文章编号: 1671-8860(2006)01-0001-04 文献标志码: A

论空间数据处理与空间数据挖掘

王新洲[1,2]

(1 武汉大学灾害监测与防治研究中心, 武汉市珞喻路129号, 430079)
(2 山东科技大学地球科学与工程学院, 青岛市经济技术开发区前湾港579号, 266510)

摘 要: 根据现行文献中反复提到的空间数据处理内涵的理解, 将空间数据处理分为空间数据处理技术和空间数据处理理论, 简要论述了各自的主要内容, 讨论了空间数据挖掘的现状和今后研究的重点, 比较了空间数据处理与空间数据挖掘的异同。

关键词: 空间数据处理; 空间数据挖掘; 空间数据不确定性
中图法分类号: P208

空间信息科学由空间数据采集、空间数据处理、空间数据表达和空间信息服务等几个方面组成。其中, 空间数据采集是基础, 空间数据处理是关键, 空间数据表达是形式, 空间信息服务是目的。作为空间信息科学之关键的空间数据处理, 其内涵与外延非常广泛, 人们难以准确理解其实质。本文针对空间数据处理和近年出现的空间数据挖掘进行了讨论。

1 空间数据处理

1.1 空间数据处理技术

仅对空间数据进行一般的操作, 以解决数据操作层面上的一系列问题的空间数据处理, 称为空间数据处理技术。如进行多种方式的栅格数据和矢量数据的裁剪; 进行数据类型的转换; 对原始数据进行拓扑处理; 对遥感图像进行彩色合成增强处理; 对海量影像数据建立影像金字塔, 进行快速显示等, 都属于空间数据处理技术的范畴。此外, 如图库操作(建立图库、修改、删除及图库漫游等)、图幅操作(图幅输入、显示、修改、删除以及调用、存取、显示、查询任一图幅等)、图幅配准(平移变换、比例变换、旋转变换和控制点变换等)、图幅接边(对图幅帧进行分幅、合幅, 并进行图幅的自动、半自动及手动接边)、图幅提取(对分层、分类存放的图形数据按照不同的层号或类别合并生成新的图件)等操作, 也都属于空间数据处理技术的范畴。

1.2 空间数据处理理论

空间数据处理技术主要是解决空间数据操作层面上的问题, 没有顾及空间数据的不确定性。由于空间数据的不确定性是客观存在的, 是任何空间数据采集方法中都不可避免的, 所以必须在考虑空间数据不确定性的基础上对空间数据进行更高层次的处理, 这种处理称为空间信息处理理论。空间信息处理理论的理论基础主要有概率与数理统计、模糊数学、仿生数学(如神经网络、遗传算法、免疫算法)等。在这些基础理论中, 占主导地位的是概率与数理统计。

1.2.1 空间数据的不确定性理论

空间数据的不确定性可以认为是空间数据的"真实值"不能被肯定的程度。它可以看作是一种更广义的误差, 既包含随机误差, 也包含系统误差和粗差, 还包含可度量和不可度量的误差, 以及数值上和概念上的误差。在形式上, 它一般包含着真实值的一个范围, 这个范围越大, 数据的不确定性就越大[1]。研究空间数据不确定性的目的之一就是要消除数据之间的矛盾。

研究空间数据不确定性的另一个目的就是要评价空间数据的质量, 包括评定原始数据和处理结果的质量。关于这方面的研究, 一般从位置不确定性和属性不确定性两者入手。对于位置不确

收稿日期: 2005-10-13。
项目来源: 国家自然科学基金资助项目(40474003); 山东省泰山学者建设工程专项经费资助项目。

定性的研究,其理论基础主要是协方差传播律,其研究成果主要集中在线元的各种误差带,如 ε 带、*E*-带、*G*-带、ε*r* 带、ε*m*-带、*H*-带、*S*-带、*C*-带、*R*-带以及等概率密度误差带[2]等。这些误差带都是根据线元上任意点 $Z_t(x(t), y(t))$ 的坐标计算式:

$$\begin{cases} x(t) = (1-t)x_0 + tx_1 \\ y(t) = (1-t)y_0 + ty_1 \end{cases} \quad (1)$$

按方差-协方差传播律求出 $Z_t(x(t), y(t))$ 的方差-协方差矩阵,然后以不同方法得到的。式中,$t = |Z_0Z_t| / |Z_0Z_1| = L_t/L$, $t \in [0, 1]$。

按以上方法得到的各种误差带存在一个共同的问题,即按式(1)求出的线元上任意点 $Z_t(x(t), y(t))$ 的方差-协方差矩阵并不是图上线元任意点的方差-协方差矩阵,这是因为图上线元任意点的坐标并不是通过式(1)计算得到的。因此,目前以式(1)为基础讨论的上述各种误差带的形状和性质仍值得商榷,应该以新的思维来研究 GIS 中的线元误差。笔者认为,首先由两个端点所确定的一条直线是一个随机向量,它的误差带就应该根据随机向量从整体上加以研究。根据这个思想,游扬声[3]提出了线元整体误差带的理论,得出了线元整体等密度误差带的形状由 Z_0 和 Z_1 的两个最大的等密度椭圆 S_0、S_1 以及 S_0 和 S_1 的两条公切线组成,并经过严密的推导,计算出了线元的真实位置落在整体等密度误差带内的概率。

1.2.2 参数估计理论

长期以来,参数估计理论领域的研究成果很多,比较有影响的参数估计有极大似然估计、最小二乘估计、极大验后估计、最优无偏估计、贝叶斯估计、稳健估计、最小二乘配置、最小二乘滤波、非线性最小二乘估计、偏参数估计、半参数估计,以及信息扩散估计[4]和基于信息扩散的极大似然估计[5]等。这些方法各有特点,但它们都具有一个共同之处,即其理论基础都是概率与数理统计。

在空间数据处理中,为了能应用现有的参数估计理论,总是将空间数据的不确定性理想化、简单化,即假定空间数据的不确定性是随机变量。事实上,由于种种原因,空间数据的不确定性并非仅仅由随机误差组成,而是多种不确定因素的综合。空间数据的不确定性在任何一次测量中是一定会出现的,它出现与否在测量前是确知的。另外,空间数据的不确定性还包括概念上的不确定性,因此,空间数据的不确定性不完全满足随机变量的定义。因为空间数据的不确定性不完全满足概率论中的随机变量这一基本假设,所以使用目前的任何一种参数估计方法来处理空间数据都是不严密的。在实际的参数估计中,人们总是根据实际情况对现有的参数估计模型进行修正。由于空间数据的不确定性既包含随机性,也包含模糊性,故现有的理论和方法根本不能全面地处理空间数据的不确定性[6]。为此,笔者将观测值看作模糊数,以模糊数学为理论基础,提出了与现有的参数估计理论和方法相并列的一类参数估计理论和方法——极大可能性估计[7]。极大可能性估计是以可能性理论为基础,对模糊数进行处理的新的估计类。它虽然是一类与极大似然估计相平行的估计,但其理论上还不够成熟,为此,文献[8]对它作了补充和完善。

1.2.3 空间数据分析与建模理论

近 10 年来,国际学术界开始研究和发展能够用于描述和表达时空变化的地理信息系统,即时态地理信息系统[9]。时态地理信息系统可以同时提供空间、时间和属性数据的建模和分析手段,能够完整地保存某一地理事件的发展历史,从而对历史状态的重建、时空变化的跟踪及未来发展态势的预测提供了可能[9]。

预测首先需要建立自变量与因变量之间的相互依赖的预测模型,传统的统计预测模型的建模方法有 p 阶自回归模型 AR(p)、q 阶滑动平均模型 MA(q)以及 p 阶自回归 q 阶滑动平均模型 ARMA(p, q)。这三种模型具有等价性,可以相互转换。

由于自变量与因变量之间的关系并非确定的,更不是线性的,而是一个复杂的非线性系统,各影响因素之间的关系非常复杂,用上述线性的确定性回归模型很难准确模拟这一复杂的非线性系统[10]。为此,神经网络被用来建立预测模型[11]。神经网络预测模型相对统计预测模型而言,具有更大的灵活性,但仍存在一些问题,如神经网络拓扑结构的确定,即隐含层的层数及节点数的确定,激活函数如何选取,神经网络结构的物理意义无法解释等。虽然国内的一些学者对这些问题进行了研究,并取得了一定的成果,如直接估计法、裁减法、生长法、双向确定法、相似相关系数法、主成分法、大样本法等[10],但这些方法都存在一定的缺陷,难以达到理想的效果[12]。为此,可将模糊逻辑与神经网络结合起来建立预测模型,这种预测模型称为模糊神经网络预测模型。应用实例表明,模糊神经网络预测模型具有训练时间短、预测精度高的特点,相对于其他变形预报模型,具有一定的优势[13]。

1.2.4 遥感影像分类理论

遥感影像分类是一种典型的模式识别问

题[14], 一般分为监督分类和非监督分类两种。根据对 $f(x|T)$ 估计方法的不同, 监督分类有若干种。如在光谱遥感影像的分类中, 主要的监督分类方法大致可分为贝叶斯判别法、原型法和函数估计法三大类[14]。近年来, 支持向量机被用来对遥感影像进行监督分类, 取得了一些很好的结果。

聚类是按照事物间的相似性进行区分和分类的过程, 是一种无监督的分类。聚类分析则是用数学方法研究和处理所给定对象的分类。由于模糊聚类得到了样本属于各个类别的不确定性程度, 表达了样本类别的中介性, 即建立起了样本对于类别的不确定性的描述, 能更客观地反映现实世界, 从而成为聚类分析研究的主流。以此为理论基础所进行的非监督分类可能更符合实际。

2　空间数据挖掘

2.1　空间数据挖掘的诞生[15]

1989 年 8 月, 在美国底特律市召开的第一届国际联合人工智能学术会议上, 首次出现了从数据库中发现知识(knowledge discovery in database, KDD)的概念。它针对的一般是非空间数据, 其研究和应用的成果势必对空间数据的利用造成影响, 引导地球空间信息学向更深的层次发展。1994 年, 在加拿大渥太华举行的 GIS 国际学术会议上, 李德仁院士首次提出了从 GIS 数据库中发现知识(knowledge discovery from GIS, KDG)的概念, 并系统分析了空间知识发现的特点和方法, 认为它能够把 GIS 有限的数据变成无限的知识, 使 GIS 成为智能化的信息系统。1995 年, 在加拿大召开的第一届知识发现和数据挖掘国际学术会议上, 又出现了数据挖掘(data mining, DM), 后又相继出现了数据发掘、数据开采、数据采掘、知识提取、信息发现、信息收获、数据考古等。由于 DM 和 KDD 较为常用且难以分离, 而且 DM 通常被认为是 KDD 中通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤, 即数据挖掘和知识发现(data mining and knowledge discovery, DMKD)。同时, 李德仁院士也把 KDD 进一步发展为空间数据挖掘和知识发现, 系统地研究或提出了可用的理论、技术和方法, 并取得了很多创新性成果[15-20], 奠定了空间数据挖掘和知识发现在地球空间信息学中的学科地位和基础。

2.2　空间数据挖掘和知识发现的进展

我国许多科研院所和高校等先后展开了对空间数据挖掘和知识发现的理论和应用研究, 国家对空间数据挖掘和知识发现也给予了极大的重视。其中, 李德仁院士的创新性研究得到了国际同行的首肯。而且, 李德仁院士和李德毅院士合作, 早在 1999 年就率先培养出了我国在空间数据挖掘和知识发现方向上的第一个博士——邸凯昌博士。之后, 王树良又进行了大量深入细致的工作, 在李德毅院士的云理论的基础上, 完善了数据场的概念, 提出了空间数据挖掘视觉的概念及实现方法, 并成功地应用于滑坡监测数据挖掘, 取得了较好的成果[21]。秦昆博士将空间数据挖掘如何应用于实践做了大量有意义的工作。他在对图像数据挖掘的理论与方法进行系统研究的基础上, 针对图像(遥感图像)数据中蕴涵的内容, 如光谱特征、纹理特征、形状特征、空间分布特征等来进行挖掘, 挖掘出抽象层次更高的知识。秦昆博士研究出了图像(遥感图像)数据挖掘软件原型系统的框架, 设计和开发了图像(遥感图像)数据挖掘软件原型系统 RSImageMiner[22]。RSImageMiner 的开发成功, 标志着空间数据挖掘已从理论研究逐步向实际应用迈进。

2.3　空间数据挖掘和知识发现的基本理论方法

文献[19]中详细讨论了空间数据挖掘的理论和方法, 概括起来有概率论、证据理论、空间统计学、规则归纳、聚类分析、空间分析、模糊集、云理论、粗集、神经网络、遗传算法、决策树、分类分析、预测、关联规则分析、时间序列分析、熵空间理论、形式概念分析理论(概念格理论)等[19,22]。这些理论和方法都是自成体系的, 根本不是空间数据挖掘自身的理论体系。因此, 关于空间数据挖掘理论的研究应重点放在构建空间数据挖掘系统的理论框架上, 不能简单地将各种现成理论统归于空间数据挖掘理论。笔者认为, 空间数据挖掘的系统理论框架应由下列三大部分构成: ①空间数据挖掘的基础理论(一定要寻求自成体系的、相对独立的一套理论); ②空间数据挖掘的技术方法(在基础理论的指导下, 形成具有特色的技术方法, 当然也可以借助现成方法); ③空间数据挖掘结果的质量评价体系。空间数据挖掘的方法应在这个理论框架下, 研制实用软件。

3　空间数据处理与空间数据挖掘的联系与区别

空间数据处理和空间数据挖掘是目前出现较多的两个名词。从内涵看, 空间数据处理是更大的概念, 广义的空间数据处理包含了空间数据挖

掘;从外延看,两者又有区别。笔者认为,其区别主要体现在三个方面。

1) 层次不同。尽管广义的空间数据处理包含了空间数据挖掘,但它们对空间数据处理的层次是不同的。一般的空间数据处理仅解决表层的、显而易见的问题,而空间数据挖掘则是解决深层次的问题,解决人们从表面无法发现的问题。

2) 数据量不同。尽管空间数据处理和空间数据挖掘所处理的都是空间数据,但其处理的数据量是不同的。空间数据处理是对一组空间数据进行处理,一般不涉及其他空间数据,即被处理的数据是极有限的。而空间数据挖掘则是对整个空间数据库进行操作,必要时甚至对多个空间数据库进行操作,所涉及的数据是大量的,而且数据量越大,挖掘效果越好。

3) 目的不同。空间数据处理技术的目的是对空间数据进行一些简单的操作,为一般用户的常规应用服务;空间数据处理理论的目的是解决由于不可避免的空间数据不确定性所引起的一系列问题;而空间数据挖掘的目的则是要寻找隐含在空间数据中的规则,并发现知识。

4 结 语

从以上讨论可知,空间数据处理可分为空间数据处理技术和空间数据处理理论。空间数据处理技术是对空间数据所作的一般性处理,属于空间数据操作层面上的问题。空间数据处理理论不是解决空间数据操作层面上的一般性问题,而是从理论上解决由于空间数据不确定性所引起的一系列问题。笔者认为,空间数据处理理论的发展趋势可能会从单一向综合发展,从简单向复杂发展,从硬计算向软计算发展。至于空间数据挖掘,目前还停留在概念阶段,对它的研究重点应放在构建空间数据挖掘的系统的理论框架上。

参 考 文 献

[1] 刘大杰,史文中,童小华,等. GIS 空间数据的精度分析与质量控制[M]. 上海:海科学技术文献出版社,1999

[2] 汤仲安. 空间线状实体等概率密度误差模型[D]. 武汉:武汉大学,2004

[3] 游扬声. 一般分布模式下 GIS 位置数据的不确定性研究[D]. 武汉:武汉大学,2005

[4] 王新洲. 基于信息扩散原理的估计理论、方法及其抗差性[J]. 武汉测绘科技大学学报,1999,24(4):240-244

[5] 游扬声,王新洲. 基于信息扩散的极大似然估计[J]. 武汉大学学报·信息科学版,2003,28(5):562-565

[6] 王新洲,史文中,王树良. 模糊空间信息处理[M]. 武汉:武汉大学出版社,2003

[7] 王新洲,史文中. 极大可能性估计[J]. 测绘学报,2003,29(3):193-197

[8] 王新洲. 最小不确定度约束下的极大可能性估计[J]. 测绘工程,2003,12(1):5-8

[9] 罗年学. 时空对象模型及其在地籍信息系统中的应用研究[D]. 武汉:武汉大学,2002

[10] 黄全义. 大坝变形预报神经网络专家系统方法研究[D]. 武汉:武汉大学,2001

[11] 胡铁松. 神经网络预测与优化[M]. 大连:大连海事大学出版社,1997

[12] 邓兴升. 大坝变形分析及预报的模糊神经网络方法研究[D]. 武汉:武汉大学,2004

[13] 王新洲,邓兴升. 大坝变形预报的模糊神经网络模型[J]. 武汉大学学报·信息科学版,2005,30(7):588-591

[14] 刘志刚. 支持向量机在光谱遥感影像分类中的若干问题研究[D]. 武汉:武汉大学,2004

[15] 王树良. 基于数据场与云模型的空间数据挖掘和知识发现[D]. 武汉:武汉大学,2002

[16] Wang Shuliang, Li Deren, Li Deyi, et al. Cloud Models-Based SDMKD, Geoinformatics'2002: GIS and Remote Sensing for Global Change Studies and Sustainable Development [J]. International Association of Chinese Professionals in Geographic Information Science (CPGIS), 2002, C54: 1-11

[17] 王树良,李德仁,史文中,等. 地学粗空间的理论与应用[J]. 武汉大学学报·信息科学版,2002,27(3):274-282

[18] 李德仁,王树良,史文中,等. 论空间数据挖掘和知识发现[J]. 武汉大学学报·信息科学版,2001,26(6):491-499

[19] 李德仁,王树良,李德毅,等. 论空间数据挖掘和知识发现的理论和方法[J]. 武汉大学学报·信息科学版,2002,27(3):221-233

[20] 王树良,李德仁,史文中,等. 地学粗空间的理论与应用[J]. 武汉大学学报·信息科学版,2002,27(3):274-282

[21] 王树良,王新洲,曾旭平,等. 滑坡监测数据挖掘视角[J]. 武汉大学学报·信息科学版,2004,29(7):608-610

[22] 秦昆. 基于形式概念分析的图像数据挖掘研究[D]. 武汉:武汉大学,2004

作者简介:王新洲,教授,博士,博士生导师。主要从事空间信息处理理论与应用方面的研究。

E-mail: whwxz@163.com

(下转第 8 页)

差的研究[J]. 测绘通报, 2004(7): 12-13
[2] 许国辉. 高精度 EDM 三角高程测量的研究[J]. 测绘通报, 2002(10): 22-24
[3] 周水渠. 精密三角高程测量代替二等水准测量的尝试[J]. 测绘信息与工程, 1999(3): 25-29
[4] 潘松庆. 根据气温变化率进行三角高程测量的折光改正[J]. 河海大学学报, 1999, 27(5): 12
[5] 姜晨光. 精密三角高程测量严密计算的理论研究与初步试验[J]. 四川测绘, 1996(3): 125-128

第一作者简介: 张正禄, 教授, 博士, 博士生导师。主要从事精密工程测量、变形监测分析与预报、测量数据处理和工程信息系统方面的科研和教学工作。
E-mail: zzl623@whu.edu.cn

Research on Precise Triangulated Height Surveying in Place of First Order Leveling

ZHANG Zhenglu[1] *DENG Yong*[1] *LUO Changlin*[1] *HU Xuqing*[2]
(1 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)
(2 Station Quality-inspected Branch of Surveying and Mapping Bureau, Hunan Province, 693 Middle Shaoshan Road, Changsha 410007, China)

Abstract: This paper analyzes the principle, error sources and precision of triangulated height surveying, points out the key problems about first order leveling replaced by triangulated height surveying; and for the first time puts forward that in some given conditions, it is not only feasible but also valuable to replace first order leveling by precise triangulated height surveying, and proves it by experimentation as well.

Key words: electronic distance measurements; triangulated height surveying; first order leveling; deformation monitoring surveying

About the first author: ZHANG Zhenglu, professor, Ph.D, Ph.D supervisor. He is concentrated on the research and education in precise engineering geodesy, deformation monitoring analysis and forecast, measurement data processing and engineering geo-information system.
E-mail: zzl623@whu.edu.cn

(上接第 4 页)

Spatial Data Processing and Spatial Data Mining

WANG Xinzhou[1,2]
(1 Research Center for Hazard Monitoring and Prevention, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)
(2 School of Geo-Information Science and Engineering, Shandong University of Science and Technology, 579 Qianwangang, Qingdao 266510, China)

Abstract: This paper summarizes the main content of technique and theory of spatial data processing respectively based on the authors own understanding on spatial data processing, discusses the current research status of spatial data mining and its emphasis in the future research, and compares the similarities and differences between spatial data processing and spatial data mining. The development direction of each one is forecasted.

Key words: spatial data processing; spatial data mining; spatial data uncertainty

About the author: WANG Xinzhou, professor, Ph.D, Ph.D supervisor. He is concentrated on the research and education in the theory and application of spatial data processing.
E-mail: whwxz@163.com