基于LSSA-XGBOOST改进算法的高体鰤鱼类体质量预测模型

俞国燕, 左仁意, 严 俊, 罗樱桐, 朱琪珩

Body mass of *Seriola dumerili* prediction model based on LSSA-XGBOOST improved algorithm

YU Guoyan, ZUO Renyi, YAN Jun, LUO Yingtong, ZHU Qiheng

在线阅读 View online: https://doi.org/10.14012/j.cnki.fjsc.2023.05.002

渔业研究　2023,45(5):427 – 437　　http://www.hyyysci.com
Journal of Fisheries Research　　DOI:10.14012/j.cnki.fjsc.2023.05.002

俞国燕,左仁意,严　俊,等. 基于 LSSA – XGBOOST 改进算法的高体鰤鱼类体质量预测模型[J]. 渔业研究,2023,45(5):427 – 437.

基于 LSSA – XGBOOST 改进算法的高体鰤鱼类体质量预测模型

俞国燕[1,2]，左仁意[1,2]，严　俊[1*]，罗樱桐[2]，朱琪珩[2]

（1. 南方海洋科学与工程广东省实验室，广东 湛江 524013；
2. 广东海洋大学机械与动力工程学院，广东 湛江 524088）

摘要：为构建利用体质量判断的精准投喂模型，需实时获取鱼群体质量状态，基于 LSSA – XGBOOST 算法，通过对工船养殖实测的高体鰤（*Seriola dumerili*）体长、体宽和体质量数据进行分析，构建以体长、体宽两项体态特征数据为输入、体质量数据为输出的高体鰤体质量预测模型。结果显示，与常规数学模型拟合相比，LSSA – XGBOOST 模型拟合的相关性系数 R^2 提高约 10%；与传统 BP 神经网络和粒子群优化 BP 相比，LSSA – XGBOOST 模型误差平方和 R^2 提升约 3%，这为构建基于体质量判断的高体鰤精准投喂模型提供了理论依据。

关键词：LSSA – XGBOOST；高体鰤；体长；体质量；关系
中图分类号：S917.4　**文献标识码**：A　**文章编号**：2096 – 9848(2023)05 – 0427 – 11

水产养殖过程中，养殖成本占比最大的是饲料成本[1]。为降低养殖成本，需搭建科学的精准投喂模型，环境、鱼群、饲料营养等均是影响模型性能的重要因素[2-3]。在工船养殖模式下，水温、pH 等环境因素趋于平稳，饲料种类在养殖开始时已被确定，但不同生长阶段的鱼群投喂饲料规格随着鱼体质量变化而有所改变。鱼群平均体质量是搭建精准投喂模型的关键要素[4]，然而养殖过程中的鱼群十分活跃，这给鱼群体质量测量带来巨大的困扰。现有学者凭借图像处理技术在水下完成了鱼的尺寸测量[5]，但如何将获取到的尺寸信息转换为体质量信息是亟待解决的关键问题，故鱼群体态特征及其体质量关系研究不可或缺。

鱼群体态特征（体长、体宽）及其与体质量的关系是一种重要的生物差异指标[6-7]，也是鱼类研究者们进行生长状态判断以及生态系统建模的重要依据[8-9]，还对鱼群生长状态及生物量的判断有较大的帮助[6]。在关于鱼群体态特征（体长、体宽）与体质量关系研究中，相关性系数 R^2 常常被用来验证模型性能，如 Sepa P 等[10]研究了在厄瓜多尔海洋水域的 4 种深海软骨鱼的体长、体质量关系，使用幂指数模型（$Y = aX^b$）完成体长、体质量关系拟合，相关性系数 R^2 达到 0.940；Najmudeen T M 等[11]为获取 3 种远洋鲨鱼体长、体质量关系及其相关系数，在阿拉伯海东南部采集了 525 组数据完成拟合，相关性系数 R^2 达到 0.901；陈锋等[12]完成察隅河及其支流贡日嘎布弧唇裂腹鱼体长、体质量关系对比研究，计算相关系数后确定其体长、体质量关

收稿日期：2023 – 03 – 09
基金项目：南方海洋科学与工程广东省实验室（湛江）科研项目（zjw – 2019 – 01）；广东省海洋经济发展（海洋六大产业）专项资金项目（GDNRC〔2021〕42）
作者简介：俞国燕（1970—），女，教授，研究方向：智能设计与制造、现代化渔业装备等。E – mail:yugy@gdau.edu.cn
通信作者：严　俊（1957—），男，研究员，研究方向：深远海渔业养殖装备。E – mail:yanj@zjblab.com

系符合 $W = 2.72 \times 10^5 SL^{2.888}$ 方程，相关性系数 R^2 达到 0.972。除使用传统数学模型的方法描述鱼体长－体质量关系外，新的研究方法也层出不穷，如林雅蓉等[13]利用绘图求积法完成中华哲水虱体长、体质量测定及其关系拟合等。此外，还有大量学者致力于寻求适用性更强、拟合度更高的新型回归方法，如张志伟等[14]使用神经网络模型对数据进行回归，搭建了具有外延性（即预测能力）、拟合性能良好的模型。然而上述拟合方法大多需要大量样本数据支撑，仅采集数据就需好几年的连续记录[11]。

随着中国深远海养殖事业的发展，大量企业开始着力构建新型养殖模式[15]，与此同时大量适养于深海的鱼类开始出现在公众视野[16]。高体鰤（*Seriola dumerili*）又名章红鱼，是一种生活在水深 20～70 m 的海洋鱼类，具有较高的食用价值，并且生长速度快、养殖周期短，是一种名贵的经济鱼类[17]。中国从 1991 年开始高体鰤养殖技术的研究[18]，至今对高体鰤的人工养殖技术研究仍未停止[19-20]。2022 年 6 月，南方海洋科学与工程实验室为验证工船养殖高体鰤的可行性，开展了高体鰤养殖实验。为降低养殖过程的饲料成本，需构建一种适用于工船养殖的精准投喂模型，而平均体质量是搭建精准投喂模型的关键要素。通过图像视频数据判断鱼体质量，可以大大地降低鱼群平均体质量获取难度，然而视频图像仅可获悉鱼群体态特征，因此搭建基于鱼群体态特征的鱼体质量预测模型十分必要。使用传统数学模型或神经网络模型搭建体态特征与体质量关系模型时，其对数据集体量要求较高[21]。因此，本研究采用有别于传统神经网络的 LSSA－XGBOOST 优化树模型完成体质量预测，在保留了决策提升树（XGBOOST）算法处理小样本数据的优良性能前提下，优化了模型结构，使 LSSA－XGBOOST 模型在仅有少量样本数据的情况下拥有更高的拟合精度，为搭建精准投喂模型提供重要的支撑。

1 材料与方法

1.1 实验材料

2022 年 6 月 22 日，第一批高体鰤苗放入养殖仓，初始平均体质量为 90 g。2022 年 8 月 25 日，养殖周期为 64 d，共取 314 条高体鰤，测量其体长、体宽和体质量数据。在数据采集过程中，分别使用直尺和电子秤进行样本鱼的体长、体宽和体质量测量，并使用棉手套擦去鱼表面水分，长度精确到 1 mm，体质量精确到 1 g。

实验地点为广西北海市银海区福成镇西村至营盘南部海域的广西精工深水网箱养殖区。实验平台为中国船舶集团广西公司负责改装修理的“银渔养 0039”游弋式实验船（图 1），该船总长 48.3 m，型宽 9.5 m，型深 2.9 m，设计吃水 1.4 m，并配备双机双桨。养殖实验期间，实验船始终沿养殖区固定航线游弋，从而保证实验期间循环水系统始终能够从外界获取优质海水。

图 1 实验船整体图

Fig.1 Overall diagram of test vessel

为保证养殖舱内水体质量，舱内四角分别设有进水口，进水流量由舱底电磁流量阀操控，并配备全套水质检测传感器。舱底中心位置为出水口，进出水时可将杂质、死鱼、残饵等养殖废料汇集，再利用出水口涡流的带动排出舱外。

1.2 数据处理

在实验过程中，往往会产生小部分异常数据点，这些异常数据点常常会造成整体数据集质量下降，不利于数据可靠性等多种负面影响[22]，也对神经网络模型训练造成影响，因此参照文献［23］对异常点数据进行预处理。

1.2.1 极端学生化偏差（Extreme studentized deviate，ESD）数据降噪方法

在实际水质监测工作中，通常有多个异常数据点，ESD 方法将单个异常数据检测（Grubbs test）方法扩展，使其能进行多个异常值检测，为了将 Grubbs’ test 扩展到 k 个异常值检测，需要在数据集中逐步删除与均值偏离最大的值

（最大值或最小值），同步更新对应的 t 分布临界值，检验原假设是否成立。算法流程如下：

计算与均值偏离最远的残差 R_j：

$$R_j = \max_i \frac{|Y_i - \bar{Y}'|}{S}, 1 \leqslant j \leqslant k \tag{1}$$

式（1）中：$\bar{Y}$ 和 S 分别为数据集均值和方差。

计算临界值 λ_j：

$$\lambda_j = \frac{(n-j) \times t_{p,n-j-1}}{\sqrt{(n-j-1+t^2_{p,n-j-1})(n-j+1)}}, 1 \leqslant j \leqslant k \tag{2}$$

式（2）中：n 为数据量；j 为预去除的第 j 个量；$t_{p,n-j-1}$ 表示 t 分布临界值。

1.2.2　传统数学模型

1）Gauss 曲线

Gauss 曲线是一种常用的拟合曲线模型，满足正态分布的高斯函数如下：

$$f(x) = \frac{1}{\sigma \sqrt{2x}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{3}$$

式（3）中：μ 为数学期望；σ^2 为标准方差。常见应用数学模型拟合鱼类体质量与体态特征关系时，大多选择体长和体质量两项参数。

2）Logistic 曲线

Logistic 曲线是一种典型的 S 型函数，又名 Sigmoid 函数，常常被用来描述生物量增长状态，生物数量增长本身应当符合指数型增长，受环境阻力（生存空间、天敌数量等）的影响，在其增长至一定数量后，达到极限数量 K 值并维持稳定。从整体曲线变化来看，前期爆炸增长及后期环境阻力减缓其增长，使曲线整体呈 S 型，即增长速率先增大后减小。其数学方程表示为：

$$P(t) = \frac{KP_0 e^{rt}}{K + P_0(e^{rt} - 1)} \tag{4}$$

式（4）中：P_0 为初始状态；K 为终值；参数 r 用于衡量变化速度。

3）幂函数曲线

幂函数曲线即指数函数，属于初等函数之一，常用于描述微生物增长状态，即拥有所有生长所需资源且无环境阻力下的生物量增长形式。方程结构调整如下：

$$F(x) = Ke^{-\frac{x}{t}} + F(0) \tag{5}$$

式（5）中：K、t 为常数；$F(0)$ 为初始状态。

目前常用的体长、体质量关系拟合方法为 Von Bertalanffy 方程[24]：

$$W = aL^b \tag{6}$$

式（6）中：W 表示体质量；L 表示体长；a、b 均为实数，可使用 SPSS 软件计算得出。

1.2.3　LSSA－XGBOOST 拟合模型

1）麻雀搜索算法（Sparrow search algorithm，SSA）及其改进

麻雀搜索算法是东华大学的薛建凯[25]于 2020 年提出的一种新型群智能寻优算法，在鸟群觅食过程中，优先找寻到食物的个体称之为发现者，发现者会向其他个体即加入者传递信息，而加入者与发现者相互竞争、抢夺资源。麻雀算法按此模式多次群体寻优，最终选出获得最高适应度个体，即算法得出的最优解。

初始化种群个体可表示为：

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \tag{7}$$

式（7）中：d 表示待优化参数量；n 为种群数量。

种群适应度为 $F(X)$，形式为个体适应度 $f(x)$ 组成的 N 行矩阵：

$$F(X) = \begin{bmatrix} f([x_{11} \quad x_{12} \quad \cdots \quad x_{1d}]) \\ f([x_{21} \quad x_{22} \quad \cdots \quad x_{2d}]) \\ \cdots \\ f([x_{n1} \quad x_{n2} \quad \cdots \quad x_{nd}]) \end{bmatrix} \tag{8}$$

发现者位置随搜寻范围变化不断更新，公式如下：

$$X^{p+1}_{i,j} = \begin{cases} X_{i,j} \times \exp(-\frac{i}{\alpha \times p_{\max}}), R_2 < ST \\ X_{i,j} + Q \times L, R_2 > ST \end{cases} \tag{9}$$

式（9）中：p 为迭代次数；i、j 分别表示个体与种群数（$X_{i,j}$ 表示第 i 个种群第 j 个个体）；$p_{\max}$ 表示最大迭代次数；α 为（0，1］区间内的随机数；R_2 表示预警值，范围取［0，1］；ST 表示安全值；范围取［0.5，1.0］；Q 为随机数；L 为维度 $1 \times d$ 的全 1 矩阵。

加入者通过观察发现者位置，并随之完成位置更新：

$$X_{i,j}^{t+1}=\begin{cases}Q\times\exp\left(\dfrac{X_{worst}-X_{i,j}^{t}}{i^{2}}\right), i>n/2\\ X_{P}^{t+1}+\left|X_{i,j}^{t}-X_{p}^{t+1}\right|\times A^{+}\times L, i\leqslant n/2\end{cases}\quad(10)$$

式（10）中：X_P 是目前发现者所占据的最优位置；X_{worst} 为全局最差位置；A 表示所有值随机为 1 或 −1 的 $1\times d$ 矩阵；$A^{+}=A^{T}(AA^{T})^{-1}$；$i>n/2$ 时，第 i 个加入者未获得食物，需重新选择觅食位置。

警觉者初始位置在群体中随机产生，其位置表示为：

$$X_{i,j}^{t+1}=\begin{cases}X_{best}^{t}+\beta\times\left|X_{i,j}^{t}-X_{best}^{t}\right|, f_i>f_g\\ X_{i,j}^{t}+K\times\left(\dfrac{\left|X_{i,j}^{t}-X_{worst}^{t}\right|}{(f_i-f_w)+\varepsilon}\right), f_i=f_g\end{cases}\quad(11)$$

式（11）中：X_{best} 是当前的全局最优解；β 为步长控制系数，其特征服从（0，1）间的正态分布；K 是区间［−1，1］下的随机数；f_i 表示当前个体适应度；f_g 表示最佳适应度；f_w 为最差适应度；ε 为常数。

（1）混沌优化（LSSA）

麻雀算法（SSA）初始种群产生方法为构成种群数量（pop）×目标参数（dim）的均匀分布的随机矩阵，这种方法在群体检索过程中会生成均匀分布在一片区域内的点，如图 2 所示。

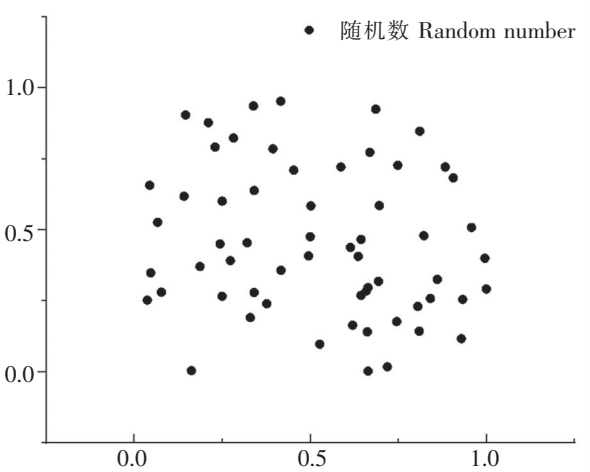

图 2 初始随机分布

Fig.2 Initial random distribution

（2）混沌随机矩阵优化

麻雀算法初始种群优化方法有多种方式，实验所用混沌随机数发生器基于 Logistic 方程，其表现形式为：

$$X(n+1)=\mu X(n)[1-X(n)]\quad(12)$$

式（12）中：参数 $u\geqslant 3.569\ 946$ 后，X 的值不再发生震荡，随后进入混沌状态。

混沌 SSA 基于该原理随机产生的随机值分布更加分散，如图 3 所示。

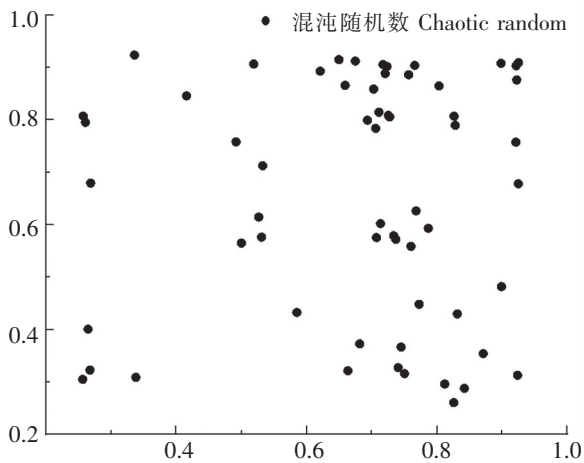

图 3 混沌随机分布

Fig.3 Chaotic random distribution

作为一种群体寻优算法，初始种群分布均匀的程度直接关系到算法的全局搜索能力[26]，对比 LSSA 初始种群和 SSA 初始种群在各范围内的分布直方图（图 4）可知，LSSA 初始种群在［0，1］区间范围内分布的数量更为平均，这将降低初始化种群时因初始化个体过于集中而漏掉关键信息的几率，提高了算法全局搜索能力。

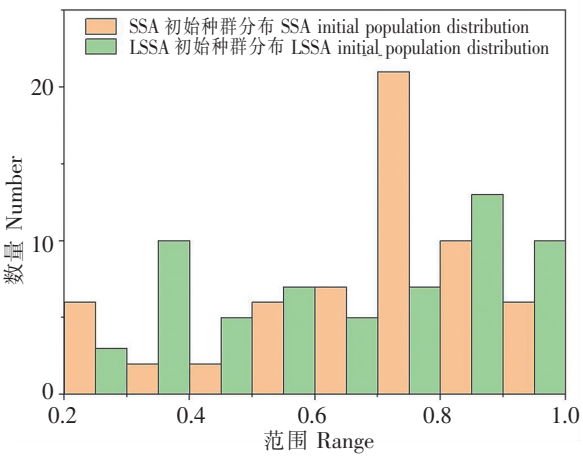

图 4 LSSA/SSA 初始种群分布直方图

Fig.4 LSSA/SSA initial population distribution histogram

2）XGBOOST 极端梯度提升树

XGBOOST 算法于 2014 年由 Chen T Q 等[27]提出，其算法核心在于将多个低准确率分类器组

合成一个高准确率模型，针对问题，将对象进行不断分类判断并打分，最终某个对象的分数是所有 XGBOOST 树评分之和。XGBOOST 算法在处理分类和回归问题中均具有十分良好的表现。

对于 XGBOOST 而言，其输出 F 是由多个评分树结果相加，表示方法如下：

$$\hat{y}_i = \Phi(X_i) = \sum_{k=1}^{k} fk(x_i), fk \in F \qquad (13)$$

式（13）中：$F = \{f(x) = w_{q(x)}\}(q:R^m \rightarrow T, \varpi \in R^T)$，$F$ 表示单个回归树空间（CART），其中 q 表示树结构，将训练集中的单组数据映射到树结构中。T 表示叶结点数量，每个回归树空间包含树结构以及其权重 w。除此之外，每个树节点中都包含有评分，表示为 W_i。树的结构 q 根据实际案例设定，以常见大小判断为例：

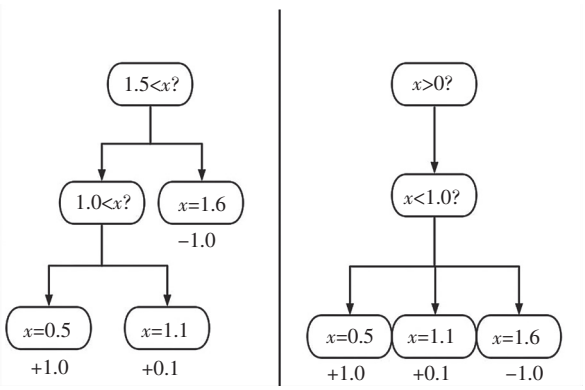

图 5　XGBOOST 树评分方式

Fig.5　XGBOOST tree scoring method

由图 5 可知，若目标是搜寻处于［0，1］的数，树模型设置了两层结构，在数据输入后对其进行打分，观察图 5（左），当输入 1.6 时，第一次判断根据其大于 1.5 直接评分为 −1.0，而输入 0.5 和 1.1 时，则分别获得 1.0 和 0.1 的评分。若运算过程涉及多个树结构，以图 5 为例，0.5、1.1 和 1.6 三个数的最终结果由左、右两侧树各末端评分分别加权求得，若两侧权重相等，则 3 个数最终评分结果为 2.0、0.2 和 −1.9，可以得出 0.5 在区间［0，1］内，1.1 在区间边缘，而 1.6 在搜索区间之外。为了模拟这个运算过程，需用到下述公式：

$$L(\Phi) = \sum_i l(\hat{y}_i, y_i) + \sum \Omega(f_k) \qquad (14)$$

式（14）中：$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2$，$l$ 表示 y_i 与 $\hat{y}_i$ 差值（损失函数）；Ω 代表回归树函数，将其作为额外的正则项，有利于降低模型过拟合概率，使学习过程更加平滑。

在实际运算中，很多关系无法通过简单累加公式拟合得出，为提高提升树的渐进能力，方程增加了二次项函数，简化后的正则公式为：

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \qquad (15)$$

式（15）中：$\begin{cases} g_i = \partial_y (t-1) l(y_i, \hat{y}^{(t-1)}) \\ h_i = \partial_y^2 (t-1) l(y_i, \hat{y}^{(t-1)}) \end{cases}$。

树结构搭建完成后，需对其结构质量进行评估，公式为：

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \qquad (16)$$

式（16）中：q 为待评估的结构；h_i、g_i 及 I_j 分别表示损失函数二阶、一阶统计量、叶节点实例集。

模型在正常运算时，由于叶节点繁多、结构的评估验证是一层一层循序推进的，单层若有左右两个节点（表示为 I_L 和 I_R），那么该层的损失函数计算将以下列公式表示：

$$L_{split} = \frac{1}{2}[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}] - \gamma \qquad (17)$$

式（17）中：I 表示左右两个实例集 I_L、I_R 的并集。

3）LSSA 优化 XGBOOST 模型

使用决策提升树（XGBOOST）模型进行高体鲴体态和体质量的预测是一个不断调整树模型各节点权值的过程，旨在使树模型函数持续逼近体态和体质量之间的关系。类似于常规的有监督学习，XGBOOST 模型的预测过程需要根据训练集（体长和体宽数据）预测目标变量（体质量数据）。由于模型无法一次性预测成功，因此每次预测结束后，XGBOOST 模型会新增一棵决策树，根据误差函数对前一棵树的预测结果进行调整和纠正，直至最终预测结果达到精度要求。传统 XGBOOST 模型最佳树深度、最佳学习率以及最佳迭代次数等 3 项超参数由用户随机定义，导

致模型效果无法保证。为提高 XGBOOST 拟合精度，使用混沌 SSA 算法对其 3 个主要参数进行寻优，获取最佳树深度、最佳学习率以及最佳迭代次数（图 6）。

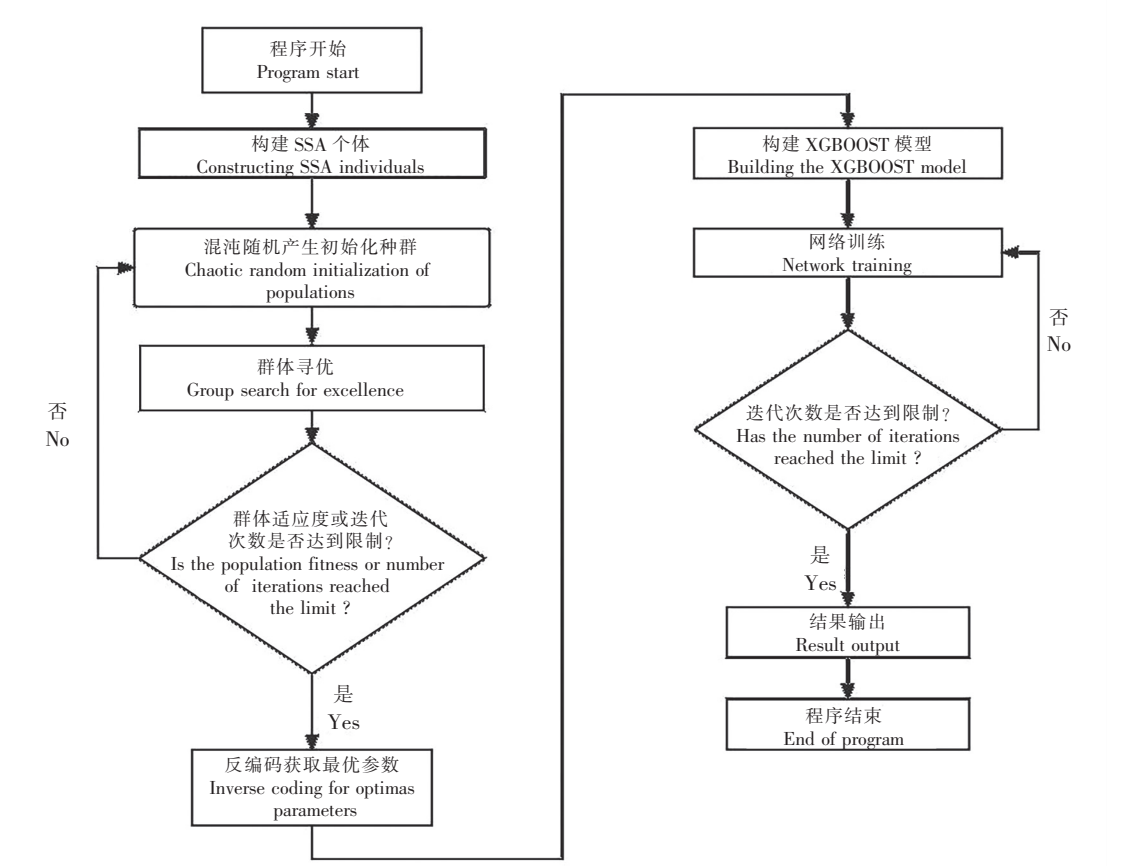

图 6 SSA–XGBOOST 算法流程图

Fig.6 SSA–XGBOOST algorithm flow chart

2 模型拟合结果

2.1 ESD 数据降噪结果

ESD 数据降噪结果如图 7、图 8 所示。采用 ESD 方法识别出 5 项异常数据，剔除了 4 个异常数据点（图中红色数据点），有效提高了模型训练精度。

将获取到的 314 组数据分别绘制体长 – 体质量、体宽 – 体质量散点图，从散点图（图 7、图 8）可以看出体长 – 体质量、体宽 – 体质量基本呈现正相关关系。样本鱼平均体长为 219 mm（标准差 σ = 2.0 mm），最大体长为 265 mm，最小体长为 155 mm；平均体宽为 62 mm（σ = 0.7 mm），最大体宽为 80 mm，最小体宽 40 mm；平均体质量为 199 g（σ = 59.0 g），最大体质量为 370 g，最小体质量仅 60 g。养殖 2 个月的单条高体鲴平均增重约 109 g。

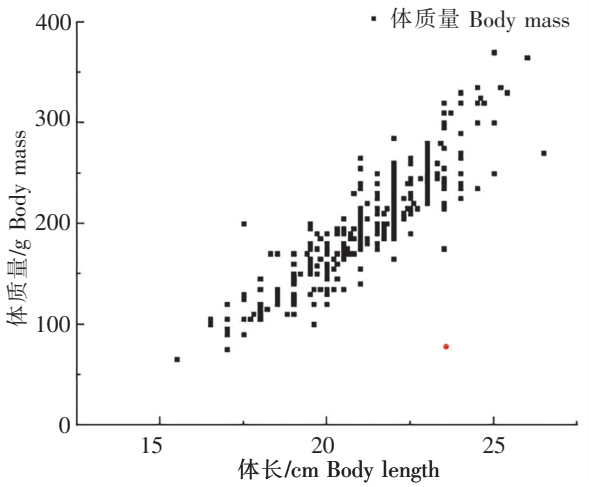

图 7 体长、体质量关系散点图(ESD 剔除)

Fig.7 Scatter diagram of body length and body mass (ESD–excluded)

注：红色点为剔除数据。图 8 同此。

Notes: The red dot represented excluded data. The same as in figure 8.

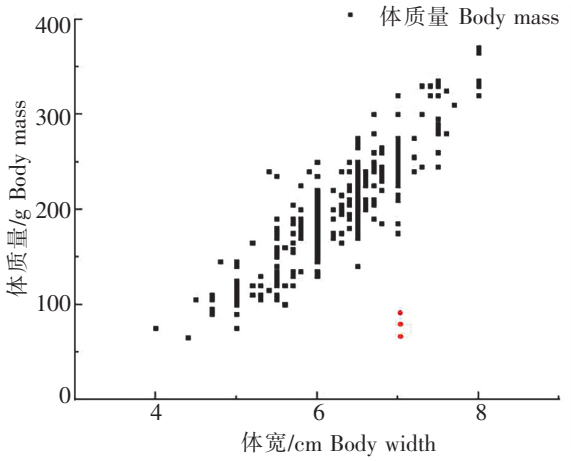

图 8　体宽、体质量关系散点图(ESD 剔除)

Fig.8　Scatter diagram of body width and body mass (ESD-excluded)

2.2　数学模型拟合结果

2.2.1　常规数学模型拟合结果

1）Gauss 曲线

使用 Gauss 曲线拟合高体鲴体长－体质量关系，拟合效果见图 9，整体数据集呈正相关趋势，数据点均匀分布在曲线两侧，曲线终点尚未达到峰值，未呈现完整的山峰形 Gauss 曲线。

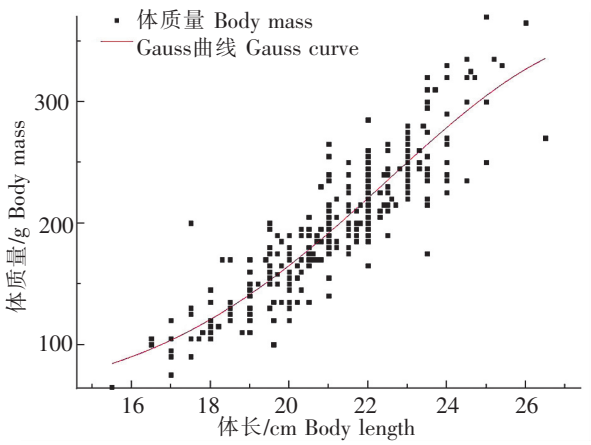

图 9　Gauss 函数拟合曲线

Fig.9　Gauss function fitting curve

2）Logistic 曲线

使用 Logistic 曲线进行高体鲴的体长、体质量关系拟合，拟合效果见图 10，整体增长较为平稳，增长速率变化不大，未呈现较为明显的 S 型曲线。

3）幂函数曲线

利用幂函数，选择体长和体质量两项因素完成高体鲴体态特征与体质量关系的拟合，拟合曲线见图 11。

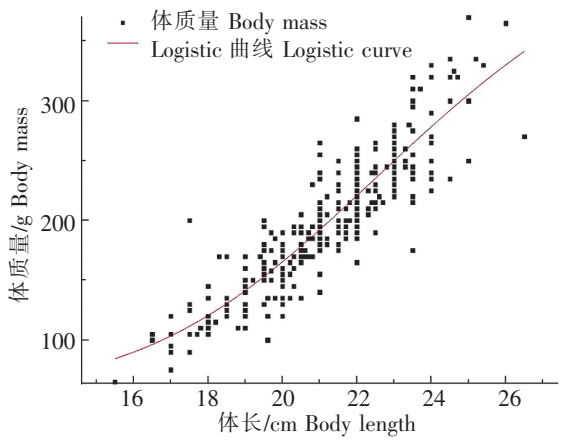

图 10　Logistic 函数拟合曲线

Fig.10　Fitting curve of Logistic function

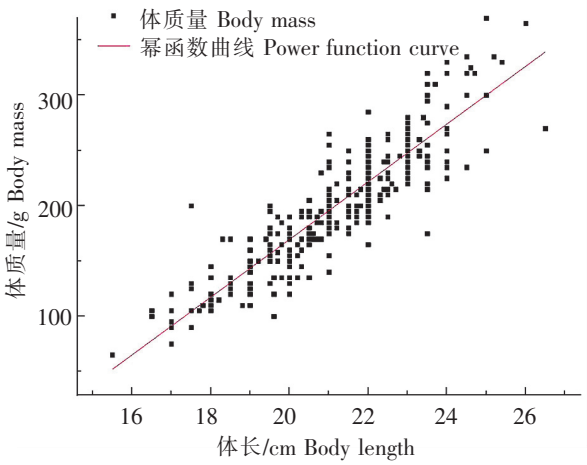

图 11　幂函数拟合曲线

Fig.11　Power function fitting curve

4）Von Bertalanffy 方程

Von Bertalanffy 方程拟合效果如图 12 所示。体长、体质量的关系式为 $W = 0.028L^{2.896\,2}$，$R^2 = 0.771\,0$。

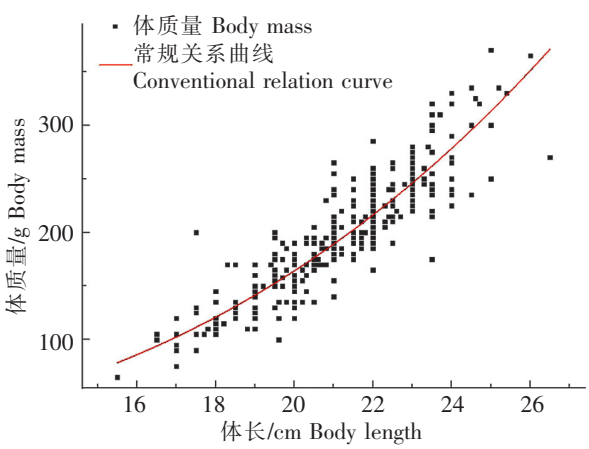

图 12　Von Bertalanffy 方程

Fig.12　Von Bertalanffy equation

2.2.2 LSSA – XGBOOST 模型拟合结果

开始实验后，将 XGBOOST 的 3 项参数作为待优化量输入 SSA 模型，SSA 模型参数设置如下：

```
fun = @ getObjValue;% 目标函数
dim = 3;% 优化参数个数
lb = [0.001, 0.001, 0.01];
% 优化参数目标下限（最大迭代次数，深度，学习率）
ub = [100, 20, 1];
% 优化参数目标上限（最大迭代次数，深度，学习率）
pop = 60;% 麻雀数量
Max_iteration = 10;% 最大迭代次数
params. objective = 'reg: linear';
% 回归函数
```

种群初始化参数设置如下：

```
Pop = 60;% 种群规模
Dim = 3;% 优化参数个数
Seed = 0.5;% 起始位置
U = 3.8;
% u 混沌序列参数，u 取 [3.569 9, 4]
```

SSA 群体适应度随迭代次数变化曲线如图 13 所示，从第三代开始，群体适应不再下降，即种群已达到最佳适应度。

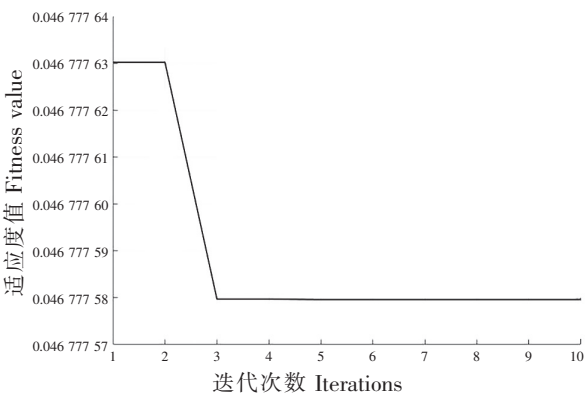

图 13 SSA 群体迭代适应度变化曲线

Fig.13 Change curve of SSA group iterative fitness

此次实验以体长、体宽两项参数为输入值预测体质量，这是由于在实验过程中发现使用体长或体宽单一参数输入预测体质量时，LSSA – XGBOOST 模型拟合度分别为 0.795 56 和 0.824 06，仅略高于部分数学模型，而使用双参数输入时拟合度有较大提升，拟合度 R^2 达到 0.944 16。预测值与真实值的拟合效果对比见图 14，在 100 个样本点的拟合跟踪中表现良好，仅丢失少量目标点。

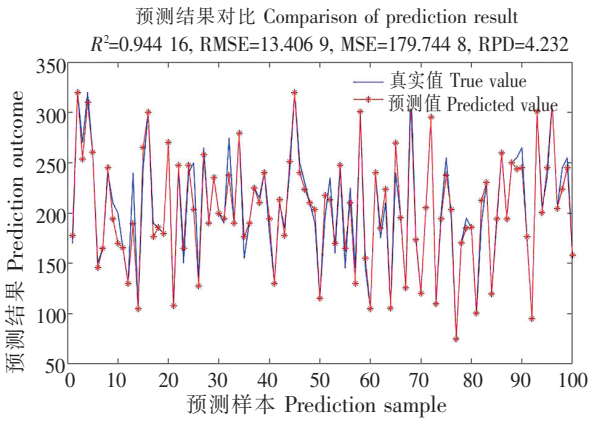

图 14 预测值与真实值对比图

Fig.14 Comparison between predicted value and real value

注：RMSE 为均方根误差；MSE 为均方误差；RPD 为相对百分比差异。

Notes: RMSE is root mean square error; MSE is mean square error; RPD is the relative percentage difference.

3 分析与讨论

3.1 与神经网络模型的对比分析

由上述数学模型拟合效果可知，针对此次高体鰤养殖实验测量数据的常规数学模型拟合并非最优方法。神经网络模型属于自适应非线性模型，大量数据表明，人工神经网络在处理常见回归拟合问题时有优异表现[28]，除传统 BP 神经网络外，多种优化 BP 模型如遗传算法优化 BP（GA – BP）、粒子群优化 BP（PSO – BP）等都具有处理回归拟合问题的能力，这些优化算法大多在 BP 神经网络初始化时采用寻优算法获取最佳的权值、阈值等初始参数，从而有效提高 BP 神经网络拟合精度。PSO – BP 是较为常见的群体寻优算法，在解决回归预测问题时常常优于 GA – BP 和传统 BP[29]。本文选用传统 BP 神经网络以及 PSO – BP 神经网络与 LSSA – XGBOOST 算法对比，结果如图 15 所示，传统 BP 神经网络拟合度 R^2 为 0.877 5，粒子群优化 BP 为 0.910 5，而本文所用 LSSA – XGBOOST 模型相关性系数 R^2 为 0.947 9。以图 15

中第 11 个点拟合效果为例，BP 和 PSO－BP 神经网络的拟合误差已经接近其最佳误差，而 LSSA－XGBOOST 每棵树模型的预测都使用 shrinkage，削弱其对结果的影响，从而提升整体模型的泛化能力，为后续训练留出更多的学习空间，有效地防止过拟合。此外，常见神经网络算法需要大量数据以支撑其算法模型的深度和训练量，从而提高预测精度，而 XGBOOST 则不需要太过庞大的数据集，这是由于决策提升树模型在训练过程中遵循确定性原则，而确定性原则使其更容易记住简单的数据变化规律，一旦规律过于复杂，其学习效果便会弱于神经网络模型。

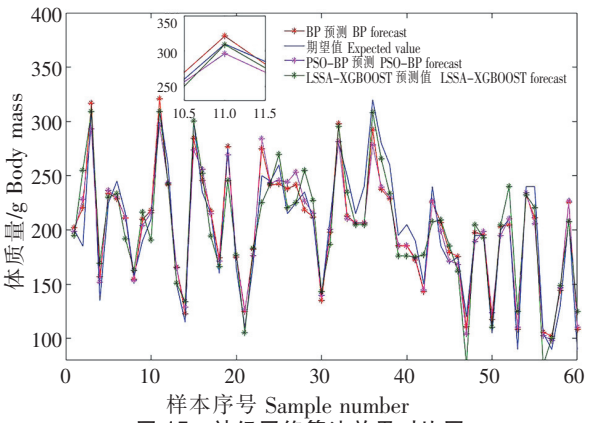

图 15　神经网络算法效果对比图

Fig.15　Comparison of neural network algorithm effects

3.2　总体对比

为使算法拟合效果对比更加直观，整理上述 7 种模型拟合度及模型输入、输出值，结果如表 1 所示。常规数学模型仅探讨单一参数输入与输出关系，故分别以体长、体宽为输入，拟合体质量关系；神经网络和改进树模型则以体长、体宽两项参数输入拟合体质量。实验对比各模型相关性系数 R^2，结果发现双参数输入的神经网络模型比单一输入数学模型的 R^2 值更高，其中优化树模型 LSSA－XGBOOST 相关性系数最高，达到 0.947 9，与 BP 神经网络和 PSO－BP 相比，其平均绝对误差（Mean absolute error，MAE）、均方误差（MSE）和均方根误差（RMSE）都有所降低，具体误差对比结果见表 2。

表 1　7 种拟合模型拟合度 R^2 对比

Tab. 1　Comparison of fitting degree R^2 of 7 fitting models

模型 Model	自变量 Independent variable	因变量 Dependent variable	相关性系数 R^2
Gauss	体长	体质量	0.822 5
	体宽	体质量	0.708 5
Logistic	体长	体质量	0.772 5
	体宽	体质量	0.708 4
幂指数 Power－exponent	体长	体质量	0.771 9
	体宽	体质量	0.709 6
$W = aL^b$	体长	体质量	0.771 0
BP	体长＋体宽	体质量	0.877 5
PSO－BP	体长＋体宽	体质量	0.910 5
LSSA－XGBOOST	体长＋体宽	体质量	0.947 9

表 2　LSSA－XGBOOST 模型与神经网络算法各项误差对比

Tab. 2　Comparison of errors between LSSA－XGBOOST model and neural network algorithm

模型 Model	平均绝对误差 MAE	均方误差 MSE	均方根误差 RMSE
BP	13.988	300.165 5	17.325 3
PSO－BP	13.716	243.015 2	16.412 4
LSSA－XGBOOST	9.042	203.533 0	14.266 5

4　结论

1）本文提出的 LSSA－XGBOOST 模型以决策提升树模型（XGBOOST）为基础进行改进，最终使得 LSSA－XGBOOST 模型在小样本数据集下有优于其他传统及改进神经网络的表现。

2）与常规数学模型拟合相比，LSSA－XGBOOST 模型拟合度相关性系数 R^2（0.947 9）提高了约 10%；与传统 BP 神经网络和 PSO－BP 相比，LSSA－XGBOOST 模型相关性系数 R^2 提升约 3%，且 MAE、MSE 和 RMSE 三项误差都有明显降低。在处理小样本数据集的回归拟合工作时，LSSA－XGBOOST 模型优于传统数学模型和常规神经网络模型，能为工船养殖高体鲫精准投喂提供理论依据，后续建议在养殖过程中扩充样本数据集，并提高混沌随机数发生器性能，将有效提高高体鲫体质量的预测精度，为饲料投喂、成鱼出仓时机判断及市场预估提供参考。

参考文献：

[1] 朱明，张镇府，黄凰，等. 鱼类养殖智能投喂方法研究进展 [J]. 农业工程学报，2022，38 (7)：38－47.

[2] 刘晓娟，沙宗尧，李大鹏，等. 基于生物能量学模型的尖吻鲈精准投喂管理辅助决策系统构建 [J]. 水生生物学报，2021，45 (2)：237－249.

[3] 沈炜皓，崔海朋，徐以军. 基于环境信息和鱼类行为的智能投喂系统研究 [J]. 中国新技术新产品，2021 (22)：33－35.

[4] 陈澜，杨信廷，孙传恒，等. 基于自适应模糊神经网络的鱼类投喂预测方法研究 [J]. 中国农业科技导报，2020，22 (2)：91－100.

[5] 林在凡，郭敬蓉，洪晓林，等. 基于 Python 的图像处理技术在鱼类尺寸测量中的应用 [J]. 福建农机，2021 (2)：38－41.

[6] Mafalda F，Pedro I，Manuel B. Length－weight relationships for eight Chondrichthyes from the north－eastern Atlantic Ocean [J]. The Egyptian Journal of Aquatic Research，2023，49 (1)：87－90.

[7] Dinh Q M，Nguyen T H D，Nguyen－Ngoc L，*et al*. Temporal variation in length－weight relationship，growth and condition factor of *Acentrogobius viridipunctatus* in the Mekong Delta，Viet Nam [J]. Regional Studies in Marine Science，2022，55：102545.

[8] Shuman L A，Selyukov A G，Nekrasov I S，*et al*. Data on length－weight and length－length relationships，mean condition factor，and gonadosomatic index of *Rutilus rutilus* and *Perca fluviatilis* [J]. Data in Brief，2022，42：108067.

[9] Kaka R M，Jung'a J O，Badamana M，*et al*. Length－weight relationships of wild penaeid shrimps in Malindi－Ungwana Bay：implications to aquaculture development in Kenya [J]. The Egyptian Journal of Aquatic Research，2019，45 (2)：167－173.

[10] Sepa P，Coello D，Herrera M，*et al*. Length－weight relationship of four deep－sea chondrichthyans (*Elasmobranchii* & *Holocephali*) in Ecuadorian oceanic waters [J]. The Egyptian Journal of Aquatic Research，2022，48：397－399.

[11] Najmudeen T M，Zacharia P U，Seetha P K，*et al*. Length－weight relationships of three species of pelagic sharks from southeastern Arabian Sea [J]. Regional Studies in Marine Science，2019，29：100647.

[12] 陈锋，刘艳超，魏聪，等. 察隅弧唇裂腹鱼仔稚鱼体长体重关系研究 [J]. 西藏科技，2020 (10)：11－12.

[13] 林雅蓉，王荣，高尚武. 胶州湾桡足类的生物学研究——Ⅱ. 小拟哲水蚤的个体重量测定及体长－体重关系 [J]. 海洋科学，1987 (5)：41－45.

[14] 张志伟，胡伍生，黄晓明. 回归拟合模型的神经网络方法 [J]. 测绘科学，2010，35 (S1)：39－41.

[15] 韩超. 深远海养殖：走向“蓝海”的朝阳产业 [J]. 农产品市场，2021 (12)：26－28.

[16] 范斌，古恒光，杨永健. 卵形鲳鲹深水网箱深远海绿色健康养殖技术研究与应用 [J]. 中国科技成果，2020，21 (19)：49－51.

[17] 陈昌生. 高体鰤人工繁殖、育苗及养殖技术的研究 [C] //中国海洋湖沼学会，中国动物学会鱼类学分会. 中国动物学会鱼类分会2004年学术研讨会摘要汇编. 重庆：中国海洋湖沼 (动物) 学会鱼类学分会，2004：77.

[18] 陈昌生，黄佳鸣，叶加松，等. 高体鰤人工育苗技术研究 [J]. 水产学报，1998，22 (1)：40－44.

[19] 王雅英. 高体鰤鱼胚胎发育过程初步报告 [J]. 中国水产，2001 (3)：59－60，53.

[20] 廖志强. 高体鰤网箱养殖技术 [J]. 中国水产，2003 (12)：60－61.

[21] 周梦，吕志刚，邸若海，等. 基于小样本数据的BP神经网络建模 [J]. 科学技术与工程，2022，22 (7)：2754－2760.

[22] 曹晨曦，田友琳，张昱堃，等. 基于统计方法的异常点检测在时间序列数据上的应用 [J]. 合肥工业大学学报 (自然科学版)，2018，41 (9)：1284－1288.

[23] 陈伟，吴布托，裴喜平. 风电机组异常数据预处理的分类多模型算法 [J]. 电力系统及其自动化学报，2018，30 (4)：137－143.

[24] 黄真理，常剑波. 鱼类体长与体重关系中的分形特征 [J]. 水生生物学报，1999，23 (4)：330－336.

[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.

[26] 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.

[27] Chen T Q，Guestrin C. XGBoost：a scalable tree boosting system[EB/OL]. [2023－03－09]. http://dl.acm.org/doi/10.1145/2939672.2939785.

[28] 唐亦舜，徐庆，刘振鸿，等. 基于优化非线性自回归神经网络模型的水质预测 [J]. 东华大学学报 (自然科学版)，2022，48 (3)：93－100.

[29] 邓森，李希建，徐畀泽. 基于PSO－BP神经网络的煤体瓦斯渗透率预测 [J]. 矿业工程研究，2022，37 (4)：35－41.

Body mass of *Seriola dumerili* prediction model based on LSSA – XGBOOST improved algorithm

YU Guoyan[1,2], ZUO Renyi[1,2], YAN Jun[1*], LUO Yingtong[2], ZHU Qiheng[2]
(1. Guangdong Provincial Laboratory of South Marine Science and Engineering, Zhanjiang 524013, China;
2. School of Mechanical and Power Engineering, Guangdong Ocean University, Zhanjiang 524088, China)

Abstract: In order to build an accurate feeding model using mass judgment and obtain the body mass state of *S. dumerili* in real time, this study built a body mass prediction model based on LSSA – XGBOOST algorithm. Firstly, the data of body length, body width and body mass measured by the culture experiment ship were detected by extreme studentized deviate (ESD) method, and the abnormal points were removed. Secondly, the chaotic random number generator was used to complete the initial population optimization of the SSA algorithm to improve its searching ability. The optimized SSA algorithm was used to optimize three parameters of the optimal tree depth, the optimal learning rate and the optimal number of iterations of the XGBOOST model. Finally, the LSSA – XGBOOST model with body length and body width as the input and body mass data as the output was constructed. The experimental results showed that, compared with the conventional mathematical model fitting, the LSSA – XGBOOST model fitting correlation coefficient R^2 increased by about 10%. Compared with the traditional BP neural network and PSO particle swarm optimization BP, the error square and R^2 were improved by about 3%, and the mean absolute error (MAE), mean square error (MSE) as well as the root mean square error (RMSE) were significantly reduced. It could be seen that LSSA – XGBOOST model was more accurate for predicting the body mass of small samples of *S. dumerili*, and the construction of LSSA – XGBOOST model was greatly significant for users to grasp the growth state of *S. dumerili* and build accurate feeding model for mass judgment.

Key words: LSSA – XGBOOST; *Seriola dumerili*; body length; body mass; relationship