文章编号: 1001-0920(2001)04-0480-04

禁忌-递阶遗传算法研究

柯 珂, 张世英
(天津大学 管理学院, 天津 300072)

摘 要: 结合禁忌搜索算法和递推遗传算法提出一种新的遗传算法——禁忌-递阶遗传算法, 该算法能在一定程度上克服早熟问题。通过求解复杂的非线性系统——分整增广 GARCH-M 模型的参数优化问题, 表明该算法具有较高的精确性和可行性。

关键词: 遗传算法; 早熟; GARCH 模型
中图分类号: TP 18 **文献标识码**: A

Research on the Tabu-Hierarchy Genetic Algorithm

KE Ke, ZHANG Shi-ying
(Management School of Tianjin University, Tianjin 300072, China)

Abstract: A new genetic algorithm based on the integrating tabu search algorithm and hierarchy genetic algorithm is proposed. The new genetie algorithm —— Tabu-Hierarchy Genetic Algorithm (THGA) can avoid the "premature" problem to a certain extent. The parameters optimization problem of a complicated nonlinear system-fractional integrated augmented GARCH-M model is solved by THGA, which shows that the algorithm has better accuracy and feasibility.

Key words: genetic algorithm; premature; GARCH model

1 引 言

为了克服遗传算法中的早熟问题, 并提高遗传算法的运算效率, 许多学者已经做了大量工作, 并提出一些改进方法。禁忌遗传算法(TSGA)就是其中一种重要方法。它根据禁忌搜索算法的思想, 对遗传算法中进化时的核心算子——交叉算子和变异算子进行改造。禁忌搜索本身就是一种有效克服局部最优的通用搜索方法, 它能有效地利用全局信息和在搜索中获得的信息, 因此求解过程较快。Glover 阐述了禁忌算法的基本原理[1,2], Faigle 和 Kern 用类似于模拟退火的方法对禁忌搜索进行数学描述, 并证明了结果的收敛性[3]。实际算例表明, 禁忌遗传算法与简单遗传算法相比, 无论是解的质量还是搜索的效率都有一定程度的提高。

另一种改进的遗传算法是递阶遗传算法[4]。它是基于下述事实提出的: 生物的染色体是由基因的变化组合形成的, 基因以一定的层次方式排列, 分为控制序列基因和结构基因。控制基因的作用在于控制构造基因是否被激活, 被激活的构造基因称为显性基因, 未被激活的基因称为隐性基因, 只有显性基因才能表现生物体特征。染色体按这种层次结构进行编码的 GA 称为 HGA, 它能同时表示解的拓扑结构和参数。递阶遗传算法的染色体中同时包含了激

收稿日期: 2000-05-19; **修回日期**: 2000-08-28
基金项目: 国家自然科学基金项目(69874028)
作者简介: 柯珂(1974—), 女, 山东淄博人, 硕士, 从事系统控制与决策的研究; 张世英(1936—), 男, 北京人, 教授, 博士生导师, 从事复杂系统控制与决策、金融计算等研究。

活基因和隐性基因。隐性基因与激活基因一起遗传给下一代, 并在遗传过程中有可能被激活。这样, 递阶遗传算法便能在一定程度上避免早熟。

可见, 禁忌遗传算法对遗传算法的改进是针对遗传算子, 并不改变算法的编码结构; 而递阶遗传算法中不需要重新定义遗传操作, 它是通过改善遗传算法中染色体的编码结构来避免早熟, 并提高算法的效率。二者在机理上存在互补性, 它们的相互借鉴将比单个算法性能有一定提高。基于这一思想, 我们构造了一种混合算法—— 禁忌-递阶遗传算法(THGA), 即在基本遗传算法的基础上, 编码结构采用递阶结构, 并在遗传算法各操作算子中融入禁忌算法的思想。从而尽可能地避免算法早熟, 并提高算法效率。

2　禁忌-递阶遗传算法的实现

2 1　THGA 编码机制

由 Holland 提出的简单遗传算法采用二进制编码, 需要将求解问题的参数表示为一定长度的二进制数串, 而用遗传算法进行实数参数的优化时, 由于受编码长度和解码数制转化精度的限制, 遗传算法的性能会有所下降。所以算法中除控制基因外, 都设计为十进制编码。实数编码直接采用实数组成染色体, 同时对应不同的遗传操作算子, 可在连续的参数空间内进行搜索, 有利于提高算法的收敛速度和稳定性。

THGA 染色体由两部分构成: ① 控制基因: 由二进制构成; ② 参数基因: THGA 的多级递阶遗传算法染色体结构如图 1 所示。

图 1 中从上到下排列为第 1 级, 第 2 级, 第 3 级。处于第 1 级的控制基因(CG-1) 控制下一级的控制基因(CG-2), 按此一层层控制, 最后一级控制基因直接控制参数基因(PG) 的活动性。在基因编码时, 控制基因通常采用二进制编码, "1"表示对应的基因处于激活状态, 与该基因相联系的低级基因串则处于有效状态; "0"表示对应的基因处于休眠状态, 与该基因联系的低级基因串则处于无效状态。

这样, THGA 的染色体中便同时包括激活基因(对应于生物学中的显性基因) 和休眠基因(对应于隐性基因)。解码显性基因得到的是所求问题的解, 而休眠基因则存在于染色体中, 并随显性基因遗传给下一代。进化过程中子代继承的隐性基因可能被激活, 而显性基因则可能进入休眠状态。染色体的这种进化方式在一定程度上避免了早熟。

2 2　遗传操作算子

遗传基本操作算子包括选择复制、交叉和变异, 它们分别模仿了自然界生物繁衍、交配和基因突变。我们在此基础上引入一个短期记忆装置, 即长度为 L 的禁忌表, 表中记录了最近进行的 L 个移动。由于这些移动在目前的迭代中是被禁止的, 所以 L 称为禁忌长度。根据这一原则, 我们重新定义了各操作算子, 分别记作禁忌交叉算子(TSCO) 和禁忌变异算子(TSMO)。由于禁忌搜索的引入, 使得 THGA 拥有了记忆功能, 从而具有较强的"爬山"能力(相对于全局最小值问题)。它能在搜索过程中跳出局部最优解, 转向其它区域进行搜索, 从而使获得更好解的概率大大增加。

1) 选择复制算子根据种群中染色体适应度的不同来分配繁殖机会, 适应度高的个体产生子代的机会就多, 而适应度低的个体繁殖受到抑制甚至被淘汰。轮赌法是一种经典的复制算子, 此外还有排序选择、随机选择、期望值模型选择等复制算子。本文采用的工具包实现了轮赌法和排序选择两种复制算子。

2) 禁忌交叉算子模仿了父代基因的混合。交叉

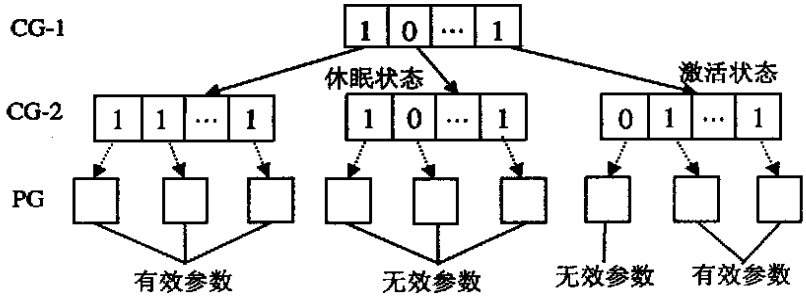

图 1　THGA 染色体结构图

算子根据交叉率 P_c 将种群中的两个个体随机地交换某些基因, 能够产生新的基因组合, 期望将有益基因组合在一起。常用的交叉算子包括一点交叉、两点交叉、多点交叉和均匀交叉。本文算法中实现了实数编码的算数交叉和混合交叉。

3) 变异算子模仿了生物繁殖过程中的基因突变。基因突变是小概率事件, 其方向是不固定的。变异算子根据一定的变异概率, 在染色体上随机选择一位基因, 然后改变该基因的特征。对二进制算子常用的变异算子有翻转变异和交换变异; 对实数编码的染色体通常采用高斯变异, 即根据变异率 P_m 在基因上加一个正态分布的随机数。本文算法使用的是实数编码的高斯变异。

禁忌变异算子的禁忌表定义如下: 在每个染色体中增设 L 个信息位, 存储最近 L 个发生变异的信息位的号码。每次发生变异时, 都将变异信息位的号码与禁忌表中的号码进行比较, 若在表中, 则属于禁忌范围; 若变异后的个体的适应值大于可望水平, 则可进入下一代。

2.3 适应值评价函数

适应值与遗传算法要解决问题的目标函数不同。通常, 目标函数值经过标定成为适应值。适当的标定可以防止遗传算法的早熟并改善其性能。常用的标定方法有模拟退火法、线性标定和 Sigmoid 标定等。

遗传算法的参数主要有种群规模、交叉率和变异率。种群规模与求解问题的非线性程度和参数规模有关, 求解问题非线性程度越强, 参数越多, 种群规模也越大。交叉率 P_c 一般取(0.75, 0.95), 变异率 P_m 一般取(0.005, 0.01)[5]。

3 实证研究

本文提出的分整增广 GARCH-M 模型如下

$$x_t = \mu_t(x_{t-1}, \sigma_t^2) + \epsilon_t \tag{1}$$

其中

$$\mu_t(x_{t-1}, \sigma_t^2) = \mu_0 + x_{t-1}[\mu_1 + \mu_2 \exp(-\sigma_t^2/\mu^2)] + \mu_3\sigma_t^2 \tag{2}$$

$$\epsilon_t = e_t\sigma_t, \quad e_t | F_{t-1} \sim N(0,1) \tag{3}$$

这里

$$(1-L)^d \epsilon_t^2 = \begin{cases} |\lambda\phi_t - \lambda + 1|^{1/\lambda} + v_t, & \lambda \neq 0 \\ \exp(\phi_t - 1) + v_t, & \lambda = 0 \end{cases} \tag{4}$$

其中

$$v_t = \epsilon_t^2 - \sigma_t^2 \tag{5}$$

取

$$\phi_t = \alpha_0 + \sum_{i=1}^{p} \alpha_1^{(i)} \phi_{t-i} + \sum_{j=1}^{q} [\alpha_2^{(j)} |e_{t-j} - c|^{\sigma} + \alpha_3^{(j)} \max(0, c - e_{t-j})^{\sigma}] \phi_{t-j} + \sum_{j=1}^{q} [\alpha_4^{(j)} f(|e_{t-j} - c|, \delta) + \alpha_5^{(j)} f(\max(0, c - e_{t-j}), \delta)] \tag{6}$$

$$f(z, \delta) = (z^{\delta} - 1)/\delta, \quad z \geq 0 \tag{7}$$

其中, L 表示滞后算子, μ^2 是 x_t^2 的样本均值, μ_1 为一阶自相关系数, $\mu_3 \neq 0$ 为 ARCH-M 模型, $\mu_3 = 0$ 为 ARCH 模型。

3.1 编码机制和适应值函数

在编码时, 专门设了一个 9 位二进制码构成递阶结构中的控制基因, 第 1 位控制 6 个参变量, 第 2 到第 9 位分别控制 8 个参变量, 这 14 个参变量都是结构基因。面向分整增广 GARCH-M 模型的递阶染色体结构如图 2 所示。

图 2 中 Par 为参变量, 其下标按顺序对应于 μ_0, μ_1, μ_2, μ_3, α_0, α_1, α_2, α_3, α_4, α_5, c, λ, σ 和 d。由于前 6 个变量的取值与模型的确立关系不大, 所以控制基因的第 1 位设为常数 1, 即前 6 个参数永远是激活的。其它控制基因位可取 1 或 0, 当某位取 0 时, 它控制的参数基因处于休眠状态, 意味着该参数也取 0, 无需估计; 反之, 当某位为1时, 它控制的参数是被激

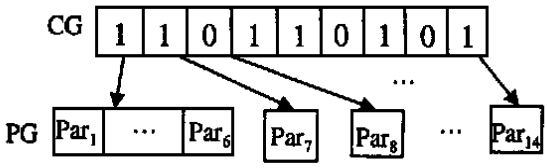

图 2 面向分整增广 GARCH-M 模型的递阶染色体结构

活的, 也就是要估计的参数。

我们编写的 ARCH 模型软件包中, 设置的适应值函数为 fitness = $N - \gamma$。其中 γ 为变异系数, 定义如下

$$\gamma = \sum_{t=1}^{N} (x_t - \hat{x}_t)^2 / \bar{x} \tag{8}$$

目标函数的优化方向为 γ 最小, 对应于适应值的增大方向。N 为任意大实数, 不失一般性, 这里设 $N = 100$。

3.2 运算过程

原始数据是 1997 年 1 月 2 日到 1998 年 6 月 15 日的英镑兑美元的日汇率收益。

算子设置包括选择复制算子、禁忌交叉算子、禁

忌变异算子和杰出个体保护算子; 参数编码采用递阶实数染色体编码。各控制参数分别选作: 种群容量为 10, 交叉率 P_c = 0.7, 变异率 P_m = 0.005。

结束条件: 由于 GA 的收敛判据是启发式的, 因此用 γ 的满意值和运行的代数联合作为控制算法的结束条件。这里取最大代数为400, γ 的满意值为 1.5。

3.3 结果及分析

拟合的优化过程如图 3 所示。其中图 3(a) 为每一代的最优适应值, 图 3(b) 为平均适应值。适应值的最大值为 100, 此为理想状况, 只有当估计值与实际值相同时才能得到。

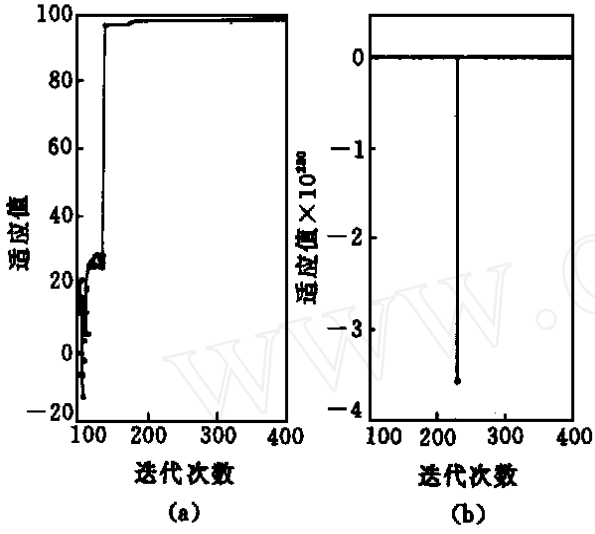

图 3　THGA 算法的适应值变化图

图 3(a) 中, 第 150 代左右适应值急剧增长, 体现了 GA 的强劲搜索能力和优化效率, 此后一直到 400 代, 适应值呈稳中有升的态势, 体现了禁忌算法和递阶算法相辅相成的作用, 在一定程度上减少了早熟的发生。图 3(b) 中, 因为表示的是每一代种群适应值的平均值, 如果这一代中有非常小的点, 则其平均值就会非常小, 所以在 220 代左右出现了一个非常小的点。这也从另一方面说明禁忌-递阶遗传算法会跳出原搜索空间, 转而向其它空间搜索, 虽然可能暂时出现极差的点, 却在一定程度上克服了早熟的弊病。

控制基因为 110001111, 解码结果为参数基因 Par8, Par9 和 Par10 是隐性基因, 所以有 $\alpha_3 = \alpha_4 = \alpha_5 = 0$。

最终确定的模型形式如下

$$x_t = -0.1637 + x_{t-1}[0.5496 + 0.5481\exp(-\sigma_t^2/2\ 695\ 1)] + 0.4240\sigma_t^2 + \epsilon_t$$

其中

$$\epsilon_t = e_t\sigma_t,\quad e_t \mid F_{t-1} \sim N(0,1)$$

$$(1-0.0683L-0.3179L(R^{0.4521})^2)\cdot(1-L)^{0.1341}\epsilon_t^2 = 0.0250 + (1-0.0683L)v_t$$

R 是平滑算子, 即 $(R^{0.4521})^2\epsilon_t^2 = (\epsilon_t - 0.4521)^2$。

上述模型为分整非线形 GARCH-M 模型。由于 d = 0.1341, 不显著为零, 说明该时间序列的二阶矩有长记忆性。可以证明, 该模型的拟合效果非常好, 与实际数据非常接近。

参考文献:

[1] Glover F. Future paths for integer programming and links to artificial intelligence[J]. Comper & Operational Research, 1986, 13(5): 533-549.

[2] Glover F. Tabu search-Part I[J]. ORSA J on Computing, 1989, 1(3): 190-206.

[3] Faigle V, Kern W. Some convergence results for probabilistic tabu search[J]. ORSA on Computing, 1992, 4(1): 32-37.

[4] 郑丕谔, 马艳华. RBF 神经网络的递阶遗传训练新方法[J]. 控制与决策, 2000, 15(2): 165-168.

[5] Schaffer J. A study of control parameters affecting online performance of genetic algorithms for function optimization[A]. Proc 3rd Conf Genetic Algorithms[C]. 1989. 51-56.