DOI:10.19652/j.cnki.femt.2305800

基于二次分解和改进沙猫群优化算法的空气质量预测*

朱菊香[1] 张诗云[2] 张 涛[2] 孙君峰[2] 张赵良[1,3]
(1.无锡学院轨道交通学院 无锡 214105;2.南京信息工程大学自动化学院 南京 210000;
3.江苏省工业环境危害要素监测与评估工程研究中心 无锡 214105)

摘 要:准确预测空气质量对人们的日常生活具有重要意义,提出了一种二次分解和改进沙猫群算法(improved sand cat swarm optimization, ISCSO)优化长短期记忆网络(long short-term memory, LSTM)相结合的预测模型。首先,利用完全自适应噪声集合经验模态分解(complete ensemble empirical mode decomposition with adaptive noise, CEEMDAN)算法将 $PM_{2.5}$ 数据分解为多个子序列,对预测效果不满意的重构序列使用变分模态分解(variational mode decomposition, VMD)方法进行二次分解;其次,引入 Cubic 混沌、螺旋搜索策略和麻雀警戒机制改进沙猫群算法,有效提高了算法的全局搜索性能和收敛速度;最后,采用改进的沙猫群算法对 LSTM 模型参数进行优化,将各个子序列导入 ISCSO-LSTM 模型预测并叠加得到最终预测结果。实验结果表明,CEEMDAN-VMD-ISCSO-LSTM 组合模型具有较低的预测误差,相比 CEEMDAN-VMD-LSTM 和 CEEMDAN-VMD-SCSO-LSTM 模型,该模型在均方根误差方面分别降低了 2.21 和 1.04 $\mu g/m^3$,在拟合度方面分别提高了 4.9%和 2.1%。

关键词:空气质量预测;二次分解;改进沙猫群算法;长短期记忆网络
中图分类号: TP393 **文献标识码**:A **国家标准学科分类代码**: 510.4030

Air quality predication based on two-layer decomposition and improved sand cat swarm optimization

Zhu Juxiang[1] Zhang Shiyun[2] Zhang Tao[2] Sun Junfeng[2] Zhang Zhaoliang[1,3]
(1. School of Rail Transportation, Wuxi University, Wuxi 214105, China; 2. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210000, China; 3. Jiangsu Provincial Engineering Research Center for Monitoring and Evaluation of Industrial Environmental Hazard Element, Wuxi 214105, China)

Abstract: Accurate prediction of air quality is of great significance to people's daily life, therefore, a predictive model based on quadratic decomposition and improved sand cat swarm optimization(ISCSO) to optimize the long short-term memory(LSTM) network was proposed. First of all, The $PM_{2.5}$ data was decomposed into multiple subsequences using complete ensemble empirical mode decomposition with adaptivenoise (CEEMDAN) algorithm, and the reconstructed sequence that are not satisfied with the prediction effect was quadratically decomposed by variational mode decomposition (VMD) method. Secondly, the sand cat swarm optimization was improved by introducing Cubic chaotic, spiral search strategy and sparrow alert mechanism to improve the global search performance and convergence speed of the algorithm. Finally, a improved sand cat swarm algorithm was used to optimize the LSTM model parameters, the individual subsequences were input into the ISCSO-LSTM model for prediction and superimposed to obtain the final prediction results. The experimental results show that the CEEMDAN-VMD-ISCSO-LSTM combination model exhibits lower prediction errors, compared to the CEEMDAN-VMD-LSTM and CEEMDAN-VMD-SCSO-LSTM, the model proposed in this article has a 2.21 and 1.04 $\mu g/m^3$ reduction respectively in root mean square error, and has a 4.9% and 2.1% higher respectively in term of fit.

Keywords: air quality predication; two-layer decomposition; improved sand cat swarm optimization; long short-term memory network

收稿日期:2023-11-29
* **基金项目**:"太湖之光"科技攻关项目(k20221050)资助

0 引 言

2000～2010年,中国有超80%的人口暴露在可吸入颗粒物($PM_{2.5}$)浓度高于35 $\mu g \cdot m^{-3}$的环境下[1],直到近几年污染情况才缓和。空气污染严重威胁人类健康,$PM_{2.5}$导致的空气污染会引起一系列疾病,例如心肺疾病,呼吸系统疾病,心脏病等。据研究,环境空气污染对死亡率的影响大于其他主要可改变的风险因素,遵守细颗粒物的空气质量标准可最大限度避免因空气污染引起的死亡[2]。空气污染物大致可分为生活污染物,物理污染物和化学污染物,这些污染物释放出的CO、CO_2、NO_2和$PM_{2.5}$等会影响空气质量,因此空气质量的准确预测对人类健康管理和政府环境管理决策具有重要意义。

空气质量预测属于时间序列研究的问题一类,时间序列预测的基础是挖掘数据特征,一直受到许多学者的广泛关注。近年来研究人员提出了多种预测模型,常用预测方法有物理模型、统计模型、神经网络模型等。Amnuaylojaroen等[3]建立两个多元线性回归(multiple linear regression, MLR)模型对泰国北部3个地方的$PM_{2.5}$等气体进行预测,预测变量是温度、相对湿度和风速等参数,结果表明该模型可以很好地预测2020年$PM_{2.5}$浓度。杨涛锋[4]等使用自回归综合移动(autoregressive integrated moving average, ARIMA)和支持向量机(support vector machine, SVM)组合模型对北京市某站点的$PM_{2.5}$浓度数据进行预测,该模型在均方根误差方面相比其他模型均有所下降。Li等[5]提出将粒子群算法优化的灰色模型(gray model, GM)用于时间序列预测,并证明预测精度和适应度都得到了一定改善。虽然上述模型预测速度很快,但是一般的污染物浓度信号具有非线性和非平稳性,所以物理模型和统计模型相比较神经网络模型而言,预测效果差一点。与传统的预测模型相比,神经网络模型等深度学习模型具有良好的学习能力和拟合能力。吴琼等[6]使用BP神经网络搭建预测模型,该模型可以在不同训练量的情况下有效完成预测任务,但其存在无法获取时间序列记忆特征的缺陷。Hou等[7]将卷积神经网络(convolutional neural network, CNN)和长短期记忆(long short-term memory network, LSTM)集成到一个网络模型,用于每小时的气温预测。与上述神经网络相比,LSTM是专门设计用于处理序列数据的神经网络结构,能够有效地捕获时间序列中的时间依赖性和模式。文献[8]提出了一种基于时空相似LSTM的预测模型,其预测精度得到提升,文献[9]提出改进的布谷鸟算法优化长短期记忆深度神经网络的预测模型。

大气污染物浓度变化通常含有大量噪声,对数据进行预处理是必不可少的一步,将信号分解技术与深度学习结合的方法逐渐应用于预测领域[10]。Erbiao等[11]基于经验模态分解(empirical mode decomposition, EMD)将原始$PM_{2.5}$序列分解为若干个子序列,且不需要选择基函数,但EMD分解得到的模态分量存在模态混叠现象。石欣等[12]将变分模态分解(variational mode decomposition, VMD)算法和NARX神经网络结合,搭建预测模型,虽然VMD算法相比EMD算法,在信号分解方面有一定的优势,但是对于利用K-means算法聚类之后的高频序列的预测效果仍然较差。Zeng等[13]采用完全自适应噪声集合经验模态分解(complete ensemble empirical mode decomposition with adaptive noise, CEEMDAN)和深度变压器神经网络相结合的混合模型提高了$PM_{2.5}$长期预测精度。鉴于数据序列的复杂性,单分解方法可能难以全面分析非线性强的信号,因此二次分解技术近年来也被许多学者用来进一步提取数据特征以提高模型精度。周尧民等[14]运用二次分解技术对北京市日均$PM_{2.5}$浓度数据进行预测,结果证明,经过CEEMDAN分解后的信号再次利用VMD技术使模型在均方根误差(RMSE)和平均绝对误差(MAE)模型评价指标中,显著优于已有的组合模型。

为了提高预测结果的准确性,模型优化问题一直是研究的焦点,例如粒子群算法(particle swarm algorithm, PSO)[15]、麻雀搜索算法[16](sparrow search algorithm, SSA)、沙猫群算法(sand cat swarm optimization, SCSO)[17]等被用来优化各种模型参数。文献[18]采用改进粒子群算法自动调整模型参数的最优值,文献[19]采用麻雀搜索算法优化HKELM网络参数,文献[20]将沙猫群算法和XGBoost集合模型相结合,有效预测了短期岩爆损伤的类型。结果表明,相比单一模型和未经优化的模型,优化后的元启发式算法模型预测精度显著提高。与其他算法相比,沙猫群算法在减少过拟合方面展现出显著的优势,稳定性更好,搜索能力更强,但是迭代后期会出现种群多样性减少,陷入局部最优陷阱等问题。

针对上述问题,本文建立了CEEMDAN-VMD-ISCSO-LSTM混合模型在南京市空气质量预测中的应用。首先使用CEEMDAN和VMD方法对原始序列进行分解;其次,引入Cubic混沌映射增加种群多样性,引入螺旋搜索策略扩展探索未知区域和算法全局搜索的能力,引入麻雀警戒机制提升收敛速度;最后,利用改进的SCSO算法优化LSTM模型权重和阈值,将各个子序列的预测值叠加得到最终预测值。

1 基本原理

1.1 CEEMDAN算法

CEEMDAN由EMD、集合经验模态分解(ensemble empirical mode decomposition, EEMD)、互补集合经验模态分解(complementary ensemble empirical mode decomposition, CEEMD)基础上演练过来,其较好地抑制了EMD模态混叠现象、重构信号中的残余噪声比EEMD更小,并且解决了CEEMD每组IMF分量分解结果差异导致最后集合难以对齐或产生误差的问题。CEEMDAN算法减少了计算量和迭代次数,其步骤如下。

1)将原始信号 $x(t)$ 添加 n 次相同长度的高斯白噪声，构造 n 次待分解序列 $x_i(t)$，其中 $i=1, 2, 3, \cdots, n$，公式如下：

$$x_i(t) = x(t) + \varepsilon_0 \delta_i(t) \tag{1}$$

式中：ε_0 为高斯白噪声权值系数；$\delta_i(t)$ 为第 i 次产生的高斯白噪声。

2)将上述序列 $x_i(t)$ 通过 EMD 分解得到模态分量 $IMF_1(t)$，通过 EMD 重复分解 n 次取平均值得到 CEEMDAN 的第 1 个模态分量 IMF_1 和残余量 $r_1(t)$，将第 1 个残余量再添加白噪声，由此类推得到 $IMF_2(t)$ 和 $r_2(t)$，其公式如下：

$$IMF_1 = \frac{1}{n}\sum_{i=1}^{n} IMF_1(t) = \frac{1}{n} EMD_1(x_i(t)) \tag{2}$$

$$r_1(t) = x(t) - IMF_1 \tag{3}$$

$$IMF_2 = \frac{1}{n}\sum_{i=1}^{n} IMF_2(t) \tag{4}$$

$$r_2(t) = r_1(t) - IMF_2 \tag{5}$$

3)与上述步骤相似，得到第 k 分量，公式如下：

$$IMF_k = \frac{1}{n}\sum_{i=1}^{n} EMD_1(r_{k-1}(t) + \varepsilon_{k-1} EMD_{k-1}(\delta_i(t))) \tag{6}$$

$$r_k(t) = r_{k-1}(t) - IMF_k \tag{7}$$

式中：IMF_k 表示 CEEMDAN 分解得到的第 k 模态分量；$r_k(t)$ 表示第 k 阶残差信号。

4)直到残差分量为单调信号即无法分解时，迭代结束，提取所有的模态分量和趋势项 $R(t)$，公式如下：

$$x(t) = \sum_{i=1}^{n} IMF_K + R(t) \tag{8}$$

1.2 VMD

VMD 是一种非递归变分模式的信号分解方法，通过变分问题进行求解，获取分解后各模态函数分量带宽限制，找到各中心频率在频域中对应的有效成分，得到模态函数[21]。假设原始输入信号为 $f(t)$，传统 VMD 模型的约束变分模型为：

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{k} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \tag{9}$$

$$\text{s.t.} \quad \sum u_k = f(t) \tag{10}$$

式中：$\{u_k\}$ 表示分解后的 k 个模态分量集合；$\{\omega_k\}$ 表示每个模态分量对应的中心频率集合；$\delta(t)$ 是脉冲函数。约束条件表明，所有模态分量之和应为原始信号。为了将约束优化问题转化为无约束变分问题，在拉格朗日展开表达式中利用二次惩罚因子 α 和拉格朗日乘数算子 $\lambda(t)$，引入增广函数，得到如下：

$$L(\{u_k\},\{\omega_k\},\lambda) = \partial \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle \tag{11}$$

式(11)的解在频域中进行，步骤如下。

1)初始化 $\{\hat{u}_k^1\}$、$\{\hat{\omega}_k^1\}$、$\hat{\lambda}^1 = 0$。

2)对所有的 $\omega \geqslant 0$，更新 $\hat{u}_k$：

$$\hat{u}_k^{n+1}(\omega) \leftarrow \frac{\hat{f}(\omega) - \sum_{i<k} \hat{u}_i^{n+1}(\omega) - \sum_{i>k} \hat{u}_i^n(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^n)^2} \tag{12}$$

3) 更新 ω_k：

$$\omega_k^{n+1} \leftarrow \frac{\int_0^\infty \omega |\hat{u}_k^{n+1}(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k^{n+1}(\omega)|^2 d\omega} \tag{13}$$

4)对所有的 $\omega \geqslant 0$，进行双重提升，更新 λ_k：

$$\hat{\lambda}^{n+1}(\omega) \leftarrow \hat{\lambda}(\omega) + \tau(\hat{f}(\omega) - \sum_k \hat{u}n + 1_k(\omega)) \tag{14}$$

其中，τ 表示噪声容限，当信号含有强噪声时，可设定 $\tau=0$ 达到更好的去噪效果。

5)重复步骤 1)～3)，直到满足迭代约束条件。

$$\sum_k \frac{\| \hat{u}_k^{n+1} - \hat{u}_k^n \|_2^2}{\| \hat{u}_k^n \|_2^2} < \varepsilon \tag{15}$$

则迭代结束。

式中：n 表示整个过程的迭代次数；u_k^n、ω_k^n、λ^n 分别表示模态分量的序列，中心频率和热乘数。

1.3 SCSO

沙猫群算法是一种模拟沙猫在自然界搜索猎物和攻击猎物两种行为的生物学算法，其具有结构简单，参数少等优点。假设沙猫群的初始种群随机生成，数量为 N，搜索空间维度为 d，则第 i 只沙猫的位置为 $X_i = \{x_1, x_2, x_3, \cdots, x_d\}$，$i = 1, 2, 3, \cdots, N$，算法步骤如下。

搜索猎物的数学模型如下所示：

$$X(t+1) = r \cdot (X_b(t) - rand(0,1) \cdot X_c(t)) \tag{16}$$

式中：t 表示当前迭代次数，X_b 和 X_c 分别表示全局最优位置和当前位置，为避免陷入局部最优，使用如下公式模拟每只沙猫灵敏度范围：

$$r = r_G \times rand(0,1) \tag{17}$$

$$r_G = S_M - \left(\frac{S_M \times i_t}{i_{t\max}} \right) \tag{18}$$

式中：r 应用于搜索和攻击阶段，表示种群中任意只沙猫对低频噪声的灵敏度范围，每只沙猫可以通过感知低于 2 kHz 低频噪声来定位地上和地下的猎物；r_G 表示灵敏度范围，值为随着迭代次数增加从 2 线性递减至 0，i_t 和 i_{tmax} 分别表示当前迭代次数和最大迭代次数，S_M 取值为 2。沙猫群算法利用参数 R 决定进入搜索行为还是攻击行为，如式(9)，示意图如图 1 所示。

$$R = 2 \times r_G \times rand(0,1) - r_G \tag{19}$$

式中：R 值大小依赖 r_G，是 $[-r_G, r_G]$ 中的一个随机值，当 $|R|>1$ 时，执行搜索任务(式(16))，当 $|R| \leqslant 1$ 时，执行攻击任务，攻击猎物的数学模型如下：

$$X(t+1)=X_b(t)-r\cdot X_{rnd}\cdot\cos\theta \tag{20}$$

$$X_{rnd}=|rand(0,1)\cdot X_b(t)-X_c(t)| \tag{21}$$

式中：X_{rnd} 表示全局最优位置和当前位置之间随机生成的一个位置；θ 是根据轮盘赌选择的随机角度，取值范围是0°～360°，通过这种方法使得群体中每只沙猫都可沿着不同的圆周方向移动，以此避免陷入局部最优陷阱。

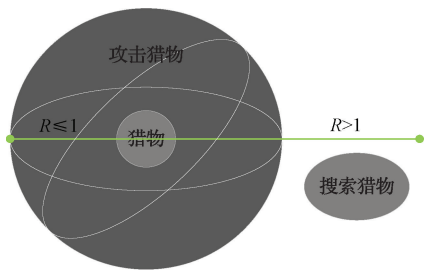


图1 沙猫群捕食过程

Fig. 1 Predation process of sand cat

1.4 LSTM网络

LSTM是一种用于处理序列数据的循环神经网络(recurrent neural network，RNN)的变体，有效缓解了RNN无法避免的梯度爆炸问题，能更好预测时间序列。LSTM通过引入记忆单元和门控机制来实现对序列信息的长期记忆和选择性遗忘。LSTM的细胞结构如图2所示。

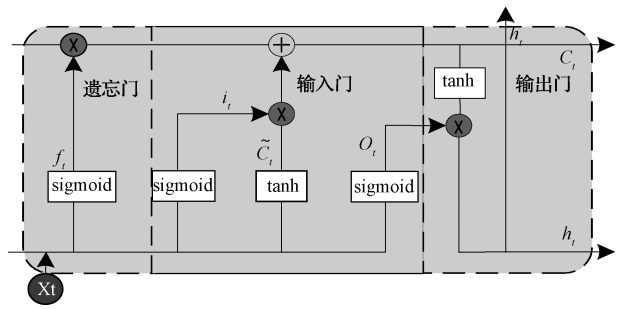


图2 LSTM原理

Fig. 2 LSTM principle

该网络在状态路径上有3个控制门，分别是遗忘门、输入门和输出门。遗忘门设计用于丢弃存储单元信息，输入门设计负责更新单元状态信息，输出门根据当前时刻内部状态来调整外部状态输出。

2 改进沙猫群算法

虽然沙猫群算法的全局搜索性能和收敛速度已经达到较高的标准，但是迭代后期会出现种群多样性减少，陷入局部最优以及寻优精度降低等情况，本文从3个方面对原始沙猫群算法进行改进。

2.1 基于Cubic混沌映射的种群初始化

多样化的初始种群可以极大地提高算法性能，然而，沙猫群算法采用随机技术生成初始化种群，这样易造成种群分布不均匀、质量差和多样性减少以及搜索效率低等问题。混沌映射有很多种，例如Tent映射、Logistic映射、Sine映射和Cubic映射等，其中Cubic的混沌性能更高，产生的混沌序列更复杂。因此，为了将群体均匀分布在搜索空间，本文将映射效果较为稳定的Cubic混沌映射应用于沙猫群算法中初始化种群的生成，计算沙猫初始化位置，随机更新沙猫群，映射公式如下。假设种群数量为100，初始化种群分布如图3所示，由图3可知，通过Cubic混沌序列产生的初始化种群分布相对来说较为均匀，遍历性良好。

$$x_{n+1}=\rho x_n(1-x_n^2) \tag{22}$$

式中：x_{n+1} 的初始值为 x_0；ρ 表示控制参数，其取值影响Cubic的映射性。参考文献[22]，$x_0=0.3$，$\rho=2.595$。

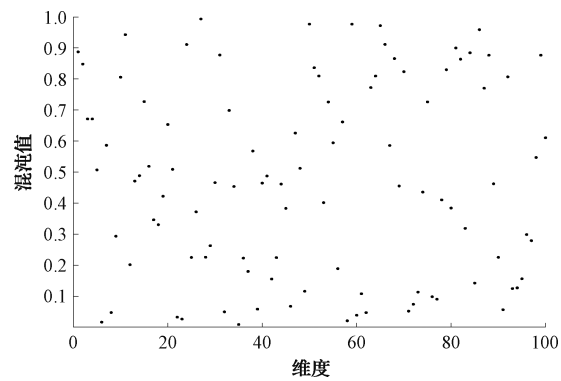


图3 Cubic混沌初始化

Fig. 3 Cubic chaos initialization

2.2 螺旋搜索策略

沙猫群在搜索步骤中，每个当前沙猫的位置更新基于搜索者位置，这种搜索方式具有单一性，部分空间可能没有被遍历到，为了扩展搜索区域，在沙猫搜索猎物阶段引入螺旋搜索策略。螺旋搜索策略是受鲸鱼优化算法螺旋操作的启发而提出的一种可螺旋改变位置的更新策略，表述鲸鱼围捕猎物行为的公式如下：

$$\begin{cases}X(t+1)=D'\cdot e^{bl}\cdot\cos2\pi l+X^*(t)\\ D'=|X^*(t)-X(t)|\end{cases} \tag{23}$$

式中：$l\in[-1,1]$的随机数，常量系数 b 对搜索过程具有很大影响，b 值过大会导致陷入局部最优，b 值过小会导致收敛速度较慢，为了解决该问题，通过引入用于动态调整螺旋搜索形状的自适应变量 z，公式如下：

$$z=e^{k\cdot\cos(\pi\cdot(1-t/t_{\max}))} \tag{24}$$

式中：k 表示变化系数，值为5。引入螺旋搜索因子 z，改进后的沙猫位置更新如下：

$$X(t+1)=e^{z\cdot l}\cdot\cos2\pi l\cdot r\cdot(X_b(t)-rand(0,1)\cdot X_c(t)) \tag{25}$$

融入螺旋搜索策略后，沙猫群将以螺旋形式在搜索空间搜索，使位置更新更加灵活，拓展搜索未知空间的能力，极大减小落入局部最优陷阱的概率，提高了算法搜索

效率。

2.3 麻雀警戒机制

沙猫是猫科动物的一种，常年生活在环境恶劣的沙漠里，虽然沙猫和家猫体型一样小巧灵活，听觉甚至比家猫更灵敏，能够检测低频噪声，是有名的猎食高手，但是仍然会遇到被天敌发现的情形。因为原始沙猫群优化算法未加入沙猫感知危险任务这一机制，所以本文参考麻雀搜索算法，在算法中融入麻雀警戒机制，表述感知危险的沙猫位置表达式如下。

$$X_{i,j}^{t+1}=\begin{cases}X_{best}^{t}+\beta\cdot|X_{i,j}^{t}-X_{best}^{t}|, & f_i>f_g\\ X_{i,j}^{t}+k\dfrac{|X_{i,j}^{t}-X_{worst}^{t}|}{(f_i-f_w)+\varepsilon}, & f_i=f_g\end{cases}\tag{26}$$

式中：X_{best}^{t} 表示全局最优位置；β 表示服从正态分布的阶跃调整因子；k 是[−1,1]的随机数；f_i、f_g 和 f_w 分别表示当前、最佳和最差适应度值。

3 CEEMDAN-VMD-ISCSO-LSTM 预测模型构建

本文基于 CEEMDAN-VMD-ISCSO-LSTM 预测模型的流程如图 4 所示。

步骤 1）使用线性插值法填补缺失的数据，然后对原始序列使用 CEEMDAN 方法进行模态分解。

步骤 2）将 CEEMDAN 分解得到的高频信号利用 VMD 算法再次分解以提取复杂分信号的潜在特征。

步骤 3）利用式（22）的 Cubic 混沌映射初始化沙猫种群。

步骤 4）计算各只沙猫适应度值，利用式（25）更新搜索者位置。

步骤 5）从沙猫群中选取部分沙猫作警戒者，利用式（26）更新位置，剩余的沙猫按照原策略更新位置。

步骤 6）将经过二次分解后的高频序列同低频序列、趋势序列输入 ISCSO 优化的 LSTM 模型预测。

步骤 7）对各分信号预测结果进行整合叠加得到最终预测结果并与原始数据做对比分析。

为了评估 CEEMDAN-VMD-ISCSO-LSTM 网络模型的预测性能，本文采用 RMSE、MAE 和拟合度 R^2 对算法模型进行评价。RMSE、MAE 反映了真实值和预测值的偏差，其值越小，模型预测效果越好，R^2 的值越接近 1，预测模型的拟合效果越好，如式（27）～（29）所示。

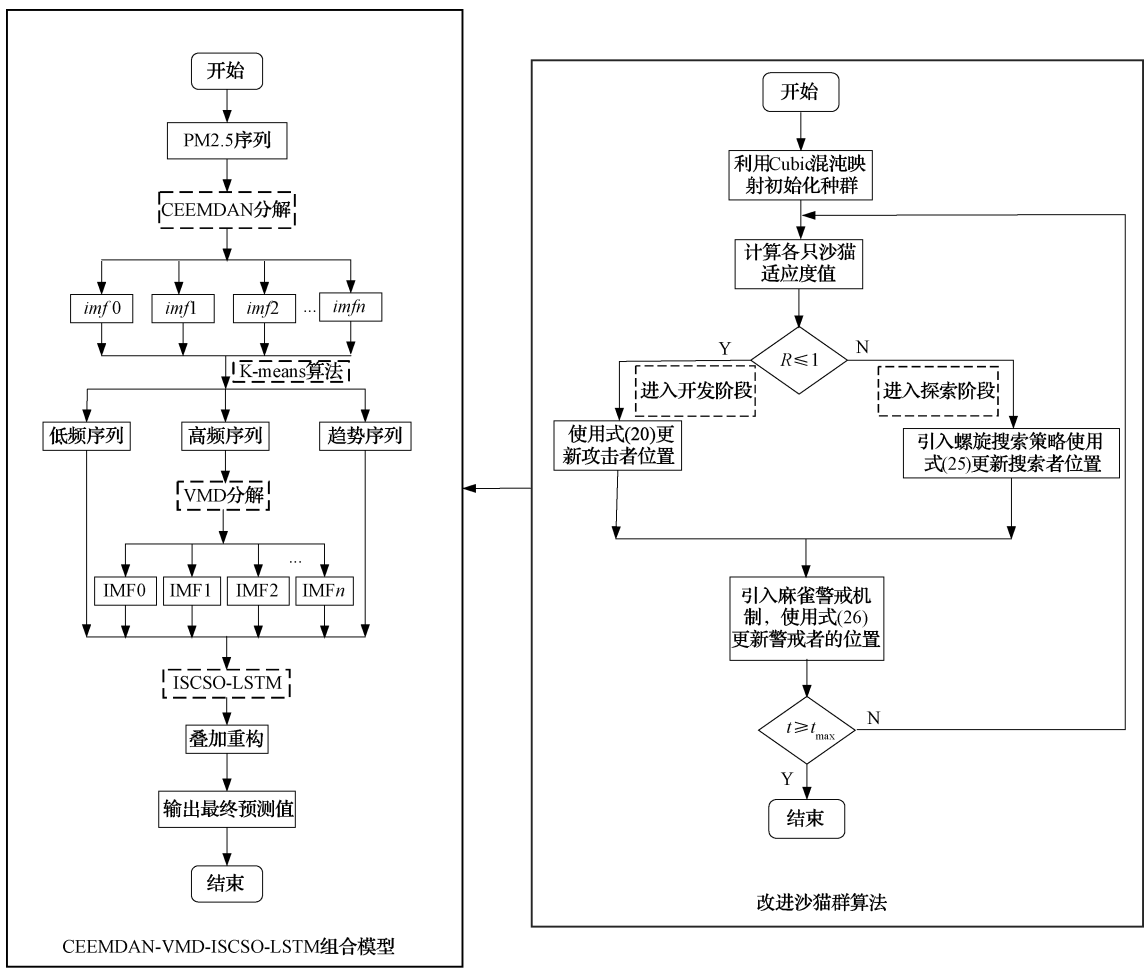


图 4 CEEMDAN-VMD-ISCSO-VMD 模型预测流程

Fig. 4 The model prediction process of the CEEMDAN-VMD-ISCSO-VMD

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_l)^2}{n}} \tag{27}$$

$$MAE=\frac{\sum_{i=1}^{n}|y_i-\hat{y}_l|}{n} \tag{28}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_l)^2}{\sum_{i=1}^{n}(y_i-\hat{y}_l)^2} \tag{29}$$

式中：y_i 和 $\hat{y}_l$ 分别是测试集的真实值和算法模型的预测值；n 为总测试次数。

4 数据处理

南京是一座工业城市，其涵盖了不同类型的工业，包括制造业、钢铁产业和化工产业等。工业生产过程涉及到燃烧、粉尘悬浮、废气排放等，这些过程会释放大量空气污染物，其中 $PM_{2.5}$ 和 CO 的排放对空气质量管理、人们的健康管理和城市的经济发展产生影响，因此，$PM_{2.5}$ 和 CO 的准确预测可以帮助提前了解南京市的环境质量和污染趋势，以便做出相应健康防护措施和城市规划。

综上，选取江苏省南京市 2015 年 11 月 19 日～2022 年 12 月 31 日的 $PM_{2.5}$ 浓度和 CO 浓度作为研究对象，各 2 600 组数据，其中 80%的数据用于训练，20%的数据用于测试，其全部来源于中国空气质量在线监测分析平台（http://www.cnemc.cn）。本文使用线性插值法对数据原有缺失的点进行补缺，原始 $PM_{2.5}$ 浓度序列如图 5 所示。

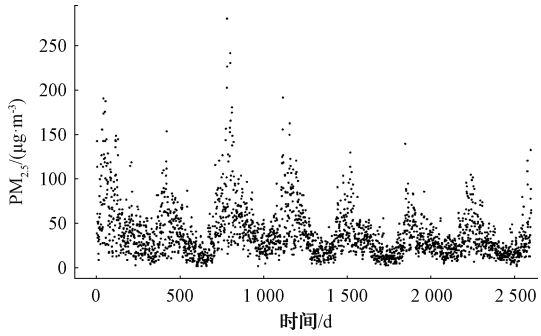


图 5 原始 $PM_{2.5}$ 序列

Fig. 5 Raw $PM_{2.5}$ sequence

从图 5 可以看出，原始 $PM_{2.5}$ 序列具有波动大、非线性、不稳定等特点，所以采用 CEEMDAN 对数据进行分解，分解结果如图 6 所示。由图 6 可知，原始序列被分解成不同尺度的 imf 分量，模态分量按照频率高低从上往下依次显示。随着频率降低，子序列的复杂性和非平稳性也相对降低[23]。利用 K-means 算法将所有分量聚类成新的模态函数，其中高频分量 co-imf0（imf0～imf1）波动较大，复杂度最高，低频分量 co-imf1（imf2～imf4）波动较小，趋势项 co-imf2（imf5～imf7）包含原始序列的趋势信息，聚类过程如图 7 所示。将高频、低频和趋势分量分别输入 ISCSO-LSTM 模型中迭代训练，其高频分量的预测结果如图 8 所示。

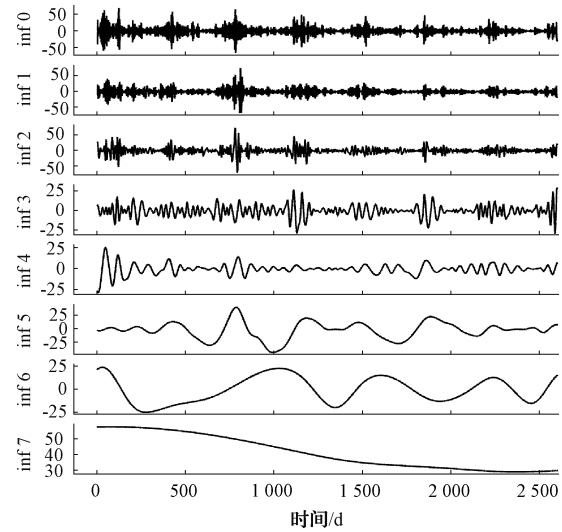


图 6 CEEMDAN 分解

Fig. 6 CEEMDAN decompose

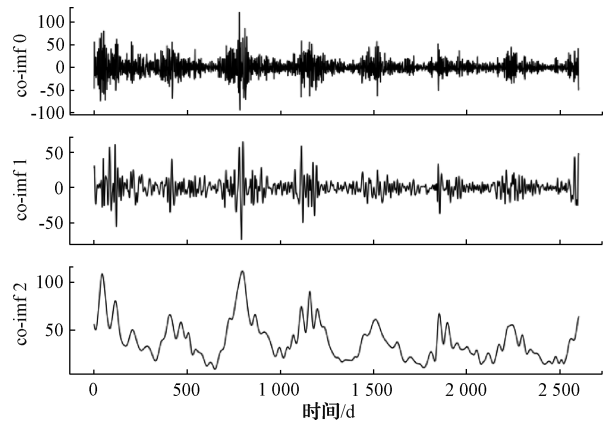


图 7 聚类结果

Fig. 7 Clustering results

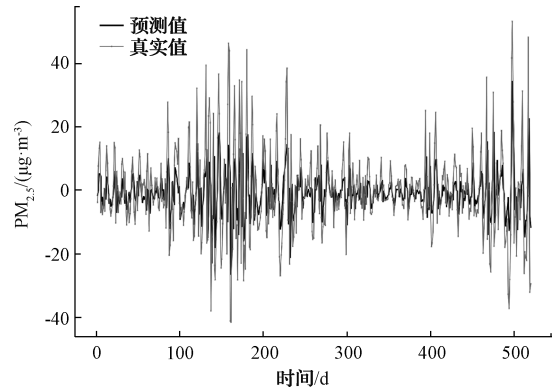


图 8 CEEMDAN-ISCSO-LSTM 高频分量预测

Fig. 8 High frequency component prediction of CEEMDAN-ISCSO-LSTM

从预测结果来看，高频分量由于高频震荡特性，CEEMDAN-ISCSO-LSTM 曲线波动与真实值相差较大，所以为了削弱其不平稳性，提高最终预测效果，本文利用 VMD 方法对难以提取非线性特征的高频分量进行二次分解。VMD 算法需要确定参数 K，令 $K=3,4,5,\cdots,10$，记录对应的中心频率，当出现相似的中心频率时，参数 K 被确定下来。通过多次调试最终确定最佳模态 $K=6$ 时的分解效果最好，使用 VMD 方法对高频序列进行分解，经过二次分解后的高频分量预测结果如图 9 所示。

从图 9 可看出，通过加入 VMD 算法的二次分解，有效地提高了模型的训练效率和预测效果，将二次分解后的 $PM_{2.5}$ 序列输入到 ISCSO-LSTM 模型中预测，并将各子序列的预测结果叠加求和得到最终预测结果。

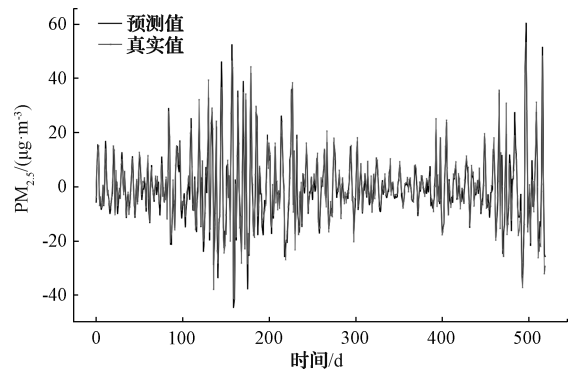


图 9 CEEMDAN-VMD-ISCSO-LSTM 高频分量预测

Fig. 9 High frequency component prediction of CEEMDAN-VMD-ISCSO-LSTM

5 实验结果与讨论

5.1 ISCSO 算法性能分析

为了验证 ISCSO 技术的搜索性能，本文使用一组测试函数对灰狼优化(grey wolf optimization，GWO)算法、哈里斯鹰优化算法(harris hawk optimization，HHO)、减法平均优化器算法(subtraction average based optimizer，SABO)、SCSO 算法和 ISCSO 算法进行比较。这组基准函数由单峰函数(F_1)、多峰函数(F_{13})和固定维多峰函数(F_{19}、F_{22})组成，测试函数如表 1 所示。

表 1 测试函数

Table 1 Test function

函数名	测试函数	n	测试范围
Sphere	$F_1(x)=\sum_{i=1}^{n}x_i^2$	30	$[-100,100]$
Generalized penalized	$F_{13}(x)=0.1\{\sin^2(3\pi x_1)+\sum_{i=1}^{n}(x_i-1)^2[1+\sin^2(3\pi x_i+1)]+(x_n-1)^2[1+\sin^2(2\pi x_{ni})]\}+\sum_{i=1}^{n}u(x_i,5,100,4)$	30	$[-50,50]$
Hartman's family	$F_{19}(x)=-\sum_{i=1}^{4}c_i\exp\left(-\sum_{j=1}^{3}a_{ij}(x_j-p_{ij})^2\right)$	3	$[1,3]$
Shekel 2	$F_{22}(x)=-\sum_{i=1}^{7}[(X-a_i)(X-a_i)^{\mathrm{T}}+c_i]^{-1}$	4	$[0,10]$

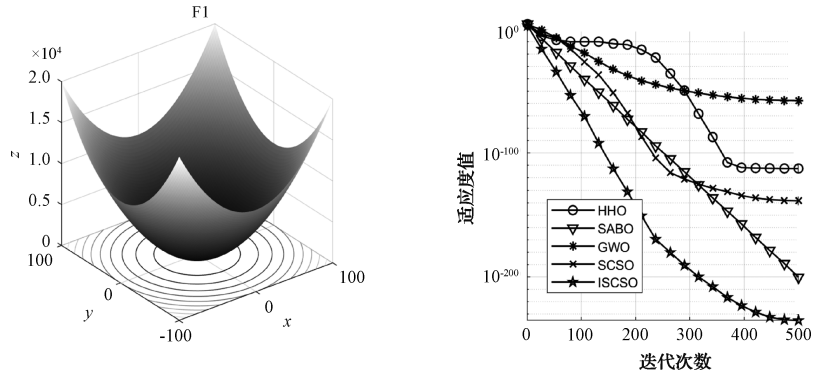
为了公平比较不同算法的综合搜索能力，实验应在相同的环境中进行，以下测试函数仿真实验均采用 MATLAB R2022b 仿真软件，操作系统为 Microsoft Win11(64 位)，处理器为 Intel(R) Core(TM) i5-10210UCPU@1.60 GHz 2.11 GHz，内存为 16 G。此外，设置各算法种群数量为 30，最大迭代次数为 500，所有测试函数独立运行 50 次进行寻优。为了评估算法稳定性，引入函数均值(Mean)和标准差(Std)，公式如下：

$$Mean=\frac{1}{P}\sum_{i=1}^{p}f_i \tag{30}$$

$$Std=\sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(f_i-Mean)^2} \tag{31}$$

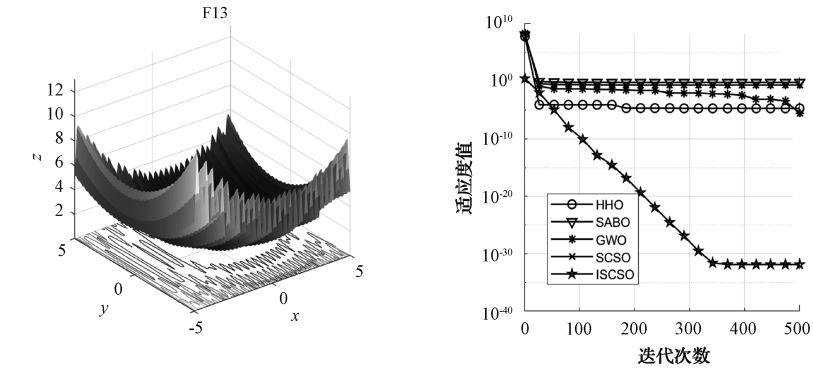
测试函数的模型及收敛曲线如图 10 所示，不同算法对基本函数的测试结果如表 2 所示。

根据种群准则适应度，由图 10 可看出，所有算法都呈现递减的收敛曲线。其中，对于单峰函数而言，改进后的沙猫群算法在收敛速度和寻优成功率方面都优于其他算法，其他算法均会不同程度的陷入局部最优，沙猫群算法引入螺旋搜索策略降低 ISCSO 陷入局部最优的概率。对于多峰函数和固定维多峰函数而言，ISCSO 算法的收敛速度仍然是最快的，即使 ISCSO 算法在前期寻优过程中陷入局部最优，但是随着迭代次数增加，能够快速跳出局部最优。由表 2 可看出，ISCSO 算法的适应度值和标准差几乎都小于其他算法，即 ISCSO 算法性能更加稳定。综上，无论是哪种测试函数，ISCSO 算法性能最好，较其他算法都是以最快的速度获得全局最优解。



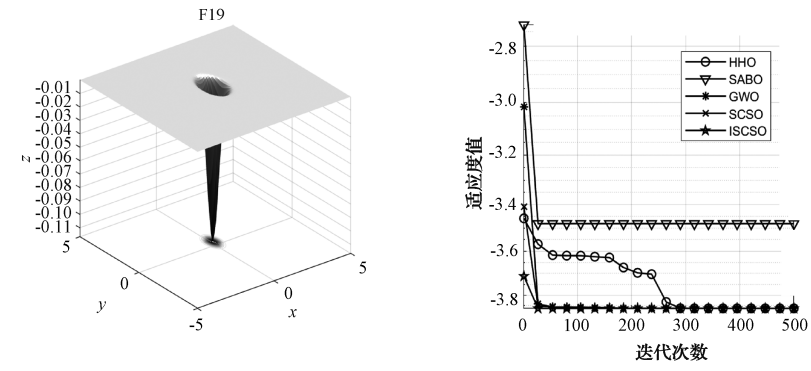
(a) Sphere函数三维模型和收敛曲线

(a) Sphere function 3D model and convergence curve



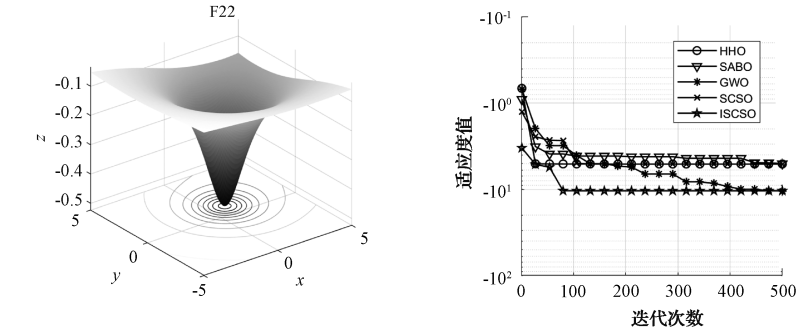
(b) Generalized penalized 函数三维模型和收敛曲线

(b) Generalized penalized function 3 D model and convergence curve



(c) Hartman's family 函数三维模型和收敛曲线

(c) Hartman's family function 3D model and the convergence curve



(d) Shekel 2 函数三维模型和收敛曲线

(d) Shekel 2 function 3 D model and convergence curve

图 10　函数模型及收敛曲线

Fig. 10　Function model and convergence curve

表 2 不同算法对基本函数的测试结果

Table 2 Test results of different algorithms on the basic functions

算法		F_1	F_{12}	F_{19}	F_{22}
GWO	Mean	2.43×10^{-57}	7.27×10^{-3}	-3.86	-1.04×10^{1}
	Std	4.36×10^{-57}	1.22×10^{-2}	2.83×10^{-3}	1.03×10^{-1}
HHO	Mean	2.97×10^{-98}	2.84×10^{-5}	-3.86	-5.08
	Std	8.17×10^{-98}	4.66×10^{-5}	2.40×10^{-3}	1.93
SABO	Mean	1.40×10^{-199}	1.06×10^{-2}	-3.61	-5.03
	Std	0.00	1.89×10^{-1}	1.52×10^{-3}	9.78×10^{-1}
SCSO	Mean	6.43×10^{-132}	2.08×10^{-2}	-3.86	-5.08
	Std	2.93×10^{-131}	2.92×10^{-2}	3.54×10^{-3}	2.95
ISCSO	Mean	2.66×10^{-229}	2.04×10^{-31}	-3.86	-1.041×10^{1}
	Std	0.00	8.48×10^{-31}	1.44×10^{-3}	4.32×10^{-2}

5.2 预测模型性能对比分析

为了突显出改进沙猫群算法的优越性能，本文将经过双分解后的 $PM_{2.5}$ 数据分别导入 LSTM、SCSO-LSTM 和 ISCSO-LSTM 模型中预测，预测结果如图 11 所示，评价指标如表 3 所示。

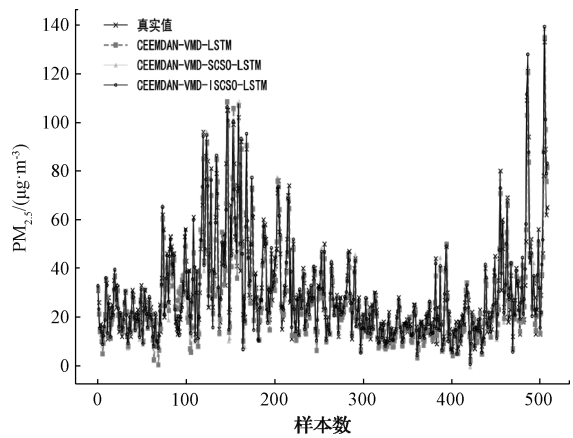


图 11 各算法模型预测结果对比

Fig. 11 Comparison of the prediction results of each algorithm model

表 3 模型评价指标

Table 3 Model evaluation index

算法模型	RMSE/($\mu g\cdot m^{-3}$)	MAE/($\mu g\cdot m^{-3}$)	R^2/%
CEEMDAN-VMD-LSTM	5.22	3.87	93.0
CEEMDAN-VMD-SCSO-LSTM	4.05	2.87	95.8
CEEMDAN-VMD-ISCSO-LSTM	3.01	2.06	97.9

由图 11 可看出，CEEMDAN-VMD-SCSO-LSTM 模型比 CEEMDAN-VMD-LSTM 模型预测效果更好，但是还有提升的空间，CEEMDAN-VMD-ISCSO-LSTM 模型的曲线波动幅度与真实值最接近，表明本文所提的改进沙猫群算法对模型预测具有良好效果。由表 3 可看出，本文所提模型的均方根误差为 3.01 $\mu g/m^3$，平均绝对误差为 2.06 $\mu g/m^3$，拟合度达到了 97.9%。对比拟合度可得，CEEMDAN-VMD-ISCSO-LSTM 模型比未使用优化算法的模型提高了 4.9%，比使用改进前的沙猫群优化算法的模型提高了 2.1%。因此，CEEMDAN-VMD-ISCSO-LSTM 模型具有最高的预测精度。

为了全面且直观的体现出本文所提出的基于二次分解和改进沙猫群算法的空气质量模型的预测优势，以 CO 数据为例，将单分解模型 VMD-ISCSO-LSTM、组合模型 CEEMDAN-VMD-GRU、CEEMDAN-VMD-AVOA-LSTM 和本文所提模型 CEEMDAN-VMD-ISCSO-LSTM 进行对比，预测结果和相对误差图分别如图 12 和 13 所示，评价指标如表 4 所示。

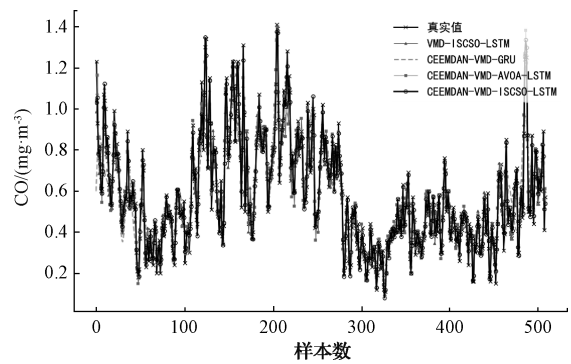


图 12 各算法模型预测结果对比

Fig. 12 Comparison of the prediction results of each algorithm model

由图 12 和 13 可以看出，CEEMDAN-VMD-ISCSO-LSTM 模型中的某些点虽然预测效果略低于其他模型，但是从总体上看，其还原原始序列变化趋势的能力最佳，预测精度最好。VMD-ISCSO-LSTM 模型比 CEEMDAN-VMD-ISCSO-LSTM 模型的拟合度差，说明虽然单分解能够降低序列的复杂性，但是效果不好，二次分解的模型利用 VMD 方法对高频序列再次分解可以降低对模型预测

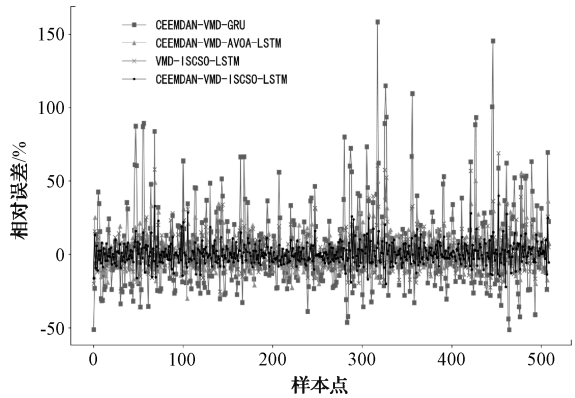


图13 各算法模型相对误差结果对比

Fig. 13 Comparison of the prediction results ofeach algorithm model

表4 模型评价指标

Table 4 Model evaluation index

算法模型	RMSE/(mg·m^{-3})	MAE/(mg·m^{-3})	R^2/%
CEEMDAN-VMD-GRU	0.13	0.10	73.5
VMD-ISCSO-LSTM	0.06	0.05	94.4
CEEMDAN-VMD-AVOA-LSTM	0.06	0.04	94.7
CEEMDAN-VMD-ISCSO-LSTM	0.03	0.03	97.8

误差的影响。由表4可得出，将本文所提CEEMDAN-VMD-ISCSO-LSTM模型与VMD-ISCSO-LSTM、CEEMDAN-VMD-GRU、CEEMDAN-VMD-AVOA-LSTM模型进行对比，均方根误差分别降低了0.03、0.1、0.03 mg/m^3；在拟合度方面，分别提高了3.4%、24.3%、3.1%。更进一步的，将CEEMDAN-VMD-ISCSO-LSTM模型与CEEMDAN-VMD-AVOA-LSTM模型进行比较，无论是在均方根误差，平均绝对误差还是拟合度方面，本文所提模型都有明显的优势，说明改进的沙猫群算法更能有效优化LSTM参数，提升预测精度。综上所述，从各个方面，CEEMDAN-VMD-ISCSO-LSTM模型都体现出了较好的预测性能。

6 结 论

本文以南京市的$PM_{2.5}$浓度和CO浓度数据为基础，提出了一种基于二次分解和改进沙猫群算法的空气质量预测，通过实验与分析，得到如下结论。1)对于CEEMDAN分解重构后的高频序列，采用二次分解技术可以进一步提取序列中的非平稳性特征；2)针对沙猫群算法在迭代后期出现种群多样性减少，陷入局部最优等问题，本文引入Cubic混沌映射，螺旋搜索策略和麻雀警戒机制有效增加了种群多样性，加快了收敛速度，提高了算法搜索性能；3)ISCSO在收敛速度、预测精度及稳定性方面与最先进的优化方法相比具有很大的竞争力，引入ISCSO算法优化LSTM模型参数使CEEMDAN-VMD-ISCSO-LSTM模型预测结果精度大幅度提升；4)本文所构建的“分解－聚类－集成”的组合模型预测准确性最高，是一个可应用的模型，$PM_{2.5}$和CO预测可为人们的日常生活和环境管理提供有价值的支撑。

参 考 文 献

[1] 朱玥，石玉胜，李正强. 2000年—2018年中国和印度的长期$PM_{2.5}$污染暴露的疾病负担研究[J]. 遥感学报，2023，27(8)：1834-1843.
ZHU Y，SHI Y SH，LI ZH Q. Studyon long-term disease burden of $PM_{2.5}$ pollution exposure in China and India from 2000 to 2018[J]. Remote Sensing Journal，2023，27(8)：1834-1843.

[2] HAYES R B，LIM C，ZHANG Y，et al. $PM_{2.5}$ air pollution and cause specific cardiovascular disease mortality[J]. International Journal of Epidemiology，2020，49(1)：25-35.

[3] AMNUAYLOJAROEN T. Prediction of $PM_{2.5}$ in an urban area of northern Thailand using multivariate linear regression model[J]. Advances in Meteorology，2022，DOI：10.1155/2022/3190484.

[4] 杨涛锋，彭艺. 基于改进PSO的ARIMA-SVM空气质量预测研究[J]. 云南大学学报(自然科学版)，2020，42(5)：854-862.
YANG T F，PENG Y. ARIMA-SVM air quality prediction study based on the improved PSO[J]. Journal of Yunnan University (Natural science edition)，2020，42(5)：854-862.

[5] LI K，LIU L，ZHAI J，et al. The improved grey model based on particle swarm optimization algorithm for time series prediction[J]. Engineering Applications of Artificial Intelligence，2016，55：285-291.

[6] 吴琼，徐锐良，杨晴霞，等. 基于PCA和GA-BP神经网络的锂电池容量估算方法[J]. 电子测量技术，2022，45(6)：66-71.
WU Q，XU R L，YANG Q X，et al. Capacity estimation method based on PCA and GA-BP neural network[J]. Electronic Measurement Technology，2022，45(6)：66-71.

[7] HOU J，WANG Y，ZHOU J，et al. Prediction of hourly air temperature based on CNN-LSTM[J]. Geomatics，Natural Hazards and Risk，2022，13(1)：1962-1986.

[8] 方伟，朱润苏. 基于时空相似LSTM的空气质量预测模型[J]. 计算机应用研究，2021，38(9)：2640-2645.
FANG W，ZHU R S. Air quality prediction model based on a spatiotemporal similar LSTM[J]. Computer Application Research，2021，38(9)：2640-2645.

[9] 王贺，陈蕨峰，熊敏，等. 融合CEEMDAN和ICS-LSTM的短期风速预测建模[J]. 电子测量与仪器学

报,2022,36(4):17-23.
WANG H, CHEN H F, XIONG M, et al. Modeling of short-term wind speed prediction by fused CEEMDAN and ICS-LSTM [J]. Journal of Electronic Measurement and Instruments, 2022, 36(4): 17-23.
[10] 高凯悦,牟莉.基于二次分解和GRU-attention的时间序列预测研究[J].国外电子测量技术,2023,42(2):80-87.
GAO K Y, MU L. Time series prediction study based on quadratic decomposition and GRU-attention [J]. Foreign Electronic Measurement Technology, 2023,42(2):80-87.
[11] YUAN E B, YANG G F. SA-EMD-LSTM: A novel hybrid method for long-term prediction of classroom $PM_{2.5}$ concentration [J]. Expert Systems With Applications,2023, 230: 120670.
[12] 石欣,张夏恒,朱雅亲,等.基于VMD-NARX的MOSFET剩余使用寿命预测方法[J].仪器仪表学报,2023,44(9):275-286.
SHI X, ZHANG X H, ZHU Y Q, et al. Residual life prediction method of MOSFET based on VMD-NARX[J]. Chinese Journal of Scientific Instrument, 2023, 44(9): 275-286.
[13] ZENG Q, WANG L, ZHU S, et al. Long-term $PM_{2.5}$ concentrations forecasting using CEEMDAN and deep Transformer neural network[J]. Atmospheric Pollution Research, 2023, 14(9): 101839.
[14] 周尧民,黄恒君.基于二次分解和深度学习的$PM_{2.5}$集成预测方法[J].统计学报,2021, 2(3): 84-94.
ZHOU Y M, HUANG H J. $PM_{2.5}$ ensemble prediction method based on quadratic dec-omposition and deep learning [J]. Statistical Journal, 2021, 2(3): 84-94.
[15] HUANG Y, XIANG Y, ZHAO R, et al. Air quality prediction using improved PSO-BP neural network[J]. IEEE Access,2020, 8: 99346-99353.
[16] GAO X Z, GUO W, MEI C X, et al. Short-term wind power forecasting based on SSA-VMD-LSTM[J]. Energy Reports,2023, 9(S10): 335-344.
[17] SEYYEDABBASI A, KIANI F. Sand cat swarm optimization: A nature-inspired algorithm to solve global optimization problems [J]. Engineering with Computers, 2023, 39(4): 2627-2651.
[18] 连莲,穆雅伟,宗学军,等.基于改进粒子群算法优化LSTM-AM的公交客流量预测[J/OL].控制工程,1-9 [2024-04-22]. https://doi.org/10.14107/j.cnki.kzgc.20220556.
LIAN L, MU Y W, ZONG X J, et al. Bus ridership flow based on LSTM-AM with improved particle swarm optimization algorithm [J/OL]. Control Engineering, 1-9 [2024-04-22]. https://doi.org/10.14107/j.cnki.kzgc.20220556.
[19] 郭建帅,崔双喜,郭建斌,等.基于VMD-SSA-HKELM的超短期负荷预测[J].国外电子测量技术,2022,41(6):105-111.
GUO J SH, CUI SH X, GUO J B, et al. Ultra-short-term load prediction based on VMD-SSA-HKELM [J]. Foreign Electronic Measurement Technology, 2022,41(6):105-111.
[20] QIU Y, ZHOU J. Short-term rockburst damage assessment in burst-prone mines: An explainable XGBOOST hybrid model with SCSO algorithm[J]. Rock Mechanics and Rock Engineering, 2023, DOI: 10.1007/s00603-023-03522-w.
[21] 刘成龙,高旭,曹明.基于VMD和BA优化随机森林的短期负荷预测[J].中国测试,2022, 48(4):159-165.
LIU CH L, GAO X, CAO M. Short-term load forecasting based on random forest with VMD and BA optimization algorithm[J]. China test, 2022, 48(4): 159-165.
[22] 张孟健,张浩,陈曦,等.基于Cubic映射的灰狼优化算法及应用[J].计算机工程与科学,2021,43(11):2035-2042.
ZHANG M J, ZHANG H, CHEN X, et al. Optimization algorithm and application of grey wolf based on cubic mapping [J]. Computer Engineering and Science, 2021,43(11):2035-2042.
[23] 闫勇志,沐年国.基于CEEMDAN-VMD-LSTM的超高频金融时间序列预测[J].计算机时代,2023(5):102-108.
YAN Y ZH, MU N G. Ultra financial time series forecasting based on CEEMDAN-VMD-LSTM [J]. Computer Generation, 2023(5):102-108.

作者简介

朱菊香,副教授,硕士生导师,主要研究方向为自动化及控制技术、检测技术。
E-mail:zjx@cwxu.edu.cn

张诗云,硕士研究生,主要研究方向为智能控制,检测技术。
E-mail:3109412978@qq.com

张涛,硕士研究生,主要研究方向为传感器应用,检测技术。
E-mail:1849173864@qq.com

孙君峰,硕士研究生,主要研究方向为环境感知、检测技术。
E-mail:850007512@qq.com

张赵良(通信作者),硕士,高级工程师,主要研究方向为控制工程、仪器仪表。
E-mail:zhangzl@cwxu.edu.cn