文章编号: 1007 4619(2006) 03 0289 05

一种面向主题的基于多层次空间概念关系的关联规则挖掘算法

陈江平[1], 李平湘[2]
(1. 武汉大学 遥感信息工程学院, 湖北 武汉 430079; 2. 武汉大学 测绘遥感信息工程国家重点实验室, 湖北 武汉 430079)

摘 要: 提出了一种面向主题的基于多层次空间概念的关联规则挖掘算法 FT_MLSAM。在 FT_MLSAM 算法中, 先根据用户感兴趣的主题确定挖掘的概念关系, 然后对所涉及的多个空间数据层进行连接, 生成空间视图, 最后进行属性泛化, 转化成一般属性关联规则的挖掘, 实验证明算法是有效的。

关键词: 数据挖掘; 空间关联规则; 面向主题; 多概念层次
中图分类号: P208 文献标识码: A

An Algorithm about Spatial Association Rule Mining Based on Thematic

CHEN Jiang-ping[1], LI Ping-xiang[2]
(1. *School of Remote Sensing and Information Engineering, Wuhan University, Hubei, Wuhan* 430079, *China*;
2. *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Hubei, Wuhan* 430079, *China*)

Abstract: In this paper, an algorithm of mining multi-level spatial association rules is presented which's based on theme. We called the algorithm as FT_MLSAM in the paper. In FT_MLSAM, it first constructs the concept relation based on the thems on which the user is interested. Second, it joins the spatial layers which are used in the data mining. Last it turns spatial association mining into association mining by attribute generalizing. It is proved right from the experiment.

Key words: data mining; spatial association rule; based on thematic; multi-layer spatial concept lattice

1 引 言

空间数据库的组织一般是按主题进行的, 如某地区建立的交通数据库包括公路信息、铁路信息、水路信息和航空信息等。然而, 空间数据本身又具有概念层次信息, 因此在空间数据挖掘中, 可以首先根据主题建立相应的概念层次关系, 进而利用这些概念层次关系确定进行挖掘的空间数据层, 然后进行属性泛化, 转化为一般关联规则的挖掘[1]。

基于事务数据库的属性关联规则的挖掘发展很快, 这点从属性关联规则挖掘的方法、挖掘的内容、挖掘的形式的多样化可以看出。由于空间数据的属性之间关系的复杂性、数据的尺度特征和数据模糊性等特点, 使得属性关联规则挖掘的方法不能完全适用于空间关联规则的挖掘。目前, 关于空间关联规则的挖掘研究还不多, 基本上是基于一般事务数据库的关联规则挖掘算法研究[2—4]。这些研究的特点是基于属性数据库, 将这些算法应用于空间数据挖掘存在的主要不足有: 空间数据量庞大及空间数

收稿日期: 2004 04 22; 修订日期: 2005 03 17
基金项目: 测绘遥感信息工程国家重点实验室开放基金资助项目(03 0101) 和武汉大学博士启动基金资助。
作者简介: 陈江平(1975—), 女, 讲师, 2003年 7月博士毕业于武汉大学地理信息系统专业。研究领域为 GIS中的数据挖掘和 3DGIS, 已发表论文 10多篇。E-mail: chen_lisa@sohu.com。

据之间存在复杂的关系 (拓扑关系、位置关系和度量关系)。为了解决空间数据量庞大问题, 本文提出了一种面向主题的基于多层次空间概念关系的关联规则挖掘算法。

2 多层次空间关联规则的定义及其空间概念关系

2 1 多层次空间关联规则的定义

定义 1设有空间数据层 $L1$和 $L2$, A ttr($L1$)表示 $L1$的附属属性集, A ttr($L2$)表示 $L2$的附属属性集, Ac=A ttr($L1$)∩ A ttr($L2$)为 $L1$和 $L2$的公共属性集, 令 $A1$=A ttr($L1$) - Ac, $A2$=A ttr($L2$) - Ac。多层次空间关联规则挖掘就是从空间数据层 $L1$和 $L2$及其附属属性集中挖掘所有满足给定支持度 m insup和置信度 m inconf的蕴含式: $X \Rightarrow Y$, ($X \in L1$, $Y \in L2$) OR($X \in L2$, $Y \in L1$), 其中 X, Y是属性的集合, 称作项集。这种规则的直观含义是: 在 $L1$与 $L2$按公共属性 Ac=1连接得到的空间数据集中, 包含 X 的元组通常也包含 Y。令 m, n分别表示 $L1$, $L2$中包含公共属性 (Ac= 1)的元组的个数, 则 $L1$, $L2$按公共属性 Ac=1连接后关系中元组的个数[5]

joinsize=$m \times n$。

若 count(X), count(Y)分别表示 $L1$ 中包含 X(Ac=1), $L2$中包含 Y (Ac= 1)的元组的个数, 则 $X \Rightarrow Y$的支持度:

$$\sup(X \Rightarrow Y) = \sup(Y \Rightarrow X) = \text{count}(X) \times \text{count}(Y) / \text{joinsize}$$

$X \Rightarrow Y$的置信度:

$$\text{conf}(X \Rightarrow Y) = \text{conf}(X \Rightarrow Ac) \times \text{conf}(Ac \Rightarrow Y) = 100\% \times \text{conf}(Ac \Rightarrow Y) = \text{count}(Y) / m$$

$Y \Rightarrow X$的置信度:

$$\text{conf}(Y \Rightarrow X) = \text{conf}(Y \Rightarrow Ac) \times \text{conf}(Ac \Rightarrow X) = 100\% \times \text{conf}(Ac \Rightarrow X) = \text{count}(X) / n$$

2 2 基于多个空间数据层的空间概念关系

在给定主题的空间关联规则挖掘中, 要全面了解一个主题的有关信息, 往往需要多个空间数据层的信息。例如, 要分析某个县市的交通情况, 至少需要从 4个空间图层上进行分析: 行政区划图层、道路图层、水系图层和航空线路图层 (图 1)。

图 1说明, 主题属于第一层信息, 位于根节点, 与主题相关的层是第二层, 第二层的具体分类是第三层。确定具体的主题后, 用户通常对跨越不同空间数据层次的项目之间能否产生规则感兴趣。例如, 人们一般对某个县市的公路总里程感兴趣, 因此用户需要分别计算该县市的高速公路总里程数、国家一级公路总里程数、国家二级公路总里程数以及通车道路但无级别的路的总里程数。

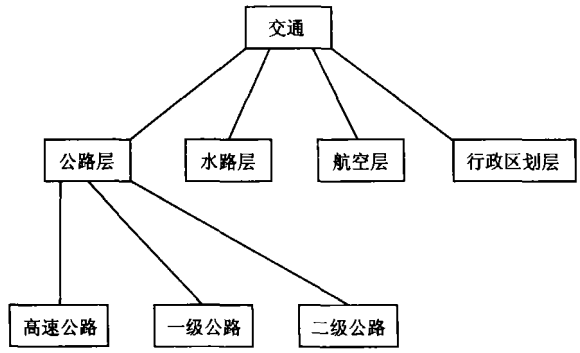

图 1 一个多空间数据层的概念层次图

Fig 1 An concept layer formulti_layer

在进行空间关联规则的挖掘时, 根据主题按照形成的概念层次关系选定需要的数据视图。如: 主题为研究水路对公路客运的影响, 则按照概念层次树在第二层的公路层和水路层的数据视图上进行挖掘; 主题为: 高速公路周边国家级公路的分布情况, 则按照概念层次树在第三层的公路层的三个子节点的数据形成的数据视图上进行挖掘。由此可见, 一旦确定了挖掘主题, 就可以确定其相应的概念层次关系, 进而可以根据此概念层次的分支和层次确定属性泛化的区域个数和中心值。

3 基于多层次空间概念关系的关联规则挖掘算法

3 1 算法思想

面向主题的空间关联规则挖掘算法 (FT_ M LSAM)分为以下几步: (1)连接与挖掘主题相关的所有空间数据层, 遍历并根据概念层次关系进行属性泛化形成挖掘数据表; (2)在挖掘数据表上按一般属性关联规则的挖掘方法进行空间关联规则的挖掘。

在算法 FT_M LSAM 的第 (1)步的工作中, 连接相关的挖掘空间数据层比较流行的有两种方法, 即空间对象的 M BR连接或对象元素空间连接。在本算法中采用空间对象的 M BR连接方法。假设有两个数据元素集合, $A = \{A_1, A_2, \cdots, A_n\}$, $B = \{B_1, B_2, \cdots, B_m\}$, 其中 A_i, B_i均可以看作是线对象, 如河流, 在 A_i和 B_i中分别包含自己的对象元素。所谓

MBR连接是计算所有满足 $MBR(A_i) \cap MBR(B_i) \neq \phi$ 的 A_i 和 B_i, 而对象元素空间连接, 则是对 A_i 和 B_i 中的具体对象元素进行计算, 以得到所有满足 $MBR(a_i) \cap MBR(b_i) \neq \phi$ 的 a_i 和 b_i。

对于属性的泛化, 文献[6]中提出了使用云模型和基于黄金分割率的方法来生成概念层次关系, 并使用云模型进行属性泛化。在本文中由于主题是用户在进行数据挖掘前就确定的, 所以可以根据主题建立一个概念层次关系 H, 然后依据 H 确定阈值并对数据库中的数据值进行属性泛化。

3.2 连接空间数据层并进行属性泛化软划分的算法描述

A 对算法第(1)步确定的空间数据层进行连接。

在算法中采用 R 树存储空间数据层; R 树中每一个非叶节点代表数据集空间中的一个矩形, 该矩形为包含其所有子节点的最小边界矩形 MBR。非叶节点由多个(MBrect, child)结构的子最小边界矩形 MBR组成, 其中 child为子节点指针(存储子节点在索引文件中的偏移地址), MBrect为与子节点 child相关的 MBR。而 R 树叶节点由多个(MBrect, Object)结构组成, 其中 Object包含了空间对象的 ID号, 通过该 ID号可以得到实际空间对象存储在数据库中的偏移地址, MBrect为对象 Object的 MBR。

Algorithm SpatialJoin

Input: 待连接的挖掘数据层 r_1, r_2

Output: 连接后的数据层 r; r中所有实体满足: $MBR r_1 \cap MBR r_2 \neq \phi$

SortandIntersectionTest(r_1, r_2, r)

For (r_1, r_2 中每一个对象)　do

if(r_1 -> level=0&&r_2 -> level=0)

then r_1 and r_2 是叶子节点

Output(r_1, r_2 中的对象的 ID)

Else if(r_1 -> level=0&&r_2 -> level≠0)

SearchforJoin(r_1. Object, r_2, child);　/将 r_1 中的叶节点插入到 r_2 的子节点中

Else if(r_1 -> level≠0&&r_2 -> level=0)

SearchforJoin(r_2. Object, r_1, child);　/将 r_2 中的叶节点插入到 r_1 的子节点中

Else /r_1 and r_2 都不是叶子节点

{

MBrect=IntersectRect(r_1. Object, r_2. Object, r);

SpatialJoin(r_1. child, r_2. child, MBrect);

}

B 属性泛化

输入: 上述 A 算法得到的连接后的相关数据集 R, 属性集 A_i ($1 \leq i \leq n$);

概念层次树 H_i ($1 \leq i \leq n$);

依据 H_i 的各个层次, 属性 A_i 的对应阈值 t_i;

输出: 泛化后的数据集 R'

$R' \leftarrow R$; /将 R' 初始化为 R,

While R' not eof do

if A_i. value$\in \{t_1, t_2, \cdots, t_i\}$ /当前记录的 A_i 属性的值落在各个阈值之中

then沿 H_i 将对应的值在 R' 中做适当替换

else从 R' 中删除 A_i

end if

从 R' 中删除重复元组;

endwhile;

4 算法实例及规则解释

4.1 算法实例

设有如下主题: 预测雷电频发区域中杆塔发生故障的规则。在雷电频发区域中杆塔发生故障有以下情况: (1)由于杆塔被雷击引起跳闸, 发生故障; (2)由于杆塔污染严重, 发生故障。

用户需要知道雷电频发区域中哪些因素导致杆塔易受雷击, 从而引起故障, 可以采取哪些措施进行改进, 同时了解污染与雷击是否有联系。综合考虑, 涉及的空间数据层有: 杆塔线路分布图、道路分布图、植被层、雷电易发区图层和污染等级图层。

具体的空间概念层次关系如图 2所示。

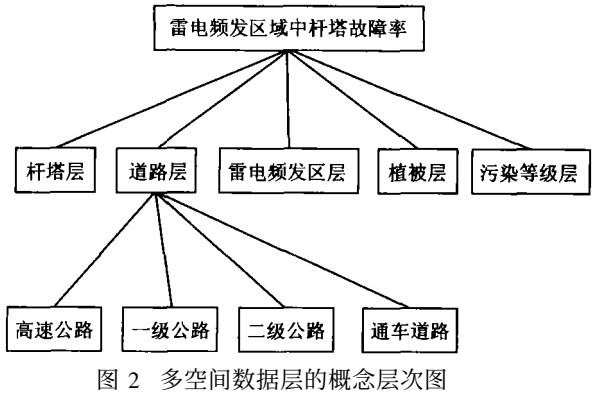

图 2　多空间数据层的概念层次图

Fig 2　An concept layer formulti_layer

设有如下经过泛化处理后的空间数据集(由用户根据挖掘主题, 选定所需要的属性)如表 1所示。

表 1 进行数据挖掘的数据表

Table 1 DataSet for data mining

植被平均高度	植被面积	与植被距离	雷击频率	杆高	车流量	污染级别	污染源名称	道路宽度	与道路距离	杆塔故障率
高	较大	较远	较高	高	大	高	水泥厂	较宽	较远	高
高	较大	较近	较高	高	大	高	煤窑	较宽	较近	高
低	较大	较远	较高	高	大	高	采石厂	较宽	较远	高
高	较小	较近	较低	高	大	高	采石厂	较宽	较近	高
低	较小	较远	较低	低	大	低	砖瓦厂	较宽	较远	低
低	中等	中	中	低	中	低	水泥厂	较宽	中	低
高	较大	中	较低	高	小	高	采石厂	较窄	中	低
低	中等	较近	中	低	小	低	煤窑	较窄	较近	低
高	较大	较近	较高	高	大	高	水泥厂	较窄	较近	高
高	较小	较近	较低	高	大	高	采石厂	较宽	较远	高
低	较小	较远	较低	低	大	低	砖瓦厂	较宽	较近	低
低	中等	中	中	高	大	低	水泥厂	较宽	中	低
高	较大	中	较高	高	小	高	采石厂	较窄	中	高
低	中等	较近	中	低	小	低	煤窑	较窄	较近	低
高	较大	较近	较高	高	大	高	水泥厂	较窄	中	高

以最小支持率 6%和最小置信度 75%(意味着在全部的记录中同时出现以上 11项至少有 6%,在这些包含以上 11项的记录中每条规则置信的概率为 75%)在空间视图的泛化数据集中进行挖掘,得到 8个大的 4项目集,因此产生 8条关联规则,以杆塔的故障率为规则后件,以其余属性为规则前件,得空间关联规则如下:

规则 1:距离植被较近∧植被较高∧杆塔雷击频率较高→故障率高

规则 2:距离植被较近∧植被面积较大∧杆塔雷击频率较高→故障率高

规则 3:距离道路较近∧车流量大∧杆塔雷击频率较高→故障率高

规则 4:距离道路较近∧车流量大∧杆塔污染严重→故障率高

规则 5:距离植被较远∧植被较低∧杆塔雷击频率较低→故障率低

规则 6:距离植被较远∧植被面积较小∧杆塔雷击频率较低→故障率低

规则 7:距离道路较远∧车流量小∧杆塔雷击频率较低→故障率低

规则 8:距离道路较远∧车流量大∧杆塔污染轻微→故障率低

4.2 算法挖掘的规则解释与评价

按本算法和用户提出的挖掘要求:预测雷电频发区域中的杆塔故障率。要挖掘得到以上空间关联规则集,共涉及 5个空间数据层,分别是:杆塔线路分布图、道路分布图、植被层、雷电易发区图层、污染等级图层。在实例中,我们预设最小支持率 $s=6\%$,最小置信度 $c=75\%$(根据用户使用规则的要求进行预设)。通过挖掘发现:

从当杆塔处于雷电频发区域中,距离较高植被较近的杆塔发生故障的频率高;距离道路较近,且道路较宽(车流量大)的杆塔的故障率较高;污染程度较重的杆塔的故障率也较高。可以认为处于雷电频发区域中的杆塔的故障率高。用户可以据此进一步分析深层次的原因,从容易遭受雷击的原因着手,如植被的类型、植被的高度、杆塔的污染等级等具有空间关联关系的角度,可以为他们提供一条有利的线索,可能因为处于某种植被附近,而该类型的植被生长得比较高,容易引雷,所以应该改变雷电频发区域中杆塔附近植被的类型。获取的规则对雷电频发区域中的植被的种植和砍伐的决策将起到很好的指导作用,人们可以有意规划出雷电频发区域中对超高压线路和杆塔受雷击影响小的植被的空间种植布局。

需要明确的是, 例中形成的空间关联规则在生产上可能并非预先有意创造的, 但是从大量的数据中通过空间关联规则, 发现了潜在的必然规律, 即如果具有这种空间关联关系, 则该关系下的与空间决策有关的特征属性值将会显著高于 (或低于)不具有该关系下的特征属性值。

5　实验结果

为了验证算法 FT_MLSAM 的性能, 我们在一台 Windows2000操作系统的内存为 256M 的 C4 1.7G 微机上做了测试, 采用合成数据进行算法测试。主题为预测雷电频发区域中的杆塔故障率。测试数据集共包括 5个数据层各含有 3—5个属性, 每个属性泛化后有 2—10个属性值。将这 3个数据层进行连接, 成为有 11个属性的一个数据集, 而各层的最低支持度均为 6%, 最低置信度均为 75%。

测试了本算法随记录数目的增加引起的时间的变化 (时间复杂性), 将测试数据库的元组数从 1000 开始, 逐渐递增到 5000, 算法执行时间的增加如图 3。

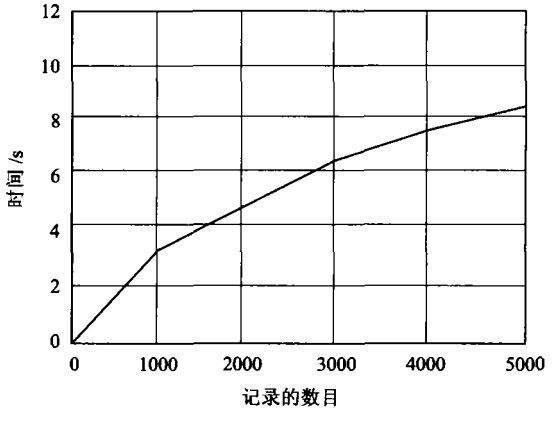

图 3　本算法的时间复杂性图

Fig. 3　Execution time with the increasing of itemsets

6　结　论

本文提出了一种 FT_MLSAM 挖掘算法, 经过实验证实, 它具有如下特点:

FT_MLSAM 根据给定的主题形成空间概念层次关系, 它是基于多个空间数据层和数据表所形成。利用定义的空间概念层次关系, FT_MLSAM 可进行属性泛化并转化成一般属性关联规则的挖掘。因此, FT_MLSAM 不仅可在单个表 (关系)内进行数据挖掘, 而且可挖掘存放在不同层中的空间数据表中蕴藏的知识和信息。

多表间的空间关联规则挖掘是富有实际意义的研究课题, 例如城市中的交通流量的分析、气象模式的分析、气候与植物分布的变化趋势等, 必须在多个空间数据层和数据集中才能发现。作者下一步的工作将是进一步地发展高效实用的多表间空间关联规则的挖掘与并行处理技术相结合的算法。

参 考 文 献 (References)

[1] Chen J P, Bian F L, Fu Z L, *et al*. An Improved Algorithm of Apriori[J]. *Geomatics and Information Science of Wuhan University*, 2003, (1): 94—99. [陈江平, 边馥苓, 付仲良等. 一种 Apriori的改进算法[J]. 武汉大学学报 (信息科学版), 2003, (1): 94—99.]

[2] Cheng J H, Shi P F. Efficient Mining Algorithm for Multiple_Level Association Rules[J]. *Journal of Software*, 1998, **9**(12): 937—941. [程继华, 施鹏飞. 多层次关联规则的有效挖掘算法[J]. 软件学报, 1998, **9**(12): 937—941.]

[3] Zuo W L. A Parallel Algorithm for Mining Association Rules Across Multi Tables[J]. *Mini Micro System*, 1999, **20**(8): 574—577. [左万利. 多表间关联规则的并行挖掘算法[J]. 小型微型计算机系统, 1999, **20**(8): 574—577.]

[4] Zhang Y L, Zhong W J, Mei S E. An Algorithm for Mining Multi_level Association Rules in a Table and It's Application in One MEIS[J]. *Computer Engineering and Applications*, 2001, **14**: 91—92, 102. [张玉林, 仲伟俊, 梅姝娥. 一类表内多概念间多层次关联规则挖掘算法及应用[J]. 计算机工程与应用. 2001, **14**: 91—92, 102.]

[5] Gao F, Xie J Y. Theoretical Foundation of Quantitative Association Rules[J]. *Computer Engineering*, 2000, **26**(11): 47—49. [高峰, 谢剑英. 多值属性关联规则的理论基础[J]. 计算机工程, 2000, **26**(11): 47—49.]

[6] Li D Y, Di K C, Li D R, *et al*. Mining Association Rules with Linguistic Cloud Mode[J]. *Journal of Software*, 2000, **11**(2): 143—158.