计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(5):137−142 [doi: 10.15888/j.cnki.csa.006887]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

卷积神经网络的短文本分类方法①

陈巧红, 王　磊, 孙　麒, 贾宇波

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 王　磊, E-mail: 641082634@qq.com

摘　要: 短文本分类是自然语言处理的一个研究热点. 为提高文本分类精度和解决文本表示稀疏问题, 提出了一种全新的文本表示 (N-of-DOC) 方法. 采用 Word2Vec 分布式表示一个短语, 将其转换成的向量作为卷积神经网络模型的输入, 经过卷积层和池化层提取高层特征, 输出层接分类器得出分类结果. 实验结果表明, 与传统机器学习 (K 近邻, 支持向量机, 逻辑斯特回归, 朴素贝叶斯) 相比, 提出的方法不仅能解决中文文本向量的维数灾难和稀疏问题, 而且在分类精度上也比传统方法提高了 4.23%.

关键词: 卷积神经网络; 短文本分类; 文本表示; 机器学习; 深度学习

引用格式: 陈巧红,王磊,孙麒,贾宇波.卷积神经网络的短文本分类方法.计算机系统应用,2019,28(5):137–142. http://www.c-s-a.org.cn/1003-3254/6887.html

Short Text Classification Based on Convolutional Neural Network

CHEN Qiao-Hong, WANG Lei, SUN Qi, JIA Yu-Bo

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Short text classification is one of the hotspots of research in natural language processing. A new model of text representation is proposed in this study (N-of-DOC), and in order to solve the problem of sparse representation in Chinese, the Word2Vec distributed representation is used, finally, it is applied to the improved Convolution Neural Network (CNN) model to extract the high level features from the filter layer, the classification model is obtained by connecting the Softmax classifier after the pooling layer. In the experiment, the traditional text representation model and the improved text representation model are used as the input of the original data, respectively. It acts on the model of traditional machine learning (KNN, SVM, logistic regression, naive Bayes) and the improved CNN model. The results show that the proposed method can not only solve the dimension disaster and sparse problem of Chinese text vectors, but also improve the classification accuracy by 4.23% compared with traditional methods.

Key words: Convolution Neural Network (CNN); short text classification; text representation; machine learning; deep learning

随着 Internet 的大规模普及和上网人数的急剧增加, 网络上每天产生的各种短文本数量也呈指数式的增长. 互联网短文本是指那些较短的文本形式, 一般不超过 500 字, 例如用户商品评论, 短博客等. 并且这种半结构或无结构化互联网文本信息具有稀疏性、实时性、不规范性、流行语不断出现等特征. 互联网短文本分类作为信息处理关键技术之一, 在信息检索和知识挖掘领域已经取得很大进展[1].

国内外众多学者针对短文本分类的研究主要体现在以下两个方面, 文本数据的特征表示和算法模型的

① 基金项目: 浙江省自然科学基金 (LY17E050028)
Foundation item: National Natural Science Foundation of Zhejiang Province (LY17E050028)
收稿时间: 2018-11-13; 修改时间: 2018-12-03; 采用时间: 2018-12-11; csa 在线出版时间: 2019-05-01

选择与改进. Kim Y[2]使用卷积神经网络对英文电影评论进行分类, 只使用了一层卷积一层最大池化最后接 Softmax 全连接得出分类模型, 该方法虽然利用了深度学习模型, 但是隐藏层太浅, 不足以提取出更高层特征. 黄文明[3]等提出将 K 近邻运用在文本加权上, 对初始文本通过一定的权重采样, 最后运用 K 近邻分类器得出分类模型, 但此方法面对现如今海量的数据集运算量过高, 训练时间太长, 实际生产效果不好.

针对上述方法存在的特征表示问题以及提取文本高层特征效果不佳问题. 本文提出一种全新的文本表示 (N-of-DOC) 方法. 该方法首先将整个训练集经过文本预处理得到词向量特征, 运用信息增益、基尼不纯度和卡方检验[4–6]从短语特征中提取出整个训练集的高质量特征, 最后在每一篇文档上提取出的短语特征必须由从全部训练集提取出的特征线性表示. 为了能进一步提取出文本的高层特征, 本文在卷积神经网络模型上进行了改进, 卷积层采用最大卷积, 池化层采用最大池化. 这种模型加强了文本数据中词与词、文本与文本之间的关系, 在短文本的分类精度和分类速度上都有显著提高, 面对现如今海量的数据进行快速且高精度的分类有着深远意义.

1 模型架构与算法改进

本文所设计的基于卷积神经网络短文本分类模型架构主要包括以下 6 个方面: 数据的收集、数据预处理、每篇文档全新的特征表示、卷积神经网络 filter 层、K-max 池化层、Softmax 分类层. 整个架构模型的流程图如图 1 所示.

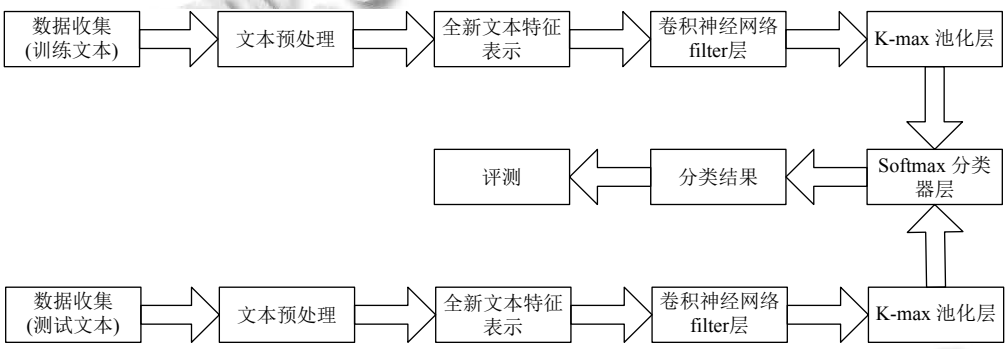


图 1 模型架构设计

1.1 数据的收集层

本文实验所用到的数据主要是 Sougou 语料库提供的文本数据 (总共包括 9 大类) 和使用爬虫库, 在电影网站爬取的一部分电影评论二分类短文本数据.

1.2 文本预处理层

通过数据收集层得到的原始数据不能直接作为卷积神经网络 (CNN) 的输入层输入给模型进行计算, 需要进行预处理操作, 预处理过程主要包括以下两部分.

(1) 中文的特殊性, 中文分词操作需要采取相应的分词算法进行分词, 本文采用的是 python 的分词库 (jieba) 来进行分词, 该分词库包含三大优势, 1、基于 Trie 树结构实现高效的词图扫描, 具有查找速度快的优势. 2、采用动态规划查找最大概率路径, 对句子进行双向切分查找, 可以有效避免过多形容词、副词等对句子切分, 计算概率带来的影响. 3、对于未登录词, 采用了汉字成词能力的 HMM 模型, 对于本文的网络数据集可以有效避免未登录词对分词造成的影响. 在分词处理过程中, 采用了并行化处理, 加速文本分词的速度[7].

(2) 去除停用词对于中文是必不可少的一件事, 因为停用词对于一篇文档来说, 它几乎不能给该篇文档带来任何信息量, 而且去除停用词可以减少文本冗余使文本分类更加准确[8].

1.3 全新文本特征表示层

为了降低文本表示的维度和减少计算的复杂度, 本文提出的全新文本特征表示模型解释如下:

(1) 本文通过将整个训练集经过文本预处理得到短语特征, 运用信息增益、基尼不纯度、和卡方检验从短语特征中提取出整个训练集的高质量特征, 最后在每一篇文档上提取出的短语特征必须由从全部训练集提取出的高质量特征线性表示. 抽象成数学模型表示如下:

$$T(D_j)=\theta_j\sum_{j=1}^{n}D_j \tag{1}$$

$T(D_j)$为具体的某一篇文档经过线性表示筛选后的短语, D_j为每一篇文档经过文本分词预处理后的短语, θ_j为线性表示这篇文档的筛选系数.

(2) 经过上述方法对每一篇文档提取出的特征短语, 本文采用了 gensim 库中的 Word2Vec[9]将其训练成一个 300 维的向量. 相对于词袋模型表征短语, 词向量模型避免了词袋模型高稀疏性的特点, 而且在卷积操作过程中, 若是词袋模型表征短语, 就意味着卷积大部分卷到的都是一些全零数字, 失去了卷积的意义, 而词向量模型则完全可以避免这方面带来的影响. 最后词向量模型还可以采用分布式训练, 加速算法的训练过程.

1.4 卷积神经网络 filter 层

经过全新的文本特征表示层后, 每一篇文档都有相应的高质量特征线性表示, 每一篇文档此时是一个 $K\times300$ 的矩阵, K 表示本篇文档提取出的高质量特征个数, 300 是本文设定的一个向量维度. 将其输入给卷积的 filter 层, 实验过程中卷积的窗口设定为 3×300、4×300、5×300、6×300 四种滤波器核, 卷积的步长设定为 1, 经过 filter 层后的文档矩阵表示分别为 (K−3+1)×1、(K−4+1)×1、(K−5+1)×1、(K−6+1)×1. 卷积公式如下:

$$B(i,j)=\sum_{m=0}\sum_{n=0}K(m,n)*A(i-m,j-n) \tag{2}$$

其中, A为被卷积矩阵, K为卷积核, B为卷积出来的结果.

1.5 K-max 池化层

在本文的实验过程中, 池化层作为一个降维操作, 进一步降低了文本的向量维度. 从特征提取的层面思考池化层, 可以认为池化层也是一层特征提取. 本文通过实验不断尝试 K-max 池化窗口的大小, 实验发现, 经过卷积之后的矩阵向量, K-max 的窗口大小调整为和卷积之后的矩阵向量大小一致最优. 因为本文的数据集是文本数据, 若池化操作还像图像处理过程一样, 采用局部池化, 这只能影响到一句短文本中的一部分词语, 失去了全局的一个信息捕获, 而且卷积操作已经是对一句短文本的局部信息卷积, 最后, 采用这种方式池化, 也可以有效避免短文本长短不一带来的影响.

1.6 分类器层

卷积神经网络的最后一层一般采用的都是全连接方式, 通过上一层的 K-max 池化层处理后的文本特征向量经过矩阵的 concat 和 reshape 之后, 送入 Softmax 分类器, 用来预测类别概率. K-max 池化后经过 concat 和 reshape 后得到 m 个训练集数据, 形式如下, $\{(X^{(1)},Y^{(1)})(X^{(2)},Y^{(2)})\cdots(X^{(m)},Y^{(m)})\}$, $X^{(i)}$代表输入的特征, $Y^{(i)}$代表文本类别. 由于本文不仅在二分类数据集上进行了实验, 也在多分类数据集上进行了实验, 所以在全连接层后的分类器公式分别如下:

(1) 二分类实验采取 sigmoid 函数, 阈值为 0.5.

$$h(\theta)=\frac{1}{1+\exp^{(-\theta^T X)}} \tag{3}$$

$$f(h(\theta))=\begin{cases}1 & h(\theta)>=0.5\\ 0 & h(\theta)<0.5\end{cases} \tag{4}$$

式中, θ代表模型参数, 通过对θ的训练可以找到最小代价函数$J(\theta)$, 其公式如下:

$$J(\theta)^1=\sum_{i=1}^{m}y^{(i)}\log h_\theta(x^{(i)}) \tag{5}$$

$$J(\theta)^2=\sum_{i=1}^{m}(1-y^{(i)})\log(1-h_\theta(x^{(i)}) \tag{6}$$

$$J(\theta)=-\frac{1}{m}[J(\theta)^1+J(\theta)^2] \tag{7}$$

(2) 多分类实验采取 Softmax 函数.

$$\sigma(z)_j=\frac{e^{z_j}}{\sum_{k=1}^{K}e^{z_k}} \tag{8}$$

$\sum_{k=1}^{K}e^{z_k}$为归一化因子, e^{z_j}为将各样本值构成单纯型, $\sigma(z)_j$为具体的某一篇文档得到的向量. 在本文实验过程中采用的是 numpy 库中的 argmax 函数, 返回对应概率值最大的索引值即为本样本的分类值.

2 评测标准及实验结果分析

2.1 评测标准

文本分类的评测标准采用了准确率、精确率、召回率、F1 值作为指标, 在计算精确率和召回率时, 需要用到混淆矩阵, 根据分类结果可建立混淆矩阵见表 1.

表 1 分类结果混淆矩阵

判别	真正属于该类别文档	真正不属于该类别文档
判别属于该类别文档	A	B
判别不属于该类别文档	C	D

(1) 准确率在文本分类算法中表示的是分类正确的文档数除以整个训练集的文档总数, 计算公式如下:

$$准确率=\frac{A+D}{A+B+C+D}\times 100\% \quad (9)$$

(2) 精确率是分类器正确判断为该类的样本数与分类器判断属于该类的样本总数之比, 体现了系统分类结果的准确程度. 计算公式如下:

$$P=\frac{A}{A+B}\times 100\% \quad (10)$$

(3) 召回率是分类器正确判断为该类的样本数与属于该类的样本总数之比, 体现了系统分类结果的完备性. 计算公式如下:

$$R=\frac{A}{A+C}\times 100\% \quad (11)$$

(4) 由于只有基本指标评价分类性能, 存在缺陷, 要全面反映分类性能需要结合基本指标衍生综合指标, 刻画分类性能, 本文采用的是 F-测度值, 计算公式如下:

$$F_\beta-测度值=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R} \quad (12)$$

其中, β是调整参数, 用于调整精确率P与召回率R在计算公式中的比重. 在本文的使用中, 取$\beta=1$, 则得到F_1值.

2.2 实验结果分析

实验使用传统机器学习 (K 近邻, 支持向量机, 逻辑斯特回归, 朴素贝叶斯) 的分类准确率与本文的分类准确率做对比, 同时也对比了精确率、召回率、F1 值. K 近邻文本分类, 除了使用本文所提到的全新的文本表示模型 (N-of-DOC) 来表示一篇文档之外, 还外加一个阈值, 即每篇文档必须有 5 个以上的相同词的训练文本, 才进行文本相似性的比较, 找出相似距离最近训练文本, 查看该篇文档属于哪类最多, 判定测试文本最终为哪个类别. 对于支持向量机和逻辑斯特回归分类, 使用本文的这种全新文本表示模式 (N-of-DOC) 方法提取词特征, 采用 Word2Vec 训练词向量作为文本特征向量, 作为模型的输入层构造分类器, 得出最后分类结果.

在卷积神经网络的短文本分类实验中, 不同模型有着不同的准确率, 精确率, 召回率, F1 值, 取所有模型各自最好的分类得分值作为实验对比, 各分类算法得分值大小如表 2 所示.

由表 2 可知采用不同的分类方法对于互联网短文本的分类结果影响很大, 使用本文提出的改进的卷积神经网络分类方法的正确率可以达到 92% 以上, 分类效果明显好于传统的分类器, 精确率、召回率、F1 值相应的也比传统分类方法效果更优. K 近邻分类器在本文使用的短文本语料库中分类效果最差, 正确率只有 60%. 而支持向量机、逻辑斯特回归、朴素贝叶斯分类器的分类准确率也是只达到了 74%-85% 之间. 相比于本文提出的改进的卷积神经网络模型算法, 不管在准确率、精确率、召回率、F1 值上都低, 这是因为传统机器学习是一种浅层的算法模型, 很大程度上依赖于特征工程的处理, 而卷积神经网络模型是一种深度学习模型, 随着隐藏层的增加, 提取更高层的特征能力也就越强.

表 2 电影评论二分类性能比较

分类算法	类别数据	评测标准			
		准确率 (%)	精确率 (%)	召回率	F1 值
K 近邻	电影评论正例	61.23	54.27	59.28	56.56
	电影评论负例	60.52	61.18	57.34	59.97
支持向量机	电影评论正例	74.85	69.26	72.02	70.96
	电影评论负例	76.31	69.24	79.68	74.08
逻辑斯特回归	电影评论正例	80.33	79.68	82.15	81.28
	电影评论负例	78.53	76.24	80.46	78.49
朴素贝叶斯	电影评论正例	84.19	82.48	84.31	83.02
	电影评论负例	83.65	81.26	83.11	82.24
本文模型	电影评论正例	92.36	93.22	95.13	94.27
	电影评论负例	90.45	94.68	96.24	91.16

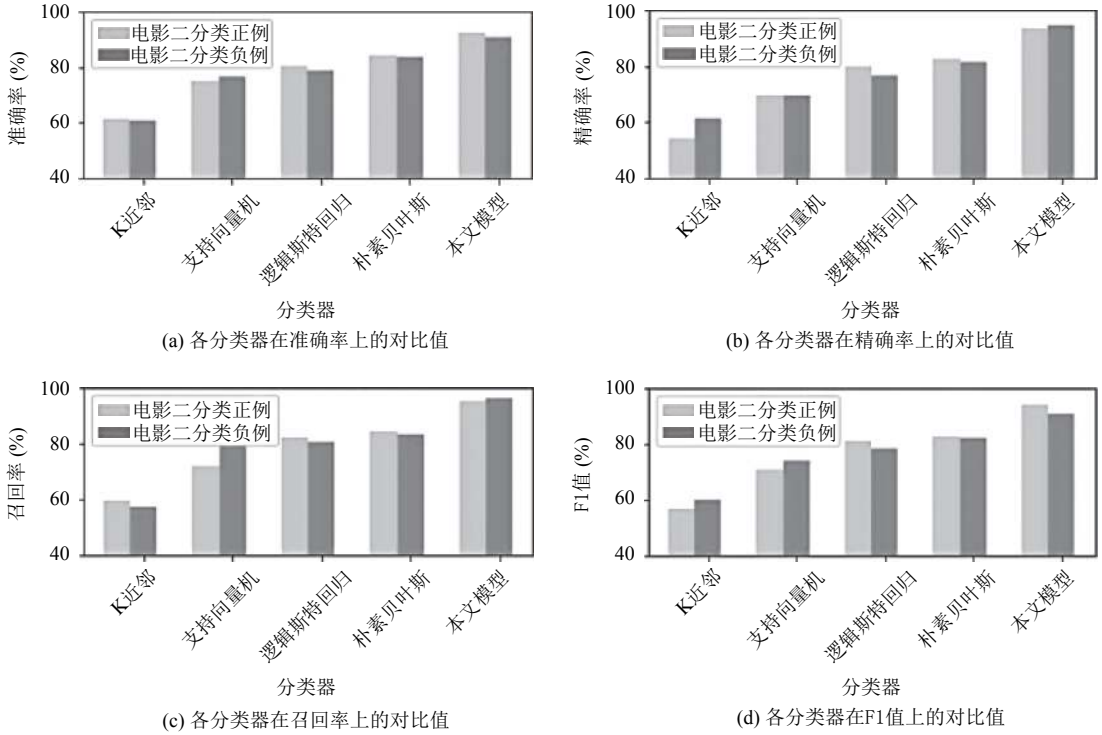
传统机器学习模型和改进的卷积神经网络模型在电影二分类数据集上的准确率、精确率、召回率、F1 值显示如图 2 所示.

传统机器学习 (K 近邻, 支持向量机, 逻辑斯特回归, 朴素贝叶斯) 和本文提出的改进的卷积神经网络模型在多分类数据集上的准确率、精确率、召回率、F1 值如表 3 所示.

传统机器学习 (K 近邻, 支持向量机, 逻辑斯特回归, 朴素贝叶斯) 和本文提出的改进的卷积神经网络模型在多分类数据集上的准确率、精确率、召回率、F1 值如图 3 所示.

通过表 3 可以看出在多分类任务上同样是使用基于本文提出的改进的卷积神经网络分类方法的准确率最高, 分类准确率平均在 90% 以上, 精确率、召回率、F1 值相比于传统机器学习效果更优. 其中 K 近邻的分类准确率平均在 70% 左右, 精确率、召回率、F1 值也是最低的. 支持向量机和逻辑斯特回归的分类准确率平均在 80% 左右, 朴素贝叶斯分类器略微提高一点, 分类准确率平均在 84% 左右. 取得上述效果的原因, 可以总结如下两点: 1) 本文使用了 gensim 工具包提供的 Word2Vec 训练词向量, 生成的词向量比简

单的词袋模型生成的向量更能代表词组之间的特征. 2) 通过本文提出的全新的文本表示模型 (N-of-DOC) 对于深度学习更加适合, 更加有利于卷积神经网络提取出更高层的特征.



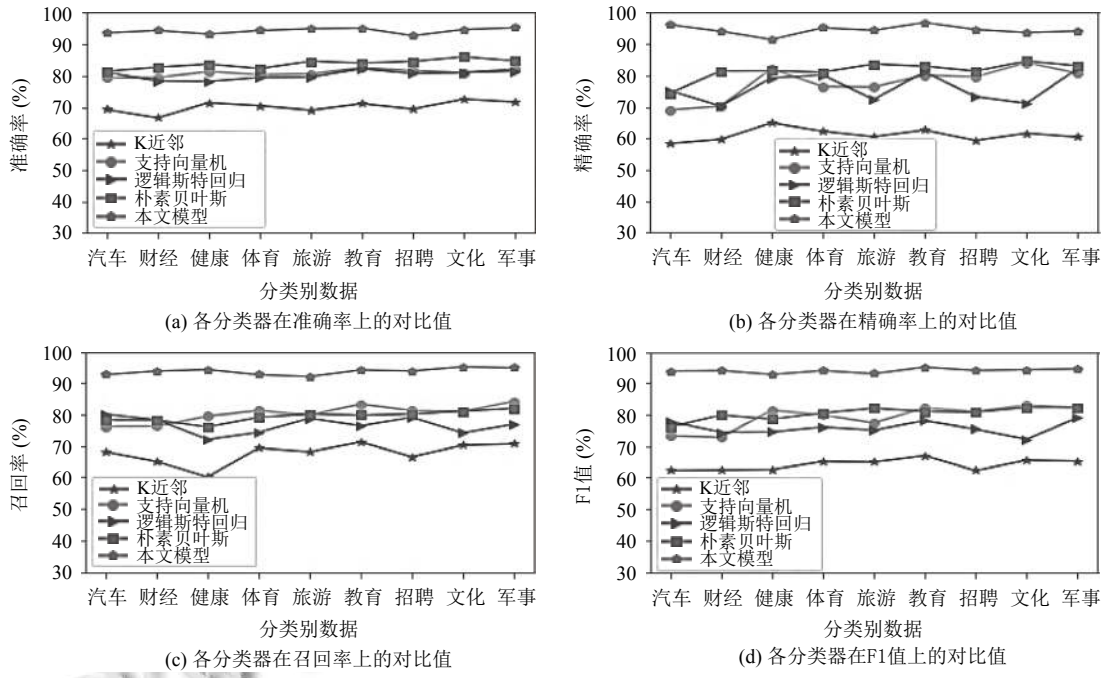
(a) 各分类器在准确率上的对比值 (b) 各分类器在精确率上的对比值

(c) 各分类器在召回率上的对比值 (d) 各分类器在F1值上的对比值

图 2 电影评论二分类性能显示

表 3 SOUGOU 语料库多分类性能比较

分类算法	评测标准	多类别数据								
		汽车	财经	健康	体育	旅游	教育	招聘	文化	军事
K 近邻	准确率	69.21	66.58	71.32	70.38	68.99	71.01	69.35	72.49	71.58
	精确率	58.31	59.68	64.88	62.24	60.54	62.62	59.33	61.64	60.42
	召回率	68.17	65.11	60.12	69.34	68.14	71.28	66.58	70.26	70.69
	F1 值	62.26	62.42	62.58	65.14	64.94	66.99	62.24	65.53	65.25
支持向量机	准确率	79.33	79.39	81.23	80.29	80.56	82.47	81.55	80.89	82.10
	精确率	69.01	70.24	82.13	76.49	76.32	80.06	79.58	84.12	80.71
	召回率	76.25	76.34	79.48	81.24	79.84	83.31	81.24	81.01	84.16
	F1 值	73.34	72.88	81.46	79.93	77.39	82.19	81.02	83.10	82.17
逻辑斯特回归	准确率	81.06	78.23	78.19	79.44	79.56	82.10	80.61	80.97	81.42
	精确率	75.24	70.26	79.21	80.14	72.39	81.10	73.26	71.13	82.16
	召回率	80.11	78.44	72.11	74.32	78.89	76.49	79.18	74.16	76.98
	F1 值	77.86	74.53	74.56	76.14	75.14	78.13	75.44	72.23	79.01
朴素贝叶斯	准确率	81.31	82.63	83.62	82.19	84.56	83.97	84.51	85.94	84.68
	精确率	74.31	81.39	81.48	81.03	83.65	82.95	81.37	84.60	83.09
	召回率	78.34	78.30	76.11	79.16	80.17	79.84	80.14	81.07	81.94
	F1 值	76.12	79.89	78.60	80.61	82.06	81.08	80.84	82.34	82.47
本文模型	准确率	93.54	94.31	93.16	94.29	94.89	95.10	92.62	94.67	95.18
	精确率	96.17	94.06	91.42	95.17	94.38	96.87	94.63	93.64	94.13
	召回率	92.64	93.82	94.30	92.63	92.05	94.16	93.86	95.20	94.86
	F1 值	93.88	94.06	92.84	94.04	93.07	95.17	94.09	94.37	94.62



(a) 各分类器在准确率上的对比值

(b) 各分类器在精确率上的对比值

(c) 各分类器在召回率上的对比值

(d) 各分类器在F1值上的对比值

图 3 Sougou 语料库多分类性能展示

3 总结

本文通过提出一种全新的文本表示模型 (N-of-DOC), 借助 Word2Vec 训练词向量, 将其表示出的词向量特征作为改进的卷积神经网络模型的输入. 该方法不仅能解决文本向量的维数灾难、局部最优解以及过学习问题, 而且有利于卷积神经网络组合低层特征形成更加抽象的高层表示. 将传统机器学习分类方法与本文提出的改进的卷积神经网络文本分类方法进行对比实验, 弥补了传统机器学习文本分类方法的不足, 提高了文本分类的准确率.

在今后的研究中, 由于本文的方法虽然在深度学习上使用了 tensorflow 的 GPU 加速功能, 但对于文本处理还需很长时间, 面对今后海量的数据分类实用性比较低. 因此, 在今后的工作中, 如何采用分布式平台进行深度学习的互联网短文本分类将是笔者的研究重点, 不仅能在分类精度上可以做到显著提高, 在分类速度上也可以提高效率.

参考文献

1 刘冬瑶, 刘世杰, 陈宇星, 等. 新闻文本自动分类技术概述. 电脑知识与技术, 2017, 13(35): 87–91.

2 Kim Y. Convolutional neural networks for sentence classification. arXiv: 1408.5882, 2014.

3 黄文明, 莫阳. 基于文本加权 KNN 算法的中文垃圾短信过滤. 计算机工程, 2017, 43(3): 193–199. [doi: 10.3969/j.issn.1000-3428.2017.03.033]

4 Beck J, Dia BM, Espath LFR, *et al*. Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain. Computer Methods in Applied Mechanics and Engineering, 2017, 334: 523–553.

5 唐伟, 刘丰年, 陈崇帮, 等. 改进的基尼指数在文本分类中的应用研究. 长沙大学学报, 2013, 27(5): 55–57, 63. [doi: 10.3969/j.issn.1008-4681.2013.05.019]

6 李军政, 黄海, 黄瑞阳, 等. 基于卡方检验和 SVM 的用户搜索画像技术研究. 电子设计工程, 2017, 25(24): 6–10. [doi: 10.3969/j.issn.1674-6236.2017.24.002]

7 陶伟. 警务应用中基于双向最大匹配法的中文分词算法实现. 电子技术与软件工程, 2016, (4): 153–155.

8 官琴, 邓三鸿, 王昊. 中文文本聚类常用停用词表对比研究. 数据分析与知识发现, 2017, 1(3): 72–80.

9 Xu CZ, Liu D. Chinese text summarization algorithm based on Word2Vec. Journal of Physics: Conference Series, 2018, 976(1): 012006.