

偏最小二乘与人工神经网络耦合模型 在酸雨 pH 值预测中的应用

周秀平¹, 王文圣¹, 曾怀金²

(1. 四川大学水利水电学院, 四川 成都 610065; 2. 四川省水利水电勘测设计研究院, 四川 成都 610074)

摘要:将偏最小二乘与人工神经网络耦合, 建立了一种新型耦合预测模型——偏最小二乘与人工神经网络耦合模型。该模型利用了偏最小二乘方法有效处理自变量之间多重相关性问题和人工神经网络可以较好地解决非线性问题的能力, 在城市酸雨 pH 值预测中的应用表明, 该模型预测精度高, 明显优于偏最小二乘回归模型和人工神经网络模型。

关键词:酸雨; pH 值; 偏最小二乘回归; 神经网络; 预测模型

中图分类号: P333.9 **文献标识码:** A **文章编号:** 1006-7647(2006)04-0050-03

Partial least square method and BP neural network coupled model and its application to prediction of pH value of acid rains//ZHOU Xiu-ping¹, WANG Wen-sheng¹, ZENG Huai-jin²(1. College of Hydraulic and Hydro-electric Engineering, Sichuan University, Chengdu 610065, China; 2. Sichuan Survey and Design Institute of Water Conservancy and Hydropower, Chengdu 610074, China)

Abstract: By taking advantages of the partial least square method for effective treatment of the multi-correlations between independent variables and BP neural network for effectively solving the nonlinear problems, a partial least square method and BP neural network coupled prediction model was developed. The application of the model to prediction of the pH value of urban acid rains shows that the precision of the model is high, and that it is superior to the partial least square regression model and the artificial neural network model.

Key words: acid rain; pH value; partial least square regression; neural network; prediction model

随着大气污染的日益严重, 世界各地均不同程度地出现了酸雨现象。酸雨污染是当今一个重要的环境问题, 准确可靠的酸雨 pH 值预测可以有效地控制和减少酸雨的危害程度。线性回归模型在酸雨 pH 值预测中得到了大量的应用, 但存在着两个问题: 一是自变量之间存在严重的多重相关性, 最小二乘方法估计的模型参数不稳定且缺乏明确的物理意义^[1-2]; 二是酸雨污染系统是非线性的, 线性关系的简单处理会带来极大的预测误差。第一个问题可以用偏最小二乘(partial least square, 简称 PLS)方法解决, 第二个问题可用具有普适性的人工神经网络(artificial neural network, 简称 ANN)来解决。为此本文提出了基于偏最小二乘与人工神经网络耦合的预测模型(简称 PLS-ANN 耦合模型)。其思路是: 用 PLS 方法提取关于自变量和因变量的主成分 t_i ($i = 1, 2, \dots, h$), 将主成分 t_i 作为 ANN 的输入, 因变量作为 ANN 的输出。将 PLS-ANN 耦合模型应用于城市

酸雨 pH 值预测, 结果表明, 该模型预测精度高, 较偏最小二乘回归(简称 PLSR) 模型和 ANN 模型优。

1 PLS 原理

PLS 是估计回归方程参数的一种比较新的方法, 集多元线性回归分析、典型相关分析和主成分分析的基本功能于一体。它利用降维的思想将系统中的多个变量指标转化为少数对系统具有最佳解释能力的新的综合变量(又称成分), 由于对变量的综合, 就可以克服多重相关性造成的信息重叠。PLS 方法不但考虑了对自变量做成分提取, 而且通过连续扩大潜在变量和因变量之间的典型相关来确定因子^[3], 这样就同时保持了不同潜在变量的相互正交。也就是说用 PLS 方法提取的成分, 对自变量和因变量都具有最佳的解释能力^[3-4]。

设有单因变量 Y 和自变量集合 $X = \{x_1, x_2, \dots, x_m\}$, 为了研究 Y 与 X 间的统计关系, 首先在 X 与 Y

基金项目: 四川省科技厅软科学基金资助项目(042R025-051)
作者简介: 周秀平(1982—), 女, 山东聊城人, 硕士研究生, 从事水文水资源与水环境系统研究。E-mail: zhouxiuping_1982@163.com

中提出成分 t_1 和 u_1 , PLS 方法在提取这两个主成分时要满足以下两个要求^[1]: ① t_1 和 u_1 尽可能多的携带它们各自数据表中的变异信息; ② t_1 和 u_1 的相关程度能够达到最大。

在第一成分 t_1 和 u_1 被提取后, PLSR 分析分别实施 X 对 t_1 的回归及 Y 对 t_1 的回归。如果回归方程已达到满意的精度, 则终止运算; 否则再利用 X 被 t_1 解释后的残余信息及 Y 被 t_1 解释后的残余信息进行第二轮的成分提取。如此反复, 一直达到满意的精度为止。若最终对 X 提取了 h 个成分 $t_1, t_2, \dots, t_h$, 再通过实施 Y 对 $t_1, t_2, \dots, t_h$ 的回归, 然后再表达成 Y 关于原变量 $x_1, x_2, \dots, x_m$ 的回归方程, 至此 PLSR 建模完成。

在 PLSR 建模中究竟应该选几个成分建立模型最合适呢? 一般采用交叉有效性作为判别标准。设 y_i 为原始数据, $t_1, t_2, \dots, t_h$ 是在 PLSR 过程中提取的主成分。 $\hat{y}_{h(i)}$ 是使用全部样本点并取 $t_1, t_2, \dots, t_h$ 回归建模后, 第 i 个样本的拟合值。 $\hat{y}_{h(-i)}$ 是在建模时删去第 i 个样本点, 取 $t_1, t_2, \dots, t_h$ 回归建模后, 再用此模型计算 y_i 的拟合值。记 $Q_h^2 = 1 - \frac{p_h}{s_{h-1}}$, 其中 $s_h = \sum_{i=1}^n (y_i - \hat{y}_{h(i)})^2$, $p_h = \sum_{i=1}^n (y_i - \hat{y}_{h(-i)})^2$ 。当 $Q_h^2 \geq (1 - 0.95^2) = 0.0975$, 说明插入的新成分对模型的预测能力有明显的改善作用。

2 PLS-ANN 耦合模型

ANN 模型由许多并行运算功能的单元组成, 这些单元类似于生物神经元。其特色在于信息的分布式存储和并行协同处理, 具有很强的自学习性、自组织性和非线性逼近功能^[5]。一般 3 层前馈网络就足够描述复杂的研究对象。

本文将 PLS 和 ANN 进行耦合, 基本思路是: ①用 PLS 方法提取关于自变量和因变量的主成分 t_i ($i = 1, 2, \dots, h$); ②将主成分 t_i 作为 ANN 的输入, 因变量作为 ANN 的输出, 构造神经网络; 设有 n 个样本, 第 j 个 ($j = 1, 2, \dots, n$) 样本输入为 t_{ji} ($i = 1, 2, \dots, h$), 对应的输出为 y_j ; ③网络学习、模型的拟合和检验, 进而进行预测。PLS-ANN 耦合模型结构见图 1。

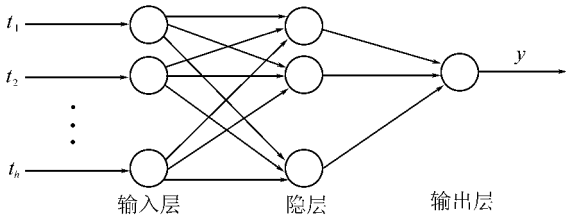


图 1 PLS-ANN 耦合模型结构

神经网络的激励函数采用 Sigmoid 函数, 网络权重采用改进的 BP 算法进行学习。需注意, PLS-ANN 耦合模型是针对多因素影响的非线性系统而言的, 而且因变量与自变量间存在着多重相关性。

3 PLS-ANN 耦合模型在酸雨 pH 值预测中的应用

酸雨是降水中各种离子综合作用的结果, 监测实例表明: 城市降水 pH 值主要受酸性离子 SO_4^{2-} , NO_3^- 和碱性离子 Ca^{2+} , NH_4^+ 影响^[2,6-7]。本文收集了我国部分城市降水中 SO_4^{2-} , NO_3^- , Ca^{2+} , NH_4^+ 离子的质量浓度和 pH 值的资料(基本数据取自文献[2, 6-7]), 见表 1。研究还表明, 酸碱离子的质量浓度对 pH 值有着重要的影响^[6]。分别以 Ca^{2+} , NH_4^+ , SO_4^{2-} , NO_3^- 的质量浓度与组合因子 ($\text{Ca}^{2+} + \text{NH}_4^+$) 和 ($\text{SO}_4^{2-} + \text{NO}_3^-$) 质量浓度的比值为自变量, 记为 x_1, x_2, x_3, x_4, x_5 , pH 值作为因变量记为 y 。用前 12 组数据建模, 后 5 组数据进行检验。

表 1 17 个城市降水中 pH 值及其影响因素

城 市	离子质量浓度/($\mu\text{mol} \cdot \text{L}^{-1}$)				x_5	y
	x_1	x_2	x_3	x_4		
北 京	151.6	163	155	39.5	1.62	6.29
长 春	256.5	61.3	157	21.2	1.79	6.71
锦 州	340.8	124	259	49.4	1.51	6.32
烟 台	289.1	39.1	183	22.8	1.6	6.95
平顶山	107.7	138	152	0.4	1.61	6.29
合 肥	110.3	117	142	31.8	1.31	4.73
苏 州	125.3	93.6	200	14.4	1.02	4.63
上 海	104.3	75.8	153	12.6	1.08	4.85
杭 州	59.9	68.2	112	13.5	1.02	4.84
南 宁	26.6	27.7	62	4.9	0.82	4.82
南宁①	131.8	84.9	197	14.4	1.03	4.76
南宁②	150.4	131	244	17	1.08	4.80
桂 林	67.2	50	107	19.7	0.92	4.83
重 庆	127.8	151	327	27.9	0.79	4.21
贵 阳	199.6	174	405	27.9	0.86	4.23
马鞍山	123	73.7	139	15.1	1.27	5.33
广 州	175.1	141	255	33.3	1.1	4.39

注: ①1988 年数据; ②1989 年数据; 其余均为 1986 年数据。下同。

3.1 多重相关性判别

分析因变量与各自变量的相关系数 r , 计算结果见表 2。从表 2 中可以看出自变量之间存在着多重相关性, 如 x_1 与 x_5 的相关系数达到 0.690, x_1 与 x_3 的相关系数达到 0.672。

表 2 自变量、因变量之间的相关系数 r

变量	x_1	x_2	x_3	x_4	x_5	y
x_1	1.000	0.080	0.672	0.658	0.690	0.736
x_2		1.000	0.459	0.416	0.322	0.076
x_3			1.000	0.473	0.235	0.175
x_4				1.000	0.451	0.360
x_5					1.000	0.913
y						1.000

表 3 pH 值预测结果

阶段	城 市	y	LSR 模型		PPR 模型		PLSR 模型		ANN 模型		PLS-ANN 模型	
			预测值	相对误差/%	预测值	相对误差/%	预测值	相对误差/%	预测值	相对误差/%	预测值	相对误差/%
拟合	北 京	6.29	5.94	5.62	6.27	-0.32	5.93	-5.71	6.25	-0.58	6.08	-3.34
	长 春	6.71	6.92	-3.18	6.70	-0.15	7.07	5.44	6.80	1.32	6.90	2.78
	锦 州	6.32	6.46	-2.38	6.31	-0.16	6.25	-1.17	6.32	-0.05	6.41	1.36
	烟 台	6.95	6.66	4.15	6.96	0.14	6.81	-2.03	6.87	-1.16	6.83	-1.67
	平 顶 山	6.29	6.37	-1.21	6.29	0.00	5.89	-6.43	6.29	0.04	6.18	-1.76
	合 肥	4.73	5.07	-7.21	4.78	1.06	5.35	13.02	4.79	1.27	5.13	8.43
	苏 州	4.63	4.64	-0.28	4.67	0.86	4.73	2.25	4.74	2.28	4.67	0.79
	上 海	4.85	4.92	-1.39	4.88	0.62	4.99	2.87	4.83	-0.46	4.86	0.27
	杭 州	4.84	4.73	-2.23	4.77	-1.45	4.85	0.16	4.70	-2.95	4.79	-1.13
	南 宁	4.82	4.83	-0.28	4.85	0.62	4.63	-4.04	4.87	0.97	4.82	0.07
检验	南宁①	4.76	4.68	1.63	4.69	-1.47	4.82	1.29	4.79	0.68	4.73	-0.72
	南宁②	4.80	4.75	0.96	4.82	0.42	4.68	-2.50	4.75	-1.12	4.59	-4.35
	桂 林	4.83	4.45	7.96	4.23	-12.42	4.74	-1.77	4.62	-4.30	4.72	-2.27
	重 庆	4.21	2.71	35.72	4.55	8.08	3.64	-13.44	4.15	-1.53	4.27	1.32
	贵 阳	4.23	2.90	31.51	4.56	7.80	3.74	-11.63	4.25	0.36	4.24	0.30
	马 鞍 山	5.33	5.35	-0.39	5.16	-3.19	5.50	3.22	5.17	-2.96	5.47	2.65
	广 州	4.39	4.50	-2.42	4.03	-8.20	4.72	7.62	4.59	4.00	4.56	3.78

3.2 用 PLS 提取主成分

用交叉有效判别法提取主成分,当主成分提取到 t_3 时 $Q_h^2 < 0.0975$,说明提取 2 个主成分即可。将提取的 2 个主成分表达如下:

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 0.5925 & 0.0613 & 0.1409 & 0.2903 & 0.7356 \\ 0.1837 & -0.4187 & -0.5752 & -0.3884 & 0.6782 \end{bmatrix} E_0^T$$

其中, E_0 为自变量原始数据标准化处理后的矩阵,即 $E_0 = (e_1, e_2, e_3, e_4, e_5)$ 。用 PLS 方法提取主成分集典型相关分析和主成分分析的功能于一体,提取出来的主成分对自变量和因变量都具有最佳的解释能力。

3.3 PLS-ANN 耦合模型的拟合和检验

将提取的 2 个主成分 $T = t_i (i = 1, 2)$ 作为 ANN 的输入,建立 2 个输入节点,3 个隐节点,1 个输出(酸雨 pH 值)节点的 ANN 模型。赋予初始权值为随机小量,学习步长取 0.35,将输出数据限制在 0.2 ~ 0.8 之间,附加冲量项取 0.85。将训练样本输入网络反复进行训练,当训练量达到 5000 次时,误差函数趋向稳定,训练结束。

为了评价 PLS-ANN 耦合模型的性能,同时利用了最小二乘回归(least square regression,简称 LSR)模型^[2]、投影寻踪回归(projection pursuit regression,简称 PPR)模型^[6]、PLSR 模型和 ANN 模型进行了对比研究,各模型拟合和检验结果及相对误差见表 3。

从表 3 中可以看出,在拟合阶段 PLS-ANN 耦合模型的精度除了没有 PPR 模型和 ANN 模型的精度高外,比另外两个模型的精度都要高,在检验阶段 PLS-ANN 耦合模型的精度最高,说明 PLS-ANN 耦合模型能较好地反映酸雨 pH 值与其各影响因素之间

的变化关系,也进一步体现了该耦合模型在处理自变量之间多重相关性和非线性问题方面的优势。

4 结 论

a. 用 PLS 方法对自变量进行处理,提取对因变量解释最强的成分,不仅可以消除自变量系统的线性相关性,而且可以降低系统维数,减少 ANN 的输入层节点数,从而简化了网络结构,增强了网络的稳定性。

b. ANN 模型能够较好地处理一般回归模型不能解决的非线性问题,但 ANN 的性能会因输入的严重自相关性使得网络的求解变得不稳定而且收敛得比较慢。PLS 和 ANN 的结合,比较好地处理了一般情况下难以同时解决好的自变量相关和非线性问题。

c. PLS-ANN 耦合模型应用于城市酸雨 pH 值预测精度高,较 PLSR 模型和 ANN 模型优。

参考文献:

[1] 王惠文. 偏最小二乘回归方法及其应用[M]. 北京: 国防工业出版社, 1999.

[2] 王文圣, 熊华康, 曹宏伟. 酸雨 pH 值预测的偏最小二乘回归模型[J]. 四川环境, 2003, 22(6): 77-80.

[3] 李洪, 代进, 蒋金泉. 偏最小二乘回归的神经网络模型在巷道围岩位移预测中的应用[J]. 煤炭学报, 2004, 29(3): 274-278.

[4] 陈南祥, 黄强, 曹连海. 基于偏最小二乘回归与神经网络耦合的岩溶泉预报模型[J]. 水利学报, 2004(9): 68-72.

[5] 张立明. 人工神经网络模型及其应用[M]. 上海: 复旦大学出版社, 1993.

[6] 李祚泳. 我国部分城市降水中离子浓度与 pH 值的关系研究[J]. 环境科学学报, 1999, 19(3): 303-306.

[7] 李祚泳. 降水中的离子浓度对降水酸度 pH 值的影响[J]. 环境工程, 1998, 16(5): 49-51

(收稿日期: 2005-10-17 编辑: 熊水斌)