文章编号：1000−8152(2009)08−0927−04

拉丁超立方体抽样遗传算法求解图的二划分问题

陈明华[1], 任 哲[2], 周本达[3]

(1. 皖西学院 计算机科学与技术系, 安徽 六安 237012;

2. 合肥学院 数理系, 安徽 合肥 230022; 3. 皖西学院 数理系, 安徽 六安 237012)

摘要: 图的二划分问题是一个典型的NP-hard组合优化问题, 在许多领域都有重要应用. 近年来, 传统遗传算法等各种智能优化方法被引入到该问题的求解中来, 但效果不理想. 基于理想浓度模型的机理分析, 利用拉丁超立方体抽样的理论和方法, 对遗传算法中的交叉操作进行了重新设计, 并在分析图二划分问题特点的基础上, 结合局部搜索策略, 给出了一个解决图二划分问题的新的遗传算法, 称之为拉丁超立方体抽样遗传算法. 通过将该算法与简单遗传算法和佳点集遗传算法进行求解图二划分问题的仿真模拟比较, 可以看出新的算法提高了求解的质量、速度和精度.

关键词: 图的二划分; 遗传算法; 拉丁超立方体抽样; 拉丁超立方体抽样遗传算法

中图分类号: TP301 **文献标识码**: A

Solving 2-way graph partitioning problem using genetic algorithm based on Latin hypercube sampling

CHEN Ming-hua[1], REN Zhe[2], ZHOU Ben-da[3]

(1. Department of Computer Science and Technology, West Anhui University, Liu'an Anhui 237012, China;

2. Department of Mathematics and Physics, Hefei University, Hefei Anhui 230022, China;

3. Department of Mathematics and Physics, West Anhui University, Liu'an Anhui 237012, China)

Abstract: The 2-way graph partitioning problem is a typical NP-hard combination optimization, and is significantly applied to many fields of science and engineering. Recently, many intelligent optimization methods including the traditional genetic algorithm(GA) are employed to solve this problem, but the result is not effective as we desired. Based on the ideal density model, we redesign the crossover operation in GA by using the Latin hypercube sampling, and combined the result with the local search strategy of the 2-way graph partitioning problem; thus, presenting a new genetic algorithm based on Latin hypercube sampling for solving the 2-way graph partitioning problem. Comparison of simulation results in solving the 2-way graph partitioning problem with this new GA, the simple GA and the good point GA shows that this new method has superiority in speed, accuracy and precision.

Key words: 2-way graph partitioning; genetic algorithm(GA); Latin hypercube sampling(LHS); genetic algorithm based on Latin hypercube sampling(LGA)

1 引言(Introduction)

图的二划分问题可描述为: 对于一个无向图G, 设其顶点的集合为V(顶点数为偶数). 将顶点集合划分为两个子集V_1和V_2, 且V_1和V_2中元素个数基本相同, 使得V_1和V_2两个顶点子集之间连接最少. 图二划分在数字电路系统的芯片划分, 网络管理, 插件划分, 分布式系统的数据流划分, 并行系统的信息流划分等许多领域有着重要应用. 图二划分问题是一个NP-hard组合优化问题, 目前已提出了许多启发式的搜索算法, 其基本思想是寻找当前二划分子集中与其它各节点连接最少的节点, 将其移动到另外的子集中, 如此反复移动, 直到可行集不能改进为止. 这种方法求解质量相对较高, 但对初始划分的依赖性很强, 陷入局部最优解的可能性较大. 近年来, 各种智能优化方法[1,2]被引入到该问题的求解中来.

遗传算法(GA)[1]作为一种全局优化搜索算法, 由于其本身所具有的全局收敛性和隐并行性, 加之其简单易用、鲁棒性强, 被广泛用于求解全局优化

收稿日期: 2008−05−17; 收修改稿日期: 2008−09−11.

基金项目: 安徽省高校省级自然科学研究项目(KJ2007B152); 安徽省教育厅自然科学研究项目(2005KJ222, 2006KJ046B); 安徽省高校青年教师资助计划项目(2007jql180).

问题. 遗传算法中的“选择、变异、交叉”等操作直接关系遗传算法的效率, 一直是遗传算法的研究热点[3,4]. 遗传操作中的交叉操作主要是想实现在高适应度模式中搜索更高适应度模式的作用, 但通常GA中的交叉操作, 是按赌轮法随机取两个染色体进行单点交叉操作(或多点交叉), 即在以高适应度模式为祖先的“家族”中随机取一点作为交叉后的后代. 这种取法当然有其片面性, 不能保证取到的后代的适应度比母体的高. 所以应用经典遗传算法求解图的二划分问题求解时间长、成功率低. 由数论方法知: 在高适应度模式空间中产生的均匀散布点集, 能很好的代表高适应度模式空间的其他点. 故吴[5]提出在以高适应度模式为祖先的“家族”中利用正交试验的方法求出几点来作为交叉后的后代, 如用$L_{16}(2^{15})$, $L_{32}(2^{31})$, $L_{36}(3^{13})$, $L_{32}(4^9)$等等. 这是一个好主意, 不过当因素的个数和等级增多时, 不但试验的规模迅速增加, 而且对应的试验正交表也很难得到. 张[6]提出了佳点集遗传算法, 即利用数论中的佳点集方法[7]求出几点来作为交叉后的后代, 这可使其精度与维数无关, 且这些点有很好的散布性, 克服了用正交设计法的不足. 但佳点集的选取在取点个数n取定后是确定的, 且佳点集布点有方向性, 不带随机性, 更不是统计意义下的抽样, 所以很多格子是取不到的, 其上的点也就不能作为交叉后的后代了, 这将影响整体搜索效果. 为了克服上述不足, 本文提出了基于拉丁超立方体抽样(LHS)的交叉操作, 在某种意义下, LHS是一种分层抽样, 就某种均匀性来说要优于MonteCarlo抽样. 而且文献[8, 9]都说明LHS具有良好的散布均匀性和代表性, 加之它是随机的, 可以取到所有的格子点, 所以搜索能力非常强, 故能比佳点集有更好的表现. 在此基础上, 再结合图二划分问题的局部优化技术[10], 提出了求解图二划分问题的拉丁超立方体抽样遗传算法(LGA).

2 局部搜索技术和LGA(Local search & LGA)

2.1 染色体编码(Chromosome code)

基于图的二划分问题, 这里采用二进制编码. 例如$x = [0\ 1\ 0\ 1\ 1\ 1\ 0\ 1]$表示将顶点集合$V = \{1,2,3,4,5,6,7,8\}$划分为$V_1 = \{1,3,7\}$和$V_2 = \{2,4,5,6,8\}$两个子集.

2.2 适应度函数(Fitness function)

根据图的二划分定义及划分原则, 因为图中所有边的权值和是一个常量, 求属于不同分块的顶点之间的边的权值之和的最小值问题, 实际上也就是求同一分块内各顶点之间的边的权值之和的最大值问题, 据此定义适应度函数如下:

$$f(x) = g(x) - r \cdot u \cdot g(x).$$

其中

$$g(x) = \sum_{1 \leqslant i \leqslant j \leqslant n} w(v_i, v_j).$$

v_i与v_j同属同个分块; $w(v_i, v_j)$表示v_i与v_j间的权值, 在简单无向图中定义为

$$w(v_i, v_j) = \begin{cases} 1, \text{如果 } v_i \text{ adj } v_j, \\ 0, \text{其他}. \end{cases}$$

式中第2项为惩罚函数, $0 < r < 1$为惩罚系数, 它根据个体违反约束条件的程度而定, r越大, 约束条件要求越严格, 否则约束条件比较宽松; u为解是否为合法解的判定系数, 可定义为:

$$u = \begin{cases} 0, \text{如果} x \text{是合法的}, \\ 1, \text{其他}. \end{cases}$$

2.3 LHS杂交算子(LHS crosser)

设在传统的GA基础上, 在进行过复制后, 对池中的染色体随机选择两个A_1和A_2, 进行LHS交叉操作. 一般情况下是令$A_1 = (a_1^1, a_2^1, \cdots, a_l^1)$, $A_2 = (a_1^2, a_2^2, \cdots, a_l^2)$, $J = \{i | a_i^1 \neq a_i^2\}$, 不妨设$A_1$和$A_2$的前$t$个分量不同, 后$l-t$个分量相同, 令模式

$$H = \{(x_1, x_2, \cdots, x_l) | i \in J, x_i = *, i \notin J, x_i = a_i^1\},$$

由A_1, A_2进行交叉(不管是单点交叉或是多点交叉)其子孙必属于H, LHS交叉操作就是要在H上利用LHS方法找出好样本来. 对图的二划分问题, 由于码串0101与1010表示的划分相同, 在这里对于两个染色体A_1, A_2首先直接比较, 记录下不同值的位置存于J_1; 然后将A_2各位取反, 再同A_1进行比较, 记录下不同值的位置存于J_2, 取J_1, J_2中长度小的为J, 不妨令其对应的模式仍为H. 不同值的位置构成一个t维立方体, 记为H', 然后在H'上进行LHS, 即要在t维单位立方体$C^t = [0,1]^t$中进行选n个点的LHS交叉操作, 具体如下:

先将每一维坐标区间$[0,1]$都分成n等分, 并用标号i记小区间$(\frac{i-1}{n}, \frac{i}{n}]$, 用$(\pi_{1j}, \pi_{2j}, \cdots, \pi_{nj})'$, 记第$j$维坐标的$n$个标号$(1, 2, \cdots, n)$的一个随机排列. 又设这$t$个随机排列相互独立, 则得到一个$n \times t$阶的随机矩阵$\pi = (\pi_{ij})_{n \times t}$, 然后令$C = (c_{ij})_{n \times t}$. 其中:

$$c_{ij} = \frac{\pi_{ij} - 0.5 + u_{ij}}{n}, i = 1, \cdots, n, j = 1, \cdots, t.$$

u_{ij}是与π独立的$[-0.5, 0.5]$上均匀分布的i.i.d.样本, 则称这样选取的n个点$c_i = (c_{i1}, c_{i2}, \cdots, c_{it})$, $i = 1, 2, \cdots, n$为一个LHS样本. 称这样的抽样方法

为在t维单位立方体$C^t = [0,1]^t$中选n个点的LHS方法.

令交叉后产生的n个后代中第k个染色体为$b^{(k)} = (b_1^k, b_2^k, \cdots, b_l^k)$, 其中:

$$b_{\mathrm{m}}^k = \begin{cases} < c_{kj} >, & m = t_j \in J\ , 1 \leqslant j \leqslant t, \\ a_{\mathrm{m}}^1, & m \notin J, \end{cases}$$

$1 \leqslant k \leqslant n, 1 \leqslant m \leqslant l$,$< a >$表示: 若$a$的小数部分小于0.5, 则$< a >= 0$; 否则$< a >= 1$.

这样在其“家族”中, 产生了n个后代, 取其中适应值最大者, 作为交叉后的后代. 上述交叉操作, 称为LHS交叉操作.

2.4 局部搜索技术(Local search)

定义顶点的适应值为: 一个顶点j的适应值$\lambda_j = \dfrac{g_j}{g_j + b_j}$, 其中$g_j$为与顶点$j$同一个子集中的顶点与$j$连接的边的权值和数; b_j为与j不同的另一个子集中的顶点与j连接的边的权值和数. 如果$g_j + b_j = 0$那么设定$\lambda_j = 1$. 则求解图的二划分问题的局部搜索算法如下:

1) 对于一个划分$< V_1, V_2 >$, 其中$V_1 \cup V_2 = V$, $V_1 \cap V_2 = \emptyset$. 令BEST $=< V_1, V_2 >$.

2) 计算V中各个顶点的适应值.

3) 选择V中适应值最小的顶点设为m, 再在不含m的子集中随机选取另一个顶点n. 交换m和n, 改写V_1和V_2, 若新得的划分好于BEST, 则改写BEST.

4) 若没达到设定的交换次数, 则转2), 否则算法终止并返回BEST.

2.5 拉丁超立方体抽样遗传算法(A genetic algorithm based on Latin hypercube sampling)

给定交叉概率p_{c}和突变概率p_{m}, LGA如下:

1) 每次进行选择遗传操作, 以概率$\dfrac{f_i}{\sum f_i}$复制A_i, 其中f_i是A_i的适应度值.

2) 以概率p_{c}对其进行LHS交叉操作.

3) 以概率p_{m}进行变异遗传操作.

4) 对新产生的染色体进行局部搜索操作.

5) 进行上述的遗传算法至第T代(T是预先给定的常数), 在算法执行过程中记录适应度最大的染色体, 即为所求的染色体, 再进行解码得到最优解.

3 图二划分问题的LGA求解(Solving 2-way graph partitioning problem using LGA)

为了说明算法的有效性, 分别用本文中提到的3种遗传算法(GA; 佳点集遗传算法(GGA); LGA)在同样条件下, 在P4 3.0G PC机器上, 采用MATLAB 7.0计算平台对国际标准数据库中[11]的算例进行仿真, 结果如表1,2所示.

表 1　仿真算例

Table 1　Example data

问题	顶点数	边数目	顶点平均度	顶点最大(小)度
Queen5_5	25	160	6.4	16(12)
David	87	406	4.67	82(1)
Miles250	128	387	3.02	16(0)
Queen10_10	100	1470	14.7	35(27)
Queen13_13	169	3328	18.03	48(36)
Queen15_15	225	5180	23.02	56(42)

表 2　3种遗传算法的仿真结果

Table 2　Experimental result of GA, GGA and LGA

问题	算法	Best	Ave	σ	Ave Gen
Queen5_5	GA	60	63.02	2.91	45.9
	GGA	60	60.12	0.84	12.52
	LGA	60	60	0	0.12
David	GA	84	92.47	4.89	146.49
	GGA	83	90.83	4.63	56.81
	LGA	82	87.03	3.02	23.94
Miles250	GA	5	33.44	8.91	135.02
	GGA	8	46.26	13.18	66.01
	LGA	4	8.71	2.95	14.84
Queen10_10	GA	495	515.93	22.87	150.71
	GGA	515	515.70	2.25	52.02
	LGA	495	495	0	13.16
Queen13_13	GA	1170	1277.3	43.25	174.17
	GGA	1266	1367.4	37.1	131.12
	LGA	1092	1092.3	1.71	8.63
Queen15_15	GA	1950	2066.9	52.67	174.9
	GGA	2166	2265.2	29.51	123.75
	LGA	1680	1686.9	13.82	15.17

其中算法参数为: 群体规模: max_pop=100; 交叉和变异概率: $p_{\mathrm{c}} = 0.9, p_{\mathrm{m}} = 0.05$; 最大迭代代数: max_gen=200. 算法中惩罚系数$r = 1 - (\dfrac{1}{2})^{|\max(|V_1|,|V_2|) - n/2|}$, $n = 10$, 当$(\max(|V_1|, |V_2|) - n/2) \leqslant 5$时$u = 0$, 否则$u = 1$, 为避免遗传算法的随机性, 对每个标准算例在同一台机器上进行连续100次计算, 记录如下结果: 100次运行中求得的最好解(记为Best); 100次运行所得解的平均值(记为Ave); 100次运行求得解的标准差(记为σ); 100次运行中每次收敛时代数的平均值(记为Ave Gen).

由上表可以得出: LGA在搜索能力、收敛速度以

及避免早熟等各项指标上均好于GA和佳点集遗传算法.

4 结论(Conclusions)

众所周知, 任何一种遗传算法的交叉操作都无非是要在以其父代为“祖先”的“家族”中寻找一个更高适应度后代. 现有的交叉操作: 如单点交叉、多点交叉等, 都只能保证求到的后代落在上述的家族中, 搜索高适应度后代的能力不强; 佳点法利用求得子集的均匀散布性, 使它们能较好地代表其“家族”性能的普遍性, 所以佳点集遗传算法构造的交叉操作提高了搜索高适应度后代的能力, 但佳点集布点带有方向性, 并且当佳点个数n取定后, 佳点集中元素是确定的, 不带随机性, 这导致了其全局搜索能力仍不够强. 而LHS方法就克服了此不足, LHS方法所得的样本具有很好的随机均匀散布性, 它是抽样, 所以可以取到所有的格子. 这样基于LHS方法的LGA具有更强的整体搜索能力, 从以上仿真结果可看出在求解的质量、速度和精度上LGA不仅优于GA, 也优于GGA.

参考文献(References):

[1] KANG S, MOON B R. A hybrid genetic algorithm for multi-way graph partitioning[C] //*Proceedings Genetic & Evolutionary Computer Conference (GECCO-2000)*. San Francisco: Morgan Kaufmann, 2000: 159 – 166.

[2] HENDRICKSON B, KOLDA T G. Graph partitioning models for parallel computing[J]. *Parallel Compute*, 2000, 26(12): 1519 – 1534.

[3] 戴朝华, 朱云芳, 陈维荣. 云自适应遗传算法[J]. 控制理论与应用, 2007, 24(4): 646 – 650.
(DAI Chaohua, ZHU Yunfang, CHEN Weirong. Adaptive genetic algorithm based on cloud theory[J]. *Control Theory & Applications*, 2007, 24(4): 646 – 650.)

[4] 李宏, 焦永昌, 张莉, 等. 一种求解全局优化问题的新混合遗传算法[J]. 控制理论与应用, 2007, 24(3): 343 – 348.
(LI Hong, JIAO Yongchang, ZHANG Li, et al. Novel hybrid genetic algorithm for global optimization problems[J]. *Control Theory & Applications*, 2007, 24(3): 343 – 348.)

[5] 吴少岩, 张青富, 陈火旺. 基于家族优生学的进化算法[J]. 软件学报, 1997, 8(2): 137 – 144.
(WU Shaoyan, ZHANG Qingfu, CHEN Huowang. A new evolutionary based on family eugenics[J]. *Journal of Software*, 1997, 8(2): 137 – 144.)

[6] 张铃, 张钹. 佳点集遗传算法[J]. 计算机学报, 2001, 24(9): 917 – 922.
(ZHANG Ling, ZHANG Bo. Good point set based genetic algorithm[J]. *Chinese Journal of Computers*, 2001, 24(9): 917 – 922.)

[7] 华罗庚, 王元. 数论在近似分析中的应用[M]. 北京: 科学出版社, 1978.
(HUA Luogeng, WANG Yuan. *Applications of Number-Theoretic Methods in Approximate Analysis*[M]. Beijing: Science Press, 1978.)

[8] STEIN M. Large sample properties of simulations using Latin hypercube sampling[J]. *Technometrics*, 1987, 29(2): 143 – 151.

[9] OWEN A B. A central limit theorem for Latin hypercube sampling[J]. *Journal of the Royal Statistical Society*, 1992, 54(B): 541 – 551.

[10] SOPER A J, WALSHAW C, CROSS M. A combined evolutionary search and multilevel optimisation approach to graph partitioning[J]. *Journal of Global Optimization*, 2004, 29(2): 225 – 241.

[11] http://www.gre.ac.uk/ c.walshaw/partition.

作者简介:

陈明华 (1954—), 男, 教授, 目前研究方向为概率统计的大样本理论、人工智能和遗传算法, E-mail: mhchen@wxc.edu.cn;

任 哲 (1957—), 女, 教授, 目前研究方向为优化算法、非参数统计和大样本理论, E-mail: renzhe@hfuu.edu.cn;

周本达 (1974—), 男, 副教授, 目前研究方向为计算智能、多Agent系统建模, E-mail: bendazhou@163.com.