文本特征和图结点混合增强的图卷积网络文本分类*

杨晓奇[1]，刘伍颖[2,3]†

(1. 广东外语外贸大学 信息科学与技术学院，广东 广州 510006；2. 鲁东大学 山东省语言资源开发与应用重点实验室，山东 烟台 264025；3. 广东外语外贸大学 外国语言学及应用语言学研究中心，广东 广州 510420)

摘　要： 在BertGCN模型的基础上改进其结构，同时结合文本特征和图结点混合增强的方法，使用新的边权重计算算法BM25+构造图的边．使用R8、R52、Ohsumed和MR这4个常用的公开数据集来验证所提方法的有效性．结果表明：与BertGCN模型及其它基线模型相比，该方法在4个文本分类数据集上的准确率评价指标均有不同程度的提升．

关键词： BM25+；文本特征增强；图结点增强；预训练模型；图卷积网络；文本分类
DOI： 10.13568/j.cnki.651094.651316. 2023.07.05.0004
中图分类号： TP183　**文献标识码：** A　**文章编号：** 2096-7675(2024)01-0069-09
引文格式： 杨晓奇，刘伍颖．文本特征和图结点混合增强的图卷积网络文本分类[J]．新疆大学学报(自然科学版)(中英文)，2024，41(1)：69-77+109.
英文引文格式： YANG Xiaoqi，LIU Wuying. Hybrid augmentation of text feature and graph node for graph convolutional networks text classification[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English)，2024，41(1)：69-77+109.

Hybrid Augmentation of Text Feature and Graph Node for Graph Convolutional Networks Text Classification

YANG Xiaoqi[1], LIU Wuying[2,3]

(*1. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou Guangdong 510006, China; 2. Shandong Key Laboratory of Language Resources Development and Application, Ludong University, Yantai Shandong 264025, China; 3. Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou Guangdong 510420, China*)

Abstract： The work will improve the structure on the basis of the BertGCN model, not only using a new algorithm to construct the edges of the graph, but also combining a hybrid enhancement of text features and graph nodes. The method not only has some optimization in the edge structure, but also makes fuller use of the extended semantic information of the text in the form of text feature enhancement and graph-enhanced nodes, while retaining the original text features. Four public datasets, R8, R52, Ohsumed and MR which are commonly used, are used to verify the effectiveness of this method. The experimental results show that compared with the BertGCN model and other baselines, the accuracy evaluation metric of the method on the four text classification data sets has been improved to varying degrees.

Key words： BM25+; text feature enhancement; graph node enhancement; pre-training model; graph convolution networks; text classification

0　引 言

文本分类是自然语言处理中的一项重要任务，它被广泛应用于各大场景，包括舆情监测、新闻分类、信息检索、情感分析[1−2]以及事件预测等，旨在根据文本的内容和语义将其自动分类到预定义的类别中，而实现这

* **收稿日期：** 2023-07-05
基金项目： 教育部人文社会科学研究规划基金“后深度学习时代低资源语言机器翻译理论与实践研究”（20YJAZH069），“自由贸易关键小语种语言资源建设理论与实践研究”（20YJC740062）；上海市哲学社会科学规划基金“‘一带一路’关键小语种千万级语言资源建设及精准机器翻译研究”（2019BYY028）．
作者简介： 杨晓奇（1998—），男，硕士生，从事文本分类的研究，E-mail：20211050027@gdufs.edu.cn.
† **通讯作者：** 刘伍颖（1980—），男，博士，教授，硕士生导师，主要从事计算语言学和自然语言处理的研究，E-mail：wyliu@ldu.edu.cn.

一过程的算法称为分类器，在这个过程中，特征提取是一个重要步骤，传统的特征提取方法主要基于词袋模型、TF-IDF和N-gram模型等，它们简单直观、易于实现，但在复杂的场景效果有限. 随着深度学习的兴起，深度学习模型在文本分类任务中取得了显著效果. 常用的深度学习模型包括卷积神经网络（CNN）[3]、循环神经网络（RNN）[4−6]等，能够自动学习文本中的语义和上下文信息. 然而它们主要关注捕捉局部连续词序列中的语义和句法信息，在一些文本中，重要的语义关系可能存在于非连续的词语之间或者跨度较长的片段中，这些传统模型无法很好地捕捉到全局的词语共现信息. 为了解决这一问题，图神经网络（GNN）[7]出现了，这是一种强大的深度学习模型，具有处理非欧氏空间数据的能力，其出现弥补了传统深度学习模型只能处理欧氏空间数据的不足. 图神经网络通过考虑数据中的复杂关系结构和全局结构信息，有效地处理具有非线性和非连续关系的数据，这不仅在文本分类任务中有着重要的研究价值，而且在知识图谱、社会网络等领域都有着广泛的应用场景.

近年来，预训练语言模型（如BERT[8]）在各种自然语言处理任务中证明了它们的巨大优势，越来越多的任务开始在原有的工作基础上融合预训练语言模型，并都获得了更好的效果. 但是单一的语义特征表示在一定程度上会限制模型对它的理解，当遇到一些特征更为丰富的文本时，模型有可能无法准确理解它们，此时，使用合适的数据增强方法[9−10]在保持标签类别不变的前提下，按照一定的策略改变文本的内容，不仅可以达到文本特征增强的效果，同时可以缓解低资源场景的局限问题.

本文提出了一种文本特征和图结点混合增强的图卷积网络文本分类方法（Maximal-features-enhancement GCN, MfeGCN）. 该方法在BertGCN[11]的基础上加以改进，一方面改进了单词-文档的边构造方式，使用BM25+算法来计算单词与文档之间的边权重，同时使用一种基于最高得分的同义词替换策略，在使用预训练模型（如BERT）进行微调时，丰富了文档结点的语义表示，在图卷积神经网络中植入使用上述增强策略增强后的文本作为图的增强结点，然后进行构图，图中的结点包括原文档结点、增强文档结点、原单词结点和增强单词结点，旨在结合文本特征方面和图结构方面加以改进，从而优化模型的分类性能.

在4个真实公开数据集进行实验，结果表明MfeGCN模型与基础BertGCN模型以及其它基线模型相比，其表现效果更为优异.

1 文本分类方法综述

文本分类方法可以划分为传统机器学习方法、深度学习方法等. 传统的方法有K-近邻、朴素贝叶斯等. 随着深度学习技术的发展，越来越多的深度学习模型（如CNN和RNN）被应用到文本分类中. 近年来，图神经网络被应用于分类任务并取得了不错的结果.

1.1 传统机器学习文本分类

传统的文本分类方法主要由特征工程和浅层分类算法构成，其中的代表性分类方法有朴素贝叶斯算法、K-最近邻算法、决策树算法和支持向量机等.

朴素贝叶斯（Naive Bayes, NB）[12]. 朴素贝叶斯算法是一种基于概率统计和特征条件独立性假设的分类算法，其核心思想是通过计算后验概率来进行分类.

K-最近邻算法（K-Nearest Neighbors, KNN）[13]. K-最近邻算法是一种基于最近邻样本的分类算法，其核心思想是通过找到与待预测样本最接近的K个训练样本来进行分类.

决策树（Decision Tree, DT）[14]. 决策树算法是一种基于树结构的机器学习算法，用于解决分类和回归问题. 其核心思想是通过构建一棵树来表示特征之间的关系，并根据特征的值进行分割和预测.

支持向量机（Support Vector Machine, SVM）[15−16]. 支持向量机使用向量空间模型来表示文档，将文档嵌入映射到高维空间，以此将文档表示为高维向量，并将每个文档抽象为高维空间中的一个点，其核心思想是通过构建一个最优的超平面来将不同类别的样本分开.

1.2 深度学习文本分类

相比传统的机器学习文本分类方法，深度学习方法避免了人工设计规则和特征，并能够自动挖掘文本的语义表示，形成一个通过学习直接将特征映射为目标输出的非线性转换过程，将特征工程集成到模型拟合的过程中. 主要的代表模型有CNN、RNN、预训练模型和GNN等.

CNN最早由Yann LeCun等于1989年提出，其基本原理是通过多层卷积和池化操作来提取输入数据的特征，并通过全连接层进行分类或回归，卷积操作利用卷积核在输入数据上进行滑动，提取不同位置的局部特征. 后来，Kim于2014年提出了TextCNN[3]，将卷积神经网络引入文本分类领域，并取得了很好的效果. TextCNN的优势在于其简单而有效的结构，能够捕捉文本中的局部特征，并且具有较好的泛化能力.

RNN是一种具有记忆功能的神经网络，其基本原理是通过在每个时间步上输入当前特征和上一个时间步的隐藏状态，来建模序列中的依赖关系，它使用相同的参数在每个时间步上进行计算，以捕捉特征的时间性质. RNN的隐藏状态可以理解为对先前信息的编码表示，在每个时间步上更新并传递到下一个时间步，因此该网络具有记忆的功能.

预训练模型是指在大规模无标签数据上进行预训练的深度神经网络模型，旨在习得数据中的统计规律和语义信息，这些模型的参数可以作为后续任务的初始参数，通过微调的方式在下游任务的有标签数据上进行训练，从而加速模型收敛、提升性能. 代表模型有BERT[8]，其通过在大规模无标签数据上进行预训练，学习了丰富的句子表示，包含了上下文相关的语义信息，在自然语言处理（NLP）任务中取得了重大突破，并成为许多NLP任务的基准模型.

近几年来，GNN[7]受到了各个研究领域的广泛关注，GNN模型基于消息传播机制，通过在图结构上进行信息传递和聚合来学习结点的表示，利用邻居结点信息更新自身结点的表示，并通过多轮迭代逐步扩展和融合全局图信息，直至达到某一个稳定状态. 而GCN[17]是一种使用了卷积操作的GNN，通过在图上进行卷积操作来更新结点的表示，利用邻居结点的特征进行卷积运算，并将卷积结果作为结点的更新表示. 在工业界，GCN也有一些实际应用，例如药物分子设计[18]、疾病诊断[19]、交通流预测[20]和推荐系统[21]等，在自然语言处理领域，实体关系抽取、文本生成、文本分类和机器翻译等方向也都取得了不错的效果.

2　方 法

本节所述MfeGCN模型基于BertGCN模型加以改进，包括构建图中边权重的算法和极大特征增强方法以及整个架构的一些步骤和细节.

2.1　BM25-PLUS（BM25+）

BM25算法是信息检索领域一种用于对给定的查询项和若干个相关的文档进行相关性计算后，根据每个文档和查询项之间的相关性得分进行排序的算法. 但当遇到一些过长的文档时，BM25算法会面临对该超长文档过度惩罚的问题[22]，于是Lyu等提出了BM25+算法[22]，该算法在原有BM25算法的基础上，为每一个查询项中出现在文本中的特征项相关性得分设置一个下界δ，此时，即使一个文档特别长，搜索项都至少贡献了一个正常数相关性得分，本文将使用该算法来计算单词-文档之间的边权重，具体的得分计算方式为：

$$BM25+(a,b)=\sum_{i} IDF(a_i)\cdot R(a_i,b) \tag{1}$$

式中：$IDF(a_i)$可以用来表示特征项a_i的权重，N表示文档集合中的所有文档数量，$n(a_i)$则表示包含特征词a_i的文档数量，其表示为：

$$IDF(a_i)=\log\left(\frac{N+1}{n(a_i)}\right) \tag{2}$$

子项$R(a_i,b)$表示特征项a_i与文档b的相关性得分，k_1和δ均为可自由调节的协调因子，一般可取值的范围为$k_1\in[0.2,4.0]$、$\delta\in[0,1.5]$，k_1用于控制词频对文档匹配得分的影响程度，δ用于缓解对超长文档过度惩罚的问题[23]，其表示为：

$$R(a_i,b)=\left(\frac{(k_1+1)\cdot \hat{tf}(a_i,b)}{k_1+\hat{tf}(a_i,b)}+\delta\right) \tag{3}$$

$\hat{tf}(a_i,b)$见式（4），$tf(a_i,b)$表示词频，β和k_1一样，是一个可调节的协调因子，其取值范围为$\beta\in[0.1,0.9]$，L_d表示文档长度，L_{avg}表示文档集合中的平均长度，其中β可用于控制文档长度对匹配得分的影响.

$$\hat{tf}(a_i,b)=\frac{tf(a_i,b)}{1+\beta\left(\frac{L_d}{L_{avg}}-1\right)} \tag{4}$$

2.2 PPMI

正点互信息（PPMI）用于计算单词-单词之间的边权重，其思想是统计两个词语在文本中同时出现的频率，频率越高，表示这两个词语的相关性越高，这一方法将全局词共现信息很好地利用了起来，具体的计算方式为：

$$PPMI(a,b)=\max\left(\log\frac{P(a,b)}{P(a)P(b)},0\right) \tag{5}$$

$$P(a,b)=\frac{win(a,b)}{wins} \tag{6}$$

$$P(a)=\frac{win(a)}{wins} \tag{7}$$

式中：$wins$表示滑动窗口的总数，$win(a,b)$表示单词a和单词b共同出现的滑动窗口个数，$win(a)$表示出现单词a的滑动窗口个数，$P(b)$与$P(a)$同理.

2.3 MfeGCN

图卷积神经网络中，需要将数据集构建成一个图$G(V,E)$[24]，V表示结点，E表示边. 当结点a表示单词、结点b表示文档时，使用BM25+来计算边的权值；当结点a和b均表示单词时，使用PPMI计算边的权值；当结点形成自连接的边时，将边的权值表示为1；其它情况定义边的权值为0，图的边权值表示为：

$$A_{i,j}=\begin{cases} BM25+(a,b), & a\text{为单词、}b\text{为文档} \\ PPMI(a,b), & a\text{、}b\text{均为单词且}a\neq b \\ 1, & a=b \\ 0, & \text{其它} \end{cases} \tag{8}$$

经过上述算法后，文本数据集中的边权重已确立完毕，然后使用预训练模型BERT对所有文档结点进行初始化获得文档嵌入，单词结点嵌入初始化为0，并将它们作为GCN结构的结点嵌入，然后将该图结构输入一个两层的GCN中，每一层的结点特征表示为：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{9}$$

式中：l表示GCN的层数，$\tilde{\boldsymbol{A}}$表示结点加上了自连接边后的邻接矩阵，也就是在邻接矩阵的基础上加上一个单位矩阵，$\tilde{\boldsymbol{D}}$表示图的度矩阵，$\boldsymbol{H}^{(l)}$表示当前层的特征向量矩阵，$\boldsymbol{W}^{(l)}$表示第l层的可学习权重矩阵，σ是一个激活函数Relu，将这些计算因子进行计算后便可得到下一层的结点特征表示. 经过多次图传播后将GCN最后一层的隐藏层向量作为$softmax$层的输入，得到文本类别的概率分布为：

$$P_{\mathrm{GCN}}=softmax(g(\boldsymbol{X},E)) \tag{10}$$

式中：当l=0时，$\boldsymbol{X}=\boldsymbol{H}^{(0)}$表示输入层的特征向量矩阵，$E$表示结点之间的边，$g$表示GCN结构，输出结果作为文本的表示，输入$softmax$层得到所有类别的概率分布.

融合BERT和GCN模块前，先构造一个BERT分类器，使用在下游任务微调后的BERT得到的文本嵌入$\boldsymbol{X}$，将$\boldsymbol{X}$输入一个$softmax$层获得最终的文本概率分布为：

$$P_{\mathrm{BERT}}=softmax(BERT(\boldsymbol{W}\boldsymbol{X})) \tag{11}$$

获得BERT的预测概率分布与GCN的预测概率分布后，使用线性预插值法融合两部分的预测. λ是一个可调节的超参数；当λ=1时，表示模型只更新GCN部分的参数；当λ=0时，表示模型只更新BERT部分的参数；当$\lambda\in(0,1)$时，根据λ的取值，两个部分将会得到不同程度的权衡. 因此，可以通过调节λ值的大小来联合优化BERT和GCN的预测结果，以这样的方式可以加快模型的收敛速度，同时获得更优性能.

$$C=\lambda P_{\mathrm{GCN}}+(1-\lambda)P_{\mathrm{BERT}} \tag{12}$$

最后，使用交叉熵损失函数同时优化BERT和GCN（式(13)），其中：C_{true}表示真实类别，C表示模型预测的类别.

$$L=-\sum C_{\text{true}}\log C \tag{13}$$

本文在文本分类模型BertGCN的基础上，除了改进其边权重的计算方式外，同时加入了一种基于最高得分的同义词增强策略[25]，包括对BERT部分和GCN部分的特征表示增强，通过增强BERT部分的文本语义表示及在GNN中以增加增强结点的方式提高模型的分类性能. 首先将训练集样本进行清洗和去除停用词后，分词处理构造出一个训练集词表l_i，也就是训练集中每一个文档皆由l_i中的单词构成，然后统计其中每个单词出现次数N_x，并计算该词的出现次数在整个训练集中的占比，得到其分值S_x.

$$S_x=\frac{N_x}{\sum_{i=1}^{n}N_i} \tag{14}$$

式中：i表示单词的索引且$i\in(1,n)$，n表示词表大小，N_i表示在词表l_i中第i个单词出现的总次数，通过计算可以得到训练数据中每一个单词的分值.

将训练集的文档个数记为H，文档的索引设置为c，$c\in(1,H)$，w表示文档中的单词，那么训练集中每一个独立的文档可以表示为$D_c=\{w_1,w_2,\cdots,w_m\}$，$m$表示组成该文档的单词个数. 同样的，将一个基于最高得分同义词替换策略处理后的文档表示为$D_c^{'}=\{w_1^{'},w_2^{'},\cdots,w_m^{'}\}$，最终的文档表示为$D_E=D_c+D_c^{'}$.

对于每一个原文档，首先使用NLTK工具包中的WordNet词典工具获得该文档中每一个单词的同义词集合$Synonyms(w_\alpha)=\{s_1,s_2,\cdots,s_k\}$. 其中：$\alpha\in(1,m)$，$k$则表示词典工具所能获得的同义词数目，然后分别计算集合中每一个同义词的得分：

$$Score_p=\begin{cases}S_x, & \text{同义词在词表}l_i\text{中}\\ 0, & \text{同义词不在词表}l_i\text{中}\end{cases} \tag{15}$$

式中：$p\in(1,k)$，如果同义词在训练集词表l_i中出现，那么将其分值置为词表中该词的得分，若同义词不在训练集词表中出现，那么该同义词的得分置为0. 随后挑选同义词集合中得分最高的同义词替换掉原单词，其定义形式为：

$$w_k^{'}=Synonym(\max(Score)) \tag{16}$$

式中：$Score=\{Score_1,Score_2,\cdots,Score_k\}$ 表示同义词得分集合，经过最高得分同义词替换策略处理的单词进行合并后得到最终的增强文档表示$D_c^{'}$，基于最高得分的增强策略为：

1）将一个训练集样本D_c做分词处理，并将其转换为一个数组；

2）对于每一个样本数组D_c，获得数组内每一个单词w_α的同义词集合$Synonyms(w_\alpha)$；

3）计算每个单词的同义词集合中每个同义词的得分$Score_p$后得到同义词得分集合$Score$；

4）选择$Score$中得分最高对应同义词$w_\alpha^{'}$替换掉原单词w_α，如果同义词集合为空，那么保持不变；

5）将整个样本数组进行以上步骤后，形成一个新的增强样本数组$D_c^{'}$；

6）将其它的训练集样本分别进行上述操作，得到一批新的增强样本；

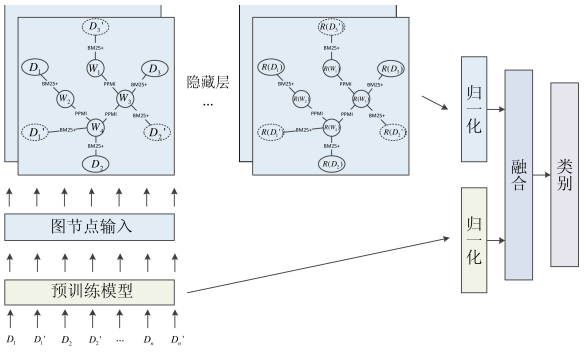

图 1　MfeGCN结构示意图

7）最后整合原样本D_c和增强样本$D_c^{'}$形成最终训练样本D_E.

本文使用的模型结构如图1所示，包括了两大类结点，分别表示文档结点和单词结点. 其中：D_c表示文档结点，$D_c^{'}$表示增强文档结点，W表示单词结点. 可以使用BERT或者ROBERTA等预训练模型将向量化的文本嵌入作为GCN的输入，经过隐藏层后形成的$R(x)$为文档或单词x的词嵌入表示. 最后，融合BERT模块和GCN模块分别经过$softmax$后的概率分布，选择概率最高的类别作为最终的预测结果.

3 实验与分析

3.1 数据集

实验使用了文本分类的4个公开数据集，分别是R8、R52、Ohsumed和MR. 与Yao等[25]使用相同的数据划分方法划分训练集和测试集，随机抽取训练集中10%的数据作为验证集.

R8：8分类的路透社数据集子集，包括5 485个训练文档和2 189个测试文档.

R52：52分类的路透社数据集子集，包括6 532个训练文档和2 568个测试文档.

Ohsumed：数据由医疗信息数据库MEDLINE中的心血管疾病摘要构成，取其中的7 400篇单标签文档，包括3 357篇训练文档和4 043篇测试文档.

MR：情感2分类的影评数据集，每一篇文档仅由一句评论构成，包括7 108篇训练文档和3 554篇测试文档.

训练模型前，先对数据集进行预处理，使用工具包NLTK去除停用词，再去除数据集中词频小于5的词，数据集中的低频词有可能是一些拼写错误的单词或其它噪声数据，去除此类词可以减少噪声对模型的影响，由于MR数据集的文档长度普遍较短，去除停用词后不再对其作删词处理.

3.2 基线模型

CNN(rand)[3]：以随机初始化的词向量作为卷积神经网络的输入.

CNN(non-static)[3]：使用预训练的词向量作为卷积神经网络的输入.

LSTM[4]：长短时记忆网络，使用随机初始化的词向量作为输入，并使用最后的隐藏层状态作为最终的文本表示.

LSTM(pretrain)[4]：使用预训练的词向量作为该网络的输入，取最后一层隐藏层向量作为文本表示.

TextGCN[26]：文本图卷积网络，将语料库中的文本数据建模成为一个包含文档结点和单词结点的异构图，实现将文本分类转换为结点分类.

SGC[27]：简单图卷积是图卷积神经网络的一种变体，通过去除非线性和在连续层之间折叠权矩阵来降低GCN的复杂性.

BERT[8]：大规模预训练模型，以文本序列作为输入，加载对应领域的预训练模型对下游任务进行微调.

BertGCN[11]：融合预训练模型BERT和图卷积网络GCN，以BERT初始化的词嵌入作为GCN中文档结点的表示，使用插值法结合BERT部分和GCN部分的预测作为最终结果.

3.3 参数设置与评价指标

为了与基础模型BertGCN作性能上的对比，以下提及的一些超参数将与BertGCN在各个数据集中的超参数设置保持一致. 首先，将微调阶段的BERT模块的学习率设置为1×10^{-4}，得到所有词嵌入后，进入联合训练阶段，设置联合训练阶段的BERT模块的学习率为1×10^{-5}，GCN模块的学习率为1×10^{-3}，dropout设置为0.5，滑动窗口的大小设置为20，GCN隐藏层的嵌入维度设置为200，使用Adam优化器[28]进行优化. 使用插值法融合BERT部分和GCN部分的概率分布时所用到的超参数λ在4个数据集中各有不同，R8数据集中λ的取值为0.7，R52数据集中λ的取值为0.3，Ohsumed数据集中λ的取值为0.9，MR数据集中λ的取值为0.4.

实验使用的评价指标为准确率（Accuracy），表示正确的预测结果个数占样本总数的百分比.

3.4 实验对比

在4个文本分类领域的公开数据集上进行对比实验，模型的性能对比如表1所示，其中加粗部分为最优结果. MfeGCN的表现在4个数据集上的测评结果优于所有的基线模型. 在Ohsumed以及MR数据集上的性能提升较为明显，分别达到了2%和1.31%，而在R8和R52数据集上的提升相对有限，我们认为有以下因素.

从边构造上来看，BM25+算法依赖文档中的词汇来计算相关性分数，如果文档的单词数较少，那么可能导致单词覆盖不充分，在计算边权重时受到一定的限制.

从语义特征上来看，得益于Ohsumed和MR数据集中较为充足的单词量，进行文档增强时所扩展的语义信息将更加广泛，而在R8和R52数据集中，增强文档得到的语义信息则相对有限.

从图结构上来看，数据集Ohsumed和MR中所构成的词汇表的单词数量远多于R8和R52数据集，这样一来，能够提供的外部知识也会相应增多. 因此，Ohsumed和MR数据集所构成的图会比R8和R52数据集所构成的图更加稠密，根据GCN的特点，图结点从其邻居结点获取到的特征也会变多.

表 1　MfeGCN和基线模型在不同数据集上的分类准确率

模型	R8	R52	Ohsumed	MR
CNN(rand)	94.02%	85.37%	43.87%	74.98%
CNN(non-static)	95.71%	87.59%	58.44%	77.75%
LSTM	93.68%	85.54%	41.13%	75.06%
LSTM(pretrain)	96.09%	90.48%	51.10%	77.33%
TextGCN	97.07%	93.56%	68.36%	76.74%
SGC	97.20%	94.00%	68.50%	75.90%
BERT	97.80%	96.40%	70.50%	85.70%
BertGCN	98.10%	96.60%	72.80%	86.00%
MfeGCN	**98.36%**	**97.08%**	**74.80%**	**87.31%**

3.5　消融实验

针对MfeGCN模型的性能效果，又进行了消融实验，首先是包括只改变构图方式而不增加增强结点的模型MfeGCN without E-nodes，还有将MfeGCN模型中构造单词-文档边权重的BM25+算法更换回原BertGCN模型中的单词-文档边权重构造方法TF-IDF，单纯进行文本特征和图结点混合增强，此处记为MfeGCN-T，以及融合了二者的MfeGCN模型本身进行消融实验，结果如表2所示.

表 2　消融实验在不同数据集上的准确率

模型	R8	R52	Ohsumed	MR
BertGCN	98.10%	96.60%	72.80%	86.00%
MfeGCN without E-nodes	98.26%	96.81%	73.63%	86.63%
MfeGCN-T	**98.40%**	96.92%	74.67%	86.94%
MfeGCN	98.36%	**97.08%**	**74.80%**	**87.31%**

由表2可知，在BertGCN的基础上改进其边构造方式，只使用BM25+算法来计算其边权重后，MfeGCN without E-nodes的性能比原方法有了一定的提升，当仅增强文本和增加增强节点而不改变其边构造方式时，MfeGCN-T的性能也比基础模型BertGCN的性能要优，而在BertGCN的基础上既使用新的边构造算法BM25+来构建单词-文档边关系权重，同时进行文本特征和图结点混合增强后，整体性能又有了一定的提升.

3.6　分析与讨论

我们认为MfeGCN取得更好的表现主要有以下原因：

1）BM25+算法是TF-IDF算法的一种改进，对于TF-IDF算法而言，当TF部分的值越大，那么整体返回的值会越大，而BM25+算法针对这一点进行改进，当其TF部分越大，那么整体返回值会趋于一个数值，同时增加了一个用于缓解对超长文档过度惩罚问题的参数. 它对于传统TF-IDF算法有一些优势，一方面，考虑了词项频率的饱和度，TF-IDF算法中词项频率的增长通常是线性的，而BM25+算法中词项频率的增长是对数级的，也就是说，对于频繁出现的词项，它们的权重增长会更加缓慢，从而避免了对高频词项的过度偏袒；另一方面，BM25+算法引入了几个可调节的参数k_1、b和δ，根据具体的应用场景进行调优，可以更好地适应不同的数据集，提供更准确的计算结果，并且BM25+算法相对TF-IDF只关注词项频率和逆文档频率而言，考虑了文档长度因素，在一些长文档场景下依旧可以表现出很好的性能. 由表2可知，仅改进构图方式，MfeGCN without E-nodes模型性能相比基础模型BertGCN有了一定的提升，说明优化构图方式对于最终的模型表现有一定的正面影响.

2）基于最高得分的同义词替换策略引入了原始训练集中不存在的单词，添加了额外的知识，分别对4个训练集进行了统计，包括统计分别组成每个训练集的原始单词个数以及引入的额外单词个数，具体指标如表3所示. 由表3可知，使用该方法后，每个训练集都引入了一定的外部单词，这些外部知识在一定程度上提升了模型的泛化性能. 由表2可知，当不改变构图方式时，仅对训练集进行特征增强，最终的模型性能得益于外部知识以及同类特征极大化聚合而有了一定的提升. 结合两个方面的改进后，模型吸收了它们各自的优势而拥有了更好的性能.

表 3 不同数据集非重复原始词个数以及外部词个数

	R8	R52	Ohsumed	MR
原始词数量	7 522	8 695	13 024	15 524
外部词数量	397	450	832	1 202

对于同类特征极大化聚合的验证，分别挑选R52和Ohsumed数据集进行实验. 所选两个数据集中分别随机选取各自所有类别中的8种类别2次进行验证，每种类别取10条数据，结果如图2和图3所示. 颜色越深代表比重越大，随机选取不同的类别后，依然符合同类特征极大化聚合的结论，也就是某个类别中样本的增强文本的替换词大部分来自其同类样本.

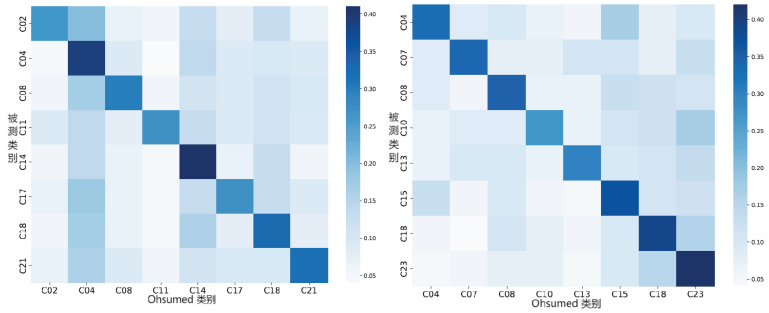

图 2 Ohsumed数据集随机选取类别组的替换词在该组各类别的占比

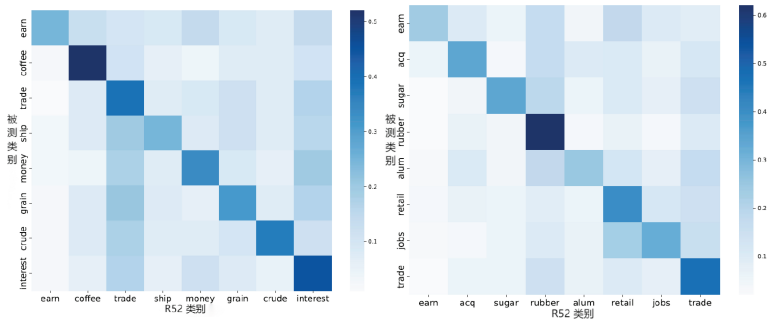

图 3 R52数据集随机选取类别组的替换词在该组各类别的占比

4 总 结

本文基于BertGCN模型进行改进，分别从其构图方式和特征增强方面入手，使用了BM25+算法来计算单词-文档之间的边权重，并且使用了文本特征和图结点混合增强策略对文本进行特征补充及增强. 一方面，BM25+算法优化了边的权重；另一方面，增强结点融合了同类知识以及外部知识.

实验结果表明，MfeGCN模型均比其它的基线模型性能好，并且在消融实验中各项结果也体现了各部分改进模块的有效性. 未来的工作中，将针对图结构和增强样本进一步改善处理，结合各自优势进一步优化模型，并加以验证分析.

参考文献：

[1] 张晓霞，王名扬，贺慧新，等. 结合情感分析的突发事件舆情网络关键节点挖掘[J]. 新疆大学学报(自然科学版)，2015，32(3)：336-341.
ZHANG X X，WANG M Y，HE H X，et al. Mining key-nodes of emergency information dissemination network based on sentiment analysis[J]. Journal of Xinjiang University(Natural Science Edition)，2015，32(3)：336-341. (in Chinese)

[2] 艾山·吾买尔，魏文琳，早克热·卡德尔. 基于 BiLSTM+Attention 的体育领域情感分析研究[J]. 新疆大学学报(自然科学版)(中英文)，2020，37(2)：142-149.
AISHAN W，WEI W L，ZAOKERE K. Sentiment analysis based on BiLSTM+Attention in sports field[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English)，2020，37(2)：142-149. (in Chinese)

[3] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1746-1751.

[4] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. July 9-15, 2016, New York, NewYork, USA. ACM, 2016: 2873-2879.

[5] CHENG J P, DONG L, LAPATA M. Long short-term memory-networks for machine reading[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 551-561.

[6] SUNDERMEYER M, SCHLÜTER R, NEY H. LSTM neural networks for language modeling[C]//Interspeech 2012. ISCA: ISCA, 2012: 194-197.

[7] WU Z H, PAN S R, CHEN F W, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. 2018: arXiv: 1810.04805. https://arxiv.org/abs/1810.04805.pdf.

[9] WEI J, ZOU K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 6382-6388.

[10] FENG S Y, GANGAL V, WEI J, et al. A survey of data augmentation approaches for NLP[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 968-988.

[11] LIN Y X, MENG Y X, SUN X F, et al. BertGCN: Transductive text classification by combining GCN and BERT[EB/OL]. 2021: arXiv: 2105.05727. https://arxiv.org/abs/2105.05727.pdf.

[12] LOWD D, DOMINGOS P. Naive Bayes models for probability estimation[C]//Proceedings of the 22nd International Conference on Machine Learning—ICML'05. August 7-11, 2005. Bonn, Germany. ACM, 2005: 529-536.

[13] ZHAO P N, LAI L F. Analysis of KNN density estimation[J]. IEEE Transactions on Information Theory, 2022, 68(12): 7971-7995.

[14] SAGI O, ROKACH L. Approximating XGBoost with an interpretable decision tree[J]. Information Sciences, 2021, 572: 522-542.

[15] DING C, BAO T Y, HUANG H L. Quantum-inspired support vector machine[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7210-7222.

[16] 杨济东，南新元，查琴. 鲁棒加权最小二乘支持向量回归的进气量预测[J]. 新疆大学学报(自然科学版)(中英文), 2022, 39(2): 189-196.
YANG J D, NAN X Y, ZHA Q. A robust weighted least squares support vector regression for air input prediction[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2022, 39(2): 189-196. (in Chinese)

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

[18] XIONG J C, XIONG Z P, CHEN K X, et al. Graph neural networks for automated *de novo* drug design[J]. Drug Discovery Today, 2021, 26(6): 1382-1393.

[19] LI Y, QIAN B Y, ZHANG X L, et al. Graph neural network-based diagnosis prediction[J]. Big Data, 2020, 8(5): 379-390.

[20] ZHANG J N, SHI X J, XIE J Y, et al. GaAN: Gated attention networks for learning on large and spatiotemporal graphs[EB/OL]. 2018: arXiv: 1803.07294. https://arxiv.org/abs/1803.07294.pdf.

[21] WANG H W, ZHAO M, XIE X, et al. Knowledge graph convolutional networks for recommender systems[C]//WWW'19: The World Wide Web Conference. May 13-17, 2019, San Francisco, CA, USA. ACM, 2019: 3307-3313.

[22] LYU Y H, ZHAI C X. When documents are very long, BM25 fails![C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. July 24-28, 2011, Beijing, China. ACM, 2011: 1103-1104.

(下转第 109 页)