第 58 卷　第 3 期
2022 年 5 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 58, No. 3
May, 2022

DOI:10.13232/j.cnki.jnju.2022.03.003

基于图神经网络的社交网络影响力预测算法

陈轶洲[1],刘旭生[2],孙林檀[2],李文中[1*],方立兵[3],陆桑璐[1]

(1.南京大学计算机软件新技术国家重点实验室,南京,210023;2.国家电网有限公司客户服务中心,天津,300309;
3.南京大学工程管理学院,南京,210023)

摘　要:近十年来,通过社交网络(如微博、推特)分享信息已经成为人们日常生活中不可缺少的一个环节,如何有效地预测信息传播的影响力成为社交网络研究中的重要课题,不论是识别病毒式营销和虚假新闻还是精确推荐和在线广告都有许多应用.目前,一些应用深度学习进行社交网络影响力预测的方法已经取得了一定进展,但在进行深度学习时仍会面临以下难点:用户通常具有不同的行为和兴趣并且他们同时通过不同的渠道进行互动;用户之间的关系难以检测和形式化表达.传统的社交网络影响力预测方法通过设计复杂的规则来手动提取用户及其所处网络的特征信息,这一方法的有效性严重依赖于设置规则的专业性,所以很难将某一领域的规则推广到其他领域的应用中去.基于深度神经网络模型,设计一种端到端的神经网络来学习用户的隐藏特征信息以预测其社交网络影响力.首先通过图嵌入的方式对用户的局部网络进行特征提取,然后将特征向量作为输入对图神经网络进行训练,从而对用户的社会表征进行预测.该方法的创新之处:运用图卷积和图关注方法,将社交网络中用户的特征属性和其所处局域网络特征相结合,大大提高了模型预测的精度.通过在推特、微博、开放知识图谱等数据集上的大量实验,证明该方法在不同类型的网络中都有较好的表现.

关键词:图嵌入,图卷积,图注意力,社交网络,深度学习

中图分类号:TP391　　**文献标志码**:A

Social influence prediction with graph neural network

Chen Yizhou[1], Liu Xusheng[2], Sun Lintan[2], Li Wenzhong[1*], Fang Libing[3], Lu Sanglu[1]

(1. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China;
2. Customer Service Center of State Grid Corporation of China, Tianjin, 300309, China;
3. School of Management & Engineering, Nanjing University, Nanjing, 210023, China)

Abstract: For a decade, sharing information through social networks (e.g. Microblog and Twitter) has become an indispensable part in our daily life. Therefore, how to effectively predict the social influence has become an important subject in the study of social network. There are many applications, such as identifying viral marketing and fake news, accurate recommendations and online advertising. In recent years, some deep learning social network influence prediction methods have made some progress, but still faces the following difficulties: users typically have different behaviors and interests and they interact through different channels at the same time, and relationships between users are difficult to detect and formally describe. Traditional social network influence prediction methods manually extract the feature information of users and their networks by designing complex rules. However, the effectiveness of this method heavily depends on the domain knowledge of the rules set, which makes it difficult to generalize the rules in one field to the application in other fields. Based on deep learning method, we design an end-to-end neural network to learn the features of users′ hidden information to predict their influence in the social network. Firstly, feature extraction is carried out on users′ local network through graph embedding, and

基金项目:国家电网有限公司科技项目(5700-202153172A-0-0-00)

收稿日期:2022－03－08

*通讯联系人,E-mail:lwz@nju.edu.cn

then the graph neural network is trained with feature vector as input, so as to predict users' social representation. Compared with previous work, we combine the feature attributes of users in social network with the local area network features by using graph convolution and graph attention method, which greatly improves the accuracy of model prediction.

Key words: graph embedding, graph convolution, graph attention, social network, deep learning

随着在线社交媒体的发展,用户通过社交网络进行信息分享的趋势愈发明显,社交网络影响力指用户的行为与观念在多大程度上受所处网络中"邻居"的影响.众所周知,一个用户的个人喜好(节点特征属性)与所处的朋友圈(图结构属性)往往在短时间内不会发生变化,通过观察用户在社交网络中的一系列行为并学习参数可以帮助我们有效地预测信息传播的影响力,这对于解决在线广告推荐、识别病毒式营销等现实问题具有重要的实际应用价值.

近几年来,随着深度学习[1]的兴起与应用,如机器翻译、语音识别这类曾经需要手工提取特征信息的机器学习任务,如今依靠各种端到端的深度学习范式(例如卷积神经网络Convolutional Neural Network,CNN)、长短期记忆(Long Short-Term Memory,LSTM)网络、循环神经网络(Recurrent Neural Network,RNN)),都在各自领域取得了极大的进展.然而,这些深度学习方法的应用往往局限于提取欧式空间的特征属性,但现实生活中的大多数数据都是从非欧式空间生成的.和图像(image)这种规则的数据结构相比,图(graph)更复杂,因为其节点往往是无序的,人们也无法规定某一节点与图中其他节点的关系,导致诸如卷积、池化之类的操作不能直接运用在图上.另一方面,传统深度学习方法的一个核心假设是各个数据样本之间保持独立,而对于图来说这一假设无法成立,因为图中的每个节点都可能与其他节点存在关系,忽略节点间的相互影响会导致模型的失真.

在社交网络影响力研究领域,传统的深度学习方法已被证明难以解决图结构中的节点无序性与样本相关性问题.这些方法要么严重依赖于基础的扩散模型,要么使用了复杂的手工特征,而这些特征难以推广到不同类型的级联中,即使最新的生成方法允许理解传播机制,但是预测准确性仍然不能令人满意,需要一种能够处理图结构数据的深度学习方法.

图神经网络(Graph Neural Network,GNN)[2]是一种新近被提出的深度学习方法,在处理图结构数据(如社交网络数据、化学图结构和引文网络)方面已经取得了一些进展.例如,通过表征学习的方式将非欧式空间数据转化为欧式空间表示并保留其中的空间信息;借鉴RNN模型,假设一个节点不停地与其邻居节点交换信息直到达到动态平衡;使用图信号处理技术定义图卷积来提取特征.利用卷积操作在非欧式空间数据上提取特征是一个值得深入研究的课题.

本文将深度学习方法应用到社交网络影响力预测,从提取用户特征与图结构特征的角度出发学习社交网络模型的隐藏参数,从而更全面地了解用户偏好与信息传播模式,可以有效预测某一事物在社交网络中的影响力大小,为目标客户提供更人性化的服务.然而,图数据在序列性与独立性上的问题导致深度学习方法难以取得理想的效果,为了解决这一问题,引入图卷积(Graph Convolutional Neural Network,GCN)与图注意力(Graph Attention,GAT)两种方法,模仿CNN中的卷积操作与RNN中的信息交换,既能兼顾用户的节点特征与所处子图的结构特征,又能有效处理有向图与无向图两种图结构,大大提升了社交网络影响力的预测精度.

实验在大规模真实世界数据集上进行,包含社交网络(微博、推特)与引用网络(Open Academic Graph,OAG)数据,但其中存在样本分布不平衡与正反例分布不平衡的问题.本文通过选择关联度高的节点与筛选减小正反例比重来解决上述问题,也提高了社交网络影响力预测的精度.

综上,本研究不仅展示了图神经网络在社交网络影响分析这一领域的光明前景,也对探索社交影响的动态变化过程打下了初步的基础.

1 相关理论

为了研究信息在社交网络中的传播,介绍一种经典的影响力传播模型:独立级联(Independent Cascade,IC)模型[3].

IC模型是一种概率模型,其思想是使处于激活状态的节点通过一个系统变量——成功概率,尝试激活其邻居节点,若失败则该影响被抛弃.IC模型的影响力传播过程:

(1)给定初始的活跃节点集合S,当节点u在时刻t被激活后,它尝试激活它的出边邻居v,成功概率为$p_{u,v}$($p_{u,v}$是系统变量,不受其他节点影响),若失败则放弃对v的影响.

(2)若节点v有多个入边邻居都是最近被激活的节点,那么这些邻居将会按照随机顺序依次激活节点v(所有的尝试都可视为发生在时刻t).

(3)若在时刻t节点v被成功影响,则在时刻$t+1$时对节点v重复上述过程.

Hootsuite[4]提供了一种利用嵌入式IC模型预测社交网络信息扩散的方法,通过研究社交网络中的网络簇结构将社交网络用户投射到一个隐藏特征空间中,再计算这些用户之间的欧氏距离来确定一个活跃节点激活其他节点的成功概率.借鉴这种经典影响力传播模型的思想,本文从深度学习的角度出发,提出了一种更加适合标签分类任务的社交网络影响力预测模型.

1.1 图表征学习 常见的机器学习模型,如神经网络,其输入都是序列化或规则的数据(例如图像、文本),但是难以处理非欧氏空间数据(例如图).为了解决这个问题,主流的处理方法有:使用基于矩阵分解的方法,在保留数据空间特性的基础上将高维数据投影到低维(例如用奇异值分解(Singular Value Decomposition,SVD)或主成分分析(Principal Component Analysis,PCA)来分解连接矩阵);使用手工构造特征的方法(如将page rank,degree等值拼接起来作为特征向量).但是上述两种方法存在明显不足:第一种方法的时间复杂度至少是二次的并且可扩展性差;第二种方法属于特征工程,需要数据处理者掌握这一领域的专业知识,也难以推广到其他级联中.结合现有工作的不足,希望有一项技术能自动学习图的特征并用于之后的任务.

先介绍自然语言处理的一项技术word2vec[5].word2vec由谷歌提出,其方法是通过大量的文本训练来获得一个单词在向量空间中的向量表示,而不是使用one-hot方法对单词进行编码.根据表征学习的理论,直接用one-hot编码训练,虽然模型会学得单词的表征,但由于训练数据集规模的限制,其表现不如word2vec.

图表征学习的思想与word2vec类似,都是捕捉原始数据信息(如图的结构信息、文本中的词素信息)并编码成向量.DeepWalk[6]是一种有效的被应用在图表征学习中的方法,在图上随机游走(Random Walk)获得多条节点序列,将每个节点看成一个词素,将节点序列看成一个语句,这样就可以使用类似word2vec的技术为这些节点生成表征向量.具体方法:对于初始节点v_i,每次无偏差地随机选择一个邻居节点作为下一个节点,不断重复上述过程直到序列W_{v_i}的长度达到t.

1.2 社交网络影响力预测模型 典型的传统社交网络影响力预测有线性预测器逻辑回归(Logistic Regression,LR)与支持向量机(Support Vector Machine,SVM)[7].首先人为设计提取图中每个节点的特征的方法,再根据图结构手动提取每个节点的特征,将复杂的非欧氏结构转化为特征向量.这一方法的优势是训练速度快,但由于其性能严重依赖于被挑选的特征,故难以推广到其他领域.同时,这两种方法都是线性分类器,预测结果不能令人满意.

Perozzi et al[8]通过设计图卷积核来运用CNN的成熟理论,即:将图结构的数据转化为易于输入神经网络的欧氏结构数据,然后用经典的深度学习框架来训练.具体流程分四步:第一步,使用Label Procedure函数对图中的节点进行降序排序,选择前w个节点作为中心点;第二步,对w个中心点分别寻找其一阶邻居并选取前k个作为该中心点卷积操作时的邻居(若一阶邻居数量不足k个则加入二阶邻居),这样就有了w个团;第三步,将w个团正则化,包括对邻居进行排序;最后,将上述w个团包含点和边的特征进行拼接,

目的是拼接成规整的特征形式以便后续使用CNN方法进行处理.

Bakshy et al[9]从基础的信息传播模型独立级联模型(IC)出发,通过构建RNN来预测每个节点最终是否会被激活,其优势在于可以利用节点自身的特征信息对参数进行训练,其缺点是需要迭代的次数较多,计算量大.

上述工作都尝试将图结构数据转化为更易于表达的欧氏空间形式,方便后续处理,然而,它们在提取特征时要么设计了复杂的手动提取规则,要么无法兼顾节点所处子图的结构特征与自身属性特征[10]. 因此,在处理社交网络这类非欧氏数据结构时,如何自动提取节点特征并最大程度保留节点所处子图的结构特征以及提取特征后采用何种方法训练,都是值得思考的问题.

2　基于图神经网络的影响力预测模型

本节从问题的形式化定义出发,首先介绍所用技术的理论基础,然后介绍模型的搭建过程,最后分析模型中参数的设置.

2.1　基本定义

定义1　自我中心网络[11]　给定一个有向图$G=(V,E)$,其中,V是节点集,E是边集. 对于一个用户v,其r-近邻的定义:

$$\Gamma_v^r=\{u:d(u,v)\leqslant r\}$$

其中,$d(u,v)$表示节点u与节点v之间的最短路径的长度. 用户v的r-自我中心网络被定义为由顶点集Γ_v^r在图G中推导得到的子图G_v^r.

定义2　社交行为[12]　社交网络中的用户往往会做出相应的社交行为(例如在微博中转发一条动态,或在开放知识图谱中引用一篇文章). 在某一时刻t观察一个处于社交网络中的用户v,为了判断其社交行为状态引入变量s_v^t,$s_v^t=1$表示在t时刻或者之前的某个时刻,该用户已经做了相应的社交行为,$s_v^t=0$表示用户在t时刻尚未做出相应的社交行为.

根据定义,需要解决的问题:指定一个用户v,在已知其r-自我中心网络结构与r-邻居的行为状态后,给出其做出相应社交行为的概率. 更规范地说,给定G_v^r和$S_v^t=\{s_u^t:u\in\Gamma_v^r\backslash\{v\}\}$,计算在一定时间间隔$\Delta t$后该用户做出相应社交行为的概率:

$$P\left(s_v^{t+\Delta t}\middle|G_v^r,S_v^t\right)\tag{1}$$

实际应用中,假设有N个实例,每个实例用一个三元组(v,a,t)表示,其中,v表示用户,a表示相应的社交行为,t表示时间戳,并且,对于每个实例用户v,掌握其r-自我中心网络G_v^r,r-邻居的社交行为状态集S_v^t以及一段时间间隔Δt后其社交行为状态$s_v^{t+\Delta t}$. 这样,社交网络影响力预测问题就转化为二元图分类问题,解决方法是最小化负对数似然函数[13]:

$$\mathcal{L}(\theta)=-\lg\left(P_\theta\left(s_{v_i}^{t+\Delta t}\middle|\bar{G}_{v_i}^r,\bar{S}_{v_i}^t\right)\right)\tag{2}$$

不失一般性,本文假设时间间隔Δt充分长,这样只需预测当信息完成在网络中传播后用户v的最终行为状态.

2.2　图神经网络理论基础

CNN[14]是一种经典的深度学习框架,其核心特点是局部连接(local connection)、权重共享(shared weights)和多层叠加(multi-layer). 在图问题中这些特性仍然适用,因为图结构是典型的局部连接结构,并且共享权重可以减少计算量,另外,多层叠加是处理分级模式(Hierarchical Patterns)的关键. 然而,CNN只能运用于欧氏空间数据(例如二维图像),对于一般的图结构,CNN中的卷积核池化操作很难迁移到图上,如图1[15]所示.

为了在图结构数据当中使用卷积方法,Defferrard et al[16]将卷积操作定义为傅里叶频域计算图拉普拉斯(Graph Laplacian)的特征值分解. 这个操作可以定义为使用卷积核$g_\theta=diag(\theta)$对输入$x\in R^N$的卷积操作:

$$g_\theta\times x=Ug_\theta(\Lambda)U^T\tag{3}$$

$$L=I_N-D^{-1/2}AD^{-1/2}=U\Lambda U^T\tag{4}$$

其中,U是来自于标准化图拉普拉斯矩阵的特征向量矩阵,D是度矩阵,A是图的邻接矩阵,Λ是以特征值为对角线上值的对角矩阵.

Su et al[17]提出图注意力网络(GAT)模型,将注意力机制引入信息传播步骤,通过对节点的邻居节点增加注意力来计算节点的隐藏状态,还定义了一个图注意力层(Graph Attentional Layer)并通过多层叠加来构建任意的图注意力网络. 该

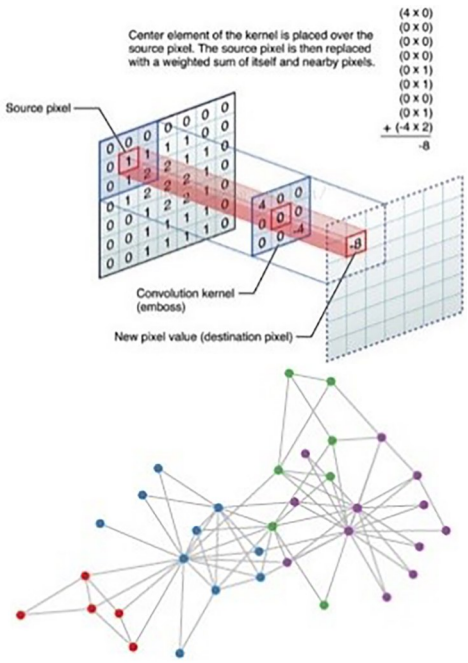

图 1 欧氏空间数据(上)与非欧氏空间数据(下)[15]

Fig. 1 Euclidean space (up) and non-Euclidean space (down)[15]

层的作用是计算节点对(i,j)的注意力系数(Attention Coefficients)[18],其计算过程如图 2 所示,计算式如下:

$$e_{ij}=a\left(\left[Wh_i\|Wh_j\right]\right),k\in N_i \tag{5}$$

$$\alpha_{ij}=\frac{\exp\left(LeakyReLU\left(e_{ij}\right)\right)}{\sum_{k\in N_i}\exp\left(LeakyReLU\left(e_{ik}\right)\right)} \tag{6}$$

其中,e_{ij}是节点对之间的相似系数[19],α_{ij}是节点j对i的注意力系数,N_i是图中节点i的邻居节点集合. 节点特征的输入集合是$h=\left\{h_1,h_2,h_3,\cdots,h_N\right\}$($h_i\in R^F$),其中,$N$是节点的个数,$F$是每个节点的特征维度. 该层会产生一个新的节点特征集(特征维度为F')$h'=\left\{h'_1,h'_2,h'_3,\cdots,h'_N\right\}$. $W\in R^{F'\times F}$是共享的线性变换权重矩阵,其作用是对顶点特征进行增维. $[\|\ \|]$对于顶点i,j变换后的特征进行拼接[20]. $a\in R^{2F'}$是单层的前向神经网络的权重向量,其作用是将拼接后的高维特征映射到一个实数上. 通过*softmax*函数对相似系数进行归一化,然后使用*LeakyReLU*非线性函数就能得到注意力系数. 最后,获得每个节点的输出特征(σ是激活函数):

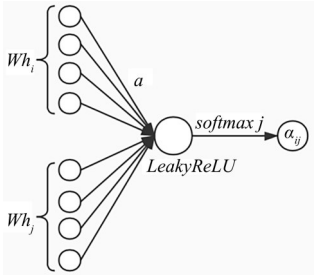

图 2 图注意力层

Fig. 2 Graph attention layer

$$h'_i=\sigma\left(\sum_{j\in N_i}\alpha_{ij}Wh_j\right) \tag{7}$$

2.3 模型构建 通过对问题的定义和基本方法的介绍,结合已有的工作,构建如图 3 所示的神经网络结构来预测社交网络影响力. 下面的章节会重点介绍此网络结构,包括每一层用到的技术以及选择这样架构的理由.

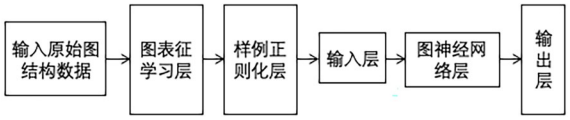

图 3 社交影响力预测模型

Fig. 3 Social influence prediction model

从问题定义出发,面对原始数据集,需要找出每个用户v的r-自我中心网络G_v^{r}[21]. 一种直接的方法是从用户v这个节点开始执行广度优先搜索,直到发现局部网络中所有距离不大于r的节点. 然而,不同的用户,其自我中心网络的大小往往不同,同时,现实生活中,由于社交网络的小圈子属性,自我中心网络的规模可能会非常大. 将这些大小不一的图结构数据进行序列化再进行特征提取非常困难,为了解决这一问题,必须对每个用户的自我中心网络进行采样,使提取的子网络具有固定的规模.

Graves et al[22]提出带重启动的随机游走(Random Walk with Restart,RWR)采样方法. 首先,随机选择中心用户或是其一个活跃邻居作为起始节点开始随机游走,每次循环中,一个节点以正相关于边权重的概率随机访问邻居,同时,每次游走开始前有一定概率回到起始节点. 算法执行完成后即可知访问的固定数量的顶点,由这些顶

点组成的集合记为$\overline{\Gamma_v^r}\left(\left|\overline{\Gamma_v^r}\right|=n\right)$. 用$\overline{\Gamma_v^r}$推导出子图$\overline{G_v^r}$来替代原问题中的子图$G_v^r$[23]. 相应地,定义新的状态集合$\overline{S_v^t}=\left\{s_u^t: u\in\overline{\Gamma_v^r}\backslash\{v\}\right\}$. 这样一来,优化目标就变为最小化负对数似然损失:

$$L(\theta)=-\sum_{i=1}^{N}\lg\left(P_\theta\left(s_{v_i}^{t_i}\middle|\overline{G_{v_i}^r},\overline{S_{v_i}^{t_i}}\right)\right) \tag{8}$$

随着表征学习[24]的兴起,利用网络嵌入技术将图结构信息降维到低维隐藏空间已经成为研究热点. 具体地,通过表征学习方法得到一个嵌入矩阵$X\in R^{D\times|V|}$,该矩阵的每一列都对应原网络中的一个节点.

受Dong et al[25]的启发,采用Inf2vec方法进行图嵌入学习,并用已经训练好的嵌入层将每个用户映射到其D维特征向量$x_u\in R^D$.

紧跟在嵌入层后,需要对数据进行正则化处理,这也是处理图像类型转换时的常用手段. 对于每个用户$u\in\overline{\Gamma_v^r}$,在获得其表征向量$x_u$后进行如下计算以求得新的表征向量$y_u$:

$$\begin{aligned}&y_{ud}=\frac{x_{ud}-\mu_d}{\sqrt{\sigma_d^2+\epsilon}},d=1,\cdots,D\\&\mu_d=\frac{1}{n}\sum_{u\in\Gamma_v^r}x_{ud}\\&\sigma_d^2=\frac{1}{n}\sum_{u\in\Gamma_v^r}\left(x_{ud}-\mu_d\right)^2\end{aligned} \tag{9}$$

其中,μ_d和σ_d分别是平均值与标准差,ϵ是用来维护数值稳定性的定值. 使用正则化表示可以消除实例指定的平均值与标准差,使得下游模型更关注用户们在隐藏表征空间中的相对位置而不是绝对位置. 同时,使用正则化的数据可以在模型训练时避免过拟合问题(overfitting)[26]的出现.

图卷积神经网络的原理已在上一节进行了介绍. 本实验中每个GCN层都会接受一个节点特征矩阵$H\in R^{n\times F}$,其中,n表示节点的数量,F表示节点的特征数量,H的每一行h_i^T都代表一个顶点的特征向量. GCN层的输出为$H'\in R^{n\times F'}$,其计算过程如下:

$$H'=GCN(H)=g\left(A_{\mathrm{GCN}}(G)\,HW^T+b\right) \tag{10}$$

其中,$W\in R^{F\times F'}$,$b\in R^{F'}$为模型参数,g为非线性激活函数,$A_{\mathrm{GCN}}(G)$为标准化图拉普拉斯矩阵.

令$A_{\mathrm{GAT}}(G)=\left[\alpha_{ij}\right]_{n\times n}$并用其替换式(10)中的$A_{\mathrm{GCN}}(G)$即可完成单头图注意力模型的定义.

输出层为每个用户输出的一个二维特征,这个特征就是对用户未来社交行为状态的预测. 通过对比计算对数似然误差[27],将误差反向传播就能达到优化目的.

3　实验评估

3.1　实验数据集　选取三个来自不同领域的大规模数据集进行实验来定量评估模型的性能[28].

(1)新浪微博:该数据集是由1776950位微博用户在2012年9月28日到2012年10月29日之间转发一条特定微博形成的社交网络,将这个社交网络中的社交行为定义为一个用户转发了该条特定的微博.

(2)推特:该数据集的形成与微博数据集类似,通过观察一篇在2012年7月4日发布的关于希格斯玻色子的文章在推特上的传播所形成. 将这个社交网络中的社交行为定义为一个用户转发了包含有“希格斯玻色子”的推特.

(3)OAG[19]:将两个大型学术图谱(微软学术图谱、AMiner图谱)进行结合,获得本实验所需的OAG(开放学术图谱)数据集. 受Niepert et al[19]的启发,从数据挖掘、机器学习、信息检索、自然语言处理、机器视觉等领域挑选了20个知名学术会议来组成一个合作者网络. 为了研究一个研究者的引用行为如何受到其合作者的影响,将这个社交网络中的社交行为定义为一个研究者从上述会议中引用了一篇论文.

为了更好地进行各个社交网络中的社交影响力预测,利用已有的方法[8]来处理上述三个数据集:首先,如果一个用户v在某一时刻t表现了社交行为a,就将这个用户标记为正例;接下来观察用户v的每个邻居,如果在观察的时间窗口内该邻居没有表现出社交行为a,就将这个邻居标记为反例. 这样,可以将关注点从社交网络中的所有顶点缩小到那些被观测到的实例用户上. 处理后的数据集特征如表1所示.

本文的实验目的:通过构建并训练神经网络来区分图中的正例与反例,即预测哪些用户在时间窗口内会做出特定的社交行为,而哪些用户不

表1 数据集介绍

Table 1 Description of datasets

	微博	推特	OAG
$\|V\|$	1776950	456626	953675
$\|E\|$	308489739	12508413	4151463
N	779164	449160	499848

会.然而,已有的数据集仍然面临数据不平衡的问题,主要表现在两个方面:

第一,正例与反例的分布不平衡.例如,在微博数据集中,观测到的反例与正例的比例高达300:1,这严重影响了训练的精度.为了解决这一问题,需要重新筛选被观测用户,最终使其中的反例与正例的比例维持在3:1左右.

第二,用户周围活跃的邻居数量不平衡.若一个用户拥有相对多的活跃邻居时,其自我中心网络中的结构特征会与其社交网络影响力产生紧密联系,然而,在大部分社交网络数据集中活跃邻居的数量分布都是不平衡的.以微博为例,78%的实例用户只拥有一个活跃的邻居,而拥有不少于三个活跃邻居的实例用户只占总数的8.57%.如果在这样不平衡的数据集中训练模型,最终的结果往往会偏向那些只拥有一个活跃邻居的实例用户.为了保证模型的鲁棒性和更好地提取用户所处局部网络的结构信息,选择剔除只有一个活跃邻居的实例用户,只考虑拥有两个及以上活跃邻居的用户.

3.2 模型训练 实验运行在一个搭载1.6 GHz双核Intel Core i5处理器和8 GB 1600 MHz DDR3内存的电脑上,操作系统为Linux16.04,使用PyTorch框架.

在数据预处理阶段,将RWR算法中回到起始点的概率设置为80%,将从自我中心网络中采样得到的节点数量固定为50.

在图嵌入层,使用Inf2vec算法提前训练好参数,并将图中每个节点映射到一个64维的网络表征向量中.

本文模型使用三层GCN/GAT的结构.无论使用的是GCN还是GAT结构,前两层每层都包含128个隐藏元,最后一层为输出层,其包含两个隐藏元,方便预测用户的最终行为状态.特别地,在GAT层中加入Multi-Head Graph Attention机制时[28],将前两层中的Attention heads个数设置为8(K=8)并且每个head负责计算16个隐藏元,这样前两层中每层总的隐藏元个数为$8\times16=128$个.

激活函数使用指数线性单元(Exponential Linear Units,ELU)[29]作为非线性激活函数,这是因为ELU具有较好的噪声鲁棒性,同时能够使神经元的平均激活均值趋近0.

开始模型训练前,使用Glorot Initialization方法对神经网络中的参数进行初始化.参数训练过程中使用Adagrad优化器,其中学习率(Learning Rate)设为0.1,权值衰减(Weight Decay)设为$5e^{-4}$,丢失率(Dropout Rate)为0.2.

为了验证结果的精度,将75%的实例作为训练集,12.5%的实例作为验证集,剩余12.5%的实例作为测试集.所有数据集上的batch size均为1024,并且在训练集上至多进行500个epoch.

3.3 结果对比 采用四种度量标准来衡量模型性能:曲线下面积(Area Under Curve,AUC)、准确度(*Precision*)、召回率(*Recall*)、$F1$测试值($F1$-measure,$F1$).

使用以下两种方法作为基准(baselines)方法并与本文搭建的模型进行比较:

(1)逻辑回归(Logistic Regression,LR)[30]是训练分类问题的经典方法.为了提取图中节点的信息,采用手动设置的方法进行提取.所需提取的特征分三类:每个用户节点自身的特征信息;提前训练好的用户的图表征向量;局部网络特征.具体内容如表2所示.

(2)Borgwardt and Kriegle[13]提出将CNN应用在图(Graph)上的一种新思路(PSCN),其具体思想:对于每张图,PSCN方法依据节点的度与中介中心度等信息将节点按降序排序并从中挑选前w个节点[31];对于每个节点,使用广度优先搜索找到它的前k个近邻并在这些节点的基础上生成邻近图,构造一个具有F(F为每个节点的特征维度)个信道的长度为$w\times k$的节点序列;最后,将这个向量输入CNN进行训练.

比较四种方法在三个数据集上的表现,实验结果如表3所示;计算AUC指标表现最好的方法

表2　采用手动提取方式得到的网络节点特征与局部图特征

Table 2　Network node features and local graph features using manual extraction

类别	描述
节点特征	核数(Coreness)
	网页排名(Pagerank)
	权威度(Authority score)
	特征向量中心(Eigenvector Centrality)
	聚类系数(Clustering Coefficient)
	稀有度(Rarity)
图表征[28]	使用Inf2vec算法提取出的64维图表征向量
子图特征	节点的活跃邻居数量
	由活跃邻居诱导出的子网密度
	由活跃邻居组成的连通分量

与表现第二的方法之间的相对增益,结果如表4所示.表中黑体字为结果最优.比较使用手动设置的特征和使用图表征对图神经网络性能的影响,实验结果如表5所示.

表3　四种方法在三个数据集上的表现

Table 3　Experimental results of four methods on three datasets

数据集	模型	*AUC*	*Precision*	*Recall*	*F*1
Weibo	LR	76.09%	41.33%	71.87%	52.55%
	PSCN	80.30%	46.71%	70.52%	56.23%
	GCN	75.84%	41.43%	70.29%	52.20%
	GAT	**81.71%**	**47.52%**	**75.08%**	**58.26%**
Twitter	LR	77.06%	44.85%	**68.80%**	54.35%
	PSCN	77.73%	46.35%	66.28%	54.58%
	GCN	75.59%	43.30%	65.73%	52.25%
	GAT	**79.21%**	**47.40%**	68.07%	**55.92%**
OAG	LR	64.54%	31.25%	68.96%	43.15%
	PSCN	68.15%	35.44%	63.63%	45.60%
	GCN	62.54%	29.27%	**73.35%**	42.02%
	GAT	**70.78%**	**39.76%**	59.96%	**47.85%**

如图4所示,无论是*AUC*还是*F*1,GAT方法都取得了远超其他三种方法的表现,这也证明了本文方法的有效性.同时,如表4所示,在OAG数据集中GAT方法的相对增益最高,为3.3%,证明其能有效揭示OAG数据集中的隐藏机制,预测局部社交网络影响力的动态变化.

表4　在*AUC*指标下GAT方法较其它三种方法的相对增益

Table 4　Relative gain of GAT method compared with other three methods under the metric of *AUC*

模型	微博	推特	OAG
LR	76.09%	77.06%	64.54%
PSCN	80.30%	77.73%	68.15%
GCN	75.84%	75.59%	62.54%
GAT	**81.71%**	**79.21%**	**70.78%**
相对增益	1.2%	1.1%	3.3%

表5　将手动提取的特征作为输入与不使用手动提取特征作为输入在GAT模型上的实验结果对比

Table 5　Performance of GAT method with and without manually extracted features as input

数据集	手动提取特征	*AUC*	*Precision*	*Recall*	*F*1
Weibo	是	81.71%	47.52%	75.08%	58.26%
	否	80.46%	45.89%	74.01%	56.70%
Twitter	是	79.21%	47.40%	68.07%	55.92%
	否	77.29%	46.23%	64.35%	53.83%
OAG	是	70.78%	39.76%	59.96%	47.85%
	否	67.06%	33.76	65.86%	44.77%

PSCN框架根据预先定义好的排序函数将节点按降序排列,并从中选取一定数量的节点组成规范序列.本文实验使用基于广度优先搜索的排序函数,目的是让每一个中心用户更加关注其附近的活跃邻居.PSCN的表现优于LR和GCN方法,但距离GAT的精度还有差距.

前人的工作[32]已经证明了GCN在处理标签分类任务上的优越性,但出人意料的是GCN是四种方法中表现最差的,不仅精确度明显劣于同为图神经网络的PSCN和GAT,甚至还不如线性预测器LR,这是因为GCN的使用建立在其同质性假设(Homophily Assumption)之上.同质性假设是指和不相似的节点相比,相似的节点有更大的概率是相互关联的.在这一假设的基础上,GCN的每个节点对它的邻居都"一视同仁",在计算中心节点的隐藏表征时,对其所有邻居的隐藏表征采取不带权重的平均相加方法.然而,本实验中同质性假设不成立,用户的自我中心网络中活跃邻居的重要性远远超过不活跃邻居,所以必须在

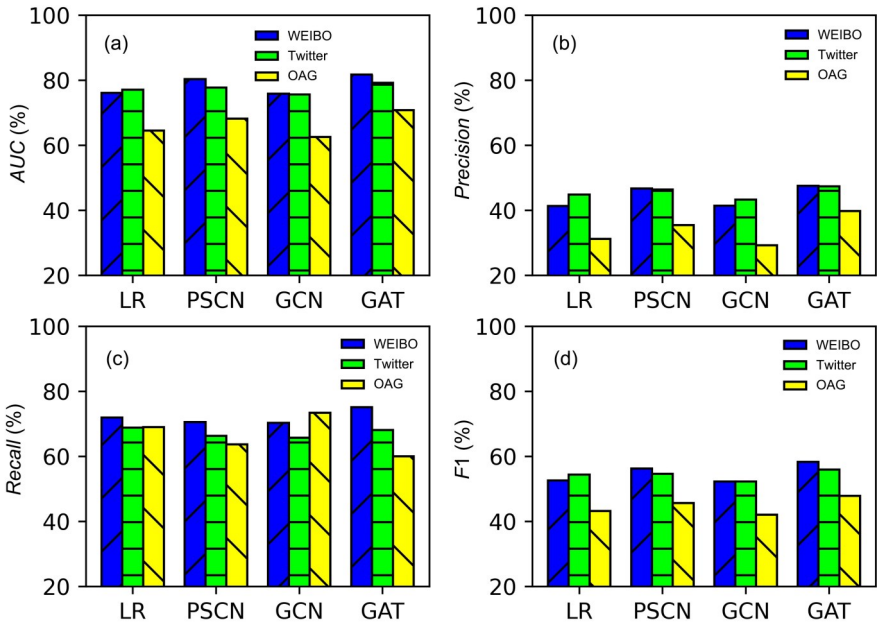

图 4　四种方法在下列指标下的实验结果对比

Fig. 4　Experimental results of the four methods compared under the following metrics

计算隐藏表征时,对活跃邻居和不活跃邻居赋予不同的权重. 否则,GCN会将有效预测信息与网络中的噪声相混合.

计算表5中的数据时,为了控制实验单一变量,统一使用表1中的手动提取特征来训练LR,PSCN,GCN和GAT. 然而,本文的重要目标是不使用手动提取特征,而是构建一个端到端的学习框架来预测社交网络影响力,为此还要研究仅使用图嵌入技术和使用手动提取特征之间的差异. 选择四种方法中表现最好的GAT,分别采用图嵌入和手动提取特征方法训练网络进行对照研究. 如表5所示,由于手动提取的特征中也包含用图嵌入方法学习的特征,所以使用手动提取特征的结果优于仅使用图嵌入学习特征的结果. 这一结果在可接受范围内,因为即使不使用手动提取特征,GAT的性能仍然优于其他三种方法. 在实际应用中,针对某一领域设计的手动提取特征方法往往很难推广到其他领域,即使它在原来的领域是有效的. 仅使用图嵌入来提取特征,是以牺牲一部分性能为代价而使方法更具有普适性和易于操作性,值得继续研究和推广.

3.4　参数敏感度分析　研究图采样阶段和神经网络构建中超参数的设置对实验结果的影响. 具体研究的超参数:RWR算法中的回归概率、对用户中心网络的采样规模、数据集中反例与正例的比、multi-head attention方法中heads的数量.

RWR算法对节点的近邻进行采样,通过设置回归概率可以控制用户自我中心网络的"范围". 图5显示了将回归概率设置为10%到90%时AUC与$F1$的变化. 由图可见,随着回归概率的提高,预测表现也有小幅度的增长,表明通过"缩小"用户自我中心的采样范围可以更好地获取社交网络中隐藏的影响力传播模式.

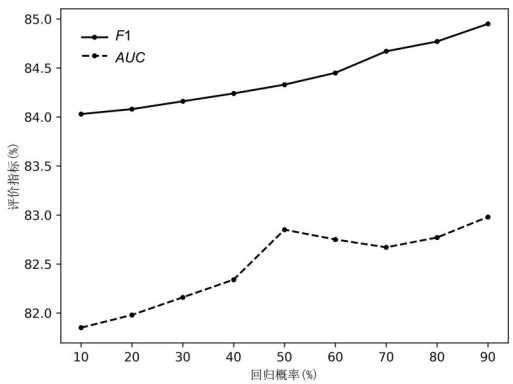

图 5　RWR算法的回归概率对算法的AUC和$F1$的影响

Fig. 5　*AUC* and *F*1 of RWR algorithm with different regression probability

另一个控制参数是采样得到的子图规模(子图中的节点数). 图6显示了采样的节点数从10个到100个时得到的AUC与$F1$. 由图可见,随着采样规模的上升,预测性能也随之上升,但采样数量达到70个时出现了拐点. 因为采样规模的上升可以获得更多的子图信息,但当采样节点过多时,算法不得不寻找离中心用户更远的节点,而这些节点不能为中心节点提供有价值的结构信息与用户特征信息,甚至会成为网络中的噪声,从而影响模型的最终表现.

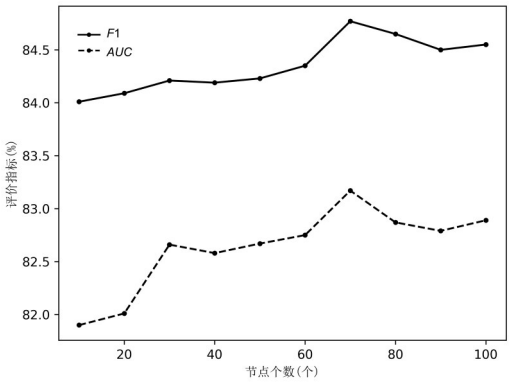

图6　采样节点的数量对RWR算法的*AUC*和*F*1的影响

Fig. 6　*AUC* and *F*1 of RWR algorithm with different size of sampled nodes

通过对数据集的分析可以发现,数据集实例中反例与正例的分布是不平衡的. 为了研究这个不平衡的比例对最终结果造成的影响,设置10个比例,分别计算相应的AUC与$F1$(图7). 由图可见,随着反例对正例的比值增大,$F1$出现明显的下降,而AUC则保持稳定.

4　结论与展望

本文研究的是针对社交网络影响力的预测. 首先从深度学习的角度将问题描述为一个标签分类任务,并使用图表征、图卷积与图注意力技术搭建了一个基于图的深度神经网络. 在三个现实世界的大规模数据集上测试了提出的模型,实验结果显示,使用图表征与图注意力方法的网络的表现远远强于其他的基准方法,但图卷积神经网络没有达到预期的结果. 本文的工作证明使用图卷积和图注意力方法可以有效提取社交网络中某一中心用户的特征信息和所处子图的结构信息,也

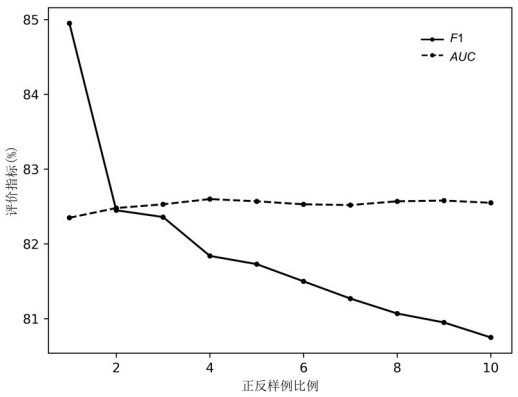

图7　数据集中正例与反例的比例对RWR算法的*AUC*和*F*1的影响

Fig. 7　*AUC* and *F*1 of RWR algorithm with different ratio of positive to negative examples

显示了图神经网络在研究社交影响力方面的潜力.

从本文架构中提炼的一般思想可以运用到许多网络挖掘任务中. 本文的方法可以有效利用用户的自我中心网络,通过研究该网络的结构与中心节点和其他节点的关系来获得局部网络信息. 这一方法可被链路预测(Link Prediction)、相似搜索(Similarity Search)、网络对齐(Network Alignment)等下游应用采用[33],具有广阔的应用前景.

参考文献

[1] Zhang J, Liu B, Tang J, et al. Social influence locality for modeling retweeting behaviors//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: AAAI Press, 2013: 2761－2767.

[2] Easley D, Kleinberg J. Networks, crowds, and markets. New York, USA: Cambridge University Press, 2010: 727.

[3] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017: 1025－1035.

[4] Hootsuite. Digital in 2019. We are social, 2019. 2019－05－10.

[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:ACM,2016:855－864.

[6] Guille A, Hacid H. A predictive model for the temporal dynamics of information diffusion in online social networks//Proceedings of the 21st International Conference on World Wide Web. Lyon, France: ACM,2012:1145－1152.

[7] Ugander J, Backstrom L, Marlow C, et al. Structural diversity in social contagion. Proceedings of the National Academy of Sciences of the United States of America,2012,109(16):5962－5966.

[8] Perozzi B, Ai-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:ACM,2014:701－710.

[9] Bakshy E, Rosenn I, Marlow C, et al. The role of social networks in information diffusion//Proceedings of the 21st International Conference on World Wide Web. Lyon,France:ACM,2012:519－528.

[10] Tang J, Qu M, Wang M Z, et al. LINE: Large-scale information network embedding//Proceedings of the 24th International Conference on World Wide Web. Florence,Italy:ACM,2015:1067－1077.

[11] Myers S A, Zhu C G, Leskovec J. Information diffusion and external influence in networks// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing,China:ACM,2012:33－41.

[12] Guo R C, Shaabani E, Bhatnagar A, et al. Toward order-of-magnitude cascade prediction//Proceedings of 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris,France:ACM,2015:1610－1613.

[13] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network// Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, USA: ACM, 2003: 137－146.

[14] Wang L R, Ermon S, Hopcroft J E. Feature-enhanced probabilistic models for diffusion network inference//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2012:499－514.

[15] Li C, Ma J Q, Guo X X, et al. DeepCas: An end-to-end predictor of information cascades//Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: ACM, 2017: 577－586.

[16] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016: 3844－3852.

[17] Su H Y, Gionis A, Rousu J. Structured prediction of network response//Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR, 2014: 442－450.

[18] Aris A, Ravi K, Mohammad M. Influence and correlation in social networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, USA: Association for Computing Machinery, 2008:7－15.

[19] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs// Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, NY, USA: JMLR, 2016: 2014－2023.

[20] Yanardag P, Vishwanathan S V N. Deep graph kernels//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia: ACM, 2015: 1365－1374.

[21] Luong M, Pham H, Manning C D. Effective approaches to attention-based neural machine translation//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: The Association for Computational Linguistics, 2015: 1412－1421.

[22] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks// Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013: 6645－6649.

[23] Wu Y C, Yin F, Liu C L. Influence and correlation in

social networks. WWW,2012,12(8):519－528.

[24] Ma H. An experimental study on implicit social recommendation//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland:ACM,2013:73－82.

[25] Dong Y X, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: ACM, 2017: 135－144.

[26] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013,arXiv:1301.3781.

[27] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas,NV,USA:ACM,2008:7－15.

[28] Yanardag P, Vishwanathan S V N. A structural smoothing framework for robust graph-comparison// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada:MIT Press,2015:2134－2142.

[29] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro,Brazil:ACM,2013:657－664.

[30] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the 25rth International Conference on Neural Information Processing Systems. Lake Tahoe,NV,USA:Curran Associates Inc.,2012:1097－1105.

[31] Borgwardt K M, Kriegel P H. Shortest-path kernels on graphs//Proceedings of the 5th IEEE International Conference on Data Mining. Houston, TX, USA: IEEE,2005:74－81.

[32] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. JMLR,2011,12(61):2121－2159.

[33] Chen S, Moore J L, Turnbull D, et al. Playlist prediction via metric embedding//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China:ACM,2012:714－722.

（责任编辑　杨可盛）