文章编号:1671 - 6833(2008)01 - 0106 - 04

基于人工神经网络与回归分析的水质预测

李亦芳,程万里,刘建厅

(华北水利水电学院 数学与信息科学学院,河南 郑州 450011)

摘 要:针对人工神经网络在预测中出现的异常值现象,采用了回归分析模型得到的预测区间来控制异常值现象的方法.并且应用在黄河三门峡河段的水质预测中,氨氮通量预测的网络模型控制前平均精度仅有50.05%,这是因为2006年6月份预测值偏离真实值太大,预测相对误差达到214.88%,超出了回归预测区间,从而影响了整体精度.控制后该月的相对精度为90.08%,平均精度达到80.79%,整体预测精度明显提高.实践表明,该方法对于消除网络模型预测中出现的异常值现象是较为有效的.

关键词:回归分析;人工神经网络;水质预测

中图分类号:O 212.5;O 29 **文献标识码**:A

0 引言

水质变化趋势预测是维护和管理当前水质状况的重要依据,通过预测可以了解当地水域环境质量演变趋势,从而及时发现水质恶化的原因并制定相应的治理措施[1].影响水质的因素有物理、化学、水力学、生物学、气象学以及人类活动等多方面的因素,在时间和空间上存在相当多的影响变量,是一个涉及多种影响因子的复杂体系[2-3].现有的基于数学表达式水质预测模型很难将这些因素都考虑进去,使得中长期时间尺度的水质预测结果仍然不能令人满意,这给水污染控制及水资源长远规划、管理的科学决策带来很大困难.

1 水质预报的联合模型

1.1 水质预报模型及置信区间

人工神经网络(Artificial Neural Networks, ANN)是模拟人脑思维与记忆的神经网络化数学模型及算法系统,主要特点是具有自学习的功能,即通过对样本信息的学习来获得信息之间复杂的关系,条件是需样本足够多、样本要具有代表性.在实际应用中往往存在样本数量有限、富含噪音等问题,训练后的网络稳定性得不到保障,预测中常常出现大部分预测精度较高,个别值偏离真实较大的现象.

回归分析是在排除其他影响因素或假定其他影响因素确定的条件下,分析某些因素(表现为自变量)是如何影响另一事物(表现为因变量)的过程.利用回归分析研究黄河断面间的水质关系是在假定断面间污染物降解、旁侧入流基本稳定的状况下,上下游断面的水质存在基本稳定的输入、输出的响应关系,所建立的回归方程是断面间水质关系的一种统计表述.在实例分析中可以看出,回归分析模型虽然整体预测精度不高,但是预测较为平稳(基本不出现异常值)且带有一定置信水平的预测区间.

设 y 与 $x_1,x_2,\cdots,x_p$ 有如下线性关系[4-5]:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \varepsilon \sim N(0,\sigma^2) \quad (1)$$

由已知样本 $y_i,x_{i1},x_{i2},\cdots,x_{ip}, i = 1,2,\cdots,n$. 求得的回归方程为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \quad (2)$$

并且回归方程与回归系数均通过检验.

又设 $x_{01},x_{02},\cdots,x_{0p}$ 为 $x_1,x_2,\cdots,x_p$ 所取的一组固定值,对应的观测值为

$$y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \varepsilon_0, \beta_0 \sim N(0,\sigma^2) \quad (3)$$

回归值为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \quad (4)$$

则 y_0 的置信水平为 $(1-\alpha)$ 的置信区间为

$$[\hat{y}_0 - \delta(x_0), x\hat{y}_0 + \delta(x_0)] \quad (5)$$

其中,

$$\delta(x_0) = \hat{\sigma}_\varepsilon t_{1-\frac{\alpha}{2}}(n-p-1) \cdot$$

收稿日期:2007 - 09 - 04;修订日期:2007 - 10 - 31

作者简介:李亦芳(1968 -)女,河南郑州人,华北水利水电学院副教授,主要从事随机分析及其应用,E-mail:liyifang@ncwu.cn.

$$\sqrt{\left[1+\frac{1}{n}+\sum_{i=1}^{p}\sum_{j=1}^{p}c_{ij}(x_{0i}-\bar{x}_i)(x_{0j}-\bar{x}_j)\right]},$$

$$\hat{\sigma}_e=\sqrt{Q_e/(n-p-1)};$$

式中：Q_e 为残差平方和；$\bar{x}_i$ 为第 i 个影响因子的样本均值；c_{ij} 为第二类正规矩阵逆矩阵的第 i 行第 j 列的元素.

1.2　联合模型的原理及其实现步骤

鉴于神经网络模型预测整体精度较高，预测往往会出现异常值现象. 回归模型精度不高，预测较为平稳且带有一定置信水平的预测区间. 我们设想利用回归分析模型的置信区间来控制神经网络的异常值现象，其步骤为：

第一步：建立水质预测的回归分析模型并计算预测值.

第二步：建立相应的人工神经网络(ANN) 模型并计算预测值.

第三步：利用公式(5) 计算置信区间.

第四步：判断 ANN 预测值是否落入置信区间，若落入置信区间则认为此次预测是可信的，否则，则认为 ANN 预测值出现了异常，用回归值代替 ANN 预测值.

2　实例分析

黄河三门峡河段地处黄河中游[3]，上自龙门水文站下至三门峡大坝，河段全长 240.4 km，是黄河流域社会、经济最发达的地区之一，是三门峡等重要城市生活及工农业生产的主要供水水源. 同时小浪底以下河段还承担着引黄济津、引黄济青等外流域调水的重要任务，正在兴建的小浪底反调节水库西霞院水利工程将担负起 2008 年向北京提供应急供水的重要使命. 鉴于该河段的重要性，若能对其水质状况作出适时准确的预测，对水污染控制及水资源规划、管理的科学决策均具有一定的参考价值，笔者以该河段水质预测为例来探讨水质预测的 ANN - 回归分析联合模型，从而为该模型进一步应用探索路子.

2.1　预测因子的选取

由文献[3] 研究结论以及黄河水质近年的监测状况可知，COD、氨氮仍然是黄河水质污染的最重要指标，基本上能够反应黄河水污染状况. 因此本次预测因子选为 COD、氨氮.

2.2　建模样本与检验样本的选取

虽然黄河水质监测数据有一定的积累，但各个断面监测起止时间、监测频次不一，为了样本的统一性起见，我们选取该河段 2003 年 1 月 ~ 2006 年 7 月的常规水质月监测数据. 其中，2003 年 1 月 ~ 2005 年 12 月的 36 组水质数据作为建模数据样本，2006 年 1 月 ~ 2006 年 7 月的 7 组水质数据作为检测样本.

2.3　研究河段的划分

黄河三门峡河段干流上共设有龙门、潼关、三门峡三个监测断面，其中，龙门 ~ 潼关段有黄河最大的支流渭河在此汇入，此外，还有污染较为严重的汾河、涑水河. 在三个支流上分别布设有华县、吊桥、河津、蒲州四个监测断面. 潼关 ~ 三门峡段其间虽有支流汇入，但缺乏相关数据资料，故本次研究将三门峡河段划分为：龙门 ~ 潼关、潼关 ~ 三门峡两段进行研究，其中，龙门 ~ 潼关段主要研究潼关与龙门、华县、河津、蒲州断面的水质关系，潼关 ~ 三门峡段主要研究三门峡与潼关断面的水质关系.

2.4　回归模型的建立

2.4.1　龙门 ~ 潼关段

由于潼关断面水质受龙门、华县、河津、蒲州断面的影响，为了简化问题，我们讨论预测潼关断面通量(水质指标浓度 × 流量)，这样只要结合相关水文资料简单计算便可得到相关水质资料. 采用逐步回归分析方法[5] 建立回归模型如表 1.

表 1　潼关断面回归模型相关参数表

Tab.1　The parameters of regression model in Tongguan section

参数名	β_0	β_1	β_2	F 统计量	R 相关系数	ρ 伴随概率
氨氮通量(因变量)	-0.15	1.46	1.25	31.57	0.81	0.000

表 1 中，β_0 为常数，β_1、β_2 分别表示华县、龙门氨氮通量相关系数，河津、蒲州断面氨氮通量影响不够显著，不以考虑. 预测结果见表 2.

2.4.2　潼关 ~ 三门峡段

由于黄河干流潼关 ~ 三门峡河段中间及其支流未设监测断面，故该模型只能模拟假定断面间污染物降解、旁侧入流基本稳定的状况下，上、下游断面的水质的输入、输出的响应关系. 同时由于影响三门峡断面(下游) 化学需氧量(COD) 浓度值不仅受潼关断面(上游) 化学需氧量(COD) 浓度值的影响，还受潼关断面背景来水量、含沙量、水温、流速等多种因素影响. 所以，我们采用逐步回归分析方法从中筛选出影响较为显著的因子建立模型，结果如表 3 所示.

表 3 中，β_0 为常数，β_1、β_2、β_3 分别为潼关断面影响因子水温、流量、COD 浓度值相关系数. 预测

结果见表4,由于该断面氨氮预测值基本正常,故本文不再讨论.

表2 潼关断面氨氮通量联合模型预测结果(置信水平95%)

Tab.2 The predictive results of ammonia flux based on union model in Tongguan section:(95% confidence level)

kg/s

2006年月份	1	2	3	4	5	6	7	平均精度/%
氨氮通量实测	2.22	1.78	2.22	1.10	2.42	2.05	0.48	—
回归预测	1.32	2.76	3.32	1.57	2.18	2.44	0.33	64.56
ANN预测(控制前)	2.00	1.74	1.71	1.19	7.60	1.21	0.76	50.05
置信上限	2.88	4.33	4.87	3.14	3.74	4.00	1.89	—
置信下限	-0.25	1.19	1.71	0.01	0.62	0.87	-1.23	—
ANN预测(控制后)	2.00	1.74	1.71	1.19	2.18	1.21	0.76	80.79

表3 三门峡断面回归模型相关参数表

Tab.3 The parameters of regression model in Sanmenxia section

参数名	β_0	β_1	β_2	β_3	F统计量	R相关系数	ρ伴随概率
COD(因变量)	17.55	-0.05	-0.005	0.32	8.72	0.67	0.000

2.5 神经网络模型

与回归分析对应,建立神经网络模型龙门-潼关段:输入因子为氨氮通量(龙门断面、华县断面),输出因子氨氮通量(潼关断面).潼关-三门峡段:潼关断面输入因子(COD浓度、流量、水温),三门峡断面输出因子(COD浓度).利用MATLAB编程采用L-M算法训练网络,预测结果如表2、表4所示.

表4 三门峡断面(COD)联合模型预测结果表(置信水平95%)

Tab.4 The predictive results of COD based on union model in Smenxia section (95% cinfidence level) mg/L

2006年月份	1	2	3	4	5	6	7	平均精度/%
COD实测	27.20	13.20	21.20	12.00	12.60	20.90	13.70	—
回归预测	25.83	19.81	18.82	18.91	22.29	21.10	25.18	59.20
ANN预测(控制前)	53.40	23.50	20.80	12.40	17.40	22.00	15.70	66.07
置信上限	32.10	26.08	25.09	25.16	28.56	27.37	31.44	—
置信下限	19.56	13.54	12.55	12.34	16.02	14.83	14.91	—
ANN预测(控制后)	25.83	23.50	20.80	12.40	17.40	22.00	15.70	80.91

2.6 预测区间控制ANN异常值

利用公式(4)~(5)计算相应预测区间结果如表2、表4所示.由表2可以看出潼关断面氨氮通量ANN控制前平均预测精度仅有50.05%,这是因为2006年5月份预测值偏离真实值太大,预测相对误差达到214.88%,超出了回归预测区间,从而影响了整体精度,控制后该月的精度为90.08%,平均精度达到80.79%,整体预测精度明显提高,异常现象基本得到控制;由表4可知三门峡断面化学需氧量(COD)ANN模型控制前平均预测精度为66.07%,这是因为2006年1月份预测值偏离真实值较大,相对误差为96.32%,预测值没有落入相应的置信区间.控制后该月份的预测相对精度为94.96%,平均精度达到80.91%,整体预测精度明显提高,异常也现象得到有效控制.

3 结束语

实践表明利用回归模型的预测区间来控制神经网络的异常值,从某种程度上会消除ANN的异常现象,利用ANN模型也能够在一定程度上弥补回归模型精度不高的缺陷,两者联合使用为该河段的水质预测提供了较为平稳又有一定精度的预测方法.利用预测区间控制ANN异常现象不仅仅限于水质预测,上述原理对于其它预测领域也是适用的.

参考文献:

[1] 邱 林,黄 鑫,李洪良.基于模糊权马尔可夫模型的综合水质预测[J].人民长江,2007,38(1):75-77.

[2] 孙才志,林学钰.降水预测的马尔可夫模型及应用[J].系统工程学报,2003,18(4):294-300.

[3] 郝伏勤,李 群,黄锦辉. 黄河水资源保护科学研究所文献[DB]. 郑州:黄河流域水资源保护局,2005.
[4] 吴 诩,李永乐,胡庆军. 应用数理统计[M]. 长沙:国防科技大学出版社,1995:161 - 186.
[5] 《现代应用数学手册》编委会. 现代应用数学手册:概率统计与随机过程[M]. 北京:清华大学出版社,1999.

The Forecast of Water Quality Based on Artificial Neural Networks and Regression Analysis

LI Yi - fang, CHENG Wan - li, LIU Jian - ting

(College of Mathematics and Information Science, North China Institute of Water Conservancy and Hydroelectric Power, Zhengzhou 450011, China)

Abstract: As to the abnormal phenomenon in the forecast of artificial neural networks, the method, in which the forecast range from the regression analysis model is used to control the abnormal phenomenon, has been adopted. In the forecast of the water quality of Yellow River in Sanmenxia, the average accuracy of the quantity of ammonia and nitrogen before the control of ANN is only 50.05%, which is because the forecast number is very different of the accurate number in June 2006, the relative error of the forecast number reach up to 214.88 percent, beyond the forecast range of regression, in order to have effect on the whole accuracy. The accuracy of this month is 90.08%, the average accuracy reaches up to 80.79%; the whole forecast accuracy is proved obviously. The practice shows that the method is effective to eliminate the abnormal phenomenon in the artificial neural networks.

Key words: regression analysis; artificial neural networks; water quality forecast

(上接第 101 页)

参考文献:

[1] 周振红,郭恒亮,张君静,等. Fortran 90/95 高级程序设计[M]. 郑州:黄河水利出版社,2005.
[2] 王丽娟,孙西超,底松茂,等. 软件复用与基于面向对象框架的软件开发方法[J]. 郑州大学学报:工学版,2003,24(3):24 - 28.
[3] 任 慧,周振红. Fortran 与 C/C ++ 共享公用外部数据[J]. 郑州大学学报:工学版,2007,28(4)63 - 65.
[4] 任 慧,周振红,张成才. Fortran 与 C/C ++ 的混合编译[J]. 计算机工程与设计,2007,28(17):4096 - 4098.
[5] 周振红,徐进军,毕苏萍,等. Intel Visual Fortran 应用程序开发[M]. 郑州:黄河水利出版社,2006.

Fortran and C/C ++ Sharing Data and Routines in Modules

BI Su - ping[1], ZHOU Zhen - hong[2]

(1. School of Civil Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. School of Environment and Water Conservancy, Zhengzhou University, Zhengzhou 450001, China)

Abstract: In mixed - language programming, routines programmed with one language are called with the other, and the data passed through calling argument lists, so no object - oriented programming is embodied in this process. In this paper, based on the module recommended by Fortran 90, a new way of object - oriented mixed - language programming with the two languages is presented. It has been proved from the experiment that not only C/C ++ is able to directly access the data and routines in the module, but also C/C ++ data and routines to be encapsulated into the module in order to be accessed by the Fortran unit using the module.

Key words: numerical computation; mixed - language programming; calling convention; object - oriented; module