

多维气象特征融合的光伏功率异常检测与修复深度学习方法

李 云^{1,2,*}, 程星月^{1,2}

1. 华北电力大学 控制与计算机工程学院, 保定 071003;
2. 河北省发电过程仿真与优化控制技术创新中心, 保定 071003



摘 要 : 光伏功率预测精度对电网稳定至关重要, 但历史数据中的异常样本会干扰预测模型训练, 降低预测精度。为此, 本文提出一种融合多气象特征的异常检测与修复方法, 首先基于相关性分析和随机森林筛选关键气象因素, 构建表征正常功率模式的参考超图像; 然后将参考超图像的体素特征作为 Reed-Xiaoli 算法的输入, 并引入高斯平滑背景, 通过计算体素与背景均值的欧氏距离来识别异常体素, 进而定位历史异常数据; 再通过迭代优化超图像, 并基于松阈值检测测试样本异常; 最后利用超图像修复异常及缺失数据。仿真实验表明, 本文所提方法的检测准确率优于传统方法, 且平均绝对误差较传统修复方法减小了 12.3%, 验证了该方法的工程实用性。

关 键 词 : 光伏功率预测; 异常; 缺失值; 修复; 超图像

中图分类号 : TP183; TM615 **文献标志码:** A **文章编号:** 2097-4981(2025)05-0736-12

DOI: 10.3724/j.issn.2097-4981.JECC-2025-0053

A Prediction-Oriented Method for Anomaly Detection and Repair of Photovoltaic Power Integrating Multi-Meteorological Features

LI Yun^{1,2,*}, CHENG Xingyue^{1,2}

1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China;
2. Hebei Innovation Center for Simulation & Optimized Control of Power Generation Processes, Baoding 071003, China

Abstract : Photovoltaic power prediction accuracy is crucial for grid stability, but outliers in historical data can interfere with the training of prediction model and reduce prediction accuracy. To this end, this paper proposes an anomaly detection and repair method that integrates multiple meteorological features, firstly, based on correlation analysis and random forest to screen key meteorological factors, and constructs a reference hyperimage to characterize the normal power mode. Then, the improved Reed-Xiaoli algorithm was used to detect and eliminate the historical anomalous data, the hyperimage was iteratively optimized, and the test sample anomaly was detected based on the loose threshold. Finally, the hyperimage was used to repair the abnormal and missing data. Simulation experiments show that the detection accuracy of the proposed method is better than that of the traditional method, and the mean absolute error is reduced by 12.3% compared with the traditional repair method, which verifies the engineering practicability of this method.

Key Words : photovoltaic power prediction; anomaly; missing values; repair; hyperimage

收稿日期: 2025-09-15; **接受日期:** 2025-10-31

基金项目: 河北省省级科技计划项目 (22567643H)

通讯作者: 李 云 (2002—), 女, 硕士研究生, 研究方向为光伏功率预测。E-mail: l_yun22419@163.com

0 引言

随着全球能源结构转型的持续发展,光伏发电作为清洁能源的重要组成部分,在电网中的渗透率日益提高^[1]。然而,受辐照度突变、云层移动等气象因素的影响,光伏发电具有不确定性和间歇性,加重了光伏系统并网的频率调节压力,对电网的安全稳定运行和规划调度构成了挑战^[2]。准确的功率预测能够预知光伏出力的变化,对于优化电网调度、避免功率突变对电网设备造成损害和提高能源利用效率具有重要意义。

在实际应用中,光伏功率的预测精度很大程度上依赖于训练模型的数据质量。由于气象因素本身的随机波动性叠加数据传输、传感器故障的影响,采集的历史数据难免会出现缺失、超过正常范围、与真实值偏差较大等异常情况^[3],其对应的异常模式会扭曲功率与气象因素的关联特征。模型可能将异常模式“正常化”,影响预测模型捕捉序列中的复杂非线性关系,轻则减慢模型训练的收敛速度,重则可能导致模型训练发散^[4]。因此,减小异常数据对预测模型的干扰是进行光伏功率预测的必要手段^[5]。

传统的无监督异常检测方法主要包括基于聚类的方法、基于距离的方法和基于统计分析的方法。基于聚类的方法^[6]通过样本相似性来划分数据簇,孤立点或边缘样本即为异常,但难以区分真实异常与辐照度突变导致的合理波动;基于距离的方法^[7]根据邻居样本的密度计算异常得分,对光伏数据中出现的局部突变易产生误判^[5];基于统计分析的方法假设数据服从特定分布^[8],经典的 3σ 阈值法虽然简便快速,但研究表明仅当处于晴天时出力才服从正态分布^[9]。随着计算机技术的发展,神经网络也在异常检测中得到了一定应用。文献^[10]以时间子序列段为样本单位并依据历史信息来构建邻域特征,均衡模型对序列局部细节与序列间关系的敏感性,但计算量较大且对全局异常不敏感。基于预测或重构的深度学习方法虽能通过神经网络隐式学习时空依赖性,但仍受训练数据中异常数据的影响^[11-12]。

在实际光伏功率数据中,异常数据与正常数据的界限难以明确区分^[13],呈现出一种交错混杂的复杂分布状态。基于统计学原理或单一特征进行异常

识别的检测效果往往不尽如人意。此外,简单地剔除异常数据可能会造成信息丢失,对于后续预测模型的学习极其不利^[14]。文献^[15]采用局部离群因子(Local Outlier Factor, LOF)算法进行离群值检测并剔除异常点,然后使用多层感知机回归算法进行修补后再进行预测,但是LOF算法对于参数的依赖性限制了其识别能力。文献^[16]通过引入松弛变量,计算多维影响因素样本到球形中心的距离来检测异常,并在预测模型中根据异常得分进行加权处理,虽提升了预测精度,但增加了模型复杂度。文献^[3]通过功率和辐照度的方差变化点和相关性分析来识别异常值,并基于线性回归模型进行修复,对由于光伏组件性能衰减、老化等非时序因素引发的异常识别能力有限。文献^[17]根据气象因素的皮尔逊相关系数确定相似时刻,再由插值法估计出缺失数据,能够适应数据短时间尺度内的波动情况,但在高比例异常情况下插值误差也随之增大。

针对现有检测方法多依赖时序统计特征,难以兼顾时序关联性与空间特征统一性,难以为数据修复提供可靠依据的问题,本文通过构建超图像将多气象因素与光伏功率的时序关系映射为体素空间的特征关系,实现时空-特征的统一表征,并提出了一种面向预测任务的光伏功率异常检测和修复方法。本文首先通过斯皮尔曼相关性分析和随机森林算法筛选出光伏功率预测的关键气象因素,并使用这些气象因素与光伏功率数据构建参考超图像;然后,基于超图像进行异常光伏功率检测,再利用超图像对异常及缺失数据进行修复,进一步将修复后的历史数据输入预测模型进行学习,从而提高待预测时段内的功率预测精度。

1 影响因素筛选

光伏功率预测精度高度依赖于环境因素的选择。若输入变量过少,模型将难以充分表征功率变化规律;若输入变量过多,则会引入噪声并降低计算效率。因此,有必要对输入模型的特征变量进行筛选,以均衡预测性能与模型复杂度。本文采用斯皮尔曼系数评估各因素与功率的关联强度,进一步结合随机森林的特征重要性分析,量化不同变量对预测的

贡献程度，从而筛选出最具解释力的特征子集。

对于两个随机变量 X 、 Y 经排序得到各自的秩次 R_X 、 R_Y ，计算斯皮尔曼相关系数。

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n (R_{X_i} - R_{Y_i})^2}{n(n^2 - 1)} \quad (1)$$

式中， n 为样本数量。 ρ_s 越接近于1，表明两个变量的变化趋势越一致。

然而，在预测任务中，特征与目标量之间的关系可能随时间变化而改变。斯皮尔曼相关系数聚焦于变量间等级顺序的关联强度，难以衡量特征与目标量在时间尺度上的动态关联性。随机森林模型作为一种强大的机器学习工具，为处理时间序列数据中的复杂关系提供了新的思路。该模型以决策树为基本单元，通过集成多棵决策树进行学习并集成预测结果。随机森林的生成过程如图1所示。

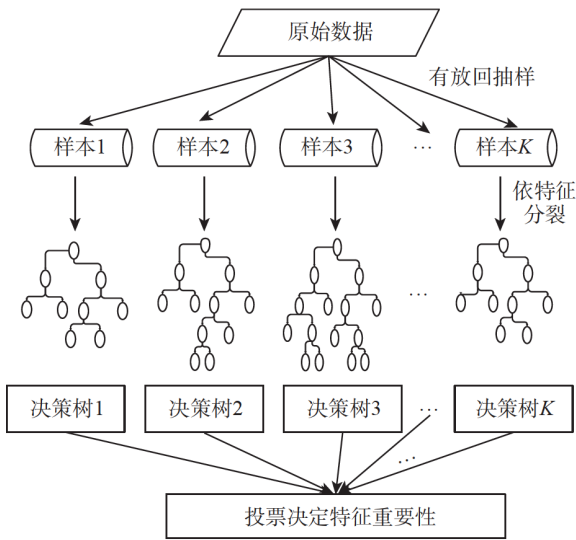


图1 随机森林模型
Fig.1 Random forest model

随机森林模型首先在原始数据集中多次有放回地重复抽样，产生随机样本集；然后根据每个样本构建决策树，通过随机选择一部分特征并选择节点分裂处的最佳分裂值来运行决策树；最后通过各个决策树的分类结果进行多数投票并集成，得出最终结果。随机森林中的特征重要性是衡量各个特征对模型预测能力影响的指标，决策树中的每个节点均可定义基尼指数的运算，特征在某一节点完成分裂

后产生的节点基尼不纯度增益的大小代表该特征分裂该节点的能力强弱，某一特征的重要性即为其在所有分裂节点上的平均基尼不纯度增益。在时间序列分析中，随机森林能捕捉特征与目标量间的动态关系，其重要性评分可量化各时间步特征对预测的贡献，弥补传统相关系数（如斯皮尔曼）在时序关联性度量上的不足。相关系数反映变量间的单调一致性，而特征重要性能够评估其对预测的影响程度，二者结合可更精准地筛选光伏功率预测的关键因素，降低无关特征的计算冗余。

2 超图像构建

预测模型的本质在于学习历史的多元因素序列与未来的目标序列间的映射关系，通过准确建立映射模型以检测异常，需耗费大量计算资源，且进行建模本身要求可见的数据能够有效反映正常模式。为避免这种迭代建模误差，本文提出在打破时间戳限制的情况下，基于插值实现对气象因素与光伏功率数据间稳定数量关系的显式建模。

假设原始数据中包含 m 个气象因素 $\mathbf{X} = \{X_k\}_{k=1}^m$ ，其中 $X_k = \{x_t^k\}_{t=1}^T$ 表示单个影响因素的时间序列；光伏功率量 $\mathbf{Y} = \{y_t\}_{t=1}^T$ ，其中 T 为时间戳。输入预测模型中的样本 $\mathbf{p}_i = [\mathbf{R}_i, \mathbf{y}_i]$ ，其中 $\mathbf{R}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^m\}$ 为过去 h 时间步的历史因素片段，且单个因素的历史片段 $\mathbf{x}_i^k = [x_{t-h}^k, x_{t-h+1}^k, \dots, x_{t-1}^k]$ ； $\mathbf{y}_i = (y_i, y_{i+1}, \dots, y_{i+l-1})$ 为未来 l 步的光伏功率向量。基于时序预测中未来目标量与历史影响因素存在规律性关联的条件假设，将因素片段映射为体素坐标，目标量映射为体素值，从而构建超图像 $MH: \mathbb{R}^M \mapsto \mathbb{R}_0^M$ 。

$$\begin{cases} \mathbf{h}_i = MH(\mathbf{v}_i) \\ \mathbf{v}_i = [v_{1,i-b}, \dots, v_{M,i-b}, \dots, v_{1,i}, \dots, v_{M,i}] \\ \mathbf{h}_i = W_i(\cdot) \end{cases} \quad (2)$$

$$\begin{cases} v_{m,i} = \frac{(x_i^m - \min_j \{x_j^m\})}{\max_j \{x_j^m\} - \min_j \{x_j^m\}} \times L_m \\ I_i = \frac{(y_i - \min_j \{y_j\})}{\max_j \{y_j\} - \min_j \{y_j\}} \times 255 \end{cases} \quad (3)$$

式中, N 为样本个数; L_m 为非负整数, 表示超图像的体素坐标范围; $v_{m,i}$ 和 I_i 分别为影响因素分量和目标量在时刻 i 处的标准化值; $W_i(\cdot)$ 表示实现由样本至体素值的映射。在单步预测任务中, 超图像为分辨率为 $\times_m = 1^M L_m$ 的8位特殊图像。为便于表述, 下文将坐标为 $\mathbf{v}_i$ 的体素称为体素 $\mathbf{v}_i$ 。

光伏功率数据的可重复性使得多点数据映射至同一体素, 因此, 定义体素的隶属样本以支持由体素向序列样本的查询。

定义1: 体素隶属样本集。因素量映射至体素位置的机制将导致多个样本映射至同一体素, 单个体素 $\mathbf{v}_k \in H$ 对应的隶属样本的时间戳集合 $\mathbf{s}_k = \{t \mid \mathbf{p}_t \in MH_{\mathbf{v}_k}, \mathbf{p}_t \in \mathbb{R}^n, t \in N^+\}$, 其对应的隶属样本 $\mathbf{S}_k = \{\mathbf{p}_t \mid t \in \mathbf{s}_k\}$ 。

根据影响因素取值与体素重心 $\mathbf{c}_k$ 的欧氏距离 $d_{\mathbf{x}^i, \mathbf{v}_k}$ 来确定各样本至目标体素值的贡献权重。体素 $\mathbf{v}_k$ 对应的一个隶属样本 $\mathbf{p}_i \in \mathbf{S}_k$, 其对体素值的权重 ω_k^i 为

$$\begin{cases} \|\mathbf{R}_i, \mathbf{c}_k\|_2 = \sum_{j=1}^m (x_i^j - c_k^j)^2 \\ \omega_k^i = \frac{\|\mathbf{R}_i, \mathbf{c}_k\|_2}{\sum_{t \in \mathbf{s}_k} \|\mathbf{R}_t, \mathbf{c}_k\|_2}, i, t \in N^+ \end{cases} \quad (4)$$

式中, $\mathbf{c}_k$ 为体素的重心。贡献权重与其影响因素到体素重心的距离成反比, 式(2)中体素值为

$$h_k = W_k(\mathbf{S}_k) = \sum_{j \in \mathbf{s}_k} \omega_k^j I_j \quad (5)$$

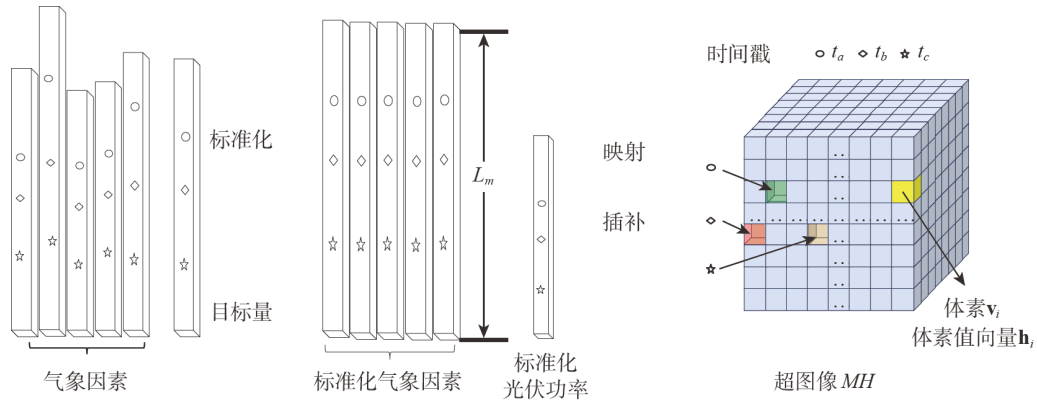


图2 超图像构建流程

Fig.2 Hyperimage generation process

超图像提供了一种考虑预测机制的时序表示方法, 体素与其体素值向量的对应关系反映了影响因素与目标量间的数值依赖关系。具体而言, 影响因素越相近的样本对应的目标量也越相近; 而邻近体素和体素向量的连续关系则保留了时序依赖, 构造超图像的具体流程如图2所示。

基于超图像可将原本时域的序列异常样本检测转化为超图像的体素分析问题。

命题1: 当 $y' = F(\mathbf{R}')$, $\mathbf{R}' = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^m\}$, 其中 $\mathbf{x}_i^k = [x_1^k, x_2^k, \dots, x_h^k]$ 。假设原始时间序列中时刻 i 处的样本满足 $F(\mathbf{R}_i, y_i) = 0$, 则 $\|\mathbf{R}_i, \mathbf{R}_j\|_2 \rightarrow 0$ 时, 有 $|y_1 - y_2| \rightarrow 0$ 。

证明如下:

$$\begin{aligned} \|\mathbf{R}_i, \mathbf{R}_j\|_2 \rightarrow 0 \text{ 时有 } |x_{i-b}^a - x_{j-b}^a| \rightarrow 0, \text{ 而} \\ |y_i - y_j| &= |F(\mathbf{R}_i) - F(\mathbf{R}_j)| \\ &= \sum_{a=1}^m \sum_{b=1}^h \frac{\partial(F)}{\partial(x_b^a)} (x_{i-b}^a - x_{j-b}^a) \end{aligned} \quad (6)$$

式中, $\frac{\partial(F)}{\partial(x_b^a)}$ 为常数值, 因此 $|y_1 - y_2| \rightarrow 0$ 。

命题1证明了样本间历史因素片段与未来功率向量的近邻一致性。基于以上命题结合超图像的构建原理, 可作如下推理:

推论1: 超图像中两个相邻体素的距离满足 $\text{dis}(\mathbf{v}_i, \mathbf{v}_j) < \varepsilon$ 时, 其对应样本的体素值 h_i, h_j 满足 $|h_i, h_j| < \eta$ 。

证明如下：

$$\begin{aligned} \text{dis}(\mathbf{v}_i, \mathbf{v}_j) &= \sqrt{\sum_{a=1}^m \sum_{b=1}^h (v_{a,i-b} - v_{a,j-b})^2} = \sqrt{\sum_{k=1}^m \sum_{t=1}^h \left[\frac{(x_i^k - \min\{x_t^k\})}{\max\{x_t^k\} - \min\{x_t^k\}} \times L - \frac{(x_j^k - \min\{x_t^k\})}{\max\{x_t^k\} - \min\{x_t^k\}} \times L \right]^2} \\ &= \sqrt{\frac{L}{\max\{x_t^k\} - \min\{x_t^k\}} \sum_k \sum_t (x_{i-t}^k - x_{j-t}^k)^2} = \sqrt{K \sum_k \sum_t (x_{i-t}^k - x_{j-t}^k)^2} \end{aligned} \quad (7)$$

$$\text{由 } h_k = \sum_{j \in s_k} \omega_k^j I_j = 255 \times \sum_{j \in s_k} \omega_k^j \frac{(y_i - \min\{y_j\})}{\max\{y_j\} - \min\{y_j\}}, \text{ 有}$$

$$\begin{aligned} |h_1 - h_2| &= 255 \times \left| \sum_{i \in s_1} \omega_1^i \frac{(y_i - \min\{y_j\})}{\max\{y_j\} - \min\{y_j\}} - \sum_{d \in s_2} \omega_2^d \frac{(y_d - \min\{y_j\})}{\max\{y_j\} - \min\{y_j\}} \right| \leq \\ &255 \left| \frac{(y_q - \min\{y_j\})}{\max\{y_j\} - \min\{y_j\}} - \frac{(y_p - \min\{y_j\})}{\max\{y_j\} - \min\{y_j\}} \right| \end{aligned} \quad (8)$$

即 $|h_1 - h_2| \leq K |y_q - y_p|$ 。其中， y_q 、 y_p 分别为 s_1 、 s_2 对应样本中目标量的最大值、最小值。由命题1可知，存在极小正数 ς 使得 $\|\mathbf{R}_q, \mathbf{R}_p\|_2 < \varsigma$ ，则由推论1有 $|y_q - y_p| < \eta_1$ ，故 $|h_1 - h_2| < \eta$ ， $\eta = K\eta_1$ 。

推论2：对于在时间上连续的两个样本 $\mathbf{p}_i = [\mathbf{R}_i, y_i]$ 、 $\mathbf{p}_{i+1} = [\mathbf{R}_{i+1}, y_{i+1}]$ ，由于序列的连续性，有 $\|\mathbf{R}_i, \mathbf{R}_{i+1}\|_2 < \varepsilon$ ，且 $|y_i - y_{i+1}| < \lambda\varepsilon$ ， λ 为常数。设其对应的体素分别为 $\mathbf{v}_a$ 、 $\mathbf{v}_b$ ，则有 $\text{dis}(\mathbf{v}_i, \mathbf{v}_j) \propto \|\mathbf{R}_i, \mathbf{R}_{i+1}\|_2$ ， $|h_a - h_b| \propto |y_i - y_{i+1}|$ 。当 ε 为足够小的数时，则有 $a = b$ 。

推论1说明了超图像中近邻体素间保留了其对应的隶属样本的相似性，并且打破了原来时序中连续样本的时间戳限制，保留了时间序列的空间相似信息。推论2说明了体素内隶属样本的相似性保留了时间序列的依赖性。因此，超图像具备反映多元时间序列中目标量与影响因素间具有的稳定数量关系与时间依赖性的能力，从而为进一步检测异常样本提供了理论依据。

3 基于超图像的异常检测

正常情况下，历史因素与未来光伏功率之间存在稳定的函数关系： $y_i = F(\mathbf{R}_i)$ ，当发生异常时 $y_i \neq F(\mathbf{R}_i)$ 。为避免预测模型学习数据集中的异常函数关系，本节针对预测任务中的训练集展开样本异常检测，并将训练集划分为离线训练集和在线测试集；同时将异常检测过程分为离线参考超图像提取和在线异常检测两个阶段，离线阶段主要利用超图像建模气象因素与光伏功率的正常模式，在线检测阶段则利用正常模式下的超图像对测试数据进行异常判别。

3.1 参考超图像构建

为避免历史数据中可能存在的无标签异常影响，本文采用无监督极大异常抑制策略实现对超图像的一致性优化，确保构建的超图像能够反映数据的正常分布。为实现异常样本的精确定位，本文定义体素内异常候选集 A_k 以支持由异常体素向时域中异常样本的查询。

定义2：体素内异常候选集。对某一体素 $\mathbf{v}_k$ 的隶属样本集按照目标量偏离均值水平的程度进行排序，即 $\mathbf{S}'_k = \text{sort}(\{\mathbf{p}_t \mid \mathbf{p}_t \in \mathbf{s}_t\}, \text{key} = \lambda_t : p_t^y)$ ， $\text{sort}(\cdot)$ 表示依据样本 $\mathbf{p}_t$ 的目标分量值 p_t^y 对 $\mathbf{p}_t$ 进行升序排列；基于 3σ 原则确定离群点为异常候选集 A_k ，定义如下：

$$\mathbf{A}_k = \begin{cases} \mathbf{S}'_k, \text{count}(\mathbf{S}'_k) < 2 \\ \mathbf{S}'_k^{(m)}, \left| p_{m+1}^y - \mu_{p_t^y} \right| > 3\sigma_{p_t^y}, p_t^y \in \mathbf{S}'_k \end{cases} \quad (9)$$

式中， $\text{count}(\cdot)$ 表示非零元素计数值； $\mu_{p_t^y}$ 和 $\sigma_{p_t^y}$ 分别表示样本集目标分量的均值和标准差。

受日内光照时间和光照强度的差异影响，光伏功率的高值区样本明显少于低值区样本，这可能导

致超图像有值体素的分布不均衡。因此，将超图像进行二值化后再进行连通域标记以定位离群体素，再根据该离群点对应的样本候选集实现对异常样本的精确定位。采用滑动窗口对超图像进行等大小的区域提取，滑动步长为1，提取标记区域中周围与中心体素连接形成的非0体素对数值，若该值 <3 ，则认为该样本对应的影响因素分布异常。

对于超图像中的有值体素连续区，首先对各体素内隶属样本的功率值提取统计特征，如极大值、极小值、均值、方差等，构建该体素点的特征向量，并展开为多层平面特征；然后利用高斯加权增强背景，并基于Reed-Xiaoli (RX) 算法检测异常体素点；最后由异常候选集来确定异常样本，分别将多层特征 $\mathbf{O}_1$ 、 $\mathbf{O}_2$ 、 $\mathbf{O}_3$ 、 $\mathbf{O}_4$ 、 $\mathbf{O}_5$ 记为

$$\begin{aligned} \mathbf{O}_1(d) &= \max(\{p\}) \\ \mathbf{O}_2(v_k) &= \min(\{p\}) \\ \mathbf{O}_3(v_k) &= \text{mean}(\{p\}) \\ \mathbf{O}_4(v_k) &= \max\{p_i - p_{i+1}\}_{i=1}^{r-1} \\ \mathbf{O}_5(v_k) &= \frac{\sum_{i=1}^r (p_i - \mu_p)^2}{r} \end{aligned} \quad (10)$$

式中， $\{p\} = \text{sort}_y(\mathbf{S}_{v_k})$ ； r 为隶属样本的数目。基于特征图进行背景增强以放大并量化异常样本导致的体素间的差异，背景增强流程如图3所示。

将各个特征图归一化后线性相加构成符合视觉显著性的显著性图。

$$\mathbf{G} = \overline{\mathbf{O}}_1 + \overline{\mathbf{O}}_2 + \overline{\mathbf{O}}_3 + \overline{\mathbf{O}}_4 + \overline{\mathbf{O}}_5 \quad (11)$$

式中， $\overline{\mathbf{O}}_i$ 表示归一化的特征图。对 $\mathbf{G}$ 进行高斯加权得到显著性权重矩阵 $\mathbf{W}$ 。

$$w(i,j) = e^{-\frac{g(i,j)-\mu}{2\sigma^2}} \quad (12)$$

式中， $g(i,j)$ 为显著性图的元素值； μ 为显著性图的均值； σ 为显著性图的方差。 $\mathbf{G}$ 对 X_h 进行显著性加权得到加权特征图 X_g 。

$$X_g = \mathbf{W} \odot X_h \quad (13)$$

式中， $\odot$ 表示逐段进行加权运算。由于加权后数据背景被增强，放大了异常前景与背景的差异，利用增强背景 X_g 进行均值和协方差的重新估计，进而对其进行异常检测。

对于经过背景增强的多段特征超图像，本文利用RX算法进行异常检测。RX算法是一种具有恒虚警率的异常点探测器， v_i 表示体素 v_i 对应的特征值，检测算子为

$$\begin{cases} \zeta_{\text{anv}}(v_i) = (v_i - \mu_b)^T C_b^{-1} (v_i - \mu_b) \\ \mu_b = \frac{1}{K} \sum_{i=1}^N v_i \\ C_b = \frac{1}{K} \sum_{i=1}^N (v_i - \mu_b)(v_i - \mu_b)^T \end{cases} \quad (14)$$

式中， $\zeta_{\text{anv}}(\cdot)$ 实现对体素的异常评分， K 为超图像中像素点个数； μ_b 为背景均值； C_b 为背景协方差。基于一般置信度的思想，当响应值 $\zeta_{\text{anv}}(v_i) > 0.997$ 时即

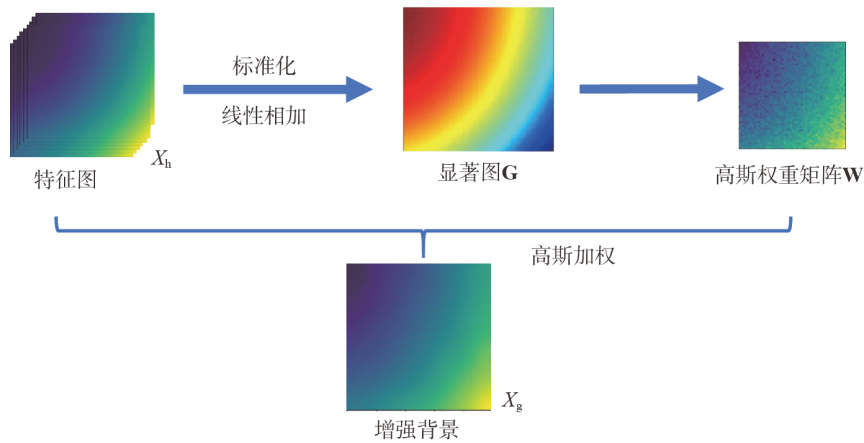


图3 基于多段特征的背景增强示意

Fig.3 Schematic of background extraction process based on multi-segment feature

可将体素 $\mathbf{v}_i$ 视为异常体素。

对于以上得到的异常体素对应的异常样本，更新该异常样本的目标量。

$$\begin{cases} p_i^{x_j*} = p_m^{x_j} \\ p_i^{y*} = \sum_{p_m \in S_k \& p_m \notin A_k} p_m^y \end{cases} \quad (15)$$

此时，更新后的历史数据集已有效剔除原始异常值的干扰。基于此构建的超图像 H^* 将为后续在线异常检测提供高可靠性的参考基准。

3.2 基于超图像的在线异常样本检测

为有效解决海量数据计算复杂度高的问题，本节定义超图像中体素的查询样本。

定义3：体素内查询样本。对于体素 $\mathbf{v}_k$ 对应的各隶属样本 $\mathbf{p}_i \in S_k$ ，有权重 ω_k^i ，对应的查询样本定义为 S_k 中权重最大即距离体素重心最近的样本。

在基于历史数据得到经优化的参考超图像 H^* 后，对于待测试样本 $\mathbf{p}_i$ ，根据该样本与参考超图像各体素的查询样本在影响因素维度上的欧氏距离来确定测试样本的映射目标体素 $\mathbf{v}_k$ ，即 $\mathbf{v}_k : \min_H \|\mathbf{x}^{i-a}, \mathbf{c}_k^*\|_2$ 。计算测试样本的异常得分：

$$z_i = \left| \frac{y_i - h_k}{\tau + h_k} \right| \quad (16)$$

式中， h_k 为参考体素值； τ 为防止分母为0的极小因子。该得分反映了测试样本点的目标量偏离正常值的情况，当得分超过设定阈值 δ 时即为异常样本点。

3.3 异常及缺失数据修复

一般地，直接删除异常数据会破坏原始数据的完整性，严重影响预测精度。考虑到设备冲击等非稳定因素导致的异常具有罕见性以及光伏功率与气象因素之间的相关性，将检测出的异常数据视为目标量缺失情况，对其进行修复。

对于在线阶段检测为异常的样本，将其目标量修正为目标体素 $\mathbf{v}_k$ 隶属样本功率值的截尾均值。

$$y_i^* = \frac{1}{s-2u} \sum_{j=u+1}^{s-u} p_j^y, \mathbf{p}_i \in S_k \quad (17)$$

式中， u 为被截去的观测值数目。

在缺失值填补方面，当缺失量为影响因素时，根据原始时间戳信息进行线性插补，并默认此时对应样本的目标量为未知情形，然后基于参考超图像 H^* 根据因素值确定映射体素 $\mathbf{v}_i$ ，再由式（17）填补其目标量。

4 实验结果与分析

本节主要针对以下3个问题开展3组实验，以验证本文方法的有效性：

- （1）在线样本异常检测任务中的性能表现；
- （2）基于超图像的缺失数据修复性能表现；
- （3）本文提出的历史数据修复方法对预测精度的影响。

4.1 数据集及其预处理

实验数据集为澳大利亚沙漠知识太阳能中心的91号集群光伏电站中1B站点采集的气象因素和光伏功率数据，数据时间分辨率为5min。本文选取2017年2月1日0:00至2017年4月10日18:00的19873条数据作为历史训练集，取其中60%的数据构建参考超图像；以2017年4月10日至15日的1658条数据作为测试集。记录的序列包括风速、温度、相对湿度、全球水平辐照度、散射水平辐照度、风向、日降水量、全局倾斜辐照度、散射倾斜辐照度等9项气象信息，以及有功功率序列。

数据预处理包括以下关键步骤：识别并去除数据集中物理意义上的明显异常值、影响因素筛选和数据标准化处理。本文首先基于光伏功率和气象因素的物理意义，将训练集中的负数转变为0值，并记录缺失数据标签，去除缺失处数据；然后采用第1节的方法来筛选影响因素量。为避免计算精度过高而影响决策树性能，将功率值放大1000倍再送入随机森林模型。统计各气象因素与有功功率序列的斯皮尔曼相关系数和特征重要性，如图4所示。

可以看出，散射水平辐照度、全球水平辐照度、风速、温度和相对湿度等因素与有功功率的斯皮尔曼相关系数较高，但辐照度特征对有功功率的预测重要性更高。为避免冗余计算，本文选取与光伏功率最为相关的全球水平辐照度和散射水平辐照度因素作

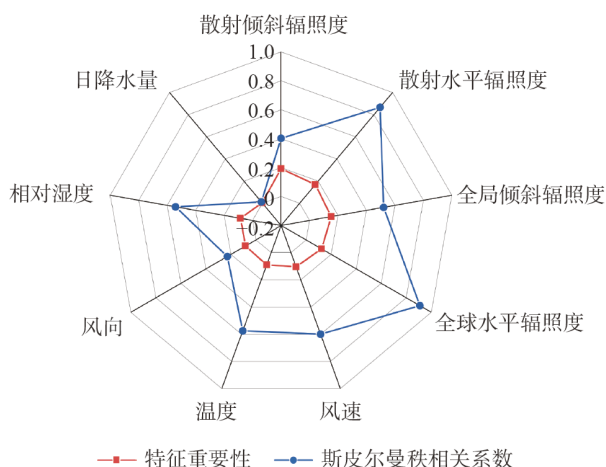


图4 影响因素与目标量的相关性分析

Fig.4 Correlation analysis between influencing factors and target indicators

为超图像的位置因素，而在预测环节再增加全局倾斜辐照度、温度、相对湿度、风速作为输入特征。

根据式(3)对历史训练集数据进行标准化处理。其中，影响因素对应的因子 L 为超图像尺寸，目标量对应因子为255，以便于进行基于图像的检测与可视化。

4.2 评价指标

为验证本文所用的异常检测方法的有效性，本文使用精确率(Precision)、召回率(Recall)和F1分数来评价异常识别的性能。精确率指检测正确的正样本数占有所有检测为正样本数的比例；召回率指模型成功预测的正样本数占有所有实际正样本数的比例；F1分数指精确率和召回率的调和均值，是最常用的异常检测综合评价指标。混淆矩阵如表1所示。

表1 混淆矩阵
Table 1 Confusion matrix

	真实为异常	真实为正常
检测为异常	真阳性 T_p	假阳性 F_p
检测为正常	假阴性 F_N	真阴性 T_N

上述异常检测评价指标的计算公式如下：

$$Precision = \frac{T_p}{T_p + F_p} \quad (18)$$

$$Recall = \frac{T_p}{T_p + F_N} \quad (19)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (20)$$

为验证本文异常修复方案的有效性，使用经过本文方法处理后的训练集来训练预测模型，并使用平均绝对误差(Mean Absolute Error, MAE)、均方根误差(Root Mean Square Error, RMSE)和决定系数(R^2)对预测结果进行分析。上述预测性能评价指标的计算公式分别为

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (21)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (22)$$

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (\hat{y}_i - \bar{y})^2} \quad (23)$$

式中， y_i 表示对未来时刻 i 处的目标量的观测值； $\hat{y}_i$ 为对应的预测值。 R^2 的取值范围为0~1，且越接近1表明模型对数据的拟合效果越好，解释能力越强。

4.3 实验分组设计和结果分析

本文所提方法涉及的超参数主要为超图像尺寸 L 和异常阈值 δ 。光伏功率数据具有长尾分布特性，在超图像中容易产生过度密集。为避免低值样本降低近邻体素内隶属样本的可分度，对于既定的历史数据，设定超图像尺寸为整10倍数，通过变化超图像的尺寸并统计其各尺寸下的适应度来得到最佳尺寸 L_m^* 。定义超图像对数据的适应度为

$$\alpha = Q \cdot g(H) \quad (24)$$

式中， Q 为超图像中非0体素个数； $g(\cdot)$ 表示对超图像进行逐体素的梯度求和。适应度越大，表明样本映射至超图像空间中的体素内一致性越强，越有利于异常检测。不同尺寸下的适应度如图5所示。可以看出，当尺寸 $L=120$ 时，适应度达到峰值。此后继续增大尺寸，适应度反而下降。因此，设定实验中超图像尺寸 $L_m^* = 120$ 。

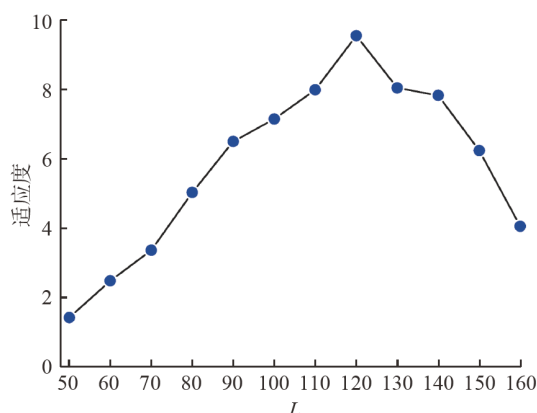


图5 超图像尺寸适应度曲线
Fig.5 Hyperimage size fitness curve

4.3.1 异常检测

为避免夜间全零功率值带来的计算冗余，本文仅考虑日间7:00—18:00的数据用于构建超图像并进行异常检测，且由于缺乏真实异常实例，对光伏功率数据进行人为添加异常来模拟。对于历史训练集，设置添加异常的比例为0.05~0.30；所加异常类型为随机放大或缩小，且数值变化幅度为原始有功功率标准差的1~3倍，以充分模拟随机异常。各组划分构建参考超图像和在线异常检测的比例均为3:2，参考超图像尺寸统一为 L_m^* ，设置异常松阈值 δ 为0.3。

极大异常抑制前后的离线阶段参考超图像如图6所示。可以看出，原始历史数据构建的超图像具有明显的离群现象，即对应影响因素的异常分布在超图像的像素值密集区域存在像素值变化明显的突出点，即违背了影响因素与目标量间稳定数量关系的异常点。极大异常抑制处理后，离群异常得到了有效去除，为在线的异常判别提供了衡量依据。

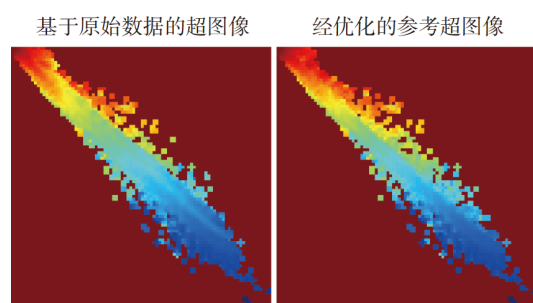


图6 参考超图像优化结果
Fig.6 Refer hyperimage optimization results

对于待在线检测的数据，统计模拟不同比例异常情况下的检测指标，如表2所示。其中， 2σ 、 3σ 分别表示所添加的随机异常幅值相比于原始数据标准差的最大倍数值。

表2 在线异常检测性能指标
Table 2 Performance index of online anomaly detection

异常比例	2σ			3σ		
	精确率	召回率	F1分数	精确率	召回率	F1分数
0.05	58.83	88.20	72.86	85.00	99.80	91.80
0.10	65.45	91.40	77.60	93.18	99.80	96.37
0.15	78.44	91.00	84.25	96.13	99.46	99.77
0.20	84.36	90.95	87.93	97.35	99.51	98.41
0.25	88.98	91.44	90.19	98.06	99.46	98.72
0.30	92.19	90.93	91.55	98.57	99.36	98.97

可以看出，松阈值设置条件下，在线检测的召回率均高于85%，且随着异常比例的增大，采用逐样本的检测优势更加突出，F1分数高达98%。

4.3.2 缺失值修复

本节通过人为添加缺失值进行算法性能评估。对于给定缺失比例 δ ，本文首先随机选取对应数目的样本数据，并随机去除该时间戳数据的任一影响因素或目标量；然后根据3.3节所提方法进行修复。对于不同缺失比例情况下的同一组测试数据，以均值填补、基于K近邻填补、最近邻填补法为对比方法，分别统计各类填补方法结果的误差指标，如表3所示。

表3 不同填补方法的误差
Table 3 Errors of different filling methods

缺失比例		0.1	0.2	0.3	0.4	0.5
本文方法	MAE	0.45	0.44	0.43	0.42	0.42
	RMSE	0.94	0.97	0.92	0.93	0.91
均值法	MAE	1.17	1.08	1.08	1.10	1.12
	RMSE	1.51	1.37	1.37	1.40	1.43
最近邻法	MAE	0.63	0.45	0.50	0.50	0.50
	RMSE	1.26	0.98	1.08	1.08	1.08
K近邻法	MAE	0.58	0.44	0.44	0.45	0.45
	RMSE	1.14	0.92	0.92	0.94	0.96

图7为对在线检测得到的异常进行修复的修复结果。可以看出，本文检测方法能够有效检测出违背

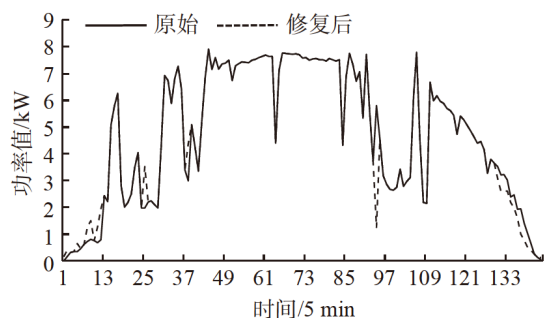


图7 异常检测及修复结果
Fig.7 Anomaly detection and remediation results

气象因素与功率之间稳定关系的异常值，并对其进行纠正。

4.3.3 预测模型及修复前后预测效果

长短期记忆 (Long Short-Term Memory, LSTM) 网络具备更少的参数，结构更为简单，且在性能相当的情况下对长期时间序列数据的拟合效果更好。因此，本文选用LSTM作为预测模型，设置模型参数隐藏单元数量为128，批次大小为32，训练轮次为100，网络层数为1。

由于真实情况下的光伏功率异常具有不可控性，且真实极大异常出现概率极低，因此本文在4.1节模拟异常的基础上设置所添加异常的数值变化幅度为原始标准差的 λ 倍，且 $\lambda \in (1, 2)$ 。分别使用基于本文方法处理后的训练集、基于LOF检测异常与线性插值 (Local Outlier Factor & Linear Interpolation, LOF_L) 处理后的训练集来训练预测模型，训练集数据修复

前后训练所得模型对同一测试集的预测指标如表4所示。

未添加异常时，经本文修复方法处理数据的预测结果相比于原始数据训练的预测结果的MAE减小了44.74%，RMSE减小了42.73%， R^2 提高了0.11%；相比于LOF_L方法训练的预测结果的MAE减小了44.74%，RMSE减小了50.39%， R^2 提高了0.11%。而在不同的异常添加比例下，与原始数据训练的预测结果相比，LOF_L方法并没有明显提升预测性能，甚至个别结果的性能有所下降，而本文修复方法能够有效提升预测性能。

表5统计了多步预测任务中各组实验的预测指标。对比表4、表5可以看出，2步预测模型的预测性能相比于单步预测有明显下降。未添加异常时，经本文修复方法处理后，MAE减小了1.63%，RMSE减小了1.14%， R^2 提高了0.0431%，而使用LOF方法几乎没有改变预测结果。在不同的异常添加比例下，对比原始数据训练结果可知，进行异常检测并修复的处理具有提升预测性能的效果。且随着异常比例增大，本文方法相比于LOF_L方法的优势更加突出。在异常比例达到0.30时，本文方法的预测结果相比于原始数据的MAE减小了34.24%，RMSE减小了33.47%， R^2 提高了12.45%。

未添加异常情况下，不同预测步数下的预测结果如图8所示。可以看出，原始数据的预测曲线与LOF_L方法的几乎完全重合，表明LOF_L处理对预测

表4 不同方法下的单步预测结果对比
Table 4 Comparison of single-step prediction results under different methods

异常比例		0	0.05	0.10	0.15	0.20	0.25	0.30
MAE	原始	0.076	0.053	0.093	0.071	0.033	0.086	0.043
	LOF_L	0.076	0.046	0.052	0.056	0.045	0.063	0.082
	本文方法	0.042	0.041	0.051	0.070	0.056	0.054	0.042
RMSE	原始	0.110	0.043	0.053	0.063	0.065	0.078	0.049
	LOF_L	0.127	0.090	0.085	0.081	0.060	0.068	0.090
	本文方法	0.063	0.041	0.047	0.051	0.065	0.067	0.048
R^2	原始	99.796	99.952	99.887	99.919	99.988	99.956	99.965
	LOF_L	99.796	99.960	99.92	99.968	99.983	99.961	99.962
	本文方法	99.904	99.967	99.923	99.912	99.964	99.968	99.971

表5 不同方法下的2步预测结果对比

Table 5 Comparison of 2-step prediction results under different methods

异常比例		0	0.05	0.10	0.15	0.20	0.25	0.30
MAE	原始	0.184	0.625	0.632	0.612	0.750	0.728	0.698
	LOF_L	0.184	0.586	0.633	0.716	0.727	0.661	0.659
	本文方法	0.181	0.346	0.450	0.488	0.435	0.490	0.459
RMSE	原始	0.440	1.523	1.538	1.524	1.845	1.748	1.688
	LOF_L	0.441	1.436	1.483	1.658	1.813	1.587	1.642
	本文方法	0.435	0.765	0.986	1.165	0.999	1.217	1.123
R^2	原始	97.437	84.479	83.853	82.870	75.098	76.893	80.922
	LOF_L	97.437	85.999	84.358	80.591	76.024	82.785	81.868
	本文方法	97.479	94.348	91.759	89.253	91.277	87.182	90.999

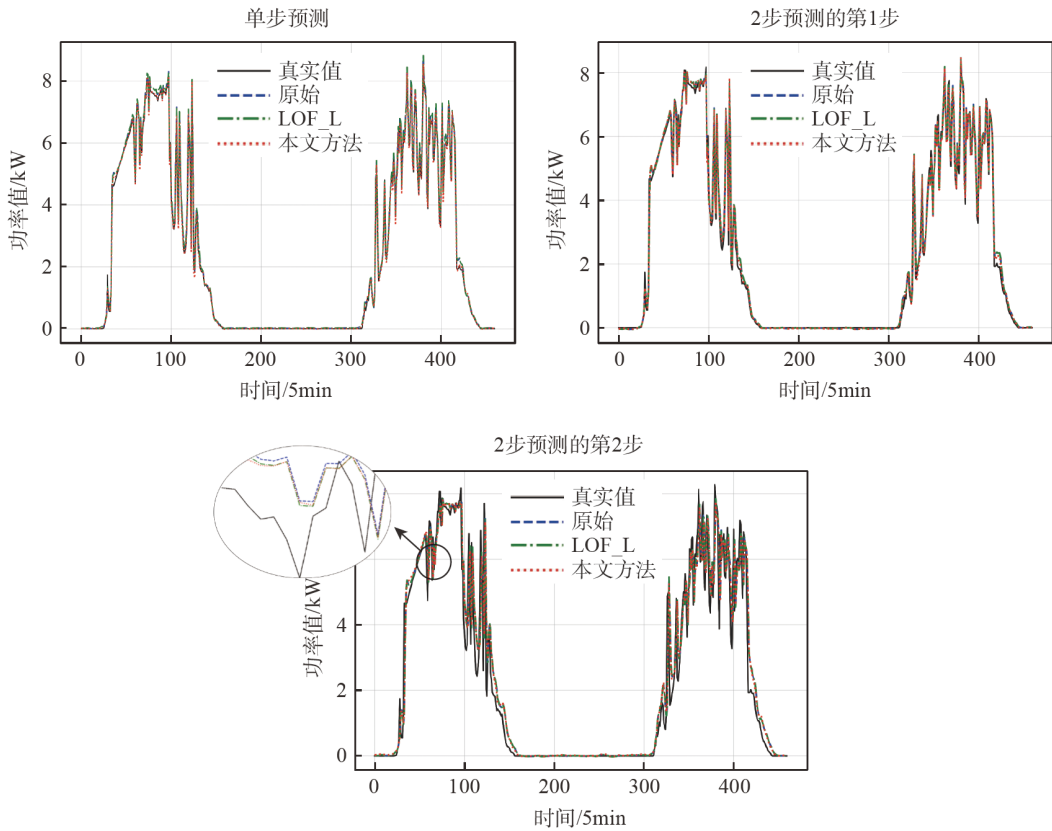


图8 不同预测步数下的预测结果对比

Fig.8 Comparison of prediction results under different prediction steps

结果几乎无影响。真实光伏功率呈现明显的波动特性，本文方法在高峰区的预测值更接近实际值，在单步预测中整体预测趋势几乎与真实曲线重合；而在2步预测中，第1步的预测精度明显高于第2步。同时，本文方法在功率低谷、高峰等关键节点的预测值更加贴近真实值，展现出更优的多步预测跟踪能力。

5 结论

本文提出了一种面向预测任务的光伏功率异常检测及修复方法，使用真实数据验证了方法的有效性，主要结论如下：

(1) 基于超图像实现光伏功率的两阶段无监督

异常检测。仿真实验表明, 离线阶段构建的参考超图像为异常检测提供了良好的解释性依据, 在线检测阶段能够实现快速有效地检测。

(2) 基于超图像进行光伏功率异常和缺失数据的修复, 相较于经典的均值填补、近邻填补方法, 其结果更为精确。

(3) 即使是在多发小幅异常情况下, 历史数据中异常的存在将明显降低模型预测性能。真实算例分析表明, 构建基于预测样本的超图像进行异常检测并修复, 在单步和多步预测任务中均具有一定的适用性, 能够有效提升预测模型的预测精度。

在后续研究中, 进一步探索异常检测阈值的自适应机制, 更准确地规避由于误判样本引进的修复误差, 增强对长时间序列的异常捕捉能力, 进而加强预测模型对有效特征的建模, 是提升预测性能的发展方向。

参考文献

- [1] 张凯, 袁家海, 丁保迪, 等. 新型电力系统灵活性资源提升效益评估[J]. 气候变化研究进展, 2024, 20(3): 316–326.
- [2] 张金平, 汪宁渤, 黄蓉, 等. 高渗透率光伏参与电力系统调频研究综述[J]. 电力系统保护与控制, 2019, 47(15): 179–186.
- [3] Jiao X Y, Sun Y Z, Peng D G, et al. Photovoltaic power abnormal data cleaning based on variance change point and correlation analysis[C]//2021 6th International Conference on Power and Renewable Energy (ICPRE). Shanghai, China, 2021: 1140–1145.
- [4] 雷柯松, 吐松江·卡日, 伊力哈木·亚尔买买提, 等. 基于WGAN-GP和CNN-LSTM-Attention的短期光伏功率预测[J]. 电力系统保护与控制, 2023, 51(9): 108–118.
- [5] 周伊, 肖先勇, 赵清华, 等. 基于组合数据清洗与改进注意力机制的光伏发电功率预测[J]. 供用电, 2024, 41(10): 31–37.
- [6] Li T, Peishan F A M, Ying T, et al. Improved dirichlet mixture model clustering algorithm for medical data anomaly detection [J]. International Journal of Bio-Inspired Computation, 2024, 1: 10064803.
- [7] Abuzaid A H. Identifying density-based local outliers in medical multivariate circular data[J]. Statistics in Medicine, 2020, 39(21): 2793–2798.
- [8] Moore A, Morelli D. conDENSE: Conditional density estimation for time series anomaly detection[J]. Journal of Artificial Intelligence Research, 2024, 79: 801–824.
- [9] 赵唯嘉, 张宁, 康重庆, 等. 光伏发电出力的条件预测误差概率分布估计方法[J]. 电力系统自动化, 2015, 39(16): 8–15.
- [10] Meng F, Gao Y, Wang H H, et al. TSLOD: A coupled generalized subsequence local outlier detection model for multivariate time series[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(5): 1493–1504.
- [11] Williams A T, Sperl R E, Chung S M. Anomaly detection in multi-seasonal time series data[J]. IEEE Access, 2023, 11: 106456–106464.
- [12] Barraz Z, Sebari I, Oufettoul H, et al. A holistic multimodal approach for real-time anomaly detection and classification in large-scale photovoltaic plants[J]. Energy and AI, 2025, 21: 100525.
- [13] 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.
- [14] Miraftebadeh S M, Longo M, Leva S, et al. Data anomaly detection in photovoltaic power time-series via unsupervised deep learning with insufficient information[J]. Sustainable Energy, Grids and Networks, 2025, 43: 101769.
- [15] 王剑斌, 傅金波, 陈博. 基于强化学习的多模型融合光伏发电功率预测方法[J]. 太阳能学报, 2024, 45(6): 382–388.
- [16] 孟安波, 陈嘉铭, 黎湛联, 等. 基于相似日理论和CSO-WGPR的短期光伏发电功率预测[J]. 高电压技术, 2021, 47(4): 1176–1184.
- [17] 纪德洋, 金锋, 冬雷, 等. 基于皮尔逊相关系数的光伏电站数据修复[J]. 中国电机工程学报, 2022, 42(4): 1514–1522.