DOI: 10.13382/j.jemi.B2205608

基于字词融合的高铁道岔多级故障诊断组合模型*

林海香[1] 赵正祥[1] 陆人杰[2] 卢 冉[1] 白万胜[1] 胡娜娜[1]
(1.兰州交通大学自动化与电气工程学院 兰州 730070;2.卡斯柯信号有限公司 上海 200071)

摘 要:为有效提升高速铁路道岔维护效率和故障定位准确率,面向其故障文本数据,提出了一种基于字词融合的高速铁路道岔多级故障诊断组合模型。首先,建立高速铁路道岔专业词库,将文本表示为字向量与词向量并进行深度融合。其次,考虑到故障文本存在类别不均衡问题,采用 Borderline-SMOTE 算法对不均衡文本数据进行处理,优化故障文本数据分布。接着使用 BiLSTM(Bi-directional long short-term memory)-CNN(convolutional neural network)的组合神经网络提取故障文本深度特征,最后通过分类器实现智能故障诊断。采用我国高速铁路道岔故障文本数据进行模型性能验证,结果显示所提模型的一级故障诊断准确率达到 95.62%,二级故障诊断准确率达到 93.81%,证明多级故障诊断精度可达到理想效果。

关键词:高速铁路道岔;多级故障诊断;字词融合;Borderline-SMOTE;组合神经网络
中图分类号:U216.42 **文献标识码**:A **国家标准学科分类代码**:520.20

Combined model for multi-level fault diagnosis of high-speed rail turnouts based on character and word fusion

Lin Haixiang[1] Zhao Zhengxiang[1] Lu Renjie[2] Lu Ran[1] Bai Wansheng[1] Hu Nana[1]
(1. School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. CASCO Signal Ltd, Shanghai 200071, China)

Abstract: To effectively improve the maintenance efficiency and fault location accuracy of high-speed railway turnouts, a combined model for multi-level fault diagnosis of high-speed rail turnouts based on character and word fusion was proposed. Firstly, a professional thesaurus of high-speed rail turnout equipment was established, and fault texts were represented as character vectors and word vectors and the character vectors and word vectors were deeply fused. Secondly, considering the problem of imbalanced categories in fault texts, the Borderline-SMOTE algorithm was used to process the imbalanced text data to optimize the fault text data distribution. Then, a combination of Bi-directional long short-term memory (BiLSTM) and convolutional neural network (CNN) was used to extract deep features of the fault text. Finally, an intelligent diagnosis of faults was achieved by means of a classifier. The model performance was validated using fault text data of China high-speed railway turnout faults. The test results show that the accuracy of the proposed model reaches 95.62% for the primary fault diagnosis and 93.81% for the secondary fault diagnosis, which proves that the multi-level fault diagnosis accuracy can reach the desired effect.

Keywords: high-speed railway turnout equipment; multi-level intelligent diagnosis; character and word fusion; Borderline-SMOTE; combined neural network

0 引 言

道岔作为高速铁路信号地面设备的重要组成部分,对于行车安全有重要影响,是高速铁路设备维护的重点[1-3]。高速铁路道岔具有高速度、高安全性、高平稳性以及高精度等性能,设备复杂,技术性能高,维修难度大。而现阶段高速铁路道岔故障主要依靠现场维修人员进行诊断和处理,效率低下且受维修人员知识和技能的限制,易造成故障延时,延误安全运营。故利用大量以中文文

收稿日期:2022-06-21 Received Date: . Received Date:2022-06-21
*基金项目:国家自然科学基金(61763023)、甘肃省科技计划项目(20YF8GA037)、甘肃省高等学校创新基金(2020B-104)项目资助

本记录高速铁路道岔故障数据,提取知识经验,驱动智能故障诊断,可辅助指导道岔维修,缩短故障延时。

采用文本挖掘的高铁道岔故障诊断存在文本特征表示、数据分布与分类问题[4-6]。如文献[4]铁路信号设备故障文本多为短文本,因此易造成特征表示稀疏,缺乏数据特异性的问题;文献[5]列控车载设备故障文本存在不均衡问题,导致少数类故障样本易被误分类;文献[6]铁路事故文本长短不一,差异性大,特征提取难度高。根据上述研究,基于故障文本数据驱动的高速铁路道岔多级故障诊断方法存在3个挑战亟待解决:故障文本特征表示稀疏、数据分布不均衡以及故障文本分类模型效果欠佳。

文本特征表示是高速铁路道岔故障诊断的关键。文献[7]采用主题模型对长文本进行文本特征表示并降维;文献[8]使用TF-IDF模型对短文本特征表示并转换为向量。随着词向量生成工具Word2vec的提出[9],文献[10]通过Word2Vec将车载设备故障文本转化为词向量,作为LSTM-BP级联网络模型的输入实现车载设备故障诊断;文献[11]通过Word2Vec生成故障文本词向量后作为卷积神经网络模型输入,实现信号设备故障诊断;文献[12]利用字向量从根本上解决了一词多义的问题,提高分类的准确度;文献[13]在词向量中加入词的字符级特征加强文本特征表示。从文献[10-13]的研究表明字向量和词向量都可表示文本特征信息,但对含义模糊易造成歧义的语句,仍无法精准全面的获取文本特征,模型准确率较低。

数据分布优化主要从数据和算法两个层面考虑。在数据层面,文献[14]针对铁路信号设备不均衡性,采用SMOTE算法自动生成少数类样本以解决数据不均衡问题;文献[15]采用G-SMOTE算法通过在几何空间内构建一个超球体为少数类样本生成人工数据。在算法层面,文献[16]利用DA方法选择不均衡文本特征改进文本特征分布的不均衡性;文献[17]采用结合代价敏感学习的随机森林算法改善了列控车载设备故障文本分布。但以上方法均未考虑样本的分布,易造成类间样本及特征的重复,导致分类结果不准确。

对于故障文本分类模型的构建,BiLSTM和CNN是在文本分类中常用的两类模型[18-22]。CNN可以有效挖掘文本的局部特征,但对长时间输入序列,单一的CNN会造成前后相关性特征的丢失,BiLSTM能提取前向与后向的语义特征信息,但不能突出局部重要信息。

综上所述,本文提出一种基于字词融合的高速铁路道岔多级故障诊断组合模型。该模型提出字词融合方法挖掘出高铁道岔故障文本更全面准确的特征并实现向量转化,可有效区分不同语境下词的语义信息,相对于单一词向量,字词融合能达到消歧的作用,并且加入字向量和位置信息能够有效避免边界词的错误划分,以此优化故障文本特征表示;再采用B-SMOTE (borderline-SMOTE)[23]算法在边界处生成少数类样本,优化故障文本数据分布;最后将BiLSTM和CNN两种深度学习算法集成,进而设计出组合神经网络模型,来提取故障文本的上下文特征和深层次特征,以实现精确诊断。

1　数据来源及数据特征

通过调研,收集我国高速铁路道岔故障文本共4 138条。按照道岔部件功能进行初步分类,将高速铁路道岔故障划分为一级故障和二级故障,一级故障指道岔中某个设备故障,共5类;二级故障则是在一级故障基础上做深度划分,直接定位到道岔的具体部件,共23类,如表1所示,其中结合部器件指铁路电务部门和工务部门有交集的道岔部件。

表1　高速铁路道岔故障分类

Table 1　High-speed rail turnout equipment fault classification

	一级故障	二级故障
道岔故障分类	道岔控制电路器材(A)	电线电缆(A1)
		变压器(A2)
		断路器(A3)
		断相保护器(A4)
		二极管(A5)
		继电器(A6)
		整流匣(A7)
	结合部器件(B)	工务病害(B1)
		滑床板(B2)
	密贴检查器(C)	弹簧(C1)
		接点组(C2)
		拐轴(C3)
		异物卡阻(C4)
	外锁闭及安装装置(D)	杆件调整(D1)
		表示杆(D2)
		密贴调整(D3)
		锁闭框(D4)
		外锁闭卡阻(D5)
		自然灾害(D6)
	转辙机(E)	液压器件(E1)
		接点组(E2)
		摩擦联结器(E3)
		电机(E4)

通过高铁道岔故障文本数据的分类统计,如图1所示,可以看出一级故障与二级故障样本均存在分布不均衡的问题,即某一文本类别中的大比例故障掩盖小比例故障,会影响实际故障定位准确性,造成诊断结果与实际问题存在较大偏差。如图1所示,各一级故障文本数目比例为9∶6∶3∶25∶7,存在一级文本分布不均衡;再

如图 1,二级故障转辙机接点组故障与摩擦联结器故障的不均衡系数高达 16.7。这种数据分布不均衡现象会导致少数类样本诊断结果倾向于多数类样本,是故障诊断研究过程中不可忽视的问题。

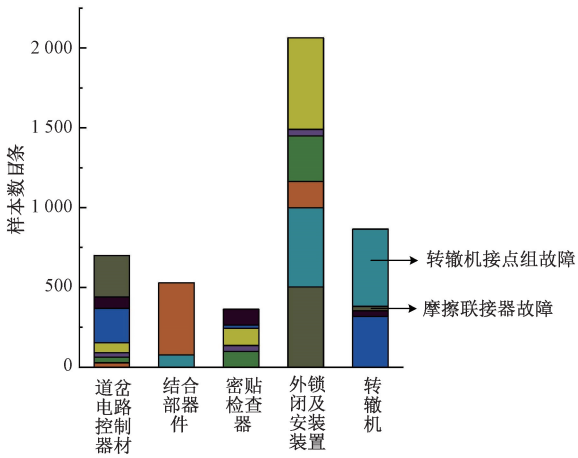

图 1　原始文本数据分布情况

Fig. 1　Distribution of original text data

2　高速铁路道岔多级故障诊断模型

本文设计的高速铁路道岔多级故障诊断模型框架如图 2 所示,主要包括如下 4 部分:高速铁路道岔故障文本数据预处理、字词融合、文本数据分布优化和基于 BiLSTM-CNN 的组合神经网络特征提取及故障文本自动分类。

2.1　故障文本预处理

1)文本分词处理

本文采用 jieba 分词工具对高速铁路道岔故障文本进行分词处理。由于高速铁路道岔故障文本中含有大量专业领域词汇,具有极强的专业性,例如“道岔锁闭器”,通用的分词方法直接将其分成“道岔”、“锁闭”、“器”,但在铁路领域该词项需作为一个整体使用。因此,需构建专有的高速铁路道岔故障词库,并使用百度停用词库去除文本中无用的符号和停用词,使分词达到更好的效果。部分高速铁路道岔故障专业词库词项如图 3 所示。

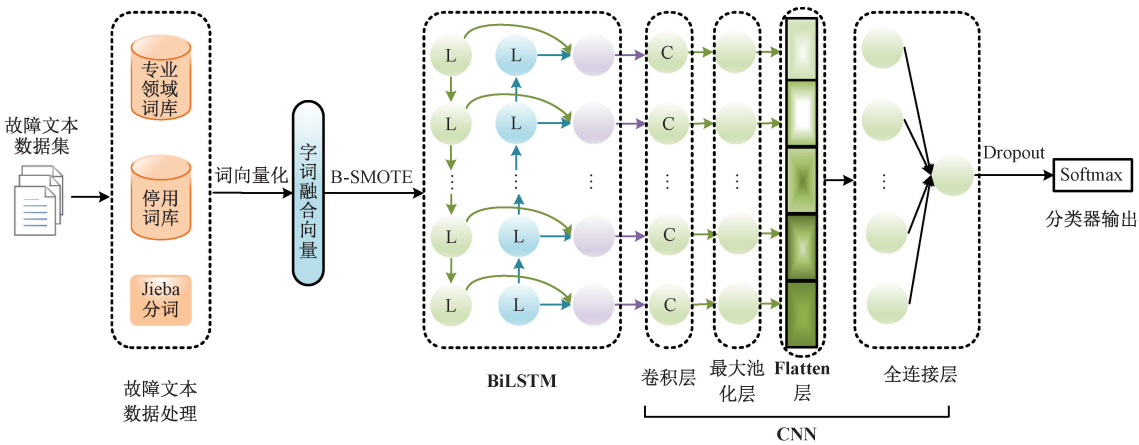

图 2　高速铁路道岔多级故障诊断模型框架

Fig. 2　Multi-level Fault diagnosis model framework for high-speed railway turnout equipment

2)文本语句填充

在进行文本处理时需统一句子的长度,即语句填充。大部分文本的处理通常会为了不舍弃任何文本信息,而选择最大句子长度。由于高速铁路道岔故障文本的句子长度差异较大,若按最长句子进行填充,不仅大幅度增加模型的向量维度和训练时间,另一方面也会引入大量噪音。本文则对文本长度进行统计,选择句子平均长度的 1.2 倍作为句子最大长度 L_{max},当句子长度大于 L_{max},多余语句被截断,长度不够的句子自动补充 0 以达到 L_{max}。如此既尽可能保留文本信息,又不会大幅增加向量维度和训练时间。

2.2　基于字词融合的特征向量表示

故障文本是以中文文本记录的,由于中文文本是靠意义连接的语言形式,包含大量的语义信息、上下文依赖信息和语序信息,故需将故障文本信息转换为多维且连续的向量形式。本文利用 jieba 分词获取位置信息,并使用 Word2vec 对高速铁路道岔故障文本进行向量化,获得字向量表示和词向量表示。词向量语义丰富,但因为中文分词限制,分词会出现错误进而影响分类效果。字向量能够表达的语义信息有限,但利用字向量及其位置信息能避免对文本中特征词的错误划分,故将字向量、词向量和位置特征融合。

将文本中的句子分别转化为 $\boldsymbol{S}_1 = [\boldsymbol{s}_1^1, \boldsymbol{s}_1^2, \cdots, \boldsymbol{s}_1^i]$,

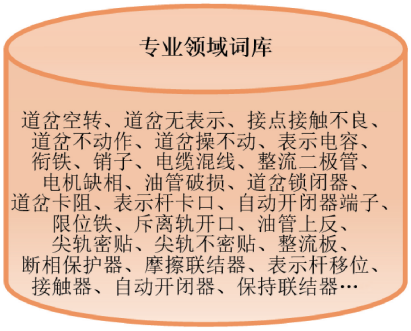

图3　高速铁路道岔故障专业词库

Fig. 3　Professional thesaurus of high-speed rail turnout equipment failure

$\boldsymbol{S}_2=[\boldsymbol{s}_2^1,\boldsymbol{s}_2^2,\cdots,\boldsymbol{s}_2^j]$，$\boldsymbol{S}_1$、$\boldsymbol{S}_2$ 分别为词向量矩阵和字向量矩阵，其中 $\boldsymbol{s}_1^i$ 是分词后第 i 个词的词向量，$\boldsymbol{s}_2^j$ 是第 j 个字的字向量。为保证字向量和词向量可融合，需将每个词的词向量重复 n 次，n 表示该词语中字的个数，如此可得与字向量对齐后的词向量序列，然后令词向量序列经过一个变换矩阵 $\boldsymbol{E}$ 使其得到和字向量相同的维度，二者融合得出字词融合结果：

$$q_i=(s_1^k\times\boldsymbol{E})\oplus s_2^k \tag{1}$$

接着再将字词融合结果 q_i 与位置信息相加：

$$f_i=q_i+\sum_m^n x_2^i \tag{2}$$

式中：(m,n) 为第 i 个词首尾字符对应的位置。最终得到 $F=(f_1,f_2,f_3,\cdots,f_n)$，构成融合后的特征矩阵。

融合过程如图4所示。

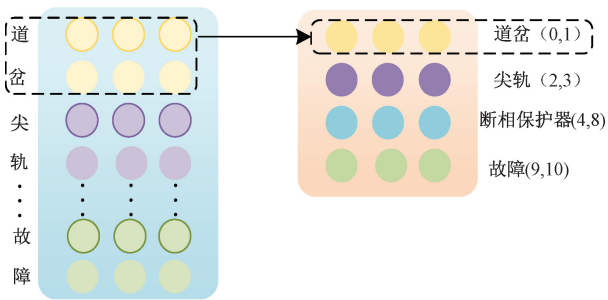

图4　向量融合过程

Fig. 4　Process of vector fusion

图4以“道岔/尖轨/断相保护器/故障”为例，分词后标记每个词的首字符和尾字符在语句中的位置，例如“尖轨(2,3)”中首字符“尖”为句中的第2个字符，“轨”为句中的第3个字符。在融合过程中每个词对应的字向量结合词向量本身得到融合向量，并使字词融合结果与位置信息结合，可以最大程度上避免边界切分错误造成歧义的问题。

2.3　文本数据分布优化

由图1可知，高速铁路道岔故障文本存在数据分布不均衡问题，增加了文本分类的难度。本文使用B-SMOTE算法作为SMOTE数据生成机制的增强，对交界处的少数类样本生成新样本，对样本分布进行优化。该算法将少数类样本分为安全类(A)、危险类(B)和噪音类(C)，但其仅对危险类，即近邻样本1/2以上为多数类样本的少数类样本进行过采样。合成样本的原理如图5所示。

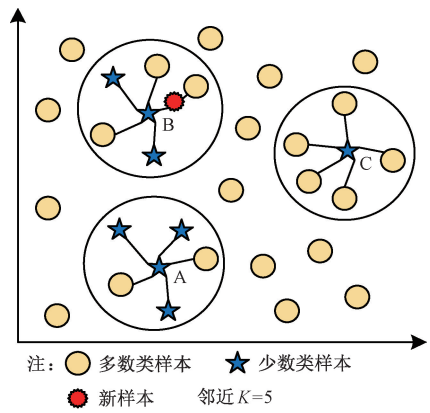

图5　基于B-SMOTE的少数类样本合成

Fig. 5　Minority sample synthesis based on B-SMOTE

B-SMOTE算法用到一些参数：S 为样本集，分别设置了邻近的多数类样本集和少数类样本集 S_{maj}，S_{min}，x_n 为近邻样本，x_i、x_{ij} 分别为该样本的全部属性和邻近样本的全部属性，近邻数 K，R_{ij} 一般取0.5或1。B-SMOTE算法的具体步骤如下描述：

步骤1)：对每一个少数类样本 x_i，确定与其最邻近的样本集 $S'\in S$。

步骤2)：对每一个少数类样本 x_i，判断最邻近属于多数样本集的个数，记为 $|S'\cap S_{\text{maj}}|$。

步骤3)：选择出符合危险类条件的 x_i，判断条件为：

$$\frac{K}{2}<|S'\cap S_{\text{maj}}|<K。$$

步骤4)：从对危险类合成少数类样本，计算 x_i 与 x_n 对应属性 j 中的差值 $d_{ij}=x_i-x_{ij}$。得出合成新的少数类样本 $h_{ij}=x_i+d_{ij}\times rand(0,R_{ij})$。

2.4　故障文本分类模型优化

1)基于BiLSTM的特征提取

本文将前向LSTM和后向LSTM组成BiLSTM模型来提取故障文本上下文信息，LSTM网络单元结构如图6所示。

LSTM网络[24]中的遗忘门机制、输入门机制和内部记忆单元分别如式(3)~(6)所示，输出门机制如式(7)和(8)所示。

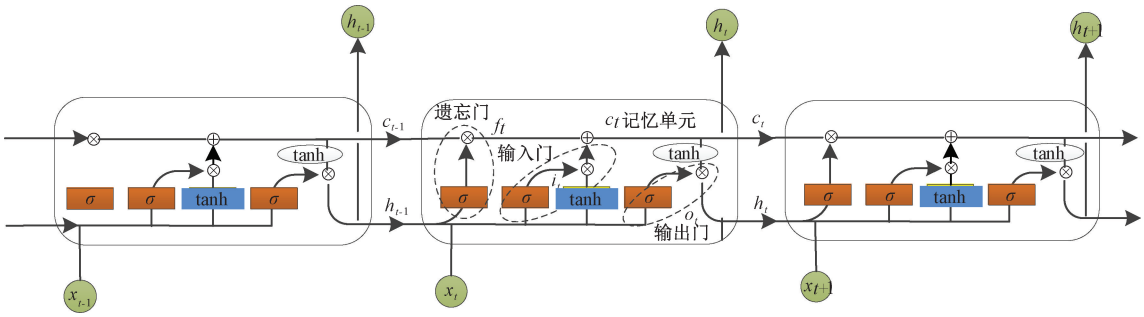

图 6　LSTM 网络单元结构

Fig. 6　LSTM network cell structure

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t + b_i) \tag{4}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b) \tag{7}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{8}$$

在上述公式中，x_t 表示当前输入信息，W、b 为权重系数矩阵和偏置项，h_{t-1} 为上一序列输出，f_t 表示遗忘门对当前细胞状态的遗忘程度，i_t 代表输入门对新特征的更新程度，$\tilde{C}_t$ 为新状态信息的候选向量，C_{t-1} 用来存储前一时刻记忆信息，C_t 为当前细胞状态，o_t 表示输出门的信息输出。

将处理后的故障文本数据输入至 BiLSTM 模型中，经过前向 LSTM 层输出 $\overrightarrow{h}$，经过后向 LSTM 层输出 $\overleftarrow{h}$。两层 LSTM 输出进行拼接后生成 BiLSTM 网络输出 $\overleftrightarrow{h}_t = [\overrightarrow{h}, \overleftarrow{h}]$。生成的特征信息通过全连接层输入至 CNN 进行下一步深度提取。

2）基于 CNN 的特征提取

不同于传统的 CNN 网络输入，本文的 CNN 对 BiLSTM 网络输出的 $\boldsymbol{h} = [\overleftrightarrow{h}_1, \overleftrightarrow{h}_2, \cdots, \overleftrightarrow{h}_t]$ 实施进一步的深度特征提取，步骤如下：

（1）卷积层。对 n 个字符的融合向量矩阵提取特征，所得的卷积结果如式（9）所示。

$$O_i = f(\boldsymbol{C} \cdot h_{i:i+n-1} + b) \tag{9}$$

式中：$\boldsymbol{C}$ 为 $n \times k$ 的实数矩阵，$h_{i:i+n-1}$ 为由第 i 个词到第 $i+n-1$ 个词组成的连续文本段，b 为偏置项，$f(\cdot)$ 为 ReLU 激活函数，有效解决梯度爆炸及消失问题。对一组故障文本进行完卷积操作后，会得到如下特征集合：

$$V = f(W \cdot O' + b) \tag{10}$$

（2）最大池化层。最大池化提取出文本最重要特征，表达式如下：

$$O' = \max[O] \tag{11}$$

（3）全连接层。在全连接层和最大池化间加入 Flatten 层把多维矩阵一维化输出，全连接层进而对输出进行整合，得到结果如式（12）所示。

$$V = f(W \cdot O' + b) \tag{12}$$

式中：W 表示权重项，b 代表偏置因子。

此外，为防止训练过程中过度依赖局部特征，避免过拟合的问题，在全连接层后嵌入 Dropout，使神经元激活数值按照概率 P 随机丢失。最后使用 Softmax 函数分析故障类别。

3）Softmax 故障文本分类器构建

本文模型引入 Softmax 函数[25]接收来自全连接层的特征数据 V，作为本次高速铁路道岔故障文本分类的输出，结果如式（13）所示。

$$Y = \text{Softmax}(W \cdot V + b) \tag{13}$$

模型最终的目标函数为：

$$loss = -\sum_i y_i' \log(y_i) \tag{14}$$

式中：y_i 为模型对故障文本的预测类别，y_i' 为故障样本的真实类别，模型采用 Adam 算法来优化此目标函数。

3　实验验证和结果分析

3.1　实验环境

本文实验环境及配置具体如表 2 所示。

表 2　实验环境

Table 2　Experimental environment

实验环境	环境配置
操作系统	Ubuntu 18.04.6 LTS
CPU	Inter(R) Core(TM) i9-12900K CPU
CUDA	版本号：11.2.162
内存	12 GB
编程语言	Python3.7
分词工具	jieba
深度学习框架	TensorFlow-GPU(版本号：1.14.0)

3.2　实验数据及评价指标

为证明模型的有效性，通过真实数据集进行实验验证。将我国近几年高速铁路道岔故障文本信息作为实验数据，将其按7∶3划分为训练集和测试集，并采用五折交叉验证法[26]来训练模型。本文采用准确率、精确率、召回率和F_1值作为评价指标[27]。

3.3　B-SMOTE生成少数类样本实验

使用B-SMOTE对少数类样本进行自动生成的结果如图7所示，各一级故障、二级故障数据的不均衡性大幅下降。与图1对比，转辙机摩擦联结器故障和转辙机接点组故障的不均衡系数由16.7降至1.3，极大优化了高速铁路道岔故障文本数据分布不均衡性。

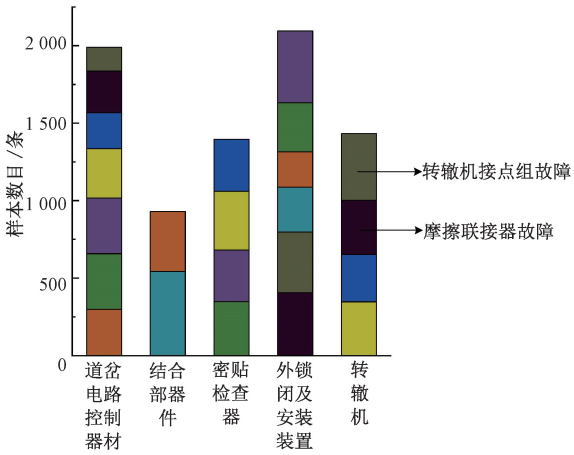

图7　B-SMOTE生成后样本分布情况

Fig. 7　Sample distribution after B-SMOTE is generated

3.4　相关超参数设置对模型性能影响

模型参数设置是否恰当直接影响模型性能，本文通过对比实验来确定模型主要超参数。首先，Dropout作为增强模型泛化能力的技术手段，选择合适的Dropout值不仅可以改善过拟合问题并且有效提高模型的训练效率。图8为各评价指标随Dropout值的变化情况，从图中可以看出，当Dropout值为0.5时，模型效果最佳。

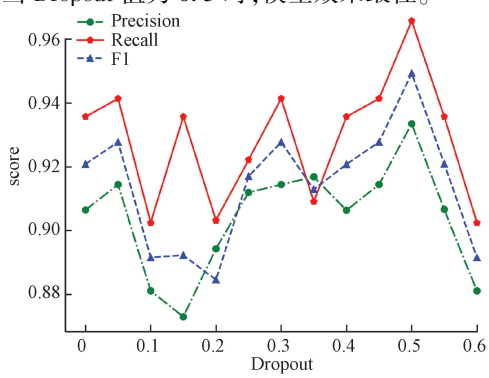

图8　评价指标随Dropout值变化情况

Fig. 8　Change of evaluation index with Dropout value

其次，迭代轮数作为本模型的关键超参数之一，若其设置过少会导致欠拟合问题；过多则会增长模型训练时间，并且导致模型泛化性下降。图9和10显示epoch在0~30次时性能指标的变化情况，从图中可看出当epoch到25次时，一级故障和二级故障的准确率与损失函数值趋于平稳且达到理想效果，因此本文选取25轮次作为最终的迭代轮次。

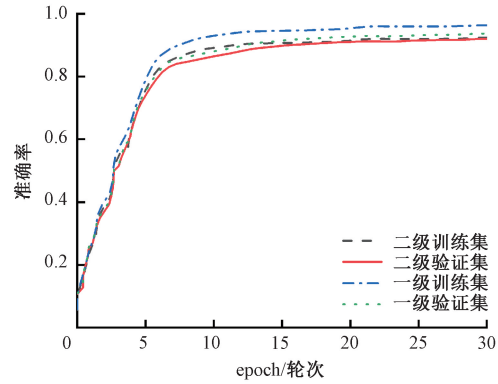

图9　准确率随epoch增加变化曲线

Fig. 9　Accuracy change curve with epoch increase

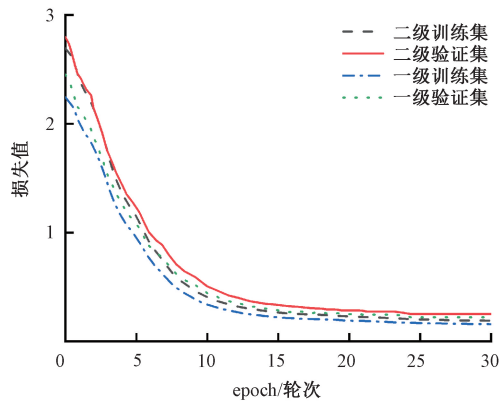

图10　损失值随epoch增加变化曲线

Fig. 10　Loss change curve with epoch increase

此外，高速道岔故障文本的长度统计如图11所示，句子平均长度L_{max}为60，句子填充(pad_size)设置为72。

根据实验的结果最终确定高速铁路道岔多级故障诊断模型的超参数如表3所示。

3.5　对比实验

将本文模型与文献[20]的CNN、文献[21]的LSTM以及文献[22]的BiLSTM等经典故障诊断模型在表2所示环境下进行对比实验，各模型均采用B-SMOTE算法处理不均衡数据，其中对比模型的关键参数如表4所示，不同故障诊断模型性能对比如表5所示。

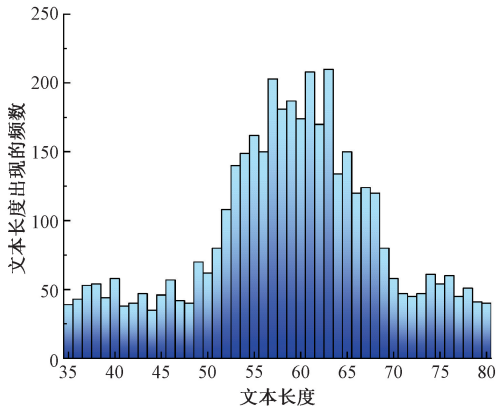

图 11　故障文本长度统计图

Fig. 11　Faut text length statistics histogram

表 3　高速铁路道岔多级故障诊断模型超参数设置

Table 3　Hyperparameter setting of multi-level fault diagnosis model for high-speed railway turnout

超参数名称	超参数值
epoch	25
学习率	0.001
Embedding	300
LSTM 隐藏节点	256
LSTM 层数	2
CNN 卷积核个数	256
CNN 卷积核大小	2, 3, 4
CNN 激活函数	ReLU
CNN 损失函数	交叉熵
优化函数	Adam
batch_size	128
pad_size	72
Dropout	0.5

表 4　对比模型参数设置

Table 4　Contrast models parameter settings

	参数	数值
CNN	卷积核大小	3,4,5
	卷积核个数	128
LSTM	LSTM 层数	1
	LSTM 隐藏节点	256
BiLSTM	LSTM 层数	2
	LSTM 隐藏节点	128

表 5 的诊断结果显示,本文模型在文本长度不一的一级故障和二级故障数据集的精确率、召回率和 F_1 值分别不小于 93.67%、91.56%和 92.60%。CNN 模型由于未考虑文本上下文语义关系,造成前后相关性特征的丢失,故针对整体故障文本特征的学习机制有所欠缺,诊断准确率较低;BiLSTM 虽不能突出局部重要信息,但却考虑了未来数据时序的影响,一级故障准确率比 CNN 模型提高了 3.18 %,二级故障准确率提高了 3.05%。本文设计出的 BiLSTM 和 CNN 组合神经网络模型在一级故障和二级故障测试集上的准确率分别达到 95.62%和 93.81%,相比 CNN 神经网络模型一级故障诊断准确率提升 6.28%,二级故障诊断准确率提升 5.14%,相比 LSTM 神经网络一级故障准确率提升 7.45%,二级故障准确率提升 3.36%,相比 BiLSTM 神经网络一级故障诊断准确率提升 3.10%,二级故障诊断准确率提升 2.09%, 效果最优。说明 BiLSTM-CNN 组合神经网络模型更能充分发挥出 BiLSTM 对维护台账上下文特征信息能力以及 CNN 对局部重点文本信息提取优势。同时由于本文将 BiLSTM 和 CNN 组合使用,模型结构相对复杂,因此训练时间比耗时最短的 LSTM 长 0.78 min,但测试时间相较于耗时最少的 CNN 仅高出 0.67 s,就高速铁路道岔故障诊断任务而言已能满足其快速辨识的要求。综上,本文模型在诊断精度和效率方面均可达到理想效果。

表 5　不同故障诊断模型性能对比

Table 5　Performance comparison of different fault diagnosis models

方法	级别	精确率/%	召回率/%	F_1 值/%	准确率/%	训练时间/min	测试时间/s
CNN	一级故障	92.67	93.34	93.00	89.34	5.34	1.36
	二级故障	88.23	90.12	89.17	88.67		
LSTM	一级故障	88.77	87.54	88.15	88.17	5.27	1.47
	二级故障	90.78	92.13	91.45	90.45		
BiLSTM	一级故障	93.61	94.08	93.84	92.52	5.76	1.86
	二级故障	89.05	93.49	91.22	91.72		
本文模型	一级故障	95.45	96.78	96.11	95.62	6.05	2.03
	二级故障	93.67	91.56	92.60	93.81		

3.6　消融实验

为验证本文的 B-SMOTE 算法和设计的字词融合向量对故障诊断性能的提升效果,本文在高速道岔故障文本数据集上设计了两组消融实验。

1)B-SMOTE 消融实验

将本文模型与未采用 B-SMOTE 算法的模型进行对

比实验。二者分类结果如表6所示。从表6可知,本文模型采用B-SMOTE算法后,处理不均衡文本的综合精确率提升4.72%,召回率提升5.73%,F_1值提升5.22%。

表6　是否采用B-SMOTE对比结果

Table 6　The comparison results of whether to use B-SMOTE

是否采用B-SMOTE	精确率/%	召回率/%	F_1值/%
否	88.73	89.45	89.09
是	93.45	95.18	94.31

图12为未使用B-SMOTE算法模型的混淆矩阵,可以看出由于故障文本分布不均衡,高速铁路道岔多级故障诊断模型中的少数类故障样本易被错分至多数类故障样本,例如少数类的滑床板故障(B2)和密检器接点组故障(C2)分类准确率较低。将图12和13的对比,可看出本文模型在采用B-SMOTE算法后,小类别样本的分类准确度有所提升,同时也保持了多数类样本的分类准确度,处理不均衡数据集的表现较优。

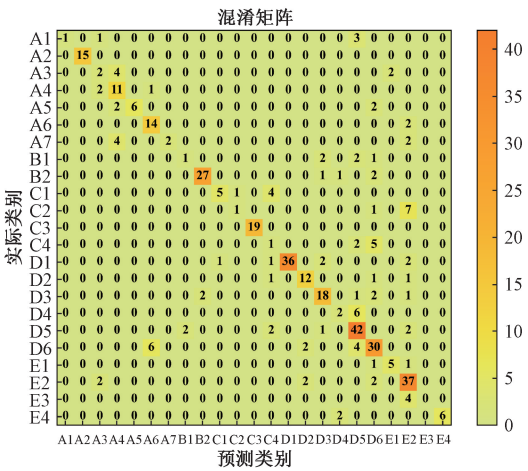

图12　未使用B-SMOTE算法的混淆矩阵

Fig. 12　Confusion matrix without B-SMOTE algorithm

2)字词融合消融实验

将本文设计的字词融合向量表达文本特征的方法与单独的字向量和词向量方法进行比较。不同故障文本特征表示方式的评价指标对比如图14所示,可得词向量的特征表示效果略优于字向量,但单独采用字向量或词向量进行特征表示效果均不如字词融合理想。字词融合相比字向量与词向量准确率分别提升5.48%、4.36%,表明字词融合结合字向量和词向量各自的优点,能够语义消歧和避免边界词划分错误,有效提高模型的诊断效果。

4　结　论

针对高速铁路道岔诊断问题,本文提出了一种基于

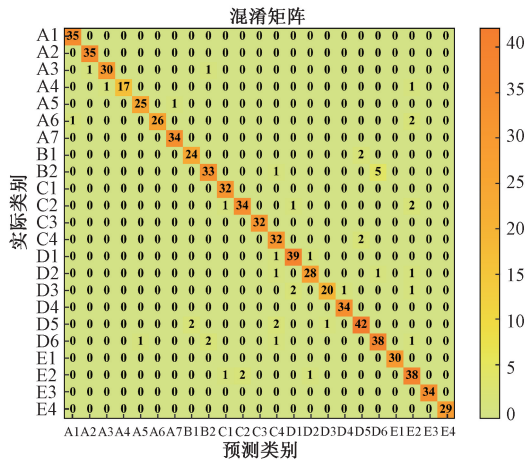

图13　本文模型的混淆矩阵

Fig. 13　Confusion matrix of this paper's model

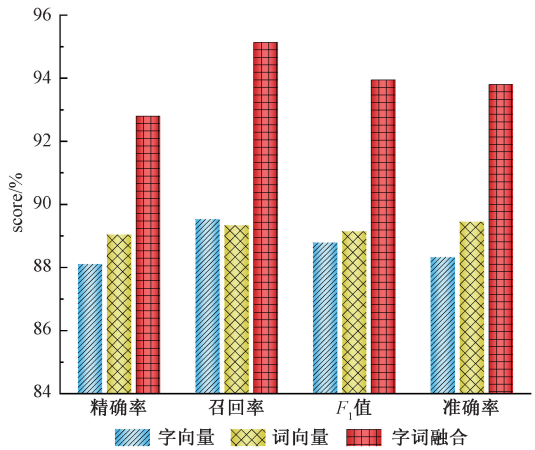

图14　不同文本特征表示方法对比

Fig. 14　Comparison of different text feature representation methods

深度故障文本挖掘的高速铁路道岔多级故障诊断模型,并以我国高速铁路道岔故障文本数据进行实验分析与性能对比,得出如下结论:

1)相比使用单独的字向量或词向量表示故障文本特征,本文所提加入位置特征的字词融合向量方法在避免边界词错误划分的同时保证包含丰富的语义,语义消歧效果显著,最高可将模型准确率提升5.48%。

2)考虑文本数据的不均衡性会对模型故障诊断性能产生极大的影响,本文提出在模型中加入B-SMOTE算法,实验结果表明,本文模型使用B-SMOTE优化故障文本数据分布后,模型综合精确率、召回率、F_1值可分别提升4.72%、5.73%、5.22%。

3)本文提出在道岔故障诊断模型中组合使用BiLSTM-CNN算法,结果表明本文模型可以满足高速铁路道岔故障快速诊断的要求,并且在故障文本长度具有

一定跨度时,相较单独使用 CNN、LSTM 和 BiLSTM 算法,本文模型的综合评价指标分别提升了 3.89%、4.77%和 2.01%,故障诊断性能最优。

综上所述,本文提出的基于深度故障文本挖掘的多级故障诊断方法可有效提升高速铁路道岔故障诊断模型性能,获得更高故障诊断准确率,提高故障定位精度,缩短故障延时。该方法也可应用于面向文本的其他铁路设备故障诊断,促使铁路设备故障诊断智能化。

参考文献

[1] 李新琴, 史天运, 李平, 等. 基于文本的高速铁路信号设备故障知识抽取方法研究[J]. 铁道学报, 2021, 43(3):92-100.
LI X Q, SHI T Y, LI P, et al. Research on knowledge extraction method for high-speed railway signal equipment fault based on text[J]. Journal of the China Railway Society, 2021, 43(3): 92-100.

[2] 武晓春, 楚昕. 基于小波包分解与 GG 模糊聚类的转辙机退化阶段划分研究[J]. 铁道学报, 2022, 44(1): 79-85.
WU X CH, CHU X. Research on division of degradation stage of turnout equipment based on wavelet packet decomposition and GG FUZZY clustering[J]. Journal of the China Railway Society, 2022, 44(1): 79-85.

[3] 林凤涛, 吴涛, 杨洋, 等. 高速铁路辙叉区钢轨打磨廓形设计方法[J]. 交通运输工程学报, 2021, 21(6):124-135.
LI F T, WU T, YANG Y, et al. Design method of rail grinding profile in frog area of high-speed railway[J]. Journal of Traffic and Transportation Engineering, 2021, 21(6): 124-135.

[4] FAN G, FAN L, WANG Z F, et al. Research on multilevel classification of high-speed railway signal equipment fault based on text mining[J]. Journal of Electrical and Computer Engineering, 2021, 2021(2): 1-11.

[5] ZHOU L J, DANG J W, ZHANG Z H. Fault classification for on-board equipment of high-speed railway based on attention capsule network[J]. International Journal of Automation and Computing, 2021, 18(5): 814-825.

[6] 韩广, 卜桐, 王明明, 等. 基于双通道双向长短时记忆网络的铁路行车事故文本分类[J]. 铁道学报, 2021, 43(9):71-79.
HAN G, BU T, WANG M M, et al. Text classification of railway traffic accidents based on dual-channel bidirectional long short term memory network[J]. Journal of the China Railway Society, 2021, 43(9): 71-79.

[7] 朱芳鹏, 王晓峰. 面向船舶工业新闻的文本分类[J]. 电子测量与仪器学报, 2020, 34(1): 149-155.
ZHU F P, WANG X F. Text classification for ship industry news[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(1): 149-155.

[8] ZHOU Y J, DENG D P, CHI J H. A short text classification algorithm based on semantic extension[J]. Chinese Journal of Electronics, 2021, 30(1): 153-159.

[9] KIM Y. Convolutional neural networks for sentence classification[J]. Computation and Language, 2014, arXiv: 1408.5882.

[10] 上官伟, 孟月月, 杨嘉明, 等. 基于 LSTM-BP 级联网络的列控车载设备故障诊断[J]. 北京交通大学学报, 2019, 43(1):54-62.
SHANGGUAN W, MENG Y Y, YANG J M, et al. LSTM-BP neural network based fault diagnosis for on-board equipment of Chinese train control[J]. Journal of Beijing Jiaotong University, 2019, 43(1): 54-62.

[11] 周庆华, 李晓丽. 基于 MCNN 的铁路信号设备故障短文本分类方法研究[J]. 铁道科学与工程学报, 2019, 16(11): 2859-2865.
ZHOU Q H, LI X L. Research on short text classification method of railway signal equipment fault based on MCNN[J]. Journal of Railway Science and Engineering, 2019, 16(11): 2859-2865.

[12] CHEN S S, DING Y D, XIE Z F, et al. Chinese Weibo sentiment analysis based on character embedding with dual-channel convolutional neural network[C]. 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018:107-111.

[13] DOS S, GATTI M. Deep convolutional neural networks for sentiment analysis of short texts[C]. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Technical Papers, 2014: 69-78.

[14] 杨连报, 李平, 薛蕊, 等. 基于不平衡文本数据挖掘的铁路信号设备故障智能分类[J]. 铁道学报, 2018, 40(2):59-66.
YANG L B, LI P, XUE R, et al. Intelligent classification of faults of railway signal equipment based on imbalanced text data mining[J]. Journal of the China Railway Society, 2018, 40(2): 59-66.

[15] GEORGIOS D, FERNANDO B. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences Volume 501, 2019: 118-135.

[16] 李晓英,杨名,全睿,等. 基于深度学习的不均衡文本分类方法[J]. 吉林大学学报(工学版),2022,52(8):

1889-1895.
LI X Y, YANG M, QUAN R, et al. Unbalanced text classification method based on deep learning[J]. Journal of Jilin University (Engineering and Technology Edition),2022,52(8):1889-1895.

[17] 周璐婕, 党建武, 王瑜鑫, 等. 基于CNN-CRSF组合模型的列控车载设备故障诊断[J]. 铁道学报, 2020, 42(11):94-101.
ZHOU L J, DANG J W, WANG Y X, et al. Fault diagnosis for on-board equipment of train control system based on CNN-CSRF hybrid mode[J]. Journal of the China Railway Society, 2020, 42(11): 94-101.

[18] 冯斌, 张又文, 唐昕, 等. 基于BiLSTM-Attention神经网络的电力设备缺陷文本挖掘[J]. 中国电机工程学报, 2020, 40(S1):1-10.
FENG B, ZHANG Y W, TANG X. Power equipment defect record text mining based on BiLSTM-attention neural network[J]. Proceedings of the CSEE, 2020, 40(S1): 1-10.

[19] ZHANG X C, QIU X P, PANG J M, et al. Dual-axial self-attention network for text classification[J]. Science China(Information Sciences), 2021, 64(12): 80-90.

[20] 林海香, 陆人杰, 卢冉, 等. 基于文本挖掘的铁路信号设备故障自动分类方法[J]. 云南大学学报(自然科学版), 2022, 44(2):281-289.
LIN H X, LU R J, LU R, et al. Automatic classification method of railway signal fault based on text mining[J]. Journal of Yunnan University (Natural Sciences Edition), 2022, 44(2): 281-289.

[21] 陈仁祥, 王帅, 杨黎霞, 等. 弓网接触力长短时记忆网络预测的受电弓主动控制与仿真[J]. 仪器仪表学报, 2021, 42(5): 192-198.
CHEN R X, WANG SH, YANG L X, et al. Active control and simulation for pantograph based on contact force prediction of long short-term memory network[J]. Chinese Journal of Scientific Instrument, 2021, 42(5): 192-198.

[22] 李卫疆, 漆芳, 余正涛. 基于多通道特征和自注意力的情感分类方法[J]. 软件学报, 2021, 32(9): 2783-2800.
LI W J, QI F, YU ZH T. Sentiment classification method based on multi-channel features and self-attention[J]. Journal of Software, 2021, 32(9): 2783-2800.

[23] CHEN Y, CHANG R, GUO J F. Effects of data augmentation method Borderline-SMOTE on emotion recognition of EEG signals based on convolutional neural network[J]. IEEE Access Volume 9, 2021: 47491-47502.

[24] 王琛, 王颖, 郑涛, 等. 基于ResNet-LSTM网络和注意力机制的综合能源系统多元负荷预测[J]. 电工技术学报, 2022, 37(7):1789-1799.
WANG CH, WANG Y, ZHENG T, et al. Multi-energy load forecasting in integrated energy system based on ResNet-LSTM network and attention mechanism[J]. Transactions of China Electrotechnical Society, 2022, 37(7): 1789-1799.

[25] SHANTHAKUMAR S, SHAKILA S, SUNETH P, et al. Environmental sound classification using deep learning[J]. Instrumentation, 2020, 7(3): 15-22.

[26] DENG J F, CHENG L L, WANG Z W. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification[J]. Computer Speech & Language, 2021, 68: 101182.

[27] ZHOU L, DANG J, ZHANG Z. Research on fault diagnosis for on-board equipment of train control system based on imbalanced text classification[J]. Journal of Applied Science and Engineering, 2021, 24(2): 167-175.

作者简介

林海香(通信作者),分别在2000年和2007年于兰州交通大学获得学士学位和硕士学位,2020年于同济大学获得博士学位,现为兰州交通大学副教授,主要研究方向为交通信息数据挖掘。
E-mail: linhaixiang@mail.lzjtu.cn

Lin Haixiang received her B. Sc. and M. Sc. degrees from Lanzhou Jiaotong University in 2000 and 2007, and Ph. D. degree from Tongji University in 2020, respectively. Now she is an associate professor at Lanzhou Jiaotong University. Her main research interest includes traffic information data mining.

赵正祥(Corresponding author),现为兰州交通大学硕士研究生,主要研究方向为自然语言处理。
E-mail: 511229689@qq.com

Zhao Zhengxiang is a M. Sc. candidate at Lanzhou Jiaotong University. His main research interest includes natural language processing.