

一种改进的时态关联规则算法

陈 官, 牛秦洲

(桂林理工大学 信息科学与工程学院, 广西 桂林 541004)

摘 要: 针对传统关联规则挖掘中可能会忽略具有时间因素的关联规则的问题, 提出了一种具有时间约束的改进时态关联规则算法. 该算法通过计算模式平均支持度、数据集平均支持度、模式集中度等参数来判断 Apriori 算法所得到的模式是否在某一时间区域数据集上具有较高支持度, 再对该数据集进行进一步挖掘, 以找到更精确的和时间相关的模式, 从而得到单个或不同时间区间上的关联规则以及跨时间区间和跨事务的关联规则. 通过实验分析, 该算法是可行的, 并在实际应用中有一定的意义.

关键词: 数据挖掘; 时态关联规则; 时态数据; 支持度; 集中度

中图分类号: TP311.13

文献标识码: A

文章编号: 1000-7180(2012)06-0189-04

An Improved Temporal Algorithm of Association Rules

CHEN Gong, NIU Qin-zhou

(College of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China)

Abstract: To question that the association rules with time factor may be neglected in data mining to the traditional related rule, an improved temporal algorithm of association rules with time-constrain is proposed, which first judges if the modes obtained by Apriori algorithm has relatively high support degree in data set at a certain time region through calculating the average supported-degree of the modes, the average supported-degree of the data set and the concentration degree of the modes, etc. Secondly this algorithm goes on data mining in the data set in order to find more accurate mode correlated with time, thus gets association rules at single time zone or different time zone or acrossing time zone or affairs. Through experimental analysis, this algorithm is feasible, and has certain meanings in practical application.

Key words: data mining; temporal association rule; temporal data; support; concentration

1 引言

关联规则用于发现交易数据库中不同商品(项)之间的联系, 反映出顾客购买行为模式, 从而用于决策支持, 如分类设计、交叉购物和贱卖分析等. Apriori 算法^[1-2]是被提出最早而且最成功的关联规则算法, 常用于挖掘顾客交易数据库中项集间的关联规则问题, 其核心方法是基于频集理论的递推方法. 此后又出现了很多关于关联规则的研究. 但传统的关联规则挖掘很少考虑关联规则的时间适用性, 而事实上, 每个关联规则都有其成立的时间区域. 因

此, 在挖掘关联规则时附加上某种时态约束会使规则能更好地描述客观现实情况, 也会更有价值. 加上某种时态约束的规则称为时态关联规则^[3-4]. 某些事务和时间是息息相关的, 例如, 采购管理中涉及到采购时间、交货时间、生产时间等. 从这些与时间紧密相连的事务中挖掘有趣的规律是很有价值的.

有些模式或知识只在某个特定的时期出现, 如果不加以时间的约束就直接挖掘的话, 由于它们的支持度在整个数据库中不够高而往往被忽略.

如果把时间做为一个普通项, 将传统关联算法 Apriori 用于时态信息挖掘时也会存在的一些问题:

(1) 如果不对时间项做处理,则时态数据库中相同的时间项极其之少,很可能大部分时间项只出现一次,因此 Apriori 算法第一步扫描数据库得出频繁项集的时候,就会导致大部分时间项因为出现次数太少而不满足最小支持度,这显然是不符合时态数据挖掘要求的。

(2) 时间间隔有起始时间和终止时间两个分值,如何处理起始时间相同或者终止时间相同的时间间隔是一个值得考虑的问题,另一个值得考虑的问题是如何处理有部分时间子集重叠的若干个事务之间的关联^[5]。

文献[6-9]中介绍了相关研究,在此基础上,提出了一个改进的时态关联规则算法,可以挖掘出单个或不同时间区间上的关联规则以及跨时间区间和跨事务的关联规则。

2 改进的时态关联规则算法

2.1 算法思想

(1) 采用分解的方法,先将数据集按照时间属性划分成 n 个小型数据集($\text{SDS}_i, i=1, 2, \dots, n$)^[10],可以减少一次挖掘的数据量,便于计算机处理数据。如何划分是由用户来决定的,用户对数据库中时态信息的类型和范围有一定的了解,根据期望获得的信息,可以按年、月、日等时间属性来划分。划分时间间隔越大,在每个时间段能够获得的关联规则就越多,关联规则所在的时间范围就大;划分时间间隔越小,在每个时间段能够获得的关联规则就越少,关联规则所在的时间范围就小。

(2) 把时间项转化为普通项,加入到事务的项中,形成[事务项集标识,事务项集,时间项]的数据结构。在 n 个数据集的某一个数据集,所有事务的时间项属于同一类,在该数据集中产生的项集所对应的时间项也是同一类,将被赋予该类时间值。另外,用户可以设定规则起始时间 BeginTime 和结束时间 EndTime,以便更好地约束挖掘,提高效率。

(3) 相关定义:

① 模式平均支持度 $\text{Ravgsupp}(p)$:是指某个候选模式(该模式满足事先设定在子数据集中的最小支持度和最小置信度)在所有子数据集中支持度的平均值,可用式(1)计算:

$$\text{Ravgsupp}(p) = 1/n * \sum_{i=1}^n \text{supp}(p_i) \quad (1)$$

式中, n 是子数据集的个数, $\text{supp}(p_i)$ 表示模式 p_i 在该子数据集中的支持度,如该模式在某个子数据

集中的支持度小于最小支持度,则将其在该子数据集中的支持度视为 0。

② 数据集平均支持度 $\text{Savgsupp}(p)$:是指某个子数据集中所有候选模式的支持度的平均值,可用式(2)计算:

$$\text{Savgsupp}(p) = 1/\text{num}(p) * \sum_{i=1}^{\text{num}(p)} \text{supp}(p_i) \quad (2)$$

式中, $\text{num}(p)$ 是该子数据集中所有候选模式的个数, $\text{supp}(p_i)$ 表示模式 p_i 在该子数据集中的支持度。

③ 模式集中度 $\text{centr}_i(p)$:是指某一个候选模式的在第 i 个子数据集中的相对其他候选模式集中出现的程度,可用式(3)计算:

$$\text{centr}_i(p) = (\text{supp}_i(p) - \text{avgsupp}_i(p)) / (1 - \text{avgsupp}_i(p)) \quad (3)$$

式中, $\text{supp}_i(p)$ 表示模式 p 在第 i 个子数据集中的支持度, $\text{avgsupp}_i(p)$ 表示第 i 个子数据集中所有候选模式的平均支持度, $\text{centr}_i(p) \leq 1$ 。

(4) 实际上,大部分事务的时间间隔都呈现“集中分布”的情况,也即在某些时间间隔区域,集中了相当大一部分的事务,而在另外一些时间间隔区域,却只有稀疏的几个事务。对于某一类事务来说,也可能集中在某些时间区域出现,因此,按时间划分数据集时,若时间间隔取值过大,则导致挖掘不精确,无法深入挖掘;若时间间隔过小,则导致挖掘效率低下,并且有不少时间间隔类上根本就没有事务。因此,对于事务密集区域,在时间上进行“细划分”;对于事务稀疏的区域,在时间上进行“粗划分”。可以根据模式平均支持度 $\text{Ravgsupp}(p)$ 和模式集中度 $\text{centr}_i(p)$ 来判断事务的集中度。先根据经验划分时间间隔,得到 n 个数据集,运行算法后,得到候选模式,若某一候选模式在某个子数据集中的支持度大于模式平均支持度 $\text{Ravgsupp}(p)$,并且其模式集中度大于设定阈值 M ,则可以将该数据集或多个相邻数据集重新划分时间间隔,再次运行算法,以得到更准确的模式。

(5) 算法可以挖掘多种与时间相关的关联规则。

① 单个或不同时间区间上的关联规则。根据规则起始时间 BeginTime 和结束时间 EndTime 以及 n 值的确定,可以在 n 个数据集上运行算法,得出 n 个模式集合,其中 $n \geq 1$ 。例如,将某一年划为 4 个季度($n=4$),运行算法可以得到 4 个季度中不同的关

联规则集合,从而得出象“第二季度无烟块煤采购次数最多”,“第一季度和第三季度采购齿轮最多”等这样的规则。

② 跨时间区间和跨事务的关联规则. 在 n 个数据集 $SDS_i(i=1,2,\cdots,n)$ 中运行算法,可以得到 n 个关联规则集合 $ARS_i(i=1,2,\cdots,n)$,其中的关联规则都满足最小支持度和置信度. 每个 ARS_i 中的项集可以表示为[项集标识,项集,时间项],如果不考虑时间项,可以直接将 n 个 ARS_i 的项集进行合并,得到整个时间区域中的关联规则. 例如,将 2010 年的采购事务分为 12 个月,项集合并后可以得到象“2010 年采购自动排气阀的时候一般都会同时采购焦亚硫酸钠”这样的跨事务的规则. 如果考虑到有些事务“集中分布”的情况,可以计算出 ARS_i 中某一模式的模式集中度 $centr_i(p)$ 和该模式的平均支持度 $Ravgsupp(p)$. 若该模式的支持度大于平均支持度 $Ravgsupp(p)$ 且模式集中度 $centr_i(p)$ 大于设定阈值 M ,则可将这个数据集或多个相邻数据集再细分时间间隔,再次运行算法,以得到更准确的模式,例如“每年的三、四月份是采购无烟煤的高峰期”等. 另外,如果项集中加上时间项,合并 n 个 ARS ,也可以得到诸如“上个月采购较多不锈钢圆钢,这个月一定会订购更多的方钢”这样的跨时间区间的规则。

(6) 最后得到的模式需要经过领域专家裁剪后,才可以成为可用的知识。

2.2 算法描述

算法: MiningTemporalPatterns
输入: 划分时间区间数 n , 起始时间 BeginTime, 结束时间 EndTime, 最小支持度 minsupp, 最小置信度 minconf, 阈值 M , 数据集.
输出: 频繁项集, 关联规则.
算法开始:
(1) 根据需求和经验, 将目标数据集按时间属性划分为 n 个子数据集, 例如, 按年、按季或按月划分;
(2) 在 n 个子数据集上运行 Apriori 算法, 可以得到 n 个关联规则集合 $ARS_i(i=1,2,\cdots,n)$, 记为 ARS_1 , 其中的频繁项集按[项集标识, 项集, 时间项]结构存储;
(3) 直接将 n 个 ARS_i 的项集进行合并, 得到整个时间区域中的关联规则集合, 记为 ARS_2 ; 如果项集中加上时间项进行合并, 可以得到跨时间区间的关联规则集合, 记为 ARS_3 ;
(4) 按式(1)~式(3)计算 ARS_i 中某一模式的

模式平均支持度 $Ravgsupp(p)$ 、数据集平均支持度 $Savgsupp(p)$ 和模式集中度 $centr_i(p)$, 若该模式的支持度大于平均支持度 $Ravgsupp(p)$ 且模式集中度 $centr_i(p)$ 大于设定阈值 M , 则可将这个数据集或多个相邻数据集再细分时间间隔, 再次运行第(1)~(3)步, 以得到更准确的模式集合, 记为 ARS_4 ;

(5) 合并 ARS_1 、 ARS_2 、 ARS_3 、 ARS_4 , 然后利用领域知识将不需要或错误的规则裁剪掉, 得到最后的规则集合 ARS .

算法结束.

3 实验

为了验证算法的有效性, 用 Microsoft Visual C++6.0 实现算法, 采用 IBM Quest Project 的数据生成程序生成实验所需要的合成数据集 DataSet, 在内存 512 MB, CPU 主频 2.6 GHz, 操作系统为 Windows XP 的计算机上进行实验。

为了进一步验证算法的有效性, 在生成 DataSet 时有意让某些项集频繁出现在某一时段内, 所有时间项均包含在 2010 年内. DataSet 的特性如表 1 所示。

表 1 数据集 DataSet 的特性

Data Set	Items	Avarage Length	Trans.
DataSet	605	5.5	195 512

第一步, 先将数据集 DataSet 按季度划分为 4 个子数据集, 划分结果如表 2 所示。

表 2 子数据集的特性

Data Set	Data1	Data2	Data3	Data4
Trans.	50 000	50 000	50 000	45 512

第二步, 用 Apriori 算法求出各子数据集 1-项集(最小支持度设置为 0.01), 在这里直接将频繁 1-项集视为一条规则, 频繁 1-项集的支持度就是该规则的支持度, 各子数据集挖掘出来的频繁 1-项集个数如表 3 所示。

表 3 各子数据集的频繁 1-项集个数

Data Set	Data1	Data2	Data3	Data4
Num. of 1-Frequest Itemset	57	74	34	65

从第二步得到的规则中选出一些有代表性的规则, 如表 4 所示。

表 4 部分规则在各子数据集的支持度

ID of rule	Data1	Data2	Data3	Data4
3	0	0.172 1	0	0
12	0.132 0	0.038 1	0.021 1	0.072 1
31	0.162 5	0.025 2	0	0
38	0	0	0.021 2	0.125 6
46	0.032 1	0.024 1	0.152 3	0.030 2
53	0.134 2	0.042 1	0.020 3	0

从表 4 中可以得到支持度比较高的规则:第一季度的规则 12、31、53,第二季度的规则 3,第三季度的规则 46,第四季度的规则 38. 很明显地可以看到这些规则在其中一个季度中的支持度远大于平均支持度 $\text{Ravgsupp}(p)$. 对于规则 3,它在 Data2 的模式集中度 $\text{centr}(p)$ 显然比较大,同样规则 38 和规则 46 在 Data4 和 Data3 的模式集中度 $\text{centr}(p)$ 也比较大,则可以将这些数据再细分时间间隔,运行算法第三、四步,再次挖掘与这几个规则相关的模式.

最后,运行算法第五步,除去不需要的规则,保留有用的规则.

实验结果表明,该算法可以挖掘出有用的时态关联规则.

4 结束语

本文提出一种改进的时态关联规则算法,将数据集按时间属性划分成若干个子数据集,时间加入到事务的项集中,并给出了模式平均支持度、数据集平均支持度、模式集中度等相关定义,可挖掘出单个或不同时间区间上的关联规则以及跨时间区间和跨事务的关联规则,并经过了实验的证明. 该算法在实际应用中有较大的意义,可对企业数据、股票数据、银行数据、气象数据等时态数据进行分析和研究.

参考文献:

(上接第 188 页)

- [5] 韩丽霞,王宇平. 图着色问题的新遗传算法[J]. 西安电子科技大学学报:自然科学版,2008,35(2):309-313.
- [6] 胡平,彭记奎. 粗糙集-遗传神经网络在挖掘机故障诊断中的应用研究[J]. 微电子学与计算机,2011,28(3):55-58.

作者简介:

陈泰伟 男,(1986—),硕士研究生. 研究方向为无线射频、

- [1] Agrawal R et al. Mining association rule between sets of items in large database[C]//Proc the ACM SIGMOD International Conference on Management of Data, Washington:IEEE,1993:207-216.
- [2] Agrawal R. Fast algorithms for mining association rules[C]//Proc the 20th International Conference on Very Large DataBases. Santiago de Chile:IEEE, 1994: 478-499.
- [3] Ale J M,Rossi G H. An approach to eiscovering temporal association rules [C] // Proc of the 2000 ACM Symp on Applied Computing. [s. l.]:ACM, 2000:294-300.
- [4] Chen X,Petrounias I,Heathfield H. Discovering temporal association rules in temporal databases[C]//Proc IADT 98,1998:371-376.
- [5] 辜炜东. 事务数据库中的时态信息挖掘[J]. 计算机工程与应用,2004(18):174-175
- [6] 欧阳为民,蔡庆生. 在数据库中发现具有时态约束的关联规则[J]. 软件学报,1999,10(5):527-532.
- [7] 李少年,孟志青. 双事件时态关联规则的知识发现问题[J]. 计算机科学,2001,29(8):71-73.
- [8] Yingjiu li,Peng Ning. Discovering calendar based temporal association rules[J]. Journal Datu Knowledye Engineering-Special issue: Temporal representation and reasoning, 2003,44(2):193-218.
- [9] Agrawal R,Srikant R. Mining sequential patterns[J]. IEEE Computer Society,1995,16(3):3-14.
- [10] 曾德胜,张师超. 大型时态数据库中的 Burst 模式挖掘[J]. 计算机应用,2006,26(10):2413-2416.

作者简介:

陈 宫 男,(1976—),硕士,讲师. 研究方向为云计算、数据挖掘.

牛秦洲 男,(1956—),博士,教授. 研究方向为网络控制.

人工智能、信息系统.

吴黎明 男,(1964—),教授,硕士生导师. 研究方向为光机电一体化关键技术、数字图像处理技术.

吴学文 男,(1985—),硕士研究生. 研究方向为嵌入式系统开发.

项 颖 男,(1968—),博士,教授,硕士生导师. 研究方向为平板显示器件、电信企业数据仓库的规划和设计.