DOI: 10.19783/j.cnki.pspc.231605

基于气候特征分析及改进 XGBoost 算法的中长期光伏电站发电量预测方法

李永飞，张 耀，林 帆，赵英杰，陈宇轩，赵寒亭，霍 巍

(陕西省智能电网重点实验室(西安交通大学电气工程学院)，陕西 西安 710049)

摘要：光伏发电在能源结构中的重要性不断凸显，而提高光伏发电量预测的准确性成为当前研究的关键问题。针对中长期光伏发电量预测问题，提出一个综合利用气候预测数据的中长期光伏发电量预测方法。首先，在基于气候预测数据的发电量预测框架中，根据气候预测数据特点和预测周期划分多重子模型以充分利用气候预测数据信息。其次，在进行数据预处理后，通过对气候特征衍生与交叉、特征筛选和选择，充分挖掘气候特征的高价值信息。然后，采取一种两重多阶段超参数寻优策略，对极端梯度增强(extreme gradient boosting, XGBoost)超参数进行调整以优化预测模型。最后，在真实光伏发电量数据上，以 MAPE 为标准评估预测水平，验证所提中长期光伏发电量预测方法的有效性。相关实验结果表明该方法可以有效提高光伏发电量预测精度。

关键词：气候预测数据；XGBoost；中长期预测；光伏发电量预测；特征工程

Medium- and long-term power generation forecast based on climate characterisation and an improved XGBoost algorithm for photovoltaic power plants

LI Yongfei, ZHANG Yao, LIN Fan, ZHAO Yingjie, CHEN Yuxuan, ZHAO Hanting, HUO Wei

(Shaanxi Key Laboratory of Smart Grid (School of Electrical Engineering, Xi'an Jiaotong University), Xi'an 710049, China)

Abstract: The importance of photovoltaic (PV) power in the energy structure is constantly highlighted, and improving the accuracy of PV power prediction has become a key issue in current research. To address the PV prediction problem, a medium- and long-term PV power generation prediction method using climate prediction data is proposed. First, multiple sub-models are divided according to the characteristics of climate prediction data and prediction period to make full use of the data. After data pre-processing, the high-value information of climate features is fully exploited through the derivation and crossover and selection of climate features. A two-fold multi-stage hyper-parameter optimization strategy is adopted to optimize the prediction model by adjusting the XGBoost hyper-parameters. Using real photovoltaic generation data, the prediction level is evaluated by MAPE, and the effectiveness of the proposed medium- and long-term PV power generation prediction method is verified by experiment. The results show that the method can effectively improve the prediction accuracy of PV power generation.

This work is supported by the National Key Research and Development Program of China (No. 2022YFB2403500).

Key words: climate prediction data; XGBoost; medium- and long-term forecasts; photovoltaic power generation forecasts; feature engineering

0 引言

光伏作为一种具有广阔应用前景的清洁能源，已成为能源转型的热门方向，并得到快速发展。然而，光伏发电固有的间歇波动特性对电力系统安全稳定运行提出了巨大挑战。间歇波动性源于多种因素，如太阳辐射强度和环境条件的变化，开发精确的光伏发电预测模型是有效应对这一难题的关键策略[1]。相对于短期预测，中长期预测面临长时间尺度下天气预报精度低、历史数据样本有限等问题[2]，但中长期预测的准确性对于电网规划优化、调度管理、提升消纳能力至关重要，是确保可再生能源大

基金项目：国家重点研发计划项目资助(2022YFB2403500)

范围实现互补调度的关键因素[3]。因此，亟需结合已有的发电量数据与气候数据，构建反映光伏发电中长期变化规律的模型，探究适用于中长期尺度的精准预测方法。

光伏发电量预测方法根据其内在逻辑可以分为物理方法、统计方法和人工智能方法[4]。物理方法作为最早的预测手段，通过整合辐照度、温度、湿度等关键气象参数，结合卫星云图与光伏电站实测数据，建立物理模型以计算光伏发电量[5]。此类方法无需依赖历史数据，但要求对光伏电站周边气象环境进行详尽且精确的监测，且对数值天气预报的依赖性较高。统计方法则利用历史时间序列数据，依据既往观测趋势预测未来光伏发电量[6]，更适合于短期与超短期预测场景。随着人工智能技术的不断进步，基于机器学习的光伏预测模型比传统的物理方法和统计方法更具有竞争力。机器学习方法可以从复杂多变的高维数据中提取影响太阳辐照量的关键特征[7]，进而对光伏发电量做到更加精准的预测。

在物理方法方面：文献[8]以逐时太阳辐射度数据和NWP数据作为输入对光电转换模型进行建模，应用卡尔曼滤波模型的动态方程对误差进行矫正来预测未来功率；文献[9]以天空图像或卫星图像为依据，构建模型以追踪云团运动趋势，并预测随后的光伏功率输出。然而这些物理方法通常依赖于详尽准确的物理参数和计算负担较大的数学模型，对气象数据质量也要求较高。

在统计方法方面：文献[10]提出一种基于外生因素及季节性的差分自回归移动平均模型，用于短期光伏发电功率预测；文献[11]设计了一种基于历史功率和天气分类的光伏功率预测模型，通过K-means 聚类和抛物线拟合进行分层建模，并分析残差分布以改进预测精度。文献[12]通过对光伏面板总发电量的季节性与非季节性变化研究，采用时间序列模型对光伏日发电量进行预测。但统计方法倚重历史数据，而且往往忽略气象数据与光伏发电量之间的高维非线性关系，导致在中长期尺度上的拟合效果欠佳。

在中长期预测场景下，机器学习方法具有对数据中非线性关系和多种特征间相互作用的捕捉能力、对数据结构更强的适应能力和模型更好的泛化能力[13]。例如：文献[14]利用反向传播的人工神经网络(artificial neural network, ANN)建立光伏发电预测模型，以气溶胶指数和温度、湿度、风速等气象参数作为输入，考虑不同天气条件的影响，但其计算量较大且主要应用于短期预测；文献[15]采用长短期记忆网络模型，利用通过插值模型得到的气象信息对光伏发电量进行预测，但在中长期尺度上的预测精度较低。

鉴于中长期光伏发电量预测面临的挑战以及现有方法存在的局限性，本文提出了一种基于气候特征分析及改进极端梯度增强(extreme gradient boosting, XGBoost)算法的中长期光伏电站发电量预测方法。以 XGBoost 模型为基础预测模型，考虑气候预测提前时间的影响，根据预测周期对数据进行合理划分，从而有效利用气候预测数据。通过特征工程挖掘气候预测数据，采用前向搜索算法筛选最优气候特征，产生光伏发电量预测结果。针对 XGBoost 超参数，本研究采用两重多阶段策略进行精细的超参数优化，旨在找到最有效的参数组合，从而最大限度地降低过拟合风险和预测误差。

1　基于气候特征的光伏发电量预测框架

1.1　气候预测数据特点

气候预测数据对于中长期光伏发电量的准确预测至关重要。此类数据涵盖了温度、湿度、风速、辐射通量等多种气象要素，反映了大气环境在一定时空范围内的变化过程[16]。

气象预报机构通过先进的观测设备、卫星技术和气象模型，系统地采集并生成各类气象数据，其时间尺度广泛，从小时级至数周不等。尤为重要的是，预报提前时间对气候预测数据质量的影响。气象数据质量通常在不同时间尺度上呈现显著差异：对于短期预测(如小时级别)，气象数据具有较高的时空分辨率，能更精确地反映近期气象变化；随着预测提前时间的增长，数据的可靠性和准确性显著下降，原因在于气象预报数据基于大气环流控制方程的数值积分生成[17]，而气象系统具有复杂的非线性动力学特征，长时间范围内的预测更易受外部因素扰动。

考虑到上述特性，本文计及了气候预测提前时间的影响，根据预测周期进行模型划分，从而充分利用气候预测数据。如此划分能使中长期预测模型更贴切地利用相应时间段内的气象数据。通过分阶段的建模，追求在不同预测时段实现最佳的预测结果，以适应不同预测周期下气象数据质量的变化。

以一份在北京时间 08：00 发布的包含未来 7×24 h 的数值天气预报文件为例，图 1 中 NWP1—NWP7 表示每天北京 08：00 发布的数值天气预测数据，将其按照时间顺序与发电量序列一一对应。

根据预测时间对气候预测数据质量的影响，第 1 天的预测精度最高，此后第 2~3 天和第 4~7 天次

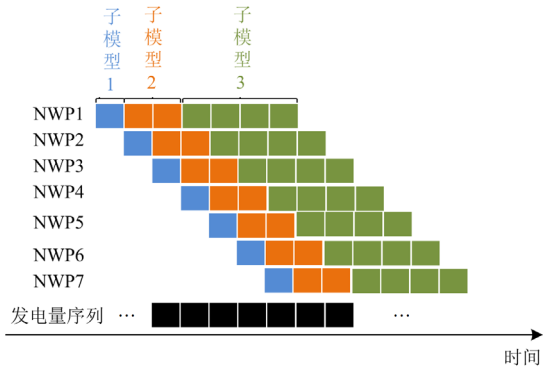

图 1 气候数值预测数据、子模型和光伏发电量序列之间的对应关系

Fig. 1 Correspondence between numerical climate prediction data, submodels, and photovoltaic power generation time series

之。所以将第 1 天的预测数据(蓝色方块)作为子模型 1 的数据集，用于建立子模型 1，第 2~3 天的数据(橙色方块)用于建立子模型 2，第 4~7 天的数据(绿色方块)用于建立子模型 3，而时间相重叠的部分选取与发布时间最近的数据。

1.2 长期光伏发电量预测整体框架

本文综合利用气候预测数据的周期特性，借助 XGBoost 算法的提升效果，构建了中长期光伏发电量预测模型。本文提出的中长期光伏发电量预测整体框架如图 2 所示。首先，根据气候预测数据的预测周期进行数据集划分，从而充分利用气候预测数据，将划分后的数据集应用于不同的子模型建立中，所有子模型均采用 XGBoost 算法。其次在各子模型建立阶段对数据进行预处理，包括异常值检测与修正以及缺失值插补，通过特征工程和特征优选对气候特征信息进行深度挖掘，通过超参数寻优提高预测模型的准确性。其中，特征工程和超参数寻优均采用多重时序交叉验证来评估预测模型的外样本性能。

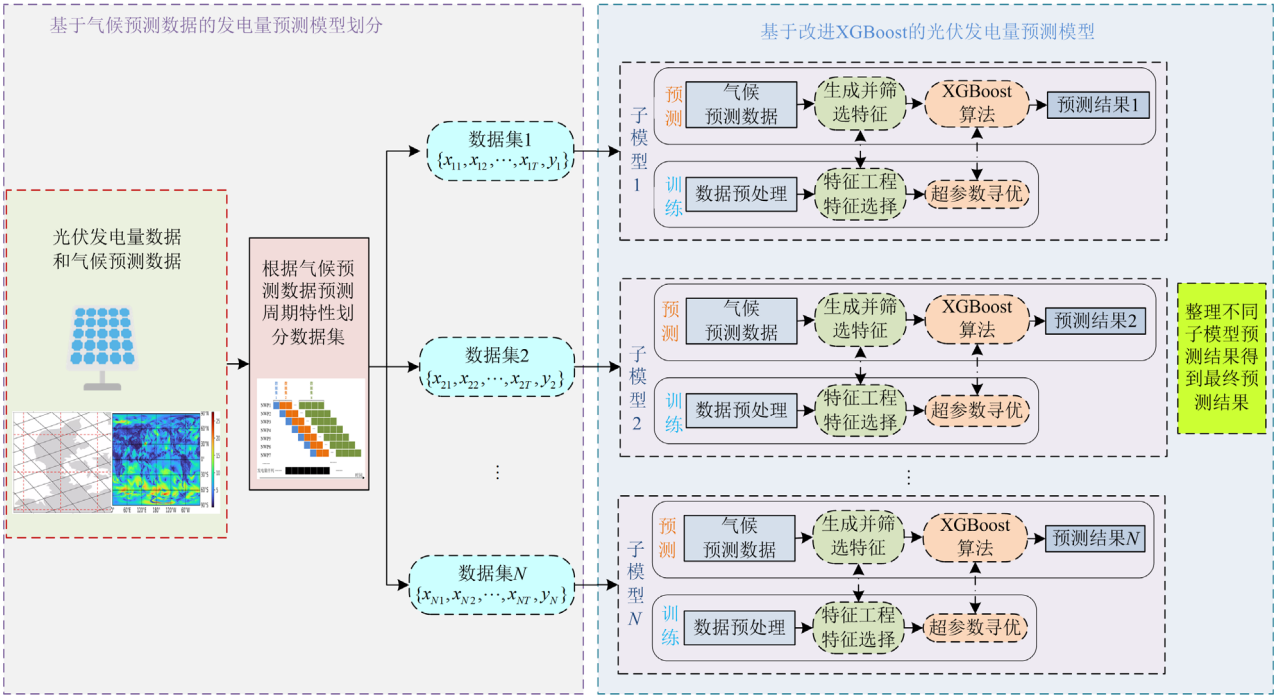

图 2 中长期光伏发电量预测整体框架

Fig. 2 Overall framework for medium- and long-term photovoltaic power generation forecast

2 基于 XGBoost 的光伏发电量预测模型

2.1 数据预处理方法

2.1.1 异常值处理方法

本文使用一种融合中值滤波、Z-score 检测与孤立森林算法的混合方法，用以识别预测模型训练数据中的异常值。在模型训练阶段，识别的异常值将被全部剔除，不参与后续预测模型的构建工作。

中值滤波通过使用邻近数据点的中位数替代每个数据点来实现降噪[18]。

$$m_i = \text{median}(p_{i-w},\cdots,p_i,\cdots,p_{i+w}) \tag{1}$$

式中：m_i 为中值滤波后的数值；p_i 为原始数据点；w 是窗口大小，表示用于计算中位数的邻近数据点的数量。

Z-score 方法可以通过统计分析来判断数据的异常程度[19]。

$$Z=\frac{p_i-\alpha_X}{\phi} \tag{2}$$

式中：Z 为 Z-score 值，表示原始数据点相对于整个数据集的偏离程度；α_X 为样本数据的平均值；ϕ 为样本数据的标准差。

而孤立森林算法则通过计算异常值分数来识别异常值[20]。

$$S(p,n)=2^{\frac{E(h(p))}{c(n)}} \tag{3}$$

$$c(n)=\begin{cases}2H(n-1)-2n(n-1) & n>2\\ 1 & n=2\\ 0 & 0<n<2\end{cases} \tag{4}$$

式中：$S(p,n)$ 为异常值分数；$c(n)$ 为 n 个样本构建树的平均路径长度；$E(h(p))$ 为样本 p 路径长度的平均值；$h(p)$ 为标准化样本 p 的路径长度；$H(n-1)$ 为调和数。

结合中值滤波、Z-score 和孤立森林算法，可以显著提高异常值检测的准确性和鲁棒性。中值滤波首先平滑数据并降低噪声，为后续的异常值检测创造更为稳定的环境。Z-score 方法通过统计分析，从全局角度判断数据点的异常程度。而孤立森林算法则通过构建个体决策树模型，从局部视角识别行为显著不同的异常值。通过综合运用这 3 种方法，能够更全面、有效地检测和处理异常值，从而确保数据分析结果的可靠性。

2.1.2 缺失值处理方法

本文通过构建 KD 树的方式，搜寻缺失值的最近邻数据，使用最近邻数据对缺失值进行填补。具体步骤如下：首先计算每个维度的方差，确定方差最大的维度为切分维度；其次以切分维度上的中位数为切分值，将数据集递归划分成左右子集，小于等于中位数的数据在左子集中，大于中位数的数据在右子集中；该过程持续递归，直至每个叶节点包含一个数据点，最后完成在 KD 树的构建。

构建 KD 树旨在寻找最近邻数据[21]，其寻找过程通过维度进行对比，从而快速定位最近邻数据。本文使用选取的 5 个最近邻数据作为备填充数据，对于天气类型使用投票法进行缺失值填充，即以备选填充数据中出现次数最多的类别作为该缺失点的分类标签，通过多数表决的方式确定缺失的分类属性；使用平均值对于气候特征数据中的连续变量和光伏发电量数据法进行缺失值填充，计算其算术平均值，将该平均值作为缺失点的填充值。

2.2 XGBoost 基本原理

2.2.1 XGBoost 集成算法

极端梯度增强[22]是一种基于决策树的集成机器学习模型。梯度提升算法是一种集成学习方法，通过迭代训练弱学习器(通常为决策树 CART)，每轮校正前一轮的残差，利用梯度下降最小化损失函数，同时引入学习率和正则化项控制模型复杂度，最终构建强大的集成模型。XGBoost 是梯度提升算法的一种优化实现，通过并行计算和特征分裂优化等技术，显著提高了训练效率和预测性能，被广泛应用于数据科学竞赛和实际问题[23]。

XGBoost 模型是基于梯度提升算法(gradient boosting decision tree, GBDT)的，梯度提升算法以决策树 CART 为基础进行分类和回归预测任务。

在 GBDT 中，按顺序构建一系列基础 CART，每个 CART 估计器都与训练过程中要调整的权重相关联，可以构建强大而稳健的集成。在回归预测任务中，一个样本的预测结果是通过对回归决策树上每个叶子节点的预测结果进行加权求和得到的。

$$\hat{y}_i=\sum_k^K \beta_k h_k(\boldsymbol{x}_i) \tag{5}$$

式中：$\hat{y}_i$ 表示预测值；K 表示树的总数量；β_k 为第 k 棵树的权重；$h_k(\boldsymbol{x}_i)$ 表示第 k 棵树的预测结果；$\boldsymbol{x}_i$ 表示第 i 个特征样本所对应的特征向量。

在 XGBoost 中，每个叶子节点都有一个权重，也被称为叶子权重，该权重就是该叶子节点对应数据的预测值。叶子权重表示在该树上所有样本在此叶子节点处的回归取值。当存在多棵树时，将所有树的预测结果加权求和得到最终的预测结果。

$$\hat{y}_i=\sum_k^K f_k(\boldsymbol{x}_i) \tag{6}$$

式中：f_k 为第 k 棵树的预测函数，$f_k(\boldsymbol{x}_i)$ 表示第 k 棵树对 $\boldsymbol{x}_i$ 的预测结果。

2.2.2 XGBoost 模型的损失函数

XGBoost 模型通过迭代的方式，结合弱学习器的组合和复杂度控制，能够发现目标模型与观测特征之间的复杂非线性统计关系，并在训练过程中逐步优化模型以提高性能。结合式(6)中的预测结果，XGBoost 模型的目标函数为

$$O_{\mathrm{bj}}=\sum_{i=1}^M l(y_i,\hat{y}_i)+\sum_{k=1}^K \Omega(f_k) \tag{7}$$

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{k=1}^K \omega_k^2 \tag{8}$$

式中：M 表示数据集中的样本总量；y_i 表示真实值；$l(y_i,\hat{y}_i)$ 为损失函数，衡量预测值与真实值之间的差

异；$\Omega(f_k)$代表模型的复杂度，即正则化项，有助于防止模型过拟合；T为叶节点数；ω_k为叶节点权值；γ和λ是预先给定的超参数，分别控制叶节点的数量和分数。

XGBoost 模型通过向前推进的方式逐步增强模型的加性训练过程。它在每一轮迭代中均训练一个新的模型，将其添加到先前模型的集合中以逐步减小损失函数。

$$\hat{y}_i^q = \sum_q^Q f_k(\boldsymbol{x}_i) = \sum_q^{Q-1} f_k(\boldsymbol{x}_i) + f_q(\boldsymbol{x}_i) \tag{9}$$

$$\sum_k^K \Omega(f_k) = \sum_q^{Q-1} \Omega(f_k) + \Omega(f_q) \tag{10}$$

$$O_{\text{bj}} = \sum_{i=1}^M \left[f_q(\boldsymbol{x}_i) g_i + \frac{1}{2} f_q(\boldsymbol{x}_i)^2 h_i \right] + \Omega(f_q) \tag{11}$$

式中：Q为迭代次数；y_i^q为第q次迭代的真实值；$\hat{y}_i^q$为第q次迭代的预测值；f_q为在第q次迭代中最优的树；g_i和h_i分别为在损失函数$l(y_i^q, \hat{y}_i^{(q-1)})$上对$y_i^{(q-1)}$所求的一阶导数和二阶导数，$l(y_i^q, \hat{y}_i^{(q-1)})$是通过对$f_q(\boldsymbol{x}_i)$的泰勒展开得来的。

XGBoost 模型是通过使用梯度提升的方法，并结合一阶导数和二阶导数的信息来逐步减小损失函数的数值，使得损失函数的优化可以转化为近似求解二次函数最小值的过程。

2.3 气候特征处理方法

2.3.1 气候特征筛选

气候预测数据包含温度、湿度、风速、辐射等多种气象变量数据，本文将这些气象变量称为基本特征，如表 1 所示。

表 1 气候预测数据的基本气象特征

Table 1 Fundamental meteorological attributes of meteorological forecast data

类型	基本特征
风特征	10 m、30 m、70 m、100 m 高处的风速、风向
辐射量特征	长波辐射通量，短波辐射通量，潜热通量，感热通量
温度特征	2 m、30 m、高处气温
降水特征	总降水、大尺度降水、对流降水
其他气象特征	海平面气压、云量、2 m 高处的相对湿度
时间特征	当天所在的年、月、周

对于所研究的气候预测数据和光伏发电量数据集，本文使用皮尔逊(Pearson)相关系数衡量温度、湿度、辐照度、风速、风向等气象特征变量与光伏发电量之间的相关性。皮尔逊相关系数可用于筛选出与发电量关联度较高的特征作为预测模型的基础输入特征，其计算公式为

$$\rho(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{12}$$

式中：$E(\cdot)$表示数学期望；μ_X和μ_Y分别表示气象特征变量X与发电量序列Y的均值；σ_X和σ_Y分别表示气象特征变量X与发电量序列Y的标准差。

2.3.2 气候特征的衍生与交叉

在筛选得到的预选基础输入特征的基础上，根据以下衍生特征与交叉特征生成规则，生成丰富的气象特征集作为光伏发电量预测模型的候选特征。

衍生特征规则如表 2 所示。此类特征被定义为基本特征经由单变量函数变换，本质上是通过对现有特征进行变换、组合或处理而产生的新特征。衍生特征旨在提供更多关于光伏发电量的气象特征信息，改善预测模型的性能和适应性。

表 2 气候特征的衍生特征

Table 2 Derivative characteristics of climatic features

特征类型	衍生特征生成规则
衍生特征	周期为 1 年、1 月、1 天的傅里叶谐波项(包括正弦项、余弦项) 日辐射比值、温度湿度指数

交叉特征规则如表 3 所示。交叉特征是通过组合两个或更多原始特征生成新的特征。组合方式可以是简单的乘法、加法，也可以是更复杂的函数关系。引入交叉特征有助于模型更深入地捕捉原始特征之间的相互影响，提升在光伏发电量预测上的准确度与适应复杂环境变化的能力。

表 3 气候特征的交叉特征

Table 3 Cross-features of climate characteristics

特征类型	交叉特征生成规则
交叉特征	不同高度处风速差值及其绝对值 云量与短波辐射乘积 风冷指数$\left(10\sqrt{\omega_{10\text{m}}} - \omega_{10\text{m}} + 10.5\right)(33 + 273.5 - T_{2\text{m}})$ ($\omega_{10\text{m}}$为10 m风速，$T_{2\text{m}}$为2 m温度)
	辐射量特征的滞后项(−1,−2,−3)、超前项(+1,+2,+3)以及长度为 3、9 的后向、中央、前向时窗内的均值、方差
	温度特征、降水特征、其他气象特征、云量−短波辐射乘积、云量−2 m 相对湿度乘积的滞后项(−1,−2,−3)、超前项(+1,+2,+3)
	海平面气压、云量的一阶后向差分

2.3.3 气候特征优选

在建立预测模型时，引入衍生特征与交叉特征虽有助于揭示复杂气象因素与发电量之间的深层次关系，但可能导致特征维度显著增加，且并非所有新增特征均对模型的性能有实质性的提升。因此，

特征优选成为避免过拟合并确保模型稳健性的关键步骤。

本文采用前向搜索方法进行特征选择，但由于待选特征数量较多，前向搜索的时间成本较高。为解决这一问题，本文引入了若干基于实践经验的优化规则，以提高搜索效率。

特征优选的主要流程如图 3 所示。首先，将待选特征按其物理意义分成若干组，对每个特征组进行独立的前向搜索。通过特征选择算法和交叉验证函数评估，选出在每个特征组中表现最佳的特征。其次将选出的特征组合成一个初始的特征子集。在初始特征子集的基础上，继续前向搜索，逐步将其他特征添加进来，并评估模型性能。此过程一直持续，直到最终完成特征优选。

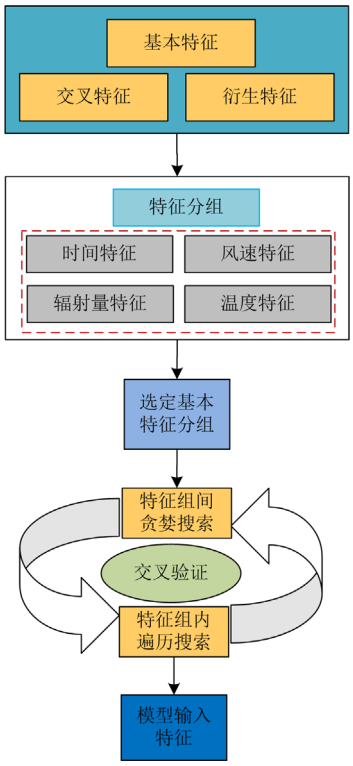

图 3 特征选择方法的主要流程

Fig. 3 Flow chart of the feature selection method

2.4 XGBoost 模型的超参数寻优策略

在寻找最优的超参数组合时，传统的穷举法或网格搜索法的计算复杂度较高，特别是当超参数数量较多时，会耗费较多的计算机资源和时间成本[24]。XGBoost 模型一般拥有 9 个超参数，如此大量的参数直接采用网格搜索法来确定最优超参数组合是不切实际的。为了更有效地进行超参数寻优，本文提出了一种两重多阶段的超参数寻优策略。

图 4 为两重多阶段寻优策略的主要流程。两重策略即首先在一个粗粒度的超参数网络上进行搜索，基于第一轮搜索的结果，确定一个细粒度的超参数网络进行第二轮搜索，最终得到最优的超参数取值。多阶段策略即将 XGBoost 模型的 9 个超参数按照其类型分为 4 组，逐步进行网格搜索。每次在一组超参数上进行搜索时，其他超参数设置为默认值，从而确保网格搜索的高效性。随着每轮网格搜索的进行，根据获得的最优网格点更新下一轮搜索的超参数设置，逐步迭代直至确定最优的超参数组合。两重多阶段策略能够在有限的算力条件下完成超参数寻优。

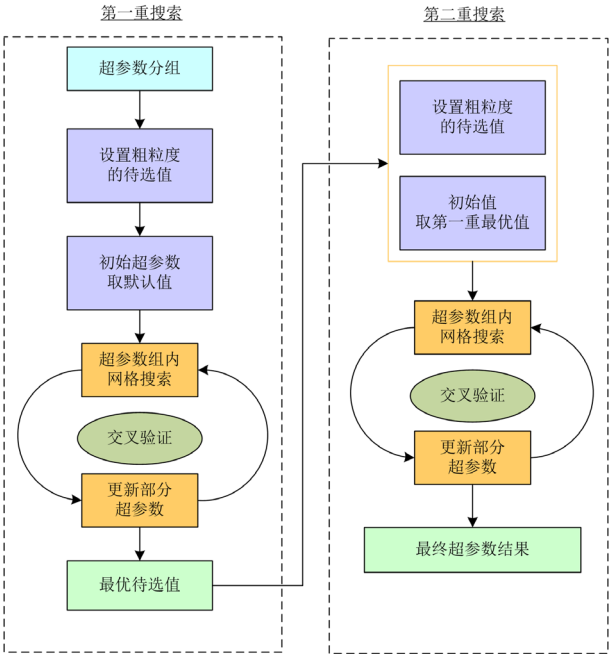

图 4 两重多阶段寻优策略的主要流程

Fig. 4 Primary procedure of a two-stage and multi-stage optimization strategy

3 实证分析

3.1 实验数据和实验设定

本文使用的数据包括中国西北某省份 5 个光伏电站的光伏发电量数据以及当地气象局提供的气候预测数据。实验数据集选取 2017 年 1 月 31 日至 2022 年 9 月 30 日的光伏发电量数据以及气候预测数据，数据时间分辨率为 1 天。每个光伏电站包含 2190 条数据，其中气候预测数据从 2017 年 1 月 31 日 00：00 时起报，每月输出自起报日后 3 个完整月份的预测数据，气候预测数据包含时间、温度、风速风向和辐射降水等 16 个气候特征的日均值。

在实验过程中，按照第 1.1 节介绍的划分方法将气候预测数据划分为 3 个数据集。每个数据集分

别作为 3 个子模型的训练数据，用于预测未来第 1 个月、第 2 个月和第 3 个月的日光伏发电量。为最大限度地利用有限的数据资源，确保模型的稳健性和泛化能力，每个数据集均按照 8:1:1 的比例分为训练集、验证集和测试集，训练集用于训练和建立预测模型，验证集用于确定模型的超参数，测试集用于评估预测模型的预测水平。本文通过 Python 语言编程实现预测模型的构建。

3.2 预测评测方法

本文以平均绝对百分比误差(mean absolute percentage error, MAPE)和均方误差(root mean square error, RMSE)为标准评价整体预测精度。

$$M_{\mathrm{APE}}=\frac{100\%}{N}\sum_{t=1}^{N}\frac{|y_t-\hat{y}_t|}{y_t} \tag{13}$$

$$R_{\mathrm{MSE}}=\sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t-\hat{y}_t)^2} \tag{14}$$

式中：N 表示样本总数；y_t 为 t 时刻真实值；$\hat{y}_t$ 为 t 时刻预测值。

MAPE 与 RMSE 的共同点是得分越低则预测质量越高。由于本文日发电量数据较大，RMSE 是在对数据进行归一化后计算得到的。

3.3 模型预测结果分析

为验证改进 XGBoost 模型的有效性、准确性和适用性，选取决策树模型、ETS 模型和 ANN 模型与其进行对比实验。图 5 为光伏电站 1 在测试集中 3 个月的模型预测结果。图中光伏发电量波动较大，但对于未来 3 个月的中长期发电量预测来说，改进 XGBoost 预测模型仍表现良好，能够实现较为准确的预测。

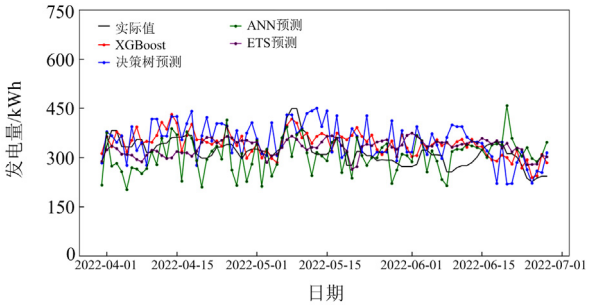

图 5 不同预测模型的光伏发电量预测曲线

Fig. 5 Forecast curves of photovoltaic power generation by different forecasting models

为了比较预测精度，表 4 和表 5 给出了 5 个光伏电站不同预测模型的 MAPE 和 RMSE。分析表中数据可知：虽然决策树模型具有直观的树状结构，易于理解和使用，但直接使用这种结构很难在复杂的时间序列和气候特征中找到可靠的非线性关系，因此与决策树模型相比，XGBoost 模型的 RMSE 分别降低了 0.0537、0.0308、0.0317、0.0522 和 0.0453，XGBoost 模型的 MAPE 分别降低了 4.15%、4.07%、6.31%、2.55%和 4.66%；ETS 模型在处理长时间序列数据时可以捕获时间序列中的趋势性、季节性和误差项，但其对非线性关系和复杂的交互效应建模能力有限。与 ETS 模型相比，XGBoost 模型的 RMSE 分别降低了 0.0142、0.0174、0.0405、0.0213 和 0.0314，而 XGBoost 模型的 MAPE 分别降低了 4.90%、2.24%、3.52%、0.98%和 2.40%；ANN 模型通常难以处理无明显规律性的时间数据，因此与 ANN 模型相比，XGBoost 模型的 RMSE 分别降低了 0.0076、0.0058、0.0115、0.0029 和 0.0209，XGBoost 模型的 MAPE 分别降低了 1.62%、1.81%、1.01%、0.61%和 0.92%。通过对比测试，改进的 XGBoost 预测模型具有最小的 MAPE 和 RMSE，预测效果最好。

表 4 不同预测模型的 RMSE 评价指标对比结果

Table 4 Comparison of RMSE scores of different forecasting models

光伏电站	RMSE			
	决策树	ETS	ANN	XGBoost
1	0.1395	0.1000	0.0934	0.0858
2	0.1763	0.1629	0.1513	0.1455
3	0.1240	0.1328	0.1038	0.0923
4	0.1999	0.1690	0.1506	0.1477
5	0.1637	0.1498	0.1393	0.1184

表 5 不同预测模型的 MAPE 评价指标对比结果

Table 5 Comparison of MAPE scores of different forecasting models

光伏电站	MAPE/%			
	决策树	ETS	ANN	XGBoost
1	15.17	15.92	12.64	11.02
2	17.08	15.25	14.82	13.01
3	18.37	15.58	13.07	12.06
4	16.56	14.99	14.62	14.01
5	17.62	15.36	13.88	12.96

3.4 模型消融实验

为了评估验证改进 XGBoost 模型功能模块的有效性，通过删除或简化集成到模型中的相应模块分别设计两种消融实验，即特征选择消融实验和超参数搜索消融实验。

3.4.1 特征选择消融实验

为验证改进的前向特征选择算法的重要性和有

效性，移除和简化模型中的气候特征选择模块，即“不考虑气候特征筛选”和“使用简单前向特征选择算法”，使用这两种模型进行预测，并与本文所提出的“改进 XGBoost 模型”进行比较，结果如表 6 所示。其中，XGBoost1 表示未使用特征选择的实验模型，XGBoost1*表示使用简单前向特征选择的实验模型。消耗时间表示特征选择所消耗的时间。

表 6 改进特征选择算法的消融实验结果

Table 6 Ablation experimental results of improved feature selection algorithm

光伏电站	预测模型	MAPE/%	消耗时间/s
1	XGBoost	11.02	20.93
	XGBoost1	13.51	0
	XGBoost1*	11.53	233.59
3	XGBoost	12.06	45.56
	XGBoost1	14.43	0
	XGBoost1*	12.26	395.63
5	XGBoost	12.96	50.13
	XGBoost1	14.65	0
	XGBoost1*	13.34	495.36

以光伏电站 1 为例，XGBoost 模型比 XGBoost1 模型的 MAPE 降低了 2.49%。未经过特征选择的 XGBoost1 模型可能包含了一些不相关或冗余的特征，导致模型复杂性提高，预测能力降低。实验表明“特征选择”模块有助于剔除不必要的特征，降低预测模型的复杂性，提高预测模型的预测能力。另一方面，XGBoost 模型比 XGBoost1*模型的 MAPE 降低了 0.51%，消耗时间减少了 212.66 s。实验表明按照实践经验和物理意义对气候特征分组后不仅提高了预测精度，而且减少了模型训练时间、提升了计算效率。本实验验证了改进前向特征选择算法的重要性和有效性。

3.4.2 超参数搜索消融实验

为了验证两重多阶段超参数搜索的必要性，简化模型中的超参数调整模块，即不采用两重多阶段策略进行超参数调整，直接使用网格搜索进行超参数调整，观察模型的训练速度和模型的预测精度。相关结果如表 7 所示，其中，XGBoost2 表示采用简单网格搜索的对比模型，训练时间表示每个模型中超参数寻优的训练时间。

由表 7 可知，在光伏电站 1 中，XGBoost 模型比 XGBoost2 模型的 MAPE 降低了 0.99%，其训练时间也减少了 538.1 s；在光伏电站 4 中，XGBoost 模型与 XGBoost2 模型的 MAPE 相同，但训练时间减少了 287.41 s；在光伏电站 5 中，XGBoost 模型对比 XGBoost2 模型的 MAPE 增加了 0.05%，但训练时间却减少了 524.46 s。综合而言，采用两重多阶段超参数寻优策略，不仅能够有效降低光伏发电量预测误差，而且大幅缩减 XGBoost 模型的训练时间。本实验验证了两重多阶段超参数搜索策略的重要性和有效性。

表 7 超参数寻优算法的消融实验结果

Table 7 Ablation experimental results of hyperparameter optimization

光伏电站	预测模型	MAPE/%	训练时间/s
1	XGBoost	11.02	72.19
	XGBoost2	12.01	610.29
4	XGBoost	14.01	40.82
	XGBoost2	14.01	328.23
5	XGBoost	12.96	69.10
	XGBoost2	12.91	593.56

4 结论

本文提出了一种基于气候特征分析及改进 XGBoost 算法的中长期光伏电站发电量预测方法，按预测周期将气候预测数据分阶段应用于不同的子模型中，从而更有效地利用相应时间段内的气候预测数据。通过特征工程、特征优选、超参数寻优等对气候特征进行深度挖掘，对 XGBoost 模型进行改进。改进 XGBoost 模型相对于其他模型在各个光伏电站上有着较好的预测效果。改进 XGBoost 模型中的特征选择算法和超参数搜索策略在提高预测精度的同时，显著减少了模型训练时间，为中长期光伏日发电量预测提供了一种有效且可行的方法。未来的工作可以进一步考虑更多或更精细的气候特征、扩展到更多地区进行应用与验证，并结合实际工程场景对预测模型开展进一步优化和提升。

参考文献

[1] XIA S, DING Z, DU T, et al. Multitime scale coordinated scheduling for the combined system of wind power, photovoltaic, thermal generator, hydro pumped storage, and batteries[J]. IEEE Transactions on Industry Applications, 2020, 56(3): 2227-2237.

[2] AMJADY N, DARAEEPOUR A. Midterm demand prediction of electrical power systems using a new hybrid forecast technique[J]. IEEE Transactions on Power Systems, 2011, 26(2): 755-765.

[3] HAN Shuang, QIAO Yanhui, YAN Jie, et al. Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network[J]. Applied Energy, 2019, 239: 181-191.

[4] 商立群, 李洪波, 侯亚东, 等. 基于 VMD-ISSA-KELM 的短期光伏发电功率预测[J]. 电力系统保护与控制, 2022, 50(21): 138-148.
SHANG Liqun, LI Hongbo, HOU Yadong, et al. Short-

term photovoltaic power prediction based on VMD-ISSA-KELM[J]. Power System Protection and Control, 2022, 50(21): 138-148.

[5] 李勇, 邱亚军, 覃茂欢, 等. 光伏组件间阴影遮挡预测模型的建立[J]. 太阳能, 2023(11): 33-38.
LI Yong, QIU Yajun, QIN Maohuan, et al. Establishment of shadow occlusion prediction model between PV modules[J]. Solar Energy, 2023(11): 33-38.

[6] 唱友义, 孙赫阳, 顾泰宇, 等. 采用历史数据扩充方法的风力发电量月度预测[J]. 电网技术, 2021, 45(3): 1059-1068.
CHANG Youyi, SUN Heyang, GU Taiyu, et al. Monthly forecast of wind power generation using historical data expansion method[J]. Power System Technology, 2021, 45(3): 1059-1068.

[7] 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.
LEI Kesong, TUSONGJIANG·Kari, YILIHAMU·Yaermaimaiti, et al. Short-term photovoltaic power prediction based on WGAN-GP and CNN-LSTM-Attention[J]. Power System Protection and Control, 2023, 51(9): 108-118.

[8] YANG Y, YU T Y, ZHAO W G, et al. Kalman filter photovoltaic power prediction model based on forecast experience[J]. Frontiers in Energy Research, 2021, 9: 682852.

[9] KIM B, SUH D, OTTO M O, et al. A novel hybrid spatio-temporal forecasting of multisite solar photovoltaic generation[J]. Remote Sensing, 2021, 13: 2605-2610.

[10] 周鑫, 李燕, 曾永辉, 等. 基于SARIMAX-SVR的光伏发电功率预测[J/OL]. 电力系统及其自动化学报: 1-8[2023-11-12]. ttps://doi.org/10.19635/j.cnki.csu-epsa.001322.
ZHOU Xin, LI Yan, ZENG Yonghui, et al. PV power prediction based on SARIMAX-SVR[J/OL]. Proceedings of the CSU-EPSA: 1-8[2023-11-12]. https://doi.org/10.19635/j.cnki.csu-epsa.001322.

[11] XIAO Bo, ZHANG Sujun, CHEN Suying, et al. A statistical photovoltaic power forecast model (SPF) based on historical power and weather data[C] // 2021 IEEE 48th Photovoltaic Specialists Conference (PVSC), June 20-25, 2021, Fort Lauderdale, FL, USA: 26-28.

[12] ATIQUE S, NOUREEN S, ROY V, et al. Forecasting of total daily solar energy generation using ARIMA: a case study[C] // 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), January 7-9, 2019, Las Vegas, NV, USA: 114-119.

[13] 王小君, 窦嘉铭, 刘嬖, 等. 可解释人工智能在电力系统中的应用综述与展望[J]. 电力系统自动化, 2024, 48(4): 169-191.
WANG Xiaojun, DOU Jiaming, LIU Zhao, et al. Review and prospect of application of explainable artificial Intelligence in power systems[J]. Automation of Electric Power Systems, 2024, 48(4): 169-191.

[14] HAYDAR D, ZOE R. Missing value imputation for short to mid-term horizontal solar irradiance data[J]. Applied Energy, 2018, 225: 998-1012.

[15] 方鹏, 高亚栋, 潘国兵, 等. 基于 LSTM 神经网络的中长期光伏电站发电量预测方法研究[J]. 可再生能源, 2022, 40(1): 48-54.
FANG Peng, GAO Yadong, PAN Guobing, et al. Research on medium and long term photovoltaic power generation forecasting method based on LSTM neural network[J]. Renewable Energy, 2022, 40(1): 48-54.

[16] KRISHNAMURTHY V. Predictability of weather and climate[J]. Earth and Space Science, 2019, 6(7): 2333-2344.

[17] TAKEYOSHI K. Prediction of photovoltaic power generation output and network operation[J]. Integration of Distributed Energy Resources in Power Systems, 2016: 77-108.

[18] LIU Wei, REN Chao. PV generation forecasting with missing input data: a super-resolution perception approach[J]. IEEE Transactions on Sustainable Energy, 2021, 12(2): 1493-1496.

[19] LI Qiaoqiao, XU Yan. An integrated missing-data tolerant model for probabilistic PV power generation forecasting[J]. IEEE Transactions on Power Systems, 2022, 37(6): 4447-4459.

[20] KIM T, KO W, KIM J. Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting[J]. Applied Sciences, 2019, 9(1): 204-211.

[21] LI Yiyan, SONG Lidong, ZHANG Si, et al. A TCN-based hybrid forecasting framework for hours-ahead utility-scale PV forecasting[J]. IEEE Transactions on Smart Grid, 2023, 14(5): 4073-4085.

[22] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C] // Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 11-13, 2016: 785-794.

[23] 罗晨, 喻锟, 曾祥君, 等. 基于高频重构信号与 Bayes-XGBoost 的低压电弧故障辨识方法研究[J]. 电力系统保护与控制, 2023, 51(13): 91-101.
LUO Chen, YU Kun, ZENG Xiangjun, et al. Low voltage arc fault identification method based on high frequency reconstructed signal and Bayes-XGBoost[J]. Power System Protection and Control, 2023, 51(13): 91-101.

[24] LI Yang, ABDALLAH S. On hyperparameter optimization of machine learning algorithms: theory and practice[J]. Neurocomputing, 2020, 415: 295-316.

收稿日期：2023-12-17； 修回日期：2024-03-27

作者简介：

李永飞(2000—)，男，硕士研究生，研究方向为可再生能源预测；E-mail: Leo_young@xjtu.edu.cn

张 耀(1988—)，男，通信作者，博士，副教授，博士生导师，研究方向为电力系统运行与规划、可再生能源预测等。E-mail: yaozhang_ee@xjtu.edu.cn

(编辑 姜新丽)