文章编号:1671-8860(2008)12-1242-04　　**文献标志码**:A

基于 Voronoi 图的空间关联规则挖掘方法研究

李光强[1]　邓　敏[1]　朱建军[1]

(1　中南大学信息物理工程学院,长沙市麓山南路,410083)

摘　要:提出了利用 Voronoi 图作为辅助工具实现连续空间离散化的方法,并给出了基于 Voronoi 图构建空间事务数据库(spatial transaction database, STD)的算法。在此基础上,采用经典的 Apriori 算法来例证如何从 STD 中挖掘空间关联规则。实际算例验证了本文方法的可行性和有效性。

关键词:Voronoi 图;空间关联规则;空间数据挖掘

中图法分类号:P208

一般地,空间数据具有多维、多时相、多尺度以及海量等特征,且隐藏着丰富的但又不明确的空间关系[1],从而导致传统的关联规则挖掘方法不能直接用于空间关联规则的挖掘。目前,常用的空间关联规则挖掘方法主要包括:① 使用空间谓词表达空间关系,实现空间数据的事务化与非空间化;② 构建空间事务数据库(STD);③ 借鉴传统关系事务数据库的关联规则挖掘方法,从 STD 中挖掘空间关联规则[2]。在创建 STD 的过程中,连续空间的离散化起着关键作用。Koperski 和 Han 提出了使用用户指定特征周围的邻域作为事务的方法[3],但这种方法的缺点是发现空间关联规则具有不完备性。Morimoto 把空间事例分成不相邻的部分,并统计它们的频繁项集[4],其主要应用于类似移动通讯等领域的与位置有关的数据处理。在国内,多矩形表示方法(MRR 方法)、基于空间影响域的空间多维关联规则模型(SMARM 模型)、共生域(S-域)模型相继被提出[5-7]。为了实现连续空间离散化,本文提出了利用 Voronoi 图作为辅助工具来实现空间分割的方法。

1　基于 Voronoi 图的 STD

空间关联规则(SAR)是空间实体之间同时出现的内在规律,如空间实体间相邻、相连、共生和包含等空间关联关系,通常采用逻辑规则来表达[8]。有效的强空间关联规则还需要考虑作用度(置信度与期望置信度的比值),只有作用度大于 1 的空间关联规则才是有用的。

通常,平面上的 Voronoi 图是以一类地物集(图层)中的每个实体作为生长核,以相同的速率向外扩张,直到彼此相遇为止。Voronoi 图将整个连续的空间划分成若干个区域,并且每个 Voronoi 图可看作是生长核的势力范围[9,10],或视为生长核的影响域。

若给定空间数据库 $SD=l_0\cup l_1\cup l_2\cup\cdots\cup l_m=\{f_1, f_2,\cdots,f_n\}$,其中,$f_k(k=1, 2, \cdots, n)$为空间要素;$l_i(i=0,1,\cdots,m)$为数据库中包含的图层(见图 1(a)),$l_0$层为 Voronoi 图生长核空间要素所在的层,$l_i$. Attr 是层 l_i的属性集。空间要素 f 可表示为一个三元组,即$<\mathrm{FID}, l_j, g>$,FID 为空间要素的标识码;l_j为 f 所属的图层;g 为空间要素的几何数据。用 l_v表示生长核空间要素生成的 Voronoi 图层,l_v和 SD 中各层的相交操作记为 $l'_k=\mathrm{Intersect}(l_v, l_k)$,那么 l'_k的属性集 l'_k. Attr 可以表达为:

$$l'_k.\mathrm{Attr}=\{\mathrm{FID}_v\}\cup\{\mathrm{FID}_k\}\cup l_k.\mathrm{Attr} \quad (1)$$

进而,所有的 l'_k图层组成空间数据库 SD_v:

$$SD_v=l'_0\cup l'_1\cup l'_2\cup\cdots\cup l'_m \quad (2)$$

图 1(b)为相交操作的结果。层 l_0中第 i 个生长核记为 c_i,对应的 Voronoi 图记为 v_i。图中,垂

收稿日期:2008-10-15。

项目来源:国家自然科学基金资助项目(40501053);辽宁工程技术大学地理空间信息技术与应用实验室开放研究基金资助项目(2007001)。

直虚线围成的虚拟多棱柱“包含”的空间要素组成一个要素集，即 $VD_i=\{\forall f_k: \mathrm{Contain}(v_i, f_k)=\mathrm{True}\}$，Contain 是空间关系的包含操作算子；层内虚线是 Voronoi 图；符号“□”是 Voronoi 图生长核地物。

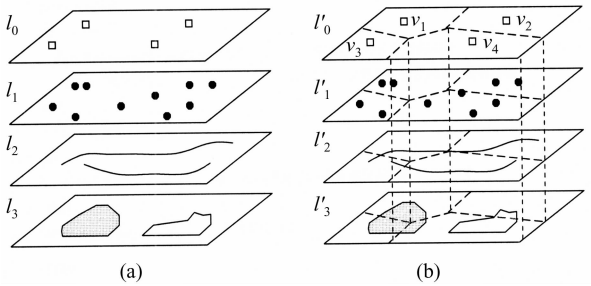

图 1　Voronoi 图与 SD 中各层的相交操作

Fig. 1　Operations Between Voronoi Diagram and Various Layers in Spatial Database

性质 1　对于 $\forall f_i \in SD_v$，$\exists v_j$、VD_j 使得 $\mathrm{Contain}(v_j, f_i)=\mathrm{True}$，$f_i \in VD_j$。

性质 2　对于 $\forall i$ 和 $j(i \neq j)$，有 $v_i \cap v_j=\varnothing$，$VD_i \cap VD_j=\varnothing$。

性质 3　对于 $\forall f_i \in VD_j$，f_i 对 c_j 的影响度都大于 f_i 对 $c_k(k \neq j)$ 的影响度。

假设 $f_i(f_i \in VD_j)$ 仅对 c_j 产生影响，而对 $c_k(k \neq j)$ 没有影响，则可将 Voronoi 图视作所有要素中具有相互影响、相互关联的约束单元“包含”在同一 Voronoi 图中的要素才有联系，否则没有联系。因此，Voronoi 图是构建 STD 的基本单元，一个 Voronoi 图可产生一条空间事务记录。

若记 $F_{ij}=\{\forall f_k: \mathrm{Contain}(v_i, f_k)=\mathrm{True} \wedge f_k \in l'_j\}$，即 $F_{ij}=VD_i \cap l'_j$，用 O_{r_i} 表示空间关系项集 R 中第 i 个空间关系项集的操作算子，项目值 $r_i=O_{r_i}(c_i, F_{ij})$。例如，某 STD 中包含“距离污染源”(即与最近污染源的距离)空间关系项目，操作算子 $\min(\mathrm{Distance}(c_i, \forall f_k: f_k \in F_{ij}))$ 返回 F_{ij} 要素集中与 c_i 最近的距离，其中，j 是污染源所在的层号；c_i 是土壤监测点的空间要素；F_{ij} 是所有落入 v_i 的污染源要素集合。

STD 中非空间属性项目数据的来源有两种：① 直接继承原空间实体的非空间属性值；② 由其他数据派生。综上所述，利用 Voronoi 图构建 STD 的步骤为：① 确定 STD 项目集 I；② 确定 Voronoi 图生长核要素层 l_0；③ 创建 Voronoi 图，并存储为层 l_v；④ l_v 和 SD 中各层的相交操作生成 SD_v；⑤ 使用 Voronoi 图创建空间事务元组。

2　从 STD 中挖掘空间关联规则的 Apriori 算法

在完成 STD 的构建后，需要对空间事务数据进行预处理。预处理包括数据归约、项目集约简等[12]。由于 STD 包含连续数值型项目(如距离)，因此还需离散化连续数值型项目。常用的离散化方法有概念分层和分箱。

为验证基于 Voronoi 图构建 STD 方法的适用性，本文采用经典的关联规则挖掘算法——Apriori 算法[11]来实现从 STD 中挖掘空间关联规则。Apriori 算法利用“在给定的事务数据库中，任意强项集的子集都是强项集；任意弱项集的超集都是弱项集”这一原理多次扫描事务数据库，第一次扫描得出频繁 1-项集 L_1，第 $k(k>1)$ 次扫描前，先利用第 $k-1$ 次扫描的结果(即频繁 $k-1$ 项集 L_{k-1})和函数 Apriori_Gen 产生候选频繁 k-项集 C_k；然后在扫描过程中确定 C_k 中每个元素的支持数，并且在每次扫描结束时计算出频繁 k-项集 L_k，算法在当候选频繁 k-项集 C_k 为空时结束；最后依据最小支持度和置信度，在 Apriori 产生的频繁项集中寻找关联规则。

3　算例分析

3.1　数据预处理与空间关联规则挖掘

为了验证基于 Voronoi 图的空间关联规则挖掘方法的可行性，本文选取华南某市的 21 个土壤重金属污染检测数据作为实验对象，挖掘重金属污染与公路距离、污染源分布、污染源距离等之间的空间关联规则。空间数据库包括 4 个图层：土壤监测点层、重点污染源分布层、公路层和河流层(见表 1)。其中，土壤监测点图层 l_0 是 Voronoi 图生长核所在的图层，Voronoi 图表示为 l_v 层。l_v 层和 l_1、l_2、l_3 各层相交操作后得到图层 l'_1、l'_2、l'_3，组成空间数据库 SD_v，图层 l'_1、l'_2、l'_3 均加入了 FID_v 字

表 1　空间数据库的组成

Tab. 1　Contents of Spatial Database

图层序号	图层名称	图层类型	主要的非空间属性
0	土壤监测点	点状	主要重金属(Cu、Zn、Ni、Cr、Pb、Cd)浓度和综合污染指数等
1	重点污染源分布	点状	污染源编号、类型、污水年排放量、废气年排放量、法人等
2	公路	线状	道路名称、等级等
3	河流	线状	河流名称、等级等

段，SD_v空间要素通过FID_v和Voronoi图建立关联。空间数据库采用ArcSDE+SQL Server2000管理，整个空间关联规则挖掘系统使用ArcMap9.2平台的VBA进行二次开发来完成。

使用Voronoi图构建的空间事务数据库列于表2。表中，主要重金属（Cu、Zn、Ni、Cr、Pb、Cd）浓度和综合污染指数直接从图层l_0中继承，其他数据项目使用空间算子从SD_v中计算得到。对表2进行维数归约和连续数值型项目数据离散化处理，结果见表3。

表2 基于Voronoi图构建的空间事务数据库

Tab.2 Spatial Transaction Database Constructed by Voronoi Diagram

SID	Cu浓度/mg·kg^{-1}	Zn浓度/mg·kg^{-1}	Ni浓度/mg·kg^{-1}	Cr浓度/mg·kg^{-1}	Pb浓度/mg·kg^{-1}	Cd浓度/mg·kg^{-1}	综合指数	污染源数量	污染源密度/个·km^{-2}	距离公路/m	河流密度/m·km^{-2}	距离电镀企业污染源/m
1	28.18	90.6	27.03	42.09	26.01	0.11	0.35	14	0.31	85	797.9	645
2	13.21	35.3	7.60	11.80	34.22	0.04	0.04	8	0.02	78	674.7	2 120
3	9.09	38.5	11.02	12.20	11.24	0.04	0.03	7	0.01	1 064	908.2	1 213
…	…	…	…	…	…	…	…	…	…	…	…	…

表3 处理后的空间事务数据库

Tab.3 Spatial Transaction Database After Processing

SID	Cu浓度	Zn浓度	Ni浓度	Cr浓度	Pb浓度	Cd浓度	综合污染指标	污染源密度	距离公路	距离电镀企业
1	高	高	高	高	中	高	高	高	近	近
2	低	低	低	低	高	低	低	低	近	远
3	低	低	中	低	低	低	低	低	远	远
4	中	低	高	高	低	低	中	低	远	远
5	低	低	低	低	低	低	低	中	远	远
6	低	中	中	中	中	低	低	低	较近	远
7	中	中	中	中	低	中	低	中	远	远
8	中	中	高	低	中	中	高	高	远	较近
9	低	中	低	低	中	低	低	低	较近	远
10	中	高	高	高	高	中	中	高	近	近
11	高	中	高	低	中	中	中	中	远	较近
12	低	低	低	中	高	中	低	中	近	远
13	高	高	中	中	高	高	高	高	较近	近
14	高	中	中	中	高	高	低	中	近	较近
15	中	高	中	高	高	中	中	高	近	较近
16	中	高	中	中	高	高	中	中	较近	较近
17	低	中	低	低	低	高	低	中	近	远
18	中	中	中	高	高	低	高	高	近	近
19	高	高	高	中	中	低	高	高	近	近
20	中	高	高	高	高	高	中	中	近	较近
21	中	中	中	高	高	中	中	低	较近	远

表4列出了经过Apriori算法挖掘得到的15个强空间关联规则（最小支持度阈值为15%，最小置信度阈值为70%，作用度大于等于1）。

3.2 空间关联规则的约简与表达

评价强关联规则就是判断挖掘出的强关联规则是否有用或有趣，评价中，需要引入应用领域的先验知识，应用领域先验知识的利用突出体现在强关联规则的约简和表达中[11]。观察表4中的15条强关联规则，可发现规则1、2、3、5、7、8隐含了规则10、11、13、14、15，所以规则10、11、13、14、15都是冗余规则。在实际环保工作中，环保工作人员更加关注企业分布或公路等对环境污染产生的影响，所以规则3、5、6、9是环保工作人员兴趣度较低的规则，因此，约简后的强关联规则仅包括规则1、2、4、7、8、12。

表4 强空间关联规则

Tab.4 Strong Spatial Association Rules

编号	关联规则内容（支持度，置信度）
规则1	污染源密度（高）⇒Zn浓度（高）（23.8%，71.4%）
规则2	污染源密度（高）⇒综合污染指数（高）（23.8%，71.4%）
规则3	污染源密度（低）⇒Cd浓度（低）（23.8%，83.3%）
规则4	距离公路（近）⇒Pb浓度（高）（33.3%，70.0%）
规则5	距离电镀企业（远）⇒综合污染指数（低）（38.1%，80.0%）
规则6	距离电镀企业（远）⇒Cu浓度（低）（33.3%，70.0%）
规则7	距离电镀企业（近）⇒Zn浓度（高）（19.0%，80.0%）
规则8	距离电镀企业（近）⇒综合污染指数（高）（19.0%，80.0%）
规则9	距离公路（较近）⇒Ni浓度（中）（19.0%，80.0%）
规则10	污染源密度（中）∧距离电镀企业（远）⇒综合污染指数（低）（19.0%，100.0%）
规则11	污染源密度（高）∧距离公路（近）⇒Zn浓度（高）（19.0%，80.0%）
规则12	污染源密度（高）∧距离公路（近）⇒Cr浓度（高）（19.0%，80.0%）
规则13	污染源密度（高）∧距离电镀企业（近）⇒Zn浓度（高）（19.0%，80.0%）
规则14	污染源密度（高）∧距离电镀企业（近）⇒综合污染指数浓度（高）（19.0%，80.0%）
规则15	污染源密度（低）∧距离电镀企业（远）⇒Cd浓度（低）（23.8%，83.3%）

约简后的规则可表达为：① 污染源密度高或者距离电镀企业近的区域，Zn浓度高、综合污染指数高；② 距离公路近的区域，Pb浓度高；③ 污染源密度高且距离公路近的区域，Cr浓度高。

4 结 语

本文提出了利用Voronoi图作为辅助工具实现连续空间离散化的方法，实际算例验证了该方

法的有效性。然而，在利用 Voronoi 图构建空间事务单元域时，假定图内空间要素仅对图的生长核产生影响，忽略了相邻图内空间要素之间的关联和影响，从而损失了一定的空间信息量。此外，在对 STD 中的连续数值型项目分组时，确定分组条件的主观因素较多。因此，进一步的研究工作应当结合领域背景知识，使用机器学习和归纳推理等方法，使得分组更加客观、合理，最终得到的关联规则更具实用性。

参 考 文 献

[1] Malerba D, Lisi F A, Appice A, et al. Mining Spatial Association Rules in Census Data: a Relational Approach[C]. The ECML/PKDD'02 Workshop on Mining Official Data, University Printing House, Helsinki, 2002

[2] Rinzillo S, Turini F. Knowledge Discovery from Spatial Transactions[J]. Journal of Intelligent Information Systems Archive, 2007, 28(1):1-22

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

[4] 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

[5] 陈江平，傅仲良，边馥苓，等. 基于空间分析的空间关联规则提取[J]. 计算机工程，2003，29(11)：29-31

[6] 黄添强，秦小麟，叶水生，等. 一种新的空间多维关联规则与算法[J]. 南京航空航天大学学报，2005，37(3)：301-306

[7] 杨晓梅，蓝荣钦，杨松. 基于 S-域的空间关联规则挖掘研究[J]. 测绘科学技术学报，2004，24(1)：10-17

[8] 李德仁，王树良，史文中，等. 论空间数据挖掘和知识发现[J]. 武汉大学学报·信息科学版，2001，26(6)：491-499

[9] Gold C M. The Meaning of "Neighbour"[J]. Lecture Notes in Computing Science ,1992(39)：220-235

[10] 陈军，赵仁亮，乔朝飞. 基于 Voronoi 图的 GIS 空间分析研究[J]. 武汉大学学报·信息科学版，2003，28(S1)：32-37

[11] Han Jiawai, Kamber M. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001

第一作者简介：李光强，讲师，博士生。主要从事空间数据挖掘、异常模式探测等方面的研究及教学工作。
E-mail：ligq168@163.com

Spatial Association Rules Mining Methods Based on Voronoi Diagram

LI Guangqiang[1]　*DENG Min*[1]　*ZHU Jianjun*[1]

(1 School of Infophysics and Geomatics Engineering, Central South University, South Lushan Road, Changsha 410083, China)

Abstract: In order to make discretization of continuous space, it is proposed to use Voronoi diagrams to partition the space of interest into the domains of spatial transaction database. The process and the algorithm to construct the spatial transaction database are discussed based on Voronoi diagrams. On the basis of this, Apriori algorithm is taken to mine spatial association rules from the constructed spatial transaction database. A practical example is given to illustrate the rationality of the proposed method.

Key words: Voronoi diagram; spatial association rules; spatial data mining

About the first author: LI Guangqiang, lecturer, Ph.D candidate, majors in spatial data mining and spatio-temporal outliers dectection.
E-mail: ligq168@163.com