基于决策树技术及在线监测的水质预测

卢金锁[1], 黄廷林[1], 韩宏大[2], 何文杰[2], 阴培军[2]
(1.西安建筑科技大学环境与市政工程学院, 陕西 西安 710055;
2.天津市自来水集团有限公司, 天津 300040)

摘 要: 利用北方某城市水源的水质在线监测系统, 建立了基于决策树技术, 具有较强可视性和实际应用, 以及能预测次日源水中叶绿素水平的决策树模型。该模型将某城市水源在线监测的溶解氧和太阳辐射照度数据转换计算为每日平均标准偏差及均值, 并与每日定时取样测定的叶绿素含量一起作为预测因子, 通过将 115组数据的前 100组数据作为训练集建立预测次日叶绿素水平决策树模型, 并采用后 15组数据进行模型的仿真预测检验, 结果只有 3 d的预测出错, 预测准确率达 80%。并讨论了模型建立对数据的要求及解读预测规则等问题。

关键词: 决策树; 水质预测; 叶绿素; 水源; 在线监测
中图分类号: X84 文献标识码: B 文章编号: 1006-2009(2006)02-0038-04

Forecast Water Quality Based on Decision-making Tree and Online Monitoring

LU Jin-suo[1], HUANG Ting-lin[1], HAN Hong-da[2], HE Wen-jie[2], YIN Pei-jun[2]
(1. *Environment and Municipal Engineering School, Xián University of Construction Science and Technology, Xián, Shanxi* 710055, *China*; 2. *Tianjin Tap Water Ltd, Tianjin* 300040, *China*)

Abstract This article set up the technology of decision-making tree used the on-line water monitoring system from the water source of a North-China city, and it has strong objective and practical application, as well as the new decision-making model can predict the chlorophyll level of water sources next day. The model convert the dissolved oxygen from on line monitoring in urban water source and the data of solar radiation illumination into the average daily standard deviation and dispersion, which as the forecast factors with the regular daily sample measurement of chlorophyll content. There are 115 sets of data, the previous 100 sets of them will be as a training group to establish the decision-making tree model for forecasting the chlorophyll levels of the following day, the follow 15 sets of them used as forecasting test results of simulation models. The result has shown only 3 d forecast was error, the accurately rate is 80%. In addition it discussed the data of model building and interpretation of forecasting rules.

Key words Decision-making tree; Water quality forecasting; Chlorophyll; Water source; Online monitoring

近年来, 各城市供水水源的藻类高发频率和程度都有所增加, 致使水厂在此期间出现产水量减少、运行费用增加及出厂水质恶化等现象。为保证出水质量, 开展了一系列针对藻类的应急处理措施, 如投加液氯、二氧化氯、硫酸铜、高锰酸钾及其复合药剂等化学杀藻工艺, 投加粉末活性碳的吸附工艺, 以及强化混凝和优化等传统工艺的应急处理工作。尽管应急处理能在一定程度上确保出水水质达标, 但部分工艺对水质及人体健康存在潜在的影响, 此外, 应急处理工艺是在滤后水的浊度升高、pH下降等水质恶化现象出现后才启动, 而且为完全确保出水达标, 应急工艺通常需持续运行一段时间, 造成运行费及对健康潜在风险的增加。

收稿日期: 2005-08-12; 修订日期: 2005-12-20
基金项目: 国家高技术研究发展计划“八六三”基金资助项目(2002AA601140)
作者简介: 卢金锁(1977—), 男, 甘肃会宁人, 博士研究生, 助教, 从事水源保护及微污染水处理技术研究。

长期的源水监测发现, 受降雨、日照等天气因素影响, 水中藻类数量变化很大, 有时出现叶绿素含量在 1 d之内降幅达 30 mg/m^3, 故依据藻类变化趋势, 启动、停止及实时调整应急工艺很有必要, 但建立具有可实时监测和预测源水水质, 并能提出相应运行方案的源水水质预警系统更有必要。现通过北方某城市水源建立的水质在线监测系统, 应用决策树技术分析高频在线监测数据, 建立便于实际应用的藻类变化“树”模型预测藻类高发, 实现保障供水安全的目的。

1 预测可行性及方法选择

基于在线监测仪的生产现状、建设费用和监测指标对水厂运行的重要性, 目前开展的环境监测指标有溶解氧、电导率、pH、ORP(氧化还原电位)、水温、浊度、氨氮和正磷酸盐, 以及水池水位及太阳辐射照度等。

当前, 反映水体质量的藻类监测指标主要是藻类和叶绿素等。藻类计数法是用显微镜观测水体中藻类的个数, 数据简单直观。所有藻类均包含叶绿素, 其含量约占有机体干重的 2%。叶绿素是藻类光合作用的重要组成物质, 叶绿素能比较准确地反映藻类生物量, 通过预测叶绿素可知源水中藻类生物量大小。

天然水中溶解氧主要是由大气通过水气界面进入, 其次是由水生植物光合作用放出氧。在标准大气压下, 水中溶解氧含量随温度升高而降低。白天光照时, 水中藻类等浮游生物大量繁殖, 光合作用加强, 在此过程释放出氧气, 水中溶解氧含量增加至最大; 夜间, 浮游生物因呼吸消耗氧气, 使水中溶解氧含量减小, 故溶解氧含量因光合作用和各种水生生物的呼吸作用而呈一定规律的周期性变化。另外, 藻类繁殖时, 大量吸收二氧化碳, 使光合水层中二氧化碳浓度降低, 引起 pH 上升[1], pH 上升可以反映水生植物的光合成和呼吸速度的日变化。此外藻类生长和大多数生物生长一致, 其绝对增加量与原生物量密切相关, 所以从预测角度看, 当前叶绿素量反映了源水中的藻类生物量, 而溶解氧的昼夜变化反映了藻类生物量的变化速率。

合适的水温能促进藻类生长, pH、溶解氧能反应藻类生长情况, 总氮、总磷通常是藻类的限制因子, 透明度会导致水体光照强度减弱而影响藻类生长, 所有这些自然规律都是预测藻类高发的重要理论基础。基于此, 许昆灿等[2]利用检测溶解氧与饱和溶解氧差值而建立了表观增氧量海水藻类即赤潮的预测模型; 王正方等[3]根据赤潮发生前期, 赤潮生物白天光照产生大量氧气溶解于水中, 夜间停止光合作用, 因呼吸作用吸收海水中溶解氧、释放二氧化碳, 使海水中溶解氧明显升高或昼夜有明显变化这一特征, 建立了溶解氧赤潮预报模式, 当溶解氧昼夜最大差值达到 5 g/m^3 时, 可以预报赤潮将要发生。黄廷林等[4,5]也就水质预测方法及适用性作了针对性研究, 认为决策树技术是数据挖掘领域中重要的分支, 具有可解读数据规则并可视化、输出结果容易理解、直接利用规则预测且精度较高等优点, 应用广泛。为此提出了将高频监测数据和决策树有机结合的叶绿素预测模型, 该模型可有效计量环境及水质等预测因子对源水中叶绿素变化的影响。这种影响是定量的, 并能直观地以“知识树”(表示预测因子对叶绿素变化影响的规律)形式显示出来, 预测人员可根据自己的知识和实践经验, 对“知识树”修正。

2 决策树方法

决策树[6]是一种以实例(训练集)为基础的归纳学习方法, 本质上是机器学习。它着眼于从一组无秩序、无规则的实例中推理出决策树表示形式的分类规则。采用自顶向下的递归方式, 在决策树内部节点进行属性值的比较, 并根据不同的属性值判断从该节点向下的分支, 在决策树的叶节点得到结论。所以从根到叶节点的一条路径就对应着一条规则, 整棵决策树可解读出一组表达式规则, 其中内部结点表示某种检验属性, 分叉表示检验结果, 叶结点表示类或某一类的分类, 顶点称为根结点。

决策树生成过程见图 1。

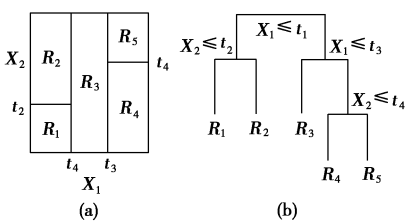


图 1 决策树生成过程

考虑具有两个特征 X_1 和 X_2 的分类问题, 其中 X_1 和 X_2 在单位区间上取值。首先将特征空间划

分为包含不同类别的两个区域,用每一个区域中的类别对特征建模,然后把这些区域中的一个或两个特征进一步分裂,直至满足要求。例如图 1(a)中,首先在特征 $X_1 = t_1$ 处分裂,然后将 $X_1 \leqslant t_1$ 的区域在 $X_2 = t_2$ 处分裂,而 $X_1 > t_1$ 的区域在 $X_2 = t_4$ 处分裂。最终,将整个区域分裂成图中显示的 5个区域 R_1、R_2、…R_5,表示 5个类别。当特征变量空间超过二维时,无法用图 1(a)表示,要用图 1(b)的决策树表示。整个特征数据集位于树的顶部,在每个节点满足条件特征时被赋给左分支,否则赋给右分支,决策树的叶节点对应于区域(或类别)R_1、R_2 … R_5。可将叶节点简单地表示为 IF…THEN 规则,如 R_5 可表示 IF: $X_1 > t_1$ 且 $X_1 > t_3$ 且 $X_2 > t_4$, THEN 观测属于 R_5。

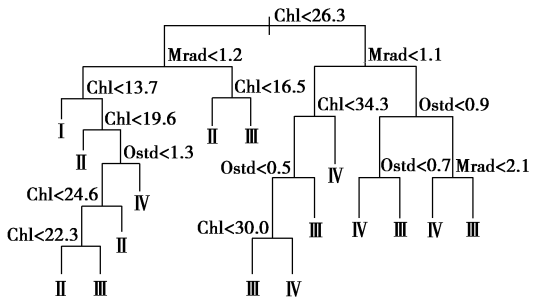
3 决策树预测模型的建立

在所有的监测指标中,选择在理论上对藻类繁殖最有影响及其值准确可靠的 3个指标:溶解氧、太阳辐射照度及当前叶绿素来预测第 2日的叶绿素水平。其中溶解氧及太阳辐射照度共 116 d在线监测的数据,日存储 94组数据,当前叶绿素值共 116个监测记录,直接采用监测值,将次日叶绿素通过 10 mg/m^3、20 mg/m^3、30 mg/m^3 3个阈值,离散化为Ⅰ、Ⅱ、Ⅲ、Ⅳ 4个水平等级,其中叶绿素含量逐级升高。

溶解氧动态变化可表征藻类生长,考虑昼夜最大差值受到溶解氧极值监测误差影响较大,而采用标准偏差,即计算每日监测值的标准偏差,其在统计上可表征溶解氧的变化特性;太阳辐射照度是反映地面亮度的值,在一定程度上反映太阳辐射到水面的能量,受天气影响,在此采用平均太阳辐射照度,可间接反映天气阴晴的变化。考虑预测及验证,共有 115×4个数据可用,其中前 100个作为训练集,用于决策树模型的建立,后 15个作为检验集,用于仿真预测对模型验证。

利用训练集,在 MATLAB环境下生成的叶绿素预测“树”模型见图 2。

图 2可以看出,模型可读性极强,从根节点至叶节点的任何一条路径,均可抽取预测因子对叶绿素变化的影响规律,如其中一条规则可表示为 IF 当前 Chl(叶绿素)监测值不大于 26.3 mg/m^3, Mrad(平均太阳辐射照度) < 1.2×10^4 lx,并且 Chl > 16.5 mg/m^3, THEN 次日叶绿素将达Ⅱ级水平,



Chl——叶绿素, mg/m^3; Ostd——溶解氧标准偏差; Mrad——平均太阳辐射照度, 10^4 lx; Ⅰ Ⅱ ⅢⅣ——次日叶绿素水平

图 2 叶绿素预测“树”模型

即次日源水中叶绿素含量将介于 20 mg/m^3 ~ 30 mg/m^{3}之间。对于新观测值,从根节点开始,逐步满足节点条件者顺左树枝向下,不满足条件者顺右树枝向下,最终到达叶节点,从而预测次日源水中叶绿素水平。对训练集中所有数据应用该“树”模型,其中 87 d“预测”正确,正确率接近 90%,说明模型对训练集数据的拟合效果很好。

4 仿真预测

将检验集中的 15组数据应用模型,对次日叶绿素水平仿真预测,预测结果见表 1。

表 1 预测叶绿素“树”模型仿真结果

日期	Chl ρ(mg·m^{-3})	Ostd	Mrad E/k	次日叶绿素水平 实际	次日叶绿素水平 预测
9月 15日	31.5	0.35	1.07	Ⅳ	Ⅳ
9月 16日	33.8	0.53	0.78	Ⅳ	Ⅳ
9月 17日	33.8	0.18	0.42	Ⅲ	Ⅳ
9月 18日	29.2	0.56	1.78	Ⅳ	Ⅳ
9月 19日	31.6	0.52	0.85	Ⅳ	Ⅳ
9月 20日	31.8	0.47	1.78	Ⅲ	Ⅳ
9月 21日	30.0	0.54	1.75	Ⅳ	Ⅳ
9月 22日	33.4	0.54	1.48	Ⅳ	Ⅳ
9月 23日	38.0	0.65	1.22	Ⅳ	Ⅳ
9月 24日	40.3	0.53	0.70	Ⅳ	Ⅳ
9月 25日	39.0	0.52	0.63	Ⅳ	Ⅳ
9月 26日	37.0	0.64	0.83	Ⅳ	Ⅳ
9月 27日	33.8	0.78	0.91	Ⅳ	Ⅲ
9月 28日	37.1	0.56	0.79	Ⅳ	Ⅳ
9月 29日	36.6	0.52	0.50	Ⅳ	Ⅳ

从表 1可见,模型预测出现 3次错误,其中 9月 17日和 9月 20日预测源水的次日叶绿素水平

为Ⅳ, 但实际含量非常接近或等于 30. 0mg/m^3, 介于水平 Ⅲ和 Ⅳ的分界线上; 9月 27日预测叶绿素水平将下降, 而实际含量上升和水平保持, 说明模型没有“学习”到这样规律, 属于完全错误预测。

5 讨论

模型预测精度较高, 说明模型可基本作规律模拟。但此仿真只利用了树中的 1条或几条树枝或规则对叶绿素水平预测, 决策树仅是部分地被检验, 需更多的数据对模型作进一步验证以确认模型预测的可靠性。

分析模型预测错误的数据, 发现尤其是 9月 27日的现象在训练集中没有出现, 即在相近的叶绿素含量水平和相同的太阳照射下, 源水叶绿素含量上升且水平保持, 说明训练集不能完全预测藻类变化所有情况, 训练数据过少。因此, 对具有较强季节性和周期性的藻类变化至少具备 1年的监测数据用于模型训练, 同时为保证模型可靠应用, 检验集的数据也应尽可能多。

决策树共有 15个节点, 但以叶绿素含量为节点的数达到 8个, 说明当前叶绿素含量是影响次日叶绿素水平的重要因素, 其和藻类现存量极大影响藻类变化的理论相一致。

决策树解读出的规则很难理论解释甚至与现存理论相背。如树节点: Ostd < 0. 7, 从理论上分析, 溶解氧变化幅度越大, 藻类繁殖和呼吸作用越强, 次日叶绿素水平越高, 但树分支与此相反。另外, 决策树节点分裂, 如 Ostd在 0. 7处分裂, 无法解释其物理意义。

6 结语

结合源水在线及人工监测数据首次试探性采用决策树技术, 按照决策树生成过程及算法, 建立预测次日源水中叶绿素水平的决策树模型。通过实测数据对模型仿真检验, 表明预测效果较好, 并分析确定了影响次日叶绿素水平的重要因子为当前源水中叶绿素量、太阳辐射强度和溶解氧变化。尽管模型预测效果较好, 但模型中出现了无法利用现有理论解释的树分支或规则。

[参考文献]

[1] 卢大远, 刘培刚, 范天俞, 等. 汉江下游突发“水花”的调查研究[J]. 环境科学研究, 2000, 13(2): 28-31.

[2] 许昆灿, 暨卫东, 周秋麟, 等. 表观增氧量在近岸海域赤潮快速评价与预警中的应用[J]. 台湾海峡, 2004, 23(4): 417-422.

[3] 王正方, 张庆, 吕海燕, 等. 长江口溶解氧赤潮预报简易模式[J]. 海洋学报, 2000, 22(4): 125-129.

[4] 黄廷林, 卢金锁, 韩宏大, 等. 地表水源水质预测方法研究[J]. 西安建筑科技大学学报(自然科学版), 2003, 36(2): 134-137.

[5] 孙英云, 何光宇, 翟海青, 等. 一种基于决策树技术的短期负荷预测算法[J]. 电工电能新技术, 2004, 23(3): 55-75.

[6] 栾丽华, 吉根林. 决策树分类技术研究[J]. 计算机工程, 2004, 30(9): 94-105.

• 简讯 •

江苏苏南地区 6县级市完成全面小康指标

地处经济发达的苏南地区的昆山、张家港、常熟、吴江、太仓和江阴 6个县级市, 已达到全面小康社会的考核指标, 近日成为江苏省首批全面小康社会达标的县级市。

江苏省建立了全面小康社会指标的评价体系, 结合江苏省省情, 制定出与经济、社会指标相适应的环保指标考核体系, 着重提高对全省各地环境质量的考核分量。其中环境保护指标成为四大类考核体系之一。在全省各地建设全面小康社会中, 重点考核和评价环境质量综合指数, 其中包括全年空气质量良好天数达标率、水域功能区水质达标率、集中式饮用水源地水质达标率和城市环境噪声达标区覆盖率等项考核指标。

近年来, 江苏省苏南地区 6个市, 紧密围绕让群众喝上干净的水、呼吸上洁净的空气、吃上放心的食品的战略目标, 认真核算本地区的现状值和差距, 制定改善环境质量的实施计划, 扎扎实实推进小康社会环保考核目标的实现。继 6个市相继建成国家环保模范城市和国家级生态示范区之后, 2004年以来又全面启动了环境保护“二次创业”工程, 矢志奋进去争创全国首批生态市。通过大力发展循环经济和清洁生产, 致力构建环境友好型和资源节约型社会, 全面提高可持续发展能力。随着各类“生态细胞”工程相继建成, 不断加快创建全国生态市的进程。其中, 昆山、常熟、张家港、江阴 4个市创建生态市已通过省级考评, 即将接受国家正式考核验收。

(记者 高杰)

摘自 www. jshb. gov. cn 2006年 3月 29日