基于免疫进化算法的贝叶斯网络预测网箱转移周期

邬华俊 耿 冰 滕丽华*
(浙江万里学院生物与环境学院,宁波 315100)

摘 要 以象山港网箱养殖区 2000～2006 年的监测数据作为训练数据,结合专家知识采用基于免疫进化的贝叶斯网络结构增量式学习算法,构建了海底网箱转移的贝叶斯网络预测模型。该模型能有效的揭示出网箱养殖环境各个指标之间的因果关系,进而可以对指定的网箱养殖的网箱转移周期进行预测和决策。结果表明,评价的准确性是 91.7%,证明该方法是有效可行的。

关键词 转移周期 网箱养殖 贝叶斯网络 免疫进化算法 增量学习
中图分类号 S955.9 **文献识别码** A **文章编号** 1000-7075(2009)06-0136-06

Predicting the shift cycle of the net-cage by the Bayesian network based on immune evolutionary algorithms

WU Hua-jun GENG Bing TENG Li-hua*
(College of Biological and Environmental Science of Zhejiang Wanli University, Ningbo 315100)

ABSTRACT By taking the monitoring data of Xiangshan Bay from the year of 2000 to 2006 as the training data and referring to the prior knowledge, a Bayesian network was constructed through the incremental learning based on the immune heredity algorithm. The model can effectively express the causal relationship among the various indicators in the net-cage aquaculture environment, and the shift cycle of the net-cage aquaculture at Xiangshan can be predicted. The result showed that the appraisal accuracy reached 91.7%, which meant that this method is feasible.

KEY WORDS Shift cycle Aquaculture Bayesian network
Immune evolutionary algorithm Incremental learning

宁波市象山港是浙江省重要的海水增养殖基地,目前港内养殖网箱已达 6.6 万只,趋于饱和,与此同时,由于在网箱养殖过程中,大量残饵加上鱼类排泄物,使水体富营养化,导致水域自身污染。同时对底质的影响也很明显,残饵等有机物不断积累的过程中,使底质变黑发臭,产生大量硫化物危及鱼类生存。为改善港区网箱转移水域环境质量,综合治理水域污染,该港除推广科学投饵和使用配合饲料外,开始尝试实行海区轮养(蔡燕

国家科技部项目(2007DFA21300)和宁波市海洋渔业局项目(甬海办 2005/331-6)共同资助
* 通讯作者。E-mail:tlh_98@163.com
收稿日期:2008-08-02;接受日期:2008-11-10
作者简介:邬华俊,男(1987-),本科,主要从事环境科学研究。E-mail:maxl@nbyz.gov.cn

红等　2002)。但养殖模式确定后网箱何时移动尚需科学论证。此外，目前国内对网箱转移周期进行预测的模型研究尚未见报道。本文采用贝叶斯网络方法进行建模，对指定的网箱养殖区的网箱转移周期进行预测和决策。贝叶斯网络是人工智能领域处理不确定性的主要方法之一，广泛应用在现代专家系统、诊断系统及决策支持系统中，其主要优势在于具有坚实的理论基础，能够有效的处理不完整数据，与其他技术相结合进行因果分析，能够使先验知识和数据有机的结合(Rish *et al*.　2000；Acid *et al*.　2005)。

1　贝叶斯网络和增量学习

贝叶斯网络是图形表示和概率知识的有机结合，是复杂联合概率分布的图形表示方式。它提供了一种自然的表示因果信息的方法，用来发现数据间的潜在关系。在这个网络中，用节点表示变量，有向边表示变量间的依赖关系，它揭示了领域对象的内在联系(滕丽华　2008 a)。贝叶斯网络的形式化定义是(Solares *et al* 2005)：贝叶斯网络是一个二元组 $S=\langle G,P\rangle$。其中，G 是有向无环图，图中节点与领域知识的随机变量一一对应；网中的有向弧表示变量间的因果关系，从节点 X 到节点 Y 的有向弧的直观含义是 X 对 Y 有直接的因果影响；$P=\{P(X|Parent(X))\}$是局部概率分布的集合(滕丽华等　2008 b)，条件概率表示因果影响的强度，其中 $Parent(X)$代表节点 X 的父节点集合。该问题域中变量集合的联合概率分布可以表示成贝叶斯网络中的每个节点的条件概率表的乘积，即：

$$P(X_1\cdots X_n)=\prod_{i=1}^{n}P(X_i\mid Parents(X_i))$$

图 1 表示了一个包含 5 个变量的贝叶斯网络结构及局部条件概率分布。

增量学习是以新数据顺序更新学习结果的在线学习过程，与批量学习不同，它不丢弃已有的工作，而是不断地利用新数据更新和求精已经学习到的结果，具有纠正存在于结果模型中的错误和适应基本概率分布发生变化的能力。增量学习算法的思想是当获得新数据时，首先检查当前模型是否能很好的反映新数据，若能，则不进行学习，直接使用当前模型作为学习结果；若当前模型不能很好的反映新数据，则将新旧数据合并，然后再对合并数据集进行学习，将学得的结果作为学习结果。而本文所涉及的学习问题，更适合采用增量式的学习，所以对于　鹏等(2006)提出的算法进行了改进，形成了改进的基于免疫遗传算法的贝叶斯网络增量学习算法。该算法在原算法的基础上改进和增加了部分免疫算子，通过进化过程中的适应度函数来使个体适应新数据，通过接种疫苗，使个体与旧数据也有较好的匹配，使得学习算法更加适合增量学习环境和网箱转移周期决策问题。

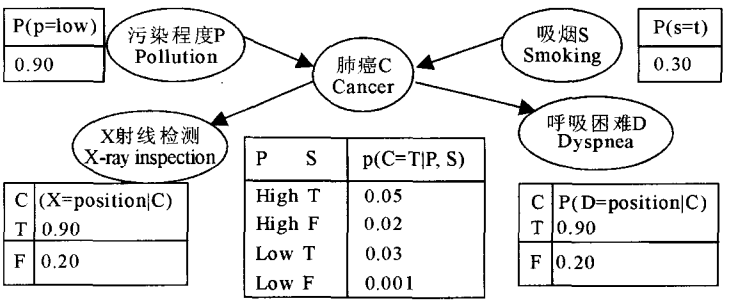

图 1　一个贝叶斯网络的示例

Fig. 1　A sample of Bayesian network

2　网箱转移周期的贝叶斯网络模型构建

对于网箱养殖区，网箱转移的环境评价指标体系由 n 个指标组成。就可以建立一个 $n+1$ 个节点的贝叶斯网络，其中 n 个指标对应网中的 n 个节点，网箱转移周期作为另一个节点，将网箱转移周期分为 5 个级别，分别是 level 1：2～3 年；level 2：3～4 年；level 3：4～5 年；level 4：5～6 年；level 5：6～7 年。将每个指标看成一个节点，首先获得所有指标的取值集合及相应的网箱转移周期，将此作为训练数据集合，某个网箱转移区的所有指标的一次取值以及网箱转移周期 $D=\{C_1,C_2,\cdots,C_n,H\}$作为一个训练数据，其中 C_i表示网箱转移的第 i 个指标的取值，H 表示网箱转移周期。这个由 $n+1$ 个变量组成的贝叶斯网络，反映了 n 个指标与网箱转移周期间的因果关系。

由于贝叶斯网络处理的是离散变量，所以需要将每个指标的取值离散化。以象山港网箱养殖区为例，本文获得的原始数据是网箱转移区在 11 个指标下的取值，根据网箱转移区连续数值属性特点，结合 GB 3097-1997《海水水质标准》和 GB 11607-1989《渔业水质标准》，采取有监督离散化算法（李 刚 2001）对指标的取值进行离散化（表 1）。

表 1 水环境指标取值的离散化分布

Table 1 The distribution of the discrete values of the environmental factors

水环境指标 Water invironment indicators	低 Low	正常 Normal	中 Mid	高 High
投饵量 Total feeding(kg/d)	<7	7～8	—	>8
水温 WT(℃)	<15	—	15～28	>29
透明度 Transparency(m)	<0.5	—	0.5～1	>1
pH	<6.5	6.5～8	—	>8
DO (mg/L)	<5	—	5～7	>7
COD (mg/L)	<1	—	1～2	>2
总磷 TP(mg/L)	—	<0.05	0.05～0.07	>0.07
总氮 TN(mg/L)	—	<1.2	1.2～1.5	>1.5
叶绿素 a Chlorophyll a(μg/L)	—	<5	5～8	>9
海水流速 Velocity (cm/s)	<20	20～40	—	>40
微生物降解能力 Bio-degradation(μmol/ml/d)	<10	10～20	—	>20

部分训练样本数据记录如表 2，表 2 中的网箱转移周期的计算根据网箱养殖年限和渔业养殖水质标准类别确定。

表 2 2006 年象山港网箱养殖区监测数据

Table 2 The monitoring data of the net-cage aquaculture at Xiangshan Bay in 2006

水温 WT (℃)	pH	DO mg/L	COD mg/L	总磷 mg/L (TP)	透明度 (m)	总氮 (mg/L)	投饵量 Total feeding (kg/d)	叶绿素 a Chlorophyll a(μg/L)	海水流速 Velocity (cm/s)	微生物降解能力 Biodegradation (μmol/ml/d)	网箱转移周期 Shift cycle (a)
7.2	9.1	9.47	1.65	0.048	0.8	0.87	7.85	0.30	40	15.6	5.0
12.7	8.01	9.70	0.81	0.054	1.2	0.86	7.85	0.20	30	19.6	5.1
13.1	8.03	9.76	0.93	0.059	0.6	0.84	7.88	0.50	20	20.8	5.2
16.1	7.96	8.10	0.79	0.058	0.3	1.23	8.21	0.41	35	13.5	5.3
20.7	7.97	6.80	0.88	0.069	1.4	1.33	8.55	0.63	60	8.2	5.4
22.5	7.94	6.94	0.7	0.059	0.7	1.42	9.56	0.84	50	6.2	5.5
29.3	7.88	6.06	0.99	0.078	0.7	0.70	9.90	4.32	70	11.0	5.6
29.6	7.93	5.64	0.99	0.082	0.9	0.82	9.80	5.04	50	17.2	5.7
27.5	7.90	6.18	0.95	0.063	0.3	0.98	9.80	2.97	35	10.1	5.8
25.1	7.94	5.92	0.8	0.053	0.5	0.88	7.60	3.24	30	10.6	5.9
15.6	8.04	6.24	1.43	0.108	0.8	1.12	6.20	1.01	40	12.9	5.9
9.2	8.76	6.27	1.63	0.089	0.9	1.23	5.60	1.61	25	12.3	6.0

贝叶斯网络是对包含定性知识和定量知识进行结构上的描述，为下一步推理提供依据。从原始数据中构造贝叶斯网络，实际上是对原始数据进行数据挖掘。构造贝叶斯网络首先找出最符合原始数据的定性网络图关系，然后根据网络图中的因果关系，计算结点间的条件概率。基于免疫遗传算法的贝叶斯网络增量学习算

法,其算法框架为:

算法 1　//增量学习

(1)读入新数据 D',使用当前 Bayesian 网 S,检测 S 是否能够很好的反映新数据;(2) 若 S 能很好的反映新数据 D',则不进行学习,直接将 S 做为本次学习结果。转(3)。否则,将新数据 D' 与旧数据 D 合并,使用算法 2 对 $D+D'$ 进行学习,得出的结果 S' 作为本次的学习结果,$S=S'$;(3) $D=D'+D$;(4) 若仍有新数据,则转 1。否则算法结束。

算法 2　//用免疫遗传算法学习最优网络结构

(1) 以 S 为基础,生成初始群体,并用其作为第 0 代群体 Pop(0),从中选出一个个体作为最优贝叶斯网络结构 Sb,$t=0$;生成免疫疫苗;(2) 对于当前前代 Pop(t)的每一个个体 Stj,计算该个体的适应度函数 $F[Stj]$;(3) 执行交叉,变异生成子代群体;(4) 对子代群体中的每一个个体接种疫苗;A:提高个体与旧数据的匹配程度,适应增量环境,B:把营养盐作为直接影响网箱转移周期的节点,C:将表示"网箱转移周期"的节点置为唯一的叶节点;(5) 进行选择,进而形成新一代群体 Pop($t+1$);(6) $t=t+1$;(7) 在新一代群体中找出适应度最高的个体作为当前最优结构 Sb;(8) 如果已经进化了 g1 代或连续 g2 代最优网络结构没有变化,则算法结束,否则返回(2)。

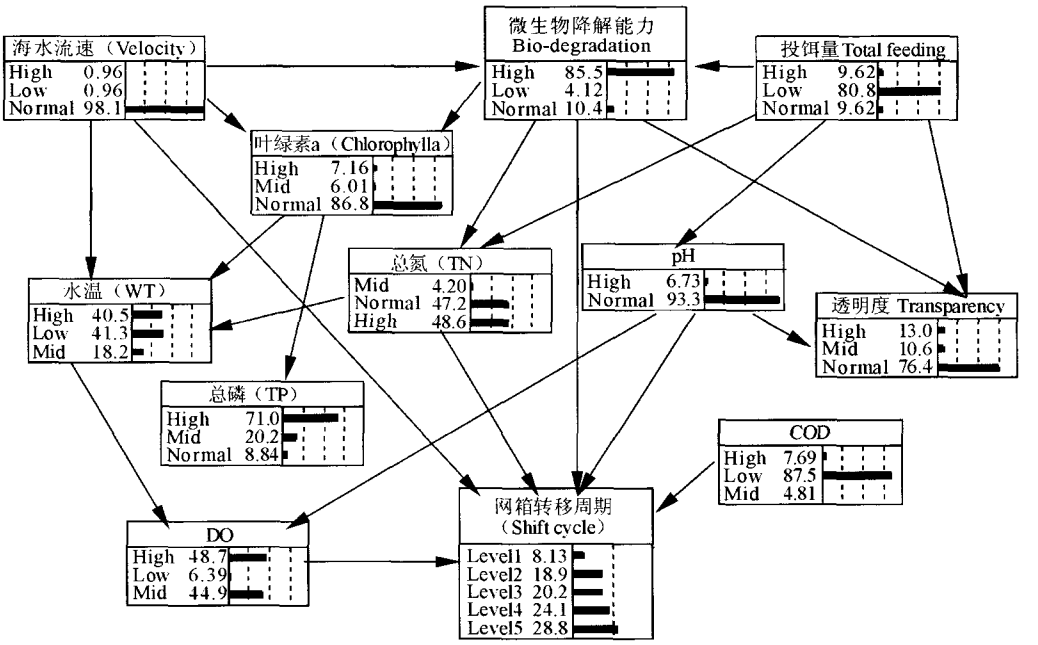

图 2　网箱转移周期的贝叶斯网络结构模型

Fig. 2　The model based on Bayesian network for the shift cycle of the net-cage

通过象山港网箱养殖区 2000～2006 年监测数据的结构学习与专家经验知识,获得了象山港网箱转移周期的贝叶斯网络模型(图 2),该结构反映了各指标之间的因果关系,条件概率表示因果影响的强度。从网箱转移周期的贝叶斯网络评价模型(图 2)可以看出,溶解氧、海水流速、微生物降解能力、总氮、COD 和 pH 这 6 个变量直接影响网箱转移周期,属于影响网箱转移的周期的主导因素,应给予较大的权重;而投饵量、总磷、透明度、水温和叶绿素 a 这 5 个变量对网箱转移周期不产生直接的影响,反映出实际存在的关联性,在判断网箱转移周期等情况时应赋予较小的权重。因此,只需测定与网箱转移周期直接相关的溶解氧、海水流速、微生物降解能力、COD、总氮和 pH 这 5 个变量的值就能通过获得的贝叶斯网络得出网箱转移周期。同时,从图 2 中可以看出,各个指标间的因果影响关系和强度,海水流速直接影响微生物降解能力,投饵量直接影响水体的 pH 值,微生物降解能力和 pH 和透明度。叶绿素 a、海水流速与水温,叶绿素 a 与总磷,水温、pH 与溶解氧都存在直接

的关联性，表明投饵量对微生物降解能力有一定影响，同时微生物降解能力还受海水流速的影响；微生物降解能力和海水流速直接影响叶绿素 a 的含量。

3 模型验证与推理

作者用 2000～2006 年的数据构建了基于网箱转移周期的贝叶斯网络预测模型，之后用 2007 年的数据对获得的模型进行检验，结果发现网箱转移周期预测精度达 91.7%。通过 Zhang 等（1996）给出的贝叶斯网络推理算法（变量消去法）来对网箱转移周期进行预测和决策。该方法的基本思想是：待计算的条件概率为 $P(X_Q \mid X_E = x_E)$，这里 X_Q 为查询变量的集合，X_E 为给定证据的变量集合，令 X_H 为全体变量中 X_Q 和 X_E 之外的变量。

$$p(X_Q \mid X_E) = \frac{P(X_Q, X_E)}{P(X_E)} = \frac{\sum_{h \in QUE} p(X_H = h, X_Q, X_E)}{\sum_{h \in E} p(X_H = h, X_E)}$$

该方法进行每次消除一个变量的循环操作，直到得到 $P(X_Q \mid X_E)$ 为止。它按照以下规则来消去变量：如果 X 是代表证据的变量，则用 $\sum_X$ 操作消除条件概率分布中的 $P(X \neq x_e \mid Pa(X))$；如果 $X = X_Q$，不进行求和操作，即不消除 X；如果 X 为其他情况，则进行求和操作，消除变量 X。以上过程一直进行，并经过最后的归一化过程，可以得到 $P(X_Q \mid X_E)$。

以 2007 年 4 月在象山港 XS06 站点的监测数据为例来预测网箱转移周期，作者只需获得与网箱转移周期直接相关的 6 个指标数据即可（表 3）。

表 3 XS06 站点一组监测数据

Table 3 A set of monitoring data at XS06

项 目 Monitoring data	COD (mg/L)	DO (mg/L)	pH	海水流速 Velocity (cm/s)	总氮 TN (mg/L)	微生物降解能力 Bio-degradation (μmol/ml/d)
监测数据 Monitoring data	1.20	7.71	7.9	35	0.85	22.8
离散后数据 Discrete data	中 Mid	高 High	正常 Normal	正常 Normal	正常 Normal	高 High

把离散后数据代入所获得的贝叶斯网络模型，得到该指标下的网箱转移周期的条件概率分布如下：

P(网箱转移周期=level 1 | COD=mid,DO= high,pH=normal,海水流速=normal,总氮=normal,微生物降解能力=high)=0.4%。

P(网箱转移周期=level 2 | COD=mid,DO= high,pH=normal,海水流速=normal,总氮=normal,微生物降解能力=high)=4.8%。

P(网箱转移周期=level 3 | COD=mid,DO= high,pH=normal,海水流速=normal,总氮=normal,微生物降解能力=high)=4.55%。

P(网箱转移周期=level 4 | COD=mid,DO= high,pH=normal,海水流速=normal,总氮=normal,微生物降解能力=high)=85.9%。

P(网箱转移周期=level 5 | COD=mid,DO= high,pH=normal,海水流速=normal,总氮=normal,微生物降解能力=high)=4.39%。

从上面的网箱转移周期的条件概率分布推理计算结果可知，网箱转移周期等于 level 4 的概率最大为 85.9%，因此，我们可以得出网箱转移周期为 5～6 年。

4 结语

将贝叶斯网络引入到网箱转移周期的预测属于首次尝试，本文提出的基于免疫进化算法的贝叶斯网络增量式结构学习的网箱转移周期预测模型，能清晰、直观地揭示出养殖水环境各指标之间以及指标与网箱转移情

况之间的内在因果关系和影响程度，该模型揭示出与网箱转移周期直接有关的 6 个指标分别为溶解氧、海水流速、微生物降解能力、COD、总氮和 pH。因此，只需测定与网箱转移直接相关的这 6 个指标的数据，就可以预测网箱转移周期的概率。本文建立的贝叶斯网络模型能够对网箱转移周期进行有效预测，但对模型的优化还需要更多的工作，还有待于在实践中进一步完善。

参 考 文 献

于　鹏，刘大有，贾海洋，杨　博. 2006. 基于免疫进化算法的 Bayesian 网结构学习. 吉林大学学报(理学版)，44(6)：919～924

李　刚. 2001. 知识发现的图模型方法. 见：中国科学院软件研究所学位论文(5 月)

蔡燕红，项有堂. 2002. 象山港海水养殖功能区环境质量评价. 海洋通报，21(4)：91～95

滕丽华. 2008a. 贝叶斯网和 MAS 在生态工业园中的应用初探. 复杂系统与复杂性科学. 5(1)：81～86

滕丽华. 2008b. 贝叶斯网在生态工业园清洁生产推进中的应用. 计算机与应用化学 25(5)：565～568

Rish，I.，and Dechter，R. 2000. Resolution versus search：Two strategies for SAT. J. Autom. Reasoning，24(1)：225～275

Acid，S.，De，C. L.，ane Castellano，J. G. 2005. Learning Bayesian network classifiers：searching in a space of partially directed acyclic graphs. Machine Learning，59(3)：213～235

Solares，C.，and Sanz，A. M. 2005. Bayesian network Classifiers：An application to remote sensing image classification. EAS Transactions on Systems，4(4)：343～348

Zhang，N. L.，and Poole，D. 1996. Probabilistic conflicts in a search algorithm for estimating posterior probabilities in Bayesian networks. Artificial Intelligence，88：69～100