计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):36−45 [doi: 10.15888/j.cnki.csa.007395]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Stacking 模型融合的光伏发电功率预测①

杨荣新, 孙朝云, 徐　磊

(长安大学 信息工程学院, 西安 710064)
通讯作者: 杨荣新, E-mail: 15829289297@163.com

摘　要: 为了提高光伏发电输出功率的预测精度和可靠性, 本文提出一种基于 Stacking 模型融合的光伏发电功率预测方法. 选取某光伏电站温度、湿度、辐照度等历史实测数据为研究对象, 在将光伏发电功率数据进行特征交叉以及基于模型的递归特征消除法进行预处理和特征选择的基础上, 以 XGBoost、LightGBM、RandomForest 3 种机器学习算法作为 Stacking 集成学习的第一层基学习器, 以 LinearRegression 作为第二层元学习器, 构建了多个机器学习算法嵌入的 Stacking 模型融合的光伏发电功率预测模型. 预测结果表明, 该方法的 R^2、MSE 分别达到了 0.9874 和 0.1056, 相较于单一的机器学习模型, 预测精度显著提升.

关键词: 光伏发电; Stacking; 模型融合; 基学习器; 元学习器

引用格式: 杨荣新,孙朝云,徐磊.基于 Stacking 模型融合的光伏发电功率预测.计算机系统应用,2020,29(5):36–45. http://www.c-s-a.org.cn/1003-3254/7395.html

Photovoltaic Power Prediction Based on Stacking Model Fusion

YANG Rong-Xin, SUN Zhao-Yun, XU Lei

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: In order to improve the prediction accuracy and reliability of photo voltaic power prediction output, this study proposes a photo voltaic power prediction method based on Stacking model fusion. The historical measured data such as temperature, humidity, and irradiance of a PV power plant are selected as the research object. Based on the feature intersection of the photo voltaic power data and the pre-processing and feature selection based on the model-based recursive feature elimination method, XGBoost and LightGBM are used. The three machine learning algorithms of Random Forest are the first layer of base learning for Stacking integrated learning. Linear Regression is used as the second layer of element learner to construct a photo voltaic power prediction model with multiple stacking models embedded in machine learning algorithms. The prediction results show that the R^2 and MSE of the method reach 0.9891 and 0.1358, respectively, and the prediction accuracy is significantly improved compared with the single machine learning model.

Key words: PV; Stacking; modelfusion; baselearner; metalearner

近年来, 光伏发电产业凭借着自身清洁、环保和无污染的诸多优势, 以及在国家相关政策的大力支持下, 实现了跨越式发展[1]. 但太阳光的光照强度、环境温度等多种因素都会对太阳能发电的输出功率产生影响, 这使得光伏发电出力表现出强烈的间接性和时间波动性. 所以, 研究如何有效提高光伏发电输出功率的

① 基金项目: 陕西省交通运输厅交通科研项目 (18-22R)
Foundation item: Transportation Research Project of Shaanxi Provincial Transportation Department (18-22R)
收稿时间: 2019-09-18; 修改时间: 2019-10-15, 2019-11-05; 采用时间: 2019-11-18; csa 在线出版时间: 2020-05-07

预测精度对光伏发电力网系统的组件调度和电力管理有着非常重要的意义[2].

国内外针对光伏发电功率预测的方法主要分为物理预测法和统计预测法. 基于物理的预测法把光伏电站的地理位置、气象条件等结合太阳能辐射传递方程和光伏组件方程来加以实现[3]. 统计预测方法将太阳辐射强度、风速、温湿度、气压等因素作为输入变量, 通过线性回归、BP 神经网络 (Back Propagation Neural Network, BPNN)、支持向量机 (Support Vector Machines, SVM) 等技术挖掘输入变量与光伏发电功率间的隐含关系, 并结合天气预报数据进行预测[4–6]. 针对神经网络初始参数的随机性缺点, 文献[7]提出了一种结合启发式算法优化 BP 神经网络权重和阈值的光伏发电功率预测方案; 文献[8]在对数据集的预处理上采用了聚类算法, 将聚类后不同类别的光伏电场数据分别建立 SVM 预测模型; 文献[9]则提出了一种选择相似日的方法, 把输出功率相似的时间段进行捆绑, 建立了最小二乘向量机的光伏阵列输出功率预测模型.

上述研究方案都是利用单一的算法模型对光伏发电的输出功率进行预测, 其预测精度不会很高, 表现出较大的局限性, 为此很多研究学者提出了组合预测的方法[10]. 组合预测方法因其可以结合各单一模型的优点于一身, 因此在很多领域也得到了广泛应用. 文献[11]提出了将相似日的光伏发电时间序列数据进行模态分解后, 对固有模态和趋势分量分别建立基于人工蜂群算法优化的支持向量机预测模型; 文献[12]为了减轻不确定性对电网的负面影响, 提出了一种灰色神经网络 (灰色−神经网络) 混合模型来预测光伏发电的短期输出功率; 文献[13,14]充分发挥各单一预测模型的优势, 并按权重将其进行优化整合以提高预测精度; 文献[15–17]将原始光伏发电数据按照信号分解的方式进行解耦, 使得特征分解为互异的模态向量, 并对这些分量构建预测模型. 采用这些组合预测模型虽然可在一定程度上提高光伏功率的输出功率预测精度, 但模型融合大多采用简单的线性加权整合, 鲁棒性无法得到保证.

基于以上研究, 本文在分析了光伏发电功率预测与人工智能技术发展背景的基础上, 利用 Stacking 集成学习框架对 XGBoost、LightGBM、Random Forest、型的优势, 进一步提高了光伏发电输出功率的预测精度.

1 算法理论介绍

1.1 XGBoost 算法机理

XGBoost 是经过优化的集成树模型, 从梯度提升树模型改进和扩展而来. 树的集成模型由式 (1) 表示:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{1}$$

其损失函数为:

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

其中, $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$, T 代表叶子的个数, w 代表叶子的权重. 并且有:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

那么损失函数可以表示为:

$$L^{(t)} = \sum_{i=1}^{n} l(\hat{y}_i^{(t-1)}, y_i + f_t(x_i)) + \Omega(f_t) \tag{4}$$

对损失函数进行泰勒展开有:

$$L^{(t)} \approx \sum_{i=1}^{n} [l(\hat{y}_i^{(t-1)}, y_i) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{5}$$

其中,

$$\begin{cases} g_i = \partial_{\hat{y}_i^{(t-1)}} l(\hat{y}_i^{(t-1)}, y_i) \\ h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(\hat{y}_i^{(t-1)}, y_i) \end{cases} \tag{6}$$

移除常数项有:

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{7}$$

将第 j 个叶子节点定义为$I_j = \{i|q(x_i) = j\}$, 即有:

$$\begin{aligned} \tilde{L}^{(t)} &= \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i\in I_j} g_i) w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda) w_j^2] + \gamma T \end{aligned} \tag{8}$$

然后将上式求导并令求导结果等于 0, 可得:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{9}$$

把 w_j 的最优解 w_j^*带入目标函数, 得到:

$$\tilde{L}_{(q)}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{10}$$

XGBoost 在进行节点分裂时, 采用一种贪心算法, 每次在已有的叶子节点中加入分裂. 假设 I_L 和 I_R 分别是分裂后的左和右叶子节点的集合. 信息增益如下:

$$Gain = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{11}$$

从式 (11) 中可知, 这个信息增益与 ID_3、$C_{4.5}$ 和分类与回归树是类似的, 都是利用分裂后的某种熵值减去分裂前的固有熵值. 同时, 为了限制树的深度, 加入了阈值γ, 这种策略以正则化的方式有效地避免了过拟合.

1.2 LightGBM 算法机理

针对传统的 Boosting 框架下的算法存在效率和可扩展性方面的不足, LightGBM 进行了两个方面的改进, 即采用梯度单边采样和互斥特征捆绑[18]. 梯度单边采样是指 LightGBM 在对模型的参数进行训练的时候, 没有使用训练集的全部样本, 而是根据梯度的大小进行了采样, 只选取那些具有高梯度的样本数据来计算信息增益. 互斥特征捆绑是指在稀疏特征空间中, 大部分特征不会同时取非 0 值, 通过将这些互斥的特征合并为单一特征, 以此达到降低特征维度的目的.

根据梯度单边采样丢弃掉某些小梯度样本数据的主要思想原则为出发点, LightGBM 在对数据集进行采样时保留梯度大的样本, 对梯度较小的样本则按照一定比例进行下采样. 为了抵消对数据分布的影响, 梯度单边采样小梯度的样本数据在计算信息增益时引入系数 $(1–a)/b$, a 表示大梯度数据的采样率, b 表示小梯度数据的采样率, 梯度单边采样的具体步骤如下:

(1) 按照数据集的梯度绝对值进行排序, 并选取最大的 $a\times100\%$ 数据集保留, 作为大梯度样本点子集.

(2) 从剩余数据集中随机选取 $b\times100\%$ 数据生成小梯度样本点集合.

(3) 将小梯度样本点集合乘以常数 $(1–a)/b$ 放大样本数据.

(4) 合并样本集, 得到一个采样集, 通过该采样集的训练, 产生一个弱学习器.

(5) 不断重复上述 4 个过程, 直至训练的模型达到提前设置的迭代次数或者出现收敛的状态.

为了提高模型训练的并行能力, LightGBM 不仅对训练样本的数据进行了按梯度采样的策略, 同时也对高维稀疏性的互斥特征进行了融合绑定, 例如, 进行独热编码后的数据, 这些数据在进行独热编码后不仅维度升高, 而且呈现出稀疏的特性. 为了减少数据的特征维度, LightGBM 采用基于直方图 (histogram) 的方式将这些高维稀疏且互斥的特征捆绑在一起, 以提高节点的分裂效率. 如图 1 所示, 这种算法首先将输入连续特征的数据离散化为 k 个整数值, 形成 k 个捆绑的结果, 每个结果内各个特征都是互斥的, 然后构建宽度为 k 的直方图, 在对训练数据进行遍历时, 只需要统计每个离散值在直方图上的累积量即可. 由于在计算分裂增益时, 是通过遍历排序直方图的离散值而得, 因此只需要计算 k 次, 且在特征分裂时, 只需要保存特征离散化后的值, 相较于 XGBoost 而言, 减小了计算和存储的成本, 提高了分裂点寻找的效率, 降低了模型的计算复杂度.

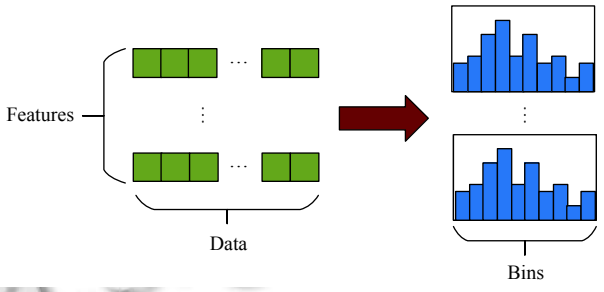


图 1 Histogram 算法基本过程示意图

1.3 基于 Stacking 的集成学习方式

如图 2 所示, 在基于 Stacking 的集成学习方式下, 整个模型的构建分为两个阶段, 通过以学习器级联的方式进行预测结果的传递, 提高预测精度[19]. 在第一阶段, 首先将原始数据集进行切分, 按照一定比例划分为训练集和测试集, 然后选取合适的基学习器以交叉验证的方式对训练集进行训练, 将训练完成后的各个基学习器对验证集和测试集进行预测, 第一阶段应该选取预测性能优秀的机器学习模型, 同时保证模型间的多元化; 在第二阶段, 将基学习器的预测结果分别作为元学习器训练和预测的特征数据, 元学习器结合上个阶段得到的特征和原始训练集的标签为样本数据进行模型构建, 并输出最终的 Stacking 模型预测结果, 该阶

段的元学习器一般选取稳定性较好的简单模型, 起到整体提升模型性能的作用.

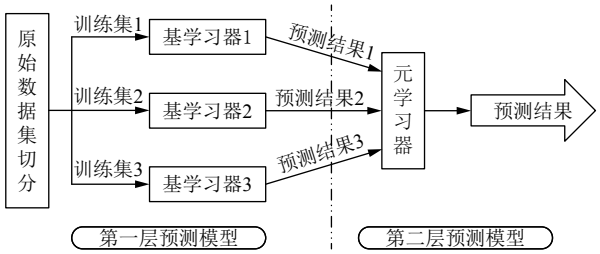


图 2 基于 Stacking 的集成学习方式

可以从两方面来看待 Stacking 方法, 第一, 它是许多集成方法的推广, 第二, 它是通过学习得到的集成方法. 在 Stacking 的训练阶段, 从第一层学习器中得到新的数据集, 如果用同一份完全相同的数据训练第一层学习器, 并用该份数据在第一层学习器上的输出作为第二层学习器的训练数据, 这会有过拟合风险. 因此, 训练第一层学习器的数据, 不能作为构造第二层学习器的数据, 所以在构造第二层学习器的训练数据时本文采用了交叉验证的方法来选取第二层学习器的训练数据.

本文选取基于五折模型的 Stacking 算法建立光伏发电功率预测模型, 其流程框图如图 3 和图 4 所示. 具体步骤如下:

(1) 将原始数据集按照一定比例切分为训练集和测试集.

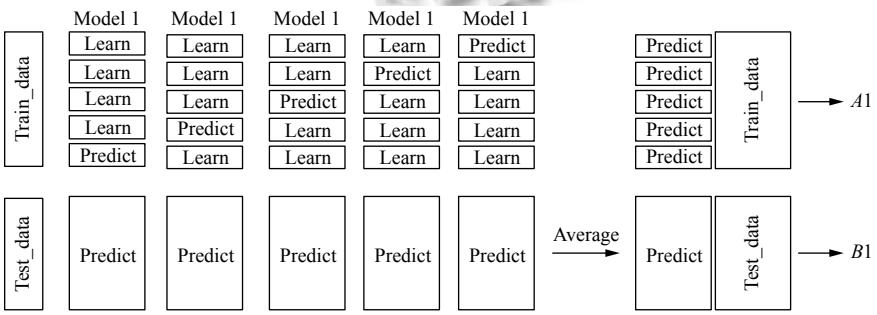


图 3 第一层单个基学习器 5 折交叉验证示意图

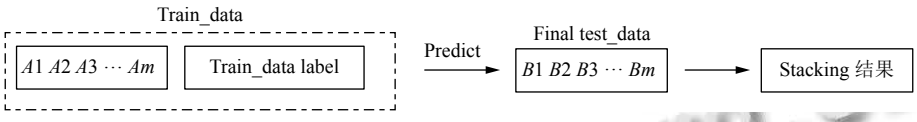


图 4 第二层元学习器模型构建示意图

(2) 将训练集平均分为 5 份, 即对于每一个基学习器进行 5 折交叉验证, 在每次交叉验证时以 4 份作为模型的训练集, 剩余的一份作为验证集, 并且在每次交叉验证完成时, 用训练的基学习器模型对验证集和测试集进行预测.

(3) 单个基学习器在完成 5 折交叉验证后, 会得到各个验证集的预测集, 也即训练集每条样本的预测值, 同时得到测试集的 5 列预测值, 然后将各个验证集的预测集整合为 1 列, 记为 $A1$, 将测试集的 5 列预测值按行取平均, 得到 $B1$.

(4) 当第一层的 m 个基学习器完成训练后, 会得到第二层元学习器模型的输入特征矩阵 ($A1, A2, \cdots, Am$), 将原始训练集的标签值作为模型的输出矩阵, 进行模型的训练. 同时会得到最终的测试集输入特征矩阵 ($B1, B2, \cdots, Bm$).

(5) 当元学习器训练完成后, 将 ($B1, B2, \cdots, Bm$) 作为特征矩阵, 利用模型输出模型融合后的最终预测结果.

2 基于 Stacking 模型融合的光伏发电功率预测模型分析

Stacking 集成学习是一种建立在统计学习理论基础之上的多算法融合的机器学习方法, 一般情况下, 对于单一的预测模型而言, 其预测准确率是呈现边际效用递减的趋势, Stacking 集成学习方式是组合来自多个预测模型的信息以生成新模型的模型集成技术. 将不同的机器学习算法通过不同的方式结合在一起, 以此获得比单一算法更优越的性能. 在 Stacking 集成学习

模型中, 要充分分析每个基学习器的单独预测能力, 使得 Stacking 集成学习模型获得最佳的预测效果.

基于 Stacking 的集成模型算法能够提高建模的精度, 但是, 由于集成模型具有融合多个模型进行建模的特性, 势必在整体建模上会牺牲一定的建模速度. 因此, 为了兼顾 Stacking 算法的预测性能和整体建模速度, 本文选取了预测精度较高的随机森林、XGBoost 以及预测性能优异且算法时间复杂度较低的 LightGBM 作为 Stacking 模型融合的第一层, 其中, 随机森林和 XGBoost、LightGBM 分别采用 Bagging 和 Boosting 的集成学习方式, 有着出色的学习能力和严谨数学理论支撑, 在各个领域得到了广泛的应用. 第二层模型采用了稳健性和泛化能力较强的 LinearRegression, 模型架构如图 5 所示.

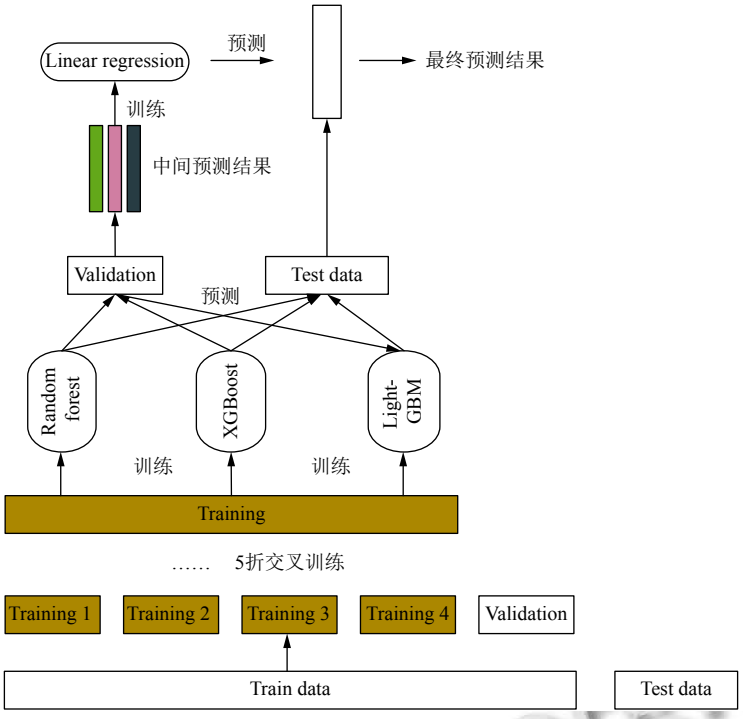


图 5 Stacking 模型融合架构图

Stacking 框架下基于多模型融合的光伏发电功率预测方法训练流程大致如下:

(1) 在数据预处理的基础上使用随机森林、XGBoost 以及 LightGBM 算法结合递归特征消除法对输入特征进行特征选择, 删除冗余特征.

(2) 划分原始数据集, 使用交叉验证方式, 优选各个模型的最优超参数.

(3) 使用划分后的数据集对 Stacking 中的第一层预测算法分别训练, 并输出预测结果, 生成新的数据集.

(4) 使用新生成的数据集, 对 Stacking 中第二层算法进行训练, 基于多模型融合的 Stacking 集成学习算法训练完毕.

3 光伏电场数据算例分析

3.1 光伏电场数据预处理

本文采用的数据集为国能日新企业下某光伏电站提供的 2016 年 4 月 1 日至 2018 年 4 月 30 日之间的连续气象历史数据和光伏电站的输出功率数据. 监测的数据每天从 0:00–23:45, 每 15 分钟进行一次数据采集, 将其中一些奇异数据剔除, 减少奇异数据在训练模型时对训练结果的影响, 共得到 66 860 条样本数据. 数据前 5 行信息如图 6 所示, 初始数据特征主要包含时间、辐照度、风速、风向、温度、湿度、压强、实际辐照度信息, 实际功率为标签值. 在模型建立过程中, 将 66 860 条样本数据随机切分出 70%, 即 46 802 条数据作为训练集, 进行模型构建, 其

余 30% 的 20 058 条数据作为测试集, 用于对模型的检验.

光伏发电主要是依靠太阳能, 因此时间特征应该作为一个重要的特征进行挖掘. 由于初始数据的时间格式无法直接作为模型的输入, 因此将时间属性进行特征提取, 切分出月份、天、小时、分钟这些有用的特征属性. 数据采集是以 15 分钟为单位进行的, 提取出的 minute 属性是 4 个离散值, 对其进行独热编码, 并剔除原来的时间以及 minute 特征, 得到的初步预处理数据集及统计信息如表 1 所示.

	时间	辐照度	风速	风向	温度	压强	湿度	实际辐照度	实际功率
0	2016/4/1 0:15	−1.0	−0.707 547	251	−0.090 909	−0.030 303	−0.157 895	0.0	−0.019 333
1	2016/4/1 0:30	−1.0	−0.716 981	250	−0.107 071	−0.030 303	−0.136 842	0.0	−0.021 000
2	2016/4/1 0:45	−1.0	−0.726 415	248	−0.123 232	0.030 303	−0.094 737	0.0	−0.022 000
3	2016/4/1 1:00	−1.0	−0.735 849	244	−0.135 354	0.030 303	−0.073 684	0.0	−0.022 000
4	2016/4/1 1:15	−1.0	−0.754 717	241	−0.147 475	0.030 303	−0.052 632	0.0	−0.022 000

图 6 光伏电场初始数据部分图示

表 1 初步预处理后数据集的描述性统计

变量	平均数	标准差	方差	最小值	最大值
辐照度	−0.576 609	0.559 381	0.312 907	−1	1
风速	−0.664 325	0.267 507	0.071 559	−1	1
风向	163.364 893	95.693 066	9157.162	0	359
温度	−0.006 276	0.380 029	0.144 422	−1	1
压强	0.110 206	0.284 216	0.080 778	−1	1
湿度	−0.088 533	0.442 732	0.196 011	−1	0.978 947
实际辐照度	236.341 804	342.082 201	117 020.2	−0.02	1303.11
hour	11.487 93	6.921 737	47.910 44	0	23
month	6.121 87	3.379 429	11.420 54	1	12
day	15.598 059	8.711 712	75.893 92	1	31
minute_0	0.249 978	0.433 003	0.187 491	0	1
minute_15	0.249 978	0.433 003	0.187 491	0	1
minute_30	0.250 022	0.433 029	0.187 514	0	1
minute_45	0.250 022	0.433 029	0.187 514	0	1
实际功率	2.105 609	2.989 093	8.934 676	−0.073 66	10.4853

3.2 光伏电场数据特征分析

为了充分挖掘气象数据对实际功率的影响, 本文采用了特征交叉的方式来扩充特征, 特征交叉是指通过将输入数据集的两个或多个特征进行相乘从而构造非线性特征的一种方式, 非线性特征针对于非线性预测模型, 其会有更好的增益贡献. 通过特征组合的方式增加特征的维度, 以求得更好的训练效果. 因此本文对风速、风向、温度、压强、湿度这几个特征进行了特征交叉, 构建了温度*温度、湿度*湿度、压强*压强、温度*湿度、湿度*压强、温度*压强、温度*湿度*压强、风向*风速这些新的特征, 同时对所有的特征进行了相关性分析, 绘制了如图 7 所示的特征相关性热度图, 该图以颜色的深浅表征特征与特征以及特征与标签之间的相关性. 可以看出, 实际功率与辐照度、实际功率与实际辐照度之间存在较高的相关性, 分别为 0.83 和 0.84. 同时辐照度和实际辐照度之间也有 0.89 的相关性, 由于实际辐照度对实际功率影响更大一些, 故删除辐照度特征, 选用实际辐照度作为特征输入. 对于特征交叉构造的新特征, 该图也表明其对实际功率预测也有很重要的作用.

通过数据预处理以及特征分析后, 利用基于模型的递归特征消除法进行特征选择, 本文分别利用随机森林、XGBoost 以及 LightGBM 3 种实现模型融合的基学习器通过交叉验证结合枚举的方式执行递归特征消除, 以此来选择最佳数量的特征, 并且在 Stacking 模型融合的过程中各自以最优特征数进行第一层预测模

型训练. 该方式具体思想为: 对于一个数量为 d 的特征属性集合, 其所有的子集的个数是 2^d–1. 指定一个学习算法, 首先通过该算法计算所有特征的重要性排名, 然后依据特征重要性得分依次构造出特征数目为 1 至 d 的所有特征子集, 并计算数据在所有特征子集上的交叉验证误差, 最后选择平均误差最小的那个子集作为所挑选的特征数量. 图 8 中 (a)、(b)、(c) 分别为基于随机森林、XGBoost 以及 LightGBM 在特征选择过程中各个特征子集的交叉验证均方误差的曲线变化图.

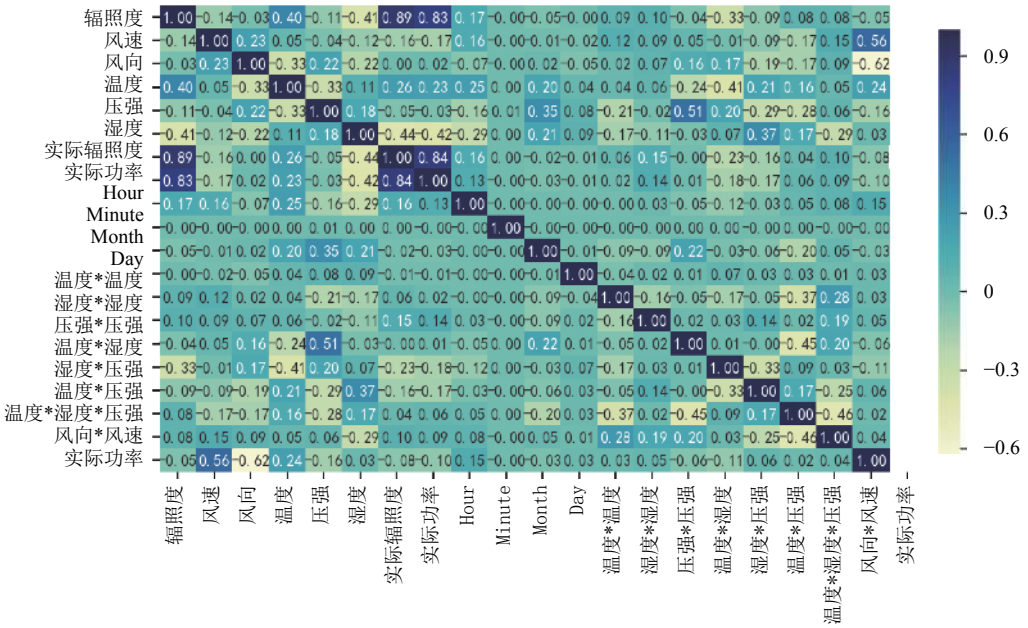


图 7 特征相关性热度图

通过图 8 可以看出, 不同的机器学习模型在建模过程中对最优特征的选择是不同的. 基于 Bagging 集成思想的随机森林选出的最佳特征个数是 10 个, 5 折交叉验证的均方误差 MSE 的平均值为 0.2785, 而基于 Boosting 集成思想的 XGBoost 和 LightGBM 选择的最佳特征个数分别为 19 和 15 个, 5 折交叉验证的 MSE 均值分别为 0.2071 和 0.1783. 对于本文的光伏电场数据, XGBoost 和 LightGBM 模型相较于随机森林模型有更好的预测效果, 并且 LightGBM 在高精度的预测前提下, 模型构建时间也显著缩短.

3.3 各单模型的超参数选择与预测性能分析

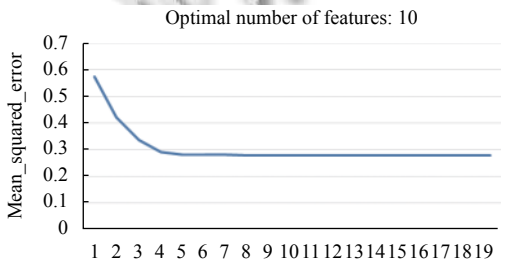
为了使 Stacking 模型融合的性能达到最优, 本节在结合 3.2 节特征选择结果的基础上, 对随机森林、XGBoost 和 LightGBM 3 个基学习器模型的学习能力进行分析. 针对各个基学习器采用网格搜索加交叉验证的方式进行超参数择优, 首先将数据集划分为训练集和测试集, 然后将划分后的训练数据进一步分为训练集与验证集, 通过交叉验证的方式, 分别观测使用不同超参数集训练后模型在验证集的预测效果, 从而选择各个模型的最优超参数集. 将构建好的各个单模型在测试集上进行预测分析, 并以决定系数 R^2 和均方误差 MSE 作为模型性能的评价标准. 各模型超参数集以及单模型预测性能如表 2 所示.

通过上表可以看出, 超参数优化后的随机森林模型 R^2 为 0.9465, 均方误差 MSE 为 0.2785, 而 XGBoost 与 LightGBM 在参数优化后的 R^2 分别为 0.9587 和 0.9632, MSE 分别为 0.2071 和 0.1783. 基于 Bagging 集成思想的随机森林和基于 Boosting 集成思想下的 XGBoost 与 LightGBM 作为 Stacking 模型融合的第一层基学习器, 其性能较好, 且 Boosting 算法模型的优越性高于 Bagging 算法模型.

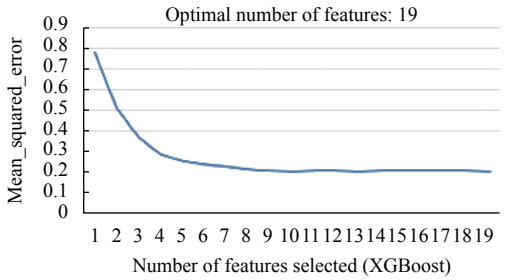
3.4 Stacking 模型融合预测性能及改进机制分析

为了验证 Stacking 集成学习模型的预测性能, 首先对 46 802 条训练集平均分为 5 份, 分别用第一层

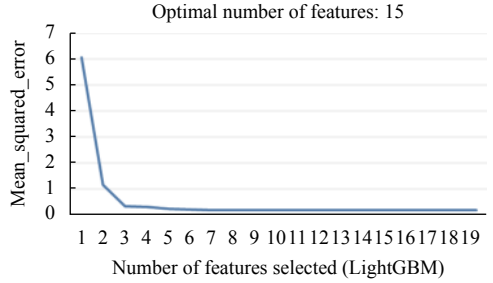
的 3 种基学习器执行 5 折交叉验证过程, 在每一次交叉验证中, 使用 4 份训练集对基学习模型进行构建, 并对剩余的一份验证集进行预测, 同时对 20 058 条测试集进行预测, 那么在 5 折交叉验证完成后, 每个基学习器模型会产生和训练集样本数量相同的一列数据, 3 个基学习器则会最终产生 3 列对训练数据预测后的数据集 *A*. 然后将每个基学习器模型在 5 次交叉验证后对测试集预测出的 5 列数据按行求均值, 即 3 个基学习模型最后会产生 3 列对测试集预测后的数据集 *B*. 最后则利用第二层的 Linear Regression 模型对数据集 *A* 和最初训练集的标签 (原始实际功率) 构成的数据集进行模型构建, 利用构建好的融合模型对数据集 *B* 进行预测, 得到最终的 Stacking 模型融合后的结果.



(a) 基于随机森林的特征选择 (最小 MSE=0.2785)



(b) 基于 XGBoost 的特征选择 (最小 MSE=0.2071)



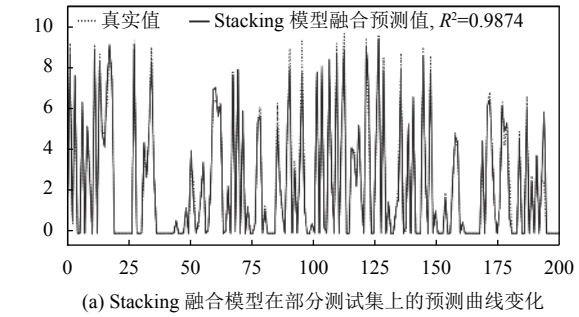
(c) 基于 LightGBM 的特征选择 (最小 MSE=0.1783)

图 8 特征子集寻优过程中交叉验证的均方误差曲线图

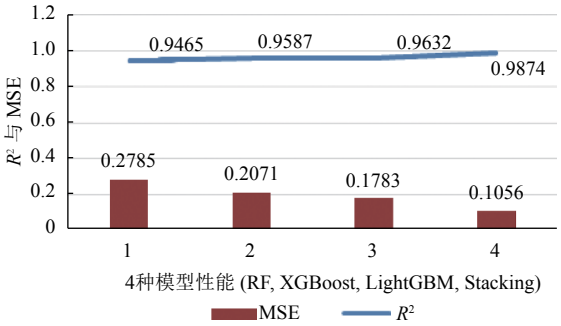
图 9(a) 比较了部分测试集在 Stacking 融合模型上的预测曲线变化情况, 图 9(b) 给出了 4 种模型在 R^2 和 MSE 上的性能评价值. 可以看出, Stacking 模型融合后的实际功率预测值无论是在拟合优度还是均方误差方面都有了较大的改善, R^2 达到了 0.9874, MSE 达到了 0.1056, 这表明本文提出的对随机森林、XGBoost 和 LightGBM 以 Stacking 集成学习方式进行模型融合, 在对光伏电场输出功率的预测方面, 具有一定的实际意义, 可为电网调度提供有益的参考.

表 2 各模型超参数集以及单模型预测性能

模型名称	超参数	预测性能	
		R^2	MSE
随机森林	n_estimators=175, max_depth =5, min_samples_split =2, min_samples_leaf=1	0.9465	0.2785
XGBoost	max_depth=6, learning_rate=0.02, n_estimators=160, min_child_weight=1, gamma=0.1, subsample=0.7	0.9587	0.2071
LightGBM	num_leaves=50, learning_rate=0.02, bagging_fraction=0.7, bagging_frequency=5	0.9632	0.1783



(a) Stacking 融合模型在部分测试集上的预测曲线变化



(b) 4种模型在 R^2 和 MSE 上的性能评价

图 9 Stacking 模型融合的性能曲线图

从理论层面分析 Stacking 集成模型优于单模型的

原因, 是因为 Stacking 模型集成了多样化的预测算法, 能够充分利用各个算法从不同的数据空间和结构来观测数据, 从而充分发挥各个算法自身优势, 摒弃了各个算法中预测效果较差的环节. 此外, 考虑到光伏发电功率预测模型训练过程中的假设空间往往很大, 可能有多个假设在训练集上达到同等性能, Stacking 集成学习的方式可以有效减少单一模型泛化性能不佳的风险. 另一方面, 从模型优化角度来看, 单一模型训练的优化过程中, 模型往往会有陷入局部最小点的风险, 有的局部极小点所对应的模型泛化性能可能较差, 而通过多个基学习器训练之后进行结合, 可有效减少陷入局部极小点的概率. 因此, 采用基于 Stacking 集成学习方式的多模型融合后预测精度有所提升.

就 Stacking 模型融合的改进机制而言, 元学习器的训练集是由基学习器的输出产生的, 直接使用基学习器的训练集来产生次级训练集的话, 可能会产生严重的过拟合. 为了防止数据被双层学习器重复使用而造成过拟合效应的发生, 本文在模型融合的过程中, 每个基学习器都对训练集执行 5 折交叉验证过程, 使用一个数据块作为验证集, 对应的其余 4 个数据块作为测试集, 在每折交叉验证完成后, 利用基学习器对验证集进行预测, 即每个基学习器在 5 折交叉验证结束后, 都会产生和原始训练集数量相同的新的数据集. 实现了所有数据从输入特征到输出特征的特征变换, 且元学习器的训练集完全来自各个基学习器的预测输出数据, 这使得元学习器能充分结合各个基学习器的模型优势来完成模型构建, 以提升模型融合的效果. 同时, XGBoost 和 LightGBM 这两类基学习器能够充分挖掘输入信息的数据内部特征, 对连续型和离散型特征都有较好应用, 而随机森林模型可以在最大程度上关注影响较大的几种特征类型, 能够充分发挥重要特征的作用. 因此, Stacking 模型融合的 3 个基学习器可以实现优势互补, 确保元学习器的输入是质量较高且无冗余特征的数据集, 这些数据集通过 Linear Regression 这一泛化能力较强的元学习器进行训练后, 可以从整体上提升 Stacking 模型的预测能力.

4 结论

为了获得更加理想的光伏发电功率预测结果, 本文建立了以历史光伏电场输出功率数据和环境气象为关键因素的基于 Stacking 模型融合的光伏发电功率预测模型. 该模型分为两层构建, 第一层基学习器分别选取预测性能较优的以 Bagging 算法为代表的随机森林和以 Boosting 算法为代表的 XGBoost、LightGBM 模型, 并在第一层进行了各个基学习器模型的超参数调优, 以充分发挥各个模型的优势, 第二层选取稳健性较好的 Linear Regression 模型进行最终的 Stacking 集成融合. 通过光伏电场数据集的试验表明, 本文建立的融合模型可以精确地预测光伏电场输出功率, 提高了光伏发电的预测精度, 对于工程上的光伏功率发电预测及建模有一定的实用意义.

在今后的工作中, 将进一步针对以下问题开展深入探讨, Stacking 的框架设计比较复杂, 对于基模型要训练多次, 即使在训练 Stacking 模型的时候对每个基学习器减少若干数据量, 计算时间仍然较长. 因此, 未来研究中有必要布置分布式计算的相关环境, 对不同的基模型分别建模, 采用分而治之的思想, 有效减小算法的时间复杂度.

参考文献

1 吕鑫, 刘天予, 董馨阳, 等. 2019 年光伏及风电产业前景预测与展望. 北京理工大学学报 (社会科学版), 2019, 21(2): 25–29.

2 瞿谊. 风光伏发电预测技术的发展. 数码世界, 2018, (4): 274. [doi: 10.3969/j.issn.1671-8313.2018.04.239]

3 张玉, 黄睿, 张振涛, 等. 基于克里格模型的光伏发电量预测. 热力发电, 2017, 46(4): 27–32. [doi: 10.3969/j.issn.1002-3364.2017.04.027]

4 Guo HP, Wu SH, Wang ZQ, *et al*. Linear regression for forecasting photovoltaic power generation. Applied Mechanics and Materials, 2014, 494–495: 1771–1774. [doi: 10.4028/www.scientific.net/AMM.494-495.1771]

5 李芬, 宋启军, 蔡涛, 等. 基于 PCA-BPNN 的并网光伏电站发电量预测模型研究. 可再生能源, 2017, 35(5): 689–695. [doi: 10.3969/j.issn.1671-5292.2017.05.009]

6 张玉, 莫寒, 张烈平. 基于模糊支持向量机的光伏发电量预测. 热力发电, 2017, 46(1): 116–120. [doi: 10.3969/j.issn.1002-3364.2017.01.116]

7 肖俊明, 韦学辉, 李燕斌. 基于 BP 神经网络和遗传算法的光伏功率预测的研究. 计算机测量与控制, 2015, 23(2):

392–393, 405. [doi: 10.3969/j.issn.1671-4598.2015.02.017]

8 张雨金, 杨凌帆, 葛双冶, 等. 基于 Kmeans-SVM 的短期光伏发电功率预测. 电力系统保护与控制, 2018, 46(21): 118–124. [doi: 10.7667/PSPC171595]

9 傅美平, 马红伟, 毛建容. 基于相似日和最小二乘支持向量机的光伏发电短期预测. 电力系统保护与控制, 2012, 40(16): 65–69. [doi: 10.3969/j.issn.1674-3415.2012.16.011]

10 Sun W, Zhang X. Application of self-organizing combination forecasting method in power load forecast. Proceedings of 2007 International Conference on Wavelet Analysis and Pattern Recognition. Beijing, China. 2007.

11 高相铭, 杨世凤, 潘三博. 基于 EMD 和 ABC-SVM 的光伏并网系统输出功率预测研究. 电力系统保护与控制, 2015, 43(21): 86–92.

12 王守相, 张娜. 基于灰色神经网络组合模型的光伏短期出力预测. 电力系统自动化, 2012, 36(19): 37–41.

13 单英浩, 付青, 耿炫, 等. 基于改进 BP-SVM-ELM 与粒子化 SOM-LSF 的微电网光伏发电组合预测方法. 中国电机工程学报, 2016, 36(12): 3334–3342.

14 Yang XY, Jie R, Hong Y. Photovoltaic power forecasting with a rough set combination method. Proceedings of the 2016 UKACC 11th International Conference on Control. Belfast, UK. 2016.

15 Li Q, Sun YQ, Yu YJ, *et al*. Short-term photovoltaic power forecasting for photovoltaic power station based on EWT-KMPMR. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(20): 265–273.

16 李多, 董海鹰, 杨立霞. 基于 EMD 与 ELM 的光伏电站短期功率预测. 可再生能源, 2016, 34(2): 173–177.

17 张立影, 刘智昱, 孟令甲, 等. 基于小波变换和神经网络的光伏功率预测. 可再生能源, 2015, 33(2): 171–176.

18 史佳琪. 区域综合能源系统供需预测及优化运行技术研究 [博士学位论文]. 北京: 华北电力大学 (北京), 2019.

19 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. 中国电机工程学报, 2019, 39(14): 4032–4042.