基于神经网络的规则提取及其渔业应用研究

袁红春[1,2], 胡倩倩[1], 沈晓倩[1], 陈新军[2]

(1. 上海海洋大学 信息学院, 上海 201306; 2. 上海海洋大学 国际海洋研究中心, 上海 201306)

摘要: 为了解决远洋渔业中过度依赖经验而产生的盲目捕捞问题, 结合海洋环境数据和历史产量数据对渔场进行有效分析, 提出了一种基于径向基函数神经网络(Radial basis function neural network, RBFNN)的栖息地指数(HSI)预测方法, 并将其应用于印度洋海域大眼金枪鱼(*Thunnus obesus*)栖息地指数的预测。在 RBFNN 训练过程中使用模糊 C 均值(Fuzzy c-means, FCM)聚类算法, 在基于神经网络的规则提取过程中首次采用了和声搜索(Harmony search, HS)算法。实验研究表明, 利用 FCM 改进后的 RBFNN, 均方误差(Mean square error, MSE)达到 0.021 6。和声搜索由于算法简单, 易于实现, 能够应用于训练后的 FCM-RBFNN 提取分类规则, 提取出的规则能够反映该渔业现状。

关键词: 印度洋大眼金枪鱼(*Thunnus obesus*); 径向基函数神经网络(Radial basis function neural network, RBFNN); 和声搜索(Harmony search, HS); 规则提取; 渔情预测

中图分类号: S934 **文献标识码**: A **文章编号**: 1000-3096(2014)09-0079-06

doi: 10.11759/hykx20121209001

渔情分析预报是海洋渔场学的重要分支, 能有效对中心渔场位置及其发展趋势进行预测。现有的渔情分析方法可归纳为三类: 经典统计学方法[1]、人工神经网络方法[2]、栖息地指数模型方法[3]。各种方法有其利弊之处。经典统计学方法大多采用多元回归方法, 然而海洋环境因子之间相互影响的动态变化特性无法满足因变量之间独立分布的研究前提。人工神经网络(Artificial neural network, ANN)方法适应性广, 但其将知识等势分布存储在连接权上, 不易被人们所理解, 从而限制了其在渔业中应用。栖息地适宜指数(Habitat suitability index, HSI)能综合多个环境分析因子, 分析它们对鱼类栖息地的影响。但建模方法大多采用回归方法, 存在传统统计方法的弊端。上海海洋大学的杨烨博[4]利用支持向量机(Support vector machine, SVM)和模糊分类规则提取方法获得渔场知识, 但训练样本只考虑年度产量数据。目前大眼金枪鱼(*Thunnus obesus*)延绳钓产量数据大多可精确至月度, SVM 明显不适用于处理大样本数据。

针对现有 ANN 存在的问题以及训练样本容量要求, 本文利用基于模糊 C 均值(Fuzzy c-means, FCM)聚类算法的径向基函数神经网络(Radial basis function neural metwork, RBFNN)训练数据, 结合栖息地适宜指数确定适应范围, 然后, 通过和声搜索(Harmony search, HS′)的方法提取 RBFNN 中隐含的规则。仿真结果表明, 提取的分类规则较符合印度洋大眼金枪鱼的渔业现状。

1 FCM-RBFNN 分类规则提取方法

1.1 HS 算法

HS 算法, 是基于现有智能算法对于生物特性的模仿, 借鉴了音乐演奏的方法。乐队中的乐器 R_i (i=1, 2,⋯, n)类比于优化问题中的第 R_i 个变量, 将各乐器的音调比作各变量的值, 各乐器音调的和声 X^k(k=l, 2,⋯, n)相当于优化问题的第 k 组解向量, 音乐效果评价类比于目标函数 $F(X^k)$(k=1, 2,⋯, n)。计算步骤如下[5]:

步骤 1: 约束条件和参数确定。确定优化问题的目标函数及 HS 基本参数。包括: 和声记忆库的大小(HMS)、和声记忆库考虑概率(HMCR)、和声音调微调概率(PAR)以及算法迭代次数(NI)。

步骤 2: 和声记忆库初始化解。随机产生 M 个优化问题的初始解放入和声记忆库 HM 内。HMS 的大小即为 M。

步骤3: 利用3种机理产生新解。(1)根据 HMCR

收稿日期: 2012-12-09; 修回日期: 2013-03-02

基金项目: 上海市教委科研创新项目(12ZZ162); 上海市科学技术委员会重点支撑项目(12510502000)

作者简介: 袁红春(1971-), 男, 江苏海门人, 教授, 博士, 主要研究方向为智能信息处理, 电话: 021-61900604, E-mail: hcyuan@shou.edu.cn

保留和声记忆库中的某些解分量; (2)随机选择产生; (3)对前2种方法的分量根据 PAR 进行微调扰动。

步骤 4: 更新记忆库。判断新解的质量, 查看其是否优于HM内的最差解, 若是, 则将新的解替换最差解; 否则, 新解被舍弃。由此, 得到新的 HM。

步骤 5: 重复步骤 3 和步骤 4, 直到达到最大迭代次数或满足停止准则后结束, 循环输出最优解。

HS 从确立初期到现在不过 10 余年, 已经在函数优化、公交线路优化、引水库调度问题、土木工程等问题上有许多成功应用的案例, 但在提取神经网络隐含规则方面的应用尚属空白。

1.2 利用 HS 算法提取 FCM-RBFNN 分类规则

ANN 的连接权蕴涵了 ANN 的知识, 也关联到它的架构和激活功能, 可以使用算法从连接权和隐含层神经元上提取规则[6-7]。该规则提取算法侧重于解决从训练后的 ANN 中提取规则, 不以任何近似系统模拟, 直接使用连接权提取属于某类的规则。连续数据需要经过离散化处理, 以显示输入项的属性划分。

RBFNN 是一种三层前馈神经网络, 其隐含层神经元激活函数是径向对称的核函数, 通常为高斯核函数。通过两阶段学习方法获得 RBFNN 的参数, 第一阶段确定隐层神经元激活函数的中心向量及其宽度参数, 第二阶段确定隐含层到输出层的权值。目前, 已经有很多的聚类方法用来选择隐函数的中心向量, 本文采用 FCM 算法。

RBFNN 的学习过程可表达为: 在 n 维空间中, 给定 p 个输入样本 X_i (i=1, 2, $\cdots$, p), 假设隐含层神经元的个数为 m, 则隐含层的第 k 个神经元输出 r_k 可以表示为:

$$r_k = \exp\left(-\frac{\|X_i - C_k\|}{2\sigma^2}\right),\ k=1, 2, \cdots, m \tag{1}$$

其中, C_k 为隐含层第 k 个神经元的中心向量, $\|X_i - C_k\|$ 为样本 X_i 与 C_k 的欧式范数, σ 为隐含层神经元高斯函数的宽度参数, 是影响 RBFNN 分类能力的重要因素, 可通过公式(2)计算获得。

$$\sigma = \frac{C_{\max}}{\sqrt{2m}} \tag{2}$$

其中, $C_{\max}$ 为中心向量之间的最大距离, m 为隐含层神经元的个数。

本研究利用 FCM 获取 C_k, 其计算步骤如下:

步骤1: 采用[0, 1]区间上的随机数初始化隶属矩阵 $\boldsymbol{U}$, 使其满足 $\sum_{k=1}^{m} u_{kj} = 1, \forall j = 1, \cdots, n$。

步骤 2: 用公式(3)计算 m 个聚类中心 C_k, k=1, 2, $\cdots$, m。

$$C_k = \frac{\sum_{j=1}^{n} \boldsymbol{u}_{kj}^{t} X_j}{\sum_{j=1}^{n} u_{kj}^{t}},\ k=1, 2, \cdots, m \tag{3}$$

这里 $t \in [1, \infty)$ 是一个加权指数。

步骤3: 根据公式(4)计算价值函数(或目标函数)。如果它小于事先给定的阈值, 或它相对于上次的改变量小于给定阈值, 则算法停止。

$$J(\boldsymbol{U}, c_1, \cdots, c_m) = \sum_{k=1}^{m} J_k = \sum_{k=1}^{m}\sum_{j}^{n} u_{kj}^{t} d_{kj}^{2},\ k=1, 2, \cdots, m \tag{4}$$

这里 $d_{kj} = \| C_k - X_j \|$ 为第 k 个聚类中心与第 j 个数据样本间的欧几里德距离。

步骤 4: 用公式(5)计算新的 $\boldsymbol{U}$ 矩阵, 返回步骤 2。

$$u_{kj} = \frac{1}{\sum_{i=1}^{m} \left(\frac{d_{kj}}{d_{ij}}\right)^{\frac{2}{t-1}}} \tag{5}$$

RBFNN 的输出层第 j 个神经元的输出 y_j 可通过公式(6)计算获得。

$$y_j = \sum_{k=1}^{m} w_{kj} r_k \tag{6}$$

其中, r_k 可参考公式(1), w_{kj} 是隐含层第 k 个神经元与输出层第 j 个神经元间的连接权值, 可以用最小二乘法直接计算得到。

为提取输入属性和输出属性之间的规则, 可寻找使 y_j 取最大值的二进制输入向量。它是一个非线性整数优化问题[8-9], 可使用 HS 算法寻优, 本文将它命名为 HS-miner 算法, 其过程可描述为:

(1) 初始化算法参数

算法参数包括: 属性变量范围值、和声记忆库的大小(HMS)、和声记忆库考虑概率(HMCR)、和声音调微调概率(PAR)以及算法迭代次数(NI)。

根据现有的栖息地适应性指数, 确定属性的取值范围和属性程度划分区间。将每个属性的程度划分区间个数相加确定输入变量的个数。然后, 对属性做格式化处理, 即将数据集转化为二进制数据。

HMS 的个数直接影响了规则提取的效果。随着 HMS 数量的增加, 其最有可能找到最优解。本次实验中, HMS 为 30。HMCR 的值极大影响了新规则的替换效率, 根据文献[10], 本文取值为 0.95。PAR 为扰动概率, 能够控制局部搜索。

(2) FCM-RBFNN 模型训练

利用处理后的样本数据, 训练 RBFNN; 利用 FCM 算法获取其中心向量, 通过公式(2)获取其宽度参数, 采用最小二乘法获取其权值; 利用公式(7)验证最佳网络参数。

$$E=\frac{1}{n}\sum_{i=1}^{n}\left(T_i-Y_i\right)^2, i=1,2,\cdots,n \tag{7}$$

其中, E 为均方误差(MSE), T_i 为第 i 个样本的观测值, Y_i 为第 i 个样本的的估计值。E 越小, 说明 FCM-RBFNN 反映现实数据的能力越强。

(3) 利用 HS 算法提取规则

确定 FCM-RBFNN 的最佳参数后, 根据公式(6)构建目标函数。利用二进制值初始化 HM, 形成 HM 矩阵。

$$\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_N^1 & Y(X^1) \\ x_2^1 & x_2^2 & \cdots & x_N^2 & Y(X^2) \\ \vdots & \vdots & & \vdots & \vdots \\ x_1^k & x_2^k & \cdots & x_N^k & Y(X^N) \end{bmatrix} \tag{8}$$

新规则生成时, 如果随机概率大于 HMCR, 则在 HM 范围外重新生成规则, 否则对于每个解向量的分量根据 PAR 进行微调。当随机概率小于 PAR 时, 每个解向量的分量的计算如公式(9), 否则, 分量保持不变。

$$X_N^{k'}=\begin{cases} X_N^{k'}+1, & X_N^{k'}=0 \\ X_N^{k'}-1, & X_N^{k'}=1 \end{cases} \tag{9}$$

如果新生成的解向量优于原来和声记忆库中的最差解向量, 则最差解向量被替换。然后, 算法运算次数叠加, 重复计算直至最大迭代次数。

(4) 规则准确度测试

根据公式(10)的最大值, 确定满足目标函数的最佳解决方案。E 的取值根据公式(7)。

$$F=\frac{1}{E} \tag{10}$$

$$A=\frac{t_1+t_2}{t_1+t_2+f_1+f_2} \tag{11}$$

利用公式(11)计算规则的准确度, 并用 A 表示, 以验证规则在测试样本中的符合性。t_1 为规则预测存在且事实也存在的事例总数, t_2 为规则预测不存在且事实不存在的事例总数, f_1 为规则预测存在但事实不存在的事例总数, f_2 为规则预测不存在但事实存在的事例总数。

2 实验结果分析

大眼金枪鱼是一种大量分布于印度洋等热带和亚热带水域的大洋性洄游鱼类。渔情预测方法研究主要探究鱼类栖息地指数(HSI)与海洋环境要素之间的关系, 本例以2006~2007年中国远洋渔业分会上海海洋大学金枪鱼技术组提供产量数据为例, 其中训练数据占80%, 预测数据占20%。数据集总共有235条记录。其中, 海面高度、海水温度(海表温度、162.5 m 海水温度、237.5 m 海水温度)、叶绿素含量是连续属性。其取值区间分别为 [–4.94, 8.01], [2.04, 19.94], [0.06, 2.11]。根据现有判断渔获量指标, 本文选取 HSI 为衡量标准, 即以单位努力渔获量(Catch per unit of effort, CPUE)作为衡量资源丰度的指数, 并结合实际的作业次数共同决定渔获量的高低。设定当 HSI≥0.5 时即有鱼群存在, 反之则无鱼群。根据现有 HSI 模型中各属性的最适宜范围, 确定各属性区间值。表1~表3分别描述了各属性离散化的范围。将 Arff 格式数据转换为 xls 格式, 再利用 Matlab 程序根据各属性的最适宜范围对属性进行离散化。

表 1 海水温度离散化范围及二进制表示

Tab.1 Discretization range of the seawater temperature and binary representation

温度程度	温度(℃)	属性分类
很低	＜4	10000
低	4~7	01000
中等	7~10	00100
高	10~15	00010
很高	＞15	00001

表 2 海面高度离散化范围及二进制表示

Tab.2 Discretization range of the sea surface height and binary representation

海面高度程度	海面高度(cm)	属性分类
低	＜0	100
中等	0~4	010
高	＞4	001

表 3　叶绿素浓度离散化范围及二进制表示

Tab.3　Discretization range of the chlorophyll concentration and binary representation

叶绿素质量浓度程度	叶绿素质量浓度(mg/m^3)	属性分类
低	＜0.2	100
中等	0.2~0.4	010
高	＞0.4	001

本实验的 RBFNN 部分和规则提取部分由 C#仿真实现，为验证本算法的有效性，实验先对 FCM-RBFNN 进行5-折交叉验证。结果显示，FCM-RBFNN 的 MSE 达到0.021 6，使用 FCM-RBFNN 直接预测 HSI 的平均准确度为96.75%，利用提取的规则进行预测，其平均准确度为88.5%。仿真系统的界面如图1和图2所示，其中图1表示 FCM-RBFNN 的学习过程，用户可根据需要确定隐含层神经元个数，并保存训练后的模型。图2表示根据模型训练后的 RBFNN，可利用 HS 算法提取规则，相关参数也可根据需要修改。

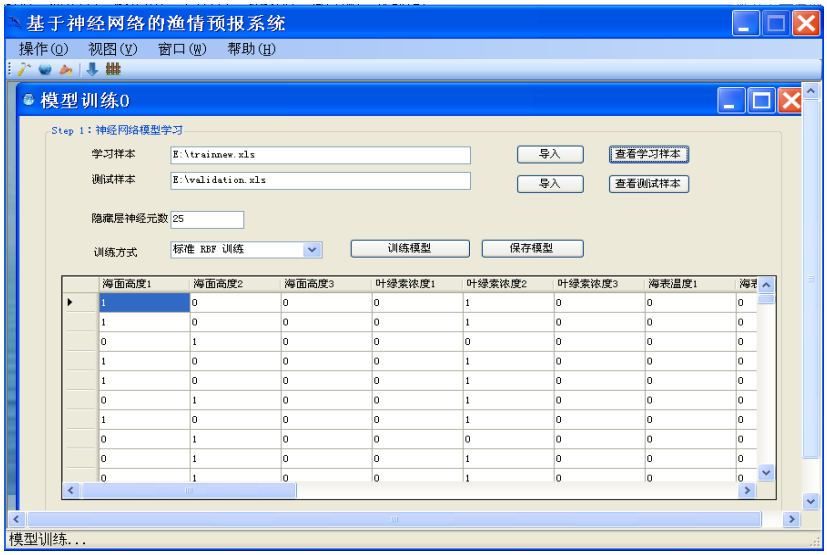

图 1　RBFNN 的训练界面

Fig.1　Interface for training the RBFNN

图 2　利用 HS 算法从 RBFNN 中提取规则的界面

Fig.2　Interface for extracting rules from RBFNN by using harmony search (HS) algorithm

遗传算法(Genetic algorithm, GA)作为最早被提出的模拟自然进化搜索算法, 已经在组合优化、人工生命、信号控制领域中有成功应用。文献[11]提出了利用GA 提取 ANN 分类器规则的方法, 其 GA 的适应度函数是取用一种包含多层前馈神经网络输入到输出模式类别传递关系的目标函数。本文将 GA 算法用于FCM-RBFNN 中, 称为 GA-miner, 以验证 HS-miner 的算法性能, 实验结果见表 4。由于新的解向量产生时, HS 算法能够处理向量中的每一个分量, 而基于遗传结构考虑, GA 需要保证其一致性, 这导致 GA-miner 不能充分考虑每个分量, 生成的规则也相对较少。

表 4 规则准确度对比
Tab.4 Comparison of rules accuracy

方法	规则数(条)	规则平均准确度(%)
HS-miner	11	88.5
GA-miner	9	75

为进一步比较本文提出算法的有效性, 利用 Weka3.5 中 NBTree、C4.5 和 PART 等传统方法挖掘样本中的规则。表 5 列出的四种不同方法, 结果比较表明, HS-miner 的性能较优于传统规则提取方法, 其提取规则最多时能达到 15 条。

根据 HS 提取的神经网络规则, 和声记忆库为 30, NI 为 10 000 次, HMCR 为 0.95, PAR 为 0.3。表 6 为利用 HS 算法从 FCM-RBFNN 中提取的部分规则。

表 5 四种规则提取方法对比
Tab.5 Comparison of four rules extraction methods

方法	测试准确度(%)	规则数(条)
NBTree	84.2	6
C4.5	78.2	7
PART	85.83	9
HS-miner	97.5	11

表 6 利用 HS 算法从 FCM-RBFNN 中提取的部分规则
Tab.6 Partial rules extracted from FCM-RBFNN by using HS algorithm

编号	规则前件	规则后件	规则准确度(%)
1	海面高度.低 AND 海表温度.高 AND 海水温度.[162.5 m].高	栖息地指数.高	97
2	海面高度.低 AND 海水温度.[162.5 m].高	栖息地指数.高	96
3	海面高度.低 AND 海表温度.中等 AND 海水温度.[237.5 m].中等	栖息地指数.高	91
4	海面高度.低 OR 中等 AND 海表温度.高	栖息地指数.高	87
5	海面高度.低 AND 叶绿素浓.低 AND 海水温度.[162.5 m].低	栖息地指数.高	85

实验结果显示各海洋因子对于印度洋大眼金枪鱼 HSI 的影响。数据分析可知, 当 HSI 较高时, 海面高度和叶绿素浓度的取值范围大多较低, 这与文献[12]指出的“SSH(-2CM-1.5CM), 叶绿素质量浓度在 0.1~0.2 mg／m^3 分布区间范围内, 大眼金枪鱼渔场 CPUE 较高”的论述较一致。当 HSI 较高时, 海表温度的范围则在中等偏上水平, 由于白天大眼金枪鱼主要在 130~300 m 之间活动, 温度在 14~17℃之间, 海水的温度随深度加深呈下降趋势。本文数据显示的海表温度较高, 而 162.5 m 海水温度和海表温度变化不大, 主要是由于 162.5 m 海水温度靠近海洋混合层, 水温较均匀。

3 结束语

本文提出利用 HS 算法提取 FCM-RBFNN 的分类规则。通过对印度洋大眼金枪鱼产量和环境数据的分析, 并提取相关规则的应用来看, HS 有较强的规则提取能力。从 FCM-RBFNN 的输入规模来看, 借助 HSI 模型划分的最适宜范围较为合理, 能够反映出属性的分类, 这不仅关系到了 FCM-RBFNN 的学习速度, 同时也关系到了 HS 算法的提取规则的准确性。基于一定误差的 FCM-RBFNN 能够保证 HS 的精度, 有一定的实用性。这为智能处理应用于渔业研究提供了理论和方法。

参考文献:

[1] 沈金鳌, 方瑞生. 浙江近海冬汛带鱼渔获量预报方法的探讨[J].水产科技情报, 1982, 5: 73-77.

[2] Laurent D, Michel P, Stretta J M. Simulation of large scale tropical tuna movements in relation with daily remote sensing data: the artificial life approach [J].Biosystems, 1997, 44(3): 167-180.

[3] 冯波, 陈新军, 许柳雄. 应用栖息地指数对印度洋大眼金枪鱼分布模式的研究[J]. 水产学报, 2007, 6: 805-812.

[4] 杨烨博. 面向渔情的智能处理模型及其应用——以印

度洋大眼金枪鱼延绳钓渔业为例[D]．上海：上海水产大学, 2008．

[5] 梁海伶. 和声搜索算法在函数优化问题中的应用研究[D]．沈阳：东北大学, 2009.

[6] Andrews R, Diederich J, Tickle A B. A survey, critique of techniques for extracting rules from trained artificial neural networks[J]. Knowledge-Based Systems, 1995, 8(6): 373-389.

[7] Hruschka E R, Ebecken N F. Extracting rules from multilayer perceptions in classification problems: A clustering-based approach [J]. Neurocomputing, 2006, 70: 384-397.

[8] Zbakir L, Baykasog Lu A, Kulluk S. Rule extraction from neural networks via ant colony algorithm for data mining applications [J]. Lecture Notes in Computer Science, 2008, 5313: 177-191.

[9] Bojarczuk C C, Lopes H S, Freitas A A. A constrained-syntax genetic programming system for discovering classification rules: Application to medical data sets[J]. Artificial Intelligence in Medicine, 2004, 30: 27-48.

[10] Omran M G H, Mahdavi M. Global-best harmony search[J]. Applied Mathematics and Computation, 2008, 198(2): 643-656.

[11] Elalfi A E, Hauque R, Elalami M E. Extracting rules from trained neural network using GA for managing E-business[J]. Applied Soft Computing, 2004, 4(1): 65-77.

[12] 陈雪冬, 崔雪森. 卫星遥感在中东太平洋大眼金枪鱼渔场与环境关系的应用研究[J]. 遥感信息, 2006, 1: 25-28.

Extracting rules based on neural network and its application in fisheries forecasting

YUAN Hong-chun[1], HU Qian-qian[1], SHEN Xiao-qian[1], CHEN Xin-jun[2]

(1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. College of Ocean Science, Shanghai Ocean University, Shanghai 201306, China)

Received: Dec., 9, 2012

Key words: the Indian Ocean big eye tuna (*Thunnus obesus*); radial basis function neural network (RBFNN); harmony search; rule extraction; fishery forecasting

Abstract: In order to solve the issue of blind fishing, which arises from over-reliance on experience in offshore fishing, marine environmental and historical production data have been used to effectively analyze the fishery. This method was proposed to forecast indices of the Indian Ocean big eye tuna's (*Thunnus obesus*) habitat based on radial basis function neural network (RBFNN). Fuzzy c-means clustering algorithm was utilized during training the neural network. While in the process of rule extraction, a harmony search algorithm was used to extract fishery rules from the trained RBFNN. Finally, the proposed method was used to forecast the fishery habitat indices of the Indian Ocean big eye tuna. Experiments showed that harmony search algorithm can extract classification rules from the trained neural network. The extracted rules reflected the status of the Indian Ocean big eye tuna fishery.

(本文编辑：刘珊珊　李晓燕)