阮敬 博士

首都经济贸易大学研究生院 副院长
首都经济贸易大学统计学院 教授

© ruanjing@msn.com

聚类分析

- “物以类聚、人以群分”往往被人们视为自然的法则。正是由于不同现象之间客观存在的共性，使得大千世界芸芸众生有了界限的划分和质的区别，而呈现出五花八门的景象。
- 在事物分类思想上最为瞩目的是生物分类学的发展， 这也成为统计分类发展的主要动力。希腊时期亚里士多德仅描述了 500 个物种， 17 世纪后， 人们知道约 6000 种植物， 而仅仅 100年后，植物学家又发现了 12000 个新种。对生物物种进行科学的分类变得极为迫切。因此有了林奈把自然界分为 3 界：即动物界、植物界和矿物界，并提出了纲、目、属、种的分类概念，人们可以依照各门类物种的典型特征，把新发现的物种归类至现有的门类当中。
- 近代统计分析中的聚类和判别分析受到了生物分类学的影响， 现实生活中需要对复杂的对象依据一定的标准进行分类，有了既定的类别之后，还可涉及到对事物进行归类。因而有了本章所要介绍的聚类分析及下一章将要介绍的判别分析。

聚类分析的基本原理

- 人们根据事物现象的一个指标或某一个方面， 可以很容易进行分类活动。 如按照收入指标把全社会人群划分为高、中、低 3 类，学生考试成绩划分为及格、不及格两类等。在进行归类时，只需考查新加入的对象在某个指标上的表现是否符合特定类别即可。
- 实际上， 需考察的事物或对象往往不是单一指标这么简单， 很可能是通过许多侧面或许多指标来进行综合考察。如按照经济发展、教育水平、面积大小、人口等诸多方面对我国地市级以上城市进行分类； 学生凭考试成绩、 社会实践、 思想品德等方面划分奖学金的等级等。这些指标在反映事物特征的作用、量纲、紧密关系等方面可能有所不同，因此很难再按照单一指标分类的原则进行分类和归类了，需要考虑多元统计分析的方法进行分类和归类。
- 多元统计分析中的聚类分析方法（Clustering Analysis）既可以对样本进行分类（记为 Q型分类） ，也可以对反映事物特征的指标或变量（记为 R 型分类）进行分类。两种分类是对等的，在算法上没有任何区别，本书主要以 Q 型分类为例进行详细讲解，在第 16.2.3 小节中对 R 型分类进行简单介绍。
- “近朱者赤， 近墨者黑” 。 人们往往可根据事物之间的距离远近或相似程度来判定类别。个体与个体之间的距离越近，其相似性可能也越大，是同类的可能性越大，聚在一起形成类别的可能性也就越大。因此就有了聚类分析的基本原则。

聚类分析的基本原则

- 首先考虑在没有进行聚类之前， 所有参加聚类过程的个体没有归入任何类别， 即对于每个个体而言，其独树一帜，自成一类。
- 有了一定的分类原则之后，人们可以根据个体与个体之间的距离大小或长短进行聚类。如首先把最近的个体聚为同类， 然后再根据最短距离继续扩大类别所涵盖的范围， 直到所有个体都聚为 1 个大类为止。 整个聚类过程就如同生活在地球上的人一样， 首先每个人都是自成一类，然后有了人种的区分，最后所有人都可以归集到“人类”这个类别当中，即所有人都是一类。在数据分析过程中，人们通常把类似上述的聚类过程称之为“系统聚类” 。
- 而聚类过程所依据的距离主要有明氏距离、 马氏距离等几大类。 那么究竟什么是距离呢？设样本数据可以用如下矩阵形式表示：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad 记为 X=\left\{x_{ij}\right\}_{n\times p}$$

聚类分析的基本原则

- 设d_{ij}表示第 i 个样本与第 j 个样本之间的距离。如果d_{ij}满足以下 4 个条件，则称其为“距离”：
  - $d_{ij} \geq 0$,对一切i, j;
  - $d_{ij} = 0$.等价于i, j:
  - $d_{ij} = d_{ji}$,对一切i, j:
  - $d_{ij} \leq d_{ik} + d_{kj}$,对一切$i, j, k$。
- 第 1 个条件表明聚类分析中的距离是非负的；第 2 个条件表明个体自身与自身的距离为0；第 3 个条件表明距离的对等性，即 A 和 B 之间的距离与 B 和 A 之间的距离是一致的；最后一个条件表明两点之间直线距离是最小的。
- 明氏距离是最常用的距离度量方法之一，其计算公式为：

$$d_{ij}(q) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{q}\right)^{1/q}$$

聚类分析的基本原则

- 明氏距离有如下几种典型情况：
  - 当 q=1 时：$d_{ij}(1)=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$，称为“绝对距离”；
  - 当 q=2 时：$d_{ij}(2)=\left(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{2}\right)^{1/2}$，称为“欧氏距离”；
  - 当 q=∞ 时：$d_{ij}(\infty)=\max_{1<k<n}|x_{ik}-x_{jk}|$，称为“车比雪夫距离”。
- 但是明氏距离的大小与个体指标的观测单位有关，没有考虑指标之间的相关性。为克服明氏距离的缺点，可以考虑采用马氏距离进行改进。马氏距离是由协方差矩阵计算出来的相对距离，具体计算公式如下：

$$d_{ij}=(X_i-X_j)'\Sigma^{-1}(X_i-X_j)$$

- 有了距离的定义，就可以在对事物现象的分类过程中，依据如前所述的距离最小原则来进行聚类分析。

聚类分析的基本原则

- 除了最短距离原则进行分类之外，还可以采用相关系数、相似系数、匹配系数等指标来衡量个体之间的相似性，以此为依据进行分类。在分类的过程当中，为了便于分析，还应当注意如下 3 个重要原则：
  - 同质性原则：即同一类中的个体之间有较大的相似性；
  - 互斥性原则：即不同类中的个体差异很大；
  - 完备性原则：每个个体在同一次分类过程中，能且只能分在一个类别当中。
- 同质性原则保证了类别之内个体特征的共性； 互斥性原则保证了类别之间的差异性； 而完备性原则则说明了每一个个体应当包含在所进行的分类当中， 同时每一个个体不能同时被分在不同的类别当中。
- 实际应用中，以最短距离原则进行的“系统聚类”比较常用。本书以此为依据进行详细的聚类过程介绍。

单一指标的系统聚类过程

- 为了更好的理解最短距离分类的基本原理，首先考察最简单的单一指标情况。
- 例16-1：为考察公司的经营业绩并对其进行分类， 可从它们的年盈利额来归类。 具体数据如表 16-1 所示。

表 16-1　公司年盈利额数据

公司	年盈利（十万元）
甲	1
乙	3
丙	9
丁	14

- 为直观的分析，把表　16-1　的数据排列在数轴上进行分析，用数轴上的点来代表各个公司相应的财务指标，如图 16-1 所示。

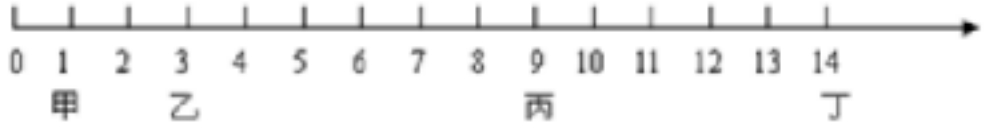


图 16-1　单一指标的分类过程（一）

单一指标的系统聚类过程

- 直观的看，哪两个点距离最近呢？自然是甲和乙，它们的距离是 2=3-1。如果按照最短距离的原则来归类， 首先就要把甲乙两点聚合成一类， 以后为了方便， 称之为 “类 （甲乙） ”。于是就把它们归为一类，如图 16-2 所示。

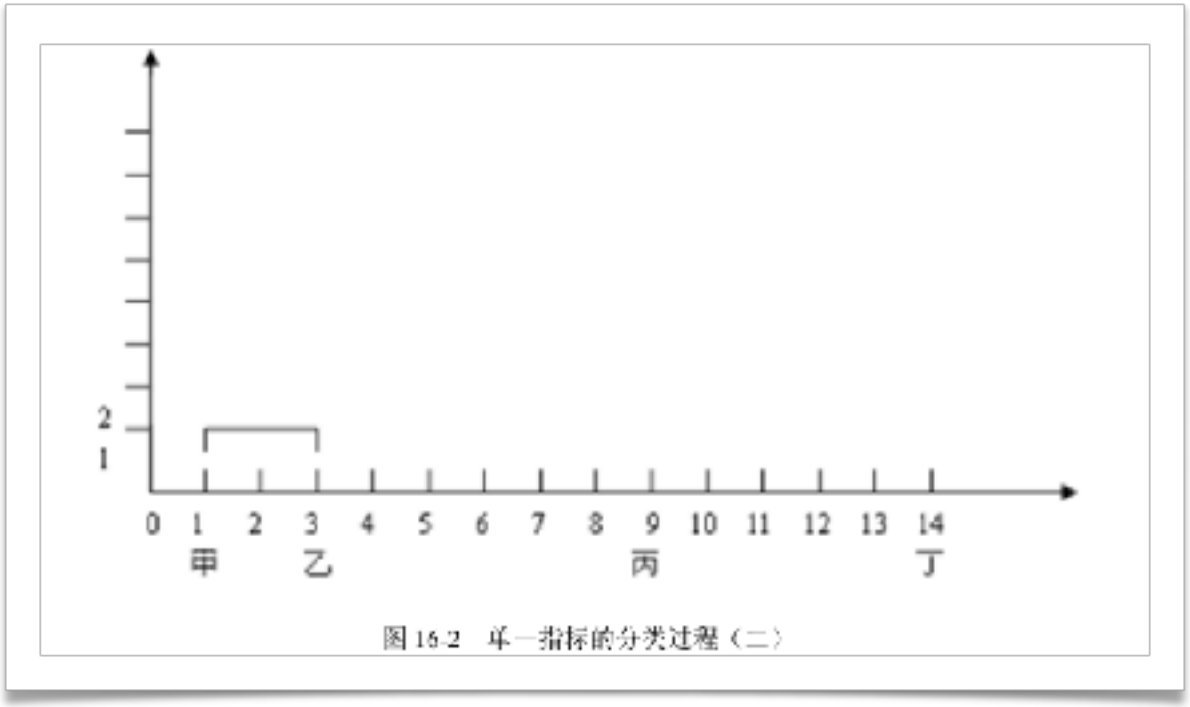


图 16-2 单一指标的分类过程（二）

- 在图 16-2 所示的分类过程中， 可增加一个维度 （即增加了纵轴） 表示两点之间的距离。如甲和乙之间的距离为 2，用横线把代表甲和乙的点连线起来，连线的高度便是纵轴所代表的“2” 。

单一指标的系统聚类过程

- 继续观察，发现剩下的点里，丙丁的距离最近，为 5=14-9。因此把二者聚为“类（丙丁） ”。于是就把它们归为一类，如图 16-3 所示。

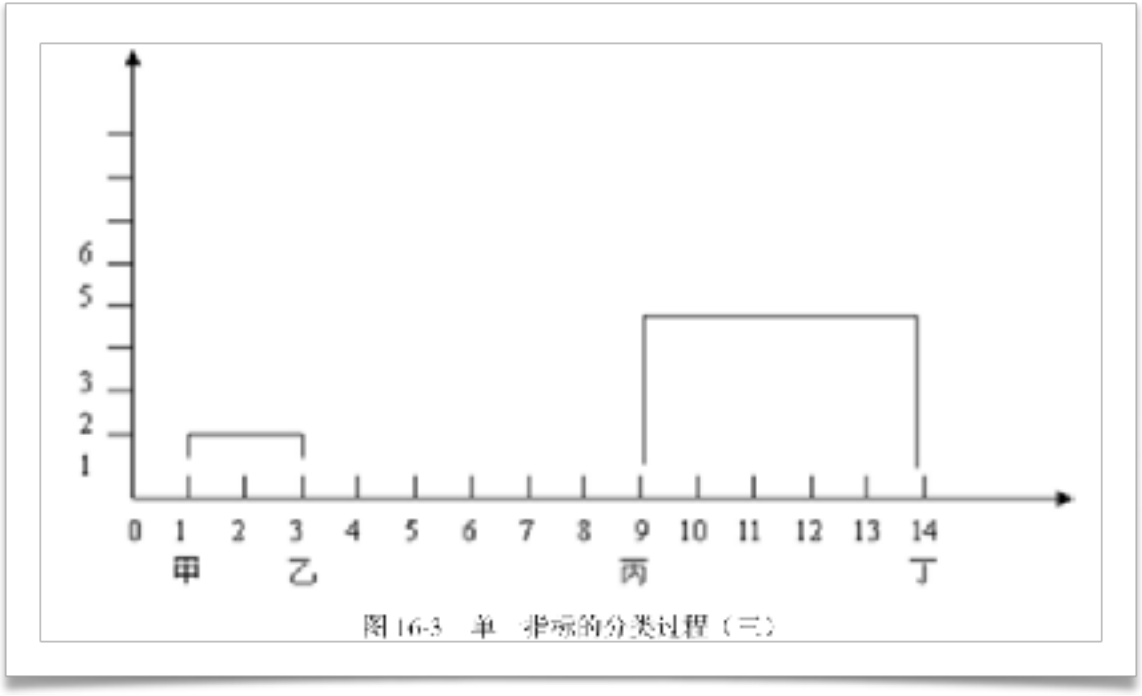


图 16-3 单一非标的分类过程（三）

- 这样，该分类过程就剩下两类，“类（甲乙） ”和“类（丙丁） ”，在这两类相互聚合的过程中可能有 4 个距离，即甲丙、甲丁、乙丙、乙丁，其距离分别是：8=9-1、13=14-1、6=9-3和 11=14-3。

单一指标的系统聚类过程

- 如果按照最短距离的原则来归类， 那么上述的乙丙之间是最近距离， 因此它就代表了“类（甲乙） ”和“类（丙丁） ”的距离。至此，这样整个聚类过程就完成了，如图 16-4 所示。

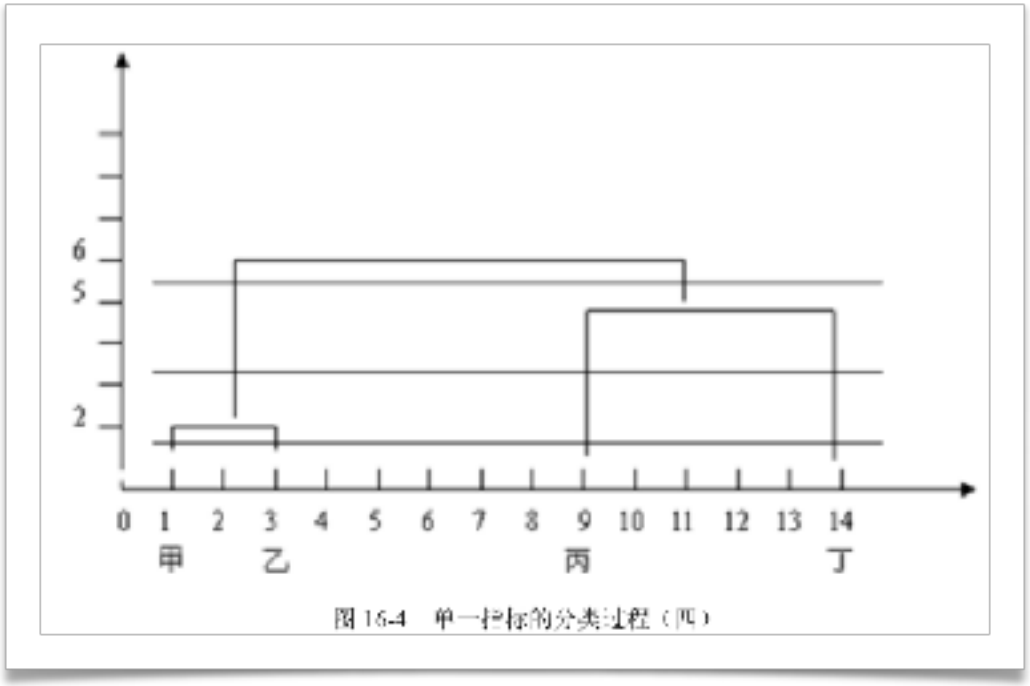


图 16-4 单一指标的分类过程（四）

- 上述过程可以形象的解答什么叫“聚类” 。

单一指标的系统聚类过程

- 聚类的结果（如图 16-4 所示）如同一棵大树的根系，最上面是根，再往下就是根的分叉，这种分叉一直分下去，最后就是这一组事物的各个个体了。通俗的说，根系的顶端，表明任何一组事物最终均可聚为一类； 反之从根系的末端来看， 如果归类达到最详细和最具体，那么各个个体自成一类，即每个个体自身都可看成是一个类。
- 而描述上述分类过程的图形可以称之为系统聚类过程中的“谱系聚类图”，简称“谱系图”，因而系统聚类又可以称之为“谱系聚类”。
- 在谱系图中，存在着若干层次的类。如图 16-4 所示，从上往下看，有 3 个层次。即在第 1 层次的水平上可以分为 2 类（在距离 5-6 之间的任意位置，画一条直线，与整个根系有2 个交点）；第 2 层次上可分为 3 类（在距离 2-5 之间的任意位置，画一条直线，与根系有 3个交点）；以此类推，第 3 层次上可以分 4 个类。这样，整个系统聚类的过程可以按照不同层次的类别对个体进行不同的聚类， 每一层次上的聚类相对其他层次而言不是独立的， 高层次的类别是在低层次类别基础上完成的。因此，从这个意义上来说，系统聚类也可以称之为“层次聚类”方法。

多指标的系统聚类过程

- 当要根据多个特征或指标对所反映的事物现象进行分类时， 其过程相对较复杂。 如对全国的大学进行分类，应当同时综合考虑学生生源、学术声誉、师资力量、学科门类齐全程度等各方面的情况。本书以例 16-2 的 2 个指标分类为例，来描述多指标的系统聚类过程。
- 例16-2：为考察投资者的盈利能力并对其进行分类， 可从资金的投入与回报两个方面来进行考察。具体数据如表16-2 所示。

表 16-2　投资者资金投入与回报情况

投资者	资金投入（万元）	回报（万元）
A	35	60
B	15	40
C	30	5
D	80	8
E	90	35

- 现在要根据“资金投入”和“回报”两个指标对 A、B、C、D、E 5 个投资者进行分门别类。根据最短距离原则，本例的问题实际上就是以二维空间中的最短距离来进行聚类。

多指标的系统聚类过程

- 首先用一个二维坐标轴分别表示 “资金投入” 和 “汇报” 两个指标， 根据对应的指标值，把 5 个投资者所代表的点描绘在如图 16-5 所示的二维坐标轴上。

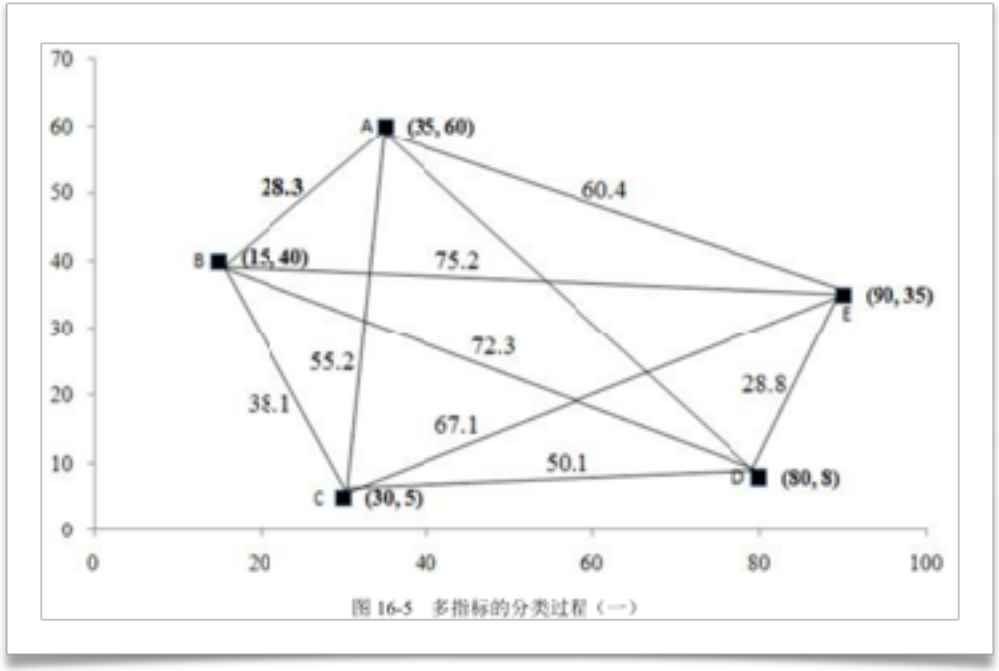


图 16-5 多指标的分类过程（一）

- 本例采用欧氏距离进行距离的计算，如 A、B 两点之间的欧式距离为：

$$d_{AB} = \sqrt{(35-15)^2 + (60-40)^2} = 28.3$$

多指标的系统聚类过程

- 其余各两点之间的欧氏距离均可计算出来并标注在图 16-5 上。从图 16-5 中可以直观看到 AB 的距离最近（28.3）。因此，按照最短距离原则，可以把 A 和 B 聚为一类，记为类 4，如图 16-6 所示。

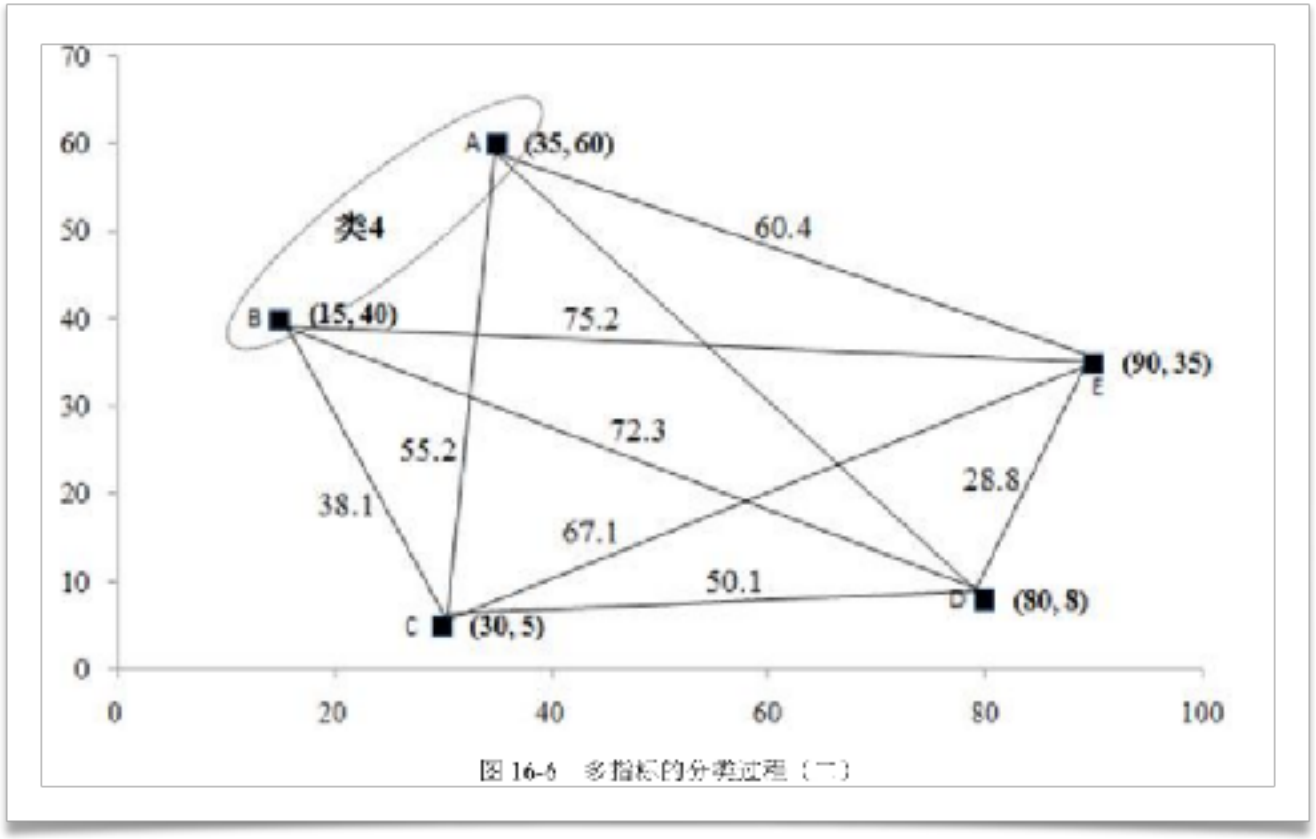


图 16-6 多指标的分类过程（一）

多指标的系统聚类过程

- 如果类别与类别之间考虑的是以最短距离原则来聚类， 类 4 与外部的距离中 （类 4 含有A、B 两个点，故其与外部的距离有两个） ，最短者才是有意义的，因此在图 16-6 中可以把不是最短距离的连线去掉，得到如图 16-7 所示的分类过程。

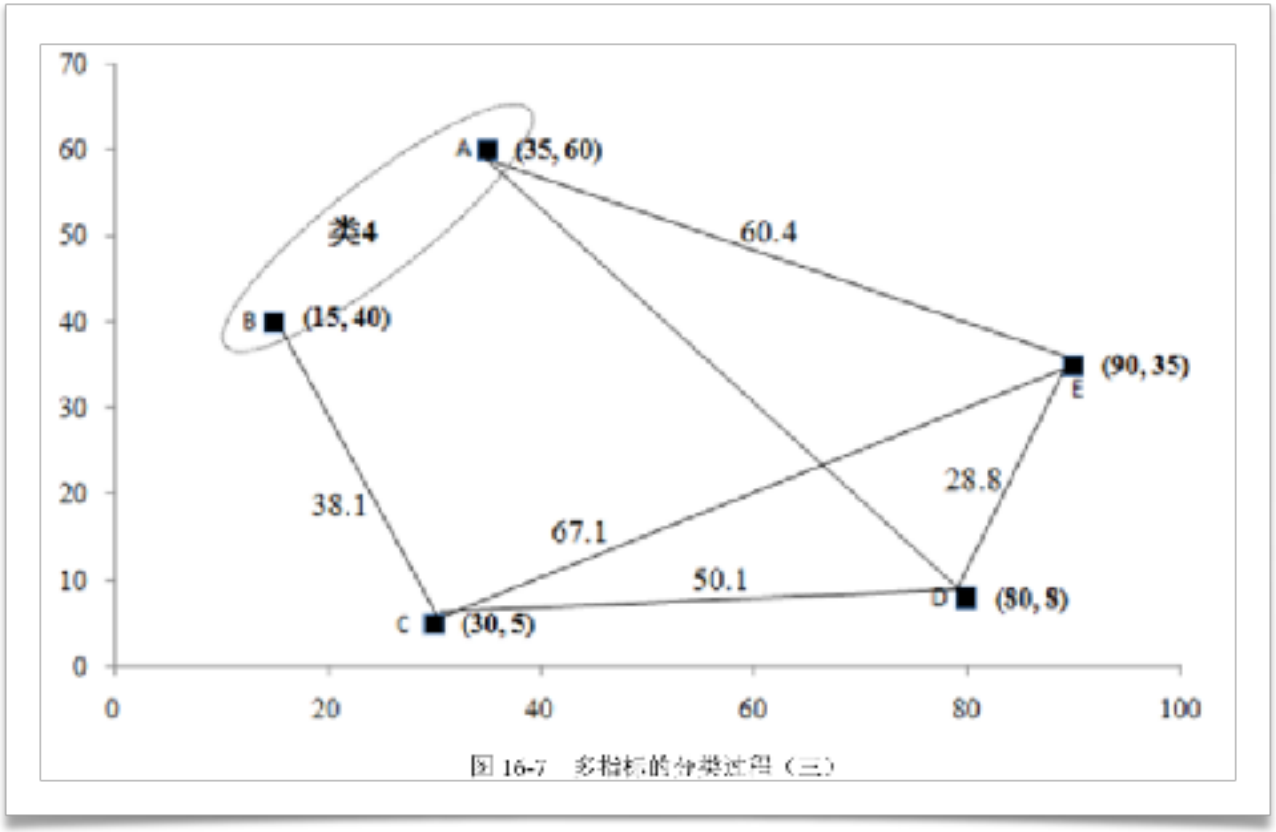


图 16-7 多指标的分类过程（三）

多指标的系统聚类过程

- 显然，在图 16-7 中剩下的所有距离中，最短距离就是 DE=28.8，因此把 D 和 E 聚为一类，记为类 3。把 D 和 E 合并之后，可得到如图 16-8 所示的过程。

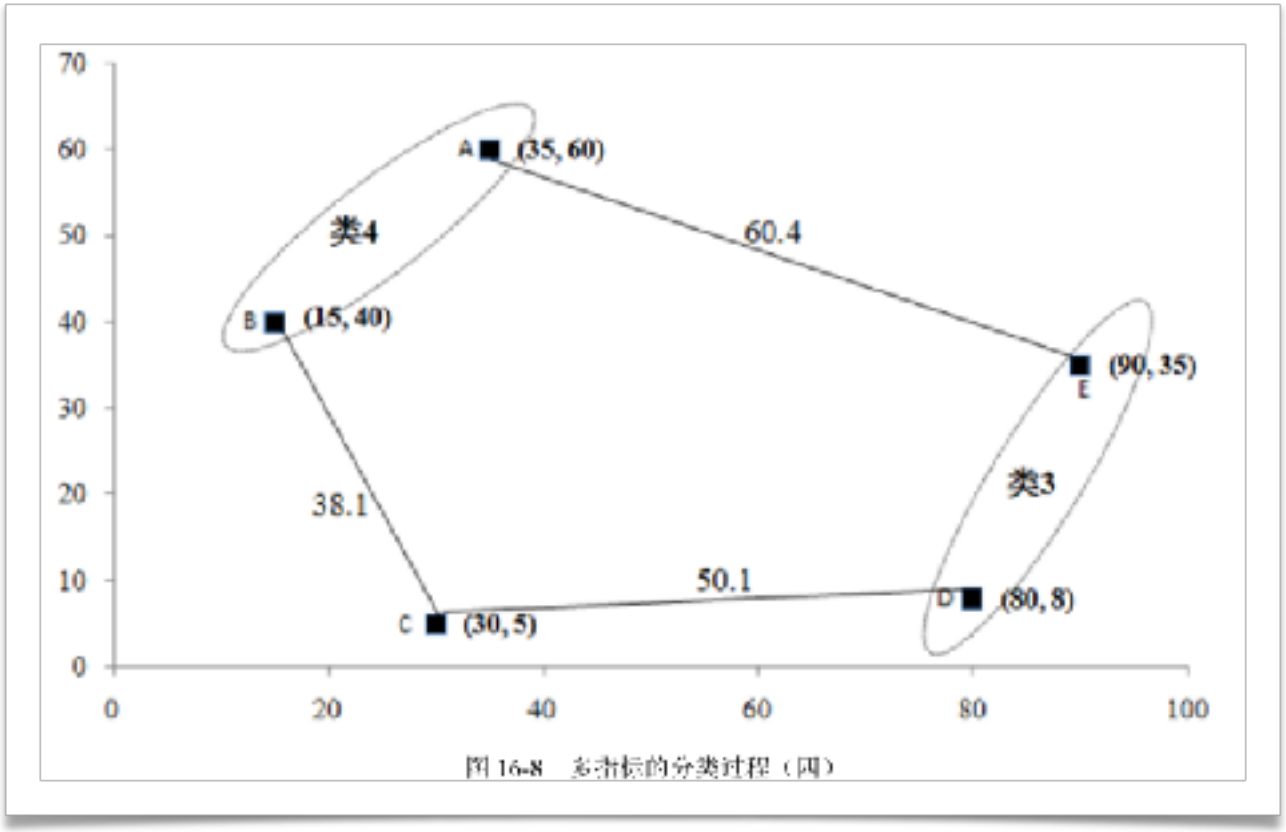


图 16-8 多指标的分类过程（四）

多指标的系统聚类过程

- 继续观测在图 16-8 中的所有距离，就是 C 到类 4 的距离是最短的（38.1），于是又可以把 C 和类 4 聚为一类，记为类 2。按照上述的原则同样可以得到如图 16-9 所示的过程。

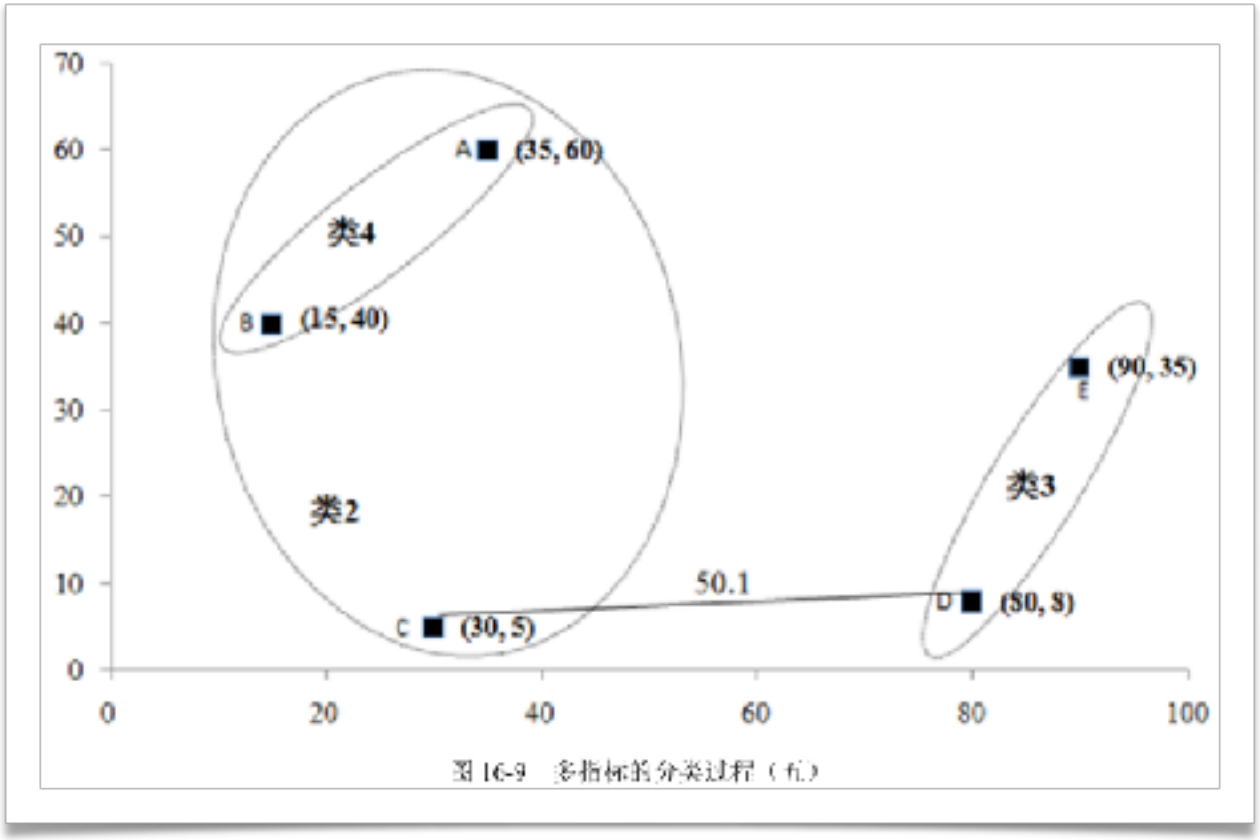


图 16-9 多指标的分类过程（五）

多指标的系统聚类过程

- 在图 16-9 中，就剩下类 2 与类 3 的一个距离 CD（50.1） 。因此，把类 2 与类 3 合并成一类。至此，所有的样本点都已经处于一定的类别当中，所有的样本归为一大类，系统聚类过程结束。整个分类过程如图 16-10 所示。

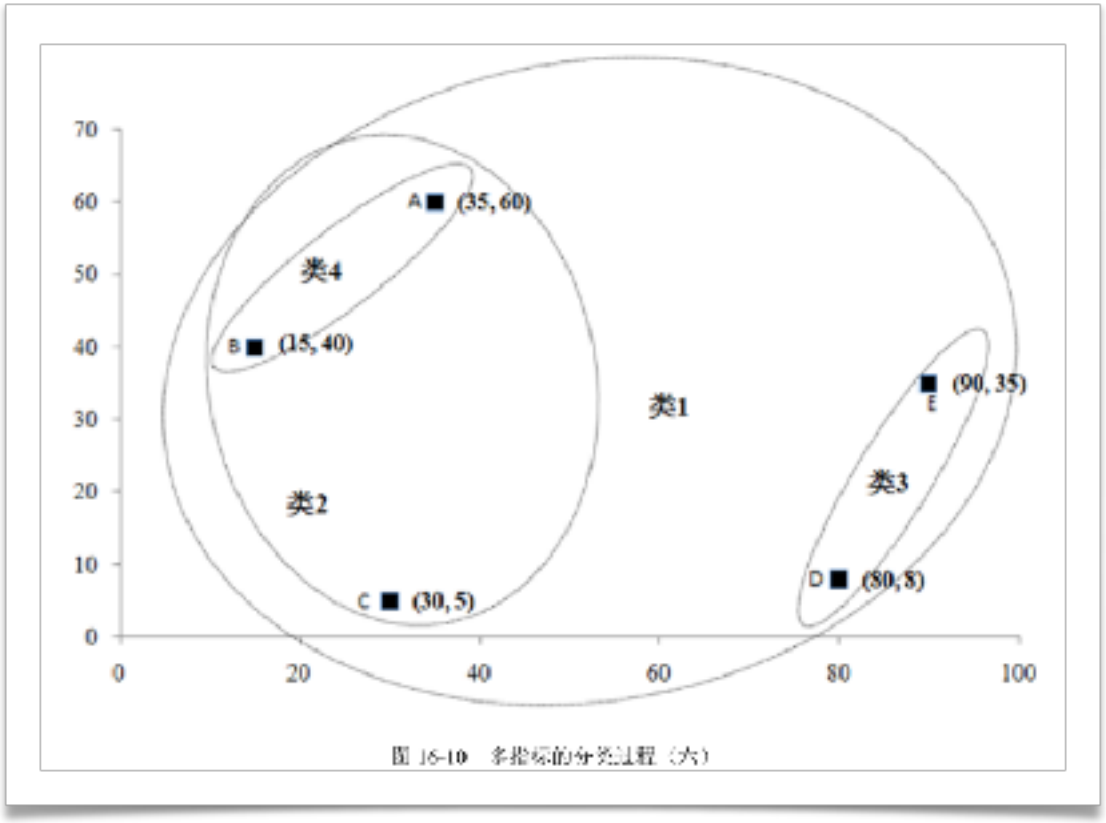


图 16-10 多指标的分类过程（六）

多指标的系统聚类过程

- 上述分类过程可以用表格的形式来表现。首先，把各样本间的距离列示在表 16-3 中，该表以对角线对称，为简单起见，省略掉重复数值。

表 16-3　样本点之间的距离（一）

	A	B	C	D	E
A					
B	28.3				
C	38.1	55.2			
D	72.3	68.7	50.1		
E	75.2	60.4	67.1	28.8	

多指标的系统聚类过程

- 观察表 16-3 中各样本点之间的距离，依据最小距离原则找出最小的距离并用灰色标注出来。显然，A、B 要聚成一类，于是表 16-3 可处理成如表 16-4 所示的形式：

表 16-4　样本点之间的距离（二）

	AB（4）		C	D	E
AB（4）					
C	38.1	55.2			
D	72.3	68.7	50.1		
E	75.2	60.4	67.1	28.8	

多指标的系统聚类过程

- 在 A、B 聚为类 4 后，其内部的距离失去意义，但其与外部的距离还有意义。按照最短距离原则，这些外部的距离中只有最短者被保留，因此表 16-4 中不是最短的距离被删除并整理如表 16-5 所示。

表 16-5　样本点之间的距离（三）

	AB（4）	C	D	E
AB（4）				
D	68.7	50.1		
E	60.4	67.4	28.8	

多指标的系统聚类过程

- 在表 16-5 中列示的所有距离中，D 和 E 的距离最小（28.8），把 D、E 聚为类 3。同理，删除掉不是最短的距离，如表 16-6 所示。

表 16-6　样本点之间的距离（四）

	AB（4）	C	DE（3）
AB（4）			
C	38.1		
DE（3）	60.4	50.1	

多指标的系统聚类过程

- 观察表 16-6，找出 AB（4）与 C 的距离最短（38.1），因此把二者归为类 2，并划掉非最短距离，得到表 16-7：

表 16-7 样本点之间的距离（五）

	ABC（2）	DE（3）
ABC（2）		
DE（3）		50.1

多指标的系统聚类过程

- 最后把 ABC（2）和 DE（3）聚成一类，如表 16-8 所示。

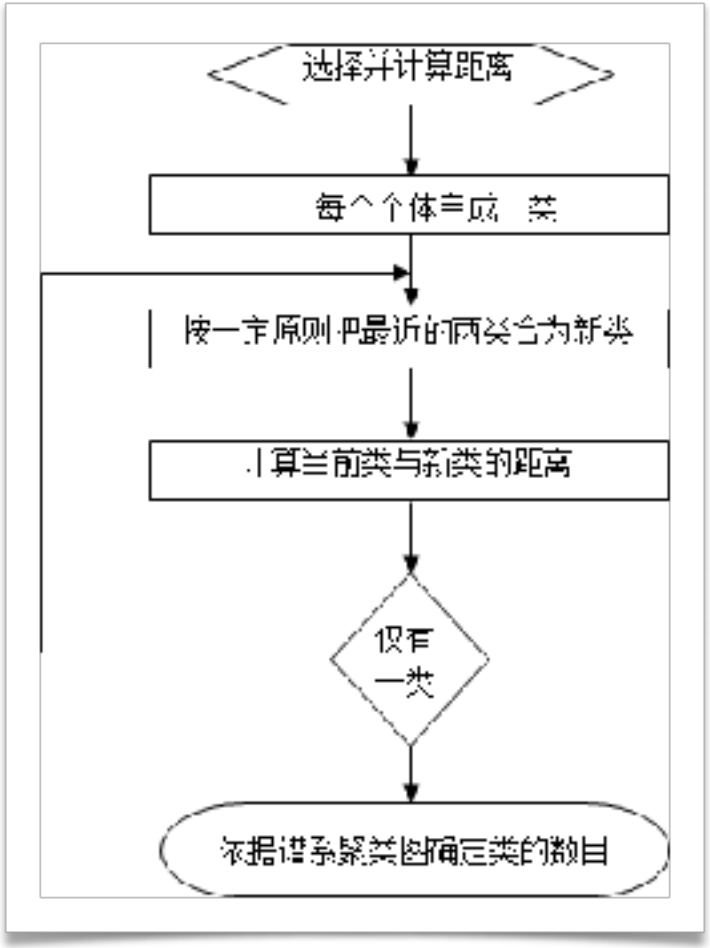
表 16-8　样本点之间的距离（六）

	ABCED（1）
ABCED（1）	

- 至此，系统聚类过程结束。

多指标的系统聚类过程

- 对于根据两个以上指标的分类过程，也类似于该过程。综合上述分类过程，本书把系统聚类的一般流程总结绘制成流程图，如图 16-11 所示。



聚类分析的步骤和过程—系统聚类

- 对于本书所介绍的聚类分析内容，在 SAS 系统中可通过 CLUSTER、FASTCLUS、VARCLUS 和 TREE 等 4 个过程来实现。
- 进行聚类分析时，可根据图 16-11 所示的流程图进行系统聚类分析。在对例 16-1 和例16-2 的分析过程中，当类别中含有多个样本点时，就会涉及到类别与类别之间的距离定义方法（因为不同类别的每两个样本点之间都会有距离） 。除了上述例题中所使用的最短距离法来确定类间距离之外，SAS 系统还提供了 11 种系统聚类过程中确定类别与类别之间距离的方法。本书主要介绍社会经济研究中如下几种最为常用的方法。

聚类分析的步骤和过程—系统聚类

- 1. 最短距离法（SINgle linkage ）
- 最短距离法又称为“单连接聚类法” 。如果有两类G_P和G_q聚成为新类G_n，在最短距离法中新类G_n与其它的任意类G_k之间的距离或相似系数由下列公式决定：

$$D_{\mathrm{kn}} = Min(D_{kp}, D_{kq})$$

- 其中，D_{kp}和D_{kq}是用来衡量原有类别G_P和G_q中各样本点与任意类G_k中各样本点的点间距离。
- 即如果新类与其它类别之间存在多个点与点之间的距离， 则取这些距离当中最小者作为两类的距离， 即在进行聚类的过程中应以最小点间距离作为并类的依据， 其具体并类过程与第 16.1 小节的分析过程相同。

聚类分析的步骤和过程—系统聚类

- 2. 最长距离法（COMplete method ）
- 该种方法也称之为“完全连接法” 。如果有两类G_P和G_q合并为新类G_n，在最长距离法中新类G_n与其它的任意类G_k之间的距离或相似系数由下列公式决定：

$$D_{\mathrm{kn}} = Max(D_{kp}, D_{kq})$$

- 即如果新类与其它类别之间存在多个点间距离， 则取这些距离当中最大者作为两类的距离。
- 在进行分类的过程中， 首先以最小距离原则把最近的两个样本合并为一个新类， 其余各样本自身自成一类。 则刚合并的新类与其它类别之间按最长距离法确定之后， 再按照最小距离原则把类别之间距离最小的合并为一个新类，以此类推，直至把所有样本归为一类。