·实验技术·

改进特征选择的光伏功率预测融合算法

苏华英[1]，王融融[1*]，张　俨[1]，廖胜利[2]，王国松[3]，代　江[4]

（1. 贵州电网有限责任公司电力调度控制中心 水调与新能源部，贵阳 550002；2. 大连理工大学 水利工程学院，大连 116024；
3. 贵州电网有限责任公司电力调度控制中心 方式部，贵阳 550002；4. 贵州电网有限责任公司电力调度控制中心 发电部，贵阳 550002）

摘要：为提高电站光伏功率预测准确率，该文提出了改进特征选择的融合预测模型。首先耦合包裹式和过滤式方法筛选特征参数；然后根据气象特征分类构建 XGBoost、LightGBM 和 MLP 的单一模型；最后使用双隐藏层多层感知器(MLP)构建融合模型进行预测。实验结果表明，通过改进特征选择以及使用对非线性描述能力更佳的 MLP 融合算法，融合预测模型相比单一模型具有更高的预测准确率以及更强的泛化能力，可较好地满足短期光伏功率预测的需求。

关　键　词：特征选择；多层感知器；融合模型；光伏功率预测

中图分类号：TM615　　**文献标志码：**A　　DOI: 10.12179/1672-4550.20220546

Photovoltaic Power Prediction Fusion Algorithm Based on Improved Feature Selection

SU Huaying[1], WANG Rongrong[1*], ZHANG Yan[1], LIAO Shengli[2], WANG Guosong[3], DAI Jiang[4]

（1. Department of Hydropower Dispatching and New Energy, Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., Guiyang 550002, China; 2. School of Hydraulic Engineering, Dalian University of Technology, Dalian 116024, China; 3. Department of Operation Mode, Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., Guiyang 550002, China; 4. Department of Power Generation, Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., Guiyang 550002, China）

Abstract: To improve the accuracy of photovoltaic power prediction, a fusion prediction model based on improved feature selection was proposed. Firstly, the Pearson correlation coefficient and the information gain method were combined to select characteristic parameters. Then, the dataset was classified to construct the single model of XGBoost, LightGBM and multilayer perceptron (MLP). Finally, a MLP with two hidden layers was used to build a fusion model. The results show that the fusion prediction model has higher prediction accuracy and stronger generalization ability than the single model, and can better meet the needs of short-term photovoltaic power prediction.

Key words: feature selection; multilayer perceptron; fusion model; photovoltaic power prediction

机器学习是涵盖决策树、贝叶斯算法、支持向量机等算法的集合。在机器学习相关课程中，教学更加侧重其原理及理论计算过程，而要更好地掌握并运用机器学习模型，需要在实际工程中加以使用并尝试解决问题。

由于金融、媒体等领域包含大量的用户数据，并且存在对机器学习建模的需求，使得机器学习在这些领域内快速发展。结合当下对能源问题的关注度不断提高，机器学习也逐渐进入光伏、风力发电等能源预测领域。

光伏能源因其分布广泛、清洁环保而被认为是最具竞争力的可再生能源之一[1]。截至 2021 年底，我国光伏并网装机容量累计达 3.06 亿千瓦，同比增长 50%，分布式光伏在电力系统中的占比逐年升高，持续增加的光伏装机给电力电量平衡及电力可靠供应带来极大挑战，也对光伏功率预测提出了更高要求[2]。

传统的光伏功率预测方法大多采用统计方法或者物理建模，如基于斜面辐照度转换进行的物理建模改进，相比水平辐照度数据的相对误差提

收稿日期：2022−09−09；**修回日期：**2022−11−11

基金项目：国家自然科学联合基金(U1765103)。

作者简介：苏华英(1981−)，女，硕士，高级工程师，主要从事新能源功率预测、电网水电调度方面的研究。

*** 通信作者：**王融融(1996−)，女，硕士，工程师，主要从事电网气象方面的研究。E-mail: 845665626@qq.com

高了 5.07%[3]，基于神经网络的预测方法也取得了较大进展[4−8]。文献 [6] 构建了深度信念网络和 T-S 模糊模型，并结合遗传算法进行时变权重加权融合预测，组合预测得到的平均绝对误差相较于 DBN 和 T-S 模糊单模型有明显降低，但并没有对特征数据进行进一步的处理。文献 [7] 利用基于模糊 C 均值样本加权卷积神经网络考虑了样本数据的影响，但对数据的精度要求较高。文献 [8] 利用 K-means 结合 XGBoost 对不同天气类型进行了预测，但在 GBDT 类预测模型中没有考虑收敛速度更快的 LightGBM 模型。文献 [9]利用粒子群算法优化 BP 神经网络模型初始参数进行预测，优化了单一模型的预测效果，但只是优化了较稳定简单的模型。文献 [10] 通过加权分位数组合结合了分位数 k 最近邻、分位数回归森林和分位数回归基本概率模型的结果，提高了概率预测峰值处的精度。

上述研究多基于简单的数据预处理及单模型预测，较少对数据和融合模型进行深入分析。本文基于对机器学习过程中特征选择、模型融合的理论知识的优化改进，提出改进特征选择的融合预测模型。

1　基于特征工程和聚类的融合预测模型框架

1.1　总体计算流程

总体计算流程包括数据预处理、特征工程模型构建、单预测模型和多层感知器（multilayer perceptron，MLP）融合模型 4 个部分，如图 1 所示。

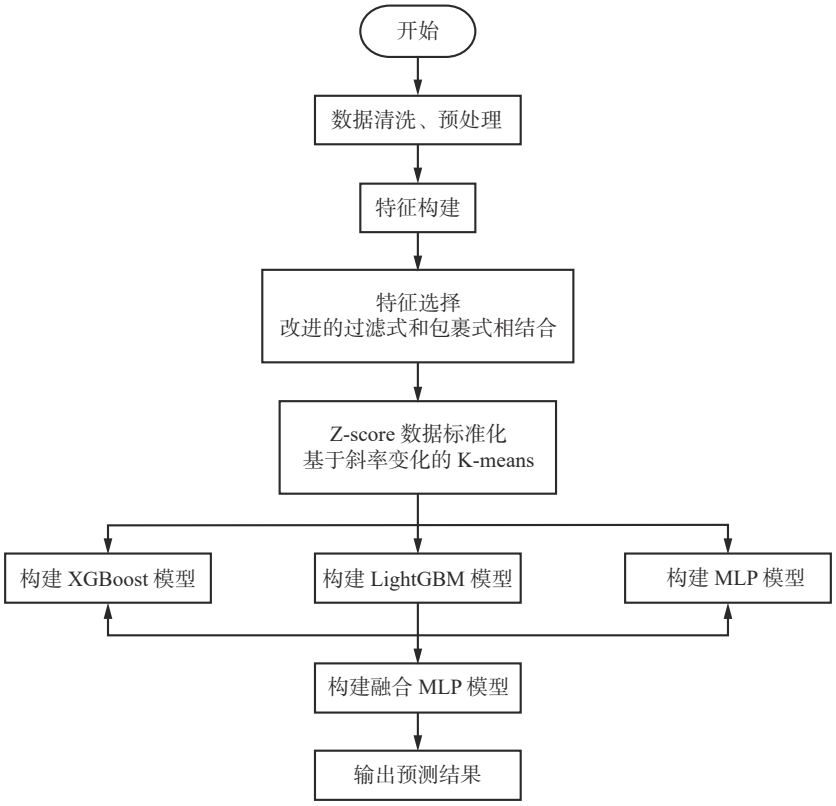

图 1　融合预测模型流程图

在预处理阶段设置阈值将可疑数据移除或修正，并进行标准化处理。进一步结合包裹式和过滤式方法来构建新特征子集。考虑到日功率数据中部分数据变化剧烈的现象，采用基于斜率变化的 K-means 算法聚类得到不同天气特征的子数据集。最后利用机器学习分别训练 XGBoost、LightGBM 以及 MLP 模型。基于双隐藏层 MLP 构建融合预测模型，最终得到一个较高精度的光伏发电预测模型。

1.2　改进特征选择的模型构建

特征工程通过对原始数据进行信息挖掘得到更多信息丰富的特征，有助于训练机器学习模型[11]。特征构建一般采用多项式特征以及数值分析的方法[12]。多项式特征指给定特征被相乘或自乘的次

数，结果会使原来特征间的非线性关系接近线性关系。数值分析是考虑数据集中包含具有强烈物理意义的时间和天气等，能够反映季节性、仪表调节等对光伏电站工作状态的影响。

特征选择方法包括过滤式、包裹式和嵌入式。为避免特征维数过大，现使用过滤式和包裹式相结合的方法。

1.2.1　过滤式

过滤式特征选择通过概率论和数理统计分别评估、选取分数高的特征，并采用信息增益[13]和皮尔逊相关系数[14]进行交替选择。

1）信息增益是一种计算原始数据集的信息熵与某个特征的条件熵之差的方法，即：

$$\mathrm{Gain}(A)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{D}\mathrm{Ent}(D^v) \tag{1}$$

$$\mathrm{Ent}(D)=-\sum_{i=1}^{m}p_i\log_2 p_i \tag{2}$$

式中：D代表原始数据集，D^v代表新条件下自变量取样本对应的值，Ent(D)用于计算数据的信息熵，p_i是标签在给定值中出现的频率。

2）皮尔逊相关系数体现了线性相关性的计算。皮尔逊相关系数在−1 和 1 之间，绝对值越大，相关性越强。当|r|>0.7 时，因变量与自变量强相关。

1.2.2　包裹式

包裹式包括子集选择和子集评价两部分。为避免遍历所有特征子集的子集爆炸问题，子集选择采用了贪心算法的前向选择。子集评估包括信息增益和学习器。算法 SFS 前向选择描述如下。

输入：设定一个特征空子集 Dq 以及初始的特征子集 Dp

1）遍历初始特征子集 Dp；

2）计算每个特征的信息熵 Ent；

3）找到最小的信息熵 Ent(D)；

4）IF：Ent(D)−Ent(Dv) > 3%Ent(D)；

将对应特征从集合 Dp 移动至集合 Dq；

5）直至不再有新的特征加入到集合 Dq。

1.3　数据标准化及 K-means 聚类

由于数据集中数据特征的数值范围差别较大，因此需要对特征和标签数据进行标准化处理。常用的标准化方法有 min-max 和 Z-score 标准化。Min-max 标准化也称为离差标准化，是对原始数据的线性变换，使得结果被映射到 0 和 1 之间，其计算式为：

$$x_i^*=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{3}$$

式中：min 和 max 分别是各个特征样本数据的最小值和最大值。

Z-score 标准化使用样本数据的均值和标准差进行处理，得到的数据符合标准正态分布，其计算式为：

$$x_i^*=\frac{x_i-\mu_i}{\sigma_i} \tag{4}$$

式中：μ_i为原始数据的均值，σ_i为原始数据的方差值，x_i^*为标准化后新的数据集。

K-means 是一种稳定性较好的聚类方法，与数据的输入顺序无关，可以很大程度上避免无序带来的问题，算法 K-means 描述如下[15]。

输入：每日的功率数据作为一条数据样本

1）随机选择数据集中的 k 个样本作为初始的聚类中心；

2）计算剩余所有的样本数据和聚类中心之间的欧氏距离；

3）比较剩余样本到聚类中心的距离，选取最小值对应的聚类中心作为其所属的数据簇；

4）遍历所有数据后，计算分类结果的中心点样本值，作为新的聚类中心；

5）重复上述步骤直至聚类中心收敛。

1.4　融合预测模型构建

1.4.1　决策树

决策树是从机器学习领域发展而来的使用概率分析的直觉算法，通过构建、修剪决策树，学习现有数据在各种情况下发生的概率，逐步构建不同属性的决策分支。使用 Keeny 指数[16]能使决策树的计算更简单、快捷，其计算式为：

$$\mathrm{Gini}(t)=1-\sum_{k}p(a_i|t)^2 \tag{5}$$

式中：Gini (t)表示在样本集合中一个随机选中的样本被分错的概率，a_i表示类别，k为决策树的个数，$p(a_i|t)$表示t条件下样本属于第i个类别的概率。

决策树构建算法易分裂过度拟合的节点，形成过拟合，降低决策树的泛化能力[17]。通常使用最小化决策树的全局损失函数的方法评估决策树的泛化能力。损失函数的计算式为：

$$C_a(T)=\sum_{t=1}^{|T|}N_tH_t(T)+\alpha|T| \tag{6}$$

式中：T表示决策树中的节点总数，H_t是T分支下样本的信息熵，N_t是该分支的训练样本数，α是惩罚系数。该损失函数中引入了基于全熵的惩罚项。

1.4.2 梯度提升决策树(GBDT)

GBDT 是一种基于集成思想的决策树算法模型[18]。GBDT 最终的预测结果由训练的众多决策树决定。假设有 k 棵树，则最终预测表示为：

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i) \tag{7}$$

式中：$\hat{y}_i$为最终的预测结果，K表示预测模型中包含树的数量，$f_k(x_i)$为各个决策树得到的预测结果。它需要按顺序生成决策树并遍历所有数据，导致训练时间延长。目前，GBDT 有 XGBoost 和 LightGBM [19] 两个衍生版本。

1) XGBoost

XGBoost 在代价函数中加入正则项以简化模型，可提高模型的泛化能力。XGBoost 增加了并行计算量，对数据进行预排序并将结果保存为一个块。后续迭代可以通过重用块结构大大减少训练时间[20]。

2) LightGBM

LightGBM 与传统 GBDT 算法相比，具有训练速度快、内存占用少、预测准确率高、支持并行学习等优势。与 XGBoost 相比，采用基于直方图的算法代替 XGBoost 中的预排序数据结构，从而提高训练速度。

1.4.3 人工神经元模型

人工神经网络[21]通过建立一个简单的神经元模型，使用权重组成多个神经元的网络结构而形成。神经元模型规定神经元接收来自其他 n 个神经元的刺激，根据相应的权重将它们相乘，相加给到下一层神经元。神经元将其与阈值进行比较，如果超过阈值，则通过激活函数对输入进行处理，并将结果输出到下一个神经元。激活函数是为了打破两层神经元之间的线性独立，加入神经元之间的非线性变换，使神经网络在理论上逼近任何非线性函数。常见的激活函数是 Sigmoid 和 ReLu，计算式分别为：

$$f(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{8}$$

$$f(x)=\max(0,x) \tag{9}$$

1.4.4 多层感知器(MLP)

MLP 是具有一个或多个隐藏层的人工神经网络[22]。最简单的 MLP 仅包含一个隐藏层。如图 2 所示，每个神经元输出权重和一个偏差，将所有权值和偏差初始化为随机值，将数据输入神经网络，根据预测结果计算梯度，然后更新神经网络的参数，直到满足特定条件。

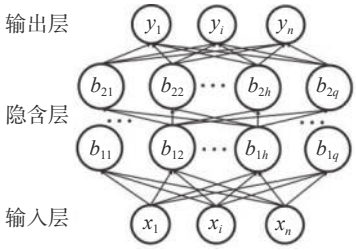

图 2 简单双隐含层 MLP 模型结构

1.5 评价指标

预测结果的指标评估包括均方误差(MSE)、均方根误差(RMSE)、平均绝对百分比误差(MAPE)、正规化方均根误差(NRMSE)、方差占(VAF)和相对方差系数(RCoV)，计算式分别为：

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i-y_i)^2 \tag{10}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{k=1}^{N}(y_k-\hat{y}_k)^2}{N}} \tag{11}$$

$$\mathrm{NRMSE}=\frac{\mathrm{RMSE}}{y_{k_{\max}}-y_{k_{\min}}}\times 100\% \tag{12}$$

$$\mathrm{MAPE}=\frac{1}{N}\sum_{k=1}^{N}\left|\frac{y_k-\hat{y}_k}{y_k}\right|\times 100\% \tag{13}$$

$$\mathrm{VAF}=1-\frac{\mathrm{var}(\hat{y}_k-y_k)}{\mathrm{var}(y_k)}\times 100\% \tag{14}$$

$$\mathrm{RCoV}=\frac{\mathrm{median}\left|\hat{y}_k-\hat{y}_{k\_\mathrm{median}}\right|}{\hat{y}_{k\_\mathrm{median}}} \tag{15}$$

式中：N表式数据量，$\hat{y}_k$表示数据预测值，y_k表示数据真实值，$y_{k_{\max}}$、$y_{k_{\min}}$分别表示真实值的最大、最小值，$\mathrm{var}(x)$表示计算x数据项的方差值，$\mathrm{median}(y)$表示计算 y 的中值，$\hat{y}_{k_\mathrm{median}}$表示测试集中实际值中值对应的预测值。

由于计算 MAPE 在实际值约为 0 时值趋于无穷大，本文引入其他指标进一步评估，其中 RMSE、NRMSE 分别反映预测值与真实值之间的误差、预

测与真实值间的总体偏差。对应的 VAF 和 RCoV 分别代表预测模型的稳定程度和预测目标的稳定程度，注重模型与预测结果的关联程度和不同气候因素对于太阳能资源预测的稳定性。

2　算例分析

2.1　数据来源及数据预处理

选用我国南方某省区 2017-01—2018-12 和 2021-05—2022-03 的实际光伏数据进行建模分析，数据时段均为 15 min，原始数据量共包含 65 760 条数据，单位数据特征包含：时间、风速、风向、温度、湿度、压力、辐照度和功率[23]。辐照度分布和功率分布分别如图 3 和图 4 所示。将数据为负的噪声数据进行修正，辐照度和功率的负样本数分别为 94 和 29 406，将负功率修改为 0。

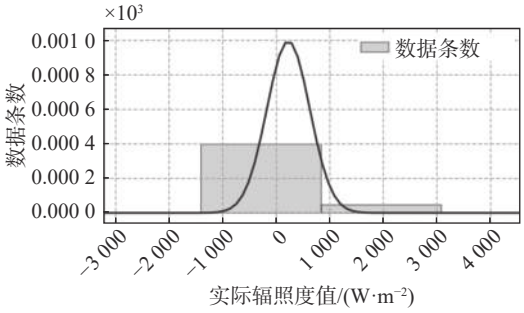

图 3　原始数据的辐照度分布

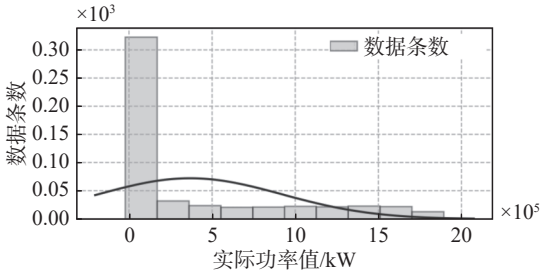

图 4　原始数据的光伏功率分布

辐照度在当月每天各时间对应的功率值应符合正态分布，并在一个受实际功率波动、仪器误差和太阳能电池板状态影响的值附近波动。将数据按月份和时间分组，对超出界限的数据进行筛选，得到正态分布 95% 百分位数的上下界：

$$\begin{cases} \sigma = \sqrt{\dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}} \\ \dfrac{|b-\mu|}{\sigma} = 1.645 \end{cases} \tag{16}$$

经过上述 95% 的辐照度和功率百分位筛选，剔除数据 5 591 条，占样本的 6.98%。当辐照度小于 6.5 单位时，大部分数据的功率小于其平均值的 0.5%。将辐照度阈值设置为 6.5 单位，并将辐照度低于 6.5 单位的数据删除。辐照度达到 1.5 倍阈值但功率小于平均功率 0.5% 的数据项被删除。作为与辐照度的正相关变量，对功率与辐照度的比值进行同样筛选，得到 14 796 个具有原始特征的数据。

2.2　特征工程参数选择

对数据集中温度、湿度、压力、风速、风向和辐照度进行二三次多项式处理，同时添加日平均值、日最大值、日最小值和数据范围，每条数据的月份、天数、日出时长与最大辐照时间和一天中的日出时间之间的差异作为新特征，得到数据集共有 268 个特征。

结合过滤式和包裹式方法对已构建的特征集合进行筛选。首先，为避免优秀的特征被筛选掉，将皮尔逊相关系数与信息增益相结合。设置皮尔逊相关系数的阈值为 0.6，得到的特征集合如图 5 所示。为计算信息增益，需将数据离散化。将连续值分成 10 个相等的区间，每个区间从小到大分配 0~9 的值。分别计算每个特征和标签的信息增益，提取信息增益的前 50 个特征，经过两种过滤式特征筛选得到由 75 个特征组成的特征子集 Dp。

然后，利用前向选择算法，计算新增某特征的信息增益，接受比当前信息熵低 3% 以上的特征为更优的特征，遍历完待选特征且特征子集相比上一次循环没有变化，最终得到 24 个特征的数据集，主要特征包含：时间、辐照度、峰谷间距等。

2.3　标准化及 K-means 聚类

Z-score 标准化适用于原始数据最大值和最小值未知，或存在数据超出取值范围的情况，便于新增数据。本文采用 Z-score 标准化并将数据集分为晴天以及阴雨天两类[24]。时间范围设置为 08：00−19：00，在此范围外的数据忽略不计，区间内部的空缺数据采用线性插值填补，最终得到 44×351 的数据，经过以相邻数据点的斜率变化率为样本的 K-means 聚类得到如图 6 和图 7 所示的两类结果。

其中，图 6 的归一化功率最大值在中午，大部分在 [1, 1.5] 的范围内，分布偏向于正态分布，有晴天的天气类型特征；图 7 显示的功率值全天呈锯齿状，上下波动情况较明显，有阴雨天的天气类型特征。

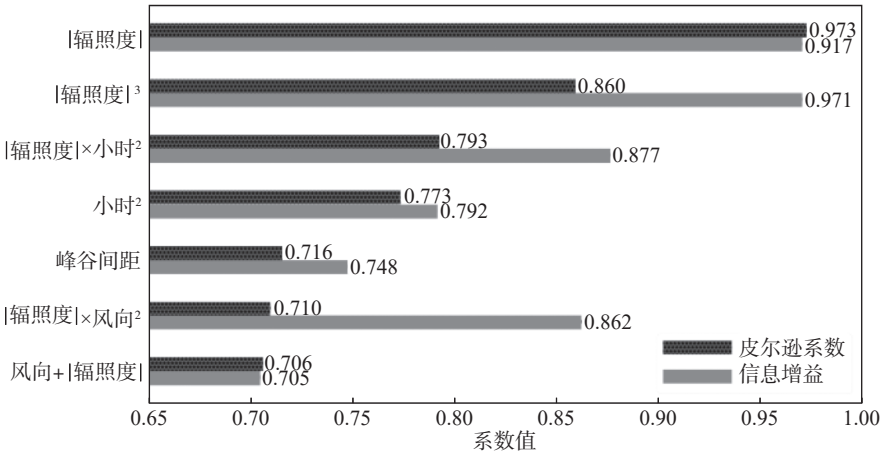

图 5　皮尔逊系数及信息增益 G>0.6 的特征

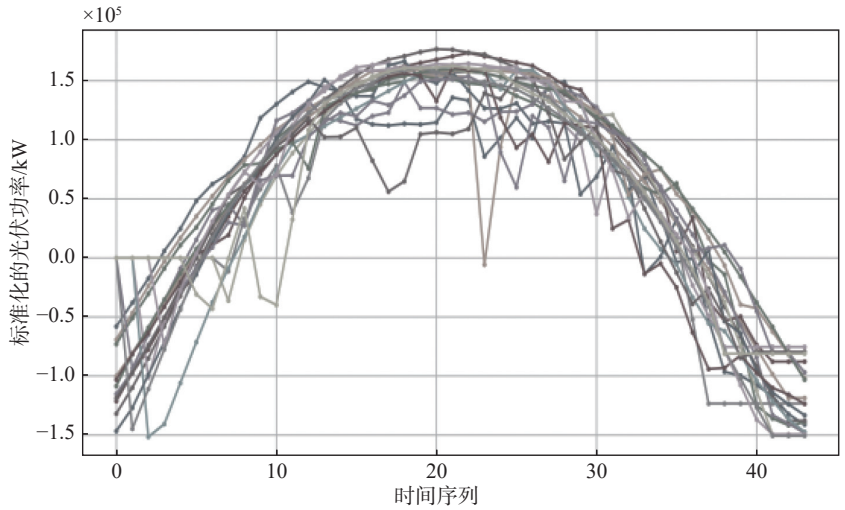

图 6　具有晴天天气特征的数据簇

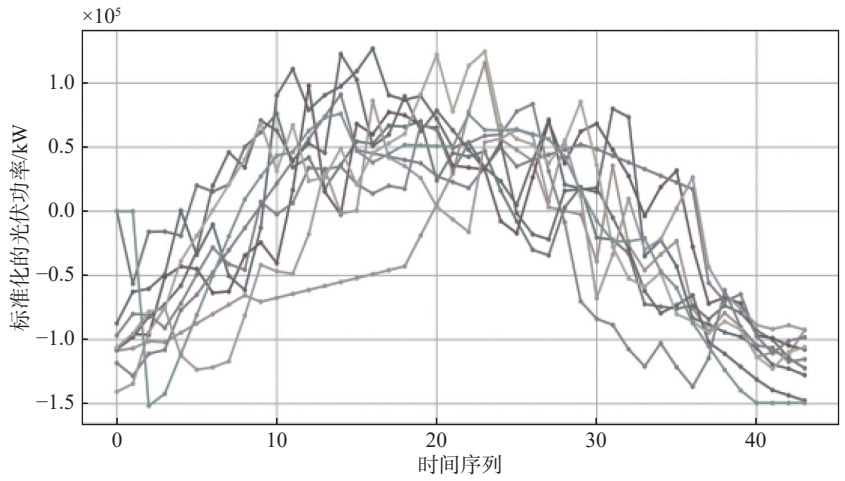

图 7　具有阴雨天天气特征的数据簇

2.4　机器学习模型的构建与训练

为了更好地拟合数据，引入了一种多模型融合机器学习模型。使用的预融合模型包括 XGBoost、LightGBM 和 MLP。XGBoost 模型的超参数设置如表 1 所示，其他模型的参数设置方法类似。

晴天和阴雨天数据集分为训练集、验证集和测试集，数据集大小比例为 6.4:1.6:2。晴天数据集的部分预测结果如图 8 和图 9 所示。

表 1　XGBoost 模型的超参数设置

参数名称	参数值
Nthread	−1
Learning rate	0.05
Objective	‘regsquarederror’
Booster	‘gbtree’
Gamma	0.3
Max depth	15
Subsample	0.75

图 8 显示了分类后晴天特征数据集预测的部分功率数据，其中 LightGBM 单模型在最大值的预测上表现不佳。多层感知器模型预测的部分功率数据效果较差，LightGBM 模型的性能介于上述两种模型之间。XGBoost 的预测效果在晴天的表现较好，多层感知器在一个简单的模型下获得了更低的 MSE。在阴雨天气特征集下 LightGBM 模型的 MAPE 最小为 15.5%，但 MSE 为 1.497、NRMSE 为 8.02%，甚至高于简单的 MLP 模型。因此，为提高阴雨天气特征下的预测效果，考虑将多个模型进行融合。

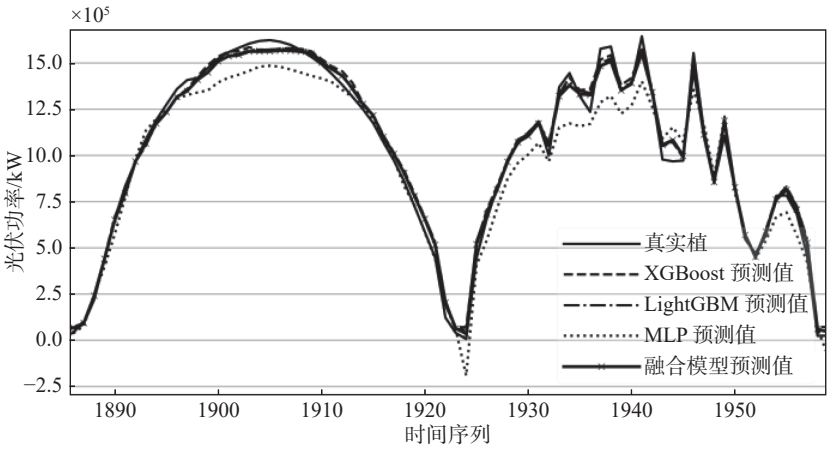

图 8　晴天特征数据集的部分预测结果

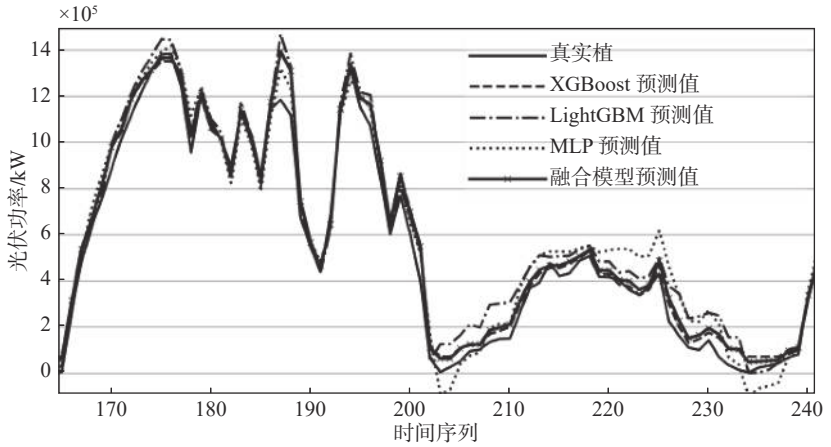

图 9　阴雨天特征数据集的部分预测结果

2.5　预报结果对比分析

现引入设计神经网络再学习以及遍历权值的模型融合方法将上述 3 个独立模型进行融合[25]。设计神经网络再学习方法以双隐含层 MLP 模型为融合基础，得到预测结果如下：晴天数据集的预测 MSE 为 3.615，MAPE 为 12.76%；阴雨天数据的预测 MSE 为 0.873，MAPE 为 15.85%。遍历权值的融合方法以预测的 MSE 为目标函数，得到最优权重。晴天数据集的预测 MSE 为 1.069，MAPE 为 7.51%；阴雨天气数据集的 MSE 为 1.054，MAPE 为 17.08%。预测结果如图 8、图 9 所示。

表 2 和表 3 包含了 5 个模型的结果，晴天特征的预测下，权值遍历的融合方式在各个指标下均得到最优的预测情况。阴雨天气特征的预测下

经 MLP 模型的融合方式在大部分指标下均得到了最优的预测情况；总体上各个模型在晴天数据集的预测效果均比阴雨天气的数据集要好。

利用融合模型在不同的数据集中表现优异，预测误差均较低，所提出的 MLP 融合模型一定程度上改进了系统的非线性，提高了阴雨天气特征下的预测精确度，针对不同的气候类型有更好的容错率。在数据稳定性较优的晴天特征集中通过 MLP 融合方式的预测模型出现过拟合的现象使得误差偏大。

表 2 晴天天气特征下各模型的预测情况

模型	MSE	MAPE/%	RMSE	NRMSE/%	VAF/%	RCoV
权值遍历融合	1.069	7.51	1.034	6.11	96.49	0.221
MLP融合模型	3.615	12.76	1.901	3.28	88.04	0.283
XGBoost模型	1.113	7.75	1.055	6.53	96.34	0.226
LightGBM模型	1.078	7.53	1.038	6.24	96.46	0.221
MLP模型	1.227	9.61	1.107	4.23	95.99	0.226

表 3 阴雨天气特征下各模型的预测情况

模型	MSE	MAPE/%	RMSE	NRMSE/%	VAF/%	RCoV
权值遍历融合	1.054	17.08	1.026	7.17	94.23	0.567
MLP融合模型	0.873	15.85	0.934	4.32	96.20	0.682
XGBoost模型	1.103	18.69	1.050	7.91	91.90	0.581
LightGBM模型	1.497	15.50	1.223	8.02	91.90	0.601
MLP模型	1.951	23.93	1.397	6.02	89.44	0.553

对于不同天气特征的两组数据集，阴雨天气数据集下的 VAF 值和 RCoV 值普遍高于晴天，说明数据分类一定程度上提升了预测效果，其中更优异的模型对应的 VAF 指标值更大，说明预测需要依据一个较为稳定适合的模型，阴雨天气下的 RCoV 指标反映出阴雨特征数据更大的波动特征和不确定性。

进一步利用南方电网的数据进行泛化能力检验，将预处理后的数据集通过相同的特征工程和模型构建及模型融合预测，得到结果如图 10 所示。

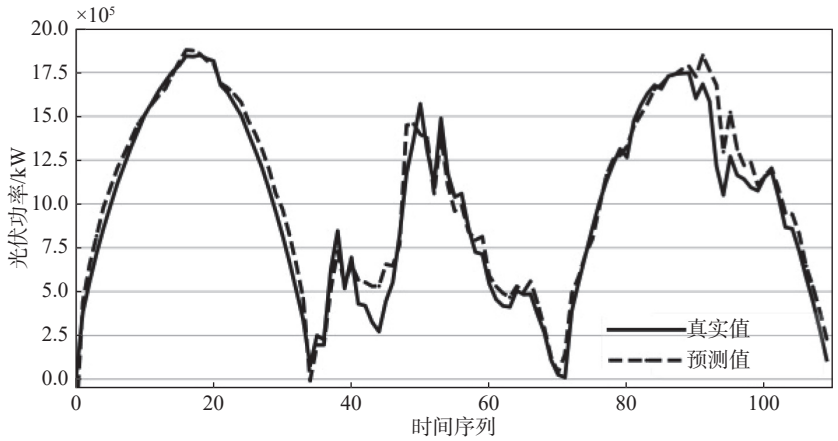

图 10 应用南方电网新数据的 MLP 融合测模型预测的部分结果

新数据集的指标计算如下：MAPE 为 13.33%，RMSE 为 1.70, NRMSE 为 8.12%, VAF 为 90.81%, RCoV 为 0.544。该电站在多数情况下数据波动剧烈，功率值分布较为分散。在阴雨天较多的情况下，融合模型得到的相对误差结果在可接受的范围内略有波动。图中时间序列 90~100 处的预测误差有所增大，造成该情况的原因是数据分类并没有考虑当日中出现短期阴雨天的情况，故对于其他的一些特殊天气状况的预测效果较差。新数据验证表明，构建的应用神经网络融合的方式在阴

雨天特征下具有较强的泛化能力，能够保证在不同数据下的预测效果。

3　结束语

本文提出了一个改进特征选择的光伏功率预测机器学习融合方法，通过离线训练的方式构造了可在线预测的光伏发电功率预测系统。方法基于正态分布消除部分可疑数据，通过交叉项和数值分析添加大量的新特征项，采用包裹式和过滤式相结合的特征筛选方式过滤相关特征，构造最优特征子集。基于聚类分析将数据分为晴天和阴雨天两类天气特征，分别构造 XGBoost、LightGBM 和 MLP 单一模型，最后通过双隐藏层 MLP 进行最优权重的计算以及权值遍历得到融合模型。

经计算，融合模型预测在晴天和阴雨天情况下的 MAPE 分别为 7.51% 和 15.85%，MSE 为 1.069 kW 和 0.873kW，与其他 3 个单一模型相比表现更好。

通过具体实践，融合模型得到较为准确的光伏功率单点预测，为电力调度决策提供相关参考并促进对机器学习算法的理解与掌握。

参 考 文 献

[1] 刘伟, 彭冬, 卜广全, 等. 光伏发电接入智能配电网后的系统问题综述[J]. 电网技术, 2009, 33(19): 1−6.

[2] 于群, 朴在林, 胡博. 基于EEMD和BP神经网络的短期光伏功率预测模型[J]. 电网与清洁能源, 2016, 32(7): 133−137.

[3] 姜文玲, 赵艳青, 王勃, 等. 基于 NWP 辐照度斜面转换的光伏功率预测方法[J]. 山东大学学报(工学版), 2021, 51(5): 114−121.

[4] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38−39.

[5] 范高锋, 王伟胜, 刘纯, 等. 基于人工神经网络的风电功率预测[J]. 中国电机工程学报, 2008, 28(34): 118−123.

[6] 谭小钰, 刘芳, 马俊杰, 等. 基于DBN与TS时变权重组合的光伏功率超短期预测模型[J]. 太阳能学报, 2021, 42(10): 42−48.

[7] 吕伟杰, 方一帆, 程泽. 基于模糊C均值聚类和样本加权卷积神经网络的日前光伏出力预测研究[J]. 电网技术, 2022, 46(1): 231−238.

[8] 常俊晓, 金之榆, 卢姬, 等. 基于集成聚类和XGBoost的短期光伏发电功率预测[J]. 浙江电力, 2021, 40(10): 102−107.

[9] 靳瑞强, 马广昭, 耿立卓. 基于PSO-BP神经网络的光伏发电功率预测方法[J]. 信息技术, 2021, 45(12): 147−152.

[10] BRACALE A, CARPINELLI G, FALCO P D. Developing and comparing different strategies for combining probabilistic photovoltaic power forecasts in an ensemble method[J]. Energies, 2019, 12(6): 1−16.

[11] FAN C, SUN Y J, ZHAO Y, et al. Deep learning-based feature engineering methods for improved building energy prediction[J]. Applied Energy, 2019, 240: 35−45.

[12] 王娟, 慈林林, 姚康泽. 特征选择方法综述[J]. 计算机工程与科学, 2005, 27(12): 68−71.

[13] 刘庆和, 梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用, 2011, 47(12): 131−134.

[14] LIU Y Q, MU Y, CHEN K Y, et al. Daily activity feature selection in smart homes based on Pearson correlation coefficient[J]. Neural Processing Letters, 2020, 51(2): 1771−1787.

[15] YANG S L, LI Y S, HU X X, et al. Optimization study on k value of kmeans algorithm[J]. Systems Engineering-Theory & Practice, 2006, 26(2): 97−101.

[16] PAXTON R J, ZHANG L F, WEI C S, et al. An exploratory decision tree analysis to predict physical activity compliance rates in breast cancer survivors[J]. Ethnicity & Health, 2019, 24(7): 754−766.

[17] FAYYAD U M, IRANI K B. On the handling of continuous-valued attributes in decision tree generation[J]. Machine Learning, 1992, 8(1): 87−102.

[18] LIU S, LYU Q, LIU X J, et al. A prediction system of burn through point based on gradient boosting decision tree and decision rules[J]. ISIJ International, 2019, 59(12): 2156−2164.

[19] LIANG W Z, LUO S Z, ZHAO G Y, et al. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms[J]. Mathematics, 2020, 8(5): 45−62.

[20] PARK N, AHN H K. Multi-layer RNN-based short-term photovoltaic power forecasting using IoT dataset[C]// 2019 AEIT international annual conference (AEIT). [S. l.]: IEEE, 2019.

[21] MUNANDAR D. Multilayer perceptron (MLP) and autoregressive integrated moving average (ARIMA) models in multivariate input time series data: Solar irradiance forecasting[J]. International Journal on Advanced Science, Engineering and Information Technology, 2019, 9(1): 220−228.

[22] 韩玲. 基于人工神经网络——多层感知器(MLP)的遥感影像分类模型[J]. 测绘通报, 2004(9): 29−30.

[23] DONAHUE E A, QUACH T T, POTTER K, et al. Deep learning for automated defect detection in high-reliability electronic parts[C]//Applications of Machine Learning. [S. l.]: SPIE, 2019.

[24] CHO Y J, LEE H C, LIM B, et al. Classification of weather patterns in the east asia region using the K-means clustering analysis[J]. Atmosphere, 2019, 29(4): 451−461.

[25] 房海腾. 多源信息融合中连续变量离散化及权重分配算法的研究[D]. 济南: 山东大学, 2017.

编辑　张俊