文章编号：2095-2163(2022)11-0134-04 中图分类号：TP391 文献标志码：A

基于奇异谱分析和长短期记忆神经网络的叶绿素 *a* 浓度短时预测研究

易 洋[1]，何先波[2]，王淳睿[1]
（1 西华师范大学 电子信息工程学院，四川 南充 637009；2 西华师范大学 计算机学院，四川 南充 637009）

摘 要：有害藻华（Harmful Algal Blooms，HABs）近年来在全球频繁发生，实时预报水体藻华的出现时间和区域，可为环保监督管理部门提供有效的参考依据。为了提高水华预测的准确性，本文提出了一种基于奇异谱分析（Singular spectral analysis，SSA）和长短期记忆神经网络 LSTM（Long Short-Term Memory，LSTM）的 SSA-LSTM 模型，将 BYK 站点的叶绿素 *a* 浓度时间序列分解重构为趋势特征和周期特征，并对其变化的趋势进行预测。分析对比了单个 LSTM、时序神经网络（Temporal Convolutional Network，TCN）、卷积神经网络（Convolutional Neural Network，CNN）的实验结果。验证了 SSA-LSTM 在叶绿素 *a* 短时预测上有更好的表现，模型的 *RMSE*、*MAE* 和 *MAPE* 分别为 0.67、0.38 和 0.09。

关键词：叶绿素 *a*；LSTM；短时预测；奇异谱分析

Study on short-term prediction of chlorophyll-*a* concentration based on singular spectrum analysis and LSTM Neural Network

YI Yang[1], HE Xianbo[2], WANG Chunrui[1]
(1 College of Electronics and Information Engineering, China West Normal University, Nanchong Sichuan 637009, China;
2 College of Computer, China West Normal University, Nanchong Sichuan 637009, China)

【Abstract】 Harmful Algal Blooms (HABs) occur frequently all over the world in recent years. Real-time prediction of the occurrence time and region of algal blooms in water bodies can provide effective reference for environmental protection supervision and management departments. In order to improve the accuracy of blooms prediction, a SSA-LSTM model based on singular spectrum analysis (SSA) and long short-term memory neural network (LSTM) is proposed in this paper. The time series of chlorophyll-*a* concentration at BYK site is decomposed and reconstructed into trend characteristics and periodic characteristics, and the changing trend is predicted. The experimental results of single LSTM, time series neural network (Temporal Convolutional Network, TCN) and convolution neural network (Convolutional Neural Network, CNN) are analyzed and compared. It is verified that SSA-LSTM had better performance in short-term prediction of chlorophyll-*a*, and the *RMSE*, *MAE* and *MAPE* of the model are 0.67, 0.38 and 0.09, respectively.

【Key words】 chlorophyll-*a*; LSTM; short-term forecasting; singular spectrum analysis

0 引 言

藻华是目前严重的水生态环境问题之一，造成水质问题和生态环境破坏。随着大数据时代的到来和人工智能的发展，数据驱动方法在藻华预测上的应用逐渐得到重视[1]。其中，神经网络算法已成功运用在多种水域的叶绿素预测上，如中国江苏省太湖区[2]、浙江省西湖[3]。LSTM 在传统 RNN 的基础上引入了改进，通过增加门控接构和记忆单元，使得网络上可以自由地选择已经丢失和保留下来的信息，从而解决了梯度下降和梯度消失的问题。目前，国内学者也尝试着把 LSTM 方法运用在对藻类变化趋势的预报上，如 Wang 等人[4]使用福建 2009～2011 年的海洋在线监测数据，构建了 LSTM 时空分布模型，用于预测叶绿素 *a* 未来的浓度变化趋势，并且在预测叶绿素 *a* 浓度的变化趋势上取得了较好的成效。Shin 等人[5]提出了基于 LSTM 和海表温度数据及光合有效辐射数据的水华预测模型。然而 LSTM 的模型效果受到输入变量可靠性的限制，藻类的在线监测数据具有离散性，在模拟藻类动态变化趋势时模型可能会受到一定的限制，并且在线监测数据由于自身的局限性，往往展现出非平稳性，因此引入奇异谱分析

基金项目：西华师范大学英才科研基金项目（17YC149）。

作者简介：易 洋（1999-），女，硕士研究生，主要研究方向：深度学习、时序预测；何先波（1971-），男，博士，教授，硕士生导师，主要研究方向：嵌入式系统研究；王淳睿（1997-），男，硕士研究生，主要研究方向：编译器开发。

通讯作者：何先波 Email：1946034057@qq.com

收稿日期：2022-05-18

对时间序列进行处理，可以使 LSTM 模型更容易捕捉到时间序列隐藏的变化的趋势。例如，Cui 等人[6]将 SSA 和 LGBM（Light Gradient Boosting Machine）算法相结合，构建了降雨时序数据的预测模型，实验结果表明，经过奇异谱分析降噪处理后的数据能够更有效地和神经网络相结合，从而提高预测性能。

1　相关理论以及方法

1.1　奇异谱分析

基本的奇异谱分析包括 2 个阶段：分解和重建。其中，分解阶段包括 2 个步骤：嵌入和奇异值分解（SVD）。重建阶段也包括 2 个步骤：分组和对角平均。考虑一个长度为 $N(N>2)$ 的实值时间序列 $F=\{x_1, x_2, \cdots, x_N\}$，这里拟给出 4 个步骤的阐释分述如下。

1.1.1　嵌入

嵌入过程将原始时间序列映射为多维滞后向量序列。设窗口长度 L 为整数，$1<L<N$，则信号向量的轨迹矩阵 $\boldsymbol{X}$ 为：

$$\boldsymbol{X}=\begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{pmatrix} \tag{1}$$

其中，$K=N-L+1$。

需要指出的是，输出轨迹矩阵是汉克尔矩阵，这意味着所有沿对角线的元素都是相等的。

1.1.2　SVD

在这一步中，对轨迹矩阵 $\boldsymbol{X}$ 进行奇异值分解，设 $\boldsymbol{S}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$，$\lambda_1, \lambda_2, \cdots, \lambda_L$ 是降序排列的 $\boldsymbol{S}$ 的特征值（$\lambda_1 \geqslant \cdots \geqslant \lambda_L \geqslant 0$），$\boldsymbol{U}_1, \cdots, \boldsymbol{U}_L$ 是对应于这些特征值的矩阵 $\boldsymbol{S}$ 的标准正交向量。设 $d=rank(\boldsymbol{X})=\max\{i, \lambda_i>0\}$（在实际序列中，通常 $d=L^*$，$L^*=\min(L, K)$。$V_i=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}_i/\sqrt{\lambda_i}(i=1, \cdots, d)$。则轨迹矩阵的奇异值分解为：

$$\boldsymbol{X}=\boldsymbol{X}_1+\cdots+\boldsymbol{X}_d \tag{2}$$

其中，$\boldsymbol{X}_i=\sqrt{\sigma_i}\,\boldsymbol{U}_i\,\boldsymbol{V}_i^{\mathrm{T}}$。

1.1.3　分组

在分组步骤中，可以选择分析周期图、右特征向量散点图或特征值函数图来区分噪声和信号。在信号重构的过程中，对于分组的方式没有具体的规则，下标 $\{1, \cdots, d\}$ 可以根据待重构时间序列的性质分为 m 个不相交的子集，即 $I_1, I_2, \cdots, I_m$。令 $I=\{i_1, \cdots, i_p\}$，则复合矩阵为 $\boldsymbol{X}=\boldsymbol{X}_{I1}+\boldsymbol{X}_{I2}+\cdots+\boldsymbol{X}_{Im}$。

1.1.4　平均对角化

SSA 的最后一步是将每个结果矩阵从分组转换为一个长度为 n 的新序列。设 $\boldsymbol{Y}$ 为 $L*K$ 矩阵，则 T_{ij} 为 T 的元素，T 可以通过以下公式转换为序列：

$$t_k=\begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} t^*_{m,k-m+1} & 1\leqslant k<L^* \\ \dfrac{1}{L^*}\displaystyle\sum_{m=1}^{L^*} t^*_{m,k-m+1} & L^*\leqslant k<K^* \\ \dfrac{1}{N-k+1}\displaystyle\sum_{m=1}^{N-K^*+1} t^*_{m,k-m+1} & K^*\leqslant k<N \end{cases} \tag{3}$$

根据式（3）可以求得长度为 N 的单一 $\boldsymbol{RC}_t$ 分量。新的 $\boldsymbol{X}$ 分量是 d 个 $\boldsymbol{RC}_t$ 分量的总和，可以表示为：

$$\boldsymbol{X}^*=\boldsymbol{RC}_1+\boldsymbol{RC}_2+\cdots+\boldsymbol{RC}_d \tag{4}$$

1.2　长短期记忆神经网络

长短期记忆模型（LSTM）是一类时间递归的神经网络，继承了大多数 RNN 模式的优点，并克服了由梯度反向传递过程所引起的梯度消失现象。LSTM 在 RNN 的基础上增加了一个记忆单元结构来判断信息是否有用。每个单元由一个输入门、一个遗忘门和一个输出门组成，如图 1 所示。这些信息都通过 LSTM 网络，并按照规则确定是否可用。只产生了合乎规则的信息，而不合乎规则的信息经由遗忘的方式而将会丢弃掉。研究可知，LSTM 对产生长期的相关性问题的方法尤其有用。对此可展开探讨论述如下。

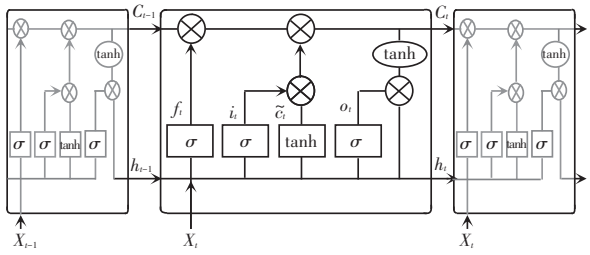

图 1　LSTM 模型

Fig. 1　LSTM model

（1）遗忘门。确定了前一时刻状态的保留情况，计算公式为：

$$F_t=\sigma(W_f\cdot[h_{t-1}, x_t]+b_f) \tag{5}$$

其中，σ 表示激活函数 *sigmoid*；W_f 表示遗忘门权重的权重；b_f 表示遗忘门的偏差；*sigmoid* 函数将输入和先前时刻的状态映射到从 0 到 1 的值；F_t 的值为 1 表示完全保留，0 表示完全丢弃。

（2）输入门。决定当前网络的输入 x_t 有多少被

更新到单元状态 c_t，此处需用到的数学公式可写为：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{7}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{8}$$

其中，W_i 和 b_i 是输入门的权值和偏差；W_c 和 b_c 表示构建候选向量时的权值和偏差，由 *sigmoid* 函数决定遗忘的比例。式(8)中的 c_t 实现了单元格状态的更新。

(3)输出门。需要用以下公式来确定输出值：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = o_t \times \tanh(c_t) \tag{10}$$

其中，W_o 和 b_o 为输出门的权值和偏差。将激活函数层进行 tanh 运算后的当前状态 c_t 乘以输出 o_t，得到当前时刻的输出 h_t。

2 相关工作

2.1 数据集

本文采用的数据主要来自某淡水湖 BYK 站点的在线监测叶绿素 *a* 浓度数据。样本包含了 2019~2020 年两年间共 6 113 条数据，采样频率为每隔 4 h 一次。

2.2 数据标准化

数据标准化可以使模型提取出更多的有效特征，本文采取式(11)对叶绿素 *a* 浓度时序数据进行极差标准化处理，将数据缩放到(0,1)之间：

$$B = \frac{R - R_{\min}}{R_{\max} - R_{\min}} \tag{11}$$

其中，R、B 分别为处理前、后的数据，$R_{\max}$、$R_{\min}$ 分别为样本中的最大值和最小值。

2.3 实验内容

本文提出的 SSA-LSTM 模型的流程如图 2 所示。由图 2 可见到，首先利用 SSA 将叶绿素 *a* 浓度时间序列分解和重构为不同的分量，并分离和去除噪声分量，留下剩余 *d* 个分量。然后，根据各个分量的贡献值进行排序。为了在突出叶绿素浓度 *a* 时间序列的趋势特征的同时，最大限度地保留时间序列信息，将 *d* 个分量分为 2 部分，再将其重构为趋势特征和周期特征。最后，LSTM 对具有不同特性的 2 个组件进行模拟，并对模拟结果进行集成，使模型实现精确预测。

在训练过程中使用贝叶斯参数优化算法，进行 50 次迭代搜索，寻找出最优参数。其中，模型主要参数包括学习率、神经元结点数、回溯时间步长、数据批处理、激活函数等。在网络结构设计中，考虑到模型的计算复杂度与计算效率，相关参数的设定范围为：回溯时间步长 5~30；神经结点数 32~128；数据批处理[64,128,256]；LSTM 激活函数 [*relu*,*sigmoid*,tanh,*elu*]；学习率 0.1 至0.000 01。叶绿素 *a* 浓度历史数据时序预测实验的具体参数设置见表 1。

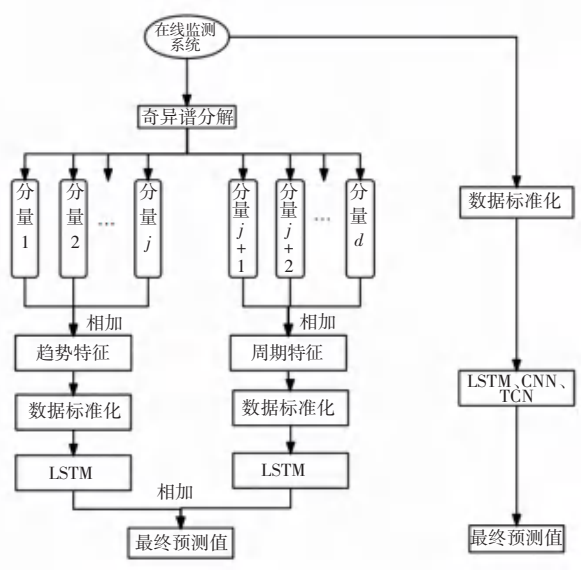

图 2 算法流程图

Fig. 2 Flow chart of the algorithm

表 1 SSA-LSTM 模型实验参数设置

Tab. 1 Configured experimental parameters of SSA-LSTM model

参数	参数描述	参数值
r	学习率	0.001
Batch_size	数据批处理	128
Activation	激活函数	*Relu*
Train:*test*	训练集:测试集	4:1
LSTM_filters	神经元个数	95
Look_back	回溯步数	13

2.4 评价指标

本文采用均方根误差（*RMSE*）、平均绝对误差（*MAE*）和绝对百分比误差（*MAPE*）对模型进行评估。其计算过程见如下公式：

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{n}} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|o_i - p_i| \tag{13}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\check{y}_i - y_i}{y_i}\right| \tag{14}$$

其中，o、p 分别表示观测数据和预测数据，n 表示观测样本的数据量。

3　实验结果

3.1　基于 SSA 的叶绿素 *a* 浓度数据分解

通过设置选取窗口长度为 15，将序列分解为 15 个不同的分量，选取前 12 个成分作为主要有用信息。在剩余的 12 个分量选取分量 1 到分量 7 作为趋势特征，剩下分量 8 到分量 12 作为周期特征，如图 3 所示。

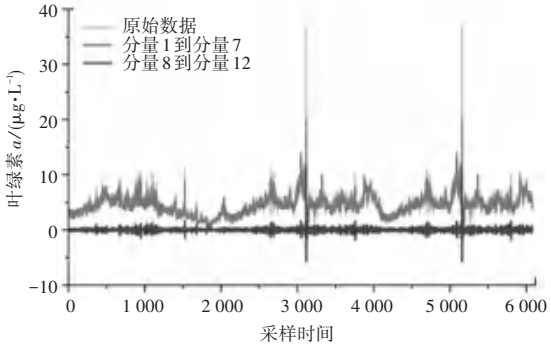

图 3　利用 SSA 重建的叶绿素 *a* 浓度序列的子序列

Fig. 3　Reconstructed sub-series of the chlorophyll-*a* concentration sequence by SSA

3.2　模型效果比较分析

为了验证 SSA-LSTM 模型的有效性和预测精度，本文实验将原始的 LSTM 模型、CNN 模型以及 TCN 模型与本文提出的 SSA-LSTM 模型进行比较。实验结果柱状图如图 4 所示，SSA-LSTM、LSTM、CNN 和 TCN 对叶绿素 *a* 浓度预测效果见表 2。

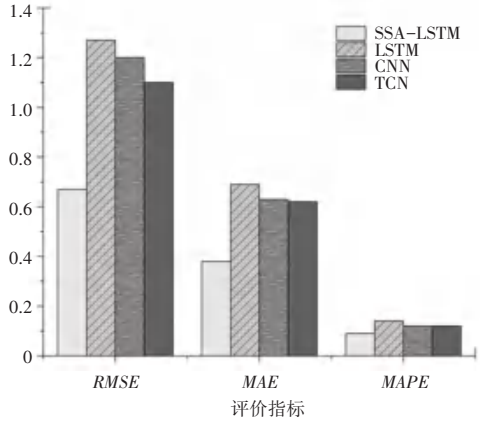

图 4　实验结果柱状图

Fig. 4　Histogram of experimental results

表 2　SSA-LSTM、LSTM、CNN 和 TCN 对叶绿素 *a* 浓度预测效果

Tab. 2　Prediction effect of SSA-LSTM, LSTM, CNN and TCN on chlorophyll-*a* concentration

模型	*RMSE*	*MAE*	*MAPE*
SSA-LSTM	0.67	0.38	0.09
LSTM	1.27	0.69	0.14
CNN	1.20	0.63	0.12
TCN	1.10	0.62	0.12

从图 4 和表 2 分析可以看出，SSA-LSTM 对叶绿素浓度的预测效果明显优于 LSTM、CNN 和 TCN，其中 *RMSE*、*MAE* 和 *MAPE* 分别为 0.67、0.38 和 0.09，这 3 种评价指标的值都优于其他 3 个模型。*RMSE* 和 *MAE* 的对比表明 SSA-LSTM 模型的预测误差较小，模型的精度高，*MAPE* 的对比表明模型更加稳定。综合以上分析可知，SSA-LSTM 的预测值更接近叶绿素浓度的真实值，体现了模型的有效性。说明相较于直接将数据输入神经网络模型中，使用 SSA 处理后的叶绿素浓度数据能够使数据驱动的模型更好地捕捉到变化趋势，使模型的预测性能得到提升。

4　结束语

本研究围绕在某湖泊 BYK 站点获取的在线监测数据，结合奇异谱分析与 LSTM 深度学习神经网络模型，探索了该模型在叶绿素浓度短期预测的应用。具体结论如下：

（1）SSA 能够有效地分离趋势项、波动项和噪声分量，克服了 LSTM 模型在处理非线性序列方面的不足，从而使建立在此基础上的 SSA-LSTM 模型具有更强的预测能力。

（2）本文提出的 SSA-LSTM 方法可以从叶绿素 *a* 浓度历史数据中训练预测模型。对未来 4 h 的叶绿素 *a* 浓度进行预测。预测后结果要明显优于纯数据驱动模型，例如 LSTM、CNN、TCN。总而言之，本文提出的 SSA-LSTM 模型能够有效地提取藻类高频监测数据的动态变化趋势，且能够对叶绿素 *a* 浓度实现短时精确预测，这为水华的治理策略的拟定提供了一定的参考和借鉴。

参考文献

[1] ROUSSO B Z, et al. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes [J]. Water Research, 2020, 182: 115959.

[2] 任黎,董增川,李少华. 人工神经网络模型在太湖富营养化评价中的应用[J]. 河海大学学报(自然科学版), 2004, 32(02):147-150.

[3] 裴洪平,罗妮娜,蒋勇. 利用 BP 神经网络方法预测西湖叶绿素 a 的浓度[J]. 生态学报, 2004, 24(02): 246-251.

[4] WANG Xiaofan, XU Lingyu. Unsteady multi-element time series analysis and prediction based on spatial-temporal attention and error forecast fusion[J]. Future Internet, 2020, 12(2):34.

[5] SHIN J, RYU J H, KIM S M, et al. Early prediction of margalefidinium polykrikoides Bloom using a LSTM Neural Network Model in the south Sea of Korea[J]. Journal of Coastal Research[J]. 2019. 90(sp1): 236.

[6] CUI Zhongjie, Qin Xiaoxia, CHAI Hongxiang, et al., Real-time rainfall-runoff prediction using light gradient boosting machine coupled with singular spectrum analysis[J]. Journal of Hydrology, 2021, 603(12).