{-1} 计算目标对象 $a_j\in A$ 和所有 b_k 的关联性 $Rel(a_j,b_k|P_R^{-1})$。然后利用 Tanimoto 系数计算 a_i 和 a_j 的关联性，计算公式如下：

$$\boldsymbol{X}=Rel(a_i,b_k|P_L) \tag{4}$$

$$\boldsymbol{Y}=Rel(a_j,b_k|P_R^{-1}) \tag{5}$$

$$DPRel(a_i,a_j|P)=\frac{\boldsymbol{X}\cdot\boldsymbol{Y}}{|\boldsymbol{X}|^2+|\boldsymbol{Y}|^2-\boldsymbol{X}\cdot\boldsymbol{Y}} \tag{6}$$

其中，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别是 a_i 和 a_j 与所有 b_k 的相关向量。

本文通过计算基于元路径的所有目标对象 author 之间的关联性最终获得相应的关联矩阵，并以此表示由目标对象 author 组成的同质网络。

3.2.2 MG2vec 模型

图卷积网络(GCN)。GCN 是卷积网络的推广，每个 GCN 层都具有一个激活函数，用于卷积节点特征的非线性变换，能同时对节点特征信息与结构信息进行端对端学习。相对于浅层模型，GCN 能提取非线性的复杂特征，达到降维的目的，同时也能捕获高度非线性的网络结构。

GCN 模型的目标是基于图 $G=(V,E)$ 学习一个特征函数，并将以下变量作为输入：

(1)特征矩阵 $\boldsymbol{X}$，每个节点 i 的特征描述为 x_i，它们组成一个 $N\times D$ 的特征矩阵 $\boldsymbol{X}$，其中 N 表示节点个数，D 表示输入特征的维度；

(2)图的矩阵表示，通常用邻接矩阵 $\boldsymbol{A}$ 来表示。

最终输出一个 $N\times F$ 的节点标签概率矩阵 $\boldsymbol{Z}$，其中 F 表示标签类型的个数。

GCN 模型的每个神经网络层都有一个非线性激活函数，定义如下：

$$\boldsymbol{H}^{(l+1)}=f(\boldsymbol{H}^{(l)},\boldsymbol{A}) \tag{7}$$

其中，$\boldsymbol{H}^{(0)}=\boldsymbol{X}$，$\boldsymbol{H}^{(L)}=\boldsymbol{Z}$，$L$ 为卷积层数。

GCN 模型的传播规则如式(8)所示：

$$f(\boldsymbol{H}^{(l)},\boldsymbol{A})=\sigma(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{8}$$

其中，$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$，$\boldsymbol{I}$ 是单位矩阵，这保证了特征矩阵与邻接矩阵相乘时，对每个节点的所有邻接节点的特征向量加和时也能加入节点本身；$\hat{\boldsymbol{D}}$ 是对角矩阵，$\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}$ 表示归一化操作，避免运算的重复操作有可能导致的数据不稳定以及梯度爆炸或消失的问题；$\boldsymbol{H}^{(l)}\in R^{N\times D}$ 表示第 l 层的激活值，$\sigma(\cdot)$ 表示激活函数。

由于在基于元路径的关联度量部分学习出来的关联矩阵包含自链接，卷积过程中不需要再进行自链接操作，因此卷积算子的定义为：

$$g*\boldsymbol{X}=\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{A}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X} \tag{9}$$

本文的分类实验采用的是 2 层 GCN 模型，卷积层采用的是 $relu$ 激活函数，因此分类目标函数如式(10)所示。$\boldsymbol{Z}$ 给出节点标签的概率分布后，使用交叉熵作为损失函数。

$$\boldsymbol{Z}=softmax(g*relu(g*\boldsymbol{X}\boldsymbol{W}^{(0)})\boldsymbol{W}^{(1)}) \tag{10}$$

其中，$\boldsymbol{W}^{(0)}\in R^{D\times H}$ 表示输入层和隐藏层之间的含有 H 维特征映射的权重矩阵；D 表示输入特征维度；$\boldsymbol{W}^{(1)}\in R^{H\times F}$ 表示隐藏层到输出层之间的权重矩阵；$\Phi=g*\boldsymbol{X}\boldsymbol{W}^{(0)}$ 作为最终学习到的节点表示。

MG2vec 模型训练算法如算法 1 所示。

算法 1 MG2vec 模型训练算法

输入：异质信息网络 G=(V,E)，元路径 P

输出：节点向量表示 $\Phi\in R^{N\times H}$

初始化 **X**、关联矩阵 **A**、权重矩阵 $\mathbf{W}^{(0)}$ 和 $\mathbf{W}^{(1)}$、隐藏层维度 H、学习率 α、迭代次数 epoch

1. 根据 P 将元路径划分成两个等长子路径 P_L 和 P_R；
2. 使用式(4)计算 Rel $(a_i,b_k|P_L)$；
3. 使用式(5)计算 Rel $(a_j,b_k|P_R^{-1})$；
4. 使用式(6)计算目标对象的关联性，并更新关联矩阵 **A**；
5. for i=1 to epoch do
6. 使用式(8)进行前向传播，最后一层逐行应用 softmax 函数；
7. 计算已知节点标签上的交叉熵损失；
8. 反向传播损失并使用 Adam 算法更新每层中的权重矩阵 **W**；
9. end

4 实验与结果分析

4.1 数据集与评价指标

本文通过对标签进行分类实验来验证 MG2vec 模型的有效性。实验中使用 DBLP 数据集和 IMDB 数据集。

(1)DBLP 是一个计算机科学书目数据集。实验从 DBLP 数据集中抽取来自数据库(Data Base)、信息检索(Information Retrieval)、数据挖掘(Data Mining)和机器学习(Machine Learning)4 个领域的 28702 位作者、13214 篇论文以及 20 个会议。

(2)IMDB 是一个互联网电影数据集。实验选取动作片(actions)、喜剧(comedy)和戏剧(drama)3 种类型的 8313 部电影(movies)、8298 个演员(actors)和 8284 个导演(directors)组成的异质信息网络。

实验采用精确率(precision)和文献[19]引入的 Macro-F1 和 Micro-F1 指标来评价分类性能。

(1)精确率是预测标签集中被正确分类的标签的比例,其值越大表明分类效果越好,计算公式如下:

$$精确率(P)=\frac{被正确分类的标签个数}{应该被分到该类的标签个数} \tag{11}$$

(2)*Macro-F*1 表示基于所有类别的 *F*1 的平均值,其值越大表明算法的分类性能越优,具体计算如式(14)所示:

$$macroP=\frac{1}{n}\sum_{1}^{n}P_i \tag{12}$$

$$macroR=\frac{1}{n}\sum_{1}^{n}R_i \tag{13}$$

$$Macro\text{-}F1=\frac{2\times macroP\times macroR}{macroP+macroR} \tag{14}$$

其中,P_i 和 R_i 指第 i 个标签的精确率和召回率,召回率指被预测到的标签在样本真实标签集中所占的比例。

(3)Micro-F1 是相对于 F1 值的微观平均,是通过对数据集中的每一个实例不分类别地的进行统计并建立全局混淆矩阵后,利用公式计算得到的,具体计算公式如下:

$$microP=\frac{\overline{TP}}{\overline{TP}+\overline{FP}} \tag{15}$$

$$microR=\frac{\overline{TP}}{\overline{TP}+\overline{FN}} \tag{16}$$

$$Micro\text{-}F1=\frac{2\times microP\times microR}{microP+microR} \tag{17}$$

其中,$\overline{TP}$,$\overline{FP}$和$\overline{FN}$分别指真正例个数之和、假正例个数之和以及假负例个数之和。*Micro-F*1 值越大表明算法的分类性能越好。

4.2 实验设置

本文将 MG2vec 算法与 DeepWalk,node2vec 和 Metapath2vec 算法作对比。

DeepWalk 通过在网络中随机游走来获取节点序列,然后利用 skip-gram 模型来学习低维的节点特征表示。实验中采用默认参数设置,窗口大小 w 为 10,随机游走序列长度 l 为 80,每个节点为起始节点的游走次数 r 为 10,训练出来的向量维度 d 为 128。

node2vec 由参数 p 和 q 控制随机游走的方向,以平衡使随机游走的深度和广度。p 和 q 通过网格搜索从{0.25,0.5,1,2,4}中择优选择[13]。当 $p=q=1$ 时,node2vec 和 DeepWalk 等价。

Metapath2vec 基于给定的一条元路径指导随机游走,从而获取节点序列,然后基于异质的 skip-gram 模型得到节点的向量表示。负采样大小为 5,其余参数设置与 DeepWalk 保持一致。

MG2vec 基于元路径获取目标对象的同质网络,然后利用图卷积网络进行深度学习以获得节点的低维表征。图卷积网络隐藏层的维度为 128,即学习到的节点向量维度为 128,学习速率 α 为 0.01。

4.3 多标签分类实验

根据节点的类别标注信息对节点进行合理地分类是网络数据分析任务中一个最常见的场景。例如,在社交网络中识别出具有相同兴趣爱好的人。兴趣爱好是对目标对象分类的类别标注信息,是其分类的依据。

本文实验在 DBLP 数据集和 IMDB 数据集上分别对 DeepWalk,node2vec,Metapath2vec 和 MG2vec 算法进行评测。首先使用完整数据集学习节点表示,然后从有标签的节点及其部分标签中随机选择 10%~90%作为训练集,将剩余部分作为测试集,每个比例重复进行分类实验 10 次,取 10 次 *presicion*,*Micro-F*1 和 *Macro-F*1 的均值进行比较。在 DBLP 数据集上,Metapath2vec 和 MG2vec 模型都选取了 APCPA 元路径进行分类实验,其中以作者所属的领域为标签对数据集进行划分,实验结果如图 3 所示。可以看出,仅利用网络结构信息的 DeepWalk 算法和 node2vec 算法的性能都较低;而 Metapath2vec 在捕获结构信息的同时利用元路径来获取异质网络中的语义信息,相比 DeepWalk 算法和 node2vec 算法,其在精确率上分别最高提升了 26.99%和 15.74%,这表明异质网络中丰富的语义信息有助于提高节点向量在分类任务中的准确性。本文算法 MG2vec 与同样考虑语义信息和结构信息的 Metapath2vec 算法相比更具优势,在 *Macro-F*1 和 *Micro-F*1 方面比 Metapath2vec 算法分别提高了 6.73%~50.32%和 6.91%~41.29%,尤其是在只有 10%的标签节点作为训练集时本文算法更具优势。这表明本文算法更能捕获网络中的结构信息,从而更好地融合到网络表示学习中。

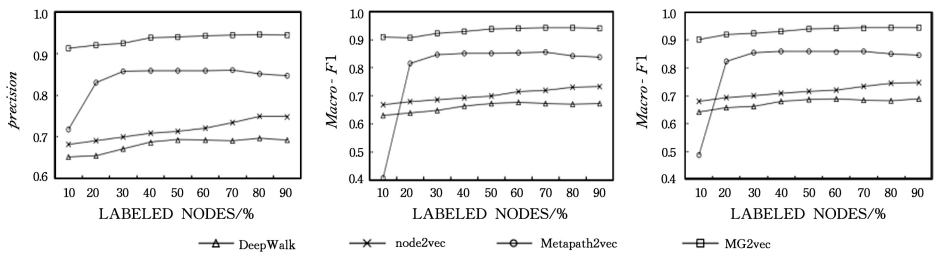

图 3 DBLP 数据集上的多标签分类结果

Fig. 3 Multi-label classification results in DBLP

在 IMDB 数据集上的分类实验中,Metapath2vec 模型和 MG2vec 模型都选取元路径 MAM(电影-演员-电影)进行关联度量,并以电影所属类型为标签对数据集进行划分,实验结果如图 4 所示。随着训练集比例的增大,各模型的性能均有提高,其中 MG2vec 的性能表现得更加稳定,且分类效果一直比 DeepWalk,node2vec 和 Metapath2vec 算法好,*Micro-F*1 值和 *Macro-F*1 值分别至少提高了 8.11%和 9.65%,而且 MG2vec 在只有 10%的标签节点作为训练集时的效果优于其他算法在 90%的标签节点作为训练集时的效果。这表明融合元路径的 GCN 模型相比 skip-gram 模型这种浅层模型能更有效地挖掘语义信息和高度非线性的网络结构信息,从而实现两方面信息的融合,学习的节点表示向量更具有鲁棒性。

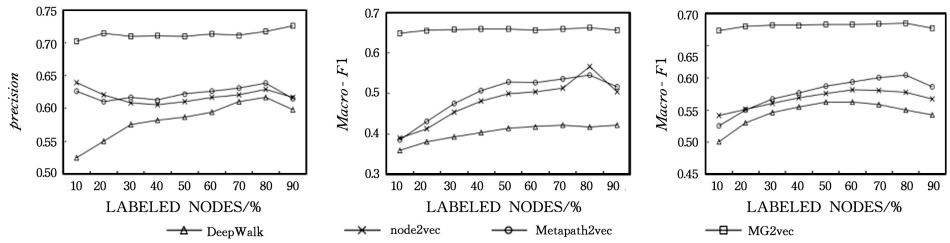

图 4 IMDB 数据集上的多标签分类结果

Fig. 4 Multi-label classification results in IMDB

结束语 本文介绍了一种基于融合元路径的 GCN 模型的异质网络表示学习模型，利用异质网络元路径的关联度量来获取目标节点的同质网络，然后利用 GCN 模型对学习到的同质网络进行深度学习，实现异质网络语义信息和网络结构信息的融合表示学习，以获得更具有鲁棒性的节点表示。在真实数据集上的实验证明了本文算法能够有效提升节点表示在多标签分类任务中的性能。但本文算法在训练过程中随着节点数量的增加，模型的训练时间开销较大，因此未来将进一步研究融合元路径的 GCN 模型的并行化计算方法，以提升算法的运行速度，并将本文方法扩展到其他网络分析任务中进行更深入的研究。

参 考 文 献

[1] TU C C, YANG C, LIU Z Y, et al. Network representation learning: an overview [J]. Scientia Sinica Informations, 2017, 47(8): 980-996.

[2] SHEIKH N, KEFATO Z T, MONTRESOR A. Semi-Supervised Heterogeneous Information Network Embedding for Node Classification using 1D-CNN[C]// 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, 2018: 177-181.

[3] YIN Y, JI L X, HUANG R Y, et al. Research and development of network representation learning[J]. Chinese Journal of Network and Information Security, 2019, 5(2): 77-87.

[4] JIANG Z L, ZHANG J L, DU Y P, et al. Hierarchical construction and node classification of heterogeneous network based on stacked denoising autoencoder[J]. Journal of Beijing University of Technology, 2018, 44(9): 1217-1226.

[5] 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. ACM, 2017: 135-144.

[6] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]// Advances in neural information processing systems. 2016: 3844-3852.

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

[8] ZHANG D, YIN J, ZHU X, et al. Network representation learning: A survey [J]. IEEE transactions on Big Data, 2017, PP(99): 1-1.

[9] PEROZZI B, AlRFOU R, SKIENA S. Deepwalk: Online learning of social representations [C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 701-710.

[10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in neural information processing systems. 2013: 3111-3119.

[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[12] 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. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[13] 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. ACM, 2016: 855-864.

[14] LE C Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[15] ZHANG J, JIANG Z, LI T. CHIN: Classification with META-PATH in Heterogeneous Information Networks[C]// International Conference on Applied Informatics. Springer, Cham, 2018: 63-74.

[16] SHI C, LI Y, ZHANG J, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 17-37.

[17] HUANG Z, ZHENG Y, CHENG R, et al. Meta structure: Computing relevance in large heterogeneous information networks [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1595-1604.

[18] GUPTA M, KUMAR P, BHASKER B. A new relevance measure for heterogeneous networks[C]// International Conference on Big Data Analytics and Knowledge Discovery. Cham: Springer, 2015: 165-177.

[19] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM computing surveys (CSUR), 2002, 34(1): 1-4.

JIANG Zong-li, born in 1956, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include network information search and processing.

LI Miao-miao, born in 1994, postgraduate. Her main research interests include network representation learning.