DOI: 10.19783/j.cnki.pspc.200367

基于深度语义学习的智能录波器自配置方法

陈 旭[1]，张 弛[1]，刘千宽[1]，彭 业[1]，周达明[2]，甄家林[2]

（1. 中国南方电网电力调度控制中心，广东 广州 510663；2. 武汉凯默电气有限公司，湖北 武汉 430023）

摘要：智能录波器的基础配置工作是将全站配置描述文件(Substation Configuration Description, SCD)中智能二次设备(Intelligent Electronic Device, IED)各运行数据输出端口的地址信息分类映射至录波器不同信息组中。目前主流配置方法是针对端口的文本描述进行人工配置。在大规模高电压等级变电站内端口文本描述繁杂，人工操作耗时长，工作量大。针对该现状，提出了基于字符级 TextCNN 深度语义学习的智能录波器自配置方法。首先利用 word2vec 模型针对高维稀疏的文本样本矩阵进行降维与稠密化处理，实现字符词向量的分布式表达。之后建立 TextCNN 模型，基于其多层次抽象化提取样本特征的结构特点进行文本语义挖掘与分类。依据文本分类结果实现端口地址信息的分类映射。案例分析表明，基于 TextCNN 模型的录波器自配置方法具有分类时间短与分类精度高的优点，提高了录波器自动化配置的准确性。

关键词：智能录波器；信息自配置；文本挖掘；词向量分布式表达；文本卷积神经网络

Automatic configuration method of intelligent recorder based on deep semantic learning

CHEN Xu[1], ZHANG Chi[1], LIU Qiankuan[1], PENG Ye[1], ZHOU Daming[2], ZHEN Jialin[2]

(1. Power Dispatching and Control Center of China Southern Power Grid, Guangzhou 510663, China;
2. Wuhan Kemov Electric Co., Ltd., Wuhan 430023, China)

Abstract: The groundwork of an intelligent recorder is to map the address information of Intelligent Electronic Device (IED) data output ports within the Substation Configuration Description (SCD) file to different recorder information groups. At present, the mainstream mapping method is manual configuration based on the text description of the output port address. In a large-scale and high-voltage substation, output port address text descriptions are numerous and diverse. The manual operation often takes a huge amount of time. Also the labor cost is high. In order to improve the situation, this paper proposes an automatic information configuration method for an intelligent recorder based on a Text Convolutional Neural Network (TextCNN), which deeply learns the semantics of tests at the character level. First, to realize the distributive expression of text character vectors, the word2vec is introduced to thicken and reduce the dimension of the sparse and high-dimensional sample vector matrix. Secondly, this paper establishes a TextCNN model for text semantic analysis. Based on its ability to obtain multi-level abstract features of samples, the text semantic mining is therefore performed and its classification is realized. The mapping of the IED output port address information is completed based on the text classification result. The case study shows that the automatic intelligent recorder information configuration based on the TextCNN model has the characteristics of short classification time and high accuracy. This improves the accuracy of the automatic configuration for the intelligent recorder.

This work is supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. (No. 000000KK 52180019).

Key words: intelligent recorder; automatic information configuration; text mining; distributive expression of word vectors; text convolutional neural network

0 引言

智能录波器集故障暂态录波、网络报文记录、二次系统可视化、智能运维、保信子站功能于一体，是实现二次设备状态监测与故障信息集总分析的核心装置。将智能二次设备 IED(Intelligent Electronic Device)各数据输出端口的地址信息映射至录波器相应信息组是录波器的基础配置工作，投运时，录波

基金项目：南方电网公司科技项目资助(000000KK52180019)

器通过解析 IED 输出端口地址信息实现对 IED 运行数据的分类监测。当前，录波器配置方法是依据 IED 端口的文本描述，人工映射端口地址信息至相关信息组。在高压大型变电站中，IED 数目繁多，人工映射工作量巨大，配置时间长达数十天。因此，实现智能录波器自动化配置势在必行。

目前，自动化配置问题在于 IED 数据输出端口描述文本的不一致，尽管已有规范对不同厂家设备描述进行一定程度约束，仍无法保证信息描述统一与文本语义无歧义。为满足自动映射的高准确性要求，提高映射系统对半结构化文本分析能力，需针对海量信息描述进行文本深度挖掘与语义分析，建立分类映射体系。

本文提出一种基于 TextCNN 模型的录波器自配置方法。首先引入文本表示模型 word2vec，实现 IED 端口描述文本的词向量表示及词间关联关系映射；随后引入基于 TextCNN 模型的分类映射单元，并对模型关键参数进行优化，提高其特征提取与泛化能力；最后通过算例证明 TextCNN 模型具有强大的语义分析能力，能有效提高录波器配置准确度。

文本分类映射流程包括文本预处理、特征提取与分类预测三个步骤。文本预处理包括文本分词与表示[1-3]；传统特征提取方法依靠特征值函数筛选特征[4-6]；传统分类模型包括决策树[7-8]、贝叶斯分类器[9-10]、支持向量机[11-12]等。但传统特征提取方法特征选取准确性低，易遗漏特征信息；传统分类模型泛化能力差且文本语义挖掘不全[13]。

卷积神经网络(Convolutional Neural Network, CNN)是深度学习理论中一类典型架构，其构造多个非线性卷积层对输入样本的多维特征向量进行局部特征提取并筛选最具表征意义的特征向量进行分类预测。文献[14]成功将 CNN 与仿射变换结合，实现对变化的图像实体识别；文献[15]将 CNN 用于恶意代码识别，提高了防御软件对恶意代码的甄别能力；文献[16]将 CNN 用于视频伪造区域检测，有效提升了伪造对象识别率。

在电力系统的应用方面，文献[17]将 CNN 结合随机森林用于三相电压暂降分类，分类结果准确率高；文献[18]将 CNN 引入继保装置状态监测，提高了状态判别准确率；文献[19]将一维 CNN 引入电能质量扰动信号分类，算法鲁棒性好，识别扰动信号能力强；文献[20]将 RCNN 引入输电线路树状障碍物辨识与建模，提高了障碍物建模效率。总体而言，CNN 的特征选择与分类学习能力强大，足以实现文本语义深度挖掘。

本文提出一种基于 TextCNN 模型的录波器自配置方法。首先引入文本表示模型 word2vec，实现 IED 端口描述文本的词向量表示及词间关联关系映射；随后引入基于 TextCNN 模型的分类映射单元，并对模型关键参数进行优化，提高其特征提取与泛化能力；最后通过算例证明 TextCNN 模型具有强大的语义分析能力，能有效提高录波器配置准确度。

1 录波器配置信息解析

录波器配置所需 IED 端口地址信息及其文本描述信息均储存于全站配置描述文件(Substation Configuration Description, SCD)中，该文件由节点与属性组成，一级节点 IED 集合了该二次设备所有数据输出端口的描述信息。IED 节点内有层层嵌套的子节点，通过不断向下检索子节点属性值可获取各端口地址信息与端口文本描述，并可实现互相匹配。本文采用 Python 语言的 xml.Etree.ElemenTree 模块解析 SCD 文件，实现节点检索、端口地址配置数据与端口文本描述读取操作，并将文本与地址信息的匹配关系以 python 字典变量形式存储，用于后续端口地址信息映射时对端口文本描述分类结果的读取。具体检索结构图如图 1 所示。各节点由 IEC61850 规约规定，图中节点包括接入点 Accesspoint、逻辑设备 LDevice、逻辑节点 LN、数据集 DataSet、功能约束数据属性 FCDA、数据对象实例 DOI 与属性实例 DAI。节点箭头所指圆点代表从属子节点，实线相连圆点代表节点属性，长虚线框内属性值用于构造地址信息，短虚线框包含目标描述文本，实线箭头代表属性值检索。

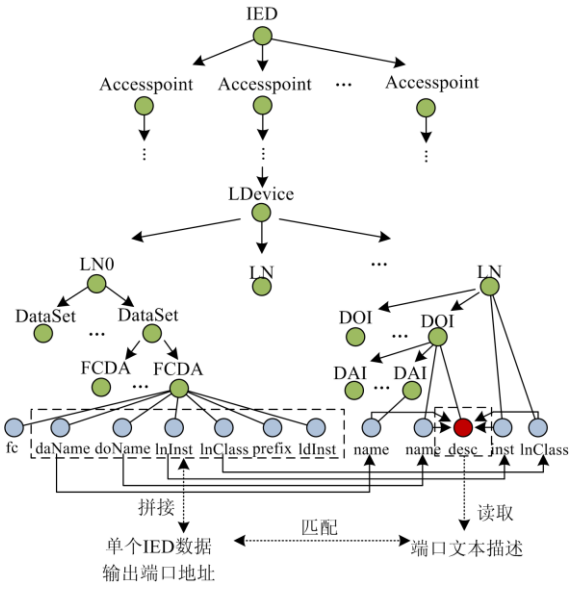


图 1 录波器配置信息解析结构图

Fig. 1 Analysis of recorder configuration datasets

2　基于 word2vec 的配置信息文本描述分布式表达

配置信息文本描述通常篇幅短小，专业词汇偏多，中英文混杂，采用分词工具包处理时通常无法识别“光口”、“硬压板”等专业词组，易出现误分词，因此本文采用字符级 TextCNN 进行文本分类处理，直接挖掘字符间的语义联系，减少误分词造成的干扰。在词向量表示方面，本文利用 word2vec 模型处理输入样本，将字符拼接表示的 One-Hot 向量矩阵作为模型输入，模型最终输出反映词义衔接的低维稠密词向量。

2.1 词袋模型

文本在输入分类模型前需通过向量化处理以表示文本语义，常用文本向量化方法是词袋模型(Bag of Words, BOW)表示法，词袋模型依据词语在词库中的索引值对语料进行单维有效编码，不同词语向量取值为 1 的维度各不相同，文本表示矩阵由词语向量拼接而成，由矢量赋值方法不同可分为 One-Hot 表示法、TF 表示法与 TF-IDF 表示法。词袋模型操作简单，但其将语料视作词语简单拼接，在向量化过程中忽略了目标词语与上下文语法语序的衔接，无法体现词语间共现关系；并且文本表示矩阵维度很高且数据稀疏，造成模型分类庞大的计算量。因此，针对词袋模型所求的高维稀疏文本表示矩阵，一定程度的降维与稠密化处理必不可少。

2.2 word2vec 分布式表达模型

词向量的分布式表达将高维稀疏文本表示矩阵映射至一低维向量空间[21]，在这一空间中，词义关联紧密的词向量之间欧式距离较小。从而克服了基于词袋模型表示的词向量相互孤立的缺陷，同时，处理所得向量矩阵低维且稠密，大大减少了文本分类模型的运行时间与内存资源。word2vec 分布式表达模型由 Miklov 等人提出[22-24]，其通过设立滑动窗口，利用局部上下文特征依次求解窗口中心词向量值，有效避免了低关联文本的重复输入，从而大幅减小计算冗余度。word2vec 包括依据上下文预测中心词的 CBOW 模型与依据中心词推测上下文的 skip-gram 模型，本文采用 CBOW 模型处理 One-hot 矩阵，利用负采样技术优化模型。

CBOW 模型示意图如图 2 所示，图中词典词汇量大小设为 V，隐藏层维度设为 N，模型输入为 C 个上下文词语的 One-Hot 向量，输入层权重设为 $\boldsymbol{W}_1$，隐藏层权重设为 $\boldsymbol{W}_2$，当模型仅输入单个上下文词语 w_i 时，设输入矢量中 $x_i=1$，可得隐藏层输出矢量 h 与输出层对应词语输入值 u_i 如下。

$$h=\boldsymbol{W}_1^{\mathrm{T}}x=\boldsymbol{W}_{(i,\cdot)}^{\mathrm{T}}=\boldsymbol{v}_{1_{w_i}} \tag{1}$$

$$u_i=\boldsymbol{v}_{2_{w_i}}^{\mathrm{T}}h \tag{2}$$

式中：$\boldsymbol{v}_{1_{w_i}}$ 为 $\boldsymbol{W}_1$ 的第 i 列矢量，称为 w_i 的中心词向量；$\boldsymbol{v}_{2_{w_i}}$ 为 $\boldsymbol{W}_2$ 的第 i 列矢量，称为 w_i 的上下文向量。

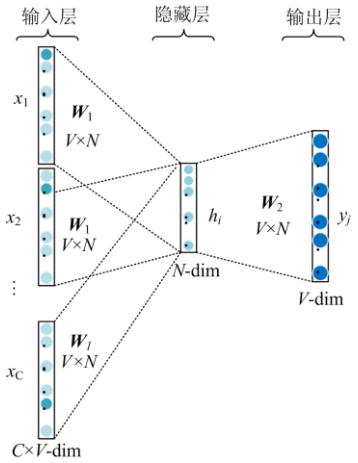


图 2 CBOW 模型示意图

Fig. 2 Structure of the CBOW model

倘若目标词语为 w_j，设 u_j 为输出层输入值，y_j 为输出层的输出值，则有式(3)，模型损失函数 E 如式(4)所示。CBOW 模型通过梯度下降算法更新权重向量 $\boldsymbol{W}_1$ 与 $\boldsymbol{W}_2$，使损失函数 E 最小化，最终，模型输出单个词的词向量矢量表示为其中心词向量与上下文向量的平均值。

$$y_j=\mathrm{p}\left(w_j\mid w_i\right)=\frac{\exp\left(u_j\right)}{\sum_{k=1}^{V}\exp\left(u_k\right)}=\frac{\exp\left(\boldsymbol{v}_{2_{w_j}}^{\mathrm{T}}\boldsymbol{v}_{2_{w_j}}^{\mathrm{T}}\boldsymbol{v}_{1_{w_i}}\right)}{\sum_{k=1}^{V}\exp\left(\boldsymbol{v}_{2_{w_k}}^{\mathrm{T}}\boldsymbol{v}_{1_{w_i}}\right)} \tag{3}$$

$$E=-\log\left(\mathrm{p}\left(w_j\mid w_i\right)\right)=\log\sum_{k=1}^{V}\exp\left(\boldsymbol{v}_{2_{w_k}}^{\mathrm{T}}\boldsymbol{v}_{1_{w_i}}\right)-\boldsymbol{v}_{2_{w_j}}^{\mathrm{T}}\boldsymbol{v}_{1_{w_i}} \tag{4}$$

当上下文包括多个词语时，隐藏层输出将对各单词向量求均值后输出：

$$h=\frac{1}{C}\boldsymbol{W}_1^{\mathrm{T}}\left(\sum_{k=1}^{V}x_k\right)=\frac{1}{C}\left(\sum_{k=1}^{V}v_{1_{w_k}}\right)^{\mathrm{T}} \tag{5}$$

此时，模型损失函数 E 定义为

$$E=-\log\left(\mathrm{p}\left(w_j\mid w_{i,1},w_{i,2}\cdots w_{i,C}\right)\right)=\log\sum_{k=1}^{V}\exp\left(\boldsymbol{v}_{2_{w_k}}^{\mathrm{T}}\boldsymbol{v}_{1_{w_i}}\right)-\boldsymbol{v}_{2_{w_j}}^{\mathrm{T}}h \tag{6}$$

在利用梯度下降法对模型权重即词向量矩阵进行更新的过程中，针对中心词向量矩阵的更新计算

比较简便，但上下文向量矩阵的更新计算需要遍历词典中各词语以求解总预测误差，用于更新词向量矩阵，更新过程计算量较大，为提高词向量更新效率，本文引入负采样技术，每次选取目标词语词向量作为正样本，并随机挑选 5 个非目标词语词向量作为负样本进行更新，设目标词语为 w_j，非目标词语为 w_i，则 w_i 被选为负样本概率如式(7)所示。

$$p(w_i)=\frac{f(w_i)^{\frac{3}{4}}}{\sum_{k=1}^{V}\left(f(w_k)^{\frac{3}{4}}\right)} \tag{7}$$

式中，$f(w_i)$ 为单词 w_i 出现的频次。此时，损失函数表示为

$$E=-\log\sigma\left(\boldsymbol{v}_{2_{w_j}}^{\mathrm{T}}h\right)-\sum_{w_i\in W_{\mathrm{neg}}}\log\sigma\left(-\boldsymbol{v}_{2_{w_i}}^{\mathrm{T}}h\right) \tag{8}$$

式中：W_{neg} 为负样本词语集合；σ 为 Sigmoid 函数。

本文 CBOW 隐藏层−输出层拥有 $10\times105=1\,050$ 个权重值，利用负采样技术每次仅需要更新 $10\times6=60$ 个权重，相当于仅需要更新 5.7%的权重，计算效率大幅提高。

3　基于文本卷积神经网络的录波器信息自配置方法

目前，卷积神经网络已成为计算机视觉领域中最强大的深度学习框架，其针对图像强大的特征抽象与提取能力得到了诸多学者青睐；在文本分类方面，由于文本语义表达有赖于关键词的遴选，因此辨析各部分文本的关联关系尤为重要，卷积神经网络可以在高于一维的词向量空间中进行多区域局部特征提取，通过不断更新卷积核权重值，实现特征提取进一步抽象化与典型化，针对这类特征值进行分类训练，可以有效提高模型预测精度。

3.1　TextCNN 模型文本分类器

1) 统计训练样本集各词语在样本中出现的频率，基于 jieba 分词词库建立字符级索引字典，利用字典检索将待训练样本语料转换为 One-Hot 矩阵，输入 CBOW 词向量处理层，获得低维稠密词向量矩阵。该矩阵将作为 TextCNN 模型输入。

TextCNN 模型架构如图 3 所示。具体分类步骤包括文本向量卷积处理、池化处理、dropout 处理与分类预测。

2) 采用文本处理卷积核，针对词向量矩阵进行卷积操作，卷积核列数与词向量大小维度一致，保证单个词不同维度特征提取的全面性；行数取决于需要考虑词语的个数，不同行数卷积核决定了局部

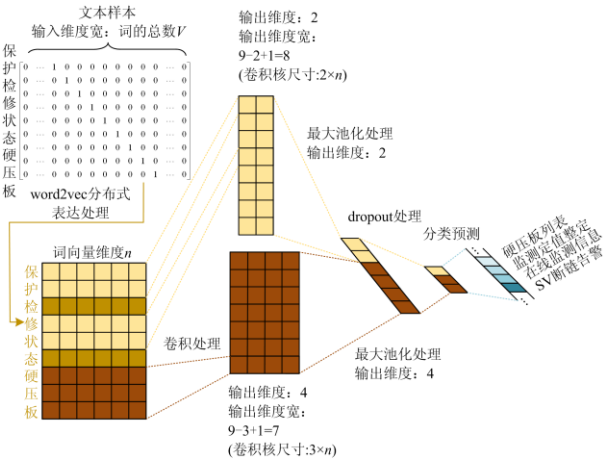


图 3　文本卷积操作图

Fig. 3 Structure of text convolution

特征考察范围。以图 3 为例，提取文本描述为“保护检修状态硬压板”，词向量矩阵行数为文本用词个数，列数为词向量维度，记为 n，图中包括尺寸为 $2\times n$ 与 $3\times n$ 两种卷积核，分别与矩阵对应值相乘得到卷积结果，卷积处理后将各元素进行偏置处理，并输入非线性函数，求得卷积层输出结果。卷积核操作可视为对不同指标值赋权的操作，设卷积核的某一单元为 w，对应的词向量处理单元为 x，卷积层输出结果为 o，模型偏置统一设为 b，则卷积计算过程为

$$o=\sigma(w\cdot x+b) \tag{9}$$

式中，σ 为某非线性函数，常用非线性函数包括 ReLU 函数、Sigmoid 函数等。本文选取 *ReLU* 函数作为 σ 函数。

3) 卷积处理后将输出结果输入池化层，本文使用的池化方法为最大池化(max-pooling)方法，该方法可以固定本层输出维度，提取局部文本中最具代表性的语义信息。池化层输出向量 y 可由式(10)表示。

$$y=\max\left(o_1,o_2\cdots,o_t\right) \tag{10}$$

式中，$o_1,o_2,\cdots,o_t$ 分别为不同卷积核处理词向量矩阵后的输出结果。

4) 池化处理后将输出结果输入 dropout 层，dropout 层将随机舍弃池化层部分输出。避免对某些局部特征过分依赖，强迫神经网络学习更加鲁棒的特征，防止模型过拟合，图中 dropout 层保留比例为 0.5。

5) 将 dropout 层输出结果输入全连接层进行分类预测，全连接层计算公式与式(1)类似，非线性函数采用 softmax 函数，可计算不同类别的隶属概率，依据最大概率原则确定文本隶属信息组类别。

在进行基于反向传播算法训练网络参数过程

中，损失函数取交叉熵函数 J，利用随机梯度下降法更新模型参数，J 的表达式如式(11)所示。

$$J=-\frac{1}{N}\sum_{i=1}^{N}y_i\log\left[\sigma\left(W\otimes X+b\right)\right]+\frac{\lambda}{2}\|W\|_2^2 \quad (11)$$

式中：$\otimes$ 表示卷积操作；$\frac{\lambda}{2}\|W\|_2^2$ 表示 L2 正则项，可以增强模型泛化能力，$\|W\|_2$ 表示卷积核 W 的第二范数；λ 为正则化系数；N 为输入文本个数，参数更新方程如式(12)、式(13)。

$$W_{ij}(k)=W_{ij}(k)-\eta\frac{\partial}{\partial W_{ij}(k)}J\left(W,b\right)-\eta\lambda W_{ij}(k) \quad (12)$$

$$b_i(k)=b_i(k)-\eta\frac{\partial}{\partial b_i(k)}J\left(W,b\right) \quad (13)$$

式中：i 表示层数；j 表示同一批次的卷积核选择；k 表示卷积核内权重值；η 表示学习率。

常用二分类模型评估指标包括查全率 R、查准率 P 以及其加权调和平均值 F_1，其表达式如下式所示。

$$P=\frac{TP}{TP+FP} \quad (14)$$

$$R=\frac{TP}{TP+FN} \quad (15)$$

$$F_1=\frac{2\times P\times R}{P+R} \quad (16)$$

式中各参数如表 1 所示。

表 1 评测指标关联关系

Table 1 Correlation between evaluation indicators

样本情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

本文信息自配置问题包括 5 个分类目录，因此本文引进宏评估综合指标 *Marco-F1*(MF_1)进行模型评估，如式(17)所示。

$$MF_1=\frac{2\times MP\times MR}{MP+MR} \quad (17)$$

$$MP=\sum_{i=1}^{n}\frac{P_i}{n} \quad (18)$$

$$MR=\sum_{i=1}^{n}\frac{R_i}{n} \quad (19)$$

式中，n 为分类目录数，本文中，n=5。

3.2 基于文本分类结果的录波器信息自配置

录波器配置文件与 SCD 文件构造类似，在各 IED 节点下囊括有不同待映射信息组子节点，这些节点又包含了数据输出地址解析子节点，投运时录波器将依据地址解析子节点中地址信息的描述，自动获取相关 IED 输出端口的运行数据。

本文在将端口地址配置数据录入前，读取端口描述文本的归类结果，据此分配端口地址配置数据的信息组节点归属。通过 ElementTree 模块遍历 IED 各信息组找寻目标类别，按照文本分类结果自动录入文本描述对应的地址配置数据，进而完成数据集的自动化映射。

4 算例分析

4.1 算例与计算参数介绍

为证明 TextCNN 在录波器信息自配置方面的优越性能，本文选取某三个变电站共 3 000 条继保装置分类文本作为分析样本。样本具有完整的文本描述及归类标注。本文将样本随机均分为 4 份，每份样本包含 750 条分类文本，选择 3 份作为训练集，1 份作为验证集进行 4 次交叉验证，将 4 次验证集所求准确度的平均值作为最终模型的准确度。

智能录波器待映射信息组包括压板信息组、告警信息组、保护监测与状态信息组，本文选取压板信息组中的硬压板列表、告警信息组中 SV 断链告警、GOOSE 断链告警、保护监测与状态信息组的监测定值整定以及在线监测信息五大典型类别作为 SCD 文件分类目录，部分文本样本如表 2 所示。

表 2 部分文本样本

Table 2 A part of samples descriptions

样本类别	样本描述
硬压板列表	远方操作硬压板、断路器就地操作硬压板
监测定值整定	装置温度上限、纵联通道接收功率下限
在线监测信息	事故照明逆变器环境温度、汇控柜湿度
GOOSE 断链告警	第一套智能终端收 5032 开关第一套保护 GOOSE 断链

本文模型采用基于 Python 语言构造的 Tensorflow 工具包，CPU 为 Intel Core i7-3537U，主频 2.0 GHz，模型超参数设置如表 3 所示。

表 3 模型参数设置

Table 3 Model hyper parameter settings

参数描述	设定值	参数描述	设定值
词向量维度	64	全连接层神经元数	128
批处理大小	64	Dropout 保留比例	0.3
卷积核数目	3	学习率	0.001
卷积核尺寸	3	最大迭代次数	30

4.2 算例计算

4.2.1 词语聚类计算

本文采用 word2vec 针对词向量矩阵进行降维处理，所得向量可以体现词语间语义衔接关系，为

直观表现这种关系，本文以词向量间的余弦相似度作为评判语义关联关系紧密程度的指标，随机挑选了 5 个语义关联相对稀疏的词，包括“警”，“地”，“定”，“断”，“光”，作为语义关联关系展示的中心词，并选取各自语义关联度最大的 5 个词，共 30 个词向量作为展示样本，采用主成分分析法(Principal Component Analysis, PCA)对上述输出词向量进行降维聚类，聚类结果如图 4 所示，PC_1 与 PC_2 分别表示原词向量矩阵中方差最大的特征方向及其正交方向，图中词间距离象征语义空间的欧氏距离，以表现词语间的关联关系。

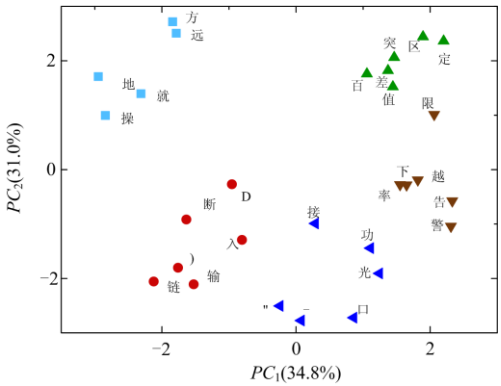


图 4 二维空间聚类图

Fig. 4 Clustering result of two-dimensional space

由图 4 可知，语义衔接紧密的单词，在聚类空间中空间距离小，例如“警”与“告”、“断”与“链”等；语义衔接不够紧密的词空间距离较大，例如“地”与“告”、“定”与“入”等，图 5 抽取词组“警告”，“断链”，“定值”，“就地”与“光口”所包含的 10 个词，计算了上述词间的语义关联度并绘制弦图，图中，不同词的连接线越宽，则关联度越大，可见，组成上述各词组的两个词之间关联度均大于两词与其他词的关联度，该图同样证明了 word2vec 具备体现词语关联关系的能力。

图 5 样本语义关联度分析弦图

Fig. 5 Chord graph of samples semantic relevance analysis

4.2.2 文本分类计算

基于 TextCNN 模型的文本分类训练集与验证集 MF_1 如图 6 所示。由图可知，随着模型权重不断更新，训练集与验证集 MF_1 逐渐收敛，当模型迭代次数为 23 次时，训练集 MF_1 逐渐收敛于 98%之上，迭代次数为 26 次时，验证集 MF_1 为 91.47%并趋于稳定，模型验证集分类预测耗时低于 0.5 s。

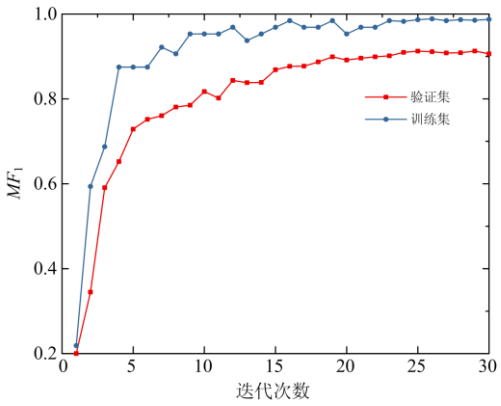


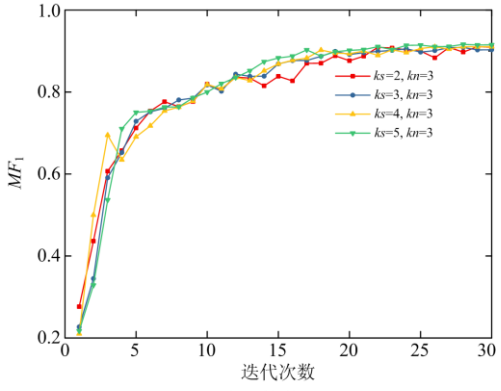
图 6 评估结果的详细变化趋势

Fig. 6 Detailed trends of the datasets evaluation results

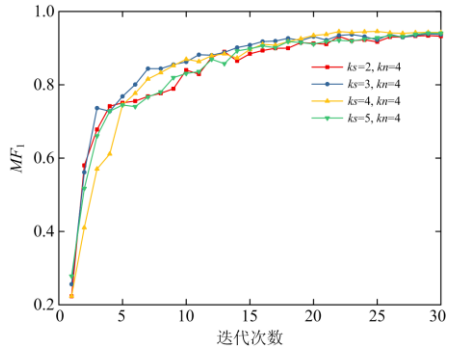
4. 3 模型优化

4.3.1 特征提取性能优化

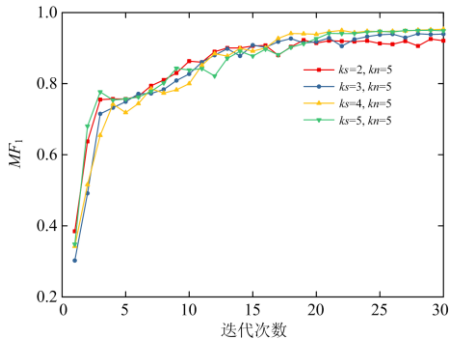
TextCNN 模型参数中，影响文本语义挖掘能力的主要因素包括卷积核大小与卷积核数目，考虑到待分类短文本半结构化特征，特征关键字通常低于 5 个字符，本文分别构造了卷积核数目为 3 至 6，卷积核尺寸为 2 至 5 的模型，取验证集 MF_1 参数作为评估指标，设卷积核数目为 kn，卷积核尺寸为 ks，评估结果如图 7 所示，当卷积核数目为 5，卷积核尺寸为 4 时，模型分类精度最高，MF_1 收敛于 95.34%，收敛迭代次数为 23 次；图 8 比较了不同卷积核数目下，验证集 MF_1 最大值与模型收敛所需迭代次数最小值，可见当卷积核数目 6 时，模型出现过拟合现象，准确度略有下降，且模型复杂度增高，训练收敛时间上升。综上所述，模型卷积核数目为 5，卷积核尺寸为 4 时模型分类能力最强，所有模型验证集分类预测耗时均在 0.5 s 以内。



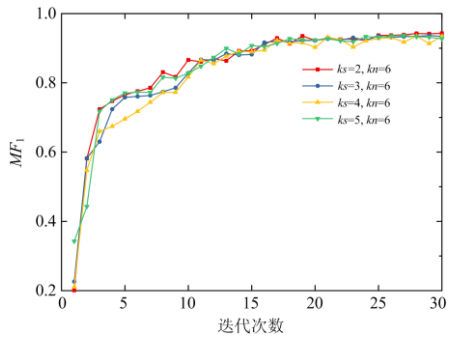
(a) 卷积核数目kn=3



(b) 卷积核数目kn=4



(c) 卷积核数目kn=5



(d) 卷积核数目kn=6

图 7 验证集评估结果的详细变化趋势

Fig. 7 Detailed results of the verified datasets evaluation

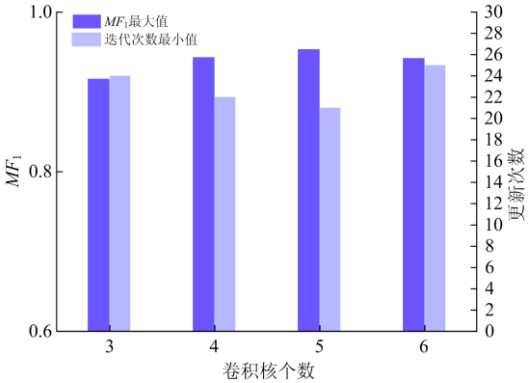


图 8 不同卷积核数目模型性能评估

Fig. 8 Analysis of models ability with different number of convolution kernels

4.3.2 泛化性能优化

TextCNN 具有强大的语义特征提取能力，但也隐藏了泛化性能差的风险，因此本文模型通过引入 dropout 层随机舍弃部分特征提取结果以降低过拟合风险，为寻找 dropout 层较优的神经元保留比例，本文分别设置 0.1、0.3、0.5、0.7、1 五种保留比例的模型，并验证模型分类能力与泛化能力，设 dropout 层保留比例为 dp，评估结果如图 9 所示。

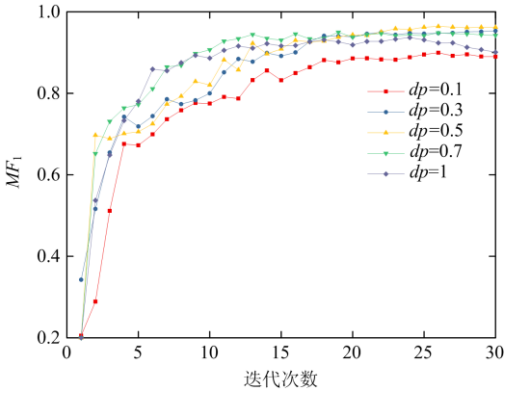


图 9 模型分类与泛化能力对比

Fig. 9 Comparison of the model generalization and classification ability

由图 9 可知：当保留比例 dp 过小，即 $dp=0.1$ 时，模型拟合能力将变差，导致验证集 MF_1 收敛所需迭代次数上升，且分类精度下降；当模型不考虑舍弃神经元提高泛化性能，即 $dp=1$ 时，验证集 MF_1 同样出现了下降的趋势，说明模型复杂度过高，在训练集上已发生过拟合现象；当池化层随机舍弃一半输出值，即 $dp=0.5$ 时，模型分类预测耗时低于 0.5 s，分类精度最高，泛化性能好，此时验证集 $MF_1=96.24\%$。

4.3.3 网络层级优化

TextCNN 模型卷积层数的加深一方面增强了局部特征提取的质量，但另一方面提高了网络复杂度，模型过拟合风险同样增高，上文针对单层特征抽取的 TextCNN 进行了性能优化，本节设置卷积层数目为 1、3、5、7、9，分别测试不同模型对验证集的分类精度及模型训练收敛速度，衡量指标为验证集 MF_1 以及模型训练迭代次数。单层卷积核尺寸为 4，数目为 5，各层级输入特征矩阵均进行区域填充，避免特征值丢失，dropout 层保留比例为 0.5，测试结果如表 4 所示。

由表 4 可知，当卷积层数为 3 层时，验证集分类精度最高，随着层数进一步加深，模型出现过拟合现象，精度逐渐下降，训练数据集迭代次数随着

网络复杂度提高而不断上升。综合考虑上述结果，本文构造卷积层深度为 3 的 TextCNN 模型。

表 4 不同层级模型性能对比

Table 4 Model hyper parameter settings

网络层数	MF_1	迭代次数/次
1	0.962 4	23
3	0.965 3	25
5	0.962 7	28
7	0.948 2	30
9	0.936 0	30

4.4 分类模型比较

为证明 TextCNN 在特征提取与文本分类方面优于传统分类模型与浅层神经网络模型，本文选取了支持向量机(SVM)、k 最近邻算法(kNN)CART 决策树与朴素贝叶斯(NaiveBayes)四种传统分类模型作为对比模型，分别采用词频-逆文档频率(TF-IDF)、潜在狄利克雷分配(LDA)、潜在语义索引(LSI)共 3 种方法进行文本向量化表示。同时选取了浅层神经网络 BPNN 作为分类器，利用 word2vec 处理文本向量作为对比。实验结果如表 5 所示，可见，传统分类模型中，采用 TF-IDF 处理输入文本的 CART 决策树模型分类性能最优，MF_1=95.43%，而 TextCNN 模型 MF_1 指标值比传统模型高 2.45%~8.58%。浅层神经网络方面，BPNN 分类精度高于传统分类模型，但低于 TextCNN 模型，且模型神经元节点较多，训练时间略长于传统分类模型。整体而言，TextCNN 模型短文本分类能力优于传统文本分类模型与浅层神经网络模型。

表 5 其他分类模型对比

Table 5 Comparison of other classification models

分类模型	向量化表示模型	MF_1
SVM	TF-IDF	0.901 6
	LDA	0.856 1
	LSI	0.916 3
kNN	TF-IDF	0.881 4
	LDA	0.877 6
	LSI	0.895 7
CART	TF-IDF	0.944 3
	LDA	0.901 8
	LSI	0.940 8
NaiveBayes	TF-IDF	0.806 4
	LDA	0.787 1
	LSI	0.817 7
BPNN	word2vec	0.957 5
TextCNN	word2vec	**0.965 3**

5 结论

本文考虑到智能录波器人工配置效率低，误差较高的问题，提出了基于 TextCNN 深度语义学习的智能录波器信息自配置方法，结论如下：

1) 利用 word2vec 中的 CBOW 模型针对配置信息描述文本的高维稀疏 One-hot 矢量矩阵进行低维稠密处理，有效解决了稀疏词向量忽略单词共现关系的问题，减小了运算复杂度。

2) 采取基于字符级词向量表示的 TextCNN 模型，进行多区域局部语义特征提取，有效挖掘了文本间语义衔接关系，进一步提升了关键语义筛选能力，实验证明，验证集文本的分类映射宏评估综合指标 MF_1 可达 95%以上。

3) 为提高模型泛化性能，本文在模型中引入了 dropout 层，模型训练时随机舍弃部分输出层神经元，避免最终模型出现过拟合情况。同时本文设置了拥有不同卷积核尺寸、数目以及卷积层深度的模型对照组针对样本进行配置实验，以寻找最优网络结构。实验证明，当各结构参数调至合适值时，验证集 MF_1 可提高至 96.53%。模型在验证集上分类时间低于 0.5 s，有效减少了人工配置的工作量，提高了录波器自配置精度。

参考文献

[1] 刘云鹏, 许自强, 李刚, 等. 人工智能驱动的数据分析技术在电力变压器状态检修中的应用综述[J]. 高电压技术, 2019, 45(2): 337-348.
LIU Yunpeng, XU Ziqiang, LI Gang, et al. Review on applications of artificial intelligence driven data analysis technology in condition based maintenance of power transformers[J]. High Voltage Engineering, 2019, 45(2): 337-348.

[2] 曾小芹. 基于 Python 的中文结巴分词技术实现[J]. 信息与电脑, 2019, 31(18): 38-39, 42.
ZENG Xiaoqin. Technology implementation of Chinese Jieba segmentation based on Python[J]. China Computer & Communication, 2019, 31(18): 38-39, 42.

[3] 杜修明, 秦佳峰, 郭诗瑶, 等. 电力设备典型故障案例的文本挖掘[J]. 高电压技术, 2018, 44(4): 1078-1084.
DU Xiuming, QIN Jiafeng, GUO Shiyao, et al. Text mining of typical defects in power equipment[J]. High Voltage Engineering, 2018, 44(4): 1078-1084.

[4] 李学相. 改进的最大熵权值算法在文本分类中的应用[J]. 计算机科学, 2012, 39(6): 210-212.
LI Xuexiang. Research of text categorization based on improved maximum entropy algorithm[J]. Computer

Science, 2012, 39(6): 210-212.

[5] 王进，金理雄，孙开伟. 基于演化超网络的中文文本分类方法[J]. 江苏大学学报：自然科学版, 2013, 34(2): 196-201.
WANG Jing, JIN Lixiong, SUN Kaiwei. Chinese text categorization based on evolutionary hypernetwork[J]. Journal of Jiangsu University: Natural Science Edition, 2013, 34(2): 196-201.

[6] 朱云霞. 结合聚类思想神经网络文本分类技术研究[J]. 计算机应用研究, 2012, 29(1): 155-157.
ZHU Yunxia. Text classification algorithm research based on clustering and neural network[J]. Application Research of Computers, 2012, 29(1): 155-157.

[7] TSANG S, KAO B, YIP K Y, et al. Decision trees for uncertain data[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(1): 64-78.

[8] 周志华. 机器学习[M]. 北京：清华大学出版社, 2016.

[9] ZHENG Zijian, WEBB G I. Lazy Learning of Bayesian rules[J]. Machine Learning, 2000, 41(1): 53-84.

[10] SCUTARI M, VITOLO C, TUCKER A. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation[J]. Statistics and Computing, 2019, 29(5): 1095-1108.

[11] 李航. 统计学习方法[M]. 北京：清华大学出版社, 2012.

[12] MURUGAN A, NAIR S H, KUMAR K P S. Detection of skin cancer using SVM, random forest and kNN classifiers[J]. Journal of Medical Systems, 2019, 43(8): 1-9.

[13] 蒋逸雯，李黎，李智威，等. 基于深度语义学习的电力变压器运维文本信息挖掘方法[J]. 中国电机工程学报, 2019, 39(14): 4162-4172.
JIANG Yiwen, LI Li, LI Zhiwei, et al. An information mining method of power transformer operation and maintenance texts based on deep semantic learning[J]. Proceedings of the CSEE, 2019, 39(14): 4162-4172.

[14] XIE Yinghong, SHEN Jie, WU Chengdong. Affine geometrical region CNN for object tracking[J]. IEEE Access, 2020, 8: 68638-68648.

[15] CHU Q, LIU G, ZHU X. Visualization feature and CNN based homology classification of malicious code[J]. Chinese Journal of Electronics, 2020, 29(1): 154-160.

[16] KOHLI A, GUPTA A, SINGHAL D. CNN based localisation of forged region in object-based forgery for HD videos[J]. IET Image Processing, 2020, 14(5): 947-958.

[17] 刘佳翰，陈克绪，马建，等. 基于卷积神经网络和随机森林的三相电压暂降分类[J]. 电力系统保护与控制, 2019, 47(20): 112-118.
LIU Jiahan, CHEN Kexu, MA Jian, et al. Classification of three-phase voltage dips based on CNN and random forest[J]. Power System Protection and Control, 2019, 47(20): 112-118.

[18] 吴迪，汤小兵，李鹏，等. 基于深度神经网络的变电站继电保护装置状态监测技术[J]. 电力系统保护与控制, 2020, 48(5): 81-85.
WU Di, TANG Xiaobing, LI Peng, et al. State monitoring technology of substation relay protection device based on deep neural network[J]. Power System Protection and Control, 2020, 48(5): 81-85.

[19] 王维博，张斌，曾文入，等. 基于特征融合一维卷积神经网络的电能质量扰动分类[J]. 电力系统保护与控制, 2020, 48(6): 53-60.
WANG Weibo, ZHANG Bin, ZENG Wenru, et al. Power quality disturbance classification of one-dimensional convolutional neural networks based on feature fusion[J]. Power System Protection and Control, 2020, 48(6): 53-60.

[20] HU Zhuangli, HE Tong, ZENG Yihui, et al. Fast image recognition of transmission tower based on big data[J]. Protection and Control of Modern Power Systems, 2018, 3(2): 149-158. DOI: 10.1186/s41601-018-0088-y.

[21] 高国骥. 基于跨语言分布式表示的跨语言文本分类[D]. 哈尔滨：哈尔滨工业大学, 2018.
GAO Guoji. Crosslingual Text classification based on grosslingual distributed representation[D]. Harbin: Harbin Institute of Technology, 2018.

[22] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C] // Proceedings of the 1st International Conference on Learning Representations, May 4, 2013, Scottsdale, Arizona, USA: 1-13.

[23] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C] // Proceedings of the 26th International Conference on Neural Information Processing Systems, Feb 15-16, 2013, Lake Tahoe, Nevada, USA: 3111-3119.

[24] LONG M A, ZHANG Yanqing. Using word2vec to process big text data[C] // 2015 IEEE International Conference on Big Data, October 29-November 1, 2015, Santa Clara, CA, USA: 2895-2897.

收稿日期：2020-04-08；　修回日期：2020-04-22

作者简介：

陈　旭(1976—)，男，博士，高级工程师，研究方向为电力系统规划研究与管理；E-mail: 473951841@qq.com

周达明(1997—)，男，通信作者，硕士，工程师，研究方向为智能电网数据挖掘。E-mail: 2941894638@qq.com

（编辑　葛艳娜）