文章编号:1671-8860(2013)01-0095-05　　文献标志码:A

利用叠置分析和面积计算实现空间关联规则挖掘

董　林[1]　舒　红[1]　牛　宵[2]

(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　山东省国土测绘院,济南市历山东路 9 号,250013)

摘　要:提出利用叠置分析和面积计算实现空间关联规则挖掘的算法 I-Apriori 及其改进算法 FI-Apriori,这两种算法不依赖于空间数据的事务化,可以直接从矢量多边形图层中提取所有强关联规则。采用实际数据对两种算法进行了检验,验证了它们的可用性与有效性,并对挖掘所得空间关联规则的筛选和可视化方法进行了探讨。

关键词:空间关联规则;叠置分析;面积;支持度;算法
中图法分类号:P208

空间关联规则的概念最早由 Koperski 和 Han 提出,指的是一种形如 $A \rightarrow B$ 的蕴涵式,其中 A 和 B 是谓词的集合($A \cap B = \varnothing$,且 $A \cup B$ 至少包含一个空间谓词)[1-2]。支持度(support)和置信度(confidence)是空间关联规则的两个基本评价指标,对于规则 $A \rightarrow B$,其支持度与置信度计算公式分别为:

$$\text{support}(A \rightarrow B) = \text{support}(A \cup B) = P(A \cup B) \tag{1}$$

$$\text{confidence}(A \rightarrow B) = P(B \mid A) \tag{2}$$

空间关联规则的提取方法主要有以下两类:① 基于事务(transaction)的方法,首先将空间数据事务化,然后对所得的事务数据集进行规则挖掘;② 基于叠置分析(overlay analysis)的方法,利用叠置分析、距离和面积计算等操作直接从空间图层中提取关联规则[3-4]。基于叠置分析的方法具有对领域知识依赖度低、挖掘结果物理含义明确等优点,但是在输入数据类型、谓词类别或谓词数量等方面有较强局限性。文献[5-6]所研究的同位模式(co-location patterns)挖掘只关注空间同位这一种谓词;文献[7]提出的聚类空间关联规则(clustered spatial association rule)提取方法是针对点图层设计的,且所得规则仅包含两个谓词;文献[8]提出的异质空间关联规则分析方法同样只能对两个谓词构成的关联规则进行分析。输入数据和谓词上的局限性使得这些方法难以处理一般性的问题。因此,本文提出基于叠置分析和面积计算的空间关联规则挖掘算法,可从多种类型空间数据中提取出多谓词(谓词类别和数量均可为多个)关联规则,并给出了对应的可视化方法。

1　基于叠置分析的挖掘算法

由式(1)、式(2)可知,空间关联规则的支持度和置信度计算都依赖于谓词集的支持度计算。本文所采用的支持度计算方法是文献[7]中方法的推广。

1.1　基于聚类图叠置的支持度计算方法[7]

对空间点图层 l 中的点进行聚类,所得簇的轮廓(矢量多边形)所覆盖的区域称为满足 l 的区域,记作 clusters_areas(l)。对于点图层集合 $S = \{l_1, l_2, \cdots, l_n\}$,称 clusters_areas($l_1$)、clusters_areas($l_2$)到 clusters_areas($l_n$)的交(intersection)为满足 S 区域,记作 clusters_areas(S)。

进行聚类空间关联规则提取时,将一个点图层看作一个谓词,将图层集合看作谓词集。称满足谓词集 S 区域的面积与研究区域 R 总面积之

收稿日期:2012-11-18。
项目来源:国家 863 计划资助项目(2008AA12Z201);国家 973 计划资助项目(2011CB707103);国家自然科学基金资助项目(41171313)。

比为 S 的聚类支持度(clustered support),记为 CS(S),即有:

$$CS(S) = \text{area}(\text{clusters\_areas}(S))/\text{area}(R) \tag{3}$$

将式(3)分别代入式(1)和式(2),就可以得到规则 $A \rightarrow B$(A 和 B 均可包含多个谓词)的聚类支持度 $CS(A \rightarrow B)$和聚类置信度(clustered confidence)$CC(A \rightarrow B)$公式:

$$CS(A \rightarrow B) = CS(A \cup B) \tag{4}$$

$$CC(A \rightarrow B) = CS(A \cup B)/CS(A) \tag{5}$$

1.2 利用多边形图层叠置计算支持度

基于聚类图叠置的支持度计算方法要求输入数据为可聚类的点图层,但实际参与支持度计算的是矢量多边形图层(簇的轮廓),不难将其推广为一种通过对多边形图层进行叠置分析来计算谓词集的支持度方法。

对于谓词集 $P = \{p_1, p_2, \cdots, p_k\}$($k \in N^+$),如果其中每一个谓词 p_i 均对应于一个矢量多边形图层 l_i($i \in \{1, 2, \cdots, k\}$),那么,该谓词集的支持度等于图层 $l_1, l_2, \cdots, l_k$ 的交 l_{com} 的面积与研究区域 R 的总面积之比,即

$$\text{support}(P) = \text{area}(l_{\text{com}})/\text{area}(R) \tag{6}$$

该方法的含义可以理解为:对于一个谓词集,其支持度就是研究区域中满足这些谓词的合取式的区域所占比例,例如对于{气温高,降水高},若气温高且降水高的区域占总面积的15%,则其支持度是15%。

该方法的输入数据(矢量多边形图层)可以通过以下两种途径获得:① 利用已有的矢量图,例如行政区划图等;② 通过空间分析生成矢量多边形图层,例如点图层聚类结果轮廓提取[7]、栅格数据矢量化、缓冲区操作所得图层。推广后的方法支持的数据类型较多,并且支持所有可用矢量多边形图层表述的谓词。

1.3 空间关联规则挖掘算法 I-Apriori

d 个谓词($d \geqslant 2$)可以组成的关联规则有 $n = 3^d - 2^{d+1} + 1$ 种[9],例如,本文实验中 $d = 47$,$n = 2.66 \times 10^{22}$。逐一检测 n 种规则找出全部强规则不现实,必须有高效的挖掘算法才能实现。

将本文支持度计算方法与经典 Apriori 算法框架相结合,得到可以直接从矢量多边形图层中挖掘关联规则的 I-Apriori(Intersect-Apriori)算法,其算法描述如下。

输入:研究区域 R,矢量多边形图层集合 $L = \{l_1, l_2, \cdots, l_n\}$,最小支持度阈值 minsup

输出:频繁谓词集 F

```
(1) map<pset, layer> map; //谓词集-图层映射
(2) for (i = 1; i <= n; i++)
(3) map.add({i}, l_i);
(4) for (i = 1; i <= n && F_{i-1} != φ; i++) {
(5) if (i == 1) C_i = {{1}, {2}, … {n}};
(6) else C_i = aprioriGen(F_{i-1});
(7) for each c in C_i
(8) if (is_frequent(c)) F_i = F_i ∪ c;
(9) }
(10) return F = ∪_i F_i;
procedure isFrequent(pset c)
(1) layers=map.getLayers(c);//获取谓词对应图层
(2) l_com = intersect(layers);
(3) if(area(l_com) / area(R) > minsup) {
(4) map.add(c, l_com); //更新 map
(5) return true;
(6) }
(7) return false;
```

其中,aprioriGen 是经典 Apriori 算法中用于连接频繁项集生成下一阶候选项集的函数;area 是面积计算函数;intersect 是图层求交函数,当输入仅含1个图层时直接返回该图层,否则,返回这些图层的交(仍以图层形式保存)。

1.4 关联规则挖掘中的快速求交方法

假设谓词 p_a、p_b、p_c、p_d 分别对应于图层 l_a、l_b、l_c 和 l_d,频繁3-谓词集 $P_1 = \{p_a, p_b, p_c\}$ 和 $P_2 = \{p_a, p_b, p_d\}$ 对应的图层分别为 $l_1 = (l_a \cap l_b \cap l_c)$ 和 $l_2 = (l_a \cap l_b \cap l_d)$。I-Apriori 算法计算由 P_1 和 P_2 连接得到的候选4-谓词集 $P = \{p_a, p_b, p_c, p_d\}$ 对应图层 $l = (l_a \cap l_b \cap l_c \cap l_d)$ 时,会直接求 l_a、l_b、l_c 和 l_d 这4个图层的交。但根据集合运算的幂等律和交换律,有 $(l_a \cap l_b \cap l_c) \cap (l_a \cap l_b \cap l_d) = (l_a \cap l_b \cap l_c \cap l_d)$,即 $l = l_1 \cap l_2$,故只需要求 l_1 和 l_2 这两个图层的交即可得到 l,如图1所示。

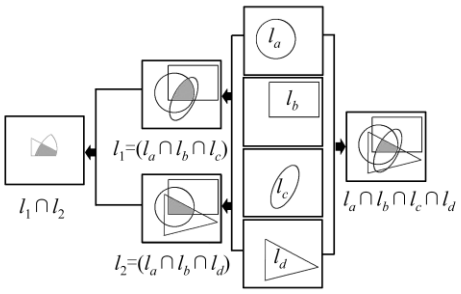


图1 图层求交的两种方法

Fig. 1 Two Approaches to Get the Intersection of Layers

一般地,记频繁($k-1$)-谓词集 P_1 和 P_2 的对应图层为 l_1 和 l_2($k \geqslant 2$),由 P_1 和 P_2 连接得到的

候选 k-谓词集为 P,P 的对应图层为 l。对于图层 l_1、l_2 和 l,恒有式(7)成立:

$$l = l_1 \cap l_2 \tag{7}$$

式(7)同样可以用幂等律和交换律加以证明,证明从略。在 I-Apriori 算法中,频繁谓词集是逐级提取的,所以在计算 l 之前必然已经求过 l_1 和 l_2,所以任意候选 k-谓词集的对应图层都可以由两个图层求交得到,而不必去计算 k 个图层的交。本文将这种利用已有图层计算候选谓词集对应图层的方法称为快速求交方法(fast intersect)。

1.5　改进算法 FI-Apriori

对于任意候选谓词集,快速求交方法最多只需要计算两个图层的交,并且所使用的图层覆盖范围较小(一组图层的交不大于这组图层中的任意一个)。由于图层数量和内容较少,这种方法通常优于直接计算的方法。显然,利用该方法对 I-Apriori 算法进行改进可以缩短关联规则挖掘耗时,改进后的算法称为 FI-Apriori(fast intersect-apriori)算法。FI-Apriori 算法描述主体部分与 I-Apriori 相同,但子程序 isFrequent 需要替换为:

```
procedure is Frequent(itemset c)
(1) layers= getFILayers(c, map);
(2) l_com = intersect(layers);
(3) if(area(l_com) / area(R) > minsup) {
(4) map.add(c, l_com); //更新 map
(5) return true;
(6) }
(7) return false;
```

其中,函数 getFILayers 在 c 中只包含 1 个谓词时直接返回 c,否则先获取连接生成候选 k-谓词集 c 的两个频繁 $(k-1)$-谓词集,然后返回它们对应的两个图层,这样就可以通过快速求交方法来得到 c 的对应图层。

当提取频繁 1-谓词集和频繁 2-谓词集时 FI-Apriori 算法与 I-Apriori 算法并无明显差别;但提取频繁 k-谓词集时($k>2$),I-Apriori 算法要计算 k 个图层的交,而 FI-Apriori 算法始终只需要计算两个图层的交,具有很好的可伸缩性。

2　实验与分析

2.1　数据准备

本文以吉林、黑龙江和辽宁等 3 省为研究区域(图 2),所用数据包括该区域的行政区划图、2005 年 81 个台站的气象记录、DEM 和 MODIS NDVI 影像等。

研究区域的行政区划图为矢量格式,按照省份拆分为 3 个图层;气象数据为月值气温和降水记录,求取季均值后通过克里金插值得到覆盖整个区域的 8 幅栅格图像,分别对应于各季度的气温和降水;DEM 数据分辨率为 90 m,对其进行了坡度和坡向提取;NDVI 数据分辨率为 500 m,提取各季度的均值后仍以栅格格式存储。DEM、坡度以及 4 个季度的气温、降水、NDVI 图像均通过模糊聚类方法分割为高、中、低 3 类并矢量化,坡向按照向阳和背阳分为两类并进行矢量化。经过以上数据准备操作共得到 16 组 47 个矢量多边形图层。

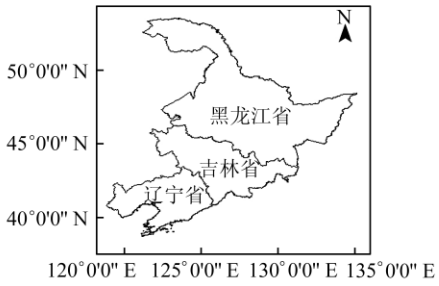


图 2　研究区域

Fig. 2　Map of Study Region

2.2　关联规则挖掘

采用 I-Apriori 算法和 FI-Apriori 算法对实验数据进行了关联规则挖掘,最小支持度阈值设为 0.1,最小置信度阈值设为 0.75,为减少规则的冗余,程序限定规则后件只包含 1 个谓词(实验所用程序及示例数据可以在 http://www.c2001.net/downloads.html 下载)。两种算法挖掘得到的频繁谓词集和关联规则数量、次序、内容均完全一致,其中频繁谓词集共 12 300 个,最大谓词数为 10;强关联规则共 36 743 条,前件最多包含 9 个谓词。两种算法耗时对比见表 1。

实验中两种算法所用时间几乎全部用于进行谓词集支持度计算,而支持度计算耗时大部分用于图层求交。FI-Apriori 算法用于相交区域求取的时间为 I-Apriori 算法的 54%左右,整体耗时低于 I-Apriori 的 60%,这表明使用快速求交方法可以有效提高挖掘速度。两种算法生成频繁谓词集时用于图层求交的平均耗时对比见图 3。

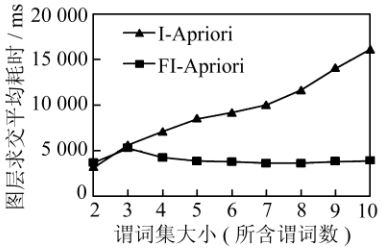


图 3　图层求交平均耗时对比

Fig. 3　Time Consumption Per Layer Intersection

表1 I-Apriori和FI-Apriori算法耗时对比

Tab. 1 Time Consumption of I-Apriori and FI-Apriori

算法	支持度计算耗时/ms				挖掘总耗时/ms
	图层求交	面积计算	其他	总计	
I-Apriori	116 001 587	11 408 722	489	127 410 798	127 477 660
FI-Apriori	63 454 542	12 572 073	7 820	76 034 435	76 104 760

由图3可知,随着谓词数(对应于图层数)的增加,I-Apriori算法求交平均耗时呈线性增长,而FI-Apriori算法则显示出较好的可伸缩性,与前面的分析相符。生成2-谓词集时FI-Apriori稍慢,这是由于快速求交方法需要保存求得的相交区域图层。此外,当存储空间不足以存储某一阶频繁谓词集对应图层时则无法使用FI-Apriori算法,此时I-Apriori算法仍有效。

2.3 关联规则的筛选

本文采用文献[10]中的方法,结合使用客观评价指标提升度和主观评价指标新颖度对所得规则进行了筛选,将其中前后件相关性较低(提升度低)以及与基础知识库不符(新颖度高)的规则剔除,筛选结果中置信度最高的6条规则见表2。由于本文主要讨论关联规则挖掘方法,故不对其评价与筛选问题进行进深入讨论,实际应用中应对此问题加以重视。

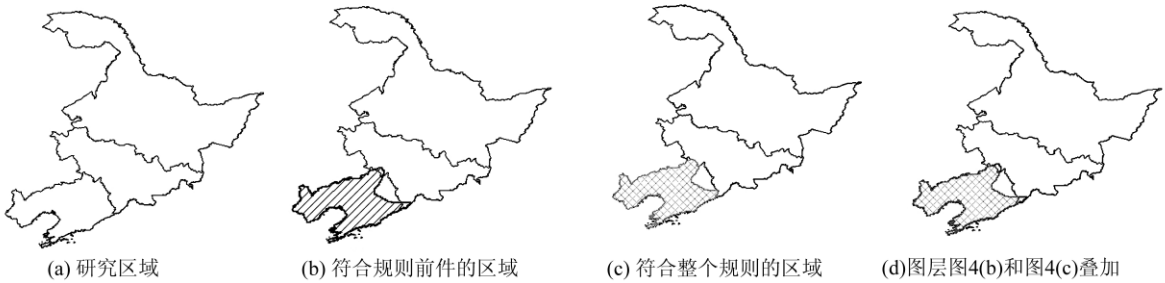
表2 筛选得到的关联规则

Tab. 2 Association Rules After Screening

前件	后件	支持度	置信度
{海拔低,第三季度气温高,第三季度降水中}	{第二季度气温高}	0.167 6	1.000 0
{海拔低,第三季度气温高}	{第二季度气温高}	0.226 2	0.999 7
{海拔低,第三季度气温高,坡度高}	{第二季度气温高}	0.206 9	0.999 7
{海拔低,第四季度NDVI低,第三季度气温高}	{第二季度气温高}	0.169 6	0.999 7
{第二季度气温高,第四季度气温高}	{辽宁}	0.162 1	0.961 7
{第二季度气温高,第三季度气温高,第四季度气温高}	{辽宁}	0.162 1	0.961 7

2.4 关联规则的可视化

采用可视化的方法可以对规则相关的空间信息进行展示,这些信息通常无法以文本形式清晰、直观地表述。本文提出的I-Apriori和FI-Apriori算法会自动生成频繁谓词集到图层的映射,可以方便地实现频繁谓词集和规则的可视化。图4是关联规则{第二季度气温高,第三季度气温高,第四季度气温高}→{辽宁}的可视化效果图。



(a) 研究区域 (b) 符合规则前件的区域 (c) 符合整个规则的区域 (d)图层图4(b)和图4(c)叠加

图4 关联规则可视化效果图

Fig. 4 Example of Rule Visualization

图4直观地展示了规则的含义,即研究区域内第二、三、四季度气温均相对较高的区域(横向对比)主要位于辽宁省境内;图4所用图层均可依据算法中的映射map获取。总之,采用I-Apriori和FI-Apriori算法便于实现空间关联规则的可视化,有利于这些规则的理解和应用。

3 结 语

本文对基于聚类图叠加的关联规则提取方法进行了拓展,提出可以直接从空间矢量图层中提取多谓词关联规则的挖掘算法I-Apriori和FI-Apriori,其中FI-Apriori算法更适用于图层较多的情况。两种算法挖掘出的关联规则都可以进行

进一步的评价和筛选,挖掘过程中生成的图层可用于规则的可视化,有利于实际应用。但是,本文提出的挖掘方法尚不能直接使用矢量多边形图层以外的数据,可使用的空间谓词种类受到一定限制,对该算法进行改进使其支持更多类型的数据是亟需解决的问题。此外,加强对时空数据的支持、实现基于限制条件的挖掘、增量式挖掘以及完善挖掘结果的筛选方法也是下一步的研究重点。

参 考 文 献

[1] Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases[C]. The 4th International Symposium on Large Spatial Databases Maine, 1995

[2] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques(3rd ed)[M]. 北京: 机械工业出版社, 2012

[3] 邓敏, 李光强. 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):1 242-1 245

[4] 马荣华, 蒲英霞, 马小冬. GIS 空间关联模式发现[M]. 北京: 科学出版社, 2007

[5] Morimoto Y. Mining Frequent Neighboring Class Sets in Spatial Databases[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2001

[6] Shekhar S, Huang Y. Discovering Spatial Co-location Patterns: a Summary of Results[C]. The 7th International Symposium on Spatial and Temporal Databases, Redondo Beach, Canada, 2001

[7] Estivill-Castro V, Lee I. Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data[C]. The 6th International Conference of Geocomputation, Brisbane, Australia, 2001

[8] 沙宗尧,李晓雷. 异质环境下的空间关联规则挖掘[J]. 武汉大学学报·信息科学版, 2009, 34(12): 1 480-1 484

[9] Tan P, Steinbach M, Kumar V. 数据挖掘导论[M]. 范明, 范宏建, 译. 北京: 人民邮电出版社, 2006

[10] 牛宵. 时空关联规则不确定性分析及质量评价[D]. 武汉: 武汉大学, 2011

第一作者简介:董林,博士生,主要研究方向为空间数据挖掘。
E-mail:dl@whu.edu.cn

Spatial Association Rule Mining Based on Overlay Analysis and Area Calculation

DONG Lin[1]　*SHU Hong*[1]　*NIU Xiao*[2]

(1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 Shandong Provincial Institute of Land Surveying and Mapping, 9 East Lishan Road, Jinan 250013, China)

Abstract: A spatial association rule mining method is proposed in this study, including support counting method based on overlay analysis and area calculation, and corresponding mining algorithms. Using this method, the extraction of spatial transactions is not needed, and association rules can be mined out directly from layers of polygon features. Besides, the results are easy to be visualized with the help of GIS. An experiment with real-world data shows that this method is valid and efficient, and the result can be further analyzed with screening and visualization methods.

Key words: spatial association rules; overlay analysis; area; support; algorithm

About the first author: DONG Lin, Ph. D candidate, majors in spatial data mining.

E-mail: dl@whu.edu.cn