第 24 卷第 1 期
2007 年 01 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 24 No. 1
Jan. 2007

文章编号: 1000−8152(2007)01−0127−05

数据库中动态关联规则的挖掘

荣　冈, 刘进锋, 顾海杰

(工业控制技术国家重点实验室 浙江大学先进控制研究所, 浙江 杭州 310027)

摘要: 关联规则能挖掘变量间的相互依赖关系, 但是不能反映规则本身的变化规律. 为此本文提出了动态关联规则. 首先将整个待挖掘数据集按时间划分成若干子集, 每个子集挖掘得到的每条规则分别生成一个支持度和一个置信度，这样每条规则在全集上就对应了一个支持度向量和一个置信度向量. 通过分析支持度向量和置信度向量, 不仅可以发现规则随时间变化的情况, 也能够预测规则的发展趋势. 本文还提出了两个挖掘动态关联规则的算法, 且对他们做了比较. 并给出了柱状图和时间序列两种方法分析这两个向量. 最后给出了一个挖掘动态关联规则的应用实例。

关键词: 动态关联规则; 关联规则; 柱状图; 时间序列

中图分类号: TP273　　**文献标识码**: A

Mining dynamic association rules in databases

RONG Gang, LIU Jin-feng, GU Hai-jie

(National Key Lab of Industrial Control Technology and Institute of Advanced Process Control, Zhejiang University, Hangzhou Zhejiang 310027, China)

Abstract: Association rules may discover the relations between variables, but are unable to reflect the variation between relations. Consequently, dynamic association rule is introduced in this paper. In our method, the entire database is divided into a series of subsets in time field, and each rule from a subset has a measure of support and confidence. As a result, there are a vector of supports and a vector of confidences for each rule. It not only helps us discover the rule variation with time by analyzing the two vectors, but also predicts the future of a rule. Two algorithms for mining dynamic association rule are proposed in this paper, and a comparison of such two algorithms is also made. Subsequently, histograms and time series are described as ways for analyzing the two vectors. Finally, the effects of dynamic association rule are shown in an instance.

Key words: dynamic association rules; association rules; histogram; time series

1　引言(Introduction)

关联规则是数据挖掘领域应用非常广泛的一种挖掘方法[1], 并且已经有许多对应的挖掘算法. 然而这些算法认为发现的关联规则在数据库中是永恒有效的, 没有考虑到规则的变化, 得到的是一种静态规则.

实际上, 规则和数据特性随着时间可能会有很大的变化, 例如: 如果用某超市一年的销售数据作为分析对象, 有可能发现”顾客在买白酒的同时也会购买礼品”这个规则, 但如果仔细分析可能发现, 支持这个规则的数据都集中在春节前后的几个月中, 而在平时的数据中支持度很小, 并没有实际的指导作用. 这说明规则是在变化的, 因此, 进一步考虑规则的变化更加符合规则的实际特性.

Agrawal R和Srikant R提出了序列模式的挖掘[2], 第一次在频繁模式挖掘中考虑到时间因素, 但仍然认为发现的序列模式是长期有效的, 还是没有考虑模式自身的变化; Dong G和Li J在文献[3]中研究了突发模式, 不过仅仅局限在支持度有剧烈变化的模式; Au WH和Chan KCC 在文献[4]中考虑了关联规则自身的变化, 但仅仅局限在对规则的预测上; 文献[5]也考虑了规则的时间特性, 它关注的只是哪些时间段的数据具有相似的关联规则.

本文在关联规则变化的基础上考虑了动态关联规则, 给出了相应的支持度和置信度表示方法, 以及动态关联规则的挖掘算法和应用途径. 应用示例表明了规则的有效性和适用性.

2　应用环境(Environment)

2.1　动态关联规则的定义(Definition of dynamic association rule)

动态关联规则是一种可以描述自身随时间变化的关联规则. 它可定义如下: 设$I = \{i_1, i_2, \cdots, i_m\}$

收稿日期: 2004−11−10; 收修改稿日期: 2006−02−23.
基金项目: 国家自然科学基金创新群体资助项目(60421002).

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, t可以分成不相交的长度为n 的时间序列, 即$t = \{t_1, t_2, \cdots, t_n\}$. 依照$t$, 数据集$D$可以划分为相应的$n$个子数据集, $D = \{D_1, D_2, \cdots, D_n\}$, 其中子数据集$D_i(i \in \{1, 2, \cdots, n\})$ 中的数据是在时间段$t_i(i \in \{1, 2, \cdots, n\})$ 内收集的. 数据集D中每个事务T是项的集合, 使得$T \subseteq I$. 每个事务有一个标识符, 称作TID. 设A是一个项集, 事务T包含A当且仅当$A \subseteq T$. 关联规则是形如$A \Rightarrow B$ 的蕴涵式, 其中$A \subset I$, $B \subset I$, 并且$A \cap B = \varnothing$. 规则$A \Rightarrow B$在事务集D 中成立, 具有支持度s, 其中s 是D 中事务包含$A \cup B$ (即A 和B 二者)的百分比, 它是概率$P_D(A \cup B)$. 设$P_D[(A \cup B)_i]$ 是子数据集D_i中所包含$A \cup B$ 的记录数与在数据集D 总记录数之比, 则s 同样可以表示为概率$\sum\limits_{i=1}^{n} P_D[(A \cup B)_i]$. 规则$A \Rightarrow B$ 在事务集D 中具有置信度c, 它是条件概率$P_D(B|A)$, 设$P_D(B_i|A)$是子数据集D_i 中包含$A \cup B$ 的记录数与在数据集D中包含A 的记录数之比, 则c 同样可以表示为概率$\sum\limits_{i=1}^{n} P_D(B_i|A)$.

2.2 支持度向量(Support vector)

本文利用支持度向量(SV)和置信度向量(CV)以及支持度s 和置信度c 4个变量共同评价一条规则.

项集A 的支持度向量定义为: $SV = [s_1, \cdots, s_n]$, 其中$s_i(i \in \{1, \cdots, n\})$是项集$A$ 在数据子集$D_i(i \in \{1, \cdots, n\})$ 中出现的频数$f_i(i \in \{1, \cdots, n\})$ 与D 中的记录数M之比, 即

$$s_i = f_i/M,\ i \in \{1, \cdots, n\}. \tag{1}$$

设项集A 的支持度为s, 则有

$$s = \sum_{i=1}^{n} s_i. \tag{2}$$

设最小支持度为$\min_\sup$, 如果$s > \min_\sup$ 成立, 则项集A 称之为频繁项集.

有时, 利用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$SV = [f_1, \cdots, f_n]$, 相应的支持度可以表示为$s = \sum\limits_{i=1}^{n} f_i$.

2.3 置信度向量(Confidence vector)

因为动态关联规则与普通关联规则在从频繁项集产生规则的过程是相同的, 不同之处在于置信度向量的计算, 所以本文仅仅关注置信度向量的生成方式.

动态关联规则$A \Rightarrow B$ 的置信度向量定义为$CV = [c_1, \cdots, c_n]$, 其中$c_i(i \in \{1, \cdots, n\})$ 是0%∼100%之间的一个百分数. 设$SV_{A\cup B} = [s_{(A\cup B)_1}, \cdots, s_{(A\cup B)_n}]$ 为$A \cup B$ 的支持度向量, $SV_A = [s_{A_1}, \cdots, s_{A_n}]$ 为A 的支持度向量, $SV_B = [s_{B_1}, \cdots, s_{B_n}]$ 为B 的支持度向量. 并且A 的支持度为s_A, 则有

$$c_i = \frac{s_{(A\cup B)_i}}{\sum\limits_{i=1}^{n} s_{A_i}} = \frac{s_{(A\cup B)_i}}{s_A},\ i \in \{1, \cdots, n\}. \tag{3}$$

设$A \cup B$ 的支持度为$s_{A\cup B}$, B 的支持度为s_B, 并且规则$A \Rightarrow B$ 的置信度为c, 则有

$$c = \frac{\sum\limits_{i=1}^{n} s_{(A\cup B)_i}}{\sum\limits_{i=1}^{n} s_{A_i}} = \frac{s_{A\cup B}}{s_A} = \sum_{i=1}^{n} c_i. \tag{4}$$

设最小置信度为$\min_\mathrm{conf}$, 如果$c \geqslant \min_\mathrm{conf}$ 成立, 则规则$A \Rightarrow B$ 是一条强动态关联规则.

2.4 动态关联规则的完整表示(Whole dynamic association rule)

一条完整的动态关联规则可以描述如下:

$$A \Rightarrow B$$
$$(SV = [s_1, \cdots, s_n], CV = [c_1, \cdots, c_n], s, c) \tag{5}$$

其中SV, CV, s 和c 一起描述了规则的特性.

如规则(5)所示, 动态关联规则中既包含了传统的支持度和置信度的信息, 还提供了普通关联规则所没有的时变特性信息.

3 两种挖掘算法(Two mining algorithms)

在准备用于动态关联规则挖掘的数据时, 需要考虑时间因素. 数据集中的每条记录必须包含一个时间指示属性, 称为time_id, 作为分割数据集的依据.

下面的两种算法主要关注频繁项集与相应支持度向量的寻找. 动态关联规则生成与普通关联规则生产相同, 置信度向量和置信度可利用式(3)和(4)计算得到.

3.1 算法1(The first algorithm)

这种算法时间消耗比较大, 但相对比较简单. 它基于成熟的关联规则挖掘算法, 并进行了相应的改进.

设整个数据集为D, 并分割成n个数据子集$D_1 \sim D_n$; 设全体频繁项集的集合为L, l_i是其中的一个项集; 设f_{ij} 是项集l_i在数据子集D_j中出现的频数. 第1步, 调用一个成熟的关联规则挖掘算法, 如Apriori或FP-growth, 在D 中找到L; 第2步, 扫描$D_1 \sim D_n$ 找出l_i在不同子集上的f_{ij}; 最后, 利用式(1)和(2)得到支持度向量和支持度. 算法描述如表1所示:

表 1　算法1伪代码

Table 1　Pseudocode of the first algorithm

输入: 数据集D与子集$D_1 \sim D_n$, min_sup
输出: L以及对应的支持度向量与支持度

```
1:  L=Association-mining-algorithm;
2:  for(j = 1; j ⩽ n; j++){
3:    for each l_i ∈ L {
4:      scan D_j for frequency f_ij;
5:      s_ij = f_ij/M; } }
6:  for each l_i ∈ L {
7:    SV_i = {s_i1, ··· , s_in};
8:    s_i = Σ_{j=1}^{n} s_ij; }
9:  return L with support vectors
```

函数Association-mining-algorithm的功能是调用一个普通关联规则挖掘算法, 如Apriori或FP-growth, 寻找数据集D中的全部频繁项集L. M是数据集D中记录数.

3.2　算法2(The second algorithm)

这个算法基于经典的Apriori算法, 并对其进行改进以便能产生规则的支持度向量. 在这个算法中, 支持度向量用出现的频数作为它的元素. 算法的改进之处描述如下.

首先, 在第1步寻找候选1-项集, 算法逐个扫描所有的子集并记录每个项在每个子集中出现的频数, 于是对每个候选1-项集都可以得到一个支持度向量. 利用这些向量可以得到每个1-项集的支持度, 因此, 就可以得到频繁1-项集.

其次, 在寻找频繁k-项集的循环过程中, 这个算法同样扫描每个子集, 从而得到利用频繁$(k-1)$-项集生成的候选k-项集在每个子集中出现的频数, 于是可以得到候选k-项集的支持度向量. 利用这些支持度向量可以计算出候选集的支持度, 从而得到频繁k-项集与它们对应的支持度向量. 算法的细节描述如表2.

函数Apriori-gen在本文中没有给出, 可以参考文献[6], 它的作用是从频繁$(k-1)$-项集产生候选k-项集. 算法中函数Scan-support-1-itemset 的作用是寻找每个1-项集在不同子集中出现的频数. 算法中函数Join-support-vector的作用是融合每个1-项集在各个子集中出现的频数得到相应的支持度向量.

算法1分成两阶段, 直观易理解. 而算法2最主要的特点是能够在寻找频繁项集的过程中间计算出支持度向量, 对数据库的扫描次数相当于算法1中第1步求时的扫描次数, 所以算法2效率更高.

表 2　算法2伪代码

Table 2　Pseudocode of The Second Algorithm

输入: 数据集D与子集$D_1 \sim D_n$, min_sup
输出: L以及对应的支持度向量与支持度

```
1:  for each D_i {
2:    C_1i=Scan-support-1-itemset(D_i); }
3:  C_1=Join-support-vector(D_1 ∼ D_n);
      //find candidate with support vectors
4:  L_1 = {c ∈ C_1| Σ_{i=1}^{n} c · frequency_i ⩾ min_sup};
      //c · frequency_i is the frequency of c in D_i
5:  for(k=2;L_{k-1} ≠ φ; k++) {
6:    C_k=Apriori-gen(L_{k-1},min_sup );
7:    for each D_i {
8:      for each transaction t ∈ D_i {
9:        C_t = subset(C_k, t);
10:       for each candidate c ∈ C_t
11:         c · frequency_i ++; } }
12:   L_k = {c ∈ C_k| Σ_{i=1}^{n} c · frequency_i
                ⩾ min_sup}; }
13: return L = ∪_k L_k with support vectors;
```

4　两种应用动态关联规则的方法(Two ways to use dynamic association rules)

动态关联规则包含支持度向量和置信度向量. 可以利用柱状图分析和时间序列分析两种方法对这两个向量进行分析得到有关规则的更详尽的信息.

4.1　柱状图分析(Histograms)

支持度向量和置信度向量的柱状图可以清楚地描述规则支持度和置信度的分布情况; 并且可以定性的反映规则支持度和置信度随时间的变化的情况. 根据定义, 可以发现支持度和置信度的变化趋势是相同的, 因此, 仅仅需要绘制其中一个向量的柱状图就可以.

如果某条动态关联规则的置信度向量CV是: CV=[3.7%, 4.744%, 7.4%, 22.2%, 21.46%, 22.94%], 则可以绘制出柱状图如图1所示.

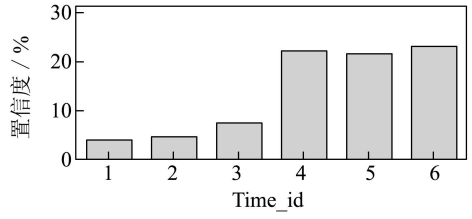

图 1　置信度CV的柱状图

Fig. 1　Histograms of CV

从图1中规则置信度在第4个time_id 处有较大的上升, 并且在之后的时间段内维持在一个比较高的水平. 这是一个上升的趋势, 它表明规则在实际应用

中将有很好的效果.

从置信度或支持度向量的柱状图中, 可能还可以找到下降的趋势、周期性的趋势等等. 一个下降的趋势表明规则的实效性不好, 应用效果将较差; 一个周期性的趋势表明规则并不是稳定的, 只有符合它的变化周期的应用才会有很好的效果.

4.2 时间序列分析(Time series analysis)

时间序列分析是在描述数据变化和预测数据趋势中应用比较广泛的一种方法. 如果一个支持度向量以规则出现的频数表示, 并且含有足够的元素, 它可能就适合时间序列分析. 例如: 如果一条从某超市12个月销售数据中找到的规则的支持度向量是: SV=[568, 574, 581, 582, 584, 586, 594, 600, 606, 612, 613], 那么, 可以建立一个自回归模型描述规则出现频数的变化过程, 如式(6)所示:

$$f_i - 1.4f_{i-1} = 8\varepsilon_t,\ \ \varepsilon_t \sim N(0,1). \tag{6}$$

从式(6)所示的模型中, 可以发现规则的支持度存在一个上升的趋势, 并且能够预测规则频数在后续几个月中的值. 如对支持度进行的3个月预测为: 617, 622, 624.

利用时间序列分析, 可以找到规则支持度或置信度向量的定量模型, 它能够给出比柱状图更加精确的信息, 最为重要的是可以预测规则的发展趋势.

5 应用实例(Application case)

本文把动态关联挖掘的思想用在某超级市场1997年的销售数据库中. 该数据库是SQL Server 7.0自带的, 含有86837条记录, 数据库按月份分为$D_1 - D_{12}$ 等12个子数据库, 在$D_1 - D_{11}$ 上进行分析, 找出动态关联规则, 并且利用时间序列分析的方法预测D_{12} 中规则出现的频数.

在分析中用到的数据库的属性如表3所示, 当Time和Customer 相同时, 分析中认为交易是在顾客在Time对应时刻的进行的一次交易, 是在物品类的层次上进行动态关联规则分析的, 这样避免了从底层物品分析时, 物品种类多, 支持度小的问题. $D_1 - D_{11}$中含78120条记录, 对于动态频繁1-项集, 设定最小支持度$\min_\sup_1 = 2\%$, 频繁2-项集, 设定最小支持度$\min_\sup_2 = 1\%$, 总体最小置信度为$\min_\mathrm{conf} = 25\%$. 在这种条件下可以找到一批动态关联规则, 其中两条与其对应CV 如表4所示.

在表4中, 99/61/58分别是3种具体的商品类, 两条规则的置信度向量CV 绘制柱状图, 分别如图2和图3所示. 从图中可以看出两条规则置信度分布都存在一定的波动, 分布呈现出相似的周期波动, 周期约为4~5个月.

表 3 分析用数据属性

Table 3 Attributes of analyzed data

属性	描述
Time	交易时间
Class	交易物品类别
Customer	顾客ID
Month	交易月份

表 4 动态关联规则

Table 4 Dynamic association rules

Rule	CV	c
99⇒61	3.18%,3.86%,3.18%,2.5%, 3.18%,3.41%,3.41%,3.18%, 2.73%,2.5%,3.41%	34.3%
58⇒61	3%,3.67%,3.33%,2.67%, 2.33%,3.33%,3.33%,3.0%, 2.67%,2.67%,3.67%	33.7%

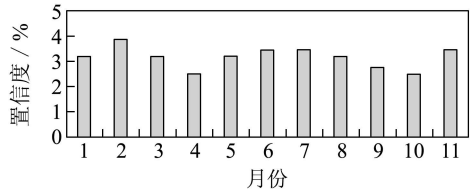

图 2 规则99⇒61的CV柱状图

Fig. 2 Histograms of CV on rule 99⇒61

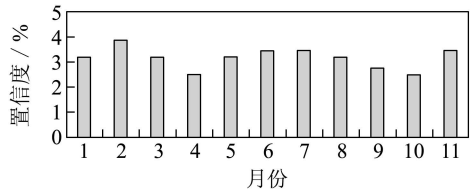

图 3 规则58⇒61的CV柱状图

Fig. 3 Histograms of CV on rule 58⇒61

对规则58⇒61在12月出现的频数进行预测, 规则支持度中各个元素转换成相应的频数后为s = {73, 82, 79, 61, 54, 76, 79, 69, 61, 64, 89}, 可以利用时间序列的分析方法建立序列的模型为

$$(1 - 0.4z^{-1} + 0.99z^{-2})s_N = 112.8 + e_N,$$

其中: z 是移位算子, e_N是白噪声. 对12月频数预测公式为$\hat{s}_{12} = 0.4s_{11} - 0.99s_{10} + 112.8$, 可以算出$\hat{s}_{12}$=85, 即12月规则出现的频数预测为85, 实际中规则出现的频数为80, 预测误差为$\dfrac{85-80}{80} = 6.25\%$, 有很好的预测效果, 说明利用支持度向量进行预测是可行的.

6 总结(Conclusion)

本文考虑了关联规则的动态特性, 通过分割挖掘数据集的方式, 挖掘不仅包含支持度和置信度, 而且

包含支持度向量和置信度向量的关联规则. 这种规则可以提供自身随时间变化的信息, 能够预测规则的发展趋势, 具有普通关联规则所不具有的功能.

参考文献(References):

[1] AGRAWALl R, MANNILA H, SRIKANT R, et al. *Fast Discovery of Association Rules*[M]// Advances in knowledge discovery and data mining. Menlo Park, CA: AAAI/MIT Press, 1996: 307 – 328.

[2] AGRAWAL R, SRIKANT R. Mining sequential patterns[C]// *Proc of the 11th Int'l Conf on Data Engineering*. Taipei: IEEE Computer Society Press, 1995: 3 – 14.

[3] DONG G, LI J. Mining border descriptions of emerging patterns from dataset pairs[J]. *Knowledge and Information Systems*, 2005, 8(2): 178 – 202.

[4] AU WH, CHAN KCC. Mining changes in association rules: a fuzzy approach[J]. *Fuzzy Sets and Systems*, 2005, 149(1): 87 – 104.

[5] GANTI V, GEHRKE J, RAMAKRISHNAN R. DEMON: Mining and monitoring evolving data[J]. *IEEE Trans on Knowledge and Data Engineering*, 2001, 13(1): 50 – 63.

[6] 韩家炜, 坎伯. 数据挖掘: 概念与技术[M]. 北京: 机械工业出版社, 2001.
(HAN Jiawei, KAN Bo. *Data Mining: Concepts and Techniques*[M]. Beijing : China Machine Press, 2001.)

作者简介:

荣　冈　(1963—), 男, 教授, 博士生导师, 研究领域为先进控制理论与应用、数据校正与数据挖掘、流程工业过程优化与调度等, E-mail: grong@iipc.zju.edu.cn;

刘进锋　(1982—), 男, 浙江大学模式识别与智能系统专业硕士, 研究领域为数据挖掘在流程工业中的应用, E-mail: jinfeng.liu.china@gmail.com;

顾海杰　(1981—), 男, 浙江大学控制理论与控制科学在读博士, 研究领域为数据挖掘及其在流程工业中的应用, E-mail: hjgu@iipc.zju.edu.cn.

(上接第126页)

[6] HAYASHI K, OTSUBO A, MURAKAMJ S, et al. Realization of nonlinear and linear PID controls using simplified indirect fuzzy inference method[J]. *Fuzzy Sets and Systems*, 1999, 105(3): 409 – 414.

[7] AMIN Haj Ali, YING Hao. Structural analysis of fuzzy controllers with nonlinear input fuzzy sets in relation to nonlinear PID control with variable gains[J]. *Automatica*, 2004, 40(9): 1551 – 1559.

作者简介:

朱　伟　(1976—), 男, 上海交大国家模具CAD工程研究中心博士生, 研究方向为板料拉深成形智能控制和工艺参数数值模拟优化, E-mail: sjtuzhuwei@sjtu.edu.cn;

董湘怀　(1955—), 男, 上海交大国家模具CAD工程研究中心教授, 博士生导师, 研究领域为材料成形过程模拟和塑性成形工艺优化, E-mail: dongxh@sjtu.edu.cn;

张质良　(1935—), 男, 上海交大国家模具CAD工程研究中心教授, 博士生导师, 研究领域为材料成形过程智能化设计, E-mail: zlzhang@mail.sjtu.edu.cn.