基于改进 K 均值聚类和加权动态时间规整的分布式光伏异常数据辨识方法

杨旺霞[1] 李本瑜[2] 翟苏巍[2] 石恒初[2] 李银银[2]
（1. 云南电网有限责任公司大理供电局，云南 大理 671000;
2. 云南电网有限责任公司，昆明 650051）

摘要 光伏发电设备故障及外界环境等多种因素导致分布式光伏在发电过程中产生大量异常数据。为了提高数据处理的准确性和效率，本文提出一种基于改进 K 均值聚类算法和加权动态时间规整（WDTW）的分布式光伏异常数据辨识方法。首先，对分布式光伏发电数据进行分析，利用同时段功率均值法对异常数据进行初步剔除。通过光照强度数据的归一化处理，提出基于改进 K 均值聚类算法的光伏数据相似日划分方法。其次，考虑光伏数据在时间维度的变化性和复杂性，引入异常数据识别最好的时段和阈值因子，提出基于 WDTW 的数据相似度分析方法。利用相似度计算轮廓系数，对光伏发电异常数据进行二次剔除。仿真结果表明，所提方法在辨识分布式光伏异常数据方面具有显著优势，相比于现有的四分位法、3-sigma 法和特征聚类法，所提方法的辨识精度分别提高 6.92%、9.00%和 8.12%，同时计算复杂度降低。

关键词：改进 K 均值聚类算法；加权动态时间规整（WDTW）；分布式光伏；异常数据辨识

A distributed photovoltaic abnormal data identification method based on improved K-means clustering algorithm and weighted dynamic time warping

YANG Wangxia[1] *LI Benyu*[2] *ZHAI Suwei*[2] *SHI Hengchu*[2] *LI Yinyin*[2]
(1. Dali Power Supply Bureau of Yunnan Power Grid Co., Ltd, Dali, Yunnan 671000;
2. Yunnan Power Grid Co., Ltd, Kunming 650051)

Abstract The failure of photovoltaic power generation equipment and various factors such as external environment lead to a large number of abnormal data during the power generation process. In order to improve the accuracy and efficiency of data processing, this paper proposes a distributed photovoltaic abnormal data identification method based on improved K-means algorithm and weighted dynamic time warping (WDTW). Firstly, the distributed photovoltaic power generation data is analyzed, and the abnormal data is preliminary eliminated by means of the simultaneous power mean method, and a photovoltaic data similarity day partitioning method based on improved K-means algorithm is proposed by normalizing the light intensity data. Secondly, considering the variability and complexity of photovoltaic data in the time dimension, a data similarity analysis method based on WDTW is proposed by introducing the best time period and threshold factor for identifying abnormal data. The similarity is used to calculate the contour coefficient, and the residual abnormal photovoltaic power generation data is culled twice. The simulation results show that the proposed method has significant advantages in identifying distributed photovoltaic abnormal data. Compared with the existing quartile method, 3-sigma method, and feature clustering method, the identification accuracy has been improved by 6.92%, 9.00%, and 8.12% respectively, while the computational complexity is reduced.

Keywords：improved K-means clustering algorithm; weighted dynamic time warping (WDTW); distributed photovoltaic; identification of abnormal data

云南电网公司科技项目（YNKJXM20230368）

0 引言

随着全球能源结构的转型和可再生能源的快速发展，太阳能作为一种重要的清洁能源，成为新能源领域的重要组成部分。分布式光伏系统因容量小、数量多、分布离散等特点，在能源供应和节能减排方面发挥了重要作用[1-3]。然而，这些特点也使分布式光伏系统的运维管理变得复杂，特别是光伏发电异常数据的辨识和处理成为一大难题。异常数据是指与大多数数据存在显著差异的数据，这些数据可能是因设备故障、测量误差或外部干扰等而产生。在分布式光伏系统中，异常数据的出现可能会影响系统的稳定性和安全性，甚至导致系统停机或损坏。因此，及时准确地辨识和处理异常数据对于保障系统的正常运行和延长系统使用寿命至关重要[4-8]。

无论是由设备故障、软件错误，还是由人为因素引起的分布式光伏数据异常，都会对分布式光伏系统产生一系列负面影响[9]。一方面，数据异常导致对光伏系统实际输出功率的误判，进而影响整个电力系统的调度和优化。不准确的数据误导运营者的决策，造成不必要的检查和维修，增加运维成本。同时，错误的数据影响光伏系统的经济效益分析，导致错误的投资决策，增加经济风险[10-13]。另一方面，数据异常掩盖系统中真正存在的问题，甚至导致设备损坏，对人员和财产构成严重威胁。分布式光伏系统往往配备环境监测设备，以实时监控温度、湿度、光照等环境参数[14-15]。数据异常会影响对环境条件的判断，进而影响光伏系统的效率和稳定性。频繁的数据异常会降低用户对分布式光伏系统的信任，尤其是依赖实时数据进行决策的商业用户，长期的数据不准确将严重影响系统的市场竞争力和用户满意度[16-18]。因此，必须加强数据监测和分析，及时发现数据异常并处理，确保光伏系统的稳定运行和电力市场的公平交易[19]。通过不断提高光伏系统的智能化和自动化水平，减少异常数据，提高电力系统的稳定性和可靠性。

分布式光伏系统的健康运行对确保能源供应的稳定性和效率至关重要。及时辨识和处理异常数据，可以预防潜在故障，维护系统的高效运行。近年来，学者们正在探索多种先进的算法和技术，以提高分布式光伏异常数据辨识的准确性和效率。文献[20]提出一种基于 Jaccard 和重叠指数的方法，用于定量评估将不同风险识别方法应用于光伏并网电力系统时可能存在的互补性和叠加性。文献[21]提出一种具有视觉几何组和挤压激励的增强型 U-net，以实现对光伏异常数据的精确识别。文献[22]提出一种基于深度残差神经网络的识别方法，解决了由光伏板不均匀灰尘引起的数据异常问题。文献[23]提出一种元启发式算法来精确计算太阳电池的温度，有效减少了光伏数据的异常率。文献[24]考虑屋顶光伏发电系统的安装区域，研究了光伏工作面板的安装区域和角度对发电数据的影响。文献[25]利用串箱级别的数据，对大规模分布式光伏发电系统中的数据缺失问题进行检测。文献[26]研究基于电流-电压转换的光伏阵列故障预诊断和类型识别方法，极大地减少了光伏异常数据的出现。文献[27]提出一种光伏分类和分段网络，提高了光伏数据分类精度，解决了新型光伏异常数据的识别问题。文献[28]提出一个光伏发电系统健康状态架构，用来预测由光伏阵列数据异常可能带来的故障。文献[29]研究基于自适应神经模糊系统的光伏系统故障实时检测、识别和消除方法，降低了光伏数据异常的可能性。需要说明的是，以上文献利用不同方法进行光伏异常数据的辨识，但未考虑光伏发电数据在时间尺度上的复杂性。

K 均值聚类算法作为一种常用的聚类分析方法，在分布式光伏异常数据辨识中得到了广泛应用[30-33]。通过设定聚类数目 K，K 均值聚类算法将样本集合划分为 K 个互不重叠的类别，使同类样本之间的相似性最大，不同类样本之间的相似性最小。在分布式光伏系统中，K 均值聚类算法可以对大量的光伏数据进行聚类分析，将正常数据和异常数据区分开来。然而，K 均值聚类算法也存在一些局限性，如聚类数目的选择、初始聚类中心的影响等，需要在实际应用中进行优化和改进。另外，在分布式光伏系统中，光照强度、温度等环境因素的变化会导致光伏功率数据波动，传统的光伏异常数据辨识方法往往难以准确处理这种波动，而动态时间规整（dynamic time warping, DTW）算法则能通过计算光伏功率曲线之间的相似度，有效辨识出异常数据[34-36]。然而，DTW 的时间复杂度较高，对存储需求大，在应用过程中需要进行改进。

综合以上分析，本文基于改进 K 均值聚类算法和加权动态时间规整（weighted dynamic time warping, WDTW）进行分布式光伏异常数据辨识，以提高分布式光伏系统数据处理的准确性和效率。通过对传

统 K 均值聚类算法进行改进，以更好地处理数据中的噪声和异常值，从而更准确地划分数据簇。引入 WDTW 算法对时间序列数据进行相似性度量，以有效捕捉数据在时间维度上的动态变化，进一步提升异常数据辨识的精度。

1　分布式光伏异常数据分析

分布式光伏异常数据的来源复杂多样，主要可归为设备故障、外部环境影响及数据采集与处理过程中的误差。一方面，光伏组件、逆变器、传感器等关键设备在运行过程中可能因老化、损坏或连接不良等问题出现故障。例如，光伏组件可能因电池损耗、封装胶开裂或玻璃破裂等导致输出功率下降或完全失效；逆变器可能因电气部件老化、散热不良或界面连接不良等引发故障，进而影响系统的正常输出。这些设备故障会直接反映在光伏出力数据上，产生异常值。含异常数据的相似光辐照度下光伏发电功率示意图如图 1 所示。

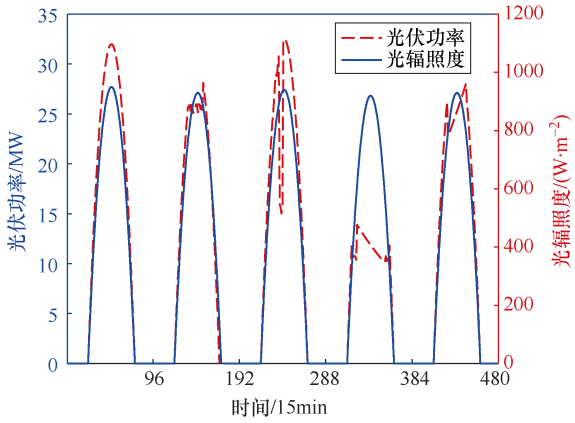

图 1　含异常数据的相似光辐照度下光伏发电功率示意图

另一方面，外部环境因素也是导致分布式光伏异常数据的重要原因。光伏系统依赖太阳光辐照进行光电转换，因此光照强度、温度、风向等环境因素的变化都会对光伏系统的发电效率产生影响。例如，在强风、暴雨等恶劣天气条件下，光伏板可能受到物理损害，导致发电量下降；同时，云层遮挡、灰尘积累等也会影响光伏板的光照接收，从而产生异常数据。

此外，数据采集与处理过程中的误差也是不可忽视的异常数据来源。在数据采集阶段，由于传感器精度不足、数据传输延迟或丢包等问题，可能导致数据记录不准确；在数据处理阶段，由于算法错误、参数设置不当或数据清洗不彻底等原因，也可能导致数据出现异常。

2　光伏异常数据辨识

光伏异常数据辨识是确保光伏系统高效运行和数据分析准确性的关键步骤。本文提出一种基于改进 K 均值聚类算法和 WDTW 的分布式光伏异常数据辨识方法。首先，利用同时段功率均值法对光伏发电异常数据进行初步剔除；其次，基于改进 K 均值聚类算法对光伏数据进行相似日划分；再次，基于 WDTW 方法对数据相似度进行分析；最后，计算轮廓系数，对异常数据进行二次剔除。基于改进 K 均值聚类算法和 WDTW 的分布式光伏异常数据辨识方法流程如图 2 所示。

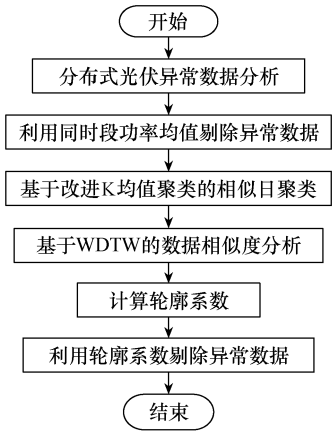

图 2　基于改进 K 均值聚类算法和 WDTW 的分布式光伏异常数据辨识方法流程

2.1　异常数据初步剔除

光伏发电产生大量数据，其中明显异常的数据可以通过对比的方法进行初步识别。本文利用同时段功率均值法对光伏发电异常数据进行初步剔除，该方法是一种数据预处理技术，用于在计算数据的统计指标之前剔除异常值。该方法首先计算同一时间段内光伏发电数据的平均值，再将实际测量值与该平均值进行比较，任何显著偏离这一平均值的数据点都可能被视为异常值，并在后续分析中被剔除。

以某地 n 天同一时段光伏数据为例，进行详细说明。假设所取时段为 t 时段，那么 n 天的 t 时段数据可表示为 $\boldsymbol{P}=[P_{t,1}\ P_{t,2}\cdots P_{t,n}]$。每个时段由 m 个数据组成，则 n 天 t 时段全部数据可表示为

$$\boldsymbol{P}_{\mathrm{pv},t}=\begin{bmatrix} P_{1,t,1} & P_{1,t,2} & \cdots & P_{1,t,n} \\ P_{2,t,1} & P_{2,t,2} & \cdots & P_{2,t,n} \\ \vdots & \vdots & & \vdots \\ P_{m,t,1} & P_{m,t,2} & \cdots & P_{m,t,n} \end{bmatrix} \tag{1}$$

式中：$\boldsymbol{P}_{\mathrm{pv},t}$ 为 n 天 t 时段光伏功率数据矩阵；$P_{m,t,n}$ 为第 n 天 t 时段第 m 个光伏功率数据。对式（1）中每列取平均值，得到 n 天 t 时段光伏数据平均值，即

$$\boldsymbol{P}_t^{\mathrm{mean}}=\begin{bmatrix}P_{t,1}^{\mathrm{mean}} & P_{t,2}^{\mathrm{mean}} \cdots P_{t,n}^{\mathrm{mean}}\end{bmatrix} \tag{2}$$

式中，$P_{t,n}^{\mathrm{mean}}$ 为第 n 天 t 时段光伏功率平均值。

利用同一时段光伏功率平均值对异常数据进行初步剔除。如果第 i 天 t 时段第 j 个数据与该时段数据平均值的差的绝对值小于允许阈值，则说明该数据为正常数据；反之则为异常数据，进行剔除。上述过程用公式表示为

$$\Psi=\begin{cases}\text{正常数据} & \left|P_{j,t,i}-P_{t,i}^{\mathrm{mean}}\right|<\delta_{\mathrm{pv}}\\ \text{剔除数据} & \left|P_{j,t,i}-P_{t,i}^{\mathrm{mean}}\right|\geqslant\delta_{\mathrm{pv}}\end{cases} \tag{3}$$

式中：Ψ 为光伏数据剔除结果；$P_{j,t,i}$ 为第 i 天 t 时段第 j 个光伏功率数据；$P_{t,i}^{\mathrm{mean}}$ 为第 i 天 t 时段光伏功率平均值；δ_{pv} 为允许阈值。

利用同时段功率均值法剔除异常数据，能有效去除那些明显偏离正常范围的数据点。使数据分布更趋于合理，提升统计分析的有效性。剔除这些异常数据后，所得到的数据更能反映光伏发电系统的常态运行情况。在数据可视化方面，异常数据可能成为图表中的离群点，干扰对整体趋势的观察。去除异常数据后，可视化图表能够更清晰地展示光伏发电系统的正常运行趋势，便于工作人员直观地了解系统的运行状态，及时发现潜在问题。

2.2 基于改进 K 均值聚类算法的相似日聚类

本文提出一种基于改进 K 均值聚类算法的光伏数据相似日划分方法。首先对光照数据进行归一化处理，然后根据不同的光照强度对光伏数据进行相似日分析，利用改进 K 均值聚类算法对光伏数据进行相似日聚类，确定在相似光照情况下进行异常数据辨识最好的时段，进一步选取最优的阈值因子。

1）归一化处理

光照数据作为光伏发电系统的主要输入变量之一，其波动性和不确定性对系统性能具有直接影响。在实际应用中，光照强度往往因时间、季节、天气条件及地理位置的不同而呈现出显著差异。这种差异不仅增加了数据分析的复杂性，还可能导致模型训练过程中的过拟合现象，降低模型的泛化能力。因此，需要对光照数据进行归一化处理，通过将其转换到一个相对稳定的数值范围，消除不同光照条件下的数据差异，提高数据的一致性和可比性。归一化处理过程可表示为

$$S_t^*=\frac{S_t-S^{\min}}{S^{\max}-S^{\min}} \tag{4}$$

式中：S_t 为 t 时段光照强度值；$S^{\min}$ 和 $S^{\max}$ 分别为光照强度最小值和最大值。

2）改进 K 均值聚类算法

K 均值聚类算法是一种基于距离度量的无监督学习算法，其核心思想是将数据集划分为 K 个不同的类别，使同一类别内的数据点之间距离最小，不同类别的数据点之间距离最大。该算法具有简单易用、计算效率高等优点，采用迭代优化的方法不断更新聚类中心点，直到满足停止条件。传统的 K 均值聚类算法易受噪声和异常值影响，并且对初始值敏感度高，因此需要对其进行改进。本文在传统 K 均值聚类算法的基础上，引入邻域半径和局部密度的思想。利用改进的 K 均值聚类算法对日照强度相似的光伏数据进行相似日划分，算法步骤如下。

步骤一：确定光伏数据聚类数量 K，随机选取初始聚类中心。利用手肘法确定聚类数量 K，可表示为

$$S_{\mathrm{SE}}=\sum_{i=1}^{K}\sum_{p\in C_i}\left|p-m_i\right|^2 \tag{5}$$

式中：S_{SE} 为手肘法的核心指标，即误差平方和；K 为聚类数量；C_i 为第 i 个簇；p 为 C_i 中的样本点；m_i 为 C_i 的质心。

步骤二：计算每个光伏数据的空间密度值。以 t 时段 n 天 m 个光伏发电数据为例，空间密度可表示为

$$\rho_{P_{j,t,i}}=\sum_{i=1}^{n}\frac{\left\|P_{t,i}^{\mathrm{mean}}-P_{j,t,i}\right\|_2}{\sum\limits_{k=1,k\neq j}^{m}\left\|P_{t,i}^{\mathrm{mean}}-P_{k,t,i}\right\|_2} \tag{6}$$

式中，$\|\cdot\|_2$ 为二范数。

步骤三：计算每个数据的邻域半径值。即

$$R_{P_{j,t,i}}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{-\rho_{P_{j,t,i}}} \tag{7}$$

步骤四：计算每个数据的局部密度值。表示为

$$D_{P_{j,t,i}}=\rho_{P_{j,t,i}}\lg R_{P_{j,t,i}} \tag{8}$$

步骤五：选取局部密度值最大的 K 个值作为聚

类中心，并将剩余数据按照距离大小划分到最近的聚类中心簇中。

步骤六：计算新的聚类中心。表示为

$$\Theta_k = \frac{1}{n_k}\sum_{s=1}^{n_k} P_{s,t,i} \tag{9}$$

式中，n_k 为第 k 个子类的数据量。

步骤七：重复步骤五和步骤六，直至聚类中心的值不再发生变化，得到最终结果。

空间密度描述样本空间中某一区域内样本点的分布密集程度。空间密度越大，说明该样本与其他样本之间的距离越小。邻域半径反映一个样本点邻域范围的最大距离。邻域半径越大，表明该样本离其他样本的平均值越远。局部密度反映样本点在特定半径内的邻居数量。由于同时考虑了空间密度和邻域半径，所以可以捕捉数据集中不同区域的密度差异，提高聚类的准确性和稳定性。同时，可以有效地检测出数据中的异常值，提高风险防范能力。

通过改进 K 均值聚类算法，得到同时段相似日下的数据分类情况。通过考虑每个数据点的邻域情况及局部密度，可以更准确地识别出真正具有相似性的日子。对于密度较高的区域，可以更细致地划分相似日群组，以反映其内部的多样性；对于密度较低的区域也能合理地将相对稀疏的数据点归为合适的类别。这种自适应性大大提高了算法的通用性和实用性，使其能够在各种不同类型的数据集中都取得良好的相似日划分效果。

3）选取最优阈值因子

光伏数据时段大小和阈值因子的选取会对异常数据的辨识效果产生影响。因此，需要筛选最优时段和最优阈值因子。本文利用控制变量法，选取光伏数据的最优时段和最优阈值因子：在确定阈值因子的情况下，通过对比不同时段下的光伏异常数据辨识效果，确定最优时段；在确定时段的情况下，通过对比不同阈值因子下的异常数据辨识效果，确定最优阈值因子。

在均值阈值因子一定的情况下，可通过式（2）所示光伏功率平均值实现异常数据均值阈值的划分，即

$$P_t = \lambda_t \max\left(P_{t,1}^{\text{mean}}, P_{t,2}^{\text{mean}}, \cdots, P_{t,n}^{\text{mean}}\right) \tag{10}$$

式中：P_t 为 t 时段均值最大值阈值；λ_t 为 t 时段均值阈值因子。通过对比不同时段异常数据的识别效果，确定相似光照情况下异常数据识别效果最好的时段。

在确定最优时段的情况下，选取 L 个阈值因子，组成集合 $\{\lambda_1, \cdots, \lambda_x, \cdots \lambda_L\}, x \in \{1, 2, \cdots, L\}$。设 η_{λ_x} 为在阈值因子 λ_x 下的异常数据剔除率，P_{t,λ_m}^{Ψ} 为剔除异常数据后的第 m 个光伏功率数据，$C_t P_{t,\lambda_m}^{\Psi}$ 为剔除异常数据后第 m 个数据光照强度与光伏功率的相关性。则最优阈值因子计算公式为

$$\lambda_t^* = \max\left(C_t P_{t,\lambda_1}^{\Psi}, C_t P_{t,\lambda_2}^{\Psi}, \cdots, C_t P_{t,\lambda_m}^{\Psi}\right) \quad \eta_{\lambda_x} < \eta_{\max} \tag{11}$$

式中，$\eta_{\max}$ 为异常数据剔除率上限。在满足异常数据剔除率上限的条件下，选取数据光照强度与光伏功率的相关性最大的值作为最优阈值因子。重复式（10）和式（11），当二者相互最优时，即可确定最优时段下的最优阈值因子。

在选取最优时段和最优阈值因子的过程中，需要注意两者之间的影响和制约关系。如果最优时段选取过大，或者最优阈值因子选取过小，会导致光伏异常数据辨识不准确，出现较大误差；如果最优时段选取过小，或者最优阈值因子选取过大，会增加计算量，降低辨识效率。仅通过单一变量无法保证辨识的合理性。非最优时段或非最优阈值因子条件下，会导致异常数据过多或过少，影响光伏数据的完整性和连续性。

2.3　基于 WDTW 的异常数据二次剔除

DTW 是一种用于计算两个时间序列之间相似度或距离的算法。基于动态规划的思想，解决时间序列因长度不一、比例不同或存在噪声而难以直接比较的问题。通过将两个时间序列按照某种最优路径进行对齐，找到一条“弯曲”的路径，使这条路径上的所有点之间的距离之和最小，从而衡量两个时间序列之间的相似度。DTW 算法可以处理不同步长的序列，具有较强的鲁棒性，但也存在不足之处，如计算复杂度高、对噪声敏感、缺乏全局结构信息等。本文提出加权动态时间规整方法，通过引入权值系数，优化计算效率，提高相似性计算准确性。

取光伏出力和光照强度两个时间序列，分别记为 $\boldsymbol{P} = \begin{bmatrix} P_1 & P_2 \cdots P_p \end{bmatrix}$ 和 $\boldsymbol{S} = \begin{bmatrix} S_1 & S_2 \cdots S_q \end{bmatrix}$。构建 p 行 q 列的距离矩阵 $\boldsymbol{D}$ 为

$$\boldsymbol{D} = \begin{bmatrix} d_{P_1,S_1} & d_{P_1,S_2} & \cdots & d_{P_1,S_q} \\ d_{P_2,S_1} & d_{P_2,S_2} & \cdots & d_{P_2,S_q} \\ \vdots & \vdots & & \vdots \\ d_{P_p,S_1} & d_{P_p,S_2} & \cdots & d_{P_p,S_q} \end{bmatrix} \tag{12}$$

式中，d_{P_i,S_j} 为光伏出力 P_i 与光照强度 S_j 的对应关系，$i\in\{1,2,\cdots,p\}$，$j\in\{1,2,\cdots,q\}$，用欧氏距离表示为

$$d_{P_i,S_j}=d(i,j)=\left\|P_i-S_j\right\|_2 \tag{13}$$

引入权值系数后，对应关系可表示为

$$d^{\mathrm{f}}_{P_i,S_j}=f_{ij}d(i,j)=\frac{2}{1-\mathrm{e}^{-(|\mu_i\mu_j|-1)}}d(i,j) \tag{14}$$

式中：f_{ij} 为光伏出力 P_i 与光照强度 S_j 之间的权值系数；μ_i 和 μ_j 分别为能够表征 $\boldsymbol{P}$ 和 $\boldsymbol{S}$ 序列点特征的量。当 $\boldsymbol{P}$ 或 $\boldsymbol{S}$ 中序列点为极大值时，μ_i 和 μ_j 的值为 1；当 $\boldsymbol{P}$ 或 $\boldsymbol{S}$ 中序列点为极小值时，μ_i 和 μ_j 的值为−1；否则为 0。

通过引入权值系数，WDTW 的递推公式表示为

$$\begin{aligned}&D_{\mathrm{TW}}(i,j)=f_{ij}d(i,j)+\\&\min\left\{D_{\mathrm{TW}}(i-1,j),D_{\mathrm{TW}}(i,j-1),D_{\mathrm{TW}}(i-1,j-1)\right\}\end{aligned} \tag{15}$$

式中，$D_{\mathrm{TW}}(i,j)$ 为第 i 个数据点和第 j 个数据点之间的 WDTW 距离。

利用欧氏距离和 WDTW 距离构建表征 $\boldsymbol{P}$ 和 $\boldsymbol{S}$ 序列的相似度函数，即

$$\begin{aligned}C(i,j)=&\alpha_1d(i,j)+\alpha_2D_{\mathrm{TW}}(i,j)+\alpha_3d\left(i^{\Delta},j^{\Delta}\right)+\\&\alpha_4D_{\mathrm{TW}}\left(i^{\Delta},j^{\Delta}\right)\end{aligned} \tag{16}$$

式中：α_1、α_2、α_3 和 α_4 为相似度权重系数；i^{Δ} 和 j^{Δ} 为光伏出力变化量和光照强度变化量，i^{Δ} 和 j^{Δ} 可分别表示为 $i^{\Delta}=\left(P_{i+1}-P_i\right)/\Delta t$，$j^{\Delta}=\left(S_{j+1}-S_j\right)/\Delta t$，其中 Δt 为采样间隔；$D_{\mathrm{TW}}\left(i^{\Delta},j^{\Delta}\right)$ 为 $\boldsymbol{P}$ 和 $\boldsymbol{S}$ 局部动态对应的 WDTW 距离，反映曲线波动性带来的影响。

通过引入权值系数，WDTW 方法能够为包含重要信息的元素分配更高的权值系数，更准确地反映这些部分在相似度计算中的重要性，从而得到更贴近实际的相似度结果。此外，WDTW 方法可以在一定程度上降低不利因素对相似度计算的影响，从而提高算法的鲁棒性。

通过相似度函数的计算，得到 $\boldsymbol{P}$ 和 $\boldsymbol{S}$ 序列的相似度。将相似度高的数据点归为一类，同时计算每一个数据点的轮廓系数，即

$$S_{\mathrm{C}}(i)=\begin{cases}\dfrac{\beta(i)-\alpha(i)}{\alpha(i)} & \alpha(i)>\beta(i)\\ 0 & \alpha(i)=\beta(i)\\ \dfrac{\beta(i)-\alpha(i)}{\beta(i)} & \alpha(i)<\beta(i)\end{cases} \tag{17}$$

式中，$S_{\mathrm{C}}(i)$为第 i 个数据点的轮廓系数；$\alpha(i)$为第 i 个数据点到所属簇中其他数据点的平均距离；$\beta(i)$为第 i 个数据点到最近簇中数据点的平均距离。

轮廓系数是衡量聚类有效性的重要指标，它结合了簇内相似度和簇间分离度两种因子，用于评估样本点在其所属簇内的紧密程度，以及与其他簇的分离程度。利用轮廓系数对异常数据进行剔除的机制如图 3 所示。由于轮廓系数的值在[−1, 1]，值越接近 1，表明数据点与其所在簇内的其他点越相似，同时与其他簇的点越远，即聚类效果越好，属于良好数据区；值越接近−1，表明数据点可能更适合被划分到其他簇中，或者是一个明显的异常值，属于异常数据区。

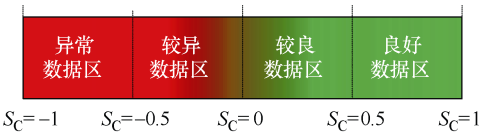

图 3 轮廓系数剔除机制

需要说明的是，利用轮廓系数剔除光伏异常数据是利用同时段功率均值法对异常数据进行初步剔除的延续，两者具有一定互补性。同时段功率均值法能够快速去除一些明显的异常值，为进一步筛选奠定基础；轮廓系数法在此基础上进行更为深入的评估和分析，能够识别并剔除一些更为复杂的异常数据。通过两种方法的结合，可以提高异常数据剔除的准确性，增强数据处理的鲁棒性，提高数据质量。

3 仿真分析

本文以某地一年（365 天）光伏数据为研究对象，该地光照充足，雨水较少，光伏发电功率较高。以 15min 为间隔，得到一年光伏数据总量为 35 040，取同时段初步剔除率为 5%，可得到有效数据总量为 33 288。以夏季数据为例，在仿真分析中，允许阈值 δ_{pv} 取 0.1，将各时段包含数量分别取为 5、10、15、20，对应时段为 1、2、3、4，阈值因子选取为 0.3、0.4、0.5、0.6，相似度权重系数之和为 1。异常数据剔除率上限选取为 0.5。仿真相关参数见表 1。

表 1　仿真相关参数

参数	取值
δ_{pv}	0.1
t	1, 2, 3, 4
λ_t	0.3, 0.4, 0.5, 0.6
η_{max}	0.5
$[\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4]$	[0.4　0.4　0.1　0.1]

首先，验证本文所提改进 K 均值聚类算法的性能。利用本文方法与传统 K 均值聚类算法和文献[30]所提改进 K 均值聚类算法进行对比，得到误差平方和如图 4 所示。

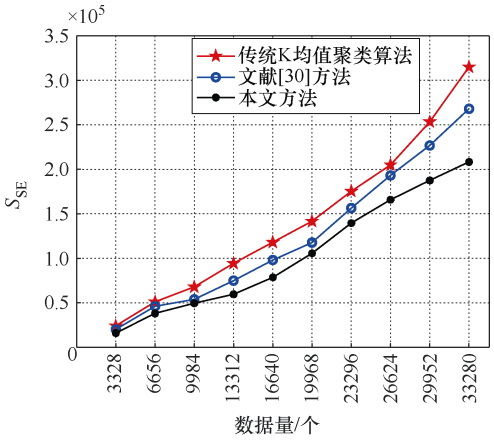

图 4　不同算法误差平方和

通过对比可知，在处理相同数据时，传统 K 均值聚类算法的误差平方和最大，说明对数据的划分存在一定不确定性，即存在数据误分的可能；本文所提方法的误差平方和最小，说明通过考虑每个数据点的邻域情况及局部密度，可以更准确地识别出真正具有相似性的数据特征。

其次，在均值阈值因子一定的情况下，分析时段的影响。选取阈值因子为 0.4，对光伏数据进行归一化处理。利用改进 K 均值聚类算法对数据进行相似日划分，得到聚类数量为 6 时不同时段下光伏异常数据辨识效果如图 5 所示。

轮廓系数表征剔除异常数据后与正常光伏数据的相关性。由图 5 可以看出，原始数据的剔除率为 0，但由于包含异常数据，轮廓系数较低。不管时段如何设置，剔除异常数据后的数据相关性都会提高。随着时段的增加，剔除率有明显的上升，轮廓系数下降，原因在于时段增加导致光伏数据增加，利用同时段功率均值法对光伏发电异常数据进行初步剔除的比例也会增加。剔除率的上升，使部分正常数据被误认为异常数据而剔除，轮廓系数下降。当时段为 1 时，轮廓系数为 0.905，达到最大，此时剔除率在四个时段中最低，说明阈值因子为 0.4、时段取 1 时，光伏异常数据辨识效果最好。

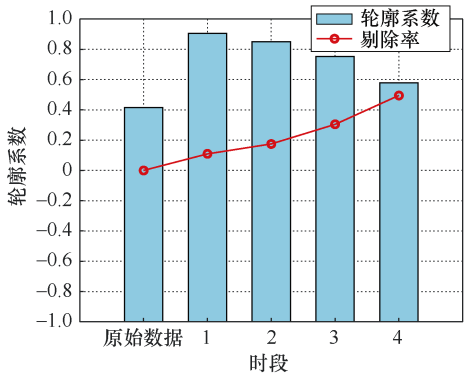

图 5　不同时段下光伏异常数据辨识效果

根据以上分析，选取最优时段为 1，分析阈值因子取 0.3、0.4、0.5 和 0.6 情况下的光伏异常数据辨识情况。图 6 为不同阈值因子下光伏异常数据辨识效果。可以看出，随着阈值因子增加到 0.4，剔除率增加，轮廓系数上升。在阈值因子由 0.4 增加到 0.6 的过程中，剔除率不断增加，但轮廓系数降低。原因在于，阈值因子选取过大，导致数据中异常数据的占比较高，当阈值因子大于 0.4 时，随着剔除率的增加，部分正常数据被识别成异常数据，导致轮廓系数下降。本文在时段为 1 时，选取最优阈值因子为 0.4。

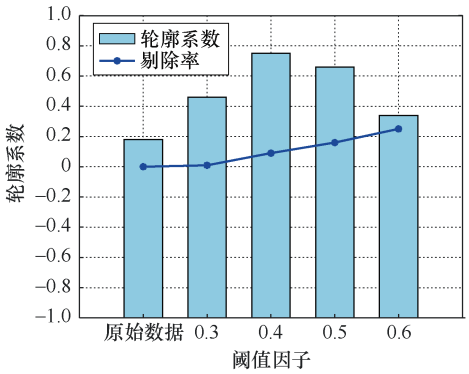

图 6　不同阈值因子下光伏异常数据辨识效果

为验证本文方法的优越性，将本文方法与四分位法、3-sigma 法和特征聚类法进行比较，得到如图 7 所示不同方法相关程度对比。

由图 7 可以看出，本文所提方法对剔除异常数据后的相关程度要明显高于上述 3 种方法。不同方法相关程度见表 2。以平均值计算，与上述 3 种方法相比，本文所提方法相关程度分别提升了 6.92%，9.00%和 8.12%。

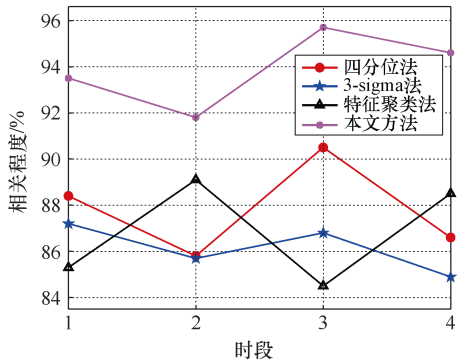

图 7 不同方法相关程度对比

表 2 不同方法相关程度

方法	相关程度/%				
	时段 1	时段 2	时段 3	时段 4	平均值
四分位法	88.4	85.8	90.5	86.6	87.825
3-sigma 法	87.2	85.7	86.8	84.9	86.150
特征聚类法	85.3	89.1	84.5	88.5	86.850
本文所提方法	93.5	91.8	95.7	94.6	93.900

光伏异常数据辨识时序如图 8 所示。与图 1 相比，图 8 剔除了正常数据中的异常数据。在相似日光辐照度情况下，光伏功率包含异常数据使部分数据分布不合理。利用本文所提方法对光伏数据进行辨识后，得到异常数据剔除后的拟合值与实际光伏发电数据更贴合。

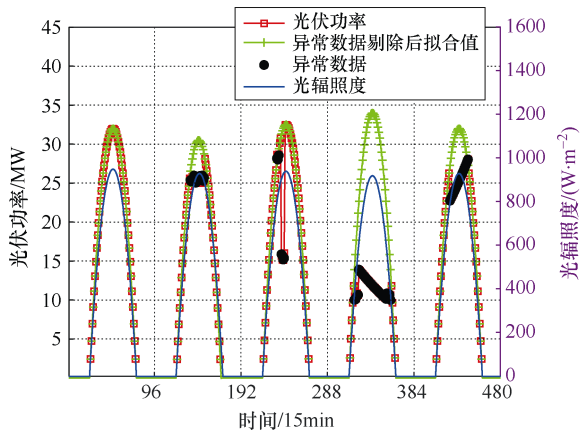

图 8 光伏异常数据辨识时序

4 结论

本文研究了基于改进 K 均值聚类算法和 WDTW 的分布式光伏异常数据辨识方法，旨在提高异常数据辨识的准确性和效率。本文提出了基于改进 K 均值聚类的光伏数据相似日划分方法，解决了初始聚类中心选择敏感、聚类结果不稳定等问题，从而提高了算法的鲁棒性和聚类效果。同时，引入 WDTW 算法对时间序列光伏数据进行相似性度量，有效捕捉了数据在时间维度上的动态变化，进一步提升了异常数据辨识的精度。仿真结果表明，本文所提方法在分布式光伏异常数据辨识中具有显著优势，相比于现有的四分位法、3-sigma 法和特征聚类法，所提方法的辨识精度分别提高了 6.92%、9.00%和 8.12%。

参考文献

[1] 廖家齐, 于若英, 刘瑜俊, 等. 基于自适应高斯混合模型的含高渗透率分布式光伏电力系统风险评估[J]. 电力系统保护与控制, 2024, 52(19): 144-156.

[2] 陈艳波, 刘宇翔, 田昊欣, 等. 基于广义目标级联法的多牵引变电站光伏-储能协同规划配置[J]. 电工技术学报, 2024, 39(15): 4599-4612.

[3] 武昭原, 刘婧宇, 周明, 等. 分散决策下分布式光伏储能系统外部性价值量化评估[J]. 电力系统自动化, 2024, 48(5): 38-47.

[4] 李斌, 罗晓伊. 分布式电源对电力系统电压无功优化影响的研究[J]. 电气技术, 2024, 25(10): 55-61, 78.

[5] MA Wenting, MA Mingyao, ZHANG Zhixiang, et al. Anomaly detection of mountain photovoltaic power plant based on spectral clustering[J]. IEEE Journal of Photovoltaics, 2023, 13(4): 621-631.

[6] 张铄, 吴丽珍. 计及坏数据辨识与修正的配电网状态估计[J]. 电气技术, 2022, 23(11): 1-6, 12.

[7] SU Binyi, ZHOU Zhong, CHEN Haiyong. PVEL-AD: a large-scale open-world dataset for photovoltaic cell anomaly detection[J]. IEEE Transactions on Industrial Informatics, 2023, 19(1): 404-413.

[8] 李阳, 沈小军, 张扬帆, 等. 基于速度-关联约束的风电机组风速感知异常数据识别方法[J]. 电工技术学报, 2023, 38(7): 1793-1807.

[9] 叶林, 崔宝丹, 李卓, 等. 光伏电站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.

[10] BENALCAZAR P, KOMOROWSKA A, KAMINSKI J. A GIS-based method for assessing the economics of utility-scale photovoltaic systems[J]. Applied Energy, 2024, 353: 122044.

[11] 顾菊平, 赵佳皓, 张新松, 等. 电力设备多参量监测数据清洗研究现状及展望[J]. 高电压技术, 2024, 50(8): 3403-3420.

[12] GAO Jianwei, WANG Yaping, GUO Fengjia, et al. A two-stage decision framework for GIS-based site selection of wind-photovoltaic-hybrid energy storage project using LSGDM method[J]. Renewable Energy, 2024, 222: 119912.

[13] 王丽朝, 孟子尧, 陈诗明, 等. 基于 GRU 神经网络的光伏电站数据预处理方法[J]. 太阳能学报, 2022, 43(11): 78-84.

[14] WANG Rui, LI Peng, YU Hao, et al. Identification of critical uncertain factors of distribution networks with high penetration of photovoltaics and electric vehicles[J]. Applied Energy, 2023, 329: 120260.

[15] JAVAID A, SHAFI I, KHALIL I U, et al. Enhancing photovoltaic systems using Gaussian process regression for parameter identification and fault detection[J]. Energy Reports, 2024, 11: 4485-4499.

[16] 郝颖, 冬雷, 王丽婕, 等. 基于数学形态学去噪的光伏发电限电异常数据识别算法[J]. 中国电机工程学报, 2022, 42(21): 7843-7854.

[17] 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.

[18] MA Mingyao, ZHANG Zhixiang, YUN Ping, et al. Photovoltaic module current mismatch fault diagnosis based on I-V data[J]. IEEE Journal of Photovoltaics, 2021, 11(3): 779-788.

[19] BADR M M, ABDEL-KHALIK A S, HAMAD M S, et al. Intelligent fault identification strategy of photovoltaic array based on ensemble self-training learning[J]. Solar Energy, 2023, 249: 122-138.

[20] SANTOS M F O, DE SOUZA MELO W JR, OLIVEIRA DE SA A, et al. A hybrid cyber-physical risk identification method for grid-connected photovoltaic systems[J]. Sustainable Energy, Grids and Networks, 2024, 39: 101490.

[21] LIU Bo, CHEN Lei, SUN Kai, et al. A hot spot identification approach for photovoltaic module based on enhanced U-net with squeeze-and-excitation and VGG19[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 3516510.

[22] FAN Siyuan, WANG Yu, CAO Shengxian, et al. A deep residual neural network identification method for uneven dust accumulation on photovoltaic (PV) panels[J]. Energy, 2022, 239: 122302.

[23] OSORIO L, MORENO M, RIVERA M, et al. A metaheuristic-based method for photovoltaic temperature computation under tropical conditions[J]. Solar Energy, 2024, 271: 112414.

[24] ASLANI M, SEIPEL S. Automatic identification of utilizable rooftop areas in digital surface models for photovoltaics potential assessment[J]. Applied Energy, 2022, 306: 118033.

[25] SIMAO T E, VALLE B C, ROSA Y C, et al. A method for the estimation of missing strings in very-large-scale photovoltaic power plants[J]. IEEE Journal of Photovoltaics, 2024, 14(5): 839-847.

[26] CHEN Xiang, JIANG Meng, DING Kun, et al. Fault prediagnosis, type identification, and degree diagnosis method of the photovoltaic array based on the current-voltage conversion[J]. IEEE Transactions on Power Electronics, 2024, 39(12): 16708-16719.

[27] CHEN Di, PENG Qiuzhi, LU Jiating, et al. Classification and segmentation of five photovoltaic types based on instance segmentation for generating more refined photovoltaic data[J]. Applied Energy, 2024, 376: 124296.

[28] MONTES-ROMERO J, HEINZLE N, LIVERA A, et al. Novel data-driven health-state architecture for photovoltaic system failure diagnosis[J]. Solar Energy, 2024, 279: 112820.

[29] MANSOURI M M, HADJERI S, BRAHAMI M. New method of detection, identification, and elimination of photovoltaic system faults in real time based on the adaptive neuro-fuzzy system[J]. IEEE Journal of Photovoltaics, 2021, 11(3): 797-805.

[30] 彭勃, 李耀东, 龚贤夫. 基于自编码的改进 K-means 光伏能源数据清洗方法[J]. 计算机科学, 2024, 51(增刊 1): 725-729.

[31] 余洋, 陆文韬, 陈东阳, 等. 光伏波动平抑下改进 K-means 的电池储能动态分组控制策略[J]. 电力系统保护与控制, 2024, 52(7): 1-11.

[32] 黄劼, 汪逸帆, 林叶青, 等. 基于 K 均值聚类算法的谐振接地系统故障区段定位方法[J]. 电气技术, 2024, 25(3): 24-31, 37.

[33] 郜彬, 黄杨珏, 沈开程, 等. 基于指标加权 K-means++ 算法的分布式光伏功率波动平抑控制方法[J]. 武汉大学学报(工学版), 2023, 56(11): 1413-1424.

[34] 邓祥力, 廖玥琳, 朱宏业, 等. 基于数字孪生模型电流动态时间规整差异度的变压器早期故障辨识[J].

（下转第 57 页）

[2] 周广猛, 刘伍权, 董素荣, 等. “以学生为中心”的工程类课程在线教学探索: 以“内燃机构造”课程为例[J]. 高等教育研究学报, 2021(44): 72-76.
[3] 刘战合, 张伟伟, 罗明强. 面向综合能力提升的连通式实践教学体系构建[J]. 高等教育研究学报, 2021(44): 113-120.
[4] 刘亚静, 段超. 全数字自适应滤波器不同离散结构的性能对比分析[J]. 电工技术学报, 2021, 36(20): 4339-4349.
[5] 谢佳, 段斌, 高婷, 等. 神经网络认知测量在工程教学课程评价中的应用[J]. 电气技术, 2023, 24(2): 52-58.
[6] 朱娟娟, 贺王鹏, 郭宝龙. 新工科电气电子类课程改革与实践: 以“信号与系统”为例[J]. 电气技术, 2024, 25(10): 72-78.
[7] 杨勇, 李红斌, 文劲宇, 等. 新工科电气工程实践教学体系重构与实践[J]. 电工技术学报, 2022, 37(19): 5074-5080.
[8] 赵玲峰. 雷达信号的数字化中频调制解调算法仿真[J]. 计算机仿真, 2019, 36(6): 16-20.
[9] 冯鸿运, 林飞, 杨中平, 等. 应用于自动导引小车无线充电系统的导航与供电一体化线圈研究[J]. 电工技术学报, 2024, 39(14): 4294-4304.
[10] 胡继云, 赫刘勤, 郑维. 基于 FPGA 和线阵 CCD 的麦粒检测系统研究[J]. 电气技术, 2017, 18(4): 84-89.
[11] 邹应全, 吴太龙, 彭榆淞, 等. 基于线下估计和线上补偿的时间交错采样 ADC 失配误差补偿方法[J]. 电子与信息学报, 2019, 41(1): 226-232.
[12] 陈顺阳. 高速 ADC 动态性能的有效评估[C]//中国电子学会电子对抗分会第十三届学术年会论文集, 北京, 2003: 674-681.
[13] 赵智兵, 袁雯, 郭倩. 数字接收机中的 ADC 性能分析[J]. 电子技术, 2016, 45(10): 25-26, 21.
[14] 申富媛, 李炜, 刘微容, 等. “自动控制原理”课程目标达成度统计分析与持续改进[J]. 电气技术, 2023, 24(1): 65-69, 75.
[15] 何杰. “半导体集成化芯片系统基础研究”重大研究计划进展综述[J]. 中国科学基金, 2005(6): 343-346.

收稿日期：2024-12-25
修回日期：2025-01-13

作者简介

李清江（1986—），男，山东省枣庄市人，博士，教授，主要从事智能信息器件与电路方面的研究工作。

（上接第 47 页）

电力系统保护与控制, 2023, 51(12): 156-167.
[35] 申江卫, 岩川, 刘永刚, 等. 基于数据挖掘与大数据分析的电池故障诊断与异常检测[J]. 电工技术学报, 2024, 39(24): 7979-7994.
[36] 阳瑞霖, 莫凡, 金艳, 等. 基于重心平均动态时间规整算法的有载分接开关机械故障诊断[J]. 高电压技术, 2023, 49(4): 1515-1525.

收稿日期：2024-11-11
修回日期：2025-01-15

作者简介

杨旺霞（1981—），女，云南省大理市人，硕士，高级工程师，主要从事电力系统继电保护研究工作。

（上接第 51 页）

2004, 24(6): 54-58.
[10] 喻锟, 兰宇婷, 曾祥君, 等. 基于零序电压区域放大的配电网高阻接地故障选线新方法[J/OL]. 中国电机工程学报, 1-14[2025-02-24]. http://kns.cnki.net/kcms/detail/11.2107.tm.20240528.1727.015.html.
[11] 张德丰. MATLAB 小波分析[M]. 北京: 机械工业出版社, 2009.

收稿日期：2024-12-23
修回日期：2025-01-24

作者简介

王建南（1987—），男，浙江宁波人，硕士，工程师，主要从事电力系统配电网研究工作。