文章编号: 1001-0920 (2012) 10-0000-00

一种面向布尔时间序列的关联规则挖掘算法

闫明月[1], 侯忠生[1], 高 颖[2]

(1. 北京交通大学 电子信息工程学院，北京 100044；2. 北京中医药大学 东直门医院，北京 100700)

摘 要: 布尔时间序列中的关联规则挖掘较难处理, 因为多数关联规则仅挖掘不同事务共同出现的规则, 却难以体现同一事件在不同时间内动态变化间的关联性. 鉴于此, 提出一种新的关联规则挖掘框架, 利用常量化表示布尔数据的时间属性, 结合聚类算法与关联分析, 提高规则的支持度, 从而解决布尔时间序列数据在关联规则挖掘中的时间值表示问题, 并使用多种指标评价规则与传统算法比较. 在真实的中风病预后好转数据预测中验证了所提出算法的有效性.

关键词: 布尔时间序列；关联规则；聚类；预后

中图分类号: TP3-05 **文献标志码:** A

Algorithm of mining association rules for binary time series

YAN Ming-yue[1], *HOU Zhong-sheng*[1], *GAO Ying*[2]

(1. School of Electronics and Information Engineering，Beijing Jiaotong University，Beijing 100044，China; 2. Dongzhimen Hospital，Beijing University of Chinese Medicine，Beijing 100700，China. Correspondent：HOU Zhong-sheng，E-mail：zhshhou@bjtu.edu.cn)

Abstract: Association analysis for Binary time series is a difficult problem, because most of association rules lay emphasis on the relation among the items, but ignore the temporal correlation in the transaction database. Therefore, a new improved algorithm of mining association rules for binary time series is presented to make use of both the relationship among the items and the temporality of association. By using the proposed algorithm, the binary data is converted to common numerical value for representing the time-value implicitly, then clustering algorithm is combined with the association analysis, which improves the supports of most association rules. Several indicators are used to evaluate the results from the proposed algorithm. The experimental results on the prognosis dataset of stroke show the effectiveness of the proposed method.

Key words: binary time series；association rules；clustering；prognosis

1 引 言

布尔时间序列是按时间次序排列的布尔型观测值的集合[1], 关联分析是通过挖掘关联规则来展示属性值频繁地在给定数据集中一起出现的条件, 从而得到大量数据中项集之间有趣的关联[2]. 布尔时间序列数据中的关联规则挖掘可揭示不同属性间的动态变化情况,在诸如生物医学、气象预测及故障诊断等领域中有重要的研究意义.

布尔时间序列中的关联规则挖掘与一般关联规则挖掘或序列模式挖掘不同: 大多数经典关联规则挖掘算法如 Apriori 算法[3]可以得到不同项 (items) 之间的关系模式, 即某些频繁项集共同出现的规律, 但这种规律是静态的, 难以表达时间信息; 序列模式挖掘[4]的目的是找出所有频繁子序列, 得到不同项之间的时间次序关系, 然而这种次序仅为时间的偏序关系[5]即简单的序列升降[6], 没有真正解决时间值表示问题.

为了解决时间值表示问题, 许多研究对关联规则挖掘算法提出各种改进. 如文献 [7] 将时间作为一个新维度, 将序列数据库从传统的 2 维扩展到 3 维, 通过改造数据库的方式表达事件发生的时间顺序, 然而这种算法却难以处理时间缺失问题; [8-9] 考虑到关联规则自身的变化, 提出动态关联规则挖掘算法, 基本思想是以 Apriori 算法为基础, 将数据集按时间划分成相应子集, 其实质是用列举时间项的方式表示时间信息; [10] 引入时间权重概念, 提出带权的布尔时间序

收稿日期: 2011-07-08；**修回日期:** 2011-09-27.

基金项目: 国家重大基础研究计划项目 (2003CB517102).

作者简介: 闫明月(1984−), 女, 博士, 从事数据挖掘及其在医学和智能交通中应用的研究；侯忠生(1962−), 男, 教授, 博士生导师, 从事控制理论与应用、智能交通等研究.

列挖掘方法, 但该算法的前提是假设后发生的事件比先发生的事件重要, 而实际数据未必都能满足这个假设, 因此对事件发生的时间准确定位并分析其与结果间的联系才有可能得到更加准确有效的规则.

综上所述, 已有算法的共同特点是时间值表示不明确, 规则中的时间规律不易体现或不能完全体现, 复杂度高, 规则解释困难.

本文针对上述问题, 结合聚类与关联规则挖掘, 提出一种新的关联算法框架. 在真实的中医中风病预后数据中进行验证, 并对其中参数进行调整测试. 结果表明, 所提出的算法可在减小计算复杂度的同时有效地解决时间值表示问题.

2 问题描述及相关定义

2.1 问题描述

一般关联规则挖掘的定义[2]如下: $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合. 设任务相关的数据 D 是数据库事务的集合, 其中每个事务 T 是项的集合, 使得 $T\subseteq I$.每个事务有一个标识符, 记作 TID. 设 X 是一个项集, 事务 T 包含 X 当且仅当 $X\subseteq T$. 关联分析就是要找到形如 $X\Rightarrow Y$, 具有支持度 s 和置信度 c, 且同时满足给定最小支持度阈值 $\min\_\sup$ 和最小置信度阈值 $\min\_\mathrm{conf}$ 的强关联规则. 显然这种规则没有表达出规则中的时间关联性.

对于布尔时间数据, 项集 I_t 中包含时间信息, 记为 $I_t=\{i_1^t,i_2^t,\cdots,i_m^t\}$, 其中 $i_k^t(k=1,2,\cdots,m)$ 是按采样时间排列的布尔向量, 事务 $T_t\subseteq I_t$, 本文要解决的问题是: 当事务 T_t 包含先决条件 X_t 当且仅当 $X_t\subseteq T_t$ 时, 如何找到规则 $X_t\Rightarrow Y_t$, 使之满足给定的最小支持度阈值和最小置信度阈值.

2.2 相关定义

对于一些基本概念, 如项集、元素、序列等, 许多文献都已给出明确定义, 本文不再赘述. 为了便于与传统算法作定量比较, 得到客观有效的规则, 实验部分涉及结果有效性的评价, 本文使用的客观性评价指标有相关度[11]、兴趣度[12]与根据专家经验及需求定义的预测准确度, 现对这些指标引用与定义如下:

定义 1 相关度是置信度与期望置信度的比值, 其计算公式为

$$\mathrm{corr}_R=\frac{P(X\bigcap Y)}{P(X)P(Y)}. \tag{1}$$

相关度描述了规则前件 X 对后件 Y 的影响力. 若 $\mathrm{corr}_R=1$, 则 X 与 Y 独立, 无相关性; 若 $\mathrm{corr}_R<1$, 则二者负相关; 若 $\mathrm{corr}_R>1$, 则二者正相关. 然而有研究表示此指标虽能显示规则中 X 对 Y 是否起到积极作用, 但未能有效说明 X 对 Y 的影响程度, 需要引入兴趣度对其结果进行补充修正.

定义 2 兴趣度定义为

$$\mathrm{Interest}_R=\frac{\mathrm{confidence}(X\Rightarrow Y)-\mathrm{support}(Y)}{\max\{\mathrm{confidence}(X\Rightarrow Y),\mathrm{support}(Y)\}}. \tag{2}$$

兴趣度是基于关联规则所表示的信息与所有原始记录支持该信息的差异来定义的, 考虑各种可能的反面示例的影响, 完善相关度的评价, 其取值范围是 $[-1,+1]$. 一条规则的兴趣度越大于0表明对该规则越感兴趣, 越小于0表明对其反面规则越感兴趣.

定义 3 准确度是为了评价所得到规则在实际预测中的准确程度, 根据医生经验与需求定义为: 测试结果中属于正确结果数占所有测试记录实际结果数的比例, 其计算公式为

$$\mathrm{Accuracy}=\frac{\left\|(X'\overset{R}{\Rightarrow}Y')\subset Y\right\|}{\|Y\|}\times 100\%. \tag{3}$$

准确度越高, 说明根据所获得规则预测的结果正确的可能性越大.

3 算法设计及其与经典算法的关系

3.1 算法提出

本文提出了面向布尔时间序列的关联规则挖掘算法, 其基本思想是用常量化表示布尔时间序列中的时间值信息, 结合聚类与传统关联规则挖掘减小计算复杂度, 得到有效的关联规则, 对处理一般的布尔时间序列问题具有广泛的适用性.

经典的关联规则挖掘步骤为: 1) 找出所有频繁项集; 2) 由频繁项集产生强关联规则. 与之不同, 对原始数据进行必要的数据清洗与预处理[13-14]之后, 本文算法由以下步骤构成:

Step 1: 常量化. 将布尔时间序列数据常量化, 生成新项集 X' 与 Y'.

Step 2: 聚类. 对 X' 与 Y' 的每个属性分别聚类.

Step 3: 规则生成. 找出聚类结果中的关联规则.

Step 4: 常量化逆运算. 将结果还原为布尔序列.

3.2 与经典算法的区别

本文算法与经典关联规则的区别主要在以下3个方面:

3.2.1 数据常量化及其逆运算

本文算法利用常量化将布尔序列按时间进行加权求和, 映射到一般的数据集, 将时间信息转化为隐含属性. 其目的, 一是可以降维, 在尽可能保持数据全部属性的前提下最大限度地减少变量数目; 二是将时间序列数据转化为一般数据处理, 可应用于最简单的挖掘算法.

设 X_t 与 Y_t 分别为 I_t 中的项集, $X_t\subset I_t$, $Y_t\subset I_t$, 且 $X_t\bigcap Y_t=\Phi$, 则常量化的一般形式为

$$X' = f(X_t),\ Y' = g(Y_t). \tag{4}$$

其中: X' 与 Y' 为常量化后的项集, f 与 g 分别为加权转换函数. 相应的逆运算形式为

$$X_t = f^{-1}(X'),\ Y_t = g^{-1}(Y'). \tag{5}$$

权重的计算方法有很多, 可根据实际数据选择合适的方法, 但应满足运算可逆的要求.本文以指数加权法为例, 运算过程与进制转换类似, 常量化公式为

$$X' = \Delta^{\mathrm{T}} \cdot X_t, \tag{6}$$

其中 $\Delta = [\ \sigma^0\ \ \sigma^1\ \ \cdots\ \ \sigma^m\ ]$, 这里 σ^i 为指数权重系数. 相应的逆运算公式为

$$x(t) = \mathrm{mod}(\mathrm{floor}(x'/\sigma^t), \sigma),\ t = 0, 1, \cdots. \tag{7}$$

其中: mod 表示取余运算, floor 表示求商运算, t 与第 t 个采样时间一一对应.

常量化及其逆运算的意义在于: 前者解决了时间值表示问题, 后者用于规则解释. 以中风预后数据为例, 若 X' 取 1 或 0 分别表示“头晕目眩”症状出现或不出现, Y' 取 1 或 0 分别表示预后情况好转或恶化, 则假设根据上述公式取 $\sigma = 2$, 时间单位为“天”, 对 X 连续 7 天的观察情况常量化值是 3, 得到规则“$X' = 3 \Rightarrow Y' = 1$”, 则可解释为: 如果患者只在入院两天内出现头晕目眩症状, 之后不再出现, 则可预测其预后情况为好转. 由此可见, 可逆运算保证了原项集与新项集的一一对应关系, 得到有趣的关联规则后能够还原其时间属性, 以便对规则进行合理解释.

3.2.2　聚　类

随着关联规则与聚类分析技术的不断发展, 人们发现在某种程度上将两种技术结合对分析结果的影响相当显著. 目前, 在各类研究文献中存在着大量聚类算法, 如划分方法、层次方法、基于密度的方法等, 本文提出的一般算法框架不受算法种类的限制. 在此框架中加入聚类过程主要是受文献 [15] 的启发, 考虑不同数据间的差异, 利用时间的阶段性特点提高规则支持度, 并且可根据具体数据的类型与特点选择一种合适的聚类算法.

对 X' 的聚类结果用 $C^{(x)}_{r_i,j}[X']$ 表示, 相应的 Y' 矩阵表示为 $C^{(x)}_{r_i,j}[Y']$. 其中: $r_i = 1, 2, \cdots, r_x$; $j = 1, 2, \cdots, p$. 对 Y' 的聚类结果用 $C^{(y)}_{r_i,j}[Y']$ 表示, 相应的 X' 矩阵表示为 $C^{(y)}_{r_i,j}[X']$. 其中: $r_i = 1, 2, \cdots, r_y$; $j = 1, 2, \cdots, q$; 且 $1 \leqslant r_x < m$, $1 \leqslant r_y < m$, $r_x + r_y \leqslant m$.

对于 X' 和 Y' 的每个属性聚类, 一方面聚类结果仍保持原有数据的时间特性, 因此可以得到某一段时间的共性, 在此基础上可能提高规则支持度. 以单列属性数据为例, 若 $X' = x_i$ 和 $X' = x_j$ 的聚类中心为 $C^{(x)}_k[X']$, $Y' = y_a$ 和 $Y' = y_b$ 的聚类中心为 $C^{(y)}_l[Y']$, 规则 $C^{(x)}_k[X'] \Rightarrow C^{(y)}_l[Y']$ 相当于将

$$X' = x_i \Rightarrow Y' = y_a,\ X' = x_j \Rightarrow Y' = y_a,$$
$$X' = x_i \Rightarrow Y' = y_b,\ X' = x_j \Rightarrow Y' = y_b$$

合并, 显然其支持度高于单条规则的支持度. 另一方面, 整体聚类可能减少规则条数, 而对每一个属性聚类却能得到更多的规则.

设有 m 个监测对象, X 是对 p 个不同属性监测 N 个时间单位得到的数据集, Y 是 q 个不同属性在一个时间单位内的监测结果. 若对 X 和 Y 的全部属性均聚为 k 类, 则得到的最高规则数为 $k \times k$ 个, 而对属性分别聚类得到的最高规则数为 $k^p \times k^q$ 个. 尽管传统的算法得到的最高规则数为 $m^{N\times p+q}$, 得到的候选规则最多, 但取值分散, 所以大多数候选规则不能满足支持度与置信度要求, 生成的规则较少而计算复杂度却很高. 因此, 对每个属性分别聚类是在传统算法与改进算法中寻求一个相对较优的结果, 既得到较多有用的规则又低了计算复杂度.

3.2.3　规则生成

关联规则挖掘解决的是发现大量数据中项集之间有趣的联系, 其形式已在多篇文献中给出, 而如何将 Step 1 中常量化表示的数据与 Step 2 的聚类结果应用于关联规则是本算法要解决的问题.

本文提出的面向布尔时间序列的关联规则是形如

$$\bigcup_{i,j=1}^{u_x,v_x} C^{(x)}_{ij} \Rightarrow \bigcup_{k,l=1}^{u_y,v_y} C^{(y)}_{kl}$$

的蕴涵式, 其中 $1 \leqslant u_x \leqslant r_x$, $1 \leqslant v_x \leqslant p$, $1 \leqslant u_y \leqslant r_y$, $1 \leqslant v_y \leqslant q$. 与经典的关联规则类似, 规则的支持度与置信度计算公式如下:

$$\mathrm{support}_R = P\Big(\Big(\bigcup_{i,j=1}^{u_x,v_x} C^{(x)}_{ij}\Big)\bigcup\Big(\bigcup_{k,l=1}^{u_y,v_y} C^{(y)}_{kl}\Big)\Big), \tag{8}$$

$$\mathrm{confidence}_R = P\Big(\bigcup_{k,l=1}^{u_y,v_y} C^{(y)}_{kl} \Big| \bigcup_{i,j=1}^{u_x,v_x} C^{(x)}_{ij}\Big). \tag{9}$$

根据最小支持度阈值 min_sup 和最小置信度阈值 min_conf 可以对得到的规则进行裁剪, 保留有用的强关联规则.

3.2.4　与经典算法的联系

本文提出的改进算法与经典的关联算法具有如下密切的联系: 两种算法中由全部属性 (并非一定是全部时间单位的全部属性) 生成的规则保持一致, 即传统算法中由最大项集生成的关联规则是改进算法中生成规则的子集. 当 X' 聚类后的类结果数与数据库事务的记录数相等时, 改进算法中的 Step 2 聚类失效, 与原数据集相比只有记录位置发生变化, 而对规则生成不产生影响. 此时, 改进后算法生成的规则是

传统算法生成关联规则的子集. 由于传统算法中由最大项集生成的关联规则又是本改进算法中生成规则的子集, 此时改进算法中生成的规则与传统算法中由最大项集生成的关联规则等价.

4 实 验

为了验证算法的可行性与有效性, 对真实的中风病四诊信息及预后数据集分别用经典算法与本文提出的算法求取关联规则. 由于经典算法的共性是时间值表示不明确, 未能体现时间对规则产生的影响. 仅以经典的 Apriori 算法为例, 为了便于比较, 对于时间值表示, 借鉴文献 [8] 的方法, 将数据库按时间划分求取关联规则. 所有程序均在 Intel(R)Core(TM)Duo CPU T2300 @1.66 GHz 1.66 GHz, GB 内存, Windows 平台下, 使用 Matlab 7.0 软件编写并执行.

4.1 数据集及其描述

本文选用的中风病四诊信息及预后数据集由北京中医药大学东直门医院提供, 共 1 023 名病人参与调查, 包括入院后连续 14 天的 187 种症状观测情况及预后结果, 收集数据共计 14 880 例次. 其中 838 名病人的 10 795 例次数据为有效数据, 根据专家知识, 挑选 85 种主要症状, 并将其归类使之分属为内风 (9 种)、内火 (21 种)、痰湿 (17 种)、血瘀 (10 种)、气虚 (18 种) 和阴虚 (10 种) 6 类证候, 选择 419 名病人的 5 866 例次数据作为训练集, 其余 419 名病人的 5 866 例次数据作为测试集. 数据选择原则见文献 [16].

4.2 实验结果

首先, 根据式 (4) 和 (6) 将数据集常量化. 为了便于计算暂取 $\sigma = 2$; 然后, 根据医生经验将预后结果分为好转、无变化和恶化 3 类. 在 Step 2 中使用 k-means 算法[19], 该算法规则简单, 易于实现, 且可根据预后结果指定聚类数 $k = 3$. 对于 k-means 算法对初值选择的强依赖性可能导致结果不稳定的问题, 可进行多次迭代直至收敛以保证得到较好且稳定的聚类结果. 在 Step 3 中, 为了便于比较且考虑到预测对置信度要求较高, Apriori 算法与本文的改进算法均取参数 $\min\_\sup = 0.05$, $\min\_\text{conf} = 0.60$.

以内风证候预后好转数据为例, Apriori 算法与本文改进算法得到的关联规则分别为 200 条和 1 065 条, 相应运行时间分别为 82.75 s和 2.39 s, 所得关联规则结果评价情况如表 1 所示.

由表 1 可见, 改进算法的运行速度更快, 且得到的全部规则数和大项集规则数远远多于 Apriori 算法的结果. 从评价指标来看, 两种算法得到的规则相关程度都满足要求, 而改进算法得到的大部分规则兴趣度更高, 整体平均准确度远远高于 Apriori 算法. 然而, 由于聚类数较少, 为了得到更多的规则, 本文又选择了较低的支持度阈值. 因此, 实验中个别准确度很低的频繁小项集规则也被选了出来. 对此可通过增大聚类参数 k 和提高支持度阈值避免这种情况. 而实际应用时根据一两种症状就预测预后结果的可能性也很低, 一般需要考虑多种症状判断预后结果. 因此, 小项集仅作为候选集用以生成大项集规则而不作为有效的预测依据.

表 1 两种算法所得关联规则性能评价

算法	项数	规则数	平均相关度	平均兴趣度	平均准确度/%
Apriori	2	1	1.017 7	0.051 9	44.15
	3	38	1.035 2	0.051 9	35.98
	4	74	1.055 5	0.052 1	36.09
	5	74	1.113 9	0.101 6	36.06
	6	13	1.253 6	0.198 9	35.71
改进算法	2	59	1.152 1	0.134 9	65.77
	3	161	1.155 7	0.135 3	79.60
	4	273	1.156 5	0.135 5	88.30
	5	301	1.157 3	0.135 9	93.83
	6	217	1.158 3	0.136 7	97.25
	7	25	1.159 7	0.137 7	99.17
	8	26	1.161 2	0.138 8	100
	9	3	1.162 9	0.140 1	100

由表 1 可以看出, 改进算法无论从得到的规则数、算法运行时间还是对结果评价的大部分指标来看都优于 Apriori 算法. 从算法本身分析, Apriori 算法无法得到频繁大项集规则是因为其对数据的特定值在数据库中的支持度与置信度进行规则修剪, 而改进后的算法利用时间加权减少了数据维度, 且其规则生成是对数据转换加权和值作属性的组合运算, 从而大大降低了运算复杂度, 而经过聚类后得到的近似值在数据库中的计算相当于对时间边界作了模糊化处理. 因此, 不仅规则支持度增加了, 相应的各项评价标准性能也提高了.

4.3 参数问题

为了分析算法中每个环节的参数对下一步运算及最终结果的影响, 本文对各项参数进行了多值测试.

首先, 测试常量化参数 σ 取不同值的情况; 其次, 测试聚类参数 k 分别取 3 和 6 的情况, 以 Silhouette 作为评价标准, 如图 1 所示, 图中灰色代表 $k = 6$, 黑色代表 $k = 3$.

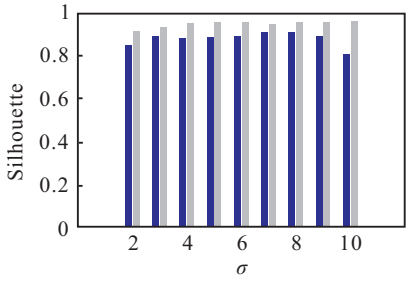

图 1 σ 和 k 对聚类结果的影响

由图 1 可知, 当 k 值一定, σ 取不同的值时, 聚类效果无太大差别, 当聚类效果基本一致时, 所得规则也比较稳定. 从算法本身来看, σ 仅仅起到标记时间的作用, 对于项集的生成及最后的规则修剪影响不大, 这与实际结果保持一致, 因此应尽可能选择较小的基数简化运算. 由图 1 可以看出, 当 $k = 6$ 时聚类效果更好些, 但在 Step 3 规则生成时发现, 聚类数越多, 满足条件的规则也就越少. 这是因为当分类越多越细, 各种性能 (支持度、置信度等) 也相应下降, 但如果增大训练样本集, 则可能消除这种差距, 所以对于聚类参数 k 及规则生成中的最小支持度阈值 $\min\_\sup$ 和最小置信度阈值 $\min\_\mathrm{conf}$, 可根据具体数据特点与专家经验选择, 也可进行多值测试后选取满足需求的最优参数.

5 结　论

本文针对传统关联规则不能有效解决布尔时间序列时间关联性的问题, 利用时间序列数据不同属性的动态变化特性, 结合常量化、聚类与传统关联规则, 提出了一种改进的面向布尔时间序列的关联规则挖掘算法框架. 该算法无论是在运行速度还是结果性能上, 相对于传统算法都有较大提高. 最后, 将该算法应用于实际的中医中风病预后数据集, 解决了症状观测序列与预后预测的关联性问题, 且对其中的参数选择作了测试分析. 分析结果表明, 该算法适用于一般的预后分析、故障预测及相似问题的研究.

本文提出的改进算法框架仅以指数加权法与 k-means 聚类为例, 可根据数据特点与用户需求选择其他常量化方法与聚类算法, 也可对多种算法比较选择最优结果. 该算法在规则生成时可能产生准确度较低的小项集规则方面还有待改进. 虽然可通过调节参数降低这种可能性, 但却是以减少规则数为代价. 如何在规则修剪时直接将这些无效规则过滤将是今后工作的一个重点.

参考文献(References)

[1] MacNeill I B, Umphrey G J. Time series and econometric modeling[M]. Tokyo: D Reidel Pub Co, 1987, 36: 147-163.

[2] Han J W, Kamber M. Data mining : Concepts and techniques[M]. The United States: Morgan Kaufmann Publishers Inc, 2005.

[3] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. Proc of the 1993 ACM SIGMOD Int Conf Management of Data. The United States: ACM Press, 1993, 22(2): 207-216.

[4] Agrawal R, Imielinski T, Swami A. Database mining : A performance perspective[J]. IEEE Trans on Knowledge and Data Engineering, 1993, 5(6): 914-925.

[5] 王金龙, 徐从富. 启发式全局偏序挖掘算法[J]. 模式识别与人工智能, 2008, 21(2): 142-147.
(Wang J L, Xu C F. A heuristic algorithm for global partial order mining[J]. Pattern Recognition and Artificial Intelligence, 2008, 21(2): 142-147.)

[6] 曾海泉, 胡勤友, 周水庚, 等. 基于互关联后继树的时序模式挖掘[J]. 模式识别与人工智能, 2003, 16(3): 299-305.
(Zeng H Q, Hu Q Y, Zhou S G, et al. Mining sequence pattern based on relevant successive trees model[J]. Pattern Recognition and Artificial Intelligence, 2003, 16(3): 299-305)

[7] Yu Chung-Ching, Chen Yen-Liang. Mining sequential patterns from multidimensional sequence data[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(1): 136-140.

[8] 荣冈, 刘进峰, 顾海杰. 数据库中动态关联规则的挖掘[J]. 控制理论与应用, 2007, 24(1): 127-131.
(Rong G, Liu J F, Gu H J. Mining dynamic association rules in databases[J]. Control Theory & Applications, 2007, 24(1): 127-131.)

[9] 沈斌, 姚敏. 一种新的动态关联规则及其挖掘算法[J]. 控制与决策, 2009, 24(9): 1310-1315.
(Shen B, Yao M. A new kind of dynamic association rule and its mining algorithms[J]. Control and Decision, 2009, 24(9): 1310-1315.)

[10] Lo S. Binary prediction based on weighted sequential mining method[C]. The 2005 IEEE/WIC/ACM Int Conf on Web Intelligence. The United States: IEEE Computer Society, 2005: 755-761.

[11] Brin S, Motwani R, Silverstein C. Beyond market baskets: Generalizing association rules to correlations[C]. The 1997 ACM SIGMOD Int Conf on Management of Data. New York: ACM Press, 2009, 24(9): 1310-1315.

[12] 周欣, 沙朝锋, 朱扬勇, 等. 兴趣度——关联规则的又一个阈值[J]. 计算机研究与发展, 2000, 37(5): 627-633.
(Zhou X, Sha C F, Zhu Y Y, et al. Interest measure — another threshold in association rules[J]. J of Computer Research and Development, 2000, 37(5): 627-633.)

[13] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11): 2076-2082.
(Guo Z M, Zhou A Y. Research on data quality and data cleaning: A survey[J]. J of Software, 2002, 13(11): 2076-2082.)

[14] Galhardas H, Florescu D, Shasha D, et al. Declarative data cleaning : Language, model, and algorithms[C]. Proc of the 27th Int Conf on Very Large Dada Bases. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 371-380.

[15] Miller R J, Yang Y. Association rules over interval data[C]. Proc of the 1997 ACM SIGMOD Int Conf on Management of Data. Tucson Arizona: ACM Press, 1997, 26(2): 452-461.

[16] Saab M E, Gotman J. A system to detect the onset of epileptic seizures in scalp EEG[J]. Clinical Neurophysiology, 2005, 116(2): 427-442.