

基于 K -medoids 聚类的分布式光伏集群方法及短期功率预测

蒋亚雪, 钱 晶, 何昊城, 张皓彦, 曹 雷, 毛西蒙

(昆明理工大学冶金与能源工程学院, 昆明 650093)

摘 要: 为实现对光伏集群的科学聚类划分, 该文融合考虑气象因素和地理位置对光伏电站聚类的影响, 将出力特性作为子集群的划分特征, 提出一种基于改进曼哈顿距离的 K 中心点聚类算法 (K -中心点)。光伏集群聚类的两个关键是气象特征的有效识别和各电站的相似性度量, 为此采用变异系数法与秩和比法相结合对气象输入特征进行提取, 可提高特征的识别有效性。另一方面将改进曼哈顿距离引入 K -中心点聚类算法中, 充分挖掘数据的动态特性, 并采用 3 种内部有效指标优化聚类划分数, 避免常规划分带来的分散性和偏差, 提高聚类的整体效果。此外, 在聚类基础上进行预测验证, 为解决预测模型的参数设置问题, 在预测模型中融入优化算法, 采用改进灰狼优化算法 (IGWO) 对 iTransformer 和极度梯度提升 (XGBoost) 模型中的参数进行优化, 功率预测结果表明所提聚类算法的有效性以及较高的精度。

关键词: 光伏发电; 聚类分析; 功率预测; 改进灰狼优化算法; iTransformer; 极度梯度提升

中图分类号: TM615

文献标志码: A

0 引 言

光伏发电已成为中国重要的战略性新兴产业^[1], 国内分布式光伏新增并网规模持续增长。由于光伏电站受气候条件、地理位置等因素的影响, 导致其发电情况差异明显, 若仅对单个光伏电站进行控制预测, 不仅任务繁重, 还会在预测过程中造成误差累积。因此, 为实现分布式光伏电站精准有效的功率预测, 对其进行合理的聚类划分成为关键。目前对于分布式光伏电站典型的集群划分方法有 K 均值 (K -means) 聚类^[2]、近邻传播 (affinity propagation, AP) 聚类算法^[3]、层次聚类算法^[4]和模糊 C 均值 (fuzzy C -means, FCM) 聚类算法^[5]等。针对 K 均值聚类方面的研究, 文献[6]提出以光伏发电特征为基准, 使用 K 均值算法对光伏集群进行聚类划分, 但该算法对偏差点较为敏感; 文献[7]采用 K 均值聚类对光伏电站的地理位置进行了区域划分, 但未考虑影响光伏电站划分的各气象因素, 可能导致划分不精细而对后续工作带来较大误差。层次聚类因其高效灵活的优势也在预测方面得到广泛应用, 文献[8]提出采用层次聚类将气象数据划分到不同天气类别; 文献[9]采用层次聚类挖掘数据信息将光伏数据划分到不同特征的数据库。但层次聚类无明确的优化目标, 聚类结果难以量化。

在光伏电站功率预测方面, 随着人工智能技术的飞速发展, 深度学习和集成学习模型获得广泛研究, 较具代表性的

模型有支持向量机 (support vector machine, SVM)^[10]、门控循环单元 (gated recurrent unit, GRU)^[11]、双向长短期记忆 (bidirectional long short-term memory, BiLSTM)^[12]等。集成学习中的极限梯度提升 (extreme gradient boosting, XGBoost) 算法具有较高的拟合精度和较快的计算速度优势, 已普遍应用在各种预测方面^[13-14]。深度学习中的 iTransformer 模型^[15]在时序预测方面也有较高精度。

本文同时考虑各光伏电站的地理位置和气象因素对光伏电站聚类划分的影响, 提出一种基于改进曼哈顿距离的 K -中心点集群聚类, 并采用改进灰狼算法 (improved grey wolf optimizer, IGWO) 优化 iTransformer 和 XGBoost 模型的分布式光伏集群短期功率预测方法。首先对预处理的数据构造气象因素和功率的数据矩阵 R , 通过变异系数法 (coefficient of variation method, CVM) 和秩和比法 (rank-sum ratio, RSR) 提取出光伏电站聚类划分的输入特征, 然后通过各光伏电站的功率数据计算光伏电站间的改进曼哈顿距离 (improved Manhattan distance, IMD), 并将其作为 K -中心点聚类的指标, 采用 3 种内部有效评价指标确定聚类划分数, 以此构建基于改进曼哈顿距离的 K -中心点聚类算法, 并对各光伏电站进行聚类, 最后采用 IGWO 算法优化 iTransformer-XGBoost 组合模型对该聚类算法进行验证分析, 并将其运用到分布式光伏集群功率预测中。

收稿日期: 2025-01-10

基金项目: 国家自然科学基金 (52269020)

通信作者: 钱 晶 (1967—), 女, 硕士、教授, 主要从事多能互补下水电机组的协同控制方面的研究。qj0117@163.com

1 分布式光伏电站聚类划分

1.1 输入特征选择

各气象因素的变化幅度较大,难以满足近似正态分布规律,而光伏发电功率易受气象因素影响,因此本文采用变异系数法计算每个气象因素在光伏功率中的权重值并将其赋予在秩和比法中进行秩和比值 Z_{RSR} 计算,这种方法能精确量化光伏发电功率和各气象因素间的相关性,进而确定光伏电站输出功率预测模型的输入气象特征。权重是反映指标在整体中重要程度的标志,采用变异系数法对评价指标进行动态赋权,能客观反映评价指标之间的变异程度^[16]。将 6 个评价指标包括太阳辐照度、直接法向辐照度、全球水平辐照度、相对湿度、大气压力、温度和评价对象即光伏发电功率构造数据矩阵 R 进行标准化处理,从而得到矩阵 R' 。

$$R' = (r'_{ij})_{m \times n} = \begin{bmatrix} r'_{11} & \cdots & r'_{1n} \\ \vdots & \ddots & \vdots \\ r'_{m1} & \cdots & r'_{mn} \end{bmatrix} \quad (1)$$

式中: r'_{ij} ——指标数据标准化后的值; r_{ij} ——第 i 个评价对象第 j 个指标上的原始数值。

指标的变异系数 V_j 计算为:

$$V_j = \frac{\sigma_j}{\bar{x}_j} \quad (2)$$

式中: $\bar{x}_j$ ——指标的平均值; σ_j ——指标的标准差。

指标的权重 W_j 计算为:

$$W_j = \frac{V_j}{\sum_{i=1}^m V_j} \quad (3)$$

在计算权重前,通过对评价指标实行非整数秩次转换。极大型指标的评价对象秩 R_{ij} 的公式为:

$$R_{ij} = 1 + (m+1) \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (4)$$

极小型指标的评价对象秩 R_{ij} 的公式为:

$$R_{ij} = 1 + (m+1) \frac{\max(x_j) - x_{ij}}{\max(x_j) - \min(x_j)} \quad (5)$$

将权重赋予在秩和比值的计算中, Z_{RSR} 越接近 1 表明该气象因素与光伏发电功率的相关性越高。计算公式为:

$$W_{\text{RSR}} = \frac{1}{m} \sum_{j=1}^n W_j R_{ij} \quad (6)$$

由于 Z_{RSR} 无量纲,最小值 $Z_{\text{RSR}} = 1/n$,最大值 $Z_{\text{RSR}} = 1$ 。列出各组频数,计算累计频数 Σf ; 计算各组秩次 R 和平均秩次 $\bar{R}$; 计算向下累计频率 $\bar{R}/n$,最后一项用 $1 - 1/4n$ 修正。根据 Σf 查询“百分数与概率单位对照表”,求其对应的概率单位 Probit 值,进行回归分析,计算公式如式(7)所示,通过对比 Z_{RSR} 值确定预测模型输入的气象特征。

$$Z_{\text{RSR}} = a + b \times c_{\text{Probit}} \quad (7)$$

1.2 分布式光伏电站集群划分方法分析

分布式光伏电站的发电功率在不同地理位置表现差异较大,本文采用基于改进曼哈顿距离的 K -中心点算法对分布式光伏电站进行聚类划分,将相似发电功率特征的分布式光伏电站划分为同一个子集群。由于各光伏电站的时间序列数据具有动态特性,反映曲线轨迹波动趋势,相比于欧氏距离度量静态数据之间的相似性^[17],改进曼哈顿距离更能反映时间序列数据间的相似程度,所以在 K -中心点聚类中采用改进曼哈顿距离进行相似性度量。本文将在 3.3 节通过仿真验证所选算法的有效性。

1.2.1 基于改进曼哈顿距离的 K -中心点聚类

采用改进曼哈顿距离 $d_{\text{IMD}}(J, K)$ 描述两个光伏电站间的距离^[18],具体方法为先计算两个光伏电站矩阵中各对对应行向量的曼哈顿距离 $d_{\text{MD}}(x, y)$,再取其平均值作为两个光伏矩阵的改进曼哈顿距离,计算公式为:

$$d_{\text{IMD}}(J, K) = \frac{1}{m} d_{\text{MD}}(x, y) \quad (8)$$

$$d_{\text{MD}}(x, y) = \sum_{i=1}^m \sum_{u=1}^n |x_{iu} - y_{iu}| \quad (9)$$

式中: x_{iu}, y_{iu} ——任两个光伏矩阵 J, K 第 i 行第 u 列的元素。

这里使用基于改进曼哈顿距离的聚类误差平方和 E 来评估聚类结果质量,定义为:

$$E = \sum_{m=1}^k \sum_{n \in A_m} |A - o_m|^2 \quad (10)$$

式中: k ——聚类数量; A ——各簇类聚类中心 o_m 中的样本。

聚类的目标是尽量减小簇内差异,使簇内差距达到最小。 K -中心点算法是采用簇中与其他对象最为相似的实际对象作为聚类簇的中心^[19]。 K -中心点聚类算法步骤为:

- 1) 将数据集中随机挑选的 K 个对象设为初始聚类中心点;
- 2) 通过改进曼哈顿距离确定非中心点与中心点的距离,并将数据集中的其他非中心点对象归入最近中心点所在的簇中;
- 3) 重新计算每个簇中心点位置,使其到该簇其他样本的距离最小;
- 4) 重复步骤 2) 和步骤 3),直到聚类误差平方和基本不变,满足聚类目标为止。

1.2.2 分布式光伏电站集群划分数的确定

聚类算法可通过自身性能评价指标进行聚类质量评估。聚类数过小可能导致数据信息缺失,过大则可能导致模型复杂,只有合适的聚类数才能提高聚类质量,保证聚类结果的准确性。本文通过采用内部有效评价指标包括平均轮廓系数 (average silhouette coefficient, E_{ASC})、Calinski-Harabaz 指数 (E_{CH}) 和 Davies-Bouldin 指数 (E_{DB}) 自动确定划分数,不用依靠主观思维预设聚类数,可避免划分结果不合理的问题,其计

算公式为：

$$\begin{cases} E_{ASC} = \frac{1}{N} \sum_{i=1}^N E_{SC} \\ E_{SC} = \frac{D_2 - D_1}{\max(D_1, D_2)} \end{cases} \quad (11)$$

$$\begin{cases} E_{DB} = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \frac{N_i + N_j}{M_{ij}} \\ N_i = \left\{ \frac{1}{\theta} \sum_{j=1}^{\theta} |Y_{ij} - X_i|^\beta \right\}^{\frac{1}{\beta}} \\ M_{ij} = \left\{ \sum_{k=1}^k |b_{ki} - b_{kj}|^\beta \right\}^{\frac{1}{\beta}} \end{cases} \quad (12)$$

$$\begin{cases} E_{CH} = \frac{\text{tr}(A_k)(N-K)}{\text{tr}(B_k)(K-1)} \\ A_k = \sum_{\beta=1}^k \theta_\beta (c_\beta - c_d)(c_\beta - c_d)^T \\ B_k = \sum_{\beta=1}^k \sum_{y \in \beta} (y - c_\beta)(y - c_\beta)^T \end{cases} \quad (13)$$

式中： N ——样本数量； E_{SC} ——其中一个样本的轮廓系数； D_1 ——样本与其所在簇内其他样本的平均距离； D_2 ——样本与其他簇内所有样本的平均距离； N_i 、 N_j ——样本 i 和 j 到相应簇中心距离的平均值； M_{ij} ——第 i 类与第 j 类中心的改进曼哈顿距离； b_{ki} ——第 i 类的中心点的第 K 个属性的值； β ——各点到中心的距离， $\beta=1$ 表示各点到中心距离的均值，

$\beta=2$ 表示各点到中心距离的标准差； Y_{ij} ——第 i 类中第 j 个数据点； X_i ——第 i 类的中心； θ ——第 i 类中数据点的数量； A_k ——类间的协方差矩阵； B_k ——类内数据的协方差矩阵； c_β ——类 q 的中心点； c_d ——数据集的中心点； θ_β ——类 q 中数据的数量； y ——类 q 的数据集合。

2 基于 IGWO 优化的 iTransformer-XGBoost 预测模型分析

本文使用 IGWO 优化 iTransformer-XGBoost 模型对分布式光伏集群聚类后的子集群进行功率预测，通过对比分析预测结果的误差指标，验证该预测模型的性能。

2.1 iTransformer 模型原理

文献[20]在2023年提出 iTransformer，是将传统 Transformer 模型的注意力模块和前馈网络模块的职责进行倒置，由自注意力机制捕获多变量相关性，并利用前馈网络对每个变量标记进行非线性表征学习。本文采用 iTransformer 构造预测组合模型的主要原因是利用 iTransformer 对分布式光伏电站集群进行特征提取，捕捉光伏数据的时序相关性和非线性关系，从而提高预测精度。iTransformer 模型结构如图1所示。图1中，图1a标记时间序列数据的不同变量独立嵌入到感受野中；图1b使用自注意力机制应用在嵌入式可变标记上以此增强可解释性；图1c共享前馈网络提取嵌入式可变标记的序列；图1d使用层归一化法减少变量之间的差异。

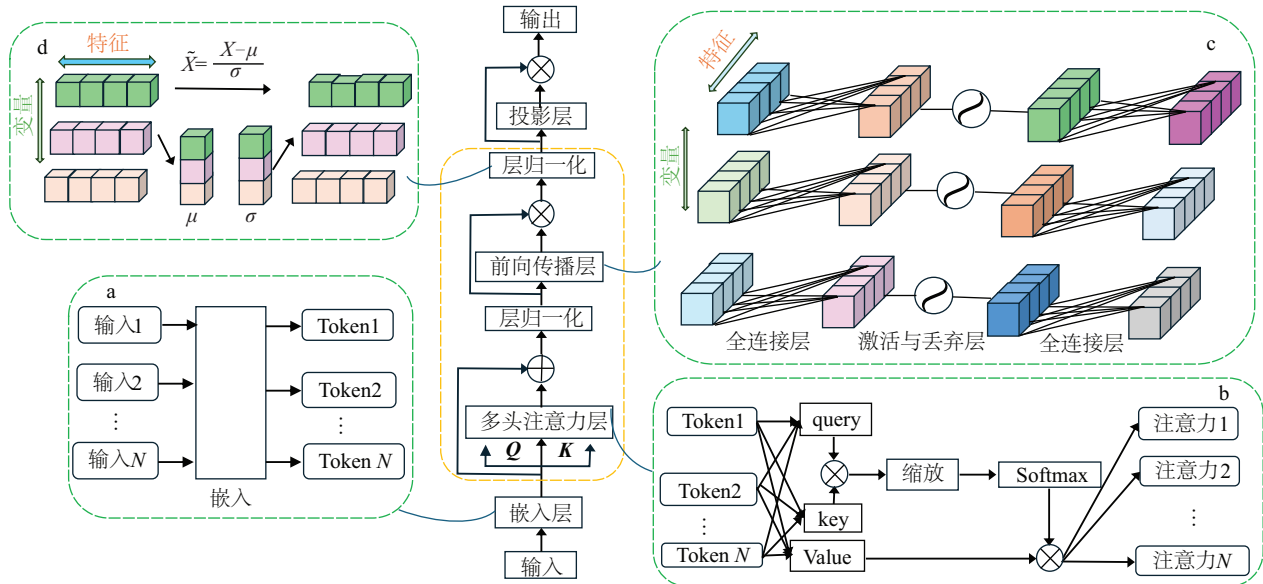


图1 iTransformer 整体结构

Fig. 1 Overall structure of iTransformer

iTransformer 模型包含嵌入层、投影层和可堆叠的 Transformer 模块(TrmBlock)，将光伏功率时间序列和气象因素构成矩阵输入嵌入层，其公式为：

$$\begin{cases} h_a^0 = \text{Embedding}(F; N) \\ H^{l+1} = \text{TrmBlock}(H^l), l = 0, 1, \dots, L-1 \\ \hat{G}_{:,n} = \text{projection}(H_N^L) \end{cases} \quad (14)$$

式中: F 、 N ——气象因素、光伏功率时间序列数据构成的矩阵; Embedding——嵌入操作; h_a^0 ——嵌入操作后的气象因素、光伏功率时间序列标记数据; H ——归一化前的矩阵; $\hat{G}_{:,m}$ ——气象因素、光伏功率时间序列的预测结果; L ——TrmBlock 的数量。

归一化层用于表示单个变量的时间序列,并将所有嵌入标记归一化为正态分布,其计算公式为:

$$\text{LayerNorm}(H) = \left\{ \frac{h_o - M(h_o)}{\sqrt{A(h_o)}} \right\}, o = \{1, 2, \dots, D\} \quad (15)$$

式中: h_o ——矩阵 H 的第 o 个序列; $M(h_o)$ ——均值函数,计算气象因素、光伏功率时间序列的均值; $A(h_o)$ ——方差函数,计算气象因素、光伏功率时间序列的方差; D ——矩阵 H 中带归一化的序列数。

前馈神经网络采用非线性映射学习时间序列,并对气象因素、光伏功率时间序列进行全局表示。通过自注意力机制揭示多个气象因素间的相关性,促进信息交换。多个自注意力机制构成的多头自注意力机制更能够提取光伏功率时间序列内部的相关信息。自注意力机制计算公式为:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (16)$$

式中: Q ——查询向量 query; K ——键向量 key; V ——值向量 value; softmax——数据归一化; QK^T ——查询向量和键向量的转置; d_k ——键向量的维度。

将 TrmBlock 模块处理后的结果映射在投影层中映得到预测结果。

2.2 XGBoost模型原理

XGBoost 算法是对提升决策树 (gradient boosting decision tree, GBDT) 算法的改进。XGBoost 是由多个基础的弱学习器组成一个具有高精度预测能力的模型,决策树通过不断迭代先前树的残差,使得预测结果趋近真实值。XGBoost 的目标函数公

式为:

$$F_{\text{obj}} = \sum_i l(y_i, \hat{y}_i) + \sum_e \Omega(f_e) \quad (17)$$

$$\Omega(f_e) = \mu T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (18)$$

式中: e ——第 e 棵树; y_i ——第 i 个样本的预测值; $l(y_i, \hat{y}_i)$ ——预测值和实际功率值的差; $\Omega(f_e)$ ——正则化项; μ, λ ——惩罚系数; T ——回归树的叶子数量; w_j ——叶子节点权重值。

迭代过程公式为:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_e(x_i) \quad (19)$$

$l(y_i, \hat{y}_i)$ 采用二阶泰勒展开,目标函数最终公式为:

$$F_{\text{obj}} = \sum_{i=1}^n \left[g_i f_e(x_i) + \frac{1}{2} h_i f_e^2(x_i) \right] + \Omega(f_e) \quad (20)$$

式中: g_i —— $l(y_i, \hat{y}_i^{(t-1)})$ 在 $\hat{y}_i^{(t-1)}$ 的一阶导数; h_i —— $l(y_i, \hat{y}_i^{(t-1)})$ 在 $\hat{y}_i^{(t-1)}$ 的二阶导数。

XGBoost 对损失函数进行二阶泰勒展开,在目标函数中引入正则化项,既有效约束 XGBoost 模型的复杂度,增强其泛化能力,又避免其在训练数据时过度拟合。

2.3 IGWO-iTransformer-XGBoost组合模型分析

改进灰狼优化算法 (IGWO) 是一种原启发的寻优算法,个体在狩猎过程中通过融合维度学习 (dimension learning based hunting, DLH) 的策略增强探索能力。本文采用 IGWO 算法对 iTransformer 和 XGBoost 进行超参数优化,从而找到 iTransformer 模型和 XGBoost 模型的最佳超参数模型,并通过组合权重值确定 IGWO-iTransformer-XGBoost 组合模型,该模型结合了 iTransformer 的时序特征提取能力和 XGBoost 快速拟合光伏发电时序数据的优势,通过 IGWO 算法优化模型权重,进一步提升预测精度,IGWO 算法优化组合模型如图 2 所示。具体步骤如下:

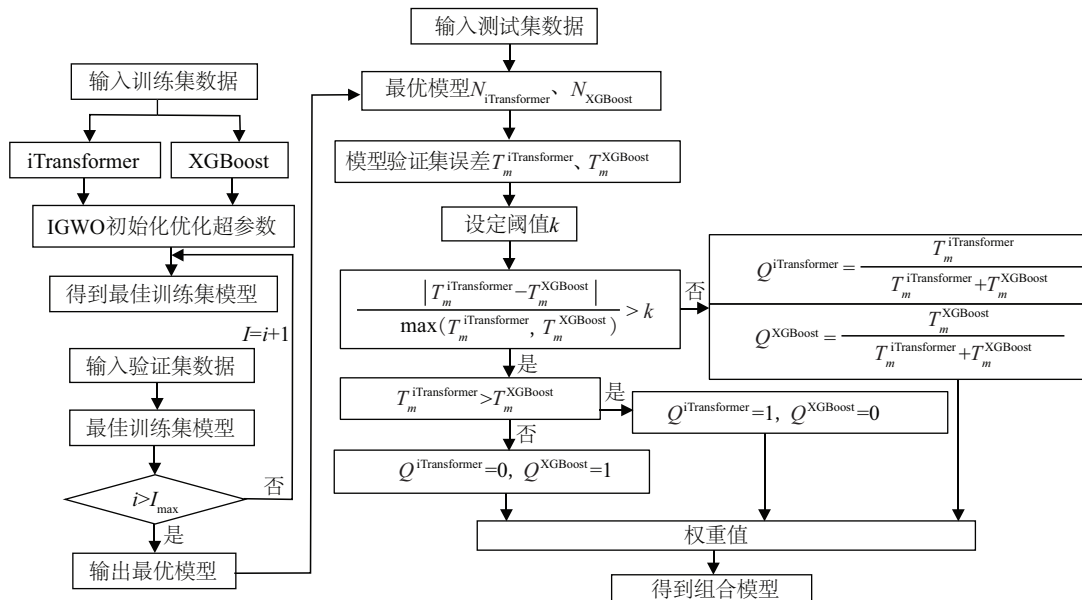


图2 IGWO优化组合模型

Fig. 2 IGWO optimized combination model

1)将预处理后的数据集以 8:1:1 的比例划分为训练集、验证集、测试集,初始化迭代次数、学习率和位置编码等超参数,用 IGWO 算法分别对 iTransformer 模型和 XGBoost 模型在训练集中进行训练。

2)在训练中分别找到 iTransformer 模型和 XGBoost 模型的最佳超参数模型 $N_{iTransformer}$ 和 $N_{XGBoost}$ 。

3)设验证集的特征为 X_m ,真实值为 Y_m ,经过验证集后得到两种模型预测值即 $Y_m^{iTransformer}$ 、 $Y_m^{XGBoost}$ 和两种模型间的误差,即 $T_m^{iTransformer}$ 、 $T_m^{XGBoost}$ 。

$$T_m^{iTransformer} = |Y_m^{iTransformer} - Y_m| \quad (21)$$

$$T_m^{XGBoost} = |Y_m^{XGBoost} - Y_m| \quad (22)$$

4)确定组合模型的权重:

$$Q^{iTransformer} = \frac{T_m^{iTransformer}}{T_m^{iTransformer} + T_m^{XGBoost}} \quad (23)$$

$$Q^{XGBoost} = \frac{T_m^{XGBoost}}{T_m^{iTransformer} + T_m^{XGBoost}} \quad (24)$$

若两种算法在验证集中的误差相差很大,取效果最好的模型作为最终模型。设定阈值 $k=0.1$,若满足 $\frac{|T_m^{iTransformer} - T_m^{XGBoost}|}{\max(T_m^{iTransformer}, T_m^{XGBoost})} > k$,则两种模型不进行结合,此时设置误差小的模型的权重为 1,另一种模型权重为 0。

5)最终组合模型的预测结果为:

$$\hat{Y}_m = Y_m^{iTransformer} \times Q^{iTransformer} + Y_m^{XGBoost} \times Q^{XGBoost} \quad (25)$$

2.4 模型误差评估指标

为反映分布式光伏集群功率预测精度,本文选用平均绝对百分比误差(mean absolute percentage error, MAPE)、平均绝对误差(mean absolute error, MAE)、均方误差(mean square error, MSE)、均方根误差(root mean square error, RMSE)和决定系数(R-square, R^2)对预测结果进行评估。其具体表达式分别为:

$$E_{MAE} = \frac{1}{p} \sum_{i=1}^p |s_i - \hat{s}_i| \quad (26)$$

$$E_{MAPE} = \frac{1}{p} \sum_{i=1}^p \left| \frac{s_i - \hat{s}_i}{s_i} \right| \quad (27)$$

$$E_{MSE} = \frac{1}{p} \sum_{i=1}^p (s_i - \hat{s}_i)^2 \quad (28)$$

$$E_{RMSE} = \sqrt{\frac{1}{p} \sum_{i=1}^p (s_i - \hat{s}_i)^2} \quad (29)$$

$$R^2 = 1 - \frac{\sum_{i=1}^p (\hat{s}_i - s_i)^2}{\sum_{i=1}^p (\bar{s}_i - s_i)^2} \quad (30)$$

式中: s_i 、 $\hat{s}_i$ 、 $\bar{s}_i$ ——第 i 个样本光伏功率的实际值、预测值和均值; p ——光伏功率样本总数。

2.5 基于K-中心点聚类的功率预测

本文提出基于改进曼哈顿距离的 K-中心点聚类 and IGWO-iTransformer-XGBoost 分布式光伏电站短期功率预测方法,该方法主要分为数据预处理、气象特征选择、光伏电站集群聚类划分和子集群功率预测 4 个部分,整体框架图如图 3 所示。

具体步骤为:

1)输入各光伏电站的历史数据,包含太阳总辐照度、直接法向辐照度、全球水平辐照度、温度、相对湿度和功率等数据,对缺失值进行识别后采用三次样条插值方法填补缺失值,并采用孤立森林检测法剔除异常值。

2)基于预处理后的数据构造气象因素和功率的数据矩阵 R ;通过变异系数法与秩和比法计算光伏输出功率与气象因素的 Z_{RSR} 值,从而筛选出光伏电站聚类算法和预测模型的气象输入特征。

3)通过各光伏电站的功率数据计算出光伏电站间的改进曼哈顿距离,并将其作为聚类划分指标引入 K-中心点集群聚类中。

4)采用内部有效评价指标自动设定聚类划分数,包含平均轮廓系数、Calinski-Harabaz 指数和 Davies-Bouldin 指数,建立基于改进曼哈顿距离的 K-中心点聚类算法,并通过该聚类算法对各光伏电站进行聚类划分。

5)各子集群使用整体法,采用 IGWO-iTransformer-XGBoost 组合模型对各子集群进行功率预测,并进行相关预测效果评价指标的对比分析,验证所提算法的聚类效果与所提模型的功率预测性能。

3 算法分析与验证

3.1 数据说明与预处理

本文选用数据集为可再生能源发电预测竞赛中使用的 8 个分布式光伏站点的 2020 年实际发电功率数据,采样时间间隔 15 min。所选 8 个太阳能发电站的发电量在 30~130 MW 之间,平均实际输出功率为 4.2~29.8 MW。由于光伏发电夜停昼发的特性,所以选取每日 08:00—18:30 的数据进行分析,共 125904 条数据。

在实际基于环境信息的能源发电功率预测过程中,可能由传感器故障、数据采集问题和数据记录问题等原因导致部分数据缺失,因此本文首先对缺失值进行识别后采用三次样条插值方法填补缺失值,然后采用孤立森林检测法剔除异常值,最终确定对 112743 条数据进行分析。

3.2 气象特征选择

本文使用变异系数法计算每个气象因素在光伏功率中的权重值并赋予在秩和比法中进行 Z_{RSR} 值计算,从而确定光伏集群聚类算法和功率预测模型的气象输入特征。由图 4 看

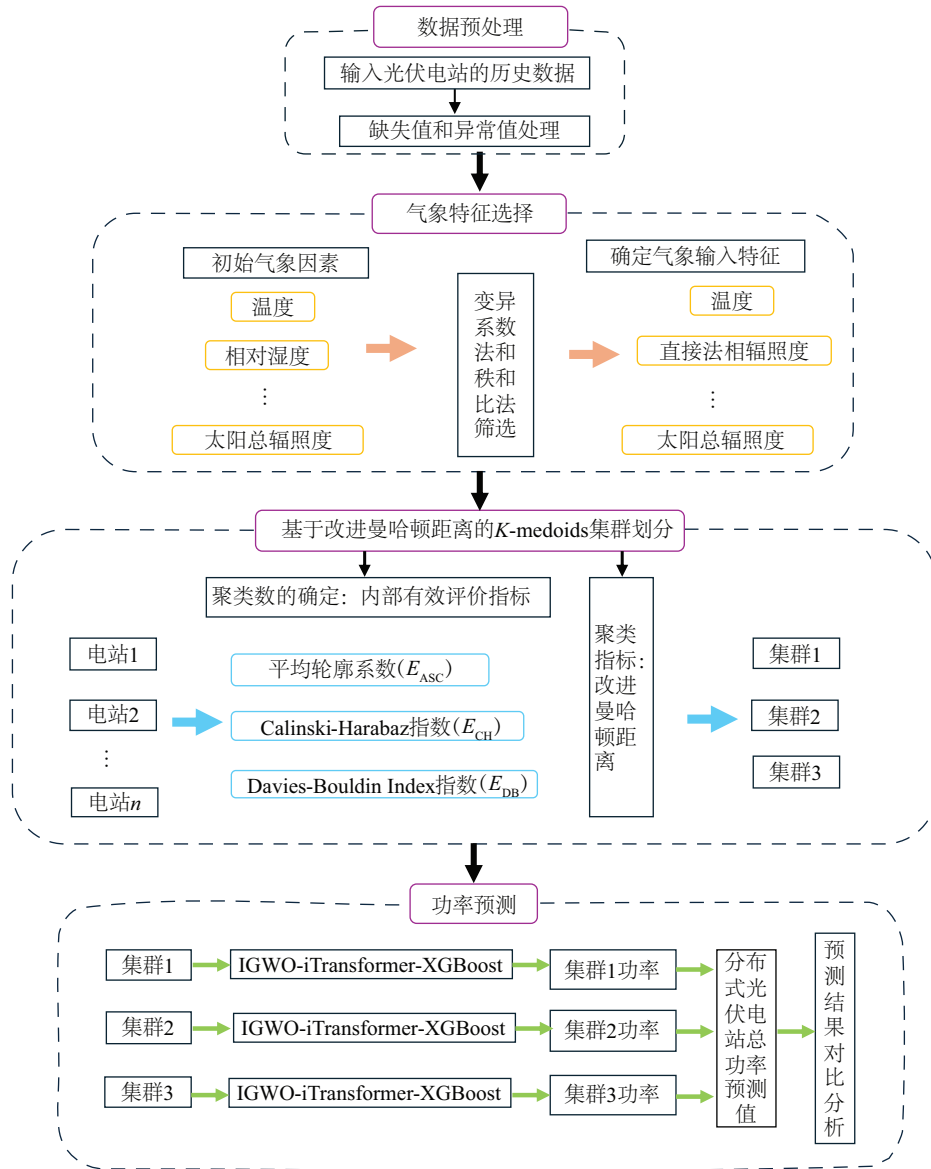


图3 基于K-中心点聚类的功率预测框架

Fig. 3 Power prediction framework based on K-medoids clustering

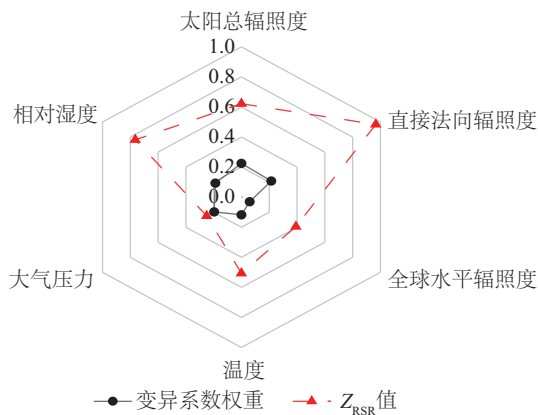


图4 气象因素与光伏功率的秩和比值

Fig. 4 Rank-sum ratio of meteorological factors and PV power

出大气压力的 Z_{RSR} 值最小,说明在这 6 项气象因素中,大气压力与光伏功率输出的相关性最小,本文选取 Z_{RSR} 值较大的 5 项气象因素作为影响光伏电站输出功率的气象特征,即太阳总辐照度、直接法向辐照度、全球水平辐照度、温度和相对湿度。

3.3 聚类划分

本文分别对基于改进曼哈顿距离的 K -中心点聚类、 K 均值聚类和层次聚类 3 种聚类算法采用内部有效评价指标包括平均轮廓系数(E_{ASC})、Calinski-Harabaz 指数(E_{CH})和 Davies-Bouldin 指数(E_{DB})进行对比分析,实现自动设置聚类划分数,搜索出最佳聚类划分数,并对聚类结果进行考察。图 5 为 3 种聚类方法针对不同聚类数时的 E_{ASC} 、 E_{CH} 和 E_{DB} 比较。可看出,采用基于改进曼哈顿距离的 K -中心点聚类方法在聚类数 $K=3$ 时, E_{ASC} 值最接近 1 说明簇内紧凑, E_{CH} 值最大表明簇间

与簇内距离比值越大, E_{DB} 值最小表示簇间距离越大, 此时聚类效果最佳, 因此本文选用基于改进曼哈顿距离的 K -中心点聚类方法将分布式光伏集群划分为 3 个子集群。

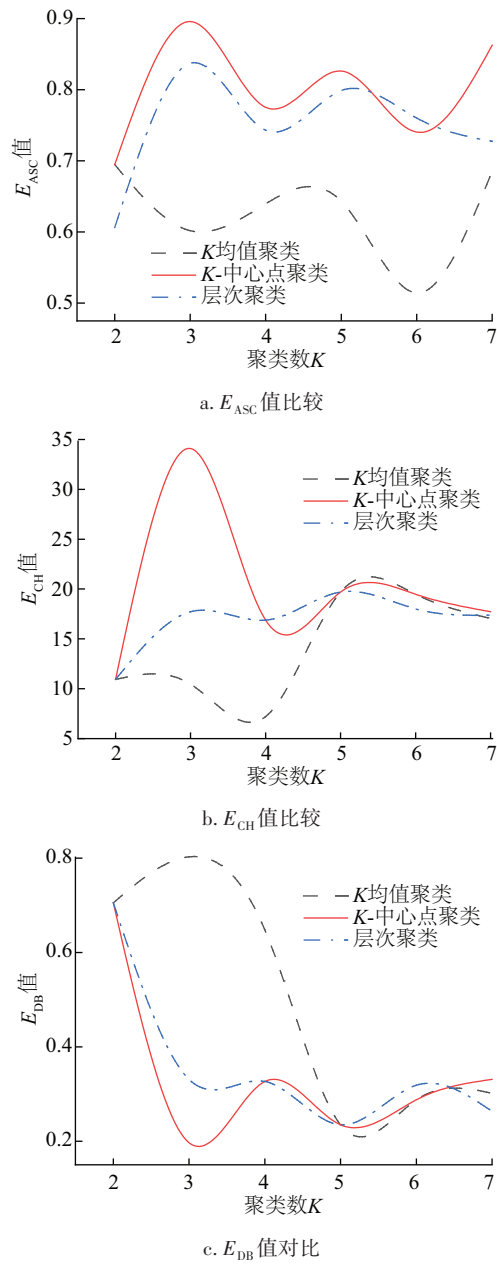


图5 不同聚类方法的内部有效评价指标比较
Fig. 5 Comparison of internal effective evaluation indexes of different clustering methods

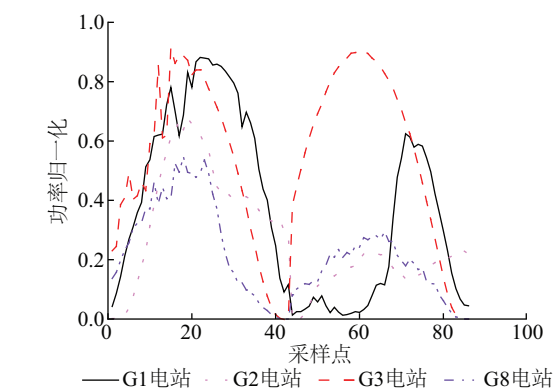
表 1 为聚类数 $K=3$ 时, 各光伏电站采用 K 均值聚类、层次聚类方法和 K -中心点聚类的具体集群划分结果。图 6 为对 K 均值聚类、层次聚类和 K -中心点聚类中的每个集群 1 进行功率归一化可视化。由图 6 对比分析, 采用基于改进曼哈顿距离的 K -中心点集群聚类算法, 更能保证将功率曲线变化趋势相似、功率特征相近的光伏电站聚集划分为同一集群, 减少了后续功率预测的误差, 从而提高预测精度。此处为验

证基于改进曼哈顿距离的 K -中心点算法划分分布式光伏集群的有效性, 通过选用在图 5 评价指标对比中聚类效果较佳的层次聚类算法与 K -中心点聚类算法分别进行集群划分, 将各子集群分别用本文所提 IGWO-iTransformer-XGBoost 组合模型作预测误差指标对比分析, 其预测误差指标百分数计算结果如图 7 所示。从图 7 可看出, 基于改进曼哈顿距离的 K -中心点算法的决定系数 R^2 最接近 1, 其余 4 项误差指标值均最小, 表明基于改进曼哈顿距离的 K -中心点算法的预测

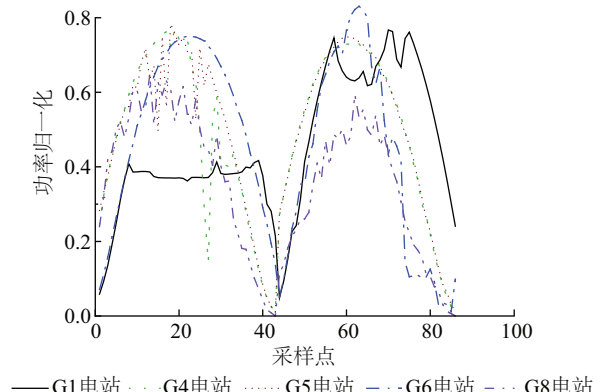
表 1 聚类划分结果

Table 1 Results of cluster division

聚类方法	子集群	所含电站
K均值	集群1	G1、G2、G3、G8
	集群2	G6、G7
	集群3	G4
K-中心点	集群1	G1、G3、G6、G7、G8
	集群2	G2
	集群3	G4、G5
层次聚类	集群1	G1、G4、G5、G6、G8
	集群2	G3、G7
	集群3	G2



a. K-means 聚类



b. 层次聚类

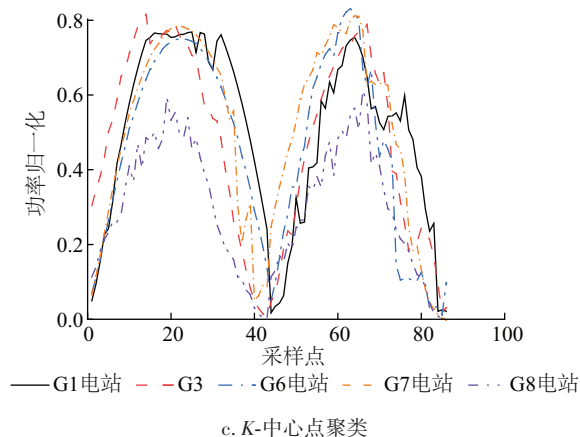


图6 3种聚类法集群1功率归一可视化

Fig. 6 Visualization of normalized power of cluster 1 for three clustering methods

性能优于基于层次聚类的预测性能,这是因为本文采用的基于改进曼哈顿距离的 K -中心点聚类算法通过对各光伏电站时间序列数据的动态特性进行特征提取,以及采用内部有效评价指标自动确定聚类数,避免了因人为主观设置划分数带来偏差的问题,从而取得较好的聚类结果。

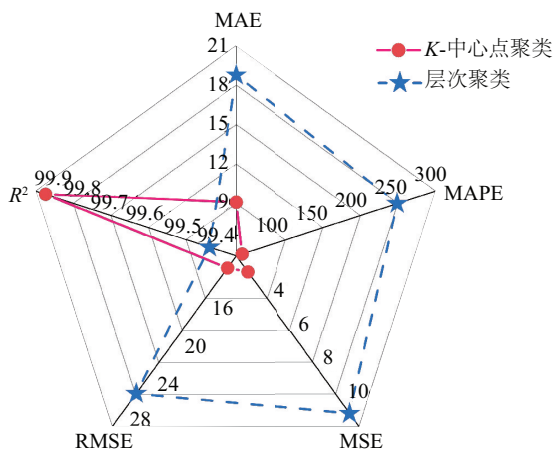


图7 两种聚类算法的预测误差指标百分数比较

Fig. 7 Comparison of prediction error indicator percentage for two clustering algorithms

图8为基于层次聚类算法与基于改进曼哈顿距离的 K -中心点聚类算法的预测值与真实值的差值比较。可看出在整体趋势上基于改进曼哈顿距离的 K -中心点聚类算法的预测值与真实值差值的波动比较平稳,变化趋势较为一致,而层次聚类的预测值与真实值差值的波动起伏剧烈,尤其在某些局部采样点的差值较大,表明基于层次聚类的预测效果较差,说明基于改进曼哈顿距离的 K -中心点聚类算法的预测模型更易较好地预测分布式光伏集群的功率输出。

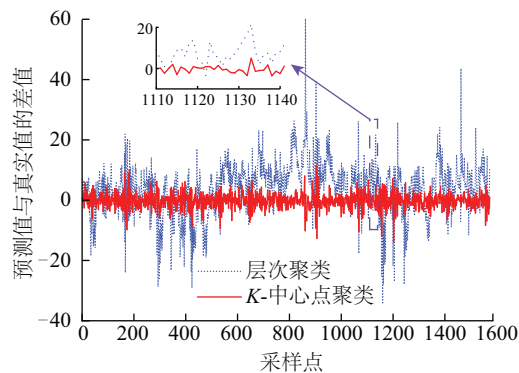


图8 两种聚类算法预测值与真实值差值比较

Fig. 8 Comparison of differences between predicted values and actual values for two clustering algorithms

3.4 实验方案设置

为验证所提模型在光伏功率预测作业中的预测效果是否有效,设置两个实验方案,与本文所提方案的预测效果进行对比分析。

方案1:为验证聚类算法在分布式光伏集群功率预测中的优越性,采用累加法、整体法与本文所提聚类算法进行对比。

方案2:对采用基于改进曼哈顿距离的 K -中心点算法划分后的子集群分别用本文所提 IGWO-iTransformer-XGBoost 组合模型和现在常用的时间卷积网络(temporal convolutional network, TCN)^[21]、门控循环单元网络(GRU)和双向长短期记忆(BiLSTM)网络模型进行功率预测,并完成预测指标分析。

3.4.1 实验方案1结果分析

图9为不同预测方法对分布式光伏电站集群进行功率预测的结果比较。可看出,与真实值相比累加法的预测效果偏差最大,而整体法和集群聚类法的效果更好,而从局部图中可明显看出集群聚类法的预测结果与实际值曲线的拟合程度最高。使用累加法的预测结果易受到单个光伏电站预测精度的影响,产生的累积误差会导致预测精度降低。整体

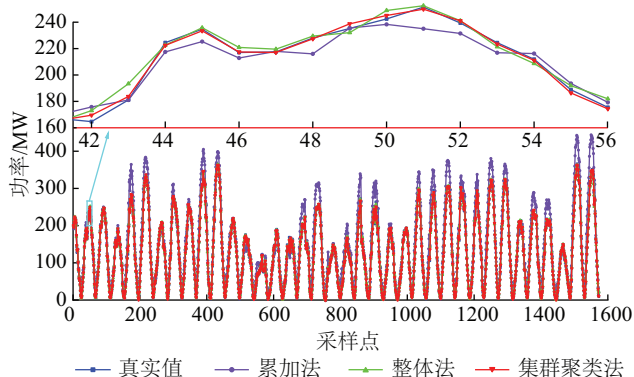


图9 不同方法预测值比较

Fig. 9 Comparison of predicted values from different methods

法更适合对单个光伏电站进行建模和预测,无法较好地捕捉到分布式光伏电站集群的局部特征,且整体法使用单一模型进行功率预测会导致该预测模型的鲁棒性和泛化能力受到限制。集群聚类法则通过聚类提取各簇特征将相关性较强的电站划分为同一子集群,能更准确地捕捉到该簇内电站的共同特征和规律,有效消除了分布式光伏电站间的冗余干扰,减少预测误差,提高预测模型的灵活性。

通过对比3种预测方法预测误差的相关指标,从而验证其预测性能。计算结果如表2所示。由表2得出,本文提出的基于改进曼哈顿距离的K-中心点聚类和IGWO-iTransformer-XGBoost预测方法的MAE、MAPE、MSE和RMSE在这3种集群预测方法中均最小。相比于累加法和整体法,本文所提方法的MAE分别平均降低12.04%和0.03%;MAPE分别平均降低38.35%和23.45%;MSE分别平均降低5.95%和0.92%;RMSE分别平均降低14.8%和5.91%; R^2 分别平均提高2.56%和1.13%,验证了基于改进曼哈顿距离的K-中心点聚类算法的优越性和有效性。

表2 不同预测方法评价指标

集群预测方法	MAE	MAPE	MSE	RMSE	R^2
累加法	21.14	95.62	7.89	28.08	97.31
整体法	9.13	80.72	2.86	19.19	98.47
集群聚类法	9.10	57.27	1.94	13.28	99.87

3.4.2 实验方案2结果分析

对比IGWO-iTransformer-XGBoost、BiLSTM、GRU和TCN预测模型的误差评估指标,并进行误差和精确度分析,从而验证本文模型的预测精度。计算结果如表3所示。由基于改进曼哈顿距离的K-中心点聚类后的各子集群分别采用IGWO-iTransformer-XGBoost模型进行预测,与另3种预测模型相比,其MAE分别平均降低25.18%、34.8%和34.86%;MAPE分别平均降低135.88%、319.47%和129.44%;MSE分别平均降低25.38%、33.85%和36.53%;RMSE分别平均降低37.59%、46.13%和48.31%; R^2 分别平均提高1.81%、2.5%和

表3 不同预测模型误差指标

组合模型	MAE	MAPE	MSE	RMSE	R^2
IGWO-iTransformer-XGBoost	9.10	57.27	1.94	13.28	99.87
BiLSTM	34.28	193.15	27.32	50.87	98.06
TCN	43.90	376.74	35.79	59.41	97.37
GRU	43.96	186.71	38.47	61.59	97.25

2.62%,证明经过IGWO优化后的iTransformer-XGBoost组合模型的精确度更高,以及该模型进行超参数优化和全局寻优的有效性。

图10为采用不同预测模型对分布式光伏集群进行功率预测的预测结果与实际功率的整体曲线对比和局部曲线对比。可看出,使用本文所提预测方法的预测结果曲线更贴近于实际功率曲线,更能追踪实际功率的走向,在各拐点处也与实际功率值更加吻合。与BiLSTM、GRU、TCN模型相比,XGBoost能够快速对光伏发电功率数据进行较好地拟合,iTransformer能有效对光伏集群功率数据进行特征提取,具有较强的泛化能力,再通过IGWO对iTransformer和XGBoost进行权重组合,以此充分利用两种模型的优势,这种方法有效提高了预测结果的精度。实验结果证明采用基于改进曼哈顿距离的K-中心点聚类和IGWO-iTransformer-XGBoost预测模型进行分布式光伏电站集群功率预测可提高预测精度。

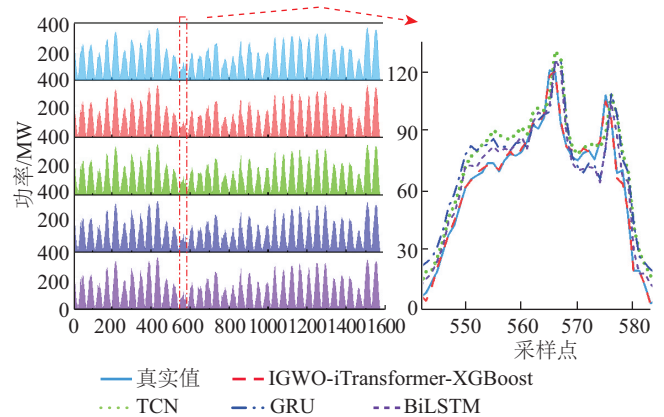


图10 不同模型预测结果对比

Fig. 10 Comparison of prediction results from different models

4 结 论

本文同时考虑各光伏电站的地理位置和气象因素对光伏电站聚类划分的影响,提出一种基于改进曼哈顿距离的K-中心点聚类方法,得到以下主要结论:

1)有效的气象特征提取对聚类结果至关重要,结合变异系数法和秩和比法识别出与发电功率相关性较强的气象特征,确保聚类的有效性。

2)采用改进曼哈顿距离进行各光伏电站间的相似性度量,更能反映各光伏电站时序数据的动态特性以及发电功率曲线轨迹的波动趋势,有效提高聚类质量。

3)采用3种内部有效指标自动设定聚类划分数,减少了人为主观设置聚类数的不确定性与偏差,保证了聚类结果的准确性。

4)与本文中聚类效果较佳的层次聚类相比,采用基于改进曼哈顿距离的K-中心点算法的各预测误差指标均更优,聚类结果更为准确。

5)与文中不采用聚类算法的累加法和整体法相比,

MAE、MAPE、MSE 和 RMSE 最大分别降低 12.04%、38.35%、5.95% 和 14.8%, R^2 最大提升 2.56%, 进一步说明了本文所提聚类算法的有效性。

[参考文献]

- [1] 中国电力网. 中国光伏产业发展路线图(2023—2024 年) [EB/OL]. (2024-2-28) [2024-12-26]. https://www.chinapv.org.cn/Industry/resource_1380.html. China Power Net. China PV Industry Development Roadmap (2023—2024) [EB/OL]. (2024-2-28) [2024-12-26]. https://www.chinapv.org.cn/Industry/resource_1380.html.
- [2] 张扬科, 李刚, 李秀峰. 基于典型代表电站和改进 SVM 的区域光伏功率短期预测方法[J]. 电力自动化设备, 2021, 41(11): 205-210.
ZHANG Y K, LI G, LI X F. Short-term forecasting method for regional photovoltaic power based on typical representative power stations and improved SVM [J]. Electric power automation equipment, 2021, 41(11): 205-210.
- [3] 王晓霞, 俞敏, 霍泽健, 等. 基于近邻传播聚类与 LSTNet 的分布式光伏电站群短期功率预测[J]. 电力系统自动化, 2023, 47(6): 133-141.
WANG X X, YU M, HUO Z J, et al. Short-term power forecasting of distributed photovoltaic station clusters based on affinity propagation clustering and long short-term time-series network [J]. Automation of electric power systems, 2023, 47(6): 133-141.
- [4] 杨锡运, 刘晗, 陈文进, 等. 基于层次聚类分场景的光伏汇聚趋势量化方法[J]. 太阳能学报, 2024, 45(3): 496-505.
YANG X Y, LIU H, CHEN W J, et al. Photovoltaic convergence trend quantification method based on hierarchical clustering and scenarios [J]. Acta energiae solaris sinica, 2024, 45(3): 496-505.
- [5] 李培强, 李欣然, 陈辉华, 等. 基于模糊聚类的电力负荷特性的分类与综合[J]. 中国电机工程学报, 2005, 25(24): 73-78.
LI P Q, LI X R, CHEN H H, et al. The characteristics classification and synthesis of power load based on fuzzy clustering [J]. Proceedings of the CSEE, 2005, 25(24): 73-78.
- [6] 卢俊杰, 蔡涛, 郎建勋, 等. 基于集群划分的光伏电站集群发电功率短期预测方法[J]. 高电压技术, 2022, 48(5): 1943-1951.
LU J J, CAI T, LANG J X, et al. Short-term power output forecasting of clustered photovoltaic solar plants based on cluster partition [J]. High voltage engineering, 2022, 48(5): 1943-1951.
- [7] 叶林, 施媛媛, 王启亨, 等. 面向暂态分析的分布式光伏集群多步分群与等值建模[J]. 电力系统自动化, 2023, 47(14): 72-81.
YE L, SHI Y Y, WANG Q H, et al. Multi-step grouping and equivalent modeling of distributed photovoltaic clusters for transient analysis [J]. Automation of electric power systems, 2023, 47(14): 72-81.
- [8] 程启明, 张强, 程尹曼, 等. 基于密度峰值层次聚类的短期光伏功率预测模型[J]. 高电压技术, 2017, 43(4): 1214-1222.
CHENG Q M, ZHANG Q, CHENG Y M, et al. Short-term photovoltaic power prediction model based on hierarchical clustering of density peaks algorithm [J]. High voltage engineering, 2017, 43(4): 1214-1222.
- [9] 张晓珂, 张辉, 戴小然, 等. 基于层次聚类和 BiLSTM 的光伏短期功率预测模型[J]. 智慧电力, 2024, 52(9): 41-48.
ZHANG X K, ZHANG H, DAI X R, et al. Photovoltaic short-term power forecasting model based on hierarchical clustering & BiLSTM [J]. Smart power, 2024, 52(9): 41-48.
- [10] 吴汉斌, 时珉, 郑焕坤, 等. 基于 EEMD-ALOCO-SVM 模型的光伏功率短期区间预测[J]. 太阳能学报, 2023, 44(11): 64-71.
WU H B, SHI M, ZHENG H K, et al. Short-term interval prediction of photovoltaic power based on EEMD-ALOCO-SVM model [J]. Acta energiae solaris sinica, 2023, 44(11): 64-71.
- [11] 陈庆明, 廖鸿飞, 孙颖楷, 等. 基于 GWO-GRU 的光伏发电功率预测[J]. 太阳能学报, 2024, 45(7): 438-444.
CHEN Q M, LIAO H F, SUN Y K, et al. Photovoltaic power prediction model based on GWO-GRU [J]. Acta energiae solaris sinica, 2024, 45(7): 438-444.
- [12] 姜建国, 杨效岩, 毕洪波. 基于 VMD-FE-CNN-BiLSTM 的短期光伏发电功率预测[J]. 太阳能学报, 2024, 45(7): 462-473.
JIANG J G, YANG X Y, BI H B. Photovoltaic power forecasting method based on VMD-FE-CNN-BiLSTM [J]. Acta energiae solaris sinica, 2024, 45(7): 462-473.
- [13] ZAMANI JOHARESTANI M, CAO C X, NI X L, et al. PM_{2.5} prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data [J]. Atmosphere, 2019, 10(7): 373.
- [14] YANG T. Sales prediction of walmart sales based on OLS, random forest, and XGBoost models [J]. Highlights in science, engineering and technology, 2023, 49: 244-249.
- [15] 牛甲俊, 张薇, 许达明. 基于 FCM 和 iTransformer-TCN 的短期风电集群功率预测[J]. 东北电力大学学报, 2024, 44(1): 9-16.
NIU J J, ZHANG W, XU D M. Short-term power

- prediction of wind power cluster based on FCM and ITransformer-TCN[J]. Journal of Northeast Electric Power University, 2024, 44(1): 9-16.
- [16] 阮楚婷, 冯启明, 赵劲民, 等. 基于变异系数法和耦合协调度模型的广西三级公立综合医院高质量发展状况研究[J]. 现代预防医学, 2024, 51(9): 1643-1648.
- RUAN C T, FENG Q M, ZHAO J M, et al. Study on high quality development of tertiary public general hospitals in Guangxi based on coefficient of variation method and coupling coordination degree model [J]. Modern preventive medicine, 2024, 51(9): 1643-1648.
- [17] LIBERTI L, LAVER C, MACULAN N, et al. Euclidean distance geometry and applications [J]. SIAM review, 2014, 56(1): 3-64.
- [18] SUN Y J, LI S H, WANG Y L, et al. Fault diagnosis of rolling bearing based on empirical mode decomposition and improved Manhattan distance in symmetrized dot pattern image [J]. Mechanical systems and signal processing, 2021, 159: 107817.
- [19] 李姣, 王秋萍, 戴芳. 基于改进HHO与K-Medoids的混合聚类算法[J]. 西安理工大学学报, 2022, 38(3): 410-420.
- LI J, WANG Q P, DAI F. Hybrid clustering algorithm based on improved HHO and K-Medoids[J]. Journal of Xi'an University of Technology, 2022, 38(3): 410-420.
- [20] LIU Y, HU T G, ZHANG H R, et al. iTransformer: inverted transformers are effective for time series forecasting [EB/OL]. 2023: arXiv: 2310.06625. <https://arxiv.org/abs/2310.06625>.
- [21] 邢晨, 张照贝. 基于改进时间卷积网络的短期光伏出力概率预测方法[J]. 太阳能学报, 2023, 44(2): 373-380.
- XING C, ZHANG Z B. Short-term probabilistic forecasting method of photovoltaic output power based on improved temporal convolutional network [J]. Acta energiae solaris sinica, 2023, 44(2): 373-380.

DISTRIBUTED PHOTOVOLTAIC CLUSTER METHOD AND SHORT-TERM POWER PREDICTION BASED ON K-MEDOIDS CLUSTERING

Jiang Yaxue, Qian Jing, He Haocheng, Zhang Haoyan, Cao Lei, Mao Ximeng

(School of Metallurgical and Energy Engineering, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: For the purpose of achieving a scientifically rigorous and methodologically sound clustering partition of photovoltaic clusters, this study incorporates the synergistic effects of meteorological factors and geographical positioning on the clustering of photovoltaic power stations. Employing output characteristics as the basis for subgroup delineation, the paper introduces a K-medoids clustering algorithm based upon an enhanced Manhattan distance metric. The two pivotal dimensions in photovoltaic cluster analysis encompass the robust identification of meteorological features and the precise quantification of inter-station similarity. Accordingly, an integrated approach combining the coefficient of variation method with the rank-sum ratio method is utilized to derive meteorological input features, thereby augmenting the efficacy of feature discrimination. Moreover, the refined Manhattan distance is embedded within the K-medoids framework to comprehensively elucidate the dynamic attributes inherent in the dataset. Three internally validated clustering metrics are leveraged to optimize cluster cardinality, effectively mitigating the dispersion and systematic bias endemic to conventional partitioning approaches while enhancing overall clustering robustness. Subsequent predictive validation is performed based on the established clusters. To resolve the challenge of parameter calibration within the forecasting model, metaheuristic optimization techniques are integrated. Specifically, an improved grey wolf optimization (IGWO) algorithm is deployed to fine-tune the hyperparameters of both the iTransformer and extreme gradient boosting (XGBoost) architectures. The power forecasting results verify the operational efficacy and superior predictive accuracy of the proposed clustering methodology.

Keywords: photovoltaic power; cluster analysis; power forecasting; improved gray wolf optimization algorithm; iTransformer; extreme gradient boosting