doi:10.16055/j.issn.1672-058X.2015.0012.009

基于频繁项集分类统计的增量式关联规则应用*

刘 绍 清
(福州职业技术学院 计算机系,福州 350002)

摘 要:针对商业交易数据构成项目繁多、动态数据增加量大、历史数据量更大的特点,根据频繁项集的商业特征,分为新生、成熟、老化、过期 4 种类型并分类统计;提出了基于分类统计增量地挖掘新增业务数据中关联规则的算法,算法只需两次扫描新增数据库,无需扫描历史数据库,算法将发现的规则按照其反应的商业特征分为 4 种类型:新生规则、成熟规则、老化规则、过期规则,在提升规则内容识别效率的同时,强化规则特点的识别能力。

关键词:频繁项集分类;统计信息;增量式更新;关联规则分类
中图分类号:TP3 **文献标志码**:A **文章编号**:1672-058X(2015)12-0043-05

商业活动是一个不断进行的过程,活动的行为模式随时间推移而变化,体现在数据上就是:商业活动每天会产生大量的数据,积累下来就是海量的数据;从新数据中也可能存在新规律需要发掘;这些新发现的规律在随着新数据的加入将逐渐失效,也可能持续有效。增量式关联规则在有效地发现商业活动中交易行为的动态规律方面有其特殊的优势,因此,很多学者对增量式关联则进行了研究,提出了 FUP[1]、IUA[2]、FUFLA[3]、FIUA[4]以及很多改进算法与应用[5-9],但是这些算法在提升规律发现效率的同时,没有考虑到随着新数据的增加,一些已经发现的商业规律可能会逐渐失去价值,而一些原先不成熟的规律会逐渐成熟,也就无法识别这些规律是处于新发现的,还是处于逐步失效的规律,还是逐步成熟的规律。鉴于此,针对商业交易数据的动态特征,提出一种基于数据频繁项集历史统计信息的增量式关联规则算法(HS_IAR),只需对本期新增数据集扫描两次,而无需对历史数据集进行扫描,就能够找出所有大于事先给定的最小支持度的关联规则,并辨别关联规则所处的生命周期阶段,将其分为 4 种类型:新生规则、成熟规则、老化规则、过期规则,在提升规则内容识别效率的同时,强化规则特点的识别能力。

1 FUP 算法简介

1.1 算法基本思想

(1) 扫描新事务数据集(d)生成频繁项目集 $L(d)$;

(2) 获得 d 中所有的频繁项集之后,将其和 D 的频繁项集合并,合并过程中,对任意项集可能存在 4 种情况:

情况 A:在 D 中是频繁项集,在 d 中是频繁项集;

情况 B:在 D 中是非频繁项集,在 d 中是频繁项集;

情况 C:在 D 中是频繁项集,在 d 中不是频繁项集;

收稿日期:2015-06-08;**修回日期**:2010-06-15.
* **基金项目**:福建省教育厅 A 类项目“数据挖掘批量建模技术”(JA12401).
作者简介:刘绍清(1974-),男,福建福州人,副教授,硕士,从事数据挖掘应用研究.E-mail:1132655585@qq.com.

情况 D:在 D 中是非频繁项集,在 d 中是非频繁项集。

① 对于情况 A,放入当前事务数据库(d+D)的频繁项目集 L(d+D)中

② 对于情况 B 和情况 C,需要判断频繁项集 $x_{D\cup d}$在合并后的数据集($D\cup d$)中是否还是频繁,大多数算法往往采用公式:

$$\text{support}(x_{|D\cup d|}) = \frac{\text{count}(x_d) + \text{count}(x_D)}{|d| + |D|} \times 100\% \tag{1}$$

公式说明:$|d|$和$|D|$分别是数据集 d 和 D 的事务数或记录数;如果支持度大于等最小支持度,则将其放入 L(d+D)中,否则不放入。

③ 对于情况 D,则不需要处理。

1.2 FUP 算法不足

在处理情况 B 和情况 C 的时候,FUP 会存在以下两个不足:

不足 1:对于情况 B 需要扫描历史交易数据集 D,扫描需要耗费大量的资源;

不足 2:对于情况 B 和 C,由于 D 中的数据记录数一般会远远大于 d 中的记录数,导致公式(1)中项集 x_d 的支持数 count(x_d)在合并后的数据集中作用太小,甚至小到可以被忽略的地步,导致 d 中的新情况(新规则出现,老规则过期)都被历史数据否定了,这种否定现象要持续相当长一段时间,直到 count(x_D)积累了足够大为止,这将导致算法发现新关联规则能力大幅度滞后,最终影响算法的应用。

2 HS_IAR 算法思路

2.1 策略 1——对发现的规则分类对待

结合着 4 种情况体现出来的商业业务特点,HS_IAR 将发现规则根据其所处生命周期位置分成四类:新生规则、成熟规则、老化规则、过期规则。类似地,频繁项集也分为新生频繁项集、成熟频繁项集、老化频繁项集、过期频繁项集,收集四类频繁项集的相关的统计信息,在后续新增交易数据集合并的时候,只要根据这些历史统计信息以及每一类的特征,而不需要扫描历史交易数据库 D,也不需要反复扫描新增交易数据库 d,就能有效地处理情况 A 到情况 D。

定义 1:一个项集 X,在 D 中是非频繁项集,在 d 中是频繁项集,则称 X 为新生频繁项集;

定义 2:D 中一个成熟频繁项集 X,在 d 中是非频繁项集,则称 X 为老化频繁项集;

定义 3:X 是 D 中一个老化频繁项集或者新生频繁项集,在连续 L 期新增交易数据集 $d_i(i=1,2,\cdots,L)$ 中,如果$\frac{FT_X}{L}$<minpercent(其中,FT_X 表示 L 期中 X 是频繁项集的次数,minpercent 是一个给定的比例值),则称 X 为过期频繁项集;如果$\frac{FT_X}{L}\geqslant$minpercent,则称为成熟频繁项集。

交易行为的变迁,导致交易规则性质发生变化,具体有:

定义 4:$X\subset I$,$Y\subset I$,且 $X\cap Y=\varnothing$,蕴涵式 $X\Rightarrow Y$ 称为关联规则。如果 $X\cup Y$ 与 X 满足:二者都是成熟频繁项集,意味着规则重复多次出现,而非偶然因素所致,为成熟关联规则;如果二者有一个是过期频繁项集,意味着规则有一段时期都不再,称其为过期规则;如果二者都不是过期频繁项集,但有一个老化频繁项集,意味着规则之前有过一段时间没有出现,但是近期又频繁出现,这是失效的征兆,则为老化规则;如果上述 3 种情况都不满足,则关联规则为新生规则。

可见,情况 A 是一种成熟交易模式的表现,相关项集可归入成熟频繁项集中重点关注;情况 B 预示着一种新的交易模式可能出现,相关项集可归入新生频繁项集中,继续跟踪;情况 C 是一个规则老化的开始,相关项集归入老化频繁项集中,继续跟踪;情况 D 不会出现规则,无需处理。

2.2 策略 2——删除事务中非频繁数据项,减少候选项集

定理 1:一个频繁项集的所有非空子集一定是频繁项集,若一个项集是非频繁项集,则该项集的超集也

一定是非频繁项集。

对新增交易数据 d 第一次扫描可得到 1-频繁项集 X_1^d,从定理 1 可知:通过 $t'_p=t_p\cap X_1^d$ 删除每个事务包含的数据项中的非频繁项,可大幅度减少候选项集数量,而不遗漏频繁项集。

2.3　策略 3——限制候选项集最大项数,减少候选项集

推论 1:对于数据集 d,对于给定的最小支持度 minsupport,如果用 n_k 表示包含 k 个数据项的事务数,则可以得到一个最大的 k($1\leqslant k\leqslant m$,m 表示 d 所有事务包含的最多数据项数),使得 $\sum n_k \geqslant |d|\times(1-\text{minsupport})$,保证 d 所有事务的(k+1)-项集都是非频繁项集。

从历史数据频繁项集统计信息中可以得到历史 1-频繁项集 X_1^D,对新增交易数据 d 第一次扫描时,通过 $t'_p=t_p\cap X_1^D$,可以统计每个事务中频繁的事务项数 $|t'_p|$,根据推论 1,可以得到 Max_k,在第二次扫描 d 的时候,无需生成 Max_k 以上的候选项集。Max_k 计算理论上应该用新增数据集 d 的 X_1^d 计算,由于第一次扫描 d 的时候,X_1^d 还不可知,用 X_1^D 代替不影响效果,这是因为对于一个成熟的规则,二者的效果是一样的,而对于偶然的因素导致的规则,即使被忽略了也影响不大。

3　算法步骤

输入:新增交易数据集 d、最小支持度 minsupport、最小置信度、分类频繁项集的历史统计信息 X_1^D 和 X_D;

输出:合并后分类频繁项集的统计信息 $X_1^{D'}$ 和 X'_D、分类关联规则;

处理过程:

(1) 第一次扫描新增交易事务数据库 d,获得以下内容:事务数 $|d|$、1-项频繁集 X_1^d 及其各项的支持数 count(X_1)、根据 X_1^D 统计的每个事务包含的频繁项数 $|t'_p|$ (应该是 tp)($p=1,2,\cdots,|d|$),以及每个项数对应的事务数 $n_k(1\leqslant k\leqslant \max(|t'_p|))$;

(2) 根据推论 1 方法,计算 Max_k,在第二次扫描 d 时,只处理 2-项集到 Max_k-项集;

(3) 第二次扫描新增交易事务数据库 d,对 d 每条事务 t_p 的处理逻辑是:$t'_p=t_p\cap X_1^d$;按照策略 2,用 X_1^d 删除非频繁项,不用 X_1^D;按照策略 3 生成项数不大于 Max_k 的候选项集,并归并入集合 C 中。

(4) 从 C 中剔除所有非频繁的项集,就得到了 d 中所有的频繁项集 X_d;

(5) 按照策略 1,分类合并 X_1^d 和 X_1^D 生成 $X_1^{D'}$,合并 X_d 和 X_D 生成 X'_D;

(6) 根据新合并后频繁集,重新生成分类关联规则集。

4　算法实际应用及效果分析

用算法对一家大型百货超市 2012 年 11 月和 12 月数据进行分析,超市有 5 家门店,大约 10 900 种商品,1 600 小类,一天 10 000 笔以上 21 d,最多 13 663 笔交易,5 000 笔以下 25 d,最少 1 424 笔交易,一笔交易 2 项以上产品的大约 5%,最多的一笔 78 种商品,设定最小支持度为 2%,最小置信度为 60%,分别用 matlab 实现 FUP 算法和 HS_IAR,效果如下:

4.1　运行效率分析

两个模型每天分析耗费时间如图 1、图 2 所示,随着历史数据的逐步增加,从图 1 可以看出 HS_IRA 每天建模时间基本稳定,而 FUP 则是逐步增加;从图 2 可知,两种算法和新增数据量都有一定关系,HS_IRA 耗时和新增数据量的关系比较稳定,而 FUP 因为访问历史数据库不确定导致二者关系不是那么稳定。

4.2　规则发现能力

表 1 是对比了一个规则(巴布豆童装(巴布豆童鞋)发现的过程,从表 1 看出,两种算法都能发现规则,但是 HS_IAR 比 FUP 早了 13 d 发现,在发现规则的及时性上更有优势。

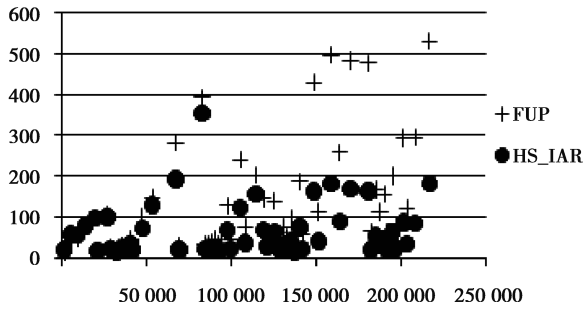

图1 历史数据量与时间

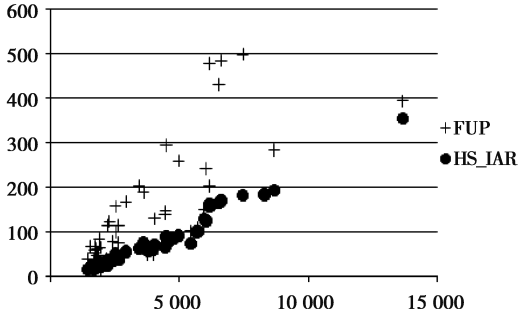

图2 新增数据量与时间

表1 规则发现能力对比表

日 期	总交易笔数/笔			规则交易笔数/笔		FUP	HS_IRA
	历史	新增	规则历史	当日	累计	(5)/(1)	(5)/(3)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
35	125 769	4 436	4 436	167	167	0.13%	3.76%
36	127 546	1 777	6 213	134	301	0.23%	4.84%
37	129 381	1 835	8 048	719	1 020	0.77%	12.67%
…	…	…	…	…	…	…	…
46	159 374	7 463	38 041	288	3 103	1.89%	8.16%
47	164 332	4 958	42 999	252	3 355	1.96%	7.80%
48	170 946	6 614	49 613	202	3 557	2.03%	7.17%

5 总 结

针对商业交易中商品项数多、历史数据量远大于新增交易数据量的特点,提出了基于频繁项集分类统计信息的增量关联规则算法,采取3种策略来减少数据库扫描次数和候选项集数量,对挖掘的关联规则按其体现商业活动特点分成四类区别对待,算法具有较高的规则内容识别效率,较强的规则特点识别能力。

参考文献:

[1] CHEUNG D W,HAN J,NG V T,et al.Maintenance of Discovered Association Rules in large Database:An Incremental Updating Technique[C]//In:Proc of the 12th Int Conf on Data Engineering,New Orleans,Louisiana,1996:106-114

[2] 冯玉才,冯剑琳.关联规则的增量式更新算法[J].软件学报,1998(4):301-306

[3] 朱玉全,孙志挥,赵传申.快速更新频繁项目集[J].计算机研究与发展,2003(1):94-99

[4] 朱玉全,孙志挥,季小俊.基于频繁模式树的关联规则增量式更新算法[J]计算机学报,2003(1):91-96

[5] 李松生,赵燕伟,顾熙仁,等.改进的FUP算法在五金产品质量分析系统中的应用[J].吉林大学学报:工学版,2012(9):251-254

[6] 唐璐,江红,上官秋子,等.一种改进的关联规则的增量式更新算法[J].计算机应用与软件,2012(4):246-248

[7] 杜焕强,俞立峰.一种高效的关联规则连续增量更新改进算法[J].哈尔滨师范大学学报:自然科学版,2015(5):49-52

[8] 郑亚军,胡学钢.基于PFP的关联规则增量更新算法[J].合肥工业大学学报:自然科学版,2015(4):500-503,551

[9] 陈丽芳.基于Apriori算法的购物篮分析[J].重庆工商大学学报:自然科学版,2014(5):84-89

Application of Incremental Association Rules Based on Statistical Information of Classified Frequent Itemsets ——Taking Supermarket as an Example

LIU Shao-qing

(Department of Computer, Fuzhou Institute of Technology, Fuzhou 350002, China)

Abstract: Business activities always generate large dataset and accumulate much larger dataset with many items in each transaction. According to the commercial character, frequent itemsets are classified into four classes: new, mature, aging, expired, and an incremental updating algorithm based on statistical information of classified frequent itemsets is put forward. The algorithm only scans newly increased transaction dataset twice without scanning original transaction dataset, and it can also classify all rules into four classes: new, mature, aging, expired with strong rule identification capability as well as higher rule identification efficiency.

Key words: classified frequent itemsets; statistical information; incremental updating; classified association rules

(上接第 7 页)

参考文献:

[1] 屈克.区间直觉模糊集熵的构造及其基本性质[J].重庆文理学院学报,2010,29(3):21-24
[2] 毛军军.基于一种新的信息熵的区间直觉模糊集多属性决策分析[J].合肥师范学院学报,2011,29(6):4-7
[3] 毛军军.基于熵和相关系数的直觉模糊多属性决策方法[J].合肥师范学院,2012,32(11):3002-3004,3017
[4] 陈启斐.皖江城市带承接能力差异性研究[D].南京:南京财经大学,2011
[5] 安徽省统计局.2010-2013 安徽统计年鉴[Z].北京:中国统计出版社,2010-2012

The Decision Making Model of Evaluating the Ability of Undertaking Industrial Transfer of Cities Along Yangtze River in Anhui Based on Interval Valued Intuitionistic Vague Entropy

DING Xin[1], TONG Wan-ning[1], CHAN Hui-xian[1]
XU Dan-qing[1], MAO Jun-jun[1,2]

(1.School of Mathematical Science, Anhui University, Hefei 230601, China; 2.Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei 230039, China)

Abstract: In this paper, a novel vague entropy is proposed. Based on the new vague entropy, a new determination of attribute weight is given so that a new interval valued intuitionistic fuzzy multiple attributes decision making model is achieved. The index system is constructed to evaluate the ability of undertaking industrial transfer. The new multiple attributes decision making method, combining with the index system, helps to evaluate the ability of undertaking industrial transfer of cities along Yangtze River in Anhui. The paper proposes the appropriate destinations of the industrial transfer through analyzing the comprehensive value of each city.

Key words: ability of undertaking industrial transfer; new vague entropy; interval intuitionistic fuzzy sets; multiple attributes evaluation method