

DOI: 10.19666/j.rlfed.202402037

基于 DTW-两阶四分位的分布式光伏发电 异常数据辨识

刘洋^{1,2}, 于海东^{1,2}, 刘文彬^{1,2}, 黄敏^{1,2}, 李立生^{1,2}, 张世栋^{1,2}

(1. 国网山东省电力公司电力科学研究院, 山东 济南 250003;

2. 山东省智能电网技术创新中心, 山东 济南 250003)

[摘要] 设备故障、天气环境等因素导致分布式光伏发电系统产生大量异常数据, 对其安全稳定运行造成严重影响。为了准确识别和剔除存在的异常数据, 提出一种基于动态时间弯曲 (DTW)-两阶四分位的分布式光伏发电异常数据辨识方法。首先, 通过对比相似辐照度下光伏功率均值实现连续型异常数据识别与剔除, 采用基于同时段光伏功率均值剔除异常数据, 并考虑光伏发电曲线的波动性, 采用基于 DTW 与欧氏距离的综合曲线相似度判定方法剔除连续型异常数据, 更全面地考虑了数据的波动特性, 提高了连续型异常数据辨识和剔除效果; 其次, 提出 DTW-两阶四分位异常数据辨识算法, 采用一阶变化率和二阶变化率对融合后的数据进行离散型异常数据剔除, 有效识别和剔除离散型异常数据; 最后, 根据异常数据识别和剔除结果判断是否出现故障。实验结果表明: 所提算法剔除异常数据后能更好地拟合正常光伏功率数据分布情况, 相较于四分位法和 3-Sigma 算法, 所提算法剔除异常数据前后线性相关程度变化分别提高了 58.15% 和 68.41%, 辨识效果更佳。

[关键词] 分布式光伏; 异常数据辨识; 动态时间弯曲; 两阶四分位

[引用本文格式] 刘洋, 于海东, 刘文彬, 等. 基于 DTW-两阶四分位的分布式光伏发电异常数据辨识[J]. 热力发电, 2024, 53(7): 34-44. LIU Yang, YU Haidong, LIU Wenbin, et al. Abnormal data identification for distributed photovoltaic generation based on DTW and two-stage quartile[J]. Thermal Power Generation, 2024, 53(7): 34-44.

Abnormal data identification for distributed photovoltaic generation based on DTW and two-stage quartile

LIU Yang^{1,2}, YU Haidong^{1,2}, LIU Wenbin^{1,2}, HUANG Min^{1,2}, LI Lisheng^{1,2}, ZHANG Shidong^{1,2}

(1. State Grid Shandong Electric Power Research Institute, Jinan 250003, China; 2. Shandong Smart Grid Technology Innovation Center, Jinan 250003, China)

Abstract: Equipment failures, weather conditions and other factors can lead to a large amount of abnormal data in distributed photovoltaic (PV) power generation systems, causing serious effects on their safe and stable operation. In order to accurately identify and remove these abnormal data, a distributed PV power generation abnormal data identification method is proposed based on dynamic time warping (DTW) and two-stage quartile. Firstly, continuous abnormal data identification and elimination are achieved by comparing the mean photovoltaic power under similar irradiance. Abnormal data are eliminated based on the comparison of the mean photovoltaic power at the same period, taking into account the fluctuation of the photovoltaic power generation curve. A comprehensive curve similarity judgment method based on DTW and Euclidean distance is used to consider the fluctuation characteristics of the data more comprehensively, thereby improving the recognition and elimination effect of continuous abnormal data. Secondly, the DTW-Two-Stage Quartile abnormal data identification algorithm is proposed, and the first-order change rate and the second-order change rate are used to eliminate discrete abnormal data from the fused data, effectively identifying and eliminating discrete abnormal data. Finally, it is determined whether a fault has occurred based on the results of abnormal data identification and elimination. Experimental results show that, after the

收稿日期: 2024-02-21

基金项目: 国家电网有限公司科技项目 (520626210014)

Supported by: Science and Technology Project of State Grid (520626210014)

第一作者简介: 刘洋 (1988), 男, 博士, 高级工程师, 主要研究方向为配电网运行控制技术, hustliuyang@sina.com。

proposed algorithm eliminates abnormal data, it can better fit the distribution of normal photovoltaic power data. Compared with the quartile method and the 3-Sigma algorithm, the linear correlation degree of the proposed algorithm before and after the elimination of abnormal data has increased by 58.15% and 68.41% respectively, with better identification results.

Key words: distributed photovoltaic; abnormal data identification; dynamic time warping; two-stage quartile approach

光伏发电作为一种清洁、可再生能源利用形式,能有效减少二氧化碳等温室气体的排放量,是实现“碳达峰、碳中和”目标的重要途径^[1-2]。同时,光伏发电技术的创新和应用,不断推动着新能源产业的发展,而研究分布式光伏运行数据是推动光伏发展不可或缺的一环^[3-5]。其中,准确、可信的光伏功率数据是光伏发电性能评估、光伏功率预测及光伏并网等研究的基础^[6-8]。然而,外界环境变化和系统设备故障都可能导致分布式光伏功率数据中含有大量异常数据,如设备故障、通信故障、暴风暴雪等情况,不仅影响光伏功率数据质量,也难以保障相关实验和测试结果的准确性^[9-11]。因此,开展光伏发电异常数据辨识研究具有重要意义。

异常数据辨识是数据分析的重要环节,现有异常数据辨识方法可分为单体异常数据辨识方法和组合异常数据辨识方法2类^[12]。常见单体异常数据辨识方法包括基于回归分析^[13-14]、聚类分析^[15-16]、距离分析^[17]、支持向量机^[18-19]、统计特征^[20]等方法。然而,单体异常数据辨识方法易受噪声干扰,在处理复杂异常数据时,可能会产生误判或漏检的情况。组合异常数据辨识方法是指将多种单体异常数据辨识方法结合起来,通过融合多种单体异常数据辨识方法的优势,充分发挥不同单体异常数据辨识方法之间的互补性,实现更为高效、精准的异常数据识别,适用于处理复杂的异常数据^[21]。

组合异常数据辨识方法相较于单体异常数据辨识方法,能够更全面地考虑多维度数据,具有全面、准确、鲁棒性高的优点,广泛应用于光伏功率异常数据辨识领域^[22-23]。目前,针对光伏功率异常数据辨识研究较多。文献[24]考虑到同一区域内光伏发电曲线具有相似度,提取刻画了区域光伏聚合特征曲线,并通过发电曲线和光伏聚合特征曲线的相似度作为辨识异常数据的判据。文献[25]利用灰色关联度法来分析气象相关程度,并根据与目标光伏电站具有相似气象条件的出力数据来对异常数据进行辨识。文献[26]提出了使用高斯拟合函数和相似日系数共同构建相似日选取函数的基本思路,用以辨识光伏发电的异常数据,实现短期光伏发电

预测模型的数据预处理。然而,上述研究中对相似度的度量均采用基于欧氏距离的方法,该方法在用于受环境影响较大的光伏功率异常数据辨识时,难以实现数据异步匹配,且受异常数据点干扰严重,导致其相似度度量准确性低。文献[27]为建立光伏出力预测模型,采用四分位法检测原始数据中的异常值。文献[28]考虑到光伏出力受到多变量影响,采用2步四分位法对光伏阵列异常数据进行清洗。文献[29]将异常数据划分为堆积型和离散型,并利用四分位-快速密度峰值聚类组合方法进行数据清洗。然而,这些研究中使用的1步或2步四分位法仅基于单个数据点与四分位数之间的差异来判断异常值,未关注异常值之间可能具有的相关性或依赖关系,对由多种因素影响而产生的光伏异常数据辨识能力弱。同时,分布式光伏运行数据通常呈现非正态分布的特点,而2步四分位法在非正态分布的数据集上的效果较差,无法有效检测出异常值。

针对上述问题,本文提出一种基于动态时间弯曲(dynamic time warping, DTW)-两阶四分位的光伏发电异常数据辨识方法,通过光伏功率连续型、离散型异常数据辨识方法分别解决由外部环境变化和系统内部故障产生的连续性异常数据和离散型异常数据。首先,通过光伏功率 K 均值聚类算法对相似日进行划分,并采用同时段均值对比算法对相似日聚类下各时段光伏功率均值进行对比;其次,采用DTW算法进行光伏出力曲线的整体形状相似度判定并计算轮廓系数,根据轮廓系数辨识并剔除连续型异常数据;接着,根据两阶光伏功率方差变化率判据法辨识并剔除堆积离散型异常数据;最后,根据四分位法辨识剩余的少量离散型异常数据。并将本文方法与现有的四分位法、3-Sigma方法等算法进行对比,比较各方法异常数据辨识前后辐照度和光伏功率的线性相关程度,证明了本文所提方法在光伏功率异常数据辨识中的高效性和可靠性。本文主要贡献总结如下。

1) 提出考虑异常值相关性的连续型异常数据辨识方法,在确定最优时段和最优阈值系数后,基于同时段光伏功率均值对比剔除异常数据并考虑

光伏发电曲线的波动性,采用基于 DTW 与欧氏距离的曲线相似度方法对光伏发电曲线的时间序列进行相似度分析。能够更全面地考虑数据的波动特性,提高异常数据辨识的全面性、准确性和鲁棒性。

2)提出基于 DTW-两阶四分位的光伏离散异常数据辨识算法,在进行离散型数据异常识别时,采用 DTW 算法滤掉一阶、二阶四分位数据样本中相似度较高的离散数据,降低数据计算复杂度,提升一阶、二阶四分位数据样本融合效率,并通过两阶四分位方法,同时采用一阶变化率和二阶变化率对融合后的数据进行离散型异常数据剔除,有效解决了传统 2 步四分位法在非正态分布的数据集上效果较差、无法有效检测出异常值的问题。提高异常数据剔除的高效性和准确性。

1 光伏异常数据类型

光伏功率数据易受到外部环境变化和系统内部故障的影响,产生大量的异常功率数据。其中,外部环境变化包括天气、光伏组件表面污染等;系统内部故障包括设备故障和通信故障等。在光伏电站的实际运行中,大风、暴雨、大雪等因素会导致光照不足或光照不均产生异常功率数据,光伏组件、逆变器、电缆等设备出现故障或老化可能导致光伏功率数据出现持续性下降或波动而产生异常功率数据,这些光伏功率数据可分为连续型异常数据和离散型异常数据。

若外部环境变化和系统设备故障导致光伏功率数据出现长时间的波动或下降,为连续型异常数据,其主要特征为长时间内的光伏功率数据偏低,甚至功率数据为 0,通常连续型异常数据对分布式光伏的运行影响较大。若系统设备或外界环境导致光伏功率出现突然变化或跳跃,则为离散型异常数据,其特征多为突发时间短、恢复快,因此离散型异常数据对光伏发电系统影响较小,可通过故障处理在短时间内恢复正常。

2 光伏功率连续型异常数据辨识

分布式光伏运行过程中,可能发生设备故障、弃光限电及机组停机等事件,导致在连续的一段时间内,光伏功率低于正常值,产生大量连续性异常数据,影响分布式光伏系统安全稳定运行。本文提出了一种光伏功率连续型异常数据辨识方法,首先,基于 K 均值聚类算法进行相似日划分,在确定最优时段和最优阈值系数后,基于同时段光伏功率

均值对比剔除异常数据;其次,考虑光伏发电曲线的波动性,采用基于 DTW 与欧氏距离的曲线相似度判定方法对光伏发电曲线的时间序列进行相似度分析;最后,基于轮廓系数进一步剔除连续型异常数据,具体流程如图 1 所示。

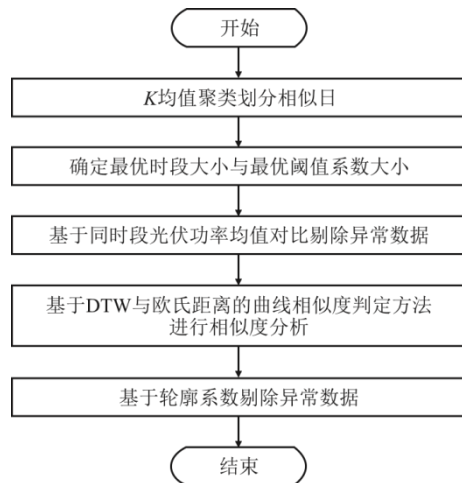


图 1 光伏功率连续型异常数据辨识方法流程
Fig.1 Flow chart of the PV power continuous abnormal data identification method

2.1 基于光伏功率均值对比的异常数据辨识方法

本文提出基于光伏功率均值对比的异常数据辨识方法。首先,根据不同的日辐照度对光伏功率数据进行相似度分析,并将日辐照度相似的光伏功率数据划分为同一相似日进行聚类。然后,先给定一个连续性异常数据的阈值系数,在辐照度相同的情况下将该阈值系数划分到不同的时段,异常数据识别效果最好的为最优时段,再基于此最优时段,分别将满足异常数据剔除率限值的所有阈值系数分配到该最优时段,通过比较此最优时段下不同阈值的异常数据识别效果,以确定最优时段下的最优阈值系数。

2.1.1 相似日划分

分布式光伏存在海量光伏功率数据,而光伏功率受太阳辐照度影响最大,为减小计算量,本文采用均值聚类算法^[30],在获得光伏功率和对应的辐照度数据后,先对辐照度数据进行归一化处理,再根据不同的日辐照度对光伏功率数据进行相似度分析,并将日辐照度相似的光伏功率数据划分为同一相似日进行聚类。归一化后的数据 S_o^* 可表示为:

$$S_o^* = \frac{S_o - S_{o,\min}}{S_{o,\max} - S_{o,\min}} \quad (1)$$

式中: S_o 为太阳辐照度数据; $S_{o,\min}$ 为最小太阳辐照

度数据； $S_{0,\max}$ 为最大太阳辐照度数据。

采用欧氏距离作为相似度的评价指标，将日辐照度相似的光伏功率数据划分为同一相似日进行聚类。考虑到聚类个数 K 值大小可能影响到相似日的准确划分，需要先确定合理的相似日聚类数目。本文采用手肘分析法来确定聚类数目，通过分析不同光伏电站在不同聚类数目下的损失函数，以确定最佳聚类数目，具体方法如下：

$$\begin{cases} \Delta S_r = S_r - 2S_{r-1} + S_{r-2} \\ \Delta S_{r,\text{first}} < \gamma_{\max}, K_{\text{best}} = r - 1 \end{cases} \quad (2)$$

式中： S_r 为聚类数目 r 的损失函数； ΔS_r 为 r 到 $r-2$ 的损失函数下降的差值； γ_{\max} 为损失函数的最大值； $\Delta S_{r,\text{first}}$ 为 S_r 中 r 的第 1 次取值，当 $\Delta S_{r,\text{first}} < \gamma_{\max}$ ，可得到最佳聚类数目 K_{best} 。

2.1.2 确定最优时段

在相似日聚类下，光伏功率数据划分的时段大小和阈值系数的选取均会对连续型异常数据的识别效果产生影响。为实现连续型异常数据的高质量识别，采用“定阈值系数，变时段大小”和“定时段大小，变阈值系数”的方法实现最优时段大小和最佳阈值系数的选取。“定阈值系数，变时段大小”是通过对比阈值系数固定的情况下，划分不同时段大小进行的连续型异常数据的识别，可实现最优时段大小的选取；“定时段大小，变阈值系数”是通过固定最优时段大小，比较不同阈值系数下连续型异常数据的识别效果。

以某相似日聚类下 n 天的光伏功率数据为例进行说明，该相似日聚类的光伏功率数据可定义为 $\mathbf{W}=[W_1, W_2, \dots, W_n]^T$ 。设每日光伏功率被划分为 t 个功率时段，且每个时段包含 c 个光伏功率数据，则相似日聚类第 t 个时段的总光伏功率数据 \mathbf{W}_t 为：

$$\mathbf{W}_t = \begin{bmatrix} W_{1,t,1} & W_{1,t,2} & \cdots & W_{1,t,c} \\ \vdots & & & \vdots \\ W_{n,t,1} & W_{n,t,2} & \cdots & W_{n,t,c} \end{bmatrix} \quad (3)$$

式中： $W_{n,t,c}$ 为第 n 天第 t 个时段的第 c 个光伏功率数据。对第 t 个时段的光伏功率数据进行均值计算，得到相似日聚类下 n 天第 t 个时段的光伏功率均值 $\bar{\mathbf{W}}_t=[\bar{W}_{1,t}, \bar{W}_{2,t}, \dots, \bar{W}_{n,t}]^T$ 。

当连续型异常数据的均值阈值系数 φ 一定时，可通过相似辐照度下的最大功率均值实现连续型异常数据均值的阈值划分，即：

$$\mathbf{W}_{s,t} = \varphi \max([\bar{W}_{1,t}, \bar{W}_{2,t}, \dots, \bar{W}_{n,t}]^T) \quad (4)$$

式中： $\max([\bar{W}_{1,t}, \bar{W}_{2,t}, \dots, \bar{W}_{n,t}]^T)$ 为某相似日聚类下第 t 个时段的最大功率均值； $\mathbf{W}_{s,t}$ 为某相似日聚类下第 t 个时段均值阈值。则通过对比阈值系数固定情况下不同时段连续型异常数据识别效果，选择连续型异常数据识别效果最好的时段作为最优时段。

2.1.3 确定最优阈值系数

在确定的最优时段下，均值阈值系数 φ 的取值大小会影响最终异常数据辨识后整体数据的线性相关程度，为提高辐照度和光伏功率的线性相关程度，设 ζ 为异常数据剔除率的限值，即异常数据剔除率的上限，避免过多数据被剔除影响光伏数据的连续性、完整性^[12]。现定义 φ 为可变值，取值集合为 $[\varphi_1, \varphi_2, \dots, \varphi_x, \dots, \varphi_m]$ ， $x \in [1, \dots, 2, \dots, m]$ ， ζ_{φ_x} 为阈值系数 φ_x 下的异常数据剔除率，则在阈值系数为 φ_x 情况下，第 x 个相似日聚类经过异常数据辨识后的正常功率数据为 \mathbf{W}_{φ_x} ，最优阈值系数的选取公式为：

$$\begin{cases} \max(a\tilde{\mathbf{W}}_{\varphi_1}, a\tilde{\mathbf{W}}_{\varphi_2}, \dots, a\tilde{\mathbf{W}}_{\varphi_m}) \\ \zeta_{\varphi_x} \leq \zeta_{\max} \end{cases} \quad (5)$$

式中： $\tilde{\mathbf{W}}_{\varphi_1}, \tilde{\mathbf{W}}_{\varphi_2}, \dots, \tilde{\mathbf{W}}_{\varphi_m}$ 分别为异常数据辨识后整体的光伏功率数据； $a\tilde{\mathbf{W}}_{\varphi_1}, a\tilde{\mathbf{W}}_{\varphi_2}, \dots, a\tilde{\mathbf{W}}_{\varphi_m}$ 分别为异常数据辨识后整体辐照度和光伏功率的线性相关程度，在满足各相似日聚类数据剔除率均小于 ζ_{\max} 的情况下，选取使整体数据线性相关程度最大的均值阈值系数 φ_x 。

相似日聚类下光伏功率数据的大小相近，划分时段大小和阈值系数均会对光伏功率连续型异常数据识别产生影响。过大的阈值系数或过小的划分时段都会降低光伏功率连续型异常数据识别效率，而过小的阈值系数或过大的划分时段会导致光伏功率连续型异常数据识别不准确，且难以得到具体时段的异常数据识别情况。仅通过时段大小确定阈值系数，无法保证时段大小划分的合理性，非最优时段的最优阈值系数可能导致识别的异常数据太少或太多，若直接剔除异常数据会影响光伏数据的连续性、完整性。

最后，将选取使整体数据线性相关程度最大的均值阈值系数定为当前实验的阈值系数，重复执行 2.1.2 和 2.1.3 小节所述步骤，当确定的时段大小和确定的阈值系数相互为最优时，则所求最大均值阈值系数为确定最优时段时选取的均值阈值系数。

2.2 基于 DTW 与欧氏距离异常数据检测方法

考虑到基于光伏功率均值大小对比的异常数

据辨识方法忽略了分布式光伏发电曲线波动性对相似光伏发电曲线的时间序列进行相似度分析,为提高异常数据识别的准确性。首先,通过DTW分析光伏出力和光照强度的时间序列在各个时间点的弯曲程度;其次,结合曲线的整体分布特性、整体动态特性和局部动态特性,将DTW与欧氏距离相结合进行曲线相似度判定,并使用轮廓系数评估聚类结果,以检测是否存在异常数据。

2.2.1 曲线相似度判定

定义 u 和 v 分别为光伏出力曲线的时间序列、光照强度曲线的时间序列,则 u 和 v 之间的DTW距离可用 $D(u,v)$ 表示,其递推公式为:

$$\begin{aligned} DTW(u,v) &= D(u,v) + \min\{DTW(u,v-1), \\ &DTW(u-1,v) + DTW(u-1,v-1)\} \end{aligned} \quad (6)$$

式中: $D(u,v)$ 为 u 和 v 间的距离, $u=1,2,3,\dots,u_i,\dots,u_n$, $v=1,2,3,\dots,v_i,\dots,v_n$,且 $D(0,0)=0$, $D(u,0)=D(0,v)\rightarrow+\infty$ 。

基于式(6),光伏出力曲线时间序列 u 、光照强度曲线时间序列 v 的相似度量函数 $s(u,v)$ 为:

$$s(u,v) = \beta_1 \sqrt{\sum_{i=1}^n (u_i - v_i)^2} + \beta_2 DTW(u,v) + \beta_3 DTW(u',v') \quad (7)$$

式中: $u'=(u^1, u^2, \dots, u^n)$,且 $u^i=(u_{i+l}-u_i)/\Delta i$,同理, $v'=(v^1, v^2, \dots, v^n)$,且 $v^i=(v_{i+l}-v_i)/\Delta i$ 。 $D(u,v)$ 作为 u 和 v 的动态时间弯曲距离,表示整体动态特性;

$\sqrt{\sum_{i=1}^n (u_i - v_i)^2}$ 为 u 和 v 的欧氏距离,表示整体分布

特性; $DTW(u',v')$ 为 u 和 v 局部动态特性对应的动态时间弯曲距离; β_1 、 β_2 、 β_3 分别为整体分布特性、整体动态特性和局部动态特性的权重,用以得到参数间重要关系并在一定程度平衡三者数量级^[28]。

$s(u,v)$ 值越小表示2条曲线之间相似度越高。

2.2.2 轮廓系数计算

轮廓系数是一种用于评估数据聚类质量的指标,可准确识别出与正常光伏发电数据有显著差异的异常数据。首先,在计算出 u 和 v 的相似度后,将相似的数据点 k 分配到同一类;其次,计算每个数据点 k 的轮廓系数,主要通过计算数据点与其所属类别内其他数据点的距离和与最近邻类别的平均距离之间的差异来衡量数据点的聚类紧密度,数据点 k 的轮廓系数计算公式为:

$$H_k = \frac{p_k - q_k}{\max\{p_k, q_k\}} \quad (8)$$

式中: p_k 为数据点 k 到其所属类别内其他数据点的平均距离,称为内聚度; q_k 为数据点 k 到最近邻类别数据点的平均距离,称为分离度。轮廓系数 H_k 的值在 $[-1,1]$ 之间, H_k 越接近1,表示数据点与其所属类别内的其他数据点距离越近,与最近邻类别的距离越远,说明数据点聚类得较好,是正常数据。 H_k 越接近-1,表示数据点与其所属类别内的其他数据点距离较远,与最近邻类别的距离较近,说明数据点与其他类别更相似,可能是异常数据。

3 光伏功率离散型异常数据辨识

光伏功率离散型异常数据中,通常存在大量的堆积离散型异常数据点和少量剩余的离散型异常数据点。本文提出了一种基于DTW-两阶四分位的光伏离散异常数据辨识算法,具体流程如图2所示。

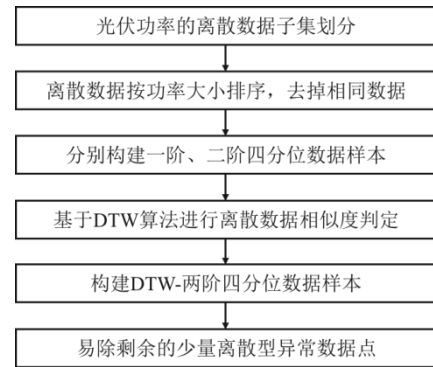


图2 基于DTW-两阶四分位的光伏离散异常数据辨识流程
Fig.2 Flow chart of the PV power discrete abnormal data identification algorithm based on DTW and two-stage quartile

由图2可见:首先,考虑数据的方差一次变化率,利用四分位法对数据进行处理,将数据样本均分为4个部分数据集,并剔除一阶变化率异常的数据,获得一阶四分位数据样本;其次,考虑方差一阶变化率的局限性,进一步对数据的一阶变化率求差分,同样利用四分位法对数据进行处理,剔除二阶变化率异常的数据,将一阶四分位数据样本处理为二阶四分位数据样本;最后,采用DTW算法滤掉一阶、二阶四分位数据样本中相似度较高的离散数据。降低数据计算复杂度的同时,提升一阶、二阶四分位数据样本融合效率,再通过两阶四分位方法对融合后的数据进行离散型异常数据剔除,实现异常数据剔除的高效性和准确性,为光伏发电异常数据的精准辨识提供保障。

3.1 方差变化率判断辨识

根据辐照度区间将光伏功率离散数据集 R 划分

为 N 个离散数据子集, 并根据光伏功率的大小对离散数据进行排序, 去掉离散数据子集中的相同数据, 得到的离散数据集表示为 $\mathbf{R}=\{R(1), \dots, R(n), \dots, R(N)\}$, 辐照度区间的划分间距一般取 $20 \text{ W}\cdot\text{m}^2$ 。其中, $R(n)=\{p_{n,1}, \dots, p_{n,d}, \dots, p_{n,D}\}$ 表示第 n 个辐照度区间的数据子集, D 为辐照度区间内的数据样本总数, $p_{n,d}$ 为 $R(n)$ 中第 d 个数据样本的光伏功率, 且满足 $p_{n,d} < p_{n,d+1}$ 。

$R(n)$ 中第 d 个数据样本的方差是否发生明显变化可根据其方差的一阶变化率来判断, 定义 $\delta_{n,d}$ 为每个辐照度区间内各个功率点的方差一阶变化率, 计算公式为:

$$\delta_{n,d} = \frac{1}{20} \left| \frac{\sum_{g=1}^d (p_{n,g} - \bar{p}_{n,d})^2}{d} - \frac{\sum_{g=1}^{d-1} (p_{n,g} - \bar{p}_{n,d-1})^2}{d-1} \right| \quad (9)$$

式中: $\bar{p}_{n,d}$ 为 $R(n)$ 中前 d 个数据样本点的光伏功率的平均值, $d=2, 3, \dots, D$ 。

针对不同的辐照度段, 根据历史经验数据设置对应的方差一阶变化率阈值 $\delta_{n,\max}^{[29]}$, 判定方差一阶变化率大于阈值的数据为异常数据, 定义 $R(n)$ 经过一阶变化率判据剔除异常数据之后的数据集为 $R_\delta(n)$, 表示为:

$$R_\delta(n) = \{p_{n,d} \mid p_{n,d} \in R(n), \delta_{n,d} \leq \delta_{n,\max}\} \quad (10)$$

考虑到方差一阶变化率的局限性, 进一步对一阶变化率求差分, 得到功率点方差的二阶变化率 $s_{n,d}$:

$$s_{n,d} = |\delta_{n,d} - \delta_{n,d-1}|, \quad d=3, 4, \dots, D \quad (11)$$

光伏功率方差二阶变化率曲线平稳部分对应光伏功率正常点, 而曲线非平稳部分对应光伏功率异常点。利用箱型图法, 根据每个辐照度区间的光伏功率数据整体分布来辨识光伏功率方差二次变化率曲线非平稳区段, 判定方差二阶变化率大于箱型图外限 $s_{n,\max}$ 的数据为异常数据, 定义 $R(n)$ 经二阶变化率判据剔除异常数据后的数据集为 $R_s(n)$:

$$R_s(n) = \{p_{n,d} \mid p_{n,d} \in R(n), s_{n,d} \leq s_{n,\max}\} \quad (12)$$

3.2 DTW-两阶四分位异常数据辨识算法

DTW-两阶四分位异常数据辨识算法融合了一阶四分位数据样本和二阶四分位数据样本, 构建两阶四分位数据样本, 再结合 DTW 相似度判定方法, 准确高效地辨识和剔除剩余少量离散型异常数据。

首先, 将 $R_\delta(n)$ 和 $R_s(n)$ 中的光伏功率离散数据按照升序重新排列, 分别得到一阶四分位数据样本

和二阶四分位数据样本; 其次, 将一阶四分位数据样本和二阶四分位数据样本分别看作 2 个序列, 基于 DTW 算法进行离散数据相似度判定, 去掉相似度较高的数据, 融合剩下的一阶、二阶四分位数据样本, 构建 DTW-两阶四分位异常数据样本; 然后, 计算 DTW-两阶四分位数据样本的中位数 $\Theta_{n,2}$ 、第 1 四分位数 $\Theta_{n,1}$ 和第 3 四分位数 $\Theta_{n,3}$, 并计算四分位距 $\Theta_{n,3} - \Theta_{n,1}$; 最后, 根据四分位距 $\Theta_{n,3} - \Theta_{n,1}$ 计算两阶四分位异常数据内限, 定义两阶四分位异常数据内限为 $[U_{n,\min}, U_{n,\max}]$, 可表示为:

$$[U_{n,\min}, U_{n,\max}] = [\Theta_{n,1} - \alpha\Theta_{n,2}, \Theta_{n,3} + \alpha\Theta_{n,2}] \quad (13)$$

式中: α 为边界系数, 可根据不同需要调整取值, 通常取 1.5。

判定处于 $[U_{n,\min}, U_{n,\max}]$ 之外的数据为异常数据并剔除, 得到剔除全部离散型异常数据的光伏功率数据集 $R_0(n)$ 。按相同方法对其他辐照度区间的光伏功率数据进行处理, 得到剔除全部连续型异常数据和离散型异常数据的分布式光伏功率数据集 $\mathbf{R}_0 = \{R_0(1), \dots, R_0(n), \dots, R_0(N)\}$ 。

3.3 算法复杂度分析

本文方法主要包括数据连续检测与排序、DTW 相似度判定、数据融合、四分位法剔除异常数据 4 个主要步骤。设 1 组按时间排序的光伏功率数据为 $P = \{p_1, \dots, p_t, \dots, p_T\}$, 其中 T 为剔除前数据总数。

数据连续检测与排序的计算复杂度通常取决于数据数量和排序算法, 常用的较为高效的排序算法, 如归并排序和快速排序等算法在最坏情况下的计算复杂度为 $O(T \log T)$ 。

DTW 相似度判定的计算复杂度同样取决于数据数量, 最坏情况下的计算复杂度为 $O(T^2)$ 。

数据融合主要对剔除部分离散型异常数据后的数据样本进行融合处理, 其计算复杂度取决于剔除部分离散型异常数据后的数据数量, 在最坏情况下的计算复杂度为 $O(T^2)$, 其中, T' 为剔除后数据总数, $1 < T' < T$ 。

四分位法剔除异常数据的计算复杂度主要来自于对数据的排序, 因此, 四分位法剔除异常数据的计算复杂度与数据连续检测与排序的计算复杂度相同, 均为 $O(T \log T)$ 。

综上, 对于一组数据总数为 T 的光伏功率数据, 本文所提方法在最坏情况下的总计算复杂度为 $O(2T \log T + T^2 + T'^2) = O(T^2)$ 。

4 故障预警

通过观察试验结果,在某些时段内,模型未能有效剔除连续型和离散型异常数据,导致辐照度与光伏功率之间的线性相关程度提升不明显,甚至在某些情况下出现了下降。因此,进一步构建故障预警模型,在剔除异常光伏数据后,计算辐照度与光伏功率之间的线性相关程度,基于正常光伏数据辐照度与光伏功率的线性相关程度计算预警阈值。当线性相关程度高于预警阈值,表明异常数据被有效剔除;若线性相关程度低于预警阈值,需进一步调整阈值系数以及 DTW 算法的相关参数,以提高异常数据剔除的准确性。图 3 为故障预警流程。

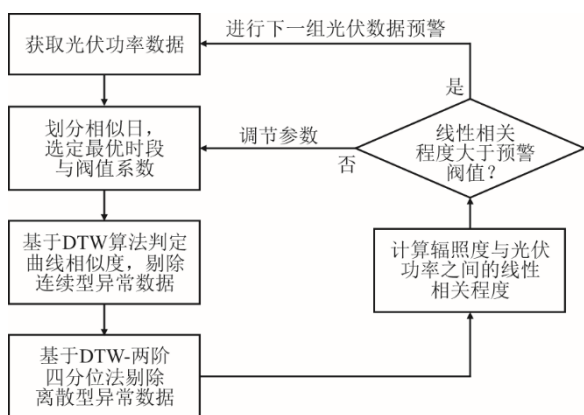


图 3 故障预警流程
Fig.3 Flow chart of fault warning

5 实验验证与分析

5.1 算例分析与对比算法

以某地市分布式光伏 1 年的实测辐照度数据和光伏数据为研究对象,测试所提方法对各分布式光伏连续型和离散型异常数据的识别和剔除效果。

该地 1 年内阳光充足、降雨较少、光伏发电功率大。考虑连续型异常功率数据随设备故障、通信故障、弃光限电及机组停机等事件产生;少量离散型异常数据由通信短时异常导致光伏功率记录数据突变产生。首先,对日辐照度与光伏功率数据进行归一化处理,将日辐照度相似的光伏功率数据划分为同一相似日进行聚类,图 4、图 5、图 6 分别为聚类数量为 6 时测得的实验结果、文献[12]异常数据剔除率限值 ζ 设定为 20%结果、文献[28]整体分布特性、整体动态特性和局部动态特性的权重 β_1 、 β_2 、 β_3 分别设定为 0.2、0.5、0.3 结果;基于文献[29],方差一阶变化率阈值 $\delta_{n,\max}$ 的取值范围设定为 [0.5, 1.5]。然后,在“定阈值系数,变时段大小”阶

段,得到最优时段大小;在“定时段大小,变阈值系数”阶段,得到最优阈值系数,其中阈值系数 $\varphi=\{0.2,0.3,0.4,0.5\}$,时段大小为 $\{1,2,3,4\}$,对应时段数量 $t=\{24,12,8,6\}$,模型参数见表 1。最后,通过与传统算法对比检验本文所提算法剔除异常数据的有效性。

表 1 模型参数
Tab.1 Model parameters

项目	数值	项目	数值
ζ	20%	β_1	0.2
$\delta_{n,\max}$	[0.5,1.5]	β_2	0.5
t	[6,8,12,24]	β_3	0.3
φ	[0.2,0.3,0.4,0.5]	K_{best}	6.0

本文通过与以下 4 种算法对比来验证所提算法性能,对比算法具体设置如下。

1) 四分位法^[30] 该算法利用方差变化率判据识别光伏功率曲线上下两侧的堆积型异常数据点,再利用四分位法识别剩余的离散型异常数据点。但是对连续型异常数据的剔除效果较差。

2) 3-Sigma^[31]算法 该算法利用正态分布 3-Sigma 原则处理异常数据以提高辐照度与光伏功率的相关性,但在对不同辐照度数据进行分段或聚类后难以保证其正态性,导致异常数据剔除过程中误差较大。

3) 2 步分位数算法^[32] 该算法将光伏样本数据按照不同种类数据,得到样本数据集。基于单一维度的分位数算法,对不同的样本数据集进行计算,最后对数据进行归类整理,得到阵列不同维度下的正常数据集与离散型异常数据集。但是对连续型异常数据的剔除效果较差。

4) 基于特征聚类的异常数据辨识算法^[33] 该算法基于特征聚类分析技术,统计各电源历史数据,提取其分布规律,基于特征聚类分析进行数据校验,从而剔除异常数据。但是无法确定最优时段与最优阈值系数,导致异常数据剔除效果较差。

5.2 实验结果分析

图 4 为在阈值系数给定的情况下,轮廓系数与数据剔除率随选取时段大小的变化情况。轮廓系数可表示剔除后数据与正常光伏发电数据的相似程度,轮廓系数越高,说明异常数据剔除效果越好。数据剔除率表示所剔除数据占全部数据的比例,在处理时应保证数据剔除率不会过高,过高的数据剔除率可能导致有效数据被剔除。由图 4 可见,相较

于未经剔除的原始数据，本文所提方法可以有效剔除异常数据，提升轮廓系数。在阈值系数一定的情况下，随着时段大小取值不断增大，数据剔除率不断提高，而轮廓系数不断下降。这是因为选取的时段大小越大，时段内光伏功率数据数值范围变大，经同时段光伏功率均值对比算法剔除的数据比例大幅度上升，数据剔除率过高，导致部分非异常数据被剔除，轮廓系数下降。

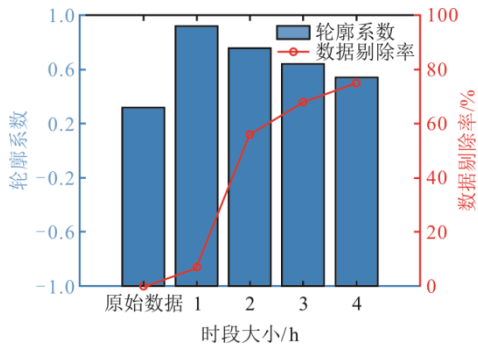


图 4 不同时段大小下连续型异常数据辨识效果
Fig.4 Identification effect of continuous abnormal data under conditions with different sizes of time periods

图 5 为在时段大小给定的情况下，轮廓系数与数据剔除率随阈值系数的变化情况。原始数据表示未处理的数据，即阈值系数为 0。由图 5 可见，随着阈值系数由 0 增加至 0.3，数据剔除率不断提高，轮廓系数随之增大。随着阈值系数由 0.3 增加至 0.5，数据剔除率不断提高，但轮廓系数却不断减小。其原因在于选取的阈值系数越大，在同时段光伏功率均值对比算法中被判定为异常的数据比例越高，在阈值系数大于 0.3 时，数据剔除率过高，将有部分正常数据被判定为异常数据，因此轮廓系数下降。本文通过实验证明阈值系数选为 0.3 时，轮廓系数最大。

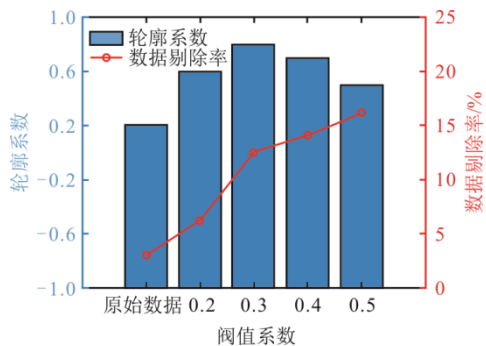


图 5 不同阈值系数下连续型异常数据辨识效果
Fig.5 Identification effect of continuous abnormal data with different threshold coefficients

图 6 通过比较在确定最优时段和最优阈值系数

的情况下，不同算法下连续型异常数据辨识效果。由图 6 可见，所提算法可以有效地识别和剔除连续型异常数据。其中，算法 1 为四分位法，算法 2 为 3-Sigma 算法，算法 3 为 2 步分位数算法，算法 4 为基于特征聚类的异常数据辨识算法。相较于四分位法、3-Sigma 算法、2 步分位数算法与基于特征聚类的异常数据辨识算法，本文所提算法在轮廓系数分别提高了 58.77%、67.73%、52.59%、72.52%，在数据剔除率分别提高了 41.52%、58.20%、39.24%、61.38%。这是因为所提算法通过对相似辐照度下光伏功率均值大小的对比和基于 DTW 与欧氏距离的综合性曲线的相似度判定，可以准确地识别和剔除连续型异常数据。同时，基于两阶四分位的方法，可以有效地识别和剔除光伏功率离散型异常数据。四分位法基于数据分布情况辨识异常数据，剔除离散型异常数据，无法有效剔除连续型异常数据。3-Sigma 算法基于正态分布 3-Sigma 原则处理异常数据，在非正态分布下的数据剔除效果较差。2 步分位数算法对不同的样本数据集按照单一维度的分位数算法步骤进行计算，最后对数据进行归类整理，得到阵列不同维度下的正常数据集与离散型异常数据集，但是对连续型异常数据的剔除效果较差。基于特征聚类的异常数据辨识算法基于特征聚类分析进行数据校验，从而剔除异常数据。但是无法确定最优时段与最优阈值系数，导致异常数据剔除效果较差。

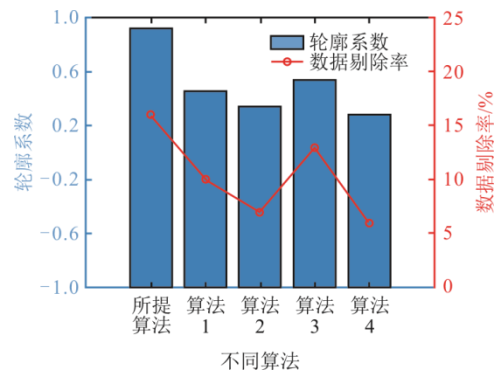


图 6 不同算法下连续型异常数据辨识效果
Fig.6 Identification effect of continuous abnormal data using different algorithms

图 7 通过散点图描述了本文所提算法对异常数据的剔除效果。由图 7 可以看出，所提算法可以有效地识别和剔除连续型异常数据和离散型异常数据。其中，连续型异常数据表现为功率较正常值偏低；离散型异常数据表现为离散分布且与其他光伏

功率数据差异较大。由于连续型异常数据在整体异常数据中的占比较大,在相同辐照度下,存在连续型异常数据的光伏功率数据均值较小,光伏功率数据的线性拟合直线斜率较小。因此光伏功率数据的线性拟合直线斜率可体现连续型异常数据剔除情况,斜率越高说明连续型异常数据剔除效果越好。

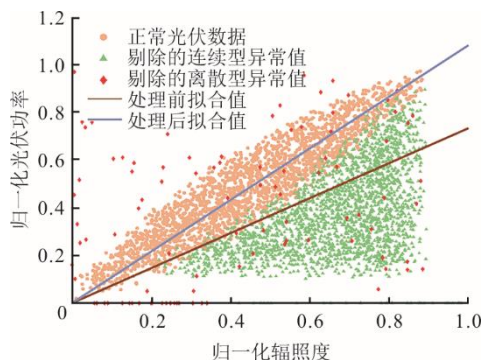


图7 所提算法对异常数据的剔除效果
Fig.7 Elimination effect of abnormal data using the proposed algorithm

由图7可以看出,经所提算法剔除异常数据后,光伏功率数据的线性拟合直线斜率变大,且更能拟合正常光伏功率数据分布情况,证明了所提算法对连续型异常数据的高效剔除能力。

表2展示了不同算法处理异常数据后辐照度与光伏功率的线性相关程度。由表2可见,所提算法对光伏功率异常数据的剔除效果明显优于四分位法和3-Sigma算法,所提算法剔除异常数据后的线性相关程度最高。相较于四分位法和3-Sigma算法,所提算法的线性相关程度变化分别提高了58.15%和68.41%。因为所提算法通过DTW算法进行光伏出力曲线的整体形状相似性判定,并计算轮廓系数剔除连续型异常数据,然后进一步通过两阶四分位法剔除离散型异常数据,提高了辐照度与光伏功率的线性相关程度。四分位法基于数据分布情况辨识异常数据,可有效剔除离散型异常数据,但对连续型异常数据剔除效果较差。3-Sigma算法利用正态分布3-Sigma原则处理异常数据,难以保障非正态分布下的数据剔除效果。2步分位数算法对不同的样本数据按照单一维度的分位数算法步骤进行计算,最后对数据进行归类整理,得到阵列不同维度下的正常数据集与离散型异常数据集,但是对连续型异常数据的剔除效果较差。基于特征聚类的异常数据辨识算法基于特征聚类分析进行数据校验,从而剔除异常数据。但是无法确定最优时段与最优阈值系数,导致异常数据剔除效果较差。

表2 不同算法处理异常数据后辐照度与光伏功率的线性相关程度 单位: %
Tab.2 Linear correlation between radiation and PV power after abnormal data processing by different algorithms

方法	剔除前线性相关程度	剔除后线性相关程度	线性相关程度变化
四分位法	81.22	85.79	4.57
3-Sigma法	81.22	84.67	3.45
2步分位数算法	81.22	86.71	5.94
基于特征聚类的异常数据辨识算法	81.22	85.13	4.07
所提算法	81.22	92.14	10.92

6 结 语

为准确识别和剔除光伏发电异常数据,本文提出了基于DTW-两阶四分位的光伏发电异常数据辨识。该方法结合了光伏发电连续型和离散型异常数据的识别和剔除。通过对相似辐照度下光伏功率均值大小的对比和基于DTW与欧氏距离的综合性曲线的相似性判定,可以准确地识别和剔除连续型异常数据。同时,基于两阶四分位的方法,可以有效地识别和剔除光伏功率离散型异常数据。基于异常数据识别和剔除结果判断是否出现故障。实验结果表明,所提方法能提高光伏发电异常数据识别的准确性,并有效剔除异常数据。本文研究成果能够为分布式光伏发电性能评估、光伏功率预测及光伏并网等方面提供技术支撑,推动“碳达峰、碳中和”目标的实现以及新能源产业的发展。

[参 考 文 献]

- [1] 陈骏喙,张娜,刘广忱,等.基于EOSSA-ELM的光伏短期输出功率预测[J].可再生能源,2022,40(7):890-898.
CHEN Junhao, ZHANG Na, LIU Guangchen, et al. Photovoltaic short-term output power forecasting based on EOSSA-ELM[J]. Renewable Energy Resources, 2022, 40(7): 890-898.
- [2] 石理,刘朋矩,杜治钢,等.数字孪生辅助的智能楼宇多模态通信资源管理方法[J].电信科学,2023,39(1):60-71.
SHI Cheng, LIU Pengju, DU Zhigang, et al. Digital twin-assisted multi-mode communication resource management methods for smart buildings[J]. Telecommunications Science, 2023, 39(1): 60-71.
- [3] 王超,胡浩,郑炼,等.基于智慧蓄电池的光伏储能系统及其控制策略[J].可再生能源,2022,40(4):506-512.
WANG Chao, HU Hao, ZHENG Lian, et al. Photovoltaic energy storage system based on smart battery and its control strategy[J]. Renewable Energy Resources, 2022, 40(4): 506-512.
- [4] 张娜,任强,刘广忱,等.基于VMD-GWO-ELMAN的光伏功率短期预测方法[J].中国电力,2022,55(5):57-65.

- ZHANG Na, REN Qiang, LIU Guangchen, et al. PV power short-term forecasting method based on VMD-GWO-ELMAN[J]. *Electric Power*, 2022, 55(5): 57-65.
- [5] 张翔, 王翌, 周振宇, 等. 考虑光伏集成 5G 基站用能模式的多主体共享储能优化配置[J]. *电测与仪表*, 2023, 60(3): 97-106.
- ZHANG Xiang, WANG Zhao, ZHOU Zhenyu, et al. Optimal configuration of shared energy storage system for multi-entities considering PV integrated 5G base station energy consumption mode[J]. *Electrical Measurement & Instrumentation*, 2023, 60(3): 97-106.
- [6] 时珉, 尹瑞, 胡傲宇, 等. 基于滑动标准差计算的光伏阵列异常数据清洗办法[J]. *电力系统保护与控制*, 2020, 48(6): 108-114.
- SHI Min, YIN Rui, HU Aoyu, et al. A novel photovoltaic array outlier cleaning algorithm based on moving standard deviation[J]. *Power System Protection and Control*, 2020, 48(6): 108-114.
- [7] 张翔, 廖海君, 周振宇, 等. 含规模化 5G 基站租赁共享储能的配电网混合博弈优化调度[J]. *电测与仪表*, 2023, 60(5): 23-32.
- ZHANG Xiang, LIAO Haijun, ZHOU Zhenyu, et al. Hybrid game optimal dispatching for distribution network with large-scale 5G base station leasing shared energy storage[J]. *Electrical Measurement & Instrumentation*, 2023, 60(5): 23-32.
- [8] 吴家葆, 曾国辉, 张振华, 等. 基于 K-means 分层聚类的 TCN-GRU 和 LSTM 动态组合光伏短期功率预测[J]. *可再生能源*, 2023, 41(8): 1015-1022.
- WU Jiabao, ZENG Guohui, ZHANG Zhenhua, et al. Dynamic combination of TCN-GRU and LSTM photovoltaic short-term power prediction based on K-means hierarchical clustering[J]. *Renewable Energy Resources*, 2023, 41(8): 1015-1022.
- [9] 刘岩, 李文文, 周丽霞, 等. 基于高斯混合模型的光伏发电出力中高比例异常数据检测方法研究[J]. *电测与仪表*, 2021, 58(9): 14-21.
- LIU Yan, LI Wenwen, ZHOU Lixia, et al. Research of high proportion of outliers detection method in photovoltaic power output data based on Gaussian mixture model[J]. *Electrical Measurement & Instrumentation*, 2021, 58(9): 14-21.
- [10] 郭泽阳, 程拥军, 温家铭, 等. 基于支持向量机的分布式光伏异常数据序列提取[J]. *中国测试*, 2023, 49(7): 156-161.
- GUO Zeyang, CHENG Yongjun, WEN Jiaming, et al. Extraction of abnormal data sequence of photovoltaic power station based on support vector machine[J]. *China Measurement & Test*, 2023, 49(7): 156-161.
- [11] 赵晋斌, 张建平, 毛玲, 等. 基于 PSO-Soft attention 双向 LSTM 算法的光伏发电量预测研究[J]. *智慧电力*, 2022, 50(3): 1-7.
- ZHAO Jinbin, ZHANG Jianping, MAO Ling, et al. Photovoltaic power generation forecasting based on PSO-soft attention bidirectional LSTM algorithm[J]. *Smart Power*, 2022, 50(3): 1-7.
- [12] 叶林, 崔宝丹, 李卓, 等. 分布式光伏高比例异常运行数据组合识别方法[J]. *电力系统自动化*, 2022, 46(20): 74-82.
- YE Lin, CUI Baodan, LI Zhuo, et al. Combined identification method for high proportion of abnormal operation data in photovoltaic power station[J]. *Automation of Electric Power Systems*, 2022, 46(20): 74-82.
- [13] 梅玉杰, 李勇, 周王峰, 等. 基于机器学习的配电网异常缺失数据动态清洗方法[J]. *电力系统保护与控制*, 2023, 51(7): 158-169.
- MEI Yujie, LI Yong, ZHOU Wangfeng, et al. Dynamic data cleaning method of abnormal and missing data in a distribution network based on machine learning[J]. *Power System Protection and Control*, 2023, 51(7): 158-169.
- [14] 黄颖祺. 基于机器学习的用户侧异常数据诊断研究[J]. *制造业自动化*, 2023, 45(8): 49-55.
- HUANG Yingqi. Research on user-side data abnormal diagnosis based on machine learning[J]. *Manufacturing Automation*, 2023, 45(8): 49-55.
- [15] 刘丰硕, 刘然, 董子慧. 基于 k-means 聚类的智能电表异常动态数据检测方法[J]. *电子设计工程*, 2023, 31(5): 84-88.
- LIU Fengshuo, LIU Ran, DONG Zihui. Abnormal dynamic data detection method for smart meters based on k-means clustering[J]. *Electronic Design Engineering*, 2023, 31(5): 84-88.
- [16] 赵天辉, 张耀, 王建学. 基于空间密度聚类 and 异常数据域的负荷异常值识别方法[J]. *电力系统自动化*, 2021, 45(10): 97-105.
- ZHAO Tianhui, ZHANG Yao, WANG Jianxue. Identification method of load outlier based on density-based spatial clustering and outlier boundaries[J]. *Automation of Electric Power Systems*, 2021, 45(10): 97-105.
- [17] 王锋, 高欣, 贾欣, 等. 一种基于对数区间隔离森林的电力调度数据异常检测集成算法[J]. *电网技术*, 2021, 45(12): 4818-4827.
- WANG Feng, GAO Xin, JIA Xin, et al. An anomaly detection ensemble algorithm based on loginterval isolation[J]. *Power System Technology*, 2021, 45(12): 4818-4827.
- [18] 王璨, 马金辉, 王松, 等. 基于大数据的配电网运行线损异常诊断模型构建[J]. *自动化与仪表*, 2023, 38(3): 96-99.
- WANG Can, MA Jinhui, WANG Song, et al. Construction of abnormal diagnosis model of distribution network operation line loss based on big data[J]. *Automation & Instrumentation*, 2023, 38(3): 96-99.
- [19] 刘震宇. 基于支持向量机的电力系统调度数据异常检测方法[J]. *自动化技术与应用*, 2023, 42(6): 24-27.
- LIU Zhenyu. Anomaly detection method of power system dispatching data based on support vector machine[J]. *Techniques of Automation and Applications*, 2023, 42(6): 24-27.
- [20] 杨茂, 黄鑫. 基于光照过程特征分析的光伏功率异常数据识别算法[J]. *电力系统自动化*, 2019, 43(6): 64-69.
- YANG Mao, HUANG Xin. Abnormal data identification algorithm for photovoltaic power based on characteristics analysis of illumination process[J]. *Automation of Electric Power Systems*, 2019, 43(6): 64-69.
- [21] 吴悦, 张忠会, 游志刚, 等. 基于可视化和数据融合的分布式光伏电气设备异常运行状态检测方法[J]. *制造业自动化*, 2021, 43(9): 16-19.
- WU Yue, ZHANG Zhonghui, YOU Zhigang. Abnormal operation state detection method of electrical equipment in photovoltaic power station based on visualization and data fusion[J]. *Manufacturing Automation*, 2021, 43(9): 16-19.
- [22] 马速良, 武亦文, 李建林, 等. 聚类分析架构下基于遗传算法的电池异常数据检测方法[J]. *电网技术*, 2023,

- 47(2): 859-867.
MA Suliang, WU Yiwen, LI Jianlin, et al. Anomaly detection for battery data based on genetic algorithm under cluster analysis framework[J]. Power System Technology, 2023, 47(2): 859-867.
- [23] 周嘉琪, 毕利. 基于 GAN 的光伏逆变器数据异常检测技术[J]. 电力系统保护与控制, 2022, 50(1): 133-140.
ZHOU Jiaqi, BI Li. Abnormal detection technology of photovoltaic inverter data based on GAN[J]. Power System Protection and Control, 2022, 50(1): 133-140.
- [24] 曹俊, 王珂, 曹瑞峰, 等. 基于区域光伏发电曲线相似度的异常运行分析[J]. 电力大数据, 2022, 25(7): 43-52.
CAO Jun, WANG Ke, CAO Ruifeng, et al. Operational anomaly analysis based on regional photovoltaic power generation curve similarity[J]. Power Systems and Big Data, 2022, 25(7): 43-52.
- [25] 郭辉, 杨国清, 姚李孝, 等. 基于综合相似日和功率相关性的分布式光伏预测功率修正[J]. 电网与清洁能源, 2018, 34(9): 52-58.
GUO Hui, YANG Guoqing, YAO Lixiao, et al. Correction of predictive power of PV plants based on integrated similar days and power correlations[J]. Power System and Clean Energy, 2018, 34(9): 52-58.
- [26] 罗智骁. 基于聚类和相似日理论的光伏预测数据预处理模型研究[D]. 北京: 华北电力大学, 2021: 1.
LUO Zhixiao. Research on photovoltaic prediction data preprocessing model based on clustering and similar day theory[D]. Beijing: North China Electric Power University, 2021: 1.
- [27] 邢晨, 温蜜, 张照贝. 基于 SA 的光伏出力概率密度预测方法[J]. 计算机仿真, 2023, 40(5): 154-161.
XING Chen, WEN Mi, ZHANG Zhaobei. Photovoltaic output probability density forecasting method based on Seq2seq-attention(SA)[J]. Computer Simulation, 2023, 40(5): 154-161.
- [28] 薛文琦. 基于深度学习的光伏数据异常修复与出力预测方法研究[D]. 长沙: 湖南大学, 2022: 1.
XUE Wenqi. Research on photovoltaic data anomaly repair and output prediction method based on deep learning[D]. Changsha: Hunan University, 2022: 1.
- [29] 吴永斌, 张建忠, 邓富金, 等. 基于方差变化率判据-四分位的风电场功率异常数据识别[J]. 电力工程技术, 2023, 42(4): 141-148.
WU Yongbin, ZHANG Jianzhong, DENG Fujin, et al. Identification of wind farm power anomaly data based on variance change rate criterion-quartile[J]. Electric Power Engineering Technology, 2023, 42(4): 141-148.
- [30] 马良玉, 耿妍竹, 袁乃正, 等. 基于四分位与 CFSFDP 的风电机异常数据清洗方法[J]. 电力科学与工程, 2023, 39(6): 9-16.
MA Liangyu, GENG Yanzhu, YUAN Naizheng, et al. Abnormal data cleaning method of wind turbine based on quartile and CFSFDP[J]. Electric Power Science and Engineering, 2023, 39(6): 9-16.
- [31] 杨茂, 孟玲建, 李大勇, 等. 基于类 3σ 准则的光伏功率异常数据识别[J]. 可再生能源, 2018, 36(10): 1443-1448.
YANG Mao, MENG Lingjian, LI Dayong, et al. Identification of abnormal data of photovoltaic power based on class 3σ [J]. Renewable Energy Resources, 2018, 36(10): 1443-1448.
- [32] 胡傲宇. 光伏发电系统异常数据清洗及其应用研究[D]. 北京: 华北电力大学, 2020: 1.
HU Aoyu. Research oil abnormal data cleaning of photovoltaic power generation and its application[D]. Beijing: North China Electric Power University, 2020: 1.
- [33] 李智勇, 周剑, 俞蕙, 等. 基于特征聚类分析的大规模发电数据异常辨识方法[J]. 南方能源建设, 2018, 5(1): 35-39.
LI Zhiyong, ZHOU Jian, YU Hui, et al. An abnormal data identification method of large-scale generation data based on cluster analysis[J]. Southern Energy Construction, 2018, 5(1): 35-39.

(责任编辑 杜亚勤)