

基于事件大数据的住宅小区火灾隐患预警模型构建研究

安 全^{1,2} 倪 俊² 杨 威³ 宋 淼² 王炳贵⁴ 张 玉⁵
AN Quan NI Jun YANG Qi SONG Miao WANG Binggui ZHANG Yu

摘 要

住宅小区火灾事故严重威胁居民生命财产安全，建立一个有效的风险隐患研判模型对于预防和减少小区火灾事故发生具有重要意义。文章基于基层大联动、数字城管、12345 热线、119 警情等事件数据，融合气象、人口、房屋、环境等信息形成大数据基础底座，综合应用自然语言识别、决策树等多种机器学习算法，构建了基于事件大数据的住宅小区火灾隐患预警模型。以成都市郫都区 2023 年 1 月—2024 年 10 月期间共计 124.3 万条事件数据和小区火灾 119 消防实证数据为基础，进行模型参数确定和效果检验，结果表明该模型线下准确率 0.935、精确率 0.813，召回率 0.742，可以为基层研判小区火灾隐患提供依据支撑。模型于 2024 年 11 月上线运行，在 2024 年 11—12 月期间的实践应用结果表明线上准确率为 64%，并且通过该模型进行火灾隐患预警事件下发，有效地提高了住宅小区火灾隐患预防和排查的准确性。

关键词

事件大数据；住宅小区；火灾隐患；机器学习；预警模型

doi: 10.3969/j.issn.1672-9528.2025.07.044

0 引言

城市运行日常中存在众多潜在风险因素，及时发现和处理潜在风险是确保城市安全的关键。火灾是人们日常生活中发生最频繁的灾害之一，无时无刻不对人类的财产及生命安全造成威胁^[1]。住宅小区具有内部结构复杂、人员流动性大、设备用电量多等特点，如出现火情会对居民的生命财产安全和社会稳定构成严重威胁^[2]。2022 年全国消防救援队伍共接报火灾 82.5 万起，其中电气火灾风险最大、电动自行车火灾持续上升^[3]。

住宅小区火灾事故的发生虽然有一定偶然性，但并非没有征兆^[4-5]，小区火灾事故的发生多与小区设施特征、居民的行为习惯、近期的天气特征、历史火灾发生情况、房屋特征、居住人口特征、周边环境特征等相关，如电气老化、消防设施异常、飞线充电、高温干燥、房屋老旧等，而这些特征通常会反映在事件大数据中。本文所称事件大数据是指由基层网格员、城管队伍、市民群众等不同系统报送的各类事件，通过数据汇聚手段集成的区域全量事件大数据。本文通过融合事件大数据，综合应用自然语言识别、梯度提升决策树等机器学习算法，构建了基于事件大数据的小区火灾隐患预警模型，

并在郫都区进行实践验证，取得了较好地预警预防效果。

1 研究思路与框架

本文以成都市郫都区 2023 年 1 月—2024 年 10 月期间共计 124.3 万条事件数据和小区火灾 119 消防实证数据为基础，提出一种基于事件大数据的小区火灾隐患预警模型，实现思路与框架如图 1 所示。

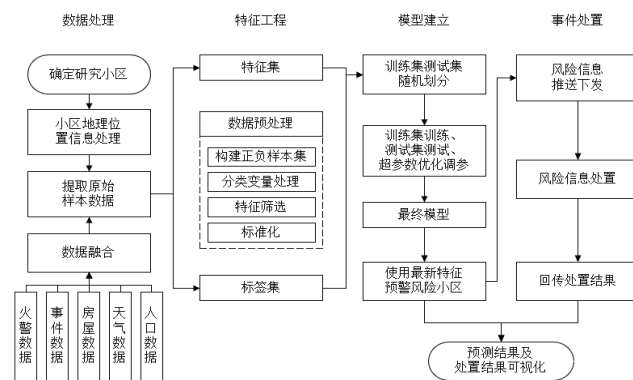


图 1 风险小区火灾隐患预警模型流程图

首先，收集数据，通过四川省成都市郫都区城市运行中心数据资源体系汇集大联动事件数据、数字城管事件数据、12345 市民热点投诉数据、119 火警数据、天气数据、人口数据、房屋数据等，通过清洗融合，形成以每天为时间维度、以各小区为空间维度的统一数据系列。其次，对数据进行预处理，通过分析火灾发生的影响因素，定义特征关键字，通过自然语言分析，从事件中提取出特征信息，同时合并天气、人口、房屋等其他特征信息，通过数据归一化等操作对提取

1. 中国电建成都勘测设计研究院有限公司 四川成都 610072
2. 郫都区智慧蓉城运行中心 四川成都 611730
3. 郫都区行政审批局 四川成都 611730
4. 成都四方伟业软件股份有限公司 四川成都 610093
5. 四川华鲲振宇智能科技有限责任公司 四川成都 610093

的特征信息进行标准化处理。最后，构建模型，运用梯度提升决策树算法（gradient boosting decision tree, GBDT）构建小区火灾隐患预警模型，根据提取的特征数据和标签数据，按 5:1 进行训练集与测试集的抽样划分^[6]，反复调优训练和检验模型，形成稳定准确的小区火灾隐患预警模型。最后，通过实时收集的城市事件大数据和相关信息，利用构建好的模型进行小区火灾风险分析和预警，形成预警信息后通过事件处置平台推送至相关的街道和社区，由街道或社区有针对性地开展隐患排查，降低火灾发生可能。

历史火灾风险研判相关研究文献中，已有多位研究者使用机器学习算法进行模型训练。其中包括贾南等人^[7]建立的森林火灾风险预警模型，徐爱明^[8]建立的单位火灾风险评估模型，以及马晓东^[9]建立的森林火灾风险等级预测模型，他们均对比了几种典型机器学习算法在火灾风险预测方面的优劣程度。其中梯度提升决策树 GBDT 算法在几篇文献中均表现优异，因此本文选用梯度提升决策树 GBDT 算法建立小区火灾隐患预警模型。

2 特征指标提取及筛选

2.1 特征指标提取

为准确评估小区火灾隐患风险，通过成因分析法从海量事件和多源数据中提取特征指标，根据因素性质不同将特征指标划分为行为特征、天气特征、火灾历史特征、人口特征、房屋特征、地区特征、管理特征 7 类 49 个维度，如表 1 所示。其中行为特征、房屋特征以及管理特征的前 6 个特征（缺乏消防知识宣传教育、电力通信设施隐患、消防设施异常、消防通道堵塞、九小场所消防隐患、群租房消防隐患）是根据海量事件的原始分类类别进行提取，当出现对应类别的事件后，该事件所属小区当天对应的特征值为 1；管理特征的其余特征（物业区域内违章搭建、闲杂人进出频繁、电动车停放杂乱、垃圾桶脏乱被引燃、电线裸露、设施老化锈蚀、电表气表损坏、房屋电梯漏水、电梯漏水、线路杂乱）是从海量事件内容信息中通过类别分类结合关键字匹配的方法进行特征提取和记录，当类别对应且在内容中识别到“火灾隐患”关键字时，对应的特征值记为 1。天气特征、人口特征直接采用融合共享的全区气象信息和人口信息，地区特征是定义小区周边 300 m 范围内是否存在加油站、医院、学校、加气站、零售药店、农贸市场等场所，通过地理坐标空间计算生成。火灾事故特征是根据历史 119 警情数据直接获取。

表 1 原始数据特征指标分类提取情况表

序号	特征类别	特征指标名称
1	行为特征	飞线充电、破坏“技防、物防、消防”设备、易燃易爆物品存放、易燃物堆放、危险用电、私接电线、危险用火、烤火、小区楼道堆物、露天焚烧

表 1（续）

序号	特征类别	特征指标名称
2	天气特征	温度、湿度、降雨量、水平能见度
3	火灾历史特征	事故历史特征
4	人口特征	人口规模、性别结构、年龄结构、家庭结构、流动人口、重点人员
5	房屋特征	围墙失修、管道破裂堵塞、墙面墙砖脱落、危房隐患、房屋出租、房屋总数
6	地区特征	加油站、医院（卫生院）、学校、加气站、零售药店、农贸市场
7	管理特征	缺乏消防知识宣传教育、电力通信设施隐患、消防设施异常、消防通道堵塞、九小场所消防隐患、群租房消防隐患、物业区域内违章搭建、闲杂人员进出频繁、电动车停放杂乱、垃圾桶脏乱被引燃、电线裸露、设施老化锈蚀、电表气表损坏、房屋电梯漏水、电梯漏水、线路杂乱

小区发生特征风险事件并非只影响当日，其影响一般会持续多日，因此本文引入了事件特征影响衰减机制，例如假设在 2024 年 9 月 9 日的事件数据中检测到包含“危险用电”类事件，并且对应匹配到的小区为“A 小区”，在特征表中，时间为 2024 年 9 月 9 日、小区为“A 小区”对应的“危险用电”特征值为 1，并且在之后的 30 天内，“A 小区 - 危险用电 - 时间衰减”不断减小，特征值为 $1/n$ ，其中 n 为当天距离 9 月 9 日的天数。

2.2 特征指标筛选

上述数据集特征指标较多，若同时作为特征变量纳入模型会造成特征冗余，对模型预测准确性造成影响，因此需要对特征进行降维。特征指标筛选是从已有的特征中选出若干最具代表性的特征，以达到数据降维的目的。本文在全面考虑 7 个方面的特征指标基础上，先使用单因素显著性检验方法，剔除对火灾事故不具备显著性影响的指标，再以 CART 决策树作为基分类器，训练机器学习模型，CART 决策树训练模型过程中可以输出各项特征分别对于是否发生火灾这一结果的重要性，依据重要性大小进行排序，保留重要性大于 0.02 的特征并纳入模型，最终形成特征指标如表 2 所示。

表 2 特征指标详情表

特征类别	特征名称
行为特征	电瓶车飞线充电 × 时间衰减 X_1
	破坏消防事件 × 时间衰减 X_2
	易燃易爆物品存放 × 时间衰减 X_3
	危险用电事件 × 时间衰减 X_4
	危险用火事件 × 时间衰减 X_5
	小区楼道堆物事件 × 时间衰减 X_6

表 2(续)

特征类别	特征名称
天气特征	温度 X_7
	相对湿度 X_8
人口特征	人口数量 X_9
	男女性别比 X_{10}
	老年人数占比 X_{11}
	未生育家庭占比 X_{12}
	流动人口占比 X_{13}
	重点关注人员占比 X_{14}
房屋特征	小区围墙失修, 服务设施老化事件 $\times$ 时间衰减 X_{15}
	管道堵塞、破裂事件 $\times$ 时间衰减 X_{16}
	危房隐患事件 $\times$ 时间衰减 X_{17}
	房屋出租率 X_{18}
地区特征	房屋总数 X_{19}
	小区周边 300 距离内是否有加油站 X_{20}
	小区周边 300 距离内是否有加气站 X_{21}
管理特征	小区周边 300 距离内是否有农贸市场 X_{22}
	近一周内是否发生消防通道堵塞 X_{23}
	近一周内是否发生消防设施异常 X_{24}
	近一周内是否存在三 / 二合一、九小场所消防隐患 X_{25}
	一周内是否收到居民火灾隐患相关投诉 X_{26}
响应变量	是否发生火警 Y

3 预警模型构建

3.1 模型原理

梯度提升决策树回归预测模型 (GBDT) 是一种集成学习算法, 模型的原理是通过单独训练多个子模型, 最后将多个子模型的结果按照一定规则进行综合, 得到一个更加准确的结果^[10]。一般结构如图 2 所示, 个体学习器通常使用机器学习的某个算法通过数据训练得到, 每个个体学习器也被称为基学习器。

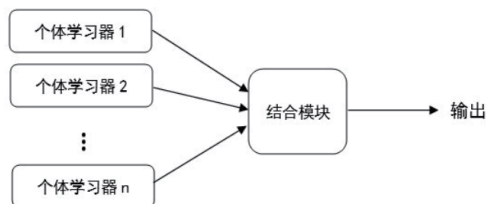


图 2 集成学习结构示意图

单个学习器往往因为数据不平衡、参数设置不当等多方面因素, 导致性能良莠不齐或是存在明显的“偏科”, 集成学习的思想是通过将多个学习器组合来实现回归或者分类, 组合后的学习器泛化性能会比单一学习器更强^[11]。

本文选用 Python3.8 中的语言机器学习库 Sklearn.ensemble 包中的 Gradient Boosting Classifier 类建立 GBDT 模型。

Boosting 是串行式集成学习的著名代表, 其核心思想是通过不断改变训练样本的权重分布, 提高分类器对样本的关注, 迭代训练后将多个弱学习器线性顺序拟合成一个强学习器, 在拟合过程中, 每个模型都将精力着眼于最难拟合的数据上, 最终训练出具有低偏置的强学习器^[12]。GBDT 是 Boosting 的代表算法之一, 它使用的决策树是 CART 树。在每一次迭代过程中, 沿当前弱学习器的损失函数负梯度方向, 训练得到新的弱学习器, 最后将所有弱学习器的输出结果按照其权重进行加强组合, 得到最终的强学习器。提升树利用加法模型和前向分布算法实现学习的优化过程^[13], 对于一般的损失函数来说, 优化过程比较艰难, GBDT 利用损失函数的负梯度来拟合基学习器, 即用负梯度替代残差^[14]。

具体算法为:

Step1: 初始化。

$$f_0(x) = \arg \min_{h_0} \sum_{i=1}^N L(y_i, h_0(x)) \quad (1)$$

式中: $L(y_i, h_0(x))$ 是损失函数; 初始化 $h_0(x)$ 是取所有样本的均值。

Step2: 迭代 t 次, 计算负梯度, 作为残差值的近似值。

$$y_i = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{t-1}(x)}, i = 1, 2, 3, \dots, N \quad (2)$$

Step3: 拟合残差近似值, 得到基学习器。

$$w_t = \arg \min_{w_t} \sum_{i=1}^N (y_i - h_t(x; w_t))^2 \quad (3)$$

Step4: 然后再计算基学习器的权重。

$$\alpha_t = \arg \min_{\alpha_t} \sum_{i=1}^N L(y_i, f_{t-1}(x_i) + \alpha_t h_t(x; w_t)) \quad (4)$$

Step5: 更新 $f_t(x)$, 并循环重复第二步。

$$f_t(x) = f_{t-1}(x) + \alpha_t h_t(x; w_t) \quad (5)$$

GBDT 的每一步残差计算增大了被错误划分样本的权重, 而被正确划分样本的权重趋于 0, 这样后面的树就能专注于那些被错误划分的样本。GBDT 算法模型结构流程如图 3 所示。相较于其他机器学习算法, GBDT 的优点在于几乎适用于所有回归预测问题, 不论是线性回归还是非线性回归, GBDT 都具有较好的预测性能, 且模型泛化能力较强^[15]。

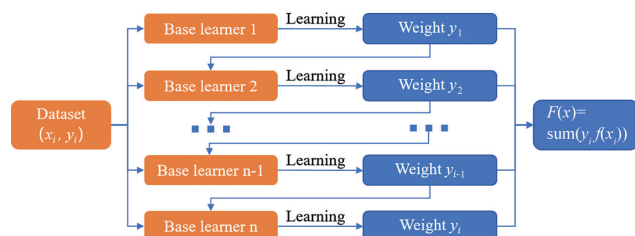


图 3 GBDT 算法模型结构流程图

3.2 模型构建

利用行为特征、天气特征、火灾事故历史特征、人口特征、

房屋特征、地区特征等数据进行小区火灾隐患研判，为保证变量之间的协方差关系以及模拟数据与实际数据的一致性，本文基于历史的实际数据进行抽样，正负样本按 5:1 比例分别随机抽样，生成模拟试验数据集，用以建立模型。

模型的学习能力和泛化能力可以通过调节树结构进行调整，即通过调节超参数从而减小偏差和方差，解决过拟合和欠拟合的问题。训练过程中，本模型使用了 5 折交叉验证的方法对模型进行训练。5 折交叉验证最终将准确度取平均值作为该模型的最终表现得分。通过这种方法确定的超参数，可以有效避免模型泛化性能力差的问题^[16]。

参数调优时，选用网格搜索的方式优化超参数^[17]，利用穷举搜索的方式对比出模型性能最好的最佳参数组合^[18]。GBDT 的分类算法中生成树的个数、树的最大深度、学习率、子采样、分割内部节点所需的最小样本数、叶子节点最少样本数这几个参数都会影响 GBDT 的分类效果。用网格搜索算法与 5 折交叉验证进行参数选择，初始设置随机森林生成树个数范围 [50, 200]，步长为 5；树的最大深度范围 [3, 6]，步长为 1；学习率范围 [0.01, 0.1, 0.5]；分割内部节点所需的最小样本数范围 [2, 12]，步长为 2；叶子节点最少样本数范围 [2, 12]，步长为 2。经过网格搜索后得到的最优参数如表 3 所示。最优参数组合下，模型训练效果最好。

表 3 GBDT 最优参数组合

参数名	参数值
n_estimators	150
max_depth	5
learning_rate	0.1
min_samples_split	6
min_samples_leaf	4

利用模型构建中的模拟数据集构建方法完成模型训练，结果显示基于 GBDT 算法构建的小区火灾隐患预警模型的准确率平均值 0.914、精确率平均值 0.82、召回率平均值 0.789，呈现出较好的风险研判效果。

4 实践应用效果

成都市郫都区位于成都市第二圈层，面积 438 km²，常住人口 139 万人，域内城乡结构复杂，小区数量众多，日常火灾风险隐患较大。据统计，2024 年 1 月—10 月，共发生小区火灾事故 250 起，严重威胁郫都区居民生命财产安全。利用本文建立的小区火灾隐患预警模型，及时识别潜在的火灾风险，并将预警结果及时推送给相关街道，相关街道有针对性地及时采取措施排查防范，将会有效降低小区火灾事故发生的可能。

本模型于 2024 年 11 月在郫都区智慧蓉城运行中心正式上线试运行，在 2024 年 11 月 1 日—12 月 30 日正式运行时间内，模型共输出 14 条预警信息，其中共有 10 例后续 15 天内真实发生了火灾事故，线上精确率约为 64%，验证了模型的有效性。具体情况如表 4 所示（其中小区名经过脱敏处理，使用代号）。

表 4 试运行期间准确预测场景

预警时间	消防警情时间	实际警情地址
2024/12/21	2025/1/3	AA 小区 5 栋 3 单元 801
2024/12/24	2024/12/28	BB 小区 1 期 4 栋 1 单元 2805 号
2024/12/27	2024/12/28	CC 小区
2024/11/24		
2024/12/19	2024/12/25	DD 小区 4 期 5 栋
2024/12/14	2024/12/25	EE 小区 3 期 12 栋
2024/12/13	2024/12/23	FF 小区 1 期 3 栋 1 单元
2024/11/23	2024/12/18	GG 小区
2024/11/23	2024/12/2	HH 小区 11 栋 3 单元
2024/11/9	2024/11/17	II 小区

5 结论

火灾事故是人类日常生活中最频发的灾害之一，住宅小区内部建筑结构、人员构成等较为复杂，一旦发生火灾事故会造成难以估量的损失。只有不断提高预防能力，从源头出发加大排查力度，才能切实降低小区火灾事故风险，维护广大人民群众的安全。本研究将机器学习算法运用到小区火灾隐患研判上来，通过汇聚的海量跨部门数据，构建出 7 类共 49 个小区火灾隐患特征，在此基础上使用 GBDT 机器学习算法构建了住宅小区火灾隐患预警模型，结果显示模型预警线下准确率为 0.914、精确率 0.82、召回率 0.789，呈现较好的风险研判效果。在实际上线运行的两个月中，火灾预警 14 次，其中 9 次预警成功，精确率为 64%，表现出了良好的预警效果。预警信息通过事件平台下发给基层小区，为相关风险隐患的排查提供了宝贵的时间窗口，具有较强的实用性。

参考文献：

[1] 杜兰萍. 火灾风险评估方法与研究案例 [M]. 北京：中国人民公安大学出版社，2011.

[2] 刘宇. 提升火灾防控消防监督管理力度的策略 [J]. 科技创新与应用，2021,11(31): 129-132.

[3] 国家消防救援局. 2022 年全国警情与火灾情况 [EB/OL].[2024-11-12].<https://www.119.gov.cn/qmxfxw/xfyw/2023/36210.shtml>.

[4] 姚正登. 现代信息技术在火灾调查中的应用 [J]. 数字通信世界，2023(3): 130-132.

（下转第 191 页）

完成复杂任务，目前已经被广泛应用于自动化任务和多场景应用中，极大提高了工作效率。Bash 已成为 Linux 环境中不可或缺的工具，为软件部署、自动化测试等提供了强大的解决方案。

本文中所提出的自定义脚本命令可视化页面的应用方法，能够解决运维人员对相关命令不熟悉等问题，从而提高运维效率。为了系统安全，防止攻击者通过管理节点对超级管理节点进行恶意操作，造成业务瘫痪问题，因此各管理节点并不能 SSH 到超级管理节点。下一步将重点考虑在集群整体性能的基础上，对各节点运行过程中的系统负载及异常情况进行可视化应用。

参考文献:

[1] 冯伟,姜远飞,孙晶,等. 高校中小型高性能集群系统的建设及管理实践 [J]. 中国现代教育装备,2024 (17):16-19.

[2] 金石声,白铁男,谭海波,等. 基于 Zabbix 的 Veritas Cluster Server 集群管理系统的监控技术 [J]. 数字技术与应用, 2023, 41(11):188-190.

[3] 贺禹鑫. 集群管理命令系统的研究与实现 [D]. 成都: 西南民族大学, 2023.

[4] 卫晓锋. 基于 Ganglia 的 Hadoop 集群监视系统研究与实现 [D]. 西安: 西安电子科技大学, 2017.

(上接第 187 页)

[5] 付小千. 城市区域火灾风险评估方法综述 [J]. 消防科学与技术, 2021, 40(11): 1622-1625.

[6] 成晓明,王磊,张鹏超,等. 区域级电力负荷数据特性研究与短期预测 [J]. 电网技术, 2022,46(3):1092-1099.

[7] 贾南,何昌原,段海鹏. 基于集成算法的森林火灾风险预警模型研究 [J]. 武警学院学报, 2021, 37(12): 5-9.

[8] 徐爱明. 基于机器学习的单位火灾风险评估系统开发研究 [D]. 南京: 东南大学, 2022.

[9] 马晓东. 基于随机森林的森林火灾风险等级预测的研究与应用 [D]. 南昌: 南昌大学, 2022.

[10] 李凯,崔丽娟. 集成学习算法的差异性及性能比较 [J]. 计算机工程, 2008(6): 35-37.

[11] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016.

[12] 邹雯. 基于改进随机森林的卷烟订购量预测的研究与应用 [D]. 南昌: 南昌大学, 2020.

[13] 王天华. 基于改进的 GBDT 算法的乘客出行预测研究 [D]. 大连: 大连理工大学, 2016.

[14] ZHANG W H, YU J Q, ZHAO A J, et al. Predictive model of cooling load for ice storage air-conditioning system by using GBDT[J]. Energy reports, 2021, 7:1588-1597.

[5] 朱质宁. 基于 Prometheus 的分布式集群监控系统设计实现及其优化 [D]. 北京: 中国地质大学, 2023.

[6] 陈旭辉,刘洋,张鸿,等. Linux 大规模集群应用管理模型的研究及实现 [J]. 电子设计工程, 2022,30(10):17-21.

[7] 陈荣鑫. 基于 NodeJS+Express 框架的学院会议室预定系统设计与开发 [J]. 信息与电脑 (理论版), 2021,33(2):95-97.

[8] 黎青霞. Node.js 在 Web 开发中的应用研究 [J]. 信息记录材料, 2024,25(10):91-93.

【作者简介】

高呈泰（1998—），男，甘肃白银人，硕士研究生，助理工程师，研究方向：气象信息技术、网络信息安全。

刘洋（1987—），男，甘肃庄浪人，硕士研究生，工程师，研究方向：气象信息技术、智能优化算法等。

陈旭辉（1968—），男，甘肃定西人，硕士研究生，高级工程师，研究方向：气象信息技术、数据挖掘等。

席晓慧（1990—），女，甘肃庄浪人，硕士研究生，工程师，研究方向：气象信息技术、智能优化算法等。

（收稿日期：2025-02-23 修回日期：2025-07-08）

[15] LIU Z, ZHANG M, LIU F, et al. Multidimensional feature fusion and ensemble learning-based fault diagnosis for the braking system of heavy-haul train[J]. IEEE transactions on industrial informatics, 2021, 17(1): 41-51.

[16] 张钧博,何川,严健等. 基于交叉验证的 XGBoost 算法在岩爆烈度分级预测中的适用性探讨 [J]. 隧道建设 (中英文), 2020, 40(S1): 247-253.

[17] YANG L, SHAMI A. On hyperparameter optimization of machine learning algorithms: theory and practice[J]. Neuro-computing, 2020, 415:295-316.

[18] 王健峰. 基于改进网格搜索法 SVM 参数优化的说话人识别研究 [D]. 哈尔滨: 哈尔滨工程大学, 2012.

【作者简介】

安全（1985—），男，陕西榆林人，硕士研究生，高级工程师，研究方向：水文水资源、人工智能等。

（收稿日期：2025-02-06 修回日期：2025-07-09）