北京大学学报(自然科学版) 第 58 卷 第 1 期 2022 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 58, No. 1 (Jan. 2022)
doi: 10.13209/j.0479-8023.2021.104

基于类别混合嵌入的电力文本层次化分类方法

陈晓娜 高鹏飞 梁越 马应龙†

华北电力大学控制与计算机工程学院, 北京 102206; † 通信作者, E-mail: yinglongma@ncepu.edu.cn

摘要 针对当前电力文本分类方法中因忽视类别标签之间潜在语义关联关系而导致分类性能低效的问题, 提出一种基于层次化分类模型的电力文本分类方法。首先, 利用采集的电力成果非结构化文档, 采用自动化信息提取技术和标注技术, 构建电力文本多标签分类训练集, 并结合领域知识分析, 构建类别标签之间的层次化关系。然后, 提出基于类别结构和标签语义混合嵌入的文本分类模型 HONLSTM-BERT, 利用类别标签之间的层次化结构关系进行自顶向下的层次化文本分类。最后, 通过实验与当前流行的文本分类模型进行对比分析, 结果表明 HONLSTM-BERT 方法具有更好的分类准确率, 可有效地提高电力文本自动分类性能。

关键词 电力信息技术; 电力文本分类; 层次化文本分类; 类别嵌入

A Category Hybrid Embedding Based Approach for Power Text Hierarchical Classification

CHEN Xiaona, GAO Pengfei, LIANG Yue, MA Yinglong†

School of Control and Computer Engineering, North China Electric Power University, Beijing 102206;
† Corresponding author, E-mail: yinglongma@ncepu.edu.cn

Abstract Aiming at the problem that the current power text classification methods ignore the latent semantic association between category labels and therefore lead to low classification performance, a hierarchical multi-label power text classification method is proposed. Firstly, a power multi-label text dataset is built using automatic information extraction based on power unstructured texts, and the hierarchical structural relationships between categories are constructed by leveraging relevant domain knowledge. Secondly, a text classification method HONLSTM-BERT is proposed based on hybrid embeddings of category structure and label semantics for hierarchically classifying power texts in a top-down manner. At last, experiments were made in comparison with some popular text classification methods, and the experimental results show that proposed HONLSTM-BERT method achieves superior classification accuracy, and can efficiently improve the performance of automatic text classification.

Key words power information technology; power text classification; hierarchical text classification; category embedding

随着智能电网的快速发展, 与人工智能、物联网、信息、通信和安全相关的各种技术广泛地应用于电力系统发电、输电、变电、配电、用户服务和电网调度等环节。创新驱动的企业发展战略也迫切要求在构建全球能源互联网和泛在电力物联网等新一代电力基础设施过程中能够采用创新性技术, 以便满足电力业务日益增长的发展需求和技术革新要求[1−2]。这对电力企业技术创新决策和管理提出更高的要求。企业管理人员需要深入了解各种电力信息技术在企业运营、维护和管理等方面的实际应用现状, 有效地维护电力信息技术资源文档库, 确保能够快速定位所需要的电力信息技术文档, 同

国家重点研发计划课题(2018YFC0831404)资助
收稿日期: 2021−05−31; 修回日期: 2021−08−14

时能够及时地分析当前电力信息技术分类, 预测面向电力领域未来发展的各种信息技术, 从而实现合理有效的电力新项目立项和新技术投资, 同时为电力系统的良性创新发展奠定基础[3]。

通过对电力信息技术现有研究成果的深入分析, 可以为准确地评估当前电力系统智能化应用水平提供有效的决策支持依据。典型的研究成果是目前已公开发表的与电力信息化和智能化研究紧密相关的学术论文, 可以通过互联网检索获得。然而, 这些已有成果大多是半结构化或非结构化的文档, 对企业管理人员来说, 用人工方式判别成果涉及的电力信息化智能化具体信息技术及其所解决的问题非常耗时耗力。采用人工智能和自然语言处理相关技术, 对采集的成果文档进行自动化信息抽取, 并构建相关模型训练集, 可以将针对成果的信息技术和电力应用问题研判归结为基于模型训练集的多标签文本分类问题。

目前已有一些研究将文本分类技术用于电力文本数据处理, 如针对电力设备缺陷预测的基于卷积神经网络的缺陷文本分类模型[4], 利用两阶段短文本分类方法提升审计问题分类性能[5], 以及基于支持向量机(SVM)[6]和双向长短期神经网络(BiLSTM)[7]等模型的电力系统应用。但是, 现有的电力领域使用的文本分类大都假设文本类别标签是平面化的(flattened), 忽略文本标签之间真实存在的潜在依赖关系。例如, 信息技术在类别上也存在类似依赖关系, 如可将“深度学习”技术看成“人工智能”技术的子类别。在对电力 ICT 系统进行故障分类[8]时考虑到类别嵌入, 但未考虑类别在层次结构中的位置对层次化分类的影响。在多标签分类过程中, 如果能充分利用潜在的标签依赖关系, 将其作为背景知识用于文本分类模型训练, 将有助于提升文本多标签分类性能。

针对电力文本的多标签分类问题, 本文提出基于层次化分类模型的电力文本分类方法, 首先基于有序神经元长短期记忆模型, 构建基于类别结构和标签语义混合嵌入的层次化分类模型 HONLSTM-BERT, 采用 BERT 模型进行类别语义的词嵌入。然后, 利用类别标签之间的层次化关系结构, 进行自顶向下的层次化多标签文本分类。最后, 通过实验与当前流行的文本分类基准算法进行对比分析, 验证本文方法的文本分类性能。

1 基于类别嵌入的层次化分类方法

1.1 层次化文本分类现状

层次化分类算法分为局部算法和全局算法。全局算法仅训练一个分类器, 无法全面地考虑到层次结构具有的一些局部信息。局部算法以逐层或逐节点的方式训练多个分类器, 能够全面地考虑到逐层或逐节点内部的标签信息。目前最新的层次化分类技术研究中, HDLTex 模型[9]是一种局部算法, 该算法为每个父节点训练一个分类器。此外, 还有基于集成学习 BR-GBDT 模型的电力文本多标签分类方法[10]、通过捕捉文本的内部图结构来提升分类性能的分类方法[11]以及面向蛋白质功能分类的方法[12]。层次化分类方法也开始应用于电力 ICT 故障类型识别[8], 但因该方法未考虑分类类别的具体图结构语义, 只对类别标签语义进行编码, 受领域分级和分类知识结构的影响较大。

与电力 ICT 故障类型识别研究[8]明显不同, 本文提出的基于层次化分类模型的电力文本分类方法 HONLSTM-BERT 采用具有更强文本特征表示能力的有序神经元长短期记忆模型。在分类模型构建过程中, 不仅考虑上一层类别标签的语义, 而且充分考虑类别在图结构中的结构化语义, 采用类别语义的混合嵌入编码方式进行电力文本的特征提取。

1.2 类别混合编码

1.2.1 层次化分类方法

层次化分类问题最大的特点是, 在进行第 l 层分类时, 必须考虑 l 层之前的分类情况。为了构建层次化结构的智能电网技术(标签)体系, 采用层次化类别混合嵌入的方法解决标签的层次化关系问题, 即将样本的文本特征与上一层的语义标签和类别标签进行拼接, 所得结果作为模型当前层的输入:

$$d_l = \mathrm{pos}\left(p_{l-1}\right) \oplus \mathrm{wse}\left(p_{l-1}\right) \oplus t\,, \tag{1}$$

p_{l-1} 为第 l–1 层的分类结果, t 为文本向量, ⊕表示向量的拼接(concatenation)运算。$\mathrm{pos}(p_{l-1})$表示类别嵌入, 用于获取标签位置, 即针对标签类别 p_{l-1} 可以获得一个类别所在位置为 1, 其余位置为 0 的 n_{l-1} 维的向量, n_{l-1} 为第 l–1 层的类别个数。$\mathrm{wse}(p_{l-1})$表示语义嵌入, 用于获取类别标签语义: 返回类别 p_{l-1} 对应的文本词嵌入向量表示。当 l=1 时, 表示第一层分类, 缺少上一层的类别嵌入, 所以 d_1=t。

最后, 使用一个全连接层和一个 softmax 层对第 l 层进行分类:

$$\begin{cases} d'_l = \tanh\left(W_{l1}d_l + b_{l1}\right), \\ p_l = \text{soft}\max\left(W_{l2}d'_l + b_{l2}\right), \end{cases} \tag{2}$$

softmax是在第l层的所有类别上进行计算，所得结果用于下一层的分类；tanh为激活函数；p_l为l层的分类结果；W为权重；b为偏置项。HONLSTM对每层分类进行上述操作，直至得到最后一层的结果。

1.2.2 HONLSTM-BERT算法

用于电力信息技术文本第l层分类的HONLSTM-BERT模型如图1所示。利用该模型对电力信息技术进行层次化多标签分类，通过有序神经元长短期记忆模型(ordered neurons long short-term memory, ONLSTM)[13]，分别对第l层和第l–1层进行特征提取，得到文本表示t，输入中的w_n表示文本的第n个词向量。根据第l–2层的分类结果p_{l-2}，经过处理获得p_{l-2}的语义标签和位置标签，然后将第l–1层得到的文本表示t与这些标签进行拼接，得到新的文本表示d_{l-1}。同样地，在第l–1层，模型将第l层的文本表示t与p_{l-1}的相应标签进行拼接，得到新的文本表示d_l。最后，使用全连接层和softmax层，基于文本表示d_l，对文本进行分类，输出第l层的类别p_l。

HONLSTM-BERT是一种局部层次化分类算法，能够进行逐层分类，并在每一层训练一个如图1所示的子模型用于该层的分类。

HONLSTM-BERT是基于ONLSTM[13]模型的层次化多标签分类的扩展模型。ONLSTM模型的主要特点是基于有序神经元，将文本序列中的语法树结构嵌入LSTM，而获得更高水平的抽象层次表示，可以在一定程度上解决长期依赖问题，提高模型的训练效率。

2 实验

2.1 数据集

对于本文提出的模型，采用3个数据集进行实验验证：两个数据集PLTD(power literature text dataset, PLTD)和WORF(work order failure record, WOFR)来自电力领域；一个通用领域数据集WOS (Web of Science)[9]。

WORF数据集已用于电力ICT故障诊断领域[8]，WOS是文本分类中常用的具有层次化类别结构类别的数据集，PLTD数据集是本文根据实际业务数据手工创建和类别标注。通过编写爬虫脚本从《电网技术》、《电机工程学报》和《电力系统自动化》等10余种与电力信息技术相关的期刊中采集11020份文本，构建电力文本训练数据集。文本中包含题目、摘要和关键词等。摘要是一种非结构

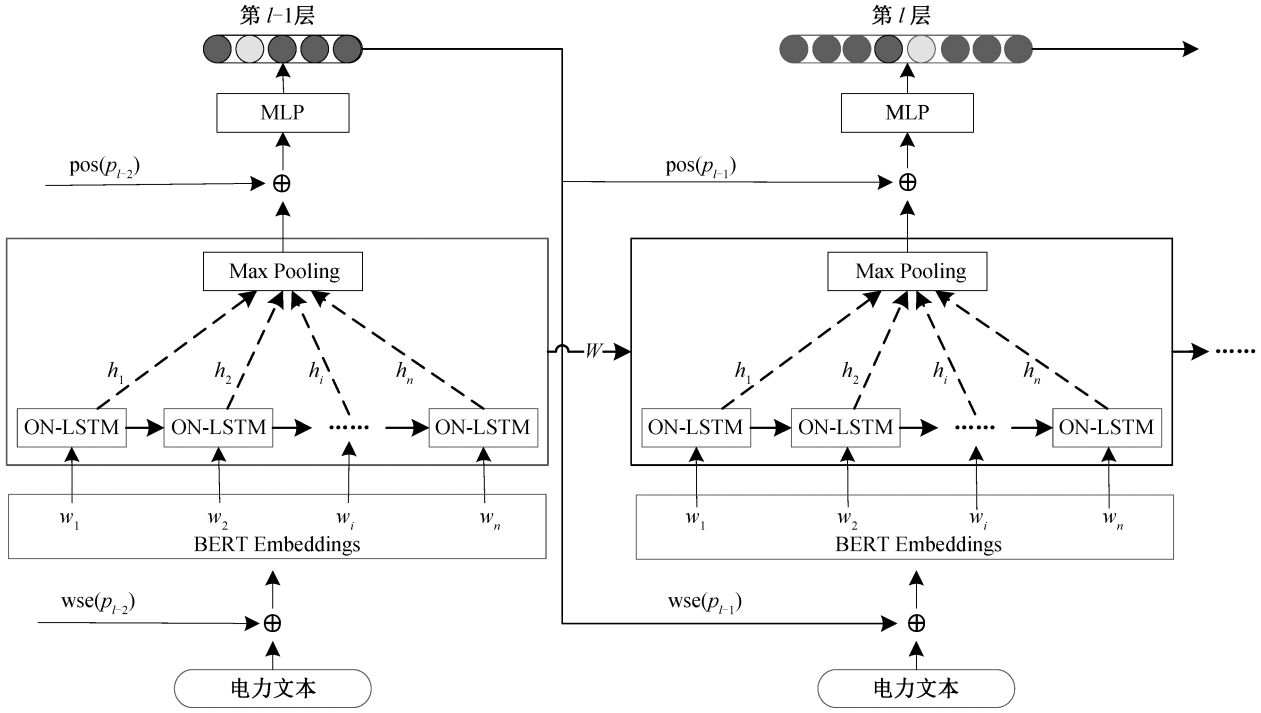

图1　电力文本第l层分类的HONLSTM-BERT模型

Fig. 1　HONLSTM-BERT model for power text classification

化文本, 不利于机器的读取与识别; 另一方面, 摘要中涉及多方面的技术, 如果以技术作为摘要的标签, 那么这个样本就是典型的多标签文本。例如, 将人工智能中的图像识别技术用于架空输电线路巡检业务[3], 该文本涉及的标签有“人工智能”、“图像识别”和“深度学习”等, 标签之间存在层次化结构关系, 即“深度学习”是“人工智能”的子标签。因此, PLDT 的样本标签采用手工标注。

由于数据集是中文文本, 需对数据集中每个样本的摘要部分进行数据清洗, 清除异常数据。采用 Jieba 分词技术对其进行分词和去停用词等操作, 完成对数据集的预处理, 用于后续的文本词嵌入。

2.2 模型训练采样优化

为了避免数据集可能存在样本不平衡问题, 优化模型训练过程, 提升模型分类性能, 本文采用 SMOTE[14]数据合成方法:

$$x_{\text{new}} = x_i + (\hat{x} - x_i) \times \delta, \tag{3}$$

其中, x_i 表示小众样本, 该方法利用样本在特征空间的相似性生成新样本, 然后从其所属小众类的 K 近邻中随机选取一个样本点 $\hat{x}$, 生成一个新的小众样本 x_{new}; $\delta \in [0, 1]$, 是一个随机数。

经 SMOTE 处理的数据集类别分布是平衡的。例如, 第 2 层有 8 个标签, 其中每个标签的样本数量占比为 8.3%~12.9%。WOFR 的处理过程与 PLTD 类似, 样本数量 2445 个, 通过该数据集所构建的标签结构有 3 层。3 个数据集的描述见表 1。

2.3 实验设置

2.3.1 分类性能指标

本文通过总体准确率和中间各层准确率两个性能指标进行算法评估。

1) 在为逐级分类提供预测的父类别时, 用总体准确率(overall accuracy, OA)评估总体分类性能:

$$\text{OA} = |C^p| \,/\, |S|, \tag{4}$$

其中, C^p 表示当父类别给定情况下, 最后一层正确分类的实例集合, S 表示用于预测的实例集合。

2) 中间各层准确率(accuracy, ACC)指在层次化结构的各层具有的分类准确率, 第 l 层的分类准确率 ACC_l 可用式(5)表示:

$$\text{ACC}_l = |C_l^t| \,/\, |S|, \tag{5}$$

其中, C_l^t 表示在提供准确的父类别的基础上, 第 l 层正确分类的实例集合。

表 1 数据集描述

Table 1 Description of datesets

数据集	标签数量			文本数量
	第 1 层	第 2 层	第 3 层	
PLTD	2	8	34	11020
WORF	2	11	35	2445
WOS	7	134	—	46985

2.3.2 对比算法

通过实验, 对比 HONLSTM 与上述各种算法的分类性能。层次化文本分类模型选择 HDLTex[9]和 HDPCNN[8]。HDLTex 是层次化文本分类中常用的算法, HDPCNN 采用 Word2vec[15]进行词嵌入, 具备更加有效的层次化电力文本分类性能[8]。平面多标签分类算法选择 TextCNN[16], BiLSTM[17] 和 BR-GBDT[10]。超参数设置如下: 用 Word2vec 进行词嵌入时生成词向量的维度是 300, 用 BERT 进行文本表示时生成向量的维度是 768, 全连接层的维度根据层类别数量设置为 256 或 512。另外, 分别基于 Word2vec[15]和 BERT[18]进行词嵌入, 与不同的文本表示模型进行消融实验, 以验证方法中各个组件的效能。

2.4 实验结果分析

2.4.1 总体准确率

表 2 显示各分类算法只考虑最后一层的分类准确率(即总体准确率), 可以看出, HONLSTM-BERT 在这 3 个数据集上总体准确率都最高。HDLTex 是为了研究层次化分类问题设计的算法, 与平面分类算法相比, 其总体分类准确率不具备明显的优势, 但 HONLSTM 算法的准确率高于平面算法。

从表 2 可以看出, HONLSTM-BERT 的分类效

表 2 平面算法和层次化算法的总体准确率(%)

Table 2 Overall accuracy on flat algorithm and hierarchical algorithm (%)

	算法	PLTD	WORF	WOS
平面算法	TextCNN	70.81	66.28	67.99
	BR-GBDT	71.22	70.37	72.98
	BiLSTM	72.79	71.16	74.23
	HONLSTM-BERT	**80.19**	**76.98**	**86.81**
层次化算法	HDLTex	74.44	70.05	77.15
	HDPCNN	77.90	72.18	81.46
	HONLSTM-BERT	**80.19**	**76.98**	**86.81**

说明: 粗体数字表示性能最佳, 下同。

果最好，这是因为HONLSTM-BERT通过对LSTM内部神经元进行特定的排序，学习了句子的句法结构，从而获得最佳的准确率。TextCNN模型难以获得文本长距离依赖关系。BR-GBDT虽然考虑了多标签的情况，但是未考虑标签之间的层次化关系。BiLSTM能够更好地捕捉双向语义依赖，所以具有较好的总体准确率。HONLSTM-BERT模型准确率最佳是因为采用有序神经元来编码文本层次结构，高阶神经元存储高层次文本信息，可长期记忆，而低阶神经元存储低层次文本信息会很快遗忘。

2.4.2 准确率

表3是层次化分类算法分别在电力文本数据集(PLTD)、WORF和WOS数据集上的分类准确率，其中l_1，l_2和l_3表示本层的分类准确率。可以看出，各层次化分类算法在PLTD数据集的l_3和WOS数据集的l_2都比对应的OA高。

HDLTex采用逐节点的方法进行层次化分类，当对第l层进行分类时，如果第$l-1$层分类错误，会导致第l层出现错误，出现错误逐层传递的问题。在WORF数据集上HDPCNN的l_1比HONLSTM-BERT高，但是HONLSTM-BERT的l_3更高，分类性能更加稳定，能有效地抽取长距离的文本依赖关系。HONLSTM-BERT模型在PLTD和WOS数据集上的总体准确率和各层准确率都有更好的效果，说明在分类时考虑类别之间的层次关系，采用类别语义混合嵌入的方式进行特征提取，可以有效地提升电力文本分类性能。

2.4.3 消融实验

消融实验(ablation study)通过移除或替换模型中的某些组件来来评估该组件的性能。本文提出的HONLSTM-BERT模型由HONLSTM和BERT两个组件构成。首先确定词嵌入方法BERT对分类准确率的影响。在文本表示模型HONLSTM不变的情况下，对Word2vec词嵌入方法(HONLSTM-W2V)与BERT词嵌入方法(HONLSTM-BERT)进行对比，结果见表4。可以看出，基于BERT进行词嵌入可以获得更高的分类准确率。同理，为了说明文本表示模型组件HONLSTM对分类性能的影响，在使用BERT词嵌入的前提下，分别利用文本表示模型HDPCNN和HONLSTM进行实验，即HDPCNN-BERT和HONLSTM-BERT。从表4看出，HONLSTM可以有效地提升文本分类准确率。

究其原因，基于BERT的词嵌入是一个句子级别的词表示，可以对一词多义建模，能够提高词向量模型的泛化能力。HDPCNN的原型是DPCNN，受卷积核窗口大小的影响较大，故HONLSTM模型的分类准确率比HDPCNN模型高。

3 结论

本文构建一个面向电力信息技术层次化类别标签体系的电力文本多标签训练集，可用于评估信息技术在泛在电力物联网领域的技术应用现状，该方法可推广到其他电力应用领域。本文提出的基于类别混合嵌入的文本分类算法HONLSTM-BERT，充

表3　层次化分类算法在数据集上各层的分类准确率(%)

Table 3　Accuracy of hierarchical classification in each layer of datasets (%)

分类器	PLTD				WORF				WOS		
	l_1	l_2	l_3	OA	l_1	l_2	l_3	OA	l_1	l_2	OA
HDLTex	87.54	85.66	79.87	74.44	86.79	83.07	70.63	70.05	91.84	85.11	77.15
HDPCNN	92.75	87.35	81.78	77.90	**92.83**	86.75	80.98	72.18	95.93	86.35	81.46
HONLSTM-BERT	**94.32**	**88.74**	**82.53**	**80.19**	91.33	**89.74**	**82.05**	**76.98**	**95.11**	**86.89**	**86.81**

表4　词嵌入方式及文本表示模型对分类准确率的影响

Table 4　Word embedding approaches and text representation models on classification accuracy

模型组件	分类器	PLTD				WORF				WOS		
		l_1	l_2	l_3	OA	l_1	l_2	l_3	OA	l_1	l_2	OA
词嵌入方法	HONLSTM-W2V	90.13	86.48	77.90	74.21	92.83	86.75	80.98	72.18	93.89	86.57	80.67
	HONLSTM-BERT	**94.32**	**88.74**	**82.53**	**80.19**	**91.33**	**89.74**	**82.05**	**76.98**	**95.11**	**86.89**	**86.81**
文本表示模型	HDPCNN-BERT	90.79	87.07	78.75	75.36	88.32	86.52	80.14	76.64	92.90	84.89	84.37
	HONLSTM-BERT	**94.32**	**88.74**	**82.53**	**80.19**	**91.33**	**89.74**	**82.05**	**76.98**	**95.11**	**86.89**	**86.81**

分考虑上一层类别标签的语义以及类别在图结构中的结构化语义，采用混合嵌入编码的方式进行电力文本的特征提取，经实验对比，验证了HONLSTM-BERT算法具有较高的分类准确率，可以作为泛在电力物联网领域信息技术现状评估和预测的有效工具。

参考文献

[1] 周东杰. 电力通信网络中的信息通信技术研究[D]. 南京: 南京大学, 2018

[2] 王顺江. 电力实时信息优化处理关键技术研究与应用[D]. 沈阳: 中国科学院沈阳计算技术研究所, 2019

[3] Cai Xinlei, Cui Yanlin, Dong Kai, et al. Framework of smart regulation system for large-scale power grid based on big data and artificial intelligence // Proceedings of International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy. Switzerland: Springer International Publishing, 2020: 440–445

[4] 刘梓权, 王慧芳, 曹靖, 等. 基于卷积神经网络的电力设备缺陷文本分类模型研究. 电网技术, 2018, 42(2): 644–651

[5] 赵雅欣, 郑明洪, 石林鑫, 等. 面向电力审计领域的两阶段短文本分类方法研究. 西南大学学报, 2020, 42(10): 1–7

[6] 汪崔洋, 江全元, 唐雅洁, 等. 基于告警信号文本挖掘的电力调度故障诊断. 电力自动化设备, 2019, 39(4): 126–132

[7] Melamud O, Goldberger J, Dagan I. Context2vec: learning generic context embedding with bidirectional LSTM // Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Stroudsburg: ACL, 2016: 51–61

[8] 李建桂, 梁越, 高鹏飞, 等. 基于层次化类别嵌入的电力ICT系统故障分类. 北京邮电大学学报, 2021, 44(4): 34–40

[9] Kowsari K, Brown D E, Heidarysafa M, et al. HDLTex: hierarchical deep learning for text classification // Proceedings of International Conference on Machine Learning and Applications. Los Alamitos: IEEE Computer Society, 2017: 364–371

[10] Xi Ziyue, Chen Xiaona, Ahmad T, et al. A novel ensemble approach to multi-label classification for electric power fault diagnosis // Proceedings of IEEE 7th International Conference on Computer Science and Network Technology. Piscataway, 2019: 278–282

[11] Peng Hao, Li Jianxin, He Yu, et al. Large-scale hierarchical text classification with recursively regularized deep graph-CNN // Proceedings of the World Wide Web Conference. Stroudsburg: ACL, 2018: 1063–1072

[12] Wehrmann J, Cerri R, Barros R. Hierarchical multi-label classification networks // Proceedings of the 35th International Conference on Machine Learning. New York: ACM, 2018, 12: 8321–8330

[13] Shen Yikang, Tan S, Sordoni A, et al. Ordered neurons: integrating tree structures into recurrent neural networks // Proceedings of the 7th International Conference on Learning Representations (ICLR' 19). La Jolla, 2019: 1–14

[14] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357

[15] Li Bofang, Drozd A, Guo Yuhe, et al. Scaling word2vec on big corpus. Journal of Data Science and Engineering, 2019, 4(2): 157–175

[16] Guo Bao, Zhang Chunxia, Liu Junmin, et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. Journal of Neurocomputing, 2019, 363: 366–374

[17] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. Journal of IEEE Transactions on Signal Processing, 1997, 45(11): 2673–2681

[18] Wang Ziniu, Huang Zhilin, Gao Jianling. Chinese text classification method based on BERT word embedding // Proceedings of the 5th International Conference on Mathematics and Artificial Intelligence. New York: ACM, 2020: 66–71