DOI:10.12158/j.2096-3203.2021.06.012

基于偏最大信息系数与组合 XGBoost 的短期风功率预测

李科，黄东晨，陶子彬，熊欢，李浩文，杜业冬
（南瑞集团（国网电力科学研究院）有限公司，江苏 南京 211106）

摘　要：作为新能源领域的课题热点之一，短期风功率预测的研究在提高预测精度的同时也应重视模型的工程化应用。据此，提出一种基于偏最大信息系数的组合 XGBoost 预测模型。首先，设计一种基于偏最大信息系数的特征选择算法，通过引入偏互信息，在挖掘出对风功率影响较大的气象特征的同时，也能消除耦合信息带来的不利影响。在此基础上，为兼顾模型的精度和计算效率，降低单个模型的预测风险，构建以 XGBoost 为底层算法的组合预测模型，进一步实现风功率预测。采用 2 个具有较大差异的风电场作为算例进行验证分析，结果表明，基于偏最大信息系数特征选择算法的组合 XGBoost 预测模型不但能提升短期风功率的预测精度，与相近的组合预测模型相比，也具备更高的计算效率，有利于工程化应用。

关键词：特征选择；组合 XGBoost；偏最大信息系数；短期风功率预测；计算效率；工程化应用
中图分类号：TM614　　**文献标志码**：A　　**文章编号**：2096-3203（2021）06-0095-08

0　引言

风能是一种可大规模商用的绿色可再生能源。由于风力发电具有强烈的间歇性和随机性，且周期规律不明显，当风电大规模并入电网时，可能对电网的安全及稳定运行产生不良影响[1]。因此，亟需发展精准的风功率预测技术。

围绕着短期风功率预测，国内外学者进行了诸多探索。自回归滑动平均模型（autoregressive moving average model，ARMA）假设当前时刻的值与前一时刻的值及随机干扰量均有关，具有一定捕捉时序信息的能力，是一种常用于风功率预测的时间序列分析模型[2]。但对于功率波动较大且无明显周期规律的风电场而言，使用 ARMA 模型会导致预测结果误差较大。人工神经网络（artificial neural network，ANN）[3—4] 和支持向量机（support vector machine，SVM）[5] 是另外 2 种经典的风功率预测模型。ANN 能够自适应、自学习，以任意精度逼近任何非线性映射，非常适合描述风功率预测模型复杂与非线性的特点。但 ANN 训练时间长，调参过程较为繁琐，容易出现过拟合的情况[6]。相较而言，SVM 有着较强的泛化能力，不易过拟合。然而，当训练样本大幅增加时，SVM 的性能提升不明显[7—8]。

近年来，极限梯度提升（extreme gradient boosting，XGBoost）[9] 算法由于在 Kaggle、KDD 等一系列大数据算法竞赛中表现优异，引发了大量的关注。XGBoost 不仅在算法精度上较传统算法表现出色，同时也支持并行化运行，减少了模型的训练时间。此外，XGBoost 还具有可移植性强、支持多数主流编程语言、集成了 Spark 等各类主流大数据平台等特性[10]，这些特性增强了 XGBoost 的普适性，使得 XGBoost 在工程化应用方面具有更大的优势。目前 XGBoost 已经在光伏发电量预测[11—14] 等多个领域有所应用。然而，仅使用单一预测模型存在一定的泛化问题，需要结合组合预测策略[15]。文献[16] 以反向传播（back propagation，BP）神经网络、线性外推和 SVM 为底层算法，构建一种动态调整权重分配的风电预测集成学习模型，获得了提高模型整体泛化能力的效果。然而，文中的方法缺少效率层面上的考虑。另外，在工程应用中，不同风电场适用的模型气象特征输入不同[17]，需要有相应的算法选择合适的气象特征作为模型的输入变量。

综上所述，文中提出一种结合偏最大信息系数（partial maximal information coefficient，PMIC）特征选择算法的组合 XGBoost 短期风功率预测模型。首先，设计基于 PMIC 的特征选择算法，对风速、风向等常用气象特征进行优选；其次，以 XGBoost 为底层算法构建组合预测模型，实现对短期风功率的预测。算例结果表明，文中方法能有效提高短期风功率预测精度及计算效率，有助于工程化应用。

1　特征选择

1.1　风功率预测问题中常用的气象特征

据已有文献及现场情况，目前常用气象特征有：

收稿日期：2021-06-17；修回日期：2021-08-21
基金项目：国家电网有限公司总部科技项目“基于大数据的电网趋势预测及操作智能预演技术研究”（5108-20214003-6A-0-0-00）

（1）风速 V_h，即风电场与地面相对高度为 h 米时的平均风速，在文中后续算例中，h 的取值范围为 {30,50,70,90,110} 。

（2）风向 D_{dir}，即风电场当地的平均风向。

（3）温度 T_{temp}，即风电场当地的平均温度。

（4）湿度 H_{hum}，即风电场当地的平均湿度。

（5）气压 P_{pres}，即风电场当地的平均气压。

在短期风功率预测问题中，风速是决定风电场输出功率的主导因素。由于地表粗糙度和大气热分层的影响，风速的分布并不完全遵循对数风廓线或指数风廓线，有时还会出现低海拔风速高于高海拔风速的情况。因此，在选择预测模型的特征输入时，可以考虑不同高度的风速，这样能够更好地表征风电场周围的大气特征[18]。

另外，根据文献[19]，风速之外的因素也可能对风电场的功率出力情况造成影响。对于不同风电场，气象特征对输出功率的影响程度也不相同。例如，当风电场内风机空间布局密集程度较大时，尾流效应对风电场出力的影响尤其突出[20]，此时风向对风功率有较大的影响。因此，需在建立模型前对特征进行选择。

1.2 基于 PMIC 的特征选择

最大信息系数（maximal information coefficient，MIC）是一种衡量变量间相关性程度的统计量[21]，不仅能够刻画变量间的线性与非线性关系，还能够捕获变量间潜在的非函数关系。其主要思想为：如果 2 个变量之间具有一定的相关关系，对相应变量的散点图进行不同方案的网格划分，计算对应的互信息（mutual information，MI）值并且进行正则化，取这些值中的最大值，则该值为这 2 个变量的 MIC。其中，MI 值是衡量变量之间相关性程度的指标。给定变量 $X=\{x_i\}$，$Y=\{y_i\}$，$i=1,2,\cdots,n$，n 为样本数目，其 MI 值定义为：

$$I_{\mathrm{mi}}(X,Y)=\sum_{x\in X}\sum_{y\in Y}\left(f(x,y)\log_2\frac{f(x,y)}{f(x)f(y)}\right) \tag{1}$$

式中：$f(x,y)$ 为 X 和 Y 的联合概率密度；$f(x)$，$f(y)$ 分别为 X 和 Y 的边缘概率密度。采用高斯函数对上述概率密度进行估计，至此，可进一步求得 MI 值。

给定一个有限二元数据集合 $D=\{(x_i,y_i)\}$，$i=1,2,\cdots,n$，将变量 X 划分为 x 个区间，Y 划分为 y 个区间，则能够得到一个 $x\times y$ 的网格划分 G。同样的 $x\times y$ 规格的网格划分方案有多种，对每一种方案，计算其 MI 值，取不同划分方案下 $I_{\mathrm{mi}}(X,Y)$ 的最大值作为划分 G 的 MI 值。至此，定义集合 D 在划分 G 下的最大 MI 值为：

$$\boldsymbol{I}^*(D,x,y)=\max_G \boldsymbol{I}(D\mid_G) \tag{2}$$

式中：$D\mid_G$ 为集合 D 在 G 上的概率分布；$\boldsymbol{I}(D\mid_G)$ 为在该概率分布下的 MI 值；$\max_G$ 为遍历所有可能的 $x\times y$ 网格 G。

将所有划分方案下的最大 MI 值进行正则化，并组成特征矩阵 $\boldsymbol{M}(D)_{x,y}$，定义为：

$$\boldsymbol{M}(D)_{x,y}=\frac{\boldsymbol{I}^*(D,x,y)}{\log_2\min(x,y)} \tag{3}$$

最大信息系数 I_{mic} 定义为：

$$I_{\mathrm{mic}}(D)=\max_{xy<B(n)}\{\boldsymbol{M}(D)_{x,y}\} \tag{4}$$

式中：$B(n)$ 为网格划分 $x\times y$ 的上限值。在文献[22]中，建议将 ε 设为 0.6，即 $B(n)=n^{0.6}$。

然而，风电场相关气象特征之间普遍存在一定程度的耦合关系。为此，文中在 MIC 的基础上进一步引入偏互信息（partial mutual information，PMI），将 MIC 改造为 PMIC，以消除耦合给特征选择带来的不利影响。

设 X 和 Z 为多输入系统中的输入变量，Y 为输出变量。若 X 和 Z 之间具有耦合关系，将导致 X 和 Y 之间最大信息系数 $I_{\mathrm{mic}}(X,Y)$ 的计算出现偏差。因此，文中应用条件期望 $m_X(z)$ 和 $m_Y(z)$ 分别对 X 和 Y 中包含 Z 的信息剔除，分别记为 U，V：

$$m_X(z)=E(x\mid Z=z)=\frac{\sum_{i=1}^{n}(x_i f(z))}{\sum_{i=1}^{n}f(z)} \tag{5}$$

$$U=X-m_X(Z) \tag{6}$$

$$V=Y-m_Y(Z) \tag{7}$$

式中：$f(z)$ 为 Z 的边缘概率密度函数。X 和 Y 的 PMIC 记为：

$$I_{\mathrm{pmic}}(X,Y)=I_{\mathrm{mic}}(U,V) \tag{8}$$

文中采用赤池信息量准则（akaike information criterion，AIC）[22]作为变量筛选的结束条件，即：

$$T_{\mathrm{AIC}}=n\log_2\left(\frac{1}{n}\sum_{i=1}^{n}r_i^2\right)+2(p+1)^2 \tag{9}$$

式中：r_i 为根据已选变量计算的 Y 回归残差；p 为已选变量个数。随着变量的筛选，T_{AIC} 不断减小，当 T_{AIC} 达到最小值时，最优自变量集合筛选完毕。

设输入变量集为 F，输入变量为 Y，最优输入变量集为 S，F_S 为最大的 PMIC 值对应的候选变量。PMIC 变量选择算法流程如下：

（1）将 S 初始化为空集。

（2）计算 F 中各变量与 Y 的最大信息系数 $I_{\mathrm{mic}}(F_i,Y)$。

（3）选择使 $I_{\mathrm{MIC}}(F_i,Y)$ 值最大的 F_S。

(4) 计算 T_{AIC} 值,并将 F_S 移入 S 。

(5) 若 $F \neq \varphi$,计算 $V = Y - m_Y(S)$;对于每一个 $F_j \in F$,计算 $U = F_j - m_{F_j}(S)$ 。

(6) 选择使 $I_{\mathrm{mic}}(U,V)$ 值最大的 F_S 。

(7) 更新 T_{AIC} 值,若 T_{AIC} 减小,则将 F_S 移入 S ,返回步骤(5),否则终止筛选。

2 短期风功率组合预测模型

2.1 XGBoost 算法原理

XGBoost 是一种以决策树为基础的梯度提升算法,计算速度快,模型表现好。给定含有 N 个样本和 M 个气象特征的训练样本集 $D = \{(x_i, y_i)\}$ ($i = 1,2,\cdots,N, x_i \in \mathbf{R}^M, y_i \in \mathbf{R}$) ,$\mathbf{R}^M$ 为具有 M 个维度的实数集。XGBoost 算法使用由 K 个回归决策树函数相加构成的集成模型对功率进行回归预测:

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in F \tag{10}$$

式中:$\hat{y}_i$ 为第 i 个样本对应的预测值;f_k 为集合 F 中第 k 个决策树函数。为了学习式(10)中的决策树函数,构造一个目标函数,如式(11)所示。

$$f_{\mathrm{obj}}(\varphi) = \sum_{i=1}^{N} l(\hat{y}_i, y_i) + \Omega(f_k) \tag{11}$$

式中:$l(\hat{y}_i, y_i)$ 用于衡量模型的预测值 $\hat{y}_i$ 与实测值 y_i 之间的误差;$\Omega(f_k)$ 为正则项,用于控制模型的复杂度,防止出现过拟合。正则项的定义见式(12)。

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{12}$$

式中:T 为树中叶子结点数量;ω_j 为第 j 个叶子的权重;γ 为 T 的系数,一般取为 1;λ 为 $\sum_{j=1}^{T} \omega_j^2$ 的系数。

传统的随机梯度下降等优化算法难以对 XGBoost 模型进行直接优化,需要在训练过程中引入加法策略,即采用增量的训练方式:每一步都在前一步的基础上增加一棵树。设 $\hat{y}_i^{(t_{\mathrm{rou}})}$ 为第 i 个样本在第 t_{rou} 轮迭代时的预测值,则第 t_{rou} 轮的目标函数可写为:

$$f_{\mathrm{obj}}^{(t_{\mathrm{rou}})}(\varphi) = \sum_{i=1}^{n} l(\hat{y}_i^{(t_{\mathrm{rou}}-1)}, y_i + f_{t_{\mathrm{rou}}}(x_i)) + \Omega(f_{t_{\mathrm{rou}}}) \tag{13}$$

式中:$f_{t_{\mathrm{rou}}}(x_i)$ 为第 t_{rou} 轮增加的决策树函数;$\Omega(f_{t_{\mathrm{rou}}})$ 为第 t_{rou} 轮对应的正则项,对以上目标函数进行二阶泰勒展开,并移除常数项,可得:

$$f_{\mathrm{obj}}^{(t_{\mathrm{rou}})}(\varphi) \simeq \sum_{i=1}^{n} [l(\hat{y}_i^{(t_{\mathrm{rou}}-1)}, y_i) + g_i f_{t_{\mathrm{rou}}}(x_i) + h_i f_{t_{\mathrm{rou}}}^2(x_i)/2] + \Omega(f_{t_{\mathrm{rou}}}) \tag{14}$$

其中:

$$g_i = \frac{\partial l(\hat{y}_i^{(t_{\mathrm{rou}}-1)}, y_i)}{\partial \hat{y}_i^{(t_{\mathrm{rou}}-1)}} \tag{15}$$

$$h_i = \frac{\partial^2 l(\hat{y}_i^{(t_{\mathrm{rou}}-1)}, y_i)}{\partial (\hat{y}_i^{(t_{\mathrm{rou}}-1)})^2} \tag{16}$$

式中:g_i,h_i 分别为损失函数的一、二阶导数。

通过对目标函数进行二阶泰勒展开,同时用到了一阶导数和二阶导数,有利于模型在训练集上更快地收敛。

综合式(12)和式(14),并定义:

$$\begin{cases} G_j = \sum_{i \in I_j} g_i \\ H_j = \sum_{i \in I_j} h_i \end{cases} \tag{17}$$

则有:

$$f_{\mathrm{obj}}^{(t_{\mathrm{rou}})}(\varphi) = \sum_{j=1}^{T} [G_j \omega_j + (H_j + \lambda)\omega_j^2/2] + \gamma T \tag{18}$$

对于一个确定的树结构 q_{tree},其对应的最优化目标函数值为:

$$f_{\mathrm{obj}}^{(*)}(\varphi) = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{19}$$

由 $\frac{\partial f_{\mathrm{obj}}^{(t_{\mathrm{rou}})}(\varphi)}{\partial \omega_j} = 0$ 可得相应的最优叶节点权重:

$$\omega_j^* = \frac{-G_j}{H_j + \lambda} \tag{20}$$

式(19)可用于衡量树结构 q_{tree} 的质量。通常所有可能的树结构不可能被完全枚举出来,故 XGBoost 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} 为分裂后左右子节点的集合,设 $I = I_{\mathrm{L}} \cup I_{\mathrm{R}}$,则分裂后产生的信息增益如下:

$$G_{\mathrm{gain}} = \frac{1}{2}\left[\frac{G_{\mathrm{L}}^2}{H_{\mathrm{L}} + \lambda} + \frac{G_{\mathrm{R}}^2}{H_{\mathrm{R}} + \lambda} - \frac{(G_{\mathrm{L}} + G_{\mathrm{R}})^2}{H_{\mathrm{L}} + H_{\mathrm{R}} + \lambda}\right] - \gamma \tag{21}$$

式(21)通常用来评价分割的候选节点。

2.2 组合 XGBoost 预测模型构建

组合预测能有效综合多个单一模型的信息,减少单个模型的预测风险,提升算法整体的预测精度[23—24]。由此,文中结合前文阐述的基于 PMIC 的特征选择算法,构建以 XGBoost 为底层算法的组合预测模型。相应的训练流程如图 1 所示。

(1) 对原始特征集执行基于 PMIC 的特征选择操作。

(2) 根据选择出的风速高度将相应的数据集分成 n 个训练子集 $\{M_1, M_2, \cdots, M_n\}$ 。

(3) 利用 XGBoost 分别对 n 个训练子集进行训

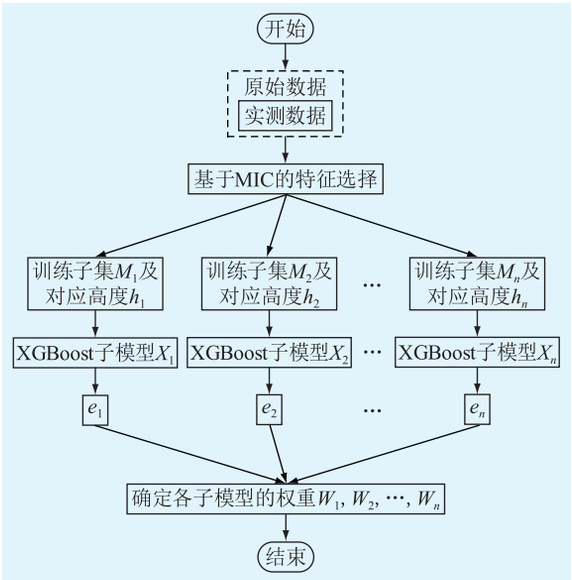

图 1　组合 XGBoost 模型训练流程

Fig.1　Training process of combined XGBoost model

练，生成 n 个子模型 $X_1, X_2, \cdots, X_n$，利用测试集及均方误差对每个子模型的预测误差 e_i 进行评估，计算公式为：

$$e_i = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2 \tag{22}$$

式中：y_t 为第 t 个样本的实测功率值；$\hat{y}_t$ 为第 t 个样本的预测功率值；N 为样本数量。

（4）利用熵权法对各子模型进行权重分配。

相应的预测流程如图 2 所示。

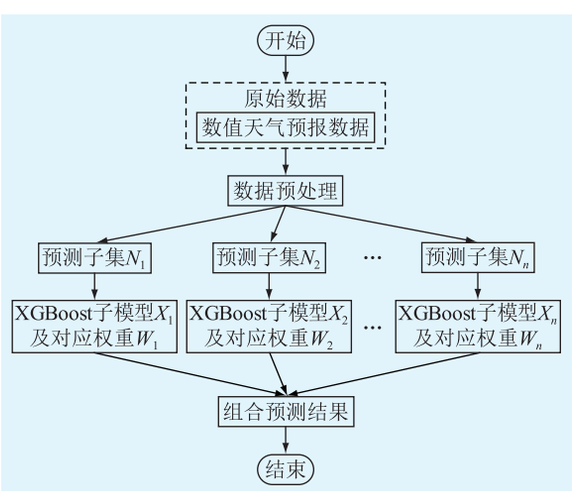

图 2　组合 XGBoost 模型预测流程

Fig.2　Forecasting process of combined XGBoost model

（1）对预测集进行包含特征选择在内的数据预处理工作，并根据风速高度将处理后的预测集分成 n 个预测子集 $N_1, N_2, \cdots, N_n$。

（2）利用训练流程生成的各 XGBoost 子模型 $X_1, X_2, \cdots, X_n$ 及对应权重 $W_1, W_2, \cdots, W_n$ 对测试样本进行计算，得出预测结果。

3　算例分析

3.1　数据集与模型评价指标

为充分论证文中所提方法性能，文中采用华东某风电场 A、西北某风电场 B 的实测数据及历史数值天气预报数据进行了相关实验。其中，风电场 A 的平均风速较小，波动相对稳定；而风电场 B 则平均风速较大，波动性较强。表 1 和表 2 分别为 2 个风电场数据的数据概况及相关风电机组的主要参数。

表 1　风电场 A 和 B 的数据概况

Table 1　The data overview of wind farms A and B

风电场	采样周期/min	数据区间	训练区间	测试区间
A	15	2017-01-01—2018-12-31	2017-01-01—2018-09-30	2018-10-01—2018-12-31
B	15	2017-08-01—2019-09-30	2017-08-01—2019-06-30	2019-07-01—2019-09-30

表 2　风电场 A 和 B 的风电机组主要参数

Table 2　Main parameters of wind turbines in wind farms A and B

风电场	装机容量/MW	风轮直径/m	风电机组轮毂高度/m
A	49.5	58	60
B	300	100	70

训练阶段的输入为测风塔不同高度的风速数据及风电场当地的平均风向、温度、湿度及气压数据，输出为实测功率。测试阶段的输入为对应的数值天气预报数据，输出为预测功率。对于短期功率预测，后一天的数值天气预报数据由预报供应商于前一天的 6 点前发布。

在数据预处理方面，采用文献[25]中数据预处理方法对训练集数据进行预处理。

在模型精度评价方面，参考国家电网颁布的《风功率预测功能规范》[26]，选用均方根误差、平均绝对误差和合格率作为风电场功率预测的精度的评价指标，各指标的具体定义如下。

均方根误差：

$$e_{\mathrm{RMSE}} = \frac{1}{S_{\mathrm{op}}}\sqrt{\frac{1}{N}\sum_{t=1}^{N}(P_t - P_{\mathrm{ft}})^2} \tag{23}$$

平均绝对误差：

$$e_{\mathrm{MAE}} = \frac{1}{N}\sum_{t=1}^{N}\frac{|P_t - P_{\mathrm{ft}}|}{S_{\mathrm{op}}} \tag{24}$$

合格率：

$$Q_{\mathrm{R}} = \frac{1}{N}\sum_{t=1}^{N}B_t \tag{25}$$

其中:

$$B_t=\begin{cases}1 & \dfrac{|P_t-P_{\mathrm{ft}}|}{S_{\mathrm{op}}}<0.25\\ 0 & \dfrac{|P_t-P_{\mathrm{ft}}|}{S_{\mathrm{op}}}\geqslant 0.25\end{cases}\tag{26}$$

式中:P_{ft} 为 t 时刻预测功率值;P_t 为 t 时刻实测功率值;S_{op} 为风电场的额定装机容量。

3.2　特征选择

为得到 2 个风电场的最优特征子集,利用 PMI 对 MIC 进行改造,构建一种基于 PMIC 的特征选择算法,相应的特征选择过程分别如图 3、图 4 所示。对于风电场 A 而言,当第 4 个特征被选出来时,T_{AIC} 最小,为-86 275。对于风电场 B 而言,当第 5 个特征被选出来时,T_{AIC} 最小,为-97 553。相应的特征选择结果如表 3 所示。

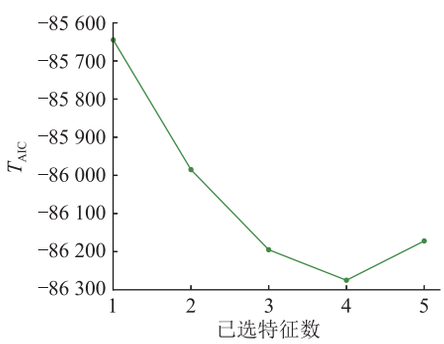

图 3　风电场 A 对应的特征选择过程

Fig.3　Feature selection process corresponding to wind farm A

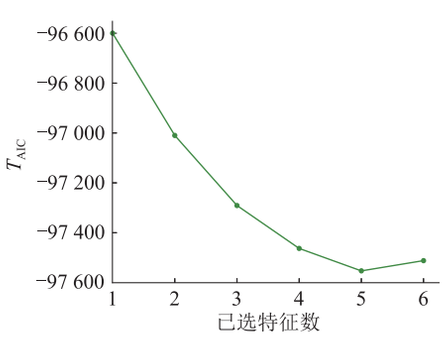

图 4　风电场 B 对应的特征选择过程

Fig.4　Feature selection process corresponding to wind farm B

表 3　风电场 A 和 B 对应的最优特征子集

Table 3　The optimal feature subsets for wind farms A and B

风电场	最优特征子集
A	$\{V_{30},V_{70},V_{90},P_{\mathrm{pres}}\}$
B	$\{V_{30},V_{70},V_{110},D_{\mathrm{dir}},T_{\mathrm{temp}}\}$

表 3 中,如文中 1.1 节所述,V_{30} ,V_{70} ,V_{90} 分别表示 30 m,70 m,90 m 层高对应的风速特征。在不同高度的风速特征选择方面,风电场 A 比风电场 B 少选了 110 m 风速,这是由于风电场 A 风机风轮扫风范围所限。在其他气候条件选择方面,风电场 B 选择了风向和温度,风电场 A 则只选择了气压,表明不同风电场的输出功率对气候条件的敏感程度不同。

3.3　算法整体预测效果分析

为了验证组合 XGBoost 模型在解决短期风功率预测问题上的有效性,文中首先将未经特征选择的单一预测模型 ARMA、SVM、BP、XGBoost、结合了 PMIC 的 XGBoost(PMIC-XGBoost)与结合了 PMIC 的组合 XGBoost(PMIC-CXGBoost)作为类比模型同时进行风功率预测,并从训练效率和模型精度两方面进行验证。

以风电场 A 中的训练数据为基础,图 5 为不同未经特征选择的单一预测模型在相同迭代次数下的训练耗时。可见,XGBoost 模型训练耗时要远低于其他模型,展现了其在模型训练效率方面的优越性。其原因在于,XGBoost 在效率上进行了多方面优化,包括基于列存储块的并行学习实现、采用缓存感知访问、外存块计算等。因此,选择 XGBoost 作为底层算法能使组合模型具备更高的计算效率。

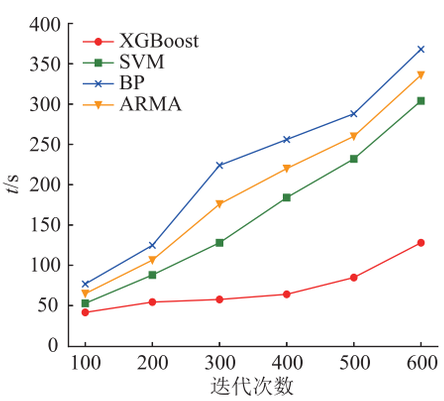

图 5　模型训练耗时与迭代次数的关系

Fig.5　Relationship between model training time and number of iterations

表 4 为不同模型对于风电场 A 和风电场 B 的预测结果。可见,风电场 A 具有远低于风电场 B 的均方根误差和平均绝对误差,同时也有更高的合格率,表明平均风速小,波动相对稳定的风电场的输出功率更好预测。当由风电场 A 切换到风电场 B 时,ARMA、SVM、BP 的预测效果均急剧下降,而基于 XGBoost 模型,即 XGBoost,PMIC-XGBoost,PMIC-CXGBoost 仍能有 10%左右的均方根误差及 7%以下的平均绝对误差,合格率则均在 87%上。统计结果表明,XGBoost 在模型精度方面相较其他单一预测模型更加优越且稳定,这是因为 XGBoost 的设计同

时考虑了学习能力与泛化能力。在学习能力方面，XGBoost 是一种由若干决策树组成的集成学习模型，决策树的个数理论上可以任意大，意味着 XGBoost 可以更灵活地对训练样本进行拟合，从样本中学习更加丰富的信息。另外，XGBoost 针对目标函数进行二阶泰勒展开，在对一阶导数进行保留的同时加入二阶导数，能够使模型在训练集上更容易收敛。在泛化能力方面，XGBoost 算法在正则项上对单颗树中叶节点的数量及叶节点的权重均进行了控制，避免模型出现过拟合，进而提升模型在训练集外的泛化能力。

表 4　不同模型的预测结果

Table 4　The forecasting results of different models

%

模型	风电场 A			风电场 B		
	e_{RMSE}	e_{MAE}	Q_R	e_{RMSE}	e_{MAE}	Q_R
ARMA	9.56	8.14	88.62	14.71	10.32	84.11
SVM	9.25	7.52	88.95	14.18	9.45	84.45
BP	9.13	7.43	89.13	13.98	9.61	84.77
XGBoost	8.04	6.54	89.88	10.39	7.67	87.38
PMIC-XGBoost	6.65	5.27	91.53	8.72	6.51	88.81
PMIC-CXGBoost	4.96	3.43	93.97	6.88	4.41	91.42

比较单一 XGBoost 和 PMIC-XGBoost：对于风电场 A，PMIC-XGBoost 的均方根误差和平均绝对误差相较单一 XGBoost 分别下降了 1.39%和 1.27%，合格率提升了 1.65%；对于风电场 B，PMIC-XGBoost 的均方根误差和平均绝对误差也相较单一 XGBoost 分别下降了 1.67%和 1.16%，合格率提升了1.43%。该结果验证了基于 PMIC 的特征选择算法的有效性。

比较 PMIC-XGBoost 和 PMIC-CXGBoost：无论是风电场 A 还是风电场 B，PMIC-CXGBoost 模型整体的均方根误差、平均绝对误差都要小于 PMIC-XGBoost，其合格率也高于 PMIC-XGBoost。其中，风电场 A 的均方根误差和平均绝对误差分别相较 PMIC-XGBoost 低了 1.69%和 1.84%，合格率相较高了 2.44%；风电场 B 的均方根误差和平均绝对误差分别相较 PMIC-XGBoost 低了 1.84%和 2.1%，合格率相较高了 2.61%。该结果表明，引入组合预测的思想后，预测效果有了进一步的提升。

表 5 比较了文中组合预测方法 PMIC-CXGBoost 与其他相近组合预测方法的性能，类比模型为文献[16]中的集成学习预测模型，底层算法的迭代次数设为 300。文献[16]中的模型也将不同高度的风速作为模型的输入，底层算法则采用了 BP 神经网络、线性外推和线性 SVM，对应风电场 A 和风电场 B。与文献[16]中的集成学习预测模型相比，文中所提组合 XGBoost 不仅具有更优的精度，也具备更短的训练时间。

表 5　文中组合预测方法 PMIC-CXGBoost 和文献[16]中的预测方法的对比

Table 5　Comparison of the proposed combination forecasting method PMIC-CXGBoost and the forecasting method in reference 16

风电场	模型	评价指标			
		e_{RMSE}/%	e_{MAE}/%	Q_R/%	训练时间/s
A	文献[16]中的方法	8.09	6.73	89.42	284
	PMIC-CXGBoost	4.96	3.43	93.97	153
B	文献[16]中的方法	10.96	7.82	87.21	302
	PMIC-CXGBoost	6.88	4.41	91.42	175

图 6 和图 7 分别为文中方法在 2 个风电场中风功率预测值和实测值的比较。可以看出，文中方法能很好地预测实测序列的变化趋势。

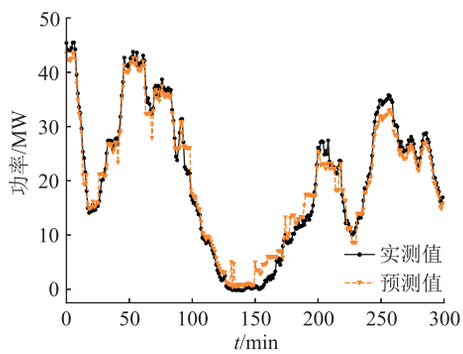

图 6　风电场 A 中组合 XGBoost 预测值与实测值的比较

Fig.6　Comparison of forecasting and measured values of combined XGBoost in wind farm A

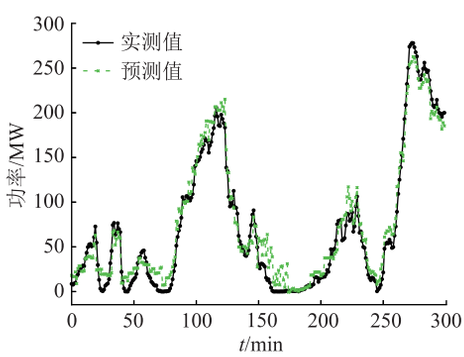

图 7　风电场 B 中组合 XGBoost 预测值与实测值的比较

Fig.7　Comparison of forecasting and measured values of combined XGBoost in wind farm B

4 结语

针对当前短期风功率预测中存在的精度以及工程化应用问题,文中提出一种将 PMIC 特征选择与组合 XGBoost 相结合的预测模型。一方面,引入 PMI 对 MIC 进行改造,使相关特征选择算法不仅能得到对风功率影响程度较大的气象特征,也有利于消除变量间的耦合关系。另一方面,为兼顾算法的精度和效率,减少单个模型的预测风险,采用 XGBoost 作为底层算法构建组合预测模型。将 2 个具有较大差异的风电场作为算例进行验证,实验结果表明,结合了 PMIC 特征选择的组合 XGBoost 模型不仅在精度方面效果理想,在计算效率方面,也较相近组合预测模型有更好的效果,便于工程化应用。

在下一步工作中,将考虑将误差修正技术引入组合 XGBoost 预测模型中,使得算法整体上具备更好的反馈能力,以进一步提升短期风功率的预测精度。

参考文献:

[1] VARGAS S A,ESTEVES G R T,MACAIRA P M,et al. Wind power generation:a review and a research agenda[J]. Journal of Cleaner Production,2019,218:850-870.

[2] 惠小健,王震,张善文,等. 基于 ARMA 的风电功率预测[J]. 现代电子技术,2016,39(7):145-148,153.
XI Xiaojian,WANG Zhen,ZHANG Shanwen,et al. Wind power forecast based on ARMA[J]. Modern Electronics Technique,2016,39(7):145-148,153.

[3] 牛哲文,余泽远,李波,等. 基于深度门控循环单元神经网络的短期风功率预测模型[J]. 电力自动化设备,2018,38(5):36-42.
NIU Zhewen,YU Zeyuan,LI Bo,et al. Short-term wind power forecasting model based on deep gated recurrent unit neural network[J]. Electric Power Automation Equipment,2018,38(5):36-42.

[4] 崔嘉,杨俊友,杨理践,等. 基于改进 CFD 与小波混合神经网络组合的风电场功率预测方法[J]. 电网技术,2017,41(1):79-85.
CUI Jia,YANG Junyou,YANG Lijian,et al. New method of combined wind power forecasting based on improved CFD and wavelet-HNN model[J]. Power System Technology,2017,41(1):79-85.

[5] LIU Y Q,SHI J,YANG Y P,et al. Short-term wind-power prediction based on wavelet transform-support vector machine and statistic-characteristics analysis[J]. IEEE Transactions on Industry Applications,2012,48(4):1136-1141.

[6] OLIVEIRA V,SOUSA V,DIAS-FERREIRA C. Artificial neural network modelling of the amount of separately-collected household packaging waste[J]. Journal of Cleaner Production,2019,210:401-409.

[7] LUO C,JIANG Z P,ZHENG Y J. A novel reconstructed training-set SVM with roulette cooperative coevolution for financial time series classification[J]. Expert Systems With Applications,2019,123:283-298.

[8] MEENAL R,SELVAKUMAR A I. Assessment of SVM,empirical and ANN based solar radiation prediction models with most influencing input parameters[J]. Renewable Energy,2018,121:324-343.

[9] CHEN T Q,GUESTRIN C. XGBoost:a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. New York,NY,USA:ACM,2016:785-794.

[10] 何龙. 深入理解 XGBoost:高效机器学习算法与进阶[M]. 北京:机械工业出版社,2020.
HE Long. Depth understanding XGBoost:efficient and advanced machine learning algorithms[M]. Beijing:China Machine Press,2020.

[11] 彭曙蓉,郑国栋,黄士峻,等. 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.
PENG Shurong,ZHENG Guodong,HUANG Shijun,et al. Multiple-feature short-term photovoltaic generation forecasting based on XGBoost algorithm[J]. Electrical Measurement & Instrumentation,2020,57(24):76-83.

[12] MUNAWAR U,WANG Z L. A framework of using machine learning approaches for short-term solar power forecasting[J]. Journal of Electrical Engineering & Technology,2020,15(2):561-569.

[13] 王桂兰,赵洪山,米增强. XGBoost 算法在风机主轴承故障预测中的应用[J]. 电力自动化设备,2019,39(1):73-77,83.
WANG Guilan,ZHAO Hongshan,MI Zengqiang. Application of XGBoost algorithm in prediction of wind motor main bearing fault[J]. Electric Power Automation Equipment,2019,39(1):73-77,83.

[14] 陈明华,刘群英,张家枢,等. 基于 XGBoost 的电力系统暂态稳定预测方法[J]. 电网技术,2020,44(3):1026-1034.
CHEN Minghua,LIU Qunying,ZHANG Jiashu,et al. XGBoost-based algorithm for post-fault transient stability status prediction[J]. Power System Technology,2020,44(3):1026-1034.

[15] 凌立文,张大斌. 组合预测模型构建方法及其应用研究综述[J]. 统计与决策,2019,35(1):18-23.
LING Liwen,ZHANG Dabin. A review of construction and application of combination forecast model[J]. Statistics & Decision,2019,35(1):18-23.

[16] 刘克文,蒲天骄,周海明,等. 风电日前发电功率的集成学习预测模型[J]. 中国电机工程学报,2013,33(34):130-135,21.
LIU Kewen,PU Tianjiao,ZHOU Haiming,et al. A short term

wind power forecasting model based on combination algorithms [J]. Proceedings of the CSEE,2013,33(34):130-135,21.

[17] 钱政,裴岩,曹利宵,等. 风电功率预测方法综述[J]. 高电压技术,2016,42(4):1047-1060.
QIAN Zheng,PEI Yan,CAO Lixiao,et al. Review of wind power forecasting method[J]. High Voltage Engineering,2016,42(4):1047-1060.

[18] 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.
FAN Gaofeng,WANG Weisheng,LIU Chun,et al. Wind power prediction based on artificial neural network[J]. Proceedings of the CSEE,2008,28(34):118-123.

[19] 梁正堂. 基于不确定性特征挖掘的风电系统预测与决策理论研究[D]. 济南:山东大学,2017.
LIANG Zhengtang. Theoretical studies on wind power system forecasting and decision-making based on characteristic mining of uncertainty[D]. Jinan:Shandong University,2017.

[20] 杨贺钧. 计及多因素的含风能电力系统可靠性评估及优化研究[D]. 重庆:重庆大学,2014.
YANG Hejun. Reliability evaluation and optimization models of power systems containing wind energy considering mutliple factors[D]. Chongqing:Chongqing University,2014.

[21] RESHEF D N,RESHEF Y A,FINUCANE H K,et al. Detecting novel associations in large data sets[J]. Science,2011,334(6062):1518-1524.

[22] MAY R J,MAIER H R,DANDY G C,et al. Non-linear variable selection for artificial neural networks using partial mutual information[J]. Environmental Modelling & Software,2008,23(10/11):1312-1326.

[23] 张妍,王东风,韩璞. 一种风电场短期风速组合预测模型[J]. 太阳能学报,2017,38(6):1510-1516.
ZHANG Yan,WANG Dongfeng,HAN Pu. Combination forecasting model of short-term wind speed for wind farm[J]. Acta Energiae Solaris Sinica,2017,38(6):1510-1516.

[24] 周滢,任海军,李健,等. 层次结构下的中长期电力负荷变权组合预测方法[J]. 中国电机工程学报,2010,30(16):47-52.
ZHOU Quan,REN Haijun,LI Jian,et al. Variable weight combination method for mid-long term power load forecasting based on hierarchical structure[J]. Proceedings of the CSEE,2010,30(16):47-52.

[25] 周永华,张国建,李科,等. 基于最佳预测步长的超短期风电功率预测[J]. 广东电力,2015,28(8):19-22,54.
ZHOU Yonghua,ZHANG Guojian,LI Ke,et al. Prediction on ultra-short-term wind power based on optimal predictive time length[J]. Guangdong Electric Power,2015,28(8):19-22,54.

[26] 风电功率预测功能规范:QGDW 10588—2015[S]. 北京:国家电网公司,2015.
Function specification of wind power forecasting: QGDW 10588-2015[S]. Beijing: State Grid Corporation of China,2015.

作者简介:


李科

李科(1982),男,硕士,高级工程师,从事电力系统自动化、新能源预测技术研究工作(E-mail:like@sgepri.sgcc.com.cn);

黄东晨(1990),男,硕士,工程师,从事新能源预测、数据挖掘技术研究工作;

陶子彬(1996),男,学士,助理工程师,从事风光发电站功率预测系统软件研发、预测算法研究工作。

Combined XGBoost short-term wind power forecasting model based on partial maximum information coefficient

LI Ke, HUANG Dongchen, TAO Zibin, XIONG Huan, LI Haowen, DU Yedong

(NARI Group (State Grid Electric Power Research Institute) Co.,Ltd.,Nanjing 211106,China)

Abstract:As one of the hot topics in the field of new energy forecasting,it is necessary for the research of short-term wind power forecasting to pay attention to the engineering application of the model while improving forecasting accuracy. Hence,a combined XGBoost forecasting model based on partial maximum information coefficient is proposed. To begin with,a feature selection algorithm based on partial maximum information coefficient is designed. By introducing partial mutual information,while mining meteorological features that have a greater impact on wind power,it can also eliminate the adverse effects of coupled information. On this basis,in order to take the accuracy and computational efficiency of the model into account and reduce the forecasting risk of a single model,a combined forecasting model with XGBoost as the underlying algorithm is constructed to further realize wind power forecasting. Two wind farms with large differences are used as examples for verification analysis. The results show that the combined XGBoost forecasting model based on partial maximum information coefficient feature selection algorithm can not only improve the forecasting accuracy of short-term wind power,but also has higher calculation efficiency compared with similar combined forecasting models,which is beneficial to engineering application.

Keywords:feature selection;combined XGBoost;partial maximal information coefficient;short-term wind power forecasting;calculation efficiency;engineering application

(编辑　钱悦)