空间关联规则的双向挖掘*)

王佐成[1,2] 汪林林[2] 薛丽霞[1,2] 李永树[1]

(西南交通大学土木工程学院 成都 610031)[1] (重庆邮电学院软件学院 重庆 400065)[2]

摘 要 空间数据库中关联规则挖掘不仅需要考虑关系元组属性之间的关系——纵向关系,更需要挖掘元组之间的关系——横向关系,如相邻、相交、重叠等。本文通过分析空间数据库的存储模式,借鉴事务数据库关联规则的挖掘方法,对空间关联规则进行完整定义,并对规则的兴趣度度量进行探讨。根据挖掘的方向将空间数据挖掘归纳为纵向挖掘、横向挖掘、双向挖掘。在双向挖掘中,提出一种新算法,该算法根据挖掘任务进行约束,缩小挖掘空间,然后通过空间计算将空间关系转化为非空间关系,经过多次循环,获取非空间项集,进而挖掘出空间关联规则。据此提出空间数据双向挖掘工作流程,并通过实例进行了验证。

关键词 数据挖掘,空间数据,关联规则,双向挖掘

Mining Spatial Association Rules in Two-direction

WANG Zuo-Cheng[1,2] WANG Lin-Lin[2] XUE Li-Xia[1,2] LI Yong-Shu[1]

(Civil Engineering College, Southwest Jiaotong University, Chengdu 610031)[1]

(Software Institute, Chongqing University of Post and Telecommunication, Chongqing 400065)[2]

Abstract Spatial data mining is different from data mining in transaction DB. In most case, the relationships between the tuples in transaction DB do not be taken into account. But in spatial database, there are relationships not only between the attributes, but also between the tuples, and most of the associations exists between the tuples—objects, such as adjacent, intersection, overlap and other topological relationships. So the tasks of spatial data association rules mining include not only mining the relationships between attributes of spatial objects, which we call vertical direction DM, but also mining the relationships between the tuples, which we call horizontal direction DM. This paper analyses the storage models of spatial data, uses for reference the technologies of data mining in transaction DB, defines spatial association rules, including vertical direction association rule, horizontal direction association rule and two-direction association rule, discusses the measurements of interestingness of spatial association rules, and propose the work flows of spatial association rules data mining. During two-direction spatial association rules mining, we propose an algorithm to get non-spatial itemsets. By spatial analysis, we can transfer the spatial relations into non-spatial associations and get non-spatial itemsets. Based on the non-spatial itemsets, we can make use of Apriori algorithm or other algorithms to get the frequent itemsets and then, spatial association rules come into being. To confirm that, we mine in the land using spatial DB to get spatial association rules to validate the algorithm. The test results show that the algorithm is efficient and can mine the interesting spatial rules.

Keywords Data mining, Spatial data, Association rule, Vertical and horizontal direction

空间数据挖掘和知识发现(SDMKD, Spatial Data Mining and Knowledge Discovery)是数据挖掘和知识发现(DMKD)的分支学科。但SDMKD不同于普通的DMKD,它的对象是空间数据库或空间数据仓库,有别于常规的事务型数据库。空间信息在空间数据库中根据特定的主题划分为主题层[1],一个主题层和关系模式中的关系类似,有一个模式和实例,一个主题层是一类空间对象的集合。空间对象与现实世界的实体相对应,具有两个方面的属性:一是非空间描述属性;二是空间属性。因此,空间对象之间的关联关系不仅包括非空间属性之间的关系,还包括空间属性之间的关系,而且后者的表现更加复杂,也是目前空间数据关联规则挖掘主要研究方面[2,3]。要对空间数据关联规则进行有效的挖掘,一个重要的问题就是根据空间数据的特点,对空间数据关联规则进行定义,然后才能根据规则定义进行有效挖掘。目前较多的研究是针对空间关联的复杂性,研究空间关联的不确定性,并据此提出了一些算法[4~8]。虽然也对空间数据关联进行了归类整理[8,9],但是根据复杂的空间关联对空间关联规则系统的定义尚未形成,不能为空间数据关联规则挖掘提供有效的依据。

在空间数据库关联规则挖掘中,我们必须研究的关联集中在3个方面:1)每一个主题层(关系模式)中的同类对象属性之间的关联;2)一个主题层中同类对象之间的关联;3)不同

*)重庆市自然科学基金资助项目(No. 2005BB2065)。王佐成 博士生,讲师,主要研究领域为空间数据库、数据挖掘;汪林林 硕士,教授,主要研究领域为空间数据库、数据结构、数据仓库与数据挖掘;薛丽霞 博士生,讲师,主要研究领域为空间数据库、数据挖掘;李永树 博士,教授,主要研究领域为空间数据结构、测量工程。

主题层之间的不同对象间的关联。其中，后两者常常作为研究的重点。在本文中，我们将第一种关联的挖掘称为纵向挖掘，后两种关联的挖掘称为横向挖掘，对3种关系同时进行挖掘称为空间数据双向挖掘。

1 空间关联规则的描述

关联规则挖掘是由R. Agrawal等人提出的，目的是要在事务数据库中发现各项目之间的关系并形成规则[10]。与此相对应，在空间数据库中存在大量的关联关系，使得空间数据关联规则挖掘成为空间数据挖掘领域的一个重要研究方向[11]。

1.1 一般关联规则的定义

为完整性起见，给出关联规则的一般形式化描述：$I=\{i_1,i_2,\cdots,i_m\}$是m个不同交易项目的集合。D为交易T的集合，$T\subseteq I$。每一个T对应有唯一的标识符TID，设$A\subseteq I$，若$B\subseteq T$，则称交易T包含A。关联规则形如$A\rightarrow B$的蕴涵式，这里$A\subset I$，$B\subset I$，且$A\cap B=\Phi$。规则$A\rightarrow B$由支持度s(support)和置信度c(confidence)约束。置信度表示规则的有效性或“值得信赖性”的确定性度量，支持度表示规则的潜在有用性。挖掘关联规则就是产生那些支持度和置信度分别大于用户给定的最小支持度和最小置信度的规则[10]。

1.2 空间关联关系的定义

空间关联规则是建立在空间数据间关联关系的基础上，下面对空间数据间的关联关系进行简单描述。空间关联关系有3种基本的二元空间关联关系：拓扑关系、距离关系和方向关系[11]。

1)*拓扑关系* 拓扑关系可以用4－intersection模型来表达[12]，该模型可以扩展到高维空间[13]。针对二维平面上的点集A和B，它们的拓扑关系也可以用9-intersection(I9)来表达，可以通过3×3矩阵来表示。在二维平面上，没有岛和洞的空间对象可以用8种关系来表达：disjoint，contain，inside，equal，meet，cover，coveredby以及overlap。这8种关系同I9的矩阵一起完整地表达了拓扑关系。

2)*距离关系* 地理空间中两点间的距离度量可以沿实际的地球表面进行，也可以沿着地球椭球体的距离量算。两对象距离可以表现为以下几种形式：大地测量距离、曼哈顿距离、时间距离、交通路径距离等。

3)*方位关系* 方位关系定义了地物对象之间的方位。在二维平面上，给定定位坐标系，点目标之间的方位关系可以定义为8种基本关系和4种扩展关系[14]。在实际的研究工作中，由于方向的不确定性和模糊性，可以引入模糊集理论来更好地表达方位关系，将8种基本关系作为语言值，求方位对其隶属度。

对于线目标和面目标，其方位关系与点目标相同，判别可以通过计算MBR(Minimum Bounding Rectangle)的方位关系进行[3]，或者对线和面目标进行抽象，这需要从目标的性质和需求考虑。

1.3 空间关联关系在空间数据库中的表现

空间数据之间的拓扑、距离和方向关系在空间数据库中表现为空间对象之间的关系，即元组之间的横向关系。考察空间数据库的关系模式如表1。

表1 空间数据库的关系模式

Apartment_house table

Name	Location	Price	Area	House_num	…
Aigleng	Str12-13	3000	200	120	
Bulic	Str11-2	2300	155	200	…
…	…	…	…	…	

School table

Name	Chairman	Population	Location	…
Wanli	John	200	Str1-23	
herla	Mark	350	Str17-2	…
…	…	…	…	

从关系模式上看，空间关联关系存在于横向元组(空间对象)之间，如表Apartment_house中Aigleng与Bulic公寓之间的拓扑关系、距离关系和方向关系。除此之外，关系表与关系表之间也存在空间关联关系，如表Apartment_house中Aigleng公寓与School关系中的Wanli学校的关联关系。

除了在横向元组(空间对象)之间存在大量的空间关联关系之外，在元组的属性方向(纵向)也存在空间和非空间的关联关系。如表Apartment_house中的Location与Price之间的空间关联关系和Area与House_num之间的非空间关联关系。

1.4 空间关联规则的定义

在对空间关联关系进行理解后，给出空间关联规则相关定义。

定义1 空间。$S=\{s_1,s_2,\cdots,s_m\}$是m个空间对象集，D_i($i=1,2,3\cdots$)为空间对象集按照空间概念模型生成的空间对象子集，$D_i\subseteq S$，$O_i\in D_i$，O_i代表一个空间对象，D_i构成子空间，S构成完整空间。

定义2 纵向关联规则。设$I=\{I_1,I_2,\cdots,I_n\}$是项的集合，$A\subset I$，$B\subset I$，$A\cap B\neq\Phi$，R_A和R_B分别是A和B的值域，$R_{A_i}\in R_A$，$R_{B_i}\in R_B$，则$A(R_{A_i})\rightarrow B(R_{B_i})(s\%,c\%)$为纵向关联规则。规则在子空间$D_i$中成立，支持度是子空间$D_i$中包含$A(R_{A_i})$和$B(R_{B_i})$的空间对象数与子空间$D_i$中的对象数之比，记为$support(A(R_{A_i})\rightarrow B(R_{B_i}))$，即$support(A(RAi)\rightarrow B(R_{B_i}))=|\{O;A(R_{A_i})\cup B(R_{B_i})\subseteq O,O\in D_i\}|/|D_i|$。置信度是指包含$A(R_{A_i})$和$B(R_{B_i})$的空间对象数与包含$A(R_{A_i})$的对象数之比，记为$confidence(A(R_{A_i})\rightarrow B(R_{B_i}))=|\{O;A(R_{A_i})\cup B(R_{B_i})\subseteq O,O\in D_i\}/|O;A(R_{A_i})\subseteq O,O\in D_i|$。

定义3 横向关联规则。设$X\subseteq D_i$，$Y\subseteq D_j$($j=1,2,3\cdots$)。其中，如果$i=j$，则为某一个关系模式内操作；如果$i\neq j$，则为多关系模式操作。则$P(X,x)\rightarrow Q(Y,y)(s\%,c\%)$为横向关联规则。规则在子空间$D_i$和$D_j$中成立。支持度是子空间$D_i$和$D_j$中包含满足$P$和$Q$谓词的$X$和$Y$的空间对象数与子空间$D_i$和$D_j$的对象总数之比，记为$support(X\rightarrow Y)=|\{O;X\cup Y\subseteq O,O\in(P(D_i\times D_j)\cap Q(D_i\times D_j))\}|/|D_i\cup D_j|$。置信度是子空间$D_i$和$D_j$中包含满足$P$和$Q$谓词的$X$和$Y$的空间对象数与子空间$D_i$的对象总数之比，记为$confidence(X\rightarrow Y)=|\{O;X\cup Y\subseteq O,O\in(P(D_i\times D_j)\cap Q(D_i\times D_j))\}|/|D_i|$。

定义4 双向关联规则。双向关联规则(包含对象属性的关联规则)设$I=\{I_1,I_2,\cdots,I_n\}$是项的集合，$X\subseteq D_i$，$Y\subseteq D_j$，A是D_i的项目子集，$A\subset I$是D_i的项目集，R_A是A的值

域，$R_{A_i} \in R_A$。则 $P(X,x) \wedge Q(Y,y) \rightarrow A(R_{A_i})(s\%,c\%)$ 为双向关联规则。规则在子空间 D_i 和 D_j 中成立。支持度是子空间 D_i 和 D_j 中包含满足 P 和 Q 谓词及 $A(R_{A_i})$ 的 X 和 Y 的空间对象数与子空间 D_i 和 D_j 的对象总数之比，记为 $support(X \rightarrow Y) = |\{O: X \cup Y \cup A(R_{A_i}) \subseteq O, O \in (P(D_i \times D_j) \cap Q(D_i \times D_j))\}| / |D_i \cup D_j|$。置信度是子空间 D_i 和 D_j 中包含满足 P 和 Q 谓词及 $A(R_{A_i})$ 的 X 和 Y 的空间对象数与子空间 Di 的对象总数之比，记为 $confidence(X \rightarrow Y) = |\{O: X \cup Y \cup A(R_{A_i}) \subseteq O, O \in (P(D_i \times D_j) \cap Q(D_i \times D_j))\}| / |D_i|$。

2 空间关联规则挖掘

空间数据库关联规则挖掘与事务数据库关联规则挖掘有很大的区别。以购物篮数据挖掘为例，由于交易不同，导致元组之间的属性项个数不同，这为 Apriori 算法实施提供了假设前提。而空间数据库中对象的属性除非是数据缺失，一般是完整的，这就使得 Apriori 算法在空间数据库中的应用部分假设条件不满足，在具体挖掘频繁项集中，其使用方法也有很大的变化。

2.1 空间关联规则的纵向挖掘

空间关联规则的纵向挖掘主要是对空间对象属性以及属性组之间进行关联规则挖掘，得到纵向关联规则。空间对象的属性可分为空间属性和非空间属性，空间属性主要是描述空间对象的位置编码相关，非空间属性描述空间对象的性质相关。对于纵向挖掘，可以在对空间属性进行数据预处理后，直接采用数据库查询语言并结合 Apriori 算法直接得到强关联规则。数据预处理方法包括基于属性的归并、基于云理论的方法等[15]。工作方式如图 1。

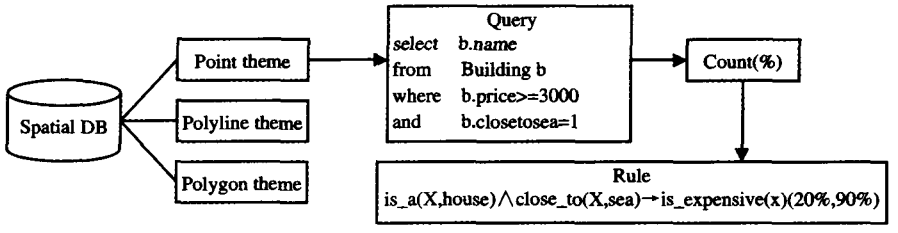

图 1 纵向空间关联规则挖掘

图中是对建筑表进行挖掘。其中，房屋被抽象表示为 0 维点，先对房屋的位置进行数据预处理，根据房屋与湖泊距离设置阈值，进行二值化处理，1 为靠近，0 为远离。然后借助空间查询进行计数，最后产生频繁项集并生成规则。当然，这里数据预处理可以采用模糊集、粗集或者云理论方法，均能够更好地表达与湖泊的关系[16,17]。本文为表述方便，仅做简单二值化处理。

2.2 空间关联规则的横向挖掘

空间关联规则的横向挖掘是对空间对象或者空间对象集之间的关系进行挖掘，得到横向关联规则。对象之间的关联关系主要包括拓扑关系、距离关系和方位关系。横向关联规则可能涉及一个空间关系模型或者多个，首先通过空间计算得到空间关联，然后采用 Apriori 算法进行规则提取，如图 2。图中的空间运算 adjacent()是定义在空间抽象数据类型上的一种求 A adjacent B 运算。

目前，对空间数据进行管理主要采取扩展关系数据库系统[3]。扩展型关系数据库系统在关系系统中增加空间抽象数据类型，扩展 SQL 使其可以像处理非空间数据一样地处理空间数据。新增空间数据类型（点、线、面域）作为基本几何类型，与其他非空间数据的表达和操作集成在一个逻辑层，同时在物理层对空间数据提供有效的存储和处理[3]。空间计算主要是在这个扩展 SQL 基础上，在空间数据库管理器中完成空间计算和挖掘中间数据存储。

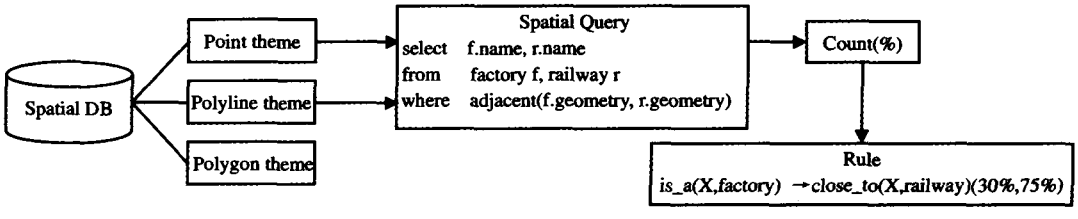

图 2 横向空间关联规则挖掘

2.3 空间关联规则的双向挖掘

空间关联规则的双向挖掘是对空间对象之间以及对象的属性之间同时进行双向关联规则挖掘，得到双向关联规则。在挖掘之前，需要对空间数据关联规则挖掘的任务进行初步定义，借此过滤掉不感兴趣的空间关联关系，以缩小空间分析的空间，减少复杂度，然后再进行关联规则挖掘。分为两步进行：第一步是得到非空间项集。根据空间关联规则挖掘的任务，对空间对象集进行空间关系运算，统计运算结果得到非空间项集，根据是否满足挖掘任务来决定是否重复以上空间关系运算来提升该项集直到满足要求，最后得到非空间项集；第二步是采用 Apriori 及其它算法在非空间项集上进行剪枝和连接，最后生成规则，或者回到空间关系运算进行多维关联规则挖掘。如图 3。

空间对象之间的关联关系在空间数据库中是隐含存储在空间拓扑表中，基于空间数据库定义的地理空间对象抽象数据模型，和抽象数据类型操作，可以完成大部分空间分析计算，部分复杂的空间分析可以借助专业的空间分析工具来完成。图中的空间查询 Meets()是定义在空间抽象数据类型上

的一种求 A *meets* B 运算。

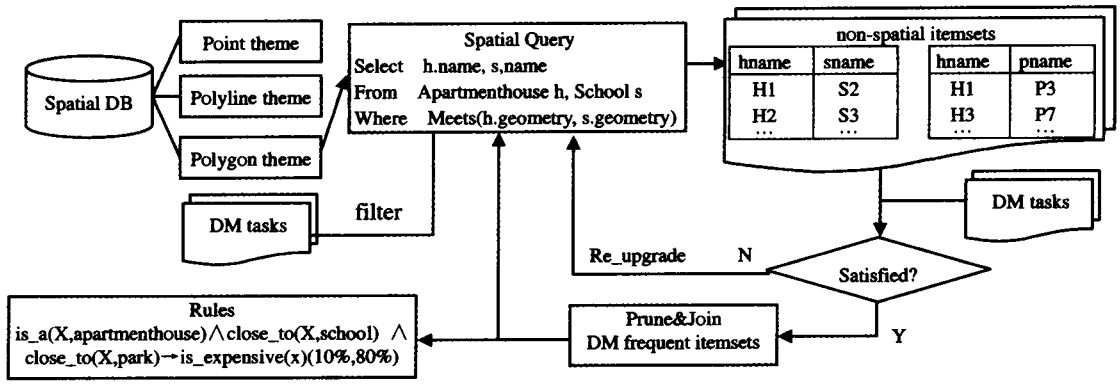

图 3 双向空间关联规则挖掘

3 实例运行

通过选用某大城市的地籍空间数据库对双向数据挖掘进行验证。该数据库是按照《城镇地籍数据库标准》建立的。该空间数据挖掘任务确定为研究公寓价格与学校、公园、小型广场的关系。在此任务的界定范围下，为了提高效率，对空间数据库进行精简，仅留下公寓、学校、公园、小型广场 4 个空间数据层。但是在数据挖掘的过程中，发现公寓靠近学校房价高这样规则的兴趣度很低，与实际经验有一定的偏差。于是对学校进行细化，将学校分为中学和小学进行挖掘而发现有趣的规则。统计表明：公寓有 1940 座，小学 285 个，中学 78 个，公园 18 个，小型广场 142 个。

第一步是对公寓、学校、公园、小型广场进行空间关系运算，得到非空间二项集。下面首先给出一个求二项集的算法描述，然后进一步说明。

```
Input: Spatial database SD, themes including a_house_tab, ele_
       Sch_tab
Output: non-spatial object sets a_house&ele_sch (A_house, ele_
        sch)
Begin
DBSizeH=COUNT(SD. a_house);
DBSizeS=COUNT(SD. ele_Sch);
Geo R1=Φ,R2=Φ;Geoset a_house&ele_sch;
for(i =1;i<=DBSizeH;i++){
    for(j=1;j<=DBSizeS;j++){
      R1 = BUFFER(a_housei);
      R2 = INTERSECT(R1. geometry, ele_schj. geometry);
      if (R2=TRUE)
    Insert a_housei, ele_schj into a_house&ele_sch;
      }
}
return a_house&ele_sch;
```

本算法是计算公寓 BUFFER 区内的小学数据集。根据挖掘任务要求，还需要计算公寓与中学、公园、小型广场的 L2。

第二步是在第一步得到的非空间二项集基础上，按照最低支持度和置信度要求，采用 Apriori 算法对非空间项集进行剪枝和连接，得到二项频繁项集。对多维数据挖掘，在二项频繁项集的基础上，需要回到第一步，增加其它维进行空间分析。重复以上步骤，直到完成挖掘任务。所得频繁项集和部分非频繁项集如表 2，规则如表 3。

规则表明，公寓靠近广场的价格较高的兴趣度较高，而公寓靠近公园的支持度和兴趣度都较低而被排除。经过实地调查分析，发现主要由于该城市的公园主要位于城市边缘，并且数量较少，该处开发的公寓由于购物、娱乐等设施没有完善，价格一直比较低。而小型的娱乐广场由于有一些绿地和健身设施，成为购房者首选。另外一条低支持度和兴趣度规则是靠近中学的公寓价格较高。经过项目组成员调查发现，居民对普通中学附近的公寓并不感兴趣导致售价低。然而，却对小学附近的公寓感兴趣，使得挖掘出公寓靠近小学价格较高的频繁项集。我们对城市有名的几所重点中学进行挖掘，无一例外地发现附近公寓价格均高。然而，由于该类重点中学太少，不能构成频繁项集而被剪枝去掉了。实例表明，该挖掘方法能够挖掘出兴趣度较高的有用的规则，从而指导城市规划和开发建设。

表 2 部分频繁项集和非频繁项集列表

k	Itemsets(L_k)			Count	Exp-count	S%	C%
2	a_house	h_sch		890	328	16.3	36.9
	a_house	ele_sch		1584	834	37.5	52.7
	a_house	pk		340	124	6.3	36.5
	a_house	sqr		1365	910	43.7	66.7
3	a_house	h_sch	sqr	459	203	9.4	44.2
	a_house	ele_sch	sqr	1189	793	33.5	66.7

表 3 空间关联规则列表

编号	关联规则
1	is_a(X,a_house)∧adjacent_to(X,sqr)→exp (X)(43.7%, 66.7%)
2	is_a(X,a_house)∧adjacent_to(X,ele_sch)→exp (X)(37.5%,52.7%)
3	is_a(X,a_house)∧adjacent_to(X,ele_sch)∧adjacent_to (X,sqr)→exp (X)(33.5%,66.7%)

总结 本文针对空间关联规则挖掘中存在的一些不甚明确的问题进行了探讨。在分析空间数据库的数据结构、数据操作、数据存储和空间关联的基础上，对空间关联规则进行了系统的定义，并据此提出空间数据关联规则挖掘可以从纵向属性之间和横向元组之间进行双向关联规则挖掘。针对横向挖掘的复杂性，首先需要对挖掘任务有一个初步定义，根据任务过滤掉一些不感兴趣的空间关系，缩小空间查询的空间，减少复杂度，然后采用两步的挖掘方式以获得规则。研究还对双向挖掘的工作模式进行了探讨，并在实践中进行验证。由于空间数据挖掘是空间信息处理领域一个新兴学科，目前仅仅取得了初步成果，尚有许多关键性技术问题需要进一步研究，如纵向挖掘空间属性数据的概化，以及在实践中表现出的横向挖掘的复杂性等方面。

参 考 文 献

1 Shekhar S, Ravada S, Fetterer A, et al. Spatial Databases: Accomplishments and Research Needs. IEEE Trans Knowledge and Data Eng, 1999,11(1):45～55
2 Nyerges T. Schema Integration Analysis for the Development of CIS Databases. Int J Geographic Information Systems, 1989,3(2):153～183
3 Gueting R H. An Introduction to Spatial Database Systems. The VLDB Journal, 1994,3(4):357～399
4 李德仁，王树良，李德毅，等. 论空间数据挖掘和知识发现的理论与方法. 武汉大学学报(信息科学版)，2002，27(3)：221～233
5 刘大有，王生生，虞强源，等. 基于定性空间推理的多层空间关联规则挖掘算法. 计算机研究与发展，2004，41(4)：565～570
6 李德毅. 知识表示中的不确定性. 中国工程科学，2000,2(10)：73～79
7 刘君强，潘云鹤. 挖掘空间关联规则的前缀树算法设计与实现. 中国图象图形学报，2003,04
8 Tung A K H, Lu Hongjun, Han Jiawei, et al. Efficient Mining of Intertransaction Association Rules. IEEE Transactions on Knowledge and Data Engineering, 2003, 15 (1)
9 裴韬，周成虎，骆剑承，等. 空间数据知识发现研究进展评述. 中国图象图形学报,2001,6(9):854～860
10 Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2000
11 Ester M, Kriegel H P, Xu X. Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification. In: Proc. 4th Int Symp on Large Spatial Databases, Portland, ME, 1995, Lecture Notes in Computer Science, Springer, 1995, 951:67～82
12 Egenhofer M. A Formal Definition of Binary Topological Relationships. In: Proc. Intl Conf On Foundations of Data Organization and Algorithms (FODO), 1989. 457～472
13 Efenhofer M, Franzosa R D. Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems, 1991,5(2):161～174
14 Theodoridis Y, Stefanakis E, Sellis T K. Efficient Cost Models for Spatial Queries Using R-Trees IEEE Trans. Knowledge and Data Eng, 2000,12(1):19～32
15 Tian Yongqing, Weng Yingjun, Zhu Zhongying. Mining association rules based on cloud model and application in prediction. In: Proceedings of the 4-th World Congress on Intelligent Control and Automation. Shanghai, P. R. China, 2002. 2203～2207
16 Matsakis P, Keller J M, Wendling L, et al. Linguistic Description of Relative Positions in Images. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2001, 31(4): 573～588
17 Kuok C, Fu A, Wong M. Mining fuzzy association rules in databases. 1998, 27(1): 41～46

(上接第 189 页)

格数 t 越少，则聚类效率越高；当数据真正所占网格(非异常点网格)数 m 越少，则聚类效率越高。

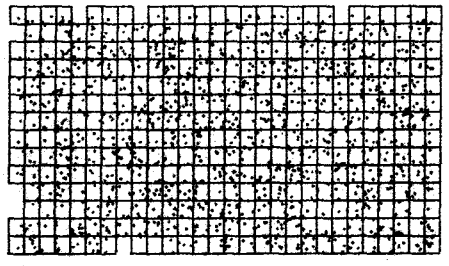

图 2 网格微聚类的结果

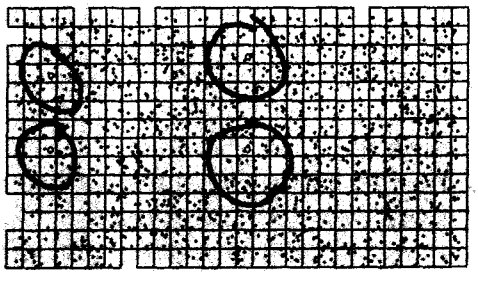

图 3 网格聚类选取的初始随机中心点

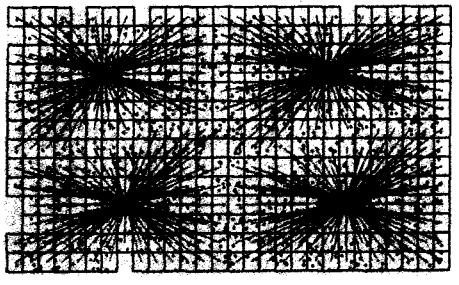

图 4 最后的聚类结果

4.3 MCC 算法评价

MCC 算法具有如下优点：

(1) 该算法克服了 K-均值算法要求事先给定类的个数的缺陷，由算法自动生成；

(2) 不必事先随机选取 K-均值算法的初始点，由算法自动选取初始点，从而避免了因初始点选择不当而导致收敛为一个局部最小准则函数的可能性；

(3) 该算法克服了网格聚类算法要求数据较集中的缺陷，可以用在数据较分散的情况中，而且可以方便有效地发现异常点；

(4) 该算法首先采用了网格聚类，所以算法的时间复杂度与数据对象数只呈一次线性关系，主要由网格的划分及数据的空间分布情况决定。如果数据的空间分布越集中，则实际所占的网格数越少，则聚类效率越高，时间复杂度低；

(5) 通过网格合并得到的聚类结果，也很容易给出其现实意义描述。

MCC 算法的主要缺点是聚类的结果依赖于异常点阈值和聚类阈值。

参 考 文 献

1 HarPeled S, Mazumdar S. Coresets for k-means and k-median clustering and their applications. In: Proc. 36th Annu. ACM Sympos. Theory Comput. ,2004. 291～300
2 Bach F R, Jordan M I. Learning spectral clustering. Advances in Neural Information Processing Systems (NIPS), 2004,16
3 LI CunHua, SUN ZhiHui. A Mean Approximation Approach to a Class of Grid－Based Clustering Algorithms. Journal of Software, 2003. 1267～1274
4 胡泱，陈刚. 一种有效的基于网格和密度的聚类分析算法. 计算机应用，2003(12)
5 Kanungo T, Mount D M, Netanyahu N, et al. A Local Search Approximation Algorithm for k－Means Clustering, Computational Geometry: Theory and Applications, 2004 (28): 89～112
6 Peng J M, Xia Y. A new theoretical framework for K－means clustering. To appear in Foundation and recent advances in data mining, Eds Chu and Lin, Springer Verlag, 2005
7 田启明，王丽珍，尹群. 基于网格距离的聚类算法的设计、实现和应用. 计算机应用，2005(2)
8 Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001(8)
9 Berson A, Smith S, Thearling K. 构建面向 CRM 的数据挖掘应用. 人民邮电出版社，2001(8)