

袁红春, 王 敏, 刘 慧, 等. 基于特征交互与卷积网络的渔场预测模型[J]. 江苏农业学报, 2021, 37(6): 1501-1509.
doi: 10.3969/j.issn.1000-4440.2021.06.019

基于特征交互与卷积网络的渔场预测模型

袁红春^{1,2}, 王 敏¹, 刘 慧¹, 陈冠奇¹

(1. 上海海洋大学信息学院, 上海 201306; 2. 农业农村部渔业信息重点实验室, 上海 201306)

摘要: 长鳍金枪鱼是南太平洋渔业生产中主要的捕捞对象, 准确预测其渔场分布对提高渔业捕捞效率具有重要意义。针对传统渔场预测方法预测精度低的问题, 本研究提出一种基于特征交互与卷积网络的渔场预测模型——CNN-Cross。该模型引入 Embedding 层对数据进行处理, 解决了 One-Hot Encoding(独热编码)带来的特征稀疏性问题以及手动特征工程对结果的影响。同时, 引入 Cross 网络提取特征之间的交互信息, 消除了单特征对目标拟合不足的问题, 并且结合 CNN 网络对 Embedding 层生成的二维特征图进行高阶隐藏信息提取, 最后将两部分网络提取到的特征融合, 输出分类结果。使用渔业数据对模型预测效果进行验证, 结果表明, 模型预测南太平洋渔场总召回率达到 87.4%, 中心渔场召回率达到 89.4%。表明, 将特征交互网络与卷积神经网络相结合可以明显提高渔场预报精度, 且精度能够较好地满足现实渔业作业需求。

关键词: 长鳍金枪鱼; Cross 网络; 卷积神经网络; 特征交互

中图分类号: S931.41 **文献标识码:** A **文章编号:** 1000-4440(2021)06-1501-09

Fishing ground prediction model based on feature interaction and convolutional network

YUAN Hong-chun^{1,2}, WANG Min¹, LIU Hui¹, CHEN Guan-qi¹

(1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs, Shanghai 201306, China)

Abstract: *Thunnus alalunga* is the main fishing target of fishery production in the South Pacific Ocean. It is of great significance to accurately predict the fishery distribution of *T. alalunga* for improving fishery efficiency. In view of lack of accuracy of traditional fishery prediction methods, this paper proposed a fishery prediction model based on feature interaction and convolutional network—CNN-Cross. In this model, the Embedding layer was introduced to process the data, which solved the problem of feature sparsity caused by One-Hot Encoding and the influence of manual feature engineering on the result. At the same time, the Cross network was introduced to extract interactive information between different features to eliminate the problem of insufficient target fitting by single feature, and the two-dimensional feature map generated by the Embedding layer was extracted with the CNN network for high-order hidden information extraction. Finally, the features extracted by two networks were fused and the classification results were output. The effect of the model was verified by fishery data. The results showed that the predicted total recall rate of the South Pacific fishery reached 87.4%, and that of the central fishing ground

reached 89.4%. The research results show that the combination of feature interaction network and convolutional neural network can obviously improve the accuracy of fishery forecast, and the accuracy can better meet the needs of practical fishery operations.

Key words: *Thunnus alalunga*; Cross network; convolutional neural network; feature interaction

收稿日期: 2021-04-08

基金项目: 国家自然科学基金项目(41776142); 国家重点研发计划项目(2018YFD0701003)

作者简介: 袁红春(1971-), 男, 江苏海门人, 博士, 教授, 主要从事专家系统、智能计算、智能信息处理等研究。(E-mail) hcyuan@shou.edu.cn

长鳍金枪鱼作为一种大洋性中上层高度洄游生物,具有较高的经济价值、营养价值^[1]。长鳍金枪鱼广泛分布于各热带海域与温带海域^[2],尤以南太平洋海域分布较广。近年来,长鳍金枪鱼已成为南太平洋金枪鱼延绳钓的主要捕捞对象之一^[3-5]。因此,准确预测南太平洋长鳍金枪鱼渔场,对于降低渔业作业成本、提高寻找渔场效率具有极大的意义。

目前国内外学者仍然使用传统的方法与模型对渔场进行预测,并没有对渔场预测与深度学习结合进行进一步探究。如 Zainuddin 等^[6]分别利用广义加性模型(GAM)和广义线性模型(GLM)对长鳍金枪鱼渔场进行回归预测。崔雪森等^[7]利用朴素贝叶斯建立渔场预报模型对西北太平洋柔鱼渔场进行预测,对高产区预报精度达到 69.9%。张孝民等^[8]利用 MAXNET 模型预测西北太平洋秋刀鱼潜在渔场。宋利明等^[9]采用 SVM 研究了水中环境因子与长鳍金枪鱼渔获关系,建立了金枪鱼栖息环境综合指数模型。但由于海洋数据存在数据维度高、环境数据复杂等情况,目前传统方法无法很好地拟合复杂数据,对特征信息提取不够充分,且在数据预处理过程中,人为因素也会对最后的预测准确度造成影响,存在预测精度较低的问题,导致预测结果对渔业作业的指导具有局限性。

现有的特征选择方法都忽略了特征之间的交互作用。某一特征可能与目标没有直接关联,如果将该特征与其他特征结合,就可能与目标密切相关^[10]。目前特征交互在广告点击率预测领域已有较多的研究。如 Guo 等^[11]提出 DeepFM 模型可以对特征进行二阶交互,解决了特征稀疏情况下的特征组合问题,成功提高了广告点击率预测精度。Cheng 等^[12]提出 wide & deep 模型,该模型使用 wide 部分提取特征交互信息,并与 DNN 相结合,明显提高了 google play 应用的获取率。由于海洋数据与广告数据都具有高维稀疏特性,故本研究将特征交互技术应用于渔场预测领域,从而进一步提高渔场预测精度。

随着渔业捕捞技术的提高与科技的发展,渔业数据规模也变得越来越大。传统方法对于大规模数据拟合效果欠佳,而深度学习可以挖掘出复杂动态场景中重要的语义特征,可以拟合高维复杂数据。CNN 作为深度学习领域的代表,已在图像识别、语

音识别等方面取得巨大成功。CNN 网络从图像中提取局部有效特征,可以有效提高模型的泛化性。因此,本研究将特征交互技术与 CNN 相结合,提出了一种全新的渔场预测模型——CNN-Cross 网络。模型采用 Cross 网络部分提取显式低阶交互特征,采用 CNN 部分提取隐藏高阶特征,然后对两部分网络的输出结果进行融合,最后通过 sigmoid 函数输出分类结果。

1 材料与方法

1.1 数据来源

本研究选取南太平洋 110°E~135°W、5°S~40°S 为研究区域,采用 2000~2015 年共 16 年数据进行试验。

其中,渔业作业数据来自于中西太平洋渔业委员会(Western and Central Pacific Fisheries Commission, WCPFC)的南太平洋延绳钓数据。该数据包括空间坐标位置、渔业作业时间、平均钩数以及平均渔获量(质量和尾数)。本研究对该数据以月为时间单位进行统计,并将产量按 5°×5°空间范围进行归并。部分渔业作业数据如表 1 所示。

表 1 渔业作业数据

Table 1 Operation data of fishery

时间 (年-月)	纬度 (°)	经度 (°)	钩数 (×10 ³ 钩)	鱼获尾数 (尾,×10 ³ 钩)
2003-02	-5	205	2 085	100
2003-02	-10	220	1 908	19
2003-02	-10	145	1 214	126
2003-02	-10	215	9 738	2 146
2003-02	-10	210	26 346	8 513
⋮	⋮	⋮	⋮	⋮
2015-12	-30	170	1 568	960

环境因子选取海面高度(SSH)、海表温度(SST)和叶绿素 a 浓度(Chla),3 种环境因子数据均来自美国国家海洋和大气管理局(Nation Oceanic and Atmospheric Administration, NOAA),并且 3 种环境因子数据以月为时间分辨率,空间分辨率为 1°×1°网格。由于渔业作业数据与环境因子数据空间分辨率不匹配,所以需要将环境因子数据取在渔区网格中心点上,归并为 5°×5°空间分辨率,消除量纲。部分数据如表 2 所示。

表2 环境因子数据

Table 2 Data of environmental factors

时间 (年-月)	纬度 (°)	经度 (°)	海面温度 (℃)	海面高度 (m)	叶绿素 a 质量浓度 (mg/m ³)
2003-02	-15	210	28	0.678	0.104
2003-07	-10	210	27	0.647	0.143
2003-07	-20	155	24	0.833	0.138
2003-03	-10	220	29	0.479	0.143
⋮	⋮	⋮	⋮	⋮	⋮
2003-11	-5	125	30	0.626	0.115

1.2 数据预处理

1.2.1 单位捕捞努力量渔获量($CPUE$) 在渔业生产当中,通常使用单位捕捞努力量渔获量($CPUE$)作为评估渔业资源密度的指标^[13]。 $CPUE$ 通过公式(1)计算:

$$CPUE_{(m,n)} = \frac{F_{(m,n)}}{H_{(m,n)}} \quad (1)$$

其中, $CPUE_{(m,n)}$ 代表经纬度为 (m,n) 渔区范围内的月平均 $CPUE$, $F_{(m,n)}$ 和 $H_{(m,n)}$ 分别代表该月该渔区渔获尾数和某月该渔区总千钩数。

1.2.2 渔场等级划分 为了适应模型的需要并使渔场等级划分更加合理,本研究采用陈雪忠等^[14]的渔场等级划分方式,计算了各月各个渔区的 $CPUE$ 三分位数,并将大于上三分位数的渔区划分为中心渔场(分类标签用1表示),小于上三分位数的渔区划分为非中心渔场(分类标签用0表示)。试验使用了2000–2015年共16年南太平洋长鳍金枪鱼作业数据,把渔区按 $5^\circ \times 5^\circ$ 空间分辨率进行划分,最终得到10 603个样本,其中,中心渔场数为3 584个,非中心渔场数为7 019个。

1.2.3 分类标准的定义 由于本试验是二分类问题,而Sigmoid函数输出(y)在0~1,优化会相对稳定方便,所以将Sigmoid函数作为最后输出层的激活函数,结果会更加合理。在样本分布均匀的试验中,定义当 $y > 0.5$ 时,为正样本(在本研究中为中心渔场)。但在本研究中,由于样本分布不均匀,其中训练集中中心渔场数为2 669个,非中心渔场数为5 182个,故分类阈值应按照公式(2)划分:

$$\frac{y}{1-y} = \frac{T}{F} \quad (2)$$

其中, T 和 F 代表中心渔场与非中心渔场个数。

故按照上式,最终的分类阈值为0.339。

1.2.4 数据归一化处理 由于不同数据的单位、数量级以及量纲不同,故需要将月份、经纬度、海面高度、海面温度和叶绿素a浓度归一化到0~1。计算公式如公式(3)所示:

$$x^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (3)$$

其中, x^* 表示某一样本归一化后的值, x_i 表示该样本初始值, x_{\max} 、 x_{\min} 分别表示样本中某一类特征值中的最大值和最小值。

1.3 模型结构

本研究模型分为4部分,分别为Embedding层、Cross网络、CNN网络和输出部分。首先Embedding层对one-hot编码后的时空因子数据进行重新空间映射,然后CNN网络对Embedding层生成的二维特征图进行高阶隐藏信息提取,同时将Embedding层生成的数据展平后与环境因子数据结合共同作为Cross网络的输入,在Cross网络部分提取所有特征之间的显式交叉信息。最后,在输出部分将CNN网络与Cross网络提取到的特征进行融合,并通过Sigmoid激活函数输出预测结果。

1.3.1 Embedding层 在回归、分类等机器学习任务中,特征值之间的距离以及相似度计算的准确性对于模型的性能有显著的影响,大部分算法计算距离和相似度都是基于欧氏空间。然而离散特征无法映射到欧氏空间导致无法计算距离和相似度,因此,对于离散特征需要进行One-Hot处理。使用One-Hot编码,可以使非偏序关系的变量取值不具有偏向性,而且到原点的距离是等距的,能够让特征之间的距离计算更加合理,但数据在One-Hot编码后会变得非常稀疏,不利于模型对相关信息的提取。针

对 One-Hot 编码带来的稀疏问题,本研究使用 Embedding 技术来解决。

Embedding 在数学上可以看作一种空间映射。该映射是单射,且映射前后结构不变,即函数 f 被称为单射函数时,对每一个 y 至多存在一个定义域内的 x 使得 $f(x) = y$ 。对应到 Embedding 就是寻找一个单射函数,将原始数据特征生成新的空间表示。

Embedding 向量是动态生成的,无需繁琐的手动特征工程, Embedding 向量会随机初始化得到一组随机向量,接着该向量会像其他权重参数一样随着网络训练最终确定,即 Embedding 向量会随着网络的训练得到优化。本研究学习的是一个 4 维的 Embedding 向量,最终每一项都会映射到 4 维空间中的一个点,这样相似项就会在该空间内彼此邻近。Embedding 向量计算公式如式(4)所示:

$$x_{\text{embed},i} = W_{\text{embed},i} x_i \quad (4)$$

其中, $x_{\text{embed},i}$ 是生成的嵌入向量, x_i 是离散特征(时空因子)生成的 One-Hot 编码, $W_{\text{embed},i} \in \mathbb{R}^{n_e \times n_v}$ 是对应的嵌入矩阵,该矩阵将会与其他网络参数一起随着训练进行优化。 n_e 是每一个特征中不同类别的个数, n_v 是嵌入的维度。

引入 Embedding 层可以避免繁琐的手动特征工程,而且还可以解决 One-Hot 编码带来的特征稀疏性问题。

1.3.2 Cross 网络模块 特征交互又名特征组合,是一种合成新特征的方法,可以在具有多维数据的数据集上很好地拟合非线性特征。由于单一特征不能很好地表达目标信息,所以数据中的各个特征往往需要以不同的维度展现不同的信息,因此,需要对特

征进行交叉,获取特征之间的交互信息,从而更充分地拟合目标。

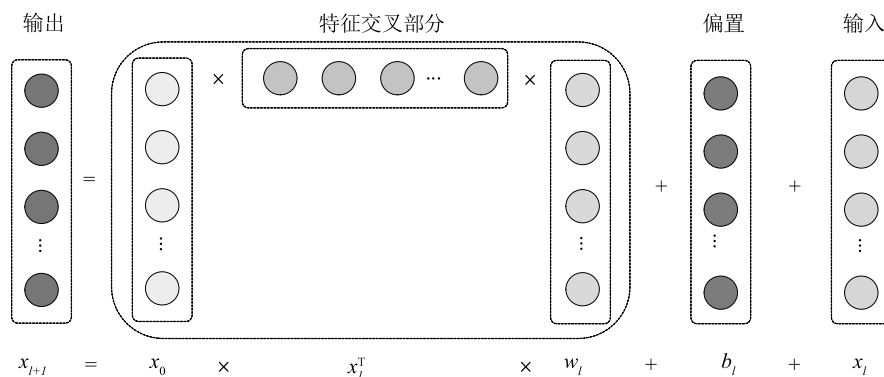
传统的特征交叉是由专业人员通过手工进行组合的,这种方法常常存在局限性,并且成本较高,难以应用。因此,许多学者开始研究使用神经网络模型自动学习交叉特征,弥补传统特征交叉方法的不足。如 Rendle^[15] 于 2010 年提出的因子分解机,该算法可以显式自动组合二阶特征交叉,但该算法无法提取更高阶的特征交互信息。2017 年 Wang 等^[16] 提出了 DCN 网络,该网络解决了因子分解机的缺陷,可以对特征进行更深层次的交互,从而可以挖掘到更深层次的信息。

Cross 网络是 DCN 网络的一个模块,该模块可以以一种显式、可控且高效的方式,自动构造有限高阶显式交叉特征。Cross 网络由交叉层组成(图 1),其中每层的计算公式见公式(5):

$$x_{l+1} = x_0 x_l^T w_l + b_l + x_l = f(x_l, w_l, b_l) + x_l \quad (5)$$

其中, x_0 表示初始输入向量, x_l, x_{l+1} 为列向量,分别代表第 l 和第 $l+1$ 个交叉层的输出, w_l 和 b_l 分别代表第 l 层的权重和偏置值。每次完成特征交叉后,交叉网络都会将原来的输入(x_l)再输入,作为下一次交叉网络的输入。这里借鉴了残差网络的思想,可以提高权重的敏感度,并且使网络更易于反向传播,提高训练效率。

经过试验验证,模型设置 5 层 Cross 网络层,可以获得最好的预测效果,因此本研究设置 5 层 Cross 网络层对经过 Embedding 层编码后的时空因子数据和环境因子数据进行特征交叉,提取 6 种因子之间的特征交互信息。



x_l, x_{l+1} : 列向量; x_0 : 初始输入向量; w_l : 第 l 层的权重; b_l : 第 l 层的偏置值。

图 1 单个交叉层示意

Fig.1 Schematic diagram of a single cross layer

1.3.3 卷积神经网络模块 卷积神经网络由 Lecun 等^[17]提出,主要用于对图像中的信息进行提取。卷积神经网络基本结构由输入层、卷积层、池化层、全连接层以及输出层组成^[18],具有权值共享、局部连接等特点。

传统的卷积神经通常会采用卷积层与池化层交替使用的结构,池化层用于选择特征和减少特征数量,以达到减少参数数量的目的,但池化操作后可能会存在信息丢失的问题。卷积层通过卷积核在特征图上进行滑动操作,提取一个局部区域的特征,不同的

卷积核相当于不同的特征提取器^[19]。假设给定 1 张图片 $X \in \mathbb{R}^{M \times N}$ (M, N 为原图像尺寸) 和 1 个滤波器 $W \in \mathbb{R}^{U \times V}$ (U, V 为滤波器尺寸),那么此时卷积按公式 (6) 计算:

$$y_{ij} = \sum_{u=1}^U \sum_{v=1}^V w_{uv} x_{i-u+v, j-v+1} \quad (6)$$

其中, w_{uv} 表示卷积核矩阵, $x_{i-u+v, j-v+1}$ 表示特征图中与卷积核对应的矩阵, y_{ij} 表示卷积之后的结果。

二维卷积层的卷积过程如图 2 所示:

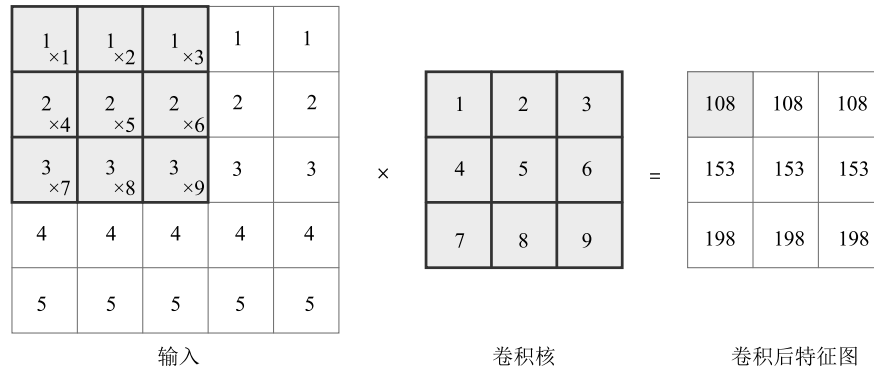


图 2 二维卷积层的卷积过程

Fig.2 Convolution process of two dimensional convolution layer

本模型卷积部分的输入为 Embedding 层生成的 $3 \times 4 \times 1$ 的矩阵,由于矩阵的尺度较小,故 CNN 部分使用全卷积网络^[20]结构,同时对特征图边缘使用 0 填充,以便能更加充分提取到特征图的边缘信息。经过多次试验,卷积层设置为 3 层,卷积核数目为 32,卷积核大小为 2×2 ,步长为 1,可以更好地适应本研究模型的需求。

1.3.4 模型总体结构 本研究模型的总体结构如图 3 所示。其中,在数据处理部分,将原始时空因子数据进行 one-hot 编码后输入到 Embedding 层,经 Embedding 层动态生成 $3 \times 4 \times 1$ 矩阵。在 Cross 网络部分,将 Embedding 层的输出与环境因子数据结合(即图 3 中的 x_0 代表 Embedding 层输出展平后与环境因子数据拼接的结果),共同作为 Cross 网络的输入,然后通过 5 层 Cross 网络进行交叉特征信息提取,并在最后输出一维特征向量。在 CNN 网络部分,输入 Embedding 层生成的 $3 \times 4 \times 1$ 矩阵,通过滑动窗口操作提取特征图的高阶隐藏信息,并在最后将二维特征图展平为一维向量。在输出部分,将 CNN 网络部分与 Cross 网络部分提取得到的两组一维向量进行融合,最后输入全连接层,通过 sigmoid 激活函数输出最后的分类结果。

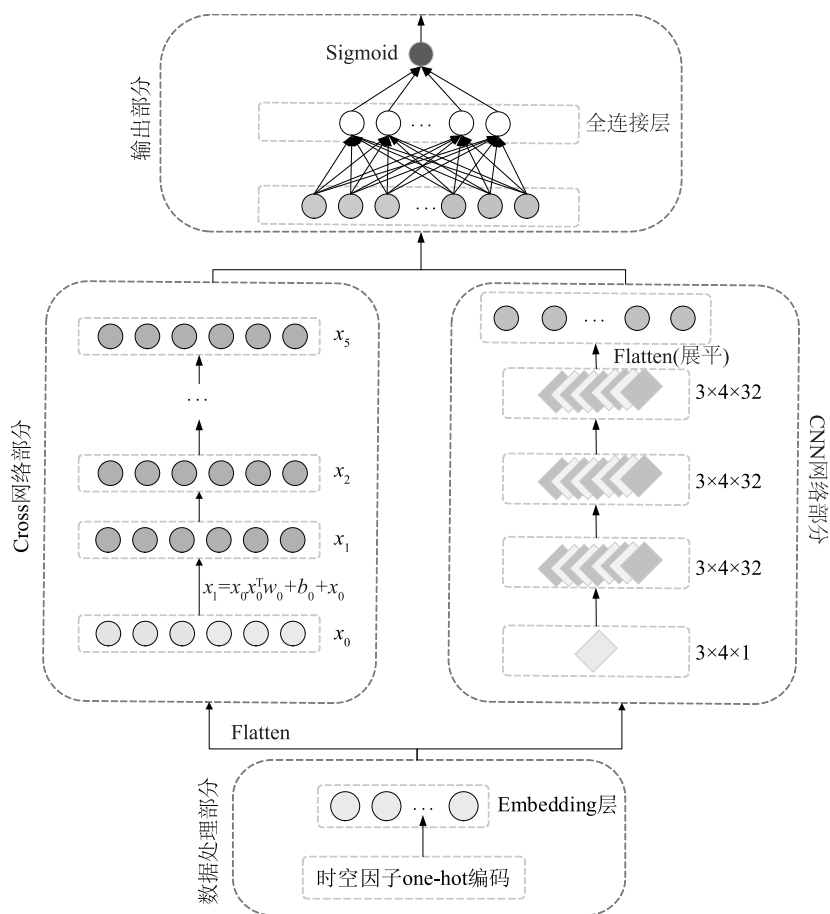
1.4 评价指标

分类任务中通常把准确率 (Accuracy) 作为衡量分类准确性的评价指标,但在样本分布不均衡的时候,该评价指标存在明显缺陷,假设正样本数占总样本数的 90%,分类器只要全部预测为正,分类器的准确率就可以达到 90%,但其实际泛化性能非常低。因此本研究采用召回率(查全率)、准确率(查准率)和 $F1$ 分数作为综合渔场等级评估标准。由于渔业实际作业中,主要选择中心渔场作为作业区域,故本研究将中心渔场看作正样本,非中心渔场看作负样本,以更好满足渔业作业需求。

1.5 试验设计

本研究所使用的工作站显卡型号为 NVIDIA GeForce RTX 2070 Super,中央处理器 (CPU) 型号为 Intel Core i7 10750H,操作系统为 Windows 10,并搭建了基于 Python 3.6 的 Tensorflow 2.1.0 框架。

试验数据集为 2000–2015 年期间南太平洋 $110^\circ \text{E} \sim 135^\circ \text{W}$ 、 $5^\circ \text{S} \sim 40^\circ \text{S}$ 时空因子、环境因子数据以及渔业作业数据。将 2000–2014 年的 9 860 条数据分别按照 15% 和 85% 划分为验证集和训练集,同时将 2015 年的 743 条数据作为测试集。



公式变量含义同公式(5)。

图3 模型总体结构

Fig.3 Overall structure of the model

为了提高整体训练速度,初始学习率(lr)=0.01,并采用学习率动态调整策略,当循环迭代1 000次后,损失值仍不下降时,按0.96的衰减率调整学习率,动量参数($momentum$)=0.9。并设置 $batch_size=32$, $epochs=150$ 。

本研究设置了3组对比试验,第一组试验研究模型中各组成部分对模型精度的影响与作用;第二组试验验证引入Embedding层对模型精度的影响;第三组试验与现有主流渔场预报模型进行对比,验证本研究模型的合理性与有效性。本研究统一采用方法1.4的评价指标对试验结果进行评估,其中中心渔场召回率和总召回率为主要参考指标。

2 结果与分析

2.1 模型各模块组成对比分析

本研究对模型组成部分合理性进行验证,通过模型自身结构对比试验可以验证本研究模型各部分

构成的合理性和有效性。其中CNN模型通过Embedding层对数据进行处理,输入到CNN网络中进行试验,Cross网络也使用Embedding层对数据进行处理,输入到Cross网络中进行试验。

分析表3可知,使用CNN网络对渔场进行预测,中心渔场召回率最低,这是因为仅通过CNN网络对特征进行提取,并不能有效提取到特征之间的交互信息,所以最后的预测结果并不理想。而使用Cross网络对渔场进行预测时,虽然中心渔场召回率高于CNN网络,但在总召回率上偏低,这是因为该网络加强了特征交互信息的提取,但忽略了对高阶隐藏信息进行提取,对数据拟合不够充分。因此本研究将两部分网络进行组合,既能有效获得特征交互信息,又能提取到高阶隐藏信息,分析表4得出,本研究模型相较于单一使用某一部分网络,在所有评价指标上均有较大幅度的提升,验证了本研究模型的合理性与有效性。

表3 模型自身结构对比

Table 3 Comparison of the model's own structure

模型	中心渔场召回率 (%)	非中心渔场召回率 (%)	总召回率 (%)	总精准率 (%)	总 F1 分数 (%)
CNN	81.0	82.9	81.9	80.1	80.8
Cross	84.1	78.4	81.3	78.6	79.2
CNN-Cross	89.4	85.4	87.4	85.0	85.8

2.2 Embedding 层对模型精度的影响

本小节共设置了 2 组试验,以验证引入 Embedding 层对模型产生的影响。一组为 BP 神经网络模型,另一组为本研究模型的 Cross 网络模块。通过

设置 2 组对比试验,可以消除试验中可能存在的偶然性,从而保证试验的真实有效。在 2 组试验中分别使用 Embedding 层和不使用 Embedding 层验证其对模型精度的影响。试验结果如表 4 所示。

表4 Embedding 层对模型精度的影响结果

Table 4 Influence of Embedding layer on model accuracy

模型	中心渔场召回率 (%)	非中心渔场召回率 (%)	总召回率 (%)	总精准率 (%)	总 F1 分数 (%)
BP	72.2	79.2	75.7	74.4	76.9
Embedding-BP	83.7	82.7	83.2	81.1	83.0
Cross	82.1	79.2	80.7	78.3	79.0
Embedding-Cross	84.1	78.4	81.3	78.6	79.2

分析表 4 可知,无论是 BP 神经网络模型,还是 Cross 网络模型,通过使用 Embedding 层对输入的时空因子特征进行处理,相较于将原始特征作为网络的输入,在评价指标上都有不同程度的提高。通过对时空因子特征进行 Embedding 处理,相较于将原始时空因子特征作为网络的输入,试验结果在各项评价指标上都有不同程度的提高。这是因为 Embedding 层对时空因子特征进行了重新的空间映射,将在欧氏空间中距离相近的特征信息聚集在一起,消除了 one-hot 编码的稀疏性问题,同时又避免了手动特征工程对结果的影响,后续特征提取将会更加充分,提取的特征信息也将更具代表性。

2.3 本研究模型与现有模型对比分析

为了验证本研究模型的优势,我们选择了 3 种主流的渔场预测模型进行对比,分别为 BP 神经网络模型、SVM 机器学习模型以及 F-CAN 深度学习模型。其中,BP 神经网络模型采用 6-512-1 网络结构,SVM 模型采用文献[21]中的模型参数,即采用径向基函数(RBF)为核函数,F-ACN 模型^[22]采用 one-hot 编码方式对数据进行处理,并使用全卷积网络结构。

如表 5 所示,BP 神经网络模型中心渔场召回率为 72.2%,总召回率为 75.7%,该模型在渔场预测中的精度最低。这是因为 BP 神经网络模型结构简单,可训练参数量较少,因此对于大规模数据的拟合能力十分有限。

SVM 模型中心渔场召回率为 83.0%,总召回率为 80.6%,相较于 BP 神经网络模型有提升,但预测精度仍然较低。SVM 模型针对小样本具有较好的拟合能力,具有较强的泛化能力,但是面对大规模复杂海洋数据,SVM 模型的效率会下降很多,而且 SVM 核函数对于高维数据的映射解释力不强,尤其是径向基函数。因此采用 SVM 模型进行渔场预测的精度并不理想。

F-ACN 模型使用了深度学习模型——全卷积神经网络,因此中心渔场召回率达到 84.0%,说明深度学习对大规模非线性复杂数据的拟合能力较强。但是,该模型使用 one-hot 编码对数据进行预处理,因此数据会变得非常稀疏,不利于特征的提取,从而使得总召回率为 78.3%,预测精度偏低。

本研究模型(CNN-Cross)在各个评价指标上相较于上述模型都有很大程度的提高。中心渔场召回

率为 89.4%, 总召回率为 87.4%, 相较于 F-ACN 模型分别提高了 5.4 个和 9.1 个百分点。在真实的渔业捕捞作业中, 渔获量和捕捞效率与渔场的等级密切相关, 通常渔业作业会选择中心渔场, 最大程度地

提高渔获量。而本研究模型中心渔场召回率达到了 89.4%, 远远高于现有模型, 这对于缩短寻找渔场时间、节省作业成本可以起到很大的作用, 可以较好地满足真实渔业作业的需求。

表 5 本研究模型与现有模型结果对比

Table 5 Comparison of the results by the model of this research with those of existing models

模型	中心渔场召回率 (%)	非中心渔场召回率 (%)	总召回率 (%)	总精准率 (%)	总 F1 分数 (%)
BP	72.2	79.2	75.7	74.4	76.9
SVM	83.0	78.2	80.6	78.0	78.6
F-ACN	84.0	75.3	78.3	81.1	78.8
CNN-Cross	89.4	85.4	87.4	85.0	85.8

3 讨论

本研究针对传统渔场预测模型预测精度低、难以拟合海洋大数据的问题, 提出了一种基于特征交互与卷积网络的渔场预测模型——CNN-Cross, 该模型引入 Embedding 层, 避免了手动特征工程以及解决了 one-hot 编码带来的特征稀疏性问题。Cross 网络通过线性时间复杂度提取特征之间的交互信息, 消除了单一特征对目标值表达不足的问题。同时通过 CNN 网络对高阶隐藏信息进行提取, 提高了模型对复杂数据的拟合能力, 提高了模型的泛化性和鲁棒性。研究结果显示, CNN-Cross 模型的南太平洋长鳍金枪鱼渔场中心召回率达到 89.4%, 总召回率达到 87.4%, 相较于现有的主流渔场预测模型预报精度有较大的提升, 且预测精度可以较好地满足现有渔业作业的需求, 可以为南太平洋长鳍金枪鱼渔业捕捞作业提供有效的指导。

CNN-Cross 模型是一种“黑盒模型”, 原理和实现较为简单, 通用性较强, 对渔业数据不需要过多手动特征工程, 仅需对数据进行简单的 one-hot 编码即可输入到模型中进行训练, 所以应用于其他渔场也不会对模型泛化性产生较大影响。但 CNN-Cross 模型仅使用了 3 种易于获取的环境因子数据对长鳍金枪鱼渔场进行预报, 而长鳍金枪鱼是一种高度洄游性生物, 其渔场分布必然会受到许多环境因子的影响, 而且渔场分布也必然会与时间和空间有紧密的联系, 所以在下一步工作中需要收集更多环境因子数据, 同时考虑渔场的时间序列因素和空间相关性, 从而进一步提高模型的精度和泛化性。

参考文献:

- [1] 范永超, 陈新军, 汪金涛. 基于多因子栖息地指数模型的南太平洋长鳍金枪鱼渔场预报[J]. 海洋湖沼通报, 2015(2): 36-44.
- [2] 郭刚刚, 张胜茂, 樊伟, 等. 基于表层及温跃层环境变量的南太平洋长鳍金枪鱼栖息地适应性指数模型比较[J]. 海洋学报, 2016, 38(10): 44-51.
- [3] 王德芬, 王玉堂, 杨子江, 等. 我国渔业多功能性的研究与思考[J]. 中国水产, 2012(1): 15-17.
- [4] 苗振清, 黄锡昌. 远洋金枪鱼渔业[M]. 上海: 上海科学技术文献出版社, 2003.
- [5] BRIAND K, MOLONY B, LEHODEY P. A study on the variability of albacore (*Thunnus alalunga*) longline catch rates in the southwest Pacific Ocean[J]. Fisheries Oceanography, 2011, 20(6): 517-529.
- [6] ZAINUDDIN M, SAITOH K, SAITOH S E I I. Albacore (*Thunnus alalunga*) fishing ground in relation to oceanographic conditions in the western North Pacific Ocean using remotely sensed satellite data[J]. Fisheries Oceanography, 2008, 17(2): 61-73.
- [7] 崔雪森, 唐峰华, 张衡, 等. 基于朴素贝叶斯的西北太平洋柔鱼渔场预报模型的建立[J]. 中国海洋大学学报(自然科学版), 2015, 45(2): 37-43.
- [8] 张孝民, 石永闯, 李凡, 等. 基于 MAXENT 模型预测西北太平洋秋刀鱼潜在渔场[J]. 上海海洋大学学报, 2020, 29(2): 280-286.
- [9] 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.
- [10] ZHOU P, LI P, ZHAO S, et al. Feature interaction for streaming feature selection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 10: 1-12.
- [11] GUO H F, TANG R M, YE Y M, et al. DeepFM: a factorization-machine based neural network for CTR prediction[C]//SIERRA C. Proceedings of the Twenty-Sixth International Joint Conference

- on Artificial Intelligence. Melbourne: International Joint Conferences on Artificial Intelligence, 2017: 1725-1731.
- [12] CHENG H T, KOC L, HARMSSEN J, et al. Wide & deep learning for recommender systems[C]//KARATZOGLOU A, HIDASI B, TIKK D, et al. Proceedings of the 1st workshop on deep learning for recommender systems. New York: Association for Computing Machinery, 2016: 7-10.
- [13] 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.
- [14] 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(中文版),2013,35(1):158-164.
- [15] RENDLE S. Factorization machines[C]//HINCHEY M, BERGMAN L A, WANG W P, et al. 2010 IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society, 2010: 995-1000.
- [16] WANG R, FU B, FU G, et al. Deep & cross network for ad click predictions[C]//ACM Special Interest Group on Knowledge Discovery in Data. Proceedings of the ADKDD'17. New York: Association for Computing Machinery, 2017: 1-7.
- [17] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [18] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229-1251.
- [19] 邱锡鹏. 神经网络与深度学习[M]. 北京:机械工业出版社, 2019:110-115.
- [20] AYSEL G, HEISTERMANN M, SOROKIN A, et al. All convolutional neural networks for radar-based precipitation nowcasting[J]. Procedia Computer Science, 2019, 150: 186-192.
- [21] 崔雪森,唐峰华,周为峰,等. 基于支持向量机的西北太平洋柔鱼渔场预报模型构建[J]. 南方水产科学,2016,12(5):1-7.
- [22] 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.

(责任编辑:陈海霞)