ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1134－1144
DOI:10. 16337/j. 1004-9037. 2022. 05. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于异构特征融合的论文引用预测方法

朱丹浩[1]，黄肖宇[2]

(1. 江苏警官学院刑事科学技术系,南京 210031;2. 江苏警官学院计算机信息与网络安全系,南京 210031)

摘　要：针对论文引用预测方法在特征稀疏时性能下降的问题,提出了基于异构特征融合的方法,可同时利用定长特征、引文网络特征和引文时序特征,有效提升了引用预测方法的精度。本文针对论文引用预测任务定义了引文属性网络,对3类异构特征进行建模;提出了面向异构特征融合的论文引用预测方法,使用图神经网络处理定长特征和引文网络特征,使用循环神经网络处理引文时序特征,基于多头注意力机制对提取到的异构特征进行融合并预测被引次数。在大规模真实数据集上的实验表明,本文方法可以有效利用多种异构特征并缓解数据稀疏问题,均方根误差(Root mean squatr error,RMSE)比最好的基准方法降低了0.31。

关键词：引用预测;循环神经网络;图神经网络;异构特征;注意力

中图分类号：G353.1　　**文献标志码**：A

Paper Citation Prediction Method Based on Heterogeneous Feature Fusion

ZHU Danhao[1], HUANG Xiaoyu[2]

(1. Department of Criminal Science and Technology, Jiangsu Police Institute, Nanjing 210031, China;2. Department of Computer Information and Network Security, Nanjing 210031, China)

Abstract: Aiming at the problem that the performance of the paper citation prediction method is degraded when the features are sparse, a method based on heterogeneous feature fusion is proposed, thus can use fixed-length features, citation network features and citation time series features at the same time, thus effectively improving the accuracy of the citation prediction method. Firstly, this paper defines a citation attribute network for the paper citation prediction task, and models three types of heterogeneous features. Secondly, a paper citation prediction method for heterogeneous feature fusion is proposed. The method uses the graph neural network to process fixed-length features and citation network features, uses the recurrent neural network to process citation time series features, and fuses the extracted heterogeneous features and predicts the number of citations based on multi-head attention mechanism. Experiments on large-scale real datasets show that the proposed method can effectively utilize multiple heterogeneous features and alleviate the problem of data sparsity, and its root mean square error(RMSE) is 0.31 lower than that of the best benchmark method.

Key words: citation forecast; recurrent neural network; graph neural network; heterogeneous features; attention

---

基金项目:国家自然科学基金(71974094);江苏省社会科学基金(19TQD002);江苏省教育厅自然科学项目(21JHB520004)。
收稿日期:2021-07-18;**修订日期**:2021-09-20

引 言

被引频次是最具代表性、最简单、最标准和最客观的度量学术影响力的指标[1],贯穿了科研活动的始终。例如,文献搜索引擎会根据被引次数调整检索结果的排序,科技期刊的分区主要依据所载论文的平均被引次数,学科热点的发现常常依赖于对引用网络进行聚类分析。然而,引用行为具有较长的滞后性,影响了各类下游任务的应用范围和性能。为解决这一问题,研究者尝试通过机器学习算法来预测论文的未来被引情况。例如,Ibáñez等[2]使用多元线性回归方法,基于摘要等文本特征预测了论文发表后4年内的引用次数;耿骞等[3]尝试了朴素贝叶斯和逻辑回归方法等。近年来,随着深度学习的发展,循环神经网络及前馈神经网络[4-5]也被引入论文被引预测算法中,预测精度取得了一定的提高。论文引用预测任务的一大特点是,存在大量的、形态各异的论文被引影响因素可作为输入特征,但单一影响因素的预测能力极弱,在具体论文上常常是稀疏的。例如,一经发表就被引用的论文常会被引更多,但一方面,也存在大量的"睡美人"文献[6],在发表后多年才突然被唤醒,成为研究的热点;另一方面,对于刚发表的新论文,并不存在早期被引,基于该特征的预测方法自然也就失效了。因此,如何充分利用异构特征,挖掘其中蕴含的复杂关联关系是建立论文引用预测方法的关键。现有的研究大多只能利用单一类型的特征,这不仅阻碍了预测精度的进一步提高,也限制了预测方法的适用范围。

基于以上考虑,本文提出了基于异构特征融合的论文引用预测方法。首先,本文针对论文引用预测任务定义了引文属性网络,对3类异构特征进行建模;其次,提出了面向异构特征融合的论文引用预测方法,使用图神经网络处理定长特征和引文网络特征,使用循环神经网络处理引文时序特征,基于多头注意力机制对提取到的异构特征进行融合并预测被引次数。本文在基于CSSCI真实数据集的实验证明了所提出方法的有效性,可以有效缓解数据稀疏问题。

1 相关研究

从使用特征的形态来看,当前的论文被引预测算法大体可分为3类:基于定长特征向量的方法、基于不定长引文时序特征的方法以及基于引文网络特征的方法。

定长的特征向量,主要是论文发表时即存在的特征,包括论文本身相关、期刊相关和作者相关3类[7]。例如,论文的摘要或标题中的关键词[8]、语言风格[9];载文期刊的影响因子[10]、载文量[11]、引用半衰期[12];作者之前的被引频次[13]以及是否获得诺贝尔奖等[14]。最常见的预测方法当属多元线性回归[2,15],其优点是可解释性较强,可以比较不同特征对论文被引的解释能力。但如果以提高预测精度为目标,该类方法有些力不从心,并不能挖掘因素间的非线性关系。耿骞等[3]尝试了朴素贝叶斯、逻辑回归、支持向量机、梯度提升决策树、XGBoost、AdaBoost和随机森林7种算法,发现XGBoost和随机森林可以取得最好的预测结果。

基于不定长引用时序特征的方法以论文发表后前若干年的逐年被引次数为输入,预测其后的被引次数。对于这一类特征,面向时间序列的序列化学习算法是自然的解决思路。Abrishami等[4]基于循环神经网络,以论文前若干年的被引次数为每一步的输入,预测论文在数年之后的被引次数。Liu等[5]结合了连续长短期记忆循环网络(Continuous-time long short-term memory,cLSTM)和神经霍克斯过程进行引用预测,他们认为该方法更能识别"睡美人"现象。

基于引用网络特征的方法将待预测论文看作引文网络中的节点,目前主要基于无监督的学习模式进行学习,不同于之前的分类或回归算法,这一类算法基于论文在引文网络中的拓扑信息,基于Pagerank或相近算法判断其在网络中的重要性,假设重要性更高的重要节点的引文排名会更高。相应的研究包括Walker等[16]、刘大有等[17]和Davletov等[18]。

现有的方法大多只能利用单一类型的特征,主要出于两个原因:(1)非经专门设计,多种类型的特

征很难兼容彼此。例如,引文网络特征是非欧几里得空间的数据,难以直接转化为定长特征。(2)方法本身只适用于单一类型的特征。例如,基于无监督网络学习的算法只能利用引文网络特征,无法建模其他两种特征类型。

尽管种类繁多,但对于具体的论文,特征常常是稀疏的。新发表的论文不存在被引网络和引用时序特征,大多数论文也不发表在重点期刊,或由知名学者发表。因此,建立能够同时利用多种特征的论文被引预测算法,可以有效缓解数据稀疏问题,提高预测精度。

2 兼容异构特征的论文引用预测任务定义

2.1 属性引文网络

本文定义了属性引文网络,可同时兼容定长特征、引文网络特征和引用时序特征,具体定义如下。

定义 1(属性引文网络) 令 $G=(V, W, X^f, X^c)$,其中,G 为属性引文网络,V 为网络中节点 $v_1, v_2, \cdots, v_n$ 的集合,节点 v_i 为第 i 篇论文,$n=|V|$ 为论文的数量。$W \in \mathbb{R}^{n\times n}$ 为节点的邻接矩阵,存储了论文之间的引用关系,其中的元素只能为0或1,如果为 $W_{i,j}=1$,表示论文 v_i 引用了 v_j。$X^f \in \mathbb{R}^{n\times f}$ 和 $X^c \in \mathbb{R}^{n\times c}$ 是节点的两类属性矩阵,分别为定长特征矩阵和引用时序特征矩阵,各自存储了论文本身的特征和历年被引用的次数。两个矩阵中,第 i 行表示论文 v_i 对应的属性向量,f 和 c 分别为两类属性的维度。尽管引用时序特征本身是不定长的,发表年份越久的论文特征维度越大,但本文使用填充技术将统一转换为同一长度,可提升定义的简洁性。

本文所使用的特征和编码方式见表1。此处重点对“期刊名称”“论文关键词”和“历年被引次数”进行介绍。“期刊名称”表示为单热点向量,即每个期刊对应于1个编号,在后续的图神经网络中,该编号将隐式地转换为稠密的期刊特征向量。由于每个期刊均会出现在多篇论文的 X^f 中,通过训练该期刊特征向量将会反映期刊本身的特性。“论文关键词”也是单热点向量,如果出现多个关键词,则多个维度的对应位置都被设为1。“历年被引次数”是论文发表后的逐年被引次数,本文根据所用数据设置长度为18,即对应于论文在1998—2015年的逐年被引次数。如果1篇论文是2014年发表的,则其对应向量在1998—2013年的维度上的值都设为0。

表1 本文所使用的特征和编码方式

Table 1 Features and coding methods used in the paper

编码位置	特 征	说 明
X^f	第一作者的发文量	1维的实数向量,存储了第一作者的历史发文篇数
	第一作者的平均被引次数	1维的实数向量,存储了第一作者的历史平均被引次数
	期刊名	单热点编码,维度等同于总期刊数
	论文关键词	单热点编码,维度等同于关键词总数
X^c	历年被引次数	18维的实数向量,论文过去若干年逐年的被引次数,在本文中设为18年;若是发表不满18年,则不存在被引的年份,用0填充

3种形式的特征对应于属性引文网络的位置如下:(1)定长特征,包括论文内容、期刊和作者等,存储于在内容属性矩阵 X^f 中;(2)引文网络特征,本文中即为 W;(3)不定长引用时序特征,对应于引用属性矩阵 X^c。

值得一提的是,限于篇幅、工作量和本文所使用数据集的特点,本文并未设计和使用更多的特征。属性引文网络具有良好的扩展性,足以编码绝大部分论文被引影响因素。例如,如果数据集中包含了学术全文本信息,则可在通过自然语言处理技术提取具体的引用行为特征后,编码至 X^c 中;期刊的影响因子、作者的H指数以及标题摘要等文本特征等也可直接附加至 X^f 中。

2.2 论文引用预测任务

本文对论文引用预测任务定义如下。

定义2(论文引用预测任务) 对于属性引文网络G,每一个节点v_i对应一个标签$y_i \in Y$,Y是标签的集合。已知属性引文网络G和一部分节点的标签$y_i \in Y_{\text{train}}$,Y_{train}指训练集的标签,论文引用预测的目标是学习出1个模型M,使得$M(v_i)=y_i$,$y_i \in Y_{\text{test}}$,Y_{test}指测试集的标签。

标签Y如果是离散的,例如高被引/低被引,论文引用预测可归类为分类任务;反之,如果Y直接是连续的被引次数,则可归类为回归任务。Dong等[19]则认为论文引用频次是长尾分布,不适用于回归预测。耿骞等[3]认为,将引用预测定义为分类问题,可以使预测粒度变粗,可利用更符合真实分布的数据,模型泛化能力更强,研究更有价值。但从机器学习模型的角度来看,分类方法是在回归预测目标后多加了一层分类层,对构建预测算法本身影响并不大。因此,本文直接以论文的被引次数为预测目标,即$Y \in \mathbb{R}^+$。

3 论文引用预测方法

3.1 总体框架

算法总体框架见图1。首先,以图的邻接矩阵和定长特征矩阵为输入,使用图神经网络学习出论文的网络特征表示;其次,以引用时序特征矩阵为输入,基于循环神经网络学习出论文的逐年引用特征表示;最后,基于多头注意力模型,融合网络特征表示和逐年引用特征表示,并预测论文的引用次数。

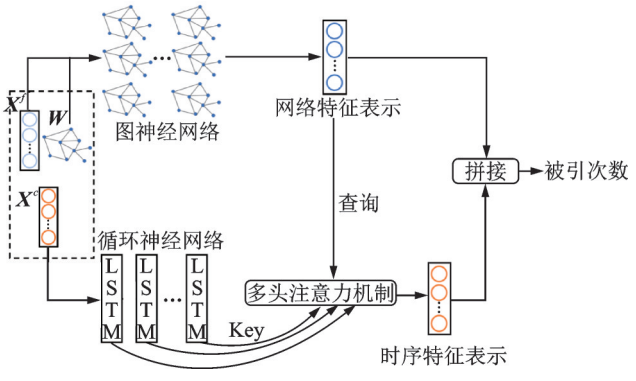

图1 本文方法总体框架图

Fig.1 Framework of the proposed method

3.2 基于图神经网络的网络特征表示学习

图神经网络系列算法是目前属性网络上最为强大的学习算法,其中最为经典的是图卷积神经网络(Graph convolution network, GCN)[20]。本文基于GCN,面向属性引文网络的特性进行了针对性的特征学习。整个GCN的输入为引文属性网络的邻接矩阵$\boldsymbol{W}$和定长特征$\boldsymbol{X}^f$,输出为所有论文的网络特征表示$\boldsymbol{S} \in \mathbb{R}^{n \times g}$,第$i$行对应于论文$v_i$的网络特征表示向量,维度为$g$。

整个算法可看作多层神经网络,在第k层中输入的节点属性矩阵为$\boldsymbol{H}^{(k)} \in \mathbb{R}^{n \times h_k}$,第$i$行对应于论文$v_i$在第$k$层的特征表示,维度为$h_k$。最初始的第0层被定义为输入层,即:$\boldsymbol{H}^{(0)}=\boldsymbol{X}^f$。每一层中,每一个节点都从其周围的邻接节点中搜集信息,并更新到下一层的节点属性特征向量中去。为了更好地利用节点本身的信息,需要首先对邻接矩阵增加自连接,使得节点可以直接利用上一层自己的信息,即

$$\bar{\boldsymbol{W}}: \boldsymbol{W}+\boldsymbol{I}_n \tag{1}$$

式中$\boldsymbol{I}_n$为对角线为1、其余位置均为0的方阵。再进行矩阵行和列的归一化处理,有

$$\bar{\bar{\boldsymbol{W}}}: \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{2}$$

式中$\boldsymbol{D}$为对角矩阵,元素为节点的度,有

$$D_{ii}=\sum_k \bar{\bar{W}}_{ik} \tag{3}$$

第k层的GCN函数为

$$\boldsymbol{H}^{(k+1)}=\mathrm{ReLU}\left(\bar{\bar{\boldsymbol{W}}}\boldsymbol{H}^{(k)}\boldsymbol{T}^{(k)}\right) \tag{4}$$

式中：$\boldsymbol{T}^{(k)}\in\mathbb{R}^{n\times h_k}$为线性转换矩阵；ReLU为非线性激活函数。

一般GCN多为2层，过多的层数会引起过平滑现象，从而导致性能下降。所以对于2层的GCN，论文的网络特征表示矩阵$\boldsymbol{S}$可由式(1)求得。$\boldsymbol{S}$的每一行对应于一篇论文的网络特征表示向量，有

$$\boldsymbol{S}=\bar{\bar{\boldsymbol{W}}}\mathrm{ReLU}\left(\bar{\bar{\boldsymbol{W}}}\boldsymbol{X}^{f}\boldsymbol{T}^{(0)}\right)\boldsymbol{T}^{(1)} \tag{5}$$

3.3 基于循环神经网络的引用特征表示学习

论文的引用特征$\boldsymbol{X}^{c}$，反映了学术界对工作的认同程度和引文曲线的形态。本文使用循环神经网络对时间序列进行建模，为解决循环神经网络的梯度爆炸和梯度消失问题，使用了长短期记忆单元(Long-short term memory unit，LSTM)[21]。为简化标记，此处令x为某篇论文的逐年被引次数，对应于$\boldsymbol{X}_i^{c}$的1行，$\boldsymbol{x}_t$表示论文第t年的被引次数。

首先，利用门函数计算遗忘门向量$\boldsymbol{f}_t$、输入门$\boldsymbol{i}_t$、输出门$\boldsymbol{o}_t$，以及单元状态更新值$\widetilde{\boldsymbol{c}}_t$，有

$$\begin{cases}\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f)\\ \boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o)\\ \widetilde{\boldsymbol{c}}_t=\sigma(\boldsymbol{W}_c\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c)\end{cases} \tag{6}$$

式中：σ为sigmoid函数；$\boldsymbol{h}_{t-1}$为论文在第$t-1$年的隐藏层向量；$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_o$、$\boldsymbol{W}_c$为线性转换矩阵；$\boldsymbol{b}_f$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$为偏置向量。

其次，基于上述4个向量对单元状态$\boldsymbol{c}_t$进行更新，并得到新的隐藏层向量$\boldsymbol{h}_t$，有

$$\begin{cases}\boldsymbol{c}_t=\boldsymbol{f}_t\times\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\times\widetilde{\boldsymbol{c}}_t\\ \boldsymbol{h}_t=\boldsymbol{o}_t\times\tanh(\boldsymbol{c}_t)\end{cases} \tag{7}$$

式中tanh为激活函数，即

$$\tanh(x)=\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}} \tag{8}$$

式中初始的$\boldsymbol{c}_0$和$\boldsymbol{h}_0$都设置为$\boldsymbol{0}$向量。

每篇论文得到一个逐年的隐藏层向量$\boldsymbol{h}_t$，按行堆叠，即可得到其逐年的隐藏层矩阵$\boldsymbol{H}\in\mathbb{R}^{c\times d_{\mathrm{h}}}$，$d_{\mathrm{h}}$表示LSTM的隐藏层维度。常见的LSTM常以$\boldsymbol{H}$的最后一列作为输出。而对于引用次数预测任务，由于“睡美人“等形态引用曲线的存在，中间步数的输出也可能具有重要的预测意义，本文将序列的每一步输出都保留下来，用于下一步针对性的特征融合。

3.4 基于多头注意力模型的特征融合和预测

注意力机制被广泛应用于机器翻译[22]、知识图谱[23]和目标识别[24]等领域，可以动态地聚焦于复杂特征的重要部分。本文使用多头注意力模型，基于论文的网络特征对其不同年份的引文时序特征进行注意力加权，从而实现不同类型特征的深度融合。

对于论文v_i，其网络特征表示向量记为$\boldsymbol{s}$，即为在2.2节所得的网络特征表示矩阵$\boldsymbol{S}$中的对应行数；对应的时序特征矩阵为$\boldsymbol{H}$，由2.3节得出。由于$\boldsymbol{H}$中包含了不同年份的论文引用时序特征，本文基于多头注意力机制，以$\boldsymbol{s}$为查询式，对不同年份的特征，也就是$\boldsymbol{H}$的不同列，赋予不同的权重，聚焦于对未来被引最具预测能力的时序特征。

首先，通过线性转换$\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$，将$\boldsymbol{s}$和$\boldsymbol{H}$转换为查询向量$\boldsymbol{q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$，有

$$\begin{cases} \boldsymbol{q}=\boldsymbol{W}_q\boldsymbol{s} \\ \boldsymbol{K}=\boldsymbol{W}_k\boldsymbol{H} \\ \boldsymbol{V}=\boldsymbol{W}_v\boldsymbol{H} \end{cases} \tag{9}$$

其次,对查询向量和键矩阵进行按列点乘,再通过softmax函数归一化后求每一列的权值,有

$$\boldsymbol{\alpha}=\mathrm{softmax}(\boldsymbol{q}^{\mathrm{T}}\boldsymbol{K}) \tag{10}$$

最后,不同时序的论文引用特征进行加权求和,其中$\boldsymbol{V}_{:,i}$表示$\boldsymbol{V}$的第i列,即有

$$\boldsymbol{v}=\sum_i \boldsymbol{\alpha}_i\boldsymbol{V}_{:,i} \tag{11}$$

由于不同的时序特征中包含着不同方面的信息,此处采用多头注意力特征机制,具体流程图见图2。即使用多组不同的$\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$,计算出不同的$\boldsymbol{v}$,记为$\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^m$,$m$为多头注意力的个数。

对多头注意力和$\boldsymbol{s}$进行拼接,再经过向量点乘后,得到了最终的预测结果

$$\widetilde{y}_i=\mathrm{ReLU}(\boldsymbol{u}^{\mathrm{T}}\mathrm{concat}(\boldsymbol{s},\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^m)) \tag{12}$$

式中:$\boldsymbol{u}^{\mathrm{T}}$为权重向量;$\widetilde{y}_i$为论文$v_i$的预测被引次数;concat是拼接函数;ReLU激活函数除了可以提供非线性转换,还能保证预测的被引次数大于等于0。

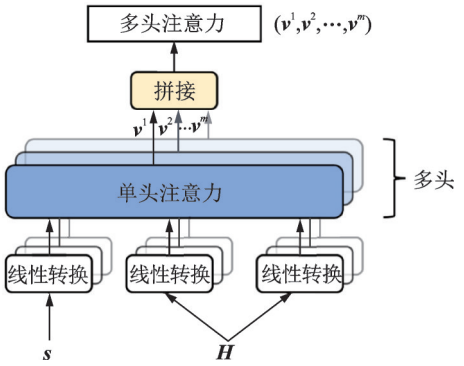

图2 多头注意力计算流程图

Fig.2 Flow chart of multi-head attention calculation

本文使用均方根误差(Root mean square error,RMSE)计算损失函数为

$$\mathrm{Loss}=\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{y}_i-y_i\right)^2 \tag{13}$$

再使用反向传播算法优化模型中的所有参数,包括GCN、LSTM和特征融合模块中的所有参数。

3.5 讨 论

在特征融合时,为何要区分$\boldsymbol{X}^f$和$\boldsymbol{X}^c$,对其分别使用GCN和FNN进行特征表示学习;而不是直接合并$\boldsymbol{X}^f$和$\boldsymbol{X}^c$输入到1个GCN中进行预测?这是由论文被引预测本身的性质决定的。在GCN中,属性通过邻接边传递到相邻的节点上去,相邻的节点常常会学习出相近的属性和标签。因此,使用GCN预测论文的学科时很容易取得成功[20]。而在预测论文被引时,相邻的2个论文节点的引用差距极大是常见的现象,比如一篇经典论文发表10年,被引数百次,而另一篇论文刚刚发表,尚未获得被引,经典论文的被引属性传递到新论文上,会严重高估新论文的预测被引次数。基于以上考虑,本文对两类特征进行区分学习,避免上述的信息传播问题。

4 实 验

4.1 数据集和评测标准

本文使用的数据库为中文社会科学引文索引(Chinese social sciences citation index,CSSCI)1998—2020年的数据,该数据库包含了中文核心期刊论文的题录和引文信息。本文以1998—2015年的数据构建了引文属性网络,并预测网络中论文在16~20年间的被引次数。按5:1:4的比例随机设置了训练节点、验证节点和测试节点。需要强调的是,本文采用的是半监督的学习模式,也就是说,整个网络在训练阶段对于模型都是可见的,但隐去了验证节点和测试节点的标签。表2给出了引文属性网络的总体统计信息。其中节点的属性由16 601维关键词的稀疏向量、672维期刊的稀疏向量、1维的作者历史被

引次数和1维的作者历史平均被引组成。本文只保留了出现频次20以上的关键词。

图3给出了属性和标签的分布。第1行的3张和第2行的第1张是节点的属性，总体上呈现明显的长尾分布，但其中期刊的分布较为平滑。第2行的第2张给出了引文的间隔，第0年的引用较少，第1、2年的引用达到高峰，之后逐年下降。第2行的第3张是待预测的标签，也就是2016—2000年的被引次数，大部分的论文被引次数均是0次，引用次数在9次以下的占了绝大部分，极少数论文会被引更多次。

表2 引文属性网络的总体统计信息

Table 2 Overall statistical information of citation attribute networks

网络参数	数值
结点总数量	686 591
边的数量	1 390 980
训练节点数量	343 295
验证节点数量	68 659
测试节点数量	274 637
节点属性维度	17 275

本文使用在测试集上的RMSE来评测算法的精准度，该指标越低，表示预测的精准度越高。

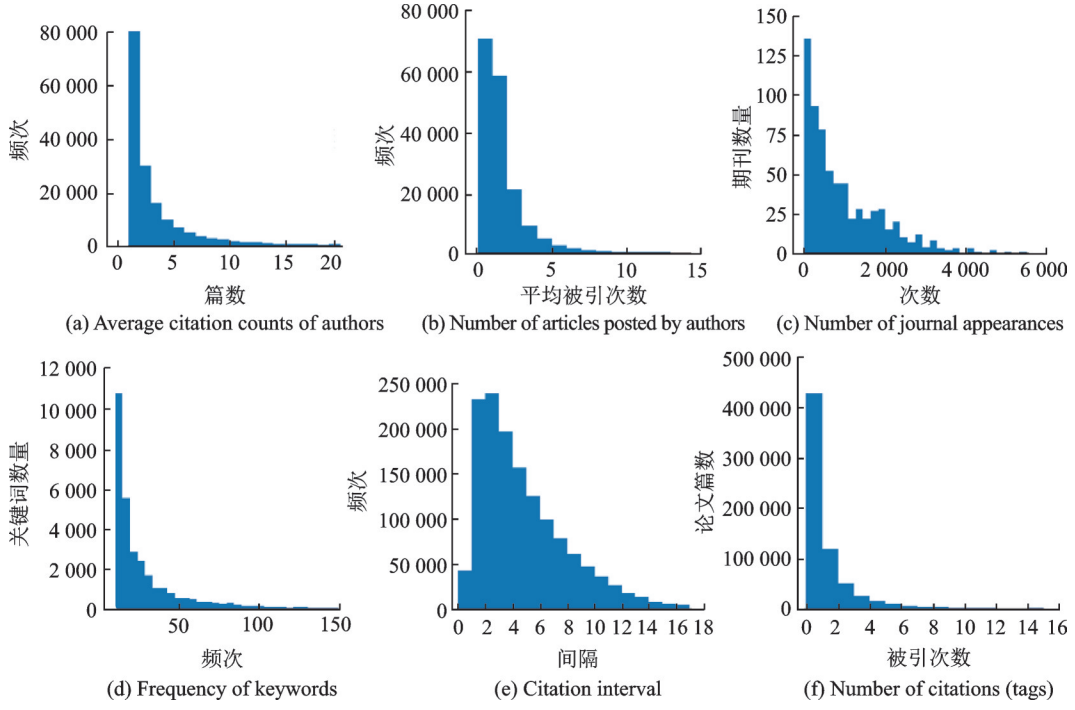

图3 属性和标签的分布图

Fig.3 Distribution of attributes and labels

4.2 基准方法和训练过程

4.2.1 基准方法

本文对比了在论文被引预测中常用的3种算法，这些方法基于不同类型的特征进行学习。本文的数据集远大于之前的研究，例如本文的训练集包含了34万篇论文，而耿骞等[3]的训练数据约包含2.6万篇论文。因此，在小数据集上常用的算法，如支持向量机、随机森林等，因内存和训练时间的限制不再适用，故本文主要选择了在大数据集上性能和表现优秀的神经网络系列算法作为基准。

随机猜测：选取测试集上所有标签的平均值作为预测结果，经统计为1.31次，该方法忽略了所有输入特征的作用。所有其他方法的结果均应优于随机猜测。

前馈神经网络(Feed-forward neural network, FNN)：前馈神经网络是经典的神经网络。深度为2层，隐藏层的单元数为512，使用Adam梯度下降[25]进行优化，初始值设为0.01，使用Dropout[26]技术避

免过拟合,概率值设为0.3,批大小为1 000,在测试集上反复训练,最多50个Epoch。

循环神经网络(Recurrent neural network, RNN):循环神经网络适用于不定长的特征序列,常在论文引用预测任务中被用来建模引用时序特征。本文基于Abrishami等[4]的设置进行了实验,使用LSTM为基本单元。深度为1层,隐藏层的单元数为512,使用Adam梯度下降进行优化,初始值设为0.01,Dropout技概率设为0.3,批大小为1 000,在测试集上反复训练,最多50个Epoch。

图卷积神经网络[20]:标准的图神经网络,层数为2,隐藏层单元数为32,Dropout概率设为0.5,使用Adam优化,初始学习率为0.01,训练了200轮。

4.2.2 实验细节

实验环境:全部代码基于Python 3.6实现,使用Pytorch深度学习框架。实验运行在2核的Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz服务器上,整个模型训练时间大约为1.5 h。GCN模块基于DGI①,目前流行的图神经网络框架实现。

评测指标:本文使用RMSE作为评测指标,可以看作是预测被引次数和实际被引次数的平均偏差次数,该指标越小,表示预测越为精确。如果不经训练,直接随机猜测的话,最低偏差为4.99次。

预测和训练过程:基于Early stop技术,在验证集上选取RMSE最小的模型作为最终模型,并汇报该模型在测试集上的结果。

具体参数设置:深度学习的结果和超参数、实验设置密切相关,表3中详述了本文的实验参数。

表3 本文方法的主要参数配置

Table 3 Main configure parameters of the proposed method

参数名	参数值和说明
LSTM层数	1
LSTM维度	16
GCN层数	2
GCN维度	16
多头注意力个数	4
Dropout概率	0.5
优化算法	Adam
最大训练轮数	200
初始学习率	0.01
批大小	1 000

4.3 实验结果

4.3.1 总体比较

实验结果见表4。RNN、FNN方法以论文的历史逐年被引次数$\boldsymbol{X}^c$为特征时,平均偏差为3.49次和3.44次,取得了较大提升。在结合$\boldsymbol{X}^c$和$\boldsymbol{X}^f$后,两种方法都取得了进一步提升,平均偏差分别下降到了3.21次和3.16次。经典的图神经网络方法GCN以$\boldsymbol{W}$、$\boldsymbol{X}^f$为输入时RMSE为3.89次,但在引入$\boldsymbol{X}^c$后,预测精度未见明显提高。本文方法综合使用了所有的3种特征,取得了最好的预测精度2.85次,偏差比第二名,使用了$\boldsymbol{X}^f$、$\boldsymbol{X}^c$特征的FNN,下降了0.31次。

综上,可以得到3点结论:(1)本文方法能够利用多种异构特征,取得了最好的预测精度;(2)引用特征$\boldsymbol{X}^c$对于预测引用次数极为关键,仅基于该特征的RNN和FNN方法都取得了很好的预测精度;(3)本文的特征融合方法是有必要的,更能适应引用预测任务的特性。相比而言,使用全部特征的GCN方法和本文方法在输入特征上是公平的,但并未比仅使用$\boldsymbol{W}$、$\boldsymbol{X}^f$的GCN有大幅度提升,这表明GCN并不能很好地利用好引用特征$\boldsymbol{X}^c$。

表4 不同方法的预测结果

Table 4 Prediction results of different methods

方法	利用特征	RMSE
随机猜测方法	无	4.99
RNN	$\boldsymbol{X}^c$	3.49
	$\boldsymbol{X}^f$,$\boldsymbol{X}^c$	3.21
FNN	$\boldsymbol{X}^c$	3.44
	$\boldsymbol{X}^f$,$\boldsymbol{X}^c$	3.16
GCN	$\boldsymbol{W}$,$\boldsymbol{X}^f$	3.89
	$\boldsymbol{W}$,$\boldsymbol{X}^f$,$\boldsymbol{X}^c$	3.77
本文方法	$\boldsymbol{W}$,$\boldsymbol{X}^f$,$\boldsymbol{X}^c$	2.85

①https://docs.dgl.ai/

4.3.2 有效性分析

通过比较不同年份上本文方法和基准方法的表现，验证了异构特征融合方法的有效性。图4(a)比较了发表年份不同时GCN方法和本文方法的结果，其中GCN方法的特征是$\boldsymbol{X}^f$和$\boldsymbol{X}^c$的拼接。也就是说，此时的GCN方法和本文方法输入的特征是完全一致的。GCN方法在1～2年时和本文方法结果很接近，此时引用特征$\boldsymbol{X}^c$的信息还较为稀疏，当年份增加时，GCN方法的RMSE的下降趋势并没有本文方法明显，这表明GCN并不能很好地利用引用特征，也验证了3.5节中的讨论。图4(b)比较了不同发表年份下RNN方法和本文方法的对比。可以看出，发表年份越大时，预测的精度越好(RMSE越小)。而论文刚发表的1～2年之间，引用数据极为稀疏，此时预测的偏差较大，但本文方法的RMSE相对提高较大。这表明，本文方法可以较好地应对数据稀疏问题。

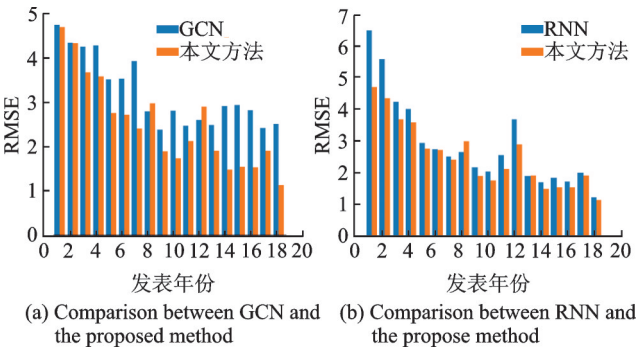

图4 不同年份下基准方法和本文方法的RMSE对比

Fig.4 Comparison on RMSE between the benchmark method and the proposed method

4.3.3 参数敏感性分析

图5给出了不同超参数设置下本文方法的性能。图5(a)显示，LSTM的隐藏层维度在16时，取得了最低的RMSE值，随着维度值增加，模型的误差上升较慢。图5(b)显示，GCN的隐藏层维度在16时，RMSE最低。图5(c)给出，随着多头注意力个数的上升，模型误差下降，到4时取得最好结果，再提升注意力个数会导致误差急速上升。

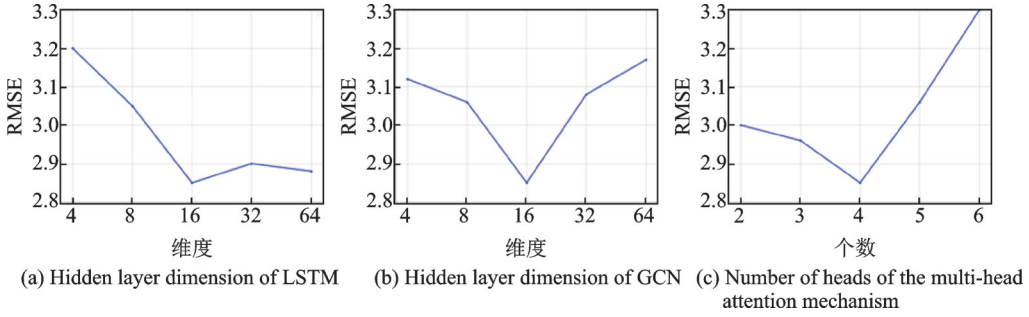

图5 参数敏感性分析

Fig.5 Parameter sensitivity analysis

在选择超参数时，本文方法的预测误差对GCN的隐藏层维度不太敏感，对LSTM的隐藏层维度较为敏感。选择多头注意力的个数要格外小心，会明显影响最终结果的精度。

4.3.4 多头注意力的可视化

图6给出了2篇典型论文的多头注意力热力图。颜色越深表示权重越高，横坐标中1对应1998年，18对应2015年。从图6(a)可知：(1)越靠后的年份对未来的被引次数预测越重要；

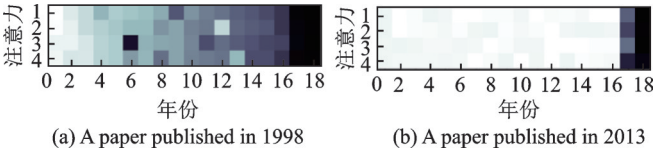

图6 多头注意力的热力图

Fig.6 Heat map of multi-head attention mechanism

(2)不同的注意力聚焦了不同年份的信息,例如,第1行和第2行相比,更注重于中间靠后部分的年份,第3行选中了第6年的引文信息用于预测。从图6(b)中可以看出,由于该论文发表于2013年,横轴编号16之前的年份不存在被引信息,注意力模型基本上不再聚焦于这些年份,这表明注意力模型可以有效聚焦于部分重要年份。

5 结束语

本文提出了一种融合异构特征的论文引用预测方法,可以有效利用定长特征、引用时序特征和引文网络特征进行被引预测。在CSSCI数据库18年的大规模数据上的实验证明,本文方法可以有效解决数据稀疏问题,提高预测精度,RMSE比最好的基准方法降低了0.31。尽管本文所提出的预测框架可以涵盖多种异构特征,但还不足以建模引文上下文信息。早期引文上下文中蕴含了学者对论文的初步评价,对未来被引具有重要指征意义。下一步工作探索使用具有边属性的引文网络来组织数据,将引文上下文信息放置边上,并设计针对性的预测方法。

参考文献:

[1] 陈仕吉,史丽文,左文革.基于ESI的学术影响力指标测度方法与实证[J].图书情报工作,2013,57(2):97-102,123.
CHEN Shiji, SHI Liwen, ZUO Wenge. Theoretical and empirical study on measure method of academicinfluence indicator based on ESI[J]. Library and Information Service, 2013,57(2): 97-102,123.

[2] IBÁÑEZ A, LARRAÑAGA P, BIELEA C. Predicting citation count of bioinformatics papers within four years of publication [J]. Bioinformatics, 2009, 25(24): 3303-3309.

[3] 耿骞,景然,靳健,等.学术论文引用预测及影响因素分析[J].图书情报工作,2018,62(14):29-40.
GENG Qian, JING Ran, JIN Jian, et al. Citation prediction and influencing factors analysis on academicpapers[J]. Library and Information Service, 2018,62(14): 29-40.

[4] ABRISHAMI A, SADEGH A. Predicting citation counts based on deep neural network learning techniques[J]. Journal of Informetrics, 2009,13(2): 485-499.

[5] LIU L, YU D, WANG D, et al. Citation count prediction based on neural hawkes model[J]. IEICE Transactions on Information and Systems, 2020,103(11): 2379-2388.

[6] 胡泽文,任萍,沈佳慧.融合K值算法与三指标的神经科学领域"睡美人"论文识别及影响因素探析[J].现代情报,2022,42(3):147-156.
HU Zewen, REN Ping, SHEN Jiahui. Identification of sleeping beauties in neuroscience through combining K value and three indicator methods and analysis on their influencing factors[J]. Journal of Modern Information,2022,42(3): 147-156.

[7] TAHAMTAN I, AFSHAR A S, AHAMDZADEH K. Factors affecting number of citations: A comprehensive review of the literature[J]. Scientometrics, 2016, 107(3): 1195-1225.

[8] FENG G, CHAO M, QINGLING S, et al. Succinct effect or informative effect: The relationship between title length and the number of citations[J]. Scientometrics, 2018,116: 1531-1539.

[9] SOHRABI B, IRAJ H. The effect of keyword repetition in abstractand keyword frequency per journal in predicting citation counts[J]. Scientometrics, 2017, 110(1): 1-9.

[10] VIEIRA E S, GOMES J A N F. Citations to scientific articles: its distribution and dependence on the article features[J]. Journal of Informetrics, 2010, 4(1): 1-13.

[11] 王群英,林耀明.影响因子、总被引频次与期刊载文量的关系研究——以资源、生态、地理方面的8个期刊为例[J].中国科技期刊研究,2012,23(1):76-79.
WANG Qunying,LIN Yaoming. A study on the relationship between impact factor, total citation frequency and journal article volume in eight journals in resources, ecology and geography[J]. Chinese Journal of Scientific and Technical Periodicals,2012,23(1): 76-79.

[12] BIGLU M H. The influence of references per paper in the SCI to impact factors and the matthew effect[J]. Scientometrics,

2008, 74(3): 453-470.

[13] FRANDSEN T F, NICOLAISEN J. The ripple effect: Citation chain reactions of a Nobel prize[J]. Journal of the Association for Information Science & Technology, 2014, 64(3): 437-447.

[14] BORNMANN L, DANIEL H D. Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at angew and techemie international edition[J]. Journal of Informetrics, 2010, 4(1): 83-88.

[15] LI X, THELWALL M, GIUSTINI D. Validating online reference managers for scholarly impact measurement[J]. Scientometrics, 2012, 91(2): 461-471.

[16] WALKER D, XIE H, YAN K K, et al. Ranking scientific publications using a model of network traffic[J]. Journal of Statistical Mechanics: Theory and Experiment, 2007(6): 06010.

[17] 刘大有,齐红,薛锐青. 基于作者权威值的论文价值预测算法[J]. 自动化学报, 2012, 38(10): 1654-1662.
LIU Dayou, QI Hong, XUE Ruiqing. The paper value prediction agorithm based on the authors authority value[J]. Acta Automatica Sinica, 2012, 38(10): 1654-1662.

[18] DAVLETOV F, AYDIN A S, CAKMAK A. High impact academic paper prediction using temporal and topological features [C]//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. Shanghai: ACM, 2014: 491-498.

[19] DONG Y, JOHNSON R A, CHAWLA N V. Will this paper increase your H-index?: Scientific impact prediction[C]// Proceedings of Machine Learning and Knowledge Discovery in Databases. Porto: Springer, 2015: 149-158.

[20] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016-01-01) [2021-08-30]. http://arxiv.org/abs/1609.02907.

[21] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL].[2021-08-30]. http://arxiv.org/abs/1706.03762.

[23] DEEPAK N, JATIN C, CHARU S, et al. Learning attention-based embeddings for relation prediction in knowledge graphs [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 4710-4723.

[24] REDMON J, FARHADI A. Yolov3: An incremental improvement[EB/OL].(2018-02-10)[2021-08-30]. http://arxiv.org/abs/1804.02767.

[25] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-07-15)[2021-08-30]. http://arxiv.org/abs/1412.6980.

[26] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

作者简介:

朱丹浩(1986-),**通信作者**,男,讲师,研究方向:自然语言处理、深度学习,E-mail: zhudanhao@jspi.edu.cn。

黄肖宇(2002-),男,本科生,研究方向:深度学习,E-mail: 1541673949@qq.com。

(编辑:刘彦东)