

基于 SVM 方法的小流域泥石流输沙量预测

王 涛,刘兴年,黄 尔

(四川大学水力学与山区河流开发保护国家重点实验室,四川 成都 610065)

摘要 :介绍 SVM 方法的基本原理及特性,建立基于 SVM 方法的小流域泥石流输沙量预测模型,用复相关分析法确定了影响流域输沙的 3 个主要因子:过程降雨量,前期降雨量,泥石流历时。对 12 组实测资料进行训练,训练值与实测值吻合较好;用训练好的模型对 5 组实测资料进行预测,预测效果优于神经网络模型。理论分析和实实验证均表明 SVM 方法可以获得整体最优效果。

关键词 :支持向量机;泥石流;输沙量预测

中图分类号 :P642.23

文献标识码 :A

文章编号 :1006-7647(2008)02-0001-03

Prediction of sediment discharge of debris flow in small watershed based on SVM analysis//WANG Tao, LIU Xing-nian, HUANG Er (*State Key Laboratory of Hydraulic and Mountain River Engineering, Sichuan University, Chengdu 610065, China*)

Abstract : The principle and characteristics of the support vector machine (SVM) method were described, and a predictive model for sediment discharge of debris flow in small watershed was developed. Three main influencing factors in this model, i. e. precipitation, ante-precipitation, and duration of debris flow, were determined based on multi-correlation analysis. The model was trained by 12 groups of experimental data, and the trained values were in good accordance with the experimental data. Then the trained model was applied to the prediction of sediment discharge based on 5 groups of experimental data, and the result of the prediction is superior to that of the neural network model. Theoretical analysis and a case study show that the SVM method is helpful to achieve an integral optimal effect.

Key words : support vector machine; debris flow; prediction of sediment discharge

以往,对泥石流输沙量的研究常采用单因子线性回归方法,这类方法虽然也能反映出某种统计特性,但不能刻画自然界复杂的非线性特性。随着系统科学、非线性科学和计算技术的高速发展,人们在泥石流输沙量预测中引进了人工神经网络模型(ANN^[1])等非线性模型,取得了一定的成果;可是 ANN 方法的最终解过于依赖初值,存在过学习现象,训练过程中存在局部极小点问题,且收敛速度比较慢。近年来,支持向量机(support vector machine,简称 SVM^[2])方法以其解决有限样本、非线性及高维识别中的优势,引起了人们的广泛关注。本文介绍 SVM 方法的基本原理和回归方法,建立了基于 SVM 方法的蒋家沟流域泥石流输沙量预测模型。比较表明, SVM 模型预测效果优于 ANN 模型,为泥石流输沙量预测提供了新途径。

1 SVM 方法

1.1 基本原理及特点

基于数据的机器学习, SVM 研究的是从观测数

据出发寻找规律,用这些规律对未来数据或无法观测的数据进行预测。以往机器学习理论的核心是经验风险最小化原则(简称 ERM)。事实上 ERM 并不能保证对未知数据的正确预测,相反会导致“过学习”问题。如果学习机器能力过强,能够无误差地适应任意的训练样本,这是因为它所采用的函数过于复杂,同时这也蕴含着预测的不可靠。Vapnik 提出了 VC 维的概念来标志函数集的复杂程度,并用这个概念给出了与分布无关的学习机器推广能力的界,即在给定的样本数量下实际风险将会在这个界的范围内,该界由两部分构成:经验风险和置信范围(以 VC 维为参数)。学习机器能力过强(VC 维很大),虽然能取得小的经验风险,但置信范围会很大;VC 维太小又会导致大的经验风险。结构风险最小化(简称 SRM)正是在这两者之间做出的折中。SRM 可以通过不同的方法实现。在实际的算法中,ANN 是通过选择一个适当构造的机器来保持固定的置信范围并最小化经验风险;SVM 则是保持经验风险固

定并最小化置信范围。

SVM方法根据有限的样本信息在模型的复杂性和学习能力之间寻求折中,以期获得最好的推广能力^[3]。其主要特点是:①SVM方法专门针对有限样本的情况,其目标是得到现有信息条件下的最优解,这个解不一定是样本无穷大时的最优解;②SVM是基于结构风险最小化原则的方法,理论上明显优于以往的经验风险最小化原则的学习方法;③SVM的学习算法是一种二次寻优方法,理论上可得到全局最优解;④算法将实际问题通过非线性变换映射到高维的特征空间,在高维空间中构造线性判别函数来实现原空间中的非线性判别函数,同时巧妙地解决了高维数问题,其算法复杂度与维数无关。因此SVM方法有着出色的学习性能,近年来已成为国际上数据挖掘如分类、回归等的流行方法。国际上已有很多关于SVM的研究报道,SVM在很多方面都有成功运用实例。本文仅介绍SVM用于回归的基本思路。

1.2 SVM的回归

已知训练样本为 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, $x \in R^m, y \in R^1$ 。 x 为输入空间向量, n 为训练样本数, y 为输出。寻求最优的估计函数 $y = f(x, w)$ (w 为参数),使得对 y 拟合或预测造成最小的损失。

SVM用于估计回归函数时有3个特点:①利用在高维空间中定义的线性回归函数集来估计回归;②利用线性最小化来实现回归估计,风险用Vapnik的 ϵ -不敏感损失函数来度量;③采用的风险系数是由经验误差和一个由结构风险最小化原则导出的正则化部分组成的。SVM估计函数如下:

$$\begin{cases} y = f(x, w) = \langle w, \varphi(x) \rangle + b \\ w, x \in R^n, b \in R \end{cases} \quad (1)$$

式中: $\varphi(x)$ 是从输入空间到高维特征空间的非线性映射; $\langle * , * \rangle$ 表示内积函数。 w 和 b 由最小化式(2)来估计:

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \frac{\lambda}{2} \|w\|^2 \quad (2)$$

式中: $R_{\text{reg}}[f]$ 为正则化风险; $R_{\text{emp}}[f]$ 为经验风险,由式(3)给出的 ϵ -不敏感损失函数来度量; $\frac{\lambda}{2} \|w\|^2$ 为正则化部分, λ 为正则化常数(一般取为1), $\|*\|$ 表示欧氏距离。 ϵ -不敏感损失函数描述为

$$L_{\epsilon}(y) = \begin{cases} 0 & |f(w, x) - y| < \epsilon \\ |f(w, x) - y| - \epsilon & \text{其他} \end{cases} \quad (3)$$

取 $R_{\text{reg}}[f] = \frac{1}{n} \sum_{i=1}^n L_{\epsilon}(y)$,即经验误差。那么式(2)

可进一步写为

$$R_{\text{reg}}[f] = C \frac{1}{n} \sum_{i=1}^n L_{\epsilon}(y) + \frac{1}{2} \|w\|^2 \quad (4)$$

式中: C 为误差惩罚因子,为一正常数,它决定着经验风险和正则化部分之间的平衡。亦即寻找 w 和 b ,需对式(4)进行最小化。

引入松弛变量 ξ_i, ξ_i^* ,式(4)的最小化等价于

$$\min(R_{\text{reg}}[f]) = \min\left(C \sum_{i=1}^n (\xi_i + \xi_i^*) + \frac{1}{2} \|w\|^2\right) \quad (5)$$

$$\text{s.t.} \begin{cases} y_i - \langle w, \varphi(x_i) \rangle - b \leq \epsilon + \xi_i \\ \langle w, \varphi(x_i) \rangle + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (6)$$

这是二次优化问题。利用对偶原理和拉格朗日乘子法,上述问题的对偶形式为

$$\max\left[-\frac{1}{2} \sum_{j=1}^n (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) \langle \varphi(x_i), \varphi(x_j) \rangle - \epsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i + \alpha_i^*)\right] \quad (7)$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \quad (8)$$

式中的 α_i 和 α_i^* 不会同时为非0,且仅有一部分 $(\alpha_i - \alpha_i^*)$ 不为0($\alpha_i - \alpha_i^*$)非0值对应的数据点就是支持向量。解二次优化,可得 $w = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \varphi(x_i)$,则支持向量机回归方程为 $f(x, w) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \langle \varphi(x_i), \varphi(x) \rangle + b$ 。

为避免高维特征空间中出现“维数灾”,SVM考虑用核函数取代内积函数,即 $K(x_i, x) = \langle \varphi(x_i), \varphi(x) \rangle = \sum_{i=1}^n \varphi(x_i) \varphi(x)$,这样同时解决了非线性映射 $\varphi(x)$ 的具体形式未知问题,得到回归方程: $f(x, w) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_i, x) + b$ 。

为避免高维特征空间中出现“维数灾”,SVM考虑用核函数取代内积函数,即 $K(x_i, x) = \langle \varphi(x_i), \varphi(x) \rangle = \sum_{i=1}^n \varphi(x_i) \varphi(x)$,这样同时解决了非线性映射 $\varphi(x)$ 的具体形式未知问题,得到回归方程: $f(x, w) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_i, x) + b$ 。

1.3 SVM模型的参数

在基于SVM的回归方程中,需要确定核函数的类型,此外,还需确定如下参数:①核函数所含的参数 σ 。本文采用常用的核函数,即径向基函数。②不敏感系数 ϵ 。 ϵ 影响支持向量的数量, ϵ 值越大则支持向量就越多,估计的函数精度也就较低。若要函数拟合、预测精度都比较高, ϵ 就必须在合适的范围内。③误差惩罚因子 C 。 C 取得越小,则对样本数据中超出 ϵ 管道的样本惩罚就较小,训练误差变大,系统的泛化能力变差; C 的取值大, $\frac{1}{2} \|w\|^2$

的权重就小,同样泛化能力下降。当上述参数确定后,求解式(7)中的 $(\alpha_i - \alpha_i^*)$ 就是求解一个二次规划(QP)问题,QP的求解方法比较成熟。

2 SVM在泥石流输沙量预测中的应用

整理了1995~1997年蒋家沟17场泥石流对应场次的输沙量、泥石流历时和降雨量较为完整的资料。泥石流发生必须具备3个条件:陡峻的地形、丰沛的降水和丰富的松散固体物质。因而坡降、降水和固体物质储量是3个非常重要的因子,它们也影响泥石流的输沙,另外输沙历时也是必须考虑的。而由于条件的限制,尚有许多因子一时无法定量化。蒋家沟为单一流域,其坡降在短时间内不会发生大的变化,可视为常量。降水是泥石流发生的激发因子,对输沙起着重要作用,用前期降雨量的大小来相对说明补给物质的饱和程度是一个简便易行的方法^[4]。

用复相关分析法^[5]对输沙量、过程降雨量、前期降雨量和泥石流历时进行复相关分析,得到表1,

表1 相关系数

变量	S	R	P	T
S	1	0.459	0.517	0.700
R	0.459	1	0.170	0.386
P	0.517	0.170	1	0.604
T	0.700	0.386	0.604	1

其中S,R,P,T分别代表输沙量、过程降雨量、前期降雨量和泥石流历时。由表1可见输沙量与过程降雨量和前期降雨量的相关系数较低,只有0.459和0.517,与泥石流历时的相关系数稍高,为0.700。据此得输沙量与过程降雨量、前期降雨量和泥石流历时的复相关系数r为0.87,因此选用过程降雨量、前

期降雨量和泥石流历时3个因子进行预测。

3 流域输沙量预测

本实例属于小样本情况。取前12组作为建模训练样本,后5组为预测检验样本。在建立SVM模型时,首先对数据进行(0,1)规一化处理。在拟合、预测精度目标控制下进行,通过组合选优,得到 $\sigma = 0.2, C = 50, \epsilon = 0.0075$ 。拟合、预测成果见表2。人工神经网络是20世纪80年代中后期迅速发展起来的一种机器学习方法,其应用已渗入到各个领域,BP神经网络模型是人工神经网络的模型之一,应用尤为广泛。本文同时用ANN方法中的BP神经网络模型进行了拟合、预测,结果见表2。

从表2可以看出:对于前12组用于学习的样本,ANN方法中的BP神经网络模型的预测效果较好,相对误差在0.15%以内,平均相对误差更是仅为0.027%,而SVM模型的模拟效果要差一些,平均相对误差为3.09%。对于后5个留待检验的样本,SVM模型的预测相对误差平均值为41%,最大相对误差不超过60%,明显优于ANN方法(预测值的相对误差平均值为67%,最大相对误差为179.3%)。出现这种情况的主要原因是BP神经网络方法是通过构造适当的模型来保持固定的置信范围并最小化经验风险,在学习过程中容易出现过拟合问题,其预测解对于初值的依赖性较强,得到的结果往往并非整体最优解,而SVM方法则是保持经验风险固定并最小化置信范围,考虑的是如何使整体最优。因此对于训练数据,SVM方法的精度较ANN方法稍差,但SVM方法的预测效果要优于ANN方法。比较17个样本的泥石流输沙量ANN与SVM预测结果的

表2 SVM模型和BP神经网络模型预测结果

序次	输沙量实测值/ 10^4 m^3	历时/ h	过程降雨/ mm	前期降雨/ mm	输沙量预测结果/ 10^4 m^3		相对误差绝对值/%	
					SVM模型	BP神经网络模型	SVM模型	BP神经网络模型
1	1103928	8.83	20.08	44.32	1083206.4	1103921	1.8771	0.0006
2	450191	3.23	32.18	16.33	435383.5	450190	3.2892	0.0002
3	887437	8.90	17.83	38.80	886623.4	887445	0.0917	-0.0009
4	652204	7.29	12.97	32.50	631489.4	652198	3.1761	0.0009
5	370696	3.97	10.17	20.77	349981.4	370720	5.5880	0.0065
6	135248	2.67	13.65	7.45	134434.4	135187	0.6016	0.0451
7	375851	10.37	17.50	20.87	375037.4	375858	0.2165	0.0019
8	146937	2.37	6.03	13.25	126222.4	146883	14.0976	0.0368
9	74575	2.33	8.77	4.77	73761.4	74683	1.0910	0.1448
10	336956	3.17	9.42	28.93	320097.9	336957	5.0031	0.0003
11	518534	7.58	15.22	28.55	517720.4	518508	0.1569	0.0050
12	42536	1.50	5.18	21.32	41722.4	42573	1.9128	0.0870
13*	972055	4.20	18.00	18.20	393233.4	336816	59.5462	65.3501
14*	296873	2.78	4.35	8.23	187793.8	422547	36.7427	42.3326
15*	343022	3.95	14.33	10.07	254494.5	268033	25.8081	21.8613
16*	278177	4.33	12.85	20.55	399846.4	348813	43.7381	25.3925
17*	71367	2.20	4.65	18.53	99351.2	199336	39.2116	179.3112

注:带*的为预测组数据,未带*的是训练组数据。

从表 2 可以看出 经验公式形式较简单 但最大相对误差最大 而且不便于记忆 ;孙建公式误差最小 但公式复杂 而且是分段函数表示 ;王正中公式最大相对误差较小 但公式形式还是不够简单 本文公式形式最为简单 容易记忆 最大相对误差小于 0.86%。因此 笔者认为本文公式是计算圆形断面临界水深的最佳公式。

4 应用举例

例 1 :某水利工程的引水隧洞设计泄流量为 $500 \text{ m}^3/\text{s}$,拟用圆形断面 初设直径为 15 m ,试计算洞内的临界水深值。

解 :由式(20)计算得 $k = 0.0364$,由式(19)计算得 $h_k = 6.4341 \text{ m}$ 。

例 2 :以文献 [4] 为例 ,某引水式电站输水隧洞为圆形断面 流量为 $8 \text{ m}^3/\text{s}$,洞径为 3 m ,试计算洞内的临界水深。

解 :由式(20)计算得 $k = 0.0291$,由式(19)计算得 $h_k = 1.2152 \text{ m}$ 。

用经验公式、孙建公式和王正中公式分别计算例 1 和例 2 ,计算结果列于表 3 中。

从 2 个算例的计算过程和表 3 中的误差比较可以看出 用本文近似计算公式求解圆形断面的临界

水深不仅求解过程简单 而且计算精度高 能够满足工程实际要求。

表 3 不同计算方法误差比较

计算条件	公式名称	临界水深 计算值/m	临界水深 精确解/m	相对 误差/%
例 1 $Q = 500 \text{ m}^3/\text{s}$ $d = 15 \text{ m}$	经验公式	6.5191	6.4275	1.425
	孙建公式	6.4267	6.4275	-0.012
	王正中公式	6.4425	6.4275	0.233
	本文公式	6.4341	6.4275	0.103
例 2 $Q = 8 \text{ m}^3/\text{s}$ $d = 3 \text{ m}$	经验公式	1.2202	1.2129	0.601
	孙建公式	1.2127	1.2129	-0.016
	王正中公式	1.2169	1.2129	0.330
	本文公式	1.2152	1.2129	0.200

参考文献 :

- [1] 武汉水利电力学院. 水力计算手册 [M]. 北京 :水利电力出版社 ,1983.
- [2] 孙建 李宇. 圆形和 U 形断面明渠临界水深的直接计算公式 [J]. 陕西水力发电 ,1996 ,1(3) :39-42.
- [3] 张文卓. 圆形断面临界水深计算 [J]. 四川水力发电 ,2002 ,2(1) :15-17.
- [4] 吕宏兴 把多铎 宋松柏. 无压流圆形断面水力计算的迭代法 [J]. 长江科学院院报 ,2003 ,20(5) :15-17.
- [5] 王正中 陈涛 万斌 等. 圆形断面临界水深的近似计算公式 [J]. 长江科学院院报 ,2004 ,21(2) :1-2.
- [6] 吴持恭. 水力学 [M]. 北京 :高等教育出版社 ,1979.

(收稿日期 2007-05-15 编辑 高建群)

(上接第 3 页)

相对误差 可以看出 SVM 方法较 ANN 方法更为稳定、可靠 表明基于 SVM 方法的泥石流输沙量预测方法在数据拟合方面有良好的性质 其数据预测精度在泥石流输沙量研究中也是可以接受的 该方法值得进一步深入研究。下一步的研究应考虑以下 2 种情况 :① 鉴于控制泥石流输沙的因子很多 可在深入分析的基础上适当增加影响因子 ;② 本文 SVM 模型预测时所选择的参数可能并不是最优的 采用其他优化算法可能会得到更好的效果。

4 结 语

SVM 方法是一种基于结构风险最小的小样本学习方法 可以较好地解决以往 BP 神经网络模型非线性方法容易出现的小样本、过学习、局部最小等难题。本文将 SVM 方法引进流域泥石流输沙量预测 为泥石流输沙量预测研究提供了一条新的思路和途径 具有较大的实用价值。采用复相关分析方法 选定过程降水、前期降水和泥石流历时为流域泥

石流输沙的主要影响因子。实例分析和对比研究表明 SVM 方法的整体预测效果要优于 BP 神经网络模型 用于泥石流输沙量预测有较好的前景 可进一步探讨泥石流输沙因子和 SVM 模型参数的选择对预测效果的影响。

参考文献 :

- [1] 金菊良 丁晶. 水资源系统工程 [M]. 成都 :四川科学技术出版社 ,2002.
- [2] VAPNIK V. 统计学习理论 [M]. 许建华 张学工译. 北京 :电子工业出版社 ,2004.
- [3] VAPNIK V. 统计学理论的本质 [M]. 张学工译. 北京 :清华大学出版社 ,2000.
- [4] 陈景武. 云南东川蒋家沟泥石流爆发与暴雨关系的初步分析 [C] // 中国科学院水利部成都山地灾害与环境研究所. 全国泥石流学术会议论文集. 成都 [出版者不详] ,1980 :93-99.
- [5] 张超 杨炳根. 计算地理学基础 [M]. 北京 :高等教育出版社 ,1993.

(收稿日期 2007-01-16 编辑 高建群)