考虑基因表达过程的进化算法

周 晴,李衍达
(清华大学自动化系,北京 100084)

摘 要: 生物的进化过程是在其基因层与表型层上同时进行的. 表型层上的进化是以环境为参考的自然选择过程,而在基因层上则是一个带随机性的自我更新、自我优化的过程,而且在某种程度上具有自组织趋向. 基于这种新的进化观点,本文提出了一种新的进化算法并将其应用于各种函数优化问题中. 此算法不但考虑了表型层上的自然选择作用,还考虑了生物在基因层上的进化过程及两个层次间的相互映射关系. 仿真结果表明,此算法不论在收敛速度、参数鲁棒性还是全局搜索能力上,都优于传统框架下的进化算法.

关键词: 进化算法; 遗传算法; 全局优化
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2002) 01-0114-04

An Evolutionary Algorithm Considering Gene Expression

ZHOU Qing, LI Yan-da
(Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract: The evolution of organisms takes place both in the genotype and the phenotype. At the phenotype level, the evolution is based on the laws of natural selection, while at the genotype level, it is a stochastic process of self-updating and self-optimization, and has a tendency of self-organizing to some degree. Based on this viewpoint of evolution, this paper proposes a new evolutionary algorithm and applies it to implement optimization. The algorithm considers not only natural selection of the phenotype, but also the evolution of genotype and the mappings between them. It has been demonstrated by simulation results that the algorithm is much better than other evolutionary algorithms based on typical frames in the aspects of convergence speed, robustness and global optimization ability.

Key words: evolutionary algorithm; genetic algorithm; global optimization

1 引言

遗传算法(GA),进化策略(ES)等进化算法借鉴生物进化中的自然选择法则,利用选择、变异等操作模拟生物进化过程以解决各种优化问题,在多峰函数的优化,神经网络的结构优化,控制器的参数优化,IIR 滤波器的设计与辨识等方面取得了十分广泛的应用[1~4]. 但是这些优化方法存在着以下缺陷: (1) 容易出现过早收敛,从而陷入局部极值点; (2) 算法后期搜索效率偏低; (3) 对程序设定的初始参数较为敏感,即参数鲁棒性较差. 为了克服上述缺点,基于 GA 和 ES 的各种改进算法层出不穷,将局部搜索方法、模拟退火机制及自适应调整交叉变异概率等思想结合到进化算法中,取得了一定的效果[5~8]. 近年来,生物进化研究有了新的进展. 对生物进化的认识不会再停留于自然选择机制之上. 我们认为,生物进化过程是在基因层与表型层上同时进行的. 表型层上的进化是以环境为参考,遵循自然选择的机制;而在基因层上的进化是一个带随机性的自我更新的过程,是自参考的. 在某种程度上可以认为基因层上的进化可能出现某种程度的自组织过程. 表型层与基因层的相互映射关系使得生物个体成为一个完整的系统而不断进化. 显然,传统的进化算法是不足以表示这种全新的进化观点. 基于新的进化理论,本文提出了一种新的进化算法:基因表达进化算法(Gene Expression Evolutionary Algorithm,简称 GEEA). 它不但考虑了生物个体在表型层上的自然选择机制,还考虑了表型层与基因层之间的相互映射关系以及基因层上自组织的定向突变. 实验结果表明,此算法的综合性能优于传统框架下的进化算法.

2 算法的基本模型

算法的基本模型可以用图 1 简明的表示. 在图中生物个体的进化体现在两个层次和四种映射关系上. 两个层次是指基因层和表型层,分别用状态空间 G 和状态空间 X 表示. 基因层状态空间 G 和表型层状态空间 X 是生物进化过程固有的两个层次. G 空间中的向量 $g = [g_1, g_2, \cdots, g_m]^T (m > 0)$,代表了此生物个体染色体上的各段基因. X 空间中的向量 $x = [x_1, x_2, \cdots, x_N]^T (N > 0)$,表示一个种群. 它的各个分量表示此种群中某一个体的表型模式,这些分量均是很多基因共

收稿日期:2000-09-08;修回日期:2001-06-26
基金项目:国家自然科学基金(No. 69935020)

同表达的结果. 四种映射关系分别是基因表达过程、表型层上的自然选择、表型特征在基因上的固化和基因层的自参考突变作用,分别用 f_1, f_2, f_3 和 f_4 表示. 它们发生在表型层与基因层中,共同作用以推动生物个体不断演变进化. 以往的进化算法的模型实际上只有图 1 中的右半部分,并未真正考虑基因层的自参考突变过程.

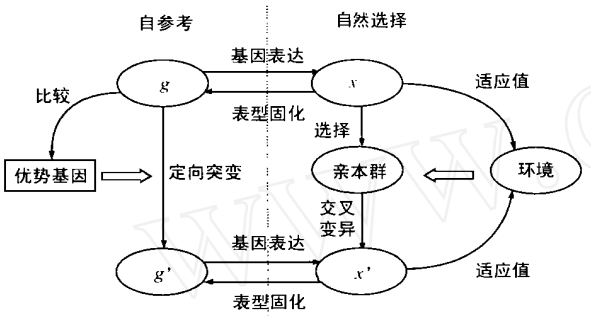

图 1　算法模型简图

2.1　基因表达

基因表达是一个由基因层到表型层的复杂过程,任何一种表型模式都是该个体许多基因共同表达的结果. 这个过程可以抽象为由 G 空间到 X 空间的一种映射关系 f_1,即

$$f_1:\quad G \longrightarrow X \qquad (1)$$

算法在实现此映射关系时,定义了基因表达矩阵 S.

定义 1　称 $N \times m$ 维的矩阵 S 是从 G 空间到 X 空间的基因表达矩阵,如果它满足

(1) $S_{ij} \geq 0$, $(i = 1, 2, \cdots, N; j = 1, 2, \cdots, m)$

(2) $Sg = x$, $(g \in G, x \in X)$

基因表达矩阵实际上反映了不同基因对形成个体表型模式的贡献,它是从 G 空间到 X 空间的线性映射,这是一种简化的映射方法.

2.2　自然选择

自然选择是在个体表型空间 X 与环境之间的相互作用中进行的,即自然选择是作用于基因表达结果之上的. 它用定义在 X 空间上的适应度函数 $f(x)$ 来衡量某种个体表型的适应能力. 个体表型模式在从亲代遗传到子代的过程中会部分的保留亲代的特征,同时发生变化,这种变化在本质上是无方向的. 它使得个体数量增多,而自然选择机制从中选出适应能力较强的 N 个个体,组成新一代群体. 这个过程是在 X 空间内的映射 f_2,即

$$f_2:\quad X \longrightarrow X \qquad (2)$$

2.3　表型特征在基因上的固化

当生物种群经过许多代的进化之后,由于自然选择的机制使得该种群在表型模式上逐渐发生变化. 这种变化的形成必定是在一定程度上改变了基因层或者是基因表达的结果. 也就是说,表型层的变化是有向基因层次上固化的过程. 这个过程是从表型空间 X 向基因空间 G 和基因表达矩阵空间 S 的一种映射. 它不能简单的用某种函数关系表示,它在实质上是一种优势的统计. 用 f_3 表示这种映射关系,即

$$f_3:\quad X \longrightarrow (S, G) \qquad (3)$$

2.4　基因的自参考突变

基因及其表达过程是一个复杂系统. 基因层次上任何一点的突变,在局部层次上(即单个基因)只是一个扰动,但反映到复杂系统上可能产生突现性效果. 自组织进化观认为,基因的突变是不完全随机的,即表型模式的积累优势在基因上的固化使得各个基因发生突变的概率不再是相等的. 对表型优势贡献小的基因,其突变概率显著性上升. 而且随着进化的进行,基因层上的突变由杂乱无章的随机突变逐渐进化成有序的定向突变. 这种变化趋势在本质上就是一个自参考的进化过程. 这个过程是在 G 空间中进行的,记为映射 f_4

$$f_4:\quad G \longrightarrow G \qquad (4)$$

上述在 G, X 空间中发生的(1)—(4)四种映射,完成了生物个体从表型层到基因层的共同进化.

3　算法的实现

本文算法实现了生物进化在两个空间上的四种映射关系. 法流程如下所述:

(1) 初始化程序参数:设定基因数目 m,个体数目 N,每代种群的亲本群数目 $n(n < N)$ 和交配亲本数 $p(p > 1)$. 选择高斯白噪声的标准差为 σ_0,设置种群进化代数 $k = 0$.

(2) 产生初始基因组与个体种群:在问题的可行解中等间距地产生 m 条基因,均匀的随机产生初始个体种群 $x^{(0)}$. 求出初始种群中各个体的适应函数值 $f(x_i^{(0)})$, $(i = 1, 2, \cdots, N)$.

(3) 判断是否停止进化.

(4) 进行个体表型的交叉和变异:选择当前种群中适应函数值最大的前 n 个个体作为候选亲本群. 第 i 个亲本被选中进行交叉的概率与式(5)成正比

$$P_{ci} \propto (f(x_i^{(k)}) - f_{\min}^{(k)}) \qquad (5)$$

式中 $f(x_i^{(k)})$ 是第 i 个个体的适应值, $f_{min}^{(k)}$ 是 N 个个体适应值的最小值. 设任意一次被选中的亲本为 $x_{i1}^{(k)}, x_{i2}^{(k)}, \cdots, x_{ip}^{(k)}$. 则按(6)产生子代.

$$x_i^{k+1} = \frac{1}{p} \sum_{l=1}^{p} x_{il}^{(k)} + \delta_x \qquad (6)$$

式中 $\delta_x \sim N(0, \sigma^{(k)})$, $\sigma^{(k)}$ 满足下式

$$\sigma^{(k)} = \min(\sigma_0 \frac{10}{10 + k}, \sigma_{\min}) \qquad (7)$$

其中 σ_0 是初始化的高斯白噪声标准差, k 是当前进化代数. 交叉完成后共产生 $2n$ 个子代.

(5) 替换:保留亲本中部分精英,其余个体由子代替换组成新一代种群 $x^{(k+1)}$.

(6) 判断是否出现种群共性优势:求出本代 N 个个体的适应值的平均 $f_M^{(k)}$,如果

$$(f_{\max}^{(k)} - f_M^{(k)}) / f_M^{(k)} \leq 0.05 \qquad (8)$$

则转(7),否则转(3).

(7) 表型优势向基因层固化:对 N 个个体中的任意一个 x_i $(i = 1, 2, \cdots, N)$,计算

$$a_{ij} = \frac{1}{|x_j - g_j| + \varepsilon} \qquad (j = 1, 2, \cdots m; \varepsilon > 0) \qquad (9)$$

ε 是一个很小的正数. 再由式(10)

$$\beta_i \sum_{j=1}^{m} a_{ij} g_j = x_i \qquad (i = 1,2,\cdots,N) \tag{10}$$

解出 β_i,从而得出新的基因表达矩阵

$$S_{ij} = \beta_i a_{ij} \qquad (i = 1,2,\cdots,N;\ j = 1,2,\cdots,m) \tag{11}$$

(8) 统计各基因对表型优势的贡献:对基因表达矩阵 S 按行进行归一化处理

$$S_{ij}^{*} = \frac{S_{ij}}{\sum_{k=1}^{m} S_{ik}} \qquad (i = 1,2,\cdots,N,\ J = 1,2,\cdots,m) \tag{12}$$

使得基因表达矩阵每行元素之和为 1. 再按式 (13) 进行纵向加权统计

$$w_j = \sum_{i=1}^{N} S_{ij}^{*} \qquad (i = 1,2,\cdots,m) \tag{13}$$

得到权向量 $w = [w_1, w_2, \cdots, w_m]^T$,它们分别对应基因 $g = [g_1, g_2, \cdots, g_m]^T$ 对表型优势的贡献.

(9) 基因层的自参考突变:令 w_r 是权向量 w 的最大分量,则它对应的 g_r 是优势基因. 若 $r > [m/2]$ ($[\ *\ ]$ 表示对 * 取整运算),则 $g_1, g_2, \cdots, g_{r-1}$ 按式 (14) 发生突变

$$g_l' = g_l + s|\ \delta_l| \tag{14}$$

式中 s 是一个离散随机变量,其分布律为 $P(s=1) = 0.618$, $P(s=-1) = 0.382$. s 实际上是决定基因突变的方向,在此情况下,基因增大的可能性比较大,即种群的基因以较大的概率向优势基因靠近. $\delta_l \sim N(0, \sigma_l^{(k)})$, $\sigma_l^{(k)}$ 的选择同 g_l 与 g_r 的距离有关. 若 $r < [m/2]$,则 $g_{r+1}, g_{r+2}, \cdots, g_m$ 按式 (14) 发生突变,式中 s 的分布律为 $P(s=-1) = 0.618$, $P(s=1) = 0.382$,此时,基因减小的可能性比较大.

(10) 求出基因突变后的个体:由 x = Sg,求出基因突变后的 N 个个体,与当前已有的 N 个个体进行选择,得到适应值最大的前 N 个个体,组成新的种群,转 (3).

在算法第 (4) 步中,高斯白噪声标准差 $\sigma^{(k)}$ 采用了十分简单的变化策略. 由式 (7) 可知,随着进化的不断进行,$\sigma^{(k)}$ 的趋势是逐渐减小,使得算法搜索的随机性逐渐减小.

在算法第 (6) 步中,利用式 (8) 判断种群是否出现共性表型优势,若出现,则进行表型优势向基因的固化过程.

在算法第 (7) 步中,式 (9) 利用各基因 g_j ($j = 1,2,\cdots,m$) 与表型个体 x_i 之间的距离衡量它们对表达此个体表型模式的相对贡献. 若 g_j 距 x_i 越近,则认为其贡献越大. 式中分母加上一个正小量 ε,是为了防止零溢出. 再根据式 (10), (11) 求出相应的基因表达矩阵. 这是由 X 空间向 (S, G) 空间的映射过程. 算法第 (8) 步首先将基因表达矩阵按行进行归一化处理,然后按列进行加权统计. 权向量 w 表示出了各基因对种群表型优势的不同贡献,然后在算法第 (9) 步中进行基因层的突变. 突变以优势基因作为参考,其余基因按概率向优势基因突变,体现了基因层的自参考进化过程. 最后利用基因表达矩阵将基因的突变映射到表型层,在表型层上继续进行自然选择.

本算法很容易推广到多维函数的寻优. 例如 D 维函数,对其每一维建立从基因层到表型层的映射. 在基因空间中,共有 D 种基因向量. 而任一个体共有 D 个表型特征,每个表型特征都是对应的 m 条基因表达的结果.

4 实验结果与讨论

为了比较算法的性能,用此算法与各种优化算法,如进化策略算法 (ES),遗传算法 (GA) 进行函数优化,考察算法的参数鲁棒性、全局搜索性与后期收敛性.

4.1 GEEA 与 GA 的比较

为了说明 GEEA 的性能,我们将它与遗传算法 (GA) 做比较. 这些遗传算法分别记为 FP, SA, PRAM1 和 PRAM2. 其中 FP 是一种由文献[9]提供的遗传算法. SA 是由文献[10]提供的自适应调整突变概率的遗传算法. PRAM1 和 PRAM2 是由文献[6]提供的两种自适应调整突变概率和交叉概率的遗传算法. 以下常用的目标函数用以测试各种算法的性能,它们是 De Jong's function1[11], Rastrigin's function[12] 和 Ackley's Path function[13]. 其中 De Jong's function1 是一个单峰函数,其余两个均为多峰函数. 以上函数的全局最小值点都取在自变量各分量全为 0 的时候.

De Jong's function 1:

$$F(x) = \sum_{i=1}^{n} x_i^2 \qquad x_i \in [-5.12, 5.12]$$

Rastrigin's function:

$$F(x) = nA + \sum_{i=1}^{n} (x_i^2 - A\cos(2\pi x_i)) \qquad x_i \in [-5.12, 5.12],\ A = 10$$

Ackley's Path function:

$$F(x) = 20 + e - 20\exp\left(-0.2\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}\right) - \exp\left(\frac{\sum_{i=1}^{n}\cos(2\pi x_i)}{n}\right)$$

$$x_i \in [-30, 30]$$

程序参数设定:基因数目 $m = 40$,种群数目 $N = 40$,亲本群数目 $n = 30$,交配亲本数 $p = 4$, $\sigma_0 = 0.01 \times$各维区间长度,算法采用保留 5 %的精英的策略,重复执行程序 20 次. 各函数所取的维数分别为 5, 4 和 4. 本文算法的最优解精度与文献[6]是等价的. 此部分主要比较算法找到最优解所用的目标函数计算次数和算法的收敛成功率.

Ackley's Path 函数的实验结果如图 2 所示,其余两个函数的优化结果与之类似. 从图 2 中可以看出, GEEA 所用的目标函数计算次数较之各种 GA 是最少的,而收敛成功率是最高的,说明 GEEA 的性能是优于 GA 的.

4.2 讨论

在这一部分,我们将讨论引入基因层的作用. 我们将 GEEA 与 ES 进行比较. ES 的实现即是本文算法中去掉 6 - 10

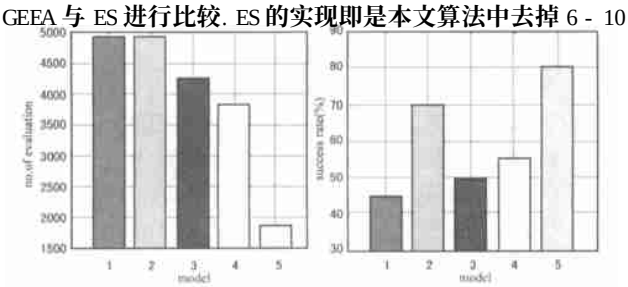

图 2 Ackley's Path function 的收敛速度与成功率

在图 2 中,model 1: FP; model 2: SA; model 3: PRAM1; model 4: PRAM2; model 5: GEEA

步,即不引入基因层的自参考突变作用. 两种算法的唯一差别就体现在基因层的引入上. 而比较的结果清楚的显示出,正是基因层的引入使算法的性能大大提高.

例

$$y(x)=\begin{cases}(5-|x|)|\cos(10\pi x)| & x\in[-5,5]\\ 0 & x\in[-5,5]\end{cases}$$

此函数共有 100 个极大值点,全局最优解为 5,可以很好的衡量算法的各项性能. 在实验中,取基因数目 $m=20$,个体数目 $N=20$,亲代群数目 $n=16$,交配亲本数 $p=2$. 设定最大进化代数为 500,函数满意解为 4.99999,高斯白噪声的初始标准差 $\sigma_0=0.3$. 每种算法重复执行 20 次实验,得到三种算法达到指定函数值时所用的平均目标函数计算次数如表 1 所示.

表 1　算法达指定目标函数值的平均目标函数计算次数比较

$y(x)$	4.99	4.995	4.999	4.9995	4.9999	4.99995	4.99997	4.99998	4.99999
ES	244	309	543	636	1118	1907	2235	2592	3132
GEEA	208	233	324	382	722	1106	1317	1662	2227

①后期收敛性　从表中数据不难看出,GEEA 与 ES 算法相比,突出优势在进化的后期. 在目标值达到 4.99 之前,两种算法所用的目标函数计算次数相差不大;而随着目标值的进一步增大,GEEA 的计算次数明显少于 ES 算法,这正是引入基因层的自参考突变的结果.

②参数鲁棒性　为了研究算法对参数 σ_0 的鲁棒性,我们令 σ_0 的取值从 0.1 均匀变化到 0.5,记录以上两种算法目标函数值达到 4.99999 时的平均目标函数计算次数如表 2 所示. 从表中数据可以明显的看到,GEEA 所用的目标函数计算次数小于 ES;当参数 σ_0 大范围变化时,GEEA 均能很好的收敛;而 ES 算法的收敛范围明显较小.

表 2　两种算法平均目标函数计算次数比较

σ_0	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
ES	1932	2908	2995	3962	3132	2971	3805	3873	4656
GEEA	1630	1886	2278	2288	2227	2443	2656	2042	2144

从此部分的比较我们可以看出,算法引入基因层后有两个明显的优点. 其一是提高了算法的收敛速度,其二是扩大了算法的有效收敛范围. 特别是明显的减小了寻找到最优解时所用的目标函数计算次数,这对于某些非常复杂的目标函数的优化问题,具有更明显的意义.

5　结论与展望

本文提出并实现了一种新的进化算法 GEEA. 它采用的进化模型既反映了自然选择过程,又反映了生物进化过程中的随机性,同时利用了复杂系统的自组织趋势,加快了种群的形成过程. 它强调了生物基因与表型,个体与种群之间的相互关系,将全新的生物进化的内在机理进一步体现在进化算法之上. 利用新的进化模型,GEEA 的性能得到了一定程度的提高. 本文所讨论的 GEEA 算法只是基于新的进化理论的基本算法,它还有很大的改进空间. 下一步研究的重点是研究进化模型中的四种映射实现方式的优化,同时考虑算法的几个可调参数的自适应调整进行,如 σ_0,亲本精英保留率等,并且进一步将算法广泛应用于各个领域的优化问题,以验证算法的普适性.

参考文献:

[1] X Yao, Y Liu. A new evolutionary system for evolving artificial neural networks [J]. IEEE Transactions on Neural Networks, 1997, 8 (3): 694 - 713.

[2] A V Sebald, J Schlenzig. Minimax design of neural net controllers for highly uncertain plants [J]. IEEE Transactions on Neural Networks, 1994, 5 (1): 73 - 82.

[3] K S Tang, K F Man, S Kwong, et al. Design and optimization of IIR filter structure using hierarchical genetic algorithms [J]. IEEE Transactions on Industrial Electronics, 1998, 45 (3): 481 - 487.

[4] S C NG, S H Leung, C Y Chung, et al. The genetic search approach [J]. IEEE Signal Processing Magazine, 1996, November: 38 - 46.

[5] K S Tang, K F Man, S Kwong. et al. Genetic algorithms and their applications [J]. IEEE Signal Processing Magazine, 1996, November: 22 - 37.

[6] C W Ho, K H Lee, K S Leung. A genetic algorithm based on mutation and crossover with adaptive probabilities [A]. Proceedings of the 1999 IEEE International Conference on Evolutionary Computation [C], IEEE Press, 1999: 768 - 774.

[7] M Srinivas., L M Patnaik. Adaptive probabilities of crossover and mutation in genetic algorithms [J]. IEEE Trans. on System, Man and Cybernetics, 1994, 24(4): 656 - 666.

[8] S Chen, R H Istepanian, B L Luk. Signal processing applications using adaptive simulated annealing [A]. Proceedings of the 1999 IEEE International Conference on Evolutionary Computation [C], IEEE Press, 1999: 842 - 849.

[9] J D Schaffer, R Caruana, L J Eshelman. et al. A study of control parameters affecting online performance of genetic algorithms for function optimization [A]. Proceedings of the 3rd International Conference on Genetic Algorithms [C], J D Schaffer, editor 1991: 51 - 60.

[10] J Smith, T Fogarty. Self-adaption of mutation rates in a steady-state genetic algorithm [A]. Proceedings of the 3rd International Conference on Evolutionary Computation [C], IEEE Press, 1996: 318 - 323.

[11] K De Jong. An analysis of the behavior of a class of genetic adaptive systems [D]. PhD. Thesis, University of Michigan, 1975.

[12] V S Gordon, D Whitley. Serial and parallel genetic algorithms as function optimizers [A]. Proceedings of the 5rd International Conference on Genetic Algorithms [C], 1993: 177 - 183, Morgan Kaufmann.

[13] H M Voigt. Soft genetic operators in evolutionary algorithms [A]. In W. Banzhaf, F H Eeckman. Eds., Evolution and Biocomputation [M]. Berlin; New York: Springer, 1995.

[14] J J Shynk. Adaptive IIR filter [J]. IEEE ASSP Magazine, 1989, April: 4 - 21.

作者简介:

周　晴　男. 1977 年 4 月出生于四川省成都市. 1995 年进入清华大学自动化系学习,1999 年获工学学士学位,2001 年获工学硕士学位. 主要研究兴趣为进化计算,生物进化的计算模型,生物信息学等.