计算机科学
COMPUTER SCIENCE

基于混合字词特征的中文短文本分类算法

刘硕，王庚润，彭建华，李柯

引用本文

刘硕，王庚润，彭建华，李柯. 基于混合字词特征的中文短文本分类算法[J]. 计算机科学，2022，49(4)：282-287.

LIU Shuo，WANG Geng-run，PENG Jian-hua，LI Ke. Chinese Short Text Classification Algorithm Based on Hybrid Features of Characters and Words[J]. Computer Science，2022，49(4)：282-287.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200027

基于混合字词特征的中文短文本分类算法

刘 硕 王庚润 彭建华 李 柯
中国人民解放军战略支援部队信息工程大学 郑州 450000
(842964176@qq.com)

摘 要 随着信息技术的迅速发展,网络中产生了海量的中文短文本数据。利用中文短文本分类技术,在低信息量的数据中挖掘出有价值的信息是当前的一个研究热点。相比中文长文本,中文短文本具有字数少、歧义多以及信息不规范等特点,导致其文本特征难以提取与表达。为此,文中提出了一种基于混合字词特征深度神经网络模型的中文短文本分类算法。首先,该算法同时计算出中文短文本的字向量和词向量,并分别对其进行特征提取;然后将提取到的字向量特征和词向量特征进行融合;最后通过全连接层和softmax层完成分类任务。在公开的THUCNews新闻数据集上的测试结果表明,该算法在精确率、召回率和F1值3种评价指标上均优于主流的TextCNN,BiGRU,Bert以及ERNIE_BiGRU等对比模型,具有较好的短文本分类效果。

关键词:中文短文本分类;预训练模型;字向量;词向量;卷积神经网络
中图法分类号 TP391.1

Chinese Short Text Classification Algorithm Based on Hybrid Features of Characters and Words

LIU Shuo,WANG Geng-run,PENG Jian-hua and LI Ke
People's Liberation Army Strategic Support Force Information Engineering University,Zhengzhou 450000,China

Abstract The rapid development of information technology has lead to massive data of Chinese short texts on the Internet. As such,using classification technology to dig out valuable information from it is a current research hotspot. Compared with Chinese long texts,short texts have the characteristics of fewer words,more ambiguities and irregular information,making text feature extraction and expression a challenge. For this reason,a Chinese short text classification algorithm based on the deep neural network model of hybrid features of characters and words is proposed. First,the character vector and word vector of Chinese short text are calculated respectively. Then,their features are extracted and fused. Last,the classification task is accomplished through the fully connected layer and the softmax layer. The test results on the public THUCNews news data set show that the algorithm is better than the mainstream TextCNN,BiGRU,Bert and ERNIE_BiGRU comparison models in terms of accuracy,recall and F1 value. It has a good effect on short text classification.

Keywords Chinese short text classification,Pre-training model,Character vector,Word vector,Convolutional Neural Network

1 概述

近年来,随着信息技术的飞速发展,网络中产生了海量的短文本数据。为了提高短文本数据的利用效率,文本分类技术得到了广泛应用。文本分类是自然语言处理(Natural Language Processing,NLP)领域中最基本的任务之一,主要是将文本数据归类为预定类别标签中的一个或多个,被广泛应用于垃圾短信分类[1]、舆情监测[2]、情感分析[3]和用户个性化推荐[4]等实际任务中。相比长文本,短文本数据存在字数少、歧义多以及信息不规范等特点,造成文本特征难以提取与表达,从而导致短文本数据的分类效果往往不佳。

目前,深度学习在NLP领域已经取得突破性的进展。相比传统机器学习算法[5-6],基于深度学习的短文本分类算法往往能够取得更优的分类效果[7-8]。当前短文本分类任务的研究主要集中在文本表示和分类器模型的搭建上。

在文本表示方面,Mikolov等[9-10]提出了静态词向量表示模型word2vec,该模型可将高维稀疏的one-hot向量表示映射成低维稠密的词向量表示,同时能够较好地考虑文本上下文的语义信息。Peters等[11]提出了动态词向量表示模型ELMo,该模型通过结合每个词的上下文信息来计算词向量,能够较好地解决word2vec模型中存在的多义词表示问题。近年来,预训练模型在文本表示上取得了优异的效果。Devlin等[12]提出了基于深度双向Transformer的Bert预训练模型。该模型采用两阶段训练任务,第一阶段利用语言模型在大规模的语料库上进行无监督预训练;第二阶段根据具体的NLP任务进行有监督微调训练。实验结果表明,该模型在文本

到稿日期:2021-02-02 返修日期:2021-05-31 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61803384)
This work was supported by the National Natural Science Foundation of China(61803384).
通信作者:王庚润(wanggengrun@gmail.com)

表示任务上能够取得较高的分数。Lan 等[13] 提出了 ALBERT 预训练模型，该模型采用矩阵分解和跨层参数共享技术对 Bert 模型进行参数缩减，在维持 Bert 性能的情况下，降低了其空间复杂度，提高了模型的训练速度，同时具有较好的扩展性。Zhang 等[14]在 Bert 模型的基础上提出了 ERNIR1.0 版本。该模型利用知识图谱中的多信息实体作为外部知识来改善语言表征，并在中文文本表示任务中取得了优异的效果。Sun 等[15]提出了可连续学习的 ERNIE2.0 版本。该模型首先在大型文本语料库上进行无监督的词法级别、语法级别以及语义级别的预训练任务，然后持续用预训练任务来更新模型，这种训练方式进一步提升了模型的语言表示能力。

在分类器模型研究上，Kim 等[16]首次将卷积神经网络应用到文本分类任务中，提出了 TextCNN（Text Convolutional Neural Network，TextCNN）模型，该模型的优点是能够有效地提取文本的局部信息。Johnson 等[17]提出了一种基于词级的深度卷积神经网络模型 DPCNN（Deep Pyramid Convolutional Neural Networks，DPCNN），该模型的复杂度低，相比基于字符级的卷积神经网络模型有更优的分类效果。Li 等[18]提出了一种用于短文本分类的 LSTM-TextCNN 联合模型，该模型将词向量分别输入到 LSTM 模型和 TextCNN 模型中，然后将两个模型提取的特征进行融合，通过 softmax 层完成分类。实验结果表明，该方法优于单个的 LSTM 模型和 TextCNN 模型。Duan 等[19]提出了基于 Bert 的中文短文本分类模型，该模型将 Bert 表示的特征向量直接输入到 softmax 层进行分类，分类结果的 F1 值相比 TextCNN 模型提高了 6%。Zheng 等[20]提出了一种用于短文本分类的 BLSTM_MLPCNN 模型，该模型联合字符向量与词向量作为模型输入，采用 BLSTM 模型构建文档特征图，最后使用多层感知器神经网络 MLPCNN 进行特征提取。实验结果表明，该模型相比 CNN，RNN 以及两者的组合模型有更高的分类精度。Hou 等[21]提出了一种多神经网络混合的短文本分类模型，该模型分别利用 FastText 和 TextCNN 对短文本数据进行特征提取，然后将提取到的特征进行深度融合，从而充分发挥 FastText 和 TextCNN 两者的优势，提高了对短文本的分类效果。

本文提出了一种混合字特征和词特征的深度神经网络中文短文本分类算法（Hybrid Features of Characters and Words CNN，HFCW-CNN），该算法首先训练出短文本的字向量与词向量，然后对字向量与词向量表示的短文本信息进行特征提取，最后将提取的高层特征进行融合，通过全连接层和 softmax 层完成分类任务。本文算法通过使用预训练模型进行文本表示，并对短文本进行字级与词级的多层次表示，在一定程度上缓解了短文本数据表示不充分的问题；同时使用改进的深度神经网络进行特征提取，通过结合两者的优点来解决短文本特征提取不充分的问题。通过解决以上两个短文本分类任务中存在的问题，本文算法在中文短文本分类任务上取得了较好的效果。

2　混合字词特征的深度神经网络模型

本文提出的 HFCW-CNN 模型结构如图 1 所示，该模型主要由编码层、特征提取层和输出层组成。

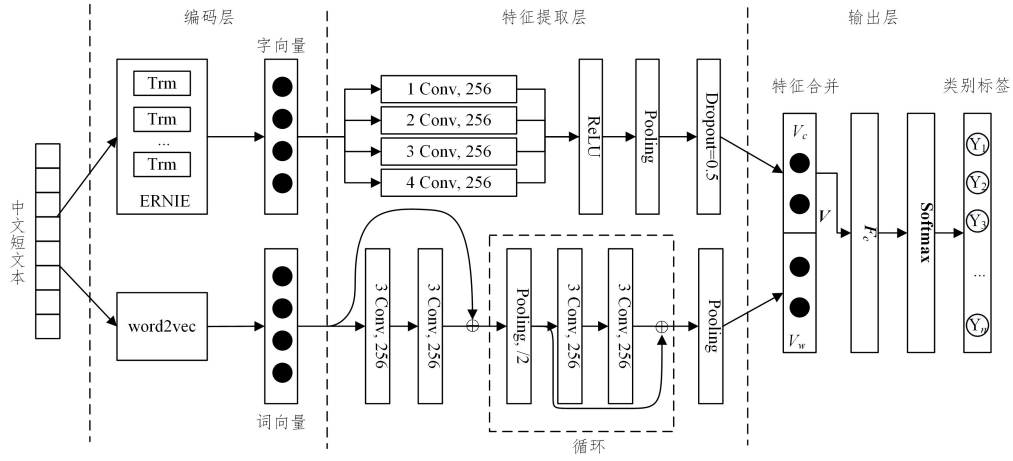


图 1　HFCW-CNN 模型的结构

Fig. 1　Structure of HFCW-CNN model

2.1　编码层

首先，中文短文本数据将进入编码层进行文本表示。为了提高对短文本数据的表示能力，编码层采用字向量与词向量结合的混合表示方法，这里输入的中文短文本内容序列为：

$$X=[x_1,x_2,\cdots,x_n]^{\mathrm{T}} \tag{1}$$

2.1.1　字向量表示

HFCW-CNN 模型使用 ERNIE 对短文本序列进行字向量表示。ERNIE 是一种将 Bert 与知识图谱相结合的增强语言表示模型。其在 Bert 训练的基础上，通过添加结构化的实体信息来增强模型的语言表示能力。基于 ERNIE 的字向量计算流程如图 2 所示。

字向量的计算过程可分为两步，第一步将短文本内容序列 X 输入到由多个 Transformer 模型构成的文本编码器中，编码器会将短文本信息转化成向量；第二步主要是将上一步得到的向量和外部实体信息输入到聚合器中，聚合器首先采用多头注意力机制进行信息提取，然后进行信息融合，最后输出短文本的字向量。

在字向量计算过程中采用了多头注意力机制。多头注意力机制即使用多个自注意力机制进行计算，以获取更多层面的语义信息，然后将每个自注意力机制计算的结果进行拼接

组合，从而得到最终的结果。注意力 *Attention* 的计算式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 代表一个字向量的 3 个矩阵表示，d_k 代表向量的维度。

本文使用 ERNIE 训练字向量时，隐藏层大小设置为 768，隐藏层层数设置为 12，多头注意力机制的头数设置为 12，随机失活率 dropout 为 0.1，词表大小为 20 000。经训练得到的字向量表示为：

$$\boldsymbol{C}=[c_1,c_2,\cdots,c_n]^{\mathrm{T}} \tag{3}$$

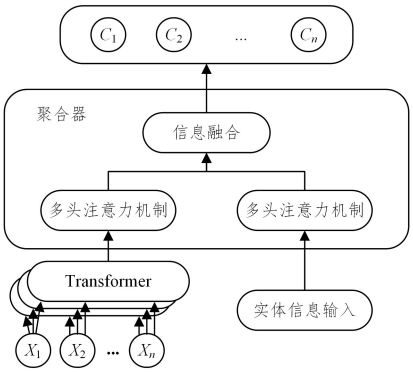


图 2 基于 ERNIE 的字向量表示

Fig. 2 Character vector representation based on ERNIE

2.1.2 词向量表示

本文采用常见的 word2vec 计算短文本的词向量。word2vec 是一种训练词向量的模型，其仅采用单层隐藏层的神经网络，即可将高维、稀疏的 one-hot 形式向量，映射成一个低维、稠密的词向量。word2vec 有两种词向量训练方式，分别是 CBOW 和 Skip-gram。本文采用 Skip-gram 方式进行词向量训练，其结构如图 3 所示。

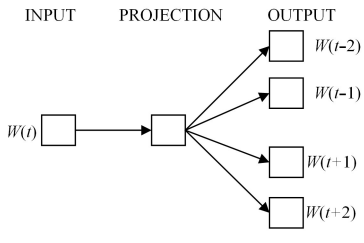


图 3 Skip-gram 模型的结构

Fig. 3 Structure of Skip-gram model

本文使用 word2vec 训练词向量时，词向量的维度设置为 300，窗口大小设置为 5，词频的采样阈值设置为 5，训练迭代次数设置为 8。经训练得到的词向量表示为：

$$\boldsymbol{W}=[w_1,w_2,\cdots,w_n]^{\mathrm{T}} \tag{4}$$

2.2 特征提取层

经过编码层后的字向量与词向量将进入特征提取层。特征提取层主要是完成短文本数据的特征提取工作。为了充分提取短文本数据中的特征信息，本模型采用改进的 TextCNN 和 DPCNN，分别对字向量和词向量表示的短文本内容进行特征提取。

2.2.1 字向量特征提取

本文使用 TextCNN 对字向量表示的短文本数据进行特征提取。TextCNN 主要由卷积层、池化层和非线性激活层组成。

(1)卷积层

卷积层是 TextCNN 的核心部分，本文采用尺寸大小分别为 1，2，3，4 的卷积核对短文本数据进行局部特征提取，卷积过程表示为：

$$\boldsymbol{Vc}_i=f(w*c_{i:i+h-1}+b) \tag{5}$$

其中，$\boldsymbol{Vc}_i$ 代表经过卷积操作得到的第 i 个特征向量，c_i 代表第 i 个输入数据，f 代表激活函数，w 为卷积核权重，b 是偏置参数。

(2)池化层

池化层的任务是对卷积层提取的特征进行再次提取，同时减少训练参数，实现降维效果。池化操作主要分为平均池化和最大池化，本文算法采用最大池化。

(3)非线性激活层

为了强化神经网络的学习能力，需要进入非线性函数进行激活操作。本文算法采用 ReLU 激活函数，表达式为：

$$f(x)=\max(0,x) \tag{6}$$

其中，x 代表输入值。

2.2.2 词向量特征提取

本文使用 DPCNN 对词向量表示的短文本数据进行特征提取。DPCNN 是一种基于词级的深度卷积神经网络，其网络结构简单，可通过增加网络深度来获得最佳精度，同时不会增加过多的计算成本。

DPCNN 主要由卷积层和池化层组成。卷积层采用两个堆叠的尺寸大小为 3 的卷积核进行等长卷积操作；池化层则采用步长为 2 的最大池化操作。本文在词向量特征提取时，首先将编码后的词向量输入到卷积层，然后采用池化与卷积交叉循环的形式进行下采样，完成基于词向量的特征提取工作。

2.3 输出层

经过两个卷积神经网络模型的特征提取，分别得到基于字向量表示的特征向量 $\boldsymbol{V}_c$ 和基于词向量表示的特征向量 $\boldsymbol{V}_w$；然后对两个特征向量进行拼接，得到合并后的特征向量 $\boldsymbol{V}$，最后将特征向量 $\boldsymbol{V}$ 输入到全连接层与 softmax 层，即可得到短文本数据在每个类别上的预测概率 $\boldsymbol{P}$，预测概率值最高的类别即为最终的预测类别，其中 $\boldsymbol{P}$ 的计算式为：

$$\boldsymbol{P}=\text{softmax}[\boldsymbol{W}_{fc}(\boldsymbol{V}_c\oplus\boldsymbol{V}_w)+b_{fc}] \tag{7}$$

其中，$\boldsymbol{W}_{fc}$ 是全连接层的参数矩阵，b_{fc} 是全连接层的偏执量，$\oplus$ 代表拼接操作。

2.4 HFCW-CNN 模型的训练方法

HFCW-CNN 模型训练以中文短文本数据和实验预设参数为输入；然后对数据分别进行字向量编码和词向量编码，并使用神经网络提取数据的字向量特征和词向量特征；最后使用融合的字词特征训练模型来完成中文短文本分类任务。

算法 1 基于 HFCW-CNN 模型的中文短文本分类

输入：中文短文本数据集，实验预设参数

输出：HFCW-CNN 模型，实验数据的分类结果

将实验数据分为 B 个批次，并做如下计算：

For i=0 to B−1：

$\mathbf{C}$=ERNIE. Embedding. from_pretrained(i)；

$\mathbf{W}$=Word2vec. Embedding. from_pretrained(i)；

$\mathbf{V}_{ci}$=TextCNN($\mathbf{C}$)；

$\mathbf{V}_{wi}$=DPCNN($\mathbf{W}$)；

$\mathbf{V}$=torch. cat ($\mathbf{V}_{ci}$,$\mathbf{V}_{wi}$)；

$\mathbf{P}$=softmax($\mathbf{W}_{fc}$V+$\mathbf{b}_{fc}$)；

End For；

Return HFCW-CNN model,$\mathbf{P}$.

不同于由单词构成的英文文本，中文短文本存在字和词两个基本语义维度，单独使用字或词来进行语义信息表达将会丢失部分信息。因此，本文模型对短文本数据进行字级和词级的表示，并在字词两层表示的基础上进行特征提取，从而获取更多的短文本语义信息，提高短文本分类的性能。

3　实验设置

3.1　实验环境

通过实验的方式，对本文算法的可行性与有效性进行验证，实验环境的配置信息如表 1 所列。

表 1　实验环境的配置

Table 1　Experimental environment configuration

Lab Environment	Specific Information
Operating System	Ubuntu 16.04
CPU	Inter Xeon E5-2650 V4
GPU	NVIDIA TITAN Xp
Development Language	Python 3.7
Development Platform	Pytorch 1.6.0

3.2　实验数据

本实验选用文献[22]提供的新闻数据作为实验数据，为了充分检验本文算法的性能，选用了两个实验数据集。

(1)THUCNews 标题数据集

该数据集由 20 万条 THUCNews 新闻标题构成，其中涉及时政、财经、科技、教育、娱乐、游戏、社会、股票、房产和体育 10 个类别，每个类别选取 16000 条新闻标题构建训练集，选取 2000 条新闻标题构建验证集，选取 2000 条新闻标题构建测试集，每条新闻标题的字数约为 25，数据样本如表 2 所列。

表 2　THUCNews 标题的数据样本

Table 2　THUCNews headlines data sample

Text content	Class Label
朱婷回归 女乒夺冠最大障碍已除	体育
贡米与李治廷合拍电影 大赞其英俊帅气	娱乐
i5 芯极致便携 ThinkPad X201i 本 8199 元	科技
手机网游《上古 II》封测日期独家爆料	游戏
天鹅堡多种类型现房房源在售低至 98 折	房产
工行开出首单跨境贸易人民币信用证	财经
专家：详解英美澳留学申请材料和流程	教育
二号限空令威力减弱 股市获超跌买盘支撑	股票
海地同意为流亡 7 年前总统签发护照	时政
货车运送 500 只小猫往广州宰杀食用被查	社会

(2)THUCNews 新闻数据集

该数据集中的数据由新闻标题和新闻内容组成，包含体育、娱乐、财经、房产、游戏、家居、教育、科技、时尚和时政 10 个类别，该数据集共有 65000 条数据，其中 50000 条作为训练集，5000 条作为验证集，10000 条作为测试集。

3.3　对比模型与参数设置

为了检验本文算法与传统分类模型在性能上的优劣，本文进行了对比实验，对比模型的具体信息如表 3 所列。

表 3　对比模型信息

Table 3　Comparison model information

Algorithm
Word2vec+TextCNN[16]
Word2vec+DPCNN[17]
Word2vec+BiGRU[23]
Word2vec+BiGRU_Att[24]
Bert+softmax[12]
ERNIE+softmax[14]
ERNIE_BiGRU[25]

本实验的主要参数设置如表 4 所列，在 THUCNews 标题数据集上，pad_size 值设置为 32；在 THUCNews 新闻数据集上，pad_size 值设置为 200。为减小模型过拟合的风险，当模型连续 2000 个 batch 的训练效果无明显提升时，提前终止训练。

表 4　主要实验参数

Table 4　Main experimental parameters

Parameter	Value
batch_size	128
learning_rate	5×10^{-5}
dropout	0.5
TextCNN_filter_sizes	1,2,3,4
DPCNN_filter_sizes	3
filter_num	256
hidden_size	756
epochs	20

3.4　评价指标

本实验采用精确率(Precision，P)、召回率(Recall，R)和综合评价指标 $F1$ 值来衡量分类模型的性能，其计算式为：

$$P=\frac{TP}{TP+FP}*100\% \tag{8}$$

$$R=\frac{TP}{TP+FN}*100\% \tag{9}$$

$$F1=\frac{2*P*R}{P+R}*100\% \tag{10}$$

其中，TP(True Positive，真阳性)表示预测是正类，实际也是正类的文本个数；FP(False Positive，假阳性)表示预测是正类，实际是负类的文本个数；TN(True Negative，真阴性)表示预测是负类，实际也是负类的文本个数；FN(False Negative，假阴性)表示预测是负类，实际是正类的文本个数。

4　实验结果分析

为了综合评估本文算法与常用分类模型性能的优劣，分别在 THUCNews 标题数据集和 THUCNews 新闻数据集上进行了实验，同时选用精确率、召回率和 $F1$ 值作为评价指标。

(1)在 THUCNews 标题数据集上的结果分析

在 THUCNews 标题数据集上进行实验，选用前述表 3 所列的模型作为对比，其中对比模型的实验数据请参考文献[15]。实验结果如表 5 所列。

表 5 在 THUCNews 标题数据集上的实验结果

Table 5 Experimental results on the THUCNews headlines data set

（单位：%）

Algorithm	P	R	F1
Word2vec+TextCNN	90.88	90.80	90.82
Word2vec+BiGRU	90.49	90.47	90.45
Word2vec+BiGRU_Attention	90.48	89.47	90.42
Bert+softmax	92.64	92.58	92.60
ERNIE_BiGRU	94.32	94.12	94.22
HFCW-CNN	**94.82**	**94.79**	**94.80**

由表 5 可知，在 THUCNews 标题数据集上，本文算法在 3 类评价指标上均优于对比模型。其中，在综合性评价指标 $F1$ 值上，本文算法相比 TextCNN，BiGRU，BiGRU_Attention，Bert 和 ERNIE_BiGRU 模型，分别提升了 3.98%，4.35%，4.38%，2.2%和 0.58%。从 $F1$ 值的提升幅度可以看出，采用预训练模型 Bert 和 ERNIE 作为输入的模型均优于使用 Word2vec 模型作为输入的模型，这也证明预训练模型的文本表示能力优于传统的词嵌入模型。本文算法通过结合预训练模型 ERNIE 和 Word2vec 模型各自的优点，对短文本内容进行字维度和词维度双层表示，能够提升模型对短文本数据的表示能力，从而有助于短文本特征的提取。

为了对比本文算法在每类新闻标题数据上的分类效果，绘制了本文算法在 10 类新闻标题数据上的精确率、召回率和 $F1$ 值分布，如图 4 所示。

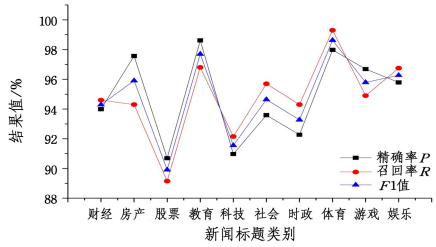


图 4 本文算法在 10 类新闻标题数据上的分类效果

Fig. 4 Classification effect of this algorithm on ten types of news headlines data

由实验数据可知，本文算法在房产、教育、体育和娱乐 4 类新闻标题数据上的分类效果较好，在股票类新闻标题数据上的分类效果最差，其原因是股票类新闻与财经、科技和时政类新闻的相关性强，导致分类难度较大。

(2)在 THUCNews 新闻数据集上的结果分析

在 THUCNews 新闻数据集上进行的实验，选用 TextCNN，DPCNN 和 ERNIE 模型作为对比模型，实验结果如表 6 所列。

表 6 在 THUCNews 新闻数据集上的实验结果

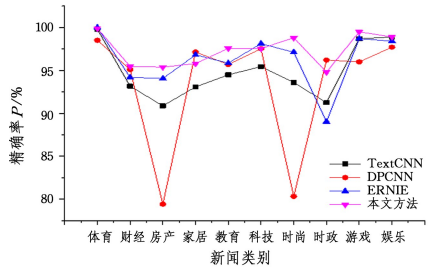
Table 6 Experimental results on the THUCNews news data set

（单位：%）

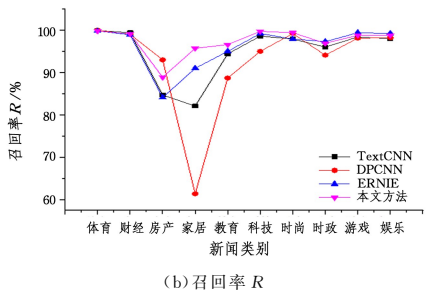
Algorithm	P	R	F1
Word2vec+TextCNN	94.93	94.82	94.87
Word2vec+DPCNN	93.37	91.53	92.44
ERNIE+softmax	96.23	96.06	96.14
HFCW-CNN	**97.37**	**97.34**	**97.35**

由表 6 可知，在 THUCNews 新闻数据集上，本文算法在 3 类评价指标上相比对比模型均有所提升。其中，相比 TextCNN 模型，本文算法的精确率、召回率和 $F1$ 值分别提升了 2.44%，2.52%和 2.48%。相比 DPCNN 模型，本文算法的精确率、召回率和 $F1$ 值分别提升了 4.0%，5.81%和 4.91%。相比 ERNIE 模型，本文算法的精确率、召回率和 $F1$ 值分别提升了 1.14%，1.28%和 1.21%。

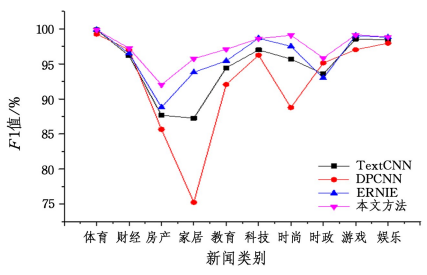
为了对比上述模型在每类新闻数据上的分类效果，绘制了各个模型在 10 类新闻数据上的精确率、召回率和 $F1$ 值的分布结果，如图 5 所示。



(a)精确率 P



(b)召回率 R



(c)$F1$ 值

图 5 各个模型在 10 类新闻数据上的分类效果

Fig. 5 Classification effect of each model on ten types of news data

由图 5(a)可知，本文算法在多类新闻数据上的精确率均高于对比模型，特别是在精确率普遍较低的房产类新闻上，取得了 4 类模型中的最高分值。在召回率评价指标上，由图 5(b)可知，本文算法在 10 类新闻数据上的成绩均比较优异，并在召回率普遍偏低的家居类新闻上取得了最高分值。由图 5(c)可知，在综合性评价指标 $F1$ 值上，本文算法在各类新闻数据上均取得较高分值。由此可知，本文算法的分类效果优于对比模型。

本文提出了基于 HFCW-CNN 的中文短文本分类算法，通过使用字向量和词向量多层次的文本表示作为输入，增强了模型对中文短文本数据的表示能力，同时结合 TextCNN 和 DPCNN 各自的优点，提升了模型对中文短文本数据的特征提取能力。该算法可以有效地提高中文短文本分类任务的效果。

结束语 针对中文短文本分类任务存在的特征信息难以提取与表示的问题，本文提出了基于 HFCW-CNN 的中文短文本分类算法。首先，该算法采用字向量与词向量结合的方式，对短文本数据进行多层次的向量表示；然后使用 Text-

CNN和DPCNN分别对字向量与词向量表示的短文本信息进行特征提取；最后将两个神经网络模型提取到的高层特征进行融合，并通过 softmax 层完成分类任务。实验结果表明，本文算法能够在短文本分类任务上取得良好的效果。

目前，在实际项目开发中，采集到的数据几乎都是无标签数据，然而利用人工的方式进行数据标注，不仅费时费力，而且准确率会受到人为因素的影响。接下来将利用半监督学习算法对大量的无标签数据进行自动标注，结合本文的优化算法提高中文短文本分类算法的效果。

参 考 文 献

[1] SHI H M. Research on Social Network Information Filtering Method Based on Long Short-term Memory Network [D]. Nanjing University of Posts and Telecommunications, 2019.

[2] ZHAO J Q. Research on Internet Public Opinion Monitoring Method Based on Automatic Classification[J]. Software Guide, 2016, 15(3): 133-135.

[3] WU S, GAO M, XIAO Q, et al. A topic-enhanced recurrent autoencoder model for sentiment analysis of short texts[J]. International Journal of Internet Manufacturing and Services, 2020, 7(4): 393-399.

[4] CHEN H. Personalized recommendation system of e-commerce based on big data analysis[J]. Journal of Interdisciplinary Mathematics, 2018, 21(5): 1243-1247.

[5] TAN C. Short Text Classification Based on LDA and SVM[J]. International Journal of Applied Mathematics & Stats, 2013, 51(22): 205-214.

[6] YIN C, SHI L, WANG J. Short Text Classification Technology Based on KNN+Hierarchy SVM[C]// International Conference on Multimedia and Ubiquitous Engineering International Conference on Future Information Technology. 2017: 633-639.

[7] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning based text classification: A comprehensive review[J]. arXiv: 2004. 03705, 2020.

[8] LI C B, DUAN Q J, JI C H, et al. Method of Short Text Classification Based on CHI and TF-IWF Feature Selection[J]. Journal of Chongqing University of Technology(Natural Science), 2021, 35(5): 135-140.

[9] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[C]// Proceedings of the International Conference on Learning Representations. ACM, 2013: 1-8.

[10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrasesand their compositionality [C]// Advances in Neural Information Processing Systems. 2013: 3111-3119.

[11] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational. 2018: 2227-2237.

[12] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceeding of the 2019 Conference of the North American Chapter of the Association for Computational Linguistices (NAACL). 2019: 4171-4186.

[13] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[EB/OL]. (2019-09-26)[2020-01-06]. https://arxiv.org/abs/1909.11942.

[14] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: enhanced language representation with informative entities[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 1441-1451.

[15] SUN Y, WANG S, LI Y, et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8968-8975.

[16] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Association for Computational Linguistics. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746-1751.

[17] JOHNSON R, ZHANG T. Deep Pyramid Convolutional Neural Networks for Text Categorization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 562-570.

[18] LI Z J, GENG C Y, SONG P. Research on Short Text Classification Based on LSTM-TextCNN Joint Model[J]. Journal of Xi'an Technological University, 2020, 40(3): 299-304.

[19] DUAN D D, TANG J S, WEN Y, et al. Research on Chinese Short Text Classification Algorithm Based on BERT[J]. Computer Engineering, 2021, 47(1): 79-86.

[20] ZHENG C, HONG T T, XUE M Y. BLSTM_MLPCNN Model For Short Text Classification [J]. Computer Science, 2019, 46(6): 206-211.

[21] HOU X L, LI X, CHEN Y P. Short Text Classification Model Based on Multi-Neural Network Hybrid[J]. Computer System Applications, 2020, 29(10): 9-19.

[22] SUN M S, LI J Y, GUO Z P, et al. THUCTC: An efficient toolkit for Chinese text classification [EB/OL]. http://thuctc.thunlp.org. 2016-12-30.

[23] HU D F, ZHANG C X, WANG S T, et al. Intelligent Prediction Model of Tool Wear Based on Deep Signal Processing and Stacked-ResGRU[J]. Computer Science, 2021, 48(6): 175-183.

[24] WANG W, SUN Y X, QI Q J, et al. Text sentiment classification model based on BiGRU-attention neural network[J]. Application Research of Computers, 2019, 36(12): 3558-3564.

[25] LEI J S, QIAN Y. Chinese text classification method based on ERNIE-BiGRU model[J]. Journal of Shanghai Electric Power University, 2020, 36(4): 329-335, 350.

LIU Shuo, born in 1996, postgraduate. His main research interests include data analysis, natural language processing and short text classification.

WANG Geng-run, born in 1987, Ph.D, assistant researcher. His main research interests include telecommunication network security and data processing.

（责任编辑：李亚辉）