文章编号: 1004－7271(2009)01－0120－04

·研究简报·

基于信息增益技术的影响渔场环境因子的选定

邓 薇[1], 张 健[2], 刘必林[3]

(1.上海海洋大学信息学院, 上海 201306; 2.上海海洋大学图书馆, 上海 201306;
3.大洋生物资源开发和利用上海市高校重点实验室, 上海 201306)

摘 要: 在渔场分析中, 需要确定影响中心渔场的海洋关键环境因子, 通过建立模型对中心渔场进行预测。海洋关键环境因子的选取直接影响着预测模型的精确度。文中利用信息增益技术, 通过计算北太平洋 150°以西海域柔鱼的产量和 CPUE(单船平均日产量)对海洋环境中的垂直温度因子 (即 5 m、45 m、95 m、205 m水层的温度, 以及 5～45 m、45～95 m、95～205 m水层的垂直梯度)的信息增益值, 获得其中影响渔场分布的关键环境因子。分析结果显示: 关键环境因子是 5 m水层温度、45 m水层温度、95 m水层温度、205 m水层温度和 95～205 m温度梯度。K－S检验表明, 利用信息增益技术得到的各个关键环境因子是适合的。

关键词: 熵; 信息增益技术; 渔场预测; 关键环境因子; 北太平洋; 柔鱼; K-S检验
中图分类号: S931.4 文献标识码: A

Selection of marine environment factors based on the information gain technology

DENG Wei[1], ZHANG Jian[2], LIU Bi-lin[3]

(1. Information College, Shanghai Ocean University, Shanghai 201306, China;
2. Library of Shanghai Ocean University, Shanghai 201306, China;
3. The Key Laboratory of Shanghai Education Commission for Oceanic Fisheries Resources Exploitation, Shanghai 201306, China)

Abstract: In the fishery analysis, the key marine environment factors, which determine the central fishing ground, are needed to be identified to establish the model and then to predict the central fishing ground. The selected key marine environment factors have a direct influence on the accuracy of the forecast model. In the paper, by the information gain technology, the information gaining values which are determined by the output and CPUE(average output per vessel in a day), of vertical temperature factors in the marine environment(the water temperature at 5 m, 45 m, 95 m, 205 m and the temperature gradient in 5—45 m, 45—95 m, 95—205 m) in the western waters 150° E of North Pacific are calculated. Its main purpose is to discover the key environment factors which affect the fishing ground distribution. The result shows that the key environment

收稿日期: 2008-05-17
基金项目: 2006年度教育部新世纪优秀人才(NCET-06-0437); 国家科技支撑计划(2006BAD09A05)和上海市捕捞学重点学科(T1101)

作者简介: 邓 薇(1983—), 女, 湖北咸宁人, 硕士研究生, 专业方向为计算机应用。Tel: 13482059234, E-mail: dengwei2006_shanghai@yahoo.com.cn
通讯作者: 张 健, E-mail: jzhang@shou.edu.cn

factors are the water temperature at 5 m, 45 m, 95 m, 205 m and the temperature gradient in 95—205 m. The K-S test indicates that each key environment factor obtained by the information gain technology is appropriate.

Key words: entropy; information gain technology; fishery forecast; key environment factors; North Pacific; Ommastrephes bartrami; K-S test

在渔情预报中, 通常通过建立预测模型对中心渔场进行预测, 而预测模型的精确度受到海洋关键环境因子的影响。海洋环境因子主要包括水温, 盐度等, 其中温度因子较为重要[1]。环境因子的选取可以使用属性相关分析, 属性相关分析在机器学习、统计、模糊和粗糙集等方面都有许多研究[2]。属性相关分析的基本思想是计算某种度量, 用于量化属性与给定类或概念的相关性。这种度量包括信息增益、Gin索引、不确定性和相关系数[3]。在如判定树算法等相关分析的诸多算法中, 都采用了信息增益来度量属性与给定类的相关性[4]。并且, 使用信息增益度量方法, 信息量越大, 效果越好[3]。因此, 文中利用信息增益技术, 通过分析水温垂直结构 (5 m、45 m、95 m、205 m水层的水温, 以及 5~45 m、45~95 m、95~205 m水层的垂直梯度), 从影响北太平洋 150°以西海域柔鱼 (Ommastrephes bartrami) 渔场形成的海洋垂直温度因子中, 挖取出起决定作用的关键环境因子, 便于以后利用水温垂直结构对柔鱼中心渔场进行预测分析。

1　材料和方法

1.1　数据来源

水温数据来自哥伦比亚大学网站 http://iridl.ldeo.columbia.edu, 时间为 1998—2004年 6—11月, 即 5 m、45 m、95 m、205 m水层的水温, 以及 5~45 m、45~95 m、95~205 m水层的垂直梯度 (℃/m)。生产数据来自上海水产大学鱿钓技术组, 范围为 140°E~150°E、39°N~45°N。所有数据均处理成空间分辨率为经纬度 1.5°×1°。

1.2　数据的概化

在文中, 因为采用的是信息增益分析技术来确定关键环境因子的挖掘算法。所以为了满足算法的需要, 对各属性值进行数据概化。数据概化 (data generalization) 是一个过程, 它将大的任务相关的数据集从较低的概念层 (如: 城市) 抽象到较高的概念层 (如: 国家)。大的数据集有效、灵活的概化方法分为两类: (1) 数据立方体 (或 OLPA) 方法; (2) 面向属性的归纳方法[2]。

在文中, 采用的是面向属性的归纳方法。根据属性概化的思想, 第一根据属性删除的规则, 删除时间属性年、月以及经纬度属性。

第二, 对垂直温度梯度变化值的衡量主要是考察其是否是温跃层, 因此概化为两类信息: 温跃层, 非温跃层 (根据规定, 在深海中, 温度梯度超过 0.05℃/m即为温跃层, 否则为非温跃层[1])。所以5~45 m、45~95 m、95~205 m水层的垂直梯度数据的分类区间如下: (−∞, 0.05℃/m); (0.05℃/m, +∞)。

第三, 对于 5m、45m、95m、205m水层的温度, 以及产量和单船平均日产量 (CPUE) 这些属性, 按照等深 (一种分类方法, 即按照属性值的范围, 将属性对应的所有值进行分类, 使得在每一个分类区间中的数值个数大致相等[2]) 对每个属性进行离散化, 每个属性分为四类, 得到每个属性的不同的分类区间 (表 1)。

2　信息增益技术的实现

一个海域的 Output (总产量) 和 CPUE都可以反映出该海域渔场的分布情况, 所以分别计算出每个属性即环境因子 (5 m、45 m、95 m、205 m水层的水温, 5~45 m、45~95 m、95~205 m水层的垂直梯度) 对产量和 CPUE这两个属性的信息增益来反映每个环境因子对渔场的影响程度。

表 1 各属性对应的分类区间
Tab. 1 The corresponding classification sectors of each attribute

属性	分类区间(每个属性对应的不同分类区间值)			
5 m水层温度(℃)	6~13	13~16	16~18	18~24
45 m水层温度(℃)	3~10	10~13	13~15	15~24
95 m水层温度(℃)	1~6	6~8	8~11	11~20
205 m水层温度(℃)	1~4.5	4.5~6	6~8	8~16
总产量(t)	0~15	15~105	105~423	423~14158
CPUE(t/d)	0~1	1~2	2~3	3~9

(1)求出 Output(或 CPUE)的信息期望 I:

$$I(S_1, S_2, \cdots, S_m) = -\sum_{i=1}^{m} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

式中: m表示 Output(或 CPUE)不同的属性值个数, S_i表示 Output(或 CPUE)值为第 i个属性值的记录条数, S表示样本总数。

(2)求出每个属性对应于 Output(或 CPUE)分类的熵 $E_1(A)$:

$$E_1(A) = \sum_{j=1}^{v} \frac{S_{1j} + \cdots + S_{mj}}{S} I(S_{1j}, \cdots, S_{mj})$$

式中: v表示属性 A不同属性值的个数, S_{ij}表示属性 A值为 A_j且 Output(或 CPUE)值为第 i个属性值的记录条数, S表示样本总数, $I(S_{1j}, \cdots, S_{mj})$表示属性 A取值 A_j为对 Output(或 CPUE)分类的信息期望。

(3)计算每个属性对应于 Output(或 CPUE)分类信息增益 $Gain_1(A)$(表 2)。

$$Gain_1(A) = I(S_1, \cdots, S_m) - E_1(A)$$

表 2 各属性分别对应于产量和 CPUE的信息增益值
Tab. 2 The value of information gaining for the output and CPUE in each attribute

属性	对应产量信息增益值	对应 CPUE信息增益值
5 m水层温度	0.086 884	0.108 225
45 m水层温度	0.086 274	0.116 003
95 m水层温度	0.055 239	0.052 182
205 m水层温度	0.084 293	0.086 929
5~45 m温度梯度	0.021 196	0.018 633
45~95 m温度梯度	0.027 381	0.004 467
95~205 m温度梯度	0.083 117	0.031 061

3 海洋关键环境因子的判定与检验

3.1 海洋关键环境因子的判定

从表 2可看出,各属性对应于总产量的影响力从强到弱的顺序依次为 5 m水层温度, 45 m水层温度, 205 m水层温度, 95~205 m温度梯度, 95 m水层温度, 45~95 m温度梯度, 5~45 m温度梯度。综合领域专家的意见,阀值定为 0.03,则影响总产量的关键因子为 5 m水层温度、45 m水层温度、205 m水层温度、95~205 m温度梯度和 95 m水层温度。同样,影响 CPUE的关键因子为 45 m水层温度、5 m水层温度、205 m水层温度、95 m水层温度和 95~205 m温度梯度。综合上述分析结果,关键环境因子是 5 m水层温度、45 m水层温度、95 m水层温度、205 m水层温度和 95~205 m温度梯度。

3.2 检验

文中利用非参数统计 K-S(Kolmogorov-Smirnov)检验方法[5],对各关键环境因子进行显著性检验。K-S检验方法如下:

$$f(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i) \qquad g(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\bar{y}} I(x_i) \qquad D = \max | g(t) - f(t) |$$

式中: n为资料个数; t为分组关键环境因子值; x_i为第 i月关键环境因子观察值; y_i为第 i月的 CPUE; $\bar{y}$为所有月的平均 CPUE;若 $x_i \leqslant t$时, $I(x_i)$值为 1,否则 $I(x_i)$值为 0。

计算所有月份 K-S检验统计量 D,并以 $\alpha = 0.10$做显著性检验。结果表明,各关键环境因子 $D < P(\alpha/2) = 0.114$,假设检验条件 $f(t) = g(t)$对各个关键环境因子均成立,没有显著性差异,即认为作业渔场的每个关键环境因子是合适的。

表 3　K-S检验结果
Tab. 3　The result of K-S test

环境因子	5 m水层温度	45 m水层温度	95 m水层温度	205 m水层温度	95 ~205 m温度梯度
D值	0. 061	0. 039	0. 047	0. 042	0. 086

4　结论与分析

本文的研究结果与柔鱼的栖息环境和生活习性是一致的。昼夜垂直移动是柔鱼的重要生活习性之一。白天柔鱼通常生活在深水层, 150°E以西海域柔鱼栖息水层为 100 ~200 m。在水深 200 m以下, 其水温几乎无变化[6]。夜间柔鱼生活在 50 m以上水层, 主要游泳层处在温跃层附近或处在温跃层与海面之间[7-8]。而在西北太平洋海域主要有黑潮和亲潮两大流系, 正是由于它们的交汇与混合作用产生了许多著名的渔场, 如秋刀鱼渔场和柔鱼渔场。根据研究, 柔鱼主要分布在黑潮的前锋海域, 冷暖水交汇处, 这些海域往往与 50 m以上水层相关[1]。所以, 5 m水层温度、45 m水层温度为主要环境因子, 而 95 ~205 m水层的温度梯度和新发现的 95 m、205 m水层温度也是关键环境因子。

利用信息增益技术来确定影响北太平洋 150°E以西海域柔鱼渔场形成的有关环境因子, 找出的 95 ~205 m水层的温度梯度与 5 m、45 m水温这三个关键环境因子与以往的研究成果相符合[6], 并且发现了另外两个强环境因子, 即 95 m以及 205 m水温的关键环境因子。这样, 挖掘出更多的强关联垂直温度因子, 在利用垂直温度结构对柔鱼中心渔场进行预测分析时, 有助于提高预测模型的精度。以往对垂直温度结构方面的分析, 采用的方法大部分都是针对小样本数据, 挖掘出的信息不是很准确, 获取的海洋环境因子也不是很全面。而信息增益技术可以处理大量的数据, 且处理的数据量越大, 效果越好[3]。文中利用信息增益技术, 通过对影响渔场分布的海洋环境数据的分析, 来确定关键海洋关键环境因子的方法是非常有效的。

利用信息增益技术来确定渔场环境因子时, 在属性概化方面还有待提高。因为属性概化没有一个严格的标准, 如果属性概化的太高或者概化不足, 产生的结果都没有多少信息。如果选择较复杂的概化方法, 则时间复杂度太大, 而选择简单的概化方法如等深、等宽, 很难结合实际的情况, 此时必须结合专家的意见进行概化才能得到较理想的结果。文中仅针对垂直温度数据对中心渔场分布进行了研究, 今后还需要结合更多的环境因素, 如表温、盐度、叶绿素等, 通过组合环境因素对中心渔场进行预测分析, 进一步提高预测的精度。

参考文献:

[1]　陈新军. 渔业资源与渔场学[M]. 北京: 海洋出版社, 2004. 116－129.
[2]　Han J W, Kamber M. Data Mining Concepts and Techniques[M]. 北京: 机械工业出版社, 2006. 8. 120－131.
[3]　陈文伟, 黄金才, 赵新星. 数据挖掘技术[M]. 北京: 北京工业大学出版社, 2002. 1－48.
[4]　Hand D, Mannila H, Smyth P. Principles of Data Mining[M]. 北京: 机械工业出版社, 2003. 4. 233－255.
[5]　魏季瑄. 数理统计基础及其应用[M]. 成都: 四川大学出版社, 1991. 184－185.
[6]　王尧耕, 陈新军. 世界大洋性经济柔鱼类资源及其渔业[M]. 北京: 海洋出版社, 2005. 135－145.
[7]　陈新军, 许柳雄. 北太平洋 150°E~165°E海域柔鱼渔场与表温及水温垂直结构的关系[J]. 海洋湖沼通报, 2004. 2. 42－43.
[8]　陈新军. 北太平洋 150°E以西海域柔鱼渔场与时空、表温及水温垂直结构的关系[J]. 上海水产大学学报, 2004. 13(1): 78－83.