# 空间数据知识发现研究进展评述

裴 韬 周成虎 骆剑承 韩志军 汪 闽 秦承志 蔡 强
(中国科学院资源与环境信息系统国家重点实验室, 北京 100101)

**摘 要** 首先对当前空间数据的复杂性特征进行了分析,提出海量的数据、空间属性之间的非线性关系、空间数据的尺度特征、空间信息的模糊性、空间维数的增高以及空间数据的缺值是当前空间数据复杂性的主要表现特征,并以其为线索将近年来在空间数据知识发现领域的研究进展及其热点进行了较为系统的归纳. 在此基础上,对空间数据知识发现与 GIS 的关系进行了阐述,并对空间数据知识发现的未来发展趋势进行了展望.

**关键词** 空间数据 知识发现 评述
**中图法分类号**: TP18 **文献标识码**: A **文章编号**: 1006-8961(2001)09-0854-07

# Review on the Proceedings of Spatial Data Mining Research

PEI Tao, ZHOU Cheng-hu, LUO Jian-cheng, HAN Zhi-jun
WANG Min, QIN Cheng-zhi, CAI Qiang
(*State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing* 100101)

**Abstract** In this paper, the authors analyze the increasing trend of spatial data and propose that the large data sets, nonlinear relationship among attributes, scaling characters of spatial data, fuzzy character of spatial information, multidimensional attributes and missing data problems are major characters of the complexity characters of spatial data. The proceedings of spatial data mining researches are systematically summarized in the clue of the complexity characters which mentioned above. In conclusion, the relationship between spatial data mining and GIS is expatiated, and the future of the relative research areas are prospected.

**Keywords** Spatial data, Knowledge discovery, Review

## 0 引 言

空间数据挖掘和知识发现(Spatial Data Mining 和 Knowledge Discovery form Spatial Database)是指从空间数据中提取信息和发现知识. 近年来,这方面的研究已然成为空间信息处理领域的热点,其中的原因主要源自两个方面:

(1)由于近年来空间信息技术领域内观测技术、网络技术的飞速发展以及台站建设的普及和不断完善,包括资源、环境、灾害的各种空间数据呈指数级数增长;

(2)专职处理空间数据的地理信息系统(GIS)在近十几年来虽得到了广泛的应用,并在空间数据的存储、查询以及显示等方面有了较快的发展,但面对数据量日益增长和种类繁多的空间数据,因其空间分析多以图形操作为主(如缓冲区操作,空间叠加,邻近分析以及空间连接等等),故而在空间信息的深入提取和知识发现等方面的功能仍相对薄弱[1].

正是上述两方面原因使得我们拥有的海量空间数据与有用知识的获取之间存在尖锐的矛盾,并因此而促进了空间数据知识发现研究的飞速发展.

基金项目:中国科学院知识创新项目(CX10G-D00-06)
收稿日期:2001-06-30;改回日期:2001-07-16

## 1　空间数据的特点及其复杂性特征

空间数据的复杂性特征在很大程度上是由其特点所决定的,并成为空间数据知识发现研究首要解决的任务.

### 1.1　空间数据的主要特点

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(不仅多了拓扑关系、方位关系,而且度量关系还与空间位置和个体之间的距离有关),使空间数据与其他类型数据的知识发现方法之间存在明显的差异.

### 1.2　空间数据的复杂性特征

随着近年来信息技术的飞速发展,空间数据具备了以下几个方面的复杂性特征:

(1) 海量的数据

海量数据常使一些方法因算法难度或计算量过大而无法得以实施,因而知识发现的任务之一就是要创建新的计算策略并发展新的高效算法克服由海量数据造成的技术困难.

(2) 空间属性之间的非线性关系

空间属性之间的非线性关系是空间系统复杂性的重要标志,其中蕴含着系统内部作用的复杂机制,因而被作为空间数据知识发现的主要任务之一.

(3) 空间数据的尺度特征

空间数据的尺度性是指空间数据在不同观察层次上所遵循的规律以及体现出的特征不尽相同. 尺度特征是空间数据复杂性的又一表现形式,利用该性质可以探究空间信息在概化和细化过程中所反映出的特征渐变规律.

(4) 空间信息的模糊性

空间数据复杂性的另一个特征就是模糊性. 模糊性几乎存在于各种类型的空间信息中,如空间位置的模糊性、空间相关性的模糊性以及模糊的属性值等等.

(5) 空间维数的增高

空间数据的属性增加极为迅速,如在遥感领域,由于感知器技术的飞速发展,波段的数目也由几个增加到几十甚至上百个,如何从几十甚至几百维空间中提取信息、发现知识则成为研究中的又一障碍.

(6) 空间数据的缺值

缺值现象源自由于某种不可抗拒的外力而使数据无法获得或发生丢失. 如何对丢失数据进行恢复并估计数据的固有分布参数,成为解决数据复杂性的难点之一.

空间数据所表现出的上述复杂性特征为相应的知识发现研究提出了更高的要求,并成为推动其发展的强大动力.

## 2　近年空间数据知识发现研究的主要进展

### 2.1　针对海量数据的算法研究

提高计算效率是针对海量数据顺利实施知识发现算法的主要手段之一. 解决算法效率的方法主要分为 3 种:

(1) 改变算法运行的策略:其主要方式为,采用并行运算环境,实施并行算法[2~5],如在大型数据库中实施决策树分类、空间聚类以及关联规则发现等算法时采用了并行策略,大幅提高了计算效率;

(2) 提高数据库查询语言的效率,如文献[6]中提出的效率和性能更好的规则提取和查询语言 MSQL;

(3) 对原有算法的结构进行改进,从而减小运算的复杂度,如 Cios *et al*. 改进决策树算法 CLILP2 后提出的 CLIP3 算法[7],不仅大大减小了存储空间,而且提高了运算效率.

### 2.2　以神经网络为代表的智能方法成为解决空间非线性关系的主要工具

#### 2.2.1　神经网络学习算法的发展

神经网络作为模拟复杂系统非线性关系的一种模型[8],按照其内部神经元连接的拓扑结构、学习规则以及传递函数的类型等标准可以分为若干种类,较常见的有:前向网络(BP)、Hopfield 网络、自组织映射网络(SOM),径向基函数网络(RBF)等等.

由于神经网络非常适用于处理空间数据的非线性复杂关系,并且在处理复杂问题时不需了解网络内部所发生的结构变化,因而被广泛地应用于空间数据挖掘和知识发现的研究中[9],并以不同的网络模型分别实现了空间聚类、分类、关联、回归、模式识别等多种算法[9~11].

当神经网络在众多领域内取得很多应用成果之后,人们又将目光转移到神经网络的内部,试图解释这一黑箱的运作机制. 当前研究的热点之一就集中在神经网络学习的优化算法上. 以前的网络学习算法(如,梯度下降法等)由于容易陷入局部极值,形成

网络结构早熟,致使计算结果出现偏差.近年来,在神经网络学习算法中引入了各种演化算法作为优化策略,其一是模拟退火算法[12](Simulated Annealing,SA),另一种为遗传算法[13,14](Genetic Algorithm,GA).模拟退火算法是模拟金属材料在加温后的退火过程.退火处理的目的是细化金属晶粒,并使材料更具韧性.遗传算法模拟的是生物的自然进化过程,其结果是最终演化成最适于环境的群体.虽然两种方法所模拟的对象大相径庭,但所得到的结果却十分相似.

神经网络在众多领域广泛应用的同时,也遇到一些难以解决的问题.如在对付数据量巨大、非线性程度很高的数据集时,神经网络存在着学习速度慢、难以收敛等问题;而对于采用自组织增量式学习方式的网络,会使其结构急剧膨胀,甚至崩溃;此外,神经网络的另一突出弱点就是当使用带有噪声的数据训练网络时,往往会因训练过程的控制不当而使网络产生过度拟合(overfitting)现象,从而影响网络预测的结果.

2.2.2 统计学习领域的研究热点——支撑向量机

支撑向量机(Support Vector Machine,SVM)是一种基于统计学习理论的一般性构造学习方法,其理论由 Vapnik 于 1995 年提出.其主要思想为,在高维空间内利用线性函数的对偶核,并通过内积空间的向量运算来处理线性不可分数据[15].

支撑向量机(SVM)的主要优点在于:该模型利用优化对偶理论使高维特征空间中的模型参数估计易于计算,并且运算的复杂度与问题的维数关系不大.

支撑向量机模型在学习效率,解决 overfitting 问题,全局最优化等方面都表现出优于神经网络的良好性质[16];在解决空间数据的分类、特征识别、图象压缩等问题方面也取得了一定进展[17].从 SVM 产生的背景和应用的效果来看,该模型特别适合处理高维、复杂的目标识别问题,其在遥感影象理解,特别是对复杂地学信息的识别等方面显示出广阔的应用前景.

2.2.3 机器学习中熵标准的应用

自从普利高津用耗散结构理论结束了热寂学说长期的"黑暗统治"之后,人们才发现"熵"原来也可以依靠"外力"而减小,于是熵的变化就成为研究复杂系统演化的"风向标".随着非线性热力学和信息理论的不断推广和普及,信息熵作为一种衡量信息量的指标之一,最终也进入了空间数据知识发现研究的领域.

信息熵之所以被广泛的应用,其主要原因是信息熵用简单的方式定义了系统的复杂性,并具有明确的物理含义,即信息熵是在平均的意义上来表征信息源的总体特征[18].

在空间数据发现的各方面研究中,信息熵的作用在机器学习领域中得到了充分的展示.机器学习按其研究范畴有狭义和广义之分.广义的机器学习泛指一切通过学习可以改进自身性能的算法(包括神经网络和支撑向量机等算法);而狭义的机器学习是特指通过学习,对数据进行归纳和泛化,从而产生规则的过程.与神经网络模型所代表的"黑箱"相比,狭义的机器学习所产生的规则是显式的,并且可以被修改、学习和使用[19].在下文中将采用狭义的机器学习定义.

机器学习按照是否需要先验知识和学习样本可分为有监督的机器学习(supervised learning)和无监督的机器学习(unsupervised learning).机器学习分为 3 种类型[20]:规则算法(Rule Algorithm)、决策树算法(Decision Tree Algorithm)以及综合前面两种算法的杂交算法(Hybrid Algorithm).

规则算法的主要代表有 AQ 和 AQ15;决策树算法的主要代表有 ID3 及其改进方法 C4.5、CID3,而杂交算法的主要代表为 CN2,CLILP2 以及 CLIP3[7,21~23].

规则算法和决策树算法实际上可以看作是一种对样本点的分割,所不同的是规则算法的分割是建立在信息函数(Information function)的基础上,而决策树则是将信息增益(Information gain)或信息熵(Information entropy)作为分割的标准.杂交算法将两种方法的概念结合起来,通过解决整数规划模型(Integer Programming)来选择最好的分割特征(feature)和总结规则.

Quinlan 将 Shannon 的信息熵概念引入到决策树的算法中,并作为寻找最明显特征的标准①.此外,在这方面的应用还有,使用神经网络创建决策树和规则算法中的决策界面等[21];利用地学信息熵作为空间数据分割的标准对土壤数据等空间数据进行了空间分割的尝试,从而将空间属性与熵标准的判定有机地结合起来[24].

① Quinlan J R. C4.5 Programs for Machine Learning, Morgan-Kaufmann,1993.

### 2.3　尺度空间概念的应用

关于空间数据的尺度特征,邸凯昌等认为,空间数据是包含了尺度维的四维发展状态空间,尺度维所反映的是空间数据由细到粗多比例尺或多分辨率的几何变换过程[25]. 要刻画空间数据的尺度性,建立一种空间数据分辨率由细到粗的序列是其关键. Witkin 和 Koenderink 最先提出了尺度空间的概念,并将之应用于图象结构的表达[26,27].

尺度空间构造的基本原理就是将空间数据集"投影"到不同分辨率的空间内,并挖掘尺度空间下的"知识". 这一过程可以利用视觉想象进行类比. 在最高的分辨率下,空间中的每一点可视为一个小光点,整个的空间数据就成为完整的一幅图象,当逐渐远离这幅图象时,小光点变得模糊,进而融合为小光斑,当图象进一步模糊时,多个小光斑又融合为大光斑……,这一过程不断重复下去直至所有的数据点都融为一个大光斑,就完成了尺度空间的构造[28].

建立尺度空间的方法多种多样,主要有以下几类:小波滤波[29](wavelet filter)、高斯平滑[30](Gaussian Smoothing)以及高斯导数滤波(Gaussian derivative filter)、Gabor 滤波等[31].

尺度空间概念应用的领域包括空间特征的识别和空间尺度聚类等方面. 在空间特征识别方面 Andrew D J Cross 和 Edwin R Hancock 借助向量场算子理论对尺度空间内实物图象的边界和对称性进行了识别[32]. 张讲社等借鉴视觉的基本理论将尺度空间应用到聚类算法中,并确定聚类结果的有效性,从而减少在聚类过程中的人为干预[28].

### 2.4　模糊集和粗集理论的应用

对于空间关系中的不确定性,通常采用模糊集理论加以描述. 模糊集理论的优势在于利用隶属函数刻画空间关系的不确定性,用部分归属代替了归属的概率. 模糊集的思想已渗透到空间数据知识发现的各种方法之中,如,模糊聚类与分类、模糊神经网络、模糊专家系统等等[33].

在实际应用中,模糊集可被应用于 GIS 中主题图的准确性评估以及面积估计[34]. 此外,在土地覆盖数据的查询中,可将模糊集的运算代替多条件的复合查询,并在 GIS 中确定符合复合查询条件的区域[35].

隶属函数虽然对不确定关系进行了成功的刻画,打破了非此即彼的传统概念,但其确定仍然需要借助先验的知识,从而导致结果的多解性[36].

近年来崛起的粗集理论取模糊概念之长,去隶属函数之短,从而成为研究模糊现象的又一有力工具. 粗集理论不需要先验假设,而是利用集合论中的上近似(up approximation)和下近似(bottom approximation)来刻画集合:当个体 $A$ 属于集合 $X$ 的下近似时,$A$ 肯定属于集合 $X$;而当 $A$ 不属于集合 $X$ 的上近似时,则 $A$ 肯定不属于集合 $X$;如果 $A$ 属于 $X$ 的上近似而不属于 $X$ 的下近似,则 $A$ 有可能属于集合 $X$.

文献[37]中将粗集理论应用到 GIS 数据的分类研究中,利用二维扩展误差矩阵(two-dimensionally Extended error matrix)进行空间数据的多次分类,并对不同分类的结果合并. 文中对瑞典 Stockholm County 的植被类型利用粗集分类算法进行了重新分类实验.

此外,文献[38]中还将粗集应用到瑞典冰川空间分布的研究中,并对 Fennoscandia 最近的冰期状态进行了重新解释.

以上是针对矢量数据应用的粗集分类方法. 对于栅格数据,Aldridge 利用 RS-GKDD(rough set based geographic knowledge induction)方法对新西兰 Dunedin 附近的滑坡影象进行了分析. 它首先利用粗集方法对所有属性进行筛选,并建立一个优化条件属性集;然后根据这些属性集找出"最近的"统计显著规则. 该项研究最终发现了高程、岩性与滑坡之间的规则关系①.

在目前的 GIS 应用中,对数据进行概括的手段主要是执行 Zoom in、Zoom out 等功能[39],但要从真正的意义上做到数字概括,仍然是比较困难的,必须发展一种抽象和浓缩数据的方法,同时还必须保证算法执行过程中数据的质量. 文献[39]中提到,粗集理论正好可以凭借不需定量化不确定性的优势来做到这一点.

除了上文谈到的利用模糊集和粗集处理空间数据的不确定以外,还有云模型[40],该模型将模糊性与随机性有机结合,并从另一角度解决了模糊集理论中隶属函数的固有缺陷.

### 2.5　高维数据的挖掘算法

要解决高维数据的挖掘算法,首先必须了解高

① Colin H Aldridge. Discerning landslide hazard using a rough set based geographic knowledge discovery methodology. Presented at SIRC 99 — The 11th Annual Colloquim of the Spatial Information Research, Centre University of Otago, Dunedin, New Zealand, December 13-15th, 1999

维数据的性质. 与低维空间的性质对比,高维空间所表现出的性质与我们日常接触的三维空间的性质完全不同[41]. 例如,超立方体与其内接球的体积之比随维数增大而趋于 0、超立方体对角线与坐标轴之间的夹角随维数的增加逐渐逼近 90 度等等[42,43].

在近期的研究中,对高维数据进行挖掘的思路一般有两条,一是将高维数据通过线性变换投影到低维空间,然后再实施其他挖掘算法;另一种就是采用适合处理高维数据的算法直接对其进行信息提取.

降维方法的主要问题在于,当维数无限增加时,由高维到低维的线性变换会掩盖数据原有的信息,而使数据呈现正态分布. 这样,原先在高维空间存在明显差异或特征的类别在低维的空间内会混杂在一起难以区别[44],因而高维空间向低维空间线性变换的关键就在于寻找合适的投影方向,将高维空间的目标特征信息尽可能忠实地投影到低维空间.

由高维向低维进行线性投影的方法也比较多,最常见的有主成分分析,空间因子分析及其相应的改进方法等[45].

在对高维数据直接进行信息提取的方面,Lee 和 Landgrebe 曾将一阶(均值)和二阶(方差)的统计方法应用于高维的数据分类中,并比较了它们之间的差异[46]. 此外,他们还发现将各维之间的相关矩阵进行对角化(diagonalizion)处理,有助于提高分类方法的准确性. 文中还创造性地利用色阶来表示类别之间相关性的大小,从而使结果更加直观.

除此之外,降维还可通过使用粗集理论精简维数而得以实现;而支撑向量机模型则能够应付处理数据因维数过高而产生的复杂性.

### 2.6 空间数据的缺值研究

对空间数据进行缺值研究的过程主要可以分为两步,首先,利用各种统计手段模拟出缺失值 $N_{mis}$,然后再利用包含已知观测值 $N_{obs}$ 和 $N_{mis}$ 的全集进行统计分布函数的参数估计. 用于产生缺失值的方法包括,均值转嫁、有补充的随机替代、无补充的随机替代、时间序列转嫁、完全均值转嫁等等[47]. 上述方法是将缺值的补充与分布函数的参数估计作为两个独立的部分来看待,而 EM 算法(Expectation-Maximization Algorithm)则将二者有机地联系起来[48],在处理非完整数据集时对其最大似然函数应用了迭代的方法,从而在模拟缺失样本的同时估计出分布函数的未知参数. 有关利用 EM 及相关算法对不同分布空间数据的缺值研究可参考文献[48]~[50].

## 3 结语与展望

空间数据本身的复杂性特征是空间数据挖掘和知识发现理论不断发展和完善的内在动力,并已在一定程度上左右着该理论前进的方向,因此在未来的一段时间内空间数据知识发现研究的主要任务仍然是解决空间数据中蕴藏的复杂性问题. 对于其他领域出现的新理论、新方法,当然是空间数据知识发现理论逐渐成熟所必须的养份,由此而产生的外力推动作用也不容忽视.

关于 GIS 的发展与空间数据知识发现研究之间的关系:一方面,GIS 技术在空间数据库功能以及图形操作等方面已取得了巨大的成功,但在空间信息提取和知识发现的功能上仍有所欠缺,急待相关理论和技术的补充和完善;另一方面,空间数据知识发现的理论和技术只有借助 GIS 强大的数据库和图形功能才能加速其前进的步伐,因而在未来的发展中,GIS 与空间数据知识发现理论的结合将成为一种必然的趋势,并为空间信息技术领域的研究注入新的活力.

## 参 考 文 献


1 王学军. 空间分析技术与地理信息系统的结合[J]. 地理研究,1997,16(3):70~73。

2 Anurag S, Han Eui-Hong, Vipin kumar *et al*.. Parallel formulation of decision-tree classification algorithm [J]. Data Mining and Knowledge Discovery, 1999, 3(3): 237~262.

3 Xu Xiaowei, Jager Jochen, Kriegel Hans-peter [J]. A fast parallel clustering algorithm for large spatial database. Data Mining and Knowledge Discovery, 1999, 3(3): 263~290.

4 Cheng David W, Xiao Yongqiao. Effect of data distribution in parallel mining of association [J]. Data Mining and Knowledge Discovery, 1999,3(3):291~314.

5 Xiang Y, Chu T. Parallel learning of belief networks in large and difficult domains [J]. Data Mining and Knowledge Discovery, 1999, 3(3):315~339.

6 Tomasz Imielinski, Aashu Virmani. MSQL: A query language for database mining [J]. Data Mining and Knowledge Discovery, 1999, 3(4):373~408.

7 Cois K J, Wedding D K, Liu N. CLIP3: Cover learning using integer programming [J]. Kybernetes, 1997, 26(4-5):513~536.

8 焦李成. 神经网络系统理论 [M]. 西安:西安电子科技大学出版社,1992.

9 周成虎,骆剑承等. 遥感影象地学理解与分析 [M]. 北京:科学出版社,1999.

10 Juha Vesanto, Esa Alhoniemi. Clustering of the selforganizing map [J]. IEEE Transactions on Neural Networks, 2000,11(3), 586～600.

11 MacKay D J C. A practical bayesian framework for backpropagation neural network [J]. Neural Computation, 1992, 4(3): 448～472.

12 Kirkpatrick S. Optimization by simulated annealing: Quantitative Studies [J]. Journal of Statistics Physics, 1984, 34(5): 975～986.

13 Holland J H. Adaptation in natural and artificial system [M]. Massa-chusetts: MIT Press, 1992.

14 Yao X. Evolving artifical neural networks [J]. Procee-dings of the IEEE, 1999, 87(9): 1423～1447.

15 Vapnik V. The nature of statistical learning theory [M]. New York: Springer Verlag, 1995.

16 Burges C J C. A tutorial on support vector machines for pattern recogn-ition [J]. Data Mining and Knowledge Discovery, 1998, 2(1): 121～167.

17 Lothar Hermes, Dieter Frieauff, Jan Puzicha *et al*. Support vector machines for land usage classification in landsat TM imagery[A]. In:Proc. of the IEEE International Geoscience and Remote Sensing Symposium[C], Hanburg,1999:348～350.

18 **沈永欢编. 数学手册[M]. 北京:科学出版社**,1997.

19 Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W. Data mining method for knowledge discovery[M]. Norwell: Kluwer Academic Publishers, 1998.

20 Quinlan J R. Simplifying decision trees[J]. International Journal of Man-Machine Studies, 1987, 27(2):221～234.

21 Cios K J, Liu N. Machine learning in generation of a neural network architecture: A continuous ID3 approach [J]. IEEE Transaction On Neural Networks, 1992, 3(2): 280～291.

22 Clark P, Niblett T. The CN2 algorithm [J]. Machine Learning, 1989, 3(2): 261～283.

23 Cios K J, Liu N. An algorithm which learns multiple covers via integer linear programming, part I-The ClILP2 Algorithm [J]. Kybernetes, 1995, 24(1): 29～50.

24 **周成虎,张健挺. 基于信息熵的地学空间数据挖掘[J]. 中国图象图形学报**,1999,4(11):946～951.

25 **邸凯昌,李德仁,李德毅. 空间数据挖掘和知识发现的框架[J]. 武汉测绘科技大学学报**,1997,22(4):328～332.

26 Witkin P. Scale-space ltering[A]. In:Proc. 8th Int. Joint Conf. Art. Intell. [C], West Germany. 1983: 1019～1022.

27 Koenderink J J. The structure of images [J]. Biological Cybernetics, 1984, 50(2):363～370.

28 **张讲社,徐宗本. 基于视觉系统的聚类:原理与算法[J]. 工程数学学报**,2000,17(**增刊**):14～20.

29 Zheng Yuan-jin, David B H Tay, Li Le-min. Signal extraction and power spectrum estimation using wavelet transform scale space filtering and Bayes shrinkage [J]. Signal Processing, 2000, 12(6):1535～1549.

30 Lin Hsin-Chih, Wang Ling-Ling, Yang Shi-Nine. Automatic determination of the spread parameter in Gaussian smoothing [J]. Pattern Recognition Letters, 1996, 17(5):1247～1252.

31 Li Xiao-ping, Chen Tong-wen. Efficient synthesis of parameterized gaussianlike filters by approximation [J]. Signal Processing, 1995,7(1):119～134.

32 Andrew D J Cross, Edwin R Hancock. Scale space vector fields for symmetry detection[J]. Image and Vision Computing, 1999, 17(2):337～345.

33 Kent J T, Merdia K V. Spatial classification using fuzzy membership model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988, 10(5): 659～671.

34 Curtis E Woodcock, Sucharita Gopal. Fuzzy set theory and thematic maps: Accuracy assessment and area estimation[J]. Int. J. geographical information science, 2000, 14(2):153～172.

35 Bruin S De. Querying probabilistic land cover data using fuzzy set theory[J]. Int. J. geographical information science, 2000, 14(4):359～372.

36 Burrough P A, McDonnell R A. Principles of geographical information systems[M]. London: Oxford University Press, 1998.

37 Ola Ahlqvist, Johannes Keukelaar, Karim Oukbir. Rough classification and accuracy assessment[J]. Int. J. Geographical Information Science, 2000,14(5):475～496.

38 Hättestrand C. The glacial geomorphology of central and northern sweden ser. Ca 85 [M], Stockholm: Sveriges Geologiska Undersökning,1998.

39 Müller J-C, Wiebel R, Lagrange J-P *et al*. Generalization: State of the art and issues [A]. In: GIS and Generalization: Methodology and Practice [C]. London: Taylor & Francis, 1995:3～17.

40 **邸凯昌,李德毅,李德仁. 云理论及其在空间数据发掘和知识发现中的应用[J]. 中国图象图形学报**,1999,4(11):930～935.

41 Luis Jimenez, David Landgrebe. Supervised classification in high dimensional space: Geometrical, statistical and asymptotical properties of multivariate data [J]. IEEE Transaction on Systems, Man, and Cybernetics, 1998, 28(1): 39～54.

42 Scott D W. Multivariate density estimation[M]. New York: John Wiley & Sons, 1992.

43 Wegman E J. Hyperdimensional data analysis using parallel coordinates [J]. Journal of the American Statistical Association, 1990, 85(3):664～675.

44 Hall P, Li K On. Almost linearity of low dimensional projections from high dimensional data[J]. The Annals of Statistics, 1993, 21(2):867～889.

45 Grunsky E C, Agterberg F P. SPFAC: A FORTRAN-77 program for spatial factor analysis of multivariate data [J]. Computer & Geosciences, 1991,17(1):133～160.

46 Lee Chulhee, David A Landgrebe. Analyzing high dimensional multispectral data[J]. IEEE Transactions on Geoscience and Remote Sensing, 1993,31(4):792～800.

47 Little R J A, Rubin D B. Statistical analysis with missing data [M]. New York: Wiley, 1987.
48 Todd K Moon. The expectation-maximization algorithm [J]. IEEE Signal Processing Magazine, NOV., 1996,47～60.
49 Barroso Lúcia P, Wilton O Bussab, Martin Knott. Best linear unbiased predictor mixed model with incomplete data [J]. Communications in Statistics: Theory and Methods, 1998, 27(1):121～129.
50 Luttrell S P. Partitioned mixture distribution: An adaptive bayesian network for low-level image processing [J]. IEEE Proceedings on Vision, Image and Signal Processing, 1994, 141(4):251～260.

**裴 韬** 1972 年生,1998 年获中国地质大学(武汉)博士学位,现为中科院资源与环境信息系统国家重点实验室副研究员. 近年从事空间数据库知识发现、GIS 应用等方面的研究工作.

**周成虎** 1964 年生,研究员,1992 年获中国科学院地理研究所博士学位,现为中科院资源与环境信息系统国家重点实验室主任. 长期从事洪水、GIS 应用以及数据挖掘等方面的研究.

**骆剑承** 1970 年生,1999 年获中国科学院地理研究所博士学位,现为中科院资源与环境信息系统国家重点实验室副研究员. 主要从事遥感数据的智能理解、GIS 应用与开发等方面的研究工作.

**韩志军** 1970 年生,博士后,2000 年获中国地质大学(武汉)博士学位. 主要从事数据库分析以及 GIS 应用等方面的研究工作.

**汪 闽** 1975 年生,博士生,2000 年获浙江大学地质系硕士学位. 主要研究方向为空间数据知识发现.

**秦承志** 1977 年生,博士生,2000 年获中国科学院兰州地质所硕士学位. 主要研究方向为空间数据知识发现.

**蔡 强** 1978 年生,硕士生,1999 年获北京大学城市环境系学士学位. 主要方向为空间数据知识发现.