文章编号:1004-1478(2011)04-0068-04

基于兴趣度剪枝的Apriori优化算法

刘上力, 杨清

(湖南科技大学 网络信息中心, 湖南 湘潭 411201)

摘要:鉴于关联规则挖掘中的Apriori算法在挖掘潜在有价值、低支持度模式时效率较低,因此提出一种优化的Apriori挖掘算法,即在频繁项集挖掘中引入项项正相关兴趣度量剪枝策略,有效过滤掉非正相关长模式和无效项集,从而扩大了可挖掘支持度阈值范围. 实验结果表明,该算法是有效和可行的.

关键词:Apriori算法;频繁项集;兴趣度;项项正相关;剪枝

中图分类号:TP392 **文献标志码**:A

Optimized Apriori algorithm based on interestingness measure pruning

LIU Shang-li, YANG Qing

(*Network Infor. Center, Hunan Univ. of Sci. and Tech., Xiangtan* 411201, *China*)

Abstract: To solve the problem that the Apriori algorithm of mining association rules in database mining is not quite effective in the process of mining potentially valuable low-support patterns, an optimized apriori mining algorithm was proposed. This algorithm exploits an efficient pruning strategy which uses the interestingness measure to filter the non-positive correlated long model and invalid itemsets. The range of support threshold is expanded. The experimental results indicated that the given algorithm was efficient and feasible.

Key words: Apriori algorithm; frequent itemset; interestingness measure; between-items positive correlation; pruning

0 引言

关联规则[1]是数据挖掘的重要研究领域,它包括频繁项集挖掘和关联规则发现2个过程,挖掘的总体性能主要由频繁项集挖掘决定. R. Agrawal等[2]在1993年首次提出Apriori算法,该算法是布尔关联规则挖掘频繁项集的原创性算法,它通过连接和剪枝操作生成频繁项集,此后出现的各种类Apriori算法大多利用这种策略来提高频繁项集挖掘效率. 但对于低支持度下大型稠密数据库的频繁项集,由于频繁项集数量和长度的急剧增长,利用Apriori算法和类Apriori算法则难以挖掘.

本文拟结合现有兴趣度方法[3]和基于约束条件的项集剪枝思想[4],提出新的兴趣度量度,即将兴趣度量嵌入到Apriori算法,实现频繁项集兴趣量度剪枝,以减少低兴趣度长模式频繁项集的生成.

收稿日期:2011-05-13

基金项目:湖南省教育厅重点科学研究项目(10A028);湖南省科技计划项目(JT3031)

作者简介:刘上力(1978—),男,湖南省湘潭市人,湖南科技大学工程师,硕士,主要研究方向为数据挖掘.

1 兴趣度量度

传统的关联规则一般采用支持度－置信度框架，这种框架容易产生无意义、冗余、甚至是误导的关联规则[5-6]. 为了解决这些问题，研究者提出了兴趣度量度 $F(A,B)$. 兴趣度量结果可反映出 A(或 B)的出现对 B(或 A)的出现的提升程度. 兴趣度一般分为客观兴趣度、主观兴趣度和基于语义的兴趣度3种[7]. 其中，客观兴趣度是研究和应用的重点.

研究者较早提出了客观兴趣度是提升度度量 $lift(A,B)=P(A,B)/(P(A)P(B))$[8]，其中 $P(A,B)$ 表示 A 与 B 并的概率(下同)，用于评估一个出现“提升”另一个的程度. 随后，研究者提出了多种兴趣度，但对于应用而言，能适用各类挖掘要求和场合的兴趣度是不存在的[7]. 本文提出一种新的兴趣度.

设 $I=\{I_1,I_2,\cdots,I_m\}$ 为 m 个不同项的集合，给定 l 条事务的数据库 $D=\{T_1,T_2,\cdots,T_l\}$，其中，$T\in I$.

定义1 对于项集 $X\in I$，X 在 D 中支持度计数是指 D 中 X 的事务数，记为 $X.count$，则称 $X.sup=P(X)=X.count/l$ 为 X 在 D 中的支持度.

定义2 若 $X.sup\geq minsup$(最小支持度阈值)，则 X 为 D 的频繁项集.

定义3 给定一个项集 $X=A\cup B$，兴趣度量 $F(A,B)$ 定义为

$$F(A,B)=\frac{(P(A,B)-P(A,\neg B)+P(A,B)-P(\neg A,B))/2}{P(A)+P(B)} \quad ①$$

它主要考察项集并集概率与其中一项不出现时的并集概率的差异，这种差异反映了项集间真正的蕴含关系. 分母作为标准因子，使得量度值仅受 A,B 和 AB 概率的影响，而不受事务总个数的影响；分子中的算术均值运算对于量度值间的比较没有意义，可忽略. 由于 $P(A,\neg B)=P(A)-P(A,B)$，$P(\neg A,B)=P(B)-P(A,B)$，可得到①的简化式

$$F(A,B)=\frac{4P(A,B)}{P(A)+P(B)}-1 \quad ②$$

为有效控制规则数量并保证规则质量，文献[8]总结了评价兴趣度度量标准的7个原则：

1)如果 A 和 B 统计独立，$F(A,B)=0$；

2)其他参数不变时，$F(A,B)$ 随 $P(A,B)$ 单调递增；

3)其他参数不变时，$F(A,B)$ 随 $P(A)$ 或 $P(B)$ 单调递减；

4)A,B 置换时，$F(A,B)$ 应保持不变；

5)零不变性，即不包含 A,B 的记录的增减对 $F(A,B)$ 结果应没有影响；

6)兴趣度阈值的范围应该是固定的；

7)兴趣度的语义表达应该是容易理解的.

这些原则可作为评价相关度量的依据. 表1为定义3给出的兴趣度量度与置信度、全置信度[3]、提升度之间依据以上原则给出的评估分值对比. 结果表明，定义3给出的兴趣度得分最高，其中，遵循原则的给1分，否则记0分.

表1 不同量度间的评估对比

原则	置信度	全置信度	提升度	本文量度
1)	0	0	0	0
2)	1	1	1	1
3)	1	1	1	1
4)	0	1	1	1
5)	1	1	0	1
6)	1	1	0	1
7)	1	0	1	1
合计	5	5	4	6

表2给出了一组事务数据集，以及对应的相依表和4种量度的值. A 和 B 在数据集 $D_1\sim D_2$ 正相关，在 D_3 中属于并发关系，统计独立，在 $D_4\sim D_7$ 负相关. 对于独立情况，除了提升度外，其余3种度量都是好的指示器，除此以外，置信度由于忽略了 $P(B)$ 的影响，在 D_6 上出现误导；全置信度表现较好，但仅考虑最小置信度，在 D_4 和 D_5 间无法区别相关程度；提升度因不具备零不变性，受零事务影响误差较大；只有本文提出的量度在所有数据集中均有较好表现，不受零事务影响，正负相关判断准确，取值范围规范.

2 频繁项集挖掘算法

挖掘者往往根据应用的不同而选择特定相关量度，例如求解关联束挖掘[9]、前后项集对称型应用等问题，需要量度规则中项项间正相关. 本文在Apriori挖掘算法中引入兴趣度，实现了项项正相关

表2 使用不同数据集的相依表比较4种量度

数据集	AB	$\neg AB$	$A\neg B$	$\neg A\neg B$	置信度	全置信度	提升度	本文量度
D_1	1 000	100	100	100 000	0.91	0.91	83.64	0.82
D_2	1 000	100	100	10 000	0.91	0.91	9.26	0.82
D_3	1 000	1 000	1 000	10 000	0.50	0.50	3.25	0.00
D_4	100	1 000	1 000	100 000	0.09	0.09	8.44	-0.82
D_5	1 000	100	10 000	100 000	0.09	0.09	9.18	-0.67
D_6	10 000	100 000	1000	100 000	0.91	0.09	1.74	-0.67
D_7	0	10 000	100	100 000	0.00	0.00	0.00	-1.00

频繁项集的挖掘,同时,通过频繁项剪枝以提高算法的效率.

2.1 项项相关兴趣度

为适应频繁项兴趣度剪枝,对②式进行改进.

定义4 给定一个项集 $X=\{i_1,i_2,\cdots,i_n\}$,$n>1$,则 X 的项项相关兴趣度 $F'(X)$ 定义为

$$F'(X)=\min\left\{\frac{4P(i_j,i_k)}{P(i_j)+P(i_k)}-1\right|\forall j,\forall k\in 1,\cdots,n;j\neq k\}$$

该量度具有良好的性质:1)它具有稳定的上下界[-1,1];2)具有良好的反单调性质,可以利用该性质实现 Apriori 算法中频繁项集挖掘的项集剪枝,以减少长模式和低兴趣度候选项集的产生;3)当 $F'(X)>0$ 时,可确保 X 中任意一项的发生均能提升 X 的其余项发生的概率,表明 X 中的项两两正相关,以该量度作为频繁项集剪枝的标准可得到项项正相关频繁项集.下面给出该量度反单调性质的证明.

性质1 当项集 $X'\subset X$,θ 为阈值,如果 $F'(X')<\theta$,则有 $F'(X)<\theta$,即 $F'(X)$ 具有反单调性质.

证明 设 $X'=\{i_1,i_2,\cdots,i_{n'}\}$,$X=\{i_1,i_2,\cdots,i_{n'},\cdots,i_n\}$,则

$$F'(X)=\min\left\{\frac{4P(i_j,i_k)}{P(i_j)+P(i_k)}-1\right|\forall j,\forall k\in 1,\cdots,n;j\neq k\}\leqslant$$

$$\min\left\{\frac{4P(i_j,i_k)}{P(i_j)+P(i_k)}-1\right|\forall j,\forall k\in 1,\cdots,n';j\neq k\}=F'(X')<\theta$$

性质成立,得证.

2.2 Apriori 频繁项集兴趣度剪枝算法

输入:事务数据库 D,最小支持度计数阈值 *minsup*,最小兴趣度阈值 θ.

输出:D 中的频繁项集 L.

```
Begin
1) L1 = find_frequent_1 - itemsets(D);
2) for (k=2;Lk-1≠φ;k++){
     Ck = apriori_gen(Lk-1);
3)   for each 事务 t∈D{//扫描 D 用于计数
       Ct = subset(Ck,t);//得到 t 的子集
       for each 候选 c∈Ct
         c.count++
       }
       //兴趣度剪枝,
       //即 F'(c)<θ 的候选 c 直接剪枝掉
4)   Lk = {c∈Ck|c.count≥minsup,F'(c)≥θ}
   }
5) return L = ∪kLk;
End
 procedure apriori_gen(Lk-1)
1) for each 项集 l1∈Lk-1
     for each 项集 l2∈Lk-1
       if(l1[1]=l2[1])∧(l1[2]=l2[2])∧…∧
         (l1[k-2]=l2[k-2])∧(l1[k-1]<l2[k-1])
         then{c=l1><l2;//连接步:产生候选
         if has_infrequent_sebset(c,Lk-1) then{
           delete c;//剪枝步:删除非频繁候选
         }else add c to Ck;
         }
2) return Ck;
 procedure has_infrequent_subset(c,Lk-1)
1) For each(k-1)-subset s of c
     If s∉Lk-1 then return TRUE;
2) return FALSE;
```

算法中,除了 Apriori 算法固有的连接和剪枝操作外,为了同步清除低兴趣度的频繁项集,在主程序的第4步中加入兴趣度剪枝,降低了频繁项集产生的规模,使得最终得到的频繁项集内都项项正相关.

3 实验测试

在 Pumsb_star 数据集[10]上进行实验. 测试环境为 P4 1.5 GHz 的 CPU 和 1 G 内存，Windows 2003 操作系统. Apriori 算法采用文献[1]的方法，考虑到程序移植性，采用 Java(jdk1.5)实现.

实验测试 Apriori 算法在引入项项正相关频繁项剪枝前后的执行性能. 如图 1 所示，针对稠密数据库 Pumsb_star，加入兴趣度前，算法在低支持度阈值（*minsup* =45%）以下难以有效挖掘频繁项集；加入兴趣度剪枝后，由于对低兴趣度的项集及时剪枝，精简了候选项集的空间，可设置的最小支持度阈值可降低到 10% 以内，另外，算法的整体执行效率也得到提高.

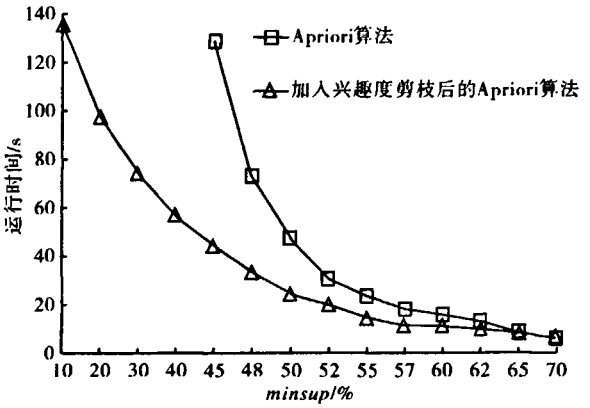

图 1　Pumsb_star 引入项项正相关兴趣度剪枝前后频繁项集挖掘效率对比

4 结语

本文提出一种具有反单调性质的项项正相关兴趣度，并将该量度引入到 Apriori 频繁项集挖掘算法中，通过 2 项集及 2 项以上集兴趣度剪枝，优化了项集空间，扩大了可挖掘支持度阈值范围，提高了挖掘效率. 在真实数据库上进行实验对比，验证了算法的可行性.

参考文献：

[1] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques[M]. Second Edition. Beijing: China Machine Press, 2006: 147 - 172.

[2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]//Proc of the ACM SIGMOD Conf on Mana of Data(SIGMOD'93), New York: ACM Press, 1993: 207 - 216.

[3] Omiecinski E. Alternative interesting measures for mining associations[J]. IEEE Trans Knowledge and Data Eng, 2003, 15: 57.

[4] 李英杰. 项约束频繁项集挖掘的新方法[J]. 计算机工程与应用, 2009, 45 (3): 161.

[5] 张玉芳，熊忠阳，彭燕，等. 基于兴趣度含正负项目的关联规则挖掘方法[J]. 电子科技大学学报, 2010, 39 (3): 407.

[6] 王艳，刘双红，李玲玲. 基于加权关联规则的选课推荐系统的构建[J]. 郑州轻工业学院学报：自然科学版, 2009, 24(5): 44.

[7] Geng L Q, Hamilton H J. Interestingness measures for data mining: A survey [J]. ACM Comp Surveys, 2006, 38 (3): 9.

[8] Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations [C]//Proc ACM SIGMOD Int Conf on Mana of Data, Tucson: ACM Press, 1997: 265 - 276.

[9] Huang Wenxue, Krneta Milorad, Lin Limin, et al. Association bundle—A new pattern for association analysis[C]// Sixth IEEE Int Conf on Data Mining Workshops(ICDMW'06) Washington: IEEE Computer Society, 2006: 601 - 605.

[10] FIMI. Frequent Itemset Mining Dataset Repository[EB/OL]. (2003 - 11 - 19)[2011 - 03 - 08]. http://fimi.cs.helsinki.fi/data/, 2003.