

Comparative Study of Cluster Analysis Method

Tingting Chen[†]

School of Science, China University of Geosciences (Beijing), Beijing 100083, China

[†]Email:954302164@qq.com

Abstract

Cluster analysis is a kind of modern multivariate statistical analysis method to research "Things of one kind come together". And cluster analysis methods have developed rapidly, and made fruitful applications in economics, management, geological exploration, weather forecast, taxonomy, archaeology, medicine, psychology and national and regional standards development and so on. First, learning the relevant knowledge of cluster analysis, using the different kinds of system clustering method to classify the instance data through the application of SPSS software, and determining the suitable number of classes by using the threshold T, observing scatter plots and using statistics, to research and compare the different kinds of system clustering method. Finally come to the conclusion that: The deficiency of "Given a threshold value T" is its subjectivity; the method "Scatter plot" is more intuitive, and the efficiency may be better than the normal clustering method; "Using statistics" is often more clear. In the effect of clustering methods, in many applications, the clustering effect of the average linkage method and the Ward method is relatively good.

Keywords: Cluster Analysis; Classification; System Clustering Method; SPSS

聚类分析方法的比较研究

陈婷婷[†]

中国地质大学（北京）数理学院，北京 100083

摘要：聚类分析是研究“物以类聚”的一种现代多元统计分析方法，而且聚类分析方法发展很快，并在经济、管理、地质勘探、天气预报、生物分类、考古学、医学、心理学以及制定国家标准和区域标准等许多方面都取得了很有成效的应用。

本文首先重点学习了聚类分析的相关知识，通过对具体实例数据用 SPSS 软件进行不同种系统聚类法的应用分类，并利用阈值 T、散点图和使用统计量确定适合的类的个数，把不同种系统聚类法进行研究和比较。最后得出结论：“给定一个阈值 T”这种方法的主观性较强；“观测散点图”这个方法较为直观，效率也许会好于正规聚类方法；“使用统计量”往往更明确。在聚类方法的效果方面，类平均法和离差平方和法的聚类效果相对较好。

关键词：聚类分析；分类；系统聚类法；SPSS

引言

生活中分类应用在很大程度上被需要，过去人们主要依靠经验和专业知识来分类，但这种分类是定性的，而且人的主观意识影响很大，带有一定的随意性，而聚类分析可以有效的克服以上弊端。系统聚类法是对象分类的常用方法，但根据不同的系统聚类法得到的分类一般是不相同的，聚类分析主要集中在基于距离的研究分析。本文使用不同系统聚类法对工厂产品数据的分类结果并进行研究，即对样品分类，所以本文着重学习研究 Q 型聚类分析。常见的聚类分析方法有系统聚类法和 k 均值聚类法，本文对常用的系统聚类法中的五种聚类方法进行比较研究，它们之间的不同在于类与类之间的距离计算方法不同。对样本进行不同种方法分类后往往会产生一个问题，选择哪一个结果更好，这时候就需要用给定一个阈值 T、直接观测散点图或者使

用统计量来确定适宜的类的个数，使用统计量包括 统计量、半偏 统计量、伪 统计量和伪 统计量，而且一般这四种使用统计量都是值越大表示聚类结果越好。聚类分析方法发展很快，并且在经济、管理、地质勘探、天气预报、生物分类、考古学、医学、心理学以及制定国家标准和区域标准等许多方面都取得了很有成效的应用，因而也使其成为国内外较为流行的多变量统计分析方法之一。在我国，聚类分析在数据挖掘、模糊控制和计算机视觉等领域具有广泛的应用，也是近年来得到迅速发展研究的一个研究热点。

1. 聚类分析知识

1.1 系统聚类法

在对样本进行分类时，常用距离来度量样本之间的相似性。用 d_{ij} 表示为第 i 个样本与第 j 个样本之间的距离。常见的距离有：

- (1) 绝对值距离：
$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$
- (2) 欧式距离：
$$d_{ij} = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}$$
- (3) 切比雪夫距离：
$$d_{ij} = \max_{1 \leq k \leq p} |x_{ik} - x_{jk}|$$
- (4) 明考斯基距离：
$$d_{ij} = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^q \right]^{1/q} \quad (\text{当 } q=1,2 \text{ 时, 为绝对值、欧式距离})$$

定义类与类之间的距离为两类最近样品间的距离，聚类步骤为：规定样品之间的距离，计算 n 个样品的距离矩阵 $D_{(0)}$ ，它是一个对称矩阵；选择 $D_{(0)}$ 中的最小元素，设为 D_{KL} ，则将 G_K 和 G_L 合并成一个新类，记为 G_M ，即 $G_M = \{G_K, G_L\}$ ；计算新类 G_M 与任一类 G_J 之间距离的递推公式；在 $D_{(0)}$ 中， G_K 和 G_L 所在的行和列合并成一个新行新列，对应 G_M ，该行列上的新距离值，由递推公式求得，其余行列上的距离值不变，这样就得到新的距离矩阵，记作 $D_{(1)}$ ；对 $D_{(1)}$ 重复上述对 $D_{(0)}$ 的两步得 $D_{(2)}$ ，如此下去直至所有元素合并成一类为止。（如果某一步 $D_{(m)}$ 中最小的元素不止一个，则称此现象为结，对应这些最小元素的类可以任选一对合并或同时合并。）

常见的系统聚类法：

- (1) 最短距离法：
$$D_{KL} = \min_{i \in G_K, j \in G_L} d_{ij}$$
- (2) 最长距离法：
$$D_{KL} = \max_{i \in G_K, j \in G_L} d_{ij}$$
- (3) 类平均法：
$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in G_K, j \in G_L} d_{ij}$$
- (4) 重心法：
$$D_{KL}^2 = d_{\bar{x}_K \bar{x}_L}^2 = (\bar{x}_K - \bar{x}_L)' (\bar{x}_K - \bar{x}_L)$$
- (5) 离差平方和法：
$$D_{KL}^2 = \frac{n_L n_K}{n_M} (\bar{x}_K - \bar{x}_L)' (\bar{x}_K - \bar{x}_L)$$

1.2 类的个数

下面介绍几种比较常用的确定的方法：

- (1) 给定一个阈值 T

给定阈值——通过观测聚类图，给出一个合适的阈值 T 。要求类与类之间的距离不要超过 T 值。例如我们给定 $T=0.5$ ，当聚类时，类间的距离已经超过了 0.5 时，聚类结束。

- (2) 观测样品的散点图

如果样品只有两个或三个变量，则可通过观测数据的散点图来确定类的个数。对于三个变量，可使用 SAS 软件通过旋转三维坐标轴从各个角度来观测散点图。如果变量个数超过三个，则可将原始变量综合成两个或者三个综合变量，然后再观测这些综合变量的散点图。

顺便，观测散点图还有一个重要的用途，就是从直觉上来判断所采用的聚类方法是否合理，甚至有时直接从散点图中进行直观的分类，效果也许会好于正规的聚类方法，特别是在寻找“自然的”类方面。

(3) 使用统计量

i) R^2 统计量
$$R^2 = 1 - P_k / W = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})' (\bar{x}_i - \bar{x}) / W$$

R^2 值在 0 与 1 之间， R^2 值越大，聚类效果越好，其值随着分类个数的减少而变小。（其中 W 为所有样品的总离差平方和；k 表示分为 k 个类。）

ii) 伪 F 统计量
$$伪F = \frac{n-k}{k-1} \frac{R^2}{1-R^2}$$

伪 F 值越大，表明此时的分类效果越好。

iii) 伪 t^2 统计量
$$伪t^2 = \frac{D_{KL}^2}{(W_K + W_L) / (n_K + n_L - 2)}$$

伪 t^2 值表示合并后的类内离差平方和相对于原两类的类内离差平方和的增量，值大说明被合并的两个类较分开，即上一次聚类效果好。（其中 D 表示表示 G_K 和 G_L 合并为新类 G_M 后的增量。）

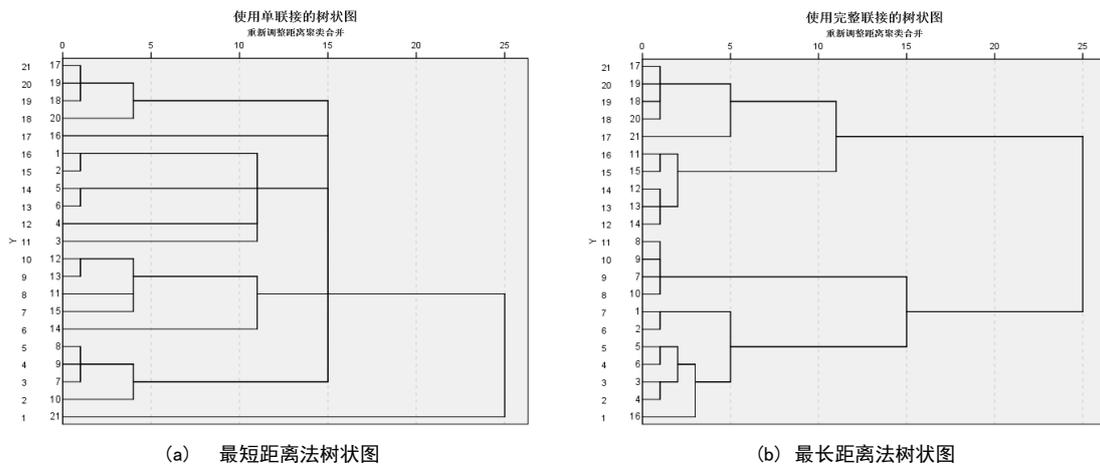
2. 数据处理

现在分析实例，将某工厂产品按照指标分类，下有 21 组数据，含有两项指标：

表 1 初始数据

| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| VAR1 | 0 | 0 | 2 | 2 | 4 | 4 | 5 | 6 | 6 | 7 | -4 | -2 | -3 | -3 | -5 | 1 | 0 | 0 | -1 | -1 | -3 |
| VAR2 | 6 | 5 | 5 | 3 | 4 | 3 | 1 | 2 | 1 | 0 | 3 | 2 | 2 | 0 | 2 | 1 | -1 | -2 | -1 | -3 | -5 |

运用 SPSS 软件，采用欧氏距离 (euclidean distance)，由原始数据得到不同种系统聚类法的树状图如下：



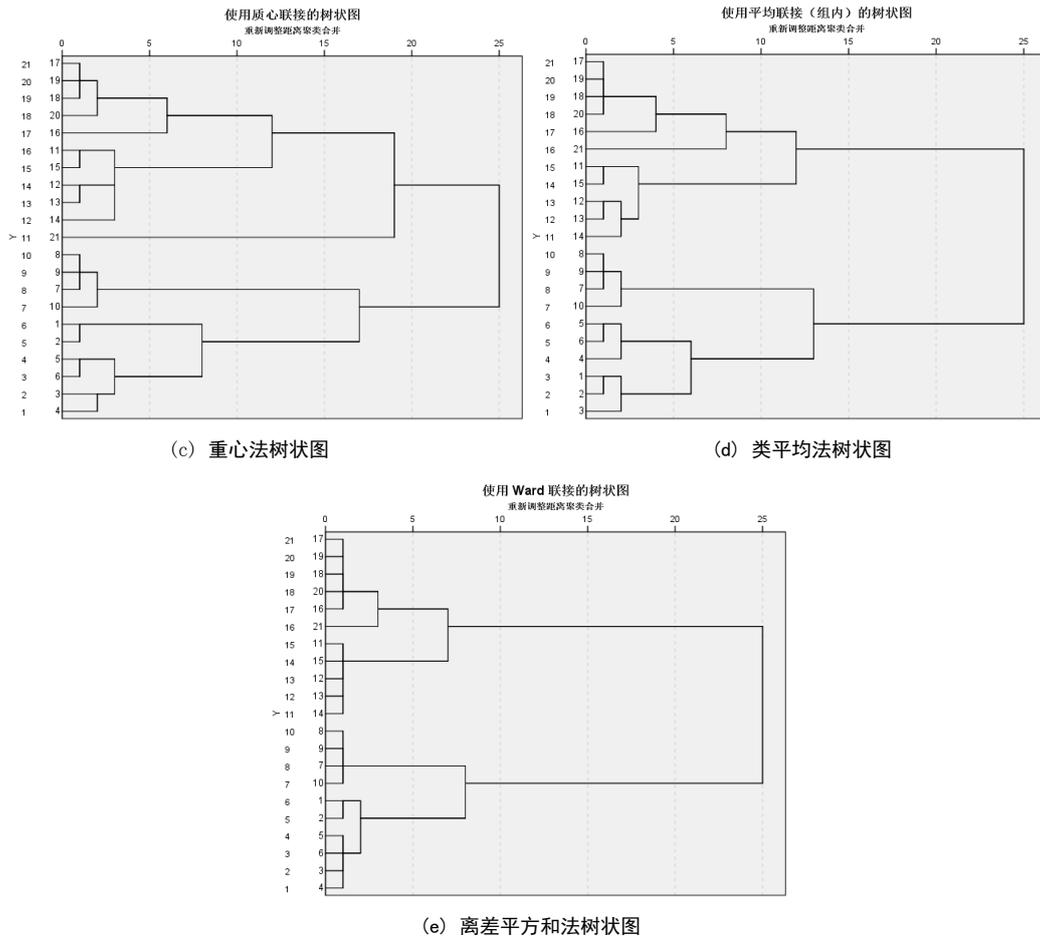
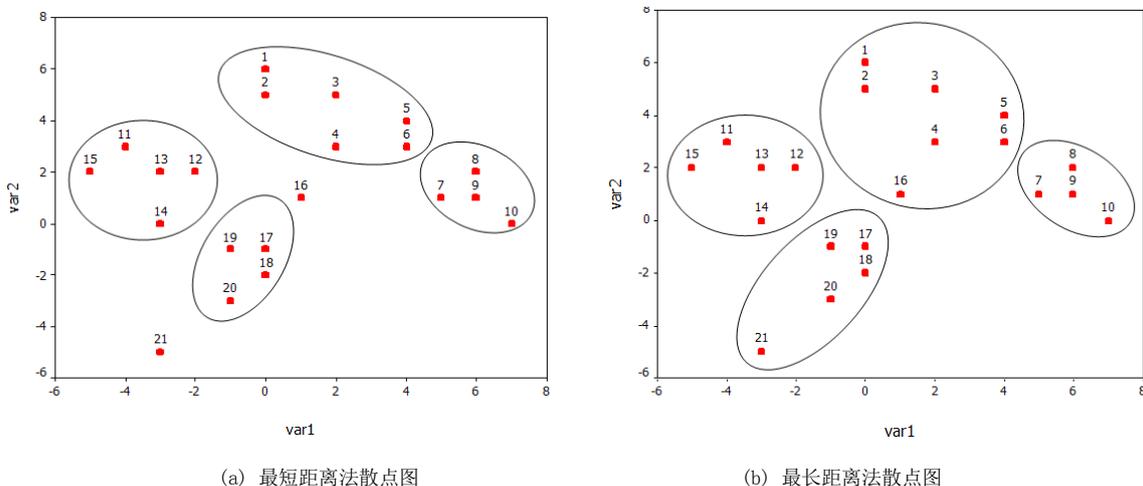


图1 系统聚类法树状图

以上五种方法的数据无效为0，即21个数据都参与分类，下面我们可以通过给定一个阈值 T 、观测样品的散点图和计算使用统计量来检验比较由以上五种系统聚类法分类的结果。

首先，通过上述图1-5，不妨分别直接给出一个阈值 T ：最短距离法阈值 $T=14$ ；最长距离法阈值 $T=8$ ；重心法和类平均法阈值 $T=10$ ；离差平方和阈值 $T=5$ 。则分别可以分为6类、4类、5类、5类、4类。但可以发现，当最短距离法阈值 $T=13$ ；最长距离法阈值 $T=10$ ；重心法和类平均法阈值 $T=9.5$ ；离差平方和阈值 $T=5.5$ 时，数据同样分别可以分为6类、4类、5类、5类、4类。因此，可知给定阈值 T 来确定适宜的类的个数具有较强的主观性，且不精确。



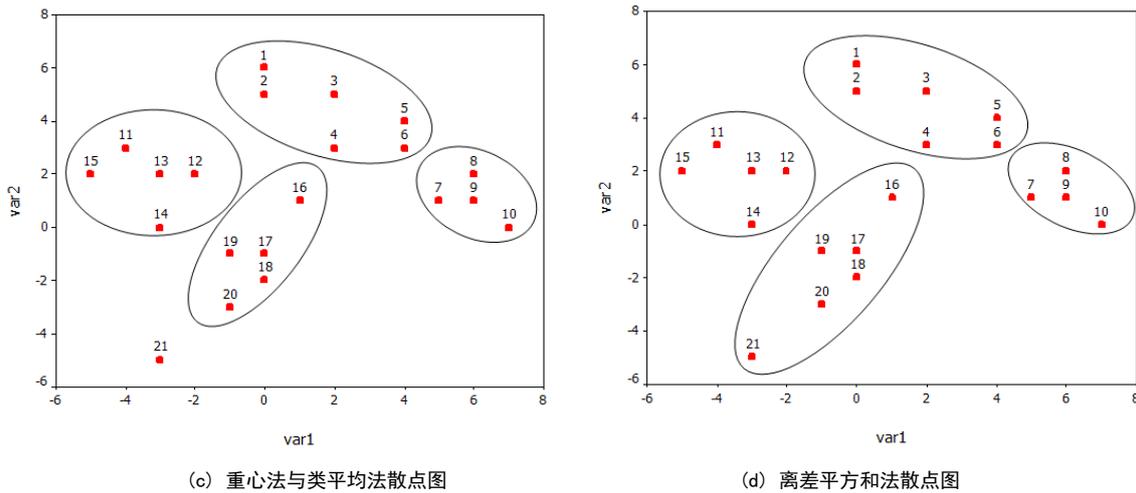


图2 系统聚类法散点图

然后，分别计算各统计量的值来确定适宜的类的个数和系统聚类方法。

表 2 不同系统聚类法统计量值表

| | 最短距离法 | 最长距离法 | 重心法和类平均法 | 离差平方和法 |
|-------------|-------|-------|----------|--------|
| R^2 统计量 | 0.898 | 0.838 | 0.878 | 0.832 |
| 伪 F 统计量 | 22.45 | 29.31 | 28.79 | 28.06 |
| 伪 t^2 统计量 | 3.11 | 22.02 | 21.32 | 21.32 |

综上，可知利用 R^2 统计量最短距离法的 R^2 最大，所以最短距离法的聚类结果更好；利用伪 F 统计量最长距离法的伪 F 值最大，最长距离法的聚类结果更好；利用伪 t^2 统计量算得最长距离法、重心法、类平均法和离差平方和法分类效果好。

3. 总结

本文学习研究了聚类分析方法，以工厂产品的 21 组数据为实例，利用 SPSS 软件采用常用的系统聚类法（包括最短距离法、最长距离法、重心法、类平均法和离差平方和法）将数据进行了分类处理得到相应的树状图，然后分别给定不同的阈值 T 和观测散点图来确定类的个数，再通过对统计量的值的计算，分析比较适宜的类的个数，最后得出结论：“给定一个阈值 T ”这种方法的不足之处是它的主观性较强；“观测散点图”这个方法较为直观，效率也许会好于正规聚类方法；“使用统计量”往往更明确。在聚类方法的效果方面，在许多应用中，类平均法和离差平方和法的聚类效果相对较好。

致谢

非常感谢导师给予我的帮助，感谢参考文献参考的作者给予我的启发，让我能更好地理解以及应用聚类分析。

REFERENCES

- [1] Xuemin Wang. Application of Multivariate Analysis (Second Edition). Shanghai: Shanghai University of Finance Economics, 2004.1.
- [2] Kaitai Fang, Enpei Pan. Cluster Analysis. Beijing: Geology Press, 1982.4.

- [3] Jun Wang, Shitong Wang and Zhaohong Deng. Some Problems in Cluster Analysis. Summary and Comment (2012 issue 03)
- [4] X.CHEN, Z.CUI, The Design and Implementation of User Clustering Based on Chameleon Algorithm. Microcomputer Development,4(15)(2005),48-50.
- [5] D.EVANS, D.C.YEN, E-government: Evolving relationship of citizens and Government, domestic, and international development. Government Information Quarterly, 23(2)(2006),207-235.
- [6] Yangyang Zhou. Simple application of cluster analysis theory. Science Chinese.2016, 1005-3573.
- [7] G.LI, J.LI, Web Log Mining Based on the Fuzzy Clustering. Journal of Computer Science, 31(12)(2004),130-131.
- [8] G. LIU, Fuzzy cluster analysis in the application of text classification. Computer Engineering and Application, 33(9) (2003), 110–111.
- [9] C.DU, G. L. JI, Fuzzy cluster analysis in the application of Chinese text classification research. Computer Engineering and Application, 8 (2006), 170–172.
- [10] Y. WANG, Y. Y. GUAN, A new judging model of fuzzy cluster optimal dividing. Fuzzy Systems and Mathematics, 20(4) (2006), 79–85.
- [11] G. JUNHUA, Clustering Analysis in Data Mining Re-search. Wuhan University of Technology, (2003), 17–28.
- [12] Lili Xu. Algorithm and Application of Cluster Analysis. Master's thesis. Changchun: Jilin University, 2010.5.

【作者简介】



陈婷婷（1994-），女，汉族，在读研究生，统计学，学习经历：2012年9月至2016年7月于中国地质大学（北京）数理学院学习，并获得学士学位；2016年9月至今于中国地质大学（北京）数理学院攻读硕士学位。Email: 954302164@qq.com