采用二叉树编码的遗传算法实现数据拟合*

刘东波　高春鸣
(湖南师范大学理学院计算研究所,湖南长沙,410081)

摘　要　针对采用数值分析方法进行数据拟合求解复杂度高、运算最大而精度较低的缺陷,本文给出一种基于二叉树编码的遗传算法来进行数据拟合,取得了较好的效果.

关键词　遗传算法　二叉树编码　数据拟合

DATA SIMULATION APPLYING GENETIC ALGORITHMOF BINARY-TREE CODING

Liu Dongbo　Gao Chuming
(Institute of Computer, Hunan Normal University, Hunan Changsha, 410081)

Abstract　This paper studies Data Simulation which is based on Genetic Algorithm solution method. Making use of the art of Genetic evolvement mechanism applying Binary- Tree Coding, we build a data analyzer that can do data simulation and some data mining works.

Keywords　Genetic Algorithm; Binary- Tree Coding; Data Simulation

1. 引　言

在科学实验和研究中,人们常常需要对收集到的数据进行分析,以期求得自变量 $x_1, x_2, \cdots, x_n$ 和因变量 y之间的近似解析表达式,即数据拟合.传统的方法通过求导和正规方程组来求解,特别是对于高阶多项式和多自变量(即多维空间)的情况,计算过程复杂、运算最大并且通常求得的结果精度不高.

鉴于上述考虑,我们尝试采用二叉树编码的遗传算法来进行数据拟合.充分利用遗传算法广搜索空间和只关注目标函数信息,而不需要推导过程及其它信息的特性[1][2],同时利用了二叉树的结构特点,极大地降低了问题的复杂性和运算量,并得到了较高的精度值.

2. 算法思想及其实现

2.1 总体思想

利用二叉树依据特定的遍历方式与函数表达式一一对应的特性,将满足一定条件的随机二叉树作为问题的一个解.同时利用二叉树的结构特点——任何结点的左、右子树仍是二叉树

* 湖南省教育厅科研项目(00C059),湖南师范大学 2001年重点科研项目资助
陈传淼教授推荐
收稿日期: 2001年 10月 30日

—— 构造特定的选择算子、杂交算子、变异算子.

首先产生一定数目的二叉树群体作为问题的初始解空间,然后对二叉树群体演化 n 代,每一次演化都经过评价、选择、杂交、变异过程,找到当前最好的二叉树个性,并将之对应的函数表达式作为最优解.同时,在演化过程中,为了保持个体的多样性,每演化一定代数 m (称之为演化周期)就新增加若干新个体(称之为外族个体)替代原有群体中其适应值处于整个建群体的中间部分的若干个体,以增加群体的多样性[3].参数的设置仍采用人工设置方式.算法的一般框架结构如下:

```
{
        t= 0;
        initialize p(t)
        while(t < n)
        {
          select p(t);
          if (t mod m= = 0)
          create nwe_ colony;
        intercross p(t);
        aberrance p(t);
        evaluate p(t) p
        t+ + ;
    }
}
```

2.2基本步骤

(1)初始化二叉树群体

产生一定数目的棵随机二叉树(一般取 100棵以上),放入二叉树组 a中.每棵二叉树满足下列要求:

每一结点由操作数和操作符组成.

操作符{一元运算符: fabs, exp, log, atan, sin, cos
二元运算符: + , - , * , /

操作数{变量: $x_1, x_2, \cdots, x_n$
常量: 随机产生的不大于某一数的随机实数

说明: 在实际当中变量的个数 n 需要根据具体情况而定;

分支结点必须为操作符,叶子结点必须为操作数;

二叉树的层数不能超过 layer层(layer一般取 8);

每棵二叉树的适应值存在(采用最小二乘误差),且要求其适应值不超过 1E+ 8.

适应值求法(适应值越小,则该二叉树越好):

$$\sum (f(x_1, x_2, x_3, \cdots, x_n) - y)^2$$

其中 f 为与二叉树对应的函数,y 为真实值.

(2)选择

I. 求出每棵二叉树的适应值,并将二叉树数组 a中的元素按适应值从小到大的顺序排

列;

Ⅱ.保留二叉树群体中最好的 bs棵和最坏的 bs棵(bs的值一般取群体规模的百分之五,我们称之为保留数),求出中间二叉树的选择概率、累积概率,然后根据累积概率确定中间的二叉树.方法如下;

为每一棵中间二叉树产生一随机小数 r,若 acc[i- 1]<i<acc[i](acc为存放累积概率的数组,i为当前二叉树在数组中的位置),则将第 i棵二叉树保留下来,经过这样的选择之后,产生的新群体中可能含有相同的二叉树,从而体现了优质个体在演化过程中看作有较强的生存能力.

选择概率的求法: 求出所有的中间二叉树的适应值的倒数总和 sum,则二叉树 a[i]的选择概率 select[i]为: (1/s[i])/sum(s[i]为第 i棵二叉树的适应值);

累积概率的求法:

$accu[i]=\sum select[j]$ (k<= j<= i,k为二叉树数组中第一棵中间二叉树的位置)

(3)群体更新

判断当前演化的代数是否为演化周期的整数倍.若是,则产生一定数目的满足条件的随机二叉树 (也即外族个体,一般为群体规模的五分之一)代替原来群体当中适应值处于中间部分的个体.

(4)杂交

Ⅰ.杂交树的选取:

给每棵二叉树产生 0~ 1之间的随机小数,若该随机小数小于杂交率,则选中该棵二叉树,然后对选中的二叉树序列按照选中的先后次序每相邻两棵进行杂交.若最后剩余一棵,则丢掉.

Ⅱ.两棵二叉树的杂交过程:

为两棵二叉树随机选择杂交点 (可以为二叉树中的任意结点).

将以两杂交点为根结点的二叉树相互交换;

若杂交后的二叉树超过 layer层,则截断,并且第 layer层的二叉树结点由产生的随机操作数代替;

若经过上述过程后,二叉树的适应值不存在或大于 1E+ 8,则丢掉该棵二叉树,并产生一棵符合要求的随机二叉树代替 (见初始化二叉树部分说明).

(5)变异

Ⅰ.变异树的选取: 给每棵二叉树产生一随机小数,若该随机小数小于变异率,则选中该棵二叉树,并对其进行变异操作;

Ⅱ.变异过程:

随机选择变异点 (可以为二叉树中中的任意结点);

变异点的处理:

ⅰ.若随机选择的变异点为叶子结点,则从下面两种处理方式中随机选择一种;

叶子结点{ 以任意的操作数代替该叶子结点;
随机产生一棵二叉树代替该叶子结点.

ⅱ.若随机选择的变异点为分枝结点,则根据其类型从该灯型的处理方式中随机选择一种进行处理.

分枝结点为一元运算符: 以不同于原来值的随机一元运算符代替.

分枝结点为二元运算符{ 以不同于原来值的随机二元运算符代替;
变换以变异点为根结点的左、右子树.

若结过变异后的二叉树的适应值不存在,或其适应值大于 1E+ 8,则丢掉该棵二叉树,并产生一棵适应值满足条件的随机二叉树代替 (见初始化二叉树部分说明).

Ⅲ.将二叉树群体按照适应值从小到大的顺序进行排序,并保留群体演化至此适应值最小的二叉树个体.然后对演化代数进行判断,若演化代数不少于 *n*,则转 (6),否则转 (2).

(6)报告结果

将演化 n代后的当前二叉树群体当中适应值最小的二叉树作为最优解,显示该二叉树对应的函数表达式、演化代数、参数设置及适应值的大小.

2.3 实验结果

为了验证算法的可行性,我们用 C语言将该算法编写成程序,并对以下三维空间上 20组实验点进行了测试 [4].实验点 (X, Y, Z):

(- 0.8, - 0.4, 0.447)　(- 0.4, - 0.4, 0.825)　(0.0, 0.8, 0.600)　(0.6, - 0.8, 0.000)
(- 0.8, - 0.2, 0.566)　(- 0.4, - 0.6, 0.693)　(0.0, 1.0, 0.000)　(0.6, - 0.6, 0.529)
(- 0.8, 0.0, 0.600)　(- 0.4, - 0.8, 0.477)　(0.2, - 0.8, 0.566)　(0.6, - 0.4, 0.693)
(- 0.8, - 0.2, 0.566)　(- 0.4, - 0.8, 0.566)　(0.2, - 0.6, 0.775)　(0.6, - 0.2, 0.775)
(- 0.8, - 0.4, 0.447)　(- 0.2, - 0.6, 0.775)　(0.2, - 0.4, 0.894)　(0.6, 0.0, 0.800)

在取参数 (群体规模= 500,二叉树最大深度= 8,二叉树的允许的最大适应值 1E+ 8,外族群体= 20,演化周期= 3,保留数= 5,杂交概率= 0.6,变异概率= 0.2)运行程序,演化第 20代得到结果,其数学表达式为: 0.628161+ x^2* y其适应值为: 7.90436E- 01.

3. 结束语

与用数值分析的方法进行数据拟合相比,采用基于二叉树编码的遗传算法极大的降低了问题的复杂度和运算量.同时,外族个体的引进及最优解的保留增加了群体的多样性,扩大了解空间,并保留了群体演化至此适应值最大的二叉树个体,使得下一代的最优个体总不比上一代差.另一方面,该算法在前 100代之内收敛较快,然而以后的演化过程中,算法收敛得并不理想,这是该算法的不足之处.在以后的研究中,我们将采用动态设置参数方法,使得参数随着遗传进程而自适应变化,从而提高算法的效率和个体解的全局最优性.

参考文献

[1] Mitchell M. An Introduction to Genetic Algorithms[M]. Cambridge, MA: MIT Press. 1996.
[2] Holland. J. H.基因算法 [J]. 科学, 1992, 1: 24- 31.
[3] 张晓缋,戴冠中,徐乃平. 遗传算法种群多样性的分析研究 [J]. 控制理论及应用, 1998, 15(1): 13- 18.
[4] 侯进军,熊令纯. 遗传程序设计在数据拟合中的应用 [J]. 长沙电力学院学报 (自然科学版), 1999, 14(2): 141- 144.