基于遗传算法的最小一乘回归新算法*

张 春 涛
(重庆三峡学院 计算机科学系,重庆 万州 404000)

摘 要:最小一乘在稳健性上比最小二乘好,使得最小一乘在工程中得到广泛的应用,但求解最小一乘的算法并不理想;本文根据最小一乘的性质,把最小一乘问题变为组合优化问题。将遗传算法用在最小一乘模型的求解上,在后面的仿真实验中得到了较好的效果。

关键词:最小一乘;最小二乘;遗传算法;线性模型
中图分类号:O241.5;O242.1 文献标识码:A 文章编号:1672-6693(2005)02-0015-03

A New Algorithm for the Least Absolute Deviation Regression Based on the Genetic Algorithm

ZHANG Chun-tao
(Dept. of Computer Science, Three Gorge College, Wanzhou Chongqing 404000, China)

Abstract: The least absolute deviations are using widely used in engineering because of it's robustness, but the algorithm solving the least absolute deviation is not efficient. Changing the least absolute deviation to the combinatorial optimization, based on it's characters, we use the genetic algorithm to solve the least absolute deviation regression. At last the numerical experimentations show that the Genetic algorithm is efficient.

Key words: least absolute deviation; least squares; genetic algorithm; linear model

在近代关于数理统计的稳健性研究中发现,基于最小二乘估计的回归分析有时并不理想,尤其是当所收集到的样本数据较少并且其中含有个别异常值(或大误差点)时,异常点有较大的偏差,其平方之值相对更大,为了压低平方和,就不能不"将就"这些点,因而虚增加了残差大的数据对回归线施加的影响,从而异常点会把回归线拉得离它更近一些,导致回归线"失真"较大[1]。因此,产生了以"残差绝对值和最小"为准则的最小一乘回归方法。

在误差不服从正态分布(比如,计量经济中误差有时服从尾部占更大比重的分布)的问题中,最小一乘估计的统计性能优于最小二乘估计[1],其具有不可替代的优越性;另外,最小一乘准则的稳健性比最小二乘准则的稳健性好,而且其受异常点的影响较小一点,所以将误差绝对值之和最小作为目标也被广泛地应用到工程实践中。

由于最小一乘回归属于不可微问题,与最小二乘相比,具有较大的难度。在给出的样本点(即原始数据)的条件下,快速正确地获得稳健的最小一乘回归模型的方法不多。目前,在已有的文献中主要有松弛算法、目标规划法和搜索算法[2~4],而这些算法都有各自的缺点,特别是当初始数据点较多时,用上述方法来求解较困难。因而快速正确地获得最小一乘回归模型的算法十分必要。根据最小一乘法确定的直线必过样本点的特征,把最小一乘法转化为从初始样本点集中选出几个样本点的组合优化问题。而近年来发展起来的非数值优化算法—遗传算法,在解决组合优化问题方面表现出良好的性能。

遗传算法(GA,Genetic Algorithm)是一种更为宏观意义下的仿生算法,它模仿的机制是一切生命与智能的产生与进化过程,激励好的结构,在迭代过程中保持已有结构的同时,寻找更好的结构。它具

* 收稿日期:2004-11-11 修回日期:2005-04-04
资助项目:重庆三峡学院科技项目资助
作者简介:张春涛(1978-),男,重庆人,硕士研究生,主要研究方向为进化算法及其应用。

有智能性搜索、并行式计算、全局优化和计算复杂度与问题规模无多大直接关系等优点,没有传统的建立在梯度计算基础上优化算法的缺点,特别适合于求解目标函数的多极值点问题和大规模组合优化问题。

1 最小一乘回归模型

1.1 最小一乘回归模型的建立

假设样本数据为(x_1,y_1)、(x_2,y_2)、…、(x_n,y_n),其中$y_n\in\mathbf{R}^1$,$x_i=(x_{1i},x_{2i},\cdots,x_{Pi})\in\mathbf{R}^P$是$P$维行向量。现将由这些数据拟合一条曲线

$$y=f(x)\qquad x\text{是}P\text{维行向量}\tag{1}$$

当用最小二乘准则时,建立的模型为

$$\min Q_2=\sum_{i=1}^{n}(y_i-f(\hat{x}_i))^2=\sum_{i=1}^{n}e_i^2\tag{2}$$

式中e_i为实际值y_i与拟合值$f(\hat{x}_i)$的残差($i=1,2,\cdots,n$),Q_2为残差平方和。

当用最小一乘准则时,建立的模型为

$$\min Q_1=\sum_{i=1}^{n}|(y_i-f(\hat{x}_i))|=\sum_{i=1}^{n}|e_i|\tag{3}$$

式中Q_1为残差绝对值之和。

假定拟合的曲线(1)是线性的,即$y=a+xb^{\mathrm{T}}$。式中,a和b为待定参数,T表示向量的转置,其中$a\in\mathbf{R}^1$,$b=(b_1,b_2,\cdots,b_P)$为P维的行向量。

用最小二乘准则可建立如下线性回归模型

$$\min Q_2=\sum_{t=1}^{n}(y_t-a-\sum_{i=1}^{P}b_ix_{it})^2\tag{4}$$

用最小一乘准则建立的线性回归模型为

$$\min Q_1=\sum_{t=1}^{n}|y_t-a-\sum_{i=1}^{P}b_ix_{it}|\tag{5}$$

由于(5)式是不可微优化问题,当n(初始数据点数)较大时,用通常的优化算法基本无能为力。

假定拟合的曲线(1)是非线性的,即建立非线性回归模型,根据文献[5]可将其线性化,所以下面只讨论线性回归模型。

1.2 最小一乘线性回归模型的性质

定理1 设有n个样本点(x_i,y_i)($i=1,2,\cdots,n$),则由最小一乘准则确定的直线$y=a+xb$经过其两个样本点。

定理2 设有$n(n>P)$个样本点$(x_{1i},x_{2i},\cdots,x_{Pi},y_i)$,则由最小一乘准则确定的直线$y=b_1x_1+b_2x_2+\cdots+b_Px_P+a$经过其$P+1$个样本点。

定理3[6] 设有$n(n>P)$个样本点$(x_{1i},x_{2i},\cdots,x_{Pi},y_i)$,则由最小一乘准则确定的直线$y=b_1x_1+b_2x_2+\cdots+b_Px_P$经过其$P$个样本点。

由上述定理可知,只需应用枚举法从n个样本点中任意选择$P+1$个样本点确定一条直线,并从中选择误差绝对值之和最小的即为相应的估计参数。由该定理得到的参数估计值是最小一乘准则下的理论值,它不存在初值的选择(即对初值无依赖性)、参数估计的稳定性等问题,这是与松驰算法、目标规划法和搜索算法求解最小一乘问题的本质区别。但当样本点n较大时,用枚举法显然不太现实,因此用遗传算法的群体搜索技术来确定从n个样本点中选择$P+1$个样本确定的一条直线,且该直线的误差绝对值之和最小。

2 遗传算法

GA是基于自然选择和自然遗传这种生物进化机制的搜索算法,把优化问题的解的搜索空间映射为遗传空间,把每一可能的解编码为一个称为染色体的二进制串(也有其它编码方法),染色体的每一位称为基因。每个染色体(对应一个个体)代表一个解,一定数量的个体组成群体。GA首先随机地产生一些个体组成初始群体(即问题的一群候选解),按预先根据目标函数确定的适应度函数计算各个体对问题环境的适应度,再根据个体适应度对个体对应的染色体进行选择,抑制适应度低的染色体,弘扬适应度高的染色体,进行交叉、变异等遗传操作产生进化了的一代群体。如此反复操作,一代一代不断向更优解方向进化,最后得到满足某种收敛条件的最适应问题环境的群体,从而获得问题的最优解。

3 GA应用于最小一乘回归模型

假设样本数据为(x_1,y_1)、(x_2,y_2)、…、(x_n,y_n),其中$y_n\in\mathbf{R}^1$,$x_i=(x_{1i},x_{2i},\cdots,x_{Pi})\in\mathbf{R}^P$是$P$维行向量。现将由这些数据拟合一条曲线:$y=b_1x_1+b_2x_2+\cdots+b_Px_P+a$。根据最小一乘线性回归模型的性质(定理1,定理2,定理3),关键是找出n个样本点中的P个点,用这P个点来确定一条直线,然后用这条直线来确定其残差绝对值之和,再根据遗传算法的实现步骤,得到如下的算法。

3.1 算法描述

(1) 编码。样本数据点编号分别为$A_1,A_2,\cdots,A_n$。采用0-1编码方法,$A_i\in\{0,1\}$($i=1,2,\cdots,n$),当$A_i=1$时表示编号为A_i的数据点被选中,反之,$A_i=0$表示编号为A_i的数据点没被选中。根据定理2,得到合法个体的标准为$\sum_{i=1}^{n}A_i=P+1$,即每个个体中只有$P+1$个数据点被选出。

(2) 初始群体的确定。设群体规模$N=100$,初

始群体中的个体用随机的方法产生,但要满足合法个体的标准。

(3) 确定适应值函数 $G(x)$。根据个体选出的 $P+1$ 个点确定出 $b_1, b_2, \cdots, b_P$ 和 a 的值,即确定了一条直线。把该直线与其他数据点的残差绝对值之和作为该个体的适应度。

(4) 选择算子。采用比例选择算子。即个体在下一代群体中的个数由该个体的适应值在种群总的适应值中的比例来决定。

(5) 交叉算子。采用两点交叉算子,要判断个体的合法性。交叉概率 P_C 为 $0.7 < P_C < 0.95$。

(6) 变异算子。根据编码的特点,采用循环移位变异算子。变异概率 P_m 为 $0.001 < P_m < 0.1$。

(7) 终止条件。取最大迭代次数 $T < 100$ 或残差绝对值之和充分小。

3.2　仿真实验

例 1　对文献[7]的例题数据用遗传算法拟合成直线 $y = ax + b$,与原方法(迭代法)进行比较见表 1。

表 1　遗传算法与迭代法的比较

遗传算法			迭代法		
a	b	$\sum \lvert e_i \rvert$	a	b	$\sum \lvert e_i \rvert$
0.297 297	19.027 027	40.378 378	0.357 1	2.939 5	42.285 205

从上表可以看出使用遗传算法的精度比使用迭代法的精度要高,是因为用遗传算法解出的是理论值,只是在最后一步才有舍入误差 迭代法的舍入误差从算法开始就产生,自然精度较差。说明用遗传算法求解最小一乘是可行的。

例 2　对 2004 高教社杯全国大学生数学建模竞赛 B 题的第一小题的数据用最小一乘拟合成 $Y = AX + B$ 的形式,得出结果如下:

$$y_1 = 0.079\ 384x_1 + 0.039\ 639x_2 + 0.051\ 164x_3 + 0.117\ 819x_4 - 0.027\ 873x_5 + 0.117\ 894x_6 + 0.125\ 255x_7 - 0.003\ 77x_8 + 112.653\ 8$$

$$y_2 = -0.056\ 082x_1 + 0.123\ 521x_2 - 0.001\ 531x_3 + 0.031\ 691x_4 + 0.085\ 645x_5 - 0.114\ 901x_6 - 0.016\ 571x_7 + 0.097\ 550x_8 + 132.544\ 695$$

$$y_3 = -0.068\ 310x_1 + 0.070\ 058x_2 - 0.155\ 637x_3 - 0.009\ 646x_4 + 0.126\ 659x_5 + 0.004\ 529x_6 - 0.005\ 848x_7 - 0.200\ 066x_8 - 110.262\ 986$$

$$y_4 = -0.035\ 336x_1 - 0.105\ 793x_2 + 0.204\ 044x_3 - 0.023\ 564x_4 - 0.014\ 093x_5 + 0.003\ 591x_6 + 0.149\ 137x_7 + 0.073\ 916x_8 + 78.775\ 698$$

$$y_5 = -0.001\ 185x_1 + 0.236\ 736x_2 - 0.066\ 161x_3 - 0.043\ 2x_4 - 0.067\ 648x_5 + 0.066\ 427x_6 + 0x_7 - 0.010\ 782x_8 + 134.788\ 440$$

$$y_6 = 0.235\ 233x_1 - 0.067\ 293x_2 - 0.079\ 587x_3 + 0.089\ 205x_4 + 0.044\ 341x_5 - 0.005\ 985x_6 + 0.172\ 965x_7 - 0.002\ 696x_8 + 122.984\ 363$$

其中 Y 表示 6 条线路的有功潮流值,X 表示 8 台发电机组的出力。算出各条线路绝对误差和分别为:1.199 751,0.818 208,0.899 666,0.846 811,1.082 345,1.132 507。

最后用松弛算法算出各条线路绝对误差和分别为:1.200 385,0.930 022,1.043 265,0.998 321,1.199 382,1.345 866。

由此可见使用遗传算法得出了较理想的结果。

4　结束语

本文提出了求解最小一乘的新方法——遗传算法法,克服了最小一乘求解难的问题,并在后面的两个数值实验中验证了该方法的有效性和可操作性,可以预见,最小一乘将同最小二乘一样在工程中得到广泛的应用。

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] 董祺." 残差绝对值和最小 "准则的松弛算法[J]. 预测,1990,9(2):16-18.

[3] 周宗放,杨春德. 应用目标规划法确定组合预测权重数初探 [J]. 重庆邮电学院学报,1994,6(2):59-64.

[4] 李显方, 李学全." 残差绝对值和最小 "准则的搜索算法 [J]. 预测, 1994, 13(4):49-50.

[5] 董建,谢开贵. 基于最小一乘准则的非线性回归模型研究[J]. 重庆师范学院学报(自然科学版),2001,18(4):71-74.

[6] 谢开贵,宋乾坤,周家启. 最小一乘线性回归模型研究[J]. 系统仿真学报,2002,14(2):189-192.

[7] 李德志. 过两个样本点的最小一乘回归线[J]. 数量统计与管理,1996,15(2):40-43.

(责任编辑　黄　颖)