水产学报
JOURNAL OF FISHERIES OF CHINA
DOI: 10.11964/jfc.20210312692
Science Press

基于集成学习的大西洋热带水域大眼金枪鱼渔情预报

宋利明[1,2*]，任士雨[1]，张　敏[1,2]，隋恒寿[3]
(1. 上海海洋大学海洋科学学院，上海　201306；
2. 上海海洋大学，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：为提高大西洋大眼金枪鱼渔场预报模型的准确率，实验利用13艘中国延绳钓渔船2013—2019年的渔捞日志数据和对应的海洋环境数据(海表面风速、叶绿素*a*浓度、涡动能、混合层深度和0~500 m水层的垂直温度、盐度和溶解氧等)，以天为时间分辨率、2°×2°为空间分辨率、以数据集的75%为训练数据建立了K最近邻(KNN)、逻辑斯蒂回归(LR)、分类与回归树(CART)、支持向量机(SVM)、人工神经网络(ANN)、随机森林(RF)、梯度提升决策树(GBDT)和Stacking集成(STK)渔情预报模型，以25%的测试数据进行模型性能测试、比较。结果显示，(1) STK (由KNN、RF、GBDT模型集成)模型的大眼金枪鱼渔场预报性能较KNN、LR、CART、SVM、ANN、RF和GBDT模型有所提高且相对稳定，上述模型对应的准确率和ROC曲线下面积(AUC)依次分别为81.62%、0.781，79.44%、0.778，72.81%、0.685，74.84%、0.717，73.67%、0.702，67.70%、0.500，80.96%、0.780和78.13%、0.747；(2) STK模型预测的中心渔场与实际中心渔场基本吻合，主要在5°N~10°N，33°W~43°W海域附近；(3)影响大西洋大眼金枪鱼渔场分布的海洋环境因子主要有300 m水层的溶解氧、500 m水层的盐度、海面风速和混合层深度，相对重要性分别为13.24%、9.12%、9.12%和8.81%。研究表明，STK模型对大西洋大眼金枪鱼渔场的预报准确率较高。

关键词: 大眼金枪鱼；延绳钓；渔情预报；集成学习；大西洋
中图分类号: S 934　　**文献标志码:** A

大眼金枪鱼(*Thunnus obesus*)广泛分布于太平洋、大西洋和印度洋较深水层的热带及亚热带水域[1]，是中国金枪鱼延绳钓渔业的主捕对象之一。目前，大眼金枪鱼的研究更倾向于分析渔场分布与海洋环境因子之间的关系。如樊伟等[2]分析大西洋大眼金枪鱼渔场的环境特征和空间分布，发现大眼金枪鱼的偏好水温范围为26~27 °C，集中分布在赤道两侧南北纬15°范围内的热带海域；杨胜龙等[3-4]基于Argo数据分析印度洋和中西太平洋大眼金枪鱼渔场分布和温跃层的关系，发现大眼金枪鱼主要在温跃层以下活动；Setiawati等[5]基于GAM模型研究发现，海表温度对印度洋爪哇南部海域和巴厘岛南部海域大眼金枪鱼渔场分布的影响较大。Schaefer等[6]认为赤道东太平洋大眼金枪鱼主要生活在200~300 m水层。周为峰等[7]利用贝叶斯分类器结合海表温度和海面高度

收稿日期：2020-12-10　　**修回日期：**2021-06-07
资助项目：国家重点研发计划(2020YFD0901205)；2016年农业农村部海洋渔业资源调查与探捕项目(D-8006-16-8045)
通信作者：宋利明(照片)，从事渔业捕捞学研究，E-mail：lmsong@shou.edu.cn

对印度洋大眼金枪鱼渔场进行预测，其准确率达到 65.96%；杨胜龙等[8]利用海表温度和叶绿素 *a* 浓度基于空间叠加方法构建了西北印度洋大眼金枪鱼预报模型；沈智宾等[9]、冯波等[10]和宋利明等[11]结合各海洋环境因子建立了大眼金枪鱼栖息环境综合指数模型。传统的金枪鱼渔情预报方法大多采用单一模型[7-15]和同质模型[16]，由于海洋环境复杂多变，使得单一模型出现过拟合现象，降低模型的泛化能力[17]。Stacking 集成学习是一种多模型的集成方法，能够利用多个学习器解决同一个问题，从而得到比单模型更准确的结果，具有结构简单、性能高、分类能力强等特点[18]。本研究根据 2016—2019 年中国捕捞船队的大西洋延绳钓渔业数据，结合海表温度、风速、叶绿素 *a* 浓度、涡动能、混合层深度以及 0~500 m 水层的垂直温度、盐度和溶解氧等环境因子，利用 K 最近邻 (KNN)、逻辑斯蒂回归 (LR)、分类与回归树 (CART)、支持向量机 (SVM)、人工神经网络 (ANN)、随机森林 (RF)、梯度提升决策树 (GBDT) 和 Stacking 集成学习 (STK) 方法建立大西洋大眼金枪鱼渔场预报模型，并比较得出预测能力最好的模型，为今后金枪鱼渔情预报模型的选用提供参考。

1　材料与方法

1.1　数据来源与匹配

研究用延绳钓渔业数据由中国捕捞船队提供，数据包括 13 艘渔船的船名、作业日期 (年/月/日)、作业位置 (经度、纬度) 和渔获信息 (渔获种类、渔获尾数和下钩数) 等。根据 13 艘渔船的作业范围和时间确定研究区域为 14°20′S~15°20′N，47°38′W~2°30′E，时间跨度为 2016—2019 年。海洋环境数据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)，数据包括海面风速 (WS)、叶绿素 *a* 浓度 (Chl-*a*)、涡动能 (EKE)、混合层深度 (MLD)、垂直温度 (T，其中，T0、T50、T100、T150、T200、T250、T300、T400、T500 分别表示 0、50、100、150、200、250、300、400、500 m 水深处的温度)、垂直盐度 (S，其中，S100、S200、S300、S400、S500 分别表示 100、200、300、400、500 m 水深处的盐度) 和垂直溶解氧 (D，其中，D0、D50、D100、D150、D200、D250、D300、D350、D400、D450、D500 分别表示 0、50、100、150、200、250、300、350、400、450、500 m 水深处的溶解氧)，共 29 个海洋环境因子。环境数据的时间分辨率为 d，空间分辨率为 0.25°×0.25°。鉴于延绳钓渔业的特性 (投绳开始到结束的位置跨度达到甚至超过 60 海里)，为了尽可能覆盖延绳钓作业的全部范围，保证数据的有效性，本实验将环境因子和渔业数据的空间分辨率统一为 2°×2°。使用 MATLAB 将某天某网格的大眼金枪鱼单位捕捞努力量渔获量 (CPUE，尾/千钩) 与当天该网格的海洋环境数据进行匹配。

1.2　数据处理

CPUE 的计算方法　将每天的大眼金枪鱼的渔获尾数划分到 2º×2º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，得出每天每个网格内的大眼金枪鱼 CPUE，计算各网格内 CPUE[19]：

$$\mathrm{CPUE}_{(i,j)}=\frac{F_{(i,j)}}{H_{(i,j)}}\times 1\,000 \tag{1}$$

式中，$\mathrm{CPUE}_{(i,j)}$、$F_{(i,j)}$、$H_{(i,j)}$ 分别表示在第 *i* 经度、*j* 纬度的网格内的 CPUE、尾数和钓钩数量。

CPUE 与各海洋环境因子的相关性分析　利用 SPSS 软件分析大眼金枪鱼 CPUE 与 29 个海洋环境因子的 Spearman 系数，得出与大眼金枪鱼 CPUE 具有相关性的海洋环境因子。

多变量共线性分析　对与大眼金枪鱼 CPUE 具有相关性的海洋环境因子进行多变量共线性分析。方差膨胀因子 (*VIF*) 是各变量之间存在多重共线性时的方差与不存在多重共线性时的方差的比值，是度量海洋环境因子是否存在共线性的重要指标[20]。计算方程：

$$y=a_1x_1+a_2x_2+a_3x_3+\cdots+a_ix_i+\cdots a_nx_n+\beta \tag{2}$$

式中，*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。第 *i* 个自变量的 *VIF* 计算公式：

$$VIF=\frac{1}{1-R_k^2} \tag{3}$$

式中，R_k 为线性方程中的决定系数。当 $|VIF|\geqslant 10$ 时，表示环境因子之间存在多重共线性，反之，各环境因子之间不存在共线性[20]。据此，逐步剔除环境因子中 *VIF* 值最高的因子，重复此过程直至剩下的环境因子 *VIF* 值小于 10。使用 *VIF* 值小

于 10 的环境因子建立渔情预报模型。

相对重要性分析　基于多重共线性分析的结果，计算无共线性环境因子与 CPUE 之间的 Spearman 相关系数，根据其相关系数得出每个环境因子对大西洋大眼金枪鱼渔场分布的相对重要性，相对重要性计算公式：

$$RE_i = \frac{|R_i|}{\sum_{i=m} |R_i|} \tag{4}$$

式中，RE_i 表示第 i 个环境因子的相对重要性，R_i 表示第 i 个环境因子与 CPUE 之间的 Spearman 相关系数，m 表示无共线性环境因子的总数。

1.3　模型建立

模型建立及参数选择方法　为建立渔情预报模型，将 CPUE 按照三分位数法划分两类，CPUE 大于 10 (第一三分位数) 的为高 CPUE 渔区，CPUE 小于 10 的为低 CPUE 渔区，将大西洋大眼金枪鱼原始数据集的 75% 和余下的 25% 分别划分为训练数据 L 和测试数据 T。系列模型的建立方法和参数选择方法：(1) KNN 是通过计算不同数据之间特征值进行分类的方法，距离为欧氏距离，近邻个数为 7；(2) LR 是通过线性回归模型的预测结果去逼近真实标记的对数几率；(3) CART 是通过计算决策树中各节点的 Gini 不纯度指标，对样本采集采用二分递归的分割，其复杂度为 0.01，最大深度为 30；(4) SVM 的主要原理是找到一个能够将所有数据样本划分开的超平面，使得样本集中的所有数据到这个超平面的距离最短，其核函数为高斯核函数，惩罚系数为 1；(5) ANN 是由大量神经单元以某种方式相互连接而成的网络拓扑结构，其隐藏层数为 1，学习率为 0.1，激活函数为 Sigmoid；(6) RF 和 GBDT 分别是一种基于分类决策树的 Bagging 和 Bosting 集成学习方法。其中 RF 子叶点数上最小样本数量为 1，分割内部节点最小样本数量为 1，决策树个数为 500；(7) GBDT 树的个数为 100，树的深度为 6，学习率为 0.1；(8) STK 是通过训练多个分类器，按照某种结合策略得到一个更强的分类器。算法框架如图 1 所示，首先从 7 个单模型中选择预测能力较强的 3 个模型作为初级学习器，将结构简单的 LR 作为次级学习器；其次将 75% 的训练样本随机划分为 K 个数量相同的训练集子集 $L_1, L_2, \ldots, L_K$，取 T_i 作为测试子集，$L_{k-1}\{L_1, L_2, \ldots, L_k\}(L_i \notin L_{k-1})$ 作为 KNN、RF 和 GBDT 的训练子集，接着将每个模型的预测结果合并作为 LR 的训练集 L'，LR 的测试集 T'为每个单模型对测试集 T 的预测结果，最终输出预测结果。

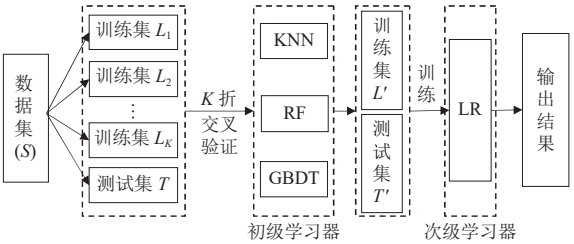

图 1　Stacking 集成学习方法

Fig. 1　Method of Stacking ensemble learning

数据归一化处理　由于建立模型需要的海表温度、风速、叶绿素 a 浓度、涡动能、混合层深度以及垂直温度、盐度和溶解氧等海洋环境因子和渔业数据的量值单位不同，数值范围差别较大，为了防止小数值量被大数值量淹没，同时提高模型的运行效率[21]，因此对所有数据进行归一化处理，公式：

$$y_z = \frac{x - x_n}{x_m - x_n} \tag{5}$$

式中，y_z 为归一化后的值，x 为实际值，x_m 为最大值，x_n 为最小值。

1.4　模型性能评价指标

接受者操作特征 (receiver operating characteristic, ROC) 曲线最初被用于雷达信号的检测，后用于模型预报性能评价。本研究中设定正确预测“高 CPUE 渔场”的个数 true positive (TP)，错误预测“高 CPUE 渔场”的个数为 false positive (FP)，正确预测“低 CPUE 渔场”的个数为 true negative (TN)，错误预测“低 CPUE 渔场”的个数为 false negative (FN)，将真正类率 true positive rate (TPR)作为纵坐标，假正类率 false positive rate (FPR) 作为横坐标绘制 ROC 曲线，其公式[22]：

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{7}$$

ROC 曲线与 FPR=0 围成的面积值则是 AUC 值，AUC 取值在 0~1 之间，值越大说明该模型的整体预测能力越好，反之预测能力越差[23]。准确率 (ACC) 作为模型性能评价指标，其公式：

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

分别将 25% 的测试数据的实际高 CPUE 渔区和实际低 CPUE 渔区与最优模型预测得到的高 CPUE 与低 CPUE 渔区进行叠加，并用 ArcGIS 软件画出其叠加图，以定性分析最优模型的预测性能。

1.5　中心渔场的确定

本实验将 25% 的测试站点的环境数据代入预测能力最优模型，计算得出“高 CPUE 渔区”的位置，利用 ArcGIS 软件的核密度分析工具计算并画出“高 CPUE 渔区”位置的密度分布图，将单位面积内高 CPUE 渔区个数大于 6.0 (高 CPUE 渔区个数/km^2) 的范围定义为中心渔场。

2　结果

2.1　大眼金枪鱼渔场分布

2016—2019 年大西洋大眼金枪鱼渔场主要分布在 15°N~15°S，3°E~50°W 的大西洋中部，其高 CPUE 渔区集中分布在 0°N~15°N，30°W~50°W 的海域 (图 2)。

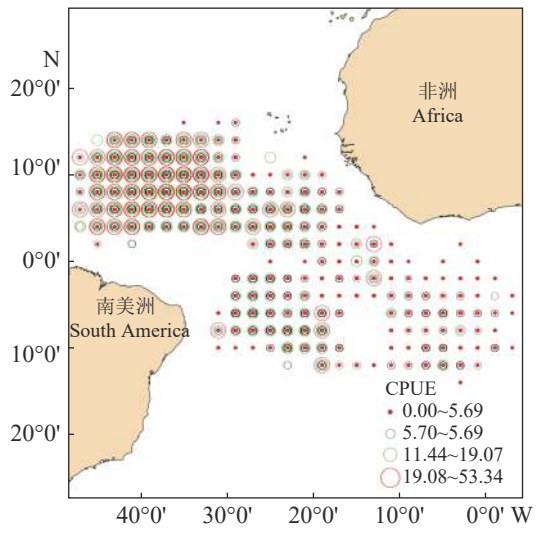

图 2　大西洋大眼金枪鱼渔场分布

Fig. 2　Distribution of fishing ground for *T. obesus* in Atlantic Ocean

2.2　CPUE 与各海洋环境因子的相关性分析

结果发现，29 个海洋环境因子均与大眼金枪鱼 CPUE 具有显著相关性 (P<0.01)，将用于多变量共线性分析 (表 1)。

2.3　共线性分析

环境变量之间的 Pearson 相关系数矩阵如图 3 所示，可以观察到深于 200 m 水层的溶解氧因子间大多具有高度相关性 (R>0.9)，200 m 水层以深的部分垂直温度和盐度之间也存在较大相关性 (R>0.8)，且 400 m 水层的温度和盐度相关系数达到 0.91，500 m 水层的温度和盐度的相关系数达到 0.93，此外，海表温度和海表溶解氧也具有较高的相关性。基于方差膨胀因子 (*VIF*) 选择环境因子 (表 2)，各变量的 *VIF* 值均小于 10，容差均大于 0.1。筛选出 21 个用于渔场预测的环境因子，分别是 Chl-a、MLD、WS、EKE、T0、T50、T100、T150、T200、T300、T500、D0、D50、

表 1　CPUE 与各环境因子的 Spearman 相关系数分析结果

Tab. 1　Results of Spearman correlation analysis on the relationship between CPUE and environmental variables

变量 variables	相关系数R correlation coefficient R	P
Chl-a	−0.129	0.000
MLD	0.312	0.000
WS	0.323	0.000
EKE	−0.105	0.000
T0	−0.028	0.014
T50	0.117	0.000
T100	0.188	0.000
T150	−0.057	0.000
T200	−0.276	0.000
T250	−0.284	0.000
T300	−0.189	0.000
T400	0.053	0.000
T500	0.232	0.000
D0	−0.093	0.000
D50	0.036	0.001
D100	−0.048	0.000
D150	0.059	0.000
D200	0.406	0.000
D250	0.473	0.000
D300	0.469	0.000
D350	0.440	0.000
D400	0.337	0.000
D450	0.268	0.000
D500	0.199	0.000
S100	0.193	0.000
S200	−0.127	0.000
S300	−0.040	0.000
S400	0.185	0.000
S500	0.323	0.000

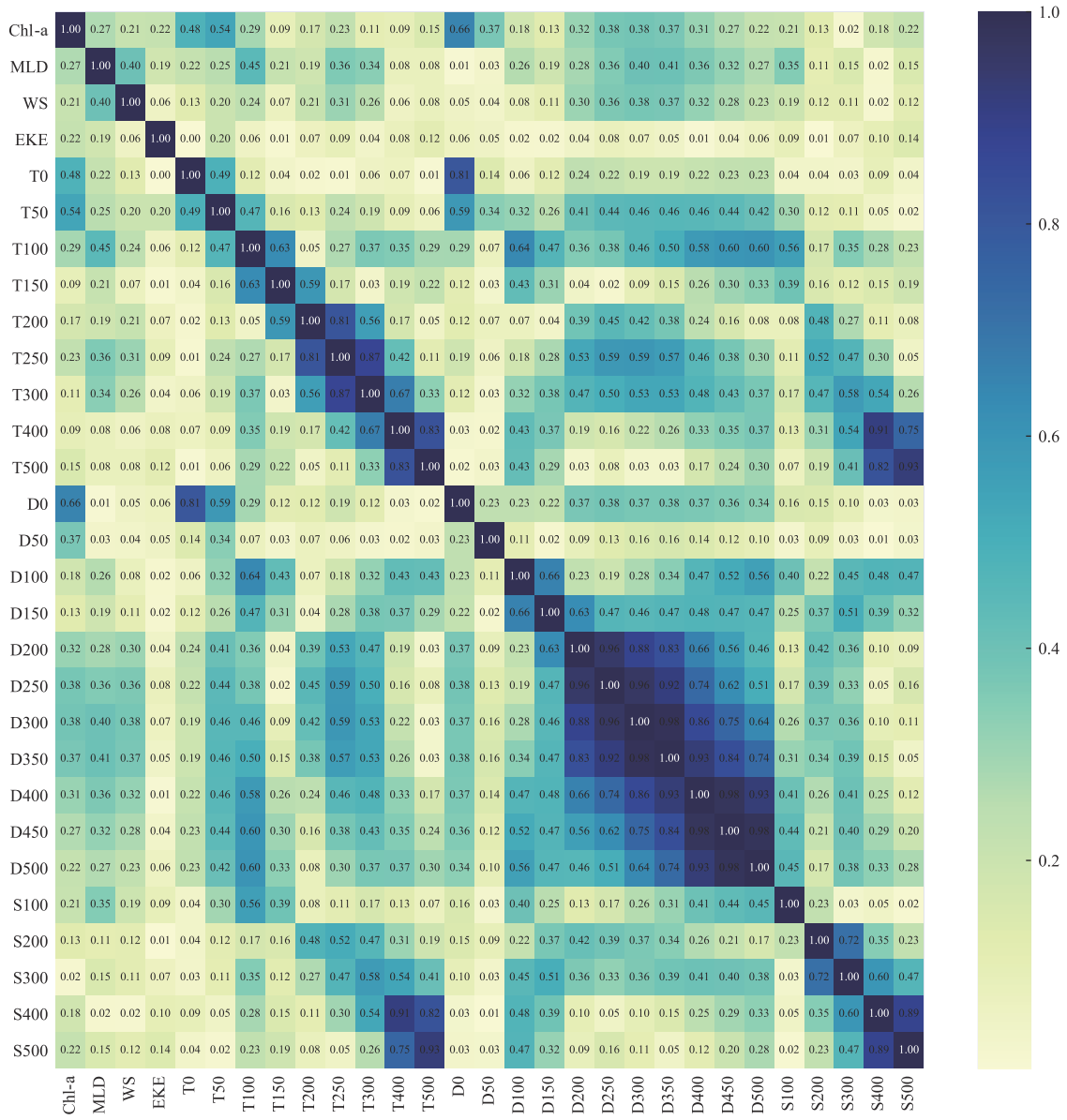

图 3　各环境变量间 Pearson 相关系数

Fig. 3　Pearson correlation coefficient among environmental variables

D100、D150、D300、D500、S100、S200、S300和 S400，大大降低了各因子间的相关性 (图 4)。

2.4　相对重要性分析

相对重要性分析结果显示，300 m 水层的溶解氧相对重要性最高，为 13.24%，海面风速、500 m 水层的盐度和混合层深度对大眼金枪鱼的影响也较大，相对重要性分别为 9.12%、9.12%和 8.81%，高于叶绿素 *a* 浓度和涡动能，海表温度的相对重要性最低，为 0.79% (图 5)。

2.5　预测结果及模型性能

各模型总体的预测精度和 ROC 曲线下面积 (AUC) 如表 3 所示。结果表明，KNN、RF、GBDT 的预报性能较好。因此，选择 KNN、RF、GBDT 模型建立 STK 模型。最终结果表明，STK 模型具有良好的预报性能，ACC 达到 81.62%，AUC 达到 0.781，其预测准确率比 KNN、LR、CART、SVM、ANN、RF、GBDT 等单一模型高出 2.18%、8.81%、6.78%、7.95%、13.92%、0.66% 和 3.49%，AUC 也均高于其他模型，但与 RF 的值相差很小。结果显示，STK 模型在高 CPUE 渔区的准确率为 72.10%，略低于 KNN 的 73.20%，但明显高于其他模型，且 STK 模型在低 CPUE 渔区的准确率最高，达到 86.84% (表 4)。将测试数据中的实际高

表 2　多重共线性诊断结果

Tab. 2　Results of multicollinearity diagnosis

变量 variables	VIF	容差 tolerance
Chl-a	2.560	0.391
MLD	1.972	0.507
WS	1.380	0.725
EKE	1.153	0.867
T0	3.993	0.250
T50	2.363	0.423
T100	4.200	0.238
T150	4.261	0.235
T200	4.595	0.218
T300	3.523	0.284
T500	3.619	0.276
D0	4.575	0.219
D50	1.349	0.741
D100	3.330	0.300
D150	2.554	0.392
D300	3.829	0.261
D500	3.132	0.319
S100	2.015	0.496
S200	3.313	0.302
S300	3.563	0.281
S400	5.215	0.192

CPUE 渔区与预测得到的高 CPUE 渔区进行叠加，二者分布基本一致 (图 6)，实际的高 CPUE 渔区主要分布在 4°N~15°N，30°W~47°W 的海域，0° W~30° W 海域的高 CPUE 渔区的误判率较高。将测试数据中的实际低 CPUE 渔区与预测得到的低 CPUE 渔区进行叠加，二者分布基本一致 (图 7)，低 CPUE 渔区分布比较均匀，且整体的误判率较低，说明 STK 模型的预测性能良好。

使用 ArcGIS 软件画出 25% 的测试数据中的“高 CPUE 渔区”位置密度分布 (图 8)，STK 模型预测得到的“高 CPUE 渔区”位置密度分布见图 9。从整体来看，预测得到的“高 CPUE 渔区”位置密度分布与实际“高 CPUE 渔区”位置密度分布基本吻合，说明 STK 模型的预测性能良好。

2.6　中心渔场分布

2016—2019 年 25% 的测试数据中高 CPUE 中心渔场集中分布在 5°N~10°N，33°W~43°W 海域附近 (图 8)，STK 模型预测得到的高 CPUE 渔区密度分布与实际中心渔场基本吻合 (图 9)。

3　讨论

3.1　模型对比分析

研究表明 STK 模型对整个渔场的预报准确率最高。这是因为 STK 算法采用了 *K* 折交叉验证，有效降低单一模型过拟合的问题。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 3 种模型的使用范围和优势。KNN 预报性能良好，但其预报性能很不稳定，当各个类别的样本数据不平衡时，预测精度则会降低，本研究中 KNN 在高 CPUE 渔区的准确率高于 STK 模型，主要因为高 CPUE 渔区的数据量较少，分布相对集中，使得样本数据在建模和预测过程中表现较好；LR 和 CART 简单易操作，但预测精度不高，容易出现过拟合问题，导致模型的泛化能力弱；SVM 对于核函数的高维映射解释力不强，尤其是径向基函数，并且对缺失数据敏感；ANN 是一种网络拓扑结构，由神经元之间相互连接，对数据量的要求很高，运行效率较低，对环境变量不具有可解释性；RF 可以高度并行化训练，大大缩短模型的训练时间，其使用的 Bootstrap 自助采样方法，使其具有良好的抗噪和泛化能力；GBDT 与 RF 相比，对参数的设置更为敏感，并且 GBDT 由于训练样本的单一性，会导致其泛化能力降低。研究结果也表明 RF 渔场预报效果优于 GBDT[24]。

3.2　共线性分析的必要性

海洋环境因子间相互关联[25]，不可避免地存在共线性，消除共线性可提高模型预测精度和运行速率。当环境因子之间的相关系数大于 0.7 时，说明存在多重共线性的可能性较大[26]。而多重共线性是影响 SVM 和 LR 预测精度和运行速率的主要原因之一[27-28]，虽然 CART、RF、ANN 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。

3.3　共线性分析结果的可靠性

筛选得出的非共线性环境因子影响大眼金枪

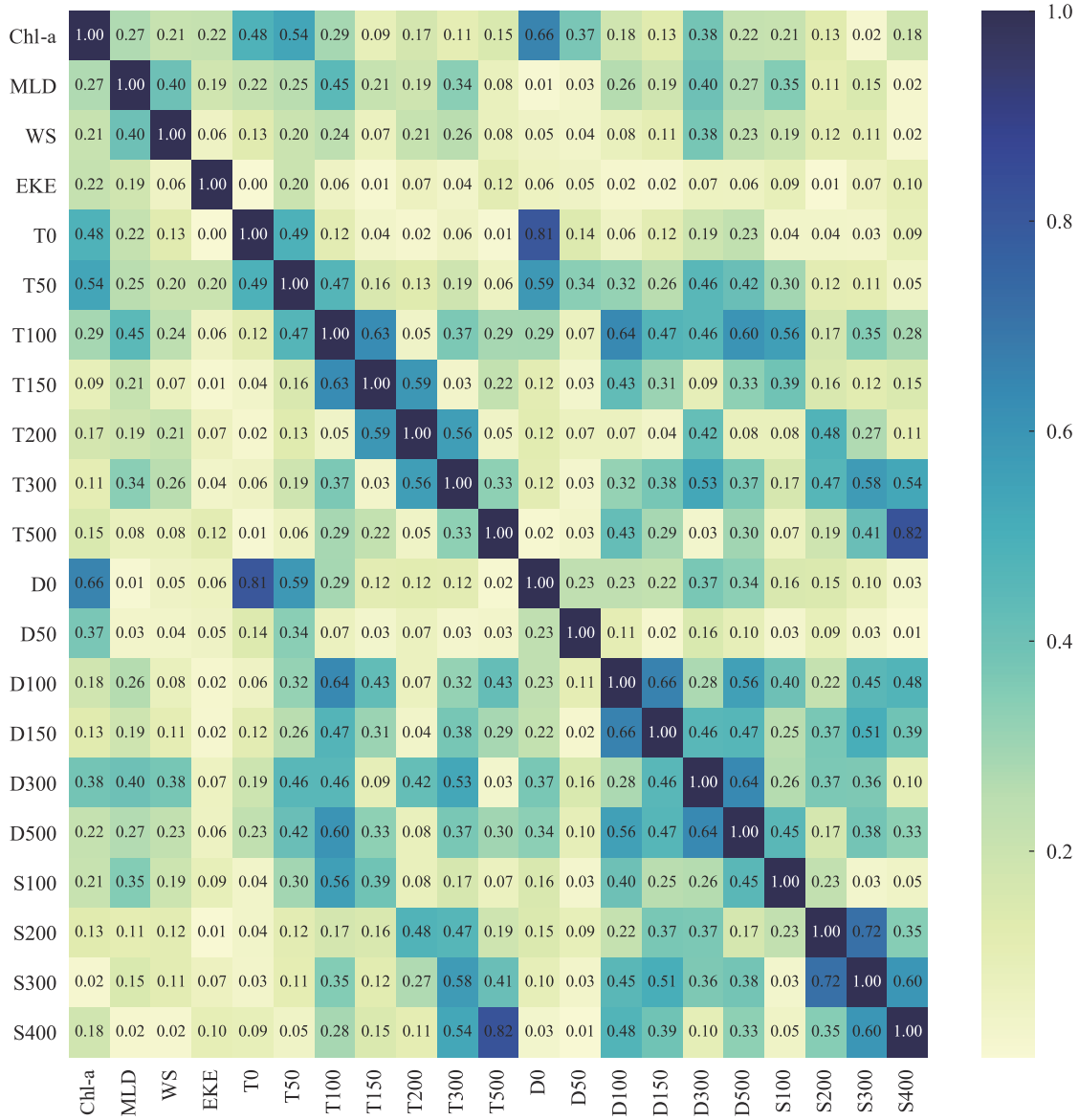

图 4　VIF 分析后各环境变量间 Pearson 相关系数

Fig. 4　Pearson correlation coefficient among environmental variables after VIF analysis

鱼的渔场分布，共线性分析结果可靠。大眼金枪鱼渔场分布主要受到饵料及海洋环境的影响。其中叶绿素 *a* 浓度与海洋中藻类等第一营养级生物的分布密切相关，是其他营养级的海洋生物获得能量的基础，通过叶绿素 *a* 浓度能够粗略判断大眼金枪鱼的空间分布[29]。海面风速能够改变叶绿素 *a* 的空间位置和浓度，从而影响大眼金枪鱼的渔场分布[30-31]，较低的风速有利于浮游植物的生长和繁衍，且浮游植物会随着海风改变其地理位置[32]。EKE 是通过影响环流、海洋温度以及叶绿素 a 的垂直和水平分布来影响黄鳍金枪鱼的资源丰度和渔场分布[33]。传统的金枪鱼渔情预报大多使用海表面环境因子[12,34]，忽略了不同水层的环境因子对金枪鱼类活动的影响，如 Matsumoto 等[35]认为温跃层会直接影响大眼金枪鱼的垂直空间分布，Cayré[36] 和宋利明等[37] 指出温度、溶解氧均是影响金枪鱼渔场分布的重要因子，但是在渔情预报模型构建的过程中却很少使用溶解氧作为环境变量。

3.4　环境因子相对重要性

溶解氧和温度是影响大眼金枪鱼垂直运动的主要原因[36]。研究表明，300 m 水层的溶解氧与大眼金枪鱼 CPUE 关系最为密切，其次是 500 m

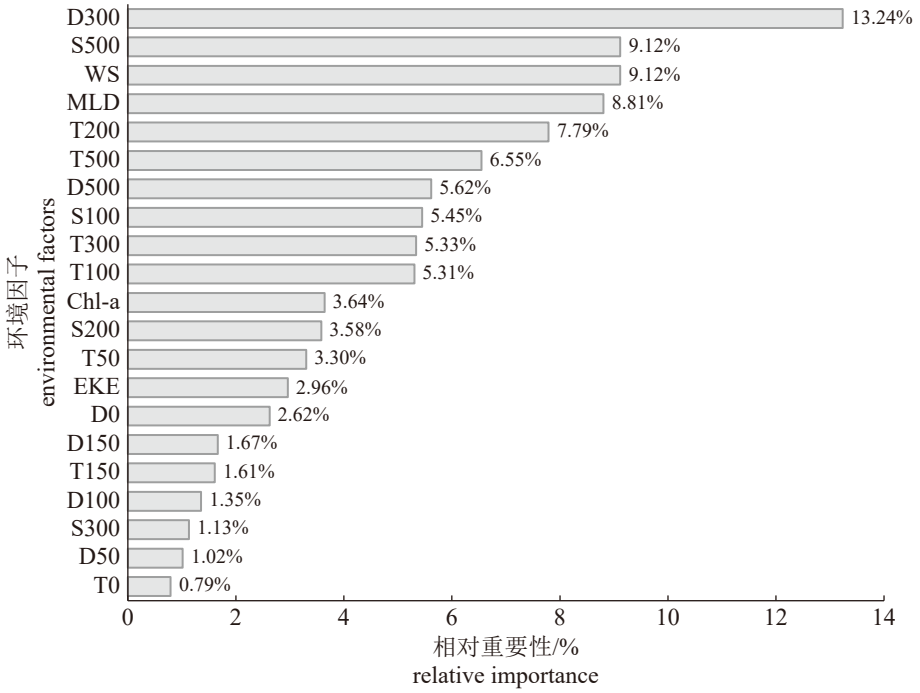

图 5　预测因子相对重要性

Fig. 5　Relative importance of the predictor variables

表 3　各个模型预测结果对比

Tab. 3　Comparison of forecast results of various models

模型 models	KNN	LR	CART	SVM	ANN	RF	GBDT	STK
ACC/%	79.44	72.81	74.84	73.67	67.70	80.96	78.13	81.62
AUC	0.778	0.685	0.716	0.701	0.631	0.780	0.746	0.781

表 4　各模型不同类别准确率比较

Tab. 4　Comparison of accuracy of each model in different categories　%

模型 models	KNN	LR	CART	SVM	ANN	RF	GBDT	STK
高CPUE渔区 high CPUE area	73.20	56.43	65.52	60.35	58.31	69.28	65.05	72.10
低CPUE渔区 low CPUE area	82.42	80.63	80.03	80.03	72.18	86.76	84.44	86.84

水层的盐度、海面风速和混合层深度。这可能与大眼金枪鱼的垂直活动有关，通常大眼金枪鱼白天会下潜至深约 450 m 的水层觅食，导致 300 m 处的溶解氧含量对其分布具有重要的影响。此外，200 和 500 m 水层的温度对大眼金枪鱼 CPUE 的影响比其他水层的温度影响更大，其中海洋表面温度的影响最小，这可能是因为大眼金枪鱼游泳速度较快，尤其是垂直方向的快速运动降低了海表温度和较浅水层的水温对金枪鱼分布的影响[38-39]。叶绿素 *a* 浓度对大眼金枪鱼分布的影响较小，这是由于较高的叶绿素 *a* 浓度会吸引更多浮游生物在附近繁殖，但大眼金枪鱼并不会直接捕食浮游生物[3]。500 m 水层的盐度对大眼金枪鱼的分布具有较大的影响，这与 Song 等[40]认为印度洋热带海域大眼金枪鱼的分布与盐度的关系有所不同，可能为海域不同造成的，但需进一步调查研究。

4　不足与展望

本研究使用了海面风速、叶绿素 a 浓度、涡

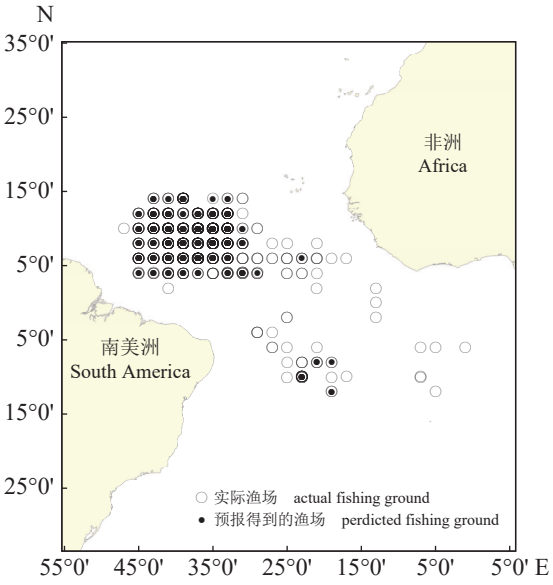

图 6　实际高 CPUE 渔区与预报得到的高 CPUE 渔区对比图

Fig. 6　Comparison of actual and predicted high CPUE areas

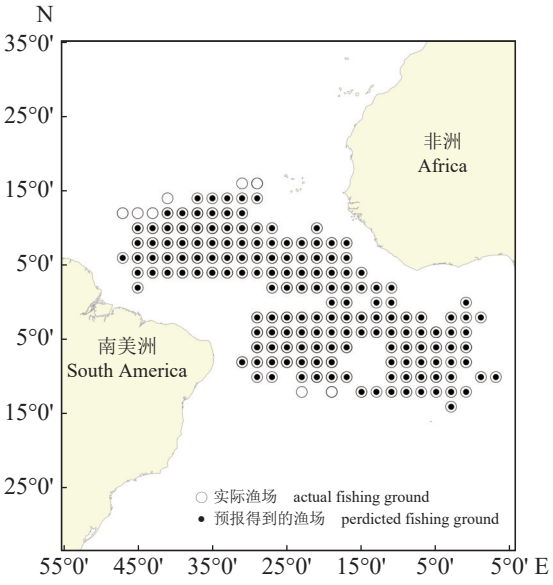

图 7　实际低 CPUE 渔区与预报得到的低 CPUE 渔区对比图

Fig. 7　Comparison of actual and predicted low CPUE areas

动能、混合层深度、温度、盐度和溶解氧等海洋环境因子建立大眼金枪鱼渔场预报模型，准确率较高，但是大眼金枪鱼分布还受到饵料生物的分布和垂直运动的影响，甚至全球气候变化的影响，今后还需进一步研究其对大眼金枪鱼渔场分布的影响。

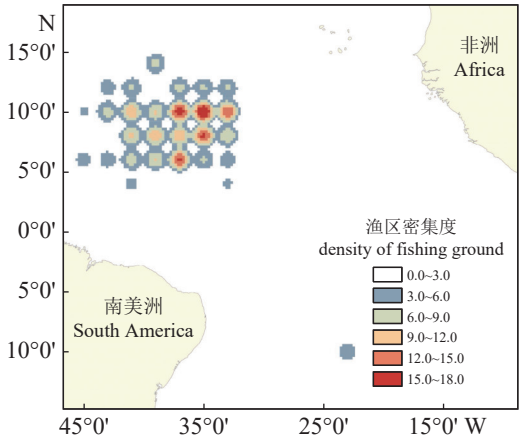

图 8　25 % 的测试数据中实际高 CPUE 渔区密度分布图

Fig. 8　Distribution map of actual fishing ground density for the high CPUE area in the 25% test data

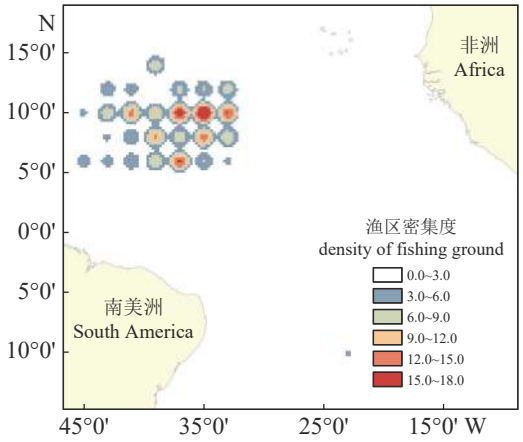

图 9　25 % 的测试数据中预测高 CPUE 渔区密度分布图

Fig. 9　Distribution map of predicted fishing ground density for the high CPUE area in the 25 % test data

感谢中水集团远洋股份有限公司宗文峰、叶少华、邓荣成和李彬先生的大力支持。

（作者声明本文无实际或潜在的利益冲突）

参考文献 (References):

[1] 宋利明, 赵海龙, 谢凯, 等. 库克群岛海域海洋环境因子对大眼金枪鱼渔获率的影响[J]. 水产学报, 2015, 39(8): 1230-1241.
Song L M, Zhao H L, Xie K, *et al*. Effects of environmental variables on catch rates of *Thunnus obesus* in waters near Cook Islands[J]. Journal of Fisheries of China, 2015, 39(8): 1230-1241 (in Chinese).

[2] 樊伟, 沈新强, 林明森. 大西洋大眼金枪鱼渔场、资源及环境特征的研究[J]. 海洋学报, 2003, 25(S2): 167-

176.
Fan W, Shen X Q, Lin M S. Study on resource, environment and fishing-ground of Atlantic bigeye tuna[J]. Acta Oceanologica Sinica, 2003, 25(S2): 167-176 (in Chinese).

[3] 杨胜龙, 张禹, 樊伟, 等. 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, 2012, 19(4): 679-689.
Yang S L, Zhang Y, Fan W, *et al*. Relationship between the temporal-spatial distribution of fish in bigeye tuna fishing grounds and the thermocline characteristics in the tropical Indian Ocean[J]. Journal of Fishery Sciences of China, 2012, 19(4): 679-689 (in Chinese).

[4] 杨胜龙, 伍玉梅, 张忭忭, 等. 中西太平洋大眼金枪鱼中心渔场时空分布与温跃层的关系[J]. 应用生态学报, 2017, 28(1): 281-290.
Yang S L, Wu Y M, Zhang B B, *et al*. Relationship between fishing grounds temporal-spatial distribution of *Thunnus obesus* and thermocline characteristics in the Western and Central Pacific Ocean[J]. Chinese Journal of Applied Ecology, 2017, 28(1): 281-290 (in Chinese).

[5] Setiawati M D, Sambah A B, Miura F, *et al*. Characterization of bigeye tuna habitat in the Southern Waters off Java-Bali using remote sensing data[J]. Advances in Space Research, 2015, 55(2): 732-746.

[6] Schaefer K M, Fuller D W. Movements, behavior, and habitat selection of bigeye tuna (*Thunnus obesus*) in the eastern equatorial Pacific, ascertained through archival tags[J]. Fishery Bulletin, 2002, 100(4): 765-788.

[7] 周为峰, 樊伟, 崔雪森, 等. 基于贝叶斯概率的印度洋大眼金枪鱼渔场预报[J]. 渔业信息与战略, 2012, 27(3): 214-218.
Zhou W F, Fan W, Cui X S, *et al*. Fishing ground forecasting of bigeye tuna in the Indian Ocean based on Bayesian probability model[J]. Fisheries Information & Strategy, 2012, 27(3): 214-218 (in Chinese).

[8] 杨胜龙, 周为峰, 伍玉梅, 等. 西北印度洋大眼金枪鱼渔场预报模型建立与模块开发[J]. 水产科学, 2011, 30(11): 666-672.
Yang S L, Zhou W F, Wu Y M, *et al*. Development of predictive model and module of fishing ground for bigeye tuna, *Thunnus obesus* and in the Northwest Indian Ocean[J]. Fisheries Science, 2011, 30(11): 666-672 (in Chinese).

[9] 沈智宾, 陈新军, 汪金涛. 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, 2015, 39(10): 45-51.
Shen Z B, Chen X J, Wang J T, *et al*. Forecasting of bigeye tuna fishing ground in the Eastern Pacific Ocean based on sea surface temperature and sea surface height[J]. Marine Sciences, 2015, 39(10): 45-51 (in Chinese).

[10] 冯波, 陈新军, 许柳雄. 多变量分位数回归构建印度洋大眼金枪鱼栖息地指数[J]. 广东海洋大学学报, 2009, 29(3): 48-52.
Feng B, Chen X J, Xu L X, *et al*. Multivariate quantile regression on habitat suitability index of *Thunnus obesus* in the Indian Ocean[J]. Journal of Guangdong Ocean University, 2009, 29(3): 48-52 (in Chinese).

[11] 宋利明, 高攀峰, 周应祺, 等. 基于分位数回归的大西洋中部公海大眼金枪鱼栖息环境综合指数[J]. 水产学报, 2007, 31(6): 798-804.
Song L M, Gao P F, Zhou Y Q, *et al*. Habitat environment integration index of *Thunnus obesus* in the high seas of the Central Atlantic Ocean based on the quantile regression[J]. Journal of Fisheries of China, 2007, 31(6): 798-804 (in Chinese).

[12] 毛江美, 陈新军, 余景. 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, 2016, 38(10): 34-43.
Mao J M, Chen X J, Yu J. Forecasting fishing ground of *Thunnus alalunga* based on BP neural network in the South Pacific Ocean[J]. Acta Oceanologica Sinica, 2016, 38(10): 34-43 (in Chinese).

[13] 宋利明, 周建坤, 沈智宾, 等. 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, 2017, 36(2): 195-208.
Song L M, Zhou J K, Shen Z B, *et al*. An integrated habitat index for albacore tuna (*Thunnus alalunga*) in waters near Cook Islands based on the support vector machine method[J]. Marine Science Bulletin, 2017, 36(2): 195-208 (in Chinese).

[14] 袁红春, 陈骢昊. 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, 2019, 46(5): 74-81.
Yuan H C, Chen C H. Prediction of *Thunnus alalunga* fishery based on fusion deep learning model[J]. Fishery Modernization, 2019, 46(5): 74-81 (in Chinese).

[15] 袁红春, 陈冠奇, 张天蛟, 等. 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,

2020, 36(2): 423-429.
Yuan H C, Chen G Q, Zhang T J, *et al*. Fishing ground forecast model of albacore tuna based on fully convolutional networks in the South Pacific[J]. Jiangsu Journal of Agricultural Sciences, 2020, 36(2): 423-429 (in Chinese).

[16] 陈雪忠, 樊伟, 崔雪森, 等. 基于随机森林的印度洋长鳍金枪鱼渔场预报[J]. 海洋学报, 2013, 35(1): 158-164.
Chen X Z, Fan W, Cui X S, *et al*. Fishing ground forecasting of *Thunnus alalung* in Indian Ocean based on random forest[J]. Acta Oceanologica Sinica, 2013, 35(1): 158-164 (in Chinese).

[17] Wolpert D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.

[18] 罗智青, 莫汉培, 王汝辉, 等. 基于Stacking模型融合的失压故障识别算法[J]. 能源与环保, 2019, 41(2): 41-45.
Luo Z Q, Mo H P, Wang R H, *et al*. Loss-of-voltage fault identification algorithm based on stacking model fusion[J]. China Energy and Environmental Protection, 2019, 41(2): 41-45 (in Chinese).

[19] Feng Y J, Chen X J, Gao F, *et al*. Impacts of changing scale on Getis-Ord Gi* hotspots of CPUE: a case study of the neon flying squid (*Ommastrephes bartramii*) in the northwest Pacific Ocean[J]. Acta Oceanologica Sinica, 2018, 37(5): 67-76.

[20] Van de Geer J P. Multivariate analysis of categorical data: theory[J]. Journal of Marketing Research, 1995, 32(1): 111-112.

[21] 袁红春, 胡光亮, 陈冠奇, 等. 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, 2019, 46(6): 96-103.
Yuan H C, Hu G L, Chen G Q, *et al*. Research on yield prediction methods of *Thunnus alalunga* in South Pacific based on particle swarm extension[J]. Fishery Modernization, 2019, 46(6): 96-103 (in Chinese).

[22] 张天蛟. 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, 2016.
Zhang T J. Ecological niche modeling and analysis of pelagic broadcast-spawning small fish[D]. Beijing: China Agricultural University, 2016 (in Chinese).

[23] 袁培森, 杨承林, 宋玉红, 等. 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, 2019, 50(11): 144-152.
Yuan P S, Yang C L, Song Y H, *et al*. Classification of rice phenomics entities based on stacking ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11): 144-152 (in Chinese).

[24] 侯娟, 周为峰, 樊伟, 等. 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, 2020, 16(5): 42-50.
Hou J, Zhou W F, Fan W, *et al*. Research on fishing grounds forecasting models of albacore tuna based on ensemble learning in South Pacific[J]. South China Fisheries Science, 2020, 16(5): 42-50 (in Chinese).

[25] He R Y, Ke C, Moore T, *et al*. Mesoscale variations of sea surface temperature and ocean color patterns at the Mid-Atlantic Bight shelfbreak[J]. Geophysical Research Letters, 2010, 37(9): L09607.

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

[27] 惠守博, 王文杰. 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, 2006, 27(8): 1385-1388.
Hui S B, Wang W J. Improvement of multi-variable's redundant attributes in classification algorithm of support vector machines[J]. Computer Engineering and Design, 2006, 27(8): 1385-1388 (in Chinese).

[28] 张玲. 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, 2010, 31(4): 19-20,38.
Zhang L. The test of multi-collinearity and the quantitative analysis of the degree of impact of prediction targets[J]. Journal of Tonghua Teachers College, 2010, 31(4): 19-20,38 (in Chinese).

[29] 朱国平, 李凤莹, 陈锦淘, 等. 印度洋中南部长鳍金枪鱼繁殖栖息的适应性[J]. 海洋环境科学, 2012, 31(5): 697-700,707.
Zhu G P, Li F Y, Chen J T, *et al*. Spawn habitat suitability for albacore (*Thunnus alaunga*) in South-central Indian Ocean[J]. Marine Environmental Science, 2012, 31(5): 697-700,707 (in Chinese).

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

[31] Al-Jufaili S A, Piontkovski S A. Seasonal and interannual variations of yellowfin tuna catches along the

Omani Shelf[J]. International Journal of Oceans and Oceanography, 2019, 13(2): 427-454.

[32] 赵娜, 王霄鹏, 李咏沙, 等. 黄渤海海域叶绿素a浓度时空特征分布及影响因子分析[J]. 科学技术与工程, 2020, 20(17): 7101-7107.
Zhao N, Wang X P, Li Y S, *et al*. Temporal-spatial distribution of Chlorophyll-a and impacts of environmental factors in the Bohai Sea and Yellow Sea[J]. Science Technology and Engineering, 2020, 20(17): 7101-7107 (in Chinese).

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

[34] 周为峰, 黎安舟, 纪世建, 等. 基于贝叶斯分类器的南海黄鳍金枪鱼渔场预报模型[J]. 海洋湖沼通报, 2018(1): 116-122.
Zhou W F, Li A Z, Ji S J, *et al*. Forecasting model for yellowfin tuna (*Thunnus albacares*) fishing ground in the South China Sea based on bayes classifier[J]. Transactions of Oceanology and Limnology, 2018(1): 116-122 (in Chinese).

[35] Matsumoto T, Kitagawa T, Kimura S. Vertical behavior of bigeye tuna (*Thunnus obesus*) in the northwestern Pacific Ocean based on archival tag data[J]. Fisheries Oceanography, 2013, 22(3): 234-246.

[36] Cayré P. Behaviour of yellowfin tuna (*Thunnus albacares*) and skipjack tuna (*Katsuwonus pelarnis*) around fish aggregating devices (FADs) in the Comoros Islands as determined by ultrasonic tagging[J]. Aquat Living Resource, 1991, 4(1): 1-12.

[37] 宋利明, 吕凯凯, 胡振新, 等. 吉尔伯特群岛海域延绳钓渔场大眼金枪鱼的环境偏好[J]. 海洋渔业, 2010, 32(4): 374-382.
Song L M, Lv K K, Hu Z X, *et al*. Environmental preferences of *Thunnus obesus* near Gilbert Islands: an application to longline fishery[J]. Marine Fisheries, 2010, 32(4): 374-382 (in Chinese).

[38] Zagaglia C R, Lorenzzetti J A, Stech J L. Remote sensing data and longline catches of yellowfin tuna (*Thunnus albacares*) in the equatorial Atlantic[J]. Remote Sensing of Environment, 2004, 93(1-2): 267-281.

[39] Maury O, Gascuel D, Marsac F, *et al*. Hierarchical interpretation of nonlinear relationships linking yellowfin tuna (*Thunnus albacares*) distribution to the environment in the Atlantic Ocean[J]. Canadian Journal of Fisheries and Aquatic Sciences, 2001, 58(3): 458-469.

[40] Song L M, Zhou J, Zhou Y Q, *et al*. Environmental preferences of bigeye tuna, *Thunnus obesus*, in the Indian Ocean: An application to a longline fishery[J]. Environmental Biology of Fishes, 2009, 85(2): 153-171.

Fishing ground forecasting of bigeye tuna (*Thunnus obesus*) in the tropical waters of Atlantic Ocean based on ensemble learning

SONG Liming [1,2*], REN Shiyu [1], ZHANG Min [1,2], SUI Hengshou [3]

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

Abstract: In order to improve the accuracy of bigeye tuna (*Thunnus obesus*) fishing ground forecast model in the tropical waters of Atlantic Ocean, a series of fishery forecast models were established based on the logbook data of 13 Chinese longliners from 2013 to 2019 and the corresponding marine environment data, *e.g.* sea surface wind speed, chlorophyll a concentration, eddy kinetic energy, upper boundary depth of thermocline, vertical temperature, salinity and dissolved oxygen in 0-500 m water layer. *T. obesus* CPUE was calculated based on the logbook data. The environmental factors related to *T. obesus* CPUE were screened out from 29 environmental factors by correlation analysis. The non-collinear environmental factors were selected by collinearity analysis based on the variance expansion factor (*VIF*) and used to build the bigeye tuna fishing ground prediction models. The Spearman correlation coefficients between non-collinear environmental factors and *T. obesus* CPUE were calculated and used to analyze the relative importance of the environmental factors to the *T. obesus* CPUE. These series of prediction models, *e g.* K-Nearest Neighbor (KNN), Logistic Regression (LR), Classification and Regression Tree (CART), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Stacking ensemble model (developed by KNN, RF and GBDT, STK) were built by using 75% of data and verified by using 25% of data. The time resolution of *T. obesus* CPUE and marine environment data was one day, and the spatial resolution was 2° × 2°. The performance of 8 models were evaluated by the area under the receiver operating characteristic curve (AUC) and prediction accuracy. The maps of the actual fishing ground and the predicted fishing ground were overlapped by ArcGIS and used to evaluate the performance of the best model. The central bigeye tuna fishing ground was determined by the nuclear density analysis tool of ArcGIS. The results show that (1) compared with the single model (KNN, LR, CART, SVM, ANN, RF and GBDT), the forecasting performance of *T. obesus* fishing ground of STK model was better and relatively stable. The accuracy (AUC) of the STK model, KNN, LR, CART, SVM, ANN, RF and GBDT were 81.62% (0.781), 79.44% (0.778), 72.81% (0.685), 74.84% (0.717), 73.67% (0.702), 67.70% (0.500), 80.96% (0.780), and 78.13% (0.747), respectively; (2) the distribution of central fishing ground predicted by STK model was basically consistent with the actual distribution of central fishing ground, all of them were mainly distributed in the area of 5 °N-10 °N, 33 °W-43 °W; (3) the marine environmental factors that affect the distribution of *T. obesus* fishing grounds in the Atlantic Ocean mainly included dissolved oxygen of 300 m layer, salinity of 500 m layer, sea surface wind speed and upper boundary depth of thermocline, and the relative importance were 13.24%, 9.12%, 9.12% and 8.81%, respectively. The results suggest that the accuracy of the STK model for *T. obesus* fishing ground forecast in the Atlantic Ocean is high.

Key words: *Thunnus obesus*; longline; fishing ground forecast; ensemble learning; Atlantic Ocean

Corresponding author: SONG Liming. E-mail: lmsong@shou.edu.cn

Funding projects: National Key R&D Program of China (2020YFD0901205); Marine Fishery Resources Investigation Project of the Ministry of Agriculture and Rural Areas in 2016 (D-8006-16-8045)