GIS 与空间数据挖掘集成在农业中的应用

王 娟[1],查良松[2]　(1. 盐城师范学院城市与资源环境学院,江苏盐城 224001;2. 安徽师范大学国土资源与旅游学院,安徽师范大学 GIS 重点实验室,安徽芜湖 241000)

摘要　随着数据挖掘的发展与日益成熟,空间数据挖掘成为研究的热点之一,可以广泛地应用于地理信息系统、遥感图像处理、计算机利用等领域。进行了 GIS 与空间数据挖掘集成技术的初探,探讨 DM 在 GIS 中应用的必要性与可能性,以解决 GIS 中面临的"数据丰富但知识贫乏"的瓶颈问题。介绍在 GIS 数据库中进行数据挖掘,可以发现的知识类型和具体应用,最后以马尔可夫预测法预测未来几年的农业收成状态的实例说明空间数据挖掘技术对 GIS 的贡献。

关键词　GIS;空间数据挖掘;集成;农业
中图分类号　S127　　**文献标识码**　A　　**文章编号**　0517 - 6611(2008)29 - 12974 - 02

Application of the Integration of GIS and SDM in Agriculture

WANG Juan et al　(College of Urban and Resources Environment, Yancheng Teachers University, Yancheng Jiangsu 224001)

Abstract　With the maturity of the data mining technology, spatial data mining became one of the research hot spots. It could widely applied in GIS, RS and computer application. The initially researches on the integration between GIS and the spatial data mining technology were conducted. the necessity and the possibility of DM in GIS application were discussed in order to solve the bottleneck problem of "rich data but deficient knowledge" in GIS. The data mining in the GIS database could discover knowledge type and concrete application. Finally, by using the Markov pre-measurement, the agricultural crop condition in several years was forecasted, which showed the contribution of spatial data mining technology to GIS.

Key words　Geography information system (GIS); Spatial Data Mining(SDM); Integration; Agriculture

数据挖掘是在计算机技术的普及、数据库和数据收集技术日益成熟的情况下发展起来的。至今,数据挖掘(DM, DataMining)没有统一的定义。根据 Jiawei Han 的定义,数据挖掘是指从大量的数据中提取出可信的、新颖的、有用的并能被人理解的模式的处理过程[1]。这一概念一经提出,引起了学者、软件开发商的极大兴趣,许多数据挖掘公司或部门进行了广泛深入的研究。到目前为止,已经形成了比较完整的数据挖掘理论和方法体系,并且出现了许多实用的数据挖掘工具,如加拿大 Simon Fraser 大学开发的 GeoMiner、美国 ESRI 公司开发的 Arcview GIS 中的 S-Plus 接口等,被应用于各行各业,产生了巨大的效益。DM 的研究主要集中在关系数据库和事务数据库[2-4],但随着研究的深入,人们发现,空间数据库隐藏着比一般关系数据库、事务数据库更丰富复杂的语义信息和知识。空间数据挖掘(SDM, Spatial DataMining)即空间知识发现(KDSD, Knowledge Discovery in Spatial Database)是指从空间数据库中发现知识的技术,具体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藏在数据库中的普遍的数据特征。GIS 是计算机技术、数据库技术、图像图形技术相结合的产物,是管理和研究空间数据的技术系统。GIS 是空间数据库发展的主体[5],包含了空间数据和非空间数据。因此,GIS 与空间数据挖掘技术集成具有广阔的应用前景。

1　GIS 与空间数据挖掘集成的必要性

Tobler 的第一地理规则描述了这样的一种空间依赖性:"所有的事物都是有联系的,一个地方发生的事件总是与它附近发生的事件有关联,并且相距近的事物之间的联系比相距远的事物之间的联系要紧些[7]。"如果能从这些数据中找出其规律或相互联系,就可以反过来推断客观世界的情况。而数据挖掘技术正为 GIS 中组织管理海量数据提供了新思路。GIS 的应用需要空间数据挖掘技术的帮助。因为 GIS 数据库中不仅包含了大量的属性数据,而且还包含了大量的空间数据。GIS 数据库中的数据挖掘可以分为对矢量空间结构对象的数据挖掘和对栅格空间结构对象的数据挖掘两大类[6]。空间数据带有空间拓扑结构和距离信息,通常用复杂的多维空间索引结构组织存放(如 R 树),并通过空间数据存取方法访问,也常常需要空间推理、几何计算和空间知识的表示技术等,其相互依赖性强。事物挖掘算法假定数据独立,所以只有把事物挖掘技术扩充到空间数据挖掘,才能更好地分析复杂的空间现象和空间对象。空间数据挖掘技术是 DM 技术的分支,是数据挖掘的主要研究方向之一,是 KDD(Knowledge Discovery in Database)技术在空间数据库方面应用的延伸。空间数据挖掘技术的应用一般可使 GIS 查询和分析技术提高到发现知识的新阶段,另一方面,从中发现的知识可构成知识库用于建立智能化的 GIS 系统,因为专家系统(ES, Expert System)中所需要的许多知识技术就隐藏在 GIS 数据库中。

地理信息系统技术与空间数据挖掘技术的集成不仅促进 GIS 的自身发展,也必将为数据挖掘的发展提供更广阔的前景。因此,GIS 与 SDM 技术集成研究是必要的,同时,也是可能的。

2　GIS 数据库中的知识发现及应用[7-8]

2.1　从 GIS 数据库中可以发现的知识类型

(1)几何信息知识。从 GIS 的图形数据库和 RS 影像中,可以很方便地得到某类目标的数量、大小、位置及结构等几何特征,通过归纳和演绎的方法就可以得出关于该类地物目标(如农田、森林、果园等)的规律性的几何信息知识。

(2)空间分布规律。指目标在地理空间的分布规律,分成在垂直向、水平向及垂直水平向的联合分布规律。垂直向即地物沿高程带的分布,如农作物沿高程带和坡度的分布规

基金项目　安徽省自然科学基金项目(050450103)。
作者简介　王娟(1981 -),女,江苏盐城人,硕士,助教,从事模型设计与 GIS 应用研究。
收稿日期　2008-07-23

律;水平向分布指地物在平面区域的分布规律,如不同区域农作物的差异;垂直向和水平向的联合分布即不同的区域中地物沿高程的分布规律。

(3)空间特征规则。即将 GIS 中的空间数据和属性数据对应起来,发现目标的几何和属性之间对应的关系。如北方以旱季作物为主,在南方则以水稻为主。

(4)空间聚类规则。指特征相近的空间目标聚类成上一级类的规则,可用于 GIS 的空间概括和综合。如精确农业中的作物产量图可聚类成高、中、低产区。

(5)空间区分规则。指两类或多类目标间几何的或属性的不同特征,即可以区分不同类目标的特征,如荞麦一般生长在北方,而甜菜一般生长在南方。

(6)空间关联规则。从 GIS 图形和属性数据库中,不难发现目标间的相连、相邻、共生及包含关系。如某种植农场 80% 的农田低产区靠近公路。

(7)空间演变规则。若 GIS 数据中存有同一地区多个时间数据的快照(snapshot),则可以发现空间演变规则,即哪些地区易变,哪些地区不易变,哪些目标易变等及怎么变,哪些目标固定不变。

(8)面向对象的知识。若 GIS 中采用了面向对象的数据类型,则可以很方便地提供超类 - 类 - 子类目标之间的知识继承、传播和集成。

(9)空间混沌模式。指空间数据库中空间数据、属性数据中存在介于确定关系和纯随机关系间的混沌关系,是一种无序中的有序关系。

(10)空间序贯模式。指空间数据库中满足用户指定最低支持的最长的空间数据时间序列或属性数据时间序列。

2.2　从 GIS 数据库中发现知识的应用　GIS 数据库中发现的知识,一方面可应用于 GIS 的智能化分析和智能系统的分析决策。SDM 获取的知识同现有 GIS 分析工具获取的信息相比更加概括、精练,并可以发现现有 GIS 分析工具无法获取的隐含的模式和规律从而广泛应用于公用事业、军事、交通领域、农业领域、土地管理、城市规划、流行病防治、环境监测与评价等诸多领域中[9-10]。另一方面,KDSD 可以应用于 RS 影像解译,提高解译的精度、速度和准确度。RS 影像解译的结果可以更新 GIS 数据库。例如,王雷等用知识发现方法挖掘 RS 影像的土地覆盖类型[10]。

3　GIS 与 SDM 集成应用于农业

3.1　GIS 支持下的空间数据挖掘技术的分类　空间数据挖掘技术按功能可分为描述、解释和预测[11]。描述性的模型将空间的分布特征化,如空间聚类。解释性的模型处理空间关系,如一个空间对象和影响其空间分布的因素之间的关系。预测型的模型用来根据给定的一些属性预测某些属性,如回归、马尔可夫等。

3.2　GIS 技术在农业中的应用　GIS 技术用于国家和地区范围内农业相关的评估已有多年。很多特定的农场系统利用 GIS 和一些相关技术来收集空间参考数据,进行空间分析和制定决策,作物状况和收成、土地能力、土壤侵蚀、土壤密度、地表和地下水污染、虫害袭击、杂草清除和气候变化影响的评估。例如,Corbett 和 Larter 生成的气候表面可以作为基因型 - 敏感性农作物模型的输入来评价特定作物品种承担的风险,他们利用 GIS 和 RS 技术在 SOYGRO 生理学的大豆生长模型中预测了美国卡罗来纳州 Orangebury 的大豆收成空间分布差异[9-10]。鉴于 GIS 的数据管理集成和演示能力,许多学者承认利用 GIS 管理农业信息达到了某些目的,如距离、面积计算、布尔叠加、缓冲和分类,但不能解决全部问题,比如忽略了大部分克里金相关地理统计,多元分析、趋势表面分析、模糊逻辑统计工具等。为更好地将 GIS 技术应用于农业,迫切地要求 GIS 与空间数据挖掘技术集成。

3.3　马尔可夫预测法挖掘农业收成状态　马尔可夫(Markov)预测法是 DM 的一种工具,就是一种预测事件发生概率的方法。它是基于 Markov 链,根据事件的目前状况预测其未来各个时刻(或时期)变化状况的一种预测方法[12]。具有无后效性的随机过程称为马尔可夫过程。时间和状态均为离散的马尔可夫过程称为马尔可夫链,简称马氏链。在应用马尔可夫模型进行预测时,转移概率矩阵 P 计算是关键,它的精度将直接影响该方法的准确性。举一个 DM 技术在农业中应用的简单实例,已知某地区 40 年的农业收成状态(记 E_1 为“丰收”状态、E_2 为“平收”状态、E_3 为“歉收”状态)变化,如表 1 所示。人们可以利用马尔可夫预测方法就可以获得未来几年的收成状况,数据挖掘的基本思想是将每年的收成状况作为系统数据单元,进行数据挖掘[13]。

表 1　某地区 1965 ~ 2004 年农业收成 状态转移情况

Table 1　State transition of agricultural yield from 1965 to 2004 in certain area

年份 Year	序号 No.	状态 State	年份 Year	序号 No.	状态 State	年份 Year	序号 No.	状态 State
1965	1	E_2	1979	15	E_3	1993	29	E_1
1966	2	E_1	1980	16	E_2	1994	30	E_2
1967	3	E_2	1981	17	E_1	1995	31	E_2
1968	4	E_3	1982	18	E_1	1996	32	E_3
1969	5	E_1	1983	19	E_2	1997	33	E_1
1970	6	E_2	1984	20	E_1	1998	34	E_3
1971	7	E_1	1985	21	E_2	1999	35	E_2
1972	8	E_2	1986	22	E_1	2000	36	E_1
1973	9	E_3	1987	23	E_1	2001	37	E_2
1974	10	E_1	1988	24	E_2	2002	38	E_3
1975	11	E_2	1989	25	E_2	2003	39	E_2
1976	12	E_1	1990	26	E_3	2004	40	E_1
1977	13	E_2	1991	27	E_3			
1978	14	E_1	1992	28	E_3			

由表 1 可知,该地区农业收成变化的状态转移概率矩阵为

$$P=\begin{bmatrix}0.133\,3 & 0.733\,4 & 0.133\,3\\ 0.562\,5 & 0.125\,0 & 0.312\,5\\ 0.444\,5 & 0.333\,3 & 0.222\,2\end{bmatrix}$$

根据马尔可夫预测法,将 2004 年的农业收成状态记为 $\pi(0)=[1,0,0]$,就可以求得 2005 ~ 2014 年可能出现的各种状态的概率,如表 2 所示。

由表 2 可以得出终极状态概率为 $\pi=[0.373\,3,0.399\,9,0.226\,6]$。这说明,使地区农业收成的变化过程在无穷多次状态转移后,“丰收”和“平收”的概率都将大于“歉收”状态出现的概率。

(下转第 12985 页)

表 2　不同土地利用类型

Table 2　Different land use types　　hm^2

1 级类 Level 1	2 级类 Level 2	2006 年	2007 年	差值 Difference	变更率//% Changing rate
农用地 Farmland	耕地 Arable land	34 733.78	43 950.77	9 216.99	26.54
	园地 Garden plot	4 137.02	4 810.53	673.51	16.28
	林地 Woodland	162 424.04	179 181.61	16 757.57	10.32
	其他农用地 Other agricultural land	12 251.07	8 245.19	-4 005.88	-32.70
建设用地 Construction land	居民点及独立工矿 Residential points and independent industrial and mining land	4 651.69	7 052.69	2 401.00	51.62
	交通运输 Transportation	645.03	526.44	-118.59	-18.38
	水利设施 Water conservancy facilities	177.94	669.52	491.58	276.26
未利用地 Unused land	未利用土地 Unused land	31 425.13	6 316.57	-25 108.53	-79.90
	其他土地 Other land	1 586.00	1 410.41	-175.59	-11.07

4　结论

(1)土地利用更新数据库的建立加快了信息化管理进程,实现了政府各管理部门之间的数据信息和技术共享。咸丰县土地利用更新调查数据库系统是在空间数据库的基础上采用先进的网络技术、GIS 技术构建的生产建库系统,大大提高了土地利用更新调查的生产速度、质量保障和管理效率。

(2)由于部分行政界线的修正,调查的辖区范围有所增加,并以此为依据作为 2008 年变更调查数据的衔接。对比分析 2007 年更新调查和 2006 年变更调查的数据得知,耕地、园地、林地、居民点及独立工矿用地面积增加,其增幅分别为 26.54%、16.28%、10.32%、51.62%,水利设施用地面积增幅最快,为 276.26%。其他农用地、交通运输用地、未利用土地、其他土地面积有所减少,其减幅分别为 32.70%、18.38%、79.90%、11.07%。

参考文献

[1] 黄照强,黄杏元.新一代土地资源信息系统的开发与设计研究[J].计算机应用研究,2003(1):113-115.

[2] 王梅.基于 ArcGIS 的土地利用更新调查数据库建设研究[J].工程勘察,2003(1):54-57.

[3] 杨军,徐世武.县(市)级土地利用数据库系统的构建和实现[J].地球科学·中国地质大学学报,2002,27(3):297-300.

[4] 廖一兰,王亚军,孙在宏.基于 GIS 系统的土地利用数据建库模式研究[J].农机化研究,2006,2(2):146-150.

[5] 中国土地勘测规划院.县(市)级土地利用数据库建设技术规范(试行)[S].2002.

[6] 中国土地勘测规划院.县(市)级土地利用数据库标准[S].2007.

(上接第 12975 页)

表 2　某地区 2005~2014 年农业收成状态概率预测值

Table 2　Probability forecast of agricultural yield from 2005 to 2014 in certain area

年份 Year	状态 E_1 概率 State E_1 probability	状态 E_2 概率 State E_2 probability	状态 E_3 概率 State E_3 probability
2005	0.133 3	0.733 4	0.133 3
2006	0.489 5	0.233 9	0.276 5
2007	0.319 8	0.480 4	0.199 8
2008	0.401 7	0.361 2	0.237 1
2009	0.362 1	0.418 8	0.219 1
2010	0.381 3	0.390 9	0.227 8
2011	0.372 0	0.404 4	0.223 6
2012	0.376 5	0.397 9	0.225 6
2013	0.374 3	0.401 1	0.224 6
2014	0.375 4	0.399 5	0.225 1

4　结语

马尔可夫预测法只是众多挖掘工具的一种,用于农作物收成状况预测也只是空间数据挖掘的一个小实例。从这个小实例可以看出,GIS 与 SDM 集成可以使有限数据的 GIS 成为无限知识的 GIS,使静态的数据变成动态的数据和知识,更进一步可用于 GIS 数据的更新。现有 GIS 数据库中存储了大量数据,数据更新通常是利用新的航空或航天 RS 数据,利用 SDM 可以获知哪些数据需要更新,并自动从 RS 影像中获得更新数据[14]。同时,SDM 可使 GIS 成为真正的“智能”空间信息系统,弥补了 ES 与 GIS 结合的不足,进而促进 GPS、DPS、RS、GIS 与 ES 的完美结合。

参考文献

[1] HAN J W,KAMBER M. 数据挖掘:概念与技术[M].范明,孟小峰,译.北京:机械工业出版社,2001.

[2] AGRAWAL R,SRIKANT R. Fast algorithms for mining association rules [M]. In Proc. 20th VLDB Conference,Santiago,Chile,1994:487-499.

[3] FAYYAD U M,PIATETSKY-SHAPIRO G,SMYTH P,et al. Advances in knoeledge discovery and data mining[M]. AAAI/MIT Press,Menlo Park,CA,1996.

[4] HAN J,CAI Y,CERCONE N. Data-driven discovery of quantitative rules in relational databases[J]. Knowledge and Data Enginnering,1993,5:29-40.

[5] 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(4):328-332.

[6] 蒋旻,梁平,贺贵明,等.GIS 系统数据挖掘功能的扩展[J].计算机工程与应用,2003(28):211-213.

[7] 蒋良孝,蔡之华.GIS 数据库中的数据挖掘[J].计算机工程与应用,2003(18):202-204.

[8] 袁红春,熊范纶,淮晓永.空间数据挖掘及其智能系统的集成框架[J].信息与控制,2002(4):304-309.

[9] LONGLEY P A,GOODCHILD M F,MAGUIRE D J. Geographical information systems,volume 2[M].唐中实,黄俊峰,尹平,等,译.北京:电子工业出版社,2004.

[10] 王雷,冯学智,郁金康.遥感影像分类与地学知识发现的集成研究[J].地理研究,2001(5):637-643.

[11] NG R T,HAN J. Efficient and effective clustering methods for spatial data mining[C]//BOCCA J B,JARKE M,ZANIOLO C. Proc Twentieth International Conference on Very Large data Bases,Santiago,Chile,Morgan Kaufmann,1994:144-155.

[12] 徐建华.现代地理学中的数学方法[M].北京:高等教育出版社,1996:93-97.

[13] 王铮,吴建平,邓悦,等.城市土地利用演变信息的数据挖掘——以上海市为例[J].地理研究,2002(6):675-681.

[14] 马荣华,黄杏元,朱传耿.用 ESDA 技术从 GIS 数据库中发现知识[J].遥感学报,2002(2):102-106.