doi:10.3969/j.issn.1001-4616.2022.01.019

基于差分进化的卷积神经网络的文本分类研究

钟桂凤[1],庞雄文[2],孙道宗[3]

(1.广州理工学院计算机科学与工程学院,广东 广州 510540)
(2.华南师范大学计算机学院,广东 广州 530631)
(3.华南农业大学 电子工程学院,广东 广州 510642)

[摘要] 为了提高文本分类的性能,采用差分进化的卷积神经网络(convolutional neural network,CNN)算法进行分类. 首先随机设置 CNN 结构参数,然后采用差分进化算法优化参数,通过交叉和选择等操作选择不断进化获得最优个体,为增强差分优化的适用性,将缩放因子变化与进化代数相关联,解决了因为缩放因子设置不合理而造成优化等级不高的问题. 卷积神经网络采用经过差分优化后的权重和阈值对文本进行分类训练,以获得稳定的文本分类结果. 实验证明,通过合理设置差分进化交叉速率和卷积神经网络的卷积核尺寸,能够获得较好的分类准确率性能,RMSE 值更低,在文本分类中的适用度高.

[关键词] 分类,差分进化,卷积神经网络,缩放因子,卷积核
[中图分类号]TP391.1 [文献标志码]A [文章编号]1001-4616(2022)01-0136-06

Research on Text Classification Based on Convolutional Neural Network of Differential Evolution

Zhong Guifeng[1], Pang Xiongwen[2], Sun Daozong[3]

(1.College of Computer Science & Engineering, Guangzhou Institute of Science and Technology, Guangzhou 530631, China)
(2.School of Computer Science, South China Normal University, Guangzhou 530631, China)
(3.School of Electronic Engineering, South China Agricultural University, Guangzhou 530631, China)

Abstract: In order to improve the performance of text classification, differential evolution convolutional neural network (CNN) algorithm was used for text classification. Firstly, the structural parameters of CNN were set randomly, then the parameters were optimized by differential evolution algorithm, and the optimal individuals were obtained by continuous evolution through crossover and selection. In order to enhance the applicability of differential optimization, the change of scaling factor was associated with evolutionary algebra, which solved the problem of low optimization level caused by unreasonable setting of scaling factor. Convolutional neural network used the weight and threshold after differential optimization to train text classification, so as to obtain stable text classification results. Experimental results showed that by setting the crossover rate of differential evolution and convolution kernel size of convolution neural network reasonably, better classification accuracy performance can be obtained, RMSE value was lower, and applicability in text classification was high.

Key words: classification, differential evolution, convolution neural network, scaling factor, convolution kernel

互联网飞速发展,大规模网络文本数据分析研究应运而生,通过分类,将网络中的数据进行归档整理,提高数据管理的有效性. 网络文本的格式不标准及编码方式的多样性[1],以及文本长度的差异性,使得文本分类的难度提升. 在采用深度学习的文本分类中,由于普通文本不同于变量属性特征,可以直接进行输入并通过深度学习训练而获得结果,在文本分类之前,需要通过向量转换,然后进行训练获得分类结果,这种方法对分类精度造成了影响,因此需要进一步优化深度学习算法,以提高深度学习对文本分类的适用度.

收稿日期:2021-05-07.
基金项目:国家自然科学基金项目(31101077)、2020 年度广东省高校科研项目(2020GXJK201).
通讯作者:钟桂凤,讲师,研究方向:数据分析与挖掘,人工智能,机器学习. E-mail:109488818@qq.com

当前关于文本分类的研究较多,于游等[2]对常见的中文文本分类方法做了系统阐述,比较了各类方法对于不同文本的适用度及优缺点,给后续文本分类算法研究提供了借鉴;郭超磊等[3]采用 SVM 方法进行中文文本分类,可以达到一定的分类效果,但分类准确度不高,且对分类样本的格式要求严格,对网络各种符号及文字混合的文本分类适用度不高;Shu 等[4]对学习推荐开展了研究,提出了基于潜在因素模型的 CNN 文本分类模型,分类精度高,但是基于内容的 CNN 结构存在分类效率及稳定性不理想的问题. 本文采用卷积神经网络对普通文本进行分类,为了提高文本分类的性能,引入差分进化(differential evolution, DE)算法对网络参数进行优化求解,并对差分算法的缩放因子采取自适应策略,以提高优化求解精度,通过差分进化优化的卷积神经网络算法,可以有效提高文本分类精度及 RMSE 性能.

1 理论基础与模型设计

1.1 自适应 DE 算法

设种群规模为 N,属性维度为 D,差分缩放因子为 F,交叉速率 CR,每个个体的取值为$[U_{\min},U_{\max}]$,则第 i 个个体 的 j 维属性可表示为[5]:

$$x_{ij}=U_{\min}+rand\times(U_{\max}-U_{\min}), \tag{1}$$

式中,$i=1,2,\cdots,N$,$j=1,2,\cdots,D$,$rand$ 为$(0,1)$随机数.

设第 G 代的个体 $x_i^G(i=1,2,\cdots,N)$的变异操作得到的 $G+1$ 代个体为[6]:

$$v_i^{G+1}=x_{r_1}^G+F\times(x_{r_2}^G-x_{r_3}^G), \tag{2}$$

式中,$i\neq r_1\neq r_2\neq r_3$,$r_1$、$r_2$ 和 r_3 为第 G 代中除了编号为 i 的个体之外的随机 3 个个体. F 常见取值$[0,2]$.

个体交叉方法为[7]:

$$u_{ij}^{G+1}=\begin{cases}v_{ij}^{G+1}, & \text{rand}(0,1)\leqslant CR,\\ x_{ij}^G, & \text{otherwise.}\end{cases} \tag{3}$$

对比 x_i^G 与 u_i^{G+1},分别求取两个个体的适应度值,选择两者中值较高的个体进行进化. 具体方法为:

$$x_i^{G+1}=\begin{cases}u_i^{G+1}, & f(u_i^{G+1})>f(x_i^G),\\ x_i^G, & f(u_i^{G+1})\leqslant f(x_i^G),\end{cases} \tag{4}$$

式中,f 表示适应度函数. 当达到最大代数 $G_{\max}$时,DE 算法停止.

F 常见取值$[0,2]$,DE 的优化过程与 F 值密切,F 值不合适将会造成差分进化算法的优化性能不高的问题,因此在计算时引入自适应 F 值[8]. $F_{\min}$和 $F_{\max}$范围为$[0,2]$,则:

$$F=F_{\min}+(F_{\max}-F_{\min})\times e^{1-\frac{G_{\max}}{G_{\max}-G+1}}. \tag{5}$$

F 值随着进化代数 G 的变化而逐渐变小,前期进化追求种群多样化,后期注重搜索能力,这样 DE 算法更容易获得最优个体.

1.2 卷积神经网络模型设计

设文本样本集 $X=(x_1,x_2,\cdots,x_N)$,m 个文本属性特征通过第 l 层卷积运算得[9]:

$$x_{lj}=f\left(\sum_{j\in m}x_{l-1}*k_{lj}+b_{lj}\right), \tag{6}$$

式中,k_{lj}和 b_{lj}分别表示 l 层对特征 j 赋予的权重及偏置,$*$ 为卷积,$f(\cdot)$为:

$$f(z)=\frac{1}{1+e^{-z}}. \tag{7}$$

对 N 个样本的 m 个特征进行卷积,卷积核尺寸(kernel size)$h\times w$,按照公式(8)进行:

$$g(x)=\max_{1\leqslant k\leqslant h\times w}(x_k). \tag{8}$$

令 $M=N/(h\times w)$,那么原样本 $X=(x_1,x_2,\cdots x_N)$经过卷积池化后重新得到的样本为 $X'=(x_1,x_2,\cdots,x_M)$.

然后 X'进行转换运算[10]:

$$x_j^l=f\left(\sum_{i=1}^{M}a_{ij}(x_i^{l-1}*k_i^l)+b_j^l\right). \tag{9}$$

限制条件为：$\sum a_{ij}=1, 0\leqslant a_{ij}\leqslant 1$.

根据公式(9)得到 CNN 所有连接层,最后选择分类器预测样本类别.

设第 k 个节点的训练输出和实际值分别为 y_k 和 d_k,则误差项 δ_k 为：

$$\delta_k=(d_k-y_k)y_k(1-y_k). \tag{10}$$

假设第 l、l+1 层分别包含 L 和 P 个节点,则第 l 层节点 j 的误差为[11]：

$$\delta_j=h_j(1-h_j)\sum_{k=1}^{P}\delta_k W_{jk}, \tag{11}$$

式中,h_j 为输出,W_{jk}为神经元 j 到 l+1 层神经元 k 的权重,更新方法为：

$$\Delta w_{jk}(n)=\frac{\eta}{1+N}(\Delta w_{jk}(n-1)+1)\delta_k h_j, \tag{12}$$

式中,η 为学习率.

偏置 $\Delta b_k(n)$ 的更新方式为[12]：

$$\Delta b_k(n)=\frac{\alpha}{1+N}(\Delta b_k(n-1)+1)\delta_k, \tag{13}$$

式中,α 为偏置更新步长,一般 $\alpha=1$. 调整后的权重为：

$$w_{jk}(n+1)=w_{jk}(n)+\Delta w_{jk}(n). \tag{14}$$

调整后的偏置为：

$$b_k(n+1)=b_k(n)+\Delta b_k(n). \tag{15}$$

所有节点的误差 E 为：

$$E=\frac{1}{2}\sum_{k=1}^{M}(d_k-y_k)^2. \tag{16}$$

当 E 满足设定的阈值,迭代停止,获得稳定的 CNN 模型.

1.3　DE-CNN 模型的分类流程

在运用 CNN 对文本进行分类之前,首先需要对待分类的样本数据进行 word2vec 转换[13],这主要是为了解决文本属性的向量化过程,转换后的 Skip-gram 便于进行 CNN 的有效输入. 建立了 CNN 文本分类模型之后,将随机权重和偏置通过 DE 算法优化求解,根据文本分类准确度函数建立适应度函数,通过 DE 的多代进化,获得权重和偏置最优个体,最后 CNN 进行分类训练获得文本分类结果.

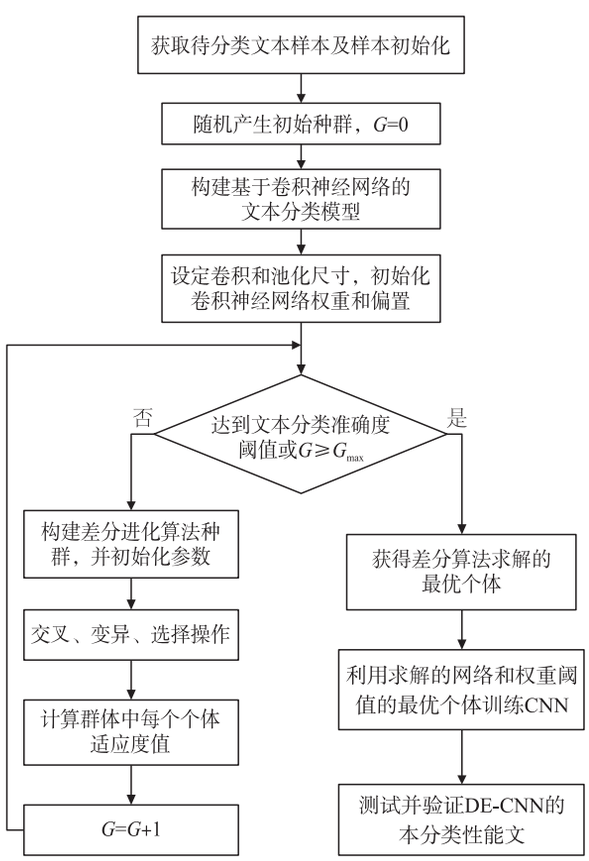

图 1　分类流程

Fig. 1　Classification process

2　实例仿真结果与分析

为了验证差分进化的卷积神经网络算法在文本分类中的性能,进行实例仿真. 首先,对不同的差分进化算法参数进行性能仿真,其次对不同卷积核尺寸的性能进行仿真,最后将本文算法与常用文本分类算法进行性能对比仿真.

文本分类仿真的数据来源为 SST(stanford sentiment treebank)和 THUCnews 新闻数据. 其中 SST 数据样本 11 852 个,5 个类别;而 THUCnews 选取了 7 类共计 10 500 个新闻样本. 通过算法对新闻文档进行分类,从而能够实现新闻自动归档. 样本具体分布结构如表 1 所示.

表 1　THUCnews 文本集

Table 1　THUCnews text set

文档类别	数量	总词数	非重复词数	文档类别	数量	总词数	非重复词数
财经	1 500	532 967	38 427	旅游	1 500	514 725	32 436
IT	1 500	315 738	23 537	文化	1 500	903 718	57 932
教育	1 500	843 781	31 269	军事	1 500	716342	33 209
体育	1 500	212 862	20 351				

对从表 1 中的文本采用 word2vec 得到 Skip-gram 结构,从而完成了本文至属性向量映射,这样文本样本就可以进行 CNN 分类训练. 在仿真过程中,THUCnews 和 SST 样本集分别按照总样本容量的 7:2:1 的比例数量进行训练、测试和验证.

本文 DE 算法设置的初值 $F_{\min}=0.2$, $F_{\max}=0.9$, $CR=0.1$, $G_{\max}=100$. CNN 卷积核默认 2 * 2.

2.1　不同卷积尺寸文本分类性能

采用不同 kernel size 的 CNN 结构分别对 THUCnews 和 SST 样本进行仿真.

从表 2 可得,选择卷积核尺寸为 3 * 3 效果最佳,THUCnews 的数据样本分类准确率到了 92.16%,而 SST 数据样本分类准确率达到了 93.27%. 当尺寸增大时,2 个数据集的分类准确率和标准差均在下降,这是因为卷积尺寸过大,造成了卷积粒度大减少了样本重要属性参与卷积及转换运算的机会.

表 2　分类准确率

Table 2　Classification accuracy

数据集	类别数	准确率	RMSE	数据集	类别数	准确率	RMSE
卷积核尺寸 2 * 2				卷积核尺寸 4 * 4			
THUCnews	7	0.916 4	0.186 2	THUCnews	7	85.193 7	0.233 4
SST	5	0.924 1	0.171 3	SST	5	87.715 3	0.210 7
卷积核尺寸 3 * 3				卷积核尺寸 5 * 5			
THUCnews	7	0.921 6	0.184 7	THUCnews	7	68.617 1	0.521 9
SST	5	0.932 7	0.168 5	SST	5	71.915 3	0.479 3

对比发现,DE-CNN 算法对 SST 的分类性能优于 THUCnews 数据集,这可能是因为 THUCnews 类别数较多而不易分类造成的. 当卷积核尺寸为 3 * 3 时,DE-CNN 算法在 THUCnews 和 SST 数据集的收敛时间性能如图 2 所示.

从图 2 得,在卷积核设置为 3 * 3 时,DE-CNN 算法在 THUCnews 数据集的分类时间约为 55s,而在 SST 数据集的分类时间约为 50s,这主要是因为 THUCnews 集的类别比 SST 集多的原因. 2 个样本收敛时的分类准确率均超过了 0.9.

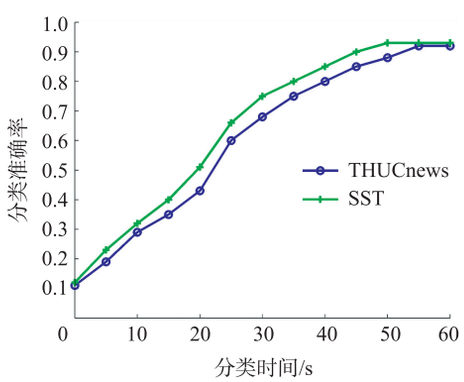

图 2　分类准确率(卷积核 3 * 3)

Fig. 2　Classification accuracy(convolution kernel 3 * 3)

2.2　DE 算法的优化性能

为了验证 DE 算法对 CNN 的优化性能,分别采用 CNN 算法和 DE-CNN 算法对 THUCnews 集和 SST 集的样本进行性能仿真.

从表 3 可以看出,在 3 种不同数据集的文本分类中,经过了 DE 优化的 CNN 算法表现出了更优的性能. 对于 3 个样本集,DE-CNN 文本分类的 3 个指标均超过了 0.9. DE-CNN 的最大分类准确率为 93.18%,CNN 最大分类准确率仅为 88.96%,准确率提升明显. 这主要是因为经过 DE 的权重优化后,CNN 获得了更优的权重和偏置初值,从而获得了更准确的文本分类性能,下面将继续对两种算法的收敛性能进行对比.

从图 3 和 4 得,DE-CNN 相比于 CNN 的收敛性能优势明显. 在 THUCnews 数据样本分类中,DE-CNN 收敛时 RMSE 约为 0.18,而 CNN 收敛的 RMSE 值约为 2.5;而在

表 3　CNN 和 DE-CNN 算法的分类性能

Table 3　Classification performance of CNN and DE-CNN algorithms

数据集	算法	准确率	召回率	F1 值
THUCnews	CNN	0.864 6	0.847 3	0.806 4
	DE-CNN	0.927 5	0.901 4	0.901 2
SST	CNN	0.889 6	0.872 5	0.823 5
	DE-CNN	0.931 8	0.916 8	0.903 7

SST 数据样本分类中,DE-CNN 收敛时 RMSE 约为 0.16,而 CNN 收敛的 RMSE 值约为 2.2,因此 DE-CNN 算法相比于 CNN 算法的分类稳定性更好. 在收敛时间方面,对于 2 种不同的样本集,CNN 比 DE-CNN 收敛的时间少 5s 左右,这可能是因为 DE 算法求解最优权重和偏置的时间消耗,但从整个 DE-CNN 分类时间来看,DE 算法消耗的时间占比很小,对文本分类时间影响较小.

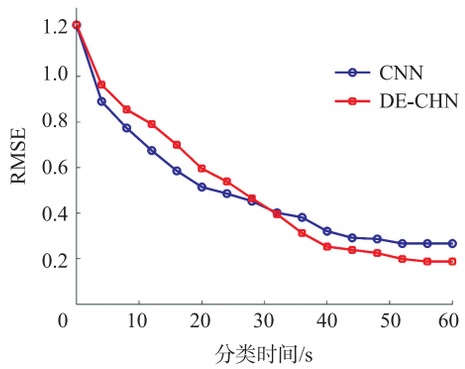

图 3　2 种算法的 RMSE 值(THUCnews 集)

Fig. 3　RMSE values of the two algorithms(THUCnews set)

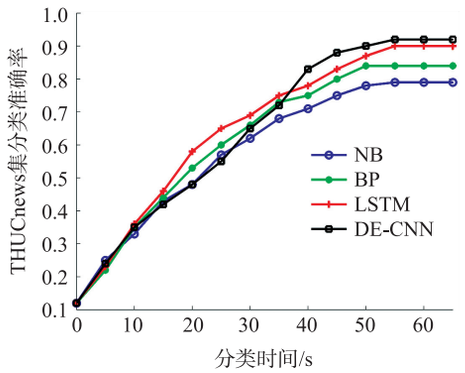

图 4　2 种算法的 RMSE 值(SST 集)

Fig. 4　RMSE values of two algorithms(SST set)

2.3　不同算法的文本分类性能

采用常用朴素贝叶斯(NB)[14]、神经网络(BP)[15]、LSTM 神经网络(LSTM)[16]和本文算法分别对 THUCnews 和 SST 数据集进行仿真.

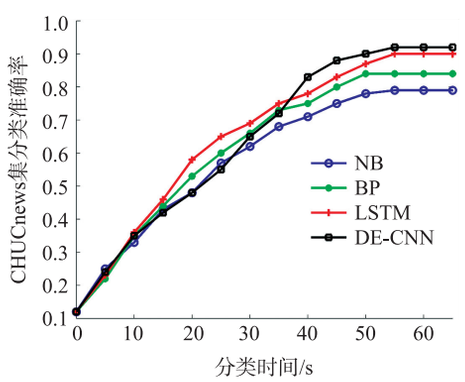

图 5　4 种算法的分类准确率(THUCnews 数据集)

Fig. 5　Classification accuracy of four algorithms (THUCnews dataset)

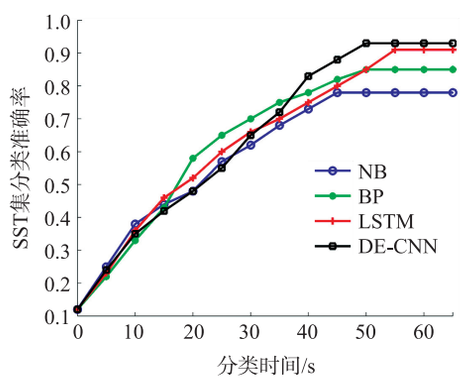

图 6　4 种算法的分类准确率(SST 数据集)

Fig. 6　Classification accuracy of four algorithms (SST dataset)

从文本的分类准确率来看,DE-CNN 和 LSTM 算法的分类准确率最高,稳定时两者的分类准确率非常接近,且均超过了 0.9,NB 的分类准确率最差,均小于 0.8. 从分类时间方面来看,对于 THUCnews 和 SST 数据集,LSTM 算法消耗时间最长,DE-CNN 算法次之,NB 算法最省时.

下面继续对 4 种算法在文本的分类稳定性进行仿真,验证 4 种算法的准确率 RMSE 性能.

从表 4 的 RMSE 性能中可以看出,对于 2 种数据集,DE-CNN 算法的分类准确率 RMSE 值最优,NB 表现最差. 相比而言,4 种算法在 SST 集的 RMSE 性能表现更优,这可能是因为 SST 集待分类的类别数较少,而 THUCnews 需要分类的类别数较多,在文本分类时,类别过多造成了分类准确率值在多次分类中波动较大,这也说明分类准确率 RMSE 值对分类类别数影响敏感. 综合而言,对于 THUCnews 新闻集和 SST 情感集的文本分类,对比常见分类算法,在获得较高分类准确率的条件下,本文算法仍能获得较好的分类时间和 RMSE 性能.

表 4　不同算法的准确率 RMSE 性能

Table 4　Accuracy and RMSE performance of different algorithms

数据集	分类 RMSE 值			
	NB	BP	LSTM	DE-CNN
THUC news	0.368 5	0.269 2	0.253 9	0.184 1
SST	0.325 1	0.237 9	0.221 7	0.160 5

3 结论

采用差分进化的卷积神经网络算法应用于文本分类,充分利用差分进化算法的权重优化求解优势,提高了卷积神经网络算法在文本分类中的适用度,相比于常用文本分类算法,本文算法在分类准确率及RMSE性能方面优势明显. 后续研究将进一步调整差分进化参数,以提高文本分类时间性能.

[参考文献]

[1] TANG X C,DAI Y S,XIANG Y P,et al. Feature selection based on feature interactions with application to text categorization[J]. Expert systems with applications,2019,120:207-216.

[2] 于游,付钰,吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报,2019(5):1-8.

[3] 郭超磊,陈军华. 基于SA-SVM的中文文本分类研究[J]. 计算机应用与软件,2019(3):277-281.

[4] SHU J B,SHEN X X,LIU H,et al. A content-based recommendation algorithm for learning resources[J]. Multimedia systems,2018,24(2):163-173.

[5] 孙越泓,王丹. 基于新约束集成的差分进化算法[J]. 南京师大学报(自然科学版),2019,42(4):1-11.

[6] 王永安,赵阳,蓝雨晨,等. 基于进化差分算法的环形电感建模及应用[J]. 南京师范大学学报(工程技术版),2020,20(3):32-37.

[7] SLW A,FM A,TFN B,et al. Insights into the effects of control parameters and mutation strategy on self-adaptive ensemble-based differential evolution-ScienceDirect[J]. Information sciences,2020,514:203-233.

[8] QLAB C,SD C,BJVW D,et al. Double-layer-clustering differential evolution multimodal optimization by speciation and self-adaptive strategies-science direct[J]. Information sciences,2021,545:465-486.

[9] AMIT K S,SANDEEP C,DEVESH K S. Sentimental short sentences classification by using cnn deep learning model with fine tuned Word2Vec[J]. Procedia computer science,2020,167:1139-1147.

[10] PAN R L,YU C,ZHAO W,et al. Multi-channel Sliced Deep RCNN with residual network for text classification[J]. Chinese journal of electronics,2020,29(5):92-98.

[11] LIU Y,SUEN C Y,LIU Y,et al. Scene classification using Hierarchical wasserstein CNN[J]. IEEE Transactions on geoscience and remote sensing,2019,57(5):2494-2509.

[12] 沈浩,江臣,陈宇文,等. 基于深度学习的钢桁架桥螺栓病害智能识别方法[J]. 南京工业大学学报(自然科学版),2020,42(5):608-614.

[13] NADERALVOJOUD B,SEZER E A. Term evaluation metrics in imbalanced text categorization[J]. Natural Language engineering,2019,26(1):1-17.

[14] 梁柯,李健,陈颖雪,等. 基于朴素贝叶斯的文本情感分类及实现[J]. 智能计算机与应用,2019(5):150-153,157.

[15] 钱鹏,陆金桂. 基于PSO-BP神经网络的红外无损检测缺陷定量识别[J]. 南京工业大学学报(自然科学版),2019,41(4):501-507.

[16] 田园,原野,刘海斌,等. 基于BERT预训练语言模型的电网设备缺陷文本分类[J]. 南京理工大学学报,2020,44(4):446-453.

[责任编辑:顾晓天]