XGBoost-LSTM 变权组合模型支持下短期 $PM_{2.5}$ 浓度预测——以上海为例

康俊锋[1],谭建林[1],方 雷[2*],肖亚来[1] (1.江西理工大学土木与测绘工程学院,江西 赣州 341000；2.复旦大学环境科学与工程系,上海 200433)

摘要：为进一步提高 $PM_{2.5}$ 浓度预测的精度,基于 XGBoost 和 LSTM 进行改进得到变权组合模型 XGBoost-LSTM(Variable).过对预测因子进行相关性分析,得到其它大气污染物和气象因素对 $PM_{2.5}$ 浓度的影响,确定最优 $PM_{2.5}$ 浓度预测因子,再将预处理后数据集输入 LSTM 模型和 XGBoost 模型分别进行预测,采用基于残差改进的自适应变权组合方法得到最终预测结果.结果表明,污染物变量的相对重要性高于气象因子变量,其中当前 $PM_{2.5}$ 和 CO 浓度的相对重要性较高,而平均风速和相对湿度重要性较低.XGBoost-LSTM(Variable)模型的 RMSE、MAE 和 MAPE 值为 1.75、1.12 和 6.06,优于 LSTM、XGBoost、SVR、XGBoost-LSTM(Equal)和 XGBoost-LSTM(Residual)模型.分季节预测结果表明,XGBoost-LSTM(Variable)模型在春季预测精度最好,而夏季预测精度较差.模型预测精度高的原因在于其不仅考虑了数据的时间序列特征,又兼顾了数据的非线性特征.

关键词：LSTM；XGBoost；$PM_{2.5}$ 预测；变权组合

中图分类号：X831 **文献标识码**：A **文章编号**：1000-6923(2021)09-4016-10

DOI:10.19674/j.cnki.issn1000-6923.20210430.004

Short-term $PM_{2.5}$ concentration prediction based on XGBoost and LSTM variable weight combination model: a case study of Shanghai. KANG Jun-feng[1], TAN Jian-lin[1], FANG Lei[2*], XIAO Ya-lai[1] (1.School of Civil and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China；2.Department of Environmental Science and Engineering, Fudan University, Shanghai 200433,China). *China Environmental Science*, 2021,41(9)：4016~4025

Abstract：In order to further improve the accuracy of $PM_{2.5}$ concentration prediction, a variable weight combination short-term 1-hour $PM_{2.5}$ concentration prediction model based on LSTM network and XGBoost model was proposed. First, analyze the predictive factors, explore the influence of air pollutant factors and meteorological factors on the $PM_{2.5}$ concentration, to determine the best $PM_{2.5}$ concentration predictive factors and analysis the variable importance. Then, after data pretreatment the LSTM prediction model and the XGBoost prediction model was built respectively, and adopt the adaptive variable weight combination method based on residual improvement to obtain the final prediction result. The results show that: The relative importance of pollutant variables is higher than the importance of meteorological factors, among which the relative importance of current $PM_{2.5}$ concentration and CO concentration is higher, while the importance of average wind speed and relative humidity is lower. The values of RMSE, MAE and MAPE of the variable weight combined XGBoost-LSTM (Variable) model proposed in this study are 1.75, 1.12 and 6.06, which are better than LSTM, XGBoost, SVR, XGBoost-LSTM (Equal) and XGBoost-LSTM (Residual) model. The combined model predicts performance best in spring but the forecast accuracy is poor in summer. The variable weight method combination model proposed in this study effectively combines the advantages of the two models, not only considers the time series information of the data but also takes into account the nonlinear relationship between the features, and has higher prediction accuracy compared with other models.

Key words：long short term memeny (LSTM)；XGBoost；$PM_{2.5}$ forecast；variable weight combination model

社会经济的快速发展导致 $PM_{2.5}$ 等空气污染问题日益突出[1-2],对 $PM_{2.5}$ 等空气污染物浓度进行精准预测和提前预警具有重要意义.$PM_{2.5}$ 浓度预测模型主要包括以 CAMQ[3](通用多尺度空气质量模型)模式、WRF-Chem[4] (区域大气动力-耦合模型)模式和 NAQPMS[5](嵌套空气质量预报模式系统)模式等为代表的机理模型,以多元统计理论、灰色预测模型(G,M)[6-7]、多元线性回归模型[8]等为代表的统计预报模型,以及以径向基神经网络(RBF)、反向传播神经网络(BP)、支持向量机(SVM)等神经网络发展到基于深度学习模型的神经网络,如:基于深度信念网络(DBNs)、长短期记忆神经网络(LSTM)等[9-12].

收稿日期：2021-01-26

基金项目：国家重点研发计划项目(2016YFC08033105);国家留学基金资助项目(201808360065);江西省教育厅科学技术研究项目(GJJ150661);国家自然科学基金青年基金资助项目(41701462)

* 责任作者, 博士, fanglei@fudan.edu.cn

随着机器学习技术的发展,有研究采用历史气象数据或历史污染数据,利用支持向量回归模型[13]、随机森林[14-16]、BP 人工神经网络[17]以及 LSTM 网络[18]等单机器学习模型,预测实时 $PM_{2.5}$ 浓度[19]、未来短期[20-21]和长期 $PM_{2.5}$ 浓度[14,22-23]及 $PM_{2.5}$ 浓度的空间变异[17]等.有研究通过构建多个单机器学习模型进行 $PM_{2.5}$ 浓度预测比较,LSTM 网络在处理非线性时序数据方面性能高效并且有更好的泛化能力[24],XGBoost 模型预测精度优于其他单机器学习模型[25].为进一步提高 $PM_{2.5}$ 浓度预测精度,有学者开始尝试组合多个机器学习模型来预测 $PM_{2.5}$ 浓度.宋国君等[26]和李建更等[27]分别建立了基于时间序列分解的 SVR 组合预测模型、Liu 等[28]构建了 DBN、LSTM 网络和多层神经网络(MLP)的三模型组合模型.虽然组合预测模型相较于单机器学习模型可以提升和改善模型预测精度[29],但已有的组合模型研究都只是简单的将一个模型预测结果输入另一模型进行二次预测,或者将多个模型的预测结果进行简单求和.其特点类似一种“机械组合”,两种或多种组合模型之间未发生真正的“化学反应”.

此外,由于 $PM_{2.5}$ 浓度变化既受气象因素影响,也受空气污染物影响[30-31],但已有基于机器学习的 $PM_{2.5}$ 浓度变化预测研究大都只采用气象数据,或者只采用污染物浓度历史数据来进行 $PM_{2.5}$ 浓度预测,预测精度受限.因此,本研究尝试将气象数据、空气污染物数据和 $PM_{2.5}$ 浓度历史数据结合,在分析空气污染物和气象因素对 $PM_{2.5}$ 浓度影响基础上,设计了一种基于残差赋权[32]改进的自适应赋权方法,构建 XGBoost 模型和 LSTM 网络变权组合模型,对未来 1h 短期 $PM_{2.5}$ 浓度进行预测,以期为环境监测部门及社会公众提供预警及精准预测.

1　研究材料与方法

1.1　研究区域与数据

上海市(30°40′~31°53′N,120°52′~122°12′E)位于中国东部沿海的长江三角洲地区,是典型的特大型城市,面积约 6340km^2,地形起伏小,属于亚热带季风气候,其空气质量一直引人关注.本研究选取上海市 10 个环境监测站点(图 1)2017 年 1 月 1 日~10 月 31 日逐小时历史空气质量浓度数据和气象数据(共 7297 组)数据,其统计性描述如表 1 所示.其中,近地面 $PM_{2.5}$ 浓度等空气质量数据来自于生态环境部部空气质量实时发布系统(http://106.37.208.233:20035/),气象数据来自于欧洲中期天气预报中心 3km×3km 再分析数据(https://www.ecmwf.int/).

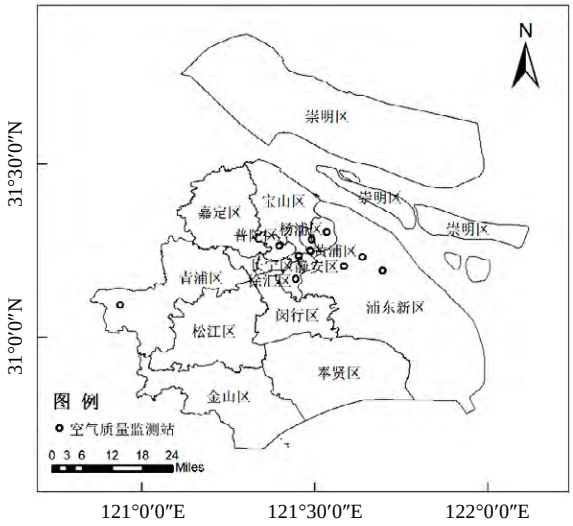

图 1　研究区域

Fig.1　Research area

表 1　上海市空气质量数据和气象数据统计性描述

Table 1　Statistical description of air quality data and meteorological data in Shanghai

指标	平均值	最大值	最小值	标准差
$PM_{2.5}$(μg/m^3)	36.9	169.9	3.67	24.49
SO_2(μg/m^3)	11.2	43.7	4.8	4.47
NO_2(μg/m^3)	39.9	146.1	6.1	20.96
CO(μg/m^3)	0.74	2.22	0.34	0.23
平均风速(m/s)	4.63	11.62	0.16	1.93
降水(mm)	1.78	95.7	0.00	6.38
平均气温(℃)	19.9	34.9	0.6	8.32
平均气压(hPa)	1014.5	1036.3	997.1	8.29
边界层高度(km)	488.5	1908.8	16.8	303.41
相对湿度(%)	79.1	98.4	30.8	10.7

1.2　研究方法

1.2.1　XGBoost

XGBoost(Extreme Gradient Boosting)是一种集成的树模型,是 GBDT(Gradient Boosting Decision Tree)的改进 boosting 算法,具有训练速度快、预测精度高等优点[33].XGBoost 集成了多棵分类回归树(CART)以弥补单棵 CART 无法满足预测精度的不足,预测结果等于所有 CART 的得分总和[34].模型表示为:

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

式中:$\hat{y}_i$表示 i 个样本的预测值;k 表示树的数量;F 是 CART 树的集合空间;x_i 表示 i 个数据点的特征向量.

XGBoost 通过对代价函数进行二阶泰勒展开,使用一阶和二阶导数,在训练集上可以更快收敛,有效提高训练速度,并且将正则化项加到损失函数上,可以降低模型的复杂度和过拟合的风险.

1.2.2 LSTM LSTM(Long Short-Term Memory)是 RNN(Recurrent Neural Network)的改进模式,由 Hpchreiter 等[35]在 1997 年提出,采用 LSTM 层替换了传统的隐藏层,通过引入输入门、输出门、遗忘门三种“门”结构实现信息的有效筛选和长期记忆.LSTM 内部结构如图 2 所示:

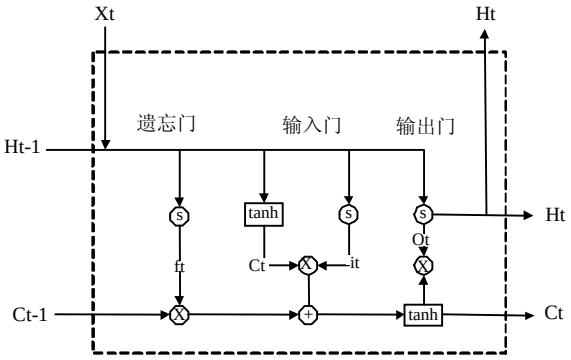

图 2 LSTM 模型结构

Fig.2 LSTM model structure

计算公式如下:

$$f_t=\sigma(w_f\cdot[h_{t-1},x_t]+b_f) \tag{2}$$

$$i_t=\sigma(w_i\cdot[h_{t-1},x_t]+b_i) \tag{3}$$

$$c_t^{'}=\tanh(w_c\cdot[h_{t-1},x_t]+b_c) \tag{4}$$

$$c_t=f_t*C_{t-1}+i_t*c_t^{'} \tag{5}$$

$$o_t=\sigma(w_0\cdot[h_{t-1},x_t]+b_0) \tag{6}$$

$$h_t=o_t*\tanh(c_t) \tag{7}$$

式中:f_t 表示遗忘门限;i_t 表示输入门限;c_t' 表示前一刻细胞状态;c_t 表示当前细胞状态;o_t 表示输出门限;h_t 表示 t 时刻单元输出;x_t 为 t 时刻的输入;σ为 sigmoid 函数;tanh 代表双曲正切函数;w_f、w_i、w_c、w_o 分别代表遗忘门、输入门、细胞状态、输出门的权重矩阵;b_f、b_i、b_c、b_o 分别为遗忘门、输入门、细胞状态和输出门的偏移向量.

1.2.3 变权组合预测模型 本文构建三个组合模型:XGBoost 和 LSTM 等值赋权组合模型 XGBoost-LSTM(Equal)、XGBoost 和 LSTM 残差赋权组合模型 XGBoost-LSTM(Residual) 以及 XGBoost 和 LSTM 变权组合模型 XGBoost-LSTM (Variable).

(1) 单机器学习模型构建 单机器学习模型的优劣决定组合模型的预测精度和性能,设置合理有效的超参数对于提高组合模型的预测性能和收敛速度具有重要意义[36].基于前人研究模型参数设置[37-38]对 LSTM 网络超参数进行设置,最终模型网络层数为 2,学习率设置为 0.001,激活函数设置为 Tanh,优化算法选用 Adam 算法,迭代训练次数设置为 100 次,并设置学习率衰减为 50 次削弱为 10%.

利用 Scikit-learn 提供的网格搜索(GridSearch)方法[39]对 XGBoost 模型的超参数寻优,模型参数最终设置为 :max_depth=4,learning_rate=0.1,n_estimators=200,subsample=0.7,colsample_bytree=0.85, silent=True,ganma=0.2.

(2)组合模型赋权 组合模型精度与单机器学习模型的赋权有直接关系,赋权方法常见的有固定赋权与自适应赋权,其中固定赋权以等值赋权和残差赋权法最为常见[40].

等值赋权将单模型赋予相同的权重,而残差赋权组合模型表达为:

$$f(x_t)=\frac{1}{n}\sum_{i=1}^{n}\omega_i(t-1)f_i(x_t) \tag{8}$$

$$\omega_i(t)=\frac{\frac{1}{\overline{\varepsilon_i}(t-1)}}{\sum_{i=1}^{n}\frac{1}{\overline{\varepsilon_i}(t-1)}} \tag{9}$$

$$s.t.\sum_{i=1}^{n}\omega_i(t-1)f_i(x_t),\omega_i(t-1)\geq 0 \tag{10}$$

式中:$\omega_i(t-1)$为 $t-1$ 时刻第 i 个模型的权重;$\overline{\varepsilon}_i(t-1)$ 为 $t-1$ 时刻第 i 个模型的预测误差平方和.

(3)改进的变权组合模型赋权方法 本文使用基于残差赋权改进的自适应赋权方法的变权[41]方法构建了 XGBoost-LSTM(Variable)模型.对于单机器学习模型在基于式(9)得到所有时刻残差赋权的权重基础上改进,计算最优 m 值,使用前 m 时刻权重平均值对本时刻模型进行初始赋权,即:

$$\omega_j(t)=\frac{1}{m}\sum_{k=1}^{m}\omega_j(t-k)(m=3) \tag{11}$$

对于 t 时刻,假设基于式(9)和式(11)得到各单机器学

习模型权重后,计算该时刻组合模型的预测值与真实值的误差绝对值分别为 $e_{i,t}$、$e_{j,t}$,则有:

$$e_{i,t}=\sum_{i=1}^{n}\omega_i(t)f_i(x_t)-f(t) \tag{12}$$

$$e_{j,t}=\sum_{j=1}^{n}\omega_j(t)f_j(x_t)-f(t) \tag{13}$$

比较 $e_{i,t}$ 和 $e_{j,t}$ 值的大小,如果 $e_{i,t}<e_{j,t}$ 则该组合模型用新的权重 $\omega_j(t)$ 代替原来的权重 $\omega_i(t)$,否则模型权重保持不变.

1.2.4　组合预测模型构建流程　组合预测模型构建流程如图 3 所示,包括数据预处理、单机器学习模型和变权组合预测模型构建以及模型评价分析.

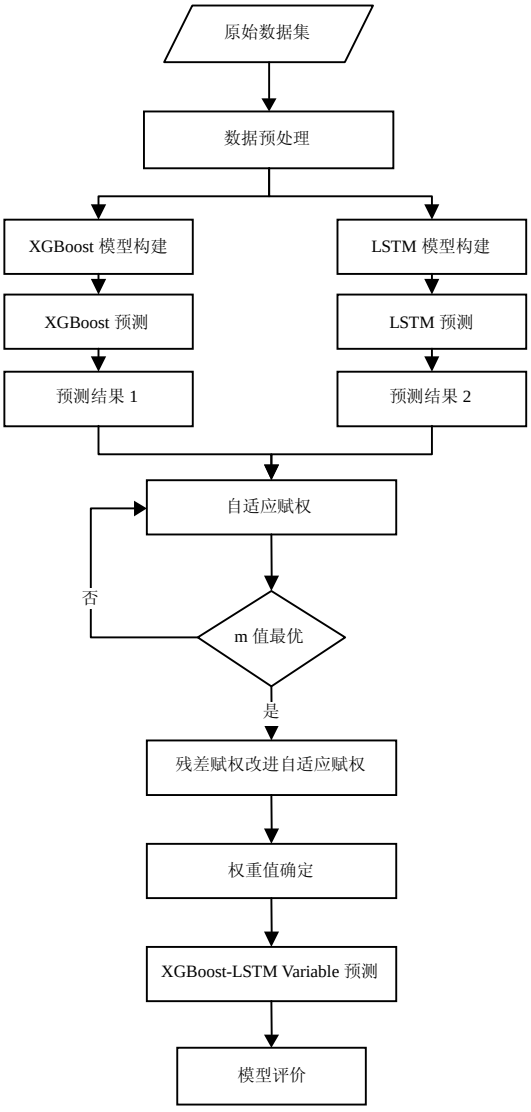

图 3　XGBoost-LSTM(Variable)模型预测流程

Fig.3　XGBoost-LSTM(Variable) model prediction process

(1)数据预处理　得到的原始数据集进行预处理,主要包括数据清洗、缺失值填充和归一化处理,本研究缺失值采用缺失前后数据均值补充.

(2)单机器学习模型构建　数据集按照训练集:测试集=9:1 比例划分后,在训练集上分别训练 LSTM 网络和 XGBoost 模型,确定模型最优超参数,保存训练模型.将测试集分别输入模型,得到各单机器学习模型预测结果.

(3)变权组合预测模型构建　采用前文所示赋权方法确定各单机器学习模型的权重,计算得到组合模型最终预测结果.

(4)模型评价分析　根据模型评价指标比较模型预测能力,分析模型预测效果.

1.2.5　评价指标　本研究采用常见的评价指标均方根误差(RMSE),平均绝对误差(MAE),平均绝对百分比误差(MAPE)以及相关系数(R^2)进行模型精度比较,指标的计算公式如下所示:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}(yo_t-ym_t)^2} \tag{14}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{t=1}^{n}|(yo_t-ym_t)| \tag{15}$$

$$\mathrm{MAPE}=\sum_{t=1}^{n}\left|\left(\frac{yo_t-ym_t}{yo_t}\right)\right|\cdot\frac{100}{n} \tag{16}$$

$$R^2=1-\frac{\sum(ym-y_0)^2}{\sum(y_0-\overline{y_0})^2} \tag{17}$$

式中:n 为样本数据的数量;ym 为预测结果;yo 为真实值; ym_t 和 $\overline{y_0}$ 分别表示预测结果和真实结果的平均值.误差越小,预测方法效果更好,模型预测精度更高.

2　结果与讨论

2.1　预测结果的影响因子分析

2.1.1　空气污染物因子对 $PM_{2.5}$ 浓度的影响　$PM_{2.5}$ 与其它空气污染物之间存在着物理化学层面的相互转化或者在传输过程之间产生相互影响[14],因此,对研究区 $PM_{2.5}$ 浓度与其它污染物变量之间进行了相关性分析.

如图 4 所示,分析 $PM_{2.5}$ 与其它大气污染物(CO,NO_2,O_3,PM_{10},SO_2)之间的相关性,可以发现,$PM_{2.5}$ 与各污染物之间均存在一定的相互关系.其中 PM_{10}、CO 与 $PM_{2.5}$ 之间的相互关系极强,而 O_3 与 $PM_{2.5}$ 之间的相关性最低,所以可以忽略 O_3 对于

$PM_{2.5}$ 的影响,这与前人[14]的研究结果相同.

综上分析,将 SO_2、NO_2 和 CO 3 个污染物变量作为预测模型的输入,其中与 $PM_{2.5}$ 有极强相关性的 PM_{10} 未加入到输入变量集中,因为经过实验分析 $PM_{2.5}$ 与 PM_{10} 相关性过高导致产生冗余,从而导致精度降低.

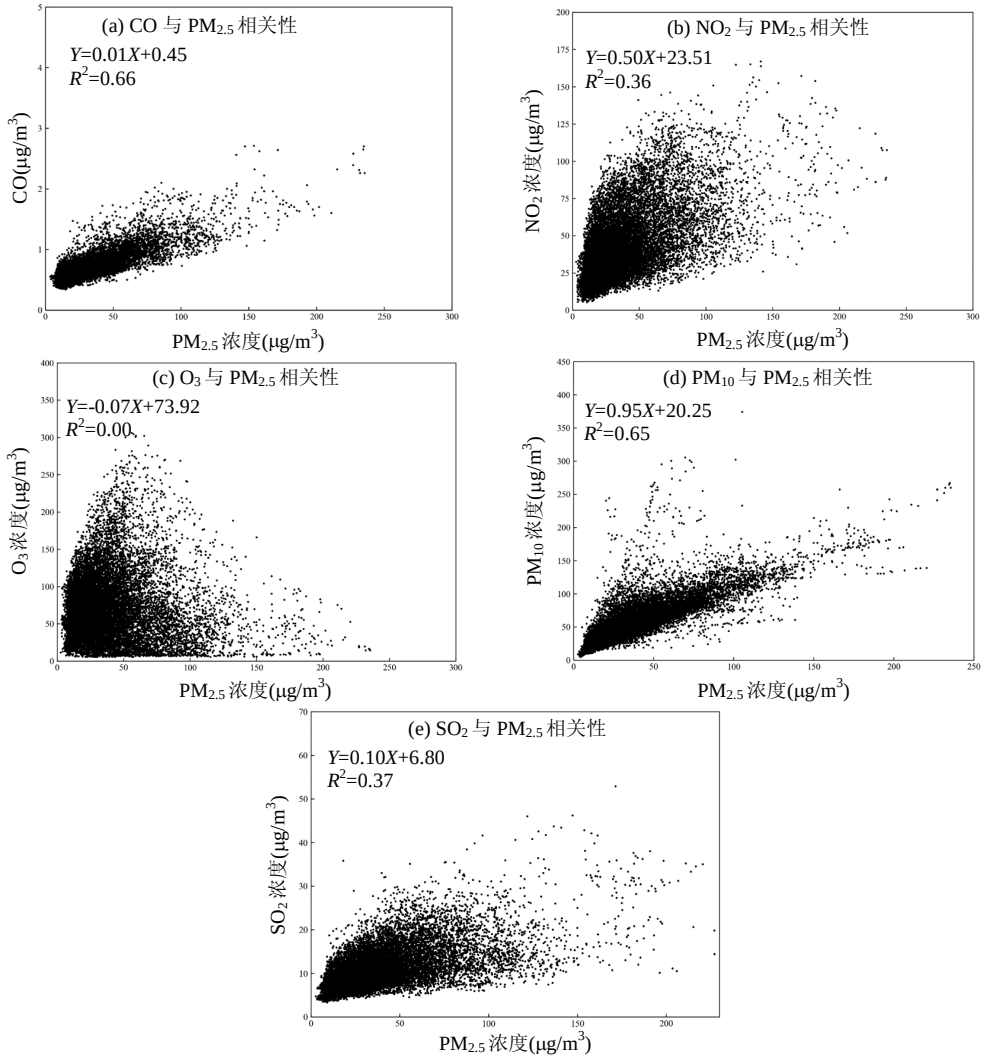

图 4 $PM_{2.5}$ 与其他大气污染物的相关性分析

Fig.4 Correlation analysis of $PM_{2.5}$ and other air pollutants

表 2 $PM_{2.5}$ 浓度与气象因素相关系数

Table 2 Correlation coefficient of $PM_{2.5}$ concentration and meteorological factors

项目	气温	气压	风速	风向	边界层高度	相对湿度	降水量
相关系数	−0.24	0.18	−0.14	0.20	−0.13	−0.12	−0.1

2.1.2 气象因素对 $PM_{2.5}$ 浓度的影响 气象因子也是影响 $PM_{2.5}$ 浓度的一个重要因子,已有大量学者证明 $PM_{2.5}$ 浓度与风速、风向、湿度、气压、气温等因素之间具有密切关系[8,42-43].对研究区 $PM_{2.5}$ 与气象因素进行皮尔逊相关性分析,结果如表 2 所示.$PM_{2.5}$ 与气象因子存在一定的相关性,其中 $PM_{2.5}$ 与气压和风向呈正相关关系,与气温、风速、边界层高度、相对湿度和降水量呈负相关关系.

在本研究中,气象因子作为辅助变量进行 $PM_{2.5}$ 浓度预测,因此气象因子所有变量均加入本文实验中.

2.2 变量重要性分析

利用训练好的 XGBoost 模型对输入变量的重要性进行评价,如图 5 所示,对于未来 1h $PM_{2.5}$ 浓度预测,变量重要性结果为污染物变量大于气象变量重要性,其顺序为当前 $PM_{2.5}$ 浓度、CO 浓度、SO_2 浓度、NO_2 浓度、降水量、边界层高度、风向角度、

平均气温、平均气压、平均风速、相对湿度.污染物变量中当前 $PM_{2.5}$ 浓度值和和 CO 浓度值重要性相对较高,而 SO_2、NO_2 浓度值重要性相对较低.气象因子变量中,降水量和边界层高度较为重要,平均风速和相对湿度的重要性相对较低.

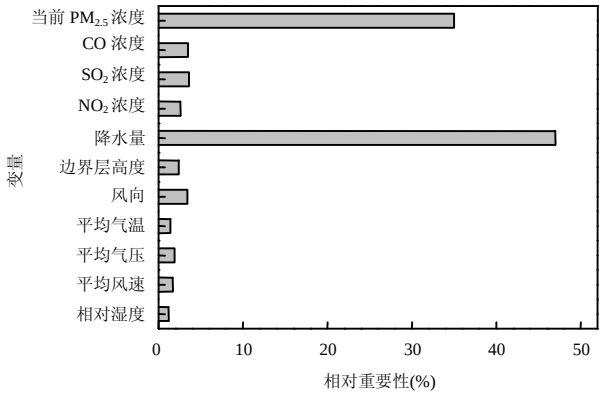

图 5　变量重要性分析

Fig.5　Variable importance

2.3　短时预测分析与对比

为了验证改进的组合模型 XGBoost−LSTM (Variable) 精度,选择 XGBoost、SVR、LSTM、XGBoost−LSTM(Equal)、XGBoost−LSTM(Residual)模型进行对比实验.不同模型预测值与实际值的对比如图 6~7 所示.

由图 6 可知,$PM_{2.5}$ 浓度值实际值处于 15~80ug/m^3 时,各模型预测值和实际值的拟合度均较高,而对于实际值小于 15ug/m^3 和大于 80ug/m^3 的拟合效果均较差.单机器学习模型的拟合效果劣于组合模型的拟合效果,组合模型中,改进的变权组合模型与实际值的拟合效果最好,起伏程度更加接近 $PM_{2.5}$ 浓度变化的实际趋势,偏差较小.

由图 7 可知,组合模型的预测精度优于单机器学习模型和传统赋权方法组合模型预测精度.其中,改进的组合模型 XGBoost−LSTM(Variable)的 MAE、MAPE 和 RMSE 值相较于 XGBoost−LSTM(Equal)模型分别提升了 27.3%、22.9%、32.7%,相较于 XGBoost−LSTM(Residual)分别提升了 20.6%、19.7%、15.1%,表明改进的变权组合方法具有更高的预测精度.

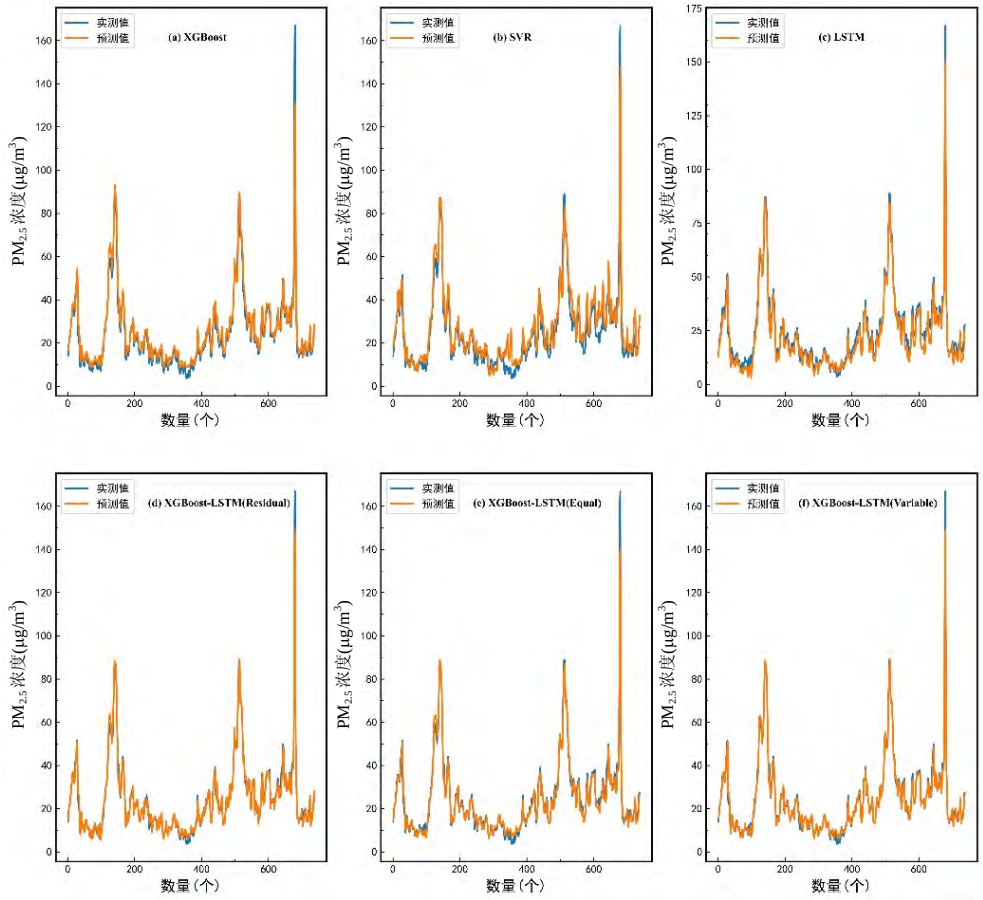

图 6　各模型预测和实测结果对比

Fig.6　Model forecast result and measurement result

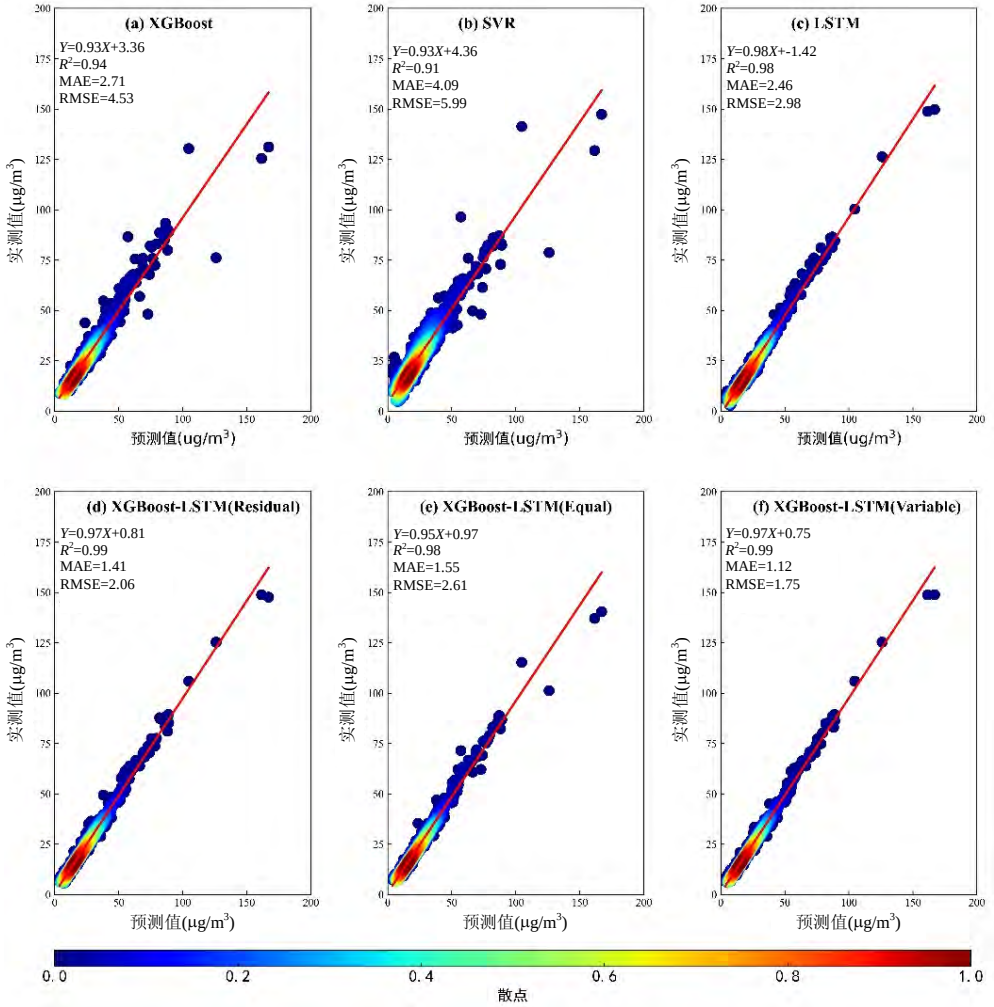

图 7 不同模型实测值与预测值

Fig.7 Scatter plot of prediction results and observation values under different models

2.4 不同季节预测分析

研究区属于典型亚热带季风性气候,季节气候存在明显差异,并且不同季节具有不同的污染物来源.因此针对不同季节选取典型月份进行预测分析,月份选取分别为春季(4 月)、夏季(6 月)、秋季(10 月)和冬季(1 月).

由图 8 可知,本研究改进的变权组合模型在春季和秋季的预测结果较好,其中春季即 4 月份为代表的预测精度最高,RMSE、MAE 和 MAPE 各指标值分别为 1.65、1.23 和 2.81,远小于其它季节的指标值;而在夏季和冬季的预测结果较差,其中夏季的预测结果最差,指标值分别为 7.56、6.04 和 15.19.对于模型不同季节典型月份预测结果分析来看,造成夏季预测结果较差原因是由于夏季强烈的大气层活动,降雨频率高以及风速快,形成了较好的大气颗粒物扩散和清除的气象条件[44].而在冬季预测结果较好是由于 $PM_{2.5}$ 浓度与影响因子的相关性更好[25,45-46].

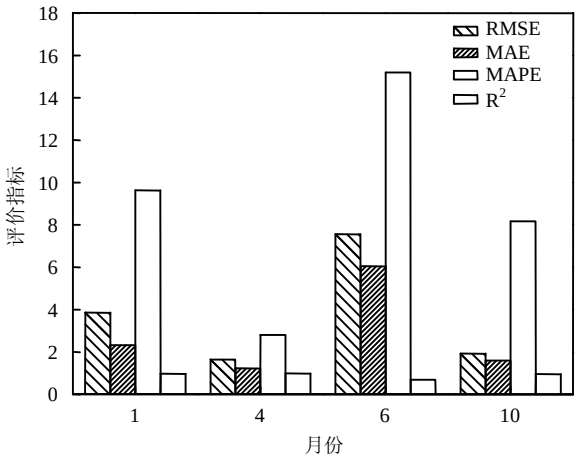

图 8 组合模型四季典型月份预测结果

Fig.8 Combination model forecast results in different seasons

2.5 讨论

使用气象数据、空气污染物数据以及 $PM_{2.5}$ 浓

度历史数据构建了变权组合模型.以上海为研究区域,进行未来 1h 短期 $PM_{2.5}$ 浓度预测.采用改进的 XGBoost-LSTM(Variable)变权组合模型,RMSE、MAE 和 MAPE 值为 1.75、1.12 和 6.06,远小于瓮克瑞[47]提出的组合模型预测值 8.901、6.774 和 8.862,以及 Liu Hui[28]提出的集成模型 4.51、2.78 和 7.79,是由于将时间序列预测模型中性能最好的 LSTM 网络和非线性模型中表现较好的 XGBoost 模型以变权组合的形式进行组合预测,该模型不仅考虑了数据的时间序列特征,又兼顾了数据的非线性特征;对于短时预测分析对比结果,本研究改进的 XGBoost-LSTM(Variable)变权组合模型优于 XGBoost-LSTM(Equal)组合模型、XGBoost-LSTM(Residual)组合模型,是由于本方法考虑到 XGBoost 模型和 LSTM 网络在不同时刻预测误差不同,通过对不同时刻采取不同的权重值,充分融合 XGBoost 模型和 LSTM 网络的优势.

研究区域选择中,不同地区的污染物组成以及气象条件具有强烈的地方性特点,因此,本研究只选取上海市作为研究区域探讨模型的表现.另外,在 $PM_{2.5}$ 预测影响因素选择上,本研究目前只将气象和空气质量污染物要素作为预测因子,未来应该考虑土地利用变化因素、经济、交通、环保政策等更多合适的因素进行预测研究,以进一步提高 $PM_{2.5}$ 预测精度.

3　结论

3.1　模型变量重要性分析可知,污染物变量的相对重要性高于气象因子变量重要性,其中当前 $PM_{2.5}$ 和 CO 浓度相对重要性高,而平均风速和相对湿度重要性较低.

3.2　由于组合模型不仅考虑了数据的时间序列特征,又兼顾了数据的非线性特征,因此,与单机器学习模型和其它组合模型结果相比,改进的变权组合模型的预测结果与真实值更加接近,误差更小,稳定性也更强,可以用于 $PM_{2.5}$ 浓度短期预警预报.

3.3　由于季节特征等差异,改进的组合模型在季节上的表现有所差异,表现为在春、秋季节预测效果较好,而在夏、冬季节预测结果较差.

参考文献：

[1] Kim Y, Manley J, Radoias V. Medium- and long-term consequences of pollution on labor supply: evidence from Indonesia [J]. IZA Journal of Labor Economics, 2017,6(1):1-15.

[2] 王庚辰,王普才.中国 $PM_{2.5}$ 污染现状及其对人体健康的危害 [J]. 科技导报, 2014,32(26):72-78.
Wang G C, Wang P C. $PM_{2.5}$ pollution in China and its harmfulness to human health [J]. Science & Technology Review, 2014,32(26):72-78.

[3] Dennis R L, Byun D W, Novak J H. The next generation of integrated air quality modeling: EPA's models-3 [J]. Atmospheric Environment, 1996,30(12):1925-1938.

[4] 周广强,谢　英,吴剑斌,等.基于 WRF-Chem 模式的华东区域 $PM_{2.5}$ 预报及偏差原因 [J]. 中国环境科学, 2016,36(8):2251-2259.
Zhou G Q, Xie Y, Wu J B, et al. WRF-Chem based $PM_{2.5}$ forecast and bias analysis over the East China Region [J]. China Environmental Science, 2016,36(8):2251-2259.

[5] Qingxin W, Qiaolin Z, Jinhua T, et al. Estimating $PM_{2.5}$ concentrations based on MODIS AOD and NAQPMS data over Beijing-Tianjin-Hebei. [J]. Sensors (Basel, Switzerland), 2019,19(5):1207.

[6] Zhang Z, Wu L, Chen Y. Forecasting $PM_{2.5}$ and PM_{10} concentrations using GMCN(1,N) model with the similar meteorological condition: Case of Shijiazhuang in China [J]. Ecological Indicators, 2020,119: 106871.

[7] Pai T, Ho C, Chen S, et al. Using seven types of GM (1, 1) model to forecast hourly particulate matter concentration in Banciao City of Taiwan [J]. Water, Air, & Soil Pollution, 2011,217(1):25-33.

[8] 方晓婷,段华波,胡明伟,等.气象因素对大气污染物影响的季节差异分析及预测模型对比——以深圳为例 [J]. 环境污染与防治, 2019, 41(5):541-546.
Fang X T, Duan H B, Hu W M, et al. The seasonal differential effects of meteorological parameters on atmospheric pollutants and the prediction model comparison: a case study of Shenzhen [J]. Environmental Pollution & Control, 2019,41(5):541-546.

[9] Liao Q, Zhu M, Wu L, et al. Deep learning for air quality forecasts: a review [J]. Current Pollution Reports, 2020:1-11.

[10] 戴李杰,张长江,马雷鸣.基于机器学习的 $PM_{2.5}$ 短期浓度动态预报模型 [J]. 计算机应用, 2017,37(11):3057-3063.
Dai L J, Zhang C J, Ma L M, et al. Dynamic forecasting model of short-term $PM_{2.5}$ concentration based on machine learning [J]. Journal of Computer Applications, 2017,37(11):3057-3063.

[11] 郑　毅,朱成璋.基于深度信念网络的 $PM_{2.5}$ 预测 [J]. 山东大学学报(工学版), 2014,44(6):19-25.
Zheng Y, Zhu C Z. A prediction method of atmospheric $PM_{2.5}$ based on DBNs [J]. Journal of Shandong University(Engineering Science), 2014,44(6):19-25.

[12] 朱晏民,徐爱兰,孙　强.基于深度学习的空气质量预报方法新进展 [J]. 中国环境监测, 2020,36(3):10-18.
Zhu Y M, Xu A L, Sun Q. New progress for air quality forecasting methods based on deep learning [J]. Environmental Monitoring in China, 2020,36(3):10-18.

[13] 谢永华,张鸣敏,杨　乐,等.基于支持向量机回归的城市 $PM_{2.5}$ 浓度预测 [J]. 计算机工程与设计, 2015,36(11):3106-3111.
Xie Y H, Zhang M M, Yang L, et al. Predicting urban $PM_{2.5}$ concentration in China using support vector regression [J]. Computer

Engineering and Design, 2015,36(11):3106-3111.

[14] 侯俊雄,李 琦,朱亚杰,等.基于随机森林的 $PM_{2.5}$ 实时预报系统 [J]. 测绘科学, 2017,42(1):1-6.
Hou J X, Li Q, Zhu Y J, et al. Real-time forecasting system of $PM_{2.5}$ concentration based on spark framework and random forest model [J]. Science of Surveying and Mapping, 2017,42(1):1-6.

[15] 任才溶,谢 刚.基于随机森林和气象参数的 $PM_{2.5}$ 浓度等级预测 [J]. 计算机工程与应用, 2019,55(2):213-220.
Ren C R, Xie G. Prediction of $PM_{2.5}$ concentration level based on random forest and meteorological parameters [J]. Computer Engineering and Applications, 2019,55(2):213-220.

[16] 夏晓圣,陈菁菁,王佳佳,等.基于随机森林模型的中国 $PM_{2.5}$ 浓度影响因素分析 [J]. 环境科学, 2020,41(5):2057-2065.
Xia X S, Chen J J, Wang J J, et al. $PM_{2.5}$ concentration influencing factors in China based on the random forest model [J]. Environmental Science, 2020,41(5):2057-2065.

[17] 王 敏,邹 滨,郭 宇,等.基于 BP 人工神经网络的城市 $PM_{2.5}$ 浓度空间预测 [J]. 环境污染与防治, 2013,35(9):63-66.
Wang M, Zou B, Guo Y, et al. BP artificial neural network-based analysis of spatial variability of urban $PM_{2.5}$ concentration [J]. Environmental Pollution & Control, 2013,35(9):63-66.

[18] 白盛楠,申晓留.基于 LSTM 循环神经网络的 $PM_{2.5}$ 预测 [J]. 计算机应用与软件, 2019,36(1):67-70.
Bai S N, Shen X L. $PM_{2.5}$ Prediction based on LSTM recurrent neural network [J]. Computer Applications and Software, 2019,36 (1):67-70.

[19] Zhang Y, Bocquet M, Mallet V, et al. Real-time air quality forecasting, part I: History, techniques, and current status [J]. Atmospheric Environment, 2012,60(1):632-655.

[20] 段大高,赵振东,梁少虎,等.基于 LSTM 的 $PM_{2.5}$ 浓度预测模型 [J]. 计算机测量与控制, 2019,27(3):215-219.
Duan D G, Zhao Z D, Liang S H, et al. Research on $PM_{2.5}$ concentration prediction based on LSTM [J]. Computer Measurement & Control, 2019,27(3):215-219.

[21] Liu D, Sun K. Short-term $PM_{2.5}$ forecasting based on CEEMD-RF in five cities of China [J]. Environmental Science and Pollution Research, 2019,26(32):32790-32803.

[22] Huang K, Xiao Q, Meng X, et al. Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain [J]. Environmental Pollution, 2018,242.

[23] Mao X, Shen T, Feng X. Prediction of hourly ground-level PM $PM_{2.5}$ concentrations 3days in advance using neural networks with satellite data in eastern China [J]. Atmospheric Pollution Research, 2017,8(6):1005-1015.

[24] 赵文芳,林润生,唐 伟,等.基于深度学习的 $PM_{2.5}$ 短期预测模型 [J]. 南京师大学报(自然科学版), 2019,42(3):32-41.
Zhao W F, Lin R S, Tang W, et al. Forecasting model of short-term concentration based on deep learning [J]. Journal of Nanjing Normal University (Natural Science Edition), 2019,42(3):32-41.

[25] 康俊锋,黄烈星,张春艳,等.多机器学习模型下逐小时 $PM_{2.5}$ 预测及对比分析 [J]. 中国环境科学, 2020,40(5):1895-1905.
Kang J F, Huang L X, Zhang C Y, et al. Hourly $PM_{2.5}$ prediction and its comparative analysis under multi-machine learning model [J]. China Environmental Science, 2020,40(5):1895-1905.

[26] 宋国君,国潇丹,杨 啸,等.沈阳市 $PM_{2.5}$ 浓度 ARIMA-SVM 组合预测研究 [J]. 中国环境科学, 2018,38(11):4031-4039.
Song G J, Guo X D, Yang X, et al. ARIMA-SVM combination prediction of $PM_{2.5}$ concentration in Shenyang [J]. China Environmental Science, 2018,38(11):4031-4039.

[27] 李建更,罗奥荣,李晓理.基于互补集合经验模态分解与支持向量回归的 $PM_{2.5}$ 质量浓度预测 [J]. 北京工业大学学报, 2018,44(12):1494-1502.
Li J G, Luo A R, Li X l. Prediction of $PM_{2.5}$ mass concentration based on complementary ensemble empirical mode decomposition and support vector Regression [J]. Journal of Beijing University of Technology, 2018,44(12):1494-1502.

[28] Liu H, Dong S. A novel hybrid ensemble model for hourly $PM_{2.5}$ forecasting using multiple neural networks: a case study in China [J]. Air Quality, Atmosphere & amp; Health, 2020:1-10.

[29] 王学梅,王凤文,陈 滔,等.基于组合模型的 $PM_{2.5}$ 浓度预测及其不确定性分析 [J]. 环境工程, 2020,38(8):229-235.
Wang X M, Wang F W, Chen T, et al. $PM_{2.5}$ concentration prediction and uncertainly analysis based on a composite model [J]. Environmental Engineering, 2020,38(8):229-235.

[30] Wang J, Shao W, Kim J. Multifractal detrended cross-correlation analysis between respiratory diseases and haze in South Korea [J]. Chaos, Solitons and Fractals: the Interdisciplinary Journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena, 2020,135:10.1016/j.Chaos.2020.109781.

[31] Chen J, Lu J, Avise J C, et al. Seasonal modeling of $PM_{2.5}$ in California's San Joaquin Valley [J]. Atmospheric Environment, 2014,92:182-190.

[32] 王新民,崔 巍.变权组合预测模型在地下水水位预测中的应用 [J]. 吉林大学学报(地球科学版), 2009,39(6):1101-1105.
Wang X M, Cui W. Application of changeable weight combination forecasting model To groundwater level prediction [J]. Journal of Jilin University (Earth Science Edition), 2009,39(6):1101-1105.

[33] Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization [J]. Machine Learning, 2000,40(2):139-157.

[34] Wu Y, Qi S, Hu F, et al. Recognizing activities of the elderly using wearable sensors: a comparison of ensemble algorithms based on boosting [J]. Sensor Review, 2019,39(6):743-751.

[35] Hochreiter S, Schmidhuber J. Long Short-Term Memory [J]. Neural Computation, 1997,9(8):1735-80.

[36] 郭立力,赵春江.十折交叉检验的支持向量机参数优化算法 [J]. 计算机工程与应用, 2009,45(8):55-57.
Guo L L, Zhao C J. Optimizing parameters of support vector machine's model based on genetic algorithm [J]. Computer Engineering and Applications, 2009,45(8):55-57.

[37] Zhai W, Cheng C. A long short-term memory approach to predicting air quality based on social media data [J]. Atmospheric Environment, 2020,237.

[38] Chang Y, Chiao H, Abimannan S, et al. An LSTM-based aggregated model for air pollution forecasting [J]. Atmospheric Pollution Research, 2020,11(8):1451-1463.

[39] Gang L, Jingying F, Dong J, et al. Spatial variation of the relationship

between $PM_{2.5}$ concentrations and meteorological parameters in China [J]. BioMed Research International, 2015,2015,684618.

[40] 刘　明,王红蕾,索良泽.基于变权组合模型的中长期负荷概率密度预测 [J]. 电力系统及其自动化学报, 2019,31(7):88-94.
Liu M, Wang H L, Suo L Z. Medium-and long-term load probability density forecasting based on variable weight combination model [J]. Proceedings of the CSU-EPSA, 2019,31(7): 88-94.

[41] 王新民,崔　巍.变权组合预测模型在地下水水位预测中的应用 [J]. 吉林大学学报(地球科学版), 2009,39(6):1101-1105.
Wang X M, Cui W. Application of changeable weight combination forecasting model to groundwater level prediction [J]. Journal of Jilin University (Earth Science Edition), 2009,39(6):1101-1105.

[42] 曲　悦,钱　旭,宋洪庆,等.基于机器学习的北京市 $PM_{2.5}$ 浓度预测模型及模拟分析 [J]. 工程科学学报, 2019,41(3):401-407.
Qu Y, Qian X, Song H Q, et al. Machine-learning-based model and simulation analysis of $PM_{2.5}$ concentration prediction in Beijing [J]. Chinese Journal of Engineering, 2019,41(3):401-407.

[43] 谢　超,马民涛,于肖肖.多种神经网络在华北西部区域城市空气质量预测中的应用 [J]. 环境工程学报, 2015,9(12):6005-6009.
Xie C, Ma M T, Yu X X. Forecasting model of air pollution index based on multi-artificial neural network in western region of Northern China [J]. Chinese Journal of Environmental Engineering, 2015,9(12): 6005-6009.

[44] 刘小真,任羽峰,刘忠马,等.南昌市大气颗粒物污染特征及 $PM_{2.5}$ 来源解析 [J]. 环境科学研究, 2019,32(9):1546-1555.
Liu X Z, Ren Y F, Liu Z M, et al. Pollution characteristics of atmospheric and source apportionment of $PM_{2.5}$ in Nanchang City [J]. Research of Environmental Sciences, 2019,32(9):1546-1555.

[45] 张淑平,韩立建,周伟奇,等.冬季 $PM_{2.5}$ 的气象影响因素解析 [J]. 生态学报, 2016,36(24):7897-7907.
Zhan S P, Han L J, Zhou W Q, et al. Relationships between fine particulate matter($PM_{2.5}$) and meteorological factors in winter at typical Chinese cities [J]. Acta Ecological Sinical, 2016,36(24):7897-7907.

[46] 朱媛媛,高愈霄,刘　冰,等.京津冀秋冬季 $PM_{2.5}$ 污染概况和预报结果评估 [J]. 环境科学, 2019,40(12):5191-5201.
Zhu Y Y, Gao Y X, Liu B, et al. Concentration characteristics and assessment of model-predicted results of $PM_{2.5}$ in the Beijing-Tianjin-Hebei Region in autumn and winter [J]. Environmental Science, 2019,40(12):5191-5201.

[47] 翁克瑞,刘　淼,刘　钱. TPE-XGBOOST 与 LassoLars 组合下 $PM_{2.5}$ 浓度分解集成预测模型研究 [J]. 系统工程理论与实践, 2020, 40(3):748-760.
Weng K R, Liu M, Liu Q. An integrated prediction model of $PM_{2.5}$ concentration based on TPE-XGBOOST and LassoLars [J]. Systems Engineering-Theory & Practice, 2020,40(3):748-760.

作者简介：康俊锋(1978-),男,江西赣州人,副教授,博士,主要从事高性能 GIS 算法及其在环境与土地中的应用研究.发表论文 10 余篇.