一种基于CatBoost优化的光伏阵列故障诊断模型

彭自然[1,2],许怀顺[1,2],肖伸平[1,2*]
(1. 湖南工业大学电气与信息工程学院,湖南株洲 412007;2. 湖南省电传动控制与智能装备重点实验室,湖南株洲 412007)

摘 要: 大部分光伏电站地处偏僻、地形复杂的区域,受到外界环境的影响,易发生各种故障. 而传统的光伏阵列故障诊断方法存在精度不高以及光伏数据利用率低等问题. 针对以上问题,本文先是通过引入Levy飞行策略和步长因子动态调整策略改进麻雀搜索算法(Sparrow Search Algorithm,SSA),降低SSA算法陷入局部最优的风险,提升SSA算法的寻优能力. 然后采用改进的Levy步长调整麻雀搜索算法(Levy Adjustment Sparrow Search Algorithm,LASSA)对CatBoost模型关键超参数进行寻优,提出了一种基于CatBoost并以LASSA为优化策略的光伏阵列故障诊断模型LASSA-CatBoost,以实现光伏阵列的短路、开路、老化和阴影遮挡故障的精确诊断. 实验结果表明,LASSA-CatBoost模型的故障诊断准确率为99.7%,相较于优化前的CatBoost模型,准确率提高了3.6%. 与现有的光伏阵列故障诊断模型相比,LASSA-CatBoost模型的准确性和稳定性更高.

关键词: 光伏阵列;故障诊断;I-V特性曲线;CatBoost;Levy adjustment sparrow search algorithm
基金项目: 国家重点研发计划基金(No.2019YFE0122600);湖南省教育厅重点科研项目(No.22A0423);湖南省自科基金(No.2023JJ60267,No.2022JJ50073)
中图分类号: TM615;TP18 **文献标识码:** A **文章编号:** 0372-2112(2024)07-2418-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240236

A CatBoost Optimization-Based Fault Diagnosis Model for Photovoltaic Arrays

PENG Zi-ran[1,2], XU Huai-shun[1,2], XIAO Shen-ping[1,2*]
(1. *School of Electrical and Information Engineering*, *Hunan University of Technology*, *Zhuzhou*, *Hunan* 412007, *China*;
2. *Hunan Provincial Key Laboratory of Electric Drive Control and Intelligent Equipment*, *Zhuzhou*, *Hunan* 412007, *China*)

Abstract: Most of the photovoltaic power stations are located in remote areas with complex terrain, which are affected by the external environment and prone to various faults. The traditional PV array fault diagnosis methods have the problems of low accuracy and low utilization of PV data. Aiming at the above problems, in this paper, we first improve the sparrow search algorithm (SSA) by introducing the Levy flight strategy and the dynamic adjustment strategy of the step factor to reduce the risk of the SSA algorithm falling into the local optimum and improve the optimization ability of the SSA algorithm. Then the improved levy adjustment sparrow search algorithm (LASSA) is used to optimize the key hyperparameters of the CatBoost model, and a photovoltaic array fault diagnosis model LASSA-based on CatBoost and using LASSA as the optimization strategy is proposed. CatBoost for accurate diagnosis of short-circuit, open-circuit, aging and shadow masking faults in PV arrays. The experimental results show that the fault diagnosis accuracy of the LASSA-CatBoost model is 99.7%, which is 3.6% higher compared to the CatBoost model before optimization. Compared with the existing PV array fault diagnosis models, the LASSA-CatBoost model has higher accuracy and stability.

Key words: photovoltaic array; fault diagnosis; I-V characteristic curve; CatBoost; Levy adjustment sparrow search algorithm
Foundation Item(s): National Key R&D Program Funded Project of China (No.2019YFE0122600); Key Research Program of Hunan Provincial Department of Education (No.22A0423); Hunan Provincial Self-Science Foundation Program (No.2023JJ60267, No.2022JJ50073)

收稿日期:2024-03-14;修回日期:2024-06-13;责任编辑:覃怀银
*通讯作者:肖伸平

1 引言

如今对可再生能源的需求日益增加,光伏发电作为一种清洁、可持续的能源备受关注. 然而,光伏系统在实际运行中可能会遇到光伏组件的老化、组件之间的短路和开路、阴影遮挡和逆变器故障等各种问题,这些故障不仅影响了发电效率,还可能导致系统性能下降甚至完全失效[1,2]. 因此,光伏阵列的故障诊断成为了一个至关重要的问题.

光伏故障诊断方法主要有,一是可视外观检测:通过观察光伏组件和系统的外观,检查是否存在损坏、污垢、影响光照接收的遮挡物等. 二是电性能测试[3,4]:包括开路电压、短路电流、最大功率点等测试,通过测量这些参数来评估光伏组件和系统的性能状态. 三是红外热成像[5,6]:利用红外热像仪检测光伏组件表面的温度分布,发现温度异常区域,从而识别可能存在的故障,如电池片热点效应、接触不良等. 四是数据监测和分析[7,8]:利用监测系统采集的实时数据,如发电量、温度、辐射等,通过数据分析技术检测故障,并提供故障诊断和预警. 以上传统方法通常只能检测到一些明显的外观缺陷或表面问题,对于内部故障或隐蔽性问题的发现能力有限. 而且需要投入大量的人力物力,且耗时较长,特别是对于大型光伏阵列而言,诊断周期较长,影响了故障的及时处理和维修. 光学检测法和红外热成像识别法虽然在某些方面提高了诊断的精度,但仍然存在一定的局限性,无法满足对光伏阵列多种故障类型的全面检测需求.

近年来机器学习技术开始应用于光伏故障检测. 利用人工智能网络对多模态光伏数据进行处理和分析,实现自动化的故障检测和诊断,大幅提高了检测效率与检测精度[9,10]. 文献[11,12]利用是一种半监督学习技术,结合了已标记数据和未标记数据来进行模型训练. 但是该方法在处理复杂的光伏阵列数据时,需要大量的计算资源来训练模型. 文献[13]结合改进蜣螂优化算法和梯度提升树模型来快速诊断光伏阵列的故障,选择最佳特征子集,使用优化后的特征子集训练梯度提升树模型,以识别各故障. 但要额外的编码和预处理步骤来处理类别数据,增加了数据处理的复杂性. 文献[14]利用哈里斯鹰优化算法(Harris Hawks Optimization,HHO)优化极限学习机的参数,结合了多尺度卷积网络和时序动态特征建模,有效地诊断光伏阵列的故障. 但模型在数据量较少或数据噪音较大的情况下,存在过拟合的问题. 文献[15]提出了基于支持向量机(Support Vector Machine,SVM)的光伏故障诊断方法,具有较高的准确度和泛化能力. 但SVM本质上是一个二分类模型,不利于处理多分类问题. 文献[16]提出了基于CatBoost算法的光伏阵列故障诊断方法,将I-V曲线的部分特征进行提取,作为模型的输入向量,实现小规模训练集下不同程度故障的诊断. 但该模型在处理复杂的光伏阵列故障时,精度不高,并且I-V曲线的利用率低.

综上所述,现有的方法仅提取I-V曲线中的部分关键数据,未能全面利用其完整信息,因此可检测的故障类型有限,并且对于复杂故障的诊断准确率低. 此外,现有研究缺乏对故障程度的细致分析,尤其是对不同阴影模式的研究不足. 针对以上问题,本文充分利用I-V曲线所包含的全部信息,提出了一种基于CatBoost优化的光伏阵列故障诊断模型,用于诊断光伏阵列中短路、开路、不同程度的老化及阴影遮挡故障.

2 光伏阵列故障分析

2.1 光伏阵列模型构建及分析

本文以大唐华银湖泉太阳能发电站为例,选择最典型的排列方式构建光伏阵列仿真模型,并分析I-V曲线特征. 如图1所示,该阵列的结构由26个相同的光伏组件串联形成一条支路,16个这样的支路再并联组成整体的光伏阵列.

根据图2(a)分析,随着辐照度G的增加,短路电流I_{sc}和开路电压U_{oc}均出现升高的趋势;从图2(b)可知,随着温度T的上升,U_{oc}下降,I_{sc}上升. 结果表明,G、T对光伏阵列的I-V特征曲线有明显的影响. 从图2(c)可知,发生开路故障时,U_{oc}仅有微小变化,而I_{sc}和最大功率点P_m则有显著变化. 从图2(d)可知,短路故障发生后,I_{sc}几乎不变,而U_{oc}和P_m的变化则较为明显,且失配程度越大,变化越显著. 从图2(e)可知,当光伏组件发生异常老化时,I_{sc}和U_{oc}变化不大,但P_m的变化幅度较大. 从图2(f)可知,在不同程度的阴影遮挡的工况下,I_{sc}和U_{oc}变化较小,而P_m则发生剧烈波动,且阴影面积越大失配越明显.

2.2 光伏阵列故障诊断流程

诊断流程如图3所示,首先,对采集到的原始故障数据进行详尽的预处理工作,并对每个样本进行Z-score归一化,以确保数据的统一性和可比性. 接下来,运用Levy步长调整麻雀搜索算法(Levy Adjustment Sparrow Search Algorithm,LASSA)对CatBoost模型的关键超参数进行优化,建立高效可靠的LASSA-CatBoost诊断模型. 然后,通过在测试集上进行交叉验证训练,验证模型的稳健性和泛化能力,输出不同故障类型的识别标签,为后续实际故障诊断提供了可靠支持. 最后,利用混淆矩阵综合计算一系列相关评估指标,对诊断模型的性能进行全面深入的评估分析,以确保其在实际应用中的可靠性和准确性.

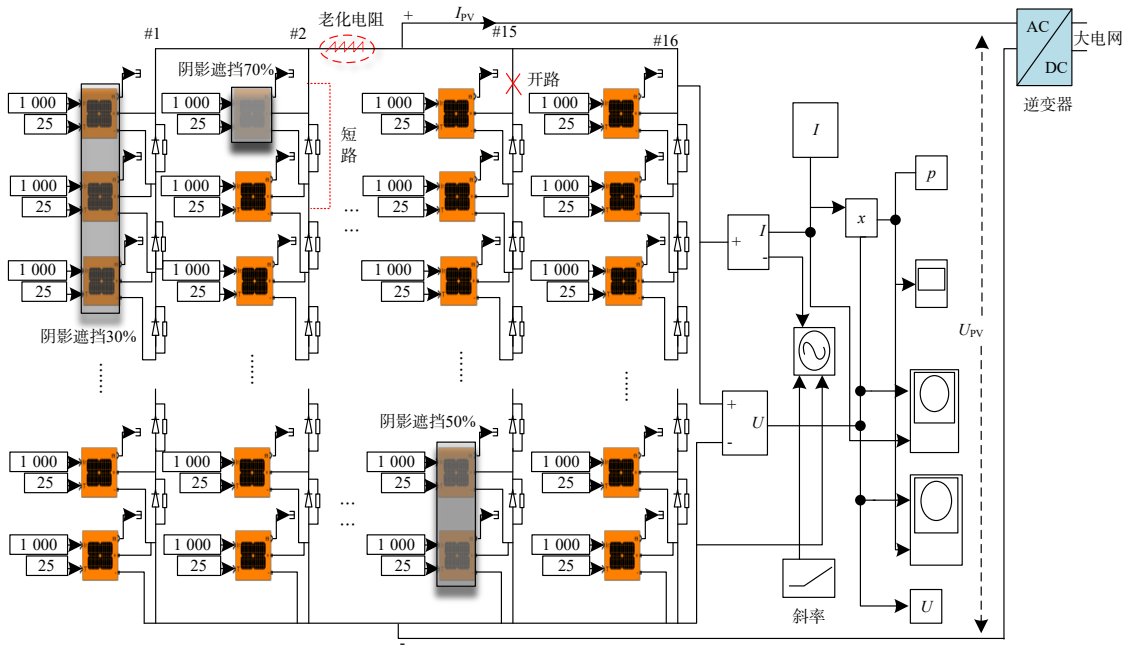


图1 光伏阵列仿真模型

3 LASSA-CatBoost诊断模型及改进

3.1 CatBoost算法原理

CatBoost是一种梯度提升树(Gradient Boosting Decision Tree,GBDT)算法,巧妙处理了单个决策树容易过度拟合的问题. 如图4所示,其核心思想是利用原始数据训练第一棵决策树作为弱学习器,然后通过迭代更新权值减小残差,直至达到设定的阈值. 在每轮迭代中,需要找到一个弱学习器,每轮的决策树选择残留减少量最大的点作为分裂点,使样本损失最小化,从而获得最优的模型[17]. 这一过程可以提升模型的性能和泛化能力.

3.2 针对麻雀搜索优化算法的改进

麻雀搜索算法(Sparrow Search Algorithm,SSA)是一种新兴的群智能优化算法,已在多个领域展现了良好的应用效果,灵感来源于麻雀在觅食时的行为. 这种算法通常应用于解决优化问题,特别是在离散空间中的问题[18,19]. 麻雀搜索算法模拟了麻雀在寻找食物时的行为,包括觅食、发现和追逐. 算法通过在解空间中进行随机搜索和局部搜索来寻找最优解. SSA算法更新方式如式(1)所示:

$$x_{i,j}(t+1)=\begin{cases} x_{\text{best}}(t)+\beta|x_{i,j}(t)-x_{\text{best}}(t)|, f_i>f_{\text{g}} \\ x_{i,j}(t)+K\left(\dfrac{|x_{i,j}(t)-x_{\text{worst}}(t)|}{(f_i-f_{\text{w}})+\text{e}}\right), f_i=f_{\text{g}} \end{cases} \tag{1}$$

式中:$x_{\text{best}}(t)$为当前全局最佳位置;β为正态分布随机数的步长控制参数;$K\in[-1,1]$为麻雀运动方向,也是步长控制参数;f_i为当前麻雀的适应度值;f_{g}和f_{w}为当前全局最优值和最差值;e为常数,避免分母为0.

麻雀搜索优化算法仍然存在一些挑战,如搜索效率低下、收敛速度较慢,以及容易陷入局部最优解的问题. 这主要源于两个方面,第一是在算法初始化阶段,麻雀种群采用随机初始化,这会影响种群的质量,低质量的初始解会降低算法的搜索效率和收敛速度[20,21]. 第二是一旦麻雀种群搜索到一个质量较高的区域,它们往往会过度聚集于此. 然而,这个区域并非一定是全局最优解所在的区域,从而导致算法陷入局部最优解的困境.

针对以上问题,提出了Levy步长调整麻雀搜索算法. 将Levy飞行策略(Levy Flight)引入式(1)中麻雀最优的位置,SSA算法会根据当前位置与麻雀最优位置新的距离来进行位置更新,改进后的SSA算法降低了麻雀陷入局部最优的风险,而且仍然能充分执行局部探索,改进如式(2)所示:

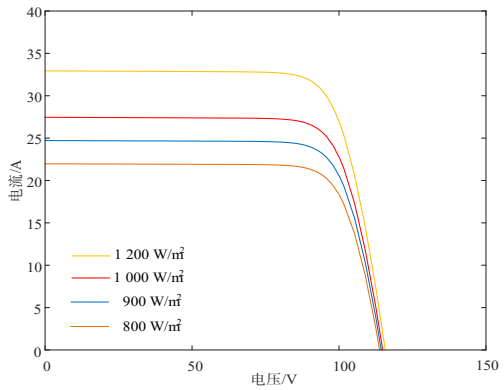
$$x_{i,j}(t+1)=\begin{cases} \text{Levy}(d)\cdot x_{\text{best}}(t)+\beta|x_{i,j}(t)-\text{Levy}(d)\cdot x_{\text{best}}(t)|, f_i>f_{\text{g}} \\ x_{i,j}(t)+K\left(\dfrac{\left|x_{i,j}(t)-x_{\text{worst}}(t)\right|}{\left(f_i-f_{\text{w}}\right)+\text{e}}\right), f_i=f_{\text{g}} \end{cases} \tag{2}$$

式(2)中,d为向量维度.

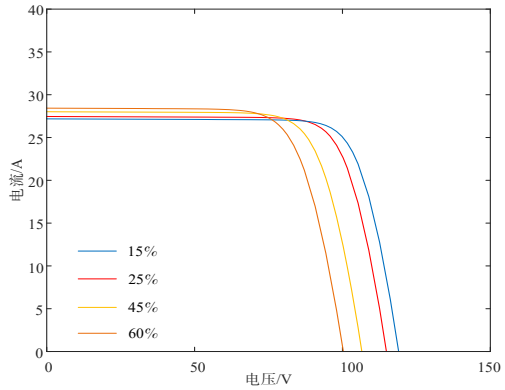
Levy计算如式(3)和式(4)所示:

$$\text{Levy}(d)=0.01\cdot\frac{r_1\cdot\sigma}{|r_2|^{1/\beta}} \tag{3}$$

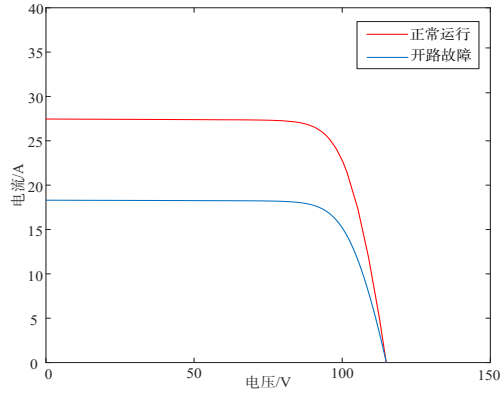
$$\sigma=\left\{\frac{X(1+\beta)\cdot\sin(\pi\beta/2)}{X[(1+\lambda)/2]\beta\cdot2^{(\beta-1)/2}}\right\}^{1/\beta} \tag{4}$$



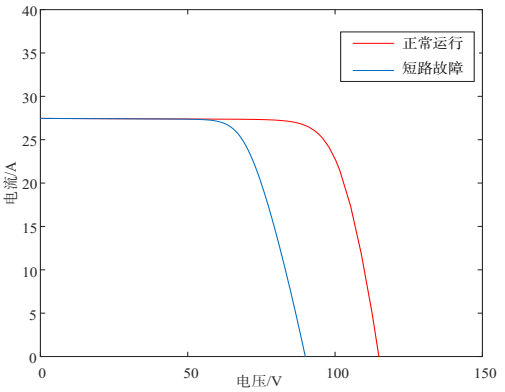
(a) 不同辐照度



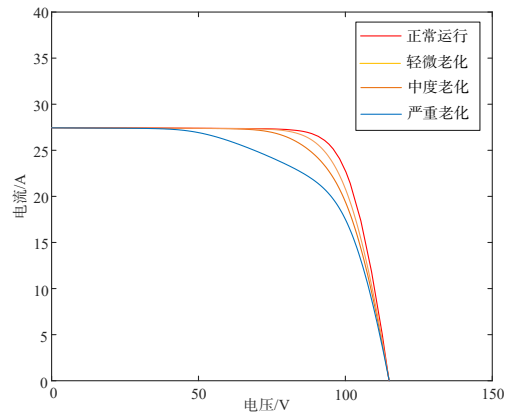
(b) 不同温度



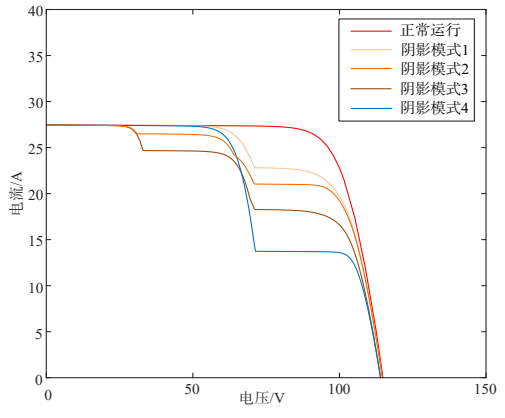
(c) 开路故障



(d) 短路故障



(e) 老化故障



(f) 阴影遮挡

图2　不同工况下的I-V特性曲线影响

其中,X为伽马函数;β为步长参数;r_1和r_2为0~1之间的随机数.

在SSA算法中,见式(1)中的步长控制参数β和K在平衡全局搜索能力与局部开发能力方面发挥重要作用,但β和K都为变量参数,无法满足算法在空间的探索,也可能导致SSA陷入局部最优,所以对步长控制参数β和K进行优化,较大的控制参数便于全局搜索,较小的控制参数促进局部开发. 对步长控制参数β和K的改进如式(5)和式(6)所示:

$$\beta=\text{fitness}_{\text{best}}-(\text{fitness}_{\text{best}}-\text{fitness}_{\text{worst}})\cdot\left(\frac{T-t}{T}\right)^{1.5} \quad (5)$$

$$K=(\text{fitness}_{\text{best}}-\text{fitness}_{\text{worst}})\cdot\exp\left[-20\cdot\tan\left(\frac{t}{T}\right)^2\right]\cdot[2\text{rand}()-1] \quad (6)$$

3.3　CatBoost参数优化

CatBoost模型的性能受很多超参数的影响,为使得

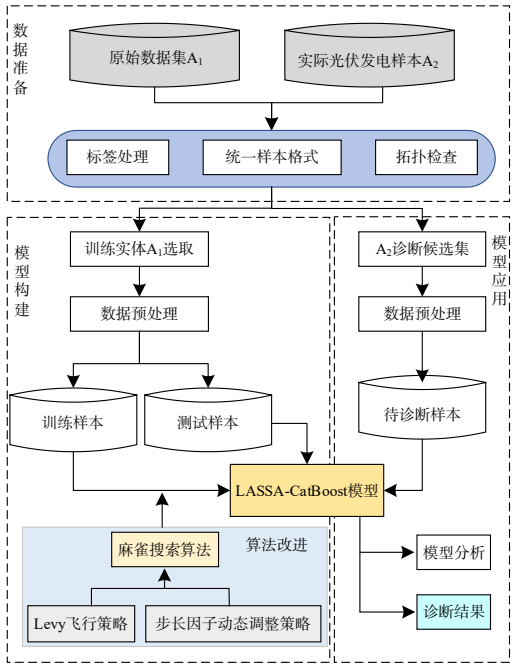


图3 光伏阵列故障诊断流程

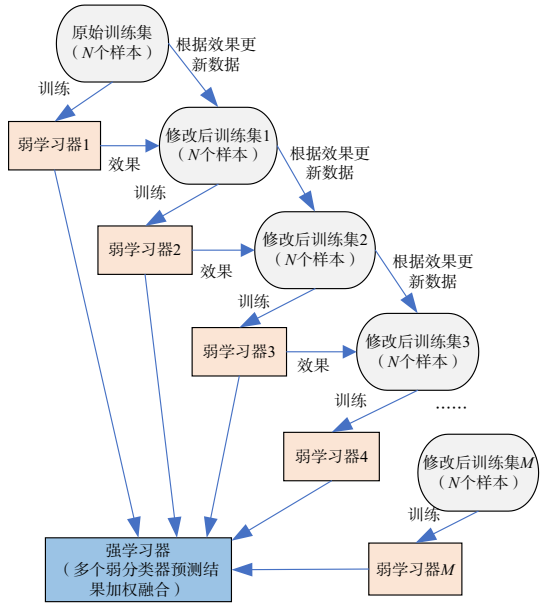


图4 CatBoost结构原理

模型收敛更快，进一步提高CatBoost模型分类的性能，本文提出利用Levy步长调整麻雀搜索算法来寻找模型的最优关键超参数. 具体被用于优化CatBoost模型的超参数包括树的数量n_{trees}及深度d_{depth}、学习率η、正则化参数λ和随机子空间r_{sm}.

麻雀搜索算法的优化过程用数学公式进行分析：设目标是最小化分类任务的损失函数$L(\theta)$，其中θ是CatBoost模型的超参数组合，定义超参数空间S，其中每个超参数θ_i都有定义域D_i. 在每次迭代中，根据目标函数的评估结果，麻雀选择一组超参数进行捕捉. 麻雀根据当前最优解θ_{best}来调整超参数，具体选择规则如式(7)所示：

$$\theta_{\text{prey}}(t)=\theta_{\text{best}}(t)+\alpha\cdot\text{rand}()\cdot(P_{\max}-P_{\min}) \tag{7}$$

其中，α是控制步长的参数，rand()是在区间[0,1)上的随机数，$P_{\max}$和$P_{\min}$分别是超参数的最大值和最小值.

初始化麻雀群体M，每只麻雀m_k表示一个超参数组合，即$m_k=(\theta_{k1},\theta_{k2},\cdots,\theta_{kn})$. 使用对应的超参数组合训CatBoost分类模型，并在验证集上计算损失函数值$L(m_k)$. 根据当前的位置m_k和邻域范围D_i，使用Levy飞行策略生成新的位置m'_k，即$m'_k=(m'_{k1},m'_{k2},\cdots,m'_{kn})$. 以一定的概率接受新位置$m'_k$，否则保持原位置不变. 对于一些较好的麻雀，根据适应度选择的策略，进行局部搜索以进一步改进超参数组合. 根据每只麻雀的适应度和更新策略，更新麻雀群体. 从最终的麻雀群体中选择具有最小损失函数值的超参数组合作为最优解θ^*. 使用最优的超参数组合θ^*训练CatBoost分类模型，并在测试集上进行性能评估.

优化流程如图5所示，算法初始化了一个包含多个麻雀的群体，每个麻雀代表了一组超参数的取值. 这些超参数包括树的数量及深度、学习率、正则化参数和随机子空间. 然后，利用当前麻雀位置构建CatBoost模型，并通过交叉验证等方法评估模型性能，得到每个麻雀的适应度值. 接下来，算法进入发现者阶段，通过更新麻雀的位置来探索搜索空间，在这个阶段，麻雀根据当前位置附近的适应度值进行更新，以便在搜索空间中寻找到最优解. 然后，算法进入加入者阶段，根据最优位置和最差适应度值引导，更新麻雀的位置，加速收敛到全局最优解. 在迭代优化的过程中，算法不断重复发现者阶段和加入者阶段的操作，直到达到预设的迭代次数或满足停止条件为止. 最终，输出最优位置对应的超参数值，即为优化后的CatBoost模型的超参数.

4 实验验证与分析

4.1 数据集准备及预处理

利用第2.1节搭建的模型采集数据集，用于训练诊断模型并分析其性能. 利用大唐华银涟源新能源湖泉光伏发电站的实际数据训练模型来分析实际中的应用效果.

本文选取了5种运行工况作为研究对象，表1展示了不同工作状态样本的数量及其对应的标签. 由于多条支路同时出现断路和短路的可能性较小，因此，在研究过程中，只考虑了单条支路的断路和短路情况. 3类老化类型为：轻度老化、中度老化和严重老化，老化电阻分别为5 Ω、10 Ω、15 Ω. 4种不同的阴影方案：方案1是将单个光伏的照度降低到50%；方案2是同一串联的

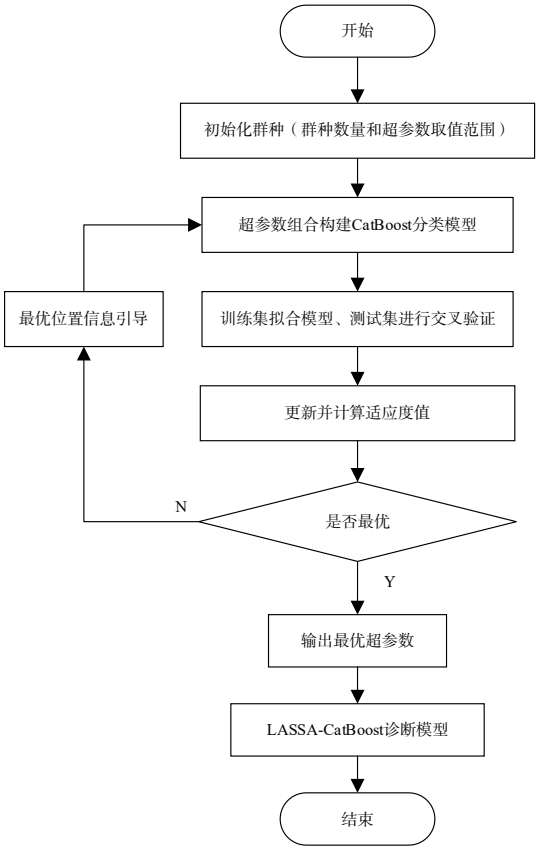


图5　LASSA算法优化CatBoost模型

3个光伏组件的照度下降到40%;方案3是每个串联组合中的其中一个组件照度下降到30%;方案4是其中一串联组的所有组件照度下降到50%,而另外一个串联组的所有组件的照度下降到70%. 对于每种工况,选取$G\in[300,1\ 200]$ W/m^2,$T\in[15,45]$ ℃,间隔5 W/m^2、1 ℃进行采样,每类运行状态采集800条作为样本,共得到4 000条伏安特性曲线,并将训练样本设为总数据集的0.8,测试样本设为0.2.

光伏阵列的I-V特征曲线易受光照和温度的影响. 为消除这些环境因素对特征曲线的干扰,使诊断模型

表1　数据集设置情况

工况	训练样本		测试样本	标签
正常状态	640		160	1
开路故障	640		160	2
短路故障	640		160	3
老化故障	轻微老化	214	53	4
	中度老化	214	53	
	严重老化	213	53	
阴影遮挡	方案1	160	40	5
	方案2	160	40	
	方案3	160	40	
	方案4	160	40	

能够更专注于捕捉与故障类型相关的信息,如图6所示,将I-V曲线数据与环境变量(温度和辐照度)结合成一个二维矩阵. 此矩阵作为诊断模型的输入,为模型提供了更加完整和准确的数据,从而增强了对光伏阵列故障的识别能力. 这种方法的优势在于,在考虑环境条件的同时,更加有针对性地分析光伏阵列的性能特征. 通过减小环境因素的影响,模型能够更好地反映与故障相关的特征,从而提高故障诊断的准确率.

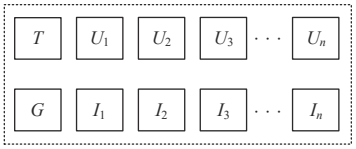


图6　GTIV矩阵

4.2　LASSA算法优化前后诊断结果与分析

CatBoost的关键超参数中,通过调整树的数量、深度以及正则化参数,可以有效提升模型在光伏阵列真实数据集上的适应性和泛化能力. 能够使得模型更加适应光伏阵列的复杂环境,并增强其对未观测数据的预测能力. 通过优化学习率,可以使模型更有效地从数据中学习特征,进而提高I-V曲线诊断的准确性. 调整随机子空间有助于模型处理不同特征之间的关联,从而避免了过度拟合现象. 同时,也能提高模型对关键特征的敏感性,使之在实际应用中表现得更为出色. 使用4.1数据集训练模型,并用Pandas库计算超参数组合与模型性能指标之间的Kendall相关系数. 图7展示了相关性热力图,用于分析超参数组合对CatBoost模型性能的影响. 结果表明,树的数量、树的深度和学习率之间的相关性最为明显,是影响诊断模型的关键超参数.

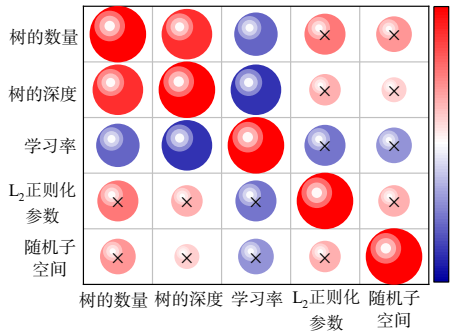


图7　超参数相关性热图

如表2所示,LASSA算法优化CatBoost模型前后的超参数值.

如图8所示,CatBoost诊断模型在训练集上优化前后的损失函数比较. 结果表明,随着迭代次数的增加,两个模型的损失函数逐渐收敛并最终趋于稳定. 但是经过LASSA算法优化的CatBoost模型最小损失函数值

表2 CatBoost参数优化表

优化参数	参数范围	默认值	LASSA寻优后
树的数量	正整数	1 000	512
树的深度	(1,16)	6	5
学习率	(0.0,1.0)	0.009	0.01
L_2正则化参数	正整数	3	3
随机子空间	(0.0,1.0]	1.0	1.0

小于0.1,且收敛速度更快,误差更小. 说明LASSA算法有效提升了模型的性能.

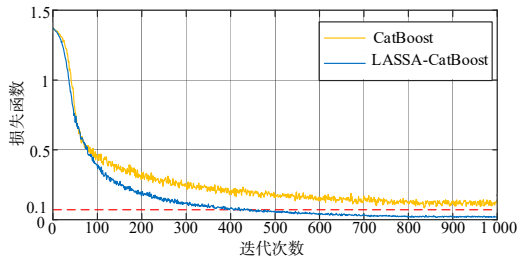


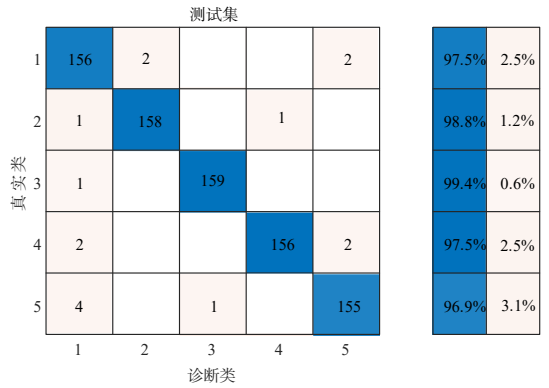
图8 CatBoost和LASSA-CatBoost训练集损失函数曲线

为了直观展示LASSA算法优化CatBoost模型前后的诊断性能,如图9所示,通过混淆矩阵可视化测试集中5种状态的诊断结果,并计算查全率R_c、精确率P_{re}、F_1评分和AUC等评估指标,其中F_1评分如式(8)所示:

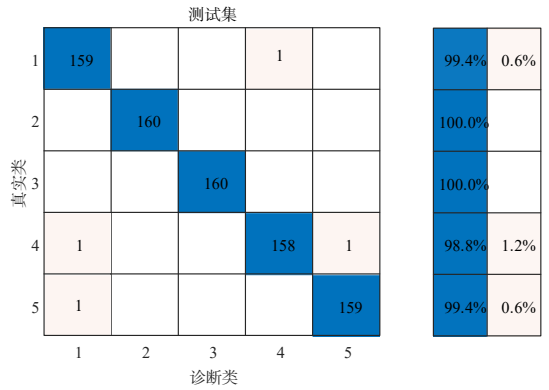
$$F_1=\frac{2P_{re}R_c}{P_{re}+R_c}\times 100\% \tag{8}$$

如图10(a)所示,查全率相比优化前有所提高,尤其提升了在正常状态和阴影遮挡状态下的性能,分别提升了1.9%和2.5%. 如图10(b)所示,根据精确率进行分析,在老化、开路和短路状态下,优化前后模型的性能相近,但对于正常状态和阴影遮挡状态下的复杂故障,LASSA-CatBoost模型的性能分别提升了3.7%和1.9%. 如图10(c)所示,通过F_1评分的对比,LASSA-CatBoost模型在正常、老化和阴影遮挡状态下的诊断效果有明显的改善. 总体而言,LASSA算法提升了CatBoost模型对各故障的诊断效果,整体性能进步明显,并且对于复杂的故障(老化故障和阴影遮挡)仍能实现较高的准确度.

LASSA-CatBoost诊断模型的ROC曲线如图11所示. 结果表明,LASSA-CatBoost模型阵列间开路和短路的AUC为1,正常、老化和阴影遮挡故障的AUC为0.98、0.98和0.99,说明LASSA-CatBoost诊断模型的总体性能已经达到诊断要求. 对于复杂故障的AUC稍低的原因是,光伏发电系统在运行时面临着复杂多变的环境,尤其在低辐照度条件下,故障特征的改变幅度很小,容易被误认为是常态,从而影响了系统的整体精度.



(a) CatBoost混淆矩阵



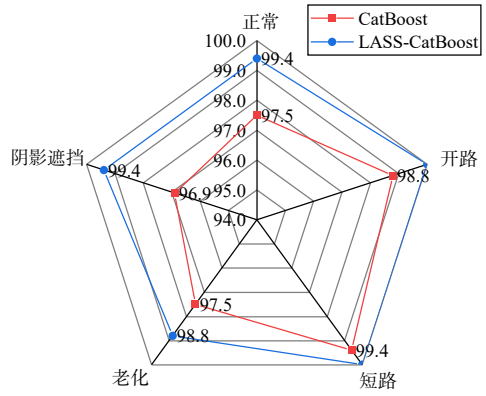
(b) LASSA-CatBoost混淆矩阵

图9 测试集混淆矩阵

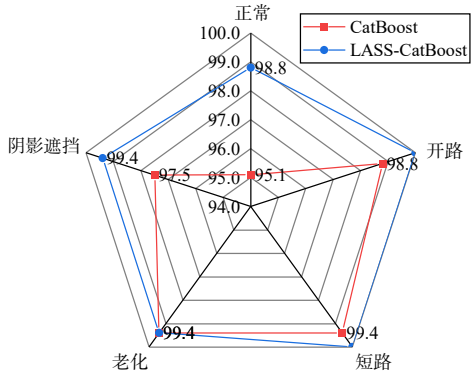
4.3 不同输入方式的诊断效果分析

为了验证全I-V曲线在光伏故障诊断中是否优于提取部分特征数据的方法. 首先,将4.1节中采集到4 000条伏安特性曲线样本进行特征提取,包括开路电压U_{oc}、短路电流I_{sc}和最大功率点P_m,以及对应的电流I_m和电压U_m等关键特征参数与辐照度和温度组合成一维矩阵. 其次,为了保证实验的准确性,同样将提取的特征进行Z-score归一化. 最后,将经过预处理的样本按照8:2的比例分配到训练集和测试集中,并将它们输入到LASSA-CatBoost模型进行训练和交叉验证. 如图12(a)所示,为测试集输出的混淆矩阵.

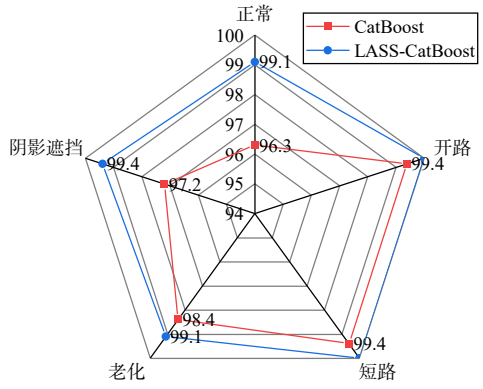
如图12(b)所示,结果表明,两种方法对于短路和开路两种故障的诊断效果相近,查全率都达到了94%以上,但是在复杂的情况下(老化故障和阴影遮挡),利用完整I-V曲线训练诊断模型有明显的优势,相较于使用部分特征数据的方法,分别提升了11.9%和11.3%. 相比之下,全I-V曲线的方法(直接I-V)提供了更详细和多维的数据支持,帮助模型更好地学习和识别复杂的故障模式,这有助于提升诊断系统的性能.



(a) 查全率



(b) 精确率



(c) F_1评分

图10　不同运行状态下两模型性能度量指标的分析结果

4.4　不同算法对优化CatBoost的性能分析

本文选取对CatBoost优化有明显效果的算法进行分析,比较不同优化算法对CatBoost模型性能的影响,以损失函数曲线为指标进行分析. 如图13所示,包括蝙蝠优化算法(Bat Algorithm, BA)、粒子群优化算法(Particle Swarm Optimization, PSO)、蚁群优化算法(Ant Colony Optimization, ACO)以及论文中提出的LASSA算法. 损失函数曲线能够直观地反映模型的拟合情况和优化效果.

实验结果表明,通过不同优化算法优化后的CatBoost模型在损失函数曲线上呈现出不同的特征. 具体而言,LASSA算法能够在较少的迭代次数下达到较低的损失函数值,表现出了更快的收敛速度和更好的拟合效果. 相比之下,其它优化算法虽然也能够使损失函数值逐渐降低,但其收敛速度和最终拟合效果不如

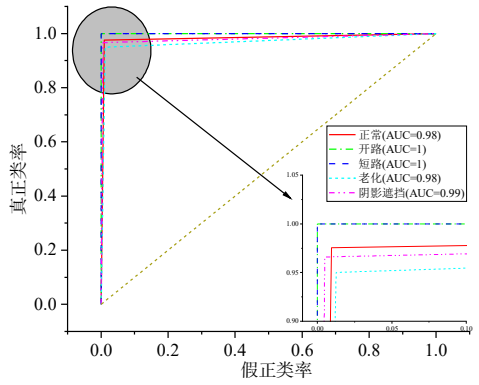
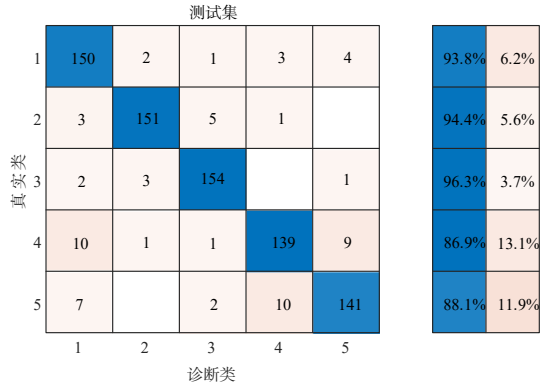
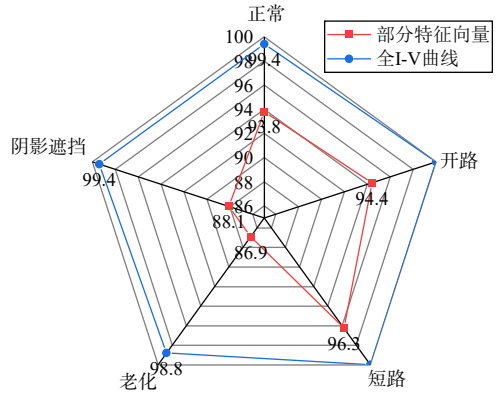


图11　LASSA-CatBoost模型在不同状态下的ROC曲线



(a) 测试集混淆矩阵



(b) 不同状态下的R_c对比

图12　不同运行状态下两输入方式性能分析

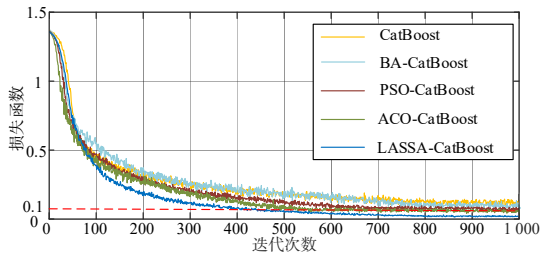


图 13 不同算法优化 CatBoost 损失函数曲线

LASSA 算法明显. 综上所述,通过损失函数曲线的分析,证明了 LASSA 算法在优化 CatBoost 性能方面的显著优势,有助于提高诊断模型的性能和泛化性.

4.5 使用实验数据诊断分析

为验证 LASSA-CatBoost 模型在实际应用中的性能表现,本研究采用湖泉光伏发电站实际运行数据进行试验,该数据包括不同天气条件下光伏阵列的运维性能,每种运行状态下获得 1 600 个样本,总计获得 8 000 个样本. 如图 14 所示,是现场的实际装置和部分实物.

图 14 大唐华银涟源新能源湖泉光伏发电站现场及装置

将 5 600 组样本作为训练集,剩余的 2 400 组样本作为验证集,利用 LASSA-CatBoost 算法进行模型的训练和验证. 如图 15 所示,是验证集下各故障诊断准确率. 实验结果表明,其中有 2 393 个样本分类正确,2 个正常工况的样本被诊断为老化故障;2 个老化故障样本被诊断为正常的工况;3 个阴影遮挡的故障样本被诊断为老化故障. 对于光伏阵列的各故障诊断准确率分别为 99.6%、100%、100%、99.6% 和 99.4%,模型的整体的准确率为 99.7%.

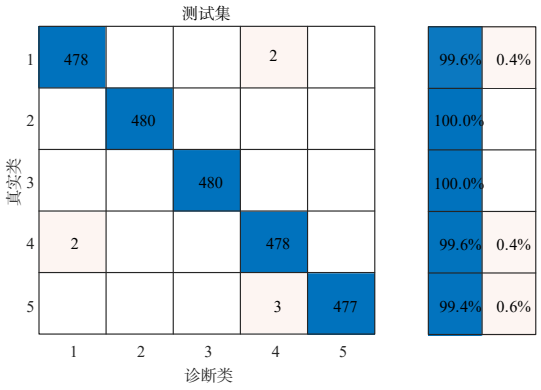


图 15 使用现场测量数据的测诊断结果

将本文的模型与传统的神经网络和其它机器学习模型进行比较,包括反向传播神经网络(BackPropagation Neural Network, BPNN)、长短期神经网络(Long Short-Term Memory, LSTM)、卷积神经网络(Convolutional Neural Network, CNN)和极限梯度提升算法(Extreme Gradient Boosting, XGBoost). 根据当前参考文献中的方法,将部分特征向量输入到神经网络诊断模型中,将 GTIV 向量作为决策树诊断模型的输入.

由表 3 可见,从整体分析,基于决策树诊断模型的查全率高于神经网络诊断模型,这表明基于决策树的机器学习模型比传统的神经网络模型在分类问题上表现更加出色. 根据复杂故障中的老化和阴影的查全率分析,CatBoost 模型相较于其它的模型具有显著优势,查全率可达到 95% 和 96%,并且经过 LASSA 算法优化后的 CatBoost 模型的查全率分别提升了 4% 和 3%. 模型整体的准确率相较于优化前提升了 3.6%.

综上所述,本次实验验证了 CatBoost 模型在故障分类问题上具有良好的表现,有较高的准确率. 利用 LASSA 算法进一步提升了 CatBoost 模型在光伏阵列故障诊断中的性能. LASSA-CatBoost 模型即使在复杂的故障(老化和阴影遮挡)情况下,仍可以实现较高精度的分类,模型具有良好的稳定性. 并且也验证了利用全 I-V 曲线进行故障诊断的可行性以及准确性.

表 3 各算法诊断结果

模型类别	模型	输入方式		查全率/%					诊断准确率/%
		GTIV	部分特征	正常	开路	短路	老化	阴影遮挡	
神经网络	BPNN		√	82	88	92	86	93	89.6
	CNN		√	94	98	97	93	94	96.5
	LSTM		√	91	95	95	93	90	93.1
决策树	XGBoost	√		94	97	95	92	91	95.9
	CatBoost	√		97	96	98	95	96	96.1
	LASSA-CatBoost	√		99	100	100	99	99	99.7

5 结论与展望

本文通过Levy步长调整麻雀搜索算法,对CatBoost模型的迭代次数、树的深度和学习率等超参数进行优化,在此基础上提出了LASSA-CatBoost光伏阵列故障诊断模型,解决了现有光伏阵列故障模型性能差以及对于复杂故障诊断准确率低的问题. 通过实验验证,相较于其它的优化算法,LASSA算法提升CatBoost模型性能的效果更加明显,模型的收敛速度更快,损失函数值更小. 并且与传统的光伏阵列故障诊断模型相比,LASSA-CatBoost有较高的准确率和稳定性. 但该模型仍然存在过拟合风险,对大数据集的扩展性有限,模型的实际应用仍处在基础阶段,后续将针对这些问题进一步展开研究.

参考文献

[1] FIRTH S K, LOMAS K J, REES S J. A simple model of PV system performance and its use in fault detection[J]. Solar Energy, 2010, 84(4): 624-635.

[2] CHEN X, GAO W, HONG C, et al. A novel series arc fault detection method for photovoltaic system based on multi-input neural network[J]. International Journal of Electrical Power & Energy Systems, 2022, 140: 108018.

[3] CHINE W, MELLIT A, PAVAN A M, et al. Fault diagnosis in photovoltaic arrays[C]//2015 International Conference on Clean Electrical Power (ICCEP). Piscataway: IEEE, 2015: 67-72.

[4] MELLIT A, TINA G M, KALOGIROU S A. Fault detection and diagnosis methods for photovoltaic systems: A review[J]. Renewable and Sustainable Energy Reviews, 2018, 91: 1-17.

[5] 胡国兵, 赵敦博, 杨莉, 等. 基于自相关函数图特征的频谱感知算法研究[J]. 电子学报, 2023, 51(5): 1327-1333.
HU G B, ZHAO D B, YANG L, et al. Research on spectrum sensing based on graphical feature of the autocorrelation[J]. Acta Electonica Sinica, 2023, 51(5): 1327-1333. (in Chinese)

[6] PEI T T, LI L, ZHANG J F, et al. Module block fault locating strategy for large-scale photovoltaic arrays[J]. Energy Conversion and Management, 2020, 214: 112898.

[7] PROKHORENKOVA L, GUSEV G, VOROBEV A, et al. CatBoost: Unbiased boosting with categorical features[J]. Advances in Neural Information Processing Systems, 2018, 31: 6638-6648.

[8] MADETI S R K. A monitoring system for online fault detection in multiple photovoltaic arrays[J]. Renewable Energy Focus, 2022, 41: 160-178.

[9] WANG M, XU X, YAN Z Y. Online fault diagnosis of PV array considering label errors based on distributionally robust logistic regression[J]. Renewable Energy, 2023, 203: 68-80.

[10] VERGURA S. Correct settings of a joint unmanned aerial vehicle and infrared camera system for the detection of faulty photovoltaic modules[J]. IEEE Journal of Photovoltaics, 2021, 11(1): 124-130.

[11] MOMENI H, SADOOGI N, FARROKHIFAR M, et al. Fault diagnosis in photovoltaic arrays using GBSSL method and proposing a fault correction system[J]. IEEE Transactions on Industrial Informatics, 2019, 16(8): 5300-5308.

[12] 徐先峰, 李芷菡, 刘状壮, 等. 基于半监督学习标签传播-极端随机树算法的光伏阵列故障诊断及定位[J]. 电网技术, 2023, 47(3): 1038-1047.
XU X F, LI Z H, LIU Z Z, et al. Fault diagnosis and localization of photovoltaic arrays based on semi-supervised learning label propagation-extra tree algorithm[J]. Power System Technology, 2023, 47(3): 1038-1047. (in Chinese)

[13] 吴亚钧, 王璐, 张金江. 基于IDBO-LightGBM的光伏阵列故障诊断方法[J/OL]. 电源学报, 2024, 1-15. DOI: https://link.cnki.net/urlid/12.1420.
WU Y J, WANG L, ZHANG J J, Fault diagnosis method of photovoltaic array based on IDBO-LightGBM[J/OL]. Journal of Power Supply, 2024, 1-15. DOI: https://link.cnki.net/urlid/12.1420. (in Chinese)

[14] 钱亮, 黄伟, 杨建卫. 基于HHO-ELM的光伏阵列故障诊断方法研究[J]. 电源技术, 2024, 48(2): 345-350.
QIAN L, HUANG W, YANG J W. Research on fault diagnosis method of photovoltaic array based on HHO-ELM[J]. Chinese Journal of Power Sources, 2024, 48(2): 345-350. (in Chinese)

[15] WANG J, GAO D, ZHU S, et al. Fault diagnosis method of photovoltaic array based on support vector machine[J]. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2023, 45(2): 5380-5395.

[16] 顾崇寅, 徐潇源, 王梦圆, 等. 基于CatBoost算法的光伏阵列故障诊断方法[J]. 电力系统自动化, 2023, 47(2): 105-114.
GU C Y, XU X Y, WANG M Y, et al. CatBoost algorithm based fault diagnosis method for photovoltaic arrays[J]. Automation of Electric Power Systems, 2023, 47(2): 105-114. (in Chinese)

[17] 郭步豪. 基于梯度提升机器学习算法的ECG身份识别[D]. 长春: 吉林大学, 2020.

GUO B H. ECG Identification Based on Gradient Enhancement Machine Learning Algorithm[D]. Changchun: Jilin University, 2020. (in Chinese)

[18] MEHMOOD A, SHER H A, MURTAZA A F, et al. A diode-based fault detection, classification, and localization method for photovoltaic array[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12.

[19] 张长胜, 张健忠, 钱斌, 等. 多策略融合的改进天鹰优化算法[J]. 电子学报, 2023, 51(5): 1245-1255.
ZHANG C S, ZHANG J Z, QIAN B, et al. Improved aquila optimization based on multi-strategy integration[J]. Acta Electonica Sinica, 2023, 51(5): 1245-1255. (in Chinese)

[20] ZHOU J, DAI Y, HUANG S, et al. Proposing several hybrid SSA—Machine learning techniques for estimating rock cuttability by conical pick with relieved cutting modes[J]. Acta Geotechnica, 2023, 18(3): 1431-1446.

[21] 张浩, 胡昌华, 杜党波, 等. 多状态影响下基于Bi-LSTM网络的锂电池剩余寿命预测方法[J]. 电子学报, 2022, 50(3): 619-624.
ZHANG H, HU C H, DU D B, et al. Remaining useful life prediction method of lithium-lon battery based on Bi-LSTM network under multi-state influence[J]. Acta Geotechnica, 2022, 50(3): 619-624. (in Chinese)

作者简介

彭自然 男, 1969年10月出生于湖南省益阳市. 现为湖南工业大学电气与信息工程学院副教授、硕士生导师. 研究方向为人工智能、信号处理、智能检测仪表、智能移动终端等方面.
E-mail: pengziran@hut.edu.cn

许怀顺 男, 1999年7月出生于山东省临沂市. 现为湖南工业大学电气与信息工程学院硕士研究生. 研究方向为人工智能、机器学习、光伏故障诊断等方面.
E-mail: 1106238410@qq.com

肖伸平 男, 1965年5月于湖南省永州市, 现为湖南工业大学电气与信息工程学院教授、博士生导师. 研究方向为时滞系统鲁棒控制理论及应用、电力时滞系统稳定性分析、工业网络控制、智能控制、过程控制等方面.
E-mail: xsp@hut.edu.cn