第19卷 第3期 郑州工业大学学报 Vol.19 No.3
1998年 9 月 Journal of Zhengzhou University of Technology Sep. 1998

基于数值属性的关联规则的挖掘

涂星原
(郑州工业大学计自系)

摘 要 研究了基于数值属性(Quantitative Attribute)的关联规则的挖掘问题,提出了挖掘算法Q-Basic和Q-AR,试验表明Q-AR是有效的。

关键词 数据挖掘;关联规则;数值属性
中图分类号 TP311.13

关联规则(Associstion Rules)是数据挖掘(Data Mining)研究的重要内容。其应用包括商场购物分析、广告邮寄分析、网络故障分析等。

文献[1～5]对关联规则的挖掘作了有意义的研究。R·Agrawal等提出了Apriori算法[2]和挖掘多层次关联规则的Culmulate,Stratify等算法[3],J·S·Park等提出了DHP算法[5],J·Han等提出了面向属性归纳(Attribute-Oriented Induction)的关联规则挖掘算法ML-T2L1等[6]。这些算法均将事务中项目的数量作为0/1值来看待,挖掘布尔关联规则BAR(Boolean Association Rules)。许多应用中,布尔关联规则不能清楚地表达数据间的关联关系。例如,股票综合指数的上升与价位上升的股票个数的关系。

本文研究了基于数值属性(Quantitative Attribute)的关联规则的挖掘问题和挖掘算法,给出了挖掘基于数值属性的关联规则的算法:基本算法Q-Basic和改进算法Q-AR,并对关联规则的有趣性(Interestingness)作了探讨。实验表明算法Q-AR是有效的。

1 基于数值属性的关联规则的挖掘问题描述

事务表示为$\{tid,<(A_1,N_1),(A_2,N_2),\cdots,(A_j,N_j)>\}$,$tid$为事务标识号,全局唯一,$A_i$是项目($A_i\in I$),$N_i$是项目$A_i$的数量值,$i=1,2,\cdots,j$,$I$为项目集合。

定义1 模式P定义为$(A_1,N_1)\cap(A_2,N_2)\cdots\cap(A_i,N_i)$,$A_k\in I$,$N_k>0(k=1,2,,i)$。事务$t$:$\{tid,<(B_1,M_1),(B_2,M_2),\cdots,(B_j,M_j)>\}$,若模式$P$中任意$(A_k,N_k)$,在$t$中有$(B_S,M_S)(S=1,2,\cdots,j)$,使$B_S=A_k$,$M_S\geqslant N_k$,称$t$包含$P$,或$t$支持$P$。

定义2 模式$P=(A_1,N_1)\cap(A_2,N_2)\cdots\cap(A_i,N_i)$和$Q=(A_1,M_1)\cap(A_2,M_2)\cdots\cap(A_k,M_k)$,$i<k$,对于任意$j(j=1,2,,i)$:$N_j<M_j$,称为$P$小于$Q$,记为$P<Q$。

引理1 若支持模式$Q=(A_1,N_1)\cap(A_2,N_2)\cdots\cap(A_i,N_i)$,模式$P<Q$,则$t$支持$P$。

模式P在事务集合D中的支持率(support)$\sigma(P/D)=\dfrac{D\text{中包含模式}P\text{的事务的个数}}{D\text{中事务总个数}}$。

关联规则$A\Rightarrow B$的可信度(confidence)$\Psi(A\Rightarrow B/D)=\dfrac{\sigma(AB/D)}{\sigma(A/D)}$,其中$A$,$B$均为模式[1,3],为挖掘有效的关联规则,定义最小支持率σ_{min}与最小可信度Ψ_{min}。支持率σ_{min}的模

收稿日期:1997－09－04;修改稿返回日期:1998－02－15
第一作者 女 1964年7月生 学士学位 讲师

式称为频繁模式，频繁模式集合记为 L。包含 i 个项目的频繁模式的集合，称为 i 频繁模式集合，记为 L_i[1]。

定义 3　若事务 t 支持模式 $P=(A_1, N_1)\bigcap(A_2, N_2)\cdots\bigcap(A_i, N_i)$，$M_i$ 为 N_i 与事务 t 中 A_i 项目的数量值的小者，称模式 $Q=(A_1, M_1)\bigcap(A_2, M_2)\cdots\bigcap(A_i, M_i)$ 为模式 P 在事务数据中的最大包含模式，记为 in-max(P, t)。若 t 不支持模式 P，模式 P 在事务中的最大包含模式 in-max(P, t) 为空。

定义 4　模式集 S 的简化集 $design(S)\{x \mid x\in S$，不存在 $y\in S, y<x\}$。模式集 S 的增长集 gen$(S)=\{y \mid y\notin S$，存在 $x\in S, x<y\}$。

定义 5　i 简化频繁模式集合 S_i 定义为：$design(L_i)$。

挖掘关联规则即是找出满足最小支持率 σ_{min} 与最小可信度 Ψ_{min} 的形如 $A\Rightarrow B$ 的规则。挖掘过程分为两步：首先求出频繁模式集合 $L=U_kL_k$，然后由 L 构造关联规则集合。R·Agwral 给出了由 L 求解关联规则的方法[2]。本文着重讨论频繁模式的求解。

2　基于数值属性的关联规则的挖掘

下面给出挖掘基于数值属性的关联规则的两个算法：Q-Basic 和 Q-AR。

2.1　基本算法 Q-Basic

算法 Apriori 根据模式 P 中的项目是否出现在事务 t 中，判断事务 t 是否支持模式 P。算法 Q-Basic 在模式的表示和事务对模式的支持的判别上，对算法 Apriori 作如下修改：模式由包含的项目和项目的数量值组成，根据定义 1 判断事务 t 是否支持模式 P，即检查模式 P 中的项目是否在 t 中出现和 P 中项目的数量值是否小于 t 中相同项目的数量值。

2.2　改进算法 Q-AR

以下假设模式集合中的模式按项目的字典序排列，相同项目的模式按对应项目的数量值按由小到大排列。模式 P 的支持事务的个数记为 $P\cdot support$。

算法 Q-AR(挖掘具有数值属性的关联规则集合)：

input：事务数据库 D，最小支持率 σ_{min}

output：关联规则集合 AR

method：

```
L1=1 频繁模式集合
S1=degen(L1)
k=2
L=S1
while Sk-1<>NIL Do {
S-Ck=simple-apriori-gen(Sk-1)
for every t∈D do
for every Q∈S-Ck do{
P=in-max(t,Q)          //P=Q 与 t 的最大包含模式
if P<>NIL then
if P∈S-Ck then P·support=P·support+1
else {S-Ck=S-Ck+{P};P·support=1}}
```

L_k=itemset-gen(S-C_k,σ min)
L=L∪L_k
S_k=degent(L_k)
k=k+1}

关联规则集合 AR={$A \Rightarrow B \mid A \cap B \in L, \frac{\sigma(A \cap B / D)}{\sigma(A / D)} > \Psi_{\min}$}

算法 itemset-gen(由候选 k 简化频繁模式集合求解 k 频繁模式集合 L_k)
input:候选 k 简化频繁模式集合 $S\text{-}C_k$
output:k 频繁模式集合 L_k
method:
L_k={}
while S-C_k<>NIL do{
P=S-C_k 中字典序的第一个模式
P-SET=S-C_k 中与 P 有相同的项目的模式集合
S-C_k=S-C_k−P-SET
P-GEN=gen(P-SET)
for every Q∈P-GEN do
Q·support=0
SET=P-SET+P-GEN
while SET<>NIL do{
Q=SET 中字典序的第一模式
Q·support=Q·support+$\sum_{Q<P}$P·support
SET=SET−Q
if Q·Support y/|D|>$\sigma_{\min}$ then L_k=L_k+{Q}}}

算法 Simple−apriori−gen 用于求解候选 k 简化频繁模式集合。输入:$k-1$ 简化频繁模式集合 S_{k-1};输出 k 简化频繁模式集合 $S-C_k$。算法 Q-AR 说明:由 L_{k-1} L_k 时,利用 simple-apriori-gen 由 $k-1$ 简化频繁模式集合 S_{k-1}构造候选频繁模式集合 $S\text{-}C_k$;然后访问事务数据 D,计算 $S\text{-}C_k$ 中的候选频繁模式的支持;利用算法 itemset-gen 由候选 $k-1$ 简化频繁模式集合 $C\text{-}S_k$ 求出 L_k;最后求出 $L=U_kL_k$。

算法 itemset-gen 将 C_k 中的模式划分为包含相同项目的子集,对每个子集 P-SET,求出其增长集 P-GEN,然后根据 P-SET 中模式的支持求出 P-GEN 中模式的支持,经过 min 的过滤,最后求出 k 频繁模式集合 L_k。

Q-AR 算法由 k 简化频繁模式集合 S_k 产生候选频繁模式集合的简化集合 $S\text{-}C_k$,节约了存储空间,而且事务 t 对模式的支持的判断个数减少,提高了 L_k 的求解速度。

3 算法分析

在事务数据个数|D|=14000,项目个数|I|=15,平均项目的数量值=8,事务平均包含的项目数=10 的试验数据下,在 Pentium−133 上的试验结果如图 1 和图 2。

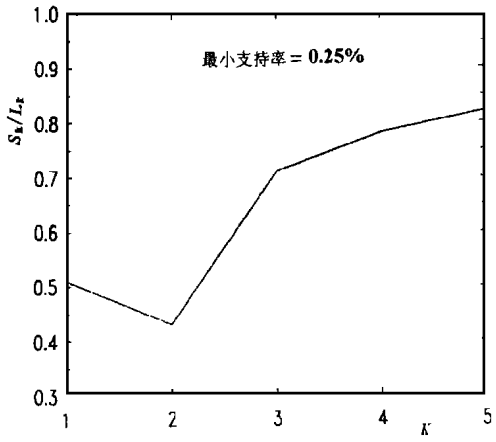

图 1　循环次数 k 与 S_k/L_k 关系

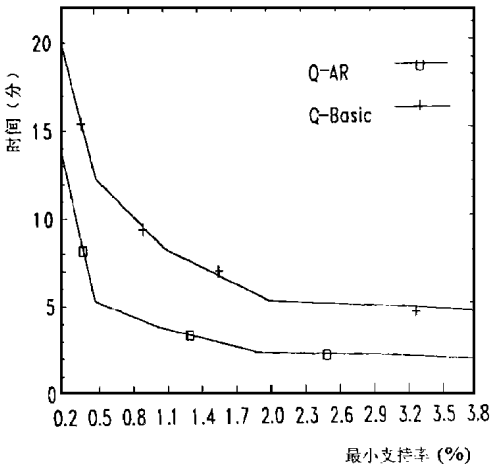

图 2　最小支持率与执行时间关系

4　结论

本文提出了基于数值属性的关联规则的挖掘问题，给出了挖掘算法 Q-Basic 和 Q-AR，并进行了比较，实验表明 Q-AR 是有效的，比 Q-Basic 快 1.5～2 倍，占用存储空间少 1/3。基于数值属性的关联规则是布尔关联规则的扩展，更精确地反映了数据间的关联关系。

参考文献

1　Agrawal, R., Imielinski, T., Iyer, B. et al. Mining association rulesbetween sets of items in larg databases. Proc. of ACM SIGMOD Conf. on managenment of data, 1993: 207～216

2　Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In the Proc. of VLDB－1994, Chile, Sep. 1994: 487～499

3　Agrawal, R., Srikant, R. mining generalized association rules. In the Proc. of VLDB－1995, Swizerland, 1995: 407～419

4　Fayyad, U., Piatesky－shapiro, G., Smyth, P. From data mining to knowledge discovery: An overview. Advances in knowledge discovery and data mining, MIT Press, 1996: 1～34

5　Park, M., Chen, M., Yu, P. An effective hash based algorithm for mining association rules. in the Proc. of ACM SIGMOD 1995: 175～186

6　Han, J., Fu, Y. Discovery of multiple－level association rules from large databases. in Proc. of VLDB－1995, Swizerland, 1995: 420～431

Efficient Mining of Association Rules Based on Quantitative Attributes

Tu Xingyuan

(*Zhengzhou University of Technology*)

Abstract　The mining of the association rules based on quantitative attributes is discussed, and two algorithms, Q-Basic and Q-AR, are proposed in the paper. The experiments show that the algorithm Q-AR is effective.

Keywords　data mining; association rules; quantitative attribute