DOI:10.19651/j.cnki.emt.2211199

基于融合 LDA 与双层 CNN 的文本分类研究*

杨 秀[1] 刘胜全[2] 贾李睿智[2] 解舒淇[2]
(1.新疆大学软件学院 乌鲁木齐 830046; 2.新疆大学信息科学与工程学院 乌鲁木齐 830046)

摘 要:针对基于主题的文本分类任务存在的主题特征表征能力不足、数据高维导致的特征维度高等问题,本文对输入的特征表示与卷积神经网络结构(CNN)做出了改进。在特征表示时提出了使用 LDA 模型计算逆主题空间频率从而得到文本的主题向量矩阵,降低了噪声主题的特征表达,增强了关键主题的权重;分别将文本的主题向量矩阵与词向量矩阵作为 CNN 模型的输入。提出了双层 CNN 网络结构,在每层 CNN 的池化层后增加一层多通道池化层,以融合每层 CNN 的池化结果,降低特征维度的同时获取更多的局部显著特征;最后使用 Attention 机制对融合的特征进行加权后输入到全连接层进行分类。由实验结果可知,改进的模型在文本分类任务上的准确率、召回率均在 98%以上,F1 值较基准实验提高了近 6%。

关键词:LDA;双层 CNN;Attention;文本分类
中图分类号:TP391 **文献标识码**:A **国家标准学科分类代码**:520.2020

Research on text classification based on fusion of LDA and two-layer CNN

Yang Li[1] Liu Shengquan[2] Jia Liruizhi[2] Xie Shuqi[2]
(1. School of Software, Xinjiang University, Urumqi 830046, China;
2. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: Aiming at the problems of insufficient topic feature representation ability and high feature dimension caused by high dimensionality of data in topic-based text classification task, this paper improves the input feature representation and Convolutional Neural Networks (CNN) structure. In the feature representation, LDA model is proposed to calculate the inverse topic space frequency to get the topic vector matrix of text, which reduces the feature representation of noisy topic and enhances the weight of key topic. The topic vector matrix and word vector matrix of the text are respectively used as the input of CNN model. A two-layer CNN network structure is proposed, and a multi-channel pooling layer is added after the pooling layer of each layer of CNN to integrate the pooling results of each layer of CNN, and obtain more local salient features while reducing the feature dimension. Finally, the Attention mechanism is used to weight the fused features and input them to the fully connected layer for classification. According to the experimental results, the accuracy and recall of the improved model in text classification tasks are above 98%, and the F1 value is increased by 6% compared with the benchmark experiment.

Keywords: LDA; two-layer CNN; Attention; text classification

0 引 言

随着信息技术的飞速发展与互联网时代的到来,新闻、微博等信息的交流与传播使得每天产生了数以万计的文本信息。这些文本信息包含了情感、用户画像、热点话题等,因此对文本进行分类有利于对海量文本进行特定的处理分析从而进行舆情预警等活动。

潜在狄利克雷(latent Dirichlet allocation,LDA)[1-2]主题模型是机器学习中典型的文本主题分类算法,同时也是常用的特征表示方法。使用 LDA 表示文本特征的研究中,Li 等[3]提出将主题-词分布考虑进 Word2Vec 训练时的最大似然公式中,得到的 Topic2Vec 学习了主题表示和单词表示。张标[4]将 TF-HF-IDF 与 LDA 结合提取特征,但依赖于 LDA 进行聚类,具有一定的局限性。颜端武等[5]将文档-主题分布与加权 Word2Vec 词向量结合,构建了微博短文本的融合特征表示。但表示的特征仅仅将文档-主题分布嵌入到文本表示矩阵中,缺乏词向量与主题概率分布的

收稿日期:2022-08-28
* 基金项目:新疆维吾尔族自治区教育厅重点基金(XJEDU2019I004)项目资助

进一步融合表示。席笑文等[6]使用文档-主题分布和主题-词分布构建专利权人-专利-技术主题三层概率分布进行技术相似可视化研究，但是仅仅使用三层概率分布的数值评判相似度缺乏了对潜在概率分布的考量。

近年来，深度学习算法在文本分类中得到了更多的关注。Zhang 等[7]提出了将神经主题模型与 DCNN 模型结合进行细粒度特征的提取，可以高效地识别出可解释的主题信息从而提高文本表示能力。Kim[8]改进了 CNN 的卷积核将之应用到自然语言处理领域，在个别文本分类任务中刷新了当时最好的记录。杨兴锐等[9]将卷积神经网络(convolutional neural networks，CNN)与带有残差网络的 BiLSTM 模型融合，使模型学习到残差信息，但只是简单的将两个模型拼接而无法评估提取的特征的重要程度。王根生等[10]使用 TF-IDF 对词向量进行加权后再输入到 CNN 中进行分类，考虑了词在类中的分布概率增强了分类的准确度，却忽略了文本主题信息和词性信息。冼广铭等[11]简单的使用 LDA 获取主题信息并输入到 BiGRU 模型中但并未考虑提取的主题信息的完整性，分类的精度仍有进步的空间。邓维斌等[12]融合了多种神经网络使得分类结果有了显著提高，但是割裂了 CNN 与 Attention，无法使提取的典型特征有更高的权重。

针对以上问题，为了在文本分类任务中增加主题特征的表达，使用 LDA 计算逆主题频率，最后计算文本的主题向量矩阵，将主题特征考虑进文本分类中同时降低了噪声主题与普通主题的权重。在使用 CNN 模型进行分类时，分别将文本的主题向量矩阵与词向量矩阵输入到两层 CNN 中，增加一层多通道池化层用以融合每层的池化结果，从而提取更显著的特征。

1　本文模型

1.1　算法流程

本文在特征表示时，使用主题-词分布与 Word2Vec 词向量结合构建主题向量；使用主题空间熵计算主题的逆主题空间频率，考虑了主题在空间的分布对特征表示的影响；将主题的逆主题空间频率与文档-主题分布共同计算得到主题在文本中的总权重；最后将主题的总权重与主题向量结合构建文本的主题向量矩阵，降低了噪声主题和普通主题的权重，增强了重要主题的特征在文本中的表达。使用 CNN 进行文本分类时，将文本的词向量矩阵与主题向量矩阵分别输入在 CNN 模型中；在两层的池化结果后再加一层池化用于融合特征表示；融合的特征输入到 Attention 层进行特征的加权；最后在全连接层使用 softMax 函数进行分类。本模型的流程如图 1 所示。

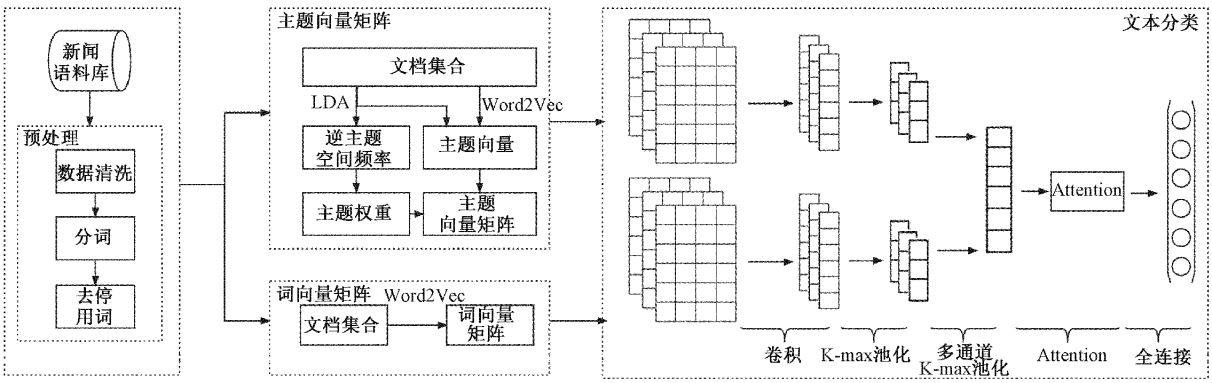

图 1　模型流程

1.2　词向量矩阵

模型使用 Word2Vec 技术生成词向量。本文使用 Word2Vec 技术中的 skip-gram 模型对文档集合 $D=\{d_1,d_2,\cdots,d_n\}$ 生成包含所有词的词向量集合 $V_w=\{v_{w_1},v_{w_2},\cdots,v_{w_v}\}$。最后将每篇文本中词的词向量使用矩阵进行储存，得到每篇文本的词向量矩阵 $\boldsymbol{WV}_{j_{n*k}}=\{WV_0,WV_1,\cdots,WV_n\}_{n\times k}$。

1.3　主题向量矩阵

模型在生成词向量、文档-主题-词分布之后，计算得到主题向量和主题空间熵；然后使用主题空间熵计算主题的逆主题空间频率，再计算主题在单个文本中的总权重；最后使用主题总权重与主题向量结合构建文本的主题向量矩阵。生成主题向量矩阵过程如图 2 所示。

本节使用到的符号如表 1 所示。

1)生成文档-主题-词分布

LDA 分布主题模型是一个三层贝叶斯分层模型。使用 LDA 主题模型计算得到文档-主题分布($p(z|d)$)、主题-词分布($p(w|z)$)。

2)计算主题向量

为每个主题选取概率最大的 topH 的词，计算每个词在其主题下所占的归一化权重。如式(1)所示。

$$\theta_{w_{t,i}}=\frac{p(w_i\mid z_t)}{\sum_{j}^{H}p(w_j\mid z_t)} \tag{1}$$

其中，$\theta_{w_{t,i}}$ 为单词 i 在主题 t 下的归一化权重，$p(w_i\mid z_t)$ 表示在主题 t 中第 i 个词的概率。

一个词在主题中概率分布越大，该词就越能够表征该主题的主题信息，也就应赋予该词更高的权重[13]。将

式(1)计算得到每个词在主题下的归一化权重结合 Word2Vec 得到的词向量通过加权求和得到主题向量,如式(2)所示。

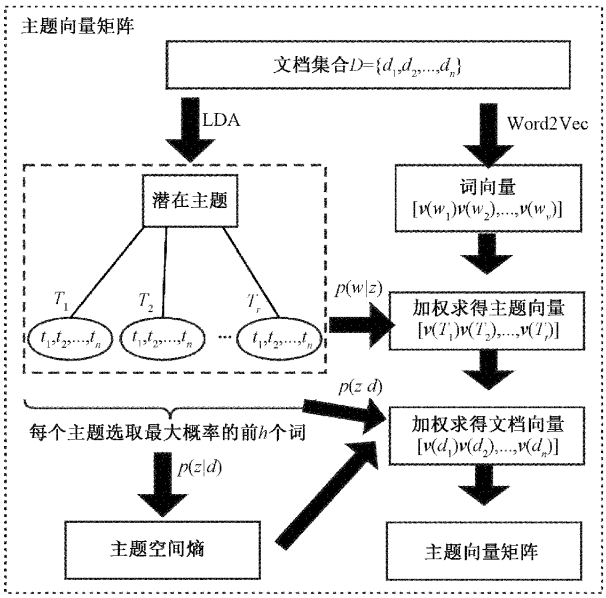

图 2　主题向量矩阵过程

表 1　本节符号及说明

符号	说明
$p(z\|d)$	文档-主题分布
$p(w\|z)$	主题-词分布
$\theta_{wi,j}$	词 i 在主题 j 下的归一化权重
V_{topic}	主题向量集合
$S(z_{space})$	主题空间熵集合
$WV_{d_i\ \ topic_t}$	主题 t 在文档 i 的总权重
$\boldsymbol{TopicMV}_{j_{m\times k}}$	文档主题矩阵

$$V_{topic_t}=\sum_{i}^{H}\theta_{w_{t,i}}\times V_{W_i} \tag{2}$$

3)生成主题向量矩阵

(1)计算主题空间熵

Shannon 提出信息熵的概念,将信息中去掉冗余且驳杂的信息后得到的平均信息量定义为"信息熵"。使用信息熵表示信息的不确定程度。信息的熵值越大,信息分布越均匀,信息的不确定程度就越大,将其弄明白所需要的信息量也就越大。数学公式如式(3)所示。

$$H(x)=-\sum p_{(x)}\log p_{(x)} \tag{3}$$

其中,x 为随机变量的值,$p(x)$表示 x 的概率函数。

李彦飞[14]利用 $p(w|z)$和 $p(z|d)$从主题完整性、主题的空间差异、主题的时间差异这 3 个方面来研究主题的质量。本文在此基础上认为主题的空间差异由主题空间熵度量。主题的空间熵越大,则该主题在空间中分布越均匀,不确定程度越大,该主题可能是噪声主题或普通主题。比如北京奥运赛事的新闻中大多数都有关于"奥组委"的字眼或主题出现,但是"奥组委"这一主题并不能表征足球、游泳等新闻,应该以较大概率认定为普通主题。主题空间熵定义如式(4)所示。

$$S(z_{space_t})=-\sum_{d\in D}p(d\mid z_t)\times\log(p(d\mid z_t)) \tag{4}$$

其中,$p(d\mid z)$是主题 z 条件下文本 d 出现的概率,由贝叶斯公式得出:

$$p(d\mid z_t)=\frac{p(z_t\mid d)\times p(d)}{p(z_t)}=\frac{p(z_t\mid d)\times p(d)}{\sum_{d\in D}p(z_t\mid d)\times p(d)} \tag{5}$$

(2)计算主题所占权重

一个主题在文档中概率越大,说明该主题越能表征该文档。所以该主题在文档中重要性与该主题在文档中的概率成正比;而一个主题在主题空间中分布越均匀说明该主题可能是噪声主题或普通主题,那么该主题在文档中重要性就要相应降低,则该主题在某文档中重要性与该主题在空间中分布均匀程度成反比,本文使用主题空间熵来度量主题在主题空间中分布均匀的程度。

通过计算某主题的空间熵值占总空间熵值的比例来表示该主题在主题空间所占的权重,如式(6)所示。

$$p_{Space_t}=\log(1/\frac{S(z_{space_t})+1}{\sum_{i}^{T}S(z_{space_i})}) \tag{6}$$

其中,p_{Space_t} 是主题 t 在空间所占权重,分子加 1 是为了防止公式运算无意义的情况发生。

本文使用 LDA 输出的文档-主题分布($p(z|d)$)来表示主题在文档中所占权重。

因为主题在文档的重要性与主题在文档中所占权重成正比,与主题在主题空间所占的权重成反比使用式(7)计算主题在文档中所占权重。

$$WV_{d_i-topic_t}=p(z_t\mid d_i)\times p_{space_t} \tag{7}$$

其中,$WV_{d_i\ \ topic_i}$ 是文档 i 中主题 t 所占的权重,$p(z_t\mid d_i)$是主题 t 在文档 i 中的概率。

(3)生成主题向量矩阵

通过式(2)求得主题向量和式(7)求得主题所占的权重后,将文档中每个主题的总权重乘以对应主题的主题向量,再使用矩阵存储,可以得到文本的主题向量矩阵,公式如式(8)所示。

$$\boldsymbol{TopicMV}_{j_{m\times k}}=\begin{bmatrix}WV_{d_j\ \ tpoic_0}\times V_{topic_0}\\WV_{d_j-tpoic_1}\times V_{topic_1}\\\vdots\\WV_{d_j-tpoic_m}\times V_{topic_m}\end{bmatrix}_{m\times k} \tag{8}$$

1.4　Double-Layer CNN

为了将主题信息融合进文本分类任务中,本文提出了两层 CNN 网络结构,增加多通道池化层提取两层 CNN 得到的特征,以获取更多的典型特征。网络结构如图 3 所示。

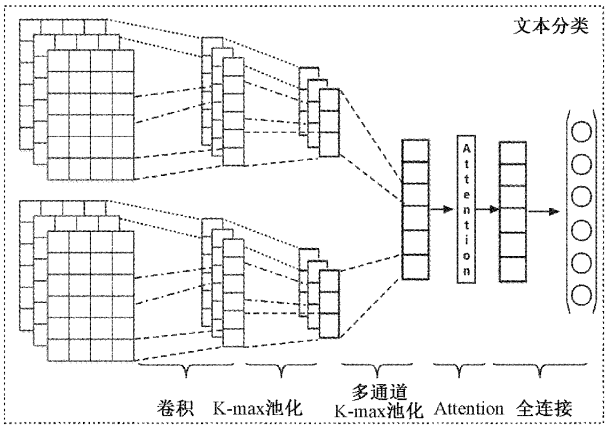

图 3　Double-Layer CNN 网络结构

网络结构分为输入层、卷积层、池化层、多通道池化层、Attention 层、全连接层。

输入层使用分别在 2.2 节、2.3 节得到的文本词向量矩阵与主题向量矩阵作为神经网络模型的输入。其中因为每个文本的词数不一，所以词向量矩阵使用 zero-padding 做填充。

卷积层中卷积核尺寸需要适应文本的处理，所以按照 Kim 提出的 textCNN[8] 模型设置卷积核尺寸。设置 3 个不同大小的卷积核分别为 2、3、4。通过卷积核卷积得到 feature map，特征计算公式如式(9)所示。

$$\sum_{1:n-h+1} c_i = (w \times T_{i:i+h-1} + b) \tag{9}$$

其中，w 为卷积核的参数，$T_{i:i+h-1}$ 是输入层的输入的矩阵第 i 行到 $i+h-1$ 行，经卷积得到了如式(10)的特征矩阵。

$$\boldsymbol{C} = f((n-h+1) \times k + b) \tag{10}$$

其中，f 为激活函数，$f(x)=\max(0,x)$，b 为偏置项。

池化层对每个通道进行池化，使用 K-max 池化方式。卷积层得到的 feature map 是大小为 $1\times(i-h+1)$ 的矩阵，经过池化后每个通道得到的是 $1\times(i-h+1)$ 维的特征向量。

多通道池化层对两层 CNN 得到所有特征向量使用 K-max 池化，得到的特征向量维度是 $1\times(2n\times k)$，其中 n 为每层 CNN 中的通道数。

Attention 层中 Attention 机制本质是学习出一个权重分布，再用学习出来的权重分布与特征进行乘积加权。在文本分类中，权重越大的特征越重要，也越能代表文本从而有利于进行文本的区别分类。计算出的权重注意力得分使用 e_i 表示。计算公式如式(11)所示。

$$e_i = \tanh(\boldsymbol{w}_i C_i + b_i) \tag{11}$$

其中，tanh 是激活函数，$\boldsymbol{w}_i$ 是权重矩阵，b_i 是偏置项。

使用 softmax 函数计算注意力权重得分得到权重向量 $\boldsymbol{a}_i$。将归一化得到的权重 $\boldsymbol{a}_i$ 与多通道池化层得到的特征向量乘积得到加权的特征向量。

全连接层使用 softmax 计算文本的类别进行分类。损失函数选择交叉熵损失函数，使用 Adam 优化模型参数，多轮迭代训练直到模型的损失值等达到收敛状态。

2　实验结果及评估

2.1　实验环境及数据集

本文的实验环境为：操作系统为 64 位 Windows10 系统；处理器为 Inter Core i7 9th；运行内存为 12 G；CPU 为 2.3 Hz。开发工具为：Pycharm 社区版 2021.2.3；python3.6；TensorFlow1.14.0；使用 gensim 库提供的 LDA 主题模型训练主题，提供的 Word2Vec 模型训练词向量；分词工具使用 jieba0.42。

数据集使用 THUCNews。THUCNews 是清华大学自然语言处理实验室构建的中文文本分类数据集，数据集中的数据种类齐全共计 14 个类别，74 万条新闻数据，且均使用 UTF-8 格式存储。本文选取 6 个类别：体育、星座、房产、科技、家居、游戏，从 6 类中随机选取筛选 12 000 条新闻数据，筛选过的每篇新闻数据的词数控制在[50,1 000]。数据集信息如表 2 所示。本文随机算选取数据集的 80% 数据作为训练集，20% 的数据作为测试集。

表 2　数据集信息

类别	体育	星座	房产	科技	家居	游戏
篇数	2 000	2 000	2 000	2 000	2 000	2 000

2.2　评价指标

分类算法常用的指标有精确率、召回率、F 值、纯度、兰德系数和调整兰德系数。本文实验使用准确率(precision，P)、召回率(recall，R)和 F1 值评价分类的效果。P 是分类预测为真且实际也为真的数据占所有预测为真的数据的比例；R 是分类预测为真且实际也为真的数据占实际为真数据的比例；F1 值是准确率与召回率的调和平均值。公式定义如式(12)～(14)所示。

$$P = \frac{n_{i,j}}{|N_j|} \tag{12}$$

$$R = \frac{n_{i,j}}{|M_i|} \tag{13}$$

$$F1 = \frac{2P \times R}{P + R} \tag{14}$$

2.3　模型参数

词向量训练时使用 skip-gram 模型训练，词向量维度 100 维，窗口数为 5，选取前 $topH=50$ 的词构成主题向量。LDA 模型训练主题时为了得到每篇文档对于所有主题的分布，设置 LDA 参数中 minimum_probability=0；对于 LDA 主题模型主题数的选择使用交叉验证法选取最适合的主题数，使用 F1 值来选取合适的主题数，令主题数取 6、11、16、21、…、91、96 这 19 个值，通过实验观察在不同主题数下训练的 LDA 主题模型对分类评价指标 F1 影响的变

化选取合适的主题数区间,如图 4 所示主题数在 66 左右时分类的 F1 值最大,故将 LDA 主题模型的主题数设置为 66 训练得到主题向量矩阵。

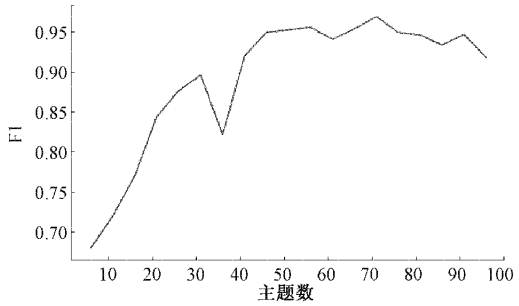

图 4　主题数与 F1 值关系

模型使用 CNN 结构结合 Attention 机制,需要设置一系列深度学习神经网络的参数,参数设置如表 3 所示。

表 3　参数设置

参数	设置	参数	设置
卷积核窗口数	2、3、4	dropout	0.5
卷积核个数	256	学习率	1×10^{-3}
池化方式	1－max	attention_size	50
全连接神经元数	128	激活函数	ReLu
批训练(btch_size)	64	优化器	adam
迭代(epoch)	10	损失函数	交叉熵

2.4　实验结果及分析

1)本文模型实验

本文提出的模型使用训练集进行训练,然后使用测试集进行验证。分类的结果使用 P、R 和 F1 值评价,并对每个指标取平均值。实验结果如表 4 所示。

表 4　本文模型实验结果　%

类别	P	R	F1
体育	98.3	99.4	98.8
星座	97.6	98.4	98.0
房产	98.6	97.9	98.2
科技	99.1	98.9	99.0
家居	97.0	97.8	97.4
游戏	98.7	98.5	98.6
平均	98.2	98.5	98.3

由表 4 可以看出,本文提出的模型在数据集中的类别数据中表现良好,模型的 P、R、F1 的平均值分别达到了 98.2%、98.5%、98.3%。说明提出的模型在文本分类任务上有良好的应用价值。

2)验证模型有效性

为了验证本文提出的模型有效性,将在同等实验环境下与以下基准模型进行对比。实验结果如表 5 所示。

(1)textCNN[8]:首次将 CNN 应用到文本分类领域,改变卷积核的尺寸使之适应文本矩阵。

(2)CTMWT[10]:基于类频方差改进 TF-IDF 算法计算词语权重从而构建词向量矩阵,最后使用 CNN 进行文本分类。

(3)FMNN[12]:使用 BERT 进行预训练,融合 CNN、BiLSTM-Attention 模型进行文本分类。

(4)MCA-CL[15]:将 CNN-Attention 与 BiLSTM-Attention 融合用于文本分类。

(5)CNN-Att:使用词向量矩阵输入,CNN-Attention 进行文本分类。

表 5　模型有效性实验结果

类别	P/%	R/%	F1/%	时间/h
textCNN	92.1	92.6	92.3	1.1
CTMWT	94.9	93.6	94.2	1.6
FMNN	97.6	98.0	97.8	2.1
MCA-CL	98.4	97.9	98.1	1.9
CNN-Att	94.3	95.8	95.0	1.3
本文模型	98.2	98.5	98.3	1.5

由表 5 可知,本文提出的模型的准确率、召回率、F1 值均优于 textCNN、CTMWT、FMNN、MCA-CL 和 CNN-Att 模型。相较于 textCNN,F1 值提高了 6%。因为本文模型将主题特征考虑进文本分类中,增加了文本的特征表达,更加有利于分类器对文本的区分;相较于 CTMWT,F1 值提高了 4.1%。因为虽然 CTMWT 改进了 TF-IDF 算法将词语在不同类别的分布情况考虑进特征中,但是 CNN 在提取特征时会提取典型的特征同时使用 K-max 池化方式,所以改进了 TF-IDF 算法对结果的增幅会被舍弃一部分,所以效果没有本文模型好;相较于 CNN-Att,F1 值提高了 3.3%。因为本文模型提出了逆主题空间频率将主题特征考虑进文本分类的同时降低了普通主题与噪声主题的权重,且跨通道池化层能获取更多的典型特征,所以分类结果更好;相较于 FMNN,F1 值提高了 0.5%;相较于 MCA-CL,F1 值提高了 0.4%。与基准实验对比分析之后可以得到结论:本文提出的模型在文本分类任务上有良好的准确性和有效性。

同时从表 5 最后一列的时间开销可以看出,本文模型相较于 CTMWT、FMNN、MCA-CL 模型有更低的时间消耗;相较于 textCNN 与 CNN-Att 模型的时间开销大但准确率等指标优于两个模型,上述可以说明本文模型有良好的应用前景。

3)消融实验

为了验证本文提出的模型的各部分对分类效果的增益,与如下组件进行了消融实验。实验结果如表 6 所示。

(1) W2V-CNN-Att：使用词向量矩阵输入，CNN-Attention 进行文本分类。

(2)LDA-CNN-Att：使用 LDA 模型计算得到主题向量矩阵输入，不考虑逆主题空间频率，CNN-Attention 进行文本分类。

(3)W2V-L-CA：如本文模型一样处理，但是将跨通道卷积层去掉，使用 Attention 机制对各层池化结果进行加权。

(4) W2V-L-CNN：如本文模型一样处理，去掉 Attention 机制。

表 6　消融实验结果　%

类别	*P*	*R*	F1
W2V-CNN-Att	94.3	95.8	95.0
LDA-CNN-Att	94.9	94.6	95.2
W2V-L-CA	96.1	96.7	96.4
W2V-L-CNN	94.7	95.3	95.0
本文模型	98.2	98.5	98.3

由表 6 可知，本文模型相较于 W2V-CNN-Att 模型，F1 值提高 3.3%，说明本文将主题特征考虑进文本分类中是明显可行有效的；本文模型相较于 LDA-CNN-Att 模型，F1 值提高 3.1%，说明考虑了逆主题空间频率来计算主题特征的权重，达到了降低普通主题和噪声主题的权重的效果；本文模型相较于 W2V-L-CA 模型，F1 值提高 1.9%，说明跨通道池化层能有效提取更多的典型特征用于文本分类。

3　结　　论

本文将主题特征考虑进文本分类任务中，同时提出逆主题空间频率降低普通主题与噪声主题的权重。使用两层 CNN 进行文本分类任务，提出跨通道的池化层融合两层 CNN 的结果用于分类，在清华大学实验的数据集 THUCNews 上有良好的表现。未来的工作主要集中在将 BiLSTM 等神经网络融合进模型中进一步提升分类性能。

参考文献

[1] HUANG Y, WANG R, HUANG B, et al. Sentiment classification of crowdsourcing Participants' reviews text based on LDA topic model[J]. IEEE ACCESS, 2021:128-136.

[2] SHAO D, LI C, HUANG C, et al. The short texts classification based on neural network topic model[J]. Journal of Intelligent & Fuzzy Systems, 2022, 42(3): 114-126.

[3] LI N, XIN D, JIAN Z, et al. Topic2Vec: Learning distributed representations of topics[C]. Proceedings of 2015 International Conference on Asian Language Processing(IALP), 2015:209-212.

[4] 张标. 基于关键词提取和 BERT 词向量的新闻文本分类研究[D]. 淮南：安徽理工大学，2021.

[5] 颜端武，梅喜瑞，杨雄飞，等. 基于主题模型和词向量融合的微博文本主题聚类研究[J]. 现代情报，2021，41(10):67-74.

[6] 席笑文，郭颖，宋欣娜，等. 基于 word2vec 与 LDA 主题模型的技术相似性可视化研究[J]. 情报学报，2021，40(9):974-983.

[7] ZHANG Z, RAO Y, LAI H, et al. TADC: A topic-aware dynamic convolutional neural network for aspect extraction[J]. IEEE transactions on neural networks and learning systems, 2021, 1248-1255.

[8] KIM Y. Convolutional neural networks for sentence classification [J]. ArXiv Preprint, 2014, ArXiv: 1408.5882.

[9] 杨兴锐，赵寿为，张如学，等. 结合自注意力和残差的 BiLSTM_CNN 文本分类模型[J]. 计算机工程与应用，2022，58(3):172-180.

[10] 王根生，黄学坚. 基于 Word2vec 和改进型 TF-IDF 的卷积神经网络文本分类模型[J]. 小型微型计算机系统，2019，40(5):1120-1126.

[11] 冼广铭，王鲁栋，曾碧卿，等. 基于 LDA 和 BiGRU 的文本分类[J]. 计算机技术与发展，2022，32(4):15-20.

[12] 邓维斌，朱坤，李云波，等. FMNN：融合多神经网络的文本分类模型[J]. 计算机科学，2022，49(3):281-287.

[13] WANG Z, MA L, ZHANG Y. A hybrid document feature extraction method using latent dirichlet allocation and Word2Vec[C]. Proceedings of IEEE 1st International Conference on Data Science in Cyberspace, 2016:98-103.

[14] 李彦飞. 基于 LDA 模型和信息熵的热门微博发现[D]. 天津：天津财经大学，2017.

[15] 李超凡，马凯. 基于多通道注意力机制的文本分类模型[J]. 微电子学与计算机，2022，39(4):33-40.

作者简介

杨雳，硕士研究生，主要研究方向为舆情分析。

刘胜全(通信作者)，硕士，教授，主要研究方向为自然语言处理、计算机网络安全。

贾李睿智，博士研究生，主要研究方向为命名实体识别。

解舒淇，硕士研究生，主要研究方向为关系抽取。

E-mail:1731218691@qq.com