

融合多特征和双向图分类的专家推荐方法

丁婧娴, 李翔, 孙纪舟, 周泓

(淮阴工学院计算机与软件工程学院, 淮安 223003)

摘要: 专家推荐是推荐系统领域的一个研究热点, 专家信息特征提取的合理性直接影响到推荐的准确性。然而多数专家推荐方法未对多源信息构建特征关系文本图, 忽略了属性特征之间的相关性, 以及无法依据关联性拓展知识领域特征。针对以上问题本文提出了一种融合多特征和双向图分类的专家推荐方法 CMFBG。首先通过多源信息融合获取专家个体多特征信息, 并对不同属性特征构建类内文本图; 然后分别使用基于 Transformer 的双向编码器表示 (Bidirectional encoder representation from transformer, BERT) 模型和图卷积神经网络 (Graph convolutional network, GCN) 模型对特征提取并融合; 最后通过双向注意力机制增强源数据对图特征的扩展, 实现图结构上的分类。在同一专家数据集上进行实验分析, 结果表明在图分类任务中 CMFBG 精确率高于其他算法, 达到了 91.71%。

关键词: 专家推荐方法; 双向图卷积神经网络; 多特征融合; 图结构分类

中图分类号: TP391 **文献标志码:** A

Expert Recommendation Method Combining Multi-features and Bi-directional Graph Classification

DING Jingxian, LI Xiang, SUN Jizhou, ZHOU Hong

(Faculty of Computer and Software, Huaiyin Institute of Technology, Huaian 223003, China)

Abstract: Expert recommendation is a research hotspot in the field of recommendation system. The rationality of expert information feature extraction directly affects the accuracy of recommendation. However, most expert recommendation methods donot build text graphs of feature relation for multi-source information, and ignore the correlation between attribute features. Additionally, most expert recommendation methods cannot expand the features of knowledge field according to the relevance of text graph. Therefore, we propose CMFBG, an expert recommendation method combining multi-features and bi-directional graph classification. Specifically, CMFBG obtains multi-feature information of experts through multi-source information fusion, and construct text graphs for different attribute features within categories. Then, CMFBG employs bidirectional encoder representation from transformer (BERT) and graph convolutional network (GCN) models to extract features and fuse them. Finally, CMFBG employs the bidirectional attention mechanism to enhance the extension of the source data to the graph features and realize the classification of the graph structure. The experimental analysis on the same expert data set

shows that the precision of CMFBG is 91.71% higher than other algorithms in the task of graph classification.

Key words: expert recommendation method; bidirectional graph convolutional neural network; multi-feature fusion; graph structure classification

引 言

专家推荐是推荐系统的应用热点之一,在企业发展、领域科研评审、在线社区问答等方面有着广泛应用。专家推荐对专家学术研究领域和专家关系等多角度进行整合分析,以不同形式向企业、政府部门呈现选择列表。由于专家的专业领域通常不限于单一学科,同时科技项目对特定领域专家的需求未获得精确的推荐,因此引入了科技项目在特定专业领域的专家推荐^[1]。科技领域专家推荐包含关系特征、属性特征及文本特征等多特征表达和建立专业领域名词的标签体系。在知识领域体系中,找出特定领域主题内知识最丰富的专家是一个具有挑战性的问题。

科技领域专家推荐系统的主要目标是确定专家的知识领域并匹配与科技领域最合适的专家。专家推荐将推荐算法实例化,目的是为了满足不同企业、科研项目等实体对专家的需求。专家推荐的实质是提取专家多特征标签信息,基于多特征实现分类,给出推荐列表^[2]。专家推荐系统分为4个过程:(1)确定专家专业信息,提取与专家专业知识相关的信息;(2)专家信息建模,以专家领域知识为基础建立专家模型;(3)查询匹配,对查询匹配机制建模,以帮助推荐对象找到最合适的专家模型;(4)结果呈现^[3]。专家推荐方法包括专家个体多维特征刻画、面向知识问答社区推荐、领域科研项目评审推荐以及科研项目与专家自动匹配等^[4-6],较少关注于使用多特征构建文本图、深入探索实体之间的关系特征。

科技领域专家数据涉及到的专业领域和术语多且复杂,导致数据处理、特征提取和文本图构建的任务量和难度增加。一方面大部分现有专家推荐技术仅考虑了单一特征,忽略了特征之间的联系以及拓展领域问题,未对多源信息构建特征和关系文本图,特别是对于拥有先进的工业领域知识和科研资源的专家数据,例如专业领域的最新知识与企业问题之间异构数据的对接,知识领域专家关系的拓展等问题。另一方面,多源渠道高校和科研机构的专家信息及相关科研成果,多源异构的数据形式加大了挖掘数据信息之间浅层以及深层关系特征的难度,对数据处理和特征提取同样带来了挑战。因此,专家数据处理的精确性以及合理提取、构建多特征文本图直接影响推荐方法的准确性和应用效果。为了更好地实现基于多源信息推荐,提取数据的多特征之后构建数据文本图,文本图与图卷积神经网络相结合的方式成为新的推荐方法。

本文针对以上问题提出一种融合多特征和双向图分类的专家推荐方法,记为CMFBG。该方法首先对专家的基础数据处理,构建实体的多特征,将多特征基于单词共现和整体特征-单个标签关系为实体构建单个文本图;然后阐述多特征嵌入和CMFBG实现过程;最后以专家信息和文献信息文本为实验数据,根据实验结果分析CMFBG的优势以及未来工作。

1 相关工作

1.1 深度学习推荐系统

深度学习推荐系统缓解了传统推荐系统中的冷启动问题和稀疏矩阵问题,提高了推荐系统的可解释性和准确性。Zhang等^[7]对深度学习推荐系统的研究成果总结归纳,将最新推荐模型分为:单模块深度学习推荐模型和深度学习混合推荐模型。其中,单模块深度学习模型在推荐对象方面,提高了推荐的普遍适用性和推荐对象针对性。例如,卷积神经网络(Convolutional neural network, CNN)在多模态

信息数据中提取局部以及全局特征表示;回归神经网络(Recurrent neural network, RNN)基于内容信息对推荐系统中的时间顺序和动态演化建模。

Pan等^[8]通过微调基于Transformer的双向编码器表示(Bidirectional encoder representation from transformers, BERT)模型针对局部缺失信息问题,使用预训练模型的输出并与文本卷积神经网络构建文本分类模型。利用用户文本类别的概率分布计算用户之间的兴趣相似度,实现Top-N好友推荐。Jiang等^[9]提出幅度有界矩阵分解,通过约束每个单独的用户和项目特征向量构建模型,使有约束的优化问题变为可以通过随机梯度下降等无约束优化算法解决,允许单个用户和项目有不同的界限,解决了传统推荐算法在预测中获取无限制的和波动的值问题。Li等^[10]在基于用户行为分析的基础上,开发了由推荐系统注意模块和捕获静态和动态用户切换行为的通道注意模块组成的深度神经网络模型,生成了实时更新的供用户考虑的频道列表,融合推荐进一步提高了推荐的准确性。Abolghasemi等^[11]提出三阶段群组推荐系统,适应不同群组成员的个性,依次使用二进制矩阵分解方法、影响图以及意见应用动态模型,以获得基于决策力影响力的群组推荐系统。Lee等^[12]提出一种基于卷积神经网络的推荐系统,使用特征的外积矩阵和交叉卷积滤波器,处理各种类型的特征同时捕获用户和项目之间有意义的高阶交互,判断重要特征并给予大的权重,缓解了过度拟合问题。Vivekanandan等^[13]将用户评分序列和用户社交影响力结合,使用长短期记忆(Long short-term memory, LSTM)网络对用户评分序列考虑用户兴趣的动态,同时利用CNN提取电影的特征,提高了基于内容过滤的有效性。以上深度学习推荐方法在推荐系统中取得了良好的效果,但面对存在拥有复杂文本的科技领域专家数据效果还有待提高。这是由于专家数据存在多维信息并且专业术语多,导致文本属性较多,特征之间的联系难以构建。

图卷积神经网络(Graph convolutional network, GCN)^[14]在图神经网络的基础上使用空域和频域两种方式提取图结构特征,可以为图形中的不同实体提取高质量的嵌入。GCN为处理图结构数据提供了新的办法,并将图像卷积方法应用于具有拓扑关系的图结构数据。GCN的特点是从图结构数据中提取特征,通过传播的方式学习图结构中每个节点之间的交互,实现多级邻居的信息传输。Yao等^[15]使用单词共现和文档-单词关系为语料库建立单个文本图,文本图卷积网络(TextGCN)构建文本和单词的异构图,在类标签监督下,联合学习单词的嵌入、预测,提高了文本分类的准确率。Fan等^[16]依据图神经网络固有的集成节点信息和拓扑结构,建立用户-用户社交图 and 用户-项目关系图,提出了一种基于注意力机制来区分社会关系的异质强度的图神经网络框架,完成了用户-项目图 and 用户-用户社交图之间的交互。Wang等^[17]开发了一种新的基于空间的深图卷积神经网络模型,将从空间图卷积层中提取的无序顶点的顶点特征转换为固定大小的局部顶点网格结构,通过以优先顺序保留指定数量的顶点来实现的。Nieper等^[18]提取并规范化一个固定大小的局部邻域,每个邻域的顶点基于相同的图标记方法和图规范化工具重新排序,规范化邻域可以作为卷积运算的根顶点的感受域,因此该过程自然会为每个图形成一个局部级别的固定大小的顶点网格结构,通过在相邻顶点上滑动固定大小的经典卷积滤波器来执行图卷积运算。Duvenaud等^[19]提出一种图卷积网络预测方法,该预测管道不受限于文本图的结构以及大小,实现了端到端学习进行操作以及对标准特征的提取方法。

Monti等^[20]提出了第一种基于GCN的推荐系统方法。GCN用于聚合来自两个辅助用户和项目图的信息。在每个聚合步骤后更新用户和项目的潜在因素,并使用GCN和矩阵分解(Matrix factorization, MF)的组合目标函数来训练模型。Guo等^[21]提出具有邻居复杂交互的图卷积推荐网络,重点关注捕获节点邻居之间可能的共生信号,将异构多关系图与特征传播融合,细化了模型预测的用户和项目表示。Mei等^[22]将GCN应用于用户项图中的邻域信息,基于评级矩阵行或列的成对余弦相似性构建用户-用户和项目图。Ma等^[23]为解决用户行为多样化而产生的问题,提出具有适当学习机制的交互式异

构图,在行为图上获得实体语义,输出图卷积神经网络和注意力机制学习语义表示和行为序列表示,同时引入衡量用户兴趣集中度的行为图的核心和相关特征,实现了不同用户的不同集中程度推荐。Gwadabe等^[24]为解决信息有限和不确定的用户行为问题,提出了基于图神经网络模型以端到端的方式学习当前会话项目之间的非顺序交互,再进行顺序转换交互的方法,提高了推荐系统性能。Achary等^[25]将用户评分的相似性和用户人口统计和位置信息相结合,构建出图模型,利用自动编码器的优势提取特征,使用组合属性的新特征聚类用户,在冷启动问题上取得较好效果。Gao等^[26]通过用户给出的评分对被评分项目类别的有用性构建类别信息交换图,量化用户在社会推荐系统中的影响分数,提出无监督的影响分数测量方法来估计用户对其他用户的影响程度,以获得用户对项目的评分可以影响其他用户对该项目的感知,最后在基于图神经网络的推荐模型中加入类别用户影响分数,执行缺失评级预测任务。Liao等^[27]用加权线性聚合器来聚合邻居的特征,得到用户和项目在交互图和社交图表示,同时使用图融合操作在图传播过程中聚合用户表示。Shang等^[28]使用图卷积完成了通过与文本的嵌入组合进行药物推荐任务,反馈到BERT中进行预测,在Transformer中丢弃无法捕获所有的本地信息文本序列。Zheng等^[29]利用图卷积网络注重多样化的用户满意度,在推荐候选项的基础上加入衡量推荐项目的多样化因素,结合多样化优化推荐候选项。Jeong等^[30]提出使用论文引文图的引文推荐任务组合,局部信息和全局信息之间相互结合,提高GCN输出和BERT输出连接起来的预测任务效率。

1.2 多特征融合

典型相关分析(Canonical correlation analysis, CCA)^[31]核心思想是探究两组变量之间关系,最终变为两组线性组合的相关分析,通过特征对齐捕获多个特征之间的关系。Xu等^[32]提出结合上下文相关特征和基于时间卷积网络(Temporal convolutional network, TCN)和CNN的多阶注意力模型,使用TCG捕获单词和概念层面的上下文相关特征,同时衡量特征的重要性,通过注意机制与CNN的融合,在每个阶段获得区分特征,解决了短文本的数据稀疏性和歧义性,提高了模型的并行性。Tan等^[33]提出双向映射的标签特定特征多标签学习方法,通过重建损失函数鉴别映射过程中信息的损失,使用因果学习集合探索标签之间的内在因果关系以挖掘标签之间的本质联系,提高了标签多特征的分类效率。Dong等^[34]提出金字塔特征融合和全局上下文注意网络,从骨干网络中提取多尺度特征,融合金字塔特征模块,跳过连接将多特征融合成分辨率,后将全局上下文注意模块应用于相邻融合特征图,改善了预测结果。Abdi等^[35]使用RNN克服单词顺序和信息消失的缺点,同时建立包含了单词嵌入、语言知识、移位规则等特征向量组成的特征集,提高了句子级别的分类精度。Huang等^[36]基于分层多标签文本提出逐级学习文本表示的混合嵌入分类,混合嵌入分别通过增加层架结构的结构特征和不同类别的语义特征,在大量密切相关类别的文本分类得到较好的结果。Zhang等^[37]提出多级特征融合,将数据定时、字节和统计特征融合,从多个角度提取了有效信息。综上可知,现有的多种特征融合方法在各领域取得了较好的效果,但面对多维信息的科技领域专家数据效果还有待提高。

2 本文推荐模型

针对传统方法特征提取模型存在的特征向量稀疏、仅考虑单一特征的问题,本文利用文本图获得特征之间联系以及多特征的融合方法建立网络模型CMFBG。其中文本图能够获得专家文本的特征联系。BERT语言模型可以有效提取深层特征,BERT语言模型和GCN并联后输入Bi-GCN(Bi-directional graph convolutional network)提取特征的方式既能发现深层特征,同时也可以增强根源特征,进而提高特征提取的准确率。CMFBG模型结构如图1所示。

2.1 GCN提取特征

通过构建专家多特征关系图,并将图融合构成特征混合图,利用GCN对结构图进行学习,获取专家与专家、专家与领域知识的内在联系。在构建异构文本图前,使用K-means对使用专家信息进行知识领域聚类后,计算每个样本距离中心点的距离 L 。使用点互信息(Pointwise mutual information, PMI)和TF-IDF(Term frequency-inverse document frequency)以 L 为辅助量计算节点之间边权重,PMI的定义为

$$\text{PMI}(i, j) = \log_2 \left(\frac{p(i, j)}{p(i)p(j)} \right) \quad (1)$$

两节点 i 和 j 之间边的权重定义为

$$A_{ij} = \begin{cases} \text{PMI}(i, j) = \log_2 \left(\frac{p(i, j)}{p(i)p(j)} \right) & i \neq j \\ \text{TF-IDF}(i, j) & \\ 1 & i = j \\ 0 & \end{cases} \quad (2)$$

式中 i 为专家,在TextGCN^[15]中,单位矩阵 $X = I_{n_{\text{ex}} + n_{\text{oth}}}$ 被用作初始节点特征,其中 n_{ex} 是专家节点数, n_{oth} 是其余节点数,其余节点包括领域知识节点、文献标签节点等。使用BERT模型来获取全部数据特征嵌入GCN,并将其作为其余节点的输入表示。其余节点嵌入表示为 $X_{\text{oth}} \in \mathbb{R}^{n_{\text{oth}} \times d}$,其中 d 是嵌入维度,则初始节点特征矩阵为

$$X = \begin{pmatrix} X_{\text{oth}} \\ 0 \end{pmatrix}_{(n_{\text{ex}} + n_{\text{oth}}) \times d} \quad (3)$$

之后将 X 输入GCN模型进行特征提取。

GCN模型在图结构数据上执行卷积操作,将节点邻域属性嵌入节点自身属性向量。GCN模型如图2所示。图中 X_1, X_2, X_3, X_4 为不同节点的特征矩阵。 X_1 通过节点之间的连接关系,卷积形成新的特征矩阵 Y_1 ,特征矩阵 Y_2, Y_3, Y_4 形成过程同 Y_1, Z_1 和 Z_4 分别为特征矩阵 Y_1 和 Y_4 的标签。考虑图的形式为 $G = (P, E)$,其中 P 和 E 分别是节点和边的集合。对于GCN的单个卷积层,表达式为

$$H = \tilde{A}XW \quad (4)$$

式中: $X \in \mathbb{R}^{n \times m}$ 为 n 个节点、 m 个特征维度的输入矩阵; \tilde{A} 为归一化邻接矩阵; $W \in \mathbb{R}^{n \times m}$ 为权值矩阵。在数据中,需要对专家标签、领域标签以及文献标签等单词进行卷积处理,因此GCN模型是根据单词而构建而不是根据整条数据构建。对于单条文本数据,定义的卷积层为

$$h = (\tilde{A}x^T)^T W = x\tilde{A}W \quad (5)$$

式中 $\tilde{A}^T = \tilde{A}$ 表示标签文本图, $x\tilde{A}$ 提取与输入特征 x 相关的文本图部分, W 为单条数据的隐藏状态向量的权重,维数为 $|V| \times h$ 。

首先,GCN对节点进行数据建模,将矩阵 X 输入GCN模型

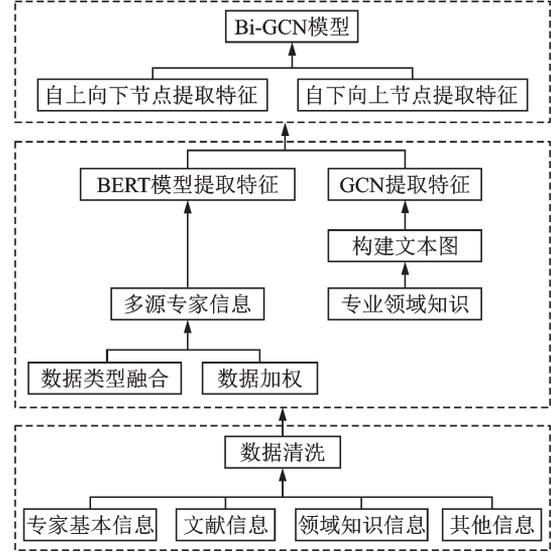


图1 整体架构

Fig.1 Overall structure

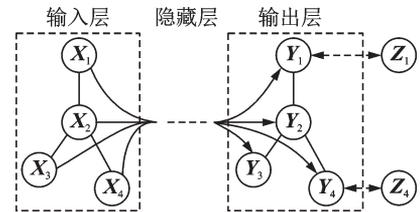


图2 GCN模型

Fig.2 GCN model

后,第*i*层的输出特征矩阵为

$$L^{(i)} = \rho(\tilde{A}L^{i-1}W^{(i)}) \tag{6}$$

式中: ρ 为激活函数; $W^{(i)} \in \mathbb{R}^{d_{i-1} \times d_i}$ 是该层的权重矩阵; $L^{(0)} = X$ 模型的输入特征矩阵。其次,具体节点更新计算为

$$x_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \frac{1}{c_{ij}} x_j^{(l)} w^{(l)} + b^{(l)}\right) \tag{7}$$

式中: $x_i^{(l)}$ 表示节点*i*在第*l*层的特征; N_i 表示节点*i*的所有邻居节点,包含节点*i*自身; c_{ij} 是一个可学习或者可提前设定的正则化常量; $w^{(l)}$ 为权值参数; $b^{(l)}$ 为偏置参数; σ 为激活函数。最后,模型将产生*N*个节点的表示矩阵 $\{X_1, X_2, \dots, X_M, X_{M+1}, \dots, X_N\}$, $i = 1, 2, \dots, N$ 。其中 $\{X_1, X_2, \dots, X_M\}$ 为专家节点, $\{X_{M+1}, X_{M+2}, \dots, X_N\}$ 为领域和文献标签节点, X_i 为专家或者领域标签的表示向量。

2.2 BERT 嵌入 GCN

专家数据经过数据处理后生成较短文本,信息量有限,且与其他维度的特征有复杂的关联性,表示向量可以为权重计算和特征关联提高精确度。BERT模型使用大量语料库通过屏蔽子词和再预测进行训练,得到先进行预训练再微调的特别训练方式,同时BERT模型使用Transformers对单词、句子以及基于上下文联系提取文本特征。使用BERT编码器^[38]来抽取给定单文本的特征,在生成词向量时可以充分考虑其长距离依赖信息,由Transformer组成BERT,其中每个Transformer用 $\text{trans}(x)$ 表示,BERT模型为

$$h_0^w = IW_s + W_p \tag{8}$$

$$h_l^w = \text{trans}(h_{l-1}) \quad l \in [1, n] \tag{9}$$

式中: I 为单词索引; W_s 为子单词的嵌入矩阵; W_p 为输入文本的位置嵌入矩阵; h_l^w 是第*l*层句子的上下文隐含状态表示,输入文本的全局上下文表示为 $ash^w = [h_1^w, h_2^w, \dots, h_n^w]$ 。将专家文本改变格式后输入BERT模型进行编码,得到向量序列 $S = \{T_1, T_2, \dots, T_n\}$,其中*n*为字向量序列长度,此向量作为嵌入GCN模型的输入。BERT模型如图3所示。

BERT模型嵌入其余模型时,现有的研究工作中主要有3类:包含单词位置信息的单词嵌入模块;采用多层多头自主意叠加的变压器模块;利用输出句子嵌入全连接层。图特征嵌入模型如图4所示。BERT模型中,attention机制为

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

式中: Q, K, V 为输入的字向量矩阵; d_k 为*Q*矩阵和*K*矩阵的维数。利用注意力机制的计算,可以获得每个单词的加权向量表示,将加权向量附着在特征计算时,提高特征之间距离的准确率以及特征利用率。在BERT中,将式(3)中得到的特征图和特征文本嵌入BERT的编码器,这样不仅获得了基于图结构数据的特征,同时也获得了文本中的特征以及这些特征更丰富的背景信息。将式(6)计算得到的结果后,在自注意力编码器逐层交互后,将局部嵌入和全局嵌入完全融合,此时嵌入定义为

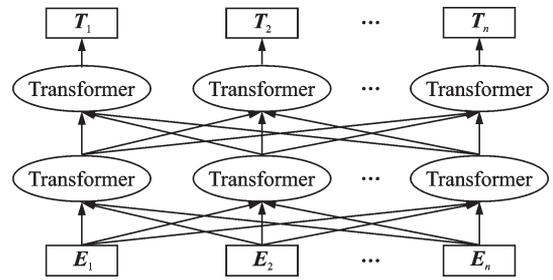


图3 BERT语言模型
Fig.3 BERT language model

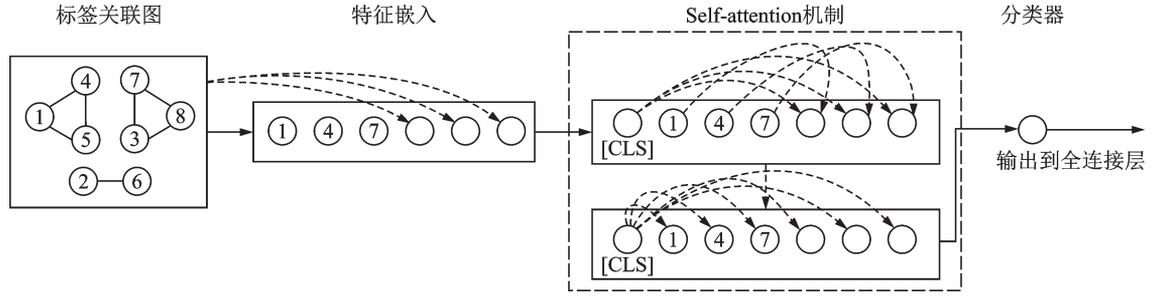


图4 图特征嵌入模型

Fig.4 Graph feature embedding model

$$G_{\text{embedding}} = \text{ReLU}\left(X_{ev} \tilde{A}_{vw} W_{vh}\right) W_v \quad (11)$$

式中： W 为单个文本图隐藏状态向量的权重，其维数与每个词嵌入的维数是相同的； e 为特征嵌入的维数； v 为特征词总数量。通过对特征的嵌入以及连接全局，得到组合特征 $\text{feature}^s = [h^w, h^c]$ ，其中 $h^m = [h_1^m, h_2^m, \dots, h_n^m]$ 表示为特征嵌入后的矩阵； h^w 是BERT提取的文本特征， h^c 是基于图结构数据提取的特征。并将组合特征输入到BiLSTM，获取到每个文本的多粒度为

$$h_i^m = \left[\overrightarrow{\text{LSTM}}(h_i^s); \overleftarrow{\text{LSTM}}(h_i^s) \right] \quad i \in [1, n] \quad (12)$$

2.3 双向图卷积注意力增强节点表示

在双向图卷积网络模型中，整体由T-GCN(自顶向下的图卷积网络)和B-GCN(自底向上的图卷积网络)两部分组成。其中，T-GCN收集信息特征是由根节点向外扩展；B-GCN收集信息特征是由子节点向根节点聚拢，其结构图如图5所示。形成由专家到领域知识和文献到领域知识再到专家的两种提取特征方式，最后使用全连接层将两种图卷积的输出矩阵连接，并得到结果。为增强领域知识对特征抽取的影响，在图卷积神经网络层结构中将领域知识节点的特征与该节点层的隐藏特征拼接，得到两部分相同的特征矩阵 X ，但邻接矩阵不同。其中T-GCN邻接矩阵为 $A^{\text{TD}} = A'$ ，BU-GCN邻接矩阵为 $A^{\text{BU}} = A'^T$ 。

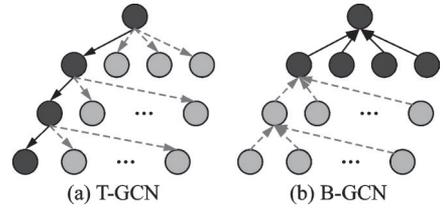


图5 T-GCN和B-GCN结构

Fig.5 Structure diagrams of T-GCN and B-GCN

在双向图卷积网络中，使用T-GCN和B-GCN对图结构卷积、提取特征后，分别得到t-down和b-top两种传播特征。T-GCN中两层的隐藏层定义为

$$H_1^{\text{TD}} = \sigma\left(\hat{A}^{\text{TD}} X W_0^{\text{TD}}\right) \quad (13)$$

$$H_2^{\text{TD}} = \sigma\left(\hat{A}^{\text{TD}} H_1^{\text{TD}} W_1^{\text{TD}}\right) \quad (14)$$

式中： $H_1^{\text{TD}} \in \mathbf{R}^{n \times v_1}$ 和 $H_2^{\text{TD}} \in \mathbf{R}^{n \times v_2}$ 是T-GCN两层的隐藏层特征表示； $W_0^{\text{TD}} \in \mathbf{R}^{d \times v_1}$ 和 $W_1^{\text{TD}} \in \mathbf{R}^{v_1 \times v_2}$ 是TD-GCN的卷积核参数矩阵，同样依据式(13)和(14)可以得到B-GCN的隐藏层特征 H_1^{BU} 和 H_2^{BU} 。在除了T-GCN和B-GCN的隐藏层特征之外，还使用了根节点特征的增强，对于T-GCN的第 k 层GCL，将每个节点的隐藏层特征向量和第 $(k-1)$ 层GCL的根节点的隐藏层特征向量拼接起来，构成一个新的特征矩阵为

$$\tilde{H}_k^{\text{TD}} = \text{concat}\left(H_k^{\text{TD}}, (H_{k-1}^{\text{TD}})^{\text{root}}\right) \quad (15)$$

最后将 T-GCN 和 B-GCN 的输出特征融合在一起,再输入全连接层和 softmax 层中,得到预测向量,其结构图如图 6 所示。

3 实验

3.1 实验数据

在实验中,本文使用了公共数据集 MR-1 和自建数据集 DING。MR-1 是对现有 MovieLens-100k 和 MovieLens-20m 两个数据集的整合,其中包含 65 223 条数据,根据电影类型,共分为 10 个类。自建数据集 DING 包含 108 715 条专家基本信息,依据研究方向分为 8 类,使用专业领域知识词语特征深入扩展。将数据划分训练集、验证集和测试集为 8:1:1 的比例进行实验。数据集节点、连边以及标签数量如表 1 所示。

3.2 评价指标

为验证 CMFBG 的精准度,本文采用精确率 Pre、召回率 Recall 和 F_1 值和 AUC 值来评估模型在分类和推荐性能,精确率 Pre 表示预测为推荐成功的项目占全部预测应该推荐项目的比率,定义方式如式(16)所示。召回率 Recall 表示推荐成功的项目占有所有应该推荐项目的比率,定义方式如式(17)所示。准确率和召回率不是同时增大或减小,为了平衡两个指标,使用平均数 F_1 对两者进行评判,定义方式如式(18)所示。准确率 ACC 和接收者操作特征曲线 ROC 下方的面积 AUC 来评价模型效果,其中 AUC 定义方式如式(19)所示。

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (17)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (18)$$

$$\text{AUC} = \frac{\sum_i^n (P_{\text{pos}} > P_{\text{neg}}) + 0.5 \times \sum_i^n (P_{\text{pos}} = P_{\text{neg}})}{M \times N} \quad (19)$$

式中:TP 表示预测为正,实际也为正的样本数;FP 表示预测为正,实际为负的样本数;TN 表示预测为负,实际为负的样本数;FN 表示预测为负,实际为正的样本数; M 、 N 分别表示正样本和负样本的数量; P_{pos} 表示正样本的分数; P_{neg} 表示负样本的分数。

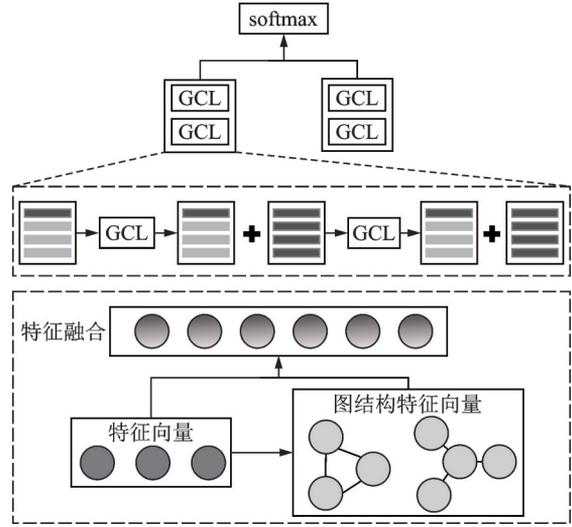


图 6 CMFBG 结构图

Fig.6 CMFBG structure diagram

表 1 实验数据

Table 1 Experimental data

数据集	Node	Edge	Label
MR-1	2 708	5 429	10
DING	100 312	333 983	8

3.3 实验分析

实验使用百分比表示本文模型的分类结果,通过对两种数据集提取领域关键词特征构建文本图,以加入注意力机制的文本分类算法模型为基线进行测试,CMFBG通过对文本和图结构两种特征提取和嵌入,生成矩阵作为神经网络模型的输入,所有模型若训练过程中效果不再提升,则提前结束训练。本文提出的CMFBG对比其他模型在专家数据集上带来的增益效果,实验结果如表2所示。相比于5种特征处理的参比模型,CMFBG模型在特征提取和嵌入后 F_1 值达到91.29%,高于其他深度学习模型。

表2 对比实验结果

Table 2 Compare experimental results

模型	MR-1				DING				%
	Pre	Recall	F_1	AUC	Pre	Recall	F_1	AUC	
LSTM	85.58	86.31	85.94	76.82	82.85	83.62	83.23	80.64	
BiLSTM	86.92	86.52	86.72	82.74	84.12	86.74	85.41	84.42	
TextGCN	80.36	79.94	80.15	79.39	81.54	80.38	80.96	83.15	
BERTGCN	88.28	87.62	87.95	86.44	86.95	85.34	86.14	85.02	
BERTGAT	88.07	89.31	88.69	86.58	89.45	88.47	88.96	86.92	
CMFBG	90.73	89.45	90.09	88.73	91.71	90.87	91.29	89.85	

为了直观表示各模型优劣,挑选加入注意力机制的深度学习模型使用验证集进行迭代测试,记录结果如图7所示。同时挑选不同程度图结构嵌入的5种模型,分别为能够单维度提取特征的LSTM模型,双向提取上下文特征的BiLSTM模型,提取图结构和节点表示后更新特征的TextGCN模型,提取图结构特征后融合文本特征的BERTGCN和BERTGAN模型,以及多特征融合的CMFBG模型。由图7可以看出,随着迭代次数的增长,CMFBG模型的准确率相较于其他模型取得较好的效果。损失值的大小表明模型训练中的收敛情况。CMFBG模型分别使用BERT模型和图神经网络提取文本和图结构特征后,嵌入特征同时使用双向图神经网络提取特征更新节点表示生成矩阵再次传入神经网络不仅能够获得文本特征和图结构特征,还能有挖掘单维度特征下的深层次关联特征,因此CMFBG模型可以在文本特征与图结构特征嵌入时达到良好效果。

为准确地表达特征嵌入对模型效果的影响,使用不同嵌入比例进行测试,记录结果如表3所示。 m 代表嵌入时图结构特征所占比例。随着嵌入特征比例的上升,准确率在 m 为0.7时,准确率达到最高,同样使模型 m 值取0.7时,达到了91.71%效果。

表3 参数 m 对准确率的影响

Table 3 Influence of parameter m on accuracy

m	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Pre/%	84.21	85.04	87.93	88.90	89.55	90.39	90.86	91.71	89.94	89.81	88.24

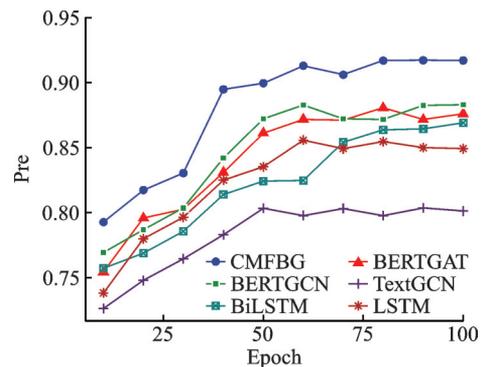


图7 准确率变化曲线

Fig.7 Accuracy rate change curves

3.4 消融实验

为分析 CMFBG 中不同模块在模型中的作用,设计了不同模块的消融实验,结果如表 4 所示。CMFBG-T 表示模型在双向提取图结构特征算法中,仅使用自上而下的算法提取图结构特征;CMFBG-B 表示模型仅使用自下而上的算法提取图结构特征;CMFBG-N 表示未构建图结构,仅对文本数据提取特征向量。CMFBG-R 表示在提取图结构特征增强节点特征时,去除图结构中根节点连接。由表 4 分析可知,与完整的 CMFBG 相比,去除双向提取图结构特征单元或者仅使用单向提取图结构特征后的模型性能都有所下降,验证了 CMFBG 中双向提取图结构特征增强源特征表示的有效性。其中去除双向提取图结构特征单元,无法在图结构中挖掘不同节点之间的隐藏特征。仅使用单向提取图结构特征无法丰富节点的隐藏特征,同时无法平衡多个隐藏特征向量对节点特征的增强。去除图结构中根节点连接模型,无法获得根节点其余扩展边的直接特征向量,失去根节点特征与增强节点特征之间的权重向量平衡。仅使用对文本数据提起特征向量时,模型性能下降显著,验证了图卷积提取特征方法的优越性。综合 CMFBG-T、CMFBG-B、CMFBG-N 和 CMFBG-R 实验结果可知,无图结构特征和单向提取图结构增强源特征表示性能均低于双向提取图结构增强源特征的模型,验证了基于图结构双向拓展并提取图结构特征增强源特征表示可以有效提高模型分类和推荐性能。

表 4 消融实验结果

Table 4 Ablation experimental results

%

模型	DING			
	Pre	Recall	F_1	AUC
CMFBG	91.71	90.87	91.29	89.85
CMFBG-T	90.29	89.64	89.96	88.21
CMFBG-B	89.84	89.29	89.56	87.89
CMFBG-R	91.14	90.05	90.59	89.22
CMFBG-N	89.06	88.28	88.66	86.74

4 结束语

针对领域专家推荐提出融合多特征和双向图分类的专家推荐方法 CMFBG,在分析多维度专家信息时,建立多特征图结构,并对图结构双向增强源特征表示。该方法通过分析处理领域专家信息建立多维度特征,采用 K-means 对多特征标签聚类,确定知识领域的类别;接着基于专家多源信息构建类别内文本图,分别使用 BERT 和 GCN 模型对文本和图结构提取特征,完成特征融合嵌入;最后使用双向图卷积神经网络对图结构进行特征增强,完成图节点分类。使用 MR-1 和 DING 数据集实验,结果表明 CMFBG 方法能有效提高图节点分类的准确率和推荐效果。在专家信息多特征构建文本图中不同的权重选择连接,以及多特征标签抽取后的细化方法对后续多特征嵌入融合存在影响。因此,对专家信息建立多特征标签后,如何利用多特征间的关系进行细化和构图,提取实体之间更深层次的隐藏关系及特征成为未来的研究重点。

参考文献:

- [1] LIN S, HONG W, WANG D, et al. A survey on expert finding techniques[J]. Journal of Intelligent Information Systems, 2017, 49(2): 255-279.
- [2] NIKZAD-KHASKHAKI N, BALAFAR M A, FEIZI-DERAKHSHI M R. The state-of-the-art in expert recommendation systems[J]. Engineering Applications of Artificial Intelligence, 2019, 82: 126-147.

- [3] SUN J, MA J, CHENG X, et al. Finding an expert: A model recommendation system[C]//Proceedings of the 34th International Conference on Information Systems (ICIS 2013). Milan: [s.n.], 2013: 1660-1669.
- [4] 张雯,张仰森,周炜翔,等. 领域科研项目评审专家推荐算法[J]. 计算机工程与设计, 2021, 42(6): 1787-1794.
ZHANG Wen, ZHANG Yangsen, ZHOU Weixiang, et al. Algorithm recommended by experts for evaluation of scientific research projects in the field[J]. *Computer Engineering and Design*, 2021, 42(6): 1787-1794.
- [5] 吕晓琦,纪科,陈贞翔,等. 结合注意力与循环神经网络的专家推荐算法[J]. 计算机科学与探索, 2022, 16(4): 877-887.
LV Xiaoqi, JI Ke, CHEN Zhenxiang, et al. Expert recommendation algorithm by combining attention and recurrent neural network[J]. *Journal of Frontiers of Computer Science and Technology*, 2022, 16(4): 877-887.
- [6] RINKE E M, DOBBRICK T, LÖB C, et al. Expert-informed topic models for document set discovery[J]. *Communication Methods and Measures*, 2022, 16(1): 39-58.
- [7] ZHANG S, YAO L, SUN A, et al. Deep learning based recommender system: A survey and new perspectives[J]. *ACM Computing Surveys*, 2019, 52(1): 1-38.
- [8] PAN N, YAO W, LI X. Friends recommendation based on KBERT-CNN text classification model[C]//Proceedings of 2021 International Joint Conference on Neural Networks(IJCNN). Shenzhen, China: IEEE, 2021: 1-6.
- [9] JIANG S, LI K, DA XU R Y. Magnitude bounded matrix factorisation for recommender systems[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2020, 34(4): 1856-1869.
- [10] LI G, QIU L, YU C, et al. IPTV channel zapping recommendation with attention mechanism[J]. *IEEE Transactions on Multimedia*, 2020, 23: 538-549.
- [11] ABOLGHAEMI R, ENGELSTAD P, HERRERA-VIEDMA E, et al. A personality-aware group recommendation system based on pairwise preferences[J]. *Information Sciences*, 2022, 595: 1-17.
- [12] LEE S, KIM D. Deep learning based recommender system using cross convolutional filters[J]. *Information Sciences*, 2022, 592: 112-122.
- [13] VIVEKANANDAN K, PRAVEENA N. Hybrid convolutional neural network (CNN) and long-short term memory (LSTM) based deep learning model for detecting shilling attack in the social-aware network[J]. *Journal of Ambient Intelligence and Humanized Computing*, 2021, 12(1): 1197-1210.
- [14] JIANG B, ZHANG Z, LIN D, et al. Semi-supervised learning with graph learning-convolutional networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 11313-11320.
- [15] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019, 33(1): 7370-7377.
- [16] FAN W, MA Y, LI Q, et al. A graph neural network framework for social recommendations[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2020, 34(5): 2033-2047.
- [17] WANG Y, WANG H, JIN H, et al. Exploring graph capsual network for graph classification[J]. *Information Sciences*, 2021, 581: 932-950.
- [18] NIEPERT M, AHMED M, KUTZKOV K. Learning convolutional neural networks for graphs[C]//Proceedings of International Conference on Machine Learning. New York, USA: ICML, 2016(48): 2014-2023.
- [19] DUVENAUD D K, MACLAURIN D, IPARRAGUIRRE J, et al. Convolutional networks on graphs for learning molecular fingerprints[J]. *Advances in Neural Information Processing Systems*, 2015, 28: 2224-2232.
- [20] MONTI F, BRONSTEIN M, BRESSON X. Geometric matrix completion with recurrent multi-graph neural networks[J]. *Advances in Neural Information Processing Systems*, 2017, 30: 3700-3710.
- [21] GUO J, ZHOU Y, ZHANG P, et al. Trust-aware recommendation based on heterogeneous multi-relational graphs fusion[J]. *Information Fusion*, 2021, 74: 87-95.
- [22] MEI T, SUN T, CHEN R, et al. GCNNIRec: Graph convolutional networks with neighbor complex interactions for recommendation[C]//Proceedings of Asia-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data. Guangzhou, China: APWeb-WAIM, 2021: 338-347.
- [23] MA M, NA S, WANG H, et al. The graph-based behavior-aware recommendation for interactive news[J]. *Applied Intelligence*, 2022, 52(2): 1913-1929.
- [24] GWADABE T R, LIU Y. Improving graph neural network for session-based recommendation system via non-sequential

- interactions[J]. Neurocomputing, 2022, 468: 111-122.
- [25] ACHARY N S, PATRA B K. Graph based hybrid approach for long-tail item recommendation in collaborative filtering[C]// Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD). [S.l.]: ACM, 2021: 426.
- [26] GAO X, FENG F, HUANG H, et al. Food recommendation with graph convolutional network[J]. Information Sciences, 2022, 584: 170-183.
- [27] LIAO J, ZHOU W, LUO F, et al. SocialLGN: Light graph convolution network for social recommendation[J]. Information Sciences, 2022, 589: 595-607.
- [28] SHANG J, MA T, XIAO C, et al. Pre-training of graph augmented transformers for medication recommendation[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019: 5953-5959.
- [29] ZHENG Y, GAO C, CHEN L, et al. DGCN: Diversified recommendation with graph convolutional networks[C]// Proceedings of the Web Conference. Ljubljana: Slovenia, 2021: 401-412.
- [30] JEONG C, JANG S, PARK E, et al. A context-aware citation recommendation model with BERT and graph convolutional networks[J]. Scientometrics, 2020, 124(3): 1907-1922.
- [31] HARDOON D R, SZEDMAK S, SHAWE-TAYLOR J. Canonical correlation analysis: an overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664.
- [32] XU J, CAI Y, WU X, et al. Incorporating context-relevant concepts into convolutional neural networks for short text classification[J]. Neurocomputing, 2020, 386: 42-53.
- [33] TAN Y, SUN D, SHI Y, et al. Bi-directional mapping for multi-label learning of label-specific features[J]. Applied Intelligence, 2022, 52(7): 8147-8166.
- [34] DONG H, SONG K, HE Y, et al. PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection[J]. IEEE Transactions on Industrial Informatics, 2019, 16(12): 7448-7458.
- [35] ABDI A, SHAMSUDDIN S M, HASAN S, et al. Deep learning-based sentiment classification of evaluative text based on multi-feature fusion[J]. Information Processing & Management, 2019, 56(4): 1245-1259.
- [36] HUANG X, CHEN B, XIAO L, et al. Label-aware document representation via hybrid attention for extreme multi-label text classification[J]. Neural Processing Letters, 2022, 54(5): 3601-3617.
- [37] ZHANG H, NIE R, LIN M, et al. A deep learning based algorithm with multi-level feature extraction for automatic modulation recognition[J]. Wireless Networks, 2021, 27(7): 4665-4676.
- [38] NIU W C, CHEN Q, ZHANG W, et al. GCN2-NAA: Two-stage graph convolutional networks with node-aware attention for joint entity and relation extraction[C]// Proceedings of 2021 13th International Conference on Machine Learning and Computing. Shenzhen, China: ICMLC, 2021: 542-549.

作者简介:



丁婧娴(1998-),女,硕士研究生,研究方向:数据挖掘、推荐系统等,E-mail: 212006560408@hyit.edu.cn。



李翔(1980-),通信作者,男,教授,研究方向:大数据、区块链、推荐系统、知识图谱,E-mail: lixiang@hyit.edu.cn。



孙纪舟(1985-),男,讲师,研究方向:海量数据计算、分布式系统、数据库技术。



周泓(1980-),女,副教授,研究方向:机器学习、数据挖掘、大数据。