基于组合-卷积神经网络的中文新闻文本分类

张 昱[1,2],刘开峰[1],张全新[3],王艳歌[1],高凯龙[1]
(1. 北京建筑大学电气与信息工程学院 & 建筑大数据智能处理方法研究北京市重点实验室,北京 100044;
2. 中国矿业大学深部岩土力学与地下工程国家重点实验室,北京 100083;
3. 北京理工大学计算机科学与技术学院,北京 100081)

摘 要: 目前的新闻分类研究以英文居多,而且常用的传统机器学习方法在长文本处理方面,存在局部文本块特征提取不完善的问题. 为了解决中文新闻分类缺乏专门术语集的问题,采用构造数据索引的方法,制作了适合中文新闻分类的词汇表,并结合 word2vec 预训练词向量进行文本特征构建. 为了解决特征提取不完善的问题,通过改进经典卷积神经网络模型结构,研究不同的卷积和池化操作对分类结果的影响. 为提高新闻文本分类的精确率,本文提出并实现了一种组合-卷积神经网络模型,设计了有效的模型正则化和优化方法. 实验结果表明,组合-卷积神经网络模型对中文新闻文本分类的精确率达到93.69%,相比最优的传统机器学习方法和经典卷积神经网络模型精确率分别提升6.34%和1.19%,并在召回率和F值两项指标上均优于对比模型.

关键词: 自然语言处理; 词向量; 组合-卷积神经网络; 中文新闻; 文本分类
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2021)06-1059-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200134

A Combined-Convolutional Neural Network for Chinese News Text Classification

ZHANG Yu[1,2], LIU Kai-feng[1], ZHANG Quan-xin[3], WANG Yan-ge[1], GAO Kai-long[1]
(1. *School of Electrical and Information Engineering & Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing* 100044, *China*;
2. *State Key Laboratory in China for Geo Mechanics and Deep Underground Engineering, China University of Mining & Technology, Beijing* 100083, *China*;
3. *School of Computer Science and Technology, Beijing Institute of Technology, Beijing* 100081, *China*)

Abstract: At present, most of the researches on news classification are in English, and the traditional machine learning methods have a problem of incomplete extraction of local text block features in long text processing. In order to solve the problem of lack of special term set for Chinese news classification, a vocabulary suitable for Chinese text classification is made by constructing a data index method, and the text feature construction is combined with word2vec pre-trained word vector. In order to solve the problem of incomplete feature extraction, the effects of different convolution and pooling operations on the classification results are studied by improving the structure of classical convolution neural network model. In order to improve the precision of Chinese news text classification, this paper proposes and implements a combined-convolution neural network model, and designs an effective method of model regularization and optimization. The experimental results show that the precision of the combined-convolutional neural network model for Chinese news text classification reaches 93.69%, which is 6.34% and 1.19% higher than the best traditional machine learning method and classic convolutional neural network model, and it is better than the comparison model in recall and F-measure.

Key words: natural language processing; word vector; combined-convolutional neural network; Chinese news; text classification

收稿日期:2020-02-03;修回日期:2020-04-05;责任编辑:梅志强
基金项目:北京建筑大学优秀主讲教师培育计划(No. 21082718041);国家重点研发计划(No. 2016YFC0600901);教育部2018产学合作协同育人项目(No. 201801113001);市属高校基本科研业务费(No. 30850919027);北京建筑大学研究生创新项目(No. PG2020051)

1 引言

如今,互联网和大数据行业蓬勃发展,新闻已经成为人们了解社会动态、获取社会信息资源的重要手段之一. 自 20 世纪 90 年代末以来,经国家正式批准建立新闻网站约 200 多家,移动端新闻 APP 也是种类繁多,由此产生了海量新闻数据[1]. 为了高效地获取和管理有价值的新闻数据,新闻文本分类俨然成为世界上一个热门的研究领域[2,3]. 新闻文本分类的实现,有助于文本信息的管理、新闻秩序的实现和新闻数据的挖掘[4~6].

因全球经济一体化和一带一路战略的影响,汉语作为世界上使用最广泛的语言,俨然在世界语言体系中占有重要地位. 然而,对中文的新闻文本分类却很少,尤其是对中文长文本的分类[7,8]. 一方面,研究中文文本分类的相关语料库较少. 另一方面,汉语比西方语言复杂得多,很难用传统的方法提取特征. 这也是中文新闻文本分类发展缓慢的原因[9,10].

针对中文新闻文本分类所遇到的问题,本文采用构造数据索引的方法,制作了适合中文文本分类的术语集,用于新闻长文本分类. 同时,通过优化经典卷积神经网络模型结构[11~13],提出了一种组合-卷积神经网络模型自动提取文本特征,提升了中文新闻文本的分类效果. 此外,本文采用 word2vec 词袋模型训练的词向量特征作为原始输入,利用提出的模型算法与传统的新闻文本分类方法进行了多组实验对比,组合-卷积神经网络对中文新闻文本的分类准确率达到 93.69%. 在进一步的实验中,去除因样本数据集太不均衡造成的影响因素,本文的算法在准确率上又有所提升. 实验结果表明,本文算法是有效的.

2 相关工作

目前,文本分类作为自然语言处理的基础问题之一,解决这一问题为自然语言处理打开了许多大门,如信息检索、机器翻译和自动文摘等. 新闻文本分类常用的机器学习算法有:朴素贝叶斯(NB)[14]、最近邻(KNN)[15]、决策树(DT)、神经网络(NNs)、最大熵模型(ME)和支持向量机(SVM)[16]等.

2003 年词的分布式表示首次被 Bengio 等[17]运用于统计语言模型,神经语言模型开始获得广泛关注. 2008 年 Collobert 等[18]提出并采用神经网络的方法将文本词汇表示成张量数据,即相似的词映射到向量空间中相近的位置,一个词的含义由其上下文的词汇决定,但是其共享单词嵌入的方式只能在矩阵协作低级信息. 2013 年 Mikolov 等[19,20]提出来两个模型,连续词袋模型(CBOW)和连续 Skip-gram 模型. CBOW 是以先验概率的方式,输入某一个特征词上下文相关的词向量,输出该特定词的词向量. 而连续 Skip-gram 模型的预测方式与 CBOW 相反,通过输入中间词的向量来预测上下文的词向量. 连续 Skip-gram 模型能够更好地处理生僻词,但是当数据量较大时,存在训练耗时太长的问题. 针对解决在百万数量级的词典和上亿的数据集上进行高效地训练的问题,Google 开源了一款用于词向量计算的工具——word2vec[21]. 该工具主要将单词映射到低维空间,使用这些较低维的词嵌入向量放入分类器. 并且,word2vec 得到的训练结果词向量(word embedding)可以很好地度量词与词之间的相似性. 同年,Barakat 等[22,23]在发表的论文中提到多层神经网络有较为强大的特征学习能力,经过训练可以更加准确地映射出原始数据的真实含义.

卷积神经网络(CNN)模型最初是为计算机视觉而发明的,后来被 Meek[24]证明对 NLP 是有效的,并在语义分析上取得了很好的效果. 此后,LeCun 等[25,26]提出了一种字符级卷积神经网络模型,用不同的分类数据集进行语义分析和话题分类任务. 但该方法用于中文文本分类的训练和工作非常缓慢,因为中文文本分类的术语集和词的 N-gram 要比英文文本分类要大得多[27]. 而且,字符级的特征处理放弃了词所具有的语义信息,对于汉语来说,词与字符之间存在很多重叠语义,该特征提取的方式存在缺陷. 在借鉴前人研究成果的基础上,本文提出了一种有监督学习的组合-卷积神经网络模型,以分别卷积再组合的方式改进经典 CNN 模型结构,增加卷积操作却没有加深神经网络层,最终取得了较好的文本分类效果,解决了中文文本分类器训练缓慢的问题,并增强了对文本局部特征的提取.

3 组合-卷积神经网络的中文新闻分类

根据英文文本分类的处理步骤,本文在中文新闻分类方面,采用的流程包括:数据集预处理、文本特征表示、文本特征提取和分类器的训练. 根据数据集文本的特点,预处理部分主要确定文本序列长度构造数据集索引和实现数据读写的批处理. 文本特征表示用于制作词汇表和数据标准化. 文本特征提取采用 word2vec 训练的词向量特征作为输入,不同大小的卷积核自动提取文本块的局部特征. 最后,利用提取的特征构造组合-卷积神经网络模型(简称组合-CNN 模型)分类器进行训练和分类,并通过调整超参数优化模型.

3.1 数据集预处理

本文使用的数据集是 THUCnews,源于新浪新闻 RSS 订阅频道的历史数据筛选过滤生成,包含 836075 篇新闻文档(2.04GB),均为 UTF-8 纯文本格式. 在原始新浪新闻分类体系的基础上,整合划分出 14 个类别:科

技、股票、体育、娱乐、时政、社会、教育、财经、家居、游戏、房产、时尚、彩票、星座.

为了更好更方便构造整个数据集索引,本文对 THUCnews 进行大数据可视化分析,从而确定并设置最优的文本序列长度,其也作为后面模型中句子填充长度的标准. 经统计,平均每篇新闻字数为 941. 由直方图 1 可以看出,绝大部分文本都在 2000 字以内,而且利用出现频数的累计分布函数图(见图 2)可知,90% 的分位点对应的文本长度为 1857,所以根据可视化分析结果,本文设置读取文本长度为 2000.

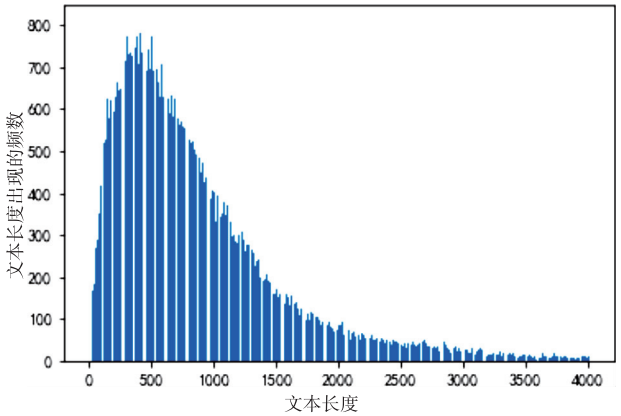

图1　文本长度的出现频数统计图

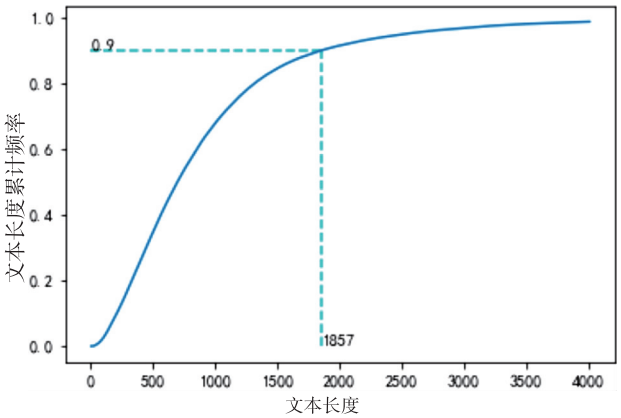

图2　文本长度的累计分布函数图

因为处理 80 多万个文本文件,读取时间较长,所以编程中采用 Python 的 pickle 标准模块存储复杂数据类型,将文本信息转变为二进制数据流. 二进制文件的加载速度非常快,加载速度是文本文件的 50 倍以上. 这样的信息存储在硬盘中,当实验读取文件数据的时候就很方便,将其反序列化即可得到原始的数据. 为避免内存溢出,所以每整合一定数量的文件保存一次.

3.2　文本特征表示

3.2.1　制作词汇表

制作词汇表是为了文本数据的标准化做准备. 2008 年 Hao Lili 等[28]提出了用于中文文本分类的标准停用词列表,以及用于识别停用词的加权卡方统计方法. 词汇表中去除停用词,原因在于这些词的使用频率过高,且语义影响不大,如“的”、“了”、“在”、“是”等. 如果词汇表中存在大量这样的词语,相当于浪费了很多资源. 添加一个关键词,特征提取就越好,所以词汇表该给予关键词更多的空间. 因此,词汇表中剔除了中文新闻里 20 个使用最频繁的停用词,如下图 3 所示.

```
#Remove stop words
stop_words = ['的', '了', '在', '是', '我', '有', '和', '就', '都', '一',
              '个', '上', '也', '到', '要', '去', '你', '会', '着', '这']
```

图3　去除停用词

汉字的数量很多,但是很难说出准确的数字. 据北京国安咨讯设备公司统计,汉字字库收入有出处汉字 91251 个. 常用汉字只有几千字,分为常用字表和次常用字表,大约是 2500 到 7000 之间. 简体与繁体的统计结果相差不大. 因此,对所有中文新闻文本的字做统计计数,出现频率排名前 7000 的字作为词汇表语料库.

3.2.2　数据标准化

(1)文本内容的数据标准化:首先,遍历词汇表索引序列,列出数据和数据下标,采用字典方法将列表强制转换. 其次,使用列表推导式和 lambda 匿名函数实现词和词 id 的映射. 最后,将词的映射用于每个样本内容获取标准化数据,即词转换成词 id 的向量数据.

(2)文本标签的数据标准化:采用分类数据广泛使用的 One-Hot 编码,将每个标签表示为全零向量,只有标签索引对应的元素为 1,使文本标签向量化.

3.3　一种改进的组合-CNN 模型

3.3.1　经典 CNN

经典 CNN 模型结构如图 4 所示,一般分为两种情况,单层卷积神经网络和多层卷积神经网络. 第 1 层是输入层,用于接收张量数据;第 2 层是卷积层,卷积核自动提取文本特征;第 3 层是池化层,常用方法有最大池化和平均池化;第 4 层选择性使用,单层卷积神经网络不需要使用,多层卷积神经网络是在单层卷积的基础上添加了额外的多个卷积层和池化层,前一个池化层的输出作为后一个卷积层的输入,属于叠加的方式;第 5 层是全连接层,可以添加 Dropout 层和常用激活函数,如 Sigmod、Tanh 和 ReLU;最终,采用 Softmax 逻辑回归模型来解决多分类问题.

3.3.2　组合-CNN

为了对新闻文本进行分类,在经典 CNN 模型的基础上,本文设计并实现了一个六层组合-CNN 模型,如图 5 所示.

第 1 层:用于接收输入的 Embedding 层. 因为新闻分类的输入数据为文本数据,文本数据需转化为实数向量数据才能进行输入. 因此,输入层中采用word2vec

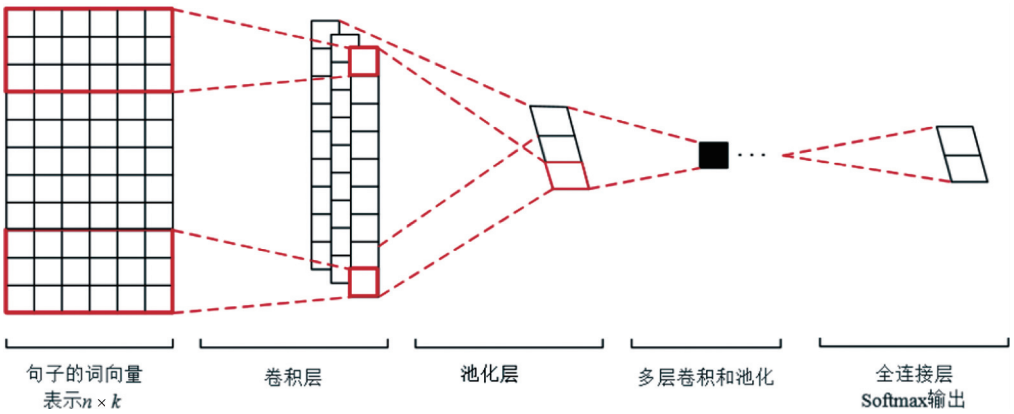

图4 经典CNN模型结构

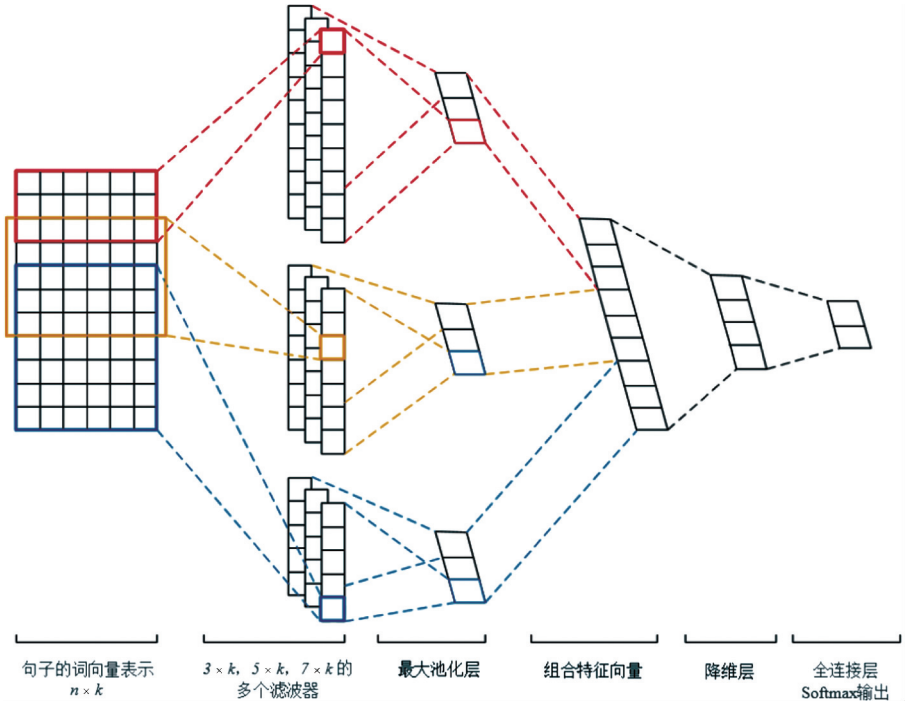

图5 组合-CNN 模型结构

将词汇表语义映射成实数向量,然后对数据标准化的样本内容做词嵌入,获得句子的词向量表示作为下一层的输入.

第 2 层和第 3 层:卷积层和池化层. 相比于经典 CNN 模型,组合-CNN 模型主要改进了卷积和池化操作的方式. 经典 CNN 模型有单层卷积和多层卷积的不同情况,单层卷积方面,一个卷积核提取的局部文本特征信息有限,并且不够完整. 多层卷积方面,多层卷积操作以叠加方式提取的文本特征往往过于抽象,不利于表达文本真实的涵义. 因此,为了提取更加完善的局部文本块特征,组合-CNN 模型利用三种不同大小的卷积核分别提取文本特征. 同时,为了抽取主要特征和减少特征参数的个数,利用最大池化层降采样的特点,对卷积的输出分别进行最大池化操作. 从而,在没有加深神经网络深度的情况下,提取到更多更重要的文本特征.

第 4 层和第 5 层:均属于中间隐藏层,在经典 CNN 中没有这两个隐藏层. 因为第 3 层的输出是三个池化操作的结果,所以采用隐藏层组合不同卷积核提取的特征向量. 本文在模型中,对每种卷积核的数量设置较多,且经过第 4 层组合特征向量输出的向量维数太大,从而添加一个隐藏层用于降维.

第 6 层:全连接层. 首先,在全连接层中添加 Dropout 层,防止模型过拟合 提升模型泛化能力. 其次,模型采用 ReLU 作为激活函数,增加神经网络模型的非线性,避免出现神经网络梯度消失的问题. 最后,利用 Softmax 对新闻文本进行分类预测.

以下为组合-CNN 模型工作原理的详细说明.

Embdding 层是一种字典查找,将整数索引映射为

密集向量. 该层接收整数作为输入,然后在内部字典中查找这些整数相关联的向量,并返回用于输出. 该层内部词向量映射用 Google 的词向量计算工具 word2vec,将输入数据做词嵌入,得到输入卷积层的词向量.

映射后向量化的中文文本,是一个 k 维的词向量 $\boldsymbol{R}^k$,假设 $\boldsymbol{x}_i$ 是 $\boldsymbol{R}^k$ 的第 i 个字的向量表示,所以长度为 n 的句子可以用式(1)表示:

$$\boldsymbol{x}_{1:n}=\boldsymbol{x}_1\oplus\boldsymbol{x}_2\oplus\cdots\oplus\boldsymbol{x}_n \tag{1}$$

其中⊕表示连接操作,$\boldsymbol{x}_{i:i+h-1}$ 为输入的第 i 个到第 $i+h-1$ 个窗口内的词向量矩阵. 卷积层利用不同大小卷积核对宽度为 k 的连续窗口进行卷积运算,卷积核为 $h\times k$ 的矩阵,本文中三种卷积核的高度 h 值分别设为 3、5、7,每种尺寸的卷积核有 r 个,值设置为 256. 权值矩阵 $\boldsymbol{W}_1$ 是一个 $h\times k$ 的实数矩阵,对 h 个字的文本块进行特征提取,由 $\boldsymbol{x}_{i:i+h-1}$ 提取的一个特征 o_i 如下:

$$o_i=f(\boldsymbol{W}_1\cdot\boldsymbol{x}_{i:i+h-1}+b_1) \tag{2}$$

$f(\cdot)$ 是非线性的激活函数,$b_1\in R$ 是一个偏置项. 卷积操作应用于一个完整新闻文本的词向量 $\{\boldsymbol{x}_{1:h},\boldsymbol{x}_{2:h+1},\cdots,\boldsymbol{x}_{n-h+1:n}\}$ 会得到以下一个特征图:

$$\boldsymbol{o}=[o_1,o_2,\cdots,o_{n-h+1}] \tag{3}$$

公式中 $\boldsymbol{o}$ 是大小为 $n-h+1$ 的实数向量. 为了抽取主要特征同时减少特征参数和计算量,采用最大池化方法取每个特征图中的最大值,作为该卷积核在文本向量上提取到的最重要特征,得到一个维度为 $1\times r$ 的特征向量. $\hat{\boldsymbol{o}}$ 表示最大池化运算后的结果,池化操作如下:

$$\hat{\boldsymbol{o}}=\max\{\boldsymbol{o}\} \tag{4}$$

以上内容介绍了一种尺寸的卷积核,进行特征提取的过程. 本文模型使用多个不同大小的卷积核来获取多个特征,所以将不同卷积核经最大池化后的结果拼接起来,得到大小为 $1\times 3r$ 的实数特征向量 $\boldsymbol{a}$,定义如下:

$$\boldsymbol{a}=\hat{\boldsymbol{o}}^3\oplus\hat{\boldsymbol{o}}^5\oplus\hat{\boldsymbol{o}}^7 \tag{5}$$

其中 $\hat{\boldsymbol{o}}^3$、$\hat{\boldsymbol{o}}^5$、$\hat{\boldsymbol{o}}^7$ 分别表示高度为 3、5、7 的卷积核经最大池化后输出的特征向量. 然后,添加一个隐藏层,用于非线性降维,变成大小为 $1\times d$ 的实数特征向量 $\boldsymbol{z}$(d 为隐藏层神经元结点数,本文设置为 128).

最后,这些特征传递到全连接层,通过 Softmax 层输出 14 个类别标签的概率分布. 取最大概率对应的类别,得到预测类别的标签值为 y_i,定义如下:

$$y_i=\max[\operatorname{softmax}(\boldsymbol{W}_2\cdot\boldsymbol{z}+b_2)] \tag{6}$$

公式中 $\boldsymbol{W}_2$ 是一个 $m\times d$ 的实数矩阵(m 为类别数),$b_2\in R$ 为偏置项. 为了加快收敛速度,采用小批量样本梯度下降,本文实验中设置批量样本数为 64. 另外,在全连接层引入 Dropout 层和 ReLU 激活函数的处理.

3.3.3　正则化和优化方法

在深度学习领域,合理划分训练集、验证集和测试集很重要. 当数据量不大(万级别及以下)的时候,划为 6∶2∶2 比较科学. 但本文数据量陡增将近百万级别,此时应将更多的样本数据给训练集,不需要太多的验证集和测试集,将比例设置为 98∶1∶1 就能很好地工作. 因此,根据自身使用数据量的规模,本文将训练集、验证集、测试集比例调整为 82∶6∶12,采用随机划分的方法,得到 686075 条中文新闻样本用于训练、50000 条验证集用于模型验证和优化,以及利用 100000 条测试集评估模型的分类效果.

第 1,验证集用于验证模型精度和损失,寻找模型开始过拟合的迭代轮次. 模型每迭代 100 轮次,输出一组精度值和损失值,绘制的精度曲线和损失曲线,如下图 6 和图 7 所示. 网络总的迭代次数为 20000 轮次,在训练第 10000 轮左右开始过拟合,即训练精度和训练损失相对稳定,且验证精度不再提高、验证损失也不再下降. 因此,去除此后的迭代训练,既能减轻电脑计算负载,也能避免模型过拟合.

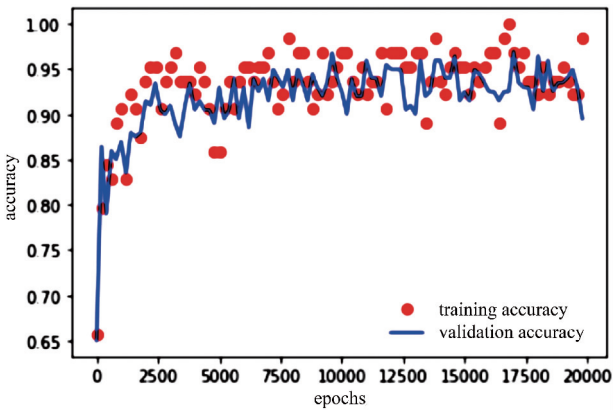

图6　训练精度和验证精度

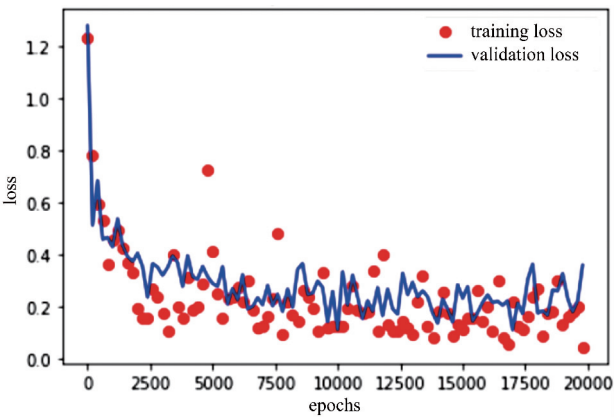

图7　训练损失和验证损失

第 2,神经网络的全连接层中,添加正则化方法 Dropout 层减少过拟合. Dropout 层是 CNN 中防止过拟合提升效果的重要方法,在每个训练批次中以一定概率 $1-p$ 将隐含层节点的输出值清零. 以这种方式减少特征检测器(隐藏层节点)间的相互作用,可以有效地

减轻过拟合现象,一定程度上达到正则化的效果.

3.3.4 模型训练

本文通过最小化训练集上的损失函数来训练组合-CNN 模型. 损失函数使用多分类交叉熵,即对数损失函数. 如式(7)所示:

$$L(Y,P(Y\mid X)) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{m=0}^{M-1} y_{i,m}\log p_{i,m} \quad (7)$$

其中,L 为损失函数,X 为输入变量,Y 为输出变量. $y_{i,m}$是一个二值指标,表示类别 m 是否为输入实例 X_i 的真实类别. $p_{i,m}$表示在 N 个实例中第 i 个实例预测为第 m 个类别的概率. 损失值用于衡量网络输出的概率分布与标签真实概率分布之间的距离,训练网络可使输出结果尽可能接近真实标签. 优化器调用 Adam 优化算法,引入了二次方梯度校正,计算每个参数的自适应学习率,是一个寻找全局最优点的优化算法. 模型训练共迭代 10000 次,训练完成大约 20min. 因此,采用 Tensorflow 中模型保存和加载的方法. 通过加载预先训练好的模型,在该模型基础上再次训练,从而在实验中节省大量时间.

4 实验与分析

本文实验环境的设置和实验平台的搭建如下:

(1)硬件方面:Windows 10 系统、CPU Inter(R) Core(TM) i7-8750H 2.20GHz、内存 8GB.

(2)软件和依赖的库:Python3.7、Jupyter notebook、Tensorflow_gpu-1.13.1、sklearn 等.

4.1 实验参数

实验过程中,组合-CNN 模型可调参数的设置是一致的,如表 1 所示. 数据被分批加载用于训练,每个批次为 64,全连接层中隐藏神经元个数为 128.

表 1 可调参数设置

参数	值
卷积核窗口长度 h	3,5,7
每种卷积核个数	256
词向量维数	128
Pooling 方法	Max pooling
学习率	0.0005
Dropout rate	0.5
激活函数	ReLU

4.2 实验设计

为了验证组合-CNN 模型算法的有效性,本文进行了多组不同模型的中文新闻文本分类实验,将其与传统且具有代表性的分类算法进行实验对比,使用各分类整体平均的精确率(precision)、召回率(recall)和 F_1 值(F-measure)评价不同模型的分类效果,并作为衡量分类器的性能指标.

(1)为验证组合-CNN 模型的分类性能,我们选择多个基准进行比较,分别将组合-CNN 与经典 CNN、传统的机器学习方法进行对比试验. 其中,经典 CNN 包括单层卷积神经网络(CNN-1)和多层卷积神经网络(CNN-3),传统机器学习方法包括朴素贝叶斯(NB)、最近邻(KNN)和支持向量机(SVM).

(2)为了进一步测试模型的有效性,减少因样本数据太不均衡对分类结果产生的影响,将数据集均衡化处理. 各类新闻样本原始占比如下图 8 所示,“星座”、“彩票”、“时尚”类别样本太少,不到总样本数的 3%,而“科技”、“股票”、“体育”类别样本又太多,仅三个类别就超过总样本数的 50% . 因此,会导致前者分类效果较差,通过图 9 混淆矩阵的标红框数据可以看出,前者的部分样本会被归类于后者. 混淆矩阵的每一行代表了数据的真实归属类别,每一列代表了预测类别. 再次经过随机划分均衡化的数据集共有 65000 个样本数据,分为 10 个类别,其中训练集 5000×10 个,验证集 500×10 个,测试集 1000×10 个. 基于不同的数据集,利用本文组合-CNN 模型的分类结果进行对比.

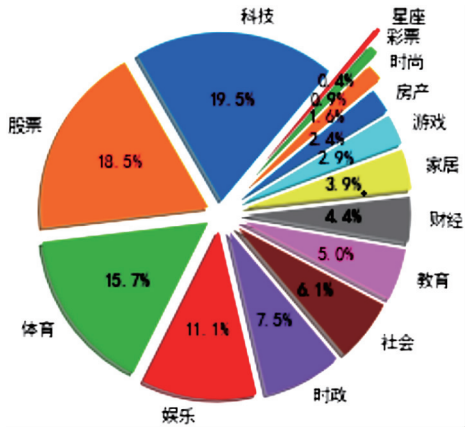

图8 各类样本占比图

	体育	娱乐	家居	彩票	房产	教育	时尚	时政	星座	游戏	社会	科技	股票	财经
体育	15585	50	10	16	2	8	11	11	0	5	17	20	5	0
娱乐	128	10381	62	0	3	26	35	56	3	21	97	145	10	0
家居	11	34	3637	0	6	13	32	13	5	1	16	91	16	2
彩票	84	2	0	747	2	3	0	8	0	1	33	8	2	1
房产	7	5	61	0	2264	5	0	15	0	2	15	15	42	4
教育	20	33	38	2	4	4490	25	89	3	1	120	94	16	3
时尚	3	64	72	1	0	6	1470	5	3	3	10	21	1	0
时政	55	42	42	1	10	71	17	6839	0	7	167	146	102	21
星座	2	11	11	0	1	14	16	5	364	3	12	6	4	0
游戏	14	28	12	0	0	5	4	2	0	2526	6	302	9	0
社会	47	101	62	12	17	147	28	206	0	4	5161	279	11	22
科技	37	103	106	2	7	58	13	187	1	116	196	18497	216	27
股票	11	29	44	0	32	7	6	236	0	5	18	448	17254	415
财经	8	27	15	6	17	6	5	54	0	1	52	70	473	3714

图9 分类结果混淆矩阵

4.3　结果分析

(1)在实验中,我们实现特征构建的方法均以预训练好的词向量作为输入,不同分类模型的分类结果如表 2 所示.

表 2　模型分类结果比较

模型	精确率	召回率	F_1 值
NB	0.8187	0.8162	0.8140
KNN	0.8548	0.8526	0.8494
SVM	0.8735	0.8763	0.8739
CNN-1	0.9250	0.9240	0.9243
CNN-3	0.9192	0.9178	0.9167
组合-CNN	0.9369	0.9368	0.9368

通过表 2 对比可以发现,第 1:采用 word2vec 词袋模型预训练词向量,进行特征构建作为模型输入,在相同的数据集上各个分类模型均取得了 80% 以上的精确率,说明词向量能够很好地描述文本特征. 第 2:不论是单层卷积神经网络还是多层卷积神经网络,取得的分类效果都优于三种传统机器学习算法,说明 CNN 模型可以学习到更多的分类特征,相比传统的机器学习模型更有优势. 第 3:多个卷积层的 CNN-3 模型比单个卷积层的 CNN-1 模型取得的分类效果差,说明在经典 CNN 模型的基础上加深卷积层并没有取得预期的效果,这也是本文改进经典 CNN 模型结构的原因. 第 4:组合-CNN 模型对中文新闻文本分类的精确率达到 93.69%,相比 NB、KNN、SVM 的分类效果,分别在分类精确率上提高了 11.82%、8.21%、6.34%,且相比于经典 CNN-1 模型的分类效果,在精确率也有 1.19% 的提升,同时召回率和 F_1 值两项指标也优于对比模型,说明采用词向量分别卷积再组合的方式,能够提取更加全面的局部文本块特征信息,在文本分类效果上有很好的提升.

为了进一步分析不同分类模型之间的差异,本文分别挑选了三类方法中分类效果最优的模型进行可视化分析. 我们分别对组合-CNN、CNN-1 和 SVM 模型进行统计对比,测试模型每迭代 100 轮次,输出一组测试精度值和损失值. 随迭代次数的变化,不同模型的测试精度和损失如图 10 所示.

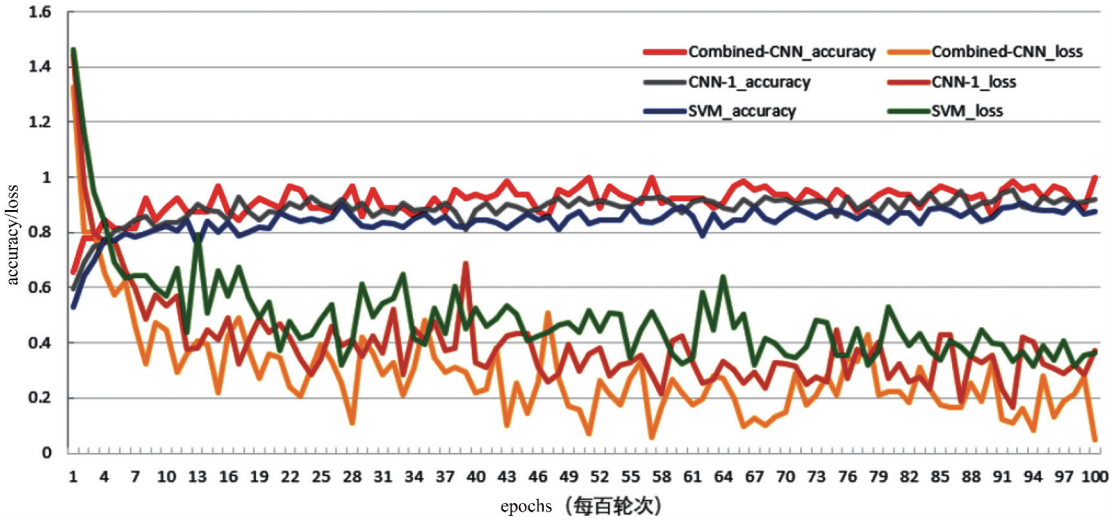

图10　不同模型的测试精度和损失

由图 10 可知,模型精度值随迭代次数的增加很快上升趋于平稳,并最终趋于稳定收敛状态. 因为梯度下降和 Adam 优化算法的作用,损失值也逐渐下降,最终稳定在一个很小的区间波动. 组合-CNN 模型的整体精度值高于 CNN-1 和 SVM 模型,说明组合-CNN 模型在经典 CNN 模型的基础上进行结构优化,改进后取得了不错的分类效果. 此外,组合-CNN 模型损失值的收敛速度明显增加,虽然浮动较大,但整体损失值还是低于经典 CNN 模型,并且明显优于 SVM 模型. 由此可见,组合-CNN 模型算法在中文新闻文本分类方面的有效性.

(2)我们又设计了不同数据集的分类实验. 分析分类结果的混淆矩阵发现,样本占比少的类别往往被错误分类成样本占比多的类别. 因此,在实验中进一步划分数据集,采用相同模型,在不同数据集上进行分类结果比较,如表 3 所示.

表 3　不同数据集分类结果

数据集	精确率	召回率	F_1 值
不均衡	0.9369	0.9368	0.9368
均衡	0.9557	0.9544	0.9540

根据表 3 可知,同样使用组合-CNN 模型的情况下,在均衡的数据集上,取得的精确率高达 95.57%. 使用均衡数据集相比不均衡的数据集,取得的分类效果更好,精确率提升了 1.88%,召回率提升了 1.76%,F_1 值

提升了 1.72%.说明对全部的不均衡数据集再次处理获得均衡数据集,可以很好地解决样本数据占比极端造成的问题,防止样本占比少的类别被错误分类成样本占比多的类别.因此,数据集太不均衡对分类结果的影响较大,对数据集的均衡化处理可以进一步提升新闻分类的精确率.

5 结论

本文利用构造数据索引的方法制作词汇表,并通过 word2vec 将词汇表语义映射到实数向量.在经典 CNN 模型的基础上,提出了一种改进的组合-CNN 模型结构,由经典叠加卷积的方式改进为分别卷积再组合的方式,这样使文本块局部特征的提取更加全面.从实验结果看出,组合-CNN 模型对文本的分类效果有了一定程度的提升,精确率达到 93.69%.另外,对数据集均衡化处理后,组合-CNN 模型进一步取得了更好的分类结果.但是,由于现实中的新闻数据不可能是均衡化的,所以此处存在数据集较理想,泛化不足的缺点.下一步工作,尝试在更多的数据集上使用该模型,并对样本数据进行权重计算用于分类模型的训练和测试,减少模型对数据集的依赖性.利用长短时记忆神经网络适用于序列长文本数据和能够表达上下文信息的特点,进行中文新闻分类,并将其与组合-CNN 模型结合,构建用于集成学习的多分类器系统.

参考文献

[1] Chung T, Xu B, Liu Y, et al. Empirical study on character level neural network classifier for Chinese text[J]. Engineering Applications of Artificial Intelligence, 2019, 80(4):1-7.

[2] J He, M Zou, P Liu. Convolutional neural networks for Chinese sentiment classification of social network[A]. 2017 IEEE International Conference on Mechatronics and Automation (ICMA)[C]. Takamatsu, Japan: IEEE, 2017. 1877-1881.

[3] 唐焕玲,窦全胜,于立萍,等.有监督主题模型的 SLDA-TC 文本分类新方法[J].电子学报,2019,47(6):1300-1308.
TANG Huan-ling, DOU Quan-sheng, YU Li-ping, et al. SLDA-TC: A novel text categorization approach based on supervised topic model[J]. Acta Electronica Sinica, 2019, 47(6):1300-1308. (in Chinese)

[4] 钟将,张淑芳,郭卫丽,等.主题特征格分析:一种用户生成文本质量评估方法[J].电子学报,2018,46(9):2201-2206.
ZHONG Jiang, ZHANG Shu-fen, GUO Wei-li, et al. TF-LA: A quality analysis framework for user generated contents[J]. Acta Electronica Sinica, 2018, 46(9):2201-2206. (in Chinese)

[5] Yang Y, Nenkova A. Combining lexical and syntactic features for detecting content-dense texts in news[J]. Journal of Artificial Intelligence Research, 2017, 60(9):179-219.

[6] Wang Y, Li H, Wu Z. Attitude of the Chinese public toward off-site construction: A text mining study[J]. Journal of Cleaner Production, 2019, 238(11):117926.

[7] Liu C, Wang X. Quality-relatedenglish text classification based on recurrent neural network[J]. Journal of Visual Communication and Image Representation, 2019, 71(8):102724.

[8] 吕品,计春雷,汪鑫,等.融合锚词抽取的海量短文本主题层次挖掘[J].电子学报,2018,46(5):1084-1088.
LU Pin, JI Chun-lei, WANG Xin, et al. Mass of short texts topical hierarchy mining integrated anchor extraction[J]. Acta Electronica Sinica, 2018, 46(5):1084-1088. (in Chinese)

[9] Liao W, Wang Y, Yin Y, et al. Improved sequence generation model for multi-label classification via CNN and initialized fully connection[J]. Neurocomputing, 2020, 382(3):88-195.

[10] 吕泽芳,马刚,孙先文,王俪蓉,史凌云,关志涛.人工智能安全的概念、分类及研究现状综述(一)[J].智慧电力,2019,8(47):32-42.
Lu Z F, Ma G, Sun X W, et al. Overview of the concept, classification and research status of AI security (I)[J]. Smart Power, 2019, 8(47): 32-42. (in Chinese)

[11] Ong Hui J L, Hoon G K, Wan Zainon W M N. Effects of word class and text position in sentiment-based news classification[J]. Procedia Computer Science, 2017, 124(11):77-85.

[12] Yang X, Xu S, Wu H, et al. Sentiment analysis of weibo comment texts based on extended vocabulary and convolutional neural network[J]. Procedia computer science, 2019, 147(2):361-368.

[13] Khan A, Sung J E, Kang J W. Multi-channel fusion convolutional neural network to classify syntactic anomaly from language-related ERP components[J]. Information Fusion, 2019, 52(12):53-61.

[14] Liu P, Zhao H H, Teng J Y, et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark[J]. Journal of Central South University, 2019, 26(1):1-12.

[15] Jiang J Y, Tsai S C, Lee S J. FSKNN: Multi-label text categorization based on fuzzy similarity and k nearest neighbors[J]. Expert Systems with Applications, 2012, 39(3):2813-2821.

[16] Malviya R, Jain P. Anovel text categorization approach based on K-means and support vector machine[J]. Inter-

national Journal of Computer Applications, 2015, 130(14):1 – 7.

[17] Bengio Y, Schwenk H, Senécal J S. Neural Probabilistic Language Models [M]. Berlin, Heidelberg: Springer, 2003.

[18] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[A]. Proceedings of the 25th international Conference on Machine learning[C]. Helsinki, Finland: ICML, 2008. 160 – 167.

[19] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [A]. Advances in Neural Information Processing Systems [C]. Lake Tahoe, NV, United States: NIPS, 2013. 3111 – 3119.

[20] Xu R, Chen T, Xia Y, et al. Wordembedding composition for data imbalances in sentiment and emotion classification[J]. Cognitive Computation, 2015, 7(2):226 – 240.

[21] Yang Z T, Zheng J . Research on Chinese text classification based on Word2vec[A]. 2016 2nd IEEE International Conference on Computer and Communications (ICCC) [C]. Chengdu, China: IEEE, 2016. 1166 – 1170.

[22] Barakat B K, Seitz A R, Shams L . The effect of statistical learning on internal stimulus representations: Predictable items are enhanced even when not predicted[J]. Cognition, 2013, 129(2):205 – 211.

[23] 杨国为,王守觉,卫成兵,等. 基于同源的同类事物连通本性的模式分类神经网络模型[J]. 电子学报,2013,41(1):52 – 55.
YANG Guo-wei, WANG Shou-jue, WEI Cheng-bing, et al. Pattern classification neural network model based on homologue connectedness [J]. Acta Electronica Sinica, 2013, 41(1):52 – 55. (in Chinese)

[24] Yih W, He X, Meek C. Semantic parsing for single-relation question answering[A]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics[C]. Baltimore, MD, United States: ACL, 2014. 643 – 648.

[25] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[A]. Advances in Neural Information Processing Systems[C]. Montreal, QC, Canada: NIPS, 2015. 649 – 657.

[26] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 2016, 116(1): 1 – 20.

[27] 任卓君,陈光,卢文科. 基于 N-gram 特征的恶意代码可视化方法[J]. 电子学报,2019,47(10):2108 – 2115.
REN Zhuo-jun, CHEN Guang, LU Wen-ke. Malware visuallzation methods based on N-gram features [J]. Acta Electronica Sinica, 2019, 47(10):2108 – 2115. (in Chinese)

[28] Hao L, Hao L. Automatic identification of stop words in Chinese text classification[A]. 2008 International Conference on Computer Science and Software Engineering [C]. Hubei, China: IEEE, 2008. 718 – 722.

作者简介

张　昱　男,1979 年 1 月生于内蒙古呼和浩特. 毕业于北京理工大学获博士学位,现为北京建筑大学电气与信息工程学院副教授、硕士生导师,主要研究方向为大数据、人工智能与岩爆.
E-mail: yuzhang@ bucea. edu. cn

刘开峰(通讯作者)　男,1996 年 1 月生于江苏淮安. 北京建筑大学电气与信息工程学院硕士研究生,主要研究方向为大数据、城市计算与人工智能.
E-mail:bigdata@ bucea. edu. cn

张全新　男,1974 年生于山东. 2003 年毕业于北京理工大学获博士学位,美国康涅狄洛大学访问学者,现为北京理工大学计算机学院讲师,主要研究方向为计算机网络、机器学习.
E-mail:zhangqx@ bit. edu. cn

王艳歌　女,1994 年 7 月生于河北衡水. 北京建筑大学电气与信息工程学院硕士研究生,主要研究方向为大数据、数据融合、可视分析.
E-mail:yangechn@ 126. com

高凯龙　男,1996 年 5 月出生于河北石家庄. 北京建筑大学电气与信息工程学院硕士研究生,主要研究方向为大数据、城市计算与人工智能.
E-mail:2489681545@ qq. com