基于改进遗传算法-偏最小二乘回归的大坝变形监测模型

杨 杰[1,2],杨 丽[1],李建伟[1],包天栋[1]

(1 西安理工大学 水利水电学院,陕西 西安 710048;2 水资源与水电工程科学国家重点实验室,湖北 武汉 430072)

[摘 要] 【目的】针对常规大坝变形监测回归模型中存在的因子多重相关性干扰和模型拟合效果欠佳问题,进行偏回归模型优化方法研究。【方法】将改进的遗传算法引入大坝变形监测偏回归模型,利用遗传算法强大的自适应全局优化搜索功能,对偏最小二乘回归模型进行优化,建立了基于改进遗传算法-偏最小二乘回归的大坝变形监测模型。【结果】工程实例研究与对比分析表明,改进遗传算法-偏最小二乘回归模型在一定程度上改善了原偏回归模型存在的拟合效果不佳的问题。【结论】改进遗传算法-偏最小二乘回归模型具有较好的拟合与预测能力,有较强的工程实用性。

[关键词] 改进遗传算法;偏最小二乘法;回归模型;大坝变形监测

[中图分类号] TV698.1 [文献标识码] A [文章编号] 1671-9387(2010)02-0206-05

Monitoring model for dam deformation based on partial least-squared regression and improved genetic algorithm

YANG Jie[1,2], YANG Li[1], LI Jian-wei[1], BAO Tian-dong[1]

(1 *Institute of Water Resources and Hydro-electric Engineering, Xi'an University of Technology, Xi'an, Shaanxi* 710048, *China*;
2 *State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan, Hubei* 430072, *China*)

Abstract: 【Objective】 Aiming at the problem of the under-fitting and multicollinearity in general dam safety monitoring model, an optimization of partial regression was put forward. 【Method】 The improved genetic algorithm was introduced in the modeling of partial least-squared regression in dam deformation monitoring to optimize the partial least-squares regression model by using its powerful adaptive global optimization search function, and the monitoring model established for dam deformation based on partial least-squared regression and improved genetic algorithm. 【Result】 Project instance studies and correlation analysis show that the model based on partial least-squared regression and improved genetic algorithm improved the under-fitting phenomenon of the original model of partial regression to some extent. 【Conclusion】 The model based on partial least-squared regression and improved genetic algorithm has good simulating effect and forecasting precision and strong engineering practicability.

Key words: improved genetic algorithm; partial least-squared; regression model; dam deformation monitoring

常规统计模型因简单直观、易于求解而得以普遍应用,目前已形成了一套完整的建模理论和方法

* [收稿日期] 2009-08-15
[基金项目] 国家自然科学基金项目(50779051);水资源与水电工程科学国家重点实验室开放基金项目(2007B037);陕西省教育厅专项科研计划项目(07JK354);西安理工大学科学研究基金项目(106-210509)
[作者简介] 杨 杰(1971—),男,四川大邑人,副教授,博士,主要从事水工结构、岩土工程及其安全监测研究。
E-mail: yjie999@xaut.edu.cn

体系。偏最小二乘法回归(Partial least-squares regression，PLSR)[1]是对多元线性回归模型的一种扩展，主要适用于多因变量对多自变量的线性回归建模，可以有效地解决用普通多元回归无法解决的许多问题，如克服变量多重相关性以及在样本容量小于变量个数的情况下进行回归建模预测分析等。PLSR 将模型式的方法和认识性的方法有机地结合起来，在一个算法下，可以同时实现多元线性回归、主成分分析以及典型相关性分析。

大坝的整体工作性能，主要反映在变形、渗流、应力变化等方面，其中变形变化状况尤为直观可靠，是评价大坝安全状况的重要依据。大坝的变形也是目前坝工专家公认的、能够说明大坝实际状况的惟一精确可靠的指标。因此，开展大坝变形监控及其应用研究，具有重要的理论价值和现实意义。以前，在大坝安全监测数据的处理分析中，国内外广泛采用统计学方法，近年来 PLSR 模型在大坝安全监控领域得到了较好的应用[2-6]，但由于影响大坝安全的因素众多且复杂，使得 PLSR 模型在安全监测资料分析中存在拟合能力欠佳的问题[7]，尤其在对长时间段原型观测资料进行拟合时，由于大坝经历的历史情况比较复杂，得到的拟合方程复相关系数 R 值普遍不高，从而影响了偏回归分析建模的准确度。因此，对 PLSR 模型进行优化改进，对于更好地实现大坝的安全监控具有重要意义。

遗传算法[8](Genetic algorithm，GA)是通过模拟自然进化过程搜索最优解的方法，其最大优点就是全局优化搜索能力强，无需对所要优选的决策变量赋于初值，算法本身会自动从其上下限范围内随机选取一组初值，通过遗传算子作用，按遗传选择策略优选出参数的全局最优解或相对全局最优解。

本研究利用改进遗传算法强大的自适应全局优化概率型搜索功能，对 PLSR 建模方法进行优化，建立基于改进遗传算法-偏最小二乘的回归模型(Genetic algorithm-partial least-squares regression，GA-PLSR 模型)，以期解决常规回归模型中存在的因子多重相关性干扰和模型拟合效果不佳的问题，进一步提高模型的拟合和预测精度。

1　PLSR 的基本思想

设有 q 个因变量$\{y_1,\cdots,y_q\}$和 p 个自变量$\{y_1,\cdots,y_p\}$。为了研究因变量与自变量的统计关系，采集 n 个样本点，构成自变量与因变量的数据表 $X=\{x_1,\cdots,x_p\}_{n\times p}$和$Y=\{y_1,\cdots,y_q\}_{n\times q}$。偏最小二乘分别在 X 和 Y 中提取主成分 t_1 和 u_1，即 t_1 是 $x_1,\cdots,x_p$ 的线性组合，u_1 是 $y_1,\cdots,y_q$ 的线性组合。在提取这 2 种成分时，为了回归分析的需要，t_1 和 u_1 应尽可能多地携带其各自数据表中的变异信息，且 t_1 与 u_1 的相关程度能够达到最大。

提取第 1 主成分 t_1 和 u_1 后，分别进行 X 对 t_1 的回归和 Y 对 u_1 的回归，然后利用 X 被 t_1 解释后的残余信息以及 Y 被 u_1 解释后的残余信息进行第 2 轮主成分的提取，得到第 2 主成分 t_2、u_2，继续进行 Y 和 X 对 u_2、t_2 的回归，如此反复，直至达到一个较满意的精度为止。若最终对 X 提取 m 个成分 $t_1,t_2\cdots,t_m$，则通过进行 y_k 对 $t_1,t_2,\cdots,t_m$ 的回归，转化为 y_k 关于原变量 $x_1,\cdots,x_p$ 的回归方程，其中 $k=1,2,\cdots,q$，这样就完成了偏最小二乘回归建模。

从其算法特点和实际应用来看，经典的 PLSR 也存在不足之处[7,9-11]，如其提取的主成分可能只对自变量有很强的综合能力，而与因变量的相关程度并非最大；对含有较多自变量的模型的拟合效果欠佳，且不能对自变量进行筛选和识别等。GA 具有鲁棒性强、易于并行化、易于与别的技术相融合等特点。为此，将改进 GA 引入偏回归模型，进行偏回归优化方法研究，建立基于改进遗传算法-偏最小二乘回归的大坝变形监测模型。

2　改进遗传算法-偏最小二乘回归模型的建立

2.1　基本原理

相对于传统遗传算法[8,12]，改进遗传算法在以下几个方面有所改进：①采用浮点编码(基因)表示优化方程中的决策变量；②采用基于排序的选择策略和最优保存策略；③变异操作采用均匀变异操作；④浮点编码的交叉算子采用算术交叉算子。改进的遗传算法提高了遗传算法的效率和性能，并可应用于偏最小二乘回归模型中。

用偏最小二乘回归法初步建立大坝变形监测偏回归模型，保留其回归因子，变回归系数为遗传优化计算的决策变量。每个个体所含的决策变量 x_i 在各自定义域$[a_i,b_i]$内随机取值，有：

$$x_{ij}=a_{ij}+(b_{ij}-a_{ij})r;$$
$$i=1,2,\cdots,N;j=1,2,\cdots,M。\qquad(1)$$

式中：x_{ij}为第 j 个个体中的第 i 个决策变量，a_{ij}、b_{ij}分别为回归系数的下限、上限，r 为(0，1)区间内的均匀分布随机数，N 为每个个体所含决策变量的个数，M 为群体规模。

根据决策变量建立目标函数,将实际结果与期望结果差的平方和作为问题的目标函数,然后将所建立的目标函数转化为求极值问题。假设有 m 组观测样本,位移为 $y_i(i=1,2,\cdots,m)$,GA-PLSR 模型得出的位移为 y_{mi},则建立的目标函数为:

$$f(x)=\min f=\sum_{i=1}^{m}(y_{mi}-y_i)^2。\tag{2}$$

遗传算法的搜索是按优胜劣汰的自然法则进行的,种群成员的优劣则通过其适应值大小来评价,原始适应函数是求解目标的直接表示,通常采用问题的目标函数作为个体的适应性度量(要求为非负值)。在采用浮点编码和基于线性排列选择策略的情况下,无论对极大值问题或极小值问题,都不需进行适应值的标准化和调节,可以直接使用原始适应值进行排名选择。由于大坝变形监测模型为求极小值问题,且目标函数值为非负值,故可通过改进的界限构造法由式(2)的目标函数 $f(x)$ 变换为个体适应度函数 $F(x)$,即:

$$F(x)=\frac{1}{1+f(x)}。\tag{3}$$

遗传算法在决策变量的定义域内随机选取一个值来产生初始种群,但由于回归系数的定义域为整个实数域,若仍按上述方法随机产生初始种群,则随机性较大,势必导致优化搜索效率下降。针对该问题,本研究在建立模型的过程中,以偏回归模型得出的各回归系数附近区域作为其定义域,选取原则为:根据初始回归系数的正负,取其与 0 之间区域作为其定义域,根据计算结果和精度逐步缩小区间,再按一定概率随机产生初始种群,以提高求解效率和计算结果的可靠性;最后依据计算结果进行调整,得到新的偏回归系数后,将其返回检验,计算复相关系数和标准差,并与原偏回归模型相比较,直至优化所得到的回归系数值在所选定的定义域内且优化结果令人满意为止[10-11,13-16]。

2.2 建模步骤

用 MATLAB 编制了大坝变形监测的 GA-PLSR 模型程序,具体建模步骤可参照文献[14,17]。这里需要说明的是,本研究在确定遗传算法的运行参数和控制参数时,群体规模 $M\in[20,100]$,交叉概率 $p_c\in[0.4,0.99]$,变异概率 $p_m\in[0.001,0.1]$;执行算术交叉算子时,本模型交叉概率 p_c 取值为0.95;执行均匀变异算子时,本模型变异概率 p_m 取值为 0.01。

3 工程应用实例与对比分析

以某拱坝监测资料[18]为例,通过 MATLAB 自编的 GA-PLSR 建模程序,对其垂线顺河向位移采用改进 GA-PLSR 方法进行建模计算,并将计算结果与偏最小二乘回归模型结果进行对比分析。

3.1 模型因子的选择与参数的确定

大坝各测点位移受库水位、温度、时效等因素的综合作用。根据大坝运行的特点,并考虑初始值对其的影响,得到坝体和坝基岩体垂线顺河向水平位移的统计模型,共选用影响因子 14 项,其中水压因子 8 项、温度因子 4 项、时效因子 2 项。模型的表达式为:

$$\delta=\sum_{i=1}^{4}[a_{1i}(H_u^i-H_{u0}^i)]+\sum_{i=1}^{4}[a_{2i}(H_d^i-H_{d0}^i)]+\sum_{i=1}^{2}\left[b_{1i}\left(\sin\frac{2\pi it}{365}-\sin\frac{2\pi it_0}{365}\right)+b_{2i}\left(\cos\frac{2\pi it}{365}-\cos\frac{2\pi it_0}{365}\right)\right]+c_1(\theta-\theta_0)+c_2(\ln\theta-\ln\theta_0)+a_0。\tag{4}$$

式中:δ 为拱坝顺河向水平位移,H_u^i、H_{u0}^i、H_d^i 和 H_{d0}^i 分别为观测日、始测日所对应的上、下游水头的 i 次方($i=1,2,3,4$),a_{1i}、a_{2i}为水压因子回归系数($i=1,2,3,4$),t 为观测日至观测基准日的累计天数,t_0 为建模资料系列第 1 个测值日至观测基准日的累计天数,b_{1i}、b_{2i}为温度因子回归系数($i=1,2$),θ 为观测日至观测基准日的累计天数除以 100,θ_0 为建模资料系列的第 1 个测值日至观测基准日的累计天数除以 100,c_1、c_2 为时效因子回归系数,a_0 为常数项因子。

回归系数 a_{1i}、a_{2i}、b_{1i}、b_{2i}、c_1、c_2 和 a_0 即为遗传算法的决策变量,故 N 取 15,群体规模 $M=80$,交叉概率 $p_c=0.95$,变异概率 $p_m=0.01$,进化终止代数 $T=200$。适应度函数为:

$$F(x)=\frac{1}{1+f(x)}=\frac{1}{1+\sum_{i=1}^{m}(\hat{\delta}_i-\delta_i)^2}。\tag{5}$$

式中:$f(x)$ 为遗传优化目标函数,$f(x)=\min f$;m 为建模序列样本组数;$\hat{\delta}_i$ 为 GA-PLSR 模型的拱坝顺河向位移计算值;δ_i 为拱坝顺河向水平位移实测值。

3.2 计算结果的对比分析

利用 MATLAB 分别编制大坝变形监测 PLSR 模型程序和 GA-PLSR 模型程序,对所选观测资料的因子进行计算,得到坝段垂线顺河向水平位移的最佳 PLSR 模型为:

$$\delta=[-0.220\,6(H_u-H_{u0})-2.689\,3\times10^{-5}(H_u^2-H_{u0}^2)+3.613\,2\times10^{-6}(H_u^3-H_{u0}^3)+3.962\,5\times10^{-6}(H_u^4-H_{u0}^4)+0.386\,6(H_d-H_{d0})+0.003\,2(H_d^2-H_{d0}^2)-1.307\,8\times10^{-4}(H_d^3-H_{d0}^3)-1.442\,5\times10^{-5}(H_d^4-H_{d0}^4)]+\left[7.063\,6\left(\sin\frac{2\pi t}{365}-\sin\frac{2\pi t_0}{365}\right)+2.819\,8\left(\cos\frac{2\pi t}{365}-\cos\frac{2\pi t_0}{365}\right)+0.020\,3\left(\sin\frac{4\pi t}{365}-\sin\frac{4\pi t_0}{365}\right)-0.333\,7\left(\cos\frac{4\pi t}{365}-\cos\frac{4\pi t_0}{365}\right)\right]+[1.158\,2(\theta-\theta_0)-6.929\,6(\ln\theta-\ln\theta_0)]+13.177\,1。$$

最佳 GA-PLSR 模型如下：

$$\delta=[0.091\,0(H_u-H_{u0})+3.612\,6\times10^{-4}(H_u^2-H_{u0}^2)+1.889\,0\times10^{-6}(H_u^3-H_{u0}^3)+1.099\,2\times10^{-8}(H_u^4-H_{u0}^4)-0.027\,1(H_d-H_{d0})-6.916\,5\times10^{-4}(H_d^2-H_{d0}^2)-2.176\,3\times10^{-5}(H_d^3-H_{d0}^3)-7.101\,0\times10^{-7}(H_d^4-H_{d0}^4)]+\left[5.412\,0\left(\sin\frac{2\pi t}{365}-\sin\frac{2\pi t_0}{365}\right)+0.148\,4\left(\cos\frac{2\pi t}{365}-\cos\frac{2\pi t_0}{365}\right)+1.388\,4\left(\sin\frac{4\pi t}{365}-\sin\frac{4\pi t_0}{365}\right)-0.122\,6\left(\cos\frac{4\pi t}{365}-\cos\frac{4\pi t_0}{365}\right)\right]+[0.293\,3(\theta-\theta_0)+3.467\,6(\ln\theta-\ln\theta_0)]+12.309\,0。$$

为检验回归效果，计算模型方程的复相关系数 R 及各统计参数，并对回归方程进行 F 检验。2 种建模方法的统计参数如表 1 所示。从表 1 可以看出，GA-PLSR 模型对 PLSR 模型的回归系数进行了优化处理，回归效果得到了改善，复相关系数由 0.993 6 提高到 0.995 8，总残差平方和由 174.80 降至 130.58，而剩余标准差由 0.790 降至 0.682。基于遗传算法的偏回归模型的回归效果和模型精度均有一定程度的提高。

表 1　GA-PLSR 和 PLSR 模型的统计参数对比

Table 1　Statistical parameter comparison of GA-PLSR model and PLSR model

模型 Model	复相关系数 Multiple correlation coefficient	调整的复测定系数 Adjustive determination coefficient	F 值 F value	总残差平方和 The total sum of residual squares	剩余标准差 Residual standard deviation
PLSR	0.993 6	0.986 5	1 431.01	174.80	0.790
GA-PLSR	0.995 8	0.991 2	2 198.90	130.58	0.682

图 1 为 2 种模型的拟合曲线对比图。

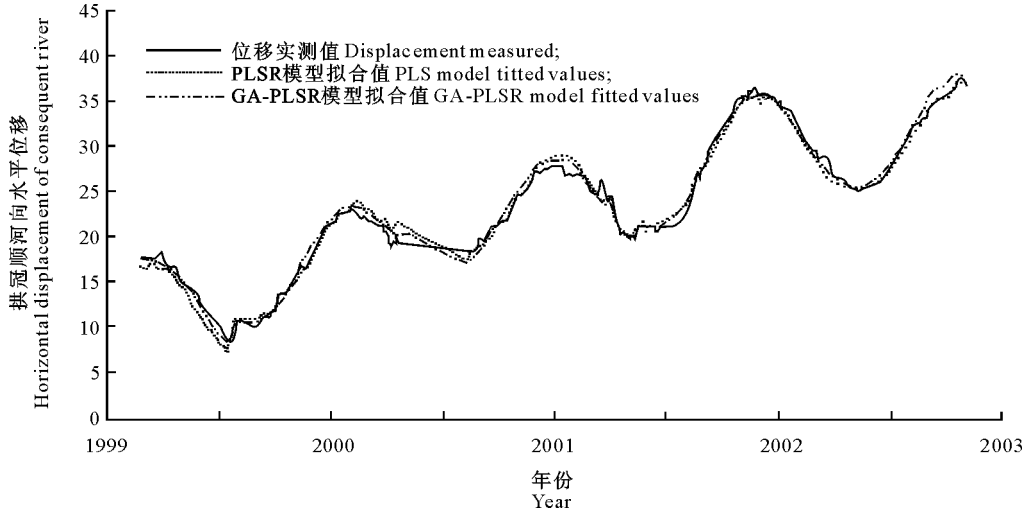

图 1　GA-PLSR 模型和 PLSR 模型的拟合曲线

Fig. 1　The curve fitting drawings of improved GA-PLSR model and PLSR model

对比图 1 中 2 种模型的拟合曲线可以看出，　GA-PLSR 模型在一定程度上改善了原 PLSR 模型

存在的拟合效果不佳的情况。与原PLSR模型的拟合效果相比,GA-PLSR模型的位移模型残差值普遍变小:1999-06的残差最大值从2.057减小到1.230;2000-06的残差最大值从2.302减小到1.194;2001-04的残差最大值从2.356减小到1.579。更重要的是,该阶段的残差值有正有负,显示出了随机变化的性态,而不再是系统性偏离,拟合效果较好。

4 结 论

研究了大坝变形监测的PLSR模型优化问题,利用GA强大的自适应全局优化搜索功能,将改进的GA算法引入偏最小二乘回归模型进行优化,建立了基于改进GA-PLSR的大坝变形监测模型,较好地解决了常规大坝变形监测回归模型中存在的因子多重相关性干扰和模型拟合效果欠佳等问题,所建模型具有较好的拟合与预测能力,有较强的工程实用性。

但由于该模型是在大坝变形监测PLSR模型基础上的优化重估,为了提高搜索效率,每个模型参数的定义域选取的是原位移偏回归模型参数附近的实数域,因此改进GA-PLSR虽然能部分解决大坝监控模型的拟合效果欠佳的问题,但其改善效果仍有一定的局限性,还有待于进一步提高。

[参考文献]

[1] 王惠文.偏最小二乘回归方法及其应用[M].北京:国防工业出版社,1999.
Wang H W. Method and application of partial least-squares regression [M]. Beijing: Publishing Company of National Defense Industry, 1999. (in Chinese)

[2] 杨 杰,方 俊,胡德秀,等.偏最小二乘回归在水利工程安全监测中的应用[J].农业工程学报,2007,23(3):136-139.
Yang J, Fang J, Hu D X, et al. Application of partial least-squares regression to safety monitoring of water conservancy projects [J]. Transactions of the CSAE, 2007, 23(3): 136-139. (in Chinese)

[3] 杨 杰,胡德秀,吴中如.大坝安全监控模型因子相关性及不确定性研究[J].水利学报,2004,35(12):99-100.
Yang J, Hu D X, Wu Z R. Multiple co-linearity and uncertainty of factors in dam safety monitoring model [J]. Journal of Hydraulic Engineering, 2004, 35(12): 99-100. (in Chinese)

[4] 李 波,顾冲时,李智录,等.基于偏最小二乘回归和最小二乘支持向量机的大坝渗流监控模型[J].水利学报,2008,39(12):1390-1391.
Li B, Gu C S, Li Z L, et al. Monitoring model for dam seepage based on partial least-squares regression and partial least square support vector machine [J]. Journal of Hydraulic Engineering, 2008, 39(12): 1390-1391. (in Chinese)

[5] 周光文,袁晓峰,黄筱蓉.递阶偏最小二乘回归在大坝安全监测中的应用[J].水电自动化与大坝监测,2008,32(4):59-61.
Zhou G W, Yuan X F, Huang X R. Hierarchical partial least-square regression and its application to dam safety monitoring [J]. Hydropower Automation and Dam Monitoring, 2008, 32(4): 59-61. (in Chinese)

[6] Reis L F, Bessler F T, Walters G A, et al. Water supply reservoir operation by combined genetic algorithm-linear programming (GA-LP) approach [J]. Water Resources Management, 2006, 20: 227-255.

[7] 王 建,阳 武,郑东健.监测数据回归分析中典型监测值欠拟合释因[J].武汉大学学报:工学版,2003,36(6):9-12.
Wang J, Yang W, Zheng D J. Reason of poor fitting in regression analysis for dam safety monitoring data [J]. Engineering Journal of Wuhan University, 2003, 36(6): 9-12. (in Chinese)

[8] 周 明,孙树栋.遗传算法原理及应用[M].北京:国防工业出版社,2000.
Zhou M, Sun S D. Principle and application of genetic algorithm [M]. Beijing: Publishing Company of National Defense Industry, 2000. (in Chinese)

[9] 吴中如.水工建筑物安全监控理论及其应用[M].北京:高等教育出版社,2003.
Wu Z R. Safety control theory and application of hydraulic structures [M]. Beijing: Publishing Company of High Education, 2003. (in Chinese)

[10] 费如君,董曾川,王德智,等.改进的加速遗传算法在梯级水电站优化调度中的应用[J].水力发电,2008,34(8):8-9.
Fei R J, Dong Z C, Wang D Z, et al. Application of modified RAGA in optimal operation of cascaded hydroelectric plants [J]. Water Power, 2008, 34(8): 8-9. (in Chinese)

[11] 陈立华,梅亚东,董雅洁,等.改进遗传算法及其在水库群优化调度中的应用[J].水利学报,2008,39(5):9-10.
Chen L H, Mei Y D, Dong Y J, et al. Improved genetic algorithm and its application in optimal dispatch of cascade reservoirs [J]. Journal of Hydraulic Engineering, 2008, 39(5): 9-10. (in Chinese)

[12] Zolfaghari A R, Heath A C, McCombie P F. Simple genetic algorithm search for critical non-circular failure surface in slope stability analysis [J]. Computers and Geotechnics, 2005, 32: 139-152.

[13] 刘学增,周 敏.改进的自适应遗传算法及其工程应用[J].同济大学学报:自然科学版,2009,37(3):303-305.
Liu X Z, Zhou M. Improved adaptive genetic algorithm and its application to backward analysis of geotechnical engineering [J]. Journal of Tongji University: Natural Science Edition, 2009, 37(3): 303-305. (in Chinese)

(下转第218页)