

一种基于改进遗传算法的医学题库自动组卷设计与实现

席卫文¹,张春辉²,王飞²,吴冬成³

1.南方医科大学生物医学工程学院,广东广州 510515; 2.南方医科大学教务处,广东广州 510515; 3.南方医科大学教育技术中心,广东广州 510515

【摘要】传统遗传算法适用于大型试题库的自动组卷,具有内在的并行性、全局寻优等特点,但是实现方法复杂,用时较长。为此本文提出一种改进的遗传算法,按照题型和知识点联合分段编码,种群中试卷个体同时满足该题型和知识点分布要求,简化了最优解的求解过程,交叉算子和变异算子采用自适应函数,采用最优保存策略,产生下一代个体。实验结果表明,本算法收敛速度加快,组卷所需时间显著提升,算法的稳定性好,组卷成功率高,已经多次应用于医学高校各课程自动组卷,具有较高的实际应用价值。

【关键词】医学题库;遗传算法;自动组卷;分段编码

【中图分类号】TP391

【文献标志码】A

【文章编号】1005-202X(2016)08-0861-04

Design and implementation of improved genetic algorithm for automatic test paper generation

XI Wei-wen¹, ZHANG Chun-hui², WANG Fei², WU Dong-cheng³

1. School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China; 2. Academic Affairs Office, Southern Medical University, Guangzhou 510515, China; 3. Educational Technology Center, Southern Medical University, Guangzhou 510515, China

Abstract: The traditional genetic algorithm which is suitable for the automatic test paper generation of large test bank is implemented in parallel and achieves global optimization. However, the tradition approach is complex and time consuming. An improved genetic algorithm was proposed in the paper. Since individual papers always satisfied the requirement of question types and knowledge point distribution, the unit could be joint sub-encoding based on question types and knowledge points. Consequently, the optimization process was simplified. The adaptive function was used for crossover operator and mutation operator. The elitist strategy was adopted to produce the next generation of individuals. Experiment results show that the convergence speed of the proposed algorithm is accelerated, and that the stability of the proposed algorithm is enhanced. The proposed method is successfully applied for automatic test paper generation of various courses in medical universities, with high application value.

Key words: medical test bank; genetic algorithm; automatic test paper generation; sub-encoding

前言

在互联网+的教育大背景下,开展网络题库与在线考试,已经成为许多医学院校教学考核的新趋势。建立良好的组卷方式,则是试题库与网考系统的核心^[1]。目前传统的组卷方式包括随机法、回溯法、遗传算法。随机法组卷是指从题库中进行不断重复随机抽取试题,直至满足组卷要求,或者无法抽取满足组卷要求算法

结束。用随机法进行组卷,其过程比较简单,但其具有不可回溯的性质。回溯法属于深度优先算法的一种,是对随机算法的一种改进。回溯法随机抽取第一状态,并记录下来,当搜索失败后,释放上次记录的状态类型,然后通过来回不断的回溯试探,直到组卷的完成。回溯法组卷算法实现较为复杂,占用内存多,空间开销大。随机法和回溯法都不太适用于大型试题库的组卷方式^[2]。遗传算法(Genetic Algorithm)是现在比较新兴和流行的一种组卷算法,遗传算法是模拟生物界的自然选择和遗传变异的机制,来求解复杂问题的随机搜索和优化算法,具有内在的并行性、全局寻优和收敛速度快的特点,因此广泛应用于大型试题库的自动组卷^[3]。

【收稿日期】2016-06-20

【基金项目】广东省科技计划项目(2015A03040103)

【作者简介】席卫文,男,硕士,讲师,主要研究方向:计算机辅助教学,
Tel:020-62789315, E-mail: xww@fimmu.com

1 组卷方法的数学描述

假定组卷时需要给出以下5个约束条件:试卷总分、答题总时间、试卷难度分布、题型分布、知识点分布,假定试卷试题数为 n ,可建立一个 $m \times n$ 的目标矩阵($M=5$),如公式(1)所示^[4]:

$$\begin{bmatrix} a[1][1] & a[1][2] & \cdots & a[1][n-1] & a[1][n] \\ a[2][1] & a[2][2] & \cdots & a[2][n-1] & a[2][n] \\ a[3][1] & a[3][2] & \cdots & a[3][n-1] & a[3][n] \\ a[4][1] & a[4][2] & \cdots & a[4][n-1] & a[4][n] \\ a[5][1] & a[5][2] & \cdots & a[5][n-1] & a[5][n] \end{bmatrix} \quad (1)$$

$a[i][j]$ 表示组卷后抽取出来的第 j 题的第 i 项约束值,分别满足以下5个约束条件:

(1)试卷总分(S):

$$S = \sum_{i=1}^n a[1][i] \quad (2)$$

$a[1][i]$ 表示试卷抽取后中第 i 题的分值,该值可由组卷老师自由设置。

(2)答题用时(D):

$$D = \sum_{i=1}^n a[2][i] \quad (3)$$

$a[2][i]$ 表示试卷抽取后中第 i 题的答题时间,该值可由组卷老师自由设置。

(3)试卷难度(T):

$$T = (\sum_{i=1}^n a[3][i]) / n \quad (4)$$

$a[3][i]$ 表示试卷抽取后中第 i 题预测难度,该值在试题库中已有固定值。

(4)题型分布(Q):

$$Q[j] = \sum_{i=1}^n a[4][i] \times c[j][i] \quad (5)$$

$Q[j]$ 指题型为 j 的分数, j 为某一指定题型,例如选择题等。具体题型分布和题型分值可由组卷老师自由设置,其中:

$$c[j][i] = \begin{cases} 1, & a[4][i] = j \text{ 题型} \\ 0, & a[4][i] \neq j \text{ 题型} \end{cases} \quad (6)$$

$a[4][i]$ 表示试卷抽取后中第 i 题的题型。

(5)知识点分布(Z):

$$Z[j] = \sum_{i=1}^n a[5][i] \times Z[j][i] \quad (7)$$

$Z[j]$ 指知识点为 j 的分数, j 为某一指定知识点,例如第一章等。具体知识点分布、分值可由组卷老师自由设置。

$$Z[j][i] = \begin{cases} 1, & a[5][i] = j \text{ 知识点} \\ 0, & a[5][i] \neq j \text{ 知识点} \end{cases} \quad (8)$$

$a[5][i]$ 表示试卷抽取后中第 i 题所属知识点。

由上述数学模型得知,组卷算法属于一个多目标求优问题,满足条件的最优解不唯一^[5]。为了便于求解,本文采用权重系数法将多目标转化成单目标,目标函数可表示如下:

$$F = \sum_{i=1}^5 a[i] \times |e[i]| \quad (9)$$

$a[i]$ 为第 i 项组卷要求的权重系数, $e[i]$ 为第 i 项组卷要求约束满足误差的绝对值。其中前2项组卷要求:试卷总分、答题总时间,可由组卷老师自由设置,因此为前2项组卷要求的权重系数可以设置为0,即 $a[1]=0$, $a[2]=0$ 。

2 改进的遗传算法设计

2.1 编码策略

传统的遗传算法是将一份试卷映射为一个染色体,组成试卷的各个试题映射为基因,基因的值,直接用试题的题号表示^[6]。也可在编码中加入分段方法,每一段反映一种题型,相同题型的试题题号在同一段编码中,每一段编码长度由试卷中该题型的题数决定^[7]。采用分段编码方式可以克服编码长度过长的缺点,同时种群中试卷个体满足题型分布要求,使得目标函数的求解问题,简化为对试题知识点分布和试题难度二维度要求,从而提高求解速度^[8]。

本文设计的题库系统较为规范,试题库中各课程的试题数量充足,能够完全满足组卷老师对试题题型分布和知识点分布的要求^[9],因此本文采用一种更加有效的编码规则,即将题型和知识点联合分段编码。假定试卷共 N 种题型, M 个知识点,采用 $N \times M$ 段,每一段编码长度等于试卷中同时满足该题型和该知识点的试题数量,具体编码方式参见采用下文公式(11)。采用此联合编码,种群中试卷个体同时满足该题型和知识点分布要求,使得目标函数的求解问题进一步简化为对试题难度一维度要求,从而大大提高了求解速度,算法的求解结果也得到最大的优化。

2.2 适应度函数

在遗传算法中,以适应度大小来区分群体中个体的优劣,因此适应度函数的设计非常重要^[8]。由上述得知,本文通过联合分段编码,目标函数的求解已经简化为对试题难度一维度要求,因此在本文中设计适应度函数为:

$$T = fn - \left| fn - \left(\sum_{i=1}^n a[i] \times b[i] \right) \right| \quad (10)$$

fn 代表设定的期望考生得分, $a[i]$ 为试卷中第 i 题难度值 ($a[i]$ 大于 0 并且小于等于 1), $b[i]$ 为试卷中第 i 题设定的分值, T 的取值越接近 fn , 个体的适应度越好。

2.3 选择算子的设计

根据用户选定的题型和知识点分布要求, 从题库中随机抽取满足条件的试题。生成初始群体, 群体的大小按经验或实验给出, 经遗传操作后产生新一代。

2.4 交叉算子和变异算子

遗传算法的交叉概率 P_c 和变异概率 P_m , 直接影响算法的收敛性^[7]。 P_c 和 P_m 越大, 算法产生新个体的能力就越强。 P_c 和 P_m 越小, 算法使个体趋于收敛的能力越强。本文采用自适应的思想, 在算法的运行过程中对 P_c 和 P_m 进行调整, 让它们随着个体适应度值的增加而变小, 随着个体适应度值的减小而增加。本文染色体个体的交叉和变异均在相同编码段内进行, 即相同题型和知识点在各自的编码组中独立进行交叉和变异操作。

2.5 最优保存策略

经过选择、交叉、变异操作后, 比较新一代的最好个体与上一代的最好个体的适应度, 如下降, 则以上一代最好个体替换新一代的最差个体^[8]。

3 算法实现与实验结果及分析

3.1 算法实现

为验证上述遗传算法的可行性和有效性, 本文采用 asp.net 开发工具编制程序, 实现《外科学》基于改进遗传算法的组卷。《外科学》试题库中题型知识点分布、组卷要求分别如表 1、表 2 所示。

表 1 原有试题库题型、知识点分布

Tab.1 Question type and knowledge point distribution

Knowledge point	A1 type	B1 type	X type	Total
Hepatobiliary surgery	200	102	70	372
Orthopedics	220	105	71	396
Anesthesiology	220	88	80	388
Urology surgery	232	160	92	484
General surgery	192	102	76	370
Total	1064	557	389	2010

其中试卷难度期望为 70 分, 具体实现算法如下:

(1) 按照题型和知识点联合分段编码, 一共分为

表 2 组卷要求

Tab.2 Requirement of test paper generation

Knowledge point	A1 type	B1 type	X type	Total
Hepatobiliary surgery	10	6	4	20
Orthopedics	10	6	4	20
Anesthesiology	10	6	4	20
Urology surgery	10	6	4	20
General surgery	10	6	4	20
Total	50	30	20	100
Each question score	1 point	1 point	1 point	100 point

$3 \times 5 = 15$ 段编码, 编码方式为:

$$G(x) = \left\{ \begin{aligned} & (x[1][1][1] \cdots x[1][1][k_{1,1}]) \cdots (x[i][j][1] \cdots x[i][j][k_{i,j}]) \\ & \cdots (x[3][5][1] \cdots x[3][5][k_{3,5}]) \end{aligned} \right\} \quad (11)$$

其中 $x[i][j][1]$ 为染色体中等于第 i 类题型、第 j 类知识点的第 1 道题在原题库中的编号, $k_{i,j}$ 为试卷中第 i 类题型、第 j 类知识点的试题总数, 即染色体中第 $[i][j]$ 这一段编码的长度。

设置种群规模 $N=20$, 初始个体使用如下 SQL 语句随机产生:

"select top k 编号, 难度 from 题库 where 题型 = $x[i]$ and 知识点 = $y[j]$ order by newid()", 其中 $x[i]$ 为第 i 类题型, $y[j]$ 为第 j 类知识点, k 为第 $[i][j]$ 段编码的长度。

(2) 适应值函数:

$$T = 70 - \left| 70 - \left(\sum_{i=1}^{100} a[i] \times b[i] \right) \right| \quad (12)$$

70 代表设定的期望考生得分, $a[i]$ 为个体试卷中第 i 题难度值 ($a[i]$ 大于 0 并且小于等于 1), $b[i]$ 为第 i 题设定的分值, T 的取值越大, 个体的适应度越好。

(3) 交叉算子初值设置为 $P_c=0.8$, 变异算子初值 $P_m=0.1$, 随着迭代次数的增加, 个体越趋于收敛, P_c 和 P_m 取值可自适应:

$$P_c = 0.8 \times P_{c2}$$

$$P_{c2} = \begin{cases} 1 - t/70, & t > 50 \\ 1, & t \leq 50 \end{cases} \quad (13)$$

t 为该个体的适应度值

(4) 经过选择、交叉、变异操作后, 最优保存策略产生新一代个体, 当适应值 $T < 1$ 或者满足最大迭代数 200, 即可结束遗传算法。

3.2 试验结果与分析

在以上组卷要求下,使用 Dell PowerEdge R710 服务器(cpu 配置为 Intel(R), Xeon(R) 2.40 GHz, 双核, 32 G 内存, 操作系统为 Windows Server 2008 R2 Enterprise), 分别用随机法组卷和标准遗传算法和本文提出的改进遗传算法进行 20 次组卷实验, 最大迭代数均为 200, 试验结果见表 3。

表3 组卷试验结果比较($n=20$)Tab.3 Comparison of experiment results of test paper generation ($n=20$)

Algorithm	Fitness value	Iteration times	Average time/s
Random method	50.22	200	7.5
Standard genetic algorithm	68.35	151	68.4
Improved genetic algorithm proposed in the paper	69.17	108	29.8

从表3可以看出, 试题库题目数量大, 题型、知识点分布合理, 使用遗传算法得到的平均适应值接近试卷期望得分 70, 成功组卷的概率非常高, 使用随机法组卷虽然算法简单, 用时较小, 但是最优解的平均适应值偏低, 成功组卷的概率偏低。对比标准的遗传算法, 本文按照题型和知识点联合分段编码, 简化了最优解的求解过程, 因此算法收敛速度加快, 组卷所需时间有了显著提升, 同时算法的稳定性好, 组卷成功率高。

4 小 结

本文提出一种改进的遗传算法, 用于医学课程的组卷问题。算法按照题型和知识点联合分段编码, 简化了最优解的求解过程, 对比传统的遗传算法收敛速度加快, 组卷用时少, 成功率高。本算法已成功应用于南方医科大学 2010 学年开始的期末考试, 1 038 场次 54 门医学课程的组卷, 具有较高的实际应用价值^[10]。

【参考文献】

- [1] 吴运明, 赵秀玲. 网络考试测评系统的研究现状与发展趋势[J]. 软件导刊, 2010(10): 415-417.

- WU Y M, ZHAO X L. Research status and development trend of network-examination evaluation system[J]. Software Guide, 2010(10): 415-417.
- [2] 乔家庆, 付平, 孟升, 等. 基于个体差异的遗传选择算子设计[J]. 电子学报, 2006, 34(12): 2414-2416.
- QIAO J Q, FU P, MENG S, et al. A genetic selection operator based on difference among individuals[J]. Acta Electronica Sinica, 2006, 34(12): 2414-2416.
- [3] 朱剑冰, 李战怀, 赵娜, 等. 基于混合遗传算法的自动组卷问题的研究[J]. 计算机仿真, 2009, 25(6): 328-332.
- ZHU J B, LI Z H, ZHAO N, et al. Research on the problem of auto-composing test paper based on hybrid genetic algorithm [J]. Computer Simulation, 2009, 25(6): 328-332.
- [4] 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): 78-79.
- YANG L M, CHEN D X. Application of an improved genetic algorithm in composing a test paper [J]. Computer & Digital Engineering, 2004, 32(5): 78-79.
- [5] 尹红卫. 一种改进的遗传算法及其在组卷系统中的应用[J]. 现代计算机, 2006(5): 28-30.
- YI H W. An improved genetic algorithm and the application of auto-composing test paper system[J]. Modern Computer, 2006(5): 28-30.
- [6] 王淑佩, 易叶青. 基于改进自适应遗传算法的组卷研究[J]. 科学技术与工程, 2006(6): 470-471.
- WANG S P, YI Y Q. Investigation of test paper auto-generation based on improved adaptive genetic algorithm [J]. Science Technology and Engineering, 2006(6): 470-471.
- [7] 孙勇, 柏云. 基于遗传算法的试题组卷策略[J]. 淄博学院学报(自然科学与工程版), 2002, 4(3): 27-28.
- SUN Y, BAI Y. Strategy of combination of test papers based on genetic algorithm [J]. Journal of Zibo University (Nat Sci and Eng Ed), 2002, 4(3): 27-28.
- [8] 石中盘, 韩卫. 基于概率论和自适应遗传算法的智能抽题算法[J]. 计算机工程, 2002, 28(1): 141-143.
- SHI Z P, HAN W. An intelligent algorithm to extract examinations based on probability and adaptive genetic algorithm[J]. Computer Engineering, 2002, 28(1): 141-143.
- [9] 文民刚, 席卫文, 张春辉, 等. 医学院校网络考试系统的功能性研究[J]. 中国高等医学教育, 2010(12): 4-5.
- WEN M G, XI W W, ZHANG C H, et al. Research on the function of network examination system in medical college [J]. China Higher Medical Education, 2010(12): 4-5.
- [10] 席卫文, 张春辉, 张立力, 等. 医学网络题库与考试系统的研究与实现[J]. 中国高等医学教育, 2012(2): 27-28.
- XI W W, ZHANG C H, ZHANG L L, et al. Study and realization of medical network question bank and examination system[J]. China Higher Medical Education, 2012(2): 27-28.

(编辑: 黄开颜)