

THE METHODS AND PROGRAMS FOR
NUMERICAL TAXONOMY

数量分类的方法与程序

钟扬 陈家宽 黄德世 编著

By

Zhong Yang, Chen Jia-kuan and Huang De-shi

武汉大学出版社

THE UNIVERSITY OF CHICAGO PRESS
530 North Dearborn Street, Chicago, IL 60610, USA
http://www.press.uchicago.edu

数量分类的方法与程序

作者：[美] 詹姆斯·H·柯林斯 著

译者：王德

Chicago, Illinois: The University of Chicago Press, 1996

ISBN 7-309-03600-0

数量分类的方法与程序

THE METHODS AND PROGRAMS FOR
NUMERICAL TAXONOMY

钟扬 陈家宽 黄德世 编著

By

Zhong Yang ,Chen Jia-kuan and Huang De-shi

武汉大学出版社

数量分类的方法与程序

钟扬 陈家宽 黄德世 编著

*

武汉大学出版社出版发行

(430072 武昌 珞珈山)

海南华美公司激光排版

武汉大学印刷厂印刷

*

850×1168 毫米 1/32 9.5 印张 240 千字

1990 年 10 月第 1 版 1990 年 10 月第 1 次印刷

印数: 1—2000

ISBN 7-307-00834-3/Q·27

定价: 5.00 元

内 容 提 要

本书系介绍数量分类学方法的入门书。主要内容为数量分类的基本概念和常用方法（如系统聚类、图论聚类、主成分分析等），以及一些正在发展的方法（如模糊聚类和数量分支分析等）。同时，介绍了这些数量分类方法的BASIC程序。

本书可作为从事生物分类学研究和有关科研工作者的工具书，亦可供大专院校有关专业的师生阅读学习。

前言

分类学是一门古老的基础学科。迄今为止，在生物学研究和应用中仍占有极其重要的地位。然而，随着生物科学的不断发展，传统分类学中存在的不足之处显得十分突出。一方面，生物物种的不断发现和研究领域的逐渐扩展导致资料积累日益增多；另一方面，长期沿用的定性描述方法受人为因素的影响较大，加之繁琐低效的工作方式，不可避免地造成分类研究中的一些混乱。因此，有关分类方法的定量化和科学化已成为生物学界一项十分重要的任务。

数量分类学 (numerical taxonomy)，或译数值分类学，是一门应用数学理论和电子计算机技术处理生物分类问题的边缘学科。它是本世纪 50 年代末期，由英国微生物学家史尼斯 (P. H. A. Sneath)、动物与人类学家克恩 (A. J. Cain) 和美国生物统计学家索卡尔 (R. R. Sokal) 等人在总结前人对定量分类方法的研究基础上创立的。此后，越来越多的生物学工作者对这门学科的理论、方法和应用产生了浓厚的兴趣。70 年代以后，数量分类方法已广泛应用于生物学及相关学科的众多研究领域中，其数学理论和分析模型也逐步得到发展。

国内数量分类学研究的开展也引起了普遍的重视。近年来，不少单位从事了这方面的工作，尤其是广大青年分类学工作者兴趣较高。为了结合生物系研究生“现代植物分类学技术”课程的教学实习及培训其他分类学工作者的需要，我们编写了本书。编写中主要参考了史尼斯博士等著《数值分类学：数值分类的原理和

应用》和阳含熙教授等著《植物生态学的数量分类方法》两本著作，以及徐克学先生在沈阳等地的数量分类学讲习班教材。考虑到大多数生物学工作者在较短时间内全面掌握众多数量分类方法的困难，本书仅综合介绍了数量分类学中的基本概念和较为成熟的分析模型，也选择了一些正在发展中的方法。力求通俗易懂，简明实用。同时，还介绍了常用数量分类方法的 BASIC 程序，供实际应用中参考。尽管书中大部分例子都与植物分类有关，然而，这些分析方法对于其它分类学科也是适合的。

本书是“数量分类方法与程序”课题的部分总结。中国科学院植物研究所系统与进化植物学开放实验室给予了专项资助。在写作过程中，得到了我们所在单位中国科学院武汉植物研究所和武汉大学生物系的鼓励与支持。李群副教授精心审编本书。张晓艳同志参加了大量工作。李伟、邹洪才等同学给予了帮助。在此，我们一并表示衷心的感谢。

虽然本书的编写历时两年有余，但由于我们水平所限，仍会有不少错误和不足之处。敬请读者批评指正。

编著者

1989年10月于

武昌磨山、珞珈山

目 录

前言	(1)
第一章 数量分类的基本概念	(1)
第一节 分类的目的与形式	(1)
第二节 数量分类学与数量分类方法	(4)
第三节 分类运算单位与分类性状	(6)
第四节 分类分析	(8)
第二章 性状的选取与量化	(14)
第一节 性状选取的原则	(14)
第二节 性状编码	(24)
第三节 数据标准化	(27)
第三章 系统聚类分析	(35)
第一节 相似性系数	(35)
第二节 系统聚合方法	(45)
第三节 聚类结果分析与方法比较	(72)
第四章 排序分析	(82)
第一节 矩阵基本知识	(82)
第二节 主成分分析	(91)
第三节 主坐标分析	(102)
第四节 典型相关分析	(107)

第五章 图论聚类分析.....	(114)
第一节 图论基本知识.....	(114)
第二节 图论聚类方法.....	(117)
第六章 模糊聚类分析.....	(126)
第一节 模糊数学基本知识.....	(126)
第二节 系统模糊聚类.....	(135)
第三节 模糊图与最大生成树.....	(144)
第七章 其它数量分类方法.....	(148)
第一节 信息聚类法.....	(148)
第二节 动态聚类法.....	(156)
第八章 数量分支分类.....	(161)
第一节 分支分类学简介.....	(162)
第二节 数量分支分析方法.....	(168)
第九章 数量分类程序.....	(184)
第一节 计算机辅助分类分析简介.....	(184)
第二节 BASIC 语言简介.....	(187)
第三节 数量分类程序.....	(195)
参考文献.....	(293)

第一章 数量分类的基本概念

第一节 分类的目的与形式

按照辛普森(G. G. Simpson)(1961)的定义,分类学(taxonomy)是“一门最初步又最广泛的生物科学”,其主要任务是“科学地研究各种生物的区别及其间的关系。”^[56]而广义的分类(classification)是指依据被研究的各种事物所反映的相似关系,将它们划分为不同的类群。因而,分类不仅是生物分类学研究的核心内容,而且在其它自然科学与社会科学领域都是一种基本的研究手段。

一般来说,各种分类的目的都是为了划分出不同等级的类群,使得同一类群中的事物尽量相同,而不同类群间的事物尽量不同,从而有利于我们观察和研究众多事物之间存在的共性与个性,掌握事物变化发展的规律。但是,由于生物物种数量庞大,生命现象繁杂纷呈,加上人们认识水平的差别,生物分类学中目前还很难完全实现一般目的的分类(general purpose classification),更多的是具有特殊目的的分类(special purpose classification)。对分类学家来说,建立分类系统的最终目标是用自然系统(natural system)取代人为系统(artificial system)。^[62]

各种分类主张或方法在形式上是多种多样的,这里仅介绍如下几种类型:^[5,58]

重叠的与非重叠的分类

当每一事物只能属于最后划分的某一类群,而不能同时属于其它类群时,则称这样的分类为**非重叠的**(non-overlapping),即“非此即彼”的分类;反之,如果一个事物同时属于两个或两个以上

的类群时,这样的分类就是重叠的(overlapping)。

重叠的与非重叠的分类在植物分类学中都是常见的。例如,被子植物中依据子叶数目划分为单子叶植物和双子叶植物,这就是非重叠的分类;而人们通常所称的水生植物、观赏植物和药用植物等,如同时置于一个系统则产生重叠的分类。因为,很多植物可能既是水生植物又是观赏植物,也许还是药用植物等等。当然,由于占有资料或见解的不同造成同一植物在不同分类系统中位置不一致的情况是十分自然的。但在同一系统中,目前均只使用非重叠的分类方式。

内在的与外在的分类

将需要分类的事物看成一个完备的集合,仅仅依据它们本身的特征进行分类的方式称为内在的(intrinsic);反之,将某些外部因素引入其中并参与分类过程的方式称为外在的(extrinsic)。

内在的分类相当于处理封闭系统,这种方式较为普遍;外在的分类则相当于处理开放系统,需要考虑的内容更加复杂化。目前,生物分类中还极少直接引入外部因素(如环境等),即使分支分类中存在外类群比较法,借助外部类群分析分类类群特征,也不能算作外在的分类方法。这一内容在后面将详细讨论。

等级的与非等级的分类

将事物集合按不同等级逐渐划分类群,结果可能形成树状结构的方式称为等级的(hierarchical);反之,不按先后次序,仅按类群内相同性状划分,结果可能形成网状结构或更复杂结构的方式称为非等级的(non-hierarchical)。等级的与非等级的分类又常被称为系统的(systematic)与非系统的(non-systematic)的分类。

例如,按《国际植物命名法规》的等级分类中主要的植物分类等级(rank)有12个,即界、门、纲、目、科、族、属、组、系、种、变种、变型。事实上,还可根据需要设立亚等级(如亚属、亚种等)或超等级(如超目等),那么等级的数量将增加一倍以上。

聚合的与分划的分类

将被分类的事物,依据相似原则两两合并,逐渐归并成较大的类群直至整个集合,这种分类方式就是聚合的(agglomerative);反之,从整个集合开始,通过逐次二分最终形成一定水平的类群直至所有事物,这样的方式就是分划的(divisive)。

现有的分类学系统几乎都是依靠聚合分类的方式所构成的;而检索表的制定与使用则属于分划的分类方式。

一元的与多元的分类

仅仅根据事物具有的某一特征或属性将其聚合或分划的方式就是一元的(monothetic);反之,根据多个特征或属性的综合性质对事物进行分类处理的方式就是多元的(polythetic)。

在分类学发展初期,林奈(C. Linnaeus)的性系统(sexual system)以每朵花雄蕊的数目去划分纲,以雌蕊的数目去划分目等,实际上就是一元的分类。目前,几乎所有的聚合分类都是多元的;而分划分类中虽然大多数是一元的,但也有多元的。

加权的与不加权的分类

多元分类中对不同特征或属性同等看待,这种方式称为不加权的(unweighted);反之,对这些特征或属性赋以不同的权重,即根据其不同的重要性进行分类的方式称为加权的(weighted)。

加权与不加权的分类方式不是绝对的。在不加权的分类中认为应当进行后验(posteriori)(即根据经验)而不是先验(priori)(根据基本推测或推导)加权。从数学方法上看,不加权的分类中也可以采用加权的算法。此外,一元分类事实上是多元加权分类中的一种极端情况,它使某一特征的权重为1而其它特征的权重都为零。

除上述几种分类方式外,还有一些具体的分类运算方式,如序贯的(sequential)与同时的(simultaneous),直接的(direct)与迭代的(iterative),适应的(adaptive)与非适应的(non-adaptive)等,以及不同的分类标准,如局部的(partial)与全局的(overall)等,有兴趣的读者可参阅文献[58]。

总之,自然现象不是孤立的和绝对的,它的种种侧面都会随着人们认识水平的提高而逐步揭示。在具体研究工作中,采用何种方式往往需要研究者根据具体分类工作的目的与要求加以选择,而电子计算机技术的发展为短时间内综合考虑与比较各类方法提供了便利的条件,也将促进新的分类方式的产生和发展。

第二节 数量分类学与数量分类方法

尽管在 18 世纪中叶的林奈时代就有了客观地及定量地处理分类学问题的设想,在本世纪初又有斯米尔诺夫(E. Smirnov)(1925, 1927)和珀尔森(K. Pearson)(1926)等人开始将统计学方法引入生物分类学研究的尝试,但直到 50 年代末期,随着电子计算机技术的迅速发展,一些复杂的数学模型逐渐克服实际应用的困难,人们才开始广泛应用数学分析方法来实现生物分类学目的的实践。特别是英美两国的一批青年分类学工作者有关细菌、动物和植物学的研究工作中应用的定量分析方法引起了人们的注意。当时,英国伦敦国立医学研究院的史尼斯,牛津大学的克恩和哈里森(G. A. Harrison),美国堪萨斯大学的索卡尔和米切纳(C. D. Michener)3 个研究小组独立地进行着这方面的研究。此外,波兰以弗洛雷克(K. Florek)为首的弗罗茨拉夫分类学派(Wroclaw School of Taxonomy)以及美国纽约植物园的罗杰斯(D. J. Rogers)和谷本(T. T. Tanimoto)等也开展了有关工作。这些都为数量分类学的产生奠定了基础。索卡尔和史尼斯合作完成的《数量分类学原理》(1963)^[60]一书的出版标志了分类学与数学相结合的产物——数量分类学的诞生。十年之后,史尼斯和索卡尔又出版了《数值分类学:数值分类的原理和应用》(1973),^[58]进一步总结了数量分类学的理论与方法,并且展示了数量分类学应用的巨大成就。总之,这门学科对生物分类学的方法论产生了十分重大的影响^[26, 27]。海伍德(V. H. Heywood)称其为 20 世纪生物分类学界最有影响的两件

大事之一。

数量分类学派认为:分类系统应当建立在被分类有机体的全面相似性(overall similarity)基础上。全面相似性意味着在一次分类中,所依据的特征越多越全面,则分类效果越好。这里假设了特征是先验等权(不加权)的,即在分类之前将每个特征的作用平等看待;任何两个分类群间的相似性是每个特征相似性的综合函数;由于特征相关程度的不同,可以依据函数值的不同将分类群予以划分;数量分类学将分类学看成一门经验的科学,在实践上以表征的(phenetic)相似性作为分类基础。

上述原理或见解一直是分类学界争议颇多的内容。对数量分类学的理解也随着学科的发展而变化。根据其基本原理,数量分类学派亦被称为“新阿当森(Neo-Adansonian)学派”,反映出它的哲学渊源为18世纪已有的阿当森原理;在应用数学分析工具的早期,也有人称之为数学分类学(mathematical taxonomy);更重要的是,由于数量分类学揭示的有机体间的亲缘关系是一种表征关系,因而它常常被称为表征分类学(phenetic taxonomy)。与此相对应,重点研究有机体间系统发育关系的分类学理论被称为分支分类学(cladistic taxonomy),而以传统分类学为主兼收上述两个学派理论或方法的进化分类学则被称为综合分类学(synthetic taxonomy)。三种学派之间依然有很多联系。例如,在数量分类学后来的观点中认为从某一类群的分类结构以及特征相关中可以得出系统发育的推论,即有关演化途径和机理的某些设想。^[29,58]

数量分类学的基本观点目前还没有被所有分类学工作者接受,其理论基础自身也有很多薄弱之处。尽管如此,应用于数量分类学实践的数量分类方法,由于既建立在较严格的数学基础上,又能为生物分类提供简明直观的结果,因而受到分类学界重视。事实上,如同成分与结构分析对于化学分类学(chemotaxonomy)^[57]和染色体核型分析对于细胞分类学(cytotaxonomy)^[63]的作用一样,一些较为成熟的数量分类方法已广泛应用于生物分类学的众多学科

领域,甚至逐渐渗透到其它分类学理论中,形成了新的交叉学科分支。例如,近来发展较快的数量分支分析(numerical cladistic analysis)就是其中之一^[6]。

从形式上看,传统数量分类方法大多是多元的、聚合的、等级的、内在的和重叠的技术。现代数量分类方法则是对上述方法的补充和扩展,如模糊聚类分析(fuzzy clustering analysis)和数量分支分析等。目前,有关重叠的、外在的和非等级的分类方法的研究正在逐步发展之中。^[10,61]

第三节 分类运算单位与分类性状

分类运算单位,或译运筹的分类单位(operational taxonomic unit, OTU)和分类性状(character)是数量分类学中最基本的两个术语。

一、分类运算单位

在传统的生物分类学中,物种(species)一直被认为是最基本的分类单位。分类学家们往往将需要鉴定的蜡叶标本与模式标本或原始描述中的特征作详细比较,来确定是否为新的物种。可见,理想的模式标本应当包含该物种的全部信息,并且有利于作客观比较。但实际并非如此。模式标本实际上也仅仅是相当多的个体中的一个随机样本而已,再加上环境差异以及采集人与鉴定人的不同等原因,不可避免地造成了分类学中的一些混乱情况。^[15]以作者之一关于慈姑属(*Sagittaria* L.)的系统研究为例。该属为水生单子叶植物泽泻科(Alismataceae)中的一个重要属,种类虽不多,但形态变异较大。其中,小慈姑(*S. potamogetifolia* Merr.)是1934年经美国植物学家梅里尔(E. D. Merrill)根据中国学者蒋英1929年采自广东信宜的标本鉴定后发表为新种的。但由于其他植物学家对该种的形态变异幅度缺乏全面的认识,导致各标本室对所藏的

标本错误鉴定,甚至存放在同一标本室的同号模式标本亦被另一位学者鉴定为慈姑(*S. sagittifolia* L.)。这种普遍存在的状况促使新的分类方法首先必须对模式标本概念加以修正。^[17]

本世纪 40 年代起,新的系统学中开始引入居群(population)的概念和方法,认为一份标本不可能代表某一物种的全部特征,而仅仅是居群中个体的样本。为了全面地认识一个物种,必须通过不同居群中大量个体的观察来获取典型的性状纪录,才有可能进行客观的有价值的分类工作。此后,居群生物学(population biology)作为物种生物学(biosystematics)的一个重要分支学科蓬勃发展起来。

虽然居群方法具有重要的意义,但是在数量分类研究中广泛应用还存在一些困难。首先,居群中个体数目较多时,统计工作变得十分复杂,而且电子计算机因容量所限,也难于处理过于庞大的数据量;其次,对于很多科、目等水平的分类工作来说,都以居群作为基本单位来统计资料似无必要。基于这些考虑,数量分类学中采取了一种更灵活的方法,即定义一个抽象的基本运算单位——分类运算单位(OTU)。

根据具体工作的需要,OTU 可以是个体、居群和种、属……等传统分类中的任一等级,也可以是上述等级中若干分类群的代表(如取平均值等)或其它抽象的概念。这样,数量分类的数学方法将更具有普遍的意义。^[58,60]

二、分类性状

数量分类与传统分类一样,必须以区分不同事物的特征或属性作为分类的依据,这些特征或属性就是分类性状。具体用于分类研究的每个性状称为单位性状(unit character),或简称性状。而 OTU 在每一个性状上所呈现出来的状况或数值称为性状状态(character state)。在生态学等领域中,它们被称为属性(attribute)和属性状态(attribute state)等。^[5]

例如,以叶片的边缘作为性状,常见的状态有全缘、波状、深波

状、钝齿状、锯齿状……、深裂状等类型。又如,若将生态类型作为性状,中国慈姑属植物中该性状有三种不同状态:冠果草(*S. guyanensis*)呈浮叶状态,慈姑(*S. trifolia*)等呈挺水状态,而矮慈姑(*S. pygmaea*)等则呈沉水状态。

有关分类性状的选取与量化方面的技术是数量分类学中至为关键的问题,下章将详细讨论。

第四节 分类分析

数量分类的发展使之逐渐形成一个新的数学分支学科——分类分析。分类分析通常包括三大类处理数据的多元分析方法:聚类分析(cluster analysis, 或 clustering analysis)、排序分析(ordination analysis)和判别分析(discriminatory analysis)。其中以聚类分析为核心内容。这里先分别简介如下:

一、聚类分析

聚类分析,或译群分析、点群分析、簇分析和簇群分析等,其中中心思想是运用数学方法研究 OTU(或性状)之间的亲疏程度,以此为依据将一批 OTU(或性状)聚合为若干 OTU 组(或性状组)。

聚类分析的研究大约始于本世纪 30 年代数学家们从相关矩阵中提取互相关组的工作。到 50 年代已逐渐形成聚类分析的雏型,同时许多学科领域中开始应用。直到数量分类学诞生以后,聚类分析研究方进入鼎盛时期。随着电子计算机技术的发展,它在实际应用中取得了显著的成绩。反过来,实际应用中出现的问题又导致越来越多的新方法出现。从数学意义上看,由于聚类分析不是从过去的数学分析中派生出来的,所以具有独特的风格。例如,它将数理统计的概念和几何学的语言结合在一起使用。当然,聚类分析的理论基础至今仍不完善,一些方法也较为粗糙,仍是一个值得深入探讨的学科领域。^[3,4,9]