

基于 LSTM—LightGBM 模型的车站 环境温度预测

张亚伟, 陈瑞凤, 徐春婕, 杨国元, 吕晓军, 方 凯

(中国铁道科学研究院集团有限公司 电子计算技术研究所, 北京 100081)

摘要: 客运火车站环境温度易受其他环境特征变量如湿度、PM2.5、二氧化碳等影响, 传统的单变量预测算法并未考虑其他环境特征变量的影响因素; 为进一步准确预测车站环境温度值, 提出了结合长短期记忆神经网络 LSTM 与梯度提升算法 LightGBM 的组合模型, 对客运站环境温度值进行预测; 首先将预处理数据输入 LSTM 模型, 对环境特征变量湿度、二氧化碳、PM2.5、PM10 进行单变量预测; 再将环境特征变量的 LSTM 输出预测值输入 LightGBM 模型得出环境温度预测值; 根据波形图与均方根误差 RMSE 对比分析, 基于 LSTM—LightGBM 的组合模型预测方法可以保留 LSTM 模型对单变量预测的周期性特点, 且可表现出环境特征变量输入 LightGBM 模型后对温度预测的非平稳变化; 结果表明基于 LSTM—LightGBM 的组合模型方法比单纯使用 LSTM 方法更接近原始波形, 具有更低的 RMSE。

关键词: 环境温度; 单变量预测; 长短期记忆神经网络; 梯度提升算法; 环境特征变量

Station Environment Temperature Prediction Based on LSTM—LightGBM Model

ZHANG Yawei, CHEN Ruifeng, XU Chunjie, YANG Guoyuan, LÜ Xiaojun, FANG Kai

(Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: The environment temperature of passenger railway station is easily affected by other environmental characteristic variables such as humidity, PM2.5, carbon dioxide, etc. The traditional univariate prediction algorithm does not consider the influence factors of other environmental characteristic variables. In order to predict the environmental temperature of the passenger station more accurately, a combined model combining LSTM neural network and LightGBM gradient lifting algorithm is proposed to predict the environmental temperature of the passenger station. Firstly, the pre-processed data were input into the LSTM model, and environmental characteristic variables such as humidity, carbon dioxide, PM2.5 and PM10 were predicted. Then input the predicted value of environmental characteristic variables output by LSTM into LightGBM model to get the predicted value of temperature. According to the comparison and analysis of the waveforms and RMSE, the combined model prediction based on LSTM—LightGBM can retain the periodicity of the univariate prediction used by LSTM model, and can show the non-stationary changes of the temperature prediction after the environmental characteristic variables input into LightGBM model. The results show that the combined model method based on LSTM—LightGBM is closer to the original waveform and lower RMSE than the method using LSTM alone.

Keywords: the environment temperature; univariate prediction; LSTM; gradient lifting algorithm; environmental characteristic variable

收稿日期: 2021-05-31; 修回日期: 2021-07-20。

基金项目: 国家自然科学基金(62001519); 国家重点研发计划(2020YFF0304100)。

作者简介: 张亚伟(1989-), 男, 北京人, 硕士, 助理研究员, 主要从事物联网方向的研究。

通讯作者: 陈瑞凤(1989-), 女, 北京人, 副研究员, 主要从事物联网方向的研究。

引用格式: 张亚伟, 陈瑞凤, 徐春婕, 等. 基于 LSTM—LightGBM 模型的车站环境温度预测[J]. 计算机测量与控制, 2022, 30(1): 20-25.

0 引言

智能高铁已成为目前铁路运输领域的发展方向, 推进智能车站的建设是构建智能高铁的重要组成部分。铁路客运站作为城市间重要的桥梁, 智能化体验、舒适旅行也成为了人们所追求的目标。某些大型客运站人流进出量大, 容易出现人流密集、拥挤的情况, 而其内部的空气环境直接影响着旅客舒适度的体验, 尤其是温度成为了环境指数中重要的物理量。站内风水系统的调节直接受环境温度影响, 随着环境温度高低的变化, 风水系统可提前调节为合适的大小, 从而达到最佳的旅客感知体验。站内温度受多方面环境因素影响, 包括湿度、二氧化碳 (CO₂) 浓度、PM2.5、PM10 等参数, 本文基于站内环境特征变量参数, 结合模型算法重点研究站内环境温度的预测问题。

基于数学建模的预测方法种类繁多, 如单耗法、弹性系数法、统计分析法、灰色预测法等^[1]。针对不同预测内容的特点会选择相应的预测方法建立预测模型。对于中短期预测, 近似于指数增长的预测, 可选用灰色预测模型。对于时间序列问题, 传统方法有指数平滑法、自回归积分滑动平均 ARIMA 模型算法, LSTM 模型算法等^[2], 这些方法对具有周期性规律的数据预测有较好的效果。对于多因子影响的预测问题, 可用 BP、GBDT 等机器学习算法^[3]。文献 [2-3] 采用分位数回归法、改变学习速率法对 LSTM 模型进行了改进, 并应用于用电负荷与股价趋势预测。文献 [4] 采用时间序列分析 ARIMA 模型对局部地区的季节性气温进行了预测。另外, 基于 GBDT 模型改进的梯度提升算法如 XGBoost、LightGBM 模型, 具有训练速度快、占用内存小的特点, 也得到了相关预测的应用。车站环境温度值有两个特点: (1) 一种时间序列, 即数据会随时间发展而展现出一种规律性, 受季节与白天黑夜的影响呈现一种周期性波动; (2) 规律之中又有其特殊性, 如会受湿度值、PM2.5、CO₂ 等环境因素的影响而发生波动, 对于车站人员密集场所, 这些因素的影响会更大, 单独使用上述算法缺乏对车站环境温度的整体预测感知能力。

针对车站环境温度的特点, 本文将 LSTM 与 LightGBM 模型进行组合, 充分利用车站环境特征变量对站内环境温度进行预测。该组合模型能够结合 LSTM 与 LightGBM 模型各自的特点, 既可挖掘长时数据之间的内在关系, 又可避免非连续性以及波动

数据对预测精度的不良影响^[4]。

1 LSTM 环境预测模型算法

LSTM 是一种特殊的 RNN 网络, RNN 会存在梯度消失和梯度爆炸的问题, LSTM 通过 cell 门开关实现时间上的记忆功能, 并防止梯度消失, 可以学习长期依赖信息, 让信息长期保存, 解决 RNN 的缺陷问题^[5], 其算法结构如图 1。LSTM 的当前输入 x^t 和上一个状态传递下来的 h^{t-1} 拼接训练可以得到四个状态, 状态公式如下:

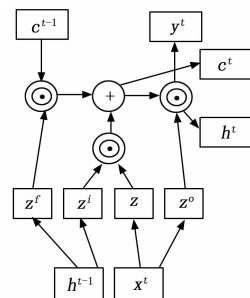


图 1 LSTM 算法结构图

$$z = \tanh(w_h^f h^{t-1}) \quad (1)$$

$$z^i = \sigma(w_h^i h^{t-1}) \quad (2)$$

$$z^f = \sigma(w_h^f h^{t-1}) \quad (3)$$

$$z^o = \sigma(w_h^o h^{t-1}) \quad (4)$$

式 (2) ~ (4) 的 z^i, z^f, z^o 是由拼接向量乘以权重矩阵之后, 再通过一个 sigmoid 激活函数转换成 0 到 1 之间的数值, 作为一种门控状态。公式 (1) 中 z 是将结果通过 $\tanh$ 激活函数转换成 -1 到 1 之间的值。 $\odot$ 是操作矩阵中对应的元素相乘, 要求两个相乘矩阵是同型的^[6]。 $\oplus$ 代表矩阵加法。 c^t, h^t, y^t 计算公式如下:

$$c^t = z^f \odot c^{t-1} + z^i \odot z \quad (5)$$

$$h^t = z^o \odot \tanh(c^t) \quad (6)$$

$$y^t = \sigma(W^f h^t) \quad (7)$$

LSTM 模型计算过程主要分为三个阶段: 忘记阶段、输入阶段、输出阶段。忘记阶段即“忘记”之前没用的信息。遗忘门会根据当前时刻节点的输入 x^t 、上一时刻节点的状态 c^{t-1} 和上一时刻节点的输出 h^{t-1} 来决定哪些信息将被遗忘^[7]。输入阶段决定当前输入数据哪些信息被留下来。主要对输入 x^t 进行选择记忆。当前输入内容由公式 (1) z 得出, 选择门控信号由 z^i 来进行控制。由公式 (1)、(2)、(3) 可得到传输给下一个状态的 c^t , 即公式 (5)。输出阶段

确定输出值。LSTM 在得到最新节点状态 c' 后, 结合上一时刻节点输出 h^{t-1} 和当前时刻节点的输入 x' 来决定当前时刻节点的输出 y' , 通过公式 (6) 的 h' 变化得到公式 (7)。其中公式 (6) 中 z'' 来进行控制, 并对状态 c' 进行了缩放 (通过 tanh 激活函数进行变化)^[8]。

2 LightGBM 模型算法

LightGBM 是基于决策树算法的分布式梯度提升 GBDT 框架, 在大训练样本和高维度特征的数据环境下, GBDT 算法的性能以及准确性会面临极大的挑战。为了解决此问题, LightGBM 应运而生。LightGBM 具有训练速度快、内存占用少、准确率高、支持并行化学习、可处理大规模数据的特点^[9]。

LightGBM 在梯度算法中主要采用了一些优化算法:

1) 单边梯度采样算法 (GOSS): LightGBM 使用 GOSS 算法进行训练样本采样的优化。GOSS 算法的基本思想是首先对训练集数据根据梯度排序, 预设一个比例, 保留在所有样本中梯度高于比例的数据样本; 梯度低于该比例的数据样本不会直接丢弃, 而是设置一个采样比例, 从梯度较小的样本中按比例抽取样本。为了弥补对样本分布造成的影响, GOSS 算法在计算信息增益时, 会对较小梯度的数据集乘以一个系数用来放大。在计算信息增益时, 算法可以更加关注“未被充分训练”的样本数据。

2) EFB 算法 (Exclusive Feature Bundling): LightGBM 算法不仅通过 GOSS 算法对训练样本进行采样优化, 也进行了特征抽取, 以进一步优化模型的训练速度。EFB 算法可以将数据集中互斥的特征绑定在一起, 形成低维的特征集合, 能够有效避免对 0 值特征的计算。在算法中, 可以对每个特征建立一个记录非零值特征的表格。通过对表中数据的扫描, 可以有效降低创建直方图的时间复杂度。

3) 直方图算法 (Histogram): LightGBM 采用了基于直方图的算法, 将连续的特征值离散化成了 K 个整数, 构造宽度为 K 的直方图, 遍历训练数据, 统计每个离散值在直方图中的累积统计量。在选取特征的分裂点时, 只需要遍历排序直方图的离散值。使用直方图算法降低了算法的计算代价, 同时降低了算法的内存消耗。

4) 按叶子生长 (leaf-wise) 算法: 大多数决策树学习算法的树生成方式都采用按层生长 (level-

wise) 的策略^[10], 如图 2 所示。

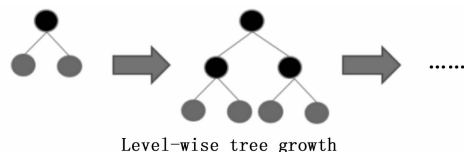


图 2 level-wise 策略图

LightGBM 采用一种更为高效的按叶子生长 (leaf-wise) 策略算法, 如图 3 所示。该策略每次从当前决策树所有的叶子节点中, 找到分裂增益最大的一个叶子节点进行分裂, 如此循环往复。该机制减少了对增益较低的叶子节点的分裂计算。与 level-wise 策略相比, 在分裂次数相同的情况下, leaf-wise 可以降低误差, 得到更好的精度。Leaf-wise 算法的缺点是可能会生成较深的决策树。因此, LightGBM 在 leaf-wise 上增加了限制最大深度的参数, 在保证算法高效的同时, 防止过拟合^[11]。

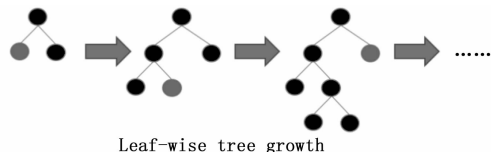


图 3 leaf-wise 策略图

3 基于 LSTM-LightGBM 组合模型预测算法

基于 LSTM-LightGBM 组合模型对温度进行预测的流程图如下, 其中环境特征变量为湿度、二氧化碳、PM2.5、PM10, 环境传感器数据除上述 4 种环境特征变量外还包含温度值。

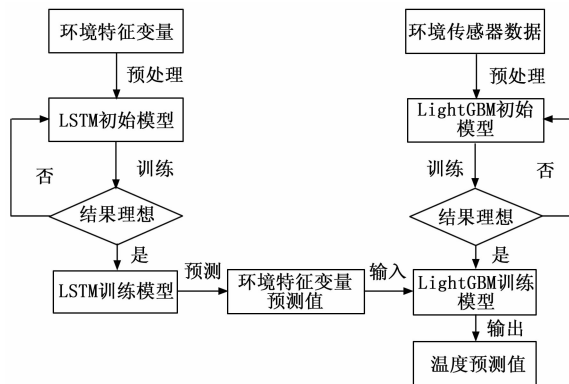


图 4 LSTM-LightGBM 算法流程图

3.1 数据预处理

数据来源为京张线清河站环境传感器获取的两日内 145 条数据, 包含温度 (Temperature)、湿度 (humidity)、二氧化碳 (CO_2)、PM2.5、PM10 数

据。表 1 所示为采集数据的前 5 行:

表 1 环境传感器采集数据

time	Humidity /%	CO ₂ /PPM	PM2.5/ (μg/m ³)	PM10/ (μg/m ³)	Tempera- ture/℃
2021/1/24 0:00	57	506	19	373	7
2021/1/24 0:05	37	521	20	412	5.2
2021/1/24 0:10	36	524	20	424	5.1
2021/1/24 0:15	38	549	18	353	3.7
2021/1/24 0:20	56	522	18	350	7.2

由于输入、输出变量值差异较大, 先做归一化处理, 以减小所得结果误差^[12]。采用 min-max 标准化方法, 将数据映射到 [0, 1] 之间, 如公式 (10) 所示:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

其中: $x_{\max}$ 为样本数据的最大值, $x_{\min}$ 为样本数据的最小值。归一化后的数据前五五行如表 2 所示。

表 2 归一化样本

humidity	CO ₂	PM2.5	PM10	Temperature
0.78	0.05	0.67	0.33	0.91
0.16	0.32	1.00	0.79	0.49
0.13	0.38	1.00	0.93	0.47
0.19	0.82	0.33	0.09	0.14
0.75	0.34	0.33	0.06	0.95

3.2 构造模型及训练

3.2.1 构造 LSTM 训练模型

首先将输入数据整理成 LSTM 所需的三维结构, 表示为 (TrainX, SeqLen, Dim_in) 第一个维度 TrainX 表示对应样本, 第二个维度 SeqLen 表示该样本所采集的序列数据 (指定序列长度), 第三个维度 Dim_in 表示对应的特征维度, 本文数据集环境特征变量为 4 个, 因此 Dim_in 取 4。按照此三维结构形式将数据集划分为训练集和测试集^[13]。

第二, 构建 LSTM 模型。模型结构表示为 (Units, Input_Shape, Activation, Recurrent_dropout)。Units 为隐含层神经元个数, Input_Shape 为输入数据集的结构形式, Activation 为激活函数, Recurrent_dropout 为学习率。分析数据集样本大小和特点, 得到最佳模型取值。本文 Units 取 45, Activation 取 “relu”, Recurrent_dropout 为 0.01。

第三, 训练 LSTM 模型。训练模型结构为 (X_Train, Y_Train, Epochs, Batch_Size, Validation

_Split)。X_Train, Y_Train 为模型训练数据; Epochs 为迭代次数; Batch_Size 为批处理样本数; Validation_Split 为训练验证集分割比例。训练模型参数的取值是基于先验知识与大量实验得出, 本文 Epochs 取值为 100, Batch_Size 为 16, Validation_Split 为 0.8^[14]。

3.2.2 构造 LightGBM 训练模型

将数据集划分为训练集和验证集, 设置训练集和验证集比例系数为 0.8。模型的参数设置是基于先验知识以及大量的实验, 针对所用实验数据, LightGBM 训练模型的核心参数设置如下^[15]:

objective: 任务类型。可选任务类型为 regression (回归)、binary (二分类)、multiclass (多分类) 等。本文任务是进行预测, 设置为 regression。

num_leaves: 叶节点的数目。该参数决定树模型的复杂度, 越大会越准确, 但可能过拟合, 该参数设置为 120。

max_depth: 控制了树的最大深度。该参数可以显式的限制树的深度。一般设置为不大于 log2 (num_leaves) 的值, 本文设置为 7。

min_data_in_leaf: 每个叶节点的最少样本数量。它是处理 leaf-wise 树的过拟合的重要参数。将它设为较大的值, 可以避免生成一个过深的树, 但是也可能导致欠拟合。本文设置为 16。

learning_rate: 训练模型的学习率。较大的学习率会加快收敛速度, 但是会降低准确率, 默认为 0.1。针对数据集的大小可调整学习率, 本文设置为 0.05。

num_boost_round: 迭代次数。设置为 1 000^[16]。

3.2.3 LSTM-LightGBM 组合模型预测

将环境特征变量湿度、二氧化碳、PM2.5、PM10 数据输入 LSTM 训练模型, 可计算出每个环境特征变量的预测值。该多维预测值作为输入变量, 输入 LightGBM 模型, 由此可得出温度预测值。

3.2.4 实验评价标准

误差标准采用均方根误差 (RMSE) 公式^[7]。RMSE 计算公式如下: 其中 y_i 为所采集的每个数据, $\hat{y}_i$ 为预测值, m 为样本数^[17]。

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \tag{9}$$

4 结果分析

选取环境特征变量 24 条数据 (即 2 小时) 输入

LSTM 模型，所得预测结果如图 5，四幅图为每个环境特征变量的预测结果。从图中可看出 LSTM 模型对于有突变的尖峰预测灵敏度较差，但是可以将数据的整体上升下降趋势预测出来^[18]。表 3 为环境特征变量预测值与真实值的 RMSE。

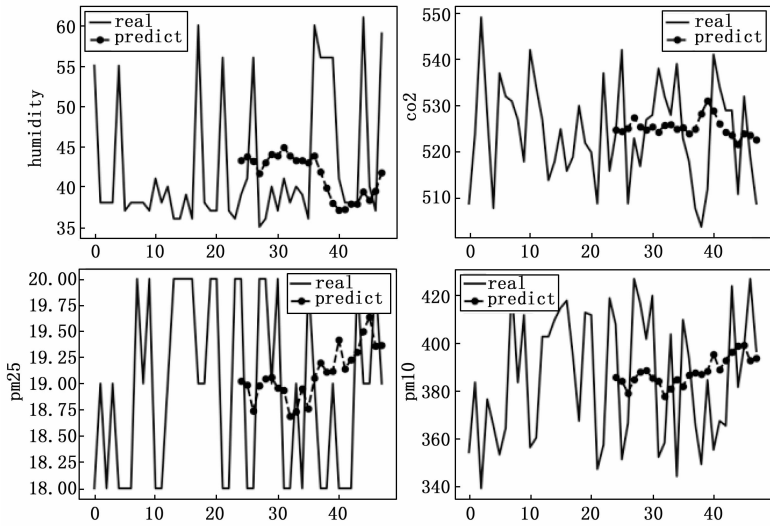


图 5 LSTM 环境特征变量预测结果

表 3 环境特征变量预测 RMSE

	humidity	CO ₂	PM2.5	PM10
RMSE	9.81	11.23	0.88	26.39

将 LSTM 模型已预测的湿度、二氧化碳、PM2.5、PM10 数据输入 LightGBM 训练模型，所得温度预测值如图 6。两模型预测结果 RMSE 分别为 0.82 和 1.35。从图中可看出，LSTM—LightGBM 组合模型比 LSTM 模型结果更接近原始波形，

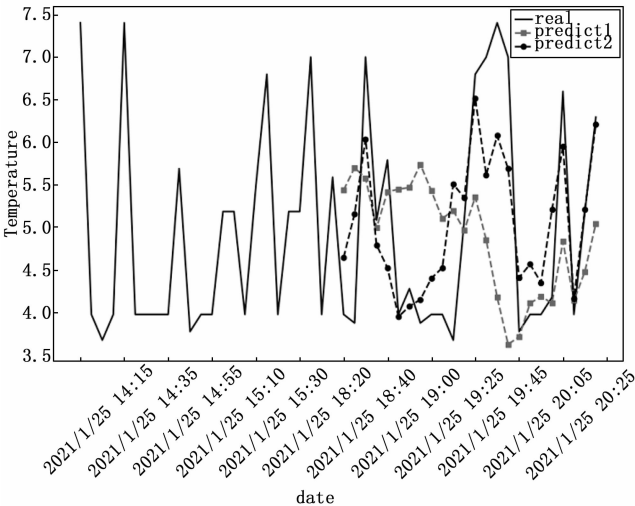


图 6 LSTM—LightGBM 温度预测结果

LSTM—LightGBM 组合模型对波形突变有更好的响应，能反应出环境特征变量所引起的温度变化。

本文利用 LightGBM 自带的输入变量重要性判别函数对环境特征变量的重要性做了对比，结果如表 4 与图 7。表 5 对环境特征变量的重要性用数值进行量化，重要性代表特征值在训练过程中使用的平均次数^[19]，图 7 以折线图形式呈现。可以看出，二氧化碳和 PM10 对预测结果有较大影响。

表 4 环境特征变量重要性比对

环境特征	重要性
humidity	132
CO ₂	268
PM2.5	29
PM10	287

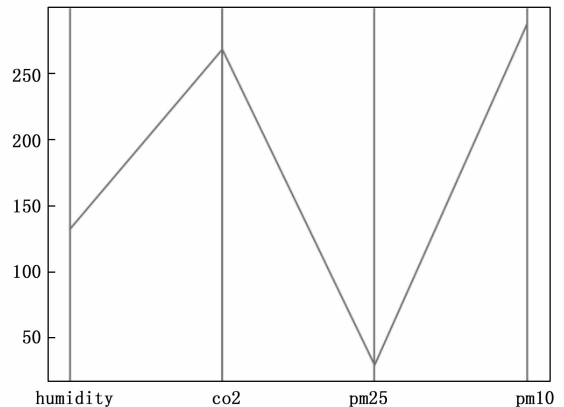


图 7 环境特征变量重要性比折线图

5 结束语

本文基于客运火车站重点区域的环境传感器实际数据，采用预测算法对车站温度值进行预测分析。首先利用 LSTM 模型分别对湿度、二氧化碳、PM2.5、PM10 四项环境特征变量进行预测。再利用 LSTM 模型预测的四项环境特征变量输入 LightGBM 模型对温度值进行预测。通过对比单独使用 LSTM 模型与使用 LSTM—LightGBM 组合模型对温度的预测结果，得出 LSTM—LightGBM 组合模型的预测值有更低的 RMSE，其预测波形趋势更接近原始波形。

本文研究内容可应用于车站重点区域环境温度的预测，可为工作人员提供辅助决策手段，如通过预测值提前设定站内风水系统的空调温度值与通风量大

小,从而提升区域舒适度,减少能耗^[20]。最后,本文的算法仍有研究空间,如训练数据量大小对模型的影响,不同环境特征变量的预测对车站节能、舒适度的影响等。

参考文献:

- [1] ZHAO Z, CHEN W H, WU X M, et al. LSTM Network: A Deep Learning Approach for Short-term Traffic Forecast [J]. IET Intelligent Transport Systems, 2017, 11 (2): 68-75.
- [2] 李彬,彭曙蓉,彭君哲,等.基于深度学习分位数回归模型的风电功率概率密度预测[J].电力自动化设备,2018,38(9):15-20.
- [3] 孙瑞奇.基于LSTM神经网络的美股股指价格趋势预测模型的研究[D].北京:首都经济贸易大学,2016.
- [4] 马力,于瑞林,李杰,等.基于时间序列分析并预测三峡库区局地环境温度的变化[J].中国农业气象,2018,39(1):9-17.
- [5] 李乐.时空序列数据预处理方法研究[D].北京:中国科学院大学,2017.
- [6] 刘蓉晖,赵才涛.基于数据挖掘技术的气温敏感负荷短期预测研究[J].电网与清洁能源,2017,33(11):32-38.
- [7] YANG Y, DONG J, SUN X, et al. A CFCC-LSTM Model for Sea Surface Temperature Prediction [J]. Geoscience and Remote Sensing Letters, 2018, 15 (2): 207-211.
- [8] KE G L, QI M, THOMAS FINLEY, et al. LightGBM: a highly efficient gradient boosting decision tree [C] // Advances in Neural Information Processing Systems, 2017: 3149-3157.
- [9] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 134-142.
- [10] 赵冬红,刘文颖.基于负荷预测模型特性的组合预测新方法[J].现代电力,2008,25(6):21-26.
- [11] 陈剑强,杨俊杰,楼志斌.基于XGBoost算法的新型短期负荷预测模型研[J].电测与仪表,2019,56(21):23-29.
- [12] 焦润海,苏辰隽,林碧英,等.基于气象信息因素修正的灰色短期负荷预测模型[J].电网技术,2013,37(3):720-725.
- [13] 杨红军,王胜辉,李潇潇.基于最优加权组合模型的光伏出力预测[J].沈阳工程学院学报,2019,15(4):300-305.
- [14] 王华勇,杨超,唐华.基于LightGBM改进的GBDT短期负荷预测研究[J].自动化仪表,2018,39(9):76-78.
- [15] 熊军华,牛珂,张春歌,等.基于小波变异果蝇优化支持向量机短期负荷预测方法研究[J].电力系统保护与控制,2017,45(13):71-77.
- [16] CHARLES DUBOUT, FRANÇOIS FLEURET. Boosting with maximum adaptive sampling [C] // Advances in Neural Information Processing Systems, 2011: 1332-1340.
- [17] 许言路,张建森,吉星,等.基于多模型融合神经网络的短期负荷预测[J].控制工程,2019,26(4):619-624.
- [18] SANJAY RANKA, V SINGH. Clouds: A decision tree classifier for large datasets [C] // Proceedings of the 4th Knowledge Discovery and Data Mining Conference, 1998: 2-8.
- [19] 周挺,杨军,周强明,等.基于改进LightGBM的电力系统暂态稳定评估方法[J].电网技术,2019,43(6):1931-1940.
- [20] RON APPEL, THOMAS J FUCHS, PIOTR DOLLÁR, et al. Quickly boosting decision trees - pruning underachieving features early [C] // ICML (3), 2013: 594-602.
- [16] 程利军,张英堂,李志宁.基于阶比跟踪及共振解调的连杆轴承故障诊断研究[J].内燃机工程,2012(5):67-63.
- [17] 周宇,陈进,董广明,等.基于循环双谱二次切片分析的滚动轴承故障诊断研究[J].振动与冲击,2012(18):24-28.
- [18] 王小玲,陈进,从飞云.基于时频的频带熵方法在滚动轴承故障识别中的应用[J].振动与冲击,2012(18):29-33.
- [19] 罗红梅,齐明侠,裴峻峰,等.滚动轴承故障诊断中精确转频的实用计算新方法[J].振动与冲击,2007(5):64-66.
- [20] 肖志松,唐力伟,王虹.时域平均在行星齿轮箱故障诊断的应用[J].河北工业大学学报,2003(6):99-102.