中国农业大学学报 1999,4(6):1~5
Journal of China Agricultural University

我国各地农机化发展水平的一种有序样本分类法①

白人朴② 田志宏
(中国农业大学农村发展研究所)

摘 要 基于动态规划思想提出了一种有序样本的分类方法,能够识别子样(分段)间的分点,按问题的需要将整个样本划分成不同子样,使整个分类达到某种目标下的最优。应用于农业机械化发展水平的分类研究,取得了与实际情况一致的效果。

关键词 有序样本;分类方法;农业机械化水平
分类号 F 224.31

A Clustering Method for Ordering Samples and Its Utilization in Study of Farm Mechanization Level

Bai Renpu Tian Zhihong
(Rural Development Institute, CAU)

Abstract A clustering method for the ordering samples is discussed based on the dynamic programming. It can be used to identify the breaking points and divide the sample into some sub-samples at where there are significant differences between the neighbor sub-samples, and obtain the optimum clustering results under a certain criterion. This method has been used in the evaluation of farm mechanization levels of different provinces in China. The results proved that this method is practical.

Key words ordering samples; clustering method; farm mechanization level

正确认识我国各地区农业机械化所处的发展阶段,是制定和执行正确的农业机械化发展方针、政策和规划的根本依据。对农业机械化的宏观指导,应在阶段分析的基础上,对各地农业机械化发展水平进行具体分析,因地制宜、分类指导,促进我国农业机械化现代化事业的持续、稳定发展。

在诸如农业机械化发展水平等社会经济系统分析的项目中,有些变量的观测值是按照一定顺序排列的,称之为有序样本,如全国各省市区农业机械化的综合水平即表征其所处的不同发展阶段。对于有序样本的分类,关键是找出分点(也称为变结构点)[1],使分类在某一目标下达到最优,即将整个样本划分成几个子样(分段),每个子样与其他子样有一定差别[2,3],各子样内部各样本点之间的差异最小,而各子样之间的差异最大,这样所形成的分类是最显著的。

在以往农业机械化水平的分类研究中,研究者们较多地使用统计聚类或模糊聚类方法[4]。与之相比,有序样本分类方法较充分地利用了顺序信息,更明晰地表达了分类点。

收稿日期:1999-03-15
①中国博士后科学基金项目
②白人朴,北京清华东路17号 中国农业大学(东校区)48信箱,100083

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1 有序样本的最优分类方法

1.1 基本定义

1)样本、分点与子样(分段)

设有 q 个变量(指标),有 N 个样本个体,第 i 个个体的第 j 个指标值为 x_{ij},则原始样本可表示为

$$\boldsymbol{X}=\begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{q1} & x_{q2} & \cdots & x_{qN} \end{Bmatrix}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N) \tag{1}$$

一般地,将 q 维变量的 N 个样本划分为 m 个子样(m 个分段或 m 类),须确定 $m-1$ 个分点,分段结果为

$$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{k_1}\},\{\boldsymbol{X}_{k_1+1},\boldsymbol{X}_{k_1+2},\cdots,\boldsymbol{X}_{k_2}\},\cdots,\{\boldsymbol{X}_{k_{m-1}+1},\boldsymbol{X}_{k_{m-1}+2},\cdots,\boldsymbol{X}_N\}$$

分点集合由 $m-1$ 个元素组成,记为 $\Omega=\{\boldsymbol{X}_{k_1},\boldsymbol{X}_{k_2},\cdots,\boldsymbol{X}_{k_{m-1}}\}$。

2)样本数据的正规化与加权

首先用 2 种方法来做样本数据的正规化,然后求得可直接用于分类计算的加权正规化矩阵。

第 1 种方法。将各个变量的数据变换到(0,1),即将式(1)中的矩阵元素做如下变换

$$y_{ij}=\frac{x_{ij}-\min\limits_j x_{ij}}{\max\limits_j x_{ij}-\min\limits_j x_{ij}} \quad i=1,2,\cdots,q;\ j=1,2,\cdots,N \tag{2}$$

第 2 种方法。将各个变量的数据变换为服从 $N(0,1)$ 分布的序列

$$y_{ij}=\frac{x_{ij}-\sum\limits_{j=1}^{N}x_{ij}/N}{\left[\sum\limits_{j=1}^{N}\left(x_{ij}-\sum\limits_{j=1}^{N}x_{ij}/N\right)^2\Big/N\right]^{1/2}} \quad i=1,2,\cdots,q;\ j=1,2,\cdots,N \tag{3}$$

即得到样本数据的正规化矩阵 $\boldsymbol{Y}=(y_{ij}),i=1,2,\cdots,q;j=1,2,\cdots,N$。

由于各变量对于样本分类的重要程度不同,取权向量 $\boldsymbol{W}=[w_1,w_2,\cdots,w_q]^{\mathrm{T}}$,其中 w_i 表示第 i 个变量的重要程度。权向量满足归一化条件,即 $\sum\limits_{i=1}^{q}w_i=1$。

用权向量来处理由式(2)或(3)得到的样本数据的正规化矩阵 $\boldsymbol{Y}=(y_{ij})$,可得加权的正规化矩阵 $\boldsymbol{Z}=(z_{ij})$,其中,$z_{ij}=w_iy_{ij},i=1,2,\cdots,q;j=1,2,\cdots,N$。

3)分类的目标函数

各分类之间的差异是用变量样本值的差异来衡量的。一般地,总体样本的任一子样$\{\boldsymbol{X}_k,\boldsymbol{X}_{k+1},\cdots,\boldsymbol{X}_l\}$的变差可定义为

$$D_{kl}=\sum_{j=k}^{l}\sum_{i=1}^{q}\left[z_{ij}-\sum_{j=k}^{l}z_{ij}/(l-k+1)\right]^2 \quad 1\leqslant k<l\leqslant N \tag{4}$$

考虑整个样本的前 p 组数据($m\leqslant p\leqslant N$),将其划分为 m 个子样,$\Omega=\{\boldsymbol{X}_{k_1},\boldsymbol{X}_{k_2},\cdots,\boldsymbol{X}_{k_{m-1}}\}$,分点的位置满足 $1\leqslant k_1<k_2<\cdots<k_{m-1}<p$。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

令 $k_0=0,k_m=p$,各子样的变差可表示为 $D_{k_{i-1}+1,k_i},i=1,2,\cdots,m$,总变差

$$S_{p,m}(k_1,k_2,\cdots,k_{m-1})=\sum_{i=1}^{m}D_{k_{i-1}+1,k_i}$$

式中:S 的下标 p 表示参与分段的样本容量,m 表示划分的子样数;括号中的 $m-1$ 个元素是各子样的分点。则将整个样本划分为 m 个子样的变差为

$$S_{N,m}(k_1,k_2,\cdots,k_{m-1})=\sum_{i=1}^{m}D_{k_{i-1}+1,k_i} \tag{5}$$

将样本前 p 组数据划分为 2 个子样的变差为

$$S_{p,2}(k_1)=D_{1k_1}+D_{k_1+1,p}\quad p=2,3,\cdots,N;k_1=1,2,\cdots,p-1$$

其中 k_1 是 2 分类的分点。

4)最优分类

对于整个样本,通过选择分点集合 Ω 中 $m-1$ 个元素的位置,使式(5)所示的 m 个子样的总变差最小即为最优。记最优分类对应的分点集合 $\Omega^*=\{\boldsymbol{X}_{k_1^*},\boldsymbol{X}_{k_2^*},\cdots,\boldsymbol{X}_{k_{m-1}^*}\}$。

找到样本前 p 组数据的最优 2 分类的分点 k_1^p,对应

$$S_{p,2}(k_1^p)=\min_{1\leqslant k_1<p}S_{p,2}(k_1)=\min_{1\leqslant k_1<p}(D_{1k_1}+D_{k_1+1,p})\quad p=2,3,\cdots,N \tag{6}$$

显然,当 $p=N$ 时,分点 k_1^p 就是整个样本最优 2 分类的分点 k_1^*。

1.2　有序样本的分类方法

最优分类方法由以下 4 个主要步骤组成。

1)样本数据的正规化。用式(2)或(3)做样本数据变换,得正规化矩阵 $\boldsymbol{Z}$。

2)计算变差矩阵。由式(4)求得各子样的变差 $D_{kl},1\leqslant k<l\leqslant N$,得变差矩阵 $\boldsymbol{D}=(D_{kl})_{N\times N}$。由 $D_{kl}=D_{lk},D_{kk}=0$ 可知变差阵 $\boldsymbol{D}$ 是一对称矩阵,故只需计算上三角阵即可。

3)求最优 2 分类与最优 m 分类

先求整个样本的前 p 组数据($2\leqslant p\leqslant N$)的最优 2 分类。取 $p=2,3,\cdots,N$,由式(6)求出相应的总变差 $S_{p,2}(k_1),k_1=1,2,\cdots,p-1$,通过比较得最优 2 分点 k_1^p,对应最小的总变差 $S_{p,2}(k_1^p)=\min_{1\leqslant k_1<p}S_{p,2}(k_1)$。当 $p=N$ 时,即得到整个样本的最优 2 分类:$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{k_1^*}\}$,$\{\boldsymbol{X}_{k_1^*+1},\boldsymbol{X}_{k_1^*+2},\cdots,\boldsymbol{X}_N\}$。

获得最优 m 分类,需满足

$$S_{N,m}(k_1^*,k_2^*,\cdots,k_{m-1}^*)=S_{p,m}(k_1^p,k_2^p,\cdots,k_{m-1}^p)\big|_{p=N}=$$
$$\min_{m-1\leqslant k_{m-1}<N}S_{N,m}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}},k_{m-1}) \tag{7}$$

其中

$$S_{N,m}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}},k_{m-1})=S_{k_{m-1},m-1}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}})+D_{k_{m-1}+1,N}$$

在已求得最优 $m-1$ 分类的条件下,最优 m 分类问题就转化成最优 2 分类问题,即可通过比较样本前 k_{m-1} 组数据的最优 $m-1$ 分类求得最优 m 分类。

上述结果可用反证法得到。若有 $S_{N,m}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}},k_{m-1})$ 是最优 m 分类,则 $S_{k_{m-1},m-1}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}})$ 必是前 k_{m-1} 组数据的最优 $m-1$ 分类,$k_{m-1}=N-1,\cdots,m-1$,否则,若另有最优 $m-1$ 分类 $S_{k_{m-1},m-1}(l_1^{k_{m-1}},l_2^{k_{m-1}},\cdots,l_{m-2}^{k_{m-1}})$,则必有

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$S_{N,m}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}},k_{m-1})>S_{N,m}(l_1^{k_{m-1}},l_2^{k_{m-1}},\cdots,l_{m-2}^{k_{m-1}},k_{m-1})$$

这与 $S_{N,m}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}},k_{m-1})$ 是最优 m 分类的前提相矛盾，故式(7)正确。

设已得到样本前 p 组数据的最优 $m-1$ 分类，$p=m-1,m-2,\cdots,N$。其总变差为

$$S_{p,m-1}(k_1^p,k_2^p,\cdots,k_{m-2}^p)\quad p=m-1,m-2,\cdots,N$$

求整个样本的第 $m-1$ 个分类点 k_{m-1}，使式(7)右端函数所表示的总变差最小

$$\min_{m-1\leqslant k_{m-1}<N} S_{N,m}(k_1^{k_{m-1}},k_2^{k_{m-1}},\cdots,k_{m-2}^{k_{m-1}},k_{m-1}),k_{m-1}=m-1,m-2,\cdots,N-1$$

则有最优 m 分类：$\{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{k_1^*}\},\{\boldsymbol{X}_{k_1^*+1},\cdots,\boldsymbol{X}_{k_2^*}\},\cdots,\{\boldsymbol{X}_{k_{m-1}^*+1},\cdots,\boldsymbol{X}_N\}$。

2 分类方法在农业机械化发展水平评判中的应用

笔者将上述方法用于评判我国各省市区农业机械化发展水平的分类研究①。

表1示出我国农业机械化水平评价的指标体系。在对全国和各地区农业机械化总体水平进行评价时，对6个方面的水平配以适当的权重并综合，权值用专家调查法得到，不同权重对分类结果有影响。

表1 我国农业机械化水平评价的指标体系

农业机械化水平	评价指标	权重
作业水平	耕、播、收综合机械化水平	0.4
效益水平	农业劳均产值	0.1
	农业劳均产粮	0.1
结构水平	第一产业从业人员占全社会从业人员的比重	0.05
	第一产业在国内生产总值中所占比重	0.05
经济水平	人均国内生产总值	0.05
	农民人均年纯收入	0.05
	农业生活费用中现金支出所占比重	0.05
规模水平	农业劳均播种面积	0.10
文化水平	从业人员中初中以上文化程度所占比重	0.05

根据1997年数据将全国各地农业机械化水平分类，结果见表2。若分为3类，则将第二、三类并为1类；若分为5类，则将第4类中海南省及其后省区分为1个新类。具体计算结果表明该方法的分类过程非常稳定。

表2 我国各省市区农业机械化水平分类结果

分类	地 区
1	北京，上海，黑龙江，天津，新疆
2	江苏，辽宁，河北，山东，吉林，内蒙古，山西，河南
3	浙江，安徽，宁夏，广东，青海，陕西
4	湖北，福建，江西，湖南，甘肃，海南，广西，四川，重庆，云南，贵州，西藏

说明：不包括台湾省。

①白人朴. 关于我国农业机械化水平的评价研究报告. 中国农业大学农村发展研究所，1998

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 1 示出分类数和总变差的关系。可以看出,分类数越大,总变差越小,但各类之间的显著性也越小。上述分类问题的分类数在 4 左右比较合适。

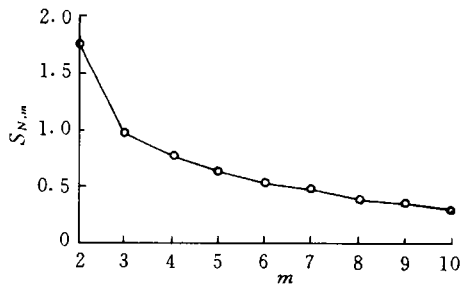

图 1　我国农机水平分类的总变差 $S_{N,m}$ 与分类数 m 的关系

3　结论与建议

1)基于动态规划思想提出了一种有序样本的分类方法。这种方法能够识别子样之间的分点,将整个样本划分成不同子样,使整个分类达到某种目标下的最优。做多指标样本分类时,样本数据的正规化可以采取线性化方法或正态变换方法。

2)将有序样本分类方法用于评判我国农业机械化发展水平的分类,把各省市区的农机化水平显著地分为 4 类;各类型在作业、效益、结构、经济、规模和文化诸方面所形成的综合水平具有显著差异。

3)用有序样本分类方法研究我国农业机械化发展水平问题,有 2 个方面的工作尚待完善:一是深入分析各类之间差异的特征;二是对多年资料进行分析,从时序的角度来考察各类型的动态变化规律。

参　考　文　献

1　黄连洪,张世英.模型结构变化点检测算法——GBV 法.应用数学学报,1987,10(3):267～275
2　黄连洪,张世英,刘　豹.社会经济系统建模的回归分类法.天津大学学报,1985(2):47～56
3　黄连洪,刘　豹,张世英.变结构模型的最优分类法(二).系统工程学报,1986,1(1):1～13
4　杨敏丽.农业机械化发展阶段性与区域不平衡性研究:[学位论文].北京:中国农业大学,1998.36～58

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net