混合股遗传算法及其在智能组卷中的应用①

A Hybrid Cohort Genetic Algorithm and Its Application in Test Paper Generation

邓新秀　张　敏　葛　斌　（大连大学 信息工程学院 辽宁 大连 116622）

摘　要：为了在尽可能短的时间内找到问题的最优解，本文在采用股遗传算法高适应度个体遗传速度快、抑制劣质基因的遗传漂移、能够保持群体的多样度、减少早熟收敛现象的发生等优点的基础上，引入了变异过程中的预选择机制来保持优良个体，避免优良基因的丢失，提高种群的整体水平，从而提高算法的性能。将改进的股遗传算法应用于智能组卷，并引入了分段的整数编码和选题过程中的变异操作来提高组卷速度和避免重题的出现。实验结果表明：改进的股遗传算法与标准遗传算法、股遗传算法相比，该算法能大幅度的提高组卷质量。

关键词：变异 预选择 股遗传算法 智能组卷 适应度

计算机辅助组卷系统是一种设计型专家系统，主要任务是对影响试卷的质量指标如：难度、内容、时间、分数及教学要求等进行控制，构成一些试卷的各种“专家曲线”（也称分布列），全部专家曲线又组成试卷模式产生合乎要求的试卷，它依赖于专家经验（所谓专家曲线）[1]。组卷系统的性能主要取决于组卷算法，如何保证生成的试卷能最大程度的满足用户的不同需要，并具有随机性，科学性，合理性是实现中的一个难点。因此，选择一个高效，科学，强壮的算法是智能组卷的关键。

组卷的目标是从题库中选取试题，组成一套满足用户要求即满足约束条件的试卷。组卷问题的求解过程就是从影响试卷质量的指标约束中寻找满足条件的最优组合，使生成的试卷满足考试要求。显然，该问题为一个目标函数与一组约束条件的组合，因此，智能组卷是一个约束满足问题的求解过程。

遗传算法是一种基于自然选择和基因遗传学原理的随机并行搜索算法 ，是一种寻求全局最优解而不需要任何初始化信息的高效优化方法。由于遗传算法能以较大概率在有限时间内求得整体最优解，同时对目标函数和约束条件不作更多要求，已经成为求解一般约束优化问题的强有力工具，但是遗传算法具有收敛速度慢、早熟、局部寻优能力差等缺点。股遗传算法[2] CGA（cohort GA）是 Holland 于 2000 年提出的一种新型遗传算法，它强调了重组算子的优点。文献[3]中证明了 CGA 在一定条件下能够减少早熟收敛现象的发生，并指出对 CGA 进行更深入的研究具有重大意义。目前国内相关研究较少。本文根据组卷问题的特点设计了一个改进的股遗传算法在适当降低股遗传算法效率的条件下，大幅度提高了算法的质量。实验表明，该方法对于解决智能组卷中的约束优化问题具有很好的性能。

1　混合股遗传算法的基本思想

Holland 提出的股遗传算法维持一组有序编号的子种群，按照编号顺序使各子种群依次进化。交叉和变异操作在各子种群内部进行，每一个父代个体均产生相同数量的新个体。新个体根据其适应度值决定要进入的子种群编号，编号的原则是使高适应度值的新个体在相继的子种群中再次获得繁殖机会，这样，高适应度的个体繁殖速度快，就会有更多的后裔。CGA 的目的是为了克服劣等基因的遗传漂移现象，使遗传算法能更有效地利用优良积木块，提高遗传算法的性能。但是在 CGA 中，采用传统的变异算子，产生的新个体

① 基金项目：辽宁省教育厅高等学校研究项目（2023901017）

完全替代父代个体,造成一些优良个体的丢失,为了避免这种情况,改进算法在变异算子中加入一种排挤技术的替换策略—预选择机制[4],即只有当新个体适应度值超过父代个体时,新个体才替换父代个体进入下一代种群。Cavicchio[5]声称在种群规模相对较小的情况下,预选择可维持较高的种群多样度。新算法还采用了保优机制,最优个体参与交叉操作,使算法能收敛到全局最优解。

改进的股遗传算法整个过程可描述如下:

Step 1 随机初始化种群,并划分为若干个子种群,从小到大依次编号(如 0,1,…)。

Step 2 计算种群中新个体的适应度。

Step 3 保留最优个体,如果满足停止准则,算法输出结果并停止运行;否则,继续。

Step 4 对当前子种群执行如下操作,操作完成后转到 Step 3。(从子种群编号为 0 的开始每个子种群依次进行,执行完最后一个子种群再从 0 开始执行。)

Step 4.1 最优个体及当前子种群中每个个体按交叉概率进行交叉操作,计算新个体的适应度值,并保留最优个体,产生的新个体完全替代父代个体。

Step 4.2 对交叉产生的新个体进行按变异概率进行变异操作,计算新个体的适应度值与父代个体进行比较,保留适应度高的个体。

Step 4.3 每个子种群生成的个体根据新个体放置策略被放入到后面子种群中。放入哪一个子种群由该个体的适应度函数和它所在的子种群决定。如:个体适应度,子种群个数为,所在的子种群号为,则该个体将被放入子种群中,由下面公式决定:

$$d = \mathrm{mod}(t + dout, nocoh)$$

$$dout = (nocoh - 1) + \lceil (u - umin) \times (2 - nocoh + 1)/(umax - umin) \rceil$$

其中,与分别为至今为止最小适应度值和最大适应度值,为该个体的适应度值。

2 组卷算法中的约束

组卷的数学模型参见文献[6],组卷中决定一道试题,取决于 n 项指标,本文考虑六项指标(题型,题分,难度,教学要求层次,章节,时间)来生成一份试卷,由于采用分段的整数编码,将题型标识加入到编码中,因此题型这一约束自动满足,不用再作为约束条件。所以生成的试卷只需满足下列约束条件:

(1) 题型分数:该类题型的分数。

(2) 难度:试题的难易程度,它主要反映试题是否符合被测试者的实际水平。

(3) 教学要求层次,反映该题要求掌握的程度(识记,理解,应用,综合)。

(4) 章节:反映知识点的覆盖程度。

(5) 估计答题时间:预计完成该题所需要的时间。

以上五个约束即用户对试卷的基本要求,为了满足特殊的要求,还可以定义其它的约束如区分度等,但是,组卷实验的结果表明,指标过多会增加组卷的难度从而降低效率。

3 混合股遗传算法在智能组卷中的应用

编码方法:文献[7]将二进制编码的遗传算法应用于智能组卷中,染色体的长度等于题库中试题个数,每个基因位以二进制表示相应的试题是否被选中。这种编码方式的优点是易于操作,简单明了,但是染色体的长度由题库中试题数决定,而题库中试题数经常发生变化,这样降低了组卷效率和灵活性。因此本文针对标准化试卷选择,填空和判断题中各小题分数相等的特点,采用分段的整数编码方式。分段编码就是每一段表示一种题型,每一段的长度由试卷中要求该题型的个数决定,编码总长度为试卷要求的题目个数,编码形式为 a1a2...an,其中 ai 为题库中的题目编号,这样缩短了编码串的长度,减少了二进制编码中编码、解码过程,明显缩短了求解时间,提高了算法的速度,在进化过程中表现出很好的搜索性能。

初始化群体:产生随机数初始化群体。实际组卷中,在题库中每种题型个数范围内产生随机数,染色体的长度为试卷要求的题目个数;为了便于从题库中抽取试题属性约束计算适应度,把随机数加上题型标识转化为题库中题号的形式作为基因值。

适应度函数:在遗传操作中,以适应度大小来区分个体的优劣。生成的试卷要满足用户的总体要求即:试题类型、试题难度、估计时间等约束,在该算法中用来表示这些属性与用户要求之间的误差,由于每个属性对试卷的影响不同,因此整套试卷的误差就是这些约束的加权和,为了各个误差不相互抵消,对每个误差

都取绝对值[8,9]，则用下式表示：

$$f=\sum_{i=1}^{5} f_i w_i \qquad 0\le w_i\le 1,\ 1\le i\le 5$$

其中：表示第个约束与用户要求差的绝对值，表示第个约束的权重，实际组卷中，有些约束很重要（如分数），有些约束可以适当放松（如考试时间），因此，根据这些约束的不同性质和不同要求，可对自行赋值。根据的定义可知，定义适应度函数为：

$$fitness=\begin{cases}1-f, if\ 0<f\\ 0, others\end{cases}\qquad 0\le fitness\le 1$$

值越大表示越满足命题的要求，当时，组卷方案完全满足命题要求。

交叉算子：由于采用分段编码方式，交叉只能在相同的段内进行，这里采用单点交叉策略，最优个体参与交叉运算按交叉概率进行交叉运算，交叉之后的新个体完全替代父代个体。

变异算子：随机产生变异点，产生一个该题型范围内的随机数，并保证与同一段内其它位不重复，将新个体与父代个体进行比较，如果其适应度值超过相应父代个体的适应度值，新个体才替换父代个体进入下一代群体。在题库中提取所选试题及属性约束时若在同一段内有重复试题则采用变异操作来重新产生新的题目替换重复的题目，直到没有重题为止。

保优策略：在算法的运行过程中，由于交叉、变异操作可能会破坏当前群体中适应度最好的个体，降低群体的平均适应度，对运行效率和收敛性产生不利的影响。本文采用保优策略，把至今为止最好的个体保留下来，并且为了充分利用最优个体中的基因，最优个体参与交叉运算，这样保证算法全局收敛的同时还提高了收敛速度。

终止条件：（1）最优个体的适应度函数值为 1 或几乎为 1；（2）达到一定的进化代数。

4　实验结果

实验中的约束条件可以人工输入，其默认设置为：试卷总分 100 分，考试时间 120 分钟，其余约束所占比例设置如表 1 所示：

题库中共有 1000 道题，试卷要求的题数为 47 道。参数设置为：最大迭代次数为 500 代，期望的误差精度为 0.05，种群数为 100，子种群个数为 10。算法用 C#实现，后台数据库采用 SQL Server2000，分别利用本文算法、股遗传算法 CGA 和 SGA 进行性能比较如表 2 所示：

表 1　约束条件表

难度	题型分数	教学层次	章节
很容易 10%	选择 40%	识记 20%	一章 20%
容易 25%	填空 10%	理解 30%	二章 25%
中等 30%	判断 10%	应用 30%	三章 25%
较难 25%	简答 30%	综合 20%	四章 15%
很难 10%	应用 10%	其它 0	五章 15%

表 2　组卷效率比较表

次数	SGA 算法		股遗传算法		改进股遗传算法	
	Num	Max	Num	Max	Num	Max
1	200	0.9768	18	0.9686	73	0.9823
2	232	0.9560	28	0.9670	75	0.9848
3	200	0.9623	29	0.9722	45	0.9897
4	368	0.9602	20	0.9664	63	0.9893
5	206	0.9538	23	0.9706	63	0.9851

表 2 中，Max 表示最优适应度，Num 表示执行代数。从组卷结果可以看出，在题库中试题足够多，分布合理的条件下，标准遗传算法组卷效率比较低，且组卷质量不高；股遗传算法的组卷速度明显提高，且组卷质量也有所提高；改进的股遗传算法，较 SGA 在效率和质量上都有明显的提高；与股遗传算法相比，改进后的算法由于引进了变异预选择机制，保留了最优个体，大幅度的提高了组卷质量，但是组卷效率相对于股遗传算法有所降低。实验表明，改进的股遗传算法在期望的误差精度 0.02 范围内很好的收敛到最优解，组卷成功率达到 100% 。

5　结束语

本文提出一个改进的股遗传算法，并应用到智能组卷中，使算法快速收敛，大幅度提高了组卷质量。实验表明，如果题库量足够大，各项指标分布合理，本文算法在组卷中能取得良好的效果，由于算法的求解精

（下转第 83 页）

(上接第92页)

度和收敛速度是相互矛盾的,本文算法提高组卷质量的同时,较股遗传算法降低了效率,要使组卷速度和精度得到进一步提高,还要进行更深入的研究。

参考文献

1 国家教委考试中心. 题库建设理论与实践. 光明日报出版社,1991:1-252.

2 Holland J H. Building blocks, cohort genetic algorithms, and hyperplane-defined functions. Evolutionary Computation, 2000, 8(4):373-391.

3 Pei H, Goodman E. A comparison of cohort genetic algorithms with canonical serial and island-model distributed gas. In Spector L et al., eds, Proceedings of the Genetic and Evolutionary, Computation Conference, San Francisco, CA, USA. Morgan Kaufmann. 2001.501-510.

4 郭观七,喻寿益. 小生态进化技术综述. 计算机工程与设计,2005,26(4):857-861.

5 Cavicchio D J. Reproductive Adaptive Plans. Proc of the ACM 1970 Annual Conf,1970.1-11.

6 谢志强. 题库系统中试卷生成与分析的研究. 湘潭大学, 2005.

7 全惠云,范国闯,等,基于遗传算法的试题库智能组卷系统研究,武汉大学学报,2000,45(5):758-760.

8 华如海,王俊普. 基于约束满足的智能组卷方法的研究与实现. 计算机应用研究,2000,11:20-22.

9 闭应洲,苏德富,陈宁江. 基于矩阵编码的遗传算法及其在自动组卷中的应用. 计算机工程,2003,29(6):73-76.