# 象山港海水养殖区贝叶斯网络水质评价及预测

滕丽华，程利江，王海丽，杨季芳
（浙江万里学院 生物与环境学院，浙江 宁波 315100）

**摘　要**：水质评价是水环境保护与管理的重要环节，传统的评价方法在处理评价中的不确定性、大量信息处理等方面存在局限性。贝叶斯网络可以有效地表达和分析不确定性问题，实现定性分析与定量分析的有机结合。以近 10 a 来象山港海水养殖区的水质监测数据为样本数据，采用贝叶斯网络技术，建立反映各水质指标及水质级别之间相互关系和相互影响强度的贝叶斯网络模型。模型结构表明直接影响水质级别的水质指标为氨氮、化学需氧量、硝酸盐、无机磷和叶绿素 a，而其他亚硝酸盐、无机氮等 4 个水质指标与水质级别存在间接的因果关系。对 200 条监测数据进行模型精度检验，结果表明，其预测精度达94.8%，Kappa 指数为 0.892，这说明采用贝叶斯网络技术对水质进行评价及预测是可行的。

**关键词**：水质；评价；贝叶斯网络
**文献标识码**：A　　**文章编号**：1000-288X(2012)01-0189-03　　**中图分类号**：X832

# Assessment and Prediction of Water Quality in Mariculture Zone of Xiangshan Harbor Based on Bayesian Network

TENG Li-hua, CHENG Li-jiang, WANG Hai-li, YANG Ji-fang
(*College of Biological and Environmental Science, Zhejiang Wanli University, Ningbo, Zhejiang* 315100, *China*)

**Abstract**: Water quality assessment plays an important role in water environmental protection and management. Traditional methods have some limitations in dealing with the uncertainty in assessment and massive information. A Bayesian network model can effectively express and analyze uncertain problems and combine qualitative analysis with quantitative analysis. Based on the mornitoring data for nearly ten years in Xiangshan harbor of Ningbo City, a Bayesian network model expressing the relationships and interactions between different water quality indexes and water quality levels was constructed by Bayesian network approach. The model structure indicated that ammonia nitrogen, COD, inorganic phosphorus, nitrate, and chlorophyll had direct effects on water quality level, whereas there was an indirect causality between other water quality indexes such as nitrite and inorganic nitrogen and water quality level. The results of the model validation using 200 monitoring data showed that the predictive precision reached 94.8% and the Kappa was 0.892, which suggests the Bayesian network is feasible for assessment and prediction of water quality.

**Keywords**: water quality; assessment; Bayesian network

水质评价与水质预测是水环境保护管理与治理的科学依据，评价结果的准确性、客观性至关重要。20 世纪 90 年代以来象山港海水养殖迅速发展，加上工农业、生活污水等陆源污染的入海，导致港内局部海域环境质量下降，影响了海水养殖业的健康可持续发展[1-2]。港内水环境保护及各项规划管理的制定离不开水质的有效评价与预测。目前对水质评估的数学模型很多，如综合指数法，模糊数学法，灰色聚类法等[3]。问题在于水体污染过程是连续的，而监测过程却是间断的，并且某些污染因子的监测数据尽管是明确的信息，但污染过程的全部信息则是不完全的，水环境的水文及排污条件具有随机性，因此水质评价具有一定的不确定性。目前研究较多的水质模型不能很好地解决评价因子与水质等级间复杂的非线性关系。贝叶斯网络具备复杂关联关系表达以及概率推理能力，是人工智能领域处理不确定性信息的有力工具，广泛应用在现代专家系统、诊断系统及决策支持系统中[4-6]。本文利用宁波市象山港海水养殖区近 10 a 的水质监测数据，应用贝叶斯网络技术，建立了反映各水质指标及水质级别之间因果关系和相互影响强度的贝叶斯网络模型，以期提供更加客观、准确的水质信息，为水环境管理与保护提供更加可靠的参考依据。

**收稿日期**：2011-01-14　　**修回日期**：2011-04-24
**作者简介**：滕丽华(1974—)，女(汉族)，吉林省长春市人，博士，副教授，主要从事环境化学与水污染治理研究。E-mail：tlh_98@163.com。

## 1 贝叶斯网络简介

贝叶斯网(Bayesian networks)是在 20 世纪 80 年代发展起来的,最早由 Pearl[7] 于 1988 年提出。贝叶斯网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,对知识结构的自然表述,灵活的推理能力及方便的决策机制使其成为人工智能领域处理不确定性的主要方法之一,在数据分析、模式分类及建模等方面具有很多优点。

贝叶斯网的形式化定义如下[8]:贝叶斯网是一个二元组 $S=<G,P>$,其中:

(1) $G$ 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;

(2) $P=\{P(X \mid \mathrm{Parent}(X))\}$ 是局部概率分布的集合,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。

给出一个贝叶斯网的例子,Pearl 教授家安装了报警铃,发生地震和盗窃时都有可能引起家里的警铃发出警报,听到警报后,Pearl 的两个邻居 Mary 和 John 有可能会打电话给警察。针对此问题的贝叶斯网结构及其条件概率表如图 1 所示,贝叶斯网的通常用法是先利用专家经验或者利用机器学习方法从数据中学习或者二者的结合来构建贝叶斯网,然后通过贝叶斯网的推理来进行某些决策。其中贝叶斯网的推理是指在给定一组证据变量确切值的情况下,计算一组查询变量的概率分布。

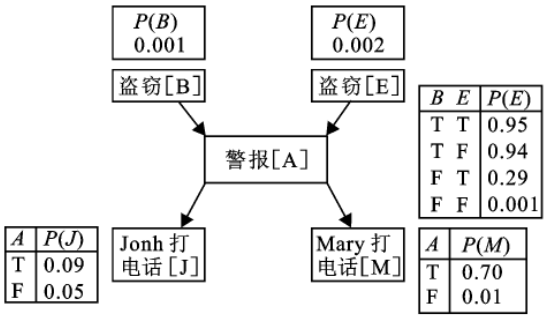


**图 1 Bayesian 网的一个简单示例**

## 2 贝叶斯网络构建

### 2.1 水质指标选取

在对水质类别进行预测评价时,如果对所有监测指标不加选择地一概用来评价分析,不仅计算量大、模型结构复杂,而且往往由于变量间自相关性的影响,使得评价结果的精度下降。主成分分析法是把多个指标化为少数几个综合指标的一种统计分析方法,通过降维,找到少数几个综合因子来代表众多的指标,使这些综合因子尽可能地反映原来众多指标的信息,并且彼此之间互不相关。以近 10 a 象山港水质监测数据为样本数据,选取了 11 项水质指标:盐度($C_1$),pH($C_2$),溶解氧(DO)($C_3$),化学需氧量(COD)($C_4$),$NO_2^-$—N($C_5$),$NO_3^-$—N($C_6$),$NH_4^+$—N($C_7$),无机氮($C_8$),无机磷($C_9$),叶绿素 a($C_{10}$)和粪大肠菌群($C_{11}$)。为构建简单、直观、准确的模型结构,首先利用主成份分析确定影响水质的主要水质指标。

表 1 是用 MATLAB 7.0 软件对水质参数数据求得的相关系数对角矩阵。通过相关系数矩阵计算特征向量、特征值和累计贡献率,表 2 为水质参数的第 1,第 2 主成分系数权重和累积贡献率。由表 2 可知,第 1,第 2 主成分累积贡献率为 88%,满足了累积贡献率大于 85%的要求。由表 2 中主成分系数大小可知粪大肠菌群的权重较小,因此,通过主成分分析法确定了用于建立水质类别模型的 9 个关键指标为:溶解氧、化学需氧量、亚硝酸盐、硝酸盐、氨氮、无机磷、无机氮、pH 值、叶绿素 a。

### 2.2 水质分类的贝叶斯网络模型

2.2.1 模型变量离散化 采取均值化的方法对原始数据进行无量纲化,根据网箱养殖区连续数值属性特点,采用离散化算法[9-10] 对指标的取值进行离散化。离散化后各个指标的取值分别如下。pH 值:低,中,高;溶解氧:低,中,高;化学需氧量:低,中,高;亚硝酸盐:低,中,高;硝酸盐:低,中,高;氨氮:低,中,高;无机氮:低,中,高;无机磷:低,中,高;叶绿素 a:低,中,高;水质分类:Ⅰ,Ⅱ,Ⅲ。

2.2.2 水质模型及分析 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,计算结点间条件概率[10]。本文利用著名的 EM 算法[11-12],通过 1 000 个数据实例的结构学习,构建出包含溶解氧、化学需氧量、亚硝酸盐、硝酸盐、氨氮、无机磷、无机氮和叶绿素 a 和水质级别共 9 个节点的贝叶斯网络模型如图 2 所示。

由图 2 模型结构中有向边的指向关系可知,各水质指标间存在着直接或间接的因果影响关系,例如对于水质指标中的化学需氧量,溶解氧、无机氮和亚硝酸盐对其有直接的影响关系,而其又直接影响水质类

别和氨氮。对于模型中的目标节点水质类别,直接对其产生因果影响关系的水质指标是氨氮、化学需氧量、硝酸盐、无机磷和叶绿素 a,而其他亚硝酸盐无机氮等 4 个水质指标与水质级别具有间接的因果关系。

表 1　象山港各水质参数的相关系数对角矩阵

| 水质参数 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ | $C_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_1$ | 1.000 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $C_2$ | 0.119 8 | 1.000 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $C_3$ | −0.389 0 | −0.571 4 | 1.000 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $C_4$ | 0.256 2 | 0.409 4 | −0.510 3 | 1.000 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $C_5$ | 0.607 6 | 0.363 4 | −0.480 9 | −0.006 0 | 1.000 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $C_6$ | −0.889 9 | −0.130 5 | 0.651 4 | −0.379 5 | −0.621 1 | 1.000 0 | 0 | 0 | 0 | 0 | 0 |
| $C_7$ | 0.293 0 | −0.016 1 | −0.044 3 | 0.716 7 | −0.077 7 | −0.249 2 | 1.000 0 | 0 | 0 | 0 | 0 |
| $C_8$ | −0.799 9 | −0.116 7 | 0.647 5 | −0.137 3 | −0.617 5 | 0.934 7 | 0.105 9 | 1.000 0 | 0 | 0 | 0 |
| $C_9$ | −0.707 3 | 0.155 7 | 0.449 0 | −0.543 5 | −0.232 9 | 0.820 3 | −0.452 1 | 0.695 4 | 1.000 0 | 0 | 0 |
| $C_{10}$ | 0.541 2 | 0.545 4 | −0.728 1 | 0.425 9 | 0.854 9 | −0.680 6 | 0.114 7 | −0.620 7 | −0.453 1 | 1.000 0 | 0 |
| $C_{11}$ | −0.265 8 | −0.058 0 | 0.072 0 | −0.386 4 | −0.091 8 | 0.239 3 | −0.222 6 | 0.166 1 | 0.220 0 | −0.166 2 | 1.000 0 |

表 2　象山港网箱养殖区各水质参数的主成分系数权重及累积贡献率

| 主成分 | 各水质参数的主成分系数 | | | | | | | | | | | | 累积 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ | $C_{11}$ | 特征值 | 贡献率 |
| 第 1 主成分 | −0.11 | 0.35 | 0.26 | 0.36 | 0.36 | 0.36 | 0.36 | 0.36 | 0.30 | −0.01 | −0.25 | 7.71 | 0.70 |
| 第 2 主成分 | 0.19 | −0.08 | −0.43 | 0.01 | 0.04 | 0.04 | 0.03 | 0.04 | 0.27 | −0.56 | −0.24 | 1.94 | 0.88 |

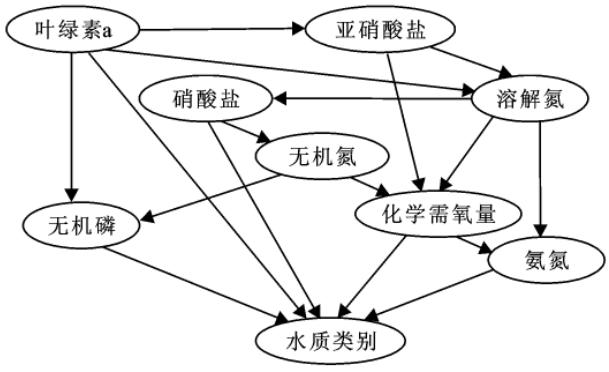


图 2　水质指标与水质级别间关系的贝叶斯网络模型

通过 Zhang 等[13]给出的贝叶斯网推理算法(变量消去法)对各指标状况进行预测和决策。根据模型的因果推理功能,在一定的水质条件下只需测定具有直接影响关系的氨氮、化学需氧量等水质指标即可获知水质级别属性的概率,当水质恶化时可利用模型的诊断推理功能,找到使水质级别低下的主要影响指标。以上这些指标间的直接影响程度可从条件概率分布表中看出,而各指标的间接影响程度可以通过推理计算得出。

## 3　模型检验

基于该模型,对其中的 200 条数据实例进行模型精度检验,表 3 是贝叶斯网络水质分类精度的混淆矩阵列表,2×2 表的 Kappa 系数:

〔(123+67)/200 −(129×127+71×73)/40 000)〕/〔1−(129×1 278+71×73)/40 000〕=0.892

其预测精度达 94.8%,这说明采用贝叶斯网络技术对水质进行评价及预测是可行的。

表 3　贝叶斯网络水质分类精度的混淆矩阵

| 水质分类预测 | 水质分类实际情况 | | |
|---|---|---|---|
| | Ⅰ类 | Ⅱ类 | 总和 |
| Ⅰ类 | 123 | 4 | 127 |
| Ⅱ类 | 6 | 67 | 73 |
| 合计 | 129 | 71 | 200 |
| 预测精度/% | 95.3 | 94.3 | — |

## 4　结论

本文对复杂、充满不确定因素的水质评价进行了贝叶斯网络建模研究,以期获得更客观、准确的水质评价信息,为水环境保护提供更可靠的依据。论文建立了包括 8 个水质指标和一个水质级别共 9 个节点的贝叶斯网络模型,该模型揭示出了各水质指标之间以及各水质指标与水质级别之间的因果关系及影响强度,得到了与水质级别具有直接因果影响关系的水质指标为氨氮、化学需氧量、硝酸盐、无机磷和叶绿素 a,而其他亚硝酸盐、pH 值等 4 个水质指标对水质级别产生间接的影响关系。模型检验表明模型精度达 94.8%,Kappa 指数为 0.892,这表明本文采用贝叶斯网络模型方法对水质进行评价与预测具有一定的可行性。

(下转第 232 页)

且与区域经济发展水平具有较强的相关关系，即经济发展水平较高的省区，协调发展度相对较高。

(4) 根据协调度和协调发展度可以将 31 个省区分为 7 类，分别体现了经济发展、自然条件和政策因素等对区域城市土地集约利用与生态环境作用的差异。

(5) 应加快转变经济发展方式，促进产业结构调整升级，优化土地利用结构，完善市场机制，推进增量用地为主的粗放用地方式向以存量用地为主的集约用地方式转变，不断提高城市土地集约利用水平；加强生态环境保护建设，加大对环境保护建设的投入，依靠科技创新，降低能耗，有效发展循环经济，促进生态环境水平的提高；开展城市生态规划，建立环境友好型土地利用模式；与此同时各个省区依据各自的协调发展程度及时采取相应的政策措施进行宏观调控，进而有效促进城市土地集约利用与生态环境协调发展。

[ 参 考 文 献 ]


[1] 姜志德. 中国土地资源可持续利用战略研究[M]. 北京：中国农业出版社，2004：1-2.

[2] 王恒伟. 基于生态健康的城市土地集约利用[D]. 重庆：西南大学，2010.

[3] 王静，邵晓梅. 土地节约集约利用技术方法研究：现状、问题与趋势[J]. 地理科学进展，2008，27(3)：68-74.

[4] 王兰霞，李巍，王蕾. 哈尔滨市土地利用与生态环境物元评价[J]. 地理研究，2009，28(4)：1001-1010.

[5] 吴郁玲，曲福田. 中国城市土地集约利用的影响机理：理论与实证研究[J]. 资源科学，2007，29(6)：106-113.

[6] 梁红梅，刘卫东，刘会平，等. 深圳市土地利用社会经济效益与生态环境效益的耦合关系研究[J]. 地理科学，2008，28(5)：636-641.

[7] 陈兴雷，李淑杰，郭忠兴. 吉林省延边朝鲜自治州土地利用与生态环境协调度分析[J]. 中国土地科学，2009，23(7)：66-70.

[8] 廖重斌. 环境与经济协调发展的定量评判及其分类体系[J]. 热带地理，1999，19(2)：171-177.

[9] 封毅，阎伍玖，崔灵周，等. 芜湖市经济与环境协调发展类型评价研究[J]. 水土保持通报，2007，27(6)：211-215.

[10] 曹利军，王华东. 可持续发展评价指标体系建立原理与方法研究[J]. 环境科学学报，1998，18(5)：526-532.

[11] 彭建超，徐春鹏，吴群，等. 长三角地区城市土地利用集约度区域分异研究[J]. 中国人口·资源与环境，2008，18(2)：103-109.

[12] 朱天明，杨桂山，苏伟忠，等. 长三角地区城市土地集约利用与经济社会发展协调评价[J]. 资源科学，2009，31(7)：1109-1116.

[13] 刘耀彬，李仁东，宋学锋. 中国城市化与生态环境耦合度分析[J]. 自然资源学报，2005，20(1)：105-112.

[14] 江红莉，何建敏. 区域经济与生态环境系统动态耦合协调发展研究[J]. 软科学，2010，24(3)：63-68.

[15] 陶晓燕，章仁俊，徐辉，等. 基于改进熵值法的城市可持续发展能力的评价[J]. 干旱区资源与环境，2006，20(5)：38-41.

[16] 郑华伟. 基于改进熵值法的耕地利用集约度评价[J]. 新疆农垦经济，2010，29(4)：53-58.


(上接第 191 页)

[ 参 考 文 献 ]


[1] 罗益华. 象山港海域水质状况分析与污染防治对策[J]. 污染防治技术，2008，21(3)：88-90.

[2] 蔡惠文，孙英兰. 象山港网箱养殖对海域环境的影响及其养殖环境容量研究[J]. 环境污染治理技术与设备，2006，7(11)：71-76.

[3] 马文涛. 支持向量机回归方法在地表水水质评价中的应用[J]. 数学的实践与认识，2007，37(9)：45-50.

[4] 仲少云，王庆，金秉福. 模糊综合评判法在芝罘湾水质评价中的应用[J]. 海洋湖沼通报，2007(2)：57-61.

[5] 邵继业，王日新，徐敏强. 贝叶斯网络在模型诊断中的应用[J]. 吉林大学学报，2010，40(1)：234-237.

[6] 胡学钢，胡春玲. 一种基于依赖分析的贝叶斯网络结构学习算法[J]. 模式识别与人工智能，2006，19(4)：445-449.

[7] 胡春玲，胡学钢. 一种具有缺失数据的贝叶斯网络结构学习方法[J]. 合肥工业大学学报：自然科学版，2007，30(4)：449-453.

[8] Pearl J. Probabilistic reasoning in intelligent systems: Networks of plausible inference[M]. San Mateo, California: Morgan Kaufmann, 1988.

[9] Solares C, Sanz A M. Bayesian network classifiers: an application to remote sensing image classification [J]. EAS Transactions on Systems, 2005, 4(4): 343-348.

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

[11] 李刚. 知识发现的图模型方法[D]. 北京：中国科学院软件研究所，2001.

[12] Cooper, Herskovits. A Bayesian method for the induction of probabilistic networks from data [J]. Machine Learning, 1992, 9(4): 309-347.

[13] Zhang N L, Poole D A. simple approach to Bayesian network computations[C]// Proceedings of the Tenth Canadian Conference on Artificial Intelligence. 1994: 171-178.