

基于包络线聚类的多模融合超短期光伏功率预测算法

杨国清¹, 张 凯¹, 王德意¹, 刘 菁², 秦美荣¹

(1. 西安理工大学 电气工程学院, 陕西 西安 710054; 2. 西安理工大学 西安市智慧能源重点实验室, 陕西 西安 710054)

摘要:针对传统功率预测方法以气象因素进行聚类划分时各气象因素权重难以分配以及单模型预测精度较差的问题,提出一种基于光伏功率包络线聚类的多模融合超短期光伏功率预测算法。对异常特征数据进行预处理,采用Pearson相关系数与XGB Feature Importance模块分析光伏功率和各特征之间的相关关系,并构建新特征;介绍包络线理论,并根据光伏功率包络线参数进行聚类划分,将聚类后的数据作为输入,借鉴Stacking集成学习框架构造XGBoost+LightGBM+LSTM融合模型对光伏功率进行预测;将所提算法与气象因素聚类 and 功率区间聚类下的各预测算法进行实验对比;为了避免训练过程中模型超参数的影响,采用K折交叉验证对数据的训练集、验证集和测试集进行划分。仿真结果表明,所提算法较气象因素和功率区间聚类法能有效提高复杂天气情况下光伏功率预测精度,且多模融合效果总体优于单独算法模型。

关键词:光伏功率预测;包络线聚类;多模融合算法;特征工程;K折交叉验证

中图分类号:TM 615

文献标志码:A

DOI:10.16081/j.epae.202101009

0 引言

光伏发电出力具有随机波动性、周期性的特点,研究光伏出力变化并做出准确的预测,关系到整个电网的安全稳定运行,能源最优管理分配对于电力负荷调度也具有重大意义^[1]。超短期光伏功率预测属于光伏功率研究中的一个细分领域,预测时间尺度为0~6 h,其因能提供短时功率瞬变信息,对未来光伏电站参与调峰^[2]、电力调度计划制定^[3]、多能互补的工业园区新能源发电与负荷之间的消纳^[4]、电力市场现货交易^[5]均具有一定指导意义。

现有的光伏功率预测法从模型方法上可分为物理方法和统计方法^[6]。统计方法基于预测模型输入、输出因素之间的统计规律进行预测,不考虑内部复杂的建模过程,其原理可以视为黑箱模型。近年来以统计方法为原理的各人工智能算法在光伏功率预测中得到了广泛的应用^[7-8]。

聚类是分类问题的一种统计方法,将具有相似特征的数据聚为一类,可以使同一类数据具有高度的相似性,从而缩短模型训练的时间,增强模型的训练效果。文献[9]通过改进Kohonen模型,将数据分为春秋夏季和冬季2种季节维度作为输入,分别得到4种和6种天气类型并进行训练。文献[10]通过Pearson相关系数分析光伏发电量与气象因子间的相关性,以温度和湿度作为输入特征,通过自组织特征映射SOM(Self-Organizing feature Map)对天气类

型进行聚类识别。文献[11]通过概率神经网络(PNN)将天气划分为晴、雨、阴、多云4种类型,通过主成分分析(PCA)对输入特征进行降维,获取影响光伏出力的主要成分并进行建模。文献[12]通过SOM聚类将天气类型划分为晴天、阴天、多云天、突变天气(晴转阴、雨转晴、阴转雨等),考虑空间相关性的净空出力和天气聚类的相对出力进行建模。上述聚类算法都取得了较好的效果,但都以气象因素进行聚类划分,而气象因素对输出光伏功率的影响程度不同,且不同天气类型的主导因素也不同,传统Pearson相关系数等指标难以衡量各气象因素对发电量的影响程度,导致基于气象因素聚类时各气象因素的权重难以分配。此外,气象因素聚类法只考虑影响光伏功率的外部因素,而忽略了影响光伏功率的内部因素,如光伏组件类型、控制器等内部参数,这可能导致具有相似气象条件的光伏电站出力产生巨大差异,从而影响聚类效果。光伏功率区间聚类法^[13]将光伏出力的值作为指标进行聚类,避免了内部参数对聚类效果的影响,但其对光伏出力进行离散化处理,割裂了光伏出力的时间序列特性,不能反映光伏功率的随机波动情况。据此,本文提出一种基于包络线理论的聚类方法,根据待预测点前一定时间维度的功率包络线参数进行聚类划分,该聚类方法能将待预测点划分到与其相属的波动簇中,且避免了气象因素聚类时权重分配与内部因素影响的问题。

在光伏超短期功率预测领域中,机器学习模型和时间序列模型被广泛应用^[14-17]。多模融合算法融合了上述2种及以上算法,从而进一步提高了预测精度,目前关于融合算法的研究较少。文献[16]结合极限梯度提升XGBoost(eXtreme Gradient Boosting)与长短期记忆网络LSTM(Long Short-Term Memory

收稿日期:2020-03-07;修回日期:2020-11-09

基金项目:国家自然科学基金资助项目(51507134);陕西省重点研发计划项目(2018ZDXM-GY-169)

Project supported by the National Natural Science Foundation of China(51507134) and the Key Research and Development Program of Shaanxi Province(2018ZDXM-GY-169)

networks)算法进行电力负荷超短期预测,结果表明融合模型较单算法模型有更高的预测精度和更低的泛化误差。文献[18-19]使用均值计算的方式求取多种算法模型或同一类型算法不同参数模型的预测均值,并按比例进行模型组合。上述多模融合算法较传统单算法模型取得了更高的预测精度,但多模融合方式只是各模型间的简单组合,未考虑不同模型在预测过程中的优劣性与差异性,也未进行单模型的训练优化。文献[20]采用Stacking集成学习方法融合多个算法模型,并选择基学习器,结果表明具有强学习能力的基学习器能显著提高负荷预测精度。

本文借鉴文献[20]中Stacking集成学习的思想,从离散性和连续性两方面构建融合模型进行光伏功率预测。离散性预测在基于包络线聚类情况下训练具有强特征学习能力的基学习器XGBoost与轻量级梯度提升机LightGBM(Light Gradient Boosting Machine)。由于聚类会打破光伏出力时间序列的连续性,因此连续性预测基于全体样本训练基学习器LSTM。最后将离散性预测和连续性预测值作为特征构建次学习器并得出预测结果。仿真结果表明,基于包络线聚类的多模融合算法较其余因素聚类的单模型预测算法有更高的预测精度。

1 特征工程

本文以广东某地区光伏电站3个月的实证数据作为研究对象,每15 min记录一次数据,每天记录96组数据。数据包括光伏逆变器采集到的电压、电流、功率等参数以及光伏电站配套气象站采集到的温度、湿度、风速等气象参数。

1.1 异常数据检测及修复

准确的光伏运行数据和气象数据是进行光伏功率预测研究的基础,但在光伏电站的实际运营中维护、限电、通信故障等原因往往导致实测数据中含高比例异常数据,引起数据间相关关系混乱,使模型对温度、光照强度等特征的拟合能力变差,影响预测模型精度^[6]。因此,进行异常数据的检测与修复是光伏功率预测的前提。根据中心极限定理,当随机变量的数量足够大时,其概率分布服从正态分布,故本文采用3-sigma原理识别异常数据,如式(1)所示,在识别出异常数据后,用前向替代法进行数据替换。

$$X - \bar{X} > 3\sigma \quad (1)$$

其中, $X = [X_1, X_2, \dots, X_n]$ 为含各特征的待识别光伏数据矩阵, $\bar{X} = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n]$ 为含各特征平均值的矩阵, $\sigma = [\sigma_1, \sigma_2, \dots, \sigma_n]$ 为各特征标准差的矩阵,其元素如式(2)所示, n 为待识别数据的特征维数。

$$\sigma_i = \sqrt{\frac{1}{m-1} \sum_{j=1}^m (X_{ij} - \bar{X}_i)^2} \quad (2)$$

其中, X_{ij} 为 X_i 的第 j 个样本数据; m 为数据的样本数量。

附录A图A1是平均功率异常数据处理前后对比图。从图中可见,原始数据在部分样本处有非常明显的峰值波动,数据修复后波动规律大致正常。

1.2 数据标准化

为避免模型训练过程中因各特征数据量纲不同而使模型优先拟合高量纲维度的数据,需对预训练数据进行标准化处理。本文调用sklearn.preprocessing机器学习模块对各特征数据进行min-max标准化,使标准化后的值处于0~1范围内,如式(3)所示。

$$x_a = (x_b - x_{\min}) / (x_{\max} - x_{\min}) \quad (3)$$

其中, x_b 为待进行标准化处理的数据; x_a 为进行标准化处理后的数据; $x_{\min}$ 、 $x_{\max}$ 分别为输入数据的最小值和最大值。

1.3 各特征对光伏功率的影响

光伏功率曲线具有周期性^[6]和波动性两大特点。日照昼夜周期性和日-地相对位置周期性变化决定了光伏功率周期性的特点,而影响光伏出力的主要因素包括内部因素和外部因素,如图1所示。其中,内部因素包括光伏组件的类型、生产的材料与工艺、设备的安装角度、系统内设备的损耗和转换效率等,外部因素包括温度、湿度、光照强度、空气污染情况、云量、经纬度、赤道角等地理气象因素。

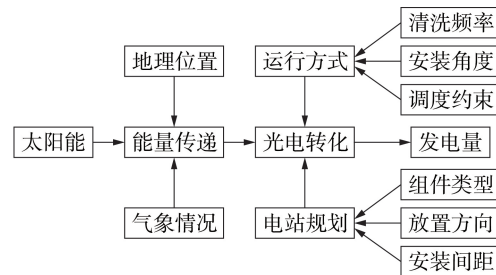


图1 影响光伏功率的因素

Fig.1 Factors affecting photovoltaic power

二维 Pearson 相关系数能够衡量2个变量之间的相关强度,如式(4)所示。

$$r_{ed} = \frac{\sum_{i=1}^n (e_i - \bar{e})(d_i - \bar{d})}{\sqrt{\sum_{i=1}^n (e_i - \bar{e})^2} \sqrt{\sum_{i=1}^n (d_i - \bar{d})^2}} \quad (4)$$

其中, $\bar{e}$ 、 $\bar{d}$ 和 n 分别为随机向量 e 、 d 中元素的平均值和个数; e_i 和 d_i 分别为 e 和 d 中的第 i 个元素值。

当相关系数分别位于区间 $[0.8, 1.0]$ 、 $[0.6, 0.8]$ 、 $[0.4, 0.6]$ 、 $[0.2, 0.4]$ 、 $(0, 0.2)$ 时,相关性依次为极强相关、强相关、中等程度相关、弱相关、极弱相关或不相关。部分特征与光伏功率的 Pearson 相关系数表如附录A表A1所示,在影响光伏功率的特征中, A 相电流、光照强度、组件板温相关性较强,其余特征为弱相关或不相关。作为衡量输入特征对光伏功率影响程度的指标, Pearson 相关系数具有局限性,其

只能反映特征和光伏功率间的正负线性相关关系,但光伏功率与各特征之间的关系往往是复杂的非线性关系,为了解决该问题,本文引入Python机器学习库XGBoost中的子模块XGB Feature Importance分析树模型中特征对光伏功率的影响。

部分特征的XGB树模型得分见附录A图A2。综合Pearson相关系数与XGB Feature Importance得分分析可知,现场温度、板温、光照强度对功率预测影响程度较高,风速和风向在2种评估方法中影响程度均较低,本文算法中包含树模型和LSTM神经网络模型两部分,具有强相关系数的A相电流特征在LSTM模型中体现出良好的拟合能力,而在树模型中影响因子较小,故其在树模型中可舍弃。

附录A图A3绘制了部分特征与光伏功率联合曲线,其中光伏功率和所有特征均进行了标准化处理。从图中可以看出:板温和光照强度能大致反映光伏功率变化趋势,但在光伏功率突变时拟合能力较差;A相电流拟合能力最好,当光伏功率突变时,A相电流变化规律与光伏功率变化规律一致,且幅值变化较为明显;风速与光伏功率拟合性最差,风速的波动与光伏功率的变化并无明显相关关系,为了简化模型,节省计算时间,可以舍弃该特征。

1.4 新特征的构造

为增强模型对输入特征的拟合能力,选取强相关特征并适当地人为构造新特征有利于加快模型学习速度、提高模型预测精度。构建新特征的步骤为:

(1)利用Pearson相关系数以及XGB Feature Importance分析各特征与光伏功率的相关关系;

(2)选择强相关特征用组合的方式构建新特征;

(3)根据步骤(1)对新特征进行相关性分析;

(4)根据步骤(3)中结果选择最终添加的新特征,并且为了规避过拟合风险,步骤(2)中的新特征应避免简单的线性组合。

图2为某时间段内光伏功率曲线。该曲线呈现日周期的变化规律,为刻画该规律,体现各样本所处日周期中的位置,构建峰距和谷距特征来衡量某样本与该样本最近的波峰和波谷间的距离,并记录该波峰、波谷的值。此外,将与光伏功率强相关的几个特征进行二阶组合,其结果见附录A表A2,根据构造特征的Pearson相关系数判断是否保留该组合特征。

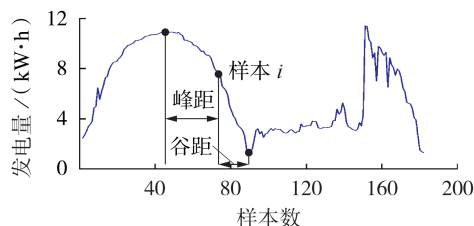


图2 峰、谷距示意图

Fig.2 Schematic diagram of peak and valley distances

为避免过多特征导致模型训练困难、计算时间增加、过拟合等风险,仅选择极强相关和强相关的组合新特征。需要注意的是,由于各特征对树模型与神经网络模型的影响程度不同,而表A2中的组合特征对树模型的影响因子较小,因此在XGBoost与LightGBM的模型训练中舍弃A相电流特征及组合特征。

2 包络线聚类算法

2.1 光伏功率包络线

由于光伏功率的随机波动性,准确衡量光伏出力曲线的波动程度并建立与波动程度对应的簇有利于提高预测精度。包络线原本为一种通信领域处理信号的方法,是描述振荡信号上下极值的曲线,如图3所示。包络线聚类算法选取预测点前一定时间维度的功率曲线构建功率包络线,将包络线的参数作为指标进行聚类。衡量曲线波动程度的指标由波动频率和波动幅度组成,波动频率通过包络线范围内的波峰和波谷数量来衡量,波动幅度主要由包络线之间的面积来衡量。图4展示了不同波动频率和波动幅度下的典型出力曲线。

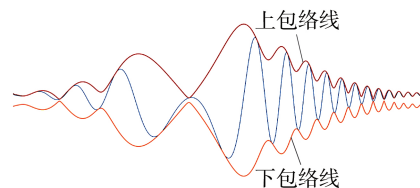


图3 信号包络线

Fig.3 Signal envelopes

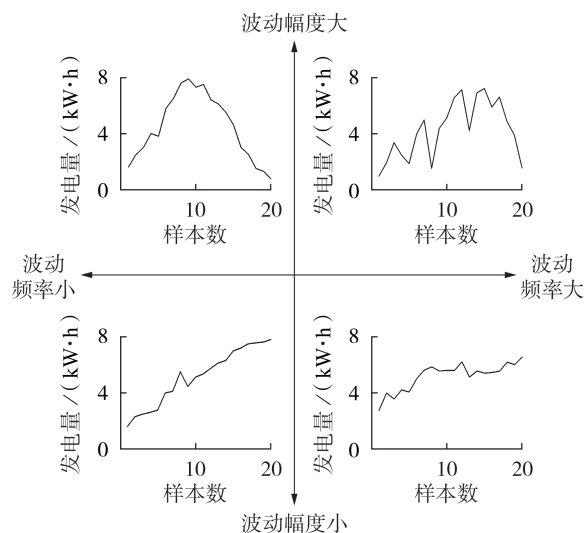


图4 不同波动频率和波动幅度下的典型出力曲线

Fig.4 Typical output curves under different fluctuation frequencies and amplitudes

波峰、波谷数和包络线之间的面积数值较大时,说明光伏出力不稳定,波动较大,此时的气象情况较为复杂多变;波峰、波谷数和包络线之间的面积数值

较小时,则说明光伏出力波动程度低,往往处于功率稳定爬坡阶段。通过这2个参数可以准确衡量光伏出力的波动情况,将待预测点划分到与其相属的波动簇中。包络线聚类法直接针对光伏电站历史出力曲线,忽略了内部参数对出力的影响,相较于气象因素聚类法,其更为简单直接,且能更有效地提取实时波动特征。由于光伏组件类型、控制器等内部参数的差异,不同光伏电站出力曲线往往存在差异性,这使得气象原始参数聚类法难以进行相似性特征的提取。且气象因素聚类法往往以天气类型作为聚类划分后的簇,将光伏数据划分为晴天、阴天、雨天等典型天气的数据集并进行单独建模,但在不同的天气类型下可能具有相似的气象条件,而光伏出力规律却完全不同。此外,由于各气象因素对光伏功率的影响程度难以衡量,因此以气象因素作为聚类指标进行聚类时各特征权重分配成为待解决的问题。

根据图5分析预测点前 z 小时曲线特征,步骤为:

- (1)绘制光伏功率曲线,找到观测范围内曲线的极大值和极小值,并分别将其定义为波峰和波谷;
- (2)利用三次样条插值法对所有波峰值进行曲线拟合,形成上包络线;
- (3)利用三次样条插值法对所有波谷值进行曲线拟合,形成下包络线;
- (4)求取上、下包络线之间的面积衡量观测范围内曲线波动的幅度;
- (5)本文数据采样间隔为15 min,若采样间隔较小,应在上述一阶包络线的基础上构建二次包络线。

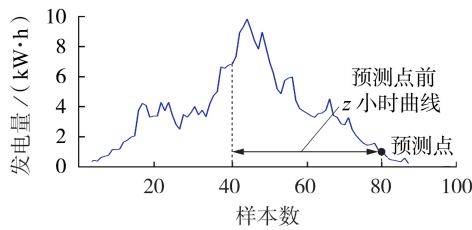


图5 预测点前 z 小时曲线特征分析

Fig.5 Characteristic analysis of curve z hours before prediction point

本文将预测点前10 h(40个数据点)的数据作为观测曲线范围,求得范围内波峰、波谷数量 f_i 和上、下包络线之间的面积 v_i ,考虑到观测范围内曲线的发电量平均维度的不同构建特征 h_i 。

$$v_i = \sum_{k=1}^{40} (S_k^{\text{upper}} - S_k^{\text{lower}}) \quad (5)$$

$$h_i = \sum_{k=1}^{40} (S_k^{\text{upper}} + S_k^{\text{lower}}) \quad (6)$$

$$f_i = p_i + q_i \quad (7)$$

其中, p_i 和 q_i 分别为预测点前10 h功率包络线的波峰数和波谷数; S_k^{upper} 和 S_k^{lower} 分别预测点前 k 点处上、

下包络线的值。

2.2 基于包络线特征的K-means聚类

K-means算法是一种常用的聚类算法,原理流程如下。

(1)给定聚类数 J 和样本数据集 D 。

(2)随机生成 J 个聚类中心。

(3)利用式(8)所示欧氏距离判断样本集中数据到 J 个聚类中心的距离。

$$d_{is}(s_i, c_j) = \sqrt{\sum_{t=1}^m (s_{it} - c_{jt})^2} \quad (8)$$

其中, s_i 为第 i 个样本数据; c_j 为第 j 个聚类中心; s_{it} 为第 i 个样本中的第 t 个特征数据; c_{jt} 为第 j 个聚类中的第 t 个特征数据。

(4)将数据点分配到距其最近的聚类中。

(5)在所有点分配完毕后根据各聚类内所有点重新计算聚类中心(取平均值)。

(6)重新计算 D 中数据到新的聚类中心的距离,重复步骤(3)一(5)。

(7)聚类中心不再变化或达到最大迭代次数时停止。

轮廓系数是衡量聚类效果的指标,如式(9)、(10)所示。计算第 i 个样本到同簇其他样本的平均距离 a_i , a_i 越小,说明第 i 个样本越应该被聚类到该簇,将 a_i 称为第 i 个样本的簇内不相似度。计算第 i 个样本到其他某簇 g_j 中所有样本的平均距离 b_{ij} ,将其称为第 i 个样本与簇 g_j 的不相似度。定义第 i 个样本的簇间不相似度 $b_i = \min \{b_{i1}, b_{i2}, \dots, b_{ij}\}$ 。若 $s(i)$ 接近1,则说明第 i 个样本聚类合理;若 $s(i)$ 接近-1,则说明第 i 个样本更应该分类到另外的簇;若 $s(i)$ 近似为0,则说明第 i 个样本在2个簇的边界上。

$$s(i) = (b_i - a_i) / \max \{a_i, b_i\} \quad (9)$$

$$s(i) = \begin{cases} 1 - a_i/b_i & a_i < b_i \\ 0 & a_i = b_i \\ b_i/a_i - 1 & a_i > b_i \end{cases} \quad (10)$$

图6是以不同特征聚类下包络线轮廓系数随聚类数的变化趋势图。在聚类之前对所有特征参数进行标准化处理,曲线1以波峰、波谷数量 f_i 和上、下包络线之间的面积 v_i 进行聚类,曲线2以 f_i 、 v_i 和上、下包络线波峰、波谷之和 h_i 进行聚类,曲线3以气象因素(环境温度、光照强度、风速、风向)进行聚类,曲线4以光伏功率区间范围进行聚类。可见,以 f_i 和 v_i 进行聚类轮廓系数最高,且聚类数为4最合理,聚类数过少时难以反映曲线波动情况,聚类数过多时轮廓系数下降,且每个聚类样本数量减少影响模型训练效果。以气象因素进行聚类时效果最差,标准化处理后的各气象因素间没有明显的相关关系,难以形成簇。综上所述,选择光伏功率包络线间的面积 v_i 和其范围内波峰、波谷数量 f_i 作为聚类指标最合适。

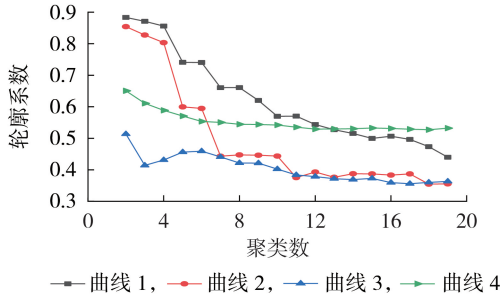


图6 不同聚类特征下包络线随聚类数的变化趋势

Fig.6 Trend of envelope curves with cluster number for different cluster features

选取聚类数为4,附录A图A4是以 f_i 和 v_i 进行聚类时的样本划分情况。由图中可见,簇1波动幅度小,波动频率较快,簇2波动幅度小,波动频率较慢,这2种簇占样本比例大,簇3波动幅度较大,波动频率无明显特点,随机分布,簇4波动幅度极大,样本数量最少。4种簇对应的典型包络线见图7,其中实线为功率曲线,虚线为功率曲线包络线。

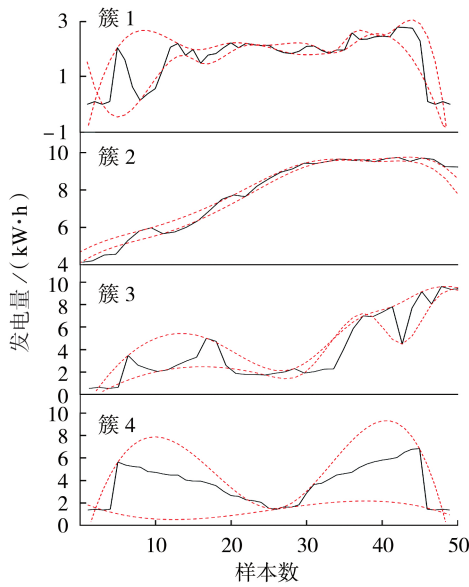


图7 簇1—4典型包络线

Fig.7 Typical envelope curves of Cluster 1-4

3 多模融合算法

3.1 Stacking集成学习

Stacking集成学习框架^[20]将训练集分为若干子数据集,并训练与之对应的基学习器,然后将第一层基学习器输出的预测结果作为第二层次学习器训练的特征进行第二层的训练,见附录A图A5。该模型第一层本质上是一种学习过程,是高度复杂的非线性特征转换器,可有效抽取原始数据中的特征,且数据集交叉验证避免了过拟合风险。相较于相互独立的预测模型,融合模型有更强的非线性表述能力和更低的泛化误差。本文借鉴Stacking集成学习的思

想,在Stacking集成学习的基础上进行如下改动:

- (1)输入数据集 $D=\{x_i, y_i\}_{i=1}^m$, x_i, y_i 分别为第 i 个样本数据的特征参数与发电量;
- (2)将 D 尽可能平分为 w 个子集 $D_1, D_2, \dots, D_w$;
- (3)分别基于数据集 $D_1, D_2, \dots, D_w$ 对第一层基学习器进行训练并输出预测结果 $t_1, t_2, \dots, t_w$;
- (4)将基学习器输出添加到第二层次学习器的特征中, $x'_i=\{x_{i1}, x_{i2}, \dots, x_{in}, t_1, t_2, \dots, t_w\}$,其中 x_{ij} 为第 i 组数据的第 j 个特征;
- (5)数据集更新为 $D_{\text{new}}=\{x'_i, y_i\}_{i=1}^m$;
- (6)进行第二层次学习器的学习,并输出预测结果。

3.2 融合模型

XGBoost模型^[16]是一种基于Boosting的改进集成树模型,通过贪心策略和最优化思想集合若干弱分类器,生成一棵新树并拟合前一棵树的残差,使模型损失函数下降,随着模型深度和迭代次数的提升,预测精度显著提高。LSTM是基于时间序列的神经网络,解决了循环神经网络(RNN)计算过程中容易梯度消失和梯度爆炸的问题,常用于处理时间序列相关问题。LightGBM^[17]是一个由微软开源的项目,相较于XGBoost具有轻量化、高效化的特点。光伏出力预测从离散性和连续性两方面进行,离散性预测光伏出力与时间序列无相关关系,只取决于当前预测参数状态,而连续性预测光伏出力基于历史出力并结合当前预测参数状态进行预测。融合模型的建立综合考虑光伏出力的离散性与连续性,构建基于包络线聚类的XGBoost+LightGBM基学习器模型与不聚类的LSTM基学习器模型,并融合各模型预测结果进行次学习器的训练,融合模型结构图见图8,图中0.2与0.3为各模型占融合模型的权重。由于神经网络模型(LSTM)和树模型(XGBoost+LightGBM)与各输入特征的相关性不同,需要差异化选择特征,其样本数据选择如表1所示,表中特征1为待预测的光伏发电量,其余特征为模型输入训练特征。

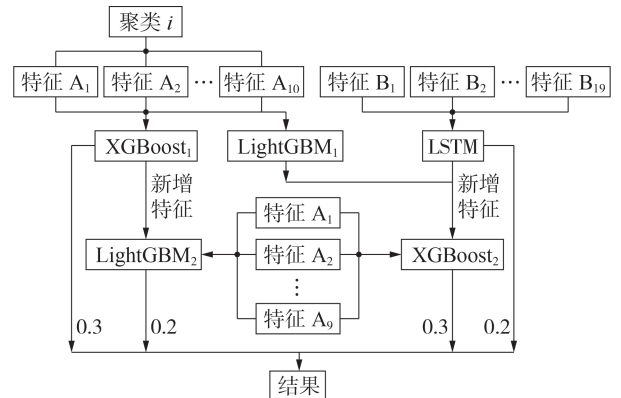


图8 融合模型结构图

Fig.8 Structure of fusion model

表1 神经网络与树模型数据点构成

Table 1 Data point composition of neural network and tree models

特征编号	神经网络模型	树模型
1	光伏发电量	光伏发电量
2、3	板温、环境温度	板温、环境温度
4、5	光照强度、风向	光照强度、风向
6	转化效率	转化效率
7—10	峰距、谷距值	峰距、谷距值
11—19	A相电流、附录A表A2组合特征	

4 算例分析

采用第1节中广东某地区光伏电站3个月的数据进行仿真分析,数据采样周期为15 min,每天采集96组数据,样本共9000组数据,将前8000组数据作为训练集,后1000组数据作为测试集,预测目标是未来15 min的光伏发电量,并与单独算法模型进行对比。实验基于python3.7环境,利用sklearn、pandas、keras等模块进行模型训练,实验平台为intel i5-4200h, GTX860M。预测评估指标选用平均绝对误差MAE(Mean Absolute Error)与平均绝对百分比误差MAPE(Mean Absolute Percentage Error),其值 e_{MAE} 、 e_{MAPE} 分别见附录A式(A1)、(A2)。

4.1 光伏功率预测流程及模型超参数选取

基于包络线聚类多模融合超短期光伏功率预测算法流程图如图9所示。合适的模型超参数能大

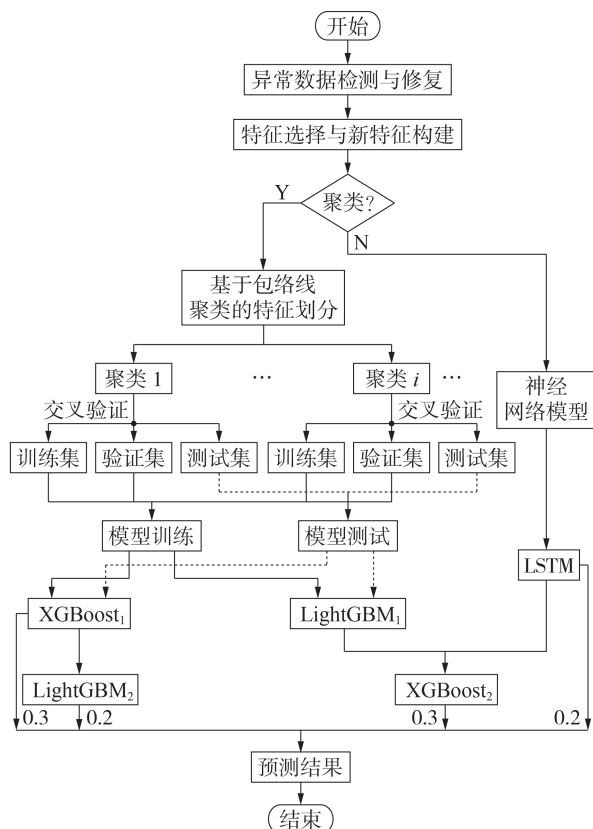


图9 光伏功率预测算法流程图

Fig.9 Flowchart of photovoltaic power prediction algorithm

幅度提高模型预测的精准度,本文先利用网格搜索对单模型进行超参数调优,再考虑模型融合后的综合性能,确定的模型超参数如附录A表A3所示。

4.2 基于光伏功率包络线聚类多模融合超短期光伏功率预测结果分析

随机抽取部分样本进行多种算法模型预测值与真实值的比较,结果见附录B图B1。当功率比较平稳时,5种算法模型均有不错的预测效果;当功率发生小波动时,融合算法与LSTM较LightGBM与XGBoost算法精准度更高;当功率发生较大波动时,融合算法体现出较高的预测精度,而LSTM与深度神经网络反向传播(DNN-BP)模型产生了明显的功率预测偏差,对于DNN-BP,这可能与该模型参加训练的样本数量较少且波动随机性较大有关,对于LSTM,这是由RNN的滞后性特点导致的。

各算法模型预测误差如表2所示。表中根据不同聚类指标将算法分为光伏功率包络线聚类法、气象因素聚类法、光伏功率区间聚类法,每种方法聚类为4簇,再按各簇样本比例进行加权得到总体样本指标。基于总体样本,融合算法使用光伏功率包络线聚类时,MAE为0.08540 kW·h,MAPE为4.0081%;使用气象因素进行聚类时,MAE为0.11990 kW·h,MAPE为8.1588%;使用光伏功率区间聚类时,MAE为0.09420 kW·h,MAPE为7.6869%。采用本文融合算法基于不同指标进行聚类时部分样本预测值与真实值如附录B图B2所示,可见光伏功率区间和气象因素聚类较光伏功率包络线聚类法预测误差更大,特别是当发电量(光伏出力)处于低水平时。

光伏功率包络线聚类法根据功率包络线曲线参数将样本分为4种典型簇,各簇在不同算法模型下的预测误差见附录B图B3。其中,簇1和簇2分别对应包络线间面积小、波动频率大以及包络线间面积小、波动频率小的2种数据样本类型,该类型样本比例最高;簇3和簇4分别对应包络线间面积较大与极大的2种数据样本类型,该类型样本比例较低。采用本文算法模型时,簇1和簇2的MAE与MAPE均小于该簇其余算法模型;簇3的MAE最小,MAPE小于LSTM与DNN-BP,略高于XGBoost与LightGBM;簇4的MAE与MAPE均小于除LSTM外的其余算法模型。

此外,由表2可以看出,在以光伏功率包络线作为聚类指标时,融合算法的MAE较XGBoost、LightGBM、LSTM、DNN-BP分别下降0.0093、0.0178、0.0185、0.1152 kW·h,MAPE分别降低0.855%、0.959%、0.212%、11.870%,证明了融合算法较单模型算法有更高的预测精度。而在以气象因素和功率区间作为聚类指标时,部分簇的融合算法预测误差大于该簇单模型预测算法,这说明了融合模型预测的精准度在一定程度上依赖聚类效果,也从侧面印证了光伏功率包络线聚类的有效性。

表2 各算法模型预测误差

Table 2 Prediction error of each algorithm model

聚类指标	聚类簇别	$e_{MAE} / (\text{kW} \cdot \text{h})$					$e_{MAPE} / \%$				
		XGBoost	LightGBM	LSTM	DNN-BP	融合算法	XGBoost	LightGBM	LSTM	DNN-BP	融合算法
光伏功率 包络线	1	0.056 80	0.062 2	0.058 64	0.079 69	0.042 25	1.741 3	1.908 7	1.987 4	2.873 2	1.489 6
	2	0.112 00	0.119 8	0.125 10	0.142 00	0.106 90	6.379 3	6.425 5	5.106 5	8.889 2	5.173 3
	3	0.078 36	0.091 2	0.160 30	4.275 80	0.061 30	7.195 2	5.907 7	15.636 6	618.330 3	8.921 0
	4	0.413 40	0.691 2	0.163 70	0.380 30	0.220 60	7.255 8	14.945 8	3.054 1	11.847 0	4.522 3
	总体	0.094 70	0.103 2	0.103 90	0.200 60	0.085 40	4.863 2	4.967 3	4.220 8	15.878 5	4.008 1
气象因素	1	0.241 80	0.251 5	0.071 60	0.044 79	0.192 30	72.241 2	78.052 3	12.387 1	8.732 3	58.207 6
	2	0.092 70	0.110 2	0.151 00	0.160 80	0.090 50	1.828 0	2.324 0	1.976 4	2.580 1	1.740 7
	3	0.193 80	0.124 5	0.131 19	0.256 20	0.161 90	5.876 1	6.949 9	3.628 3	8.279 3	4.916 7
	4	0.071 30	0.075 4	0.084 80	0.090 50	0.064 50	7.130 8	7.310 0	8.879 3	10.170 3	6.618 4
	总体	0.137 20	0.121 1	0.127 40	0.174 30	0.119 90	9.736 0	10.812 3	4.486 3	6.294 7	8.158 8
光伏功率 区间	1	0.067 60	0.070 0	0.083 60	0.075 60	0.057 01	1.677 1	1.764 0	2.152 9	1.864 0	1.425 2
	2	0.125 30	0.119 3	0.220 80	0.126 70	0.130 30	1.392 1	1.285 4	2.191 7	1.281 3	1.295 6
	3	0.119 60	0.132 0	0.137 70	0.172 40	0.104 30	1.779 8	1.967 6	1.972 2	2.639 5	1.555 0
	4	0.091 40	0.091 2	0.072 30	0.078 90	0.086 00	15.417 3	16.159 0	13.629 4	11.805 7	22.624 2
	总体	0.092 60	0.102 9	0.128 10	0.118 40	0.094 20	5.921 7	6.006 3	4.643 3	5.884 8	7.686 9

5 结论

本文针对传统以气象因素和功率区间聚类的单模光伏预测模型预测精度较差的问题,提出一种基于光伏功率包络线聚类的多模融合模型,引入通信领域包络线和机器学习领域集成学习的前沿算法技术,详细阐述异常数据的处理、特征筛选和构建、包络线聚类算法及聚类特征的选择、多模型融合集成学习算法等预测流程。算例测试结果表明:光伏功率包络线聚类法较气象因素和功率区间聚类法能更准确反映光伏功率的随机波动情况,预测误差更低;融合模型综合考虑光伏出力的离散性和连续性,在合适的聚类划分情况下较单算法模型预测精度有较大提升。因此,基于光伏功率包络线聚类的多模融合算法能提高复杂天气情况下的光伏功率预测精度,在超短期光伏功率预测方面有较高的应用价值。

在今后的研究中,将进一步开展如下工作:进行特征的深度挖掘,构建能极大提高模型拟合能力的强特征;尝试改进三次插值函数来提高包络线对光伏功率曲线的拟合能力;基于Stacking框架融合模型最优设计,寻找高性能的基学习器,简化模型,缩短计算时间,提高预测精度。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] 王继东,庞文杰. Fisher判别分类法在光伏并网系统稳态电能质量评估中的应用[J]. 电力自动化设备,2017,37(3):50-54.
WANG Jidong, PANG Wenjie. Application of Fisher discriminant analysis in steady-state power quality evaluation of grid-connected photovoltaic system[J]. Electric Power Automation Equipment, 2017, 37(3): 50-54.
- [2] 杨永标,于建成,李奕杰,等. 含光伏和蓄能的冷热电联供系统调峰调蓄优化调度[J]. 电力系统自动化,2017,41(6):6-12,29.
YANG Yongbiao, YU Jiancheng, LI Yijie, et al. Optimal load leveling dispatch of CCHP incorporating photovoltaic and storage[J]. Automation of Electric Power Systems, 2017, 41(6): 6-12, 29.
- [3] 曾鸣,杨雍琦,向红伟,等. 兼容需求侧资源的“源-网-荷-储”协调优化调度模型[J]. 电力自动化设备,2016,36(2):102-111.
ZENG Ming, YANG Yongqi, XIANG Hongwei, et al. Optimal dispatch model based on coordination between “generation-grid-load-energy storage” and demand-side resource[J]. Electric Power Automation Equipment, 2016, 36(2): 102-111.
- [4] 李宏仲,王磊,林冬,等. 多主体参与可再生能源消纳的Nash博弈模型及其迁移强化学习求解[J]. 中国电机工程学报,2019,39(14):4135-4150.
LI Hongzhong, WANG Lei, LIN Dong, et al. A Nash game model of multi-agent participation in renewable energy consumption and the solving method via transfer reinforcement learning[J]. Proceedings of the CSEE, 2019, 39(14): 4135-4150.
- [5] 刘佩云,丁涛,贺元康,等. 基于综合需求响应的负荷聚合商最优市场交易策略[J]. 电力自动化设备,2019,39(8):224-231.
LIU Peiyun, DING Tao, HE Yuan Kang, et al. Optimal trading strategy for load aggregator based on integrated demand response[J]. Electric Power Automation Equipment, 2019, 39(8): 224-231.
- [6] 龚莺飞,鲁宗相,乔颖,等. 光伏功率预测技术[J]. 电力系统自动化,2016,40(4):140-151.
GONG Yingfei, LU Zongxiang, QIAO Ying, et al. An overview of photovoltaic energy system output forecasting technology[J]. Automation of Electric Power Systems, 2016, 40(4): 140-151.
- [7] 张程熠,唐雅洁,李永杰,等. 适用于小样本的神经网络光伏预测方法[J]. 电力自动化设备,2017,37(1):101-106,111.
ZHANG Chengyi, TANG Yajie, LI Yongjie, et al. Photovoltaic power forecast based on neural network with a small number of samples[J]. Electric Power Automation Equipment, 2017, 37(1): 101-106, 111.
- [8] AGOXA X G, GIRARD R, KARINIOTAKIS G. Short-term spatio-temporal forecasting of photovoltaic power production[J]. IEEE Transactions on Sustainable Energy, 2018, 9(2): 538-546.
- [9] 李燕青,杜莹莹. 基于双维度顺序填补框架与改进Kohonen天气聚类的光伏发电短期预测[J]. 电力自动化设备,2019,39(1):60-65.
LI Yanqing, DU Yingying. Short-term photovoltaic power forecasting based on double-dimensional sequential imputation framework and improved Kohonen clustering[J]. Electric Po-

- wer Automation Equipment, 2019, 39(1):60-65.
- [10] 代倩, 段善旭, 蔡涛, 等. 基于天气类型聚类识别的光伏系统短期无辐照度发电预测模型研究[J]. 中国电机工程学报, 2011, 31(34):28-35.
- DAI Qian, DUAN Shanxu, CAI Tao, et al. Short-term PV generation system forecasting model without irradiation based on weather type clustering[J]. Proceedings of the CSEE, 2011, 31(34):28-35.
- [11] 王昕, 黄柯, 郑益慧, 等. 基于PNN/PCA/SS-SVR的光伏发电功率短期预测方法[J]. 电力系统自动化, 2016, 40(17):156-162.
- WANG Xin, HUANG Ke, ZHENG Yihui, et al. Short-term forecasting method of photovoltaic output power based on PNN/PCA/SS-SVR[J]. Automation of Electric Power Systems, 2016, 40(17):156-162.
- [12] 王晶, 黄越辉, 李驰, 等. 考虑空间相关性和天气类型划分的多光伏电站时间序列建模方法[J]. 电网技术, 2020, 44(4):1376-1384.
- WANG Jing, HUANG Yuehui, LI Chi, et al. Time series modeling method for multi-photovoltaic power stations considering spatial correlation and weather type classification[J]. Power System Technology, 2020, 44(4):1376-1384.
- [13] LI K W, WANG R, LEI H T, et al. Interval prediction of solar power using an improved Bootstrap method[J]. Solar Energy, 2018, 159:97-112.
- [14] SHI Z C, LIANG H, DINAVAH V. Direct interval forecast of uncertain wind power based on recurrent neural networks[J]. IEEE Transactions on Sustainable Energy, 2018, 9(3):1177-1187.
- [15] LI Z L, XIA J, LIU A, et al. States prediction for solar power and wind speed using BBA-SVM[J]. IET Renewable Power Generation, 2019, 13(7):1115-1122.
- [16] 陈振宇, 刘金波, 李晨, 等. 基于LSTM与XGBoost组合模型的超短期电力负荷预测[J]. 电网技术, 2020, 44(2):614-620.
- CHEN Zhenyu, LIU Jinbo, LI Chen, et al. Ultra short-term power load forecasting based on combined LSTM-XGBoost model[J]. Power System Technology, 2020, 44(2):614-620.
- [17] JU Y, SUN G Y, CHEN Q H, et al. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting[J]. IEEE Access, 2019, 7:28309-28318.
- [18] LI Zhiyi, LIU Xuan, CHEN Liyuan. Load interval forecasting methods based on an ensemble of extreme learning machines[C]//2015 IEEE Power & Energy Society General Meeting. Denver, CO, USA:IEEE, 2015:1-5.
- [19] ZHANG R, XU Y, DONG Z Y, et al. Short-term load forecasting of Australian National Electricity Market by an ensemble model of extreme learning machine[J]. IET Generation, Transmission & Distribution, 2013, 7(4):391-397.
- [20] 史佳琪, 张建华. 基于多模型融合Stacking集成学习方式的负荷预测方法[J]. 中国电机工程学报, 2019, 39(14):4032-4042.
- SHI Jiaqi, ZHANG Jianhua. Load forecasting based on multi-model by Stacking ensemble learning[J]. Proceedings of the CSEE, 2019, 39(14):4032-4042.

作者简介:



杨国清

杨国清(1979—),男,江苏南京人,副教授,博士,主要研究方向为电力系统新能源消纳、高电压新技术(E-mail: yanggq@xaut.edu.cn);

张凯(1996—),男,陕西西安人,硕士研究生,主要研究方向为人工智能算法在电力系统中的应用、电力系统新能源消纳控制(E-mail: 183600393@qq.com)。

(编辑 王锦秀)

Multi-mode fusion ultra-short-term photovoltaic power prediction algorithm based on envelope clustering

YANG Guoqing¹, ZHANG Kai¹, WANG Deyi¹, LIU Jing², QIN Meirong¹

(1. School of Electrical Engineering, Xi'an University of Technology, Xi'an 710054, China;

2. Xi'an Key Laboratory of Smart Energy, Xi'an University of Technology, Xi'an 710054, China)

Abstract: Aiming at the problems that the weights of meteorological factors are difficult to assign and the prediction accuracy of single model is poor when traditional power prediction methods are used for meteorological factor cluster division, a multi-mode fusion ultra-short-term photovoltaic power prediction algorithm based on envelope clustering of photovoltaic clustering is proposed. The abnormal feature data is preprocessed, Pearson correlation coefficient and XGB Feature Importance module are adopted to analyze the correlation between photovoltaic power and each feature, and new features are constructed. The envelope theory is introduced, the photovoltaic envelope parameters are used for cluster division, the data after clustering is taken as input, and the fusion model of XGBoost+LightGBM+LSTM is constructed using Stacking integrated learning framework to predict photovoltaic power. The prediction algorithm is compared with each forecasting algorithm under meteorological factor clustering and power interval clustering. In order to avoid the influence of model hyperparameters during training process, K fold cross validation is used to divide the training set, validation set and test set of the data. The simulative results show that the proposed algorithm can effectively improve the prediction accuracy of photovoltaic power under complex weather conditions compared with the meteorological factor and power interval clustering methods, and multi-mode fusion effect is totally better than single algorithm model.

Key words: photovoltaic power prediction; envelope clustering; multi-mode fusion algorithm; feature engineering; K-fold cross validation

附录 A:

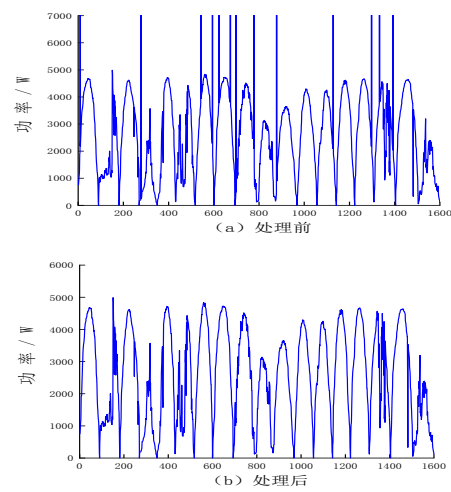


图 A1 光伏电站平均功率异常数据处理前后波形图

Fig.A1 Waveforms before and after restoration of abnormal average power data of photovoltaic power plants

表 A1 影响光功率的部分特征的 Pearson 系数表

Table A1 Pearson coefficient table of partial characteristics affecting optical power

变量	相关系数	相关性
A 相电流	0.909 233 496	极强相关
光照强度	0.879 283 275	极强相关
板温	0.519 033 444	中等程度相关
风向	0.217 038 381	弱相关
平均功率	0.168 906 929	极弱相关或不相关
风速	0.128 217 446	极弱相关或不相关

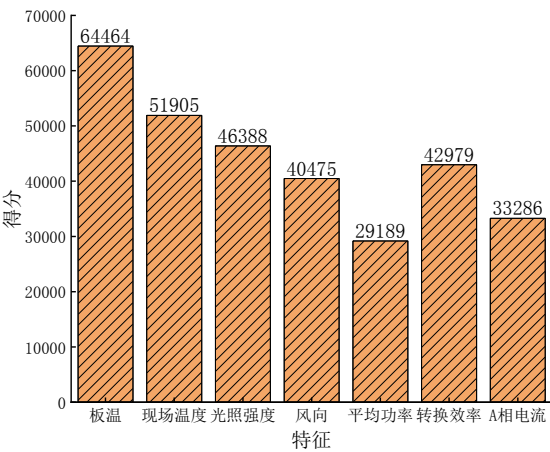


图 A2 XGB 树模型特征得分

Fig.A2 Feature score XGB tree model

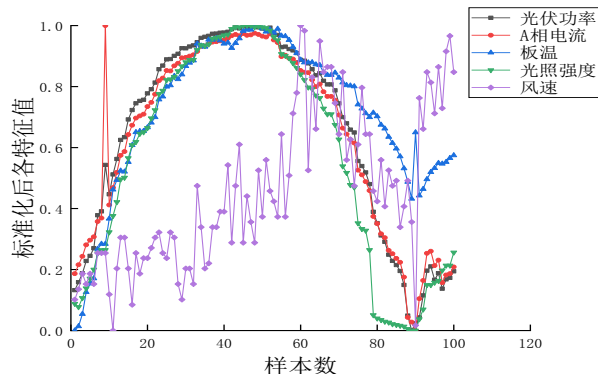


图 A3 部分特征与光伏功率

Fig.A3 Partial features and photovoltaic power

表 A2 组合特征相关性分析

Table A2 Correlation analysis of combined features

变量	相关系数	相关性（是否舍弃）
A 相电流乘光照强度	0.899 40	极强相关（否）
A 相电流乘板温	0.629 00	强相关（否）
A 相电流乘风向	0.812 50	极强相关（否）
A 相电流乘转换效率	0.767 40	强相关（否）
光照强度乘板温	0.585 32	中等程度相关（是）
光照强度乘风向	0.736 10	强相关（否）
光照强度乘转化效率	0.991 50	极强相关（否）
板温乘风向	0.490 60	中等程度相关（是）
板温乘转换效率	0.492 10	中等程度相关（是）
风向乘转换效率	-0.018 20	或不相关（是）
A 相电流平方	0.949 08	极强相关（否）
光照强度平方	0.814 70	极强相关（否）
板温平方	0.294 10	弱相关（是）
风向平方	0.197 30	不相关（是）
转换效率平方	-0.143 80	不相关（是）

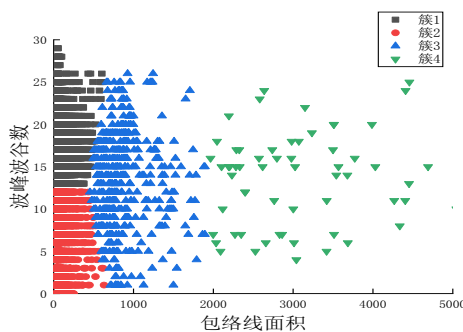


图 A4 二维聚类样本划分

Fig.A4 Two-dimensional clustering sample partition

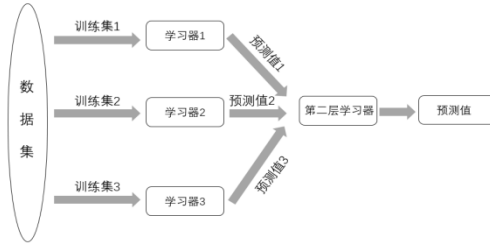


图 A5 Stacking 集成学习框架

Fig.A5 Ensemble learning framework based on Stacking

$$e_{\text{MAE}} = \frac{1}{m} \sum_{i=1}^m |\hat{y}_i - y_i| \quad (\text{A1})$$

$$e_{\text{MAPE}} = \frac{100\%}{m} \sum_{i=1}^m \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (\text{A2})$$

其中， m 为数据的样本数量； $\hat{y}_i$ 为样本 i 的预测值。

表 A3 模型超参数

Table A3 Model hyperparameters

算法模型	超参数集
XGBoost	树深度 6，学习率 0.01，决策树的数目 15000，最小叶子节点样本权重 1，每个树上的叶子数 20，每次分裂对列数的采样的占比 0.4，稳定终止迭代数 300
LightGBM	学习率 0.005，决策树的数目 12000，最小叶子节点样本数 20，每次迭代所用特征比例 0.8，稳定终止迭代数 2000
LSTM	迭代次数 1000，隐含层 1，时间步长 20，1 次训练所选取的样本数 200
DNN-BP	迭代次数 1000，隐含层 7，隐含层神经元数 64，优化器 adam，L2 正则化，激活函数 relu

附录 B:

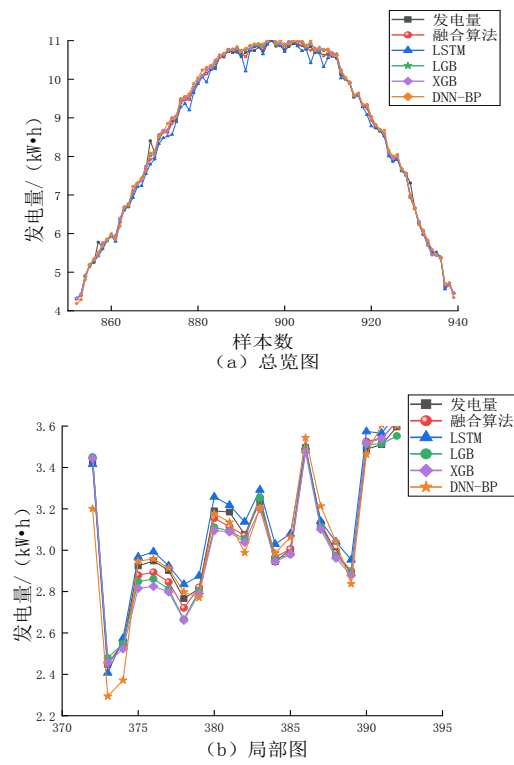


图 B1 各算法模型预测值与真实值曲线图

Fig.B1 Curves of predicted value and actual value of each algorithm model

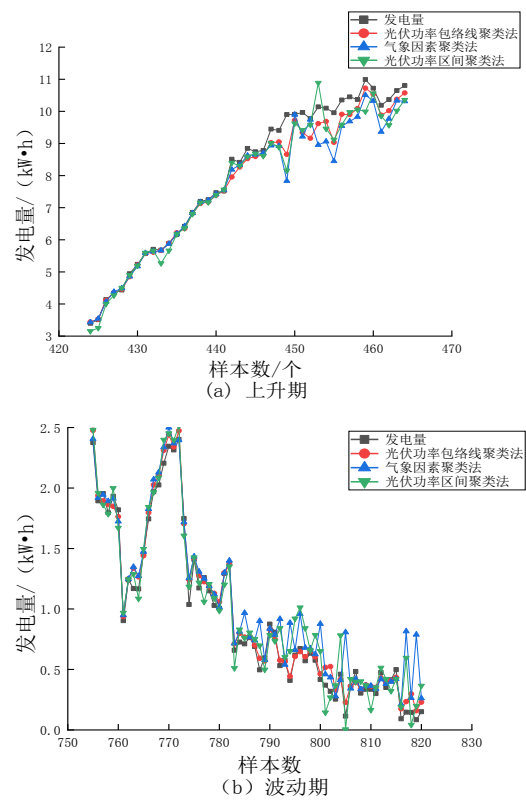


图 B2 不同聚类方法预测结果对比图

Fig.B2 Comparison of prediction results among different clustering methods

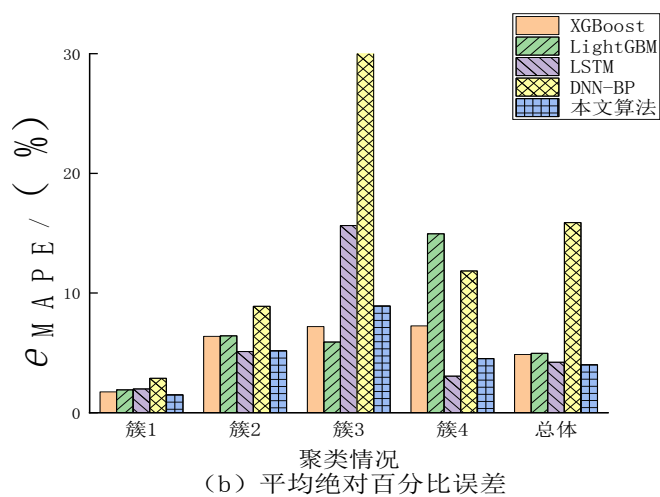
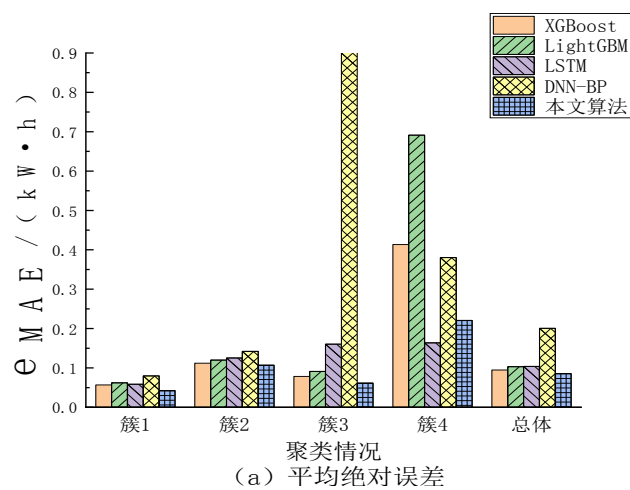


图 B3 各簇在不同算法模型下的预测误差

Fig.B3 Prediction error of each cluster under different algorithm models