doi:10.6041/j.issn.1000-1298.2022.01.029

基于 LightGBM－SSA－ELM 的新疆羊舍 CO_2 浓度预测

尹　航[1]　吕佳威[1,2]　陈耀聪[1,3]　岑红蕾[4]　李景彬[4]　刘双印[1,5]

(1. 仲恺农业工程学院信息科学与技术学院，广州 510225；2. 北京市农林科学院信息技术研究中心，北京 100097；3. 广东省农产品安全大数据工程技术研究中心，广州 510225；4. 石河子大学机械电气工程学院，石河子 832003；5. 仲恺农业工程学院广州市农产品质量安全溯源信息技术重点实验室，广州 510225)

摘要：为减少肉羊集约化养殖过程中因环境恶化产生的应激反应，精准调控 CO_2 质量浓度，提出了基于分布式梯度提升框架(LightGBM)、麻雀搜索算法(SSA)融合极限学习机(ELM)的 CO_2 质量浓度预测模型。首先利用 LightGBM 筛选出与 CO_2 质量浓度相关的重要特征，降低预测模型的输入维度；然后选择 Sigmoid 为激活函数，使用具有较强非线性处理能力的单隐含层 ELM 神经网络算法构建 CO_2 质量浓度预测模型；最后通过麻雀智能优化算法对 ELM 模型中所需要的超参数进行优化，并将优化后模型应用于新疆玛纳斯集约化肉羊养殖基地。试验结果表明，该模型预测均方根误差(RMSE)、平均绝对误差(MAE)和决定系数(R^2)分别为 0.021 3 mg/L、0.013 6 mg/L 和 0.988 6，综合性能指标优于支持向量回归(SVR)、反向传播神经网络(BPNN)、长短记忆神经网络(LSTM)、门限循环单元(GRU)和 LightGBM 等；CO_2 质量浓度预测曲线贴近真实曲线，具有良好的预测效果，能有效满足集约化肉羊养殖过程中 CO_2 质量浓度精准预测及调控要求。

关键词：羊舍；集约化养殖；CO_2 质量浓度预测；极限学习机；麻雀搜索算法；分布式梯度提升框架

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1000-1298(2022)01-0261-10　　**OSID：**

Prediction of CO_2 Concentration in Xinjiang Breeding Environment of Mutton Sheep Based on LightGBM－SSA－ELM

YIN Hang[1]　LÜ Jiawei[1,2]　CHEN Yaocong[1,3]　CEN Honglei[4]　LI Jingbin[4]　LIU Shuangyin[1,5]

(1. *College of Information Science and Technology*, *Zhongkai University of Agriculture and Engineering*, *Guangzhou* 510225, *China*
2. *Information Technology Research Center*, *Beijing Academy of Agricultural and Forestry Sciences*, *Beijing* 100097, *China*
3. *Guangdong Provincial Agricultural Products Safety Big Data Engineering Technology Research Center*, *Guangzhou* 510225, *China*
4. *College of Mechanical and Electric Engineering*, *Shihezi University*, *Shihezi* 832003, *China*
5. *Guangzhou Key Laboratory of Agricultural Products Quality and Safety Traceability Information Technology*, *Zhongkai University of Agriculture and Engineering*, *Guangzhou* 510225, *China*)

Abstract: Air quality plays an important role in mutton sheep breeding environment, in order to reduce the stress response of CO_2 to the growth of large-scale mutton sheep and ensure the healthy growth of mutton sheep in the appropriate environment, the key is to accurately control the CO_2 in the mutton sheep breeding environment. A CO_2 prediction model of mutton sheep breeding environment was proposed based on light gradient boosting machine (LightGBM), sparrow search algorithm (SSA) and extreme learning machine (ELM). Firstly, LightGBM was used to screen out the important characteristics of carbon dioxide concentration and reduce the input dimension of the prediction model. Then, ELM neural network algorithm with single hidden layer with strong nonlinear processing ability was used to build the CO_2 prediction model. Finally, through the sparrow intelligent optimization algorithm, the super parameters needed in ELM model were optimized to obtain the best prediction model. The prediction

收稿日期：2021－07－15　修回日期：2021－11－12

基金项目：国家自然科学基金项目(61871475)、广东省自然科学基金项目(2021A1515011605)、现代农业机械兵团重点实验室开放项目(BTNJ2021002)、广州市创新平台建设计划项目(201905010006)、广州市重点研发计划项目(20210300003)和广东省科技厅重点领域研发计划项目(2020B0202080002)

作者简介：尹航(1978—)，男，副教授，博士，主要从事数据挖掘和智能信息处理研究，E-mail：736028008@qq.com

通信作者：刘双印(1977—)，男，教授，博士，主要从事人工智能和智能信息处理研究，E-mail：hdlsyxlq@126.com

model was applied to a large-scale mutton sheep breeding base in Manas County, Changji Hui Autonomous Prefecture, Xinjiang Uygur Autonomous Region, and good prediction results were obtained. The experimental results showed that the prediction model had good prediction effect, and the root mean square error (RMSE) of ELM was higher than that of SVR, BPNN, LSTM, GRU and LightGBM. The RMSE, mean absolute error (MAE) and R^2 were 0.021 3 mg/L, 0.013 6 mg/L and 0.988 6, respectively. The results showed that the combined model can not only achieve accurate control of carbon dioxide in sheep house, but also meet the needs of fine decision-making for mutton sheep breeding. It also can help farmers make decisions and reduce farming risks.

Key words: sheep house; intensive culture; CO_2 concentration prediction; extreme learning machine; sparrow search algorithm; light gradient boosting machine

0 引言

新疆是我国重要的肉羊养殖基地,其中规模化养殖占比约47.1%,规模化集约化养殖得到快速发展[1]。赵彦超等[2]对新疆地区集约化肉羊养殖环境空气质量调研时发现,新疆集约化羊舍在夏季和冬季的空气质量差异巨大,每天不同时段有害气体浓度也呈显著变化,随着集约化程度提高,养殖密度大、有害气体富聚严重影响到羊畜生长与繁衍[3-4],其中舍内空气龄、CO_2浓度是关键指标,集约化羊舍内 CO_2浓度预测极有必要[5-7]。

目前,不同网络模型已应用于禽畜舍内环境预测[8-12],于此同时,郭昱辰等[13]采用随机森林进行环境变量重要性排序并以此选择输入变量,构建 RF-LSTM 模型来实现更准确的氨气浓度预测。目前主要研究基于理化统计和智能算法两大类模型,理论统计模型需要大量试验数据支撑,工作量大、预测模型泛化性能不足;智能算法模型多影响因子分析能力强,但是在多种环境因素影响下,往往计算复杂度高、预测实时性较弱[14]。

极限学习机(Extreme learning machine, ELM)属于单隐含层前馈神经网络,与传统神经网络相比 ELM 可以避免陷入局部极值的现象,学习精度及速度均表现优秀[15],已在工业、农业各领域获得广泛应用[16-19],并在状态评估、诊断预测等方面收到良好效果[20-21]。徐龙琴等[22]将极限学习机用于水产养殖领域环境预测,通过 ELM 对多尺度分解后模态分量分别训练建模,然后将分量预测结果叠加重构后完成水温预测,EMD-ELM 组合模型具有较高的预测精度和泛化性能;施珮等[23]在此基础上,采用改进遗传算法(IGA)优化 ELM 构建了 EMD-IGA-SELM 模型,提升了水产养殖温度预测精度;偏最小二乘算法(Partial least square, PLS)被用于优化 ELM 网络中隐含层共线性问题,以保障输出权值的稳定性,构建 SPLS-ELM 模型运用到水体溶解氧浓度预测[17]。由于 ELM 初始输入权值和阈值具有随机性,训练精度和时间受随机性影响,所以多种智能优化算法被用于 ELM 优化[24-26]。律方成等[27]使用遗传算法(Genetic algorithm, GA)优化极限学习机并用于短期电力负荷预测;唐延强等[28]通过改进粒子群算法优化 ELM 的初始权值及阈值;廖康等[29]使用灰狼优化(Grey wolf optimization, GWO)算法优化极限学习机模型并用于位移预测。XUE 等[30]受到麻雀群体觅食和反捕食行为的启发,提出麻雀搜索算法(Sparrow search algorithm, SSA),并与改进遗传算法(GA)、粒子群优化算法(Particle swarm optimization, PSO)、灰狼优化算法进行比较,试验证明其在精度、收敛速度、稳定性和鲁棒性上优于其它算法[25];李雅丽等[31]对目前主流的群智能优化算法开展了对比研究,发现 SSA 算法具有很好的探索全局最优潜在区域的能力,可有效避免局部最优问题;马晨佩等[32]将 SSA 算法用于支持向量机(Support vector machine, SVM)的惩罚参数与核参数的优化上,并在轴承故障诊断上获得优异性能。此外,针对多环境因素影响问题,许国艳等[33]采用分布式梯度提升框架(Light gradient boosting machine, LightGBM)对多环境因素进行特征选择以简化多环境因素的影响。

综上分析,本文选取新疆集约化肉羊养殖环境监测数据作为输入量,使用极限学习机对 CO_2质量浓度进行预测。考虑到多环境因素,先用 LightGBM 特征选择简化多环境因素影响,再使用极限学习机对强相关性特征数据进行非线性建模,最后用 SSA 算法对模型参数进行优化调整,构建 LightGBM-SSA-ELM 预测模型。同时选择不同预测模型在真实数据集下进行对比试验,以验证提出模型的合理性和预测能效。

1 材料与方法

1.1 试验区域

本研究以新疆维吾尔自治区昌吉回族自治州玛纳斯新澳畜牧养殖基地(44.27°N,86.10°E)为试验

现场，基地总面积 1 367 m^2，是一个典型的以萨福克肉羊养殖、育种为主的集约化肉羊养殖基地。试验数据采集环境选取面积 422 m^2 的半封闭式标准羊舍（长 33.75 m，宽 12.5 m，均高 3.9 m），羊舍按《畜禽养殖小区建设管理规范》设计，四面墙体采用砖混结构，顶面采用钢板结构，地面采用泥土结构，主要包括主体区域（休息区）、遮阳区域和活动区域（饮食区），各区域间都有可开闭通道，两侧形成空气对流。在夏季，采用自然通风以及遮阳棚进行避暑；在冬季，则在主体区域进行封闭式养殖，并采用换气扇通风。试验对象为萨福克肉羊，试验羊舍羊只年存栏数约为 300 只，饲养密度 0.19 只/m^2；饲喂方式为人工定时饲喂，每日上午、下午各投料一次，自由饮水，人工清粪。

考虑现场环境，传感器选择在主体区域中心檐下安装，其中 CO_2 质量浓度传感器和总悬浮颗粒物质量浓度传感器距地面 2.4 m，其它传感器距地面 3.0～3.1 m，安装位置如图 1 所示。

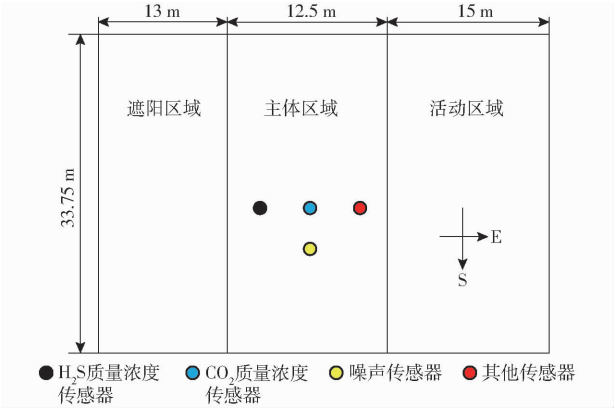

图 1　试验区域示意图

Fig. 1　Experimental area schematic

为保证不同季节和时段样本的一致性，选用广州海睿信息科技有限公司的环境监测设备，包括物联网传输网络、集线器、温度传感器、相对湿度传感器、CO_2 质量浓度传感器、$PM_{2.5}$ 质量浓度传感器、PM_{10} 质量浓度传感器、光照强度传感器、噪声传感器、总悬浮颗粒物质量浓度传感器和 H_2S 质量浓度传感器。监测设备的响应时间小于等于 30 s、重复性在 ±2% 范围内、线性误差在 ±2% 范围内、零点漂移在 ±1% 范围内，具体参数如表 1 所示。

1.2　数据获取与预处理

本试验通过安装在主体区域内的环境监测设备，实时获取羊舍主体区域环境数据，并传输至物联网监测平台数据中心，物联网监测平台拓扑架构如图 2 所示。

试验采集周期为 2021 年 2 月 8 日至 2021 年 3 月 26 日，采集间隔为 10 min，环境数据包括：温度、相对湿度、噪声、光照强度、$PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、总悬浮颗粒物质量浓度、CO_2 质量浓度和 H_2S 质量浓度，有效样本数据共计 6 640 组。有效样本按照 8∶2 划分为训练集和测试集。采集的羊舍主体区域原始数据（部分）如表 2 所示。

考虑数据采集过程中传感器故障、传输故障等因素，采用插值法、均值平滑法对异常数据进行预处理[8]，并对各个变量采用标准化处理器进行归一化处理。

表 1　传感器技术参数

Tab. 1　Technical data of sensors

检测指标	量程	精度	协议
空气温度/℃	－40～105	±0.4	IIC
空气相对湿度/%	0～100	±5	IIC
CO_2 质量浓度/（μg·m^{-3}）	0～50 000	±50	PWM
$PM_{2.5}$ 质量浓度/（μg·m^{-3}）	0～999.9	±7%	Modbus
PM_{10} 质量浓度/（μg·m^{-3}）	0～999.9	±7%	Modbus
光照强度/lx	0～65 535	±5	IIC
噪声/dB	30～120	±5	IIC
总悬浮颗粒物质量浓度/（μg·m^{-3}）	0～999.9	±7%	PWM
H_2S 质量浓度/（μg·m^{-3}）	0～10	±3%	PWM

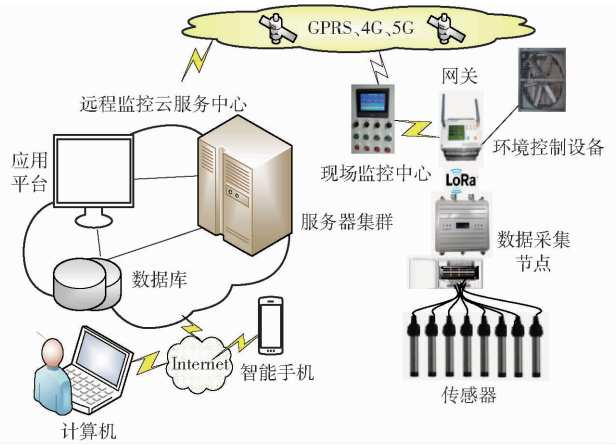

图 2　物联网监测平台拓扑架构图

Fig. 2　Topology structure diagram of Internet of things monitoring platform

2　羊舍 CO_2 质量浓度预测模型构建

2.1　LightGBM 特征选择

羊舍中的 CO_2 质量浓度变化具有一定规律性，并受到温度、相对湿度及多种环境因素的影响，如把所有环境因素都作为特征参数放入模型，会使得网络结构复杂、性能降低。本研究用基于决策树的分布式梯度提升框架（LightGBM）来评估特征重要性，以精简参数[33]。LightGBM 是梯度提升决策树（Gradient boosting decision tree，GBDT）算法的轻量级框架，它利用决策树迭代训练得到最优模型，具有训练效果好、不易过拟合等优点。与极限梯度提升

表 2 2021 年 2 月 8 日—3 月 26 日采集的部分原始数据

Tab. 2 Partial data collected from February 8 to March 26, 2021

检测时间	空气温度/℃	空气相对湿度/%	CO_2 质量浓度/($\mu g \cdot m^{-3}$)	$PM_{2.5}$ 质量浓度/($\mu g \cdot m^{-3}$)	PM_{10} 质量浓度/($\mu g \cdot m^{-3}$)	光照强度/lx	噪声/dB	总悬浮颗粒物质量浓度/($\mu g \cdot m^{-3}$)	H_2S 质量浓度/($\mu g \cdot m^{-3}$)
2021-02-08 17:10	4.2	78.0	1 355	16.3	81.3	146	31.1	120.7	5.2
2021-02-08 17:20	3.9	78.4	1 330	26.5	107.1	97	72.7	165.3	5.2
2021-02-08 17:30	3.7	78.5	1 130	28.9	115.0	134	73.7	178.1	1.7
2021-02-08 17:40	3.4	78.7	1 100	25.5	94.5	122	74.4	148.6	2.3
2021-03-26 23:20	14.9	86.5	2 135	3.0	7.8	0	30.1	10.8	0
2021-03-26 23:30	14.9	87.0	2 175	3.2	4.3	0	32.4	9.3	0
2021-03-26 23:40	15.0	87.2	2 125	3.7	52.6	0	32.0	69.3	0
2021-03-26 23:50	15.0	87.9	2 180	5.1	9.6	0	37.9	18.2	0

(eXtreme gradient boosting, XGBoost)算法相比,LightGBM 使用基于梯度的单边采样(Gradient-based one-side sampling,GOSS)算法配合 EFB 预处理稀疏数据,能够在不损害准确率的条件下加快 GBDT 模型的训练速度[34]。

2.2 极限学习机

极限学习机是一种前馈神经网络,具有良好的泛化性能及快速学习能力[24]。其网络架构由输入层、单隐含层和输出层组成,如图 3 所示。

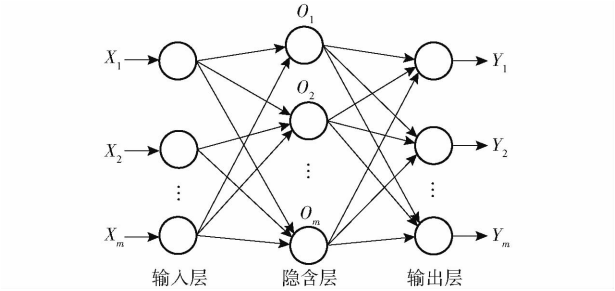

图 3 单隐含层网络结构图

Fig. 3 Single layer hidden layer network

由于网络架构简单,在求取网络输出权重时只需要计算一个逆,且不需要通过反向传播来更新网络模型的权重和偏置参数,因此 ELM 可以克服传统神经网络需要多次迭代学习的缺陷,学习速度快。

对于一个单隐含层网络,假设给定训练集为 $\{x_i, t_i \mid x_i \in \mathbf{R}^D, t_i \in \mathbf{R}^m, i = 1, 2, \cdots, N\}$,其中 x_i 为第 i 个数据示例,t_i 为第 i 个数据示例对应的标记,$\mathbf{R}^D$ 和 $\mathbf{R}^m$ 为所有训练的数据集合。

则有 L 个隐含层节点的网络可以表示为

$$\sum_{i=1}^{L} \beta_i g(\boldsymbol{W}_i \boldsymbol{X}_j + b_i) = c_j \quad (j = 1, 2, \cdots, N) \tag{1}$$

式中 c_j——第 j 个隐含层节点的输出

$g(\cdot)$——激活函数 $\boldsymbol{W}_i$——输出权重

$\boldsymbol{X}_j$——输入的特征矩阵

β_i——加权权重

b_i——第 i 个隐含层节点的偏置

通过拟合可减少误差,其过程可表示为

$$\sum_{j=1}^{N} \| c_j - t_j \| = 0 \tag{2}$$

式中 t_j——第 j 个数据示例对应的标记

即存在 β_i、$\boldsymbol{W}_i$ 和 b_i 使得

$$\sum_{i=1}^{L} \beta_i g(\boldsymbol{W}_i \boldsymbol{X}_j + b_i) = t_j \tag{3}$$

式(3)可写为矩阵形式,表示为

$$\boldsymbol{H}\beta = \boldsymbol{T} \tag{4}$$

式中 $\boldsymbol{H}$——ELM 的隐含层输出

$\boldsymbol{T}$——期望

为了训练单隐含层神经网络,得到 $\hat{\boldsymbol{W}}_i$、$\hat{b}_i$ 和 $\hat{\beta}_i$,公式表示为

$$\| \boldsymbol{H}(\hat{\boldsymbol{W}}_i, \hat{b}_i)\hat{\beta}_i - \boldsymbol{T} \| = \min \| \boldsymbol{H}(\boldsymbol{W}_i, b_i)\beta_i - \boldsymbol{T} \| \tag{5}$$

式中 $\hat{\boldsymbol{W}}_i$——训练后输入权重

$\hat{\beta}_i$——训练后输出权重

$\hat{b}_i$——训练后第 i 个隐含层节点的偏置

接着,训练过程转为搜索最小二乘解,其中隐含层和输出之间连接权重 β 的计算公式为

$$\beta_{\min} = \| \boldsymbol{H}\beta - \boldsymbol{T} \| \tag{6}$$

其解为

$$\hat{\beta} = \boldsymbol{H}^{+}\boldsymbol{T} \tag{7}$$

式中 $\beta_{\min}$—— 最小输出权重

$\hat{\beta}$——训练后的输出权重

$\boldsymbol{H}^{+}$——输出矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 的广义逆,且求得解 $\hat{\beta}$ 的范数最小且唯一

2.3 麻雀搜索算法

麻雀搜索算法(SSA)[30]是 2020 年提出的一种智能优化算法,相比遗传算法、粒子群算法、灰狼算法和引力搜索算法,SSA 具有较强的全局搜索能力,在高维函数计算搜索中具有很好的效果,可以有效避免陷入局部最优的问题[31-32]。SSA 模拟麻雀觅食过程,将不同个体分为发现者(Discoverer)和追随者(Follower),发现者积极发现目标,追随者跟随发

现者，期间有捕食者和警戒者加入。算法流程图如图 4 所示。

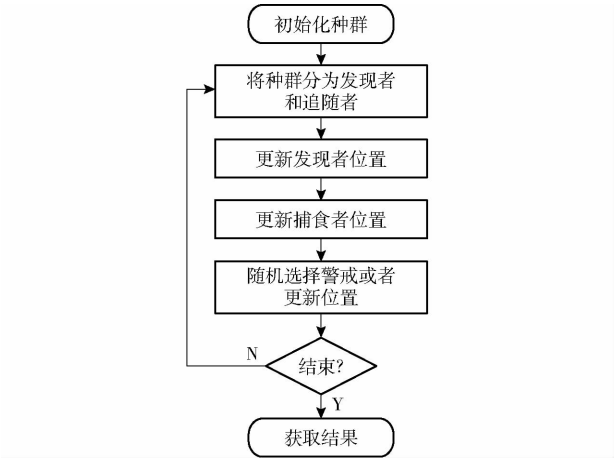

图 4　麻雀搜索算法流程图

Fig. 4　Flow chart of sparrow search algorithm

首先，将数量为 n 的麻雀种群表示为

$$\boldsymbol{X}=\begin{bmatrix} X_1^1 & X_1^2 & \cdots & X_1^d \\ X_2^1 & X_2^2 & \cdots & X_2^d \\ \vdots & \vdots & \vdots & \vdots \\ X_n^1 & X_n^2 & \cdots & X_n^d \end{bmatrix} \tag{8}$$

式中　n——群体中麻雀的总数

d——需要优化的维度

则群体中所有麻雀的适应度可表示为

$$\boldsymbol{X}=\begin{bmatrix} f([X_1^1 \quad X_1^2 \quad \cdots \quad X_1^d]) \\ f([X_2^1 \quad X_2^2 \quad \cdots \quad X_2^d]) \\ \vdots \\ f([X_n^1 \quad X_n^2 \quad \cdots \quad X_n^d]) \end{bmatrix} \tag{9}$$

式中　f——不同发现者的适应度

发现者需要为追随者提供觅食坐标，且所有发现者比追随者具有更大搜索范围及能力。根据式(8)、(9)，发现者位置更新为

$$X_{i,j}^{t+1}=\begin{cases} X_{i,j}+Q\boldsymbol{L} & (R_2 \geqslant S_T) \\ X_{i,j}\exp\left(-\dfrac{i}{\alpha_{iter_{\max}}}\right) & (R_2 < S_T) \end{cases} \tag{10}$$

式中　t——当前迭代数

$\alpha_{iter_{\max}}$——最大迭代次数

$X_{i,j}$——第 i 个麻雀在第 j 维中的位置

S_T——安全值，$S_T \in [0.5,1]$

R_2——预警值，$R_2 \in [0,1]$

$\boldsymbol{L}$——$1 \times d$ 维矩阵，其所有元素初值都为 1

Q——随机数，满足正态分布

当 $R_2 \geqslant S_T$ 时，触发预警，有麻雀已经发现了捕食者，此时所有麻雀都要离开预警区域；当 $R_2 < S_T$ 时，安全状态，发现者可以继续大范围搜索。

此时追随者位置更新为

$$X_{i,j}^{t+1}=\begin{cases} Q\exp\left(-\dfrac{\boldsymbol{X}_{worst}-\boldsymbol{X}_{i,j}^{t}}{i^2}\right) & \left(i>\dfrac{n}{2}\right) \\ \boldsymbol{X}_P^{t+1}+|\boldsymbol{X}_{i,j}-\boldsymbol{X}_P^{t+1}|\boldsymbol{A}^{+}\boldsymbol{L} & \left(i \leqslant \dfrac{n}{2}\right) \end{cases} \tag{11}$$

其中　$\boldsymbol{A}^{+}=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$

式中　$\boldsymbol{X}_{worst}$——全局最差位置

$\boldsymbol{X}_P^{t+1}$——发现者目前所占据的最优位置

$\boldsymbol{A}^{+}$——逆矩阵

当 $i>n/2$ 时，表示第 i 个追随者需要飞往其他地方搜寻觅食。

为提高模型预测性能，将 LightGBM、SSA 和 ELM 模型相互结合，构建基于 LightGBM－SSA－ELM 的羊舍 CO_2质量浓度非线性组合预测模型，预测流程图如图 5 所示。首先对采集的环境数据进行预处理和归一化处理，然后进行重要特征提取，划分训练集和测试集，使用麻雀搜索算法优化。

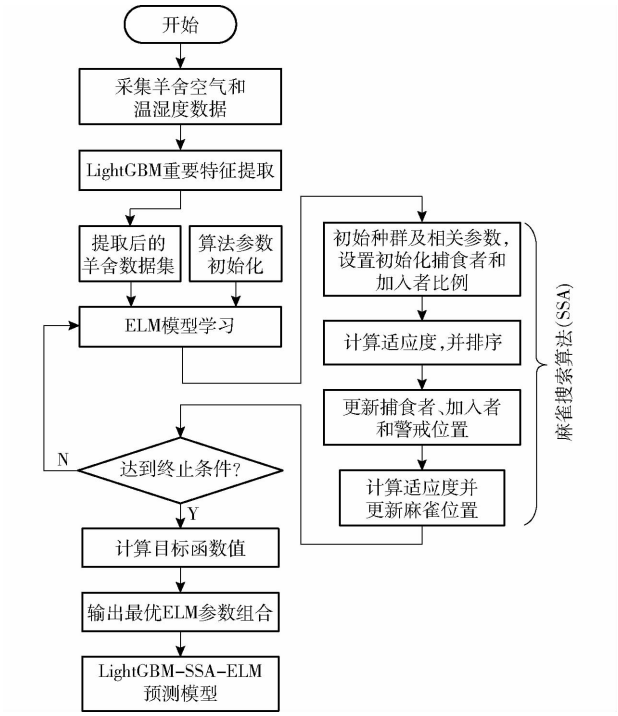

图 5　基于 LightGBM－SSA－ELM 的 CO_2质量浓度预测流程图

Fig. 5　CO_2 concentration prediction flow chart

具体步骤及整体模型流程如下：

(1)通过羊舍内置传感器设备在线采集羊舍环境数据，并实时传输至云平台，导出数据并进行修复。

(2)采用 LightGBM 对修复后的数据进行重要特征提取，筛选出与 CO_2质量浓度相关性高的环境因素，并对筛选后的数据进行归一化处理，按照 8∶2 划分为训练集和测试集。

(3)初始化极限学习机的模型参数后,将训练集输入模型中训练学习,并采用测试集对模型进行测试评分,查看预测精度是否达标,如果达标则模型训练完成,反之利用麻雀搜索算法对模型参数进行寻优。

2.4 性能评价指标

为评价羊舍 CO_2 质量浓度预测模型性能,本文选用均方根误差(RMSE)、平均绝对误差(MAE)和决定系数(R^2)作为主要评价指标。

3 试验与结果分析

3.1 数据准备

本试验以新疆维吾尔自治区昌吉回族自治州玛纳斯县肉羊集约化养殖基地中的半封闭式标准羊舍中主体区域的 CO_2 质量浓度为研究对象,采用物联网在线监测平台,每间隔 10 min 采集一次数据,将 2021 年2 月 8 日至 2021 年 3 月 26 日的环境参数(温度、相对湿度、CO_2 质量浓度、$PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、总悬浮颗粒物质量浓度和 H_2S 质量浓度)总计6 640 条作为试验样本(训练集和测试集按 8∶2划分),预测主体区域 CO_2 质量浓度。其中基于物联网采集 CO_2 质量浓度原始数据呈现较明显的周期性特点,变化曲线如图 6 所示。

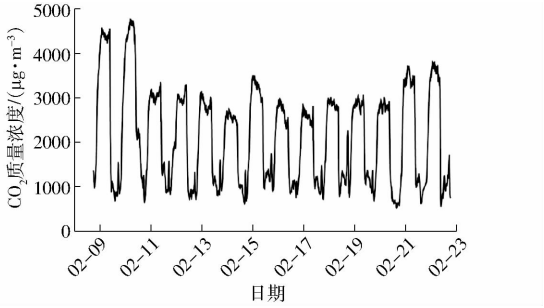

图 6 CO_2 质量浓度原始数据变化曲线

Fig. 6 Variation curve of original CO_2 concentration data

3.2 试验环境与参数选择

试验环境为 I5 - 6200U、2.3 GHz CPU、8 GB 内存、Windows 10 系统、Python3.8 及 Anaconda3 集成开发环境,并使用了多种对比模型。其中 SVR、BPNN、LSTM、GRU 和 LightGBM 采用了基于 Anaconda 的 Sklearn 程序包,ELM 模型基于 Matlab 开发工具箱实现,麻雀优化算法使用 Python 编写,初始参数设置种群数量为 50、最大迭代次数为 10。并对上述模型参数进行寻优,得到基于 ELM 的羊舍 CO_2 质量浓度预测模型。

3.3 特征选择

采集到的环境参数共计 9 类,其中部分参数与 CO_2 质量浓度存在较强相关性,为选择强关联性特征,试验将温度、相对湿度、光照强度、$PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、总悬浮颗粒物质量浓度和 H_2S 质量浓度等 8 种参数输入 LightGBM 模型,得到各参数对 CO_2 质量浓度影响并进行特征重要性排序,得分如表 3 所示。

表 3 基于 LightGBM 模型的 8 种参数特征重要性排序

Tab. 3 Importance scores of eight parameters based on LightGBM model

重要性排序	参数	特征重要性得分
1	光照强度	362.677 986
2	相对湿度	78.038 279
3	$PM_{2.5}$ 质量浓度	51.039 347
4	温度	42.590 839
5	H_2S 质量浓度	21.586 573
6	PM_{10} 质量浓度	20.824 779
7	噪声	19.595 671
8	总悬浮颗粒物质量浓度	11.987 807

从表 3 可以看出,对 CO_2 质量浓度影响最大的特征参数是光照强度,其次是相对湿度。为选取重要的样本特征,按排名依次选择不同数量的参数加入模型,进行对比试验。对比试验得到均方根误差变化曲线如图 7 所示。

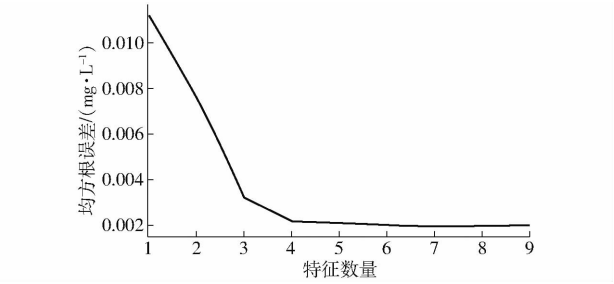

图 7 不同特征数量时 LightGBM 均方根误差变化曲线

Fig. 7 RMSE variation curve of different characteristic numbers

由图 7 可见,选择特征参数的数量为 4 时,均方根误差下降趋势呈现显著拐角,即排名前 4 的特征变量对模型拟合效果影响较大。所以试验选取排名前 4 的参数(光照强度、相对湿度、$PM_{2.5}$ 质量浓度和温度)作为模型输入以降低计算复杂度,并在此基础上使用麻雀搜索算法进行参数优化,构建基于 LightGBM - SSA - ELM 的 CO_2 质量浓度预测模型。同时选择相同参数及优化算法用于 SVR、BPNN、LSTM、GRU 等对比模型,以验证算法有效性。

3.4 极限学习机仿真

试验选取极限学习机建立预测模型,对隐含层神经元数在 5 ~ 65 之间按 10 步长进行遍历,不同神经元数量下的预测误差如图 8 所示,预测效果如图 9 所示。

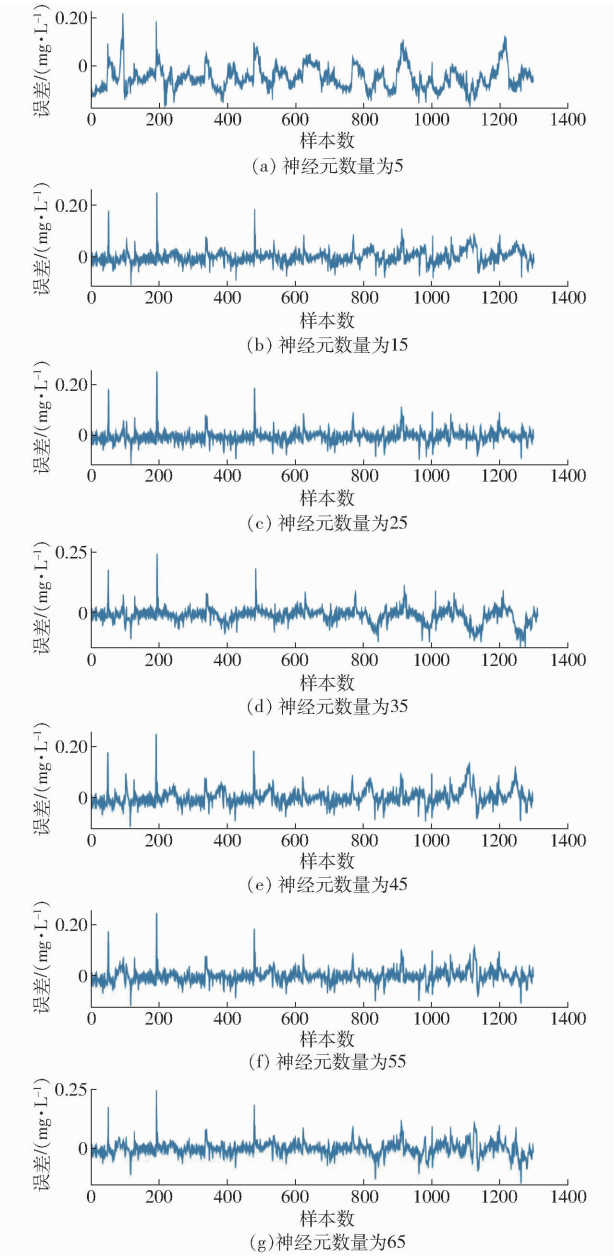

图 8　不同隐含层神经元数时的预测误差变化曲线

Fig. 8　Prediction error of different hidden layer neuron numbers

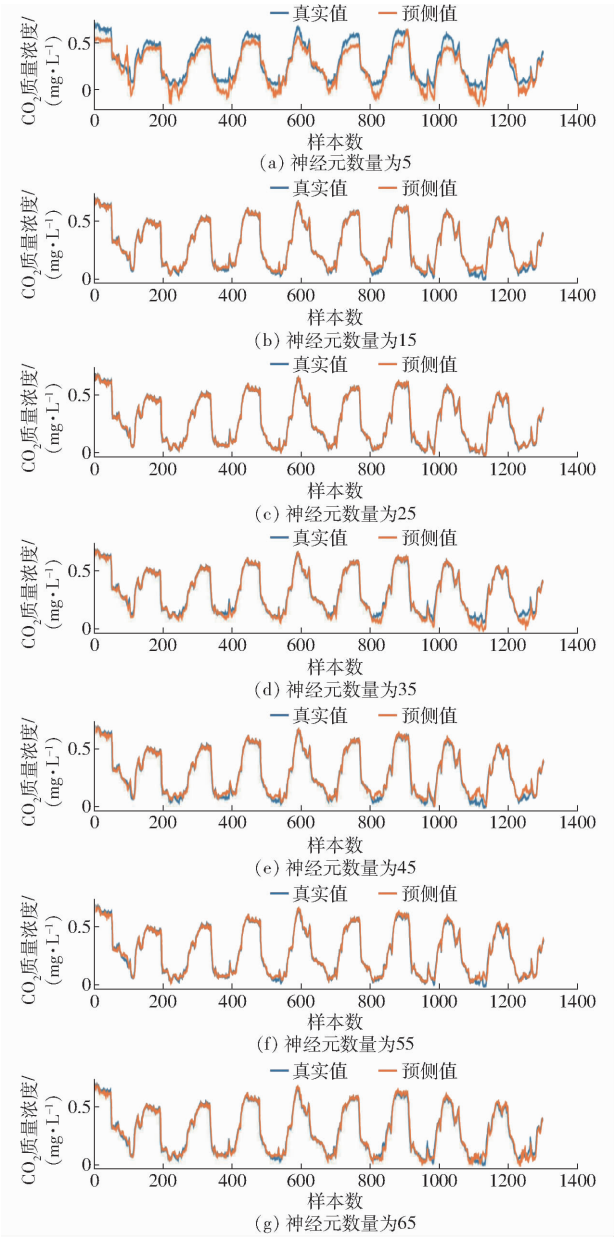

图 9　不同隐含层神经元数时的预测结果曲线

Fig. 9　Prediction results of different hidden layer neuron numbers

从图 8b、8c 可见，在神经元数量为 15 和 25 时，误差波动更快趋于稳定；从图 9 中 ELM 预测值和实际值变化趋势可见，当神经元数量为 25 时（图 9c）预测效果及拟合性能表现最好。

为了对不同神经元结构下模型的预测结果进行定量分析评价，本文采用 RMSE、MAE、R^2 3 个指标进行评价，表 4 为不同数量时 ELM 预测模型的预测效果。从表 4 可知，当隐含层神经元数量为 25 时，RMSE、MAE、R^2 分别达到 0.022 mg/L、0.014 mg/L 和 0.988，均为最优值；取后 6 次试验 RMSE 均值为 0.027 mg/L，MAE 均值为 0.018 mg/L，R^2 均值为 0.981，接近最优值。从定量分析可见，模型可很好地拟合羊舍内环境因素与 CO_2 质量浓度之间复杂的非线性关系，展现了较好的泛化能力。

表 4　不同神经元数量时预测精度分析

Tab. 4　Accuracy analysis of prediction model with different number of neurons

神经元数量	RMSE/(mg·L⁻¹)	MAE/(mg·L⁻¹)	R^2
5	0.083	0.072	0.824
15	0.026	0.018	0.982
25	0.022	0.014	0.988
35	0.033	0.022	0.975
45	0.029	0.020	0.976
55	0.025	0.016	0.984
65	0.028	0.019	0.981
平均值	0.027	0.018	0.981

3.5 对比试验

为验证特征选择后的预测效果,试验选择不同基础模型进行对比试验。ELM、GRU 两种基础模型以及 LightGBM 特征选择后的 LightGBM - BPNN、LightGBM - SVR、LightGBM - LSTM、LightGBM - GRU、LightGBM - LightGBM 和本文 LightGBM - ELM 模型在相同数据集下开展试验,结果如表 5 所示。

表 5 不同模型预测性能对比

Tab. 5 Accuracy analysis of prediction results of each model

模型	RMSE/ ($mg \cdot L^{-1}$)	MAE/ ($mg \cdot L^{-1}$)	R^2
GRU	0.022 8	0.016 1	0.986 3
ELM	0.023 1	0.016 5	0.985 5
LightGBM - BPNN	0.054 6	0.034 3	0.904 3
LightGBM - SVR	0.082 9	0.068 1	0.625 3
LightGBM - LSTM	0.023 4	0.016 4	0.985 5
LightGBM - LightGBM	0.027 6	0.019 7	0.979 6
LightGBM - GRU	0.022 0	0.014 7	0.987 3
LightGBM - ELM	0.022 0	0.014 2	0.987 7

由表 5 可见,经过 LightGBM 特征选择后的 LightGBM - ELM 和 LightGBM - GRU 模型表现优于 ELM 和 GRU 模型,说明提取重要特征简化输入参数,可以提高预测精度;其中本文模型与次优的 LightGBM - GRU 模型相比,RMSE 指标持平,而 MAE 和 R^2 分别提升了 0.04%、3.40%。

为进一步提高预测精度,在上述基础上使用麻雀搜索算法对 LightGBM - ELM 和 LightGBM - GRU 模型进行优化,对 ELM 的隐含层神经元数、GRU 的隐含层神经元数、学习率、批次等分别进行调参优化。

经过 SSA 优化后 LightGBM - ELM 模型 RMSE、MAE、R^2 分别为 0.021 3 mg/L、0.013 6 mg/L 和 0.988 6,与 SSA 优化后的 LightGBM - GRU 模型相比,RMSE、MAE 和 R^2 分别降低了 0.93%、4.23% 和 0.000 5。可见基于 LightGBM - ELM 的 CO_2 质量浓度预测模型明显优于 ELM 和 GRU 等基础模型,LightGBM 特征选择可以提高预测精度,SSA 优化后的 LightGBM - SSA - ELM 模型预测精度也高于 LightGBM - ELM、LightGBM - GRU 和 LightGBM - SSA - GRU 模型,对采集到的时间序列数据具有较好的预测精度,可用于集约化养殖羊舍中 CO_2 质量浓度预测。

4 结束语

利用 LightGBM 提取环境数据的重要特征,去除与 CO_2 质量浓度相关性小的环境变量以简化模型复杂度;使用极限学习机拟合 CO_2 质量浓度与其他环境因子的非线性关系,能够较准确预测 CO_2 质量浓度;使用麻雀算法探索全局最优潜在区域,能更快地搜寻模型最优参数,可提高模型稳定性和鲁棒性。LightGBM - SSA - ELM 模型与 BPNN、LSTM 等模型相比,可有效拟合多环境参数、提高 CO_2 质量浓度预测精度,对集约化肉羊养殖环境预测及精准调控具有积极意义。

参 考 文 献

[1] 徐光艳. 新疆肉羊产业生产效率及影响因素研究[D]. 石河子:石河子大学, 2020.
XU Guangyan. Study on production efficiency and influencing factors of Xinjiang meat sheep industry[D]. Shihezi: Shihezi University, 2020. (in Chinese)

[2] 赵彦超,王瑞宁,李祥龙,等. 不同季节不同类型羊舍空气质量分析[J]. 家畜生态学报, 2020,41(3):57 - 61.
ZHAO Yanchao, WANG Ruining, LI Xianglong, et al. Analysis on air quality in sheep shed in different seasons[J]. Journal of Domestic Animal Ecology, 2020,41(3):57 - 61. (in Chinese)

[3] 周丹,刁亚萍,高云,等. 猪舍内 CO_2 的排放研究进展[J]. 中国农业科学, 2018,51(16):3201 - 3213.
ZHOU Dan, DIAO Yaping, GAO Yun, et al. Research review on CO_2 production in pig house[J]. Scientia Agricultura Sinica, 2018,51(16):3201 - 3213. (in Chinese)

[4] 徐菁,张明新,赵云辉,等. 环境因素对羊繁殖性能影响的研究[J]. 家畜生态学报, 2019,40(4):85 - 88.
XU Jing, ZHANG Mingxin, ZHAO Yunhui, et al. Effects of environmental factors on the reproduction performance of sheep [J]. Journal of Domestic Animal Ecology, 2019, 40(4):85 - 88. (in Chinese)

[5] 曹孟冰,杨婷,宗超,等. 进风口高度与导流板角度对猪舍空气龄和 CO_2 分布的影响[J]. 农业机械学报, 2020, 51(增刊2):427 - 434.
CAO Mengbing, YANG Ting, ZONG Chao, et al. Influence of air inlet height and angle on air age and CO_2 distribution in pig house[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020,51(Supp.2):427 - 434. (in Chinese)

[6] 王美芝,刘继军,赵婉莹,等. 北京夏季机械通风育肥猪舍 CO_2、NH_3 排放和耗水量研究[J]. 农业工程学报, 2017, 33(17):152 - 160.
WANG Meizhi, LIU Jijun, ZHAO Wanying, et al. CO_2 and NH_3 emissions and water consumption at mechanically ventilated

finishing pig house in Beijing equipped with pull-plug manure removal system[J]. Transactions of the CSAE, 2017, 33(17): 152－160. (in Chinese)

[7] 焦伟娜,李卓程,刘虹家,等.新疆南北疆两羊场春季养殖环境指标差异比较及羔羊健康状况调查[J].中国畜牧兽医, 2020,47(1):306－314.
JIAO Weina, LI Zhuocheng, LIU Hongjia, et al. Comparative study on spring environmental indices and lamb's health status of two intensive sheep farms in south and north of Xinjiang[J]. China Animal Husbandry & Veterinary Medicine, 2020,47(1): 306－314. (in Chinese)

[8] 谢秋菊,郑萍,包军,等.基于深度学习的密闭式猪舍内温湿度预测模型[J].农业机械学报, 2020, 51(10):353－361.
XIE Qiuju, ZHENG Ping, BAO Jun, et al. Thermal environment prediction and validation based on deep learning algorithm in closed pig house[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020,51(10):353－361. (in Chinese)

[9] RODRIGUEZ M R, LOSADA E, BESTEIRO R, et al. Evolution of NH_3 concentrations in weaner pig buildings based on setpoint temperature[J]. Agronomy, 2020, 10(1): 107.

[10] ROBERTO B,TAMARA A, ORTEGA J A, et al. Prediction of carbon dioxide concentration in weaned piglet buildings by wavelet neural network models[J]. Computers and Electronics in Agriculture, 2017,143:201－207.

[11] 刘春红,杨亮,邓河,等. 基于 ARIMA 和 BP 神经网络的猪舍氨气浓度预测[J].中国环境科学, 2019,39(6): 2320－2327.
LIU Chunhong, YANG Liang, DENG He, et al. Prediction of ammonia concentration in piggery based on ARIMA and BP neural network[J]. China Environmental Science, 2019,39(6):2320－2327. (in Chinese)

[12] 丁涛,蒋雷生,施正香,等.基于随机森林的寒区奶牛舍环境因素与产奶量关系研究[J].农业机械学报,2021,52(12):378－385.
DING Tao, JIANG Leisheng, SHI Zhengxiang, et al. Relationship between environmental factors and milk yield of dairy farm in cold region based on random forest[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(12): 378－385. (in Chinese)

[13] 郭昱辰,杨亮,刘春红,等. 基于 RF－LSTM 的鸡舍恶臭气体预测研究[J].中国环境科学,2020,40(7):2850－2857.
GUO Yuchen, YANG Liang, LIU Chunhong, et al. Prediction of stench gas in chicken house based on RF－LSTM[J]. China Environmental Science, 2020,40(7):2850－2857. (in Chinese)

[14] 郭祥云,连京华,李慧敏,等. 基于双阶段注意力机制和 LSTM 的鸡舍氨气浓度预测算法[J].中国农业大学学报, 2021,26(6):187－195.
GUO Xiangyun, LIAN Jinghua, LI Huimin, et al. Ammonia concentration forecasting algorithm in layer house based on two-stage attention mechanism and LSTM[J]. Journal of China Agricultural University, 2021,26(6):187－195. (in Chinese)

[15] 于海洋. 在线预测的极限学习机方法研究[D].长春:吉林大学,2019.
YU Haiyang. Research on extreme learning machine methods for online prediction[D]. Changchun: Jilin University,2019. (in Chinese)

[16] 王振阳,江德文,王新东,等.基于支持向量回归与极限学习机的高炉铁水温度预测[J].工程科学学报, 2021,43(4): 569－576.
WANG Zhenyang, JIANG Dewen,WANG Xindong, et al. Prediction of blast furnace hot metal temperature based on support vector regression and extreme learning machine[J]. Chinese Journal of Engineering, 2021,43(4):569－576. (in Chinese)

[17] 施珮,匡亮,袁永明,等.基于改进极限学习机的水体溶解氧预测方法[J].农业工程学报, 2020,36(19):225－232.
SHI Pei, KUANG Liang, YUAN Yongming, et al. Dissolved oxygen prediction for water quality of aquaculture using improved ELM network[J]. Transactions of the CSAE, 2020, 36(19): 225－232. (in Chinese)

[18] 国婷婷,殷廷家,杨正伟,等.基于 WPT－IAF－ELM 的小麦储存年限电子舌检测分析[J].农业机械学报, 2019,50(增刊):404－410.
GUO Tingting, YIN Tingjia, YANG Zhengwei, et al. Detection and analysis of wheat storage year using electronic tongue based on WPT－IAF－ELM[J]. Transactions of the Chinese Society for Agricultural Machinery,2019,50(Supp.):404－410. (in Chinese)

[19] 王雪松,赵季娟,程玉虎.基于极限学习机的无监督领域适应分类器[J].控制与决策, 2020,35(4):861－869.
WANG Xuesong,ZHAO Jijuan, CHENG Yuhu. Unsupervised domain adaption classifier via ELM[J]. Control and Decision, 2020,35(4):861－869. (in Chinese)

[20] 葛兴来,张鑫. 采用奇异能量谱与改进 ELM 的轴承故障诊断方法[J].电机与控制学报, 2021,25(5):80－87.
GE Xinglai, ZHANG Xin. Bearing fault diagnosis method using singular energy spectrum and improved ELM[J]. Electric Machines and Control, 2021,25(5):80－87. (in Chinese)

[21] 周凯啸,林文辉,孙建坤,等.基于改进极限学习机的转炉出钢合金化锰收得率预测模型[J].中南大学学报(自然科学版), 2021,52(5):1399－1406.
ZHOU Kaixiao, LIN Wenhui,SUN Jiankun, et al. A prediction model to calculate Mn yield during BOF alloying process using improved extreme learning machine[J]. Journal of Central South University (Science and Technology), 2021,52(5):1399－1406. (in Chinese)

[22] 徐龙琴,张军,李乾川,等.基于 EMD 和 ELM 的工厂化育苗水温组合预测模型[J].农业机械学报, 2016,47(4):265－271.

XU Longqin, ZHANG Jun, LI Qianchuan, et al. Combined prediction model of water temperature in industrialized cultivation based on empirical mode decomposition and extreme learning machine[J]. Transactions of the Chinese Society for Agricultural Machinery, 2016,47(4):265 - 271. (in Chinese)

[23] 施珮,袁永明,匡亮,等. 基于 EMD - IGA - SELM 的池塘养殖水温预测方法[J].农业机械学报, 2018,49(11):312 - 319.
SHI Pei, YUAN Yongming, KUANG Liang, et al. Water temperature prediction in pond aquaculture based on EMD - IGA - SELM neural network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018,49(11):312 - 319. (in Chinese)

[24] 徐睿,梁循,齐金山,等.极限学习机前沿进展与趋势[J].计算机学报, 2019, 42(7):1640 - 1670.
XU Rui, LIANG Xun, QI Jinshan, et al. Advances and trends in extreme learning machine[J]. Chinese Journal of Computers, 2019,42(7):1640 - 1670. (in Chinese)

[25] 薛建凯. 一种新型的群智能优化技术的研究与应用[D]. 上海:东华大学, 2020.
XUE Jiankai. Research and application of a novel swarm intelligence optimization technique: sparrow search algorithm[D]. Shanghai: Donghua University,2020. (in Chinese)

[26] 冯茜,李擎,全威,等.多目标粒子群优化算法研究综述[J].工程科学学报, 2021,43(6):745 - 753.
FENG Qian, LI Qing, QUAN Wei, et al. Overview of multi-objective particle swarm optimization algorithm[J]. Chinese Journal of Engineering, 2021,43(6):745 - 753. (in Chinese)

[27] 律方成,刘怡,亓彦珣,等.基于改进遗传算法优化极限学习机的短期电力负荷预测[J].华北电力大学学报(自然科学版), 2018,45(6):1 - 7.
LÜ Fangcheng, LIU Yi, QI Yanxun, et al. Short-term load forecasting based on optimized learning machine using improved genetic algorithm[J]. Journal of North China Electric Power University, 2018,45(6):1 - 7. (in Chinese)

[28] 唐延强,李成海,宋亚飞,等.基于改进粒子群优化和极限学习机的网络安全态势预测[J].计算机应用, 2021,41(3):768 - 773.
TANG Yanqiang, LI Chenghai, SONG Yafei, et al. Network security situation prediction based on improved particle swarm optimization and extreme learning machine[J]. Journal of Computer Applications, 2021,41(3):768 - 773. (in Chinese)

[29] 廖康,吴益平,李麟玮,等.基于时间序列与 GWO - ELM 模型的滑坡位移预测[J].中南大学学报(自然科学版),2019, 50(3):619 - 626.
LIAO Kang, WU Yiping, LI Linwei, et al. Displacement prediction model of landslide based on time series and GWO - ELM [J]. Journal of Central South University (Science and Technology), 2019,50(3):619 - 626. (in Chinese)

[30] XUE J, SHEN B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. Systems Science & Control Engineering, 2020,8:22 - 34.

[31] 李雅丽,王淑琴,陈倩茹,等.若干新型群智能优化算法的对比研究[J].计算机工程与应用, 2020,56(22):1 - 12.
LI Yali, WANG Shuqin, CHEN Qianru, et al. Comparative study of several new swarm intelligence optimization algorithms [J]. Computer Engineering and Applications, 2020, 56(22):1 - 12. (in Chinese)

[32] 马晨佩,李明辉,巩强令,等.基于麻雀搜索算法优化支持向量机的滚动轴承故障诊断[J].科学技术与工程, 2021, 21(10):4025 - 4029.
MA Chenpei, LI Minghui, GONG Qiangling, et al. Fault diagnosis of rolling bearing based on sparrow search algorithm optimized support vector machine[J]. Science Technology and Engineering, 2021,21(10): 4025 - 4029. (in Chinese)

[33] 许国艳,周星熠,司存友,等. 基于 GRU 和 LightGBM 特征选择的水位时间序列预测模型[J].计算机应用与软件, 2020,37(2):25 - 31.
XU Guoyan, ZHOU Xingyi, SI Cunyou, et al. A water level time series prediction model based on GRU and LightGBM feature selection[J]. Computer Applications and Software, 2020,37(2):25 - 31. (in Chinese)

[34] CHEN Cheng, ZHANG Qingmei, YU Bin, et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier[J]. Computers in Biology and Medicine, 2020,123:103899.