文章编号:1671-8860(2009)12-1485-05　　**文献标志码**:A

一种基于免疫算法的空间关联规则挖掘方法

朱　玉[1]　张　虹[1]　孔令东[1]

(1　中国矿业大学环境与测绘学院,徐州市解放南路延长段 1 号,221008)

摘　要:针对海量数据空间关联规则挖掘的不足,提出了一种基于免疫算法的空间关联规则挖掘方法。算法充分利用了免疫识别、免疫记忆及克隆选择特性,把要挖掘的空间关联规则作为抗原,候选项目集作为抗体,把挖掘的关联规则存入记忆库,加快了关联规则的挖掘速度。以杆塔故障的空间要素的关联关系为例,验证了算法的有效性。

关键词:免疫算法;空间数据挖掘;空间关联规则;频繁项集

中图法分类号:P208

自 1995 年 Koperski 将传统关联规则拓展到空间数据挖掘领域以来,很多学者对空间关联规则的概念、挖掘算法、不确定性的表达和挖掘结果的可视化等方面进行了深入的研究并取得了一系列的成果[1-3]。目前,关于空间关联规则挖掘的研究基本上是基于一般事务数据库的关联规则挖掘算法进行[4-8]。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为空间数据量庞大及空间数据之间存在复杂的关系(拓扑关系、位置关系和距离关系),导致算法效率较低。

免疫算法是利用人工免疫系统(artificial immune system, AIS)的仿生机制设计出的一种新型算法[9]。AIS 具备强大的识别、学习和记忆能力,及分布式、自组织、多样性和没有中心控制点等特性。它的自体/非自体识别能力正是空间关联规则挖掘良好而又天然的解决方法。同时,免疫算法具备高效率并行搜索的能力,是处理大规模数据项目集的有效方法[10]。为了解决海量空间数据的关联规则挖掘问题,本文提出了一种基于免疫算法(immune algorithm,IA)的空间关联规则挖掘方法,挖掘过程不需要生成大量的频繁项集,从而提高关联规则挖掘的总体性能。实验表明,算法具有良好的性能。

1　基于免疫算法的空间关联规则挖掘

1.1　相关概念

表 1 为某雷电频发地杆塔故障监测的数据样本。

1) 抗原。通常对应要解决的问题,根据不同的问题有不同的数据编码形式。本文将用户感兴趣的属性值作为抗原,例如,对于表 1,若认为杆塔故障的最主要影响因素为 d_1,则可以将 d_1 作为抗原。

表 1　某雷电频发地杆塔故障监测数据样本

Tab. 1　Pole and Power Fault Monitoring Data Samples of a Frequent Thunder District

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
a_1	b_1	c_1	d_1	e_1	f_1	g_1	h_2	i_3	j_1	o_1
a_2	b_2	c_2	d_2	e_2	f_2	g_2	h_3	i_1	j_3	o_2
a_3	b_2	c_1	d_2	e_1	f_3	g_1	h_1	i_2	j_2	o_3

2) 抗体。通常对应要解决的问题的优化解。本文中抗体即是最终挖掘的空间关联规则,例如,

收稿日期:2009-10-18。

项目来源:江苏省自然科学基金资助项目(BK200502);江苏省普通高校研究生科研创新计划资助项目(CXB-392)。

对于表1,设抗原为 d_1,则抗体是 $d_1 \Rightarrow o_1$。

3) 信息熵。假设某抗体有 M 个基因,每个基因位上采用的字符集大小为 S(若采用二进制编码,字符集就为{0,1},$S=2$),则该抗体的信息熵为:

$$H(\eta) = \frac{1}{M}\sum_{j=1}^{M} H_j(\eta) \tag{1}$$

式中,$H_j(\eta)=\sum_{i=1}^{S} -p_{ij}\lg p_{ij}$,$H_j(\eta)$为该抗体第 j 个基因的信息熵,p_{ij} 是字符集中第 i 个符号出现在第 j 个基因座上的概率。

4) 亲和力。亲和力表示抗原抗体之间的结合强度,亲和力越高则抗原与抗体结合强度越高。

5) 抗体的适应值。抗体的适应值用于评价抗体的优劣。选取恰当的适应度函数是空间关联规则挖掘的一个关键问题,因为基于免疫算法的空间关联规则挖掘算法的主要目标就是优化种群中抗体的适应值。确定所有满足阈值条件的关联规则,并删除负相关的规则,所得到的各关联规则的支持度与置信度之和即为该抗体的适应值,即

$$f(i) = S(i) + C(i) \tag{2}$$

式中,$f(i)$为抗体 i 的适应值;$S(i)$、$C(i)$分别表示抗体 i 的支持度和置信度。

1.2 免疫算法流程

1) 抗原识别,将需要解决的问题抽象成符合处理的抗原形式。抗原识别对应于求解问题。

2) 产生初始抗体群体,如果记忆库中有记忆抗体,则将记忆库中的抗体看成是初始抗体的一部分,不足的部分,系统随机产生。

3) 亲和力计算,计算抗原与抗体之间的亲和力。

4) 记忆细胞分化,与抗原有最大亲和力的抗体加入记忆库。由于记忆细胞数目有限,新产生的与抗原具有更高亲和力的抗体替换较低亲和力的抗体。

5) 抗体的促进和抑制,高亲和力抗体受到促进,高浓度抗体受到抑制。

6) 群体更新,群体更新保证了抗体群的多样性。该操作将群体中亲和力较小的部分抗体用产生的等量新抗体取代。

7) 抗体产生,通过交叉和变异操作来产生新的抗体。

8) 若满足结束条件(结束条件一般指迭代周期达到最大代数)则结束,否则转步骤3)。

1.3 算法原理

一般而言,由空间数据库挖掘关联规则是一个两步的过程:① 找到所有支持度大于最小支持度的项目集,这些项目集被称为频繁项集。② 使用找到的频繁项集产生期望的强关联规则。根据定义,这些规则必须满足最小支持度和最小置信度。

在具体应用中,往往存在着仅对空间数据库组成事务的某一个或几个项集感兴趣,仅需要对这几个项集进行关联规则挖掘即可,而不必要对整个数据库进行全面的规则提取。因此,只需要将重点放在几个重点的项集上,对其相应的关联规则进行挖掘即可。故可将免疫算法的思想引入空间关联规则挖掘中,进行关联规则提取,其具体步骤如下所示。

1) 以要挖掘的空间关联规则作为抗原,可设定一个或多个抗原。

2) 抗体抗原编码,一般采用字符编码或实数编码方式。

3) 确定抗体种群规模 M,免疫选择阈值 Th 及终止条件最大迭代次数 N。

4) 在解空间随机产生 n 个候选解作为抗体种群和记忆库中的抗体,共同组成初始种群 $P_t(t=0)$。

5) 计算种群中每个抗体的适应值,从当前种群中选择适应值最高且适应值互不相同的 N_1 个抗体,进行克隆操作,每个抗体克隆的数目 N_c 与其适应值成正比。对于克隆新生成的抗体,实施超变异操作(超变异操作就是在变异操作之前先进行倒位操作),抗体的适应值越高,其对应的变异率越小。

6) 分别计算抗原 Ag 与抗体 v 之间的亲和力 ax_v 及抗体 v 与抗体 w 之间的亲和力 A_{vw}。

根据信息熵的定义,可得到任意两个抗体之间的亲和力:

$$A_{vw} = \frac{1}{1+H(2)} \tag{3}$$

式中,A_{vw} 的取值范围是(0,1);$H(2)$是抗体 v 和 w 之间的信息熵,$H(2)=0$ 时说明抗体 v 和 w 的所有基因都是相同的。

由于关联规则中最重要的参数是置信度与支持度,所以把抗原 Ag 与抗体 v 之间的亲和力定义为:

$$ax_v = W_s \times \frac{\text{support}(X \Rightarrow Y)}{\text{minsup}} + W_c \times \frac{\text{confidence}(X \Rightarrow Y)}{\text{minconf}} \tag{4}$$

式中,$W_s+W_c=1$,$W_s \geqslant 0$,$W_c \geqslant 0$;minsup 是支持度的阈值;minconf 是置信度的阈值。

7) 把抗体种群中,同时满足最小支持度和最

小置信度条件的抗体加入到记忆库中。由于在关联规则挖掘过程中,事先不能知道挖掘的关联规则的数目,因此不设定记忆库的规模。

8)抗体的促进和抑制。计算抗体的期望值,期望值低的将受到抑制。抗体在种群中的浓度为:

$$C_v = \frac{1}{n}\sum_{w=1}^{n} ac_{vw} \tag{5}$$

式中,$ac_{vw}=\begin{cases}1, A_{vw}>Th\\0, \text{其他}\end{cases}$。

抗体 V 的期望值定义为:

$$S_v = \frac{ax_v}{C_v} \tag{6}$$

由上式可知,与抗原亲和力高的抗体和低浓度的抗体生存的机率较大,高亲和力的抗体得到促进而高浓度的抗体受到抑制,从而体现了免疫控制的多样性。

9) 从抗体种群中未进行克隆的个体中随机选择 N_d 个抗体,用经过抗体抑制保留的抗体中适应值最高的 N_d 个抗体进行替换,得到下一代抗体种群 P_t, $t=t+1$。

10) 若满足终止条件,把编码还原为规则,结束挖掘过程;否则转步骤 5)。终止条件为迭代周期达到最大迭代次数。

在记忆库中设置关联规则表和有趣规则表,关联规则表和有趣规则表分别存放关联规则和有趣规则。每次挖掘的结果都存放在关联规则表和有趣规则表中,表中记录有规则结构、支持度、置信度等项目。

1.4　空间关联规则挖掘算法分析

本文算法从种群中选择适应值最高且适应值互不相同的 N_1 个抗体,作为进行克隆候选个体。这种选择策略能够保证搜索区域的多样性,避免陷入未成熟收敛。每个抗体克隆的数目 N_c 与其适应值成正比,即

$$N_c = \max\left\{N_{\min}, \text{round}\left(\frac{N_{\max}\times f(i)}{f_{\max}}\right)\right\} \tag{7}$$

式中,$N_{\min}$ 和 $N_{\max}$ 分别为最小和最大克隆数目;$f_{\max}$ 为当前种群中抗体的最大适应值。

在更新抗体种群时,随机选择 N_d 个未进行克隆操作的抗体,用经过抗体抑制保留的抗体中适应值最高的 N_d 个抗体进行替换,每代种群的规模保持恒定。本文算法在选择克隆候选抗体时,既考虑了抗体的适应值,又考虑了抗体之间的相似性,该过程包含了对相似抗体的抑制,它能与种群更新机制共同保持种群中个体的多样性。

空间关联规则频繁项集挖掘的主要开销在空间数据库的多次扫描时间上,导致算法的时间代价非常高。本文算法充分利用了免疫算法的免疫记忆特性,通过免疫学习把挖掘的关联规则保存在记忆库中,加快了关联规则的挖掘速度,同时算法具有较强的鲁棒性和快速、有效的全局搜索能力。

2　实例及分析

对雷击频发地杆塔故障监测的数据集进行空间关联规则提取,软件工具选取 VC^{++} 与 Matlab。杆塔故障监测部分数据如表 2 所示。

首先对数据库进行泛化处理,泛化处理后的空间数据集如表 3 所示。再对泛化后的空间数据集的属性值进行编码,部分编码结果如表 4 所示。然后应用算法对空间关联规则进行挖掘,取种群规模为 1 000,免疫选择阈值为 90%,最大迭代次数为 500 代。当以支持度阈值为 30%、置信度阈值为 90%进行空间关联规则挖掘测试时,挖掘出关联规则 126 条,挖掘时间为 19.8 s,得到的部分空间关联规则如下:

表 2　某雷击频发地杆塔故障监测数据

Tab. 2　Pole and Power Fault Monitoring Data of a Frequent Thunder District

植被平均高度/m	植被面积/m^2	与植被距离/m	雷击频率/(次/a)	杆高/m	车流量/(辆/h)	污染级别	污染源名称	道路宽度/m	与道路距离/m	杆塔故障率/(次/a)
8	35	3	3	27	3 000	高	煤窑	30	3	6
2	3	15	1	6	2 500	低	砖瓦厂	35	12	1
…	…	…	…	…	…	…	…	…	…	…

表 3　泛化后的空间数据集

Tab. 3　Generation Spatial Datasets

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	较近	较高	高	大	高	煤窑	较宽	较近	高
低	较小	较远	较低	低	大	低	砖瓦厂	较宽	较远	低
…	…	…	…	…	…	…	…	…	…	…

表 4 部分属性值编码结果表

Tab. 4 Part of Attribute Encoding

属性名	属性值	编码值	属性名	属性值	编码值
植被面积	较大	1	雷击频率	较高	1
	中等	2		中	2
	较小	3		较低	3
植被平均高度	高	1	杆高	高	1
	低	2		低	2
	较远	1	车流量	大	1
与植被距离	中	2		中	2
	较近	3		小	3
…	…	…	…	…	…

与植被距离=3∧植被面积=1∧雷击频率=1⇒杆塔故障率=1 [support=33.7%, confidence=92.5%];

与道路距离=3∧车流量=1∧雷击频率=1⇒杆塔故障率=1 [support=35.6%, confidence=99.5%];

植被平均高度=2∧车流量=3∧污染级别=2⇒杆塔故障率=2[support=36.4%, confidence=97.2%]。

从以上的挖掘结果来看,结合该雷击频发地的实际情况,基于免疫算法的空间关联规则挖掘算法是有效的。

为了便于同本文提出的算法相比较,同时采用经典 Apriori 算法和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 对上述数据集进行空间关联规则提取,并分别对提取的质量和时间进行分析。

图 1 表明,当支持度比较大时,3 种算法的时间代价相差不大,而当支持度的阈值不断降低时,由于经典 Apriori 算法和 FT_MLSAM 算法均要不断地扫描数据库,时间聚集增加,产生大量的候选集合。因此,基于免疫算法的空间关联规则挖掘算法的效率优于经典 Apriori 算法和 FT_MLSAM 算法。

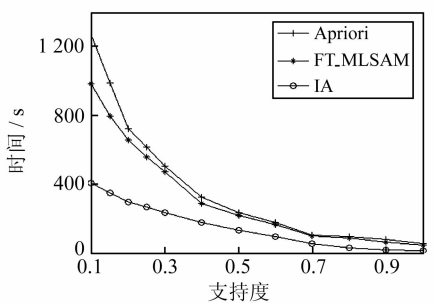

图 1 算法时间性能比较

Fig. 1 Time Comparison of Three Algorithms

固定置信度阈值为 0.7,支持度阈值从 0.3 增加到 0.6 时,3 种算法所得到的满足阈值条件的关联规则数目如图 2 所示。可以看出,随着最小支持度的增加,频繁项集的数目逐渐减少,所得到的空间关联规则数目也会相应减少,但对于不同的阈值条件,本文算法都能得到更多的满足阈值条件的空间关联规则。

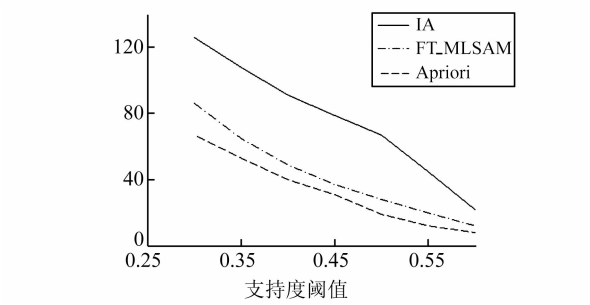

图 2 不同支持度阈值得到的空间关联规则数目

Fig. 2 Number of Spatial Association Rules of Different Support Thresholds

固定支持度阈值为 0.5,置信度阈值从 0.6 增加到 0.9 时,3 种算法所得到的满足阈值条件的关联规则数目如图 3 所示,本文算法仍能得到更多的满足条件的关联规则。

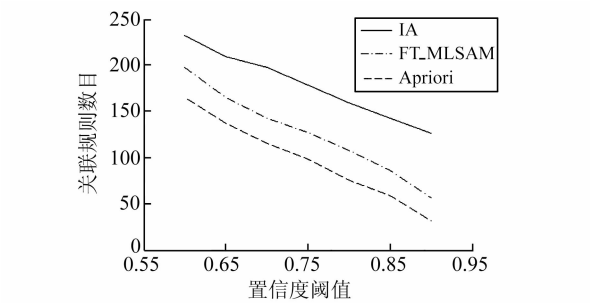

图 3 不同置信度阈值得到的关联规则数目

Fig. 3 Number of Spatial Association Rules of Different Confidence Thresholds

3 结 语

空间关联规则是传统关联规则在空间数据挖掘领域的延伸,但又与传统关联规则存在着较大的不同。空间数据库中包含有大量的对象和对象间的关系,所以空间关联规则本质上是空间实体间相邻、相连、共生和包含等基于多关系数据格式的关联规则。在传统关联规则挖掘中,尽管采用支持度和置信度阈值可过滤出高频强模式,但是往往还会出现大量的规则冗余、无意义模式,同时,空间谓词计算复杂度高,时间代价大。目前,大多数算法采用空间 Apriori 算法或它的变形。

本文提出了一种基于免疫算法的空间关联规

则挖掘算法,该算法充分利用了免疫算法的免疫识别、免疫记忆和克隆选择特性,通过免疫学习把挖掘的关联规则保存在记忆库中,克隆选择有利于群体的相对稳定,加快了关联规则的挖掘速度,同时算法具有较强的鲁棒性和快速、有效的全局搜索能力。通过实际数据集与相关算法进行了比较,实验结果表明了本文算法具有更优的性能,适用于海量空间数据库的关联规则挖掘。

参 考 文 献

[1] Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases [M]//Egenhofer M J, Herring J R. Advances in Spatial Databases. Berlin: Springer-Verlag, 1995: 47-66

[2] Ester M, Kriegel H P, Sander J. Spatial Data Mining: a Database Approach[C]. The 5th International Symposium on Spatial Database, Berlin, 1997

[3] Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases[J]. Lecture Notes In Computer Science, 1995, 95: 47-66

[4] 陈江平, 李平湘. 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[J]. 遥感学报, 2006, 10(2): 289-293

[5] 刘小生, 任海峰, 陈棉. 用空间分析方法进行空间关联规则提取[J]. 测绘通报, 2007(5): 19-21

[6] 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

[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 Geo-computation, Brisbane, Australiz, 2001

[8] 马荣华, 何增友. 从空间数据库中挖掘频繁邻近类别集的一种新算法[J]. 武汉大学学报·信息科学版, 2007, 32(2): 112-114

[9] 王磊, 潘进, 焦李成. 免疫算法[J]. 电子学报, 2000, 28(7): 74-78

[10] 王新洲, 许承权. 免疫算法及其在测量数据处理中的应用[J]. 武汉大学学报·信息科学版, 2006, 31(10): 887-890

第一作者简介:朱玉,博士生,主要从事人工免疫、空间数据挖掘等方面的研究。

E-mail:zhuyujs@139.com

A New Spatial Association Rules Mining Method Based on Immune Algorithms

ZHU Yu[1]　*ZHANG Hong*[1]　*KONG Lingdong*[1]

(1 School of Environment Science and Spatial Informatics, China University of Mining and Technology, 1 Extension Section, South Jiefang Road, Xuzhou 21008, China)

Abstract: On the basis of analyzing the now-generally-used spatial association rules algorithm, aiming at the shortage of the very large database spatial association rules mining, a spatial association rules mining algorithm based on immune algorithms is proposed. This algorithm makes use of the immune recognition mechanism, immune memory characters and clonal selection characters. In the process of spatial association mining, spatial association rules are regarded as the antigens, candidate itemsets are looked upon as the antibodies. The spatial association rules are stored in memory, and speed of mining spatial association rules is accelerated. We take the incidence relation of special data of pole and tower fault as an example, to verify the algorithm. Experiment results show that the proposed algorithm is effective. The algorithm is able to be more quickly and efficiently search in the whole global, and extremely be used for the mining spatial association rules to very large database.

Key words: immune algorithm; spatial data mining and knowledge discovery; spatial association rules; frequent itemset

About the first author: ZHU Yu, Ph.D candidate, majors in artificial immune system and spatial data mining.

E-mail: zhuyujs@139.com