

基于改进支持向量回归的岩溶天窗水位预测模型

王 恩,束龙仓,刘丽红,黄币娟

(河海大学水文水资源学院,江苏 南京 210098)

摘要 利用主成分分析对影响变量进行特征提取,选择通用性较强的径向基核函数,应用遗传算法对影响支持向量回归模型的 2 个重要参数惩罚因子 C 和核函数参数 γ 进行优化,建立了基于遗传算法优化的支持向量回归模型,并以提取特征作为模型输入应用于后寨地下河流域平山天窗水位预测.预测结果表明,与传统偏最小二乘回归模型相比,优化后的模型具有更高的精度和更强的泛化能力.

关键词 支持向量回归 主成分分析 遗传算法 地下河 天窗水位预测

中图分类号 P641.2 **文献标志码** A **文章编号** 1000-1980(2011)01-0020-04

地下河是我国南方碳酸盐岩地区最重要的岩溶现象之一,也是地下水赋存的一种特有形式,已经成为南方多省岩溶地区主要供水水源^[1-3].天窗作为地下河发育地区联系地表水和地下水的中间纽带,对岩溶水模型研究起着相当重要的作用.然而,天窗水位变化物理机制十分复杂,国内外研究不深^[4],很难用准确的数学模型刻画,通常采用基于传统大样本理论的统计模型,如多元无偏线性回归、偏最小二乘回归、神经网络等方法进行建模预测^[5-6].但是,多元无偏线性回归和偏最小二乘回归建模时,要求因变量和自变量间存在较为显著的线性关系,而实际岩溶地区存在多重介质且高度各向异性和非均质性,很难满足线性回归建模要求,相比神经网络等传统非线性建模方法,虽可以较好反映变量间非线性问题,但受其结构及人为设计因素影响,建模过程中较容易陷入局部最优值,且初始条件和模型参数选择对模型影响较大^[7].

支持向量回归(support vector regression,SVR)算法是 Vapnik 于 1998 年提出的^[8],其建立在统计学习理论的 VC 维理论和结构风险最小原理基础上.其中 VC 维是表示函数集的复杂程度和学习能力的重要指标,而结构风险最小指寻找介于给定样本逼近精度与逼近函数之间的折中^[9-10].支持向量回归的精髓在于惩罚因子 C 和核函数的引入,而对核函数的选择,现在还缺乏指导原则,但经验表明,径向基核函数(RBF)具有广泛的适用性.支持向量回归算法参数对支持向量回归的性能有着重要影响.本文利用遗传算法(genetic algorithm,GA)对重要参数进行优选,利用主成分分析(principal component analysis,PCA)提取对方差解释强的信息作为模型输入,建立遗传算法优化支持向量回归耦合(PCA-GA-SVR)模型,以期提高在小样本条件下回归模型的精度与泛化能力,并应用于贵州省普定县后寨地下河流域平山天窗水位预测.

1 PCA-GA-SVR 模型原理

1.1 主成分分析原理

主成分分析能在保持原始数据信息丢失最少的情况下,对高维变量空间进行降维处理,利用少数解释性强的综合变量替代原始高维变量,在满足建模精度要求基础上,提高效率.其简要步骤如下:

设原始变量 $x_1, x_2, \dots, x_p$,主成分分析后得到的综合变量为 $t_1, t_2, \dots, t_k (k < p)$,称 $t_1, t_2, \dots, t_k$ 空间为 k 维主超平面.其中,第一主成分 t_1 对应于数据变异最大方向,其贡献率为 e_1 ,而 $t_2, t_3, \dots, t_k$ 的贡献率依次递减.

a. 原始数据标准化:

$$x'_{ij} = (x_{ij} - \bar{x}_i) / \sigma_i \quad (i = 1, 2, \dots, p; j = 1, 2, \dots, n) \quad (1)$$

式中: n ——样本数; x_{ij} ——第 i 个变量第 j 个样本的值; $\bar{x}_i, \sigma_i$ ——第 i 个指标的样本均值和标准差.

b. 计算相关系数矩阵 R :

$$R = (r_{ij})_{p \times n} \tag{2}$$

其中

$$r_{ij} = \frac{1}{n} \sum_{k=1}^n (x'_{ki} - \bar{t}_i)(x'_{kj} - \bar{t}_j) / \sigma_i \sigma_j \tag{3}$$

c. 计算 R 的特征值与特征向量. 根据方程 $|\lambda E - R| = 0$ 计算特征根 λ , 并从大到小排列 $\lambda_1, \lambda_2, \dots, \lambda_p$, 对特征根进行标准正交化, 得对应的特征向量 $u_1, u_2, \dots, u_p$.

d. 根据公式(4)计算贡献率 e_i 和累计贡献率 E_m .

$$e_i = \lambda_i / \sum_{i=1}^p \lambda_i \quad E_m = \sum_{i=1}^m \lambda_i / \sum_{i=1}^p \lambda_i \tag{4}$$

e. 根据公式(5), 判断选取主成分个数, 通常取较小的 m, λ 取 0.8.

$$x_m = E_m \geq \lambda \tag{5}$$

1.2 GA-SVR 模型

遗传算法是一种模仿生物进化过程的随机算法^[11-12]. 与传统搜索算法不同, 遗传算法是从代表问题可能潜在解集开始搜索过程, 该解集被称为种群. 种群中的每个带有染色体特征的个体是问题的一个解. 这些染色体在后续迭代中不断进化, 该过程称为遗传. 在每一代中通过相应的适应度值来评价染色体的好坏, 生成的下一代染色体称为子代. 子代主要由前一代染色体通过选择、交叉和变异运算形成. 在子代形成过程中, 根据适应度值的大小选择部分后代. 适应度值高的染色体被选中的概率相对较高, 这样经过若干代之后, 算法收敛于最好的染色体, 它很可能就是问题的最优解.

支持向量回归的基本思想是通过某种特定的非线性映射 φ , 利用满足 Mercer 条件的对称正实数核函数 $k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ 将输入数据映射到一个高维特征空间, 并在这个高维特征空间求取最优线性回归函数. 径向基核函数^[13-14]为

$$k(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2) \quad (\gamma > 0) \tag{6}$$

式中 γ 为核函数参数.

设数据集 $G = \{(x_i, y_i), \dots, (x_n, y_n)\}$, 其中 x_i 是 n 维输入向量, y_i 是 1 维的目标向量, 标准的支持向量回归为

$$f(x) = \langle \omega, \varphi(x) \rangle + b \tag{7}$$

式中 ω 和 b 表示回归因子, 可通过最小化正则化风险泛函获得

$$R(\omega) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n L(f(x_i), y_i) \tag{8}$$

式中 C 称为惩罚因子, L 代表损失函数, 本文选取一次不敏感损失函数:

$$L_\epsilon(f(x), y) = \max(|f(x) - y| - \epsilon, 0) \tag{9}$$

最小化 $R(\omega)$ 可得 $\omega = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \varphi(x_i)$, 其中 α_i 和 α_i^* 是最小化 $R(\omega)$ 对偶问题的解. 将 ω 代入式(8)得到回归函数:

$$f(x) = \sum_{i=1}^n (-\alpha_i + \alpha_i^*) k(x_i, x) + b \tag{10}$$

1.3 PCA-GA-SVR 模型算法步骤

利用主成分分析对天窗水位变化影响因素进行特征提取, 采用遗传算法对支持向量回归模型中惩罚参数 C 和核函数参数 γ 进行优化, 采用 31 位二进制编码, 种群中的个体解码成一组参数 C 及 γ 的集合, 利用模型的均方误差的倒数作为适应度函数. 算法具体步骤如下: (a) 利用主成分分析对影响岩溶天窗水位变化因素进行特征提取; (b) 设置最大进化代数、种群规模、交叉概率、变异概率; (c) 随机产生二进制编码的初始种群 P , 其个体代表一组参数 C 及 γ ; (d) 建立支持向量回归模型, 计算个体适应度值; (e) 对父代种群进行选择、交叉和变异操作产生子代种群; (f) 将子代种群中的个体解码后建模训练, 计算子代种群中个体的适应度值, 若达到最大进化代数, 转(g), 否则转(d); (g) 选出最优个体解码, 该个体即代表遗传算法优化后的最佳的参数 C 及 γ 值, 重新训练支持向量回归模型并进行预测.

2 模型应用

以下用 PCA-GA-SVR 模型预测贵州省普定县后寨地下河流域平山天窗水位变化^[4]. 选用 1990 年 5 月 24 日至 6

月 27 日降水-水位数据为拟合数据(图 1(a)),以 1995 年 6 月 24 日至 7 月 17 日降水-水位数据为检验数据(图 1(b)).图 1 中每天对应 2 个实测水位,任一实测水位对应的降水量为从该水位实测时间开始累计 12 h 的降水量.

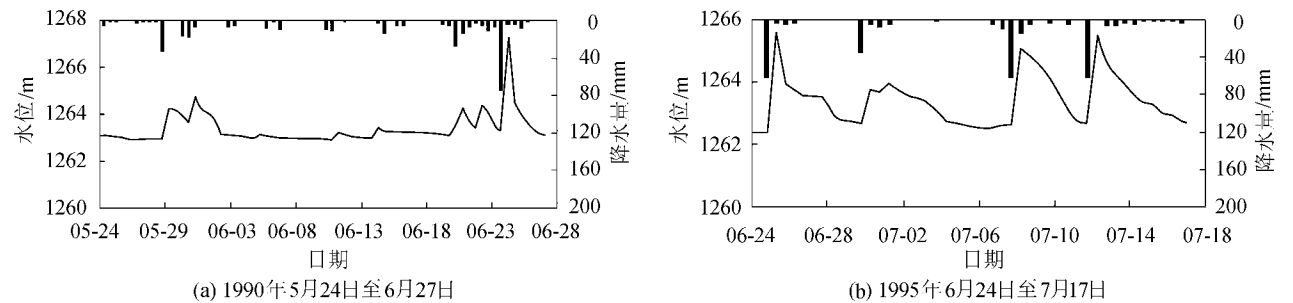


图 1 降水水位变化过程

Fig. 1 Variation of water levels due to precipitation

2.1 数据分析

拟合数据水位变化时间序列自相关系数如图 2 所示,显著水平为 0.05,从图 2 水位前 4 阶自相关系数变化可知,前 4 阶与水位都有一定相关性,且前 4 阶自相关系数值差异较小,因此有必要考虑前 4 阶水位 $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}$ 对当前水位 y_t 的影响. 根据文献[4]中降水-水位互相关系数结论,选取对水位影响相对较大的前 3 阶降水序列 $p_{t-1}, p_{t-2}, p_{t-3}$. 因此,建立回归方程表达式如下

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, p_{t-1}, p_{t-2}, p_{t-3}) \tag{11}$$

2.2 预测结果

利用主成分分析提取对方差解释较强的成分,经计算前 4 个主成分累计解释方差达到 91%,满足主成分选取条件. 利用遗传算法优化支持向量回归模型中的惩罚参数 C 和核函数参数 γ ,遗传算法的参数设置值及优化后的 C, γ 值见表 1,模型拟合均方误差为 0.02,确定系数 $R^2 = 0.93$. 拟合结果如图 3 所示,拟合均方误差为 0.065. 与传统偏最小二乘模型[4]相比,精度和泛化能力都有一定的提高.

表 1 遗传算法设置参数值及 C 和 γ 优化值

Table 1 Values of parameters of genetic algorithm and optimal values of C and γ

种群规模	C 取值范围	γ 取值范围	最大进化代数	交叉概率	变异概率	C 优化值	γ 优化值
20	[0.1,100]	[0.01,1000]	150	0.5	0.02	94.35	0.41

3 结 论

- a. 岩溶地下河系统中,天窗水位动态变化受诸多非线性因素影响,系统建模时,利用主成分分析对变量进行主成分提取,从而克服原变量间复相关性对建模的影响,提高模型精度.
- b. 应用主成分分析进行主成分提取,利用遗传算法对支持向量回归模型重要参数优化后建立的模型相对传统的偏最小二乘模型具有更好的精度和更强的泛化能力.
- c. 支持向量回归模型的建立依赖于自身参数和核函数的选取,针对不同的应用领域确定参数和选取合适的核函数是支持向量回归发展的重要方向.

参考文献：

[1] 杨立铮. 中国南方地下河分布特征[J]. 中国岩溶,1985,16(1) 92-101. (YANG Lu-zheng. Distribution of subterranean rivers in south China[J]. Carsologica Sinica,1985,16(1) 92-101. (in Chinese))

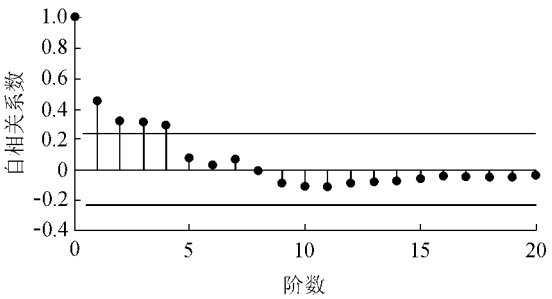


图 2 水位自相关系数

Fig. 2 Autocorrelation function of water levels

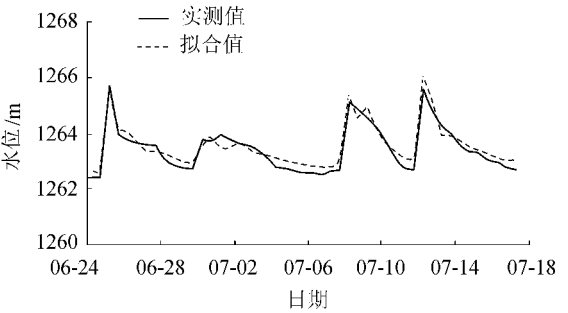


图 3 水位拟合结果

Fig. 3 Fitting results of observed water levels

- [2] 袁道先. 对南方岩溶石山地区地下水资源及生态环境地质调查的一些意见[J]. 中国岩溶, 2000, 19(2): 103-108. (YUAN Dao-xian. Aspects on the new round land and resources survey in karst rock desertification areas of South China[J]. Carsologica Sinica, 2000, 19(2): 103-108. (in Chinese))
- [3] 章程, 袁道先. 典型岩溶地下河流域水质变化与土地利用的关系: 以贵州普定后寨地下河流域为例[J]. 水土保持学报, 2004, 18(5): 134-137. (ZHANG Cheng, YUAN Dao-xian. Hydrochemical variation of typical karst subterranean stream basin and its relationship with landuse change: a case study of Houzhai Subterranean Stream Basin, Puding County, Guizhou Province[J]. Journal of Soil Water Conservation, 2004, 18(5): 134-137. (in Chinese))
- [4] 束龙仓, 董贵明, 陶玉飞, 等. 地下河天窗水位变化分析及预测[J]. 水利学报, 2009, 40(5): 529-534. (SHU Long-cang, DONG Gui-ming, TAO Yu-fei. Forecast and analysis on water level fluctuation in sinkhole of underground rivers[J]. Journal of Hydraulic Engineering, 2009, 40(5): 529-534. (in Chinese))
- [5] 王惠文. 偏最小二乘回归方法及应用[M]. 北京: 国防科技出版社, 1999.
- [6] 罗批, 郭继昌, 李铿, 等. 基于偏最小二乘回归建模的探讨[J]. 天津大学学报, 2002, 35(6): 783-786. (LUO Pi, GUO Ji, LI Qiang, et al. Modeling construction based on partial least squares regression[J]. Journal of Tianjin University: Science and Technology, 2002, 35(6): 783-786. (in Chinese))
- [7] 姚林, 阳建宏, 何飞, 等. 基于核偏最小二乘的锌层重量预测模型[J]. 控制工程, 2008, 15(2): 154-157. (YAO Lin, YANG Jian-hong, HE Fei. Forecasting model for zinc coating weights based on kernel partial least squares[J]. Control Engineering of China, 2008, 15(2): 154-157. (in Chinese))
- [8] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.
- [9] CHAPELLE O, VAPNIK V, BOUSQUET O. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1): 131-159.
- [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. (ZHANG Hao-ran, HAN Zhang-zhi, LI Chang-gang. Support vector machine based nonlinear model predictive control[J]. Systems Engineering and Electronics, 2003, 25(3): 330-335. (in Chinese))
- [11] 肖伟平, 何宏. 基于遗传算法的数据挖掘方法及应用[J]. 湖南科技大学学报: 自然科学版, 2009, 24(3): 82-86. (XIAO Wei-ping, He Hong. Data mining algorithm based on genetic algorithm and its application[J]. Journal of Hunan University of Science & Technology: Natural Science Edition, 2009, 24(3): 82-86. (in Chinese))
- [12] 罗固源, 郑剑锋, 许晓毅, 等. 基于遗传算法的次级河流回水段水质模型多参数识别[J]. 中国环境科学, 2009, 29(9): 962-966. (LUO Gu-yuan, ZHENG Jian-feng, XU Xiao-yi, et al. Parameters identification of water quality model in branch backwater reach based on genetic algorithm[J]. China Environmental Science, 2009, 29(9): 962-966. (in Chinese))
- [13] 熊建秋, 李泳. 基于支持向量机的水流挟沙力预测研究[J]. 水利学报, 2005, 36(10): 1171-1175. (XIONG Jian-qi, LI Yong, SUN Jing-sheng. Sediment-carrying capacity forecasting based on support vector machine[J]. Journal of Hydraulic Engineering, 2005, 36(10): 1171-1175. (in Chinese))
- [14] POURBASHEER E, RIAHI S, GANJALI M R, et al. Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity[J]. European Journal of Medicinal Chemistry, 2009, 44(12): 5023-5028.

Prediction model for water level of sinkholes in karst region based on improved SVM

WANG En, SHU Long-cang, LIU Li-hong, HUANG Bi-juan

(College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract: The principal component analysis was employed to extract features of independent variables. The universal RBF kernel function was selected. The genetic algorithm (GA) was applied to the optimization of two important parameters of penalty factor C and kernel parameter γ for the support vector regression model. Accordingly, a GA-based support vector regression model was established. The extracted features of independent variables were regarded as the model input to predict the water level of Pingshan sinkhole in Houzhai underground watershed. The predicted results show that, compared with the traditional PLS model, the proposed coupling model has higher precision and strong generalization ability.

Key words: support vector regression; principal component analysis; genetic algorithm; underground river; water level prediction of sinkhole