文章编号: 1001-0920(2007)03-0278-06

持续时态数据挖掘的研究

潘　定[1,2], 沈钧毅[2]
(1. 暨南大学 管理学院, 广州 510632; 2. 西安交通大学 计算机系, 西安 710049)

摘　要: 基于一阶线性时态逻辑,形式化定义时态数据挖掘中的主要概念,利用线性状态结构对每个时间点上的一阶语言符号进行赋值,并度量公式的真值范围. 按照挖掘段概念,开发持续挖掘过程模型,用于归纳局部一阶规则与推导高阶规则. 基于信息扩散原理,提出一阶规则的度量值估计方法和规则泛化算法. 最后通过算例说明了扩散估计和算法的有效性.

关键词: 时态数据; 持续数据挖掘; 信息扩散; 高阶挖掘
中图分类号: TP18　　**文献标识码**: A

Research on continuous temporal data mining

PAN Ding[1,2], *SHEN Jun-yi*[2]
(1. School of Management, Ji'nan University, Guangzhou 510632, China; 2. Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China. Correspondent: PAN Ding, E-mail: pandingcn@gmail.com)

Abstract: The definitions of main notions used in temporal knowledge discovering are proposed in a formal way, which is based on first-order linear temporal logic. The concept of linear state structure allows associating each time moment with an valuation of all symbols of a first-order language, and measures the extent of truth of a formula. According to the notion of session mining, a continuous data mining process model is developed for inducing the local first-order rules and inferring higher order rules. Based on the principle of the information diffusion, the estimation for the measures and an algorithm for rule generalization are presented. The simulations show the effectiveness of the methods.

Key words: Temporal data; Continuous data mining; Information diffusion; High order mining

1 引　　言

客观世界中的数据是不断扩充和持续变化的,人们对规律的认识和评价也在不断发展. 带有时间维度的数据形成时态数据,如股市交易、超市销售、Web 点击流等. 持续的时态数据挖掘(TDM)跟踪时态数据变化,试图及时反映数据的演化特征.

Agrawal 等[1]首先提出了持续挖掘的概念,描述了一种主动挖掘过程. Cheng 等[2]提出增量挖掘算法. Gupta 等[3]提出一种以用户为中心的挖掘过程模型,分为模式设计、累积、挖掘 3 个阶段. 高阶数据挖掘[4]试图从已发现的规则中发现有趣的高层次语义规则,即高阶规则. 对 TDM 的一般框架研究[5]主要着重于挖掘方法和算法,很少涉及理论研究. Cotofrei 等[6]基于一阶时态逻辑给出时间序列规则的形式化定义. 黄崇福等[7]提出的信息扩散原理是一种模糊信息优化处理技术.

传统的挖掘过程局限于按照特定挖掘目的,组织数据并执行挖掘的过程模型. 随着数据的急速增长,需要开发支持自动、持续时态数据挖掘的过程模型. 本文基于一阶线性时态逻辑,扩展文献[6]的工作,对时态数据挖掘的主要概念进行形式化,描述一阶公式赋值及其度量的定义,给出数据表示、规则和分段挖掘方式的抽象视图,其形式描述不依赖于具体的表示和挖掘算法;然后基于挖掘段概念,提出一种新颖的数据挖掘过程模型,试图实现持续的挖掘过程;最后,提出利用信息扩散原理估计一阶规则的度量值和规则泛化的算法.

2 时态序列问题

收稿日期: 2005-11-11; **修回日期**: 2006-02-16.
基金项目: 国家自然科学基金项目(70372024, 60173058); 广州市科技计划项目(2004Z3-D0351, 2006Z3-D3101).
作者简介: 潘定(1963—),男,江苏宝应人,博士,从事数据挖掘、数据仓库等研究; 沈钧毅(1939—),男,江苏常熟人,教授,博士生导师,从事数据库、数据挖掘等研究.

时态数据涉及的时间表示和记录方式可有多种形式,本文选择一种线性有序的离散结构 $T = (T_D, <)$ 作为时态域,其中: T_D 是可数时间点集合, $<$ 是线性序关系. 为方便处理,假设 T 中的时间点升序排列且 $t_{i+1} - t_i = \Delta$ 为常数.

定义 1　对时间域 T,非空状态属性集 $\{A_1, \cdots, A_k\}$ 及其对应值域 D_{A_i},一个时态序列是有序项队列 $X = \{X_1, X_2, \cdots, X_m\}$,其中 X_i 是一个 $k+1$ 元组 $(t, a_1, \cdots, a_k)$, $t \in T$, $a_i \in D_{A_i}$.

当 $D_{A_i} \subseteq R$ 时, X 是一个时间序列;当 $D_{A_i} \subseteq \Sigma$, Σ 是字母表时, X 是一个事件序列;当 $D_{A_i} \subseteq R^+$ 或 $D_{A_i} \subseteq \{1,0\}$ 时, X 是一个交易序列. 某类序列的所有实例构成序列空间 W_x.

状态属性集 $\{A_1, \cdots, A_k\}$ 描述有限实体集 Q(如客户、股票集) 的各种属性. 对于交易序列,状态属性集是 k 个交易项目的状态,相应的值域是交易量或成交标志. TDM 的挖掘对象序列集 $w_x \subseteq Q \times W_x$.

在运行挖掘算法前,假设经过序列的预处理已将 w_x 中的时态序列 X 变换成由若干典型、有趣形状或符号串构成的线性状态序列 $S = (S_1, \cdots, S_k)$.

定义 2　给定有限形状或符号串的标识集合 D_e,特征函数集 $\{f_1, \cdots, f_p\}$ 及其对应值域 D_{f_i}, $p \geq 0$,则 w_x 的状态集 $E_s = \{(e, b_1, \cdots, b_p) \mid e \in D_e, b_i \in D_{f_i}, b_i = f_i\}$.

在归纳分类时,存在一个有限类标签集合 D_g. 当以有监督方式归纳时,应事先指定映射 $w_x \rightarrow D_g$, w_x 为训练集;当以无监督方式归纳时,算法导出映射.

例 1　对股票日交易均价数据库,有趣形状集 $D_e = \{$高峰,低谷,平坦$\}$,特征函数 f_1 为 5 日均值, f_2 为均差,则某股票序列经变换后的状态序列为 ((平坦,3, 1.5),(高峰,10,2.4), $\cdots$, (高峰,8, 1.5)). 对股票分类 $D_g = \{$高成长,波动,高风险$\}$,设有规则(其中符号意义参见下节)

$$R_1: X_{-4}(t_e = \text{高峰}) \wedge X_{-4}(t_1 > 8.0) \wedge X_{-3}(t_e = \text{高峰}) \Rightarrow X_0(t_e = \text{平坦}). \quad (1)$$

3　时态数据挖掘的形式化

按形式化 TDM 的需要,使用一种有约束的一阶线性时态逻辑语言 L,包括常量符号、函数符号、谓词符号、关系符号 $\{=, <, \leq, >, \geq\}$、逻辑连接符号 $\{\wedge\}$、时态连接符号 X_k ($k \in Z, k > 0$ 表示未来, $k < 0$ 表示过去, $k = 0$ 表示现在[6]).

3.1　语　　法

由语言 L 的表达式可定义术语、原子公式(简称原子), 公式的集合, 分别记为 Term(L)、Atom(L) 和 Form(L).

定义 3(Term(L))　$t \in$ Term(L),当且仅当 t 由(有限次使用) 以下规则形成:

1) $a, u \in$ Term(L),其中 a 和 u 分别是常量、变量符号;

2) 若 $t_1, t_2, \cdots, t_n \in$ Term(L),且 f 是 n 元函数符号,则 $f(t_1, t_2, \cdots, t_n) \in$ Term(L).

定义 4(Atom(L))　$A \in$ Atom(L),当且仅当 A 由以下任一规则形成:

1) $R(t_1, t_2, \cdots, t_m)$, 其中: R 是谓词, $t_i \in$ Term(L);

2) $t_1 \rho t_2$,其中: $t_1, t_2 \in$ Term(L), $\rho \in \{=, <, \leq, >, \geq\}$,也称为关系原子.

定义 5(Form(L))　$F \in$ Form(L),当且仅当 F 由(有限次使用) 以下规则形成:

1) Atom(L) $\in$ Form(L);

2) $(F_1 \wedge F_2) \in$ Form(L), 其中 $F_1, F_2 \in$ Form(L);

3) $X_k F \in$ Form(L),其中: $F \in$ Form(L), X_k 是时态连接符号.

在 Horn 子句中,逻辑蕴涵 ($\rightarrow$) 符左边的蕴涵式为假时,子句仍为真,所以对本文的时态规则不适用. 因此,引入连接符号 $\Rightarrow$,形成类 Horn 子句: $A_1 \wedge A_2 \wedge \cdots \wedge A_k \Rightarrow A_{k+1}$,其中 A_i 是正原子公式. 此时类 Horn 子句与 $A_1 \wedge A_2 \wedge \cdots \wedge A_k \wedge A_{k+1}$ 等价.

定义 6　一个事件是一个 $p+1$ 元谓词 $E(e, f_1, \cdots, f_p)$,其中: e 是常量符号表示事件名称, $f_1, \cdots, f_p$ 是函数符号, $p \geq 0$. 事件 $E(e, f_1, \cdots, f_p) \in$ Atom(L).

定义 7　一个事件 $E(e, f_1, \cdots, f_p)$ 的约束公式是一个合取式 $C_1 \wedge C_2 \wedge \cdots \wedge C_m$,其中: C_i 是关系原子 $t_e = a$ 或 $t_j \rho b$, t_e 是事件名称的变量符号, t_j 是对应 f_j 的变量符号, $1 \leq j \leq p$, $\rho \in \{=, <, \leq, >, \geq\}$, a 和 b 是常量符号. 时态约束公式 H_k 表示 $X_k(C_1 \wedge C_2 \wedge \cdots \wedge C_m)$, $k \in Z$.

定义 8　一个子序列是一系列事件的合取式 $E_{i_1} \wedge E_{i_2} \wedge \cdots \wedge E_{i_m}$,其中 E_{i_k} 表示 $X_{i_k}(E(e, f_1, \cdots, f_p))$, $e \in D_e$, $1 \leq k \leq m$, $i_1 < i_2 < \cdots < i_m$.

定义 9　一个序列模式(或模式) 是一系列时态约束公式的合取式 $H_{i_1} \wedge \cdots \wedge H_{i_m}$,其中 $i_1 < \cdots < i_m$.

定义 10　一个时态规则的形式为 $H_{i_1} \wedge H_{i_2} \wedge \cdots \wedge H_{i_{m-1}} \Rightarrow H_{i_m}$,其中: $i_1 < \cdots < i_m$,且 H_{i_m} 中含有关系原子 $t_e = a$, $H_{i_1} \wedge \cdots \wedge H_{i_{m-1}}$ 称为规则体, H_{i_m} 称为规则头.

对全局数据特征,一个类别是一个谓词 $G(g)$

$\in \mathrm{Atom}(L)$,其中 $g \in D_g$ 是有限类标签.

定义 11 一个分类规则的形式为 $H_{i_1} \wedge H_{i_2} \wedge \cdots \wedge H_{i_m} \Rightarrow G(g)$,其中 $i_1 < \cdots < i_m$.

对子序列、模式、时态规则和分类规则统称为一阶规则 R_F,$|i_m - i_1|$ 是 R_F 的时间间隔. 当仅需考虑事件发生的顺序时,R_F 中可省略符号 X_i.

3.2 语 义

对一阶语言的公式必须通过解释来表示命题. 通常,一阶语言的解释是基于结构 $U = (D, \{a^i\}, \{f^i\}, \{R^i\})$ 的. 其中:D 表示论域;a^i, f^i, R^i 分别表示 D 中的常量、全函数和谓词. 对语言的解释,即使用解释函数分别将常量符号、函数符号和谓词符号映射到相应的 a^i, f^i, R^i 上. 此外,还需将 D 中的个体指派给经解释后的自由变量符号. 若将解释和指派统称为赋值,则对一阶语言的赋值 V 有:1) $V(a), V(u) \in D$; 2) $V(f): D^n \to D$; 3) $V(R): D^m \to \{\mathrm{ture}, \mathrm{flase}\}$. 赋值可扩展到任意表达式,如 $V(f(t_1, t_2, \cdots, t_n)) = V(f)(V(t_1), \cdots V(t_n))$. $V| = p$ 表示 $V(p)$ 取真值.

在 TDM 环境中,对给定的 $D = w_x \cup D_e \cup D_f \cup D_g$,特征函数集 $\{f_1, \cdots, f_p\}$ 是定义在 D 上的全函数. 为定义语言 L 的一阶线性时态逻辑,还需要具有时态维度、能对特定时点赋值的结构.

定义 12 对语言 L 和论域 D,L 的线性状态结构是 5 元组 $M = (U, E_s, Tr, \Omega, V)$. 其中:$U = (D, \{a^i\}, \{f^i\}, \{R^i\})$,$E_s$ 是状态集合,$Tr: w_x \to N \times E_s$ 将序列 X 映射成状态序列 $(S_{(1)}, \cdots, S_{(i)}, \cdots)$,$\Omega$ 是状态序列集,V 是为每个状态 $S_{(i)}$ 点所有符号赋值的函数.

给定线性状态结构 M,有 $\Omega = \{S^1, \cdots, S^k, \cdots, S^n\}$,将序列 S^k 在状态 $S_{(i)}$ (或 X_i) 时的 $V| = p$ 表示为 $(M, S^k, i)| = p$. 对时态规则,当规则体和头分属于序列 S^j 和 S^k 时,记为 $(M, S^{j,k}, i_1)| = H_{i_1} \wedge H_{i_2} \wedge \cdots \wedge H_{i_{m-1}} \Rightarrow H_{i_m}$. 当不用区分 M 和 $S^{j,k}$ 时可简单地记为 $i_1| = p$. 以下仅就规则体和头属同一序列时展开讨论.

当且仅当 $i| = p$ 且 $i| = q$ 时,$i| = p \wedge q$. 当且仅当 $i + k| = p$ 时,$i| = X_k p$. 因此,$i| = E(e, f_1, \cdots, f_p)$ 表示在状态 $S_{(i)}$ 序列出现了一个名称为 e,特征 $f_1, \cdots, f_p$ 的事件. 类似地,当且仅当所有 $i| = C_j$ 时,时态约束公式在状态 $S_{(i)}$ 为真;当且仅当所有 $i| = E_{i_k}$ 时,子序列在状态 $S_{(i)}$ 为真;当且仅当所有 $i| = H_{i_k}$ 时,模式在状态 $S_{(i)}$ 为真;当且仅当 $i| = H_{i_1} \wedge \cdots \wedge H_{i_{m-1}}$ 且 $i| = H_{i_m}$ 时,时态规则在状态 $S_{(i)}$ 为真;$i| = G(g)$ 表示在状态 $S_{(i)}$ 时序列属于 g 类;当且仅当 $i| = H_{i_1} \wedge \cdots \wedge H_{i_{m-1}}$ 且 $i| = G(g)$ 时,分类规则在状态 $S_{(i)}$ 为真,表示当具备 $H_{i_1} \wedge \cdots \wedge H_{i_m}$ 特征时,序列属于 g 类.

基于序列集 Ω,可建立一些度量来衡量 $V| = p$ 的范围程度. 假设对语言 L 的 p,存在一个算法,可通过对现有数据的有限步计算获得赋值 $V(p)$.

定义 13 给定语言 L 和线性状态结构 M,对每个 p,序列集合 Ω,定义实值集函数 $P(p) = |A|/n$. 其中 $n = |\Omega|$,$A = \{k \in \{1, \cdots, n\} | (M, S^k, i)| = p\}$.

定理 1 对于语言 L 和线性状态结构 M,p 的实值集函数 $P(p) = \frac{|A|}{n}$ 是在 Ω 中出现 $V| = p$ 的概率.

证明 将基于 M 获得的序列集合 $\Omega = \{S^1, \cdots, S^k, \cdots, S^n\}$ 作为样本空间,取 $F = 2^{\Omega}$,从而 $\Omega \in F$,Ω 的任意子集 $Q \in F$,则 F 是一个 σ 代数.

对于公式 p,设 $Q = \{S^k | (M, S^k, i)| = p\}$,$A = \{k | S^k \in Q\}$,则 $|A| \geq 0$,$P(p) \geq 0$;又因 $Q = \Omega$ 时,$|A| = n$,则 $P(p) = 1$. 又设 $|A_j| = K_j (\leq n)$,则对应的 $P(p_j) = K_j/n (j = 1, 2, \cdots, m)$. 若这些 A_j 不相交,则对相应的 p_j,有

$$P\left(\sum_{j=1}^{m} p_j\right) = \left(\sum_{j=1}^{m} K_j\right)/n = \sum_{j=1}^{m} \frac{K_j}{n} = \sum_{j=1}^{m} P(p_j). \quad (2)$$

因此,(Ω, F, P) 是一个概率空间. □

定义 14 给定语言 L 和线性状态结构 M,公式 p 的 $V| = p$ 的一个范围度量是函数 $\mathrm{supp}(p) = P(p)$. 这个度量通常称为 p 的支持度.

对时态规则,还可以定义另一个度量来描述规则体和头之间蕴涵的范围程度.

定义 15 给定语言 L 和线性状态结构 M,对时态规则 p 的 $V| = p$,有范围度量函数 $\mathrm{conf}(p) = P(p)/P(p_b)$,其中 p_b 是规则体,当 $P(p_b) = 0$ 时,$\mathrm{conf}(p) = 0$. 这个度量通常称为规则 p 的信任度.

子序列、模式和时态规则通过支持度、信任度说明序列局部数据特征. 分类规则对全局数据特征进行归纳,且若干条分类规则可对应一种分类. 分类规则的支持度度量类 $g_i \in D_g$ 的数据范围. 对一个序列属于多种分类的情况,按规则优先度确定.

3.3 分段有限模型

在现实环境中,经常难以获得整个序列数据,或只能在有限区间的序列上进行挖掘. 因此,应在分段有限状态子序列上估计度量函数.

定义 16 给定语言 L 和线性状态结构 M,M 的模型是一个结构 $\tilde{M} = (\tilde{Q}, s_1)$,其中 s_1 是分段长度,$\tilde{\Omega}$

$= \{\widetilde{S}^k = \{S^k_{i_1}, S^k_{i_2}, \cdots, S^k_{i_{s_1}}\} \mid \widetilde{S}^k \subseteq S^k, 1 \leq k \leq n\}$.

定义 17　给定语言 L 和 M 的模型 $\widetilde{M}$,公式 p 的 $\mathrm{supp}(p)$ 的估计量定义为 $\mathrm{ES}(p, \widetilde{M}) = |A^e| / n$,其中 $|\widetilde{\Omega}| = n, A^e = \{k \in \{1, \cdots, n\} \mid (\widetilde{M}, S^k, i) \mid = p\}$.

定义 18　给定语言 L 和 M 的模型 $\widetilde{M}$,时态规则 p 的 $\mathrm{conf}(p)$ 的估计为 $\mathrm{EC}(p, \widetilde{M}) = \mathrm{ES}(p, \widetilde{M}) / \mathrm{ES}(p_b, \widetilde{M})$,其中 p_b 是规则体. 当 $\mathrm{ES}(p_b, \widetilde{M}) = 0$ 时,$\mathrm{EC}(p, \widetilde{M}) = 0$.

定义 19　给定语言 L 和 M 的模型 $\widetilde{M}$,基于 $\widetilde{M}$ 的挖掘段是一个 6 元组 $DM_F = (\mathrm{Task}, T_{st}, w_x, \mathrm{Stat}, \mathrm{DK}, R)$,其中: Task 是挖掘任务, T_{st} 是挖掘起始点, w_x 是数据集, Stat 是估计量阈值, DK 是领域知识, $R = \{r \in R_F \mid \mathrm{Stat}(r) \wedge \mathrm{DK}(r)\}$.

$\widetilde{M}$ 将 Ω 分为多个长度为 s_1 的子序列集合,因此,对公式 p 将形成支持度估计量序列 $\mathrm{ES}_1, \mathrm{ES}_2, \cdots, \mathrm{ES}_r, \cdots$ 和规则信任度的估计量序列 $\mathrm{EC}_1, \mathrm{EC}_2, \cdots, \mathrm{EC}_r, \cdots$.

定义 20　给定语言 L 和线性状态结构 M,若对序列 S 和公式 p,极限 $\lim\limits_{m\to\infty} |B| / m$ 存在,其中 $B = \{i \in \{1, \cdots, m\} \mid (M, i) \mid = p\}$,则序列 S 对 p 一致. 当序列集中序列皆为对 p 一致时,称为 p 一致序列集.

当公式满足一致性,支持度估计量是相合的.

定理 2　给定语言 L 和线性状态结构 M,若公式 p 有一致序列集 Ω,则对公式 p,当模型 $\widetilde{M}$ 的 s_1 足够大时, $\lim\limits_{r\to\infty} \mathrm{ES}_r = P(p) = |A| / n$,其中 $\Omega = \{S^1, \cdots, S^n\}, A = \{k \in \{1, \cdots, n\} \mid (M, S^k, i) \mid = p\}$.

证明　因为 $\Omega = \{S^1, \cdots, S^n\}$ 是一致序列集,则公式 p 对 Ω 中任意 S^k,有 $\lim\limits_{m\to\infty} |B| / m = \alpha_k$,其中 $B = \{i \in \{1, \cdots, m\} \mid (M, i) \mid = p\}$. 这样,对 Ω 有 $\alpha(p) = \{\alpha_1, \cdots, \alpha_n\}$. 设 $\alpha = \min(\{\alpha_j \in \alpha(p) \mid \alpha_j > 0\})$,取 $s_1 = \max(1/\alpha, p$ 的时间间隔$)$.

设 $\widetilde{M} = (\widetilde{\Omega}, s_1)$,支持度估计序列是 $\mathrm{ES}_1, \cdots, \mathrm{ES}_r, \cdots$: 显然 $A^e_r \subseteq A$. 对任意 S^j,若 $j \in A = \{k \in \{1, \cdots, n\} \mid (M, S^k, i) \mid = p\}$,即 $\alpha_j > 0$,当 r 足够大时,对子序列 $\widetilde{S}^j = \{S^j_{i_1}, S^j_{i_2}, \cdots, S^j_{i_{s_1}}\}$ 必有 $|i \in \{r_1, \cdots, r_{s_1}\} \mid (\widetilde{M}, \widetilde{S}^j, i) \mid = p\}| > 0$,则 $j \in A^e_r$,即 $A \subseteq A^e_r$. 因此

$$\lim_{r\to\infty} \mathrm{ES}_r = \lim_{r\to\infty} \frac{|A^e_r|}{n} = \frac{1}{n} \lim_{r\to\infty} |A^e_r| = \frac{|A|}{n} = P(p). \qquad \square$$

3.4　高阶数据挖掘

基于 $\widetilde{M}$ 的挖掘段结果产生出一阶规则的估计量序列,这些估计量由数据库的大量相关数据归纳计算而来,具有特定的含义,从不同的角度刻画一阶规则的特征,称此类序列为度量值序列. 度量值序列实际上是一阶规则度量值的时间序列.

在一般情况下,度量值序列具有上升、下降、波动和平行 4 种基本演化趋势. 高阶规则可用于描述度量值序列的动态特征,其语法形式与以上所定义的一阶规则相同. 高阶数据挖掘的主要任务有趋势分析、关联分析、聚类和结构演化分析[4].

定义 21　给定语言 L 和 M 的模型 $\widetilde{M}$,高阶数据挖掘的论域 $D = w_S \cup D_{es} \cup D_{fs} \cup D_{gs}$. 其中: $w_S = \{S(p) \mid p \in R_F\}$,估计量序列 $S(p) = \{\mathrm{ES}_1, \mathrm{ES}_2, \cdots, \mathrm{ES}_r\}$, D_{es} 是有限形状标识集, D_{fs} 是特征函数值域, D_{gs} 是有限类标签集合.

规则的多层次归纳如图 1 所示. 高阶规则的语义通常应结合一阶规则的归纳任务来确定. 相应于一阶规则的关联、分类和聚类,有以下非形式解释:

1) 趋势分析归纳关联、分类和聚类的变化趋势,如 20 岁年龄组客户中购买书籍的人明显增多;

2) 关联分析显示出关联关系间变动的相关性和多种分类间的关联,如白领购书行为变得难以区分;

3) 聚类识别一阶规则变化趋势的相似性,如 50 岁组白领与 30 岁组女性的购买习惯趋于相似;

4) 规则结构演化总结各种规则构成的演化趋势,如买面包和牛奶的客户变成购买水果和牛奶.

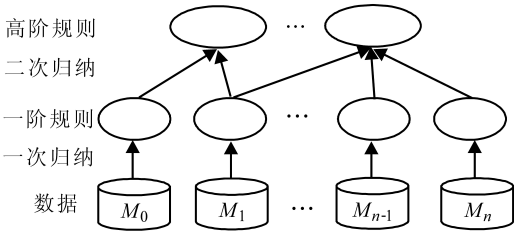

图 1　规则的多层次归纳

形成度量值序列时必须保证: 1) 序列项是可比较的,应按分段进行挖掘并标准化; 2) 每个挖掘段都存在序列项,若有缺失,可用缺省值填充.

4　持续时态数据挖掘的过程模型

基于挖掘段的思路,建立一种内在的机制,实现在不断扩充的数据环境中的持续数据挖掘过程,同时利用本体服务集成领域知识[8].

持续数据挖掘过程 (C-DM) 模型如图 2 所示. 挖掘过程由 4 个阶段构成: 计划、期间挖掘、合并挖掘和后处理. C-DM 强调分段、局部的自动挖掘,并通过规则合并 (高阶挖掘) 快速提供结果.

在计划阶段,挖掘过程开始于数据探索. 经过交互探索和试验,识别挖掘目标、业务数据和后续处

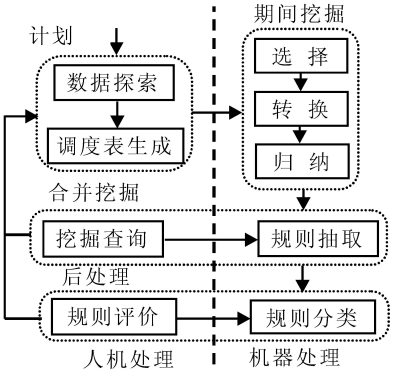

图 2 G-DM 处理模型

理步骤,生成挖掘任务调度表(TS).

期间挖掘执行选择-转换-挖掘段处理,实现局部挖掘,其重点是一阶规则归纳.这些步骤按 TS 规定的触发条件或频率周期性地对增量数据运行,自动执行后(每次)结果是一个规则箱(RB).

合并挖掘功能由挖掘查询/应用启动,以联机和交互方式进行全局、动态的规则发现(高阶挖掘),从多个 RB 中合并、求精跨期间的时态规则.

后处理将发现的规则与领域知识进行匹配,过滤无用规则,自动排序有趣规则,最后将规则返回给用户.当不能满足需求时,可重返计划阶段.

本体服务(OS)可提供多 agent 环境中处理显式的、陈述性表示本体的处理机制[9].在此利用 OS 集成挖掘需要的领域知识.为了实验该框架的可行性,构造一个原型系统[8],限于篇幅,不再详述.

基于挖掘段思路的期间挖掘,其结果形成度量值序列.确定分段长度(s_1)与具体应用相关.

例 2 对例 1 的股票日交易均价数据库,将分段长度(或期间)定为月,则对每月股票均价归纳形成一阶规则,如例 1 的 R_1 形式.若干月份的 R_1 支持度形成度量值序列,如(1,0.43),(2,0.45),…,(9,0.55).对此时间序列进行合并挖掘可获得高阶规则.以下将讨论一些合并挖掘技术.

5 度量值的扩散估计与规则泛化

由期间挖掘获得的度量值序列提供了用统计方法估计度量值(如支持度)的有限样本.利用信息扩散原理的扩散估计将获得较好的估计结果.

设 W 为母体 U 的样本,当由 W 不能完全精确地认识 U 的概率密度 $f(x)$ 时,称 W 是非完备的.信息扩散原理[7]为:设 $W=\{w_1,w_2,\cdots,w_n\}$ 是样本,V 是基础论域,w_j 的观测值为 v_j,令 $x=\phi(v-v_j)$,则 W 非完备时,存在函数 $\mu(x)$,使 v_j 点获得的量值为 1 的信息可按 $\mu(x)$ 的量值扩散到 v,且扩散所得的原始信息分布 $Q(x)=\sum_{j=1}^{n}\mu(\phi(v-v_j))$ 能更好地反映 U 的规律.

定义 25[7] 设 $\mu(x)$ 为定义在 $(-\infty,\infty)$ 上的一个波雷尔可测函数,$d>0$ 为常数,n 为样本数,则称

$$\hat{f}(v)=\frac{1}{nd}\sum_{j=1}^{n}\mu\left[\frac{v-v_j}{d}\right] \tag{3}$$

为母体的概率密度函数 $f(x)$ 的一个扩散估计.式(3)中,$\mu(x)$ 称为扩散函数,d 称为窗宽.

将规则 p 的估计量序列 $S(p)=\{ES_1,ES_2,\cdots,ES_r\}$ 看成是非完备样本,$V=[0,1]$.设在 V 上有控制点集 $C=\{c_1,c_2,\cdots,c_m\}$,$0\le c_1<c_2<\cdots<c_m\le 1$ 且 $c_{i+1}-c_i=\Delta$ 为常数.令 μ 为正态扩散函数,则控制点 c_i 接收到由 r 个样本 $ES_1,ES_2,\cdots,ES_r$ 扩散出的信息总量为

$$Q(c_i)=\frac{1}{rd}\sum_{j=1}^{r}\mu\left[\frac{c_i-ES_j}{d}\right]=\frac{1}{\sqrt{2\pi}rh}\sum_{j=1}^{r}\exp\left(-\frac{(c_i-ES_j)^2}{2h^2}\right). \tag{4}$$

其中 $h=\sigma d$.

若取样本落在 c_i 处的频率值 $P_i=Q(c_i)/\sum_{i=1}^{m}Q(c_i)$ 作为概率的估计值,则一阶规则 p 估计量的期望值为 $\mu=\sum_{i=1}^{m}c_iP_i$,从而,μ 可作为度量值的估计.利用扩散估计,当样本值不含粗差时,估计结果与最小二乘估计一样,是最优无偏估计;当样本值含有粗差时,也能较好地抵御粗差的影响[10].

对用户查询时给出度量值阈值 ω,利用扩散估计的合并挖掘应返回 $\mu\ge\omega$ 的所有规则.

例 3 假设有规则 $R_1\sim R_5$ 的度量值序列如表 1 所示.各规则分别在 $M_0\sim M_9$ 分段中有信任度估计值.假设有 11 个控制点,取经验值 $h=1.4208(\max(样本)-\min(样本))/(样本个数-1)$[10].最后一列均值表示按以上扩散估计方法获得的 μ 值.这些 μ 值显示了此方法综合考虑各个样本点扩散信息的特性.

在按分段有限模型获得的时态规则中,可能存在一些不相关的约束关系式,即删除这些关系式后规则的信任度保持不变.因而,通过约简不相关的约束关系式的方式,可达到对时态规则的求精.

估计量的方差为 $\sigma=E(C^2)-\mu^2=\sum_{i=1}^{m}c_i^2P_i-\mu^2$.设时态规则的信任度 $EC\sim N(\mu,\sigma^2)$,则其估计的置信区间下限 $La=\mu-z_{\alpha/2}\sigma$,上限 $Ha=\mu+z_{\alpha/2}\sigma$,其中 $z_{\alpha/2}$ 是置信水平 $1-\alpha$ 的标准正态分布上 α 分位点.时态规则泛化是通过分别约简规则头和规则体来实现的,目标是经约简后的置信区间应落在原置

表 1　一阶规则的扩散估计

规则	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	均值
R_1	0.42	0.34	0.25	0.45	0.55	0.2	0.6	0.53	0.65	0.55	0.45
R_2	0.41	0.35	0.25	0.44	0.52	0.4	0.55	0.53	0.52	0.35	0.43
R_3	0.41	0.3	0.21	0.4	0.35	0.2	0.35	0.22	0.45	0.5	0.34
R_4	0.45	0.53	0.4	0.85	0.48	0.3	0.75	0.25	0.43	0.42	0.49
R_5	0.45	0.53	0.4	0.1	0.48	0.3	0.15	0.25	0.43	0.42	0.35

信区间之中. 以下是约简时态规则 TR 算法.

Reduction TR(TR, La, Ha)

For each 关系原子 in 关系头(TR) do

If 关系原子中不含事件名 then

If 约简后 TR 区间 $\leq Ha$ then 删除关系原子

End for

For each 关系原子 in 关系体(TR) do

If La ≤约简后 TR 区间 $\leq Ha$ then 删除关系原子

End for;

If La ≤约简后 TR 区间 $\leq Ha$ then

约简成功,输出 TR.

Reduction TR 算法的复杂性是 $O(knl)$, k 是时态规则中的约束关系式个数, n 是时态序列数, l 是分段长度. 规则泛化是在置信水平为 1 - α的估计区间中寻找最小约束关系式集合的过程. 该算法按逐个约简的方式可获得局部最小值,但不能保证找到全局最小值. 假设对例 1 中的 R_1 执行算法结果如表 2 所示. 对 R_1 的信任度估计如表 1 所示, R_2 和 R_3 是 R_1 约简后的规则, $\alpha = 0.05$.

表 2　时态规则的约简

规则	删除式	均值	方差	La	Ha
R_1		0.454	0.025	0.404	0.503
R_2	$X_{-4}(t_e$ = 高峰)	0.431	0.011	0.409	0.452
R_3	$X_{-4}(t_1 > 8.0)$	0.338	0.012	0.314	0.362

在表 2 中, R_2 的均值区间落于 R_1 区间中,可约简;而 R_3 的均值小于 R_1 的 La,不可约简.

6　结　　语

本文基于一阶线性时态逻辑,提出时态数据挖掘的形式理论,并讨论了一种期间挖掘和合并挖掘相结合的挖掘过程模型 C-DM. 理论框架对序列事件、子序列、模式和规则等概念进行形式化,以类 Horn 子句的形式表示规则,定义对应公式赋值的度量,证明了当公式满足一致性特征时,其支持度估计量是相合的. 基于信息扩散原理,对挖掘结果的度量序列值进行扩散估计,并提出规则约简算法. 未来的工作将基于粒度计算理论,考虑不同时间区间的规则发现和融合问题.

参考文献(References)

[1] Agrawal R, Psaila G. Active data mining [C]. Proc of the KDD'95. California: AAAI Press, 1995: 3-8.

[2] Cheng H, Yan X, Han J. IncSpan: Incremental mining of sequential patterns in large database [C]. Proc of ACM SIGKDD'04. New York: ACM Press, 2004: 527-532.

[3] Gupta S K, Bhatnagar V, Wasan S K. Architecture for knowledge discovery and knowledge management [J]. Knowledge and Information Systems, 2005, 7(3): 310-336.

[4] Spiliopoulou M, Roddick J F. Higher order mining[C]. Proc 2nd Int Conf on Data Mining Methods and Databases. Cambridge, 2000: 309-320.

[5] Last M, Klein Y, Kandel A. Knowledge discovery in time series databases[J]. IEEE Trans on Systems, Man and Cybernetics — Part B, 2001, 31(1): 160-169.

[6] Cotofrei P, Stoffel K. From temporal rules to temporal meta-rules[C]. Proc of the DaWaK 2004. Heidelberg: Springer, 2004: 169-178.

[7] Huang C F, Shi Y. Towards Efficient fuzzy Information processing —Using the principle of information diffusion [M]. Heidelberg: Physica-Verlag, 2002.

[8] Pan D, Shen J. Incorporating domain knowledge into data mining process[J]. Wuhan University J of Natural Sciences, 2006, 11(1): 165-169.

[9] XC00086D. FIPA ontology service specification[S].

[10] 王新洲. 基于信息扩散原理的估计理论、方法及其抗差性[J]. 武汉测绘科技大学学报, 1999, 24(3): 240-244.
(Wang X Z. The theory, method and robustness of the parameter estimation based on the principle of information spread [J]. J of Wuhan Technical University of Surveying and Mapping, 1999, 24(3): 240-244.)