遗传算法编码策略研究*

解 庆[1]，赵小强[2]
(1. 甘肃蓝科石化高新装备股份有限公司，甘肃 兰州 730070; 2. 兰州理工大学 电信学院，甘肃 兰州 730050)

摘 要: 遗传算法是一类基于自然选择和自然遗传机制的自适应全局优化概率搜索算法，编码策略是设计遗传算法的一个重要步骤，通过研究二进制码和格雷码的编码策略，分析了编码差异、个体差异和适应度差异之间的关系，得到了两种不同编码对遗传算子搜索能力的影响和它们的特性。

关键词: 遗传算法; 编码策略; 二进制码; 格雷码
中图分类号: TP18

遗传算法是一类借鉴生物界自然选择和自然遗传机制的自适应全局优化概率搜索算法，它起源于 60 年代对自然和人工自适应系统的研究，主要特点是群体搜索策略和群体中个体之间的信息交换和搜索不依赖于梯度信息，它尤其适用于处理传统搜索方法难于解决的复杂和非线性问题。作为一种全局优化搜索算法，以其简单、通用、较强的自适应性和鲁棒性，以及适于并行处理等特点，被广泛应用于组合优化、机器学习、自适应控制、规划设计和人工生命等领域[1]。

按照遗传算法的工作流程，当用遗传算法求解问题时，必须在目标问题实际表示与遗传算法的染色体位串结构之间建立联系，即确定编码，遗传算法以决策变量的编码作为运算对象，三种基本的遗传算子的选择和设计都依赖于编码的方式[2]，编码策略对于遗传算子，尤其是对交叉和变异算子的功能和设计有很大的影响。编码作为遗传算法流程的第一步，在遗传算法中起着重要作用，因而提出了许多不同的编码方法[3]。

主要分析了两种主要的编码方式(二进制码和格雷码)自身具有的特点，以及这些特点对于算法的搜索能力的影响。

1 遗传算法流程

遗传算法将 m 维决策向量 $X=[x_1,x_1,\cdots,x_m]^T$ 用 m 个记号 $X_i(i=1,2,\cdots,m)$ 所组成的符号串来表示:

$$X=X_iX_2\cdots X_m\Rightarrow X=[x_1,x_1,\cdots,x_m]^T$$

把每一个 X_i 看作一个遗传基因，X 就可以看做是由 m 个遗传基因所组成的一个染色体。一般情况下，染色体的长度是固定的，但对于一些问题 m 也可以是变化的，根据不同的情况，等位基因可以是一组整数，也可以是某一范围内的实数值，或者是一个纯粹的记号。最简单的等位基因是由 0 和 1 这两个整数组成的，相应的染色体就可表示为一个二进制符号串。这种编码所形成的排列形式 X 是个体的基因型，与它对应的 X 值是个体的表现型。对于每一个个体 X，要按照一定的规则确定出其适应度，个体的适应度与其对应的个体表现型 X 的目标函数值相关联，X 越接近于目标函数的最优点，其适应度越大; 反之，其适应度越小。

遗传算法中，决策变量 X 组成了问题的解空间，对问题最优解的搜索是通过对染色体 X 的搜索过程来进行的，由所有的染色体 X 就组成了问题的搜索空间。遗传算法的运算对象是由 N 个个体所组成的集合，称为群体，遗传算法的运算过程也是一个反复迭代过程，第 t 代群体记做 $P(t)$，经过一代遗传和进化后，得到第 $t+1$ 代群体，它们是由多个个体组成的集合，记做 $P(t+1)$，这个群体不断地经过遗传和进化操作，并且过次都按照优胜劣汰的规则将适应度较高的个体更多的遗传到下一代，这样最终在群体中将会得到一个优良的个体 X，它所对应的表现型 X 将达到或接近于问题的最优解 X^*。

2 编码策略

编码策略是设计 GA 的一个重要步骤。三种基本 GA 算子的选择和设计都依赖于编码的形式。编码成为 GA 应用中的首要问题，因而对编码策略的研究也成为研究 GA 的热点之一，见表 1。

* 项目资助: 甘肃省自然科学基金(1112RJZA028)。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1 编码策略的研究与结论[4-6]

编码策略	研究者	研究结论
二进制编码	Holland	根据模式定理建议采用,并给出了最小字符集编码规则
动态变量编码	Schraudolph	通过对 De - Jong 的 5 函数进行测试表明: 动态变量编码比普通二进制编码的优化效果好,可以克服早熟现象
双倍体编码	Goldberg Smith	双倍体具有长期记忆作用,采用动态 Knapsack 问题进行比较研究表明: 双倍体比单倍体的跟踪能力强
浮点数编码	Michalewicz Qi,Palmieri	Michalewicz 等比较了两种编码的优缺点,浮点数编码具有精度高、便于大空间搜索的优点; Qi 和 Palmieri 基于 Markov 链,假设群体无穷大,分析了浮点数编码的遗传算法的全局收敛性
大字符集编码	Antonisse Bosworth	Antonisse 从理论上证明了 Holland 在推导最小字符集编码原则时存在错误,并指出大字符集编码可提供更多模式
编码的同构性	Vose	扩展了 Holland 的模式概念,揭示了不同编码之间的同构性

3 编码分析

以前对编码策略的研究,多以仿真对比来说明不同编码策略对算法的影响。并没有从编码本身的区别上说明原因。Schraudolph 的动态编码研究、Goldberg 双倍体编码分析等研究就属于此类。对编码本身特点的分析是设计遗传基本算子的基础,因此,结合遗传算子分析编码的本身特点来研究。

3.1 二进制码分析

个体编码长度主要由搜索空间和问题精度要求决定。一般函数优化问题的编码精度是唯一的,考虑到实际应用中,不同决策变量可能有不同的精度要求,因此,将编码长度的结论扩展到多精度的情形。

定义决策变量个数为 m,变量 $x_i(i=1,2,\cdots,m)$ 满足 $x_i \in [u_i,v_i]$,x_i 的搜索精度为 δ_i,即:

$\forall$ 变量 x_i 的两个值 $x_i^1,x_i^2 \in [u_i,v_i]$,如果 $|x_i^1 - x_i^2| < \delta_i/2$,则认为 $x_i^1 = x_i^2$,则有:

$$L = \sum_{i=1}^{n}\left(\frac{v_i - u_i}{\delta_i} + 1\right) \tag{1}$$

二进制编码不能保持群体稳定性。这是由于个体和编码之间的映射关系决定的。即对于任意给定的第 t 代第 $i(i \in \{1,2,\cdots,n\})$ 个个体的二进制编码 X_t^i,如果只变异一位,产生新个体 X_t^*,不妨假设第 k $(k \in \{1,2,\cdots,m\})$ 个变量 x_k 发生变异,且第 k 个变量的编码长度为 l,定义变异前后个体的差异为 s,记第 k 个变量发生变异前后两个个体的数值差为 s^k:

$$s^k = |X_t^i - X_t^*| = \sum_{j=1}^{m}|x_j^i - x_j^*| = |x_k^i - x_k^*|$$

则 s 的最大值为:

$$s_{max}^k = \frac{v_k - u_k}{2^l - 1}2^{l-1} \tag{2}$$

s 的最小距离为:

$$s_{min}^k = \frac{v_k - u_k}{2^l - 1} \tag{3}$$

以上说明了对于二进制编码,变异操作不能保证父个体与新个体充分接近。即编码相近,但个体未必相近,这就导致了变异后的个体与原个体的差异不可预估,种群的稳定性较差。但是这个结论只说明了编码相差 1 位的个体的数值差的最大和最小值,通过下面的分析,将给出编码差异和数值差所有可能的对应关系,首先考虑只有一个连续变量 x 的情形,多变量编码时可以通过简单的组合得出类似结论。变量 x 满足 $x \in [u,v]$,在搜索精度 δ 时,根据(1)式得出的编码长度为 l,变量 x 对应的编码用 X 表示,对二进制编码而言,采用 B 表示,设二进制编码 B 可以表示为: $B = b_l b_{l-1}\cdots b_2 b_1$(其中 $b_{i,i=1,2,\cdots,l} \in \{0,1\}$)。任给变量 x 的两个相异值 x^1、x^2,他们分别对应的编码是 B^1、B^2。

由二进制数和整数之间的转换公式知,二进制编码从高位到低位的权值依次为: $2^{l-1},2^{l-2},\cdots,2^l,2^0$。则有个体 x 的编码 $B = b_l b_{l-1}\cdots b_2 b_1$ 对应的解码公式是:

$$x = u + \left(\sum_{i=1}^{l} b_i \cdot 2^{i-1}\right) \cdot \frac{v-u}{2^{l-1}} \tag{4}$$

其中,$\frac{v-u}{2^{l-1}} = \delta$。

数值差异为:

$$d_x(x^1,x^2) = \text{round}(|x^1 - x^2|/\delta) \tag{5}$$

其中,round( ) 表示就近取整函数。将解码公式(4)代入式(5)得:

$$d_x(x^1,x^2) = \left|\sum_{i=1}^{l}(b_i^1 - b_i^2) \cdot 2^{i-1}\right| \tag{6}$$

式(6)可以看作二进制码对应的数值差异定义,可以看出编码精度和变量的取值区间已经对数

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

值差异不存在影响。编码差异用 $d_H(X^1,X^2)$ 表示，定义为两个编码 X^1、X^2 的海明距离(Hamming Distance)。对于二进制码，编码差异为:

$$d_H(B^1,B^2)=\sum_{i=1}^{l}|b_i^1-b_i^2| \tag{7}$$

[定理 1] 当 $d_H(B^1,B^2)=1$ 时，如果固定其中一个个体编码 B^1，则有:

1) B^2 存在的种类个数为 $C_l^1=l$;

2) $d_x(x^1,x^2)$ 的可能取值分别为 $2^{l-1},2^{l-2},\cdots,2^l,2^0$;

3) $P\{d_x(x^1,x^2)=2^{k-1}\}=1/l$ (8)

其中($\forall k\in\{1,2,\cdots,l\}$)

证明: $d_H(B^1,B^2)=1\Rightarrow\sum_{i=1}^{l}|b_i^1-b_i^2|=1$

又$\because b_{i,i=1,2,l}\in\{0,1\}$

$\therefore$ $\exists$ 唯一的 k($k\in\{1,2,\cdots,l\}$)，

$$st\begin{cases}b_i^1\neq b_i^2, i=k\\ b_i^1=b_i^2, i\neq k\end{cases} \tag{9}$$

(8) 代入(6) 得:

$$d_x(x^1,x^2)=\left|\sum_{i=1}^{l}(b_i^1-b_i^2)\cdot 2^{i-1}\right|=|(b_i^1-b_i^2)\cdot 2^{k-1}|=2^{k-1}k\in\{1,2,\cdots,l\} \tag{10}$$

再由 k 在集合$\{1,2,\cdots,l\}$ 内取值的任意性知(8) 式成立。

下面以十进制数 0 ~ 7 的数值差异(数值差) 和二进制编码差异的对比为例来说明以上结论，见表 2，其中，表的首行和首列的数据格式为“个体数值/二进制编码”，表内的数据格式为“数值差异/编码差异”。从表中得不出数值差异和编码差异的有规律性的联系。但是从式(10) 可得如下结论: 编码差异为 1 的两个个体的差值可能为$\{4,2,1\}$，或者表示为$\{2^{3-1},2^1,2^0\}$。即，一发生变异产生的新个体与父个体的数值差异是无法预估的。表 2 的* 部分就说明了这一点。

表 2　对于二进制编码方式数值差异与编码差异的关系

	0/000	1/001	2/010	3/011	4/100	5/101	6/110	7/111
0/000	0/0	1/1	2/1	3/2	4/1	5/2	6/2	7/3
1/001	1/1*	0/0	1/2	2/1*	3/2	4/1*	5/3	6/2
2/010	2/1*	1/2	0/0	1/1*	2/2	3/3	4/1*	5/3
3/011	3/2	2/1*	1/1*	0/0	1/1*	2/2	3/2	4/1*
4/100	4/1*	3/2	2/2	1/1	0/0	1/1*	2/1*	3/2
5/101	5/2	4/1*	3/3	2/2	1/1*	0/0	1/2	2/1*
6/110	6/2	5/3	4/1*	3/2	2/1*	1/2	0/0	1/1*
7/111	7/3	6/2	5/3	4/1*	3/2	2/1*	1/1*	0/0

二进制编码因为具有编解码操作简单易行、遗传操作便于实现、符合最小字符集编码原则、便于利用模式定理对算法进行理论分析等优点而得到广泛的应用。但是，由于二进制编码使得编码差异的大小不能代表数值差异的大小，导致变异算子的局部搜索能力不足。

3.2　格雷码分析

二进制编码不便于反映所求问题的结构特征，对于一些连续的函数优化问题等，也由于遗传运算的随机特性而使其局部搜索能力较差。为改进这个特性，人们提出了格雷码(Gray Code) 来对个体进行编码。例如，十进制数 0 ~ 7 之间的二进制码和相应的格雷码见表 3。

格雷码是这样的一种编码方法: 其连续的两个整数所对应的编码值之间仅仅只有一个码位是不相同的，其余码位都完全相同。

表 3　二进制码与格雷码对比

十进制数	0	1	2	3	4	5	6	7
二进制码	000	001	010	011	100	101	110	111
格雷码	000	001	011	010	110	111	101	100

假设个体 x 的二进制编码为 $B=b_lb_{l-1}\cdots b_2b_1$，其对应的格雷码为 $G=g_lg_{l-1}\cdots g_2g_1$($g_{i,i=1,2,\cdots,l}\in\{0,1\}$)。由二进制编码到格雷码的转换公式为:

$$\begin{cases}g_l=b_l\\ g_i=b_{i+1}\oplus b_i, i=l-1,l-2,\cdots,l\end{cases} \tag{11}$$

由格雷码到二进制编码的转换公式为[5]:

$$\begin{cases}b_l=g_l\\ b_i=b_{i+1}\oplus g_i, i=l-1,l-2,\cdots,l\end{cases} \tag{12}$$

上面两种转换公式中，$\oplus$表示异或运算符。

格雷码的编码差异同样用编码的海明距离来表示:

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$d_H(G^1,G^2)=\sum_{i=1}^{l}|g_i^1-g_i^2| \tag{13}$$

[定理 2] $d_x(x^1,x^2)=1\Rightarrow d_H(G^1,G^2)=1$ (14)

证明: 由 $d_x(x^1,x^2)=1$ **及**(6)**得:**

$$d_x(x^1,x^2)=\left|\sum_{i=1}^{l}(b_i^1-b_i^2)\cdot 2^{i-1}\right|=1 \tag{15}$$

并且 $b_{i,i=1,2,\cdots,l}\in\{0,1\}$**,不妨设** $x^1>x^2$**,并令两个个体的二进制编码** B^1 **和** B^2 **的编码存在差异的最高位是** k**(其可能取值范围为** $k\in\{1,2,\cdots,l\}$**),则有:**

$$\begin{cases}b_i^1-b_i^2=0, i>k\\ b_i^1-b_i^2=1, i=k\\ b_i^1-b_i^2=-1, i<k\end{cases} \tag{16}$$

(16)代入(11)并由 $g_{i,i=1,2,\cdots,l}\in\{0,1\}$ **得:**

$$\begin{cases}g_i^1=g_i^2, i>k\\ g_l^1\neq g_l^2, i=k\\ g_l^1\neq g_l^2, i<k\end{cases} \tag{17}$$

由式(17)代入式(13)得,

$$d_H(G^1,G^2)=\sum_{i=1}^{1}|g_i^1-g_i^2|=1 \tag{18}$$

下面以整数 0 ~ 7 的格雷码编码为例,分析一下数值差异(整数值的差)与海明距离的关系,见表 4。其中,表的首行和首列的数据格式为"个体数值/格雷码",表内的数据格式为"数值差异/编码差异"。

表 4 对于格雷码编码方式数值差异与编码差异的关系

	0/000	1/001	2/011	3/010	4/110	5/111	6/101	7/100
0/000	0/0	1/1*	2/2*	3/1	4/2	5/3	6/2	7/1
1/001	1/1*	0/0	1/1*	2/2*	3/3	4/2	5/1	6/2
2/011	2/2*	1/1*	0/0	1/1*	2/2*	3/1	4/2	5/3
3/010	3/1	2/2*	1/1*	0/0	1/1*	2/2*	3/3	4/2
4/110	4/2	3/3	2/2*	1/1*	0/0	1/1*	2/2*	3/1
5/111	5/3	4/2	3/1	2/2*	1/1*	0/0	1/1*	2/2*
6/101	6/2	5/1	4/2	3/3	2/2*	1/1*	0/0	1/1*
7/100	7/1	6/2	5/3	4/2	3/1	2/2*	1/1*	0/0

从表 4 的* 部分可知,数值差异为 1 时,其编码差异全部是 1,验证了(13)的成立。从定理 2 可以看出,格雷编码方式使相近个体具有相似编码,这是采用格雷码的遗传算法局部搜索能力强的原因。类似于上面的分析,可以给出数值差异为任何一个值时的编码差异。但是,从遗传算法的运行过程可知,需要的编码方式是满足编码相似可以推导出个体相近的编码策略。定理 2 表明了格雷码数值差异和编码差异的关联性要好于二进制码。

4 结论

主要分析了两种主要的编码方式(二进制码和格雷码)自身具有的特点,以及这些特点对于算法的搜索能力的影响。这种分析方法也适用于十进制编码和符号编码,但浮点数编码不同于二进制码等编码方式,需要结合具体的遗传算子通过搜索能力的对比,来分析浮点数编码和其他编码方式对某一类问题的优劣对比。通过分析个体差异和编码差异的关系,得出了格雷码的一些有利于遗传算法局部搜索的特性,这是格雷码得到广泛应用的原因。

参考文献:

[1] Chao Chen, Jianghai Xia, Jiangping Liu, Guangding Feng. Nonlinear inversion of potential – field data using a hybrid – encoding genetic algorithm [J]. Computers & Geosciences, 2006, 132(2): 230-239.

[2] 李敏强,寇纪淞,林丹,李书全. 遗传算法的基本理论与应用 [J]. 北京: 科学出版社, 2002.

[3] Alan Crispin, Paul Clay, Gaynor Taylor etl. Genetic Algorithm Coding Methods for Leather Nesting [J]. Applied Intelligence 23, 0 – 20, 2005.

[4] Serkan Bekiro lu, Tayfun Dede, Yusuf Ayvaz. Implemen tation of different encoding types on structural optimization based on adaptive genetic algorithm [J]. Finite Elements in Analysis and Design, 2009, 45(11): 826-835.

[5] Ying – Hua Chang, Young – Chang Hou. Dynamic programming decision path encoding of genetic algorithms for production allocation problems [J]. Computers & Industrial Engineering, 2008, 54(1): 53-65.

[6] Carlos Fernandes, Agostinho C. Rosa. Self – adjusting the intensity of assortative mating in genetic algorithms [J]. Soft COMPUT, 2008(12): 955-979.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net