文章编号:1000-6788(2005)12-0057-05

基于整数编码遗传算法的均匀设计表构造

张礼兵[1,2],程吉林[2],金菊良[1],蒋晓红[2]
(1. 合肥工业大学土木与建筑工程学院,安徽 合肥 230009;2. 扬州大学水利科学与工程学院,江苏 扬州 225009)

摘要: 均匀设计的核心问题是均匀设计表的合理构造,其实质是一个以某类均匀度为目标的优化问题.将寻优能力极强的遗传算法引入该构造过程使均匀设计表的自动构造成为可能.研究结果表明,与方幂生成向量法,正交设计扩展法,拉丁方法以及门限接受法等常用方法相比,基于整数编码遗传算法的均匀设计表构造法可获得各种混合水平,n 较大的 $U_n(q^s)(q\leq n)$ 表,且计算精度高,速度快,稳定性好,并易与其它智能算法相结合,具有较高的应用价值.

关键词: 试验设计; 均匀设计表; 整数编码遗传算法
中图分类号: O224;TP18　　**文献标识码**: A

A Uniform-designs-table Construction Method Based on Integer-coded Genetic Algorithm

ZHANG Li-bing[1,2], CHENG Ji-lin[2], JIN Ju-liang[1], JIANG Xiao-hong[2]
(1. College of Civil Engineering, Hefei University of Technology, Hefei 230009, China; 2. College of Conservancy and Hydraulic Engineering, Yangzhou University, Yangzhou 225009, China)

Abstract: The key point of the uniform designs is the reasonable construction of uniform designs tables. The process of constructing a uniform designs table is actually an optimization problem with the object of certain uniform index. The genetic algorithm (GA) with excellent optimizing ability is introduced to solve the problem, which tries to make it possible to construct a uniform designs table automatically. The result of studying shows that, compared with prime power generating-vectors methods (PP-GVM), extended orthogonal designs (EOD), Latin hypercube (LH) and threshold accepting (TA), the methods of generating-vectors based on integer-coded genetic algorithm (IGA-GVM) has characters of good stability, high calculating precision and efficiency. With the value of application, IGA-GVM can be used to construct any large-scale-n $U_n(q^s)(q\leq n)$ with any levels, and can also be combined with some other intelligent optimization approaches.

Key words: experiment designs; uniform designs table; integer-coded genetic algorithms

1 引言

试验选优是进行复杂大系统优化的一种常用方法,而传统的试验设计常以随机方法或定态形式形成信息样本空间,其表达的精度完全依赖于随机数的均匀性和独立性.目前较为广泛应用的试验设计方法主要有正交设计(orthogonal designs)和均匀设计(uniform designs, UD),它们都是部分试验法[1~4].其中均匀设计自王元、方开泰 1981 年提出以来,无论在理论上还是在实践中都得到了充分的研究和发展[5~13],并在我国的航空航天、电子、医药、化工、纺织、冶金等领域取得了丰硕的成果和巨大的社会经济效益,得到了国内外同行的高度重视.均匀设计方法基于数论中的一致分布理论,借鉴了"近似分析中的数论方法"这一领域的研究成果,将数论和多元统计相结合,是属于蒙特卡罗方法的范畴.虽然与正交设计一样同属部分试验方法,但均匀设计更着重在实验范围内考虑试验点散布均匀以通过最少的试验来获得最多的信息,因而试验次数比正交设计法明显减少,这使得均匀设计特别适合于多因素多水平的试验和系统模型完全未知的

收稿日期:2004-12-13
资助项目:国家自然科学基金(70471090,50579009),国家"十五"科技攻关项目(2004BA608B-02-02)
作者简介:张礼兵(1972-),男,安徽肥东人,博士生,讲师,主要从事灌溉排水系统工程研究;程吉林(1963-),男,江苏常熟人,博士生导师,教授,主要从事水资源系统、农业系统工程研究.

情况,它已成为统计试验设计的重要方法之一.

一般文献所给的均匀设计表数量不多,且各表的水平数和因素数都较小,试验者大多只能在有限的均匀设计表及其使用表中查用,不仅无法得到任意试验点数 n、因素数 s 和水平数 q 的均匀设计表 $U_n(q^s)$ $(q\le n)$,而且使用起来极为不便,更不利于试验设计的程序化运行,这极大制约了均匀设计这一优良设计方法在更广阔应用领域的拓展.鉴于此,有必要根据均匀设计原理和方法进行均匀设计表的自动构造.目前常用的构造方法有方幂生成向量法(prime power generating-vectors method, PP-GVM)、正交设计扩展法(extended orthogonal designs, EOD)、拉丁方法(Latin hypercube, LH)和门限接受法(threshold accepting, TA)等[3~5],这些方法各有优点及不足.构造均匀设计过程实质上是一个求解某均匀度指标极值的优化问题,因此可以考虑将寻优能力很强的基于整数编码遗传算法(integer-coded genetic algorithm, IGA)与生成向量法(GVM)相结合进行均匀设计表的构造(即 IGA-GVM),这就是本文的基本思想.

2 构造均匀设计表的生成向量法

设一个要进行 n 次试验的实验,含 s 个因素,它们各自取 $q(q\le n)$ 个水平,则均匀设计表为 $U_n(q^s)$ $(q\le n)$,为方便,这里取等水平 $q=n$.该设计方案 $U_n(n^s)$ 可用一个 n 行 s 列的矩阵表示,并称这一矩阵为均匀设计的方案阵.方案阵的每一行代表一次试验,每一列代表一个因素,各列是$\{1,2,\cdots,n\}$的一个置换(即 $1,2,\cdots,n$ 的重新排列),每个元素代表在每次试验中对应因素所取的水平值.从几何角度看,如果我们将每个因素用一个坐标轴表示,因素的水平值变为对应的坐标值,则一个 s 因素 n 水平的均匀设计方案又可用散布在 s 维欧氏空间中$[1,n]^s$ 立方体内的 n 个点表示.均匀设计方法的目的就是在试验区域内生成均匀散布的信息样本,使试验者能有效地实现对搜索空间进行信息提取,从而得到 s 维空间(c^s)上的具有统计意义的最大概率点.过去的 20 多年中,生成向量法一直是构造均匀设计表的常用方法之一[3~9],其构造过程简述如下.

由数论的知识可知,对于给定的正整数 n,小于 n 且与 n 互素的自然数共有 $m=\phi(n)$ 个,这里 $\phi(n)$ 是著名的 Euler 函数,即任一正整数 n 存在惟一的素数分解 $n=p_1^{r_1}p_2^{r_2}\cdots p_t^{r_t}$,则 $\phi(n)$ 由下式确定:

$$\phi(n)=n\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\cdots\left(1-\frac{1}{p_t}\right). \tag{1}$$

令 $H=\{h_1,h_2,\cdots,h_m\}$为一个包含 m 个元素的正整数集合,其中任一元素 $h_j<n$,且 h_j 和 n 的最大公约数为 1.同时令 $u_{ij}=ih_j\,(\mathrm{mod}\ n)$,这里$(\mathrm{mod}\ n)$是同余运算,则 $U=(u_{ij})$是一个大小为 $n\times m$ 的均匀矩阵 U-矩阵.给定 $s<m$,则 U-矩阵取任意 s 列组成的矩阵仍为 U-矩阵,共有$\binom{m}{s}$个大小为 $n\times s$ 这样的子阵,各子阵组成的向量 $h_j=(h_{j_{v1}},h_{j_{v2}},\cdots,h_{j_{vs}})$,$v_1,v_2,\cdots,v_s\in\{1,2,\cdots,m\}$称为该均匀设计的生成向量.

由于对给定的因素数 n 和水平数 s,可产生$\binom{m}{s}$种均匀设计方案的 U-矩阵,而方案的选择不同,试验的效果也大不一样,因此有必要给出评价各设计方案好坏的指标,即均匀性度量指标,也称均匀度函数.均匀度越小表示方案的均匀性越好,常用的均匀度函数有:偏差 D,L_2-偏差 D_2,修正 L_2-偏差 MD_2,中心化 L_2-偏差 CD_2,对称 L_2-偏差 SD_2 和可卷 L_2-偏差 WD_2 等[3,4,10~12],文献[4]对这 6 个偏差值进行了分析比较,认为 MD_2,CD_2,SD_2 和 WD_2 较为合理,不失一般性,取 CD_2 为度量指标,将 u_{ij} 变换到 $x_{ij}=(2u_{ij}-1)/2n$,则 CD_2 可由下面简化式获得:

$$CD_2(h)=\Bigg[\left(\frac{13}{12}\right)^s-\frac{2^{1-s}}{n}\sum_{k=1}^{n}\prod_{i=1}^{s}\left(2+\left|x_{ki}-\frac{1}{2}\right|-\left|x_{ki}-\frac{1}{2}\right|^2\right)+\frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}\left(1+\frac{1}{2}\left|x_{ki}-\frac{1}{2}\right|+\frac{1}{2}\left|x_{ji}-\frac{1}{2}\right|-\frac{1}{2}\left|x_{ki}-x_{ji}\right|\right)\Bigg]^{\frac{1}{2}}. \tag{2}$$

由于一个生成向量 h_j 只能唯一地确定一个 n 水平的均匀设计方案 $U_n^*(h)$,故构造均匀设计实质就是

以 h_j 作为自变量,在共有 $\begin{pmatrix} m \\ s \end{pmatrix}$ 个组合中寻找均匀度函数最好的子阵 $U_n^*(n^s)$,就得到最优均匀设计的一个近似解,因此它是一个组合优化问题. 当 m、s 较大,如 $n = 31$, $s = 14$ 时, $m = \phi(31) = 30$,则比较 $\begin{pmatrix} 30 \\ 14 \end{pmatrix} \approx 1 \times 10^9$ (文献[4]为 3×10^{21} 有误) 个设计的工作量还是较大的. 为此文献[4]通过寻找 U-矩阵生成元 a 的方法以减少需考察的生成向量数,即方幂生成向量法(PP-GVM). 但 PP-GVM 的向量选取过程较复杂,同时在生成元 a 对 n 的次数判断上耗费大量计算时间,甚至可能由于幂乘数值过大造成内存溢出,因此有较大的局限性.

3 基于 IGA 的均匀设计表构造

如前所述,均匀设计过程实际是一个等价于以某种均匀度函数为目标的组合优化问题,因此可以引入寻优能力很强的基于整数编码遗传算法(IGA)来求解. 关于 IGA 算法的优良性已有许多研究[14,15],这里不再赘述. 下面简述 IGA-GVM 构造均匀设计表步骤.

目标函数即求均匀设计方案中心偏差 $CD_2(h)$ 最小

$$\min CD_2(h), \quad h \in H_l. \tag{3}$$

步骤 1 优化个体编码. 由前知,生成向量集合 $H_l = \{h_1, h_2, \cdots, h_m\}$ 由 n 唯一确定,令正整数集合 $J = \{1, 2, \cdots, m\}$, J 的元素与 H_l 中元素一一对应,则每个生成向量 $h_j = (h_{j_{v1}}, h_{j_{v2}}, \cdots, h_{j_{vs}})$, $v_1, v_2, \cdots, v_s \in \{1, 2, \cdots, m\}$ 对应于 J 的一个子集构成的正整数数列 $p = (j_k | j_k \in J, k = 1, 2, \cdots, s)$, p 即为遗传算法的一个个体,则 j_k 为该个体的第 k 个基因段,通过这样的编码,使得遗传算法对生成向量集合 $H_l = \{h_1, h_2, \cdots, h_m\}$ 的各种组合,转换为对个体 x 的基因段进行各种遗传操作过程.

例如 $n = 21$、$s = 5$ 时,由欧拉函数得 $m = \phi(n) = 12$,则 $H_{21} = \{1, 2, 4, 5, 8, 10, 11, 13, 16, 17, 19, 20\}$,对生成向量 $h' = \{2, 5, 16, 17, 19\}$ 这里编码为 $p' = (2, 4, 9, 10, 11)$.

步骤 2 父代群体初始化. 设群体规模为 M,利用均匀随机数易产生 s 个不大于 m 且互异的正整数并将其按升序排列,即得到一个个体,这样的个体共产生 M 个以作为父代群体.

步骤 3 父代群体适应度评价. 将个体解码为相应的生成向量 h_j 并利用 $u_{ij} = ih_j \pmod n$ 和 $x_{ij} = (2u_{ij} - 1)/2n$ 得到该个体的分布矩阵 x_{ij},将其代入式(2)即得该个体的均匀度函数 $CD_2(\mathrm{h})$,个体的 $CD_2(\mathrm{h})$ 值越小,表示该个体的适应值越高,基于此将遗传算法的适应度函数定义为

$$F = \frac{1}{[CD_2(h)]^2 + 0.001} \tag{4}$$

式中 0.001 是为了避免 $CD_2(\mathrm{h})$ 值为零时出错而设的经验值.

步骤 4 选择操作. 对父代群体按适应度进行依概率 P_s 选择和依概率 P_{ex} 精英保留[15],共获得 M_s 个子代个体.

步骤 5 杂交操作. 即在 M 个父代个体中依杂交概率 P_c 任取两个体进行单点(或多点)交叉操作,杂交后新个体内的基因应保持互异并按基因值以升序排列.

步骤 6 变异操作. 对 M 个父代个体进行依概率 P_m 变异操作,同样应保持新个体的基因互异并对基因值按升序排列.

步骤 7 演化迭代. 由步骤 4 至步骤 6 得到 $M_s + 2M$ 个子代个体,按适应度降序排列,并取前 M 个个体作为新的父代群体,算法进入步骤 3,进入下一轮次的演化,直到指定的演化代数 N_p.

以上构成了 IGA-GVM 求解均匀设计表优化的全过程.

4 试验结果

现试用 IGA-GVM 进行均匀设计表的优化构造,这里取群体规模 $M = 100$,进化代数 $N_p = 10$,精英个体

保留概率 $P_{ex}=0.1$,选择概率 $P_s=0.8$,杂交概率 $P_c=1$,即采用全体杂交,变异概率 $P_m=0.05$,经试验获得了任意 n、s 组合的最优生成向量,表 1 给出了部分试验结果,为方便比较将文献[4]的设计方案也一同列入表内.

表 1 不同方法构造 $U_n(n^s)$ 的中心化偏差 CD_2 值及生成向量

n	$s=4$		$s=5$	
	PP-GVM	IGA-GVM	PP-GVM	IGA-GVM
21	0.0917 (1 5 8 19)	0.0912 (2 5 8 20)	0.1310 (1 4 10 13 16)	0.1309 (2 5 8 17 20)
22	0.0808 (1 5 7 13)	0.0761 (6 8 10 22)	0.1157 (1 3 5 7 13)	0.1052 (2 13 14 15 17)
23	0.0745(1 7 18 20)	0.0744 (3 5 16 22)	0.1088 (1 4 7 17 18)	0.1086 (2 7 14 15 22)
24	0.0679 (1 11 17 19)	0.0631 (1 13 18 23)	0.1422 (1 5 7 13 23)	0.1422 (1 5 11 17 19)
25	0.0672 (1 6 11 16)	0.0671 (3 13 18 23)	0.0946 (1 6 11 16 21)	0.0946 (3 8 13 18 23)
26	0.0700 (1 5 11 17)	0.0699 * (1 11 17 19)	0.1049 (1 3 5 11 17)	0.1048 * (3 5 7 17 25)
27	0.0670 (1 8 20 22)	0.0670 (2 13 17 19)	0.1013 (1 8 20 22 23)	0.1013 (4 10 11 14 26)
28	0.0703 (1 9 11 15)	0.0703 (13 17 19 27)	0.0993 (1 9 11 15 23)	0.0993 (1 11 15 23 25)
29	0.0606 (1 8 17 18)	0.0605 * (8 9 14 17)	0.0901 (1 7 16 20 24)	0.0900 * (4 5 13 22 28)
30	0.0559 (1 17 19 23)	0.0531 (15 18 19 21)	0.1301 (1 7 11 13 29)	0.1188 (1 6 13 14 27)

注: *栏中 IGA 取 $M=300$、$N_p=30$ 以提高优化效果.

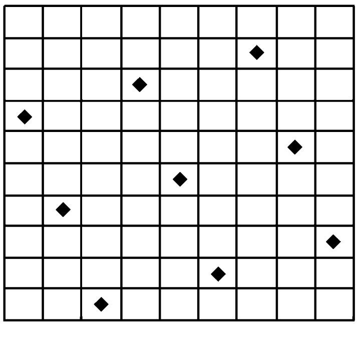
由表 1 的试验结果可知:

1) IGA-GVM 能够以较小的群体规模 M 和较少的进化代数 N_p 得到不劣于 PP-GVM 的均匀设计表,计算快速、结果稳定,并且通过适当增大 N_p、M 可进一步提高计算精度,当然计算时间相应有所增加.

2) PP-GVM 的向量选取过程较复杂,在生成元对 n 的次数判断上耗费大量计算时间,甚至由于幂乘数值过大造成内存溢出,而 IGA-GVM 无须进行生成元判断,且理论上可以生成一切 $s\leq n$ 的均匀设计表,当 n 较大(如 $n>50$)时该法的优越性体现更明显.

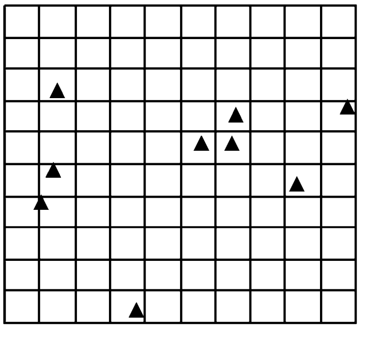
3) 因为具有等价的均匀设计表往往不唯一,而过去文献中一般仅给出一个,且多含有不可实施的试验点、因子与水平组合,IGA-GVM 可以同时获得若干个等价试验方案(这里仅示例一个),给试验者以更大的备择空间.

众所周知,遗传算法的搜索效率与其初始样本点的分布有很大的关系,而一般遗传算法的初始样本均匀布点依赖于随机数的均匀性.为了比较这里给出了 $s=2$, n 分别等于 9、29 时,IGA-GVM 与均匀随机数的布点结果,见图 1～图 4.



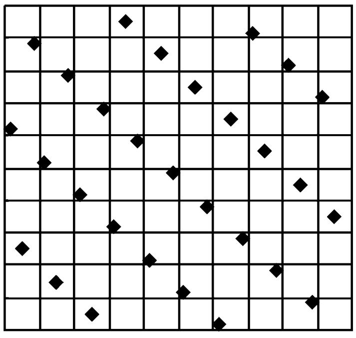
$(n=9, s=2, CD_2=0.0614)$

图 1 IGA-GVM 布点结果



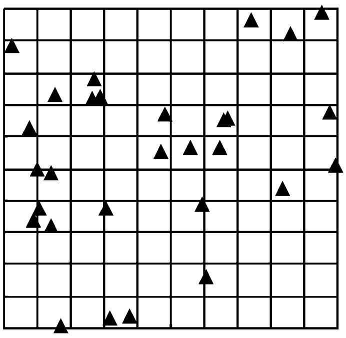
$(n=9, s=2, CD_2=0.0671)$

图 2 均匀随机数布点结果



($n = 29$, $s = 2$, $CD_2 = 0.0205$)

图 3　IGA-GVM 布点结果



($n = 29$, $s = 2$, $CD_2 = 0.0440$)

图 4　均匀随机数布点结果

由图 1～图 4 知: ①IGA-GVM 分布点群充分体现了均匀分散性,且不会产生重复试验点,而均匀随机数的分布则相对杂乱、缺乏均匀性. ②IGA-GVM 的每个样本点集不仅有一维和二维投影空间上的均匀性,还有任意 s 维投影空间上的均匀性. 与均匀随机数相比,在固定样本大小的前提下 IGA-GVM 使样本的经验分布能很好地整体逼近样本空间总体的分布. 试验还说明随着 n、s 增大,IGA-GVM 均匀性明显好于均匀随机数的分布.

5　结语

本文提出的基于整数编码遗传算法的均匀设计构造法,根据均匀设计原理和方法可以自动生成任意因素水平的均匀设计表,这为均匀设计方法拓展更大的应用空间提供可能. 例如,可以将均匀设计方法与现代智能优化算法如遗传算法、人工神经网络、模拟退火算法等相结合以增强优化性能.

参考文献:

[1] 正交试验法编写组. 正交试验设计法[M]. 上海:上海科学技术出版社,1978,1 - 17.
Group of compiling the orthogonal experiment. The Orthogonal Experiment Method[M]. Shanghai: Shanghai Science & Tech. Press, 1978,1 - 17.

[2] 程吉林. 大系统试验选优理论和应用[M]. 上海:上海科技出版社,2002,25 - 39.
Chen J L. Theory and Application of Experimental Optimization for Large-scale System[M]. Shanghai: Shanghai Science & Tech. Press,2002, 25 - 39.

[3] 方开泰. 均匀设计与均匀设计表[M]. 北京: 科学出版社, 1994, 1 - 78.
Fang K T. Uniform Design and the Uniform Design Table[M]. Beijing: Science Press, 1994, 1 - 78.

[4] 方开泰,马长兴. 正交与均匀试验设计[M]. 北京: 科学出版社,2001, 83 - 102.
Fang K T, Ma C X. Orthogonal and Uniform Design for Experiment[M]. Beijing: Science Press,2001, 83 - 102.

[5] Hua L K,Wang Y. Applications of Number Theory to Numerical Analysis[M]. Springer and Science Press, Berlin and Beijing, 1992,15 - 77.

[6] 张润楚,王兆军. 均匀设计抽样及其优良性质[J]. 应用概率统计, 1996,(12): 337 - 347.
Zhang R C, Wang Z J. Uniform design sampling and its good properties[J]. Chinese Journal of Applied Probability & Statistics, 1996,(12): 337 - 347.

[7] 刘洪谦,麻德贤,陈慧灵,等. 增强并行均匀序贯寻优方法及其全局寻优性能研究[J]. 北京化工大学学报, 1999,26(4):11 - 14.
Liu H Q, Ma D X, Chen H N, Li Z X. An enhanced parallel uniform sequential optimization method and studies on its ability for global optimization[J]. Journal of Beijing University of Chemical Technology, 1999,26(4):11 - 14.

[8] Kai-tai Fang, Xuan Lu, Yu Tang. Constructions of uniform designs by using resolvable packing and coverings[J]. Discrete Mathematics,2004,(274): 25 - 40.

(下转第 82 页)

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341 - 356.

[2] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data [M]. London: Kluwer Academic Publishers, 1991.

[3] Kryszkiewicz M. Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 112: 39 - 49.

[4] Kryszkiewicz M. Rules in incomplete information systems [J]. Information Sciences, 1999, 113: 271 - 292.

[5] Leung Y, Li D. Maximal consistent block technique for rule acquisition in incomplete information systems [J]. Information Sciences, 2003, 153: 85 - 106.

[6] 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展, 2002, 39(10), 1238 - 1243.
Wang Guo-yin. Extension of rough set under incomplete information systems [J]. Journal of Compute Research and Development, 2002, 39(10), 1238 - 1243.

[7] 黄兵, 周献中. 不完备信息系统中基于联系度的粗集模型拓展[J]. 系统工程理论与实践, 2004, 24(1), 88 - 92.
Huang Bing, Zhou Xian-zhong. Extension of rough set model based on connection degree under incomplete information systems[J]. Journal of Systems Engineering - Theory & Practice, 2004, 24(1), 88 - 92.

[8] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.
Zhang Wen-xiu, Wu Wei-zhi, Liang Ji-ye, et al. Rough Set Theory and Approach [M]. Beijing: Science Press, 2001.

[9] 刘清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.
Liu Qing. Rough Sets and Rough Reasoning [M]. Beijing: Science Press, 2001.

(上接第 61 页)

[9] Kai-tai Fang, Hong Qin. A note on construction of nearly uniform designs with large number of runs[J]. Statistics & Probability Letters, 2003, (61): 215 - 224.

[10] 徐洪泉, 沈世镒. 基于计算机试验的均匀设计[J]. 高校应用数学学报, 1998, 13 (2) A: 167 - 174.
Xu H Q, Shen S Y. The uniform design based on computer experiments[J]. Appl Math-JCU, 1998, 13(2)A: 167 - 174.

[11] 孙先仿, 范跃祖, 宁文如. U* 均匀设计的均匀性研究[J]. 应用概率统计, 2001, 17(4): 341 - 345.
Sun X F, Fan Y Z, Ning W R. On the uniformity of U* uniform designs[J]. Chinese Journal of Applied Probability & Statistics, 2001, 17(4): 341 - 345.

[12] 马长兴. 均匀性的一个新度量准则-对称偏差[J]. 南开大学学报(自然科学), 1997, 30(1): 31 - 37.
Ma C X. A new criterion of uniformity-symmetrical discrepancy[J]. Acta Scientiarum Naturalium Universitatis Nankaiensis, 1997, 30(1): 31 - 37.

[13] Hickernell F J. A generalized discrepancy and quadrature error bound[J]. Math Comp, 1998(67): 299 - 322.

[14] 陈国良, 王煦法, 等. 遗传算法及其应用[M]. 北京: 人民邮电出版社, 1996, 28 - 111.
Chen G L, Wang X F. Genetic Algorithm and Its Application[M]. Beijing: People's Post & Telecom Press, 1996, 28 - 111.

[15] 金菊良, 丁晶. 水资源系统工程[M]. 成都: 四川科学技术出版社, 2002, 68 - 71.
Jin J L, Ding J. Water Resource Systems Engineering [M]. Chendu: Sichuan Science & Tech Press, 2002, 68 - 71.