

基于改进遗传算法的智能组卷方法

马德良, 陆昌辉, 王小乐

(国防科学技术大学 信息系统与管理学院, 长沙 410073)

(madeliang@gmail.com)

摘 要:组卷问题是一个多约束多目标优化问题。建立了一种新的试卷矩阵数学模型,提出了改进的遗传算法编码方式,并通过改进初始群体的产生方法和遗传算子,有效提高了遗传算法的收敛速度,并较好地避免了局部收敛现象。实验结果表明,在试题库试题数量适中、分布合理的情况下,本算法产生的试卷能够很好满足各项组卷指标。

关键词:遗传算法;智能组卷;数学模型

中图分类号: TP301.6 **文献标志码:** A

Intelligent test paper generation based on improved genetic algorithm

MA De-liang, LU Chang-hui, WANG Xiao-le

(College of Information Systems and Management, National University of Defense Technology, Changsha Hunan 410073, China)

Abstract: The test paper generation is a multi-constraint and multi-objective optimization issue. A new mathematical model of intelligence test paper generation system was set up. To avoid slow-convergence and local convergence of simple genetic algorithm (SGA), a kind of improved genetic algorithm was proposed in this paper. The experimental results show that the new method is more efficient and effective to deal with the problem of intelligent test paper generation.

Key words: Genetic Algorithm (GA); intelligent test paper generation; mathematical model

0 引言

随着各种自动化考试系统的普及和各类题库建设的逐步完善,自动组卷应用日益广泛。一份高质量的试卷要满足多种约束。自动组卷就是按照考试的要求,由计算机自动从试题库中选择试题,组成一份满足多重约束的试卷^[1]。如何保证生成的试卷能最大限度地满足用户的不同需要,并具有随机性、科学性、合理性,是现实中的一个难点。因此,选择一种高效、科学、合理的算法是自动组卷的关键^[2]。当前具有代表性的组卷算法有优先权策略、弱并行策略^[3]、误差补偿策略^[4]、随机抽题法及回溯试探法^[5]等。以上各种算法各有其优点,但是,当试题库规模较大和组卷指标复杂时,这些算法在组卷质量和组卷速度方面,都难以达到令人满意的效果。

遗传算法(Genetic Algorithm, GA)是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法,具有并行性、通用性、全局优化性、稳健性、操作性与简单性等特点。遗传算法的基本特征是通过在代与代之间维持由潜在解组成的种群来实现多样性和全局搜索^[6]。遗传算法的群体搜索策略为多目标优化提供了较好的解决方案。近年来,将遗传算法用于解决智能组卷问题^[7-8],取得了很好的进展。但在处理自动组卷中的知识点重复、后期收敛速度慢、多个约束条件之间的冲突等问题方面仍无法达到理想效果。本文给出了一种基于矩阵编码的组卷算法,并改进了遗传算子,提出了利用遗传算法进行智能组卷的新思路。

1 编码的基本思想和试卷的数学模型

1.1 编码方案

编码方案的选择依赖于问题的性质,并影响到算法内操

作的设计,是影响算法性能的重要因素。常见的编码方案有二进制编码、十进制编码、实数编码等。自动组卷可描述为从一定题量的试题库中抽取满足组卷目标要求的一组试题组合,从而组卷问题转化为一个多重约束目标优化问题。组卷涉及的试题以记录的形式保存在试题库中。对于单个试题,常有的属性有:考查的知识点、试题题型、试题难度、估计用时、试题区分度、试题分值、试题考查的能力层次等。在题库中可以根据需要扩充这些属性。

本算法采用实数矩阵编码方案,可以解决采用二进制编码搜索空间过大和编码长度过长的缺点,同时取消了个体的解码时间,提高了求解速度。具体来说,就是对试题库中每个试题独立进行实数编码,假设试题具有 n 个属性,每道试题就对应一个 $n+1$ 维向量,其中第一维为试题编号,这个编号由知识点决定,第二维为题型,第三维为难度、第四维为估计用时等。每个知识点下有题型、难度、答题时间、区分度等等不同的试题。可以根据题库的实际情况决定每一维的编码位数。假设题库中的知识点有99999个,题型有99种,难度有9类,区分度有9类等。则知识点编号12345下的,题型编号为06,难度为7,区分度为8,(其他属性……)的试题编号为:12345 06 7 8 (其他属性编码……)。

考虑到实际组卷中的应用,对每道试题在成功组卷前并不直接赋予分数值属性。在成功组卷以后,再根据题型、难度、答题时间等属性对试题赋予分数值。

1.2 试卷的数学模型

在本算法中,将一份试卷映射为一个矩阵染色体,组成试卷的每个试题映射为一个基因向量。设每份试卷有 m 道题,每题有 n 个属性,试卷结构可以表示为 $(m+1) \times (n+1)$ 矩阵形式。其中第1列对应于试题在题库中的编号,其余各列代

收稿日期:2009-01-13;修回日期:2009-03-02。

作者简介:马德良(1980-),男,河南项城人,硕士研究生,主要研究方向:信息系统、智能决策; 陆昌辉(1976-),男,湖南武冈人,讲师,博士,主要研究方向:数据仓库、决策支持; 王小乐(1983-),男,陕西西安人,博士研究生,主要研究方向:信息系统、智能决策。

表试题的各个属性值。第1行对应这份试卷的评价值, a_0 等于这份试卷的适应值, a_k 等于第 k 列属性的值。如第2列指题型, a_{12} 指第 i 题的题型代码, a_2 指整份试卷对题型要求的满足情况。第3列指难度, a_3 指整份试卷的难度;第4列指答题时间, a_4 指整份试卷的答题时间等。试卷编码矩阵如图1。

$$\begin{pmatrix} a_0 & a_1 & \cdots & a_n \\ a_{10} & a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m0} & a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{matrix} \text{—— 试卷的评价值} \\ \text{—— 试卷 } a_{10} \\ \vdots \\ \text{—— 试卷 } a_{m0} \end{matrix}$$

图1 试卷编码矩阵

1.3 适应值函数设置

本算法中,适应值的计算采取为每个目标分配权重并将其组合成为一个目标函数。该方法的基本思想由文献[9]首先提出。本算法中预先设定理想结果下组卷各个目标的属性值 e_i 及其允许误差范围 m_i 。同时设定各个属性所占权重, $\sum w_i = 1$ 。各个属性的值为其和理想值差值然后除以理想值,当其小于允许误差时,取实际值;当其大于允许误差时,取1。各属性值的目标函数计算公式为:

$$f_i = \begin{cases} \left| \frac{a_i - e_i}{e_i} \right|, & \left| \frac{a_i - e_i}{e_i} \right| < m_i \\ 1, & \left| \frac{a_i - e_i}{e_i} \right| \geq m_i \end{cases}$$

目标函数设为 $f = \sum_{i=1}^n w_i f_i$ 。而适应度函数是越大越好,所以要将目标函数转换成适应度函数。因指数比例在保持优良个体高复制机会的同时又限制其复制数目,提高了相近个体间的竞争,保持种群多样性,故采用指数比例变换方法将其转换为适应度函数^[10]。目标函数 f 变换为适应度函数 F : $F = \exp(-af)$, 这种变换法的基本思想来源于模拟退火过程(Simulated Annealing, SA), 其中的系数 a (一般可取值 0.05) 决定了复制的强制性, 其值越小, 复制的强度就越趋向于那些具有最大适应度的个体。

2 智能组卷算法的具体实现

2.1 初始种群的产生

本算法中,初始种群的产生不采用完全随机的方法,而是以知识点为基本依据,产生满足知识点、题型分布、答题时间等三项基本要求的试卷初始群体,以加快遗传算法的收敛速度并减少迭代次数。

初始群体中单份试卷产生的具体算法如下:

- 1) 输入组卷要求。
- 2) 初始化算法,置题型 i 下的试题数目 $t_i = 0$, 试卷答题时间 $T = 0$ 。
- 3) 如果某知识点下的试题未被抽取,则在试题库中随机抽取该知识点下题型为 i 的答题时间为 T_s 试题。
- 4) $t_i \leftarrow t_i + 1$; $T \leftarrow T + T_s$ 。
- 5) 判断,如满足基本条件,转6);否则转3)。
- 6) 计算本份试卷的各个属性值和整份试卷的适应值,并记入试卷染色体矩阵的第一行。

2.2 遗传算子的改进

2.2.1 选择算子

本算法采用基于适应值排序的选择方法。本方法在遗传进化过程中具有适应值比例变换的效果,同时避免了比例变换参数的选取^[11]。对父代中的个体按照适应值进行排序,然

后选取前面的 25% 直接复制进入子代。

2.2.2 杂交算子

在遗传算法中,杂交操作是获取新的优良个体的重要手段,只要初始种群中包含足够的模式信息,杂交操作有能力搜索到全局最优解,并且当前种群的多样性越大,交叉操作的搜索能力越强^[12]。本算法采用杂交的方法产生子代 50% 的个体。采用分段单点杂交策略:对随机选择的两个染色体,在同一题型内采用随机产生杂交位置的单点杂交,若杂交后出现知识点重复,则重新选择杂交点。父代中的每个染色体都杂交一次且仅杂交一次。两个父代染色体杂交产生两个新染色体,适应值高的进入子代种群,适应值低的被舍弃。

2.2.3 变异算子

本算法在父代群体中随机选取 25% 的个体进行变异,产生子代 25% 的个体。采用变异产生新的子代的比率比较高,可以较好保持种群的多样性。本算法采用双点变异策略:随机选择一个基因在保持题型不变的情况下进行知识点编码段的变异,由于变异产生了知识点冲突,将有冲突的另一个基因在保持题型不变的情况下相应进行知识点编码的变异。变异的范围必须保持在题库中已有的编码内。

2.3 遗传算法的实现

根据需要组成的试卷份数决定群体规模大小。设定算法的终止条件为:达到进化代数,或者连续 n 代中的最大个体适应值增长率低于设定值。具体算法如下:

- 1) 初始化。 $t = 0$, 产生初始群体 $G(t)$;
- 2) 选择和复制。产生新群体 $G(t+1)$ 的 25% 的个体;
- 3) 杂交。产生新群体 $G(t+1)$ 的 50% 的个体;
- 4) 变异。产生新群体 $G(t+1)$ 的 25% 的个体;
- 5) 对新一代 $G(t+1)$ 中的个体根据适应值进行排序;
- 6) 判断。如果满足条件终止进化;否则 $t = t + 1$, 返回2);
- 7) 按题型和其他预设条件对试卷中的试题排序,给试题赋予分数值,输出。

赋予分数值时,先对固定分数值的题型进行赋值(如选择题、填空题、判断题等客观题),再把剩余的分数根据试题难度和答题时间等因素合理分配给主观题,需要时可以进行人工调整。

3 实验结果

实验平台计算机配置:CPU: Pentium IV 2.4 GHz; 主板芯片组: Intel 865G; 内存: 1 GB; 操作系统: Windows XP sp2。算法实现语言采用 Microsoft Visual Studio 2005; 试题库使用 SQL SEVER 2005 随机生成。生成的试题库中试题包含知识点编码 00~99 共 100 个知识点; 题型编码为 01~05, 分别设定为对应判断、选择、填空、简答、计算共 5 种题型; 难度编码从易到难为 1~5 共 5 种; 区分度编码从差到好为 11~55 共 5 种; 最后曝光时间使用年月日表示, 设定为 1980 年 1 月 1 日到 2008 年 12 月 31 日, 如 2000 年 1 月 1 日编码为 20000101, 选取试题时, 最后曝光时间之后曝光过的试题不得选取。每道题的答题时间根据题型和难度的不同设定为 1~30 个单位时间, 每个单位时间设定为 0.5 分钟。试题库中试题数量为 10 万条。

生成的试卷含 5 种题型, 每种题型含 10 道题, 答题时间设定为 300 个单位时间, 最后曝光时间设定为 2006 年 12 月 31 日, 试卷难度设定为 3, 试卷区分度设定为 3。初始群体规模分别设置为 10、20、50、100、150, 最大进化代数为 1000。

不同种群规模下进化代次与最优试卷适应度的关系如图 2。

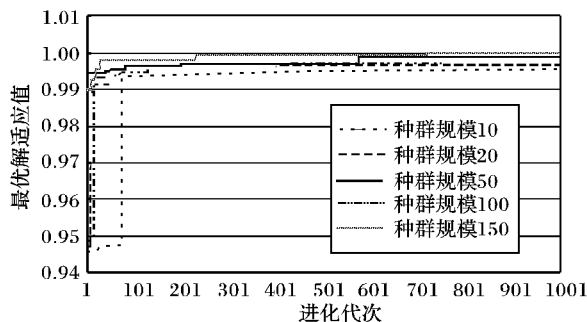


图2 进化代数与最优试卷适应度关系

种群规模为50时,进化代数与最优试卷各项指标误差关系如图3。

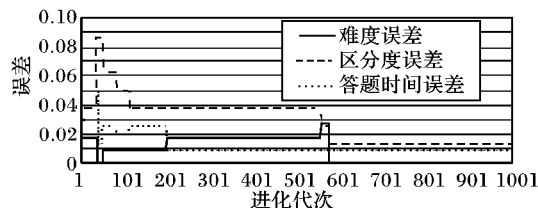


图3 进化代数与最优试卷各项指标误差关系

不同种群规模下的初始化时间、相关代数进化时间如图4。

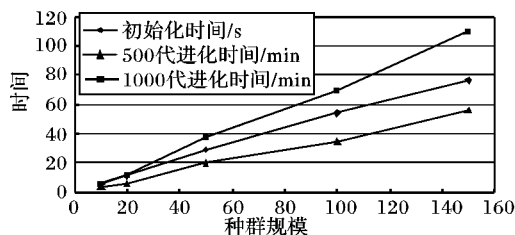


图4 种群规模和初始化时间、进化时间关系

实验结果证明:相对于原始遗传算法存在收敛速度慢,容易陷入局部收敛,很难产生满足多种指标要求的试卷的问题,改进后的遗传算法具有以下优点:1)相同种群规模下,进化代次和时间存在近似线性关系;2)相同进化代次下,进化时间和种群规模存在近似线性关系;3)在进化代次达到500代时,可以组成令人满意的符合多种设定指标的试卷。

4 结语

本文利用一种改进的遗传算法来解决智能组卷问题。给

出了一种基于矩阵编码的组卷模型。改进了编码方式,使每个试卷多维矩阵个体能够记忆其各个属性值和适应值,避免了对个体适应值的重复计算。改进了产生初始群体的方法,通过预有控制地产生较好的初始群体,保证了进化的起点比较高。在进化过程中,利用基于适应值排序的选择方法进行复制,保证父代中的优良个体进入子代。改进了杂交算子,实行分段单点杂交,实现试卷在杂交前后都能够满足题型要求。改进了变异算子,采用双点变异策略,避免了变异产生的试题知识点冲突。同时利用较大的杂交概率和变异概率不断补充新个体,保持群体的多样性,从而避免了搜索空间的迅速缩小,提高了全局寻优性能,加快了整个算法向全局最优值的逼近速度。可以同时生成多份符合需要的试卷,并可以根据需要扩展试卷的指标项。所设计的组卷方法具有收敛速度快、性能好、效率高等特点。

参考文献:

- [1] 吴美娟. 网络考试系统的组卷算法及安全策略研究[D]. 长沙: 中南大学, 2006.
- [2] 魏平, 于海光, 熊伟清. 基于进化稳定策略的单亲遗传算法求解组卷问题[J]. 微电子学与计算机, 2005, 22(1): 105-109.
- [3] 戴亚非, 李晓明, 唐朔飞. 计算机自动组卷演算法分析[J]. 小型微型计算机系统, 1995, 16(9): 51-55.
- [4] 胡维华, 梁荣华, 江虹. 多目标选题策略研究与应用[J]. 杭州电子工业学院学报, 1999, 19(2): 37-41.
- [5] 杨路明, 陈大鑫. 改进遗传算法在试题自动组卷中的应用研究[J]. 计算机与数字工程, 2004, 32(5): 76-79.
- [6] GEN MITSUO, CHENG RUNWEI. 遗传算法与工程优化[M]. 于歆杰, 周根贵, 译. 北京: 清华大学出版社, 2004: 83-86.
- [7] 路景. 基于改进遗传算法的智能组卷研究[D]. 长沙: 中南大学, 2007.
- [8] 朱黎明. 基于单亲遗传算法的试题生成及其应用研究[D]. 长沙: 湖南大学, 2005.
- [9] ZADEH L. Optimality and non-scalar-valued performance criteria[J]. IEEE Transactions on Automatic Control, 1963, 8(1): 59-60.
- [10] 李晋宏, 陈锋, 程楠楠, 等. 基于遗传算法的智能组卷研究与应用[J]. 科技信息, 2007(12): 37-38.
- [11] REEVES C. Diversity and diversification in genetic algorithms: Some connections with tabu search [C]// Artificial Neural Nets and Genetic Algorithms. New York: Springer-Verlag, 1993: 344-351.
- [12] 张文修, 梁怡. 遗传算法的数学基础[M]. 西安: 西安交通大学出版社, 2001.

(上接第1873页)

似移动特征的节点分到一个群,从而提高了群的稳定性,并通过限制节点充当群首的时间长度,避免其过早地消耗完能量。仿真实验表明,在节点呈“组移动”特征的网络中,GMBC算法产生的群数目适中,群首改变次数较少,节点充当群首的公平性程度合理。是一种具有实用价值的分群算法。

参考文献:

- [1] CHUNGUNG R, GERLA M. Adaptive clustering for mobile wireless networks[J]. IEEE Journal on Selected Areas in Communications, 1997, 15(7): 1265-1275.
- [2] BASAGNI S. Distributed clustering for Ad Hoc networks[C]// Proceedings of the 1999 International Symposium on Parallel Architectures, Algorithms, and Networks. Washington DC: IEEE Computer Society, 1999: 310-315.
- [3] CHATTERJEE M, DAS S K, TURGUT D. WCA: A weighted clustering algorithm for mobile Ad Hoc networks [J]. Journal of Clustering Computing IEEE, 2002, 5(2): 193-204.
- [4] MAINAK C, SAJAL K. An on-demand weighted clustering algo-

- rithm (WCA) for Ad Hoc networks[C]// IEEE Global Telecommunications Conference. New York: ACM, 2000: 1697-1701.
- [5] HONG X, GERLA M, PEI G, et al. A group mobility model for Ad Hoc wireless networks [C]// Proceedings of ACM/IEEE MSWIM'99. New York: ACM, 1999: 53-60.
- [6] 董超, 杨盘龙, 田畅. 一种Ad Hoc网络组移动模型[J]. 系统仿真学报, 2006, 18(7): 1879-1883.
- [7] MICHELLE X, SCOTT F. A self-Organized clustering algorithm for UWB Ad Hoc networks[C]// WCNC 2004/IEEE Communications Society. New York: IEEE, 2004: 1860-1811.
- [8] INN E R, WONSTON K. Mobility-based d-Hop clustering algorithm for mobile Ad Hoc networks[C]// WCNC 2004/IEEE Communications Society. Piscataway NJ: IEEE, 2004: 2359-2364.
- [9] PRITHWISH B, NAVED K. A mobility based metric for clustering in mobile Ad Hoc networks[C]// IEEE Distributed Computing Systems Workshop 21 International Conference. Washington DC: IEEE Computer Society, 2001: 413-418.
- [10] 王海涛, 田畅, 郑少仁. 一种新型的Ad Hoc网络分簇算法及其性能仿真[J]. 系统仿真学报, 2003, 15(2): 193-197.