遗传算法在非线性最小一乘回归中的应用

景 继[1,2], 王 建[1,2], 谷艳昌[1,2]

(1. 河海大学 水利水电工程学院, 江苏 南京 210098; 2. 河海大学 水资源高效利用与工程安全国家工程研究中心, 江苏 南京 210098)

摘要: 针对含粗差的大坝监测数据非线性回归问题,引入了最小一乘法进行非线性抗差回归,并采用遗传算法解决非线性最小一乘回归的计算. 实例计算表明,该方法能够获得较好的抗差效果,其回归结果受粗差的影响程度小于传统的最小二乘法. 对其适用性的探讨表明,该方法适用于诸如施工期等短期监测数据的抗差回归分析.

关 键 词: 最小一乘; 遗传算法; 非线性; 抗差; 监测; 回归分析
中图分类号: TV698.1 **文献标识码**: A **文章编号**: 1009-640X(2008)02-0043-05

Application of genetic algorithm to nonlinear LAD regression

JING Ji[1,2], WANG Jian[1,2], GU Yan-chang[1,2]

(1. *College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing* 210098, *China*; 2. *National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing* 210098, *China*)

Abstract: In order to solve the problem of nonlinear regression of dam monitoring data that contains outliers, a nonlinear LAD regression model is established and the genetic algorithm is applied to solve it. The analysis of an example shows that the nonlinear LAD regression has the advantage of robustness over the traditional LS regression. This method is suitable for short term monitoring data analysis.

Key words: least absolute deviations; genetic algorithm; nonlinear; robust; monitoring; regression

统计模型回归分析是当前大坝监测数据处理的常用方法,而统计模型回归分析中通常采用的求解方法是最小二乘法,最小二乘法在理论上较为成熟,计算较为方便,通常情况下可以取得较满意的结果. 但实际监测工作中,由于仪器故障、人为因素等不可避免地夹杂有粗差,而最小二乘法对粗差较为敏感,少量的粗差就会导致回归结果的不可靠. 由于最小一乘法对粗差的敏感性较小,因此可以将其用于含有粗差数据的回归分析. 最小一乘的求解属于无约束最优化范畴,但由于其目标函数是非光滑的,因而不便采用传统方法求解. 线性模型的最小一乘回归已有学者提出了不少求解方法[1,2],而非线性模型的最小一乘回归尚没有较好的求

收稿日期: 2007-06-18
基金项目: 国家自然科学基金资助项目(50579010);水利部"948"项目(CT200612)
作者简介: 景 继(1981-),男,江苏海安人,博士研究生,主要从事水工结构工程安全监控研究. E-mail: iamj@163.com

解方法. 本文将遗传算法应用于非线性模型的最小一乘回归求解,以实现大坝监测数据非线性模型的抗差回归,并进一步讨论了其适用情况.

1 基于遗传算法的非线性最小一乘回归方法

1.1 传统回归方法的不足

传统的最小二乘回归以残差平方和极小为准则,其准则函数为:

$$\sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \min \tag{1}$$

式中:$v_i = y_i - f(x_i)$ 为残差; y_i 为因变量实测值; $f(x_i)$ 为拟合值; x_i 为自变量实测值; n 为测值个数. 可见,最小二乘法的目标函数值以残差平方的速率递增,因而对粗差很敏感. 对于这种情况,可采用抗差估计,抗差估计又称稳健估计,其目标函数对粗差敏感程度低于最小二乘法,受粗差影响较小.

1.2 最小一乘回归

最小一乘回归准则要求模型的残差绝对值之和达到最小. 假设有一系列样本数据$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$,用一个函数$f(x)$对其进行拟合,按照最小一乘准则,

$$\sum_{i=1}^{n} |v_i| = \sum_{i=1}^{n} |y_i - f(x_i)| = \min \tag{2}$$

可见,由于最小一乘法只考虑残差的一次方,因而对粗差的敏感性较最小二乘法要小得多[1]. 若拟合函数$f(x)$是线性的,则为线性最小一乘回归;若$f(x)$非线性,则为非线性最小一乘回归. 对于线性最小一乘回归,文献[1,2]已给出了部分计算方法,如线性规划法、直接算法和 GAUSS 定理等. 大坝监测数据分析中常常会用到非线性回归模型. 非线性的最小一乘回归由于方程的非线性而无法采用以上的计算方法. 本文对非线性模型的最小一乘回归计算问题进行了探讨.

非线性模型的最小一乘回归可作为无约束最优化问题来求解. 以(2)式作为目标函数,采用优化算法搜索最小值,其目标函数的导数为:

$$\frac{\partial\left(\sum_{i=1}^{n} |v_i|\right)}{\partial \theta_j} = \frac{\partial\left(\sum_{i=1}^{n} |y_i - f(x_i)|\right)}{\partial \theta_j} = \sum_{i=1}^{n} -\operatorname{sign}(y_i - f(x_i)) \frac{\partial f(x_i)}{\partial \theta_j} \tag{3}$$

式中:$\operatorname{sign}(y_i - f(x_i))$为残差 $y_i - f(x_i)$的正负号取值(+1 或-1);θ_j 为第 j 个待估计回归系数($j = 1, 2, \cdots, n$),n 为测值个数.

由于 $\operatorname{sign}(y_i - f(x_i))$是非连续函数,故(3)式最小一乘法目标函数是非光滑的. 传统的优化算法通常要根据目标函数的导数信息确定下一步的搜索方向,适用于连续光滑函数的优化问题,但不适合最小一乘法的求解.

1.3 基于遗传算法的非线性最小一乘回归求解

与传统优化算法相比,遗传算法利用进化过程中获得的信息进行自组织搜索,具有自组织、自适应、自学习性;遗传算法按并行方式搜索一个种群数目的点而不是单点,具有并行性;遗传算法不需要求导,而只需要目标函数值和适应度信息;遗传算法强调概率转换规则,而不是确定的转换规则等. 因此,遗传算法已得到越来越多的应用[3-6].

由于遗传算法仅需被优化目标函数的函数值信息而无需求导,适合于非光滑函数的优化问题. 因此,可将遗传算法用于非线性模型的最小一乘回归求解.

对一系列样本数据$(x_1, y_i), (x_2, y_2), \cdots, (x_n, y_n)$,建立非线性回归模型:

$$\hat{y}_i = f(x_i, \theta) \tag{4}$$

式中:$\hat{y}_i$ 为拟合值; $f(x_i, \theta)$为一个非线性函数; x_i 为自变量; θ 为回归系数.

按最小一乘准则进行回归计算，即可求解以下最小值问题：

$$\sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - f(x_i,\theta)| = \min \tag{5}$$

式中：y_i 为实测值；$\hat{y}_i = f(x_i,\theta)$ 为拟合值；n 为测值个数.

这是一个无约束最优化问题，因此可将(5)式作为目标函数，将回归系数 θ 作为决策变量，用遗传算法搜索得出其最优解 θ^* 作为该非线性最小一乘回归问题的解. 基本运算步骤如下：

Step 1：初始化遗传算法运行参数，设置种群大小、决策变量个数、编码方式、长度及进化终止准则等信息；编码方式通常有二进制编码和实数编码，二进制编码由符号集{0,1}组成，实数编码每一个基因用某一范围内的一个实数表示，对应一个决策变量的真实值；

Step 2：随机生成初始种群；

Step 3：计算种群中各个个体适应度，个体适应度大小与个体被遗传到下一代的机会大小成正比. 适应度必须大于或等于零. 对于不同的问题，需要确定好由目标函数值到适应度之间的转化规则；

Step 4：执行选择运算，选择适应度较高的个体构成下一代种群；选择运算建立在对个体适应度的评价之上，针对具体的问题可以采用不同的选择算子；

Step 5：对选出的下一代个体进行交叉运算，交叉运算模拟自然界生物的基因重组现象，从配对后的两个父代个体产生出新的子代个体，它是产生新个体的主要方法，决定了遗传算法的全局搜索能力[2]；

Step 6：对选出的下一代个体进行变异运算，变异运算改变个体编码串中的某些基因值，从而生成新的个体，它是产生新个体的辅助方法，决定了遗传算法的局部搜索能力[2]；

Step 7：终止条件判断，若当前步满足 Step 1 中设置的终止准则，则算法终止，将进化过程中得到的具有最大适应度的个体作为最优解输出；否则，转到 Step 3.

2　实 例 计 算

以某碾压混凝土重力坝施工期某钢筋计的监测资料分析为例，建立统计回归模型. 由于是施工期，水库尚未蓄水，因此只考虑伴测温度和非线性的指数时效，建立统计回归模型表达式[1]如下：

$$\sigma = b_0 + b_1(1 - e^{-b_2\theta}) + b_3 T \tag{6}$$

式中：σ 为钢筋应力(MPa)；b_0, b_1, b_2, b_3 为需要求解的回归系数，用遗传算法进行搜索；θ 为观测日至始测日累计天数除以 100；T 为温度(℃).

采用遗传算法求解，取目标函数：

$$\mathrm{obj}(b_0,b_1,b_2,b_3) = \sum_{i=1}^{n} |\sigma_i - [b_0 + b_1(1 - e^{-b_2\theta_i}) + b_3 T_i]| \tag{7}$$

式中：n 为测值个数；σ_i 为应力实测值序列中第 i 个测值；θ_i 为第 i 个测值至始测日累计天数除以 100；T_i 为伴测温度实测值序列中第 i 个测值.

2.1　算法的改进

求目标函数关于常数项 b_0 的偏导数并令其等于零：

$$\frac{\partial \mathrm{obj}(b_0,b_1,b_2,b_3)}{\partial b_0} = -\sum_{i=1}^{n} \mathrm{sign}(\sigma_i - [b_0 + b_1(1 - e^{-b_2\theta_i}) + b_3 T_i]) = 0 \tag{8}$$

显然，中位数 $b_0 = \mathrm{med}\{\sigma_i - [b_1(1 - e^{-b_2\theta_i}) + b_3 T_i]\}$ 是(8)式的解. 因此，在搜索空间中给定一组 b_1, b_2, b_3 的值，则 $b_0 = \mathrm{med}\{\sigma_i - [b_1(1 - e^{-b_2\theta_i}) + b_3 T_i]\}$ 对应的目标函数达到最小. 因此，在遗传算法搜索过程中可以只搜索回归系数 b_1, b_2, b_3，而常数项 b_0 直接取为中位数 $\mathrm{med}\{\sigma_i - [b_1(1 - e^{-b_2\theta_i}) + b_3 T_i]\}$，目标函数重新表示为：

$$\begin{cases} \mathrm{obj}(b_1,b_2,b_3)=\sum_{i=1}^{n}\left|\sigma_i-\left[b_0+b_1(1-\mathrm{e}^{-b_2\theta_i})+b_3T_i\right]\right| \\ b_0=\mathrm{med}\left\{\sigma_i-\left[b_1(1-\mathrm{e}^{-b_2\theta_i})+b_3T_i\right]\right\} \end{cases} \tag{9}$$

这样,遗传算法的目标函数表达式中就省去了常数项,从而大大缩小了算法的搜索范围.

设置最大遗传代数为 500,代沟为 0.8(子种群大小为父种群的 80%). 根据以往的监测数据分析经验,确定各决策变量搜索范围:$b_1\in[-15,15]$、$b_2\in[-10,20]$、$b_3\in[-10,10]$. 设置种群大小为 200,由于本优化问题属于函数优化,各决策变量均为连续变量,而二进制编码只能产生有限的离散点阵,且有可能产生额外的最优点[7]. 因此本文采用实数编码.

目标函数按(9)式计算,由于是求解最小化问题,因此要使目标函数值较小的个体获得较大的适应度以获得较大的遗传机会. 本文采用基于排序的适应度计算方法,将目标值从大到小排列后,按下式计算适应度[4]:

$$\mathrm{Fit}(i)=\frac{2(\mathrm{POS}_i-1)}{\mathrm{NIND}-1} \tag{10}$$

式中:$\mathrm{Fit}(i)$为第 i 个个体的适应度;POS_i 为第 i 个个体在排序中的位置;NIND 为种群大小.

这种基于排序的适应度分配方法可以避免算法的早熟,防止算法收敛到一个局部最优点,而不是全局最优点.

选择操作采用轮盘赌算子(比例选择算子),个体按一定机率被选中,适应度越大,被选中的机率越大. 交叉运算采用实数编码的算术交叉,由两个父个体进行线性组合得到子个体. 变异运算采用实值变异,按设定的变异概率对个体中每个变量附加上随机的扰动. 对子代进行交叉和变异操作后,按适应度将其插入到父代种群中,替换掉父代中适应度较小的个体. 设置代沟为 0.8,这样子代将替换掉父代中 80% 的个体,而父代种群中最优的 20% 的个体仍得以保留而不被破坏,保证算法的收敛性.

2.2 计算结果分析

在数据中加入了 4 个粗差数据见表 1. 对不含粗差和含粗差两种情况分别进行最小一乘回归计算. 为了便于比较,两种情况也均采用最小二乘法进行回归分析,计算结果见表 2. 在含粗差数据的最小一乘回归遗传算法求解过程中,各进化代中的最优值变化过程见图 1,实测和拟合曲线见图 2.

由表 2 可见,加入粗差前后,最小一乘法得到的各回归系数变化很小,与无粗差情况下的最小二乘计算结果也较为接近;而加入粗差前后的最小二乘法回归结果变化较大. 这说明最小二乘法对粗差较敏感.

表 1 粗 差 数 据

Tab. 1 Gross errors

日 期	加入粗差前的测值/ MPa	加入粗差后的测值/ MPa	粗差/ MPa
2005-07-17	5.573	10.1	4.527
2005-10-13	8.952	24.2	15.248
2006-01-03	15.326	10.22	-5.106
2006-03-27	15.262	24.3	9.038

表 2 计 算 结 果

Tab. 2 Results of calculation

回归系数	最小一乘		最小二乘	
	加入粗差前	加入粗差后	加入粗差前	加入粗差后
b_0	49.2	49.209	47.854	35.85
b_1	-4.49	-4.496	-4.228	3.658
b_2	1.06	1.06	0.665	12.665
b_3	-1.566	-1.566	-1.532	-1.247

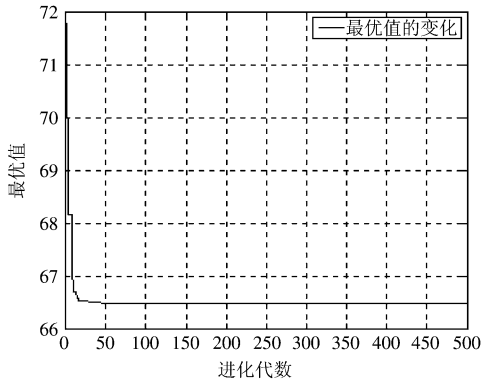

图 1　最优值变化过程线

Fig. 1　Variation curve of optimum value

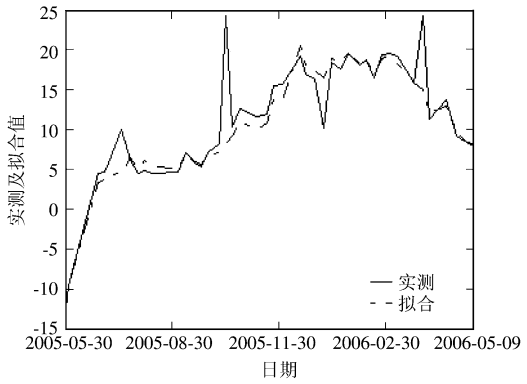

图 2　实测和拟合曲线

Fig. 2　Curves of observed values and fitting values

从算法原理看，每增加 1 个回归系数，搜索空间就扩大一维，从而使计算量急剧增加. 整个计算过程需要计算目标函数的次数=种群大小×(进化代数+1)，每次目标函数的计算量与测值个数成正比，若种群较大、进化代数较多，则测值系列越长，计算量越大. 综上分析，该方法适用于测值系列较短、考虑回归因子较少(建议回归因子少于 5 个)的问题，如施工期的大坝监测资料回归分析.

3　结　语

(1)由于最小一乘法相对最小二乘法具有较好的抗差性，因此，对于含有粗差的大坝监测数据可采用最小一乘法进行抗差回归分析. 非线性最小一乘回归由于目标函数的非光滑性和回归方程的非线性，因而难以使用传统的方法求解. 由于遗传算法仅用到目标函数的函数值信息而无需求导，因而适合于非光滑函数的优化问题. 本文采用遗传算法对非线性最小一乘回归问题进行了求解，取得了较好的效果.

(2)最小一乘法的计算量受回归系数个数和测值系列长短影响较大，尤其是回归系数的个数，该算法适用于测值系列较短，回归因子较少的问题.

参　考　文　献：

[1]　陈希孺. 最小一乘线性回归(上)[J]. 数理统计与管理，1989，(5)：48-55.
[2]　陈希孺. 最小一乘线性回归(下)[J]. 数理统计与管理，1989，(6)：48-56.
[3]　周　明，孙树栋. 遗传算法原理及应用[M]. 北京：国防工业出版社，1999：1-60.
[4]　王小平，曹立明. 遗传算法—理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002：1-50.
[5]　雷英杰，张善文，李续武，等. Matlab 遗传算法工具箱及应用[M]. 西安：西安电子科技大学出版社，2005：62-106.
[6]　金菊良，丁　晶. 遗传算法及其在水科学中的应用[M]. 成都：四川大学出版社，2000：25-66.
[7]　吴中如. 水工建筑物安全监控理论及其应用[M]. 南京：河海大学出版社，1990：36-120.