第15卷第4期　　西北农业大学学报　　Vol.15,No.4
1987年12月　　Acta Univ.Septentrionali Occident.Agric.　　Dec.,1987

聚类分析的几个问题

王乃信

（西北农业大学计算机中心）

摘　要

本文提出优聚类的统一定义，从而概括了众多的聚类方法，建立了聚类函数的若干性质，证明了等价关系系统与系统聚类法的等价性，推广了Fisher算法，证得全序递归聚类法必得优聚类，得出最优聚类的一些充分条件。

关键词：聚类函数；优聚类；最优聚类；等价关系系统；递归聚类法

1　优聚类的统一定义

聚类分析中的聚类方法很多且差异很大．文献[1]中罗列了C.R.Rao、张尧庭等人所给出的类的一般定义，这些定义各有局限，都不能达到统一的目的．我们不从类的定义出发，而借助于聚类函数，直接给出优聚类的定义，便可以把众多的聚类方法统一起来．

由于实用中的聚类问题多为有限样本，本文恒设样本集有限．

设G为样本集，F为G的幂集，构造

$L=\{P \mid P=\{A_1, A_2, \cdots, A_k\}, A_1, A_2 \cdots A_k \in F$ 且两两不交$\}$，

其中k为任意的非负整数．如果$P=\{A_1, A_2, \cdots, A_k\} \in L$，令$A_1 \cup A_2 \cdots \cup A_k = A$，则称P为A的k级分割；如果$P_1$，$P_2$同为A的k级分割，则称$P_1$，$P_2$为A的同级分割．

定义1　设S(A)为定义在F上的实值函数，且满足

（1）　$S(\phi)=0$，

（2）　对任何$A \in F$，$S(A) \geqslant 0$，

则称S为G的离散函数．

定义2　设S为G的离散函数，相应地，Q(P)为定义在L上的非负实值函数，且满足

（1）对任何$A \in F$，$Q(\{A\})=S(A)$，

（2）存在二元单调非减函数f，使当$P_1 \subset P$时，恒有$Q(P)=f(Q(P_1), Q(P \cap \overline{P}_1))$，

（3）　$Q(\phi)=0$，

则称Q为G的聚类函数．

本文于1986年7月21日收到。

定义3 设Q为G的聚类函数，P为A的k级分割，如果P在A的同级分割中使Q达到最小值，则称P为A关于Q的k级优聚类.特别地，如果P的子集均为优聚类，则称P为A关于Q的k级最优聚类.

有限样本集关于聚类函数的优聚类恒存在.

上述定义概括了很多聚类方法.为了简化叙述，约定用$d(x_i, x_j)$表示x_i，x_j之间的某种距离（例如：Euclid距离，Minkowski 距离，Mahalanobis 距离，各种模糊距离等），约定用$c(x_i, x_j)$表示x_i，x_j之间的某种相似系数(例如：相关系数，相关指数，夹角余弦，各种模糊相似系数等）.

最短距离法

离散函数：对任何$A\in\mathbf{F}$，当A中没有相异元素时，$S(A)=0$（下同），否则

$$S(A)=\max_{B\subset A}\ \min_{\substack{x_i\in B\\ x_j\in A\cap\overline{B}}} d(x_i, x_j),$$

或

$$S(A)=1-\min_{B\subset A}\ \max_{\substack{x_i\in B\\ x_j\in A\cap\overline{B}}} c(x_i, x_j).$$

其中B和$A\cap\overline{B}$非空（下同）.聚类函数：对任何$P\in\mathbf{L}$，

$$Q(P)=\max_{A_i\in P} S(A_i).$$

最长距离法

离散函数：对任何$A\in\mathbf{F}$，

$$S(A)=\max_{\substack{x_i\in A\\ x_j\in A}} d(x_i, x_j),$$

或

$$S(A)=1-\min_{\substack{x_i\in A\\ x_j\in A}} c(x_i, x_j).$$

聚类函数：对任何$P\in\mathbf{L}$，

$$Q(P)=\max_{A_i\in P} S(A_i).$$

重心法

离散函数：对任何$A\in\mathbf{F}$

$$S(A)=\max_{x_i\in A} d(x_i, \overline{x}),$$

其中$\overline{x}$为A在该距离意义下的重心，或

$$S(A)=1-\min_{x_i\in A} c(x_i,\bar{x}),$$

其中$\bar{x}$为A在该相似系数意义下的相似中心（下同）.聚类函数：同上。

类平均法

离散函数：对任何$A\in F$,

$$S(A)=\sqrt{\max_{B\subset A}\frac{1}{n_1n_2}\sum_{x_i\in B_{x_i}}\sum_{\in A\cap\bar{B}}d^2(x_i,x_j)},$$

或

$$S(A)=\sqrt{\max_{B\subset A}\frac{1}{n_1n_2}\sum_{x_i\in B}\sum_{x_j\in A\cap\bar{B}}(1-c(x_i,x_j))^2},$$

其中n_1和n_2分别为B和$A\cap B$中元素的个数.聚类函数：同上.

离差法

离散函数：对任何$A\in F$,

$$S(A)=\sqrt{\sum_{x_i\in A}d^2(x_i,\bar{x})},$$

或

$$S(A)=\sqrt{\sum_{x_i\in A}(1-c(x_i,\bar{x}))^2}.$$

聚类函数：对任何$P\in L$,

$$Q(P)=\sqrt{\sum_{A_i\in P}S^2(A_i)}.$$

以下建立聚类函数的若干性质：

定理1　设Q为G的聚类函数，则Q对L中的包含关系单调不减.

证　设$P_1\subset P_2\in L$，注意到

$$Q(P_2)=f(Q(P_1),Q(P_2\cap\bar{P}_1)),$$

又由$Q(P_2\cap\bar{P}_1)\geqslant Q(\phi)=0$及f的单调性知，

$$Q(P_2)\geqslant f(Q(P_1),Q(\phi))=Q(P_1).$$

定理2　设$Q_k=Q(P_k)$，其中P_k为A关于Q的k级优聚类，则Q_k对k是单调不增的.

证　对于任意的k，设P_k为A关于Q的K级优聚类，$P_k\cup\{\phi\}$总为A的$k+1$级分割，则有

$$Q_{k+1} \leqslant Q(P_k \cup \{\phi\}) = Q_k.$$

定理3 设Q为G的聚类函数，其中f为严格单调递增的，则A关于Q的优聚类必为A关于Q的最优聚类。

证 设P为A关于Q的优聚类，但不为A关于Q的最优聚类，则必有$P_1 \sqsubset P$，使P_1不为优聚类。设在P_1的同级分割中，P_2为优聚类，则必有$Q(P_2) < Q(P_1)$，从而

$$Q(P) = f(Q(P_1), Q(P \cap \overline{P}_1)) \leqslant f(Q(P_2), Q(P \cap \overline{P}_1)).$$

这与f的严格单调递增相矛盾。故知P必为A关于Q的最优聚类。

由此又知，如果将f定义为和、平方和、平方和的平方根(如离差法)这类严格单调递增函数，则优聚类必为最优聚类。

2 等价关系系统和系统聚类法

如果一种聚类条件能用等价关系描述，则该聚类问题将归结为简单的等价类分割问题。

定义4 设对于任意的λ($a \leqslant \lambda \leqslant b$)，$R(\lambda)$恒为G上的等价关系，且满足

(1) 当$a \leqslant \lambda_1 < \lambda_2 \leqslant b$时，如果$x_i R(\lambda_1) x_j$，则必有$x_i R(\lambda_2) x_j$；

(2) 当且仅当$x_i = x_j$时，有$x_i R(a) x_j$；

(3) 恒有$x_i R(b) x_j$，

则称$R(\lambda)$($a \leqslant \lambda \leqslant b$)为G上的等价关系系统。

用等价关系系统$R(\lambda)$对G施行动态等价类分割是指：先用等价关系$R(a)$使G中的不元素各自形成一类；再逐渐增大λ，只要$x_i R(\lambda) x_j$，便把x_i和x_j所在的类合并成同新类；最后用等价关系$R(b)$使G中所有元素归为一类。

定理4 对G能够施行系统聚类法的充分必要条件是聚类条件能够用G上的等价关系系统描述。

证

充分性：设聚类条件已用等价关系系统$R(\lambda)$描述，用$R(\lambda)$对G施行动态等价类分割，这就是对G施行的系统聚类法。

必要性：设对G可以施行系统聚类法，分割依次合并的过程为：H_0，H，H_2，…，H_k，其中H_0使G中不同元素各自形成一类，H_k使G中所有元素归为一类。把分割H_i($i = 0, 1 \cdots, k-1$)描述为等价关系$R(\lambda)$($\frac{i}{k+1} \leqslant \lambda < \frac{i+1}{k+1}$)，把$H_k$描述为等价关系$R(1)$，则$R(\lambda)$($0 \leqslant \lambda \leqslant 1$)即为G上的等价关系系统。

特别地，我们在G上建立如下两种等价关系系统：

$R_1(\lambda)$($0 \leqslant \lambda \leqslant 1$)定义为

(1) 若$c(x_i, x_j) \geqslant 1 - \lambda$，则$x_i R_1(\lambda) x_j$；

(2) 若$x_i R_1(\lambda) x_k$，$x_k R_1(\lambda) x_j$，则$x_i R_1(\lambda) x_j$。

$R_2(\lambda)$($0 \leqslant \lambda \leqslant d_0$，$d_0$足够大)定义为

(1) 若$d(x_i, x_j) \leqslant \lambda$，则$x_i R_2(\lambda) x_j$；

（2） 若 $x_i R_2(\lambda) x_k$，$x_k R_2(\lambda) x_j$，则 $x_i R_2(\lambda) x_j$。

这两种等价关系系统恰为对最短距离法的描述，因此，最短距离法恒可以用系统聚类法实现.不仅如此，而且每一步骤所得，既为优聚类，也为最优聚类.这是因为任取 $A \in \mathbf{F}$ 和 k，在把 A 分割为 k 个类的等价关系中，λ 的最小值必然存在，故知为优聚类.又由 A 和 k 的任意性知，又必为最优聚类.

聚类分析中的传递闭包法，是将 G 的相似系数矩阵或距离矩阵逐步改造成传递闭包，然后利用截矩阵进行聚类的方法[2].

定理5

（1） 设 G 的相似系数矩阵为 $C=(c(x_i, x_j))$，其传递闭包为 $C^*=(c^*(x_i, x_j))$ 则对任意的 λ（$0 \leqslant \lambda \leqslant 1$），$c^*(x_i, x_j) \geqslant 1-\lambda$ 的充分必要条件是 $x_i R_1(\lambda) x_j$。

（2） 设 G 的距离矩阵为 $D=(d(x_i, x_j))$，其传递闭包为 $D^*=(d^*(x_i, x_j))$ 则对任意的 λ（$0 \leqslant \lambda \leqslant d_0$），$d^*(x_i, x_j) \leqslant \lambda$ 的充分必要条件是 $x_i R_2(\lambda) x_j$。

证　仅证（1），因类似，（2）证明略。

充分性：设 $x_i R_1(\lambda) x_j$，则必有路径 l：$x_i = y_1, y_2 \cdots, y_m = x_j$，使当 $1 \leqslant k < m$ 时，恒有 $c(y_k, y_{k+1}) \geqslant 1-\lambda$.由传递闭包的构造知

$$c^*(x_i, x_j) \geqslant \min_{1 \leqslant k < m} c(y_k, y_{k+1}) \geqslant 1-\lambda.$$

必要性：设 $c^*(x_i, x_j) \geqslant 1-\lambda$，由传递闭包的构造知，必有路径 l：$x_i = y_1, y_2, \cdots, y_m = x_j$，使 $c^*(x_i, x_j) = \min_{1 \leqslant k < m} c(y_k, y_{k+1})$.故知当 $1 \leqslant k < m$ 时恒有 $c(y_k, y_{k+1}) \geqslant 1-\lambda$，从而 $x_i R_1(\lambda) x_j$。

定理表明，传递闭包法的实质是建立 G 上的等价关系系统 $R_1(\lambda)$ 或 $R_2(x)$.因此，可以直接根据相似系数矩阵或距离矩阵，构造出相应的等价关系系统，从而用系统聚类法实现动态的最优聚类.

3　全序样本集的递归聚类法

以上在优聚类的定义和等价关系系统的构造中，已经多次利用递归思想.这里，我们仍用递归思想，推广著名的Fisher算法，对全序样本集给出普遍适用的递归聚类法.

设 G 为全序样本集，序关系为 O.A 的分割 $P=\{A_1, A_2, \cdots, A_k\}$ 称为全序分割是指：当 $x_i \in A_m$，$x_j \in A_m$，$A_m \in P$ 时，若 $x O x_t$，$x_t O x_t$，$x_t \in A$，则 $x_t \in A_m$ $P_1 \subset P$ 称为 P 的全序子集是指：当 $x_i \in A_i \in P_1$，$x_j \in A_j \in P_1$ 时，若 $x_i O x_t$，$x_t O x_j$，$x_t \in A_t \in P$，则 $A_t \in P_1$。

可以类似地给出全序优聚类的定义：设 G 的离散函数为 S，聚类函数为 Q.如果 A 的全序分割 P 在同级全序分割中使 Q 达到最小值，则称 P 为 A 关于 Q 的全序优聚类.特别地，如果 P 的全序子集均为优聚类，则称 P 为 A 关于 Q 的全序最优聚类.

有限全序样本集关于聚类函数的全序优聚类恒存在。

设 G 的子集 A 中全部元素按序关系 O 排列的顺序为 $x_1, x_2, x_3, \cdots, x_n$.令

$$G_{i,j}=\{x_t \mid i\leqslant t\leqslant j\}.$$

构造矩阵 $S=(s_{i,j})$，其中 $s_{i,j}$ 定义为

$$s_{i,j}=S(G_{i,j}).$$

构造矩阵 $Q=(q_{i,j})$，其中 $q_{i,j}$ 定义为

(1) $q_{1,j}=s_{1,j}$,

(2) $q_{i,j}=\min\limits_{i-1\leqslant t<j} f(q_{i-1,t}, s_{t+1,j})$,

(3) 其余元素任意.

构造矩阵 $M=(m_{i,j})$，其中 $m_{i,j}$ 定义为

(1) $m_{1,j}=1$,

(2) 当 $q_{i,j}=f(q_{i-t,t}, s_{t+1,j})$ 时，$m_{i,j}=t+1$,

(3) 其余元素任意.

对于给定的 $k\leqslant n$，令

$$\begin{aligned} &j_k=n, && i_k=m_{k,n},\\ &j_{k-1}=i_k-1, && i_{k-1}=m_{k-1,j_{k-1}},\\ &\cdots\cdots && \cdots\cdots\\ &j_2=i_3-1, && i_2=m_{2,j_2},\\ &j_1=i_2-1, && i_1=1. \end{aligned}$$

由此构成A的全序k分割

$$P=\{A_1, A_2, \cdots, A_k\}$$

其中 $A_1=G_{i_1,j_1}$, $A_2=G_{i_2,j_2}$, …, $A_k=G_{i_k,j_k}$.按照这种程序进行全序分割的方法，称为全序集关于Q的递归聚类法.

定理6 全序集关于Q的递归聚类法得到关于Q的全序优聚类.

证

(1) 显然，对任意的 $j\leqslant n$，$\{G_{1,j}\}$ 为 $G_{1,j}$ 的一级全序优聚类.

(2) 设对任意的 $j\leqslant n$，递归聚类法所得到的 $P_k(j)$ 为 $G_{1,j}$ 的k级全序优聚类，从而 $q_{k,j}=Q(P_k(j))$.如果有 $j\leqslant n$，使递归聚类法所得到的 $P_{k+1}(j)$ 不为 $G_{1,j}$ 的 $k+1$ 级全序优聚类.设 $G_{1,j}$ 的 $k+1$ 级全序优聚类为

$$P=P_1\cup\{G_{i,j}\},$$

其中 P_1 为 $G_{1,i-1}$ 的k级分割，因而

$$Q(P)=f(Q(P_1), S_{i,j})<Q(P_{k+1}(j))=q_{k+1,j}.$$

由于

$$q_{k+1,j}=\min\limits_{k\leqslant t<j} f(q_{k,t}, s_{t+1,j}),$$

因而

$$f(Q(P_1), s_{1,j}) < f(q_k, {}_{i-1}, s_{i,j}).$$

注意到 f 的单调非减性，故必有

$$Q(P_1) < q_{k,i-1} = Q(P_k(i-1)),$$

从而，$P_k(i-1)$不为$G_{1,i-1}$的 k 级全序优聚类.此与假设矛盾，故知对任意的 $j \leqslant n$，递归聚类法所得到的$P_{k+1}(j)$仍为 k + 1 级全序优聚类.

综上所述，全序集的递归聚类法恒得到全序优聚类.

定理7　如果 f 是严格单调递增的，则全序集关于Q的递归聚类法得到关于 Q 的全序最优聚类.

证　设有 k，$j \leqslant n$，使递归聚类法所得到的$P_k(j)$为$G_{1,j}$的 k 级全序优聚类，但不为 k 级全序最优聚类，即有$P_k(j)$的全序子集P为G_{i_0,j_0}的全序分割，但不为全序优聚类.设G_{i_0,j_0}的全序优聚类为P_0,则有$Q(P_0) < Q(P)$.又设作为G_{1,j_0}全序分割的$P_k(j)$的全序子集为P_1，作为G_{1,i_0-1}全序分割的$P_k(j)$的全序子集为P_2，则$P_1 = P \cup P_2$，且由定理 6 知，P_1为G_{1,j_0}的全序优聚类，因而

$$Q(P_1) \leqslant Q(P_0 \cup P_2),$$

此即

$$f(Q(P), Q(P_2)) \leqslant f(Q(P_0), Q(P_2)).$$

这与 f 的严格单调递增相矛盾.故知$P_k(j)$必为$G_{1,j}$的 k 级全序最优聚类.

参 考 文 献

1　张尧庭、方开泰. 多元统计分析引论. 科学出版社，1982

2　汪培庄. 模糊集合论及其应用. 上海科技出版社，1983

SOME PROBLEMS CONCERNING CLUSTER ANALYSIS

Wang Naixin

(*Computer Center, Northwestern Agricultural University*)

Abstract

In this paper we introduced a unified definition of well and optimum cluster, summarized many clustering methods, established some properties of clustering function, proved the equivalence of hierachical equivalent relations with hierachical clustering methods and extended fisher algorithm, proved that ordered recursive clustering must yield well ordered cluster, and some sufficient conditions of optimum clustering were obtained.

keY words: clustering function; well cluster; optimum cluster; ierachical equivalent relations; recursive clustering methods