文章编号:1001 - 9081(2011)02 - 0435 - 03
doi:10.3724/SP.J.1087.2011.00435

Apriori 算法低频规则的有效性及实现

张春生,庄丽艳,李　艳
(内蒙古民族大学 计算机科学与技术学院,内蒙古 通辽 028043)
(zhangcs_817@sina.com)

摘　要:针对经典 Apriori 算法基于全局、高频两个条件的缺陷,指出事务数据库低频规则的有效性,并通过对 C4.5 决策树的规则构造,进一步证明事务数据库存在低频规则,在此基础上,给出了一种 Apriori 低频规则挖掘算法。该算法与经典的 Apriori 算法兼容,但不是对 Apriori 算法简单的扩展,而是从理论上打破了 Apriori 算法基于全局和高频两个条件。最后通过实例用 Apriori 低频规则挖掘算法和 C4.5 算法对实例数据库进行挖掘,证明两者的一致性和 Apriori 低频规则的有效性,同时也证明了 Apriori 低频规则挖掘算法的有效性。

关键词:Apriori 算法;低频规则;有效性;C4.5 算法;数据挖掘
中图分类号: TP311　　**文献标志码**:A

Effectiveness and implementation of low frequency rule based on Apriori algorithm

ZHANG Chun-sheng, ZHUANG Li-yan, LI Yan
(*College of Computer Science and Technology, Nei Mongol University for Nationalities, Tongliao Nei Mongol* 028043, *China*)

Abstract: Firstly, the defects of classical Apriori algorithm based on global view and high frequency were pointed out, and the effectiveness of low frequency rule of transaction database was presented. By constructing the rules of C4.5 decision tree, that the low frequency rule exists in transaction database also was proved. On the foundation of this, a mining algorithm based on low frequency rule of Apriori algorithm was given, which was compatible with classical Apriori algorithm. However, it was not a simple extension of Apriori algorithm, it had broken theoretically Apriori algorithm view based on global view and high frequency. Finally, case database was mined by mining algorithm based on low frequency rule of Apriori and C4.5 algorithms, and the consistency of two methods and the effectiveness of low frequency rule were proved. Moreover, the effectiveness of mining algorithm based on low frequency rule of Apriori algorithm was validated.

Key words: Apriori algorithm; low frequency rule; effectiveness; C4.5 algorithm; data mining

0　引言

传统的关联规则算法只有一个最小支持度,其中隐含了 3 个假设:一是针对事务数据库全局而言的;二是项的出现频率大致相同;三是项出现的频率比较高[1-3]。

然而,现实生活中一个事务数据库中存在的规则并不一定都是高频的,若一个事务数据库中某一特定群体的内部存在高频规则,那么这个规则也应该是整个事务数据库的规则;但传统的 Apriori 算法不能或很难发现这个规则的存在,这不能不说是 Apriori 算法的缺陷。

目前,关于 Apriori 算法的研究工作大部分是围绕如何提高 Apriori 算法的效率而开展的,针对关联规则的研究还比较少[4-7]。

在发现潜在规则方面,Liu 等人[13]提出了多支持度关联规则发现的概念,对不同的项给出不同的最小支持度,以此来解决项出现的频率不同的问题;但仍然存在没有真正摆脱单支持度的理论,没有对多支持度算法中各支持度定义选取规则,2-项集以后的项集生成中,仍然取各个项支持度的最小值,使得多支持度的概念退化[8-12]。

本文从 Apriori 算法低频规则出发,首先说明 Apriori 算法低频规则的有效性,并对 C4.5 决策树分类算法进行规则构造,通过构造的 C4.5 分类算法的规则证明 Apriori 算法低频规则的有效性,并且给出 Apriori 算法低频规则的挖掘算法,进一步通过实验数据进行证明。

1　Apriori 算法低频规则的有效性

现实生活中事务数据库都来自实际的社会活动,往往体现局部群体的特征,存在着一些潜在的规则。

对内蒙古民族大学某年计算机本科班 61 名学生的部分成绩进行挖掘,选取字段为数学、专业课、英语、体育、德育、去向,其中有 7 名学生考取了研究生。

表 1 表示了“毕业去向:研究生”与“专业课成绩:好”的记录分布。

表 1　“毕业去向:研究生”与“专业课成绩:好”的记录分布

范围	研究生		专业成绩优	
	人数	比例/%	人数	比例/%
全局	7	11	9	15
局部(研究生)	7	100	6	86

从表中可以看出,在“去向(研究生)”的局部群体中,存在潜在的规则:“去向(研究生)→专业成绩好”。因为在局部

收稿日期:2010 - 07 - 27;**修回日期**:2010 - 09 - 07。
基金项目:内蒙古人才基金资助项目(第 8 批);内蒙古教育科研项目(NJZY07140)。
作者简介:张春生(1965 -),男,河北乐亭人,教授,硕士,主要研究方向:数据库、数据挖掘、软件理论;　庄丽艳(1976 -),女,内蒙古通辽人,讲师,硕士,主要研究方向:人工智能;　李艳(1979 -),女,内蒙古通辽人,讲师,硕士,主要研究方向:数据库。

群体中数据项〈去向(研究生),专业成绩好〉的支持度是 min{86%,100%,}=86%。

同样的问题在全局情况下,数据项〈去向(研究生),专业成绩好〉的支持度是 min{11%,15%}=11%,经典 Apriori 算法只有当最小支持度小于等于 11% 时才能挖掘出来,这已经违背了 Apriori 算法的高频规则,而且计算量较大,产生的关联规则多而且很难进行分析。

2 C4.5 算法对 Apriori 算法低频规则的有效性

决策树算法和决策规则是解决实际应用中分类问题的数据挖掘方法,著名的有 C4.5 算法。

C4.5 算法用增益标准选择需要的属性,即熵的概念。

设事务集 T 分为 k 类,$T\rightarrow\{C_1,C_2,\cdots,C_k\}$,训练样本分成 n 个子集 $T\rightarrow\{T_1,T_2,\cdots,T_n\}$,$S$ 是任意的样本集,$freq(C_i,S)$ 代表 S 中属于 C_i 的样本数量,$|S|$ 为集合 S 中样本数量。

$$info(S)=-\sum_{i=1}^{n}(freq(C_i,S)/|S|)\cdot \mathrm{lb}(freq((C_i,S)/|S|)) \tag{1}$$

其中 $info(S)$ 为 S 的熵。

$$info_x(T)=-\sum_{i=1}^{n}((|T_i|/|T|)\cdot info(T_i)) \tag{2}$$

其中 $info_x(T)$ 为 T 的熵。

$$Gain(x)=info(T)-info_x(T) \tag{3}$$

其中:$Gain(x)$ 为增益准则,可选择 $Gain(x)$ 最大的属性作分类属性。

定义 1 设 T 有 m 个属性,属性为 $x_i(1\leqslant i\leqslant m)$,$\rho_i$ 是属性 i 的分支条件,则从决策树的根节点出发到叶子节点形成一个规则,即 $(x_1\rho_1,x_2\rho_2,\cdots,x_h\rho_h)\rightarrow C_i$,其中 h 为根到叶子节点的分支数,C_i 为第 i 个分类。

定理 1 对于包含分类 C_i 的事务集 T_i 中的属性值 $x_i=count$ 出现的频率较高($count$ 为常量),则一定存在规则 $x_i\rho_i\rightarrow C_i$。

证明 对指定的事务集 T,增益准则 $Gain(x)=info(T)-info_x(T)$ 中的 $info(T)$ 是一个固定值。

由式(3)知 $info_x(T)$ 的公式,其中 $\sum_{i=1}^{n}(|T_i|/|T|)$ 对于 x_i 属性来说,T_i 也是固定的,因而 $\sum_{i=1}^{n}(|T_i|/|T|)$ 的值也相对固定。下面考虑 $info(T_i)$,由式(2)知:

$$info(T_i)=-\sum_{i=1}^{n}(freq(C_i,T_i)/|T_i|)\cdot \mathrm{lb}(freq((C_i,T_i)/|T_i|))$$

根据条件,包含分类 C_i 的事务集 T_i 中的属性值 $x_i=count$ 出现的频率较高,则 $freq((C_i,T_i)/|T_i|)\rightarrow 1$,所以 $\mathrm{lb}(freq((C_i,T_i)/|T_i|)\rightarrow 0$,这必然导致 $info(T_i)$ 的值减小,从而使 $info_x(T)$ 的值减小,由式(4)知 $Gain(x)=info(T)-info_x(T)$ 的值增大,亦即 x_i 属性的增益增大,x_i 作为决策属性,故 $x_i\rho_i\rightarrow C_i$ 存在。

推论 1 若决策树的根节点到 C_i 的分支中规则 $x_1\rho_1\rightarrow C_i,x_2\rho_2\rightarrow C_i,\cdots,x_m\rho_m\rightarrow C_i$ 成立,则规则 $(x_1\rho_1,x_2\rho_2,\cdots,x_m\rho_m)\rightarrow C_i$ 一定成立。

证明 若 $x_1\rho_1\rightarrow C_i,x_2\rho_2\rightarrow C_i,\cdots,x_m\rho_m\rightarrow C_i$ 成立,根据定理 1 知,x_i 是决策属性,且出现在从决策树的根节点出发到叶子节点 C_i 的分支中,由定义 1,$(x_1\rho_1,x_2\rho_2,\cdots,x_m\rho_m)\rightarrow C_i$ 一定成立。

推论 2 包含分类 C_i 的事务集 T_i 中的属性有 $x_1,x_2,\cdots,x_m$,若 $x_i=count_i(1\leqslant i\leqslant m)$ 出现的频率较高($count_i$ 为常量),则一定存在规则 $(x_1\rho_1,x_2\rho_2,\cdots,x_m\rho_m)\rightarrow C_i$,这个规则是整个事务集 T 的规则,与 T_i 出现的频率无关。

证明 包含分类 C_i 的事务集 T_i 中的属性有 $x_1,x_2,\cdots,x_m$,若 $x_i=count_i(1\leqslant i\leqslant m)$ 出现的频率较高($count_i$ 为常量),由定理 1 知,$x_i\rho_i\rightarrow C_i$ 一定成立,进而由推论 1 知,$(x_1\rho_1,x_2\rho_2,\cdots,x_m\rho_m)\rightarrow C_i$ 一定成立,对决策树来说,这是全局的,而非局部的,所以这个规则是整个事务集 T 的规则,与 T_i 出现的频率无关。

综上所述,若事务集中存在局部规则,则决策树算法可以发现。

3 Apriori 算法低频挖掘算法

以群体的比例范围为基础,确定最小支持度的范围,以此范围作为产生规则的标准,可有效过滤不相关规则,降低规则数量,提高群体规则的准确性。

设某一事务数据库中,某一群体所占的比例为[$p\%,q\%$](这只是估计,群体事先不一定明确),定义频繁项集的最小支持度为 $c\%$,则低频挖掘的支持度范围为[$p\%\times c\%,q\%\times c\%$],由于 k-项集的频率都在[$p\%\times c\%,q\%\times c\%$]范围之内,所以最小信任度应与经典的 Apriori 算法的选取原则相同。

$$c(A\rightarrow B)=\frac{s(A\cup B)}{s(A)}=\frac{s(A\cup B)\times c\%}{s(A)\times c\%} \tag{4}$$

本文算法兼容经典的 Apriori 算法,当群体比例为[100%,100%]时,算法退化成经典的 Apriori 算法。

4 Apriori 低频规则挖掘算法描述

设最小支持度的范围为[$minsup\_l,minsup\_h$],最小支持数的范围为[$minsup\_con\_l,minsup\_con\_h$],最小信任度为 $minconf$[1]。

1) 扫描数据库,形成频繁项集。

```
//输入:数据集 D,最小支持数范围[ minsup_con_l,minsup_con_h ]
//输出:频繁项目集 L
for all 1-itemsets do begin
  if sup ( 1-item ) > = minsup_con_l and sup(1-item) < =
  minsup_con_h
   L1 = L1 ∪{1-item};
end
L1 = {large 1-itemsets};
//得到所有[ minsup_con_l,minsup_con_h] 范围内的 1-项集
for( k = 2;L_{k-1} ≠∅; k ++) do begin
  C_k = aprior-gen( L_{k-1});          // C_k是 k 个元素的候选集
  for all transactions t ∈ D do begin
    C_t = subset( C_k,t);          // C_t 是所有 t 包含的候选集元素
    for all candidates c ∈ C_t do c. count ++;
  end
  L_k = {c ∈ C_k | c. count > = minsup_con_l and c. count < =
  minsup_con_h};
end
L = ∪ L_k;
// aprior-gen()函数:由 (k - 1)- 项集产生 k- 项集
// has_infrequest_subset (c,L_k - 1) 函数:
// 输入 k- 候选项目集 c,k - 1 频繁项集 L_{k-1},判断 c 是否从
// 候选项集中删除
```

2) 关联规则的生成。

```
// 对于每一个频繁项集 l,生成其所有的非空子集
// 对于 l 的每一个非空子集 x,计算 confidence (x),
// 如果 confidence (x) ≥ minconfidence,那么 x → (l - x) 成立。
rule-generate (L,minconf)
```

for each frequent itemset l_k in L
genrules (I_k,I_k)
// genrules( )递归产生一个频繁项集的关联规则

5　实验证明

对第 1 章中的事务集,应用 Apriori 低频规则挖掘算法和经典的 C4.5 算法进行数据挖掘。

1) Apriori 低频规则挖掘。

估计考研学生的群体人数为 5 ~ 10,经典支持度 80%,则群体支持度区间[(5/63) ×80%,(10/63) ×80%] ≈[6.3%,12.7%] ≈[5%,15%],选取置信度为 80%。

挖掘结果如图 1 所示,产生频繁项 7 个,最大频繁项 5 个,关联规则一个:"去向(研究生)→专业成绩好"。

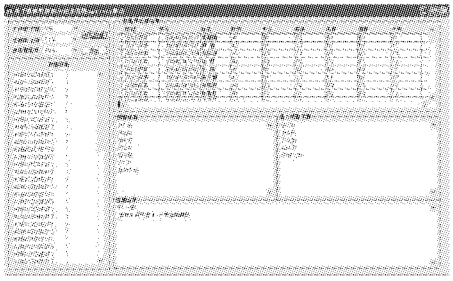

图 1　应用 Apriori 低频挖掘算法的结果

2) C4.5 算法挖掘。

golf. names 文件内容(命名文件)

A, B, C, D.
　　//输出(毕业去向): A 研究生; B: 公务员; C: 就业; D: 降级
math: continuous.　　//数学: 连续值
speciality: continuous.　　//专业: 连续值
english: continuous.　　//英语: 连续值
phedu: continuous.　　//体育: 连续值
moral: continuous.　　//德育: 连续值

golf. data 文件内容(事务数据文件 61 个样本)

71, 64, 65, 9, 25, C
79, 79, 77, 10, 25, C
⋮
74, 56, 52, 9, 25, D

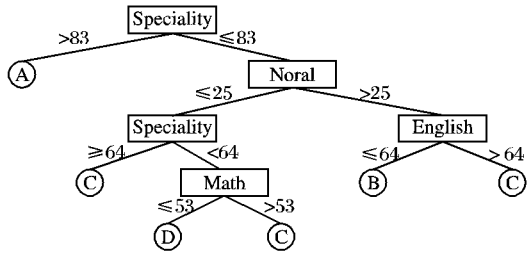

图 2　C4.5 决策树输出

从 C4.5 决策树上可以看出,存在规则(Spaciality > 83→A),在 1)的属性定义中,"专业课 > =85"为"专业课成绩好",2)中定义"输出(毕业去向): A 研究生",所以此决策树输出存在规则:"专业成绩好→去向(研究生)",这与 1)的挖掘结果完全吻合。

6　结语

通过实例对比可以看出,本文提出的 Apriori 低频挖掘算法克服了经典 Apriori 算法的缺陷,可按群体特征对数据库进行数据挖掘,发现经典 Apriori 算法不能发现的或很难发现的强关联规则,发现了事务数据库中潜在的规则。

本文不是简单地对经典 Apriori 算法进行扩展或改进,因为它从理论上破坏了 Apriori 算法全局、高频两个性质,从算法实现过程来看,算法的效率与经典 Apriori 算法相同。

参考文献:

[1] 毛国君, 段立娟, 王实, 等. 数据挖掘原理与算法[M]. 2 版. 北京: 清华大学出版社, 2007.

[2] KANTARDZIC M. 数据挖掘——概念、模型、方法和算法[M]. 闪四清, 陈茵, 程雁, 译. 北京: 清华大学出版社, 2003.

[3] DUNHAM M H. 数据挖掘教程[M]. 郭崇慧, 田凤占, 靳晓明, 译. 北京: 清华大学出版社, 2005.

[4] HAN JIAWEI, PEI JIAN, YIN YIWEN. Mining frequent patterns without candidate generation [C]// Proceedings of the 2000 ACM SIGMOD Internal Conference on Management of Data. New York: ACM, 2000: 1 - 12.

[5] BERZAL F, CUBERO J C, MARIN N, *et al*. TBAR: An efficient method for association rule mining in relational databases[J]. Data & Knowledge Engineering, 2001, 37(1): 47 - 64.

[6] 皮德常, 秦小麟, 王宁生. 基于动态剪枝的关联规则挖掘算法[J]. 小型微型计算机系统, 2004, 25 (10): 1850 - 1852.

[7] 刘以安, 羊斌. 关联规则挖掘中对 Apriori 算法的一种改进研究[J]. 计算机应用, 2007, 27(2): 418 - 420.

[8] 宫雨. 分组多支持度关联规则研究[J]. 计算机工程与设计, 2007, 28(5): 1205 - 1207.

[9] 史原, 鲁汉榕, 罗菁, 等. 基于规模约简和多支持度的关联规则挖掘[J]. 计算机工程与设计, 2006, 27(21): 4105 - 4107.

[10] 阮璐, 肖冬荣, 周杰, 等. 利用组合支持度进行关联规则挖掘[J]. 微计算机信息, 2008(9): 239 - 240.

[11] 楼晓鸿, 丁宝康. 一种多支持度的关联规则采集算法[J]. 计算机工程, 2001, 27(6): 102 - 103.

[12] 李刚, 董祥军. 多支持度关联规则的研究[J]. 广西轻工业, 2007, 23(5): 60 - 61.

[13] LIU BING, HSU W, MA Y. Mining association rules with multiple minimum supports[C]// Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 1999: 143 - 150.

(上接第 431 页)

[11] ZHANG YUFANG, XIONG ZHONGYANG, MAO JIALI, *et al*. The study of parallel *K* - means algorithm [C] // Proceedings of the 6th World Congress on Intelligent Control and Automation. Washington, DC: IEEE Computer Society, 2006: 5868 - 5871.

[12] 韩晓红, 胡彧. K-means聚类算法的研究[J]. 太原理工大学学报, 2009, 40(3): 236 - 239.

[13] SHI Y H, EBERHART R C. Parameter selection in particle swarm optimization[J]. Lecture Notes in Computer Science, 1998(1447): 591 - 600.

[14] 陈国良. 并行计算——结构·算法·编程[M]. 北京: 高等教育出版社, 2001.

[15] 都志辉. 高性能之并行编程技术——MPI 并行程序设计[M]. 北京: 清华大学出版社, 2001.

[16] MPI 文档[EB/OL]. [2010 - 05 - 10]. http://www. mpi-forum. org/docs/docs/html.

[17] MPICH 文档[EB/OL]. [2010 - 05 - 10]. http://www. mcs. anl. gov/research/projects/mpich2/.

[18] 王华秋, 廖晓峰. 微粒群并行聚类在客户细分中的应用[J]. 计算机应用研究, 2008, 25(10): 2987 - 2990.

[19] UCI 数据库[EB/OL]. [2010 - 06 - 20]. http://archive. ics. uci. edu/ml/machine-learning-databases/.