快速二变量边缘分布算法及其应用研究*)

Research on the Fast Algorithm of the Bivariate Marginal Distribution and it's Application

杨小林

(湖南大学计算机系　长沙410082)

Abstract This paper discusses the fast calculation problem of the bivariate marginal distribution algorithm (BMDA). A fast BMDA is proposed. An experimental case with the multi-constraints Knapsack NP-hard problem is solved using the algorithm. The results show that the algorithm has quick and accurate performance.

Keywords Evolutionary algorithm, Probabilistic model, Bivariate marginal distribution, Knapsack NP-hard problem

1. 引言

近年来,一些研究者从统计学的观点出发,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[1~3],文献中也称这类算法为分布评价算法(EDA),概率分析构造遗传算法(PMBGA)等名称,本文统一称之为概率分析进化算法,简称为 PMEA(Evolutionary Algorithm based on Probability Modeling)。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合中提取信息,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。

依据采用概率模型的复杂性,PMEA 算法可分为一阶、二阶和高阶算法[4],本文着重讨论两变量边缘分布算法[5](Bivariate Marginal Distribution Algorithm, BMDA)的快速计算问题,提出一种适合 BMDA 的快速算法,并利用这种快速算法求解多约束背包问题,实验结果表明该算法快速、精确。

2. 概率分析进化算法

为了克服基本遗传算法因为交叉重组导致的构造块破坏问题,人们提出是否可以不采用重组操作,而是通过从优选的解集合中提取信息,然后利用这种信息构造概率统计模型,再利用这种概率统计模型产生新的解。这一思想形成了概率分析进化算法。PMEA 算法的概念可以追溯到 Ackley 于1987年提出的一个学习算法[6],在此算法中他引入了一个具有重要意义的数据结构即基因向量,通过来自群体的正负反馈信息操作此向量。文中有一个直观的比拟,即把长度为 n 的基因向量比作由 n 个成员组成的政府,群体数据比作选民的投票结果,依据此结果评判群体(选民)对向量(政府)的满意程度。这一算法实际上应用了群体的概率分布分析和评价解的质量。令 $X=(X_1,X_2,\cdots,X_n)$ 表示群体的染色体,其中的随机变量 X_i 对应染色体中的第 i 个基因位,一般而言,X 的概率分布应为各变量的联合分布,即:

$$P(X)=P(X_1|X_2,\cdots,X_n)P(X_2|X_1,X_3,\cdots,X_n)\cdots P(X_n|X_1,\cdots,X_{n-1}) \quad (1)$$

显然,前述的 Ackley 算法实际上是假定基因变量彼此独立,在此条件下(1)式变为:

$$P(X_1,X_2,\cdots,X_n)=\Pi P(X_i) \quad (2)$$

其中 $P(X_i)$ 表示基因变量 X_i 的边缘分布,算法通过统计样点(群体的染色体)基因位的频率可以计算上述分布,然后利用此分布生成新的群体。我们将这种假定基因变量彼此独立的 PMEA 算法,称为一阶 PMEA 算法。一阶 PMEA 在求解线性问题(如 OneMax 问题)或一阶构造块问题上具有和 GA 相当或更好的效率和质量,且实现简单。但对于变量间具有相互作用的高阶构造块问题,这些算法在求解问题时,其群体规模和算法所需的函数评价次数随问题规模呈指数增长。

若考虑两个变量之间的依赖关系,则(1)式可简化如下:

$$P'(X)=\Pi P(X_{m(i)}|X_{m(R(i))}) \quad (3)$$

上式中,$m=(m(1),m(2),\cdots,m(n))$ 是整数集合(1,2,…,n)的一个未知排列,$m(1)\leqslant R(i)<m(i)$。显然,由(3)式决定的概率分布是二阶概率分布,因此需要寻找合适的二阶概率分布模型予以评价,算法的目的是寻找最佳的排列,使得 P'(X)最接近 P(X)。注意到(3)式的约束条件,可知二阶概率分布对应的图结构是树。目前提出的二阶概率分布算法主要有 MIMIC[7]、依赖树[8]和 BMDA[5]算法。二阶 PMEA 求解问题的能力强于一阶 PMEA,但对于更高阶的问题,需要能够描述多变量依赖关系的概率模型。自然地,人们想到可以用更为复杂的图结构来表示变量间的相关性。Bayes 优化算法(Bayesian Optimal Algorithm, BOA)便是在这种背景下产生的[9]。Bayes 网络适合求解多变量依赖关系的问题,但其概率模型也更为复杂。PMEA 算法随着概率模型的精确性提高,其求解问题的能力也随之增强,但由于复杂的概率模型,其求解问题引入的计算也明显增加。在进化计算领域内,称能够快速、精确和可靠求解问题的算法为胜任算法(Competent algorithm),研究胜任算法是进化计算的主要目的[11]。

3. BMDA 算法

BMDA 算法采用的模型是 PeasonX 平方统计,即计算两两变量 X_i 和 X_j 之间的 X 平方:

$$X^2(X_i,X_j)=\Sigma\frac{(Np(X_i,X_j)-NpX_i)p(X_j))^2}{Np(X_i)P(X_j)} \quad (4)$$

上式中的 N 表示样本空间的大小。$p(X_i,X_j)$ 和 $p(X_i)$ 分别表示 X_i 和 X_j 的联合分布和边缘分布。BMDA 利用(4)式计算的结果构造最大生成树,得到两两变量之间的最大依赖关系。注意上式中的和式是对所有 X_i 和 X_j 的组合求和,如果 $X^2(X_i,$

*)本文研究得到湖南省自然科学基金资助。

$X_j)<3.84$，则 X_i 和 X_j 的不相关性为95%[51]。因此，算法在实现上利用了这一性质，构成的实际上是森林，森林中的每一棵树是由 $X^2(X_i,X_j)$ 不小于3.84的相关结点构成的最大生成树。

依据(4)式，可以构造二变量边缘分布算法，其基本思想是利用(4)式发现基因变量两两之间的相关性，这种相关性通常利用一种依赖图表示，图中的结点对应一个基因变量。初始时，图为空，然后，选择任意一个结点 X_0 加入图。在此基础上，选择与图中已有结点相关性最大(即选择 $X^2(X_0,X_j)$ 的 X_j 结点加入图，并在这两个结点之间建立一条边，其过程类似最大生成树的构造，直到所有结点加入到图。显然这种方法保证生成的依赖图是无环的，而且所有边结点之间的 X^2 统计值最大。

利用第 t 代群体构造的依赖图，便可以产生第 t+1代群体。对于 BMDA 算法，依据森林中的每一棵树进行操作。首先根据树中根结点(不妨假定为 X_i)的边缘分布 $p(X_i)$ 产生 X_i，然后依据条件概率计算 X_j 对应的子结点变量 X_j 之值：

$$p(X_j|X_i)=p(X_i,X_j)/p(X_i) \tag{5}$$

依次类推，可以产生第 $t+1$ 代群体得新个体。具体的 BMDA 算法描述如下：

算法1 BMDA 算法

1. $t\leftarrow 0$，随机生成群体 $P(t)$；
2. 从 $P(t)$ 中选择子集 $S(t)$；
3. 计算 $S(t)$ 的单变量频率 $p(X_i)$ 和二变量频率 $p(X_i,X_j)$；
4. 依据 $p(X_i)$ 和 $p(X_i,X_j)$ 计算 $X^2(X_i,X_j)$ 统计值；
5. 依据 $X^2(X_i,X_j)$ 统计值，构造子树为最大生成树的森林；
6. 利用最大生成树和条件概率产生 $P(t+1)$ 代群体；
7. 若不满足终止条件，$t\leftarrow t+1$，转2；
8. 最后群体为所求解。

分析上述基本算法，可以发现其计算量主要体现在每一代中的概率模型计算，即算法都要统计群体的单变量边缘分布和两变量的联合分布，对于问题规模为 n，群体规模为 N 的算法，这两种统计量分别是 n 和 $(n-1)/2$。在假定二元染色体的条件下，对于两变量联合分布，其组合分别有四种(00,01,10,11)，因此至少需要统计三种统计量(依据全概率公式，第四种可以直接计算)。由此可知其计算量很大，加上算法中建立最大生成树的计算以及利用树结构生成新的群体的计算，算法计算量在问题规模和群体规模增大时，其计算复杂性随之明显增大。然而，这些计算都具有可分解计算的特点，即大量的计算是统计求和。利用这一特点，可以进行算法上的改进，加快算法的速度。

对于基本遗传算法，第 $t+1$ 代群体一般是在第 t 代群体的基础上通过选择、交叉、变异或其他操作构成的，因此在第 $t+1$ 代群体中保留第 t 代群体的部分优秀个体是合理的。令 $P_i(t)$ 表示第 t 代群体第 i 个基因的边缘分布频率；$P_{ij}(t)$ 表示第 t 代群体第 i 个和第 j 个基因的联合分布的频率；类似地定义 $Q_i(t)$ 和 $Q_{ij}(t)$ 分别为第 t 代新生成样点的第 i 个基因边缘分布频率，第 i 个和第 j 个基因的联合分布的频率。定义：

$$P_i(t+1)=P_i(t)+\alpha(Q_{ij}(t)-P_i(t)) \tag{6}$$

$$P_{ij}(t+1)=P_{ij}(t)+\beta(Q_{ij}(t)-P_{ij}(t)) \tag{7}$$

上式中 α,β 为大于0小于1的因子。利用(6)和(7)表示的群体构成形式，结合 BMDA 算法可分解计算的特点，可以对算法进行改进，减少算法的计算量。若令 $\alpha=\beta=0.5$，即新的群体保留旧群体的一半个体，则前述的计算可以减少一半。由此，我们得到以下的快速 BMDA 算法：

算法2 快速 BMDA 算法

1. $t\leftarrow 0$，随机生成群体 $P(t)$；
2. 从 $P(t)$ 中选择最好的 M 个个体，形成子集 $S(t)$；
3. 计算 $S(t)$ 的单变量频率 $p(X_i)$ 和二变量频率 $p(X_i,X_j)$；
4. 依据 $p(X_i)$ 和 $p(X_i,X_j)$ 计算 $X^2(X_i,X_j)$ 统计值；
5. 依据 $X^2(X_i,X_j)$ 统计值，构造最大生成树的森林；
6. 利用最大生成树和条件概率，产生 K 个新的个体，形成子集 $O(t)$，利用 $S(t)$ 和 $O(t)$ 形成 $P(t+1)$ 代群体；
7. 若不满足终止条件，$t\leftarrow t+1$，转2；
8. 最后群体为所求解。

在第 $t+1$ 代计算单变量的边缘分布和两变量的联合分布时，算法只需要统计新生成的 K 个样点的对应统计频度即 $Q_i(t)$ 和 $Q_{ij}(t)$，这些统计值构成整个群体的统计值。若 $K=M=N/2$，算法可减少一半的计算量。

4. 快速 BMDA 算法求解背包问题性能分析

本节讨论快速 BMDA 算法求解多约束背包问题的性能。多约束背包问题描述如下：给定 n 个物体和一个背包，背包具有 m 种不同的测量方式，第 i 种测量方式下的容量 b_i，每个物体针对 m 种测量方式有 m 种权值，记第 j 个物体在第 i 种测量方式下的权值为 $w_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$，如果物体 j 可以放进背包(必须满足 m 种测量方式)，则可获得利润 p_j。问题是求得某一组合方式，使得总利润最大。上述背包问题可以形式地表示为求向量 $X(X_1,X_2,\cdots,X_n)$，使得：

$$Z=\max\sum_{j=1}^{n}p_jX_j \tag{8}$$

并满足：

$$\sum_{j=1}^{n}w_{ij}X_j\leq b_i \tag{9}$$

不失一般性，设 p_j、w_{ij} 和 b_i 都为非负整数。为了便于测试算法，令 $p_j=w_{ij}$，X_j 随机取0或1，其中，若物体 j 放进背包则 $X_j=1$，否则 $X_j=0$。并初始化背包容量为：

$$b_i=\sum_{j=1}^{n}w_{ij}X_j \tag{10}$$

则可以预先得知问题的最优解为 b_i。可以证明上述背包问题是 NP 难的[10]。

由于多约束背包问题存在(9)式的约束条件，利用基本遗传算法求解此问题难以收敛，因为当群体在进化时，交叉算子十分容易破坏现有的构造块而导致群体逼近无效解。而概率分析进化算法是通过分析基因之间的相关性进化，因此群体逼近时具有比较好的稳定性。

本文利用快速 BMDA 算法对上述背包问题求解，有关算法的其它参数描述如下：

1)解的表示(即染色体结构)：为固定长度的二元字串，对应向量为 $(X_1,X_2,\cdots,X_n)$；

2)群体规模：随问题规模变化，一般在200-300之间；

3)选择机制(快速 BMDA 算法步骤2)：采用块选择(truncation selection)，该选择方法选择当前群体中最好的部分；

4)K 值选择为群体规模的一半；

5)适应度函数:由(9)式约束下的(8)式确定。

实验中,取约束条件值 m 为10;问题规模 n 从10到100以增量10变化;w_{ij}取0-500之间的均匀分布随机数。实验结果如表1所示:

表1 快速 BMDA 求解多约束背包问题性能

问题规模(n)	实际最优值	算法求解值	BMDA 运行时间(秒)	快速 BMDA 运行时间(秒)
10	2040	2040	0.49	0.28
20	2851	2851	7.69	5.17
30	3040	3040	13.08	3.79
40	4465	4465	22.96	6.37
50	6729	6729	53.72	15.16
60	8206	8206	53.06	53.17
70	9437	9437	111.23	96.95
80	10513	10513	190.65	110.22
90	12062	12062	317.57	126.33
100	13179	12775	362.18	165.22

从表1可知,快速 BMDA 求解此问题具有很好的性能,除问题规模为100外,其它情况下,算法都得到了最优解,与文[10]利用改进的遗传算法求解多约束问题结果比较,本文得到了更好的结果。说明了利用 BMDA 概率模型算法求解此类问题的精确性。

另一种评价进化算法求解问题性能的指标是算法执行过程所经历的函数评价次数[1],这种评价比简单地给出算法经历的进化代数更有说服力,因为它综合考虑了群体规模和进化代数两种因素。图1给出了快速 BMDA 求解问题规模和函数评价次数之间的关系,可以看出,随着问题规模的增加,函数评价次数基本上呈一种线性关系,而不是一种指数关系。

表1给出了快速 BMDA 和 BMDA 在相同硬件环境下求解问题的速度,采用快速 BMDA 显著减少了求解问题的计算复杂性,从而提高了算法的运行速度。应该指出这种时间的统计只是针对两种算法同等条件进行比较,因此只具有相对比较意义,由于不同的硬件和软件条件,难以和其它算法比较。

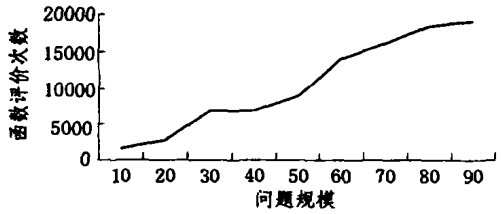

图1 快速 BMDA 求解多约束背包问题的函数评价次数

结束语 针对基本遗传算法存在的连锁问题,本文讨论了基于概率分析模型的遗传算法,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。本文结合 BMDA 算法讨论了二阶概率分析进化算法的快速计算问题,提出了一种快速 BMDA 算法,文中分析了快速计算的原理,并利用这种快速算法对多约束背包问题进行了求解,结果表明,该算法具有胜任进化算法所要求的精确性和快速性。本文认为在这一方面还有许多值得进一步探讨的问题,包括研究高阶概率进化算法的快速计算问题;研究如何将现有的一些降阶算法引入概率分析进化算法,使得可以利用低阶概率分析进化算法求解高阶问题,达到快速、精确和可靠地求解问题的目的。

参考文献

1 Muehlenbein H. From recombination of genes to the estimation of distributions I. Binary parameters. Parallel Problem Solving from Nature, 1996, 4: 178～187

2 Baluia S. Fast probabilistic models for combinatorial optimization. In: Proc. of the 15th National Conf. on Artificial Intelligence (AAA-98) AAA Press, 1998. 469～476

3 Pelikan M, Muehlenbein H. Marginal distributions in evolutionary algorithms. http://www-illigal.ge.uiuc.edu/~Pelikan/

4 林亚平. 概率进化算法及其研究进展. 计算机研究与发展, 2001, 38(1): 43～49

5 Pelikan M. The bivariate marginal distribution algorithm. Advanced in soft computing-Engineering Design and Manufacturing, London: Springer-Verlag, 1998. 521～535

6 Ackley D H. A Connectionist Machine for Genetic Hill climbing. Boston: Kluwer Academic, 1987

7 Bonet J D, et al. MIMIC: Finding Optima by Estimating Probability Densities, Advances in Neural Information Processing System. The MIT Press, Cambridge, 1997

8 Baluia S, Davies S. Using optimal dependency-trees for combinatorial optimization learning the structure of the search space. In: Proc of the 14th Intl. Conf. on Machine Learning. Morgan Kaufmann, 1997. 30～38

9 Pelikan M, Goldberg D E. BOA: The Bayesian Optimization Algorithm: [IlliGAL Report No. 98013]. UIUC, IlliGAL Genetic Algorithm Aboratory, 1998

10 Lin F T, et al. Applying the Genetic Approach to Simulated Annealing in Solving Some NP-Hard Programs. IEEE Trans. on Systems, MAN and Cybernetics, 1993, 23(6): 1752～1767

11 林亚平, 杨小林. 快速概率分析进化算法及其性能研究. 电子学报, 2001, 29(2): 178～181

(上接第113页)

4 Page L B, Perry J E. Reliability of directed networks using the factoring theorem. IEEE Trans Reliability, 1989, 39(5): 556～561

5 Satyanarayana A, Chang M K. Network reliability and factoring theorem. Networks, 1983, 13: 107～120

6 Ke W J, Wang S D. Reliability evaluation for Distributed Computing Networks with imperfect nodes. IEEE Trans on Reliability, 1997, 46(3): 342～349

7 Kumer A, Agrawal D P. A generalized algorithm for evaluating distributed program reliability. IEEE Trans Reliability, 1993, 42(3): 416～426

8 Wood R K, Factoring algorithm for computing K-terminal network reliability. IEEE Trans Reliability, 1986, 35(3): 269～278

9 Factoring and reductions for networks with imperfect vertices. IEEE Trans Reliability. 1991, 40(2): 210～217