电网技术
Power System Technology
ISSN 1000-3673,CN 11-2410/TM

《电网技术》网络首发论文

题目：基于特征工程和集成学习与模型融合的输电走廊实时山火风险评估模型
作者：张可颖，吴新桥，赵继光，刘岚，覃平，王昊，詹谭博驰
DOI：10.13335/j.1000-3673.pst.2022.2235
收稿日期：2022-11-11
网络首发日期：2023-01-17
引用格式：张可颖，吴新桥，赵继光，刘岚，覃平，王昊，詹谭博驰．基于特征工程和集成学习与模型融合的输电走廊实时山火风险评估模型[J/OL]．电网技术．https://doi.org/10.13335/j.1000-3673.pst.2022.2235

基于特征工程和集成学习与模型融合的输电走廊实时山火风险评估模型

张可颖*，吴新桥，赵继光，刘岚，覃平，王昊，詹谭博驰
（南方电网数字电网研究院有限公司，广东省 广州市 510700）

A Real-time Wildfire Risk Assessment Model for Transmission Corridors Based on Feature Engineering, Ensemble Learning and Model Fusion

ZHANG Keying, WU Xinqiao, ZHAO Jiguang, LIU Lan, QIN Ping, WANG Hao, ZHAN Tanbochi
(China Southern Power Grid Digital Grid Research Institute Co., Ltd, Guangdong Province, Guangzhou 510700, China)

ABSTRACT: Wildfire disasters are easy to cause tripping of overhead transmission lines and threaten the stable operation of power grid. In this paper, a real-time bushfire risk assessment model for transmission corridor based on feature engineering and Boosting algorithm is proposed. First of all, the original data of four categories of 20 characteristics, including human activity behavior, geography, real-time meteorology and regional historical bushfire disaster, which affect the occurrence of bushfire in the transmission corridor, are collected, extracted and cleaned. Then the quadratic polynomial is used to derive the features and generate 236 features. Considering the time complexity of model calculation, the Wrapper method combined with the five-fold cross-validation method was used to iteratively obtain the 100 features with the highest importance to construct a feature subset as the model input to construct a real-time bushfire risk assessment model of transmission corridor based on Boosting algorithm. In order to minimize the log-loss function as the optimization objective, the Bayesian optimization algorithm was used to search the parameter space and obtain the optimal model parameters. Finally, the accuracy, precision and recall of the model are verified on the test set. The model performs well with the accuracy of 96.4%, and recall of 88.1%, which can effectively evaluate the real-time risk of the transmission corridor in real time.

KEY WORDS: Wildfire; Transmission Corridors; Ensemble Learning; Feature Engineering; Risk Assessment

摘要：山火灾害易造成架空输电线路跳闸，威胁电网的稳定运行。本文提出了一种基于特征工程和 Boosting 集成学习框架与模型融合的输电走廊实时山火风险评估模型。首先，对影响输电走廊山火发生产生影响的人类活动行为、地理、实时气象、区域历史山火灾害情况 4 类 20 个特征的原始数据进行收集、提取和清洗。然后，利用二次多项式对特征进行衍生，生成 236 个特征，考虑到模型计算的时间复杂度，再使用 Wrapper 方法结合五折交叉验证法迭代获取重要度最高的 100 特征构建特征子集作为模型输入构建基于 Boosting 算法的输电走廊实时山火风险评估模型，以最小化对数损失函数为优化目标，利用贝叶斯优化算法对参数空间进行搜索，得到最优模型参数。最后，在测试集上对模型准确率、召回率和调和均值进行验证，其预测准确率达 96.4%，召回率达 88.1%，调和均值达 85.5%，在正负样本差异极大的现实场景中可以有效实时评估输电走廊的实时风险。

关键词：山火；输电走廊；集成学习；特征工程；风险评估

DOI：10.13335/j.1000-3673.pst.2022.2235

0 引言

南方电网输电线路总长度超 30 万 km，且大多位于偏远位置，所处地形复杂，运行条件严峻，容易受到各种外部因素的影响。受烧田、祭祖等人类活动的影响，这些地区的架空输电线路容易发生大规模火灾，造成输电线路跳闸，严重时甚至威胁到重要交叉跨越、密集通道、西电东送主通道，对电网安全稳定运行造成负面影响。火灾引起的架空输电线路跳闸在各类跳闸和停电事故中占相当大的比例，南方电网所辖的广东、广西、云南、贵州、海南等地每年森林火灾次数占全国 80%以上[1]。据统计，南方电网 2015—2019 年输电线路因山火灾害跳闸总数逾 300 次，其中重合闸成功 120 余次，占比仅 40%[2]。2020 年山火造成南方电网全网

基金项目：南方电网数字电网研究院有限公司科技项目（210005KK52220019）。
Project Supported by China Southern Power Grid Digital Power Grid Research Institute Co., Ltd（210005KK52220019）.

110kV 及以上输电线路跳闸接近 100 次，与 2019 年同比增加 20 余次。山火已成为引起输电线路跳闸停电的主要原因之一，严重影响电网安全稳定运行。准确评估输电线路山火跳闸风险分布，有助于进一步制定差异化防山火措施，提升电网抵御山火灾害的能力[3]。

为提高架空输电线路山火灾害防治水平，国内外学者已开展包括山火分布规律[4-7]、跳闸机理[8-11]、监测告警[12-16]和风险评估[17-21]等多维度研究，但距离电网山火风险事前预警和应急响应的精准要求还有不小差距。目前，山火风险评估建模方法主要以选取山火影响因子构建评估模型为主[1]。文献[22]提出了一种考虑降水、卫星监测火点和工业用火因素的输电线路山火预测模型，但仅仅考虑了少量的山火影响要素，且模型构建中需要依赖专家主观经验，难以全面、客观描述山火风险。文献[23]提出了一种基于 Relief 特征选择和朴素贝叶斯的输电走廊山火风险评估模型，但模型准确度仅有 80.9%，模型准确度有待进一步提高，实用性有待验证。文献[5]结合模糊层次分析法和熵权法，形成主观赋值权重和客观计算权重结合的山火评估指标组合权重，并在此基础上建立物元可拓模型，评估山火灾害风险等级，但仅在 2019 年 11 月至 2020 年 5 月 7 次跳闸案例中进行运用，模型实用性有待进一步验证。

本文首先筛选了影响架空输电走廊山火发生的人类活动行为、地理、实时气象、区域历史山火灾害情况 4 类 20 个特征，以南方电网所辖广东、广西、云南、贵州和海南五省（区）为研究对象，以 1km×1km 网格精度为标准，对 20 个特征的数据进行采集、清洗，对历史卫星山火热点样本进行标注。然后利用特征衍生技术和 Wrapper 方法结合五折交叉验证法获取重要度最高的 100 特征构建特征子集。继而提出一种基于 Boosting 算法的架空输电线路山火风险概率模型，对山火风险概率进行分级，并利用贝叶斯优化算法对模型参数空间进行搜索，得到最优模型。最后使用五折交叉验证对模型适用性进行验证。

与同领域研究相比，本文主要有以下三个重要贡献。首先，本文提出了一种基于特征工程和 Boosting 集成学习与模型融合的输电走廊实时山火风险评估模型，考虑了实时温度、降雨、风速和湿度数据，在文献[24-26]的基础上更进一步，实现了实时山火风险评估预警。经验证表明，本文提出的模型算法准确率、召回率更高。此外，本文对模型计算时间效率也进行了评估，针对高准确率和高时效性两种不同工程化场景，分别给出了不同的推荐模型。最后，本文对提出的模型采用真实正负样本差异极大的大数据集进行了严格验证。

1 输电走廊山火灾害风险指标体系

1.1 研究区域

本文选择南方电网所辖广东、广西、云南、贵州和海南五省（区）作为研究区域。

1.2 影响输电走廊山火灾害发生特征

输电线路山火灾害的致因复杂，通常是多种因素相互作用结果，主要可以划分为人为因素和自然因素，其中由人为因素引起的火灾比例高达 90%以上[27-31]。本文收集整理人类活动、地理信息、实时气象和区域历史山火灾害情况 4 类影响山火发生的特征，构建输电线路实时山火灾害风险评估模型。

1.2.1 人类活动特征

研究显示 90%以上架空输电走廊山火由人为野外用火导致，以 1-4 月冬春季烧荒、春节、清明祭祖、上坟等为主。本文选择了距铁路距离、距高速公路距离、距一级公路距离、距居民点距离、人口密度和国民生产总值 GDP（Gross National Product, GDP）6 个特征来描述人类行为活动对山火风险的影响，如表 1 所示。

表 1 人类活动特征

Tab. 1 Features of Human Activity

编号	类别
1	距铁路距离
2	距高速公路距离
3	距一级公路距离
4	距居民点距离
5	人口密度
6	人口密度和国民生产总值

1.2.2 地理信息特征

地理信息特征可分为植被特征和地地理特征。地形的不同会影响气候和植被状况，引起林火程度的不同。不同植被类型着火难易程度不同，山火灾害多发生在野外植被覆盖率高的针叶林、针阔混交林区域，可燃物风险等级和 NDVI 植被指数可以对植被分布情况进行描述。可燃物风险等级数据来源于国家气象中心，NDVI 植被指数可以反映植被生长状态、植被覆盖度，从中国科学院资源环境科学与数据中心获取。

地形地貌不仅会影响植被类型，也会影响火焰燃烧和蔓延速度，相关特征主要包括经度、纬度、高程、坡度、坡向。

因此，本文选择经度、纬度、高程、坡度、坡

向、可燃物风险等级、植被类型风险等级和 NDVI 植被指数 8 个特征描述地理特征对山火风险的影响，如表 2 所示。数据分辨率均为 1km×1km。

表 2 地理信息特征

Tab. 2 Features of Geographic Information

编号	类别
1	经度
2	纬度
3	高程
4	坡度
5	坡向
6	可燃物风险等级
7	植被类型风险等级
8	NDVI 植被指数

1.2.3 气象特征

在气象特征中，降水量、气温、湿度、风速等因素对于林火的发生有明显影响。降水量大小直接影响林区可燃物的含水量。如果一个地区的年降水量超过 1500 毫米，或月降水量超过 100 毫米，一般不发生或少发生森林火灾。日最高气温往往是该地着火与否的主要指标。山火往往多发于白天气温最高的时段。风速也是山火蔓延的重要影响因子，风速越大，火烧面积也越大。而对于湿度，相对湿度越大，可燃物含水率也随之增大。故本文选取实时时段温度、一小时内降水量、实时风速和相对湿度描述气象特征对山火风险的影响，如表 3 所示。

表 3 气象特征

Tab. 3 Meteorological Features

编号	类别
1	实时时段温度
2	1 小时降水量
3	风速
4	相对湿度

1.2.4 区域历史火点特征

历史山火数据可以有效指导未来山火防治工作，本文选取 2021 年火点密度和 2021 年山火风险等级作为南方五省（区）历史山火分布情况的描述，如表 4 所示。数据来源于《南方电网山火风险等级分布图（2021 版）》，该分布图是南方电网 2021 年统计的历史 10 年火点密度和山火风险等级，网格精度为 1km。

表 4 区域历史火点特征

Tab. 4 Regional Historical Fire Features

编号	类别
1	2021 火点密度
2	2021 山火风险等级

2 基于特征工程和 Boosting 集成学习框架模型融合的输电走廊实时山火风险评估模型

2.1 算法总体流程

本文提出基于特征工程和Boosting集成学习框架模型融合的输电走廊实时山火风险评估模型整体流程如图 1 所示。该模型主要有 4 个阶段：阶段 1 数据预处理；阶段 2 特征工程；阶段 3 建立基于 Boosting 算法的集成学习模型融合输电走廊山火灾害风险预测模型并进行超参数优化；阶段 4 模型评估和结果分析。这 4 个阶段又分为 10 个步骤，详细介绍如下。

2.2 阶段 1：数据预处理

2.2.1 步骤 1：特征提取

基于 1km×1km 精度的栅格数据和全国路网矢量数据，提取火点和非火点地理信息如高程、坡度、坡向，可燃物风险等级数据；提取人类活动特征如人口密度和区域国民生产总值，计算火点到铁路、高速公路、一级公路、最近居民点的欧氏距离，提取历史火情数据如火点密度和山火风险等级；根据火点和非火点经纬度和时间戳关联实时气象数据如温度、1 小时降雨量、风速和相对湿度数据，构建特征集。

2.2.2 步骤 2：数据正确性校验

检查是否所有取值和数据列名表述一致，绘制箱型图观察数据分布，判断各属性是否存在明显离群异常值，如存在，则对异常值所在行执行删除操作。

2.2.3 步骤 3：缺失值处理

识别空值，并使用列均值对空值进行填补。

2.3 阶段 2：特征工程

2.3.1 步骤 4：数据标准化

虽然数据标准化不影响 XGBoost 和 LightGBM 等树模型选择分裂点，不影响树的最终结构，但是数据标准化可以消除不同数据量纲的影响，提升如 SVM、KNN 等对比模型的收敛速度和准确度，从而使各类算法在同一数据集上可以更好的进行对比，故采用基于原始数据均值（Mean）和标准差（Standard Deviation）的正规化方法对样本空间各属性的各序列 $x_1, x_2, \ldots, x_n$ 进行变换：

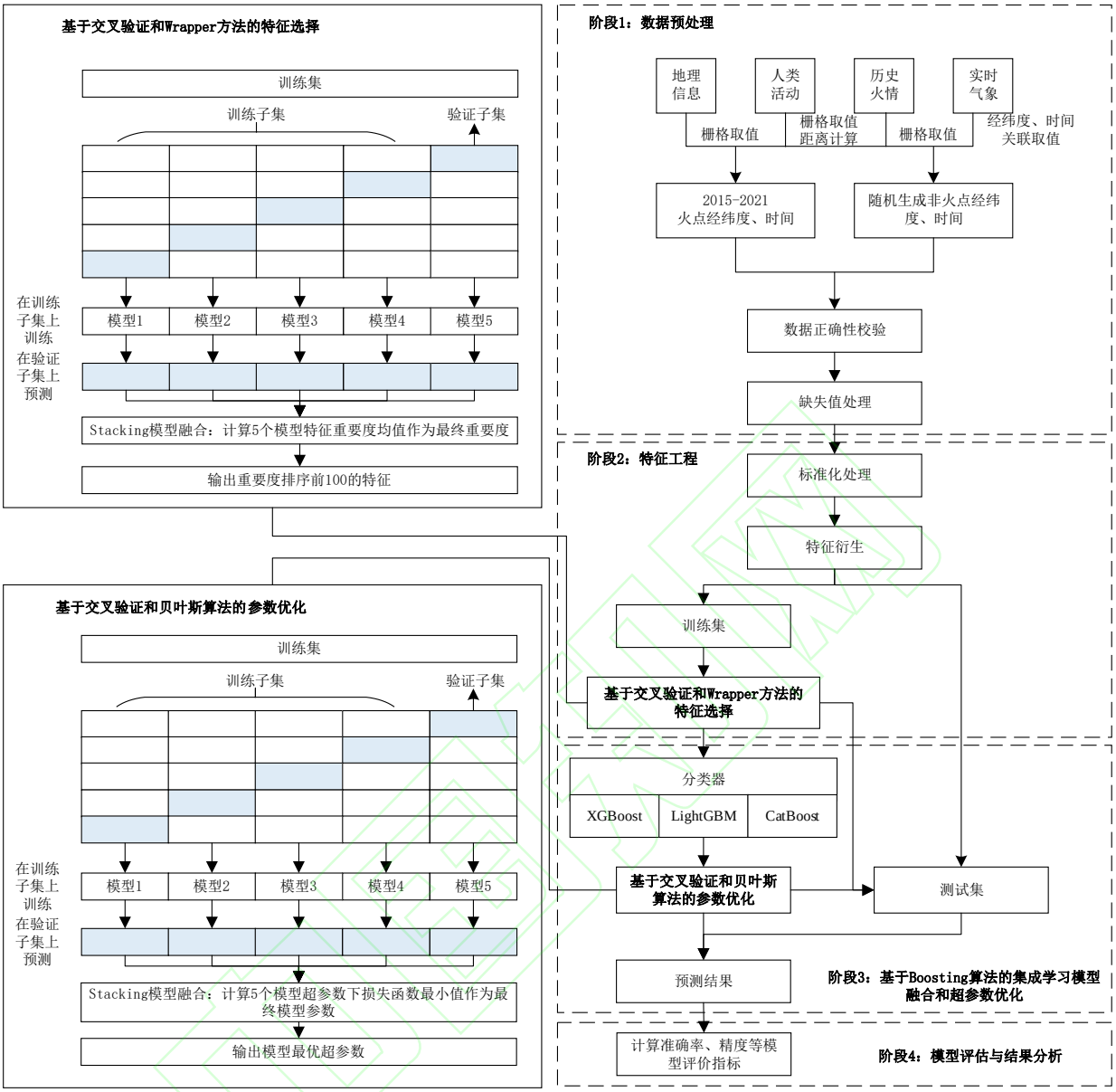

图 1 总体流程图

Fig. 1 Overall Flowchart

$$y_i = \frac{x_i - \bar{x}}{s} \tag{1}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{2}$$

2.3.2 步骤 5：特征衍生

特征衍生可以从已有数据集中尽可能多的挖掘数据的价值，本文利用多项式定理对 20 个连续变量特征 $x_1, x_2, ..., x_{20}$ 进行自动衍生：

$$\left(x_1 + x_2 + \cdots + x_{20}\right)^n = \sum \frac{n!}{n_1!n_2!\cdots n_{20}!} x_1^{n_1} x_2^{n_2} \cdots x_{20}^{n_{20}} \tag{3}$$

2.3.3 步骤 6：基于交叉验证和 Wrapper 方法的特征选择

（1）特征选择流程

常见的特征选择方法包括 Filter、Wrapper 等方法。本文采用的 Wrapper 方法是一种直接把最终使用模型性能作为特征子集评价标准的特征选择方法。相比绝大部分 Filter 方法，Wrapper 方法虽然计算开销更大，但是可以捕获特征之间相互作用的关系，从而使选择的特征子集让模型获得更好的表现。由于集成学习模型在选择特征分裂时具有一定随机性，所以本文使用交叉验证法对多个模型特征重要度取平均后再排序。在本文中，选择 Wrapper 方法结合 5 折交叉验证评估特征重要度，详细步骤如下：

Step1：将整个训练集随机分成 5 个数目近似相等的子集；

Step2：使用一个子集进行验证，其余子集用于训练集，执行 5 次进行训练，即每次将 80%数据用于训练，20%数据进行预测；

Step3：计算每次重复后的模型各属性重要度，将结果累加后取均值，得到平均重要度，并进行排序。

本文采用基于交叉验证和 Wrapper 技术的特征选择方法，流程如图 2 所示。

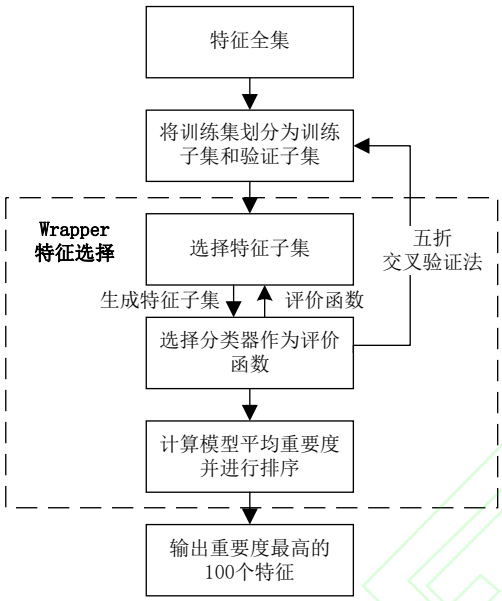

图 2 特征选择流程图

Fig. 2 Feature Selection Process

（2）不同分类器特征评价指标

大部分树模型在训练过程中通过计算特征被选为分裂特征的次数或总（平均）信息增益来衡量特征的重要程度，如 XGBoost、LightGBM 和 DT。而 CatBoost 使用预测变量变换值（Prediction Values Change, PVC）和损失函数变换值（Loss Function Change, LFC）描述特征值的变换，特征值越重要，则预测变量变换值和损失函数变换值越大，详细论述参照原论文[33]。本文选用 Boosting 算法中的 XGBoost、LightGBM 和 CatBoost 作为实验模型，选择决策树（Decision Tree, DT）作为基模型进行对比，4 种模型的特征评价备选指标和本文选择使用的评价指标如表 5 所示。

表 5 特征评价指标

Tab. 5 Feature Evaluation Metrics

编号	分类器	备选评价指标	本文选择评价指标
1	XGBoost	该特征被选为分裂特征的次数、总（平均）信息增益、特征平均覆盖率	该特征被选为分裂特征的次数
2	LightGBM	该特征被选为分裂特征的次数、总（平均）信息增益	该特征被选为分裂特征的次数
3	CatBoost	预测值改变量和损失函数变换值	预测值改变量和损失函数变换值
4	Decision Tree	总（平均）信息增益、Gini 系数、信息增益率	总（平均）信息增益

（3）特征重要度排序结果

根据上述流程，在 XGBoost、LightGBM 和 CatBoost 模型中计算得到的特征重要度及排序如图 3、4、5 所示，由于篇幅原因，只展示重要度最高的前 10 个特征，特征名称和对应编码如表 6 所示。对于 XGBoost 算法，最近居民点距离、纬度、坡向、经度和最近一级公路距离对模型预测贡献程度最大；对于 LightGBM，火点密度、湿度、实时降雨量、湿度、火点密度、纬度和经度与高程构成的复合指标对模型预测贡献程度最大；对于 CatBoost 算法，实时降雨量、湿度、火点密度、经度与高程构成的复合指标和降雨量与火点密度构成的复合指标对模型预测贡献程度最大。

表 6 特征和对应编码

Tab. 6 Features and Corresponding Codes

编号	类别
x0	经度
x1	纬度
x2	一小时降雨量(mm/h)
x3	风速(m/s)
x4	温度(℃)
x5	相对湿度(%)
x6	高程(m)
x7	坡向(度)
x8	坡度(度)
x9	最近铁路距离(m)
x10	最近高速公路距离(m)
x11	最近一级公路距离(m)
x12	最近居民点距离(m)
x13	火点密度(火点 km^2)
x14	可燃物风险等级
x15	山火风险等级
x16	植被类型风险等级
x17	POP2019（人 km^2）
x18	GDP2019（值 km^2）
x19	2021_NDVI（值 km^2）

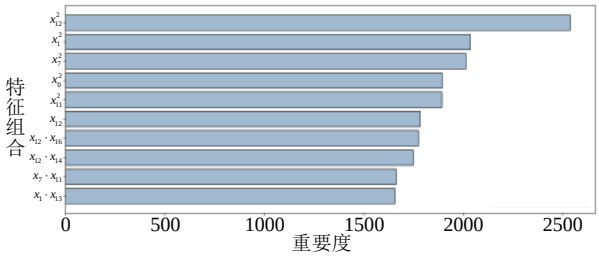

图 3 XGBoost 特征重要度

Fig. 3 XGBoost Feature Importance

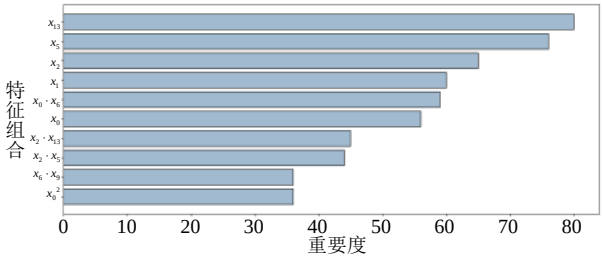

图 4 LightGBM 特征重要度

Fig. 4 LightGBM Feature Importance

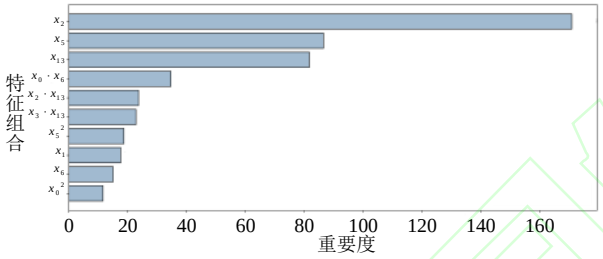

图 5 CatBoost 特征重要度

Fig. 5 CatBoost Feature Importance

2.4 阶段 3：建立基于 Boosting 算法的集成学习模型融合输电走廊山火灾害风险预测模型并进行超参数优化

2.4.1 步骤 7：选择分类器在训练集上进行训练

（1）XGBoost 算法核心原理

XGBoost 是一个基于决策树的集成机器学习算法，使用了梯度提升框架。在涉及非结构化数据(图像、文本等)的预测问题中，人工神经网络的表现往往优于所有其他算法或框架。然而，当处理中小型结构化表格数据时，基于树模型的算法往往表现最好，其结果是多棵决策树的求和，计算公式为：

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{n} f_k(x_i), \quad f_k \in F \tag{4}$$

式中，$\hat{y}_i$ 为模型预测值，x_i 为第 i 个样本，k 为决策树的数目，F 为所有树的集合。

在 XGBoost 中可以根据特征分裂次数、特征平均增益和特征平均覆盖率来衡量特征重要度，计算公式为：

$$F_{split} = |X| \tag{5}$$

$$AverageGain = \frac{\sum \mathrm{Gain}_X}{\mathrm{F}_{split}} \tag{6}$$

$$AverageCover = \frac{\sum Cover_X}{\mathrm{F}_{split}} \tag{7}$$

其中，X 是所有特征在叶子节点的集合；Gain 是 X 中每个叶子节点在分割时的节点增益值；Cover 是 X 中落在每个节点的样本个数。

（2）LightGBM 算法核心原理[32]。

LightGBM 是一种基于梯度提升决策树 GBDT（Gradient Boosting Decision Tree, GBDT）的集成回归树近似最终模型。针对 XGBoost 算法在特征维度度极高，数据量极大数据场景下表现欠佳提出的优化算法。LightGBM 不论在时间复杂度还是内存消耗上都优于 XGBoost。GOSS（Gradient-based One-Side Sampling）和 EFB（Exclusive Feature Bundling）技术是 LightGBM 论文中提出的两种采样策略。GOSS 策略用于减少样本维度，EFB 策略用于合并稀疏特征，减少特征维度，由于本文特征数目有限，特征值并不稀疏，故采用 GOSS 策略保留梯度较大样本，对梯度值小的数据实例进行随机采样，不使用 EFB 策略。LightGBM 基础弱分类器为决策树，表示如下：

$$L = f_{\mathrm{T}}(X) = \sum_{t=1}^{T} f_t(X), \ f_t \in \Theta \tag{8}$$

f_t 为第 t 棵决策树，Θ 为所有树的集合空间。x_s 为样本空间，假设有数据量为 n 的训练集 $\{x_1, \cdots, x_n\}$，其中 x_i 为 x_s 中第 i 个维度为的向量。

构造损失函数 $loss = L(y, H(x))$，y 为输出，$H(x)$ 为估计函数，$H^*(x)$ 为最小损失函数，表示如下：

$$\begin{aligned} H^*(x) &= \arctan \min_H E_{y,x}(L(y, H(x))) \\ &= \arctan \min_H E_x(E_y(L(y, H(x))) \mid x) \end{aligned} \tag{9}$$

在每次梯度提升的迭代中，当前损失函数输出值为 $\{g_1, \cdots, g_n\}$，g_i 为 x_i 在当前迭代中对应损失函数的负梯度，决策树在信息增益最大的特征点分裂，方差用于度量信息增益。

假设 O 为基模型的一个数据集，此节特征 j 处特征在分割点 d 的信息增益如下：

$$V_{j|O}(d)=\frac{1}{n_O}\left(\frac{\left(\sum_{x_i\in O:x_i z d} g_i\right)^2}{n^j_{l||O}(d)}+\frac{\left(\sum_{\langle x_i\in O:x_i d} g_i\right)^2}{n^j_{r|O}(d)}\right) \quad (10)$$

n_O 为训练集样本数，且 $n_O=\sum I\left[x_i\in O\right]$；$n^j_{l|O}(d)$ 为第 j 个特征中值小于等于 d 的样本数，且 $n^j_{l|O}(d)=\sum I\left[x_i\in O:x_{ij}\leqslant d\right]$，$n^j_{r|O}(d)$ 为第 j 个特征中值大于 d 的样本数，且 $n^j_{r|O}(d)=\sum I\left[x_i\in O:x_{ij}\leqslant d\right]$。

LightGBM 遍历每个特征的每个分裂点，计算信息增益，在信息增益最大处分裂。

与 XGBoost 类似，LightGBM 也是通过计算特征分裂次数和特征总（平均）增益来评估特征的重要程度。

（3）CatBoost 算法核心原理[33]。

CatBoost 与 LightGBM、XGboost 算法都是基于 GBDT 算法实现。在 GBDT 基础上，Catboost 对标称属性处理进行改进；并且提出了排序提升策略对预测偏移进行了优化，有效减少模型过拟合。

CatBoost 算法中，排序提升策略替代了传统的梯度估计，可以有效减少梯度估计偏差，提高模型泛化能力。为得到无偏梯度估计，CatBoost 算法会用每一个样本 x_i 训练出一个单独的模型 M_i，M_i 由不包含样本 x_i 的训练集训练获得，算法步骤如下[33]：

输入：$\{(x_k,y_k)\}_{k=1}^{n},I$

1. σ-*randompermutation* of $[1,n]$;
2. $M_i\leftarrow 0$ for $i=1..n$;
3. **for** $t-1$ **to** I **do**
4. **for** $i-1$ **to** n **do**
5. $r_i-y_i-M_{\sigma(i)-1}(x_i)$;
6. **end for**
7. **for** $i-1$ **to** n **do**
8. $\Delta M\leftarrow$
9. $LearnModel\left(\left(x_j,r_j\right):\sigma(j)\leq i\right)$
10. $M_i\leftarrow M_i+\Delta M$
11. **end for**

12. end for

输出：最强模型 M_n

2.4.2 步骤 8：基于交叉验证和贝叶斯算法的超参数优化

人工手动调参不仅费时费力，而且往往效果差强人意。比较常见的参数搜索法有网格搜索、随机搜索和贝叶斯优化。网格搜索即为暴力搜索，通过穷举遍历所有的可能解获得最优值，计算量极大，很容易陷入组合爆炸。与网格搜索相比，随机搜索并未尝试所有参数值，而是从指定的分布中采样固定数量的参数设置，搜索速度通常快于网格搜索，但是搜索效果随机性较强。相比于上述两种搜索策略，贝叶斯优化算法采用高斯过程，能够考虑之前搜索的参数信息，迭代次数少，收敛速度快，并且对于非凸优化问题表现稳健，能够比较容易找到局部最优解。故本文利用贝叶斯算法搜索超参数空间，寻找局部最优参数，调优后各模型参数设置情况见表 7。

表 7 各模型超参数搜索空间

Tab. 7 Hyperparameter Search Space of Each Model

模型	超参数名称	最优超参数最佳取值
XGBoost	max_depth	11
	learning_rate	0.43
	min_child_weight	1
	colsample_bytree	0.75
	gamma	0.30
	n_estimators	120
LightGBM	learning_rate	0.49
	bagging_fraction	5
	feature_fraction	0.67
	num_leaves	50
	reg_alpha	9
	reg_lambda	8.40
	bagging_freq	5
	min_child_samples	26
CatBoost	depth	6
	border_count	254
	l2_leaf_reg	3
	learning_rate	0.23
	leaf_estimation_iterations	10
	bayesian_matrix_reg	0.10
	model_size_reg	0.5

2.4.3 步骤 9：在测试集上进行预测

分别使用超参数优化后的 XGBoost、LightGBM 和 CatBoost 模型对测试集数据进行预测，将预测结果和实际分类结果进行对比。

2.5 阶段 4：模型评估与结果分析（步骤 10）

2.5.1 模型评估指标

采用混淆矩阵对模型的性能进行评估,如表 8 所示。对于山火预测的二分类问题,将预测是否发生山火和实际是否发生山火的结果进行直观对比。T_P

（True Positive）为正确被预测的山火样本;T_N（True Negative）为正确被预测的未发生山火的样本;F_P（False Positive）为错误预测的山火样本;F_N（False Negative）为错误预测的为发生山火的样本。

表 8 混淆矩阵

Tab. 8 Confusion Matrix

实际\预测	预测不会发生山火	预测会发生山火
实际未发生山火	T_N	F_N
实际发生了山火	F_P	T_P

基于混淆矩阵,本文引入准确率 P_a、精确率 P_p、召回率 P_r 对模型的预测效果进行衡量,表示如下：

$$P_a = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{11}$$

$$P_p = \frac{T_P}{T_P + F_P} \tag{12}$$

$$P_r = \frac{T_P}{T_P + F_N} \tag{13}$$

$$F = \frac{(1+\beta^2)P_p P_r}{\beta^2 P_p + P_r} \tag{14}$$

准确率 P_a 反映所有预测正确的样本占所有样本的比例;精确率 P_p 反映预测结果为正例的样本占真实为正例的比例;召回率 P_r 反映真实为正例的样本中预测结果为正例的比例。

由于召回率和精确率在一定程度上相对，而在实际日常工作中，相比于排查的精准性，运维人员更重视排查的全面性，所以本文选择准确率、召回率作为评价指标，选择召回率和精确率的调和均值 $\boldsymbol{F}$ 作为参考，在 $\boldsymbol{F}$ 中赋予召回率更大权重，β 取 4。

模型效果的性能评价指标有很多，除分类模型的准确率、召回率和调和均值外，本文还着重关注模型的运行效率，可以辅助预测算法工程落地的运行效果和评估所需硬件资源。

2.5.2 评估模型准确率、召回率和调和均值

由图 6、7 对比可知，总体而言 CatBoost 模型性能在准确率、召回率和调和均值方面都是最优，不论是否选择进行特征衍生。虽然决策树准确率达到 98%，但召回率和调和均值都只有 76%左右，相差悬殊，说明模型受正负不平衡样本影响程度大，过拟合严重。值得注意的是，相比于其他算法，特征衍生对 LightGBM 模型准确率和调和均值提升显著，经过衍生后准确率从 84.3%上升至 94.3%，调和均值从 75%上升至 83.5%。

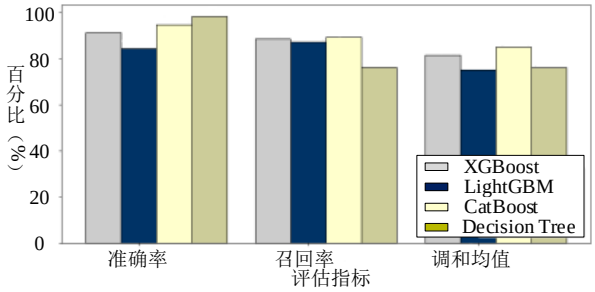

图 6 四种模型准确率、召回率和调和均值对比（不进行特征衍生）

Fig. 6 Comparison of Accuracy, Recall and Harmonic Mean of the Four Models(Characteristics without Derivative)

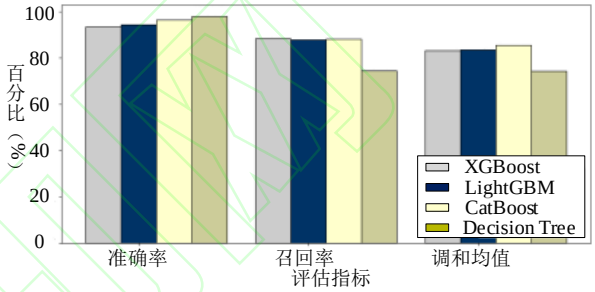

图 7 四种模型准确率、召回率和调和均值对比（特征衍生）

Fig. 7 Comparison of Accuracy, Recall and Harmonic Mean of the Four Models(Characteristics of the Derivative)

2.5.3 评估模型运行效率

不同算法的训练时间和预测时间对比如图 8、9 所示。从图 8 可得出，特征衍生后，除去 LightGBM，其余模型的运行效率显著降低，LightGBM 受特征维度数目影响极小。在经过特征衍生后的模型中，CatBoost 也展示出不俗的运行效率，训练时间仅次于 LightGBM，同样远快于其他两种模型。从图 9 可得出，4 种算法在是否进行特征衍生的选择上没有显著差异。

综合考虑所有指标，对于 XGBoost 和 CatBoost 算法，可以选择不进行特征衍生，特征衍生对这两种算法准确率的提升仅在 2%左右，对召回率和调和均值更影响甚微，意味着特征衍生不能提高这两种模型山火样本预测的准确性，只能小幅提高非山火样本预测的准确性。但对于 LightGBM，特征衍生效果显著，经过衍生后准确率上升 10%，调和均值上升 8.5%，并且特征维度十余倍的增长并未增加算法的时间复杂度，模型训练时间甚至小幅降低，在四种算法中是工程落地最优选。如果仅考虑准确率、召回率和调和均值指标，则 CatBoost 算法表现最优。

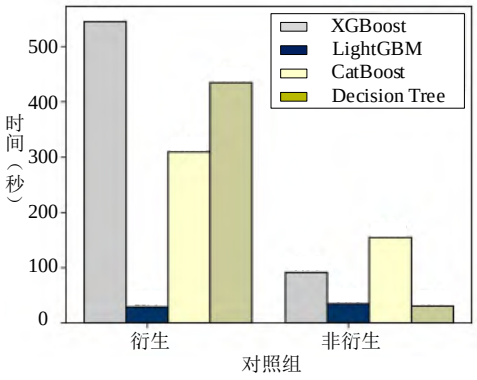

图 8 训练时间对比

Fig. 8 Comparison of Training Time

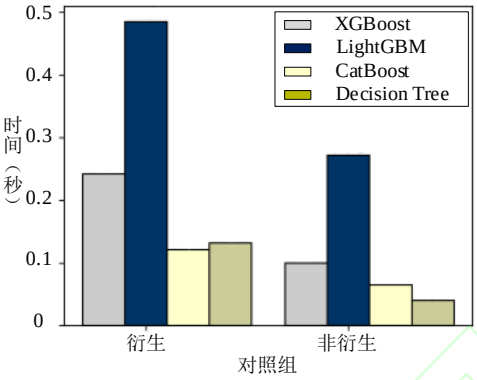

图 9 预测时间对比

Fig. 9 Comparison of Predict Time

2.5.4 南方五省卫星山火热点实例验证分析

本文采用 2021-2022 南方五省实时卫星山火和非山火数据作为测试集对模型进行实例验证，总计 187042 条实时输电走廊通道数据，其中卫星山火告警数据 6620 条，非山火实时数据 180422 条，山火和非山火样本比例为 1 比 27，每条数据 20 个特征，其中时段温度、1 小时降水量、风速和相对湿度四个特征为实时特征。

南方电网山火风险分布图将山火风险等级分为四个等级，6620 条实时卫星山火数据中 80.8%落在风险等级为三级和四级的区域。180422 条非山火数据中 60.35%落在了三级和四级风险区域。由于南网山火分布图是静态的，不具备实时性，单一地理位置坐标便决定了风险等级，准确率难以直接对比，此数据仅供参考。

如章节 2.5.2 所述，本文提出的基于特征衍生的 XGBoost、LightGBM 和 CatBoost 算法准确率分别为 93.4%，94.3%和 96.4%，召回率分别为 83.1%，83.3%和 85.5%。由于三个集成学习模型准确率和召回率都十分接近，故以基于特征衍生和 LightGBM 的模型预测结果为例进行呈现分析，预测效果如图 10 所示。

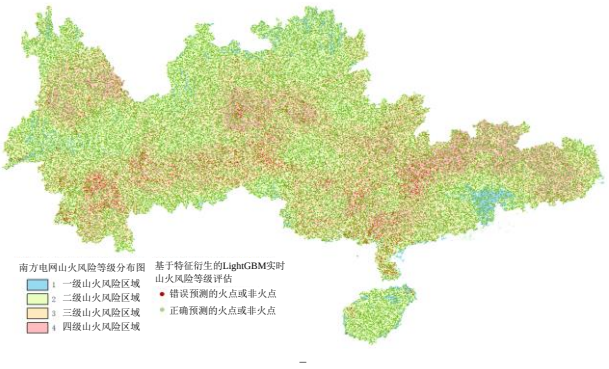

图 10 南方五省实时卫星山火热点实例验证分析

Fig. 10 Validation and Analysis of Real-time Satellite Hot Spots in Five Southern Provinces

从图中结合准确率和召回率指标可以得出，本文提出的算法误分类点集中在山火分布图高风险等级区域，将模型误分类数据筛选后如图 11 所示，94.7%误分类为非山火误识别为山火，6620 条卫星山火告警中仅有 11.7%分类错误，相比于南方电网山火分布图，火点的误分率降低 7.5%，在正负样本悬殊的真实数据场景下，模型泛化能力表现优异。在真实运维工作中，相比与“查准”山火，运维管理人员更加看重“查全”山火，本文提出的基于特征工程和集成学习的模型相比于静态山火分布图有着更优异的查全率和查准率，可以更有效的指导班组开展山火防治工作。

图 11 南方五省实时通道走廊山火风险误分类分析

Fig. 11 Misclassification Analysis of Real-time Corridor Wildfire Risk in Five Southern Provinces

2.5.5 南方五省跳闸实例验证分析

本文使用南方五省近年 80 例山火跳闸案例进行山火跳闸实例验证，南方电网山火分布图和本文提出的基于特征衍生和 LightGBM 的模型预测结果对比如图 12 所示。

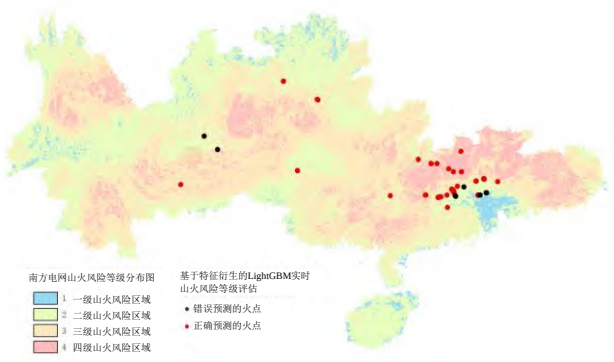

图 12 南方五省跳闸实例验证分析

Fig. 12 Verification and Analysis of Trip Cases in Five Southern Provinces

80 例真实跳闸案例中，55%的跳闸坐标落在南方电网分布图中山火风险等级为三级或四级的区域，抽取跳闸坐标的静态特征和实时气象特征，每一例跳闸案例构成 20 维特征向量作为模型输入，88.75%的案例可以被基于特征衍生和 LightGBM 的模型正确预测。本文提出的基于特征工程和集成学习的实时山火风险评估模型相比于分布图，不论是在实时卫星山火热点验证分析还是在跳闸实例验证分析中都表现更佳，可以有效指导运维人员开展山火防治工作。

3 结论

本文研究了不同特征影响山火是否发生的重要程度，充分结合人工智能和机器学习领域前沿算法，基于 Wrapper 技术、交叉验证和三种 Boosting 算法，提出了一种输电走廊实时山火风险评估模型，并在十八万余条实时卫星山火和非山火数据集上进行了验证分析，山火和非山火样本比例为 1 比 27，结论如下

（1）采用特征工程和 Boosting 集成学习框架模型融合算法对山火风险进行预测时，采用特征衍生的 CatBoost 算法模型准确率最高，为 96.4%，召回率为 88.1%，调和均值为 85.5%。

（2）采用特征工程和 Boosting 集成学习框架模型融合算法对山火风险进行预测时，采用特征衍生的 LightGBM 算法运行效率最佳，训练所需时间不到采用特征衍生的 CatBoost 算法 10%，模型准确率至 94.3%，召回率至 87.8%，调和均值至 83.5%，略低于采用特征衍生的 CatBoost 算法。

（3）本文对数据进行了特征工程，融合了不同训练模型的优势，相较于朴素贝叶斯，模型准确率有大幅提升，提高 16%，相比于决策树，模型准确率提高 20%。

（4）从准确率、召回率和调和均值三个角度分析，经过特征衍生，能有效提高模型效果，LightGBM 算法提升最为显著，其准确率、召回率和调和均值分别提高了 10%、0.7%和 8.5%。

（5）相比于静态的南方分布图，本文提出的基于特征工程和集成学习模型融合的输电走廊山火实时风险评估模型可以提取实时的气象特征，不论是在实时卫星山火热点数据还是在近年真实跳闸案例的预测中都有着更高的准确率和召回率，可以实时指导班组开展日常山火防治工作。

本文所采用的集成学习算法具有天然并行优势，未来可以研究本文提出的算法在 Spark、MapReduce 等分布式大数据处理框架下的实现，以进一步提高模型训练效率，减少算法时间复杂度。

参考文献

[1]. 何诚,舒立福,刘柯珍,周生瑞.广西地区山火引起高压线路跳闸环境特征研究[J].消防科学与技术,2020,39(12):1626-1629.
HE Cheng, SHU Lifu, LIU Kezhen, ZHOU Shengrui. Study on environmental characteristics of high voltage line tripping caused by mountain fire in guangxi [J]. Fire Science and Technology,2020,39(12):1626-1629(in Chinese).

[2]. 周恩泽,樊灵孟,黄勇,周游,周文涛,陈维捷.基于火焰燃烧模型的输电线路山火跳闸风险分布评估[J].电网技术，2022,46(07) :2778-2785.DOI:10.13335/j.1000-3673.pst.2021.1039.
ZHOU Enze, FAN Lingmeng, HUANG Yong, ZHOU Zuo, ZHOU Wentao, CHEN Weijie. Risk distribution assessment of bushfire tripping in transmission lines based on flame combustion model [J]. Power Grid Technology, 222,46(07):2778-2785.DOI:10.13335/ J.1000-3673.pST.2021.1039(in Chinese).

[3]. 陆佳政,周特军,吴传平,李波,刘毓,朱远.输电线路差异化防山火技术与策略[J]. 高电压技术，2017,43(08): 2524-2532. DOI: 10. 13336/j. 1003-6520. hve.20170731012a
LU Jiazheng, ZHOU Tejun, WU Chuan-ping, LI Bo, LIU Yu, ZHU Yuan.HighVoltageTechnology,2017,43(08):2524-2532.DOI:10.13336/j.1003-6520.hve.20170731012(in Chinese).

[4]. 周恩泽,胡思雨,张录军,魏瑞增,王华婴,杨凡.电网山火灾害特征及风险预警技术[J].电力工程技术,2020,39(03):58-64.
ZHOU Enze, HU Siyu, ZHANG Lujun, WEI Ruizheng, WANG Hua Zhao, Yang Fan. Characteristics and risk warning technology of bushfire disaster in power grid [J]. Electric Power Engineering Technology,2020,39(03):58-64(in Chinese).

[5]. 陆佳政,周特军,吴传平,李波,刘毓,朱远.输电线路差异化防山火技术与策略[J].高电压技术,2017,43(08):2524-2532.DOI:10.13336/j.1003-6520.hve.20170731012(in Chinese).
LU Jiazheng, ZHOU Tejun, WU Chuan-ping, LI Bo, LIU Yu, ZHU Yuan. High Voltage Technology,2017,43(08):2524-2532.DOI:10.13336/j.1003-6520.hve.20170731012(in Chinese).

[6]. 周恩泽,黄勇,陈洁,魏瑞增,王彤,隋三义.广东地区山火识别方法及其在电网中的应用[J].气象科技，2020，48(01)：132-140. DOI：10.19517/j. 1671-6345.20190025.
ZHOU Enze, HUANG Yong, CHEN Jie, WEI Ruizeng, WANG Tong, SUI Sanyi. Method of mountain fire identification in guangdong and its application in power grid [J]. Meteorological Science and Technology,2020,48(01):132-140.DOI:10.19517/j.1671-634

5.20190025(in Chinese).

[7]. 周恩泽,全玉生,吴昊,房林杰,叶海峰.基于故障特征谱的输电线路故障状态评估[J].电力科学与工程,2017,33(03):26-30.
Zhou Enze, Quan Yusheng, Wu Hao, Fang Linjie, Ye Haifeng. Fault state evaluation of transmission lines based on fault feature spectrum [J]. Electric Power Science and Engineering,2017,33(03):26-30(in Chinese).

[8]. 周恩泽,龚博,刘淑琴,向谆,史晓桢,黄道春,陈鑫.南方电网架空线路因山火跳闸故障统计分析[J].广东电力,2022,35(04):80-86.
ZHOU Enze, GONG Bo, LIU Shuqin, XIANG Zhun, SHI Xiaozhen, HUANG Daochun, CHEN Xin. Statistical analysis of overhead line tripping fault due to mountain fire in china southern power grid [J]. Guangdong Electric Power, 222,35(04):80-86(in Chinese).

[9]. 吴田,阮江军,胡毅,刘兵,陈成.500kV 输电线路的山火击穿特性及机制研究[J]. 中国电机工程学报,2011,31(34):163-170.DOI:10.13334/j.0258-8013.pcsee.2011.34.017.
WU Tian, RUAN Jiangjun, HU Yi, LIU Bing, CHEN Cheng. Study on the breakdown characteristics and mechanism of 500kv transmission line by bushfire [J]. Proceedings of The Csee,2011,31(34):163-170.DOI:10.13334/J.0258-8013.Pcsee.2011.34.017(in Chinese).

[10]. 叶立平,陈锡阳,何子兰,谢从珍,黄健华,夏云峰,戴栋.山火预警技术在输电线路的应用现状[J]. 电力系统保护与控制,2014,42(06):145-153.
YE Liping, CHEN Xiyang, HE Zilan, XIE Congzhen, HUAN Jianhua G, XIA Yunfeng, DAI Dong. Application status of hill fire early warning technology in transmission lines [J]. Protection and control of power systems,2014,42(06):145-153(in Chinese).

[11]. 周志宇. 山火灾害下电网输电线路跳闸风险评估研究[D].华北电力大学(北京),2019.DOI:10.27140/d.cnki.ghbbu.2019.000060.
ZHOU Zhiyu. Power grid transmission line trip under fire disaster risk assessment research [D]. North China electric power university (Beijing), 2019. The DOI: 10.27140 /, dc nki. Ghbbu. 2019.000060(in Chinese).

[12]. 周游,隋三义,陈洁,周恩泽,黄勇,王彤.基于 Himawari-8 静止气象卫星的输电线路山火监测与告警技术[J]. 高电压技术,2020,46(07):2561-2569.DOI:10.13336/j.1003-6520.hve.20190498.
ZHOU You, SUI Sanyi, CHEN Jie, ZHOU Enze, HUANG Yong, WANG Tong. Monitoring and Alarm technology of transmission line mountain fire based on Himawari-8 Geostationary meteorological satellite [J]. High Voltage Technology,2020,46(07):2561-2569.DOI:10.13336/j.1003-6520.hve.20190498(in Chinese).

[13]. 陆佳政,吴传平,杨莉,张红先,刘毓,徐勋建.输电线路山火监测预警系统的研究及应用[J].电力系统保护与控制,2014,42(16):89-95.
LU Jiazheng, WU Chuanping, YANG Li, ZHANG Hongxian, LIU Yu, XU Xunjian. Research and application of bushfire monitoring and early warning system for transmission lines [J]. Protection and Control of Power Systems,2014,42(16):89-95(in Chinese).

[14]. 陆佳政,刘毓,徐勋建,杨莉,章国勇,何立夫.架空输电线路山火预测预警技术[J]. 高电压技术,2017,43(01):314-320.DOI:10.13336/j.1003-6520.hve.20161227041.
LU Jiazheng, LIU Yu, XU Xun-jian, YANG Li, ZHANG Guoyong, HE Li-fu. Prediction and early warning technology of bushfire in overhead transmission line [J]. High Voltage Technology,2017,43(01):314-320. (in Chinese) DOI:10.13336/j.1003-6520.hve.20161227041(in Chinese).

[15]. 陆佳政,刘毓,吴传平,张红先,周特军.输电线路山火卫星监测与告警算法研究[J]. 中国电机工程学报,2015,35(21):5511-5519.DOI:10.13334/j.0258-8013.pcsee.2015.21.015.
LU JiaZheng, LIU Yu, WU Chuanping, ZHANG Hongxian, ZHOU Tejun. Research on satellite monitoring and alarm algorithm of transmission line mountain fire [J]. Proceedings of the CSEE, 2015,35(21):5511-5519.DOI:10.13334/ j.0258-8013.pcse.2015.21.01 (in Chinese).

[16]. Shi, Shuzhu & Yao, Chunjing & Wang, Shiwei & Han, Wenjun. (2018). A Model Design for Risk Assessment of Line Tripping Caused by Wildfires. Sensors. 18. 1941. 10.3390/s18061941.

[17]. 周恩泽,黄勇,陈洁,田翔,魏瑞增,周游.基于图模型的架空输电线路山火风险等级预测模型[J]. 南方电网技术,2020,14(04):8-16.DOI:10.13648/j.cnki.issn1674-0629.2020.04.002.
ZHOU Enze, HUANG Yong, CHEN Jie, TIAN Xiang, WEI Ruizeng, ZHOU Zuo. Overhead transmission lines based on graph model fire risk rating forecast model [J]. Journal of Southern Power Grid Technology, 2020, 14 (4) : 8-16. DOI: 10.13648 / j.carol carroll nki issn1674-0629.2020.04.002(in Chinese).

[18]. 熊小伏,曾勇,王建,李浩然.基于山火时空特征的林区输电通道风险评估[J].电力系统保护与控制,2018,46(04):1-9.
XIONG Xiaofu, ZENG Yong, WANG Jian, LI Haoran. Risk assessment of transmission channel in forest area based on temporal and spatial characteristics of bushfire [J]. Protection and Control of Power Systems,2018,46(04):1-9(in Chinese).

[19]. 艾欣,周志宇.山火灾害下电网输电线路关键性评估方法[J].高电压技术,2018,44(08):2433-2441.DOI:10.13336/j.1003-6520.hve.20180731001.
Ai Xin, Zhou Zhiyu. Key evaluation method of power grid transmission line under bushfire disaster [J]. High Voltage Technology, 2018,44(08):2433-2441.DOI:10.13336/j.1003-6520.hve.20180731001 (in Chinese).

[20]. Liu, Yu & Li, Bo & Wu, Chuanping & Chen, Bao-Hui & Zhou, Tejun. (2021). Risk warning technology for the whole process of overhead transmission line trip caused by wildfire. Natural Hazards. 107. 1-18. 10.1007/s11069-021-04579-y.

[21]. Dian, Songyi & Cheng, Peng & Ye, Qiang & Wu, Jirong & Luo, Ruisen & Wang, Chen & Hui, Dafeng & Zhou, Ning & Zou, Dong & Gong, Xiaofeng. (2019). Integrating Wildfires Propagation Predic-tion Into Early Warning of Electrical Transmission Line Outages. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2894141.

[22]. 陆佳政,刘毓,徐勋建,杨莉,章国勇,何立夫.架空输电线路山火预测预警技术[J]. 高电压技术,2017,43(01):314-320.DOI:10.13336/j.1003-6520.hve.20161227041.
LU Jia-zheng, LIU Yu, XU Xunjian, YANG Li, ZHANG Guo-yong, HE Li-fu. Prediction and early warning technology of bushfire in overhead transmission line [J]. High Voltage Technology,2017,43(01):314-320. (in Chinese) DOI:10.13336/j.1003-6520.hve.20161227041(in Chinese).

[23]. Chen, Weijie & Zhou, You & Zhou, Enze & Xiang, Zhun & Zhou, Wentao & Lu, Junhan. (2021). Wildfire Risk Assessment of Trans-mission-Line Corridors Based on Naïve Bayes Network and Remote Sensing Data. Sensors. 21. 634. 10.3390/s21020634.

[24]. 周恩泽,黄勇,龚博,魏瑞增,向谆,陈维捷,周游.基于朴素贝叶斯网络的输电走廊山火风险评估模型[J].南方电网技术，2021，15(08):120-129.DOI:10.13648/j.cnki.issn1674-0629.2021.08.015.
ZHOU Enze, HUANG Yong, GONG Bo, WEI Ruizeng, XIANG Zhun, CHEN Weijie, ZHOU Zuo. Based on the simple bayesia

n network transmission corridor fire risk assessment model [J]. Journal of Southern Power Grid Technology, 2021 (8) : 120-129. The DOI: 10.13648 / J.Carol Carroll Nki Issn1674-0629.2021.08.015(in Chinese).

[25]. 周恩泽,黄勇,陈洁,田翔,魏瑞增,周游.基于图模型的架空输电线路山火风险等级预测模型[J].南方电网技术, 2020, 14(04): 8-16. DOI: 10. 13648/j.cnki.issn1674-0629.2020.04.002.
ZHOU Enze, HUANG Yong, CHEN Jie, TIAN Xiang, WEI Rui zeng, ZHOU Zuo. Overhead transmission lines based on graph model fire risk rating forecast model [J]. Journal of Southern Power Grid Technology, 2020, 14 (4) : 8-16. DOI: 10.13648 / J.Carol Carroll Nki Issn1674-0629.2020.04.002(in Chinese).

[26]. 周恩泽,黄勇,向谆,罗颖婷,魏瑞增,向坤轩,周游.基于物元可拓的输电线路山火风险评估模型[J].南方电网技术,2022,16(01):145-154.DOI:10.13648/j.cnki.issn1674-0629.2022.01.016.
ZHOU Enze, HUANG Yong, XIANG Zhun, LUO Yingting, WEI Ruizeng, XIANG Kunxuan, ZHOU Zuo. Based on the matter-element extension fire risk assessment model of transmission line [J]. Journal of Southern Power Grid Technology, 2022 (01) : 145-154. The DOI: 10.13648 / j.carol carroll nki issn1674-0629.2022.01.016(in Chinese).

[27]. 周恩泽,龚博,刘淑琴,向谆,史晓桢,黄道春,陈鑫.南方电网架空线路因山火跳闸故障统计分析[J].广东电力,2022,35(04):80-86.
ZHOU Enze, GONG Bo, LIU Shuqin, XIANG Zhun, SHI Xiaozhen, HUANG Daochun, CHEN Xin. Statistical analysis of overhead line tripping fault due to mountain fire in china southern power grid [J]. Guangdong Electric Power, 222,35(04):80-86(in Chinese).

[28]. 李隆基,文清丰,周文涛,郁晓光,张弛.基于大数据的输电线路通道智能风险预控技术研究[J].电测与仪表,2020,57(06):82-87.DOI:10.19753/j.issn1001-1390.2020.06.013.
LI Longji, WEN Qingfeng, ZHOU Wentao, XI Xiaoguang, ZHANG Chi. Based on large data transmission channel intelligent risk precontrol technology research [J]. Electric Measurement and Instrumentation, 2020,57 (6) : 82-87. The DOI: 10.19753 / j.issn 1001-1390.2020.06.013(in Chinese).

[29]. 陆佳政,周特军,吴传平,李波,谭艳军,朱远.某省级电网 220kV 及以上输电线路故障统计与分析[J].高电压技术, 2016, 42(01): 200- 207. DOI: 10.13336/j.1003-6520.hve.2016.01.026.
LU Jiazheng, ZHOU Tejun, WU Chuanping, LI Bo, TAN Yanjun, ZHU Yuan. Fault statistics and analysis of 220kv and above transmission lines in a provincial power grid [J] High Voltage Technology,2016,42(01): 200-207.Doi :10.13336/J.1003-6520.Hve.2016.01.026(in Chinese).

[30]. 陆佳政,刘毓,杨莉,徐勋建,吴传平.输电线路山火发生规律分析[J].消防科学与技术,2014,33(12):1447-1451.
LU Jiazheng, LIU Yu, YANG Li, XU Xunjian, WU Chuanping. Fire Analysis of Bushfire occurrence in Transmission Line [J]. Science and Technology,2014,33(12):1447-1451(in Chinese).

[31]. Chen, Bao-Hui. (2021). Study on Distribution Regularity of Over-head Transmission Line nearby Wildfires.

[32]. Meng, Qi. (2018). LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

[33]. Dorogush, Anna & Ershov, Vasily & Gulin, Andrey. (2018). Cat-Boost: gradient boosting with categorical features support.


张可颖

收稿日期：2022-11-11。

作者简介：

张可颖(1997)，女，硕士，工程师，研究方向为数据科学输电防灾减灾应用研究， E-mail：993592473@qq.com；

吴新桥(1979)，男，教授级工程师，研究方向为输电防灾减灾、跳闸分析等。

（责任编辑　马晓华）

 DOI：10.13335/j.1000-3673.pst.2021.0947

A Real-time Wildfire Risk Assessment Model for Transmission Corridors Based on Feature Engineering, Ensemble Learning and Model Fusion

ZHANG Keying, WU Xinqiao, ZHAO Jiguang, LIU Lan, QIN Ping, WANG Hao, ZHAN Tanbochi

(China Southern Power Grid Digital Grid Research Institute Co., Ltd, Guangdong Province, Guangzhou 510700, China)

KEY WORDS: wildfire; transmission corridors; ensemble learning; feature engineering; risk assessment

Wildfire disasters are easy to cause tripping of overhead transmission lines and threaten the stability of power grid. In this paper, a real-time wildfire risk assessment model for transmission corridor based on feature engineering and Boosting algorithm is proposed. First of all, the original data of 20 features, including human activity behavior, geography, real-time meteorology and regional historical wildfire disaster factors, which affect the occurrence of wildfire in the transmission corridor, are collected, extracted and cleaned. Then the quadratic polynomial is used to derive the features and finally generate 236 features. Considering the time complexity of model calculation, the Wrapper method combined with the five-fold cross-validation method is used to iteratively obtain the 100 features with the highest importance to construct a feature subset as the model input to construct a real-time wildfire risk assessment model of transmission corridor based on Boosting algorithm. In order to minimize the log-loss function as the optimization objective, the Bayesian optimization algorithm was used to search the parameter space and obtain the optimal model parameters. Finally, the accuracy, precision and recall of the model are verified on the test set. The model performs well with the accuracy of 96.4%, and recall of 88.1%, which can effectively evaluate the real-time risk of the transmission corridor in real time. The flow diagram is shown in Fig. 1.

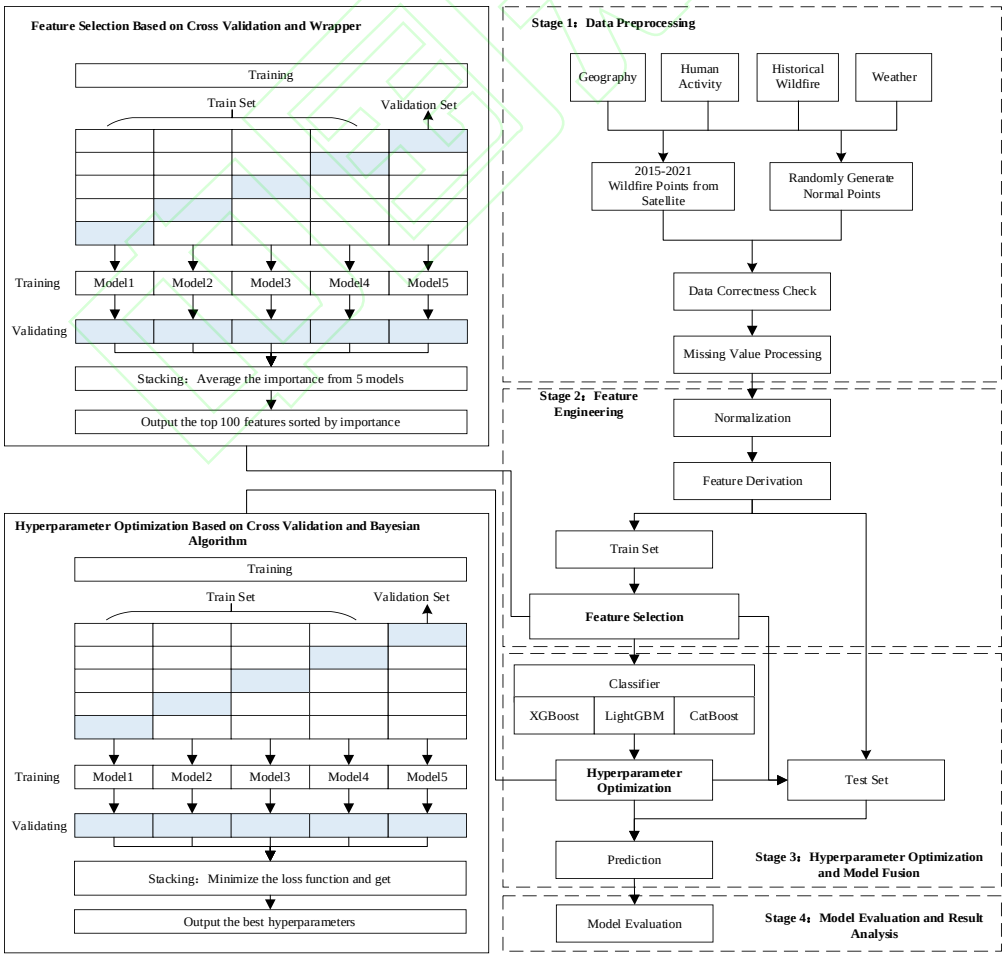

Fig. 1 Overall Flowchart