

DOI:10.3880/j.issn.1004-6933.2022.06.015

基于 ARIMA 和 Prophet 的水质预测集成学习模型

嵇晓燕¹, 杨凯¹, 陈亚男¹, 姚志鹏¹, 王正², 安新国²

(1. 中国环境监测总站, 北京 100012; 2. 北京金水永利科技有限公司, 北京 100012)

摘要:将时间序列模型 ARIMA 和 Prophet 作为基学习器, 结合 BP 神经网络模型构建了水质预测集成学习模型。选取长江流域某断面 2019—2020 年的 DO、COD_{Mn}、NH₃-N、TP 和 TN 等 5 个水质指标的监测数据对该模型的有效性进行了检验, 结果表明: 5 个水质指标集成学习模型预测结果的平均绝对百分比误差比时间序列模型的预测误差分别低 35.0%、29.9%、4.1%、40.6% 和 17.1%, 模型预测值和监测值的皮尔逊相关系数大于 0.8。集成学习模型预测精度高于单一模型, 可以更精确地进行水质预测。

关键词:水质预测; ARIMA 模型; Prophet 模型; 集成学习

中图分类号: X832 **文献标志码:** A **文章编号:** 1004-6933(2022)06-0111-05

An ensemble learning model for water quality forecast based on ARIMA and Prophet // JI Xiaoyan¹, YANG Kai¹, CHEN Ya'nan¹, YAO Zhipeng¹, WANG Zheng², AN Xinguo² (1. China National Environmental Monitoring Center, Beijing 100012, China; 2. Golden Water Technology (Beijing) Ltd., Beijing 100012, China)

Abstract: An ensemble learning model for water quality forecast was constructed by using time series models including ARIMA and Prophet model as base learners, combined with BP neural network model. The monitoring data of five water quality indicators, including DO, COD_{Mn}, NH₃-N, TP and TN, for a section of the Yangtze River Basin from 2019 to 2020 were selected to test the validity of the model. The results show that the mean absolute percentage errors of the predicted results of the ensemble learning model for the five indicators were lower than those of the time series model by 35.0%, 29.9%, 4.1%, 40.6% and 17.1%, respectively. The prediction accuracy of the ensemble learning model is higher than that of the single model, which can be more accurate for water quality forecast.

Key words: water quality forecast; ARIMA model; Prophet model; ensemble learning

地表水质自动化监测预警是当前水环境管理的切实需要, 是现代管理“安全第一、预防为主”理念的具体体现^[1]。水质预测作为地表水质自动化监测预警的基础, 已经成为水环境污染防治工作的重要研究内容。水质预测就是针对目标水体在建立基于水动力的机理模型或基于历史数据分析的数理模型的基础上, 实现对该水体未来一段时间水质情况的预判。根据模型的预测结果, 可以及时发现潜在的污染风险, 避免重大水污染事件的发生, 为水污染防治工作的决策提供数据支持, 因此水质预测具有非常重要的实际意义。

国内外学者对水质预测方法进行了大量的研究, 包括基于水动力学模型的水质预测方法^[2-3]和

基于数理模型的预测方法。基于数理模型的常见预测方法包括时间序列法^[4-5]、回归分析法^[6-7]、灰色预测法^[8]以及神经网络预测法^[9-11]等。水质数据作为时间序列数据, 具有很强的自相关性以及趋势性, 适用于建立时间序列模型进行预测, 但常用的单一时间序列分析模型难以捕捉到水质数据序列中的非线性部分。而对于神经网络这类模型而言, 因为其强大的非线性拟合能力, 导致模型预测的泛化能力较差。考虑到影响水质环境的因素较多, 且水质监测数据具有非线性、随机性以及周期性等复杂特性, 建立单一模型进行预测会因为模型自身的限制使得模型预测精度及健壮性普遍不高^[12]。

本文采用基于时间序列分析模型的集成学习方

基金项目: 长江生态环境保护修复联合研究项目(2019-LHYJ-01-0301); 国家水环境监测监控及业务化平台技术研究课题(2017ZX07302002)

作者简介: 嵇晓燕(1981—), 女, 正高级工程师, 博士, 主要从事水环境质量和评价研究。E-mail: jixy@cnemc.cn

通信作者: 安新国(1988—), 男, 工程师, 硕士, 主要从事机器学习在水环境预测预警中的应用研究。E-mail: anxinguo@gwatertech.com

法进行水质预测,通过融合多个时间序列模型的预测结果,利用神经网络模型对预测结果进一步进行非线性处理,从而能够显著提升预测精度。

1 模型构建

集成学习是使用多个基学习器进行学习,并根据规则整合每个学习器输出的结果,从而获得比单一基学习器更好效果的一种机器学习方法^[13]。本文利用 Blending 集成学习的方式,将时间序列模型 ARIMA 和 Prophet 作为基学习器,将 BP 神经网络模型作为整合学习器结果的集成模型,从而构建本文的集成学习模型。集成学习模型中,ARIMA 模型可以很好地获取时间序列中的周期性和趋势信息,Prophet 模型可以很好地处理时间序列数据中存在的异常值和缺失值;同时 BP 神经网络模型可以很好地处理非线性问题^[14-15]。

集成学习模型建模流程(图1)如下:①将各个指标的水质历史数据集按照监测时间顺序分割为训练、验证以及测试数据集;②针对训练数据集分别建立 ARIMA 模型和 Prophet 模型,计算两个模型对验证数据集的预测结果,将其作为输入数据,结合验证数据集中的输出数据形成学习器结果数据集;③将该结果数据集分割为训练、验证数据集,针对训练数据集建立 BP 神经网络模型,并根据验证数据集的预测效果判断模型是否提前结束训练;④评估每个模型在测试数据集上的预测精度。

1.1 ARIMA 模型

ARIMA 模型由自回归(auto regression, AR)模型和移动平均(moving average, MA)模型组成,包括 p 、 d 、 q 3 个参数,其中 p 为 AR 模型参数, q 为 MA 模型参数, d 为时间序列转换为平稳序列所做的差分

次数^[16]。

a. AR 模型。该模型可以描述当前时刻水质数据与历史水质数据之间的关系,也就是利用历史数据对当前时刻数据进行预测。带有随机扰动项的 p 阶 AR 模型可以写为

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + u_t = \sum_{i=1}^p \alpha_i X_{t-i} + u_t \quad (1)$$

式中: X_t 为 t 时刻的水质数据; p 为用于预测的时刻数; u_t 为随机扰动项; α_i 为模型参数。

b. MA 模型。该模型认为水质序列数据是平稳的,也就是将 AR 模型中的随机扰动项 u_t 看作一个 q 阶的移动平均项,就有下面的形式:

$$u_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q} = \varepsilon_t + \sum_{j=1}^q \beta_j \varepsilon_{t-j} \quad (2)$$

式中: ε_t 为白噪声; q 为受白噪声序列影响的时刻数; β_j 为模型参数。

c. ARIMA 模型。对于不平稳的水质序列数据经过差分得到平稳序列,其中差分次数为 d 。将 AR(p)和 MA(q)结合得到 ARIMA(p, d, q)模型的表达式为

$$X_t = \varepsilon_t + \sum_{i=1}^p \alpha_i X_{t-i} + \sum_{j=1}^q \beta_j \varepsilon_{t-j} \quad (3)$$

1.2 Prophet 模型

Prophet 模型是一种开源的可处理时间序列数据的模型,可以很好地处理时间序列中的缺失值以及异常值对预测的影响,适用于水质数据的预测分析^[17]。

Prophet 模型可将水质时间序列数据写为如下关于时间 t 的函数形式:

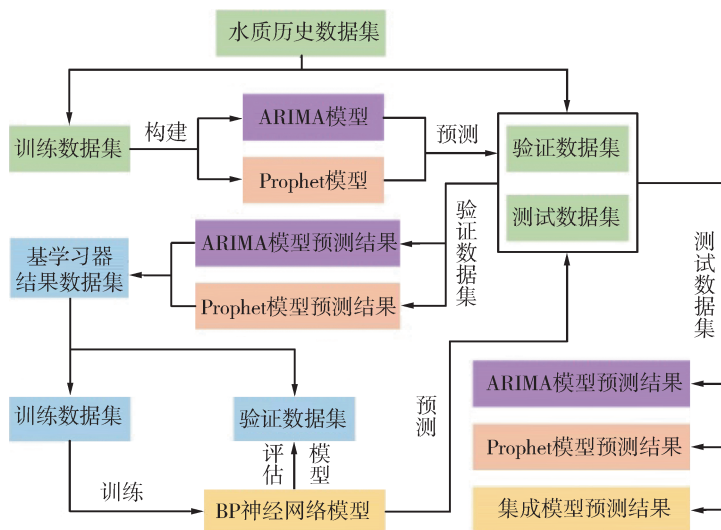


图1 集成学习模型建模流程

Fig.1 Modeling flowchart of ensemble learning model

$$y(t) = g(t) + s(t) + h(t) + \xi_t \quad (4)$$

式中: $g(t)$ 为趋势项,是时间序列中非周期性变化趋势的部分; $s(t)$ 为周期项,是时间序列中呈现周期性变化的部分; $h(t)$ 为节假日项,是该序列中受节假日影响的部分,因为在水质预测中节假日的影响较小,在本文中并不考虑这一项; ξ_t 为剩余项。

a. 趋势项。在 Prophet 算法中趋势项可通过逻辑回归函数和分段线性函数两种不同的函数来计算。考虑到水质数据自身的特性,本文采用分段线性函数来模拟趋势项。基于分段线性函数的趋势项形如:

$$g(t) = (k + a(t)\delta)t + (m + a(t)^T\gamma) \quad (5)$$

$$\text{其中 } \gamma = (\gamma_1, \gamma_2, \dots, \gamma_s)^T$$

$$\gamma_j = -s_j\delta_j \quad (j = 1, 2, \dots, S)$$

式中: k 为增长率; $a(t) \in \{0, 1\}^S$ 为指示函数; $\delta \in \mathbf{R}^S$ 为增长率的变化量; δ_j 为在时间戳 s_j 上的增长率的变化量; m 为补偿参数; S 为突变点的个数。

b. 周期项。使用傅里叶级数来模拟时间序列的周期性,周期项可以写为下面的形式:

$$s(t) = \sum_{n=1}^N \left[\alpha_n \cos\left(\frac{2\pi n t}{H}\right) + b_n \sin\left(\frac{2\pi n t}{H}\right) \right] = \mathbf{w}(t)\boldsymbol{\beta} \quad (6)$$

$$\text{其中 } \boldsymbol{\beta} = (\alpha_1, b_1, \alpha_2, b_2, \dots, \alpha_N, b_N)^T$$

式中: H 为时间序列的周期; $\boldsymbol{\beta}$ 为 Prophet 模型的初始化参数, $\boldsymbol{\beta}$ 的初始化为 $\boldsymbol{\beta} \sim N(0, \sigma^2)$, σ 越大,表示季节效应越明显, σ 越小,表示季节效应越不明显,本文 $\sigma = 0.7$ 。

在水质预测中设置以年为周期,也就是 $H = 365.25$, $N = 10$,此时

$$\mathbf{w}(t) = \left(\cos\left(\frac{2\pi(1)t}{365.25}\right), \sin\left(\frac{2\pi(1)t}{365.25}\right), \dots, \cos\left(\frac{2\pi(10)t}{365.25}\right), \sin\left(\frac{2\pi(10)t}{365.25}\right) \right) \quad (7)$$

2 实例验证

2.1 数据来源

选择长江流域某监测断面 2019—2020 年溶解氧(DO)、高锰酸盐指数(COD_{Mn})、氨氮(NH₃-N)、总磷(TP)、总氮(TN)等 5 个指标的水质自动监测数据(共 4386 条)来验证模型。其中 2019 年 1 月 1 日至 12 月 31 日的 2190 条数据作为训练数据集,2020 年 1 月 1 日至 10 月 26 日的 1800 条数据作为验证数据集,2020 年 10 月 27 日至 2020 年 12 月 31 日的 396 条数据作为测试数据集。构建集成学习模型阶段,2020 年 1 月 1 日至 9 月 26 日的 1620 条数据作为训练数据集,2020 年 9 月 27 日至 2020 年 10 月

26 日的 180 条数据作为验证数据集。因为 DO 的自动监测频次为 1 h 1 次,其他指标均为 4 h 1 次,为了统一时间精度,DO 数据选用其他指标监测时刻的数据。

2.2 ARIMA 模型验证结果

2.2.1 序列平稳性检验

利用 ADF 检验进行序列平稳性判断。如果该检验统计量小于置信度 1% 对应的值,并且显著性 P 值小于 0.05,则表示序列是平稳的。由表 1 可知,除 COD_{Mn} 外,其他指标 ADF 检验结果均小于置信度 1% 对应的值,并且 P 值均小于 0.05,表明数据序列是平稳的。对 COD_{Mn} 序列进行一阶差分,计算得到检验统计量为 -14.461,小于置信度 1% 对应的值 -3.431 852, P 值为 6.842×10^{-27} ,表明经过一阶差分后数据变为平稳序列。

表 1 数据平稳 ADF 检验

Table 1 Data stationary ADF test

指标	检验统计量	P	置信度 1% 对应的值	是否平稳
DO	-4.607	1.254×10^{-4}	-3.433 374	是
COD _{Mn}	-3.054	3.018×10^{-2}	-3.433 377	否
NH ₃ -N	-8.587	7.453×10^{-14}	-3.433 353	是
TP	-5.423	2.983×10^{-6}	-3.433 374	是
TN	-5.114	1.317×10^{-5}	-3.433 366	是

2.2.2 序列随机性检验

Ljung-Box 检验可评判时间序列数据是否为随机序列,如果为随机序列则不能构建 ARIMA 模型。设置最大延后期为 40,经过计算,5 个水质指标数据序列得到的 P 值均接近于 0,表明序列不是随机序列。

2.2.3 确定模型参数

针对每个水质指标,根据赤池信息准则(Akaike information criterion, AIC)和贝叶斯信息准则(Bayesian information criterion, BIC),采用参数组合的方式,将使得 AIC 与 BIC 的和最小的 p 、 q 作为该水质指标 ARIMA 模型的参数。本文 p 、 q 的选择范围均为 0~6, d 的取值范围为 0~2。各水质指标的最佳模型参数见表 2。

表 2 ARIMA 模型参数

Table 2 ARIMA model parameters

指标	p	d	q	模型
DO	5	0	4	ARIMA(5,0,4)
COD _{Mn}	6	0	6	ARIMA(6,0,6)
NH ₃ -N	1	0	4	ARIMA(1,0,4)
TP	6	1	5	ARIMA(6,1,5)
TN	6	1	6	ARIMA(6,1,6)

2.2.4 模型构建与验证

ARIMA 模型的残差应该满足均值为 0 且方差为常数的正态分布,并且不具备自相关性。根据各

水质指标对应的最佳模型进行训练数据的拟合,对训练数据的模型预测值与监测值之间的残差序列进行正态分布检验和 Durbin-Watson (DW) 检验(表3),其中正态分布检验的统计量 p 应大于 0.05, DW 检验的值应接近于 2。如果残差近似正态分布并且不满足自相关性,说明构建的模型是适用于该指标数据序列分析的。

表 3 残差检验

Table 3 Residual test

指标	p	DW 检验值
DO	0.056	1.946
COD _{Mn}	0.089	2.054
NH ₃ -N	0.124	1.992
TP	0.077	2.047
TN	0.135	1.992

由表 3 可知,各水质指标模型的残差序列均满足正态分布并且不存在自相关性,表明建立的 ARIMA 模型适用于这 5 个水质指标的时间序列数据分析。本文的预测方式为静态预测,也就是用真实的监测数据预测下一时刻的数据。采用平均绝对误差(mean absolute error, MAE)、平均绝对百分比误差(mean absolute percentage error, MAPE)以及皮尔逊相关系数(r)进行模型预测精度评价,其中 MAE 是预测值与真实值之差的绝对值的均值,可以评价预测结果的误差情况;MAPE 是预测值与真实值之差的绝对值与真实值比值的均值,可以用于计算模型误差占真实值的比例;皮尔逊相关系数可以用来判断预测序列和真实序列趋势的一致性,接近于 1 说明模型预测精度较高。表 4 为模型在测试数据集上的预测效果。

表 4 ARIMA 模型预测精度评价

Table 4 Evaluation of prediction accuracy of

ARIMA model

指标	MAE/(mg · L ⁻¹)	MAPE/%	r
DO	0.986	12.97	0.859
COD _{Mn}	0.470	9.40	0.904
NH ₃ -N	0.157	18.67	0.800
TP	0.022	13.38	0.876
TN	0.261	10.36	0.823

由表 4 可知,ARIMA 模型预测结果的 MAPE 均在 20% 以下,相关系数均在 0.8 以上,表明构建的模型预测效果较好。

2.3 Prophet 模型验证结果

对各水质指标的训练数据集构建 Prophet 模型,表 5 为各水质指标在测试数据集上的预测精度评价结果。

因为 Prophet 模型是将原始的水质监测序列分解为不同的项,导致其预测值并不能很好地反应监

表 5 Prophet 模型预测精度评价

Table 5 Evaluation of prediction accuracy of

Prophet model

指标	MAE/(mg · L ⁻¹)	MAPE/%	PIC/%
DO	1.055	13.02	98.2
COD _{Mn}	0.328	6.52	94.7
NH ₃ -N	0.235	21.24	88.8
TP	0.030	24.29	91.6
TN	0.648	20.77	87.1

注:PIC 为区间覆盖占比,是监测数据落在预测值区间内的个数与监测数据总数的比值。

测数据的时间变化,因此和监测值序列的皮尔逊相关性较低,因此用 PIC 作为评价模型预测性能的指标。从表 5 可知,5 个指标的 PIC 值均在 90% 左右,并且 MAPE 在 20% 上下,说明该模型的预测具有一定的精度。

2.4 集成学习模型验证结果

利用 BP 神经网络模型对两个模型的结果进行融合。选用 6 层的神经网络,其中包含 4 个隐层,每个隐层有 80 个神经单元。将 ARIMA 模型针对验证数据集的 5 个水质指标预测结果、Prophet 模型的预测结果以及预测区间的上限值、下限值一起作为 BP 神经网络的输入数据,将水质的监测数据作为输出数据,构成模型数据集,并按照 9:1 的比例将模型数据集分割为训练和验证数据集。

模型迭代次数设置为 5 000 次,优化求解器采用 Adamax。模型训练过程中随着迭代次数的增加,训练、验证数据集的误差逐渐降低,为避免模型过拟合,需要根据验证数据集的预测结果对模型进行评估,当迭代次数为 3 430 时模型的预测性能最佳。

从表 6 可看出,DO、COD_{Mn}、NH₃-N、TP 和 TN 5 个水质指标集成学习模型的 MAPE 比 ARIMA、Prophet 模型中较低的分别低 35.0%、29.9%、4.1%、40.6% 和 17.1%,皮尔逊相关系数和 ARIMA 模型相比相差不大,基本在 0.8 ~ 1.0 之间,说明集成学习模型的预测精度较单一模型精度高。

表 6 集成学习模型预测精度评价

Table 6 Evaluation of prediction accuracy of

ensemble learning model

指标	MAE/(mg · L ⁻¹)	MAPE/%	r
DO	0.668	8.43	0.895
COD _{Mn}	0.224	4.57	0.838
NH ₃ -N	0.117	17.91	0.973
TP	0.009	7.95	0.981
TN	0.234	8.59	0.972

3 结 语

针对水质数据预测建立了一种基于 ARIMA 和 Prophet 的集成学习模型。集成学习模型中的时间

序列分析模型可以很好地捕捉水质数据中存在的趋势性信息,利用 BP 神经网络模型进行集成可以对数据施加非线性的处理。这种通过融合不同原理模型预测结果的方式,集成了不同模型的优点,可以学习到更多数据序列存在的信息,从而使得预测结果更为精确。采用长江流域某监测断面 5 个水质指标的监测数据进行了验证,结果显示集成学习模型的评估指标均优于单一时间序列模型,表明集成学习模型的水质预测精度更高,可以应用于实际水环境管理中的水质预测,为实现更科学、精准的水质自动化预警提供数据支持。

参考文献:

- [1] 嵇晓燕,肖建军,杨凯. 复合河流系统健康监测和预警理论初探[C]//2020 中国环境科学学会科学技术年会议论文集(第三卷). 北京:中国环境科学学会,2020:473-485.
- [2] 尤宾,臧红霞,李永丽. 淮河干流信阳段水质预测模型[J]. 水资源保护,2013,29(1):41-45. (YOU Bin, ZANG Hongxia, LI Yongli. Water quality prediction model for Xinyang section of main stream of Huaihe River[J]. Water Resources Protection, 2013, 29(1): 41-45. (in Chinese))
- [3] 张影,杨晓东,贾子璇,等. 基于改进重叠投影法的空间耦合水污染评估模型[J]. 水资源保护,2021,37(2):128-134. (ZHANG Ying, YANG Xiaodong, JIA Zixuan, et al. Space coupling water pollution evaluation model based on improved overlapping projection method[J]. Water Resources Protection, 2021, 37(2): 128-134. (in Chinese))
- [4] 谢建辉,闫学军,孙开争,等. ARIMA 模型在河流水质预测中的应用[C]//2015 年中国环境科学学会学术年会议论文集(第一卷). 北京:中国环境科学学会,2015:1239-1243.
- [5] 龙军,冀海峰,黄志尧,等. 时间序列分析在水质预测中的应用[C]//中国仪器仪表学会第十一届青年学术会议论文集. 北京:中国仪器仪表学会,2009:354-356.
- [6] 黄廷林,卢金锁,韩宏大,等. 地表水源水质预测方法研究[J]. 西安建筑科技大学学报(自然科学版),2004,36(2):134-137. (HUANG Tinglin, LU Jinsuo, HAN Hongda, et al. Forecasting methods of surface water quality[J]. Xi'an University of Architecture and Technology, 2004, 36(2): 134-137. (in Chinese))
- [7] 向速林,杨柳春,冉全,等. 回归分析法在地下水水质动态预测中的应用[J]. 长江大学学报(自科版)理工卷,2006,3(1):40-42. (XIANG Sulin, YANG Liuchun, RAN Quan, et al. Application of regression analytical method in dynamic prediction of groundwater quality[J]. Journal of Yangtze University (Natural Science Edition), 2006, 3(1): 40-42. (in Chinese))
- [8] 刘阳. 灰色模型在水库水质预测中的应用研究[J]. 水土保持应用技术, 2019(5): 23-24. (LIU Yang. Application of grey model in water quality prediction of reservoir[J]. Technology of Soil and Water Conservation, 2019(5): 23-24. (in Chinese))
- [9] 崔东文. 极限学习机在湖库总磷、总氮浓度预测中的应用[J]. 水资源保护, 2013, 29(2): 61-66. (CUI Dongwen. Application of extreme learning machine to total phosphorus and total nitrogen forecast in lakes and reservoirs[J]. Water Resources Protection, 2013, 29(2): 61-66. (in Chinese))
- [10] 袁健,树锦. 改进多元回归法与神经网络应用于水质预测[J]. 水资源保护, 2008, 24(3): 46-48. (YUAN Jian, SHU Jin. Application of modified multivariate nonlinear regression analysis model and neural network model to water quality prediction[J]. Water Resources Protection, 2008, 24(3): 46-48. (in Chinese))
- [11] 吴娟,朱跃龙,金松,等. 三种机器学习模型在太湖藻华面积预测中的应用[J]. 河海大学学报(自然科学版), 2020, 48(6): 542-551. (WU Juan, ZHU Yuelong, JIN Song, et al. Area prediction of cyanobacterial blooms based on three machine learning methods in Taihu Lake[J]. Journal of Hohai University (Natural Sciences), 2020, 48(6): 542-551. (in Chinese))
- [12] 崔东文. 多组群教学优化算法-神经网络-支持向量机组合模型在径流预测中的应用[J]. 水利水电科技进展, 2019, 39(4): 41-48. (CUI Dongwen. Combined model of multi-group teaching optimization algorithm-neural network-support vector machine in runoff prediction application[J]. Advances in Science and Technology of Water Resources, 2019, 39(4): 41-48. (in Chinese))
- [13] 涂吉昌,陈超波,王景成,等. 基于深度学习的水质预测模型研究[J]. 自动化与仪表, 2019, 34(6): 96-100. (TU Jichang, CHEN Chaobo, WANG Jingcheng, et al. Research on water quality prediction model based on deep learning[J]. Automation & Instrumentation, 2019, 34(6): 96-100. (in Chinese))
- [14] CAI Rong, HAN Taihao, LIAO Wenyu, et al. Prediction of surface chloride concentration of marine concrete using ensemble machine learning[J]. Cement and Concrete Research, 2020, 136(10): 1061-1084.
- [15] YU Xiaohan, XIONG Shengwu, HE Ying, et al. Research on campus traffic congestion detection using BP neural network and Markov model[J]. Journal of Information Security and Applications, 2016, 31(11): 54-60.
- [16] BOX G E P, JENKINS G M. Time series analysis: forecasting and control[J]. Journal of the American Statistical Association, 2010, 31(3): 84-95.
- [17] TAYLOR S J, LETHAM B. Forecasting at scale[J]. The American Statistician, 2018, 72(1): 37-45.

(收稿日期:2021-08-18 编辑:熊水斌)