第 34 卷 第 6 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.6 Computer Engineering March 2008
·软件技术与数据库· 文章编号：1000—3428(2008)06—0051—03 文献标识码：A 中图分类号：TP311.13

数据挖掘中加权时态关联规则的构造

朱建平，乐燕波
(厦门大学经济学院计划统计系，厦门 361005)

摘 要：传统的关联规则很少考虑规则的时间适用性，而时态关联规则中每条关联规则都有其成立的时间区域，对上述问题进行了一定的改进。该文在此基础上，构造了一种体现数据时间价值的加权时态关联规则，以使规则的发现体现一种时间趋势，并对同一组数据采用不同关联规则挖掘的结果进行比较，取得了良好的效果。

关键词：数据挖掘；加权时态关联规则；事件生命周期

Construction of Weighted Temporal Association Rules in Data Mining

ZHU Jian-ping, LE Yan-bo
(Department of Planning and Statistics, College of Economics, Xiamen University, Xiamen 361005)

【Abstract】 The fitness of time is seldom illustrated by traditional association rules. Temporal association rules are improved by regarding every association rule with valid time area. Weighted temporal association rule is presented in this paper based on these researches, which can reflect the time value of data and the time tendency of discovered rules, and the results of different association rules mining on the same data are also compared and achieve a fine performance.

【Key words】 data mining; weighted temporal association rules; affair life circle

1 概述

关联规则是数据挖掘研究中的一个重要课题，由文献[1]首次提出，目的是在交易数据库中发现各项目间的关系。关联规则发现的经典算法是由文献[2]提出的 Apriori 算法。该算法将关联规则的发现分为 2 步：(1)识别所有的频繁项集，即其支持度不低于用户设定的最低支持度的项目集；(2)从频繁项集中构造其信任度不低于用户设定的最低信任度的规则。以后诸多研究人员对该算法进行了优化扩展，但其基本框架没有变化。

传统的关联规则研究很少考虑规则的时间适用性。事实上，时间是数据本身固有的因素，许多数据库中的记录都带有时间标记，如交易记录中的交易时间、病历信息库中的诊断时间等。时态数据库的出现必然要求在知识发现过程中考虑时间因素[3]。目前，规则都假定为永远有效，而不表明其何时开始有效，何时开始无效；无效的规则也没有说明它在过去或将来是否有效[4]。因此，附加上某种时态特征的规则能更好地描述实际情况，这样的规则就称为时态关联规则。它能有效挖掘出一些全局支持度较低，但在某些时段却有较高支持度和信任度的规则，如中秋节的月饼、圣诞节礼品等容易在传统关联规则挖掘中被忽略但对用户有重要价值的规则。目前国内外研究的主要内容有：带有一般时态约束的关联规则挖掘，周期关联规则挖掘，趋势性挖掘，序列模式挖掘等。

本文对时态关联规则的一些相关研究进行了剖析，在此基础上，提出了加权时态关联规则的概念，初步研究了其应用的可行性，即给予较近发生的事件较大的权值，使规则的发现体现一种时间趋势。用本方法挖掘时态数据库能更好地反映事物发生发展的过程，有助于揭示事物发展的本质规律。最后，对不同关联规则挖掘得到的结果进行了比较和分析。

2 时态关联规则的相关研究工作

带有时态特征的关联规则称为时态关联规则。时态关联规则挖掘就是要发现事件与时间之间的关联以及基于时域的事件与事件之间的关系。本文主要研究带有一般时态约束的关联规则。时态关联规则的一些基本概念如下：

定义 1 设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项目的集合，时态数据可表示为 $D=\{D_1,D_2,\cdots,D_n\}$，其中，$D_i=\langle tid,itemset,T_i\rangle$，$tid$ 是事务的标识，$itemset\subseteq I$，T_i 是事务发生的时刻[5]。

定义 2 设项集 $X\subseteq I$，X 在 D 内从最初出现到最后出现的时间区域为 $[T_1,T_2]$，$T_1<T_2$，称 $X[T_1,T_2]$ 为事件 X 的生命周期。$|X[T_1,T_2]|$ 为 $[T_1,T_2]$ 时间段上包含事件 X 的个数。$D[T_1,T_2]$ 是 D 在 $[T_1,T_2]$ 上的子集，$|D[T_1,T_2]|$ 表示 D 在 $[T_1,T_2]$ 上总的事务个数。

时间粒度是最小的时间区域，是衡量时域长度的基本单位，时域长度是包含时间粒度的个数[5]。一些项集成立的时域太小，因此设立时态阈值 τ 以剔除无实用价值的规则。

若项集 X 和 Y 成立的有效时间区域分别为 $[T_1,T_2]$ 和 $[T_1',T_2']$，令 $Z=X\cup Y$，则项集 Z 成立的时间区域是

$$[T_1,T_2]\cap[T_1',T_2']$$

定义 3 项集 X 的支持度为

$$Support(X[T_1,T_2])=|X[T_1,T_2]|/|D[T_1,T_2]|$$

基金项目：国家教育部新世纪优秀人才计划基金资助项目(NCET-04-0608)；国家教育部社科研究规划基金资助项目(06JA910003)
作者简介：朱建平(1962－)，男，教授、博士、博士生导师，主研方向：数理统计，数据挖掘；乐燕波，硕士研究生
收稿日期：2007-05-20 **E-mail：**leyanbo007@sina.com

简写为 $S(X)$。

规则 $X \Rightarrow Y$ 的支持度为

$$Support(X \cup Y) = \left| X \cup Y\left[(T_1,T_2) \cap (T_1',T_2')\right] \right| \Big/ \left| D\left[(T_1,T_2) \cap (T_1',T_2')\right] \right|$$

信任度 $C(X \Rightarrow Y)$ 为

$$Confidence(X \cup Y) = \left| X \cup Y\left[(T_1,T_2) \cap (T_1',T_2')\right] \right| \Big/ \left| X\left[(T_1,T_2) \cap (T_1',T_2')\right] \right|$$

定义 4 给定最小支持度 σ、最小信任度 ε、时态阈值 τ，则 $X \Rightarrow Y\left([T_1,T_2] \cap [T_1',T_2'], S, C\right)$ 是一条时态关联规则，当且仅当在 $[T_1,T_2] \cap [T_1',T_2']$ 上，$S \geq \sigma$，$C \geq \varepsilon$，$\left|[T_1,T_2] \cap [T_1',T_2']\right| \geq \tau$。

在给定 D, σ, ε 和 τ 的情形下，时态关联规则挖掘就是在时态数据库 D 中找出所有的强时态关联规则，与经典的 Apriori 算法相似，一般分为 2 步：(1)找出所有时态频繁项集；(2)应用时态频繁项集生成强时态关联规则。下面举例说明时态关联规则挖掘的过程。

设一个交易数据库，如表 1 所示，各交易项按发生的时间顺序排列。设定最小支持度 σ =0.6，最小信任度 ε =0.6，时态阈值 τ =3。以定义 1~定义 4 为基础，采用文献[6]的时态关联规则挖掘算法，挖掘结果如图 1 所示。

表 1　交易数据库 *D*

T	*tid*	*item set*
1	100	ABCE
2	200	ACDF
3	300	CD
4	400	AC
5	500	BCDE
6	600	ADF

*C*1			*L*1			*C*2			*L*2			*C*3
项集	支持度	生命周期	项集	支持度	生命周期	项集	支持度	生命周期	项集	支持度	生命周期	项集
A	0.67	[1,6]	A	0.67	[1,6]	AC	0.60	[1,5]	AC	0.60	[1,5]	∅
B	0.40	[1,5]	C	1.00	[1,5]	AD	0.40	[2,6]	CD	0.75	[2,5]	
C	1.00	[1,5]	D	0.80	[2,6]	CD	0.75	[2,5]				
D	0.80	[2,6]										
E	0.40	[1,5]										
F	0.40	[2,6]										

图 1　时态关联规则挖掘结果

图 1 中各项举例说明如下：如项集 A 的生命周期为 $[1,6]$，在该时间范围内，A 事务出现次数为 4，总事务数为 6，因此，$S(\mathrm{A}) = 4/6 = 0.67$。以此类推可得事务 A~事务 F 的支持度，如图 1 中 *C*1 项所列。在给定最小支持度 0.6 的条件下，可从 *C*1 即一阶候选频繁项集中选出 *L*1，一阶频繁项集 A, C, D，依照经典的 Apriori 算法，这 3 个项目 $\{\mathrm{A,C}\}, \{\mathrm{A,D}\}$ 和 $\{\mathrm{C,D}\}$ 组合成候选二阶频繁项集 *C*2。考察 $\{\mathrm{A,C}\}$ 这条规则，A 的生命周期为 $[1,6]$，*C* 的生命周期为 $[1,5]$，则 $\{\mathrm{A,C}\}$ 的生命周期取其交集为 $[1,5]$，有 $S(\mathrm{A} \Rightarrow \mathrm{C}) = 3/5 = 0.6$，其他同理可得。在给定最小支持度 0.6 的基础上，得出 *L*2 即二阶频繁项集 $\{\mathrm{A,C}\}, \{\mathrm{C,D}\}$。因为 $\{\mathrm{A,C,D}\}$ 中的一个子集 $\{\mathrm{A,D}\}$ 不在 *L*2 中，所以 $\{\mathrm{A,C}\}$ 和 $\{\mathrm{C,D}\}$ 不能组成三阶频繁项集 *C*3，算法到此结束。最后得到 2 条时态关联规则 $\{\mathrm{A,C}\}$ 和 $\{\mathrm{C,D}\}$，分别计算其信任度：

$$C(\mathrm{A} \Rightarrow \mathrm{C}) = \left|\mathrm{A} \Rightarrow \mathrm{C}[1,5]\right| \Big/ \left|\mathrm{A}[1,5]\right| = 3/3 = 1$$

$$C(\mathrm{C} \Rightarrow \mathrm{D}) = \left|\mathrm{C} \Rightarrow \mathrm{D}[2,5]\right| \Big/ \left|\mathrm{C}[2,5]\right| = 3/4 = 0.75$$

它们都满足最低信任度的要求，因此，都是有效的时态关联规则。

3　加权时态关联规则的提出

已有很多学者提出关联规则的加权思想[7]，其主要思想是区别数据库中不同项目的重要性，据此给予各项目不同的权重，并挖掘出最能引起用户兴趣的规则。本文的加权思想与其有所不同，在动态更新的时态数据库中给予较近发生的事务以较大的权重，体现时间的价值，从而尽可能地挖掘出最新最有用的时态关联规则。加权时态关联规则的主要思想是把数据库中最早发生的事务其发生时刻标志为 1，最近发生的事务时刻标志为 T，分别给予权重 W_i，其中，$W_1 < W_2 < \cdots < W_i < \cdots < W_T$。因此，对数据库作如下新定义：

定义 5 设 $I = \{i_1, i_2, \cdots, i_m\}$ 是项目的集合，时态数据可表示为 $D = \{D_1, D_2, \cdots, D_n\}$，其中，$D_i = \langle tid, itemset, T_i, W_i \rangle$，*tid* 是事务的标识，$itemset \subseteq I$，$T_i$ 是事务发生的时间表达式，W_i 为该事务的权重。

挖掘加权时态关联规则过程中，在计算事件的支持度和信任度时，不是简单地累加发生次数，而是依据该事件发生时刻的权重来加权累加。例如，事件A发生的生命周期为 $[T_m,T_n]$，其权重区间为 $[W_m,W_n]$，其中必有 $W_m<\cdots<W_i<\cdots<W_n$，发生A事件的时刻为 T_j，$j \in (m,n)$，权重为 W_j，则事件A的支持度为

$$S(\mathrm{A}) = \sum_{j \in (m,n)} W_j \Big/ \sum_{i=m}^{n} W_i$$

表 2 仍延续表 1 的结构，最后一列是给予不同项目集的权重，并按时间顺序递增。

表 2　带加权的交易数据库

T	*tid*	*item set*	W_i
1	100	ABCE	0.1
2	200	ACDF	0.2
3	300	CD	0.3
4	400	AC	0.4
5	500	BCDE	0.5
6	600	ADF	0.6

通过计算得

$$S(\mathrm{A}) = \sum_{j \in (1,6)} W_j \Big/ \sum_{i=1}^{6} W_i = \frac{0.1+0.2+0.4+0.6}{0.1+0.2+0.30+0.4+0.5+0.6} = 0.62$$

则规则 $\mathrm{A} \Rightarrow \mathrm{C}$ 的支持度 $S(\mathrm{A} \Rightarrow \mathrm{C}) = 0.467$。依此类推，得到如图 2 所示的挖掘结果。

*C*1			*L*1			*C*2			*L*2			*C*3
项集	支持度	生命周期	项集	支持度	生命周期	项集	支持度	生命周期	项集	支持度	生命周期	项集
A	0.620	[1,6]	A	0.620	[1,6]	AC	0.476	[1,5]	CD	0.714	[2,5]	∅
B	0.400	[1,5]	C	1.000	[1,5]	AD	0.400	[2,6]				
C	1.000	[1,5]	D	0.800	[2,6]	CD	0.714	[2,5]				
D	0.800	[2,6]										
E	0.400	[1,5]										
F	0.400	[2,6]										

图 2　加权时态关联规则挖掘结果

该例中的数据个数较少，因此，可采用逐个事务加权的方法。在大型时态数据库中，可采用分时段加权的方法，如对年分月来加权也能体现时间趋势。

相比用时态关联规则挖掘得到的结果，加权时态关联规则得到一条规则 $C \Rightarrow D$，其信任度为

$$C(C \Rightarrow D) = \sum_{j\in(2,5)} W_j \Big/ \sum_{i=2}^{5} W_i = 0.714$$

由于大于最低信任度 0.6，因此它是一条强规则。

为了体现加权时态关联规则的时间趋势性，本文尝试对不同时段数据进行挖掘，以表 2 带加权的交易数据库为例，前述过程挖掘了其中前 6 条事务中的规则，现在挖掘前 5 条事务中的规则，得到如图 3 所示的挖掘结果。

*C*1			*L*1			*C*2			*L*2			*C*3
项集	支持度	生命周期	项集	支持度	生命周期	项集	支持度	生命周期	项集	支持度	生命周期	项集
A	0.700	[1,4]	A	0.700	[1,4]	AC	0.700	[1,4]	AC	0.700	[1,4]	∅
B	0.400	[1,5]	C	1.000	[1,5]	AD	0.220	[2,4]	CD	0.714	[2,5]	
C	1.000	[1,5]	D	0.714	[2,5]	CD	0.714	[2,5]				
D	0.714	[2,5]										
E	0.400	[1,5]										
F	1.000	[2,2]										

图 3　不同时段的加权时态关联规则挖掘结果

其中，项目 F 虽然支持度较高但由于时态阈值的限制应被舍去；时态规则 $\{A \Rightarrow D\}$ 由于支持度小于阈值也被舍去，最后得到 2 条规则：$\{A \Rightarrow C\}$ 和 $\{C \Rightarrow D\}$，且有 $S(A \Rightarrow C)=0.7$，$S(C \Rightarrow D)=0.714$，都高于最小支持度，计算得 $C(A \Rightarrow C)=1$，$C(C \Rightarrow D)=0.714$。

比较图 2 和图 3 的挖掘结果可以发现，在 5 项事务的数据库中，挖掘出支持度分别为 0.7 和 0.714 的 2 条规则：$\{A \Rightarrow C\}$ 和 $\{C \Rightarrow D\}$。增加了一项事务后再次挖掘，发现 $\{A \Rightarrow C\}$ 的支持度降低，$\{C \Rightarrow D\}$ 的支持度不变。最后只挖掘出了一条符合要求的规则 $\{C \Rightarrow D\}$，该规则是当时最有时间价值的规则。从 $\{A \Rightarrow C\}$, $\{C \Rightarrow D\}$ 到 $\{C \Rightarrow D\}$ 的结果说明规则 $\{A \Rightarrow C\}$ 在数据库中的重要性逐渐降低。因此，频繁挖掘加权时态关联规则，不仅不会遗漏重要的关联规则，而且能体现规则的变化趋势，提供决策依据。

4　基于 3 种规则的挖掘结果比较

对文中所采用的数据表使用传统的关联规则发现算法 Apriori进行挖掘[8]，即不考虑事件的生命周期和时间权重，得到如图 4 所示的结果。可以看出，传统的关联规则算法没有挖掘出一条有效的规则，这说明该算法忽略项目集的时间属性可能造成的不良后果，是不能挖掘出数据库中一些带有时态约束的重要关联规则。时态关联规则对这一缺点进行了改进。而本文构造的加权时态关联规则更进一步。如果能在动态更新的时态数据库中定期且高频率地作加权时态关联规则挖掘，就能不断发现新的有用规则。

*C*1		*L*1		*C*2		*L*2
项集	支持度	项集	支持度	项集	支持度	项集
A	0.67(4/6)	A	0.67	AC	0.50(3/6)	∅
B	0.33(2/6)	C	0.83	AD	0.33(2/6)	
C	0.83(5/6)	D	0.67	CD	0.50(3/6)	
D	0.67(4/6)					
E	0.33(2/6)					
F	0.33(2/6)					

图 4　传统的时态关联规则挖掘结果

5　结束语

在知识发现的过程中，挖掘结果发现某些规则的支持度、信任度逐渐降低，说明该规则正在逐渐过时或失效，可能是该事件的周期性原因或其他外部环境变化的影响，应予以重视。用户无疑更重视新挖掘出的规则，因为它们体现一种新的趋势变化，用户可以借此对其商业策略作调整。需要说明的是，对于某些规律变动不大的数据库，如在医疗诊断信息库中，疾病和症状表现出的关联较强、规则较稳定，加权的思想不会影响该种数据库的挖掘结果。

由于时间是数据本身固有的因素，因此在挖掘关联规则时附加上某种时态约束会使发现的规则更好地描述实际情况，而对时间加权能更好地体现规则的时间适用性和趋势性，因而也会更有价值。今后的研究将对时态数据库的动态挖掘以比较发现关联规则的演进趋势及在大型数据库中更复杂有效的时间加权方法进行深入探讨。

参考文献

[1] Agrawal R, Mielinsi I T, Swami A. Mining Association Rules Between Sets of Items in Large Database[C]//Proc. of 1993 ACM SIGMOD Conference. Washington D C, USA: ACM Press, 1993.

[2] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules[C]//Proceedings of the 20th International Conference on Very Large Databases. Santiago, Chile: [s. n.], 1994-09.

[3] Ale J M, Rossi G H. An Approach to Discovering Temporal Association Rules[C]//Proc. of the 2000 ACM Symposium on Applied Computing. [S. l.]: ACM Press, 2000: 294-300.

[4] 欧阳为民, 蔡庆生. 在数据库中发现具有时态约束的关联规则[J]. 软件学报, 1999, 10(5): 527-532.

[5] 邓大权, 李　磊. 时态关联规则研究与应用[J]. 大连理工大学学报, 2003, 43(S3): S150-S154.

[6] 董祥军, 宋瀚涛, 姜　合, 等. 时态关联规则的研究[J]. 计算机工程, 2005, 31(15): 24-26.

[7] 欧阳为民, 郑　诚, 蔡庆生. 数据库中加权关联规则的发现[J]. 软件学报, 2001, 12(4): 612-619.

[8] Dunham M H. 数据挖掘教程[M]. 郭崇慧, 田凤占, 靳晓明, 等, 译. 北京: 清华大学出版社, 2005: 145-149.