中国矿业大学博士学位论文出版基金资助

空间数据挖掘不确定性理论及其应用

KONGJIAN SHUJU WAJUE BUQUEDINGXING LILUN JIQI YINGYONG

何彬彬 著

中国矿业大学出版社

China University of Mining and Technology Press

中国矿业大学博士学位论文出版基金资助
中国博士后科学基金项目(20060390326)
国家自然科学基金项目(60275021)

空间数据挖掘不确定性理论及其应用

何彬彬 著

中国矿业大学出版社

内容简介

本书重点论述了空间数据挖掘的不确定性理论及其应用，主要内容包括：空间数据挖掘不确定性的理论与技术框架；不确定性空间数据挖掘算法模型；多光谱遥感图像分类的不确定性度量和表达；基于客观分析和梯度倒数加权滤波提高时序 MODIS LAI 数据产品质量；顾及不确定性的区域环境质量评价。

本书可供高等院校和科研院所从事遥感、地理信息系统、数据挖掘以及资源环境等领域的科研工作者参考，也可作为高年级本科生和研究生的教学参考书。

图书在版编目(CIP)数据

空间数据挖掘不确定性理论及其应用/何彬彬著. —徐州：中国矿业大学出版社，2007.2

ISBN 978-7-81107-613-4

Ⅰ.空… Ⅱ.何… Ⅲ.数据采集—计算机应用—地理信息系统 Ⅳ.P208-39

中国版本图书馆 CIP 数据核字(2007)第 049847 号

书　　名　空间数据挖掘不确定性理论及其应用
著　　者　何彬彬
责任编辑　潘俊成
出版发行　中国矿业大学出版社
（江苏省徐州市中国矿业大学内　邮编 221008）
网　　址　http://www.cumtp.com　**E-mail**：cumtpvip@cumtp.com
排　　版　中国矿业大学出版社排版中心
印　　刷　徐州中矿大印发科技有限公司
经　　销　新华书店
开　　本　850×1168　1/32　**印张** 5.375　**字数** 140 千字
版次印次　2007 年 2 月第 1 版　2007 年 2 月第 1 次印刷
定　　价　28.00 元

前　言

空间数据挖掘技术是当今空间信息智能处理的重要方法之一。从20世纪90年代开始，各国学者就开始展开对空间数据挖掘的理论和方法的研究，并取得了大量的学术成果。但是，相对于空间数据不确定性方面的研究而言，空间数据挖掘不确定性方面的研究还远远不够。事实上，空间数据自身具有不确定性，空间数据挖掘过程中也会带来一系列的不确定性，这些不确定性在空间数据挖掘过程中会不断传播和累积，从而可能导致挖掘出来的知识有较大的误差甚至毫无意义。而传统的空间数据挖掘并未将这些特性考虑进去，并且一般认为挖掘出来的知识都是有用的和确定的。因此，如何从大量的、含有多种不确定性的空间数据中挖掘出隐含的、有价值的知识，具有重要意义。研究空间数据挖掘不确定性的主要目标和意义有两个：一方面，通过分析空间数据和空间数据挖掘过程中各阶段存在的不确定性及其特性，寻找有效方法来降低其不确定性，以达到提高空间数据挖掘质量的目的；另一方面，我们不可能消除空间数据挖掘过程中所有的不确定性，但通过评价空间数据挖掘质量，研究不确定性在空间数据挖掘过程中的传播规律，建立其质量评价指标和不确定性传播模型，对挖掘出的空间知识有一近似度量，以便人们更好地利用所挖掘的知识。

本书的主要内容是2002年至2005年我攻读博士学位期间，在两位导师——中国矿业大学郭达志教授和上海交通大学方涛教授的指导下，在国家自然科学基金项目“空间数据挖掘的若干关键

理论与技术研究”(60275021)的支持下取得的主要研究成果。主要内容包括:① 空间数据挖掘不确定性的理论与技术框架;② 不确定性空间数据挖掘算法模型,包括不确定性的空间聚类算法和不确定性空间数据关联规则挖掘模型;③ 多光谱遥感图像分类的不确定性度量和表达;④ 基于客观分析和梯度倒数加权滤波提高时序 MODIS LAI 数据产品质量;⑤ 顾及不确定性的区域环境质量评价。

在研究过程中,得到了中国矿业大学杜培军教授和南京师范大学盛业华教授的热情帮助。此外,中国矿业大学邓喀中教授、张书毕教授、汪云甲教授、吴侃教授、张海荣副教授等对论文的撰写也提出了很多有益的建议;上海交通大学唐宏博士和霍宏讲师、陕西师范大学员疆博士、闽江学院的谢储辉教授、中国矿业大学(北京)孙庆先博士等,也给予我大量帮助,与他们的学术讨论,总会给我启发!

本书的出版得到了中国矿业大学研究生院和环境与测绘学院的大力支持和协助。成都理工大学的陈翠华博士提供了部分地球化学数据。在此一并表示诚挚的谢意!

最后,谨向中国矿业大学研究生院、中国博士后科学基金会、国家自然科学基金委和支持、帮助过我的老师、同事与家人致以我崇高的敬意。特别向给予我悉心指导的郭达志教授和方涛教授表示衷心的感谢!

由于作者水平有限,对空间数据挖掘的不确定性研究还有待深入,有些内容正在继续进行深入研究,书中存在不足和谬误之处,敬请读者批评指正!

何彬彬

2006 年国庆于电子科技大学

目　录

1 绪 论

1.1 研究背景、目的及意义

遥感(RS)、地理信息系统(GIS)和全球定位系统(GPS)已成为现代对地观测技术中信息获取、存储管理、更新、分析和应用的三大支撑技术。随着空间数据获取手段的自动化程度的不断提高,空间数据呈指数级地增长,传统的专题制图和空间分析已经远远不能满足社会对其需求。专职处理空间数据的 RS 和 GIS 软件在分析功能上的不足,使得海量空间数据与有用知识获取之间存在尖锐的矛盾,致使"空间数据爆炸但知识贫乏"[1]。前美国副总统戈尔于 1998 年发表的题为"数字地球:21 世纪认识我们这颗星球的方式(The Digital Earth: Understanding our planet in the 21st Century)"的演讲中清楚地描述了这种挑战,"充分利用这些海量数据的困难在于把这些数据变得有意义——即把原始数据变成可理解的信息。今天,我们经常发现我们拥有很多数据,却不知如何处理(The hard part of taking advantage of this flood of geospatial information will be making sense of it-turning raw data into understandable information. Today, we often find that we have more information than we know what to do with)"[2]。

目前,绝大多数遥感和地理信息系统对空间数据的利用主要是查询、专题制图、空间分析和简单的统计,这仅可以满足某些低

层次的需求，人们迫切需要的是从大量数据中挖掘出对决策具有指导意义的知识。这些知识比简单的查询和统计获取的信息更加概括、更加浓缩和精练，是对数据的更深刻的认识。从空间数据中挖掘出的知识一方面可用于决策支持，提高 GIS 数据分析和利用的智能化水平，另一方面用以支持遥感图像处理的自动解译，提高解译的自动化水平，从而促进 GIS 与 RS 的智能化集成[3]。因此，如何从大量的、含有多种不确定性的空间数据中挖掘出隐含的、有价值的知识，是一个重要的前沿课题，具有重要意义。

20 世纪 80 年代末，计算机信息处理领域开始兴起数据挖掘技术。从数据库中发现知识(KDD)首次出现在 1989 年 8 月召开的第十一届国际联合人工智能会议上。KDD 被明确地定义为“从数据中发现隐含的、先前不知道的、潜在有用的信息的非平凡过程”[3,4]，随后国际上兴起了研究数据挖掘的热潮，美国和欧洲相继举办了一年一次的数据挖掘国际会议，包括 IEEE 的 ICDM、ACM 的 SIGKDD 等。1997 年国际上第一本数据挖掘杂志——*Data Mining and Knowledge Discovery* 创刊，其影响因子在 2003 年已经上升到 2.5。与此同时，国际上很多软件开发商也相继推出了数据挖掘软件，目前有影响力的有 IBM 的 Intelligent Miner、SAS 公司的 Enterprise Miner、SGI 的 MineSet、Simon Fraser 的 DBMiner、Megaputer Intelligence 的 PolyAnalyst 等。

随着数据挖掘技术的兴起及其在商业应用中的良好前景，引起了空间信息领域学者的关注和兴趣。20 世纪 90 年代中期开始，加拿大 Simon Fraser 大学、德国慕尼黑大学、芬兰赫尔辛基大学以及美国、澳大利亚等国家的许多大学和研究所，都有空间数据挖掘研究成果的报道[5~10]。这些研究者大多具有计算机科学背景，他们把空间数据挖掘作为数据挖掘的一个应用领域，研究的重点是提高原有数据挖掘算法在空间数据库中的执行效率。在我国，测绘遥感、地理学的学者们在特征提取、模式识别等研究中已

经做了许多空间数据挖掘的工作。武汉大学的邸凯昌博士(1999)在李德仁院士和李德毅院士的指导下,率先在国内系统展开空间数据挖掘的研究[11],并出版专著《空间数据发掘与知识发现》;中国科学院资源与环境信息系统国家重点实验室的杨存建博士(1999)和张健挺博士(1999)在陈述彭院士和周成虎研究员的指导下,进行地学空间数据挖掘方面的研究[12,13]。随后,中国科学院地理所[14,15]、中国科学院遥感所[16]、解放军信息工程大学[17~20]、武汉大学[21~25]、中国地质大学[26]、中国科技大学[27]、中国矿业大学[28~33]等都有空间数据挖掘方面的成果报道。在空间数据挖掘系统的开发方面,国际上有代表性的空间数据挖掘系统有:加拿大Simon Fraser大学计算机科学系数据挖掘小组在MapInfo平台上建立了空间数据挖掘的原型系统——GeoMiner[34];德国Fraunhofer大学自动智能系统研究所主持的空间数据挖掘与知识发现原型系统——SPIN[35],该系统用Java语言开发,集空间数据分析和数据挖掘于一体,很多功能还处于研发过程中。

目前,空间数据挖掘的研究主要集中在:传统的数据挖掘算法(如ID3、C5、Apriori、粗集等)在空间数据挖掘中的应用;面向海量空间数据的挖掘算法研究,以韩家伟教授为首的数据挖掘研究组在这方面已做了大量工作,其主要成果在K. Koperski博士的博士学位论文中得到详尽体现[36];遥感数据挖掘(信息提取)方面,美国NASA和德国航空中心(DLR)的学者们近几年进行了较深入的研究,并已取得一些重要成果[37~40]。另外,J. Dong、W. Perrizo和Q. Ding等结合TM图像数据和地面农作物产量数据进行了遥感数据关联规则挖掘方面的研究[41~43];A. Kitamoto采用数据挖掘技术对卫星遥感数据中台风图像数据进行分析,并提取各种模式和特征[44~45]。总之,空间数据挖掘已经不是一个崭新的方向,但这方面的研究远未达到成熟和实用的阶段,还有很多理论难题。因为目前的大量空间数据挖掘工作还是停留在应用传统

的数据挖掘算法做一些简单的示例，而对空间数据自身的特点和数据挖掘算法局限性的考虑还远远不够，例如空间数据固有的不确定性、多尺度、空间自相关性以及空间数据挖掘的模糊性等问题。鉴于此，我们将研究重点放在空间数据挖掘不确定性问题的探讨方面。因为空间数据自身具有不确定性，空间数据挖掘过程中也会带来一系列的不确定性，这些不确定性在空间数据挖掘过程中会不断传播和累积，从而可能导致挖掘出来的知识有较大的误差甚至毫无意义。而传统的空间数据挖掘并未将这些特性或问题考虑进去，并且一般认为挖掘出来的知识都是有用的和确定的，这显然是不科学的和不妥当的。

研究空间数据挖掘不确定性的主要目标和意义有两个：一方面，通过分析空间数据和空间数据挖掘过程中各阶段存在的不确定性及其特性，寻找有效方法来降低其不确定性，以达到提高空间数据挖掘质量的目的；另一方面，我们不可能消除空间数据挖掘过程中所有的不确定性，但通过评价空间数据挖掘质量，研究不确定性在空间数据挖掘过程中的传播规律，建立其质量评价指标和不确定性传播模型，对挖掘出的空间知识有一近似度量，以便人们更好地利用所发掘的知识。

1.2 国内外相关领域的研究综述

1.2.1 空间数据不确定性研究概述

空间数据质量与不确定性是目前 GIS 研究的重要基础理论之一。关于空间数据不确定性的论述始于 1968 年 Blakney 发表的《地形图的精度标准》和 Webster 和 Beckett 发表的《土壤图的质量和用途》两篇文献，并认为是 GIS 空间数据不确定性研究方面最早的文献资料[46,47]。在此之后，空间数据不确定性方面的研

究成果大量涌现，尤其是海内外的中国学者在空间数据不确定性方面进行了大量研究[48~62]。美国国家地理信息和分析中心(NCGIA)在制定20世纪90年代研究计划时，提出十二个研究课题，其中第一个就是“空间数据库精度”。NCGIA于1988年12月主持召开的空间数据库精度科学大会是国际GIS空间数据不确定性理论研究史上的一个里程碑。近年来，空间数据不确定性的研究成果越来越多地出现在一些大型国际会议上，影响较大的有Auto-Carto会议、国际空间数据处理会议(SDH)和国际空间数据质量会议(ISSDQ)。其中ISSDQ到目前已召开了三届，报道了几年来空间数据质量与不确定性方面取得的众多最新成果。第三届国际空间数据质量专题研讨会于2004年4月15～17日在奥地利维也纳科技大学召开，我们发表了题为*Uncertainty in Spatial Data Mining and Its Propagation*的学术论文[29]。总结第二届和第三届空间数据质量专题研讨会的主要研究内容有[63~64]：① 位置误差模型；② 空间对象和拓扑关系的不确定性模型；③ 空间认知的不确定性；④ 语义不确定性；⑤ DEM的精度分析；⑥ 空间分析中的不确定性；⑦ 土地资源监测的不确定性；⑧ 遥感图像分类的精度估计；⑨ 高分辨率卫星遥感数据的测量精度；⑩ 多传感器和多角度遥感数据误差分析；⑪ 元数据管理及其标准化；⑫ 空间数据基础设施的质量管理等。

目前，对空间数据的位置不确定性及其传播已进行了大量研究，主要是研究点、线、多边形和面状物体的位置不确定性，而点的不确定性是研究线和多边形不确定性的基础。主要不确定性度量模型有：点位置不确定性的标准椭圆模型[65]和圆形正态模型[46]、线位置不确定性的Epsilon带模型[66]和误差带模型[67]。多边形位置不确定性的处理方法主要有基于误差传播定律的方法[65]和Monte Carlo模拟方法[68]。多年来，空间数据质量控制问题主要集中在位置不确定性上，而对属性数据不确定性的研究较少。属

性数据是度量被测量对象知识缺乏的程度，可表现为空间属性数据的误差、不精确性、随机性和模糊性，且受尺度、分辨率、抽样等因素影响。近几年，国内外学者对 GIS 属性数据的不确定性进行了一些研究[69~71]。史文中[72]和张景雄[51]分别建立了结合位置与属性统一的 S—带模型和场模型。葛咏等[47]从遥感成像机理的角度分析了 SAR 图像数据的不确定性。承继成等[73]系统论述了遥感数据不确定性的有关问题。

1.2.2 数据挖掘概述

数据挖掘(DM)就是对观测到的大量数据集进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据[74]。通过数据挖掘过程所推导出的关系和结构经常被称为模型或模式。例如线性方程、规则、聚类图、树结构以及时间表示的循环模式。

DM 经常被视为 KDD 的同义词。KDD 这个术语来源于人工智能(AI)领域。KDD 由以下几个主要步骤组成：① 数据清理，消除噪声或不一致数据；② 数据集成，多种数据源可以组合在一起；③ 数据选择，从数据库中检索与分析任务相关的数据；④ 数据变换，数据变换或统一成适合挖掘的形式，如通过汇总或聚积操作；⑤ 数据挖掘，使用智能方法提取数据模式；⑥ 模式评价，根据某种兴趣度量，识别表示知识的真正有趣的模式；⑦ 知识表示，使用可视化和知识表示技术，向用户提供挖掘的知识[75]。

典型的数据挖掘系统结构如图 1-1 所示，主要包括：数据库、数据仓库或其他信息库，数据库或数据仓库服务器，知识库，数据挖掘引擎，模式评估模块，图形用户界面等。

DM 是一门跨学科的技术，统计学、数据库技术、机器学习、模式识别、人工智能、可视化技术都在其中起着重要作用(图 1-2)，而且就像难以定义这些学科间的严格界限一样，也很难定义这些

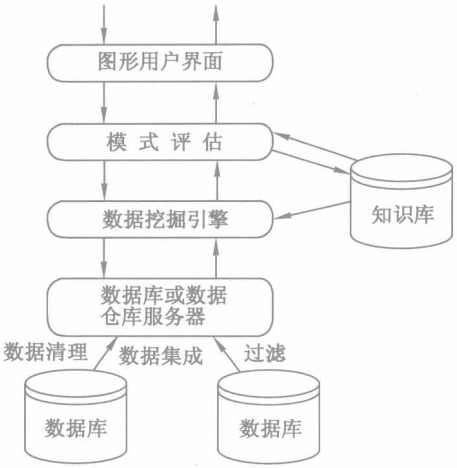

图 1-1 典型的数据挖掘系统结构图

学科和数据挖掘间的界限。在边缘上,一个人的数据挖掘问题可能是其他人的统计、数据库或机器学习问题。

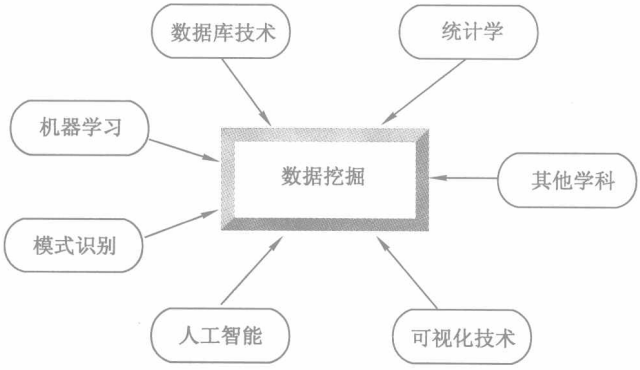

图 1-2 数据挖掘的多学科交叉

不同数据分析目标有不同数据挖掘任务的类型划分。数据挖掘任务可以总结为如下几个方面：探索性数据分析(Exploratory Data Analysis，EDA)、描述建模、预测建模、寻找模式和规则、根据内容检索。针对这些任务的数据挖掘算法具有以下四个基本组件：① 模型或模式结构，决定要从数据中寻找的潜在结构或函数形式；② 评分函数，鉴定一个已拟合模型的质量；③ 优化和搜索方法，优化评分函数并对不同的模型和模式结构进行搜索；④ 数据管理策略，在搜索和优化期间高效地处理数据访问问题[74]。

1.2.3 空间数据挖掘研究概述

空间数据挖掘(Spatial Data Mining，SDM)，或称从空间数据库中发现知识(Knowledge Discovery from Spatial Databases，KDSD)，是指从空间数据库中提取用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其他一些隐含在数据库中的普遍的数据特征[76,3]。本书采用空间数据挖掘的广义观点，空间数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的实际应用空间数据中提取隐含的、未知的、潜在的、有用的知识的过程[11]。

1.2.3.1 空间数据挖掘的特点

由于空间属性的存在，空间的个体才具有了空间位置和距离的概念，并且距离邻近的个体之间存在一定的相互作用，空间数据之间的关系类型因此也就更为复杂。空间数据的复杂性特征主要表现在以下几个方面[14]：① 空间属性之间的非线性关系；② 空间数据的多尺度特征；③ 空间数据的模糊性；④ 空间维数的增高；⑤ 空间数据的缺值。此外，几乎所有的空间数据都具有空间自相关性。

空间数据挖掘是数据挖掘的一个分支，由于空间数据的复杂性，空间数据挖掘又不同于一般的事务数据挖掘。与关系数据库

和事务数据库中的数据挖掘技术相比，空间数据挖掘具有如下几个特点：

(1) 处理的对象更加复杂。关系数据库主要是各种结构化的关系数据表，由元组、属性构成，便于数据的处理和知识的发现；但空间数据挖掘的处理对象，不仅仅包含复杂的空间数据，同时有非空间数据参与。此外，空间数据本身也存在比例尺的多样性、表达方式的不一致性、图形数据的非结构化等特点。

(2) 空间数据挖掘存在粒度问题——数据处理的元组。根据不同的挖掘目的和数据组织方式，用户处理数据的分辨率不一致，要根据需要进行选择。在以矢量方式存储的数据中，对象实体是以图元的方式存储的，在处理时可以图元作为最小对象，即处理粒度；以栅格方式存储的数据，其处理的最小粒度是像元，如遥感影像的信息处理，可以充分利用像元的特征信息(如像元位置、高程、坡度等)进行运算。

(3) 空间数据挖掘具有尺度维的变化。所谓尺度是指在研究某一现象或事件时采用的空间或时间单位，同时又可指某一现象或过程在空间和时间上所涉及到的范围和发生的频率，有时是指人们观察世界的窗口[77,78]。在不同的学科领域中，对尺度有不同的表述或含义[76]：在测绘学、地图制图学和地理学中通常把尺度表述为比例尺；在遥感科学技术中，尺度一般相应于分辨率；在生态环境中，大尺度或粗尺度(Coarse Scale) 是指大空间范围或长时间幅度，它往往对应于小比例尺、低分辨率，而小尺度或细尺度(Fine Scale) 一般指小空间范围或短时间幅度，往往对应于大比例尺、高分辨率。空间图形数据的表达在不同的比例尺上表现出不同的空间尺度。尺度的变化影响着参与运算数据的质量和数量以及数据提供的信息量。

(4) 时态数据的处理和分析困难。在地理空间现象中，许多现象是随时间而发展变化的，这些与时间相关的时态数据在空间