doi:10.6041/j.issn.1000-1298.2025.10.018

基于滞后因果分析和改进 SSIM 的溶解氧含量数据插补算法

刘世晶[1,2]　张佳鹏[1]　钱　程[1]　涂雪滢[1]　聂鹏程[3]

(1. 中国水产科学研究院渔业机械仪器研究所, 上海 200092; 2. 中国海洋大学三亚海洋研究院, 三亚 572011;
3. 浙江大学生物系统工程与食品科学学院, 杭州 310058)

摘要: 面向精细化和智慧化水产养殖领域水环境数据系统性或意外丢失问题,本文提出了一种基于滞后因果分析和改进 SSIM(Sequence-to-sequence imputation model)的溶解氧含量数据插补算法。首先,采用固定采样频率收集水质和气象等环境数据,基于 GC(Granger causality)理论,利用 cLSTM(Component-wise long-short term memory)方法分析不同环境变量与溶解氧含量时序数据的滞后相关性,选择因果关系显著的环境变量构建训练样本集。其次,采用 SSIM 为基础框架实现溶解氧含量缺失数据插补,并提出一种结合两层 BiLSTM(Bidirectional long short-term memory)结构和 Dropout 正则化的 SSIM 模型优化方法,提升模型对复杂特征的表征能力。试验结果表明,本文方法有效提升数据插补精度,1 h 缺失数据修复平均绝对误差、均方根误差分别达到 0.04、0.05,3 h 缺失数据误差分别为 0.16、0.17,5 h 缺失数据误差分别为 0.43、0.45,为水产养殖数据质量控制提供了有效的技术支撑。

关键词: 智慧水产养殖; 溶解氧含量预测; SSIM; NGC; BiLSTM; 数据插补算法

中图分类号: S951.2　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)10-0175-09　　**OSID**:

Incomplete Dissolved Oxygen Data Interpolation Based on Lagged Causality Analysis and Improved SSIM

LIU Shijing[1,2]　ZHANG Jiapeng[1]　QIAN Cheng[1]　TU Xueying[1]　NIE Pengcheng[3]

(1. *Fishery Machinery and Instrument Research Institute*, *Chinese Academy of Fishery Sciences*, *Shanghai* 200092, *China*
2. *Sanya Oceanographic Insitution*, *Ocean University of China*, *Sanya* 572011, *China*
3. *College of Biosystems Engineering and Food Science*, *Zhejiang University*, *Hangzhou* 310058, *China*)

Abstract: Based on the industrial development demands of intelligent aquaculture production operations in aquaculture facilities, an incomplete dissolved oxygen data imputation algorithm was proposed based on lag causality analysis and an improved sequence-to-sequence imputation model (SSIM) to address issues related to systematic or accidental loss of aquaculture environmental data. Firstly, environmental data such as water quality and meteorological data were collected at fixed sampling frequencies. Using the granger causality (GC) theory, the component-wise long short-term memory (cLSTM) method was employed to analyze the lag correlation between different environmental variables and dissolved oxygen time series data. The environmental variables with significant causal relationships were selected to construct the training sample set. Secondly, the SSIM framework was used to implement the imputation of missing dissolved oxygen data, and an optimization method of the SSIM model was proposed by combining a two-layer bidirectional long short-term memory (BiLSTM) structure and Dropout regularization to improve the model's ability to represent complex features. Experimental results showed that the proposed method effectively improved the data imputation accuracy. The mean absolute error (MAE) and root mean square error (RMSE) for 1-hour missing data reached 0.04 and 0.05, respectively; for 3-hour missing data, the errors were 0.16 and 0.17, and for 5-hour missing data, the errors were 0.43 and 0.45. The research result can provide effective technical support for data quality control in aquaculture.

Key words: smart aquaculture; dissolved oxygen content prediction; SSIM; NGC; BiLSTM; data interpolation algorithm

收稿日期: 2025-07-04　修回日期: 2025-07-29

基金项目: 农业农村部科技项目

作者简介: 刘世晶(1982—),男,副研究员,主要从事渔业信息化、图像处理和机器视觉研究,E-mail: liushijing@fmiri.ac.cn

通信作者: 钱程(1995—),男,研究实习员,主要从事渔业信息化、图像处理和机器视觉研究,E-mail: qiancheng@fmiri.ac.cn

0 引言

水产养殖水体溶解氧含量是反映水质情况、生物生存状况的重要指标[1]。目前,利用传感技术获取溶解氧含量等关键水质数据已经成为现代水产养殖重要组成部分[2]。然而,在实际使用过程中,大多数传感活动都存在系统性或意外的数据丢失问题,增加了后续数据处理和分析的难度。因此,开展水产养殖溶解氧含量数据修补方法研究,及时全面掌握溶解氧含量历史数据,对指导养殖生产、辅助养殖决策、减少养殖风险具有重要意义。

数据插补是基于已知数据点构造缺失值的过程[3]。当时间序列几乎没有离散缺失值时,使用统计数据进行插补[4-6]可以获得良好的缺失数据估计。此外,利用同一时间相关数据恢复缺失数据同样可以取得较好的效果[7-8]。在水产养殖实际生产过程中,受传输网络和断电等问题影响,感知数据往往在同一时间段内同时丢失,导致缺失值存在连续非强相关特性,因此传统的基于统计数据和同时相关数据的预测方法无法满足类似完全随机缺失数据插补的需要[9]。

在利用时间序列数据开展预测分析时,探明时间序列因果关系是一个基础性工作,加入不相关的时序特征会降低预测精度并产生额外的计算成本。传统的主成分分析、随机森林等面向数据关联性的降维方法,无法有效揭示时间序列之间的依赖性[10]。GC 是理解序列之间因果关系的常用框架,可以有效量化分析不同时间序列在预测另一个时间序列的未来演变方面的帮助程度[11],已经在经济学[12]、气象科学[10]、神经科学[13]等不同领域得到应用。但传统 GC 方法适用性通常受到线性和稳定性假设的限制,无法满足水质、气象等具有明显非线性特征的时序数据分析要求。NGC(Neural granger causality)通过整合神经网络学习技术能够有效地捕捉数据中的非线性关系和复杂的动态模式,已经成为非线性时间序列分析重要研究方向[14],但相关研究主要集中在经济学和生命科学场景,面向渔业领域研究较少。

近年来,基于序列到序列的循环神经网络(Recurrent neural network,RNN)在机器翻译[15]、图像字幕生成[16]、语音识别[17]等数据分析应用领域取得了良好的表现。序列到序列模型的大多数实现都使用长短时记忆网络(Long short-term memory,LSTM)作为隐藏单元[18],该体系结构能够利用更多的上下文和以前的序列规则,有效增强了网络对长期时间依赖性的模拟能力,使其在时序数据分析中表现出更强的适应性[19]。在小样本场景下,LSTM 凭借其参数效率高的优势表现出更强的稳定性,这一特性使其在资源受限的应用场景中具有独特价值,一些学者对此进行了研究[20-22]。SSIM 模型在传感器数据修补中取得了较好的效果,但是针对实际养殖环境,水体中溶解氧含量不仅受到气温、气压、降雨量、风速、风向等外在环境因素影响,还与 pH 值、水温甚至溶解氧本身有着紧密的相关性,使其时序数据具有多维度、大时滞等特点,增加原有时间序列的复杂程度。相较而言,基于 Transformer 的模型虽然能通过自注意力机制更好地捕捉这类复杂依赖关系[23],但其注意力权重矩阵难以直观理解,且需要大量训练数据,这对水产养殖等数据有限的领域构成显著挑战。此外,环境数据往往受季节变化、气候变化、突发天气事件等多种因素影响,数据分布非固定、非线性特征明显,基于浅层结构预测模型无法有效处理和利用复杂的时间依赖关系,制约了技术的深度应用。

针对上述问题,本文提出一种基于 NGC 分析和改进 SSIM 的池塘溶解氧含量数据插补方法。该方法选用 cLSTM 方法构造 NGC 因果分析框架,分析不同环境数据与溶解氧的滞后相关性,并根据相关性筛选关键影响因子,以期更精准地利用影响池塘溶解氧的环境参数;然后,构造一种结合两层 BiLSTM 结构和 Dropout 正则化的 SSIM 改进模型,通过增加网络的深度,进一步提高模型对复杂时间序列数据的建模能力,面向网络层数增加可能引起的过拟合风险,在输入层、隐藏层引入了多维 Dropout 正则化,通过随机丢弃网络中的部分神经元,减少过拟合的风险,并利用改进 SSIM 方法构造池塘溶解氧含量数据插补模型,以期提高溶解氧含量缺失数据修补精度;最后,通过与不同插补方法比较,验证本文方法的有效性。

1 数据获取与预处理方法

1.1 数据获取

试验数据采集自“渔机 1 号”试验平台,试验时间为 2024 年 8 月 6 日—9 月 3 日。试验区域选择平台 1 号养殖仓作为试验对象,养殖品种为大黄鱼。为了避免增氧和投喂过程对水质监测数据造成影响,水质传感器采用浮标形式放置在养殖舱远离投喂口处,传感器数据采用无线网络传输,传输数据包括溶解氧含量、pH 值以及水温,采集频率为 15 min/条,采样水深为 40 cm。气象站放置在平台顶部,同样采用无线信号传输,上传数据包括气温、气压、风速、风向、大气压、光照强度、降雨量等 7 种气象参数,采

集频率与水质传感器一致。采用自行开发的组态程序按照标准时间戳将控制数据、水质数据和气象数据存储到数据库内。

1.2 数据预处理

在实际采样过程中，人为固定了传感器采样频率，保证采集的水质数据和气象数据具有相同的分布特征。针对传感器网络，受信号干扰、网络延迟以及人为因素影响，难免会出现随机数据缺失、异常等状况，为了保证数据修复时间序列的完整性，对缺失值少于等于 2 个采样时间间隔的数据采用线性插值[24]方法进行修复，并对数据进行归一化处理，即

$$x_{i\text{norm}} = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}} \tag{1}$$

式中　x_i——原始单因子数据

$x_{i\text{norm}}$——归一化单因子数据

$x_{i\min}$——数据中的最小值

$x_{i\max}$——数据中的最大值

2 插补模型构建

2.1 非线性格兰杰因果关系

传统的 GC 方法主要分析变量间的线性因果关系，无法满足水质、气象等非线性动态变化趋势分析要求。为了解决多维环境时间序列变量的 Granger 因果发现问题，构造了非线性自回归线性模型，即

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{F}^{(k)} \boldsymbol{x}_{t-k} + \boldsymbol{e}_t \tag{2}$$

其中

$$\begin{cases} \boldsymbol{x}_t = (x_{t1}, x_{t2}, \cdots, x_{tp}) + \boldsymbol{\varepsilon}_t \\ x_{ti} = g_i(x_{t-1}, x_{t-2}, \cdots, x_{t-k}) \end{cases} \tag{3}$$

式中　$\boldsymbol{F}^{(k)}$——系数矩阵

$\boldsymbol{x}$——p 维非平稳环境变量时间序列

t——时间序列长度

$\boldsymbol{x}_t$——t 时刻各个时间序列的值

K——时间序列滞后阶数

$g_i(\cdot)$——每个时间序列 i 对应的非线性函数

$\boldsymbol{e}_t$、$\boldsymbol{\varepsilon}_t$——零均值噪声

本文采用层数为 1 的 cLSTM 网络构造训练模型[14]，层权重表示为 $W_i = \{W_i^{(1)}, W_i^{(2)}, \cdots, W_i^{(K)}\}$。对于所有的滞后阶数 K，如果第 1 层对应的权重都为零，则认为输入序列 i 与输出之间没有格兰杰非因果性。因此为了有效评价不同序列在特定滞后阶数下的耦合相关性，采用 group lasso 损失函数，即

$$\min Y = \min_{W} \sum_{t=2}^{T} (x_{ti} - g_i(x_{<t}))^2 + \gamma \sum_{i=1}^{p} \| W_i \|_2 \tag{4}$$

式中　Y——目标函数

γ——损失权重

T——预测所覆盖的最大时间步数

本文利用水质和气象等多维数据进行因果分析，多个变量的组合可能产生大量的假阳性，为了有效提升高维环境数据比较准确性，采用近端梯度下降法优化式(4)，即

$$W_{n+1} = \text{sign}(W_n - \varphi_n \nabla Y_n) \cdot \max(0, \| W_n - \varphi_n \nabla Y_n \|_2 - \varphi_n \gamma) \tag{5}$$

式中　φ_n——第 n 次迭代的学习率

∇Y_n——目标函数的梯度

2.2 改进 SSIM 模型

改进 SSIM 结构如图 1 所示，与 SSIM 相同，改进模型仍然使用基于注意力机制的序列到序列架构，且编码器和解码器是两个关键功能组件。本文主要利用气象和水质数据预测在季节变化、气候变化、突发天气事件和人为操作(如增氧和投喂)影响下的复杂过程，由于不同因素存在隐藏的影响属性，因此采用双层时间序列结构构造编码器，用于处理复杂时间序列输入并将其映射到高维向量。同时，为了降低层数增加带来的过拟合风险，在输入层和隐藏层引入了 Dropout 正则化。

2.2.1 双层双向 LSTM

与基于过去数据的一般预测问题不同，数据插补是对缺失数据进行修复，可以利用过去和未来数据重构缺失值。因此，选择 BiLSTM 作为编码器，以充分利用已有数据。BiLSTM 输入是可变长度序列 $y = \{\boldsymbol{y}_i\}$，其中 $\boldsymbol{y}_i$ 为特定时间长度索引 i 处的特征向量。不同于单向 LSTM，BiLSTM 具有前向和反向两个隐藏状态，因此在每个时间索引 t 处的隐藏状态为 $h_t = \{\overrightarrow{h}_t, \overleftarrow{h}_t\}$，由此构建的隐藏序列 $h = \{h_t\}$。则时间索引 t 处的隐藏状态更新为

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + b_f) \tag{7}$$

$$\boldsymbol{c}_t = f_t \otimes \boldsymbol{c}_{t-1} + i_t \otimes \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \tag{8}$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o) \tag{9}$$

$$h_t = \{\overrightarrow{o}_t \otimes \tanh(\overrightarrow{\boldsymbol{c}}_t), \overleftarrow{o}_t \otimes \tanh(\overleftarrow{\boldsymbol{c}}_t)\} \tag{10}$$

式中　i、f、o——输入门、遗忘门、输出门

$\boldsymbol{c}$——单元向量

i_t、f_t、o_t——t 时刻下输入门、遗忘门和输出门

$\boldsymbol{c}_t$——t 时刻下单元向量状态

σ——sigmoid 函数

$\otimes$——元素乘法运算符

b——偏置变量

本文中，缺失数据存在非线性、大时滞特征，增加了时间序列数据的复杂度，因此采用如图 1 所示双层 BiLSTM，以提高模型处理复杂数据的能力。

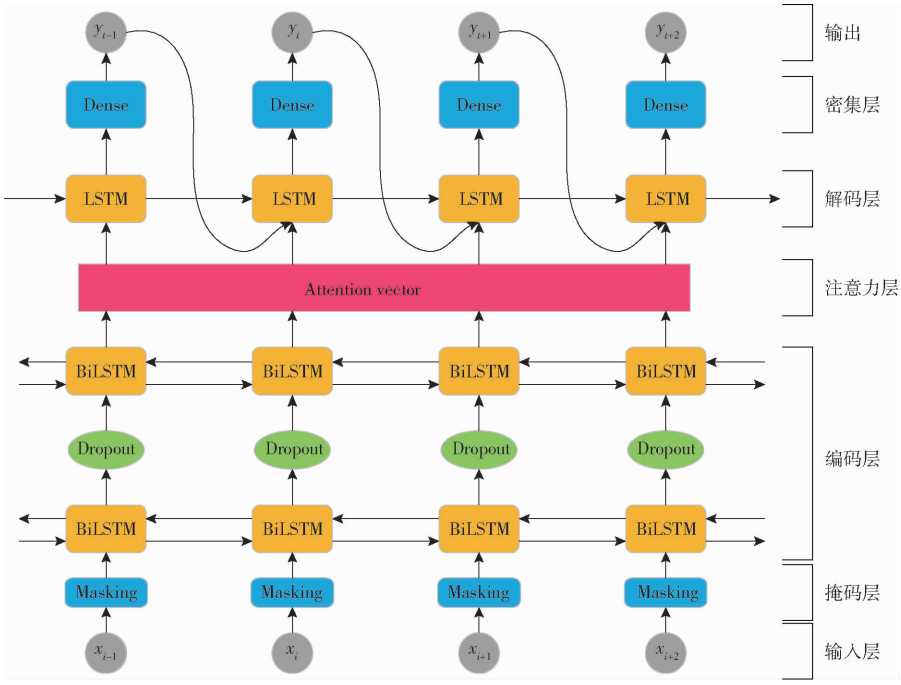

图 1 改进 SSIM 模型网络结构

Fig. 1 Improved network structure of SSIM model

2.2.2 Dropout 策略

Dropout 是一种正则化技术,通过在训练过程中随机忽略神经网络中部分神经元及其连接,从而减少过拟合。这一过程有效地模拟了不同的网络结构,从而增强了模型的泛化能力。为应对增氧和投喂过程中产生的非典型时间序列数据,Dropout 被应用于两个 BiLSTM 层之间(如图 1 所示),减少神经元之间的过度依赖,然后结合 SSIM 原有的 Dropout 共同形成本文的 Dropout 策略。

2.2.3 解码器

解码器负责递归生成输出序列 $y=\{y_1,y_2,\cdots,y_m\}$,是在 LSTM 层的顶部添加一个具有线性激活函数的完全连接层,以生成具有连续值的预测。对于解码器,在时间 t 处的估计 y_t 计算式为

$$y_t=\text{Linear}(W[s_t,c_t]+b) \tag{11}$$

$$s_t=\text{LSTM}(y_{t-1},s_{t-1},c_t) \tag{12}$$

式中 s_t——解码器在时间索引 t 处的隐藏状态

c_t——编码器传递的隐藏状态加权和

2.3 模型评价指标

为了评价模型对缺失数据的插补性能,本文选用均方根误差(RMSE)、平均绝对误差(MAE)、均方根误差困惑度(PPL_{RMSE})以及平均绝对误差困惑度(PPL_{MAE})作为评价指标[25]。

3 结果与讨论

3.1 试验平台

试验平台配置如下:处理器为 Intel W-2245,CPU 主频为 4.7 GHz;内存容量为 32.00 GB;程序设计语言为 Python 3.7,集成开发环境为 Anaconda 3。程序设计中,改进 SSIM 模型由基于 PyTorch 的神经网络框架实现和优化,BP 神经网络和最小二乘支持向量回归均结合 Python 所提供的工具箱实现。

3.2 数据集构建

受水体体量和水层交换影响,水质变化具有低频特性,因此在试验基地实际养殖生产中,将水质和气象采样时间间隔设定为 5 min。基地内增氧机和投喂机全部采用工控软件实现定时自动投喂,在保证开关时长的情况下,为了不失一般性采用随机时间设定设备开关时间,例如增氧机开启时间为 12:30,关闭时间为 14:35。针对标准化水产养殖平台,由于其良好的内控机制,使得发生内部断电等意外情况修复时间不会超过 3 h,因此在常规数据修复数据构建时,数据缺失时间定为 3 h。此外,为了评估本文方法恢复缺失数据的性能,增设 1 h 和 5 h 两个缺失时间段作为试验案例。

由于实际生产过程中存在设备维护更换、网络故障、电网断电等影响数据采集的不可预知因素,本文在 28 d 试验周期内,共获得有效试验数据 2 601 组。为了增加训练样本数量,增加数据利用率,并采用滑动窗口方法构造数据集,其中历史及未来时间的单位取样时长等于缺失时间长度,且最小采样时间为 1 个单位取样时长,最大采样时间为 3 个单位取样时长。上述取样方法能够有效增加数据利用率,可满足体量不大的数据集数据修复的需要。不

同试验案例构造数据集如表 1 所示。

表 1　不同试验案例数据集数量

Tab. 1　Number of data sets of different trial cases

时间跨度/h	数据集样本总量	训练集样本数	测试集样本数
1	57 792	46 792	11 000
3	25 536	20 536	5 000
5	14 784	11 784	3 000

3.3　关键影响因子

影响池塘溶解氧的因素众多，若将所有数据不加分别的直接利用，会导致模型庞大，增加数据分析利用难度。在实际生产场景中，不同环境变量时间序列的过去与未来之间主要存在非线性依赖关系，不相干数据可能会导致局部收敛。因此，为了探寻影响养殖水体溶解氧含量变化的关键因素，需对影响因素进行关联性分析，去除干扰变量，明确主要影响因子。本文利用 cLSTM 方法实现全变量耦合分析，并引入近端梯度下降提升模型的稀疏表达能力，并根据全变量相关性矩阵，筛选出池塘溶解氧关键影响因子。不同滞后阶数相关性矩阵如图 2 所示。

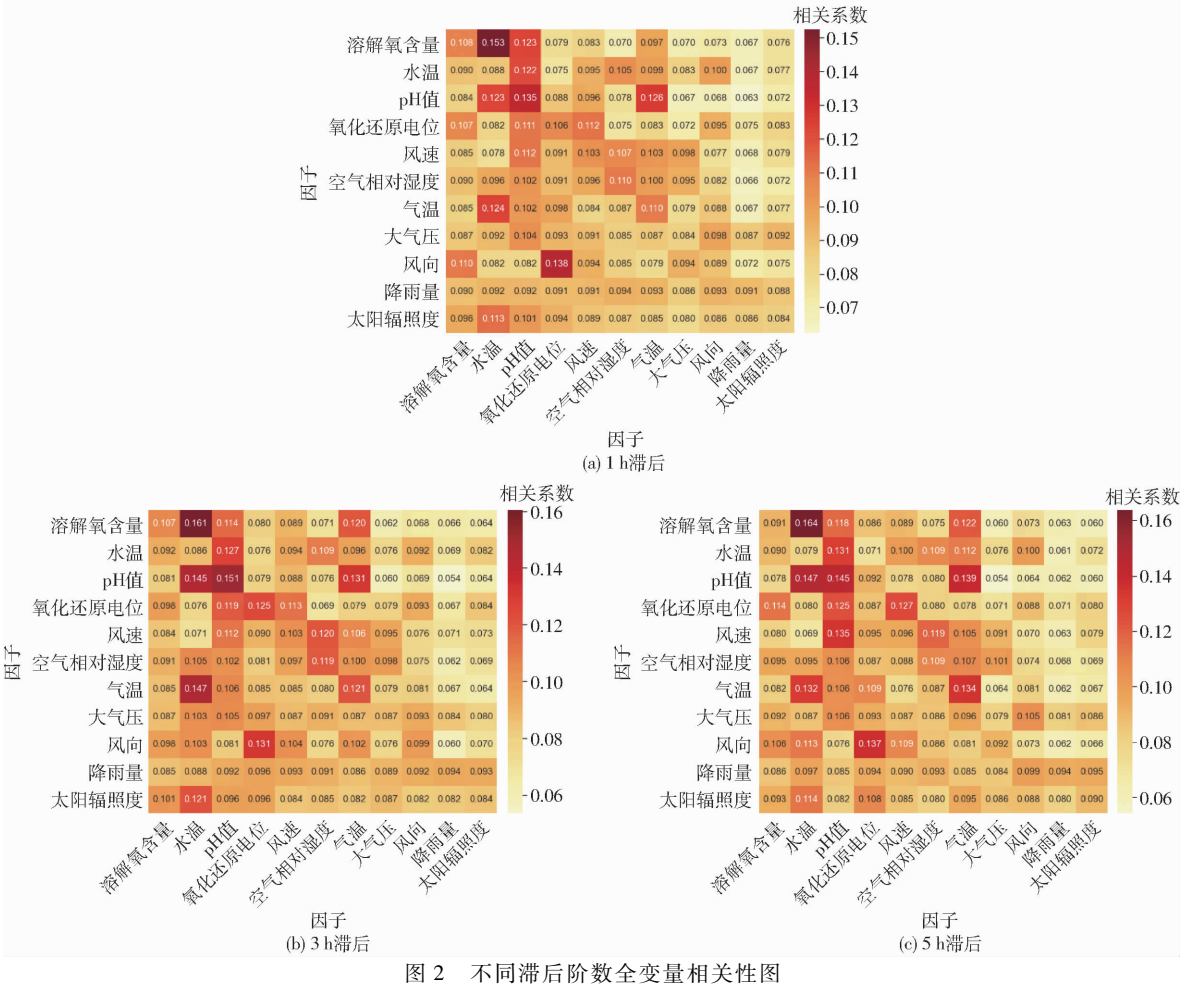

图 2　不同滞后阶数全变量相关性图

Fig. 2　Correlation plot of all variables for different lags

3.3.1　自相关性分析

如图 2 所示，在 3 个滞后时间段内，水温、空气湿度、pH 值与溶解氧含量之间的滞后相关性较为显著，且这些变量在不同滞后期的相关性数据序列保持一致。相比之下，溶解氧含量的自相关性较低，相关系数仅约为 0.09，这一数值显著低于水温、空气湿度和 pH 值之间的互相关系数。表明溶解氧含量在时间序列中的变化趋势缺乏较强的惯性或一致性，其变化更多地受到外部多种环境因素的综合作用。因此，单纯依赖溶解氧含量自相关性进行预测可能无法满足精准预测的要求，在建立相关滞后预测模型时，应考虑更多的环境变量及其交互作用。

3.3.2　互相关性分析

溶解氧含量与水温、空气相对湿度、pH 值的相关系数均超过 0.1，其中水温与溶解氧含量的相关系数最高，达 0.16。水温升高不仅加剧氧分子逸出[26]，还加速生物呼吸耗氧[27]，导致溶解氧供需失衡，本文数据结果支持了这一机制在本研究场景下的主导作用。虽然水温的直接相关性显著，通过本文数据进一步揭示了其在某些特定情境下可能存在的更复杂非线性影响。空气相对湿度与溶解氧含量的相关系数为 0.12，尽管空气相对湿度对溶解氧含

量的直接影响较弱,但其通过调节水体的蒸发速率,可能间接影响溶解氧含量。pH 值与溶解氧含量的相关系数为 0.11,表明在一般情况下,pH 值变化对溶解氧含量的影响较小。然而,在某些特定水体环境中,极端的 pH 值变化可能会加剧氧气的消耗或释放,这一效应已被模型有效捕捉,表明尽管 pH 值的直接相关性较弱,其非线性影响在特定情况下仍可能显著。氧化还原电位(ORP)和气压与溶解氧含量的相关系数均略高于 0.08,但低于前述 3 个变量。尽管相关研究表明大气压直接影响氧气溶解度,但本研究区域的气压变化范围可能不足以产生强烈的溶解氧波动,导致其在本数据集中的直接统计相关性较弱。ORP 作为水体氧化还原状态的综合指标,其与溶解氧的关系复杂,可能受到多种生物化学过程的综合影响。其他 6 个变量与溶解氧含量的相关系数普遍低于 0.08,表明溶解氧含量的变化与水温、空气湿度、pH 值之间的关系较为显著,而与其他环境因子(如风向、风速等)的关系较为微弱。因此,基于水温、空气湿度和 pH 值等关键变量建立的预测模型,可能具有较高的有效性,而将其他相关性较低的变量纳入模型,则可能不会显著提升预测效果。

此外,本研究构建的预测模型聚焦于水温、空气湿度和 pH 值等常用可测变量。验证结果表明,该模型能取得良好的预测效果。这并非忽视实际生产中众多未监测因素(如特定水质参数和生物活动)对溶解氧含量的影响,而是表明在当前研究的数据条件下,模型能够有效学习水温、湿度和 pH 值动态中所蕴含的溶解氧含量关键变化模式及部分隐含的复杂影响,从而实现对溶解氧含量的准确预测。未来研究在获取更全面数据后,可进一步整合这些关键因素以提升模型的普适性和解释性。

3.3.3 交互效应分析

为了进一步探讨变量之间的交互效应,本文构建了一个基于 LSTM 框架[25]的单层预测模型,按照图 2 滞后相关性图,分别选择与溶解氧含量相关性最强的 4 参数、6 参数、8 参数和全参数对 1、3、5 h 滞后时间内的溶解氧进行预测,预测结果如表 2 所示。

表 2 不同变量组合和不同滞后时间溶解氧含量预测均方误差

Tab.2 Prediction MSE for different variable combinations and lag times mg/L

时间跨度/h	4 参数	6 参数	8 参数	全参数
1	2.71	2.31	2.76	3.48
3	2.33	2.07	2.47	3.11
5	2.13	1.48	2.21	3.09

从表 2 中可以看出,不同的变量组合和滞后时间对溶解氧含量预测的均方误差(MSE)[23]有明显的影响。整体而言,6 参数组合在 1、3、5 h 的预测效果均优于其他变量组合,尤其是在 5 h 滞后期,6 参数组合的 MSE 最低,仅为 1.48,这表明关键环境变量在较长的滞后时间下影响作用逐渐显现,因此在面对不同滞后分析时应慎重选择预测对象。进一步分析这些预测结果,可以看出,尽管溶解氧含量与水温、空气湿度和 pH 值的线性相关性较强,而氧化还原电位和气压的线性相关性相对较低,但基于交互效应分析,增加两个变量后模型预测效能得到有效提升,说明这些非线性关系对预测结果具有重要影响。然而,增加更多的环境变量(如风速、降雨量等)并未显著改善模型的预测效果,反而增加了模型的复杂度,导致了预测性能的下降。这与之前的分析结果一致,即溶解氧含量主要受到水温、空气湿度和 pH 值等关键因素的影响,但从交互作用的角度来看,仍需引入少量相关性较低的变量,以进一步提升模型的预测能力。此外,由于其他相关系数低于 0.08 的变量关联性较弱,因此对模型预测效果的提升作用有限,过多的变量不仅未能显著提高预测精度,反而可能使模型面临过拟合的风险,从而降低其泛化能力。因此,在构建预测模型时,除了考虑预测变量与不同变量间的互相关作用,还应更加关注多变量间的非线性交互效应。综上,本文在综合分析的基础上,选择滞后相关性最高的 6 个参数作为降维结果,用于后续溶解氧含量数据修复。

3.4 改进 SSIM 结构参数

在本研究中,模型的主要超参数设置如下:编码器部分采用双层结构,包含两个隐藏层,解码器部分设置为单层隐藏层。每层 LSTM 的隐藏单元数量为 30,丢失率(Dropout)设定为 0.3。优化器选用 Adam,损失函数采用均方误差(MSE),匹配尺寸设定为 60。均方误差具有数学性质上的良好表现,其函数形式平滑可微,梯度易于计算,因此被选作训练过程中改进 SSIM 的损失函数。

3.5 试验结果分析

3.5.1 数据降维的有效性

针对多维数据耦合利用分析问题,不同的数据维度往往包含冗余或噪声特征,可能导致模型在训练过程中出现过拟合风险,降低模型的泛化能力。本文收集了池塘水质、气象等不同维度的信息,并利用 NGC 分析提取了与溶解氧关联度较高的数据变量,从表 3 中可以看出,经过特征选择后,模型表现出更高的预测精度,MAE 和 RMSE 结果比没有使用降维数据时分别降低了 20% 左右。此外,从训练集

和测试集的预测结果来看，使用降维数据后，SSIM 训练集和测试集差异存在小幅降低，其中 SSIM 差异在 0.4 左右，而 NGC + SSIM 方法为 0.3 左右，说明剔除关联度较低数据后，模型泛化能力得到进一步提升。

表 3　消融试验结果（时间间隔：5 h）

Tab.3　Results of ablation test (time interval: 5 h)

方法	数据集	平均绝对误差/（mg·L^{-1}）	均方根误差/（mg·L^{-1}）	平均绝对误差困惑度	均方根误差困惑度
SSIM	训练集	0.88	0.92	2.41	2.50
	测试集	0.92	0.96	2.50	2.60
NGC + SSIM	训练集	0.68	0.71	1.98	2.02
	测试集	0.71	0.74	2.07	2.12
NGC + 改进 SSIM	训练集	0.47	0.49	1.59	1.63
	测试集	0.51	0.53	1.69	1.75
NGC + 改进 SSIM + Dropout	训练集	0.36	0.38	1.43	1.46
	测试集	0.43	0.45	1.54	1.58

3.5.2　增加 BILSTM 层的重要性

针对多维复杂数据耦合利用问题，由于不同维度数据关联性和可解释性存在差异，单一网络结构可能不能有效处理复杂的依赖关系，特别在存在长期依赖时，单层网络捕捉和融合复杂上下文信息能力不足，降低了模型预测准确性和泛化能力。针对开放养殖环境，水质变化不但受到外部环境数据影响，还与人为生产作业存在紧密联系，因此，为了有效修复养殖溶解氧含量数据，本文构建了集水质、气象等两个维度的数据向量，增加了数据复杂度。此外，面向开放环境的实际养殖生产过程，经常存在停电、设备损坏、断网等情况，且相关故障修复时间往往较长，导致存在长距离数据丢失。从表 3 中可以看出，当数据丢失时长为 5 h 时，本文提出的增加 BiLSTM 层的 SSIM 改进方法预测结果明显优于原始 SSIM 方法，MAE 和 RMSE 分别达到 0.51 和 0.53，比 SSIM 模型分别降低了 0.2 和 0.21，说明改进 SSIM 方法的长距离依赖捕捉能力得到显著提升。此外，困惑度指数也有明显提升，进一步反映了模型预测稳健性。

3.5.3　Dropout 的作用

SSIM 方法在编码器和解码器输出时，增加了 Dropout 正则化处理以降低模型对单一特征的依赖，降低过拟合现象，提升模型的鲁棒性。为了提升模型对复杂数据的处理能力，本文在原有 BiLSTM 层的基础上，增加了一个 BiLSTM 层，其中第 1 层 BiLSTM 提取低级特征，第 2 层 BiLSTM 基于这些特征进一步建模，捕捉更复杂的模式和依赖关系。从图 2 中可以发现，水温、pH 值、空气相对湿度等数据与溶解氧含量关联度较高，经过第 1 层 BiLSTM 时生成的低维特征可能对关联度较高的数据产生较多依赖，可能引起模型局部收敛，导致过拟合现象，因此在两层 BiLSTM 中间增加了一个 Dropout 正则化处理，从表 4 中可以看出，增加 Dropout 后，改进 SSIM 模型训练和预测精度得到一定改善，其中测试集 MAE 和 RMSE 降低 0.8 左右，分别达到 0.43 和 0.45。

表 4　不同方法比较结果

Tab.4　Comparison results of different methods

方法	时间跨度/h	平均绝对误差/（mg·L^{-1}）	均方根误差/（mg·L^{-1}）
RF	1	0.20	0.30
	3	0.17	0.28
	5	0.19	0.29
SVM	1	0.21	0.39
	3	0.25	0.44
	5	0.27	0.48
LSTM	1	0.34	0.38
	3	0.74	0.81
	5	0.96	1.06
BiLSTM	1	0.46	0.48
	3	0.67	0.75
	5	0.97	1.05
SSIM	1	0.10	0.11
	3	0.32	0.34
	5	0.92	0.96
NGC + 改进 SSIM + Dropout	1	0.04	0.05
	3	0.16	0.17
	5	0.43	0.45

3.5.4　与其他方法比较

为了验证本文方法的有效性，利用降维后数据，并选择目前具有代表性的数据插补方法进行比较。这些比较方法需要不同的数据输入格式，其中 RF（Randaom Forset）、SVM[28-29] 使用所有剩余数据集来弥补缺失数据，LSTM[30] 采用单向输入数据，BiLSTM[31] 采用双向输入数据，后两者都采用文献[22]中的方法构造训练集和测试集。试验结果如表 4 所示，本文提出的基于滞后相关性降维和改进 SSIM 的方法优于其他方法，且所有方法的性能都随着缺失率的增加而降低。其中 RF、SVM 等机器学习算法预测精度在数据丢失时间较少时与其它方法接近，而随着数据丢失时间变大，机器学习方法预测误差变化幅度明显高于 LSTM、BiLSTM。BilSTM 方法预测结果优于 LSTM 方法，这主要是由于双向网络结构能够更有效地利用过去和将来两种上下文信息，有效提升了模型对数据的理解和建模

能力。本文方法比传统 SSIM 方法在长距离数据修复上表现更好,而在短距离数据修复上表现差异不大,证明在面向水产养殖此类存在较大时长数据丢失的场景,本文方法具有更好的数据修复表现。

4 结束语

本文提出一种基于滞后因果分析和改进 SSIM 的养殖水体溶解氧缺失数据修复方法。面向户外开放养殖环境,水体中溶解氧含量不仅受到外界气象环境影响,还与人为操作关系紧密,使其时序数据具有非稳态、大时滞等特点,增加了数据修复的难度。为了解决这个问题,本文通过分析不同环境变量神经网络格兰杰因果关系,找到与溶解氧含量紧密相关的水质和气象参数,然后结合互相关性和交互效应分析,优化了关键参数变量。本文通过增加编码器层数,提升模型对复杂时间序列数据的建模能力,在两个 BiLSTM 层之间增加 Dropout 正则化降低模型过拟合风险。相关试验结果验证了本文方法的有效性。与其他现有典型方法相比,基于滞后因果分析降维和改进的 SSIM 方法通过提高模型对复杂特征表征能力获得具有竞争力的性能,为保障精准化和智慧化养殖系统运行提供数据完备性支撑。

参 考 文 献

[1] ROY S M, JAYRAIJ P, MACHAVARAM R, et al. Diversified aeration facilities for effective aquaculture systems a comprehensive review[J]. Aquaculture International, 2021(4): 1181 – 1217.

[2] RAFAEL A, CORDOVA L, NUEZ J, et al. Implementation and evaluation of open-source hardware to monitor water quality in precision aquaculture[J]. Sensors, 2020, 20(21):6112.

[3] YUAN H, XU G, YAO Z, et al. Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks[J]. ACM, 2018(18): 1293 – 1300.

[4] MORRIS T, WHITE I, ROYSTON P. Tuning multiple imputation by predictive mean matching and local residual draws[J]. Bmc Medical Research Methodology, 2014, 14(1):75.

[5] ZUBIR A, HUDHA K, KADIR Z. Impact behaviour modelling of magnetorheological elastomer using a non-parametric polynomial model optimized with gravitational search algorithm[J]. Journal of Mechanical Engineering, 2023, 69(11/12): 471 – 482.

[6] JAHANGIRI M, KAZEMNEJAD A, GOLDFELD K S, et al. A wide range of missing imputation approaches in longitudinal data: a simulation study and real data analysis[J]. BMC Medical Research Methodology, 2023, 23(1):161.

[7] FEKADE B, MAKSYMYUK T, KYRYK M, et al. Probabilistic recovery of incomplete sensed data in IoT[J]. IEEE Internet of Things Journal, 2017, 5(4): 2282 – 2292.

[8] LEKE C, TWALA B, MARWALA T. Missing data prediction and classification: the use of auto-associative neural networks and optimization algorithms[J]. arxiv Preprint, arxiv: 1403.5488, 2014.

[9] FIERO M H, HUANG S, OREN E, et al. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review[J]. Trials, 2016, 17: 1 – 10.

[10] SHAN S, WANG Y, XIE X Y, et al. Analysis of regional climate variables by using neural Granger causality[J]. Neural Computing and Applications, 2023, 35(22): 16381 – 16402.

[11] SHOJAIE A, FOX E B. Granger causality: a review and recent advances[J]. Annual Review of Statistics and Its Application, 2022, 9(1): 289 – 319.

[12] MAZIARZ M. A review of the Granger-causality fallacy[J]. The Journal of Philosophical Economics, 2015, 8(2): 86 – 105.

[13] STOKES P A, PURDON P L. A study of problems encountered in Granger causality analysis from a neuroscience perspective[J]. Proceedings of the National Academy of Sciences, 2017, 114(34): E7063 – E7072.

[14] TANK A, COVERT I, FOTI N, et al. Neural Granger causality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4267 – 4279.

[15] VATHSALA M, HOLI G. RNN based machine translation and transliteration for Twitter data[J]. International Journal of Speech Technology, 2020, 23(3): 499 – 504.

[16] KIRAN S, PATIL U, SHANKAR P, et al. Subtitle generation and video scene indexing using recurrent neural networks[C]// 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2021: 847 – 854.

[17] SUNDER V, THOMAS S, KUO H, et al. Fine-grained textual knowledge transfer to improve RNN transducers for speech recognition and understanding[C]// ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1 – 5.

[18] CAHUANTZI R, CHEN X, GUTTEL S. A comparison of LSTM and GRU networks for learning symbolic sequences[C]// Science and Information Conference. Cham: Springer Nature Switzerland, 2023: 771 – 785.

[19] ZAHEER S, ANJUM N, HUSSAIN S, et al. A multi parameter forecasting for stock time series data using LSTM and deep learning model[J]. Mathematics, 2023, 11(3): 590.

[20] YUAN H, XU G, YAO Z, et al. Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks[C]// Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018:1293 – 1300.

[21] VERMA H, KUMAR S. An accurate missing data prediction method using LSTM based deep learning for health care[C]// Proceedings of the 20th International Conference on Distributed Computing and Networking, 2019: 371 – 376.
[22] ZHANG Y, THORBURN P, XIANG W, et al. SSIM—a deep learning approach for recovering missing time series sensor data [J]. IEEE Internet of Things Journal, 2019, 6(4): 6618 – 6628.
[23] AHMED S, NIELSEN I E, TRIPATHI A, et al. Transformers in time-series analysis: a tutorial[J]. Circuits, Systems, and Signal Processing, 2023, 42(12): 7433 – 7466.
[24] 陈红, 夏青, 左婷, 等. 基于机器视觉的花菇分选技术[J]. 农业机械学报, 2014, 45(1): 281 – 287.
CHEN Hong, XIA Qing, ZUO Ting, et al. Determination of shiitake mushroom grading based machine vision [J]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(1): 281 – 287. (in Chinese)
[25] ARCHITH S, YUKTA C, ARCHANA H, et al. Analysis of M – SEIR and LSTM models for the prediction of COVID – 19 using RMSLE[M]// SINGH P. Fundamentals and methods of machine and deep learning: algorithms, tools and applications. Hoboken: John Wiley & Sons, 2022: 101 – 119.
[26] WANNINKHOF R, LEDWELL J R, BROECKER W S. GAS exchange-wind speed relation measured with sulfur hexafluoride on a lake[J]. Science, 1985, 227(4691): 1224 – 1226.
[27] GALLAGHER B K, GEARGEOURA S, FRASER D J. Effects of climate on salmonid productivity: a global meta-analysis across freshwater ecosystems[J]. Global Change Biology, 2022, 28(24): 7250 – 7269.
[28] 鲁植雄, 王雨彤, 王琳,等. 基于 IGWPSO – SVM 的 HMCVT 湿式离合器摩擦副温度预测[J]. 农业机械学报, 2023, 54(10):407 – 415.
LU Zhixiong, WANG Yutong, WANG Lin, et al. Temperature prediction of friction pair of HMCVT wet clutch based on IGWPSO – SVM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10):407 – 415. (in Chinese)
[29] 罗琦, 茹晓雅, 姜元,等. 基于机器学习与气象灾害指标的苹果相对气象产量预测[J]. 农业机械学报, 2023, 54(9): 352 – 364.
LUO Qi, RU Xiaoya, JIANG Yuan, et al. Relative meteorological yield prediction of apples based on machine learning and meteorological hazard indicators[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(9):352 – 364. (in Chinese)
[30] BHANDARI H N, RIMAL B, POKHREL N R, et al. LSTM – SDM: an integrated framework of LSTM implementation for sequential data modeling[J]. Software Impacts, 2022, 14: 100396.
[31] MOHAPATRA A, THOTA N, PRAKASAM P. Fake news detection and classification using hybrid BiLSTM and self-attention model[J]. Multimed Tools Appl., 2022, 81: 18503 – 18519.

(上接第 164 页)

[19] 魏子茹,卢延辉,王鹏宇,等. 基于 CRITIC 法的灰色关联理论在无人驾驶车辆测试评价中的应用[J]. 机械工程学报, 2021,57(12):99 – 108.
WEI Ziru, LU Yanhui, WANG Pengyu, et al. Application of grey relational theory based on CRITIC method in autonomous vehicles test and evaluation[J]. Journal of Mechanical Engineering, 2021,57(12): 99 – 108. (in Chinese)
[20] 靳亚红,吴鑫淼,甄文超,等. 基于采样点光谱信息窗口尺度优化的土壤含水率无人机多光谱遥感反演[J]. 农业机械学报,2024,55(1):316 – 327.
JIN Yahong, WU Xinmiao, ZHEN Wenchao, et al. UAV multispectral remote sensing inversion of soil moisture content based on window size optimization of spectral information at sampling points[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024,55(1): 316 – 327. (in Chinese)
[21] 刘爽,崔修杰,吕超. 考虑多影响因素的渔船单船捕捞能力评估方法[J]. 农业工程学报,2024,40(5):90 – 98.
LIU Shuang, CUI Xiujie, LÜ Chao. Method for assessing single-vessel fishing capacity considering multiple influencing factors [J]. Transactions of the CSAE, 2024,40(5): 90 – 98. (in Chinese)
[22] 徐立鸿,黄志尊,龙伟,等. 基于 MobileViT-CBAM-BiLSTM 的开放式养殖环境鱼群摄食强度分类模型[J]. 农业机械学报,2024,55(11):147 – 153.
XU Lihong, HUANG Zhizun, LONG Wei, et al. Classification model of fish feeding intensity based on MobileViT-CBAM-BiLSTM[J]. Transactions of the Chinese Society for Agricultural Machinery ,2024,55(11):147 – 153. (in Chinese)
[23] FLUNKERT V, SALINAS D, GASTHAUS J. DeepAR: probabilistic forecasting with autoregressive recurrent networks[J]. International Journal of Forecasting, 2020, 36(3):114 – 128.
[24] 李浩,王卓健,李哲,等. 基于堆栈自编码器和 DeepAR 的航空发动机剩余寿命预测[J]. 推进技术,2022,43(11):72 – 80.
LI Hao, WANG Zhuojian, LI Zhe, et al. Prediction of remaining useful life of aero-engine based on stacked autoencoder and DeepAR[J]. Propulsion Technology, 2022,43(11): 72 – 80. (in Chinese)
[25] 何旭辉,段泉成,严磊. 基于 DeepAR 的短期风速概率预测[J]. 铁道学报,2023,45(7):152 – 160.
HE Xuhui, DUAN Quancheng, YAN Lei. Short term wind speed probability prediction model using DeepAR[J]. Journal of the China Railway Society, 2023,45(7): 152 – 160. (in Chinese)