网络首发时间：2021-04-08 10:18:30
网络首发地址：https://kns.cnki.net/kcms/detail/37.1389.N.20210407.1345.004.html
山东大学学报(理学版) 2021 年 5 月 第 56 卷 第 5 期
Journal of Shandong University (Natural Science), Vol.56, No.5, 2021
E-mail: xblxb@sdu.edu.cn
http://lxbwk.njournal.sdu.edu.cn
Tel: +86-531-88366917


文章编号: 1671-9352(2021)05-0057-09 DOI: 10.6040/j.issn.1671-9352.1.2020.060

# 基于半监督图神经网络的短文本分类

张斌艳，朱小飞*，肖朝晖，黄贤英，吴洁
(重庆理工大学 计算机科学与工程学院，重庆 400054)

**摘要:** 文中提出了在短文本建模过程中引入词项与词项之间、词项与文档之间的全局结构关系来增强短文本的表示。由于有标签训练数据的缺乏，使得现有的全局结构关系建模方法，如 TextGCN，无法学习到高质量的词项和文档全局结构表示，因此，文中进一步提出采用半监督学习思想来解决有标签训练数据不足的问题。实验结果表明，在基准数据集 MEDUI 上，与现有相关模型进行对比，文中提出的方法比最好的基准模型在 $F_1$ 指标上提高了 1.91%。

**关键词:** 图神经网络; 半监督学习; 短文本分类
**中图分类号:** TP391 **文献标志码:** A
**引用格式:** 张斌艳，朱小飞，肖朝晖，等.基于半监督图神经网络的短文本分类[J]. 山东大学学报(理学版)，2021，56(5)：57-65.

# Short text classification based on semi-supervised graph neural network

ZHANG Bin-yan, ZHU Xiao-fei*, XIAO Zhao-hui, HUANG Xian-ying, WU Jie
(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

**Abstract:** This paper proposes to introduce the global structural relationship between terms and terms and between terms and documents in the process of short text modeling to enhance the representation of short text. Due to the lack of labeled training data, existing global structural relationship modeling methods, such as TextGCN, cannot learn high-quality terms and document global structure representations. Therefore, we further propose to adopt the idea of semi-supervised learning to solve the problem of insufficient training data. On the benchmark dataset MEDUI, we compare with the existing related models. The experimental results show that the method proposed in this paper improves the $F_1$ index by 1.91% compared with the best benchmark model.
**Key words:** graph neural network; semi-supervised learning; short text classification

## 0 引言

随着社交媒体、电子商务和在线通信的飞速发展，互联网已经产生了越来越多的信息，包括文本、搜索词条、微博评论等。文本分类是自然语言处理中的一项基本且重要的任务，在情感分析[1]、问答系统[2]以及垃圾邮件检测[3]等应用中发挥着重要作用。在现有的研究中，文本分类已经取得了很大的进展，包括基于人为设计特征的传统方法和基于深度架构的神经网络[4]，但是，这些方法更适合处理长文档和段落，对短文本分类的效果仍然有限。近年来，基于深度架构的神经网络[5]被广泛应用于包括卷积神经网络(CNN)[6]和递归神经网络(RNN)的短文本分类任务中，如长短期记忆网络(LSTM)[7]等。由于 CNN 和 RNN 优先考虑位置和顺序关系[8]，因此这些模型可以很好地捕获本地连续词项序列中的上、下文语义和短语信息，但对于非连续的、长距离语义信息的建模能力不足[9]。由于短文本包含的词项少，因此导致文本内容缺乏足够的上、下文信息[9]，极大地限制了短文本分类任务的完成。

收稿日期: 2020-08-18
基金项目: 国家自然科学基金资助项目(61702063, 61502065)；重庆市基础科学与前沿技术研究项目(cstc2017jcyjBX0059, cstc2017jcyjAX0339)；重庆市教委语言文字科研项目重点项目(yyk20103)
第一作者简介: 张斌艳(1996—)，女，硕士研究生，研究方向为机器学习和自然语言处理. E-mail: 1576579348@qq.com
* 通信作者简介: 朱小飞(1979—)，男，博士，教授，研究方向为机器学习、信息检索和自然语言处理. E-mail: zxf@cqut.edu.cn

为了解决上述问题,一些研究人员提出将知识库加入传统方法或神经网络中,以克服这些局限性。引入外部资源信息可以为短文本的特征提取提供丰富的语义信息,但这些方法的性能在很大程度上取决于所引入知识库的质量,并且构造大规模的知识库即费时又费力。另一种策略是挖掘文本的潜在主题或聚类特征,并将其作为特征输入到一些分类器中,这种方法可以减少高维数和术语稀疏分布的问题,但其不足之处在于过度依赖预先训练的主题或聚类特征,导致难以捕捉聚类和分类之间的潜在关联。

近年来,图神经网络(Graph Neural Network, GNN)引起了研究人员的广泛关注[8]。受TextGCN[10]可学习词项与文档在文档集上全局结构关系的启发,我们提出了一种新的基于半监督图神经网络的短文本分类方法BSGNN(Based on Semi-supervised Graph Neural Network)。首先,我们将文档集构造成一个文本图Text Graph,为了缓解TextGCN建模短文本数据的局限,我们基于半监督学习的思想提出SemiTextGCN,通过从外部引入部分无标注样本参与文本图的构建,帮助增强词项与词项、词项与文档之间的结构关系表示,提升文本全局结构信息的建模能力。在学习到词项和文档全局特征表示之后,我们将其与现有的局部短文本建模方法进行融合。具体而言,我们将SemiTextGCN学习到的词项和文档的嵌入表示分别对应拼接到一个主要以双向门控循环(BiGRU)和多头自注意力组成的基础模型上。该基础模型首先使用BiGRU学习词项表示并捕获文本中的局部上、下文信息,然后将SemiTextGCN学习到的词项表示对应拼接到BiGRU的隐藏层输出上,目的是获取更为准确的词项特征表示。在拼接后的特征表示上使用一个多头自注意力层,挖掘不同方面的文本上、下文依赖关系。最后,在多头自注意力的输出上使用一个最大池化层,选取最显著的特征信号,并将该信号与SemiTextGCN学习到的文档级表示拼接。我们将该信号送入全连接层,并用softmax输出分类结果,并在基准数据集MEDUI上,与现有相关模型进行对比。

本文将适用于长文本分类任务的TextGCN模型进行改进,解决了在短文本分类任务中数据稀疏且没有足够上、下文的问题,为研究短文本分类任务提供新的思路;利用半监督学习的思想,从外部选取部分同类型无标注样本辅助训练并得到更加准确的词项和文档的特征表示,验证了半监督学习在短文本分类任务中的有效性;实验结果证明,本文提出的方法可显著提升短文本分类任务的效果且始终优于基准方法。

## 1 相关工作

传统的文本分类研究主要集中在特征工程和分类算法上,特征工程通常依赖于手动功能,其中最常用的特征是词袋特征和N-gram[11]。最新研究为特定应用设计了更复杂的功能,例如Lazaridou等[12]在情感分类的贝叶斯模型中考虑了连接词,Post等[13]使用多种显式和隐式语法功能(例如单字、双字和语法树模式)进行文本分类,Zhang等[4]将word2vec学习的词嵌入集成到支持向量机模型中,但特征工程非常费时费力,且对人工的专业性要求比较高。

最近,深度学习方法已被证明对文本分类有效,Kim于2014年提出卷积神经网(CNN)体系结构[6],该体系结构使用具有不同过滤器窗口大小的多个并行卷积层,并将选定的重要特征连接到一个密集的softmax层中,以进行文本分类。Lai等[14]应用递归结构来学习每个词项的上、下文信息,并采用最大池化层来捕获文本中最重要的特征。另一种先进的方法是用于文档分类的分层注意力网络[5],该方法基于文档的层次结构,并在BiLSTM提取的词项和文档表示上加一层注意机制,由于该方法更关注文档内部的局部信息,因此对文本全局结构信息并没有很好的建模能力。

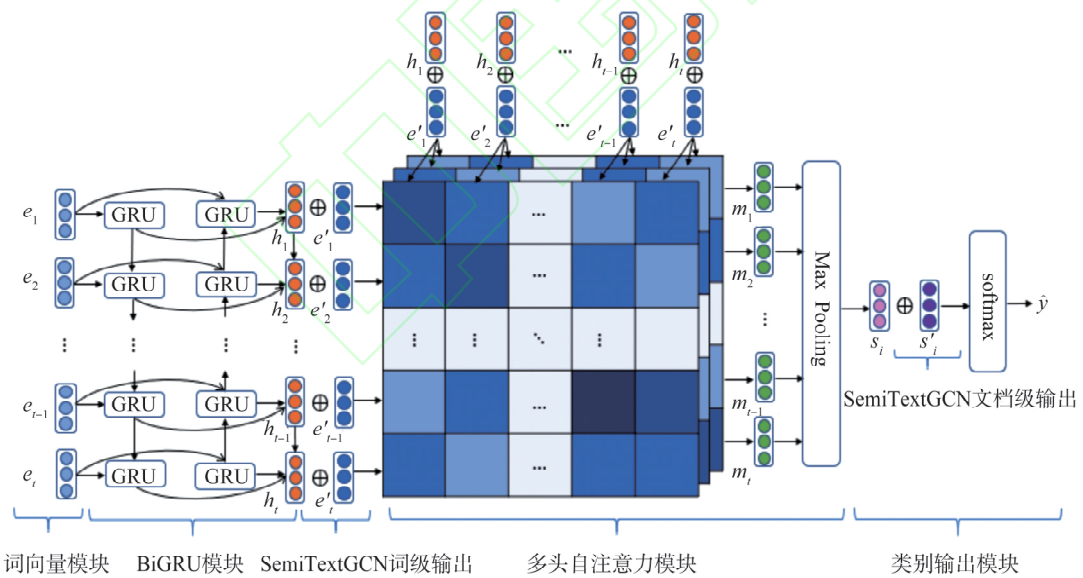
图神经网络最近受到越来越多的学者关注。Kipf等[15]提出了一种简化的图神经网络模型,称为图卷积网络(GCN),许多基准图数据集在该模型上均实现了最新的分类结果。GCN也被应用于其他NLP任务中,例如语义角色标签[16]、关系分类[17]和机器翻译[18],主要用于编码句子的句法结构。最新的一些研究探索了图神经网络在文本分类中的应用,图神经网络可以在图嵌入中保留全局结构信息,该网络被证明对具有丰富关系结构的任务非常有效。近期,基于图神经网络的方法TextGCN在长文本分类任务中取得了很好的效果。TextGCN将整个语料库中的词项和文档作为节点,通过词项与文档之间的关系构建成一整张图,并使用图神经网络(GCN)对图建模。图卷积网络是捕获高阶邻域信息简单且有效的图神经网络,此时的文本分类问题就转换成了节点分类问题;但由于短文本具有数据稀疏且没有足够上、下文的特点,因此导致该方法更适合于长文本,对短文本并没有很好的效果。

以上这些方法对长文本,特别是对长文档和段落来说,具有良好的性能,但是这些模型主要建模文本局部语义信息,忽略了文本的全局结构信息。图神经网络方法中的 TextGCN 往往更适合处理长文本,且该模型面临有监督数据缺乏的问题,因而在处理短文本分类问题时效果并不理想。我们提出 BSGNN 模型,其中的 SemiTextGCN 模块在 TextGCN 构图的基础上添加部分同类型无标注样本,然后将 SemiTextGCN 学习到的特征表示对应拼接到模型预先学习到的词级和文档级的嵌入上,目的是深度挖掘文本全局结构信息,学习到更为精准的词项和文档的特征表示。

## 2　模型描述

模型的输入是短文本 $s$,输出是类别标签的概率分布。我们用 $p(y|s,\phi)$ 表示短文本 $s$ 为 $y$ 类的概率,其中 $\phi$ 是网络中的参数。

该模型包括 5 个模块: (1) 词向量表示模块。本文在表示词向量时采用的词向量词典是 gensim 由 word2vec 训练的维基百科词向量,若词项不存在于该词典中,我们将用词典中"0"元素所对应的向量替代; (2) SemiTextGCN 模块。在 TextGCN 构图时加入部分同类型无标注样本,目的是捕捉到文本全局结构信息并得到更为精准的特征表示; (3) BiGRU 模块。将获得的词向量经过一层双向 GRU,获得每个词语与其上、下文的关联,初步建模文本的语义信息,并将 SemiTextGCN 学到的词级嵌入对应拼接到 BiGRU 的隐藏输出上; (4) 多头自注意力模块。由多头自注意力捕捉词与词之间的交互作用,构建高质量词的上、下文依赖关系,并使用最大池化层捕获最重要的特征; (5) 类别输出模块。首先对应拼接池化后的输出和 SemiTextGCN 学到的文档级嵌入,并将这些信息送到全连接层中,由 softmax 激活函数输出每个类标签的概率。模型结构如图 1 所示。




$e_i$是词项经过 SemiTextGCN 得到的词级嵌入,$s_i$是句子经过 SemiTextGCN 后得到的文档级嵌入。

图 1　BSGNN 模型总体架构图

Fig.1　Overall architecture diagram of the BSGNN model

### 2.1　词向量表示模块

词是提取语义信息的基本单位,因此构建词表示通常是基于神经网络的前提。该模型的输入是长度为 $t$ 的短文本,在此模块中我们使用预训练的词向量[19]来获得每个词的词嵌入,并将其映射到高维向量空间得到 $d$ 维词向量序列,表示为 $(e_1,e_2,\cdots,e_{t-1},e_t)$。

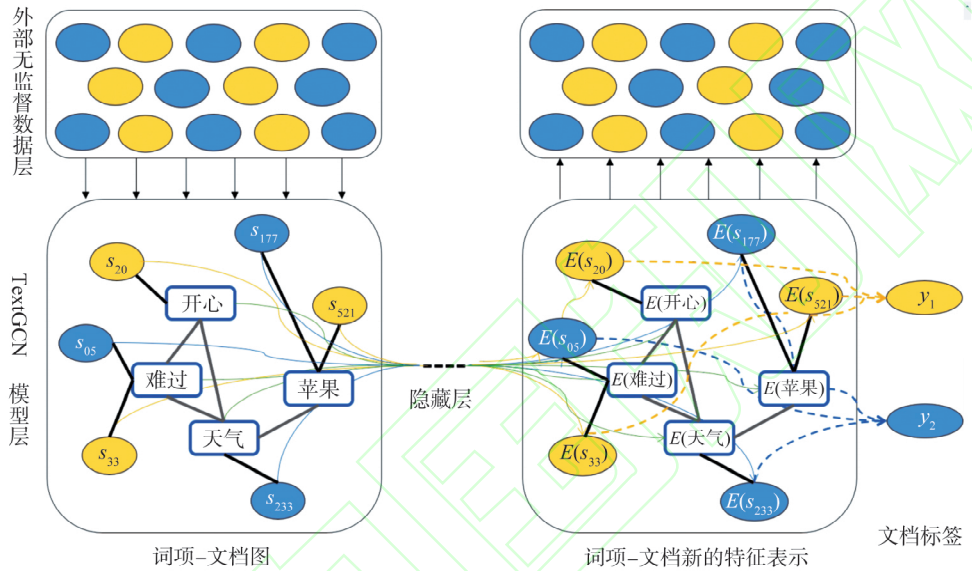
### 2.2　SemiTextGCN 模块

我们的目标是通过提取词项的全局结构信息来捕捉文档的全局依赖注意权重。首先在 TextGCN 构图前加入部分同类型无标注样本参与构图(如图 2 所示),接着在构成的文本图上用 GCN 进行训练,从全局的角度学习特征表示,最后在学习到的所有特征表示中单独抽取出本文数据集的表示。文本图是由整个语料

库中的词项和文档以及外部无标注样本的词项和文档作为节点构成的,利用点互信息(PMI)建立词项与词项节点之间的边,利用文档中词项的 TF-IDF 构建词项与文档节点之间的边。然后我们在文本图上应用一个双层的 GCN,如式(1)所示:

$$\hat{y}_{\text{semi}}=\text{softmax}(\overline{A}'\sigma(\overline{A}'XW_0)W_1)。\tag{1}$$

令 $A'=A+I$,其中 $A$ 是通过 PMI 与 TF-IDF 计算得到的矩阵,$I$ 是单位矩阵。$\overline{A}'=D^{-\frac{1}{2}}A'D^{-\frac{1}{2}}$,$\hat{y}_{\text{semi}}$ 是 SemiTextGCN 的输出,$W_0$ 和 $W_1$ 为可训练矩阵,$D$ 是将矩阵 $A'$ 各行元素作为对角元素,即 $D_{ii}=\sum_{j=1}^{n}A'_{ij}$。$X\in\mathbf{R}^{m\times 2n}$ 是一个包含 $m$ 个词项和文档节点及其特征的矩阵,$e_i'\in\mathbf{R}^{2n}$ 是每个词项特征向量 $e_i$ 经 SemiTextGCN 学习得到的新词级表示,$s_i'\in\mathbf{R}^{2n}$ 是每个文档特征向量 $s_i$ 经 SemiTextGCN 学习得到的新文档级表示。




$s_i$ 是文档节点,其他是词项节点;黑色粗体边缘为文档与词项边缘,灰色细边缘为词项与词项边缘;$E(x)$ 表示经过 Text-GCN 之后的 $x$ 的嵌入,公式中词级嵌入用 $e_i$ 表示,文档级嵌入用 $s_i$ 表示;不同的颜色代表不同的文档类别;为简洁明了,例子中仅展示 2 个示例类别。

图 2 SemiTextGCN 模型图

Fig.2 SemiTextGCN model graph

### 2.3 BiGRU 模块

BiGRU [20] 由前向和后向网络组成,用以处理短文本。将上层模块得到的词向量序列采用 BiGRU 捕获序列的定点信息,可进一步提高网络的表达能力。

$$\overrightarrow{h_i}=\overrightarrow{\text{GRU}}(e_i,\overrightarrow{h_{i-1}}),\tag{2}$$

$$\overleftarrow{h_i}=\overleftarrow{\text{GRU}}(e_i,\overleftarrow{h_{i+1}})。\tag{3}$$

将每个 $\overrightarrow{h_i}$ 和 $\overleftarrow{h_i}$ 连接起来,获得隐藏状态 $h_i$。设每个单向 GRU 的隐藏单元数为 $u$,我们将所有 $h_i s$ 表示为 $H\in\mathbf{R}^{t\times 2u}$:

$$H=(h_1,h_2,\cdots,h_t);\tag{4}$$

再将 SemiTextGCN 模块训练得到的词级嵌入 $e_i'$ 对应拼接到 BiGRU 的隐藏层输出上,表示为 $H'\in\mathbf{R}^{t\times 4u}$:

$$H'=(h_1\oplus e_1',h_2\oplus e_2',\cdots,h_t\oplus e_t')。\tag{5}$$

### 2.4 多头自注意力模块

多头自注意力机制的原理是缩放点积注意力 [21],输入由矩阵 $Q\in\mathbf{R}^{n\times d_k}$、$K\in\mathbf{R}^{m\times d_k}$、$V\in\mathbf{R}^{m\times d_v}$ 组成,缩放点积注意力可通过以下公式计算出 Attention 的分值:

$$M=\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V,\tag{6}$$

其中 $Q=K=V=H'$,$1/\sqrt{d_k}$ 表示可用于调节缩放的程度(避免 $Q$、$K$ 的内积过大)。该模块的输出是一个矩

阵,表示为 $M \in \mathbf{R}^{t\times 4u}$。接下来在 $M$ 上使用最大池化层获取短文本的特征表示 $s_i \in \mathbf{R}^{4u}$,目的是在向量的每个维度上选择最大值以捕获最重要的特征。

### 2.5　类别输出模块

首先将 2.4 模块的输出对应拼接上 SemiTextGCN 训练得到的文档级嵌入 $s_i'$,如公式(7)所示:

$$S_i' = (s_i \oplus s_i'); \tag{7}$$

接着使用带有 ReLU 激活函数的全连接层转换短文本的隐藏表示,如公式(8)所示:

$$G_i' = \mathrm{ReLU}(W_i \times S_i' + b_i); \tag{8}$$

其中 $W_i$ 和 $b_i$ 分别为第 $i$ 个特征表示的权重和偏置项,然后用 softmax 分类器对其进行分类,计算方式如公式(9)所示:

$$\hat{y} = \mathrm{softmax}(W_i' \times G_i' + b_i')。 \tag{9}$$

本文使用的损失函数 $\mathscr{L}$ 如公式(10)所示:

$$\mathscr{L} = -\lambda \frac{1}{|S|} \sum_{y\in S} \sum_{i=1}^{2} y_k \log(\hat{y}_i), \tag{10}$$

其中 $S$ 代表数据集,$y_i$ 代表数据集的标签,$\lambda$ 是超参。

## 3　实验

### 3.1　数据集

本文基于微博构建一个新的、含有情绪的短文本数据集 MEDUI,该数据集选取了 200 位活跃度较高的微博用户,爬取了大约 1 万条微博文本,并对数据集进行人工情感标注与筛选,最终筛选出 2 741 条带有积极、消极情绪的微博文本。实验随机抽取 80%的语句(共 2 193 条语句)作为训练集,剩下的 20%(共 548 条语句)作为测试集[22]。为验证我们的方法,我们还从微博搜集了 10 万条同类型样本,并去掉样本的标签信息。

### 3.2　实验参数设置

本文所有的模型均采用 Adam[23] 进行学习,模型参数设置如表 1 所示。

表 1　模型参数设置表

Table 1　Model parameter setting table

| 参数名 | 值 |
|---|---|
| 词向量维度 | 200 |
| 学习率 | 0.01 |
| 权重正则限制 | 2 |
| dropout | 0.5 |
| 批处理大小 | 64 |
| 外部补充数据 | · 50 000 |

### 3.3　对比模型介绍

为了更好地验证本文的性能,在 MEDUI 数据集上,本文与 8 种短文本分类算法进行实验对比。

LR: 线性回归(Linear Regression, LR)模型首先使用 TF-IDF 对短文本进行表示,然后对短文本使用传统的回归分析方法对语句进行标签分析,该方法不考虑句子整体的结构信息。

SVM[24]: 支持向量机(Support Vector Machine)是一种二分类模型,使用 word2vec 和 TF-IDF 相结合的方法表示短文本的文本特征,最后用 SVM 分类器进行短文本分类。

W2V+CNN[6]: 该方法将卷积神经网络用于句子级分类任务,首先使用 word2vec 训练词向量,并将短文本看成一个词向量的序列,最后利用卷积神经网络学习短文本分类模型并进行分类。

TNA[25]: 该模型提出了一种具有注意机制的新型双通道神经网络结构,首先利用序列 LSTM 通道捕获语义信息,并应用 Tree-LSTM 通道获取语法信息,通过建模句子结构特征来加强深层语义学习,最后采用全连接层来合并语义和句法信息。

EV-CNN[26]: 该模型利用 Wiki 中文数据集和网络术语来扩展原始词汇,训练新的词嵌入,并基于卷积神经网络实现句子级文本分类,同时提出了根据池层语句长度的优化方法。

UA-LSTM[27]: 该方法在提取文本特征时不仅考虑了传统的基于情感词典的方法,还通过建模用户自身的情感标志得分来帮助识别语句的情感特征,最后用 softmax 层实现对短文本的分类,实验结果表明该方法在短文本情感分析方面具有显著的作用。

HAN[5]:该方法提出了一种分层的文档分类注意力网络(Hierarchical Attention Network, HAN),该模型分别在词级和文档级使用2个层次的注意力机制,使它能够在构建文档表示时区别地关注文本信息。在6个长文本分类任务上的实验表明,所提出的体系结构在很大程度上优于以前的方法。

NPA[27]:该模型通过观察到不同的用户通常有不同的兴趣、同一用户可能有不同的兴趣这一现象,提出一种具有个性化注意的神经网络推荐模型(Neural news recommendation with Personalized Attention, NPA),并建议在词级和新闻级分别应用注意力机制来帮助模型关注重要的词汇和新闻,不仅注重提取文本上、下文信息,还将用户情感倾向信息嵌入到注意力机制中来,用于生成文字和新闻级注意的查询向量。实验结果表明,该模型在MSN真实新闻数据集上明显优于其他最新方法。

### 3.4 实验结果分析

本文采用的评价模型性能指标主要有精确率(precision,$P$)、召回率(recall,$R$)、$F_1$-measure($F_1$),这3个指标是机器学习、自然语言处理中最常用的3个指标。

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$F1=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}},$$

其中TP是将正类正确预测为正类数,FP是将负类错误预测为正类数,FN是将正类错误预测为负类数。

表2是9种模型在数据集MEDUI上的评测结果。从表2可知,使用TF-IDF的LR方法分类效果最差,$F_1$值仅为0.70;SVM方法效果要好过LR方法,可达到0.78,这是因为SVM方法能够建模非线性数据;基于卷积神经网络模型的方法W2V+CNN比SVM方法提高了6.4%,这是由于深度学习拥有良好的建模能力;TNA的分类效果高于EV-CNN,$F_1$值达到0.84,该方法利用了注意力机制和Tree-LSTM可深度挖掘语句结构特征的特点;EV-CNN则是将词语的情感得分和权重得分进行建模,在模型中加入情感信息来帮助模型提高短文本分类性能,其$F_1$值达到了0.88;UA-LSTM引入用户自身的情感倾向,其分类结果超过了EV-CNN,达到了0.91;HAN将2个层次的注意力机制分别作用于词级和文档级,联合挖掘文档信息,由于本文使用的数据集是短文本,因此本文在复现HAN的方法时,仅在词级上应用一层注意力机制,该注意力层选择了具有定性信息的词项,最终$F_1$值超过了UA-LSTM,达到了91.67%;NPA提出了一种将用户情感倾向信息嵌入到词级和新闻级注意力机制中,帮助模型关注重要的词汇和新闻,最终NPA的$F_1$值高于HAN 0.74%,因为NPA不仅注重提取文本上下文信息,另将用户情感倾向信息建模到注意力机制中去。

表2 不同模型在3个指标($P$、$R$、$F_1$)上的测试结果

Table 2 Test results of different models on three indicators ($P$, $R$, $F_1$) %

| 模型 | $P$_1 | $P$_0 | $R$_1 | $R$_0 | $F_1$_1 | $F_1$_0 | $P$_Avg | $R$_Avg | $F_1$_Avg |
|---|---|---|---|---|---|---|---|---|---|
| LR | 0.67 | 0.73 | 0.57 | 0.80 | 0.61 | 0.76 | 0.70 | 0.71 | 0.70 |
| SVM | 0.76 | 0.80 | 0.69 | 0.85 | 0.72 | 0.82 | 0.78 | 0.78 | 0.78 |
| W2V+CNN | 0.85 | 0.81 | 0.74 | 0.90 | 0.79 | 0.85 | 0.83 | 0.83 | 0.83 |
| TNA | 0.91 | 0.80 | 0.72 | 0.94 | 0.80 | 0.86 | 0.85 | 0.84 | 0.84 |
| EV-CNN | 0.91 | 0.86 | 0.79 | 0.95 | 0.84 | 0.90 | 0.88 | 0.88 | 0.88 |
| UA-LSTM | 0.92 | 0.91 | 0.88 | 0.94 | 0.90 | 0.92 | 0.91 | 0.91 | 0.91 |
| HAN | 88.97 | 96.55 | 97.78 | 83.69 | 93.16 | 89.65 | 92.22 | 91.77 | 91.67 |
| NPA | 90.22 | 96.53 | 93.81 | 85.13 | 93.81 | 90.43 | 92.84 | 92.5 | 92.41 |
| **BSGNN** | **93.73** | **95.28** | **96.76** | **90.98** | **95.22** | **93.07** | **94.38** | **94.34** | **94.32** |

### 3.5 无标注样本对模型性能影响

我们的模型采用SemiTextGCN来获取词项和文档的全局特征表示。为了分析添加不同数量的无标注样本对SemiTextGCN模型分类性能的影响,我们以步长5 000依次从0条添加到6万条无标注样本。值得注意的是,当添加0条无标注样本时,SemiTextGCN等价于TextGCN。

由图3可知,添加数据数量从0到1万时,模型准确率有较明显的提升,初步判定是由于添加外部数据

补充了短文本的全局结构信息; 而从 1 万添加到 5 万条数据时,模型效果提升速度逐渐平缓,并在添加数量为 5 万时达到最优解,这可能是由于在这个阶段虽然外部信息有助于补充短文本数据稀疏的问题,但同时加进来的噪音也较多,干扰到了模型学习特征表示的效果; 当添加数据大于 5 万时,模型分类效果呈下降趋势,这可能是添加的数据过多导致噪音过大,反而干扰了模型对特征表示的学习。

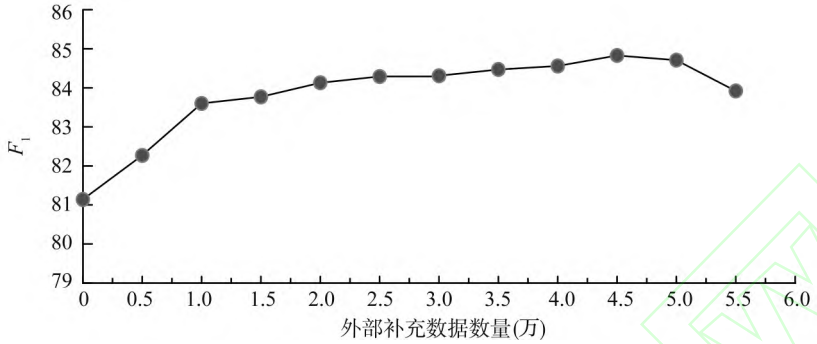



图 3　不同补充数量在 $F_1$ 值上的表现

Fig.3　The performance of different supplements in $F_1$

### 3.6　模块对比性实验

本文在基础模型 Base 上融合了经 SemiTextGCN 处理后得到的词级和文档级的特征嵌入,其中基础模型是使用 BiGRU 和多头自注意力联合提取文本内部的结构信息,接着用最大池化层提取最强的特征表示并将其送入全连接层,最后由 softmax 分类器输出类别概率分布,最终本文的实验结果比最好的基准模型 NPA 高出了 1.91%,表 3 中 Base+W、Base+S 表示分别在基础模型的词级和文档级拼接经过 TextGCN 学习得到的词级和文档级的嵌入,Base+WS 表示同时在基础模型的词级和文档级添加经过 TextGCN 学习得到的词级和文档级的嵌入,Base+W* 、Base+S* 表示分别在基础模型的词级和文档级拼接经过 SemiTextGCN 学习得到的词级和文档级的嵌入,BSGNN 表示同时在基础模型的词级和文档级添加经过 SemiTextGCN 学习得到的词级和文档级的嵌入。

表 3　本文不同组件在 3 个指标( $P$、$R$、$F_1$) 上的测试结果

Table 3　Test results of different components on three indicators ( $P$, $R$, $F_1$)　%

| 对比模型 | 指标 | 积极 | 消极 | 总体 |
|---|---|---|---|---|
| Base | $P$ | 90.62 | 95.17 | 92.52 |
| | $R$ | 96.87 | 86.03 | 92.34 |
| | $F_1$ | 93.64 | 90.37 | 92.27 |
| Base+W | $P$ | 92.50 | 89.91 | 91.42 |
| | $R$ | 92.79 | 89.52 | 91.42 |
| | **$F_1$** | **92.64** | **89.72** | **91.42** |
| Base+S | $P$ | 95.38 | 87.76 | 92.19 |
| | $R$ | 90.60 | 93.89 | 91.97 |
| | **$F_1$** | **92.39** | **90.72** | **92.00** |
| Base+WS | $P$ | 93.29 | 94.09 | 93.63 |
| | $R$ | 95.92 | 90.39 | 93.61 |
| | **$F_1$** | **94.59** | **92.20** | **93.59** |
| Base+W* | $P$ | 91.57 | 93.06 | 92.19 |
| | $R$ | 95.30 | 87.77 | 92.15 |
| | **$F_1$** | **93.39** | **90.34** | **92.12** |
| Base+S* | $P$ | 93.17 | 91.59 | 92.51 |
| | $R$ | 94.04 | 90.39 | 92.52 |
| | **$F_1$** | **93.60** | **90.99** | **92.51** |
| *BSGNN* | $P$ | 93.73 | 95.28 | 94.38 |
| | $R$ | 96.76 | 90.98 | 94.34 |
| | **$F_1$** | **95.22** | **93.07** | **94.32** |

由实验结果可知,在 Base 模型中仅添加由 TextGCN 训练得到的词级或文档级嵌入,其结果相对于 Base 并没有提升,但同时添加词级和文档级嵌入,则模型的分类效果要高于 Base 1.37%,这可能是由于分别添加信息导致词级和文档级信息不匹配造成的结果偏差。分别将 SemiTextGCN 训练得到的词级或文档级嵌入对应添加到 Base 模型中,可看出文档级的嵌入比词级的嵌入更能帮助模型分类。将 SemiTextGCN 训练得到的词级或文档级嵌入同时添加到 Base 模型中去,可较好地解决短文本数据稀疏的问题,最终模型的 $F_1$ 值达到了 94.32%。

## 4 结论

本文针对现有短文本存在的数据稀疏且缺乏足够上下文的特点,且现有模型不能充分挖掘短文本全局结构信息这一问题,提出了一种新的基于半监督图神经网络的短文本分类方法。该方法首先用 BiGRU 和多头自注意力充分挖掘文本内部的词依赖关系,并捕捉到序列的内部结构,然后在基础模型中添加 SemiText-GCN 学习得到的文本特征表示。实验结果表明,本文提出的模型在基准数据集上取得了较好的分类效果,但由于该模型在给新的数据进行分类时,需要再将全部的样本重新构图并输入到 GCN 中进行计算,因此导致该模型的灵活性不佳。在下一步工作中,我们将重点解决模型灵活性的问题,探索如何把已处理过的文本图的参数直接应用到新的文本图中去。

**参考文献:**


[1] WANG Fang, WANG Zhongyuan, LI Zhoujun, et al. Concept-based short text classification and ranking [C] // Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. New York: ACM, 2014: 1069-1078.

[2] LEE Jiyoung, DERNONCOURT F. Sequential short-text classification with recurrent and convolutional neural networks [C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016: 515-520.

[3] CAI Hongyun, ZHENG V W, CHANG Kevin Chen-chuan. A comprehensive survey of graph embedding: problems, techniques and applications [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616-1637.

[4] ZHANG Dongwen, XU Hua, SU Zengcai, et al. Chinese comments sentiment classification based on word2vec and SVMperf [J]. Expert Systems with Applications, 2015, 42(4): 1857-1863.

[5] YANG Zichao, YANG Diyi, DYER Chris, et al. Hierarchical attention networks for document classification [C] //Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016: 1480-1489.

[6] KIM Yoon. Convolutional neural networks for sentence classification [J]. arXiv, 2014: 1746-1751. https://arxiv.org/abs/1408.5882.

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

[8] BATTAGLIA P W, HAMRICK J B, BAPST V, et al. Relational inductive biases, deep learning, and graph networks [J]. arXiv, 2018. https://arxiv.org/pdf/1806.01261.pdf.

[9] PENG Hao, LI Jianxin, HE Yu, et al. Large-scale hierarchical text classification with recursively regularized deep graph-cnn [C] // Proceedings of the 2018 World Wide Web Conference. Lyon: WWW, 2018: 1063-1072.

[10] YAO Liang, MAO Chensheng, LUO Yuan. Graph convolutional networks for text classification [R/OL]. AAAI, 2019: 7370-7377. https://arxiv.org/abs/1809.05679.

[11] WANG Sida, MANNING Christopher D. Baselines and bigrams: simple, good sentiment and topic classification [C] // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju: ACL, 2012: 90-94.

[12] LAZARIDOU A, TITOV I, SPORLEDER C. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations [C] //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: ACL, 2013: 1630-1639.

[13] POST Matt, BERGSMA Shane. Explicit and implicit syntactic features for text classification [C] //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: ACL, 2013: 866-872.

[14] LAI Siwei, XU Liheng, LIU Kang, et al. Recurrent convolutional neural networks for text classification [C] //Proceedings of the 29th AAAI Conference on Artificial Intelligence. [S.L]: [s.n.], 2015: 2267-2273.

[15] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [R/OL]. 2017. https://arxiv.org/pdf/1609.02907v4.pdf

[16] MARCHEGGIANI D, TITOV I. Encoding sentences with graph convolutional networks for semantic role labeling [C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. [S.l.]: [s.n.], 2017: 1506-1515.

[17] LI Yifu, JIN Ran, LUO Yuan. Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-Gcrns) [J]. Journal of the American Medical Informatics Association, 2019, 26(3): 262-268.

[18] BASTINGS J, TITOV I, AZIZ W, et al. Graph convolutional encoders for syntax-aware neural machine translation [C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017: 1957-1967.

[19] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013. https://arxiv.org/pdf/1301.3781v3.pdf.

[20] HAO Yanchao, ZHANG Yuanzhe, LIU Kang, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge [C] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL, 2017: 221-231.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] //Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: [s.n.], 2017.

[22] 吴洁, 朱小飞, 张宜浩, 等. 基于用户情感倾向感知的微博情感分析方法 [J]. 山东大学学报(理学版), 2019, 54(3): 46-55.
WU Jie, ZHU Xiaofei, ZHANG Yihao, et al. Microblog sentiment analysis method based on user's emotional orientation perception [J]. Journal of Shandong University (Natural Science), 2019, 54(3): 46-55.

[23] KINGMA D, BA J. Adam: a method for stochastic optimization [R/OL]. 2017. https://arxiv.org/pdf/1412.6980.pdf.

[24] LILLEBERG Joseph, ZHU Yun, ZHANG Yanqing. Support vector machines and word2vec for text classification with semantic features [C] // Proceedings of the 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing. Beijing: IEEE, 2015: 136-140.

[25] DAI Yuanfei, GUO Wenzhong, CHEN Xing, et al. Relation classification via LSTMs based on sequence and tree structure [J]. IEEE Access, 2018, 6: 64927-64937.

[26] YANG Xiaoyilei, XU Shuaijing, WU Hao, et al. Sentiment analysis of weibo comment texts based on extended vocabulary and convolutional neural network [J]. Procedia Computer Science, 2019, 147: 361-368.

[27] WU Chuhan, WU Fangzhao, AN Mingxiao, et al. NPA: neural news recommendation with personalized attention [C] //Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [S.l.]: ACM, 2019: 2576-2584.

(编辑: 于善清)

(上接第 56 页)

[18] LI Y, YANG W, XU X, et al. Micro-/mesoporous carbon nanofibers embedded with ordered carbon for flexible supercapacitors [J]. Electrochim Acta, 2018, 271: 591-598.

[19] LI H, YU H, QUAN X, et al. Improved photocatalytic performance of heterojunction by controlling the contact facet: high electron transfer capacity between $TiO_2$ and the {110} facet of $BiVO_4$ caused by suitable energy band alignment [J]. Advanced Functional Materials, 2015, 25(20): 3074-3080.

[20] ZHOU G, HU Y, LONG L, et al. Charged excited state induced by ultrathin nanotip drives highly efficient hydrogen evolution [J]. Applied Catalysis B: Environmental, 2020, 262: 118305.

[21] KIM Y, COY E, KIM H, et al. Efficient photocatalytic production of hydrogen by exploiting the polydopamine-semiconductor interface [J]. Applied Catalysis B: Environmental, 2021, 280: 119423.

[22] LU Y, YIN P, MAO J M, et al. A stable inverse opal structure of cadmium chalcogenide for efficient water splitting [J]. Journal of Materials Chemistry A, 2015, 3(36): 18521-18527.

[23] CHENG C, KARUTURI S, LIU L, et al. Quantum dots sensitized $TiO_2$ inverse opal for photoelectrochemical hydrogen generation [J]. Small, 2012, 8(1): 37-42.

(编辑: 于善清)