

Beetle swarm optimisation for solving investment portfolio problems[J]. The Journal of Engineering, 2018 (16):1600-1605.

[6] 沈涵,都海波,周俊.自适应变异的天牛群优化算法[J]. 计算机应用, 2020,40(S2):1-7.

[7] 王振东,曾勇,王俊岭,等.基于改进天牛群算法优化的 BP 神经网络的入侵检测[J].科学技术与工程, 2020, 20(32):13249-13257.

[8] 祝晓燕,张金会,付士鹏,等.基于改进 PSO 的 SVM 参数优化及其在风速预测中的应用[J].中国电力, 2013, 46(11):105-108.

[9] 殷丽娟,赵熙临,梅真.基于混沌粒子群优化算法的电网优化运行技术[J].电力系统及其自动化学报, 2016,28(5):55-61.

[10] JUN WANG,JUNXING CAO,SHAN YUAN. Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network[J]. Journal of Petroleum Science and Engineering, 2020(194):107466.

[11] LIN HSIAOCHUNG,WANG PING,LIN WENHUI, et al. A multiple-swarm particle swarm optimisation scheme for tracing packets back to the attack sources of botnet[J]. Applied Sciences, 2021,11(3):1139.

[12] 孙锋利,何明一,高全华.引入欧椋鸟群飞行机制的改进粒子群算法[J].计算机应用研究, 2012,29(5):1666-1669+1697.

[13] DE MONTES O C A,STUTZLE T,BIRATTARI M, et al. Frankenstein's PSO: a composite particle swarm optimization algorithm[J].IEEE Trans on Evolutionary Computation, 2009,13(5):1120- 1132.

Application of Distributed Beetle Swarm Optimization Algorithm in Classification

HUANG Song, CHEN Hongwei, BIAN Fan, YANG Weiwei, YANG Zhihui
(School of Computer Science, Hubei Univ. of Tech., Wuhan 430068, China)

Abstract: Logistic regression is commonly used in sentiment classification. Based on the limitations of traditional logistic regression classifier parameter adjustment, some parameters may not reach the global optimum, resulting in the low performance of the classifier. To solve this problem, we propose an improved beetle swarm optimization algorithm (IM-BSO) to optimize the hyperparameters of logistic regression to improve classification accuracy. The IM-BSO algorithm adopts an adaptive adjustment strategy of learning factors and inertia weights. The inertia weight of each beetle is different and changes with the change of the fitness value. In addition, the IM-BSO algorithm incorporates the K-means clustering and topology mechanism, which increases the diversity of the beetle swarm. Due to the large amount of data to be processed and the long calculation time of the IM-BSO algorithm, we propose a new distributed and improved beetle swarm optimization algorithm (DIBSO), combined with logistic regression to form a new classification model: the DIBSO-LR model. Finally, use the model to classify the sentiment of the Twitter comment data set at different numbers of nodes, comparing the speedup ratio, the experimental results show that within a certain range, the larger the amount of data, the more obvious the speedup effect will be as the number of nodes increases.

Keywords: beetle swarm optimization algorithm; logistic regression; sentiment classification; speedup ratio

[责任编辑：张岩芳]

[文章编号] 1003-4684(2022)01-0024-05

基于改进 SARIMA-LSTM 的海上风速预测方法

余聪聪, 熊才权, 徐仕强, 古小惠

(湖北工业大学计算机学院, 湖北 武汉 430068)

[摘 要] 为了提高海上风速预测的精度,提出了一种基于局部加权回归的周期趋势分解(STL)改进的季节性差分自回归移动平均模型(SARIMA)和长短时记忆(LSTM)神经网络的海上风速预测方法。首先通过 STL 分解原始风速时间序列,提高 SARIMA 模型季节性差分步长的准确性,再使用 SARIMA 模型对观测的风速序列数据进行预测,得到预测值以及预测值与观测值之间的残差;然后用残差样本集训练长短时记忆神经网络并对残差进行预测;最后将两部分得到的预测值求和得到风速序列的预测值。选定 3 个不同地点分别进行仿真实验并与改进前方法进行比较,结果表明改进后模型的预测精度更高,误差更小。

[关键词] 海上风速预测;差分步长;季节性差分自回归移动平均

[中图分类号] TP18 [文献标识码] A

风速不仅决定了船只的航行路线,还对船只的航行安全产生很大的影响。如果船只能够在出海前比较精确的掌握未来一段时间相应海域的风速情况,那将对船只的航行线路规划产生积极影响。

针对近海海域风速变化的特点,风速预报大多在一些临近海边的风场中。如张增海等^[1]通过地表的粗糙度指数和大气的稳定度给出了相应的海上风速和沿海风速观测站风速关系的经验公式,此经验公式适用于该观测站附近的海域风速预测。考虑到风速序列中既有线性的影响又有非线性因素的影响和单一预测模型自身的局限性,研究者们又提出一系列组合预测的风速预测模型^[2-4]。如田中大等^[5]将 ARIMA^[6]和回声状态网络(Echo State Network, ESN)相结合、李蓉蓉等^[7]将时间序列分析和 LSTM 相结合、高桂革等^[8]将经验模态分解和极限学习机相结合、王耀庆等^[9]将小波变换与 LSTM 相结合,他们的做法都是将原始风速序列分解为线性自相关和非线性残差两部分,使用组合模型分别对两部分进行建模预测,充分利用好每一个模型优势,提高风速预测^[10]精度。

现有研究使用季节性差分移动自回归平均模型(Seasonal Autoregressive Integrated Moving Average Model, SARIMA)进行时间序列预测时,对于季节性参数的判定大都是通过人为估计,存在一定的误差。本文提出将 STL^[11](Seasonal-Trend de-

composition procedure based on Loess)方法用于 SARIMA 模型季节性差分步长的判定,以提高 SARIMA 的预测精度,进而提升和长短时记忆神经网络模型(Long Short-term Memory, LSTM)相结合后对海上风速预测的准确性。实验结果表明,通过对 SARIMA 模型的改进,可以有效提高海上风速预测的精度。

1 基础算法

1.1 季节性差分自回归移动平均

在众多时间序列中,由于月度、季度等因素影响,如某景点的旅游人次数据,某些序列常常呈现出一种周期性变化,这类序列统称为季节性序列,同时也衍生出了季节性 ARIMA 模型,用 SARIMA 表示,它对数据变量建立序列回归,并根据数据周期项和随机项对序列未来趋势做测算。SARIMA 模型^[12]源于 ARIMA 模型。将原始的时间序列记为 y_t , SARIMA 模型首先是对 y_t 进行差分处理,消除序列当中的趋势性,然后通过季节性差分消除季节性,经过处理后,模型可以表示为 $SARIMA(p, d, q)(P, D, Q)_s$, 记作^[10]:

$$\varphi_p(B^s) \varphi_P(B) (1 - B^s)^D (1 - B)^d y_t = \Theta_Q(B^s) \theta_q(B) a_t \quad (1)$$

其中: $\varphi_p(B^s)$ 为季节 P 阶自回归算子多项式、 p 为自回归阶数、 P 为季节性自回归阶数、 $\varphi_P(B)$ 为非季

[收稿日期] 2021-07-03

[基金项目] 国家重点研发计划(2017YFC1405403);国家自然科学基金(61075059);湖北工业大学绿色工业科技引领计划(CPYF2017008)

[第一作者] 余聪聪(1995-),男,湖北黄冈人,湖北工业大学硕士研究生,研究方向为人工智能,计算机仿真

[通信作者] 熊才权(1966-),男,湖北鄂州人,工学博士,湖北工业大学教授,研究方向为人工智能,辩论模型

节自回归多项式、 D 为季节差分阶数、 d 为逐期差分阶数、 s 为季节差分步长、 $(1-B^s)^D$ 为季节差分算子、 $(1-B)^d$ 为差分算子、 Q 为季节移动平均阶数、 $\Theta_Q(B^s)$ 为季节 Q 阶移动平均算子多项式、 $\theta_q(B)$ 为非季节移动平均多项式、 a_t 为白噪声序列。

1.2 循环神经网络

典型循环神经网络结构一般如图 1 所示,主体结构的输入包括输入层 x_t ,循环边上所提供的上一时刻隐藏状态 s_{t-1} 。在某一个时刻 t ,网络在读取了 x_t 和 s_{t-1} 之后还会生成新的隐藏状态 s_t 和产生当前时刻的输出 o_t 。

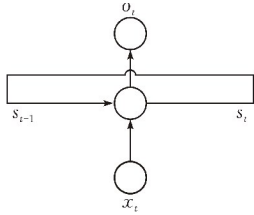


图 1 循环神经网络结构图

从理论上来说,循环神经网络可以处理任意长度的序列,但在实际的训练过程中,当序列过长时,一方面可能会导致出现梯度消失和梯度爆炸的问题,另一方面网络展开后会占用过大的内存,因此在实际使用时会规定一个序列的最大长度,当序列的长度超过这个长度时,应该进行分段处理。

1.3 LSTM 神经网络结构

循环神经网络与传统的神经网络相比可以通过历史保存的信息来辅助当前的决策,但在某些问题中,模型只需要短期的信息来处理当前任务,因此循环神经网络存在长期依赖问题。长短时记忆神经网络是一种特殊的循环神经网络,与传统的循环神经网络相比,LSTM 在其基础上添加了一些“门”结构,它可以选择性保留网络信息,有效地解决了一些无效数据的依赖问题,提高了神经网络的效率,LSTM 单元结构如图 2 所示。

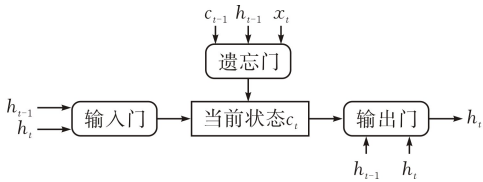


图 2 LSTM 网络结构图

图 2 中的“遗忘门”和“输入门”是 LSTM 结构中最为核心的部分。其中“遗忘门”会依据当前输入的 x_t 和上一时刻输出的 h_{t-1} 来决定遗忘哪一部分记忆,而“输入门”依据 x_t 和 h_{t-1} 决定哪些信息应该加入到状态 c_{t-1} 中从而生成新的状态 c_t 。此时“输出门”会根据当前时刻的输入 x_t ,上一时刻的输出 h_{t-1} 和 c_t 来决定此刻的输出。

2 STL 对 SARIMA 模型的改进

STL 是一种通用稳健基于 Loess 的分解时间序列方法。估计某个响应变量值时,优先选择预测变量附近的一个数据子集,通过采用二次回归或加权最小二乘法进行线性回归,使离该响应变量较远点的权重变小后通过局部回归模型估算出响应变量的值。它通过提取时间序列中的部分局部数据,从而使得回归曲线变得平滑且让数据在一定的范围内的趋势和变化规律变得更加明显。影响海上风速序列变动的因素包括季节变动、趋势变动和不规则变动,SARIMA 模型季节性差分步长通常是人为估计,会存在误差,使得模型的预测结果不准确。为了能够更好的确定 SARIMA 模型的季节性差分步长,使用 STL 将海上风速序列进行分解,通过分解后的风速季节分量来确定该参数值。分解后的表达式可以表示如下:

$$T_t = S_t + C_t + R_t, (t | 0 \leq t \leq |T|, t \in Z) \quad (2)$$

其中, T 为原始海上风速序列、 S 为季节分量、 C 为趋势分量、 R 为剩余分量、 $|T|$ 为序列的长度。STL 分解过程由内循环和外循环两部分组成。每次内循环都包含季节性平滑,用来更新季节性分量,而外循环则是在内循环完成之后计算稳健的权重,以减少下一次内循环中异常值对更新季节性分量的影响。STL 内循环过程见图 3。步骤分别为:

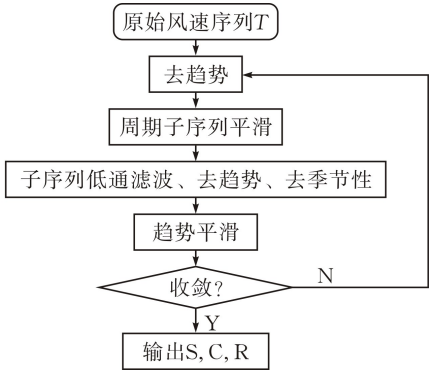


图 3 STL 内循环过程图

- 1)内循环迭代 $n+1$ 次,通过计算原始风速序列 T 与第 n 次迭代中所得到的估计趋势分量 C_t^n 之差来去趋势,即: $T_t - C_t^n = T_t^{\text{detrend}}$;
- 2)使用 Loess 来对周期子序列 T_t^{detrend} 进行平滑处理,得到初步的季节分量 \tilde{N}_t^{n+1} ;
- 3)对 \tilde{N}_t^{n+1} 使用低通滤波器进行处理,然后利用 Loess 得到 L_t^{n+1} ;
- 4)通过计算低通值和季节分量的差值来平滑周期子序列,即: $\tilde{N}_t^{n+1} - L_t^{n+1} = S_t^{n+1}$;
- 5)使用原始序列 T 减掉季节分量 S^{n+1} 以去季

节性,即: $T_t^{\text{deseason}} = T_t - S_t^{n+1}$;

6)利用 Loess 对 T_t^{deseason} 平滑后获得趋势分量 C_t^{n+1} ;

在外循环中,使用内循环得到的趋势分量 C 和季节分量 S 来计算剩余分量 S 。分析季节分量 S 得到 SARIMA 模型的季节性差分步长。

3 改进后的 SARIMA 与 LSTM 组合模型的建立

海上风速序列不仅具有线性特征,还具有非线性、随机性和非平稳性等特征。改进后的 SARIMA 模型对风速序列中的线性部分拟合得更好,而长短时记忆网络模型(LSTM)的优势在于拟合较为复杂的非线性,非平稳性数据,二者的优势互补。假设存在风速时间序列 Y_t 由两部分组成,分别为线性自相关的 L_t 和非线性的残差 N_t ,则有: $Y_t = L_t + N_t$, 本文将采用改进后的 SARIMA 和 LSTM 的组合模型进行风速预测。

首先对海上风速序列进行 STL 分解后确定 SARIMA 模型的季节性差分步长,模型其他参数通过网格搜索的方法确定,然后使用得到的 SARIMA 模型对风速序列的线性部分进行建模得到预测结果 \hat{L}_t 和对应的残差 N_t ,其中 N_t 包含风速序列中的非线性关系,其次对得到的 N_t 序列进行重构得到 LSTM 模型的训练样本集,再利用 LSTM 对残差进行预测,得到残差预测结果 \hat{N}_t ,最后将线性预测的结果和非线性残差预测的结果组合,得到最终的风速预测结果 $\hat{Y}_t = \hat{L}_t + \hat{N}_t$ 。组合预测原理见图 4。

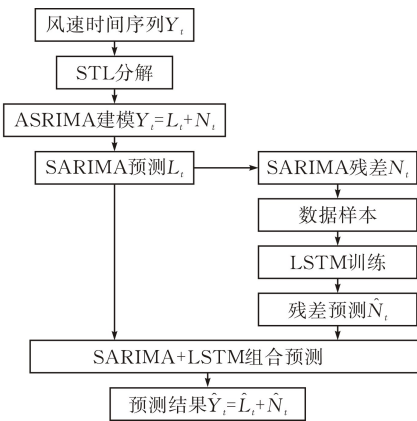


图 4 SARIMA 与 LSTM 组合预测原理图

4 实验验证与结果分析

4.1 数据来源及处理

实验数据的提供方是中国科学院南海海洋研究所,选取我国南海 17.5°N/110.5°E、17.8°N/110.7°E、18°N/111°E 分别表示为 A、B、C 的三个点从 2020 年 4 月 20 日中午 12 时至 2020 年 4 月 25 日中午 12 时为时长 5 d 的风速数据分别进行实验,前 4 d 数据作为模型的训练数据,最后一天数据作为测试数据。

4.2 SARIMA 模型的构建和检验

以 A 点为例,首先对风速时间序列进行 STL 分解,结果见图 5。

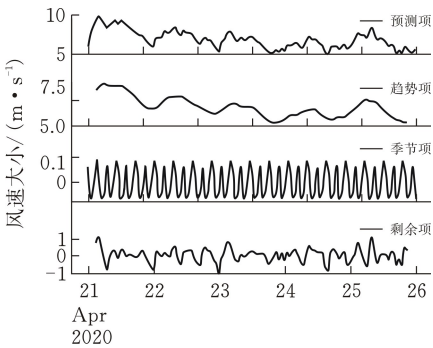


图 5 风速时间序列分解示意图

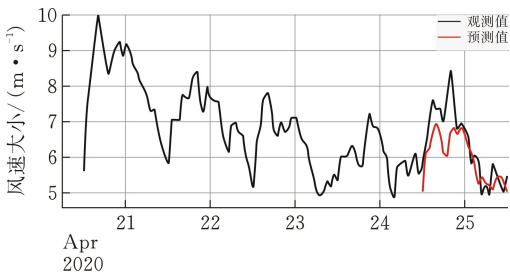
从图 5 中可以看出观测的风速序列整体在前 4 d 存在明显的下降趋势,最后一天稍微有所上升,每一天内都呈现出先上升后下降的趋势。由于实测的数据是以小时为单位进行观测得到,因此确定 SARIMA(p,d,q)(P,D,Q) s 模型中的季节性差分步长 s 的值为 24。

模型的参数确定通过网格搜索的方式进行,参数搜索范围确定为 0~2,确定了模型的参数后,需要对模型进行检验,主要是进行模型的显著性检验和参数的显著性检验。模型参数及其显著性检验信息见表 1,检验方法采用 Z 值检验。结果表明该模型参数均显著非零,模型参数均通过检验。

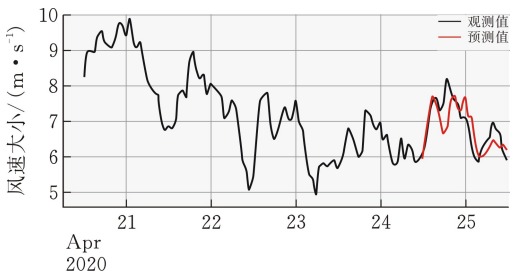
采用相同的方式对 B、C 两点的风速数据进行同样的建模处理,最后使用得到的模型对观测到的风速数据进行提前一天的预测,得到的风速实际预测值和风速观测值对比结果见图 6。

表 1 模型参数与参数显著性检验

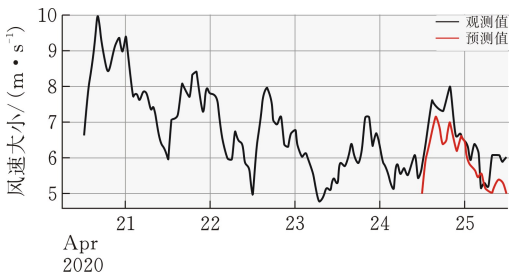
参数	系数值	标准误差	Z 值	P 值	95%置信下限	95 置信上限
ar.L1	-0.583	0.322	-1.814	0.07	-1.214	0.047
ma.L1	0.8159	0.235	3.472	0.001	0.355	1.277
ar.S.L24	-0.526	0.133	-3.945	0	-0.787	-0.265
sigma2	0.1558	0.037	4.199	0	-0.083	0.229



(a) A 点



(b) B 点



(c) C 点

图 6 预测结果与真实值对比

观察以上三个地点风速的观测值和预测值的结果曲线,预测值曲线的整体波动趋势与观测值曲线的波动趋势基本一致,但在风速波动较大时,预测精度还有待提升。

4.3 SARIMA-LSTM 模型风速预测

在 SARIMA-LSTM 组合模型中,先使用风速的观测值进行 SARIMA 模型建模,将模型得到的残差作为 LSTM 模型的输入。为了实现对模型残差进行准确的预测,设计了图 7 所示的残差预测 LSTM 网络结构。

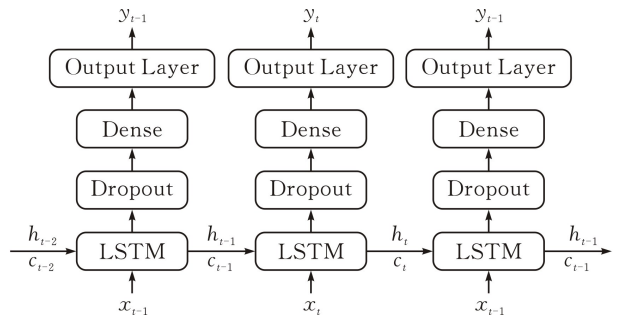


图 7 残差 LSTM 网络结构

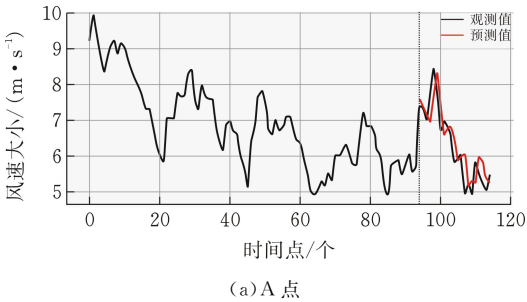
设置 LSTM 网络的输入维度和输出维度都是 1 维,隐藏层神经元节点的个数为 120,损失函数设置为均方误差函数(MSE),优化器选择 Adam,对模型进行训练。网络的输出信息见图 8。加入 LSTM 网络后,组合风速预测模型在 A、B、C 三点的预测效果见图 9。

Model:"sequential_1"

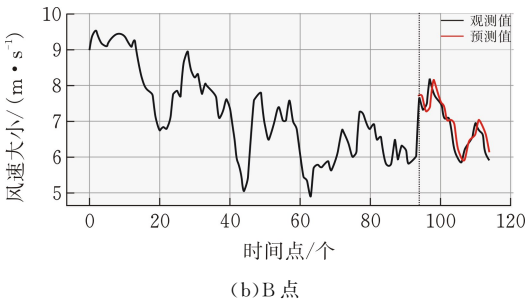
Layer (type)	output Shape	Param #
lstm_1 (LSTM)	(None,120)	58560
dropout_1 (Dropout)	(None,120)	0
Dense_1 (Dense)	(None,4)	484
Dense_2 (Dense)	(None,1)	5

Total params:59,049
Trainable params:59,049
Non-trainable params:0

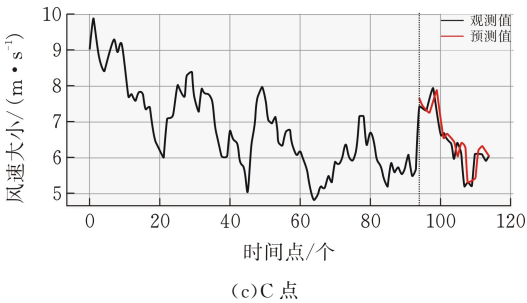
图 8 LSTM 网络信息输出



(a) A 点



(b) B 点



(c) C 点

图 9 风速真实值与预测值对比

从实验结果可以看出,在 SARIMA 模型的预测基础之上增加长短时记忆网络后的预测结果值与原始的风速值更为接近,预测曲线的变化趋势与实际观测风速曲线的变化趋势也基本一致。

4.4 SARIMA-LSTM 模型风速预测

为了验证改进后组合模型在风速预测中的有效性,在相同的实验条件下,分别使用 SARIMA 模型、BP 网络模型、LSTM 网络模型、LSTM-SARIMA 组合模型和本文的 STL-SARIMA-LSTM 组合模型进行实验对比。预测误差选取 3 个不同地点平均绝对误差 MAE、均方根误差 RMSE 和平均绝对百分比误差 MAPE 的均值作为评价标准。定义见公式(3)~(5)。

MAE = \frac{1}{N} \sum_{t=1}^N |x_t - \hat{x}_t| \tag{3}

RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (x_t - \hat{x}_t)^2} \tag{4}

MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{x_t - \hat{x}_t}{x_t} \right| \times 100\% \tag{5}

其中 x_t 表示 t 时刻的风速观测值, \hat{x}_t 为 t 时刻的风速预测值。预测结果对比见表 2。

表 2 预测结果对比

预测模型	RMSE/ (m · s ⁻¹)	MAE/ (m · s ⁻¹)	MAPE/%
BP	0.6035	0.4635	7.188
LSTM	0.5609	0.4747	7.203
SARIMA	0.5546	0.4303	6.47
SARIMA-LSTM	0.4563	0.3257	5.52
STL-SARIMA-LSTM	0.3714	0.2706	4.62

由表 2 可知:单一预测模型中,SARIMA 模型预测效果较好,LSTM 神经网络模型和 BP 神经网络模型的预测精度相差不大。由于组合模型对海上风速数据特征的提取更加充分,预测精度高于单一预测模型。使用 LSTM 和 SARIMA 组合时预测误差较小,而当在 SARIMA 模型季节性差分参数确定时,考虑结合 STL 后,再结合 LSTM 进行海上风速的预测时,预测风速曲线与实际风速曲线最为接近,预测精度最高。

5 结论

在 STL 分解海上风速序列后,SARIMA 模型的季节性差分步长确定变得更为准确,从而使其预测精度得到了提升。从一系列的对比实验中可以看出改进后的 SARIMA 与 LSTM 组合后对于海上风速预测的精度更高。加入 STL 后模型变得更为复杂,导致数据处理的时间变长且由于海上风速的不稳定性,在风速波动较大点预测精度还不够好。后续研究应进一步优化模型并考虑诸如气压、温度等

额外因素对于海上风速大小的影响。

[参 考 文 献]

[1] 张增海, 曹越男, 赵伟. 渤海湾海域风况特征分析与海-陆风速对比分析[J]. 海洋预报, 2011, 28(6): 33-39.

[2] 林涛, 刘航鹏, 赵参参, 等. 基于 SSA-PSO-ANFIS 的短期风速预测研究[J]. 太阳能学报, 2021, 42(3): 128-134.

[3] 苏盈盈, 李翠英, 王晓峰, 等. 风电场短期风速的 C-C 和 ELM 快速预测方法[J]. 电力系统及其自动化学报, 2019, 31(07): 76-80+87.

[4] 姚万业, 黄璞, 姚吉行, 等. 一种基于深度学习的 FRS-CLSTM 风速预测模型[J]. 太阳能学报, 2020, 41(9): 324-330.

[5] 田中大, 李树江, 王艳红, 等. 基于 ARIMA 与 ESN 的短期风速混合预测模型[J]. 太阳能学报, 2016, 37(6): 1603-1610

[6] NAYEREH ESMAEILZADEH, MOHAMMAD-TAGHI SHAKERI, MOSTAFA ESMAEILZADEH, et al. Arima models forecasting the SARS-COV-2 in the islamic republic of iran[J]. Asian Pacific Journal of Tropical Medicine, 2020, 13(11): 521-524.

[7] 李蓉蓉, 戴永. 基于 LSTM 和时间序列分析法的短期风速预测[J]. 计算机仿真, 2020, 37(3): 393-398.

[8] 高桂革, 原阔, 曾宪文, 等. 基于改进 CEEMD-CS-ELM 的短期风速预测[J]. 太阳能学报, 2021, 42(7): 284-289.

[9] 王耀庆, 孙建平, 李冰, 等. 基于小波变换和 LSTM 的短期风速预测研究[J]. 计算机仿真, 2021, 38(2): 438-443.

[10] LIU LING, JI TIANYAO, LI MENGSHI, et al. Short-term local prediction of wind speed and wind power based on singular spectrum analysis and locality-sensitive hashing[J]. Journal of Modern Power Systems and Clean Energy, 2018, 6(2): 317-329.

[11] ROJO J, RIVERO R, ROMERO-MORTE J, et al. Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing[J]. International journal of biometeorology, 2017, 61(2): 335-348.

[12] 杨振昊, 张俊波, 杨晨星, 等. 基于 SARIMA 模型的我国水产品消费价格指数预测[J]. 海洋湖沼通报, 2021, 43(2): 131-138.

(下转第 53 页)