改进遗传算法对医学题库组卷问题的应用研究

肖桂霞,彭春富
(常德职业技术学院 现代教育技术中心,湖南 常德 415000)

摘 要：对含分支题的医学题库进行研究,并对遗传算法做了改进,提出了占位符编码方案、扩位交叉算子和重题优化策略。占位符编码方案能分段定长编码的同时累计各题型段的实际分支题量;扩位交叉算子能智能扩展落在分支题段的交叉点，避免因分支题段局部交叉而出现重题和实际分支题量与条件不符等情况;重题优化策略能快速替换重题,有效缩短组卷时间。仿真结果表明,改进的算法能适应不同题型，在不影响一般题型段抽取与进化的同时，精确控制分支题段的总分支题量和质量,是解决医学题库智能组卷问题的一种有效途径。

关键词：智能组卷;遗传算法;医学题库;重题优化策略;占位符编码;扩位交叉算子
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1674-7720(2013)06-0072-03

Research of generating test paper for medicine question bank by improved genetic algorithm

Xiao Guixia, Peng Chunfu
(Modern Education Technology Center, Changde Vocational Technical College, Changde 415000, China)

Abstract: In this paper, an improved genetic algorithm is proposed for medicine question bank which includes problems with branches. The improved algorithm contains three parts; they are placeholder coding method, extended crossover operator and iterant problems optimization strategy. The placeholder coding method codes the individual into subsections according by the type and the required number of branches of questions. The extended crossover operator can extend the crossover points which locating in the subsections, thus avoiding bringing iterant questions and will not lead to the discrepancy between the actual number of branches and the required one. The iterant problems optimization strategy can replace iterant problems of test paper quickly. Simulation shows that this improved algorithm can applies to questions of all kinds of types, especially for questions with branches, for it can precisely control the number of branches of subsection of this type of questions, and what's more important, it's an effective technique of generating test paper for medicine question bank.

Key words: generating test paper intelligently; genetic algorithm; medicine question bank; iterant problems optimization strategy; placeholder coding; extended crossover operator

随着国家新医改政策的推行与深入，从事医疗卫生事业的人才队伍不断壮大，给这类考试的组考和阅卷带来很大的困难。目前国家医师、护士执业资格考试等理论考核部分仍采用纸质试卷考核方式，各大高校医学专业考核中仅极少数采用了在线抽题组卷的方式，且题型简单暂未扩展到病例分析和标准配伍等分支题题型。设计符合医学题库特点的智能组卷算法对未来医学类网络考试系统的建立有着重要的应用价值和现实意义。

遗传算法是一种模拟生物进化过程和自然遗传机制的过程搜索最优解的算法,与传统的算法相比,遗传算法具有内在并行性、高鲁棒性、全局寻优和收敛速度快的特点。它尤其适用于处理传统搜索方法难以解决的非线性、多约束等复杂问题[1-2]。智能组卷问题是一个典型的多约束问题,遗传算法已经广泛应用于求解该类问题[3-6]。

本文主要对含一个病例描述、多个分支子题的医学混合题库进行研究,针对分支题型和总分支题量难以控制的问题,对遗传算法进行了改进。提出了占位符编码方案、扩位交叉算子和重题优化策略,仿真结果表明,改进后的遗传算法能有效地解决上述问题,完成各种复杂

《电子技术应用》 www.ChinaAET.com

题型题库的抽题组卷任务，具有很好的实用价值。

1 改进的遗传算法

标准遗传算法是针对只含一般题型的题库设计的，如果应用于医学类专业含一般题型和分支题型的混合题库，将出现个体编码长度不定、个体编码中题型分段界限难以判断、不同题型间可能进行了交叉变异操作而导致染色体混乱等问题；更令人困扰的是，分支题段因局部交叉会破坏该编码段题目的总分支题量，导致出现重题。

综上所述，一种更通用、能适应各种不同题型的智能抽题算法对医学题库来说将是非常具有现实意义的。为说明问题，此处约定题数和题量的含义，题数指题目个数，题量指一个题目所含分支数，此分支要记入试卷总题量。只含一般题型的题库其题数和题量是一致的，但含分支题的题库却并非如此，组卷时应该严格按照题量进行组卷而不是题数。

1.1 占位符编码方案

为解决上述问题，提出一种基于占位符的编码方案。该方案根据组卷条件中要求的每种题型的题量统一各题型段的染色体长度。在分题型段进行定长编码的基础上，加入一种占位符。初始化种群时，边抽题边累计某题型段的实际题量，如果该段染色体还未达到指定长度时实际题量已经满足要求，则后面的题就用占位符代替。为方便操作，一般选取题库中不存在的题号作为占位符。

以下举例以更好地阐述。题库中 T_1~T_6 为选择题，$T_7(2)$、$T_8(5)$、$T_9(3)$、$T_{10}(4)$、$T_{11}(2)$、$T_{12}(2)$为分支题(括号中的数字为每题题量，即分支数)，T_{13}~T_{20} 为计算题。抽取 10 题，分题型段题量要求分别为 3、5、2。首先随机抽取 10 题为 T_2、T_5、T_1、T_9、T_7、T_8、T_{12}、T_{11}、T_{19} 和 T_{20}；设占位符为 Z，编号为 -1。则试卷组成为 $T_2T_5T_1T_9T_7ZZZT_{19}T_{20}$，个体编码为 2 5 1 9 7 -1 -1 -1 19 20。

1.2 扩位交叉算子

简单的两点交叉算子容易使某题型编码段产生局部交叉，如果局部交叉落在一般题型段，其结果往往产生重题，这一点可以通过重题优化策略[7]来解决；然而如果局部交叉落在了分支题段，会大大破坏该编码段的结构和质量，使该编码段总分支题量无法达标，同时出现重题。

为此设计了一种扩位交叉算子，这种交叉算子不影响一般题型段的交叉，同时分支题型段交叉后也不会打乱原有题量和出现重题。具体步骤如下：

(1)随机生成染色体长度以内的两个交叉点 P_1 和 P_2。通过交换交叉点的操作确保 P_1 小于 P_2；

(2)获取当前题型段的起始位置 L_1 和 L_2，并判断此题型段是否为分支题型；

(3)若是，则继续判断 P_1 是否落在$[L_1, L_2]$区间。是则扩位 P_1 为 L_1；若否，则转步骤(5)；

(4)判断 P_2 是否在$[L_1, L_2]$区间内，若是则扩位 P_2 为 L_2；

(5)判断题型段指针是否已经指向末尾，若是则结束交叉点的扩位，否则指针向后移，并且转步骤(2)。

图 1~图 3 为交叉点被扩位的示例图。其中，实心点为当前分支题型段的起始界标 L_1 和 L_2，空心点为交叉算子的交叉点 P_1 和 P_2。从图中可以发现，交叉点被扩位后再进行交叉操作使得分支题型段只要有交叉，一定是该题型段的整体交叉。这样一来，该题型段的总分支题量和题目结构都是整体交换的，只要父代是合法的，子代也一定是合法的。

图 1 左交叉点被扩示例

图 2 右交叉点被扩示例

图 3 左右交叉点被扩示例

扩位交叉操作后，变异算子也要做相应改进。为了控制各题型段的题量，需选择与某题题量相等、题型一致，且在当前题型分段中还未出现过的题目进行变异替换，这样可以保留扩位交叉的成果，同时实现分支题型段的微调与更新。

1.3 重题优化策略

上述算法设计中，交叉和变异操作可能导致重题出现。如父代个体 Parent1、Parent2 在位置 3 和位置 7 处进行两点交叉，交叉后得到子代个体 Son1、Son2：

Parent1=T_3 T_5|T_1 T_6 T_9 T_{10} T_{12}|T_{15} T_{18} T_{19}

Parent2=T_2 T_4|T_3 T_7 T_9 T_{11} T_{10}|T_{13} T_{19} T_{20}

Son1=T_3 T_5|T_3 T_7 T_9 T_{11} T_{10}|T_{15} T_{18} T_{19}

Son2=T_2 T_4|T_1 T_6 T_9 T_{10} T_{12}|T_{13} T_{19} T_{20}

其中 Son1 个体出现了重题 T_3，假设在此基础上再对 Son2 在位置 8 进行变异，取与 T_{13} 同题型的 T_{12} 进行替换，得到的 Son2 也将出现重题。

对于重题的处理，大多在每次进化后检查重题，并进行替换[6]。每代种群更新后都必须对整个种群进行一次重题检测与替换，这样会耗费大量的时间；另外算法在进化过程中，不进行去重题干予能最大限度地保护进化成果，带重题的试卷和不带重题的试卷同属于一个进化空间[6]，减少干予能帮助算法搜索更大范围的进化空间；如果只选择同一张试卷中未曾出现的新题进行替换，有可能会破坏现有的进化成果，降低原本优秀个体的适应值。

为解决上述问题，本文提出一种重题优化策略，该策略包含重题甄别和重题替换两部分。重题甄别能快速甄别并替换试卷中的重题。重题替换只对进化后的最优

《电子技术应用》 www.ChinaAET.com

解进行去重题操作[7]。

2 仿真结果及分析

为检验本文提出的改进遗传算法的实用性，分别对不含分支题的一般题库和含分支题的医学题库进行仿真实验。并采用标准遗传算法（算法 1）和本文提出的改进算法（算法 2）分别从题库抽取一定题量的题组成试卷做对比实验，仿真结果均为算法嵌入系统运行 30 次所得的平均性能指标。

2.1 一般题库的仿真结果

首先对 1 000 题的数学题库进行实验。题库中题型 1 占 500 题，题型 2 占 300 题，题型 3 占 200 题，且所有题均只有一个分支。组卷时要求难度级别“易”占 30%、“中”占 50%、“难”占 20%；章节分布可自行设置也可选择按照在题库中所占比例抽取。

表 1 是从中抽取 100 题的仿真结果，其中 3 种题型比例为 5:3:2，卷面分 100 分。表 2 是抽取 200 题的仿真结果，题型比例不变，卷面分 200 分。

表 1 各算法抽取 100 道试题的性能指标对比

	算法 1	算法 2
耗费时间/ms	100.520 8	67.958 3
最优解适应值	0.910 1	0.933 7

表 2 各算法抽取 200 道试题的性能指标对比

	算法 1	算法 2
耗费时间/ms	205.729 2	130.432 3
最优解适应值	0.921 8	0.944 9

从表中可以发现，两个算法最终获得的最优解适应值差不多，并且都满足组卷要求的题量。算法 2 的时间耗费明显低于算法 1，这是因为算法 2 采取了重题优化策略[7]的结果。

2.2 医学题库的仿真结果

构建内科护理 1 000 道题的题库，其中一般题型题 700 道，分支题型题 300 道，且这些分支题所含的分支数均不超过 5，分支数为 2、3、4、5 的题目在题库中的比例为 12:8:3:2；组卷时的难度级别要求、章节分布要求等均与 2.1 节中相同。

如果将题量要求综合到适应值计算中，算法 1 将无法收敛，因此在下列对比中，题量要求从适应值计算中分解出来，单独作为一个评价指标列出来以供分析。表 3 和表 4 分别为从内科护理题库中抽取 100 道试题（分支题题量占 30 道）和 200 道试题（分支题题量占 60 道）的仿真结果。

表 3 算法抽取 100 道试题的性能指标对比

	算法 1	算法 2
耗费时间/ms	121.175 2	102.562 5
最优解适应值	0.919 2	0.928 9
最优解题量	154	100

表 4 算法抽取 200 道试题的性能指标对比

	算法 1	算法 2
耗费时间/ms	210.375 5	182.291 6
最优解适应值	0.925 7	0.946 9
最优解题量	305	200

从表 3、表 4 中可以看出，算法 2 的时间耗费明显低于算法 1，这主要归功于算法 2 采用的重题优化策略[7]，并且这种优势在试卷题量更大时体现得更明显。两种算法虽然都能够收敛，但题量指标并不符合组卷要求。

对医学题库的仿真结果表明，算法 2 的综合性能较好，本文提出的编码方案和遗传算子对解决医学类含多分支题题库的组卷问题非常有效。

从不同题库的仿真结果来看，本文提出的改进遗传算法不仅可以广泛应用于一般题库的组卷问题，更重要的是它解决了含分支题题库抽题组卷时如何精确控制总分支题量的难题，算法对各种不同类型的题库以及各种复杂的题型都具有很好的适应和求解能力。改进的算法已经集成于我院护理系题库系统并投入使用，主要用于期末组卷和自主招生的专业组考。今后的工作将致力于设计动态的变异池来提高算法的收敛速度与精度。

参考文献

[1] 陈国良，王煦法，庄镇泉，等.遗传算法及其应用[M].北京：人民邮电出版社，1996.

[2] 郑金华.多目标进化算法及其应用[M].北京：科学出版社，2007.

[3] 董敏，霍剑青，王晓蒲.基于自适应遗传算法的智能组卷研究[J].小型微型计算机系统，2004，25(1)：82-85.

[4] 朱剑冰，李战怀，赵娜.基于混合遗传算法的自动组卷问题的研究[J].计算机仿真，2009，26(5)：328-331.

[5] 全惠云，范国闯，赵霆雷.基于遗传算法的试题库智能组卷系统研究[J].武汉大学学报(自然科学版)，1999，45(5)：758-760.

[6] 黄小明，刘长安.改进遗传算法在自动组卷系统中的应用[J].科学技术与工程，2010，10(8)：1999-2006.

[7] 肖桂霞，赵武初，朱伟，等.基于遗传算法智能组卷的去重题方法[J].计算机工程，2012，38(11)：150-152.

（收稿日期：2012-11-27）

作者简介：

肖桂霞，女，1983 年生，硕士研究生，信息系统项目管理师，主要研究方向：进化计算，智能算法。

彭春富，男，1976 年生，硕士研究生，信息系统项目管理师，主要研究方向：嵌入式系统，智能算法。