第 17 卷 第 6 期　　控 制 与 决 策　　2002 年 11 月
Vol. 17 No. 6　　Control and Decision　　Nov. 2002

文章编号: 1001-0920(2002)06-940-04

遗传规划中的基因内区研究

云庆夏, 卢少华
(西安建筑科技大学 管理学院, 陕西 西安 710055)

摘　要: 利用试验证实遗传规划的个体中存在基因内区。它们是冗余的表达式, 附加在算法树上, 使个体变得臃肿, 但对实际的输出结果没有影响。它对进化的收敛既有积极作用, 也有消极作用。通过对遗传操作方法和参数的研究, 揭示出基因内区发生的规律, 并提出扬长避短的途径。

关键词: 遗传规划; 基因内区; 进化; 试验
中图分类号: TP 18　　**文献标识码**: A

Study on introns of genetic programming

YUN Qing-xia, LU Shao-hua
(College of Management, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: After a set of experiments, it is proved that there exists intron in the individuals of genetic programming. Intron is a redundant expression for individual and attached to its algorithm tree, which makes the individual swelling but has no effect to the final output. Intron gives impact on evolution both positively and negatively. Based on the study of genetic operations and their parameters, emerging patterns of intron are discovered and some methods are proposed to develop the positive role and discard the negative one.

Key words: genetic programming; intron; evolution; experiment

1 引　　言

遗传规划是在遗传算法基础上发展起来的一种新型搜索寻优技术。它和遗传算法一样, 是仿效生物界的进化和遗传, 遵循达尔文"优胜劣汰"的原则, 从一组随机生成的初始可行解出发, 经过复制、交换、突变等遗传操作, 逐渐逼近问题的最优解。遗传规划与遗传算法的差别主要在问题的表达方式上。前者用动态可变的程序式语言表达所研究的问题, 而后者用定长的字符串表达问题, 不如前者灵活[1]。

在遗传规划的进化过程中, 某些个体表达式是多余的。如

$$X + 0, X * 1, \mathrm{NOT}(\mathrm{NOT}(X)), \cdots$$

在进化过程中, 这些多余的表达式附加在算法树上, 使个体变得臃肿, 但对实际的输出结果并没有影响。遗传规划中称这种多余的表达式为基因内区。这一术语来自生物学, 它把染色体内对生物性态没有直接效用的基因也称作基因内区[2]。

在遗传规划中, 基因内区既有积极作用, 也有消极作用。其中积极作用表现为:

1) 保护作用。基因内区的存在, 使个体变得臃肿, 当交换和突变在此多余的基因内区中发生时, 不

收稿日期: 2001-08-07; **修回日期**: 2001-10-29
基金项目: 国家自然科学基金项目(59874019)
作者简介: 云庆夏(1937—), 男, 广东广州人, 教授, 博士生导师, 从事人工智能及优化技术的研究; 卢少华(1971—), 男, 湖北黄石人, 博士生, 从事人工智能的研究。

会改变个体或构造块(优良基因组) 的适应度, 从而缓解交换和突变的破坏作用。

2) 节俭作用。由于基因内区的存在, 个体的节点可分为有效节点和无效节点两部分, 后者就是基因内区。由于它的存在, 使个体的有效部分更加简洁精悍, 从而避开交换和突变的破坏作用。

基因内区的消极作用表现在:

1) 进化停滞。若基因内区大量充斥在群体中, 交换(或突变) 往往发生在基因内区, 不能改善个体的素质, 从而使遗传规划的进化过程陷于停滞。

2) 计算冗余。基因内区的存在, 也浪费计算机资源, 使计算机的存贮空间和计算时间白白消耗在对基因内区的无效处理上。

总之, 遗传规划中的基因内区有利也有弊, 应该设法扬长避短。本文正是基于此目的对基因内区展开研究。

2　基因内区的存在

为演示基因内区的存在, 我们采用符号回归的方法, 对下列函数进行曲线回归, 其原始数据按 $\Delta x = 1$ 于 $x = -10 \sim 10$ 区间内选取。

函数 1

$$f(x) = \frac{1}{x} + \sin 2x + \frac{x-1}{3} + 1 \qquad (1)$$

函数 2

$$f(x) = 0.5 + \frac{\sin\sqrt{x^2+1} - 0.5}{(1 + 0.001(x^2+1)^2)^2} \qquad (2)$$

函数 3

$$f(x) = x^4 + x^3 + x^2 + x + 1 \qquad (3)$$

试验中遗传规划的计算参数如下: 群体规模为 500～1000; 迭代代数为100～500; 复制概率为 0.1～0.3; 交换概率为0.1～0.3; 突变概率为0.05。

初始群体产生方法: 生长法, 完全法, 混合法; 复制方法: 精英选择复制, K-竞技选择复制, 轮盘选择复制; 突变方法: 点突变, 增删子分支; 函数符集: 四则运算, 三角函数, 对数和指数函数, 开方运算; 迭代终止标准: 规定最大迭代代数, 或精度 $\delta \leq 0.01$; 适应度: 拟合误差, 越小越好。

试验中, 为尽量减少遗传规划在进化过程中随机因素的影响, 每组试验都分别运行 10 次以上, 然后综合评价其结果。图 1 是一组有代表性的试验。从图中可以看出, 随着遗传规划迭代次数的增加, 每代群体中最优个体的节点数(曲线 ②) 以及单个个体的平均节点数(曲线 ③) 都不断增加, 而最优适应度(误差) 不断减小(曲线 ①, 放大 1 000 倍)。特别是在迭代后期, 每代群体适应度的改善极为缓慢, 但节点数却迅速增长, 其增长速度远大于适应度的改善速度。而节点数的增加, 正说明基因内区膨胀, 使进化过程湮没在基因内区中, 从而延滞了群体素质的改善。此外, 臃肿多余的节点也浪费计算机的存贮空间和 CPU 时间。这些都是基因内区的消极作用。

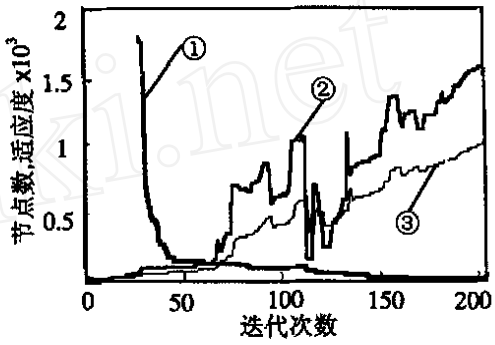

图 1　基因内区的存在

另一方面, 图 1 也反映出基因内区的积极作用。图中表明, 最优个体的节点数一般都大于群体的平均节点数。这说明最优个体中冗余的基因内区可防止遗传操作对优良构造块的破坏, 起到保护优良个体的作用。

类似的其他试验都反映出遗传进化先快后慢及节点数不断增加的现象, 说明基因内区对遗传规划既有消极作用又有积极作用。

3　复制对基因内区的影响

遗传规划中, 复制是将优良个体拷贝入下一代, 体现了优胜劣汰的原则。不同的复制方法对基因内区有不同的影响。图 2 记录 3 种复制方法下个体平均节点数的差别。图中曲线 ① 是精英复制法, 它只复制群体中的绝对优良者, 因此, 含有基因内区的臃肿的优良个体每次都被复制到下一代, 导致每代个体的平均节点数最大; 曲线 ③ 是轮盘复制法, 它除了复制优良个体外, 一些欠佳或劣质个体也有可能入选到下一代, 正是由于后者的存在, 加上被复制的优良个体数目少于精英复制法, 因此每代个体的平均节点数变化不大, 基因内区的膨胀不明显; 曲线 ② 是 K-竞技选择复制, 它在任意选择的 k 个个体中选出最优者进入下一代, 由于 k 个个体选择的任意性, 导致被复制的个体素质时好时坏, 反映在个体平均节点数上, 该法正好介于前两种复制方法之间。总之, 复制过程中个体的选择力度对基因内区的成长

有很大影响: 精英复制法选择力度最大, 基因内区膨胀最快; 轮盘复制法选择力度最小, 基因内区膨胀不明显; 竞技选择复制法则介于两者之间。

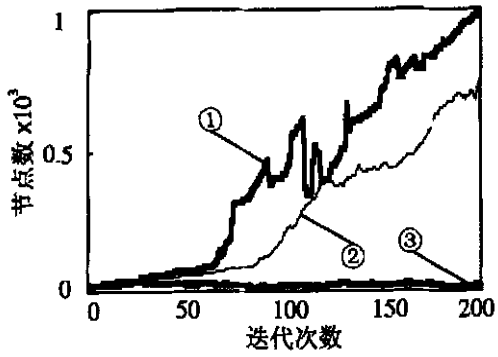

图 2 复制方法对基因内区的影响

复制对基因内区的影响还表现在复制概率的选取上。图 3 是采用轮盘复制法时 10 次运行的平均结果。从图中可见, 随着复制概率的增加, 臃肿的优良个体被大量移入新群体, 因此计算耗时(曲线 ①)明显增加, 充分体现基因内区的消极作用。此外, 从适应度(曲线 ④, 放大 100 倍) 来看, 随着复制概率的增加, 基因内区不断膨胀, 导致进化停滞。当复制概率取 0. 5 时, 适应度最差, 进化终止也最早(曲线 ②为迭代收敛代数), 这时基因内区的消极作用达到极点。当复制概率再增加时, 基因内区膨胀更快, 一方面基因内区对优良个体构造块的保护作用更加明显, 另一方面, 交换和突变对其他基因部分的改良几率也有所增加, 使进化又重新加速, 体现出基因内区的积极作用。由图可见, 此时复制概率宜取 0. 1～0. 3。曲线 ③ 为个体平均节点数, 变化不大。

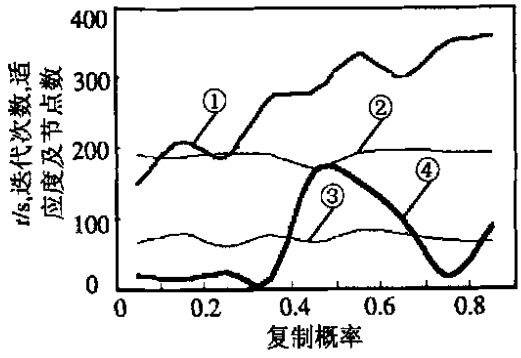

图 3 复制概率对基因内区的影响

4 交换对基因内区的影响

遗传规划中, 交换是产生新个体的主要手段, 也是滋长基因内区的主要原因。为了研究交换对基因内区的影响, 我们对 9 个交换概率(从 0. 1～ 0. 9) 各运行 10 次, 取平均结果示于图 4。图中表明, 当交换概率较小或较大时, 遗传规划皆以较快的速度收敛(曲线 ② 为迭代收敛代数), 最优个体的节点数(曲线 ④) 也较小; 当交换概率处于 0. 5 左右时, 计算的各项指标都要恶化, 尤其是计算时间(曲线 ①)及达到收敛所需迭代代数(曲线 ②) 达到峰值。

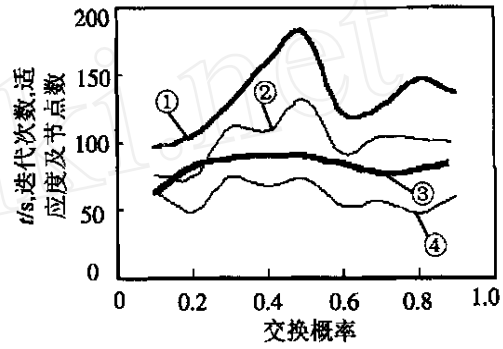

图 4 交换概率对基因内区的影响

图 4 的结果也体现了基因内区的作用。当交换概率小时, 基因内区膨胀缓慢, 因此计算时间较短, 进化速度较快(收敛所需代数较小)。随着交换概率的增大, 基因内区迅速增长, 其消极作用明显体现, 使计算时间加长, 进化速度变慢, 最优个体节点数也增加。当交换概率约为 0. 5 时基因内区的消极作用达到极点。

当交换概率继续增加时, 基因内区以更快的速度增长, 充分地保护优良个体或个体中的优良基因(构造块), 改善了遗传进化的速度和效率, 充分地体现了基因内区的积极作用。

5 限制节点数的效果

为防止基因内区无限度地扩散, 可在交换(或突变) 中规定个体的最大节点数目。 一旦新个体的节点数超过规定值, 此次交换(或突变) 作废, 重新执行原操作。图 5 是限制最大节点数为 100 与没有限制时的试验结果的对比。

从图 5 可以看出, 限制节点数后, 个体的平均节点数(曲线 ①) 增长缓慢, 并逐渐趋于平稳, 说明基因内区的膨胀得到抑制。另一方面, 从适应度来看, 限制节点数后适应度的变化情况(曲线 ②) 与无限制时(曲线 ③) 差别不大。尽管进化缓慢一些, 但计算时间可明显节约。因此, 在遗传规划中常常要限制个体的最大节点数。

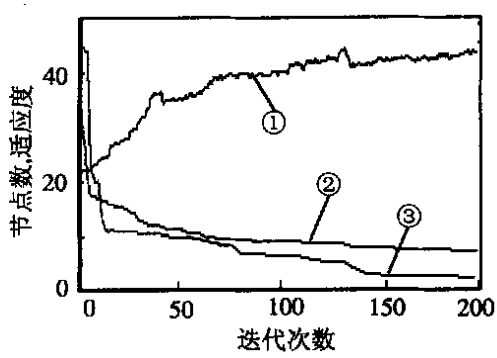

图 5 限制节点数的效果

6 罚函数的效果

为降低基因内区的消极作用, 可对节点数较多的个体实行惩罚, 使个体的适应度与节点数联系在一起[3]。试验中, 采用以下的适应度

$$f = f_0 + N/C \tag{4}$$

式中, f_0为原来遗传规划的适应度, 本文为拟合误差, 越小越好; N 为个体的节点数; C 为惩罚因子, 本文取 200。

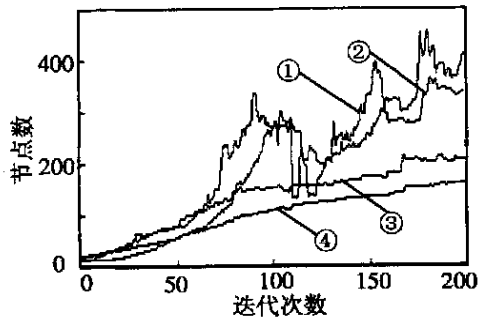

图 6 罚函数对节点数的影响

图 6 表示惩罚前后节点数的变化。从图中可见, 施加惩罚后个体的平均节点数(曲线 ④) 及最大节点数(曲线 ③) 都较惩罚前(曲线 ② 和 ①) 明显降低, 从而节省了计算机资源。

另一方面, 从适应度衡量, 施加惩罚前后适应度的变化情况差别不大, 说明罚函数对进化收敛没有明显的不良影响。因此, 施加惩罚是抑制基因内区消极作用的一个有效手段。

顺便指出, 惩罚后最优个体节点数大于个体的平均节点数, 其情况类似于图 1 的曲线 ② 高于曲线 ③, 说明施加惩罚后基因内区仍发挥积极作用。

7 结 论

遗传规划在迭代过程中会出现冗余的基因内区, 对进化收敛有积极作用也有消极作用。它们可以从个体的节点数、适应度的变化和迭代收敛时间上得到证实。遗传操作对基因内区的形成和扩散有重要影响, 有必要正确选择复制方法、复制概率及交换概率, 并限制个体最大节点数。在适应度中添加惩罚函数可以限制个体内节点数的增加, 从而抑制基因内区的发展。

参考文献(References):

[1] 云庆夏, 黄光球, 王战权. 遗传算法和遗传规划[M]. 北京: 冶金工业出版社, 1997.

[2] 云庆夏. 进化算法[M]. 北京: 冶金工业出版社, 2000.

[3] Banzhaf W, Nordin P, Keller R E, et al. *Genetic Programming—An Introduction on the Automatic Evolution of Computer Programs and Its Application*[M]. USA: Morgan Kaufmann Publishers, 1998.

第六届国际粉体检测与控制学术会议(MCGM 2003)征文通知

国际粉体检测与控制联合会(IFMCGM)联合中国颗粒学会、中国仪器仪表学会、陶瓷学会、冶金自动化学会, 定于 2003 年 8 月 20 日至 22 日在中国上海召开第六届国际粉体检测与控制学术会议(MCGM 2003)。会议由东北大学、上海宝山钢铁股份有限公司主办。

征文范围: 颗粒、粉体、块(状)、浆(液)等材料参数的测量: 成分、尺寸、形状、重量、密度、粘性、湿度、温度、压力、流量、含水量、膨胀度等; 采矿和矿物工业中原料开采、粉碎、运输、筛分、浮选、精选等工艺的过程控制; 冶金工业中原料制备、火冶法、烧结等工艺的过程控制; 水泥及陶瓷工业中原料制备、焙烧、填料、烘干等工艺的过程控制; 煤炭制备中研磨、输送、流化床等工艺的过程检测及控制; 石油、化工、医药、造纸、食品、纺织、环保等工业中有关粉体材料的过程控制; 工业过程成像技术及其应用; 其它有关粉体、颗粒、浆(液)材料的检测与控制的理论研究或实践。

征文要求: 英文摘要一式三份(400~600 个单词)

截稿日期: 2003 年 1 月 15 日

联系地址: 110004 沈阳市东北大学 321 信箱(国际粉体检测与控制联合会秘书处)

联 系 人: 李新光

电 话: 024-23891977, 024-83685464

传 真: 024-23891977

E-mail: mcgm@mail.neu.edu.cn

Lxguang@mail.sy.ln.cn