DOI:10.11737/j.issn.1003-0239.2021.04.001

基于串联深度神经网络的Chl-*a*浓度短期预报方法研究

何恩业[1]，李尚鲁[2]，杨静[1]，季轩梁[1]，高姗[1]，王丹[1]

（1.国家海洋环境预报中心 自然资源部海洋灾害预报技术重点实验室，北京 100081；2.浙江省海洋监测预报中心，浙江 杭州 310007）

摘 要： 以浙江海洋保护区2019年5月生态浮标监测数据为基础，对叶绿素a（Chl-*a*）与各理化因子进行Pearson相关性分析，发现研究海域的Chl-*a*与溶解氧和pH呈显著正相关（*P*=0.01），与硝氮和磷酸盐呈显著负相关（*P*=0.05）。在此基础上，建立了一种串联深度神经网络（DNN）的Chl-*a*短期预报模型，该模型以5层神经网络为基本单元，采用前后串联方式构建了拥有6个隐层的DNN。实验结果显示：DNN模型能够较为准确地预测Chl-*a*浓度短期变化趋势，24 h和48 h预报结果的RMSE分别为1.25 μg/L和2.43 μg/L，MAE分别为1.03 μg/L和1.99 μg/L，相比于浅层网络预测精度更高。

关键词： DNN；神经网络；深度学习；串联神经网络；叶绿素a

中图分类号：X55 **文献标识码：**A **文章编号：**1003-0239（2021）04-0001-10

1 引言

叶绿素a（Chl-*a*）作为一个生物量指标可以表征水体初级生产力状况，研究Chl-*a*的变化趋势对赤潮早期预警和海水富营养化潜势研究具有重要意义[1]。众多学者针对Chl-*a*含量的预报和研究开展了大量工作。预报方法有很多种，大体可以归纳为4类：（1）单要素指标预测法。如郭文景等[2]利用有滞后变量参与的格兰杰因果关系检验和向量自回归模型，分析了太湖水质参数对浮游植物生物量的影响；阮华杰等[3]对赤潮发生前后生态浮标各监测要素的变化进行分析，认为可以通过监测要素指标进行藻华预测。（2）传统统计学预测法。如金衍健等[4]利用舟山近岸水质监测数据建立了Chl-*a*多元线性回归方程；林祥[5]利用主成分线性回归分析方法建立了诏安湾Chl-*a*统计方程。（3）数值模拟方法。如杨德周等[6]基于POM（Princeton Ocean Model）模型模拟长江口Chl-*a*分布状况；崔玉洁[7]利用CE-QUAL-W2模型对三峡库区藻类水华生消过程进行模拟。（4）人工智能预测法。如张娣等[8]建立了自回归滑动平均-反向传播（AutoRegressive Moving Average-Back Propagation，ARMA-BP）模型对太湖藻类Chl-*a*浓度进行预测；石绥祥等[9]根据海洋各要素与Chl-*a*浓度之间的长短期依赖程度构建了长短期记忆网络（Long Short Term Memory network，LSTM）预测模型，预测精度大幅提高。

由于Chl-*a*变化成因复杂，其与环境因子之间模糊和不确定性的高度非线性关系，造成传统统计方法预测效果较差，而数值模拟对具体站点Chl-*a*含量预测的精确度不高，难以应用于预报实践。随着计算机技术的飞速发展，神经网络模型的智能预报方法逐渐体现出独特的优势。神经网络是一种对人脑结构和功能进行模拟的数学模型，它是由大量且互相连接的处理单元组成的复杂系统，具有分布式存储和处理以及自组织自学习的能力，特别适合处理因素众多、机制不明晰和信息缺失的复杂问

收稿日期：2020-04-14；**修回日期：**2020-06-27。

基金项目：国家重点研发计划（2016YFC1401605、2016YFC1401800）；广东省海洋遥感重点实验室（中国科学院南海海洋研究所）开放课题（2017B030301005-LORS2011）。

作者简介：何恩业（1981-），男，助理研究员，学士，主要从事海洋生态环境、生态灾害和水文气象预报研究。E-mail：heenye@163.com

题[10]。神经网络诞生至今主要经历了感知器、浅层学习和深度学习3个阶段。Taylor等[11]研究发现大脑具有逐层处理信息的能力;人工智能之父Hinton等[12]提出具有多层次的深度神经网络(Deep Neural Networks,DNN)更易从低层信息提取高层语义特征。但是由于DNN参数众多,造成模型运算效率较低,甚至难以收敛于全局最优,陈旭伟等[13]提出了一种串联BP神经网络结构,不但可以减少模型参数,实现对多个非线性函数的拟合,还可以实现特征信息逐级提取和传递,在实验中取得了较大成功。Sutskever等[14]针对DNN容易陷入局部最小的缺陷提出了优化调整方案。这些研究成果极大地促进了深度学习的长足发展。近几年来,基于深度学习智能模型,如卷积神经网络(Convolutional Neural Network,CNN)、循环神经网络(Recurrent Neural Network,RNN)和DNN的各种研究应用已逐渐成为最热门的课题。但是当前针对Chl-*a*浓度预报的智能模型大都以浅层学习为主,鲜有将深度学习应用于预报研究之中。本文以浙江海洋保护区生态浮标监测数据为依据,建立了一种串联式DNN的Chl-*a*短期预报模型,并针对预报结果进行了检验。

2 数据来源和研究方法

2.1 数据来源

本文采用2019年5月浙江省海洋保护区3个生态浮标NB03、TZ01和WZ02的连续监测资料作为样本数据进行建模分析(见图1)。浮标监测数据主要由浙江省近岸海域浮标实时监测系统省级数字化监控平台提供。海上浮标系统由浮体、标架、供电设备、防护设备、锚系、传感器和数据采集传输等部分组成。浮体上加载的水质多参数传感器(置于水面以下0.5~0.8 m处)可获取间隔1 h的水温、盐度、pH、溶解氧及其饱和度、Chl-*a*浓度、浊度和电导率等常规水质参数数据,以及间隔4 h的氨氮、硝酸盐氮、亚硝酸盐氮和磷酸盐等数据。海上浮标系统每月1次例行维护,不定期开展应急维护,保证浮标系统运行稳定性,定期开展人工采样比对监测,确保监测数据质量。

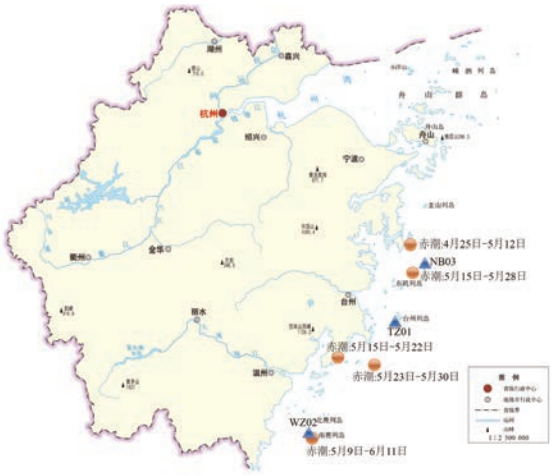

图1 2019年5月浙江沿海生态浮标位置(▲)和赤潮发生时间及中心位置(●)

2.2 基于DNN的Chl-*a*短期预报模型构建方法

DNN是具有多个隐层的神经网络,层与层之间采用全连接的结构,隐层中的任一神经元与前后层任一神经元相连。DNN具有优异的特征学习能力,对复杂函数的逼近能力极强,具有能够从少量样本集合中挖掘高阶本质特征的优势[12]。在DNN架构研究方面,Wang等[15]基于两阶段深度学习的综合推荐框架,利用潜因子向量作为深度学习推荐模型的输入,不仅捕捉到高阶交互特征,而且减轻了隐层的负担,避免了模型训练陷入局部最优。结果表明串联结构的DNN在预测精度、参数空间和训练速度等方面都体现出更好的性能。本文采用5层神经网络为基本单元,以串联方式构建了24 h和48 h的Chl-*a*浓度预报模型。

2.2.1 子神经网络

具有5层结构的子神经网络模型包含1个输入层、3个隐层和1个输出层(见图2)。输入层有$n0$个变量输入节点,隐层分别有$n1$、$n2$和$n3$个神经元节点,输出层有m个神经元节点。前后层节点之间通过耦合权值矩阵进行全连接,上一层神经元的输出作为下一层神经元的输入,神经元采用单极性Sigmoid激活函数处理信息,在输出层采用线性方式输出预测变量。Sigmoid激活函数为:

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

其一阶导数函数为:$f'(x)=f(x)\cdot(1-f(x))$。

设w_{ij}为任意两个神经元之间的耦合权值,g为神经元的输入,h为神经元的输出,则有:

$$h = f\left(\sum(w_{ij}\cdot g) - \theta\right) \tag{2}$$

式中,θ为偏置量,表示神经元的阈值。

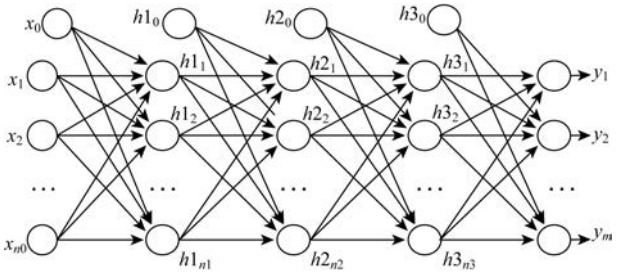

图2　具有5层结构的DNN

2.2.2　模型损失函数和参数的优化调整方案

设DNN模型期望输出为$T = \left(t_1, t_2, \cdots, t_m\right)^{\mathrm{T}}$,共有α个仿真训练样本,第$p$个样本输入模型后的输出方差为$E_P = \sum_{k=1}^{m}(t_k^p - y_k^p)^2$,α个样本输入模型后的输出总方差为$E_\Sigma = \sum_{p=1}^{\alpha}\sum_{k=1}^{m}(t_k^p - y_k^p)^2$,定义模型损失函数为:

$$E_\Sigma = \frac{1}{2}\sum_{p=1}^{\alpha}\sum_{k=1}^{m}(t_k^p - y_k^p)^2 \tag{3}$$

损失函数的值即为网络模型误差,若其值达不到目标值ε(期望损失),则进行误差反馈,对网络耦合权值按照损失函数负梯度方向从输出层至输入层进行全局调整,调整量为:

$$\Delta w_{ij} = -\eta\frac{\partial E_\Sigma}{\partial w_{ij}} \tag{4}$$

式中,η为学习率,大取值可以加快学习速度,但易导致w_{ij}调整量过大造成模型震荡难以收敛,小取值会导致模型运算时间加长,效率变差。本文对算法进行了优化,采用可变学习率,网络每迭代100次η乘以系数0.98,且设置η最小值为0.01以保证学习速度,优化后的网络模型可以在初期采用较大的学习率加快收敛速度,在后期以较小的学习率解决参数调整过大产生的模型震荡难题。此外,为了进一步提高模型运算效率和稳定性,本研究引入可变动量项$m_c\Delta w$,其作用在于记忆上一次w_{ij}的调整方向,m_c为动量系数,若本次调整与上次调整方向一致,增大为$1.1\times m_c$,加快收敛速度;若调整方向相反网络产生了震荡,则减小为$0.9\times m_c$,起到平滑作用。设n为迭代次数(调整次数),经过优化后各层之间的耦合权值调整公式可表达为:

$$\begin{cases} w4_{(k,j3)}{}^{n+1} = w4_{(k,j3)}{}^{n} + \Delta w4_{(k,j3)}{}^{n} + m_c\Delta w4_{(k,j3)}{}^{n-1}, \\ \qquad k = 1,2,\cdots,m;\ j3 = 0,1,2,\cdots,n3 \\ w3_{(j3,j2)}{}^{n+1} = w3_{(j3,j2)}{}^{n} + \Delta w3_{(j3,j2)}{}^{n} + m_c\Delta w3_{(j3,j2)}{}^{n-1}, \\ \qquad j3 = 1,2,\cdots,n3;\ j2 = 0,1,2,\cdots,n2 \\ w2_{(j2,j1)}{}^{n+1} = w2_{(j2,j1)}{}^{n} + \Delta w2_{(j2,j1)}{}^{n} + m_c\Delta w2_{(j2,j1)}{}^{n-1}, \\ \qquad j2 = 1,2,\cdots,n2;\ j1 = 0,1,2,\cdots,n1 \\ w1_{(j1,i)}{}^{n+1} = w1_{(j1,i)}{}^{n} + \Delta w1_{(j1,i)}{}^{n} + m_c\Delta w1_{(j1,i)}{}^{n-1}, \\ \qquad j1 = 1,2,\cdots,n1;\ i = 0,1,2,\cdots,n0 \end{cases} \tag{5}$$

2.2.3　Chl-*a*短期预报模型结构和参数设置

本文采用前后串联的方式将两个5层结构的神经网络进行桥接建立了一个DNN(见图3),该网络拥有1个前端输入层、1个中间桥接层、6个隐含层和1个终端输出层。该模型将24 h预报日期前2 d的生态浮标Chl-*a*敏感性理化因子作为自变量进行信息输入,前一个子神经网络输出未来24 h Chl-*a*浓度预报结果,并将该结果作为后一个子神经网络的输入参与运算,模型终端输出未来48 h Chl-*a*浓度预报量结果。

模型设置所有耦合权值w_{ij}的初值为随机小量,以保证单极性Sigmoid激活函数处于灵敏区间,加快调整速度。设置初始学习率η为0.2,初始动量系

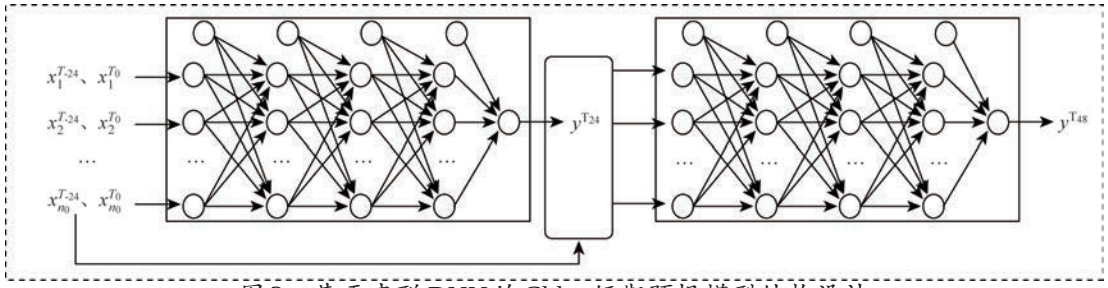

图3　基于串联DNN的Chl-*a*短期预报模型结构设计

数m_c为0.8。

本文采用Fortran软件进行数学建模与编程，利用SPSS和Excel软件分别进行统计学分析和绘图。

3 结果与分析

3.1 DNN模型验证

不同的数据预处理方法会导致模型预测结果差距较大[10]。本文利用目标函数$y=x^3$评价几种不同样本处理方法对网络模型预测效果产生的差异，设计了4套实验方案对子神经网络模型的相关特性进行测试验证，根据测试结果优劣为后续Chl-*a*短期预报模型提供最优样本数据处理方式。目标函数$y=x^3$的样本选取见表1，(x^p,y^p)表示样本对，p为样本序号，自变量x取值间隔为0.1。自变量方面额外增加了3个随机干扰变量（v1、v2和v3），共选取36条样本。

表1 目标函数$y=x^3$测试样本集合

样本序号	x取值	随机干扰$v1$	随机干扰$v2$	随机干扰$v3$	y实际值
1	0	14	2	1	0
2	0.1	17	1	9	0.001
3	0.2	10	9	4	0.008
…	…	…	…	…	…
35	3.4	6	4	7	39.304
36	3.5	84	9	0	42.875

输入层按照实验方案设置输入节点数（变量x或者变量x与干扰变量的组合），3个隐层均配置10个神经元节点，输出层为预测变量y，模型损失函数目标值为3×10^{-4}，4套测试实验方案为：

方案a：所有样本按照顺序排列，自变量x不加入干扰变量，取前30个作为仿真训练集，剩余后6个作为预测检验样本集，测试模型对未涉猎知识领域的预测效果。

方案b：打乱样本集顺序作类间交叉处理，自变量x不加入干扰变量，随机取6个样本进行预测检验，其他30个样本作为仿真训练集，测试样本类间均衡对模型预测效果产生的作用。

方案c：在方案b的基础上，对自变量x加入3个随机干扰变量$v1$、$v2$和$v3$，测试神经网络模型容错能力和剔除噪音的能力。

方案d：在方案c的基础上大幅减少仿真训练样本至20个，预测检验样本不变，测试学习样本多寡对预测效果的影响程度。

方案a实验结果显示（见图4a），由于仿真训练集未包含自变量x的所有区间，模型丧失了对后部区间学习的机会，系统学习不完整，对于未涉猎的知识处理缺乏经验，导致预测效果不理想。预测检验显示，给定自变量x值，其值偏离训练集样本区间越远，则对因变量y的预测能力就会变得越差。

方案b实验结果显示（见图4b），样本集合进行了类间交叉处理后，仿真训练样本基本包含了所有自变量区间，模型学习信息较为系统，因此在预测检验时，随机给定自变量x，模型能够根据仿真训练建立的知识库对变量y做出较准确的预测，效果提升明显。

方案c实验结果显示（见图4c），虽然在自变量信息中随机加入了$v1$、$v2$和$v3$等干扰变量，但是DNN经过自学习能够有效的剔除噪音信号，对关键信息提取效果较好。对比来看，预测效果好于方案a，相较于方案b的预测误差略有增大。该实验结果表明DNN特别适合对变量之间映射关系模糊不清或充满各种干扰噪音的复杂问题进行建模研究。

方案d实验结果显示（见图4d），减少训练样本后，模型整体预测误差稍有增大，较大误差主要分布在仿真训练集中学习类别较少的区间，对于学习样本密集度较高的区域，模型对于预测的能力仍然较强。该实验结果表明即使仿真训练样本较少，但只要做到样本类间均衡，仍能有效减小预报误差。

表2为4种实验方案的仿真训练和预测结果的均方根误差（Root Mean Square Error，RMSE）和平均绝对误差（Mean Absolute Error，MAE）的对比，4种方案模型的仿真RMSE差别较小，说明DNN对任何训练样本都能做到较高的拟合精度。但是从预测检验来看，4种方案表现出了明显的差距，即使没有加入干扰变量，方案a的预测效果仍然最差，表明样本类间均衡对于模型最终执行效果起到最为关键性的作用。虽然方案b最优，但是在解决实际问题方面，变量之间的映射关系并不清晰，很难准确

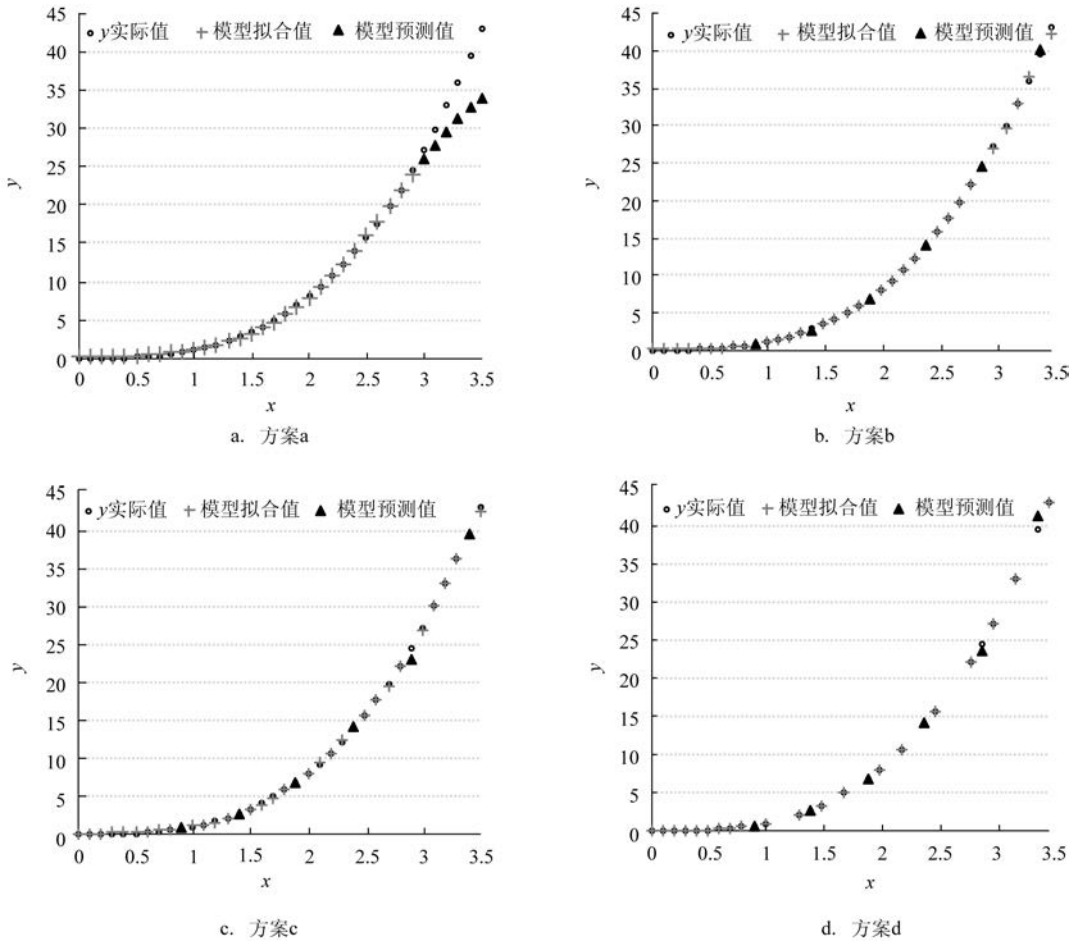

图4　不同方案DNN对目标函数模拟和预测结果比较

界定因变量是由哪种或哪几种自变量引发,所以方案c和d成了现实中最多的选择。

表2　各实验方案的DNN模型误差对比

实验方案	模型仿真 RMSE	模型仿真 MAE	模型预测 RMSE	模型预测 MAE
a	0.19	0.17	5.30	4.53
b	0.19	0.10	0.24	0.13
c	0.19	0.15	0.60	0.33
d	0.23	0.60	0.84	0.60

针对方案c和d变量过多导致的模型参数增加、学习缓慢和预测效果变差的问题,国内外众多学者也提出了多种解决方案,如利用相关性分析法、聚类算法和小波分析法等对模型输入信息进行降维去噪和特征提取,DNN的预测准确率得到了较大提高[16-18]。本文选用Pearson法筛选Chl-*a*含量相关的敏感性因子作为模型的自变量输入信息。

3.2　基于DNN的Chl-*a*短期预报模型仿真和预测结果分析

3.2.1　原始数据处理和相关性分析

由于原始数据存在大量的噪音、冗余和不完整的信息,所以在输入模型之前需要进行清洗以达到改进数据质量的目的[19-20]。本文采用如下方案对生态浮标原始数据进行处理。

对于异常值的剔除采用3σ准则:

$$P(|x-\mu|>3\sigma)\leqslant 0.003 \tag{6}$$

式中,x表示观测要素变量;σ、μ分别为标准差和均值,观测要素值超出区间$[(\mu-3\sigma),(\mu+3\sigma)]$的离群点只占0.3%,以异常值处理。对于缺失值采用五点等权滑动平均滤波法(Moving Average)进行插值填充:

$$f_k = y_k = \frac{1}{5}\sum_{k-2}^{k+2} y_k \tag{7}$$

式中，f_k为5个相邻数据y_{k-2}、y_{k-1}、y_k、y_{k+1}和y_{k+2}的平滑数据，y_k为等效监测数据。

由于各要素监测频率不一致以及部分监测要素值变化剧烈，其瞬时值与其他要素的步调性并不好。为了利于后续分析和应用，我们对各要素进行日平均处理，共形成93条记录（见图5）。Chl-*a*浓度的变化直接表征浮游植物数量变动状况，例如：2019年5月9日在生态浮标WZ02海域发现以东海原甲藻为优势种的赤潮，2019年5月15日在生态浮标NB03毗邻水域发现以东海原甲藻和夜光藻等为优势种的赤潮。图5显示出WZ02和NB03浮标的Chl-*a*浓度监测值分别在赤潮发生日期有较明显的大幅度上升，而未有赤潮发生海域的TZ01浮标处Chl-*a*变化较为平稳，波动不大。3个浮标监测数据既包括赤潮发生前后各理化要素的连续记录，同时也涵盖了正常水体Chl-*a*含量变化的连续记录，具有明显的样本均衡特性，特别适合神经网络建模样本数据。

相关性分析是机器学习样本数据预处理的核心工具，本文采用Pearson相关法分析Chl-*a*与各环境因子的相关关系，以此衡量各要素变动的一致程度，为预测模型的输入信息筛选敏感性因子。

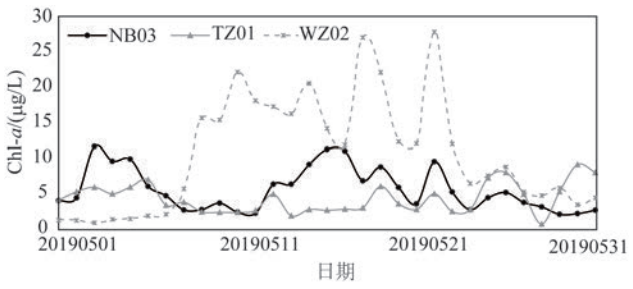

图5 2019年5月浙江海洋保护区各生态浮标Chl-*a*浓度日变化

Pearson相关法不但可以做到数据降维、减小模型参数提高学习速度，还可以有效改善预测效果，提高预测准确率[21]。Pearson相关系数公式为：

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}} \tag{8}$$

利用IBM SPSS Statistics21进行相关性分析，结果见表3。分析发现研究海域$T_{0\,h}$时刻Chl-*a*浓度与T_{-24h}和T_{-48h}（−24 h、−48 h表示0 h时刻前1 d和前2 d）时刻溶解氧浓度、pH值和Chl-*a*浓度在P=0.01水平上均呈现显著性正相关，且时间越接近相关性越强，说明Chl-*a*浓度变化与这3种要素的变化具有较好的同步性；$T_{0\,h}$时刻Chl-*a*浓度与$T_{-24\,h}$和$T_{-48\,h}$时刻的硝氮和磷酸盐在P=0.05水平上均呈现显著的负相关，且过去2 d的相关性整体高于当天，说明营养盐对浮游植物生长的影响具有滞后性效应；Chl-*a*浓度与其他水质和营养盐要素在5月的相关性不大。相关性分析结果表明，监测样本偏赤潮发生初期，处于藻类密度不大、爆发性增殖前或开端时期，藻类生物量繁殖增长时吸收表层CO_2释放氧气，而营养盐无机态也处于被消耗状态，氮和磷成为浮游植物生长的关键性限制因子[22-23]。所以，建立的模型更适用于赤潮早期或将发生期的预测。

3.2.2 模型仿真和预测结果分析

采用与Chl-*a*浓度变化具有显著相关性的环境因子作为预报因子，Chl-*a*浓度作为预测变量，两者分别作为模型的输入和输出信息在进入学习之前进行类间交叉处理。一般要求预测检验样本占总体样本的10%左右[10]，因此本文随机预留10个样本作为预测检验集，剩余样本做仿真训练集。

24 h预报方案：选取T_{-24h}和T_{0h}时刻的溶解氧、pH、硝氮、磷酸盐和Chl-*a*作为预报因子，对$T_{24\,h}$时刻的Chl-*a*浓度进行预测；

48 h预报方案：在24 h预报因子基础上加入$T_{24\,h}$时刻Chl-*a*预报值作为48 h预报因子，对$T_{48\,h}$时

表3 T_{0h}时刻Chl-*a*与不同时刻理化因子的相关性

时间	表层温度/℃	盐度	溶解氧/(mg/L)	pH	浊度/NTU	氨氮/(μg/L)	硝氮/(μg/L)	亚硝氮/(μg/L)	磷酸盐/(μg/L)	叶绿素/(μg/L)
T_{0h}	0.104	−0.063	0.673**	0.531**	−0.045	−0.008	−0.257*	0.089	−0.232*	1.000**
T_{-24h}	0.035	0.085	0.533**	0.461**	0.015	−0.013	−0.270**	0.091	−0.259*	0.748**
T_{-48h}	−0.017	0.112	0.421**	0.404**	0.048	−0.041	−0.255*	0.114	−0.307**	0.593**

** 在0.01的水平上显著；* 在0.05的水平上显著。

刻的Chl-*a*浓度进行预测。

变量之间由于量纲和数值大小不同,各度量之间的特征很难具有可比性,同时对目标函数影响权重也不一致。为了消除这些影响,本文对输入变量统一进行标准归一化处理:

$$y_i = \frac{x_i - \bar{x}}{\sigma},\ z_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \tag{9}$$

式中,σ为变量标准偏差,y_i和z_i分别为变量x_i标准化和归一化数值。经过处理后所有变量处于同等地位,也符合神经网络Sigmoid激活函数定义域的要求。

网络神经元节点数设置过少会造成信息的挖掘能力不足,设置过多又因为出现过拟合现象,即将原始数据的噪音转变为特征信号,而造成预测误差偏大[24]。通过对隐层神经元节点不同设置的实验可知(见表4),前一子网络结构为10-9-9-9-1,后一子网络结构为11-6-6-6-1时,模型预测效果最优,24 h Chl-*a*浓度预报RMSE达到最小值1.25 μg/L,MAE为1.03 μg/L,48 h预报RMSE达到最小值2.43 μg/L,MAE为1.99 μg/L。

此外,实验过程中发现,虽然DNN能够对仿真训练集进行任意精度的拟合,但是过于拟合会将原始数据中的噪音转变为网络学习特征,导致模型测试集预测精度降低(泛化效果变差)。本研究中当目标损失值(训练集拟合误差RMSE)为0.5～1.0 μg/L时,测试集的损失较小,预测效果达到较高的精度(见图6)。

为了对比深层学习与浅层学习在预测效果上的差异,本文构建了经典的单一隐层BP神经网络模型。经过同样方法进行测试,BP模型的前一子网络结构为10-8-1,后一子网络结构为11-11-1时,模型预测效果最优,24 h Chl-*a*预报RMSE达到最小值1.78 μg/L,MAE为1.42 μg/L,48 h预报RMSE达到最小值3.09 μg/L,MAE为2.20 μg/L(见表5)。

对比显示,深层DNN相比浅层人工神经网络(Artificial Neural Network,ANN),24 h Chl-*a*预报的RMSE减小了0.53 μg/L,MAE减小了0.39 μg/L;48 h Chl-*a*预报的RMSE减小了0.66 μg/L,MAE减小了0.21 μg/L,一定程度上反映了深度学习在挖掘高阶特征上比浅层学习更具有优势。另一方面,无论是深层DNN还是浅层ANN,48 h预报的RMSE比24 h的RMSE大幅增加,深层DNN增加了1.18 μg/L,浅层ANN增加了1.31 μg/L,显示出神经网络对于临近预报更有优势,随着时间跨度的加大不确定因素也会增加。该模型未将气象和水动力等对Chl-*a*含量产生重要影响的因素加入考虑,一定程度上也降低了模型在较长时间预测方面的精度。

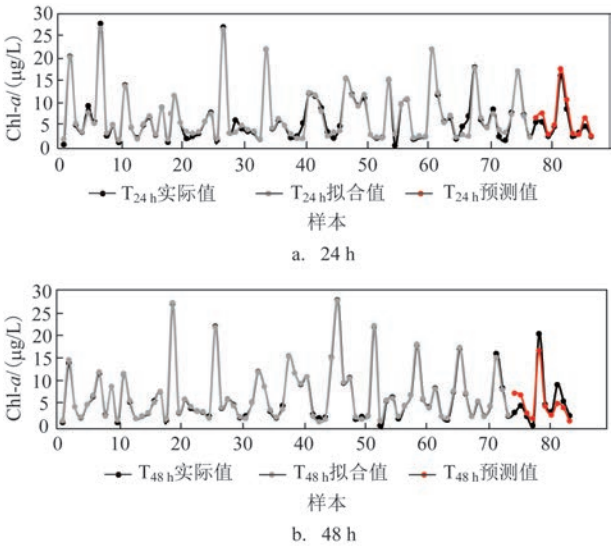

图6　串联DNN模型的Chl-*a*浓度仿真训练和预测结果

表4　不同节点设置的DNN模型预测效果对比

实验方案	24 h预报实验结果/(μg/L)			48 h预报实验结果/(μg/L)		
	子网络结构	RMSE	MAE	子网络结构	RMSE	MAE
EXP1	10-5-5-5-1	2.15	1.66	11-5-5-5-1	3.04	2.22
EXP2	10-6-6-6-1	1.43	1.11	11-6-6-6-1	2.43	1.99
EXP3	10-7-7-7-1	1.78	1.18	11-7-7-7-1	3.9	2.5
EXP4	10-8-8-8-1	2.62	1.79	11-8-8-8-1	2.77	1.86
EXP5	10-9-9-9-1	1.25	1.03	11-9-9-9-1	3.29	2.74
EXP6	10-10-10-10-1	6.1	3.77	11-10-10-10-1	7.5	5.26

表5 浅层ANN与深层DNN模型对Chl-*a*浓度的预测结果对比（单位：μg/L）

24 h预报结果对比					48 h预报结果对比				
Chl-*a*实际值	浅层ANN		深层DNN		Chl-*a*实际值	浅层ANN		深层DNN	
	预测值	偏差	预测值	偏差		预测值	偏差	预测值	偏差
5.82	5.86	0.04	6.80	0.98	3.49	3.00	−0.49	7.63	4.14
5.86	2.58	−3.28	7.83	1.97	4.93	3.48	−1.45	7.25	2.32
2.69	2.2	−0.49	3.10	0.41	2.60	3.15	0.55	3.32	0.72
4.87	1.98	−2.89	5.21	0.34	0.65	1.84	1.19	2.14	1.49
16.21	14.02	−2.19	17.70	1.49	20.49	15.15	−5.34	16.87	−3.62
8.73	7.44	−1.29	10.83	2.10	5.19	5.08	−0.11	4.85	−0.34
2.69	1.65	−1.04	3.30	0.61	3.60	1.67	−1.93	2.92	−0.68
3.49	2.9	−0.59	3.10	−0.39	9.47	3.30	−6.17	5.38	−4.09
4.93	7.01	2.08	6.75	1.82	5.88	10.46	4.58	4.49	−1.39
2.60	2.87	0.27	2.76	0.16	2.75	2.93	0.18	1.61	−1.14
RMSE	1.78		1.25		RMSE	3.09		2.43	
MAE	1.42		1.03		MAE	2.20		1.99	

4 结论

本文尝试建立一种串联DNN的Chl-*a*短期预报模型，通过对目标函数的测试验证了DNN的相关特性，利用浙江海洋保护区生态浮标数据对Chl-*a*浓度进行了仿真和预报实验。结果表明：

（1）对训练样本进行类间均衡处理比单纯增加样本数量更为有效，且具有更好的预测效果。采用传统统计方法对输入信息进行去噪和降维预处理，有利于提升预测精度。

（2）2019年5月浙江海洋保护区生态浮标水质和营养盐要素相关性分析结果显示，T_{0h}时刻的Chl-*a*与T_{-24h}和T_{-48h}时刻的溶解氧、pH和Chl-*a*在P=0.01水平上均有明显的正相关，与T_{-24h}和T_{-48h}时刻的硝氮和磷酸盐在P=0.05水平上均有明显的负相关，氮和磷是浮游植物生长的关键性限制因子。

（3）本文所建立的串联DNN Chl-*a*浓度短期预报模型24 h预报的RMSE为1.25 μg/L，MAE为1.03 μg/L，48 h预报的RMSE为2.43 μg/L，MAE为1.99 μg/L，预报精度比浅层学习提升明显，体现了深度学习从原始数据中挖掘高阶语义特征的优势。该方法不但可以减少模型参数，实现对多个非线性函数的拟合，还可以实现特征信息的逐级提取和传递，具有一定的通用性和可移植性。

本研究只考虑了水质和营养盐要素对Chl-*a*浓度变化的影响，实际上淡水输入、环流形势和上升流等水动力因子以及气温、光照和风等气象因子都会对Chl-*a*浓度产生重要影响，如果将这些影响因子一并考虑无疑会提高预测的精度，这也是未来进一步研究需要开展的工作[25-28]。此外，DNN虽然具有很强的仿真和预测性能，但是网络不同的参数设置也会对结果产生不同影响，如模型大小、结构和训练细节等设置目前没有统一的标准，仍需要不断优化[29]。总之，随着人工智能技术的不断完善和发展，基于深度学习的Chl-*a*预报模型将具有广阔的应用前景。

参考文献：

[1] 张雪, 郑小慎. 基于BP神经网络渤海湾表层叶绿素浓度反演方法探讨[J]. 海洋技术学报, 2018, 37(6): 79-87.

[2] 郭文景, 符志友, 汪浩, 等. 水华过程水质参数与浮游植物定量关系的研究——以太湖梅梁湾为例[J]. 中国环境科学, 2018, 38(4): 1517-1525.

[3] 阮华杰, 马骏, 何志强. 生态浮标预测赤潮暴发的分析[J]. 声学与

电子工程, 2014(2): 44-46, 49.
[4] 金衍健, 卓丽飞. 舟山沿岸海域叶绿素a时空分布及与水质因子的相关分析[J]. 浙江海洋大学学报(自然科学版), 2017, 36(5): 389-395.
[5] 林祥. 诏安湾养殖区叶绿素a与水质因子的主成分线性多元回归分析[J]. 环境影响评价, 2018, 40(5): 88-90, 96.
[6] 杨德周, 尹宝树, 俞志明, 等. 长江口叶绿素分布特征和营养盐来源数值模拟研究[J]. 海洋学报, 2009, 31(1): 10-19.
[7] 崔玉洁. 三峡水库香溪河藻类生长敏感生态动力学过程及其模拟[D]. 武汉: 武汉大学, 2017.
[8] 张娣, 景元书, 李亚春, 等. 基于ARMA-BP集成的藻类叶绿素a预测研究[J]. 气象科学, 2015, 35(3): 312-316.
[9] 石绥祥, 王蕾, 余璇, 等. 长短期记忆神经网络在叶绿素a浓度预测中的应用[J]. 海洋学报, 2020, 42(2): 134-142.
[10] 陈祥光, 裴旭东. 人工神经网络技术及应用[M]. 北京: 中国电力出版社, 2003, 9: 22-31.
[11] Taylor J G. On Intelligence, Jeff Hawkins, Sandra Blakeslee, Times Books (2004)[J]. Artificial Intelligence, 2005, 169(2): 192-195.
[12] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[13] 陈旭伟, 傅刚, 陈环. 基于串联BP神经网络多函数拟合的研究设计[J]. 现代电子技术, 2013, 36(22): 14-16.
[14] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]//Proceedings of the 30th International Conference on Machine Learning. Atlanta: JMLR. org, 2013: 1139-1147.
[15] Wang R Q, Jiang Y L, Lou J G. TDR: Two-stage deep recommendation model based on mSDA and DNN[J]. Expert Systems with Applications, 2020, 145: 113116.
[16] 王瑞琴, 吴宗大, 蒋云良, 等. 一种基于两阶段深度学习的集成推荐模型[J]. 计算机研究与发展, 2019, 56(8): 1661-1669.
[17] JayaLakshmi A N M, Kishore K V K. Performance evaluation of DNN with other Machine learning techniques in a cluster using Apache Spark and MLlib [J]. Journal of King Saud University－Computer and Information Sciences, 2018. Doi：https://doi. org / 10.1016/j.jksuci.2018.09.022.
[18] 刘胜辉, 张人敬, 张淑丽, 等. 基于深度神经网络的切削刀具剩余寿命预测[J]. 哈尔滨理工大学学报, 2019, 24(3): 1-8.
[19] 卢勇夺, 王朝阳, 王豹, 等. 我国海洋锚系浮标数据异常值检测方法研究——以QF110和QF306为例[J]. 海洋预报, 2019, 36(6): 37-43.
[20] 祖子清, 朱学明, 王辉, 等. Argo数据处理系统设计与应用[J]. 海洋预报, 2019, 36(4): 1-12.
[21] 何沁波. 龙景湖叶绿素a浓度预测模型敏感性分析[D]. 重庆: 重庆大学, 2015.
[22] Kudryavtseva E, Aleksandrov S, Bukanova T, et al. Relationship between seasonal variations of primary production, abiotic factors and phytoplankton composition in the coastal zone of the south-eastern part of the Baltic Sea[J]. Regional Studies in Marine Science, 2019, 32: 100862.
[23] Liu L L, Dong Y C, Kong M, et al. Towards the comprehensive water quality control in Lake Taihu: Correlating chlorphyll a and water quality parameters with generalized additive model [J]. Science of the Total Environment, 2020, 705: 135993.
[24] 王嵘冰, 徐红艳, 李波, 等. BP神经网络隐含层节点数确定方法研究[J]. 计算机技术与发展, 2018, 28(4): 31-35.
[25] Yu X Y, Xu J, Long A M, et al. Carbon-to-chlorophyll ratio and carbon content of phytoplankton community at the surface in coastal waters adjacent to the Zhujiang River Estuary during summer[J]. Acta Oceanologica Sinica, 2020, 39(2): 123-131.
[26] Lu Z B, Liu D D, Liao J S, et al. Characterizing spatial distribution of chlorophyll a in the Southern Ocean on a circumpolar cruise in summer[J]. Science of the Total Environment, 2020, 708: 134833.
[27] Yang P P, Fong D A, Lo E Y M, et al. Circulation patterns in a shallow tropical reservoir: Observations and modeling[J]. Journal of Hydro-environment Research, 2019, 27: 75-86.
[28] Gao S, Wang H, Liu G M, et al. Spatio-temporal variability of chlorophyll a and its responses to sea surface temperature, winds and height anomaly in the western South China Sea[J]. Acta Oceanologica Sinica, 2013, 32(1): 48-58.
[29] Maas A L, Qi P, Xie Z A, et al. Building DNN acoustic models for large vocabulary speech recognition[J]. Computer Speech & Language, 2017, 41: 195-213.

Study on short-term prediction method of Chl-*a* based on cascade DNN model

HE En-ye[1]，LI Shang-lu[2]，YANG Jing[1]，JI Xuan-liang[1]，GAO Shan[1]，WANG Dan[1]

（*1. Key Laboratory of Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Ministry of Natural Resources, Beijing 100081 China; 2. Marine Monitoring & Forecasting Center of Zhejiang Province, Hangzhou 310007 China*）

Abstract：Based on the monitoring data of ecobuoys in Zhejiang marine protected area in May 2019, this paper

analyses the correlation between Chl-*a* and physicochemical factors. Statistics shows that Chl-*a* is positively correlated with dissolved oxygen and pH at the level of P=0.01, while it is negatively correlated with nitrate and phosphonate at the level of P=0.05. In addition, a Chl-*a* short-term prediction model is established, which constructs a cascade deep neural network (DNN) with 6 hidden layers in series by taking 5-layer neural network as the basic unit. The experimental results show that the cascade DNN model can accurately predict the short-term variation trend of Chl-*a* with higher prediction accuracy compared to the shallow neural network. The RMSE of 24 h and 48 h prediction is 1.25 μg/L and 2.43 μg/L, respectively. The MAE of 24 h and 48 h prediction is 1.03 μg/L and 1.99 μg/L, respectively.

Key words: DNN; neural network; deep learning; cascade neural network; Chl-*a*