

数据空间自相关性对关联规则的挖掘与实验分析

陈江平, 黄炳坚

(武汉大学遥感信息工程学院, 武汉 430079)

摘要: 传统的空间关联规则挖掘, 一般是使用属性关联规则的挖掘算法, 对空间数据进行泛化处理, 不考虑空间数据的空间自相关性, 也没有考虑空间自相关与空间关联规则的关系。本文运用改进的 Apriori 算法对某一数据进行空间关联规则挖掘, 并对同一数据进行空间自相关分析, 比较两种方法反映的属性的相关性, 探讨了数据的空间自相关性对空间关联规则挖掘的影响。论文采用 2000 年英国的 HAYFEVE 患病数据集和当时的气温、降雨数据作为实验数据。采用两种方法处理相同的数据集, 即 Apriori 方法和空间自相关方法, 发现二者的结果中所得的一项关联规则和二项关联规则一致, 证明了通过研究数据的空间自相关性也能获得准确的关联规则, 且数据的空间自相关性对关联规则的挖掘存在作用和影响。如何定量度量一元空间自相关对空间关联规则的影响, 以及利用二元空间自相关结果作为空间关联规则候挖掘的候选频繁项集, 进而提高挖掘效率是本文的进一步工作。

关键词: 空间自相关; 关联规则挖掘; 空间数据挖掘; Apriori

DOI: 10.3724/SP.J.1047.2011.00109

1 引言

空间数据挖掘(SDM)^[1]是一个从空间数据中提取出有效的、新颖的、潜在有用的、并能最终被人理解的模式的非平凡过程^[2], 揭示出蕴含在数据背后的客观世界的本质规律、内在联系和发展趋势, 实现知识的自动获取, 从而提供技术决策与经营决策的依据^[3]。空间关联规则挖掘是空间数据挖掘的重要内容, 其目的是发现现实世界中空间对象之间的有趣的关联模式或相互关系^[4]。空间关联规则挖掘是空间数据挖掘的一个重要组成部分。其一般形式是:

$$A_1 \wedge A_2 \wedge \cdots \wedge A_m = B_1 \wedge B_2 \wedge \cdots \wedge B_n$$

谓词 $A_1, A_2, \dots, A_m, B_1, B_2, \dots, B_n$ 是空间和非空间谓词的集合, 其中至少有一个是空间谓词; 令 $A = A_1 \wedge A_2 \wedge \cdots \wedge A_m$, 称为规则的前件; 令 $B = B_1 \wedge B_2 \wedge \cdots \wedge B_n$, 称为规则的后件, $A \wedge B = \emptyset$; $S\%$ 是规则的支持度(support), $C\%$ 表示规则的置信度(confidence)^[5]。

空间关联规则本质上也是地理现象的多个因

子的相互关系与作用的规律。构成地理现象的各种地理因子都不是独立出现的, 它们的关系是复杂的, 可能是相互抑制, 也可能是相互促进的。对地理现象的多因子分析, 有助于我们了解空间因子间的关联模式或者相互关系, 即空间关联规则^[6]。但是, 多因子交互作用识别是一个理论难题, 缺少有效方法。目前有效的方法是王劲峰等提出的建立地理探测器模型, 分析各因子对模型的影响, 能有效识别因子间的相互关系^[6]。探测各因子对模型的贡献率能从庞大的空间数据库中提取有用的空间关联规则。

如何衡量影响地理现象的各因子的关系呢? 定性的有: 地理学第一定律地表所有事物和现象在空间上都是关联的, 距离越近, 关联程度就越强, 距离越远, 关联程度就越弱^[7-8]。定量的衡量方法一般有 Moran's I ^[9-10] 和 Geary's C ^[11] 等方法。Moran's I 和 Geary's C 常用来度量时间序列相邻数值间的相关关系^[12-13]。

空间自相关的存在, 产生的空间差异、空间依赖、空间回归等^[14], 加之属性数据与空间数据不可分性^[15], 使得空间关联规则的挖掘不同于一般的关

收稿日期: 2010-03-16; 修回日期: 2010-10-27。

基金项目: 国家自然科学基金青年科学基金项目(40801152); 教育部留学科研基金项目(213153249)。

作者简介: 陈江平(1975-), 女, 湖北洪湖人, 副教授。研究方向为空间分析, 数据挖掘等。E-mail: chenjp_lisa@163.com

系数据库的关联规则挖掘^[16]。在空间关联规则挖掘中考虑空间自相关,国内外学者提出了很多方法。如引入空间权重矩阵,空间自相关和空间关联的度量函数,并结合空间数据的地理位置构造 Voronoi 图, Delaunay 图,通过直观的方法来发现空间关联规则^[17]。或者将空间信息泛化后转化成属性关系数据库,将空间自相关用数据的形式表达,然后采用属性关联规则的挖掘算法进行挖掘^[18]。再是将空间数据进行分类,分类的依据是类标签属性,即类的唯一标识,继而用决策树等办法将空间数据库的目标以叶子节点的方式置于各类上,根据决策树得到空间关联规则^[19]。

将空间数据泛化为属性数据的过程中,数据之间的空间关系可能会被削弱或者忽略。当数据量很大时,通过图表等直观的方式进行空间关联规则的挖掘显然不现实。采用决策树的方式进行挖掘时,类标签属性的选择决定了类的属性,也就是分类的结果,决策树选择的失误会极大影响关联规则挖掘的结果。本文通过对同一数据集进行两种方法(即空间相关性的分析法和空间关联规则挖掘方法)的实验,将结果进行对比,讨论空间自相关对空间关联规则挖掘的指导作用及影响。

2 研究方法与数据

2.1 方法

传统的关联规则挖掘中,假设数据之间是没有关联的、独立的,在进行关联规则挖掘时,要不停地扫描数据库,得到候选频繁项集,再与最小支持度和最小可信度进行对比,得出频繁项集,最终得到关联规则,这也是经典的 Apriori 算法的思想^[17]。国内外不少学者提出了对 Apriori 算法的改进,使其不需重复对数据库进行扫描,提高算法的运行效率。

在空间关联规则挖掘中,单个变量的空间自相关是不容忽视的因素。当某个变量的自相关系数很大时,说明了该变量存在很强的空间聚集性,即该变量同高或同低的概率高,在空间数据库中表现为该变量相同或者相近的数据出现的次数较多,也就是空间数据库中的记录数出现次数较多。除了单个变量自身的空间自相关外,研究两个变量间的二元自相关,可以得到两个变量之间的空间相关

性,从而知道对主变量影响最大的变量,在进行关联规则挖掘时,相关系数较大的变量间的关系,往往是感兴趣的空间关联规则。按照这个思路,定义一个主变量,通过空间自相关分析,找出对因变量影响大的变量,可得到:

(1)形成感兴趣的空间关联规则;

(2)为关联规则的进一步挖掘提供频繁二项集。

最后可通过扫描数据库,验证所得到的空间关联规则的支持度与置信度。

论文通过两种不同的方式,对同一数据集进行空间关联规则挖掘和空间相关性分析,旨在研究空间自相关在空间关联规则挖掘中的作用,探讨空间关联规则挖掘时的属性变量之空间自相关,以及将二元空间相关的结果作为空间关联规则挖掘的先验知识,以提高传统的空间关联规则挖掘的效率。

(1)使用文献[20]提供的改进的 Apriori 算法对属性数据库进行关联规则挖掘,得出一项、二项、三项、四项频繁项集组成关联规则,用于与自相关分析得到的规则进行对比分析。

(2)使用 GeoDA 软件进行空间自相关分析,包括一元自相关分析和二元自相关分析,可按照各因素,以及各因素之间的相关系数的大小,包括各因素一元自相关的大小、因素间的二元自相关系数的大小。

将改进的 Apriori 算法中的一项、二项关联规则与空间自相关分析中的一元自相关、二元自相关的结果进行对比。观察关联规则中置信度大的频繁一项集与一元自相关系数大的因素之间的关系。通过观察上述各组合的关系,可以得出关联规则与自相关强度的对应关系,换言之,通过单个或者多个因素之间自相关的强度可以作为先验知识判断该因素间是否能构成关联规则,以及关联规则的强度,为空间关联规则挖掘提供背景知识。

具体的技术路线图如图 1 所示。

2.2 实验数据

本文所使用的实验数据是英国 2000 年的花粉热患病人数,以及相关影响因素数据,主要包括:

(1)英国 2000 年各郡的花粉热患病人数数据,授权从网站 <http://www.esds.ac.uk/findingData/snDescription.asp?sn=5558> 下载;

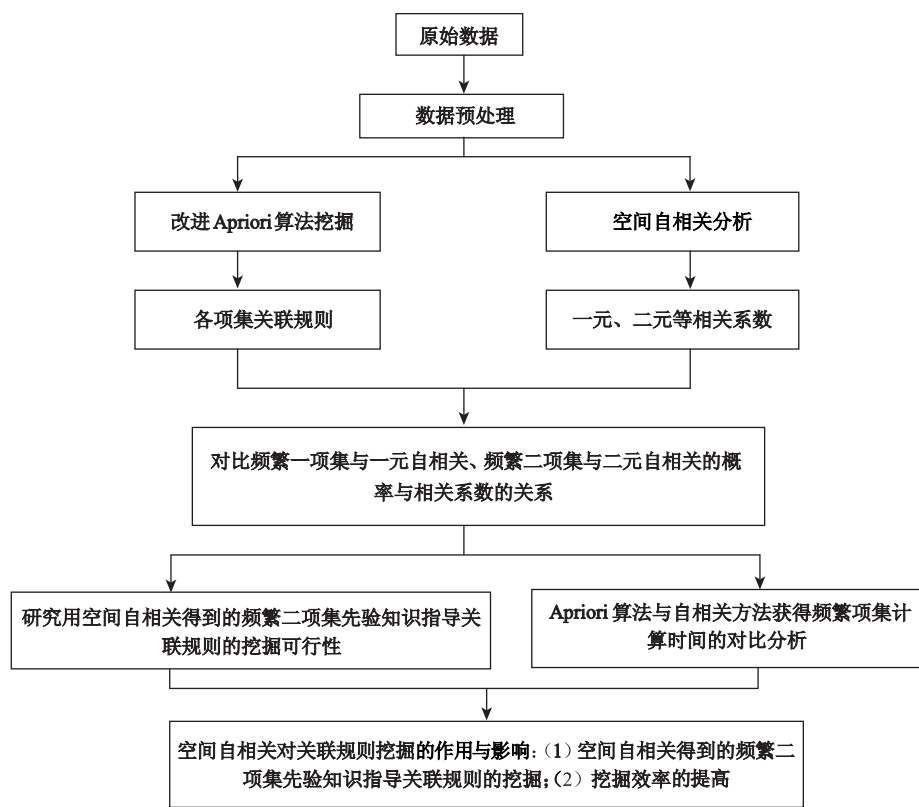


图 1 技术路线图
Fig. 1 Workflow map

(2)英国 2000 年各郡的气温数据(1-12 月), 授权从网站 [http://www. metoffice. org. uk/](http://www.metoffice.org.uk/)下载;

(3)英国 2000 年各郡的降水数据(1-12 月), 授权从网站 [http://www. metoffice. org. uk/](http://www.metoffice.org.uk/)下载;

(4)英国 2000 年各郡的土地/植被类型覆盖率数据,包括耕地、阔叶林、针叶林、改良草地、山地沼泽,以及半天然草地覆盖率数据,授权从网站 [http://192. 171. 153. 213/sections/seo/lcm2000. ht- ml](http://192.171.153.213/sections/seo/lcm2000.html)(Centre for Ecology& Hydrology Uk website) 下载。

实验数据是以 SHP 文件的形式存储的,其存储属性数据的是 DBF 文件。使用 Apriori 算法进行挖掘时,数据源要从 DBF 格式的文件转换为 EXCEL 工作表格式,生成的文件中,每一行代表一个郡,共 100 行;而第 1 列为花粉热发病数目(hayfever),第 2-13 列为 1-12 月的气温数据(tr_jan1-tr_dec1),第 14-25 列为 1-12 月的降水数据(tr_jan3-tr_dec3),第 26-31 列为 6 中植被类型覆盖率的数据(landuse1-landuse6),即耕地、阔叶林、针叶林、改良草地、山地沼泽和半天然草地,第 32-36

列为所属地区数据,分别是东北地区、东南地区、西北地区、西南地区和中部地区的数据。为了数据更有统计意义,对数据进行了对数处理,使其满足正态分布。对空间自相关分析时,直接使用经对数处理后的 SHP 文件进行分析。

本文对 Apriori 挖掘部分所使用的工具,是作者编写的程序,空间分析部分所使用的分析软件是 GeoDA 095i。

3 空间关联规则挖掘的实验与分析

对原数据进行对数处理后,根据每一列的平均数,将数据进行二值化处理。经过以上处理后,运行 Apriori 程序,对上述处理结果进行关联规则挖掘,得到频繁一、二、三、四、五项集。

3.1 Apriori 改进算法的关联规则挖掘

根据参考文献[20]中改进的 Apriori 算法进行关联规则挖掘,运行改进的 Apriori 算法,其部分挖掘结果如表 1 所示。

表 1 改进的 Apriori 算法挖掘结果

Tab. 1 Mining results of improved Apriori algorithm				
Item1	Item2	Probability/ Item3	Probability/ Item4	Probability
hayfever	landuse2	0.33473115		
hayfever	Landuse1	0.33473115		
hayfever	tr_aug3	0.330945822		
hayfever	tr_jun1	0.327160494		
hayfever	tr_feb3	0.323375166		
hayfever	tr_nov3	middle	0.396299	
hayfever	tr_mar1	south_west	0.395941	
hayfever	tr_apr1	south_west	0.395941	
hayfever	Landuse4	south_west	0.395941	
hayfever	Landuse5	south_west	0.395941	
hayfever	tr_may1	south_east	0.394595	
hayfever	Landuse6	middle	0.394372	
hayfever	landuse2	south_west	0.393987	
hayfever	Landuse1	south_west	0.393987	
hayfever	middle	landuse2	south_east	0.355856
hayfever	middle	Landuse1	south_east	0.355856
hayfever	tr_feb1	tr_jul3	south_east	0.355389
hayfever	tr_feb1	tr_oct1	south_east	0.355389
hayfever	tr_feb1	tr_oct3	south_east	0.355389
hayfever	tr_dec1	tr_jul3	south_east	0.355389

表 1 中,每一行最后的小数即为变量在数据库中同时出现的概率,如 HAYFEVER 与中部(middle)同时出现的概率为 0.396299。

3.2 空间自相关的空间关联规则挖掘

使用 GeoDA 对图层进行空间自相关分析。为了使数据与 Apriori 算法挖掘的数据一致,这里也对 DBF 文件的各列数据进行对数处理。本文所使用的空间自相关计算方法是 Moran’s I 统计值,包括一元自相关 Univariate Moran 和二元自相关分析 Multivariate Moran。

使用 Univariate Moran 进行各变量的自相关分析,其相关系数结果如表 2 所示:

从表 2 可以看出,除了 3、4、11 月,其他各月气温的空间自相关性都比较高,除了 2、7、10、11 月,其他各月的降水的空间自相关性也都比较高,除了草地以外,其他各土地利用类型的空间自相关性也都比较高。事先知道这些属性的空间自相关性较高,在进行空间关联规则挖掘时,在计算其可信度时应该有所考虑。

表 2 一元自相关分析结果

Tab. 2 Results of univariate Moran analysis							
气温(月)	自相关系数	降水(月)	自相关系数	植被(类型)	自相关系数	HAYFEVER	自相关系数
1	0.4228	1	0.5070	耕地	0.2930		0.0568
2	0.7077	2	0.1137	阔叶林	0.2254		
3	0.0163	3	0.5061	针叶林	0.5461		
4	0.1180	4	0.2364	改良草地	0.0823		
5	0.5335	5	0.5911	山地沼泽	0.4693		
6	0.6247	6	0.6729	半天然草地	0.0777		
7	0.5978	7	0.1597				
8	0.5064	8	0.4938				
9	0.5328	9	0.5393				
10	0.5505	10	0.0741				
11	0.3931	11	0.0586				
12	0.6407	12	0.3086				

变量的一元自相关从地物分布的角度来看,表达的是地物的某一属性的聚集程度,将变量的自相关系数表现在地图上可以直观地观察到地物的聚集。另一方面,从地统计学的角度看,变量的一元自相关则表达了相邻地物的相似程度,相关系数越大,则表明相似程度大,反之则相似程度小。更进

一步,对于相似程度大的地物属性,它们在数据库中属性区别小,则进行关联规则挖掘时,某一属性出现的频率高,从而可根据自相关系数大小提取频繁一项集,而无需扫描数据库。如 2 月的气温自相关系数较大(0.7077),其在 Apriori 挖掘中的概率也较大(0.495),3 月的气温自相关系数较小

(0.0163), 其在 Apriori 挖掘中的概率也较小 (0.464)。在进行关联规则挖掘的过程中, 可以通过获得一元相关系数的大小而获得挖掘的先验知识。

使用 Multivariate Moran 进行二元自相关分析, 其中主变量是花粉热患病人数 HAYFEVER, 所有其他变量与 HAYFEVER 进行二元自相关分析, 分析结果如表 3。

表 3 HAYFEVER 与气温、降水、植被类型的二元自相关系数
Tab. 3 Results of multivariate Moran analysis(HEYFEVER with other variables)

气温(月)	二元自相关系数	降水(月)	二元自相关系数	植被(类型)	二元自相关系数
1	0.1185	1	−0.1305	耕地	0.0688
2	0.1102	2	−0.0188	阔叶林	0.0322
3	0.0318	3	−0.1288	针叶林	−0.0410
4	0.0666	4	0.1044	改良草地	0.0681
5	0.1486	5	0.1475	山地沼泽	−0.0190
6	0.1215	6	−0.1975	半天然草地	0.0665
7	0.1436	7	−0.0529		
8	0.1529	8	−0.1303		
9	0.1803	9	−0.1051		
10	0.1407	10	−0.0323		
11	0.1285	11	−0.1249		
12	0.1569	12	−0.7810		

从表 3 可以看出, HAYFEVER 发病率与 3 月、4 月的气温相关性小于与其他各月的相关性; 只与 4 月、5 月的降水呈正相关, 与其他各月的降水呈负相关, 与耕地和草地的相关性大于与针叶林的相关性。

在二元自相关分析过程中, 可以观察到 HAYFEVER 发病率与各发病因素的相关关系, 且相关大小一目了然。相关程度大的两变量在地统计上表现为两变量同时发生的概率高, 而在空间数据库记录中, 则表现为二者同时出现的概率高。因此, 在进行关联规则的挖掘时, 则可根据两变量的相关系数的大小判断二者能否成为一条关联规则, 为挖掘过程提供方便和先验条件。如 HAYFEVER 发病率与 5 月气温相关性较大(0.1486), 其与 5 月气温在数据库中出现的概率也高(0.357)。

3.3 实验结果对比分析

本文的实验是对同一数据集采用的两种不同的数据分析方法, 理论上两种分析方法得到的分析结果, 即知识、规则应该是一致的, 但从实际的实验结果可以看出还是有一些不同。

3.3.1 改进的 Apriori 算法实验结果分析

从改进的 Apriori 算法中, 可以提取以下规则及其概率:

(1) 三项集

①5 月份的气温高 \wedge 降水少 \Rightarrow 发病率高, 概率 0.41

②4 月份的气温高 \wedge 降水少 \Rightarrow 发病率高, 概率 0.40

③6 月份的气温高 \wedge 降水少 \Rightarrow 发病率高, 概率 0.38

④7 月份的气温高 \wedge 降水少 \Rightarrow 发病率高, 概率 0.43

⑤阔叶林多 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.39

⑥4 月气温高 \wedge 阔叶林多 \Rightarrow 发病率高, 概率 0.40

⑦7 月气温高 \wedge 阔叶林多 \Rightarrow 发病率高, 概率 0.40

⑧6 月气温高 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.39

⑨7 月气温高 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.39

⑩8 月气温高 \wedge 耕地多 \Rightarrow 发病率高, 概率

0.39

⑪5、6、7 月气温高 \wedge 西南部地区 \Rightarrow 发病率高, 概率 0.38~0.39

⑫5、6、7 月气温高 \wedge 中部地区 \Rightarrow 发病率高, 概率 0.38~0.39

⑬4 月降水少 \wedge 西南、东南部地区 \Rightarrow 发病率高, 概率 0.30~0.34

⑭阔叶林、改良草地多 \wedge 西南、中部地区 \Rightarrow 发病率高, 概率 0.38

⑮2 月气温低 \wedge 中部地区 \Rightarrow 发病率低, 概率 0.33

⑯1 月气温低 \wedge 1 月降水多 \Rightarrow 发病率低, 概率 0.33

⑰1 月气温低 \wedge 西南部地区 \Rightarrow 发病率低, 概率 0.32

⑱11 月气温低 \wedge 中部地区 \Rightarrow 发病率低, 概率 0.39

(2) 四项集

①5 月份的气温高 \wedge 降水少 \wedge 东南部地区 \Rightarrow 发病率高, 概率 0.46

②6 月份的气温高 \wedge 降水少 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.45

③7 月份的气温高 \wedge 降水少 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.38

④8 月份的气温高 \wedge 降水少 \wedge 耕地多 \Rightarrow 发病率高, 概率 0.42

⑤8 月份的气温高 \wedge 阔叶林多 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.41

⑥4 月份的气温高 \wedge 阔叶林多 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.40

⑦4 月份的降水少 \wedge 阔叶林多 \wedge 改良草地多 \Rightarrow 发病率高, 概率 0.40

⑧5、6、7 月气温高 \wedge 降水少 \wedge 西南部地区 \Rightarrow 发病率高, 概率 0.40~0.44

⑨5、6、7 月气温高 \wedge 阔叶林多 \wedge 西南部地区 \Rightarrow 发病率高, 概率 0.40~0.42

⑩5、6、7 月气温高 \wedge 改良草地多 \wedge 西南部地区 \Rightarrow 发病率高, 概率 0.39~0.41

⑪9 月气温高 \wedge 降水少 \wedge 中部地区 \Rightarrow 发病率高, 概率 0.30

⑫9 月降水少 \wedge 东南部地区 \wedge 阔叶林多 \Rightarrow 发病率高, 概率 0.30

3.3.2 空间自相关的属性之间相关性结果分析

对同一数据集采用空间数据分析的方法, 得到的结果如下:

(1) 一元自相关的结果

由一元自相关的结果可以看出, 各月的气温和降水的自相关系数较大, 各植被类型的自相关系数较大。也就是说, 该区域内各月的气温、降水、植被的类型都有很强的空间聚集性, 这说明了气温、降水及植被等分布都比较集中。

从 Apriori 挖掘和一元自相关的结果知道, 相关系数大的变量说明聚类中心显著且邻域属性值差别小, 在数据库中某一段属性值出现的频率就高。该变量就可以作为频繁一项集, 可以为关联规则的挖掘提供先验条件。实验中, 一元自相关的分析结果与改进 Apriori 运算所得的频繁一项集结果存在对应的关系, 即相关系数与出现的概率存在对应的关系, 相关系数大的变量在数据库中出现的概率也大。

(2) 二元自相关的结果

从相关系数的正负性来看, HAYFEVER 与各月的气温相关系数全部为正, 这说明了 HAYFEVER 与气温是正相关的, 气温高时发病率也高, 气温低时发病率也低。HAYFEVER 与各月的降水基本上是呈负相关的, 只有两个月(4 和 5)是成正相关的, 也就是说, HAYFEVER 与降水是负相关的, 降水多时发病率就低, 降水少时发病率就高。对于植被类型来说, 耕地、阔叶林、改良草地, 以及半天然草地与 HAYFEVER 的相关系数为正, 说明该植被类型越多, 发病率越大。而针叶林和山地沼泽与 HAYFEVER 的相关系数为负, 说明当针叶林和山地沼泽面积大时, 发病率低。

从 Apriori 挖掘结果的频繁二项集与二元自相关的对比发现, 凡相关系数大的两个变量, 它们在数据库中出现的概率也高, 如前文所说的 HAYFEVER 发病率与 5 月气温相关性较大(0.1486), 其与 5 月气温在数据库中出现的概率也高(0.357), 二者可以自成一条关联规则。同时, 也可以为关联规则的挖掘提供频繁二项集, 即先验知识, 如可直接将 HAYFEVER 患病率与气温、降水作为候选的频繁二项集, 减少在关联规则挖掘时的数据库扫描次数。

3.3.3 数据的空间自相关性对关联规则挖掘的影响分析

(1) 由上述的实验分析可知, 无论是一元自相关结果还是二元自相关结果, 都与改进 Apriori 挖

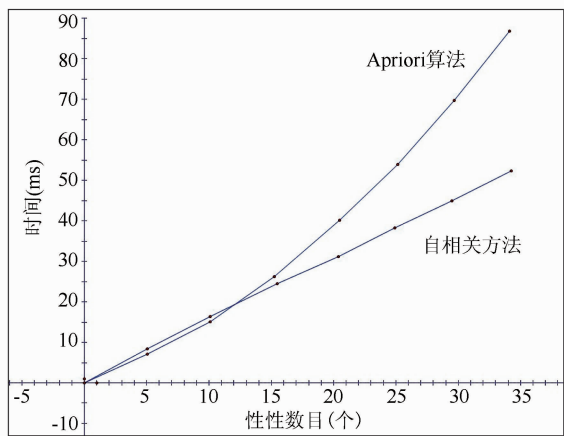


图2 两种方法挖掘的时间对比

Fig. 2 Comparison of efficiency for the two methods

掘的一项集和二项集比较一致,而改进 Apriori 运算的结果,则可以通过扫描数据库得到验证,这证明了数据的空间自相关性确实对空间关联规则的挖掘有很大的作用。空间自相关分析的结果,一方面为关联规则的挖掘提供频繁项集,即先验知识;另一方面,相关性强的变量本身就是很强的关联规则。如二元自相关中可以看出,HAYFEVER 与 6、7、8、9 月气温的相关系数较大,二者有显著的相关性,这意味着它们在数据库中同时出现的概率也大,即它们的组合是关联规则的频繁二项集,可认为是关联规则;HAYFEVER 与降水是成负相关的,即降水多一般与发病率低同时发生,降水少一般与发病率高同时发生,同理,它们的组合是关联规则的频繁二项集,可以认为是关联规则。

(2)在运行效率上看,面对海量的地理数据库,使用空间自相关的方法也有很大的优势,它无需进行笛卡尔积的运算,使用现有的分析软件即可简单快捷地计算相关系数。当属性的数目较大时,笛卡尔积的组合随属性数目的增长是呈非线性的增长的,它的计算会极大地降低挖掘的速度。

(3)从得到的规则上看,Apriori 算法进行挖掘可得到三元关联规则,四元关联规则,以及多元关联规则,但是对于空间自相关分析而言,海量数据的多元相关性分析还缺乏有力、有效的工具软件,所以,目前还只能为空间关联规则提供一、二元关联规则的先验知识。

4 结论

(1)空间自相关的分析结果与改进 Aprori 算

法关联规则挖掘的结果一致。空间自相关的分析结果中,与发病率相关性大的因素在改进 Aprori 算法挖掘中出现的频率也高,这说明了空间自相关性大的因素可以部分转换成空间关联规则,为关联规则的挖掘候选提供频繁项集,即研究空间自相关对研究空间关联规则挖掘有指导性的作用。

(2)空间自相关的分析方法在效率上优于改进的 Apriori 算法。计算自相关系数的时间比计算笛卡尔积的时间短,且随着属性数目的增加,计算相关系数的时间优势尤为明显。

本文研究证实了空间自相关与空间关联规则两种方法分析的结果比较一致。如果在空间关联规则的挖掘中考虑空间自相关,将会使空间关联规则挖掘更有意义,规则更好地反映事实,从而提高其挖掘效率。但是挖掘的精度问题还没有进行分析,如何选取自相关系数的阈值,才能包含所有的强关联规则,中等强度关联规则或弱关联规则等,并且如何判断通过自相关计算而来的关联规则的可信度问题,将是本文未来的研究方向。另外,目前空间数据分析对多元属性的相关性分析还缺乏有力的方法,如何利用空间相关性分析对多属性关联规则的挖掘提供先验知识也是一个未来的研究方向。

致谢:感谢英国生态与水文中心的 Diane Unwin 女士和英国数据档案中心的 Beate Lichtwardtand 女士,为本论文提供了英国 2000 年的植被覆盖数据和花粉热病情数据,使本实验得以顺利进行。特别感谢 Haining 教授,为本研究提供了宝贵的意见。

参考文献:

- [1] 李德仁,王树良,李德毅,王新洲. 论空间数据挖掘和知识发现的理论与方法[J]. 武汉大学学报(信息科学版), 2002(3):222-233.
- [2] Fayyad U M, Piatetsky Shapiro G, Smyth P. Advances in Knowledge Discovery and Data Mining[C]. London : AAAI/MIT Press, 1996.
- [3] 李德仁,王树良,李德毅. 空间数据挖掘理论与应用[M]. 北京:科学出版社,2006.
- [4] 张建峰,王泳,王剑. 关联规则在空间数据挖掘中的应用及实现[J]. 计算机技术与发展,2007, 17(8): 208-211.
- [5] 黄旭峰,邹菁. 空间数据挖掘中关联规则的研究与实现[J]. 科技信息,2009(7):481-482.
- [6] Wang J-F, Li X-H, Christakos G, Liao Y-L, Zhang T, Gu X & Zheng X-Y. Geographical Detectors-based

- Health Risk Assessment and Its Application in the Neural Tube Defects Study of the Heshun Region, China [J]. International Journal of Geographical Information Science, 2010, 24(1): 107 - 127.
- [7] Tobler W. A Computer Movie Simulating Urban Growth in the Detroit Regional Economic Geography [J]. Economic geography, 1970, 46 (2): 234 - 2401.
- [8] Tobler W. On the First Law of Geography: A Reply [J]. Annals of the Association of American Geographers, 2004, 94(2): 304 - 310.
- [9] Moran PAP. The Interpretation of Statistical Maps[J]. Journal of the Royal Statistical Society B, 1948(10): 243 - 251.
- [10] Moran PAP. Notes on Continuous Stochastic Phenomenal[J]. Biometrika, 1950, 37: 17 - 33.
- [11] Geary R C. The Contiguity Ratio and Statistical Mapping[J]. The Incorporated Statistician, 1954, 5: 115 - 145.
- [12] 王永, 沈毅. 空间自相关方法及其主要应用现状[J]. 中国卫生统计, 2008, 25(4): 443 - 445.
- [13] 陈彦光. 基于 Moran 统计量的空间自相关理论发展和方法改进[J]. 地理研究, 2009, 28(6): 1449 - 1463.
- [14] 王劲峰. 地图的定性和定量分析[J]. 地球信息科学学报, 2009, 11(2): 169 - 175.
- [15] 廖顺宝, 张赛. 属性数据空间化误差评价指标体系研究[J]. 地球信息科学学报, 2009, 11(4): 176 - 182.
- [16] Chen Jiangping, Tan Xiaojin. Mining Spatial Association Rules with Geostatistics[C]. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 2008.
- [17] 何彬彬, 郭达志, 方涛. 基于空间统计学的空间关联挖掘[J]. 计算机工程, 2006, 32(5): 20 - 22.
- [18] Han Jiawei. Mining Knowledge at Multiple Concept Levels[C]. Proceedings of the Fourth International Conference on Information and Knowledge Management, 1995.
- [19] Krzysztof Koperski, Junas Adhikary, Han Jiawei. Spatial Data Mining: Progress and Challenges Survey Paper[J]. SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96), Montreal, Canada, June 1996.
- [20] 陈江平, 付仲良, 徐志红. 一种 Apriori 的改进算法[J]. 武汉大学学报(信息科学版), 2003, 28(1): 94 - 99.

Application and Effects of Data Spatial Autocorrelation on Association Rule Mining

CHEN Jiangping, HUANG Bingjian

(School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China)

Abstract: Spatial autocorrelation is a very general statistical property of spatial variables, it indicates correlation of a variable with itself through space. Spatial association rule mining, discovery of interesting, meaningful rules in spatial databases, ignores autocorrelation of spatial data, or just generalizes the spatial data into attribute data currently. In most of the ways on spatial association rules mining, they transferred the spatial relations into non-spatial relations by virtue of spatial analysis. This means the separation of spatial autocorrelation from spatial association rule mining. In order to study the relations between spatial autocorrelation and spatial association rule mining, in this paper, the spatial association rules were mined by developed Apriori algorithm. Then, spatial autocorrelation analysis was implemented in the same spatial data set. A basic assumption of many spatial association rules mining is lacking for a priori information about spatial attributes. The two dimensional spatial autocorrelation results were used as priori knowledge in spatial association rules mining in this paper. The experimental data is about the amount of the hay fever (disease caused by pollen allergic rhinitis) patients and its factors, including temperature, precipitation and vegetation types of each county in the United Kingdom in 2000. The obtained frequent itemsets and the spatial association rules prove that factors have stronger correlation with hay fever (correlation coefficient is larger) appear with hay fever simultaneously more frequently in the spatial database, which con-

firm the existence of the effects that spatial autocorrelation has on spatial association rule mining. The analysis results not only point out the relation between spatial autocorrelation and spatial association rule mining, but also provide priori knowledge in the process of spatial association rule mining, making the mining process more targeted. Besides, without calculating the Cartesian in developed Apriori algorithm, spatial autocorrelation analysis can get the correlation coefficients efficiently, making the mining process more effectively. Further work would focus on how to evaluate the effects of the spatial autocorrelation on spatial association rules mining, how to find out the candidate frequent spatial itemsets from the results of spatial autocorrelation analysis in practical application.

Key words: spatial autocorrelation; association rule mining; spatial data mining; Apriori