基于特征优化和 ISSA-LSTM 的脱硝系统入口 NO*x* 浓度预测模型

王渊博，金秀章

（华北电力大学　控制与计算机工程学院，河北　保定　071003）

摘　要：针对电厂脱硝系统入口 NO*x* 浓度受较多因素的影响波动较大，且 CEMS 检测仪表有很大迟延难以精准测量的问题，提出了一种基于随机森林算法（RF）和改进麻雀搜索算法（ISSA）优化长短时记忆神经网络（LSTM）的脱硝系统入口 NO*x* 浓度预测模型。首先，通过机理和相关性分析确定与 SCR 入口 NO*x* 质量浓度相关的初始辅助变量，并利用 RF 算法对辅助变量进行特征优化选择，然后通过互信息（MI）对各辅助变量与输出变量之间进行迟延估计并提取时序特征，并通过小波滤波对输入变量进行降噪处理，建立 LSTM 神经网络预测模型。利用 ISSA 算法确定 LSTM 模型的最优组合参数，最后与传统的 LSSVM、RBF、BP 模型的预测结果进行对比。实验结果证明，特征优化后的 ISSA-LSTM 神经网络预测模型的决定系数（R^2）最大，均方根误差（RMSE）和平均绝对百分比误差（MAPE）最小，具备很强的拟合和泛化能力，可以精准预测脱硝系统入口氮氧化物的质量浓度。

关键词：NO*x* 浓度预测；特征优化；互信息；麻雀搜索算法；LSTM 神经网络；随机森林算法

中图分类号：TP183　　**文献标识码**：A　　**DOI**：10. 19358/j. issn. 2097－1788. 2023. 04. 012

引用格式：王渊博，金秀章. 基于特征优化和 ISSA-LSTM 的脱硝系统入口 NO*x* 浓度预测模型［J］. 网络安全与数据治理，2023，42(4)：70－77，84.

Prediction model of NO*x* concentration at the inlet of the denitration system based on feature optimization and ISSA-LSTM

Wang Yuanbo，Jin Xiuzhang

(School of Control and Computer Engineering，North China Electric Power University，Baoding 071003，China)

Abstract：Aiming at the problem that the NO*x* concentration at the inlet of the denitrification system in power plants is greatly affected by many factors and fluctuates greatly，and the CEMS detection instruments have great delays and are difficult to accurately measure，a prediction model for the NO*x* concentration at the inlet of the denitrification system based on the random deep forest algorithm (RF) and the improved sparrow search algorithm (ISSA) optimized long-term and short-term memory neural network (LSTM) was proposed. Firstly，the initial auxiliary variables related to the mass concentration of NO*x* at the SCR inlet were determined by mechanism and correlation analysis，and the auxiliary variables were selected for feature optimization using the RF algorithm，then the delay between each auxiliary variable and the output variables were estimated by mutual information (MI) and the timing features were extracted，and the LSTM neural network prediction model was established by denoising the input variables through wavelet filtering. The modified sparrow search algorithm was used to determine the optimal combination parameters of the LSTM model and finally contrasted with the prediction results of the traditional LSSVM，RBF and BP models. The experimental results proved that the ISSA-LSTM neural network prediction model after feature optimization had the largest coefficient of determination (R^2) and the smallest root mean square error (RMSE) and mean absolute percentage error (MAPE)，which exhibited strong fitting and generalization ability to accurately predict the mass concentration of NO*x* at the inlet of the denitrification system.

Key words：NO*x* concentration prediction；feature optimization；mutual information；sparrow search algorithm；LSTM neural network；random forest

0 引言

为了实现碳中和的目标，我国近年来积极推进能源转型，优化能源结构[1]。根据国家统计局最新公布的数据，2022 年火电的装机容量仍然占比 52% 左右，是我国发电领域中的领头羊[2]。火力发电机组的主要燃料来源是煤炭，而煤炭在燃烧过程中会产生大量的 NOx，NOx 是造成大气污染的主要污染物之一[3]。

当前我国电厂常用的烟气脱硝方法主要分为两种，分别为选择性催化还原（SCR）脱硝系统和选择性非催化还原（SNCR）脱硝系统。两种方法各有优劣，前者具有工艺成熟、安全稳定且脱硝效率超过 90% 等优点，是当前电厂烟气脱硝技术的首选[4]，后者由于脱硝效率低，在烟气脱硝中一般只用作辅助手段。本文研究的燃煤电站采用 SCR 技术对尾部烟气中的氮氧化物进行脱销处理。

由于燃煤电站锅炉燃烧系统是一个具有大延迟、大惯性的非线性系统，SCR 入口 NOx 浓度容易受不同因素的影响而波动较大，使得精准 SCR 入口氮氧化物浓度的获取变得困难，进而很难对喷氨量进行精准的控制。喷氨量过低，脱销效果不好，会造成 NOx 排放不达标；过量喷氨不但影响脱硝效率，又造成巨大的资源消耗，提高运行成本。因此，建立精准有效的脱硝系统 SCR 入口氮氧化物预测模型，不仅可以帮助脱硝系统精准调控喷氨量，提升脱硝品质，又可以降低电厂的脱硝成本。

针对预测模型的建立问题，随着近几年深度学习技术的高速发展，许多先进预测模型被提出。余廷芳等人[5]提出了基于向量机和径向基神经网络的 NOx 预测模型，但在建模过程中未考虑迟延时间的问题。于静等人[6]提出了基于结构改进的 RBF 神经网络预测模型，利用互信息估计辅助变量的延迟时间，利用调整时序后的辅助变量建立模型，提高了模型的泛化能力。刘岳等人[7]提出了利用 LSTM 建立预测模型，并通过改进的粒子群算法对预测模型的参数进行优化，提高了预测模型的精准度。姚宁等人[8]提出利用 Bi-LSTM 建立脱硝系统模型，并通过优化的鲸鱼算法对预测模型进行超参数选取，使收敛因子非线性递减，提高了模型的搜索速度和泛化能力。

除了建模方法，辅助变量特征的选取也对模型的预测精度有很大的影响，邢红涛等人[9]提出了利用 CNN 网络的卷积层和池化层提取辅助变量与目标变量的高维映射关系，构造高维时序特征向量，然后再利用 LSTM 建立预测模型。金秀章等人[10]通过利用互信息筛选出相关性较高的辅助变量，以实现辅助变量的降维。

基于上述研究，本文提出了一种基于特征优化和 ISSA 优化 LSTM 的 SCR 入口氮氧化物浓度预测模型，首先，通过机理和相关性分析确定与 SCR 入口 NOx 浓度相关的初始辅助变量，再利用 RF 算法对初始辅助变量进行特征优化选择，以模型的均方误差（MSE）和拟合优度（R^2）作为评价函数，判定各个辅助变量在决策过程中的重要性程度，筛选出重要性大的辅助变量作为模型输入，然后通过 MI 估计模型各输入变量与目标变量之间的延迟时间并提取时序特征，最后通过小波滤波算法对模型变量进行降噪处理，建立脱硝系统入口 NOx 浓度预测模型。最后利用 ISSA 算法确定 LSTM 神经网络模型的最优组合参数，并与传统的 LSSVM、RBF、BP 神经网络模型进行对比验证。仿真结果表明，经过特征优化和时序调整后的 ISSA-LSTM 预测模型与经典的神经网络预测模型相比预测效果更好，具有很强的泛化能力。

1 辅助变量选择和特征优化

本次研究选用保定某 660 MW 电厂 DCS 系统现场运行 2 天多的 20 万组历史数据，模型输入数据的采样周期为 10 s，共选取 2 万组数据作为模型的初始变量数据集。

1.1 选择初始相关辅助变量

该燃煤电站 SCR 脱硝系统采用高飞尘布置，其工艺流程如图 1 所示。

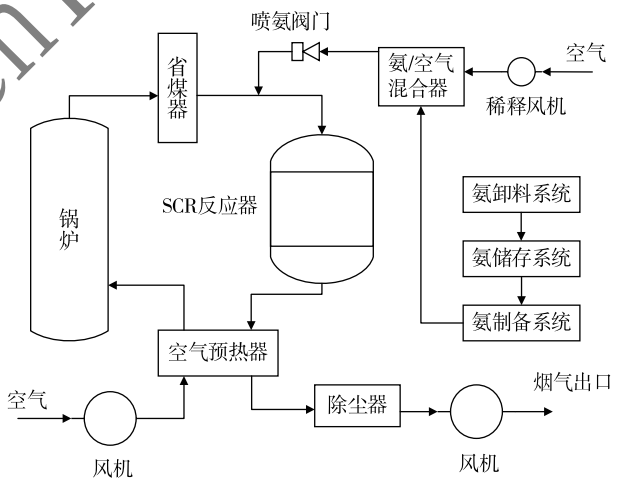

图 1 SCR 脱硝系统工艺流程图

通过图 1 可以看出，燃煤机组 SCR 反应器入口的 NOx 由锅炉产生，并通过省煤器流通到 SCR 反应器上方，与喷入的还原剂进行氧化还原反应，使烟气中的 NOx 通过化学反应生成为 N_2 和 H_2O，进而达到脱硝的目的。该电厂 NOx 质量浓度测点分为甲侧和乙侧且有相同的变化趋势，本次选用甲侧测点进行数据建模分析。通过机理分析可以初步筛选出总风量、一次风量、二次风量、机组负荷、总煤量 、氨流量、SCR 入口烟气流量、SCR 管道压力、SCR 入口压力、SCR 入口烟气温度、主蒸汽流

量、主蒸汽温度、锅炉氧量等 13 个与 SCR 入口 NOx 浓度相关的初始辅助变量。

1.2 基于 RF 的特征优化

由于燃煤机组锅炉燃烧情况复杂多变，DCS 采集的各个辅助变量数据存在严重的耦合和冗余，单纯地机理分析不能满足要求，还需要对各个初始辅助变量进行数学算法分析。

1.2.1 Bagging 算法

随机采样就是在某个样本数据集合中抽取一定数量的数据，然后每抽取一个数据之后，再将之放回原本样本数据集合中。那么假设原始数据集合中有 n 个样本数据，对于任何一个数据集合中的样本，其每次被采集到的概率是 $1/n$。如此这般有放回地从 m 个原始数据集合中随机抽取 n 个样本数据，可以计算原本数据集合中每个样本数据未被抽中的概率为：

$$p = \left(1 - \frac{1}{m}\right)^m \tag{1}$$

当 m 逐渐趋于无穷大时，未被抽中的概率 p 约等于 0.368，即 Bagging 每轮随机采样中，没有被抽取到的样本数据占整个原始样本数据集合的 36.8% 左右，这些数据称为袋外数据（Out Of Bag，OOB），它们可以用做测试集来检测模型的泛化能力。

1.2.2 RF 特征优化

随机森林（RF）是一种综合了 Bagging 算法和决策树算法的机器学习算法。在传统的 Bagging 算法基础上使用 CART 决策树作为弱学习器，相比普通的决策树模型，RF 在每个节点进行分裂时，并非比较所有输入变量（特征），而是随机从中选择 m 个特征并比较其中的最优点进行节点分裂，这个过程为随机特征选取。该方法在决策树模型的训练过程中进一步保证了每棵决策树模型的差异性和多样性。

通过变量置换法可以得出各个辅助变量的变量重要性评分（Variable Importance Measure，VIM），计算过程如下：

（1）用数据样本建立随机森林模型，评估所有袋外样本的错误率。

（2）对所有袋外样本的辅助变量 K 进行置换调整，得到新的袋外数据（OOB′），评估所有袋外数据样本的错误率。

（3）计算两次袋外数据样本的错误率变化值，并将所有的袋外数据样本的错误率变化均值作为变量 K 的变量重要性评分[11]。

变量 K 的 VIM 计算如下：

$$\mathrm{VIM}_K^{\mathrm{ER}} = \frac{1}{N_{\mathrm{tree}}}\sum_{t=1}^{N_{\mathrm{tree}}}(\mathrm{ER}_{kt} - \mathrm{ER}'_{kt}) \tag{2}$$

式中：N_{tree}为决策树的个数，ER_{kt}为变量 K 置换之前第 t 棵树对应的错误率，ER'_{kt}为变量 K 置换之后第 t 棵树对应的错误率。

如果变量 K 与预测变量无关，随机置换该变量后对应的袋外数据错误率不会发生改变，理论上 $\mathrm{VIM}_K^{\mathrm{ER}} = 0$；反之，如果 $\mathrm{VIM}_K^{\mathrm{ER}} > 0$，则说明变量 K 与预测变量有一定相关性。即 VIM 越小，变量对预测结果的重要性越低；相反，VIM 越大，变量的重要性越高。因此可以剔除重要性低的辅助变量，对变量特征进行优化选择。

从 20 000 万组初始数据集中随机抽取 3 000 组连续数据进行辅助变量特征优化处理，选取 3 次，最后的变量重要性评分取三次结果的平均值，图 2 为运算结果。

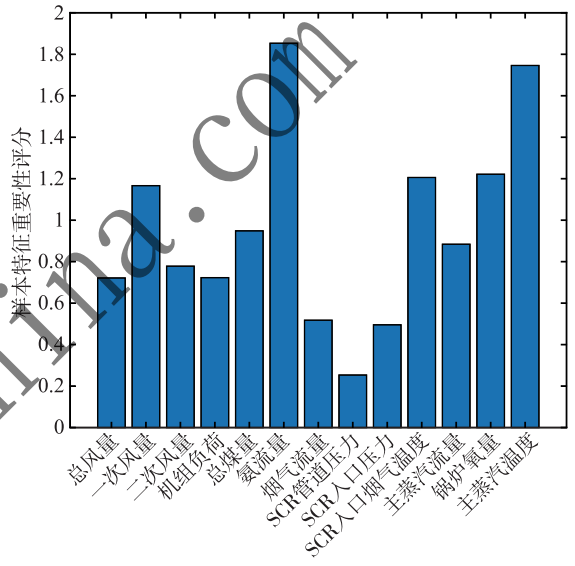

图 2　样本特征重要性评分分布

从图 2 可以看出，烟气流量、SCR 入口压力、SCR 管道压力的 VIM 值最低且均小于 0.7，与目标变量关联性不大，故选取总风量、一次风量、二次风量、机组负荷、总煤量 、氨流量、SCR 入口烟气温度、主蒸汽流量、锅炉氧量、主蒸汽温度等 10 个辅助变量作为模型的输入变量。

1.3 基于互信息的时序校正

由图 1 的 SCR 脱硝系统工艺流程可以看出，锅炉燃烧之后产生的 NOx 混在烟气中，需要经过省煤器以及很长的管道才能到达 SCR 反应器入口，所以 SCR 入口 NOx 浓度与各辅助变量之间存在很大的时间延迟，于是提出了一种基于最大互信息的迟延分析方法。两变量 x、y 彼此之间的互信息 I（x；y）定义如下[12]：

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3}$$

式中：p（x）和 p（y）分别为 x 和 y 的概率密度分布函

数；p（x，y）表示 x 和 y 的联合概率密度分布函数。

电厂锅炉燃烧一个周期为 10 min 左右，因此变量的最大延迟时间不超过 600 s。因为变量的采样周期为 10 s，所以评估得到的迟延时间为 10 的整数倍。分别计算一个周期内各个输入变量与输出变量之间的互信息，将互信息最大的时刻当作变量之间的迟延时间[13]，对模型输入变量进行迟延补偿，实现输入变量与目标输出变量之间的时序统一。输入变量与目标变量之间的最大互信息和最优延迟时间如表 1 所示。

表 1　输入变量的最大互信息及其延迟时间

变量单位	最大互信息	延迟时间/s
总风量	0.859 6	320
一次风量	0.873 4	320
二次风量	0.874 6	320
机组负荷	0.798 8	300
总煤量	0.874 1	320
氨流量	0.880 7	50
SCR 入口烟气温度	0.862 0	60
主蒸汽流量	0.831 6	320
锅炉含氧量	0.812 1	300
主蒸汽温度	0.802 0	240

1.4　基于小波阈值的降噪

小波阈值滤波算法是一种对含噪信号进行降噪处理的方法。该方法的主要思想是先通过小波分解算法对包含噪声的数据样本进行变换处理，然后通过计算得到小波分解系数和阈值函数，将样本数据中绝对值低于阈值的小波系数视为含噪信息，将其进行置零处理；相反，将样本数据中绝对值高于阈值的小波系数视为有效信息予以保留，从而达到数据降噪的目的。

常用的小波阈值函数有两种，分别为硬阈值法和软阈值法，硬阈值能够对数据变量的局部特性进行比较好的保护，软阈值对数据变量的处理会相对平滑一些，但是数据的边缘处理会相对模糊。

硬阈值函数的数学表示式为：

$$W_{\text{new}}=\begin{cases}w, & |w|\geqslant T\\ 0, & |w|<T\end{cases} \tag{4}$$

软阈值函数的数学表示式为：

$$W_{\text{new}}=\begin{cases}\text{sign}(w)(|w|-T), & |w|\geqslant T\\ 0, & |w|<T\end{cases} \tag{5}$$

固定阈值的数学表示式为：

$$H=\sigma\sqrt{2\log(N)} \tag{6}$$

$$\sigma=\text{MAD}/0.674\ 5 \tag{7}$$

式中：W 为小波系数，H 为阈值，σ 为数据信号的噪声标准差，MAD 为小波系数幅度的中值，N 为数据信号的序列长度。

图 3 为总风量与氨流量经过小波滤波降噪后的数值对比。

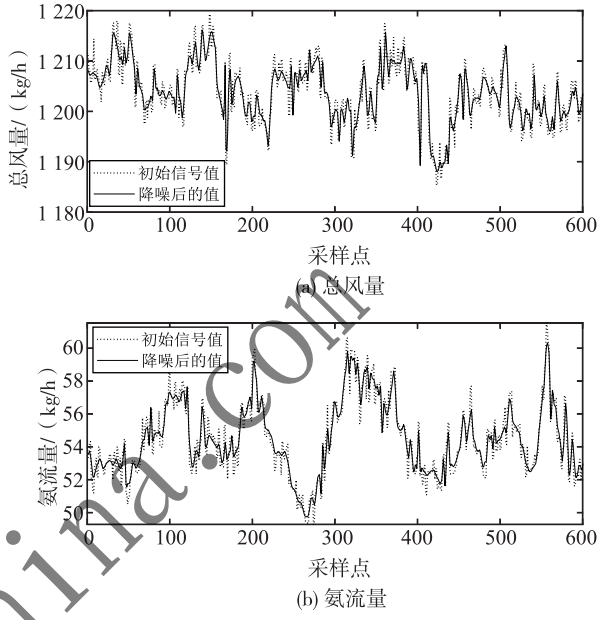

图 3　小波滤波降噪效果

2　ISSA-LSTM 网络预测模型

2.1　LSTM 网络预测模型

长短时记忆网络本质上就是一种特殊的 RNN 神经网络，由德国学者于 20 世纪 90 年代提出[14]，其特点是引入了门控循环单元机制，增加了长短记忆功能，有效地解决了传统循环神经网络的梯度爆炸或消失问题。

LSTM 神经网络拥有三个门控循环单位，分别为遗忘门、输入门和输出门，分别用 $\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{O}_t$ 表示，三个门控制着神经元信号的传递，目的是对输入信号和记忆状态的更新和遗忘进行调控。图 4 为 LSTM 神经网络的基本结构图。

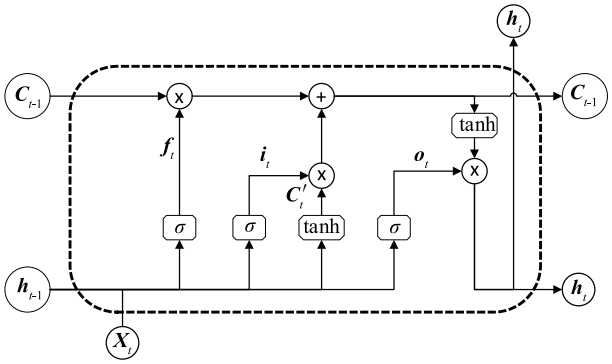

图 4　LSTM 神经网络的基本结构图

LSTM 神经网络的训练过程如下：

（1）遗忘门 $\boldsymbol{f}_t$。遗忘门的作用是对上一时刻得到的记忆单元中信息 $\boldsymbol{C}_{t-1}$ 进行选择性保存，确保重要的特征信息始终保留，遗忘那些不太重要的信息，具体数学表达式如下：

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{W}_f\left[\boldsymbol{h}_{t-1},\ \boldsymbol{x}_t\right]+\boldsymbol{b}_f\right) \tag{8}$$

式中：$\boldsymbol{W}_f$ 为权重矩阵，$\boldsymbol{b}_f$ 为偏置向量。

通过遗忘门的定义可以看出，LSTM 神经网络在 t 时刻的输入信息不仅仅是当前输入 $\boldsymbol{X}_t$，还包括上一时刻的长期记忆信息 $\boldsymbol{C}_{t-1}$ 以及上一时刻的短时记忆信息 $\boldsymbol{h}_{t-1}$，遗忘门控单元的作用就是对长期记忆的信息 $\boldsymbol{C}_{t-1}$ 进行筛选并为当前时刻新输入的信息预留空间。

（2）输入门 $\boldsymbol{i}_t$。输入门的作用是对当前时刻的输入信息 $\boldsymbol{X}_t$ 以及上一时刻得到的短期记忆信息 $\boldsymbol{h}_{t-1}$ 进行选择性保存，并更新细胞状态信息，数学表达式如下：

$$\boldsymbol{i}_t=\sigma\left(\boldsymbol{W}_i\left[\boldsymbol{h}_{t-1},\ \boldsymbol{x}_t\right]+\boldsymbol{b}_i\right) \tag{9}$$

$$\boldsymbol{C}'_t=\tanh\left(\boldsymbol{W}_c\left[\boldsymbol{h}_{t-1},\ \boldsymbol{x}_t\right]+\boldsymbol{b}_c\right) \tag{10}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{C}'_t \tag{11}$$

式中：$\boldsymbol{W}_i$、$\boldsymbol{W}_c$ 为权重矩阵，$\boldsymbol{b}_i$、$\boldsymbol{b}_c$ 为偏置向量，⊙为矩阵相乘。

（3）输出门 $\boldsymbol{O}_t$。输出门的作用是对当前时刻细胞单元的状态值进行更新，并对下一时刻的输入信息进行确认，具体数学表达式如下：

$$\boldsymbol{O}_t=\sigma\left(\boldsymbol{W}_o\left[\boldsymbol{h}_{t-1},\ \boldsymbol{x}_t\right]+\boldsymbol{b}_o\right) \tag{12}$$

$$\boldsymbol{h}_t=\boldsymbol{O}_t\odot\tanh\left(\boldsymbol{C}_t\right) \tag{13}$$

式中：$\boldsymbol{W}_o$ 为权重矩阵，$\boldsymbol{b}_o$ 为偏置向量。

2.2 改进 SSA 算法

为了提高模型的预测精度，需要对预测模型的各个参数进行优化算法寻优。

2.2.1 麻雀搜索算法（SSA）

麻雀搜索算法（SSA）是 Xue 和 Shen 于 2020 年提出的一种新型群体智能优化算法，其主要思想是依据麻雀的觅食行为和反捕食行为[15]。由于结构简单、易于实现，且控制参数较少，因此麻雀优化算法在收敛速度以及探寻全局最优等方面具有较大的优势。

SSA 算法将捕食中的麻雀分为发现者和跟随者两类。发现者自身拥有较高适应度以及较广的搜索范围，能够对种群的搜索和觅食提供指引，跟随者则紧跟发现者并通过与其争夺来获得食物，同时，为了提高自身的生存环境，部分跟随者会对发现者进行监视，以便进行争夺食物或在其周围进行觅食。与此同时，当天敌临近或意识到危险的时候，种群会及时进行反捕食行为。SSA 算法的具体步骤如下：

（1）初始化麻雀种群位置、最大迭代次数。麻雀种群可表示为：

$$\boldsymbol{X}=\begin{bmatrix}x_{1,1} & x_{1,2} & \cdots & x_{1,d}\\ x_{2,1} & x_{2,2} & \cdots & x_{2,d}\\ \cdots & \cdots & \cdots & \cdots\\ x_{n,1} & x_{n,2} & \cdots & x_{n,d}\end{bmatrix} \tag{14}$$

式中：n 为麻雀种群个数，d 为优化参数维数。

（2）计算每只麻雀的适应度值，并按照适应度值大小降序排列，选取当前最佳适应度值 f_g 和最差适应度值 f_w 以及对应的位置 X_{best} 和 X_{worst}。所有麻雀的适应度值可表示为：

$$\boldsymbol{f}_x=\begin{bmatrix}f\left[x_1^1 & x_1^2 & \cdots & x_1^d\right]\\ f\left[x_2^1 & x_2^2 & \cdots & x_2^d\right]\\ \cdots\\ f\left[x_n^1 & x_n^2 & \cdots & x_n^d\right]\end{bmatrix} \tag{15}$$

（3）更新发现者位置，更新公式如下：

$$X_{i,j}^{t+1}=\begin{cases}X_{i,j}^t\cdot\exp\left(-\dfrac{i}{\alpha\cdot\text{iter}_{\max}}\right), & R_2<\text{ST}\\ X_{i,j}^t+Q\cdot\boldsymbol{L}, & R_2\geqslant\text{ST}\end{cases} \tag{16}$$

式中：$X_{i,j}$ 表示第 i 只麻雀在第 j 维中的位置，t 代表当前迭代数，$j=1,\ 2,\ 3,\ \cdots,\ d$；$\text{iter}_{\max}$ 为最大迭代次数，α 为［0，1］之间随机分布的数，Q 为服从标准正态分布的随机数；$\boldsymbol{L}$ 表示元素为 1、大小为 $1\times d$ 的矩阵；$R_2\in$［0，1］表示预警值；$\text{ST}\in[0.5,\ 1]$ 表示安全值。

当 $R_2<\text{ST}$ 时，种群未察觉天敌的存在或未意识到危险，生存环境比较安全，发现者可以在广泛的空间范围内进行搜索和捕食，并引导种群内其他麻雀捕食，使种群获得更高的适应度；当 $R_2\geqslant\text{ST}$ 时，种群中麻雀意识到危险或察觉捕食者的存在，并向种群传递危险信号，种群立即调整搜索策略并迅速向安全区域靠拢。

（4）更新跟随者位置。在觅食过程中，跟随者为了获得更优的适应度，部分跟随者会对发现者进行监控，当发现者搜索到更好的食物，跟随者会与其进行抢夺，若抢夺成功，则跟随者随机得到该发现者的食物，若抢夺失败，则跟随者位置更新如下：

$$X_{i,j}^{t+1}=\begin{cases}Q\cdot\exp\left(\dfrac{X_{\text{worst}}^t-X_{i,j}^t}{i^2}\right), & i>\dfrac{n}{2}\\ X_p^{t+1}+\left|X_{i,j}^t-X_p^{t+1}\right|\cdot\boldsymbol{A}^+\boldsymbol{L}, & \text{other}\end{cases} \tag{17}$$

式中：X_{worst}^t 表示第 t 次迭代中适应度最差的麻雀位置，X_p^{t+1} 表示发现者所拥有的最佳位置，$\boldsymbol{A}$ 表示一个 $1\times d$ 的矩阵，其元素都为［－1，1］之间的随机数，同时满足 $\boldsymbol{A}^+=\boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\right)^{-1}$。当 $i>n/2$ 时，意味着第 i 只跟随者的适应度值较低，说明这只麻雀的生存环境很差，它需要向其他位置移动来寻找食物。

（5）更新麻雀种群中的警戒者，麻雀中警戒者数量为种群数量的 10% ~20%，其初始位置是在种群中随机产生，其位置更新的表达式如下：

$$X_{i,j}^{t+1}=\begin{cases}X_{\text{best}}^{t}+\gamma\cdot\left|X_{i,j}^{t}-X_{\text{best}}^{t}\right|, & f_i>f_g\\ X_{i,j}^{t}+K\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{\text{worst}}^{t}\right|}{(f_i-f_w)+\varepsilon}\right), & f_i=f_g\end{cases}\tag{18}$$

式中：γ 表示步长控制参数，是服从均值为 0 方差为 1 的正态分布的随机数，K 为一个［−1，1］之间的随机数，f_i 表示当前麻雀的适应度值，ε 为一个防止分母为零的常数。

当 $f_i>f_g$ 时，说明该麻雀此时位于种群的边缘，适应度差，非常容易受到天敌的袭击，需要向安全区域转移才能得到更好的觅食位置；当 $f_i=f_g$ 时，表示在种群中间的麻雀察觉到了危险，为了躲避天敌的捕食，需要及时向其他麻雀位置移动。

（6）重新计算麻雀种群的适应度值，更新所有麻雀位置，并更新种群最佳适应度值 f_g 和最差适应度值 f_w 以及对应的位置 X_{best} 和 X_{worst}。

（7）判断是否满足停止条件，若满足则退出，输出结果，否则重复执行步骤(3)~(6)。

2.2.2 基于 Levy 飞行的改进麻雀搜索算法（ISSA）

虽然 SSA 算法有寻优能力强、收敛速度快等优点，但是当发现者经过一定次数的迭代且适应度值不变时，此时跟随者就成了发现者，在搜索全局最优时，容易陷入局部最优。为避免此问题，在跟随者更新公式中引入 Levy 飞行策略，提高全局搜索能力[16]。改进后跟随者位置更新的数学表达式如下：

$$X_{i,j}^{t+1}=\begin{cases}Q\cdot\exp\left(\dfrac{X_{\text{worst}}^{t}-X_{i,j}^{t}}{i^2}\right), & i>\dfrac{n}{2}\\ X_{p}^{t+1}+X_{p}^{t+1}\oplus\text{Levy}(d), & \text{other}\end{cases}\tag{19}$$

Levy 飞行的计算公式为：

$$\text{Levy}(d)=\frac{\mu}{|v|^{1/\beta}}\tag{20}$$

$$\begin{cases}\mu\sim N(0,\sigma_\mu^2)\\ v\sim N(0,\sigma_v^2)\end{cases}\tag{21}$$

$$\begin{cases}\sigma_\mu=\left\{\dfrac{\Gamma(1+\beta)\cdot\sin(\pi\beta/2)}{\Gamma[(1+\beta)/2]\cdot\beta\cdot2^{\frac{\beta-1}{2}}}\right\}^{1/\beta}\\ \sigma_v=1\end{cases}\tag{22}$$

式中：β 为取值范围是［0，2］的常数，参数 μ、v 为符合式（21）的正态分布随机数，式（22）为其所对应的正态分布的标准差的计算公式。

2.3 基于 ISSA-LSTM 网络的预测模型

将原始数据进行变量选取和特征优化后，再对数据进行归一化作为模型的输入变量，图 5 为模型总体算法流程。

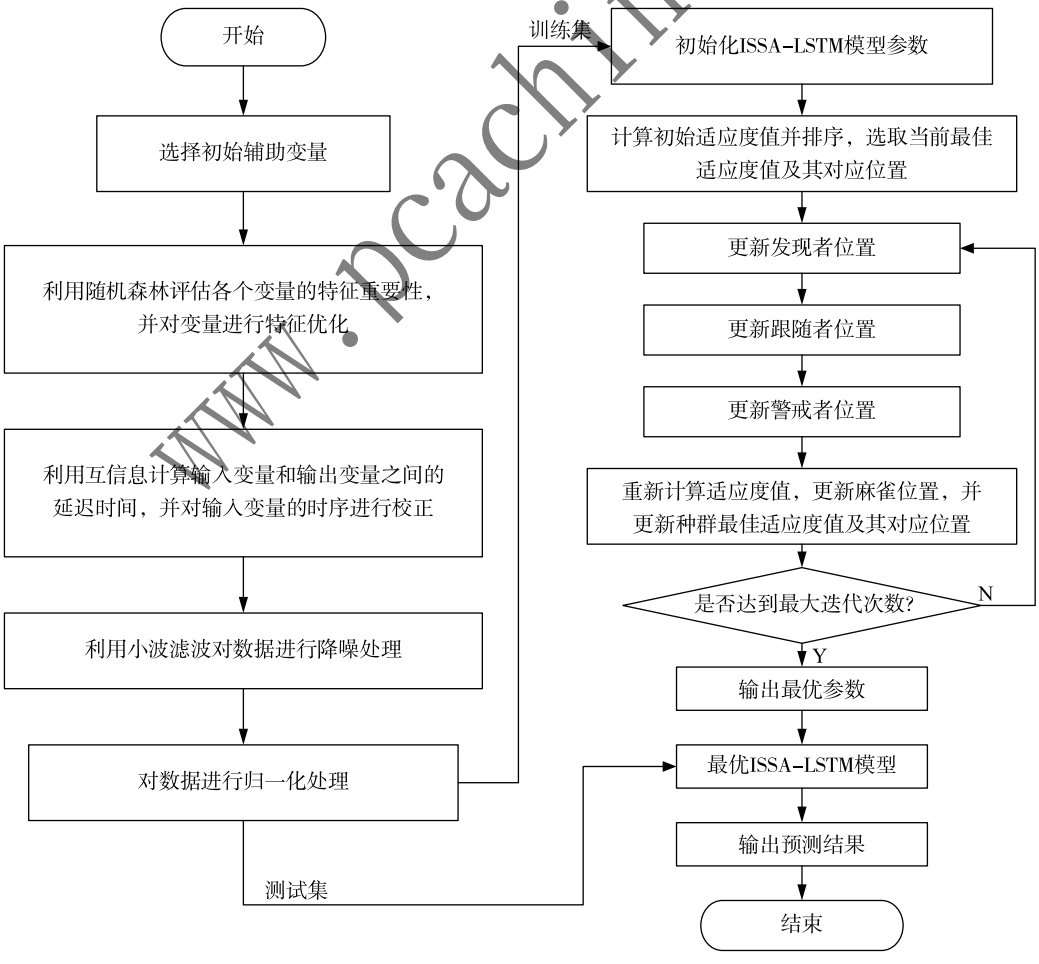

图 5　基于 ISSA-LSTM 预测模型的算法流程

3 实验设计与结果分析

本次实验的硬件配置是 Intel（R）Core（TM）i5 + Windows10 系统，软件配置为 MATLAB2019b。因为 SCR 入口 NOx 浓度受时间影响较大以及 LSTM 神经网络模型的固有特性，所以加入上一时刻 SCR 入口 NOx 浓度作为输入变量，最终选取总风量、机组负荷、总煤量、氨流量、锅炉含氧量、上一时刻 SCR 入口 NOx 浓度等 11 个辅助变量作为模型的输入变量。为验证模型的准确性，从 20 000 组电厂运行 DCS 数据中选取了 8 000 组连续的包含有稳定工况和变工况的 DCS 数据。其中训练集 5 000 组，验证集 2 000 组，测试集 1 000 组。

3.1 模型的评价指标

本文模型评价指标为均方根误差 δ_{RMSE}、平均绝对误差 δ_{MAPE} 和决定系数 R^2。δ_{RMSE} 和 δ_{MAPE} 越小表示模型预测精确度越高；R^2 越接近 1 表示模型预测适应程度越好，与真实值变化趋势越接近。其计算公式如下：

$$\delta_{\mathrm{RMSE}} = \sqrt{\frac{\sum_{i=1}^{M}(y_i - \hat{y}_l)^2}{M}} \tag{23}$$

$$\delta_{\mathrm{MAPE}} = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{y_i - \hat{y}_l}{y_i}\right| \times 100\% \tag{24}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \hat{y}_l)^2}{\sum_{i=1}^{M}(y_i - \bar{y}_i)^2} \tag{25}$$

式中：y_i 为真实值，$\hat{y}_l$ 为预测值，$\bar{y}_i$ 为真实值序列的均值，M 为样本数量。

3.2 模型的参数选取

LSTM 模型的隐含层参数采用自适应运动估计（Adaptive Moment Estimation，Adam）的方法进行梯度迭代。最大训练轮数为 250，梯度阈值为 1，L2 正则化参数为 0.001，LSTM 模型的参数包含两层隐含层的层数和初始学习率 3 个超参数。两个隐含层层数的初始范围设为［1，100］，初始学习率的初始范围设为［0.001，0.1］，通过改进麻雀算法对超参数进行优化，得到最优超参数分别为 15、8 和 0.01。

3.3 实验结果分析

3.3.1 模型输入变量特征优化对预测结果的影响

将经过特征优化处理后的数据与未进行特征处理的原始数据分别进行模型的训练和预测，得到不同特征选择下的预测结果如图 6、图 7 所示，表 2 为不同特征训练集的模型预测精度。

由图 6、图 7 和表 2 可以看出，经过特征优化后的数据集比原始数据的 RMSE 降低了 9.68%，MAPE 降低了 12.76%，R^2 提高了 1.28%，说明冗余变量的存在，降低了模型的精度和泛化能力。经过时序校正后的数据集比未经过时序校正的数据的 RMSE 降低了 12.97%，MAPE 降低了 5.66%，R^2 提高了 1.36%，说明时序校正能提高模型的精度。经过滤波处理后的数据集比未经过滤波处理的数据的 RMSE 降低了 14.19%，MAPE 降低了 14.37%，R^2 提高了 1.10%，说明噪声对模型的精度有很大影响。对输入变量进行特征优化和时序、滤波处理能有效提升模型的预测精度和泛化能力。

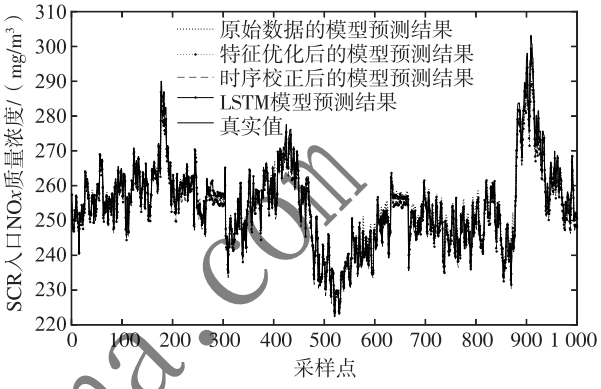

图 6 变量特征优化对模型预测结果的影响

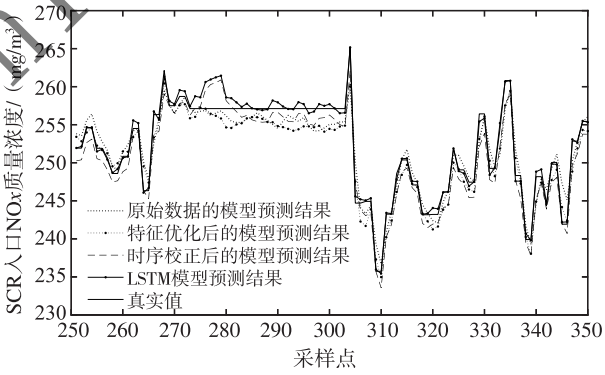

图 7 部分序列变量特征优化对模型预测结果的影响

表 2 特征优化对模型预测结果的评价指标对比

项目	δ_{RMSE}	δ_{MAPE}	R^2
LSTM 预测模型	2.119 7	0.650 5	0.970 3
时序校正的 LSTM 预测模型	2.470 2	0.759 7	0.959 7
特征优化的 LSTM 预测模型	2.838 3	0.805 3	0.946 8
原始数据的 LSTM 预测模型	3.142 6	0.923 1	0.934 8

3.3.2 不同预测模型对预测结果的影响

为了对比验证 LSTM 神经网络预测模型的特点，分别利用 BP、RBF 和 LSSVM 这 3 种具有代表性的建模方法搭建了预测模型，模型的最优超参数通过 ISSA 算法寻优确定。图 8、图 9 和表 3 为几种不同模型的预测结果和模型的预测精度。

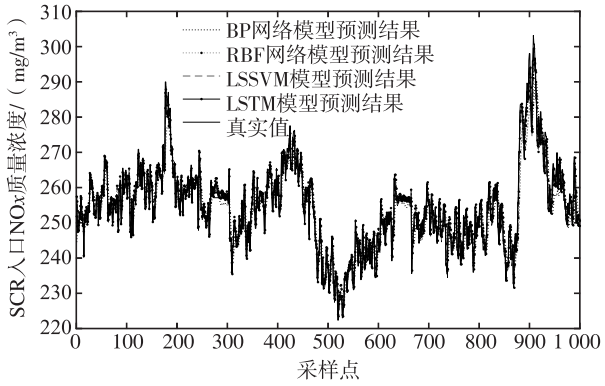

图 8　不同预测模型对预测结果的影响

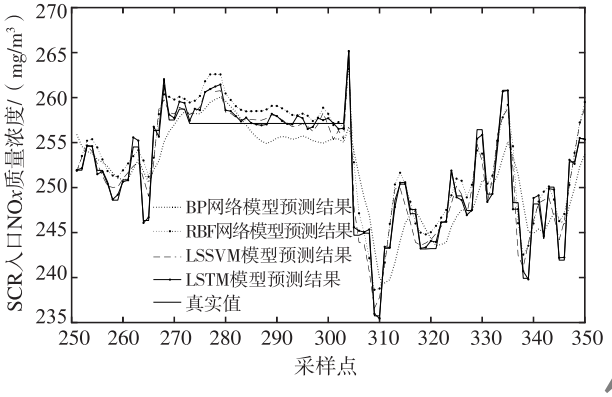

图 9　部分序列不同预测模型对预测结果的影响

表 3　不同模型对模型预测结果的评价指标对比

项目	δ_{RMSE}	δ_{MAPE}	R^2
LSTM	2. 119 7	0. 650 5	0. 970 3
LSSVM	2. 827 6	0. 876 1	0. 947 2
RBF	3. 524 1	1. 055 8	0. 917 9
BP	4. 326 3	1. 271 0	0. 876 4

由图 8、图 9 和表 3 可以看出，从模型的预测精准度来看，LSTM 神经网络模型的预测精准度最好，其均方根误差 RMSE 和 MAPE 均比传统的 LSSVM 模型、RBF 神经网络模型和 BP 神经网络模型更低；从模型的相关性来看，对比其他几种模型，LSTM 神经网络模型的决定系数 R^2 最大，也表明了 LSTM 神经网络模型的预测值与真实值更接近，模型适应程度更好。

在其他条件一致的情况下，LSSVM 模型能够跟踪入口 NO*x* 质量浓度的变化趋势，但是存在一定的偏差；RBF 神经网络模型和 BP 神经网络模型的预测结果震荡严重。

综上所述，基于 ISSA-LSTM 的预测模型相比于传统的 LSSVM 模型、RBF 神经网络模型和 BP 神经网络模型具有更强的学习能力、预测精度和泛化能力。

4　结论

本文针对燃煤电厂 SCR 反应器入口处 NO*x* 质量浓度受较多不同因素的影响波动较大，且 CEMS 检测仪表有很大迟延难以精准测量的问题，提出了一种基于特征选择和 ISSA-LSTM 神经网络的脱硝系统入口 NO*x* 浓度预测模型。首先利用机理分析筛选出与目标变量相关的初始辅助变量，然后利用随机森林算法对模型输入变量进行特征优化选择，再使用互信息对辅助变量进行延迟计算和时序校正，并采用小波滤波对模型输入数据进行降噪处理，最后利用基于 Levy 飞行的改进麻雀算法对模型的超参数寻优，提升了算法的全局搜索能力，并对预测模型的预测能力有明显的提高。基于电厂运行数据的实验结果表明，经过特征优化的输入变量，剔除了冗余变量，提高了模型的泛化能力，经过时序校正和降噪处理的输入变量，有效提升了模型的预测精度。与传统 LSSVM 模型和 BP、RBF 神经网络模型相比，ISSA-LSTM 模型的预测误差更小，精度更高，泛化能力更强。

参考文献

[1] 习近平. 在第七十五届联合国大会一般性辩论上的讲话［N］. 人民日报，2020－09－23（2）.

[2] 国家统计局. 中华人民共和国 2022 年国民经济和社会发展统计公报［N］. 人民日报，2022－02－28（10）.

[3] 朱法华，许月阳，孙尊强，等. 中国燃煤电厂超低排放和节能改造的实践与启示［J］. 中国电力，2021，54（4）：1－8.

[4] 牛玉广，潘岩，李晓彬. 火力发电厂烟气 SCR 脱硝自动控制研究现状与展望［J］. 热能动力工程，2019，34（4）：1－9.

[5] 余廷芳，张浩杰. 基于 SVM 和 RBF 神经网络的 CFB NO*x* 生成预测模型[J]. 计算机仿真，2020，37（9）：209－213，316.

[6] 于静，金秀章，刘岳. 基于结构改进 RBF 神经网络的 NO*x* 预测模型比较［J/OL］. 控制工程：1－8［2022－02－22］. https：//doi. org/10. 14107/j. cnki. kzgc. 20210150.

[7] 刘岳，于静，金秀章. 基于特征优化和改进长短期记忆神经网络的 NO*x* 质量浓度预测［J］. 热力发电，2021，50（7）：162－169.

[8] 姚宁，金秀章，李阳峰. 基于改进鲸鱼算法优化 Bi-LSTM 的脱硝系统 NO*x* 建模［J］. 华北电力大学学报（自然科学版），2022，49（6）：76－83.

[9] 邢红涛，郭江龙，刘书安，等. 基于 CNN-LSTM 混合神经网络模型的 NO*x* 排放预测［J］. 电子测量技术，2022，45（2）：98－103.

[10] 金秀章，于静，刘岳. 基于人工鱼群-径向基神经网络的 NO*x* 预测模型［J］. 动力工程学报，2021，41（7）：551－557.

（下转第 84 页）

的自然资源要素综合观测平台构建［J］. 资源科学，2020，42（10）：1965－1974.
［14］夏红军，安燕娜. 数据中台视角下供电企业数据资产管理模型构建［J］. 情报科学，2021，39（10）：70－75.
［15］施俊君. 基于智能运维的城市轨道交通轨道专业数据中台［J］. 城市轨道交通研究，2021，24（S1）：105－107，112.
［16］杨进. 基于数据中台和 GIS 的可视化固定资产管理模式探析［J］. 财务与会计，2021（3）：70－72.
［17］张雯，周子航，周明升. 基于物联网和人工智能的园区安全运营管理平台［J］. 计算机时代，2023（2）：132－136.
［18］雷鸣，姜罕盛，武国良，等. 基于 HBase 的大数据架构下负载平衡技术［J］. 计算机与现代化，2021（6）：91－95.
［19］周明升，韩冬梅. 上海自贸区金融开放创新对上海的经济效应评价——基于“反事实”方法的研究［J］. 华东经济管理，2018，32（8）：13－18.
［20］韩冬梅，周明升. 上海自贸区金融开放创新的宏观效应模拟［J］. 统计与决策，2019，35（9）：155－159.
［21］周明升，韩冬梅. 基于 Rossle 混沌平均互信息特征挖掘的网络攻击检测算法［J］. 微型机与应用，2016，35（14）：1－4.

（收稿日期：2023－01－05）

作者简介：

张雯（1983－），女，硕士，经济师，主要研究方向：大数据分析和预测。

周明升（1981－），男，博士，高级工程师，主要研究方向：智慧城市、决策支持。

（上接第 61 页）

［9］陈晓安. 计算机网络入侵检测系统的研究［J］. 电子测试，2021(18)：76－77，73.
［10］NAM K，KIM K. A study on SDN security enhancement using open source IDS/IPS Suricata［C］//2018 International Conference on Information and Communication Technology Convergence（ICTC）；2018：1124－1126.
［11］WONG K，DILLBAUGH C，SEDDIGH N，et al. Enhancing Suricata in－trusion detection system for cyber security in SCADA networks［C］//2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering（CCECE）. IEEE，2017：1－5.
［12］REN H，NIAN M. Dpdk-based high-speed packet acquisition method［J］. Computer System Applications，2018，27(6)：242－245.
［13］凌质亿. 面向高速网络环境的实时入侵检测系统的研究与实现［D］. 南京：东南大学，2016.
［14］李毅飞，杨进. 一种基于平衡二叉树的 CDP 数据备份及重构方法［J］. 数据通信，2019(2)：13－17.

（收稿日期：2023－03－07）

作者简介：

宗学军（1970－），男，硕士，教授，主要研究方向：工业过程控制、工业信息安全等。

刘欢欢（1997－），通信作者，男，硕士研究生，主要研究方向：工业信息安全。E-mail：1965991358@ qq. com。

何戡（1978－），男，硕士，副教授，主要研究方向：工业过程控制、机器学习等。

（上接第 77 页）

［11］张晓凤，侯艳，李康. 基于 AUC 统计量的随机森林变量重要性评分的研究［J］. 中国卫生统计，2016，33(3)：537－540，542.
［12］Xiang Xinrong，Jin Baisuo，Wu Yuehua. Change-point detection in a high-dimensional multinomial sequence based on mutual information［J］. Entropy，2023，25(2).
［13］李悦，唐振浩，曹生现，等. 基于动态时延分析和典型样本筛选的 NO*x* 排放浓度预测［J/OL］. 中国电机工程学报：1－10［2023－02－06］. https：//doi. org/10. 13334/j. 0258－8013. pcsee. 213189.
［14］HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997（9）：1735－1780.
［15］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.
［16］毛清华，张强，毛承成，等. 混合正弦余弦算法和 Lévy 飞行的麻雀算法［J］. 山西大学学报（自然科学版），2021，44(6)：1086－1091.

（收稿日期：2023－02－23）

作者简介：

王渊博（1993－），男，硕士研究生，主要研究方向：先进控制策略在大型火电机组的应用。

金秀章（1969－），男，副教授，主要研究方向：先进控制策略在大型火电机组的应用和信息融合技术等。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部