第 27 卷第 11 期
2010 年 11 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 27 No. 11
Nov. 2010

文章编号: 1000−8152(2010)11−1585−05

多细胞基因表达式编程的函数优化算法

彭昱忠[1], 元昌安[1], 陈建伟[1], 吴信东[2], 王汝凉[1]

(1. 广西师范学院 科学计算与智能信息处理广西高校重点实验室, 广西 南宁 530001;

2. 美国佛蒙特大学 计算机科学系, 佛蒙特州 05405)

摘要: 针对处理复杂的函数优化问题时传统演化算法易出现收敛性能不佳、搜索冗长和精度不高等问题, 提出了一种基于多细胞基因表达式编程的函数优化新算法. 该算法引入了同源基因和细胞系统思想, 设计了相应新的个体编码方案、种群生成和遗传操作策略. 通过对8个Benchmarks函数的对比实验, 验证了该算法具有很强的全局寻优能力、较佳的收敛性能和更高的解精度.

关键词: 函数优化; 演化算法; 基因表达式编程; 同源基因; 细胞系统

中图分类号: TP311　　**文献标识码:** A

Multicellular gene expression programming algorithm for function optimization

PENG Yu-zhong[1], YUAN Chang-an[1], CHEN Jian-wei[1], WU Xin-dong[2], WANG Ru-liang[1]

(1. Key Lab of Scientific Computing &Intelligent Information Processing in Universities of Guangxi, Guangxi Teachers Education University, Nanning Guangxi 530001, China;

2. Department of Computer Science, University of Vermont, Vermont 05405, USA)

Abstract: In dealing with complex function optimization problems, many existing evolutionary algorithms have performance limitations such as inability of convergence, poor searching efficiency and low precision. To cope with this problem, we adopt the idea of homeotic genes and cellular system, and propose a new algorithm based on multicell genetic expression programming. In addition, a new relevant individual coding method and new schemes of population generating and genetic operation are designed. Compared with other algorithms on eight Benchmark functions testing, the proposed algorithm shows higher precision, improved convergence ability and global search ability.

Key words: function optimization; evolutionary algorithm; gene expression programming; homeotic genes; cellular system.

1 引言(Introduction)

函数优化是一类具有广泛的工程应用背景而又非常难求解的问题. 其求解方法越来越受到人们的重视. 模仿自然界物质进化过程的演化算法在解这类优化难题中显示出了优于传统优化算法的性能. 但对复杂问题, 它们也常存在或易趋于早熟收敛而陷于局部最优解, 或存在收敛速度慢, 或计算量大, 或解精度不高等问题[1~3].

基因表达式编程(gene expression programming, GEP)是演化计算家族中的新成员, 融合了遗传算法(GA)和遗传编程(GP)的优点, 有很强的解决实际问题的能力. 其解决同样问题比传统演化算法快2∼4个数量级, 具有广阔的应用前景和深厚的应用潜力[4,5]. 2001年12月, GEP首次被正式发表[4]以来受到了学术界的高度关注, 目前已成为演化计算领域的研究热点[5~13]. 但GEP在函数优化领域研究极少. 文献[10]首次将GEP应用于求解函数优化问题, 并设计了Hzero算法和GEP-PO算法, 效果都很好, 尤其是GEP-PO算法较其他演化算法更出色. 文献[11]利用类似GEP-PO算法原理设计了多目标优化算法GEPMO, 并取得了不错的效果. 但是这些现有的基于GEP的函数优化算法依然存在不少缺陷, 如模式结构容易被破坏, 解的搜索空间相对有限等.

本文提出了一种基于多细胞基因表达式编程函

收稿日期: 2009−03−26; 收修改稿日期: 2010−03−15.

基金项目: 国家自然科学基金资助项目(60763012); 广西自然科学基金资助项目(桂科自0731028); 广西高等学校优秀人才资助计划资助项目(RC2007022).

数优化算法(GEP-MCFO). 该算法能够很好地保持优秀的结构延续到后代, 并可通过拓展搜索空间增加解的多样性. 8个典型算例的对比仿真实验验证了该算法具有很强的全局寻优能力、较佳的收敛性能和更高的解精度, 适合求解复杂函数优化问题.

2 GEP-MCFO算法(GEP-MCFO algorithm)

2.1 GEP-MCFO 基本思想(The mind of GEP-MCFO)

定义 1 细胞(cells)是一个4元组, 记为$C = (G, g, H, S)$, 其中: G为普通基因的集合; $g = |G| \geqslant 2$为细胞包含的普通基因数目; H为同源基因的集合; S为细胞对特定数据集的适应值.

定义 2 细胞系统(cellular system)是一个3元组, 记为$MC = (C, CS, n)$. 其中: C是上述定义中细胞的集合, 每个细胞对应函数的一个决策变量; $n = |C|$为细胞系统包含的细胞数目; CS为常量集合. 在GEP-MCFO中, 一个细胞系统相当于种群中的一个个体.

由定义可知细胞是通过每个同源基因的头部的函数符灵活地将多个普通基因动态地连接组合而成的. 因此, 可以把细胞中的各普通基因看作为染色体的基因片段. 显然这种通过细胞表达个体的方式有利于染色体中优秀的结构得到保护和重用. 此外, 这种表达还可通过拓展搜索空间增加解的多样性, 从而提高搜索质量和效率. 本文在GEP的基础上引入了同源基因和细胞系统思想来求解函数优化问题. 每个细胞对应函数的一个决策变量, 整个细胞系统表示解空间中的一个可行解, 从而将函数优化问题转换为优化细胞系统(即GEP-MCFO的个体)问题.

2.2 个体编码(Individual coding designing)

2.2.1 终端字符(Terminal)

在求解某些问题时, 常需产生结果为整数的模式, 以利于问题的求解. GEP-PO算法的染色体编码方法通常需通过ET树的若干个节点和大量遗传操作随机组合才能产生整数, 耗费不少的系统资源. GEP-MCFO允许将常整数作为终端字符集的元素. 这样, 仅用ET树中的一个节点和单一的变异操作即可得到该模式, 既节省了基因位, 又减少了遗传操作和算术运算, 可提高算法的效率.

2.2.2 常量集合(Constants set)

GEP-PO算法的个体编码方案中, 每个基因单独用一个对应的分常量集合. 在GEP-MCFO中, 每个细胞系统中的各基因自由共享一个全局的常量集合. 其中, 该常量集合的维数与GEP-PO个体的各分常量集合的维数总和相等. 这样, 可增大系统表达空间、增加个体编码的多样性. 即

定理 1 相同的条件下, 各基因共用一个常量集合比各基因独自用一个对应分常量集合所得的系统表达空间大.

篇幅所限, 略去证明过程.

2.2.3 细胞系统编码(Cells system encoding)

假设需求解二元函数$f(x, y)$的最优值, 设普通基因和同源基因头长各取3, 普通基数为2, F为$\{+, -, *, /\}$, T为$\{?, 0, 1, \cdots, 9\}$, $F_{\rm H} = F$, 则可构造细胞系统$MC = (G, 2, H, S), CS, 2$. 两个细胞C_1和C_2分别表示两个变元, 其系统(个体)编码(基因型)如图1所示.

01234567890 01234567890 **0123456 0123456**
++?b??AFDB ++???fCEAD ***+/0101 -/*1001**
CS={0.1 , 3.2 , 4.8 , 0.4 , 2.0 , 1.4 , 1.3 , 2.7}

图 1 GEP-MCFO细胞系统编码结构

Fig. 1 Structure of GEP-MCFO cellular system

其中: 为了区分同源基因用小写字母代表对应整数, 大写字母表示常量的位置, 黑体部分为同源基因. 解码得$C_1 = 2.247$, $C_2 = -18.318$, 则目标函数为$f(2.247, -18.318)$的值.

2.3 系统的搜索空间(The search space of system)

定义 3 基因系统(genes system)是一个3元组, 记为$MG = (G, CS, n)$. 其中: G是普通基因的集合, 每个普通基因对应函数的一个决策变量; $n = |{\rm G}|$为基因系统的普通基因数目; CS为常量集合. 在GEP-PO中, 一个基因系统相当于一个染色体.

定义 4 系统表达空间是系统编码所能表示的最大空间. 它直接影响算法的搜索空间大小和搜索能力.

多样性是是进化计算能够搜索到全局最优解的基本条件. 而重组过程的成熟化效应会使经过若干代进化后种群的多样度逐渐趋于零, 从而导致过早收敛. 拓宽算法的搜索空间, 增加群体的多样性, 可保证遗传算子正常发挥进化和重组效应, 大大减小算法由于过早收敛而陷入局部最优解的可能性. 本文发现, 同源基因除了具有利于重用和保存染色体优秀结构的作用外, 还可有效地拓宽系统的搜索空间. GEP-MCFO利用这一性质, 在染色体较短的编码长度下, 使系统获得较大的搜索空间, 增强算法的全局搜索能力, 避免陷入局部最优. 下面对此性质予以证明.

引理 1 设基因系统含有m个普通基因, 且每个基因位的取值范围相同, 普通基因头部长为h, 函

数符集合为F, F的最大操目数为n, $f=|F|$, 终端字符集为T, $t=|T|$, 常量集的维数为u, 常量的取值区间长度为r, 常量的精度为p, 则系统表达空间

$$D_{\mathrm{g}}=s^m u^{r/p},$$

其中

$$s=(f+t)^h t^{h(n-1)+1}U^{h(n-1)+1}.$$

证　略.

引理 2　设细胞系统含有i个同源基因和m个普通基因, 且每个普通基因的基因位取值范围相同, 普通基因头部长为h, 普通基因函数符集合为F, F的最大操目数为n, $f=|F|$, 普通基因终端字符集为T, $t=|T|$, 同源基因函数符集合为F_{H}, F_{H}的最大操目数为n_{h}, $f_{\mathrm{h}}=|F_{\mathrm{H}}|$, 同源基因终端字符集为$T_{\mathrm{H}}$, $t_{\mathrm{h}}=|T_{\mathrm{H}}|$, 常量集合的维数为$u$, 常量的取值区间长度为$r$, 常量的精度为$p$, 则系统表达空间

$$D_{\mathrm{c}}=s^m k^i u^{r/p},$$

其中:

$$s=(f+t)^h t^{h(n-1)+1}U^{h(n-1)+1},$$
$$k=(f_{\mathrm{h}}+s^{t_{\mathrm{h}}})^h (s^{t_{\mathrm{h}}})^{h(n_{\mathrm{h}}-1)+1}.$$

证　略.

定理 2　对于拥有相同长度的多细胞系统染色体和多基因系统染色体, 若每个普通基因的基因位取值范围相同, 普通基因及同源基因的头部长相同, 普通基因和同源基因的函数符集合相同, 普通基因终端字符集相同, DC域长度相同, 常量集合的维数相同, 常量的取值区间大小相同, 常量的精度相同, 则多细胞系统染色体的个体表达空间D_{c}大于多基因系统染色体的个体表达空间D_{g}.

证　设多细胞系统的染色体MC和多基因系统的染色体MG, 若MC由$m-i(m-i\geqslant 2)$个普通基因和i个同源基因组成, MG由m个普通基因组成, 且它们的普通基因及同源基因的头部长都为h, 普通基因和同源基因的函数符集合都为F, $f=|F|$, 最大操目数为n, 普通基因终端字符集都为T, $t=|T|$, 常量集合的维数都为u, 常量的取值区间长度都为r, 常量的精度都为p. 因普通基因及同源基因的头部和尾部都遵循$t=h(n-1)+1$关系, 显然染色体MC和MG的长度是相等的. 又根据前面引理1和引理2可得MG的表达空间大小

$$v=s^m u^{r/p}=s^{m-i}s^i u^{r/p},$$

MC的表达空间大小

$$w=s^{m-i}k^i u^{r/p},$$

其中:

$$s=(f+t)^h t^{h(n-1)+1}U^{h(n-1)+1},$$
$$k=(f+s^t)^h (s^t)^{h(n-1)+1}.$$

根据GEP的原理和它染色体构造规则必有$f\geqslant 1$, $t\geqslant 1, h\geqslant 1, U\geqslant 1, n\geqslant 1$. 所以有

$$k=(f+s^t)^h (s^t)^{h(n-1)+1}>(s^t)^h (s^t)^{h(n-1)+1},$$

则

$$k/s>(s^t)^h (s^t)^{h(n-1)+1}/s,$$

即

$$k/s>s^{th-1}(s^t)^{h(n-1)+1}.$$

所以$D_{\mathrm{c}}>D_{\mathrm{g}}$成立, 且是指数级倍递增.　　证毕.

2.4　遗传操作和种群策略(Strategies of genetic operations and population)

遗传算子操作在遗传演化算法中是极为重要的. 本文根据算法个体编码结构特点, 新增了DC变异DM和常数变异CM算子分别对常量域DC和候选常量进行遗传操作, 并根据各遗传算子的特性将其归结为3类: 选择操作、交叉类(包括单点重组和两点重组)、泛变异类(包括变异和根插串)和数字变异类(包括DC变异异和常数变异). 在GEP中, 最有效的遗传算子是变异, 而交叉类的算子比泛变异类对求解问题效果贡献都差[10]. 标准GEP的遗传操作存在前面的遗传操作得到好的个体模式结构容易被后面的遗传操作破坏的不足. 为了克服此不足, GEP-MCFO对算法遗传操作进行了新的组合设计. 主要思想为: 先将遗传算子按类分组, 然后按随机概率满足与否对被选染色体依次进行各组遗传操作, 且对于每代中同一个体, 交叉操作仅在单点重组和两点重组之间选一进行(较优先两点重组), 变异类的操作仅在泛变异类和数字变异类所有算子中选一进行(较优先变异); 各项遗传算子操作采用了循环方式完成, 首先确定当前代的个体, 然后根据选择执行遗传操作, 最后根据种群策略产生下一代并返回.

文献[12]证明了采用最优保留的机制, 可以保证算法具有全局收敛性. GEP-MCFO用轮盘赌选择法选择父代一定比例的个体形成临时的群体, 并对临时群体中的个体进行遗传操作, 然后统计临时群体的各个体的适应值, 最后将父代种群和临时种群中的个体按适应度排序, 并取当前种群规模大小个最优个体作为子代的父体.

2.5　算法主框架(The main algorithm)

根据前4小节, GEP-MCFO算法步骤概括为:

Step 1　加载算法初始配置;

Step 2 初始化种群并生成第一代个体;

Step 3 评价第1代群体适应值. 如果未满足停机条件则转Step 4, 否则转Step 8;

Step 4 从当前代群体中用轮盘赌选择构建一定规模的临时群体, 按照2.4节的方法对临时种群进行遗传操作;

Step 5 评价临时群体适应值;

Step 6 对当前种群和临时群体按适应值从大到小排序, 并取排序结果前面种群规模大小个个体作为下一代;

Step 7 进化代数加1;

Step 8 输出当前目标函数的最优解, 算法结束.

3 实验与结果分析(Experiments and result analysis)

本文选择了以下函数优化领域的8个典型Benchmarks函数对算法的函数优化能力、效率和精度进行测试:

De Jong's F1函数:

$$f_1(x_i)=\sum_{i=1}^{3}x_i^2,$$

其中$-5.12\leqslant x_i\leqslant 5.12$;

De Jong's F2函数:

$$f_2(x_1,x_2)=100(x_1^2-x_2)^2+(1-x_1)^2,$$

其中$-2.048\leqslant x_i\leqslant 2.048$;

Shaffer's F6函数:

$$f_3(x_1,x_2)=0.5+\frac{\sin^2\sqrt{x_1^2+x_2^2}-0.5}{(1.0+0.001(x_1^2+x_2^2))^2},$$

其中$-100\leqslant x_i\leqslant 100$;

Shaffer's F7函数:

$$f_4(x_1,x_2)=(x_1^2+x_2^2)^{0.25}[\sin^2(50(x_1^2+x_2^2)^{0.1})+1.0],$$

其中$-100\leqslant x_i\leqslant 100$;

六峰驼返回函数:

$$f_5(x_i)=4x_1^2-2.1x_1^4+1/3x_1^6+x_1x_2-4x_2^2+4x_2^4,$$

其中$-5\leqslant x_i\leqslant 5$;

$$f_6(x)=\exp(-0.001x)\cos^2(0.8x),$$

其中$x\in[0,18]$;

$$f_7(x,y)=-x\sin(4x)-1.1y\sin(2y),$$

其中$0\leqslant x\leqslant 10,\ 0\leqslant y\leqslant 10$;

$$f_8(x)=x\sin(10\pi x)+1.0,$$

其中$x\in[-1,2]$.

以上的函数中有部分是震荡剧烈的多峰函数, 函数场景非常复杂, 常规的方法容易陷入局部最优. 为了更好测试GEP-MCFO的算法性能, 本文将函数仿真实验分为两组. 部分参数设置如下: 临时群体规模为父代规模的80%, 同源基因函数集与普通基因函数集一致, 每个函数运行50次. 限于篇幅, 其余的GEP参数设置方法及参数性能分析请参考文献[10]. 实验环境如下: 笔记本, CPU为PIII850 MHz、内存为128 MB, 开发工具为VC++6.0.

3.1 实验1: GEP-MCFO搜索最优解(Testing for GEP-MCFO searching optimum)

根据函数的变元数等因素差异, 本实验对部分参数进行了调整. 同源基因数: f_1设置为3, f_6设置为1, 其余函数都为2. 对函数f_5, 父代种群大小为200, 最大进化代数限制为800; 对其余函数, 父代种群大小为100, 最大进化代数限制为300. 各次测试的染色体长度为90, DM为0.2, CM为0.1. 算法认为在最大进化代数内达到指定精度即为成功进化. 仿真结果见表1. 其中, 平均进化代数是指成功进化情况下的各次进化的代数平均值.

分析表1可知, GEP-MCFO对表1列出的函数的测试均能在平均进化代数小于500的情况下达到100%成功进化. 其中, 它对函数f_6和f_3平均仅需进化1代和1.2代便能成功进化到10^{-12}的高精度, 虽对f_4的精度最差, 但也能达到10^{-4}. 表明该算法有很强的全局寻优能力、较快的收敛速度和较高的解精度.

表 1 实验1结果

Table 1 Result of the first group experiment

	f_1	f_2	f_3	f_4	f_5	f_6
指定精度	10^{-9}	10^{-10}	10^{-12}	10^{-4}	10^{-6}	10^{-12}
成功次数	50	50	50	50	50	50
成功率	100%	100%	100%	100%	100%	100%
平均进化代数	117.85	93.88	1.20	257.05	484.17	1.00
平均每次运行时间/s	64.731	54.708	6.122	781.653	802.415	6.047

3.2 实验2与其他GEP函数优化算法比较 (Comparingwith other GEP algorithms)

为了与Ferreira的GEP函数优化算法结果直接比较, 本实验选取Ferreira实验中仅用的2个函数f_7和f_8进行测试, 均求最大值. Hzero, GEP-PO, GEP-MCFO共有的参数取相同的值, 其中GEP-PO和GEP-MCFO的染色体总长度基本相同, 详见文献[10]. GEP-MCFO独有的同源基因头长度取4和同源基因数目取函数变元个数, 变异率为0.2, DM为0.2, CM为0.1. 实验结果如表2所示.

从表2结果可知, GEP-MCFO能更好地跳出局部峰值找到全局峰值, 且能在较少的运行次数和进化代数中找到最优值. 该算法与Hzero相比的最优解精度误差减小率可达3.46%到667.82%, 与GEP-PO相比的最优解精度误差减小率可达49.51%到60%. 这表明, GEP-MCFO在函数优化时的全局搜索能力、搜索效率和精度都明显优于Hzero和GEP-PO.

表 2　GEP-MCFO与其他GEP算法比较结果

Table 2　Comparison of GEP-MCFO with other GEP algorithms on benchmark

	f_7			f_8		
项目名	Hzero[10]	GEP-PO[10]	GEP-MCFO	Hzero[10]	GEP-PO[10]	GEP-MCFO
最优函数值	2.8502737087	2.8502737665	2.8502737668	18.5546969067	18.5547140746	18.5547210428
历次最好值平均	2.8271147281	2.8360257076	2.8455127548	18.4243180252	18.4547542812	18.4811463854
最优值产生于	第17次的43代	第12次的38代	第7次的40代	第37次的73代	第20次的155代	第8次的732代

4 结论(Conclusion)

本文在GEP的基础上提出了一个有效的函数优化算法GEP-MCFO. 该算法引进了同源基因和细胞系统的思想, 并设计了有利于增大系统表达空间和个体多样性的个体编码方法, 以及能较快全局收敛的遗传操作和种群策略. 两组实验结果表明了GEP-MCFO算法具有较好的收敛性, 且能较好地避开选择与变异的两难问题, 更容易跳出局部最优, 有效地改善了算法的全局搜索能力, 并提高了解的精度.

参考文献(References):

[1] LOZANO M, HERRERA F, KRASNOGOR N, et al. Real-coded memetic algorithms with crossover hill-climbing[J]. *IEEE Transactions on Evolutionary Computation*, 2004, 12(3): 273 – 302.

[2] UYAR A S, HARMANCI A E. A new population based adaptive dominance change mechanism for diploid genetic algorithms in dynamic environments[J]. *Soft Computing*, 2005, 9(11): 803 – 815.

[3] 李宏, 焦永昌, 张莉, 等. 一种求解全局优化问题的新混合遗传算法[J]. 控制理论与应用, 2007, 24(3): 346 – 348.
(LI Hong, JIAO Yongchang, ZHANG Li, et al. Novel hybrid genetic algorithm for global optimization problems[J]. *Control Theory & Applications*, 2007, 24(3): 346 – 348.)

[4] FERREIRA C. Gene expression programming: a new adaptive algorithm for solving problems[J]. *Complex Systems*, 2001, 13(2): 87 – 129.

[5] ZUO J, TANG C J, ZHANG T. Mining predicate association rule by gene expression programming[C] //*Proceedings of the 3rd International Conference on Advances in Web-Age Information Management, Lecture Notes In Computer Science*. Berlin: Springer-Verlag, 2002, 2419: 92 – 103.

[6] HARDY Y, STEEB W H. Gene expression programming and one-dimensional chaotic maps[J]. *International Journal of Modern Physics*, 2002, 13(1): 25 – 30.

[7] ZHOU C, XIAO W M, NELSON P C, et al. Evolving accurate and compact classification rules with gene expression programming[J]. *IEEE Transactions on Evolutionary Computation*, 2003, 7(6): 519 – 531.

[8] LOPES H S, WEINERT W R. EGIPSYS: an enhanced gene expression programming approach for symbolic regression problems[J]. *International Journal of Applied Mathematics and Computer Science*, 2004, 14(3): 375 – 384.

[9] 彭京, 唐常杰, 李川, 等. M-GEP: 基于多层染色体基因表达式编程的遗传进化算法[J]. 计算机学报, 2005, 28(9): 1459 – 1466.
(PENG Jing, TANG Changjie, LI Chuan, et al. M-GEP: a new evolution algorithm based on multi-layer chromosomes gene expression programming[J]. *Chinese Journal of Computer*, 2005, 28(9): 1459 – 1466.)

[10] FERREIRA C. *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*[M]. New York: Springer-Verlag, 2006.

[11] 向勇, 唐常杰, 曾涛, 等. 基于基因表达式编程的多目标优化算法[J]. 四川大学学报(工程科学版), 2007, 39(4): 124 – 129.
(XIANG Yong, TANG Changjie, ZENG Tao, et al. Multiobjective optimization based on gene expression programming[J]. *Journal of Sichuan University(Engineering Science Edition)*, 2007, 39(4): 124 – 129.)

[12] YUAN C A. Markov chain analysis of gene expression programming[J]. *Journal of Information and Computational Science*, 2008, 5(5): 2027 – 2034.

作者简介:

彭昱忠　(1980—), 男, 讲师, 主要研究方向为进化计算, Email: jedison@163.com;

元昌安　(1964—), 男, 博士, 教授, 主要研究方向为进化计算、数据挖掘, E-mail: yca@gxtc.edu.cn;

陈建伟　(1968—), 男, 讲师, 主要研究方向为人工智能、数学建模;

吴信东　(1963—), 男, 教授, 长江学者, 博士生导师, 主要研究方向为基于知识的系统、数据挖掘, E-mail: xwu@cems.uvm.edu;

王汝凉　(1963—), 男, 博士, 教授, 主要研究方向为自动化控制理论, E-mail: wrl@gxtc.edu.cn.