水力发电学报
Journal of Hydroelectric Engineering
ISSN 1003-1243,CN 11-2241/TV

《水力发电学报》网络首发论文

题目：基于字符级 CNN 的调水工程巡检文本智能分类方法
作者：刘婷，张社荣，李志竑，关炜
收稿日期：2020-12-01
网络首发日期：2021-01-14
引用格式：刘婷，张社荣，李志竑，关炜．基于字符级 CNN 的调水工程巡检文本智能分类方法．水力发电学报.
https://kns.cnki.net/kcms/detail/11.2241.TV.20210113.1847.004.html

水 力 发 电 学 报

Journal of Hydroelectric Engineering

基于字符级 CNN 的调水工程巡检文本智能分类方法

刘婷[1]，张社荣[1]，李志纮[2]，关炜[2]

（1. 天津大学 水利工程仿真与安全国家重点实验室，天津 300350；2. 水利部南水北调规划设计管理局，北京 100038）

摘　要：日常安全巡检是维护长距离调水工程安全运行的重要手段。目前巡检采集的非结构化文本数据主要依靠人工进行安全等级评判，在工作效率和准确率方面存在明显不足。本研究基于自然语言处理技术，提出了一种面向字符层面的卷积神经网络的巡检安全文本智能分类方法。该方法通过引入预训练的单个字符向量改进卷积神经网络的输入层，使得分类模型直接从原始文本中提取特征信息，不仅避免了传统分类方法对专业词库的依赖，而且不易受文本中出现的口语化表达和错别字的影响。以国内某调水工程的巡检文本为案例，通过与多种深度学习算法进行全面比较对比验证了所提方法的有效性和优越性。结果表明，字符级的分类方法明显优于传统基于词的分类方法，且卷积神经网络在巡检文本分类方面明显优于其他深度学习网络，该方法具有较高的分类准确率，以此为调水工程安全维护提供新的智能化手段。

关键词：调水工程；文本分类；字符向量化；卷积神经网络；自然语言处理

中图分类号：TV68　　**文献标志码：**A

Intelligent text classification method for water diversion project inspection based on character level CNN

LIU Ting[1], ZHANG Sherong[1], LI Zhihong[2], GUAN Wei[2]

1. State Key Laboratory of Hydraulic Engineering Simulation and Safety , Tianjin University, Tianjin 300350;

2. Bureau of South to North Water Transfer of Planning, Designing and Management, Ministry of Water Resources, Beijing, 100038

Abstract: Daily safety inspection is an important means to maintain the safe operation of long-distance water diversion project. At present, unstructured text data collected by patrol inspection mainly rely on manual safety level evaluation, which has obvious deficiencies in work efficiency and accuracy. Based on the natural language processing technology, this paper proposes a character oriented convolutional neural network intelligent text classification method. This method improves the input layer of convolutional neural network by introducing a pre-trained single character vector,

收稿日期：2020-12-01　　**接受日期：**2021-01-12

基金项目：国家重点研发计划项目（2018YFC0406905）

which makes the classification model directly extract feature information from the original text. It not only avoids the dependence of traditional classification methods on the professional lexicon, but also is not easily affected by the colloquial expression and typographical errors in the text. Taking the inspection text of a domestic water diversion project as an example, the effectiveness and superiority of the proposed method are verified by comprehensive comparison with various deep learning algorithms. The results show that the character level classification method is obviously better than the traditional word based classification method, and the convolution neural network is obviously better than other deep learning networks in the text classification of patrol inspection. The method has high classification efficiency and accuracy, which provides a new intelligent means for the safety maintenance of water diversion project.

Key words: water diversion project; text classification; character vectorization; convolution neural network; natural language processing

0 引言

调水工程由于其距离长、影响范围广的特点，建筑物的日常安全巡检极为重要。建筑物长期运行过程中通过巡检手段记录并积累了大量的非结构化巡检安全文本数据，这些数据实时存入数据库中通常作为建筑物险情处理的依据。巡检安全文本数据通常包括问题描述和问题等级等信息，现阶段巡检安全数据的等级分类主要依靠人工作业，不仅费时费力，而且由于巡检人员的知识储备和经验有限，对于一些模糊性较强的数据难以精确分类，因此需要采用智能化手段对巡检文本等级进行自动化判别。随着自然语言处理技术的快速发展，采用文本分类技术使得调水工程巡检过程中记录的大量安全文本数据进行自动分类成为可能。该技术的应用不仅能够减轻甚至消除人工分类的工作量、提高分类效率，而且能够在提高分类准确度的同时保证险情能被及时处理和上报，从而保证建筑物安全运行。

文本分类技术是自然语言处理中一项经典工作，它可以将非结构化的数据按照一定的分类体系或者标准分为有意义的结构类[1]。目前该技术已经应用到电力[2,3]、新闻[4,5]、医学[6,7]、金融[8]、农业[9]等领域，但在水利领域内鲜有研究。传统的中文文本分类方法往往采用浅层机器学习算法，如决策树[10]、支持向量机[11]、最大熵[12]、K 近邻算法[13]、贝叶斯[14]等，通过人工来设计特征选择方法进而对特征进行提取，然而此类方法不仅人工成本高、耗时长、训练难，而且难以应用到海量数据场景中。随着工程的持续运行，文本数据日益增长，如何对海量非结构化数据进行高效的自动分类成为亟需解决的问题。

基于深度学习的文本分类技术能快速从海量文本数据中自动进行特征学习与提取，大幅度降低人工成本且易于训练[15]。其中卷积神经网络(CNN)[16]、循环神经网络（RNN）[17]、深度金字塔卷积神经网络（DPCNN）[18]、FastText[19]、Transformer[20]等是广泛应用于文本分类等自然语言处理任务中的方法，并且取得了比传统机器学习算法更优越的性能[21]。现有的文本分类方法一般都是以上几种方法的组合或者改进，例如汪嘉伟等[22]基于词级别的 CNN 融合自注意力机制提出了一种新的文本分类模型 Word-CNN-Att，算例结果表明该模型的分类准确率优于传统 CNN。

Du Jie 等[23]提出了一种新型高效的 RNN 和类 LSTM 架构，通过同时学习序列信息和单词重要性等多种信息，可以实现文本分类的高精度。Li 等[24]通过在 DPCNN 中引入了自注意机制来提取文本的局部特征，提高了分类的准确性。Ling 等[25]通过对 Facebook 2016 年提出的 FastText 模型的分类精度和参数进行研究分析，得出了 FastText 的优化规则。唐庄等[26]提出了一种 transformer-capsule 集成模型，通过集成来更全面的提取文本序列的多层次特征，提高分类性能。

然而以上方法都不能直接应用于水利领域，主要存在以下几点局限性：1）目前大多数文本分类任务都是基于词的（例如 Glove[27]、Word2vec[28]等），需要预训练好的词向量和专业词库信息，而目前面向水利领域的自然语言处理技术尚未有公开的开源专业词库。2）自训练的专业词库往往泛化性较差，随着科技的日新月异，海量文本中常出现一些未能被及时收录进专业词库的未登录词，进而影响分类的准确性。3）工程巡检安全文本往往由巡检人员手工录入，不可避免的存在无法通过专业词库识别的口语化表达和错别字，此外巡检安全文本通常是由几个到几十个词短文本组成，长度短、特征稀疏，直接采用常规分类方法准确率较低。

为了解决以上问题本文提出了一种字符级卷积神经网络（Char-CNN, Character-level convolutional neural network）进行调水工程巡检安全文本智能分类。通过预训练单个字符向量嵌入卷积神经网络的输入层来帮助模型从原始文本中提取特征信息，此过程无须预训练和收录好的词向量。Char-CNN 一方面能够避免依赖专业词库和中文分词等辅助工作，另一方面其在噪声较大的短文本中表现优良，表明该方法能在一定程度上能解决未登录词问题。最后通过与传统 CNN 和嵌入字符级向量的其他深度学习方法的全面对比，证明了本文提出的 Char-CNN 的可行性和优越性。

1 文本分类流程

文本分类是指给定需要分类的序列文本 $D=\{d_1,d_2,\ldots,d_m\}$，在给定类别集合 $C=\{c_1,c_2,\ldots,c_n\}$ 中，存在一个函数 R: $D\times C\rightarrow\{T,F\}$，其中 T 值表示对于 (d_i,c_j) 来说文档 d_i 属于类别 c_j，反之用 F 值表示。根据调水工程巡检安全记录文本的特点，文本分类的一般流程如图 1 所示，主要分为四步，具体为：

（1）预处理：不同于英文文本词语之间有空格划分，中文文本没有汉字或词语之间的分割符号，因此首先需对训练集和测试集的语料库进行分词、分字，同时去除语气词、标点符号等干扰词汇。

（2）字符向量化：字符向量化是指将文本信息映射到计算机可识别和读取的数字化语义空间中，本研究采用预训练好的嵌入式（pre-train embedding）字符向量对文本进行向量化表示。

（3）特征提取：文本数据集特征的集合通常巨大，而且存在对分类效果有影响的噪声特征，本研究在 CNN 的卷积层采用了字符级别的文本进行特征提取和特征降维，不仅避免了由于缺乏水利领域专业词库对分类效果产生的影响，还能够增强文本特征表示的细粒度。

（4）分类建模和效果评估：通过构造 Char-CNN 分类器对训练集进行训练，并选取性能评价指标，选择最优的匹配对文本进行智能分类。

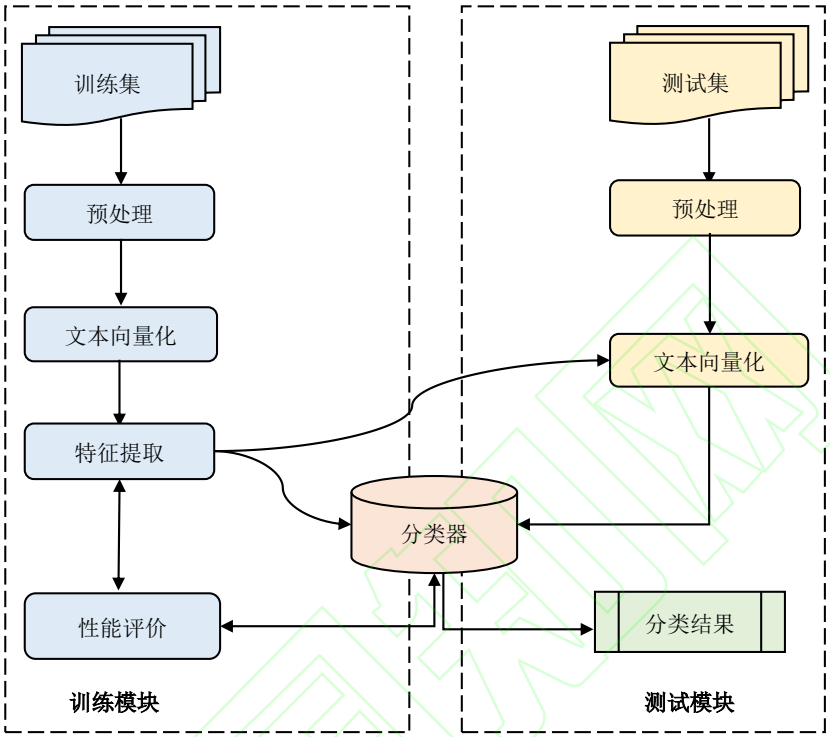

图 1 文本分类流程图

Fig. 1 Flow chart of text classification

2 字符级卷积神经网络文本分类方法

2.1 Char-CNN 模型架构

作为深度学习的经典算法之一，卷积神经网络得到了快速发展，并被普遍应用于计算机视觉和自然语言处理等领域[29,30]。近年来随着智慧水利的大力推进，已有研究将卷积神经网络应用到工程裂缝的图像识别中，并取得了一定进展[31-33]。传统 CNN 具有参数数目少，训练速度快的优点，在处理中文文本分类问题时表现优异。然而该方法在模型训练过程中通常需要预训练的词向量和专业词库的信息，目前在自然语言处理领域尚无水利专业的开源专业词库。因此本文基于传统 CNN 模型架构提出了一种面向字符级的 CNN 方法实现巡检文本的自动分类。Char-CNN 通过预训练单个字符向量帮助模型从原始文本中提取特征信息，此过程将无须预训练和收录好的词向量，避免了传统分类方法对专业词库的依赖。模型架构如图 2 所示。

Char-CNN 模型可分为五层，分别为字符嵌入层、卷积层、池化层、全连接层和输出层。其

中嵌入层相当于传统 CNN 模型中的输入层，只是在文本分类任务中将字符以向量的形式嵌入，因此输入的数据是字符向量堆叠形成的二维矩阵。卷积层用来提取字符向量的特征信息，其维度是二维矩阵与嵌入层相同，卷积核宽和词向量维度相同，长可以自己设置，图 2 中设置了长分别为 2、3、4 三个不同的卷积核来提取字符向量的特征信息，同时能够表达字符的上下文信息。池化层取每一个卷积输出的最大值，通过级联得到最后的特征表达。全连接层将前边提取到的局部特征综合起来。最后通过 softmax 分类器输出最终的分类标签。

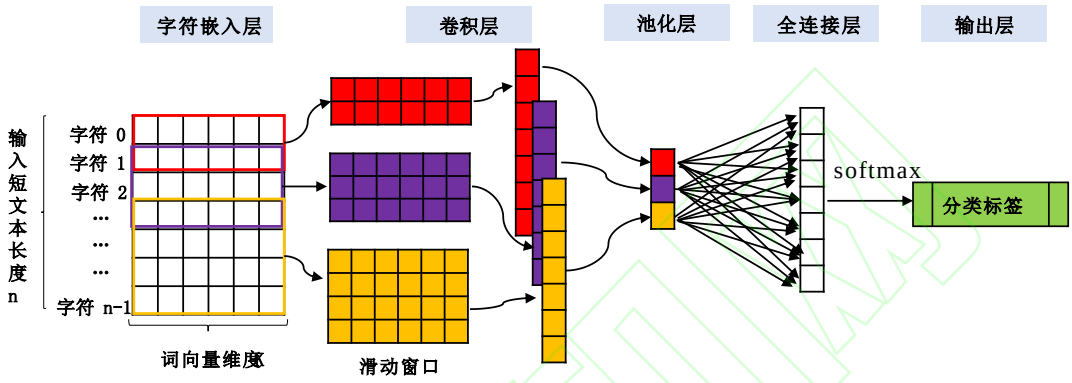

图 2 Char-CNN 模型架构图

Fig. 2 Chart of char CNN model architecture

2.2 预处理和字符向量化

字符嵌入层是作为模型的第一层，能够在训练模型的同时得到该语料库的字符向量。图 3 展示了水利工程巡检不同等级的短文本字符向量化流程，主要分为文本预处理、构建字典和生成字符向量模型三步，具体步骤如下：

（1）文本预处理：去除原始文本的多余的符号、空格和停用词（如的、了、在等无实义的字符），减少文本噪声对分类效率的影响，并对数据集中的每一行文本按照字符进行分割。

（2）构造字典：统计所有字符出现的频率，再将频率大于自定义最小频率的元素按照从高到低的顺序进行排序，最后按照频率降序构建{字符：索引}格式的字典。字典中的最后两个元素是‘<UNK>’和‘<PAD>’，其中 < UNK>和< PAD>是两个初始化的标签，< UNK >用来替代语料中未出现过的单词，< PAD >用来做句子填补，保证句子有相同的长度。与此同时，原始数据集的{一般，重要，紧急}三类标签被符号化表示为{0，1，2}。判断标准是根据该工程印发的《工程巡查技术标准》。图 3 中 S 即指代符号化表示后的标签对应的数字，其中一般被表示为 0，重要被表示为 1，紧急被表示为 2。

（3）生成向量模型：对于给定的输入文本 $Sen=\{c_i,1\le i\le n\}$，其中 c_i 是字典中的字符索引，n 为字符所在的序列长度。在预训练的向量矩阵 M 执行查找表操作（Lookup Table），根据索引获得文本中每个字符的向量表示 $\{v_i,1\le i\le n\}$，对于字典中不存在的字符用’<UNK>’元素表示，由此即将原始文本转换为计算机可识别的编码。构建的字符向量矩阵 $V^{d\times n}=\{v_i,1\le i\le n\}$ 即可

作为 Char-CNN 文本分类模型的输入层，d 为自设定的字符向量的维度，由此完成字符向量化。

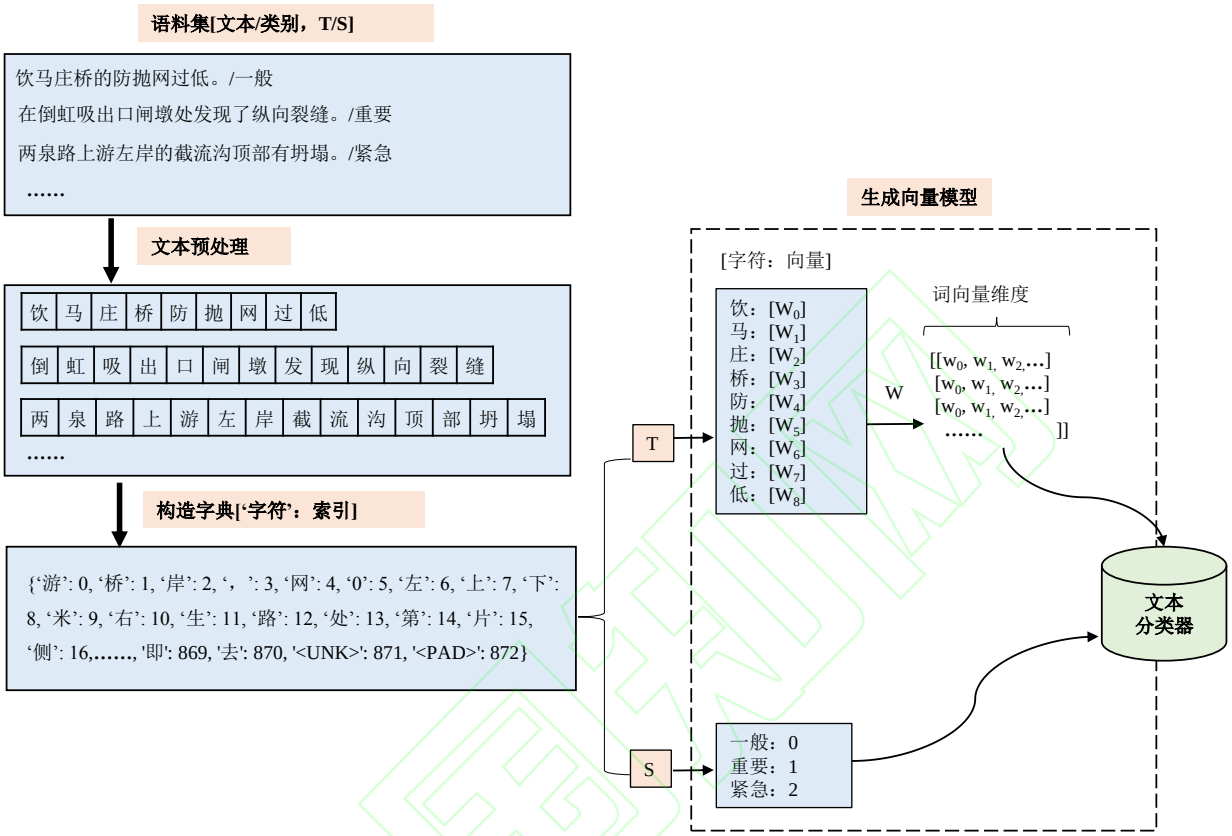

图 3 字符向量化流程图

Fig. 3 Char vectorization flow chart

2.3 特征提取与分类预测

卷积神经网络具有稀疏交互和权值共享的特点，在特征自提取方面表现优异，因此本文选取 Kim 等人在句子分类模型中提出的 CNN 非静态架构（CNN-non-Static）对向量化的字符进行特征提取和分类预测，CNN-non-Static 指在训练时嵌入层跟随整个网络一起训练。

特征提取主要在卷积层和池化层进行，设文本嵌入序列 S：$n=(s_1,s_2,\ldots,s_n)$，其中 n 表示文本长度。定义操作符 $\oplus(s_{i:i+k-1})$ 表示将向量 $s_i,s_{i+1},\ldots,s_{i+k-1}$ 进行拼接，s_i 表示第 i 个字符向量，k 为卷积核大小，则第 i 个窗口的特征提取可表示为：

$$\begin{cases} Y_i = g(x_i \cdot W + b) \\ x_i = \oplus(s_{i:i+k-1}) = [s_i, s_{i+1}, \ldots, s_{i+k-1}] \end{cases} \tag{1}$$

式中：Y_i 表示卷积输出的第 i 个向量；x_i 表示第 i 个窗口的拼接输入向量；W 表示权重矩阵；b 表示偏置向量。

卷积核（即滑动窗口）不断的在输入序列上向前滑动，直到滑动至输入字符序列的末端为止完成文本特征提取。定义 $Conv_{U,b}^{k,m}$，其中 k 为卷积核大小，m 为滑动步长，则整个特征提取过程可表示为：$Y_{1:m}=Conv_{U,b}^{k,m}(S_{1:n})$。

将卷积后得到的 m 个输出向量 $Y_{1:m}$ 进行最大池化见式（2），即每一维度取最大值，从而筛选出文本序列中最重要的特征。

$$v[j]=\max_{1\le i\le m} Y_i[j] \quad \forall j \in m \tag{2}$$

式中：$v[j]$表示卷积输出的向量 v 的第 j 维元素；$Y_i[j]$表示向量 Y_i 的第 j 维元素。

将池化层的输出通过带有 Dropout 和 ReLU 的全连接层进行输出，Dropout 通过随机减少中间特征的数量提高模型的泛化能力；ReLU 非线性激活函数减少梯度消失问题，缓解了过拟合；最后通过 Softmax 层进行归一化文本分类。

2.4 文本分类评价指标

为了评估本研究提出的文本分类的有效性，使用目前常用指标准确度（A）、精确度（P）、召回率（R）、F 值（F）和宏平均（MA）等[34]来衡量模型分类的准确性。首先针对分类体系中的任意类别 X 构建混淆矩阵（CM，Confusion Matrix），如表 1 所示。

表 1 混淆矩阵

Table 1 Confusion Matrix

预测值 / 真实值	正例	负例
正例	TP(True Positive)	FN(False Negative)
负例	FP(False Positive)	TN(True Negative)

表中：TP 表示 X 类别的样本被正确分类到 X 类；FP 表示其他类的样本被错误分类到 X 类；FN 表示 X 类样本被错误分到其他类；TN 表示其他类样本被正确分类。因此，各项评价指标可计算如下：

$$A=\frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

$$P=\frac{TP}{TP+FP} \tag{4}$$

$$R=\frac{TP}{TP+FN} \tag{5}$$

$$F=\frac{2PR}{P+R} \tag{6}$$

$$MA=\frac{1}{n}\sum_{i=1}^{n} Y_i, Y_i \in \{P_i, R_i, F_i\} \tag{7}$$

式中：n 表示待分类文本的类别个数；A 表示被正确预测的样本占总样本的比例；P 表示模型识别出的正确文本数与识别的文本总数的比率；R 表示正样本在被正确划分的样本中所占的比例；F 和 MA 属于综合性指标：F 表示精确率与召回率的调和平均数；MA 指所有类别的每一个统计指标的算数平均数。

3 方法验证

3.1 实验环境和参数设置

本研究采用国内某调水工程 2018-2019 年的 4890 条巡检安全记录数据，共 120352 个字符，经过文本预处理和去除重复字符后构造的字典共含 4762 个字符。每条数据均记录了安全信息和人工分类的标签(即安全等级)，标签分为三类，分别为：{一般，重要，紧急}，符号化表示为{0，1，2}。进行文本分类任务的训练和测试时，先将 4890 条平均分为 5 份，每份含 978 条数据，选取其中三份数据作为训练集（2934 条数据），一份为验证集，一份为测试集，所有数据均打乱处理。各个集合中不同类别样本的具体分布见图 4，以测试集为例，其含标签为一般、重大、紧急的样本数量分别为 469、447、62 条。本研究采用的巡检文本数据集示例如表 2 所示。

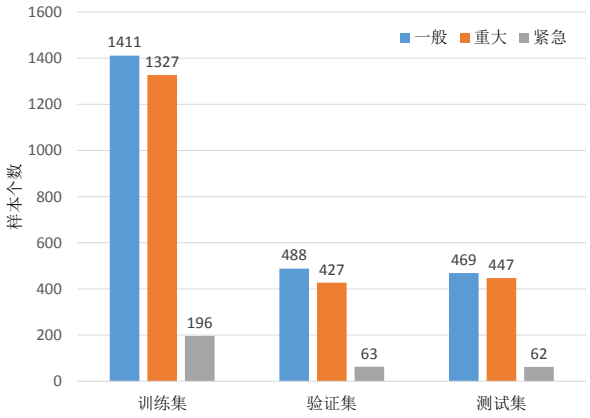

图 4 样本数量分布

Fig. 4 Sample size distribution

表 2 巡检文本数据集示例表

Table 2 Sample table of patrol text dataset

问题描述	标签
金灯寺北桥到山彪西北桥左岸截流沟内杂草过高，需要清理。	0
君子村东北公路桥下游侧左岸第二个防抛网掉漆皮、锈蚀。	1
鹤壁思德河渠道内坡出现滑塌坑洞。	2

文中实验采用中央处理器：Intel(R) Core(TM) i7-8700 CPU，主频 3.20 GHz 的硬件设备。软件环境主要包括：Windows 10 操作系统、Python 3.6 编程语言、Pytorch 开源深度学习框架和 Keras 开源深度学习库。

由于该调水工程巡检安全短文本句子长度在 9~48 个字符之间，因此训练过程中每句话处理的长度选为 50，字符长度不足 50 的句子使用<PAD>进行填补；选取三个尺寸的卷积核进行特征提取，即（2，3，4），每个尺寸的卷积核有 256 个；子向量维度为 300；每次训练样本的批大小为 128；激活函数选择 ReLU，学习率为 0.001。训练过程中每一层的输入和输出具体参数和模型结构如图 5 所示。

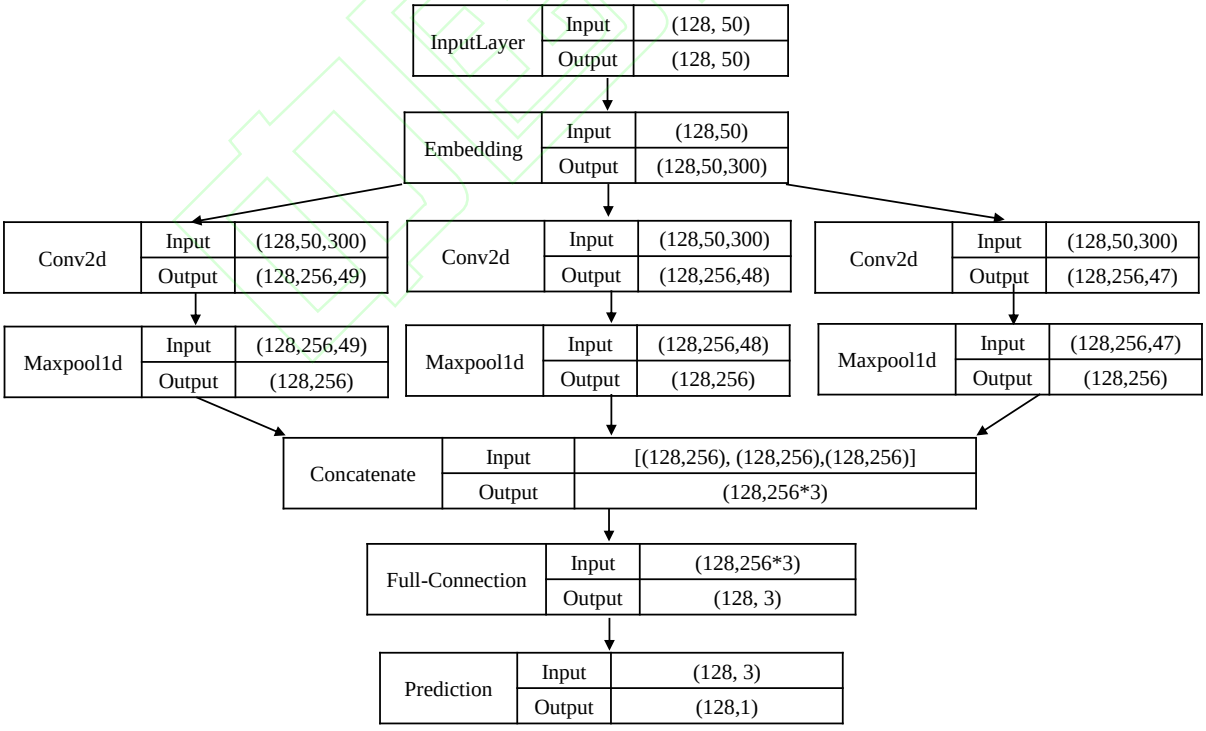

图 5 模型参数结构图

Fig. 5 Model parameter structure diagram

3.2 实验结果与分析

通过本文提出的 Char-CNN 训练巡检安全文本的数据集得到的各类分类结果如表 3 所示，从表中可以看出三种类别的 F1 值和宏平均均达到 95%以上，表明该模型的分类准确度较高，可以满足实际工程应用的需求。从混淆矩阵

$$CM=\begin{bmatrix}452 & 17 & 0\\ 13 & 434 & 0\\ 0 & 5 & 57\end{bmatrix}$$

（CM 的具体参数意义见表 4，其中对角线表示正确预测的结果）的结果来看，虽然{一般，重要，紧急}三种类别中均有少量样本被错误分类，但从实际应用的过程来考虑，巡检安全问题上报过程中会优先处理{紧急}类别下的工程问题，处理过程往往会消耗大量的人力和物力，一旦出现错报不可避免的会出现大量资源的浪费，而 Char-CNN 智能分类模型能够有效的避免{紧急}类别的误分类，{紧急}类别的分类精确度达到 100%，表明该方法具有较强的工程适用性。

表 3 Char-CNN 分类结果统计表

Table 3 Statistical table of char-CNN classification results

指标 / 类别	P	R	F1
一般	97.20%	96.38%	96.79%
重要	95.18%	97.09%	96.12%
紧急	100.00%	91.94%	95.80%
MA	97.46%	95.13%	96.24%

表 4 Char-CNN 模型混淆矩阵结果

Table 4 Char-CNN model confusion matrix results

预测值 / 实际值	一般	重要	紧急	总计	TP	FN
一般	452	17	0	469	452	17
重要	13	434	0	447	434	13
紧急	0	5	57	62	57	5
总计	465	456	57	978		
TN	452	434	57			
FP	13	22	0			

为验证模型的准确性，将其迭代 20 次计算训练数据集和验证数据集的准确率和损失值随迭代次数的变化情况，可以看出迭代达到 13 次的时候，训练集的准确率达到 0.99，损失值为 0.064，验证集的准确率达到 96%，损失值为 0.16，最终趋于稳定，表明该模型性能表现较好。对于测试数据集，通过其预测准确率和损失值随迭代次数的变化曲线（如图 6、7 所示）可得该方法迭代到第 20 次的时候达到最大准确率 96.42%，之后准确率稍有减少，最终在稳定在 96.22%。因此基于字符级的文本分类模型只需迭代 20 次就能达到很好的分类的效果，大大缩短了分类时间，提高了分类的效率。

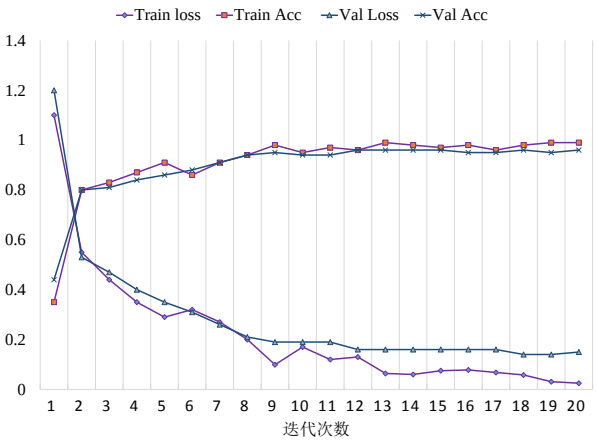

图 6 训练集和验证集的准确率和损失值随迭代次数变化

曲线

Fig. 6 Curve of accuracy and loss of training set and test set with iteration times

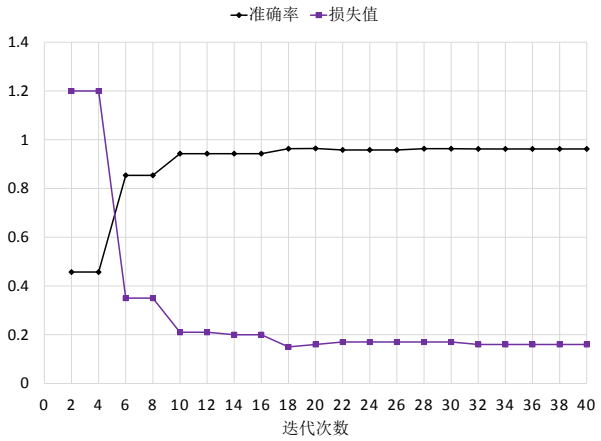

图 7 测试集的准确率和损失值随迭代次数变化曲线

Fig. 7 Curve of accuracy and loss of test set with iteration times

3.3 多方法性能对比

为验证 Char-CNN 的优越性，将本文提出的字符级向量模型嵌入目前常用的 RNN、循环卷积神经网络（RCNN）、深层金字塔卷积网络（DPCNN）、FastText、Transformer 等基于深度学习的文本分类器中，构建 Char-RNN、Char-RCNN、Char-FastText、Char-DPCNN、Char-Transformer 等分类模型，同时引入传统的基于词向量的 CNN 分类模型。通过与 Char-CNN 相同的数据集、实验环境和参数训练分类器，并将这些模型的训练结果与 Char-CNN 进行比较。采用指标为测试集的准确率（TA, Test Accuracy）和损失值（TL, Test Loss），多方法对比结果如表 5 所示。

表 5 多方法对比结果统计表

Table 5 Statistical table of comparison results of multiple methods

方法	TA	TL	CM
CNN	80.16%	0.48	$\begin{bmatrix}398 & 55 & 16\\ 66 & 369 & 12\\ 2 & 20 & 40\end{bmatrix}$
Char-CNN	96.42%	0.16	$\begin{bmatrix}452 & 17 & 0\\ 13 & 434 & 0\\ 0 & 5 & 57\end{bmatrix}$
Char-RNN	85.89%	0.45	$\begin{bmatrix}418 & 40 & 11\\ 61 & 378 & 8\\ 1 & 17 & 44\end{bmatrix}$
Char-RCNN	94.38%	0.22	$\begin{bmatrix}428 & 39 & 2\\ 9 & 437 & 1\\ 0 & 4 & 58\end{bmatrix}$
Char-DPCNN	93.97%	0.26	$\begin{bmatrix}449 & 20 & 0\\ 27 & 418 & 2\\ 4 & 6 & 52\end{bmatrix}$
Char-FastText	90.39%	0.31	$\begin{bmatrix}433 & 35 & 1\\ 33 & 412 & 2\\ 4 & 19 & 39\end{bmatrix}$
Char-Transformer	85.28%	0.42	$\begin{bmatrix}405 & 58 & 6\\ 51 & 385 & 11\\ 0 & 18 & 44\end{bmatrix}$

与传统的 CNN 相比，Char-CNN 在分类准确率方面提升了约 16%。其它字符级深度学习分类方法与基于词级别的 CNN 相比，在准确率方面嵌入了字符层面文本向量的深度学习方法普遍优于 CNN，大部分准确率均能达到 90%以上，表明引入字符级的分类方法在巡检安全文本这类数据集上具有其独到的优势。Char-CNN 与其它字符级深度学习分类方法相比，其准确率均优于其他分类方法，可以看出，不论是结合了 CNN 和 RNN 的 Char-RCNN 分类方法，还是深层金字塔卷积网络的 Char- DPCNN 方法，其准确性的均低于 Char-CNN，说明采用 CNN 分类器在运行准确率上是最优选择。此外，从多方法各自的混淆矩阵对比看出，其他类型的方法均会出现错分为{紧急}类别的情况，在工程实际应用中会造成人力物力资源的浪费，因此从工程应用角度证明了 Char-CNN 的可行性。综上，针对调水工程巡检安全数据集 Char-CNN 能显著提升文本分类的准确率，且具有较强的工程适用性。

4 结论

本文通过将预训练字符级向量嵌入卷积神经网络输入层，实现了调水工程巡检安全记录文本的智能分类，得到以下结论：

（1）本文提出的 Char-CNN 在应用时无需进行分词，不依赖专业词库，基于字符层面的嵌入向量不受巡检安全文本中存在的口语化表达、错别字以及未登录词等的影响，因此具有一定的普适性。

（2）Char-CNN 与传统基于词向量的 CNN 相比分类准确率提升了约 16%，同时其它融合字符级的深度学习方法的分类准确率均优于 CNN，因此本文提出的字符级文本分类方法在巡检安全数据集上表现优异。

（3）Char-CNN 在分类准确率方面均优于其它基于字符级的深度学习神经网络，且无错分为{紧急}类别的文本，因此采用卷积神经网络结构的分类方法能显著提高调水工程巡检安全文本分类的准确性，具有较强的工程适用性。

然而，根据巡检安全文本进行字符嵌入向量预训练的字符仅仅包括了 4762 个字符，尚未涵盖所有汉字，后续可以通过增加语料库来对字符向量进行补足。此外，文本所提的模型目前仅在巡检安全类的短文本上效果较好，后续将进一步研究针对事故案例库等长文本的分类模型。

参考文献（References）

[1] 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(5):1-8.

Yu You, Fu Yu, Wu Xiaoping. Summary of text classification methods[J]. Journal of Network and Information Security, 2019, 5(5):1-8. (in Chinese)

[2] 陈平，匡尧，胡景懿，等. 增强领域特征的电力审计文本分类方法[J].计算机应用，2020，40(S1)：109-112.

Chen Ping, Kuang Yao, Hu Jingyi, et al. Text categorization method with enhanced domain features in power audit field[J]. Computer Application, 2020, 40(S1): 109-112. (in Chinese)

[3] 廖胜兰，殷实，陈小平等. 面向电力业务对话系统的意图识别数据集[J]. 计算机应用，2020，40(9):77-82.

Liao Shenglan, Yin Shi, Chen Xiaoping, et al. Intent recognition dataset for dialogue systems in power business[J]. Computer Application, 2020, 40(9):77-82. (in Chinese)

[4] 许诺，唐锡晋.基于百度热搜新闻词的社会风险事件 5W 提取研究[J]. 系统工程理论与实践，2020，40(2):334-342.

Xu Nuo, Tang Xijin. Research on 5W extraction of social risk events based on Baidu hot search news words[J]. Theory and Practice of System Engineering, 2020, 40(2):334-342. (in Chinese)

[5] 胡万亭，贾真. 基于加权词向量和卷积神经网络的新闻文本分类[J]. 计算机系统应用，2020，29(5):275-279.

Hu Wanting, Jia Zhen. News text classification based on weighted word vector and convolution neural network [J]. Application of Computer System, 2020, 29(5):275-279. (in Chinese)

[6] 於张闲，胡孔法. 基于 BERT-Att-BiLSTM 模型的医学信息分类研究[J]. 计算机时代，2020,(3):1-4.

Yu Zhang Xian, Hu Kong method. Classification of medical information based on the BERT-Att-BiLSTM model [J]. Computer Age, 2020,(3):1-4. (in Chinese)

[7] 李强，李瑶坤，夏书月,等. 一种改进的医疗文本分类模型:LS-GRU[J]. 东北大学学报:自然科学版，2020(7):938-942.

Li Qiang, Li Yaokun, Xia Shuyue, et al. An improved medical text classification model: ls-gru [J]. Journal of Northeast University: Natural Science Edition, 2020 (7): 938-942. (in Chinese)

[8] 罗明，黄海量. 一种基于语义标注特征的金融文本分类方法[J].计算机应用研究，2018，35(8):2281-2284+2288.

Luo Ming, Huang Hailiang. A financial text classification method based on semantic annotation features [J]. Computer Application Research, 2018, 35(8):2281-2284+2288. (in Chinese)

[9] 陈鹏，郭小燕. 基于LSTM-Attention的农业短文本信息分类研究[J]. 软件导刊，2020，19(9):21-26.

Chen Peng, Guo Xiaoyan. Classification of agricultural short text information based on LSTM attention [J]. Software Guide, 2020, 19 (9): 21-26. (in Chinese)

[10] 朱远平，戴汝为. 基于 SVM 决策树的文本分类器[J].模式识别与人工智能，2005，18(4):412-416.

Zhu Yuanping, Dai Ruwei. Text classifier based on SVM decision tree [J]. Pattern Recognition and Artificial Intelligence, 2005, 18 (4): 412-416. (in Chinese)

[11] 车君华，冯毅雄，谭建荣，等.基于决策支持向量机的产品设计知识文档分类研究[J].计算机集成制造系统，2007(5):891-897.

Che Junhua, Feng Yixiong, Tan Jianrong, et al. Classification of product design knowledge documents based on decision support vector machine [J]. Computer

Integrated Manufacturing System, 2007(5): 891-897. (in Chinese)

[12] 黄文明, 孙艳秋. 基于最大熵的中文短文本情感分析[J].计算机工程与设计, 2017, 38(1):138-143.

Huang Wenming, Sun Yanqiu. Sentiment analysis of Chinese short text based on maximum entropy [J]. Computer Engineering and Design, 2017, 38 (1): 138-143. (in Chinese)

[13] 余鹰, 苗夺谦, 刘财辉, 等.基于变精度粗糙集的KNN分类改进算法[J].模式识别与人工智能, 2012, 25(4): 617-623.

Yu Ying, Miao Duoqian, Liu Caihui, et al. Improved KNN classification algorithm based on variable precision rough set [J]. Pattern Recognition and Artificial Intelligence, 2012, 25 (4): 617-623. (in Chinese)

[14] 庞秀丽, 冯玉强, 姜维.贝叶斯文本分类中特征词缺失的补偿策略[J].哈尔滨工业大学学报, 2008, 40(6):956-960.

Pang Xiuli, Feng Yuqiang, Jiang Wei. Compensation strategies for missing feature words in Bayesian text categorization [J]. Journal of Harbin Institute of Technology, 2008, 40 (6): 956-960. (in Chinese)

[15] 邓丁朋, 周亚建, 池俊辉,等. 短文本分类技术研究综述[J].软件, 2020, 478(2):149-152.

Deng Dingpeng, Zhou Yajian, Chi Junhui, et al. Summary of short text classification technology [J]. Software, 2020, 478 (2): 149-152. (in Chinese)

[16] Kim Y. Convolutional neural networks for sentence classification[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP 2014). 2014: 1746-1751.

[17] Liu P, Qiu X, Huang X. Recurrent Neural Network for Text Classification with Multi-Task Learning[J]. 2016.

[18] 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.

[19] 张超超, 卢新明. 基于 FastText 的新闻文本多分类研究[J].软件导刊, 2020, 019(3):44-47.

Zhang Chaochao, Lu Xinming. Research on multi classification of news text based on fast text [J]. Software Guide, 2020, 019 (3): 44-47. (in Chinese)

[20] Chan K H, Im S K, Ian V K, et al. Enhancement Spatial Transformer Networks for Text Classification[C]// ICGSP 2020: 2020 The 4th International Conference on Graphics and Signal Processing. 2020.

[21] 何力, 谭霜, 项凤涛, 等.基于深度学习的文本分类技术研究进展[J/OL].计算机工程:1-15.

He Li, Tan Shuang, Xiang Fengtao, et al. Research progress of text classification technology based on deep learning [J / OL]. Computer Engineering: 1-15. (in Chinese)

[22] 汪嘉伟, 杨煦晨, 琚生根, 等.基于卷积神经网络和自注意力机制的文本分类模型[J]. 四川大学学报(自然科学版), 2020, 57(3):469-475.

Wang Jiawei, Yang Xuchen, Ju Shenggeng, et al. Text classification model based on convolutional neural network and self attention mechanism [J]. Journal of Sichuan University (NATURAL SCIENCE EDITION), 2020, 57 (3): 469-475. (in Chinese)

[23] Du J, Vong C M, Chen C L P. Novel Efficient RNN and LSTM-Like Architectures: Recurrent and Gated Broad Learning Systems and Their Applications for Text Classification[J]. IEEE Transactions on Cybernetics, 2020, PP(99):1-12.

[24] Li X, Ning H. Deep Pyramid Convolutional Neural Network Integrated with Self-attention Mechanism and Highway Network for Text Classification[J]. Journal of Physics: Conference Series, 2020, 1642(1):012-008.

[25] Ling-Ling D, Kan J. Chinese Text Classification Based on FastText[J]. Computer and Modernization, 2018.

[26] 唐庄,王志舒,周爱,冯美姗,屈雯,鲁明羽.面向文本分类的transformer-capsule集成模型[J/OL].计算机工程与应用:1-7.

Tang Zhuang, Wang Zhishu, Zhou Ai, Feng Meishan, Qu Wen, Lu Mingyu. Transformer capsule integrated model for text classification [J / OL]. Computer Engineering and Application: 1-7. (in Chinese)

[27] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C].Conference on Empirical Methods in Natural Language Processing, 2014:1532-1543.

[28] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [C].1st International Conference on Learning Representations. Scottsdale, AZ, United states, 2013.

[29] Vodrahalli K, Bhowmik A K. 3D computer vision based on machine learning with deep neural networks: A review[J]. Journal of the Society for Information Display, 2017, 25(10-12):676-694.

[30] Li X, Wu W, Su H. Convolutional Neural Networks Based Multi-task Deep Learning for Movie Review Classification[C]. 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2017.

[31] 任秋兵, 李明超, 沈扬, 等.水工混凝土裂缝像素级形态分割与特征量化方法研究[J/OL].水力发电学报:1-15.

Ren Qiubing, Li Mingchao, Shen Yang, et al. Research on pixel level morphological segmentation and feature quantification of hydraulic concrete cracks [J / OL]. Journal of Hydropower: 1-15. (in Chinese)

[32] 王超, 贾贺, 张社荣, 等.基于图像的混凝土表面裂缝量化高效识别方法[J/OL].水力发电学报:1-11.

Wang Chao, Jia He, Zhang Sherong, et al. Quantitative and efficient identification method of concrete surface cracks based on image [J / OL]. Journal of Hydropower Generation: 1-11. (in Chinese)

[33] 陈波, 张华, 汪双, 等.基于全卷积神经网络的坝面裂纹检测方法研究[J].水力发电学报,2020,39(7):52-60.

Chen Bo, Zhang Hua, Wang Shuang, et al. Study on dam surface crack detection method based on full convolution neural network [J]. Journal of Hydropower Generation, 2020,39 (7): 52-60. (in Chinese)

[34] 宋胜利, 王少龙, 陈平. 面向文本分类的中文文本语义表示方法

[J]. 西安电子科技大学学报, 2013, 40(2):89-97+129.

Song Shengli, Wang Shaolong, Chen Ping. Chinese text semantic representation for text classification [J]. Journal of Xi'an University of Electronic Science and Technology, 2013, 40 (2): 89-97 + 129. (in Chinese)