文章编号: 1001-0920(2004)05-0506-04

遗传算法中的自相似现象

张 伟, 吴智铭, 李树刚, 杨根科
(上海交通大学 自动化系, 上海 200030)

摘 要: 遗传算法(GA)的执行过程可看作复杂自适应系统的演化过程. 以 GA 运行过程中输出的适应值序列为研究对象, 通过功率谱和重标定域两种方法发现 GA 的输出序列存在自相似行为. 研究表明, 算法执行过程中最优解的输出与 Hurst 指数的变动密切相关, 算法在全局搜索阶段呈现明显的自相似性, 而在到达最优值附近则呈现明显的白噪声特征. 这一发现为深刻理解 GA 运行机理和优化 GA 设计带来了新的思路.

关键词: 遗传算法; 自相似性; Hurst 指数; 功率谱幂律; 演化行为
中图分类号: TP18 **文献标识码**: A

Self-similarity in genetic algorithm

ZHANG Wei, WU Zhi-ming, LI Shu-gang, YANG Gen-ke
(Deptartment of Automation, Shanghai Jiaotong University, Shanghai 200030, China. Correspondent: ZHANG Wei, E-mail: zhang_wi@sjtu.edu.cn)

Abstract: The evolution of GA is a kind of complex adaptive system evolution. By introducing the power spectrum density analysis and re-scaled range method, the self-similar behavior is revealed in GA's fitness series. Further investigation shows the Hurst exponent has intimate relationship with GA's fitness outputs. The evolution of GA exhibits strong self-similar characteristic in its global search stage and much noisy characteristic in its local search stage. This improves the understanding of GA dynamics and will help to better implementation of GA algorithm.

Key words: genetic algorithm; self-similarity; Hurst estimator; power-law; evolutionary dynamics

1 引 言

近年来, 以遗传算法(GA)为代表的演化计算在许多领域, 特别是 NP-Hard 问题的数值化求解方面, 取得了非常成功的应用[1], 但仍存在很多问题, 如 GA 收敛速度的预测与评估. 尽管 Holland[2]提出了积木块原理, 直观地解释了 GA 为什么具有寻优的能力; 其他学者证明了在保留最优值的情况下, 标准遗传算法(SGA)可依概率 1 收敛到最优解[3], 但这些工作并没有揭示出遗传算法运行过程的规律.

从复杂自适应系统(CAS)的角度看, 一个 GA 算例相当于一个微型复杂系统. 其中, GA 中的个体相当于复杂系统中的智能体, 个体之间的交叉算子相当于智能体之间的信息交换过程, 个体的变异算子则可理解为单个智能体对环境的探索过程; 相应地, 整个 GA 的运行过程便可理解为复杂系统的演化过程. 因此, GA 很可能具有 CAS 系统中常见的一些规律和现象. 本文通过考察标准 GA 输出的适应值序列发现, GA 在寻优过程中存在自相似现象, 且自相似程度随着 GA 由全局搜索到局部搜索的转变有降低的趋势. 这一发现揭示了 GA 作为复杂系统演化所具有的运行规律, 有助于我们更好地理解 GA 的动态运行过程, 并改进现有的 GA 算法.

收稿日期: 2003-05-19; **修回日期**: 2003-07-07.
基金项目: 国家自然科学基金资助项目(59889505, 70071017).
作者简介: 张伟(1975—), 男, 河北沧州人, 博士, 从事复杂网络、非线性时间序列分析等研究; 吴智铭(1936—), 男, 河北沧州人, 教授, 博士生导师, 从事生产计划与调度、智能计算方法等研究.

本文介绍了复杂系统中的自相似行为及其评估标准——Hurst 指数, 以 Rosen-Brock 函数优化问题的标准遗传算法求解为例, 判定了 GA 运行中自相似行为的存在, 并以 TSP 问题的遗传算法为例, 考察了自相似行为随演化代数而变化的规律

2　自相似现象

自相似性(分形性)是复杂系统中常见的一种现象[4]. 直观上看, 自相似系统经常在某种测度的不同尺度上表现出相似的特征. 这里的测度可以是空间、时间或任何其他合理的定义

2.1　自相似随机过程的定义

定义 1(自相似随机过程的连续时间定义)[4]　一个随机过程 $x(t)$ 在统计意义上是以参数 H ($0.5 < H < 1$) 自相似的, 如果对于任意实数 $\alpha > 0$, 随机过程 $\alpha^{-H}x(\alpha t)$ 与 $x(t)$ 有相同的统计特征. 该关系可用以下 3 项条件表达:

1) 均值

$$E[x(t)] = \frac{E[x(\alpha t)]}{\alpha^{H}}; \tag{1}$$

2) 方差

$$Var[x(t)] = \frac{Var[x(\alpha t)]}{\alpha^{2H}}; \tag{2}$$

3) 自相关

$$R_x(t,s) = \frac{R_x[\alpha t, \alpha s]}{\alpha^{2H}}. \tag{3}$$

参数 H 称为 Hurst 指数, 在本质上是一种随机现象持续性或长程依赖程度的度量, 反映了系统在不同尺度上自相似的程度. H 一般可通过重标定域法(R/S)计算, $H = 0.5$ 表示没有自相关性; H 越接近于 1, 持续性或长程依赖性的程度越大, 自相似的程度也越高

上述定义是基于连续时间变量直接尺度变换的定义, 相应的离散时间定义如下:

定义 2(自相似随机过程的离散时间定义)　对于一个平稳时间序列 x, 首先定义 m 重聚集时间序列 $x^{(m)} = \{x_k^{(m)}, k = 0, 1, \cdots\}$, 将原时间序列分成大小为 m 的块, 然后在各块中求出平均值, 这一过程可表示为

$$x_k^{(m)} = \frac{1}{m}\sum_{i=km-(m-1)}^{km} x_i \tag{4}$$

多重聚集时间序列可视为原时间序列在时间尺度上的压缩, 如果原始序列经压缩而相应的统计特性(均值、方差、自相关)保持不变, 则可认为这是一个自相似过程. 因此, 自相似性在直观上意味着原始序列在不同尺度上的聚集序列具有相同的变化程度或突发性

2.2　功率定律与双对数功率谱图

自相似过程在频域内的重要特征之一是具有 $1/f^{\beta}$ 信号的特点, 服从功率定律, 即对随机过程 x, 其功率谱密度为

$$\Gamma(\nu) \approx c|\nu|^{-\alpha}, \nu \to 0. \tag{5}$$

其中: c 为常数, $0 < \alpha = 2H - 1 < 1$, H 为 Hurst 指数. 式(5)两边取对数, 有 $\lg\Gamma(\nu) \propto \lg(\nu)$. 可见, 若在功率谱图上取双对数坐标, 则具有直线特征的功率谱曲线就表明所分析序列服从功率定律. 当然, 对于实际系统而言, 功率定律一般只能在一个有限的频段 $\nu_{min} < \nu < \nu_{max}$ 内成立. 本文用双对数功率谱图作为检测自相似性是否存在的方法之一.

2.3　自相似程度的度量——Hurst 指数及 R/S 算法

如定义 1 所述, Hurst 指数 H 是自相似程度的一种度量, H 的值介于 0.5～1 之间, H 的值越大, 表明自相似程度越高. 一般情况下, H 的值都是利用系统在某种测度下不同尺度的不变性进行估计. 本文采用重标定域法[5] 计算 Hurst 指数. 算法描述如下:

算法: 重标定域法(R/S 法)估计 Hurst 指数;

输入: 时间序列 $\{X_n, n \in Z^+\}$.

1) 计算时间序列 X_n 的均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 方差 $S^2(n)$, 距离均值的差值 $\Delta X(k) = X_k - \overline{X}$. 显然, $\Delta X(k)$ 的均值为 0, 故这个动作称为重标定(或归一化).

2) 计算累加值 $\Delta_j = \sum_{i=1}^{j}\Delta X(i)$, $j = 1, 2, \cdots, n$.

3) $R_n = \max(0, \Delta_1, \Delta_2, \cdots, \Delta_n) - \min(0, \Delta_1, \Delta_2, \cdots, \Delta_n)$.

4) 计算 R_n/S_n, 这一过程相当于将 R_n 归一化(标准化). 称 R_n/S_n 为重标定域, 它与 n 有幂关系, 即 $R_n/S_n \propto n^H$. 两边取对数, 有

$$\log(R_n/S_n) = H\log(n) + \log(c), \tag{6}$$

故在对数平面上选取 n 为水平轴, R_n/S_n 为垂直轴绘出曲线, 应近似为一条直线, 该直线的斜率即为 H.

在实际估计 H 时, 应取直线段部分进行估计, 同时视情况舍弃直线段边缘误差较大的点

3　标准遗传算法中的自相似现象

本实验选取标准遗传算法(SGA)[3] 在运行过程中输出的群体平均值序列和随机选择的某个体值

序列为研究对象 算法采用二进制编码, 原算例[3] 群体大小M = 80, 交叉概率P_c = 0.6, 变异概率P_m = 0.001, 以Rosen Brock 函数优化目标:

$$\max f(x_1, x_2) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2,$$

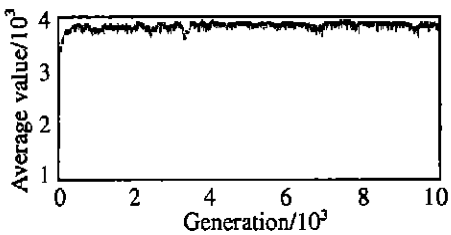
$$\text{s.t.} \quad -2\,048 \leq x_i \leq 2\,048,\ i = 1, 2$$

该算例是GA 的一个最小实现, 仅仅包含了GA 运行所必需的选择、交叉和变异等算子. 考虑到原算法选择算子中的最优个体保留策略会导致算法迅速收敛, 所以特意放弃该策略, 并且缩小群体规模为10, 延长最大演化代数到10 000, P_c 调整为0.2, P_m 调整为0.01, 使算例输出的值序列不致迅速收敛, 以充分反映GA 演化过程中的动态行为

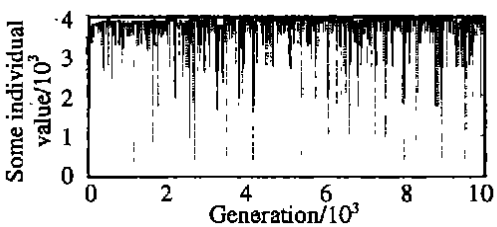
经过上述修正后, 虽然算法不能从理论上保证一定依概率1 收敛到最优, 但这并不违背GA 的基本原理, GA 仍然保留了在整个搜索空间中搜索最优的能力; 而且这种最简化的GA 可以充分代表各种GA 变种算法, 使本实验的结果更具代表性和普遍性

图1(a) 和(b) 分别是该简化SGA 算例演化10 000 代输出的平均值序列和个体值序列 从图中可以看出, 作为整体而言, GA 可以迅速搜索到最优值邻近区域, 但并不能避免在最优值附近的长期波动, 这一点在图1(b) 中看得更清楚 GA 中单个个体的适应度并不因为整体适应度的相对稳定而有所收敛, 相反, 它依然保持了大范围的波动性质

图2 是适应值序列的功率谱图双对数化后的结果, 其中功率谱的计算采用Welch 方法, 横坐标采用对数坐标 从图中可以看出, 群体平均值序列的功率

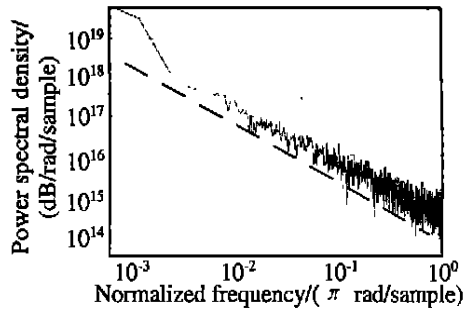


(a) 群体的平均值序列

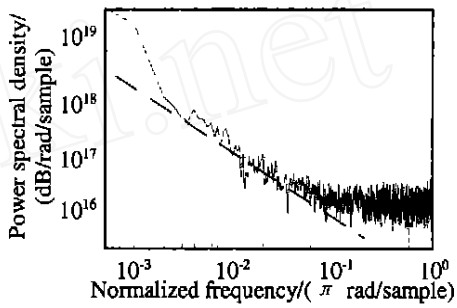


(b) 个体值序列

图1 SGA 的输出序列



(a) 平均值序列



(b) 个体值序列

图2 SGA 输出值序列的功率谱

谱图在取对数之后呈现非常明显的直线特征, 这是自相似性存在的结果 相比图2(a) 和(b) 中的个体适应度序列, 仅在中低频段呈现出直线特征, 而在高频段更多地呈现随机白噪声特征, 这反映了GA 中单个个体的适应度演化规律 虽然短期内的波动(对应图2(b) 高频部分) 更多地呈现随机波动特征, 但从长期演化的趋势看, 与系统平均的变化规律是一致的(对应图2(a) 和(b) 的中低频部分).

适应度序列中是否存在自相似性特征的进一步判据是H 指数 通常, H 指数可通过R/S 法进行估计. 一般认为, 当0.5 < H < 1 时, 被分析序列中存在自相似特征 以R/S 法估计SGA 平均值序列和个体值序列的H 指数分别为0.92 和0.88, 表明在一定尺度范围内, 上述适应值序列中存在较强的自相似性

为什么适应度序列的自相似特征在平均意义上反映得更明显呢?这是因为作为GA 中的个体, 其演化行为受交叉和变异算子的影响, 而这两大算子内含的随机性会使个体适应度的变化在小尺度上更多地呈现噪音特征; 而大量看似在作噪音运动的个体的综合, 又使得GA 在整体上呈现出明显的非随机行为(即在解空间中寻优的过程). 这便是GA 动态运行过程的外在表现

4　GA 演化过程中自相似行为的变化

上节实验验证了 GA 运行过程中自相似行为的存在, 本节进一步考察自相似性随 GA 演化而变化的规律 鉴于前述算例以较少的进化代数就能到达最优值附近, 不适合用来考察 Hurst 随演化代数变化的规律, 故引入 GA 求解 TSP 问题 其中, 算法实现参照上节作同样简化, 但考虑到 TSP 问题本身的特殊性, 仍然保留了"个体修正"操作, 否则, GA 运行中交叉和变异算子会产生大量的不可行解, 严重干扰算法的执行过程

图 3 的数据是 GA 演化 76 141 代输出的最优值序列 由图可以清楚地看到, 虽然 GA 的输出可以较快地到达最优值附近, 但输出值的波动并不因在最优值附近而有所减弱 这分别对应于 GA 的高效全局搜索和局部搜索缓慢两种特征

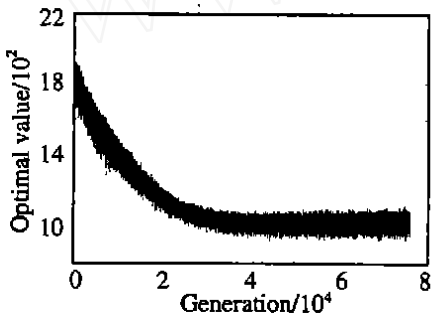


图 3　GA 求解 TSP 问题输出的最优值序列

根据图 3, 分别选取 10 000～ 20 000 代, 30 000 ～ 40 000 代和 50 000～ 60 000 代输出的最优值序列估计 Hurst 指数, 计算结果分别为: 0. 95, 0. 91, 0. 67, 它们分别对应于 GA 演化过程中的 3 个阶段: 1) 全局寻优阶段, 这时 GA 的搜索效率和自相似程度都是最高的, Hurst 指数可以超过 0. 9; 2) 由全局搜索到局部搜索的过渡阶段, 这时 GA 的输出趋向最优, 并具有较强的自相似性; 3) 局部搜索阶段, 这时 GA 的输出已到达最优值附近, 并且处于上下波动无法稳定的状态, 呈现出明显的白噪音特征, Hurst 指数也随之明显降低, 只有 0. 67 左右, 自相似性已经很不明显

图 4 进一步展示了 Hurst 指数随演化代数变化的情况 相邻 Hurst 指数的计算相差 5 000 代, 每次计算均取 10 000 代演化输出的最优值 与图 3 对照可以清楚地看到, GA 的全局搜索阶段具有较高程度的自相似性(Hurst 指数在 0. 8 以上), 局部搜索阶段具有较强的噪音特性(Hurst 指数低于 0. 7), 而在中间的过渡阶段, Hurst 指数出现大幅的波动

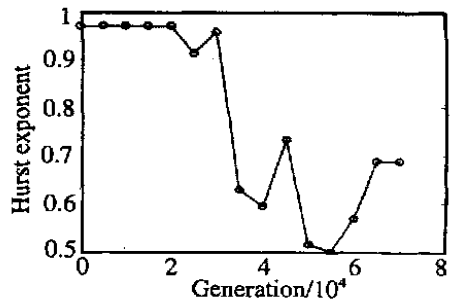


图 4　Hurst 指数随 GA 演化代数变化情况

由此可见, GA 的运行阶段与其自相似程度之间存在密切联系 在算法运行初期的全局搜索阶段, GA 中的每个个体通过彼此之间的信息交换(以交叉算子和变异算子实现)完成解空间的高效搜索, 导致最优解的输出呈现明显的自相似现象; 当 GA 到达稳定解附近时, 个体之间的信息交换所起的作用已不明显, 交叉变异算子所起的作用更象是随机噪音干扰, 这也是 GA 局部搜索能力较弱的一个原因 对自相似程度的判断有助于确定何时在 GA 中引入高效的局部搜索策略

5　结　　语

本文从复杂自适应系统的角度看待 GA 的运行过程, 以 GA 运行过程中的群体平均值、个体值和最优值序列为研究对象, 通过对数功率谱计算和 R/S 算法两种方法, 确认了 GA 中自相似特性的存在 实验结果表明, 群体序列的自相似性表现得比单个个体更为突出, 在相当大的尺度上都存在; 自相似程度与演化代数密切相关, 伴随着 GA 从全局搜索向局部搜索的转变阶段呈明显下降趋势 这些发现对于深刻理解 GA 的动态运行机制、评判 GA 的运行效果以及设计更好的算法实现, 提供了新的思路和方向

参考文献(References):

[1] Zbigniew Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs* [M]. New York: Springer Verlag, 1996. 119-167.

[2] Holland J. 隐秩序[M]. 上海: 上海科技教育出版社, 2000. 75-83.

[3] 周明, 孙树栋. 遗传算法原理及应用[M]. 北京: 国防工业出版社, 1999. 18-31, 118-120.

[4] Kihong Park, Walter Willinger. *Self-similar Network Traffic and Performance Evaluation* [M]. New York: John Wiley & Sons Inc, 2000. 30-34.

[5] 郑维敏. 正反馈[M]. 北京: 清华大学出版社, 1998. 118-120.