

具有回路约束的有序样品的聚类方法

费荣昌

(基础课部)

摘要

本文研究了具有回路约束的有序样品的聚类方法, 该法可以在计算机上实现, 并给出了两个计算实例。

关键词: 有序样品; 回路约束; 聚类方法

Fisher提出的最优分割法^[1,2]是解决有序样品聚类问题的常用方法, 本文把最优分割法推广到样品具有回路约束的情形。

设 n 个样品 X_1, X_2, \dots, X_n (都是 m 维向量)是有序的, 且 X_n 不仅与 X_{n-1} 相邻, 还与 X_1 相邻, 称这 n 个样品具有回路约束。不改变样品的次序, 把具有回路约束的 n 个样品分成 k 类, 一切可能的分法有 C_n^k 种。设某一种分法是 $p(n, k): \{X_{i1}, \dots, X_{i2-1}\}, \{X_{i2}, \dots, X_{i3-1}\}, \dots, \{X_{ik}, \dots, X_{i1-1}\}$, 这里约定 X_0 就是 X_n , X_{n+1} 就是 X_1 。定义这种分类的目标函数为

$$e[p(n, k)] = \sum_{j=1}^k D_j,$$

其中 D_j 是第 j 类 $G_j = \{X_{i1}, \dots, X_{i(j+1)-1}\}$ 的离差平方和:

$$D_j = \sum_{X_i \in G_j} (X_i - \bar{X}_j)'(X_i - \bar{X}_j),$$

求分法 $p(n, k)$, 使目标函数达到最小。

1 算法

具有回路约束的 n 个有序样品有 n 个间隔, 分成 k 类就是在这 n 个间隔中选择 k 个间隔进行分割, 有下列 $n-k+1$ 种情况。

(1)先在 X_n 和 X_1 之间分割, 化成有序样品 X_1, X_2, \dots, X_n , 然后在 $n-1$ 个间隔中选择 $k-1$ 个间隔进行分割, 共有 C_{n-1}^{k-1} 种分法, 从这些分法中求出目标函数取值最小的分法 $p_1(n, k)$, 可按下述递推公式计算:

$$e[p_1(i, j)] = \min_{j \leq l \leq i} \{e[p_1(l-1, j-1)] + D_j\}, \quad \begin{matrix} i = j, \dots, n, \\ j = 2, \dots, k. \end{matrix}$$

其中 $e[p_1(l-1, 1)]$ 是 $G_1 = \{X_1, \dots, X_{l-1}\}$ 的离差平方和, D_j 是 $G_j = \{X_j, \dots, X_i\}$ 的

离差平方和。

(2)先在 X_{n-1} 和 X_n 之间分割,并把 X_n 和 X_1 看作一个样品,即化成有序样品 $(X_n, X_1), X_2, \dots, X_{n-1}$,然后在 $n-2$ 个间隔中选择 $k-1$ 个间隔进行分割,共有 C_{n-2}^{k-1} 种分法,从这些分法中求出目标函数取值最小的分法 $p_2(n-1, k)$ 。可按下述递推公式计算:

$$e[p_2(i, j)] = \min_{j \leq l \leq i} \{e[p_2(l-1, j-1)] + D_j\}, \quad \begin{matrix} i = j, \dots, n-1; \\ j = 2, \dots, k. \end{matrix}$$

其中 $e[p_2(l-1, 1)]$ 是 $G_1 = \{X_n, X_1, \dots, X_{l-1}\}$ 的离差平方和, D_j 是 $G_j = \{X_1, \dots, X_j\}$ 的离差平方和。

(3)先在 X_{n-2} 和 X_{n-1} 之间分割,并把 X_{n-1}, X_n 和 X_1 看作一个样品,即化成有序样品 $(X_{n-1}, X_n, X_1), X_2, \dots, X_{n-2}$,然后在 $n-3$ 个间隔中选择 $k-1$ 个间隔进行分割,共有 C_{n-3}^{k-1} 种分法,从这些分法中求出目标函数取值最小的分法 $p_3(n-2, k)$ 。可按下述递推公式计算:

$$e[p_3(i, j)] = \min_{j \leq l \leq i} \{e[p_3(l-1, j-1)] + D_j\}, \quad \begin{matrix} i = j, \dots, n-2; \\ j = 2, \dots, k. \end{matrix}$$

其中 $e[p_3(l-1, 1)]$ 是 $G_1 = \{X_{n-1}, X_n, X_1, \dots, X_{l-1}\}$ 的离差平方和, D_j 是 $G_j = \{X_1, \dots, X_j\}$ 的离差平方和。

如此做下去。

$n-k+1$ 先在 X_k 和 X_{k+1} 之间分割;并把 $X_{k+1}, X_{k+2}, \dots, X_n$ 和 X_1 看作一个样品,即化成有序样品 $(X_{k+1}, X_{k+2}, \dots, X_n, X_1), X_2, \dots, X_k$,然后在 $k-1$ 个间隔中进行分割,这时只有 $C_{k-1}^{k-1} = 1$ 种分法 $p_{n-k+1}(k, k)$,并且 $e[p_{n-k+1}(k, k)]$ 就是 $G_1 = \{X_{k+1}, X_{k+2}, \dots, X_n, X_1\}$ 的离差平方和。

最后由

$$e[p(n, k)] = \min \{e[p_1(n, k)], e[p_2(n-1, k)], \dots, e[p_{n-k+1}(k, k)]\}$$

即得具有回路约束的 n 个样品分成 k 类的最优分法 $p(n, k)$ 。由于

$$C_{n-1}^{k-1} + C_{n-2}^{k-1} + \dots + C_{k-1}^{k-1} = C_n^k,$$

所以 $p(n, k)$ 是在比较一切可能的 C_n^k 种分法后求出的最优分法,它是回路约束有序样品聚类问题的精确最优解。

2 实例

例1 某市市区环行河道各监测点的水质(单位: ppm)如表1。

表 1 各监测点的水质

项目 监测点	pH	色度	COD	DO	NH ₃ -N	石油类	挥发酚	Tcr
1	7.53	25	12.13	0.89	5.09	0.54	0.007	0.018
2	7.59	28	15.80	0.99	5.20	0.96	0.018	0.022
3	7.80	20	7.43	4.03	1.91	0.46	0.010	0.010
4	7.83	15	6.47	4.90	1.46	0.34	0.002	0.011
5	7.88	20	5.42	5.93	1.18	0.30	0.003	0.013
6	7.83	18	7.81	4.98	1.68	0.42	0.002	0.014
7	7.60	24	9.77	2.44	3.78	0.38	0.014	0.023

如分成两类(重污染区和严重污染区), 使用本算法($n=7, m=8, k=2$), 得分类结果如下:

类	重污染区 (5级水体)	严重污染区 (6级水体)
监测点	3, 4, 5, 6	7, 1, 2

例2 古运河饮料公司某工厂的饮料月销售量(单位: 箱)如表2。

表 2 某饮料厂月销售量

月 品种	1	2	3	4	5	6	7	8	9	10	11	12
鲜桔	7837	20214	25237	30609	50474	128747	139543	99772	24941	9474	5263	3429
柠檬	1404	5337	5656	21310	44102	82627	81257	63711	19537	13252	2786	1337

如分成两类(旺季和淡季), 使用本算法($n=12, m=2, k=2$), 得分类结果如下:

类	旺 季	淡 季
月	6, 7, 8	9,10,11,12,1,2,3,4,5

如分成四类(旺季、淡季和两个过渡期), 使用本算法($n=12, m=2, k=4$), 得分类结果如下:

类	过渡期	旺 季	过渡期	淡 季
月	5	6, 7	8	9, 10, 11,12, 1, 2, 3, 4

上两实例的计算是在张建华同志的协助下完成的, 表示感谢。

参 考 文 献

[1] Fisher W D. On Grouping for Maximum Homogeneity, J Amer. Statist,

Assoc, 1958, 53, 789~798

[2] 张尧庭.方开泰, 多元统计分析引论.科学出版社, 1982

The Clustering Method for the Order Sample with Subject to Circle

Fei Rongchang

Abstract

In this paper, the clustering method is studied for the order sample with subject to circle. The calculation of these methods can be easily done by the computer, and two numerical examples are given.

Subjectwords, order sample, subject to circle, clustering method