大西洋热带海域长鳍金枪鱼渔场预报模型的比较*

宋利明[1, 2] 任士雨[1] 洪依然[1] 张天蛟[1] 隋恒寿[3] 李 彬[3] 张 敏[1, 2]

(1. 上海海洋大学海洋科学学院 上海 201306; 2. 国家远洋渔业工程技术研究中心 上海 201306;
3. 中水集团远洋股份有限公司 北京 100032)

摘要 为提高大西洋热带海域长鳍金枪鱼(*Thunnus alalunga*)渔场预报的准确率, 对 K 最近邻(k nearest neighbor, KNN)、逻辑斯蒂回归(logistic regression, LR)、决策与分类树(classfication and regression tree, CART)、梯度提升决策树(gradient boosting decision tree, GBDT)、随机森林(random forest, RF)、支持向量机(support vector machine, SVM)和 Stacking 集成(stacking ensemble learning, STK)共 7 个模型的预报性能进行了对比分析。该 7 个模型利用 2016~2019 年在大西洋公海海域(19°16′S~16°21′N; 46°27′W~2°09′E)作业的 13 艘中国远洋延绳钓渔船的渔业数据, 结合 0~500 m 不同水层的温度、盐度、溶解氧、叶绿素 *a* 浓度、海表面风速、涡动能和混合层深度数据建立。各模型取 75%数据作为训练数据, 25%为测试数据, 采用预报准确率(accuracy, ACC)与接受者操作特征曲线下面积(area under curve, AUC)评价建立的单位努力量渔获量(catch per unit of effort, CPUE)与海洋环境因子关系模型的性能。结果表明: (1) STK 模型对大西洋长鳍金枪鱼渔场的预报性能相比其他模型明显提高, ACC 为 75.92%, AUC 为 0.742; (2) 基于 STK 模型预测得到的中心渔场范围总体上与实际作业渔场一致; (3) 影响大西洋长鳍金枪鱼渔场分布的海洋环境因子主要为 100 m 水层的温度、盐度, 以及 100、150、500 m 水层的溶解氧。研究表明 STK 模型对大西洋长鳍金枪鱼渔场的预报准确率较高, 性能良好。

关键词 长鳍金枪鱼; 渔场预报模型; 模型性能比较; 大西洋热带海域
中图分类号 S934 doi: 10.11693/hyhz20211000253

长鳍金枪鱼(*Thunnus alalunga*)广泛分布于热带、亚热带及温带海域, 为高度洄游性鱼类, 是我国远洋延绳钓渔业重要的捕捞对象之一。近年来, 国内外学者研究了长鳍金枪鱼分布与海洋环境之间的关系(Chen *et al*, 2005; Domokos *et al*, 2007; Zainuddin *et al*, 2008; Briand *et al*, 2011; 林显鹏等, 2011; 杨嘉樑等, 2014; Goñi *et al*, 2015; Lehodey *et al*, 2015; Williams *et al*, 2015; 储宇航等, 2016; 郭刚刚等, 2016; 宋利明等, 2017a), 宋利明等(2017b)研究得出温度是影响库克群岛海域长鳍金枪鱼分布的主要环境因子; 闫敏等(2015)认为海表面温度和叶绿素 *a* 浓度与长鳍金枪鱼渔获率分布有关; 张嘉容等(2020)认为温度对长鳍金枪鱼分布的影响最大, 叶绿素 *a* 浓度的影响最小。但大部分模型仅分析海洋表层环境与长鳍金枪鱼分布的关系, 缺乏使用海洋立体环境因子的分析研究。

目前长鳍金枪鱼渔场预测大多使用较为传统的单一预报模型, 如一元非线性回归模型(范江涛, 2011)、栖息地指数模型(任中华等, 2014)和神经网络模型(毛江美等, 2016)等。随着渔业生产对渔场预报精度要求的提高, 深度学习开始在长鳍金枪鱼渔场预测中得到应用, 如袁红春等(2019a, 2020)。Stacking 集成学习是一种多模型的集成方法, 能够得到比单模型更准确的结果(罗智青等, 2019; 侯娟等, 2020)。但海洋立体环境因子间的共线性会影响如逻辑斯蒂回归(logistic regression, LR)模型的预测精度(Raykov *et al*, 2010), 且环境因子过多容易导致 K 最近邻(k

* 国家重点研发项目, 2020YFD0901205 号; 中水集团远洋股份有限公司技术研发项目, D-8006-20-0180 号。宋利明, 博士生导师, 教授, E-mail: lmsong@shou.edu.cn
收稿日期: 2021-10-23, 收修改稿日期: 2021-12-20

nearest neighbor, KNN)、梯度提升决策树(gradient boosting decision tree, GBDT)、随机森林(random forest, RF)和支持向量机(support vector machine, SVM)等模型的复杂性, 造成过拟合问题, 降低模型可解释性(Jou *et al*, 2014; Assegie *et al*, 2021)。目前大部分机器学习模型均未针对海洋立体环境因子间的共线性进行分析及因子筛选。

本研究根据 2016~2019 年中国船队在大西洋公海作业的延绳钓渔业数据, 结合不同深度的海洋环境数据进行共线性分析与因子筛选, 建立 KNN、LR、决策与分类树(classfication and regression tree, CART)、GBDT、RF、SVM 和 Stacking 集成(stacking ensemble learning, STK)模型; 各模型取 75%站位的数据作为训练数据, 25%站位的数据为测试数据, 采用预报准确率(accuracy, ACC)与接受者操作特征(receiver operating characteristic, ROC)曲线下面积(area under curve, AUC)进行性能评价, 以提高大西洋热带海域长鳍金枪鱼渔场预报的准确性。

1　材料与方法

1.1　数据来源

本研究选取的渔业数据来源于中水集团远洋股份有限公司, 时间为 2016~2019 年, 区域为 19°16′S~16°21′N; 46°27′W~2°09′E 范围内的大西洋中部热带海域, 数据信息包括延绳钓捕捞渔船(共 13 艘)船名、作业时间(年-月-日)、作业地点(经纬度)、不同鱼种的产量、渔获尾数及下钩数等。将每天长鳍金枪鱼的渔获尾数划分到 1°×1°的网格内, 并算出 CPUE (尾/千钩)。

海洋环境因子主要包括表层因子和 0~500 m 深度因子共 29 个。其中, 表层因子包括: 叶绿素 *a* 浓度(chl_*a*)、混合层深度(mixed layer depth, MLD)、海表面风速(wind speed, WS)、涡动能(eddy kinetic energy, EKE); 0~500 m 深度因子包括: 0、50、100、150、200、250、300、400、500 m 水深处的温度(采用 T_0、T_{50}、T_{100}、T_{150}、T_{200}、T_{250}、T_{300}、T_{400}、T_{500} 表示); 0、50、100、150、200、250、300、350、400、450、500 m 水深处的溶解氧(采用 D_0、D_{50}、D_{100}、D_{150}、D_{200}、D_{250}、D_{300}、D_{350}、D_{400}、D_{450}、D_{500} 表示); 100、200、300、400、500 m 水深处的盐度(采用 S_{100}、S_{200}、S_{300}、S_{400}、S_{500} 表示)。以上因子数据除 WS 来源于美国国家海洋和大气管理局(National Oceanic and Atmospheric Administration, NOAA)的数据库(https://oceanwatch.pifsc.noaa.gov/)外, 其他数据均来源于哥白尼海洋环境监测服务中心(copernicus marine environment monitoring service, CMEMS)网站 (http://marine.copernicus.eu)。环境因子数据的初始时间分辨率为 d, 初始空间分辨率为 0.25°×0.25°。本文将环境因子和渔业数据进行了匹配, 最终的空间分辨率统一为 1°×1°, 以 d 为单位。以上数据预处理过程通过 MATLAB 软件完成。

1.2　数据处理

1.2.1　长鳍金枪鱼 CPUE 的计算方法　　CPUE 作为评价渔场的指标(沈智宾等, 2015)。将每天的长鳍金枪鱼的渔获尾数划分到 1°×1°的网格内, 根据每天的船位数据等得到每天每网格内的总钓钩数, 算出每天每个网格内的长鳍金枪鱼 CPUE (尾/千钩), 计算各网格内 CPUE 的公式为

$$E_{kmnij}=\frac{F_{kmnij}}{H_{kmnij}}\times 1000\,, \tag{1}$$

式(1)中: E_{kmnij}, F_{kmnij} 和 H_{kmnij} 分别表示在第 k 年、第 m 月、第 n 天、第 i 经度、第 j 纬度的网格内的 CPUE、尾数和下钩数量。

1.2.2　海洋环境因子的共线性分析　　由于建立模型所用的叶绿素 *a* 浓度、混合层深度、海表面风速、涡动能以及 0~500 m 水层的温度、溶解氧、盐度等海洋环境因子的量级单位不同, 数值范围差别较大, 为防止对模型产生数值影响且为了提高模型运行的准确度, 对所有环境变量进行归一化处理(张天蛟, 2016), 公式为

$$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}}\,, \tag{2}$$

式(2)中: X'、X、$X_{\max}$、$X_{\min}$ 分别为归一化后的值、实际值、最大值、最小值。

针对 29 个海洋环境因子之间存在的多重共线性, 本文采用方差膨胀因子(variance expansion factor, VIF)(Akinwande *et al*, 2015)进行分析。

对于方程:

$$Y=\alpha_1X_1+\alpha_2X_2+\alpha_3X_3+\ldots+\alpha_iX_i+\ldots+\alpha_nX_n+\beta, \tag{3}$$

式(3)中, y 为因变量, x 为自变量, α_1、α_2、α_3、…、α_n 为回归系数, β 为常数项。则第 i 个自变量的 VIF 计算公式为:

$$F=\frac{1}{1-R_k^2}\,, \tag{4}$$

式(4)中, F 表示方差膨胀因子 VIF 的取值; R_k 为线性方程中的决定系数。当 VIF 值较大时, 表示环境变量之间存在多重共线性, 即 VIF 值越大, 表示环境变量之间多重共线性越严重, 一般认为 F<10, 即表示各环境变量之间没有多重共线性。选取 F<10 的海洋环

境因子用于建立大西洋热带海域长鳍金枪鱼的渔场预报模型。

1.2.3 海洋环境因子相对重要性分析 海洋环境因子相对重要性的计算方法为: 利用 SPSS 软件计算长鳍金枪鱼 CPUE 与筛选后各环境因子的 Pearson 相关系数; 根据 Pearson 相关系数绝对值判断其相对重要性。

1.3 模型建立

渔场分类时, 若 CPUE 等于 0, 为“非渔场”; CPUE 大于 0, 则为“渔场”, 其中零值比列为 40.11%。从数据集中随机选取 75%站点的数据作为训练数据, 25%站点的数据为测试数据; 并使用训练数据分别建立 KNN、LR、SVM、CART、RF、GBDT 和 STK7 种模型。系列模型的基本原理和参数选择方法如表 1。

Stacking 算法框架如图 1 所示, 第一层由 3 个基学习器组成, 并输入原始训练集, 即第一步将与各个海洋环境因子匹配好的大西洋热带海域长鳍金枪鱼 CPUE 原始数据集 $\boldsymbol{S}$, 基于非共线性海洋环境因子的筛选, 划分为 75%的训练数据 $\boldsymbol{D}$ 和 25%测试数据 $\boldsymbol{T}$; 第二步从 7 个模型中选择预测效果较好的 3 个模型作为初级学习器, 选取结构简单的 LR 作为次级学习器; 第三步将 75%的训练样本随机划分为 k 个数量相同的训练集子集 $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, …, $\boldsymbol{D}_k$, 取 $\boldsymbol{T}_i$ 作为测试子集, $\boldsymbol{D}_{k-1}\{\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_k\}(\boldsymbol{D}_i \notin \boldsymbol{D}_{k-1})$作为 KNN、RF 和 GBDT 的训练子集, 接着将各个模型的预测结果统一为 LR 模型的训练集 $\boldsymbol{D}'$, 各个单模型对测试集 $\boldsymbol{T}$ 的预测结果合并为 LR 模型的测试集 $\boldsymbol{T}'$, 最终输出预测结果。

1.4 模型性能评价

将 25%测试数据代入各模型得出预测结果, 对模型性能进行评价。本文采用 AUC 和 ACC 作为模型性能评价指标。其中 AUC 取值范围为 0~1, 值越大说明该模型的预测性能越好, 反之预测性能越差(张天蛟, 2016); 同理, 所得 ACC 值越大, 说明该模型整体预测效果越好, 反之预测效果越差, 所以本文按照两者结果, 对比并选取预测准确度较高且性能稳定的模型。

表 1 各模型的基本原理与参数选择

Tab.1 The basic principle and the parameter selection of each model

模型名称	参考文献	基本原理	参数选择
K 最近邻模型(k nearest neighbor, KNN)	(Hwang *et al*, 1998; 王超学等, 2012; 张莹, 2015; Zhang *et al*, 2018; 吴昊等, 2019)	为分类模型, 主要是针对测试数据, 找到与该测试数据距离最近的 k 个训练数据, 即对该训练数据进行分类	取邻近点的个数 K=7
逻辑斯蒂回归(logistic regression, LR)	(Dahlem *et al*, 1989; Zhang *et al*, 2018)	在线性回归模型的基础上使用 sigmoid 函数, 将线性模型的结果压缩到[0,1]之间, 使其拥有概率意义	惩罚系数 C=1; 最大的迭代次数 100; 当目标函数(样本集最大似然函数)导数的第 j 个分量小于 0.000 1 时, 迭代停止
支持向量机(support vector machine, SVM)	(Suykens *et al*, 2000; 邵元海等, 2020)	SVM 的中心思想是通过二分类方法找到使数据集到分隔超平面的几何间隔最远的点集合面, SVM 通过引入核函数, 实现从非线性学习到线性学习的过渡	惩罚系数为 1
决策树算法(classfication and regression tree, CART)	(Kristensen *et al*, 1998; 谢金梅等, 2008; 田欣, 2017)	主要通过计算 Gini 指数不纯度, 再根据最小的 Gini 指标作为分裂属性, 最终使测试样本数据集形成二分类递归的分割树权	复杂度为 0.01, 最大深度为 30
随机森林算法(random forest, RF)	(Ou *et al*, 2002; 陈雪忠等, 2013; 吕红燕等, 2019)	对于原始测试样本随机抽取训练子集, 生成分类树, 再随机抽取 q 个特征作为分裂属性, 至每棵决策树完整形成随机森林模型	最小样本数量为 1, 节点最小样本数量为 1
梯度提升决策树(gradient boosting decision Tree, GBDT)	(Friedman, 2003; 田欣, 2017)	是一种对决策树进行 Boosting 集成的学习算法, 由集中选取相对简易的决策树来提高模型的预测性能	树的个数为 100, 树的深度为 6, 学习率为 0.1
Stacking 集成学习(stacking ensemble learning, STK)	(Džeroski *et al*, 2004; 张春霞等, 2011; 袁培森等, 2019)	通过集合训练多个分类器来得到较单个分类器 更强、更优化的分层模型集成框架	初级学习器 classifiers=[RF, GBDT, KNN]; 次级学习器 meta_classifier=LR; 交叉验证折数 R=5

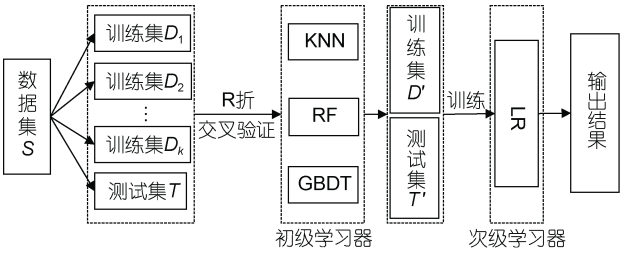

图 1　Stacking 集成学习方法

Fig.1　Method of stacking ensemble learning

注: KNN: k nearest neighbor, K 最近邻模型; RF: random forest, 随机森林算法; GBDT: gradient boosting decision tree, 梯度提升决策树; LR: logistic regression, 逻辑斯蒂回归

将 25%的测试数据代入预测能力最佳的模型, 计算得出“渔场”和“非渔场”并使用 ArcGIS 软件画出实际“渔场”和“非渔场”与模型预测得到的“渔场”和“非渔场”叠图, 定性评价最佳模型的预测能力。

1.5　中心渔场的确定

本研究把 25%的测试站点的环境数据代入预测能力最佳的模型, 计算得出“渔场”位置, 利用 ArcGIS 软件的核密度分析工具计算并画出“渔场”位置密度分布图, 把密度大于最大密度的 60% (8 个/km^2)以上的范围定义为中心渔场。

2　结果

2.1　长鳍金枪鱼渔场分布

通过 ArcGIS 软件画出长鳍金枪鱼 CPUE 分布图(图 2), 得出长鳍金枪鱼渔场主要分布在 15°N~15°S, 15°~45°W 的大西洋中部热带海域, 其高 CPUE 渔区集中分布在 4°~10°N, 30°~45°W 海域和 5°~10°S, 18°~30°W 的海域。

2.2　海洋环境因子筛选与相对重要性分析

按照各环境因子的$|F|$<10, 筛选得出 20 个用于渔场预测的环境因子, 分别是 chl *a*、MLD、WS、EKE、T_0、T_{50}、T_{100}、T_{150}、T_{250}、T_{400}、D_0、D_{50}、D_{100}、D_{150}、D_{200}、D_{500}、S_{100}、S_{200}、S_{300}和 S_{500} (表 2)。相关系数分析结果表明, 100 m 水层温度(T_{100})与 CPUE 的相关系数最高, 达到 0.399; 100、150、500 m 水层的溶解氧、100 m 水层的盐度与 CPUE 的相关系数较高, 分别达到 0.372、0.334 、0.322 和 0.322; 相关系数较小的是 MLD 和叶绿素 *a* 浓度, 分别为 0.21 和–0.148; 海表面温度(T_0)和海表面风速(WS)的相关系数分别为 0.140 和 0.069; 而涡动能(EKE)的相关系数最低, 为–0.036, 影响程度最小(表 2)。

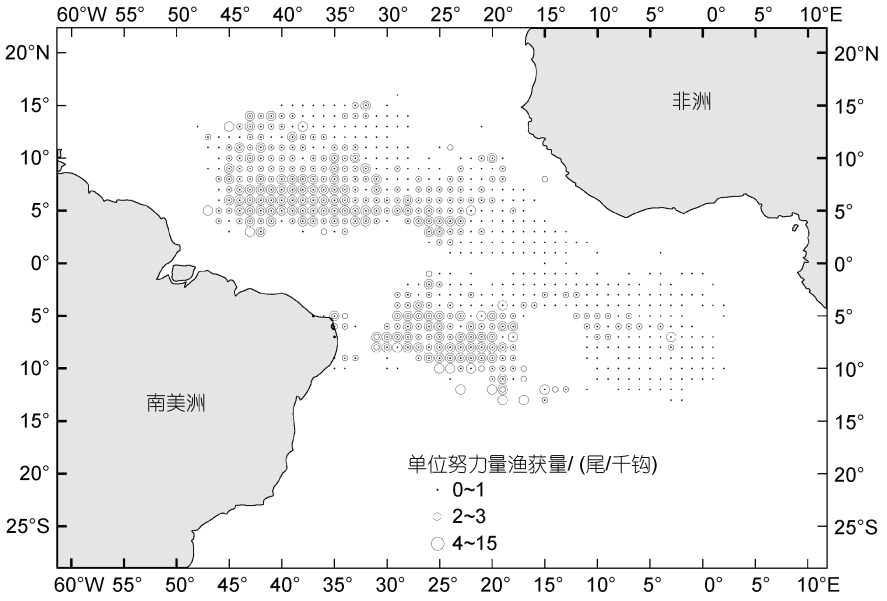

图 2　2016~2019 年大西洋热带海域 13 艘渔船长鳍金枪鱼 CPUE 分布

Fig.2　CPUE distribution of *Thunnus alalunga* from 13 fishing vessels in the tropical waters of Atlantic Ocean in 2016~2019

2.3　模型性能评价

各个模型总体的 ACC 和 AUC 如表 3 所示, 单个模型预测结果中 RF 的准确度最高, 为 75.57%, KNN 和 GBDT 的准确度并列为第二(73.92%), 因此选取 RF、KNN 和 GBDT 用于 STK 模型; CART 最低, 为 66.85%。STK 模型的 ACC 最高, 达到 75.92%, AUC 也达到了 0.742。综合上述结果, 表明 STK 模型预测性能良好。

表 2 多重共线性和相关性分析结果
Tab.2 Results of multi-collinearity diagnosis and correlation analysis

变量	容差	VIF	相关系数 R
chl_a	0.433	2.310	–0.148
EKE	0.810	1.235	–0.036
MLD	0.494	2.022	0.210
WS	0.726	1.377	0.069
S_{100}	0.265	3.769	0.322
S_{200}	0.174	5.758	0.067
S_{300}	0.177	5.656	–0.105
S_{500}	0.223	4.483	–0.113
T_0	0.160	6.251	0.140
T_{50}	0.240	4.173	0.228
T_{100}	0.152	6.561	0.399
T_{150}	0.234	4.266	0.313
T_{250}	0.131	7.612	–0.037
T_{400}	0.223	4.481	–0.079
D_0	0.112	8.964	0.046
D_{50}	0.179	5.593	0.062
D_{100}	0.229	4.365	0.372
D_{150}	0.178	5.617	0.334
D_{200}	0.184	5.425	0.230
D_{500}	0.388	2.574	0.322

对于 STK 模型，将测试数据中得到的实际渔场与预测渔场进行叠加，得出渔场主要分布在 2°~14°N, 32°~45°W 和 2°~10°S, 18°~28°W 的海域，极少分布在 5°~10°W 附近，且预测的 0°~18°W 海域的渔场的误判率很高(图 3)；将测试数据中的实际非渔场与预测非渔场进行叠加，实际非渔场主要分布在 5°~14°N, 30°~45°W 和 2°S~7°N, 12°~30°W 的海域，极少分布在 2°~12°S, 2°~10°W 海域附近，且预测的 2°~8°S, 20°~30°W 海域的实际非渔场的误判率较高(图 4)。各模型对渔场和非渔场的判别准确率比较结果见表 4。说明 Stacking 集成模型的预测性能良好。

表 3 各个模型预测结果对比
Tab.3 Comparison of forecast results of each model

模型	ACC/%	AUC
LR	71.93	0.692
RF	75.57	0.737
GBDT	73.92	0.715
KNN	73.92	0.709
CART	66.85	0.656
SVM	71.62	0.684
STK	75.92	0.742

注: ACC: accuracy, 预报准确率; AUC: area under curve, 曲线下面积

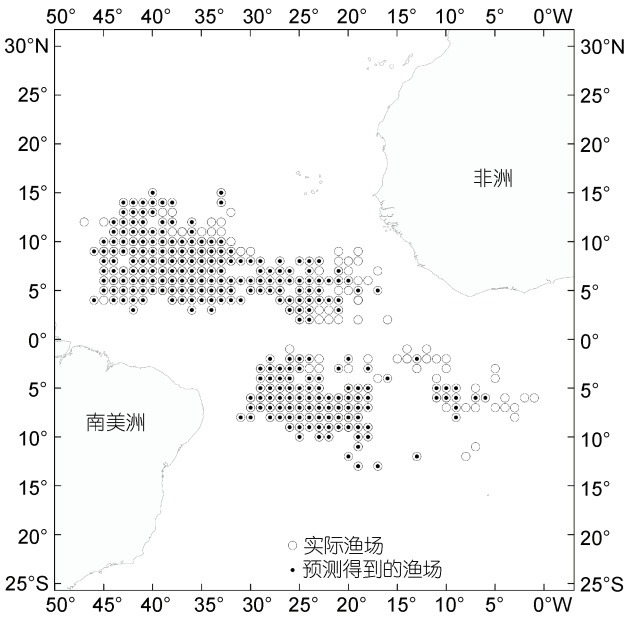

图 3 实际渔场与预报得到的渔场对比图
Fig.3 Comparison of actual fishing ground and predicted fishing ground

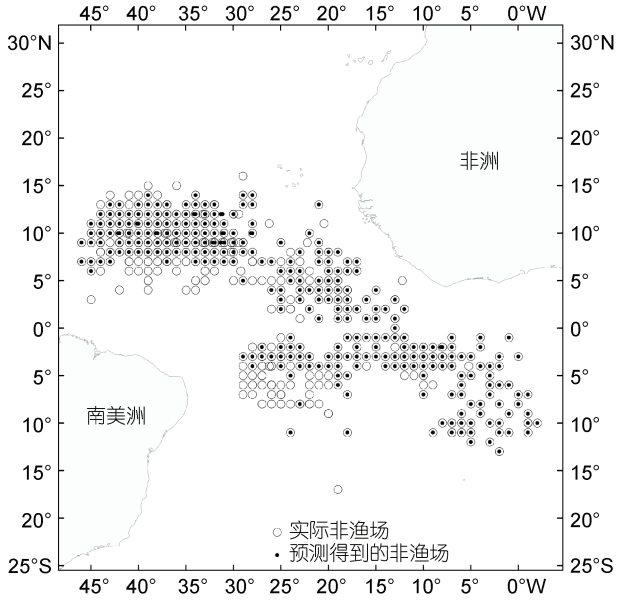

图 4 实际非渔场与预报得到的非渔场对比图
Fig.4 Comparison of actual non-fishing ground and predicted non-fishing ground

2.4 中心渔场

2016~2019 年 25%的测试数据中实际渔场密度分布如图 5a 所示，其中心渔场集中分布在 4°~10°N, 33°~43°W 海域，以及 5°~8°S, 27°~29°W 和 7°~9°S, 22°~24°W 的海域。STK 模型预测渔场密度分布如图 5b 所示，从整体来看预测的中心渔场比实际中心渔场略大，但基本吻合。

表 4　各模型对渔场和非渔场的判别准确率(%)比较
Tab.4　Comparison of discrimination accuracy (%) for fishing ground and non-fishing ground of each model

模型	渔场	非渔场
KNN	80.52	61.21
LR	84.08	54.31
SVM	84.78	51.05
CART	71.60	57.76
RF	82.96	63.79
GBDT	83.81	58.14
STK	85.09	62.75

3　讨论

3.1　模型对比分析

STK 模型的渔场预报性能比 6 种单一模型(KNN、LR、CART、SVM、RF、GBDT)对大西洋热带海域长鳍金枪鱼渔场预报的性能要好，为最佳模型。CART 的渔场预报性能最差。这是因为 STK 是一种堆叠集成方法，能够将 KNN、RF 和 GBDT 预测结果再次进行次级训练，克服单一模型的局限性，能综合三种模型的优势，而 CART 容易过拟合，导致泛化能力不强，且某些自变量的占比过高时，容易导致预测能力降低。RF 对长鳍金枪鱼渔场的预报性能与 STK 模型相近，这可能是因为 RF 在构建模型时通过自助采样选取训练样本，增强其泛化和抗噪能力，在一定程度上降低了不良数据对模型预测能力的影响(方匡南等, 2011)。GBDT 与 RF 都是由多个分类树通过不同算法构建的模型，但 GBDT 的预测性能明显低于 RF，这是因为 GBDT 训练样本选取的方法与 RF 不同，其在模型构建过程中使用同一个训练样本，导致模型泛化能力降低。LR、KNN、SVM 的预测性能都比较低，这几个模型不能有效克服构建模型所使用的样本数据质量不高的问题(如各类别的样本数据不平衡、数据缺失等)。

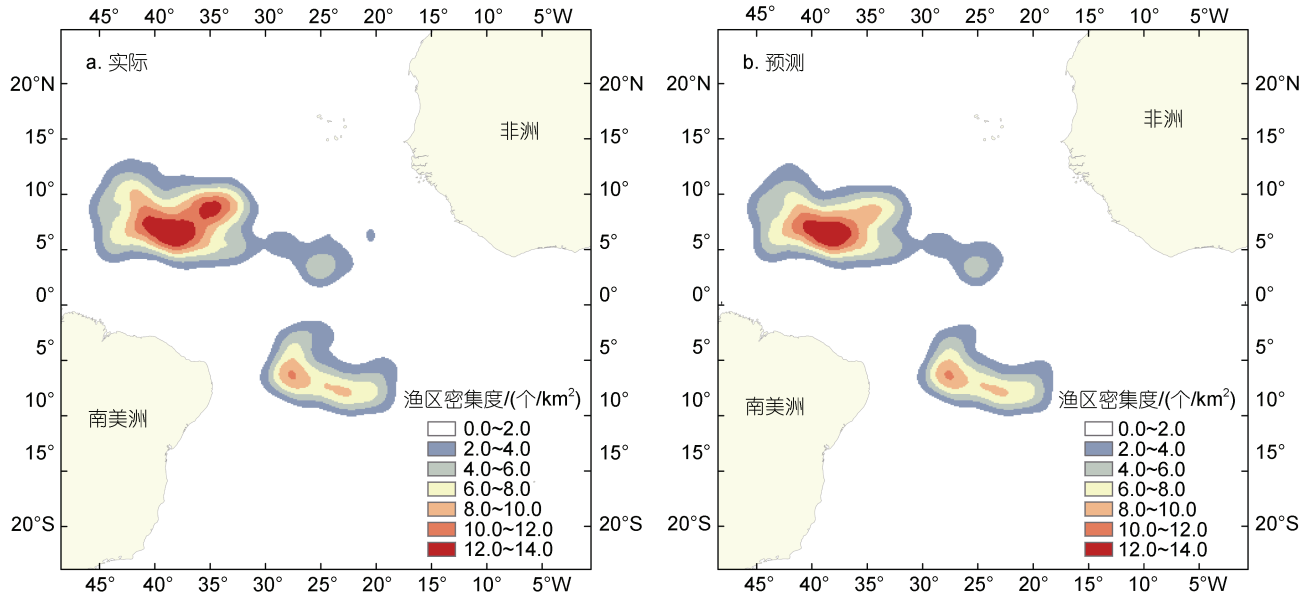

图 5　25%的测试数据中渔区密度分布图
Fig.5　Distribution of fishing ground density in the 25% test data

3.2　环境变量共线性分析

由于海洋环境非常复杂，环境因子相互影响，可能导致环境变量之间存在多重共线性，会对预测结果产生偏差。Dormann *et al* (2013)认为当环境变量之间的相关系数大于 0.7 时，说明存在共线性的可能性较大。多重共线性是影响 SVM 和 LR 预测精度和运行速率的主要原因之一(惠守博等, 2006; 张玲, 2010)，虽然 CART、RF 和 GBDT 在模型训练的过程中会自动消除多重共线性对预测精度的影响，但是共线性使得环境变量的大部分信息相互叠加，导致数据大量冗余，模型的运行效率降低以及误判解释变量的贡献率。故消除环境变量之间的多重共线性，在一定程度上提高了模型的预测精度、稳定性和可靠性。

VIF 方法虽然可以消除共线性，但也有可能导致原本生态学意义重要的因子被删除，本研究从 29 个潜在的影响因子中删除了 9 个具有共线性的因子(T_{200}、T_{300}、T_{500}、D_{250}、D_{300}、D_{350}、D_{400}、D_{450}、S_{400})。这些被删除的因子均为 200 m 以深的因子，由于长鳍

金枪鱼偏好觅食的水层为 160~240 m 水层，且在该水层主要受饵料分布及水温的影响(杨嘉樑等，2014)，因此本研究未删除对长鳍金枪鱼具有生态学重要意义的因子。

3.3 环境因子的相对重要性及中心渔场

研究认为温度直接或间接影响鱼类分布、集群和洄游(陈雪忠等，2013)，其垂直水温是影响长鳍金枪鱼渔场分布主要原因之一(郭刚刚等，2016)。研究表明，100 m 水层的温度、溶解氧和盐度与长鳍金枪鱼 CPUE 关系较为密切，可能的原因是该水层是水温急剧下降的温跃层，温度和盐度等环境因子适宜长鳍金枪鱼生存，并且具有丰富的饵料生物资源，较高的溶解氧含量(张嘉容等, 2020)。张嘉容等(2020)研究分析得到南太平洋中对长鳍金枪鱼 CPUE 具有显著影响的环境因子是 120 m 水层的温度和盐度，这与本研究结果基本一致。叶绿素 *a* 浓度和 WS 对长鳍金枪鱼渔场分布的影响较小，这是由于海面风速能够改变叶绿素 *a* 的空间位置和含量(Pickett *et al*, 2006; Jufaili *et al*, 2019)，而较高的叶绿素 *a* 浓度会吸引更多浮游生物在附近繁殖(杨胜龙等,2012)，但长鳍金枪鱼并不会直接捕食浮游生物。EKE 对长鳍金枪鱼渔场分布的影响最小，这可能是因为 EKE 是通过影响环流、海洋温度以及叶绿素 *a* 的垂直和水平分布(Tussadiah *et al*, 2018)间接影响长鳍金枪鱼渔场的分布。2016~2019 年 25%的测试数据中预测得到的中心渔场比实际中心渔场范围略大，可能是由于渔业生产作业不能均匀、完全覆盖整个区域，而预测是根据海洋环境数据进行平滑计算得出的，预测得到的中心渔场范围可靠。另外，准确预测中心渔场能够使渔船缩短搜索渔场的时间、节省燃油，增加长鳍金枪鱼的渔获量，提高经济效益。

4 展望

本研究根据 29 种海洋环境因子，建立 6 种模型并筛选最佳的三个预测模型构成 STK 模型，提高了大西洋热带海域长鳍金枪鱼渔场预报的精度，但长鳍金枪鱼渔场预报的精度还与数据的空间分辨率、渔捞日志记录的数据的准确度、饵料资源分布、洋流、流速等其他环境变量有关，还会受到气候的年代际，如北大西洋涛动等大尺度海洋事件的影响，本研究中 CPUE 为 0 的比例较高，还需要进一步收集数据开展研究。另外，本研究结果适用于大西洋热带海域长鳍金枪鱼渔场的预报，对于大西洋温带水域的长鳍金枪鱼渔场的预报还需要进一步收集数据开展相关研究。

致谢 本研究得到了中水集团远洋股份有限公司宗文峰、叶少华和邓荣成先生的大力支持，谨致谢意。

参 考 文 献

王超学, 潘正茂, 马春森, 等, 2012. 改进型加权 KNN 算法的不平衡数据集分类[J]. 计算机工程, 38(20): 160-163, 168.

毛江美, 陈新军, 余景, 2016. 基于神经网络的南太平洋长鳍金枪鱼渔场预报[J]. 海洋学报, 38(10): 34-43.

方匡南, 吴见彬, 朱建平, 等, 2011. 随机森林方法研究综述[J]. 统计与信息论坛, 26(3): 32-38.

田欣, 2017. 决策树算法的研究综述[J]. 现代营销(1): 36.

吕红燕, 冯倩, 2019. 随机森林算法研究综述[J]. 河北省科学院学报, 36(3): 37-41.

任中华, 陈新军, 方学燕, 2014. 基于栖息地指数的东太平洋长鳍金枪鱼渔场分析[J]. 海洋渔业, 36(5): 385-395.

闫敏, 张衡, 樊伟, 等, 2015. 南太平洋长鳍金枪鱼渔场 CPUE 时空分布及其与关键海洋环境因子的关系[J]. 生态学杂志, 34(11): 3191-3197.

杨胜龙, 张禹, 樊伟, 等, 2012. 热带印度洋大眼金枪鱼渔场时空分布与温跃层关系[J]. 中国水产科学, 19(4): 679-689.

杨嘉樑, 黄洪亮, 宋利明, 等, 2014. 基于分位数回归的库克群岛海域长鳍金枪鱼栖息环境综合指数[J]. 中国水产科学, 21(4): 832-851.

吴昊, 秦立春, 罗柳容, 2019. 基于提升度的 KNN 分类子的分类原则改良模型[J]. 广西师范大学学报(自然科学版), 37(2): 75-81.

沈智宾, 陈新军, 汪金涛, 2015. 基于海表温度和海面高度的东太平洋大眼金枪鱼渔场预测[J]. 海洋科学, 39(10): 45-51.

宋利明, 周建坤, 沈智宾, 等, 2017a. 基于支持向量机的库克群岛海域长鳍金枪鱼栖息环境综合指数[J]. 海洋通报, 36(2): 195-208.

宋利明, 谢凯, 赵海龙, 等, 2017b. 库克群岛海域海洋环境因子对长鳍金枪鱼渔获率的影响[J]. 海洋通报, 36(1): 96-106.

张玲, 2010. 多重共线性的检验及对预测目标影响程度的定量分析[J]. 通化师范学院学报, 31(4): 19-20, 38.

张莹, 2015. 基于自然最近邻居的分类算法研究[D]. 重庆: 重庆大学.

张天蛟, 2016. 产漂流性卵小型鱼类的生态位建模及分析[D]. 北京: 中国农业大学.

张春霞, 张讲社, 2011. 选择性集成学习算法综述[J]. 计算机学报, 34(8): 1399-1410.

张嘉容, 杨晓明, 戴小杰, 等, 2020. 南太平洋长鳍金枪鱼延绳钓渔获率与环境因子的关系研究[J]. 南方水产科学, 16(1): 69-77.

陈雪忠, 樊伟, 崔雪森, 等, 2013. 基于随机森林的印度洋长鳍金枪鱼渔场预报[J]. 海洋学报, 35(1): 158-164.

邵元海, 刘黎明, 黄凌伟, 等, 2020. 支持向量机的关键问题和展望[J]. 中国科学: 数学, 50(9): 1233-1248.

范江涛, 2011. 南太平洋长鳍金枪鱼延绳钓渔业渔情预报研究[D]. 上海: 上海海洋大学: 21-22.

林显鹏, 郭爱, 张洪亮, 等, 2011. 所罗门群岛海域长鳍金枪鱼的垂直分布与环境因子的关系[J]. 浙江海洋学院学报(自然科学版), 30(4): 303-306.

罗智青, 莫汉培, 王汝辉, 等, 2019. 基于 Stacking 模型融合的失压故障识别算法[J]. 能源与环保, 41(2): 41-45.

侯娟, 周为峰, 樊伟, 等, 2020. 基于集成学习的南太平洋长鳍金枪鱼渔场预报模型研究[J]. 南方水产科学, 16(5): 42-50.

袁红春, 陈冠奇, 张天蛟, 等, 2020. 基于全卷积网络的南太平洋长鳍金枪鱼渔场预报模型[J]. 江苏农业学报, 36(2): 423-429.

袁红春, 陈骢昊, 2019a. 基于融合深度学习模型的长鳍金枪鱼渔情预测研究[J]. 渔业现代化, 46(5): 74-81.

袁红春, 胡光亮, 陈冠奇, 等, 2019b. 基于粒子群可拓的南太平洋长鳍金枪鱼产量预测方法研究[J]. 渔业现代化, 46(6): 96-103.

袁培森, 杨承林, 宋玉红, 等, 2019. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J]. 农业机械学报, 50(11): 144-152.

郭刚刚, 张胜茂, 樊伟, 等, 2016. 南太平洋长鳍金枪鱼垂直活动水层空间分析[J]. 南方水产科学, 12(5): 123-130.

惠守博, 王文杰, 2006. 支持向量机分类算法中多元变量共线性问题的改进[J]. 计算机工程与设计, 27(8): 1385-1388.

储宇航, 戴小杰, 田思泉, 等, 2016. 南太平洋延绳钓长鳍金枪鱼生物学组成及其与栖息环境关系[J]. 海洋渔业, 38(2): 130-139.

谢金梅, 王艳妮, 2008. 决策树算法综述[J]. 软件导刊, 7(11): 83-85.

AKINWANDE M O, DIKKO H G, SAMSON A, 2015. Variance inflation factor: as a condition for the inclusion of suppressor variable(s) in regression analysis [J]. Open Journal of Statistics, 5(7): 754-767.

ASSEGIE T A, SUSHMA S J, BHAVYA B G, *et al*, 2021. Correlation analysis for determining effective data in machine learning: detection of heart failure [J]. SN Computer Science, 2(3): 213.

BRIAND K, MOLONY B, LEHODEY P, 2011. A study on the variability of albacore (*Thunnus alalunga*) longline catch rates in the southwest Pacific Ocean [J]. Fisheries Oceanography, 20(6): 517-529.

CHEN I C, LEE P F, TZEND W N, 2005. Distribution of albacore (*Thunnus alalunga*) in the Indian Ocean and its relation to environmental factors [J]. Fisheries Oceanography, 14(1): 71-80.

DAHLEM A M, HASSAN A S, SWANSON S P, *et al*, 1989. A model system for studying the bioavailability of intestinally administered microcystin‐LR, a hepatotoxic peptide from the cyanobacterium *Microcystis aeruginosa* [J]. Pharmacology & Toxicology, 64(2): 177-181.

DOMOKOS R, SEKI M P, POLOVINA J J, *et al*, 2007. Oceanographic investigation of the American Samoa albacore (*Thunnus alalunga*) habitat and longline fishing grounds [J]. Fisheries Oceanography, 16(6): 555-572.

DORMANN C F, ELITH J, BACHER S, *et al*, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance [J]. Ecography, 36(1): 27-46.

DŽEROSKI S, ŽENKO B, 2004. Is combining classifiers with stacking better than selecting the best one? [J]. Machine Learning, 54(3): 255-273.

FRIEDMAN J H, 2003. Stochastic gradient boosting [J]. Computational Statistics & Data Analysis, 38(4): 367-378.

GOÑI N, DIDOUAN C, ARRIZABALAGA H, *et al*, 2015. Effect of oceanographic parameters on daily albacore catches in the Northeast Atlantic [J]. Deep Sea Research Part II: Topical Studies in Oceanography, 113: 73-80.

HWANG W J, WEN K W, 1998. Fast KNN classification algorithm based on partial distance search [J]. Electronics Letters, 34(21): 2062-2063.

JOU Y J, HUANG C C L, CHO H J, 2014. A VIF-based optimization model to alleviate collinearity problems in multiple linear regression [J]. Computational Statistics, 29(6): 1515-1541.

JUFAILI S A, PIONTKOVSKI S A, 2019. Seasonal and interannual variations of Yellowfin tuna catches along the Omani Shelf [J]. International Journal of Oceans and Oceanography, 13(2): 427-454.

KRISTENSEN P, JUDGE M E, THIM L, *et al*, 1998. Hypothalamic CART is a new anorectic peptide regulated by leptin [J]. Nature, 393(6680): 72-76.

LEHODEY P, SENINA I, NICOL S, *et al*, 2015. Modelling the impact of climate change on South Pacific albacore tuna [J]. Deep Sea Research Part II: Topical Studies in Oceanography, 113: 246-259.

OU J J, JIN X D, MA I, *et al*, 2002. CMOS RF modeling for GHz communication IC's [C] // Proceedings of 1998 Symposium on VLSI Technology Digest of Technical Papers. Honolulu, HI, USA: IEEE: 94-95.

PICKETT M H, SCHWING F B, 2006. Evaluating upwelling estimates off the west coasts of North and South America [J]. Fisheries Oceanography, 15(3): 256-269.

RAYKOV T, PENEV S, 2010. Testing multivariate mean collinearity via latent variable modelling [J]. British Journal of Mathematical and Statistical Psychology, 63(3): 481-490.

SUYKENS J A K, LUKAS L, VANDEWALLE J, 2000. Sparse approximation using least squares support vector machines [C] // Proceedings of 2000 IEEE International Symposium on Circuits and Systems. Geneva, Switzerland: IEEE: 757-760.

TUSSADIAH A, PRANOWO W S, SYAMSUDDIN M L, *et al*, 2018. Characteristic of eddies kinetic energy associated with yellowfin tuna in Southern Java Indian Ocean [J]. IOP Conference Series: Earth and Environmental Science, 176(1): 012004.

WILLIAMS A J, ALLAIN V, NICOL S J, *et al*, 2015. Vertical behavior and diet of albacore tuna (*Thunnus alalunga*) vary with latitude in the South Pacific Ocean [J]. Deep Sea Research Part II: Topical Studies in Oceanography, 113: 154-169.

ZAINUDDIN M, SAITOH K, SAITOH S I, 2008. Albacore

(*Thunnus alalunga*) fishing ground in relation to oceanographic conditions in the western North Pacific Ocean using remotely sensed satellite data [J]. Fisheries Oceanography, 17(2): 61-73.

ZHANG S C, LI X L, ZONG M, *et al*, 2018. Efficient KNN classification with different numbers of nearest neighbors [J]. IEEE Transactions on Neural Networks and Learning Systems, 29(5): 1774-1785.

COMPARISON ON FISHING GROUND FORECAST MODELS OF *THUNNUS ALALUNGA* IN THE TROPICAL WATERS OF ATLANTIC OCEAN

SONG Li-Ming[1, 2], REN Shi-Yu[1], HONG Yi-Ran[1], ZHANG Tian-Jiao[1], SUI Heng-Shou[3], LI Bin[3], ZHANG Min[1, 2]

(1. *College of Marine Sciences, Shanghai Ocean University*, *Shanghai* 201306, *China*; 2. *National Engineering Research Center for Oceanic Fisheries*, *Shanghai* 201306, *China*; 3. *CNFC Overseas Fisheries Co*, *Ltd*, *Beijing* 100032, *China*)

Abstract To improve the accuracy of the forecast model for albacore tuna (*Thunnus alalunga*) fishing ground in the tropical waters of Atlantic Ocean, seven fishery forecast models, *e.g.* k-nearest neighbor (KNN), logistic regression (LR), classfication and regression tree (CART), support vector machine (SVM), random forest (RF), gradient boosting decision tree (GBDT), and stacking ensemble learning (STK) model were used and compared based on the data of 13 tuna longliners of Chinese fishing enterprises from 2016 to 2019 in the high seas of the Atlantic Ocean (19°16′S~16°21′N; 46°27′W~2°09′E). Using environmental factors (temperature, salinity and dissolved oxygen) at different water layers from 0 to 500 m, as well as chlorophyll-*a* concentration, sea surface wind speed, eddy kinetic energy, and mixed layer depth, the relationship between albacore tuna CPUE and the environmental factors were analyzed. Seventy-five percent of the data were taken as training data and 25% as test data. The performance of each model was evaluated by prediction accuracy (ACC) and area under receiver operating characteristic curve (AUC). Relationships between CPUE (catch per unit of effort) and marine environmental factors were established. Results show that: (1) the prediction performance of STK model was obviously better compared with other models and its ACC and AUC is 75.92% and 0.742, respectively; (2) the areas of central fishing ground predicted by STK model for albacore tuna is consistent with the actual fishing ground generally; (3) the marine environmental factors that affect the distribution of albacore tuna fishing grounds in the Atlantic Ocean included mainly temperature and salinity of 100 m layer, and dissolved oxygen at 100, 150, and 500 m layer. The accuracy and the prediction performance of the STK model is high for albacore tuna fishing ground forecast in the tropical waters of Atlantic Ocean.

Key words *Thunnus alalunga*; fishing ground forecast model; comparative study of model performance; tropical waters of Atlantic Ocean