

# 快速概率分析进化算法及其性能研究

林亚平, 杨小林

(湖南大学计算机科学系, 长沙 410082)

**摘要:** 概率分析进化算法是最近几年发展起来的一类新的构造性进化算法. 本文介绍了其产生背景和基本原理, 分析了复杂概率模型导致的计算复杂性, 在此基础上提出快速概率分析进化算法的概念和算法模型. 文中给出了 Bayes 网络作为概率模型, 爬山算法作为快速搜索技术的算法实例, 实验结果验证了快速概率进化算法具有很好的性能.

**关键词:** 进化算法; 概率模型; 快速计算

**中图分类号:** TP18 **文献标识码:** A **文章编号:** 0372-2112 (2001) 02-0178-04

## Research on Fast Evolutionary Algorithms Based on Probabilistic Models

LIN Ya-ping, YANG Xiao-lin

(Dept. of Computer Science, Hunan University, Changsha 410082, China)

**Abstract:** This paper discusses the probabilistic model evolutionary algorithms which are a class of new algorithms proposed recently. After introducing the original and the principles of these classes of algorithms, the paper analyzes the complex computation caused by the complex models. Based on this, a new fast evolutionary algorithm based on probabilistic models is proposed. Also, a fast algorithm which uses the Bayes network model combined with the climbing-hill searching is realized. The experimental results show that the new algorithm has quick, reliable and accurate performance.

**Key words:** evolutionary algorithms; probabilistic models; fast computing

### 1 引言

近年来, 一些研究者从概率统计的观点出发, 将构造性模型引入进化算法的研究, 形成一类基于概率分布的进化算法<sup>[1~4, 9~12]</sup>, 本文统一称为概率分析进化算法 (Evolutionary Algorithm based on Probability Modeling), 简称为 PMEA. 和传统的进化算法不同, PMEA 的基本思想是通过从当前优选的解集中提取信息, 然后依据这些信息建立概率分布模型, 再利用这种分布产生新的解, 如此重复, 直到满足算法的终止条件. 在进化计算领域内, 称能够快速 (quickly), 可靠 (reliably) 和精确 (accurately) 求解问题的算法为“胜任算法” (Competent Algorithm)<sup>[5]</sup>, 因此 PMEA 总的研究趋势是朝这一目标发展. 目前提出的 PEMA 算法在求解质量上, 或者说是算法的精确性和可靠性方面表现了很好的性能, 但由于算法采用的概率模型一般比较复杂, 这些算法在求解问题时引入的计算复杂性也随之增加. 当问题的规模很大时, 算法计算速度是必须考虑的因素, 即必须研究快速算法. 本文提出一种新的快速 PMEA (简称 Q-PEMA) 算法模型, 其基本思想是利用快速搜索算法搜索到好的解, 再对这些解建立概率分布, 并依据建立的分布产生新解. 新的算法保留了概率分析模型的基因连锁学习的功能, 且明显地加快了算法的速度, 具有信任算法要求的快

速、精确和可靠性质. 文中给出了爬山算法作为快速搜索技术的应用实例, 实验结果验证了快速概率进化算法具有信任算法的特点.

### 2 概率分析进化算法

#### 2.1 连锁问题与模式定理分析

由 Holland 首先提出的基本遗传算法 (GA), 是基于自然选择机制和交叉、变异等操作构成的一种优化方法. 其基本依据是“模式定理”和“构造块” (Building Block) 假设理论<sup>[8]</sup>. 由于构造块中的确定位之间存在连锁依赖关系, 而 GA 在进行交叉操作时不具备分辨和学习构造块中这种关系的能力, 不可避免地造成构造块被损坏的问题. 通常也称构造块损坏问题为连锁 (linkage) 问题<sup>[6]</sup>. 由于高阶、长距离构造块在交叉操作时的损坏概率较大, GA 在求解高阶、长距离构造块的问题时容易陷入局部最优或早熟. 为了克服 GA 因为交叉重组导致的连锁问题, 人们提出是否可以不用重组操作, 而是通过从优选的解集中提取信息, 然后利用这种信息建立合适的概率分布, 再利用这种概率分布产生新的解, 由此避免构造块破坏, 达到扩大搜索解空间范围的目的. 这种将具有连锁学习的构造性概率统计方法引入进化计算的思想产生了概率分析进化算法.

收稿日期: 2000-05-23; 修回日期: 2000-10-28

基金项目: 湖南省自然科学基金 (No. 99JJ Y20060)

本文下面应用概率论方法分析模式定理. 注意到本文讨论的 PMEAs 算法不考虑交叉和变异算子, 因此只需分析选择算子对算法性能的影响. 对于采用比例选择的基本遗传算法, 第  $t+1$  代的新点可用下面的概率分布产生:

$$p(X, t+1) = p(X, t)f(X)/F(t) \quad (1)$$

上式中  $X = (x_1, x_2, \dots, x_n)$  表示基因随机变量向量, 其中的  $n$  表示结构(染色体)的长度,  $p(X, t)$  表示  $X$  在第  $t$  代的分布;  $f(X)$  表示  $X$  的适应度; 而  $F(t)$  表示第  $t$  代群体平均适应度. 即有:

$$P(X, t+1) = \sum_X p(X, t)f(X) \quad (2)$$

式(2)中的求和是对群体中所有  $X$  求和. 因为式(1)计算直观上涉及的样点有  $2^n$  个, 计算复杂性为指数级. 因此需要寻找近似式(1), 但计算复杂性为多项式时间复杂度的方法求解. 事实上, GA 和其它的一些试探性算法都是在做这种努力.

本文利用一种显式的概率分析方法讨论上述问题. 若给定一个模式  $H$ , 按照式(1)对所有匹配  $H$  的  $X$  求和, 得到:

$$p(H, t+1) = \sum_{X \in H} p(X, t)f(X)/F(t) \quad (3)$$

按照概率论的术语, 式(3)是求一个边缘分布. 因此, 可以用概率分布的方法探讨下述问题: 是否存在某些边缘分布, 它们可以很好地近似式(1), 且可以在多项式时间内计算.

## 2.2 概率分析进化算法

PMEA 算法的基本思想是从当前解空间抽取信息, 然后对这些信息的分布进行评价, 在此基础上再产生新的群体. 本质上, 这类算法改变了传统 GA 通过重组操作产生群体的途径, 由此改善传统 GA 存在的连锁问题. PMEAs 算法的难点是如何评价分布, 不同的评价方法形成了一些不同的算法, 依据算法处理变量连锁依赖关系的能力, 可将这些算法分为单变量边缘分布算法<sup>[3, 10]</sup>, 两变量边缘分布算法<sup>[2, 4]</sup>和多变量边缘分布算法. 多变量边缘分布算法试图计算所有变量之间的连锁依赖关系, 代表性的算法有 Bayes 优化算法(BOA)<sup>[11]</sup>, 扩展压缩遗传算法(ECGA)<sup>[9]</sup>和比例分布算法(FDA)等<sup>[12]</sup>. 由于 Bayes 网络具有许多特点, 本文将在下一节重点讨论这一算法. 尽管 Bayes 等多变量算法可以有效地求解高阶构造块问题, 但由于模型的复杂性, 其引入的计算量很大, 因此当问题规模增大时, 需要寻找快速的算法简化其计算. PMEAs 算法求解问题的基本框架可用算法 1 描述如下:

### 算法 1: PMEAs 算法

- (1)  $t \leftarrow 0$ , 随机产生初始群体  $P(t)$ ;
- (2) 利用某种选择机制从  $P(t)$  中选择部分好的解构成  $S(t)$ ;
- (3) 用一种概率模型建立  $S(t)$  的概率分布;
- (4) 依据概率分布产生新的解集  $O(t)$ ;
- (5) 利用  $O(t)$  替换  $P(t)$  中的某些解构成  $P(t+1)$ ;
- (6) 若终止条件不满足,  $t \leftarrow t+1$ , 转 2;
- (7) 群体  $P(t)$  为所求解.

## 3 快速概率分析进化算法

### 3.1 快速概率分析进化算法

本文提出一种新的快速概率分析进化算法. 基本思想是利用快速搜索算法搜索好解, 然后建立概率分析模型, 概率分布模型又为快速搜索算法提供开始点, 如此循环, 以达到快速、可靠和精确求解问题的目的. 算法 2 给出了快速概率分析进化算法的基本步骤:

### 算法 2: 快速概率分析进化算法

- (1)  $t \leftarrow 0$ , 随机产生第一代  $P(t)$ ;
- (2) 利用块选择从  $P(t)$  中选取适应度最高的子集  $S(t)$ ;
- (3) 用一种概率模型建立  $S(t)$  的概率分布;
- (4) 利用概率分布产生  $K$  个解;
- (5) 以前述  $K$  个解中的最好解为基础, 执行快速搜索算法, 得到新的  $M$  个新解;
- (6) 用  $M$  个新解替换  $P(t)$  中的部分解, 构成  $P(t+1)$ ;
- (7) 如果未达到终止条件,  $t \leftarrow t+1$  转 2;
- (8) 最后的群体为所求解.

与文献中其他一些混合算法不同, 本文提出的快速 PMEAs 算法中的快速搜索不是一种分离的预处理手段, 而是将快速搜索和概率分析在循环体内紧密结合, 即快速搜索为概率分析提供建立概率分布的样点; 用概率分布生成的新点为快速搜索提供好的起始点. 假定算法中当前群体为  $P(t)$ , 采用的概率分布模型为  $p(x_1, x_2, \dots, x_n)$ , 则  $p$  用于产生一系列的候选解, 其中最好的解用来初始化快速搜索算法; 快速搜索算法终止后, 产生的  $M$  个新解用于替换  $P(t)$ .

算法中  $M$  的选择对算法性能具有重要影响,  $M$  越大, 则算法给快速搜索算法的权重越大, 容易导致算法早熟逼近; 反之, 算法给概率分布算法的权重太多, 导致算法计算时间增加.

为了研究快速 PMEAs 算法的性能, 本文利用 Bayes 网络作为基本概率模型, 采用爬山法作为快速搜索算法. Bayes 网络是一无环有向图, 它即可以对所给的数据进行描述, 又可以产生与所给定数据性质相同的数据, 因此常用于对离散或连续变量的多项式数据建模<sup>[7]</sup>. Bayes 网络图中一个结点代表一个变量; 两个结点之间的边代表变量之间的关系, 若结点  $x_i$  存在一条指向  $x_j$  的边, 则称  $x_i$  是  $x_j$  的父结点. 上述 Bayes 网络可以用数学公式表示如下:

$$p(X) = \prod_{i=1}^n p(x_i | \text{pa}(x_i)) \quad (4)$$

式(4)中  $X = (x_1, x_2, \dots, x_n)$  是随机变量向量,  $\text{pa}(x_i)$  表示结点  $x_i$  的父结点(即有边指向  $x_i$  的结点)集合;  $p(x_i | \text{pa}(x_i))$  是  $x_i$  的条件概率;  $p(X)$  是联合概率分布. 由 Bayes 规则可知:

$$p(x_i | \text{pa}(x_i)) = p(x_i, \text{pa}(x_i)) / p(\text{pa}(x_i)) \quad (5)$$

在建立 Bayes 网络的过程中, 用给定的数据和式(5)就可以得到各变量的条件概率. 利用 Bayes 网络求解问题的关键在于学习网络, 学习网络指的是找到一个网络, 使之在约定的评价标准下, 最好地匹配现有训练数据集. 网络的评价标准说明网络所分析的数据的好坏程度, 也就是对网络质量的评价; 搜索过程则是要寻找具有最高评价标准值的网络. BOA 算法利用 BD 评价标准和贪婪算法(已经证明, 当网络的入度大于

1 时,搜索网络是一 NP 难题)建立和搜索网络<sup>[7,11]</sup>. BOA 算法求解许多测试函数,包含高阶构造快问题都得到了很好的优化结果,但算法的复杂程度也随之增加. 实际计算量很大. 因此结合快速搜索算法,使得 BOA 算法不频繁地调用,将实质性地降低计算成本.

算法中使用的爬山算法具有以下特点:算法中允许新解的适应度值高于或等于原起始点的适应度,其原因是对于结构复杂的搜索空间,需要这种平移搜索.

### 3.2 实验结果

利用前述的快速 PMEA 算法,本文对以下几个有代表性的函数进行了测试(下面三个函数中变量  $\mu$  表示输入字符串中 1 的个数):

(1) 三阶欺骗函数:

$$f(u) = \begin{cases} 0.9, & u = 0 \\ 0, & u = 1 \\ 0, & u = 2 \\ 1, & \text{其它} \end{cases} \quad (6)$$

(2) 五阶欺骗函数:

$$f(u) = \begin{cases} 5 - u, & u < 5 \\ 5, & \text{其它} \end{cases} \quad (7)$$

(3) 六阶双极函数:

$$f(u) = \begin{cases} 0.9, & u = 3 \\ 0.8, & u = 2, u = 4 \\ 0, & u = 1, u = 5 \\ 1, & u = 0, u = 6 \end{cases} \quad (8)$$

上述函数都是由  $k$  阶单一函数 (unitation function, 即函数值只与输入字符串中 1 的个数有关的函数) 构成的复杂叠加可分解函数 (additively decomposable function)<sup>[12]</sup>, 这些函数属于 GA 难的函数. 也就是说用 GA 求解时, 由于所有低于  $k$  阶的模式容易误导 GA 远离全局最优值而陷入局部最优. 因此也称这类函数为欺骗函数. 欺骗函数表明染色体中的基因之间具有紧密的连锁关系, 利用欺骗函数可以评价函数某个基因在对函数适应度产生贡献时, 依赖其它基因的程度. 反之, 高度的连锁性问题意味着问题具有欺骗性. 当这种问题规模扩大时, 为了获得一个特定的解, GA 求解的计算复杂性呈指数增长. 如果上述函数中考虑子函数间的变量相互重叠, 则问题的复杂性更高. 本文主要考虑非重叠的叠加可分解函数, 但作为一种初步的测试, 给出了三阶欺骗函数中, 相邻子函数存在的一位重叠函数情况的测试结果. 例如, 对于 6 位的非重叠三阶欺骗函数, 我们可以构造其编码为  $x_1 x_2 x_3 x_4 x_5 x_6$ , 其适应度的计算依据式 (7), 分为两个子函数分别计算, 然后将其结果相加. 因此计算可用下式表示:

$$g(u) = f(u_1) + f(u_2) \quad (9)$$

其中  $\mu_1$  和  $\mu_2$  分别表示  $x_1 x_2 x_3$  和  $x_4 x_5 x_6$  字符串中 1 的个数.

当考虑一位重叠问题时, 上述函数的编码应有 7 位, 其计算可用下式表示:

$$g(u) = f(u_1) + f(u_2) + f(u_3) + f(u_4) \quad (10)$$

其中  $\mu_1, \mu_2, \mu_3$  和  $\mu_4$  分别表示  $x_1 x_2 x_3, x_3 x_4 x_5, x_4 x_5 x_6$  和

$x_5 x_6 x_7$  字符串中 1 的个数. 测试结果如表 1 所示.

表 1 Q-PMEA 与 BOA 实验数据比较

函数 (规模, 最优值)	算 法		BOA		Q-PMEA	
			最优值	评价次数	最优值	评价次数
三阶欺骗函数	$n = 30$	10.0	10.0	8500	10.0	6510
	$n = 60$	20.0	20.0	27300	20.0	25200
	$n = 90$	30.0	30.0	57000	30.0	45300
三阶重叠欺骗函数	$n = 31$	15.0	15.0	19950	15.0	14710
	$n = 61$	30.0	30.0	59500	30.0	40270
	$n = 91$	45.0	45.0	86800	45.0	76120
五阶陷阱函数	$n = 30$	30.0	30.0	14300	30.0	7150
	$n = 60$	60.0	60.0	41250	60.0	39620
	$n = 90$	90.0	90.0	75450	90.0	67620
六阶双极函数	$n = 30$	15.0	15.0	9000	15.0	3150
	$n = 60$	60.0	10.0	36000	9.4 *	13010
	$n = 90$	90.0	15.0	45900	14.0 *	24310

表中给出了函数实际最优值和算法得到的最优值, 以及得到算法最优值所花费的适应度函数评价次数. 为了便于比较, 表中给出了 BOA 算法相应的求解结果. 对于同一函数给出了三种不同规模的实验结果 (用  $n$  表示问题的规模, 即函数输入字符串的长度). 算法中选用的其他参数如下: 对于爬山算法,  $M$  选择为群体规模的 10%;  $K$  选择为 100; 对于 Bayes 概率模型, Q-PMEA 与 BOA 算法采用相同的群体规模和评价模型. 从实验和表 1 可以得到以下结论:

(1) 和 BOA 算法类似, 随着问题规模的增大, Q-PMEA 可以线性地逼近最优值, 除了个别情况 (见表中 \* 号标识), Q-PMEA 都达到了最优值. 这是 GA 算法所不能达到的, 说明 Q-PMEA 保证了概率分析进化算法的特点;

(2) 比较 BOA 算法, Q-PMEA 逼近解所需要的函数评价次数较少, 引入的计算量 (实际运行时间) 明显减少, 具有快速逼近的特点;

(3) 对于三阶重叠欺骗函数 Q-PMEA 表现了很好的性能, 这一现象值得做进一步的研究. 实验结果验证了本文提出的快速算法的思想.

## 4 结语

本文讨论快速概率分析进化算法, 这种算法的特点是把自然进化算法和构造性数学分析方法结合, 以指导对问题空间的有效搜索. 在介绍了概率分析进化算法的产生背景和基本原理的基础上, 提出了快速概率分析进化算法的概念和算法模型. 文中给出了 Bayes 网络作为概率模型, 爬山算法作为快速搜索技术的算法实例, 对有代表性的几种欺骗函数进行了测试, 实验结果验证了快速概率分析进化算法具有胜任算法的性能.

## 参考文献:

- [1] Pelikan. M et al. A survey of optimization by building and using probabilistic models [R]. Illi GAL Report No. 99018, UIUC, Illi GAL Genetic Algorithm Laboratory, 1999.
- [2] Baluia S, Davies S. Using optimal dependency-trees for combinatorial

- optimization learning the structure of the search space [A]. Proc. of the 14th International Conference on Machine Learning [C], Morgan Kaufmann, 1997:30 - 38.
- [ 3 ] Baluia S. Population-based incremental learning :a method for integrating genetic search based function optimization and competitive learning [R]. CMU Technical Report No. CMU-CS-94-163, 1994.
- [ 4 ] Pelikan M. The bivariate marginal distribution algorithm [A]. Advanced in Soft Computing: Engineering Design and Manufacturing [C]. London :Springer-Verlag, 1998 :521 - 535.
- [ 5 ] Goldberg D E. A meditation on the application of genetic algorithms [R]. IlliGAL Report No. 98003, UIUC, IlliGAL Genetic Algorithm Laboratory, 1998.
- [ 6 ] .Harik G. Goldberg D E. Linkage learning [A]. Foundations of Genetic Algorithms 4 [C]. 1996 :247 - 262.
- [ 7 ] Heckerman D. A Tutorial on learning with bayesian networks ,microsoft Research Advanced Technology Division [R]. Microsoft Technical Report No. MSR-TR-95-06, 1995.
- [ 8 ] Holland J. H. Adaptation in Natural and Artificial Systems [M]. Ann Arbor, MI. Univ. Michigan Press, 1975.
- [ 9 ] Harik G. Linkage learning via probabilistic modeling in the ECGA [R]. IlliGAL Report No. 99010, UIUC, IlliGAL Genetic Algorithm Laboratory, 1999.
- [ 10 ] Muehlenbein H. The equation for response to selection and its use for prediction [J]. Evolutionary Computation, 1997 :5 (3) :303 - 346.
- [ 11 ] Pelikan M. Goldberg D E. BOA: The Bayesian Optimization Algorithm [R]. IlliGAL Report No. 98013, UIUC, IlliGAL Genetic Algorithm Laboratory, 1998.
- [ 12 ] Muehlenbein H. Convergence theory and application of factorized distribution algorithm [R]. GMD-AS-98-03, 1998.

### 作者简介:



林亚平 1955 年生, 1982 年和 1985 年分别获湖南大学学士学位和国防科技大学硕士学位。现为湖南大学计算机科学系教授, 主要从事计算机网络和进化计算领域的研究。

杨小林 1961 年生, 1996 年获湖南大学硕士学位。现为湖南大学计算机科学系讲师, 主要从事计算机应用方面的教学和科研工作。

## 征订启事

由《电子学报》编辑部编辑出版的《第三代移动通信与网络技术专辑》, 已于 2000 年 12 月出版了。有对本专辑感兴趣的作者、图书馆, 请速与编辑部王辉联系征订事宜。本专辑每册 20 元, 来信请写明地址、邮编和数量。

### 学术论文

信道估计误差与估计长度关系的频域分析  
衰落信道下基于线性预测的迭代解调及译码技术: Turbo DPSK  
Ricean 衰落信道下一种正交多载波 CDMA 系统的性能  
一种基于一阶扰动自适应特征值分解的多用户检测算法  
多径 CDMA 信道下的盲空时多用户检测  
宽带 CDMA 通信的主动式多用户动态检测方法  
一种新型的多用户检测算法  
TCP 在不可靠网络环境中的性能分析及其服务质量控制  
一种区分服务综合方案的模型与性能分析  
一种自适应位置更新方案  
3G IMT-2000 的流动性管理分析  
一个 2D-Mesh 上的完全自适应路由器设计  
一种支持多媒体通信 QoS 的拥塞控制机制  
BICM 在 DS/CDMA 中的应用  
一种新的抗远近效应的盲 DS-CDMA 接收机  
空时发射分集 CDMA 的多径分集接收  
WCDMA 下行信道 RAKE 接收机的性能分析  
一种自适应阶距的自动频率校正算法  
一类新的性能优异的伪随机序列—GMW 相控序列  
随机排序的优化算法  
基于双向逐幸存路径处理的空时格形码软判决自适应解码  
应用于在线测试的状态判定算法  
基于参数估计的随机早期探测 (RED) 改进算法

WCDMA 中两种天线闭环发分集模式的性能分析  
基于独立分量分析的 DS-CDMA 系统接收机  
CDMA 蜂窝系统中基于模糊神经网络的反向链路功率控制

### 综述评论

高性能交换体系结构及其调度算法分析  
空时码  
CDMA 蜂窝移动通信系统中的联合功率控制

### 科研通信

用于 WCDMA 编码复接的一种新的算法  
高性能安全路由器中快速路由查找算法的研究与实现  
基于 MLS 的流量工程分布实时网络承载能力估计与分配模型  
近似镜像网页检测算法的研究与评价  
一种支持带宽预约的 crossbar 调度算法  
综合网管系统的设计方法及应用实例  
基于 X.500 目录服务实现网络即插即用  
基于双窗口测量的蜂窝移动通信系统越区切换算法性能分析  
传输分集技术在第三代移动通信系统中的应用  
一种 CDMA 蜂窝通信系统中的功率控制算法  
第三代移动通信系统中通用高速维特比译码器的设计与实现  
一种新的自适应并行干扰抵消接收机  
一种新的 CDMA2000 多载波解调方法  
一种近似解相关多用户检测器的性能分析  
Bluetooth 系统基带关键算法的研究及其仿真