基于家族优生学的进化算法*

吴少岩　张青富　陈火旺

（国防科技大学计算机系　长沙　410073）

摘要　模拟进化有几种典型方法，分别强调自然进化过程的不同侧面. 人们已意识到恰当地结合这些方法将推进该领域的研究. 本文将现有进化方法的差别归结为“遗传链”与“进化链”之间的差别，提出一种新的进化模型，称之为“基于家族优生学的进化 FEBE(family eugenics based evolution)”. 新的模型用家族优生学的思想将遗传链与进化链溶为一体，将正交设计技术引入家庭的子代培植以加强个体的行为改进. 本文将 FEBE 模型用于求解困难的 Goldberg 欺骗问题，取得了满意的实验结果.

关键词　模拟进化，遗传算法，优生学，正交设计，欺骗问题.

近 30 年来，模拟进化的研究遵循了 3 种不同方法：遗传算法(GA)、进化规划(EP)和进化策略(ES). 这些方法有明显的类似性：(1)维持可行解群体；(2)解的性能(适应值)由已知目标函数来评估；(3)探索机制(算子)决定新试验解的生成. 其差别在于它们在不同层次上模拟进化过程，每种方法都强调了自然进化的不同侧面. GA 的倡导者[1,2]认为染色体的各种行为特性(“表现型”)取决于染色体的结构(“基因型”). 一般地，好的基因型产生好的表现型，而好的表现型又归咎于好的基因型. 根据这一信念，在 GA 的迭代过程中，选择机制让好的解有多次参与繁殖子女的机会，并淘汰差的解. 换言之，选择机制让好的基因型有更大的机会向后代遗传. “交配”是主要的遗传算子，它通过交换 2 个染色体(父代)的遗传材料产生新的子女. 子代的性能并不保证高于父代. 进化的优化信念只是建立在父子之间的“遗传链”上. 简言之，“好的父代一般会产生好的子代”. 与 GA 相对照，EP 和 ES[3]强调种群或个体层次上的行为进化. “行为链”意味着好的子女有资格生存，无论其父代性能如何. 基于遗传链的算法与基于行为链的算法也可用下述方法认识其差别：GA 中，适应值用于指导选择父代；而在 EP 和 ES 中适应值用于指导选择子女. 实际上，没有证据表明单一进化过程必须遵循其一. 在文献中，两类方法在求解许多困难的搜索优化问题时(甚至同一问题)都有出色表现. 遗传链与进化链的模拟并非互相排斥.

本文提出一种新的进化方法：“基于家族优生学的进化 FEBE(family eugenics based

* 本文研究得到国家 863 高科技项目基金和国防科技大学校预研项目资金资助. 作者吴少岩，1964 年生，博士，讲师，主要研究领域为进化计算，并行处理. 张青富，1965 年生，博士，讲师，主要研究领域为神经网络，信号处理，进化计算. 陈火旺，1936 年生，教授，博士导师，主要研究领域为软件自动化，计算机科学理论，人工智能.

本文通讯联系人：吴少岩，长沙 410073，国防科技大学计算机系

本文 1996-02-15 收到修改稿

evolution)"模型. FEBE 模型基本上是遗传系统的模拟,但它在"家庭"的层次上将行为进化溶入子代的培植. 使用FEBE 模型构造的各种算法已用于组合优化、数值优化与遗传程序设计. 限于篇幅,本文描述这一模型并求解具有 10 个子函数的 Goldberg 欺骗问题.

1 基于家族优生学的进化

我们将模拟进化看作函数优化器,目标是在复杂的问题空间中确定最优解或近似最优解. DeJong 认为常规遗传算法(CGA)[1]不是函数优化器,尽管已经提出作为函数优化器的许多 CGA 变形. Holland 的最初动机是为鲁棒的自适应系统设计一般框架. "……较好的方式是将 GA 当作顺序决策过程的优化器. 这种过程包含不确定性,缺乏先验知识,有噪声反馈和随时间变化因素. 具体而言,给定有限的试验次数,如何分配这些试验使累计获益最大?". [4]显然,函数优化器与 GA 的目标并不完全一致:前者关注单个最优点的发现,而 GA 将最大获益作为所付试验代价的函数. 因此改进函数优化器的进化方法不应束缚于常规遗传算法框架.

基于 GA 的函数优化器的现有工作一般是改进 GA 的某个组成部分. 例如,文献[5]用"级别比例"选择取代"按适应值比例"选择;"精英策略"(Elitist Strategy)常用于保留群体的当前最好解. "遗传表示"受到更多关注[6,7],除二进制编码外已提出各种数据结构及相应的遗传操作. 尽管多数 CGA 的变形缺乏理论依据,但大量实验结果支持了这些 CGA 的改进策略.

然而,这些基于 GA 的优化器基本上是遗传链的模拟. 实验与分析已证明基于行为链的进化规划与进化策略也是很有价值的优化方法. 一个自然的设想是将两类模拟进化的方法相结合,期望产生更强的优化工具. 本文的目的在于将遗传进化与行为进化在统一的模型中相结合,即所谓"基于家族优生学的进化"(FEBE).

优生学是通过控制遗传特性改进群体的科学,它研究个体与群体改良的过程和手段. 例如,人类优生学研究生育更强壮、更聪慧人的方式.

FEBE 模型可描述为下述基于群体的迭代过程(设 M 为偶数):

(1)随机生成由 M 个可行解(个体)组成的初始群体;

(2)计算群体中每一个体的适应值;

(3)基于适应值建立群体中每一个体的选择概率,依此概率分布,对群体作 M 次随机选取,将选出的个体放入"配对池";

(4)配对池中个体两两结合形成 $M/2$ 个家庭(每一家庭中高适应值个体为父本,另一个为母本);

(5)对每一家庭,双亲以某种优化的交配方式产生若干"中间子女",通过家庭内部竞争,仅 2 个子女成长为下一代的合法成员;

(6)由合法子女组成的群体接受变异操作;

(7)若已获得一满意解,中止;否则,继续第(3)步,进入下一次迭代.

该模型的关键步骤是第(5)步. 除群体中遗传材料的选择(第(3)步)之外,每一家庭作为基本单位对群体进化作出积极贡献. 家庭中,双亲以优化方式交配,培育若干中间子女. 家庭

形成一个局部竞争环境,使未成熟子女相互竞争,争取成为 2 个合法成员之一. 竞争的结果使 2 个最好子女保留而淘汰其余子女. 较之双亲,子女的行为明显地被进化了. 中间子女是否也同双亲竞争,在该模型中是可选方式. 若双亲参与竞争且取胜,则双亲被复制到下一代. 按优生学,这相应于双亲没有生育更好子女的能力,因而被剥夺生育权. 当然,这样的父本或母本在下一代群体仍有机会组成各自的新家庭.

FEBE 模型的思想源于生物界的某些证据. 很多生物组织产生多个受精卵,而后在演变中挑选子女,这种选择方式是为了减少亲代在较差子女上投入的资源.[8]我们的工作与 Tackett 的观点[8]有相似性,但方法是不同的.

图 1 解释了 FEBE 模型中代 t 向代 $t+1$ 的进化过程,虚框内指明一个家庭的进化. 在 FEBE 模型中,群体的进化是由亲代的选择和家庭子女的优化共同促进的. 这体现了遗传链同进化链的自然结合.

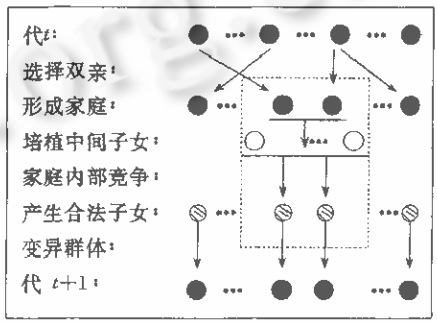

图1 FEBE模型的一次迭代

实际上,FEBE 模型也可用搜索技术的术语来阐述. 在 CGA 中,发生在 2 个选定解上的交配操作意味着:在给定的 2 个(可能的)好点附近随机采样 2 个新试验点,对新点性能的信念取决于 2 个给定点. 换言之,好点周围蕴含好点. 然而,在进化算法面临的多态(Multimodal)空间中,通常不是这种情况. 对空间中远离的 2 点,这种预测更不可靠. FEBE 模型改善了这种预测,它在 2 点决定的范围内以某种控制的方式采样几个点,并挑选 2 个好点作试验结果.

一个重要的问题是:如何以优化(控制)的方式生育子女或如何在给定点周围采样试验点? 实验优化技术为这一问题提供了答案. 早在 20 年代,英国统计学家 Fisher 使用拉丁方设计成功地解决了农业实验中的非均匀条件问题,并建立"实验设计"学科.[9]正交设计[10]是实验设计方法之一,它采用拉丁方与正交表安排试验,直接寻求优化解. 正交表源于拉丁方,是正交设计的基本工具,它是建立于均衡分布思想上的二维表.

1.1　正交表的概念

人们经常用实验来研究多种因素对产品性能、成本或数量的影响. 这里,性能、成本或数量是实验指标. 因素对指标的影响反映在指标值随因素状态或水平的改变而变化. 习惯上,$A,B,C\cdots$ 常用于表示多因素实验中的各因素. 假定 2 个因素实验中因素 A 有 r 个水平,B 有 s 个水平,则共有 $r\times s$ 个水平组合. 如果在实验中对每一水平组合都做一试验,则称实验为完全试验,否则称为部分试验.

现实问题中,实验往往是多因素多水平的. 人们自然提问:如何从全部水平组合中选取部分组合使部分试验代表完全试验? 正交表恰好按这一要求来安排多因素多水平实验. 表 1 给出一简单正交表 $L_4(2^3)$,它安排 3 因素 2 水平实验. 图 2 的均衡分布图解释 $L_4(2^3)$ 如何分配空间中的试验点. 按 $L_4(2^3)$,一个 3 因素 2 水平实验可安排如下:因素 A,B 和 C 分别对应表 1 中的列 A,B 和 C;数字'0'和'1'代表因素的 2 个不同水平. 这种安排从 2^3 种可能组合中选 4 种组合进行试验.

正交表可形式地定义为:

表1 正交表 $L_4(2^3)$

Trial No.	A	B	C
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

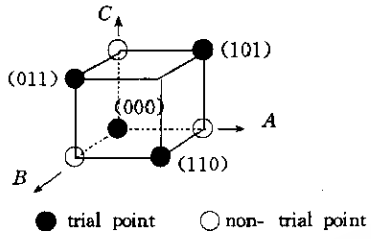

图2 $L_4(2^3)$安排3因素2水平实验

定义(正交表). 称矩阵 $H=(h_{ij})_{n*m}$是一个 $L_n(r_1\times\ldots\times r_m)$型正交表,如果

(i)$\forall\ j(j=1,\ldots,m)$,$h_{ij}\in\{0,1,\ldots,r_j-1\}$,$i=1,2,\ldots,n$.

(ii)$\forall\ j(j=1,\ldots,m)$, 在 $h_{ij}(i=1,\ldots,n)$中每一不同非负整数 $k(k\in\{0,1,\ldots,r_j-1\})$出现的次数都等于 n/r_j.

(iii)$\forall\ j_1,j_2(1\leqslant j_1,j_2\leqslant m)$由 H 的第 j_1,j_2 列组成的 $n\times 2$ 子阵中,每对不同非负整数数偶(k_1,k_2)出现的次数都等于 $n/(r_{j1}*r_{j2})$,$k_1\in\{0,1,\ldots,r_{j1}-1\}$,其中 $n,m,r_1,\ldots,r_m$ 是正整数,$r_j\geqslant 2(j=1,\ldots,m)$.

1.2 正交设计的优良性

正交表用于因素实验的优良性可用不同方式解释,这里仅通过总体与子样间的关系做简要分析.

按正交表定义,当利用它做部分试验时,保证了 2 条基本准则:(1)对每一因素,各水平参加试验的次数一样多;(2)对任意两因素,实现了包括所有水平组合的完全试验. 因此,就整体趋势而论,多因素多水平的部分试验对完全试验有一定代表性. 为了定量地评估这种代表性,下面的讨论引入随机变量与系统子样的概念. 将任一水平组合下的试验指标视作一随机变量 ζ. 由正交表确定的 n 个指标值形成一子样 $\zeta_1,\cdots,\zeta_n$,这不是简单随机子样. 因 n 次试验在完全水平组合中具有均衡分散性,不是随机选取的,称这种子样为系统子样. 我们关心的问题是:假定系统子样的极值为 λ,那么,在完全试验的所有指标值中 λ 名列第几? 若 $\zeta_1,\cdots,\zeta_n$ 是简单随机子样,有下述结论成立:

定理. 假定 ζ 的分布函数 $F(x)$连续,$\zeta_1,\cdots,\zeta_n$ 是简单随机子样,记 $\zeta_1^*\leqslant\zeta_2^*\ldots\leqslant\zeta_n^*$ 为其顺序统计量,则 $$E[F(\zeta_k^*)]=k/(n+1)\ (k=1,\ldots,n)$$

其中 $\zeta_1^*=\min\limits_{1\leqslant i\leqslant n}\zeta_i$,$\zeta_n^*=\max\limits_{1\leqslant i\leqslant n}\zeta_i$.

定理蕴含着子样的极大值 $\zeta_n^*=\max\limits_{1\leqslant i\leqslant n}\zeta_i$ 同另一子样极大值 $\eta_n^*=\max\limits_{1\leqslant i\leqslant n}\eta_i$ 之间的关系为 $\eta_n^*=F(\zeta_n^*)$,而 $E(\eta_n)^*=n/(n+1)$. 可见,η_n^* 的数学期望只与子样容量 n 有关,与总体 ζ 服从什么分布无关. η_n^* 的期望也反映了解的观察值分布规律的平均值. 这表明简单随机子样 $\zeta_1,\cdots,\zeta_n$ 中的最大值 ζ_n^* 优于总体中占比例 $n/(n+1)$的试验结果,而劣于总体中占比例 $1/(n+1)$试验结果. 一般地,系统子样中的最优者不次于简单随机子样中的最优者. 具体而言,例如,使用正交表 $L_8(2^7)$做部分试验时,那么所做的 8 个水平组合中的最优者,它在 2^7 个水平组合中优于占比例 8/9 的水平组合,即优于 113 个水平组合,就是说它在 128 个水平组合中名列前 15 名.

1.3 正交交配算子(OCX)

正交设计提供了家庭内部优育子女的方法. FEBE 模型中,正交表用于指导双亲的结合方式. 由正交表构造的新的交配算子称为"正交交配"算子(OCX). 一般地,OCX 依赖于具体的遗传表示. 本文我们以 $L_8(2^7)$(见表 2)为例说明二进制串上的 OCX 构造方式.

表2　正交表 $L_8(2^7)$

Trial No.	A	B	C	D	E	F	G
1	0	0	0	0	0	0	0
2	0	0	0	1	1	1	1
3	0	1	1	0	0	1	1
4	0	1	1	1	1	0	0
5	1	0	1	0	1	0	1
6	1	0	1	1	0	1	0
7	1	1	0	0	1	1	0
8	1	1	0	1	0	0	1

设一可行解 $P=(b_1,\ldots,b_n)$为二进制串,即 $b_i\in\{0,1\}$,$i=1,\ldots,n$. 随机划分 P 的下标集$\{1,\ldots,n\}$为 7 个不相交子集: $F_j=\{i_1^{(j)},\ldots,i_{m_j}^{(j)}\}$,使得 $\bigcup_{j=1}^{7}F_j=\{1,\ldots,n\}$ 且 $F_j\cap\bigcap_{1\leqslant j,k\leqslant 7\wedge j\neq k}F_k=\varphi$. 这样,$F_j$ 代表 P 的一部分二进制位(不一定为子串). 可行解的如此划分形成 7 个实验因素且 F_j 表示因素 j. 我们以下述方式将 2 个可行解的结合看作因素实验:2 个可行解 $P_1=(a_1,\ldots,a_n)$和 $P_2=(b_1,\ldots,b_n)$ 做相同因素划分(如前述). 因素 F_j 被赋予 2 个水平:当 F_j 的元素索引 P_1 时,F_j 取 0-水平;当 F_j 的元素索引 P_2 时,F_j 有 1-水平. 这样,P_1 与 P_2 的结合恰为 7 因素 2 水平的实验.

由 $L_8(2^7)$定义的 OCX 定义如下(记 $L_8(2^7)=(h_{ij})_{8\times 7}$):

(1)随机生成 2 可行解的因素划分:$F_1,\cdots,F_7$.

(2)为子女生育预留足够缓冲区.

(3)for (i:=1 to 8)

　for (j:=1 to 7)

　　{如果h_{ij}为 0-水平,则令 F_j 索引 P_1 的位并将这些位复制到子女 i 的相应位;

　　否则

　　如果h_{ij}为 1-水平,则令 F_j 索引 P_2 的位并将这些位复制到子女 i 的相应位;

　　}

(4)由(3)产生的子女为中间子女. 计算它们的适应值.

(5)依据适应值,中间子女相互竞争,2 个最优者作为合法成员取代其双亲.

从正交表中可发现,第 1 水平组合(试验号 1)全为 0-水平,这表明双亲之一被复制,成为中间子女之一. 正因如此,在 OCX 中我们令 P_1 的适应值高于 P_2,使双亲中的优越者自动参加竞争.

2　欺骗问题及其实验结果

由 FEBE 模型构造的算法已用于解决许多困难问题,如组合优化、非凸数值优化、遗传程序设计等. 本文将应用二进制遗传表示的 FEBE 算法解著名的 Goldberg 欺骗问题.

2.1　Goldberg 三阶欺骗问题

Holland 使用模式(超平面或子空间)建立 GA 的理论基础.[1]他提出了模式定理—GA 的基本定理:短的、低阶的且高出平均适应值的模式,在遗传算法的后续代中以指数方式增加它的试验次数. 由模式定理又进一步提出组块假设(Building Block Hypothesis):GA 通过短的、低阶的高性能模式(称为构造块)的拼接来搜寻最优解. 然而,组块假设并未得到证明,对某些问题该假设并不成立,欺骗问题正是这类问题. 在欺骗函数中,低阶模式将搜索引

向局部极值而非全局最优值. 因此,欺骗问题被普遍用作测试例子来评价 GA 性能.

Goldberg 等人设计了一个三阶欺骗问题(如表 3),已被许多人采用. 该问题具有挑战性,因为它的局部"吸收子"是一局部极小且在海明空间中优于任何相邻点. 这也是一"充分"欺骗问题,因为局部极小的吸收范围覆盖大部分空间而全局优化点是单个"尖点".

表 3 三阶欺骗函数

串	适应值	串	适应值
000	28	100	14
001	26	101	0
010	22	110	0
011	0	111	30

当然,优化三阶欺骗函数不足以说明搜索策略的优劣. 复杂的欺骗函数可由多个三阶函数组合而成,文献[11]使用 5 个三阶函数的顺序并置作欺骗函数,公布的实验结果中 30 次算法运行有 9 次落入局部极值. 我们求解 Goldberg 等人构造的较大问题:由 10 个三阶函数组成. 组合后的函数由 30 个二进位组成,问题空间包含 2^{30}(约 10 亿)个点,有 1 024 个局部极值(每个子函数有 2 个极值),全"1"是唯一全局最优点. 函数值定义为 10 个子函数值之和.

上述欺骗问题的难度也与子函数的联接方式有关. 若每个子函数的三位都分开放置,我们称之为弱联接(Weak-linkage);若顺序并置则称强联接(Strong-linkage). 一般认为弱联接函数比强联接函数有更大难度. 图 3 解释 2 种联接方式.

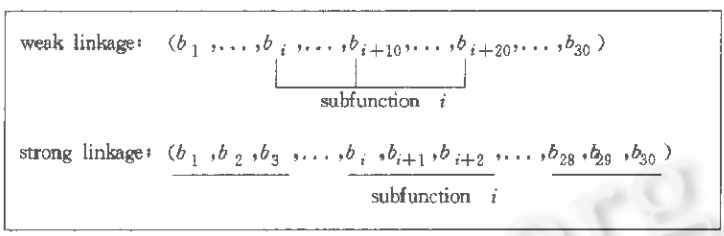

图3 三阶子函数的强联接与弱联接

组合欺骗与联接顺序给简单 GA 带来了难度. Goldberg 等人指出对该问题 CGA 不能收敛于全局优化点:每一子函数收敛于 000 而非 111.[7]

2.2 实验结果

在求解上述欺骗问题的 FEBE 算法中,设置了如下控制参数:相当小的群体 $M=50$;OCX 概率 $Pc=1.0$;变异概率 $Pm=0.03$;算法完成的最大代数 $G=3\ 000$.

FEBE 算法的主要步骤概括如下:

(1)随机生成一初始群体(大小为 50),每一个体都是均匀生成的二进制串(长度为 30).

(2)计算初始群体中每一个体的适应值(注意弱联接欺骗函数同强联接欺骗函数有不同计值方式).

(3)采用"转轮模式"(Roulette Wheel Schema)即按适应值比例机制从当前群体选择 M 个个体到配对池.

(4)配对池中个体两两结合形成 $M/2$ 个家庭.

(5)对每一家庭,双亲以概率 Pc 完成 OCX 操作,生育 2 个优秀的子孙.

(6)个体的位以概率 Pm 接受变异. 重新计算被变异的个体适应值.

(7)若已获得最优解或迭代达最大代数,停止;否则转第(3)步进入下一次迭代.

我们对弱联接与强联接欺骗函数分别运行算法 20 次. 对所有运行,算法均能获得全局最优点. 表 4 归纳了计算代价.

表 4　FEBE 算法计算代价摘要

统计项	强联接	弱联接
平均代数	2 089.8	2 100.3
函数的平均计算次数	419 687	421 401
试验点子空间/全部空间	$4.2*10^{-4}$	$4.2*10^{-4}$

从表 4 中,我们未发现弱联接欺骗比强联接欺骗更难. 图 4 绘出在典型运行中群体中最佳个体相对代数的行为变化曲线. 图 5 绘出子函数个数与获得最优解的平均代数间的关系.

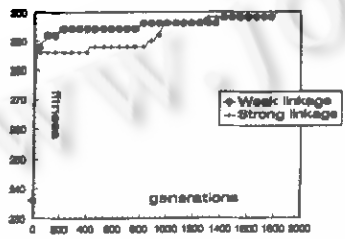

图 4　群体最佳个体相对代数的行为变化

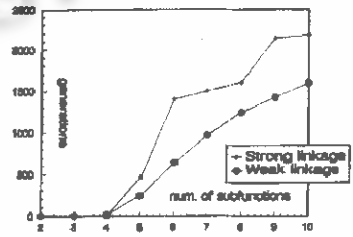

图 5　子函数数目与平均代数的关系

3　结束语

遗传算法中,群体散布性(Diversity)同选择的定向性是相冲突的. 保持群体散布性也就保证 GA 有较大的搜索范围,使群体不易落入局部优化区域. 另一方面,选择压力将群体引入有前途的子空间以推进进化过程. 这一两难问题给 GA 使用者带来困难. 本文提出的 FEBE 模型用家族优生学的观点将 GA 的遗传链同 EP,ES 的进化链结合起来. 该模型强化了个体行为改善,同时维持了群体散布性. FEBE 模型既包含了选择机制引入的群体收缩,同时又兼有行为进化的群体发散. 由于这里的行为进化基于局部区域(家庭),因此各种传统优化策略(包括正交设计)容易与该模型相结合.

Goldberg 的欺骗问题不是实际的工程优化问题,但它是优化算法的合适考题. 一个进化算法的性能需要由各类适应值函数加入评测,而欺骗问题便是其中最困难的一类. FEBE 模型对 Goldberg 欺骗问题的实验结果在一定程度上显示出该模型具有的潜力.

参考文献

1　Holland J H. Adaptation in natural and artificial system. Ann Arbor: University of Michigan Press, 1975.

2　Goldberg D E. Genetic algorithms in search, optimization and machine learning. MA: Addison Welsey, 1989.

3　Fogel D B. An introduction to simulated evolutionary optimization. IEEE Trans. Neural Networks, 1994,5(1):3~14.

4　DeJong K A. Are genetic algorithms function optimizers? In: Manner R, Manderick B eds. Proceedings of the Second Conference Parallel Problem Solving from Nature, Brussels, Netherlands: Elsevier Science Publishers, 1992. 3~13.

5 Whitley D. The genitor algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In: Schaffer J D ed. Proceedings of the Third International Conference on Genetic Algorithms, San Mateo, Los Altos: Morgan Kaufmann Publishers, 1989. 116～121.

6 Michalewicz Z. Genetic algorithms + data structures = evolution programs. 2nd ed., Berlin Heidelberg: Springer-Verlag, 1994.

7 Goldberg D E. Messy genetic algorithms: motivation, analysis, and first results. Complex Systems, 1989,3:493～530.

8 Koza J R. Genetic programming II: automatic discovery of reusable programs. MA: MIT Press, 1994.

9 Montgomery D C. Design and analysis of experiments. 3rd ed., John Wiley & Sons, Inc., 1991.

10 马希文. 正交设计的数学理论. 北京: 人民教育出版社, 1981.

11 Srinivas M, Patnaik L M. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans. on Sys., Man, and Cyber., 1994,24(4):656～667.

A NEW EVOLUTIONARY ALGORITHM BASED ON FAMILY EUGENICS

WU Shaoyan ZHANG Qingfu CHEN Huowang

(*Department of Computer Science National University of Defence Technology Changsha 410073*)

Abstract Several popular approaches of simulated evolution have been developed separately. These approaches emphasize different facets of the natural evolutionary processes, respectively. One has recognized that the simulated evolution will benefit from the adequate combination between the approaches. This paper characterizes the primary difference among existing approaches as the difference between genetic link and behavioral link. A new model of simulated evolution, called FEBE(family eugenics based evolution), is proposed, which combines the genetic link with the behavioral link in light of the idea of family eugenics. In the FEBE model the orthogonal design technique is introduced into offspring's breeding inside a family so as to enhance the behavioral improvement of individuals. The FEBE model is applied to solve Goldberg's deceptive problem that is challenging to most evolutionary algorithms. The exciting experimental results are achieved.

Key words Simulated evolution, genetic algorithms, eugenics, orthogonal design, deceptive problems.