高技术通讯 2021 年 第31 卷 第12 期:1229-1238

doi:10.3772/j.issn.1002-0470.2021.12.001

基于局部图互信息最大化的异构图神经网络方法①

朱志华②*** 范鑫鑫** 毕经平** 武 超***

(*中国科学院大学 北京 100049)

(**中国科学院计算技术研究所 北京 100190)

(***中国电子科技集团公司电子科学研究院 北京 100041)

摘 要 针对现有的基于互信息最大化的异构图神经网络(HGNN)方法因图读出操作的单射限制、粗粒度的特征保留而无法适用于现实网络的问题,提出一种基于局部图互信息最大化的、无监督的异构图神经网络方法。该方法使用元路径对异构图中涉及到的语义关系进行建模,并利用图卷积模块和语义级别的注意力机制来捕获单个节点的局部表征。该方法通过最大化单个节点与局部子图间的互信息,有效地学习高阶节点表征。实验结果表明,该方法相比基于全局图互信息的方法,可以将数据集 DBLP/IMDB 上的节点分类任务的微值 F1(micro-F1)提高大约 3%/9%,同时将 DBLP/IMDB 上的节点聚类任务的调整兰德系数(ARI)提高约 23%/46%。

关键词 异构图(HG);图神经网络(GNN);互信息;无监督方法;图表示学习

0 引 言

异构图(heterogeneous graph, HG)作为数据挖掘中一个新的发展方向[1],为研究者提供了一种融合多种异质信息的有效工具。同时,图表示学习[2]作为一种学习节点低维向量表征的便捷工具,为下游各种应用,如推荐[3]、检索[4]、用户去匿名化[5]等,提供有效的支持。相比于传统的异构图表示学习方法,异构图神经网络(heterogeneous graph neural network, HGNN)由于其强大的表达能力及有效结合节点属性特征与结构信息的特点,开始成为研究重点。然而,当前大部分的异构图神经网络都是半监督模式的,即需要充足的带标签的样本进行模型的训练。但是,在现实场景中,通常无法获得充足的带标签的数据,从而限制了这些算法的使用。

为了应对训练样本稀缺的问题,无监督的异构图神经网络引起了学者们的广泛研究兴趣。现有的无监督的异构图神经网络主要分为两类,即基于近邻的方法[6-7]和基于互信息的方法[8]。其中,基于近邻的方法仅可以保留有限范围(低价)的节点相似度,缺乏保留高价甚至是全局结构信息的机制。为了保留图的全局结构信息,深度图互信息最大化(deep graph infomax, DGI)[8]与深度异构图互信息最大化(heterogeneous deep graph infomax, HDGI)[9]等方法提供了一种同时考虑全局和局部图结构的新方向,即最大化节点局部表征与全局图表征之间的互信息,并获得了很好的效果。但是,全局图表征通常只能够对粗粒度的结构信息进行保留,无法表达节点局部结构中近邻的特征及其分布的信息,易导致节点表征发生过平滑(over-smoothing);同时,DGI 与 HDGI 中使用的图读出操作(readout)需要满足单射(injective)限制,但在实际情况下该限制过于严格。如果图读出操作不是单射的,则全局图表征中

① 国家重点研发计划(2017YFC0820700)和国家自然科学基金(61702470)资助项目。

② 男,1992 年生,博士生;研究方向:图数据挖掘,推荐系统,图神经网络;联系人,E-mail: zhuzhihua@ict.ac.cn。

(收稿日期:2020-10-10)

包含的输入图信息将随着图大小的增加而减少,从而导致节点局部表征质量下降。

针对该问题,Peng 等人[10]提出图互信息(graphical mutual information,GMI)的概念,通过比较由节点 k 阶近邻组成的子图与每个节点的表征向量直接获得互信息,实现对近邻的特征及其分布等细粒度信息的提取。然而,该概念仅针对同构图提出,无法直接应用到异构图当中。换句话说,GMI 无法适应异构图中异质性(heterogeneity)产生的各异节点分布与节点输入特征。此外,异构图中节点间通常存在不同语义的关系,并且这些关系之间表现出不同程度的兼容性。在没有先验知识的指导下,会使得模型更倾向于最大化某些特定关系上的图互信息,从而忽略其他可能存在的语义关系,即使得模型发生语义层面上的过拟合问题。

针对上述问题,本文提出了一种无监督的异构图神经网络方法,即基于局部异构图互信息最大化(heterogeneous graphical mutual infomax,HGMI)的方法。该方法首先利用元路径(meta-path)[1]对异构图中涉及的语义关系进行建模,然后利用图卷积模块和语义级别的注意力机制来融合不同的关系语义,并为每个节点生成有效的局部表征。该方法将图互信息应用到异构图中,通过最大化单个节点与局部子图间在拓扑以及输入特征上的互信息,来处理无监督的设置;同时通过在目标函数中共享语义级别的注意力权重,使得模型对所有语义关系均保持一定的关注度,以解决语义层面上可能发生的过拟合问题。本文的主要贡献如下:(1)提出了一种无监督的、基于局部图互信息的异构图神经网络模型;(2)提出了一种注意力平衡机制,用于防止语义层面过拟合的发生;(3)基于真实的异构图数据集进行了实验,相比基于全局图互信息的方法,可以将数据集 DBLP/IMDB 上的节点分类任务的 micro-F1 提高大约 3%/9%,同时将 DBLP/IMDB 上的节点聚类任务的调整兰德系数(adjusted Rand index, ARI)提高约 23%/46%。

本文剩余部分总结如下。第 1 节介绍了异构图表示学习与异构图神经网络的相关工作。第 2 节介绍了本文中使用的基本符号和相关问题定义,包括异构图与图互信息的定义。第 3 节详细描述了本文提出的基于局部图互信息最大化的异构图神经网络模型 HGMI。第 4 节通过充分的实验对本研究中提出的方法进行了有效的验证。第 5 节对全文内容进行了总结。

1 相关工作

现实世界中图结构具有普遍性,图表示学习已成为一个备受关注的主题[2]。作为包含丰富结构信息的数据类型,许多模型[11-12]基于图的结构学习节点的向量表征。DeepWalk[13]利用 Skip-Gram,通过在图上进行一组随机游走来学习节点嵌入。此外,一些方法[14-15]则通过矩阵分解来提取结构信息。但是,以上所有方法只能用于同构图,无法解决异构图中的图表示学习问题。

为了处理图的异质性,metapath2vec[16]利用预先定义的元路径指导随机游走进行采样,并通过异构图中的 Skip-Gram 学习节点的表征。HIN2Vec[17]则在执行预测任务的同时,学习节点和元路径的表征向量。Wang 等人[18]通过添加注意力机制,使得模型可以有效地学习来自多个、由元路径定义的同构图的信息。从属性图的角度进行考虑,SHNE[19]通过异构 Skip-Gram 和深度语义编码的联合优化来捕获结构紧密性和非结构化语义关系。另外,许多面向知识图谱的方法[20-22]通常也可以应用于其他异构图。

随着深度学习的成功,图神经网络在图表示学习中取得了巨大的进展。图神经网络的核心思想是通过神经网络聚合邻居的特征信息,学习结合节点独立信息和图中相应结构信息的新的特征。大多数的图神经网络是基于半监督/监督学习的,包括图卷积网络(graph convolutional network,GCN)[23]、图注意力网络(graph attention network, GAT)[24]、GraphRNN[25]和 SplineCNN[26]。而无监督的图神经网络主要分为基于随机游走的方法[27-28]和基于互信息的方法[8]。

与传统的图神经网络不同,异构图神经网络需要解决异构图中异质性带来的一系列问题,如不同

类型、不同语义的节点与边。同样,大多数的异构图神经网络也是基于半监督/监督学习的,包括关系图卷积网络(relational graph convolutional network, RGCN)[20]和异构图注意力网络(heterogeneous graph attention network, HAN)[18]等。而无监督的异构图神经网络则主要分为基于近邻的方法和基于互信息的方法。

2 问题定义

2.1 异构图

一个异构图可以表示为节点与边的集合 $G=(V,E)$,该图具有一个节点类型映射函数 $\phi:V\rightarrow T$ 和一个边类型映射函数 $\psi:E\rightarrow R$,并且满足 $|T|+|R|>2$。另外,节点的属性和内容可以编码为初始特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{|V|\times D}$。

异构图表示学习任务旨在学习包含 G 的结构信息和 X 的节点属性信息的低维节点表征 $\boldsymbol{H}\in\mathbb{R}^{|V|\times d}$。本文使用 V_t 表示目标类型的节点集合。为了简化问题设置,利用对称且无向的元路径来表示目标类型节点 V_t 之间的紧密度。形式上,路径 $v_{t1}\xrightarrow{R_1}v_{t2}\xrightarrow{R_2}\cdots\xrightarrow{R_{n-1}}v_{tn}$ 被定义为节点 v_{t1} 和 v_{tn} 之间的元路径。进一步地,本文将使用的元路径集表示为 $\boldsymbol{\Phi}=\{\boldsymbol{\Phi}_1,\boldsymbol{\Phi}_2,\cdots,\boldsymbol{\Phi}_P\}$,其中 $\boldsymbol{\Phi}_i$ 表示第 i 个元路径类型。基于定义的元路径可以生成相应的邻接矩阵集合 $\boldsymbol{A}^{\Phi}=\{\boldsymbol{A}^{\Phi_1},\boldsymbol{A}^{\Phi_2},\cdots,\boldsymbol{A}^{\Phi_P}\}$,其中,$\boldsymbol{A}^{\Phi_i}\in\mathbb{R}^{|V_t|\times|V_t|}$。

2.2 图互信息

形式上,节点 v_i 的表征 $\boldsymbol{h}_i$ 和其局部子图 $G_i=(\boldsymbol{X}_i,\boldsymbol{A}_i)$ 之间的图互信息可以表示为局部互信息(即节点与一个近邻间的互信息)的加权和[10]:

$$I(\boldsymbol{h}_i;G_i)=\sum_j^{i_n}w_{ij}I(\boldsymbol{h}_i;\boldsymbol{x}_j)+I(w_{ij};a_{ij}) \quad (1)$$

$$w_{ij}=\mathrm{sigmoid}(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{h}_j) \quad (2)$$

其中,i_n 表示 $\boldsymbol{X}_i$ 中节点的数目,a_{ij} 是邻接矩阵 $\boldsymbol{A}_i$ 中的边权重,w_{ij} 表示局部互信息 $I(\boldsymbol{h}_i;\boldsymbol{x}_j)$ 对全局互信息 $I(\boldsymbol{h}_i;G_i)$ 的贡献。

相应地,在异构图中,给定邻接矩阵集合 $\boldsymbol{A}^{\Phi}$,异构图互信息可以表示为不同邻接矩阵中给定节点 v_i 与其对应子图 $G_i^{\Phi_t}$ 间互信息的和:

$$\tilde{I}(\boldsymbol{h}_i;G_i)=\sum_t^P\tilde{I}(\boldsymbol{h}_i;G_i^{\Phi_t}) \quad (3)$$

$$\tilde{I}(\boldsymbol{h}_i;G_i^{\Phi_t})=\sum_j^{i_n^{\Phi_t}}w_{ij}I(\boldsymbol{h}_i;\boldsymbol{x}_j)+I(w_{ij};a_{ij}^{\Phi_t}) \quad (4)$$

其中,$i_n^{\Phi_t}$ 表示节点 v_i 在邻接矩阵 $\boldsymbol{A}^{\Phi_t}$ 中的 k 阶近邻,$a_{ij}^{\Phi_t}$ 是邻接矩阵 $\boldsymbol{A}^{\Phi_t}$ 中的节点 v_i 与节点 v_j 间的边权重。

3 基于局部图互信息最大化的异构图神经网络模型

3.1 模型框架

基于局部图互信息最大化的异构图神经网络模型主要由 2 个模块组成,即基于元路径的局部表征编码器与局部图互信息计算模块,整体框架如图 1 所示。

首先,给定由一组元路径定义的邻接矩阵,局部表征编码器将分别在每个邻接矩阵中利用图卷积模块生成目标类型节点的表征。然后,通过语义级别的注意力机制整合各个邻接矩阵中生成的节点表征。之后,局部图互信息计算模块将利用生成的节点表征与采样到的、各个邻接矩阵中的局部子图,计算相应的局部图互信息。最终,以最大化互信息作为目标函数,实现对模型参数的训练,并得到优化后的节点表征。

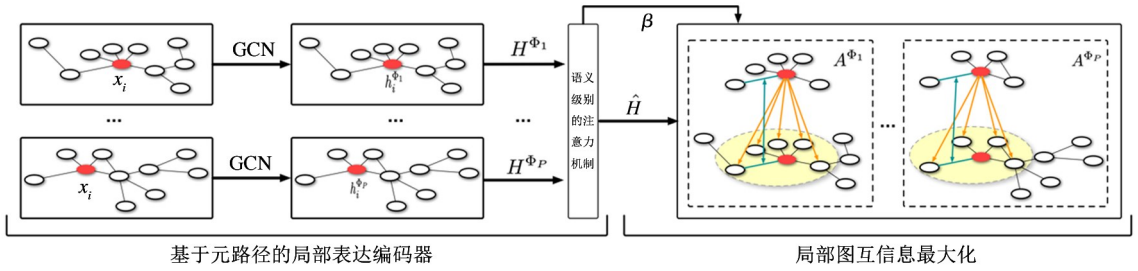

图 1 HGMI 的模型框架

3.2 基于元路径的节点局部表征

在邻接矩阵集合 $\boldsymbol{A}^{\Phi}$ 中,每一个邻接矩阵表示一个同构图,因此,使用一个节点级的编码器生成包含初始节点特征 $\boldsymbol{X}$ 和 $\boldsymbol{A}^{\Phi_i}$ 信息的节点表征:

$$\boldsymbol{H}^{\Phi_i} = f_{\Phi_i}(\boldsymbol{X}, \boldsymbol{A}^{\Phi_i}) \tag{5}$$

其中,$f_{\Phi_i}(\cdot)$ 表示节点级的编码器。为了能够获得更大的感受野,以获得更多参与运算的信息量,同时有效地整合节点熟悉特征与局部结构特征,选择图卷积网络(GCN)作为节点级编码器,来生成每个邻接矩阵中的节点表征:

$$\boldsymbol{H}^{\Phi_i} = ((\boldsymbol{D}^{\Phi_i})^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}^{\Phi_i}(\boldsymbol{D}^{\Phi_i})^{-\frac{1}{2}})\boldsymbol{X}\boldsymbol{W}^{\Phi_i} \tag{6}$$

其中,$\widetilde{\boldsymbol{A}}^{\Phi_i} = \boldsymbol{A}^{\Phi_i} + \boldsymbol{I}$,$\boldsymbol{D}^{\Phi_i}$ 表示 $\boldsymbol{A}^{\Phi_i}$ 的节点度的对角矩阵,而 $\boldsymbol{W}^{\Phi_i} \in \mathbb{R}^{d\times F}$ 则表示过滤参数矩阵。在为每个邻接矩阵进行卷积操作后,将获得一个节点表征集合 $\{\boldsymbol{H}^{\Phi_i}\}_{i=1}^{P}$,$\boldsymbol{H}^{\Phi_i}$ 表示目标节点 V_t 关于元路径 $\boldsymbol{\Phi}_i$ 的表征向量。考虑到传统的平均池化(mean pooling)或最大池化(max pooling),无法有效衡量不同元路径(语义关系)的重要程度,因此,通过语义级别的注意力机制对集合 $\{\boldsymbol{H}^{\Phi_i}\}_{i=1}^{P}$ 进行聚合,生成包含多种关系语义的节点表征。

基于特定邻接矩阵学习的节点表征仅包含特定的语义信息。为了获得包含多种关系语义的节点表征,一种直观且有效的解决方案是探索每个元路径应为最终节点表征贡献多少,然后将各自的贡献作为权重聚合各个独立的节点表征。这里通过添加一个语义注意力层 L_{att} 来学习相应权重/贡献:

$$\{\beta^{\Phi_1}, \beta^{\Phi_2}, \cdots, \beta^{\Phi_P}\} = L_{att}(\boldsymbol{H}^{\Phi_1}, \boldsymbol{H}^{\Phi_2}, \cdots, \boldsymbol{H}^{\Phi_P}) \tag{7}$$

具体通过式(8)~式(10)来计算元路径 $\boldsymbol{\Phi}_i$ 的重要性。

$$e^{\Phi_i} = \frac{1}{N}\sum_{n=1}^{N}\tanh(\boldsymbol{q}^{\mathrm{T}}[\boldsymbol{W}_{sem}\cdot\boldsymbol{h}_n^{\Phi_i} + b]) \tag{8}$$

其中,$\boldsymbol{W}_{sem}$ 表示线性变换参数矩阵,$\boldsymbol{q}$ 表示需要学习的注意力语义向量。然后,利用 softmax 函数对生成的集合 $\{e^{\Phi_i}\}_{i=1}^{P}$ 进行正则化,以获得元路径 $\boldsymbol{\Phi}_i$ 的重要性权重 β^{Φ_i}:

$$\beta^{\Phi_i} = \mathrm{softmax}(e^{\Phi_i}) = \frac{\exp(e^{\Phi_i})}{\sum_{j=1}^{P}\exp(e^{\Phi_j})} \tag{9}$$

最终,异构图节点表示 H 将通过节点表征集合 $\{\boldsymbol{H}^{\Phi_i}\}_{i=1}^{P}$ 的线性组合获得:

$$\boldsymbol{H} = \sum_{i=1}^{P}\beta^{\Phi_i}\cdot\boldsymbol{H}^{\Phi_i} \tag{10}$$

虽然本文的语义注意力层是受到 HAN[18] 的启发,但在模型优化上仍存在着差异。HAN 利用分类交叉熵作为损失函数,学习方向将由训练集中标签样本指导。由于对标签样本的依赖,HAN 容易受到训练集中标签分布的影响,使得模型优化方向向有利于部分占比大的标签的方向偏移,进而造成语义级别注意力权重的分配失衡,并最终影响节点表征的质量。

而在本文的方法中,模型学习的注意力权重是由二元交叉熵损失(binary cross-entropy loss)指导的,即指导模型判断给定节点是否属于指定的局部子图。因此,模型学习到的权重有助于衡量节点在不同分布下与其近邻节点的相似程度,即节点输入特征与其近邻节点的输入特征越相似,分配的权重越大。同时,由于不涉及分类标签,因此权重不会因已知标签而产生偏差。

元路径之间通常表现出不同程度的兼容性,换句话说,不同元路径间可能存在相似的节点分布,同样也可能存在极大差异的节点分布。例如在学术社交网络中,以论文作目标节点,论文涉及的领域作为标签。那么,“论文引用关系”与“论文共作关系”之间的兼容性要强于“论文引用关系”与“术语共用关系”之间的兼容性。这是因为,同一作者的论文更大概率上是关注同一个研究领域的,而相同术语可以被多个领域的论文共用。因此,在没有先验知识的指导下,注意力机制会使得模型更倾向于关注出现频率较高的语义所代表的元路径,从而忽略其他出现频率较低的语义所代表的元路径,即使得模型发生语义层面上的过拟合问题。针对该问题,本文提出了一种注意力平衡机制,用于防止语义层面过拟合的发生,详细内容将在下节进行描述。

3.3 局部图互信息最大化

考虑到语义级别注意力机制可能导致的语义过拟合问题,设计了一种注意力平衡机制,使得模型对所有元路径均保持一定的关注度,而不是仅关注一部分特定的元路径。具体通过将局部表征编码器中注意力模块生成的注意力权重以 $\{1-\beta^{\Phi_i}\}_{i=1}^{P}$ 的形

式加入到式(3)中,使得从不受关注的元路径获得的互信息可以对模型训练产生一定的影响。换句话说,注意力平衡机制可以在模型优化的过程中,根据生成的注意力权重 β^{Φ_i} 实时调整互信息损失所占比重,使得模型可以在一个较为全面的感受野中进行参数更新,直到收敛。

局部图互信息中主要计算的是节点表征与其近邻输入特征间的互信息。如果将 $\{1-\beta^{\Phi_i}\}_{i=1}^{P}$ 添加为 $I(\boldsymbol{h}_i;\boldsymbol{x}_j)$ 的权重,则会干扰注意力权重的选取,导致节点表征聚合过多的噪音信息,使得模型无法得到有效的收敛。相反,如果将 $\{1-\beta^{\Phi_i}\}_{i=1}^{P}$ 添加为 $I(w_{ij};a_{ij}^{\Phi_t})$ 的权重,一方面,可以使得节点表征保留不同元路径下的结构信息;另一方面,避免了节点表征在注意力权重的干预下聚合过多不必要的噪音信息。因此,式(3)可以变换为

$$\tilde{I}(\boldsymbol{h}_i;G_i^{\Phi_t})=\sum_j^{i_n^{\Phi_t}}w_{ij}I(\boldsymbol{h}_i;\boldsymbol{x}_j)+(1-\beta^{\Phi_i})I(w_{ij};a_{ij}^{\Phi_t}) \tag{11}$$

参考 MINE[29]的方法,直接最大化式(11)。需要注意的是,MINE 采用 Donsker-Varadhan[30]表示联合分布概率与边缘概率乘积之间的 KL 散度(Kullback-Leibler divergence)来估计互信息的下界。然而,当更多地关注最大化互信息而不是获得其特定值时,可以使用其他非 KL 替代方案,例如 Jensen-Shannon 互信息估计器(JSD)[31]和噪声对比估计器(infoNCE)[32],来代替 KL 散度。在本文中,参考 GMI 的实验结果[10],出于有效性和效率的考虑,采用 JSD 估计器来最大化式(11)。换句话说,可以通过训练一个判别器/双线性函数 D 来对采样的正负样本集合进行区分,即判断一个节点的表征是否属于给定的局部子图,以此来估计和最大化互信息。

具体地,利用式(12)来计算 $I(\boldsymbol{h}_i;\boldsymbol{x}_j)$。

$$I(\boldsymbol{h}_i;\boldsymbol{x}_j)=-sp(-D_w(\boldsymbol{h}_i,\boldsymbol{x}_j))-\mathbb{E}_{\mathbb{P}}[sp(D_w(\boldsymbol{h}_i,\boldsymbol{x}'_j))] \tag{12}$$

其中,$D_w:D\times D'$ 表示由一个参数为 $\boldsymbol{w}$ 的神经网络构成的判别器,$\boldsymbol{x}'_j$ 为从假设的经验概率分布 $\mathbb{P}$ 中采样的负样本,$sp(x)=\log(1+e^x)$ 表示 softplus 函数。考虑到不同元路径构成的邻接矩阵中节点的分布不同,使用同一判别器将不利于建模每个元路径的语义信息。因此,本文分别构建不同的判别器对不同邻接矩阵中节点与局部子图间的关系进行判断。给定一个邻接矩阵 $\boldsymbol{A}^{\Phi_t}$,节点 v_i 与其邻居节点的互信息 $I(\boldsymbol{h}_i;\boldsymbol{x}_j)$ 可以表示为

$$I(\boldsymbol{h}_i;\boldsymbol{x}_j)=-sp(-D_{w^{\Phi_t}}(\boldsymbol{h}_i,\boldsymbol{x}_j))-\mathbb{E}_{\mathbb{P}}[sp(D_{w^{\Phi_t}}(\boldsymbol{h}_i,\boldsymbol{x}'_j))]\quad j\in i_n^{\Phi_t} \tag{13}$$

为了有效捕获节点的结构特征,本文将邻接矩阵定义为无权重的邻接矩阵,然后利用交叉熵替代 JSD 估计器来最大化 $I(w_{ij};a_{ij}^{\Phi_t})$:

$$I(w_{ij};a_{ij}^{\Phi_t})=a_{ij}^{\Phi_t}\log w_{ij}+(1-a_{ij}^{\Phi_t})\log(1-w_{ij}) \tag{14}$$

综上所述,结合式(11)~式(14),可以得到最终的目标函数:

$$L=\sum_t^P\sum_i^{|V_t|}\sum_j^{i_n^{\Phi_t}}I(\boldsymbol{h}_i;\boldsymbol{x}_j)+(1-\beta^{\Phi_i})I(w_{ij};a_{ij}^{\Phi_t}) \tag{15}$$

其中,$I(\boldsymbol{h}_i;\boldsymbol{x}_j)$ 用以计算节点表征向量与近邻属性特征向量之间的互信息,通过最大化该互信息将促使节点表征捕获子图中属性特征的分布,进而在全局视角中,使得具有相似属性特征分布的节点生成相似的表征;而 $I(w_{ij};a_{ij}^{\Phi_t})$ 则计算 2 个节点间存在边链接的概率。通过最大化此概率,可以保证节点表征保留低价近似度(low-proximity),进而在局部视角中,使得相连节点间具有相似的表征。因此,通过对目标函数式(15)进行优化,既可以保证全局视角中具有相似属性特征的节点表征的相似性,又可以保留局部视角中结构的近似性。此外,通过注意力平衡机制权衡多个元路径下的损失,有利于节点表征捕获语义上下文信息。

4 实验与结果分析

4.1 数据集

分别在 DBLP 与 IMDB 2 种异构图数据集上评估本文提出的 HGMI 方法,相关统计数据如表 1 所示。

DBLP 数据集是一种研究论文集,其中每篇论文包含相应的发表会议、作者与关键词等信息。作者节点可划分为 4 个研究领域,即数据库、数据挖

掘、信息检索和机器学习。本文选择作者作为目标节点,并使用作者所属的研究领域作为标签。最初的特征则是根据作者的个人资料利用词袋模型生成的。

IMDB 数据集是关于电影的知识图数据,可以分为 3 种类型,即动作、喜剧和戏剧。本文选择电影作为目标节点,并使用电影的类型作为标签。电影的特征则由色彩、标题、语言、关键字、国家、评分、年份以及 TF-IDF 编码组成。

表 1 实验数据统计信息

数据集	节点	节点数目	边	边数目	特征维度	元路径
DBLP	作者(A)	4057	AP	19 645	334	APA
	论文(P)	14 328	PC	14 328		APCPA
	会议(C)	20	PT	88 420		APTPA
	术语(T)	8789				
IMDB	电影(M)	4275	MA	12 838	6344	MAM
	演员(A)	5431	MD	4280		MDM
	导演(D)	2082	MK	20 529		MKM
	关键词(K)	7313				

4.2 对比方法与相关设置

本文将对比方法分成两类,分别是无监督的图表示学习方法和有监督的图表示学习方法。

其中,无监督的图表示学习方法包括:(1)Raw Feature,即将初始的输入特征作为节点表征;(2)3 个异构图表示学习方法,即 Metapath2vec (M2V)、$HDGI_C$ 与$HDGI_A$,其中$HDGI_C$ 表示使用 GCN 作为特征生成模块的 HDGI,而$HDGI_A$ 则表示使用 GAT 作为特征生成模块的 HDGI;(3)2 个同构图表示学习方法 DGI[8]与 GMI[10]。

有监督的图表示学习方法包括 2 个异构图神经网络模型 RGCN[20]与 HAN 和 2 个同构图神经网络模型 GCN 和 GAT。

需要注意的是,对于专为同构图而设计的方法,即 DGI、GMI、GCN、GAT,不考虑图的异质性,而是构造基于元路径的邻接矩阵,报告其中最佳的结果。

本文提出的 HGMI 方法使用 Adam 优化器进行优化,并设定学习率为 0.01。同时设定节点表征的维度为 512,注意力表征的维度为 8。使用 Pytorch 来实现本文的模型,并在带有 2 个 GTX-1080ti GPU 的服务器中进行实验。

4.3 实验结果与分析

在节点分类任务中,本文为无监督学习方法训练逻辑回归分类器进行分类,而有监督方法则作为端到端模型直接输出分类结果。分别取数据集的 20% 和 80% 作为训练集进行实验。另外,选择 10% 的数据作为验证集,以及 10% 的数据作为测试集。为了保证结果的稳定性,将分类任务重复 10 次,计算平均的宏 F1 值(macro-F1)和微 F1 值(micro-F1)。

考虑到实验所用的数据集以及相关的评估方法、指标均与 HGDI[9]相同,因此,直接与文献[9]中的实验结果进行比较。实验结果如表 2 所示。

表 2 节点分类任务结果

输入数据			X	A	X, A, Y				X, A				
数据集	训练集	评价指标	Raw	M2V	GCN	RGCN	GAT	HAN	DGI	GMI	$HDGI_C$	$HDGI_A$	HGMI
DBLP	20%	Micro-F1	0.7552	0.6985	0.8192	0.1932	0.8244	0.8992	0.8975	0.9060	0.9157	0.9062	0.9305
		Macro-F1	0.7473	0.6874	0.8128	0.2132	0.8148	0.8923	0.8921	0.8984	0.9094	0.8988	0.9263
	80%	Micro-F1	0.8325	0.8211	0.8383	0.2175	0.8540	0.9100	0.9150	0.9170	0.9226	0.9192	0.9465
		Macro-F1	0.8152	0.8014	0.8308	0.2212	0.8476	0.9055	0.9052	0.9097	0.9153	0.9106	0.9421
IMDB	20%	Micro-F1	0.5112	0.3985	0.5931	0.4350	0.5985	0.6077	0.5728	0.5212	0.5893	0.5482	0.6416
		Macro-F1	0.5107	0.4012	0.5869	0.4468	0.5944	0.6027	0.5690	0.5222	0.5914	0.5522	0.6396
	80%	Micro-F1	0.5900	0.4203	0.6467	0.4476	0.6540	0.6600	0.6003	0.5480	0.6592	0.5861	0.6782
		Macro-F1	0.5884	0.4119	0.6457	0.4527	0.6550	0.6586	0.5950	0.5367	0.6646	0.5834	0.6781

从表 2 中不难看出,基于异构图的方法,即 HAN、HDGI 和 HGMI,通常要优于面向同构图的方法,即 GCN、GAT、DGI 和 GMI,这说明挖掘与保留异构图中的丰富语义信息有利于提高节点表征的质量。同时,对比以输入特征直接作为节点表征的实验结果,可以有效排除输入特征是导致模型获得较好性能的主要因素的可能性。同样,对比仅利用语义关系/网络结构进行表示学习的 M2V,有效结合输入特征与结构信息的异构图表示学习方法通常可以获得更好的节点表征。

与有监督的图神经网络方法的实验结果相比,基于互信息的无监督图神经网络方法同样可以获得较好的实验效果,甚至是表现得更好,如 HGMI 与 HDGI。这表明在缺少监督信息的场景下,基于互信息的无监督图神经网络方法会是很好的选择。该观察结果还表明,通过有监督的方式在图结构中学习到的特征可能存在局限性,即易受来自数据标签的分布或是下游任务表现出的偏好的影响。而这些局限性可能严重影响表示学习方法在真实场景中的应用。

此外,对比 HGMI 与 HDGI 的结果,可以发现本文方法的分类效果在 2 个数据集中均有提升,这充分反映了在同时考虑多个、具有不同结构的邻接矩阵时,引入局部图互信息以及注意力平衡机制的必要性。一方面,局部图互信息可以使模型更关注节点近邻的信息,而不是全图的信息,从而避免引入不必要的噪音信息;另一方面,注意力平衡机制可以使得模型对所有的元路径均保留一定的关注,而不是过度关注某个/部分元路径,从而使得节点表征获得来自其余语义关系的信息。

另外,在 IMDB 数据集中,GMI 的效果要差于 DGI 的效果。这主要是因为,在 IMDB 数据集中,目标节点间通过元路径构建的关联关系往往是弱相关的,例如,同一个导演可能指导不同类型的影片,而同一个演员也可能出演不同类型的影片。因此,这种弱相关性往往会引入较多的噪音,即具体不同输入特征的邻居节点。不同于 GMI,HGMI 通过注意力机制聚合不同邻接矩阵中近邻的信息,从而过滤噪音信息,保证节点表征的质量。

在节点聚类任务中,利用 K-mean 算法对生成的节点表征进行聚类。其中,聚类的簇数被设定为目标节点的类别种类的数目。在该任务中,仅比较无监督的方法,即 Raw、M2V、DGI、GMI、HDGI 与 HGMI。同样重复进行 10 次聚类任务,并在表 3 中展示平均的标准互信息(normalized mutual information,NMI)和调整兰德系数(ARI)。

表 3　节点聚类任务结果

数据集	DBLP		IMDB	
方法	NMI	ARI	NMI	ARI
Raw	0.1121	0.0698	0.0106	0.0117
M2V	0.3430	0.3754	0.0115	0.0151
DGI	0.5923	0.6185	0.0056	0.0260
GMI	0.6971	0.7423	0.0489	0.0391
$HDGI_C$	0.6076	0.6267	0.0187	0.0370
$HDGI_A$	0.5212	0.4986	0.0080	0.0129
HGMI	0.7749	0.8270	0.0774	0.0541

从表 3 中不难看出 HGMI 始终要优于其他的对比方法。结合节点分类任务的结果,发现对比方法均存在不同程度的过平滑问题,即局部结构中的节点表征变得过于相似。换句话说,相似的节点表征在一定程度上有利于分类器对节点进行分类;反之,在进行节点聚类时,相似的节点表征则会使得节点聚集在一起,从而变得无法区分。而通过综合考虑多个邻接矩阵下的近邻的分布情况,以及有选择地从中提取有用的信息,HGMI 可以有效地防止过平滑问题的发生。

为了进一步说明注意力平衡机制起到的作用,分别可视化 HGMI、HDGI 以及去除注意力平衡机制的$HGMI_{na}$在 IMDB 数据集上最终的注意力权重,结果如图 2 所示。

从图 2 中可以看出,HDGI 主要关注 MKM 关系;相反,$HGMI_{na}$则主要关注 MDM 与 MAM 关系。这不难理解,MKM 通过电影间相同的关键词构建关系,由于多数关键词的通用性,使得电影节点连接得更为紧密、表征变得更为相似,导致图读出操作(readout)生成的全局图表征与节点表征具有更大的互信息,从而使得MKM获得较大的关注。反之,

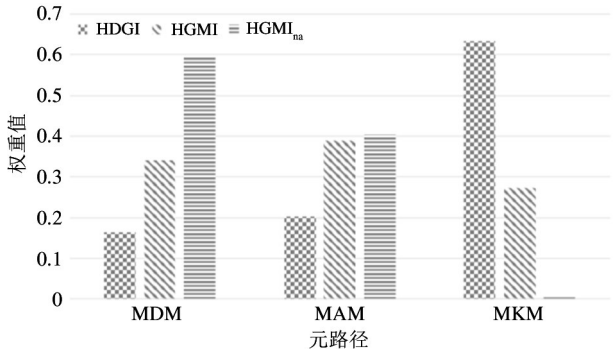

图 2 IMDB 数据集中不同元路径的注意力权重

MDM 与 MAM 往往会使得少量、具有相似属性/特征的节点聚集在一起,使得局部子图与节点表征间的互信息变大。

而在加入了注意力平衡机制后,HGMI 不仅可以对 MDM 与 MAM 保持较高的注意力权重,同时,也会为 MKM 分配一定的权重,而不是直接将其忽略。通过这种方式,HGMI 可以聚合到在 MDM 与 MAM 中接触不到的节点的特征。

5 结 论

本文主要讨论了利用图互信息进行无监督的异构图表示学习的方法。首先通过元路径将异构图转化为多个具有特定语义的同构图;然后在每个同构图中进行图卷积操作,并利用注意力机制对相同节点的不同表征进行融合;在此基础上,最大化每个图中局部子图与节点表征间的互信息,使得节点表征可以有效聚合不同语义关系下近邻的输入特征。同时,为防止语义过拟合的发生,引入注意力平衡机制,使得模型对所有语义关系均保持一定的关注度。实验结果表明,本文方法相比于其他方法,可以在节点分类与节点聚类任务中获得更好的效果。

参考文献

[1] Shi C, Li Y T, Zhang J W, et al. A survey of heterogeneous information network analysis[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017,29(1): 17-37

[2] Cui P, Wang X, Pei J, et al. A survey on network embedding[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2019,31(5):833-852

[3] 何昊晨,张丹红. 基于多维社交关系嵌入的深层图神经网络推荐方法[J]. 计算机应用, 2020,40(10): 2795-2803

[4] Tang J, Qu M, Mei Q, et al. PTE: predictive text embedding through large-scale heterogeneous text networks [C]// Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 2015: 1165-1174

[5] Chen T, Sun Y Z. Task-guided and path-augmented heterogeneous network embedding for author identification [C]// Proceedings of the 10th ACM International Conference on Web Search and Data Mining, Cambridge, UK, 2017: 295-304

[6] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, USA, 2019: 793-803

[7] Fu X, Zhang J, Meng Z, et al. MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding[C]// The Web Conference 2020, Taibei, China, 2020: 2331-2341

[8] Velickovic P, Fedus W, Hamilton W L, et al. Deep graph infomax[C]// Proceedings of the 7th International Conference on Learning Representations, New Orleans, USA, 2019: 1-7

[9] Ren Y, Liu B, Huang C, et al. Heterogeneous deep graph infomax[EB/OL]. http://arXiv.org/abs/1911.08538v2,(2020-02-04),[2020-10-10]

[10] Peng Z, Huang W, Luo M, et al. Graph representation learning via graphical mutual information maximization [C]// The Web Conference 2020, Taibei, China, 2020: 2341-2351

[11] 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, San Francisco, USA, 2016: 855-864

[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, Florence,

Italy, 2015: 1067-1077

[13] 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, New York, USA, 2014: 701-710

[14] Ou M, Cui P, Pei J, et al. Asymmetric transitivity preserving graph embedding[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016: 1105-1114

[15] Wang X, Cui P, Wang J, et al. Community preserving network embedding[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017: 203-209

[16] 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, Halifax, Canada, 2017: 135-144

[17] Fu T, Lee W C, Lei Z. HIN2vec: explore meta-paths in heterogeneous information networks for representation learning[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 2017: 1797-1806

[18] Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network[C]// The World Wide Web Conference, San Francisco, USA, 2019: 2022-2032

[19] Zhang C, Swami A, Chawla N V. SHNE: representation learning for semantic-associated heterogeneous networks [C]// Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 2019: 690-698

[20] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]// European Semantic Web Conference, Portorož, Slovenia, 2018: 593-607

[21] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: a survey of approaches and applications[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(12): 2724-2743

[22] Zhang J. Graph neural networks for small graph and giant network representation learning: an overview[EB/OL]. http://arXiv. org/abs/1908. 00187, (2019-08-01), [2010-10-10]

[23] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[EB/OL]. http://arXiv. org/abs/1609.02907,(2016-09-09),[2020-10-10]

[24] Velikovi P, Cucurull G, Casanova A, et al. Graph attention networks[EB/OL]. http://arXiv. org/abs/1710. 10903, (2017-10-30),[2020-10-10]

[25] You J, Ying R, Ren X, et al. GraphRNN: generating realistic graphs with deep auto-regressive models[EB/OL]. http://arXiv. org/abs/1802. 08773, (2018-02-24), [2020-10-10]

[26] Fey M, Eric Lenssen J, Weichert F, et al. SplineCNN: fast geometric deep learning with continuous b-spline kernels[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, USA, 2018: 869-877

[27] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs[C]// Advances in Neural Information Processing Systems, Long Beach, USA, 2017: 1025-1035

[28] Garcia Duran A, Niepert M. Learning graph representations with embedding propagation[J]. *Advances in Neural Information Processing Systems*, 2017,30: 5119-5130

[29] Belghazi M I, Baratin A, Rajeswar S, et al. Mine: mutual information neural estimation[EB/OL]. http://arXiv. org/abs/1801.04062, (2018-01-12),[2020-10-10]

[30] Donsker M D, Srinivasa Varadhan S R. Asymptotic evaluation of certain Markov process expectations for large time[J]. *Communications on Pure and Applied Mathematics*, 1983,36(2):183-212

[31] Nowozin S, Cseke B, Tomioka R. f-GAN: training generative neural samplers using variational divergence minimization[J]. *Advances in Neural Information Processing Systems*, 2016, 29: 271-279

[32] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[EB/OL]. http://arXiv. org/abs/1807.03748, (2018-07-10),[2020-10-10]

Local graphical mutual information maximization based heterogeneous graph neural network

Zhu Zhihua[*,**], Fan Xinxin[**], Bi Jingping[**], Wu Chao[***]

([*] University of Chinese Academy of Sciences, Beijing 100049)

([**] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([***] China Academic of Electronics and Information Technology, Beijing 100041)

Abstract

Aiming at the shortcomings of the injective ability of readout function and coarse-grained feature preservation in traditional mutual information maximization based heterogeneous graph neural networks (HGNN), which make them inadequate to use in the real-work networks, a new local graphical mutual information maximization based unsupervised heterogeneous graph neural network is presented. The model uses the meta-path to model the structure involving semantics in heterogeneous graphs and utilizes graph convolution module and semantic-level attention mechanism to capture individual node local representations. By maximizing the mutual information between the individual node embedding and the local graph, the proposed model effectively learns high-level node representations. The experimental results show that compared with HDGI which is based on the global graph mutual information maximization, the proposed method can increase the micro-F1 of the node classification task on DBLP/IMDB up to about 3%/9%, as well as the adjusted Rand index (ARI) of the node clustering task on DBLP/IMDB up to about 23%/46%.

Key words: heterogeneous graph (HG), graph neural network (GNN), mutual information, unsupervised method, graph representation learning