文章编号: 1000-050X(2002)03-0221-13 文献标识码: A

论空间数据挖掘和知识发现的理论与方法

李德仁[1] 王树良[1] 李德毅[2] 王新洲[3]

(1 武汉大学测绘遥感信息工程国家重点实验室, 武汉市珞喻路 129 号, 430079)
(2 中国电子系统工程研究所, 北京市万寿路 6 号, 100039)
(3 武汉大学测绘学院, 武汉市珞喻路 129 号, 430079)

摘 要: 首先分析了空间数据挖掘和知识发现(SDMKD)的内涵和外延; 然后分别研究了用于 SDMKD 的概率论、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、可视化、决策树、空间在线数据挖掘等理论和方法及其进展; 最后展望了 SDMKD 的发展前景。

关键词: 空间数据挖掘; 知识发现; 理论方法
中图法分类号: P208; TP309.13; TP18

由于雷达、红外、光电、卫星、电视摄像、电子显微成像、CT 成像等各种宏观与微观传感器的使用, 空间数据的数量、大小和复杂性都在飞快地增长, 已经远远超出了人的解译能力。终端用户不可能详细地分析所有的这些数据, 并提取感兴趣的空间知识, 致使"空间数据爆炸但知识贫乏"。因此, 利用空间数据挖掘和知识发现[1](SDMKD, spatial data mining and knowledge discovery)从空间数据库中自动或半自动地挖掘事先未知却潜在有用的空间模式变得十分必要。空间数据基础设施为 SDMKD 构造了大环境。SDMKD 当前相当于数据库技术在 70 年代所处的地位, 迫切需要理论和方法指导, 需要类似于关系模式、DBMS 系统和 SQL 查询语言等模型和工具, 才能使 SDMKD 的应用得以普遍推广。其中, 理论方法的好坏, 将直接决定 SDMKD 所发现知识的优势。

1 SDMKD 的基本涵义

SDMKD 挖掘的空间知识主要包括空间的关联、特征、分类和聚类等规则及例外。一般表现为一组概念、规则、法则、规律、模式、方程和约束等形式的集合, 是对数据库中数据属性、模式、频度和对象簇集等的描述。SDMKD 是计算机技术、数据库应用技术和管理决策支持技术等发展到一定阶段、多学科交叉的新兴边缘学科, 汇集了来自机器学习、模式识别、数据库、统计学、人工智能以及管理信息系统等各学科的成果[2]。SDMKD 因解决"空间数据爆炸但知识贫乏"的现象而发展。

SDMKD 是数据挖掘和知识发现(DMKD, data mining and knowledge discovery)的分支学科, 但 SDMKD 不同于普通的 DMKD, 它的对象是空间数据库或空间数据仓库, 有别于常规的事务型数据库, 比一般数据挖掘的发现状态空间理论[3]增加了尺度维(scale)。机器学习侧重于设计新的方法从数据库中提取知识的技术行为, 而 SDMKD 是从已经存在于空间数据库中的数据内挖掘知识的过程。与传统的地学数据分析相比, SDMKD 更强调在隐含未知情形下对空间数据本身分析上的规律挖掘, 空间知识分析工具获取的信息更加概括、精练。高于空间数据库的空间数据仓库, 遵循一定的原则用多维数据库来组织和显示数据, 将不同数据库中的数据粗品汇集精化成为半成品或成品(数据件), 稍加整理可被直接用于 SDMKD。在数字地球中, 空间数据挖掘的对象一般为空间数据仓库[4]。

收稿日期: 2002-01-09。
项目来源: 国家重点基础研究发展规划(973 计划)资助项目(G19980305084); 国家高技术研究发展计划(863 计划)资助项目(2001AA135081); 国家自然科学基金资助项目(40023004; 49874002); 教育部博士点基金资助项目(98049801); 香港理工大学科研基金资助项目(1.34.37.9709)。

2 SDMKD 的理论和方法

目前，用于表示空间数据的方法多数是基于经典的确定集合理论（如概率论），每一个空间实体都与单一的属性说明有关，属性之间被表示为清晰的边界。可是，复杂多变的现实世界并非总是如此，故当研究不能精确描述的空间实体的空间数据时，经典集合理论被扩展（如云理论），用于实现定量和定性的相互转换。

2.1 概率论

概率论（pobability theory）根据随机概率挖掘含有不确定性的空间数据库，发现的知识被表示成给定条件下某一假设为真的条件概率，常用作背景知识。在用误差矩阵描述遥感分类结果的不确定性时，可以用这种背景知识表示不确定性的置信度。Lenarcik 和 Piasta[5] 把概率论和粗集相结合，利用条件属性推理决策知识，开发了 ProbRough 系统（probabilistic rough classifiers generation）。利用基于决策树的概率图模型，Frasconi、Gori 和 Soda[6] 对带有图形属性的数据库进行挖掘，得到了用于指导机器学习的知识。区域土壤属性的空间分析常通过概率样本来实现。对于无概率样本的区域，Brus 和 Gruijter[7] 设计了利用概率样本内插回归的方法估求其土壤属性的一般模式；Sester[8] 利用基于导师的机器学习技术，从给定样本空间数据库中获取了指导空间数据自动解译和操作的知识。

2.2 证据理论

证据理论（evidence theory）是概率论的一个扩展（又称 dempster-shafer 理论），是由可信度函数（度量已有证据对假设支持的最低程度）和可能函数（衡量根据已有证据不能否定假设的最高程度）所确定的一个区间[9]。当证据未支持部分为空时，证据理论等同于传统概率论。证据理论将实体分为确定部分和不确定部分，可以用于基于不确定性的空间数据挖掘。利用证据理论的结合规则，可以根据多个带有不确定性的属性进行决策挖掘[10]。两两比较法也用于属性不确定性的知识发现[11,12]。证据理论发展了更一般性的概率论，却不能解决矛盾证据或微弱假设支持等问题。

2.3 空间统计学

空间统计学（spatial statistics）是依靠有序的模型描述无序事件，根据不确定性和有限信息分析、评价和预测空间数据[13]。它主要运用空间自协方差结构、变异函数或与其相关的自协变量或局部变量值的相似程度实现基于不确定性的空间数据挖掘。基于足够多的样本，在统计空间实体的几何特征量的最小值、最大值、均值、方差、众数或直方图的基础上，可以得到空间实体特征的先验概率，进而根据领域知识发现共性的几何知识。空间统计学拥有较强的理论基础和大量的成熟算法，能够改善 GIS 对随机过程的处理，估计模拟决策分析的不确定性范围，分析空间模型的误差传播规律，有效地综合处理数值型空间数据，分析空间过程，预测前景，并为分析连续域的空间相关性提供理论依据和量化工具等。所以，空间统计学是基本的数据挖掘技术，特别是多元统计分析（如判别分析、主成分分析、因子分析、相关分析、多元回归分析等）。

Cressie[13] 利用地理统计数据、栅格数据和点数据三种空间数据描述现实世界，并据此提出了一个通用模型。由于大部分空间数据挖掘的研究偏重于提高静态数据查询的效率，所以 Wang、Yang 和 Muntz[14,15] 基于统计信息，研究了一种由用户定义的主动空间数据挖掘的方法。应用空间统计学的克吕格方法，由一组已分类的观测点直接估计未观测点位的属于各类别的验后概率，求得类别变量在任一位置上所观测到的各类别的概率知识，就可以从影像上获取模糊分类信息[16,17]。冯建生[18] 也利用空间统计学揭示了影响冲击韧性的因素知识。

但是，空间统计学的数据不相关假设，在空间数据库或空间数据仓库中常常难以得到满足，当实际数据互相依赖时将引起问题。回归模型可以缓和这一矛盾，但建模过程非常复杂。空间统计学能够有效地处理数值型数据，却难以分析字符型数据。在应用空间统计学时，需要同时具备空间领域知识和统计学知识，故不适合于终端用户。统计方法也不能对非线性规划或符号值精确建模，不能处理不完整、不确定性数据，计算代价昂贵。同时，当知道非匀质实体的某种属性可能发生，但却不知道也难以构建其概率分布模型时，模糊集比空间统计学更利于发现隐藏在这种不确定性中的知识。

2.4 规则归纳

规则归纳（rules induction）是在一定的知识背景下，对数据进行概括和综合，在空间数据库或空间数据仓库中搜索和挖掘以往不知道的规则和规律，得到以概念树形式（如 GIS 的属性概念树和空间关系概念树）给出的高层次的模式或特征。背景知识可以由 SDMKD 的用户提供，也可以作

为 SDMKD 的任务之一自动提取。在推理方法中,归纳不同于基于公理和演绎规则的演绎,以及基于公认知识的常识推理,而是根据事例或统计的大量事实和归纳规则进行的。决策规则是数据库中总的或部分的数据之间的相关性,是归纳方法的扩充,其条件为归纳的前提,结果为归纳的结论,大致包括关联规则、顺序规则、相似时间序列、If-Then 规则等。

空间关联规则发现是 SDMKD 的重要内容。目前的研究主要集中在提高算法的效率和发现多种形式的规则两方面,并以逻辑语言或类 SQL 语言方式描述规则,以使 SDMKD 趋于规范化和工程化。一条空间关联规则可表示为 $\boldsymbol{X}\rightarrow\boldsymbol{Y}$ ($c\%$, $s\%$, $i\%$),其中, $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是空间或非空间谓词的集合, $c\%$、$s\%$和 $i\%$分别是规则的可信度、支持度和兴趣度。Koperski 和 Han[19] 提出了一种在地理信息数据库中挖掘强空间关联规则(空间数据库中使用频率较高的模式或关系)的算法,并给出了两步式的空间优化技术。程继华和施鹏飞[20] 提出了多层次关联规则的挖掘算法,利用集合"或"、"与"运算求解频繁模式,提高了挖掘的效率。许龙飞和杨晓昀[21] 分析了广义关联规则模型的挖掘方法、挖掘策略和规则挖掘语言。Han、Karypis 和 Kumar[22] 提出了挖掘关联规则的智能数据分配和混合分类两个 Apriori 并行算法。Eklund、Kirkby 和 Salim[23] 在土壤盐度分析中把决策支持系统和 GIS 数据相结合,发现了用于环境规划和二级土壤盐碱化监测的关联规则。Aspinall 和 Pearson[24] 把风景生态学、环境模型和 GIS 结合在一起,通过综合地理评估,研究了美国黄石国家公园的汇水处环境条件,发现了用于环境保护的关联规则。涂星原[25] 研究了基于数值属性的关联规则的挖掘。Clementini、Felice 和 Koperski[26] 在宽边界的空间实体中挖掘出了多层次的空间关联规则。左万利[27] 研究了在含有类别属性的数据库中提取关联规则的类型转换技术。丁祥武[28, 29] 在关联规则模型中增加了描述关联规则时效性的时态信息。丁祥武[30] 根据数据记录之间的时间间隔和相邻记录中项目的类别合并同类记录,肖利等[31] 用时间窗刻画时间约束。程继华等[32] 提出了基于概念的关联规则的挖掘算法。肖利等[33] 提出了一个基于关系操作的挖掘广义关联规则算法,在多概念层上交互挖掘关联规则。面向属性归纳(attribute oriented induction,简称 AOI),亦称概念提升,适于数据分类[34]。但对于涉及不同主题地图信息的系统,要求面向属性归纳方法能够分析不同主题地图上的不同空间特征之间的关系。

当数据之间的规律无法用关联规则描述时,肖利等[35] 挖掘的转移规则描述系统此时期到下个时期的状态按照一定的概率进行转移,下个时期的状态取决于前期的状态和转移概率。朱明、王俊普和蔡庆生[36] 在实例特征矩阵的基础上,提出了一个最优特征集的启发式搜索算法,并将其与特征选择的贪心算法[37] 相比较。

序列规则和时间紧密相关。Kriegel 等[38] 在分析巴西 Logoa de Araruama 的盐渍海岸礁湖的时间序列的空间数据时,发现了保持水文和盐分平衡的知识。丁祥武[30] 提出了序列规则中相邻项目集之间的时间间隔约束,欧阳为民和蔡庆生[39] 将序列模式的发现从单层概念扩展到多层概念,提出了自顶向下逐层递进的方法,在不同概念层发现序列模式。偏离检测是数据挖掘的一种启发式方法,欧阳为民和蔡庆生[40] 将使数据序列突然发生大幅度波动的数据认作例外,提出了一种线性的偏离检测算法。在数据库变化不大时,渐进式序列规则挖掘算法能够利用前次的结果加速本次挖掘过程[41]。序列规则挖掘还有序列规则的维护等问题尚待解决。

此外,杨学兵等[42] 的实时数据挖掘算法能在实时过程控制中自动挖掘,并根据挖掘的知识预测趋势。Levene 和 Vincent[43] 发现了关系数据库的功能独立和包含独立的规则,信息处理使用了基于知识规则挖掘的分类方法[44, 45]。当没有背景知识时,空间数据挖掘应该考虑聚类分析。

2.5　聚类分析

聚类分析(clustering analysis)主要是根据实体的特征对其进行聚类或分类,按一定的距离或相似测度在大型多维空间数据集中标识出聚类或稠密分布的区域,将数据分成一系列相互区分的组,以期从中发现数据集的整个空间分布规律和典型模式[46]。聚类分析是统计学的一个分支,与规则归类不同的是,聚类算法无需背景知识,能直接从空间数据库中发现有意义的空间聚类结构。已有的聚类算法多为模式识别设计,用特征表示的目标为多维特征空间的一个点,在特征空间中聚类。空间数据库中的聚类是对目标的图形直接聚类,聚类形状复杂,数据量庞大,使用经典的基于多元统计分析的聚类法则速度慢、效率低。这对空间数据挖掘中的聚类算法提出了更高要求,如能处理点、线、面等任意形状,计算效率高,算法需要的参数能自动确定或用户易确定。Murray

和 Estivill-Castro[47] 回顾了探测性空间数据分析的聚类发现技术。

聚类算法主要有分割和层次两类。分割算法根据目标到聚类中心的距离迭代聚类，适用于聚类为凸形、类间相距较远且直径相差不悬殊的情况，否则会分割错误[48]。为了改善分割算法，在 CLARANS(cluster analysis algorithms)的基础上，Ng 和 Han[49] 提出了随机搜索的改进 K-medoid 算法，Ester 等[50] 用基于 R 树的数据聚焦法进一步提高其效率。周成虎和张健挺[51] 则提出了基于信息熵的时空数据分割聚类模型。层次算法将数据集分解成树状图子集，直到每个子集只包含一个目标，可用分裂或合并的方法构建。它无需参数，但要定义停止条件。

概念聚类是分割算法的一种延伸，它用描述对象的一组概念取值将数据划分为不同的类，而不是基于几何距离来实现数据对象之间的相似性度量[52]。概念聚类能够输出不同类以确定其属性特征的覆盖，并对聚类结果给予解释。当利用概念聚类实行空间数据挖掘时，需对数值型字段数据概念化。Han 和 Fu[53] 先用相同的小间隔将数值字段中的数据分段，然后将数据段合并成期望的数据个数基本相同的概念段，实现较简单，效率较高。但是他用于分割每个数值型字段的数据段的 Interval 是一个不变的量，用于概念分段的标准只有数据个数，没有考虑数据的分布。为此，李世祥和李涛涛[54] 采用变间隔分割数据，考虑了概念分段时的数据个数和数据分布。

Ester、Kriegel 和 Xu[55] 使用聚类技术研究了在大型空间数据库中挖掘类别判读知识的技术。Knorr 和 Ng[56] 分析了空间数据挖掘中的聚类和特征关系，提出了发现聚合亲近关系和公共特征的算法。Edwin 等[57] 通过构造地理信息系统中的聚类器，发现了空间物体的边界形状匹配关系的部分规律。Lin、Zhou 和 Liu[58] 根据类别和特征，研究了空间数据库中的临近关系匹配算法。Tung、Hou 和 Han[59] 提出了一种在空间数据挖掘中实行空间聚类时，处理河流、高速公路等阻隔的算法。Murray 和 Shyy[60] 在分布显示和空间数据挖掘中集成了属性和空间特征，提出了一种交互的探测性空间数据聚类分析技术。

此外，还有基于密度的 DBSCAN 算法[48]、针对栅格数据的基于数学形态学的算法[61]、模糊聚类[63] 和神经网络聚类方法[18] 等。

2.6 空间分析

空间分析(spatial analysis)是利用一定的理论和技术对空间的拓扑结构、叠置、图像、空间缓冲区和距离等进行分析的方法总称，目的在于发现有用的空间模式。探测性的数据分析(exploratory data analysis, EDA)采用动态统计图形和动态链接技术显示数据及其统计特征，发现数据中非直观的数据特征和异常数据。Ester、Kriegel 和 Sander[63] 在空间数据库管理系统的基础上，基于邻图和邻径，提出了针对空间数据库的挖掘空间相邻关系知识的算法。邸凯昌[61] 把探测性的数据分析与空间分析相结合，构成探测性的空间分析(exploratory spatial analysis, ESA)，再次与 AOI 结合，则形成探测性的归纳学习(exploratory inductive learning, EIL)，它们能在 SDMKD 中聚焦数据，初步发现隐含在数据中的某些特征和规律。Murray 和 Estivill-Castro[47] 对探测性空间数据分析的聚类发现技术作了回顾。图像分析可直接用于发现含有大量图形图像数据的空间数据挖掘，也可作为其他知识发现方法的预处理手段。

Reinartz[64] 给出了他关于现实世界的数据挖掘方案及其实验结果。高光谱成像获取的地表图像包含了丰富的空间、辐射和光谱三重信息。王晋年等[65] 认为高光谱信息挖掘技术是高光谱数据应用延拓与深入的重要环节，其核心在于光谱信息的挖掘。他们基于高光谱遥感信息的特点，探讨分析了以地物识别与分类为目标的高光谱数据挖掘技术，包括基于模式识别的高光谱信息挖掘技术、基于光谱波形特征的挖掘技术以及亚像元光谱信息挖掘。马建文和马超飞[66] 分析了地面物质和结构光谱与卫星遥感信息之间的关系，建立了空间角度模型，通过对 TM 卫星数据的挖掘说明了基于空间角度算法在处理多波段遥感数据时的数学能力。布和敖斯尔[67] 提出了基于知识发现和决策规则的盐碱地 GIS 和遥感分类的方法，把盐碱地分类的地学专家思想和区域专家的思想应用到 GIS 数据挖掘中，并把从 GIS 数据库中发现的知识，按一定的规则应用到华北平原地区的盐碱地分类的决策中，能够简化数据运算过程，减少或避免分类过程中人为误差的产生。陈春香[68] 应用机器学习中的数据驱动发现学习方法处理广东云浮-阳春地区的地球化学数据的实践证明，可以挖掘出隐含在数据间的各参数间的相互关系及参数组合规律，加强人们对数据的理解，为地球化学找矿提供更合理的决策信息。周成虎和张健挺[51] 从信息熵的基本概念出发，认为地学空间数据子集划分产生的互信息或熵减源于子集划分，使得各个子集的不确定性或模糊性

降低, 并且子集之间的差异性增大, 因此具有最大熵减的子集划分方案代表一定的地学模式和地学规律。并以此为基础分别探讨了地学数据属性要素的子集划分产生多维属性关联规则, 以及通过空间和时间的子集分割来进行聚类的方法。Ester 等[69]以空间的点为基本单位, 研究了多空间物体的相邻关系的处理技术, 集成了空间数据挖掘算法和空间数据库管理系统, 同时利用相邻图形和路径以及小型的初始数据库操作挖掘空间模式, 使用相邻索引来提高初始数据库的处理效率。Mouzon、Dubois 和 Prade[70]在空间可能因果关系的属性异常诊断索引中, 使用一致和诱导的算法挖掘了属性不确定性对异常诊断影响的知识。

2.7　模糊集

模糊集(fuzzy sets)用隶属函数确定的隶属度描述不精确的属性数据, 重在处理不精确的概率[71, 72]。模糊性是客观的存在, 系统的复杂性愈高, 对它的精确化能力就愈低, 模糊性愈强。在空间数据挖掘中, 模糊集可用作模糊评判、模糊决策、模糊模式识别、模糊聚类分析、合成证据和计算置信度等。模糊集在 GIS 中把类型、空间实体分别视为模糊集合、集合元素, 空间实体对备选类别论域的连续隶属度区间为[0, 1]。每个空间实体与一组元素的隶属度有关, 元素隶属度用于表示实体属于某类型的程度, 它越接近于 1, 实体就越属于该类型。具有类型混合、居间或渐变不确定性的实体可用元素隶属度描述, 如一块含有土壤和植被的土地, 可以由两个元素隶属度表示。传统的集合具有精确定义的界线, 为 0、1 二值逻辑。给定一个元素, 要么完全属于集合, 要么完全不属于。因反映空间非匀质分布的地理属性不确定性的概率是可变的, 类别变量的不确定性主要源自定性数据所固有的主观臆断性、易混淆性和模糊性, 故没有明确定义的界线的模糊集合论, 较传统集合论更适于研究非匀质分布和模糊类别。对于遥感图像的计算机分类处理, 模糊类别域的生成可藉所使用的分类器不同而输出不同的中间结果, 如统计分类器中有某像素隶属于各备选类别的似然值及神经元网络分类器中的类别激活水平值[73]。

在应用模糊集研究基于属性不确定性的空间数据挖掘的过程中, Burrough[74]讨论了不确定性数据的模糊布尔逻辑模型; Canters[75]评价了从模糊土地覆盖分类中估计面积的不确定性规律; Vazirgiannis 和 Halkidi[76]利用模糊逻辑处理数据挖掘的空间不确定性; 全斌和马智民[77]借助模糊关系数据模型发现了土地适宜性的评价知识; 王新洲和王树良[62, 78]提出了融模糊综合评判和模糊聚类分析为一体的模糊综合法, 基于绝对均值距离的模糊聚类分析, 分别用于挖掘土地的地价和级别属性不确定性; 模糊隶属度知识也用于表达遥感影像中的不确定相邻边界的像素类别[79]。

模糊隶属度一旦确定, 模糊集合的后续数值计算实际上已经把不确定性抛开, 并没有继续向前传送至结果, 而且模糊集合主要处理具有模糊性的属性不确定性, 对于同时含有模糊性和随机性的不确定性空间数据挖掘, 它只能丢弃随机性, 这是不合适的。

2.8　云理论

云理论(cloud theory)是一个分析不确定信息的新理论, 由云模型、不确定性推理和云变换三部分构成[80]。云理论把定性分析和定量计算结合起来, 可以用于处理 GIS 中融随机性和模糊性为一体的属性不确定性。云在空间由系列云滴组成, 远观像云, 近视无边。云具有期望值、熵和超熵三个数字特征。期望值是概念在论域中的中心值, 完全隶属于该定性概念; 熵是定性概念模糊度的度量, 其值越大, 概念所接受的数值范围越大, 概念越模糊; 超熵反映云滴的离散程度, 其值越大, 隶属的随机离散度越大。云理论构成定性和定量相互间的映射, 处理 GIS 中容模糊性(定性概念的亦此亦彼性)和随机性(隶属度的随机性)为一体的属性不确定性, 解决了作为模糊集理论基石的隶属函数的固有缺陷[61, 80]。云理论已经用于空间关联规则的挖掘、空间数据库的不确定性查询[61]。

2.9　粗集

粗集(rough sets)由上近似集和下近似集组成, 是一种处理不精确、不确定和不完备信息的智能数据决策分析工具[81, 82], 较适于基于属性不确定性的空间数据挖掘。粗集从集合论的观点出发, 在给定论域中以知识足够与否作为实体分类的标准, 并给出划分类型的精度。上近似集中的实体具有足够必要的信息和知识, 确定属于该类别; 论域全集以内且下近似集以外的实体没有必要的信息和知识, 确定不属于该类别; 上近似集和下近似集的差集为类别的不确定边界, 其中的实体没有足够必要的信息和知识, 无法确切地判断是否属于该类别, 为类别的边界。若两个实体有完全相同的信息, 则它们为等价关系, 不可区分。根据利用统计信息与否, 现存的粗集模型及其延伸可以分为代数型和概率型两大类[83]。粗集的

基本单位为等价类，类似于栅格数据的栅格、矢量数据的点或影像的像素。等价类划分越细，粗集描述实体越精确，但存储空间和计算时间也越大。基于粗集的决策规则推理具有演绎、归纳和常识等推理的原理，也有其自身的特点。决策规则是演绎推理规则和归纳方法的扩充，不同点在于决策规则强调优化，而归纳则不必关心它的优化形式。粗集的决策规则从条件出发作出恰当的或近似的决策，常识推理是从区域专家共享的知识开始推导出区域中有趣的、公认的知识。粗集不排斥不确定性，力求按照实体的原型来研究实体，为基于不确定性的空间数据挖掘提供了一个新的理论基础。

在空间分析时，粗集的数学基础是近似域，不同于模糊集的模糊隶属度、证据理论的证据函数、云理论的隶属度概率空间分布等。模糊集重在模糊性，基础为模糊隶属度；云理论兼容模糊性和随机性，基础为云变换；粗集重在不完备性，基础为上、下近似集。在自变量集 x 和因变量集 $\mu(x)$ 之间，模糊数学是一一对应关系，即对一个特定的 x_k，只存在惟一的隶属度 $\mu(x_k)$，且 $\mu(x_k)\in[0,1]$；粗集是一对区域关系，即对一个特定的 x_k，存在隶属区域 $[\{\min\text{-}\mu(x_k)\},\{\max\text{-}\mu(x_k)\}]$；云理论则是一对多云滴关系，云滴根据隶属度在空间随机离散分布，聚集到一定程度成为一朵云。

粗集自从被 Pawlak 提出后，已经突破原来的医疗诊断领域，被广泛应用在机器学习、人工智能、模式识别、近似推理、知识发现等领域。在此过程中，粗集日臻完善，已经从初始的偏重定性分析(如上、下近似集的描述、最小决策集的生成)发展到现在的定性定量并重(如粗概率、粗函数和粗微积分的表示与计算)，并且与模糊集、概率论和证据理论等互相交叉，形成粗模糊集、粗概率集和粗证据理论等。

在空间数据挖掘中，利用粗集可以分析空间数据库中的属性重要性、属性不确定性、属性表一致性和属性可靠性，研究属性可靠性对决策的影响，简化数据、属性表和属性依赖，发现数据相关性，评估数据的绝对不确定性和相对不确定性，由数据产生决策算法，发现数据中的范式及因果关系，生成最小决策和分类算法等，指导不确定影像分类、模糊边界划分等。如全国农业原始数据经过统计归纳得到普遍化的数据后，粗集可对其再次简化，生成最小决策算法和多种知识[61]。粗集已被用于描述属性 ROSE 不确定模型，分辨不精确的空间影像和面向目标的软件评估[84]，实现空间数据清理(data cleaning)中的数据转换[85]，用于基于属性不确定性的银行粗选址，从数据库中发现不确定属性的知识[61]，集成多源不确定的属性数据，实现定性和定量语言值的粗转化[86]。结合模糊隶属函数的遥感影像粗分类、粗邻域和粗属性精度[87]等，卞学海[88]基于信息系统，在粗集环境下，提出了一种适应非一致性数据的自增长必然规则学习算法。刘清等[89]通过粗集软计算使决策表中的属性简化和属性值区间化，从中挖掘出的数据隐含格式，删去了冗余规则，具有广泛的表达能力和代表性，并保持了决策表的原有用途和原有性能。然后分别用基于统计或专家经验方法计算带可信度因子的产生式规则和基于 Rough 集方法计算带 Rough 算子的决策规则两种不同方法开发同一个系统，后者比前者更加理论化和实用化。专著《粗集和含混数据的数据挖掘分析》[90]、《知识发现中的粗集》(1，2)[91,92]、《粗-模糊的交叉：决策的新趋势》[93]、《粗集方法和应用：信息系统中的知识发现新进展》[94]等系统地总结了粗集在数据挖掘中的理论和技术。

基于粗集的数据挖掘系统有 GROBIAN、RSDM、LERS、TRANCE、ProbRough、ROSETTA、RSL、RoughFamily、TAS、RoughFuzzyLab、PRIMEROSE、KDD－R 等[92]。

此外，还可以在粗集的基础上，发展专门针对空间信息学的地学粗空间理论。利用粗集理论、模糊数学和插值函数等技术，基于属性不确定性，在空间数据库或空间数据仓库中，可以挖掘和发现用于影像分类和分析、地价评估和空间表达[95]、城乡结合部用地分析和规划的知识。

2.10 神经网络

神经网络(neural network)是由大量神经元通过极其丰富和完善的连接而构成的自适应非线性动态系统，并具有分布存储、联想记忆、大规模并行处理、自学习、自组织、自适应等功能[96]。神经网络由输入层、中间层和输出层组成。大量神经元集体通过训练来学习待分析数据中的模式，形成描述复杂非线性系统的非线性函数，适于从环境信息复杂、背景知识模糊、推理规则不明确的非线性空间系统中挖掘分类知识。神经网络对计算机科学、人工智能、认知科学以及信息技术等都产生了重要而深远的影响，在空间数据挖掘中可用来进行分类、聚类、特征挖掘等操作。以 MP 和 Hebb 学习规则为基础，存在的神经网络可分为三类：用于预测、模式识别等的前馈式网络，如感知机(perceptron)、反向传播模型、函数型网络和模

糊神经网络等; 用于联想记忆和优化计算的反馈式网络, 如 Hopfield 的离散模型和连续模型等; 用于聚类的自组织网络, 如 ART 模型和 Koholen 模型等。Lee[97] 在空间统计学中用模糊神经网络估计了处理空间分布异常的规则。此外, 神经网络与遗传算法结合, 也能优化网络连接强度和网络参数。

神经网络具有鲜明的"具体问题具体分析"特点, 其收敛性、稳定性、局部最小值以及参数调整等问题尚待更深入的研究, 尤其对于输入变量多、系统复杂且非线性程度大等情况。

2.11　遗传算法

遗传算法(genetic algorithms)是模拟生物进化过程, 利用复制(选择)、交叉(重组)和变异(突变) 三个基本算子优化求解的技术[98]。在空间数据挖掘中, 把数据挖掘任务表达为一种搜索问题, 利用遗传算法的空间搜索能力, 经过若干代的遗传, 就能求得满足适应值的最优解规则。当实施遗传算法时, 首先要对求解的问题进行编码, 产生初始群体, 然后计算个体的适应度, 再进行染色体的复制、交换、突变等操作, 产生新的个体。重复以上操作直至求得最佳个体。

Jiang 等[99] 研究了解译染色体空间结构的计算工具, 以沟通代间信息。陈栋和徐洁磐[100] 基于遗传算法, 用信息论的思想开发出了知识挖掘系统 Knight。向国全[101] 采用逐步添加训练数据和隐节点避开局部极小点, 在数据挖掘中改进了前向人工神经网络算法。骆剑承、周成虎和马江洪[102] 研究了遥感影像特征发现的稳健统计模型, 认为高斯混合密度降解模型(GMDD)是一种基于稳健统计理论的层次结构的聚类模型。他们首先假设特征空间由一组混合的高斯(Gaussian)分布组成, 然后通过一定的优化算法来获得特征空间中与预先假设相符合的特征分布, 并逐步分离, 直到特征空间全部降解为一组混合特征模式的分布集。他们在 GMDD 模型基础上, 对空间数据中的特征进行分层提取, 提出用遗传算法进行 GMDD 的空间搜索的优化模型, 并从遥感影像数据中进行了特征发现的实例分析。

2.12　可视化

可视化(visualization)通过研制计算机工具、技术和系统, 把实验或数值计算获得的大量空间抽像数据(如信息模式、数据的关联或趋势等)转换为人的视觉可以直接感受的具体计算机图形图像, 以供数据挖掘和分析。空间数据挖掘中的数据立方法、多维数据库或 OLAP 也是可视化技术的一种。地理可视化系统中的不同物理位置及地理表示都与数据仓库中的数据相关, 根据地理环境比较相同产品在不同地域的差异, 或相同地域不同产品的差异, 可分析数据仓库中数据的关系。SDMKD 涉及复杂的数学方法和信息技术, 可视化是空间数据的视觉表达与分析, 借助图形、图像、动画等可视化手段对于形象地指导操作、定位重要的数据、引导挖掘、表达结果和评价模式的质量等具有现实意义。可视化拓宽了传统的图表功能, 使用户对数据的剖析更清楚, 有助于减少建模的复杂性, 决策者则可通过可视化技术交互分析数据关系。

SDMKD 可视化分为二维(x, y)、三维(x, y, z)和四维(x, y, z, t), 如果分别对它们按时间序列实时处理, 就可以形成较全面地反映数据挖掘过程和知识的动画。在空间数据挖掘中, 定性和定量数据的相互转换内容较多, 也较为抽像, 较适合把可视化作为研究工具。今后, 建立在可视化基础之上的 DMKD 可视化理论和技术, 将对空间信息可视表达、分析的研究与实践产生更大的影响。Kriegel 等[103] 利用可调的多参数函数分段逼近空间物体表面, 然后以此为基础挖掘空间分布知识, 并用误差椭圆评估可视化的知识。Ravantti 和 Bamford[104] 用三维可视化的空间数据挖掘技术分析了用于表示高分子结构的密度图, 兼顾了感兴趣的确定部分和可能的扩展部分。Ankerst 等[105] 分析了空间目标的形状属性, 利用 3D 形状的直方图表示空间数据库中的相似搜寻和分类。Maceachren 等[106] 集成了地理可视化和空间数据挖掘, 从结构化的多元时空数据集中构筑知识。

2.13　决策树

决策树(decision tree)根据不同的特征, 以树型结构表示分类或决策集合, 产生规则和发现规律[107]。在空间数据挖掘中, 首先利用训练空间实体集生成测试函数; 其次根据不同取值建立树的分支, 在每个分支子集中重复建立下层结点和分支, 形成决策树; 然后对决策树进行剪枝处理, 把决策树转化为据以对新实体进行分类的规则。ID3(interactive dichotomizer 3)方法根据信息论原理建立决策树或者决策规则树, 它计算数据库中各字段的信息量, 寻找数据库中具有最大信息量的字段, 建立决策树的一个结点。再根据字段的不同取值建立树的分支, 在每个分支子集中重复建树的下层结点和分支, 叶结点为正例或反例[63]。顾及决策树邻近对象的非空间聚合值, 基于分类对象的非空间属性、描述被分类对象和邻

近特征的空间关系的属性、谓词和函数，Koperski 等[19]提出了空间数据的两步决策分类法。在查找样本对象的粗略描述后，利用机器学习的 Relief 算法提取空间谓词，合并空间谓词和非空间谓词为分类决策知识。Marsala 和 Bigolin[108]利用模糊决策树在面向目标的空间数据库中挖掘区域分类规则。POSS 系统[109]使用决策树方法对天空图像中的星体对象进行分类，并通过分辨率、背景等级或平均强度等属性参数对图像进行规范化，以提高分类的准确性。著名的 C4.5 系统也是基于决策树的。

2.14 空间在线数据挖掘(SOLAM)

空间在线数据挖掘(spatial online analytical mining，SOLAM)建立在多维视图基础之上，是基于网络的验证型空间数据挖掘和分析工具。它强调执行效率和对用户命令的及时响应，直接数据源一般是空间数据仓库。网络是巨大的分布式并行信息空间和极具价值的信息源，但因网络所固有的开放性、动态性与异构性，又使得用户很难准确、快捷地从网络上获取所需信息。空间在线数据挖掘的目的就在于解决如何利用分散的异构环境数据源，及时得到准确的信息和知识。它突破了局部限制，发现的知识也更有普遍意义。

空间在线数据挖掘通过数据分析与报表模块的查询和分析工具(OLAP、决策分析、数据挖掘)完成对信息和知识的提取，以满足决策的需要。它建立在客户/服务器的结构之上，由用户驱动，支持多维数据分析，在用户的指导下验证设定的假设。空间在线数据挖掘的传输层使用了刷新与复制技术，数据传输、传送网络和中间件等构件，在硬件/软件平台间架起了必要的通信桥。其中刷新与复制技术包括传播和复制系统、数据库网关内定义的复制工具、数据仓库指定的产品。数据传输和传送网络包括网络协议、网络管理框架、网络操作系统、网络类型等。客户/服务器代理和中间件包括数据库网关、面向消息的中间件、对象请求代理等。这里，空间数据仓库居于核心地位，是网络空间数据挖掘的基础。

基于网络数据的空间在线数据挖掘可以利用搜索引擎[110]，向量空间模型和改进 Robot 技术[111]，面向多站点、多种数据库、多类数据源的分布式数据挖掘[112]，虚拟数据库技术[113]等。Zhou、Truffet 和 Han[114]研究了用于空间在线数据挖掘的多边形合并方法，认为基于占有的算法比基于邻接的算法更节省计算量。赵霈生、杨崇俊和刘冬林[115]基于网络环境，根据地理信息系统的应用整合和数据整合，提出了在分布式计算平台上以空间信息显示、服务、获取、存储为基础的层次化 GIS 软件框架，以便提供数据分析、共享、知识发现等不同水平的空间信息服务。为了支持数据挖掘时文档层次结构间的超链链接和媒体引用，卢坚等[116]按照 WWW 文档协同写作系统的 HTML 文档的层次式结构包装技术的要求，实现了扩展 HTML 文档的数据挖掘可视化编辑和浏览导航。邹涛、黄源和张福炎[117]则提出了基于 Internet 的文本信息挖掘算法。Web 热点的日志数据正以每天数十兆的速度增长，基于网络服务器日志数据的 SDMKD 也受到重视。周斌、吴泉源和高洪奎[118]基于 E-OEM 模型，提出了考虑服务器的应用逻辑设计、页面拓扑结构及用户的浏览路径等多个数据源的用户访问模式的挖掘算法。分形系统利用混沌理论来指明模式，然后用分形将多维数据库提供的分析信息存储于数据仓库，能够为基于大型空间数据仓库的空间数据挖掘提供 OLAP 的并行分布式响应。美国 Business Objects 公司的 Business Objects (BO)就是采用 Data Warehouse＋OLAP＋Ddata Mining 方案推出的第一个集多数据源查询、任意报表生成和 OLAP 及数据挖掘技术为一体的决策支持工具软件包。

此外，研究空间数据挖掘的还有发现状态空间理论[3]、基于灰色分析的灰色系统[119]、基于信息无序互动的混沌理论[120]，基于信化概念的未确知数学[121]等。

当然，上述理论和方法不是孤立的，为了在空间数据挖掘和知识发现中得到数量更多、精度更高的可靠结果，常常要综合应用它们。例如，使用空间统计学对数据进行分析后，再用粗集理论泛化初步的结果，然后由云理论实现知识的概括和定性定量的转化。因 SDMKD 需要考虑的因素很多，故应根据特定的需求选择数据挖掘的理论、方法和工具[122]。

3 空间数据挖掘和知识发现的理论和方法展望

SDMKD 具有广泛的应用前景和潜在的综合效益，随着空间数据量的增加及软硬件技术的发展，其应用正日益渗透到人们认识和改造空间世界的各个学科，如地理信息系统、信息融合、遥感、图像数据库、医疗图像处理、导航、机器人等使用空间数据的领域。SDMKD 发现的知识将会促进

这些学科的自动化和智能化。

但是, SDMKD 毕竟是空间信息科学的新兴领域, 目前只是取得了一定的初步成果, 仍有大量的理论与方法需要深入研究, 其中, 主要包括多源空间数据的清理、基于空间不确定性(位置、属性、时间等)的数据挖掘、递增式数据挖掘、栅格矢量一体化数据挖掘、多分辨率及多层次数据挖掘、并行数据挖掘、新算法和高效率算法的研究、空间数据挖掘查询语言、遥感图像数据库的数据挖掘、多媒体空间数据库的知识发现、网络空间数据的挖掘等方向。开发实现 SDMKD 理论和方法的计算机软件系统时, 还要研究多源空间数据的集成、多算法的集成、存储空间和计算效率的降低、人机交互技术、可视化技术、SDMKD 系统与地理信息系统、空间数据仓库、空间决策支持系统和遥感解译专家系统的集成等问题。

此外, SDMKD 除了发展和完善自己的理论和方法, 也要充分借鉴和汲取数据挖掘和知识发现、数据库、机器学习、人工智能、数理统计、可视化、地理信息系统、遥感、图形图像学、医疗、分子生物学等学科领域的成熟的理论和方法。

4　结　论

1) 空间数据挖掘和知识发现目的在于从数据库中挖掘事先未知且潜在有用的知识。

2) 基于经典的确定集合理论, 概率论和空间统计学研究含随机性的空间数据挖掘。证据理论是概率论的一个扩展, 规则归纳、聚类分析和空间分析是空间统计学的延伸。规则归纳在一定的知识背景下, 对空间数据进行概括和综合, 得到以概念树形式给出的高层次的模式或特征。与规则归纳不同的是, 聚类算法无需背景知识, 能根据实体的特征直接从空间数据库中发现有意义的空间聚类结构。空间分析包括拓扑结构分析、叠置分析、图像分析、模式识别、空间缓冲区和距离分析等。

3) SDMKD 的扩展集合论方法包括模糊集、粗集和云理论。目前用于表示空间数据的方法多数是基于经典的集合理论, 每一个空间实体都与单一的属性说明有关, 属性之间被表示为清晰的边界, 这和复杂多变的现实世界并非一致。因此, 经典集合理论也被扩展, 用于研究不能精确描述的空间实体, 如云理论可以在空间数据挖掘中实现定量和定性的相互转换。

4) 神经网络和遗传算法是空间数据挖掘和知识发现的仿生学方法。神经网络是由大量神经元通过极其丰富和完善的连接而构成的自适应非线性动态系统。遗传算法是模拟生物进化过程, 在空间数据挖掘中利用复制、交叉和变异三个基本算子优化求解的技术。

5) SDMKD 涉及海量数据、复杂的数学方法和信息技术, 可视化是空间数据的视觉表达与分析, 能够把实验或数值计算获得的大量空间抽象数据转换为人的视觉可以直接感受的具体计算机图形图像。决策树根据不同的特征, 以树型结构表示分类或决策集合, 产生规则和发现规律。空间在线数据挖掘建立在多维视图基础之上, 是基于网络的验证型空间数据挖掘和分析工具, 它强调执行效率和对用户命令的及时响应, 直接数据源一般是空间数据仓库。

参　考　文　献

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

2　Li D R, Cheng T. KDG—Knowledge Discovery from GIS. Proceedings of the Canadian Conference on GIS, Ottawa, 1994

3　李德毅. 发现状态空间理论. 小型微型计算机系统, 1994, 15(11): 1~6

4　李德仁, 关泽群. 空间信息系统的集成与实现. 武汉: 武汉测绘科技大学出版社, 2000

5　Lenarcik A, Piasta Z. Rough Sets in Knowledge Discovery 2: Applications, Case Studies and Software Systems. In: Polkowski L, Skowron A, eds. Studies in Fuzziness and Soft Computing. Heidelberg: Physica-Verlag, 1998. 569~571

6　Frasconi P, Gori M, Soda G. Data Categorization Using Decision Trellises. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(5): 697~712

7　Brus D J. Using Nonprobability Samples in Design-based Estimation of Spatial Means of Soil Properties. Proceedings Accuracy 2000, Amsterdam, 2000

8　Sester M. Knowledge Acquisition for the Automatic Interpretation of Spatial Data. International Journal of Geographical Information Science, 2000, 14(1): 1~24

9　Shafer G. A Mathematical Theory of Evidence. Princeton: Princeton University Press, 1976

10　Yang J B, Madan G, Singh. An Evidential Reasoning Approach for Multiple-Attribute Decision Making with Uncertainty. IEEE Transactions on System, Man, and Cybernetics, 1994, 24(1): 1~18

11　Lipski W J. On Databases with Incomplete Information. Journal of ACM, 1981, 28: 41~70

12　Chrisman N C. Exploring Geographic Information Sys-

tems. New York: John Wiley & Sons, 1997
13 Cressie N. Statistics for Spatial Data. New York: John Wiley & Sons, 1991
14 Wang J, Yang J, Muntz R. An Approach to Active Spatial Data Mining Based on Statistical Information. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(5): 715~728
15 Wang J, Yang J, Muntz R. STING+: An Approach to Active Spatial Data Mining. Proceedings of the 15th International Conference on Data Engineering, 1999
16 郭达志, 胡召玲, 陈云浩. GIS 中空间对像的不确定性研究. 中国矿业大学学报(自然科学版), 2000, 29(1): 20~24
17 张景雄, 杜道生. 位置不确定性与属性不确定性的场模型. 测绘学报, 1999, 28(3): 244~249
18 冯建生. KDD 及其应用. 宝钢技术, 1999(3): 27~31
19 Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer M J, Herring J R, Portland M E, eds. Proceedings of the 4th International Symposium on Spatial Databases (SSD'95). Berlin: Springer-Verlag, 1995
20 程继华, 施鹏飞. 多层次关联规则的有效挖掘算法. 软件学报, 1998, 12(9): 937~941
21 许龙飞, 杨晓昀. KDD 中广义关联规则发现技术研究. 计算机工程与应用, 1998(9): 32~35
22 Han E H S, Karypis G, Kumar V. Scalable Parallel Data Mining for Association Rules. IEEE Transaction on Knowledge and Data Engineering, 2000, 12(3): 337~352
23 Eklund P W, Kirkby S D, Salim A. Data Mining and Soil Salinity Analysis. International Journal of Geographical Information Science, 1998, 12(3): 247~268
24 Aspinall R, Pearson D. Integrated Geographical Assessment of Environmental Condition in Water Catchments: Linking Landscape Ecology, Environmental Modeling and GIS. Journal of Environmental Management, 2000, 59: 299~319
25 涂星原. 基于数值属性的关联规则的挖掘. 郑州工业大学学报, 1998, 19(3): 72~75
26 Clementini E, Felice P D, Koperski K. Mining Multiple-level Spatial Association Rules for Objects with a Broad Boundary. Data and Knowledge Engineering, 2000, 34: 251~270
27 左万利. 含有类别属性数据库中联系性规则的挖掘. 吉林大学自然科学学报, 1999(1): 33~37
28 丁祥武. 挖掘时态关联规则. 武汉交通科技大学学报, 1999, 23(4): 365~367
29 丁祥武. 序列模式的一种模型及其挖掘. 中南民族学院学报(自然科学版), 1999, 18(2): 44~48
30 丁祥武. 挖掘关联规则的一种预处理: 合并交易. 中南民族学院学报(自然科学版), 1999, 18(3): 21~25
31 肖 利. 挖掘序列型式的模型研究. 计算机科学, 1998(专刊): 135~136
32 程继华, 施鹏飞, 魏暑生. 基于概念的关联规则的挖掘. 郑州大学学报(自然科学版), 1998, 30(20): 27~30
33 肖 利, 金远平, 徐宏炳, 等. 一个新的挖掘广义关联规则算法. 东南大学学报, 1997, 27(6): 76~81
34 Han J W, Cai Y, Cercone N. Data Driven Discovery of Quantitative Rules in Relational Databases. IEEE Transaction on Knowledge and Data Engineering, 1993, 5(1): 29~40
35 肖 利, 王能斌, 徐宏炳, 等. 挖掘转移规则: 一种新的数据挖掘技术. 计算机研究与发展, 1998, 35(10): 902~906
36 朱 明, 王俊普, 蔡庆生. 一种最优特征集的选择算法. 计算机研究与发展, 1998, 35(9): 803~805
37 陈 彬, 洪家荣, 王亚东. 最优特征子集选择问题. 计算机学报, 1997, 20(2): 133~138
38 Kriegel H P. 3D Similarity Search by Shape Approximation. In: Scholl M, Voisard A, eds. Proceedings of the 5th International Symposium on Spatial Databases (SSD'97). Berlin: Springer-Verlag, 1997
39 欧阳为民, 蔡庆生. 在大型数据库中多层序贯模式的发现. 计算机研究与发展, 1998, 35(10): 916~920
40 欧阳为民, 蔡庆生. 一种在数据库中发现偏离模式的线性算法. 计算机研究与发展, 1998, 35(10): 911~915
41 周 斌, 吴泉源. 序列模式挖掘的一种渐进算法. 计算机学报, 1999, 22(8): 882~887
42 杨学兵, 刘胜军, 蔡庆生. 一种实时过程控制中的数据挖掘算法研究. 计算机应用, 1999, 19(9): 8~10
43 Levene M, Vincent M W. Justification for Inclusion Dependency Normal Form. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(2): 281~291
44 覃振兴, 袁曾任. 在 KDD 和 Data Mining 中我们的部分工作和看法. 信息与控制, 1999, 28(4): 255~261
45 石 冰, 郑燕峰. 信息检索中的数据挖掘技术. 情报学报, 1999, 18(增刊): 103~106
46 Kaufman L, Rousseew P J. Finding Groups in Data: an Introduction to Cluster Analysis. New York: John Wiley & Sons, 1990
47 Murray A T, Estivill-Castro V. Clustering Discovery Techniques for Exploratory Spatial Data Analysis. International Journal of Geographical Information Science, 1998, 12(5): 431~443
48 Ester M. A Density—based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. The 2nd International Conference on Knowledge Discovery and Data Mining, Portland, 1996
49 Ng R T, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining. The 20th Very

Large Databases Conference, Santiago, Chile, 1994
50 Ester M. A Database Interface for Clustering in Large Spatial Databases. The 1st International Conference on Knowledge Discovery and Data Mining, Montreal, 1995
51 周成虎, 张健挺. 基于信息熵的地学空间数据挖掘模型. 中国图像图形学报, 1999, 4[A](11): 943~951
52 Pitt L, Reinke R E. Criteria for Polynomial Time (conceptual) Clustering. Machine Learning, 1998, 2(4): 371~396
53 Han J, Fu Y. Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. AAAI' 94 Workshop on Knowledge Discovery in Databases (KDD' 94), Seattle, WA, 1994
54 李世祥, 李涛涛. 一种数据库数值型字段概念化算法的介绍及讨论. 微型电脑应用, 1999, 15(11): 24~26
55 Ester M, Kriegel H P, Xu X. Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification. In: Egenhofer M J, Herring J R, Portland M E, eds. Proceedings of the 4th International Symposium on Spatial Databases (SSD' 95). Berlin: Springer-Verlag, 1995
56 Knorr E M, Ng R T. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 884~897
57 Edwin. Finding Boundary Shape Matching Relationships in Spatial Data. In: Scholl M, Voisard, eds. Proceedings of the 5th International Symposium on Spatial Databases (SSD' 97). Berlin: Springer-Verlag, 1997
58 Lin X, Zhou X, Liu C. Efficiently Matching Proximity Relationships in Spatial Databases. In: Güting R H, Papadias D, Lochovshy F, eds. Proceedings of the 6th International Symposium on Spatial Databases (SSD' 99). Berlin: Springer-Verlag, 1999
59 Tung A K H, Hou J, Han J. Spatial Clustering in the Presence of Obstacles. IEEE Tramsactions on Data Engineering, 2001, 11: 359~369
60 Murray A T, Shyy T K. Integrating Attribute and Space Characteristics in Choropleth Display and Spatial Data Mining. International Journal of Geographical Information Science, 2000, 14(7): 649~667
61 邸凯昌. 空间数据发掘和知识发现的理论和方法: [博士论文]. 武汉: 武汉测绘科技大学, 1999
62 王新洲, 王树良. 土地评价中的模糊聚类分析. 武汉大学学报(自然科学版), 1999(模糊理论与工程专刊): 103~107
63 Ester M, Kriegel H P, Sander J. Spatial Data Mining: a Database Approach. In: Scholl M V, eds. Proceedings of the 5th International Symposium on Spatial Databases (SSD' 97). Berlin: Springer-Verlag, 1997
64 Reinartz T. Focusing Solutions for Data Ming: Analytical Studies and Experimental Results in Real World Domains. Berlin: Springer, 1999
65 王晋年. 以地物识别和分类为目标的高光谱数据挖掘. 中国图像图形学报, 1999, 4A(11): 957~964
66 马建文, 马超飞. 基于空间角度理论的卫星光学遥感数据认知与挖掘. 中国图像图形学报, 1999, 4[A](11): 918~923
67 布和敖斯尔. 基于知识发现和决策规则的盐碱地遥感分类方法研究. 中国图像图形学报, 1999, 4[A](11): 965~969
68 陈春香. 数据发现在地球化学数据处理中的应用. 桂林工学院学报, 1999, 19(3): 230~234
69 Ester M. Spatial Data Mining: Databases Primitives, Algorithms and Efficient DBMS Support. Data Mining and Knowledge Discovery, 2000(4): 193~216
70 Mouzon O D, Dubois D, Prade H. Using Consistency and Abduction Based Indices in Possibilistic Causal Diagnosis. IEEE, 2001, 729~734
71 Zadeh L A. Fuzzy Sets. Information and Control, 1965
72 Shi W Z. Modeling Positional and Thematic Uncertainties in Integratio of Remote Sensing and Geographic Information Systems. Enschede: ITC Publication, 1994
73 张景雄, 杜道生, 孙家抦. 用随机模拟方法建立矢量数据的误差模型. 武汉测绘科技大学学报, 2000, 25(1): 49~54
74 Burrough P A, Frank A U. Geographic Objects with Indeterminate Boundaries. Basingstoke: Taylor and Francis, 1996
75 Canters F. Evaluating the Uncertainty of Area Estimates Derived from Fuzzy Land-cover Classification. Photogrammetric Engineering & Remote Sensing, 1997, 63(4): 403~414
76 Vazirgiannis M, Halkidi M. Uncertainty Handling in the Data Mining Process with Fuzzy Logic. IEEE, 2000, 393~398
77 全　斌, 马智民. GIS 中空间数据属性不确定性研究. 西安工程学院学报, 1998, 20(1): 67~71
78 王新洲, 王树良. 模糊综合法在土地定级中的应用. 武汉测绘科技大学学报, 1997, 22(1): 42~46
79 史文中. 空间误差处理的理论和方法. 北京: 科学出版社, 1998
80 李德毅, 史雪梅, 孟海军. 隶属云和隶属云发生器. 计算机研究与发展, 1995, 42(8): 32~41
81 Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences, 1982, 11(5): 341~356
82 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. London: Kluwer Academic Publishers, 1991
83 Yao Y Y, Wong S K M, Lin T Y. A Review of Rough Set Models. In: Lin Y, Cercone N K, eds. Rough Sets and Data Mining Analysis for Imprecise Data. London: Kluwer Academic Publishers, 1997

84 Gunther R. Rough Set-based Analysis in Goal-oriented Software Measurement. Proceedings of METRICS' 96, 1996

85 Stepaniuk J, Maj M. Data Transformation and Rough Sets. In: Zytkow J M, Quafafou M N, eds. Proceedings of the 2th European Symposium PAKDD' 98. Berlin: Springer, 1998

86 Wang S L, Wang X Z. Data Mining Applied in Land Use Control in City-Country Combinative Area. International Archives of Photogrammetry and Remote Sensing, 2000, 33(B7): 1 677 ~ 1 683

87 Ahlqvist O, Keukelaar J, Oukbir K. Rough Classification and Accuracy Assessment. International Journal of Geographical Information Science, 2000, 14(5): 475 ~ 496

88 卞学海. 非一致性数据的必然规则学习. 华东船舶工业学院学报, 1998, 12(1): 25 ~ 30

89 刘 清. 带 Rough 算子的决策规则及数据挖掘中的软计算. 计算机研究与应用, 1998, 36(7): 800 ~ 804

90 Lin Y, Cercone N. Rough Sets and Data Mining Analysis for Imprecise Data. London: Kluwer Academic Publishers, 1997

91 Polkowski L, Skowron A. Rough Sets in Knowledge Discovery 1: Methodologies and Applications. Heidelberg: Physica-Verlag, 1998

92 Polkowski L, Skowron A. Rough Sets in Knowledge Discovery 2: Applications, Case Studies and Software Systems. Heidelberg: Physica-Verlag, 1998

93 Pal S, Skowron A. Rough-Fuzzy Hybridization: a New Trend in Decision-Making. Singapore: Springer-Verlag, 1999

94 Polkowski L, Tsumoto S, Lin T Y. Rough Sets Methods and Applications: New Developments in Knowledge Discovery and Information Systems. Heidelberg: Physica-Verlag, 1998

95 王树良, 孙春生, 严 春. 基准地价中的土地利用类型探讨. 中国土地科学, 1999, 13(1): 5 ~ 9

96 Müller B, Rrinhardt J. Neural Networks: an Introduction. Berlin: Springer-Verlag, 1997

97 Lee E S. Neuro-fuzzy Estimation in Spatial Statistics. Journal of Mathematical Analysis and Applications, 2000, 249: 221 ~ 231

98 Buckless B P, Petry F E. Genetic Algorithms. California: IEEE Computer Press, 1994

99 Jiang W. Bridging the Information Gap: Computational Tools for Intermediate Resolution Structure Interpretation. Journal of Molecular Biology, 2001, 308: 1 033 ~ 1 044

100 陈 栋, 徐洁磐. Knight: 一个通用知识挖掘工具. 计算机研究与发展, 1998, 35, (4): 338 ~ 343

101 向国全. 前向网络 BP 算法在数据挖掘中的运用. 河南大学学报(自然科学版), 1999, 29(3): 42 ~ 45

102 骆剑承, 周成虎, 马江洪. 遥感影像特征发现的稳健统计模型研究. 中国图像图形学报, 1999, 4[A](11): 952 ~ 956

103 Kriegel H P. 3D Similarity Search by Shape Approximation. In: Scholl M, Voisard A, eds. Proceedings of the 5th International Symposium on Spatial Databases (SSD' 97). Berlin: Springer-Verlag, 1997: 11 ~ 28

104 Ravantti J J, Bamford D H. A Data Mining Approach for Analyzing Density Maps Representing Macromolecular Structures. Journal of Structural Biology, 1999, 125: 216 ~ 222

105 Ankerst M. 3D Shape Histograms for Similarity Search and Classification in Spatial Databases. In: Güting R H, Papadias D, Lochovshy F. eds. Proceedings of the 6th International Symposium on Spatial Databases (SSD' 99). Berlin: Springer-Verlag, 1999: 207 ~ 225

106 Maceachren A M. Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographical Visualization with Knowledge Discovery in Database Methods. International Journal of Geographical Information Science, 1999, 13(4): 311 ~ 334

107 Quinlan J. Introduction of Decision Trees. Machine Learning, 1986, (5): 239 ~ 266

108 Marsala C, Bigolin N M. Spatial Data Mining with Fuzzy Decision Trees. In: Ebecken N F F, eds. Data Mining. Ashurst Lodge: WIT Press/Computational Mechanics Publications, 1998. 235 ~ 248

109 Fayyad U M. Advances in Knowledge Discovery and Data Mining. Menlopark CA: AAAI/MIT Press, 1996

110 翁惠玉. 网络搜索引擎的现状分析. 情报学报, 1999, 18(增刊): 100 ~ 102

111 邹 涛. WWW 上的信息挖掘技术及实现. 计算机研究与发展, 1999, 36(8): 1 019 ~ 1 024

112 何炎祥. 基于网络环境的分布式 KDD 及 Data Mining 研究. 小型微型计算机系统, 1999, 20(10): 744 ~ 746

113 陈 莉. 数据挖掘与虚拟数据库. 四川师范大学学报(自然科学版), 1999, 21(6): 657 ~ 661

114 Zhou X, Truffet D, Han J. Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining. In: Güting R H, Papadias D, Lochovshy F, eds. Proceedings of the 6th International Symposium on Spatial Databases (SSD' 99). Berlin: Springer-Verlag, 1999. 167 ~ 187

115 赵霈生, 杨崇俊, 刘冬林. 基于网络环境的地理信息系统整合与知识发现. 中国图像图形学报, 1999, 4[A](11): 940 ~ 945

116 卢 坚. WWW DOC 系统中 HTML 文档的可视化编辑与浏览技术的实现. 计算机辅助设计与图形学学报, 1999, 11(6): 559 ~ 562

117 邹 涛, 黄 源, 张福炎. 基于 WWW 的文本信息挖

掘. 情报学报, 1999, 18(4): 289～293
118 周　斌, 吴泉源, 高洪奎. 用户访问模式数据挖掘的模型与算法研究. 计算机研究与发展, 1999, 36(7): 870～875
119 邓聚龙. 灰色系统基本方法. 武汉: 华中工学院出版社, 1987
120 Awrejcewicz J. Bifurcation and Chaos in Simple Dynamical Systems. Singapore: World Scientific, 1989
121 刘开第. 未确知数学. 武汉: 华中理工大学出版社, 1997
122 郑宏珍, 柳明欣. 数据挖掘及其工具的选择. 计算机应用, 1999, 19(10): 109～110

作者简介: 李德仁, 教授, 博士生导师, 中国科学院院士, 中国工程院院士, 欧亚科学院院士。现主要从事遥感、全球定位系统、地理信息系统和多媒体网络通信及其集成研究。代表成果: 高精度摄影测量定位理论与方法; GPS 辅助空中三角测量; SPOT 卫星像片解析处理; 数学形态学及其在测量数据库中的应用; 面向对象的 GIS 理论与技术; 空间数据挖掘和知识发现的理论与方法; 影像理解及像片自动解译以及多媒体通信等。已发表论文 300 余篇, 出版专著 8 部。
E-mail: dli@wtusm.edu.cn

Theories and Technologies of Spatial Data Mining and Knowledge Discovery

LI Deren[1]　*WANG Shuliang*[1]　*LI Deyi*[2]　*WANG Xinzhou*[3]
(1 National Laboratory for Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079)
(2 China Institute of Electric System Engineering, 6 Wanshou Road, Beijing, China, 100039)
(3 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079)

Abstract: The good methods and technologies of spatial data mining and knowledge discovery (SDMKD) may get excellent knowledge. This paper presents an overview on SDMKD. First, the concept of SDMKD is discussed. Then, this paper describes the theories and technologies on SDMKD, such as probability theory, evidence theory, spatial statistics, rules induction, clustering analysis, spatial analysis, fuzzy sets, rough sets cloud theory, neural network, genetic algorithms, visualization, decision tree, spatial online analytical mining. Finally, how to study SDMKD is forecasted.

Key words: spatial data mining; knowledge discovery; theories and technologies

About the author: LI Deren, professor, Ph. D supervisor, member of the Chinese Academy of Sciences, member of the Chinese Academy of Engineering, member of the Euro-Asia International Academy of Sciences. He is concentrated on the research and education in spatial information science and technology represented by remote sensing (RS), global positioning system (GPS), geographic information system (GIS) and the intergration of multimedia network communications. He has made unique and original contribution to the areas of theories and methods for high precision photogrammetric positioning, GPS aerotriangulation, analysis and processing of SPOT imagery, mathematical morphology and its application in spatial databases, theories of spatial data mining and knowledge discovery, theories of object-oriented GIS, image understanding and automatic photointerpretation, multimedia communication and mobile mapping systems, etc. The research findings have promoted the progress of the technology directly and are being turned into products. His published papers are more than 300 and books 8.
E-mail: dli@wtusm.edu.cn