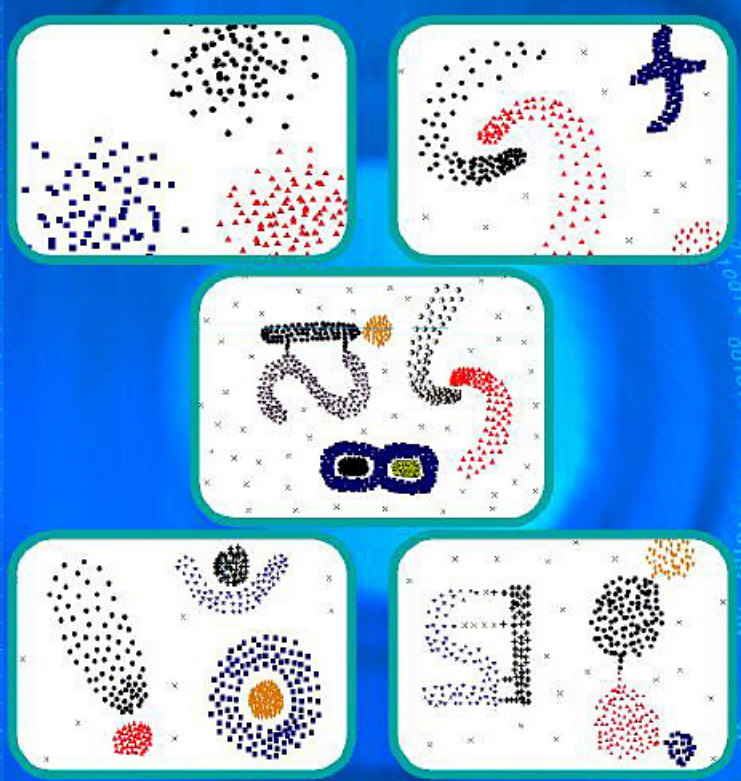
地球观测与导航技术丛书

空间聚类分析及应用

邓 敏　刘启亮　李光强　黄健柏 著



科学出版社

地球观测与导航技术丛书

空间聚类分析及应用

邓　敏　刘启亮　李光强　黄健柏　著

科学出版社

北　京

内 容 简 介

空间聚类分析是空间数据挖掘与知识发现的主要手段之一,已广泛应用于地理学、地质学、气象学、地图学、天文学及公共卫生等诸多领域。本书系统阐述了空间聚类分析的理论框架,并对当前国内外空间聚类分析领域研究的主要内容与进展进行了介绍。书中首先阐述了空间聚类分析研究的重要意义,明确了空间聚类分析研究中的基本问题,建立了空间聚类分析的理论框架,并据此对空间聚类分析的各个主要研究内容分别进行阐述,主要包括空间数据清理与聚类趋势分析、空间相似性度量、空间点实体聚类算法、空间面实体与动态轨迹聚类算法及空间聚类有效性评价方法等内容,同时介绍了空间聚类分析方法在地震模式分析、气象、环境、社会经济等领域的具体应用实例。

本书可供地理、地质、测绘、计算机、环境等相关领域的科研人员与研究生阅读参考。

图书在版编目(CIP)数据

空间聚类分析及应用 / 邓敏等著. —北京:科学出版社,2011.10
(地球观测与导航技术丛书)

ISBN 978-7-03-032533-4

Ⅰ.①空… Ⅱ.①邓… Ⅲ.①地理信息系统:空间信息系统—数据采集
Ⅳ.①P208

中国版本图书馆 CIP 数据核字(2011)第 206611 号

责任编辑:孙 芳 / 责任校对:鲁 素
责任印制:赵 博 / 封面设计:鑫联必升

科学出版社 出版
北京东黄城根北街 16 号
邮政编码:100717
http://www.sciencep.com
中国科学院印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2011 年 10 月第 一 版 开本:787×1092 1/16
2011 年 10 月第一次印刷 印张:11 3/4
印数:1—2 000 字数:261 000

定价:50.00 元
(如有印装质量问题,我社负责调换)

《地球观测与导航技术丛书》编委会

《地球观测与导航技术丛书》出版说明

地球空间信息科学与生物科学和纳米技术三者被认为是当今世界上最重要、发展最快的三大领域。地球观测与导航技术是获得地球空间信息的重要手段，而与之相关的理论与技术是地球空间信息科学的基础。

随着遥感、地理信息、导航定位等空间技术的快速发展和航天、通信和信息科学的有力支撑，地球观测与导航技术相关领域的研究在国家科研中的地位不断提高。我国科技发展中长期规划将高分辨率对地观测系统与新一代卫星导航定位系统列入国家重大专项；国家有关部门高度重视这一领域的发展，国家发展和改革委员会设立产业化专项支持卫星导航产业的发展；工业与信息化部和科学技术部也启动了多个项目支持技术标准化和产业示范；国家高技术研究发展计划（863 计划）将早期的信息获取与处理技术（308、103）主题，首次设立为“地球观测与导航技术”领域。

目前，“十一五”计划正在积极向前推进，“地球观测与导航技术领域”作为 863 计划领域的第一个五年计划也将进入科研成果的收获期。在这种情况下，把地球观测与导航技术领域相关的创新成果编著成书，集中发布，以整体面貌推出，当具有重要意义。它既能展示 973 和 863 主题的丰硕成果，又能促进领域内相关成果传播和交流，并指导未来学科的发展，同时也对地球观测与导航技术领域在我国科学界中地位的提升具有重要的促进作用。

为了适应中国地球观测与导航技术领域的发展，科学出版社依托有关的知名专家支持，凭借科学出版社在学术出版界的品牌启动了《地球观测与导航技术丛书》。

丛书中每一本书的选择标准要求作者具有深厚的科学研究功底、实践经验，主持或参加 863 计划地球观测与导航技术领域的项目、973 相关项目以及其他国家重大相关项目，或者所著图书为其在已有科研或教学成果的基础上高水平的原创性总结，或者是相关领域国外经典专著的翻译。

我们相信，通过丛书编委会和全国地球观测与导航技术领域专家、科学出版社的通力合作，将会有一大批反映我国地球观测与导航技术领域最新研究成果和实践水平的著作面世，成为我国地球空间信息科学中的一个亮点，以推动我国地球空间信息科学的健康和快速发展！

李德仁

2009 年 10 月

前　言

聚类分析是人类认识客观事物最朴素、最常用的手段之一。早在几千年前的《易经》、《战国策》等经典著作中，就已经出现了聚类分析的思想。聚类分析作为一个专门的研究领域提出至今仅有七十年左右，但其发展速度确实极为惊人，几乎遍及了所有数据处理的领域。伴随着地理信息技术的提出与发展及对地观测能力的迅速提升，1994 年，李德仁院士在国际上首先提出了从空间数据库中挖掘知识的思想，即空间数据挖掘与知识发现。自此，聚类分析在空间信息领域得到了空前的发展，空间聚类分析应运而生并已经成为空间数据挖掘与知识发现的主要技术手段之一。

自 20 世纪 90 年代，第一个专门用于空间聚类分析的算法——CLARANS 被提出以来，空间聚类分析已成为地理信息科学与计算机科学领域共同关注的热点研究问题之一。目前，国内外学者已经在空间聚类分析领域取得了可喜的研究成果，其应用领域也正逐步扩大，但需要注意的是，空间数据的复杂特性（如空间数据的几何形态特征、空间关系、空间数据的相关性、异质性及多尺度特性）导致空间聚类分析研究要比传统的聚类分析研究更加复杂。因此，空间聚类分析并不是传统的聚类分析技术在空间信息领域的简单套用，而需要开展专门的研究。目前，国内外已出版多部有关聚类分析的研究著作，但还没有专门针对空间聚类分析的学术著作出现。为此，结合作者多年来在空间数据挖掘领域的研究心得与成果，尝试撰写一本专门介绍空间聚类分析的学术著作。本书旨在明确空间聚类分析的基本问题与理论框架，同时对现有国内外学者的研究成果进行梳理，并介绍作者在空间聚类分析领域的一些研究成果与见解，希望能够促进空间聚类分析研究的深入与应用。

本书基于空间数据的基本特征、性质及聚类分析的内涵，首先提出了以“空间数据清理与聚类趋势分析-空间聚类算法设计-空间聚类有效性评价”为核心的空间聚类分析研究框架，进而针对具体内容分别进行阐述，同时对空间聚类分析所涉及的专门技术（如空间相似度量）进行了介绍。在内容设置上，为使读者对空间聚类分析的研究领域有较为全面的了解，本书专门回顾了国内外学者针对该领域的各个研究内容所取得的代表性成果，在此基础上，介绍了作者针对新的应用需求所发展的空间聚类分析算法与软件成果。本书从数据和应用两个角度分门别类地对空间聚类算法进行阐述，其中，依据数据的特征介绍了点、线（如动态轨迹）、面的空间聚类算法；并从应用的角度介绍了二维空间实体空间聚类、顾及空间障碍的空间聚类及顾及专题属性的空间聚类；还介绍了大量的算法实例及在地震、气象、环境、社会经济等领域的应用实例，以及自主开发的一款空间聚类分析软件系统 EasyCluster。

本书出版受到了国家 863 计划项目（2009AA12Z206）、教育部新世纪人才支持计划（NCET-10-0831）、中南大学前沿研究计划（2010QYZD002）及中南大学升华育英计划优秀人才资助项目的联合资助。感谢石岩、孙前虎、彭东亮、林雪梅等硕士研究生为本书所

作的部分算法实现与软件开发工作。感谢梅小明老师、赵玲老师、刘慧敏老师、赵彬彬博士、陈杰博士在撰写过程中给予的有益建议。感谢英国伦敦大学程涛教授、王佳璆博士对本书出版的帮助。

本书的出版也得到了中南大学各级领导的关心与支持。感谢中南大学人事处唐忠阳副处长、科技部吴厚平副部长、李启厚副处长、研究生院刘少军副院长、陈立章主任给予的鼓励与帮助！感谢中南大学图书馆馆长、地球科学与信息物理学院副院长朱建军教授、地球科学与信息物理学院副院长柳建新教授、刘兴权教授、邹峥嵘教授等学院领导在本书撰写过程中给予的指导、关心和支持！

空间聚类分析研究方兴未艾，本书的出版仅能起到抛砖引玉的作用。虽然本书撰写过程中力求尽善尽美，但限于作者的学识与经验，不妥之处在所难免，敬请读者批评指正。

作　者

2011 年 6 月

目　　录

第 1 章　绪　　论

1.1　空间聚类分析的产生

从人类诞生之日起，其认识、适应、改造自然的步伐就从未停止。聚类与分类是人类认识自然最基本的、最有效的技能之一，在人类社会的发展历程中发挥了重要的作用(Everitt et al.，2001；Anderberg，1973)。当人们试图去认识一类新事物或新现象时，往往会首先采用一定的特征去描述它们，然后根据一定的准则去跟其他已知的事物或现象进行比较(Xu et al.，2009)。例如，自然界大致可以分为动物、植物和矿物三界，动物界下设有纲、目、科、属、种等 5 个级别，我们通过这样的分类可以很容易发现其共性特征，如鱼类在水中生存，鸟类能够飞翔，同时也可以通过这种分类去推测具体动物的生活习性，如当我们看到喜鹊落在屋顶时，可以自然联想到它在天空中飞行的情景，而不需要看到它在飞行。

历史上，我国劳动人民最早将聚类分析的思想应用到实践中，五千多年前的《易经》就已经提出了“物以类聚，人以群分”的认识思想。公元前 4 世纪，齐国著名谋士东莱人(今山东省龙口市)淳于髡进一步阐述了这一观点，并通过实例说明了聚类思想的重要价值。在长期的生产实践中，聚类分析的思想一直是以一种经验的形式出现在人类的日常生活中。例如，水果根据颜色、大小分成不同的种类，并且通常可以卖出不同的价钱。聚类分析真正得到空前的发展还要得益于数学方法的引入。1939 年，Tryon 首次采用聚类分析的思想从相关矩阵中提取互相关的组，标志着聚类分析学科的正式提出。与传统的分类学不同，聚类分析是采用数学工具来研究类的划分及各类间的异同，而传统的分类则多是借助经验或专业知识。在随后的几十年里，聚类分析技术得到了飞速发展，一些沿用至今的经典算法被相继提出，如 50～60 年代提出的 K-means 聚类算法在 2006 年 IEEE 国际数据挖掘大会组织的评选中被评为十大最具影响力的数据挖掘算法之一(Wu et al.，2008)。聚类分析在 70 年代初首先由数学地质学家引入我国，中国科学院方开泰等(1982)首先开展了较为系统的研究，并于 80 年代初编纂出版了我国第一本系统介绍聚类分析技术的学术著作《聚类分析》。

20 世纪 80 年代，伴随着数据库技术与数据采集技术的突破性进展，数据的爆炸性增长，使人们面临了所谓“数据丰富，但信息贫乏”的困境，人们迫切需要强有力的工具从存储在大型数据库中的海量数据获取有用的信息或知识(Han et al.，2005；Tan et al.，2005)。为走出这种困境，数据挖掘技术应运而生。1989 年，在美国底特律市召开的第 11 届国际人工智能学术会议首次提出了“从数据库中发现知识(knowledge discovery in databases，KDD)”的概念。数据挖掘即从数据集中识别出有效的、新颖的、潜在有用的、最终可理解模式的信息，其主要手段包括聚类分析、关联规则挖掘、分类与预测、异常探测

及演变分析等(Han et al. ,2005;Fayyad et al. ,1996)。聚类分析既可以作为一种独立的数据挖掘工具,又可以与其他数据挖掘方法结合使用,挖掘更深层次的知识,提高数据挖掘效率,其已成为数据挖掘研究中的热点课题。

实际上,人们日常生活所接触和利用的现实世界数据中,大约有 80% 与地理位置、属性及其空间分布有关(李德仁等,2000)。伴随着计算机、网络、全球定位系统(global positioning system,GPS)、遥感(remote sensing,RS)及地理信息系统(geographical information system,GIS)等技术的迅猛发展和应用,空间数据的数量、复杂性及传输速度都在快速增长,其膨胀速度也极大超过了常规事务型数据(李德仁等,2006),尤其是近十年来对地观测技术与空间数据基础建设的突破性进展,空间数据的爆炸性增长已成为数据处理与分析领域的一个重要特征。空间数据除了具有属性特征外,还具有空间位置特征、空间关系特征和时间特征(王家耀,2001)。因此,与传统的事务性数据相比,空间数据更为复杂,具体表现为空间数据的海量性、空间属性间的非线性关系、尺度特性、模糊性、高维特性及数据的缺值(裴韬等,2001)。传统的空间分析主要是针对空间实体及其属性特征,采用统计分析的手段进行分析,其仅仅考虑了与样本性相关的统计量,没有顾及这些样本在地理空间的分布特征和相互间的位置关系,从而并不非常适用于处理空间相关的数据(Koperski,1999;郭仁忠,1997)。因此,虽然以 GIS 数据库为主体的空间数据库得到了极大发展,但由于缺乏高效、精确、科学的手段分析这些数据,使得空间数据同样面临了数据丰富而分析不足的尴尬局面,造成了空间数据的极大浪费(马荣华等,2007)。

鉴于数据挖掘技术在事务性数据库中的成功应用,从空间数据中挖掘知识已引起了国内外学者的广泛关注。1994 年,李德仁院士在加拿大渥太华举行的 GIS 国际学术会议上首次提出了从 GIS 数据库中发现知识的概念,并系统地分析了空间数据挖掘的特点和方法,这标志着空间数据挖掘理论的正式提出(Li et al. ,1994)。国际上,Han、Miller、Ester及 Shekhar 等学者也对空间数据挖掘开展了较早且持续的研究(Miller et al. ,2009;Shekhar et al. ,2005,2003;Ester et al. ,2000,1997;Han et al. ,1997)。1994 年,Ng 和 Han 提出了一个专门针对空间数据库的聚类算法——CLARANS(clustering large applications based upon RANdomized search),这标志着空间数据挖掘研究的正式兴起。空间数据挖掘是从空间数据库中提取隐含的、用户感兴趣的空间和非空间的模式、普遍特征、规则和知识的过程,已成为数据挖掘领域的一个崭新分支(邸凯昌,2000)。空间聚类、空间关联规则挖掘、空间分类、空间演变与预测、空间异常探测等构成了空间数据挖掘的主要研究内容。因此,空间聚类是空间数据挖掘的一个主要研究方向,也是传统聚类分析技术在空间数据库中的进一步应用,尤其是从空间数据中挖掘聚集模式符合人类对客观世界的认知方式,可以显著提高人们对空间数据的分析和认识水平。

1.2 空间聚类分析的研究概况与基本问题

1.2.1 空间聚类分析的研究概况

空间聚类分析既可以发现隐含在海量数据中的聚类规则,又可以与其他的空间数据挖掘方法结合,挖掘更深层次的知识,提高空间数据数据挖掘的效率和质量(杨春成,

2004)。空间实体的自然聚集现象经常反映一定的规律或趋势。《战国策·齐策三》中,淳于髡在解释“物以类聚,人以群分”时用到这样一个例子:“人们要寻找柴胡、桔梗这类药材,如果到水泽洼地去找,恐怕永远也找不到;要是到梁文山的背面去找,那就可以成车地找到。”柴胡、桔梗这种聚集现象实际上反映了周围环境的特殊性,继而对于人们认识、利用这种特性提供了重要的指示作用。1854 年,琼·斯诺博士采用空间聚集分析的手段发现伦敦霍乱病起源的案例堪称空间聚类分析最早的成功应用(Miller et al.,2009),当琼·斯诺博士将霍乱病死者居住位置标注在一张 1∶6500 比例尺的城区地图上后(图 1.1),发现死者大多集中在一口名为“布洛多斯托”的水井附近(图 1.1 中圆圈处),当关闭这口井后,新的霍乱病例也就没有再出现。

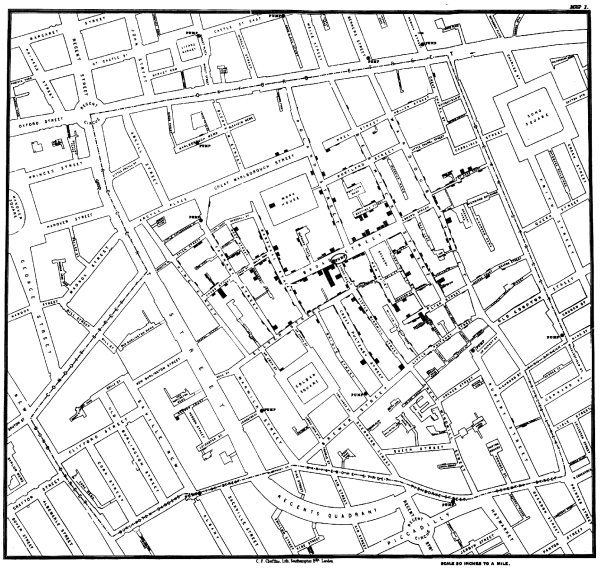


图 1.1　1854 年伦敦霍乱病空间分布图

空间聚类分析的一个重要作用在于能够发现空间实体自然的空间集聚模式,对于揭示空间实体的分布规律、提取空间实体的群体空间结构特征、预测空间实体的发展变化趋势具有重要的作用。进一步,结合空间实体的非空间属性在空间上的分布与差异,对于解释复杂的地理现象具有重要的意义。在城市规划领域,空间聚类在公共设施选址中具有明显的优势,并且已经得到了成功的应用(Liao et al.,2008;毛政元等,2004)。在制图综合领域,空间聚类已被广泛应用于点群特征简化、点群空间特征提取、建筑物聚合操作及等高线简化(武芳等,2008;Qi et al.,2008;Li,2007;郭庆胜等,2007;卢林等,2005)。在地震分析领域,空间聚类在提取地震空间分布特征及地质构造方面也体现出了独特的优势(Pei et al.,2009,2006;Xu et al.,1998)。在地价评估领域,空间聚类技术已被成功用于地价的分级(焦利民等,2009;邓羽等,2009)。在图像处理领域,空间聚类方法同样成功应用于遥感影像分类、分割研究中(秦昆等,2008;骆剑承等,1999;Sander et al.,1998)。在全球气候变化研究领域,借助空间聚类手段发现对陆地气候具有显著影响的极地、海洋大气压力模式、海表气温分布对于理解全球气候具有重要的价值(Birant et al.,2007;Tan et al.,2005)。在公共安全领域,犯罪热点分析是空间聚类分析对社会安全的又一个贡献,

可以有力地帮助警察对地方治安维护做出决策(Estivill-Castro et al. ,2002)。近年来,空间动态轨迹聚类已成为空间聚类技术的一个新的应用,借助空间聚类技术可以发现热带风暴等空间轨迹数据的空间分布模式,这对于理解局部气候变化具有重要的意义(Lee et al. ,2007)。

空间聚类不仅可以单独作为一种数据分析的手段,而且还可以作为其他空间数据挖掘方法的重要基础。例如,采用空间聚类分区预处理后构建神经网络的精度比全局构建的神经网络具有更强的预测能力(李光强等,2009;王海起等,2008)。空间聚类结果可以作为空间关联分析的输入来挖掘空间关联规则(Malerba et al. ,2002;Koperski et al. ,1995)。例如,采用空间聚类分析获取居民区边界,再分析聚类边界与高尔夫球场的邻接关系,继而预测房屋价格走向(Knorr et al. ,1996)。空间聚类分析也可以用来指导空间分类,建立分类模型,再对遥感影像进行分类,其效率和精度将大大提高(Cihlar et al. ,2003;Faber,1994)。空间聚类也是挖掘空间异常的一种有力手段(邓敏等,2010;李光强等,2008;林甲祥等,2008)。

此外,空间聚类算法在大多数情况下可以直接或稍加修改后进行传统事物型数据的聚类分析,在智能计算、机器学习、模式识别、生物学、心理学、信息检索、经济学等领域进行应用。鉴于聚类分析的巨大应用潜力,当前针对聚类分析的研究已经进入了高速发展的时期。据统计,在过去 10 年间,全球有超过 12000 篇期刊或会议论文在题目、摘要或关键词中包含聚类分析的字眼(Xu et al. ,2009),同时,这些论文涉及的范围也是极其广泛的,囊括了超过 200 个主要学科及 3000 多种杂志。从 1996 年到 2006 年的 10 年间,与聚类分析有关的论文几乎呈指数增长(图 1.2)。此外,目前有近 80 种主要期刊在刊登关于聚类分析及空间聚类的文章,近 50 个国际会议接收聚类分析有关的论文(Gan et al. ,2007)。作为聚类分析研究的一个重要分支,空间聚类分析已成为地球信息科学与计算机科学领域共同关注的热点。

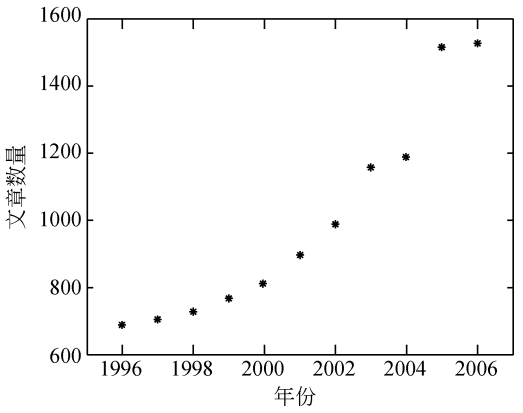


图 1.2　1996～2006 年每年发表的与聚类分析有关的科技论文(Xu et al. ,2009)

1.2.2　空间聚类分析的定义

空间聚类分析是传统聚类分析研究的一个延伸与发展,因此,有必要回顾聚类分析的

相关定义。目前,尚缺乏一个正式的、公认的聚类分析定义,下面选取了4种较有代表性的定义:

(1)聚类分析旨在将一组实体依据一定的相似性度量准则划分成一系列较为均匀的子类,同一类中实体间的相似度要尽可能大于不同类间的实体(Bacher et al.,1981)。

(2)聚类分析旨在发现 K 个簇或者一种包含 K 个簇的划分方式,相同类中的实体互相相似,而不同类中实体是不相似的(Bacher,1996)。

(3)聚类分析的目的在于将一个有限的、未标记的数据集分解成一系列有限的、自然的潜在数据结构,而不需要提供一个由同概率分布获得的未观测样本的精确分类(Baraldi et al.,2002;Cherkassky et al.,1998)。

(4)将物理或抽象对象的集合划分成相似对象的过程称为聚类(Han et al.,2005)。

首先要明确一个非常容易混淆的概念,即聚类与分类,Cherkassky 等(1998)对聚类分析的定义很好地回答了这个问题。分类系统分为监督分类与非监督分类两种,两者的根本区别在于是否将实体分配到一个预先设计好的分类系统中去。聚类分析不需要预设的分类系统,因此属于非监督分类;而我们通常所熟悉的分类则属于监督分类的范畴。综合上述定义,可以给出一个更为完整的空间聚类分析定义,即空间聚类旨在将一组具有相关性的空间实体依据一定的相似性度量准则划分成一系列由若干空间实体构成的、具有一定意义的空间簇,同一空间簇中实体尽可能相似,不同空间簇内的实体尽可能相异。可见,空间聚类分析的定义并不是传统聚类分析定义的简单套用,两者之间具有明显的差别,其集中体现在实体的定义、相似性定义及类的定义三个方面,下面将分别进行比较分析。

在传统的聚类分析中,对象的概念有多种不同的说法,如数据、元组、记录、观测资料、项目等,通常是对应了数据库中一条包含多个属性的记录,习惯性地抽象为空间中的一个点(Gan et al.,2007),当属性维数超过3后,这种抽象就失去了物理意义,因此,聚类结果也很难可视化。然而,在地理空间中,空间对象(亦称实体)总是有明确的物理意义的,并具有一定的空间位置,通常采用特定的坐标表示,同时也具有几何特征,即大小、形状、分布等。属性特征是附加于空间实体之上的,一般称之为专题属性。因此,不管属性的维数多高,空间实体本身总是可以在地理空间中唯一表示,故空间聚类的结果总是可以进行可视化表达。同时需要指出的是,不是所有的空间数据库都可以运用空间聚类方法,空间实体进行空间聚类的先决条件是它们之间存在一定的相关性。空间聚类必须在满足地理学第一定律(Tobler,1970)的前提下才能进行,即空间实体之间具有一定的依赖关系。相似性的定义在聚类分析中起关键作用,在传统的聚类分析中多采用各种距离、相关系数等度量实体间的相似性;而在空间聚类中,相似性的定义包含了两方面的意义:一种是属性上的相似,这与传统的聚类分析类似;另一种是空间关系上的相似,即要求空间实体在位置上接近或相邻,这是传统的聚类分析所不考虑的。簇也称为组、类,当前还没有一个公认的定义,但从根本上簇的定义是基于相似性的。定义簇的主要原则是要求同一个簇中的实体要尽可能相似,而不同簇中的实体要存在较大差异,内部均匀性和外部分离性是簇的主要特征(Everitt et al.,2001;Gordon,1999;Hansen et al.,1997)。空间聚类中,簇的定义的一个重要特点是要顾及空间关系与空间自相关,即实体间必须满足直接或间接的邻近关系,同时还要求簇内实体要满足空间自相关的条件,对空间不相关的实体进行聚类是

没有意义的。

可以将空间聚类形式化描述为：令 $\boldsymbol{S}=\{\boldsymbol{S}_1,\cdots,\boldsymbol{S}_i,\cdots,\boldsymbol{S}_n\}$ 表示一组具有空间相关性的空间实体集合；$\boldsymbol{S}_i=\{s_{i1},\cdots,s_{ij},\cdots,s_{im}\}$ 表示空间实体的特征向量；s_{ij} 表示空间实体 i 的一维属性；空间聚类获得 K 个空间簇，$\boldsymbol{S}=\boldsymbol{C}_1\cup\boldsymbol{C}_2\cup\cdots\boldsymbol{C}_i\cdots\cup\boldsymbol{C}_k$，$\boldsymbol{C}_i=\{\boldsymbol{S}_{i1},\cdots,\boldsymbol{S}_{ij},\cdots,\boldsymbol{S}_{it}\}$；$\mathrm{Similar}(\boldsymbol{S}_{mi},\boldsymbol{S}_{nj})$ 表示第 m 个空间簇中第 i 个实体与第 n 个空间簇中第 j 个实体的相似度。因此，对于空间聚类结果 $\boldsymbol{C}_1,\cdots,\boldsymbol{C}_i,\cdots,\boldsymbol{C}_k$，需满足下列条件：

(1) $\bigcup_{i=1}^{k}\boldsymbol{C}_i=\boldsymbol{S}$。

(2) 对于 $\forall\,\boldsymbol{C}_m,\boldsymbol{C}_n\subseteq\boldsymbol{S},m\neq n$，需要同时满足：

① $\boldsymbol{C}_m\cap\boldsymbol{C}_n=\varnothing$(仅针对硬聚类)；

② $\mathrm{MAX}\forall s_{mi}\in c_m,\forall s_{nj}\in c_n(\mathrm{Similar}(\boldsymbol{S}_{mi},\boldsymbol{S}_{nj}))<\mathrm{MIN}\forall s_{mx},s_{my}\in c_m(\mathrm{Similar}(\boldsymbol{S}_{mx},\boldsymbol{S}_{my}))$。

根据空间实体特征向量 $\boldsymbol{S}_i$ 的特点，又可以将空间聚类区分为以下三种类型：

(1) $\boldsymbol{S}_i$ 仅包含了空间位置属性。

(2) $\boldsymbol{S}_i$ 既包含空间位置属性，又包含专题属性。

(3) $\boldsymbol{S}_i$ 仅包括专题属性。

第(1)种类型的空间聚类分析可以用来发现空间实体的空间分布模式与规律；第(2)种类型综合考虑了空间位置与专题属性特征的双重意义，可以用于发现更深层次的地学规律；第(3)种类型仅考虑了专题属性的差异，需要结合空间实体的空间分布进行分析，在很大程度上退化为传统的聚类分析手段，在本书中将不做详细讨论。因此，本书将重点研究前两种类型的空间聚类分析方法。

1.2.3 空间聚类分析的基本框架

根据空间聚类的定义，一个完整的空间聚类分析过程包括以下 6 个部分：空间数据清理、空间聚类趋势分析、属性提取与相似性度量、空间聚类算法选择与设计、空间聚类有效性评价、空间聚类结果解释与应用。具体流程如图 1.3 所示。

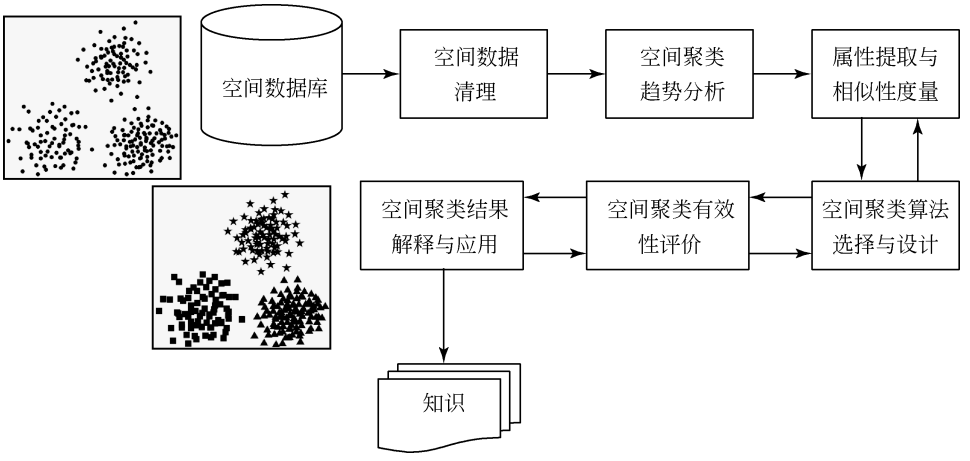


图 1.3 空间聚类的基本流程

空间聚类分析中的6个主要步骤具体可以描述如下：

(1)空间数据清理。空间数据可能有不同类型，格式也可能不尽统一，在不同的应用领域可能采用了不同的单位或格式，记录中的各个数据项可能不完整或失效，也可能有重复、错误和异常等质量问题(Koperski，1999)。空间数据的质量对数据挖掘的效果具有重要的影响，需要引起足够的重视(Wang et al.，1995)。空间数据清理主要包括不完备空间数据填补、不准确空间数据清理、重复记录空间数据清理、不一致空间数据处理等内容(李德仁等，2006)。

(2)空间聚类趋势分析。空间聚类趋势分析是目前空间聚类分析研究中涉及较少的一个内容，因为人们通常主观假设了数据是具有可聚性的，然而实际上却并非都是如此。因此，研究数据的可聚性对于合理运用空间聚类分析技术及获取有用的聚类信息大有裨益。与空间聚类问题相对应，空间聚类趋势分析也分为两种类型：① 二维空间点集聚类趋势分析；② 顾及专题属性的空间聚类分析。二维空间点集聚类趋势分析仅考虑空间实体的空间位置属性，旨在分析空间数据在空间上的位置聚集效应。空间点集的分布复杂，通常可以区分为均匀分布、随机分布及聚集分布三种情况，并且只有表现为聚集分布的数据集才适合进行空间聚类分析操作。现有空间数据分析中，点模式的分析方法可以为此类空间聚类趋势问题提供技术与方法支撑(Shekhar et al.，2009)，如样方法、核密度估计法、最邻近指数法、K-函数法及G-函数法等(王远飞等，2007)。此外，基于可视化的聚类趋势分析方法也可以进行直接应用。顾及专题属性的空间聚类趋势可以采用现有的空间自相关指数进行度量(Aldstadt，2009)，如Moran's I指数、Geary's C指数、LISA指数及G指数等(王远飞等，2007)。

(3)属性提取与相似性度量。空间聚类分析中的属性提取主要包括两方面内容：一方面，根据不同的研究目的确定参与聚类分析的属性，分为三种情况：① 仅是空间属性；② 仅是专题属性；③ 两者兼顾。另一方面，从原始专题属性中能最大程度反映其特征属性，这些属性可能是多个专题属性的集成，也可能是专题属性的一个特征子集。相似性度量即根据属性变量的特征选择相应的度量准则，如各种距离、相关性等。

(4)空间聚类算法选择与设计。针对不同的应用目的，目前国内外学者已经提出了众多的空间聚类算法，但没有一种方法可以解决所有的空间聚类问题。空间聚类算法的选择具有很大的主观性，因而需要对空间聚类中的问题进行归纳和总结，有利于选择合适的算法或者发展针对性的算法(Xu et al.，2009)。

(5)空间聚类有效性评价。针对同一个空间数据集，不同的算法或者同一算法不同参数的设置通常会获得不同的聚类结果。对空间聚类的结果进行有效地量化评价，有利于用户更好地选择和使用空间聚类的分析手段。空间聚类结果的有效性评价必须遵循客观的原则，当前空间聚类有效性评价的方法主要可以分为内部评价法、外部评价法及相对评价法三类。

(6)空间聚类结果解释与应用。空间聚类的最终目的是使用户更好地认识和发现空间数据库中隐含的信息。结合相关领域的专家知识对空间聚类的结果进行分析和解释，有助于解决相关问题并辅助决策。同时，空间聚类分析所获得的对空间数据的划分并不是数据分析过程的终点，在此基础上可以进行其他空间数据挖掘方法或进行更深层次的空间数据分析。

1.2.4 空间聚类算法分类

对空间聚类算法进行分类有助于更好地认识和分析各种算法的特征与适用性,对于空间聚类算法设计与应用具有重要的促进作用。针对传统的事务性聚类分析方法,国内外学者(Xu et al.,2009;孙吉贵等,2008;Gan et al.,2007;Han et al.,2005;Tan et al.,2005;Kolatch,2001)曾从不同角度对其进行过分类。本书一方面借鉴已有分类方法,另一方面结合空间聚类的特点与最新发展,对空间聚类方法进行了分类(图 1.4),其主要依据了以下两条原则:

(1)空间实体的维数。空间实体根据维数不同可以分为点(0 维)、线(1 维)、面(2 维)、体(3 维),空间实体几何上的复杂性直接影响空间聚类算法的设计。虽然点的形式是最常见的,但在有些情况或特殊应用下,空间实体的几何形状是不能忽略的。因此,本书根据空间实体的维数特征将聚类算法区分为空间点实体聚类方法与空间扩展实体(线、面、体)聚类方法。

(2)空间聚类的主要思想与工具。依据空间聚类的主要思想和工具,如划分的方法借助实体函数与质心的概念、基于密度的方法的关键是密度的概念,可以将空间聚类算法区分为基于划分的算法、基于层次的算法、基于密度的算法、基于图论的算法、基于模型的算法、基于格网的算法及混合的聚类算法。此外,一些附加的要求(如空间障碍约束、用户约束等)实际上也是基于上述思想的扩展,故不单独进行区分。

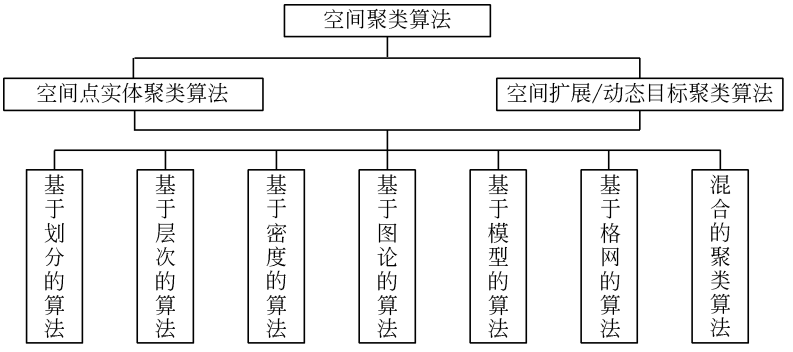


图 1.4 空间聚类方法分类

1.3 本书研究的主要内容

近十多年来,空间聚类分析在空间数据分析中的重要应用价值已经引起了国内外学者的广泛关注,并且已成为空间数据挖掘领域的重要基础研究内容。然而,现有的空间聚类研究脱胎于传统的信息科学领域,缺乏地学视角下的研究框架,需要从地球信息科学的角度对空间聚类研究的基本定义和意义进行明确。此外,研究空间聚类方法在地球信息科学领域的特殊要求、应用范围和特点,对于发展针对性的空间聚类算法,拓宽空间聚类的应用领域具有重要的价值。为此,本书的主要研究内容(图 1.5)分为如下几个部分:

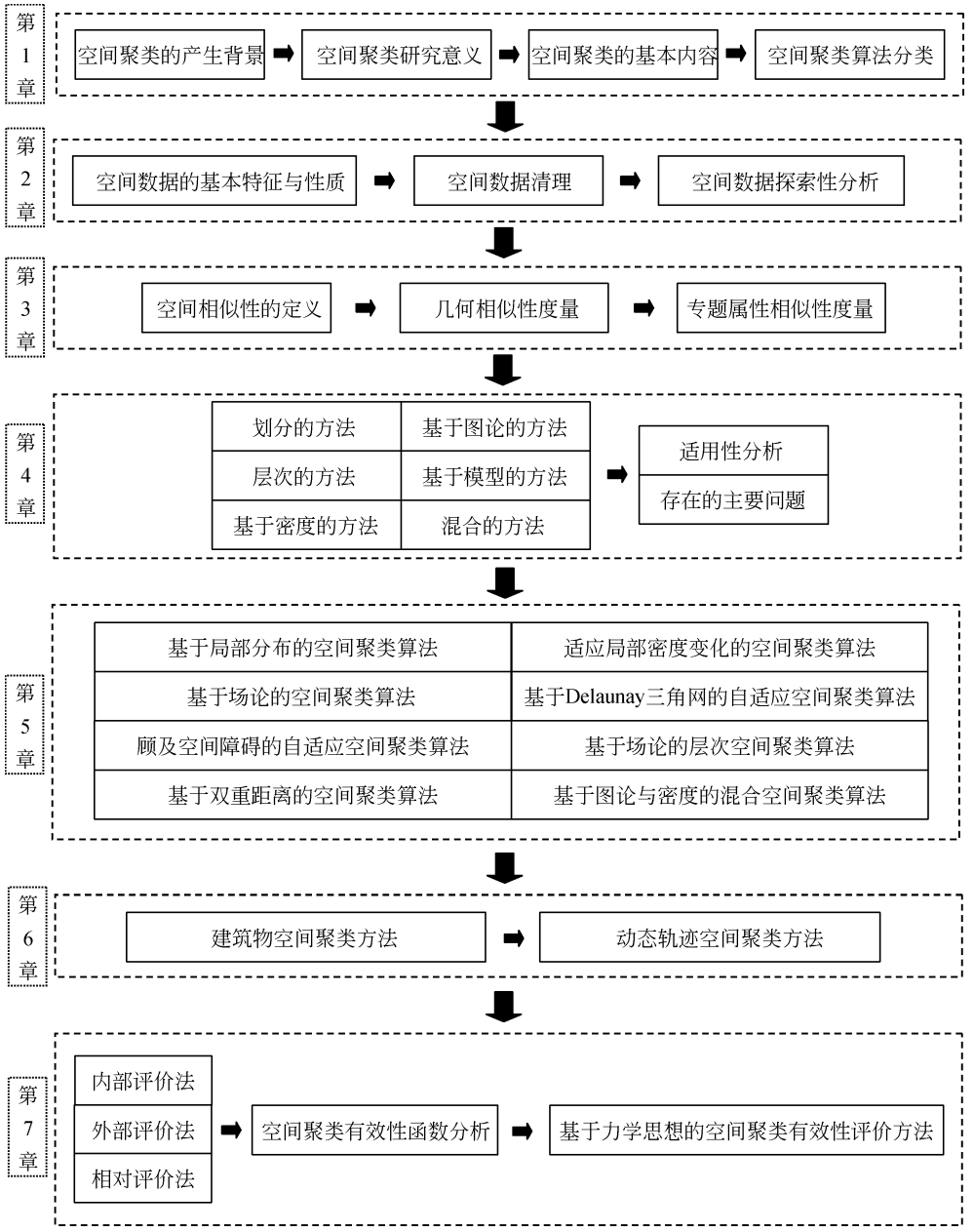


图 1.5　本书的基本内容与结构安排

(1) 空间聚类分析中的基本问题。阐述了空间聚类分析的产生背景和研究意义,分析了空间聚类分析与传统聚类分析方法的区别与联系,进一步给出了空间聚类的定义和基本研究内容,最后对现有的空间聚类方法进行了系统的归纳与分类。上述内容为空间聚类方法的深入研究奠定了重要的理论基础,同时也是第 1 章的主要内容。

(2) 空间数据探索性分析与预处理。根据空间聚类分析的流程,空间数据清理与空间自相关分析是空间聚类分析的重要准备环节,也是第 2 章的主要介绍内容。空间数据的基本特征与性质将首先进行阐述,在此基础上进一步介绍空间数据清理与探索性分析

的具体方法。

(3)空间相似性度量。空间聚类算法的一个核心环节便是空间实体间的相似性度量,这部分内容将在第3章中作具体介绍,主要包括空间相似性的定义、几何相似性度量与属性相似性度量三部分内容。

(4)现有空间聚类算法分析。这是第4章的主要内容,空间点实体的聚类方法是当前研究最广泛、成果最多的领域,本书将从两个方面进行介绍:一方面研究了当前8类主要空间聚类方法的特点与性质,并对各类方法中的典型算法进行分析,归纳空间聚类算法中的主要技术指标,继而对各类算法进行适用性分析,指出了当前空间聚类算法中存在的若干问题。

(5)空间点实体聚类算法。针对当前空间点实体聚类方法存在的局限,介绍了作者近年来开展的部分工作。对于空间二维点实体聚类分析、顾及空间障碍的空间聚类分析及顾及专题属性的空间聚类分析等不同应用需求,发展了8种改进的空间聚类算法。

(6)建筑物与动态轨迹空间聚类算法。第6章中主要介绍了空间扩展实体(线、面)的空间聚类方法,重点介绍了面实体层次约束自动聚群方法,并研究了空间动态轨迹聚类分析的理论与方法。

(7)空间聚类有效性分析。系统介绍了内部评价法、外部评价法与相对评价法等三类空间聚类有效性评价方法,并分析现有空间聚类有效性函数的构造方法与具体性能,同时介绍了一种基于力学思想的空间聚类有效性评价方法,这些内容构成第7章的主要研究内容。

此外,本书还给出了空间聚类分析方法在地震分析、空间分布模式挖掘、环境保护、地图自动综合、气候变化及社会经济发展等领域的应用实例,介绍了自主开发的空间聚类分析软件EasyCluster的主要功能与特点。

1.4 本章小结

本章首先介绍了聚类分析的发展历史及空间聚类分析的产生背景,进而分析了空间聚类分析技术在地学领域的研究现状及应用价值。在此基础上,对空间聚类分析研究中的基本问题进行了定义和归纳,如空间聚类分析的定义、类型、研究内容及分类。最后简要介绍了本书的主要研究内容及结构安排。

本章旨在对空间聚类分析的研究体系进行系统梳理,对空间聚类分析研究的目的、研究内容、定义及概念进行明确,这对于进一步深入研究空间聚类分析的理论、方法及应用问题具有重要的意义。

参考文献

邓敏,刘启亮,李光强.2010.采用空间聚类技术探测空间异常.遥感学报,14(5):951—958.

邓羽,刘盛和,张文婷,等.2009.广义多维云模型及在空间聚类中的应用.地理学报,64(12):1439—1447.

邸凯昌.2000.空间数据挖掘与知识发现.武汉:武汉大学出版社.

方开泰,潘恩沛.1982.聚类分析.北京:地质出版社.

郭庆胜,黄远林,郑春燕,等.2007.空间推理与渐进式地图综合.武汉:武汉大学出版社.

郭仁忠 .1997 .空间分析 .武汉 :武汉测绘科技大学出版社 .
焦利民 ,刘耀林 ,刘艳芳 .2009 .区域城镇基准地价水平的空间自相关格局分析 .武汉大学学报(信息科学版),34(7):873—877 .
李德仁 ,关泽群 .2000 .空间信息系统的集成与实现 .武汉 :武汉大学出版社 .
李德仁 ,王树良 ,李德毅 .2006 .空间数据挖掘理论及应用 .北京 :科学出版社 .
李光强 ,邓敏 ,程涛 ,等 .2008 .一种基于双重距离的空间聚类算法 .测绘学报 ,37(4):482—488 .
李光强 ,刘启亮 ,邓敏 .2009 .一种基于 BP 神经网络的空间异常探测方法 .测绘科学技术学报 ,26(6):439—448 .
林甲祥 ,陈崇成 ,樊明辉 ,等 .2008 .基于 MST 聚类的空间数据利群算法挖掘 .地球信息科学 ,10(5):586—591 .
卢林 ,吴纪桃 ,柳重堪 .2005 .基于特征的等高线数据聚类方法 .测绘学报 ,34(2):138—141 .
骆剑承 ,梁怡 ,周成虎 .1999 .基于尺度空间的分层聚类方法及其在遥感影像分类中的应用 .测绘学报 ,28(4):319—324 .
马荣华 ,蒲英霞 ,马晓冬 .2007 .GIS 空间关联模式发现 .北京 :科学出版社 .
毛政元 ,李霖 .2004 .空间模式的测度及其应用 .北京 :科学出版社 .
裴韬 ,周成虎 ,骆剑承 ,等 .2001 .空间数据知识发现研究进展评述 .中国图像图形学报 ,6(9):854—860 .
秦昆 ,徐敏 .2008 .基于云模型和 FCM 聚类的遥感图像分割方法 .地球信息科学 ,10(3):302—307 .
孙吉贵 ,刘杰 ,赵连宇 .2008 .聚类算法研究 .软件学报 ,19(1):48—61 .
王海起 ,王劲峰 .2008 .基于分区的局域神经网络时空建模方法研究 .遥感学报 ,12(5):707—715 .
王家耀 .2001 .空间信息系统原理 .北京 :科学出版社 .
王远飞 ,何洪林 .2007 .空间数据分析方法 .北京 :科学出版社 .
武芳 ,钱海忠 ,邓红艳 ,等 .2008 .面向地图自动综合的空间信息智能处理 .北京 :科学出版社 .
杨春成 .2004 .空间数据挖掘中的聚类分析算法研究[博士学位论文].郑州 :中国人民解放军信息工程大学 .
Aldstadt J .2009 .Spatial Clustering :Handbook of Applied Spatial Analysis .Berlin :Springer :279—300 .
Anderberg M .1973 .Cluster Analysis for Applications .New York :Academic Press .
Bacher E ,Jain A .1981 .A clustering performance measurement based on fuzzy set decomposition .IEEE Transactions on Pattern Analysis and Machine Intelligence ,PAMI-3(1):66—75 .
Bacher J . 1996 . Cluster Analyse . Anwendungsorientierte Einführung . 2nd ed . Munich/Vienna : Oldenbourg .
Baraldi A , Alpaydin E . 2002 . Constructive feedforward ART clustering networks-Part I and II . IEEE Transactions on Neural Networks ,13(3):645—677 .
Birant D ,Kut A .2007 .ST-DBSCAN :An algorithm for clustering spatial-temporal data .Data & Knowledge Engineering ,60(1):208—221 .
Cherkassky V ,Mulier F .1998 .Learning for Data :Concepts ,Theory and Methods .New York :Wiley .
Cihlar J ,Guindon B ,Beaubian J ,et al .2003 .From need to product :A methodology for completing a land cover map of Canada with landsat data .Can .J .Remote Sensing ,29(2):171—186 .
Ester M ,Frommelt A ,Kriegel H P ,et al .2000 .Spatial data mining :Database primitives ,algorithm ,and efficient DBMS support .Data Mining and Knowledge Discovery ,4(2—3):193—216 .
Ester M ,Kriegel H P ,Sander J .1997 .Spatial data mining :A database approach //Proceedings of SSD'97 :47—66 .
Estivill-Castro V ,Lee I .2002 .Multi-level clustering and its visualization for exploratory spatial analysis . GeoInformation ,6(2):123—152 .
Everitt B ,Landau S ,Leese M .2001 .Cluster Analysis .4th ed .London :Arnold .
Faber V .1994 .Clustering and the continuous k-means algorithm .Los Alamos Science ,22 :138—144 .

Fayyad U M, Piatetsky-Shapiro G, Smyth P. 1996. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining. New York: AAAI/MIT Press.

Gan G J, Ma C Q, Wu J H. 2007. Data clustering: Theory, algorithm and applications//ASM-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia.

Gordon A. 1999. Classification. 2nd ed. London: CRC Press.

Han J W, Kamber M. 2005. Data Mining: Concepts and Technique. 2nd ed. San Francisco: Morgan Kaufmann.

Han J W, Koperski K, Stefanovic N. 1997. GeoMiner: A system prototype for spatial data mining//Proceedings of the SIGMOD'97: 553—556.

Hansen P, Jaumard B. 1997. Cluster analysis and mathematical programming. Mathematical Programming, 79(1): 191—215.

Knorr E M, Ng R T. 1996. Finding aggregate proximity relationships and commonalities in spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 8(6): 884—897.

Kolatch E. 2001. Clustering algorithm for spatial databases: A survey. http://citeseer.ist.psu.edu/436843.html.

Koperski K. 1999. A progressive refinement approach to spatial data mining [PhD Dissertation]. Burnaby: Simon Fraster University.

Koperski K, Han J W. 1995. Discovery of spatial association rules in geographic information databases// Proceedings of the 4th International Symposium on Large Spatial Databases: 47—66.

Lee J G, Han J W, Whang K Y. 2007. Trajectory clustering: A partition and group framework//Proceedings of 2007 ACM-SIGOD International Conference on Management of Data, Beijing: 593—604.

Li D R, Cheng T. 1994. KDG-knowledge discovery from GIS//Proceedings of the Canada Conference on GIS, Ottawa: 1001—1012.

Li Z L. 2007. Algorithmic Foundation of Multi-scale Spatial Representation. New York: CRC Press.

Liao K, Guo D S. 2008. A clustering-based approach to the capacitated facility location problem. Transactions on GIS, 12(3): 323—339.

Malerba D, Esposito F, Lisi F A. 2002. Mining spatial association rules in census data. Research in Official Statistics, 5(1): 19—44.

Miller H J, Han J W. 2009. Geographic Data Mining and Knowledge Discovery. 2nd ed. New York: CRC Press.

Ng R, Han J W. 1994. Efficient and effective clustering method for spatial data mining//Proceeding of the 1994 International Conference on Very Large Data Bases: 144—155.

Pei T, Jasra A, Hand D J, et al. 2009. DECODE: A new method for discovering clusters of different densities in spatial data. Data Mining and Knowledge Discovery, 18(3): 337—369.

Pei T, Zhu A X, Zhou C H, et al. 2006. A new approach to the nearest-neighbor method to discover cluster features in overlaid spatial point processes. International Journal of Geographical Information Science, 20(2): 153—168.

Qi H B, Li Z L. 2008. An approach to building grouping based on hierarchical constraints. The International Archives of the Phtogrammetry, Remote Sensing and Spatial Information Science: 449—454.

Sander J, Ester M, Kriegel H P, et al. 1998. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery, 2(2): 169—194.

Shekhar S, Chawla S, Ravada A, et al. 2005. Spatial databases-accomplishments and research needs. IEEE Transactions on Knowledge and Data Engineering, 11(1): 45—55.

Shekhar S ,Lu C T ,Zhang P S .2003 .A unified approach to detecting spatial outliers .GeoInformation ,7 (2):139—166 .

Shekhar S ,Vatsavai R R ,Celik M .2009 .Spatial and Spatiotemporal Data Mining :Recent Advances .Next Generation of Data Mining .New York :CRC Press .

Tan P N ,Steinbach M ,Kumar V .2005 .Introduction to Data Mining .New York :Addison Wesley .

Tobler W .1970 .A computer movie simulating urban growth in the detroit region .Economic Geography , 46(2):234—240 .

Tryon R C .1939 .Cluster Analysis ,Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality .New York :Edwards Brother ,Inc . ,Litho Printers and Publishers .

Wang R ,Storey V ,Firth C .1995 .A framework for analysis of data quality research .IEEE Transactions on Knowledge and Data Engineering ,7(4):623—640 .

Wu X D ,Kumar V ,Quinlan J ,et al .2008 .Top 10 algorithms in data mining .Knowledge Information System ,14(1):1—37 .

Xu R ,Wunsch D II .2009 .Clustering .New Jersey :Wiley .

Xu X W ,Ester M ,Kriegel H P ,et al .1998 .A distribution-based clustering algorithm for mining in large spatial databases //Proceedings of the 14th International Conference on Data Engineering :324—331 .

第 2 章　空间数据清理与聚类趋势分析

2.1　引　　言

空间数据本身具有的特性导致了空间聚类分析与传统的聚类分析研究的巨大区别。理解和认识空间数据的基本特征与性质是进行空间聚类分析操作和设计空间聚类算法的重要基础。本章首先简要介绍了空间分析领域对空间数据本身特性的研究认识,进而针对空间聚类分析研究中的两个先决问题进行分析,即空间数据清理与空间聚类趋势分析。空间数据清理实际上是任何空间分析或空间统计研究所必需的一个环节,数据中的不完备或错误信息通常对分析结果具有较大的影响。空间聚类趋势分析是目前空间聚类分析领域较少涉及的一方面内容,因为人们通常主观假设了数据是具有可聚性的,然而实际上却并非都是如此。因此,研究数据的可聚性对于正确运用空间聚类分析技术是非常关键的,这里充分借鉴了传统空间数据分析的研究成果,系统地总结了可适用于空间聚类趋势分析的理论与工具,同时也对现有聚类分析趋势分析的专用方法在空间聚类分析中应用的可行性与使用性进行了分析。下面将首先阐述空间数据的基本特征与性质。

2.2　空间数据的基本特征与性质

空间数据用来表达具有确定的位置和形态特征,并具有地理意义的地理空间物体(郭仁忠,2001),亦可以认为空间数据是现实中存在的地理实体在地理空间中的投影。与传统的事务型数据相比,空间数据不仅包含表达空间位置的信息,还包含了极为丰富的内在关联与规律,这也是空间聚类研究的重要特色与难点所在。因此,要进行空间聚类分析研究,首先需要对空间数据的特征和性质进行全面的了解。

2.2.1　空间数据的基本特征

现有研究一般认为,空间数据具有 4 个最基本的特征,即空间特征、属性特征、时间特征与尺度特征(李光强,2009;王远飞等,2007;王劲峰,2006;李德仁等,2006;李霖等,2005;邬伦等,2002)。空间数据的这些特征直接导致了空间聚类分析与传统聚类分析的区别。

空间特征是空间数据最主要的特征,其表示空间实体的位置、几何特征及与其他空间实体的空间关系。空间实体的位置通常采用不同的坐标系统进行描述,如常用的大地坐标、空间直角坐标、极坐标及平面直角坐标等。空间实体的几何特征表示了空间实体的大小、形状及空间维度,据此可将空间实体区分为点、线、面、体及表面等 5 种类型。空间关

系描述了空间实体间的相互关系,在空间聚类研究中将起到关键作用,主要包括拓扑关系、方向关系及距离关系。空间拓扑关系主要描述了实体间包含、相交、相离等;空间方向关系如东西南北,上下左右等;空间距离关系如实体间的距离等。

属性特征描述了与空间实体紧密联系的用于表达空间实体本质特征的数据或变量,其可以从不同角度进行定义,通常分为定性与定量两种。de Smith 等(2007)从空间分析的角度区分了 5 种属性,分别为:① 标称属性,如果某个属性不需要任何排序与数学操作就可用于区分不同位置的空间实体,则其属于标称属性。例如,采用数字对土地类型命名,1 代表林地,2 代表草地,3 代表耕地等。② 次序属性,若一个属性代表了一种排序,即类别 1 比类别 2 好,则表示了次序属性。最常见的即土地等级的区分。③ 间距属性,如温度或高程领域的测量值,不同的值具有不同的含义,则可认为是间距属性。④ 比值属性,若两个测量值相除具有一定的含义,如一个人的体重是另一个人的两倍是有意义的,则属于比值属性。⑤ 周期属性,如角度、日期等属性通常具有周期性。前两种属于定性的范畴,后面三种是定量的。

时间特征描述了空间数据随时间变化的特性。通常使用的空间数据一般是某个时间点上的特征,不同时间点上空间实体主要包含两种情况:① 空间属性不变,专题属性发生变化,如各种环境监测站点数据;② 空间属性与专题属性同时发生变化,如台风监测数据。顾及时间特征的空间数据构成了更为复杂的时空数据,空间聚类分析一般不考虑空间数据在时态上的演变。

尺度特征是空间数据的另一个重要特征,其具体表现在不同的观察层次上,空间实体及其分布形态不尽相同(李霖等,2005)。尺度在不同的学科和应用领域具有不同的含义,这也是地理信息科学领域研究中一直以来的难点问题之一。广义的尺度包括了空间尺度、时间尺度与语义尺度。在地理信息科学领域,狭义的空间尺度主要指幅度与粒度两方面内容,前者表示了研究区域的空间幅度,后者表示了地理空间中的最小可辨识单元。本书在研究空间聚类分析的过程中,主要涉及空间尺度问题。

2.2.2 空间数据的基本性质

空间数据包含的众多特殊性质决定了空间聚类分析研究的特殊性,这里将重点介绍空间数据的三个基本性质:空间依赖性、空间异质性及可塑面积单元问题(modifiable area unit problem,MAUP)。

空间依赖性被认为是空间数据最基本的性质。正是由于空间依赖性的存在,直接导致了地理世界的有序性与连续性。空间依赖性产生的原因是极其复杂的,一般认为是由于空间数据的相互作用、扩散及各种测量的误差等造成的(王劲峰等,2010;王远飞等,2007)。Tobler(1970)的地理学第一定律在空间上的解释“空间上距离近的实体间的相似性比距离远的实体的相似性大”是对空间依赖性的定性描述。空间自相关可以对空间数据的空间依赖性进行定量的描述,常用的方法包括 Moran's I、Geary's C、Ripley's K、Join count 指数及半变异函数等(王远飞等,2007;王劲峰,2006;Haining,2003)。传统的统计学是基于独立同分布假设的,然而由于空间依赖性的存在,导致这种假设通常并不成立,这是造成空间统计学特殊性的根本原因。也正是由于空间依赖性的存在,导致空间聚类

分析与传统聚类分析研究具有重要的区别。因此,空间依赖性是空间聚类分析研究的基础。Tobler(1970)的地理学第一定律在地学领域的意义是极其深远的,其构成了空间分析及空间聚类分析的理论基础。

空间异质性是与空间依赖性相对应的另一个重要性质。空间异质性表现在空间数据间的差异性之中,即每个空间实体都有区别于其他空间实体的属性。空间数据所表现出的多样性、复杂性及非均质性均属于空间异质性的范畴。也有学者试图从空间异质性的角度提出地理学第二定律,与 Tobler 的地理学第一定律相对应(Goodchild,2003)。空间异质性定量分析方法主要包括局部 Moran's I、局部 Getis's C、地理加权回归等(王劲峰等,2010)。空间异质性源于局部的特殊性,表示空间现象在空间上是非平稳的,因此,空间异质性也称为空间非平稳性。空间异质性的存在导致在空间分析或空间聚类过程中需要提高对局部性质的分析和识别能力,否则很难保证结果的可靠性,甚至得出错误的结论。空间异质性的产生从根本上源自宇宙及地球系统在演化过程的分异结果,也是造成环境与物种多样性的重要原因。

可塑面积单元问题是空间数据表现出的一类特殊的性质,其表现为空间数据分析的结果随面积单元定义的不同而发生变化(邬建国,2007)。可塑面积单元问题主要体现在两个方面:① 尺度效应,即空间数据通过聚合操作使粒度发生变化时,其分析结果也随之变化;② 划区问题,即在同一粒度或聚合水平上,不同的划区方案将导致不同的分析结果(王远飞等,2007;邬建国,2007)。可塑面积单元问题最著名的记载当属西方政治选举过程中,通过修改选举区域的边界来改变选举结果的实例,而在空间数据分析领域,从 20 世纪 70 年代开始也逐渐受到重视。邬建国(2007)曾指出:可塑面积单元问题不应理解为一个"问题",因为它可能是真实系统的多尺度结果在空间上的反映。汪闽(2003)在其研究中指出,层次聚类分析与可塑面积单元问题存在内在的联系,然而在实际应用中,采用不同的相似性度量准则及聚类规则得到的层次聚类结果可能是不同的,因而也需要对可塑面积单元问题给予充分的重视。

2.3 空间数据清理

空间数据的来源是多方面的,主要方式可以分为原始数据与二手数据(龚健雅,2001)。原始数据一般是通过仪器直接采集的,如全站仪、GPS 及其他电子传感设备等;二手数据来源于地图、图书,国家(地区)的社会、经济、人口等统计数据,遥感影像等经过人为整理或处理。各种空间数据来源不同、格式也可能不统一,而且记录中也可能存在各种缺失、冗余及错误,由此带来的误差对于空间数据分析具有重大的影响,空间数据清理在空间数据挖掘研究中重要性已经受到了重视(李德仁等,2006)。因此,在进行各种空间数据分析时,需要首先对空间数据的质量进行控制。针对空间聚类分析的特点,空间数据清理的主要任务包括空间数据的完备化、冗余与重复消除及数据不一致性消除。

空间数据的完备化主要针对空间数据中的缺失记录,如由于监测设备故障造成的缺值、云雾遮蔽造成的遥感影像不完整。缺失的记录主要可以分为三类:完全随机缺失、不完全随机缺失及不可忽视的缺失。其中,不可忽视的缺失一般是由变量本身造成的,不能采用通常的缺值估计方法(王劲峰,2006)。空间数据中缺失记录的主要处理方法包括基

于完整记录的方法、基于替代的方法、权重的方法及基于模型的方法(Hentges et al.,1998)。此外,也可以直接采用空间插值技术对缺失记录进行填补,空间插值方法的具体细节这里不做详细介绍。需要注意的是,在空间聚类分析研究时,处理缺失数据需要特殊的要求,对于大范围的缺失情况,要谨慎使用插补措施;对于较大范围的缺失数据,需要必要的舍弃。

空间数据中的重复与冗余主要指数据集中对同一空间实体的信息记录有重复,或包含不需要的成分,尤其是不同来源的空间数据进行整合后,重复和冗余的记录比较常见。识别和消除冗余、重复的数据不仅可以节省存储空间,有时甚至直接影响分析结果。例如,利用 Delaunay 三角网(TIN)进行数字地形分析或邻近分析时,若存在重复的地理坐标则无法进行 Delaunay 三角网的构建。消除重复与冗余的常用方法是合并和清除,其基本策略一般为由用户实现制定相似的匹配规则,进而自动匹配出可能对应同一实体的记录选择合并或清除操作,常用的算法有排序邻居(sorted-neighborhood)、模糊匹配/合并(fuzzy match/merge)等(李德仁等,2006)。

空间数据的不一致性是数据误差的一种形式,但更加复杂,其主要产生在数据库合并与多源数据的集成过程中。空间数据的不一致性主要包括两种类型:① 上下文相关的冲突,即不同的格式、编码、数据类型的空间数据从不同的数据源进行集成时造成的语义上的冲突。例如,纸质地图矢量化后存储在计算机内的空间数据不能完全满足拓扑上的一致。② 上下文无关的冲突,即由于人为的错误记录、软硬件故障及其他偶然因素所导致的同一系统的空间数据在时态、单位或位置的差别或不一致。空间数据的不一致性检测与消除是数据库研究中的难点问题,但通过在数据库进行操作前的合理检查,理论上可以大大降低空间数据的不一致性。

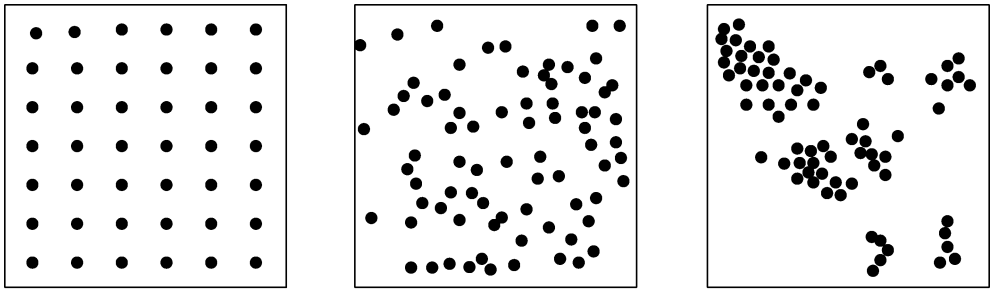
2.4 空间聚类趋势分析

现有的空间聚类分析操作实际上都隐含了一个预定的假设,即空间数据一定是可聚的,因此,针对某个数据集进行空间聚类操作,总可以获得一个空间聚类结果。这种现状可能带来两个直接的后果:① 若数据即是不可聚的,则得到的聚类结果是不可解释的,也是无用的;② 对于数据集缺乏必要的先验知识,给算法与参数的选择带来了巨大的困难。因此,在进行空间聚类分析之前首先对数据集的可聚性(即空间聚类趋势)进行分析是十分必要的。空间聚类趋势分析的目的在于判断空间数据库中是否存在空间簇,即检验数据集是否具有可聚性,但这不是进行聚类分析操作。进而将空间聚类趋势分析问题区分为两种类型,即二维空间点集聚类趋势分析及顾及专题属性的聚类趋势分析,下面将分别进行阐述。

2.4.1 二维空间点集聚类趋势分析

二维空间点集仅考虑空间实体的空间位置属性,其聚类趋势分析旨在分析空间数据在位置上的聚集效应。空间点集的分布虽然复杂,但可以区分为均匀分布、随机分布及聚集分布三种情况,只有表现为聚集分布的数据集才适合进行空间聚类分析操作。图 2.1

分别表示了呈均匀分布、随机分布及聚集分布的三个空间点集数据。



(a) 均匀分布　　(b) 随机分布　　(c) 聚集分布

图 2.1　空间点集分布模式

判断空间点集空间分布的集聚性是空间聚类趋势分析的一个主要任务，其需要从定量的角度对空间数据的聚集或分散程度进行度量，而不需给出聚集的空间分布范围与形态。计量地理学中点模式分析技术为空间聚类趋势分析研究提供了重要的支撑，空间点模式分析方法主要可以分为两类（王远飞等，2007）：① 基于聚集性的方法，如样方法与核密度估计法；② 基于分散性的方法，如最邻近指数法、K-函数法及 G-函数法等。下面首先探讨空间点模式分析技术方法，然后总结近年来发展的基于可视化技术的聚类趋势分析方法。

1. 样方法

样方法是较早提出的一种空间点模式分析方法，其基本思想为：将随机分布模式作为理论上的标准分布模式，进而将通过样方分析得到的空间点分布密度与标准分布模式进行比较，判断空间分布模式。样方分析的基本步骤为：首先使用一个格网结构对研究区域进行分割，统计每个格网中空间实体的频数。格网的形状可以为正方形、正六边形或圆形，一般采用正方形格网，也有使用固定大小的随机格网。进而统计包含不同数量空间实体格网的概率分布。最后，将得到的概率分布与已知的或理论上的概率分布（均匀分布、随机分布）进行比较，一般采用 K-S 检验方法判断空间点实体的分布模式。此外，还可以采用方差-均值比的方法来判断空间点模式的类型。

$$\mathrm{ICS}=\left[\frac{S^2}{x}\right]-1 \tag{2.1}$$

式中，ICS 表示聚集性指数，ICS＝0 表示随机分布，ICS＜0 表示均匀分布，ICS＞0 表示聚集分布。

格网的选择及格网大小的确定是样方法的关键问题，虽然有学者提出了部分经验的最优选择方法，但其对于某些特殊的情况，如空间点分布不均匀的情况仍然可能是不准确的。此外，样方法不考虑空间点实体间的空间关系，因此可能带来一定的误差。

2. 核密度估计法

核密度估计的基本思想在于地理事件在空间点密度大的区域发生的概率大，在空间点密度低的区域发生的概率低。对于每个空间实体 p，其密度在 p 的中心处最大，随着与

p 的距离增大而逐渐减小，达到一定距离阈值后变为 0。某个空间位置 x 的核密度为其窗口范围内所有实体密度之和，表示为

$$f_n(x)=\frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{d(x,x_i)}{h}\right)\right] \tag{2.2}$$

式中，n 表示距离阈值范围内包含的空间实体数量；$K(\cdot)$表示核密度方程；h 表示距离阈值；$d(x,x_i)$表示两点之间的欧氏距离。图 2.2 为核密度估计的示意图。

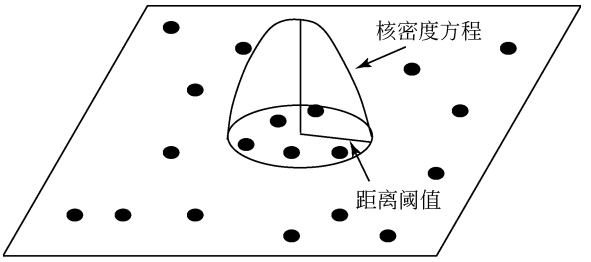


图 2.2　核密度估计

常用的核函数主要有高斯核函数及四次多项式核函数，其中，高斯核函数为

$$K_{\text{Gauss}}(d)=\frac{3}{2\pi h^2}e^{-d(x,x_i)/2h^2} \tag{2.3}$$

而四次多项式核函数为

$$K_{\text{M}}(d)=\frac{3}{\pi h^2}\left[1-\left(\frac{d(x,x_i)}{h}\right)^2\right]^2 \tag{2.4}$$

与样方分析相比，核密度估计是一种可视化的方法，其分析结果与距离阈值密切相关。较大的阈值表示了整体的分布模式，而较小的阈值更强调局部差异。在使用核密度估计法时，距离阈值的选择与边缘效应需要引起特别的重视。

3. 最邻近指数法

最邻近指数法的基本思想为采用空间点实体间的距离来描述点实体的分布模式。最邻近指数计算时，首先计算每个空间实体与其最邻近实体的距离，然后将最邻近距离的平均值作为检验指标与已知的分布模式进行比较。最邻近指数 NNI 为

$$\text{NNI}=\frac{d(NN)}{d(\text{ran})}=\frac{\sum_{i=1}^{N}d_i^{\min}/N}{d(\text{ran})} \tag{2.5}$$

式中，$d_i^{\min}$ 表示每个实体与其最邻近实体的欧氏距离；N 表示数据集中空间实体的数量；$d(NN)$表示最邻近距离均值；$d(\text{ran})$表示随机分布假设下，实体最邻近距离的平均值，其取值一般为 $d(\text{ran})=0.5\sqrt{A/N}$，$A$ 表示研究区域面积。

若 NNI<1 时，空间数据呈聚集分布；若 NNI>1 时，呈均匀分布。为检验计算的可靠性，可以用 z 检验方法，即

$$z=\frac{d(NN)-d(\text{ran})}{\text{SE}} \tag{2.6}$$