DOI:10.11737/j.issn.1003-0239.2025.05.012

机器学习算法在有害藻华早期预警模型的应用进展

丁文祥[1,2]，林晨旭[1]，张彩云[1*]

（1.厦门大学海洋与地球学院 水声通信与海洋信息技术教育部重点实验室，福建 厦门 361102；2.浙江海洋大学海洋科学与技术学院，浙江 舟山 316022）

摘 要： 综述了经典机器学习算法、人工神经网络和深度学习算法在有害藻华预警模型中的应用进展；并针对模型在数据稀缺、泛化能力不足以及精度提升方面的具体挑战，详细探讨了多源数据的应用、模型结构和参数优化、以及集合预报等策略对有害藻华早期预警模型准确性提升的作用。

关键词： 机器学习；人工神经网络；深度学习；早期预警；有害藻华

中图分类号：X55 **文献标识码：**A **文章编号：**1003-0239(2025)05-0120-14

0 引言

有害藻华，又称赤潮，是在特定环境条件下，海水中某些浮游植物、原生动物或细菌爆发性增殖或高度聚集引发的有害生态现象[1]。近年来，随着气候变化以及沿海环境污染的加剧，近海赤潮的发生频率和影响范围明显增加，严重威胁渔业生产、滨海旅游和公众健康，尤其是有毒藻类引发的赤潮危害最为严重[2-3]。例如，2015年春季，北美西海岸发生的有毒拟菱形藻（*Pseudo-nitzschia*）藻华，导致经济性海产品行业关闭，造成巨大经济损失[4]。据统计，过去30年间，中国近海因有害藻华造成的经济损失高达59亿元[5]。因此，准确预测有害藻华爆发显得尤为重要。构建有害藻华早期预警模型不仅有助于有害藻华的治理防控和风险管理，还能减轻潜在损失，具有重要的现实意义。

有害藻华预警模型主要包括数据驱动模型和机理模型。数据驱动模型通过分析和学习大量历史数据，利用统计方法或机器学习算法建立预警模型。这类模型无需依赖物理或生态系统的基本理论，而是从数据中提取规律，在缺乏明确机理理解的情况下，可借助大数据和复杂算法实现高效的预警。常见的数据驱动方法包括简单的阈值法、线性和非线性统计模型，以及更加复杂的贝叶斯统计方法和人工神经网络模型等[6-8]。而机理模型通常通过构建物理-生态耦合模型，从物理动力学和生态动力学角度来探索藻华的过程和机制，以及影响有害藻华种群动态的关键因素[9-10]。例如，区域海洋生态耦合模式ROMS-CoSiNE基于物理-生物学耦合机制，通过差分法求解数值模型中的偏微分方程，定量模拟生态系统各组成部分及其影响因素的变化[11]，进而为生态灾害预警提供理论支持。机理模型往往不需要依赖大量历史数据的统计分析，而是通过求解偏微分方程模拟系统的动态演化，其结果也具有更强的解释能力。

机理模型的复杂性和计算需求通常与模型维度有关。一维模型通常用于简化场景，例如模拟垂直方向的变化；二维模型考虑水平和垂直方向的交互作用，适用于较为复杂的区域生态模拟；三维模型则进一步纳入空间变化细节，包括水体流动和生态过程的空间分布，能提供更精确的生态预警，尤其适用于复杂海域的模拟。然而，由于海洋生态系统的复杂性，物理-生态模型往往需要高精度数据、复杂参数以及特定的初始和边界条件，因此目前仍

收稿日期：2024-07-29。
基金项目：国家自然科学基金联合基金项目（U22A20579）；广西重点研发计划（桂科AB25069453）；福建省促进海洋与渔业产业高质量发展专项（FJHYF-L-2025-07-009）；福建省海洋与渔业综合服务专项（FYZF-YJYB-2025-1-2）。
作者简介：丁文祥（1991-），男，讲师，博士，主要从事机器学习海洋参量预报、赤潮早期预警技术研究。E-mail:wxding@zjou.edu.cn
*通信作者：张彩云（1972-），女，副教授，博士，主要从事海洋遥感应用、赤潮早期预警技术研究。E-mail:cyzhang@xmu.edu.cn

中国知网 https://www.cnki.net

难以对模型中影响生物生长的变量进行准确、定量的描述[12-13]。

随着计算机技术、人工智能和生物技术的迅猛发展,机器学习算法作为数据驱动模型之一,在有害藻华预警预测研究中得到广泛应用[14-15]。机器学习模型模拟人类的某些学习过程,发展出各类学习理论和方法,在解决潜在物理和生物关系不明的多元和非线性问题上具有独特优势[16-17]。众所周知,有害藻华的发生是气候、生物、物理和化学等因素综合作用的结果[18],具有突发性和非线性特点[2, 19]。此外,导致近海海域有害藻华发生的因素及藻种相当复杂,不同区域有害藻华的原因种、主要影响因素、变化规律等均存在差异[2]。这些复杂影响因素限制了传统预报技术在近海海域有害藻华预警预测中的发展和应用。利用机器学习方法等人工智能技术,可以有效处理变量间复杂的交互作用和非线性关系。机器学习算法构建的"黑匣子"模型,可解决对未知物理、化学和生物过程构建复杂数学方程的需求[20],从而显著提升近海有害藻华信息提取和预警能力。因此,机器学习模型已成为目前有害藻华预警技术的重要研究方向。

尽管机器学习模型在有害藻华预警中取得显著进展,但仍面临数据稀缺、模型解释性不足及泛化能力较差等挑战[21]。因此,本文旨在通过综述有害藻华机器学习预警模型的研究进展,探讨其在实际应用中的优势与局限性,并提出可能的改进方向和未来研究重点,为有害藻华的早期预警与防控提供理论基础和技术支持,进而有效减少有害藻华对生态环境和人类社会的负面影响。

1　机器学习方法在有害藻华预警上的应用

1.1　经典机器学习算法

经典机器学习算法广泛应用于有害藻华预警领域,涵盖了从结构较简单的回归算法到较复杂的支持向量机(Support Vector Machine,SVM)和随机森林(Random Forest,RF)等多种方法。许多经典算法自20世纪90年代以来不断得到优化和应用。线性回归是最简单的基础算法,旨在寻找输入与输出之间的线性关系[22]。多项式回归和逐步回归是稍复杂的变种,主要用于捕捉数据中的非线性关系[23-24]。支持向量机和随机森林等非线性模型能更好地处理复杂的非线性关系,显著提高模型的预测精度,尤其在气候、海洋等领域的应用效果显著[25-26]。分类任务是机器学习中的另一个主要任务,支持向量机和随机森林既可以用于回归分析,也可以应用于分类任务,展现出处理多种类型问题的灵活性。在有害藻华预警中,分类任务常用来识别藻华是否发生,或对藻华级别进行分级预警。

SVM是一种基于统计学习理论的监督学习算法,主要用于分类和回归任务。其基本原理是通过寻找最优超平面分离将不同类别的样本数据,并使两类数据点到超平面的距离最大化,从而提高分类泛化能力。核心算法包括通过优化问题求解最优超平面,利用核函数(如线性核、径向基核等)将数据映射到高维空间,使低维空间中非线性可分的数据在高维空间实现线性可分。SVM泛化能力较强,适用于高维空间,能够有效处理复杂和非线性问题,且不易陷入局部最优解。但SVM对内存和计算时间要求较高,尤其在大规模数据集上,训练时间可能较长。此外,SVM对核函数选择敏感,不同核函数可能显著影响模型性能,调参过程较为复杂。SVM在水生态系统中的应用,尤其是有害藻华预警方面,表现出优越性能。例如,Miura等[6]利用过去7天的营养盐、径流量和气象数据,基于SVM模型成功预警日本4个大坝水库未来7天内微囊藻(*Microcystis spp.*)和蓝绿藻(*Dolichospermum spp.*)的浓度,并基于浓度阈值预测藻华发生,准确率分别达92.3%和71.4%。相关研究表明,SVM结合叶绿素a浓度、海岸环境条件、物理化学水质数据以及水动力和气象数据时,预测准确性高于人工神经网络[27-28]。此外,一些改进方法如Su等[14]在基于SVM模型开展北京密云水库月平均叶绿素浓度预报研究时,采用遗传算法特征选择提取叶绿素最相关的4个影响因子(总磷、总氮、高锰酸盐指数和水库库容),简化了模型结构,使其在环境管理中更加实用和高效。

RF是一种基于决策树的集成学习算法,通过构建多棵决策树并对其结果进行投票或平均以提高预测精度。其核心算法是通过随机选择样本和特征生成多个决策树,最终整合各树输出结果。随机

中国知网　https://www.cnki.net

森林的优势在于能够有效避免过拟合，提高模型鲁棒性，且能够处理高维数据和缺失数据。但模型训练时间较长，且因多棵决策树组合导致结果不易直观理解，可解释性较差。在多个生态问题中，RF表现出较好的预警效果，尤其在处理大规模数据时，能有效提升模型鲁棒性[8,29]。

为了进一步提升经典机器学习算法的预警性能，研究者们探索了经典机器学习算法与其他算法的结合。例如，Hill等[30]将SVM、RF、多层感知机（Multilayer Perceptron，MLP）与深度学习网络相结合构建了HABNet藻华检测系统（见图1），充分发挥不同机器学习模型的优势，利用不同中分辨率成像光谱仪（Moderate-Resolution Imaging Spectroradiometer，MODIS）数据参量和地形数据，采用分类预报方式（直接判断藻华是否发生）成功实现对弗罗里达州沿岸短凯轮藻（*K. brevis*）的检测。该方法最大检测准确率达91%，为实际应用提供了更便捷高效的预警工具。

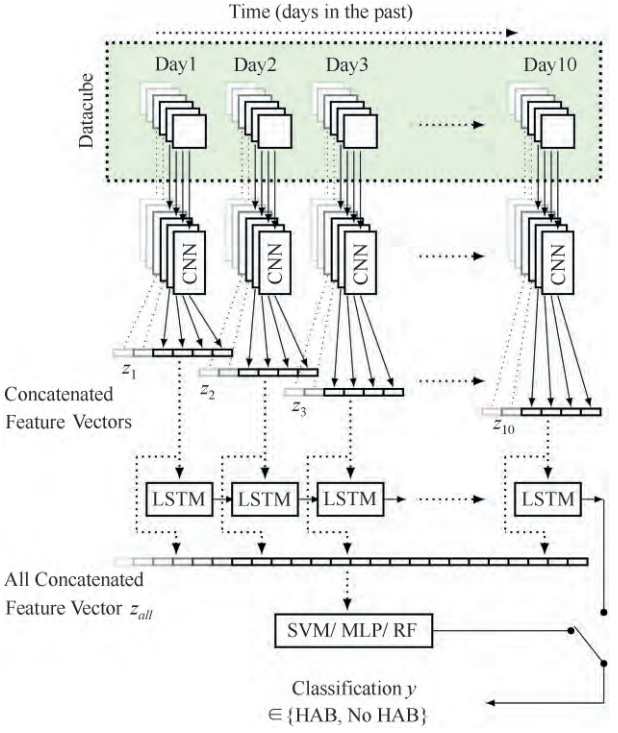

图1 HABNet机器学习藻华检测系统的结构示意图[32]

Fig.1 Structure of HABNet machine learning system[32]

1.2 人工神经网络模型

人工神经网络是机器学习的重要组成部分，其研究正迅速发展[31]。人工神经网络模仿生物神经系统的结构和功能，从信息处理角度对人脑神经网络进行抽象建模[32-33]。它通过不同连接方式的组合，形成多层次网络结构，模拟大脑神经网络处理、学习和记忆信息的过程[32-33]。它采用与传统人工智能和信息处理技术完全不同的机制，克服了基于逻辑的人工智能在处理直观和非结构化信息方面的缺陷，具备自适应、自组织和实时学习的优点[31,34]。人工神经网络已广泛应用于模式识别、智能机器人、自动控制和环境参量预报等领域，解决了现代计算机无法解决的实际问题，显示出良好的优势[35]。

人工神经网络的基本单元是神经元，它模仿生物神经元的输入输出过程[35-36]（见图2）。神经元通过加权和函数接收输入，并通过激活函数生成输出信号。常见的神经网络结构包括前馈神经网络（Feedforward Neural Network，FNN）和反向传播神经网络（Backpropagation Neural Network，BP），其中，BP神经网络通过反向传播算法优化权重参数，逐步降低预报误差，广泛用于回归和分类问题。另一类重要的神经网络是径向基函数神经网络（Radial Basis Function Network，RBF），其通过径向基函数作为激活函数，有效解决了高维数据下的分类与回归问题。

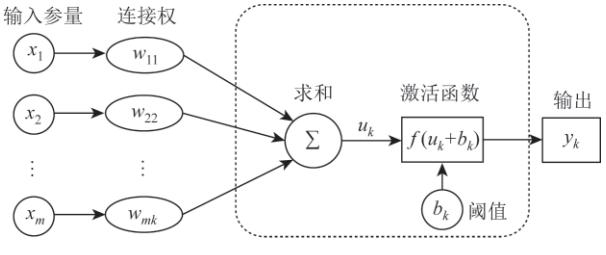

图2 人工神经元结构图

Fig.2 Artificial neuron structure diagram

早在20世纪90年代，人工神经网络开始应用于赤潮相关研究，并取得显著成果。Recknagel[36]采用人工神经网络成功预测了蓝绿藻丰度，这一成果推动了人工神经网络在水质监测和藻类预警领域的应用。杨建强等[37]采用BP和RBF人工神经网络开展大亚湾拟菱形藻和辽东湾丹麦细柱藻（*Leptocylindrus danicus*）密度预报，验证了人工神经网络在模拟和预测方面相较于传统统计回归模型的优势，最优预报结果的相对误差从多项式回归模

中国知网 https://www.cnki.net

型的23.86%降至BP模型的7.02%和RBF模型的1.7%。Deng等[28]在香港吐露港海域的研究结果证实,人工神经网络和SVM均具有较强的适应性,能有效学习海岸环境变量与藻类动态间的复杂关系,进而实现对藻类生长趋势和规模的准确预测。人工神经网络的研究结果也可以指导对机理的认识,如许阳春等[38]利用BP神经网络研究叶绿素预报模型输入参量的最优组合,结果表明以气温、溶解氧浓度、日照时长为输入参量时,BP模型误差最小,为平潭海域以叶绿素浓度作为判定指标的赤潮预警研究提供重要参考。

人工神经网络的主要优势在于强大的学习能力,尤其在面对复杂和非线性关系时,能够捕捉传统统计方法难以识别的模式。此外,神经网络在大数据环境下表现出良好的鲁棒性和自适应能力。然而,人工神经网络的缺点也不容忽视,主要体现在训练过程的计算复杂度较高、存在过拟合问题且对大量标注数据具有依赖性。尤其在缺乏足够数据支持时,网络性能可能受到限制。

1.3　深度学习模型

深度学习是机器学习的重要分支,源自人工神经网络并在其基础上发展而来[39]。它通过模拟人脑神经元的结构和工作原理,构建具有多层次结构的神经网络模型。每一层网络都可以对输入数据进行逐层抽象和特征提取,逐渐将低级的原始特征转化为更复杂的高级特征,从而捕捉数据中的深层次模式和关系[40-41]。深度学习的优势在于其强大的自动特征学习能力,无需人工设计特征,可从大量数据中自动发现潜在规律,特别适合处理复杂且高维的数据集[41-42]。它在图像识别、语音识别、自然语言处理等多个领域取得突破性进展,并在人工智能应用中发挥日益重要的作用[40-41]。深度学习的核心特点是通过多层次神经网络结构处理和表示数据,使其在解决非线性、复杂问题时具有显著优势[43-45]。卷积神经网络(Convolutional Neural Network,CNN)和长短时记忆网络(Long - Short Term Memory,LSTM)是两种广泛应用的深度学习模型,二者各具特点,在多个领域中都表现出优异的应用效果[30,43,45]。

CNN是一种深度学习模型,特别适用于处理图像、语音、视频等具有网格结构的数据。CNN的设计灵感源自生物神经系统的结构,尤其是视觉皮层的处理机制,能够通过卷积操作自动提取输入数据的局部特征[46]。与传统全连接神经网络不同,CNN通过局部连接、共享权重和池化操作,显著减少参数数量,提高模型计算效率。CNN的基本结构包括卷积层、池化层和全连接层(见图3)。卷积层通过卷积操作提取输入数据的局部特征,卷积核(滤波器)在输入数据上滑动,并通过加权和生成特征图。池化层通常用于下采样,减少特征图的空间维度,以降低计算复杂度,同时保留重要特征。常见的池

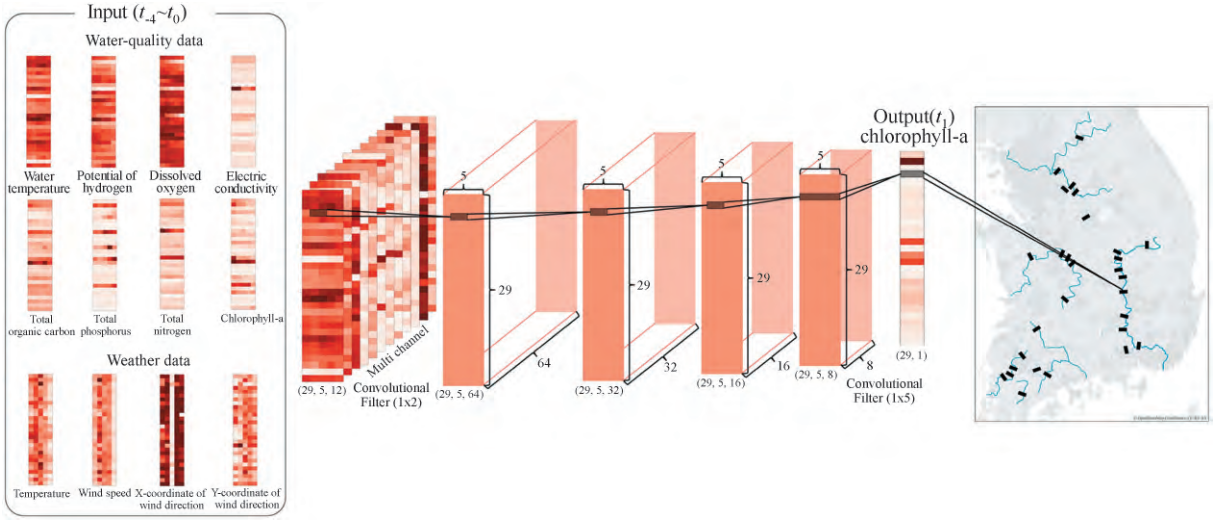

图3　基于CNN的叶绿素综合预报模型结构[49]

Fig.3　Architecture of the chlorophyll-a integrated prediction model[49]

中国知网　https://www.cnki.net

化方法包括最大池化和平均池化。全连接层位于网络末端,通常用于将特征图映射到最终的分类结果或回归输出。CNN的优势在于强大的特征提取能力,能够自动学习数据中的重要模式[47-48],避免了人工特征提取的困难和局限性。通过多层卷积和池化操作,CNN逐步提取从低级到高级的特征,形成对输入数据的深度理解,这使其在图像分类、物体识别、语音识别、自然语言处理等领域取得显著成果。

LSTM是一种特殊的递归神经网络(Recurrent Neural Network, RNN)结构,专门用于解决传统RNN在处理长序列数据时面临的梯度消失和梯度爆炸问题[50]。LSTM通过引入门控机制,能够有效捕捉和记住长期依赖关系,这使其在时间序列预测、自然语言处理、时间序列分析等任务中表现优异。LSTM的核心组件是其内部的记忆单元,它包含3个主要的"门"——输入门、遗忘门和输出门(见图4)。每个门均为神经网络层,通过决定保留和丢弃的信息来管理记忆单元的状态:输入门控制哪些新信息应该被存入记忆单元;遗忘门控制哪些旧信息应从记忆中删除,输出门决定当前记忆单元的输出值。通过这些门控机制,LSTM在处理长时间跨度序列时,能有效保留重要信息并抑制无关内容,从而解决传统RNN在长序列学习中出现的"记忆衰减"问题[50-51]。

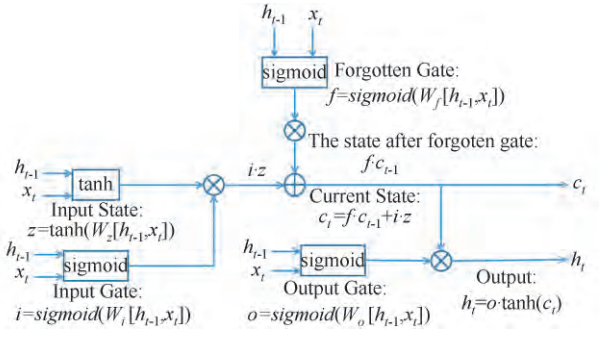

图4 LSTM单元示意图[52]

Fig.4 Schematic diagram of the LSTM unit[52]

随着深度学习技术的不断发展,CNN和LSTM在有害藻华相关研究中得到了广泛应用[7, 43, 52]。Lee等[41]对比了MLP、RNN、LSTM与传统回归模型在有害藻华预警中的潜力,模型以水文、水质和生态参量为输入,通过预报叶绿素浓度实现有害藻华预报;16个监测站的测试结果均显示LSTM模型的预测准确率最高,叶绿素预报的平均均方根误差从回归模型的17.75 μg/L下降到16.09 μg/L,揭示了LSTM和深度学习在有害藻华预测中的潜力。Baek等[7]尝试结合CNN模型的回归和分类预报能力,采用CNN分类模型预报韩国巨济岛链状亚历山大藻(*Alexandrium catenella*)藻华的发生情况,采用CNN回归模型预报其藻华生物密度,预报精度分别达到96.8%和1.20 log(cells/L);同时基于深度学习模型探讨藻华的可能影响机制,发现盐度和温度对藻华爆发贡献较大,而NH_4N会影响藻华生长。

由于LSTM模型在处理时间序列数据时具有结构设计上的独特优势,很多研究发现其在赤潮时间序列预报中明显优于其他机器学习模型[43, 45],例如,Yussof等[45]在沙巴西海岸开展遥感叶绿素时间序列预报研究时发现,LSTM模型的预报效果明显优于CNN模型,RMSE从4.4 mg/L下降至3.4 mg/L。将CNN模型与LSTM模型相融合也是赤潮预报的常用方法[44, 53],例如,Ding等[52-53]以卫星监测的气象、水文和叶绿素数据为输入,通过预报叶绿素浓度的时空分布开展舟山渔场藻华预报研究,结果显示采用CNN-LSTM混合模型的预报效果优于单一的CNN或LSTM模型,预报结果的决定系数由0.31和0.40提升至0.43。无论是将深度学习模型与传统回归模型对比、不同深度学习模型间比较,还是通过融合多模型构建混合模型,核心目的都是寻求更精确的预报结果。在这一过程中,机器学习模型的优化尤为重要,其性能的提升直接关系到预测结果的准确性与可靠性。优化方法通常包括调整模型结构、优化算法、特征选择与数据预处理等,旨在提高模型的泛化能力和处理复杂非线性关系的能力。因此,深入研究和探索机器学习模型优化技术,是实现更精确、有效预测的关键步骤,也是推动相关领域应用研究的重要基础。

2 有害藻华机器学习预警模型的优化

2.1 多源数据应用与处理

机器学习模型通过学习历史事件掌握事件发生规律,并利用最新数据预报未来事件,其本质是

中国知网 https://www.cnki.net

统计模型,因此对数据源的要求极高。然而,在有害藻华机器学习预警模型的研究中,数据源往往受限,只能依赖有限的观测数据和参量探索最佳预报结果。随着观测技术的提升和观测数据的积累,更长的时间序列数据和更完善的相关参量将会进一步推动机器学习模型在有害藻华预警方面的应用。

在赤潮预警中,通常采用赤潮优势种浓度或相应的表征因子来评估赤潮的发生及危害[54]。赤潮的发生通常是依据优势种浓度是否超过阈值来认定,因此,直接预报优势种的浓度被认为是最直接有效的预警方式。然而,赤潮的优势种鉴定过程相对复杂,往往受到采样频率和采样时空分布的限制,难以获取高频次、长时段的监测数据。对于机器学习模型而言,数据量不足成为显著挑战,因为机器学习模型的有效性和准确性往往依赖大量数据进行训练与验证。缺乏足够的数据会导致模型训练不充分,影响预测精度和可靠性,从而限制了基于机器学习的赤潮预警系统的实际应用效果。采用赤潮的表征因子进行赤潮预报是当前研究中的一种常用方法。叶绿素浓度是反映藻类生物量的重要指标,其浓度快速增加与藻华爆发起始时间一致[55]。因此,叶绿素浓度被广泛应用于赤潮预警模型中,因为它能够提供及时的藻类繁殖信息,辅助预警赤潮发生。许多研究表明,通过对叶绿素浓度的监测与预测,能够有效预警赤潮的发生,进而减少赤潮带来的生态和经济损失[45, 53, 56]。

固定浮标或定点采样数据能够提供机器学习所需的长时间序列数据,这使其成为许多赤潮预警研究中的重要数据来源。许多研究利用定点浮标监测数据或定时采样得到的时间序列数据,开展机器学习模型的训练和预报,并取得显著效果[57-59]。然而,这些数据的获取往往需要较高的成本,并且监测站点地理位置固定,这就限制了其空间监测范围。而赤潮的发生位置通常具有高度的不确定性,可能随机发生,并随着时间的推移而扩散或偏移。因此,尽管固定浮标和定点采样能为赤潮预警提供重要数据支持,但其空间覆盖的局限性意味着需要结合其他监测手段和数据源,以提高预警的时效性和准确性。

遥感数据凭借高空间覆盖率优势,能够有效弥补固定浮标或定点采样数据在空间范围上的局限性。随着航天技术的迅速发展,海洋卫星及其探测技术不断取得进展,观测精度也在持续提高[60-61],这使得遥感数据在赤潮预警研究中成为重要数据来源。例如Song等[61]为解决缺乏足够现场数据的问题,利用MODIS和中等分辨率成像光谱仪(Medium Resolution Imaging Spectrometer, MERIS)卫星数据建立了蒙特利湾随机森林赤潮模型,取得了较好的预报结果,并通过了现场数据的验证。然而,尽管遥感数据具备空间覆盖优势,但是其采样频率较低的问题限制了其在赤潮这种快速变化的生态现象中的动态监测能力;同时,遥感数据所提供的监测参数相对较少,这使得其在深入挖掘赤潮发生机制和规律上的应用面临一定的挑战。

随着技术的进步,海洋数值模型获得长足发展,可实现对近海水文参量更精细化的描述[62-63],各种海洋生态耦合模型也得到了广泛应用[64-66],这使得数值模型数据成为近海有害藻华机器学习预警模型数据源的潜力显著提升。将数值模型数据与遥感数据结合开展有害藻华机器学习预警模型研究也是目前常用的方法之一。例如,Jin等[56]将动力模型数据与高频的静止轨道和海洋色成像仪(Geostationary and Ocean Color Imager, GOCI)遥感数据共同作为深度学习模型的训练数据,以遥感叶绿素浓度为模型输出实现对叶绿素的空间分布预报,显著提高了深度学习模型的时空预报效能。赤潮的发生和发展是复杂的生物过程与化学过程交织作用的结果,尽管现代海洋数值模型已能模拟一些主要过程,但对许多生物地球化学过程的精确刻画仍然有限,特别是在处理微观尺度的藻类生长动力学、营养物质循环和气候变化的交互影响时,这些过程的复杂性和时空变异性使得模型的参数化存在较大的不确定性。因此,这些参数化的不确定性可能导致模型预报结果出现偏差,从而影响赤潮预警的准确性和可靠性。

除了数据源问题,数据不平衡也是机器学习赤潮预警模型面临的挑战之一。赤潮发生的时间远少于不发生的时间,导致监测数据中为常规时段数据为主,赤潮时段数据较少,形成不平衡的数据集。这种数据不平衡使得模型可准确预报常规事件,但在赤潮事件的预警上可能精度不足。由于高浓度赤潮事件比低浓度事件更关键,因此增加赤潮期间数据的比例,以解决数据不平衡问题,是提高模型

中国知网　https://www.cnki.net

准确性和实用性的关键。针对这一问题，已有一些研究尝试通过不同方法开展优化。苏新红等[67]将赤潮发生前后的一段时间气象监测数据作为一个赤潮样本，利用收集的219个赤潮样本作为BP网络的训练数据来预报赤潮危害等级，大幅提高了赤潮期间数据的占比；Kim等[58]在预报依据藻类细胞浓度划分的赤潮危害等级时，采用自适应合成采样方法生成合成数据，解决了原始数据中高细胞浓度等级数据失衡的问题，提高了机器学习模型的预测性能。

2.2 模型结构和参数的调整

机器学习模型在众多领域展现了巨大的应用价值，但其应用仍有待完善。在某些场景下，机器学习模型的预报精度仍有提升空间[68-70]。因此，有必要针对不同需求开发和优化机器学习模型，以实现更小误差、更短训练时间、更高精度的最佳模型配置。对机器学习模型的优化涉及多个方面，主要包括对模型结构和超参数、输入、输出和训练过程的优化。

传统上，最优的机器学习模型是通过穷举式试错调整其结构和超参数获得的[71-72]，但这种方法效率低下，有时甚至不可行，且忽略了超参数之间的交互影响，因此无法保证获得最优结果[73]。寻优算法能够有效提高模型优化效率，避免传统穷举试错法的低效性，同时考虑超参数间的交互影响，从而更有可能获得最优模型性能。因此，有害藻华机器学习模型通常结合各类寻优算法实现对模型超参数的优化。例如，研究人员采用粒子群算法（Particle Swarm Optimization，PSO）筛选SVM的超参数，提升了模型对螺旋藻（*Spirulina platensis*）的预报性能[74]；他们还采用改进的基于线性种群规模缩减的差分进化算法（Linear Population Size Reduction Differential Evolution Algorithm，L-SHADE）优化梯度增强回归树模型（Gradient Boosting Regression Tree，GBRT）的超参数，实现了对西班牙水库叶绿素异常增殖过程的预测[75]。贝叶斯优化算法[76-77]、遗传算法[78-79]等也是常用的超参数优化算法，在其他领域取得了显著成果，但在有害藻华预警模型超参数优化方面仍需更多案例验证。寻优算法也存在不足，例如优化过程中可能需要大量的计算资源，尤其是当超参数空间较大时，计算开销会显著增加；此外，部分寻优算法可能面临收敛速度较慢的问题，特别是处理复杂模型或高维度问题时，可能需要较长时间才能获得理想的结果。

机器学习模型输入的优化也是改进模型性能的常见方式，主要包括对模型输入参量的筛选和输入数据结构的优化。影响藻华或叶绿素的因素通常复杂多样[2, 80]，由多参量组成的高维特征向量时间序列中常隐藏不相关和冗余信息[81]，导致模型结构复杂，从而降低机器学习模型的分析精度和应用效率[14]。为避免冗余信息干扰，采用最相关的影响因子作为输入向量可获得更准确可靠的预测结果。比较容易理解的方式是从数据分析角度实现对输入参量的筛选，如Shin等[82]采用正向选择法分析不同参量对叶绿素预报效果的影响，剔除无关参量，以减少冗余信息对机器学习模型的干扰。主成分分析是回归模型和机器学习模型输入参量特征提取和降维的常用方法[83-84]，通常选取前几个主成分作为模型输入，实现输入参量降维。遗传算法是另一种更智能的降维方法，通过筛选模型输入参量的最优组合提升模型预报精度[85]。例如，Su等[14]在叶绿素影响因子分析结果的基础上，采用遗传算法进行输入参量特征选择，显著提高了模型对北京密云水库叶绿素浓度的预报精度和效率。输入参量的输入方式也会对模型预报结果产生重大影响。例如，Shin等[82]以业务部门监测的水文、水质和气象数据为基础，研究韩国洛东江监测站点叶绿素预报模型时发现，将RNN模型与滚动窗口输入方法相结合预测叶绿素浓度的效果最优，并发现在机器学习模型中提前一步递归预测是提升模型预测性能的重要过程。对模型输入的优化虽然能提高模型预测精度，并在一定程度上增强模型的可解释性，但也存在局限性：首先，输入参数筛选和降维过程中可能丢失部分重要信息，特别是在处理复杂的非线性关系时，过度简化可能会影响模型表现；其次，特征选择和降维方法依赖现有数据分析技术，可能导致忽略未知因素或对某些潜在变量作出错误假设，进而影响模型泛化能力。

从模型输出角度进行优化，通常需要结合对赤潮机理的深入理解。多数叶绿素或浮游生物机器学习预报模型直接预报其浓度[82, 86]，假设机器学习方法可以在不需要人工辅助的情况下学习浮游生物变化的动态机理，进而给出准确的预测结果。然

中国知网 https://www.cnki.net

而,有研究指出,浮游生物的影响因素首先作用于其变化过程,并且需要一定时间才会显著影响其浓度[2,68]。此外,研究表明,在藻华动态研究中,生物量相对变化率的重要性高于绝对浓度[87-88],这一认识推动了一些新优化方法的提出,以改进有害藻华机器学习模型的预测输出。例如,Tian等[68]以华东地区水库出水口的定时监测水文和水质数据为基础开展叶绿素预报研究时发现,以叶绿素变化率作为人工神经网络模型的输出,效果优于直接预报叶绿素浓度,预报结果与观测结果的相关系数可以从0.75提升到0.83;Ding等[52]基于定点浮标监测数据进行厦门湾叶绿素预报研究时发现,以叶绿素相对变化率作为深度学习模型的输出,预报结果的均方根误差更低、相关系数更高。这种优化方式也存在局限性,如过度依赖对赤潮机理的深入理解,而目前这种理解仍较为有限,尤其在复杂的海洋环境中,许多影响因素间的相互作用尚未完全明晰。

对模型训练过程的优化主要包括初始权值和阈值的优化、泛化能力的提升以及集成学习的应用。机器学习模型在初始训练时,其权值和阈值通常随机设定,但这种随机性可能会增加模型的不稳定性。因此,一些研究常采用遗传算法优化机器学习模型的初始权值和阈值[89-91]。例如,向先全等[91]利用遗传算法优化BP模型初始权值和阈值,以提高对渤海湾叶绿素的预测精度。在野外环境中,水体的营养条件和藻类动态具有多变性和不稳定性[92-93],不同时期驱动叶绿素变化的主要机制可能存在差异,这导致有害藻华机器学习模型常面临泛化能力不足的问题。Tian等[69]采用迁移学习方法增强深度学习叶绿素预报模型的泛化能力,有效解决了模型在长期应用中性能随时间推移而下降的缺陷。随机失活(Dropout)是一种常用的正则化技术,通过在训练过程中随机丢弃部分神经元来防止模型过拟合,从而提升模型的泛化能力[51,94],目前已广泛应用于深度学习叶绿素预报模型的优化[69,95]。集成学习是通过结合多个弱学习器构建强学习器的方法,旨在提高模型的准确性和鲁棒性。如Shin等[96]以韩国三大水库为研究对象,探讨数据采样不平衡对机器学习藻华分类预报模型的影响时发现,采用集成学习优化方法能有效解决训练数据采样不平衡问题,使预报精度提升2.12%。

机器学习赤潮预警模型的常见优化方法见表1。

2.3　集合预报

越来越多的研究开始采用机器学习方法预测有害藻华的发生。当单一机器学习算法难以实现准确预报时,不少研究开始探索多机器学习算法的结合方法。不同预报方法的有机结合或集合预报,是目前有害藻华机器学习预报中较为常用的的优化方法。美国加利福尼亚州的C-HARM(California Harmful Algae Risk Mapping)系统利用独特的混合数值模型、生态预警模型和卫星水色影像,预测拟菱形藻(*Pseudo-nitzschia*)赤潮发生的可能性和毒素等级[97]。欧洲的ASIMUTH(Applied Simulations and Integrated Modelling for the Understanding of Toxic and Harmful Algal Blooms)预警系统则通过卫星观测产品与现场监测、区域模型的结合,实现对有害藻华的短期预警[2]。Qin等[17]提出一种结合自回归集成滑动平均(Autoregressive Integrated Moving Average, ARIMA)模型和深度信念网络(Deep

表1　机器学习赤潮预警模型的常见优化方法

Tab.1　Common optimization methods for machine learning-based algal bloom early warning models

优化角度	优化方法	文献
结构和超参数	结合寻优算法:粒子群算法、差分进化算法、遗传算法等	[71]、[75]
输入	输入参量的筛选和输入数据结构的优化:遗传算法、主成分分析、滚动窗口输入法等	[14]、[82]
输出	结合对赤潮机理的理解:叶绿素变化率、叶绿素相对变化率	[3]、[68]
训练过程	初始权值和阈值的优化:遗传算法;泛化能力的提升:迁移学习、随机失活;集成学习	[69]、[91]、[95]、[96]

中国知网　https://www.cnki.net

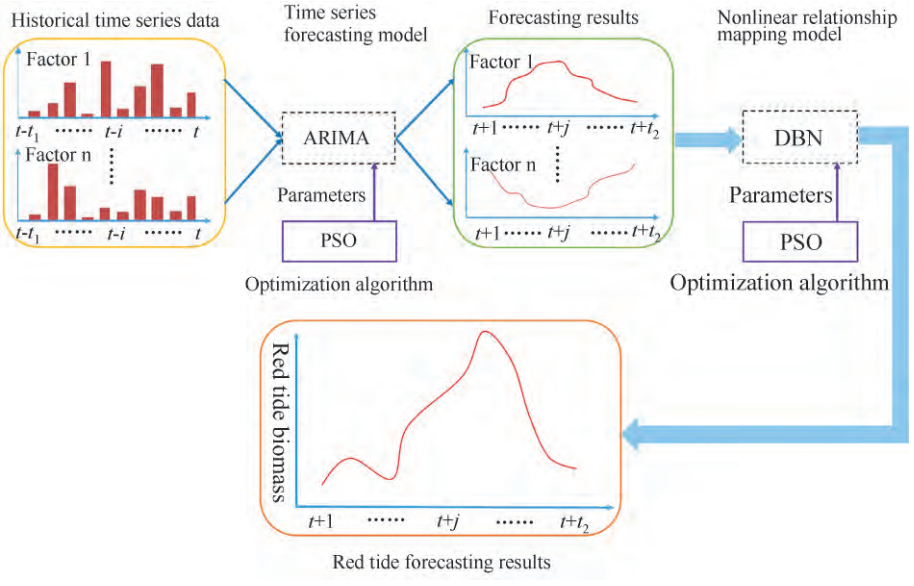

图5 ARIMA-DBN赤潮预警模型结构示意图[17]

Fig.5 The framework of ARIMA-DBN[17]

Belief Net,DBN)的混合赤潮预警模型(见图5),该模型充分利用ARIMA在捕捉时间相关性和空间异质性方面的优势,同时结合DBN对环境因子与赤潮生物量之间复杂非线性关系的强大表达能力。该模型以船舶采样监测的水文和生态数据为基础,针对不同沿海地区的环境因子构建相应的ARIMA模型,而DBN则用于捕捉环境因子与赤潮生物量间的非线性关系。通过舟山和温州海域的测试结果表明,单独的BP模型和DBN模型预报结果的相关系数分别为0.598和0.654,ARIMA-BP模型的相关系数为0.716,而ARIMA-DBN模型的相关系数达0.798,可见ARIMA与DBN结合的预报效果明显优于其他模型。

福建近岸赤潮短期预警模型也是一种集成多种机器学习网络的集合预警模型,从2019年起在福建省海洋预报台实现业务化运行。研究发现,采用BP和RBF两种人工神经网络进行组合预报,可有效避免单一模型带来的偶然误差[98]。该模型基于生态浮标监测的水文、水质和气象参量,构建了赤潮发生概率等级的业务化预警系统。模型以赤潮发生前后15天的监测数据构成的赤潮样本为训练数据,并采用自组织映射(Self-Organizing Map,SOM)神经网络对赤潮样本进行严格筛选。模型融合BP和RBF两种人工神经网络,同时利用遗传算法优化模型的输入参量以及初始权值和阈值。为应对近岸环境的高动态变化,模型每日更新最新15天的监测数据重新训练,并采用提前一步递归预测方式进行预报,显著提升了模型的精度和泛化能力。模型充分利用生态浮标0.5 h高频采样的特点,构建多个预报因子,通过分类预报的方式每日生成数百个预报结果,每个结果直接判断赤潮是否发生。根据所有预报结果中判断赤潮发生的比例,确定赤潮发生概率等级。这种方法有效避免了因单一模型、单一数据组、单一预报结果等因素带来的偶然误差。该模型自2019年5月起在福建省海洋预报台投入业务化运行,2019—2021年24 h时效预报结果的赤潮识别率分别达到了60%、55%和60%。

由于近岸水体的营养条件和藻类动态具有高度不稳定,且不同区域存在明显差异,因此不同区域所采用的预警模型和优化方法各不相同。针对不同问题,采用不同优化方法对机器学习各过程进行优化,是当前研究的重要方向,也是提升机器学习模型在有害藻华预警中应用效果的重要途径。

3 结论

本文综述了机器学习算法在有害藻华早期预警模型中的应用进展,重点探讨了经典机器学习算

中国知网 https://www.cnki.net

法、人工神经网络及深度学习模型在该领域的应用情况。与传统方法相比,机器学习方法在处理复杂非线性关系和大数据分析方面具有显著优势。尽管如此,当前研究仍面临数据稀缺、泛化能力不足及精度有待提升等挑战。针对这些挑战,本文详细论述了多源数据的应用、模型结构的优化与参数调整、以及集合预报策略对提升模型准确性的作用。

本研究回顾了机器学习在赤潮预报中的巨大潜力,但要实现更精确全面的预报效果,仍面临多个挑战。未来的研究可以在以下几个层面进行突破和拓展:

①数据层面的提升

机器学习赤潮预报模型的效果在很大程度上取决于数据的质量和多样性。目前使用的数据源在时空分辨率、监测范围及参数种类上均存在一定局限性。为进一步提高预报精度,需加强赤潮的实时监测能力,尤其要增加对营养盐、浮游植物、浮游动物等关键生物及环境参数的监测。随着遥感技术和自动化监测技术的发展,我们有望获取更全面、高质量的数据,为机器学习模型提供更丰富的输入,助力模型更准确地捕捉赤潮发生的规律和趋势。因此,探索更优质的数据源和全面的数据积累,将是机器学习赤潮预报取得重大突破的关键。

②机制层面的深入研究

深入理解赤潮发生机制是提升机器学习模型预报能力的重要基础。赤潮的发生受多种自然和人为因素影响,包括水温、盐度、营养盐浓度、光照条件等环境因素,以及水体中的生物种群动态。当前赤潮数据收集往往侧重于某些表面特征,却忽视了深层次的机制信息。若收集的数据无法全面反映赤潮发生的核心驱动机制,机器学习模型的预测能力将受到限制。因此,深入研究赤潮的物理、化学和生物学机制,尤其是通过实验和观测数据揭示不同类型赤潮的发生机制,将为数据参量的选取和机器学习模型的优化提供重要指导,进而提升模型的精准性和泛化能力。同时,需开发具有更高可解释性和更强泛化能力的深度学习模型,探索如Transformer等先进模型在赤潮预报中的应用,以进一步提高深度学习模型对赤潮预报的准确性。

③实时更新模型和数据

随着沿海人类活动的日益频繁,近岸生态系统的变化对赤潮的发生和演变产生了重要影响。例如,工业污染、农业排放和气候变化等因素可能改变赤潮藻种的种群结构和水体环境条件,进而影响赤潮的发生和发展。为应对这些动态变化,机器学习模型需要不断更新和优化。这要求实时更新数据、模型和算法,以反映最新的环境变化和赤潮趋势。通过将实时监测数据与机器学习模型相结合,可有效调整模型参数,保证模型的适应性和时效性,从而维持或提升赤潮预报的准确性和可靠性。

综上所述,提升机器学习在赤潮早期预警中的应用效果,需要从数据、机制和模型更新3个层面开展持续研究和探索。只有在这些方面取得突破,才能实现更精准的赤潮预警和防控,减少赤潮造成的生态和经济损失,为沿海地区的可持续发展提供有力支持。此外,跨学科合作与多方资源整合,将是推动赤潮预警模型不断前进的重要动力。

参考文献:

[1] ANDERSON D M. Turning back the harmful red tide[J]. Nature, 1997, 388(6642): 513-514.

[2] ANDERSON C R, MOORE S K, TOMLINSON M C, et al. Living with harmful algal blooms in a changing world: strategies for modeling and mitigating their effects in coastal marine ecosystems [M]//SHRODER J F, ELLIS J T, SHERMAN D J. Coastal and Marine Hazards, Risks, and Disasters. Amsterdam: Elsevier, 2015: 495-561.

[3] DING W X, ZHANG C Y, SHANG S P. The early assessment of harmful algal bloom risk in the East China Sea[J]. Marine Pollution Bulletin, 2022, 178: 113567.

[4] MCCABE R M, HICKEY B M, KUDELA R M, et al. An unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions[J]. Geophysical Research Letters, 2016, 43(19): 10366-10376.

[5] YAN T, LI X D, TAN Z J, et al. Toxic effects, mechanisms, and ecological impacts of harmful algal blooms in China[J]. Harmful Algae, 2022, 111: 102148.

[6] MIURA Y, IMAMOTO H, ASADA Y, et al. Prediction of algal bloom using a combination of sparse modeling and a machine learning algorithm: automatic relevance determination and support vector machine[J]. Ecological Informatics, 2023, 78: 102337.

[7] BAEK S S, KWON Y S, PYO J, et al. Identification of influencing factors of A. catenella bloom using machine learning and numerical simulation[J]. Harmful Algae, 2021, 103: 102007.

[8] CRISCI C, GHATTAS B, PERERA G. A review of supervised machine learning algorithms and their applications to ecological

中国知网 https://www.cnki.net

data[J]. Ecological Modelling, 2012, 240: 113-122.

[9] 乔方利, 袁业立, 朱明远, 等. 长江口海域赤潮生态动力学模型及赤潮控制因子研究[J]. 海洋与湖沼, 2000, 31(1): 93-100.
QIAO F L, YUAN Y L, ZHU M Y, et al. Study on HAB dynamical model and HAB limitation factors for the sea area adjacent to Changjiang River estuary[J]. Oceanologia et Limnologia Sinica, 2000, 31(1): 93-100.

[10] MCGILLICUDDY D J JR, TOWNSEND D W, HE R, et al. Suppression of the 2010 Alexandrium fundyense bloom by changes in physical, biological, and chemical properties of the Gulf of Maine[J]. Limnology and Oceanography, 2011, 56(6): 2411-2426.

[11] ZHOU F, CHAI F, HUANG D J, et al. Coupling and decoupling of high biomass phytoplankton production and hypoxia in a highly dynamic coastal system: the Changjiang (Yangtze River) Estuary[J]. Frontiers in Marine Science, 2020, 7: 259.

[12] SELLNER K G, DOUCETTE G J, KIRKPATRICK G J. Harmful algal blooms: causes, impacts and detection[J]. Journal of Industrial Microbiology and Biotechnology, 2003, 30(7): 383-406.

[13] GLIBERT P M, SEITZINGER S, HEIL C A, et al. The role of eutrophication in the global proliferation of harmful algal blooms [J]. Oceanography, 2005, 18(2): 198-209.

[14] SU J Q, WANG X, ZHAO S Y, et al. A structurally simplified hybrid model of genetic algorithm and support vector machine for prediction of chlorophyll a in reservoirs[J]. Water, 2015, 7(4): 1610-1627.

[15] MOHEBZADEH H, LEE T. Spatial downscaling of MODIS Chlorophyll-a with machine learning techniques over the West Coast of the Yellow Sea in South Korea[J]. Journal of Oceanography, 2021, 77(1): 103-122.

[16] ZHENG W, SHI H H, SONG X K, et al. Simulation of phytoplankton biomass in Quanzhou Bay using a back propagation network model and sensitivity analysis for environmental variables[J]. Chinese Journal of Oceanology and Limnology, 2012, 30(5): 843-851.

[17] QIN M J, LI Z H, DU Z H. Red tide time series forecasting by combining ARIMA and deep belief network[J]. Knowledge-Based Systems, 2017, 125: 39-52.

[18] GUALLAR C, BACHER C, CHAPELLE A. Global and local factors driving the phenology of *Alexandrium minutum* (Halim) blooms and its toxicity[J]. Harmful Algae, 2017, 67: 44-60.

[19] SUGIHARA G, MAY R, YE H, et al. Detecting causality in complex ecosystems[J]. Science, 2012, 338(6106): 496-500.

[20] LARY D J, ALAVI A H, GANDOMI A H, et al. Machine learning in geosciences and remote sensing[J]. Geoscience Frontiers, 2016, 7(1): 3-10.

[21] PARK J, PATEL K, LEE W H. Recent advances in algal bloom detection and prediction technology using machine learning[J]. Science of the Total Environment, 2024, 938: 173546.

[22] KUMAR M, SRIVASTAVA V K. Pitman nearness and concentration probability comparisons of the sample coefficient of determination and its adjusted version in linear regression models[J]. Communications in Statistics-Theory and Methods, 2004, 33(7): 1629-1641.

[23] 江兴龙, 宋立荣. 泉州湾赤潮藻类优势种细胞密度回归方程研究[J]. 海洋与湖沼, 2010, 41(3): 341-347.
JIANG X L, SONG L R. Forecast equations for cell density of the dominant red-tide algae at the Quanzhou Bay[J]. Oceanologia et Limnologia Sinica, 2010, 41(3): 341-347.

[24] 吴玉芳. 赤潮高发期间厦门海域叶绿素值预报方程建立及应用于灾害性赤潮预报模式的研究[J]. 海洋预报, 2012, 29(2): 39-44.
WU Y F. Establishment of a chlorophyll forecast equation and its application in red tide forecasting in Xiamen offshore area[J]. Marine Forecasts, 2012, 29(2): 39-44.

[25] POPOV A, SAUTIN A. Selection of support vector machines parameters for regression using nested grids[C]//Proceeding of the Third International Forum on Strategic Technologies. Novosibirsk: IEEE, 2008: 329-331.

[26] BÉJAOUI B, ARMI Z, OTTAVIANI E, et al. Random Forest model and TRIX used in combination to assess and diagnose the trophic status of Bizerte Lagoon, southern Mediterranean[J]. Ecological Indicators, 2016, 71: 293-301.

[27] PARK Y, CHO K H, PARK J, et al. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea[J]. Science of the Total Environment, 2015, 502: 31-41.

[28] DENG T A, CHAU K W, DUAN H F. Machine learning based marine water quality prediction for coastal hydro-environment management[J]. Journal of Environmental Management, 2021, 284: 112051.

[29] SEGURA A M, PICCINI C, NOGUEIRA L, et al. Increased sampled volume improves Microcystis aeruginosa complex (MAC) colonies detection and prediction using Random Forests [J]. Ecological Indicators, 2017, 79: 347-354.

[30] HILL P R, KUMAR A, TEMIMI M, et al. HABNet: machine learning, remote sensing-based detection of harmful algal blooms [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3229-3239.

[31] WU Y C, FENG J W. Development and application of artificial neural network[J]. Wireless Personal Communications, 2018, 102 (2): 1645-1656.

[32] 朱大奇. 人工神经网络研究现状及其展望[J]. 江南大学学报(自然科学版), 2004, 3(1): 103-110.
ZHU D Q. The research progress and prospects of artificial neural networks[J]. Journal of Southern Yangtze University (Natural Science Edition), 2004, 3(1): 103-110.

中国知网 https://www.cnki.net

[33] HAN S H, KIM K W, KIM S, et al. Artificial neural network: understanding the basic concepts without mathematics[J]. Dementia and Neurocognitive Disorders, 2018, 17(3): 83-89.

[34] KASABOV N, SCOTT N M, TU E M, et al. Evolving spatio-temporal data machines based on the NeuCube neuromorphic framework: design methodology and selected applications[J]. Neural Networks, 2016, 78: 1-14.

[35] WU Y, WANG S. A new algorithm of improving the learning performance of neural network by feedback[J]. Journal of Computer Research and Development, 2004, 41(9): 1488-1492.

[36] RECKNAGEL F. ANNA-Artificial Neural Network model for predicting species abundance and succession of blue-green algae [J]. Hydrobiologia, 1997, 349(1-3): 47-57.

[37] 杨建强, 罗先香, 丁德文, 等. 赤潮预测的人工神经网络方法初步研究[J]. 海洋科学进展, 2003, 21(3): 318-324.
YANG J Q, LUO X X, DING D W, et al. A preliminary study on artificial neural network method for predicting red tide[J]. Advances in Marine Science, 2003, 21(3): 318-324.

[38] 许阳春, 张明峰, 苏玉萍, 等. 基于BP人工神经网络平潭海域赤潮叶绿素a浓度模型演算研究[J]. 海洋科学, 2020, 44(3): 34-41.
XU Y C, ZHANG M F, SU Y P, et al. Calculation of the Chlorophyll-a concentration of red tide in the Pingtan Coastal Zone by a BP artificial neural network model[J]. Marine Sciences, 2020, 44(3): 34-41.

[39] RAVÌ D, WONG C, DELIGIANNI F, et al. Deep learning for health informatics[J]. IEEE Journal of Biomedical and Health Informatics, 2017, 21(1): 4-21.

[40] MU R H, ZENG X Q. A review of deep learning research[J]. KSII Transactions on Internet and Information Systems, 2019, 13(4): 1738-1764

[41] DONG S, WANG P, ABBAS K. A survey on deep learning and its applications[J]. Computer Science Review, 2021, 40: 100379.

[42] MANUCHARYAN G E, SIEGELMAN L, KLEIN P. A deep learning approach to spatiotemporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence[J]. Journal of Advances in Modeling Earth Systems, 2021, 13(1): e2019MS001965.

[43] LEE S, LEE D. Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models[J]. International Journal of Environmental Research and Public Health, 2018, 15(7): 1322.

[44] 余璇, 石绥祥, 徐凌宇, 等. 基于深度学习的赤潮发生预报方法研究[J]. 海洋通报, 2021, 40(5): 566-577.
YU X, SHI S X, XU L Y, et al. Research on red tide occurrence forecast based on deep learning[J]. Marine Science Bulletin, 2021, 40(5): 566-577.

[45] YUSSOF F N, MAAN N, REBA M N M. LSTM networks to improve the prediction of harmful algal blooms in the West Coast of Sabah[J]. International Journal of Environmental Research and Public Health, 2021, 18(14): 7650.

[46] YAO G L, LEI T, ZHONG J D. A review of Convolutional-Neural-Network-based action recognition[J]. Pattern Recognition Letters, 2019, 118: 14-22.

[47] LEI X Y, PAN H G, HUANG X D. A dilated CNN model for image classification[J]. IEEE Access, 2019, 7: 124087-124095.

[48] ILESANMI A E, ILESANMI T O. Methods for image denoising using convolutional neural network: a review[J]. Complex & Intelligent Systems, 2021, 7(5): 2179-2198.

[49] LEE D, KIM M, LEE B, et al. Integrated explainable deep learning prediction of harmful algal blooms[J]. Technological Forecasting and Social Change, 2022, 185: 122046.

[50] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[51] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[52] DING W X, ZHANG C Y, SHANG S P, et al. Optimization of deep learning model for coastal chlorophyll a dynamic forecast[J]. Ecological Modelling, 2022, 467: 109913.

[53] DING W X, LI C L. Algal blooms forecasting with hybrid deep learning models from satellite data in the Zhoushan fishery[J]. Ecological Informatics, 2024, 82: 102664.

[54] GUAN W B, BAO M, LOU X L, et al. Monitoring, modeling and projection of harmful algal blooms in China[J]. Harmful Algae, 2022, 111: 102164.

[55] SIEGEL D A, DONEY S C, YODER J A. The North Atlantic spring phytoplankton bloom and Sverdrup's critical depth hypothesis[J]. Science, 2002, 296(5568): 730-733.

[56] JIN D, LEE E, KWON K, et al. A deep learning model using satellite ocean color and hydrodynamic model to estimate chlorophyll-a concentration[J]. Remote Sensing, 2021, 13(10): 2003.

[57] DE AMORIM F D L L, RICK J, LOHMANN G, et al. Evaluation of machine learning predictions of a highly resolved time series of chlorophyll-a concentration[J]. Applied Science, 2021, 11(16): 7208.

[58] KIM J H, SHIN J K, LEE H, et al. Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method[J]. Water Research, 2021, 207: 117821.

[59] AMANI M, GHORBANIAN A, ASGARIMEHR M, et al. Remote sensing systems for ocean: a review (part 1: passive systems)[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 210-234.

[60] AMANI M, MOHSENI F, LAYEGH N F, et al. Remote sensing systems for ocean: a review (part 2: active systems)[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 1421-1453.

[61] SONG W L, DOLAN J M, CLINE D, et al. Learning-based algal

中国知网 https://www.cnki.net

bloom event recognition for oceanographic decision support system using remote sensing data[J]. Remote Sensing, 2015, 7 (10): 13564-13585.

[62] CHAO Y, FARRARA J D, ZHANG H C, et al. Development, implementation, and validation of a California coastal ocean modeling, data assimilation, and forecasting system[J]. Deep Sea Research Part Ⅱ: Topical Studies in Oceanography, 2018, 151: 49-63.

[63] LAYEGHI B, BIDOKHTI A A A, GHADER S, et al. Numerical simulations of oceanographic characteristics of the Persian Gulf and Sea of Oman using ROMS model[J]. Indian Journal of Geo Marine Sciences, 2019, 48(12): 1978-1989.

[64] GUO M X, CHAI F, XIU P, et al. Impacts of mesoscale eddies in the South China Sea on biogeochemical cycles[J]. Ocean Dynamics, 2015, 65(9-10): 1335-1352.

[65] LIU Q Q, CHAI F, DUGDALE R, et al. San Francisco Bay nutrients and plankton dynamics as simulated by a coupled hydrodynamic-ecosystem model[J]. Continental Shelf Research, 2018, 161: 29-48.

[66] TIAN D, ZHOU F, ZHANG W Y, et al. Effects of dissolved oxygen and nutrients from the Kuroshio on hypoxia off the Changjiang River estuary[J]. Journal of Oceanology and Limnology, 2022, 40(2): 515-529.

[67] 苏新红, 金丰军, 杨奇志, 等. 基于BP神经网络模型的福建海域赤潮预报方法研究[J]. 水产学报, 2017, 41(11): 1744-1755.
SU X H, JIN F J, YANG Q Z, et al. Red tide forecasting model based on BP neural network in Fujian sea area[J]. Journal of Fisheries of China, 2017, 41(11): 1744-1755.

[68] TIAN W C, LIAO Z L, ZHANG J. An optimization of artificial neural network model for predicting chlorophyll dynamics[J]. Ecological Modelling, 2017, 364: 42-52.

[69] TIAN W C, LIAO Z L, WANG X. Transfer learning for neural network model in chlorophyll-a dynamics prediction[J]. Environmental Science and Pollution Research, 2019, 26(29): 29857-29871.

[70] BENTO P M R, POMBO J A N, MENDES R P G, et al. Ocean wave energy forecasting using optimised deep learning neural networks[J]. Ocean Engineering, 2021, 219: 108372.

[71] MAIER H R, DANDY G C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications[J]. Environmental Modelling & Software, 2000, 15(1): 101-124.

[72] DEDECKER A P, GOETHALS P L M, GABRIELS W, et al. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium) [J]. Ecological Modelling, 2004, 174(1-2): 161-173.

[73] LUJAN-MORENO G A, HOWARD P R, ROJAS O G, et al. Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study[J]. Expert Systems with Applications, 2018, 109: 195-205.

[74] GARCÍA NIETO P J, GARCÍA-GONZALO E, ALONSO FERNÁNDEZ J R, et al. A hybrid PSO optimized SVM-based model for predicting a successful growth cycle of the Spirulina platensis from raceway experiments data[J]. Journal of Computational and Applied Mathematics, 2016, 291: 293-303.

[75] GARCÍA-NIETO P J, GARCÍA-GONZALO E, ALONSO FERNANDEZ J R, et al. Modeling algal atypical proliferation in La Barca reservoir using L-SHADE optimized gradient boosted regression trees: a case study[J]. Neural Computing and Applications, 2021, 33(13): 7821-7838.

[76] JÖRGES C, BERKENBRINK C, STUMPE B. Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks[J]. Ocean Engineering, 2021, 232: 109046.

[77] TANG S N, ZHU Y, YUAN S Q. Intelligent fault diagnosis of hydraulic piston pump based on deep learning and Bayesian optimization[J]. ISA Transactions, 2022, 129: 555-563.

[78] SANCHEZ-MASIS A, CARMONA-CRUZ A, SCHIERHOLZ M, et al. ANN hyperparameter optimization by genetic algorithms for via interconnect classification[C]//25th IEEE Workshop on Signal and Power Integrity. Siegen: IEEE, 2021: 1-4.

[79] RAJI I D, BELLO-SALAU H, UMOH I J, et al. Simple deterministic selection-based genetic algorithm for hyper-parameter tuning of machine learning models[J]. Applied Sciences, 2022, 12(3): 1186.

[80] SOURISSEAU M, LE GUENNEC V, LE GLAND G, et al. Resource competition affects plankton community structure; Evidence from trait-based modeling[J]. Frontiers in Marine Science, 2017, 4: 52.

[81] LIU C L, TANG D L. Spatial and temporal variations in algal blooms in the coastal waters of the western South China Sea[J]. Journal of Hydro-Environment Research, 2012, 6(3): 239-247.

[82] SHIN Y, KIM T, HONG S, et al. Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods[J]. Water, 2020, 12(6): 1822.

[83] CHO K H, KANG J H, KI S J, et al. Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea[J]. Science of the Total Environment, 2009, 407(8): 2536-2545.

[84] ZHOU Y, YU L, LIU M S, et al. Network intrusion detection based on kernel principal component analysis and extreme learning machine[C]//2018 IEEE 18th International Conference on Communication Technology. Chongqing: IEEE, 2018: 860-864.

[85] MUTTIL N, CHAU K W. Machine-learning paradigms for selecting ecologically significant input variables[J]. Engineering Applications of Artificial Intelligence, 2007, 20(6): 735-744.

[86] COAD P, CATHERS B, BALL J E, et al. Proactive management of estuarine algal blooms using an automated monitoring buoy

中国知网 https://www.cnki.net

coupled with an artificial neural network[J]. Environmental Modelling & Software, 2014, 61: 393-409.
[87] SVERDRUP H U. On conditions for the vernal blooming of phytoplankton[J]. Journal du Conseil, 1953, 18(3): 287-295.
[88] BEHRENFELD M J, BOSS E S. Resurrecting the ecological underpinnings of ocean plankton blooms[J]. Annual Review of Marine Science, 2014, 6: 167-194.
[89] SHEN X R, ZHENG Y X, ZHANG R F. A hybrid forecasting model for the velocity of hybrid robotic fish based on back-propagation neural network with genetic algorithm optimization [J]. IEEE Access, 2020, 8: 111731-111741.
[90] BAI Y L, RONG Y L, SUN J H, et al. Seamount age prediction machine learning model based on multiple geophysical observables: methods and applications in the Pacific Ocean[J]. Marine Geophysical Research, 2021, 42(3): 31.
[91] 向先全, 陶建华. 基于模糊识别和遗传神经网络的渤海湾叶绿素a预测研究[J]. 海洋环境科学, 2011, 30(2): 239-242.
XIANG X Q, TAO J H. Prediction of chl-a in Bohai Bay by genetic neural network and fuzzy recognition[J]. Marine Environmental Science, 2011, 30(2): 239-242.
[92] BERSENEVA G P, CHURILOVA T Y, GEORGIEVA L V. Seasonal variability of chlorophyll and phytoplankton biomass in the western part of the Black Sea[J]. Okeanologiya, 2004, 44(3): 389-398.
[93] FINENKO Z Z, SUSLIN V V, KOVALEVA I V. Seasonal and long-term dynamics of the chlorophyll concentration in the Black Sea according to satellite observations[J]. Oceanology, 2014, 54 (5): 596-605.
[94] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[95] YAMAMOTO R, HARADA M, HIRAMATSU K, et al. Three-layered Feedforward artificial neural network with dropout for short-term prediction of class-differentiated Chl-a based on weekly water-quality observations in a eutrophic agricultural reservoir[J]. Paddy and Water Environment, 2022, 20(1): 61-78.
[96] SHIN J, YOON S, KIM Y, et al. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms[J]. Ecological Informatics, 2021, 61: 101202.
[97] ANDERSON C R, KUDELA R M, KAHRU M, et al. Initial skill assessment of the California Harmful Algae Risk Mapping (C-HARM) system[J]. Harmful Algae, 2016, 59: 1-18.
[98] 李星, 丁文祥, 李雪丁, 等. 基于人工神经网络构建的赤潮短期预报模型及应用[J]. 海洋预报, 2023, 40(2): 67-76.
LI X, DING W X, LI X D, et al. Short-term forecasting model of red tide based on artificial neural network and its application[J]. Marine Forecasts, 2023, 40(2): 67-76.

Advances in the application of machine learning algorithms in early warning models for harmful algal blooms

DING Wenxiang[1,2], LIN Chenxu[1], ZHANG Caiyun[1*]

(1. Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Ministry of Education, Xiamen 361102, China; 2. Marine Science and Technology College, Zhejiang Ocean University, Zhoushan 316022, China)

Abstract: This paper reviews the progress in the application of classical machine learning algorithms, artificial neural networks, and deep learning models in harmful algal blooms early warning models. Addressing specific challenges such as data scarcity, limited generalization capability, and the need for improved accuracy, this paper provides a detailed discussion on the role of multi-source data integration, model structure and parameter optimization, and ensemble forecasting strategies in improving model performance.

Key words: machine learning; artificial neural networks; deep learning; early warning; harmful algal blooms

中国知网 https://www.cnki.net