第 28 卷 第 1 期
2002 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 1
Jan., 2002

短文

遗传交叉运算的可达性研究[1)]

张军英　许 进　保 铮
（西安电子科技大学电子工程研究所　西安　710071）
（E-mail：jyzhang@xidian.edu.cn）

摘　要　定义了遗传交叉运算的可达性及其达概率的概念，指出传统的单点交叉运算只使得参与交叉运算的个体对所张成子空间的边缘是可达的，且为非均匀可达的，从而大大地限制了该运算的搜索能力. 为此，讨论了一致交叉运算的可达性，指出它使得参与交叉运算的个体对所张成子空间的全空间都是可达的，且可以构造交叉字串使得它是均匀可达的，从而有效提高算法的搜索能力. 同时讨论了个体对交叉运算的可达性与群体进行交叉运算的可达性的关系.

关键词　标准交叉运算，一致交叉运算，个体，群体，可达集合，可达概率
中图分类号　O229

ATTAINABILITY OF GENETIC CROSSOVER OPERATOR

ZHANG Jun-Ying　XU Jin　BAO Zheng
(*Electronic Engineering Research Institute, Xidian University, Xi'an*　710071)
(E-mail: jyzhang@xidian.edu.cn)

Abstract　Attainability of genetic crossover operation is introduced, and the attainability of both canonical and uniform crossover operations are analyzed. The result is that the two offspring from their parents, upon which the canonical crossover is undertaken, are only on the rim of the subspace their parents span, while the two offspring are on the overall such subspace when uniform crossover is undertaken; also, the rim of the subspace is attained in an uneven probability for canonical crossover, while for uniform crossover the crossover string is designed to make the overall subspace attainable in an even probability. This attainability and attainable probability analysis for crossover operation indicates that the uniform crossover is much powerful than canonical crossover in problem space searching ability.

Key words　Attainable set, attainable probability, canonical crossover, uniform crossover, individual, population

1) 国家自然科学基金(69971018)资助.
收稿日期　1999-10-14　　收修改稿日期　2000-11-19

1　引言

遗传算法作为一种统计优化方法,在控制、人工神经网络训练、组合优化等领域得到了很好的应用.关于遗传运算性质的研究,文献[1]运用有限马尔可夫模型分析标准遗传算法的收敛性,文献[2]对遗传算法的适应度设计进行了研究;文献[3]通过理论分析和模拟实验研究了不同实数编码交叉操作的搜索效率;在应用方面,文献[4]通过引入粒度概念同时对前馈网络的连接权和结构进行遗传训练;文献[5,6]分别运用遗传算法求解二次分配问题和图的划分等问题.本文则研究二值编码遗传交叉运算的可达集合的位置、大小和概率分布情况,结果表明,传统的单点交叉运算只是"在已有空间的一个特定局部上做非均匀搜索"[7],而一致交叉运算更符合人们对交叉运算的"在已有空间中均匀搜索"[7]的初衷.这里提出的对遗传交叉运算的可达性质的研究方法同样适用于其它(如变异)遗传运算的研究.

2　解空间的可达性

在运用遗传算法进行统计寻优之前,通常是将问题的可行域进行编码,当采用二值编码形式,设码长为 n,则问题的可行空间(也称为解空间)为 $B^n=\{0,1\}^n$, 即为 n 维超立方体的所有顶点,则每一顶点 $v\in B^n$ 都是解空间中的一个个体.若将对群体 $V=\{v_i\}$ 上所进行的遗传运算统一记为 g,将在 V 上所进行的交叉运算和变异运算分别记为 c,m, 注意 g,c,m 均为统计遗传运算,即以一定概率发生的遗传运算,我们有如下定义.

定义 1. 对于群体 $V=\{v_i:v_i\in B^n,\ i=1,2,\cdots,m\}$, V 经过遗传运算 g 可能得到的子代个体的集合为 $G=\{g(v_i):i=1,2,\cdots,m\}$,则$\forall\ v\in G$,称 v 是 $g(V)$可达的,称 G 是 V 的 g 可达集合;对于$\forall\ v\notin G$,称 v 是 g 不可达的.特别的,对于交叉运算,$\{c(v_1,v_2)\}$为(v_1,v_2)的交叉可达集合;对于变异运算,$\{m(v)\}$是 v 的变异可达集合.

作者认为不能得出可达集合越大搜索能力就越强的结论,如果这样的话,随机搜索的能力就最强,也远不用研究其它的搜索算法了.但是我们还认为,一定不能得出可达集合越小搜索能力就越强的结论,即可达集合大小的研究对搜索能力的提高是有好处的,尤其在搜索过程的初期阶段.在这个阶段,一般说,遗传运算的可达集合越大,亦即其可达顶点覆盖解空间的程度越高,对应的遗传运算对解空间的搜索能力越强.本文重点研究单点交叉运算和一致交叉运算的可达集位置、大小和概率分布等有关问题.

2.1　单点交叉运算的可达性

例 1. 有两个码长为 4 的个体 0000 和 1111 已被选做参与交叉运算,交叉点的选取只有以下几种可能:选在第 1,2,3 位后,并以一定概率进行交叉,因此交叉运算的结果 c(0000, 1111)分别为(0000,1111), (1000,0111), (1100,0011), (0001,1110),即(0000,1111)的可达集合为{0000, 1111,1000, 0111, 1100, 0011, 0001, 1110},可达集合中共有 8 个个体,示于图 1 中(打三角的点).

定义 2. 设群体为 $V=\{v_i,\ i=1,2,\cdots,p\}$,称

$$B(V)=\{(x_1x_2\cdots x_n):x_j=v_{1j},\ \text{if}\ \oplus_{i=1}^{p}v_{ij}=1;\ x_j\in B,\ \text{if}\ \otimes_{i=1}^{p}v_{ij}=1;\ j=1,2,\cdots,n\}\tag{1}$$

为由 V 张成的子空间 B^d,其维数为 $d=|\otimes_{i=1}^{n}v_i|$. $B(V)$, 也记做 $B(v_1,v_2,\cdots,v_p)$. 其中$\oplus$,

⊗分别表示同或运算和异或运算.

定理 1. 设(v_1,v_2)所张成的子空间为B^d,则(v_1,v_2)经单点交叉运算的可达集合大小为$2d$.

证明. 设(v_1,v_2)所张成子空间为B^d,则$d=|v_1\otimes v_2|$,总可以通过坐标变换将v_1变换为分量全为0的向量,即$v_1'=00\cdots 0$,这时,对于$v_{1j}\otimes v_{2j}=1$的位j,有$v_{1j}'\otimes v_{2j}'=1$,即$v_{2j}'=1$,再经过下标换序(同时交叉点也发生了移动),则v_2'总可以写成前d个分量全为1而其余分量全为0的向量.尽管实际交叉点可处于v_1,v_2的第$1\sim n-1$位后的共$n-1$个位置上,我们在其第0位后加一个交叉点以生成v_1,v_2来表示不发生交叉运算所得到的结果(因为交叉是以一定的概率发生的),这样交叉点就可处于v_1,v_2的第$0\sim n-1$位后的共n个位置上,即v_1',v_2'的第$0\sim n-1$位后共n个位置上.明显地这样的坐标变换不改变可达集合的大小,即(v_1,v_2)的可达集合大小与(v_1',v_2')的可达集合大小相同.在新的坐标系下,对于$d<n$的情况,参与交叉的两个个体可表示为

$$v_1'=\{\overbrace{0,0,\cdots,0}^{d},0,0,\cdots,0\},\quad v_2'=\{\overbrace{1,1,\cdots,1}^{d},0,0,\cdots,0\} \tag{2}$$

当交叉点选在第$1,2,\cdots,d-1$位后时可以分别生成一对新的个体,而当交叉点选在第$0,d,\cdots,n$位后时生成一对新的相同个体,且其所生成的个体与不交叉的结果相同,因此$c(v_1,v_2)$的可达个体总数为$2d$. 对于$d=n$的情况,同理可得$c(v_1,v_2)$的可达个体总数为$2d$. 证毕.

从定理1可以看到:1)我们知道,通常$|v_1\otimes v_2|$较大,这时$B(v_1,v_2)$的维数$d=|v_1\otimes v_2|$较高,其中共有2^d个个体.而单点交叉运算使得这2^d(指数级)个个体中只有$2d$(线性级)个个体是可达的,其余个体是不可达的,从而单点交叉运算大大限制了遗传进化的搜索能力. 2)如果我们将一个超立方体图最外圈连线上的顶点集合称为该超立方体图的边缘,该超立方体中其余顶点的集合称为该超立方体的内部,则两个个体进行单点交叉运算使得可达个体仅处于它们所张成的子超立方体图的边缘,而无法进入它的内部.以图1为例,(0000,1111)经过一次单点交叉运算的可达集合为图1中打三角的个体,这些个体正好处在图1所给出的4维超立方体的最外圈连线上,即是由0000和1111所张的子超立方体(在这里正好是解空间的全空间)的边缘,而其余个体{0100,0010,1010,1001,0110, 0101, 1101, 1011}则都不在该子体的最外圈连线上,即为其内部,这些个体是(0000,1111)经一次单点交叉运算是不可达的. 3)在实际的的遗传进化过程中,经选择后得到的群体参加交叉运算,则这一群体的可达集合为各对交叉个体的可达集合的并,由于每对个体的可达集合仅为这对个体所张成的子超立方体的边缘,因此群体的可达集合是各对参与交叉运算的个体所张成子超立方体边缘的并,示于图2(a)中.通常每对个体张成的子超立方体比其边缘要大得多,因此,即使最优解在群体所张成的子空间上,只要不在各对个体所张成的子空间的边缘,这时最优解是不可达的,则在群体上的这次交叉运算是一定无

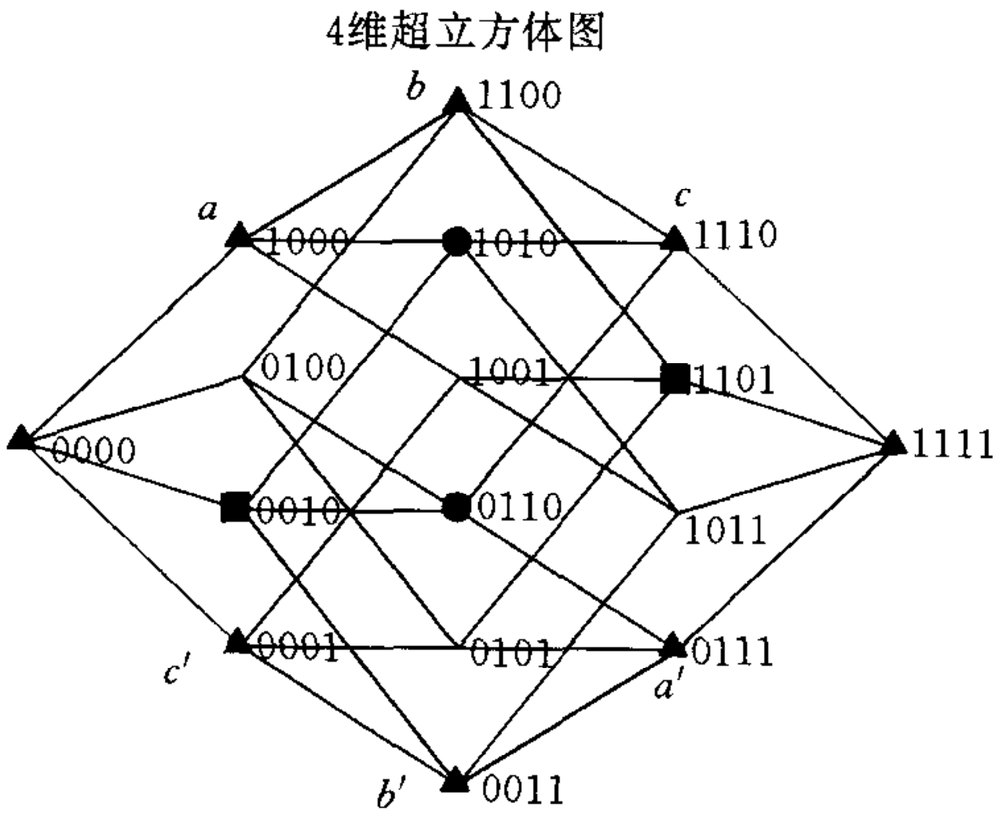


图 1 个体交叉运算的可达集合

注. 打三角的点表示(0000,1111)经单点交叉运算的可达集合,打圆圈/方块的点表示(0000,1111)经交叉字串为0110/0010的一致交叉运算的一对可达个体.

法搜索到这个最优解的.

2.2　一致交叉运算的可达性

定义 3. 对于长度为 n 的两个个体 $v_1, v_2 \in B^n$，设置长度为 n 的二进交叉字串 $t \in B^n$，则 (v_1, v_2) 的一致交叉运算，记为 $(v_1', v_2') = c_t(v_1, v_2)$，定义为

$$\begin{cases} v_{1i}' = v_{1i},\ v_{2i}' = v_{2i}, & \text{if } t_i = 0, \\ v_{1i}' = v_{2i},\ v_{2i}' = v_{1i}, & \text{if } t_i = 1, \end{cases} \quad i = 1,2,\cdots,n \tag{3}$$

当交叉字串以一定的概率选取，式(3)的交叉也以一定概率发生时，称为统计一致交叉运算，亦称一致交叉运算.

定理 2. 若 $(v_1', v_2') = c_t(v_1, v_2)$，有 $v_1', v_2' \in B(v_1, v_2)$.

证明. 明显地，对于 $v_{1i} = v_{2i}$ 即 $v_{1i} \oplus v_{2i} = 1$ 的位 i，$t_i = 1$ 或 $t_i = 0$ 都使得 $v_{1i}' = v_{2i}' = v_{1i}$，即 $v_{1i} \oplus v_{2i} = 1$；对于 $v_{1i} \neq v_{2i}$ 即 $v_{1i} \otimes v_{2i} = 1$ 的位 i，$t_i = 1$ 或 $t_i = 0$ 都使得 $v_{1i}' \neq v_{2i}'$，由 $B(v_1, v_2)$ 的定义(定义 2)，故有 $v_1', v_2' \in B(v_1, v_2)$. 证毕.

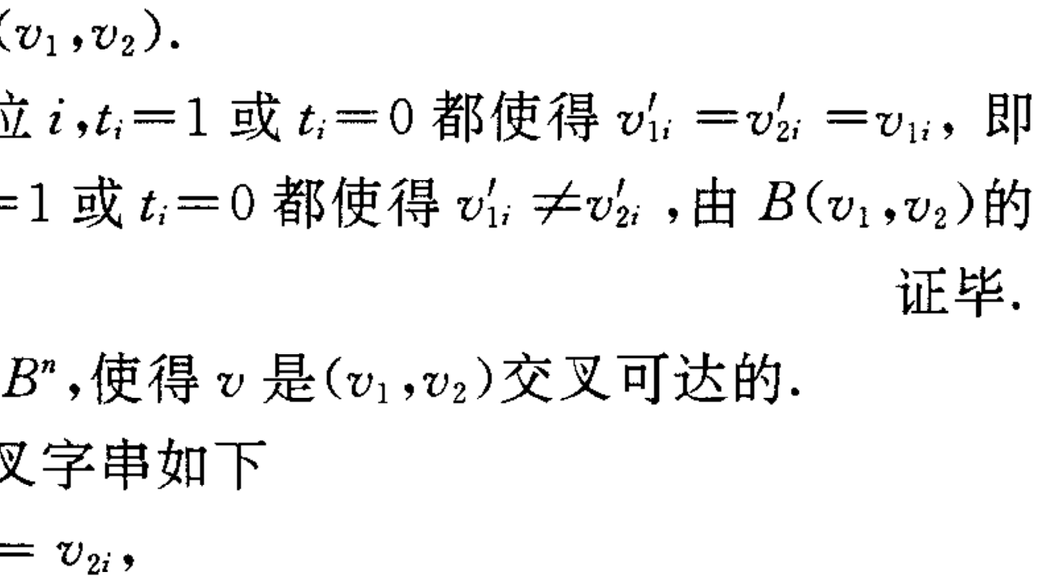
定理 3. $\forall\ v \in B(v_1, v_2)$，总存在交叉字串 $t \in B^n$，使得 v 是 (v_1, v_2) 交叉可达的.

证明. 对于 $\forall\ v = v_1' \in B(v_1, v_2)$，只要构造交叉字串如下

$$t_i = \begin{cases} 0 \text{ or } 1, & \text{if } v_{1i}' = v_{1i} = v_{2i}, \\ 0, & \text{if } v_{1i}' = v_{1i} \neq v_{2i}, \\ 1, & \text{if } v_{1i}' = v_{2i} \neq v_{1i}, \end{cases} \quad i = 1,2,\cdots,n \tag{4}$$

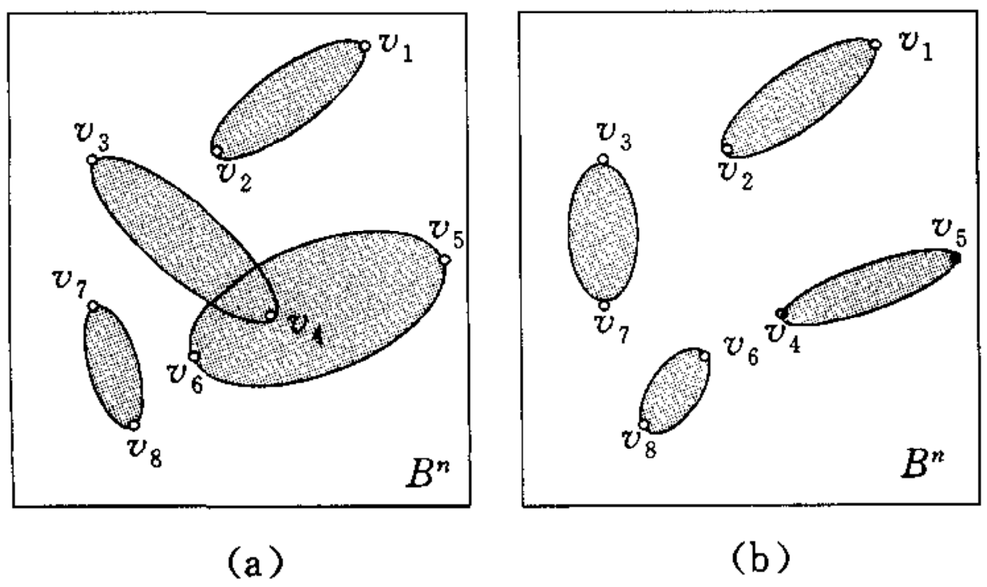
则可保证 $(v_1', v_2') = c_t(v_1, v_2)$，即 v_1' 是可达的. 证毕.

实际上单点交叉是一致交叉在交叉字串取为前 m 位(m 可为 $1,2,\cdots,n-1$)均为 1 其余位均为 0 的特殊情况. 由定理 3 可以看到：1)就两个个体进行一致交叉运算而言，对 (v_1, v_2) 的一致交叉运算使得 $B(v_1, v_2)$ 的每一个顶点都是可达的，因此对 (v_1, v_2) 进行一致交叉运算是在 (v_1, v_2) 所张成的子空间 $B(v_1, v_2)$ 上的全空间搜索，而不像单点交叉运算那样仅在 $B(v_1, v_2)$ 的边缘上搜索，从而更加符合交叉运算的“在已有空间上搜索”的初衷. 当两个参加交叉运算的个体之间的 Hamming 距离较大时，它们所张的子体空间中的个体数目就比这个子体空间的边缘上的个体数目要大得多得多. 因此，一致交叉运算与单点交叉运算相比，可以大大提高其运算的可达性. 2)从群体搜索过程上讲，群体经一致交叉运算的可达集合是各对参与交叉的个体对的可达集合的并. 尽管各对个体的可达集合是其所张成子空间的全空间，但实际上群体的可达集合通常不是群体所张子空间的全空间，图 2 给出了一个示意图. 3)群体可达集合的大小与参与交叉的个体配对有直接的关系，图 3(a)示出了 (v_1, v_2)，(v_3, v_4)，(v_5, v_6)，(v_7, v_8) 分别参与交叉运算的群体可达集合，而图 3(b)示出了 (v_1, v_2)，(v_3, v_7)，(v_4, v_5)，(v_6, v_8) 分别参与交叉运算的群体可达集合，很明显，这两个群体是一样的，但群体的可达集合却是不同的，其原因就是参与交叉运算的



(a) 单点交叉运算　　(b) 一致交叉运算

图 2　群体的可达集合



(a)　　(b)

图 3　相同群体交叉运算的配对不同时的可达集合

个体配对情况不同.

3 可达集合的达概率

由于交叉运算是统计运算,可达的个体只是以一定的概率获得,我们有如下定义

定义 4. 若 V' 是 $g(V)$ 的可达集合的某一个子集,则称 $g(V)$ 达到 V' 的概率为 V' 的 $g(V)$ 达概率,记为 $p_{g(V)}(V')$ 或简记为 $p(V')$.

3.1 单点交叉运算达概率的非均匀性

定理 4. 若经交叉点处于第 j 位后的单点交叉运算,则交叉后的个体 v 的达概率为

$$p(v)=\frac{l+1}{n-1}p_c,$$

其中 l 为与交叉点连续相邻且 v_1,v_2 的值相同的位数.

该定理的证明是显而易见的. 用一个例子来说明,若参与交叉运算的两个个体分别是 $v_1=1011110110$, $v_2=0101110101$,交叉点取为 5 时有 $l=5$. 因交叉点处于 3,4,5,6,7,8 位后的单点交叉运算结果均为 $v_1'=1011110101$, $v_2'=0101110110$,因此 $\{v_1',v_2'\}$ 的达概率为 $6/9p_c$,而交叉点取为 1,2,9 位后的交叉运算各生成一对新的个体,它们的达概率分别仅为 $1/9p_c$. 由命题可见,与交叉点连续相邻且 v_1,v_2 的值相同的位数越多,则交叉后所获得的个体对的达概率越大. 那么造成这一可达集合上可达个体达概率的非均匀性的原因是由于 v_1,v_2 处于 $B(v_1,v_2)=***11101**$ 造成的,很明显,子空间 $***11101**$ 比子空间 $*0*0*0*0*0$ 所造成的可达集合的非均匀性要大得多,因此其可达个体达概率的非均匀程度与参与交叉运算的一对个体所处的是 B^n 空间中的哪一个子空间有密切联系. 明显地,由于 $B(v_1,v_2)$ 处于 B^n 的某一特定的子空间而造成可达个体达概率的非均匀性是没有任何生物进化基础的.

3.2 一致交叉运算的达概率

定理 5. 若取 $t=v_1\otimes v_2$,且交叉概率 p_c 取为 1/2,则 $B(v_1,v_2)$ 是 (v_1,v_2) 经一致交叉运算均匀可达的.

证明. 不失一般性,v_1,v_2 总可以通过坐标变换和下标交换表示为

$$v_1'=\{\overbrace{0,\cdots,0}^{d\text{个}0},0,\cdots,0\},\quad v_2'=\{\overbrace{1,\cdots,1}^{d\text{个}1},0,\cdots,0\},$$

则 $B(v_1',v_2')=(\overbrace{*,\cdots,*}^{d\text{个}*},0,\cdots,0)$ 即为 (v_1,v_2) 在新坐标系下的可达集合,维数为 $d=|v_1\otimes v_2|$.

因 $t=v_1\otimes v_2$, 在新坐标系下对应的交换字串为 $t'=v_1'\otimes v_2'=(\overbrace{1,\cdots,1}^{d\text{个}1},0,\cdots,0)$,且每位的交叉概率 $p_c=1/2$,从而,$B(v_1',v_2')$ 中的每个个体的达概率均为 $\left(\frac{1}{2}\right)^d$,而 $B(v_1',v_2')$ 中共有 2^d 个个体,从而有 $p(B(v_1',v_2'))=2^d\times\frac{1}{2^d}=1$, 故 $B(v_1,v_2)$ 是 (v_1,v_2) 经一致交叉运算均匀可达的. 证毕.

在实际实现这个均匀可达的一致交叉运算的过程中,只需对位值不同即 $v_{1k}\otimes v_{2k}=1$ 的位 k,以概率 1/2 选择 0,1/1,0,并将其赋给 v_{1k},v_{2k},而无需真正设置交叉字串 t,操作简单方便.

定理 6. 若对于 $v_1\otimes v_2$ 为 1 的每一位,t 的对应位取 1 的概率为 p_t,且交叉概率为 1/2,

则与 v_1 距离为 k 的可达个体的达概率为 p_t^k.

证明. 若 $(v_1',v_2')=c_t(v_1,v_2)$,因为 $d(v_1,v_1')=d(v_2,v_2')=\sum_{j=1}^{n}t_j|v_{1j}-v_{2j}|=|t|$, 又 $|t|=1,2,\cdots,d$ 的概率分别为 $p_t,p_t^2,\cdots,p_t^d$,故与 v_1 距离为 k 的可达个体的达概率为 p_t^k. 证毕.

上述定理表明,若按定理 6 来构造交叉字串,则与 v_1 距离越近的个体,其达概率越大,反之,与 v_1 距离越远的个体,其达概率越小.

4 结论

本文定义了遗传运算的可达性和达概率的概念,指出传统的单点交叉运算只使得参与交叉运算的个体对所张成子空间的边缘是可达的,且为非均匀可达的,而不是其全空间是可达的,从而大大地限制了该运算的搜索能力,为此,本文讨论了一致交叉运算的可达性,指出它使得参与交叉运算的个体对所张成子空间的全空间都是可达的,且可以构造交叉字串使得它是均匀可达的,从而有效提高算法的搜索能力.讨论了个体对交叉运算的可达性与群体进行交叉运算的可达性的关系.从这些讨论中可以看到,一致交叉运算更符合人们对交叉运算的"在已有空间中均匀搜索"的初衷,具有更优良的搜索特性,而单点交叉运算则只是"在已有空间的一个特定局部上非均匀搜索"的结果.

参 考 文 献

1 Gunter Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE Trans. Neural Networks*, 1994, **5**(1):96～101

2 辛绯等. 遗传算法的适应度函数研究. 系统工程与电子技术, 1998, (11):58～62

3 黄晓峰,潘立登,陈标华,李成岳. 实数编码遗传算法中交叉操作的效率分析. 控制与决策, 1998-S1

4 Vittorio Maniezzo. Genetic evolution of the topology and weight distribution of neural networks. *IEEE Trans. Neural Networks*, 1994, **5**(1):39～53

5 Volker Nissen. Solving the quadratic assignment problem with clues from nature. *IEEE Trans. Neural Networks*, 1994, **5**(1):66～72

6 陈国良等. 遗传算法及其应用. 北京:人民邮电出版社,1996

7 Goldberg D E, Lingle R. Alleles, the travelling salesman problem. In: Proceedings of International Conference on Genetic Algorithms and Their Applications, 1985. 154～159

张军英　教授、博士. 1982 年获陕西理工大学自动控制专业学士学位,1985 年获西安电子科技大学计算机应用专业硕士学位,此后获西安电子科技大学信号与信息处理专业博士学位. 目前主要从事人工神经网络、遗传算法、智能信息处理等方面的研究工作.

许　进　教授、博士. 西安交通大学(管理工程)工学博士,北京理工大学(应用数学)理学博士. 主要研究方向为电路与系统、神经网络、图论、管理工程等.

保　铮　1953 年毕业于解放军通信工程学院. 现在是中国科学院院士、中国电子学会会员和雷达信号处理重点实验室学术委员会主任. 研究方向为雷达信号处理与检测.