文章编号：2095-2163(2023)12-0165-05 中图分类号：TP391 文献标志码：A

基于 SIDW-SSA-LSTM 的门诊量预测

樊 冲
(锦州市大数据管理中心，辽宁 锦州 121000)

摘 要：医院门诊量本质上是一种具有潜在规律的时间序列，通过对门诊量进行有效分析和预测，可以更加科学、合理地配置医疗资源。针对门诊量波动幅度较大的时间序列预测问题，提出 SIDW-SSA-LSTM 模型。首先，通过标幺化反距离加权(SIDW)插值修正原始数据，提高了门诊量数据集的可靠性；然后，采用在时序问题处理上具有良好性能的长短期记忆(LSTM)神经网络，并通过寻优能力强、稳定性好的麻雀搜索算法(SSA)对 LSTM 网络超参数进行优化，得到 SIDW-SSA-LSTM 模型。对比实验证明，本文提出的方法可以更加精准地对门诊量进行预测和分析，为医院更好地运营管理提供了重要依据和决策支持。

关键词：门诊量；麻雀搜索算法；LSTM；标幺化反距离加权

Outpatient volume prediction based on SIDW-SSA-LSTM

FAN Chong
(Jinzhou Big Data Management Center, Jinzhou Liaoning 121000, China)

Abstract: Hospital outpatient volume is essentially a time series with potential laws. Through effective analysis and prediction of outpatient volume, medical resources can be more scientifically and reasonably allocated. The SIDW-SSA-LSTM model is proposed to predict the time series with large fluctuation of outpatient volume. First, the reliability of the outpatient volume data set is improved by modifying the original data with standardized inverse distance weighting (SIDW) interpolation. Then, the long and short term memory (LSTM) neural network with good performance in time series problem processing is adopted, and the super parameters of LSTM network are optimized by Sparrow Search Algorithm (SSA) with strong optimization ability and good stability, and the SIDW-SSA-LSTM model is obtained. Finally, through comparative experiments, the proposed method can more accurately predict and analyze the outpatient volume, providing an important basis and decision support for better operation and management of the hospital.

Key words: outpatient volume; Sparrow search algorithm; LSTM; standardized inverse distance weighting

0 引 言

门诊工作是现代医疗工作里非常重要的一环，日常的门诊量也反映着医院实时的运行状态，准确地对医院门诊量做到有效预测，既能为医院管理人员进行资源合理配置提供重要保障，也能为医院的运营管理起到积极的作用[1-5]。

门诊量预测本质上是一种时间序列预测，文献[6]采用灰色理论和径向基函数构建组合预测模型，开展对医院门诊量预测；文献[7]在进行门诊量预测过程中，引入了移动平均季度指数法；文献[8]依据最小二乘法和变动系数，对医院月门诊量开展预测；文献[9]分析了门诊量的影响因素，并且依据测算获取预测结果；文献[10]采用深度信念网络构建预测模型，深入挖掘和分析医院各科室门诊量的数据特征，并经无监督学习实现门诊量的预测；文献[11]应用时间序列分解法构建门诊量预测模型，实现月门诊量的预测；文献[12-15]采用灰色系统构建门诊量预测模型，实现了门诊量的预测。

虽然上述研究构建了不同的预测模型，实现了医院门诊量的预测，但现有的门诊量预测模型性能与预测精度均有待提升。针对这一问题，本文从时序角度出发，采用人工智能方法对门诊量进行预测。通过标幺化反距离加权(Standardized Inverse Distance Weighting，SIDW)插值，对门诊量数据进行修正处理，再采用长短期记忆(Long-Short Term Memory，LSTM)神经网络模型充分利用门诊量数据间时序相关性进行预测，并通过麻雀搜索算法

作者简介：樊 冲(1981-)，男，硕士，高级工程师，主要研究方向：数据分析、虚拟化。Email：dillon_fc@126.com

收稿日期：2023-09-13

(Sparrow Search Algorithm,SSA)对网络参数进行优化,有效提升了门诊量预测模型性能,获取的门诊量预测结果为医院的医疗资源合理配置提供了更有效的参考依据。

1 相关技术

1.1 长短期记忆神经网络(LSTM)

由于循环神经网络在反向传播时,不能学习到序列中早期的数据信息,并且随着时间序列的长度逐渐增加,神经网络会遇到梯度消失的问题。针对于此,长短时记忆(Long-Short-Term Memory, LSTM)神经网络可有效解决上述局限[16]。关于门诊量预测的时序问题,LSTM 能够有效挖掘历史累积门诊量数据时序间耦合性[17-18]。LSTM 结构如图 1 所示。

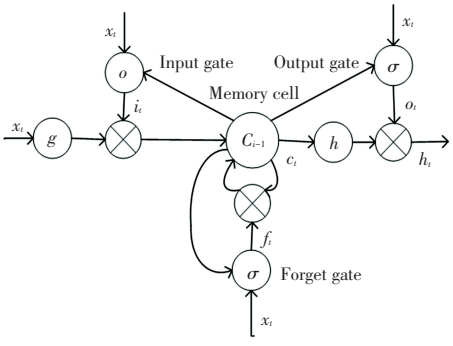

图 1 LSTM 单元结构

Fig. 1 LSTM unit structure

LSTM 网络由输入门、遗忘门和输出门组成,用于将信息更新和删除到存储单元。其中,遗忘门可以决定过去的信息是否从一个单元状态中删除;输入门可以更新信息,输出门决定单元输出。不同门的表达式如下所示:

x_t、h_t 为 t 时刻单元的输入和输出,$t = 1,2,\cdots$。h_t 由下列公式计算得出:

$$i_t = \sigma(\boldsymbol{W}_{ix}x_t + \boldsymbol{W}_{ih}h_{t-1} + \boldsymbol{W}_{ic}c_{t-1} + \boldsymbol{b}_i) \quad (1)$$

$$f_t = \sigma(\boldsymbol{W}_{fx}x_t + \boldsymbol{W}_{fh}h_{t-1} + \boldsymbol{W}_{fc}c_{t-1} + \boldsymbol{b}_f) \quad (2)$$

$$o_t = \sigma(\boldsymbol{W}_{ox}x_t + \boldsymbol{W}_{oh}h_{t-1} + \boldsymbol{W}_{oc}c_{t-1} + \boldsymbol{b}_o) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t\varphi(\boldsymbol{W}_{cx}x_t + \boldsymbol{W}_{ch}h_{t-1} + \boldsymbol{b}_c) \quad (4)$$

$$h_t = o_t\varphi(c_t) \quad (5)$$

式中:i_t 表示输入门输出,f_t 表示遗忘门输出,o_t 为记忆门的输出,c_t 表示输出门输出,x_t、h_t 为 LSTM 网络的输入、输出,$\boldsymbol{W}$ 为权重矩阵,$\boldsymbol{b}$ 为偏置向量,σ 为激活函数。

1.2 麻雀搜索算法(SSA)

SSA 模拟麻雀觅食和反捕食行为,具有寻优能力强的优势[19]。在 $t+1$ 次迭代中,发现者位置更新为:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \cdot \exp(-\dfrac{i}{a \cdot T}), & R_2 < S \\ X_{i,j}^t + Q \cdot \boldsymbol{L}, & R_2 \geqslant S \end{cases} \quad (6)$$

式中:t、T 分别为迭代和最大迭代次数,$X_{i,j}^t$ 为位置,a、Q 为随机数,R_2 为预警值,且 $R_2 \in [0,1]$,S 为安全值,且 $S_2 \in [0.5,1.0]$,$\boldsymbol{L}$ 为矩阵。

如果 $R_2 < S$ 时,在此场景下发现者开展搜索;否则,将加入者位置更新为:

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp(\dfrac{X_{\text{worst}} - X_{i,j}^t}{t^2}), & i > n/2 \\ X_p^{t+1} + \left| X_{i,j}^t - X_p^{t+1} \right| \cdot \boldsymbol{A}^+ \cdot L, & \text{otherwise} \end{cases} \quad (7)$$

式中:X_P、X_{worst} 分别为最优、最劣位置,$\boldsymbol{A}$ 为矩阵,且 $\boldsymbol{A}^+ = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$。

若 $i > n/2$ 时,第 i 个加入者无法觅食。当危险被警戒者感知时,反捕食行为产生。表达式如下:

$$X_{i,j}^{t+1} = \begin{cases} X_{\text{best}}^t + \beta \cdot \left| X_{i,j}^t - X_{\text{best}}^t \right|, & f_i > f_g \\ X_{i,j}^t + k \cdot \left(\dfrac{\left| X_{i,j}^t - X_{\text{best}}^t \right|}{(f_i - f_w) + \varepsilon} \right), & f_i = f_g \end{cases} \quad (8)$$

式中:X_{best} 为全局最优位置,k、β 是步长控制参数,β、k 为随机数,f_g、f_w 为最优、最劣适应度值,ε 为常数,f_i 为警戒者 i 的适应度值。

2 预测模型

2.1 模型原理

门诊量历史累计数据由于人为等因素影响,会产生大量异常数据。这些异常值降低了门诊量数据的时序相关性,严重影响预测模型的精度。为了提高门诊量数据集的可靠性,采用将计算出的标幺值与地理科学领域的反距离加权法相结合的标幺化反距离加权插值(SIDW)先对原始数据进行修正,提升数据的有效性;之后选择具有强泛化性的 LSTM 作为门诊量的预测模型。预测模型内,学习率以及神经元个数等超参数是影响门诊量预测精度的重要参数,如果模型的超参数选取不当,将无法有效获取最优的门诊量预测模型,且预测结果精度低。为了确定 LSTM 门诊量预测模型的最优参数,采用 SSA 算法对其超参数开展寻优,从而构建了基于 SIDW-SSA-LSTM 的门诊量预测模型。模型流程如图 2 所示。

2.2 超参数优化过程

(1)依据 LSTM 门诊量预测模型的超参数取值范围,初始化麻雀位置;

(2)采用 MSE 指标,获取每只麻雀适应度值,并依据获取的适应度结果进行排序;

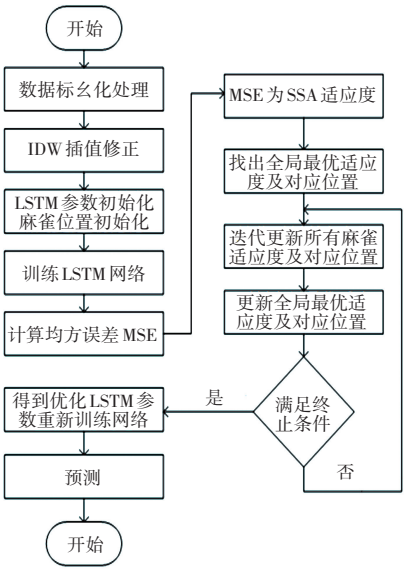

图 2　SIDW-SSA-LSTM 流程

Fig. 2　SIDW-SSA-LSTM process

(3)取前 20%为发现者,其余为加入者,随机选取 20%的麻雀并使其具有感知危险的行为机制;

(4)根据 SSA 算法公式在迭代中更新发现者、加入者、警戒者位置并计算适应度值;

(5)当 SSA 寻优过程结束,得到全局最优麻雀位置,由此确定 LSTM 门诊量预测模型的最优参数。

3　实验设计与结果分析

3.1　数据采集

本文门诊量预测研究所需的数据来自某医院 2020 年全年门诊量的实测数据,数据的分辨率为 15 min。进行门诊量预测时,将待预测日前五天的门诊量作为预测模型的输入。同时,将门诊量实测数据集按照 4∶1 的比例分为训练集和测试集。

3.2　数据预处理

通过标幺化处理,门诊量数据归算到 0~1 之间,体现了不同时间门诊量的对比情况,便于进行插值修正。在标幺化处理后,再进行反距离加权插值。假设 t 时刻日期 j 门诊量数据异常,则修正方法如下:

$$\lambda_{ij}=\frac{\frac{1}{d_{ij}}}{\sum_{i=1,\,i\neq j}^{n}\frac{1}{d_{ij}}} \tag{9}$$

$$P_j=\sum_{i=1,\,i\neq j}^{n}\lambda_{ij}P_i \tag{10}$$

式中:d_{ij} 为日期 i 到 j 的距离,λ_{ij} 为 P_i 的插值权重,P_i 为 t 时刻日期 i 标幺门诊量,P_j 为 t 时刻日期 j 标幺门诊量修正值。

通过 SIDW 插值修正门诊量数据集,既简化了训练 LSTM 网络的归一化步骤,又提高了门诊量数据的时序相关性,便于 LSTM 网络充分发挥其长时记忆功能。

3.3　模型参数设置与寻优

设置合理的模型参数,对于提升门诊量预测模型的性能至关重要。SSA 算法参数设置见表 1,LSTM 模型参数设置见表 2。

表 1　SSA 算法参数

Table 1　SSA algorithm parameters

参 数	参 数 值
麻雀种群数	5
发现者占比	20%
警戒者占比	20%
最大迭代次数	20
适应度函数	MSE
第一层神经元寻优维度	[1,100]
第二层神经元寻优维度	[1,100]
训练次数寻优维度	[1,50]
学习率寻优维度	[0.001,0.01]

表 2　LSTM 模型参数

Table 2　LSTM model parameters

超 参 数	参 数 值
第一层神经元	200
第二层神经元	200
训练次数	20
学习率	0.005

将 SIDW 修正后的数据集前 80%用于训练,后 20%用于测试,然后使用 MATLAB 进行仿真。首先通过 SSA 对 LSTM 网络参数进行寻优,隐含层第一层神经元个数、隐含层第二层神经元个数、训练次数、学习率迭代过程如图 3 所示。

在 SSA 寻优过程中以 MSE 为适应度函数,目的是找到一组超参数使 LSTM 网络的 MSE 最低,适应度值变化如图 4 所示。

LSTM 网络参数与经过 20 轮寻优后的 SIDW-SSA-LSTM 网络参数,见表 3。

表 3　网络参数对比

Table 3　Network parameter comparison

	第一层神经元数	第二层神经元数	训练次数	学习率
LSTM	200	200	20	0.005 0
SIDW-SSA-LSTM	74	46	43	0.009 5

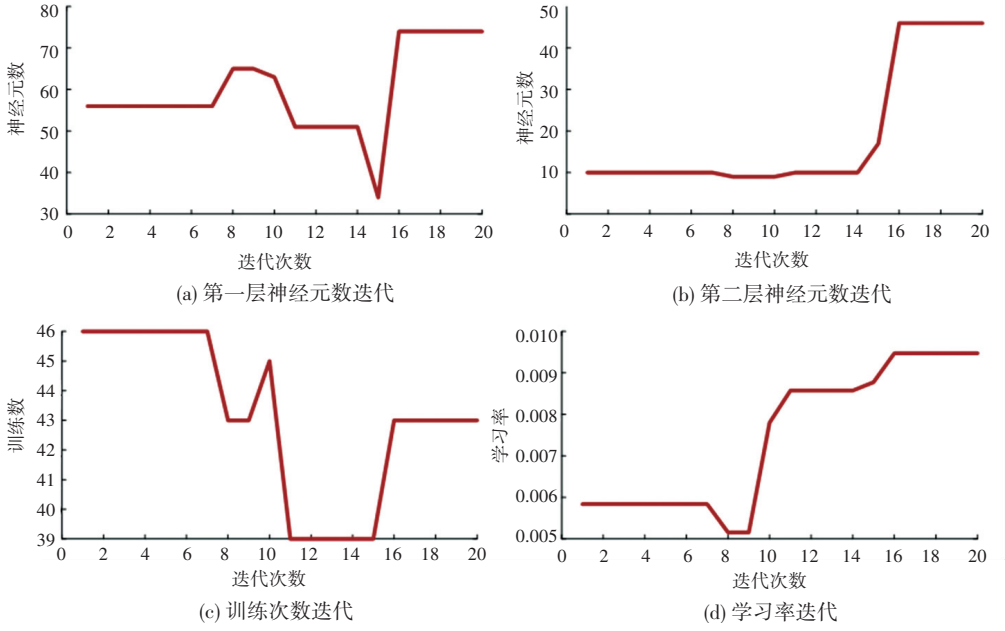

图 3 LSTM 超参数迭代过程

Fig. 3 LSTM hyperparameter iteration procedure

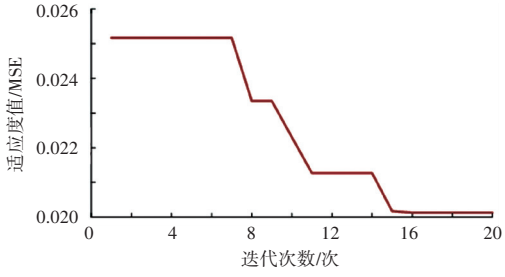

图 4 SSA 适应度值迭代过程

Fig. 4 SSA fitness value iteration process

3.4 对比分析

为了验证本文方法的有效性,将其与 LSTM 模型、SIDW-LSTM 模型进行对比实验研究。采用 3 个预测模型分别进行门诊量预测后,从 4 个季节中各选择一个典型日展示各模型的门诊量预测结果,其结果如图 5 所示。从图 5 中展示的结果可以看出,在 4 个季节中,本文提出方法获取的门诊量预测结果与实际值更贴合,并且预测效果优于其它预测模型,表明了所提方法的有效性。

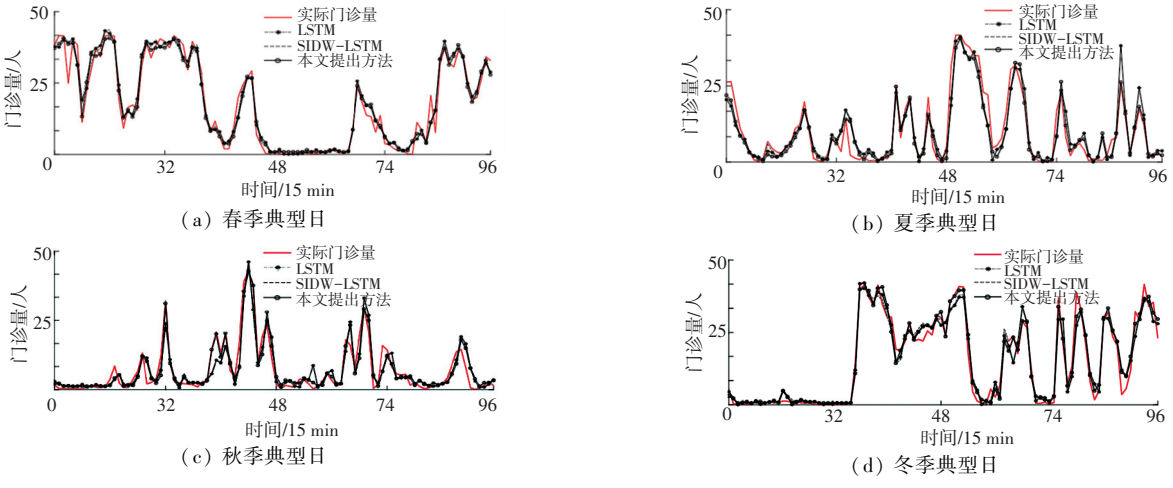

图 5 各模型门诊量预测结果对比图

Fig. 5 Comparison of outpatient volume prediction results of each model

为了进一步验证所提预测模型的性能优势,以及量化对比 LSTM、SIDW-LSTM 模型和提出方法模型的性能,本研究采用绝对值平均误差(Mean Absolute Percentage Error,MAPE)[20-22]、相对平均绝对误差(relative Mean Absolute Error, rMAE)[23-25]、相对均方根误差(relative Root Mean Square Error, rRMSE)[26-28]3 个预测精度指标值进行评估,各指标计算公式如下所示:

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\frac{|f_i - f_i'|}{f_i} \tag{11}$$

$$rRMSE = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(f_i - f_i')^2}}{\frac{1}{m}\sum_{i=1}^{m} f_i} \times 100\% \quad (12)$$

$$rMAE = \frac{\sum_{i=1}^{m}|f_i - f_i'|}{\sum_{i=1}^{m} f_i} \times 100\% \quad (13)$$

其中，f_i、f_i' 分别为门诊量的真实值与预测值。

各季节典型日门诊量预测结果的 3 个指标值见表 4。从表 4 可以看出，对比模型中 3 个评价指标的最优值分别为 3.54 MW、4.65%、2.12%；本文所提方法的 3 个评价指标的最优值分别为 2.81 MW、2.98%、1.89%。预测结果显示本文方法相较对比模型数值更低，精准度更高，有效说明了本文方法较强的泛化性，有效的克服门诊量强随机性对预测精度的影响，实验结果证明本文方法更加高效和准确。

表 4　各模型门诊量预测结果评价指标值

Table 4　Evaluation index value of outpatient volume prediction results of each model

季节	模型	评价指标		
		rRMSE/MW	MAPE/%	rMAE/%
春季	LSTM	5.43	5.46	2.12
	SIDW-LSTM	6.49	6.89	2.56
	本文方法	3.16	3.57	1.89
夏季	LSTM	5.46	6.04	5.68
	SIDW-LSTM	3.54	4.65	3.95
	本文方法	3.23	3.23	2.16
秋季	LSTM	7.21	7.57	5.24
	SIDW-LSTM	6.25	5.84	4.26
	本文方法	5.19	4.12	3.14
冬季	LSTM	5.93	4.95	3.84
	SIDW-LSTM	4.99	5.12	4.58
	本文方法	2.81	2.98	2.17

4　结束语

针对医院门诊量预测的强非线性和复杂性问题，提出了基于 SIDW-SSA-LSTM 的医院门诊量预测模型。经对比验证分析，验证了本文提出方法获取的门诊量预测结果精度更高。本研究构建的门诊量预测模型，可以对医院运营管理提供科学的理论支撑和依据，同时也能够为医院管理者提供有效的决策支持。

参考文献

[1] 桑泉红，徐培文. 基于时序序列模型预测医院门诊人次[J]. 中国医院统计，2022，29(1)：25-28.

[2] 刘焰，卢萍萍. 基于移动平均季节指数法的门诊量分析及预测[J]. 医学信息，2021，34(23)：156-158.

[3] 吴磊，徐凯. 基于深度神经网络的医院门诊量预测[J]. 微型电脑应用，2021，37(7)：108-110，130.

[4] 唐路，宋萍，谢冰珏，等. 基于 R 语言的 ARIMA 乘积季节模型对重庆某儿童医院门诊量的预测分析[J]. 医学信息，2021，34(11)：19-22.

[5] 焦晨，黄艳然，赵钦风，等. 时间序列分解模型在山东省糖尿病门诊量预测中的应用[J]. 中国农村卫生事业管理，2021，41(2)：93-97.

[6] 张筠莉，杨祯山. 现代医院门诊量的灰色 RBF 神经网络预测[J]. 计算机工程与应用，2010，46(29)：225-228.

[7] 陈辉，周雄辉，朱燕，等. 移动平均季节指数法在预测门诊量和出院人数中的运用[J]. 中国卫生统计，2012，29(2)：312.

[8] 胡蓉. 某院门诊量动态分析及预测[J]. 中国卫生事业管理，2010，27(S1)：39-41.

[9] 许崇伟，沈俊学，邓光璞，等. 医院门诊量影响因素及预测方法[J]. 中国卫生经济，2015，34(3)：74-75.

[10] 杨旭华，钟楠祎. 基于深度信念网络的医院门诊量预测[J]. 计算机科学，2016，43(S2)：26-30.

[11] 黄美林，刘世科，胡丹标，等. 时间序列分解法在预防接种门诊接种量预测的应用[J]. 中国疫苗和免疫，2017，23(6)：681-684.

[12] 王琦，郑静，吴清香，等. 灰色 GM(1，1)预测模型在门诊量预测中的应用[J]. 中国医院管理，2007，27(2)：26-27.

[13] 张珊珊，尚莉丽. 基于灰色系统理论的中医医院门诊工作量分析及预测研究[J]. 合肥学院学报(自然科学版)，2013，23(4)：24-28.

[14] 马春柳，刘海霞，李小升，等. 灰色预测模型 GM(1，1)在医院门诊量预测中的应用[J]. 中国病案，2012，13(12)：23-25.

[15] 孔超. 基于灰色预测模型的门诊量预测——以上海市浦东新区门诊总量为例[J]. 中国卫生资源，2008，11(6)：267-268，277.

[16] 王鑫，吴际，刘超，等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报，2018，44(4)：772-784.

[17] 朱乔木，李弘毅，王子琪，等. 基于长短期记忆网络的风电场发电功率超短期预测[J]. 电网技术，2017，41(12)：3797-3802.

[18] 张群，唐振浩，王恭，等. 基于长短时记忆网络的超短期风功率预测模型[J]. 太阳能学报，2021，42(10)：275-281.

[19] 吕鑫，慕晓冬，张钧，等. 混沌麻雀搜索优化算法[J]. 北京航空航天大学学报，2021，47(8)：1712-1720.

[20] 张帅可，罗萍萍. 基于混合分布模型的风电功率超短期预测误差分析[J]. 电力科学与技术学报，2020，35(5)：111-118.

[21] 杨茂，杨春霖，董骏城. 基于预测误差分布优化模型的风电功率超短期概率区间预测研究[J]. 太阳能学报，2019，40(10)：2967-2978.

[22] 周凡桂，王晓光，高忠信，等. 双目视觉绳系支撑飞行器模型位姿动态测量[J]. 航空学报，2019，40(12)：44-54.

[23] 王芬，马涛. 基于小波神经网络的短时交通流预测[J]. 宁夏师范学院学报，2012，33(6)：60-62，86.

[24] 文莉娟，吕世华. 陆面数据同化方法在绿洲农田土壤温湿度模拟中的应用[J]. 农业工程学报，2010，26(7)：60-65.

[25] 成向荣，黄明斌，邵明安. 基于 SHAW 模型的黄土高原半干旱区农田土壤水分动态模拟[J]. 农业工程学报，2007，23(11)：1-7.

[26] 王建国，陈帅，张超. 噪声参数最优 ELMD 与谱峭度在滚动轴承故障诊断中的应用[J]. 机械传动，2017，41(5)：170-175.

[27] 章颖，梁漫春，黎岢，等. 基于遗传-模拟退火算法的源项反演方法研究[J]. 核电子学与探测技术，2014，34(4)：451-455，473.

[28] 于凤鸣，李喜仓，宋进华，等. 基于中尺度模式与神经网络的风电功率预测[J]. 气象科技，2013，41(4)：784-790.