DOI: 10.19650/j.cnki.cjsi.J2412838

深度优化的集成学习模型 EKSSA-CatBoost：实现光伏阵列故障高精度智能诊断*

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

摘 要：光伏阵列在运行过程中，可能会受到多种因素的影响，导致不同类型的故障。通过机器学习算法，可以实现光伏阵列数据的实时监测、故障诊断和预测性维护，这种方法不受地理环境的限制，能够提高系统的可靠性和效率。光伏阵列的电流-电压(*I-V*)曲线是一项重要的指标，包含了大量关于光伏组件健康状况的信息，对于及时发现故障、评估健康状况至关重要。然而，现有方法只对来自 *I-V* 曲线的部分信息提取进行诊断分析，没有更深入地挖掘 *I-V* 曲线中的所有信息，能检测到的光伏阵列故障十分有限。针对以上问题，首先提出一种 *I-V* 曲线校正算法用于修正辐照度和温度对同一故障类型特征表现的影响，有效消除环境变量对故障特征表征的耦合效应。然后，利用 CatBoost 模型实现光伏阵列小样本高精度的实时故障智能诊断，并且利用麻雀搜索算法对模型的关键超参数进行优化。最后，为了进一步提升麻雀搜索算法的寻优能力，通过引入融合精英反向学习策略和柯西高斯变异策略改进麻雀搜索算法，使其在优化 CatBoost 模型中达到最佳效果。结果表明，利用模拟数据和现场数据分别进行模型的训练及故障诊断，测试集出现仅一个和两个误诊的样本，深度优化的集成学习模型 CatBoost 的分类准确率均达到 99.9%。

关键词：光伏阵列；故障诊断；电流-电压曲线；机器学习；CatBoost
中图分类号：TH7 TM615 **文献标识码**：A **国家标准学科分类代码**：470.4017

Deeply optimized integrated learning model EKSSA-CatBoost: Towards highly accurate intelligent diagnosis of PV array faults

Peng Ziran[1,2], Xu Huaishun[1,2], Xiao Shenping[1,2], Pan Changning[1,2]
(*1. School of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou 412007, China;*
2. Hunan Provincial Key Laboratory of Electric Drive Control and Intelligent Equipment, Zhuzhou 412007, China)

Abstract: Photovoltaic (PA) arrays may be affected by a variety of factors during operation, leading to different types of failures. Real-time monitoring, fault diagnosis, and predictive maintenance of PV array data can be realized through machine learning algorithms, an approach that is not limited by geography and can improve system reliability and efficiency. The current-voltage (*I-V*) curve of a PV array is an important metric that contains a great deal of information about the health of the PV module, which is crucial for timely fault detection and health assessment. However, existing methods only extract part of the information from the *I-V* curve for diagnostic analysis, without digging deeper into all the information. As a result, the range of detectable PV array faults remains limited. To address the problems, an *I-V* curve correction algorithm is proposed to correct the effects of irradiance and temperature on the characterization of the same fault type, effectively eliminating the coupling effect of environmental variables on the characterization of fault features. Then, the CatBoost model is used to realize real-time, high-accuracy fault intelligent diagnosis of PV arrays with small samples. The model's key hyperparameters are optimized using the sparrow search algorithm. Finally, in order to further enhance the optimization ability of the sparrow search algorithm, the sparrow search algorithm is improved by introducing the fusion elite inverse learning strategy and the Cauchy Gaussian variation strategy, so that it achieves the best effect in optimizing the CatBoost model. The results show that when using simulated data for model training and field data for fault diagnosis, only one and two misdiagnosed samples appear in the test set,

收稿日期：2024-05-14 Received Date: 2024-05-14
*基金项目：湖南省教育厅重点科研项目(22A0423)、湖南省自科基金项目(2022JJ50073)资助

respectively. The classification accuracy of the deeply optimized integrated learning model CatBoost reaches 99.9% in both cases, demonstrating its exceptional diagnostic performance.

Keywords: photovoltaic array; fault detection; current-voltage curve; machine learning; CatBoost

0　引　　言

传统的光伏阵列故障诊断方法通常依赖于人工检查或简单的监测设备,这种方法存在着诊断速度慢、准确性低等缺点。为了解决这些问题,近年来基于全电流-电压(*I-V*)曲线和机器学习的光伏阵列故障诊断方法逐渐引起了研究者的关注。基于对光伏组件在不同环境参数下的电流-电压特性曲线的系统测量,能够全面解析其运行特性[1]。在此技术框架下,常规诊断方法主要涵盖:人工可视检查和外观检测、光学检测、红外热成像识别和电子学分析[2]。现有方法在光伏系统故障识别层面存在显著效能瓶颈,首先在检测维度方面,主流技术多聚焦于表观异常特征识别,对于材料老化引发的电性能衰减或电池单元微观结构损伤等深层故障机制缺乏有效诊断能力[3]。其次传统诊断流程需配置专业技术人员实施密集型人工排查,造成时间成本高昂,形成故障响应迟滞现象[4]。

在新能源技术迭代加速的背景下,光伏系统健康状态监测领域正经历以深度学习为核心的技术革新浪潮。基于深度神经网络架构的智能诊断系统,通过融合电致发光图像、热斑分布图谱及 *I-V* 特性曲线等多源异构数据,构建起动态特征提取与故障模式识别的双重解析机制。文献[5]是通过基于图的半监督学习(group-based semi-supervised learning,GBSSL)方法对光伏阵列进行故障诊断,将有标签与无标签数据相融合,提高了模型的性能,但当涉及到大规模的光伏阵列数据时,其所需的计算资源将会增加。文献[6]提出的通过故障参数辨识实现光伏阵列故障诊断的技术方案,在标准工况下能够显著提升诊断结果的精确性与稳定性。然而该方案在复杂运行条件下存在一定局限性,特别是对于未在特征参数库中定义的隐性故障模式,无法通过现有参数匹配机制进行有效检测,导致诊断结果存在漏判风险。文献[7]基于直接利用电流重采样向量或通过 Gramian 角差场或递归图进行变换,提取断层特征。然后分别利用机器学习技术对光伏阵列进行分类,提高了准确率的同时也增加了数据处理的复杂性。文献[8]提出了基于支持向量机的故障诊断方法,利用数据采集和特征提取对模型训练实现故障诊断,在处理大规模数据集时,支持向量机(support vector machine, SVM)模型的计算复杂度与存储开销会显著增加。文献[9]提出的基于贝叶斯框架与最小二乘支持向量机的融合诊断模型,利用先验知识和观测数据来进行推断,从而提高模型的鲁棒性和可靠性,但如果先验知识不准确或不完备,可能会影响到模型的性能和诊断结果。文献[10]提出了堆叠自动编码器和 *I-V* 曲线聚类的光伏阵列故障诊断方法,通过对 *I-V* 曲线进行聚类分析,但在无监督预训练阶段,如果数据量不足或者数据质量不好,会影响到模型的性能。文献[11]提出采用半监督学习框架,有效融合少量标注数据与海量无标注历史数据,在降低人工标注成本的同时提升模型的环境适应性。但对于大规模数据集,模型的训练时间可能会比较长,特别是在使用复杂模型时,需要消耗大量的计算资源。

实际 *I-V* 曲线中包含了多维度光伏阵列的健康状况信息,上述的人工智能诊断方法只将 *I-V* 曲线中部分关键的数据进行提取,未能充分利用其中所包含的全部信息,所以可检测到的故障非常有限。此外,现有的研究只考虑了相对简单的故障,而没有细致考虑不同程度的故障,尤其是忽略了不同程度的阴影遮挡模式[12]。当前基于 *I-V* 曲线的公共训练样本集极为稀缺,现有的训练样本集标签数据有限,通用深度学习模型训练强度不够,对这类复杂的故障类型进行诊断,很难达到高的准确率[13]。为解决上述问题,提出一种基于 EKSSA-CatBoost 的智能故障诊断方法,该方法充分利用了光伏阵列的 *I-V* 曲线。主要贡献包括:

1)针对光伏阵列在不同环境条件下(如温度 *T* 和光照强度 *G*)导致的 *I-V* 曲线表现差异,在 IEC 60891 文件给出的 *I-V* 曲线校正算法的基础上,提出了一种新的校正算法,通过对环境因素的动态调整,提高了光伏阵列在不同条件下的诊断可靠性,具有较好的适应性和实用性。

2)提出的智能诊断方法基于 CatBoost 模型,并通过麻雀搜索算法(sparrow search algorithm,SSA)进行超参数优化。具体通过对 CatBoost 模型的关键参数(如树的深度、学习率等)的精确调优,进一步提升了诊断精度。SSA 优化过程在于它能够有效避免常见的过拟合问题,同时通过探索广泛的参数空间,确保诊断模型的稳健性和高效性。

3)为了进一步提升优化效果,在 SSA 的基础上引入了精英反向学习策略和柯西高斯变异策略,形成了改进的麻雀搜索算法(enhanced and K-improved SSA, EKSSA)。主要体现在通过引入反向学习和变异策略,避免了 SSA 算法陷入局部最优解的局限性,增强了全局搜索能力和算法的收敛速度。这使得在 CatBoost 模型的超参数优化中,EKSSA 能够更有效地找到全局最优解,显著提高了模型的性能。

1 光伏阵列故障分析

利用大唐华银湖泉地区的实际观测资料，在 Matlab/Simulink 环境下构建了光伏阵列的数学模型。如图 1 所示，使用了一个 16×26 的光伏电池阵列，也就是 26 个光伏模块组成一条分支，总共 16 条分支并联。通过分析光伏阵列的 *I-V* 曲线特性，深入理解实际系统的工作情况，并为故障诊断提供理论依据[14]。

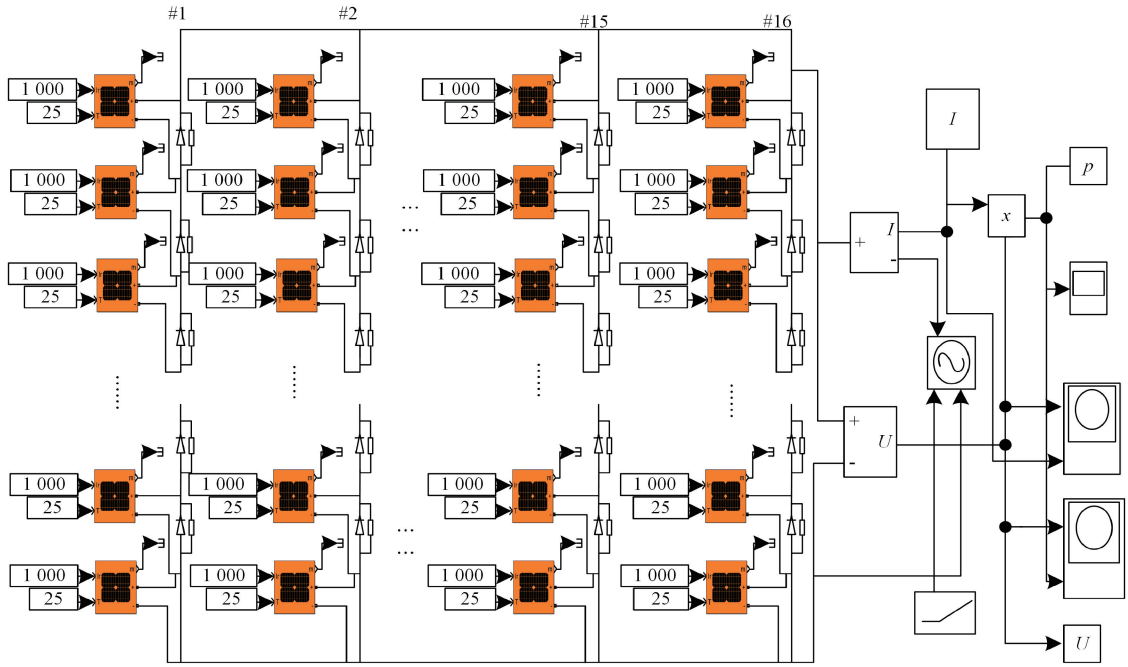


图 1 光伏阵列仿真模型

Fig. 1 Simulation model of photovoltaic array

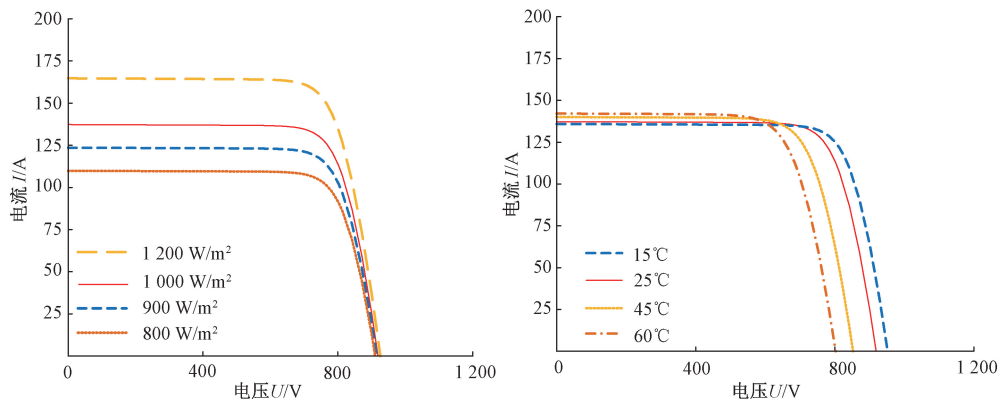
如表 1 所示，在标准工作状态（25℃、1 000 W/m^2）下，电池组件的技术和光伏电池阵列系统参数[14]。

从图 2(a)可以看出，随着辐照度 G 的增加，短路电流 I_{sc} 以及开路电压 U_{oc} 均有升高的趋势；如图 2(b)所示，随着温度的上升，U_{oc} 减小，I_{sc} 增大。所以，在研究电池的 *I-V* 曲线时，一定要考虑到光照、温度因素的影响[15]。

表 1 光伏阵列参数

Table 1 PV array parameters

规格参数	数值
最佳工作电压(U_{mp})/V	1 099.54
最佳工作电流(I_{mp})/A	215.68
开路电压(U_{oc})/V	1 301.82
短路电流(I_{sc})/A	228.00



（a）不同光照强度

（a）Varying light intensity

（b）不同温度

（b）Different temperature

图 2 不同光照强度和温度条件下的伏安特性曲线

Fig. 2 Volt-ampere characteristic curves under different light intensity and temperature conditions

光伏发电系统的故障形式多种多样,主要有开路、短路、老化及阴影遮挡等。在各种故障情况下,其 *I-V* 曲线也各不相同[16]。如图 3 所示,多个分支同时发生故障的概率是很低的,所以仅对其中一个分支的开路进行了分析。如图 4 所示,当断路故障出现时,其开路电压 U_{oc} 变化不大,但在短路电流 I_{sc} 和最大功率点上有很大的改变[17]。

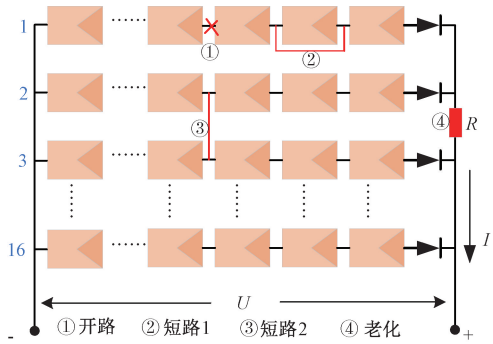


图 3　光伏阵列故障

Fig. 3　PV array fault

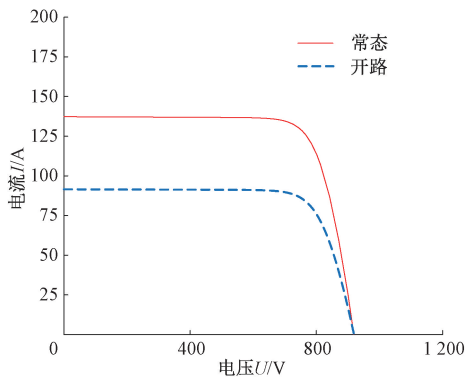


图 4　开路故障

Fig. 4　Open circuit fault

如图 5 所示,当出现短路故障时,其短路电流 I_{sc} 基本不变,但在最大功率和开路电压 U_{oc} 上却有很大的改变,并且随着失配度的增加,这种改变更加明显[18]。采用 3 种老化方式:轻微老化、中等老化和重度老化,老化电阻分别为 3、8 和 10 Ω。通过对其特征曲线的分析,可以看出,当电池出现异常老化时,其电容值和 U_{oc} 值都没有明显的变化,但其峰值电压值却有很大的波动,如图 6 所示。随着老化程度的不同,*I-V* 曲线的形状可能会发生变化,曲线的斜率可能会减小,起点和顶点位置可能会发生偏移[19]。

如图 7 所示,给出了 6 种不同的防阴影方法:方案 1 是把单块太阳能电池板的 *G* 降至 50%;在方案 2 中,3 块太阳能电池板在相同的条件下,其 *G* 降低至 40%;在

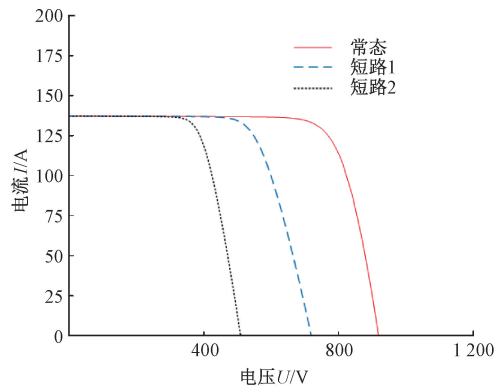


图 5　短路故障

Fig. 5　Short circuit fault

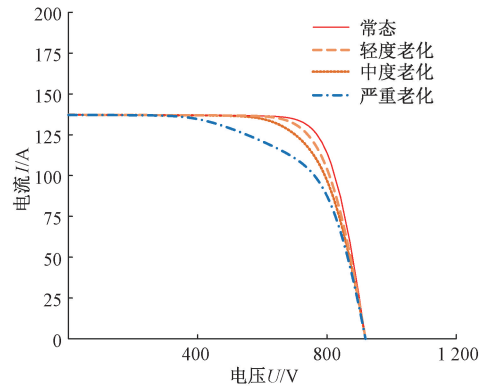


图 6　老化故障

Fig. 6　Aging fault

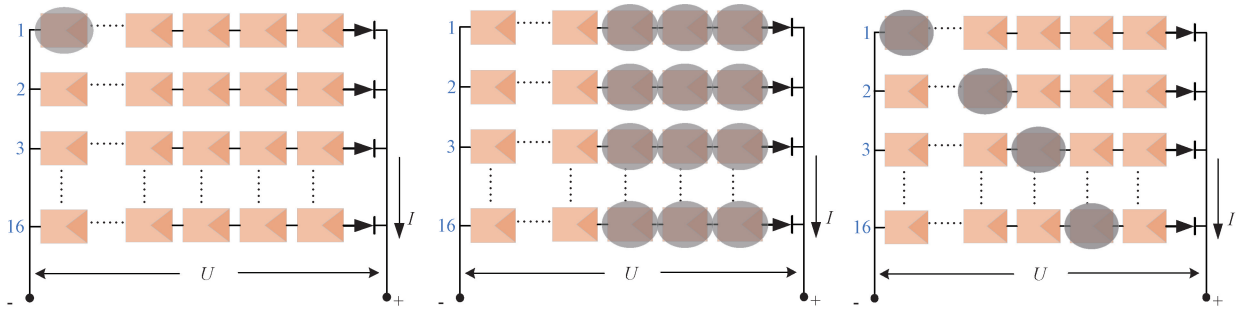
方案 3 中,在每一种串联组合中,某一块电池板的 *G* 降低至 30%;在方案 4 中,一排电池板的 *G* 降低至 50%,另一组电池板的 *G* 降低至 70%;方案 5 是将各光电板的 *G* 减小至 0;方案 6 使太阳能电池板的 *G* 达到 80%。

如图 8 所示,在不同程度的遮断故障下,其 *I-V* 曲线上可以有多个波峰,并且在最大功率点上有较大的改变[20]。

2　数据集准备及预处理

2.1　数据准备

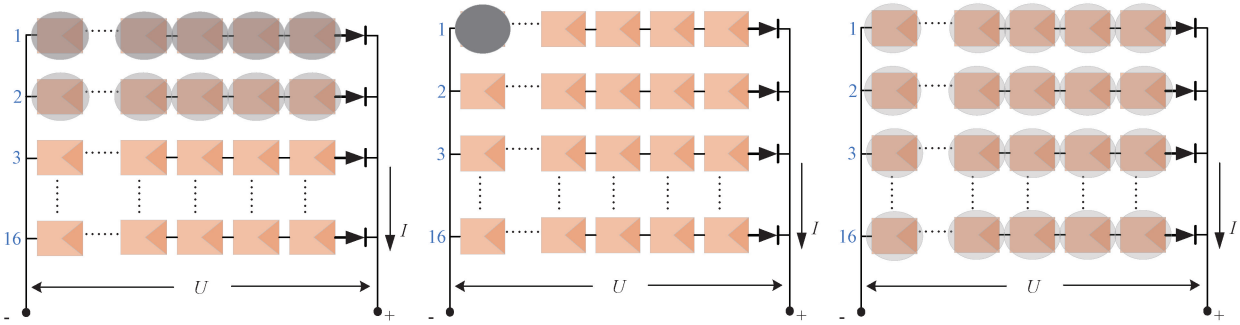
针对光伏系统实际运行环境的多样性特征,通过设定涵盖大跨度辐照度与温域的实验参数,系统性覆盖光伏组件潜在存在的各类工况条件。并采用了较小的采样间隔,以获得高分辨率的数据,有助于捕捉工况参数变化对 *I-V* 曲线的细微影响,以及模型的精细调节和性能优化。根据表 2 所示的设置,将辐照强度设定在 200 ~ 1 200 W/m^2 连续区间,温度参数扩展至 45℃ 动态范围,采用 0.1℃ 以及 10 W/m^2 的高精度采样间隔。针对各运



（a）方案1
（a）Program 1

（b）方案2
（b）Program 2

（c）方案3
（c）Program 3



（d）方案4
（d）Program 4

（e）方案5
（e）Program 5

（f）方案6
（f）Program 6

图 7 阴影遮挡的方案

Fig. 7 Scheme of shadow masking

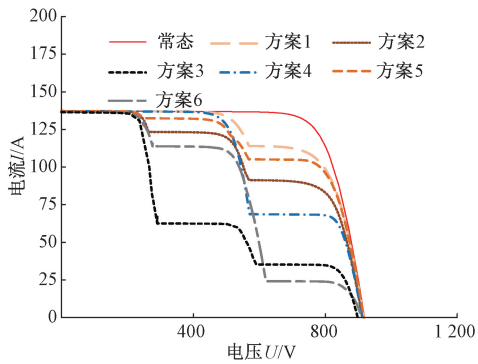


图 8 阴影遮挡

Fig. 8 Shadow masking

行工况的实验设计，通过设置单工况千组数据采集量，最终构建包含 5 000 组 *I-V* 特性曲线的实验数据库。为了评估模型的性能及其泛化能力，采用标准的数据集划分比例，将训练样本设为总数据集的 80%，测试样本设为 20%。

表 2 数据集设置情况

Table 2 Data set setup

标签	故障类型	测试样本	训练样本
1	正常	800	200
2	开路故障	800	200
3	短路 1	400	100
	短路 2	400	100
4	微老化	266	66
	中度老化	267	67
	重老化	267	67
5	阴影 1	133	33
	阴影 2	133	33
	阴影 3	133	33
	阴影 4	133	33
	阴影 5	134	34
	阴影 6	134	34

2.2 数据的预处理

I-V 曲线的预处理主要包括 2 个操作：校正和归一化。基于前期实验数据解析，光伏组件电流-电压特性呈现对光照强度与环境温度的高度敏感性。因此，在进行光伏阵列故障诊断时，必须考虑到不同环境条件对同一故障类型特征表现的影响，以确保诊断模型的准确性和可靠性。通过抑制环境变量对伏安特性曲线的干扰，确保诊断模型聚焦于故障特征的有效解析。因此，将 *I-V* 曲线校正为相同的环境条件，这里以 STC（25℃，

1 000 W/m^2)作为目标条件。

根据文献[21-23],*I-V* 曲线校正方法是 IEC 60891 标准中的程序 1 和 2。然而,这些方法在存在故障时,性能有限。因此,提出了一种改进的校正方法,只需要实际电流、电压和相应的环境条件参数,在实际应用中易于实现和操作。温度系数 α 是一个可调参数,可以根据实际测量数据或者制造商提供的参数进行调整,因此能够适应不同类型和规格的光伏模块,通过动态参数调整机制增强其跨工况匹配能力与普适特征。如式(1)和(2)所示,校正后的电流 I_{STC},校正后的电压 U_{STC}。

$$I_{\mathrm{STC}} = I_{\mathrm{measured}} \times \left(\frac{G_{\mathrm{STC}}}{G_{\mathrm{measured}}}\right) \times \left(\frac{T_{\mathrm{measured}}}{T_{\mathrm{STC}}}\right)^{\alpha} \tag{1}$$

$$U_{\mathrm{STC}} = U_{\mathrm{measured}} + (U_{oc,\mathrm{STC}} - U_{oc,\mathrm{measured}}) \tag{2}$$

式中:I_{measured} 是测量到的实际电流;G_{measured} 是测量到的实际光照强度;T_{measured} 是测量到的实际温度;G_{STC} 是 STC 条件下的光照强度;T_{STC} 是 STC 条件下的温度;α 是温度系数,指在不同温度下其电压和电流的变化率,通常取决于光伏模块的类型和制造工艺;U_{measured} 是测量到的实际电压;$U_{oc,\mathrm{STC}}$ 是 STC 条件下的开路电压;$U_{oc,\mathrm{measured}}$ 是测量到的实际开路电压。

为确保诊断模型能够快速进行迭代训练且训练结果真实有效,对训练样本数据做 *Z*-score 归一化处理[24]。如式(3)所示,这一选择是基于其在处理存在异常值和最大最小值不确定的情况下,为数据提供稳定而可靠的标准化方式的优势。

$$x' = \frac{x - \mu}{\sigma} \tag{3}$$

式中:μ、σ 分别为样本数据的均值和标准差。

3　集成学习算法的优化与改进

为实现光伏阵列小样本和高精度的故障诊断,采用 CatBoost 作为诊断模型。CatBoost 作为集成梯度提升框架,在有限样本条件下的数据分析与高准确度推断任务中展现出显著的技术优势。在实际应用中,光伏数据可能存在缺失值或者不完整的数据情况,CatBoost 能够自动处理这些缺失值,而无需额外的数据处理步骤,简化数据预处理流程,CatBoost 框架有效提升了模型收敛效率与泛化性能。但该模型效能受核心超参数配置制约,需采用 SSA 算法进行动态调优。但传统 SSA 存在初始种群分布异质性不足、群体多样性衰减及局部收敛倾向等技术缺陷,严重制约了参数空间的全局探索效能。因此对 SSA 算法运用两种策略进行改进,进一步提升 CatBoost 模型的性能。

诊断流程如图 9 所示,针对原始故障数据的特征工程处理流程,首先实施标准化预处理操作,随后通过多算法协同优化框架实现核心超参数的全局搜索,构建最优诊断架构。采用测试集实施交叉验证机制,以故障类型标签为监督信号驱动模型训练过程。最终基于混淆矩阵的多维度性能指标计算,完成对诊断系统预测精度与泛化能力的系统性效能评估。

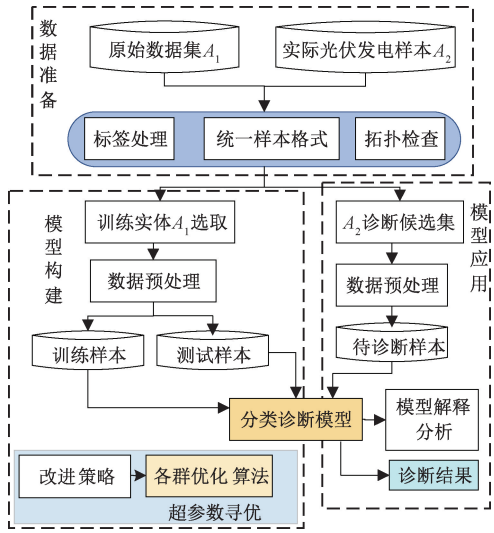


图 9　光伏阵列故障诊断流程

Fig. 9　PV array fault diagnosis process

3.1　CatBoost 模型参数优化

针对梯度提升树模型 CatBoost 的超参数敏感性特征,其诊断性能易受深度参数与学习速率等关键参数配置的制约。在数据拟合过程中可能引发梯度激增现象,导致模型收敛波动与预测稳定性下降。为此,设计基于 SSA 算法的超参数优化框架,重点针对决策树深度和学习速率构建动态寻优机制,通过平衡模型复杂度与训练效率实现收敛加速。

参数优化机制遵循以下路径:以最小化分类损失函数 $L(\theta)$ 为核心优化目标,其中 θ 表示 CatBoost 模型的超参数向量组合。首先建立超参数可行域 S,其中每个参数 $\theta_i \in D_i$ 满足定义域约束。在迭代优化阶段,算法根据适应度评估结果执行参数组合优选机制[25]。具体而言,基于全局最优解 θ_{best} 动态更新参数向量,如式(4)所示的动态调整规则。

$$\theta_{\mathrm{prey}}(t) = \theta_{\mathrm{best}}(t) + \alpha \times \mathrm{rand}() \times (P_{\max} - P_{\min}) \tag{4}$$

式中:α 系数用于调节迭代步距;rand( )生成[0,1)区间的均匀分布随机变量;$P_{\max}$ 和 $P_{\min}$ 分别表征超参数空间的上下边界约束。

初始化包含 $\boldsymbol{M}$ 个搜索代理的群体,每个代理个体表征一组超参数配置,即 $\boldsymbol{m}_k = (\theta_{k1}, \theta_{k2}, \cdots, \theta_{kn})$。针对每个代理个体,构建对应参数配置的 CatBoost 分类器,通过验证集计算其损失函数值 $L(\boldsymbol{m}_k)$ 作为适应度指标。在迭代更新阶段,基于当前参数位置 $\boldsymbol{m}_k$ 与邻域半径 D_i,采用柯

西-高斯混合变异算子生成候选解 $\boldsymbol{m}'_k$,否则保留原始参数配置。通过精英保留策略更新群体分布,最终选取适应度最优的超参数组合 θ。将该最优配置应用于 CatBoost 模型重构,并在独立测试集上验证其诊断性能。

3.2 引入精英反向学习策略优化 CatBoost 模型

CatBoost 模型的超参数通常存在相互影响的情况,这使得搜索空间更加复杂,在优化过程中,如果种群过于集中在某些局部区域,并且无法充分探索搜索空间中的其他潜在解,SSA 算法不能很好地考虑到这些相互影响,可能引发初始种群异质性不足与群体多样性衰减的缺陷[26]。

针对 SSA 算法在初始种群异质性不足与群体多样性衰减方面的技术缺陷,构建基于精英反向学习的优化框架。其核心思想是在每一代中选择适应度较低的个体进行特殊处理,记为 $\boldsymbol{P}^t_{low}$,在 CatBoost 模型的超参数优化过程中,将精英逆向学习机制整合至超参数集合的演化过程中,实现群体多样性的动态增强。假设选择了 N 个适应度较低的个体 $\boldsymbol{P}^t_{low}$。然后,用群体多样性维持机制,对适应度较低的个体实施定向优化操作,包括参数空间重采样与创新性搜索策略集成。具体而言,通过构建原始解的逆向映射解集,建立原解与其逆向解的适应度双维度评估框架,进而采用精英保留策略选择最优候选解作为迭代基向量。该优化原理如式(5)和(6)所示的逆向映射关系。

$$x_j^* = a_j + b_j - x_j \tag{5}$$

$$\begin{cases} x_{ij}^* = a_j(i), & x_{ij}^* < a_j(t) \\ x_{ij}^* = b_j(i), & x_{ij}^* < b_j(t) \end{cases} \tag{6}$$

式中:$a_j(t)$、$b_j(t)$为精英群体所构造的区间,当反向解越过边界 $a_j(t)$、$b_j(t)$时,可以用进行重置。

具体的优化和改进的流程,将 CatBoost 模型的超参数空间作为输入,这些超参数将被麻雀搜索算法搜索以找到最佳组合。在迭代优化过程中,首先对每个麻雀个体所代表的 CatBoost 模型进行训练,并使用验证集评估模型的性能。然后,根据适应度函数对个体进行排序,并保留表现优秀的个体。接着,剩余的个体将通过麻雀搜索算法进行迭代优化,以尝试找到更优的超参数组合。在更新种群时,引入一定数量的表现优秀的个体,以保持全局搜索的多样性,并防止陷入局部最优解。这有助于保持种群的多样性。最终输出是优化后的 CatBoost 模型及其对应的超参数组合,这是使优化目标最大化或最小化的最佳组合。

3.3 引入柯西高斯变异策略优化 CatBoost 模型

CatBoost 模型的超参数优化问题通常是非凸的,因此可能存在许多局部最优解。如果麻雀搜索算法在初始解附近找到了一个局部最优解,很容易陷入其中,而无法找到更好的解。特别是在高维空间中,由于搜索空间的复杂性,算法可能需要更长的时间才能收敛到全局最优解[27]。

为了解决麻雀搜索算法陷入局部最优的问题,引入柯西高斯变异策略。对个体进行变异操作,引入随机性,增加搜索的多样性。对于第 i 个个体 θ_i^t,其变异操作如式(7)所示。

$$\theta_i = \theta_i^t + \delta \times N(0,\sigma) \tag{7}$$

式中:$N(0,\sigma)$为随机扰动项服从均值为 0、σ 标准差的正态分布;δ 为调控参数。

采用精英保留策略的优化机制,优先对当前全局最优个体执行柯西-高斯混合突变操作。如式(8)所示,为动态适应度评估函数,系统对比突变前后候选解的适应度指标,基于改进准则选择优势解 U^t_{best} 作为下一代迭代基向量。

$$U^t_{best} = X^t_{best}[1 + \lambda_1 \text{cauchy}(0,\sigma^2) + \lambda_2 \text{Gauss}(0,\sigma^2)] \tag{8}$$

式中:σ^2 为征柯西高斯混合扰动机制的强度参数;$\text{cauchy}(0,\sigma^2)$是满足柯西分布的随机变量;$\lambda_1$ 和 λ_2 是随迭代次数自适应调整的动态参数[28]。

在这个改进的迭代优化过程中,首先对每个麻雀个体所代表的 CatBoost 模型进行训练,并使用验证集评估模型的性能。然后,在群体进化算法的迭代优化阶段,采用柯西高斯混合扰动机制对每个候选解实施参数空间探索操作。接着,保留表现优秀的个体,并对剩余个体进行基于麻雀搜索算法的迭代优化。最终输出是优化后的 CatBoost 模型及其对应的超参数组合,这是使优化目标最大化或最小化的最佳组合。

3.4 性能实验分析

为了进一步验证精英反向策略和柯西高斯变异策略相较于其他策略对 SSA 算法的改进效果,考虑了一系列不同的基准函数,用以评估各种策略组合的收敛速度和最优解的质量表现,如表 3 所示。

表 3 优化策略

Table 3 Optimization strategies

组合方式	名称	改进策略
①	RASSA	反向学习策略+步长因子动态调整策略
②	RLSSA	反向学习策略+Levy 飞行策略
③	RKSSA	反向学习策略+柯西高斯变异策略
④	EASSA	精英反向策略+步长因子动态调整策略
⑤	ELSSA	精英反向策略+Levy 飞行策略
⑥	EKSSA	精英反向策略+柯西高斯变异策略

为比较不同的策略组合在解决特定基准函数上的性能表现,选择的 3 个基准函数如表 4 所示。

表 4　基准函数

Table 4　Benchmark functions

测试函数名称	基准函数	维数 d	搜索范围
Shifted Sphere Function(F1)	$f(x)=\sum_{i=1}^{n}x_i^2$	30	[-100,100]
Schwefel's Problem 1.2(F2)	$f(x)=\sum_{i=1}^{d}\lvert x_i\rvert+\prod_{i=1}^{d}\lvert x_i\rvert$	30	[-10,10]
Schwefel 2.21(F3)	$f(x)=max_i\lvert x_i\rvert$	30	[-100,100]

通过循环多次运行,对每个优化算法进行初始化,并运行指定最大迭代次数为 500 以及群种数量为 30。如图 10 所示,在每次迭代中,记录收敛曲线,通过绘制收敛曲线的结果的方式来比较不同策略组合的性能表现。在基准函数的对比中,EKSSA 算法展现出显著的收敛特性优势。该算法通过精英逆向学习机制动态提升种群质量,促使搜索过程向高适应度区域高效收敛。相较而言,传统优化方法在迭代后期易出现搜索停滞现象,其优化效能与收敛稳定性均存在明显局限。值得注意的是,EKSSA 的适应度曲线呈现独特的阶梯式下降特征,这源于柯西-高斯混合扰动机制在维持搜索方向稳定性的同时,有效突破局部最优解的约束边界。理论分析与实验验证表明,该混合策略通过双重优化机制的协同作用,在解空间探索与开发能力的平衡性、算法鲁棒性以及收敛轨迹的平滑性等方面均呈现显著优势。

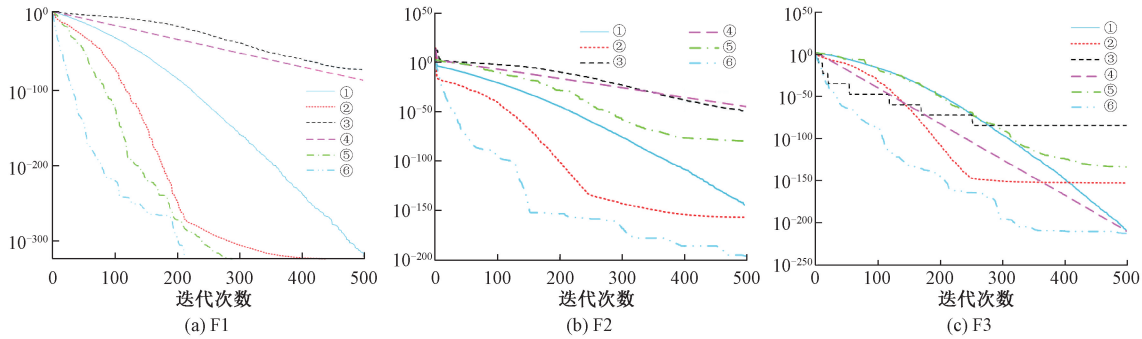


图 10　基准函数收敛曲线

Fig. 10　Convergence curve of the benchmark function

3.5　模型训练结果分析

1) CatBoost 模型损失函数分析

为进一步评估各种优化策略对 CatBoost 诊断模型的影响与效果,对改进优化前后的损失函数进行比较分析。如图 11 所示,各类算法的损失函数均呈现渐进收敛趋势并最终达到稳定状态。但采用 EKSSA 优化框架的 CatBoost 模型展现出更优异的收敛特性,其损失函数不仅实现快速趋稳,更获得更优的极值定位,在收敛速率与误差抑制维度均产生质的提升。实验验证表明,该混合优化架构通过增强种群多样性与突破局部极值约束的双重机制,显著提升诊断模型的泛化能力与鲁棒性,为解决复杂故障识别问题提供了新性技术范式。

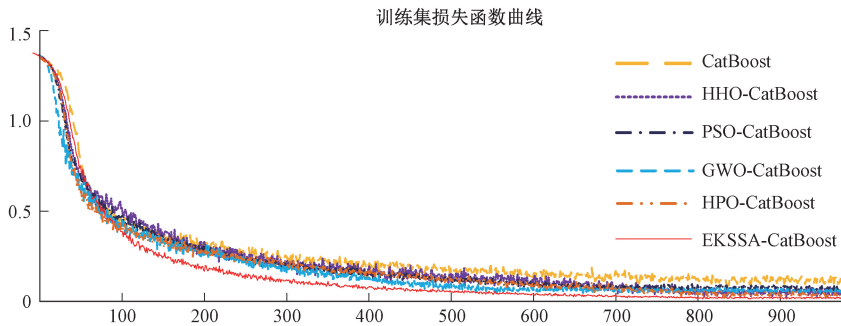


图 11　CatBoost 训练集损失函数曲线

Fig. 11　Loss function curve of CatBoost training set

2) 模型评估指标

为了更直观地展示 EKSSA 算法优化 CatBoost 模型前后对于不同故障类型的诊断性能,选择的模型性能度量指标包括查全率 R_e、精确率 P_{re} 和 F_1 评分,对 EKSSA-

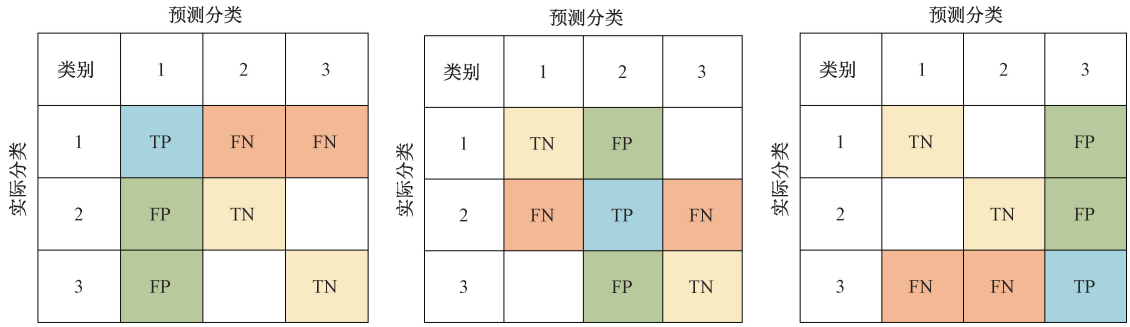
CatBoost 模型在光伏阵列故障诊断任务中的表现进行评估。为了更好的理解这些指标，如图 12 所示，利用三分类混淆矩阵进行分析。

$$R_c = \frac{TP}{TP + FN} \tag{9}$$

$$P_{re} = \frac{TP}{TP + FP} \tag{10}$$

$$F_1 = \frac{2 \cdot R_c \cdot P_{re}}{R_c + P_{re}} \tag{11}$$

式中：实际类别与预测结果均为正类的样本量定义为真阳性（true positives，TP）；实际为正类但被误判为负类的样本量称为假阴性（false negatives，FN）；实际为负类却错误标记为正类的样本量归为假阳性（false positives，FP）；而真实与预测结果均为负类的样本量则计为真阴性（true negatives，TN）。



(a) 类别1
(a) Category 1

(b) 类别2
(b) Category 2

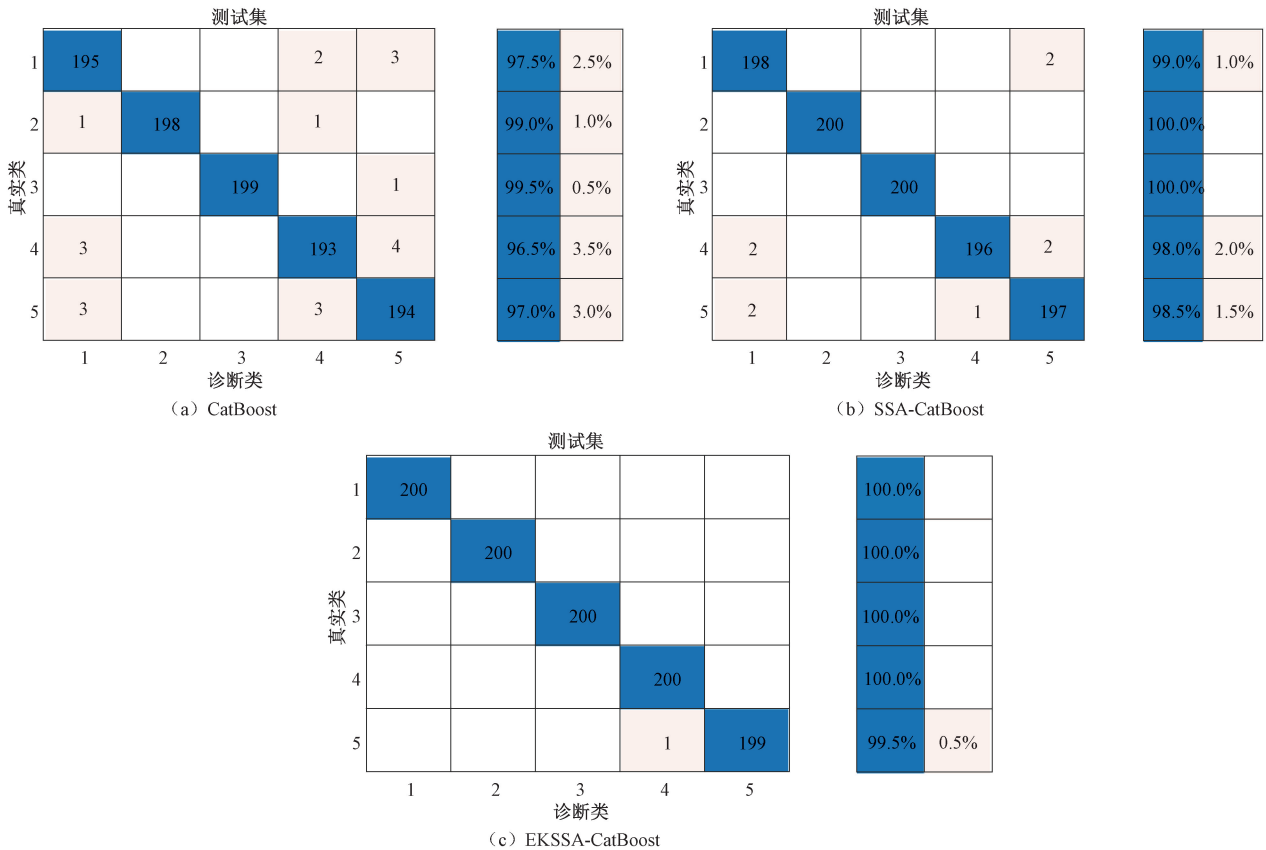
(c) 类别3
(c) Category 3

图 12 三分类各指标计算

Fig. 12 Calculation of indicators for each of the three classifications

3）模型指标分析

如图 13 所示，将测试集下 5 种状态的诊断结果通过混淆矩阵的形式可视化，并根据混淆矩阵计算评估指标。



（a）CatBoost

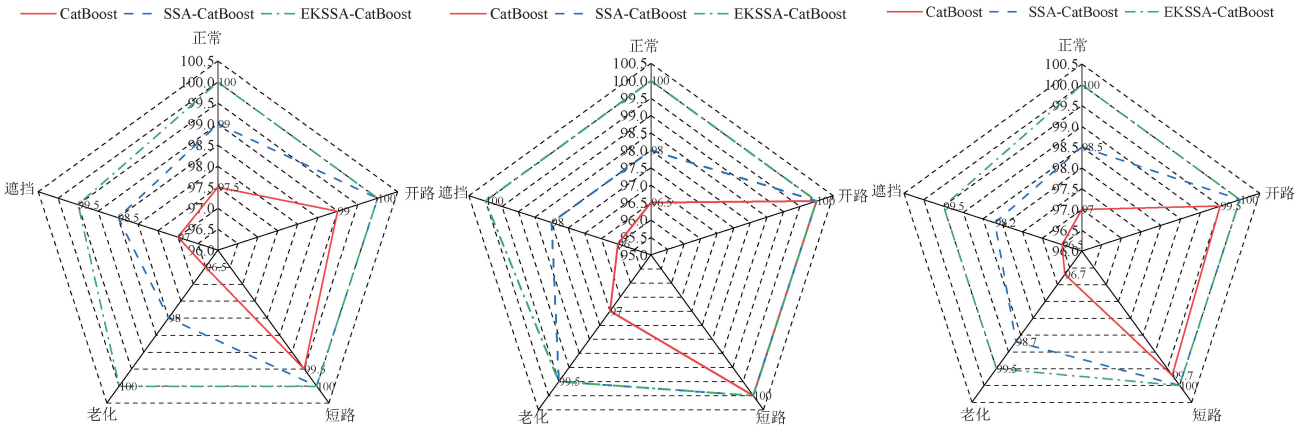
（b）SSA-CatBoost

（c）EKSSA-CatBoost

图 13 测试集混淆矩阵

Fig. 13 Test set confusion matrix

如图 14 所示,对于开路和短路这 2 种状态的诊断,3 个模型的诊断效果是相同的都能达到 100%。根据优化前后的查全率相进行分析,经过改进的 EKSSA 算法优化后的 CatBoost 模型的诊断性能达到最优,仅出现 1 次误诊,并提高了对复杂故障的诊断效果。根据精准率进行分析,除了在开路和短路状态下 3 个模型的诊断效果相同外,EKSSA-CatBoost 模型在其他状态下的诊断表现最优。通过对 F_1 值的比较,EKSSA-CatBoost 模型在正常情况、老化情况和阴影遮挡情况下的识别能力都有显著提高。总体而言,EKSSA-CatBoost 模型的整体性能达到最优,并且对于诊断老化和阴影遮挡这两种复杂的故障状态依然可以达到高准确率。



（a）查全率R_c　（b）精确率P_{re}　（c）F_1评分

（a）R_c for completeness　（b）Precision rate P_{re}　（c）F_1 score

图 14　不同运行状态下模型性能度量指标的分析结果

Fig. 14　Analysis results of model performance metrics in different operating states

4）不同输入方式的诊断效果分析

设计对比验证实验,基于 2.1 节构建的电流-电压特性曲线数据库,实施多维特征工程处理。如图 15 所示,提取涵盖开路电压、短路电流等关键电气参数,以及最大功率点的电流、电压及功率参数。将这 5 个关键特征与环境变量(T、G)组合成一维矩阵作为诊断模型的输入。最后,采用随机分层抽样策略将特征工程处理后的数据集实例按训练集与测试集 8∶2的划分方案进行分配,并将其分别输入到支持向量机(support vector machine, SVM)、k-邻近算法(k-nearest neighbors, KNN)、卷积神经网络(convolutional neural network, CNN)和 EKSSA-CatBoost 模型进行训练和交叉验证。

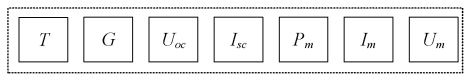


图 15　诊断模型的输入

Fig. 15　Inputs to the diagnostic model

如图 16 所示,在测试集的混淆矩阵可视化分析中,四类诊断模型的综合性能与预设目标存在显著偏差。虽然基于 EKSSA-CatBoost 的模型诊断效果比其他的 3 个模型整体效果更好,但是通过对比图 13(c),实验结果表明基于全 I-V 特性曲线作为诊断模型的输入,模型的诊断效果达到最优,整体的准确率最高。

（a）SVM　（b）KNN

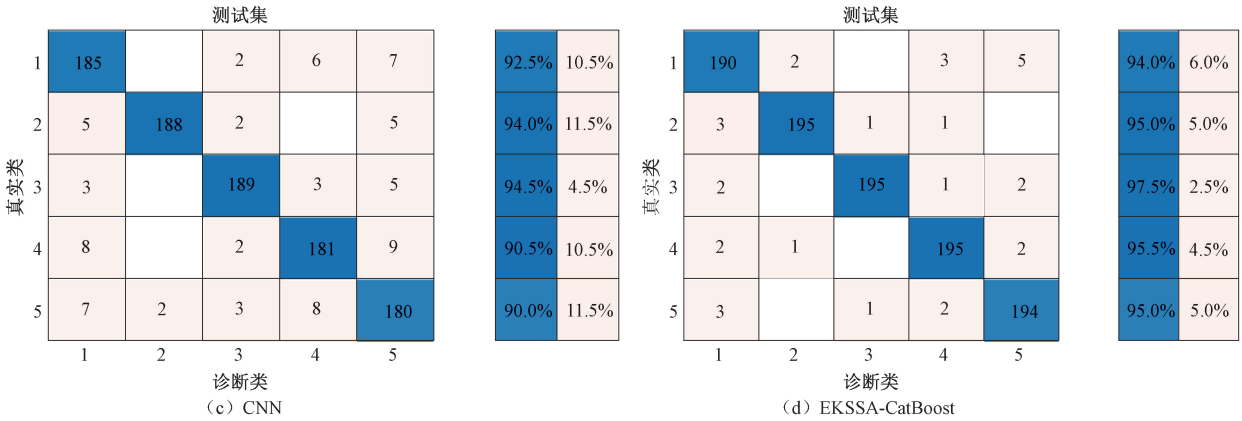


图 16 不同诊断模型测试集混淆矩阵

Fig. 16 Confusion matrix for the test set of different diagnostic models

4 使用实验数据诊断分析

构建光伏阵列试验平台，并利用该数据集对所提出的算法进行验证。基于大唐湖泉光伏电站运行参数监测的实证数据集，该数据集完整记录了多气象和各种工况下光伏阵列输出特性，单工况采样量 1 600 组，累计获取 8 000 组带有时空分布特征的伏安特性曲线样本。通过对试验平台的建立以及算法的应用，旨在评估所提算法在实际应用中的可行性和实用性，以进一步提升光伏发电系统的性能和效率。如图 17 所示，太阳电池阵列、直流汇流箱、逆变器、数据采集模块是光伏系统实验平台的主要组成，是现场的实际装置和部分实物。

图 17 大唐华银涟源新能源湖泉光伏发电站现场及装置

Fig. 17 Datang huayin lianyuan new energy huquan photovoltaic power station site and installation

如图 18 所示，展示了 4 种不同诊断模型的 t-SNE 图，这些图形可提供有关模型分类效果的直观反馈。深入分析每种模型的聚类效果，以及这些效果所反映的模型性能差异。首先，CNN 模型的聚类效果并不理想，观察图像可以看出，不同类别的样本点几乎混杂在一起，这表明该模型在区分不同故障类型方面存在挑战，其特征提取和分类能力可能不够充分。其次，SVM 和 KNN 模型的聚类效果相对有所改善，但仍存在一定的混淆情况，特别是在区分正常、老化故障和阴影遮挡工况时。这些模型在处理特定类型的故障时存在一定的困难，需要更多的训练样本或者更优化的特征表示来提高分类准确性。EKSSA-CatBoost 模型聚类效果最佳，能够清晰地区分出 5 种不同工况，这表明了该模型在分类任务上的出色性能。

如图 19 所示，根据验证集混淆矩阵进行分析，结果表明基于 EKSSA-CatBoost 的诊断模型的表现最好，只出现了 2 个误差，SVM、KNN、CNN 等传统方法的局限性是在高维数据、复杂故障类型、计算复杂度和小样本数据处理等方面的不足，尤其是对于复杂工况诊断效果查准率并不高。如表 5 所示，呈现了多算法架构在典型故障模式下的查准率指标，以及训练阶段与测试阶段的综合判别精度对比数据。提出的方法与传统方法相比有一定提升，但在复杂故障诊断、环境变化等实际应用场景中，微小提升可能带来显著的效益，尤其在减少误诊和漏诊、提升系统可靠性方面。提出的方法通过集成学习和超参数优化有效解决了这些问题，提高了故障诊断的精度和稳定性。根据召回率进行分析，4 种模型对于开路和短

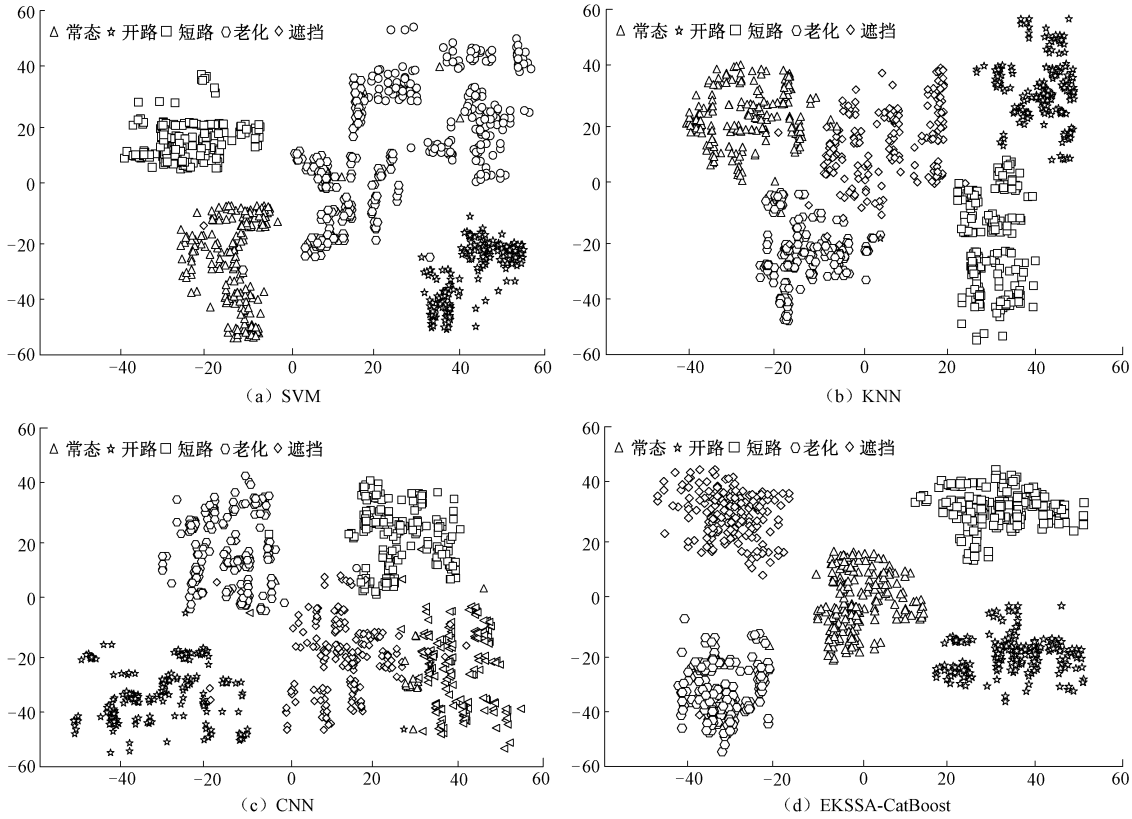


图 18　4 种诊断模型的 *t*-SNE 图

Fig. 18　*t*-SNE plots for the four diagnostic models

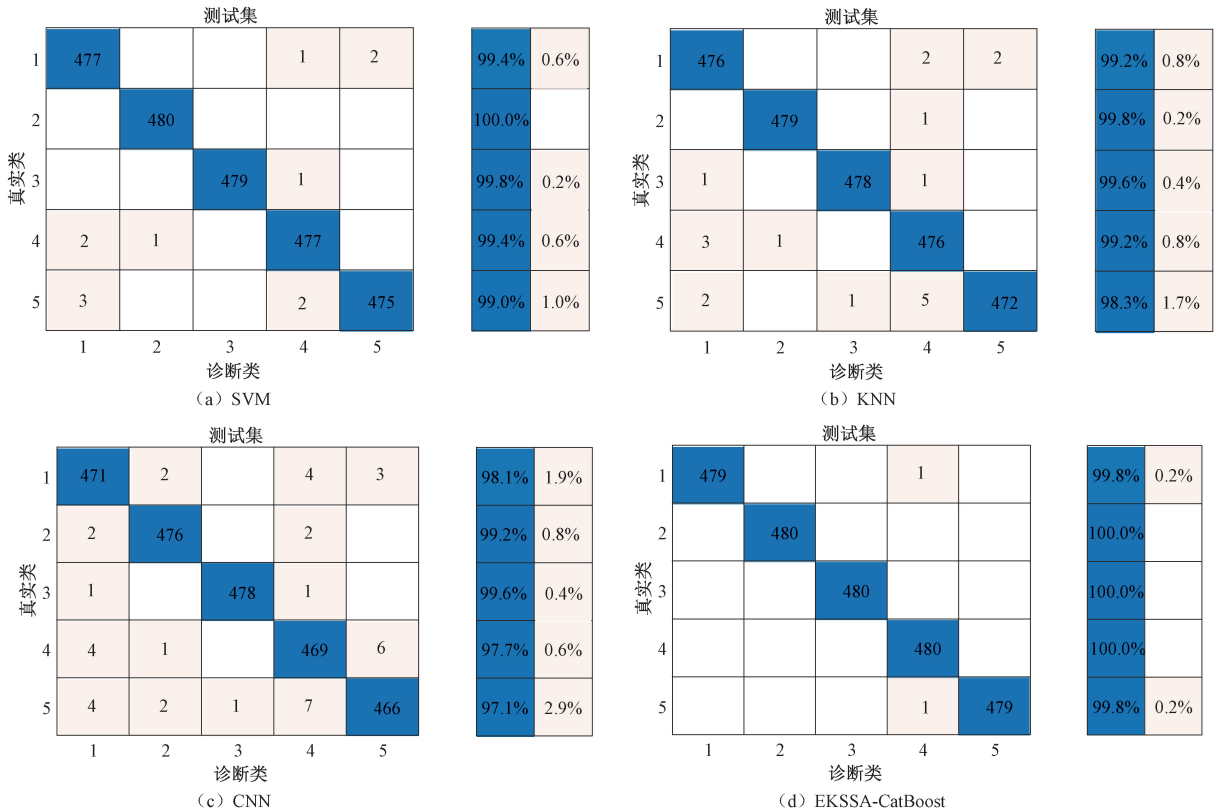


图 19　不同诊断模型测试集混淆矩阵

Fig. 19　Confusion matrix for different diagnostic model test sets

表 5 不同诊断模型结果比较

Table 5 Comparison of results of different diagnostic models (%)

模型	查全率 R_c					准确率	
	正常	开路	短路	老化	阴影遮挡	训练集	测试集
SVM	99.4	100.0	99.8	99.4	99.0	97.9	99.5
KNN	99.2	99.8	99.6	99.2	98.3	97.1	99.2
CNN	98.1	99.2	99.6	97.7	97.1	99.3	98.3
EKSSA-CatBoost	99.8	100.0	100.0	100.0	99.8	99.8	99.9

路故障的诊断召回率都可以达到 99%以上,但对于复杂的工况故障诊断(老化故障和阴影遮挡),基于 EKSSA-CatBoost 模型相对稳定,也可以达到 99%以上。根据模型的整体准确率进行分析,EKSSA-CatBoost 模型在训练集和测试集上的诊断效果都是最优,达到 99.8%和 99.9%,在实际应用中,以大唐湖泉光伏发电站为代表的光伏阵列故障诊断系统,其故障的误诊率大约在 7%左右。而提出的方法通过引入优化的机器学习模型和超参数优化策略,将故障诊断的准确率提升至 98%以上,相较于传统方法有显著提升。这种提升不仅减少了故障漏检和误检的可能性,还显著提高了系统的可靠性和运维效率。

5 结 论

结合了全 *I-V* 曲线和深度优化的集成学习模型(EKSSA-CatBoost),实现光伏阵列故障小训练样本高精度智能诊断。在研究中,提出了两个新的 *I-V* 曲线修正公式,旨在将不同环境条件下的曲线修正为标准测试条件下的状态,有效削弱辐照度波动与温度变化对算法训练阶段的特征耦合干扰。为使 CatBoost 模型在分类时达到最佳效果,在 SSA 算法中引入精英反向策略和柯西高斯变异策略对其加以改进。结果表明基于 EKSSA-CatBoost 模型的光伏阵列故障诊断表现相对稳定,在模拟和现场测试中均实现了高分类准确率,即使在复杂工况下,EKSSA-CatBoost 模型仍能实现稳定精确的诊断。

尽管现有方法在特定条件下表现优异,但仍存在一些局限性。在处理极端环境变化(如大范围温度、光照变化等)时,模型的鲁棒性有待提高。面对大规模光伏阵列数据时,可能会面临计算效率和存储瓶颈。未来的研究工作:考虑将深度学习方法(如卷积神经网络)与集成学习方法结合,通过迁移学习增强模型在不同环境条件下的适应能力。利用增强学习等技术优化模型在实际运行中对异常数据的处理能力。

参考文献

[1] 姚蜀军,张春强,刘刚, 等. 光伏发电单元电磁暂态解耦与快速仿真方法[J]. 电力系统自动化, 2022, 46(21): 170-178.
YAO SH J, ZHANG CH Q, LIU G, et al. Electromagnetic transient decoupling of photovoltaic power generation units and fast simulation method[J]. Automation of Electric Power Systems, 2022, 46(21): 170-178.

[2] 范思远,王煜,曹生现, 等. 积灰对光伏组件输出特性影响建模与分析[J]. 仪器仪表学报, 2021, 42(4): 83-91.
FAN S Y, WANG Y, CAO SH X. et al. Effect modeling and analysis of dust accumulation on output characteristics of photovoltaic modules [J]. Chinese Journal of Scientific Instrument, 2021, 42(4):83-91.

[3] BRAUN H, BANAVAR M, SPANIAS A, et al. Signal processing for solar array monitoring, fault detection, and optimization[M]. Springer Nature, 2022.

[4] 贾科,顾晨杰,毕天姝, 等. 大型光伏电站汇集系统的故障特性及其线路保护[J]. 电工技术学报, 2017, 32(9):189-198.
JIA K, GU CH J, BI T SH, et al. Failure characteristics and line protection of large-scale photovoltaic power station compilation system [J]. Transactions of China Electrotechnical Society, 2017,32(9):189-198.

[5] 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.

[6] LI Y L, DING K, ZHANG J W, et al. A fault diagnosis method for photovoltaic arrays based on fault parameters identification[J]. Renewable Energy, 2019, 143: 52-63.

[7] LI B J, DELPHA C, MIGAN-DUBOIS A, et al. Fault diagnosis of photovoltaic panels using full I-V

characteristics and machine learning techniques [J]. Energy Conversion and Management, 2021, 248: 114785.

[8] WANG J J, GAO D D, ZHU SH K, 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.

[9] SUN J M, SUN F J, FAN J Q, et al. Fault diagnosis model of photovoltaic array based on least squares support vector machine in Bayesian framework [J]. Applied Sciences, 2017, 7(11):1199.

[10] LIU Y J, DING K, ZHANG J W, et al. Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with IV curves [J]. Energy Conversion and Management, 2021, 245:114603.

[11] KUMAR U, MISHRA S, DASH K. An IoT and semi-supervised learning-based sensorless technique for panel level solar photovoltaic array fault diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72:3521412.

[12] APPIAH A Y, ZHANG X H, AYAWLI B B K, et al. Review and performance evaluation of photovoltaic array fault detection and diagnosis techniques[J]. International Journal of Photoenergy, 2019, 2019: 6953530.

[13] CHINE W, MELLIT A, PAVAN A M, et al. Fault diagnosis in photovoltaic arrays[C]. 2015 International Conference on Clean Electrical Power, 2015: 67-72.

[14] 王元章,李智华,吴春华. 光伏系统故障诊断方法综述[J]. 电源技术,2013,37(9):1700-1705.
WANG Y ZH, LI ZH H, WU CH H. Fault diagnosis technologies for photovoltaic system[J]. Chinese Journal of Power Sources, 2013,37(9):1700-1705.

[15] LI B J, MIGAN-DUBOIS A, DELPHA C, et al. Evaluation and improvement of IEC 60891 correction methods for 556I-V curves of defective photovoltaic panels[J]. Solar Energy 2021, 216:225-237.

[16] PILIOUGINE M, SANCHEZ F P, SPAGNUOLO G. Comparative of IEC 60891 and other procedures for temperature and irradiance corrections to measured I-V characteristics of photovoltaic devices [J]. Energies, 2024, 17(3):566.

[17] PHANG J C H, CHAN D S H. A review of curve fitting error criteria for solar cell I-V characteristics[J]. Solar Cells 1986, 18(1): 1-12.

[18] 彭自然,张颖清,肖伸平. 基于 YOLOv5 的太阳电池表面缺陷检测[J]. 太阳能学报,2024,45(6):368-375.
PENG Z R, ZHANG Y Q, XIAO SH P. Detection of surface defects in solar cells at YOLOv5 [J]. Acta Energiae Solaris Sinica, 2024,45(6):368-375.

[19] DIRNBERGER D, KRÄLING U. Uncertainty in PV module measurement—Part I: Calibration of crystalline and thin-film modules[J]. IEEE Journal of Photovoltaics, 2013, 3(3):1016-1026.

[20] LI CH X, YANG Y H, ZHANG K J, et al. A fast MPPT-based anomaly detection and accurate fault diagnosis technique for PV arrays[J]. Energy Conversion and Management, 2021, 234:113950.

[21] 彭自然,许怀顺,肖伸平. 一种基于 CatBoost 优化的光伏阵列故障诊断模型[J]. 电子学报,2024,52(7):2418-2428.
PENG Z R, XU H SH, XIAO SH P. A CatBoost optimization-based fault diagnosis model for photovoltaic arrays[J]. Acta Electonica Sinica, 2024,52(7):2418-2428.

[22] XU L CH, PAN ZH H, LIANG CH D, et al. A fault diagnosis method for PV arrays based on new feature extraction and improved the fuzzy C-mean clustering[J]. IEEE Journal of Photovoltaics, 2022, 12(3):833-843.

[23] 彭自然,杨肖阳,肖伸平. 基于 EKF-HInformer 模型估计汽车动力电池的 SOC&SOH[J]. 电子测量与仪器学报, 2025,39(3):21-33.
PENG Z R, YANG X Y, XIAO SH P. The SOC and SOH of the battery are estimated based on the EKF-HInformermodel[J]. Journal of Electronic Measurement and Instrumentation, 2025,39(3):21-33.

[24] 彭自然,王顺豪,肖伸平, 等. 一种精确估算电动汽车动力电池 SOC&SOH 的循环门控模型[J]. 电子测量与仪器学报,2024,38(9):11-23.
PENG Z R, WANG SH H, XIAO SH P. Cycle gating model for accurate estimation of SOC&SOH ofpower battery in electric vehicles [J]. Journal of Electronic Measurement and Instrumentation, 2024,38(9):11-23.

[25] 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 ZH H, LIU ZH ZH, 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.

[26] 彭自然,王思远,张颖清, 等. 基于 YOLOv5 的无人机航拍红外图像的微弱光伏阵列热斑检测[J/OL]. 太阳能学报, 1-9[2025-03-27].
PENG Z R, WANG S Y, ZHANG Y Q. Detection of

weak photovoltaic array hotspots in uavaerial infrared images based on YOLOv5[J/OL]. Acta Energiae Solaris Sinica, 1-9[2025-03-27].

[27] 吕游,郑茜,齐欣宇, 等. 基于改进 EfficientNet 的红外图像光伏组件故障识别研究[J]. 仪器仪表学报, 2024,45(4):175-184.
LYU Y, ZHENG X, QI X Y, et al. A study on fault recognition of photovoltaic module with infraredimages based on improved EfficientNet[J]. Chinese Journal of Scientific Instrument, 2024,45(4):175-184.

[28] 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.

作者简介

彭自然(通信作者),2004 年于中南大学获得学士学位,2008 年于中南大学获得硕士学位,2017 年于中南大学获得博士学位,现为湖南工业大学副教授,主要研究方向为人工智能、智能检测仪表、智能移动终端等方面。
E-mail:pengziran@ hut. edu. cn

Peng Ziran(Corresponding author) received his B. Sc. degree from Central South University in 2004, M. Sc. degree from Central South University in 2008, and Ph. D. degree from Central South University in 2017. Now he is an associate professor at Hunan University of Technology. His main research interests are artificial intelligence, intelligent detection instruments, and intelligent mobile terminals.

许怀顺,2018 年于青岛农业大学获得学士学位,现为湖南工业大学硕士研究生,主要研究方向为光伏发电的故障智能诊断。
E-mail:1106238410@ qq. com

Xu Huaishun received his B. Sc. degree from Qingdao Agricultural University in 2018. Now he is a M. Sc. candidate at Hunan University of Technology. His main research direction being intelligent fault diagnosis of photovoltaic power generation.

肖伸平,1988 年于东北大学获得学士学位,2002 年于中南林业科技大学获得硕士学位,2008 年于中南大学获得博士学位,现为湖南工业大学教授,主要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

Xiao shenping received his B. Sc. degree from Northeastern University in 1988, B. Sc. degree from Central South University of Forestry and Technology in 2002, and Ph. D. degree from Central South University in 2008. Now he is a professor at Hunan University of Technology. His main research interests include robust control theory and application of time-delay systems, stability analysis of power time-delay systems, industrial network control, intelligent control, process control, etc.

潘长宁,1999 年于湖南师范大学获得学士学位,2007 年于湖南师范大学获得硕士学位,2014 年于湖南大学获得博士学位,现为湖南工业大学教授,主要研究方向为热电材料中的电子输运、热输运及其热电转换效应研究。
E-mail:panchangning2000@ sina. com

Pan Changning received his B. Sc. degree from Hunan Normal University in 1999, M. Sc. degree from Hunan Normal University in 2007, and Ph. D. degree from Hunan University in 2014. Now he is a professor at Hunan University of Technology. His main research interests are electron transport, heat transport and thermoelectric conversion effect in thermoelectric materials.