

基因学习算法及其在集合覆盖问题中的应用

Gene Learning Algorithm and Application on Set Covering Problem

金炳亮

022

(浙江师范大学计算机系 浙江金华321004)

Abstract A new algorithm called Gene Learning (GL) which is developed from PBIL and Selfish Algorithm is presented in this paper. GL begin with the initial alleles probability of every gene location and followed by a series of repeated processes of sampling, selecting and local search, probability learning until the final result is obtained. Application on Set Covering Problem (SCP) is presented in this paper.

Keywords Evolutionary computation, Gene Learning algorithm (GL), Set Covering Problem (SCP)

在用传统方法解决一些复杂而规模较大的组合优化问题,尤其是NP难题,出现困难时,一些近似算法相继推出。启发式搜索法、模拟退火算法及进化算法等的出现,为解决这些优化问题提供了非常好的手段。近年来,出现了一种概率学习的进化计算模型,如Baluja的PBIL算法^[1]与Corno的自私基因算法^[2]。概率学习的进化计算模型通过不断地学习每一代的最优个体,最终收敛于最优或较优的解的等位基因概率,其过程描述如下:

- 1 初始化等位基因概率
- 2 While(进化计算的结束条件=false)
- 3 通过基因重组产生一个评价群体
- 4 比较评价群体目标函数值,选取最优个体
- 5 用最优个体修正等位基因概率
- 6 End

本文在此概率学习进化模型的基础上提出一个新的算法:基因学习算法。该算法在原概率学习算法的基础上,定义了适应范围更广的等位基因结构;在最优个体获取方法中引进了局部搜索算法,解决了原算法局部搜索能力差的缺点;在解决规模更大的组合化问题中使用了一个概率加速算法,使该算法的收敛性得到较好地解决。本文将基因学习算法用于求解集合覆盖问题,并获得较为满意的计算结果。

1. 基因学习算法原理

1.1 等位基因结构、概率表示及初始化

定义1 设可行解 $s' \in S'$ (S' 为解的基因空间) 由 l 个基因位组成,其中的第 i 位表示为 s'_i ,其值可以为集合 $\{a_{i1}, a_{i2}, \dots, a_{in_i}\}$ 中的任意一个或多个。集合 $\{a_{i1}, a_{i2}, \dots, a_{in_i}\}$ 称为 s'_i 的等位基因。如果 s'_i 只能且必须取 $\{a_{i1}, a_{i2}, \dots, a_{in_i}\}$ 中的一个值,则称 s'_i 为单等位基因位。如果

s'_i 只能且必须取 $\{a_{i1}, a_{i2}, \dots, a_{in_i}\}$ 中确定个数的多个值,则称 s'_i 为复等位基因位。

定义2 解空间 S' 的 l 个基因位中每个基因位取与该基因位对应的等位基因值的概率由矢量 $P^i = (P_1, P_2, \dots, P_l)$ 来表示。第 i 基因位取 n_i 个等位基因值的概率用矢量 $P_i = (p_{i1}, p_{i2}, \dots, p_{in_i})$ 描述。 p_{ij} 为第 i 个基因位取等位基因集合 $\{a_{i1}, a_{i2}, \dots, a_{in_i}\}$ 中第 j 个值的概率, $j=1, 2, \dots, n_i$, 且每个基因位取等位基因的概率之和都为1,即概率矢量 $P_i (i=1, 2, \dots, l)$ 都是归一化的:

$$\sum_{j=1}^{n_i} p_{ij} = 1, \text{ 其中 } i=1, 2, \dots, l$$

等位基因概率的初始化一般是采用等位基因等概率的初始化方法。对于有 m 个等位基因的基因位,每个等位基因的初始概率为 $p=1/m$ 。

1.2 基因重组及每代最优个体的获得方法

定义3 在基因空间 S' 中,通过概率 P^i 采样得到的可行解 $s' \in S'$ 称为个体或人口,该概率采样的过程称为基因重组。

基因重组算法因问题的基因结构不同而有所不同。对于最简单的无约束的单等位基因问题,其基因重组算法可以用简单的赌盘选择算法:即把第 i 个基因位的 n_i 个等位基因的概率看成是赌盘上的 n_i 个不同区域,区域的大小为其概率值。这 n_i 个区域恰好组成一个完整的赌盘(n_i 个概率是归一化的)。每次采样时用0-1均匀分布随机数来选择每个基因位的等位基因。

在基因学习算法中,每一代最优个体的获得可以使用三种方法:群体选择法、局部搜索法及这两种方法的混合算法。群体选择法是通过基因重组获得一个群体,计算并比较群体的每个个体的目标函数值,从中选

取最优者。局部搜索法是通过基因重组获得一个单独的个体,然后通过多次局部搜索操作来获得更优的个体。混合法是通过基因重组产生 n 个个体,根据其目标函数值选择最优的个体。然后对该最优个体再进行 m 次局部搜索操作;或通过基因重组产生 n 个个体,对每个个体分别进行 m 次局部搜索操作,然后再根据其目标函数值选择最优的个体。混合法是两者方法的结合从而兼得两者的优点,可以使计算效果更好。在求解如 TSP、调度问题等较为复杂的组合优化问题中,效果尤为明显。

1.3 等位基因概率的修正方法

等位基因概率修正是基因学习算法中对基因学习的累积操作,其方法可以是多种多样的。最为方便的是直接将最优个体中出现的等位基因所对应的概率加上一个修正因子,然后进行等位基因的概率归一化处理(如等比归一化,即等位基因同乘以一个归一化系数)。概率修正因子取值的范围一般在 $10^{-3} \sim 10^{-1}$ 量级。对于规模较小的问题,在计算时间允许的情况下概率修正因子相对来说可以取较小的值,这有利于获得较好的计算结果;对于一些规模较大的组合优化问题,则概率修正因子相对来说就要取稍大一点的值,才能保证在适当的时间内收敛并获得较好的计算结果。

1.4 收敛加速方法

在解决许多规模较大的组合优化问题时,收敛速度很慢,在适当的时间内难以获得满意的结果。本文提出一个概率指数采样法来加速收敛。在采样的过程中,不直接使用概率 p_i ,而是使用概率的指数 p_i^α ($\alpha > 1$)。由于 $\alpha > 1$,因此 p_i^α 可以使赌盘中各等位基因所对应的区域的大小对比加大,从而加速了概率收敛的速度。由于指数的计算量非常大,在实际应用中可以采用部分基因位的概率幂乘(平方、立方)来控制收敛速度,这里使用概率幂乘的基因位在每次采样中都是随机选取的,并且可以用一个称为加速率的参数来控制;加速率为 0 时,完全不加速;加速率为 1 时,全部基因位都采用概率幂乘加速法。

2. 集合覆盖问题的描述

集合覆盖问题描述如下^[4]:给定集合 $V = \{v_1, v_2, \dots, v_n\}$ 及 V 中每个元素对应的价值为 $C = \{c_1, c_2, \dots, c_n\}$ 。 V 中每个元素本身是一个集合,记 $I = \bigcup \{v_j: 1 \leq j \leq n\}$, 并设 I 是由 m 个元素组成,记为 $I = \{1, 2, \dots, m\}$ 。现要从 V 中找一子集 $\{v_j: j \in J^*\}$, 其中 J^* 是集合 $J = \{1, 2, \dots, n\}$ 的子集,使该子集既能覆盖 I , 即 $I = \bigcup \{v_j: j \in J^*\}$, 同时又使该子集的代价最小: $\min \sum_{j \in J^*} (c_j)$ 。

集合覆盖问题是许多实际应用问题的一个抽象,

同时也是一个著名难题。事实上该问题已被证明是一个 NP 完全问题^[5]。对于 NP 完全问题,尤其是规模较大时,一般很难获到最优解,而只能求得次最优解或较优解。求解集合覆盖问题的方法可采用线性规划法^[6]、启发式搜索算法^[7,8]等。

由于集合覆盖问题是一个有约束的优化问题。在用基因学习算法求解该问题时,采用如下的罚函数方法来处理该约束条件:

$$\min \left(\sum_{j \in J^*} c_j - \lambda(m - |J^*|) \right)$$

这里 $|J^*|$ 为可行解集合 J^* 中覆盖的元素个数, λ 为一取值适当的常数。

3. 计算实例

实例 1 这是一个 $n=9, m=7$ 的小规模计算例子^[9]。其集合与价值数据如下表 1。表中子集的信息用 0、1 表示,其含义为 1 表示该子集(列)含有该元素(行), 0 则表示不含。如子集 4 = {4, 7}, 子集 8 = {1, 2, 3, 7}。

表 1 集合覆盖问题(一)原始数据

子集编号	1	2	3	4	5	6	7	8	9
子集价值	10	5	8	6	9	13	11	5	6
子集信息	1	1	1	0	1	0	1	1	0
	0	1	1	0	0	0	0	1	0
	0	1	0	0	1	1	0	1	1
	0	0	0	1	0	0	0	0	0
	1	0	1	0	1	1	0	0	1
	0	1	1	0	0	0	1	0	1
	1	0	0	1	1	0	0	1	1

表 2 集合覆盖问题(二)子集信息表

100000001000000000100100100001
01000000000100100010101000101111
001000000100000000100001010000101
000100000000000000001000100100001
0000100000100001000000110001000010
000001000000000000001000101000101
00000011000100001000000010000110
00000001000000001001001100100001
0000000100001100000000001001110
0000000010000000001000100000011
00000000000001100100001100100101
00000000000000000110010101001010
00000000001000000001000010010101
000000000000000010010001000000010
00000000100000000100000101010011

用基因学习算法计算 50 次,每次解得的最优解都为 {子集 4, 子集 8, 子集 9}, 其相应的最小代价为 16, 与

(下转封四)

(上接第66页)

文献值相同。计算时,每代样本数为50,概率修正因子为0.01,罚函数因子 $\lambda=20$ 。

实例2 这是一个中等规模的集合覆盖问题^[6]。 $n=32, m=15$,这32个子集的价值分别为:1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,6,6,6,7,8,9;其各子集的组成集合信息如表2(含义同表1)。用基因学习算法计算50次,得到一个最优解{4,6,8,10,11,18,19}和三个次最优解{23,25,26}、{5,6,8,18,19,25}、{4,5,6,8,10,11,18,19},其相应的最小代价分别为13,14,14,14,50次平均最小代价为13.2,比文献值的最优解14(上面的第3个次最优解)结果要好。计算时,每代样本数为50,概率修正因子为0.01,罚函数因子 $\lambda=5$ 。

在如上述的中小规模的集合覆盖问题的求解中,不采用概率加速算法,计算速度也非常快。另外,局部搜索法对问题的影响也较小。但在对更大规模的集合覆盖问题的计算结果显示,用基因学习算法可以发挥加速收敛及局部搜索的特点,获得更为满意的结果。

参考文献

- 1 Baluja S. Genetic Algorithms and Explicit Search Statistics. In: Advances in Neural Information Processing System. MIT Press, 1996. 319~325
- 2 Corno F, et al. The Selfish Gene Algorithm, A New Evolutionary Optimization Strategy. In: Proc. (CD-ROM) of the Intl. Conf. of Evolutionary Computation (ICEC'98). 1998. 575~580
- 3 Garfinkel R S, Nemhauser G L. Integer Programming. New York, John Wiley & Sons. 1972
- 4 Aho A V, et al. The Design and Analysis of Computer Algorithms. Mass: Addison-Wesley, 1974
- 5 Lemke C E, et al. Set Covering by Single-Branch Enumeration with Linear-Programming Subproblems. Operations Research, 1971, 19(4): 998~1022
- 6 Chvatal V. A Greedy Heuristic For The Set Covering Problem. Mathematics of Operations Research, 1979, 4(3): 233~235
- 7 集合覆盖问题的启发式函数算法. 软件学报, 1998, 9(2): 156~159
- 8 Bellmore M, Ratliff H D. Set Covering And Involuntary Bases. Man. Sci., 1971, 18(1): 194~206

国际会议消息

第35届面向对象语言与系统技术国际会议(TOOLS Asia'2000)暨第4届中国面向对象技术及应用学术会议(OOT China'2000)将于2000年8月28日~31日在西安召开。有关会议征文范围、征文要求、重要日期请见《计算机科学》1999年第12期(封四)或主页: <http://toolsa2k.xidian.edu.cn>

第三届亚洲太平洋互联网国际会议(AP Web'2000)将于2000年8月31日~9月2日在西安召开。有关会议征文范围、征文要求、重要日期请见《计算机科学》1999年第12期(封四)或主页: <http://apweb2k.xidian.edu.cn>

上列会议欢迎投稿,欢迎莅临。联系方式是:

710071 西安市太白路2号西安电子科技大学168信箱

TOOLS Asia'2000 程序委员会/AP Web'2000 程序委员会 陈平教授

联系电话:029 8202457 传真:029 8202458 E-Mail: apweb2k@mail.xidian.edu.cn

计算机科学

(1974年1月创刊)

第27卷第1期(月刊)

2000年1月25日出版

中国标准刊号: ISSN 1002-137X
CN50-1075/TP

定价: 9.00元 国外定价: 5美元

邮发代号: 78-68

发行范围: 国内外公开

主管单位: 国家科学技术部

主办单位: 国家科技部西南信息中心

编辑出版: 《计算机科学》杂志社

重庆市渝中区胜利路132号 邮政编码: 400013

电话: (023) 63500828 传真: (023) 63502473

主 编: 朱宗元

印刷者: 国家科技部西南信息中心印刷厂

总发行处: 重 庆 市 邮 政 局

订购处: 全 国 各 地 邮 政 局

国外总发行: 中国国际图书贸易总公司(北京399信箱)

国外代号: 6210M