第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·软件技术与数据库· 文章编号：1000—3428(2009)15—0094—03 文献标识码：A 中图分类号：TP391

基于时间序列模型的动态关联规则元规则挖掘

刘 俊，张忠林，谢彦峰，米 伟
(兰州交通大学电子与信息工程学院，兰州 730070)

摘 要：针对现有关联规则挖掘算法大多是挖掘一种静态关联规则的情况，介绍动态关联规则的定义，给出动态关联规则元规则的形式化定义，解决规则随时间的推移可能会有很大变化的情况下为规则建立元规则的问题，描述一种基于时间序列模型的预测和分析动态关联规则的元规则的方法，从而较好地拟合历史数据，给出满足一定显著性水平预测趋势模型的方程，挖掘规则的变化趋势，为规则建立元规则。
关键词：动态关联规则；元规则；时间序列；预测

Meta-association Rule Mining for Dynamic Association Rule Based on Time Series Model

LIU Jun, ZHANG Zhong-lin, XIE Yan-feng, MI Wei
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070)

【Abstract】 Almost all of association rules mining algorithms are considered as static ones. In fact, it is possible that a rule will change greatly with time, this paper introduces a rule called dynamic association rule and the definition of the meta-association rules for dynamic association rule, focuses on introducing the method which by the model of time series to mine the meta-association rules for dynamic association rule, and proves that this method is fit on the historical data. It can establish the equation to forecast the tendency of changing rules and mine the meta-association rules for dynamic association rule.

【Key words】 dynamic association rule; meta-association rule; time series; forecast

为了描述关联规则随时间变化的性质，文献[1]提出了数据库中动态关联规则新技术，文献[2]在分析文献[1]原有动态关联规则定义的不足的基础上，提出了一种新的动态关联规则，该定义的支持度向量，置信度向量与经典定义相吻合，可以更好地反映规则随时间变化的动态信息，本文在此基础上给出了动态关联规则元规则的形式化定义，并利用 时间序列模型等综合方法对动态规则的元规则进行了分析和预测。

1 相关定义和定理

1.1 动态关联规则定义

动态关联规则[1-2]是一种能够描述自身特性随时间变化的关联规则，描述如下：

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项集合，任务相关的事务数据集 D 是在时间段 t 内收集到的，同时根据 t 的划分，整个数据集 D 可以被分为 n 个数据子集：$D=\{D_1,D_2,\cdots,D_n\}$，其中，数据子集 $D_i(i\in\{1,2,\cdots,n\})$ 的数据是在 $t_i(i\in\{1,2,\cdots,n\})$ 时段内收集的。项集 T 满足 $T\subseteq I$。若 A 和 B 为项集，$A\subset I,B\subset I$，并且 $A\cap B=\varnothing$，则有如下动态关联规则相关定义。

支持度向量($\boldsymbol{SV}$)：动态关联规则 $A\Rightarrow B$(或者项集 $A\cup B$)的支持度向量具有如下的表示形式：

$$\boldsymbol{SV}=[s_{(A\cup B)_1},s_{(A\cup B)_2},\cdots,s_{(A\cup B)_n}]$$
$$(A\cup B)_i=f_{(A\cup B)_i}\Big/|D_i|(i\in\{1,2,\cdots,n\}) \tag{1}$$

其中，$f_{(A\cup B)_i}$ 为项集 $A\cup B$ 在数据子集 $D_i(i\in\{1,2,\cdots,n\})$ 中出现的频数；$|D_i|$ 为 D_i 中的事务数。设项集 $A\cup B$ 的支持度为 S，则有

$$s=s_{(A\cup B)}=f_{(A\cup B)}/M=\sum_{i=1}^{n}f_{(A\cup B)}/M \tag{2}$$

其中，M 是 D 中的事务数。有时项集出现的频数表示支持度更为合适，这样项集的支持度向量为

$$\boldsymbol{SV}=[f_1,f_2,\cdots,f_n]$$

置信度向量($\boldsymbol{CV}$)：动态关联规则 $A\Rightarrow B$ 的置信度向量具有如下的表示形式：

$$\boldsymbol{CV}=[c_{(A\cup B)_1},c_{(A\cup B)_2},\cdots,c_{(A\cup B)_n}]$$
$$\text{s.t.}\quad c_{(A\cup B)_i}=s_{(A\cup B)_i}\Big/s_{A_i} \tag{3}$$

其中，$s_{(A\cup B)_i}$ 为项集 $A\cup B$ 的 $\boldsymbol{SV}$ 中的第 i 个元素，S_{Ai} 为项集 A 的 $\boldsymbol{SV}$ 中的第 i 个元素。

这样一条完整的动态关联规则就可以表述为：具有支持度向量 $\boldsymbol{SV}$、置信度向量 $\boldsymbol{CV}$、支持度 s、置信度 c 4个参数的关联规则。它具有如下表示形式：$A\Rightarrow B(\boldsymbol{SV},\boldsymbol{CV},s,c)$，其中，$\boldsymbol{SV},\boldsymbol{CV},s,c$ 可以分别根据式(1)~式(3)计算，并一起描述关联规则的动态性质。

1.2 元规则的定义

一般，元规则形成一个用户希望探察或证实的、感兴趣

作者简介：刘 俊(1985－)，男，硕士研究生，主研方向：数据挖掘；张忠林，教授、博士；谢彦峰，高级工程师；米 伟，硕士研究生
收稿日期：2008-12-20 **E-mail**：liujun_1024@163.com

联系的假定。元规则[3]是形如：$P_1 \wedge P_2 \wedge \cdots \wedge P_l \Rightarrow \ Q_1 \wedge Q_2 \wedge \cdots \wedge Q_r$ 的规则模板。其中，$P_i(i=1,2,\cdots,l)$ 和 $Q_j(j=1,2,\cdots,r)$ 是示例谓词或谓词变量。

动态关联规则元规则的形式化定义如下：

定义 在数据集 $D=\{D_1,D_2,\cdots,D_n\}$ $(i\in\{1,2,\cdots,n\})$ 上，规则 $A\Rightarrow B$ 的支持度向量元规则定义为：$\boldsymbol{SV}=[s_{(A\cup B)_1},s_{(A\cup B)_2},\cdots,s_{(A\cup B)_n}]$ 或 $\boldsymbol{SV}=[f_1,f_2,\cdots,f_n]$，当且仅当，存在 $i(1\leqslant i\leqslant n)$，使 $s_{(A\cup B)_i}\geqslant \min\_sup$，其中，$\min\_sup$ 为给定的最小支持度，则此动态规则在数据集 D 上的支持度元规则为

$A\Rightarrow B:\boldsymbol{SV}$

相似地，在数据集 $D=\{D_1,D_2,\cdots,D_n\}$ $(i\in\{1,2,\cdots,n\})$ 上的置信度元规则为

$A\Rightarrow B:\boldsymbol{CV}$

元规则是“规则的规则”，可以指导知识的发现，帮助提高挖掘过程的性能。元规则可以根据分析者的经验、期望或对数据的直觉或者数据库模式自动生成。

2 动态关联规则元规则挖掘描述

现有的关联规则挖掘算法多为静态挖掘，它们的前提假设是：数据集中各项具有近似的性质和作用，即重要性相同或相近；数据集中各项的分布是均匀的，即出现的频率相近或相似[4]。然而，不同项的价值和重要程度是不一样的，不同的项在数据集中出现的频繁程度也不一样，因此，利用统一的最小支持度进行整个数据集不同时段的关联规则挖掘是不充分的。动态关联规则能够提供规则与时间相关的信息，这样便能在一定程度上解决上述不足之处。但动态关联规则仍然存在如何选取支持度的问题，而元规则挖掘的目的之一是发现关联规则序列随时间变化的趋势，通过对趋势的分析，预测下一个时间段的支持度和置信度的可能值。因此，为动态关联规则建立了元规则，在元规则的指导下，可以对数据集中的关联模式进行更加准确有效的挖掘。

目前建立元规则的方法主要有基于概率统计的方法和基于模糊决策树的方法[5]。基于概率的方法主要采用回归分析等对规则的支持度进行曲线拟合，这在处理不确定数据上效果欠佳；而从批量处理规则的角度出发，由于需要较多的专家信息，因此基于模糊决策树的方法明显无法满足要求。在历史数据较少时，可以采用基于灰色模型的方法，对于有较多历史数据的情况，本文提出一种基于时间序列模型的方法，可以很好地拟合历史数据，给出满足一定的显著性水平下计算趋势模型的方程。

2.1 时间序列模型基本组成

将支持度向量视为一个时间序列，利用时间序列方法进行建模分析并预测。支持度向量的时间序列模型可用叠加形式表示：

$$\hat{S}_t=\hat{f}_t+\hat{p}_t+\hat{x}_t \tag{4}$$

其中，$\hat{S}_t$ 为支持度向量中元素 S_t 的估计值；$\hat{f}_t$ 为趋势分量 f_t 的估计值；$\hat{p}_t$ 为周期分量 p_t 的估计值；$\hat{x}_t$ 为随机分量 x_t 的估计值。

建立模型的过程就是从已知序列 S_t (t=1,2,⋯,n)中提取各分量的过程，提取的顺序为趋势分量、周期分量和随即分量。在建立各分量的数学模型后，将其线性叠加，就得到式(4)形式的元规则预测模型。

2.2 时间序列模型各分量的确定

2.2.1 趋势分量的确定

对于趋势分量 $\hat{f}_t$ 可用多项式逼近，即

$$\hat{f}_t=c_0+c_1t+c_2t^2+c_3t^3+\cdots+c_kt^k=\sum_{k=0}^{k}c_kt^k \tag{5}$$

对式(5)可采用多元回归的方法确定系数 $c_0,c_1,c_2,\cdots,c_k$ 和阶数 k。具体求解方法可以借用 Excel 软件中县城的回归分析模板来实现。为检验拟合结果，需计算趋势曲线拟合的相关系数 R，即

$$R^2=1-\sum_{t=1}^{n}(S(t)-\hat{S}(t))^2/\sum_{t=1}^{n}(S(t)-\hat{S}(t))^2$$

其中，n 为历史数据序列 $S(t)$的总个数；$\overline{S_t}$ 为序列 $S(t)$的均值。R 越接近 1，$\hat{S}(t)$ 与 S_t(t=1,2,⋯,k)的线性关系越密切。对给定可信度 a 及不同的自由度，可以求出 R 的临界值(查表)，只有当 R 值大于相应临界值时，回归方程才有实用意义。

2.2.2 周期分量的确定

趋势函数确定后，再对扣除趋势分量后的部分 y_t 进行周期项分析，即

$y_t=S_t-\hat{f}_t$, t=1,2,⋯,n

本文主要采用谐波分析方法进行周期分量的分析提取。对序列 y_t 用 L 个波叠加的形式表示其估计值：

$$\hat{p}_t=\frac{a_0}{2}+\sum_{k=1}^{L}\left(a_k\cos\frac{2\pi kt}{n}+b_k\sin\frac{2\pi kt}{n}\right) \tag{6}$$

其中，$\hat{p}_t$ 是序列 y_t 的估计值，将其作为式(4)的周期分量的估计值；L 为谐波的个数，一般取 n/2 的整数部分；k 为谐波序号，一般认为 k 个分波各有 n/1,n/2,⋯,n/k 的周期，即第 k 个分波的频率为 k/n；a_k,b_k 为傅里叶系数，其计算公式为

$$a_k=\frac{2}{n}y_t\cos\frac{2\pi kt}{n},\ b_k=\frac{2}{n}y_t\sin\frac{2\pi kt}{n},\ k=0,1,\cdots,L \tag{7}$$

为节省工作量，通常在 L 个波中选取波动比较明显的几个谐波相加来估计，在实际应用中只需选取前 6 个显著谐波就能满足精度要求了。若 $s_k^2=a_k^2+b_k^2>4s^2\frac{\ln k}{n}$，则认为第 k 个波显著，否则不显著。其中，s^2 为系列的方差，其计算式为 $s^2=\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\overline{x}_t)^2$；$a$ 为检验的显著性水平(一般取 5%)。

2.2.3 随机分量的确定

在消除趋势分量和周期分量后，得到随机分量的时间序列，设为 x_t，即

$$x_t=S_t-\hat{f}_t-\hat{p}_t$$

x_t 一般为随机序列，可使用自回归模型(AR)拟合，x_t 均值一般为 0，设 $\hat{x}_t$ 为 x_t 的估计值，则其自回归模型为

$$\hat{x}_t=\phi_0+\phi_1x_{t-1}+\phi_2x_{t-2}+\cdots+\phi_px_{t-p} \tag{8}$$

其中，p 为模型阶数；ϕ_i 为模型自回归系数，i=0,1,⋯,p。

本文模型阶数的确定采用 AIC 准则，对参数的估计采用最小二乘法原理估计。

3 实例分析

为了直观地说明上述方法在动态关联规则的元规则挖掘中预测支持度值的具体过程，下面将针对一实例用基于时间序列模型方法对其支持度序列的值进行分析和预测。

实例：本文把文献[2]中的动态关联规则的算法做了修改后应用到某超级市场的 2004 年~2007 年的销售数据库中。该数据库是 SQL Server 2005 自带的。挖掘出某规则 $A \Rightarrow B$，即表示，顾客在购买 A 的前提下购买 B，它在 2004 年~2007 年这 4 年内的每个月支持度计数(即顾客同时购买了 A 和 B 发生的次数，包含小于最小支持度计数的月份数据)为 f_i，对 f_i 建立时间序列模型，考察此规则的变化趋势，并对该规则在 2008 年~2010 年这 3 年的支持度计数进行预测。

为了验证此模型的预测准确性，用此模型预测 2004 年~2006 年的支持度计数的预测值，将 2007 年的预测值与实际值相比较后，可以说明所预测的 2008 年~2010 年的数据是可靠的。

3.1 模型参数识别

依式(4)~式(8)进行计算，依次求取趋势项 C_0~C_7 和 R，其结果分别为 242.962, -300.355, 149.358, -29.757, 2.993 6, -0.166 3, 0.005 2, -0.000 1 和 0.872 2。

周期项参数值如下：

傅氏系数	显著周期序号 k				
	0	3	4	6	8
a_i	-1.133 9e-010	-5.262 6	12.773 2	-3.938 6	-3.456 9
b_i		-3.554 8	9.678 9	-5.945 0	-6.129 4

随机项参数值如下：

ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7
ϕ_j	-0.433 5	-0.868 3	-0.829 9	-0.823 6	-0.460 6	0.036 3	0.766 7
AIC	-769.725						

将趋势分量、周期分量和随机分量叠加即可得到支持度计数元规则预测模型，即

$$S_t=242.962\ 7-300.355\ 6t+149.358\ 8t^2-29.757\ 2t^3+2.993\ 6t^4-0.166\ 3t^5+0.005\ 2t^6-0.000\ 1t^7-0.566\ 95\times10^{-10}-5.262\ 6\cos\frac{\pi t}{6}-3.554\ 8\sin\frac{\pi t}{6}+12.773\ 2\cos\frac{2\pi t}{9}+9.678\ 9\sin\frac{2\pi t}{9}-3.938\ 6\cos\frac{\pi t}{3}-5.945\ 0\sin\frac{\pi t}{3}-3.456\ 9\cos\frac{4\pi t}{9}-6.129\ 4\sin\frac{4\pi t}{9}-0.433\ 5-0.868\ 3\,x_{t-1}-0.829\ 9\,x_{t-2}-0.823\ 6\,x_{t-3}-0.460\ 6\,x_{t-4}+0.036\ 3\,x_{t-5}+0.766\ 7 \quad (9)$$

3.2 精度检验

支持度计数元规则预测模型建立后，需对其精度进行检验。对实际值与预测值进行对比，拟合结果见图 1。

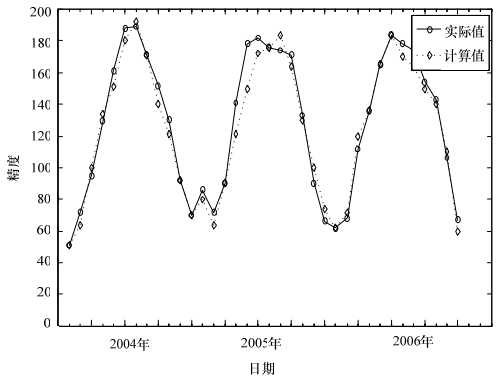

图 1 预测值和实际值拟合曲线

在把已建立的模型用于预报前，也要进行精度检验，本文采用 2007 年的数据进行后验预测检验，其预测结果见表 1。

表 1 2007 年后验预测检验数据

时间	实际值	预测值	绝对误差	相对误差/(%)	时间	实际值	预测值	绝对误差	相对误差/(%)
1 月	62	60	-2	3.20	7 月	180	184	4	2.22
2 月	82	79	-3	3.60	8 月	178	174	-4	2.25
3 月	107	105	-2	1.87	9 月	154	160	6	3.90
4 月	136	140	4	2.94	10 月	142	140	-2	1.41
5 月	165	170	5	3.03	11 月	108	111	3	2.78
6 月	182	180	-2	1.10	12 月	66	69	3	4.55

经后验差计算检验，所建立的支持度计算元规则预测模型的后验差比值 c 和小误差频率 p 分别为：2.737 5 和 1。从图 1 中也可以看出其拟合精度较高。

3.3 预测分析

曲线拟合及后验预测检验表明此模型满足精度要求。现依据式(9)，预测此规则 2008 年~2010 年的每个月的支持度计数值，所得结果见表 2。

表 2 2008 年~2010 年逐月预测的支持度元规则数据

年份	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
2008 年	52	60	105	120	164	186	175	179	142	128	105	70
2009 年	56	84	98	133	180	188	187	175	160	143	107	61
2010 年	53	119	126	163	160	195	177	171	161	143	97	50

从表 2 可见，利用本模型预测所得的此规则 2008 年~2010 年逐月预测的支持度计数仍然呈波形按周期有规律地变动，同时每年的 1 月和 12 月的数据与往年的相比有所降低，可根据实际情况分析原因，进行决策分析，其他各月数据与往年基本保持一致。

4 结束语

本文利用基于时间序列的方法解决了动态关联规则过程中如何选取支持度的问题，克服了静态关联规则挖掘算法的数据集中各项差异较大或分布不均匀的缺点。并通过一个实例说明利用本模型挖掘动态关联规则的元规则的一般过程，结果表明，其计算简单，拟合精度和预测精度均较高，是一种较好的模拟预测模型。利用本模型进行元规则挖掘适用于历史数据较多且无规律可循时，若历史数据较少则可以使用灰色模型进行建模预测。通过对规则变化趋势的拟合和对未来数据的预测能更准确地把握规则的变化趋势，从而使动态关联规则挖掘在合理的元规则指导下得到更精确的结果。

参考文献

[1] Liu Jinfeng, Rong Gang. Mining Dynamic Association Rules in Databases[C]//Proc. of International Conf. on Computational Intelligence and Security. Xi'an, China: [s. n.], 2005: 668-695.

[2] 沈 斌, 姚 敏. 一种新的动态关联规则及其挖掘算法研究[Z]. (2007-12-02). http://www.paper.edu.cn.

[3] Han Jaiwei, Kamber M. 数据挖掘: 概念与技术[M]. 范 明, 孟晓峰, 译. 北京: 机械工业出版社, 2007.

[4] 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.

[5] Wai Hoau, Chan K C C. Mining Changes in Association Rules: A Fuzzy Approach[J]. Fuzzy Sets and Systems, 2005, 14(1): 87-104.

编辑 张正兴