基于集成模型的光伏发电预测精度对比研究

欧耀春，李志生*

广东工业大学土木与交通工程学院，广东广州 510006

摘要： 短期光伏发电预测对改善电力系统并网的稳定性具有重要意义。为了提高短期光伏发电量预测的准确性，提出了一种基于集成模型的预测方法。首先，使用四分位法检测原始数据是否存在离群值，然后结合特征贡献度和 Pearson 相关系数，筛选出影响短期光伏发电量的多个特征变量。其次，构建集成模型的结构，并采用 k-fold 交叉验证方法对初级学习器和次级学习器中的模型进行训练。将初级学习器的结果作为次级学习器的新特征。最后，使用实际数据集验证了该方法的有效性。仿真结果表明，对离群值具有较高兼容性的初级学习器，有利于提高光伏发电量的预测精度。两种特征筛选方法的结合使用，比单一的特征提取方法，可以更有效地提取原始数据中的信息，提高模型的泛化能力。并且集成模型可以实现比单个模型更高的预测精度。

关键词： 光伏发电；太阳辐射因素；特征工程；集成模型；交叉验证

DOI: 10.57237/j.res.2023.02.007

Comparative Study on the Precision of Photovoltaic Power Generation Prediction Based on Integrated Models

Ou Yaochun, Li Zhisheng*

School of Civil and Transportation Engineering, Guangdong University of Technology, Guangzhou 510006, China

Abstract: Short-term photovoltaic power generation prediction is of great significance for improving the stability of power system grid connection. In order to improve the accuracy of short-term photovoltaic power generation prediction, a prediction method is proposed based on integrated models. Firstly, the quartile method is used to detect whether there are outliers in the raw data, and then multiple feature variables that affect short-term photovoltaic power generation are selected by combining feature contribution and Pearson correlation coefficient. Secondly, construct the structure of the integrated model and use the *k-fold* cross validation method to train the models in the primary and secondary learners. It is a new feature of the secondary learner with the results of the primary learner. Finally, the effectiveness of this method was verified using actual datasets. The simulation results show that a primary learner with high compatibility with outliers is beneficial for improving the prediction accuracy of photovoltaic power generation. The combination of two feature selection methods can more effectively extract information from the original data and improve the generalization ability of the model compared to a single feature extraction method. Moreover, the integrated model can achieve higher

*通信作者：李志生, Chinalzs@sina.com

收稿日期: 2023-07-15; 接受日期: 2023-09-06; 在线出版日期: 2023-09-08

prediction accuracy than a single model.

Keywords: Photovoltaic Power Generation; Solar Radiation Factors; Feature Engineering; Integrated Model; Cross Validation

1 引言

光伏发电是一种清洁的可再生能源，通过将太阳能转换成电能，在满足能源消耗的同时，可以减少使用化石能源，对减缓温室气体和空气污染物的排放具有显著作用。根据中国光伏行业协会提供的数据，2019 年中国新增光伏并网装机容量达到 30.1GW，累计光伏并网装机量达到 204.3GW，新增和累计光伏装机容量仍继续保持全球第一。但随着光伏发电产业规模的不断扩大，光伏发电具有明显的间接性和波动性等特点，使得光伏发电并网的稳定性对于电力供应显得十分重要。高精度的短期预测技术是提高光伏发电并网稳定性的重要手段。但光伏发电容易受到环境因素和气候因素的影响，使得发电的短期预测受到复杂的耦合作用[1]。根据中国气象局提供的上海地区多年太阳辐射观测数据，评估当地太阳能资源，结果表明上海地区太阳能资源丰富、稳定度较高，适合建设光伏电站[2]。但上海市处于中国太阳能资源的中等地区，相对于辐射资源丰富的一类地区，巨大的市场需求使得光伏发电产业发展迅速的同时，也受到东南地区季风气候的严峻挑战。而发电功率和有效的太阳辐射是决定光伏发电量的直接因素。气候因素通过影响太阳辐射量和光伏实际发电功率来间接影响光伏发电量。由于光伏发电功率可以通过实际发电量来得知，因此研究太阳辐射量与实际发电量之间的关系，从而提高光伏电量的预测精度，具有至关重要的意义。

近年来，许多学者提出了用于短期光伏发电预测的模型，并在预测精度上得到较大的提高。常用的光伏发电预测方法可以概括为物理原理法和机器学习法。(1)物理方法主要是根据太阳能转换为电能的原理，建立光伏电能转化系统的物理模型[3, 4]。由于物理预测的方法主要的依据是假设性理论，需要确定大量的参数和满足假设性的前提，因此无法深入研究光伏发电量受到内部系统和外部环境因素的双重作用对发电量的影响，难以直接应用于提高光伏发电量预测精度。(2)机器学习法通过使用大量的历史数据集来挖掘相关特征变量与光伏发电量之间的非线性关系，然后通过学习各个因素与光伏发电量之间的变化规律，得到良好的预测性能。目前，基于机器学习的主要预测方法有：支持向量机（SVM），决策树型和神经网络型三类模型。SVM 算法在解决非线性、高维以及局部极小等实际问题中表现出特有的优势，但无法支持庞大的数据量的运算[5]。决策树类型的算法主要通过多个决策树的投票机制，进行快速，高准确率的预测。常用的决策树类模型有极端梯度提升决策树（XGBoost）模型和梯度提升决策树（GBDT）模型。GBDT 模型对原始数据的兼容性较大，有预测速度快且精度高的优点，但是需要仔细调参，有时易于过度拟合[6, 7]。神经网络模型通过长时间运算和仔细调节参数，来构建复杂的模型，但是通常在处理具有相似特征的数据表现出最好的性能[8]。

集成学习是为了提升自动决策系统下预测结果的准确性，目前包括 bagging、boosting 和 stacking3 种学习方法[9]。Stacking 是通过训练多个初级学习器获得相应输出结果，并将其结果作为次级学习器的输入，来寻找集成学习的最优组合。其中，初级学习器作的结果，可以采取的平均法和投票法来输入次级学习器，以达到矫正偏差和方差的目的，并通过交叉验证提升训练效率，从而降低模型的过拟合风险，提升模型的精度[10]。集成模型可以有效提高各种机器学习模型的准确性。例如，唐科等，使用 Stacking 集成方法将 XGBoost、轻量化梯度促进机（LightGBM）以及随机森林（RF）算法集成，构建一个多模型集成的气态亚硝酸预测模型[11]。丁斌等，探索了 RF 与多层神经网络 MLP 的模型集合对 PM2.5 预测精度的提高作用[12]。金晓等，将神经网络，SVM 和遗传编程结合到一个集成模型中用以提高能耗预测精度[13]。杨荣新等，使用 Stacking 集成方法将以 XGBoost、LightGBM、Random Forest 为初级学习器，Linear Regression 为次级学习器，构建一个多模型集成的光伏发电功率预测模型[14]。史佳琪等人，采用 Stacking 集成学习方法融合多个算法模型，结果表明具有强学习能力的次级学习器能显著提高负荷预测精度[15]。然而，模型集成技术在光伏功发电量预测领域中的应用较少。因此，探究集成模型应用于提高光伏发电量预测的准确性，具有十分显著的意义。

为了提高光伏发电量的预测精度，提出了一种基于集成模型的光伏发电量预测方法。本文的主要贡献如下：

(1) 使用四分位数法检验原始数据集中的离群值，从而选择合适的初级学习器来提高集成模型的数据兼容性。

(2) 以特征贡献度为主、皮尔逊系数为辅的特征筛选

方法以测量输入变量与光伏发电量之间的相关性，从而为预测模型选择输入特征变量。

(3) 该模型集成技术可以充分吸收不同算法的优点，弥补单一算法的缺点，大大提高了预测精度。

本文的其余部分安排如下：第 2 节简要介绍了模型集成技术的框架。第 3 节解释了用于集成模型的短期光伏发电预测模型的原理。第 4 节讨论了仿真和结果。结论在第 5 节中描述。

2 模型集成的框架

模型集成技术分为过程集成和结果集成两种。本文使用的是过程集成技术，通过在预测过程中，综合各个子模型的优点，来提高预测结果的精度。

本文将 SVM、GBDT、MLP 三个常用的模型进行 Stacking 集成，通过 k-fold 交叉验证进行模型训练，以预测光伏发电量。其原理如图 1 所示。

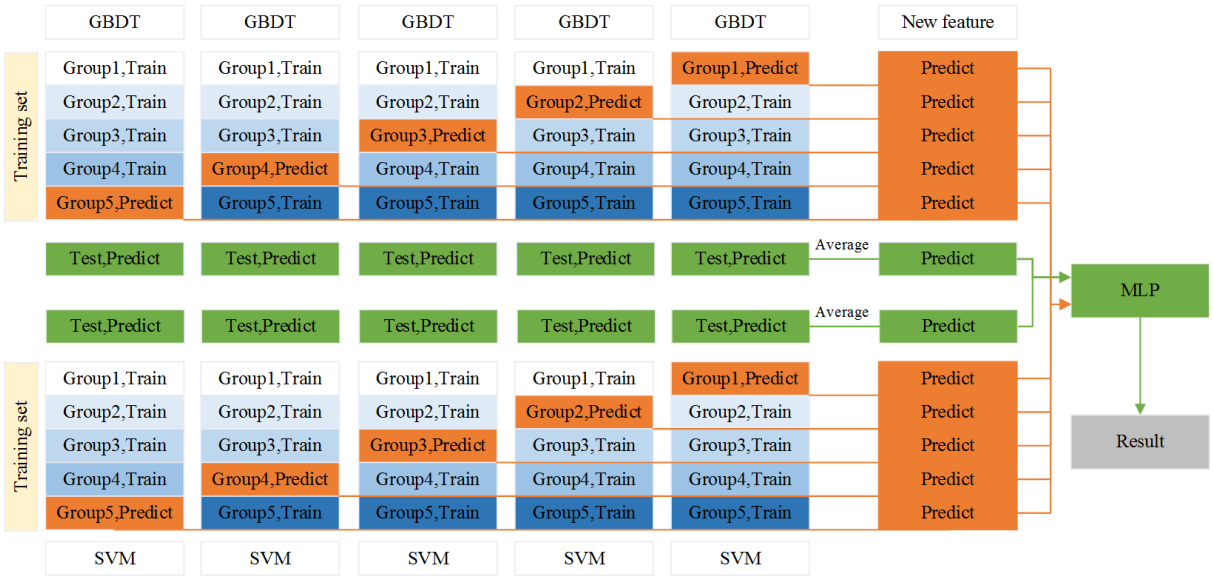

图 1 模型框架原理

学习器模型的选择和次级学习器的输入特征方法分别对于 Stacking 集成模型的数据兼容性和泛化性能有很大的影响[16]。以选择 MLP 作为次级学习器，SVM 和 GBDT 作为初级学习器为例，阐述 Stacking 集成方法的原理。首先，使用 SVM 和 GBDT 对所有原始特征进行基于交叉验证的训练，根据训练集和测试集对于训练后的 SVM 和 GBDT 进行性能测试，进行性能优化，提高对原始数据的信息提取。初级学习器中 SVM 和 MLP 的预测发电量的平均值和交叉验证中的预测值分别作为 MLP 的新输入特征。最后，用新的输入特征对 MLP 进行训练，并将检查训练后的 MLP 的性能，进行性能优化。测试集上的 MLP 模型的预测发电量即为最终的预测结果。

上述三个模型的初级学习器和次级学习器的子模型可以根据测试结果进行调整。SVM，MLP 和 GBDT 的训练方法相同。k-fold 交叉验证的训练过程如图 1 所示。

表 1 辐射因素

变量名称	简称	变量名称	简称
发电量	GE	总辐射日最大辐照度	TRE
总辐射量	TR	总辐射日最大辐照度出现时间	TRMIOT
净辐射日总量	TNR	净辐射日最大辐照度	MNRI
散射辐射日总量	TSR	净辐射日最大辐照度出现时间	MITNR
水平面直接辐射日总量	THDR	垂直面直接辐射日最大辐照度	DRMI
反射辐射日总量	TRR	垂直面直接辐射日最大辐照度出现时间	MITDR
垂直面直接辐射日总量	TVDR		

首先，原始数据集平均分为 K 个组。选择第 k 组数据作为测试集样本，其余 k-1 组数据用作训练集。训练集训练结果用于预测测试集的光伏发电量。然后，选择第（k-1）组数据作为测试集样本，其余 k-1 组数据用作训

练集。训练集训练结果用于预测第（k-1）组的光伏发电量。第三，重复以上步骤，将有 k 个预测数据组，这将构成训练集的预测能力。另外，训练集的预测能力将用作新的特征来训练下一个模型。最后，基于这 k 个训练集测试了 k 个训练过的模型的性能。测试集的平均光伏发电量是模型的结果。

3 方法

3.1 数据检验

所使用的数据集来自中国气象数据网和上海市某发电站的日发电量数据。数据集包含 2017-2019 的 1095 天样本。原始数据集包括光伏日发电量和太阳辐射变量，如表 1 所示。

为了直观地分析数据的是否存在离群值，可视化了 9 个变量，如图 2 所示。最小值是数据的“下限”，最大值是数据的“上限”，位于这些定义范围之外的任何数据都可以视为离群值。光伏发电量的数据检验中，整体的发电量、季度发电量都没有离群值，只有月发电量中的 7 月份出现少量的离群值。在原始变量中，只有 THDR，TRMIOT，MITNR 和 MITDR，有异常值，其中 THDR 只有少量离群值。离群值的存在影响模型的训练和预测性能。为了消除这些离群值的负面影响，有必要对变量的离群值并进行一些处理。

消除离群值对预测精度的影响，常用的做法为删除离群值、使用平均值或中位数代替异常值，但是这两种做法会在不同程度上造成模型对原始数据提取信息损失[17]，降低模型的泛化性。此外，还可以通过对数变换和利用对离群值鲁棒性较高的预测模型[18]，如 RT、GBDT、SVM 等。这两种方法，不会产生信息损失。综合模型集成中包含 GBDT、SVM，所以本次研究中，采取模型对离群值高鲁棒性的方法来降低离群值对于模型精度的影响。

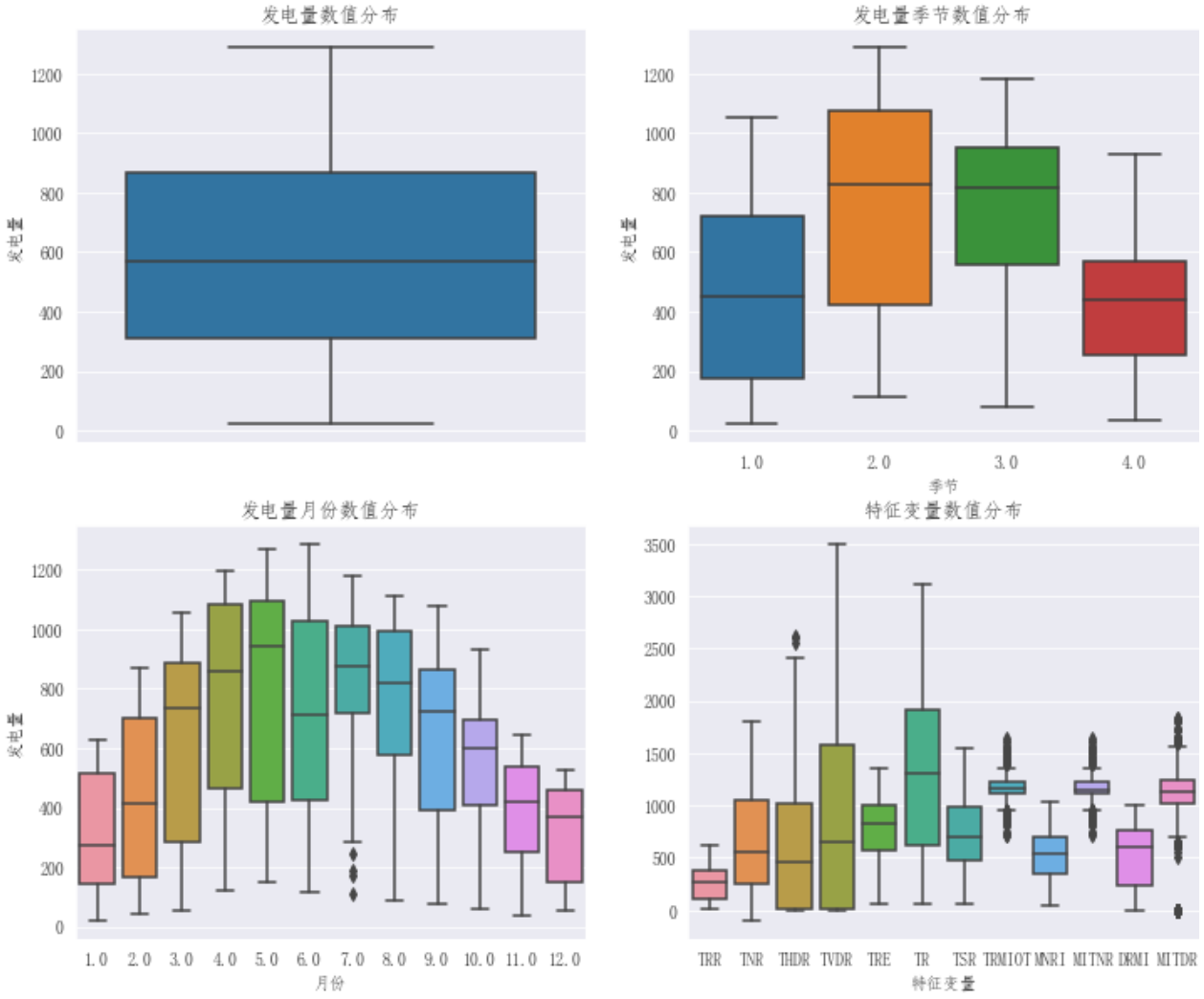

图 2 原始变量的检验

3.2 特征变量的选择

对于光伏发电而言，由于其影响因素多，相互间作用关系复杂，因此，需要探究挖掘影响光伏发电的显著性因素。特征工程是把原始数据转化为模型的训练数据的过程，其目的是最大限度的从原始数据中提取特征以供模型使用，一般包括特征构建、特征提取和特征选择三个部分[19]。特征选择就是从原始特征中选取一些最有效的特征来降低维度，提高模型泛化能力减低过拟合的过程，主要目的是剔除掉无关特征和冗余特征，选出最优特征子集。对于光伏发电量而言，从两个方面考虑来选择特征变量，特征变量是否发散以及与目标的相关性。当一个特征变量不发散，方差接近于 0，样本数据的时间维度在这个特征变量上没有明显的差异，那么这个特征变量对于目标变量的解释接近于 0。而目标相关性高的特征，应当优选选择。由于辐射因素对光伏发电的影响机理复杂，往往很难通过线性相关性来判断变量对于光伏发电量的影响程度大小。因此，本文使用 RF 算法对光伏发电量的特征变量，进行重要度分析。对于光伏发电的特征工程，从非线性相关因素入手，筛选光伏发电量贡献较大的辐射变量作为预测模型的主要输入变量。然后考虑与光伏发电量具有强线性相关性的因素，作为补充的输入变量。

特征重要度是一种利用训练好的有监督分类器来选择特征的技术。当我们训练分类器时，我们计算每个参数以创建分割；我们可以使用这个度量作为特征选择器。随机森林由于其相对较好的准确性、鲁棒性和易用性，用作本文重要度分析的训练分类器。随机森林由许多决策树组成。决策树中的每个节点都是一个基于单个特征的条件，其设计目的是将数据集分割成两个，以便相似的响应值最终出现在相同的集合中。选择（局部）最优条件的度量叫做杂质。而对于回归树，它是方差。因此，当训练一棵树时，可以通过每个特征减少的树中加权杂质的多少来计算。对于森林，可以对每个特征的杂质减少量进行平均，并根据该方法对特征进行排序。

首先我们使用基尼指数（Gini index）作为评价指标来衡量。特征贡献度用 VIM 来表示，基尼指数用 GI 来表示。第 j 个特征的 GI 评分 $VIM_j^{(Gini)}$，即第 j 个特征在 RF 所有决策树中的平均改变量。

基尼指数的计算公式为：

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \quad (1)$$

K 表示由 K 个类别，p_{mk} 表示节点 m 中类别 k 所占的比例。

特征 X_j 在节点 m 的重要性为：

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \quad (2)$$

GI_l 和 GI_r 分别表示分枝后两个新节点 Gini 指数。

特征 X_j 在第 i 棵树的重要性为：

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \quad (3)$$

假设 RF 中共有 n 棵树：

$$VIM_j^{(Gini)} = \sum_{i=1}^{n} VIM_{ij}^{(Gini)} \quad (4)$$

把所有求得的重要性评分做归一化处理即可得到每个特征变量的评分：

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \quad (5)$$

由于在原始数据中，很难直接比较变量与光伏发电量之间的线性相关的强弱，因此，通过 Pearson 相关系数进行统计检验，判断两个变量之间相关关系是否显著，其取值在[-1，1]之间，绝对值大小表示两变量间相关关系程度的强弱，越接近 1，表明两变量间相关程度越高。通过 Pearson 相关系数，衡量各个气象因素和光伏发电量相关性的强弱，作为筛选影响光伏发电量的关键气象因素的依据。Pearson 相关系数的公式如下：

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (6)$$

其中$\bar{X}$是 X 的平均值，$\bar{Y}$是 Y 的平均值。

光伏发电量由辐射强度、实际转换功率所决定。光伏发电所受到环境因素和气候因素的影响，都是通过影响这两个因素，使得发电量产生波动性和随机性的不稳定性。对于光伏发电而言，辐射分为有效辐射和无效辐射。总辐射量与光伏发电的关系探究，如图 3 所示。

通过对原始数据进行归一化处理后，光伏发电与总辐射量呈现高度吻合的变化趋势，光伏发电随着总辐射量的季节性变化，呈现高度一致的周期性。但是总辐射量中，包含净辐射和散射等多种辐射因素在内，仅仅通过总辐射量无法直接挖掘影响光伏发电与有效辐射之间的关系。因此，需要通过特征工程来挖掘影响光伏发电的重要辐射因素，以此来提高光伏发电预测的精度和泛化性。为了解决单一特征筛选方法对某些特征数据不敏感的问题[20]，因此，我们选择 Pearson 相关系数作为特征贡献度的补充特征筛选方法。

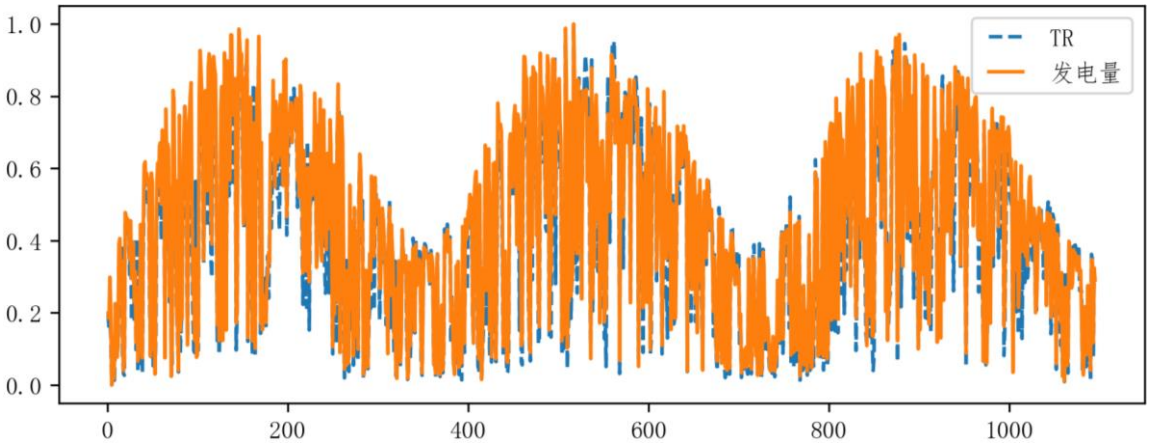

图 3 总辐射量与光伏发电的关系探究

特征重要度用于评估辐射因素与光伏发电量之间的非线性相关性的主要特征筛选方法，Pearson 相关系数用于评价辐射因素与光伏发电量之间的线性相关性的补充特征筛选方法。先用特征重要度来筛选特征重要度大于 0.2 的非线性特征变量，然后将剩余的变量通过 Pearson 相关系数进行筛选处线性关系大于 0.8 的强线性变量，筛选结果示于图 4。

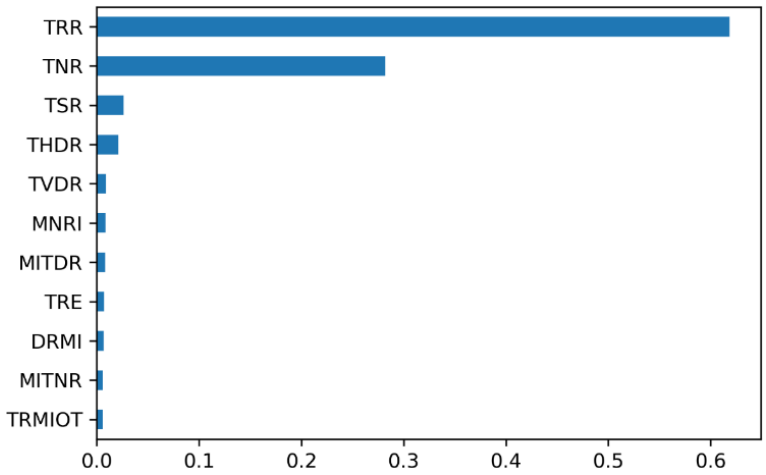

图 4 特征重要度排序（纵坐标的英文简称见表 1，因是英文软件输出，无法转为中文）

表 2 辐射因素与光伏发电量的 Pearson 相关系数

Variables	Pearson	Variables	Pearson	Variables	Pearson	Variables	Pearson	Variables	Pearson
DRMI	0.78	TSR	0.67	THDR	0.85	MITDR	-0.38	TVDR	0.80
TRE	0.81	TRMIOT	0.02	MNRI	0.78	MITNR	0		

从图 3 和表 2 得出以下结论：

(1) 辐射因素的特征贡献度呈现极端分布，存在显著的关键性影响变量。净辐射日总量和反射辐射日总量是光伏发电的重要的影响因素，反射辐射日总量对发电量起到至关重要的影响。

(2) 除了净辐射日最大辐照度与光伏发电呈现负相关，其余因素与光伏发电基本上都呈正相关关系。

(3) 综合上述的特征筛选结果，选择 TRR、TNR、THDR、TRE、TVDR 作为本文的输入特征变量。

各特征变量关系见图 5。

3.3 预测算法原理

GBDT 、MLP 和 SVM 网络被用作集成模型的三个子模型。本节主要介绍这三种算法的基本原理。

3.3.1 GBDT Model

梯度提升回归模型是一种基于 Boosting 思想的集成

算法，通过合并多个决策树来构建一个解释能力更加强的模型。梯度提升采用连续的方式构造决策树，每颗树都试图纠正前一棵树的错误，减少迭代计算时的错误积累。通过预剪枝、集成模型中决策树的数量和用于控制纠正前一步错误强度的学习率，来增加模型学习能力，使模型对于训练集上的错误有更强的纠正能力。梯度提升树通常使用深度很小的决策树，每颗树只能对部分数据做出很好的预测，随着决策树的数量增加，通过不断迭代计算，逐次减少上一次计算的残差，提高模型性能的同时，也加快预测速度。对于输入特征变量，GBRT 具有强大的包容性，对于数据的结构和数据的形式并不敏感。GBRT 的具体的运行过程如下：

(1) 初始化一棵仅包含根节点的决策树，并寻找到一个常数 Const 能够使损失函数达到极小值；

(2) 计算损失函数的负梯度值，用作残差的估计值 r_{mi}：

$$r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad (7)$$

(3) 利用数据集（x_i, r_{mi}）拟合下一轮基础模型，得到对应的 J 个叶子节点 R_{mj},计算每一个叶子节点 R_{mj} 的最佳拟合值：

$$c_{mj} = \mathrm{argmin}_c \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \quad (8)$$

(4) 进而得到第 m 轮的基础模型 $f\mathrm{m}(x)$，再结合前 m-1 轮基础模型，得到最终的梯度提升模型：

$$\begin{aligned} F_M(x) &= F_{M-1}(x) + f_m(x) \\ &= \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I \end{aligned} \quad (9)$$

(5) $F_M(x)$表示由 M 个基础模型在样本点 x_i 处的输出值 c_{mj} 之和。

3.3.2 MLP Network Model

MLP 是一种前向构造的人工神经网络，它将一组输入向量映射到一组输出向量。可以将其视为有向图，它由多个节点层组成，每个节点层都连接到下一层。经典 MLP 网络的结构，包括输入层，隐藏层和输出层。除输入节点外，每个节点都是具有非线性激活函数的神经元。

对于前向传播过程，隐藏层节点的输出可以表示为

$$z_k = f_1\left(\sum_{i=0}^{n} v_{ki} x_i\right), k = 1,2,\ldots q \quad (10)$$

其中 n 是输入层节点的数量，q 是隐藏层节点的数量，v_{ki} 是输入层的第 i 个节点和隐藏层的第 k 个节点之间的权重，而 f_1 是隐藏层的激活函数。

同样，输出层和隐藏层之间的关系可以表示为

$$y_j = f_2\left(\sum_{i=0}^{q} w_{jk} z_k\right), j = 1,2,\ldots m \quad (11)$$

其中 m 是输出层节点的数目，w_{jk} 是输出层的第 j 个节点和隐藏层的第 k 个节点之间的权重，f_2 是输出层的激活函数。

MLP 网络根据公式(10)和公式(11)将 n 维输入特征映射到 m 维输出变量y_j。通常，通过反向传播方法来训练 MLP 的网络权重。

3.3.3 SVM Model

SVM 主要是将低维线性不可分的空间转化为高维的线性可分空间，从而提高预测的准确性，该模型可以在一定程度上可以避免过拟合和陷入局部最优。为了使 SVM 回归模型具有更加强的泛化能力，需要加入松弛因子，使其满足：

$$|y_i - f(x_i)| - \xi^{(*)} \leqslant \varepsilon \quad (12)$$

$\xi^{(*)}$为松弛因子，ε为预测值和实际值的误差。

SVM 回归预测函数可以表示为：

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \cdot k(x, x_i) + b \quad (13)$$

α_i，α_i^*为 Lagrange 算子，$k(x, x_i)$为核函数，b 为位置的偏移量。

3.4 评价指标

为了验证集成模型的有效性，使用均方根误差（RMSE），平均绝对误差（MAE）和平均绝对百分比误差（MAPE）来评估不同方法的性能。它们的数学公式如下：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (14)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (15)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (16)$$

其中 n 是测试集中的样本数，$\hat{y}_i$是预测的光伏功率，y_i是实际光伏功率。

3.5 集成模型预测光伏功率的过程

使用模型集成技术进行短期光伏发电量预测的步骤如下：

(1) 使用基于四分位数法的箱型图用于找到每个

变量的离群值，将发电量的检验精确到月份发电量检验。

(2) 计算每个输入因子与光伏发电量之间的相关性。基于特征贡献度和 Pearson 相关系数的复合特征筛选方法，选择特征作为模型的输入。

(3) 使用归一化方法对选定的特征和光伏发电量进行归一化，以消除不同单位的负面影响。

(4) 训练集用于训练 MLP 网络，SVM 网络和 GBDT。然后，使用这三个模型来预测测试集的光伏发电量。

(5) 根据提出的集成框架对三个模型的预测结果进行组合，并输出最终的预测结果。

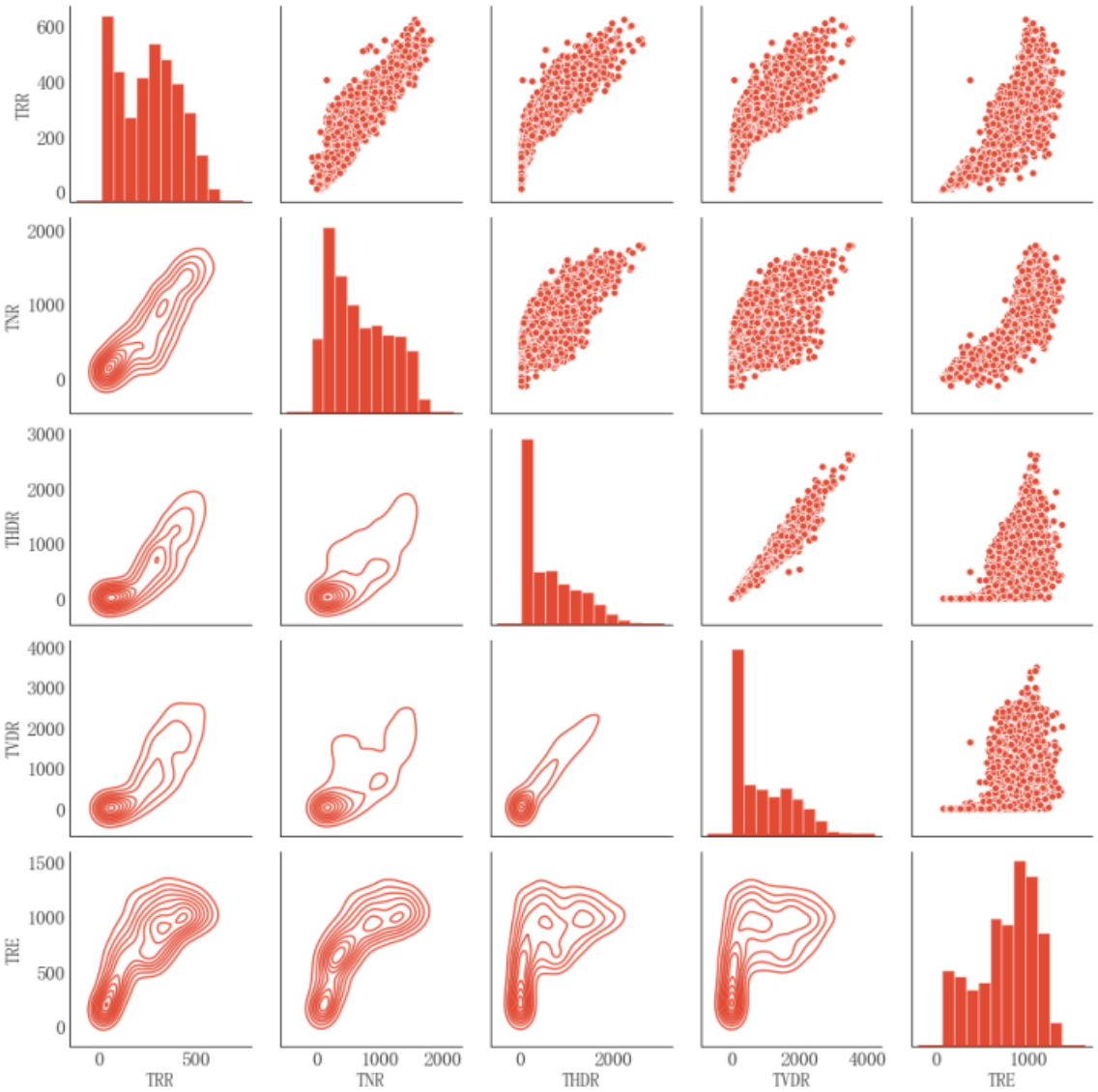

图 5 特征变量关系图

4 案例研究

4.1 测试设置

本研究以 1095 天数据作为研究样本。本次研究在 cpu：i7-8565U，gpu：1050Ti 4G，内存：16G 的计算机条件下，利用 Spyder（python3.7）实现。

在对模型的结构和参数使用网格搜索法进行多次调整后，每个模型的参数设置如下：(1)对于 GBRT，最大迭代次数为 3000。学习率为 0.05。一棵树的最大深度为 10。(2)对于 MLP 网络，输入层中的神经元数量等于输入要素的数量。隐藏层由具有 64 个神经元的三个完全连接的层组成，激活函数为 Relu。输出层是完全连接的层，其神

经元数为 1。(3)对于 SVM，目标函数的惩罚系数 C 为 100，核函数的系数 gamma 为 0.1，核系数为 RBF。

仿真主要由三部分组成：(1)分析了复合特征筛选方法对预测精度的影响。(2)测试了所提出的训练方法和传统训练方法的性能。(3)探讨了次级学习器的选择和初级学习器的模型位置对预测精度的影响。

4.2 特征筛选方法对预测准确性的影响

我们通过非线性和线性的特征筛选方法相结合，来挖掘对光伏发电量具有重要影响的辐射变量。首先使用特征贡献度来整体查看辐射变量对于光伏发电量的贡献程度，然后通过 Pearson 相关系数进行线性特征筛选。我们通过图 4 可视化了 5 个变量，发现通过特征贡献度筛选出来的特征，两者之间具有较高的线性关系，虽然三个模型对于输入变量是否具有线性相关性不存在偏好，但是输入变量存在线性相关会导致对于模型的信息提取不充分。通过偏度和峰度来描述图 4 的特征变量的数据分布情况。偏度是用来描述数据分布的对称性，正态分布时偏度为 0，其绝对值越大，出现极端值的可能性较高；峰值是用来描述数据分布陡峭或平滑的情况，正态分布时，峰值为 0 [21]。经过算法计算，偏度的绝对会介于 0.1 和 0.8 之间，THDR 和 TVDR 出现了‘长尾’；峰值在-0.2 到-1.1 之间，数据分布略为陡峭。整体数据分布呈现非统计性正态分布。为了进一步，说明特征筛选方法结合的必要性，我们计算了两者在三个模型中的预测精度。这两种特征筛选方法的结合，相对于单一的特征贡献度而言，其对预测精度的影响如表 3 所示。

通过该方法对数据特征的提取，极大地提高了 GBDT，MLP，SVM 的预测精度，显示了该方法的优越性。就评价指标而言，多方法提取特征对于模型方 MAPE 性能提高最大。此外，就单个模型而言，复合特征筛选方法对于 MLP 的预测精度的提高，最为显著。整体上，无论是复合特征筛选方法还是单一的特征筛选方法，GBDT 的预测精度在 MAE 和 MAPE 两个评价指标中，均高于 MLP 和 SVM。

表 3 单一模型的预测精度

	复合特征筛选方法			特征贡献度		
Model	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
GBDT	0.240018	0.170966	17.10%	0.240070	0.171022	27.73%
MLP	0.235380	0.174064	17.41%	0.247038	0.183786	38.36%
SVM	0.233262	0.181359	18.14%	0.243766	0.187678	42.94%

4.3 次级学习器的选择对于模型精度的影响

为了分析次级学习器的选择和初级学习器中模型的顺序对模型预测准确性的影响，我们计算了三个不同次级学习器和不同初级学习器的顺序对预测精度，统计结果如表 4 所示。

Table 4 Impact of model selection and model order on results

Scheme	**Mete learn machine**	**Primary learning machine**	**RMSE**	**MAE**	**MAPE**
Case 1	GBDT	MLP, SVM	0.074285	0.052217	15.6538%
Case 2	GBDT	SVM, MLP	0.074265	0.052228	15.6546%
Case 3	SVM	GBDT, MLP	0.067265	0.053252	18.7253%
Case 4	SVM	MLP, GBDT	0.066841	0.053009	18.9068%
Case 5	MLP	SVM, GBDT	0.047162	0.035891	11.8034%
Case 6	MLP	GBDT, SVM	0.051974	0.039443	12.7675%

显然，每个模型在集成模型中的位置对预测精度都有很大的影响。表 2 显示，就单个模型而言，经过网格搜索迭代优化参数后，三个模型的精度比较接近。不同的评价指标对于三个单一的模型，带来不同的性能结果。其中就 MAPE 评价指标方面 GBDT 的精度最高，而 SVM 的精度最差。从表 3 中可以看出，案例 3 和案例 4 的预测精度在 MAPE 指标中出现预测精度低于 GBDT 模型，这表明模型融合的预测精度无法保证高于单个模型，因为预测精度受到影响通过整体模型的结构和子模型的位置。另外，从案例 5 和案例 6 可以看出，由于 SVM 和 GBDT 对于离群值具有较好的兼容性，将两者放在初级学习器中，有利于避免使用数据替换等方法消除离群值的干扰，从而减少模型的信息缺失，将 SVM 和 GBDT 模型放在初级学习机有助于提高预测精度。这表明集成模型通过调整结构可以实现比传统单个模型更高的预测精度。

5 结论

准确地预测短期光伏发电量，对于改善电力系统的运行，提高光伏发电的并网的稳定性具有重要的现实意义。在本文中，我们形成了多个模型的集合以预测短期光伏发电量。该模型集成技术可以充分吸收不同算法的优点，大大提高了预测精度。经过仿真实验，可以得出以下结论：

(1) 基于四分位法的箱型图可以发现离群值，有利于选择合适的模型来提高预测精度。

(2) 两种特征筛选方法的结合使用，比原先单一的特征提取方法，可以有效地提高模型的预测精度。

(3) 当我们形成多个模型的集合时，模型的位置对预测精度有很大的影响。具体而言，将具有强大数据兼容性的单个模型放在初级学习器中，可以提高整体集成模型的数据兼容性，从而实现处理原始数据中存在的离群值，减少模型的信息缺失。通过调整集成模型的结构，集成模型可以实现比单个模型更高的预测精度。

参考文献

[1] 仲仕晶, 赵书杰. 多因素耦合对光伏发电性能影响的试验研究 [J]. 科学技术与工程, 2020, 20 (07): 2727-2732.

[2] 李芬, 马年骏, 刘邦银, et al. 上海地区太阳能资源评估与散射辐射推算方法研究 [J]. 水电能源科学, 2015, 33 (05): 207-210.

[3] Ma T, Yang H, Lu L. Solar photovoltaic system modeling and performance prediction [J]. Renewable and Sustainable Energy Reviews, 2014, 36.

[4] Lorenz E, Hurka J, Heinemann D, et al. Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2009, 2 (1): 2-10.

[5] 张静, 褚晓红, 黄学安, et al. 一种基于加权马尔科夫链修正的 SVM 光伏出力预测模型 [J]. 电力系统保护与控制, 2019, 47 (19): 63-68.

[6] Yin P-Y, Tsai C-C, Day R-F. PSO active learning of XGBoost and spatiotemporal data for PM2.5 sensor calibration [J]. IOP Conference Series: Earth and Environmental Science, 2019, 227 (5): 052048.

[7] 曹颖超. 改进的 GDBT 迭代决策树分类算法及其应用 [J]. 科技视界, 2017 (12): 105+149.

[8] 焦李成, 杨淑媛, 刘芳, et al. 神经网络七十年:回顾与展望 [J]. 计算机学报, 2016, 39 (08): 1697-1716.

[9] 王国薇, 黄浩, 周刚, et al. 集成学习在短文本分类中的应用研究 [J]. 现代电子技术, 2019, 42 (24): 140-145.

[10] 赵珂雨, 陈婉莹. 一种基于 stacking 集成学习的 DGA 域名检测方法 [J]. 数据通信, 2020 (06): 19-24.

[11] 唐科, 秦敏, 赵星, et al. 基于 Stacking 集成学习模型的气态亚硝酸预测 [J]. 中国环境科学, 2020, 40 (02): 582-590.

[12] 蒋洪迅, 田嘉, 孙彩虹. 面向 PM_(2.5) 预测的递归随机森林与多层神经网络集成模型[J]. 系统工程, 2020, 38 (05): 14-24.

[13] Xiao J, Li Y, Xie L, et al. A hybrid model based on selective ensemble for energy consumption forecasting in China [J]. Energy, 2018, 159.

[14] 杨荣新, 孙朝云, 徐磊. 基于Stacking模型融合的光伏发电功率预测 [J]. 计算机系统应用, 2020, 29 (05): 36-45.

[15] 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.

[16] 徐耀松, 段彦强, 王雨虹, et al. 基于相似日选择与改进 Stacking 集成学习的短期负荷预测 [J]. 传感技术学报, 2020, 33 (04): 537-545.

[17] Mallor F, León T, Boeck L D, et al. A method for detecting malfunctions in PV solar panels based on electricity production monitoring [J]. Solar Energy, 2017, 153.

[18] 俞娜燕, 李向超, 费科, et al. 基于改进高斯过程回归的光伏电站监测数据修复研究 [J]. 自动化与仪器仪表, 2020 (05): 56-58, 62.

[19] 彭曙蓉, 郑国栋, 黄士峻, et al. 基于 XGBoost 算法融合多特征短期光伏发电量预测 [J]. 电测与仪表, 2020, 57 (24): 76-83.

[20] 邓威, 郭钇秀, 李勇, et al. 基于特征选择和 Stacking 集成学习的配电网网损预测 [J]. 电力系统保护与控制, 2020, 48 (15): 108-115.

[21] 崔书华, 李果, 刘军虎, et al. 基于偏度与峰度的数据质量评估 [J]. 弹箭与制导学报, 2015, 35 (06): 98-100, 105.

作者简介

李志生

1972 年生，博士，副教授，研究生导师. 研究方向为室内环境与建筑节能.

E-mail: Chinalzs@sina.com