基于模糊模式识别的支持向量机的回归预测方法

李庆国 , 陈守煜

(大连理工大学土木水利学院 , 辽宁 大连 116024)

摘要: 尝试把最近发展起来的支持向量机引入水文预测中 , 建立了支持向量机水文回归预测模型 , 为小样本情况下水文预测提供一种行之有效的可选择的方法。在此基础上 , 为了更好地处理水文系统中广泛存在的不确定、模糊信息 , 进一步把模糊模式识别理论引入支持向量机 , 提出一种模糊模式识别核函数。该核函数具有更明确合理的物理意义。冰凌预测实例表明了 SVM 水文回归预测方法及模糊模式识别核函数的有效性和可行性。

关 键 词: 水文预测; 支持向量机; 模糊模式识别

中图分类号: P338 **文献标识码**: A **文章编号**: 1001-6791(2005)05-0741-06

由于水文系统的复杂性 , 所涉及的参数通常是不确定的、模糊的 , 影响水文系统特性的各要素之间存在着非常复杂的非线性关系 , 难以用确定的数学模型描述 , 传统的方法存在不足。在预测领域 , 近年来一些数学方法得到较大的发展 , 如神经网络。神经网络方法存在一定的学习和泛化误差 , 同时影响人工神经网络拓扑结构的因素众多 , 且参数优选等问题使其应用推广存在一定的困难[1]。同时 , 神经网络方法是基于大样本的一种学习方法。

支持向量机(Support Vector Machine , SVM)方法建模不必知道因变量和自变量之间的关系 , 通过对样本的学习即可获得因变量和自变量之间非常复杂的映射关系; 与传统的神经网络不同 , SVM 应用的是结构风险最小化原则 , 而不是经验风险最小化原则 , 即 SVM 寻求的是一般误差上界的最小化而不是单纯训练误差的最小化; 同时 , 它是基于小样本的一种学习方法 , 不必知道太多的数据即可建模[2]。因此本文尝试把 SVM 引入水文预测中。

核函数一直是 SVM 研究的重点 , 不同的核函数会导致 SVM 的泛化能力有很大的不同 , 如何根据所给数据及实际问题选择合适的核函数具有重要意义。水文系统存在不确定性与模糊性[3] , 为了更好地处理这些不确定、模糊信息 , 本文提出了一种模糊模式识别核函数 , 从而把模糊模式识别理论与 SVM 结合起来。模糊模式识别核函数具有更加明确合理的物理意义。实例检验都表明了该核函数的有效性和可行性。

1 基于模糊模式识别的支持向量机的回归预测方法

首先 , 根据成因分析 , 确定水文预测因子。设 m 个预测因子组成预测因子集

$$x = \{x_1, x_2, \cdots, x_m\} \tag{1}$$

预测对象只有一个 , 即水文要素 , 设为 y。

有 n 个学习样本 , 组成样本集 S:

$$S = \{x, y\}_j \qquad j = 1, 2, \cdots, n \tag{2}$$

收稿日期: 2004-06-29; **修订日期**: 2004-09-30

基金项目: 水利部科技创新资助项目 (SCXC2005-01)

作者简介: 李庆国(1976 -) , 男 , 山东平邑人 , 博士研究生。主要从事水文水资源研究。

E-mail: chensydlut @sina. con

其中，$x \in R^m$，$y \in R$。

设存在一个映射 $f(x)$，并假设所有训练数据都可以在精度 ε 下无误差地用函数 $f(x)$ 拟合。

$$\begin{cases} y_j - f_j(x) \leq \varepsilon \\ f_j(x) - y_j \leq \varepsilon \end{cases} \qquad j = 1,2,\cdots,n \tag{3}$$

考虑到允许拟合误差的情况，引入松弛因子 $\xi_j \geq 0$ 和 $\xi_j^* \geq 0$，则式(3)变成

$$\begin{cases} y_j - f_j(x) \leq \varepsilon + \xi_j \\ f_j(x) - y_j \leq \varepsilon + \xi_j^* \\ \xi_j \geq 0 \\ \xi_j^* \geq 0 \end{cases} \tag{4}$$

首先，考虑 SVM 线性回归情况，然后再推广到复杂的非线性问题。线性回归为了计算学习样本的 SVM 和式(5)中的 C，需要解下面的以式(4)为约束的以下优化问题：

$$\mathrm{Min}\,\phi(w, \xi, \xi^*) = \frac{1}{2}(w^{\mathrm{T}}w) + C\left(\sum_{j=1}^{n}\xi_j + \sum_{j=1}^{n}\xi_j^*\right) \tag{5}$$

式(5)的第 1 项使样本到超平面的距离尽量大，从而提高泛化能力；第 2 项则使误差尽量小。为求解这个优化问题，引入拉格郎日函数，得到其对偶优化问题

$$\mathrm{Max}\,W(\alpha,\alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) + \sum_{i=1}^{n}y_i(\alpha_i^* - \alpha_i) - \sum_{i=1}^{n}\varepsilon(\alpha_i^* + \alpha_i) \tag{6}$$

其中

$$0 \leq \alpha_i, \alpha_i^* \leq C \qquad i = 1,2,\cdots,n \tag{7}$$

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{8}$$

对于非线性回归，与非线性分类相似，同样使用一非线性映射 ω 把数据从原空间 R^m 映射到一个高维特征空间 Ω，再在高维特征空间进行线性回归，其关键问题也是核函数 $K(x, x')$ 的采用。

$$K(x, y) = \Omega(x) \cdot \Omega(x') \tag{9}$$

其中，$K(x, y)$ 为核函数。

根据分析比较验证选取合适的核，从而得到回归预测模型

$$f(x) = \mathrm{sign}\left[\sum_{\text{支持向量}}\alpha_j y_j K(x, x_j) + b\right] \tag{10}$$

采用二次优化方法的内点算法，运用 Matlab 可以较为容易的求解上述优化问题[4~6]。

根据具体问题选择合适的核函数把线性回归推广到非线性回归非常重要。常用的核函数有：①多项式核函数：$K(x, x') = (x \cdot x' + 1)^d$，$d = 1, 2, \cdots, n$；②径向基函数(RBF)核函数：$K(x, x') = \exp\left[-\|x - x'\|^2/2\sigma^2\right]$；③Sigmoid 核函数：$K(x, x') = \tanh[b(x \cdot x') - c]$

从物理意义上，核函数 $K(x, x')$ 本质上是向量 x，x 之间的一种相似性度量[5]。当前常用的核函数如 RBF 核函数物理意义不十分明确。陈守煜教授在模糊优选理论的基础上建立了一种新的人工神经网络，网络的激励函数采用模糊优选及模糊模式识别模型[7,8]，网络在径流预报等水文水资源领域的应用中取得了较好的效果，证明了模型的优越性。而文献[6]从 SVM 与人工神经网络的关系方面论述了两者的等价性。因此，考虑上述原因，在模糊模式识别理论[3]的基础上，本文引入以下核函数：

$$K(x, x') = \begin{cases} 1 & x = x' \\ 1\Big/\sum_{k=1}^{n}\dfrac{\sum_{i=1}^{m}[w_i(x_i - x_i')]^2}{\sum_{i=1}^{m}[w_i(x_i - x_{ik}')]^2} & x \neq x' \end{cases} \tag{11}$$

其中，$x=(x_1, x_2, \cdots, x_m)$，$x_k=(x_{1k}, x_{2k}, \cdots, x_{mk})$，$k=1, 2, \cdots, n$。称上述核函数为模糊模式识别核函数(Fuzzy Recognition Kernel，FRK)。考虑 m 个预测因子对预测对象的影响程度不同，模型中引入预测因子的权向量

$$w=(w_1, w_2, \cdots, w_m) \tag{12}$$

模糊模式识别理论建立在相对隶属度概念[3]之上，因此模糊模式识别核函数更加全面的描述了向量 x，x' 之间相似程度。考虑到在预测领域实际应用中，各预测因子对预测对象的贡献程度往往是不相同的。模糊模式识别核函数引入了指标权向量概念，从而体现出这种差别。因而，与传统核函数相比，模糊模式识别核函数具有更加合理的物理意义。

式(11)和径向基函数(RBF)核函数类似，是 $K(x-x')$ 的形式。一维 $K(x)$ 函数可以表示为

$$K(x)=1\Big/\sum_{k=1}^{n}\frac{\|x\|^2}{\|x-x_k\|^2}=1\Big/\sum_{k=1}^{n}\frac{x^2}{(x-x_k)^2} \tag{13}$$

$K(x)\geq 0$，可以证明它满足文献[9]中关于支撑矢量核函数条件 2 的要求。

模糊模式识别核函数中预测因子的权向量 w 可以根据统计相关理论确定[10]。

$$w=\left(|\rho_1|\Big/\sum_{i=1}^{m}|\rho_i|, |\rho_2|\Big/\sum_{i=1}^{m}|\rho_i|, \cdots, |\rho_m|\Big/\sum_{i=1}^{m}|\rho_i|\right)=(w_1, w_2, \cdots, w_m) \tag{14}$$

其中 ρ_i 为样本与预测因子 i 间的相关系数，根据相关分析，有

$$\rho_i=\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)(y_j-\bar{y})\Big/\sqrt{\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\sum_{j=1}^{n}(y_j-\bar{y})^2} \tag{15}$$

式中 $\bar{x}_i$ 为预测因子 i 特征值的均值；$\bar{y}$ 为样本均值。

2　算例分析

为验证支持向量机的水文回归预测方法的有效性，以黄河内蒙古段开河、封河历时预测为例，编写程序，进行学习预测。影响凌汛的三大因素是热力因素、动力因素与河势因素。通过相关分析确定封河历时预测因子：x_1 累计负气温、x_2 水位及 x_3 流量；开河历时预测因子：x_1 累计正气温、x_2 水位、x_3 流量及 x_4 封河期最大冰厚[11]。

2.1　黄河内蒙古段三湖河口站开河历时预测

三湖河口站各开河历时预测因子相关系数为 x_1：-0.2409；x_2：0.1369；x_3：-0.1241；x_4：-0.4860。三湖河口站预报因子特征值与开河历时见表 1。

由式(14)及开河历时 y 与预测因子 (x_1, x_2, x_3, x_4) 之间的相关系数，计算得到 4 个预测因子的权向量：$w=(w_1, w_2, \cdots, w_4)=(0.24, 0.14, 0.13, 0.49)$。

表 1　三湖河口站预测因子特征值、开河历时实测值[11]及预测模型学习检验值

Table 1　Training sample data and forecasting value of ice flood at Sanhuhekou

序号	年度	累计正气温/ ℃	水位/ m	流量/ ($m^3 \cdot s^{-1}$)	封河期最大冰厚/ m	开河历时/ d	开河历时预测值/ d	相差天数/ d
1	1968-1969	46.2	1019.55	1240	0.83	37	35	2
2	1969-1970	46.7	1019.34	1150	0.92	26	34	8
3	1970-1971	20.5	1019.44	713	0.75	37	37	0
4	1971-1972	61.0	1019.97	1150	0.92	44	36	8
5	1972-1973	8.6	1019.49	816	0.68	42	38	4
6	1973-1974	16.1	1019.81	893	0.70	34	39	5
7	1974-1975	19.5	1018.77	1150	0.78	40	34	6
8	1975-1976	8.0	1019.31	760	0.70	37	37	0
9	1976-1977	59.2	1019.44	849	0.82	32	36	4
10	1977-1978	8.4	1019.61	774	0.57	42	41	1

续表 1

序号	年度	累计正气温/ ℃	水位 / m	流量 / ($m^3 \cdot s^{-1}$)	封河期最大冰厚/ m	开河历时 / d	开河历时预测值/ d	相差天数 / d
11	1978 - 1979	8.4	1 018.97	692	0.59	42	39	3
12	1979 - 1980	140.5	1 019.15	981	0.95	34	34	0
13	1980 - 1981	38.4	1 019.53	851	0.83	44	36	8
14	1981 - 1982	15.7	1 018.17	779	0.74	42	35	7
15	1982 - 1983	26.4	1 018.82	1 210	0.66	34	36	2
16	1983 - 1984	24.7	1 019.45	1 100	0.76	32	36	4
17	1984 - 1985	25.6	1 018.95	745	0.80	35	35	0
18	1985 - 1986	28.2	1 019.44	1 080	0.72	35	37	2
19	1986 - 1987	16.4	1 018.20	690	0.70	36	36	0
20	1987 - 1988	14.7	1 018.70	680	0.65	34	37	3
21	1988 - 1989	18.4	1 019.21	1170	0.55	39	39	0
22	1989 - 1990	14.5	1 019.22	900	0.60	49	39	10
23	1990 - 1991	45.1	1 018.43	820	0.50	39	39	0
24	1991 - 1992	18.6	1 018.59	680	0.53	39	39	0
25	1992 - 1993	62.3	1 020.20	1270	0.60	42	42	0
26	1993 - 1994	15.1	1 019.14	290	0.53	41	41	0
27	1994 - 1995	2.5	1 019.44	760	0.51	43	41	2
28	1995 - 1996	10.4	1 020.25	1300	0.63	36	40	4
29	1996 - 1997	49.5	1 019.43	1 340	0.60	42	39	3
30	1997 - 1998	63.8	1 019.96	1 850	0.58	40	39	1
31	1998 - 1999	3.5	1 019.81	820	0.48	44	41	3
32	1999 - 2000	37.6	1 019.94	1020	0.52	40	43	3
33	2000 - 2001	12.8	1 019.70	550	0.43	43	45	2

以上 33 年的观测数据，前 28 年作为学习样本建立预测模型，后 5 年作为检验样本对预测模型进行检验。首先对学习样本预测因子特征值进行规格化得到矩阵 R

$$R = \begin{bmatrix} 0.6833 & 0.6797 & 0.8696 & 0.5761 & 0.9558 & 0.9014 & 0.8768 \\ 0.4231 & 0.5625 & 0.6106 & 0.8654 & 0.6346 & 0.7885 & 0.2885 \\ 0.0845 & 0.2113 & 0.8268 & 0.2113 & 0.6817 & 0.5732 & 0.2113 \\ 0.2667 & 0.0667 & 0.4444 & 0.0667 & 0.6000 & 0.5556 & 0.3778 \\ 0.9601 & 0.5891 & 0.9572 & 0.9572 & 0.0000 & 0.7399 & 0.9043 \\ 0.5481 & 0.6106 & 0.6923 & 0.3846 & 0.4712 & 0.6538 & 0.0000 \\ 0.7606 & 0.6352 & 0.7408 & 0.8563 & 0.4493 & 0.6324 & 0.7338 \\ 0.5556 & 0.2889 & 0.8444 & 0.8000 & 0.0000 & 0.2667 & 0.4667 \\ 0.8268 & 0.8391 & 0.8326 & 0.8138 & 0.8993 & 0.9116 & 0.8848 \\ 0.3125 & 0.6154 & 0.3750 & 0.6106 & 0.0144 & 0.2548 & 0.5000 \\ 0.1268 & 0.2817 & 0.7817 & 0.3099 & 0.8592 & 0.8732 & 0.1831 \\ 0.6444 & 0.4222 & 0.3333 & 0.5111 & 0.5556 & 0.6667 & 0.8889 \\ 0.9130 & 0.6913 & 0.8833 & 0.5667 & 0.9087 & 1.0000 & 0.9428 \\ 0.5048 & 0.1250 & 0.2019 & 0.9760 & 0.4663 & 0.6106 & 1.0000 \\ 0.5634 & 0.6761 & 0.8732 & 0.0423 & 1.0000 & 0.7606 & 0.0000 \\ 0.7778 & 1.0000 & 0.9333 & 0.7778 & 0.9333 & 0.9778 & 0.7111 \end{bmatrix}$$

采用径向基函数核函数，C 取 100，ξ 取 0.001，通过对样本的学习，取得了较好的效果。最后，得到 20 个支持向量(VR)。建立预报模型，对学习样本的检验结果见表 1。把检验样本输入模型，得到 5 个检验样本预测结果见表 1，可以看到平均相差天数为 2.4 d。

同样，对模糊模式识别核函数下的 SVM 径流回归预报模型进行验证。C 取 100，ξ 取 0.001 对学习样本进

行学习，同样得到 20 个支持向量，建立预测模型，得到模型检验结果见表 2。

本文同时对 SVM 径流回归预测模型与模糊优选 BP 神经网络(Fuzzy Optimal Neural Network，FONN)预测方法[11]进行了分析比较。SVM 回归预测(RBF Kernel)和 SVM 回归预测(FRK)预测检验平均相差天数为 2.4 d，而 FONN 的预测检验平均相差天数为 3 d。

2.2　黄河内蒙段巴彦高勒站封河历时预测

以黄河内蒙段巴彦高勒站 1968 年至 2002 年的凌汛资料为例，采用本文提出的预测方法模型，对其封河日期进行预测。封河历时 y 与预测因子(x_1，x_2，x_3)之间的相关系数为 x_1:0.84；x_2:0.46；x_3:0.09。

同样，计算得到 3 个预测因子的权向量:$w = (w_1, w_2, w_3) = (0.60, 0.33, 0.07)$

巴彦高勒站预报因子特征值与封河历时相关资料略[11]。同样对于 34 年的观测数据，前 29 年的作为学习样本建立预测模型，后面的 5 年作为检验样本对预测模型进行检验。分别采用 RBF 核函数和模糊模式识别核函数(FRK)，对样本进行学习得到支持向量(VR)，建立预报模型。把检验样本输入模型，得到两种核函数下检验样本预测结果见表 3。

表 2　模糊模式识别核函数 SVM 预测结果检验

Table 2　Forecasting value of the test samples at Sanhuhekou

序号	年度	观测值 /($m^3 \cdot s^{-1}$)	预测值 /($m^3 \cdot s^{-1}$)	相差天数/d	平均相差天数/d
1	1996 - 1997	42	40	2	
2	1997 - 1998	40	38	2	
3	1998 - 1999	44	41	3	2.4
4	1999 - 2000	40	43	3	
5	2000 - 2001	43	45	2	

表 3　黄河内蒙段巴彦高勒站封河历时 SVM 预测结果检验

Table 3　Forecasting value of the test samples at Bayangaole

序号	年度	观测值 /($m^3 \cdot s^{-1}$)	预测值/($m^3 \cdot s^{-1}$)		相差天数/d		平均相差天数/d	
			RBF	FRK	RBF	FRK	RBF	FRK
1	1996 - 1997	41	43	40	2	1		
2	1997 - 1998	51	52	50	1	1		
3	1998 - 1999	54	50	51	4	3	2.5	2.2
4	1999 - 2000	54	58	58	4	4		
5	2000 - 2001	45	46	43	1	2		

对 SVM 径流回归预测模型与模糊优选 BP 神经网络预测方法[11]进行分析比较。SVM 回归预测(RBF Kernel)和 SVM 回归预测(FRK)预测检验平均相差天数分别为 2.5 d 与 2.2 d，而模糊优选 BP 神经网络(Fuzzy Optimal Neural Network，FONN)的预测检验平均相差天数为 2.7 d。

综上所述可以看到，本文给出的 SVM 回归预测模型方法取得较好的预测结果，满足水文预报的精度要求。模糊模式识别核函数取得了相对较优的预测结果，尽管两种核函数预测平均误差接近，但模糊模式识别核函数的 SVM 预报模型误差比较平均，表现出更好的泛化性能。

SVM 水文回归预测模型方法的拟合学习样本相对误差值处于 0.5 %～30 %之间。根据文献[11]，SVM 方法对学习样本拟合误差大于人工神经网络方法，这说明 SVM 拟合能力相对稍逊于人工神经网络。但 SVM 的预测误差与拟合误差相对于神经网络方法较为均衡，这也体现了 SVM 较强的泛化(预测)能力的特点与优点。还应指出，SVM 水文回归预测模型方法操作较为简单，样本学习计算十分快捷，体现出比神经网络等方法较为优越的一面。

3　结　　语

无论是 SVM 方法还是人工神经网络方法都属于集中参数预测方法[6]，这类方法对于参数变化不大的系统具有广泛的适用性，对于水文资料缺乏，系统结构研究尚欠清晰的条件下这类模型往往能获得可接受的模拟和预测效果。正是基于此，本文把支持向量机方法引入到水文预报预测中来。进一步，在模糊模式识别理论基础上，给出了 SVM 的模糊模式识别核函数。模型方法取得了较好的预测结果。SVM 还在不断的发展完善中，作为一个有力的模型工具，其在水文领域的应用还需要进一步研究与检验，本文工作是在这方面的有益尝试。

参考文献:

[1] 田盛丰, 黄厚宽. 基于支持向量机的数据库学习算法[J]. 计算机研究与发展, 2001, 37(1):17 - 22.
[2] Vapnik V N. The Nature of Statistical Learning Theory[M]. NY: Springer-Verlag, 1995.
[3] 陈守煜. 工程水文水资源系统模糊集分析理论与实践[M]. 大连: 大连理工大学出版社, 1998.
[4] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2001, 26(1):32 - 42.
[5] 吴 涛, 贺汉根, 贺明科. 基于插值的核函数构造[J]. 计算机学报, 2003, 26(8):990 - 996.
[6] 张 铃. 基于核函数的 SVM 机与三层前向神经网络的关系[J]. 计算机学报, 2002, 25(7):696 - 700.
[7] 陈守煜. 工程模糊集理论与应用[M]. 北京: 国防工业出版社, 1998.
[8] 陈守煜. 复杂水资源系统优化模糊识别理论与应用[M]. 长春: 吉林大学出版社, 2002.
[9] 张 莉, 周伟达, 焦李成. 尺度核函数支撑矢量机[J]. 电子学报, 2002, 30(4):527 - 529.
[10] 陈守煜. 中长期水文预报综合分析理论模式与方法[J]. 水利学报, 1997, 8:15 - 21.
[11] 陈守煜, 冀鸿兰. 冰凌预报模糊优选神经网络 BP 方法[J]. 水利学报, 2004(6):114 - 118.

A SVM regress forecasting method based on the fuzzy recognition theory

LI Qing-guo, CHEN Shou-yu

(Dalian University of Technology, Dalian 116024, China)

Abstract: This paper first introduces the support vector machine (SVM) regression forecasting method into hydrological forecasting. Further, based on the fuzzy recognition theory proposed by Prof. Chen Shou-yu, a new kind of kernel function of SVM is proposed in the paper. The kernel function has a more reasonable physical significance. At the end, the results of a study case show that the SVM regression hydrological forecasting method and the kernel function of fuzzy pattern recognition is reasonable and feasible.

Key words: hydrology forecasting; support vector machine; fuzzy pattern recognition