文章编号：2096-1618(2020)06-0604-06

基于 XGBoost 的 10 m 风速订正研究

毛开银， 赵长名， 何 嘉
(成都信息工程大学,四川 成都 610225)

摘要：基于当前气象预报模式,风速预报的精确度存在一定的误差,国内外研究者对风速的预报订正做了大量的研究。提出一种 CD-XGBoost(clustering and double XGBoost)算法,该算法针对现有的机器学习风速预报订正算法进行改进,主要包含以下 3 个改进方向：第一,提出利用天气元素与订正元素之间的相关性进行聚类的思路,通过对簇间站点进行基于不同机器学习模型的训练,提高风速订正结果的准确性;第二,算法突出空间因素对风速预报的影响,选取气象观测站点的 K 个邻近预报网格点的预报元素构建数据集,相较传统插值进行订正的方式更多地考虑站点与网格点之间的空间因素。第三,提出使用2 个不同起报时刻数据独立训练 XGBoost 模型,对模型的输出再进行线性权重加和得到最终订正结果的算法。仿真实验中,采用中国 2552 个气象观测站的逐3 h观测数据与欧洲中期气象预报中心(ECWMF)的数值模式的逐3 h预报的数据,对 ECWMF 预报的地面10 m风速进行观测站点订正。使用改进后的算法构建的模型与目前算法构建的模型进行比较,结果证明文中算法在风速预测准确度明显优于目前的订正算法,其中3 h预报时效时准确率高于 85%,168 h预报时效是其准确率高于 60%,具有很好的应用前景。

关 键 词：机器学习;风速订正;聚类;XGBoost
中图分类号：TP301.6 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2020.06.004

0 引言

目前天气预报的主要方式是采用数值天气预报(NWP)产品对天气元素进行预报。数值天气预报产品的预报方式是将气象观测元素数据,地理信息数据等作为模式产品的输入,模式系统经过物理学、大气学、空气动力学等构建的方程组进行预演推算得到天气元素的预报结果[1]。但是由于大气系统的混沌性、不确定性,数值天气预报总存在不同程度的误差。为提高数值天气预报模式产品的预报准确率,集合预报是常用的手段。目前计算技术的不断发展也使数字模式预报越来越精确,但是整个数字预报过程中也存在许多干扰因素,最终使预报结果因为系统错误或初始场数据的错误导致预报结果存在不同程度的误差。所以在数值天气预报产品的预报结果上对预报元素做进一步的订正有必要性。

风速预测是气象预报的重要预报之一,其预报准确性对日常生活、工业生产具有重要意义。风速预报是数值天气预报的常规预报元素,由于其预报偏差一直存在,所以国内外学者对风速预报做了大量的研究。目前基于数值天气预报的订正方法主要包括时序法[2]、自回归差分移动平均模型[3-4]、高斯统计方法[5-6]、卡尔曼滤波[7]、神经网络[8-11]、支持向量机[12-13]、极限机学习法(extreme learning machine, ELM)[14-16]等。这些方法又可分为统计与机器学习类方法、滑动平均法与卡尔曼滤波法等。滑动平均法与卡尔曼滤波法等对历史资料的要求相对较小且对计算效率高,主要考虑了风速变化也存在一定的连续性,对系统错误订正效果好,但对于转折性变化效果较差。统计与机器学习的方法对历史资料数量与计算机计算能力有一定要求,通过历史资料建模,寻找各气象元素及气象元素的时空信息等存在的联系,进一步订正模式预报结果的误差。随着计算机性能提升、机器学习算法的改进、气象历史资料的积累,机器学习算法应用于气象订正的研究越来越广泛。Abo-Khalil A G 等基于支持向量回归算法使用预测数据与 17 个风电场观测数据对风电场的最大风速进行预报,并验证了该模型的有效性,但对 17 个风电场的数据仅有一个模型进行回归缺少考虑地理环境影响气象元素之间的关系。邓华等[17]采用了 PCA 结合径向基神经网络(RBF)使用 WRF 模式预报数据与观测数据对风电场的风速进行了订正,验证了经过 PCA 降维后订正效果得到了提高,但仅考虑了 WRF 预报的元素。孙全德等[8]采用了 Lasso 回归进行特征选择并且使用 3 层人工神经网络对 ECWMF 预报的中国华北地区风速进行订正,对比随机森林等算法,得出进行特征选择后的模型对风速订正效果更好,但是实验只是用了 ECWMF 数据,使用模式初始场数据作为标签数据,而模式初始场的数据是经过在分析的数据并非实际风速值,缺乏实际应用性。文献[18]通过 BP 神经网络和 PCA 对中国与西班牙的两个风场数据做风速订正,提高了风速的预测精度,但仅采用当前时刻气象元素对风速进行预测。

收稿日期：2019-12-13
基金项目：国家重大专项资助项目(2017YFG502203);国家重点研发资助项目(2019YFG0212);四川省科技计划资助项目(2019YFG0212);四川省科技计划资助项目(2018GZ0814)

文献[19]使用聚类方法结合 SVM 对大型风场进行基于风速订正的风能预报研究,但其聚类的过程采用动力公式参数与风速作为聚类参数,欠缺其他气象元素对风速的影响。

提出采用聚类的方式将站点聚类,建立多个机器学习模型,提高模型对数据的拟合度进而提高风速的订正效果。在聚类算法上,采用原始特征对风速订正的贡献度作为聚类数据集,更充分考虑站点气象元素与风速之间的关系。采用站点邻近网格点预报数据与观测站点的气象元素观测值合并的算法,构建原始数据集。改进气象观测站点与预报网格点之间因为位置不同采用插值的算法,因为插值过程仅仅考虑水平距离的预报值的影响,而忽略了不同位置之间的其他因素对预报元素的影响。但构建的原始数据集的样本特征较多,为减少相关度较低的元素影响模型的拟合度,采用特征选择算法对构建的原始特征进行选择。最后使用构建的数据样本对 XGBoost 模型进行训练,得到站点风速订正模型。

1　相关基础理论

CD-XGBoost 算法中基于机器学习的数值模式预报的后处理任务是一个典型的数据挖掘类任务,它包含 3 个阶段:数据预处理,模型训练,模型评估,在这些阶段中使用一些相关的算法与模型。在数据预处理阶段采用 KMeans 聚类算法对站点进行聚类,在模型训练阶段使用 XGBoost 构建机器学习模型。在评估阶段为比较插值运算法与文中算法,使用反距离插值算法。

1.1　KMeans 聚类算法

KMeans 是一种基于分割聚类算法的无监督学习算法[18]。它的基本思想是在给定的样本集 $S=\{s_1,s_2,s_3,\cdots,s_n\}$ 中按照样本之间的距离,将样本分割为 k 个簇 $C=\{c_1,c_2,c_3,\cdots,c_n\}$。聚类完成后的理想结果是簇内样本的距离近,但簇与簇之间的距离尽量大。所以聚类的目标就是最小化簇内误差平方 SSE (within-cluster sum of squared errors),其公式为

$$\mathrm{SSE}=\sum_{i=1}^{n}\sum_{j=1}^{m}w^{(i,j)}=\parallel x^{(i)}-\mu^{(j)}\parallel_2^2 \tag{1}$$

其中 μ 是簇 j 的均值向量,也可以称为质心。KMeans 算法流程,如图 1 所示。

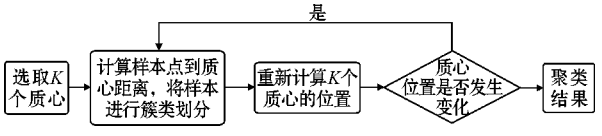

图 1　KMeans 算法流程

KMeans 算法中 K 值是一个超参数,需要进行人为的指定,K 值的合理性对于聚类结果有较大的影响。K 值的选择可以通过手肘法,手肘法的基本思想是随着聚类数 K 的增大,样本划分会更加精细,簇内的聚合程度就会提高,平方误差和变小;反之随着 K 的减小簇内的距离会增大,平方误差和也会随之变大。当 K 值小于合理的簇类个数时,K 值增大 SSE 下降的幅度会较大;当 K 值大于合理簇数时,K 值变大 SSE 下降的幅度会比较平缓。

1.2　XGBoost 机器学习算法

XGBoost (extreme gradient boosting)是陈天奇等开发的一个开源机器学习项目[20],高效地实现了 GBDT (gradient boosting decision tree)算法并进行了算法和工程上的许多改进,是一种集成学习算法,既可以用于分类任务也可以用于回归任务。

XGBoost 通过建立 k 个回归树,使得树群的预测值尽量接近真实值而且有尽量大的泛化能力,从数学角度看这是一个泛函最优化,其目标函数为

$$L(\varphi)=\sum l(\hat{y}-y_i)+\sum\Omega(f_x) \tag{2}$$

其中 i 表示第 i 个样本,$l(y-y_i)$ 表示第 i 个样本的预测误差,误差越小越好,$\Omega(f_x)$ 表示第 x 颗树的正则项,即树的复杂度的函数,越小复杂度越低,泛化能力越强。其表达式为

$$\Omega(f_x)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{3}$$

其中 T 为树模型的叶子个数,w_j 为得第 j 个叶子节点上的得分。

XGBoost 是以 CART 树中的回归树作为基分类器,它是一种加法模型,将模型上次预测(由 $t-1$ 棵树组合而成的模型)产生的误差作为参考进行下一棵树(第 t 棵树)的建立。以此,每加入一棵树,将其损失函数不断降低。其第 t 棵的函数表达为

$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{4}$$

其中 f_k 表示第 k 棵树,$\hat{y}_i^{(t)}$ 表示组合 t 树模型对样本 x_i 的预测结果。将式(4)带入到式(2)得到目标函数为

$$\mathrm{Obj}^{(t)}=\sum l[y_i,\hat{y}_i^{(t-1)}+f_t(x_i)]+\Omega(f_t)+C \tag{5}$$

将以上目标函数进行泰勒展开,并引入正则项,最终得到目标函数为

$$\mathrm{Obj}'(\theta)\cong\sum_{j=1}^{T}[g_if_i(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t)+C \tag{6}$$

其中 g_i 和 h_i 分别是在损失函数 $l(y^t,y^{(t-1)})$ 上对 $\hat{y}_i^{(t-1)}$ 所求的一阶导数和二阶导数,统称为每个样本的梯度统计量。

1.3　反距离插值

反距离加权法是非规则分布点变成规则分布点常用的网格化方法之一。该方法的基本思想是离所估算的网格点距离越近的离散点对该网格点的影响越大,越远的离散点影响越小,甚至可以认为没有影响。在估算某一网格点的值时,假设离网格点最近的 N 个点对其有影响,那么这 N 个点对该网格点的影响与他们之间的距离成反比。算法步骤如下:

需要计算所有离散数据点与所求网格点的距离,在二维平面空间,离散点(x_i,y_i)到网格(x_0,y_0)的距离

D_i 为

$$D_i=\sqrt{(x_0-x_i)^2+(y_0-y_i)^2} \tag{7}$$

计算 N 个离散点上的值差值到点(x_0,y_0)上的值，计算公式为

$$V=\sum_{i=1}^{N}\frac{(\frac{1}{D_i})^p}{\sum_{j=1}^{N}(\frac{1}{D_j})^p}V_i \tag{8}$$

其中 V_i 是离散点 i 上的值，V 即为插值到(x_0,y_0)上的估算值。D_i 为插值点与第 i 个离散点间的距离，p 是距离的幂，一般取 2。

1.4 检验方法

使用常用于气象预报领域的检验方法如均方根误差(RMSE)、平均绝对误差(MAE)与预报准确率。

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i-o_i)} \tag{9}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=0}^{N}|p_i-o_i| \tag{10}$$

其中，p 为风速的预测值，o 为风速观测值。风速预报的准确率是指风速预报的绝对误差不大于1 m/s的百分率，其公式表达为

$$\mathrm{Acc}=\frac{N_r}{N}\times 100\% \tag{11}$$

其中 N_r 为预测风速值与气象站观测风速值之间绝对误差不大于1 m/s的样本个数，N 为预报的样本个数。

2 系统模型与算法

提出的 CD-XGBoost 风速订正算法的系统模型主要基于 4 个子算法模块，如图 2 所示。各子模块采用的算法和简单描述如下：

(1)数据构建模块。构建算法时采用了 ECWMF 的网格预报数据和气象观测站点观测数据作为原始数据，预报数据采用临近网格点合并的方式，观测数据的构建采用滑动时间窗口算法。

(2)模型选择器模块。引入了模型选择器的改进理念，主要是为解决不同地区匹配的地域模型特征差异较大的问题。在模型选择器的构建上采用了特征贡献度与 KMeans 聚类等算法。

(3)数据预处理模块。主要是解决数据本身可能存在量纲不同、特征维度过高等问题。

(4)订正模型模块。ECWMF 在同一预报时刻存在 2 组起报时刻数据。提出的订正模型对 2 组预报数据分别采用 XGBoost 机器学习算法进行模型训练，得到中间订正结果。对 2 个独立训练出的中间订正结果再采用线性权重加和的方式得出最终订正结果。此方法可有效地对预报数据进行进一步订正。

以上 4 个系统模型模块与提出的改进算法相辅相成。提出算法的第一个改进点体现在数据构建模块中针对现有插值算法进行改进，充分考虑其他预报元素对风速预报的影响权重。算法的第二个改进点体现在模型选择器模块中元素相关性的引入，使模型对元素特征拟合度更高。算法的第三个改进点体现在订正模型中充分考虑了数值模式对同一时刻的 2 次预报中初始场的影响。

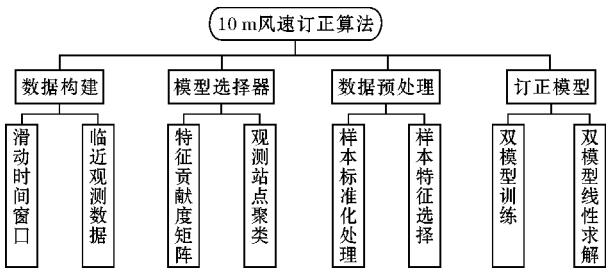

图 2 风速订正算法模型结构图

2.1 数据构建模块功能描述

模块是采用 ECWMF 的预报数据与中国气象观测站数据，构建适合机器学习模型需要的单一结构化数据，即包含样本特征数据 X，与样本标签值 y。

因为预报数据与观测数据存在分辨率不同的问题，即两种数据集描述的气象元素值对应的地理位置不同，目前研究都采用各类插值算法进行处理，而文中认为插值算法仅考虑了地理位置上的距离问题，忽略掉元素之间的联系，故采用直接选取观测站近邻的 K 个网格点预报值合并的方式提取预报数据。观测数据提取采用滑动时间窗口算法，即从预报点对应的起报时刻回滚一个时间窗口，提取在该时间窗口内的所有观测数据作为样本的属性值。综上，提取后的样本特征集可以表示为

$$X=(O_1,O_2,\cdots,O_t,P_1,P_2,\cdots,P_k) \tag{12}$$

其中，O 表示提取的观测数据，P 表示提取的预测数据。t 表示观测数据的时间滚动窗口大小，即历史观测时间点的个数。以 0 时数据为例，当窗口大小 $t=5$ 时观测数据集包含了 0 时、21 时、18 时、15 时、12 时的观测数据。K 表示距离观测站点最近的站点数。

2.2 模型选择器模块功能描述

为更加合理地拟合各站点气象元素对未来天气元素变化的影响，采用多个机器学习模型完成多个站点的预报结果订正，并采用基于站点构建数据特征对预报元素的贡献作为聚类算法的数据集。其计算过程的流程如图 3 所示。步骤详细描述如下：按观测站点归类数据集，分别采用 XGBoost 进行训练；获取各站点的样本特征对训练好的模型的贡献度；组合各站点的样本特征贡献度，构建一个以样本特征贡献度为属性的数据集；使用 KMeans 算法对构建的数据集进行聚类，将观测站点聚到不同类簇中。

图 3　气象观测站聚类流程

具体算法内容如下。

算法 1　模型选择器构建.

输入:站点数据集,DataSet;站点集合,Stations。

输出:站点模型选择器,Map<StationId,ClusterId>。

```
1:feature_importance = Array( );//初始化一个数组
2:for all Station such that Station ∈ Stations do
3:    Data = DataSet. get( Station);
4:    model = X G Boost. train( data. X, data. y);
5:    feature_importance. append( model. feature_importance_);//得到模型的特征贡献度
6:end for;
7:kmeans = K M eans( n_clusters = 100). fit( feature_importance);
8. selector = dict( Stations, kmeans, label_);
9:return selector。
```

2.3　数据预处理模块功能描述

在机器学习中数据预处理是一个不可或缺的阶段。构建的数据集存在样本量纲不同、样本特征维度较多以及样本噪声较大等问题。提出的算法采用了数据标准化算法对特征进行标准化,并使用特征选择算法[21]对数据特征进行降维。

2.4　订正模型模块功能描述

根据 ECWMF 预报数据具有不同起报时刻的特点,算法提供了一种基于双模型线性加权的数据订正方法,其理论架构如图 4 所示。算法中涉及的双模型概念就是指 ECWMF 在同一时刻点的预报包含格林威治时间(UTC)0 时和 12 时起报的预报数据。因此可基于 XGBoost 对 2 组预报数据独立进行订正模型的训练,得到 2 组中间结果。算法对于 2 组中间结果再进行线性加权求和得到最后的订正结果,其公式为

$$P=\varepsilon P_0+(1-\varepsilon)P_{12} \tag{13}$$

其中,P 为最后的订正结果,P_0P_{12}分别为基于 0 时与 12 时起报数据训练的模型订正预测值。ε 为超参数,实验中采用网格搜索算法对其进行调参。

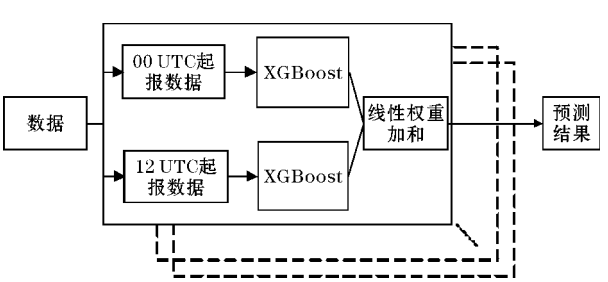

图 4　基于双模型线性加权的算法结构

2.5　ECWMF 对站点插值算法

在实验阶段多次用到插值算法对文中算法进行验证,采用反距离插值法对模式预报的数据进行站点位置风速插值。插值计算公式为

$$ws=\sqrt{\left(\sum_{i=0}^{k}\frac{v_i}{w_i}\right)^2+\left(\sum_{i=0}^{k}\frac{u_i}{w_i}\right)^2} \tag{14}$$

其中,k 表示选取的站点邻近网格点个数;v 表示经向风分量,即南北方向分量,值为正则表示南风,方向指向北,为负则方向相反;u 表示纬向风分量,即东西方向分量,值为正时表示西风,方向由南指向北,为负时则方向相反;w 为网格点的反向距离权重。

3　数据及实验

3.1　原始数据

实验选取欧洲中期预报中心预报数据集与中国 2552 个站点的观测数据集。数据集中时间区间为 2017 年 7 月到 2019 年 9 月,其预报与观测周期都是3 h,预报时长为 168 h。其中数值模式预报数据中选取了 28 个气象元素其全球个点为 481×801 个网格点,中国区站点中观测有 16 个气象元素。数据集简要说明见表 1。

表 1　风速订正原始资料信息

名称	数据大小	元素个数	观测/预报周期
ECWMF 预报数据	2017-07-01 至 2019-09-31	28	3 h
气象站观测数据	2017-07-01 至 2019-09-31	16	3 h

3.2　实验流程

首先验证 ECWMF 预报风速对中国的预报效果,确定其预报的风速经过插值运算得到的结果与观测值之间的误差在合理范围。为验证文中提出算法的有效性,实验采用传统算法与文中算法构建的数据模型在风速订正任务上的表现进行对比实验。实验中将数据集划分为训练数据集和测试数据集。为了保证每个站点的数据被划分为训练数据与测试数据,划分方式按站点分别按占比2∶8的比例进行划分。使用 2 组不同算法构建的训练集对 XGBoost 机器学习模型进行订正, 最后训练好的模型对测试数据进行预测,2 组预测结果采用均方根误差进行对比评估。最后为验证模型对风速进行订正后的预报准确率,对比与分析了 2 种算法建立的模型与 ECWMF 插值结果对风速预报在各预报时刻上的平均预报准确率。

图 5 为 ECWMF 预报风速插值到中国气象观测站点位置后的均方根误差与平均绝对误差随预报时效变化而变化的情况。其中 y 轴表示预报误差值,其 x 轴表示预报时效,例如 t+27 表示 t 时刻预报未来第27 h 的风速误差。从中可以看出插值后预报的均方根误差在2 m/s以下,平均绝对误差在1.3 m/s以下。从分析

结果来看,ECWMF 预报的风速在气象观测站点上的插值结果在合理范围内,具有一定的研究价值和订正空间。风速的数值模式系统预报的误差值基本随着预报时长的加长,误差值也随之变大。

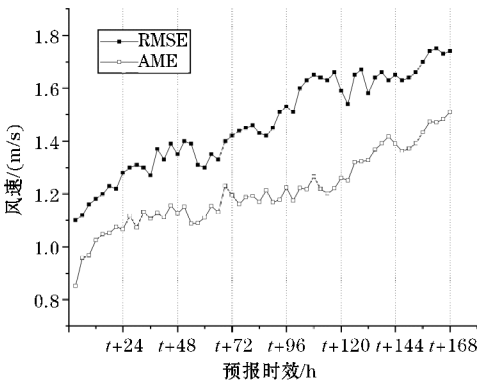

图 5 ECWF 预报风速插值站点后的均方根误差与绝对误差

图 6 为风速预报的 RMSE 误差。其中,x 轴为预报时效,y 轴为风速预报对应的 RMSE 值。CD-XGBoost、XGBoost 分别表示采用文中算法与传统算法构建数据集,基于 XGBoost 模型建模后,对测试数据风速订正后的 RMSE 值。EC 表示采用插值算法得到的 RMSE 值。EC 表示 NWP 预报结果对站点进行插值后的预报 RMSE 值。从图 5 可以看出 2 种算法订正后,相对 EC 插值结果的 RMSE 值较低,说明采用 XGBoost 模型对风速订正具有可行性。CD-XGBoost 订正后的 RMSE 值相较 XGBoost 订正后的 RMSE 值低,说明采用文中算法对风速进行订正效果更佳。CD-XGBoost 模型对测试数据风速订正后的误差值低于了1.4 m/s。订正后的风速也随着预报时长的加长,其 RMSE 值越来越大,体现了起报时刻的气象元素特征与预报点风速之间关系,随着预报时长的加长其相关性逐渐变低。

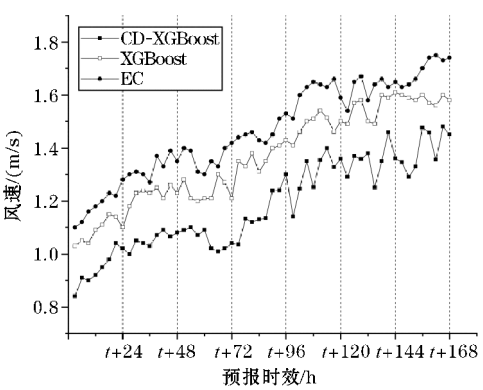

图 6 风速预报误差

图 7 为风速预报的准确率。x 轴为预报时效,y 轴为预报准确率,使用 EC 表示订正前预报准确率,使用 CD-XGBoost 表示提出的风速订正算法,XGBoost 表示传统的风速订正算法。可以看出 3 种情况下预报的准确率都随着预报时效的变长而降低。改建后的模型(CD-XGBoost)预报的准确率明显高于未改建的模型。且 CD-XGBoost 模型在预报时效为3 h时预报准确率高于 85%,在预报时效为168 h时其平均准确率高于 60%。

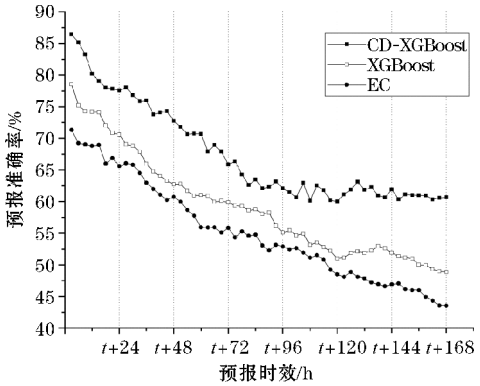

图 7 风速预报准确率

4 结束语

提出一种针对风速预报进行订正的 CD-XGBoost 算法。相较于现有风速订正算法而言,该算法充分考虑空间环境因素对风速预报的影响。并且算法对于当前难以采用单一的模型对大型风场风速进行精确订正建模的问题,提出采用模型选择器对大型风场中不同地域的风速选用具有不同气象特征的模型进行匹配。该模型选择器采用特征贡献度聚类技术,依据风速与其他气象特征之间的物理联系对站点的影响进行独立建模,从而提高模型的数据拟合度。最后,为提高模型的泛化能力,算法还对数据进行特征选择操作,降低数据集的特征维度。仿真部分采用了 ECWMF 预报的中国风速数据进行实验,实验中计算测量了传统算法与文中算法订正后的风速 RMSE 值,证实文中算法具有明显较好的结果,模型的风速预报准确率得到大幅提高。其中预报风速最大 RMSE 下降到1.2 m/s,最低预报准确率高于 60%。但受到目前收集的数据集较少的限制,目前还没有对 1 年周期的测试数据进行订正效果评估,后续研究中将针对这一问题进行深入研究,且还会考虑随着数据资料的增加,从时间周期维度分别进行机器学习建模,从而进一步研究季节等要素对风速与其他气象参数之间相关性的影响。

参考文献:

[1] 肖寅. 基于 WRF 模式的短期风速预测及订正方法研究[D]. 南京:南京信息工程大学,2016.

[2] N D Bokde, A Feijóo, N Al-Ansari, et al. A Comparison Between Reconstruction Methods for Generationof Synthetic Time Series Applied to Wind Speed Simulation [J]. IEEE Access, 2019, 7: 135386-135398.

[3] R L Kashyap. Optimal choice of AR and MA parts in autoregressive moving average models[J]. IEEE Trans. Pattern Anal. Mach. Intell., 1982, 2(2): 99-104.

[4] G P Zhang. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003, 50: 159-175.

[5] Aaron M R Culver, Adam H Monahan. The Statistical Predictability of Surface Winds over Western

and Central Canada[J]. Journal of Climate,2013, 26(21):8305-8322.
[6] Monahan, A H, The Gaussian Statistical Predictability of Wind Speeds[J]. Climate,2013,26:5563-5577.
[7] Galanis, George, Papageorgiou, et al. A hybrid Bayesian Kalman filter and applications to numerical wind speed modeling[J]. Journal of Wind Engineering & Industrial Aerodynamics, 1997, 56(167):1-22.
[8] 孙全德,焦瑞莉,夏江江,等.基于机器学习的数值天气预报风速订正研究[J].气象,2019,45(3):426-436.
[9] Gang Zhang, Lei Zhang, Tuo Xie. Prediction of short-term wind power in wind power plant based on BP-ANN[C]. 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference(IMCEC), Xi'an, 2016:75-79.
[10] J Jin, B Wang, M Yu, et al. The short-term wind speed prediction based on HF-EMD and BP neural network model[C]. 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), Siem Reap, 2017:489-494.
[11] Gang Zhang, Lei Zhang, Tuo Xie. Prediction of short-term wind power in wind power plant based on BP-ANN[C]. 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Xi'an, 2016:75-79.
[12] A Prasetyowati, D. Sudiana, H Sudibyo. Comparison Accuracy W-NN and WD-SVM Method In Predicted Wind Power Model on Wind Farm Pandansimo[J]. 2018 4th International Conference on Nano Electronics Research and Education(ICNERE), Hamamatsu, Japan, 2018:1-4.
[13] H Xue, L Li, K Chao, et al. Short-Term Wind Power Prediction Based on Improved Chicken Algorithm and Support Vector Machine[J]. 2018 International Symposium on Computer, Consumer and Control(IS3C), Taichung, Taiwan, 2018:137-140.
[14] 张颖超,肖寅,邓华.基于 ELM 的风电场短期风速订正技术研究[J].气象,2016,42(4):466-471.
[15] D Kang, Y Su, X Liu, et al. Short-term wind speed forecasting in wind farm based on C-C and ELM [C]. 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2017:2832-2836.
[16] X Luo. Short-Term Wind Speed Forecasting via Stacked Extreme Learning Machine With Generalized Correntropy[J]. in IEEE Transactions on Industrial Informatics, 2018, 14(11):4963-4971.
[17] 邓华,张颖超,顾荣,等.基于 PCA-RBF 的风电场短期风速订正方法研究[J].气象科技,2018,46(1):10-15.
[18] Y Zhang, B Chen, Y Zhao, et al. Wind Speed Prediction of IPSO-BP Neural Network Based on Lorenz Disturbance[J]. in IEEE Access, 2018, 6:53168-53179.
[19] W Teng, X Wang, Y Meng, et al. An improved support vector clustering approach to dynamic aggregation of large wind farms[J]. in CSEE Journal of Power and Energy Systems, 2019, 5(2):215-223.
[20] Chen T, Tong H. Higgs boson discovery with boosted trees [C]. International Conferenceon High-energy Physics & Machine Learning. 2014.
[21] Yang A, Zhang J, Pan L, et al. Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination[C]. 2015 International Symposium on Security and Privacy in Social Networks and Big Data(SocialSec). IEEE, 2015.

A Research for 10 m Wind Speed Prediction based on XGBoost

MAO Kaiyin, ZHAO Changming, HE Jia
(Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Based on the current meteorological forecasting model, there is a certain error in the accuracy of wind speed forecasting, and domestic and foreign researchers have done a lot of work in revising wind speed forecasting. This paper proposes a CD-XGBoost (Clustering and Double XGBoost) algorithm, which is an improvement to the existing algorithm of machine learning wind speed forecast correction. It mainly includes the following three improvement directions: First, propose the idea of clustering the correlation between weather elements and correction elements and training the intercluster sites based on different machine learning models to improve the accuracy of wind speed correction results. Second, the algorithm highlights the impact of spatial factors on wind speed forecasting, and selects the forecast elements of *K* nearby forecast grid points of the meteorological observation station to build a data set. Compared with the traditional interpolation correction method, more consideration is given to the spatial factors between the stations and grid points. Third, an algorithm is proposed to independently train the XGBoost model using data from two different reporting points, and then add linear weights to the outputs of the two models to obtain the final correction result. In this paper's simulation experiments, three-hour observation data from 2552 meteorological observatories in China and 3 h forecast data from the numerical model of the European Medium-Term Weather Forecasting Center (ECWMF) are used to revise the ECWMF's 10 m surface prediction. The model constructed using the improved algorithm is compared with the model constructed by the current algorithm, and the results show that the accuracy of the wind speed prediction algorithm proposed in this paper is significantly better than that of the current revised algorithm, in which the accuracy rate of 3 h forecast time is over 85%. The accuracy rate within 168 h is higher than 60%, which has a good application prospect.

Keywords: machine learning; wind speed correction; clustering; XGBoost