

支持向量机在湄池站洪峰水位预报中的应用

刘艳伟^{1,2}, 闵惠学², 郦 英²

(1.浙江大学, 浙江 杭州 310058; 2.浙江省水文局, 浙江 杭州 310009)

摘 要:支持向量机(SVM, Support Vector Machines)是在统计学习理论中发展起来的一种处理非线性分类和非线性回归的有效方法。感潮河段洪水位是复杂的洪、潮非线性组合问题, 本文尝试将SVM方法应用于感潮河段湄池站洪峰水位预报, 通过选取湄池站历史洪水中分别反映上游来水和下游顶托作用的预报因子, 建立湄池站洪峰水位的SVM回归模型, 获得了较好预报效果。

关键词:支持向量机; 非线性回归; 感潮河段; 水位预报

中图分类号:P731 **文献标识码:**A **文章编号:**1003-0239(2011)01-0072-05

1 引言

感潮河段的洪水过程受上游径流和下游潮汐双重影响, 进行准确的水位预报十分困难。由于洪潮相互作用机理复杂, 人们对其认识尚有局限, 还不能进行有效的数学模型描述。基于统计理论的回归分析方法由于对断面资料要求不高, 不需考虑河段洪水过程的物理意义, 简单实用, 目前在感潮河段洪峰水位预报中应用较为广泛。但是感潮河段水位系列在时序上具有相依性、突变性和随机性等复杂非线性特征^[1], 而目前大多数回归分析方法是建立在线性相关基础上的(如近年来使用较多的多元逐步回归方法)^[2-3], 用于处理此类复杂的非线性问题时具有较大的局限性。

支持向量机(SVM, Support Vector Machines)是20世纪90年代提出的一种处理非线性分类和非线性回归的新的通用学习方法^[4-5], 它是建立在Vapnik等人提出的统计学理论^[6-7]的VC维(Vapnik-Chervonenks Dimension)理论和结构风险最小原理(Structure Risk Minimization)基础上的, 能较好地解决样本空间中的高度非线性分类和回归等问题, 已成功的应用于分类、函数逼近和时间序列预测等方面。本文尝试将SVM应用于感潮

河段湄池站的洪峰水位预测, 通过选取湄池站分别反映上游来水和下游顶托作用的预报因子, 利用历史洪水样本资料构造湄池站洪峰水位的SVM回归预报模型, 预测湄池站洪峰水位, 预报精度优良。

2 SVM回归算法原理^[8]

SVM解决非线性回归问题的基本思路是基于Mercer核展开定理^[9], 通过非线性变换 Φ , 将样本空间映射到一个高维乃至无穷维的线性特征空间(Hilbert空间), 在高维特征空间通过一个线性超平面实现线性回归, 使在特征空间中可以应用线性学习的方法解决样本空间中的高度非线性分类和回归等问题。

假设给定的样本数据集为:

$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)$, 其中, $x_i \in R^N$, 为N维向量, $y_i \in R$ 。在SVM回归中, 输入样本 x_i 首先用非线性变换映射到一个m维的特征空间, 然后在这个特征空间中建立一个线性模型, 用公式表示为:

$$f(x, w) = w \cdot \Phi(x) + b \quad (1)$$

式中, $f(x, w)$ 是估计函数; w 为权向量;

$\Phi(x)$ 为非线性映射集合; b 为阈值。

其正确性是由损失函数来衡量的, SVM 回归引入一种新的损失函数叫做 ϵ (不敏感损失函数), 它是由 Vapnik 提出的。同时考虑到允许拟合误差的情况, 引入松弛因子 $\xi_i \geq 0, \xi_i^* \geq 0$, 用于表示引入训练集的误差。根据结构风险化最小原则即要寻求最优回归超平面使:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (2)$$

约束条件为:

$$\begin{cases} y_i - w \cdot \Phi(x) - b \leq \epsilon + \xi_i^* \\ w \cdot \Phi(x) + b - y_i \leq \epsilon + \xi_i \\ \xi_i \geq 0 \\ \xi_i^* \geq 0 \end{cases} \quad i=1, 2, \dots, k \quad (3)$$

式中, 常数 $C(C>0)$ 为惩罚系数, 表示控制训练误差的代价。这样, 问题转化为求解如公式(2)一个二次凸规划问题。由于目标函数和约束条件都是凸的, 根据最优化理论, 这一问题存在唯一全局最小解。公式(2)的优化问题通过引入拉格朗日函数将其转化为对偶问题, 通过解对偶问题得到公式(2)的解:

$$\begin{cases} f(x) = \sum_{i=1}^{n_{SV}} (\alpha_i - \alpha_i^*) (\Phi(x_i) \cdot \Phi(x)) + b \\ s.t. \quad 0 \leq \alpha_i \leq C \quad 0 \leq \alpha_i^* \leq C \end{cases} \quad (4)$$

根据 Mercer 核定理, 核函数 $K(x, y) = \sum_i \lambda_i \phi_i(x) \phi_i(y) = \Phi(x) \cdot \Phi(y)$, 则公式(4)转换为:

$$\begin{cases} f(x) = \sum_{i=1}^{n_{SV}} (\alpha_i - \alpha_i^*) K(x_i, x) + b \\ s.t. \quad 0 \leq \alpha_i \leq C \quad 0 \leq \alpha_i^* \leq C \end{cases} \quad (5)$$

这就是 SVM 方法最终确定的非线性回归函数。其中, α_i 、 α_i^* 为拉格朗日乘子, n_{SV} 为支持向量的个数。 α_i 、 α_i^* 和 b 通过约束条件求得, 为确定最优超平面的参数。

支持向量机理论只考虑高维特征空间的点积运算 $K(x_i, x) = \Phi(x_i) \cdot \Phi(x)$, 而不直接使用函数 Φ , 从而巧妙地解决了因 Φ 未知而 w 无法显式表达的问题。已经证明, 只要满足 Mercer 条件的对称函

数即可作为核函数, 常用的核函数有:

(1) 多项式核函数

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad d=1, 2, 3, \dots \quad (6)$$

(2) 径向基函数(RBF)核函数

$$K(x_i, x_j) = \exp \left[-\gamma \|x_i - x_j\|^2 \right] \quad (7)$$

(3) Sigmoid 核函数

$$K(x_i, x_j) = \tanh [b(x_i \cdot x_j) + c] \quad (8)$$

上述求解过程有多种高效算法和成熟的计算机程序可资利用^[10]。本文所应用的核心软件为台湾大学林智仁 (Lin Chih-Jen) 副教授等开发设计的 LibSVM^[11], 它是一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包, 不但提供了编译好的可在 Windows 系列系统的执行文件, 还提供了源代码, 方便改进、修改以及在其它操作系统上应用。该软件为非商业应用的免费软件。

3 应用^[12~14]

3.1 构造样本资料

湄池站位于浦阳江下游感潮河段, 该站洪水位受浦阳江上游来水和钱塘江下游洪、潮水顶托的共同影响。诸暨水文站位于浦阳江中游, 距离湄池站约 30 km, 其实测资料能准确反映浦阳江流域来水情况。闻家堰水文站位于浦阳江与钱塘江汇合口附近, 距离湄池站大约 40 km, 观测资料能较好的反映钱塘江洪、潮顶托影响。根据资料情况, 构建预报因子如下:

(1) 诸暨洪峰水位 X_1 和涨幅 X_2 , 反映上游来水影响;

(2) 诸暨洪峰出现前的闻家堰最高潮位因子 X_3 , 反映下游洪潮顶托影响;

(3) 闻家堰高潮位前涨潮潮差因子 X_4 。历史资料表明, 相同量级的诸暨站洪水与下游闻家堰站不同洪、潮以及江道情况组合, 会使湄池洪峰水位出现较大差异。据分析, 闻家堰高水位主要由潮水所致时, X_4 的值较大, 由富春江来水所致时, X_4 的值较小。根据闻家堰涨潮历时, 当潮型消失时 X_4 取峰前 3 h 的涨幅。

本文以湄池站洪峰水位作为预报对象, 以 X_1 、 X_2 、 X_3 和 X_4 为预报因子, 选取 1984 年 (考虑

到大型水利工程对河道状况的影响,选取陈蔡水库建成后的资料)至2010年共26 a的洪水资料,构建样本数据集,共计选出38场洪水样本。其中32场洪水作为训练集,另外的6场洪水作为检验集,不参加建模过程。

3.2 构建模型

模型构建的过程也就是选择合适的核函数,并通过对样本集的训练最终确定最优模型参数的过程。模型的优劣以均方差(MSE)和相关系数

(r^2)作为判别指标,公式如下:

3.2.1 选择核函数

核函数的选择^[15]目前国际上还没有形成一个统一的模式,一般凭借经验。LibSVM提供了常用的3种非线性核函数,本文根据试验对比最终选定径向基函数作为核函数建立SVM模型。

3.2.2 确定模型参数

LibSVM软件对SVM所涉及的参数调节相对比较少,提供了很多的默认参数,对于径向基核函数的回归问题需要人为优选确定的参数主要有

$$MSE = \frac{1}{l} \sum_{i=1}^l (f(x_i) - y_i)^2, \quad r^2 = \frac{\left(l \sum_{i=1}^l f(x_i) y_i - \sum_{i=1}^l f(x_i) \sum_{i=1}^l y_i \right)^2}{\left(l \sum_{i=1}^l f(x_i)^2 - \left(\sum_{i=1}^l f(x_i) \right)^2 \right) \left(l \sum_{i=1}^l y_i^2 - \left(\sum_{i=1}^l y_i \right)^2 \right)} \quad (9)$$

惩罚系数C、损失函数 ϵ 和核参数 γ 。LibSVM提供了交互检验(Cross Validation)的参数优选功能:将原始数据分成K组(一般均等分,K大于等于2),将每组数据分别做一次验证集,其余的K-1组数据作为训练集,这样会得到K个模型,用某组参数训练K个模型得到的K个评分结果(MSE和 r^2)的平均值作为最终的评分结果,并以此为判断参数优劣的指标,通过逐步筛选的方法调试参数,直至达到最优结果。交互检验方法可以有效的避免过学习以及欠学习状态的发生,最后得到的结果也比较具有说服力。

通过以上方法,取 $k=5$ 对训练集进行参数逐步逼近训练,得到最佳评分结果的SVM回归模型,表1列出了模型参数以及评分结果。可以看出,对训练集的回报效果和对检验集的预测效果

都很理想,相关系数都在0.99以上,均方差都不到0.02,这表明建立的SVM回归预报模型具有较好的稳定性。

3.2.3 精度评价

表2列出了SVM模型对训练样本的回报和检验样本的预测结果,以洪峰水位绝对误差小于0.2 m作为合格标准。由表可知,对训练集回报的合格率为30/32,平均绝对误差0.006 m,最大误差为0.29 m;对检验集预报的合格率为5/6,平均绝对误差0.017 m,最大误差为0.24 m。根据水文情报预报规范的有关规定,本方案达到甲级精度。

4 小结

SVM方法是一种处理高度非线性问题的有效

表1 用SVM回归方法建立涪池站洪峰水位预测模型的评价结果

最终参数组	训练结果	对训练集的回报评分结果	对检验集的预测评分结果
C=8	训练样本32个	MSE=0.006	MSE=0.017
$\gamma=0.5$	支持向量28个	$r^2=0.992$	$r^2=0.991$
$\epsilon=0.05$	迭代次数231次		

的机器学习方法,本文是支持向量机方法应用于洪水位预测的初步尝试,结果显示:在受洪、潮等复杂因素影响的涪池站,通过选择适当的预报因子建立的SVM洪峰水位预报模型,具有较好的拟合精度,达到规范规定的甲等预报方案标准,

可用于作业预报。

由于SVM方法是通过对历史样本的不断学习来建立预报模型的,SVM预报模型的建立过程就是对历史样本的自学习记忆过程。因此,用尽可能多的样本参与训练,使预报模型包含比较完备

表2 SVM回归模型预测湄池站洪峰水位的结果

样本 序号	洪峰时间	实测洪峰 水位(m)	预报因子 X_1 (m)	预报因子 X_2 (m)	预报因子 X_3 (m)	预报因子 X_4 (m)	预测洪峰 水位(m)	绝对误差 (m)	精度 评定
1	1986-4-11 16:00	7.69	9.9	6.57	0.52	2.73	7.74	0.05	平均绝对误差 0.06 m 合格率 93.75 %
2	1987-9-11 19:55	7.72	9.75	6.95	1.35	2.42	7.77	0.05	
3	1988-6-20 6:30	8.87	11.5	5.96	0.25	4.69	8.85	0.02	
4	1989-5-23 23:50	7.66	10.58	5.87	0.26	3.56	7.71	0.05	
5	1989-6-18 21:00	7.53	10.46	5.47	1.23	3.32	7.48	0.05	
6	1989-7-4 21:40	8.73	10.47	6.89	0.17	3.23	8.68	0.05	
7	1989-8-23 14:00	7.37	9.77	5.05	1.48	2.86	7.42	0.05	
8	1989-9-16 22:00	8.16	10.88	6.41	1.84	4.08	8.11	0.05	
9	1990-9-1 6:00	9.77	12.46	6.14	0.23	5.48	9.72	0.05	
10	1991-4-18 23:50	7.6	10.68	6.19	0.4	3.42	7.83	0.23	
11	1992-3-26 22:25	7.39	9.57	6.68	0.1	2.27	7.44	0.05	
12	1992-6-27 2:00	8.01	10.78	5.98	0.1	3.64	7.99	0.02	
13	1992-7-5 0:00	9.18	11.45	7.07	0.67	4.38	9.13	0.05	
14	1992-9-24 8:00	7.49	10.32	5.78	0.65	3.12	7.44	0.05	
15	1993-3-28 16:30	7.21	9.4	6.13	0.1	1.73	7.26	0.05	
16	1993-7-2 8:39	7.98	10.99	5.31	0.43	3.76	7.95	0.03	
17	1993-7-5 7:40	9.03	11.03	7.22	0.6	2.93	8.98	0.05	
18	1994-6-14 2:00	9.72	12.66	6.77	0.5	2.99	9.67	0.05	
19	1994-6-17 20:38	10.10	12.55	6.64	0.83	4.23	10.05	0.05	
20	1994-8-22 9:40	7.39	10.69	7.07	3.03	3.18	7.44	0.05	
21	1995-4-30 9:00	8.42	11.15	6.85	0.68	3.78	8.47	0.05	
22	1995-6-26 5:28	8.51	10.6	6.79	0.39	3.04	8.52	0.01	
23	1995-7-5 11:00	7.29	10.25	5.95	0.9	3.18	7.34	0.05	
24	1996-3-20 19:10	7.42	9.22	6.58	0.42	2.34	7.37	0.05	
25	1996-6-6 8:50	7.39	9.49	6.92	0.69	1.7	7.44	0.05	
26	1996-7-1 22:00	8.62	10.35	7.00	0.25	2.62	8.57	0.05	
27	1997-7-9 16:16	10.48	12.7	7.94	0.24	5.2	10.43	0.05	
28	1997-8-19 21:32	9.14	11.71	7.04	1	4.7	9.19	0.05	
29	1998-1-15 7:57	7.63	9.81	6.32	1.08	2.52	7.58	0.05	
30	1998-6-19 20:00	8.11	10.64	6.08	0.33	3.38	7.82	0.29	
31	1999-6-19 1:00	8.51	11.06	6.54	0.52	3.15	8.46	0.05	
32	2001-6-27 0:30	9.1	10.66	7.47	0.2	3.2	9.15	0.05	
33	2002-4-25 5:50	7.22	9.22	5.98	0.29	1.69	7.37	0.15	平均绝对误差 0.11 m 合格率 83.3 %
34	2002-6-29 7:07	7.45	10.15	6.12	0.7	2.96	7.51	0.06	
35	2002-7-2 6:00	8.16	10.55	6.13	0.31	2.25	8.21	0.05	
36	2006-5-19 14:00	7.28	9.61	5.75	0.23	1.61	7.52	0.24	
37	2007-11-9 9:28	8.92	11.17	6.63	0.38	2.84	8.90	0.02	
38	2010-3-6 23:00	8.89	10.44	7.46	0.1	2.3	8.74	0.15	

的支持信息,可以大大提高预报精度。建议在实
际预报中适时增加新的样本参与训练,不断优化
预报模型。

参考文献:

- [1] 张小琴,包为民.感潮河段预报方法浅析[J].水电能源科学,2009,27(3):8-10.
- [2] 卢金利,俞昌都.逐步回归法在临海站洪水预报中的应用[J].浙江水利科技,2001,(S):120-121.
- [3] 周文斌,车倩.多元线性回归法在水文预报中的应用[J].山西建筑,2009,35(1):359-360.
- [4] Cristianini N and Shaw-Taylor J. An Introduction of Support Vector Machines and Other Kernel_based Learning Methods[M]. Cambridge:Cambridge University Press, 2000.
- [5] Burges C J. A tutorial on support vector machines for pattern recognition[J].Data Mining and Knowledge Discovery,1998,2: 127-167.
- [6] Vapnik V N. Statistical Learning Theory[M].New York:John Wiley & Sons,Inc.,1998.
- [7] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag, 2000.
- [8] 陈永义,俞小鼎,高学浩.处理非线性分类和回归问题的一种新方法(I)-支持向量机方法简介[J].应用气象学报,2004,15(3): 345-354.
- [9] Courant R and Hilbert D.Method of Mathematical Physics[M]. New York:Springer Verlag,1953.
- [10] <http://www.kernel-machines.org>
- [11] <http://www.csie.ntu.edu.tw/~cjlin/index.html>
- [12] 卢敏,张展羽,冯宝平.支持向量机在径流预报中的应用探讨[J].人民长江.2005,36(8):38-39.
- [13] 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.
- [14] 李智才,马文瑞,李素敏等.支持向量机在短期气候预测中的应用[J].气象.2006,32(5):57-61.
- [15] 朱树先,张仁杰.支持向量机核函数选择的研究[J].科学技术与工程.2008,8(16):4514-4517.