第13卷第3期
1996年6月

控制理论与应用
CONTROL THEORY AND APPLICATIONS

Vol. 13, No. 3
Junl., 1996

遗传算法的运行机理分析*

恽为民　席裕庚

(上海交通大学自动化系·上海,200030)

摘要：遗传算法是一种自适应启发式群体型迭代式全局搜索算法，正受到许多学科的重视. 本文首先以函数优化为例分析了遗传算法的运行过程，然后着重探讨了遗传算法的全局收敛性和效率问题，提出了有效基因的新概念及有效基因突变操作，推导出每次遗传搜索产生$O(2^{l-1})$数量级的新模式，最后给出了结论.

关键词：遗传算法；全局收敛性；搜索效率

1　引　　言

遗传算法(Genetic algorithm)是由密切根大学Holland等创立的[1]，来源于进化论和群体遗传学，原先用于模拟自然系统的自适应现象，后来被引向了广泛的工程问题. 遗传算法发展成一种自适应启发式概率性迭代式全局搜索算法，以其解决不同非线性问题的鲁棒性、全局最优性、不依赖于问题模型的特性、可并行性及高效率，具有独特的吸引力，正引起越来越大的研究及应用热潮. 八十年代以来，遗传算法在许多方面得到了应用，如自动控制[4]、计算机科学[6]、机器人学[10,11]、模式识别[8]、工程设计[2]和神经网络[5]等领域.

遗传算法的运行过程较为简单，但它的整体行为是复杂的，其关键问题是：为什么遗传算法在解决非线性问题时具有全局收敛性?为什么全局收敛性又存在一定概率的不稳定?为什么遗传算法的搜索效率很高?Holland提出的模式定理[1]部分说明了第三个问题，对前两个问题的解释是牵强的[2]. 本文将首先以函数优化问题为例分析遗传算法的运行过程和作为遗传算法理论基础的模式定理，然后以有效基因的概念来说明前两个问题，并提出一种新的对遗传算法效率的估计方法来进一步说明第三个问题.

2　遗传算法的运行过程

遗传算法的结构与基因操作方式多种多样，难以从形式上明确定义，它的标志在于其内在特征，下面以函数优化为例来阐述遗传算法的主要特征和运行过程.

函数优化问题表述如下：有一n维未知函数$f(x)$:$R^n \to R$，但输入一自变量，能知道相应的函数值，求$\max f(x)$. 这是一个黑箱问题，没有函数在连续性等方面的任何信息，好的搜索算法必须有较好的能适用于不同函数的鲁棒性以及较高的效率.

遗传算法把该问题中的自变量当作生物体，将其转化为由基因构成的染色体，相应的函数值定义为适合度，未知函数为环境，生物体的目标是进化成具有最佳适合度的基因型. 本文以后称染色体为串，用遗传算法求解上述函数优化问题的步骤如下：

Step 1　选择编码策略，把参数转换成串；

* 国家自然科学基金与上海市自然科学基金资助项目.

本文于1993年12月15日收到. 1995年3月30日收到修改稿.

编码策略有二进制编码和实数编码等，若采用二进制码表达实数，每个二进制位即为一基因，若一维参数 $x \in [a,b]$，则

$$x = a + \frac{\sum_{i=1}^{l} g_i 2^i}{2^{i+1} - 1}(b - a). \tag{2.1}$$

其中，l 是串的长度，g_i 为第 i 个基因. 本例中，若 $a = 0, b = 1$，串长 $l = 5$，串 10101 表示实数 0.6674；

Step 2 定义串的适合度函数为 $F = f(x)$；

适合度函数是目标函数或耗散函数的映射，它包含了对黑箱所需的所有信息；

Step 3 选择群体大小 n_p 等遗传参数，随机产生 n_p 个串构成群体；

Step 4 计算群体中串的适合度. 由串解码所得的解越好，则适合度值越高；

Step 5 根据串的复制概率 $p_r(F)$ 选择2个串，各复制一份，适合度越高，则复制概率越大；

Step 6 在串上随机选择一个位置，在复制的 2 个串上标记为交位 $cs1$；

Step 7 以交换概率 p_c 交换 $cs1$ 后的基因段；

Step 8 对两个串中的基因按突变概率 p_m 进行翻转；

Step 9 跳至 Step 5，直至已复制 n_p 个串；

Step 10 从 Step 4 始重复进行，直到满足某一性能指标或规定的遗传代数.

以上遗传过程描述了最简单的进化模型. Step 1 和 Step 2 是实际应用中的关键. Step 5 至 Step 8 进行三种基本基因操作，复制实施了适者生存的原则；交换的作用是组合父代中有价值的信息，产生新的后代，以实现高效搜索；突变的作用是保持群体中基因的多样性.

下面以最简单的一维线性函数 $f(x) = x$ 为例，$x = 0,1\cdots,31$，求 $\max f(x)$. 串长为 $l = 5$，群体大小 $n_p = 4$，遗传过程见表 1. 其中，第(2)栏是随机产生的群体，第(3)栏是串对应的参数值，第(4)栏是串对应的函数值和适合度，适合度函数定义为 $F = f(x)$，第(5)栏是各种串的复制概率，由 $F_i/\sum F_i$ 求得，第(6)栏是复制数，其中第 3 个串消失，复制概率高的第(4)个串则复制 2 个，复制后的群体见第(7)栏，第(8)栏随机产生配对序号，这里 1、4，2、3 配对，第(9)栏随机产生 2 个交换位 cs1，cs2，然后交换 cs1 和 cs2 间的基因，产生的新群体见第(10)栏，第(12)栏是新群体的适合度，总和从 62 上升至 83，最大值从 27 上升至 31. 交换操作的概率为 1.0，突变操作的概率为 0.003，第一代遗传共处理 20 个基因，没有突变现象.

下面具体考察一下遗传算法的运行性能. 对如下函数进行优化，$f(x) = |\sin 30x|(1 - x/2), x \in [0,1]$，求 $\max f(x)$. 用穷举法求得 $x = 0.0517900$ 时，$\max f(x) = 0.9739626$，函数图形见图 1. 取串长 $l = 32$，群体大小 $n_p = 40$，第一代随机产生的群体见图 2，分布是均匀的，这时 $\max f(x)$ 在第 2 个峰，一般寻优方式可能陷入局部最优解，图 3 显示的第 30 代群体分布却发现，这时的 $\max f(x)$ 处在了最高峰，附近聚集了 12 个个体，显示了遗传算法是全局寻优的，且群体的整体得到优化. 遗传算法的计算精度也令人满意，经 50 次迭代 90% 可达 0.9739，误差为 6.5×10^{-5}，计算速度为 250 代 / 秒(486/66MHz).

表 1　遗传过程

序　号	初始群体	X	适合度 $F=f(x)$	复制概率	复制数
(1)	(2)	(3)	(4)	(5)	(6)
1	10100	20	0.32	0.32	1
2	01010	10	0.16	0.16	1
3	00101	5	0.08	0.08	0
4	11011	27	0.43	0.43	2
总　数		62	1.00	1.00	4
平均值		15.5	0.25	0.25	1
最大值		27	0.43	0.43	2

复制后的群体	配对	交换位置		新群体	x	适合度
(7)	(8)	$cs1$ (9)	$cs2$	(10)	(11)	(12)
10100	4	2	3	10000	16	16
01010	3	1	4	01010	10	10
11011	2	1	4	11011	27	27
11011	1	2	3	11111	31	31
总　数						84
平均值						21
最大值						31

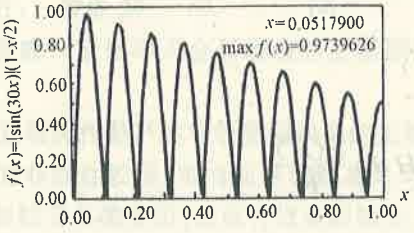

图 1　函数图形

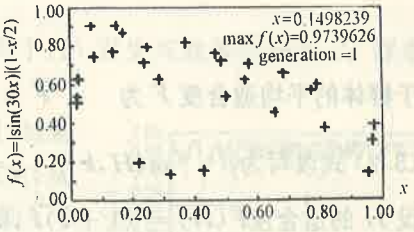

图 2　第一代群体分布

3　模式定理

遗传算法使我们直觉感到它模拟了生物的进化机理，简单而有效，但其中的许多随机性操作使我们疑虑它的性能，模式定理[1]是遗传算法中的基本理论，提供了一种解释遗传算法机理的数学工具，还蕴含着发展编码策略和基因操作策略的一些准则.

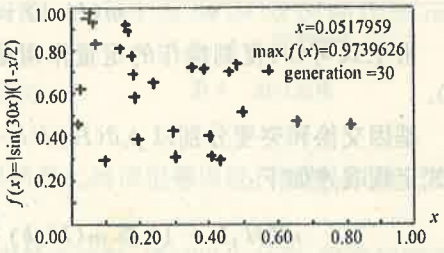

图 3　第三十代群体分布

遗传算法涉及四个空间：问题参数空间、串空间、模式空间和适合度空间，其功能的循环关系见图 4. 遗传操作是在串空间进行的，这是 GA 的显著特点之一. 在一定程度上，我们不再把串仅仅视作数的另一形式，因为高适合度串中的相似性有利于引导搜索，这意味着串空间的操作蕴含模式空间的隐形处理.

首先，把串中的相似样板称为模式(Schema)，模式表示基因串中某些特征位相同的结构，二进制串中的模式是如下的形式：

$$(a_1,a_2,\cdots,a_i,\cdots,a_n),\ a_i\in\{0,1,*\}.$$

其中“*”是任意符，既可以是“1”，也可以是“0”. 模式是串的集合，这些串与模式在所有基因位上与不是“*”的基因匹配. 因此，模式中“*”越多，则描述的串越多. 一个长度是5的模式 $*111*$，描述如下串的集合：

$$*111* = \{(11110),(01110),(01111),(11111)\}.$$

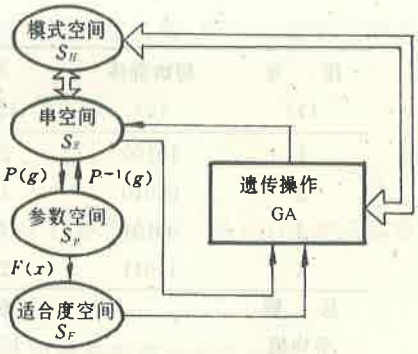

图4 遗传过程与四个空间

一个模式 H 包含四个参数：原始长度 l，度 $O(H)$，定义长度 $\delta(H)$ 及模式的维数 $D(H)$，l 即是串的长度，$O(H)$ 是模式中固定位的个数，$\delta(H)$ 是最前和最后固定位间的距离，例如，$O(*111*)=3$，$\delta(*111*)=4-2=2$，$D(H)$ 表示模式中包含串的个数，为

$$D(H)=2^{l-O(H)}. \tag{3.1}$$

假设第 k 代遗传时，在群体 $p(k)$ 中有 m 个某特定模式 H，写成 $m(H,k)$，复制过程中，含 H 的串按 $p_r=F(H)/\sum F$ 的概率复制，其中 $F(H)$ 是含 H 的串的适合度，复制数为 $nF(H)/\sum F_j$；令 $\overline{F(H)}$ 为所有含 H 的串的平均适合度，$m(H,k+1)$ 为：

$$m(H,k+1)=m(H,k)\frac{n\overline{F(H)}}{\sum_{j=1}^{n}F_j}. \tag{3.2}$$

由于群体的平均适合度 $\overline{F}$ 为

$$\overline{F}=\frac{\sum_{j=1}^{n}F_j}{n}. \tag{3.3}$$

故(3.2)式改写为

$$m(H,k+1)=m(H,k)\frac{\overline{F(H)}}{\overline{F}}. \tag{3.4}$$

假设 H 的适合度 $\overline{F(H)}=(1+c)\overline{F}$，则

$$m(H,k+1)=(1+c)m(H,k). \tag{3.5}$$

若从 $k=0$ 始，c 恒定，则可得

$$m(H,k)=m(H,0)(1+c)^k. \tag{3.6}$$

由上式可知：复制操作的定量作用是使超出(或低于)$\overline{F}$ 的模式按指数形式增加(或减少).

基因交换和突变分别以 $p_c\delta(H)/(l-1)$ 和 $O(H)p_m$ 的概率破坏模式，综合二者影响，模式定理表述如下：

$$m(H,k+1)\geqslant m(H,k)\frac{\overline{F(H)}}{\overline{F}}\left(1-p_c\frac{\delta(H)}{l-1}-O(H)p_m\right). \tag{3.7}$$

其中 l 是串长，p_c，p_m 分别是交换操作和突变操作的概率.

模式定理说明：定义长度 $\sigma(H)$ 短、度 $O(H)$ 低且适合度高于平均适合度的模式的数量在遗传过程中将以指数形式增加. 这个结果是指导遗传算法设计的重要原则. 选择的编码策略须使 $\delta(H)$ 短 $O(H)$ 低的模式对应于所求的解，由于编码是遗传过程中的基石，不合适的

编码会极大影响遗传算法的性能.

4　遗传算法的性能分析

4.1　全局收敛性

传统的基于模式定理的对全局收敛性的定性分析认为遗传算法是全局收敛的.最近基于马尔科夫链的定量的数学证明认为简单遗传算法不是全局收敛的,而带最优保存的遗传算法是全局收敛的[13],但后者的结论建立在搜索时间趋向无穷且把突变操作作为主要搜索工具的基础上.实质上的遗传算法的主要搜索工具是交换操作,下面以文献[13]中的初步结论为前提,借助于有效基因的概念更实际地说明遗传算法的全局收敛性.

定义 4.1　设全局最优串为 s^*,则 s^* 中每一基因位上的基因称之为有效基因.

定义 4.2　群体中某一基因位上,若所有串都没有该基因位的有效基因,则称为群体有效基因缺失.

命题 4.1[13]　对于两个互补的特征串,交换操作能够进行遍历搜索.

群体中若不存在有效基因缺失,则群体可抽象成两个互补的特征串,则命题 4.1 可表述为:有效基因不缺失的群体,交换操作能够进行遍历搜索,借助于文献[13]中的结论,在带有最优串保存时,有效基因不缺失的群体的遗传过程具有全局收敛性.

随机产生的初始化群体的有效基因缺失概率是:

$$p_{ag} = 1/2^n. \tag{4.1}$$

其中 n 是群体大小,若 $n = 30$,$p_{ag} = 9.3 \times 10^{-10}$,有效基因缺失的概率是很小的,这说明了实际应用中遗传算法具有较好的全局收敛性,式(4.1)也说明,群体数越大,全局收敛性越好,一般取 20 ～ 60.

造成一定概率不全局收敛的主要原因有两个:1) 发生有效基因缺失,2) 搜索时间不够.

复制操作是产生有效基因缺失的主要原因,如果问题的非线性较强,则当前最优解易是局部最优,从而复制较多此类基因,造成有效基因缺失,对与图 1 同类非线性较强的函数 $f(x) = |\sin 100x|(1 - x/2), x \in [0,1]$ 进行优化,取同样的遗传参数,遗传算法的运行过程见图 5,其运行过程不稳定,陷入局部最优的概率增大,原因是造成了基因缺失.发生有效基因缺失时,只有突变操作才能消除,因此,得如下有效基因突变的策略:

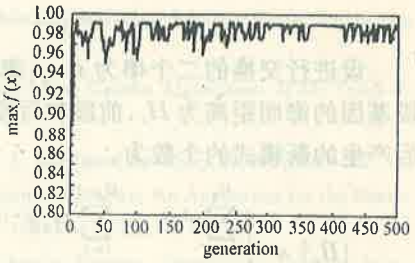

图 5　运行过程

当发现有效基因缺失时,对适合度低于平均值的串上的该位基因进行翻转,以保证有效基因的下限值.

第二节曾指出突变的主要作用是保持群体基因的多样性,通过以上分析,突变的明确作用是避免有效基因缺失,也即保证遗传算法的全局收敛性.但常规的位突变是随机进行的,不能有效地避免有效基因缺失.

遗传算法实现全局收敛的时间复杂度是一个更为复杂艰巨的问题,也是应用中更为关心的问题,目前尚无明确的定量结论.下小节基于模式理论进一步分析遗传算法的搜索效

率.

4.2 基于模式的效率分析

因为模式由固定基因1,0及任意符 * 组成,所以串长为 l 的模式空间 Π 的维数为 $|\Pi| = 3^l$,从模式的角度看,遗传算法的运行过程是在模式空间中搜索一个最优模式的过程.搜索时间取决于 $|\Pi|$ 的大小和每次遗传的搜索效率.$|\Pi|$ 的大小取决于具体问题,因此每次遗传的搜索效率是关键的,取决于以下两点:

1) 每次遗传产生的新模式.

2) 每次遗传有效保留的模式.

这两点又是矛盾的,如果每次产生很多新模式,但高适合度的模式得不到有效保留,或者每次保留很多模式,而产生较少新模式,则易陷于局部解,搜索效率均不高,遗传算法需在二方面进行较好的协调,尤其需要根据遗传过程进行动态的协调才能有高的效率.

由于串之间有重复的模式,因此,一个群体中包含的模式数量为:

$$2^l \leqslant |\Pi_b| \leqslant n2^l. \tag{4.2}$$

从模式的角度看,群体中包含的信息量十分巨大,而不仅仅是 n 个串.文献[1,2]得出了每次遗传有效保留的模式为 $O(n^3)$ 的估计,这就是所谓的"隐形平行性",文献[12]推广了这方面的结果,这些模式是 $O(H)$ 和 $\delta(H)$ 均小的模式,基因操作中不易被破坏,如果这些模式具有高的适合度,则根据模式定理,这些有效模式数量在复制操作的作用下以指数形式增加,从而加快这些模式附近空间的搜索.

遗传过程中,主要由交换和突变操作产生新的模式.

设群体中的任意串 s_i 按突变 p_m 突变后的串为 s'_i,s_i 和 s'_i 之间的海明距离为 $H(s_i,s'_i) = lp_m$,s'_i 具有 s_i 中没有的新模式为:

$$|\Pi_m| = \sum_{i=1}^{H(s_i,s'_i)} 2^{l-i}. \tag{4.3}$$

设进行交换的二个串为 s_i,s_j,海明距离为 $H(s_i,s_j)$,若采用两个交换点的操作,设交换段基因的海明距离为 H_c,前段和后段的海明距离分别为 H_b 和 H_a,经推导,两个串交换操作后产生的新模式的个数为:

$$|\Pi_c| = \begin{cases} \sum_{h=1}^{H_b} 2^{l-h} + \sum_{i=1}^{H_c-1} H_b 2^{l-H_b-i} + \sum_{j=1}^{H_a} 2^{l-H_b-j} + \sum_{k=1}^{H_c-1} H_a 2^{l-H_b-H_a-l}, & H_c \geqslant 1, \\ 0, & H_c = 0. \end{cases} \tag{4.4}$$

例如以下交换操作:

$$s_i = 1;1\ 1;1 \Rightarrow 1\ 0\ 0\ 1, \quad s_j = 0;0\ 0;1 \Rightarrow 0\ 1\ 1\ 0,$$

共产生18个新模式,$H(s_i,s_j) = 4$,$H_b = 1$,$H_c = 2$,$H_a = 1$,求得 $|\Pi_c| = 18$,计算相符.

在随机产生的群体中,基因1和0的产生概率均为0.5,则2个串的平均海明距离 $H(s_i,s_j) = 0.5l$,由(4.4)式,只要 $H_c \geqslant 1$,则有:

$$|\Pi_c| > 2^{l-1}. \tag{4.5}$$

若 $p_c = 1.0$,则群体中共有 $n/2$ 对串进行操作,有

$$2^{l-1} < n2^{l-2}. \tag{4.6}$$

令 $n = l$，并综合突变和交换的作用，产生新模式的下限为

$$|\Pi| > 2^{n-1}. \tag{4.7}$$

(4.7) 式说明每次遗传产生 $O(2^{n-1})$ 数量级的新模式. 每次遗传产生的新串的数目和串空间的比值为 $n/2^l$，每次遗传产生的新模式的数目和模式空间大小的比值为 $2^{n-1}/3^l$，若 $n = l$，则有

$$\frac{n}{2^l} < \frac{2^{n-1}}{3^l}, \quad l > 8. \tag{4.8}$$

式(4.8) 说明，随着 l 增加，遗传算法的相对搜索效率指数增加，这个结果进一步说明了遗传算法高效的搜索能力，使得遗传算法的隐形平行性机理更为完整.

5 总　　结

本文首先以函数优化为例分析了遗传算法的结构和运行过程，接着建立了遗传算法蕴含的四个空间的功能关系，阐述了遗传算法的数学基础之一 —— 模式定理，最后分析了遗传算法的全局收敛性和搜索效率. 本文提出了有效基因的概念，发展了有效基因突变作为保证全局收敛的基本策略，本文得到了每次突变操作和交换操作产生的新模式空间的维数公式，每次遗传至少产生 $O(2^{n-1})$ 数量级新模式的结论进一步阐明了遗传算法的搜索效率和隐形平行性机理.

参 考 文 献

[1] Holland. J. H.. Adaptation in Natural and Artificial System. Ann Arbor: The University of Michigan Press, 1975

[2] Goldberg, D. E.. Computer-Aided Gas Pipeline Operation Using Genetic Algorithms and Rule Learning. Engineering with Computers, 1985, 35—58

[3] Goldberg, E. E.. Genetic Algorithms in Search, Optimization, and Machine learning. Addison-Wesley Publishing Company, 1989

[4] Kristisson, K. and Dument, G. A.. System Identification and Control Using Genetic Algorithms. IEEE Trans on SMC. 1992, 22(5): 1033—1046

[5] Yao, X.. A Review of Evolutionary Artificial Neural Networks. Int. J. Intelligent Systems, 1993, 8: 539—567

[6] Raghavam, V. V. and Agarual, B.. Optimal Determination of Uesr-Oriented Clusters: An Application for the Reproductive Plan. Proc. of the Second Int. Conf. on Genetic Algorithms, 1987, 241—246

[7] De Jong, K. A.. An Analysis of the Behavior of a Class of Genetic Adaptive Systems. Dissertation Abstracts International, 36110J, 5140B, 1975

[8] Stanyk, I.. Schema Recombination in Pattern Recognition Problems. Proc. of the Second Int. Conf. on Genetic Algorithms, 1987, 27—35

[9] Versek, A.. Genetic Algorithms in Controller Design and Tuning. IEEE Trans. on SMC, 1993, 23(5): 1330—1339

[10] Parker, J. K. and Goldberg, D. E.. Inverse Kinematics of Redundant Robots Using Genetic Algorithms. IEEE Int. Conf. on Robotics and Automation, 1989, 271—275

[11] Pearce, M.. The Learning of Reactive Control Parameters Through Genetic Algorithms. IEEE/RSJ Int. Conf. on Intelligent Robots and Automation, 1992, 709—712

[12] Bertoni, A. and Dorigo, M.. Implicit Parallelism in Genetic Algorithms. Artificial Intelligence. 1993, 61: 307—314

[13] 恽为民. 基于遗传的机器人运动规划的研究. 上海交通大学博士论文, 1995

The Analysis on Running Mechanism of Genetic Algorithm

YUN Weimin and XI Yugeng

(Department of Automation, Shanghai Jiao Tong University · Shanghai, 200030, PRC)

Abstract: Genetic algorithm (GA) is a kind of adaptive, heuristic, probabilistic and iterated searching algorithm, which has attracted many subjects. The architecture and running process of genetic algorithm are analyzed first as the example of function optimization in the paper, then, the global optimum and seaching efficiency are studied. The new concept of effective gene and effective gene mutation are proposed. The new schema produced during one genetic operation are derived as $O(2^{n-1})$. Finally, the conclusion is given.

Key words: Genetic algorithm; global optimum; searching effciency

本文作者简介

恽为民 1968年生. 1992年获哈尔滨工业大学硕士学位, 1995年获上海交通大学博士学位. 研究方向为: 机器人学、遗传算法及人工生命等.

席裕庚 1946年生. 1968年毕业于哈尔滨军事工程学院, 1984年在慕尼黑工业大学获德国工学博士学位, 现为上海交通大学教授, 自动控制理论与应用博士生导师. 目前主要研究方向是复杂工业过程及智能机器人控制的理论和方法.