DOI:10.11737/j.issn.1003-0239.2021.01.006

基于双隐层ANN模型的叶绿素a浓度智能预报方法

何恩业[1]，杨静[1]，李尚鲁[2]，高姗[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： 利用2019年5月WZ02生态浮标监测数据,建立了两种不同隐层人工神经网络(ANN)模型的叶绿素a(*Chl-a*)智能预报方法,并对单隐层和双隐层模型的预测结果做了对比。结果表明:双隐层结构预测结果精度更高,泛化能力更强,一定程度上说明了深层学习比浅层学习对信息的主要特征提取能力更有优势。同时,对数据样本集合进行了系统预处理。结果显示:*Chl-a*浓度与溶解氧、pH、浊度和氨氮都有显著的相关性,与表层温度、盐度、亚硝氮和磷酸盐在限定时间段内的相关性不大。通过对模型预测结果的对比验证,发现数据预处理对数据质量的改进、数据挖掘执行效率和执行效果(预测结果)都起到明显的正向作用。

关键词： 神经网络;智能预报;深层学习;叶绿素a;数据预处理

中图分类号: X55 **文献标识码:** A **文章编号:** 1003-0239(2021)01-0043-11

1 引言

作为衡量海水初级生产力及富营养化程度的一项基本指标,叶绿素a(Chlorophylla,*Chl-a*)不仅可以表征藻类现存量,同时也是海洋水体各项生物性指标、物理性指标和化学性指标的综合体现。研究海水中*Chl-a*浓度的分布、变化以及未来发展趋势对控制海水富营养化、藻类生物量以及揭示富营养化内在的一般规律起到不可替代的作用。但是*Chl-a*浓度与环境因子之间经常呈现出复杂和不确定的非线性映射关系,应用传统方法预测*Chl-a*浓度的效果较差。

近些年来,国内外学者针对水体*Chl-a*含量预测方法展开了多方面的研究。如刘建萍等[1]构建了环境因子与*Chl-a*的多元统计回归和单隐层反向传播模型-人工神经网络(Back Propagation-Artificial Neural Network,BP-ANN)模型,通过对比发现BP模型预测结果具有很大优势;黄文超等[2]以三明地区某水源地为例,分析了水体*Chl-a*和环境因子之间的动态响应关系,筛选出敏感性因子后建立了多元回归方程,并对*Chl-a*的变化趋势进行预测;许云峰等[3]应用支持向量机回归(Support Vector Machine,SVR)算法,构建了程海富营养化水体水质指标与*Chl-a*浓度的最佳SVR预测模型;仝玉华等[4]根据天津市于桥水库常规监测的水生生态数据,提出了一种结合时序方法的可自优化的径向基函数(Radial Basis Function,RBF)神经网络智能预测模型,结果显示对*Chl-a*浓度的预测比较满意,基本反映了其未来的发展趋势;王玲玲等[5]采用数值方法,基于水动力模拟预测*Chl-a*浓度的输运扩散分布,结合实际观测值进行分析,建立了包络线方程并以此预测*Chl-a*浓度的演变趋势;张颖等[6]根据杭州湾某海域监测数据的特点以及降阶的切实需要,构建了最近邻模糊聚类和最优模糊逻辑系统相结合的算法模型对海水*Chl-a*含量状态进行预测。

当前,基于ANN的智能预报方法逐渐应用到海洋生态环境预报研究之中。ANN是对人类大脑结构和功能的数学模拟,是一个大规模的非线性自适

收稿日期: 2020-01-20; **修回日期:** 2020-03-03。
基金项目: 国家重点研发计划(2016YFC1401605、2016YFC1401800)。
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

应系统，具有极强的动态处理能力，尤其是对缺损不完整或混沌不清的信息具有很强的容错能力，可以采用自适应的模式识别方法对目标变量进行预测。但是以往针对*Chl-a*浓度预测的智能模型研究大都采用单一隐层的ANN，其固有缺陷是拟合能力虽强，但泛化能力较弱，对信息主要特征的提取存在不足。实际上，人脑对信息的处理并非依靠单层神经细胞，而是具有分层提取的能力，它能够逐层剔除干扰主特征的噪音信号，最终获得高层语义信号。随着大数据和计算机技术的高速发展，近年来基于类似人脑分层模型结构的多隐层（隐含层超过2个的ANN）深度学习算法蓬勃发展[7]，Hinton等[8]提出多隐层神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，同时针对深度学习存在的网络缺陷给出了解决方案。本文利用2019年5月温州南麂马祖岙生态浮标WZ02监测数据，分别建立了单隐层和双隐层两种不同隐层的ANN结构模型对*Chl-a*浓度进行预测，对两种不同隐层的预测结果做了对比，以此评价浅层学习和深层学习在水体*Chl-a*预测效果上的优劣。同时，针对以往文献在模型构建方面阐述较多而在数据预处理方面介绍相对较少的不足，本文专门对数据处理的流程和方法做了较为详细的阐述。从本文实验结果对比验证来看，数据预处理与ANN模型的结合，不仅改进了数据的质量，而且对数据挖掘执行效率和执行效果都起到了明显的正向作用。

2 材料和方法

2.1 数据来源

利用浙江省海洋监测预报中心提供的2019年5月1—31日温州南麂马祖岙生态浮标WZ02的连续监测资料作为样本数据进行处理分析（见图1）。WZ02监测要素主要包含水质要素（表层温度、盐度、溶解氧、pH、*Chl-a*、浊度）和营养盐要素（磷酸盐、氨氮、硝氮和亚硝氮）等。水质要素监测频率为1h/次，营养盐要素监测频率为4 h/次。

*Chl-a*是衡量藻类等浮游生物分布的最基本指标之一。*Chl-a*含量升高，一般意味着浮游植物数量增大，藻华暴发最显著的表征量是*Chl-a*浓度的急剧增加[1]。2019年5月9日—6月11日，在温州市南

图1 生态浮标WZ02布放位置（▲）和2019年5月浙江沿海赤潮发生时间及中心位置（○）

麂列岛-北麂列岛-洞头列岛以东海域发生以东海原甲藻为优势种的赤潮，最大面积达800 km^2，中心位置为（121.0548°E，27.4652°N）。生态浮标WZ02的位置（121.0476347°E，27.47545277°N）正处于上述赤潮发生海域，监测数据既包括赤潮发生前各理化要素的连续记录，同时也涵盖了赤潮发生和发展过程的连续记录，这对于分析研究各理化因子和*Chl-a*浓度在赤潮发生前后的变化、相互影响以及建立它们之间的映射关系具有重要的科学和现实意义。

2.2 数据预处理

在数据挖掘中，原始数据存在着大量不完整、有噪声和偏离的数据，数据预处理目的就是通过填充缺失值、光滑噪声并识别离群点剔除异常值以及纠正数据中的不一致性等过程，改善数据样本集合的质量，进而优化其后数据挖掘执行效率和结果。

2.2.1 数据样本集合的清洗

（1）失真值的处理

因监测设备维护或故障会造成监测数据成为失真值，数据记录中会对各监测要素状态进行描述，若状态显示非“正常”即是失真值。2019年5月，生态浮标WZ02水质监测要素共有744条记录，其中因监测设备维护或故障造成浊度（NTU），有16个记录值失真，其他水质要素（表层温度、盐度、溶解氧、pH和*Chl-a*）各有15个记录值失真；营养盐监测

要素共有195条记录,其中有4条记录失真(含磷酸盐、氨氮、硝氮和亚硝氮)。对失真数据做剔除处理,剔除失真值后的各要素时间序列见图2。

(2)异常值的处理

异常值是指数据存在不合理的或明显偏离其余观测值的个别数据,在实际数据处理阶段,一般是对超出特定范围或区域的离群点作为影响数据质量的异常点或噪声进行处理。本研究利用拉依达准则(3σ准则)剔除异常值。3σ准则公式为:

$$P(|x-\mu|>3\sigma)\leqslant 0.003 \quad (1)$$

式中:σ为标准差;μ为均值;观测样本要素数值分布在区间[$(\mu-3\sigma),(\mu+3\sigma)$]内的概率超过了99.7%,超出这个范围的离群点仅占不到0.3%。针对生态浮标WZ02各监测要素,未落在此区间的数据以异常值处理,将其从监测数据中剔除(剔除标准见表1)。在各理化因子中,浊度、*Chl-a*浓度变化比较剧烈,其均方差σ都超过了样本平均值μ;表层水温、盐度和pH变化比较平稳,其均方差σ均未超过样本平均值μ的5%。生态浮标WZ02各监测要素剔除异常值后的时间序列见图3。

(3)缺失值的处理

数据集缺失值的处理是数据预处理工作中必不可少的一步。处理方式有3种:一是删除缺失数据;二是进行数据替换;三是进行数据插补。在样本集合样本偏少的状况下,直接去除缺失值会耗损大量资源,同时还会导致分析结果偏离真实性甚至导致错误结论。由于生态浮标WZ02监测数据样本数量有限,缺失点占比较少,且分布较为离散,本文选用常用的滑动平均滤波法(Moving Average)对缺失的监测数据做插值处理。一般公式为:

$$f_k=y_k=\sum_{i=q}^{p}w_i y_{k+1},k=q+1,q+2,\cdots,N-p \quad (2)$$

式中:f_k为m个相邻数据y_q, y_{q+1}, …, y_k, …, y_{p-1}, y_p的平滑数据;y_k为缺失点的等效监测数据;w_i

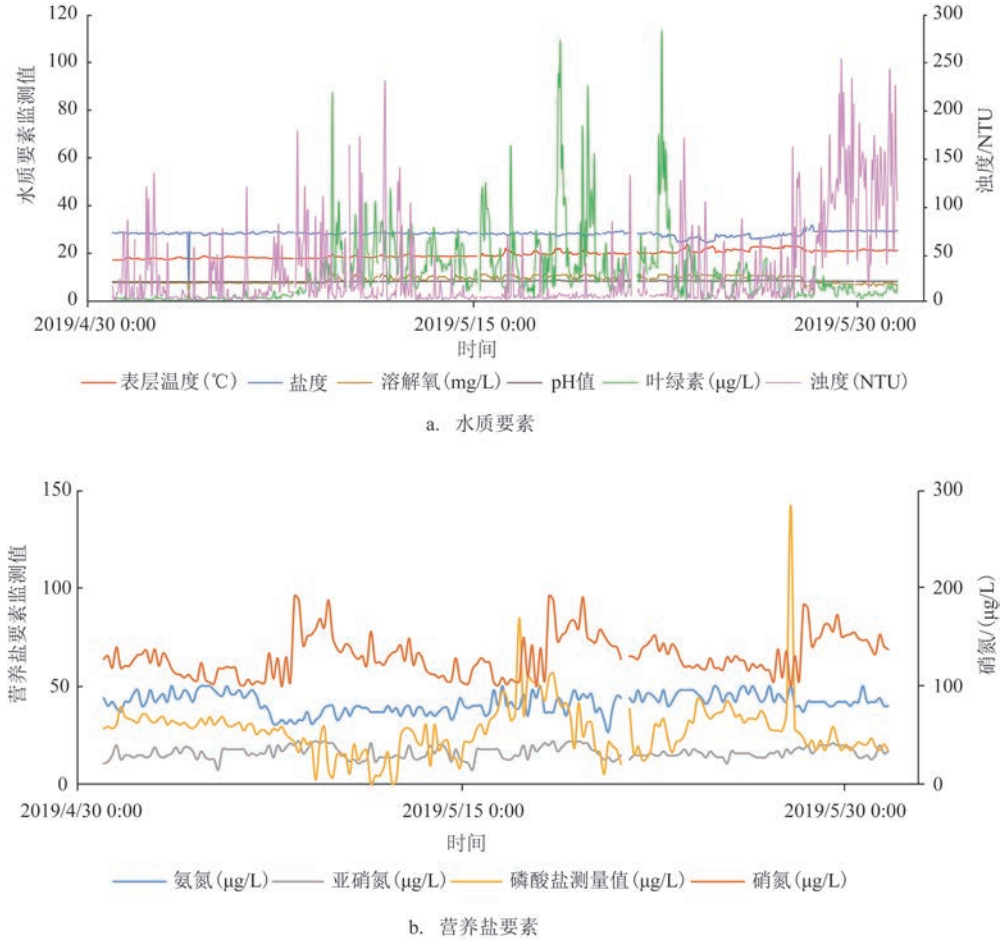

图2　2019年5月生态浮标WZ02各要素的时间序列变化

表1 2019年5月生态浮标WZ02各监测要素平均值、均方差和3σ剔除标准

监测要素	表层温度/℃	盐度	溶解氧/(mg/L)	pH	叶绿素/(μg/L)	浊度/NTU	氨氮/(μg/L)	硝氮/(μg/L)	亚硝氮/(μg/L)	磷酸盐/(μg/L)
平均值μ	19.84	28.40	9.17	8.50	11.33	30.23	41.65	131.18	16.36	28.19
均方差σ	1.54	1.22	1.45	0.18	14.12	45.91	4.93	20.96	2.92	14.14
$\mu-3\sigma$	15.22	24.74	4.82	7.96	−31.03	−107.5	26.86	68.3	7.6	−14.23
$\mu+3\sigma$	24.46	32.06	13.52	9.04	53.69	167.96	56.44	194.06	25.12	70.61
剔除后剩余/个	729	726	730	729	715	711	185	185	185	183

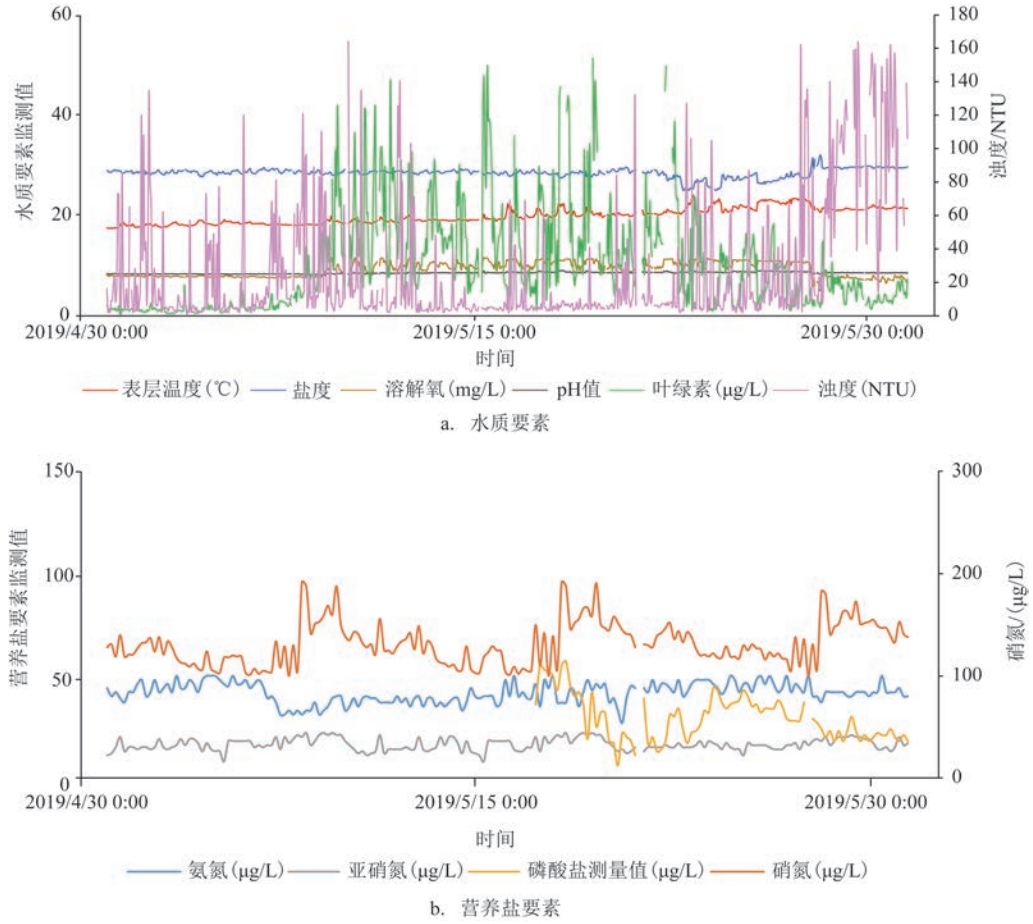

图3 剔除异常值后生态浮标WZ02各监测要素时间序列变化

为权系数，且$\sum_{i=q}^{p} w_i = 1$；p、q为任一自然数，且$p + q = m - 1$；N为变量序列个数。本文利用五点等权中心平滑法（最近邻居平均算法）插值处理生态浮标WZ02监测数据（已完成异常值剔除处理的数据），插值后形成完整的744条水质要素记录和186条营养盐要素记录（见图4）。

（4）数据的整合

数据整合是将多个数据源中的数据利用通用的标准和规范，结合并存放在一致的数据存储中。数据整合在关键性指标分析、关键性要素分析和深入层级化要素分析等方面起到重要作用。温州南麂马祖岙生态浮标WZ02各要素监测的时间点和频率存在一定差异，造成时间序列的不一致；而且由于部分监测要素一天内变化非常剧烈，经常出现跳变，其瞬时值与其他要素的同步性并不好，不利于

后续分析和应用。因此本文将通过插值填充处理后的WZ02站点各要素数据进行日平均处理,整合至统一的数据文件,共形成31条样本记录(见表2)。

2.2.2　相关分析和数据变换

(1)相关性分析

相关性分析是量化不同因素间变动状况一致程度的重要指标。在样本数据降维(通过消元减少降低模型复杂度并提高模型泛化能力)、缺失值估计和异常值修正方面发挥着极其重要的作用,是机器学习样本数据预处理的核心工具。本文采用皮尔森(Pearson)相关法分析*Chl-a*与水质和营养盐要素的相关关系。Pearson相关系数公式为[9]:

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}} \quad (3)$$

X的离均差平方和:

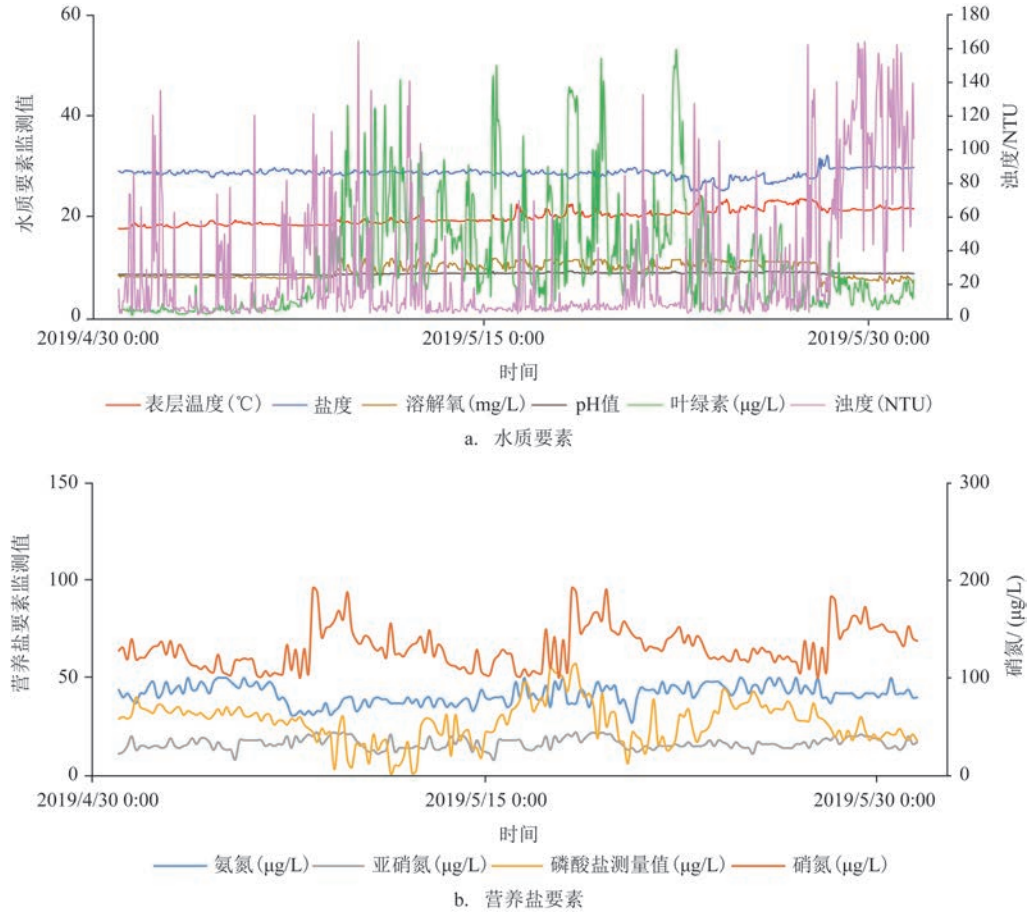

图4　插值填充后各监测要素时间序列值

表2　数据整合后WZ02各监测要素日平均值

时间	要素									
	表层温度/℃	盐度	溶解氧/(mg/L)	pH	叶绿素/(μg/L)	浊度/NTU	氨氮/(μg/L)	硝氮/(μg/L)	亚硝氮/(μg/L)	磷酸盐/(μg/L)
20190501	17.82	28.77	7.89	8.31	1.18	21.19	40.67	126.33	14.33	32.33
20190502	17.93	28.53	7.9	8.29	1.15	34.17	44.33	129	14.67	34
…	…	…	…	…	…	…	…	…	…	…
20190530	21.41	29.68	7.26	8.52	3.38	100.42	43	151.17	17.17	20
20190531	21.44	29.56	7.34	8.52	4.38	109.01	41.67	141.83	16	20.33

表3 T_0时刻Chl-a与各理化因子的相关性(2019年5月WZ02监测数据)

时间	要素								
	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)
T_0	0.094	−0.152	0.711**	0.407*	−0.367*	−0.315	0.298	0.068	−0.177
T_{-24}	−0.023	0.072	0.518**	0.225	−0.262	−0.536**	0.177	−0.003	−0.205
T_{-48}	−0.102	0.136	0.370*	0.056	−0.211	−0.678**	0.191	0.164	−0.265
T_{-72}	−0.199	0.202	0.236	−0.051	−0.14	−0.573**	0.265	0.331	−0.296

注:**在0.01的水平上显著;*在0.05的水平上显著。

$$l_{XX}=\sum(X-\bar{X})^2 \tag{4}$$

Y的离均差平方和:

$$l_{YY}=\sum(Y-\bar{Y})^2 \tag{5}$$

X与Y间的离均差积和:

$$l_{XY}=\sum(X-\bar{X})(Y-\bar{Y}) \tag{6}$$

显著性检验:总体相关系数用ρ表示;样本相关系数用r表示。

a.提出假设 H0:ρ=0 无关

H1:$\rho\neq0$ 相关

b.确定显著性水平α=0.05

利用SPSS Statistics21软件进行统计学分析。结果显示:T_0时刻*Chl-a*浓度与T_0—T_{-72}(0～滞后72 h,下同)时刻溶解氧浓度都呈显著正相关,与T_0时刻的pH值呈现显著正相关,与T_0时刻的浊度呈显著负相关,与T_{-24}—T_{-72}时刻(滞后24～72 h)的氨氮呈现显著的负相关,且都通过了0.05显著性水平的检验;与硝氮具有弱正相关性,但未通过0.05显著性水平的检验。*Chl-a*浓度与表层水温、海水盐度、亚硝氮和磷酸盐在限定的时间段内相关性不大(见表3)。相关性分析结果表明:藻类的增殖(*Chl-a*浓度增加)伴随着溶解氧浓度和pH值的升高,同时对环境水体的氨氮进行消耗,致使其浓度走低。浊度与*Chl-a*浓度呈现反相位变化趋势,原因之一可能是浊度增加造成光照强度下降,水体藻类的光合作用功能减弱,藻类生长繁殖受到抑制。

(2)数据变换

数据中不同特征的量纲存在不一致现象,各变量数值间的差距也很大,不进行处理会影响到数据分析的结果。其次,在ANN模型中也要求对各变量进行初始化,以迫使各变量处于同等地位[10]。因此,为了减轻模型仿真训练的难度必须将输入数据标准归一化到一个较小的范围之内。另一方面,S型对数激活函数(ANN模型中神经元的激活函数)输出值在0～1之间,为了防止数值溢出,也必须将输入数据信息进行标准化和归一化处理。因此,需要对数据按照一定比例进行缩放,使之落在一个特定区域,达到适用于数据挖掘的目的[10]。设x_i为变量原始值,$\bar{x}$为变量平均值,变量标准化y_i算法为:

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n},\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n},y_i=\frac{x_i-\bar{x}}{\sigma} \tag{7}$$

式中:$i=1,2,\cdots,n$,n为样本数量;σ为变量标准偏差。

归一化公式为:

$$z_i=\frac{y_i-y_{\min}}{y_{\max}-y_{\min}} \tag{8}$$

式中:$y_{\min}$为y_i的最小值,$y_{\max}$为y_i的最大值;当$z_i=0$时,令$z_i=0.000\,001$,当$z_i=1$时,令$z_i=0.999\,999$。这样经过标准归一化变换后,变量在数学上能达到非线性函数定义域的要求(生态浮标WZ02各监测要素标准归一化相关参数见表4)。

表1和表4对比显示,生态浮标WZ02各监测要素经过数据预处理后,集合样本中浊度值的均方差已经由原始数据集的45.91 NTU下降至3.8 NTU,硝氮浓度的均方差由原始数据集的20.96 μg/L下降至1.99 μg/L,磷酸盐浓度的均方差由原始数据集的14.14 μg/L下降至7.94 μg/L,叶绿素浓度的均方差由原始数据集的14.12 μg/L上升至28.35 μg/L,氨氮浓度的均方差由原始数据集的4.93 μg/L上升至15.69 μg/L,其余要素变动不大。这说明数据预处理一方面能够改善部分监测要素数据跳变给后续分析带来的困难,另一方面对叶绿素和氨氮等与藻华

表4　WZ02各监测要素(日平均数据)标准化参数

监测要素	表层温度/℃	盐度	溶解氧/(mg/L)	pH	叶绿素/(μg/L)	浊度/NTU	氨氮/(μg/L)	硝氮/(μg/L)	亚硝氮/(μg/L)	磷酸盐/(μg/L)
平均值	19.84	28.44	9.18	8.50	27.75	41.67	131.19	16.35	27.31	10.49
均方差	1.45	0.80	1.25	0.17	28.35	3.80	15.69	1.99	8.93	7.94
标准化后最小值	−1.39	−2.77	−1.53	−1.48	−0.83	−2.46	−1.45	−1.52	−1.96	−1.22
标准化后最大值	2.07	1.54	1.41	2.04	2.94	1.71	2.17	2.34	2.58	2.18

相关性较大的要素变化表征更为显著,有利于提升后续数据挖掘的执行效果。

2.3　基于BP算法的ANN模型构建

ANN模型是一种非线性映射人工神经网络,具有强大的信息处理能力。ANN由信息输入层、结果输出层和中间一个或多个隐层组成(见图5),各层神经元之间通过耦合权值进行连接。BP-ANN模型算法是一种通过误差逆向反馈,应用梯度法(最速下降)调整各层耦合权值的学习算法。网络的信号传送分为两部分:第一部分为正向传送,仿真训练集进入输入层,经网络模型计算处理后,最终结果提交至输出层;第二部分为逆向传播,如果输出结果与期望结果(实际值)之间存在较大偏差,则通过误差逆向反馈逐层调整耦合权值,如此迭代直到实际输出结果与期望结果之差小于预先设定的误差为止[10-11]。

2.3.1　ANN网络结构模型

具有两个隐层的BP-ANN网络结构如图5所示。该网络设有两个隐层,首层为信息输入层,第2、3层为隐层,第4层为输出层。所有隐含层的神经元节点采用Sigmoid激活函数[10]。

Sigmoid函数:

$$f(x)=\frac{1}{1+e^{-x}} \tag{9}$$

层1为信息输入层,$x_j, j=1,2,\cdots,n_0$,n_0为自变量的个数。

层2构建$n1$个节点,其输出向量为$g=(g_0,g_1,g_2,\cdots,g_{n1})$。

层3构建$n2$个节点,其输出向量为$h=(h_0,h_1,h_2,\cdots,h_{n2})$。

层4为预测量输出层,设有m个节点,网络的输出向量为$y=(y_1,y_2,\cdots,y_m)$。

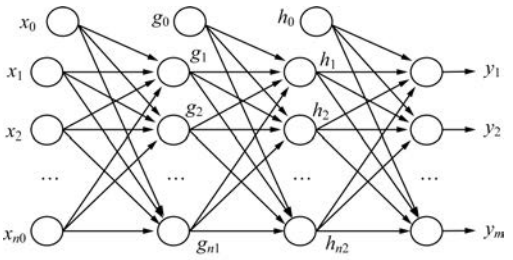

图5　双隐层ANN网络结构

隐层1神经元j输出量为:

$$g_j=f\left(\sum_{i=0}^{n_0}w_{ji}x_i-\theta_j\right),\ j=1,2,\cdots,n_1 \tag{10}$$

隐层2神经元k输出量为:

$$h_k=f\left(\sum_{j=0}^{n_1}v_{kj}g_j-\theta_k\right),\ k=1,2,\cdots,n_2 \tag{11}$$

输出层神经元l输出量为:

$$y_l=f\left(\sum_{k=0}^{n_2}r_{lk}h_k-\theta_l\right),\ l=1,2,\cdots,m \tag{12}$$

式中:x_0、g_0和h_0为各层Sigmoid激活函数的阈值;w、v和r是各层神经元之间的耦合权值;θ为各神经元的阈值。

2.3.2　网络耦合权值的动态调整算法

输入α个样本对(x^p,t^p),经模型正向传送运算后其输出结果与期望结果之间形成的总误差定义为:

$$E_\Sigma=\frac{1}{2}\sum_{p=1}^{\alpha}\sum_{l=1}^{m}(t_l^p-y_l^p)^2 \tag{13}$$

式中:y_l^p表示ANN模型的输出结果;t_l^p表示期望结果。设定E_Σ为目标函数,采用梯度法(最速下降法)逆向调整各层耦合权值,使E_Σ达到期望值。设w_{uv}为ANN模型中任意两个神经元间的耦合权值,沿梯度向量的反方向调整w_{uv}:

$$\Delta w_{uv}=-\eta\frac{\partial E_\Sigma}{\partial w_{uv}} \tag{14}$$

η(通常取值0.01～1.0)为学习率。η若取值过

小，会致使网络耦合权值在误差逆向调整过程中修改量过小，造成学习速度缓慢，网络收敛时间过长；η取值过大，可能导致网络在局部最小误差附近持续震荡，难以收敛至全局最优，甚至致使网络仿真训练陷入死循环，达不到期望的结果。

网络耦合权值在调整的过程中，误差函数E_{Σ}可能会产生平坦区。为了加快脱离误差曲面的平坦区，提高训练速率，在网络参数每次调整过程中引入动量项$m_c\Delta w_{uv}$，其功能是附加之前时刻耦合权值的调整方向，发挥其惯性效应。一是可以在平坦区逐渐加大调整力度，减少网络运算时间；二是克服可能出现的局部震荡（脱离在局部最小误差的循环跳动）。网络模型经过优化后，各层神经元间的耦合权值历次修正量为：

$$\begin{cases} r_{lk(n+1)} = r_{lk(n)} + \Delta r_{lk} + m_c \Delta r_{lk(n-1)} \\ v_{kj(n+1)} = v_{kj(n)} + \Delta v_{kj} + m_c \Delta v_{kj(n-1)} \\ w_{ji(n+1)} = w_{ji(n)} + \Delta w_{ji} + m_c \Delta w_{ji(n-1)} \end{cases} \quad (15)$$

式中：n为迭代次数；m_c为动量因子；动量项的引入还可以增加学习率η可调范围的扩大，利于提高网络学习速度[10-11]。

2.3.3 建模实现过程

基于BP算法的ANN模型构建主要包含：

（1）参数初始化处理。设置网络联接权值和模型参数为均匀分布的随机小数，使任一神经元的净输入值较小，以确使Sigmoid激活函数能够输出灵敏，提高模型运算效率。

（2）信号正向传递和误差逆向传播的计算。对ANN模型输入仿真训练样本（x^p,t^p），正向计算输出值，然后从输出层至输入层反向计算任一神经节点的等效误差。

（3）按照耦合权值修正量公式调整网络的耦合权值。

（4）以新调整的耦合权值再次进行正向计算，若所有α个学习样本均满足$\left|t_i^p - y_i^p\right|^2 < \varepsilon$，则网络仿真训练完成（$\varepsilon$为预先设定的目标误差）。

3 结果与分析

3.1 ANN模型仿真与预测结果

根据数据预处理结果，选取对浮游植物生长相关性较大的溶解氧、pH、浊度、氨氮和硝氮等5个监测要素作为ANN模型优选后的输入变量，将*Chl-a*浓度作为预测量建立映射关系，一共选取了31个日平均值样本（见表5）。

在ANN模型中，若同类样本集中在一起，则样本仿真训练后，其对耦合权值集中调整的结果会摧毁权值对前一类样本的映射关系，导致网络震荡和仿真训练时间延长。为了得到理想的仿真训练集，对仿真训练集交叉输入，打乱样本序列，达到类间交叉效果，提高网络收敛速度。针对31个样本，每次取26个样本做仿真训练集，其余样本作为测试集，并对数据输入模型前进行标准归一化处理。模型参数学习率为0.15，动量系数为0.8，网络仿真训练最大误差预设值E_{Σ}为3.0×10^{-4}，网络最大迭代次数为100万次。

试验过程中，网络隐含层神经元节点数的不同设置会对模型输出结果造成显著差异。隐层神经元节点设置数量过少，对信息的提取能力会产生不足，造成预测误差偏大；节点设置过多，对信息处理又会出现过拟合现象，将原始数据的噪音信号转变为特征信号，造成模型最终泛化能力不强（见表6）。通过对网络隐含层节点数以及网络参数反复试验[12]，单隐层ANN网络结构为5-5-1时预测检验效果最优（即输入节点5个，隐层神经元节点5个，输出节点1个），测试集的总误差达到最小值0.006 7，双隐层ANN网络结构为5-10-10-1时预测检验效果

表5 *Chl-a*与各理化因子日平均值（2019年5月WZ02监测数据）

时间	要素					
	溶解氧/(mg/L)	pH	浊度/NTU	氨氮/ (μg/L)	硝氮/ (μg/L)	Chl-a/(μg/L)
20190501	7.89	8.31	21.19	40.67	126.33	1.18
20190502	7.90	8.29	34.17	44.33	129.00	1.15
…	…	…	…	…	…	…
20190530	7.26	8.52	100.42	43.00	151.17	3.38
20190531	7.34	8.52	109.01	41.67	141.83	4.38

表6　ANN模型不同隐层及节点设置对测试结果的影响

单隐层结构	测试集总误差 E_Σ^*	双隐层结构	测试集总误差 E_Σ^*
5-5-1	0.006 7	5-5-5-1	0.121 5
5-6-1	0.038 4	5-6-6-1	0.111 8
5-7-1	0.015 7	5-7-7-1	0.063 3
5-8-1	0.029 3	5-8-8-1	0.005 6
5-9-1	0.021 4	5-9-9-1	0.040 8
5-10-1	0.007 3	5-10-10-1	0.005 1
5-11-1	0.018 7	5-11-11-1	0.007 7
5-12-1	0.021 2	5-12-12-1	0.008 9

注：E_Σ^*为标准归一化数据误差.

表7　单隐层和双隐层ANN模型预测效果对比

Chl-a 实际值/(μg/L)	单隐层		双隐层	
	预测值/(μg/L)	预测偏差率/%	预测值/(μg/L)	预测偏差率/%
6.39	7.29	−14.08	8.26	29.26
1.31	1.51	−15.27	1.37	4.58
15.32	13.6	11.23	15.06	−1.70
4.38	2.52	42.47	3.45	−21.23
5.62	4.03	28.29	7.37	31.14
RMSE	1.40		1.22	
MAE	1.25		0.97	

最优,测试集的总误差达到最小值0.005 1。

利用效果最优的单隐层(5-5-1)和双隐层ANN(5-10-10-1)模型对测试样本进行预测检验(见图6)。结果显示,无论从仿真值和实际值的对比,还是预测检验值和实际值的对比,单隐层和双隐层ANN都能够较好地拟合并预测*Chl-a*浓度的变化情况。

单隐层ANN模型26条样本的仿真训练集均方根误差(Root Mean Square Error,RMSE)为0.24 μg/L,平均绝对误差(Mean Square Error,MAE)为0.07 μg/L;5条测试样本集的RMSE为1.40 μg/L,MAE为1.25 μg/L,预报平均偏差为22.27%(见表7)。

双隐层ANN模型26条样本的仿真训练集RMSE为0.04 μg/L,MAE为0.02 μg/L;5条测试样本集的RMSE为1.22 μg/L,MAE为0.97 μg/L,预报平均偏差为17.58%。

两种不同隐层的ANN模型预测结果对比显示:双隐层ANN模型无论是对训练集的仿真还是对测

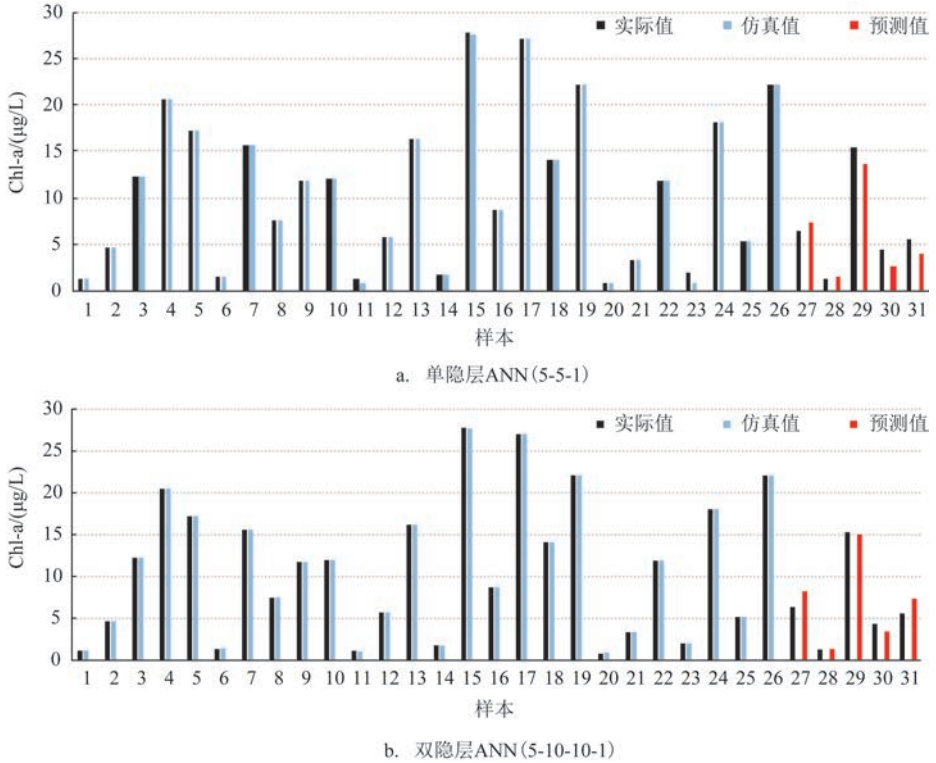

图6　ANN模型仿真值和预测值与实际值对比

试样本集的预测检验效果都更优。双隐层仿真训练集的RMSE比单隐层减小0.20 μg/L，MAE减小0.05 μg/L。在模型预测检验方面，双隐层模型预测检验集的RMSE比单隐层减小0.18 μg/L，预报平均偏差比单隐层减小4.69%。这说明双隐层ANN模型对信息的主要特征量提取要比单一隐层的ANN更加准确、预测精度更高、泛化能力更强，从而也反映了深层学习比浅层学习在实际应用中更有优势。

3.2 变量优选与否对预测效果的影响

为了检验数据预处理对数据挖掘执行效率和执行效果的影响，本文专门对未经过数据预处理和经过数据预处理的样本分别作为模型输入信息，并对输出结果做了对比实验。选择了表层温度(℃)、盐度、溶解氧(mg/L)、pH、浊度(NTU)、氨氮(μg/L)、硝氮(μg/L)、亚硝氮(μg/L)和磷酸盐(μg/L)等9个未经变量优选的监测要素作为ANN模型的输入变量，将*Chl-a*浓度(μg/L)作为预测量建立映射关系，分别使用单、双隐层ANN建立模型，并选择最优结构进行评价，结果如表8所示。

最优结构ANN模型的隐含层节点数的确定并不是固定不变的，输入变量个数的变动会导致最优结构模型的隐层节点数发生变化，它会随着输入信息维度的改变而改变。本研究中当输入变量增加至9个时，经过反复试验，单隐层最优结构为9-11-1，测试样本集的RMSE为3.24 μg/L，MAE为2.82 μg/L。双隐层最优结构为9-11-11-1，测试样本集的RMSE为1.78 μg/L，MAE为1.46 μg/L。

表8 未经变量优选的ANN模型预测效果

*Chl-a*实际值/(μg/L)	单隐层最优结构: 9-11-1		双隐层最优结构: 9-11-11-1	
	预测值/(μg/L)	预测偏差率/%	预测值/(μg/L)	预测偏差率/%
6.39	3.72	−41.78	7.57	18.47
1.31	1.43	9.16	1.45	10.69
15.32	11.32	−26.11	12.32	−19.58
4.38	1.85	−57.76	2.2	−49.77
5.62	0.82	−85.41	6.41	14.06
RMSE	3.24		1.78	
MAE	2.82		1.46	

使用未经变量优选预处理的数据作为模型输入信息，无论是单隐层还是双隐层ANN模型，其预测检验误差都要比经过变量优选预处理的要更大一些，预测效果要差一些。其中，单隐层ANN模型预测的RMSE上升了1.78 μg/L，预测偏差由22.27%上升到44.04%；双隐层ANN模型预测的RMSE上升了0.56 μg/L，预测偏差由17.58%上升到22.51%。这说明正确的数据预处理流程及方法对数据挖掘执行效率和执行效果有着积极的正向作用。

结果还显示，在未经变量优选处理的情况下，双隐层ANN模型预测结果的RMSE比单隐层模型减小1.46 μg/L，双隐层预测效果仍然好于单隐层效果，并且优势进一步扩大。同时，未经变量优选的双隐层ANN模型预测效果已经逼近了变量优选处理情况下的单隐层ANN模型，两者RMSE仅相差0.32 μg/L(见表7、8)，这再次印证了深层学习比浅层学习对录入信息噪音的剔除以及对数据关键性特征的逐层提取更有深度学习的优势。

4 小结

本文基于两种不同隐层的人工神经网络(ANN)模型尝试建立了*Chl-a*浓度的智能预报方法，对单隐层和双隐层两种ANN模型的预测效果做了对比。此外，本文还对数据预处理的流程、方法以及对模型输出结果的影响做了对比分析，结果表明：

(1)*Chl-a*浓度变化，尤其在藻华暴发前后的变化较为剧烈，它与各理化因子之间呈现极其复杂的非线性关系，基于BP算法的ANN模型能够较好地映射这种模糊的内在联系，具有较好的*Chl-a*浓度预测能力。

(2)多隐层和单隐层都有很强的表达能力，但多隐层的ANN比单隐层ANN模型对信息主特征提取能力更强，泛化能力更好，预测效果更优，具有较好的深度学习优势。在经过变量优选后，多隐层比单隐层ANN模型对*Chl-a*浓度预测的均方根误差减小了0.18 μg/L；在未经变量优选情况下，多隐层比单隐层ANN模型对*Chl-a*浓度预测的均方根误差减小了1.46 μg/L；未经变量优选情况下的多隐层ANN模型预测效果逼近了变量优选情况下的单隐层ANN模型。

(3)经过数据预处理的优选变量,不但能够对大量数据进行降噪、降维处理,提高模型的运算效率,还能够明显改善预报结果,提高预报精度和准确度。优选变量后的双隐层ANN模型预测的均方根误差下降了0.56 μg/L,预测偏差由22.51%下降至17.58%。

浙江沿海赤潮大部分由浮游植物引发,而*Chl-a*浓度与浮游植物密度的变化趋势吻合度是一致的,利用深层学习的ANN模型对*Chl-a*浓度进行预测,能够较好地抓住理化因子对*Chl-a*影响的关键特征信息,在赤潮预测中具有广阔的应用前景。

参考文献:

[1] 刘建萍, 张玉超, 钱新, 等. 太湖叶绿素a浓度预测模型初探[J]. 环境保护科学, 2009, 35(4): 46-49.
[2] 黄文超. 叶绿素a与环境因子关系的研究——以三明地区某水源地为例[J]. 环境保护科学, 2013, 39(4): 57-60, 96.
[3] 许云峰, 马春子, 霍守亮, 等. 以程海为例用支持向量机回归算法预测叶绿素a浓度[J]. 环境工程技术学报, 2012, 2(3): 207-211.
[4] 仝玉华, 周洪亮, 黄浙丰, 等. 一种自优化RBF神经网络的叶绿素a浓度时序预测模型[J]. 生态学报, 2011, 31(22): 6788-6795.
[5] 王玲玲, 戴会超, 蔡庆华. 香溪河水动力因子与叶绿素a分布的数值预测及相关性研究[J]. 应用基础与工程科学学报, 2009, 17(5): 652-658.
[6] 张颖, 李鹏, 邬益川. 基于模糊最近邻聚类学习算法的海水藻类繁殖状态预测研究[J]. 东南大学学报(自然科学版), 2011, 41(S1): 32-35.
[7] 余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.
[8] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[9] 何恩业, 王丹, 黄莉, 等. 西太平洋副热带高压的变动对我国赤潮发生的影响分析[J]. 海洋预报, 2015, 32(4): 83-89.
[10] 何恩业, 李海, 任湘湘, 等. BP人工神经网络在渤海湾叶绿素预测中的应用[J]. 海洋预报, 2008, 25(2): 1-10.
[11] 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: 22-31.
[12] 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.

Intelligent prediction method for *Chl-a* based on the artificial neural network

HE En-ye[1], YANG Jing[1], LI Shang-lu[2], GAO Shan[1]

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

Abstract: Using the monitoring data of ecological buoy WZ02 in May 2019, two intelligent prediction methods for Chlorophyll are established based on the artificial neural network (ANN), and the prediction results of the single hidden layer model and double hidden layer model are compared. It is found that the result of the double hidden layer model is more accurate with higher generalization capability, which indicates the advantages of deep learning in extracting key characteristics compared to shallow learning. The results of data preprocessing reveals that the concentration of Chlorophyll a has significant correlation with dissolved oxygen, pH, turbidity and ammonia-nitrogen, while it shows no significant correlation with surface temperature, salinity, nitrous nitrogen and phosphate. Meanwhile, the data preprocessing plays a positive role in the improvement of data quality, data mining efficiency and prediction accuracy.

Key words: artificial neural network; intelligent prediction; deep learning; Chlorophyll a; data preprocessing