

基于偏最小二乘回归与支持向量机耦合的 咸潮预报模型^{*}

刘德地, 陈晓宏

(中山大学水资源与环境研究中心, 广东 广州 510275)

摘 要: 利用偏最小二乘回归对影响咸潮的因素进行分析, 提取出对因变量影响强的成分, 从而克服了变量之间的多重相关性问题; 同时利用支持向量机在解决小样本非线性问题上的优势, 采用将偏最小二乘回归与支持向量机耦合的方法, 建立了咸潮预报模型 (PLS-SVM), 并应用该模型对珠海市平岗站盐度的变化进行了模拟和预测。研究结果表明, 所提出的 PLS-SVM 模型模拟和预测精度明显优于常用的 BP 人工神经网络、多元回归模型, 可更好地应用于咸潮预报。

关键词: 偏最小二乘回归 (PLS); 支持向量机 (SVM); 咸潮预报; 珠海市

中图分类号: TV213.1 **文献标识码:** A **文章编号:** 0529-6579 (2007) 04-0089-04

咸潮发生的物理机制十分复杂^[1-2], 目前难以用数学模型准确地描述, 而统计模型可根据现有的实测资料, 建立影响咸潮发生因素和咸潮发生规律之间关系, 从而达到对咸潮预报的目的。而在一般统计模型中, 常采用回归分析方法, 对事先拟定的因子进行筛选和系数求解, 但当拟定的因子样本数较少且因子之间存在严重的相关性时, 会导致分析失效, 而偏最小二乘回归 (Partial least squares regression, PLS) 可充分利用已有样本信息和有效地解决变量之间的多重相关性问题^[3-4]。偏最小二乘回归^[5]是于 1983 年由伍德 (S. Wold) 和阿巴诺 (C. Abano) 等人首次提出的, 在生物数据分类、化学分光镜预测化学样本成分中获得了成功的应用^[6-7]。它是一种新型的多元统计数据分析方法, 集成了多元线性回归分析、变量的主成分分析和变量间的典型相关分析三者的基本功能于一体, 将建模预测类型的数据分析方法与非模型式的数据认识性分析方法有机结合起来, 利用对数据信息进行分解和筛选的方式, 有效地提取对系统解释性最强的综合变量, 剔除多重相关信息和无解释意义信息的干扰, 从而克服了变量多重相关性在系统建模中的不良作用^[4-5], 另外, 通过主成分分析和综合变量的提取, 允许样本个数少于自变量个数的条件下进行回归建模^[3], 一定程度上克服了由于样本数量少而不能进行回归分析的问题。

咸潮 (本文以盐度表示) 的发生主要受上游来水量、潮汐等作用的影响^[8], 当因变量和自变量的关系比较复杂且样本数据较少时, 采用线性回

归模型难以反映客观规律, 并影响预报精度和效果。支持向量机 (support vector machine, SVM) 具备自组织、自学习和联想记忆功能, 不仅克服模型必须是基本观测数据的线性和非线性组合的局限, 而且由于遵循统计学习理论中的结构风险最小化原理, 也克服了一般机器学习中 (如人工神经网络) 出现的“过学习”、局部优化、样本数量要求多等问题^[9], 已经在非线性模型预测控制、模式识别等方面得到了广泛的应用^[10-12]。因此, 本文提出通过偏最小二乘回归与支持向量机方法的耦合建立咸潮预报模型 (PLS-SVM), 提高在样本数量较少条件下模型预报的准确度, 并应用于珠海市平岗站的咸潮预报。

1 PLS-SVM 模型

1.1 偏最小二乘回归的原理^[3, 5]

设有 q 个因变量 $\{Y_1, \dots, Y_q\}$ 和 P 个自变量 $\{X_1, \dots, X_P\}$ 。为了研究因变量与自变量的统计关系, 采集 n 个样本点, 构成自变量与因变量的数据表 $X = [X_1, \dots, X_P]_{n \times P}$ 和 $Y = [Y_1, \dots, Y_q]_{n \times q}$ 。偏最小二乘分别在 X 和 Y 中提取出成分 t 和 u 。即 t 是 $X_1, \dots, X_P$ 的线性组合, u 是 $Y_1, \dots, Y_q$ 的线性组合, 在提取这两种成分时, 为了回归分析的需要, t 和 u 应尽可能大地携带它们各自数据表中的变异信息; 且 t 与 u 的相关程度能够达到最大。

* 收稿日期: 2006-09-01

基金项目: 国家自然科学基金资助项目 (50579078); 广东省自然科学基金资助项目 (04009805)

作者简介: 刘德地 (1980 年生) 男, 博士生; 通讯联系人: 陈晓宏, E-mail: eescxh@mail.sysu.edu.cn

提取第一主成分 t_1 和 q_1 后, 分别实施 X 和 Y 对 t_1 的回归, 如果此时回归方程已经达到满意的精度, 则算法停止; 否则, 将利用 X 被 t_1 解释后的残余信息以及 Y 被 t_1 解释后的残余信息进行第二主成分 t_2 的提取, 继续实施 Y 和 X 对 t_1, t_2 的回归。如此反复, 直到能达到一个较满意的精度 (用交叉有效性确定) 为止。

1.2 PLS-SVM的原理

用偏最小二乘回归从 X 和 Y 中提取出的主成分 $t_1, \dots, t_m$ 和 $q_1, \dots, q_m$ (本文采用 PLS 模型, 故 q 和原因变量序列等同即 Y) 分别作为支持向量机的样本输入和输出。而模拟预测问题就是找到输入和输出的函数关系, 即找出一个函数 f 使之通过样本训练后, 对于样本以外的 t 通过 f 找出对应的 y 。支持向量机利用非线性映射 $\varphi(\cdot)$ 将训练数据集非线性地映射到一个高维特征空间 (Hilbert 空间), 使得在输入空间中的非线性函数估计问题转化为高维特征空间中的线性函数估计问题, 并根据统计学习理论中的结构风险最小化原理, 求出拟合样本集的估计函数 $\hat{f}(t)$ 的解析表达式。

2 模型求解步骤

(1) 将 X 与 Y (只取 Y_1) 进行标准化处理, 得到标准化后的自变量矩阵 E_0 和因变量矩阵 F_0 。

(2) 从 E_0 中抽取一个成分, $t = E_0 w$, 其中 $w = \frac{E_0 F_0}{\|E_0 F_0\|}$, 且实施 E_0 和 F_0 在 t 上的回归

$$E_0 = t p' + E_1 \quad F_0 = t f' + F_1$$

式中, p, f 是回归系数 (f 是标量), 即

$$p = \frac{E_0' t}{\|t\|^2} \quad f = \frac{F_0' t}{\|t\|^2}$$

记残差矩阵 $E_1 = E_0 - t p', F_1 = F_0 - t f'$ 检查收敛性, 若 Y 对 t 的回归方程已达到满意的精度, 则进行下一步; 否则, 令: $E_0 = E_1, F_0 = F_1$, 回到第 (2) 步, 对残差矩阵进行新一轮的成分提取和回归分析。

(3) 在第 h 步 ($h=2, \dots, m$), 满足交叉有效性判别^[5] 即 $Q \geq (1 - 0.95^2) = 0.0975$ 得到 m 个成分 $t_1, t_2, \dots, t_m$ 。

(4) 把步骤 (3) 得到的 m 个成分带入以下优化问题^[13-14], 并利用二次型规划方法进行求解参数 α_i, α_i^* 。

$$\max_{\alpha, \alpha^*} J = -\frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(t_i, t_j) - \epsilon \sum_{i=1}^m (\alpha_i - \alpha_i^*) + \sum_{i=1}^m Y_i (\alpha_i - \alpha_i^*)$$

$$s.t. \begin{cases} \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i - \alpha_i^* \in [0, C] \end{cases}$$

式中核函数 $k(t, t') = \langle \varphi(t), \varphi(t') \rangle$, 它是满足 Mercer 条件的任意对称函数。常用的核函数是径向基函数: $k(t, t') = \exp(-\|t - t'\|^2 / 2\sigma^2)$; 为样本个数; C 是常数, 表示函数复杂度和样本拟合度之间的折衷; ϵ 不灵敏损失函数。

(5) 通过 KKT (Karush-Kuhn-Tucker) 条件^[13], 即

$$b = Y_i - \sum_{i=1}^m (\alpha_i - \alpha_i^*) k(t_i, t) - \epsilon, \alpha_i \in (0, C) \quad \text{or}$$

$$b = Y_i - \sum_{i=1}^m (\alpha_i - \alpha_i^*) k(t_i, t) + \epsilon, \alpha_i^* \in (0, C)$$

计算偏置量 b

$$(6) \text{ 由 } \omega = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \varphi(t_i);$$

$$\hat{f}(t) = \omega^T \varphi(t) + b = \sum_{i=1}^m (\alpha_i - \alpha_i^*) k(t_i, t) + b$$

求出拟合样本集的估计函数 $\hat{f}(t)$ ^[13]。

3 PLS-SVM模型的建立与应用

珠海市位于广东省南部, 面临南海, 西江经上游马口水文站后, 分别流至磨刀门、鸡啼门、虎跳门等, 其中磨刀门水量最大, 也是珠海市重要的取水来源 (见图 1), 所以本文选取位于磨刀门水道上的平岗站盐度超标 (以含氯度大于 250 mg/L 计) 时间数作为 PLS-SVM 模型的因变量。而平岗站盐度一方面受上游来水量大小的影响 (上游来水量越小, 盐度值偏高的可能性越大, 反之亦然), 另一方面还与涨落潮紧密相关。由于联石湾水闸外盐度监测资料较长, 而且已有的研究表明^[15], 联石湾水闸外的盐度和潮汐站三灶站的潮差值之间的正相关关系非常好, 能够反映出涨落潮规律, 因此, 选取马口站的径流量 (反映上游来水量大小) 和联石湾水闸外的盐度 (反映涨落潮规律) 作为 PLS-SVM 模型的自变量。根据相关性检验, 建立珠海市月时段 PLS-SVM 模型如下:

$$Y_t = f(x_{t-1}, x_{t-2}, z_{t-1}, z_{t-2}, Y_{t-1})$$

式中: Y_t, Y_{t-1} 为平岗站第 $t, t-1$ 个月盐度超过 250 mg/L 的小时数; x_{t-1}, x_{t-2} 为第 $t-1, t-2$ 个月马口站的月径流量; z_{t-1}, z_{t-2} 为第 $t-1, t-2$ 个月联石湾水闸外的盐度超过 250 mg/L 的小时数。

根据 1998 年 10 月—2002 年 3 月实测资料, 首先选用 1998 年 10 月—2001 年 6 月的数据, 对建立平岗站 PLS-SVM 模型进行训练和模拟, 然后用

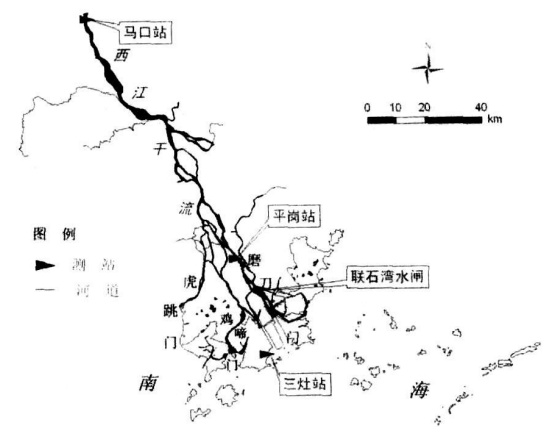


图 1 研究范围
Fig 1 Research region

2001年 7月—2002年 3月共 9个月的数据进行模型检验。

利用偏最小二乘回归提取成分并计算交叉有效性可以得出： Q^2_0 、 Q^2_1 、 Q^2_2 、 Q^2_3 分别为 0.4290、0.1474、0.1676、0.0113。由于 $Q^2_0 < 0.0975$ ，所

以选择 $h=3$ ，也就是提取 x_1 、 x_2 、 x_3 三个成分即可，3个成分表达式为

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -0.4086 & 0.4177 & 0.3524 \\ -0.3542 & 0.5848 & 0.1513 \\ 0.5938 & 0.3942 & 0.5023 \\ 0.4197 & -0.9189 & 0.4178 \\ 0.4229 & 0.5371 & -0.6527 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_{t-2} \\ z_{t-1} \\ z_{t-2} \\ y_{t-1} \end{pmatrix}$$

将提取的 x_1 、 x_2 、 x_3 作为支持向量机的输入， y_t 作为输出，进行训练学习，模拟和实测的结果见表 1 所示。

为检验模型的预报效果，运用前面已训练过的模型对 2001年 7月—2002年 3月共 9个月的咸潮情况进行预报，预报值和实测值见表 2。为了进一步分析 PLS-SVM模型的效果，表 1和表 2中还分别给出了对样本数据直接进行学习的 BP人工神经网络（BP artificial neural network 计为 BP-ANN）和多元回归分析（Multi regression Models 计为 MRM）模型的模拟和预报结果。

表 1 平岗站盐度每月超标 (大于 250 mg/L)时间实测和模拟数据表 h

Tab 1 Comparison between measured and simulated hours of monthly salinity in excess of the standards at Pinggang station

时间 (年月)	199810	199811	199812	199901	199902	199903	199904	199905	199906	199907	199908	199909
实测值	0	0	158	174	116	238	33	0	0	0	0	0
PLS-SVM模拟值	1.25	1.25	156.76	175.34	114.71	236.77	34.20	1.25	-1.26	-0.20	-1.09	-1.26
BP-ANN模拟值	0.00	23.82	136.14	181.52	131.64	223.24	68.41	2.18	0.01	0.00	0.00	0.00
MRM模拟值	-23.11	63.80	112.33	117.19	160.43	90.53	85.45	58.76	-2.53	-17.21	9.01	11.26
时间 (年月)	199910	199911	199912	200001	200002	200003	200004	200005	200006	200007	200008	200009
实测值	0	0	11	81	205	100	0	0	0	0	0	0
PLS-SVM模拟值	-0.84	-1.26	9.72	82.27	203.70	98.71	1.25	-1.26	-0.34	-1.26	0.64	-1.26
BP-ANN模拟值	0.00	0.00	24.78	83.57	204.22	73.27	6.36	0.02	0.00	0.00	0.00	0.00
MRM模拟值	-0.06	15.74	26.92	115.14	114.81	82.51	53.85	-4.27	-17.00	2.21	0.10	-13.71
时间 (年月)	200010	200011	200012	200101	200102	200103	200104	200105	200106			
实测值	0	0	46	90	84	0	0	0	0			
PLS-SVM模拟值	1.25	1.25	44.74	88.73	82.77	1.25	-1.26	1.25	1.25			
BP-ANN模拟值	0.00	0.10	15.55	89.69	60.74	36.22	2.04	0.00	0.00			
MRM模拟值	-12.44	2.48	38.50	88.44	88.03	81.12	39.46	-12.53	-19.23			

表 2 平岗站盐度月超标时间预报值和实测值对比表 h

Tab 2 Comparison between measured and predicted hours of monthly salinity in excess of the standards at Pinggang station

时间 (年月)	200107	200108	200109	200110	200111	200112	200201	200202	200203
实测值	0	0	0	0	0	5	121	14	0
PLS-SVM计算值	-1.26	1.25	1.25	1.25	-1.26	6.21	122.24	12.74	1.25
BP-ANN计算值	0.00	0.00	0.00	0.00	0.00	30.20	1.97	18.31	14.80
MRM计算值	8.02	33.83	21.90	-6.17	-12.00	22.81	41.93	62.55	57.88

分析表 1和表 2可以得出：（1）BP人工神经网络模型对学习样本的模拟效果也较为理想，说明它的学习能力较强，但由于样本数目小，容易出现“过学习”、泛化能力较差等问题，因此使得其预

报效果较差；（2）MRM是建立在自变量和因变量之间线性关系的基础上，虽总体上能反映出变化趋势，但不能满足定量化的精度需求；（3）PLS-SVM模型一方面利用偏最小二乘考虑了因子之间多重相

关的影响。另一方面又利用 SVM 考虑统计学习理论中的结构风险最小化原理, 所以模型对物理机制十分复杂且观测数据较小的咸潮发生规律的模拟和预报其结果均较好; 说明本文所建立的 PLS-SVM 模型可应用于咸潮预报; (4) PLS-SVM 咸潮预报模型计算结果中出现少量负值的原因是由于在计算中采用对原序列数据标准化后的序列, 所以在折算回原序列后存在一定误差造成的, 但对结果的影响不大; 而 MRM 计算结果中出现的负值主要是由于模型自身原因造成的, 如模型假设的线性关系不成立、各影响因素之间存在多重相关性等。

4 结 论

(1) 咸潮发生的物理机制十分复杂, 受到多种因素的影响, 且各种因素之间存在相关性, 从而影响模拟预报的精度, 而用偏最小二乘回归模型, 提取出对因变量影响最强的成分, 可提高模拟和预报精度。

(2) 由于咸潮实测资料较少, 一般是属于小样本问题, 用支持向量机对用偏最小二乘提取的成分和因变量进行模拟预测, 其模拟预报的效果均好于人工神经网络、多元回归等模型, 可为实际中的咸潮预报服务。

(3) 本文对珠海市平岗抽水站的盐度进行了模拟预测, 精度较高, 其结果可作为珠海市水资源的合理科学配置和管理提供技术支持。但由于资料较少, 还存在提高精度的空间。此实例说明, 所建立的 PLS-SVM 咸潮预报模型是简单、可行的。

参考文献:

- [1] 肖成道, 沈焕庭. 长江河口盐水入侵影响因子分析 [J]. 华东师范大学学报, 1998(3): 74-80

- [2] 宋志尧, 茅丽华. 长江口盐水入侵研究 [J]. 水资源保护, 2002(3): 27-30
- [3] 张恒喜, 郭基联, 朱家元, 等. 小样本多元数据分析方法及应用 [M]. 西安: 西北工业大学出版社, 2002
- [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] 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
- [6] TAN Y X, SHI L M, TONG W D, et al. Multi class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models [J]. Computational Biology and Chemistry 2004 28: 235-244
- [7] CHEN Q S, ZHAO J W, ZHANG H D, et al. Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration [J]. Analytica Chimica Acta 2006 572: 77-84
- [8] 罗小峰, 陈志昌. 径流和潮汐对长江口盐水入侵影响数值模拟研究 [J]. 海岸工程, 2005 24(3): 1-6
- [9] Vapnik V 著. 统计学习理论的本质 [M]. 张学工译. 北京: 清华大学出版社, 2000
- [10] 张浩然, 韩正之, 李昌刚. 基于支持向量机的非线性模型预测控制 [J]. 系统工程与电子技术, 2003 25(3): 330-334
- [11] 刘江华, 陈佳品, 程君实. 基于支持向量机的非线性系统辨识 [J]. 测控技术, 2002 21(11): 54-58
- [12] 王景雷, 吴景社, 孙景生, 等. 支持向量机在地下水位预报中的应用研究 [J]. 水利学报, 2003(5): 122-128
- [13] 熊建秋, 李祚泳. 基于支持向量机的水流挟沙能力预测研究 [J]. 水利学报, 2005 36(10): 1171-1175
- [14] 邓乃扬, 田英杰. 数据挖掘中的新方法: 支持向量机 [M]. 北京: 科学出版社, 2004
- [15] 闻平. 珠江三角洲咸潮入侵规律研究 [D]. 广州: 中山大学, 2006

Model for Prediction of Saltwater Intrusions Based on Coupling of Support Vector Machine and Partial Least Square Method

LU De-di CHEN Xiao-hong

(Center for water resource and environment, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: The factors affecting the saltwater intrusions are analyzed by means of partial least square (PLS) method to extract the most important components for solving multiple correlation among variables. Besides, the support machine vector (SVM) method is used for its advantage of solving non-linearity and little samples. A model for prediction of saltwater intrusions (PLS-SVM) based on the coupling of SVM and PLS is established and also successfully used in simulation and prediction of saltwater intrusion at Pinggang station in Zhuhai city. The results show that the PLS-SVM model is better than BP artificial neural network and Multi regression Models in the result of simulating and prediction. So it can be used for forecasting of saltwater intrusions.

Key words: partial least square method; support vector machine; forecasting of saltwater intrusion; Zhuhai city