

具有相关特性数据的随机化

袁永生¹, 董祖引¹, 张永良²

(1. 河海大学数理系, 江苏 南京 210098; 2. 南京金融高等专科学校, 江苏 南京 211800)

摘要 区别于通常确保随机性的随机化方法, 针对具有相关特性数组的随机化, 提出了一种能有效保留原有相关特性的随机化方法. 这一方法基于某个大样本按特定的方式寻找所需的随机数序列, 不需传统随机化方法中已知分布及种子数的要求. 在这一随机化方法下, 给出了一种简洁而有效的算法. 该算法中引进一个参照数组, 记录统计范围内相同的元素个数, 进而得出统计规律. 此外还提供了部分程序框图.

关键词 随机化; 相关性; 熵; 算法; 频数

中图分类号: TN911.72; O212 文献标识码: A 文章编号: 1000-1980(2001)05-0046-03

在实际工程应用中, 常需寻找若干同型数组在每个位置上元素值出现的统计规律(如数字化图像各像素上熵的计算)^[1], 这些数组的元素都在某个特定的有限集内取值, 数组的元素之间(尤其相邻位置间)具有一定的相关性. 由于工程问题的特殊性, 经常不能得到多个同类样本数组, 但能得到一个行数和列数均较大的样本数组(如遥感遥测数字化图像)^[2]. 此时若采用传统的随机化方法^[3], 即对数组每个元各由一个种子数按一定分布产生一个随机数序列, 产生的数序列确实具有较好的随机性, 但往往破坏了数组相邻位置间的固有统计相关性和数组元素的实际统计特性. 即用通常随机化方法产生的数组不一定能在样本集中, 也就是说实际问题很可能不会有这样的样本. 本文针对这样一类背景问题, 从工程应用角度出发给出一种符合工程实际的有效解决这一问题的随机化方法.

现设有 $m \times n$ 矩阵 $A = (\cdot)_{m \times n}$, 另有一数组 $B = (b_{ij})_{M \times N}$, B 中每个元素都在有限集 $\Omega = \{x_1, \dots, x_r\}$ 中取值, 相邻元素间有一定的相关性, 且 $M \geq m, N \geq n$. 现试图由数组 B 随机化出一列与 A 同型的样本数组 $A_1, \dots, A_s$ (s 为样本容量), 进而由这些数组找出 A 中各元取 $x_1, \dots, x_r$ 的统计规律.

1 随机化基本思想

若 $M \geq m, N \geq n$ (等号不同时成立), 如图 1 所示, 设想将形状矩阵 A (只有大小, 没有具体元素值) 放在 B 中, 此时 A 所对应的 B 中左上角 $m \times n$ 个元素为 A 的第一个随机数组. 再将 A 向右移动一列, 此时 A 所对应的 B 中 $m \times n$ 个元素为 A 的第二个随机数组. 依此类推, 逐步向右移动, 直至 A 的第 n 列重合于 B 的第 N 列时为止, 移动过程中每次 A 所对应的 $m \times n$ 个 B 中的元素均作为 A 的随机数组. 这样得到 A 的 $N - n + 1$ 个随机数组. 进一步, 先将 A 在 B 中下移一行 (即 B 的第二行作为 A 的第一行所处位置), 再同样让 A 的第一列自 B 的最左侧向右移动, 每次移动一列, 直至 A 的第 n 列重合于 B 的第 N 列时为止, 每次 A 所对应的 B 中 $m \times n$ 个元素均作为 A 的新随机数组, 从而得到 A 的另 $N - n + 1$ 个随机数组. 依此类推, 每次下移一行, 直至 A 的最后一行重合于 B 的最后一行时为止, 每次下移一行后的操作与前述相同. 这样共可得 A 的 $(M - m + 1) \times (N - n + 1)$ 个随机数组组成的随机数组序列.

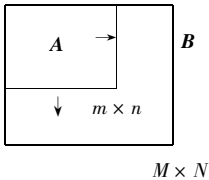


图 1 数组取法示意图
Fig. 1 Delineation of matrix selection

实质上, 按此随机化方法不需已知分布, 也不需设定种子数, 数组 A 的第一个元 a_{11} 的随机数序列是 B

收稿日期 2000-06-13
基金项目 江苏省自然科学基金资助项目(BK99106); 水利部科技创新基金资助项目(SCX2000-11)
作者简介 袁永生(1964—)男, 江苏海安人, 副教授, 应用数学专业, 主要从事随机数学应用研究.

中这样的子块

$$\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1,N-n+1} \\ b_{21} & b_{22} & \cdots & b_{2,N-n+1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{M-m+1,1} & b_{M-m+1,2} & \cdots & b_{M-m+1,N-n+1} \end{pmatrix}$$

依此类推, A 中第 i 行第 j 列元的随机数序列是 B 中这样的子块

$$\overline{B_{ij}} = \begin{pmatrix} b_{ij} & b_{i,j+1} & \cdots & b_{i,j+N-n} \\ b_{i+1,j} & b_{i+1,j+1} & \cdots & b_{i+1,j+N-n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{i+M-m,j} & b_{i+M-m,j+1} & \cdots & b_{i+M-m,j+N-n} \end{pmatrix} \quad i=1,\dots,m; j=1,\dots,n$$

这里若以 $\overline{B_{ij}}(k,l)$ 表示 $\overline{B_{ij}}$ 中的第 k 行 l 列的元素, 则有关系式

$$\overline{B_{ij}}(k,l) = B(i+k-1,j+l-1) \quad i=1,\dots,m; j=1,\dots,n$$

基于这种随机化思想, 如果数组 B 原有相邻元素之间的相关性较强, A 中相邻各元之间的相关性将得到较好的保留. 如果工程实际问题是采用较短时间间隔连续采样, 可以按这一随机化过程, 用一次采大样来代替. 这一随机化方法特别适用于形状矩阵 A (随机变量矩阵) 中各元之间相关性较强的背景问题, 至于相关性判定, 可以依据统计学的方法, 或者依据 A 中相邻各元上出现何值是否有关联来作出判断.

2 求统计规律的算法技巧

由于本随机化方法的特殊性, 因而在计算 A 中各元统计规律时算法实现上有较强技巧. 依据频率的定义, 找 A 中各元统计规律实质上是要找 A 中各元分别取 x_k ($k=1,\dots,r$) 的次数 (频数), 用这一次数除以该元的全部取值数 (重复的取值要重复计数) 即为该元取 x_k ($k=1,\dots,r$) 的频率. 本段以二维数组为例介绍一种算法.

设 B 中各元素均在有限集 $\Omega = \{x_1, \dots, x_r\}$ 中, 根据以上随机化方法, 现求随机化后 $A = (a_{ij})_{m \times n}$ 中各元分别取 $x_1, \dots, x_r$ 的频率. 这里 A 中各元 a_{ij} ($i=1,\dots,m; j=1,\dots,n$) 的取值情况在前述 $\overline{B_{ij}}$ 中. 由于数组 $\overline{B_{ij}}$ 中不同元素不一定是 $x_1, \dots, x_r$ 中全部, 即不需对每个 $x_k \in \Omega$ ($k=1,\dots,r$) 统计 $\overline{B_{ij}}$ 中 x_k 的个数, 也不需对每个 x_k 都检测一遍全部 $\overline{B_{ij}}$ 中元素. 本文引进一个参照数组而巧妙地解决了完成所需计算与提高效率的问题.

引进二维参照数组 KAB [以 $KAB(k,l)$ 表示 KAB 中第 k 行 l 列的元素], 其行数、列数分别与上述 $\overline{B_{ij}}$ 的行数、列数完全相同, 且 KAB 中的各元素都置为 0, 将 $\overline{B_{ij}}$ 与 KAB 对照使用.

步骤 1 对 KAB 按先行后列考察其元素是否为 0, $KAB(1,1)=0$, 则将 $\overline{B_{ij}}(1,1)=B(i,j)$ 作为基数逐个与 $\overline{B_{ij}}$ 中其后面及下面各行上元素比较, 直至 $\overline{B_{ij}}$ 的最后一行的最后一个元素. 比较过程中, 若后面某位置上元素值与基数相同, 则 KAB 中相应于该位置上的值置为 -1, 同时将 KAB 中相应于此时基数位置上的值加 1 [例如, $\overline{B_{ij}}(1,1)=\overline{B_{ij}}(k1,l1)$, 则 $KAB(k1,l1)=-1$, 且 $KAB(1,1)=KAB(1,1)+1$]. 全部比较完 $\overline{B_{ij}}$ 中元素后再作一次 $KAB(1,1)=KAB(1,1)+1$, 此修正后 $KAB(1,1)$ 的值即为值 $\overline{B_{ij}}(1,1)=B(i,j)$ 在 $\overline{B_{ij}}$ 中出现的频数.

步骤 2 考察 $KAB(1,2)$ 是否为 0, 若不为 0, 则检查 $KAB(1,3)$; 若为 0, 则将 $\overline{B_{ij}}(1,2)=B(i,j+1)$ 作为基数, 与 $\overline{B_{ij}}$ 中 $\overline{B_{ij}}(1,2)$ 后面及下面各元素作比较 (直至 $\overline{B_{ij}}$ 的最后一个元素). 比较过程中的操作及比较完成后的操作与步骤 1 相同.

依此类推, 一直考察到 KAB 的最后一个元素, 每考察一个 KAB 的值都需进行与上述步骤 2 中相同的处理.

最后, 检查 KAB , 若 KAB 某位置上值等于 0, 则 $\overline{B_{ij}}$ 中对应位置上的值即为 a_{ij} 的可能取值, 而 a_{ij} 取该值的频数即为 KAB 中该位置上的值. 即若 $KAB(k1,l1)>0$ 则

$$\overline{B_{ij}}(k1,l1) = B(i+k1-1,j+l1-1)$$

的值为 a_{ij} 的可能取值之一, 而 a_{ij} 取 $B(i+k1-1,j+l1-1)$ 的频数是 $KAB(k1,l1)$ 的值, 频率是 $KAB(k1,$

$l1)(M-m+1)(N-n+1)]$. 可以定义一个 2 行 r 列的矩阵用于存放 a_{ij} 的全部取值及取每一值的频率或直接由熵的计算公式求出所需要的熵.

3 算法结构框图^[4]

这里只给出求 $a_{ij}(i=1, \dots, n; j=1, \dots, m)$ 的取值及取各值频数的程序框图(图 2).

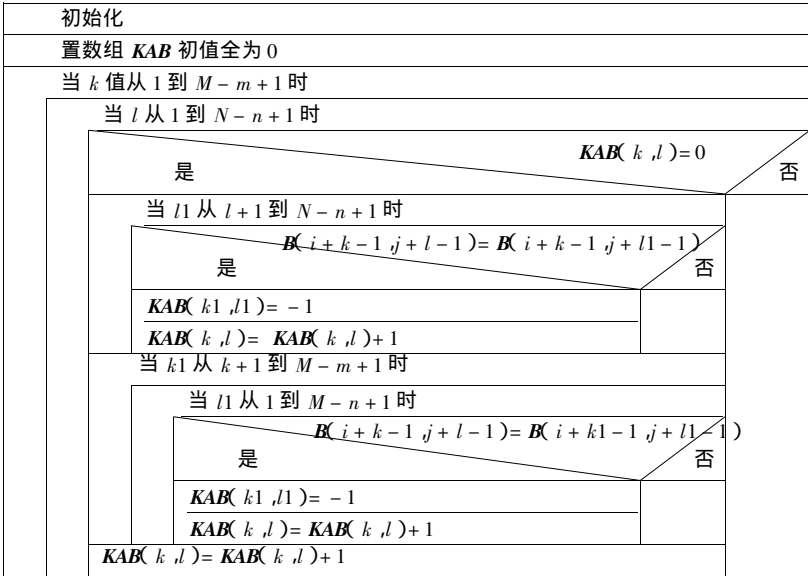


图 2 部分程序框图
Fig.2 Part of program block

参考文献 :

[1] 袁永生, 孟庆生. 数字信号压缩与滤波先后次序的分析[J]. 通信学报, 1998, 19(12): 93~96.
[2] 雅罗斯拉夫斯基著. 数字图像处理[M]. 施有秋, 袁述译. 北京: 电子工业出版社, 1990. 69~70.
[3] 刘宝碇, 赵瑞清. 随机规划与模糊规划[M]. 北京: 清华大学出版社, 1998. 38~39.
[4] 谭浩强, 田淑清. FORTRAN77 结构化程序设计[M]. 北京: 高等教育出版社, 1985. 81~84.

Randomization of Data of Related Property

YUAN Yong-sheng¹, DONG Zu-ying¹, ZHANG Yong-liang²

(1. Dept. of Math. and Phys., Hohai Univ., Nanjing 210098, China;

2. Dept. of Basic Science, Nanjing Higher Finance College, Nanjing 211800, China)

Abstract :For the data of related property ,a new randomization method ,which can keep the original relativity ,is proposed based on a large sample from which the random number sequence is searched. The method is different from the conventional randomization method in that the distribution and seeded number need not be known. Finally ,a valid algorithm is deduced ,in which statistic regularity is obtained based on a reference array and the number of the same elements and the main program block is also presented.

Key words :randomization ;related property ;entropy algorithm ;frequency