文章编号:1671-6833(2009)04-0123-05

基于聚类的支持向量机在洪水预报中的应用

胡彩虹,王艳菊,吴泽宁

(郑州大学 水利与环境学院,河南 郑州 450001)

摘 要:半干旱地区的特殊特点使其径流模拟计算难度增大,且难以获得较详细的资料,因而洪水预报难度大,尤其是洪峰流量的预报.若应用所有样本进行模型参数确定并预报,不能完全反映洪水的不同特性.因此采用了基于聚类分析的支持向量机模型,以半干旱半湿润地区的岚河流域为例,进行了模拟检验,结果表明,效率系数大部分达到85%以上,平均相对误差绝对值多数都小于1.5%.另外洪峰流量相对误差绝对值均在15%以内,特别洪峰流量较大的几场洪水,相对误差小于1%.洪峰流量和峰现时差合格率均达100%.

关键词:系统聚类;支持向量机;洪水预报

中图分类号:TV 122^{+}.5 **文献标识码**:A

0 引言

由于受到气象气候、水文及水力学等复杂过程和地形地貌等流域下垫面系统和人类活动等因素的综合影响,使洪水形成过程是一个复杂的高度非线性系统.降雨径流模型、回归分析模型、时间序列模型、神经网络模型、模糊数学及灰色系统模型等在过去的水文预报中都有广泛应用[1],但也都存在一些不足,用于小样本资料洪水预报时不能得到十分满意的结果.而支持向量机(SVM)方法能够较好地解决小样本、非线性、高维数等实际问题[2],且具有较强的泛化能力.此外,传统的洪水预报是根据流域产汇流原理和河道洪水波的运动原理寻求其关系进行演算,SVM所具有的自学习能力,决定了它具有对复杂非线性关系的识别和处理能力.其在水文中的应用也在不断地发展,国内外亦已将SVM应用于水文预报中[3-5],但现有洪水预报模型应用时,多采用全部样本训练并预报,由于洪水特性不同,就可能导致个别年份洪水预报精度较低,尤其是洪峰流量预报精度较差的现象,会影响防洪决策.

另外,在半干旱地区,目前国内外尚无比较完善适用的洪水预报模型,因此一直是研究的热点和难点[6-8].

笔者针对以上情况,在半干旱地区,提出基于聚类考虑分级进行洪水预报,合理地完成降雨和径流的非线性映射,寻求其合理的内在关系.聚类后采用SVM模型进行洪水预报.

1 基于系统聚类的支持向量机模型构建

1.1 系统聚类[9]

系统聚类法的基本思想是:先将n个样品自成一类,然后规定样品之间的距离和类与类之间的距离.分类开始时,因每个样品自成一类,类与类之间的距离与样品之间的距离是相等的,选择距离最小的一对并成一个新类,计算新类和其它类的距离,再将距离最近的两类合并,这样每次减少一类,直至所有样品都成一类为止.

聚类步骤如下:1)数据标准化处理;2)计算洪水样品之间的距离;3)计算类与类之间的距离,并选择距离最小的合成一个新类;4)计算新类和其它各类之间的距离,重复步骤3,至所有样本归为一类为止;5)画出聚类树;6)根据聚类树及实际情况,确定分类结果.

1.2 支持向量机原理[2]

假设已知训练样本数据为(x_1,y_1),(x_2,y_2),…,(x_n,y_n),n为训练样本个数,$x_i\in R^m$为模型输入,$y_i\in R$为模型输出.通过训练寻求输出y_i对

收稿日期:2009-04-16;修订日期:2009-07-27

基金项目:河南省教育厅自然科学研究资助计划(2009A57008)

作者简介:胡彩虹(1968-)女,山西平遥人,郑州大学副教授,博士,主要从事水文水资源方面的研究,E-mail:hucaihong@zzu.edu.cn.

于输入 x_i 的最优拟合函数,使拟合或预报偏差最小. 设回归函数 $y=f(x)=\langle w,\phi(x)\rangle+b$,由结构风险最小化准则,将问题转化为

$$\text{Min } J=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}R(f(x_i),y_i) \quad (1)$$

约束条件:

$$\begin{cases} y_i-\langle w,\phi(x_i)\rangle-b\leqslant\varepsilon+\xi_i^* \\ \langle w,\phi(x_i)\rangle+b-y_i\leqslant\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*\geqslant 0 \end{cases} \quad (2)$$

式中:J 为经验误差;C 为误差惩罚因子,是一个正常数;R 为惩罚函数,ε 为不敏感损失函数度量;$\langle\rangle$ 为内积函数;$\phi(x)$ 为输入空间到高维空间的非线性映射;ξ_i 和 ξ_i^* 为松弛变量. 最终确定的回归方程为:

$$f(x,w)=\langle w,x\rangle+b=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x)+b \quad (3)$$

式中:$K(x_i,x)=\langle\phi(x_i),\phi(x)\rangle$ 为核函数,代替内积函数;α_i,α_i^* 为 Lagrange 乘子.

1.3 基于聚类的支持向量机方法构建

提取洪水样本特征值标准化. 以相似系数作聚类指标,借助 MATLAB 来实现洪水聚类. 样品间的距离采用欧氏距离计算,即

$$d_{ij}=\sqrt{\sum_{t=1}^{p}(x_{it}-x_{jt})^2} \quad (i,j=1,2,\cdots,n) \quad (4)$$

类与类之间的距离采用类平均距离法,即

$$D_{pq}^2=\frac{1}{n_p n_q}\sum_{i\in C_p,j\in C_q}d_{ij}^2 \quad (5)$$

式中:n_p 和 n_q 分别为类 G_p 和类 G_q 中的洪水样本个数,选择距离最小的合成一个新类. 再计算新类与其它类之间的距离,直至所有样本归为一类为止,最后确定聚类结果.

对各类洪水分别建模并预报. 通过计算比较,确定预报因子为提前 1 h、2 h 和 3 h 的流量及同时刻降雨. 建模前对所有数据做归一化处理,并分割,将全部样本数据的 75% 作为训练样本集,20% 为试验样本集,5% 为检验样本集.

需确定的参数主要有 C、ε 和核函数类型及其参数. 笔者采用径向基核函数的一种 $K(x,x_i)=\exp(-g\|x-x_i\|^2)$,其余参数通过试算确定,回归模型的择优标准采用绝对差,模型的评价标准采用效率系数和平均相对误差. 洪峰流量平均相对误差绝对值≤30% 时为合格,峰现时差≤±3 h 为合格.

2 应用实例

2.1 研究区域概况

岚河流域地处半干旱地区,是汾河流域的一个子流域. 汾河水库蓄泄水直接影响下游工农业、城市发展以及人民生活. 而位于岚河上的上静游水文站是上游洪水进入汾河水库的主要控制测站之一. 采用岚河流域上静游水文站 1992 年到 2003 年实测降水流量资料建模分析.

2.2 洪水聚类结果

笔者仅考虑洪水总量、洪峰流量和洪水历时 3 个主要洪水特征值. 缺少的数据资料将其延伸使洪水过程线连续. 另外在此忽略蒸发对流量的影响. 根据已知实测资料提取洪水特征值如表 1,应用 MATLAB 统计工具箱得到聚类树见图 1.

表 1 岚河流域洪水特征值

Tab 1 The flood characteristics of the Lanhe River

编号	年份	洪水总量 /($m^3\cdot s^{-1}$)	洪峰流量 /($m^3\cdot s^{-1}$)	洪水历时/h
1	1992	5 531.12	261.00	1 272
2	1993	936.81	161.40	71
3	1994	1 335.35	75.00	333
4	1996	3 762.54	41.60	212
5	1998	502.68	24.40	165
6	2000	96.80	6.87	53
7	2001	253.75	7.40	207
8	2002	1 039.64	27.20	261
9	2003	125.25	5.22	130

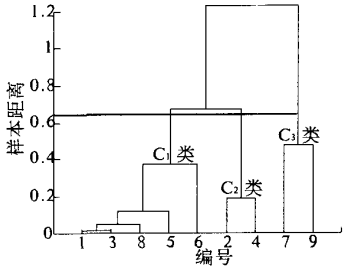

图 1 岚河流域洪水聚类结果

Fig. 1 Flood clustering results of the Lanhe River

依据聚类结果中各样本之间的距离差异及实际情况,将洪水样本分为三类:

$C_1=\{1,3,8,5,6\}=\{1992,1994,1998,2000,2001,2002\}$,

$C_2=\{2,4\}=\{1993,1996\}$,

$C_3=\{7,9\}=\{2001,2003\}$

2.3 洪水预报模拟结果

2.3.1 分类预报模拟结果

用 CMSVM2.0 软件建模,得 C_1、C_2、C_3 三类洪水最优参数分别为 $C_1=2\ 000$,$g_1=10$,$\varepsilon_1=$

0.1；C_2 = 1 001，g_2 = 25，ε_2 = 1.1；C_3 = 4 001，g_3 = 0.001 5，ε_3 = 0.000 5. 拟合预报结果见图 2(a)、(b)、(c)及表 2 和表 3.

由图 2 可看出，基于系统聚类的 *SVM* 模型对洪水动态过程的模拟较好，尤其是洪峰流量拟合精度较高. 由表 2 可知，效率系数大部分达 85% 以上，平均相对误差多数绝对值都在 1.5% 以内. 个别的效率系数与平均相对误差不太一致，如 C_2 类和 C_3 类洪水预报阶段效率系数分别为 98.99%、63.2%，而平均相对误差分别为 21.37%、1.29%. 这主要是因为半干旱地区洪峰陡涨陡落历时短，峰现时间难以准确预报，其稍有误差，洪峰前后几个时段预报流量的误差会很大，DC 就会低[10]，但平均的相对误差可能不大. 从图 2(c)和表 3 均可看出，C_3 类洪水的预报峰现时间均延迟了 1 h，C_1 和 C_2 类洪水个别也有类似现象. 表 3 中洪峰流量相对误差绝对值均小于 15%，特别洪峰流量较大的几场洪水相对误差小于 1%. 洪峰流量和峰现时差合格率均达 100%.

2.3.2　分类前预报模拟结果

全部洪水样本最优参数为 C = 81，g = 0.85，ε = 0.05，拟合预报结果见图 2(d)及表 2、表 3.

由表 2 可知，效率系数达 80% 以上，平均相对误差绝对值小于 1.5%. 但由图 2(d)可看出，分类前 SVM 模型对洪峰流量较大的洪水的预报结果并不理想. 从表 3 可知，1992 年 8 月 28 号和 1993 年 7 月 25 号的两场大洪水中，其相对误差绝对值超过了 30%. 洪峰流量合格率为 89.5%，峰现时差合格率为 100%.

2.3.3　分类后与分类前预报结果比较

表 2 中分类后的效率系数和平均相对误差除个别情况外，均较分类前有所改善. 图 2 中分类后的预报拟合情况尤其洪峰流量要比分类前好得多. 由表 3 也可看出，洪峰流量相对误差的合格率由分类前的 89.5% 提高到了分类后的 100%. 大部分洪水洪峰流量的相对误差均较分类前有明显降低. 特别地 1992 年 8 月 28 号和 1993 年 7 月 25 号的两场大洪水，相对误差绝对值由分类前的超过 30% 降低到了分类后的小于 1%，其余也均在允许范围内. 由此可见，基于聚类的 SVM 模型用于半干旱地区时，能较好地模拟洪水过程，同时也大大提高了洪峰流量的预报精度，弥补了分类前的缺陷.

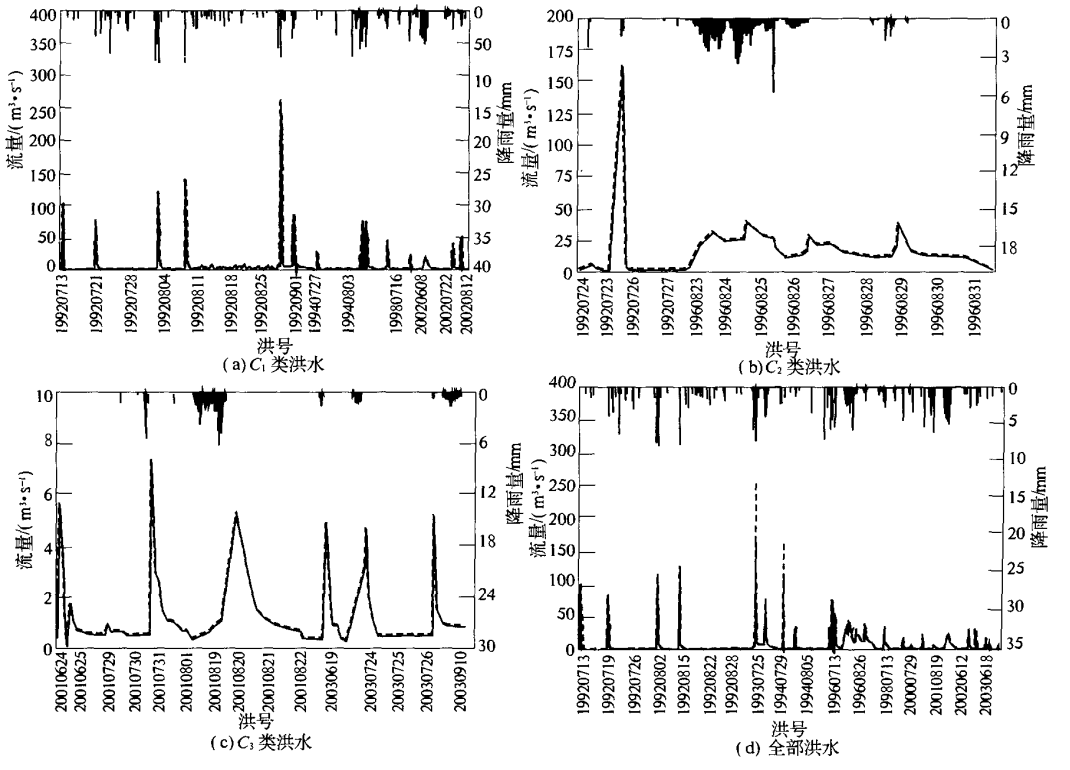

图 2　岚河流域分类前后洪水实测流量与预报流量拟合情况

Fig. 2　The validation results of observed and forecasted flows before and after the classification of the Lanhe River

(主坐标轴中实线表示预报流量，虚线表示实测流量)

表2 基于聚类的支持向量机模型的检验和预报结果

Tab.2 The testing and forecasting results of the support vector machine model based on clustering %

	样本	效率系数 DC	平均相对误差	相对误差绝对值落在下列区间的百分比 [0,5]	[0,15]	[0,25]	[0,25]以上
分类后	C1类检验阶段	76.25	1.35	66.35	85.58	91.35	100
	C1类预报阶段	85.32	-0.75	76.69	89.35	92.73	100
	C2类检验阶段	95.37	7.60	42.86	92.86	100.00	100
	C2类预报阶段	98.99	21.37	47.86	78.93	82.14	100
	C3类检验阶段	100.00	-0.88	93.75	100.00	100.00	100
	C3类预报阶段	63.20	1.29	67.89	83.49	92.66	100
分类前	检验阶段	98.99	1.06	82.24	97.06	99.26	100
	预报阶段	81.21	-1.31	87.47	97.14	99.03	100

表3 洪峰流量误差评定表

Tab.3 Error assessment of flood peak

类别	洪号	实测洪峰流量/($m^3 \cdot s^{-1}$)	预报洪峰流量/($m^3 \cdot s^{-1}$) 分类前	分类后	相对误差/% 分类前	分类后	实测峰现时间	峰现时差/h 分类前	分类后
C_1	19920713	91.40	99.67	98.64	9.05	7.92	19920713T17	0	-1
	19920720	74.40	82.06	68.60	10.30	-7.80	19920720T23	0	-1
	19920802	115.00	110.95	114.90	-3.52	-0.09	19920802T22	0	0
	19920808	131.00	123.93	131.10	-5.40	0.08	19920808T18	0	-1
	19920828	261.00	144.45	260.90	-44.66	-0.04	19920828T18	0	0
	19920831	76.50	76.83	76.60	0.43	0.13	19920831T12	0	0
	19940805	75.00	70.74	67.03	-5.68	-10.63	19940805T12	0	0
	19980713	24.40	24.02	24.08	-1.56	-1.31	19980713T12	+1	0
	20000727	6.87	7.53	6.97	9.61	1.46	20000727T14	0	-2
	20020810	27.20	30.88	28.90	13.53	6.25	20020810T07	0	-2
C_2	19930725	161.40	111.59	160.30	-30.86	-0.68	19930725T16	0	0
	19960824	41.60	42.70	42.70	2.64	2.64	19960824T14	0	-1
	19960828	33.04	34.82	35.10	5.39	6.23	19960828T21	0	-1
C_3	20010624	5.72	6.38	4.97	11.54	-13.11	20010624T19	0	-1
	20010730	7.40	8.24	6.41	11.35	-13.38	20010730T20	0	-1
	20010819	5.51	5.47	5.09	-0.73	-7.62	20010819T19	0	-1
	20030618	5.14	5.72	4.47	11.28	-13.04	20030618T19	0	-1
	20030723	4.79	4.92	4.36	2.71	-8.98	20030723T20	0	-1
	20030726	5.22	5.65	4.62	8.24	-11.49	20030726T19	0	-1
	合格率				89.5%	100%		100%	100%

注:表3中19920713T17表示1992年07月13日17时,其它类推.

3 结语

洪水分类预报可以明显提高洪峰流量的预报精度,洪峰流量相对误差的合格率由分类前的89.5%提高到了分类后的100%.特别地大洪水洪峰流量相对误差绝对值由分类前的大于30%降低到了分类后的小于1%.各类洪水预报拟合情况也要比分类前好得多.因此,证明了应用基于聚类的SVM方法进行半干旱地区的洪水预报能够得到更好的结果,可作为半干旱地区一个相对较好的洪水预报模型.

参考文献:

[1] 王 文,马 骏. 若干水文预报方法综述[J]. 水利水电科技进展,2005,25(1):56-59.

[2] 廖 杰,王文圣,李跃清,等. 支持向量机及其在径流预测中的应用[J]. 四川大学学报,2006,38(6):24-28.

[3] 张土乔,俞亭超. 提高支持向量机洪水峰值预报精度研究[J]. 水力发电学报,2005,24(2):35-39.

[4] 林剑艺,程春田. 支持向量机在中长期径流预报中

的应用 [J]. 水利学报,2006,37(6):681 - 686.
[5] CHEN S T, YU P S. Pruning of support vector networks on flood forecasting[J]. Journal of Hydrology, 2007,347:67 - 78.
[6] 李　琪. 全国水文预报技术竞赛流域水文模型分析 [J]. 水科学进展,1998,9(2):187 - 195.
[7] 胡彩虹,郭生练,彭定志,等. 半干旱半湿润地区流域水文模型分析比较研究 [J]. 武汉大学学报:工学版, 2003,36(5):38 - 42.
[8] HU C H, GUO S L, XIONG L H, et al. A modified Xinanjiang model and its application in Northern China [J]. Nordic Hydrology, 2005.36(3):175 - 192.
[9] 张尧庭,方开泰. 多元统计分析引论[M]. 北京:科学出版社,1982:313 - 361.
[10] 陈玉林,韩家田. 半干旱地区洪水预报的若干问题 [J]. 水科学进展,2003,14(5):612 - 616.

Application of Support Vector Machine Based on Clustering in Flood Forecast

HU Cai - hong, WANG Yan - ju, WU Ze - ning

(School of Water Conservancy and Environment Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: The difficulty of the runoff simulation has been increased due to the specificity of the semi - arid region, which is still a hot and difficult topic in current study. The detailed information is hard to obtain because of the complicated process of the runoff, so it is difficult to forecast flood, especially the flood peak forecast. The different characteristics of flood can not be reflected completely if all samples were used to calibrate the parameter of the model. Thus the support vector machine method based on clustering is used. And the Lan River basin in the semi - arid region is taken as an example to be simulated and tested. The results have shown that most efficient coefficient is beyond 85%, and the modulus of the relatively average error is mostly less than 1.5%. Additionally, the peak flow modulus of the relatively average error is less than 15%, particularly in the flood of large peak flow, the relatively average error is less than 1%. The qualification rate of the peak flow and the peak time difference is 100%.

Key words: system clustering; support vector machine; flood forecasting

(上接第 122 页)

Research on Impact of Termination Impedance to Crosstalk of Transmission Line

LUO Ying - hong, ZHANG Bo

(College of Automatic and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: This paper establishes the mode of corresponding equivalent circuit for transmission line coupling in homogeneous and inhomogeneous media based on the theory of multi - conductor transmission line, simulates and calculates the crosstalk among three conductor transmission lines on the PCB board using Matlab software. Firstly, the paper analyzes the voltage change of main loop of string on the impact of crosstalk voltage in the case of the terminal is matched load. Then we analyze the change of the crosstalk voltages in the terminal access resistance, terminal access reactance, terminal access impedance of the three different terminal load cases, and gets the appropriate conclusions. The results show that crosstalk voltage as the main string loop voltage increases, matched resistance value can reduce the crosstalk voltage, and capacitive load is more significant than inductive load on the impact of crosstalk voltage.

Key words: multi - conductor transmission line; termination impedance; crosstalk