谢雨茜,李路,朱明,等.基于EMD与K-means的ILSTM模型在池塘溶解氧预测中的应用[J].华中农业大学学报,2022,41(3):200-210.
DOI:10.13300/j.cnki.hnlkxb.2022.03.023

基于EMD与K-means的ILSTM模型在池塘溶解氧预测中的应用

谢雨茜[1],李路[1,2],朱明[1,2],谭鹤群[1,3],李家庆[1],宋均琦[1]

1.华中农业大学工学院,武汉 430070;
2.长江经济带大宗水生生物产业绿色发展教育部工程研究中心,武汉 430070;
3.农业农村部水产养殖设施工程重点实验室,武汉 430070

摘要 为提高池塘溶氧量预测精度并改善预测结果滞后的情况,本研究提出基于经验模态分解(empirical mode decomposition,EMD)与K-means聚类的改进长短期记忆神经网络(improved long short-time memory,ILSTM)模型。利用皮尔森相关性分析与主成分分析结合的方法对原始数据进行特征提取,对溶氧量进行EMD分解,将选出的环境参数与溶氧量各分量一起生成样本集,并对其进行K-means聚类。针对同类中不同分解分量建立相应ILSTM预测模型,并用网格搜索、五折交叉验证与早停法进行超参数选取。对未来1 h池塘溶氧量进行预测,并与LSTM、ILSTM、LSTM-SVR、EMD-LSTM、EMD-ILSTM模型进行对比试验。结果显示,ILSTM与LSTM模型相比,RMSE、MAE与MAPE分别下降了50.46%、63.20%与68.96%,证明ILSTM模型能缓解传统LSTM模型预测的滞后情况。EMD-ILSTM模型与ILSTM模型相比,RMSE、MAE与MAPE分别下降了53.22%、46.74%与38.19%,证明EMD算法能提高预测精度。EMD-KILSTM模型的RMSE、MAE、MAPE分别为0.109 9 mg/L、0.074 9 mg/L、9.327 8%,与EMD-ILSTM模型相比,分别下降了4.35%、7.42%与8.09%,证明K-means聚类能提高预测精度,并且EMD-KILSTM模型是对比模型中预测效果最好的模型。以上结果表明,EMD-KILSTM模型能从时间尺度与历史环境类别两个方面深度分析溶氧量的特征,拥有更高的预测精度与更好的泛化能力。

关键词 池塘养殖;溶解氧;长短期记忆神经网络;经验模态分解;K-means聚类;预测模型
中图分类号 S931.3;TP391 **文献标识码** A **文章编号** 1000-2421(2022)03-0200-11

溶解氧(dissolved oxygen,DO)含量(简称溶氧量)是衡量水质的最重要指标之一,不仅反映了水中生物产氧过程和耗氧过程之间的动态平衡,还直接影响养殖对象的产量和品质。目前水产养殖中大多是根据当前溶氧量决定增氧设备的启停[1],但水体环境系统具有较大惯性,如果仅根据当前数据进行调节,不仅难以及时改善恶化的水质,还会加重水质指标的震荡,不利于水产养殖对象的健康。因此,及时准确地进行池塘溶氧量预测,对提高水质调控精度、增加水产养殖效益具有重要意义。

近年来国内外很多学者对水体溶氧量预测方法进行了研究。其中,神经网络预测方法是运用最广泛的溶氧量预测方法,其包括反向传播神经网络、极限学习机、循环神经网络等。反向传播神经网络容易得到局部最优解,因此一般与遗传算法[2]或者粒子群优化相结合[3]使用。极限学习机结构简单,易获得全局最优解,且学习速度快、泛化性能好,若将其与K-means聚类结合,则能提高预测精度[4]。循环神经网络算法适合处理时间序列,它强调研究对象时间上的相关性。常用的循环神经网络是长短期记忆神经网络(long short-time memory,LSTM)与它的变体门控神经网络模型,特别适合预测溶氧量这种受多因素影响且时间依赖性强的数据,但若输入因素间关系复杂或预测时长过长,易导致预测结果滞后、误差增大的问题[5]。

本研究针对上述问题,提出一种基于经验模态

收稿日期:2022-01-28
基金项目:中央高校基本科研业务费专项(2662020SCPY003,107-11041910103);国家自然科学基金项目(31972797)
谢雨茜,E-mail:656974457@qq.com
通信作者:李路,E-mail:taiyangfeng@126.com

分解（empirical mode decomposition，EMD）与 K-means 的改进长短期记忆神经网络（improved long short-time memory neural network model based on empirical modal decomposition with K-means clustering，EMD-KILSTM）对池塘溶氧量进行预测。首先利用皮尔森相关性分析与主成分分析结合的方法对原始数据进行特征选择，然后利用EMD算法对溶氧量时间序列进行分解。之后，将选出的环境参数与溶氧量各分量一起生成样本集，并对其进行K-means聚类，最后对同类中不同分解分量建立相应ILSTM预测模型，并用网格搜索、五折交叉验证与早停法进行超参数选取。以期减少LSTM模型预测延迟现象、提高预测精度。

1　材料与方法

1.1　仪器设备

为了精准预测溶氧量，必须明确与其相关的环境参数，因此需要尽量全面地收集该池塘的水质与气象信息，使用相关性分析提取对溶氧量影响较大的参数。本研究根据相关文献[6]，选出10个影响溶氧量的环境参数，并使用基于物联网的远程监测系统采集池塘的水质数据与气象数据。

每种水质传感器都自带温度测量功能，相应的类型及详细参数：（1）荧光溶氧量传感器（NS-120ZGS）精度为±2.0%；（2）pH传感器（NPH-1000Z）精度为±1.7%；（3）氨氮传感器（NHNG-5000Z）精度为±4.0%。

气象传感器类型及详细参数如下：

（1）空气温湿度传感器（HMP155A-L），当温度为−80～20 ℃时，其精度为±（0.226−0.0028×温度）℃；当温度为20～60 ℃时，其精度为±（0.055+0.0057×温度）℃。（2）风速风向传感器（034B），当风速＜10.14 m/s时，精度为0.1 m/s；当风速＞10.14 m/s时，精度为±1.1%。（3）气压传感器（CS100），量程为600～1 100 hPa，精度为±1.5 hPa。（4）太阳辐射计（LI200X-L），量程为0～3 000 W/m^2，精度为±5%。（5）雨量计（TE525-L），精度为1%。

1.2　研究区域

以湖北省武汉市华中农业大学水产学院实验基地的8号圈养池塘[7]为试验场地。该池塘面积约为1 166.66 m^2，水深约2.8 m，在池塘内搭建了8个直径为4 m、高3.1 m的圈养桶，用以圈养鱼类。水质与气象传感器位置分布俯视图如图1。水质传感器在池塘正中心水深1 m处，气象传感器位于池塘西北角。水质传感器采样周期为0.5 min，采集4个参数分别为水温、氨氮、pH、溶氧量。气象传感器采样周期为5 min，采集7个参数分别为气温、大气压强、湿度、雨量、太阳辐射强度（solar radiation intensity，SRI）、风速、风向。数据采集时间为2021年6月26日−8月17日。

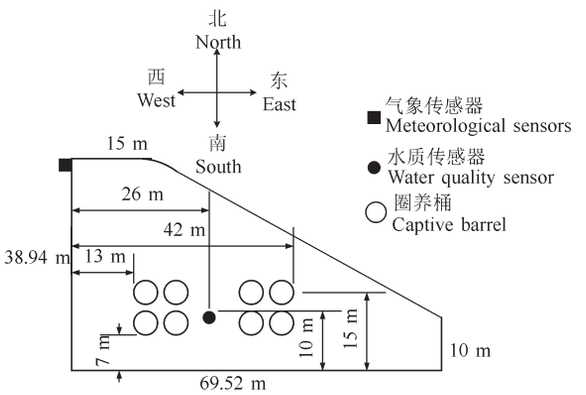

图1　水质与气象传感器的分布图

Fig.1　Distribution of water quality and weather sensors

1.3　试验数据预处理

水质数据需要进行填充、修正、滤波、合并、归一化。而气象数据已经进行过降噪处理，只需对其进行合并、归一化即可。

1）数据的填充与修正。由于水质传感器自身的测量原理的局限性，造成在天然水域中容易产生异常值。同时水质传感器需要定期擦拭与校准，其间产生空缺值。针对这些问题，对采集水质数据进行填充与修正。因为水质数据在时间上具有连续性且采样周期是0.5 min，在短时间内池塘水质数据发生剧烈变化的可能性小，所以采用线性插值法填补丢失的数据，采用均值法修正异常值[8]。

2）移动平均滤波。在复杂的池塘养殖环境中，因水流波动、藻类附着等原因，导致采集的水质数据存在一定噪声干扰，因此要对水质数据进行滤波降噪。由于水质数据中的噪声频率相对稳定，可用移动平均滤波器来实现数据降噪。

3）数据合并。因气象数据与水质数据采集周期不一样，需要将两者在时间轴上与气象数据合并成采用周期5 min的数据。

4）归一化。利用Z-score标准化方法对数据进行归一化处理，使模型输入参数介于[0，1]之间，从而提升预测模型收敛速度与精度。

1.4　皮尔森相关性分析与主成分分析结合方法

将本文“1.2中所述”11个参数全部输入预测模

型中，会增加模型训练时间、结构复杂程度与预测误差，所以需要在建模前进行特征提取。本研究选择皮尔森相关性分析与主成分分析[9]相结合的方法进行特征提取，具体步骤为：

①对经过预处理后的环境参数进行皮尔森相关性分析，将与溶氧量相关性最大的参数选为除溶氧量外第一个特征参数，相关性过小的参数淘汰。

②将其余m个参数组成一个特征空间，得到相关系数矩阵$R=\left[r_{ij}\right]_{m\times m}$，并计算其对应的特征值及特征向量。

③计算各个主成分贡献率τ_i如式(1)，贡献率τ_i表示第i个主成分表征特征空间的程度。

$$\tau_i=\frac{\lambda_i}{\sum_{k=1}^{m}\lambda_k},\ (i=1,2,...,m) \tag{1}$$

而累计贡献率η_i由多个主成分贡献率τ_i叠加而成，计算公式如式(2)：

$$\eta_i=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{m}\lambda_k},\ (i=1,2,...,m) \tag{2}$$

④原始参数线性组合成主成分的系数求法如式(3)：

$$\omega_i=\xi_i\sqrt{\lambda_i} \tag{3}$$

式(3)中，ω_i为第i个主成分的系数，λ_i为该主成分对应的特征值，ξ_i为该主成分对应的特征向量。

选取特征值大于1且累计贡献率大于70%的主成分来表征原始参数特征空间。在对应主成分的成分矩阵(特征向量表)中，筛选出最能解释样本空间数据的原始参数，从而完成特征提取。

1.5 模型建立方法

1)改进的长短期记忆神经网络。传统的LSTM神经网络的预测结果曲线与实测值曲线有一定滞后。原因在于，当采样间隔为step，时间窗长为$d=3$*step时，传统LSTM模型样本形式如式(4)：

$$\left[x^{(t-2\text{step})},x^{(t-\text{step})},x^{(t)}\right]\vdots\left[y^{(t+\text{step})}\right] \tag{4}$$

式(4)中，虚线左侧为输入样本，右侧为输出样本。$y^{(t+\text{step})}$表示$t+$step时刻溶氧量，$x^{(t)}$为t时刻环境参数(溶氧量及其相关参数)。而当输入样本导入模型时，由于t时刻溶氧量与$t+$step时刻预测目标高度相似，导致LSTM神经网络给t时刻溶氧量分配过高的权重，最后使模型主要学习到时间序列的一阶自相关性，造成预测曲线的滞后。

解决滞后现象，有2种思路：①将预测目标从未来时刻数值改成未来时刻数值和当前时刻数值的差分，直接预测一阶差分，防止模型学习到一阶相关性；②对目标时间序列进行分解，将其简化为若干简单波形再导入不同预测模型，分解形成的新波形因自身规律简单，更容易被预测模型学习。针对思路①，提出一种改进的LSTM模型。将预测目标变为溶氧量的一阶差分，并使用滑动窗口法生成更多样本。ILSTM神经网络整体结构如图2。

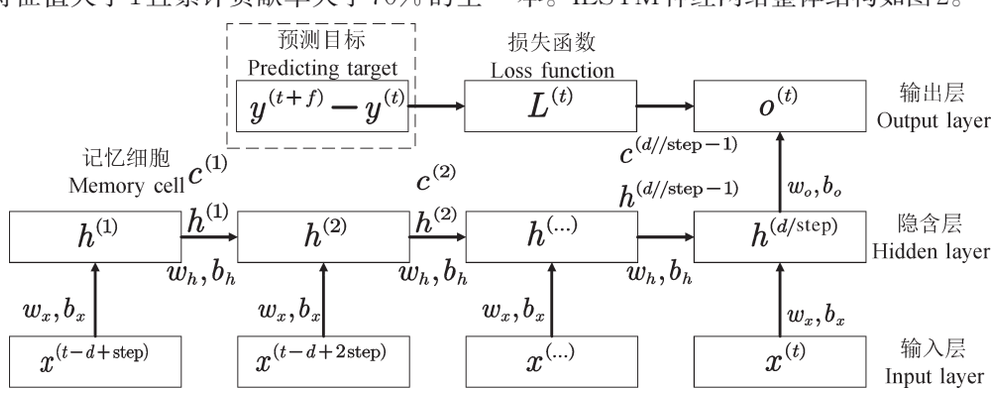

图2 ILSTM的整体结构图

Fig.2 The overall structure of ILSTM

图2中，w_x,b_x为输入层到隐含层的权重与偏置向量；w_h,b_h为隐含层内部单元的权重与偏置向量；w_0,b_0为隐含层到输出层的权重与偏置向量。图2中带入ILSTM模型的输入输出为式(5)：

$$\begin{cases}X_L^{(t)}=\left[x^{(t-d+\text{step})},x^{(t-d+2\text{step})},...,x^{(t)}\right]\\ Y_L^{(t)}=y^{(t+f)}-y^{(t)}\end{cases} \tag{5}$$

式(5)中，step为样本间隔时间步数，单个时间步长为5 min；d为滑动窗口大小；f为预测未来时间步数；$X_L^{(t)}$为一个样本的输入参数向量；$Y_L^{(t)}$为一个样本的预测目标；$x^{(t-d+\text{step})},x^{(t-d+2\text{step})},...,x^{(t)}$为$(t-d+\text{step}),...,t$时刻的环境参数；$y^{(t+f)},y^{(t)}$为$t+f,t$时刻溶氧量。

该模型将预测目标从$y^{(t+f)}$变成$y^{(t+f)}-y^{(t)}$，直接消除训练模型过程中t时刻环境参数对$t+f$时刻预测的溶氧量影响大的问题，缓解预测结果的滞后现象。ILSTM神经网络的隐含层内部单元结构与LSTM神经网络一致[10]。

2）EMD算法。该算法不像传统分解算法需要设定基函数，可直接根据数据在时间尺度上的特征进行分解[11]，因此它非常适合像溶氧量这样的非平稳时间序列。本研究用EMD将复杂的溶氧量时间序列分解为若干个单一频率的本征模函数（intrinsic mode function，IMF）与残余分量（residual，RES）如式（6）：

$$S_{DO}=\sum \mathrm{IMF}(t)+\mathrm{RES}(t) \tag{6}$$

每个IMF蕴含溶氧量时间序列在不同时间尺度的局部特征信息，并且具有如下特性：①IMF极值点数与过零点数最多相差1；②局部最大值与局部最小值形成的上下包络线的均值等于0。EMD分解流程，如图3所示。

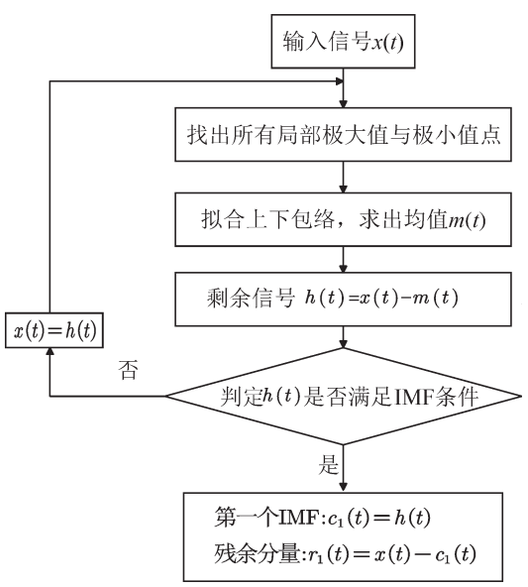

图3　EMD算法流程

Fig.3　EMD algorithm flow

3）K-means聚类算法。该算法属于无监督学习[12]，适合分类未知类别的数据，缺点是需人工确定聚类数K。由于溶氧量受其环境影响大，可以利用K-means聚类来对环境参数生成的样本集进行分类，将具有相似历史环境的样本分为一类。

4）超参数优化细节如下：①网格搜索。设定各个超参数调节范围，利用网格搜索算法对其排列组合。

②交叉验证。用5折交叉验证对选取的超参数组进行评价。本研究设定的评价指标为平均绝对误差（mean absolute error，MAE），其值越低，说明模型在训练集中表现得越优秀。

③早停法。为了缩短模型每次训练时间，本研究使用早停法提前退出迭代轮回。设定步数为5，即如果连续迭代5轮，验证集的损失函数都没下降即退出迭代。该方法可能会导致5次交叉验证的迭代轮数不同，所以选择5次中最大轮数代表该组超参数的轮数。

5）EMD-KILSTM预测模型。流程如图4所示，具体步骤如下：①溶氧量时间序列分解。其简化了溶氧量时间序列复杂度，得到n个IMF和1个RES。

②对分解分量与环境因素进行K-means聚类，导入聚类分析的样本形式如公式（7）。然后评估聚类算法的优劣，选出最优聚类情况。

$$X_{clu}^{(t)}=\begin{bmatrix} x_{1\sim m}^{(t-d+\mathrm{step})} & \mathrm{IMF}_{1\sim n}^{(t-d+\mathrm{step})} & \mathrm{RES}^{(t-d+\mathrm{step})} \\ x_{1\sim m}^{(t-d+2\mathrm{step})} & \mathrm{IMF}_{1\sim n}^{(t-d+2\mathrm{step})} & \mathrm{RES}^{(t-d+2\mathrm{step})} \\ \cdots & \cdots & \cdots \\ x_{1\sim m}^{(t)} & \mathrm{IMF}_{1\sim n}^{(t)} & \mathrm{RES}^{(t)} \end{bmatrix} \tag{7}$$

式（7）中，$X_{clu}^{(t)}$表示输入K-means算法的一个样本，其格式为$(m+n+1,\frac{d}{\mathrm{step}})$；$m$代表除溶氧量外特征参数数目；$x_{1\sim m}^{(t)}$表示$t$时刻各个环境参数值（除溶氧量外）；$\mathrm{IMF}_{1\sim n}^{(t)}$表示$t$时刻溶氧量各IMF数值；$\mathrm{RES}^{(t)}$表示$t$时刻溶氧量残余分量数值。

③在聚类得到的同类中对不同分解分量建立相应ILSTM预测模型，进行超参数优化，然后将各分量的差分预测结果$\widehat{Y_{\mathrm{IMFD}}^{(t+f)}}$与该分量当前时刻值$\mathrm{IMF}^{(t)}$相加得到一个该分量未来值预测结果$\widehat{Y_{\mathrm{IMF}}^{(t+f)}}$，最后将每个分量$\widehat{Y_{\mathrm{IMF}}^{(t+f)}}$相叠加成最终预测结果$\widehat{Y^{(t+f)}}$。

1.6　模型评价指标

采用戴维森堡丁指数（Davies-Bouldin index，DBI）来衡量聚类数K值的合理性，其定义如式（8）。DBI越小，说明类内距离越小、相似度越高，且类间距离越大、相似度越低[13]。

$$\mathrm{DBI}=\frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\frac{D_j+D_i}{d_{ij}} \tag{8}$$

式（8）中，D_i，D_j分别为第i，j类内平均距离；d_{ij}为第i类与第j类的质心距离。

采用均方根误差（root mean square error，RMSE）、MAE、平均绝对百分比误差（mean absolute percentage error，MAPE）[14]3个指标来衡量各预测

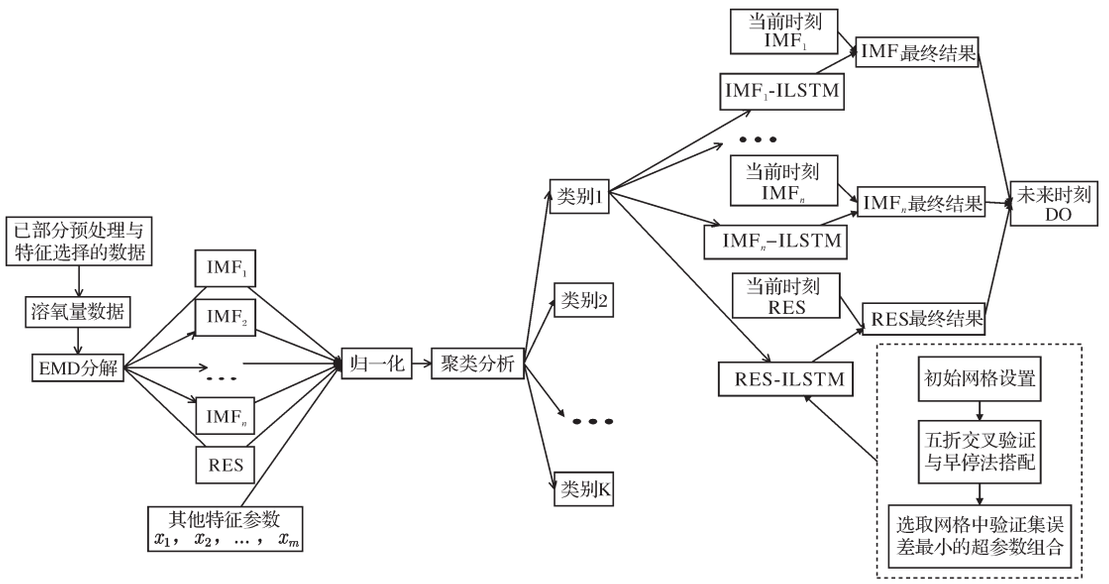

图4 EMD-KILSTM模型流程图

Fig.4 Flow chart of EMD-KILSTM model

模型的性能。

1.7 算法实现

本研究采用SPSS软件做主成分分析。用Python3.6语言编写模型主要程序，详情见表1。

表1 Python编程信息

Table 1 Python programming information

序号 Serial number	函数库/模块 Function library/module	功能 Function
1	openpyxl	打开Excel存储的数据和保存最终预测结果 Open the data stored in Excel and save the final prediction results
2	PyEMD	编写EMD算法 Writing EMD algorithm
3	sklearn	编写K-means与交叉验证 Write K-means and cross validation
4	numpy	预处理数据 Pretreatment data
5	Keras	编写ILSTM模型 Writing ILSTM model
6	itertools	编写网格搜索 Write grid search

2 结果与分析

2.1 试验数据采集与预处理结果

对2021年06月26日—2021年08月17日共53 d所采集的140 000条水质数据进行缺失值填补、异常值剔除与降噪的预处理，再与气象数据在时间维度上合并，最后得到15 264条有效数据。各水质参数预处理结果（不包括归一化）如图5所示。从图5A、图5C的方框可见，预处理方法使数据抖动显著变小，噪声与异常值被有效剔除。预处理后各参数数据描述性统计见表2。

2.2 相关性分析结果

皮尔森相关性分析结果显示，溶氧量与雨量、风速、风向、SRI、气压、水温、氨氮、湿度、气温、pH这10个参数的皮尔森相关性系数分别为0.046、0.128、0.134、0.241、−0.335、0.454、−0.538、−0.657、0.666、0.926。可见，pH与溶氧量相关性最高，确定其为特征参数。而风速、风向、雨量相关性均低于0.2，因此将它们淘汰。剩下的6个环境参数进行主成分分析，主成分贡献率与特征值见表3。从中可见有2个特征值大于1的主成分，并且贡献率高达74.095%。由主成分的成分矩阵（表4）可见，气温与湿度对第一个主成分影响较大，SRI对第二个主成分影响较大，最终选择的参数为DO、pH、气温、湿度、SRI。

2.3 EMD分解结果

利用EMD算法对溶氧量进行时间尺度上的分解，结果如图6所示。由图6可知，该算法将溶氧量分解为9个IMF和1个RES。整体来看，池塘溶氧量具有明显的时间多尺度特点。其中IMF具有一定周期性，能反映外部环境因素对溶解氧的周期性影响。IMF1～IMF4频率较高，体现了随机因素对溶解氧的影响。RES变化较平稳，反映了池塘溶解氧的总体变化趋势。

2.4 K-means聚类结果

选前51 d数据生成训练集，最后2 d数据生成测试集，用于对比不同类型模型性能，预测目标为未来

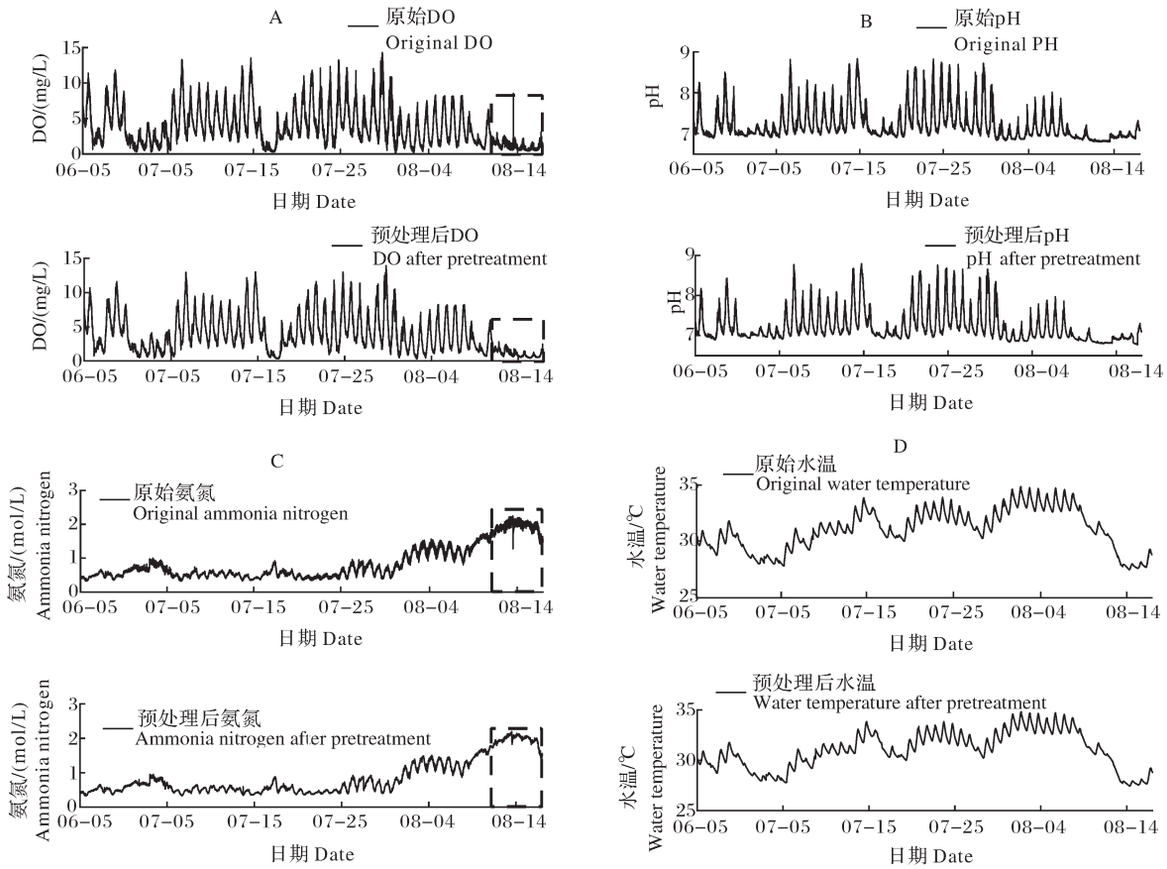

图5　溶氧量(A)、pH(B)、氨氮(C)、水温(D)预处理对比

Fig.5　Dissolved oxygen (A), pH (B), ammonia nitrogen (C), water temperature (D) pretreatment comparison

表2　参数描述性统计

Table 2　Parameter descriptive statistics

参数 Parameter	最小值 Minimum	最大值 Maximum	平均值 Mean	标准差 Standard deviation
气温/℃ Air temperature	21.730	38.370	29.579	3.276
湿度/% Humidity	42.550	98.300	76.410	14.213
SRI/(W/m^2)	0.000	705.500	106.926	165.271
风速/(m/s) Wind velocity	0.000	4.098	0.923	0.817
风向/(°) Wind direction	0.000	360.000	102.739	99.454
气压/hPa Atmospheric pressure	992.000	1 010.000	1 000.430	3.387
雨量/mm Rainfall	0.000	1.200	0.002	0.0182
pH	6.772	8.792	7.231	0.401
水温/℃ Water temperature	27.436	34.816	31.040	1.819
DO/(mg/L)	0.416	13.925	3.906	2.772
氨氮/(mg/L) Ammonia nitrogen	0.326	2.198	0.848	0.506

1 h溶氧量。样本间隔时间步数step＝12，单个时间步长为5 min；滑动窗口长度$d=24\times$step；预测未来时间步数$f=12$，因此15 264条有效数据可生成14 976个样本集，前14 400个样本用于模型训练，后576个样本用于测试模型性能。样本输入参数形式如式(7)，代入K-means聚类中，不同的聚类数K的样本分类与其性能评价如表5。可见聚类数K为2时，分类效果最好，每个样本点类别分布如图7所示，将图7与图5A对比，发现低溶氧量波形被分为了一

表3　主成分贡献率与特征值

Table 3　Principal component contribution rate and eigenvalue

成分 Component	初始特征值 Initial eigenvalue			提取载荷平方和 Sum of load squares		
	总计 Grand total	方差百分比/% Variance proportion	累积/% Accumulation	总计 Grand total	方差百分比/% Variance proportion	累积/% Accumulation/%
1	3.355	55.909	55.909	3.354	55.904	55.904
2	1.091	18.191	74.095	1.091	18.191	74.095
3	0.787	13.109	87.204			
4	0.504	8.398	95.603			
5	0.198	3.303	98.906			
6	0.066	1.094	100.000			

表4 成分矩阵

Table 4 Component matrix

参数 Parameter	成分 Component 1	2
气温 Air temperature	0.940	0.176
湿度 Humidity	−0.934	−0.188
SRI	0.580	0.682
气压 Barometric pressure	−0.605	0.592
水温 Water temperature	0.781	−0.203
氨氮 Ammonia nitrogen	−0.535	0.410

表5 不同的聚类数*K*的样本分类与其性能评价

Table 5 Sample classification with different number of clusters *K* and its performance evaluation

K	1	2	3	4	5	6	DBI
2	9 713	5 263					2.142 0
3	4 968	5 184	4 824				2.708 1
4	3 274	3 323	5 109	3 270			2.447 5
5	4 680	1 348	3 035	2 967	2 946		2.393 9
6	2 260	2 299	4 664	2 280	2 251	1 222	2.527 6

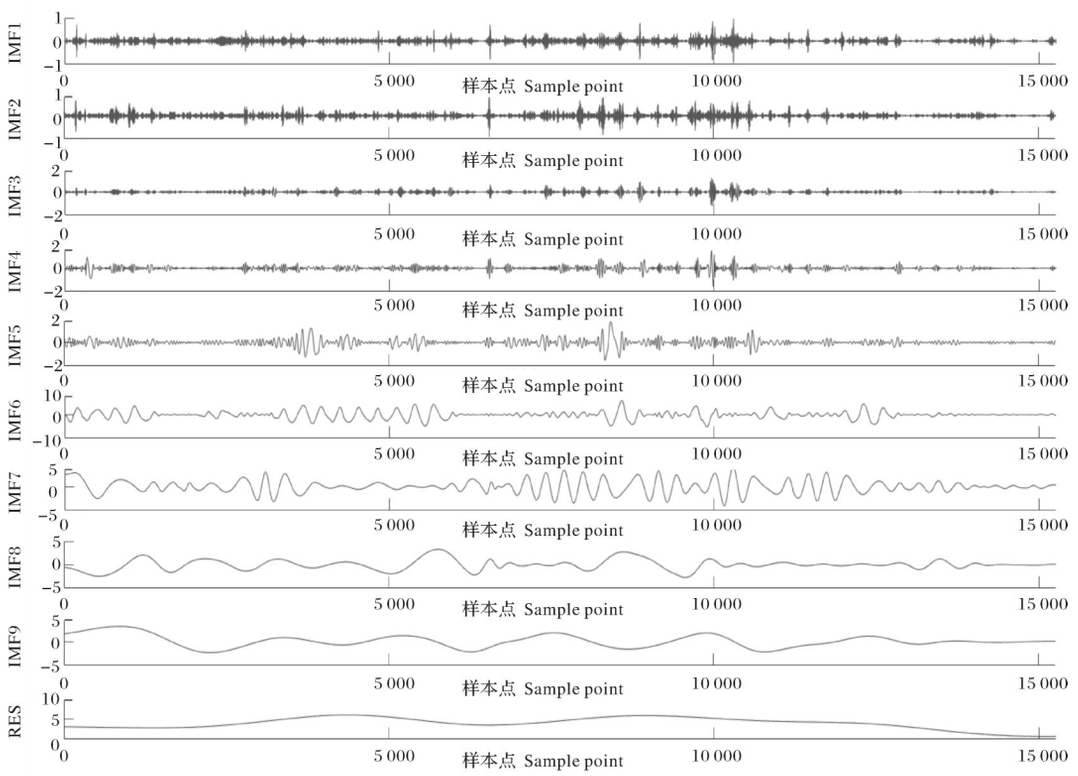

图6 EMD分解溶氧量的结果

Fig.6 Results of EMD decomposition of dissolved oxygen

类,高溶氧量波形为一类。

2.5 EMD-KILSTM预测结果及性能对比

将经过分解与分类的数据导入ILSTM模型进行预测,此处输入参数的仅有溶氧量的分解分量,输入样本形式为(24,1),由于EMD将溶氧量时间序列分解为10个分量,K-means聚类算法将每个分量都分为2类,因此需要得出20个不同情况的预测模型。

在超参数优化方面,网格搜索的信息见表6。有64种超参数组合用于模型训练,20个预测模型最优超参数与交叉验证评价见表7。从表7中可知,频率越高的分解分量交叉验证的误差越大,这是因为频率越高,对应的IMF中的随机噪声成分越多。

表6 网格搜索的信息

Table 6 Information about grid search

超参数 Hyper parameter	范围 Range	超参数 Hyper parameter	范围 Range
优化器 Optimizer	Adam	最大迭代轮数 Maximum iteration rounds	100
学习率 Learning rate	0.001,0.01	激活函数 Activation function	relu,tanh
神经元个数 The number of neurons	8,10,12,14,16,20,30,40	Dropout正则化 Dropout regularization	0.1
批量数 Batch number	16,32		

表 7　20 个预测模型最优超参与交叉验证评价

Table 7　Evaluation of cross-validation and optimal super-reference for 20 prediction models

分解分量 Decomposed component	类别 Category	学习率 Learning rate	神经元个数 The number of neurons	Dropout 正则化 Dropout regularization	激活函数 Activation function	迭代轮数 Number of iterations	批量数 Batch number	MAE
IMF1	1	0.01	8	0.1	relu	20	32	0.607 4
	2	0.001	16	0.1	relu	26	16	0.747 8
IMF2	1	0.001	20	0.1	relu	16	16	0.486 3
	2	0.001	16	0.1	relu	29	16	0.721 1
IMF3	1	0.01	40	0.1	tanh	70	32	0.394 2
	2	0.001	30	0.1	relu	33	32	0.687 6
IMF4	1	0.01	40	0.1	tanh	65	16	0.248 6
	2	0.01	40	0.1	tanh	49	32	0.435 3
IMF5	1	0.01	40	0.1	tanh	67	16	0.082 3
	2	0.01	40	0.1	tanh	50	32	0.153 7
IMF6	1	0.01	30	0.1	tanh	47	16	0.019 6
	2	0.01	40	0.1	tanh	32	32	0.039 5
IMF7	1	0.01	30	0.1	tanh	36	32	0.017 7
	2	0.01	40	0.1	relu	24	16	0.022 8
IMF8	1	0.001	40	0.1	relu	31	16	0.008 4
	2	0.001	40	0.1	relu	36	16	0.006 7
IMF9	1	0.001	20	0.1	relu	24	32	0.003 0
	2	0.001	30	0.1	relu	24	16	0.003 3
RES	1	0.001	16	0.1	relu	24	16	0.001 2
	2	0.001	10	0.1	relu	38	32	0.001 2

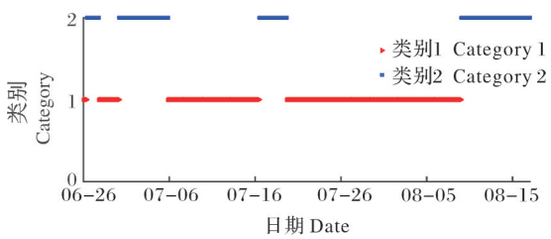

图 7　每个样本点类别分布

Fig.7　Distribution of categories per sample point

将本研究提出的EMD-KILSTM模型与LSTM、ILSTM、LSTM-SVR、EMD-LSTM、EMD-ILSTM等进行对比，各模型预测曲线如图 8 所示，模型性能对比见表 8，其中的时间复杂度与空间复杂度是对测试集使用时的复杂度；EMD-LSTM 模型、EMD-ILSTM 模型、EMD-KILSTM 模型均为单变量（仅分解分量）导入 LSTM 模型中预测。其他模型是多变量输入 LSTM 模型中预测。从图 8A 方框可知 LSTM 预测有一定滞后现象，从图 8B 与表 8 可知，ILSTM 与 LSTM 模型相比，RMSE、MAE 与 MAPE 分别下降了 50.46%、63.20% 与 68.96%，说明 ILSTM 模型能减少传统 LSTM 模型预测滞后现象。从 8C 可见 ILSTM-SVR 比 ILSTM 预测效果更好，但从其方框可看出，它在测试集的第 2 天，预测误差较大。图 8D

表 8　模型性能对比

Table 8　Comparisons of model performance

模型 Model	RMSE/(mg/L)	MAE/(mg/L)	MAPE/%	时间复杂度 Time complexity	空间复杂度 Space complexity
LSTM	0.495 8	0.412 8	52.906 6	O(1)	O(n)
ILSTM	0.245 6	0.151 9	16.421 3	O(1)	O(n)
LSTM-SVR	0.243 6	0.137 7	15.209 9	O(1)	O(n)
EMD-LSTM	0.338 5	0.324 9	48.976 6	O(1)	O(n)
EMD-STM5	0.440 7	0.388 0	51.308 0	O(1)	O(n)
EMD-ILSTM	0.114 9	0.080 9	10.149 2	O(1)	O(n)
EMD-ILSTM	0.109 9	0.074 9	9.327 8	O(n)	O(n)

中，EMD-LSTM 模型比 EMD-LSTM5 模型精度高，说明溶氧量序列经过 EMD 分解后的各分解分量在时间相关性上与其他参数不匹配，所以进行 EMD 分解后的各分量仅能单独导入 LSTM 模型。从表 8 可得 EMD-ILSTM 模型预测效果优于 ILSTM 模型，RMSE、MAE 与 MAPE 分别下降了 53.22%、46.74% 与 38.19%，说明 EMD 算法能提高预测精度。从表 8 和图 8 可知，EMD-KILSTM 模型是 7 个预测模型中精度最高的，说明 K-means 聚类能提高预测精度。EMD-ILSTM 模型预测未来 1 h 溶氧量的 RMSE、

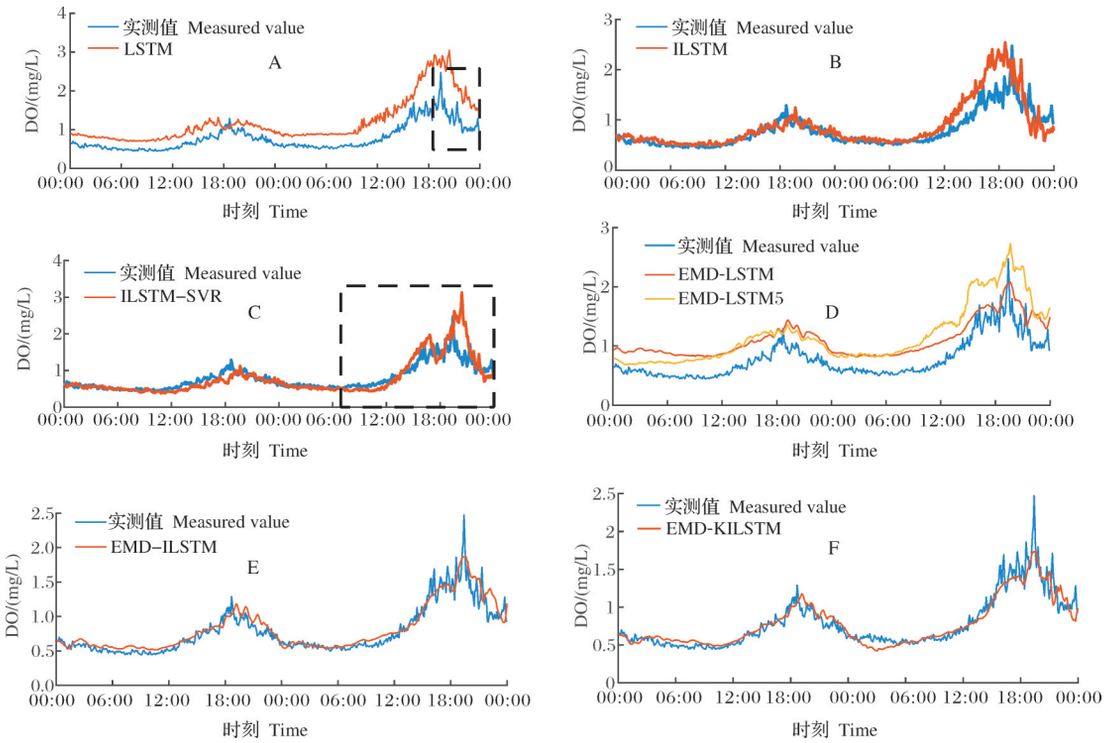

图8 LSTM模型(A)、ILSTM模型(B)、ILSTM-SVR模型(C)、EMD-LSTM模型与EMD-LSTM5模型(D)、EMD-ILSTM模型(E)、EMD-KILSTM模型(F)的预测结果

Fig.8 Prediction results of LSTM model (A), ILSTM model (B), ILSTM-SVR model (C), EMD-LSTM model and EMD-LSTM5 model (D), EMD-ILSTM model (E), EMD-KILSTM model (F)

MAE、MAPE分别为0.109 9 mg/L、0.074 9 mg/L、9.327 8%，其中MAPE较大的原因是测试集属于低溶氧量时段，分母数值太小，MAPE较敏感。EMD-KILSTM与其他6个模型的误差下降率见表9。

表9 EMD-KILSTM与其他模型误差下降率

Table 9 Error decline ratio of EMD-KILSTM to other models

%

模型 Model	RMSE	MAE	MAPE
LSTM	77.83	81.86	82.37
ILSTM	55.25	50.69	43.20
ILSTM-SVR	54.89	45.61	38.67
EMD-LSTM	67.53	76.95	80.95
EMD-LSTM5	75.06	80.70	81.82
EMD-ILSTM	4.35	7.42	8.09

3 讨 论

EMD-KILSTM模型是一种通过精细分类来预测溶氧量的方法。它能让养殖人员提前了解未来1 h池塘的溶氧量，更精确地调控增氧系统工作状态，对减小溶氧量波动，提升养殖对象环境舒适度并减少病害，提高养殖效益具有重要意义。

本研究提出的EMD-KILSTM池塘溶氧量预测模型与自回归移动平均模型[15]相比，能同时考虑环境因素与历史溶氧量对未来溶氧量的影响，而自回归移动平均模型仅根据溶氧量线性自相关关系进行预测；与灰色预测模型[16]相比，EMD-KILSTM池塘溶氧量预测模型能对溶氧量进行精准预测，而灰色预测模型只能估计溶氧量趋势；与支持向量机回归[17]相比，EMD-KILSTM池塘溶氧量预测模型考虑了溶氧量在时间轴上的自相关性和各个环境参数的互相关性，而支持向量机回归只能考虑其中一种相关性；与LSTM模型相比[18]，本模型不仅减轻了传统LSTM模型预测结果滞后的情况，还能将溶氧量依据时间尺度特征与历史环境情况自动分类，从而提高预测精度。后续拥有若干年数据时，也可以运用该方法，自动将类似环境的数据分为一类，从而做到在春夏秋冬、晴阴雨雪等各种天气模式下都能精准预测池塘溶氧量。

但EMD-KILSTM模型也存在一些缺点需要改

进：(1)虽然本研究提出的超参数优化方法涉及的超参数类型全面，但需要人为设定网格范围，寻找最优超参数组合速度慢，后续可能会与粒子群优化算法结合，提高模型训练速度；(2)仅用一种聚类算法进行分类，后续对多种聚类算法进行对比，择优确定最佳分类方案。

参考文献 References

[1] 黄建清，王卫星，姜晟，等．基于无线传感器网络的水产养殖水质监测系统开发与试验[J]．农业工程学报，2013，29(4)：183-190．HUANG J Q，WANG W X，JIANG S，et al．Development and test of aquacultural water quality monitoring system based on wireless sensor network[J]．Transactions of the CSAE，2013，29(4)：183-190(in Chinese with English abstract)．

[2] 崔雪梅．基于灰色遗传算法的LM-BP的河流溶解氧预测[J]．水文，2013，33(5)：46-51．CUI X M．Predicting dissolved oxygen in river based on grey LM-BP network of random genetic algorithm[J]．Journal of China hydrology，2013，33(5)：46-51(in Chinese with English abstract)．

[3] WU J，LI Z B，ZHU L，et al．Optimized BP neural network for dissolved oxygen prediction[J]．IFAC papersonline，2018，51(17)：596-601．

[4] 宦娟，刘星桥．基于K-means聚类和ELM神经网络的养殖水质溶解氧预测[J]．农业工程学报，2016，32(17)：174-181．HUAN J，LIU X Q．Dissolved oxygen prediction in water based on K-means clustering and ELM neural network for aquaculture[J]．Transactions of the CSAE，2016，32(17)：174-181(in Chinese with English abstract)．

[5] ZOU Q H，XIONG Q Y，LI Q D，et al．A water quality prediction method based on the multi-time scale bidirectional long short-term memory network[J]．Environmental science and pollution research，2020，27(9)：16853-16864．

[6] WU Y H，SUN L Q，SUN X B，et al．A hybrid XGBoost-ISSA-LSTM model for accurate short-term and long-term dissolved oxygen prediction in ponds[J]．Environmental science and pollution research，2022(3)：18142-18159．

[7] 何绪刚，侯杰．池塘圈养模式研究进展[J]．华中农业大学学报，2021，40(3)：21-29．HE X G，HOU J．Research progress on pond Juanyang mode[J]．Journal of Huazhong Agricultural University，2021，40(3)：21-29(in Chinese with English abstract)．

[8] KAHYA E．A new unidimensional search method for optimization：linear interpolation method[J]．Applied mathematics and computation，2005，171(2)：912-926．

[9] 李鑫飞．水环境溶解氧预测及控制方法研究[D]．南宁：广西大学，2018．LI X F．Study on prediction and control method of dissolved oxygen in the water environment dissolved[D]．Nanning：Guangxi University，2018(in Chinese with English abstract)．

[10] 李优柱，杨鸿宇，刘进思，等．我国中药材价格指数预测研究[J]．华中农业大学学报，2021，40(6)：50-59．LI Y Z，YANG H Y，LIU J S，et al．Predicting price index of Chinese herbal medicines in China[J]．Journal of Huazhong Agricultural University，2021，40(6)：50-59(in Chinese with English abstract)．

[11] 戴邵武，陈强强，刘志豪，等．基于EMD-LSTM的时间序列预测方法[J]．深圳大学学报(理工版)，2020，37(3)：265-270．DAI S W，CHEN Q Q，LIU Z H，et al．Time series prediction based on EMD-LSTM model[J]．Journal of Shenzhen University (science and engineering edition)，2020，37(3)：265-270(in Chinese with English abstract)．

[12] CAO X K，LIU Y R，WANG J P，et al．Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network[J/OL]．Aquacultural engineering，2020，91(11)：102122[2022-01-28]．https：//doi.org/10.1016/j.aquaeng.2020.102122．

[13] 朱秋圳，邬群勇，姚铖鑫，等．基于DBI和稀疏轨迹数据的交通状态精细划分与识别[J]．地球信息科学学报，2022，24(3)：458-468．ZHU Q Z，WU Q Y，et al．Fine classification and identification of traffic states based on DBI and sparse trajectory data[J]．Journal of geo-information science，2022，24(3)：458-468(in Chinese with English abstract)．

[14] 胡衍坤，王宁，刘枢，等．时间序列模型和LSTM模型在水质预测中的应用研究[J]．小型微型计算机系统，2021，42(8)：1569-1573．HU Y K，WANG N，LIU S，et al．Research on application of time series model and LSTM model in water quality prediction[J]．Journal of chinese computer systems，2021，42(8)：1569-1573(in Chinese with English abstract)．

[15] FARUK D O．A hybrid neural network and ARIMA model for water quality time series prediction[J]．Engineering applications of artificial intelligence，2010，23(4)：586-594．

[16] 谢乃明，刘思峰．离散GM(1，1)模型与灰色预测模型建模机理[J]．系统工程理论与实践，2005(1)：93-99．XIE N M，LIU S F．Discrete GM(1，1) and mechanism of grey forecasting model[J]．Systems engineering-theory & practice，2005(1)：93-99(in Chinese with English abstract)．

[17] 刘双印，徐龙琴，李道亮，等．基于时间相似数据的支持向量机水质溶解氧在线预测[J]．农业工程学报，2014，30(3)：155-162．LIU S Y，XU L Q，LI D L，et al．Online prediction for dissolved oxygen of water quality based on support vector machine with time series similar data[J]．Transactions of the CSAE，2014，30(3)：155-162(in Chinese with English abstract)．

[18] ZHOU J，WANG Y Y，XIAO F，et al．Water quality prediction method based on IGRA and LSTM[J/OL]．Water，2018，10(9)：1148[2022-01-28]．https：//10.3390/w10091148.org/．

Application of ILSTM model based on EMD and K-means in prediction of dissolved oxygen in pond

XIE Yuxi[1], LI Lu[1,2], ZHU Ming[1,2], TAN Hequn[1,3], LI Jiaqing[1], SONG Junqi[1]

1.*College of Engineering, Huazhong Agricultural University, Wuhan* 430070, *China*;

2.*Engineering Research Center of Green Development for Conventional Aquatic Biological Industry in the Yangtze River Economic Belt, Ministry of Education, Wuhan* 430070, *China*;

3.*Key Laboratory of Aquaculture Facilities Engineering, Ministry of Agriculture and Rural Affairs, Wuhan* 430070, *China*

Abstract In order to improve the prediction accuracy of dissolved oxygen in pond, and improve the lag of prediction results, this study proposed an improved long short-term memory (ILSTM) model based on empirical mode decomposition (EMD) and K-means clustering. A combination of Pearson correlation analysis and principal component analysis was used to extract features from the original data, EMD was used to decompose dissolved oxygen, and the selected environmental parameters were combined with each component of dissolved oxygen to generate a sample set to be clustered by K-means. The corresponding ILSTM prediction models were established for different decomposition components in the same kind, and the hyperparameters were selected by grid search, five-fold cross-validation and early stop method. The dissolved oxygen in the future 1 h pond was predicted and compared with models of LSTM, ILSTM, LSTM-SVR, EMD-LSTM, and EMD-ILSTM. The results showed that the RMSE, MAE and MAPE decreased by 50.46%, 63.20% and 68.96%, respectively, compared with the LSTM model, which proved that the ILSTM model could alleviate the prediction lag of the traditional LSTM model. Compared with ILSTM model, RMSE, Mae and MAPE of EMD-ILSTM model, decreased by 53.22%, 46.74% and 38.19% respectively, which proved that EMD Algorithm can improve the prediction accuracy. The RMSE, MAE and MAPE of the EMD-KILSTM model were 0.109 9 mg/L, 0.074 9 mg/L and 9.327 8%, respectively, and its RMSE, MAE and MAPE decreased by 4.35%, 7.42% and 8.09%, respectively, compared with the EMD-ILSTM model, which proved that K-means clustering could improve the prediction accuracy and the EMD-KILSTM model was the best one among the compared models. The above results show that the EMD-KILSTM model can deeply analyze the characteristics of dissolved oxygen from both time scale and historical environmental categories, and has higher prediction accuracy and better generalization ability, which provides scientific basis for intelligent water quality control.

Keywords pond farming; dissolved oxygen; long short-term memory neural networks; empirical mode decomposition; K-means clustering; predicting model

(责任编辑:边书京)