聚类分析与中药质量研究

张福良　编著

孟宪纾　审阅

人民卫生出版社

聚类分析与中药质量研究

张福良　编著
孟宪纾　审阅

人民卫生出版社

（京）新登字081号

图书在版编目（CIP）数据

聚类分析与中药质量研究/张福良编著。—北京：人民卫生出版社，1993

ISBN 7-117-02005-9

Ⅰ.聚…

Ⅱ.①张…

Ⅲ.①聚类分析-应用-中药学②中药学-应用-聚类分析

Ⅳ.R911

聚类分析与中药质量研究

张福良　编著

人民卫生出版社出版

（北京市崇文区天坛西里10号）

北京市卫顺印刷厂印刷

新华书店北京发行所发行

787×1092毫米32开本　11印张　6插页　243千字

1994年6月第1版　1994年6月第1版第1次印刷

印数：00 001—1 300

ISBN 7-117-02005-9/R·2006　定价：10.40元

〔科技新书目310—183〕

前　言

随着数学方法在医药研究中的应用，中药的分类与质量研究也趋于定量化。这标志着药学研究已从定性分析进入了定量分析阶段。作为多元统计分析的重要分支——聚类分析正是这种定量研究中不可缺少的工具之一。它已被更多的药学工作者所掌握。

基于上述原因，作者进行了初步尝试：对聚类分析在药学研究中的应用课题做了较为系统的归纳、总结和提炼，通过再现大量的应用实例、数学模型和计算机程序，试图为医药科研人员及高校有关教师提供参考。本书还侧重于方法介绍，着眼于实用价值，并借助于应用举例帮助读者理解吸收。具备微积分初步知识的读者可阅读此书。此外，全书各章彼此相对独立，读者也可依需要选读有关章节。

在本书编写过程中，我院孟宪纾教授编写了第四章和第六章；张福良副教授编写了第一章、第二章、第三章、第五章、第七章以及第八章，并研制了全部计算机程序；沈阳师范学校数学教研室孟晓燕老师绘制了插图五十幅。此外，还参阅了国内外已发表的有关论文及著作，在此一并致谢。

由于作者水平有限，书中错误与不足在所难免，恳请各位读者不吝赐教。

张福良

1993.12.于沈阳药学院

目　录

第一章 绪 论

聚类分析起源于生物学的一个分支，生物学家为了研究生物的演变规律，根据各种生物的特征将它们归属于不同的界、门、纲、目、科、属、种之中。在考古学中，为了研究样品所属的年代，常常需要了解组成样品物质的特征，从而判别样品的归属。在中药材及中成药的分类及质量研究中，也常常根据药品的某些特征、特性，某种含量的测定值等将其分类。

人类认识世界的一个重要的方法是将事物进行分类。在以往的分类学中，上述这些分类方法，多半是凭借经验和专业知识来进行的，很少与数学联系。也就是说过去的分类，多数是按定性来进行的，很少用它们的特征数值定量地进行分类。中药的化学成分十分复杂，一种中成药是多味中药材组成的复方制剂，而每一种中药材的质量又受产地、采集季节、生长年限等诸多因素的影响，所以单凭经验或专业知识来定性的分类是远远不够的。利用数学方法进行定量的、科学的分类，已成为发展的必然趋势。

聚类分析作为多元分析的一个重要分支，发展异常迅速。自它诞生几十年来，已取得了甚为丰硕的科研成果。目前，聚类分析在我国地质、考古、天气预报、工农业、医学等领域均有广泛的应用，在药学方面的应用，还刚刚起步。由于学科间的相互渗透，新边缘学科的兴起，数学与电子计算机在药学中的应用日益广泛，已使得聚类分析方法逐步成为了中药研究中不可缺少的手段。

聚类分析、回归分析和判别分析被称为多元分析的三大方法。特别是聚类与判别分析在药学中有着更加广泛的应用，将二者有机结合起来，鉴别中药质量，效果更好。

第一节　聚类分析简介

一、类与聚类的含义

我们的目的是将所研究的事物聚类。那么什么叫做类呢？显然这里所说的类与所研究的问题有关。在不同的问题中，类的含义是不尽相同的。例如，有一批不同厂家生产的同一品种的中成药，就生产厂家来分，可分为若干类；而就产品质量而言，也可分为若干类。显然后者的类与前者的类不同。要给出类的一种严格定义是非常困难的。因为事物中类与类的分界，往往是不甚明确的，或者说是模糊的。下面给出几种类的定义。

用符号G表示一个集合，设G中有 k 个元素，并用 i，j 表示G中的元素，i，$j=1$，$2\cdots k$。

定义1：设T为一取定的阈值(T为一实数)，若对任意的 i，$j\in G$，有 $d_{ij}\leqslant T$（d_{ij} 为 i 和 j 的距离），则称G为一个类。也就是说如果一个样品集中的任意两样品的距离都不大于某一个取定的实数T，则说G构成一个类。这里所说的样品距离 d_{ij} 的定义请见本章第三节。

定义2：对阈值T，若对样品集G中的任一样品 i，有 $\frac{1}{k-1}\sum_{j\in G} d_{ij}\leqslant T$，则称G为一个类。

定义3：对阈值T，若对G中任意一个样品 i，存在G中一个样品 j，使得 $d_{ij}\leqslant T$，则称G为一个类。

可见，定义1所要求的条件比较严格，凡符合定义1的

类。也一定符合定义2和定义3。在分类时应明确分类的准则，这有助于加深对所研究问题的理解。

什么是聚类？所谓聚类，就是把事物性质相同或相近的对象聚在一起，并按照这些对象的定性或定量特征数值将其分组归类。下面通过例子加以阐述。

二、引　　例

例1.1　现将不同产地，不同生长年数的人参进行分类，统计结果见表1.1。

在这里“生长年数”、“每支重量”、“总糖含量”、“总皂甙含量”称为指标（变量），每支人参称为样品。指标的选择要根据分类的目的来确定。比如说在本例中要考察这批人参的质量，就要将人参按质量分类，则指标的选择应是与人参质量有密切关系的，即哪些特征能反映人参的质量，就取其为指标。又比如，在评定中成药质量中，往往需要将一批同类产品分类，从而判别产品归属。这就需要取与该药品质量密切相关的特征作为指标，然后按照聚类分析的

表1.1

样品号	产地	生长年数	每支重(克)	总糖含量	总皂甙含量
1	浑江	1	20	0.426	1.44
2	浑江	2	36	0.593	1.99
3	集安	2	38	0.342	3.35
4	集安	3	45	0.304	3.21
5	桓仁	3	47	0.438	4.43
6	桓仁	4	54	0.414	4.48

方法将它们分类。显然指标选择得越准确，分类结果就越真实。

例1.2 这是一个考古学中的经典例子[1]，通过对样品进行分析，找出组成样品物质的特征，取两种微量元素的浓度为样品的指标。对每一样品找出其相应的两种微量元素浓度，由于是二指标（二维），故可将样品在平面中用点表

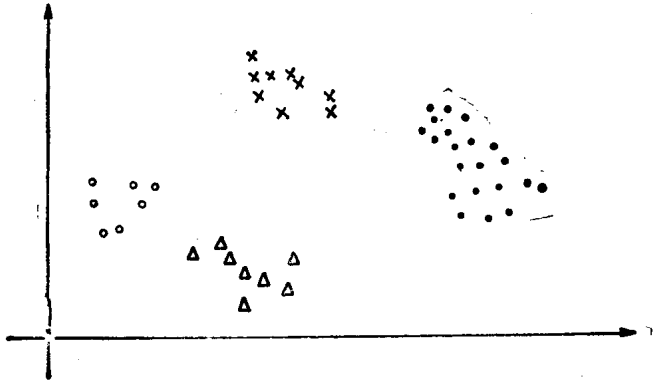

图1.1 45个黑曜岩样品的二维图示

示。图1.1给出了45个样品的散点图，在图中可直接看出一批考古样品（黑曜岩）中存在四个不同的类，且类与类之间界线清楚。实际上，图1.1是第六章中将要介绍的聚类图示法的结果。本例黑曜岩样品在文献〔1〕中是以10种微量元素的不同含量来区分的。这里我们采用第六章所讲的“降维法”，将十微数据约化成二维，从而可使其分类情况由二维平面图直观地反映出来。

值得注意的是，尽管每一个高维数据都可用第六章的办法将其投影到二维或三维空间上，但约化后的二维（三维）数据并不一定能很好地反映原高维空间中的分类情况。这是

因为，在数据由高维向低维约化的过程中，原数据要损失一定的信息。关于这方面内容，我们将在第六章详细讨论。

例1.3　此例是有关生物化学方面的问题[2]

在三个星期中从10只不同品种的山羊体内收集羊奶样品，将收集的奶样分别混合，测定10个奶样中的支链脂肪酸，共14种脂肪酸的浓度（用百分数表示），目的是要对14种脂肪酸分类。这是一个简单的聚类分析问题。文献〔2〕用系统聚类分析法将14种链脂肪酸分类，所得结果与生物化学的独立研究获得的分类结果完全一致。按文献〔2〕中的数据，用系统聚类中的最短距离法聚类，其结果为 {1，3，6，7，11}；{2}；{10}；{4，5，8，9，12，13，14} 四类。若取文献中第9号、10号山羊绘制14种酸的二维图形，则得结果图1.2，其中数字i($i=1$，2…14)表示第10号及第9号山羊在第 i 种链脂肪酸上的取值。

由图1.2 可以看到，用两只山羊也能粗略地探明14种酸

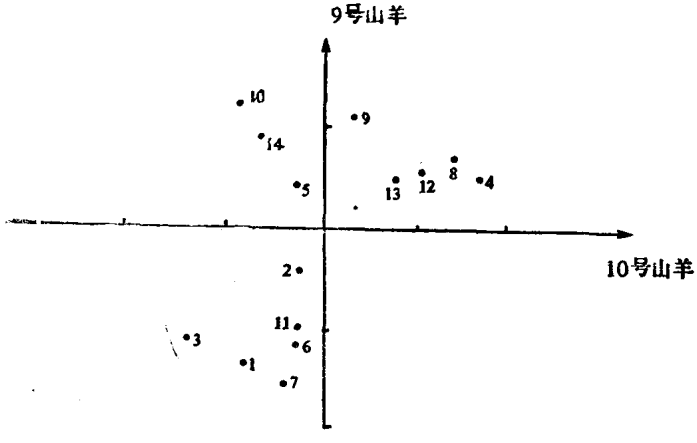

图1.2　两只山羊乳液中14种支链脂肪酸的标准化浓度

的大概分类情况，如图，数字1、3、6、7、11也聚在一起。这说明在聚类时，可首先找出两原始数据相差较大的样品，然后以这两样品为坐标轴，在二维平面上找出与其对应的若干个指标的点，由这些点分布的情况来预测指标的大概分类情况。图1.2中点的分布情况与最短距离法所得的分类结果并不完全一致。这说明要进行准确的分类，需要较多的样品或指标，而解决多样品、多指标的分类问题无法在二维或三维空间中实现。这就必须寻找一种新方法，聚类分析法就是将多样品、多指标分类的重要工具。

三、应用概况

有过聚类分析尝试的读者，一定遇到过这样的问题，在使用不同的聚类方法时往往得到不同的分类结果。究竟哪一个分类结果更能真实地反映客观事物的类别呢？至今还没有一个准确的、通用的判别方法及检验手段。正因如此，聚类分析的理论还不够完善，它仍是一个比较年轻的数学分支，尽管这样，聚类分析做为处理多样品、多指标的分类工具，仍不失为一种有实用价值，结果尚可满意的分类方法，并有着强大的生命力和发展前景。

聚类分析作为科学研究的工具，其应用已遍及各个领域，并且越来越引起人们的兴趣和关注。以聚类分析为工具的研究论文每年增长的速度更是惊人。图1.3所示是1964至1975年发表的关于聚类分析应用研究的论文情况[3]。论文数量之多，内容之广，增长速度之快，是其它多元分析方法所不及。

聚类分析作为一种通用的科学研究手段，在不同的科学领域有着广泛的应用。本书侧重于聚类分析在中药及中成药

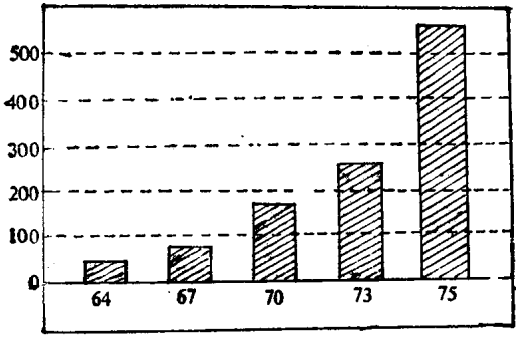

图1.3 聚类分析研究论文的增长情况

方面的应用，力图向药学工作者提供令人感兴趣的资料和方法，并将作者及其他药学工作者在聚类分析方面的应用科研成果收集于本书中，做为应用范例，供读者在阅读时参考。

第二节 数据处理

一、样本、指标和原始数据

样本也称样品。把有待于分类的事物（所研究问题的对象）叫做样本，而把样本的特征值叫做指标。一般说来，样本和指标都不是一二个，而是若干个。习惯上用 x_{ij} 表示第 i 个样本的第 j 项指标的数值（$i=1, 2\cdots n$，$j=1, 2\cdots m$），在这里有 n 个样本，m 个指标。

将 n 个样本和 m 个指标的测定值排成一个数表，称之为矩阵。矩阵是线性代数中非常重要的概念，详细的介绍读者可参阅本章第五节。

矩阵中的每一个数叫做矩阵的元素，矩阵的行数 n 及列数 m 不一定相等，且 $n\geqslant 1$，$m\geqslant 1$。于是前面提及的 x_{ij}（$i=1, 2\cdots n$，$j=1, 2\cdots m$）可用一矩阵表示：