决策支持系统的自适应建模研究及应用实例分析①

葛志远，王永县，奕 晋
（清华大学经济管理学院，北京 100084）

摘要：在进行决策的过程中，建立正确的系统模型是解决问题的关键. 由于多方面的原因，有时要针对研究对象选用适用的模型来进行分析是很困难的. 本文根据遗传算法在程序设计中的应用，探讨了基于遗传算法的自适应建模方法，并对实际应用中自适应建模的算法存在的问题提出了改进方法. 最后应用自适应建模方法对全国的集装箱与外贸数据进行了建模分析，应用实例表明自适应建模方法对于解决难以选择合适的模型进行建模分析时是很有效的.

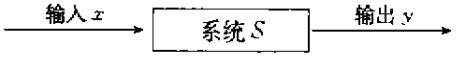
关键词：决策支持系统；自适应建模；遗传算法；二叉树结构
分类号：C934； **文献标识码**：A； **文章编号**：1007-9807(2000)01-0073-06

0 引 言

0.1 问题背景

在决策支持系统中的预测、计划、决策问题中，建立系统的模型是分析和解决问题的前提，系统常常需要对历史资料和现有的信息进行分析，建立起其量化分析模型，来分析系统的行为特性等. 现有的系统解决建模问题的方法有很多种，如回归分析、数据拟合及逼近论等，这时问题将转化为确定某一给定函数(模型)$f(x,\omega)$中的参数$\omega=(\omega_1,\omega_2,\cdots,\omega_m)$去拟合、插值或逼近已知数据对. 在这种情况下，模型的选择是至关重要的，不适合的模型很难很好地描述系统的状态行为，使建模过程失去意义. 例如，对于回归分析，要求数据有较好的分布规律，在实际的应用中还需要根据对象本身的变化特点，并且对分析结果进行评价，比较不同方法效果来选用适用的模型. 因此，这就要求决策者对于各类分析方法和研究对象有深入的了解，并掌握其应用背景，才能正确运用它来分析和解决问题. 然而，现实中很多的情况是：对于研究对象的了解很少，从而难以选择出合适的模型来对问题进行分析，有时即使对问题的背景知识比较了解，如何选择一个合适的模型和进行参数估计也是很困难的.

很多实际问题都可以归结为如下图所示的一个有输入和输出的系统：



对该系统，若已知有限个输入输出对$\{(x_i,y_i);\ i=1,2,\cdots,n\}$，要求模拟系统的行为，这就是一个建模问题. 本文重点讨论的是具有明确数学表达式的数学模型的建模问题.

0.2 问题描述

设$\{(x_i,y_i);\ x_i\in X,\ y_i\in Y,\ i\in I\}$是给定的输入输出对，$X,Y$都是实有限空间的子集，$I$是指标集；又若$F$是$C(X)$($X$上所有连续函数全体)的一个子集，$\rho$是定义在乘积空间$\prod_{i\in I}Y$上的一个距离，则建模问题便是确定一个函数$f^*\in F$使得对任意$f\in F$有

$$\rho(\{f^*(x_i)\},\{y_i\})\leqslant\rho(\{f(x_i)\},\{y_i\}) \quad (1)$$

成立，或使得对某给定的$\varepsilon>0$有

$$\rho(\{f^*(x_i)\},\{y_i\})\leqslant\varepsilon \quad (2)$$

实际上，建模问题可看成如下一个优化问题：

$$\min_{f\in F}\rho(\{f(x_i)\},\{y_i\}) \quad (3)$$

① 收稿日期：1999-05-12.
基金项目：95国家重点科技攻关项目(96-920-35-02).
作者简介：葛志远(1974-)，男(汉族)，湖南双峰县人，清华大学经济管理学院博士生.

遗传算法(genetic algorithms, GAs)是一种基于自然界生物机制的概率搜索方法. 作为一种全局优化搜索方法,遗传算法具有简单通用、鲁棒性强、适于并行处理以及应用范围广等特点,对于一些复杂问题,它表现出极强的解决问题的能力. 本文探讨利用遗传算法求解优化问题的能力,研究自适应建模问题.

1 自适应建模

1.1 函数模型表示

对于建模问题,搜索优化模型是在函数空间 F(称为模型族)上进行的,通常 F 可以有一些简单的函数经过有限次运算及符号生成的. 因此,F 中的函数可以像用基本初等函数表示初等函数一样用一些简单的函数及运算来表示它,而函数表达式可以树形结构来表示. 例如函数表达式 $a*(b+c/d)-(e**g)*f(x,y,z)$ 对应的一棵有序树如图 1 所示.

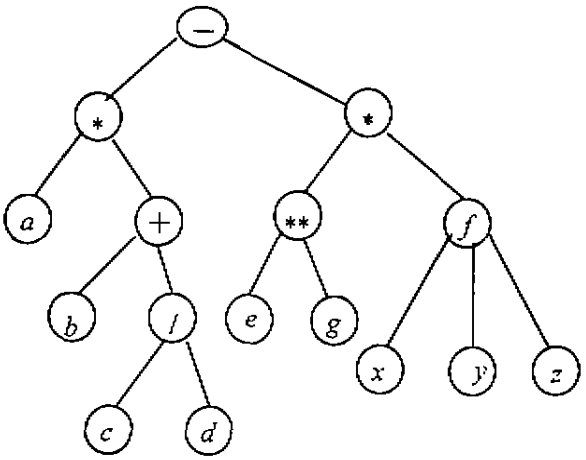


图 1 有序树表示示例

定义 1 设 F 是一函数空间,P 是 F 的一个子集,M 是有限个从 F 到 F 或从 $F\times F$ 到 F 的映射组成的集合,则称 F 到 F 的映射为 F 上的一元或二元运算组成的集合. 如果 F 中的每一个函数都可以通过 P 中的函数及有限此 M 中的运算得到,并且反之亦然,则称 P 是 F 的一个原型集,M 为 F 的一个运算集,记为 $F=\mathrm{Span}(P,M)$.

例如:若 $P=\{x,a;a$ 是 R^1 上的恒等函数,$a\in R^1$ 表示常函数$\}$,$M=\{+,-,\times,\div\}$,则 $\mathrm{Span}(P,M)$ 为所有有理函数全体.

从定义 1 易知,如果 P 是 F 的一个原型集,M 是 F 的一个运算集,则 F 中的每一个函数 f 都可以表示为集合 $P\cup M$ 中元素的一个串. 然而,$P\cup M$ 中元素组成的任何一个串却未必表示 F 中的一个函数.

定义 2 集合 $P\cup M$ 中元素的一个串若对应 F 中的一个函数 f,则称该串是 f 的串表示,若 f 的串表示的某个子串也对应 F 中的一个函数,则称该子串为 f 的一个合法子串.

易知,每一个函数 f 的串表示都对应一个二叉树,这些二叉树具有叶节点的元素属于 P 而非叶节点的元素属于 M.

定义 3 对应于 f 的串表示的二叉树称为 f 的(二叉)树表示.

由此,可以利用遗传算法的结构编码来求解式(3)所表述的建模问题.

1.2 基于遗传算法的自适应建模[4,5]

在以上分析的基础上,就式(3)所描述的优化问题进行研究,来设计遗传算法的基本算子对问题进行求解,提出自适应建模算法.

通常对模型 f 的评价是根据观测值 $\{y_i\}$ 与模型的期望值 $\{f(x_i)\}$ 的误差检验来度量的,因此式(3)中的目标函数直接采用模型 f 的方差:

$$q(f)=\rho(\{f(x_i)\},\{y_i\})=\sum_{i=1}^{n}[f(x_i)-y_i]^2 \quad (4)$$

遗传算法的编码:采用上面描述的二叉树结构式,每一个二叉树个体表示一个函数 f. 由于与标准遗传算法的编码不同,因此遗传算子的设计与标准遗传算子也有很大的不同. 下面对自适应建模过程中遗传算子进行描述.

适应性的度量:原始适应函数直接采用个体 f 的方差.

选择策略:采用线性排名选择策略. 若 $q(f_i(t))$ 是第 t 代时第 i 个模型 $f_i(t)$ 的目标函数值,$n_i(t)$ 表示 $\{i=1,2,\cdots,N\}$ 按从小到大 $q(f_i(t))$ 的排序序号,则分配给 $f_i(t)$ 的选择概率为

$$p(f_i,t)=\frac{2[N-n_i(t)+1]}{N(N+1)}$$

交叉算子:在两个父代个体中独立随机设定各自的交叉点,实行交叉时,两个个体以其交叉点作为根节点的部分子树进行交叉互换,生成两个

新个体,例如图2中所示的两个父代通过交叉算子获得两个子代.

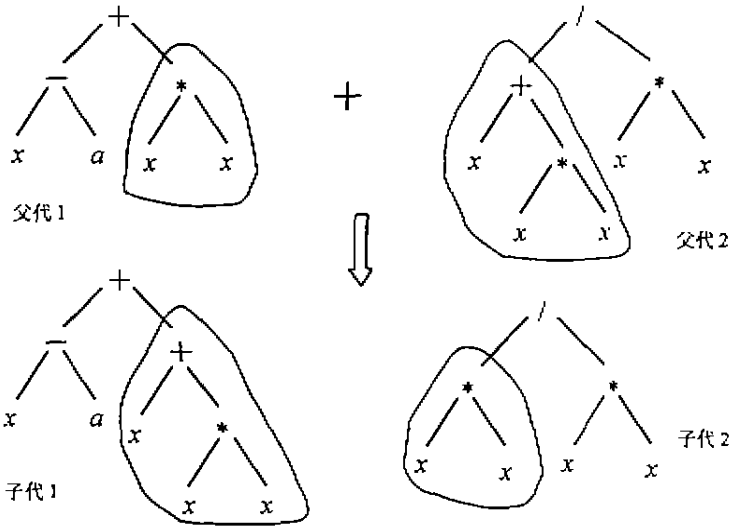


图2　交叉算子示例

实际上,这种交叉算子与标准遗传算子的交叉算子有类似之处.图中父代1的波兰表达式(后序)为 $xa-\underline{xx*}+$,父代2的波兰表达式为 $\underline{xxx*+}xx*/$,父代1中以交叉点为根节点的子串为 $xx*$,父代2中以交叉点为根节点的子串为 $xxx*+$,二者交叉的结果相当于将这两个子串进行切断和拼接,重新组合为两个新的个体.

变异算子:在父代个体中随机选择变异点进行变异,如图3所示的例子.

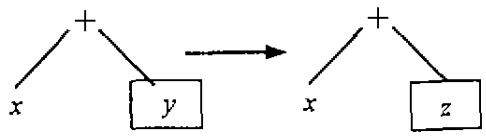


图3　变异算子示例

算法结构:上述对树结构的遗传操作实际上是基本遗传操作的一种扩展,其遗传操作的完整算法框架与标准的遗传算法基本类似.如采用非重叠种群的算法:

```
{
随机初始化种群 P(0),t = 0;
计算 P(0) 中个体的适应值;
While(不满足终止准则)do
    {
    根据个体的适应值及选择策略,计算种群内个体的选择概率 P_i;
    根据 P_i 从 P(t) 中选择个体进行遗传操作,用后代替换整个群体,产生新种群 P(t+1);
    计算 P(t+1) 中个体的适应值,t = t+1;
    }
}
```

2　算法的改进

通过试验和实际的应用,上述定义的基于遗传算法能够解决实际应用中的建模问题,并且与传统的解决建模问题的方法相比有很多的优点,表现出了很大的应用潜力,文[6]对此作了一些的研究.但其遗传算子是从标准遗传算法扩展而来,由于结构编码与标准编码有很大的不同,对遗传操作功能有很大的影响,在应用时表现出很多的缺点和局限性,如很容易出现未成熟收敛现象,获得的有较好结果的函数 f 的表达式太长而过于复杂,并且不符合实际,等等问题,因此需要对其进行改进.在上述定义了基于遗传算法的自适应建模方法的基础上,作了深入的研究与试验,对算法进行了改进,使算法在性能与应用上都取得了进展.

1) 对结构编码的修改:对树结构进行限制,将树结构限制在一定的深度之内,即树结构有一个最大深度.实际上,如果不限制树结构的深度的

话，遗传算法的搜索是在一个无限空间中进行的，因为树的深度是可以无限大. 限制了树的最大深度后，遗传算法的搜索就限制在一个有限空间中进行了.

2) 对目标函数的修改：添加控制树结构复杂程度的参量. 如果通过遗传算法搜索得到的函数 f 的表达式过于复杂，在一般情况下很难反映实际的问题，复杂的模型对于分析原问题没有什么帮助. 实际上这就是建模结果的模型评价问题，得到的模型的好坏，除了模型应该有较小的误差外，模型的简洁性也是一个评价因素. 因此，在目标函数中添加反映模型复杂度的参量 $C*N(f)$，其中 C 为一个常数，$N(f)$ 表示 f 所对应的树的结点的个数. 实际上，模型复杂度的参量 $C*N(f)$ 相当于每一个二叉树的节点都具有一定的能量，增加一个节点需要消耗系统的能量才能实现. 添加了这一项是为了使系统获得的结果不需要消耗较多的能量，即最终模型不会太复杂.

在优化的开始阶段，个体 f 的误差都比较大，此时模型的复杂度不成为主要的评价因素，当进行到优化的后期阶段时，大部分个体的最小二乘误差均有较大的改善，此时模型的复杂度成为一个重要的评价因素. 但随着遗传算法的进化，获得的个体适应值越来越好，即目标函数的误差越来越小，如果上述的模型复杂度参量 $C*N(f)$ 保持一定的值不变，会使得算法停留在一个简单但不太理想的个体上，从而阻碍算法的继续寻优. 因此，将算法的复杂度参量进一步修正为一个自适应的值，修正的方案是在复杂度参量基础上乘一个修正因子 $u(t)$，$u(t)$ 表示第 t 代时模型复杂度的修正因子，随着算法的进化而变化. 此时，模型复杂度参量为 $C*N(f)*u(t)$，修正因子 $u(t)$ 用以调整不同阶段时模型复杂度的评价值. $u(t)$ 还要使得模型复杂度的修正值与模型的最小二乘误差匹配. $u(t)$ 可取为线性函数或者与最优个体的最小二乘误差相关取值.

同时，在计算过程中所产生的函数有可能没有定义或者计算时溢出，比如对负数取对数或者指数函数的指数很大，目标函数也需要处理这些情况. 因此，在目标函数中添加一惩罚项 $g(f)$，当 $g(f)$ 充分大时可使得这些意外的个体被淘汰.

这样，目标函数由式(4)修正为：

$$q(f)=\rho(\{f(x_i)\},\{y_i\})=\sum_{i=1}^{n}\|f(x_i)-y_i\|^2+C*N(f)*u(t)-g(f) \tag{5}$$

3) 选择策略的修改：为了保证算法的收敛性，与标准的遗传算法的改进类似，也采用在选择后保留当前最好解的选择策略.

4) 交叉算子的修改：由于对树结构有了深度的限制，因此在交叉的结果也要求满足这个限制，杂交点的选择要使得两个后代模型对应的树结构的深度不超过最大深度的限制.

5) 变异算子的修改：上面所定义的变异算子是随机选择树的一个结点，然后用其它的元素替代. 但这样进行的变异有可能产生不合法的子串，因此需要进行修正. 树的内部结点(即非叶结点)由 F 的运算集 M 中的元素组成，叶结点由 F 的原型集 P 中的元素组成，在变异时需要区分这两种结点. 变异树的内部结点有多种，如下所示：

重新产生新的子树来替代旧子树；

交换结点的左右子树；

对以结点为根节点的子树进行化简等.

对子树的化简一般来说比较复杂，化简包括合并同类项、计算结果、简化互逆函数(如 $\exp(\ln x)=x$)等，由于遗传算法搜索的遍历性，树的化简可以略去，也可以保留以加快进化的速度. 可以以等概率的方法执行上述树的内部结点的变异操作.

变异叶结点也有多种，如果元素为变量，可以用 P 中其它的元素替换；如果元素为数 a，采用非一致性变异，即变异策略为

$$a'=\begin{cases}a+\Delta(t,UB_k-a) & \text{if and}(2)=0\\ a-\Delta(t,a-LB_k) & \text{if and}(2)=1\end{cases}$$

其中 $[LB_k,UB_k]$ 为 a 的取值范围，t 为当前代数，$\text{rand}(n)$ 为随机生成从 0 到 $n-1$ 的整数，函数 $\Delta(t,y)$ 的值域为 $[0,y]$，可取

$$\Delta(t,y)=y*(1-r^{(1-t/T)^2})$$

r 为随机数，$0\leqslant r\leqslant 1$，T 为最大代数，t 为当前代数. 可以看出随着 t 的增加，$\Delta(t,y)$ 减少. 当 t 小的时候，这种变异可以均匀地搜索整个空间，它能体现一种自适应的遗传性，即在一定代数后的子孙不会离它太远.

3 应　用

利用上述的自适应建模的方法对全国港口进出口集装箱量与全国进出口贸易额的关系,根据历史数据进行建模分析.其数据列入表1中,其原始数据分布如图4中的曲线C所示.

表1　进出口集装箱量和外贸进出口额数据

年份	外贸进出口额(十亿元)	进出口集装箱量(万TEU)
1985	69.60	50.313 3
1986	73.85	63.058 3
1987	82.65	70.070 9
1988	102.79	96.490 4
1989	111.68	117.685 4
1990	115.44	142.722 6
1991	135.63	204.900 0
1992	165.53	259.544 6
1993	195.70	335.325 2
1994	236.62	436.789 9
1995	280.85	608.997 3
1996	289.90	771.400 0
1997	325.06	983.700 0

数据来源:中国交通统计年鉴

利用一元线性回归分析获得的模型为 $f(x) = 3.286 * x - 233.829$(其中,$x$ 表述全国进出口贸易额,$f(x)$ 表示全国港口进出口贸易集装箱量的回归曲线),其图形如图4中的直线L所示.比较上面图中两条曲线,可看出,一元线性回归模型并不能很好地反映实际的模型,其误差较大.为了寻找二者之间的关系,采用了前面所述的自适应建模方法.

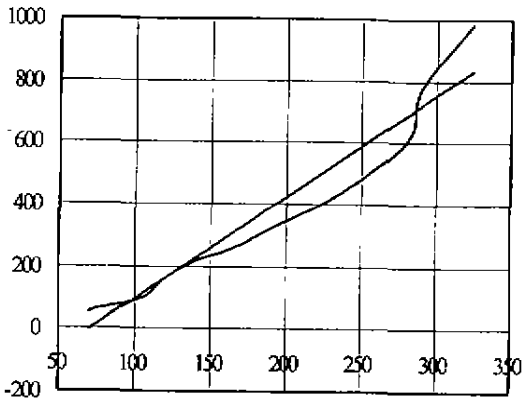


图4　一元线性回归分析结果

在研究中,算法采用的是通过改进后的基于遗传算法的自适应建模.选取 $P = \{x, a; a \in R^1\}$, $M = \{+, -, \times, \div, \mathrm{SIN}, \mathrm{COS}, \mathrm{EXP}, \mathrm{LN}\}$,这样以 P 为原型集,M 为运算集所生成的模型族 $\mathrm{span}(P, M)$ 是有理函数集合的一个子集.

应用实例中的选择策略、交叉算子、变异算子均采用改进后的算法.初始群体随机生成,根据模型的编码表述,每个模型对应一个二叉树,树的中间节点和叶节点的元素从运算集和原型集中按均匀分布随机选取.模型的控制参数取:群体大小 $N = 100$,最大代数 $T = 500$,繁殖概率 $p_r = 0.1$,杂交概率 $p_c = 0.6$,变异概率 $p_m = 0.3$,树的最大深度 $D = 6$,复杂度系数 $C = 400$,自适应修正因子 $u(t) = 1 - t/(1.1 * T)$.下面的一些模型是通过基于遗传算法的自适应建模获得的一些模型:

$$f_1(x) = 28.196\ 2 + 0.007\ 949\ 83\ x^2$$

$$f_2(x) = 63.375\ 4 - \frac{4\ 262.004\ 9}{x} + 0.007\ 949\ 83\ x^2$$

$$f_3(x) = 29.354\ 3 + 0.007\ 491\ 57 x^2 - \frac{x}{-676.07 - 0.002\ 212\ 22(-426.343\ 6 + x) + 0.008\ 009\ 41\ x^2}$$

这些模型的最大误差 $u_{\max}$、最小误差 $u_{\min}$、最大相对误差 $r_{\max}$、最小相对误差 $r_{\min}$、最小二乘误差 q 等数据分别列于表2中.

表2　模型的比较

模型	q	$u_{\max}$	$u_{\min}$	$r_{\max}$	$r_{\min}$
$f(x)$	58 348.853 0	149.400 2	2.778 0	1.101 9	0.019 5
$f_1(x)$	11 137.092 9	75.084 1	2.662 5	0.325 8	0.007 9
$f_2(x)$	11 644.712 6	66.259 5	3.956 9	0.222 6	0.015 8
$f_3(x)$	3 597.833 0	37.477 7	2.691 3	0.306 9	0.003 5

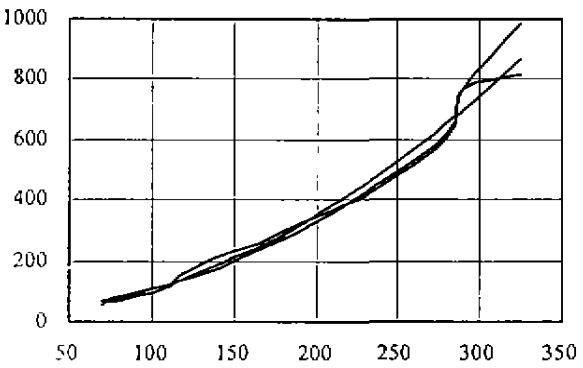


图 5 自适应建模结果

比较以上几个模型，非线性模型 $f_1(x)$ 比线性回归模型 $f(x)$ 获得的结果更好. 其中 $f_1(x)$ 和 $f_2(x)$ 如图 5 中 F_1、F_2 曲线所示. 从模型分析，进出口集装箱量与外贸进出口额呈二次非线性关系. $f_2(x)$ 增加了一个修正项，使得模型的误差指标中除了最大相对误差外，其余指标比其他模型好. 实际上，$f_3(x)$ 模型中的最大相对误差是 1985 年时的误差，1986 年以后的最大相对误差为 0.182 9，说明集装箱运输市场早期的不成熟. 更进一步的分析说明，我国的集装箱运输市场是一个新兴的市场，市场从 80 年代中期开始开拓，外贸运输中适合集装箱运输的货物采用集装箱运输方式的比例越来越大，这同我国改革开放以来，经济的发展和技术的进步是密切相关的.

4 结论

通过上面的实例，说明了基于遗传算法的自适应建模方法与传统建模方法相比，具有很多的优点. 自适应建模不需要建模者具有很强的专业背景知识，就能够建立较好的系统模型，并且在建模过程中可以提供多个较好的模型供决策者进行分析和参考. 但由于自适应建模方法的研究还不成熟，在实际的应用中存在一些问题. 首先对于结构编码的遗传算法本身的收敛性分析在理论上尚未得到很好的证明，而且自适应建模的结果是提供多个模型供分析和参考，对这些模型缺乏较好的评价标准. 如何判断和选择自适应建模所提供的模型，以及模型是否能够很好地反映系统等方面的问题，需要进行进一步的研究.

参 考 文 献：

[1] Dasgupta D. Michalewicz Z (Editors). Evolutionary algorithms in engineering applications[M]. Berlin Heideberg Springer-Verlag, 1997

[2] Herrera F. Verdegay J L., eds. Genetic algorithms and soft computing[M]. Heidelberg : Physica-Verlag, 1996

[3] Angeline P J, Kinnear K E(Editors). Advances in genetic programming Ⅱ[M]. MIT Press, Cambridge, MA. 1996

[4] Koza J R. Genetic Programming[M]. Cambridge, : MIT Press. MA. 1992

[5] Koza J R. Genetic Programming —2[M]. Cambridge: MIT Press. MA. 1994

[6] 潘正君等. 演化计算[M]. 北京：清华大学出版社，南宁：广西科学技术出版社. 1998

A study on adaptive modeling of DSS and its application

GE Zhi-yuan, WANG Yong-xian, YI Jin

School of Economics and Management, Tsinghua University, Beijing 100084

Abstract: This paper studies adaptive modeling on genetic algorithms. The modeling problem can be regarded as an optimizing problem with a simple conversion. Based on the research of genetic algorithms in automatic programming, this paper discusses adaptive modeling method on genetic algorithms, and puts forward the improved algorithms of adaptive modeling. Finally, a concrete case using adaptive modeling on genetic algorithms is presented.

Keywords: DSS, adaptive modeling, genetic algorithms, bintree