软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(5):899−913 [doi: 10.3724/SP.J.1001.2011.03768]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

用态势模型预测基因表达式编程的进化难度*

郑皎凌[1,2], 唐常杰[1+], 徐开阔[1], 杨 宁[1], 段 磊[1], 李红军[1]

[1](四川大学 计算机学院,四川 成都 610065)

[2](成都信息工程学院 软件工程系,四川 成都 610225)

Gene Expression Programming Evolution Difficulty Prediction Based on Posture Model

ZHENG Jiao-Ling[1,2], TANG Chang-Jie[1+], XU Kai-Kuo[1], YANG Ning[1], DUAN Lei[1], LI Hong-Jun[1]

[1](College of Computer Science, Sichuan University, Chengdu 610065, China)

[2](Department of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

+ Corresponding author: E-mail: cjtang@scu.edu.cn

Zheng JL, Tang CJ, Xu KK, Yang N, Duan L, Li HJ. Gene expression programming evolution difficulty prediction based on posture model. *Journal of Software*, 2011,22(5):899−913. http://www.jos.org.cn/1000-9825/3768.htm

Abstract: Fitness Distance Correlation (FDC) can hardly predict the evolution difficulty of Gene Expression Programming (GEP) because problems with different hardness would result in very similar FDC values in GEP. To solve the problem, the authors propose a posture model and region density to predict GEP's evolution difficulty. This study made the following contributions: (1) It introduces the concepts of the chromosomes' distance and posture model in GEP; (2) It proposes region density of a posture model; (3) It proves that the posture model is a mapping from the original searching space, and the mapping preserves the population's dynamic migration property in the original searching space; (4) It demonstrates the validity of using posture model and region density to predict GEP's evolution difficulty; (5) It conducts extensive experiments to show that the new model can precisely predict the evolution difficulty of GEP.

Key words: gene expression programming (GEP); evolution difficulty; posture model; region density; space mapping

摘 要: 在基因表达式编程(gene expression programming,简称 GEP)中,由于不同问题得到的适应度-距离相关系数(fitness-distance correlation,简称FDC)值很相近,所以难以用FDC预测GEP求解不同问题的进化难度.为了解决该问题,提出了态势模型及其区间密度指标来预测GEP的进化难度.主要工作包括:(1) 提出了GEP染色体之间的距离和态势模型的新概念;(2) 提出了态势模型中的区间密度指标;(3) 从动力学角度证明了态势模型是对 GEP 原搜索空间的一种映射,并且该映射保持了种群在原搜索空间中移动的动力学性质;(4) 分析了用态势模型区间密度预测GEP进化难度的合理性;(5) 用实验验证了区间密度能够准确预测GEP求解问题的进化难度.

关键词: 基因表达式编程(GEP);进化难度;态势模型;区间密度;空间映射

* 基金项目: 国家自然科学基金(60373000); 国家科技支撑计划(2006BAI05A01); 中国博士后科学基金(20090461346); 教育部人文社会科学研究青年基金(10YJCZH117); 中央高校基本科研业务费专项资金科技创新项目(SWJTU09CX035); 成都信息工程学院引进人才项目(KYTZ201110)

收稿时间: 2009-01-02; 修改时间: 2009-06-01; 定稿时间: 2009-11-04

中图法分类号: TP183 文献标识码: A

进化计算在本质上是对种群空间极优个体的搜索过程,由于搜索空间庞大以及搜索的随机性,很难通过对整个搜索空间的少量采样预测搜索到极优个体的难度并进行量化.目前,一般用成功率或进化代数来衡量进化难度[1−5],但这需要反复运行程序多次.本文旨在通过对整个搜索空间进行少量采样来预测搜索到极优个体的难度,即 GEP 的进化难度.

定义 1(进化难度 *Hard*(*P*)). 对特定问题 *P*,用进化算法求解 *P* 的进化难度记为

$$Hard(P)=1-Optimal\_Number/Total\_Number,$$

其中,*Total_Number* 是算法总运行次数,*Optimal_Number* 是得到极优解的运行次数.如果 *Hard*(*P*)为 1,则表明在已经进行的实验中,没有任何一次实验找到了极优解,说明问题非常难;反之亦然.

定义 2(GEP 搜索空间). 给定问题 *P*,设 *c* 是 GEP 求解 *P* 过程中的任意合法染色体,*fitness*(*c*)是 *c* 的适应度,称所有$\langle c,fitness(c)\rangle$二元组构成的集合(即$\{\bigcup\langle c,fitness(c)\rangle\}$)为问题 *P* 的 GEP 搜索空间,记为 *Space*(*P*).

定义 3(进化难度预测问题). 设有一个问题集合$\{P_1,P_2,\ldots,P_n\}$,用 GEP 求解的进化难度为$\{Hard(P_1),Hard(P_2),\ldots,Hard(P_n)\}$,搜索空间为$\{Space(P_1),Space(P_2),\ldots,Space(P_n)\}$,对每个问题的搜索空间进行少量采样得到采样空间$\{Space'(P_1),Space'(P_2),\ldots,Space'(P_n)\}$.进化难度预测问题旨在通过采样空间找出一种量化指标 *x*,使得对任意问题 $P_i,P_j(1\leqslant i,j\leqslant n)$,如果有 $Hard(P_i)<Hard(P_j)$,则都有 $x_i<x_j$ 或都有 $x_i>x_j$.

本文提出了基于态势模型的区间密度指标来预测 GEP 的进化难度,通过实验证实了区间密度满足问题定义 3 中的要求.又由于种群的进化本质上是一种动力学过程,本文通过引入适应度增长系数,证明区间密度反映了 GEP 进化过程的动力学本质.

1 相关工作

基因表达式编程(gene expression programming,简称 GEP)[1]融合了遗传算法(GA)和遗传编程(GP)的优势.其富有特色的染色体头和尾定义,能够保证 GEP 个体在进行各种遗传操作时始终产生有效语义个体,使 GEP 进化速率比 GP 平均快 2~4 个数量级.文献[2−5]研究了用 GEP 求解不同问题的性能.

在对 GA,GP 求解问题进化难度预测的研究中,基于其各自的搜索空间建立了适应度景观模型(fitness landscape),并在此基础上提出了基于适应度-距离相关系数(fitness-distance correlation,简称 FDC)的量化指标.认为问题的 FDC 越小(即越接近−1)越简单,反之亦然.文献[6−9]中分别讨论了用 GA,GP 求解问题的进化难度与 FDC 的关系,都得出了适应度与距离负相关性越强,进化难度越小的结论.由于种群在地形上的移动是由变异算子引起的,文献[10,11]进一步研究了变异算子与 FDC 及进化难度的关系,在不同的进化算子作用下,仍然有上述结论.另外,由于 FDC 表达的是一种强线性相关关系,文献[12−15]又对适应度-距离的一般分布形式与进化难度的关系进行了研究,仍发现 GA,GP 中的极优点彼此都离得较近,而离极优点越远的点适应度越低.故 FDC 能够较为准确地预测 GA,GP 求解特定问题的进化难度.

2 GEP 与 GA,GP 搜索空间的区别

把适应度景观模型和 FDC 用于 GEP 进化难度的预测中发现,不同难度 GEP 问题的 FDC 值很相近,难以起到预测的作用.比如,在实验中,我们分别对 7 个采用 GEP 进行求解并且 GEP 进化难度已知的问题进行了实验,发现它们的 FDC 值不但很相近,而且部分结果甚至与 FDC 的预测结论相反.所以,无法采用 FDC 来衡量 GEP 的进化难度.

经分析发现,GEP 与 GA,GP 在搜索空间的全局极优点数量上以及种群在搜索空间中的迁移能力上存在本质不同,故无法用适应度景观模型和 FDC 来预测 GEP 的进化难度.为此,本文基于 GEP 自身的特点提出了一种新的模型(态势模型,将在第 3 节给出)和衡量指标(区间密度,将在第 3 节给出).通过实验分析,证实区间密度能够很好地预测 GEP 的进化难度.例 1 分析了 GEP 与 GA,GP 在各自搜索空间中,全局极优点数量以及种群在搜索

空间中迁移能力上的区别.

例 1:设有数值优化问题,只有 1 个全局极优值(0,0,0).下面给出 GEP,GA,GP 求解该问题的区别.

区别 1. 搜索空间中全局极优点的数量.

(1) GA:数量少,只存在 1 个全局极优点(0,0,0).

(2) GP:数量多,如(0,0,0),((1−1),(2×3)−(3×2),(1−(3/3))或(0,(3−3),((1×1)−(1×1))).

(3) GEP:数量多,与 GP 类似.

区别 2. 种群在搜索空间中的迁移能力.

(1) GA:迁移能力弱,是一种渐变过程类似梯度搜索,如从(0,0,5)变到(0,0,0),一旦接近极优点则很难再跳到更远的搜索区域.

(2) GP:迁移能力较弱,由于 GP 中操作符节点与终结符节点只能变异成同种类型的节点,如可以从图 2(a)变为图 2(b),但不能发生图 1(a)到图 1(b)这种较大的变异.

(3) GEP:迁移能力强,如图 1(a)是一个很复杂的 GEP 染色体,但只需一次单点变异就可以变为图 1(b).

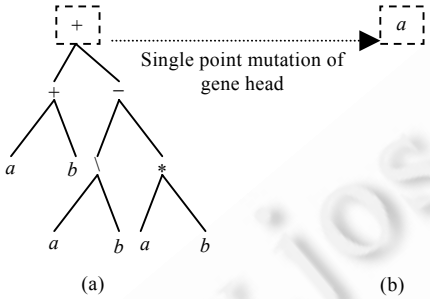


Fig.1 GEP single point mutation

图 1 GEP 单点变异

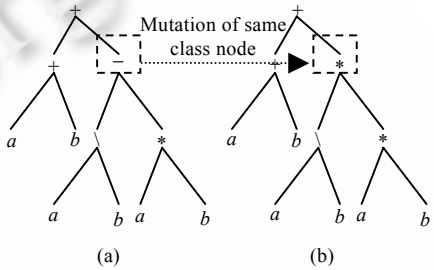


Fig.2 GP single point mutation

图 2 GP 单点变异

区别 1、区别 2 除了表达 GEP 与 GA,GP 的区别以外,还表达了特定问题用 GEP 求解的困难程度与由该问题构成的搜索空间之间的关系.得到如下观察,该观察已经在文献[2]中得到了实验证实:

观察 1. 搜索空间中全局极优点数量越多,且在[0,1]区间分布得越均匀,问题求解越简单;反之亦然.

解释:由于适应度景观模型和 FDC 衡量的是适应度与种群个体之间的相关性,但这种相关性却无法衡量上述观察中的特征,所以无法用 FDC 来解释 GEP 的进化难度.因此,我们提出了态势模型及建立在其上的区间密度指标来预测 GEP 的进化难度.

本文第 3 节定义 GEP 染色体之间的距离,并在此基础上提出 GEP 态势模型.第 4 节提出量化指标“区间密度”来预测用 GEP 求解特定问题的进化难度,并且在实验中验证该指标的有效性.第 6 节研究态势模型对搜索空间的映射性质,并从动力学角度分析这种映射的合理性.

3 GEP 态势模型

3.1 GEP染色体距离

GEP 中的染色体可用 *K*-表达式和 *K*-表达式树两种方式表达,个体的适应度由 *K*-表达式树解析得到.虽然 GEP 个体在进行各种遗传操作(如变异、交叉、插串等)时,是基于字符串形式的 *K*-表达式进行的,但个体适应度是通过 *K*-表达式树得到的.由于体现 GEP 个体真正差异的是它们的适应度而不是编码形式,故采用 GEP 中 *K*-表达式树间的距离而不是 *K*-表达式的距离来代表个体的距离.

关于 GEP 的基本概念和属性,如头部长度、基因长度、*K* 表达式、*K* 表达式树及其高度等,参见文献[1].在文献[9]中给出了 GP 中染色体所形成的树的距离,本文将其修改、扩展,用来衡量 GEP 中染色体距离.为了准

确、简捷地描述,本文将采用表 1 中列出的符号.

Table 1 Character table of posture model

表 1 态势模型的符号表

$root(T)$	Root node of K-expresson tree	$fitness(c)$	Fitness of chromosome c
$td(T)$	Depth of tree(T)	$P=\{p_1,p_2,...,p_k\}$	Investigated problem set
$n(T)$	The number of root node's child node	$h_{\max}$	The maximum height of K-expression tree in P
$\mathrm{Max}(n(T_1),n(T_2))$	The larger value of $n(T_1)$ and $n(T_2)$	$s_i(T)$	The i-th sub tree of tree(T)'s root node

定义 4(GEP 染色体距离). GEP 中两个染色体对应的 k-表达式树 T_1,T_2 的距离 $d(T_1,T_2,m)$为:

(1) 如果 $root(T_1)\neq root(T_2)$,则 $d(T_1,T_2,m)=m+|td(T_1)-td(T_2)|$;

(2) 如果 $td(T_1)=td(T_2)=0$,则 $d(T_1,T_2,m)=0$;

(3) 其他情况规定,$d(T_1,T_2,m)=\sum_{i=1}^{\mathrm{Max}(n(T_1),n(T_2))}\frac{d(s_i(T_1),s_i(T_2),m-1)}{\mathrm{Max}(n(T_1),n(T_2))}$, 其中,$m=h_{\max}$.

公式采用加权的方式计算两个染色体的 k 表达式树的距离,计算中对不同深度的节点赋予不同的权重,根节点的权重为 m,然后每一层递归地减少 1.由于 $m=h_{\max}$ 所以权重在递归下降时不会变为负数.

例 2(GEP 染色体距离):设 GEP 染色体头部长度为 3,非终结符为{+,–},终结符为{a,b,c},有两个单基因的 GEP 染色体+–$cabab$ 和++–$abab$,其对应的 k-表达式树 T_1,T_2 如图 3 所示.由于 $root(T_1)=root(T_2)$,需要进行递归计算,因为 $m=3$,故 $d(T_1,T_2,m)=(3-1+2-2)/2+(3-1+2-1)/2=5$.其中,(3–1+2–2)/2 是 T_1,T_2 两棵左子树的距离,(3–1+2–1)/2 是 T_1,T_2 两棵右子树距离.

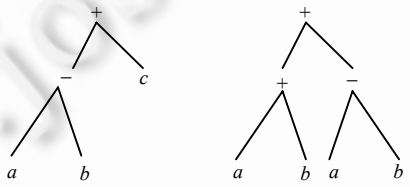


Fig.3 Ddistance between GEP chromosomes

图 3 GEP 染色体的距离

定理 1. 定义 4 描述的树距离满足度量空间中距离的 4 项准则,即:

(1) 非负性:如果 $T_1\neq T_2$,则 $d(T_1,T_2,m)>0$.

(2) 反身性:如果 $T_1=T_2$,则 $d(T_1,T_2,m)=0$.

(3) 对称性:$d(T_1,T_2,m)=d(T_2,T_1,m)$.

(4) 三角不等式:$d(T_1,T_2,m)+d(T_2,T_3,m)\geqslant d(T_1,T_3,m)$.

证明:限于篇幅,此处从略,详见附录. □

定理 2. 设 T_1,T_2 是 GEP 中两个染色体对应的 k-表达式树,其头部长度为 h,则 $d(T_1,T_2,m)$的最大值为 $m+h$.

证明:设 T_1 和 T_2 为单基因染色体:(1) 如果 $root(T_1)\neq root(T_2)$,最大值使得$|td(T_1)-td(T_2)|$最大,此时,只需要其中一个的深度为最小深度 1,而另一个的深度为最大深度 h+1,则最大距离为 $m+h$;(2) 如果 $root(T_1)=root(T_2)$,且 T_1,T_2 不为空,求解定义 4 中的递归方程,并设递归在第 $L(1<L\leqslant h+1)$层停止.可得 $d(T_1,T_2,m)<(m-L)+h<m+h$.由此证得 $d(T_1,T_2,m)$的最大值为 $m+h$. □

3.2 GEP染色体距离和适应度的正规化

由于不同的问题可能选择不同的 GEP 参数和适应度函数,为了消除由参数和适应度函数的不同导致距离的差异,对染色体的距离和适应度进行了规范化.具体如算法 1 所示.

算法 1. 规范化染色体的距离及染色体的适应度.

输入:染色体原始距离集合 $dist[n]$,适应度集合 $fit[n]$极优适应度 $bestfit$,染色体头部长度 h,定义 4 给定的 m,

m=10.

输出:规范化了的染色体距离集合 *normalized_dist*[*n*],适应度集合 *normalized_fit*[*n*].

1. for (*i*=0; *i*<*n*; *i*++)
2. {*normalized_dist*[*i*]=*dist*[*i*]/(*m*+*h*);
3. *normalized_fit*[*i*]=*fit*[*i*]/*bestfit*;}

值得指出的是,在定义 3 中规定要使 m 的值在递归计算染色体距离时非负,需要 $m=h_{\max}$.其中,$h_{\max}$ 是当染色体头部长度为 h 时,k-表达式树的最大树高,显然,$h_{\max}=h+1$.而由于在实验部分求解各个问题的 GEP 头部长度设置均小于 10,故在算法 1 中将 m 的值定为 10.

3.3 GEP态势模型

态势模型是在采样空间基础上得到的,主要是去掉了采样空间中到极优个体距离相等但适应度较小的个体.直观来说,态势模型类似于等高线地图,由各个染色体彼此之间的相对距离及其适应度构成.第 3.3.1 节给出了采样算法,第 3.3.2 节给出了态势模型的具体定义.

3.3.1 GEP 搜索空间采样

算法 2 给出了对搜索空间进行采样的算法.整个采样过程通过运行一次 GEP 算法完成,当采集的样本个数达到阈值时停止.为了能够均匀地对整个搜索空间进行采样,采用 Execute_Evolution 方法来指导 GEP 算法的每一次种群进化.Execute_Evolution 中各种变异算子的变异率均为 1,即每一代种群都是前一代的随机变异,通过这种方式来实现最大程度的均匀采样,即 Execute_Evolution 的目的不是进化出极优个体,而是实现均匀采样.

算法 2. GEP 搜索空间采样算法.

输入:样本集合 *S*,初始为空,*num* 为需要采集的样本个数,*population* 为 GEP 进化过程中的种群.

输出:采样后的样本集合 *S*.

```
1. init (population);                          //初始化种群
2. count=0;
3. while(count<num)
4.     {  int k=integer(random());
5.        S←population[k],                      //随机选择种群中的第 k 个个体放入 S
6.        Execute_Evolution(population);        //种群进化一次
7.        count++;}
```

3.3.2 GEP 的态势模型

定义 5(态势模型). 给定搜索空间的样本集 $S=\{p_1,p_2,...,p_n\}$,$p_1,p_2,...,p_n$ 是采样得到的 GEP 个体:

(1) 给定阈值 *fitness_threshold*,对 S 中任意个体 c,若 $fitness(c)\geqslant fitness\_threshold$,则称 c 为极优个体.

(2) 称 $X=\{d(p_i,c,m)|\ p_i\in S\wedge c$ 是任意极优个体$\}$为样本集 S 中所有个体到极优个体 c 的距离集合.

(3) 记 $M=\{\langle x_i,y_i\rangle|x_i\in X,y_i=\mathrm{Max}\{fitness(p_i)|p_i\in S\wedge d(p_i,c,m)=x_i\}\}$.$M$ 称为样本集 S 上的态势模型,c 是 S 中任意一个极优个体,y_i 是所有到极优个体 c 距离均为 x_i 的个体中最大的适应度值.

例 3:设样本集 S 有 8 个个体 $S=\{p_1,p_2,...,p_8\}$,极优个体为 p_1,其他个体到 p_1 的距离和适应度为$\{\langle 0.1,1\rangle,\langle 0.2,0.7\rangle,\langle 0.2,0.9\rangle,\langle 0.5,0.3\rangle,\langle 0.5,0.8\rangle,\langle 0.5,1\rangle,\langle 0.9,1\rangle\}$,则最后由 S 构成的态势模型为

$$M=\{\langle 0.1,1\rangle,\langle 0.2,0.9\rangle,\langle 0.5,1\rangle,\langle 0.9,1\rangle\}.$$

4 基于态势模型的区间密度指标预测 GEP 进化难度

由第 3 节可知,态势模型实际上是将整个 GEP 的搜索空间映射到了一个二维空间中,y 轴表示个体的适应度,x 轴表示每个个体到极优个体的相对距离.并且,由于算法 1 对个体的适应度和相对距离进行了规范化,所以 x 轴和 y 轴构成了一个长度均为 1 的区间.

定义 6(区间密度). 给定搜索空间样本集 $S=\{p_1,p_2,...,p_n\}$,S 的态势模型 M,适应度阈值 *fitness_threshold*,其中,

适应度不小于 *fitness_threshold* 的个体称为极优个体,

(1) 区间分辨率 d.$d>0$ 且 $d\in N$,直观来说,d 是 M 的 x 轴被等分的数量.

(2) 子区间 $Sub_i(M)$.$\{\langle x,y\rangle|i\in N\wedge i\in[0,d-1],x\in[1/d\times i,1/d\times(i+1)],y=fitness(x)\}$.

(3) 给定态势模型 M,M 的区间密度为 $Density(M)=\sum_{i=1}^{d}\frac{|\{c=\langle x,y\rangle\in Sub_i(M)\,|\,y>fitness\_threshold\}|}{|\{c=\langle x,y\rangle\in Sub_i(M)\}|}$.其中,分子是第 i 个子区间 $Sub_i(M)$中极优个体的数量,分母是 $Sub_i(M)$中所有个体的数量.

注意,当极优个体越多时分子越大,区间密度值也越大;当非极优个体越少时分母越小,区间密度值也越大.同时,由于区间分辨率的限制,极优个体在所有子区间分布得越多,区间密度值越大.故区间密度能够很好地量化观察 1 中的结论.并且易知,$Density(M)\in[0,d]$.

例 4:图 4(a)~图 4(c)给出了 3 种不同的态势模型,区间分辨率为 3.适应度为 1 的点为极优个体.对图 4(a)而言,$Density(M)=1+1+1=3$;对图 4(b)而言,$Density(M)=1+1/2+1/2=2$;对图 4(c)而言,$Density(M)=1+0+0=1$.

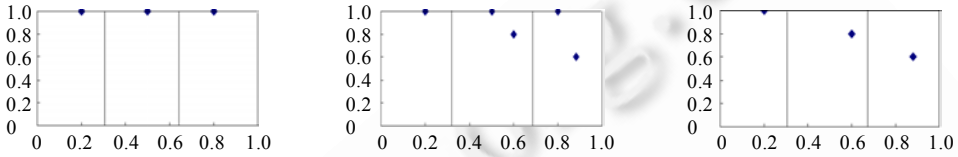


Fig.4 Posture model

图 4 态势模型

观察 2. 区间密度反比于 $Hard(P)$,即给定问题集合$\{P_1,P_2,\ldots,P_n\}$,设用 GEP 求解进化难度为$\{Hard(P_1),Hard(P_2),\ldots,Hard(P_n)\}$,则对任意问题 $P_i,P_j(1\leqslant i,j\leqslant n)$,如果 $Hard(P_i)<Hard(P_j)$,都有 $Density(M_i)>Density(M_j)$,$M_i$,$M_j$ 分别是 P_i,P_j 的态势模型.

解释:(1) 对于图 4(a),由于在第 2 节中已经指出 GEP 个体的迁移能力很强,可以轻易地从一个区间跳到另一个任意的区间,而图 4(a)中每个区间都有一个极优点,故种群能够很容易地找到全局极优;(2) 对于图 4(b),虽然该态势模型在每个区间仍然都有极优点,但每个区间中还存在非极优点,这就形成了干扰,使得进化难度增加;(3) 对于图 4(c),由于在第 2、第 3 子区间没有极优点,种群只能在第 1 子区间找到极优点,所以,图 4(c)的态势模型对种群来说是进化难度最大的.值得指出的是,如果采用 GA,GP 中的 FDC 来衡量例 4 中 3 种态势模型的进化难度,则结果恰好相反.这说明,FDC 无法预测 GEP 种群的进化难度,也说明了用态势模型及区间密度来预测 GEP 进化难度的合理性.

5 区间密度对极优个体选择的不敏感性

由于样本集 S 可能有多个极优个体,而在定义 5 中极优个体的选取是随机的,当选择不同的极优个体 c_1,c_2 作为原点时,原态势模型 M_0 会变成新的态势模型 M_1,原因如下:(1) 对 M_0 中的一个个体 c,如果 c 到 c_1,c_2 的距离不同,则 c 会从 M_0 中一个子区间移动到 M_1 中的另一个子区间;(2) 对 M_0 中的一个个体 $c=\langle x,y\rangle$,如果 c 在 M_1 中变为 $c=\langle x_1,y\rangle$,且在 M_1 中存在另一个个体 $c'=\langle x',y'\rangle$,且有 $x'=x_1$ 和 $y'>y$,则根据态势模型的定义,c 在 M_1 中将消失掉.定理 3 证明了样本集 S 在态势模型中的分布方式对极优个体的选取不敏感.为了严格地表述性质和简化问题,本文将使用表 2 中列出的符号和(合理的)假设.

假设. 有样本空间 S,用其中任意两个极优个体作为原点得态势模型 M_i,M_j,并设区间分辨率为 d,则有:

(1) $N\_global(M_i)=N\_global(M_i)=C_1\times N\_global(S)$,$N\_local(M_i)=N\_local(M_i)=C_2\times N\_local(S)(0<C_1,C_2\leqslant 1)$;

(2) 对 M_i,M_j 及其中的任意两个子区间 s_1,s_2,有 $P(M_j,s_2|M_i,s_1)=P(M_i,s_1|M_j,s_2)=1/d$.

解释:(1) 假设(1)的合理性.设有态势模型 M_i,M_j,如果任意染色体 p_1,p_2 在 M_i 中到原点 o_i 的距离相同,则有很

大可能它们在 M_j 中到原点 o_j 的距离也相同,根据态势模型定义有,假设(1)是合理的.(2) 假设(2)的合理性.设 c 是 S 中任意个体,如果 c 从 M_i 的任意子区间 s_1 移动到 M_j 的任意子区间 s_2,其可能移动到 M_j 的任意子区间,而 M_j 有 d 个子区间,故移动到每个子区间的概率为 1/d.

Table 2 Character table of GEP chromosome

表 2 GEP 染色体的符号表

M_i	Assume that there are totally n local best chromosome in the sample space, M_i is the posture model that is created by choosing the i-th ($1\leqslant i\leqslant n$) local best chromosome as the origin
$P(M_j,s_2\|M_i,s_1)$	For any point q, $q\in S$, $P(M_j,s_2\|M_i,s_1)$ is the probability of q moving from M_i's any sub region s_1 to M_j's any sub region s_2 after we change the posture model from M_i to M_j
$N\_global(S)$	The number of local best chromosomes in sample set S
$N\_local(S)$	The number of chromosomes in sample set S which are not the local best ones
$N\_global(M_i)$	The number of local best points in M_i
$N\_local(M_i)$	The number of points in M_i which are not the local best points
$N\_global(s_k,M_i)$	The number of local best points in M_i's sub region s_k
$N\_local(s_k,M_i)$	The number of points in M_i's sub region s_k which are not the local best points

定理 3. 设样本集 S 有 k 个不同的极优个体,则 S 有 k 种态势模型,分别为 $M_1,M_2,...,M_K$,设每个模型的区间分辨率均为 1/d,则各态势模型的区间密度取下列值的概率最大:

(1) 当 d=1 时,$Density(M_1)=Density(M_2)=...=Density(M_k)=C_1\times N\_global(S)/(C_1\times N\_global(S)+C_2\times N\_local(S))$;

(2) 当 d>1 时,$Density(M_1)=Density(M_2)=...=Density(M_k)=d\times C_1\times N\_global(S)/(C_1\times N\_global(S)+C_2\times N\_local(S))$.

证明:限于篇幅,此处从略,详见附录. □

根据定理 3,对由不同极优点的态势模型,其各自区间密度取值 $d\times C_1\times N_global(S)/(C_1\times N_global(S)+C_2\times N_local(S))$的概率最大,上式虽然与 d 有关,但只是都将 D=1 时的值乘以 d 倍,故不影响对问题难度相对大小的衡量.故可以看作区间密度只与样本集 S 有关.

例 5(区间密度对极优个体选择的不敏感性):对例 4 中给出的 3 种态势模型,当区间密度 d=1 时,计算得到图 4(a)~图 4(c)的区间密度分别为 1,3/5,1/3.当区间密度 d=3 时,图 4(a)~图 4(c)的区间密度分别为 3,2,1.如果按照定理 3,通过 d=1 时的区间密度推导 d=3 时的区间密度,可得图 4(a)~图 4(c)的区间密度分别为 3,9/5 和 1.与实际值的误差分别为 0,1/5 和 0.可见,例 5 符合定理 3 的结论.

6 动力学分析

第 3 节已指出,态势模型实际上是将整个 GEP 搜索空间映射到一个二维空间.用区间密度指标来预测 GEP 进化难度,本质上是用映射空间(即态势模型)的指标(即区间密度)来衡量原空间的进化难度.而由于种群在搜索空间中的进化本质上是一种动力学过程,如果两个空间中的种群遵循相同的进化原则进行搜索并且表现出相同的动力学性质,则认为用映射空间的量化指标来解释另一个空间的进化难度是合理的.我们把上述映射空间称为保持动力学的映射空间,简称保动映射空间.下面给出保动映射空间的形式化定义,并验证对实验中选出的一个有代表性的问题集合,态势模型是各个问题原搜索空间的保动映射空间.

6.1 用动力学解释态势模型的合理性

为了解释态势模型的动力学性质引入如下一些概念:

定义 7(映射搜索空间). 给定问题 P,设 c 是 GEP 求解 P 过程中的任意合法染色体,$fitness(c)$是 c 的适应度,设 $Space(P)=\{\bigcup\langle c,fitness(c)\rangle\}$ 是 P 的 GEP 搜索空间,则 P 的 GEP 映射搜索空间为 $\{\bigcup f(\langle c,fitness(c)\rangle)\}$. 其中,$f$ 可以是任意映射函数.

定理 4. 对任意问题 P,态势模型是一个 P 的映射搜索空间.

证明:设 $c_0=\langle x_0,y_0\rangle$和 $c_i=\langle x_i,y_i\rangle$分别是问题 P 的 GEP 搜索空间 S 中的极优个体和任意合法个体,则态势模型 M 可以写成 $M=\{\bigcup f(\langle c,fitness(c)\rangle)\}$, 其中,$f$ 可以写成 $f(\langle c,fitness(c)\rangle)=\langle d(c,c_0,m),fitness(c')\rangle$,其中,$c'$是所有到 c_0 距离为 $d(c,c_0,m)$的个体中具有最大适应度的个体.易知,等式右端满足定义 7,故命题得证. □

定义 8(适应度增长系数$\alpha(Space(P))$). 给定问题 P 及其搜索空间 $Space(P)$,种群在 t 代达到的最大适应度为 $\mathrm{Max}(fitness(G^t))$,称满足等式 $\mathrm{Max}(fitness(G^t))=\alpha\times\ln t+\mathrm{Max}(fitness(G^1))$的实数$\alpha$为 $Space(P)$的适应度增长系数.

注意,定义 8 是基于圣塔菲学派关于进化过程的研究[6],即所有进化种群的适应度变化是按时间作指数衰减.由于 t 取对数后会大大减小,故 $\mathrm{Max}(fitness(G^t))$很大程度上依赖于增长系数$\alpha$,$\alpha$越大进化难度越小;反之亦然.

定义 9(保动映射空间). 给定问题集合$\{P_1,P_2,\ldots,P_n\}$,记相应的 GEP 搜索空间的集合为$\{Space_1(P_1),Space_1(P_2),\ldots,Space_1(P_n)\}$,记 GEP 映射搜索空间集合为$\{Space_2(P_1),Space_2(P_2),\ldots,Space_2(P_n)\}$.

(1) 若对两个空间中的种群施加相同进化策略,则称两空间中的种群具有相同进化原则;

(2) 若两个空间中的种群遵循相同进化原则,且对任意问题 $P_i,P_j(1\leqslant i,j\leqslant n)$存在关系$\alpha(Space_1(P_i))<\alpha(Space_1(P_j))$,都有$\alpha(Space_2(P_i))<\alpha(Space_2(P_j))$,则称 $Space_2(P)$为 $Space_1(P)$保持动力学性质的映射空间,简称保动映射空间.

因为保动映射空间是与特定问题相关的,由于不可能对一切问题验证定义 9 中的 3 个条件,故选择 6 个有代表性的问题,目的是验证这 6 个问题的态势模型是各问题原搜索空间的保动映射空间.

整个验证过程分为 3 步:(1) 验证对任意问题 P,态势模型是 P 的映射搜索空间,这在定理 4 中已证明;(2) 本节旨在验证定义 9 中的第 1 个条件,即给出种群在态势模型中的进化策略,验证其与种群在原空间中的进化策略是相同的;(3) 第 7.3 节将验证定义 9 的第 2 个条件,即对任意问题 P_i,P_j 存在关系$\alpha(Space_1(P_i))<\alpha(Space_1(P_j))$,都有$\alpha(Space_2(P_i))<\alpha(Space_2(P_j))$,其中,$Space_1$ 是问题的原搜索空间,$Space_2$ 是态势模型空间.

因为在定理 3 中已经证明态势模型中种群的分布方式对极优个体的选取不敏感,故可以选取任意一个极优个体为原点,得到问题的态势模型.图 5 描述了种群在态势模型空间的进化过程,具体过程如算法 3 所示.

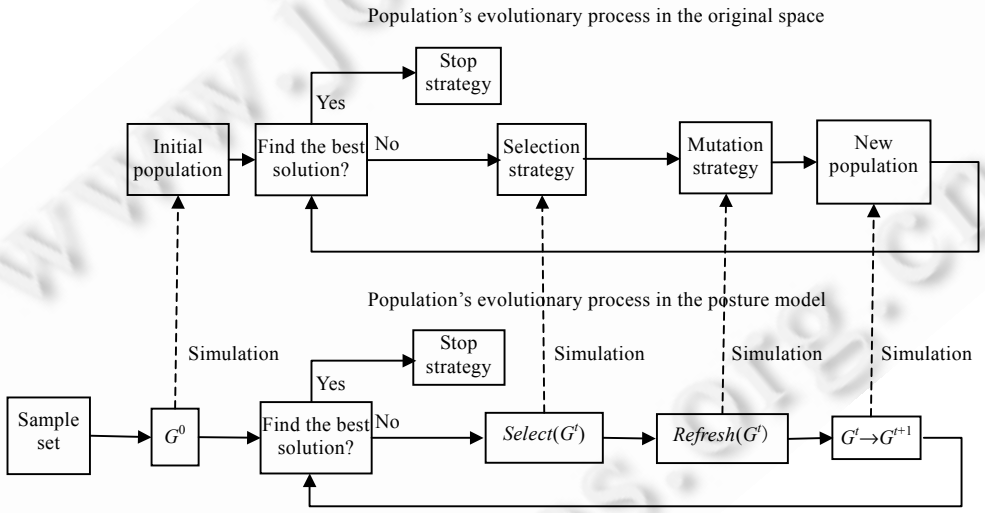


Fig.5 Using population's evolution in posture model to simulate population's evolution in the original space

图 5 用种群在态势模型上的进化过程模拟种群在原空间中的进化过程

算法 3. 动力学模拟算法.

输入:初始种群 G^0,态势模型 M.

输出:$list(t,\mathrm{Max}(fitness(G^t)))$,$list$ 的每个元素包含两个值,分别为定义 8 中的 t 和 $\mathrm{Max}(fitness(G^t))$.

1. t=0; $G^t=Init(S)$;
2. While (!$best(G^t)$)
3. { $G^t=Select(G^t)$;

4. G^t=*Refresh*(G^t);
5. t++;
6. }
7. Return t;

算法 4. *Refresh*(G^t).

输入:种群 G^t,样本集合 S,变异率 v.

输出:更新后的 G^t.

1. double *probabilities*[|G^t|][|S|];
2. for each p=⟨x_p,y_p⟩ in G^t; //x_p,y_p 是 p 到极优个体的距离和适应度
3. { int m_0=*random*();
4. if (m_0<v)
5. { for each q=⟨x_q,y_q⟩ in S
6. { int m_1=*random*(); int m_2=*random*();
7. if (m_0<|x_p−x_q| && m_1<|y_p−y_q|)
8. { p←q; break;}}}}

(1) *Init*(S)表示从态势模型 M 中随机选择一定数量的点放入 G^t 中.

(2) *Select*(G^t)表示用概率选择机制对 G^t 进行选择.其中,为体现精英策略,将 G^t 中的极优点直接放入 G^{t+1} 中.为体现优胜劣汰的概率选择机制,采用轮盘赌机制按适应度大小决定个体进入下一代的概率大小.故 *Select* 与 GEP 中的选择策略是相同的.

(3) *Refresh*(G^t)表示对 G^t 的更新.由于 GEP 各种变异本质上是从一个个体变异成另一个个体,故 *Refresh* 用态势模型平面上的点到另一个点的移动来模拟这种变异.对 G^t 中的每一个点 p(第 2 行)都会产生一个 0,1 之间的随机数 m_0(第 3 行),如果 m_0 小于变异率 v,则需要将 p 更新成 M 中的另一个点.更新方法在第 5 行~第 8 行中给出.对于 M 中的任意一点 q,p,q 的距离和适应度越近,p 被更新成 q 的概率越大(第 8 行).这与 GEP 中的变异率策略以及变异后适应度的改变趋势是相同的.综上所述,种群在态势模型中的进化策略与种群在原搜索空间中的进化策略是相同的.

6.2 用动力学解释区间密度的合理性

动力学原理可以清晰地解释观察 2.观察 2 指出,区间密度反比于 *Hard*(P).设 M_i,M_j 是问题 P_i,P_j 的态势模型,因为在 M_i,M_j 中,若 *Density*(M_i)>*Density*(M_j),则有α(*Space_M*(P_i))>α(*Space_M*(P_j)),其中,*Space_M*(P)是问题 P 的态势模型空间,而前面已经验证,*Space_M*(P)是原搜索空间 *Space*(P)的保动映射空间,故有α(*Space*(P_i))>α(*Space*(P_j));又因为α反映了种群进化的快慢,故可推得 *Hard*(P_i)<*Hard*(P_j).综上所述,整个进化过程可作如下解释:因为区间密度正比于适应度增长系数α,α反比于 *Hard*(P),故有区间密度反比于 *Hard*(P).

7 实验和性能分析

实验中,用 GEP 求解的问题均取自文献[1].通过不同的参数设置,可以使求解问题的进化难度不同.第 7.1 节对调节每个问题进化难度相关的参数进行了分析.第 7.2 节验证了区间密度能够准确预测 GEP 求解问题的进化难度.第 7.3 节验证了实验中 6 个问题的态势模型是原问题搜索空间的一个保动映射空间.

7.1 问题难度的调节

(1) 三次多项式回归 $y=a^3+a^2+a+1$.该问题的求解目标是用 GEP 进化出与上式相同的一个表达式,下面的情况(a)、情况(b)两种情况给出了通过调节 GEP 的各种参数设置,而使得求解该问题的进化难度不同.值得指出的是,求解该问题的 GEP 个体中含有多个基因.图 6、图 7 表达了在每个个体含有的基因个数以及基因连接的关系不同时,GEP 的进化难度:

(a) 基因个数不同导致难度不同(用加连接基因).图 6 显示了在其他设置相同以及用加法进行基因连接的情况下,基因数量的不同导致进化难度的不同.当基因数量为 3 时,难度最小;当基因数为 1 时,难度最大.实验部分对基因个数为 3 和 1 得到的态势模型进行对比分析.

(b) 基因个数不同导致难度不同(用乘进行连接).图 7 显示了在其他设置相同以及用乘法进行基因连接的情况下,基因数量的不同导致进化难度的不同.当基因数量为 2 时,难度最小;当基因个数为 8 时,成功率为 0,难度显然最大.这主要是由于当基因个数为 8 时,整个基因的长度达到了 104(即 8×13=104,13 为单个基因长度),即使基因每一位只有两种变化情况,整个搜索空间也会达到 2^{104}.显然,要在如此大的搜索空间找到最优解几乎是不可能的,故成功率为 0.实验部分对基因个数为 2 和 8 的态势模型行对比分析.

(2) 奇校验:f_1=*odd*2(*a*,*b*);f_2=*odd*3(*a*,*b*,*c*);f_3=*odd*4(*a*,*b*,*c*,*d*).f_1~f_3 是符号回归函数中的奇校验函数,当输入参数包含奇数个 1 时,结果为 1;反之,结果为 0.极优个体是能够满足其真值表中所有组合的逻辑表达式,f_1~f_3 的真值表分别包含 2^2=4 个、2^3=8 个和 2^4=16 个训练数据.由于训练数据量依次增大,显然,它们的难度是依次增大的.实验部分将对 f_1~f_3 的态势模型进行对比分析.

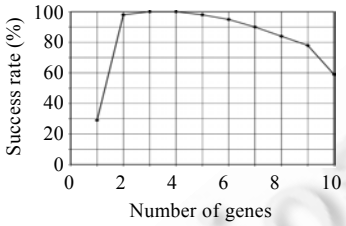


Fig.6 Different success rate with different number of genes (linked with plus)

图 6 基因个数不同导致进化难度不同(用加进行连接)

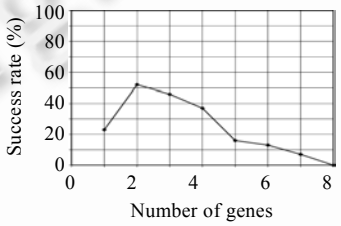


Fig.7 Different success rate with different number of genes (linked with plus)

图 7 基因个数不同导致进化难度不同(用乘进行连接)

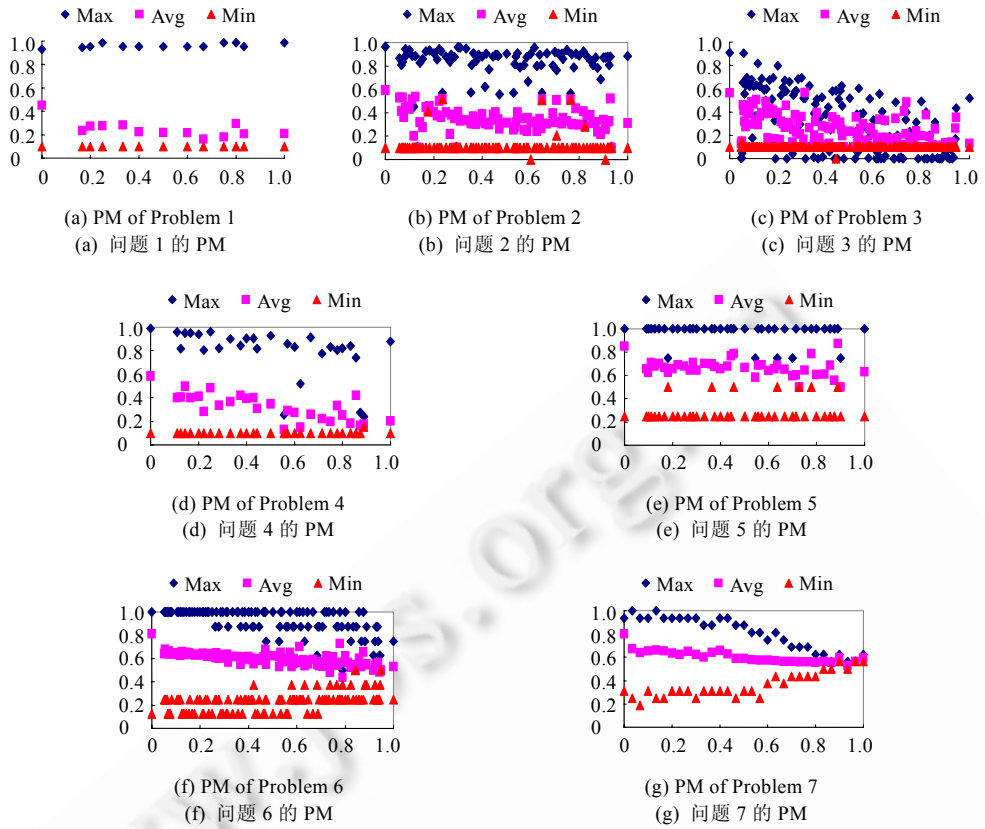
7.2 各问题的态势模型及区间密度

本节首先描绘各问题的态势模型(简称 PM),图 8 中纵坐标表示染色体的适应度,横坐标表示染色体到极优个体的距离,max,*avg*,min 分别表示到原点距离相同点的最高、平均和最低适应度.然后,计算出每个态势模型的区间密度.表 3 列出了与态势模型相关的实验数据,可知数据符合定义 3 的要求和观察 2 的结论,也即区间密度能够准确地预测 GEP 的进化难度.为了准确、简捷地描述,把本节的 7 个问题用问题 1~问题 7 表示:(1) 问题 1:三次多项式回归,3 个基因,用加连接基因;(2) 问题 2:三次多项式回归,1 个基因,用加连接基因;(3) 问题 3:三次多项式回归,8 个基因,用乘连接基因;(4) 问题 4:三次多项式回归,2 个基因,用乘连接基因;(5) 问题 5:f_1;(6) 问题 6:f_2;(7) 问题 7:f_3.

(1) 问题 1 和问题 2.直观来说,图 8(a)、图 8(b)的态势模型差距不是很大,由于各个区间均存在适应度很高的个体,体现出其成功率都在 30%以上,即进化难度都不大,表 3 中其区间密度差异也较小.在这两种情况下,其 FDC 差距很小.

(2) 问题 3 和问题 4.图 8(c)、图 8(d)的态势模型差距很大,这是由于在图 8(c)中,当区间离原点较远时,适应度高的个体几乎是呈线性地减少.体现为当基因个数为 8 时,成功率是 0,即进化难度非常大.而当基因数为 2 时,由于态势模型的各个区间均存在适应度很高的个体,体现出成功率大于 50%.这种明显的难度差距也清晰地反映在表 3 的区间密度上.同样地,在这两种不同情况下,其 FDC 值的差距很小,并且预测结果与实际情况相反.

(3) 奇校验.图 8(e)~图 8(g)分别是其态势模型,由表 3 所给出的区间密度中,知其区间密度下降得很快,体现出 f_1~f_3 的难度上升很快,说明区间密度能够很好地预测难度变化趋势.其 FDC 值差距很小,并且 f_1 和 f_2 预测结果与实际情况相反.



(a) PM of Problem 1
(a) 问题 1 的 PM

(b) PM of Problem 2
(b) 问题 2 的 PM

(c) PM of Problem 3
(c) 问题 3 的 PM

(d) PM of Problem 4
(d) 问题 4 的 PM

(e) PM of Problem 5
(e) 问题 5 的 PM

(f) PM of Problem 6
(f) 问题 6 的 PM

(g) PM of Problem 7
(g) 问题 7 的 PM

Fig.8 Posture models of Problem 1~Problem 7

图 8 问题 1~问题 7 的态势模型

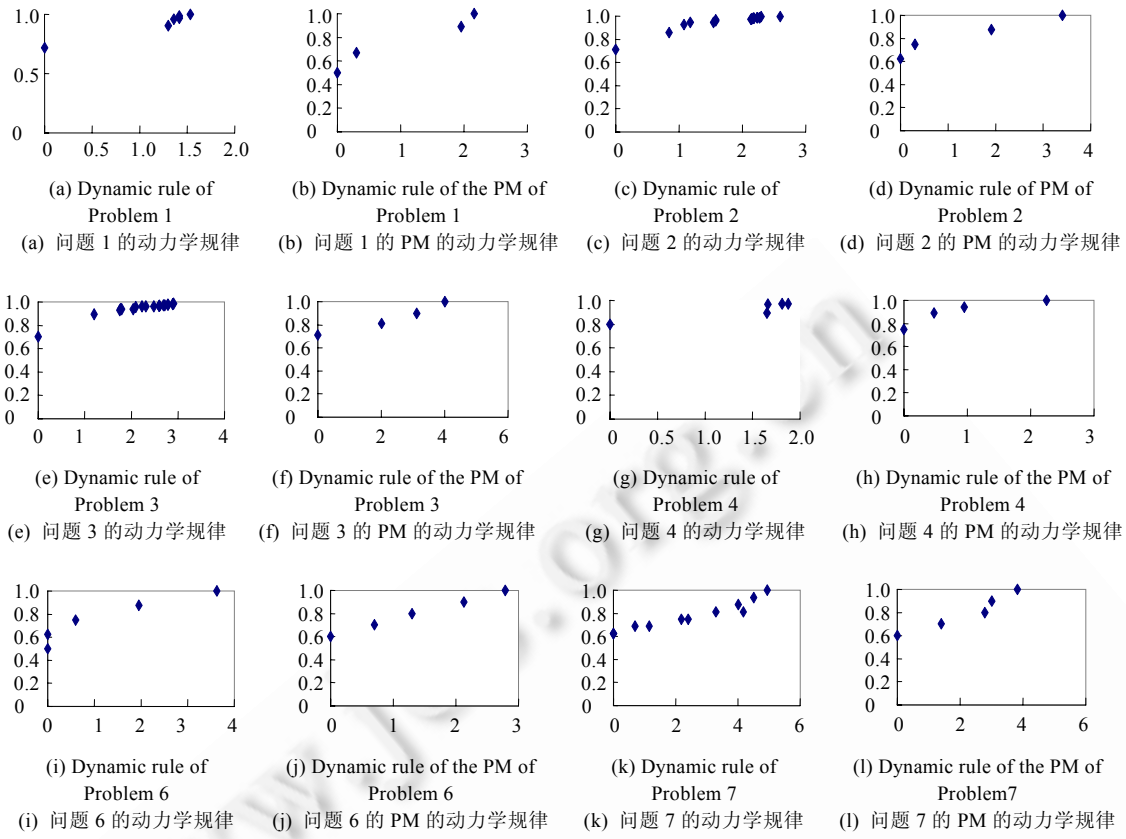
Table 3 Posture model parameters and evolutionary results of Problem 1~Problem 7

表 3 问题 1~问题 7 的态势模型参数及进化结果

Problem ID	1	2	3	4	*odd*2	*odd*3	*odd*4
Hard(*P*)	0.0	0.7	1.0	0.48	0.0	0.32	0.97
Cardinality of sample region	5 000	5 000	5 000	5 000	5 000	10^6	10^7
Region resolution	10	10	10	10	10	10	10
Region density	10	2.8	0.1	6.7	9.02	7.3	0.55
Threshold of local best point	0.9	0.9	0.9	0.9	1	1	1
FDC value	−0.23	−0.18	−0.27	−0.15	−0.35	−0.37	−0.31
Incresing index α	0.181	0.111	0.1	0.167	–	0.104	0.076
α of the simulating algorith *m*	0.227	0.113	0.083	0.136	–	0.143	0.105

7.3 动力学分析

本节的实验是验证对第 7.2 节中的问题 1~问题 4、问题 6、问题 7(由于第 5 个问题比较简单,故没有在实验中进行考察),其各自的态势模型空间(简记为 *Space_M*(*P*))是原 GEP 搜索空间(简记为 *Space*(*P*))的保动映射空间.即对任意 P_i,P_j,若存在关系$\alpha(Space(P_i))<\alpha(Space(P_j))$,则都有$\alpha(Space_M(P_i))<\alpha(Space_M(P_j))$.图 9 各图的横坐标表示群进化代数的对数值,纵坐标表示对应代中种群的极优适应度.适应度增长系数α就是各图的斜率.表 3 列出了$\alpha(Space(P))$和$\alpha(Space_M(P))$的关系,可知,数据符合定义 3 的要求和观察 2 的结论.



(a) Dynamic rule of Problem 1
(a) 问题 1 的动力学规律
(b) Dynamic rule of the PM of Problem 1
(b) 问题 1 的 PM 的动力学规律
(c) Dynamic rule of Problem 2
(c) 问题 2 的动力学规律
(d) Dynamic rule of PM of Problem 2
(d) 问题 2 的 PM 的动力学规律
(e) Dynamic rule of Problem 3
(e) 问题 3 的动力学规律
(f) Dynamic rule of the PM of Problem 3
(f) 问题 3 的 PM 的动力学规律
(g) Dynamic rule of Problem 4
(g) 问题 4 的动力学规律
(h) Dynamic rule of the PM of Problem 4
(h) 问题 4 的 PM 的动力学规律
(i) Dynamic rule of Problem 6
(i) 问题 6 的动力学规律
(j) Dynamic rule of the PM of Problem 6
(j) 问题 6 的 PM 的动力学规律
(k) Dynamic rule of Problem 7
(k) 问题 7 的动力学规律
(l) Dynamic rule of the PM of Problem7
(l) 问题 7 的 PM 的动力学规律

Fig.9 Dynamic rule
图 9 动力学规律

8 总结及未来工作

针对 FDC 能够准确地预测 GA,GP 的进化难度而无法预测 GEP 的进化难度的问题,提出了一种量化指标——态势模型的区间密度,能够准确地预测 GEP 求解特定问题的进化难度:(1) 分析了 FDC 的局限性.由于进化计算本质上是一种搜索过程,GA,GP 在解空间上的迁移能力较弱,若其优质解都分布在一个临近范围,则 GA,GP 能够较快地找到最优解,进化难度小.而这种优质解的分布特征能很好地被 FDC 所反映,故其进化难度能够准确地被 FDC 预测.而 GEP 个体的迁移能力很强,其更适合搜索优质解在整个搜索空间均匀分布的情况,但这种优质的分布特征是无法用 FDC 衡量的.(2) 提出了态势模型的区间密度指标.态势模型能够很好地反映具有不同适应度的个体在整个搜索空间的分布情况,而态势模型的区间密度能够很好地反映优质个体在整个搜索空间的分布情况.由于优质个体越多,分布越均匀,GEP 越能快速找到优质个体,而这种情况下的区间密度也越大,故能准确预测 GEP 的进化难度.(3) 从动力学角度保证了采用态势模型的区间密度预测 GEP 进化难度的正确性.由于 FDC 和区间密度都是从种群在搜索空间中的迁移能力出发的,故证明了种群在原搜索空间和在态势模型构成空间中的迁移具备相同的动力学性质.

下一步的工作主要有:(1) 由于种群在态势模型上的移动由变异算子引起,下一步将分析采用不同进化算子对种群进化难度的影响;(2) 由于 GEP 的参数设置对其进化性能有很大的影响,下一步将力图用 GEP 难度的衡量指标来优化其进化参数的设置.

References:

[1] Candia Ferreria Gene Expression Programming Mathematical Modeling by an Artificial Intelligence. Berlin: Springer-Verlag, 2006.

[2] Xu KK, Liu YT, Tang R, Zuo J, Zhu J, Tang CJ. A novel method for real parameter optimization based on gene expression programming. Journal of Applied Soft Computing, 2009,9(2):725−737. [doi: 10.1016/j.asoc.2008.09.007]

[3] Zhu MF, Tang CJ, Qiao SJ, Dai SC,Chen Y. Genetic neutrality in naive gene expression programming. In: Wang T, ed. Proc. of the Engineering Services and Knowledge Management. Berlin: Springer-Verlag, 2008. 1−4.

[4] Peng J, Tang CJ, Li C, Hu JJ. A new evolutionary algorithm based on chromosome hierarchy network. Int'l Journal of Computers and Applications, 2008,30(2):1−9.

[5] Duan L, Tang CJ, Zhu J, Zuo J, Liu YT, Wu J, Dai L. The strategies of initial diversity and dynamic mutation rate for gene expression programming. In: Huan Y, ed. Proc. of the Int'l Conf. on Natural Computation 2007, Vol.4. Berlin: Springer-Verlag, 2007. 265−269. [doi: 10.1109/ICNC. 2007.748]

[6] Kauffman S, Wrote; Li SM, Xu B, Trans. At Home in the Universe: The Search for the Laws of Self-Organization. Changsha: Hunan Science and Technology Press, 1995 (in Chinese).

[7] Jones T, Forrest S. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Eshelman L, ed. Proc. of the 6th Int'l Conf. on Genetic Algorithms. San Francisco: Morgan Kaufmann Publishers, 1995. 184−192.

[8] Smith T, Husbands P, Layzell P, O'Shea M. Fitness landscapes and evolvability. Evolutionary Computation, 2002,2(2):1−34. [doi: 10.1162/106365602317301754]

[9] Vanneschi L, Tomassini M. A study on fitness distance correlation as a difficulty measure in genetic programming. Evolutionary Computation, 2005,13(2):213−239. [doi: 10.1162/1063656054088549]

[10] Vanneschi L, Tomassini M, Collard P, Clergue M. Fitness distance correlation in structural mutation genetic programming. Genetic Programming, 2003,5(2):455−471. [doi: 10.1007/3-540-36599-0_43]

[11] Gustafson S, Vanneschi L. Operator-Based distance for genetic programming: Subtree crossover distance. In: Keijzer M, Tettamanzi A, Collet P, van Hemert J, Tomassini M, eds. Proc. of the 8th European Conf. on Genetic Programming. LNCS 3447, Lausanne: Springer-Verlag, 2005. 178−189. [doi: 10.1007/978-3-540-31989-4_16]

[12] Wineberg M, Oppacher F. Distance between populations. In: Proc. of the Genetic and Evolutionary Computation Conf. LNCS 2724, Berlin: Springer-Verlag, 2003. 1481−1492.

[13] Moraglio A, Poli R. Topological interpretation of crossover. Genetic Programming, 2004,12(1):1377−1388.

[14] Borenstein Y, Poli R. Fitness distributions and GA hardness. In: Yao X, *et al.*, eds. Proc. of the Parallel Problem Solving in Nature. LNCS 3242, Springer-Verlag, 2004. 11−20. http://springerlink.metapress.com/content/7w86kgy6bux98m08/fulltext.pdf [doi: 10.1007/978-3-540-30217-9_2]

[15] Vanneschi L, Tomassini M, Collard P, Vérel S. Negative slope: A measure to characterize genetic programming. In: Giacobini M, ed. Proc. of the Euro Genetic Programming, Vol.3905. Berlin: MIT, 2006. 178−189. [doi: 10.1007/11729976_16]

附中文参考文献:

[6] Kauffman S,著;李绍明,徐彬,译.宇宙为家.长沙:湖南科学技术出版社,2006.

附录. 用态势模型预测基因表达式编程进化难度的若干定理证明细节

定理 1 的证明细节.

定理 1. 定义 4 描述的树距离满足度量空间中距离的 4 项准则,即:

(1) 非负性:如果 $T_1 \neq T_2$,则 $d(T_1,T_2,m)>0$.

(2) 反身性:如果 $T_1=T_2$,则 $d(T_1,T_2,m)=0$.

(3) 对称性:$d(T_1,T_2,m)=d(T_2,T_1,m)$.

(4) 三角不等式:$d(T_1,T_2,m)+d(T_2,T_3,m) \geqslant d(T_1,T_3,m)$.

证明:准则(1)~准则(3)均很容易证明,这里略去.对于准则(4),设有 3 棵树 T_1,T_2,T_3,这里分两种情况进行证明:

(1) 如果 $n(T_1)=n(T_2)=n(T_3)$,则有公式(1).通过公式(1),易证 $d(T_1,T_2,m-1)+d(T_2,T_3,m-1)\geqslant d(T_1,T_3,m-1)$.故可以得到公式(2),而公式(2)实际上表达的就是 $d(T_1,T_2,m)+d(T_2,T_3,m)\geqslant d(T_1,T_3,m)$,故第 1 种情况得证.

(2) 如果 $n(T_1)\neq n(T_2)\neq n(T_3)$,不失一般性,设 $n(T_1)\geqslant n(T_2)\geqslant n(T_3)$,可以得到公式(3),同理可得公式(4)和公式(5).由于 $n(T_1)\geqslant n(T_2)\geqslant n(T_3)$,故有公式(6)和公式(7).将公式(6)、公式(7)相加,第 2 种情况即得证.

$$\begin{aligned} d(T_1,T_2,m)+d(T_2,T_3,m) &= \sum_{i=1}^{n(T_1)}\frac{d(s_i(T_1),s_i(T_2),m-1)}{n(T_1)}+\sum_{i=1}^{n(T_1)}\frac{d(s_i(T_2),s_i(T_3),m-1)}{n(T_1)} \\ &= \sum_{i=1}^{n(T_1)}\left(\frac{d(s_i(T_1),s_i(T_2),m-1)}{n(T_1)}+\frac{d(s_i(T_2),s_i(T_3),m-1)}{n(T_1)}\right) \end{aligned} \tag{1}$$

$$\sum_{i=1}^{n(T_1)}\left(\frac{d(s_i(T_1),s_i(T_2),m-1)}{n(T_1)}+\frac{d(s_i(T_2),s_i(T_3),m-1)}{n(T_1)}\right)\geqslant\sum_{i=1}^{n(T_1)}\frac{d(s_i(T_1),s_3(T_2),m-1)}{n(T_1)} \tag{2}$$

$$d(T_1,T_2,m)=\sum_{i=1}^{n(T_2)}\frac{d(s_i(T_1),s_i(T_2),m-1)}{n(T_2)}=\sum_{i=1}^{n(T_1)}\frac{d(s_i(T_1),s_i(T_2),m-1)}{n(T_2)}+\sum_{i=1}^{n(T_2)-n(T_1)}\frac{m-1}{n(T_2)} \tag{3}$$

$$d(T_2,T_3,m)=\sum_{i=1}^{n(T_2)}\frac{d(s_i(T_2),s_i(T_3),m-1)}{n(T_3)}+\sum_{i=1}^{n(T_3)-n(T_2)}\frac{m-1}{n(T_3)} \tag{4}$$

$$d(T_1,T_3,m)=\sum_{i=1}^{n(T_1)}\frac{d(s_i(T_1),s_i(T_3),m-1)}{n(T_3)}+\sum_{i=1}^{n(T_3)-n(T_1)}\frac{m-1}{n(T_3)} \tag{5}$$

$$\sum_{i=1}^{n(T_1)}\frac{d(s_i(T_1),s_i(T_2),m-1)}{n(T_2)}+\sum_{i=1}^{n(T_2)}\frac{d(s_i(T_2),s_i(T_3),m-1)}{n(T_3)}\geqslant\sum_{i=1}^{n(T_1)}\frac{d(s_i(T_1),s_i(T_3),m-1)}{n(T_3)} \tag{6}$$

$$\sum_{i=1}^{n(T_2)-n(T_1)}\frac{m-1}{n(T_2)}+\sum_{i=1}^{n(T_3)-n(T_2)}\frac{m-1}{n(T_3)}\geqslant\sum_{i=1}^{n(T_3)-n(T_1)}\frac{m-1}{n(T_3)} \tag{7}$$

证明完毕. □

定理 3 的证明细节.

定理 3. 设样本集 S 有 k 个不同的极优个体,则 S 有 k 种态势模型,分别为 $M_1,M_2,...,M_k$,设每个模型的区间分辨率均为 $1/d$,则各态势模型的区间密度取下列值的概率最大:

(1) 当 $d=1$ 时,$Density(M_1)=Density(M_2)=...=Density(M_k)=C_1\times N\_global(S)/(C_1\times N\_global(S)+C_2\times N\_local(S))$;

(2) 当 $d>1$ 时,$Density(M_1)=Density(M_2)=...=Density(M_k)=d\times C_1\times N\_global(S)/(C_1\times N\_global(S)+C_2\times N\_local(S))$.

证明:

(1) 当 $d=1$ 时,由于只有 1 个子区间,由假设(1)得 $Den(M_i)=C_1\times N\_global(S)/(C_1\times N\_global(S)+C_2\times N\_local(S))$.

(2) 当 $d>1$ 时,对任意态势模型 M_i 有公式(8)~公式(10).公式(10)表达了从 M_i 变到 M_j 时,子区间 s_k 中极优个体数量的变化情况.根据假设(2),任意点 q 移动到另外各个子区间的概率均为 $1/d$,则 $N\_global(s_k,M_i)\times 1/d$ 表示原来在 s_k 中的点只剩下 $1/d$, $\sum_{r=1,r\neq k}^{d}N\_global(s_r,M_i)\times 1/d$ 表示其他子区间都可能有 $1/d$ 的点转移到 s_k 中.同理,对 M_j 有公式(11).于是,公式(9)去掉求和符号后可变为公式(12),而公式(12)经变形可依次得到公式(13)、公式(14).最后,根据公式(7),命题得证.

$$P(M_i)=\sum_{k=1}^{d}\frac{N\_global(s_k,M_i)}{N\_global(s_k,M_i)+N\_local(s_k,M_i)} \tag{8}$$

$$P(M_j)=\sum_{k=1}^{d}\frac{N\_global(s_k,M_j)}{N\_global(s_k,M_j)+N\_local(s_k,M_j)} \tag{9}$$

$$N\_global(s_k,M_j)=N\_global(s_k,M_j)\times 1/d+\sum_{r=1,r\neq k}^{d}N\_global(s_r,M_i)\times 1/d \tag{10}$$

$$N\_local(s_k,M_j)=N\_local(s_k,M_j)\times 1/d+\sum_{r=1,r\neq k}^{d}N\_local(s_r,M_i)\times 1/d \tag{11}$$

$$P(M_j)=d\times\frac{N\_Global(M_i)\times 1/N\times 1/d}{N\_Global(M_i)\times 1/N\times 1/d+N\_local(M_i)\times 1/N\times 1/d} \tag{12}$$

$$P(M_j)=d\times\frac{C_1\times N\_Global(S)\times 1/N\times 1/d}{C_1\times N\_Global(S)\times 1/N\times 1/d+C_2\times N\_local(S)\times 1/N\times 1/d} \tag{13}$$

$$P(M_j)=d\times\frac{C_1\times N\_Global(S)}{C_1\times N\_Global(S)+C_2\times N\_local(S)} \tag{14}$$

证明完毕. □

郑皎凌(1981－),女,重庆人,博士,讲师,CCF 会员,主要研究领域为数据库与知识工程,机器 学习.

杨宁(1974－),男,博士,讲师,主要研究领域为数据库与知识工程,数据流挖掘.

唐常杰(1946－),男,教授,博士生导师,主要研究领域为数据库与知识工程,数据挖掘.

段磊(1981－),男,博士,讲师,主要研究领域为数据库与知识工程,基因表达式编程.

徐开阔(1983－),男,博士,讲师,主要研究领域为数据库与知识工程,进化计算.

李红军(1977－),男,博士生,讲师,主要研究领域为数据库与知识工程,基因表达式编程.