基于遗传算法的一种数据拟合方法

张善文，刘建都，韩小斌
(空军工程大学 导弹学院，陕西 三原 713800)

摘　要:基于遗传算法,提出了一种数据拟合方法,通过对函数参数进行编码,多次进行交叉变异操作,最终得到参数估计值。实验结果表明,该方法是有效可行的。

关键词:数据拟合;遗传算法;函数参数
中图分类号: TP301　　**文献标识码**:A　　**文章编号**:1009-3516(2006)01-0066-03

在科学数据的处理过程中,如管理与决策活动中往往需要对大量的数据进行分析并做出正确的预测。时间序列数据是最常见、十分重要的数据类型之一,时间序列建模及其预测技术多年来得到了广泛应用。对于建模及其预测,常常需要建立变量之间的数学模型f。由于现实数据的复杂性,观察数据一般含有噪声,往往无法获取足够的信息来预测的f中的参数,如果片面的或局部的硬性处理将会导致大的预测偏差。因此需要从全局上对f的参数值进行优化,遗传进化算法的广泛可用性和全局最优性为数据预测和拟合提供了可靠的工具。

遗传算法模拟自然界生物进化过程,从随机产生的一群初始解开始,按优胜劣汰的自然选择机制,通过复制、杂交、变异等遗传操作逐步改善直到找到满意解。遗传规划[1]将字符串推广到计算机程序,尤其是其符号回归技术直接对数学表达式进行操作,按树的遍历进行编码,通过对输入输出数据的分析建立描述复杂系统的数学表达式模型。

1　标准遗传算法

遗传算法(GA)是模仿自然生物进化的随机全局搜索和优化方法。遗传算法操作使用适者生存的原则,在潜在的解决方案种群中逐次产生一个近似最优的方案。在GA的每一代,根据个体在问题域中的适应度值和从自然遗传学中借鉴来的再造方法进行个体选择产生一个新的近似集。这个过程导致进化的个体比创造它们的个体能更好地适应它们的环境,就像自然适应一样。通过对生物遗传和进化过程中选择、交叉和变异机制的模仿,来完成对问题最优解的自适应搜索过程。

标准遗传算法是一种群体型操作[2,3],该操作以群体中的所有个体为对象。选择、交叉、变异是遗传算法的3个主要操作算子,它们构成了所谓的遗传操作,使遗传算法具有了其它传统方法没有的特点。遗传算法中包含了4个步骤:

1)染色体编码与解码。基本遗传算法使用固定长度的二进制符号串来表示群体中的个体,其等位基因是由二值{0,1}所组成。初始群体中各个个体的基因可用均匀分布的随机数来生成。

编码:设某一参数的取值范围为$[U_1,U_2]$,用长度为k的二进制编码符号来表示该参数,则它总共产生2^k中不同的编码,可使参数编码时的对应关系为

$$\begin{aligned}&000000\cdots0000=0\rightarrow U_1\\&000000\cdots0001=1\rightarrow U_1+\delta\\&\quad\vdots\quad\quad\vdots\quad\quad\vdots\\&111111\cdots1111=2^k-1\rightarrow U_2\end{aligned}$$

收稿日期:2006-04-06
基金项目:陕西省自然科学基金资助项目(2006F18)
作者简介:张善文(1965-),男,陕西阎良人,副教授,博士,主要从事智能信息处理研究.

其中,$\delta = U_2 - U_1/2^k - 1$。

解码:假设某一个体的编码为 $b_k b_{k-1} b_{k-2} \cdots b_2 b_1$,则对应的解码公式为 $U_1 + (\sum_{i=1}^{k} b_i \cdot 2^{i-1}) \cdot \frac{U_2 - U_1}{2^k - 1}$。

2) 个体适应度的检测评估。基本遗传算法按与个体适应度成正比的概率来决定当前群体中各个个体遗传到下一代群体中的机会多少。为了正确估计这个概率,要求所有个体的适应度必须为非负数。所以,根据不同种类的问题,需要预先确定好由目标函数值到个体适应度之间的转换规律。

3) 遗传操作。基本遗传算法使用下列 3 种遗传算子进行操作:①选择使用比例选择算子;②交叉使用单点交叉算子;③变异使用基本位变异算子或均匀变异算子;

4) 基本遗传算法的运行参数。基本遗传算法有下列 4 个运行参数需要预先设定 M, T, p_c, p_m,M 为群体大小,即群体中所含个体的数量,一般取为 20 - 100;T 为遗传算法的终止进化代数,一般取为 100 - 500;p_c 为交叉概率,一般取为 0.4 - 0.99;p_m 为变异概率,一般取为 0.000 1 - 0.1。

2　时间序列数据拟合的方法

时间序列数据中蕴涵着丰富的系统信息,时间序列分析就是要发掘出隐含的知识(关系、规则等)。前后因果关系是其中最重要的一种,就是根据历史序列,找出反应系统演化规律的函数 $f(n)$,建立预测模型。对等时间间隔时间序列 $\{u_t\}$,就要得到如下形式的预测模型:

$$u_t = f(u_{t-1}, u_{t-2}, \cdots, u_{t-p}) \tag{1}$$

式中,p 为输入历史信息时步数,反映了对系统当前状态影响最大的最临近的历史状态数。如果将长度为 p 的窗口在序列中移动,按式(1)就形成滚动多步预测。显然式(1)预测能力的好坏完全取决于 f 和 p 的正确确定。一种做法是根据经验或简单数据拟合分析并采取一定的简化假设确定一种或几种模型(函数)作为 f 的待定形式,如线性函数,幂函数等,再由最小二乘法等等回归技术,确定其中的参数[4]。在实际问题中,f 的形式可以根据已有知识预定,设

$$y = f(a_1, a_2, \cdots, a_m; x_1, x_2, \cdots, x_n) \tag{2}$$

其中,$a_1, a_2, \cdots, a_m$ 为 f 的待定参数,$x_1, x_2, \cdots, x_n$ 为 f 中的自变量。

已知 J 组自变量为 $x_1^j, x_2^j, \cdots, x_n^j, j = 1, 2, \cdots, J$ 对应的观察函数值为 $y^j, j = 1, 2, \cdots, J$。下面给出根据遗传算法[5]对待定参数 $a_1, a_2, \cdots, a_m$ 进行预测的步骤:

1)对 $a_1, a_2, \cdots, a_m$ 进行编码得到初始种群。

2)对初始种群进行解码。

3)计算目标函数值。由式(2)计算 $x_1^j, x_2^j, \cdots, x_n^j, j = 1, 2, \cdots, J$ 处的函数值,记为 $Y^j, j = 1, 2, \cdots, J$。

4)计算适应度值。$p_i = \sum_{j=1}^{J} (y^j - Y^j)^2$,其中,$p_i$ 表示第 i 个个体的适应度值。

5)遗传操作。使用 3 种遗传算子选择、交叉、变异进行遗传运算。

6)设定遗传终止条件。若满足条件,遗传终止,输出 p_i 最小时对应的个体的编码并编码得参数 $a_1, a_2, \cdots, a_m$;若不满足条件,返回 3)继续。

3　仿真实验

实验 1 为了验证本文提出拟合方法的有效性,已知一个二次多项式 $f(x) = 3x^2 + 5x + 9$,给出自变量 x 一组值 X,计算对应的函数值 Y,对 Y 增加白噪声,信噪比为 2 dB,不妨仍记为 Y。

现在由 X 和 Y 拟合一个二次多项式:$f(x) = a_1 x^2 + a_2 x + a_3$。

假设多项式的系数 a_1, a_2, a_3 的区间范围都为[0,20],使用格雷码对其进行编码,长度都为 15。群体大小为 40;遗传算法的终止代数 20;交叉概率 0.5;变异概率为 0.02。　　通过本文提出的拟合方法,遗传 20 代时得到多项式系数为[3.000 0, 5.000 0, 9.000 0],与已知多项式系数完全相同。说明该拟合方法是可行的,而且对噪声不敏感。

实验 2 ENSO 数据由复活岛和澳大利亚达尔文之间的月平均大气压差组成[6],见图 1。

从图 1 可以看出,ENSO 数据具有明显的周期性,因此可用傅里叶级数进行描述,设

$$y(x) = a_0 + \sum_{j=1}^{N} [a_i\cos(\frac{2\pi x}{c} + b_i\sin(\frac{2\pi x}{c})] \qquad (3)$$

其中,c 为周期,N 一般是已知的常数。需要确定级数中的参数 $a_0, a_1, \cdots, a_N, b_1, b_2, \cdots, b_N, c$ 。由 ENSO 数据的范围估计这些参数的范围。

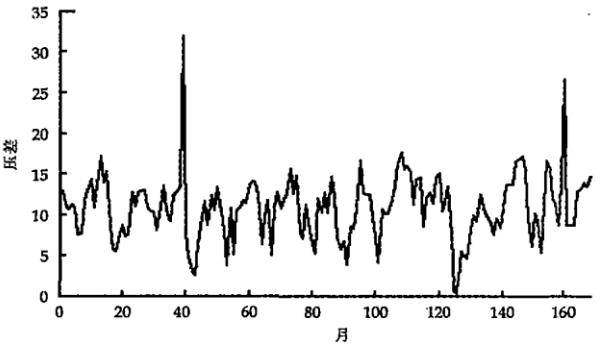

图 1 ENSO 数据

利用本文提出的拟合方法,对 $a_0, a_1, \cdots, a_N, b_1, b_2, \cdots, b_N, c$ 进行拟合。N 取为 3,编码取为浮点数编码,系数 $a_0, a_1, \cdots, a_N, b_1, b_2, \cdots, b_N$ 的范围都取为[-5,30], c 的范围取为[0,36]。群体大小为 60;遗传算法的终止代数 200;交叉概率 0.5;变异概率为 0.02。在 200 代时计算的结果为:

a_0, a_1, a_2, a_3 分别为(10.450 1, 2.810 1, -0.799 2, -1.611 0); b_1, b_2, b_3 分别为(1.477 1, 0.744 3, 0.171 0); c 为 12。

计算拟合标准误差为 0.001 5,说明拟合效果比较好。

4 结束语

遗传算法以其较强的全局搜索能力和鲁棒性,近几年在工程智能优化中得到了广泛的应用。遗传算法以决策变量的编码作为运算对象,直接以目标函数值转换为适应度值,作为搜索信息,同时使用多个搜索点的搜索信息,使用概率搜索技术,为求解复杂系统优化问题提供了一种通用框架。本文在遗传算法的基础上,将时间序列建模看作是函数、变量和参数的组合优化过程,提出了基于遗传算法的函数参数拟合优化搜索方法。实例结果表明,该算法具有较高的预测精度和推广预测能力。

参考文献:

[1] Kpza J. Genetic programming: On the programming of computers by natural selection[M]. MA:MIT Press,1992.

[2] 雷英杰,张善文,李续武,等. MATLAB 遗传算法工具箱及其应用[M]. 西安:西安电子科技大学出版社,2005.

[3] Liu yong,Kang lishan,Chen yuping. Nor Numerical Parallel algorithms Genetic algorithm[M]. Beijing:Science Press ,1998.

[4] 曾 毅. 改进的遗传算法在求解非线性方程组的应用[J]. 华东交通大学学报,2004,29(4) :39 -41.

[5] 张 雷,郑泽席,宋万德. 一种基于遗传算法的决策支持系统建模方法[J]. 空军工程大学学报(自然科学版),2000,1(3):27 -29.

[6] 苏金明,阮沈勇,王永利. MATLAB 工程数学[M]. 北京:电子工业出版社,2005.

(编辑:田新华)

A Data Fitness Method Based on Genetic Algorithm

ZHANG Shan - wen, LIU Jian - dou

(The Missile Institute, Air Force Engineering University, Sanyuan 713800, Shaanxi, China)

Abstract: Based on the genetic algorithm, a data fitness method is presented in this paper. And satisfactory evaluation values of parameters are obtained through coding the function parameters and many times of crossover and mutation operation. The experiment results indicate that this method is feasible and effective.

Key words: data fitness; genetic algorithm; function parameter