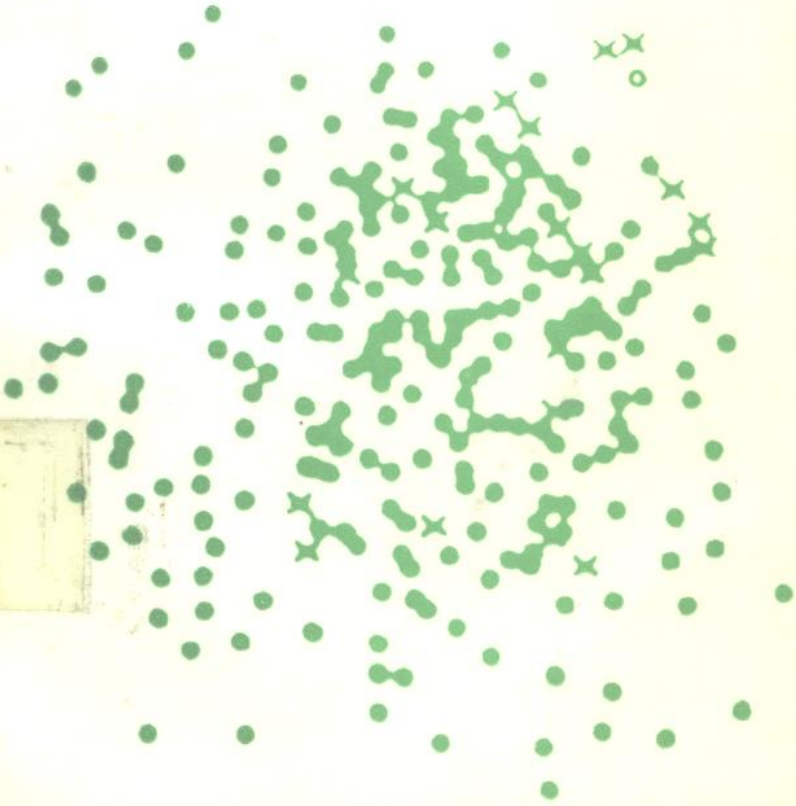
聚类分析法
解析分析化学数据

〔比利时〕 D.L.马萨特 L.考夫曼 著

化学工业出版社



聚类分析法解析分析化学数据

〔比利时〕 D. L. 马萨特 著
L. 考夫曼
刘昆元 译
俞汝勤 校

化学工业出版社

聚类分析法解析分析化学数据

D. LUC MASSART
LEONARD KAUFMAN
The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis
JOHN WILEY & SONS New York·Chichester·Brisbane·Toronto·Singapore 1983

聚类分析法解析分析化学数据

刘昆元 译
俞汝勘 校

责任编辑：任惠敏
封面设计：任 辉

化学工业出版社出版发行
（北京和平里七区十六号楼）
化学工业出版社印刷厂印刷
亘各庄装订 厂装订
新华书店北京发行所经销

开本 787×1092 1/32 印张8 字数184千字
1990年4月第1版 1990年4月北京第1次印刷
印 数 1—2,300
ISBN 7-5025-0530-X/O·14
定 价4.80元

内 容 提 要

本书介绍了模式识别的重要方法之———聚类分析，以及该方法在分析化学和有关方面的应用。全书共分八章：前五章叙述聚类分析的基本原理、方法和多元统计分析技术；第六至八章介绍该方法在优化分析化学程序、解析分析化学数据和其他化学数据等方面的应用，提供了聚类分析软件程序的文献资料来源，并用实例演示了聚类分析的全过程。本书将聚类分析这一数学方法与分析化学有机地结合起来，内容丰富，系统性强，是化学计量学方面一本有价值的参考书。

本书可供高等学校化学与分析化学专业师生和其他专业从事化学和分析化学的科技人员参考。对于需要利用化学与分析化学数据解决有关实际问题的其他领域，如环境科学、地质学、考古学、临床诊断和药物设计等方面的科技工作者，本书亦有很大的参考价值。

代 译 序

美国Howery博士等曾指出，70年代化学学科发展的最重大事件是计算机进入化学科学，并认为80年代化学教育的一项首要任务是引导化学工作者进入计算机时代〔D. G. Howery and R. F. Hirsch, J. Chem. Educ., 60,656 (1983).〕。化学（包括分析化学）与计算机的结合，最集中地体现在化学计量学的发展上。化学计量学可认作化学科学与计算机科学的接口。在化学计量学众多的分支中，化学模式识别占有重要位置。聚类分析则是模式识别中一个相对独立的分支——无监督的模式识别，它与因子分析等其他化学计量学分支又有着密切的联系。比利时Vrije大学D. L. Massart教授与L. Kaufman教授所著《聚类分析法解析分析化学数据》("The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis",1983.)一书是为对化学计量学感兴趣的分析化学工作者及其他化学工作者撰写的一部重要专著。

1978年这两位教授及其合作者曾出版了《实验室方法和分析规程的评价与优化》("Evaluation and Optimization of Laboratory Methods and Analytical Procedures",Elsevier Amsterdam, 1978.)。这本书曾被当作化学计量学的教科书使用。由于化学计量学的迅速发展，该书已难长期胜任这一任务。D. L. Massart和L. Kaufman教授的这本新书，虽然涉及的范围显著缩小并专门化了，但其深度正反映了聚类分析法在分析化学中应用的发展趋势。

刘昆元同志在湖南大学攻读博士学位期间将此书译出，并

做了若干补充，如增补了国内有关文献，有序样本的聚类分析及信息群分等内容。对于原书中的印刷错误及其他问题，均一一予以订正。可以预期此书的出版对推动化学计量学方法在分析化学中的应用将会有所裨益。

1985年美国国家标准局主持的一次化学计量学学术讨论会上曾提出，撰写化学计量学方面的参考读物和教材是一项极迫切的任务。国内外这方面的专著与教材均感缺乏。翻译有关专著能在一定程度上弥补国内这方面之不足，同时我们更期望有化学计量学方面的新著作出版。

俞汝勤

1986年8月

序

当今分析化学家们获取数据的卓越才能使自己面临着这样一个问题：解析这些越来越经常遇到的多维数据阵列显得更加困难。聚类分析就是帮助解决这些困难的一个方法。

聚类分析在许多科学领域中，尤其是行为科学和生物分类学中，是一个普遍采用的技术。在分析化学中应用这个技术已有十来年之久，但不及与之密切相关的监督模式识别（supervised pattern recognition）和因子分析等方法普及，其原因之一是由于大多数分析化学工作者似乎只注意了最简单的谱系聚类方法（见本书第三章）和这些方法的一些缺点。我们认为，适应性更广的非谱系聚类方法（见第四章）结合显示方法（第二章）和第五章中简述的克服聚类分析中所遇困难的一些方法，将使聚类方法在分析化学中获得更广泛的应用。

本书第六章介绍的许多应用实例表明了聚类分析在分析化学中的应用潜力。这一章谈不上是有关文献的详尽综述（本书的其余章节亦如此），我们的意图是选择能够说明聚类分析的应用和展示聚类分析有应用价值的广泛领域的例子。第七章列举了一些可供采用的计算机程序包。从这一章可以看出，已编出可供选用的大量计算机程序使得这一技术能为任何希望采用的读者所采用。我们希望本书能吸引更多的分析化学工作者来进行这方面的工作。

本书列入 Wiley 的《化学分析丛书》，当然主要是为分析化学工作者编写的，但对所有采用分析数据的读者，如临床化学工作者、环境化学工作者和考古计量学工作者以及采用各种

化学数据的其他读者，如药物化学工作者都是有用的。

我们的目的有两个：

1．为对聚类分析的基础理论有较大兴趣的读者提供一个良好的数学基础；

2．向要求使用这一方法而又希望避开数学细节的读者阐明基本原理，指出方法的利弊和难点。

为此，我们采用双层次编写方法，将涉及数学内容的第一至四章分成较侧重于定性叙述的部分和若干个数学部分。定性叙述部分阐述各种方法的基本原理，用文字和少数关键式子简述数学概念，给出示例并讨论方法的优缺点。这些部分能使读者正确运用聚类方法而又避开数学细节。数学部分则是为具有一定数学基础的化学工作者编写的，这一部分较详细、完整地叙述了有关数学知识。

第一至四章的定性叙述部分及第六章主要由马萨特（Massart）执笔；第一至四章数学部分及第七章由考夫曼（Kaufman）执笔，第五章及第八章由二人合写。特此说明，以便于读者就本书有关内容与作者联系。Solange Peeters 及 Gerda De Boeck为大部分手稿打字，Annie De Schrijver协助绘图，谨此致谢。

D. L. 马萨特

L. 考夫曼

1983年2月

目　　录

第一章　聚类分析的初步介绍

运用聚类方法的一般前提是存在大的数据集。分析化学家们发展了强有力的研究多元素或多组分体系的技术，电子计算机使数据易于获取，这使得遇到大型数据集的机会比几年以前要多得多，因而要求发展相应的解析技术来解析这些数据，而解析这些数据的第一步就是理出这些数据的一些头绪，了解数据集合的内在结构。聚类分析就是解决这类问题的一系列方法的总称，它可定义为“按照对象的定性或定量特征将其分组归类”。

将研究对象分类归于相应的范畴是科学研究的一种基本方法。生物学家为解决大量生物的分类问题，建立了系统分类学；医生在诊治病人时，首先必须将每一患者按其病情进行分类，即所谓诊断。这些分类在生物化学中是根据生物的形态特征进行的。通常，取得分析数据就是要用某种方式给出或描述被分析物质的特性，因此，原则上在许多情况下可以根据这些被分析物质进行分组或归类。实际上，要理出由许多对象特征数据对构成的数据集的头绪，就必须进行分类。聚类分析就是能够达到这一目的的定量规范方法。而本书就是向分析化学工作者介绍这一方法。

应当指出，采用聚类分析所取得的分类结果与研究有关数据集的专门学科所取得的结果，一般没有很大程度的差别。例如，铁陨石的分类，采用聚类分析方法取得的分类结果比陨石专家们的分析结果不一定好很多。因为聚类运算中以常规与显

式方法运用的分类依据就是（或者应该是）专家们运用的那些准则。但是在一个陨石专家手中，铁陨石分类如果采用聚类分析，将取得该铁陨石的最优分类。聚类分析确实有许多优点。必须明确分类的准则，这有助于加深对所研究问题的理解。聚类分析是经济、有效的，因为这个方法比人的头脑能考虑多得多的数据及其相互关系，使用简便，对数据不挑剔。靠人脑分类，通常也可以找到进行满意分类所需的重要特征，但这将需要花费很多时间，需要进行许多诸如绘制图表之类的单调工作。

第一节　引例与本书内容概述

聚类（clustering）这一术语源于过程的图示，如图1.1-1.4所示的情况。例如在图1.1中，用小圆圈标记的是一些考古学数据，这些数据形成若干个类〔或簇、群〕❶（clusters，法语称nuées，意即“云”），每一个簇构成一个类❷。在图1.1中，这些类可以是根据其地理来源或年代划分而成的。

图1.1是一个双变量图，图中所比较的是两种微量元素的浓度。在这种情况下，通常容易找出类。可是，当存在许多个变量时，问题就属于多元情形了。由于除二元和三元外，更多维的图无法画出，因此直观地找出类就不可能了。

---

❶ 方括号〔 〕中的内容是译者加的（下同）。——译者注

❷ 本书将由聚类分析得到的类（cluster）按通用名称译作类，但这容易与分类（Clasify）得到的类(Class)混淆。严格地讲，聚类与分类是不相同的，聚类在整个过程中采用的是同一标准（如相似性系数，见本章第三节），而在分类过程中采用的标准可能不是一致的。例如，图书分为自然科学和社会科学类等，自然科学则分为数学、物理、化学……，化学又分为无机、有机、分析……，社会科学也同样可以分为许多门类。显然，这里在各学科及各层次中所采用的分类标准是不一致的，故只能叫分类。本书在必要时将这两种类用英文注出。——译者注

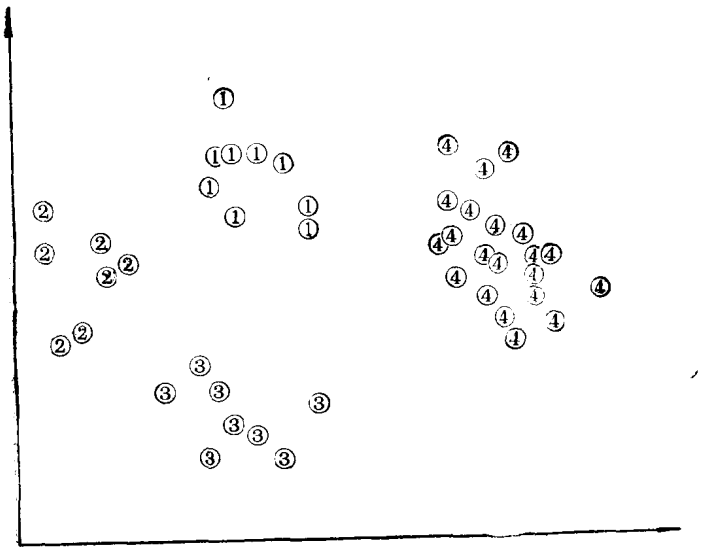


图 1.1 45个黑曜岩样品的二维图示
这些黑曜岩样品可分成四个不同来源的族（引自文献〔1〕）

在详细讨论聚类方法之前，我们先考察几个有关分析化学方面的例子。这些例子中由于出现了多维数据，因而定义一些类（class），即找出多维空间中的类（cluster）是有意义的。

第一个例子是考古计量学中的一个例子[1]。通过对样品进行分析，我们可以了解组成样品物质的特征。同一来源的粘土或同一地方制造的玻璃具有相近的组成，这个组成不同于其他来源的样品。暂且可将图1.1看作关于两个微量元素的双变量图，我们在图中可以直接看出一批考古样品（本例中就是黑曜岩玻璃）中存在的若干个类。很清楚，存在四个不同的类（或四个不同的来源）。实际上，图1.1就是第二章中将要介绍的

显示方法所给出的结果，图中黑曜岩样品是以10种元素的不同含量（10个指标）来区分的。第二章中所介绍的方法之一就是用来将十维数据约化至二维，以直接观察组成相近的样品所形成的类。

图 1.2 所示是色谱分析中的一个例子。物质可根据它们与固定相的不同作用能力用气-液色谱方法进行分离。为了更清楚地了解气-液色谱行为，我们可以测定若干种溶质在各种不

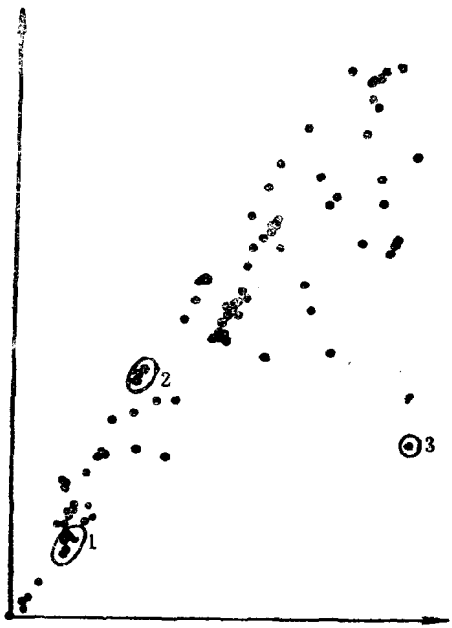


图 1.2

点代表固定相，两坐标轴分别代表某些溶质
在这些固定相上的保留指数（参见正文），数字 1 和 2
表示两个可能的类，点 3 是一个奇异点 （引自文献[2]）

同固定相上的保留指数，通过分类来找出数据的结构以便解析

所取得的这些数据，例如需要确定那些固定相对某些溶质的作用行为“相似”。图1.2是对100个固定相进行试验得到的保留指数的二维图，两坐标轴中每个表示对5种溶质的保留指数求和所得的数据，例如其中之一表示苯、正丁醇、吡啶、2-甲基-2-戊醇和顺二氢化茚五种溶质保留指数的和。这里得到一些扁长的类，如类1和类2（图1.2）。我们可以得出结论，类1中各固定相的保留行为不同于类2的行为。这里要注意，类可以是任意形状的。这一点在分析化学中特别重要，也是要求研究聚类分析的专家们解决的问题之一。我们还注意到，某些固定相（如图1-2中的点3）与其它任何固定相都无相似性质，这样的点称为奇异点〔或野点(outlier)〕。

另一个是宇宙化学中的例子[3,4,5]。在地球上的不同地方发现了大约600个所谓铁陨石。研究这些铁陨石时，首先要知道是否可以认为它们形成一个同类的组，或者相反，即是否可

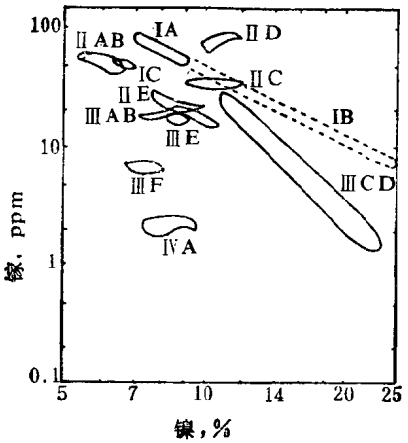


图 1.3 铁陨石的二维图

图中陨石的类用实线标记(引自文献〔4〕)

以认为它们属于不同种类的陨石。也就是说，是否有些陨石与其它一些比较其性质相互更“相似”些。在这种情况下，必须区别出陨石的组或类。每个铁陨石具有特定的分析化学参量（镍和微量元素的含量）和物理参量，因而我们可以比较得出的模式并进行分类。在这种特例中，用双变量图形来表示类的形成情况仍属可行，二维坐标分别对应于镍和镓的含量（图1.3）。但要进行完全分类，需要两个以上的参量，为此文献〔5〕提供的规范方法是有用的。

最后一个例子是有关生物化学的一个问题，如图1.4所示。哺乳山羊能在乳房中产生或者通过血液聚集一些脂肪酸，这些过程可能通过不同的代谢路径进行。现在要问，哪些脂肪酸是

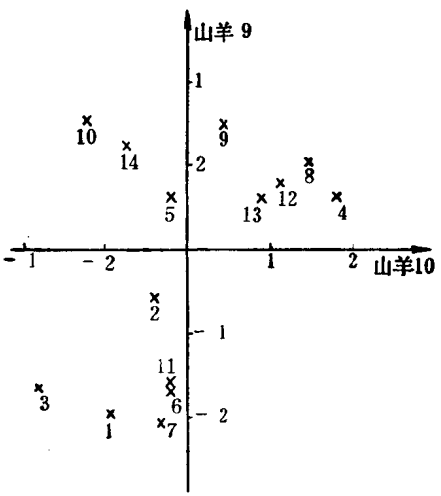


图 1.4　两只山羊乳液中14种支链脂肪酸的标准化浓度用聚类分析进行分类将在第六章和第八章中讨论(引自文献〔6〕)