文章编号:1005－0523(2005)02－0086－03

基于信任度的增量时态关联规则算法设计

罗来鹏,刘二根,王广超

(华东交通大学 基础科学学院,江西 南昌 330013)

摘要:对时态关联规则及其支持度和信任度进行了一般描述,推导了时态关联规则的一些增量信任度性质,这些性质在增量式设计的开采算法中,有着一定的指导作用.

关 键 词:数据挖掘;时态关联规则;增量式更新;信任度

中图分类号:TP311.13　　**文献标识码**:A

0 引 言

关联规则挖掘作为数据挖掘的一种重要模式,已成为数据挖掘领域的一个非常重要的研究课题.所谓关联规则挖掘是:从海量数据库中提取给定数据项集的有趣模式.在实践中,由于时间是现实数据库本身固有的因素,所以在数据中常常会发现时态语义问题,时态数据的出现使得有必要在关联规则挖掘过程中考虑时间因素,即时态的约束问题,这种关联规则,称为时态关联规则.为了解决这个问题,通过在数据模型中引入数据的时间属性,提出了时态数据库,支持有效时间和事务时间,支持时态信息查询和处理[3].在这样的情况下,体现关联规则有效性的两阈值支持度和信任度将是时间的函数,也就是说,关联规则的实用性和有效性将随时间变化而发生变化,因此,对于具有时态意义数据的规则挖掘附加上某种时态约束是很必要.

自从时态数据库提出以来,已有多种模型描述方法以及相应的挖掘算法[3,4,7,].由于时间是一个永远变化的量,因此如何去挖掘增量式的时态关联规则就显得尤为重要.它可适应大规模、动态数据、可实现并行处理等好处.目前一般有两种方法:一是接受所有的新数据,连同过去的旧数据一起,重新运用普通的关联规则的发现算法.该方法的缺点是要重新处理已经处理过的数据,不能有效地利用已经获得的结果;一是随着新数据的产生增量式地更新关联规则集,尽可能地只处理新的数据[1,6,7,8].显然,后者更可取.但也会产生挖掘的规则前后不兼容等问题.这与支持度和信任度的计算有关.本文从各时间段规则的信任度的计算角度出发,推导了增量信任度的一些性质,这些性质在增量式设计的开采算法中,为进一步提高了算法的效率,有着一定的指导作用.

1 时态关联规则[2]

设 $T=[t_1,t_2](t_2\leqslant+\infty)$ 为所考虑的时间段,并且 $T=\bigcup_{i=1}^{n}T_i$, $T_i\cap T_j=\Phi;(i\neq j;j=1,2,\cdots n)$, D 是该段时间内产生的所有事务(或交易)的集合,D_{T_i} 为在时间段 T_i 所发生的事务(或交易)的集合,显然 $D_{T_i}\subset D$,若 $x\subset D_{T_i}$, $y\subset D_{T_i}$ 则在 T_i 时间段关联规则 $x\Rightarrow y$,记为 $T_i(x\Rightarrow y)$

令 $S_{T_i}(X)(i=1,2\cdots)$ 表示在 T_i 时段内所有包

收稿日期:2004－05－10

作者简介:罗来鹏(1973－)男,江西吉水人,硕士, 讲师,研究方向:智能信息处理、数据挖掘.

含 X 的事务总数，$S^k(X)$表示在$\bigcup_{i=1}^{k} T_i$ 时间内所有包含 X 的事务总数，则

1) $\mathrm{Sup}_{T_i}(x\Rightarrow y)=\dfrac{S_{T_i}(xy)}{|D_{T_i}|}$ 为规则 $T_i(x\Rightarrow y)$的支持度.

2) $C_{T_i}(x\Rightarrow y)=\dfrac{S_{T_i}(xy)}{S_{T_i}(x)}$为规则 $T_i(x\Rightarrow y)$的信任度.

类似的，可以定义在 T 时间段关联规则 $x\Rightarrow y$ 为 $T(x\Rightarrow y)$，相应的：

1) $\mathrm{Sup}_T(x\Rightarrow y)=\dfrac{S_T(xy)}{|D_T|}(xy)$为规则 $T(x\Rightarrow y)$的支持度，或者 $Sup_T(x\Rightarrow y)=\dfrac{S^n(xy)}{|D_T|}$

2) $C_T(x\Rightarrow y)=\dfrac{S_T(xy)}{S_T(x)}$为规则 $T(x\Rightarrow y)$的信任度.或者 $C_T(x\Rightarrow y)=\dfrac{S^n(xy)}{S^n(x)}$

上述所描述的知识(规则)与事务数据库中的关联规则最大的不同是规则有着相应的时态约束，以表明所发现的的知识何时是有效的.

2　时态关联规则的性质

性质 1　$C_T(x\Rightarrow y)=\sum_{i=1}^{n} C_{T_i}(x\Rightarrow y)\dfrac{S_{T_i}(x)}{S_T(x)}$

性质 2　设 $\overline{C}=\max\{C_{T_i}(x\Rightarrow y)\mid i=1,2,\cdots,n\}$，$\underline{C}=\min\{C_{T_i}(x\Rightarrow y)\mid i=1,2,\cdots,n\}$，则 $\underline{C}\leqslant C_T(x\Rightarrow y)\leqslant\overline{C}$.

性质 2 说明在 T 时段规则 $x\Rightarrow y$ 的可信度 $C_T(x\Rightarrow y)$是位于在 $T_1,\cdots T_n$ 时段内规则 $x\Rightarrow y$ 可信度的最大值与最小值之间.

性质 3　$C_{T_i\cup T_{i+1}}(x\Rightarrow y)=\eta C_{T_i}(x\Rightarrow y)+(1-\eta)C_{T_{i+1}}(x\Rightarrow y)$.这里，$\eta=\dfrac{\mathrm{Sup}_{T_i}(x)}{\mathrm{Sup}_{T_i\cup T_{i+1}}(x)}$

证明类似文[9]

该性质反应了两相邻时间段同一关联规则的信任度和整个时间段的信任度的关系.在具体的算法设计中，只需计算新时间段规则的信任度，然后直接由性质 3 可得整体的信任度，这显然比重新进行搜索计算要简洁得多.

性质 4　设 T 由无穷多个时间段构成：$T=\bigcup_{i=1}^{+\infty} T_i$，对任意的 $T_i(i=1,2,\cdots)$和 $x\in D$，有 $0<s<S_{T_i}(x)/|D_{T_i}|$，s 为用户定义的最小支持度，若 $\lim_{k\to+\infty} C_{T_k}(x\Rightarrow y)$存在，则：

$$\lim_{k\to+\infty} C_{\bigcup_{i=1}^{k} T_i}(x\Rightarrow y)\,C_{T_k}(x\Rightarrow y)$$

证明：由 $0<s<S_{T_i}(x)/|D_{T_i}|(i=1,2,\cdots)$知：

$$\lim_{k\to+\infty} S_{\bigcup_{i=1}^{k} T_i}(x)=\lim_{k\to+\infty}\sum_{i=1}^{k} S_{T_i}(x)=\lim_{k\to+\infty} S^k(x)\to+\infty$$

设 $\lim_{k\to+\infty} C_{T_k}(x\Rightarrow y)=c$，则对于 $\forall\,\varepsilon>0$，存在正整数 N，当 $k>N$ 时，有

$$\left|\frac{S_{T_k}(xy)}{S_{T_k}(x)}-c\right|<\varepsilon$$

即 $S_{T_k}(x)(c-\varepsilon)<S_{T_k}(xy)<S_{T_k}(x)(c+\varepsilon)$

当 $k>N$ 时，得到：

$$\left(\sum_{i=1}^{k} S_{T_i}(x)-\sum_{i=1}^{N} S_{T_i}(x)\right)(c-\varepsilon)<\sum_{i=1}^{k} S_{T_i}(xy)-\sum_{i=1}^{N} S_{T_i}(xy)<\left(\sum_{i=1}^{k} S_{T_i}(x)-\sum_{i=1}^{N} S_{T_i}(x)\right)(c+\varepsilon)$$

即

$$\left|\frac{\dfrac{S^k(xy)}{S^k(x)}-\dfrac{S^N(xy)}{S^k(x)}}{1-\dfrac{S^N(x)}{S^k(x)}}-c\right|<\varepsilon \tag{1}$$

由 $S_{T_k}(xy)\leqslant S_{T_k}(x)(k=1,2,\cdots)$，因此有：

$$\lim_{k\to\infty}\frac{S^N(xy)}{S^k(x)}=\lim_{k\to\infty}\frac{S^N(x)}{S^k(x)}=0 \tag{2}$$

1)和 2)式令 $k\to+\infty$取极限得：$\lim_{k\to\infty} C_{\bigcup_{i=1}^{k} T_i}(x\Rightarrow y)=c$.

这是在许多商业数据开采算法中常会发生的一种时态数据特性.

事实上，$C_{\bigcup_{i=1}^{k} T_i}(x\Rightarrow y)$可以看成是时间段$\bigcup_{i=1}^{k} T_i$上的各个时间段的信任度的一种均值，即：

$$C_{\bigcup_{i=1}^{k} T_i}(x\Rightarrow y)=\frac{S^k(xy)}{S^k(x)}=\frac{1}{k}\sum_{i=1}^{k} S_{T_i}(xy)\div\frac{1}{k}\sum_{i=1}^{k} S_{Y_i}(x)$$

也就是说，在一定条件下，随着时间的变化，规则 $x\Rightarrow y$ 若在每个时间段信任度稳定地趋向某一个固定值，则在整个时间段上的信任度趋向于同一个值.

性质 3,4 反应了在新的时间段信任度和整个时间段的信任度的关系.这种关系使得在按增量式设计的开采算法中，能充分利用了已挖掘的知识信息，直接进行计算.

3 结束语

如何在时态数据库上进行知识挖掘，是一个十分前沿的课题.时态的引入将极大地扩充了关联规则在实际中的应用，比如可以通过对多种股票进行单属性挖掘分析，这样所得到的时态关联规则对于短期的预测或决策是有着非常重要的作用.本文就时态约束的增量关联规则挖掘问题，在理论上证明了相邻时间段信任度之间的关系和具有稳定信任度的时态关联规则在增量式算法设计中，计算到一定的时间后可以停止进行计算.这些性质为在增量开采的算法设计中进一步提高算法的效率提供了一定的理论基础.

参考文献：

[1] 马元元，孙志军，高红梅. 时态数据库中增量关联规则的挖掘[J].计算机研究与发展，2000，37(12)：1446～1451

[2] 欧阳为民，蔡庆生.基于时间窗口的增量式关联规则更新技术[J].软件学报，1999，10(4)：426～429.

[3] 唐常杰，于中华，等.时态数据的变粒度分段存储策略及其效益分析[J].软件学报，1999，10(10)：1085～1090.

[4] 贾超.不明确时间间隔的表示及时态运算的扩展[J].计算机工程，2002，28(8)：123～124.

[5] Orlando S，Palmerini P，Perego R.Enhancing the Apriori Algorithm for Frequent Set Counting. DaWak 2001，LNCS2114，2001：512～521.

[6] 任家东，任东英.基于时间戳数据库的分布式多层时态关联规则挖掘[J].计算机工程，2004，30(6)：63～64.

[7] 毛国君，刘椿年.时态约束下的数据挖掘问题及算法[J].电子学报，2003(11)：1690～1694.

[8] 辜炜东，汤庸，等.事务数据库中的时态信息挖掘[J].计算机工程与应用，2004(18)：174～177.

[9] 罗来鹏，刘二根，等.基于包含度的事务数据库关联规则挖掘[J].华东交通大学学报，2004，21(5)：23～25.

Algorithm Design of Incremental Temporal Association Rules Based on Confidence Degree

LUO Lai-peng，LIU Er-gen，WANG Gang-chao

(School of Natural Sciences，East China Jiaotong University ，Nanchang 330013 ，China)

Abstract：Temporal association rules and its support degree and confidence degree are described in this paper，and some properties of incremental confidence degree about temporal association rules is deduced. These properties have leading function for mining algorithm design of incremental data.

Key words：data mining；temporal association rules；incremental update；confidence degree