中国水产科学 2021 年 8 月, 28(8): 1069–1078
Journal of Fishery Sciences of China

研究论文

DOI: 10.12264/JFSC2020-0568

基于集成学习的大西洋热带海域黄鳍金枪鱼渔情预报

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. 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: 利用 2016—2019 年中国渔业企业在大西洋热带海域(14°20′S~15°20′N; 47°38′W~2°30′E) 13 艘延绳钓作业渔船渔捞日志记录的黄鳍金枪鱼(*Thunnus albacares*)单位捕捞努力量渔获量(CPUE)数据, 结合海表面风速、叶绿素 a 浓度、涡动能以及 0~500 m 水层的垂直温度、盐度等海洋环境变量和空间因子(经纬度)建立了一系列黄鳍金枪鱼渔场预测模型。模型的时间分辨率为天(d), 空间分辨率为 0.25°×0.25°。该系列模型利用 75%的数据训练得到朴素贝叶斯(NB)、k 最近邻(KNN)、随机森林(RF)、分类与回归树(CART)、逻辑斯蒂回归(LR)、支持向量机(SVM)、梯度提升决策树(Xgboost)和 stacking 集成(由 NB、CART 和 LR 模型集成, STK)模型, 将 25%测试数据代入系列模型进行验证, 结果显示: (1) 黄鳍金枪鱼渔场预测准确率(ACC)分别为 61.62%、62.03%、66.37%、63.0%、63.26%、64.97%、64.08%、68.72%; (2) 其对应的 ROC 曲线下面积(AUC)分别为 0.64、0.67、0.72、0.66、0.68、0.70、0.69、0.72; (3) STK 模型的预测准确率较其他的方法均有所提高。建议使用 STK 模型对大西洋热带海域黄鳍金枪鱼渔场进行预测。

关键词: 黄鳍金枪鱼; 渔情预报; 集成模型; 大西洋
中图分类号: S972 **文献标志码:** A **文章编号:** 1005–8737–(2021)08–1069–10

黄鳍金枪鱼(*Thunnus albacares*)作为全球金枪鱼渔业第二大捕捞对象[1], 是大西洋热带海域延绳钓主要捕捞对象之一。准确的渔情预测可以大大减少寻找渔场的时间, 提高金枪鱼延绳钓捕捞效率。近年来, 国内外学者对海洋环境因子和金枪鱼渔场之间的关系做了大量研究, 如周为峰等[2]使用贝叶斯分类器对南海黄鳍金枪鱼渔场进行分类预测, 发现双环境因子比单环境因子的预测性能要高; 宋利明等[3]利用不同水层的环境因子, 建立长鳍金枪鱼(*Thunnus alalunga*)栖息环境综合指数模型; 毛江美等[4]搭建了 BP 人工神经网络, 利用海表温度、海面高度、月份以及经纬度等时空变量, 预测南太平洋长鳍金枪鱼渔场。随着金枪鱼渔业的不断发展, 生产数据和海洋环境数据越来越多, 对渔情预报精度要求越来越高, 新的机器学习方法被应用于金枪鱼渔情预报。为处理高维复杂的海洋环境因子对渔情预报精度的影响, 袁红春等[5-6]提出了一种融合深度学习模型 CNN-GRU-Attention 和全卷积神经网络, 实现了对长鳍金枪鱼 CPUE 的预测。集成模型是利用多个学习器解决同一问题, 得到比单模型更优异的结果, 目前主流的集成学习有 Bagging[7]、Boosting[8]和 Stacking[9], 在渔情预报方面预测效果较好的是 Bagging 和 Boosting 集成学习, 陈雪忠等[10]利用海表温度、叶绿素 a 浓度、表温距平、海表温梯度强度、叶绿素 a 浓度距平、海面高度异常等因子, 基于随机森林模型对印度洋大眼金枪鱼(*Thunnus obesus*)渔场进行预测; 高峰等[11]利

收稿日期: 2020-12-10; **修订日期:** 2021-01-03.
基金项目: 国家重点研发计划项目(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).
作者简介: 宋利明(1968–), 博士, 教授, 研究方向为捕捞学. E-mail: lmsong@shou.edu.cn

用提升回归树模型对东、黄海鲐(*Scomber japonicus*)渔场时空分布进行分析。单一模型和同质集成模型容易出现泛化能力不强、过拟合等问题[12], Bagging 和 Boosting 的每个基分类器都是同质弱分类器, Stacking 是利用异质分类器构建的算法, 具有结构简单、性能高、分类能力强等优点[13]。本研究根据 2016—2019 年大西洋热带海域黄鳍金枪鱼 CPUE 数据, 结合海表面风速、叶绿素 a 浓度、涡动能以及 0~500 m 水层的垂直温度和盐度、空间因子(经纬度)等数据, 利用朴素贝叶斯(NB)、k 最近邻(KNN)、随机森林(RF)、分类与回归树(CART)、逻辑斯蒂回归(LR)、支持向量机(SVM)、梯度提升决策树(Xgboost)和 Stacking 集成(由 NB、CART 和 LR 模型集成, STK)8 种模型预测大西洋黄鳍金枪鱼渔场, 并对比选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

渔业数据来自 2016—2019 年中水集团远洋股份有限公司 13 艘远洋延绳钓渔船渔捞日志, 包括船名、作业日期(年/月/日)、作业位置(经纬度)、渔获信息(鱼种、产量、尾数和下钩数), 研究海域范围为 14°20′S~15°20′N; 47°38′W~2°30′E。环境数据选用海表面风速、叶绿素 a 浓度、涡动能以及 0~500 m 水层的垂直温度和盐度。海表面风速数据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/)。其他数据来源于哥白尼海洋环境监测服务中心(Copernicus Marine Environment Monitoring, CMEMS)的网站(http://marine.copernicus.eu)。环境数据的时间分辨率为天, 空间分辨率为 0.25°×0.25°。使用 MATLAB 将某天某网格的黄鳍金枪鱼单位捕捞努力量渔获量(CPUE)与当天该网格的海洋环境数据进行匹配。

1.2 数据预处理

1.2.1 单位捕捞努力量渔获量(CPUE)计算 将每天的黄鳍金枪鱼的渔获尾数划分到 0.25°×0.25°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黄鳍金枪鱼 CPUE(尾/千钩), 计算各渔区内 CPUE 的公式为[14]:

$$\mathrm{CPUE}_{(i,j)}=\frac{F_{(i,j)}}{H_{(i,j)}}\times 1000 \tag{1}$$

式中, $\mathrm{CPUE}_{(i,j)}$、$F_{(i,j)}$、$H_{(i,j)}$分别表示在第 i 经度、j 纬度渔区的 CPUE、尾数和下钩数量。

1.2.2 CPUE 与各环境因子的相关性分析 采用 Python 的 seaborn 库来计算 CPUE 与各环境因子的 Pearson 相关系数并进行显著性检验(假设显著性水平为 0.05), 得出与 CPUE 具有相关性的因子为叶绿素 a 浓度(Chl a)、海表面风速(WS)、涡动能(EKE)、海表温度(T0)、50 m 水层温度(T50)、100 m 水层温度(T100)、200 m 水层温度(T200)、250 m 水层温度(T250)、300 m 水层温度(T300)、200 m 水层盐度(S200)、500 m 水层盐度(S500)和空间因子(经纬度)(表 1)。

1.2.3 非共线性海洋环境因子提取 筛选后的环境变量之间的 Pearson 相关系数矩阵如图 1 所示, 可以看出部分垂直温度和盐度之间相关性较大, 可能存在共线性问题。使用 SPSS 对 Chl a、WS、EKE、T0、T50、T100、T200、T250、T300、S200、S500 和空间因子(经纬度)进行共线性诊断分析, 计算方差膨胀因子(VIF)值, 如果 VIF 值小于 10, 认为该环境因子与其他环境因子不存在共线性[15], 结果如表 2 所示, VIF 值都小于 10。基于 VIF 的选择程序提取 Chl a、WS、EKE、T0、T50、T100、T300、S200、S500、LON 和 LAT 等 11 个变量, 各变量间的相关性大大降低(图 2)。

1.2.4 数据归一化 由于各环境数据和渔业数据的量级不同会对训练模型产生影响, 将每个环境变量和目标变量进行归一化[16], 其公式为:

$$x_m=\frac{x_n-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

式中, x_m、x_n、$x_{\max}$、$x_{\min}$ 分别为归一后的值、实际值、最大值和最小值。

1.3 建模方法

渔场分类时, 若 CPUE 等于 0, 为“非渔场”; CPUE 大于 0, 则为“渔场”。选取 2016—2019 年

表 1　黄鳍金枪鱼 CPUE 与各环境变量及时空因子的相关性分析结果

Tab. 1　Results of correlation analysis on the relationship between *Thunnus albacares* CPUE and environmental variables and spatio factors

变量符号 variable symbol	变量 variable	*P*	相关系数 correlation coefficient, *R*
LON	经度 longitude	0.000	−0.158
LAT	纬度 latitude	0.000	0.126
Chl a	叶绿素 a 浓度 chlorophyll a concentration	0.000	−0.078
WS	海表面风速 wind speed	0.000	0.050
EKE	涡动能 eddy kinetic energy	0.020	0.028
T0	海表温度 sea surface temperature	0.000	0.130
T50	50 m 水层温度 temperature under 50 m	0.000	0.101
T100	100 m 水层温度 temperature under 100 m	0.000	0.049
T150	150 m 水层温度 temperature under 150 m	0.362	0.008
T200	200 m 水层温度 temperature under 200 m	0.000	−0.047
T250	250 m 水层温度 temperature under 250 m	0.000	−0.057
T300	300 m 水层温度 temperature under 300 m	0.000	−0.033
T400	400 m 水层温度 temperature under 400 m	0.189	−0.012
T500	500 m 水层温度 temperature under 500 m	0.754	−0.003
S100	100 m 水层盐度 salinity under 100 m	0.395	−0.008
S200	200 m 水层盐度 salinity under 200 m	0.000	−0.036
S300	300 m 水层盐度 salinity under 300 m	0.089	−0.016
S400	400 m 水层盐度 salinity under 400 m	0.236	0.011
S500	500 m 水层盐度 salinity under 500 m	0.047	0.018

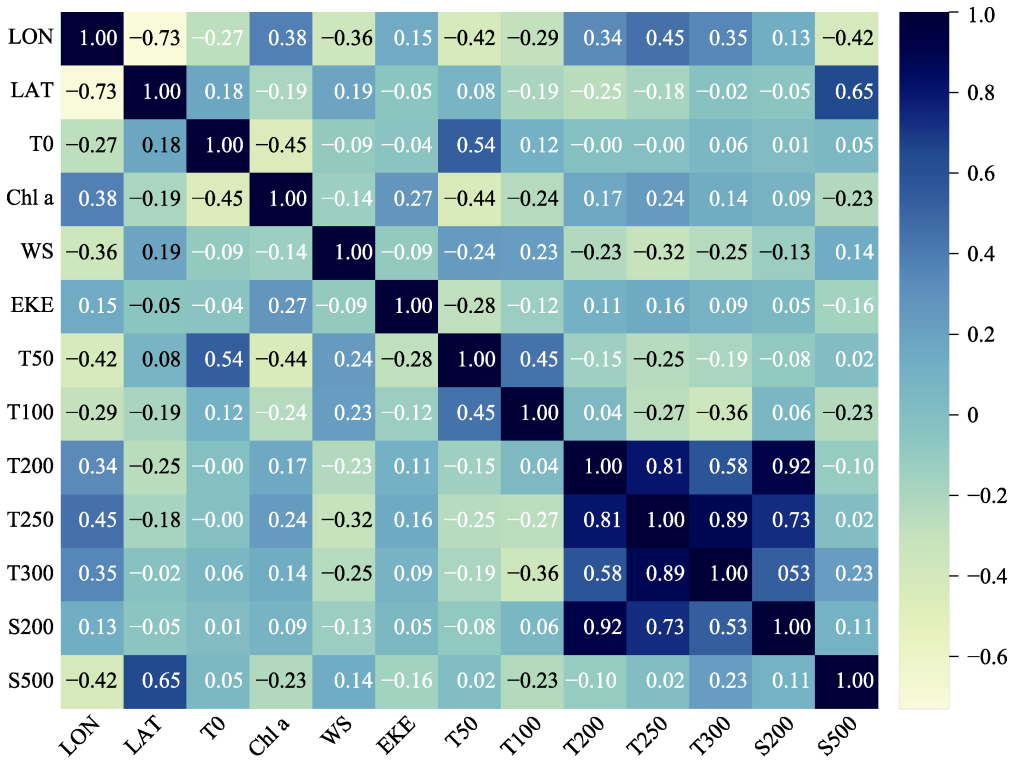

图 1　各环境变量间的 Pearson 相关系数

对各变量的解释见表 1.

Fig. 1　Pearson correlation coefficient among environmental variables

Explanation for the variables is shown in Tab.1.

表 2 多重共线性诊断结果

Tab. 2 Results of multicollinearity diagnosis

变量 variable	LON	LAT	Chl a	WS	EKE	T0	T50	T100	T300	S200	S500
VIF	4.96	4.55	1.68	1.31	1.20	1.92	2.22	2.11	2.21	1.61	2.22

注: 对各变量的解释见表 1.

Note: Explanation for the variables is shown in Tab.1.

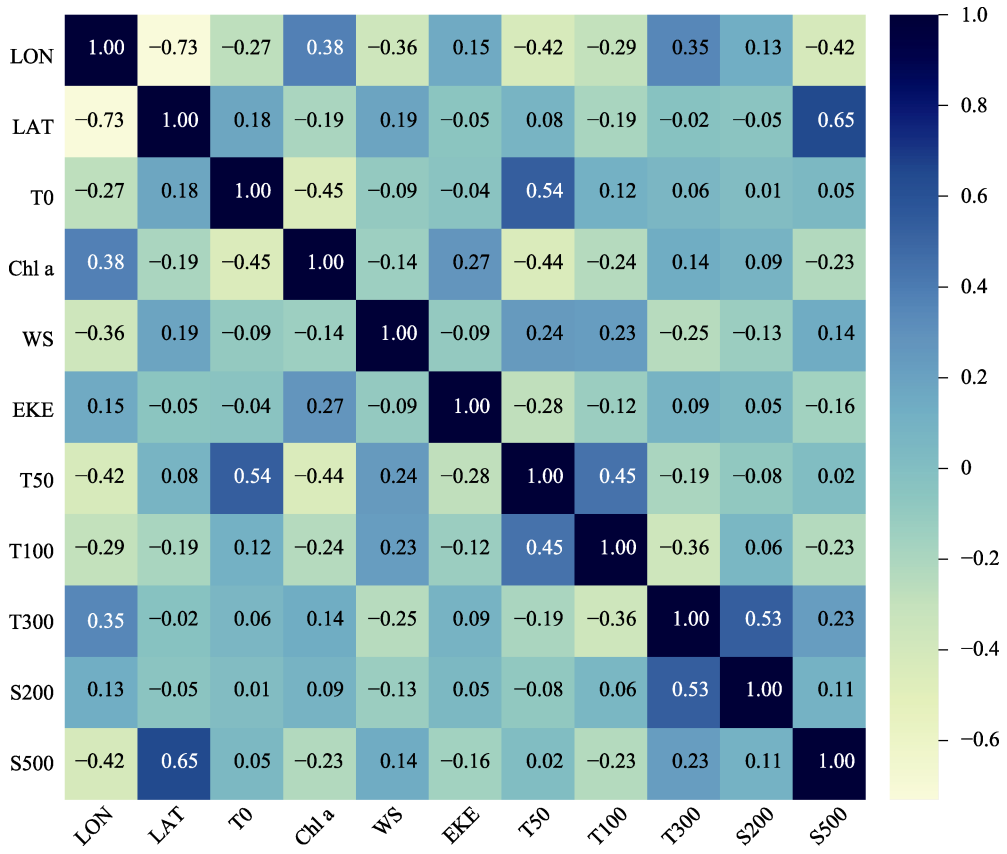

图 2 VIF 分析后各环境变量间的 Pearson 相关系数

对各变量的解释见表 1.

Fig. 2 Pearson correlation coefficient among environmental variables based on VIF analysis

Explanation for the variables is shown in Tab.1.

数据集的 75%作为训练集, 并使用训练集分别建立 NB、KNN、RF、CART、LR、SVM、Xgboost 和 STK8 种模型; 数据集的 25%为测试数据。朴素贝叶斯是利用贝叶斯原理结合先验概率和条件概率得到的后验概率; 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; KNN 是通过计算不同数据之间特征值进行分类的方法, 距离计算使用欧氏距离, 近邻个数为 7; 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; 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; RF 是一种基于决策分类树的 Bagging 集成学习方法。该模型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; Xgboost 是一种基于决策分类树的 Boosting 集成学习算法, 其树的个数为 100, 树的深度为 6, 学习率为 0.1。

本研究使用 Stacking 算法, 算法流程如图 3 所示, 首先利用原始数据采用 5 折交叉验证的方法训练 NB、CART 和 LR3 个基模型。然后将 3

个模型的预测结果构成新的数据集, 作为第二层模型 RF 的训练数据, 从而得到最终的预测结果。算法的具体步骤如下:

(1) 将大西洋热带海域黄鳍金枪鱼 CPUE 数据集 S 划分为训练集 D (75%)和测试集 T (25%)。

(2) 将训练集 D 按照 5 折交叉验证的方法随机均等划分为 D_1、D_2、D_3、D_4 和 D_5 五个子集, 依次选取其中一个子集 D_i (i=1, 2, 3, 4, 5)作为测试子集, 剩下的 4 份作为训练子集 D_j。使用训练子集 D_j 训练 NB、CART 和 LR 模型。对测试子集 D_i 进行预测, 合并每个模型的预测结果, 作为 RF 的训练集 D'。

(3) 每个基分类器对测试集 T 进行预测, 将预测结果作为 RF 的验证集。

(4) RF 从 NB、CART 和 LR 模型中得到新的训练集 D'和验证集 T', 训练 RF, 输出最终结果。

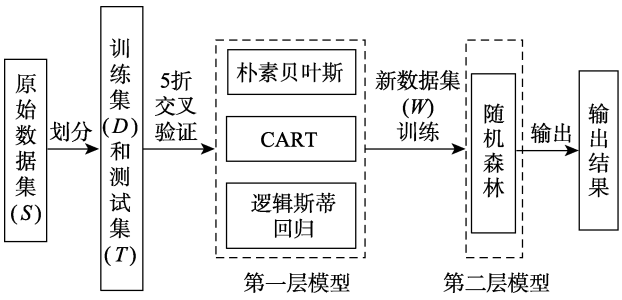

图 3　Stacking 集成学习方法

Fig. 3　Method of stacking ensemble learning

1.4　模型性能判断指标

本研究使用 ROC 曲线下面积(AUC)和准确率(ACC)作为模型性能评价指标。对于二分类问题, 会出现实际为正类预测是正类(TP)、实际为负类预测是正类(FP)、实际为正类预测是负类(FN)和实际为负类预测是负类(TN) 4 种结果。其中真正类率(true positive rate, TPR)和假正类率(false positive rate, FPR)的计算公式[17]为:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{4}$$

以 FPR 为横坐标, TPR 为纵坐标绘制 ROC 曲线, AUC 值是 ROC 曲线与横坐标围成的面积, 值域在 0~1 之间, AUC 的值越大说明该模型的预测性能越好。此外准确率(ACC)也是常用的评价模型预测能力的指标之一, 其公式为[18]:

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{5}$$

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好的模型, 计算得出“渔场”位置, 使用 Arcgis 软件画出“渔场”位置密度分布图, 把密度大于 5.6 个/km^2 的范围定义为中心渔场。

2　结果与分析

2.1　黄鳍金枪鱼 CPUE 的分布

2016—2019 年黄鳍金枪鱼 CPUE 的分布如图 4 所示, 黄鳍金枪鱼高 CPUE 主要分布在大西洋中部 0°N~15°N, 20°W~50°W 区域以及 0°N~10°S, 20°W~30°W 区域。

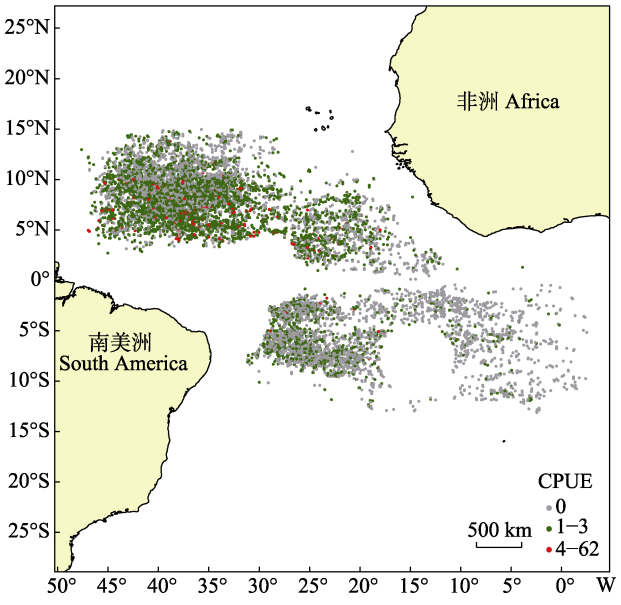

图 4　大西洋热带海域黄鳍金枪鱼渔场分布

Fig. 4　Distribution of fishing ground for *Thunnus albacares* in tropical waters of Atlantic Ocean

2.2　预测结果及模型性能

将 25%的测试数据分别代入 NB、KNN、RF、CART、LR、SVM、Xgboost 和 STK 模型。预测结果如表 3 所示, 与单一分类器相比较, STK 模型的预测能力最好, 其预测精度为 68.72%, 分别比

NB、KNN、RF、CART、LR、SVM、Xgboost 的预测精度高 7.10%、6.69%、2.35%、5.66%、5.46%、3.75%和 4.64%; AUC 为 0.72, 与 RF 模型相等, 比其他模型均高。各模型“渔场”和“非渔场”的准确率如表 4 所示, 其中 STK 模型对“渔场”分类的准确率为 68.91%, 对“非渔场”分类的准确率为 68.53%。结果显示无论是“渔场”的分类, 还是“非渔场”的分类, STK 模型都优于其他模型。将 25%测试数据的实际“渔场”和预测的“渔场”叠加, 如图 5 所示, 测试的“渔场”主要分布在 3°N~15°N, 20°W~47°W, 在 10°W 附近海域也有分布, 但是 10°W 附近海域对“渔场”的误判率很高; 将实际的“非渔场”和预测的“非渔场”叠加(图 6), 发现 0°~20°W 海域的“非渔场”预测准确率较高。

表 3 各个模型预测结果对比

Tab. 3 Comparison of forecast results of various models

指标 item	NB	KNN	RF	CART	LR	SVM	Xgboost	STK
AUC	0.64	0.67	0.72	0.66	0.68	0.70	0.69	0.72
ACC/%	61.62	62.03	66.37	63.06	63.26	64.97	64.08	68.72

表 4 各模型对不同类别渔场的判别准确率比较

Tab. 4 Comparison of discrimination accuracy for different fishing ground categories by using various models

类别 class	NB	KNN	RF	CART	LR	SVM	Xgboost	STK
渔场 fishing ground	60.07	63.81	67.22	65.68	61.76	66.44	65.58	68.91
非渔场 non-fishing ground	63.08	60.35	65.57	60.59	64.68	63.58	62.66	68.53

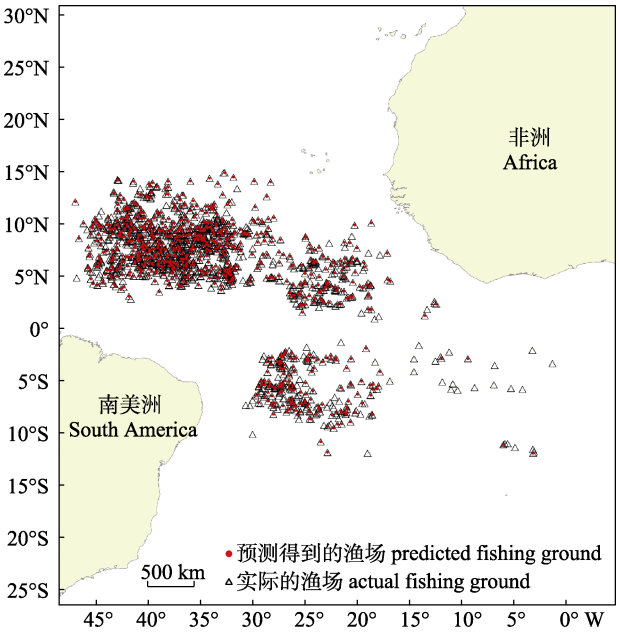

图 5 实际黄鳍金枪鱼渔场与预报得到的黄鳍金枪鱼渔场对比图

Fig. 5 Comparison of actual fishing ground and predicted fishing ground of *Thunnus albacares*

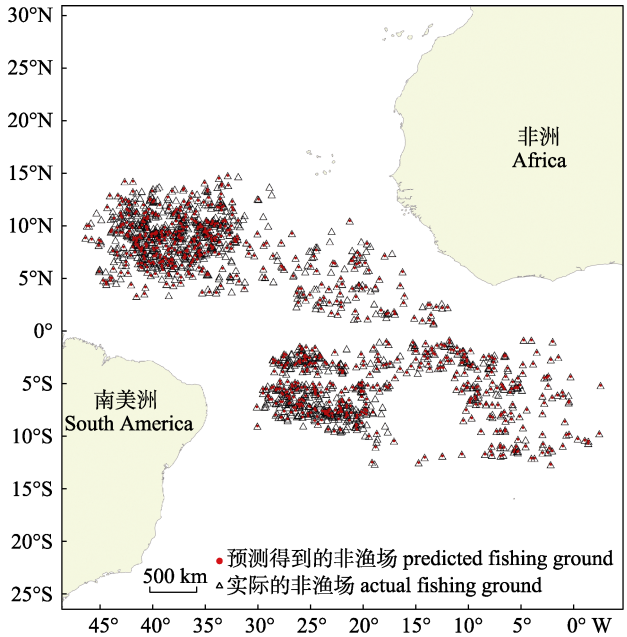

图 6 实际非黄鳍金枪鱼渔场与预报得到的非黄鳍金枪鱼渔场对比图

Fig. 6 Comparison of actual non-fishing ground and predicted non-fishing ground of *Thunnus albacares*

2.3 中心渔场

把 25%的测试站点的环境数据代入预测能力最好的 STK 模型, 计算得出“渔场”位置, Arcgis 软件画出的“渔场”位置密度分布如图 7 所示, 中心渔场主要分布在 5°N~10°N, 33°W~43°W 和 3°S~8°S, 26°W~28°W 的海域。

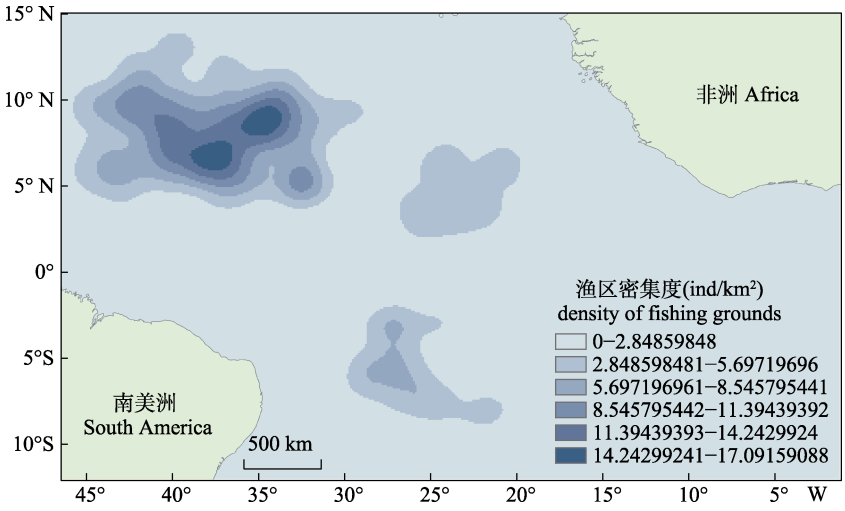

图 7　黄鳍金枪鱼中心渔场分布图

Fig. 7　Distribution map of *Thunnus albacares* fishing ground density

3　讨论

3.1　模型对比分析

STK 模型的预测准确率较其他方法均有所提高, 这是因为 Stacking 集成学习使用交叉验证方法训练 NB、CART 和 LR 模型, 产生新的训练集, 再用 RF 模型训练, 能够结合各模型的优势, 提高集成学习的泛化能力, 避免过度拟合。Xgboost 能够针对样本进行学习, 该方法能够显著提高学习效果, 但是该模型在每轮的训练中都会使用同一训练集, 训练样本集的单一性, 会降低模型的泛化能力, 也很容易受到噪声的影响产生过拟合现象, 且每个基学习器只能顺序生成, 训练效率相对较差[19], 当样本数据的质量不高时会影响模型的预测性能。RF 是通过 Bootstrap 自助采样和并行式集成算法, 能够有效降低决策树分类器的方差, 具有良好的泛化能力和抗噪能力, 提高运行效率, 使得本研究的样本数据表现较好。预测结果表明 RF 的预测效果优于 Xgboost, 这与侯娟等[20]的结果一致。KNN 在样本数据不平衡时, 预测偏差比较大, 且每一次分类时都会重新进行一次全局运算。NB 需要假设样本属性相互独立, 如果样本属性有关联时预测效果不好。LR 和 CART 容易过拟合, 导致泛化能力不强, 且使用 CART 算法时, 如果某些特征样本的比例过大, 生成的决策树容易偏向这些特征。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, 所以与 Stacking 相比对渔场的预报效果较差。

3.2　环境变量的选择

3.2.1　环境变量共线性分析的必要性　训练数据中环境变量的共线性问题普遍存在, 消除多重共线性有利于提高模型预测精度和运行效率。由于海洋环境之间的相互影响, 导致各变量之间存在相关性(图 1), 200 m、250 m 水层温度和 200 m 水层盐度的相关系数分别为 0.92, 0.73。Dormann 等[21]认为相关系数大于 0.7 说明环境变量之间可能存在共线性, 其中多重共线性对 NB、SVM 和 LR 的预测精度影响较大。目前关于 NB 在金枪鱼渔场预报上的研究常采用单因子分析[22-23], 这是因为 NB 要求各环境变量相互独立[24], 一般采用主成分分析除去变量之间的相关性和消除变量之间的线性相关以提高预测精度[2,25]。惠守博等[26]认为各变量之间存在较强的共线性是造成 SVM 预测精度降低的主要原因之一, 且多重共线性也会影响 LR 模型权重的准确性和稳定性[27]。各变量间的共线性不会影响 CART、RF 和 Xgboost 的预测精度, 是因为在模型的训练过程中会消除变量之间共线性的影响, 但是变量之间较强的相关性会使得大部分环境因子的信息相互叠加, 导致数据出现大量冗余, 即使用相似的环境因子数据会造成模型运行效率的降低和解释变量对被解释变量贡献率的误判。

3.2.2 共线性分析方法的可靠性 本研究采用相关系数检验法结合 VIF 检验法判断变量间是否存在共线性, 两种方法判断结果基本一致, 共线性分析结果可靠。当两个自变量的相关系数大于 0.7 或 VIF 大于 10 时说明可能存在共线性[21]。相关性分析法只考虑了两个变量之间简单的相关关系, 而影响渔场的海洋环境因子众多且因子之间的关系非常复杂, 所以两变量之间有较强的相关性不一定存在共线性, VIF 能够对海洋环境变量进行综合考虑[28]。

3.2.3 建模所用的环境因子 本研究最终使用 Chl a、WS、EKE、T0、T50、T100、T300、S200、S500 共 9 个环境变量建立黄鳍金枪鱼渔场预测模型, 得到的预测结果准确率较高。Chl a 是通过食物链影响金枪鱼渔场分布的, 在黄鳍金枪鱼 CPUE 和海洋环境因子关系的研究中常常将 Chl a 作为影响因子[29]。WS 则会影响海洋中的上升流[30], 能够将海底的营养物质带到表层, 为黄鳍金枪鱼提供良好的生存环境[31]。但 Bakun 等[32]认为过高的海表面风速会造成海水湍流加大和海水的浑浊程度增加, 致使浮游植物不能有效利用周围的营养盐, 造成食物短缺。EKE 是通过影响环流、海洋温度以及 Chl a 的垂直和水平分布, 从而影响黄鳍金枪鱼的资源丰度和渔场分布[33]。以往的黄鳍金枪鱼渔情预报大都采用海洋表层的环境数据, 很少利用不同水深的环境因子, 研究发现黄鳍金枪鱼具有明显的垂直活动现象[29,34], 经常在 150~250 m 水层活动[35], 且不同水层的温度会影响黄鳍金枪鱼渔场分布[36], Chen 等[37]认为温度和盐度影响黄鳍金枪鱼的产卵, 研究中 200 m 和 500 m 水层的盐度也与黄鳍金枪鱼 CPUE 相关。且从表 1 可以看出 CPUE 与使用的环境变量均呈显著相关。

3.3 展望

本研究只将海表面风速、Chl a 浓度、涡动能以及 0~500 m 水层的垂直温度和盐度环境因子用于大西洋热带海域黄鳍金枪鱼渔场研究, 在今后的研究中还需要进一步探究溶解氧、温跃层等环境因子对黄鳍金枪鱼分布的影响。此外, 在今后的研究中还应继续完善、提高 Stacking 集成学习的预测能力。

参考文献:

[1] Torres-Faurrieta L K, Dreyfus-León M J, Rivas D. Recruitment forecasting of yellowfin tuna in the Eastern Pacific Ocean with artificial neuronal networks[J]. Ecological Informatics, 2016, 36: 106-113.

[2] 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. [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.]

[3] 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. [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.]

[4] 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]. Haiyang Xuebao, 2016, 38(10): 34-43. [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.]

[5] 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. [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.]

[6] 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. [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.]

[7] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[8] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2): 197-227.

[9] Breiman L. Stacked regressions[J]. Machine Learning, 1996, 24(1): 49-64.

[10] 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 (in Chinese), 2013, 35(1): 158-164. [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.]

[11] Gao F, Chen X J, Guan W J, et al. Fishing ground forecasting of chub mackerel in the Yellow Sea and East China Sea

using boosted regression trees[J]. Haiyang Xuebao, 2015, 37(10): 39-48. [高峰, 陈新军, 官文江, 等. 基于提升回归树的东、黄海鲐鱼渔场预报[J]. 海洋学报, 2015, 37(10): 39-48.]

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

[13] 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. [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.]

[14] 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.

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

[16] Yuan H C, Zhao Y T, Liu J S. Ammonia nitrogen level forecasting based on PCA-NARX neural network[J]. Journal of Dalian Ocean University, 2018, 33(6): 808-813. [袁红春, 赵彦涛, 刘金生. 基于 PCA-NARX 神经网络的氨氮预测[J]. 大连海洋大学学报, 2018, 33(6): 808-813.]

[17] Zhang T J. Ecological niche modeling and analysis of pelagic broadcast-spawning small fish[D]. Beijing: China Agricultural University, 2016. [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.]

[18] 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. [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.]

[19] Xu J W, Yang Y. A survey of ensemble learning approaches[J]. Journal of Yunnan University (Natural Sciences Edition), 2018, 40(6): 1082-1092. [徐继伟, 杨云. 集成学习方法: 研究综述[J]. 云南大学学报(自然科学版), 2018, 40(6): 1082-1092.]

[20] 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. [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.]

[21] 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.

[22] 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]. Fishery Information & Strategy, 2012, 27(3): 214-218. [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.]

[23] Fan W, Chen X Z, Shen X Q. Tuna fishing grounds prediction model based on Bayes probability[J]. Journal of Fishery Sciences of China, 2006, 13(3): 426-431. [樊伟, 陈雪忠, 沈新强. 基于贝叶斯原理的大洋金枪鱼渔场速预报模型研究[J]. 中国水产科学, 2006, 13(3): 426-431.]

[24] Wang G C. Research and application of naive Bayesian classifier[D]. Chongqing: Chongqing Jiaotong University, 2010. [王国才. 朴素贝叶斯分类器的研究与应用[D]. 重庆: 重庆交通大学, 2010.]

[25] Xiao Y D, Ji P. Application of Bayesian method in forest fire occurrence model base on Poisson distribution[J]. Forest Science and Technology, 2018(12): 8-11. [肖云丹, 纪平. 基于泊松分布的森林火灾发生数贝叶斯估计模型[J]. 林业科技通讯, 2018(12): 8-11.]

[26] 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. [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.]

[27] Zhang L. The test of multi-collinearity and the quantitative analysis of the degree of impact of prediction targets[J]. Journal of Tonghua Normal University, 2010, 31(4): 19-20, 38. [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.]

[28] Zhu Y, Zheng Y R, Yin M. Multicollinearity test under statistical significance[J]. Statistics & Decision, 2020, 36(7): 34-36. [朱钰, 郑屹然, 尹默. 统计学意义下的多重共线性检验方法[J]. 统计与决策, 2020, 36(7): 34-36.]

[29] Song L M, Shen Z B, Zhou J K, et al. Effects of environmental variables on catch rates of yellowfin tuna (*Thunnus albacrares*) in waters near Cook Islands[J]. Journal of Shanghai Ocean University, 2016, 25(3): 454-464. [宋利明, 沈智宾, 周建坤, 等. 库克群岛海域海洋环境因子对黄鳍金枪鱼渔获率的影响[J]. 上海海洋大学学报, 2016, 25(3): 454-464.]

[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, 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] Bakun A, Black B A, Bograd S J, et al. Anticipated effects of climate change on coastal upwelling ecosystems[J]. Current Climate Change Reports, 2015, 1(2): 85-93.

[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: 012004.

[34] Xu G Q, Zhu W B, Zhang H L, et al. Relationship between fishing grounds of *Thunnus obesus* and *Thunnus albacores* with environmental factors in the Indian Ocean based on generalized additive model[J]. Haiyang Xuebao, 2018, 40(12): 68-80. [徐国强, 朱文斌, 张洪亮, 等. 基于 GAM 模型分析印度洋大眼金枪鱼和黄鳍金枪鱼渔场分布与不同环境因子关系[J]. 海洋学报, 2018, 40(12): 68-80.]

[35] Schaefer K M. Spawning time, frequency, and batch fecundity of yellowfin tuna, *Thunnus albacares*, near Clipperton Atoll in the Eastern Pacific Ocean[J]. Fishery Bulletin, 1996, 94: 98-112.

[36] Yang S L, Zhang B B, Zhang H, et al. A review: Vertical swimming and distribution of yellowfin tuna *Thunnus albacares*[J]. Fisheries Science, 2019, 38(1): 119-126. [杨胜龙, 张忭忭, 张衡, 等. 黄鳍金枪鱼垂直移动及水层分布研究进展[J]. 水产科学, 2019, 38(1): 119-126.]

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

Fishing ground forecasting models for yellowfin tuna (*Thunnus albacares*) in the tropical waters of the 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 201306, China;
3. CNFC Overseas Fisheries Co., LTD., Beijing 100032, China

Abstract: To predict the yellowfin tuna (*Thunnus albacares*) fishing ground in the tropical waters of the Atlantic Ocean accurately, and to select the optimal prediction model, a series of yellowfin tuna fishing ground prediction models were developed based on catch per unit effort data from the logbooks of 13 Chinese longliners operating in the tropical waters of the Atlantic Ocean from 2016 to 2019. The marine environmental variables (sea surface wind speed, chlorophyll a concentration, eddy kinetic energy, vertical temperature, and salinity in the 0–500 m water layer) and spatial factors (latitude and longitude) were included in the models. The time resolution of each model was one day, and the spatial resolution was 0.25° × 0.25°. The series of models, comprising Naive Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Classification and Regression Tree (CART), Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting Decision Tree (Xgboost), and stacking ensemble (STK) model (developed by the NB, CART, and LR models), were constructed using 75% of the data and verified using 25% of the data. The results showed that in the NB, KNN, RF, CART, LR, SVM, Xgboost, and STK models, (1) the forecast accuracy values for the yellowfin tuna fishing ground were 61.62%, 62.03%, 66.37%, 63.0%, 63.26%, 64.97%, 64.08%, and 68.72%, respectively; (2) the corresponding areas under the ROC curve were 0.64, 0.67, 0.72, 0.66, 0.68, 0.70, 0.69, and 0.72, respectively; and (3) the prediction accuracy of the STK model was greater than that of other methods. These results suggest that the STK model should be used to predict yellowfin tuna fishing grounds in the tropical waters of the Atlantic Ocean.

Key words: *Thunnus albacares*; fishing ground forecast; ensemble model; Atlantic Ocean

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