

基因表达式编程初始种群的多样化策略

胡建军^{1),2),3)} 唐常杰¹⁾ 段磊¹⁾ 左劼¹⁾ 彭京^{1),4)} 元昌安^{1),5)}

¹⁾(四川大学计算机学院 成都 610065)

²⁾(广东商学院信息学院 广州 510320)

³⁾(华南理工大学计算机科学与工程学院 广州 510641)

⁴⁾(成都市公安局科技处 成都 610017)

⁵⁾(广西师范学院信息技术系 南宁 530001)

摘要 基因表达式编程(Gene Expression Programming, GEP)算法是遗传家族的新成员,被广泛用于知识发现,其初始种群的质量对进化效率和进化结果至关重要. 为了产生优势初始种群,提出了基因空间均匀分布策略(Gene Space Balance Strategy, GSBS),证明了描述编码空间量化性质的 GEP 编码空间定理. 实验表明,GSBS 提高进化效率超过 20%. GSBS 算法的思想还可以应用于其它进化计算中.

关键词 遗传编程;遗传算法;基因表达式编程;函数挖掘
中图法分类号 TP311

The Strategy for Diversifying Initial Population of Gene Expression Programming

HU Jian-Jun^{1),2),3)} TANG Chang-Jie¹⁾ DUAN Lei¹⁾ ZUO Jie¹⁾ PENG Jing^{1),4)} YUAN Chang-An^{1),5)}

¹⁾(School of Computer Science, Sichuan University, Chengdu 610065)

²⁾(Information Science School, Guangdong University of Business Studies, Guangzhou 510320)

³⁾(School of Computer Science & Engineering, South China University of Technology, Guangzhou 510641)

⁴⁾(Department of Science and Technology, Chengdu Public Security Bureau, Chengdu 610017)

⁵⁾(Department of Information & Technology, Guangxi Teachers Education University, Nanning 530001)

Abstract Gene Expression Programming (GEP) is a new genetic algorithm for knowledge discovery. The diversification of initial population is very important to the evolution efficiency and result. In order to produce excellent initial population of GEP, Gene Space Balance Strategy (GSBS) is proposed. The theorem of describing the space of GEP encoding is proved. The simulation experiments show that GSBS can increase the evolutionary efficiency by 20%. The idea of GSBS algorithms can be used in other evolutionary computation else.

Keywords genetic programming; genetic algorithm; gene expression programming; function mining

1 引言

Ferreira 于 2001 年提出基因表达式编程算法

(Gene Expression Programming, GEP)^[1], 该算法继承了传统遗传算法(GA)编码简单和遗传编程(GP)可解决复杂问题的优点,引起了众多学者的关注,取得了丰硕成果^[2-3],并已应用到很多领域^[4-6].

收稿日期:2005-09-22;修改稿收到日期:2006-06-30. 本课题得到国家自然科学基金(60473071)、四川省青年软件创新基金(816)、国家“九七三”重点基础研究发展规划项目基金(2002CB111504)、高等学校博士学科点专项科研基金 SRFPD(20020610007)、广西自然科学基金(桂科自 0339039)资助. 胡建军,男,1970 年生,博士,研究方向为数据挖掘及商务智能. 唐常杰(通信作者),男,1946 年生,教授,博士生导师,研究领域为数据库与知识工程. E-mail: tangchangjie@cs.scu.edu.cn. 段磊,男,1981 年生,博士研究生,研究方向是数据库与知识工程. 左劼,男,1977 年生,博士,研究方向是数据库与知识工程. 彭京,男,1973 年生,博士,研究方向是数据库与知识工程. 元昌安,男,1966 年生,博士,教授,研究领域是数据库与知识工程.

在遗传算法中,要求初始种群内基因多样化,以保证较高的进化效率^[7-8].目前普遍采用的随机初始化策略,简单易行,占用资源少,但产生的种群多样性有限.在 GEP 算法中,随机初始化策略有时还会产生最高适应度为负的种群,从而导致进化很难开始.

本文提出了基因空间均匀分布(Gene Space Balance Strategy, GSBS)初始种群产生策略.该策略可以提高初始种群基因多样性,并且产生的初始种群的最高适应度一般都为正数.

2 基本概念

为了正确表达 GSBS 算法,本文引入下列形式化描述.

定义 1(GEP 模式). GEP 模式是七元组, $GEP = \langle p, g, h, Fs, Ts, M, F \rangle$. 其中 p 为种群大小, g 为染色体所含基因组个数, h 为头长, Fs 为函数符集合, Ts 为终结符集合, M 为选择范围, F 为连接函数.

定义 2. 设有 GEP 模式 $GEP = \langle p, g, h, Fs, Ts, M, F \rangle$, 用 C_j 表示种群 P 中第 j 个染色体, G_{ji} 表示染色体 C_j 中的第 i 个基因, 其中 $0 \leq j < p, 0 \leq i < (h+t), t$ 为尾长,

(1) G_{ji} 和 G_{ki} 称为等位基因;

(2) 如果基因 $G \in (Fs \cup Ts)$, 对于任一 j , 存在 $G \neq G_{ji}$, 则称 G 是种群 P 在位置 i 上的丢失基因组;

(3) 如果 $C_j = C_k$, 则称 C_j, C_k 是种群 P 上的重复个体;

(4) 如果存在 $G_0 \in Ts$, 表示基因组的第一个基因, 则称这样的基本组为死基因.

GEP 编码的头部可以包含函数符和终结符, 而尾部只能包含终结符. 当基因组头部第一个基因为终结符时, 整个基因组后边的编码成为非编码区^[1], 这样的基因组称为死基因组. 在译码成表达式树时, 后边的基因都不起作用. 这样的染色体适应度一般都很低.

我们采用策略 S 对 GEP 初始种群进行优化, 丰富其所含基因的多样性, 提高个体的适应度, 在本文称这类种群为优势种群. 关于策略 S, 将在第 3 节详细介绍.

定义 3. GEP 编码所表示的最大空间, 称为 GEP 的编码空间; GEP 编码翻译成表现型所对应的空间, 称为 GEP 译码空间.

由于 K 表达式与表达式树一一对应, 所以 GEP

编码空间与译码空间也一一对应.

引理 1. 设 Se 是 GEP 译码空间, Sp 是待解决问题的解空间, 则

(1) 当 $Sp \subseteq Se$ 时, GEP 可以进化出问题的解;

(2) 当 $(Sp \cap Se) \neq \emptyset$ 时, GEP 有可能进化出问题的解;

(3) 当 $(Sp \cap Se) = \emptyset$ 时, GEP 不可能进化出问题的解.

证明. 设 P 是问题的解, 则 $P \in Sp$,

(1) 由于 $Sp \subseteq Se$, 所以 $P \in Se$, 所以 GEP 可以在 Se 中搜索到 P ;

(2) 当 $(Sp \cap Se) \neq \emptyset$ 时, 如果 $P \in (Sp \cap Se)$, 则 GEP 可以在 Se 中搜索到 P ; 如果 $P \notin (Sp \cap Se)$, 则 GEP 无法在 Se 中搜索到 P ;

(3) 当 $(Sp \cap Se) = \emptyset$ 时, $P \notin Se$, 因此 GEP 无法在 Se 中搜索到 P .

对于特定问题, Sp 是固定的, 而 Se 的大小可以通过设置 GEP 的参数调整. 当 Se 的大小与 Sp 越接近, 搜索范围越小, 进化效率越高. 因此在设计 GEP 编码时要尽量使 Se 能最小覆盖 Sp .

由于多数实际问题不能提供充分的先验信息, 所以一般很难把编码空间设计到最佳状态. 通常的做法主要是靠经验. 对于简单的问题, 常用较小的头长和含有较少基本组的染色体; 对于复杂问题, 常用较长的基因组头长和含有较多基因组的染色体. 文献[9]给出了一些解决一般问题的经验值.

定理 1(编码空间定理). 设 GEP 基因组头长为 h , 函数符集合的最大运算目数为 n , 染色体所含基因组个数为 g , 函数符集合大小 $|Fs| = f$, 终结符集合大小 $|Ts| = t$, 基因组间连接函数为加, 则编码空间为 $M! / (g!(M-g)!)$, 其中 $M = (f+t)^h \times t^{h(n-1)+1}$.

证明. 因为头部元素 $Eh \in (Fs \cup Ts)$, 头长为 h , 所以头部有 $(f+t)^h$ 种编码;

尾长 $= h(n-1)+1$, 尾部元素 $Et \in Ts$, 所以尾部有 $t^{h(n-1)+1}$ 种编码;

由此可知每个基因组共有 $(f+t)^h \times t^{h(n-1)+1}$ 种编码;

因为染色体由 g 个基因组构成, 连接函数为加, 所以染色体共有 $M! / (g!(M-g)!)$ 种编码方式, 其中 $M = (f+t)^h \times t^{h(n-1)+1}$.

3 基因空间均匀分布策略

当 GEP 的初始种群具备优势种群的特征, 即种

群中包含了更丰富的基因和基因片断时,则在进化过程中,经过交叉、变异、组合等基因操作,将发生“组合效应”,因而能产生更多的模式,因此可以提高搜索效率,可以更快地跳出局部最优解.并且由于优势种群具有较高的最高适应度,因此可以减少搜索

过程,较快地搜索到全局最优解.

由于优势种群具有上述优良特性,因此需要改进初始种群产生方法,使之产生的初始种群具有优势种群的特性.

表 1 随机产生的初始种群统计

| 基因 | 位置 | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| + | 11 | 7 | 4 | 3 | 8 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| - | 13 | 4 | 9 | 6 | 7 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| * | 8 | 10 | 9 | 8 | 8 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| / | 8 | 12 | 8 | 8 | 16 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q | 2 | 4 | 7 | 7 | 5 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 6 | 10 | 7 | 11 | 5 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 2 | 12 | 5 | 9 | 7 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 2 | 3 | 5 | 8 | 9 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 1 | 10 | 6 | 6 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 7 | 5 | 5 | 7 | 7 | 10 | 17 | 18 | 19 | 20 | 20 | 25 | 24 |
| b | 9 | 5 | 8 | 7 | 6 | 7 | 16 | 23 | 21 | 20 | 17 | 17 | 18 |
| c | 8 | 7 | 5 | 8 | 9 | 4 | 22 | 17 | 26 | 16 | 19 | 18 | 16 |
| d | 9 | 7 | 11 | 8 | 9 | 7 | 19 | 28 | 22 | 18 | 24 | 20 | 20 |
| e | 14 | 4 | 11 | 4 | 3 | 6 | 26 | 14 | 12 | 26 | 20 | 20 | 22 |

表 1 为随机产生的一个初始种群的统计结果.其中 $GEP = \langle g = 1, p = 100, h = 6, Fs = \{+, -, *, /, Q, E, S, T, C^{\text{①}}\}, Ts = \{a, b, c, d, e\}, M = 100 \rangle$.从表 1 可以看出,随机产生的基因在种群中分布不均匀.如函数符 T 在第 4 个基因位中只含有 1 个,而函数符 $/$ 在第 4 个基因位中却包含 16 个.特别是当种群较小时,甚至出现丢失基因的情况.当种群较大时,常会出现重复个体.这些都减弱了种群多样性.为了改变这种情况,本文提出了基因空间均匀分布策略(Gene Space Balance Strategy, GSBS)来产生 GEP 的初始种群.其主要思想是让所有基因尽可能均匀分布在编码空间中,从而使初始种群基因多样化.具体算法如下.

算法 1. 基因空间均匀分布策略 GSBS.

输入: $\langle p, g, h, Fs, Ts, M, F \rangle$

输出: 初始种群

1. 随机产生初始种群;
2. For (检测种群中每一基因位的组成) {
3. If 某基因在该位所占比例大于平均比例
4. Then 把这种基因变异成在该位所占比例最低的基因
5. }
6. While (种群中有重复个体) {
7. 变异重复个体;
8. For (检测种群中每一基因位的组成) {
9. If 某基因在该位所占比例大于平均比例
10. Then 把这种基因变异成在该位所占比例最低的基因

11. }

12. }

4 实验和性能分析

仿真实验是在 Eclipse 环境中用 Java 编程模拟 GEP 实现函数挖掘过程.在实验中模拟了 3 个函数的挖掘:一元二次函数 $F1 = \pi r^2$,一元多次函数 $F2 = 5n^4 + 4n^3 + 3n^2 + 2n + 1$,复杂三角函数 $F3 = \frac{\sin(p_1) \cos(p_2)}{\sqrt{e^{p_3}}} + \tan(p_4 - p_5)$.以上函数来自 <http://www.gene-expression-programming.com/gep/Gep-Book/Chapter4/Section1/SS2.htm>.在实验中,首先产生以上 3 个函数的训练数据集.对每个函数的每个自变量分别随机产生 100 个 $-50.0 \sim 50.0$ 之间的随机数作为训练数据集的参数值,然后分别求出以上 3 个函数在各个参数的值作为训练数据集的目标值.对每个数据集重复 100 次挖掘实验,最后取统计结果的平均值作为最后的实验结果.

在实验中首先比较了含有死基因组的初始种群和不含死基因组的种群的进化性能.

图 1 表明,上述两种种群在进化代数、运行时间、进化成功率方面没有明显区别,但前者的种群平均适

① ‘Q’代表开方运算;‘E’代表 exp 运算;‘S’代表 sin 运算;‘T’代表 tan 运算;‘C’代表 cos 运算.

应度低于后者. 这是由于前者含有适应度很低的个体. 由于这两种初始种群的平均适应度都远低于最高适应度, 所以二者的进化效率差别不大^[13]. 因而在后边的实验中, 产生的初始种群都允许含有死基因组.

图 2 比较了随机方式、基因空间均匀分布策略和选优法 3 种方式产生的初始种群的性能. 图(a)、(b)表明, 均匀策略产生的初始种群在挖掘 3 个函数时, 在进化代数和运行时间方面都表现出了较好的性能. 只是在挖掘 F1 函数时, 均匀策略进化代数劣于选优法. 在挖掘 F3 函数时, 由于 GEP 进化参数设置的原因, 几种方式产生的初始种群的进化性能接近. 图(c)显示了几种种群挖掘不同函数时的成功

率, 均匀策略的成功率明显高于其它两种方式. 特别是在挖掘 F3 时, 均匀策略达到了 100% 的成功率. 图(d)比较了几种方式产生的初始种群的最高适应度. 选优法产生的初始种群最高适应度最高. 而均匀方式和随机方式在这方面的表现差别不大.

图 3 比较了几种方式产生初始种群的时间. 从图(a)可以看出, 选优法所用时间最长. 而均匀策略和随机方式产生初始种群速度较快.

图 2 和图 3 表明, 均匀策略总体性能要优于其它两种方式. 这是由于均匀策略通过增强基因多样性提高了进化效率. 在成功率和进化代数方面, 均匀策略与随机方式相比, 提高进化效率超过 20%.

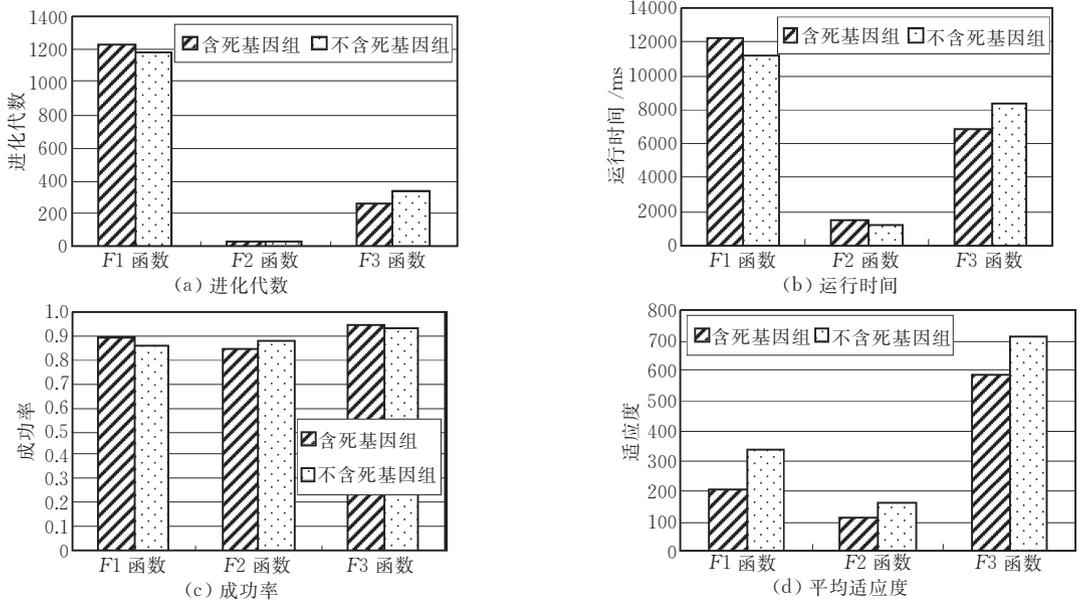


图 1 是否含有死基因组的种群性能比较

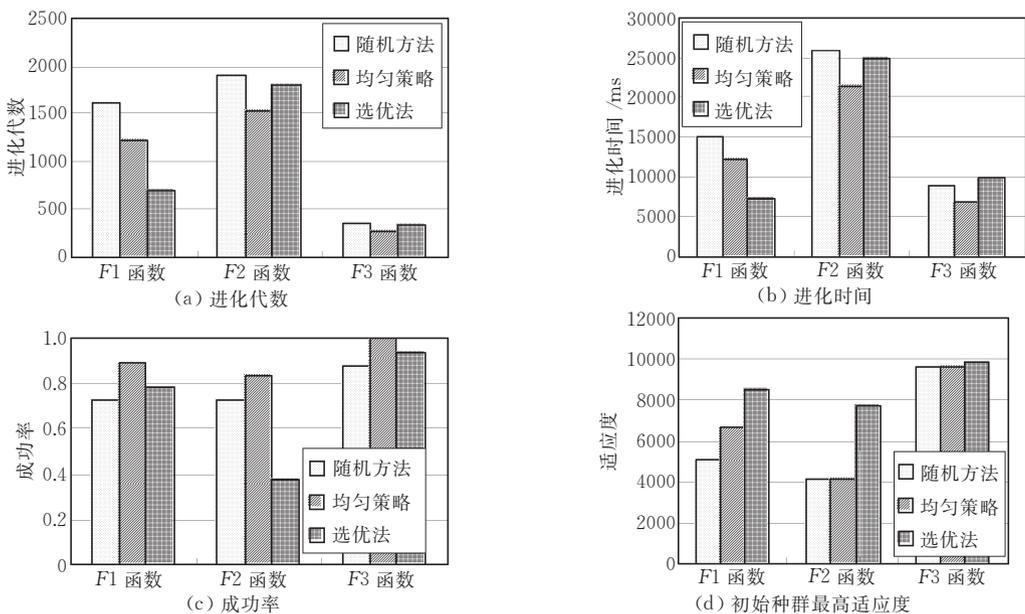


图 2 不同产生方式的初始种群的进化性能比较

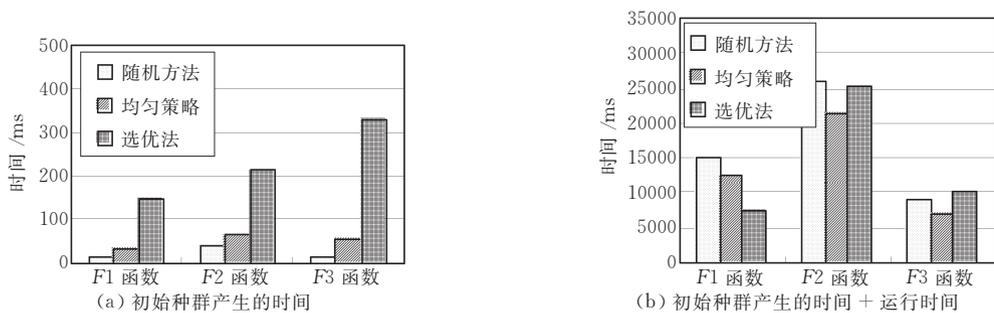


图 3 不同初始种群产生与运行时间的比较

5 结 论

GEP 算法继承了 GA 和 GP 算法的优点, 具有较高的进化速度, 并且不易陷入局部最优点. 和遗传家族的其它算法一样, 初始种群是影响进化效率和进化结果的重要因素. 本文证明了 GEP 的编码空间定理、基因空间均匀分布策略(GSBS). 最后, 通过 GEP 模拟 3 个函数的挖掘过程, 比较了 GSBS 策略产生的初始种群与其它方式产生的初始种群的进化效率. 实验表明, GSBS 提高进化效率超过 20%.

参 考 文 献

- [1] Ferreira C. Gene expression programming: A new adaptive algorithm for solving problems. *Complex Systems*, 2001, 13(2): 87-129
- [2] Yuan Chang-An, Tang Chang-Jie, Zuo Jie, Xie Fang-Jun, Chen An-Long, Hu Jian-Jun. Function mining based on gene expression programming—Convergency analysis and remnant-guided evolution algorithm. *Journal of Sichuan University (Engineering Science Edition)*, 2004, 36(6): 100-105(in Chinese)
(元昌安, 唐常杰, 左劼, 谢方军, 陈安龙, 胡建军. 基于基因表达式编程的函数挖掘——收敛性分析与残差制导进化算法. *四川大学学报(工程科学版)*, 2004, 36(6): 100-105)
- [3] Jia Xiao-Bin, Tang Chang-Jie, Zuo Jie, Chen An-Long, Duan Lei, Wang Rui. Mining frequent function set based on gene expression programming. *Chinese Journal of Computers*, 2005, 28(8): 1247-1254(in Chinese)

(贾晓斌, 唐常杰, 左劼, 陈安龙, 段磊, 汪锐. 基于基因表达式编程的频繁函数集挖掘. *计算机学报*, 2005, 28(8): 1247-1254)

- [4] Zuo Jie, Tang Chang-Jie, Li Chuan, Yuan Chang-An, Chen An-Long. Time series prediction based on gene expression programming//*Proceedings of the International Conference for Web Information Age 2004 (WAIM04)*. LNCS 3129. Berlin: Springer Verlag, 2004: 55-64
- [5] Duan Lei, Tang Chang-Jie, Zuo Jie, Chen Yu, Zhong Yi-Xiao, Yuan Chang-An. An anti-noise method for function mining based on GEP. *Journal of Computer Research and Development*, 2004, 41(10): 1684-1689(in Chinese)
(段磊, 唐常杰, 左劼, 陈宇, 钟义啸, 元昌安. 基于基因表达式编程的抗噪声数据的函数挖掘方法. *计算机研究与发展*, 2004, 41(10): 1684-1689)
- [6] Huang Xiao-Dong, Tang Chang-Jie, Li Zhi, Pu Dong-Hang, Liao Yong. Mining functions relationship based on gene expression programming. *Journal of Software*, 2004, 15(Supplement): 96-105(in Chinese)
(黄晓冬, 唐常杰, 李智, 普东航, 曾令明, 廖勇. 基于基因表达式编程挖掘函数关系. *软件学报*, 2004, 15(增刊): 96-105)
- [7] Mayr E. Change of genetic environment and evolution//Huxley J, Hardy A C, Ford E B. *Evolution As a Process*. London: Allen and Unwin, 1954: 157-180
- [8] Mayr E. *Animal Species and Evolution*, Cambridge, Massachusetts: Harvard University Press, 1963
- [9] Ferreira C. Mutation, Transposition, and Recombination: An Analysis of the Evolutionary Dynamics//*Proceedings of the 6th Joint Conference on Information Sciences, 4th International Workshop on Frontiers in Evolutionary Algorithms*. Research Triangle Park, North Carolina, USA, 2002: 614-617



HU Jian-Jun, born in 1970, Ph. D. . His research interests include data mining and business intelligence.

TANG Chang-Jie, born in 1946, professor, Ph. D. supervisor. His research interests are in the area of database and knowledge engineering.

DUAN Lei, born in 1981, Ph. D. candidate. His research interests are database and knowledge engineering.

ZUO Jie, born in 1977, Ph. D. . His research interests are database and knowledge engineering.

PENG Jing, born in 1973, Ph. D. . His research interests are database and knowledge engineering.

YUAN Chang-An, born in 1966, Ph.D., professor. His research interests are database and knowledge engineering.

Background

This paper is supported by the National Natural Science Foundation of China under grant No. 60473071.

Gene Expression Programming (GEP) is an effective evolutionary algorithm and is widely used in function mining. The initial population of GEP is an important factor to the evolutionary efficiency. The random way is usually used to produce the initial population. By this way the gene diversity in it is very poor. In order to produce excellent initial popula-

tion of GEP, Gene Space Balance Strategy (GSBS) is proposed. GSBS increases the diversity in the initial population. The simulation experiments show that GSBS can increase the evolutionary efficiency by 20%. The idea of GSBS algorithm can be used in other evolutionary computation else.

This paper is focused on producing excellent initial population of GEP.