

38-42

遗传算法

逻辑结构

全局收敛性

人工智能

TJ

⑧

计算机科学 1998Vol. 25No. 5

广义遗传算法的逻辑结构及全局收敛性的证明^{*}

Logic Structure of Generalized Genetic Algorithm and A Proof of Its Global Convergence

董 聪 郭晓华

0242-23

TP18

(清华大学 北京 100084) (中科院模式识别国家重点实验室 北京 100080)

摘 要 Generalized genetic algorithm is based on some famous modern biologics theories such as the genetic theory by Morgan, the general system theory by Bertalanffy, and et al, so it is superior in biologics to classical genetic algorithm. In this paper, the logic structure of the algorithm is presented, and its biological mechanism is revealed from a view points of micro-and-macro interaction. Besides, a exact proof is made on the ability of realizing global optimization in finite time of the algorithm.

关键词 Generalized genetic algorithm, Genetic theory, General system theory, Logic structure, Global optimization

20 世纪 70 年代前后, 北美和欧洲的一些科学家开始探索用模拟生物进化的方式求解复杂优化问题。其中, 以 Holland 提出的遗传算法 (genetic algorithm, GA) 最具代表性^[1]。80 年代以来, Hopfield 解决了反馈型人工神经网络的迭代稳定性问题, Rumelhart 等人重新发现了 BP 算法^[2]; 董聪等人对 BP 算法做了推广, 揭示了多层前向网络的逼近机理, 建立了相应的拓扑结构学习算法^[3,4]……。人工神经网络 (ANN) 在探索智能形成机制方面所取得的成就给人以启示: 研究智能的最好方式是向人类自身学习, 在几十亿年的进化过程中, 生物体形成了一种优化自身结构以适应环境变化的内在机制。ANN 对 AI 的主要贡献是解决了智能系统如何从环境中自主学习的问题, GA 的新一代支持者则试图揭示学习过程在基因层次上究竟是如何实现的。在生物基础方面, 经典遗传算法以达尔文的进化论 [1859] 和门德尔的遗传因子理论 [1865] 为依据; 在算法实现方面, 它同时具备了结构上的隐含并行性、计算原理上的随机性和自适应性。正是这些优良特性的组合使得遗传算法的研究和拓展迅速成为国际学术界和工程界关注的热点^[5-9]。在文献 [7] 中, 董聪对经典遗传算法的结构和操作系统进行了系统的剖析, 纠正了作为经典遗传算法数学基础的 Schema

定理中的一些错误, 发现并证明了广义 Schema 定理。通过引入一些新的概念和算子, 建立了广义遗传算法的理论体系和算法结构; 在文献 [8] 中, 则进一步阐述了广义遗传算法的数学结构, 并在宏观上探讨了它的生物基础, 通过在目前世界上单跨最长的铁路、公路两用桥——香港青马大桥传感器群最优布点设计方面的成功应用, 验证了广义遗传算法的单调收敛性和所具备的实现全局最优化的功能^[9]。

广义遗传算法以 Morgan 的基因理论 [1920] 及 Gould 与 Eldredge 的间断平衡理论 [1972] 为依据, 同时融合了 Mayr 的边缘物种形成理论和 Bertalanffy 一般系统理论的一些思想, 这是其在生物学原理上优于经典遗传算法的本质所在。本文重点阐述了广义遗传算法的逻辑结构, 并从宏观和微观协同作用机制上进一步明确了广义遗传算法的生物原理, 同时对该算法在有限步内找到问题全局最优解的能力做了严格的证明。

1 算法的逻辑结构

求解以下非线性优化问题:

$$\max f(X), X = (x_1, x_2, \dots, x_n) \in K \subset R^n \quad (1)$$

其中, K 为因子集。定义 $S = x_1 x_2 \dots x_n$ 为串, S 中的任一固定段为模式 (Schema), x_i 为染色体, x_i 中的

^{*} 国家自然科学基金、航空基金和攀登计划重大项目资助课题。

固定为基因, $f(X)$ 为适应值函数。从形式上看, 串 S 对应于 Morgan 基因理论中的基因型 (genotype), 适应值函数 $f(X)$ 对应表现型 (phenotype), Schema 对应于基因族 (gene cluster) 或基因组 (genome)。采用上述符号, 广义 Schema 定理可表示为^[7,8]:

$$m(H, k+1) \geq m(H, k) \frac{\bar{f}(H)}{\bar{f}} \left(1 - P_c(K) \frac{d(H)}{l-1}\right)^{O_c(H)} \left(1 - P_m(K) \frac{d(H)}{l}\right)^{O_m(H)} \quad (2)$$

其中, $m(H, k)$ 为第 k 代种群中含有给定模式 H 的串的总数; $\bar{f}(H)$ 为种群中含有给定模式 H 的串的平均适应值; $\bar{f}$ 为种群的平均适应值; $d(H)$ 为模式 H 的长度 (模式 H 最外侧的两个给定基因间的距离); $d(H)$ 为模式 H 所包含的基因总数; l 为串的长度; P_c 为交换操作概率; $O_c(H)$ 为交换阶次; P_m 为突变操作概率; $O_m(H)$ 为突变阶次。

达尔文进化论的核心在于自然选择, 而自然选择的基本素材是生物界普遍存在着的变异现象。达尔文认为, 可遗传的变异是实现定向进化的前提。从形式上看, 广义 Schema 定理没有说明有利的变异在进化过程中如何通过选择进行累积, 变异 (基因交换和突变) 在其中的作用似乎只是破坏性的, 而不是它本应具有的积极作用。另外, 受达尔文进化论的影响, 经典遗传算法更强调选择的重要性, 而忽略了突变在进化, 特别是在定向进化过程中的积极作用。广义 Schema 定理证明了可通过选择使优良个体的繁衍机会增加, 从而有利于实现局部最优化; 但对交换和突变在确保解空间各点可达性方面的作用, 却未能提供任何有利的依据。换句话说, 无论从经典达尔文进化论的角度还是从其他广泛的角度来看, 经典遗传算法在逻辑结构上均存在缺陷。

与 Holland 以归纳逻辑的方式从达尔文的进化论出发建立经典遗传算法的思路不同, 广义遗传算法是通过演绎逻辑的方式建立起来的: 首先建立一个逻辑上合理的算法框架, 然后证明这种框架在数学上是高效的, 最后是检验该算法是否为生物系统所采用。

从逻辑上讲, 一种算法要想具备实现全局最优化的功能, 它只需满足两个条件: 1) 它具有实现局部最优化的能力; 2) 它具有从一个局部最优状态向下一个更好的局部最优状态定向转移的能力。

在生存选择中, 广义遗传算法采用了四 (n 点交换时实际为 2^{n+1})。为论述简洁, 以下均称为“四”) 分之二择优选择的方式, 即在基因交换或基因突变的

过程中, 允许父辈和子代进行竞争, 并让其中的优良者进入下一轮的竞争环境。此举措保证了算法的迭代稳定性, 并使其具备了实现局部最优化的功能。已经证明, 对任意串进行有限次突变操作能够实现解空间的遍历搜索^[7,8]。广义遗传算法借助于以基因型的多点突变操作为主, 以基因交换为辅的策略, 实现了从一个局部最优状态向下一个更好的局部最优状态的定向转移。由于采用了局部最优状态的定向转移, 使算法为获得全局最优, 其搜索域不必覆盖整个解空间, 而只需覆盖感兴趣的有限解空间即可。换句话说, 广义遗传算法放弃了包括经典遗传算法在内的传统随机优化方法普遍采用的遍历搜索策略, 转而采用定向演化模式。

文献[8,9]用广义遗传算法成功地解决了香港青马大桥传感器群的最优布点设计问题, 并进而推测, 在通常情况下, 广义遗传算法也具有在有限步内实现问题的全局最优化的功能。下文将从数学上严格证明这一推测的正确性。

设式(1)存在 $m(m \geq 1)$ 个取值不同 (相同者进行归并) 的极大值点, 并根据其取值大小按升阶排列: $f(X_1) < f(X_2) < \dots < f(X_i) < \dots < f(X_m)$ 。设算法目前正处于状态 X_i , 定义:

$$X_i^+ = \{X: f(X) > f(X_i)\} \quad (3)$$

将 X_i^+ 所包含的状态数定为 Ω_i , 则 Ω_i 为 i 的单调减函数。由于对任意串进行有限次突变操作能够实现解空间的遍历搜索^[7,8], 所以

$$\text{Prob}(X_i \rightarrow X_i^+) \geq \text{Prob}(X_i \rightarrow X_j) > 0 (j > i) \quad (4)$$

也就是说, 当前状况为非全局最大值时, 广义遗传算法可在有限步内实现由当前状况向存在更大极值点区域的定向转移。

从式(4)可以看出, 状态的单向 ($X_i \rightarrow X_j$) 可达性是算法实现的充分条件; 更进一步可以推知, 最大值状态的单向可达性也是算法实现的充分条件。因此, 解空间的遍历搜索并不必要。换句话说, 广义遗传算法实现全局最优化的条件非常宽松。

对于组合优化问题, 其状态数是有限的, 由于局部极大值点的数恒少于其状态数, 因此, 广义遗传算法必可保证在有限步内实现问题的全局最优化。

在进化程序上, 广义遗传算法和经典遗传算法有所不同。经典遗传算法的进化程序为: 双亲选择 → 基因交换 → 基因突变 → 生存选择 → 下一代种群; 广义遗传算法的进化程序为: 双亲选择 → 基因交换 → 一家四口 → 四分之二生存选择 → 基因突变 → 一家四口 → 四分之二生存选择 → 下一代种群。也就是说, 广

义遗传算法将选择原则贯彻于整个生命周期。

2 算法的操作程序

从形式上看,广义遗传算法是针对由染色体顺序连接而成的基因型进行操作,根据表现型进行取舍的一种迭代运算。在操作环节和进化进程上,有以下鲜明特征:

1. 选择细分为父辈挑选和生存选择两种。父辈挑选采用不放回随机抽样,以确保每个父辈在子代中均有繁衍的机会;生存选择采用允许父辈和子代进行竞争,并让其中的优良者进入下一轮竞争环境的四分之二择优选择方式。两种选择方式的协同作用确保了进化过程的局部稳定性。

2. 在进化过程的每一次循环,交换和突变操作为一必然事件而非或然事件。

3. 交换和突变操作发生在染色体上,哪条染色体被选中是一非均匀随机事件,其概率依赖于上一次该染色体发生改变的后果,采用奖优罚劣的策略。

4. 以主群的稳定作为进化过程的收敛准则,使算法的收敛性有了明确的判据。

5. 将进化过程分为渐进和骤变两个阶段。渐进阶段,主要通过交换和选择的协同作用逼近局部最优状态;骤变阶段,主要通过突变和选择的协同作用实现局部最优状态的定向迁移,使下一次循环可在更好的起点上开始。

6. 交换和突变的顺序及方式依赖于种群所处的环境状况。渐进阶段,采用单点或少点交换和突变,其顺序为先交换,后突变;骤变阶段,采用多点交换和突变,其顺序为先突变,后交换。

7. 整个进化过程以渐进方式为主,骤变方式为辅。当进化过程局部收敛时,由反馈环节自动实现渐进方式向骤变方式的转换;同样,一旦群首被更换,则反馈环节自动实现骤变方式向渐进方式的转换。

8. 骤变阶段,采用均匀随机交换和突变的方式;渐进阶段,采用反馈型非均匀随机交换和突变的方式,且随着进化进程的展开,非均匀性由反馈环节进行自适应调节。

3 算法的进化策略及其生物学原理

上一节,我们系统阐述了广义遗传算法的操作程序,下面则进一步从宏观和微观协同作用机制上明确算法的进化策略及其生物学原理。

进化思想上,广义遗传算法以 Morgan 的基因理论及 Eldridge 与 Gould 的间断平衡理论为依据,

同时融合了 Mayr 的边缘物种形成理论和 Bertalanffy 一般系统理论的一些思想。

Morgan 的基因理论认为:1)基因是遗传物质的基本单位;2)基因位于细胞的染色体上;3)基因在染色体上排列为直线;4)同源染色体配对期间,两条染色体单体之间会发生交叉现象;5)两条染色体单体交叉时,部分基因可能发生交换和重组;6)基因突变在空间上是随机发生的;7)同一染色体的基因之间可能存在连锁现象。

1)~6)作为事实直接引入广义遗传算法的操作程序,而 7)则是算法执行过程中出现的必然现象。从仿真实验中我们得到启示:基因连锁是物种维持生存和稳定的有效方式。连锁现象本身就是生物进化的产物,原因是基因连锁的程度通常会随着进化过程的进行逐步强化。

完整的进化理论包括两个方面:进化的事实和进化的方式与动力。在承认进化是客观事实的前提下,关于进化的方式与动力,现代进化论者的观点并不统一。如同达尔文反对 Lamarck 的“获得性遗传”理论一样,新达尔文主义者不相信“定向变异和定向突变”的存在,他们认为,基因的突变是不定向的,只有选择作用是定向的。种群基因型的定向变异,是由选择作用造成的,而不是由于个体的定向变异造成的。在进化方式上,新达尔文主义者普遍赞成渐进模式,并认为自然选择是进化的唯一动力。综合进化论者 Mayr 虽然是达尔文自然选择学说的忠实捍卫者,但并不完全赞同进化的渐进模式。Mayr 认为,新物种的形成,首先是隔离机制的建立。在新物种形成初期,进化过程通常具有更快的速率。

广义遗传算法中,所谓新物种的形成过程,就是新局部最优状态的形成过程;隔离机制就是禁止参数空间中相隔较远的群落之间进行交配。隔离机制的作用是保证多局部最优状态的同步形成,也就是确保进化过程的并行化。

在仿真实验中发现,整个进化过程以渐进模式为主,其间偶尔被突发性的骤变模式所打断。骤变模式以表现型,而不是以基因型的急剧变化为特征,换句话说,骤变模式是以形态学为依据进行划分的。惟其如此,广义遗传算法摒弃了渐进进化模式,转而采纳美国古生物学家 Eldridge 与 Gould 在 70 年代创立的间断平衡理论(The Punctuated Equilibrium Theory)作为其生物学依据。间断平衡理论认为^[10],物种的演化以平缓变化的停滞期为主,其间偶尔被骤变的进化模式所打断。间断平衡理论是以下述现

代古生物学的重大发现为依据的:1)几乎所有的进化演变都集中发生在物种形成初期,换句话说,小的创始者种群通常具有更快的进化速率;2)物种形成过程结束后,物种就进入一段比较漫长的相对停滞期。间断平衡理论突破了达尔文进化论中单一渐变模式的束缚,认为渐变模式和骤变模式的交替作用才真正反映了进化过程的本质。

从仿真实验中受到启示:骤变模式的存在有其内在的发生机制,即基因型和表现型之间的非线性关系,致使基因型的微小变化在临界点外导致表现型的骤变。从现象学上看有两种表现形式:一种是在急剧变化环境中产生的骤变模式,由此产生的新生物种,需不断发生变化才可能生存。因此,环境剧变早期或中期产生的新生物种,延续下来的可能性很小。大量延续下来的新生物种,通常是在环境剧变的晚期或末尾产生的。此时,周围环境渐趋稳定,新生物种比较容易适应环境而得以生存和延续。由此推测,在环境剧变晚期或末尾的化石记录中,通常会有大量新生物种的涌现。事实是,在距今670万年至1万1千年间发生的11次冰期的晚期或末尾,均发现有大量新生物种的涌现。另外一种是在相对稳定的环境中随机发生的骤变模式。由此产生的新生物种的生存概率相对较小,且发生时间无规律可寻。至此,我们已成功再现和解释了现代古生物学的重大发现。

70年代以来,分子生物学对生物大分子(蛋白质、核酸)进化速率的研究发现了与现代古生物学研究似乎是相反的现象,这些现象本质上可归纳为两点:

现象1. 生物大分子的进化改变速率是相对恒定的。

现象2. 功能上对生物生存制约性弱的基因(中性基因)发生突变的频率大于功能上对生物生存制约性强的基因(功能基因)发生突变的频率。

以上述现象为背景,70年代初,日本学者木村创立了分子进化中性学说,并从此拉开了自然选择和中性学说近30年激烈争论的序幕。

现象1提出一个很尖锐的问题:微观上的匀速演化怎么会导致宏观上的骤变? **现象2**提出了一个更尖锐的问题:对于中性突变,自然选择如何起作用?

我们认为,即使以上证据都确凿无疑的话,木村的分子进化中性学说和现代古生物学研究成果的对立也只是一表象,构不成对后者真正的挑战。原因

是,对于几乎所有的生物,其基因型和表现型之间的关系是非线性关系而不是线性关系。因此,基因型的匀速变化在临界点处导致表现型的骤变实属必然,并没有什么不合理或者说存在矛盾的地方。上文提到,间断平衡理论认为,物种的演化以平缓变化的停滞期为主,其间偶尔被骤变的进化模式所打断。我们的仿真实验也证明了这种现象确实存在。Mayr虽然不清楚为什么会出现第2种现象,却大致正确地回答了第2个问题。Mayr指出^[12]:选择的单位应该是个体,而不是单个的基因,也不是种群和物种;并且,基因型是作为一个整体而不是单个的基因在发挥作用。Mayr显然忽略了一个本可以使其回答更直接、更有利的事实:选择不仅意味着择优,同时也意味着罚劣。中性基因发生突变的频率大于功能基因发生突变的频率可以看成是罚劣的后果。

广义遗传算法的选择单位是表现型,和表现型直接关联的是基因型,而不是单个的基因。随着算法逐步向局部最优状态逼近,基因型中的某些基因组会渐趋稳定——这就是Morgan发现的基因连锁现象和Mayr发现的染色体间的内聚约束力。此时,另外一些基因的变化对表现型的特性影响较弱,且越来越弱。这些对表现型的特性影响较弱的基因即是所谓的中性基因,而那些对表现型的特性影响较强的渐趋稳定的基因组中的基因即是所谓的功能基因。在仿真实验中发现,由于选择作用的存在,在幸存下来的个体中,中性基因发生突变的频率大于、甚至远大于功能基因发生突变的频率。在广义遗传算法中,特征2的目的在于加快演化进程,它不影响演化特性;特征3和特征8之所以采用非均匀随机抽样方式,是因为在仿真实验中发现,如果采用均匀随机抽样方式(均匀变异和突变),在幸存下来的个体中,中性基因发生突变的频率远远大于功能基因发生突变的频率。采用非均匀随机抽样后,虽不能消除这类现象,但可以使问题缓解,有利于演化进程的加速。换句话说,中性基因发生突变的频率大于功能基因发生突变的频率是一种必然规律。广义遗传算法中,间断平衡实际上就是分岔点。我们知道,迭代系统是典型的动力系统。当迭方程为非线性方程时,在迭代过程中出现分岔现象毫不奇怪。换句话说,我们之所以采纳间断平衡理论,是因为它所发现的现象是广义遗传算法执行过程中出现的必然现象。

广义遗传算法不仅完整地再现了Morgan发现的基因连锁现象和Mayr发现的染色体间的内聚约束力,同时也完整地再现了Eldredge与Gould发现

的间断平衡现象,以及木村等人发现的中性基因发生突变的频率大于功能基因发生突变的频率这一现代分子生物学中令人迷惑的未解之谜。

广义遗传算法中的表现型对应的实际上是生物界中的个体。上文已经说过,物种的演化以平缓变化的停滞期为主。在停滞期,种群逐步趋向于渐近局部最优状态。此时,由功能基因突变而产生的变异个体,由于基本特性变化过大,个体的适应能力远离近优状态,因此通常难以适应环境而被自然选择所淘汰,结果是这类个体无法留下化石记录;作为对比,由中性基因突变而产生的变异个体,由于基本特性变化不大,个体的适应能力在近优状态的邻域,因此通常可以较好地适应环境而得以生存,结果是这类个体会留下大量的化石记录。

综合以上分析,可以推测:在停滞期的化石记录中,中性基因发生突变的频率应该大于、甚至远大于功能基因发生突变的频率。这就是我们对木村的分子进化中性学说所表述的第2种现象的解释。同样,有理由推测:在骤变期的化石记录中,功能基因发生突变的频率应该大于中性基因发生突变的频率。后者是一个有待证实的猜测。

Morgan发现的基因连锁现象和Mayr发现的染色体间的内聚约束力表达了同样的观念,即变异在空间上不可能是均匀的。由于变异和突变之间有时存在耦联关系,因此有理由推测,突变在空间上可能也是不均匀的。从逻辑上讲,突变在空间上的不均匀性必然导致定向突变的存在。

Bertalanffy认为,自然选择和随机突变在进化中起了重要作用,但绝不是全部作用。自然选择无法解释由简单到复杂的进化趋势,因为这一趋势无疑和有机体内部的自主进化机制有关。Bertalanffy发现,开放系统的演化会伴随着逐步机械化这一事实。换句话说,原始物种往往具有全能性,或者说无限制的可能性。但随着进化过程的展开,这些特征逐渐丧失,代之而起的是功能精确化、固定化,无限制的可能性被有限的现实性所替代。我们知道,可能性维度上的压缩实际上意味着定向变异或定向突变的存在。

定向突变实际上有两种含义:突变点的不均匀性和突变后果的方向性。Morgan和Mayr认为前者为生物系统实际采用;Bertalanffy推测,后者同样有为生物系统实际采用的可能性。从逻辑上讲,突变点的不均匀性必然导致突变后果的方向性。换句话说,我们没有理由否定向向突变存在的合理性。

偶然性、随机性和非均匀性事实上是3个不同的概念。偶然性排除了规律性的存在;随机性蕴涵着存在概率意义上的规律性;非均匀性是一种独立的属性,它和随机性或确定性之间均不存在必然的联系,这一点往往被人们所忽视。事实上,只要承认变异或突变在空间上是不均匀的,必然导致承认定向变异或定向突变的存在,这一结论和突变或变异随机与否并无关系。由此推测,Bertalanffy所谓的随机突变指的应该是随机均匀突变。

采用广义遗传算法进行仿真实验时发现:均匀随机变异和突变也可以使系统进化,但定向变异和突变有助于进化过程的加速。作为一个完全的进化论者,我们有理由推测,生物系统将逐步倾向于采用定向变异和突变的方式实现快速进化。

与新Lamarck主义者忽视自然选择的作用不同,我们认为,自然选择在进化过程中发挥了极其重要的作用:它规定和制约了现实进化的可能方式和倾向。换句话说,自然选择是作为监督机构和反馈环节作用于生物系统的,适应或者不适应决定了是采用正反馈还是采用负反馈。反馈作用的渐近结果便是生存或毁灭。

新达尔文主义学说认为,突变以恒定的速率发生,这一速率与突变的后果无关。广义遗传算法摒弃了新达尔文主义学说对突变作用的想法,相信选择、定向交换和定向突变的协同作用是实现快速进化的合理方式。广义遗传算法首次将结果反馈引入操作程序,并通过它对选择、交换和突变的工作方式进行自适应调节。

参考文献

- [1] J. H. Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, 1992
- [2] 董聪、郇正能、夏人伟、何庆芝, 多层前向网络研究进展及若干问题, *力学进展* 25(2), 1995
- [3] 董聪、刘西拉, 广义BP算法及网络容错性和泛化能力的研究, *控制与决策*, 13(2), 1998
- [4] 董聪, 多层前向网络的逼近机理与拓扑结构学习方法, *通信学报*, 19(3), 1998
- [5] 董聪, 前向网络全局最优优化问题研究, *中国科学基金*, (1), 1997
- [6] T. N Bui, et al., *Genetic Algorithm and Graph Partitioning*, *IEEE Trans. on Computers*, 45(7), 1996
- [7] 董聪, 广义遗传算法, *大自然探索*, 16(1), 1998
- [8] 董聪, 广义遗传算法的数学结构与生物基础, *中国科学基金*, (2), 1998
- [9] 董聪、秦权、陈一珉, 基于广义遗传算法的青马桥传感器群最优布点设计, *CCNS'97*, 南京, 人民邮电出版社, 1997
- [10] E. 迈尔, *生物学哲学*, 涂长晟等译, 辽宁, 辽宁教育出版社, 1992