遗传算法的一种特例——正交试验设计法*

吴浩扬, 常炳国, 朱长纯

(西安交通大学 电信学院,陕西 西安 710049)

E-mail: wuhaoyang@263.net

摘要: 简要介绍正交试验设计法与遗传算法的基本原理,分析它们之间的内在关系,指出正交试验设计法可以认为是遗传算法的一种特例,即它是一种初始种群固定的、只使用定向变异算子的、只进化一代的遗传算法.计算结果表明,正交试验设计法可以解决一般遗传算法中的最小欺骗问题.

关键词: 正交试验设计法;遗传算法;最小欺骗问题

中图法分类号: TP18 **文献标识码**: A

正交试验设计法是一种试验的优化设计技术,由田口玄一等人于1949年提出.该方法利用正交表来安排试验,以寻找试验的优化方案,其优点是它能够以相当少的试验次数、非常短的试验时间和很低的试验费用得到满意的试验结果.

遗传算法是一种借鉴自然界中"适者生存,优胜劣汰"思想的全局优化算法,由美国学者J.H. Holland于20世纪70年代提出.它把求解问题的可能解看做是用染色体串编码的个体,大量个体(或称为染色体)构成种群,用个体的适应度函数值作为个体优劣的评价指标,在种群的进化过程中,不断使用选择、交叉、变异这3种遗传算子,使问题的解不断进化,直至产生全局最优解.遗传算法特别适合处理传统搜索方法所不能解决的复杂问题和非线性问题.

目前,尽管这两种方法均已分别得到了广泛的研究,但有关它们之间的比较研究还尚未有文献报道.本文首先简要介绍正交试验设计法和遗传算法的基本原理,然后分析它们之间的内在关系,指出正交试验设计法可以认为是遗传算法的一种特例,并说明正交试验设计法可克服一般遗传算法中的最小欺骗问题.

1 正交试验设计法与遗传算法概述

1.1 正交试验设计法[1]

在实际试验场合中,多因素、多水平试验是常见的情形.如果对每个因素的每个水平都相互搭配进行全组合试验,然后从所有的搭配方案中找出最优方案,则总的试验次数将会爆炸式地增长.正交试验设计法是解决多因素、多水平试验问题的一种有效方法,它利用正交表安排试验,对试验方案作最优设计.

我们以4因素、3水平的无交互作用试验为例.首先,构造如表1所示的正交表$L_9(3^4)$,表中字母L表示正交表;数字4表示试验中要考虑4种因素(这里没有考虑因素间的交互作用);数字3表

* 收稿日期: 1999-07-02; 修改日期: 1999-10-22

基金项目: 国家自然科学基金资助项目(69776004)

作者简介: 吴浩扬(1976--),男,江苏泰兴人,博士生,主要研究领域为电子鼻,神经网络;常炳国(1965-),男,宁夏银川人,博士生,主要研究领域为智能传感器,数据融合;朱长纯(1936-),男,辽宁辽阳人,教授,博士生导师,主要研究领域为真空微电子.

示每个因素有3种可能的取值(水平);数字9表示该表有9行,对应于9个试验方案,每个试验方案的效果用试验效果值来表征.对于一般的正交表$L_n(s^r)$而言,各因素、各水平之间的搭配是均衡分散的,即每个因素的每个水平均出现n/s次,任意两列中各种组合水平对出现的个数为n/s^2个.这样,对每个因素的每个水平而言,都进行了n/s次试验,这n/s次试验的效果平均值就代表了该因素的相应水平对试验效果的影响.比较该因素的所有水平对试验效果的影响,挑选出对试验效果影响最大的水平,即为该因素的最优水平.将每个因素的最优水平组合在一起,就构成了最优试验方案.对4因素、3水平无交互作用问题而言,正交试验设计法需要做9次试验,全组合法需要做$3^4=81$次试验.可见,正交试验设计法大大减少了试验次数,且因素数和水平数越大,该方法的优越性就越明显.

Table 1 Orthogonal array

表1 正交表

Order①	Factor② A	Factor B	Factor C	Factor D
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

①试验号,②因素.

1.2 遗传算法

遗传算法是一种全局并行的搜索寻优方法,它通过选择复制、交叉、变异等3种遗传算子的作用,使优化群体不断进化,最终收敛于最优状态[2].标准遗传算法的实现步骤为:(1)设置种群规模、个体编码方案、交叉概率、变异概率等控制参数;(2)利用3种遗传算子,对当前代的个体作繁殖,产生新个体;(3)淘汰父代中适应度较低的个体,并将适应度高的新个体与父代中保留下来的个体组合为新种群;(4)若达到计算精度或超过预设遗传进化次数,则结束计算,否则返回步骤(2).

遗传算法以Holland提出的模式定理为数学基础.考虑用三元字母表$V=\{0,1,*\}$编码的串$A=a_1,a_2,\ldots,a_L$,$*$代表不确定字母,模式H是指编码空间的一个子集.模式的定义长度$\delta(H)$是指模式H中第1个确定位置与最后一个确定位置之间的距离,模式的阶$O(H)$是指模式H中确定位置出现的个数,模式的适应度$f(H)$是指属于模式H的个体的平均适应度.设种群X,种群规模N,种群个体x_i,x_j被编码为$a_1^ia_2^i\ldots a_L^i$和$a_1^ja_2^j\ldots a_L^j$,L是个体的编码长度,随机设定交叉点r、变异点s,则个体x_i,x_j在r处单点交叉产生的新个体为$x'_i=a_1^i\ldots a_r^i\ a_{r+1}^j\ldots a_L^j$,$x'_j=a_1^j\ldots a_{r-1}^ja_r^i\ldots a_L^i$,个体$x_i$在$s$处随机变异产生的新个体为$x''_i=a_1^i\ldots a_{s-1}^ia''a_{s+1}^i\ldots a_L^i$,其中$a''$等于$a_s^i$随机取补.个体$x_i$在$s$处定向变异产生的新个体为$x'''_i=a_1^i\ldots a_{s-1}^ia_0a_{s+1}^i\ldots a_L^i$,其中$a_0$为指定的变异值.

根据以上定义,模式定理可表述为:具有短的定义长度、低阶且适应度在群体平均适应度之上的模式在遗传算法迭代过程中将按指数增长率被采样.

定理的数学表示式为

$$m[H(t+1)]\geqslant m[H(t)]\frac{f(H(t))}{\bar{f}(t)}\left[1-\frac{P_c\delta(H)}{L-1}\right](1-P_m)^{O(H)},$$

其中 $m[H(t+1)]$是种群在第 $t+1$ 次迭代时模式 H 的期望个数;$m[H(t)]$是种群在第 t 次迭代时模式 H 的期望个数;P_c 是杂交概率;P_m 是变异概率;$\tilde{f}(t)$是整个种群的平均适应度.误差项$\{1-[P_c\delta(H)/(L-1)]\}(1-P_m)^{O(H)}$来自于杂交、变异算子对模式 H 的破坏,当它很小时,如果 $f(H)>\tilde{f}(t)$,则 $m[H(t+1)]$变大.如此迭代下去,$m[H(t)]$将以指数增长率增加.

由模式定理可知,定义长度短的、低阶和适应度好的模式将以较大的概率被采样、重组而形成具有潜在的更高适应度的串.遗传算法不是通过逐一测试各个模式组合来建立高适应度的串,而是利用过去的染色体中最好的部分来构造越来越好的串.由于具有短的定义长度、低阶和高适应度的模式在遗传算法中所起的巨大作用,它们被特别称为基因块.遗传算法的基因块假设是指如果把种群中不同个体的不同优良基因块组合在一起,则有望产生出更优良的个体.

Holland 还证明,遗传算法的一大优点是它的隐并行性,即算法在处理 N 个个体的时候,也同时处理 $O(N^3)$个模式,因此每代只需执行与种群规模成比例的计算量,就可以同时收到对 $O(N^3)$个模式进行处理的目的,且无需额外的存储.

2 正交试验设计法与遗传算法的内在关系

从前面的讨论可知,正交试验设计法和遗传算法都是优化方法,两者都是试图用尽可能少的搜索次数来找到求解问题的优化解.考虑一般的含交互作用项的正交试验,如果把整个正交表看做是一个种群,正交表的每一行看做一个用染色体串表征的个体,则正交试验设计要考虑的因素就对应于染色体上的基因,n 个因素的交互作用项对应于 n 阶模式,试验方案的效果值对应于个体的适应度.这样可以看出,正交试验设计法与遗传算法之间有下面介绍的一些联系.

2.1 正交试验设计法是遗传算法的特例

正交试验设计法在做完正交表所安排的试验后,对每个因素及交互作用项的所有可能水平进行分析,挑出对试验效果影响最大的那些因素及交互作用项,把它们的最优水平组合在一起构成一个新的试验方案,并认为这个方案就是最优试验方案.用遗传算法的术语来说,正交试验设计法是对染色体中的所有模式进行分析,挑出对个体适应度影响最大的那些模式,把这些模式组合在一起,构成一个新的染色体,并认为这个染色体就是种群的最优个体,即试验的最优方案.根据第 1.2 节中基因块的假设可知,对试验效果影响最大的那些因素及交互作用项在正交试验设计法中的作用就相当于基因块在遗传算法中的作用.

由上述分析可知:(1) 在正交试验设计法中,初始种群是固定的(由正交表确定),种群中的个体不进行选择复制和交叉操作,算法只对整个种群作定向变异操作,定向变异的结果是产生一个新的个体(即所谓的最优试验方案);(2) 因为种群在进化一代后就得到了最优试验方案,所以可以说,在正交试验设计法中种群只进化一代.基于这两点,我们可以认为正交试验设计法实际上是遗传算法的一种特例,即正交试验设计法是一种初始种群固定的、只使用定向变异算子的、只进化一代的遗传算法.比较而言,一般的遗传算法随机初始种群,利用选择复制、交叉、变异这 3 种遗传算子,对种群作多代进化,显然,它产生的优化解要优于正交试验设计法产生的优化解,但遗传算法的步骤比正交试验设计法要复杂,所需的试验次数也要多.

2.2 处理交互作用项的比较

由于正交试验设计法在利用正交表安排有交互作用的试验时,不同的交互作用项有可能会被安排在正交表的同一列,即出现混杂现象.因此,要避免混杂现象的出现,就必须选用更大的正交

表.但是因素间可能的交互作用项的个数将随着因素个数的增加而呈组合爆炸式地增大,使得对交互作用项作全面分析将变得非常困难而近乎不可能.相反,遗传算法的隐含并行性使得遗传算法能在处理 n 个个体的时候同时处理 $O(n^3)$ 个模式.由前述讨论可知,n 个因素的交互作用项对应于 n 阶模式,所以遗传算法能在处理 n 个个体的时候同时处理 $O(n^3)$ 个交互作用项,即遗传算法处理交互作用项的效率高于正交试验设计法.

3 利用正交试验设计法求解遗传算法的欺骗问题

遗传算法的欺骗问题[3,4]是指,将遗传算法的低阶、高适应度的基因块组合成高阶模式后,高阶模式的适应度反而降低,使得遗传算法不能发现高阶、高适应度的基因块,最终导致算法发散,找不到最优解.一般的遗传算法的最小欺骗问题表述如下:

设有一组 4 个 2 阶模式,它们具有两个定义位置,每个模式及相应的适应度如下所示:

$$**0*****0*\ f_{00},\qquad **0*****1*\ f_{01},$$

$$**1*****0*\ f_{10},\qquad **1*****1*\ f_{11},$$

设模式 11 满足最优条件 $f_{11}>f_{00}$,$f_{11}>f_{01}$,$f_{11}>f_{10}$,但是由于存在着欺骗条件 $f_{00}+f_{01}>f_{10}+f_{11}$,算法将会认为 0* 模式优于 1* 模式,根据模式定理,0* 模式将比 1* 模式以更大的概率被采样、重组,这就是欺骗现象.根据最优条件和欺骗条件可以推出 $f_{00}>f_{10}$,$f_{01}>f_{10}$.根据 $f_{01}>f_{00}$ 和 $f_{01}<f_{00}$,可将欺骗问题分为两类:第 I 类欺骗问题的条件是 $f_{11}>f_{01}>f_{00}>f_{10}$ 和 $f_{00}+f_{01}>f_{10}+f_{11}$;第 II 类欺骗问题的条件是 $f_{11}>f_{00}>f_{01}$ 和 $f_{00}+f_{01}>f_{10}+f_{11}$.

考虑一个特殊的最小欺骗问题,即种群的个体只包含两个基因.相应地,这时在正交试验设计法中只包含两个因素 A 和 B 以及一个两因素交互作用项 $A\times B$,构造正交表 $L_4(2^3)$ 如下:

Table 2 Orthogonal array $L_4(2^3)$

表 2 正交表 $L_4(2^3)$

Experiment①	Factor②A	Factor B	Interaction③$A\times B$	Quality④
1	0	0	0	f_{00}
2	0	1	1	f_{01}
3	1	0	1	f_{10}
4	1	1	0	f_{11}
K_0	$(f_{00}+f_{01})/2$	$(f_{00}+f_{10})/2$	$(f_{00}+f_{11})/2$	
K_1	$(f_{10}+f_{11})/2$	$(f_{01}+f_{11})/2$	$(f_{01}+f_{10})/2$	
Difference⑤	d_A	d_B	$d_{A\times B}$	

①试验,②因素,③交互作用项,④试验效果,⑤极差.

其中 K_0 行的值对应于因素取 0 水平的有关试验的效果平均值,K_1 行的值对应于因素取 1 水平的有关试验的效果平均值,极差是 K_0 与 K_1 差的绝对值.

$$d_A=\frac{1}{2}|f_{11}+f_{10}-f_{00}-f_{01}|,\quad d_B=\frac{1}{2}|f_{11}+f_{01}-f_{00}-f_{10}|,$$

$$d_{A\times B}=\frac{1}{2}|f_{11}+f_{00}-f_{01}-f_{10}|.$$

极差越大表明对应因素或交互作用项对试验效果的影响就越强.

对于第 I 类欺骗问题:$f_{11}>f_{01}>f_{00}>f_{10}$,$f_{00}+f_{01}>f_{10}+f_{11}$.

$$d_{A\times B}=\frac{1}{2}|f_{11}+f_{00}-f_{01}-f_{10}|=\frac{1}{2}|(f_{11}-f_{01})+(f_{00}-f_{10})|,$$

$$d_A=\frac{1}{2}|f_{11}+f_{10}-f_{00}-f_{01}|-\frac{1}{2}|(f_{11}-f_{01})-(f_{00}-f_{10})|,$$

因为$(f_{11}-f_{01})>0$,$(f_{00}-f_{10})>0$,所以极差$d_{A\times B}>d_A$.

$$d_{A\times B}=\frac{1}{2}|f_{11}+f_{00}-f_{01}-f_{10}|=\frac{1}{2}|(f_{11}-f_{10})-(f_{01}-f_{00})|,$$

$$d_B=\frac{1}{2}|f_{11}+f_{01}-f_{00}-f_{10}|=\frac{1}{2}|(f_{11}-f_{10})-(f_{01}-f_{00})|,$$

因为$(f_{11}-f_{10})>0$,$(f_{01}-f_{00})>0$,所以极差$d_B>d_{A\times B}$.

可见,因素B、交互作用项$A\times B$起主要作用.对因素B来说,$K_1=(f_{01}+f_{11})/2>(f_{00}+f_{10})/2=K_0$,所以因素$B$取1水平为好(即对应于2、4号试验);对交互作用项$A\times B$来说,$K_0=(f_{00}+f_{11})/2>(f_{01}+f_{10})/2=K_1$,所以交互作用项$A\times B$取0水平为好(即对应于1、4号试验).综合判断后得出,4号试验(即11模式)为最优方案,故对第Ⅰ类欺骗问题而言,欺骗被克服了.

对于第Ⅱ类欺骗问题:$f_{11}>f_{00}>f_{01}>f_{10}$,$f_{00}+f_{01}>f_{10}+f_{11}$.

$$d_{A\times B}=\frac{1}{2}|f_{11}+f_{00}-f_{01}-f_{10}|=\frac{1}{2}|(f_{11}-f_{10})+(f_{00}-f_{01})|$$

$$d_B=\frac{1}{2}|f_{11}+f_{01}-f_{00}-f_{10}|=\frac{1}{2}|(f_{11}-f_{10})-(f_{00}-f_{01})|,$$

因为$(f_{11}-f_{10})>0$,$(f_{00}-f_{01})>0$,所以极差$d_{A\times B}>d_B$.

$$d_A=\frac{1}{2}|f_{11}+f_{10}-f_{00}-f_{01}|=\frac{1}{2}|(f_{11}-f_{00})-(f_{01}-f_{10})|,$$

$$d_B=\frac{1}{2}|f_{11}+f_{01}-f_{00}-f_{10}|=\frac{1}{2}|(f_{11}-f_{00})+(f_{01}-f_{10})|,$$

因为$(f_{11}-f_{00})>0$,$(f_{01}-f_{10})>0$,所以极差$d_B>d_A$.

可见,还是交互作用项$A\times B$和因素B起主要作用.类似前面的推理可知,交互作用项$A\times B$应取0水平(即对应于1、4号试验),因素B应取1水平(即对应于2、4号试验).综合判断后得出,4号试验(即11模式)为最优方案,故对第Ⅱ类欺骗问题而言,欺骗也被克服了.

以上就个体只包含两个基因的情况进行讨论.一般地,当个体包含n个基因时,对应的正交试验设计法就要包含n个因素、$n(n-1)/2$个两因素交互作用项.类似于上述步骤,可以对n个因素中的任意两个因素及对应的一个两因素交互作用项进行讨论,同理可以证明正交试验设计法能解决两类最小欺骗问题,给出正确结果(即11模式为最优解).综上所述,正交试验设计法可以解决一般遗传算法中的最小欺骗问题.

4 结束语

正交试验设计法与遗传算法是两种分别于20世纪40年代和70年代由日本和美国学者提出的优化方法,并已在实际中得到了广泛的应用.虽然这两种方法看似无关,然而经过详细分析,我们认为两者之间有许多联系,具体结论如下:

(1) 正交试验设计法是遗传算法的一种特例,即正交试验设计法是一种初始种群固定的、只使用定向变异算子的、只进化一代的遗传算法.

(2) 遗传算法的步骤比正交试验设计法复杂,所需的试验次数也要多于正交试验设计法的试验次数,但它产生的解要优于正交试验设计法产生的解.

(3) 遗传算法的隐并行性使得它在处理交互作用项时,效率比正交试验设计法要高.

(4) 本文计算结果表明,正交试验设计法可解决一般遗传算法中的最小欺骗问题.

References:

[1] Chen, Kui. Design and Analysis of Experiments. Beijing: Tsinghua University Press, 1996. 94~120 (in Chinese).

[2] Liu, Yong, Kang, Li-shan, Chen, Yu-ping. Non-Numerical Parallel Algorithms (Ⅱ)—Genetic Algorithms. Beijing: Science Press, 1995. 22~60 (in Chinese).

[3] Chen, Jian-an, Guo, Da-wei, Xu, Nai-ping, *et al*. A review on the theory for the genetic algorithm. Journal of Xidian University, 1998,25(3):363~368 (in Chinese).

[4] Huang, Yan, Jiang, Pei, Wang, Jia-song, *et al*. A solution to deceptive problems in genetic algorithm based on an adjustable mutation operator. Journal of Software, 1999,10(2):216~219 (in Chinese).

附中文参考文献:

[1] 陈魁.试验设计与分析.北京:清华大学出版社,1996.94~120.

[2] 刘勇,康立山,陈毓屏.非数值并行算法(第2册)——遗传算法.北京:科学出版社,1995.22~60.

[3] 陈建安,郭大伟,徐乃平,等.遗传算法理论研究综述.西安电子科技大学学报,1998,25(3):363~368.

[4] 黄焱,蒋培,王嘉松,等.基于可调变异算子求解遗传算法的欺骗问题.软件学报,1999,10(2):216~219.

A Special Case of Genetic Algorithm——Orthogonal Experimental Design Method

WU Hao-yang, CHANG Bing-guo, ZHU Chang-chun

(*School of Electronics and Information, Xi'an Jiaotong University, Xi'an* 710049, *China*)

E-mail: wuhaoyang@263.net

Received July 2, 1999; accepted October 22, 1999

Abstract: In this paper, the basic principles of orthogonal experimental design method and genetic algorithm are discussed briefly, and the relation between them is analyzed in detail. The paper indicates that orthogonal experimental design method is a special case of genetic algorithm, i.e. a genetic algorithm with a fixed initial population, an oriented mutation operator and one evolution epoch. It is shown that the orthogonal experimental design method can overcome minimum deception problems existing in genetic algorithm.

Key words: orthogonal experimental design method; genetic algorithm; minimum deception problem