第5卷 第1期 应用数学学报 Vol. 5, No. 1
1982 年 1 月 ACTA MATHEMATICAE APPLICATAE SINICA Jan., 1982

有序样品的一些聚类方法

方 开 泰
(中国科学院应用数学研究所)

SOME CLUSTERING METHODS FOR THE ORDER SAMPLE

Fang Kai-tai
(*Institute of Applied Mathematics, Academia Sinica*)

Abstract

In this paper, the authors generalize Fisher's[6] clustering method for the order sample. First, several distances are used for this clustering. Then, the hierachical clustering procedure is applied to that of the order sample. With this procedure both the computed time and the memory storage are considerably saved. Finally, the concept of order sample in two dimensions is presented.

有序样品[2]就是样品的次序不能打乱的样品，在地质勘探，天气预报，天体演化等领域是经常出现的，并且需要将它进行聚类. 目前国内外流行的有序样品聚类法是 Fisher[6]提出的，国内称做最优分割法. 本文将最优分割法作了一些推广，并将系统聚类法[1]推广到有序样品的情况.

一、最优分割法的推广

设样品为 $x_1, x_2, \cdots, x_n$，每个均为 m 维向量，样品是有序的. 将这 n 个样品分成 k 类，设某一种分法是:

$$P(n,k)_i\{x_{i_1}, x_{i_1+1}, \cdots, x_{i_2-1}\}, \{x_{i_2}, x_{i_2+1}, \cdots, x_{i_3-1}\}, \cdots,$$
$$\{x_{i_k}, x_{i_k+1}, \cdots, x_n\},$$

其中 $1 = i_1 < i_2 < \cdots < i_k < n$，用 $D(i,j)$ 表示类 $\{x_i, x_{i+1}, \cdots, x_j\}$ $(i < j)$ 的直径，目前采用的直径为[7]

$$D_1(i,j) = \sum_{l=i}^{j} (x_l - \bar{x}_{ij})'(x_l - \bar{x}_{ij}), \tag{1.1}$$

其中

$$\bar{x}_{ij} = \frac{1}{j-i+1} \sum_{l=i}^{j} x_l, \tag{1.2}$$

本文 1979 年 3 月 21 日收到.

当 $m=1$ 时，有时用直径

$$D_2(i,j)=\sum_{l=i}^{j}|x_l-\tilde{x}_{ij}|, \tag{1.3}$$

$\tilde{x}_{ij}$ 是 $(x_i, x_{i+1},\cdots,x_j)$ 的中位数. 正如样品间的距离和类与类之间有多种定义一样，直径也应有多种定义方法，否则不能适应千变万化的实际情况，本节将讨论直径的其它的定义法.

对分类 $P(n,k)$ 定义误差函数

$$e[P(n,k)]=\sum_{l=1}^{k}D_j(i_l,i_{l+1}-1)\quad(j=1,2), \tag{1.4}$$

最优分割法就是选择 $P(n,k)$ 使 (1.4) 式达到极小，从而给出精确最优解. 但是 (1.1)—(1.4) 式定义的误差函数对有些情况并不合适，见下例.

例 1　$m=1$，$n=23$，其 $x_1,\cdots,x_n$ 的数值是：1, 2, 3,……,20, 24, 27, 30，即前 20 个样品的指标按等差级数上升，差数为 1，最后三个样品也按等差级数上升，但差数为 3，将它们点成图 1. 欲将它们分类.

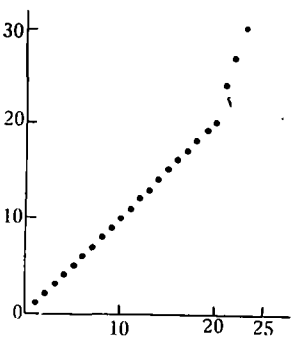


图 1

直观上很明显，应分成两类 $\{x_1,x_2,\cdots,x_{20}\}$，$\{x_{21},x_{22},x_{23}\}$，如果用最优分割法，由 (1.1)—(1.4) 定义的误差函数，则两类的最优分割（简称最优二分割）为 $\{x_1,x_2,\cdots,x_{13}\}$，$\{x_{14},x_{15},\cdots,x_{23}\}$. 显然这是很不合适的，产生的原因是误差函数定义得不合适，因此需要有适合于类似例 1 (显然这是一个人为的例子)的误差函数. 回想起在系统聚类法中最短距离法和其它方法相比有其独特之处，它适用于条形的甚至 S 形的类，这正适用于例 1 的情况.

m 维空间中的 n 个点 $y_1,\cdots,y_n$ 之间如果定义了距离(记作 $d(y_i,y_j)$)，则可求得它们的最小支撑树 MST，有关最小支撑树的概念和算法可参阅 [3] 的第三章或图论方面的书，最小支撑树的最大边长和全部边长分别记作 MMST 和 SMST，利用这些概念可以定义类的直径，即

$$D_3(i,j)=\mathrm{MMST}\{x_i,x_{i+1},\cdots,x_j\} \tag{1.5}$$

或

$$D_4(i,j)=\mathrm{SMST}\{x_i,x_{i+1},\cdots,x_j\} \tag{1.6}$$

相应地定义两种误差函数

$$e[P(n,k)]=\max_{1\leqslant l\leqslant k}\mathrm{MMST}\{x_{i_l},x_{i_l+1},\cdots,x_{i_{l+1}-1}\}, \tag{1.7}$$

$$e[P(n,k)]=\sum_{l=1}^{k}\mathrm{SMST}\{x_{i_l},x_{i_l+1},\cdots,x_{i_{l+1}-1}\}, \tag{1.8}$$

如果在 x_i 与 x_j 之间定义了距离 $d(x_i,x_j)$，仿照最长距离法可以定义

$$D_5(i,j)=\max_{i\leqslant l<t\leqslant j}d(x_l,x_t), \tag{1.9}$$

$$e[P(n,k)]=\sum_{l=1}^{k}D_5(i_l,i_{l+1}-1). \tag{1.10}$$

对于误差函数 (1.8) 和 (1.10) 除了类的直径算法不同外，其余均同最优分割法. 对于目标函数 (1.7)，其误差函数的递推公式变为

$$e[P(n,2)] = \min_{2\leqslant j\leqslant n} \max\{D_3(1, j-1), D_3(j, n)\}, \tag{1.11}$$

$$e[P(n,k)] = \min_{k\leqslant j\leqslant n} \max\{e[P(j-1, k-1)], D_3(j, n)\}, \tag{1.12}$$

其余均可仿最优分割法. 因此一旦将各类的直径算好后，上述的各种方法均可编成一个统一的计算机程序.

为了计算 $\{D_3(i, j)\}$, $\{D_4(i, j)\}$，需要计算各类的 MST，由于 $1 \leqslant i < j \leqslant n$，对每个 $i < j$，独立地计算 $\{x_i, x_{i+1}, \cdots, x_j\}$ 的 MST 是很麻烦的，自然希望有某种递推算法，这就是：(1) 如果已有了 $\{x_1, \cdots, x_n\}$ 的 MST，当加入一个新点 x_{n+1} 后，求新的 MST；(2) 已有了 $\{x_1, \cdots, x_n\}$ 的 MST，当从中剔除某个点后，求剩下 $(n-1)$ 个点的 MST. 下面我们分别来解决.

问题 (2) 比较简单，设从 $\{x_1, \cdots, x_n\}$ 中剔除 x_i，这时在 MST 中通过 x_i 的边 (即连线，下同)全部除去，于是原来的 MST 被剖分成 l 个不相交的部分 $(2 \leqslant l \leqslant n-1)$，记作 $P_1, \cdots, P_l$，然后对 $1 \leqslant i < j \leqslant l$ 中任意 i, j，计算 P_i 与 P_j 的最近距离，即找出

$$\min_{x_s\in P_i, x_t\in P_j} d(x_s, x_t)$$

并把达到极小的相应连线添加进去，即为新的 MST，图 2 示意地反映了这个过程. 证明是容易的. 从略.

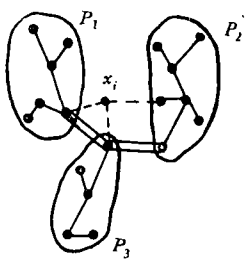


图 2

—— 从原 MST 去掉的边

━━ 新的 MST 的边

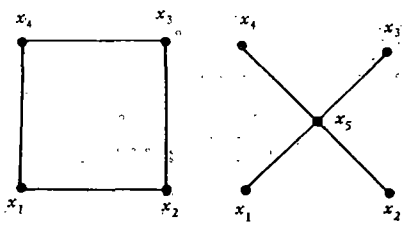


图 3

(a) 原 MST (b) 新 MST

对于问题(1)，稍为复杂一点，因为加进一个点，可能使原 MST 的边全部换掉，如图 3 中原有四点组成正方形，当加进中心点后，MST 的边全部换了(当然剔除一个点后也有全部换边的可能)；另外如何保证产生的新支撑树是 MST，比剔除时要麻烦一些. 加入 x_{n+1} 后产生新 MST 的步骤是：

(i) 计算距离 $d(x_{n+1}, x_i)$, $(i = 1, 2, \cdots, n)$. 在原图上任添一条边，比如 $\overline{x_1x_{n+1}}$.

(ii) 在图上依次添上边 $\overline{x_ix_{n+1}}(i = 2, 3, \cdots, n)$，这时一定存在一个回路，如果 $\overline{x_ix_{n+1}}$ 是回路中的最长边，则该边不能添上；如果 $\overline{x_ix_{n+1}}$ 不是回路中的最长边，则将它添上，而将最长边去掉(参看下面例 2).

经过步骤 1 和 2，所得的联结图即为新的 MST，其证明如下：由 MST 的定义只要考虑加进边 $\overline{x_ix_{n+1}}(i = 1, 2, \cdots, n)$ 后来调整联结图就可以了. 显然用步骤 1 和 2 得到的

联结图(记作 μ)肯定是支撑树，它是不是最小支撑树呢？因为凡 $\{\overline{x_i x_{n+1}}, i=1, \cdots, n\}$ 中不在 μ 中的边，按步骤 2，这些边一定是某个回路中的最大边长，因此如果我们证明了下面的引理，μ 就是新的 MST.

引理　如果边 d 是某个联结图中某个回路的最大边长，则 d 一定不在该联结图的 MST 中

证　如若不然，边 d 属于 MST，将 d 丢掉，MST 分成两个不相交的部分 P 和 Q，由于 d 是某回路中的最大边长，在回路中一定存在一个边 e，它的端点一个属于 P，另一个属于 Q，而 e 的长度小于 d，这说明 $d \notin \lambda(P, Q)$，其中 $\lambda(P, Q)$ 表示 P 和 Q 的“截集”中达到边长最小的集合[3]，由 [3] § 3.1 的定理 2，d 应属于 $\lambda(P, Q)$，矛盾，故 d 必不在 MST 中，证毕.

例 2　图 4(a) 中原有 A、B、C、D、E 五个点，图中画着它的 MST，现增加一个点 F，欲求新的 MST.

经计算，点 F 至各点的距离为 $\overline{AF}=9$，$\overline{BF}=10.5$，$\overline{CF}=8$，$\overline{DF}=3$，$\overline{EF}=2$，如果首先添上边 $\overline{AF}$，随后添上边 $\overline{BF}$ (见图 4(b))，这时产生一回路 ABF，其中 $\overline{BF}$ 是最长边，将它丢掉，再加上边 $\overline{CF}$，这时产生一回路 $ABCF$ (见图 4(c)) 舍掉其中的最长边 $\overline{AF}$. 然后加上 $\overline{FD}$，丢掉 $\overline{AE}$；最后加上 $\overline{FE}$，丢掉 $\overline{DE}$，得到了新的 MST，见图 4(d).

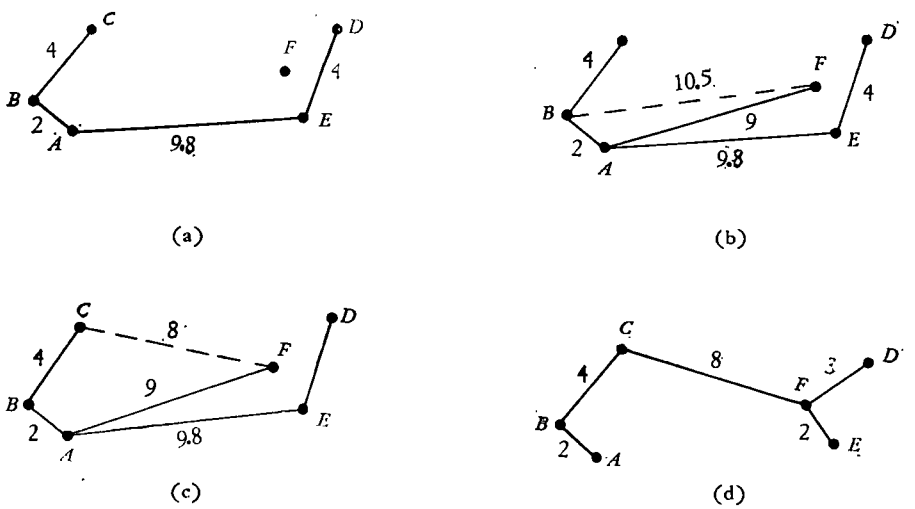


图 4

容易看出对例 1 运用 (1.7)或 (1.8) 或 (1.10) 定义的误差函数，其最优二分割正好是

$$\{x_1 \sim x_{20}\},\ \{x_{21}, x_{22}, x_{23}\}.$$

顺便指出，最短距离法得到的类是在误差函数 (1.7) 的意义下的精确最优解，这将另文讨论.

对例 1 的情况，误差函数 (1.7)、(1.8)、(1.10) 比 (1.4) 合适，但在其它的场合可能正相反，因此要具体问题具体分析，在实际中通常采用的方法是几种误差函数同时计算，然后选择符合实际情况的分类.

二、有序样品的系统聚类法

最优分割法由于要计算直径阵 $\{D(i,j)\}$ 及误差函数阵 $\{e[P(j,i)]\}$，当 n 比较大时，计算量还是很大的，而且占用了大量的内存，因此希望有其它简便的方法来克服上述困难．我们可将系统聚类法用于有序样品的情况，其步骤是：

1．开始 n 个样品各成一类，组成 $G_1,\cdots,G_n$，计算相邻两类的距离 $D(G_i, G_{i+1})$，$i=1,2,\cdots,n-1$．

2．将距离最近的两类合并，然后计算新类与相邻类的距离．如果全部类都已组成一类，则过程中止，否则回到 2．

运用这个方法可大大节省计算和内存．我们先看一个例子，然后再进一步讨论．

例 3 这是 [2] 中的例 6.1，即男孩体重增长的例子，其数据是：

年 龄	1	2	3	4	5	6	7	8	9	10	11
增加重量(公斤)	9.3	1.8	1.9	1.7	1.5	1.3	1.4	2.0	1.9	2.3	2.1

如用最短距离法，首先计算相邻两类的距离，如样品间采用绝对值距离，则相邻两类的距离分别为：7.5，0.1，0.2，0.2，0.2，0.1，0.6，0.1，0.4，0.2，以 0.1 最小，故将 x_2 与 x_3，x_6 与 x_7，x_8 与 x_9 合并，合并的类记成 G_{12}，G_{13}，G_{14}．然后计算新类与相邻类的距离，运用最短距离法定义 $D(G_1, G_{12})=7.5$，$D(G_{12}, G_4)=0.1$，$\cdots$，$D(G_{14}, G_{10})=0.3$．这时再选择最小的相邻类距离是 $0.1=D(G_{12}, G_4)=D(G_5, G_{13})$，故将 G_{12} 与 G_4，G_5 与 G_{13} 合并……，其并类过程由图 5 表示，其中的数字表示并类的距离，由图看出，大致以分为三类或四类为宜．

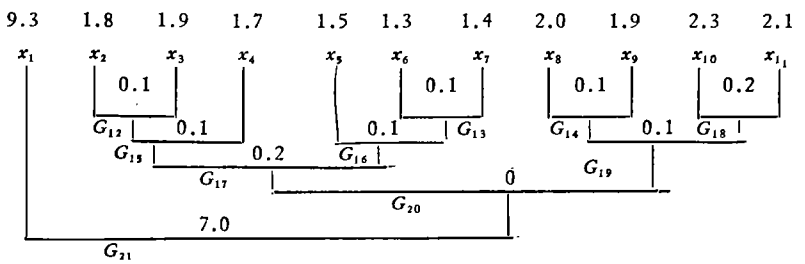


图 5

从这个简单的例子，我们看出系统聚类法用于有序样品有其矛盾的特殊性表现在：

(i) 并类距离不一定有单调性．用 D_1 表示第一次合并两类时这两类的距离，第二次合并的两类距离记作 $D_2,\cdots$，如果 $D_1\leqslant D_2\leqslant\cdots\leqslant D_{n-1}$，则称并类距离具有单调性，[3] 的 §2.7 证明了最短距离法、最长距离法、类平均法、离差平方和法都具有并类距离的单调性，而重心法不一定有单调性，当并类距离有单调性时，可以画出聚类图(谱系图)，并且可以定量地决定分类的个数．而有序样品采用上述方法时，一般均不一定有并类距离的单调性，从而又能示意地画出并类过程，画不出聚类图，例 3 就表明了这一点，这是这种方法的一个缺点．

克服这缺点的方法可用间隙法，它定义相邻两类的距离为两类中最相邻样品的距离，

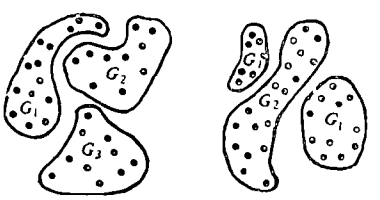
即两类间的间隙. 例如例 3 中类 $G_{16}=\{x_5, x_6, x_7\}$ 与类 $G_{19}=\{x_8, x_9, x_{10}, x_{11}\}$ 的间隙是 x_7 与 x_8 的距离，等于 0.6. 间隙法具有并类距离的单调性[4]. 但间隙法没有利用类中其它元素的信息，有时分的类并不十分合适.

(ii) 在通常的系统聚类法中，计算新类与旧类的距离时采用递推公式比较方便[1]，而在有序样品的情况，采用递推公式并不合算，因为那样需要存贮整个的距离阵，不但占用了大量的内存而且计算了许多用处极小的中间结果. 因此对有序样品，计算新类与旧类的距离时直接用各种类之间距离的定义.

一般来说，系统聚类法用于有序样品，比最优分割法的计算量小，占用内存少，且容易确定分类的个数. 对例 1 的情况，采用系统聚类法得到合理的分类，而最优分割法(用目标函数 (1.4)) 得到的是不合理分类，这是系统聚类法的优点. 但是系统聚类法只能得到局部最优解，而最优分割法可得精确最优解.

三、二维有序样品

在地质勘探中，对某个地区进行普查，根据矿的情况要将地区分类，这时同一类的样品在地理上必须是互相邻接的，邻接的概念在直观上是很清楚的，如图 6(a) 的分类是互相邻接的，而图 6(b) 则不然，现用数学方法来描述这个概念.



(a)　　(b)

图 6

设每个样品测了 m 个指标 $x_1,\cdots,x_m$，每个样品在平面上的坐标为 $\mathscr{Y}_1,\mathscr{Y}_2$，向量 $(\mathscr{Y}_1,\mathscr{Y}_2,\mathscr{X}_1,\cdots,\mathscr{X}_m)'$ 就全面代表了每个样品的信息，样品仍记成 $x_1,\cdots,x_n$，每个均是 $m+2$ 维向量. 欲将它们分成 k 类 $G_1,\cdots,G_k$.

定义 3.1 设平面上有 n 个点 $y_1,\cdots,y_n$，如果对任意 $k(2\leqslant k\leqslant n-1)$ 将其分成 k 类 $\tilde{G}_1,\cdots,\tilde{G}_k$，要求它们的最小支撑树 MST $(\tilde{G}_1),\cdots,$MST $(\tilde{G}_k)$ 是互不相交的，则称 $y_1,\cdots,y_n$ 为二维有序样品，类 $\tilde{G}_1,\cdots,\tilde{G}_k$ 内的样品是互相邻接的[5].

显然这是一维有序样品的推广. 如果 $\tilde{G}_1,\cdots,\tilde{G}_k$ 的凸包是互不相交的，则它们的 MST 也是互不相交的，反之不然.

如果样品 $x_1,\cdots,x_n$ 的前两个分量组成的向量是二维有序样品，则它们的聚类问题称为二维有序样品的聚类问题，记 $x_i=\begin{pmatrix}x_{i1}\\x_{i2}\end{pmatrix}$，$i=1,\cdots,n$，$x_{i1}$ 是二维向量，x_{i2} 是 m 维向量. 设将样品 $\{x_i\}$ 分成了 k 类 $G_1,\cdots,G_k$，记 $G_i^1=\{x_{l1}: x_l\in G_i, l=1,\cdots,n\}$，

$$G_i^2=\{x_{l2}: x_l\in G_i, l=1,\cdots,n\}; i=1,\cdots,k.$$

或者先有 $\{G_i^2\}$，用类似方式可得 $\{G_i^1\}$ 和 $\{G_i\}$. 所谓二维有序样品的聚类，就是寻求一种分法 $P(n,k)$，对某个误差函数(定义在 $\{G_i^2\}$ 上) $e[P(n,k)]$ 使其达到极小，而相应的 $\{G_i^1\}$ 满足定义3.1的有序条件. 这个问题在实际中很重要，但至今未能解决，这里仅仅给出一种初步处理的两个方法.

方法一：这是借用系统聚类法，首先 n 个样品各自成一类，记成 $G_1(0),\cdots,G_n(0)$；然后将其中的最近两类 (在有序的约束下) 归并，得到 $(n-1)$ 个类，记作 $G_1(1),\cdots,$

$G_{n-1}(1)$; 再将其中最近的两类归并，记作 $G_1(2),\cdots,G_{n-1}(2);\cdots$ 直至最后所有的元素都成一类为止. 按上述办法，当然同时也有类

$$\{G_i^l(0),\ 1\leqslant i\leqslant n\},\ \{G_i^l(1),\ 1\leqslant i\leqslant n-1\},\cdots,l=1,2.$$

(i) 给一个阈值 T. 首先每个样品各自成一类，在类 $\{G_i^1(k),\ 1\leqslant i\leqslant n-k\}$, $k=0,1,\cdots,n-1$ 之间采用最短距离法中的定义，记作 $D_1[G_i^1(k),\ G_j^1(k)]$; 在类 $\{G_i^2(k),\ 1\leqslant i\leqslant n-k\}$ 之间的距离采用系统聚类法中八种定义的任一种，记作

$$D_2[G_i^2(k),\ G_j^2(k)].$$

(ii) 对每个 k(开始 $k=0$，以后每步 k 加 1)，定义

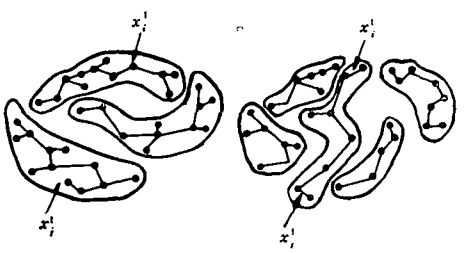
$$\mathscr{I}_i^1(k)=\{G_j^1(k):D_1[G_i^1(k):G_j^1(k)]\leqslant T,\ j\neq i,\ j=1,2,\cdots,n-k\},\quad(3.1)$$

然后在 $\mathscr{D}_k=\{D_2[G_i^2(k),\ G_j^2(k)];\ i=1,\cdots,n-k;\ G_j^1(k)\in\mathscr{I}_i^1(k)\}$ 中选择达到极小的两类，比如是 $G_{i_1}^2(k)$ 和 $G_{j_1}^2(k)$，如果 MST $\{G_{i_1}^1(k),\ G_{j_1}^1(k)\}$ 与 MST $\{G_i^1(k)\}i=1,\cdots,n-k, i\neq i_1, j_1$ 互不相交，则将 $G_{i_1}(k)$ 与 $G_{j_1}(k)$ 合并，得到

$$\{G_i(k+1),\ 1\leqslant i\leqslant n-k-1\};$$

如果新来的 MST 与旧类相交，则选择 $\mathscr{D}_k$ 中达次小的两类，按方才原则处理，如新类的 MST 与旧类的 MST 相交，在 $\mathscr{D}_k$ 中再选择达第三小的两类，如此下去. 只要其中有一步新类的 MST 与旧类均不交，就可得到 $\{G_i(k+1),\ 1\leqslant i\leqslant n-k-1\}$，这时如 $k<n$ 则回到 (ii)，否则停止. 如果 $\mathscr{D}_k$ 中所有的对并成新类后的 MST 均与旧类相交，则过程也中止.

如果 $\{x_i^1\}$ 是一维的数，问题就化为一维的有序样品，当 $T\geqslant\{x_i^1\}$ 的最大间隙时，显见这里的方法与本文第二节的方法是一致的，以 $T=\{x_i^1\}$ 的最大间隙为计算量最小. 对二维有序样品，T 也不应取得太大，如图 7(a)，T 取得很大，$T\geqslant\max\limits_{i\neq j}D_1[G_i^1(0),\ G_j^1(0)]$，图中画的类是实际应分的类，但由于 $D_2[G_i^2(0),\ G_j^2(0)]$ 特别小，而 $D_1[G_i^1(0),\ G_j^1(0)]$ 又特别大，第一次并类时将 $G_i(0)$ 与 $G_j(0)$ 合并了，从而得到图 7(b) 的不合理分类. 当然，为了全部样品都能归类，应有



(a) (b)

图 7

$$T\geqslant\text{MMST}\{x_i^1,1\leqslant i\leqslant n\},$$

在使用这个方法时可让 T 分成 n 个等级，分别计算，来决定分类的取舍.

方法二：在最优分割法中，为了节省内存和计算，可采用多次二分割的办法[3,6]，现在我们就采用这个办法. 前面曾经说，如 $\{G_i^1(k),\ 1\leqslant i\leqslant n-k\}$ 的凸包互不相交，则一定保证它们的 MST 互不相交. 平面上 n 个点，分成两类的一切可能有 $2^{n-1}-1$ 种，当要求这两类的凸包是互不相交时，只有 $n(n-1)$ 种可能[8]，且可用如下的方法来确定这些分类：记这 n 个点是 $y_1,\cdots,y_n$，n 个点中任取二个点 y_i, y_j 可连一条直线，将平面上的点分成两部分 H_1 和 H_2，先将 y_i 归入 H_1，y_j 归入 H_2；然后将 y_i 归入 H_2，y_j 归入 H_1(y_i, y_j 同归入一类的分法与用其它直线的划分有重复，故不考虑这种情况)共得两种分类，故总

共有 $2\times\frac{n(n-1)}{2}=n(n-1)$ 种分法．将这个思想用到$\{x_i', 1\leqslant i\leqslant n\}$上得到如下的方法，其步骤是：

(i) 将$\{x_i^1, 1\leqslant i\leqslant n\}$用上述方法得到 $n(n-1)$种分法，对应到$\{x_i^2, 1\leqslant i\leqslant n\}$也有 $n(n-1)$ 种分法，计算它们的误差函数(选择本文第一节讨论的任一种)，取达到极小的一种分法，所分的两类记作 G_1 和 G_2 (相应的有 G_1^1, G_2^1 和 G_1^2, G_2^2).

(ii) 然后对 G_1 和 G_2 分别应用 (i)，试图将它们分成两类，看那个的误差函数更小，譬如分割 G_2 误差函数小，就将它按 (i) 的方法分成两类，记作 G_3、G_4，这时有三类 G_1, G_3, G_4.

(iii) 设已分成 k 类，将它们分别用 (i) 试分成两类，然后选择使误差函数达到极小的一个进行分割，得到 $k+1$ 类.

控制这个过程的结束有多种方法，例如：

(a) 直分割到 n 类为止，然后将分割过程画成聚类图，以决定分类个数；

(b) 事先确定类的数目 k，当分割到 k 类后过程即停止.

(c) 给误差函数一个阈值 T，因误差函数是分类个数的(严格)单调下降函数，故当误差函数一旦小于 T 后过程即停止.

显然，这里的概念和方法可直接推广到更多维的有序样品.

本工作曾得到田丰同志的帮助，特此致谢.

参　考　文　献

[1] 方开泰，聚类分析 I，数学的实践与认识，1978 年第 1 期，66—80.
[2] ——，聚类分析 II，数学的实践与认识，1978 年第 2 期，54—62.
[3] ——，聚类分析，多元分析资料汇编 (III)，中国科学院数学研究所概率统计室印，1976 年.
[4] 张启锐，有序样品的费歇分段法与简单分段法，数学地质会议资料，中国科学院地质研究所印，1978 年.
[5] L. Fisher and J. W. V. Ness, Admissible Clustering Procedurer, *Biometrika*, 58:1 (1971), 91—104.
[6] W. D. Fisher, On Grouping for Maximum Homogeneity. *J. Amer. Statist, Assoc.*, 53(1958), 789—798.
[7] J. A. Hartingan, Clustering Algorithms, Wiley, 1975.
[8] A. J. Scott and M. J. Symons, On the Edwards and Cavalli-Storza Method of Cluster Analysis, *Biometris*, 27(1971), 217—219.