

# 十进制整数编码遗传算法的模式定理研究<sup>\*</sup>)

On Scheme Theorem of the Integer-Coded Genetic Algorithm

唐 飞 滕弘飞

(大连理工大学机械工程系 大连 116024)

(中科院现代制造 CAD/CAM 技术开放实验室 沈阳 110003)

**Abstract** This paper concentrates on the scheme theorem of integer-coded genetic algorithm. The basis theorem of binary-coded genetic algorithm is scheme theorem. However, there is not enough research on the scheme theorem of decimal coded genetic algorithm. We define the corresponding concepts and put forward the corresponding scheme theorem of the integer-coded genetic algorithm referring to the basic concepts of binary coded genetic algorithm.

**Keywords** Scheme theorem, Decimal coding, Integer coding, Genetic algorithm

## 1. 引言

本世纪 50 年代中期创立了仿生学,许多科学家从生物中寻求新的用于人造系统灵感,一些科学家分别独立地从生物进化的机理中发展出适合于现实世界复杂问题优化的模拟进化算法(Simulated Evolutionary Optimization),主要有 Holland, Bremermann 等创立的遗传算法,Rechenberg 和 Schwefel 等创立的进化策略以及 Fogel, Owens, Walsh 等创立的进化规划。遗传算法、进化策略及进化规划均来源于达尔文的进化论,但三者侧重的进化层次不同,其中遗传算法的研究最为深入、持久,应用面也最广<sup>[1]</sup>。遗传算法的编码策略包括至今仍在争论的两派,一派根据模式定理建议用尽量少的符号编码(如二进制编码);一派以数值优化计算的方便和精度为准采用一个基因一个参数的方法(如十进制编码),并把相应的基因操作改造成适合十进制数操作的方式<sup>[2,3]</sup>。对于离散的优化问题,采用二进制编码可以准确地描述问题的解群体,但是,对于连续优化问题,如函数优化和约束布局优化问题,效果不理想,这是由于:一方面,采用二进制编码,不可避免带来一定甚至是较大的精度损失;另一方面,解群体中个体的长度太大,使得遗传算法的效率变低,甚至在可

以容忍的时间内或进化代数内无法收敛。因此,在实践中采用十进制数编码来求解问题的情况越来越多。但由于十进制编码一直没有类似二进制编码遗传算法的模式定理,因此一直难于对其进化机制和性能进行精确分析。本文借鉴二进制编码遗传算法中的概念,对十进制编码的模式定理进行研究,并给出了以十进制整数编码遗传算法的基本定理,称之为十进制整数编码模式定理。

## 2. 二进制编码的模式定理

### 2.1 有关定义<sup>[2,4]</sup>

设二元基因表  $V = \{0, 1\}$ , 采用二元基因编码的染色体串可以用带下标的字母来形式地表示,其中下标代表其位置顺序,记为:

$$A = a_1 a_2 a_3 a_4 a_5$$

其中:  $a_i$  表示一个二元基因,可取值 1 或者 0。

**定义 1** 模式就是一个相同的构形。它描述的是一个染色体串的子集。这个集合中的染色体串之间在某些位上相同。考虑由三元基因表  $V^+ = \{*, 0, 1\}$  表示的模式,其中添加的 \* 代表不确定的基因,即在一特定位置上与 0 或 1 相匹配。

**定义 2** 模式的阶是指出现在模式中的确定位置的数目,记为  $\Delta(H)$ 。在二进制编码的染色体串中

<sup>\*</sup>) 本文得到国家自然科学基金资助(69573004)。唐飞 博士生,研究方向为智能 CAD,优化理论与航天器布局优化设计。滕弘飞 教授,博士生导师,研究方向为智能 CAD/CAM 及优化,计算机算法,航天器布局优化设计。

一个模式的阶就是所有 0 或 1 的个数。例如,模式  $0 * 10 *$  的阶  $\Delta=3$ ,而模式  $1 * * * *$  的阶为  $\Delta=1$ 。

**定义 3** 模式定义距是指模式中第一个确定位置与最后一个确定位置之间的距离,记为  $d(H)$ 。如模式  $0 * 10 *$  的定义距  $d=3$ ,而模式  $1 * * * *$  的定义距为 0。

**定义 4** 适应度值是指为群体中每个染色体串指定的一个数值,记为  $f$ ,它经常是问题本身所具有的。适应度值必须有能力计算搜索空间中每个染色体串的性能值。

控制遗传算法的主要参数有群体规模  $N$ 、算法执行的最大代数  $M$ 、复制概率  $p_r$ 、交叉概率  $p_c$  和变异概率  $p_m$  等参数。

### 2.2 二进制编码遗传算法的模式定理<sup>[2,4]</sup>

**模式定理** 具有短的定义距、低阶并且适应度值在群体平均适应度值以上的模式在遗传算法迭代过程中将按指数增长率被采样。也就是说,在使用遗传算法时,染色体群体中那些短的低阶模式是按照指数增加还是减少的数目进行采样,依赖于模式的平均适应度值。

## 3. 十进制整数编码的模式定理

### 3.1 基本概念定义

参考二进制编码遗传算法基本概念的定义,给出十进制整数编码遗传算法相应的基本概念的定义如下。

对于十进制整数编码的遗传算法,我们采用 10 元基因表  $V = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  对染色体串编码,每个染色体串同样可以用带下标的字母来表示,其中下标代表位置顺序。例如,一个 5 位染色体串  $A=14679$  可以表示为:

$$A = a_1 a_2 a_3 a_4 a_5$$

其中:  $a_i$  表示一个基因,且  $a_i \in V$ 。

**定义 5** 模式就是一个相同的构形。它描述的是一个染色体串的子集。这个集合中所有的染色体串之间在某些位上相同。考虑由 11 元基因表  $V^+ = \{ * \} \cup V = \{ *, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$  表示的模式,其中添加的  $*$  代表不确定的基因,即在一特定位置上与  $V$  中某一元素相匹配。例如考虑染色体串长为 5 的模式  $H = * * 6 7 *$ ,则前述的染色体串  $A = 14679$  是模式  $H$  的一个具体表示形式。

定义在染色体串长为  $l$  上的二进制染色体串的模式共有  $3^l$  个。一般地,对于基数为  $k$  的基因表,共有  $(K+1)^l$  个模式。因此定义在染色体串长为  $l$  的

十元基因表  $V$  上的模式共有  $(10+1)^l = 11^l$  个模式。在  $n$  个十进制整数编码的染色体串群体中至多有  $n \cdot 10^l$  个模式包含在其中。由此可以得出,采用十进制整数编码表示染色体串时,群体中的模式的数目仅与群体大小和染色体长度有关。

**定义 6** 模式的阶是指出现在模式中的确定位置的数目,记为  $\Delta(H)$ 。在十进制整数编码的染色体串中一个模式的阶就是所有  $V$  中元素在该模式中出现的个数。例如,模式  $14 * 7 *$  的阶  $\Delta=3$ ,而模式  $1 * * * *$  的阶  $\Delta=1$ 。

**定义 7** 模式定义距是指模式中第一个确定位置与最后一个确定位置之间的距离,记为  $d(H)$ 。如模式  $14 * 7 *$  的定义距  $d=3$ ,而模式  $1 * * * *$  的定义距为 0。

### 3.2 模式定理及推导

文献[2,4~7]中介绍并讨论了二进制编码遗传算法的模式定理,本文在此基础上推导出十进制整数编码遗传算法的模式定理。

控制遗传算法的主要参数有群体规模  $N$ 、算法执行的最大代数  $M$ 、复制概率  $p_r$ 、交叉概率  $p_c$  和变异概率  $p_m$  等参数。模式、模式阶和模式定义距对于严格讨论和区分染色体串的相似性是一个有力的工具,因此,要通过讨论复制、杂交和变异算子对模式的影响来得出该模式定理。

假定有一个由  $n$  个染色体串构成的群体  $P$ ,在给定的时间步  $t$ ,一个特定的模式  $H$  有  $m$  个代表染色体串包含在群体  $P(t)$  中,记为  $m = m(H, t)$ 。在复制阶段,每个染色体串根据它的适应度值进行复制,或者更确切地说,一个染色体串的复制概率为:

$$p_n = f_i / \sum_{j=1}^n f_j, i, j = 1, 2, \dots, n \quad (1)$$

当采用非重叠的  $n$  个染色体串的群体替代群体  $P(t)$  后,我们期望在时间步  $(t+1)$ ,模式  $H$  在群体  $A(t+1)$  中有  $m(H, t+1)$  个代表染色体串,这可以用下面的方程给出:

$$m(H, t+1) = m(H, t) \cdot n \cdot f(H) / \sum_{j=1}^n f_j \quad (2)$$

其中:  $f(H)$  是在时间步  $t$  表示模式  $H$  的染色体串的平均适应度值。由于整个群体的平均适应度值可记为

$$\bar{f} = \sum_{j=1}^n f_j / n$$

因此模式的复制生长方程可以表示为:

$$m(H, t+1) = m(H, t) \cdot f(H) / \bar{f} \quad (3)$$

令  $f(H) = (1+\lambda)\bar{f}$ , 为一常数,则模式的复制

生长方程变为:

$$m(H, t+1) = m(H, t) \cdot (1 + \lambda) \quad (4)$$

从  $t=0$  开始, 假设  $\lambda$  是一个固定值, 则有:

$$m(H, t+1) = m(H, t) \cdot (1 + \lambda)^t \quad (5)$$

这表明, 一个特定的模式按照其平均适应度值与群体平均适应度值之间的比率增长。平均适应度值以上(以下)的模式将会按照指数增长(衰减)的方式被复制。

由于复制过程不能检测搜索空间中新的区域, 因此需要采取杂交操作。杂交就是在两个染色体串之间进行信息交换。为叙述方便起见, 我们仅采用简单的一点杂交算子。

假定有一个染色体串长为 5 的特定的染色体串和包含在其中两个具有代表性的模式如下:

$$A = 1 \quad 4 \quad 6 \quad 7 \quad 9$$

$$H_1 = 1 \quad * \quad * \quad * \quad 9$$

$$H_2 = * \quad * \quad 6 \quad 7 \quad *$$

假定染色体串  $A$  被选择用来杂交, 杂交位置在 2 和 3 之间, 由一点杂交算子对模式  $H_1$  和  $H_2$  的作用效果如下:

$$A = 1 \quad 4 \quad \left| \begin{array}{ccc} 6 & 7 & 9 \end{array} \right.$$

$$H_1 = 1 \quad * \quad \left| \begin{array}{ccc} * & * & 9 \end{array} \right., \quad d(H_1) = 4$$

$$H_2 = * \quad * \quad \left| \begin{array}{ccc} 6 & 7 & * \end{array} \right., \quad d(H_2) = 1$$

可以看出, 除非染色体串  $A$  的交配染色体串在模式  $H_1$  的确定位置上与  $A$  相同, 否则模式  $H_1$  将被破坏, 而对于相同杂交位置的模式  $H_2$  将生存下来。即模式  $H_1$  比起模式  $H_2$  来更不易生存。这是由于模式  $H_1$  比模式  $H_2$  的定义距要长的缘故。

一般地, 对任意模式可计算出其杂交生存概率  $p_c$  的下界。考虑在简单一点杂交算子作用下, 对于长度为  $l$  的模式的生存概率为  $p_c = 1 - d(H)/(l-1)$ , 当杂交位置落在定义距长度之外时, 这个模式就可以生存。否则, 当杂交位置一旦落在定义距之内时, 则模式极易被破坏。若杂交本身也是按照随机选取方式进行, 即以概率  $p_c$  进行杂交, 则生存概率有下面的估计式:

$$p_c \geq 1 - p_c \cdot d(H)/(l-1) \quad (6)$$

现在考虑复制和杂交结合在一起时对模式的作用效果。这里假定复制和杂交是不相关的, 则有下面估计:

$$m(H, t+1) \geq m(H, t) \cdot [1 - p_c \cdot d(H)/(l-1)] \cdot f(H)/\bar{f} \quad (7)$$

由上式可以看出, 模式  $H$  增长或衰减依赖于一个乘积因子。在复制和杂交作用下, 这个因子依赖于

两个因素: 模式适应度值和模式的定义距。显然, 那些既在群体平均适应度值之上同时又具有短的定义距的模式将按照指数增长率被采样。

最后, 考虑变异算子的作用效果。变异算子以概率  $p_m$  随机地改变一个基因位上的值, 为了使得模式  $H$  能够生存下来, 所有特定位必须存活。因为单个基因存活的概率为  $(1 - p_m)$ , 并且由于每次变异都是统计独立的, 因此当模式  $H$  中  $\Delta(H)$  个确定位置都存活时, 这个模式才存活, 因而在变异算子的作用下, 存活概率可以近似地表示为  $(1 - p_m)^{\Delta(H)}$ 。对于很小的  $p_m$ , 模式的存活概率可以近似地等于  $(1 - \Delta(H) \cdot p_m)$ 。因此, 在复制、杂交和变异算子作用下, 一个特定模式  $H$  在下一代中期望出现的次数可以近似地表示为:

$$m(H, t+1) \geq m(H, t) \cdot [1 - p_c \cdot d(H)/(l-1) - \Delta(H)p_m] \cdot f(H)/\bar{f} \quad (8)$$

由上式可以看出, 增加变异基本上不改变先前的结论, 即变异前后相差很小。

### 3.3 十进制整数编码模式定理

综上所述, 十进制整数编码遗传算法的模式定理为: 采用十进制整数编码的遗传算法的群体中模式的数目仅与群体大小和染色体长度有关, 其中具有短的定义距、低阶并且适应度值在群体平均适应度值以上的模式在遗传算法迭代过程中将按指数增长率被采样。

**结束语** 十进制整数编码较二进制编码的遗传算法的染色体所能表示的模式数目大, 隐含并行性强。十进制编码的遗传算法在实践中得到广泛而成功地应用, 但是其模式定理的相关研究尚欠缺, 研究将二进制模式定理推广到十进制实数编码有助于对其进化机制和性能进行深入分析。

本文在二进制数编码遗传算法的模式定理基础上, 根据十进制整数编码遗传算法的遗传算子对其模式的影响, 推导出十进制整数编码遗传算法的模式定理。十进制整数编码的遗传算法的模式定理在结论上与采用二进制编码的遗传算法模式定理相似。

本文研究了二进制模式定理在十进制实数中的整数编码的推广, 至于推广到十进制任意实数编码的模式定理, 尚需慎重, 因为十进制实数编码遗传算法的编码方式、遗传算子的设计及其影响与二进制编码有诸多不同。当然采用将带小数的实数放大若干倍并取整的方法可将任意实数近似转化成整数,

(下转第 53 页)

- rithm on transputer based parallel processing systems. *Parallel Problem Solving from Nature*, 1991. 145~149
- 10 Goldberg D E. Genetic and evolutionary algorithms come of age. *CACM*, 1994, 37(3): 113~119
  - 11 Goldberg D E. Personal communication. June 1996
  - 12 Gordon V S. Locality in genetic algorithms. In: Schaffer J D, et al. eds. *Proc. of the First IEEE Conf. on Evolutionary Computation, Volume 1* Piscataway, NJ: IEEE Service Center, 1991. 428~432
  - 13 Gordon V S, Whitley D. Serial and parallel genetic algorithm as function optimizers. In: Forrest S. ed. *Proc. of the 5th Intl. Conf. on Genetic Algorithms*. same to [1], 1993. 177~183
  - 14 Gorges-Schleuter M. Asparagos96 and the traveling salesman problem. In: Back, T. ed. *Proc. of the 4th Intl. Conf. on Evolutionary Computation*. New York: IEEE Press, 1997
  - 15 Gruau F. Neural network synthesis using cellular encoding and the genetic algorithm. Unpublished doctoral dissertation, L'Universite Claude Bernard-Lyon I, 1994
  - 16 Harik G, et al. The gambler's ruin problem, genetic algorithms, and the sizing of populations. In: Back, T. ed. same to [14], 1997. 7~12
  - 17 Hauser R, Manner R. Implementation of standard genetic algorithm on MIMD machines. In: Davidor Y, et al. eds. *Parallel Problem Solving from Nature, PPSN III*. Berlin: Springer-Verlag, 1994. 504~513
  - 18 Kroger B, et al. Parallel genetic packing on transputers. In: Stender J. ed. *Parallel Genetic Algorithms: Theory and Applications*. Amsterdam: IOS Press, 1993. 151~185
  - 19 Levine D. A parallel genetic algorithm for the set partitioning problem. [Tech. Rep. No. ANL-94/23]. Argonne, IL: Argonne National Laboratory, Mathematics and Computer Science Division, 1994
  - 20 Lin S-H, et al. Investigating parallel genetic algorithms on job shop scheduling problem. In: Angeline P, et al. eds. *6th Intl. Conf. on Evolutionary Programming*. Berlin: Springer Verlag, 1997. 383~393
  - 21 Manderick B, Spiessens P. Fine-grained parallel genetic algorithms. In: Schaffer J D. ed. *Proc. of the 3rd Intl. Conf. on Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann, 1989. 428~433
  - 22 Marin F J, et al. Genetic algorithms on LAN-message passing architectures using PVM: Application to the routing problem. In: Davidor Y, et al. eds. *Parallel Problem Solving from Nature, PPSN III*. Berlin: Springer-Verlag, 1994. 534~543
  - 23 Sarma J, Jong K D. An analysis of the effects of neighborhood size and shape on local selection algorithms. In: *Parallel Problem Solving from Nature IV*. Berlin: Springer-Verlag, 1996. 236~244
  - 24 Talbi E-G, Bessi'ere P. A parallel genetic algorithm for the graph partitioning problem. In: *Proc. of the Intl. Conf. on Supercomputing*. Cologne, 1991
  - 25 Tanese R. Parallel genetic algorithm for a hypercube. In: Grefenstette J J. ed. *Proceedings of the 2nd Intl. Conf. on Genetic Algorithms*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987. 177~183

(上接第 56 页)

沿用上述的推导过程,也可得出十进制实数编码的模式定理,但总感觉不很完善,尚有待进一步深入研究。

### 参 考 文 献

- 1 席裕庚,等. 遗传算法综述. *控制理论与应用*, 1996, 13(6): 697~708
- 2 刘勇,康立山,陈毓屏. *非数值并行算法—遗传算法*. 北京:科学出版社, 1995
- 3 Srinivas M, Patnaik L M. *Genetic Algorithms: A Survey*
- 4 Vose M D. Generalizing the Notion of Schema in Genetic Algorithms. *Artificial Intelligence*, 1991, 50: 385~396
- 5 Goldberg D E. Real-Coded Genetic Algorithm. *Virtual Alphabets and Blocking*. *Complex Systems*, 1991, 5: 139~167
- 6 Wright A H. Genetic Algorithms for Real Parameter Optimization. In: Rawlins G J E, ed. *Foundations of Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann, 1991. 205~218
- 7 Michalewicz Z, et al. A Modified Genetic Algorithm for Optimal Control Problems. *Computers Math. Applic.*, 1992, 23(12): 83~94