

# 混合编码遗传算法在测试数据生成中的应用

李喜艳, 张文宁, 周清雷

(郑州大学信息工程学院 郑州 450001)

**摘要:** 针对传统遗传算法存在局部搜索能力差、未成熟收敛和多峰优化时常有漂移现象的缺陷, 充分利用格雷码来消除 Hamming 悬崖问题, 实数编码来获得大空间搜索任意精度的优势, 从而获得混合编码的遗传算法. 引入黄金分割点实现 2 种编码方法的转换, 设计了测试数据生成的混合编码遗传算法. 通过实例说明测试数据如何自动生成, 结果表明生成测试数据的效果较好.

**关键词:** 测试数据; 遗传算法; 格雷码; 实数编码; 黄金分割点

**中图分类号:** TP 311

**文章编号:** 1671-684X(2009)03-0022-04

## 0 引言

软件测试是保证软件质量和可靠性的主要手段, 而测试数据的设计是软件测试的重要环节, 测试数据的自动生成是其关键和难点. 人工智能中的遗传算法是公认较好的一种测试数据生成技术, 这种方法的关键在于适应度函数的设计. 迄今为止, 很多学者在基于遗传算法的测试数据生成方面获得了很多成果<sup>[1-3]</sup>. 由于遗传算法本身存在局部搜索能力差、未成熟收敛和多峰优化过程中常有漂移现象发生, 基于改进遗传算法的测试数据生成研究得到了很好发展<sup>[3-5]</sup>. 本文采用混合编码的策略改进遗传算法, 引入黄金分割点实现格雷码和实数编码 2 种编码方法的转换, 把混合编码的遗传算法应用到测试数据自动生成中, 使得测试数据生成的效率和质量都比较高.

## 1 基于混合编码遗传算法的测试数据生成模型

基于混合编码遗传算法的测试数据生成模型结构如图 1 所示( $g$  代表进化代数, 初始值为 0,  $mg$  为最大进化代数,  $a = 1 - 0.618$  是 2 种编码方法转换的调节参数,  $F$  为评价函数), 主要由驱动模块、被测模块和混合编码遗传算法模块组成.

### 1.1 驱动模块

驱动模块的作用是用来模拟被测模块的上级调用模块, 功能比真正的上级模块简单得多, 它只完成接受测试数据, 以上级模块调用被测模块的格式驱动被测模块. 在该模型中, 驱动模块接受混合编码遗传算法随机生成的选定路径的参数, 驱动被测模块的运行.

### 1.2 被测模块

被测模块的输入由输入空间  $D$  的参数集  $x(x_1, x_2, \dots, x_n)$  所组成, 主要利用程序插装技术, 对被测程序模块进行插装, 通过“分支函数叠加法”构造针对路径覆盖的评价函数. 设被测程序在  $x$  上通过路径  $p$ , 该目标可简化为一系列的子目标, 用函数极小化搜索技术可解决每个子

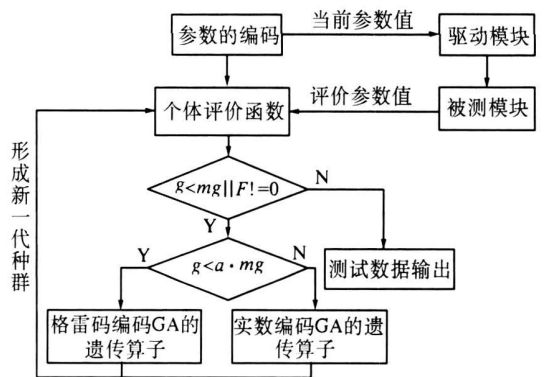


图 1 基于混合编码遗传算法的测试数据生成模型

Fig. 1 Test data generation model based on hybrid coded genetic algorithm

收稿日期: 2009-03-09

基金项目: 国家高技术研究发展计划(863)项目, 编号 2007AA010408.

作者简介: 李喜艳(1983-), 女, 硕士研究生, 主要从事软件工程、信息安全研究. E-mail: xiyanni2006@163.com.

目标.

假定分支谓词是简单关系表达式,具有如下形式: $A \text{ op } B$ .其中  $A$  和  $B$  是算术表达式,关系运算符  $\text{op} \in \{<, \leq, >, \geq, =, \neq\}$ . 目标函数的构造方法是检测条件语句的真假值关系,若能满足给定的真假值,则评价函数的适应度函数值为 0, 否则进行如表 1 所示的运算.

分支函数是一个分支谓词到实际值的映射,可以量化地反映在测试数据的驱动下,若下一条路径所经过的条件判断语句不止一个,则将每个适应度函数值进行累加,最后的和为此个体的评价函数值. 选定路径上有  $m$  个分支点,  $n$  个参数  $x_i (1 \leq i \leq n)$ , 则  $m$  个分支函数为:  $\varphi_1 = f_1(x_1, x_2, \dots, x_n)$ ,  $\varphi_2 = f_2(x_1, x_2, \dots, x_n)$ ,  $\dots$ ,  $\varphi_m = f_m(x_1, x_2, \dots, x_n)$ , 该路径的评价函数值为

$$F = \Phi(\varphi_1) + \Phi(\varphi_2) + \dots + \Phi(\varphi_m).$$

其中,  $x \leq 0$  时,  $\Phi(x) = 0$ ;  $x > 0$  时,  $\Phi(x) = x$ .

表 1 评价函数值计算方法<sup>[9]</sup>

表达式	评价函数值	
	$\Phi$ (表达式为true)	$k$ (表达式为false)
$A = B$	0	$\text{Abs}(A - B) + k$
$A \neq B$	0	$k$
$A < B$	0	$(A - B) + k$
$A \leq B$	0	$(A - B) + k$
$A > B$	0	$(B - A) + k$
$A \geq B$	0	$(B - A) + k$
Boolean	0	$k$

### 1.3 混合编码遗传算法模块

混合编码遗传算法模块是通过搜索参数空间  $D$  来寻找评价函数  $F$  的最小值,使得能够覆盖所选路径. 每迭代一次,分支函数点列随着参数的变化相应发生改变. 随着迭代次数的增加,评价函数的值距 0 越来越近,最后取值为 0, 这时找到的一组参数就是所需要的测试数据.

### 1.4 模块交联

被测模块和混合编码遗传算法模块之间的交联关系是:混合编码遗传算法模块通过把当前的参数传送给被测模块,被测模块在驱动模块驱动下运行,运行完成后得到评价函数  $F$  的值,然后传送给混合编码遗传算法模块进行评价与演化.

### 1.5 停止准则

基于混合编码遗传算法的测试数据生成的停止准则:评价函数值为 0, 或者进化到最大迭代数  $mg$ .

## 2 应用于测试数据生成的混合编码遗传算法设计

标准的遗传算法运用模式定理分析编码机制,建议采用二进制编码,但二进制编码存在着连续离散化时的映射误差,也不便反映所求问题的结构特征,且存在着 Hamming 悬崖问题. 对于一些连续优化的问题,由于遗传算法的 2 个主要的遗传算子(交叉和变异)都是在一定的概率下随机的、无指导的操作的,从而使得局部搜索能力比较差,尤其是在进行数值优化时难以达到精度要求,而实现时又需要频繁的解码,使运算过程变慢. 为了克服二进制编码的缺点,可以消除 Hamming 悬崖问题的格雷码编码方法,以及便于大空间搜索、可取得搜索空间任意精度的浮点数编码策略越来越受到重视.

用混合编码的遗传算法在实现匹配追踪算法中取得了很好的效果<sup>[9]</sup>:不容易陷入局部收敛,具有很强的跳出局部极值的能力,且收敛速度较快. 本文的遗传算法亦采用混合编码策略,先用格雷码进行较大空间的遗传搜索,进化后期再用浮点数编码进行小范围的寻优,由于采用了 2 种编码方法,借助黄金分割点在科学实验中的完美分割优势,引入黄金分割点来调节 2 种编码方法的转换,调节参数设为  $\alpha = 1 - 0.618$ , 当进化代数小于最大进化代数的调节参数倍时,采用格雷码编码方法,当进化代数不小于最大进化代数的调节参数倍时,采用实数编码方法.

基于测试数据生成的混合编码遗传算法本身参数主要有 5 个:种群大小  $n$ 、基因链长度  $l$ 、交叉概率  $p_c$ 、变异概率  $p_m$  和调节参数  $\alpha$ . 种群大小  $n$  太小时难以求出最优解,太大则增长了问题求解的收敛时间,一般取  $n = 30 \sim 160$  为宜;交叉概率  $p_c$  太小时难以向前搜索,太大则容易破坏具有高适应度值的个体结构,一般取交叉概率  $p_c = 0.2 \sim 0.6$  为宜;变异概率  $p_m$  太小时难以产生新的基因结构,太大则使遗传算法变成了单纯的随机搜索,一般取  $p_m = 0.005 \sim 0.01$  为宜<sup>[9]</sup>. 基因链长度  $l$  会影响遗传操作的速度,基因链越长,遗传操作越复杂,处理速度越慢;反之,基因链越短,所包含的问题参数信息越少,遗传计算速度越快.



代,有输出结果.

对实验过程中生成的种群数据分析发现,传统遗传算法生成的测试数据易陷入局部最优解,而混合编码的遗传算法能够很快地收敛到问题的最优解,并且保持种群的多样性,有效地改善了未成熟收敛.

## 4 结论

混合编码的遗传算法改善了传统遗传算法的不足,在上述实例阐述的过程中,基于混合编码遗传算法的测试数据自动生成取得了很好的效果,但对于输入的数据需计算给定路径上谓词中变量的当前值,测试数据自动生成需进一步研究.

## 参考文献:

- [1] 金虎,李志蜀,张磊,等.基于面向路径的遗传算法的测试用例自动生成[J].计算机工程,2007,33(3):21-23.
- [2] 周献中,孙勇成,江金龙.基于使用模型和遗传算法的测试数据自动生成技术[J].兵工学报,2006,27(6):1051-1055.
- [3] 傅博.基于模拟遗传算法的软件测试数据自动生成[J].计算机工程与应用,2005,25(12):82-84.
- [4] 赵正文,康耀红,方磊坤,等.信息检索中的遗传算法应用研究[J].郑州大学学报:理学版,2006,38(4):64-68.
- [5] 夏芸,刘峰.基于免疫遗传算法的软件测试数据自动生成[J].计算机应用,2008,28(3):723-725.
- [6] 邢恺,伦立军.测试数据自动生成方法[J].计算机技术与发展,2006,16(9):53-55.
- [7] 范虹,孟庆丰,张优云.用混合编码遗传算法实现匹配追踪算法[J].西安交通大学学报,2005,39(3):295-299.
- [8] 云庆夏.进化算法[M].北京:冶金出版社,2000.

# Application of Hybrid coded Genetic Algorithm in the Test Data Generation Study

LI Xi yan, ZHANG Wen ning, ZHOU Qing lei

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

**Abstract :** The traditional genetic algorithm has poor local search capability, premature convergence and multi peak optimization problem often drifting phenomenon. The advantage of Gray Code is made full use of, which can eliminate hanging shore Hamming problem, and real coded, which can get arbitrary accuracy in the access of large space search, to get hybrid coded genetic algorithm. The golden section point is introduced to implement two coding methods conversion, and test data generation hybrid coded genetic algorithm is designed. How to automatically generate test data is illustrated by examples. The experimental result shows that the generated test data is better.

**Key words :** test data; genetic algorithm; Gray Code; real coded; golden section point