计算机科学 1999Vol. 26№. 7

遗传程序设计模式理论新进展*)

New Advances in the Genetic Programming Scheme Theorem

陈定君+ 周志强++ 刘积仁+

(东北大学软件中心新技术部 沈阳 110006)+ (重庆邮电学院计算机系 重庆 400065)++

Abstract Genetic Programming, an automatic programming technology based on evolutionary computing, is recently growing up. GP scheme theorem (GPST) is the basis of interpreting GP's search behavior. This paper gives full summarization of the advances in GP schema theorem in recent years, mainly discusses the GPST's schema definition in subspaces of programs of different sizes and shapes, and in subspaces of programs of fixed size and shape, some differences between them are described briefly.

Keywords Genetic programming, Building block hypothesis, Crossover operator, Mutation

1. 引言

在模式识别、细胞自动控制、太空卫星控制、机器人控制、电路设计等领域,遗传程序设计已成功地解决许多难以解决的难题[1,2,3],然而,用于解释它的运行机制的理论却相当少。自 John Holland 在 70 年代中期提出了其著名的模式定理以来,模式定理就一直作为解释遗传算法 GA (Genetic Algorithm) 工作机制的理论基础。GA 采用确定长度的染色体编码方案,GP 通常是使用规模和形状能够动态变化的不确定分层计算机程序。二者操作的类似性和差异性促使沿着 GA 理论方法形成 GP 的理论,以便寻找遗传程序设计在进化过程中可能蕴藏某种内在的规律。目前,建立 GP 模式定理最自然方式就是定义模式概念来扩充 GA 模式定理,探讨新的模式定理,在选择,交叉和变异等算子作用下,是如何生成新一代的[4]。

获得 GP 理论的困难之一是 GP 模式定理的定义比 GA 的模式定理更缺少直观性。在早期模式定义中,模式被定义成没有根结点的树,这样在同一个程序内,模式可以以多种方式来表示。随着匹配同一模式的程序规模和形状的变化,使模式遭破坏的概率计算是非常复杂的,这也充分说明有必要建立新的 GP 模式理论[5]。在近来的模式定义中,模式被表示成有根结点的树片段(tree fragment),这种定义使模式理论计算稍容易些[6]。而且,这些模式通常将程序空间划分成两种类型完全不同的子空间,动态可变化的与不可变化的规模和形状的程序子空间。本文主要介绍这两种不同程序子空间的模式理论的最新进展,并对它们之间不同之处作了简要的描述。

2. 变长 GP 模式定理的进展

2.1 Koza 的 GP 模式理论

Holland[1975]的 GA 模式是通过定义它所包含的比特和比特的位置来确定的,因此,GA 模式是一个双元集,GA 模式定理常常用来描述群体内某些比特流(模式中构件 component)的实例数变化,而不是用来描述由模式表示程序子空间的串样本数变化[7]。

由于 Holland 模式定义十分流行,Koza[1994]提出了一些非正式的想法,申明 Holland 的模式理论稍作修改同样适用于 GP。Koza 是第一个大胆作这种尝试的人,他的观点是基于把模式定义为树的子空间这一思想,这些树是事先包含了预定义的子树集合。Koza 的模式 H 能够表示为 S-表达式集合,Koza 的模式定义仅仅给出了模式中定义构件的有

*)863 计划资助、重庆市应用基础基金资助。陈定君 讲师,主要研究方向:嵌入式软件仿真开发、分布式计算,遗传程序设计。周志强 讲师,主要研究方向:进化计算,进化数字电路设计。刘积仁 教授,博士生导师,主要研究方向:分布式多媒体,协议工程。

关信息，却没有指定构件的位置信息，所以同一模式在同一程序中能够以不同方式实例化多次[4]。

2.2 O'Reilly 的 GP 模式理论

O'Reilly[1995]在 Koza 的模式定义基础上作了进一步的研究，推导出基于适应值比例的选择和交叉的 GP 模式理论[6]，O'Reilly 的模式理论不包含变异对进化所产生的影响。

定义 1 模式 H 是一个双元集，一元是一个唯一具有共同特征的 LISP S-表达式树或一些叶子作为通配符的不完全 S-表达式树的集合，另一元为一个相应整数，指明前一元的数量。

定义 2 GP 模式 H 的阶是在树中对应于 S-表达式中已确定的结点数，用符号 $\|H\|$ 表示。

定义 3 GP 模式 H 固定长度 $D_{fixed}(H)$ 是在树中每一个 S-表达式中不包含与通配符相连的边的边总数。

定义 4 GP 模式 H 可变长度 $D_{var}(h,H)$ 是树中与 S-表达式连在一起的边数（h 表示对应的某一程序）。

定义 5 GP 模式定义长度 D 是它的可变长度和固定长度的和，即：

$$D(inst(h,H),H)=D_{fixed}(H)+D_{var}(inst(h,H),H) \tag{1}$$

其中 $inst(h,H)$ 表示程序 h 中模式 H 的实例化集合。

当在 h 中一个结点被选择为交叉点并与来自另一程序子树在子树根结点处交换后，原有 GP 模式 H 实例化集 $inst(h,H)$ 遭破坏，以致它不再属于模式 H 的一个实例，这种遭破坏的概率上限可由下式来表示：

$$P_d(H,h,t)=\frac{D_{fixed}(H)+D_{var}(h,H)}{Size(h)} \tag{2}$$

其中 $Size(h)$ 表示程序 h 中可用的交叉位置数量。这样，很容易得到模式 H 生存概率下限 $P_s(H,h,t)=1-P_d(H,h,t)$

O'Reilly 模式的定义长度不是一个常量，它依赖于该模式在程序取样中被示例化的方式。另外，在每个树中总的连接数是可变的，这暗示了由于交叉使模式 H 遭破坏的概率 $P_d(H,h,t)$ 依赖于与模式相匹配的树 h 的形状、规模和成分。为了克服这个困难，O'Reilly 推导的模式定理中 $P_d(H,h,t)$ 取其最大值，即 $P_d(H,t)=\max\{P_d(H,h,t)\}$，O'Reilly 模式定理如下：

$$E[\alpha(H,t+1)]\geqslant\alpha(H,t)\cdot\frac{f(H,t)}{\overline{f(t)}}\cdot(1-P_cP_d(H,t)) \tag{3}$$

其中，$\alpha(H,t)$ 是在 t 代模式 H 的实例数，$f(H,t)$ 是 t 代 H 实例的平均适应值，$\overline{f(t)}$ 是 t 代群体的平均适应值，P_c 是交叉概率，$E[\alpha(H,t+1)]$ 表示模式 H 在 $t+1$ 代模式 H 期望实例数，

O'Reilly 模式定理描述的是一个模式所表示的构件从当前代到下一代的繁殖方式，而不是一给定模式的程序样本数在当前代到下一代的改变方式。O'Reilly 的模式定理揭露了为什么对实际繁殖没有作出假设和在 GP 中使用基因块（building blocks）（短的、低价的、适应值相当高的模式）这一概念就在于 $P_d(H,t)$ 的内在可变性。

2.3 Rosca 的 GP 模式理论

Rosca[1997]在 O'Reilly 所做的工作基础上，对 O'Reilly 模式理论进行部分修正与扩充，提炼并归纳出更通用的新的模式定义[5]，其定义如下：

定义 6 一个有根结点的 K 阶模式是一 K 个函数和端点标识来指定的有根结点彼此邻近的树片段。

这个定义利用了程序表示的层次化本性，建设性地指明了模式表示和由模式定义的程序子集间的相互关系。考虑某一 Rosca GP 模式 H 和与 H 匹配的子群体，$\|H\|$ 表示模式的阶，$m(H,t)$ 表示在 t 代与模式 H 所能匹配的实例数，$f_H(t)$ 表示 t 代群体中与模式 H 匹配的所有树的平均适应值，$f(t)$ 表示 t 代群体平均适应值，交叉和变异概率分别为 P_c 和 P_m，并且令 $P_d=P_c+P_m$，s 表示整个群体的平均复杂度，s_H 表示模式 H 树结构的复杂度，M 表示群体规模。Rosca GP 模式定理结果如下：

$$m(H,t+1)\geqslant m(H,t)\times\frac{f_H(t)}{f(t)}\times\left[1-\left(\sum_{h\in H}\left(P_d\times\frac{f_h(t)}{s_h}\right)\right)/\sum_{h\in H}\frac{f_h(t)}{\|H\|}\right] \tag{4}$$

Rosca 并没有给出模式定义长度的定义，但考虑了变异对进化所产生的影响。另外，Rosca 给出了其基因块假设 BBH 成立的必要条件。

$$m(H,t)\geqslant\theta(H,f,t) \tag{5}$$

其中 $\theta(H,f,t)=$

$$\left(\|H\|\times\sum_{h\in H}\left(P_d\times\frac{f_h(t)}{s_h}\right)\right)/(f_H(t)-f(t))$$

对于一个给定的模式 H，$\|H\|$ 是一定值，P_d 是一常量，则有：

$$\theta(H,f,t)=\left(\|H\|\times P_d\times\left(\frac{f(t)}{s}\right)_H\right)/(f_h(t)-f(t)) \tag{6}$$

这样，只要 $m(H,t)$ 大于或等于阈值 $\theta=\theta(H,f,t)$，那么模式 H 的实例数将会递增。

由(6)式，不难看出，通过增加个体 $h\in H$ 的适应值或减少它的复杂度 s_h，均可使阈值 θ 减少，从而增加了模式 H 的生存机率。

GP 模式定理描述的是在下一代模式 H 实例数增加或减少的行为，并没有指出在多代反复迭代情形下，GP 模式定理是否仍然成立。根据 GP 模式定理，模式 H 的增长是依赖于繁殖因子，繁殖因子可由下式表示：

$$\text{繁殖因子}=\frac{f_H(t)}{f(t)}\times[1-P_d(H,t)]$$

其中 $P_d(H,t)=$

$$\left(\sum_{h\in H}\left(P_d\cdot\frac{f_h(t)}{s_h}\right)\right)/\sum_{h\in H}\frac{f_h(t)}{\|H\|}$$

在繁殖因子$\geqslant 1$ 情形下，在下一代中模式 H 实例数才有可能增加。然而，如果在繁衍过程中，出现后续几代繁殖因子保持不变情况的话，即有：

$$\frac{f_H(t)}{f(t)}\times[1-P_d(H,t)]=\frac{f_H(t+1)}{f(t+1)}\times[1-P_d(H,t+1)] \tag{7}$$

进一步化简：

$$\Delta\log\frac{f_H(t)}{f(t)}=-\Delta\log(1-P_d(H,t)) \tag{8}$$

那么基因块假设 BBH 将不会成立，这种可怕情形是可能存在的。更糟糕的情况有可能发生，由于被评估的模式适应值相对于群体适应值和模式的遭破坏的概率最大上限是随进化过程不断改变，在程序进化一段时期后，某些基因块可能根本不存在[16]。

另外，在 GP 初始化群体时，程序树的规模和最大深度被设为较低的值，在进化过程中，程序树规模及深度不断增大，在每一代中程序树的增大可能会引起因遗传操作而导致遭破坏概率的减少，这种无法控制程序规模过程将会使程序进化不稳定。

3. 定长 GP 模式定理的进展

Riccardo Poli 和 W. B Langdon[1997]提出了固定规模和形状的程序子空间模式定义，并给出了基于一点交叉和变异的模式定理，该模式定理与遗传算法 GA 模式的原始概念较接近[11]。

定义 7 如果 F 和 T 是 GP 运行中所使用的函数集和端点集，那么模式是由来自 $F\cup T\cup\{=\}$ 集合中函数集和端点集构成的一个树(或 S-表达式)。

算符"="是一个通配符，与前文所述的通配符"#"类似，但稍有不同，"="仅表示一单个函数或端点，"#"可用作表示由多个函数或端点构成的整个子树。

定义 8 阶是一个模式 H 中非"="符号的数量，记作 $Q(H)$。

定义 9 模式 H 长度是模式 H 中结点的总数，记作 $N(H)$。

定义 10 模式 H 的定义长度是模式 H 中包含所有非"="符号的最小树片段的连接数，记作 $L(H)$。

与传统二进制 GA 类似，GS 模式的阶、长度和定长与实际群体中程序的形状和规模无关。低阶且大长度的模式能够表示较大数量的程序。如果函数集和端点集是有限的，那么定长 GP 模式所表示的程序数量也是有限的，且很容易计算出来。而变长的模式如不加约束所表示的程序却是无穷多，尽管在程序的深度或规模被指定为有限的约束条件下，模式所表示的程序总数是有限的，但也难以用一种简单相近的形式来表示。

从某种程度上讲，定长的模式定义比变长的模式定义较低级，因为如果采用 O'Reilly 或 Rosca 的模式定义所表示的一个较小数量的树能够用相同数的通配符来表示，那么这个模式就可以由定长的模式定义来表示，反过来却不行。

定长的模式定义，与原始 GA 的模式定义一样，提供模式中构件的位置信息，这样，在群体中一个模式的实例数与从模式中取样的程序数是一致的，而变长的模式定义是不提供构件位置信息的。

为了能够更好地理解一点交叉对模式的影响，Riccardo Poli 和 W. B. Langdon 给出了模式的几何解释定义。

定义 11 如果模式 G 中不含有任何确定的结点，即 $O(G)=0$，那么模式 G 就是一个超空间(hyperspace)。

定义 12 如果模式 H 中至少含有一个确定的结点，那么模式 H 就是一个超平面(hyperplane)。

定义 13 如果用通配符替换超平面模式 H 中所有确定结点，那么 $G(H)$ 就表示与 H 相关的超空间。

每个超空间表示一给定形状的所有程序，如果对程序规模和形状不加约束，那么将会有无穷多超空间。然而，函数 F 和端点 T 是有限的，属于每个超空间的程序数是一定的。超平面是超空间程序中一个较小的子集。

在适应值比例选择作用下，假定采用赌轮(roulette-wheel)算法建立"交配池"，$\{h\in H\}$ 表示从模式 H 中抽取程序 h 实例来形成"交配池"事件，那么事件的发生概率为 $P_r\{h\in H\}$，其中，$m(H,t)$ 表示在 t 代与模式 H 所能匹配的实例数，$f(H,t)$ 表示在 t 代群体中与模式 H 匹配的所有树的平均适应

值，$\overline{f(t)}$表示 t 代群体平均适应值，M 表示群体规模。$D_c(H)$表示当从模式 H 中取样的程序 h 与程序 $\bar{h}$交叉后模式 H 遭破坏的事件。$G(H)$表示与 H 相关的超空间，令 $p_{diff}(t)=P_c\{D_c(H)\mid \bar{h}\in G(H)\}$，Riccardo Poli 和 W. B. Langdon[1997]的模式定理如下：

$$E[m(H,t+1)]\geqslant m(H,t)f(H,t)/\overline{f(t)}\cdot(1-p_m)^{O(H)}\cdot\left\{1-p_c\left[p_{diff}(t)\left(1-\frac{m(G(H),t)f(G(H),t)}{M\overline{f(t)}}\right)+\frac{L(H)}{(N(H)-1)}\frac{m(G(H),t)f(G(H),t)-m(H,t)f(H,t)}{M\overline{f(t)}}\right]\right\} \tag{9}$$

在 GP 运行的早期阶段，在交叉作用下，树 h 和不同形状树$\bar{h}$以完全相同的形状交换子树是不太可能的，这是因为初始化群体中多样性，仅有少量的个体有包含端点在内的共同部分，所以有 $p_{diff}(t)\approx 1$，$m(G(H),t)f(G(H),t)$一项表示含有结构 $G(H)$程序的总适应值。$M\overline{f(t)}$表示群体总适应值。在运行开始，有 $m(G(H),t)f(G(H),t)\ll M\overline{f(t)}$，模式遭破坏的概率接近于 1。所以，在运行开始，交叉算子完全可以抵消选择算子的作用。

在 GP 运行的后期阶段，越来越类似 GA 的运行。如果变异率较小，基于一点交叉的 GP 群体将开始收敛，规模和形状的多样性将会减少。在交叉作用下，程序之间的共同部分将会增加，程序将包含越来越多的端点。当和不同形状的程序交叉时，用相同结构交换子树的概率将会增加，$p_{diff}(t)$将会减少。随着多样性逐渐丢失，$p_{diff}(t)$将会变得更小。此时，$m(G(H),t)=M$，$f(G(H),t)=\overline{f(t)}$，Riccardo Poli 和 W. B. Langdon 的 GP 模式定理与 John Holland 的 GA 模式定理，式(10)，完全相同。

$$E[m(H,t+1)]\geqslant m(H,t)\cdot\frac{f(H,t)}{\overline{f(t)}}\cdot(1-p_m)^{O(H)}\cdot\left[1-p_c\frac{L(H)}{N-1}\left(1-\frac{m(H,t)f(H,t)}{M\overline{f(t)}}\right)\right] \tag{10}$$

即有：$\lim_{t\to\infty}GP_{1,p_c}(t)=GA$ (11)

结束语 变长模式定理和定长模式定理均能在一定程度上解释 GP 运行内部机制，即：

在遗传算子作用下，具有低阶、短定义长度以及平均适应值高于群体平均适应值的模式在子代中将得以指数增长。

一般来说，对变长模式定理，无论是 Koza 和 O'Reilly 的模式定理，还是 Rosca 的，三者都难以很好地支持基因块假设 BBH。然而，Riccardo Poli 和 W. B. Langdon 的定长模式定理却能较好地支持 BBH。需要进一步研究用表示超空间和在超空间内表示超平面的两种基因块来解释 GP 是如何构造其解这一难题。

Riccardo Poli 和 W. B. Langdon 的 GP 模式定理是 Holland 的 GA 模式定理的一种更通用的表达，实现了由 GA 到 GP 模式理论的平滑过渡。当 t →∞时，定长 GP 模式定理变为 GA 模式定理。作者认为，定长模式定理比变长模式定理更能较好地反映遗传程序设计 GP 的内部运行机制。

参 考 文 献

1 Scott B. Using genetic programming to evolve recursive programs for tree search. In: Proc. of the Fourth Golden West Conference on intelligent systems. Raleight, NC: International Society for Computers and Their Applications, 1995. 60～65

2 Koza John R, et al. Automated design of circuits using genetic programming. In: Gero John S., Sudweeks Fay edits. Artificial Intelligence in Design' 96. Dordrecht, Kluwer. 1996. 151～170

3 Angeline P J. Genetic programming and emergent intelligence. In: Kenneth E. edits. Advances in Genetic Programming 1. Chapter 4. MIT Press. 1994

4 Langdon William B. A bibliography for genetic programming. In: Peter J. Angeline edits. Advances in Genetic Programming 2, chapter B, Cambridge, MA, USA: MIT Press. 1996. 507～532

5 Altenberg L. The schema theorem and Price's theorem. In: L. D. Whitley edits. Foundations of Genetic Algorithms. Volume 3. San Mateo. CA: Morgan Kaufmann 1995

6 O'Reilly Una-May. An Analysis of Genetic Programming: [PH. D thesis]: Carleton University. Ottawa. 1995

7 De Jong K A. An Analysis of the Behavior of a Class of Genetic Adaptive Systems: [PH. D thesis]. University of Michigan. 1975

8 刘勇. 非数值并行算法—遗传算法. 科学出版社出版, 1995

9 Rosca J. Hierarchical Learning with Procedural Abstraction Mechanisms: [PH. D thesis]: University of Rochester 1997

10 O'Reilly Una-May, Oppacher F. The troubling aspects of a building block hypothesis for genetic programming. In: L. Darrel Whitley edits, Foundations of Genetic Algorithms 3, Estes Park, Colorado, USA: Morgan Kaufmann. 1994—1995, 73～78

11 Poli R, Langdon W B. A New Schema Theorem for Genentic Programming with One-point Crossover and Point Mutation. [Technical Report, CSRP-97-3], The University of Birmingham, UK 1997