计算机科学1998Vol. 25№. 6

进化算法的选择机制分析

The analysis of selection mechanism in evolutionary algorithms

莫纯欢[1,△] 石纯一 陈 青[△] 周代琪[△] 史忠植[*]

(深圳市华为技术有限公司[△] 清华大学计算机系[1] 中国科学院计算技术研究所[*])

摘 要 This paper gives an analysis and comparison for selection mechanism in evolutionary algorithms, random selection, e. g. roulette wheel selection; competitive selection, e. g. (μ,λ)selection, q-tournament selection. If the initial fitness function of q-tournament selection is normal distribution, we can give out the behavior of the q-tournament selection. To run in parallel EA, the key is the design and implementation of useful migration mechanism, but decentralized selection with global or local gene pools can solve this problem. If the selection variance is higher, population size is smaller, the search performance will be poor, to improve performance, one must reduce selection variance or increase population size.

关键词 Evolutionary algorithms, Selection density, Loss of diversity

一、引言

进化算法(EA)是受自然进化所启发的搜索和优化技术,它包括:遗传算法(GA),进化规划(EP),进化策略(ES),和遗传编程(GP)。

GA 是 John Holland[1]等人于1975年在美国密歇根大学应用自然选择过程来解决机器学习问题时提出的,它在于搜索有高的适应值的基因结构;ES 在于搜索与适应值函数相匹配的行为;EP 采用了比 GA 和 ES 更抽象的进化模型,它模拟了在持续的后代物种行为之间的繁殖关系;GP 是属于 GA 的最新衍生物,其中群体的个体是计算机程序,例如,C, Pascal. ,Fortran,Lisp 等编程语言的程序。

进化算法(EA)的运行过程如下:

1. 初始化:随机产生一个群体(第一代),评价群体中每个个体的适应值。

2. 重复执行下述步骤直到满足停机标准:①从群体(第 N 代,N=1,2…)中选择父母亲;②用重组算子(交叉,突变,反转等)产生孩子;③计算孩子的适应值;④选择出个体组成第 N+1代的群体。

3. 结束。

遗传算法的详细过程很难跟踪,即使是对单纯的选择方法。这是因为存在着大量的可能状态,但可以用 Markov 链[9]或者动态系统理论[11]进行形式分析。在[11]中提出了新的选择模型,普通投票选择机制,通过研究虚拟的平均群体 VAP 的进化,容易发现渐近的群体行为。当然,使用 VAP 并不局限于普通投票选择。

研究进化涉及两个问题,一是如何把个体引进群体中的方法;二是,一旦把有兴趣的个体引进群体后,使之保留在群体中的办法。单纯地考虑选择,而不考虑交叉,对于研究进化有简化的意义,通过研究选择方法本身可以更好地了解某种平衡状态的收敛速度[7-9]。

选择机制是通过把适应值高的个体复制到下一代而改善群体的平均适应值。所以,选择集中于在搜索空间有希望的区域中搜索。通过交叉或者突变的重组则改变了群体中的遗传材料,以便在搜索空间中开发新的点。作为优化应用的进化算法的收敛速度基本上是由选择算子所决定的,更精确地说,是通过选择算子施加于搜索过程的选择压力。

开发和探测之间的平衡可以通过调节选择算子的选择压力,或者调节交叉算子和突变算子的概率。该平衡对于 EA 的行为是很关键的,所以,了解选择算子和重组算子的性质,就可以理解这些算子对收敛速度的影响。

[3]引进占优势时间(takeover time)的术语,占优势时间就是:假如不用重组算子,单个的最好的个

体充满整个群体所需要的后代的数量。而[4]分析了最著名的选择机制，该机制使用根据"占优势时间"的进化算法。在[5]中，复制遗传算法 BGA(breeder GA)中的选择密度用来测量群体的进步，选择密度起源于比例选择和切断选择。

[12]使用阶统计来分析(μ,λ)选择和锦标赛选择；展示了选择密度和(μ,λ)选择的参数 μ,λ,锦标赛选择的锦标规模 q 之间的数学关系。

基于最好的个体的行为的分析，或者基于平均群体适应值的分析仅仅描述了选择方法的一个侧面，选择机制也可以用适应值分布的交互来描述。

一般来说，选择机制的行为仅仅依赖于群体中个体的适应值。所以假设选择和重组是顺序进行的，首先，选择阶段创造了临时的群体，带有确定概率的重组算子作用于该临时的群体，而得到下一代群体。这样的描述与从重组的个体中选择的方法是不同的。但在数学上是等价的，可以分别考虑。

二、随机选择机制

2.1 滚轮选择机制

群体中每个个体被选择作为双亲的概率正比于该个体的适应值。也就是说，设 t 是代数，在第 t 代中个体 i 的适应值是 f_i，则个体 i 被选择作双亲的概率 Pr_i。

$$Pr_i=\frac{f_i}{\sum_{i=1}^{N} f_i},$$

详情参见[1,2]。

2.2 普通投票选择机制

详情见[11]。

三、竞争性选择机制

3.1 q 锦标赛选择

锦标赛选择是个竞争性选择方法，当前普遍用在 GA 和 EP 中，一旦发现了每一个个体的适应值，则通过比较适应值的方法来选定下一代的父亲。

在 q 锦标赛选择中，从群体中随机选出 q 个个体，并对它们的适应值进行比较。有最高适应值的个体声称是赢者，把赢者复制到下一群体中。重复该过程，直到把下一群体充满为止。

在进化规划(EP)的 q 锦标赛选择中，每一个群体成员的竞争成绩是通过把该成员的适应值与其它 q−1 个群体成员的适应值比较而创造出来的。在每一个群体成员都获得成绩之后，有最高竞争成绩的那些群体成员将变成下一群体的双亲。

锦标赛适应值使用单个的切断、二进制的锦标赛决定在群体中相对的适应值排序，初始化时，整个群体都参加锦标赛。随机选择两个成员，相互竞争，仅仅赢者才有机会参加下一轮的锦标赛，第一轮竞争完成后，赢者随机配对以决定下一轮赢者，继续进行锦标赛选择直到只有一个赢者。

在进化算法 EA 中，群体表示了差异性的自然资源，竞争的环境有可能太没有结构，很难指导一个群体朝着特定的目标，除非该目标是适合模糊的，或者是不存在的。同时需要对群体成员做很多评价，以决定精确的排序，必须综合考虑进化过程的自然动态。

下面我们假设 q 锦标赛选择的初始适应值函数是正态分布的情况下，讨论 q 锦标赛选择的行为。

3.1.1 q 锦标赛选择的初始适应值函数是正态分布

3.1.1.1 定义 假设适应值 f_i(i=1,2,…,λ,λ 是群体的固定大小)是正态分布的，f_i 是群体的个体 i 在 t 时的大小，$f'=(1/\lambda)\sum_{i=1}^{\lambda} f_i$ 是适应值的平均值，标准偏差记为 σ，适应值 f_i 根据 $N(f',\sigma_i)$ 分布，也就是说，可以解释为带有公共的概率密度函数的样板随机变量 $F_i\sim N(f',\sigma_i)$ 以不增的排序安排 F_i，重新排序 $F_{1,\lambda}\leqslant F_{2,\lambda}\leqslant\cdots F_{\lambda,\lambda}$。

定义1(选择方法) 选择方法 g 就是一个函数，它把一个适应值分布 s 变换为另一个适应值分布 s^*，$s^*=g(s)$。

定义2(繁殖率) 繁殖率 R(f)记为在选择后和选择前的个体的数目与一个确定的适应值 f 的比率。

$$R(f)=s^*(f)/s(f):s(f)\geqslant 0,R(f)=0:s(f)=0$$

定义3(差异性丢失) 差异性丢失 p_d 是在选择期间群体中没有被选择的个体的比例。

选择密度或称选择压力，用在一个选择方法的不同性质，使用占优势时间来定义选择压力。群体的平均适应值的变化是选择密度的一个合理量度，但依赖于初始适应值的分布。

定义4(选择密度) 选择密度 I 是在对正态高斯分布 $G(0,1)(f)=(2\pi)^{(-1/2)}e^{-f^2/2}$ 应用了选择方法之后，群体的期望平均适应值：

$$I=\int_{-\infty}^{\infty} f\omega^*(G(0,1))(f)df。$$

定义5(选择方差) 选择方差 V 是对正态高斯

分布 $G(0,1)$ 应用了选择方法之后，群体的适应值分布的期望方差：

$$V=\int_{-\infty}^{\infty}(f-I)^2\omega^*(G(0,1))(f)df。$$

选择方差和差异性丢失率有一点差别。差异性丢失给出了没有被选择的个体的比例，而不管适应值是多少；而选择方差定义为假设高斯初始适应值分布时适应值分布的新的方差，通过指定适应值可以知道选择方差和选择密度。

3.1.1 2 锦标赛选择分析 对于单个的 q 锦标，期望的适应值：

$$f_{\mu}^{t+1}=(1/\lambda)\sum_{i=1}^{\lambda}E(F_{q,q}^{i}=E(F_{q,q}^{i}))$$

在标准化和正态化以后，

$f_{\mu}^{t+1}-f^{t}=\sigma_t E(Z_{q,q})$

选择密度

$I=E(Z_{q,q})$。

现在知道 $q=1,2,3,4,5$ 的解析式，$q=1,I=0$；$q=2,I=(\pi)^{(-1/2)}$；$q=3,I=(3/2)(\pi)^{(-1/2)}$；$q=4,I=6(\pi)^{(-3/2)}\arctan 2^{(-1/2)}$；$q=5,I=10(\pi)^{(-1/2)}[(3/2)(\pi)^{(-1/2)}\arctan 2^{(-1/2)}-(1/4)]$；但 $q\geqslant 6$ 的解析式还不知道。

3.2 (μ,λ)选择

在 (μ,λ) 选择中，$\lambda\geqslant\mu$，从 λ 中选择出最好的(适应值最高)的 μ 个个体，作为下一代的双亲，而 $\lambda-\mu$ 个个体被丢弃。[6]使用术语切断选择来代表一个等价的方法，基于选择 T 个最好的个体，对应于 (μ,λ) 选择中的 $T=\mu/\lambda$。

选择压力完全由 λ/μ 来决定，选择密度可以近似为 $x\geqslant x_s$ 的基础的概率密度函数的切断的期望，经计算，选择密度 I：

$I=(\lambda/\mu)\phi(x_s)$，

这里 ϕ 是正态密度函数。

3.3 分散选择算法

3.3.1 为什么要引进分散选择算法 并行计算结构可用性的不断增加提供了用 EA 来求解复杂问题的可能性。为了有效地开发细粒度的并行结构，EA 的控制结构必须分散化，所以，有必要研究分散选择算法。

EA 的经常讲述的一个长处就是"自然"的并行性，然而，人们用到 EA 时，一般都是显式和稳式的中心化控制。让 EA 来处理粗的和细粒度的并行结构，对于粗粒度的并行的最自然的适应就是"孤岛"模型[13]，这时有很多个中心化的 EA 在并行运行，而关键在于设计和实现有用的"迁移"机制。

允许独立进化的局部群体之间的信息交换，在于用 EA 来有效地开发细粒度的结构，在于把算法并行化，很多传统的选择算法，例如，与适应值成比例的，与排序成比例的，或者切断选择。这时全局的计算要求很高的通信开销，可以通过把它们分散化，以产生比中心化不同的选择压力(即选择密度)，从而导致不同的进化行为。

一个例外就是锦标赛选择，因为不用保持任何全局统计或者分配排序，所以锦标赛选择可以用在分散化的进化算法中。

3.3.2 带有全局基因池的分散选择 带有全局基因池的分散选择算法包括两个步骤：一个选择池，以及该池中的选择概率分布。所谓池，是整个群体。选择概率分布用该个体的适应值相对于池中的其它的个体的适应值来定义。

锦标赛的例外就是以正态分布的概率从选择池中随机选择 k 个个体，其中有最高适应值的个体算是赢者，该赢者被选择为父亲。

显然，二进制锦标赛选择($k=2$)等价于标准的线性排序机制的期望值，从两个个体中找出做好的个体，而坏的就不要。

所以，二进制选择可以作为分散选择，把群体的每一个成员指定给个别的处理器，并行地在每一个处理器中执行各自的代码，而在整个池使用二进制机制来选择双亲。适当数目的双亲可以产生一个孩子，以代替指定给该处理器的当前个体。

实现分散选择有两个问题：

①通信代价：不同的大多数细粒度的结构通信代价比邻居的要高。所以，存在着这样的问题，是否任何形式的全局选择池在这样的结构中都是有效的实现方法。

②假如忽略通信代价，则与传统的最佳曲线做比较，二进制锦标赛选择比等价的线性排序机制要差，虽然两种机制的选择压力是一样的。

3.3.3 带有局部基因池的分散选择 之所以使用局部的邻域作为选择池，主要是从减少通信开销以及生物合理性的观点来考虑。这时需引进某种距离测量，或者群体中的拓扑，这样才能定义邻域的概念，度量涉及到基因型或者表现型空间的距离(例如，共享函数)。

最简单的拓扑是两维的，格子是方的，围绕该特定格点的邻域是从该格点的上，下，左，右的一定步数而定义的，每一个格点，都有一个邻域，该邻域可以用附近格点的邻域来重叠。

为了取得完全的分散化，修改的 EA 在每一个格点并行运行，从邻域选择双亲，产生孩子，若可能的话，代换该格点指定的当前个体，重叠邻域提供隐式的机制以便在方格周围迁移遗传材料。若邻域太大，则要承担与3.3.2节同样高的通信开销，所以，一般的研究都是考虑小的邻域。每一个邻域定义了一个局部选择池。在这些小的邻域中考察选择算法的作用，例如，比例选择，排序选择，二进制锦标赛选择。

选择算法的有效分散化，可以开发细粒度的并行结构，若没有假设选择方差，则需要假设选择算法有等价的期望选择压力，以产生类似的搜索行为。

若方差较高，则当涉及到小的群体规模时，与差的搜索性能有较强的相关性；若要改善性能，则必须减少选择方差，或者增加群体的规模。

对于三个局部选择压力机制的分析，指出需要较强的局部选择压力来诱发适当的全局选择压力。在所研究的三个局部选择机制中，无论是从全局选择的观点还是通信开销来看，二进制选择机制都是最合符需要的性质。除此之外，由局部选择算法所诱发的研究全局选择分布表明了优越的策略怎样与二进制锦标赛选择结合起来以便改善性能，减少通信开销。

参考文献

[1] J. H. Holland, Adaptation in natural and artificial systems, Ann Arbor, MI: the University of Michigan Press, 1975

[2] D. E. Goldberg, Genetic algorithms in search, optimazation, and machine learning, Reading, MA: Addison-Wesley publishing company, Inc. 1989

[3] E. Goldberg David et al., A comparative analysis of selection schemes in genetic algorithms. In G. Rowlins (editor), Foundation of genetic algorithms, San Maeto, Morgan Kaufmann, 1991

[4] Thomas Back, Selective pressure in evolutionary algorithms: A characterization of selection mechanisms. In proc. of the first IEEE conference on evolutionary computation IEEE world congerss on computational intellgence(WCCI), 1994

[5] Muhlenbein, Schlierkamp, Voosen, Predictive models for the breeder genetic algorithm. Evolutionary computation 1(1), 1993

[6] Muhlenbein, Schlierkamp, Voosen, The science of breeding and its application to the breeder genetic algorithm (BGA) Evolutionary computation 1(4), 1993

[7] D. E. Goldberg et al., Comparative analysis of selection schemes used in genetic algorithms, Same to [3]

[8] S. W. Mahfoud, Crowding and preselection revised, Proc. parallel problem from nature conference (Brussels, Belgium), 1992

[9] D. Thierens et al., Convergence models of genetic algorithm selection schemes, Lecture notes in computer science, 866, 1994

[10] D. E. Goldberg et al., Finite markov chains analysis of genetic algorithms, Proc. 2nd int. conf. on genetic algorithms(Cambridge, MA), 1987

[11] J. Shapiro etc, A statistical mechanical approach to model the dynamics of genetic algorithms. Lecture notes in computer sciences, 865, 1994

[12] D. E. Goldberg, et al., Messy genetic algorithms: motivation, analysis, and first results. Complex systems 3(5), 1989

[13] Levine David, A parallel genetic algorithm for the set partitioning problem, Ph. D dissertation in computer science of illinois institute of technology, May 1994