

doi: 10.11835/j.issn.1000-582X.2025.04.009

引用格式:黄威,贾若然,钟坤华,等.基于XGB-KF模型的农业温室温度预测[J].重庆大学学报,2025,48(4): 108-114.



基于XGB-KF模型的农业温室温度预测

黄威^{1,2},贾若然³,钟坤华^{1,2},刘曙光^{1,2}

(1. 中国科学院重庆绿色智能技术研究院,重庆 400714; 2. 中国科学院大学,北京 100049; 3. 科大讯飞股份有限公司,合肥 230031)

摘要:针对农业温室大棚温度测量受噪声影响不易直接预测的问题,提出一种将XGBoost (extreme gradient boosting) 和 Kalman filter 相结合的集成预测模型 XGB-KF (extreme gradient boosting with Kalman filter)。该模型首先基于XGBoost对温室内部当前时刻的温度值进行初步估计,使用卡尔曼滤波(Kalman filter)对得到的估计结果进行动态修正,得到最终的预测结果。基于涿州地区农业温室大棚的传感器数据进行了数值实验,以均方根误差(root mean square error, RMSE)作为主要指标对模型进行性能评估。与XGBoost、Bi-LSTM 和 Bi-LSTM-KF 方法相比较, XGB-KF 的 RMSE 分别降低 5.22%、10.85%、7.45%。

关键词:集成模型;机器学习;时间序列;温室温度

中图分类号:TP399

文献标志码:A

文章编号:1000-582X(2025)04-108-07

Agricultural greenhouse temperature prediction based on the XGB-KF model

HUANG Wei^{1,2}, JIA Ruoran³, ZHONG Kunhua^{1,2}, LIU Shuguang^{1,2}

(1. Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, P. R. China; 2. University of Chinese Academy of Sciences, Beijing 100049, P. R. China; 3. Iflytek Co., Ltd., Hefei 230031, P. R. China)

Abstract: To address the challenge of agricultural greenhouse temperature measurement being highly susceptible to noise, which limits direct prediction accuracy, this study proposes an integrated prediction model, XGB-KF, combining XGBoost and the Kalman filter. First, the model estimates the current greenhouse temperature using XGBoost. Then, the Kalman filter dynamically adjusts the estimated result to refine the prediction. Numerical experiments are conducted using sensor data from a greenhouse in Zhuozhou, with root mean square error (RMSE) as the main evaluation metric. Compared with XGBoost, Bi-LSTM, and Bi-LSTM-KF methods, the XGB-KF model reduces RMSE by 5.22%, 10.85% and 7.45% respectively.

Keywords: integrated model; machine learning; time series; greenhouse temperature

收稿日期:2021-01-22

基金项目:中国科学院重点资助项目(E351600201)。

Supported by Key Research Programs of Chinese Academy of Sciences (E351600201).

作者简介:黄威(1998—),男,硕士研究生,主要从事数据挖掘、时间序列预测方向研究。

通信作者:刘曙光,男,硕士生导师,(E-mail)liushuguang@cigit.ac.cn。

温度作为农作物生长的重要环境因素,影响着作物的生长速度、产量和质量,如何准确预测和调控温度成为现代农业的重要问题。温室大棚拥有可调控的环境,促进作物生长和提高产量。中国温室建设起步较早,经历了改良型日光温室、大型玻璃温室和现代化温室3个阶段,但受各地区生产状况和经济条件的影响,至今各阶段不同类型的温室依然并存^[1]。温室温度可以直接通过温度传感器来测量,但温室中的各种环境因子对温度测量会产生干扰^[2],并带来较多噪声。农业温室温度预测可以看作是一个时间序列预测问题。时间序列预测的方法目前大致可分为4类:传统时间序列分析法、机器学习法、深度学习法及混合模型方法。

1)传统时间序列分析方法,如自回归移动平均模型(autoregressive moving average, ARMA)^[3]、差分自回归移动平均模型(autoregressive integrated moving average, ARIMA)^[4]等;

2)机器学习方法。针对温室温度或湿度的预测,极限学习机及其改进方法^[5-6]被研究和应用。Yu等^[7]采用粒子群算法对最小二乘支持向量机做优化,在中国农业大学寿光蔬菜产业集团技术应用与示范区的园艺监测和管理系统的数据上进行实验,结果表明,其对最高和最低温度的预测比标准支持向量机和反向传播神经网络的预测更为准确。

3)深度学习方法。周翔宇^[2]等针对浅层神经网络面对温室复杂多变环境因子表征能力弱、学习时间长的的问题,提出一种基于改进深度信念网络并结合经验模态分解与门控循环单元的温室预测方法。Jung等^[8]使用深度学习方法进行温室预测和管控,在人工神经网络(artificial neural network, ANN)、带外生输入的非线性自回归模型(nonlinear autoregressive with exogenous inputs, NARX)以及长短时记忆循环神经网络(recurrent neural network with long short-term memory, RNN-LSTM)3种方法的对比中,RNN-LSTM取得最好效果。文献[9-11]则分别基于循环神经网络(recurrent neural network, RNN)和长短时记忆网络(long short-term memory, LSTM)方法对温室内部的多维元素进行学习。

4)混合模型法。Song等^[12]使用结合卡尔曼滤波的长短时记忆网络(long short-term memory with kalman filter, LSTM-Kalman)模型对空气质量数据进行静态预测和动态调整,实验证明加入卡尔曼滤波方法后模型效果更好。Yan等^[13]针对厄尔尼诺-南方涛动(El Niño-southern oscillation, ENSO)超前预测的问题使用了集合经验模态分解-时间卷积网络(ensemble empirical mode decomposition with temporal convolutional network, EEMD-TCN),结果表明了方法的优越性。Hu等^[14]使用双向长短时记忆卷积神经网络(convolutional neural network-bidirectional long short-term memory, CNN-Bi-LSTM)对城市供水量进行预测,在与LSTM、Bi-LSTM、卷积神经网络(convolutional neural network, CNN)、堆叠自编码器(Stacked Autoencoder, SAE)以及长短时记忆卷积神经网络(convolutional neural network with long short-term memory, CNN-LSTM)4种方法的比较中取得较低误差和更快收敛速度。

卡尔曼滤波作为一种最优线性状态估计方法,由卡尔曼在1960年提出^[15],可以在诸多不确定性的情况下估计动态系统的状态,即通过数学方法寻求与观测数据最佳拟合的状态向量,预测存在噪声的数值。

针对农业温室温度预测问题,研究提出一种将XGBoost(extreme gradient boosting)和Kalman filter相结合的集成模型XGB-KF。该模型首先利用温室内外相关协变量数据,基于XGBoost对温室内当前时刻的温度进行初步预测,再通过Kalman filter对估计结果进行动态修正。实验结果表明,XGB-KF法在温室温度预测上更准确有效。

1 方法理论

1.1 XGBoost

XGBoost是一种基于梯度提升决策树(gradient boosting decision tree, GBDT)的改进算法^[16]。GBDT是集成模型,训练时每一次迭代都学习一棵CART树来拟合之前 $t-1$ 棵树的预测结果与训练样本真实值的残差。已知训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$,损失函数 $l(y_i, \hat{y}_i)$,正则化项 $\Omega(f_k)$,整体函数可以表示为

$$F(X) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad (1)$$

其中: $F(X)$ 是线性空间上的表达; i 表示第*i*个样本; k 表示第*k*棵树; \hat{y}_i 是第*i*个样本 x_i 的预测值: $\hat{y}_i = \sum_{k=1}^K f_k(x_i)$; $\sum_{i=1}^n \Omega(f_k)$ 表示*k*棵树的复杂度, f_k 表示第*k*棵树的函数。

由于

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), \quad (2)$$

则目标函数可以转化成如下形式

$$F = \sum_{i=1}^n l[y_i, \hat{y}_i^{(t-1)} + f_t(x_i)] + \sum_{k=1}^K \Omega(f_k). \quad (3)$$

1.2 Kalman filter

卡尔曼滤波适合使用在具有不确定性的动态系统中,基于系统中的观测值和估计值得到一个比任何依据自身估计更好的结果,从而对系统的下一步走向做出预测。

1.2.1 基本模型

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k + \boldsymbol{\omega}_{k-1}, \quad (4)$$

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k, \quad (5)$$

其中: \mathbf{x}_k 表示系统状态矩阵; \mathbf{z}_k 表示状态阵的观测值; \mathbf{A} 表示状态转移矩阵; \mathbf{B} 表示控制输入矩阵; \mathbf{u}_k 表示外界对于控制系统的输入; \mathbf{H} 表示状态观测矩阵; $\boldsymbol{\omega}_{k-1}$ 表示过程噪声, \mathbf{v}_k 表示测量噪声。

1.2.2 状态更新

根据上一时刻的最优估计值和控制量得到当前时刻的最优估计值,根据上一时刻最优估计值的协方差得到当前时刻的最优估计协方差^[15]

$$\hat{\mathbf{x}}_k = \mathbf{F}_k \hat{\mathbf{x}}_{k-1} + \mathbf{B}_k \mathbf{u}_k, \quad (6)$$

$$\mathbf{P}_k = \mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^T + \mathbf{Q}_k, \quad (7)$$

在温室温度预测的问题上不存在外部控制量,所以没有 $\mathbf{B}_k \mathbf{u}_k$ 这一部分。通过卡尔曼增益对估计值和观测值作加权融合,作出“更可信”的决策,同时更新协方差阵

$$\hat{\mathbf{x}}_k' = \hat{\mathbf{x}}_k + \mathbf{K}'(\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_k), \quad (8)$$

$$\mathbf{P}_k' = \mathbf{P}_k - \mathbf{K}' \mathbf{H}_k \mathbf{P}_k, \quad (9)$$

$$\mathbf{K}' = \mathbf{P}_k \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^T + \mathbf{R}_k)^{-1}. \quad (10)$$

2 模型及评价指标

2.1 模型流程

XGB-KF模型的流程如图1所示:

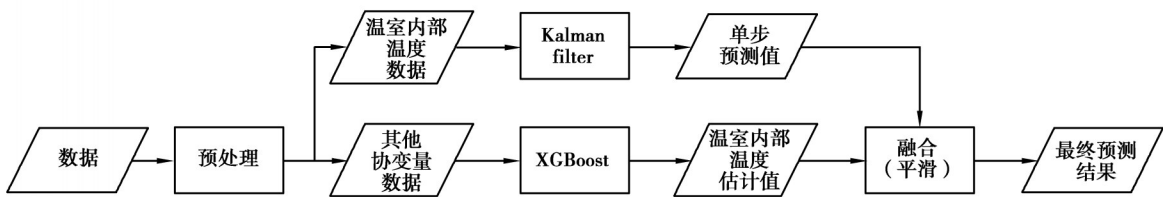


图1 XGB-KF模型流程

Fig. 1 Flow chart of XGB-KF model

原始数据经预处理后,协变量数据(温室内、外部湿度、气压)作为XGBoost的输入数据,温室内部温度值作为XGBoost的预测目标。训练好的XGBoost被用于估计当前时刻的温室内部温度。与此同时,Kalman filter使用温室内部温度数据作为输入,通过数据的历史观测值得到当前时刻的单步预测结果。最后,将

Kalman Filter得到的结果作为系统观测值,将XGBoost得到的结果作为系统估计值,融合这2个值得到最终预测结果。

2.2 评价指标

为评估模型的性能和方法的有效性,实验采用评价指标:均方根误差(RMSE)和拟合优度(R^2)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}. \quad (12)$$

3 数值实验

3.1 数据集

数据来自科大讯飞开放平台提供的中国农业大学涿州实验站的温室温度数据,数据的时间范围为2019年3月14日1点-2019年4月3日0点,每条数据由时间戳和温室内、外的温度、湿度、气压构成,采样频率为每min 1次。实验中,数据按4:1的比例划分训练集和测试集,最终的实验结果由各方法在测试集上评估得到。

3.2 数据预处理

温室内、外的温度变化情况可视化如图2所示。温室内、外温度变化范围较大,且温室内部温度和外部温度强相关。为更好地获得预测结果,对温度进行平稳处理,分别对温室内、外的温度取log平滑再进一步做差,处理后温室内温度和温室内、外温度对数差的变化情况如图3所示。

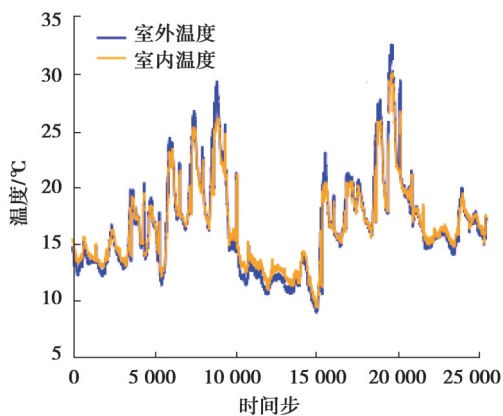


图2 温室内、外温度变化情况

Fig. 2 The change of temperature inside and outside the greenhouse

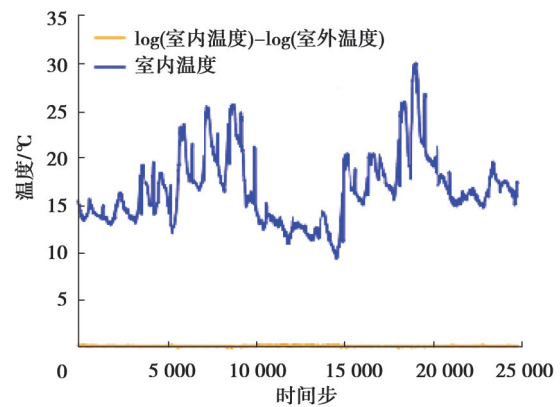


图3 室内温度平稳化结果

Fig. 3 Results of indoor temperature stabilization

删除温室内部温度的缺失数据,以免填充方法不当造成更大误差,然后采用 3σ 法则检测气压特征数据中存在的异常值,使用前后数据进行线性修正。最后,通过皮尔逊相关系数筛选出相关性较高的3个特征(温室外部和温室内部的湿度差值、比值以及湿度的乘积)添加进数据集。

3.3 实验内容及环境

3.3.1 实验内容

对划分过的数据集使用XGBoost、Bi-LSTM、XGB-KF以及Bi-LSTM-KF 4种方法分别进行实验,在训练集上训练XGBoost和Bi-LSTM单模型,在测试集上使用单模型进行预测。XGB-KF和Bi-LSTM-KF 2种模型则在测试集上对单模型的预测结果使用Kalman filter做进一步的调整和修正,修正后的结果作为集成模型的最终预测结果。

3.3.2 实验环境

实验在 Windows 系统下使用 anaconda、jupyter-notebook 及 python 进行。其中, anaconda 版本 4.8.4, jupyter-notebook 版本 6.0.0, python 版本 3.7.3。

3.4 结果及对比分析

3.4.1 XGBoost

最佳状态下, XGBoost 单模型取得的均方根误差(RMSE)为 0.217 07。预测效果如图 4 所示。

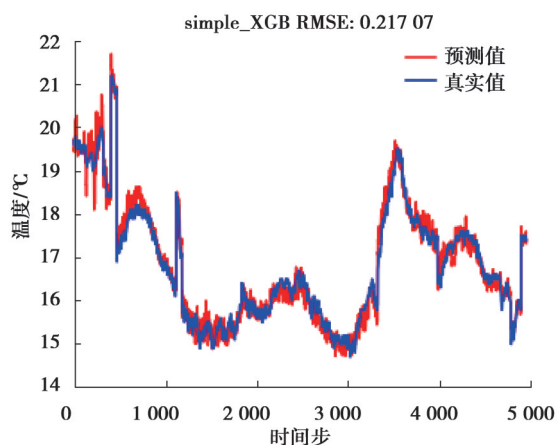


图4 XGBoost的预测结果

Fig. 4 Prediction result of XGBoost

3.4.2 Bi-LSTM

Bi-LSTM 取得的最低 RMSE 为 0.230 77, 预测效果如图 5 所示。

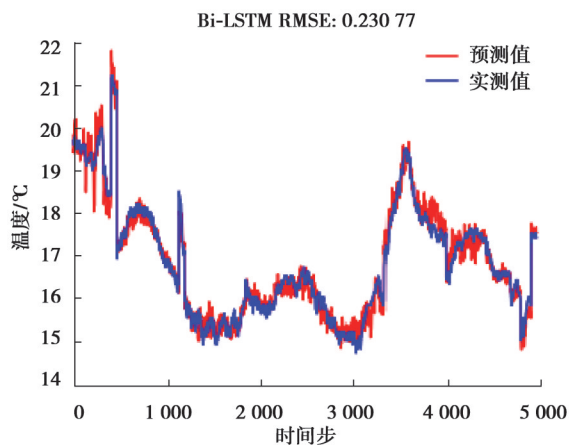


图5 Bi-LSTM 预测结果

Fig. 5 Prediction result of Bi-LSTM

3.4.3 XGB-KF

加入 Kalman filter 方法对 XGBoost 的预测结果进行修正, 修正后 RMSE 和 R^2 变化趋势如图 6~7。选取实验中的最佳参数运行模型, 相较于 XGBoost 单模型, XGB-KF 模型的 RMSE 值从 0.217 07 降至 0.205 73, 降低 5.22%。 R^2 值从 0.974 33~0.976 94, 提升 0.27%。

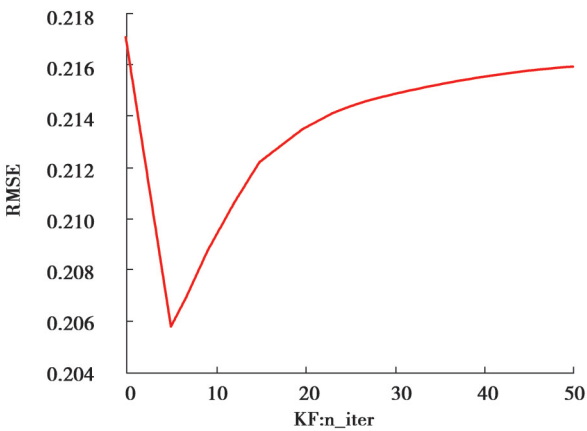


图 6 XGB-KF 模型的 RMSE 变化趋势
Fig. 6 RMSE trend of XGB-KF model

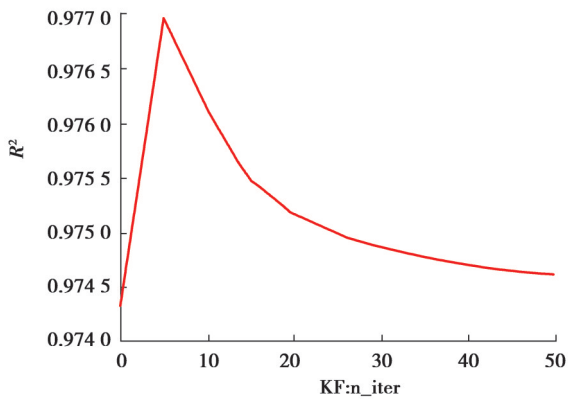


图 7 XGB-KF 模型的 R^2 变化趋势
Fig. 7 R^2 trend of XGB-KF model

3.4.4 Bi-LSTM-KF

加入 Kalman filter 对 Bi-LSTM 的预测结果进行修正后,Bi-LSTM 的 RMSE 从 0.230 77 降至 0.222 29,下降 3.67%。 R^2 从 0.993 79 降至 0.973 08,下降 2.08%。RMSE 和 R^2 的趋势变化情况如图 8 和图 9 所示。

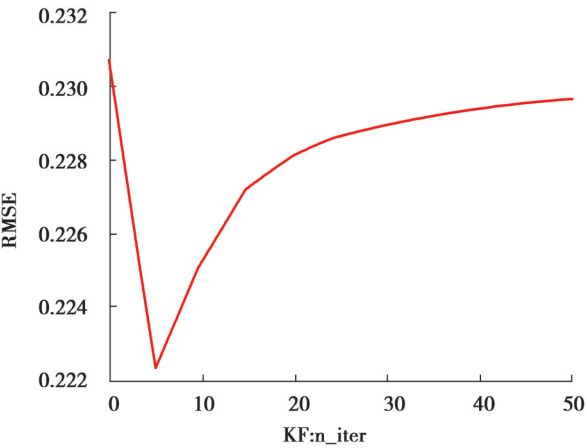


图 8 Bi-LSTM-KF 模型的 RMSE 变化趋势
Fig. 8 RMSE trend of Bi-LSTM-KF model

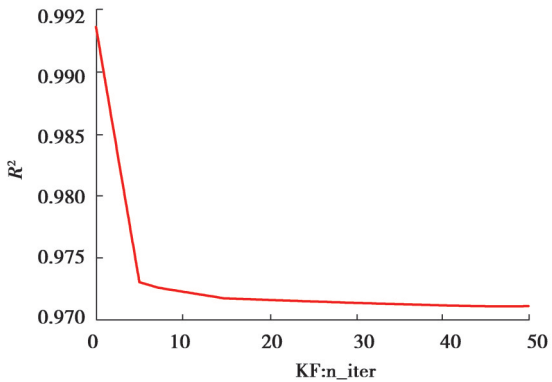


图 9 Bi-LSTM-KF 模型的 R^2 变化趋势
Fig. 9 R^2 trend of Bi-LSTM-KF model

3.4.5 R^2 对比分析

各方法的指标评估结果如表 1 所示,考虑到拟合优度(R^2)值并不能完全代表模型的预测能力,且这 4 种方法的拟合优度(R^2)值均较高,所以实验选取均方根误差(RMSE)作为模型的主要评估指标。最后,在与 XGBoost、Bi-LSTM、Bi-LSTM-KF 3 种方法的比较中,XGB-KF 模型取得了最低均方根误差(RMSE),证明其在农业温室温度预测问题中的有效性。同时,实验在仅使用 CPU 运行程序情况下对 XGBoost 和 Bi-LSTM 的训练时间进行记录和对比,得到 XGBoost 方法的运行时间为 4.8 s,Bi-LSTM 方法的运行时间为 108.3 s。在约 2×10^4 条数据的训练集上,XGBoost 展现出明显的速度优势。

表 1 各方法的指标评估结果

Table 1 Indicator evaluation results of each method

方法	RMSE	R^2
Bi-LSTM	0.230 77	0.993 79
XGBoost	0.217 07	0.974 33
Bi-LSTM-KF	0.222 29	0.973 08
XGB-KF	0.205 73	0.976 94

4 结 语

针对温室温度测量受噪声影响大的问题,提出集成模型 XGB-KF。经过在真实数据集上的实验,模型在与 XGBoost、Bi-LSTM、Bi-LSTM-KF 方法的对比中取得了最低的均方根误差。实验虽然只基于温室温度数据,但 XGB-KF 模型引入协变量预测的思路对时序预测的其他应用场景也有一定参考价值,如区域气温预测、城市供水量预测等。在未来的研究中,长短时记忆网络、时间卷积网络等优异方法可以被考虑用于时序预测的自回归问题中,同时它们的改进和组合方法也值得进一步尝试。

参考文献

- [1] 葛志军,傅理.国内外温室产业发展现状与研究进展[J].安徽农业科学,2008,36(35):15751-15753.
GE Z J, FU L. Development status and research progress of the greenhouse industry at home and abroad[J]. Journal of Anhui Agricultural Sciences, 2008, 36(35): 15751-15753. (in Chinese)
- [2] 周翔宇,程勇,王军.基于改进深度信念网络的农业温室温度预测方法[J].计算机应用,2019,39(4):1053-1058.
ZHOU X Y, CHENG Y, WANG J. Agricultural greenhouse temperature prediction method based on improved deep belief network[J]. Journal of Computer Applications, 2019, 39(4): 1053-1058. (in Chinese)
- [3] 左志宇,毛罕平,张晓东,等.基于时序分析法的温室温度预测模型[J].农业机械学报,2010,41(11):173-177.
ZUO Z Y, MAO H P, ZHANG X D, et al. Forecast model of greenhouse temperature based on time series method[J]. Transactions of the Chinese Society for Agricultural Machinery, 2010, 41(11): 173-177. (in Chinese)
- [4] Wang H, Huang J J, Zhou H, et al. An integrated variational mode decomposition and arima model to forecast air temperature[J]. Sustainability, 2019, 11(15): 4018.
- [5] Liu Q, Jin D D, Shen J, et al. A WSN-based prediction model of microclimate in a greenhouse using extreme learning approaches [C]//2016 18th International Conference on Advanced Communication Technology (ICACT). Piscataway: IEEE, 2016: 730-735.
- [6] Zou W D, Yao F X, Zhang B H, et al. Verification and predicting temperature and humidity in a solar greenhouse based on convex bidirectional extreme learning machine algorithm[J]. Neurocomputing, 2017, 249: 72-85.
- [7] Yu H H, Chen Y Y, Hassan S G, et al. Prediction of the temperature in a Chinese solar greenhouse based on LSSVM optimized by improved PSO[J]. Computers and Electronics in Agriculture, 2016, 122: 94-102.
- [8] Jung D H, Kim H S, Jhin C, et al. Time-serial analysis of deep neural network models for prediction of climatic conditions inside a greenhouse[J]. Computers and Electronics in Agriculture, 2020, 173: 105402.
- [9] Jung D H, Kim H J, Park S H, et al. Recurrent neural network models for prediction of the inside temperature and humidity in greenhouse[C]//Proceedings of the Korean Society for Agricultural Machinery Conference. South Korea: Korean Society for Agricultural Machinery, 2017: 135-135.
- [10] Song Y E, Moon A, AN S Y, et al. Prediction of smart greenhouse temperature-humidity based on multi-dimensional LSTMs[J]. Journal of the Korean Society for Precision Engineering, 2019, 36(3): 239-246.
- [11] Wang H K, Li L, Wu Y, et al. Recurrent neural network model for prediction of microclimate in solar greenhouse[EB/OL]. (2018-09-12)[2021-01-01]. <https://www.sciencedirect.com/science/article/pii/S2405896318312151>.
- [12] Song X J, Huang J J, Song D W. Air quality prediction based on LSTM-Kalman model[C]//2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Piscataway: IEEE, 2019: 695-699.
- [13] Yan J N, Mu L, Wang L Z, et al. Temporal convolutional networks for the advance prediction of ENSO[J]. Scientific Reports, 2020, 10(1): 8055.
- [14] Hu P, Tong J, Wang J C, et al. A hybrid model based on CNN and Bi-LSTM for urban water demand prediction[C]//2019 IEEE Congress on Evolutionary Computation (CEC). Piscataway: IEEE, 2019: 1088-1094.
- [15] Kalman R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1): 35-45.
- [16] Chen T Q, Guestrin C, et al. XGBoost[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

(编辑 侯 湘)