华中农业大学学报(社会科学版),(总 84 期)2009(6)
Journal of Huazhong Agricultural University(Social Sciences Edition)

基于概念树的空间关联规则挖掘算法及其在土地利用分析中的应用*

许 红[1)],严 静[2)],张群洪[1)]
([1)]厦门大学 管理学院,福建 厦门 361005;[2)]福建行政学院,福建 福州 350002)

摘要 从地理信息系统(GIS)的角度研究空间关联规则的挖掘算法,以 GIS 智能分析和辅助决策为主要应用,从单一数据层中的空间概念层次关系研究空间关联规则的挖掘算法,利用该算法对龙海市土地利用进行空间拓扑关系挖掘,得到一些有意义的空间关联规则,例如 is_a 园地 ∧adjacent_to 交通用地 →有居民区,以及 is_a 水域 ∧adjacent_to 耕地 ∧intersects 交通用地 →有居民区。

关键词 空间关联规则;概念层次;土地利用;拓扑关系
中图分类号:F061.2 **文献标识码**:A **文章编号**:1008-3456(2009)06-0046-05

A Spatial Association Rule Algorithm Based on Concept Tree and its Application in Land Use Analysis

XU Hong[1)], YAN Jing[2)], ZHANG Qun-hong[1)]
(*[1)] School of Management, Xiamen University, Xiamen, Fujian*, 361005; *[2)] Fujian Administration Institute, Fuzhou, Fujian*, 350002)

Abstract Spatial association rule is one of the most fundamental rules in the result of spatial data mining. It emphasizes particularly on confirming the relation of data in different fields. It tries to find out the dependence of data in multi-fields. An important research field of land use analysis is to determine the spatial topology relations in the land use. This paper introduces a research on the algorithms of spatial association rule from the point of GIS based on the application of assistant decision-making in intelligent GIS. An efficient method for mining spatial association rule that uses the relation of spatial concept in the same layer has been tested and practically used in the land use of Longhai, and some very realistic and significant association rules have been determined such as: is_a gardenland ∧adjacent_to trafficland →residential, and is_a water ∧adjacent_to growland ∧intersects trafficland →residential.

Key words spatial association rule; spatial concept; land use; spatial topological analysis

土地利用/土地覆盖分析是土地开发和管理研究的核心内容,土地利用分析不仅体现为土地资源的数量、质量上,同样还表现为土地利用类型的空间拓扑关系。以往对土地利用研究多集中在以下几个方面:区域自然与社会经济条件分析,土地利用结构、布局与历史趋势分析,土地利用程度与效益分析,土地利用存在问题与对策分析。比如,李栓等[1]使用地理信息系统(GIS)的空间分析和数据统计功能,分析研究区域土地类别的数量变化,土地利用类型之间的转化情况和分布状况。而对土地利用类型在空间上的拓扑关系的研究较少,在探讨空间格局时也主要采用一些空间统计方法,因此如何准确地描述土地利用类型在空间上的关系,是土地利用研究领域的一个关键问题。

收稿日期:2009-07-06
* 国家自然科学基金资助项目(70902041);教育部人文社会科学研究项目(08JC630072)。
作者简介:许红(1967-),女,博士研究生;研究方向:企业信息化、数据挖掘。

随着信息技术和 GIS 技术的发展与普及,现有的土地利用空间数据库不仅包含地理要素的属性数据,还包含地理要素的空间定位数据。具有拓朴关系的空间地理实体是地理信息系统研究的主要对象,它们具有点、线、面 3 种基本要素特征和一定的空间作用域。但是这些土地利用 GIS 系统在功能上仅可以满足一些低层次的需求,如进行数据的收集、查询和简单的统计,人们无法从这些大量的空间数据中挖掘出对决策具有指导意义的知识[2]。利用 GIS 的空间分析方法研究区域土地利用状况,无法获取隐含在土地利用图层中的特征,例如各地类的空间拓扑关系。各种空间数据挖掘算法在土地 GIS 系统中的应用是土地利用研究需要努力的一个方向。空间数据挖掘就是从空间数据库中提取非平凡的、隐式的、未知的及有潜在应用价值的信息[3],提取的信息包含了复杂的空间关系。空间关联规则是空间数据挖掘的结果的一种最主要的知识规则,其侧重于确定数据中不同领域之间的联系,找出满足给定支持度和可信度阈值的多个域之间的依赖关系。采用空间关联规则挖掘算法作为提取用户感兴趣的知识的过程是一个独立于 GIS 的计算功能模块。空间关联规则挖掘正广泛应用于各个领域,取得令人满意的成果,而且已经有了比较现成完整的算法,但是将其应用于土地利用空间关系分析尚不多见,具有一定的理论和实践意义。

一、空间关联规则

1. 关联规则算法

设 $I=\{i_1, i_2, \cdots, i_m\}$ 是二进制文字的集合,其中的元素称为项。记 D 为交易 T 的集合,这里交易 T 是项的集合,并且 $T \subseteq I$。对应每一个交易有唯一的标识,如交易号,记作 TID。设 X 是一个 I 中项的集合,如果 $X \subseteq T$,那么称交易 T 包含 X。一个关联规则是形如 $X \Rightarrow Y$ 的蕴涵式,这里 $X \subset I$, $Y \subset I$,并且 $X \cap Y = \Phi$。规则 $X \Rightarrow Y$ 在交易数据库 D 中的支持度(support)是交易集中包含 X 和 Y 的交易数与所有交易数之比,记为 support$(X \Rightarrow Y)$,即

$$\text{support}(X \Rightarrow Y) = |\{T: X \cup Y \subseteq T, T \in D\}| / |D|$$

规则 $X \Rightarrow Y$ 在交易集中的可信度(confidence)是指包含 X 和 Y 的交易数与包含 X 的交易数之比,记为 confidence$(X \Rightarrow Y)$,即

$$\text{confidence}(X \Rightarrow Y) = |\{T: X \cup Y \subseteq T, T \in D\}| / |\{T: X \subseteq T, T \in D\}|$$

所有支持度大于最小支持度的项集称为频繁项集,或简称项集。同时满足最小支持度阈值和最小可信度阈值的规则称为强规则[4]。

给定一个交易集 D,挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度(minsupp)和最小可信度(minconf)的关联规则。它可以分为两个子问题[5]:(1)找出事务数据库 D 中所有大于等于用户指定最小支持度的项目集,具有最小支持度的项目集成为最大项目集,即寻找在给定的交易集上的所有频繁项集;(2)利用频繁集产生关联规则。

Agrawal 等[6]提出了一种基于两阶段频集思想的关联规则方法,包括两个子问题:(1)找到所有支持度大于最小支持度的项集;(2)使用第 1 步找到的频集产生期望的规则,引入了修剪技术(Pruning)来减小候选集的大小,可以显著地改进生成所有频集算法的性能。还可以引入杂凑树(Hash Tree)方法[7]来有效地计算每个项集的支持度。

2. 空间关联规则的描述

为了发现反映空间对象结构以及空间对象与空间对象之间、空间对象与非空间对象之间的关联规则,设置了一组表示空间关系的空间谓词,空间关联规则就是表示对象/谓词之间的联系[8]。

空间关联规则形如[9]:$P1 \wedge \cdots \wedge Pm -> Q1 \wedge \cdots \wedge Qm$(s%,c %),其中:$P1, \cdots, Pm, Q1, \cdots, Qm$ 中至少有一个是空间谓词,c %为此规则的信任度,其含义为满足规则前件的对象中有 c %的对象同时满足规则的后件。令 $P = P_1 \wedge \cdots \wedge P_k$,谓词合取 P 在集合 S 中的支持度,定义为 S 中满足 P 的对象数量与 S 中对象总数之比,记为 $\sigma(P/S)$;规则 $P \rightarrow Q$ 在 S 中的信任度,定义为 $\sigma(P \wedge Q/S)$ 与 $\sigma(P/S)$ 之比,即 S 中满足 P 的元素同时满足 Q 的概率,记为 $\psi(P \rightarrow Q/S)$。

在空间数据库中,可能存在大量的对象间的关联,但其中大部分只适用于少量的空间对象,或者规则的可信度较低。显式地存储在关系数据表中信息只有空间对象的非空间属性,而空间规则必须经过 GIS 的空间运算和空间分析才能得到。但是如果要计算所有空间对象的规则和空间谓词,开销将是非常大,代价过高,而且也是没有必要的。

空间关联规则的方法使用两个阈值:最小支持度和最小可信度,以过滤出描述少量空间对象的关联和具有较低可信度的规则。在对象非空间描述的

不同层次上这两个阈值均不相同,因为如果使用相同的阈值,在低的概念层次上可能找不到有趣的规则。许多谓词和概念可能在相对高的概念层次上才具有关联规则,因此,应该在不同的概念层次上定义不同的阈值,即在较高的概念层次上定语比较大的阈值,而在较低层次上,为了推导出更有用的规则,通常需要降低最小支持度和最小可信度,才能挖掘出比较有意义的关联规则。空间关联规则挖掘可以借鉴属性关联规则挖掘的思想和方法,可以从两个方面着手工作:将空间数据库中的数据进行查询、去噪等一系列预处理后,然后泛化,将挖掘的空间数据集转化成属性数据集,然后运用属性关联规则的挖掘方法进行挖掘;运用空间分析、空间计算等方法将空间数据之间的复杂关系(拓扑关系、度量关系和方位关系)的空间谓词求出后,再运用属性关联规则挖掘方法进行挖掘。比如李光强等[10]给出了基于Voronoi图构建空间数据库的算法,并采用经典的Apriori算法来例证如何从空间数据库中挖掘空间关联规则。

二、基于概念树的空间关联规则挖掘算法

1. 空间数据概化

空间数据概化是一个将与任务相关的空间数据集从较低的概念层抽象到较高概念层的过程。概念树可用于表达控制归纳进程的必要的背景,不同层次概念通常根据从一般到特殊的顺序,排序组织成为一类概念。使用概念树中高层次的概念并以简单明了的形式表达学到的规则是数据挖掘研究的重要方向之一。概念数据挖掘需要的背景知识,即概念层次结构,常以概念树的形式给出,在空间数据库中,概念层次有三种,一是:空间概念层次结构,如行政区划;二是:非空间概念层次,如土地{交通用地[公路(国道,省道,…),铁路(…)],林地(…),耕地(…)};还有一个是空间拓扑关系(即空间谓词)的概念层次。

2. 算法描述

本文提出的基于概念树的关联规则挖掘算法分为两个部分,第一部分利用GIS中的空间分析技术和空间关系运算等对地理空间中的目标和对象的空间关系以及空间行为进行描述,生成以空间属性为主的空间概念层次关系;第二部分提出利用此空间概念层次关系进行空间关联规则挖掘的算法。这种算法有效地利用了GIS中的空间分析技术,较好地处理了空间数据间的空间关系,通过复杂的空间计算和分析求得空间属性之间的关联规则,大大提高了挖掘效率。基于概念树的空间关联规则算法挖掘流程图如图1,包括以下五个步骤。

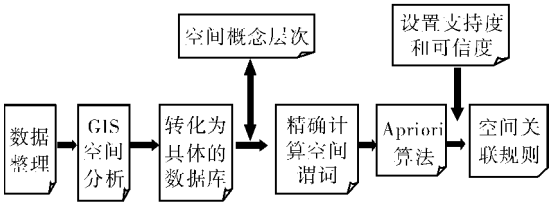

图1 空间关联规则挖掘流程图

第一步:对矢量数据图进行数据预处理,包括空缺数据填充、噪音数据处理、数据约简、连续属性数据的离散化、空间概化等。经过提取、转换、预处理后的空间数据库SDB;经过专家认可的描述空间对象层次的概念树集合,包括空间概念层次结构,空间拓扑层次结构。

第二步:通过执行空间查询和空间分析,将所有与目标相关的空间对象、目标空间对象的参照集S、与挖掘目标任务相关的空间关系的集合收集到空间数据库Task_DB中。空间信息抽取过程主要包括两个方面:空间分析和空间谓词的计算,抽取出来的空间信息以定性化描述的形式存入数据库中,这样就可以像对待属性数据一样执行传统的关联规则挖掘算法。在计算目标空间对象的最小限定矩形的交,抽取MBR间距离落在预设阈值之内的对象,并将描述对象间空间关系的谓词存贮在空间数据库Coarse_DB中,其属性值是单个值或一组值。计算空间谓词的时间复杂度很高,在提高效率方面,Koperski提出采用一种逐步求精的方法计算空间谓词。该方法首先用一种快速的算法粗略地对一个较大的数据集进行一次挖掘,然后在处理过的数据集上用代价较高的算法进一步改进挖掘的效果。然后把提取出来的空间谓词用符号化的方式存在特定的空间数据库。

第三步:为Coarse_DB中的每个谓词计算支持度和可信度,并过滤支持度低于最小支持度和可信度阈值的对象,形成数据库Frequent_coarse _DB。为了挖掘于空间谓词close-to有关的空间分类或空间关联规则,可以通过下面方法收集一些候选数据:(1)使用最小边界矩形(MBR)结构进行近似空间运算;(2)计算粗略的空间谓词,如:根据空间谓词的概

念层次可以知道 g_close_to 包括了,close,insert,contain 的结果。如果两个空间对象紧密相邻,则其最小边界矩形也一定相邻,即满足 g_close_to。但反之,如果最小边界矩形紧密相邻,两个空间对象可能相邻也可能不相邻。

第四步:根据概念层次树在 Frequent_coarse_DB 上执行精确空间计算,即采用 MRR 技术对经过第三步骤后的空间谓词关系进行检查,滤去与实际不相符合的空间谓词关系,形成新的拓扑关系数据表,并计算这些谓词的支持度,滤去支持度小的项目形成空间数据库 Fine_DB。

第五步:采用 Apriori 算法在 Fine_DB 上抽取强空间关联规则并提取出关联规则。在目标数据集中计算粗略空间谓词,然后计算得到的每个谓词对应于概念层次顶层的支持度,去掉支持度小于最小支持度阈值的谓词,得到所有频繁的 1 - 谓词。对得到的谓词进行求精,在处理后的数据集中执行空间计算,在概念层次的每一层,计算频繁的 k - 谓词,生成对应于该层的关联规则。

三、应用实例 ——以龙海市土地利用为例

以 2002 年龙海市土地利用覆盖图层(coverage 格式如图 2)为实例,利用基于概念书的空间关联规则挖掘方法进行空间数据挖掘。本研究没有考虑土地利用变化情况,而目的在于挖掘出土地利用类型空间上的关系,即挖掘出八大类土地利用类型之间的空间关联规则。龙海市土地分类主要包括八大类(耕地、园地、林地、牧草地、居民及工矿用地、交通用地、水域、未利用土地)。

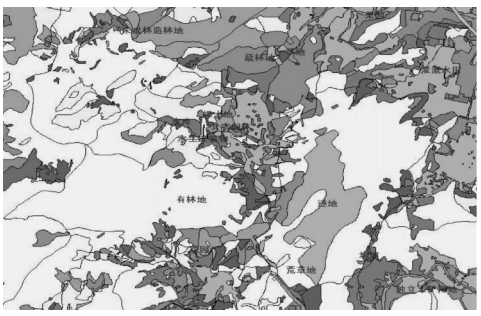

图 2　龙海市土地利用矢量图

地理信息系统可以为空间信息系统的数据收集、处理和预测结果的表示提供一个良好的平台。首先从 coverage 矢量图中提取信息到空间数据库中,然后从 GIS 数据库中提取出空间数据后,再经过上述预处理以及重新代号信息转化,便可生成进行挖掘的空间数据基础。具体挖掘实施有以下五个步骤。

第一步:充分利用 GIS 软件 ArcGIS9.0 进行空间多边形分析,利用土地利用通过 ArcSDE 将空间拓扑关系提取到数据库中,接着按照土地覆盖概念树进行概化,并进行数据预处理,使得到的数据能够适合关联规则挖掘。即选定一种土地利用类型作为挖掘的目标对象,挖掘出其它土地利用类型(如水域)与目标类型在空间位置上的关联情况,以支持度、可信度、信用可信度、作用度这四个指标作为评价标准,如水域与居民地在空间位置上的相邻情况。

第二步:确定挖掘任务所涉及的数据集,同时形成空间概念层次,进行数据预处理,形成特定空间数据库。

涉及耕地、园地、林地、牧草地、居民及工矿用地、交通用地、水域、未利用土地八个空间对象,搜集和查询相关的空间对象,通过属性查询或空间查询得到和挖掘问题相关的数据对象。如与交通用地相邻的土地利用情况。一般说来,在进行知识发现时首先需要有背景知识的概念层次信息。在空间数据挖掘中背景知识的概念层次包括空间属性的概念层次信息和(非空间)属性知识的概念层次信息。概念层次信息既可以由领域专家给出也可以通过数据分析得到。关系数据库中属性概念层次的建立通过对属性值进行归纳和概念攀升即可得到,图 3 是一个龙海市土地利用属性概念层次图。

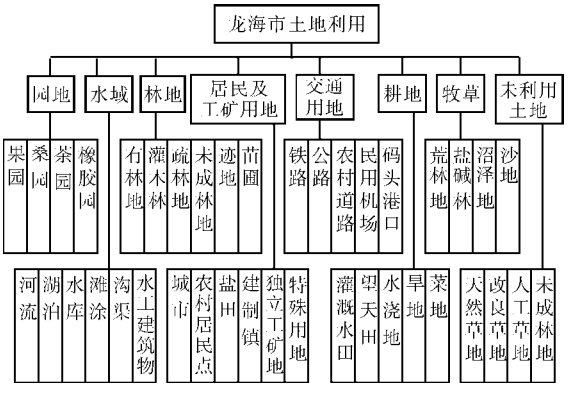

图 3　龙海市土地概念层次图

第三步:计算空间谓语(邻近),并根据设定的支持度和可信度缩减空间挖掘的数据范围。

空间信息抽取过程主要包括两个方面:空间分

析和空间谓词的计算,抽取出来的空间信息以定性化描述的形式存入数据库中,这样就可以像对待属性数据一样执行传统的关联规则挖掘算法。空间谓词的类型主要有:表示拓扑关系的谓词,如相交、覆盖等;表示空间方位和排列次序谓词,如东、西、左、右等;表示距离的谓词,如接近、远离等。本研究所需执行的空间查询主要是为了空间相邻,计算空间对象的相邻关系。系统采用 SDE 空间数据引擎以提高空间数据挖掘的效率。

从空间对象耕地、园地、林地、牧草地、居民及工矿用地、交通用地、水域、未利用土地计算空间拓扑关系,抽取出相应地相关空间谓语。计算这些谓语的支持度,滤去支持度小的项目。只有通过初始测试的数据才需要计算代价更高的算法做进一步的处理,即通过这一预处理,只有在近似阶段频繁出现的模式,才可能被更精细、更复杂的空间运算方法加以处理。

第四步:调用 Apriori 算法,根据支持度和信度,挖掘出关联规则,并选择有效关联规则,并根据专家经验进行分析。关联规则算法只是针对数据的一种算法,挖掘结果需要经过专家进一步分析,以确定其合理性。

第五步:将挖掘结果并以专题图的形式呈现,以反映数据的空间特性。挖掘结果如表 1 所示。

表 1 空间关联规则在土地利用分析中的挖掘结果

关联规则	支持度	可信度	期望可信度	作用度
园地 T 居民用地	0.081	0.217	0.355	0.611
园地 T 水域	0.069	0.182	0.316	0.585
园地 T 耕地	0.123	0.329	0.481	0.683
居民用地 T 园地	0.081	0.229	0.374	0.612
居民用地 T 水域	0.071	0.200	0.316	0.632
居民用地 T 耕地	0.116	0.326	0.481	0.677
未利用土地 T 耕地	0.054	0.300	0.481	0.623
林地 T 耕地	0.063	0.301	0.481	0.625
水域 T 园地	0.069	0.220	0.384	0.588
水域 T 居民用地	0.071	0.224	0.355	0.630

设定空间对象(多边形)之间接近关系(g_close_to)距离阈值为 5 公里,最小支持度阈值为 0.3,最小可信度阈值为 0.5,可得到如下的关联规则挖掘结果。

is_a 园地 ∧adjacent_to 交通用地 →有居民区;(支持度:0.414,可信度:0.517,期望可信度:0.355)。

is_a 水域 ∧adjacent_to 耕地 ∧intersects 交通用地 →有居民区;(支持度:0.692,可信度:0.817,期望可信度:0.365)。

is_a 交通用地 ∧adjacent_to 耕地 ∧adjacent_to 园地 →有居民区;(支持度:0.332,可信度:0.747,期望可信度:0.425)。

四、结论

本文讨论了在 GIS 的单一空间数据层中以空间拓扑属性为主生成空间数据概念层次关系,采用概念层次树对形成的拓扑关系进行概化后形成新的拓扑关系数据表,利用关联规则算法提取出具有现实意义的空间关联规则。本文分析了空间关联规则挖掘的一般理论和方法以及空间关联规则挖掘对象的特点。为了提高空间关联规则挖掘算法的效率,提出了利用空间数据的概念关系进行空间关联规则的挖掘,以及一种基于概念树的空间关联规则挖掘算法,并以龙海市 2008 年矢量图为例,在设定距离阈值下挖掘出不同土地利用类型间的空间关系,比如园地和交通用地相邻,则出现居民用地的支持度达到 0.414,可信度达到 0.517,具有一定的现实意义。但本文仅考虑单一图层的矢量数据挖掘,如何提高多尺度、多图层的大型 GIS 或空间数据仓库中空间关联规则挖掘的效率问题有待进一步研究。

参 考 文 献

[1] 李栓,王红梅. 基于 GIS 的哈尔滨市土地利用动态变化分析[J]. 地理信息世界, 2008(6):73-77.

[2] 王秀兰,包玉海. 土地利用动态变化研究方法探讨[J]. 地理科学进展,1999,18(1):81-86.

[3] 史培军,陈晋. 深圳市土地利用变化机制分析[J]. 地理学报,2000,55(2):151-159.

[4] 朱会义,何书金,张明. 土地利用变化研究中的 GIS 空间分析方法及其应用[J]. 地理科学进展,2001,20(2):101-110.

[5] 陆丽娜,陈亚萍,魏恒义等. 挖掘关联规则中 Apriori 算法研究[J]. 小型微型计算机系统,2006,21(9): 940-942.

[6] AGRAWAL T, SWAMI A. Mining association rules between sets of items in large databases[C]. New York: ACM, 1993: 207-216.

[7] PARK J S, CHEN M S, YU P. An effective hash based algorithm for mining association rules[J]. ACM SIGMOD Record, 1995, 24(2): 175-186.

[8] 郭仁忠. 空间分析[M]. 武汉:武汉测绘科技大学出版社,1998:120.

[9] 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,1997,2(4):328-332.

[10] 李光强,邓敏,朱建军. 基于 Voronoi 图的空间关联规则挖掘方法研究[J]. 武汉大学学报:信息科学版,2008,33(12):1142-1145.

(责任编辑:刘少雷)