文章编号：1001-0920(2009)10-1504-05

基于 TD-FP-growth 的模糊关联规则挖掘算法

霍纬纲[1,2]，邵秀丽[1]

(1. 南开大学 信息技术科学学院，天津 300071；2. 中国民航大学 计算机科学与技术学院，天津 300300)

摘 要：提出一种基于 TD-FP-growth 的模糊关联规则挖掘算法. 首先，使用 3 种 t-模算子以及由其产生的蕴涵算子计算模糊频繁项的支持度和规则的蕴涵度，产生的关联规则能表示模糊项间的确定性和渐近性逻辑语义；然后，以事务的惟一标识为键值，散列存储每个事务相对 FP-tree 中每个结点所表示模糊项的隶属度，使 TD-FP-growth 适用于模糊频繁项的挖掘，并分析了算法的时间和空间复杂度；最后，实验结果表明该算法比基于 apriori 的模糊频繁项挖掘算法在时间方面更加有效.

关键词：模糊关联规则；模糊蕴涵；TD-FP-growth；t-模算子

中图分类号：TP18 **文献标识码**：A

Algorithm for mining fuzzy association rule based on TD-FP-growth

HUO Wei-gang[1,2], *SHAO Xiu-li*[1]

(1. College of Information Technical Science, Nankai University, Tianjin 300071, China; 2. College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China. Correspondent: HUO Wei-gang, E-mail: huowg2003@126.com)

Abstract: An algorithm based on TD-FP-growth is proposed for mining fuzzy association rule, which uses three kinds of t-norm operator to calculate the support degree of fuzzy frequent items, and adopts corresponding implication operator to measure implication degree of fuzzy association rule. The association rule mined by the algorithm can express the logic semantic of graduality and certainty between fuzzy items. Each transaction's membership degree versus fuzzy item denoted by FP-tree's node is stored by hash technology, and each transaction's identifier is regarded as key value, which adapts TD-FP-growth to mine fuzzy frequent items. The time and space complexity of the algorithm are analyzed. The experimental results show that the algorithm is more effective than the fuzzy frequent item mining algorithm based on apriori in term of time.

Key words: Fuzzy association rule; Fuzzy implication; TD-FP-growth; t-norm operator

1 引 言

为了解决在数量型属性的关联规则挖掘过程中边界划分过硬的问题，研究人员将模糊集合的概念引入关联规则挖掘，提出了很多模糊关联规则挖掘算法[1-5]. 但模糊集的引入产生了新的问题，模糊关联规则的逻辑语义不同于传统的关联规则：传统的关联规则表示的是不同数据项在同一事务中同时出现的逻辑语义，而模糊关联规则除了表示模糊数据项同时出现的语义外，还可表示模糊数据项之间的确定性和渐近性的逻辑语义[6,7]，模糊数据项间逻辑语义是由具体模糊蕴涵算子确定的. 文献[8]提出了基于 Apriori 的模糊蕴涵关联规则挖掘算法及基于蕴涵度的模糊关联规则度量方法；[9]提出基于 TD-FP-growth 的模糊关联规则算法用于生物基因数据的分析，但其生成 FP-tree 的结点间没有关联，造成 FP-tree 的头表过于复杂. 以上方法都没从语义的角度考虑挖掘出的规则. [7]提出了具有确定性和渐近性语义的模糊规则，并提出了一种能够学习模糊数据项之间逻辑语义的方法，但该方法没有模糊频繁项的产生步骤，无法与关联规则挖掘方法相融合；[10]提出了基于归纳逻辑编程和模糊蕴涵算子的模糊规则算法，该算法产生规则的后件只能包含一个模糊项.

本文对文献[9,11]提出的 TD-FP-growth 算法进行扩充，使之适用于具有逻辑语义的模糊关联规则挖掘，用散列技术存储每个事务相对 FP-tree 结

收稿日期：2008-12-17；**修回日期**：2009-03-09.

基金项目：国家自然科学基金委与民航局联合项目(60672174,60776806).

作者简介：霍纬纲(1978—)，男，山西洪洞人，博士生，从事数据挖掘的研究；邵秀丽(1963—)，女，江苏南通人，教授，博士生导师，从事并行与分布式系统、CSCW 等研究.

点表示的模糊项的隶属度，由 3 种 t-模算子计算模糊频繁项的支持度以产生具有不同逻辑语义的关联规则. 实验结果表明了模糊关联规则语义的必要性，以及本文提出的算法比目前模糊关联规则挖掘中常用的 Apriori 算法在时间方面的有效性.

2　模糊关联规则的相关定义

设 $I=\{I_1,I_2,\cdots,I_m\}$ 是数据库 $D=\{t_1,t_2,\cdots,t_n\}$ 中由 m 个不同的数量属性组成的集合，其中 $t_i(1\leqslant i\leqslant n)$ 为数据库 D 的第 i 个事务. 在 $I_j(1\leqslant j\leqslant m)$ 的取值范围 Δ_j 上可定义模糊集 $\{I_{j1},I_{j2},\cdots,I_{jq_j}\}$. 将 I_j 扩展到与 Δ_j 相关联的模糊集上，可得到扩展项目集 $I_f=\{I_{11},I_{12}\cdots,I_{1q_1},\cdots,I_{m1},\cdots,I_{mq_m}\}$. 由此，可将 D 转换为所有属性取值在区间[0,1]上的模糊数据库 D_f. 模糊关联规则是形如 $X\Rightarrow Y$ 的蕴涵式. 其中 $X\subset I_f,Y\subset I_f,X\cap Y=\varnothing$，且 $X\cup Y$ 不同时包含从同一个属性扩展而来的任何两个项目.

模糊关联规则 $X\Rightarrow Y$ 在数据库 D_f 的支持度定义为

$$\mathrm{Dsupp}(X\Rightarrow Y)=\frac{\sum_{i=1}^{n}T_i(x_1,x_2,\cdots,x_n)}{n}. \quad (1)$$

其中：x_i 为第 i 个事务对模糊项 $X\cup Y$ 的隶属度值，n 为 D_f 中事务的个数，T 为 t- 模算子.

$\mathrm{Dimp}_i(X\Rightarrow Y)=I(\mathrm{Dsupp}_i(X),\mathrm{Dsupp}_i(Y))$ 为在 D_f 的第 i 个事务对 $X\Rightarrow Y$ 的蕴涵度，其中 I 为模糊蕴涵算子[7]. 模糊关联规则 $X\Rightarrow Y$ 的蕴涵度[8] 定义为

$$\mathrm{Dimp}(X\Rightarrow Y)=\frac{\sum_{i=1}^{n}\mathrm{Dimp}_i(X\Rightarrow Y)}{n}. \quad (2)$$

定义大于等于最小支持度 $D_{\min\_\mathrm{supp}}$ 和最小蕴涵度 $D_{\min\_\mathrm{imp}}$ 的规则 $X\Rightarrow Y$ 为强模糊关联规则.

3　模糊关联规则的语义

模糊关联规则的语义由模糊蕴涵算子确定，设有模糊项 X,Y. 模糊关联规则 $X\Rightarrow Y$，根据文献[10]可表示 3 种逻辑语义：同时出现，表示隶属模糊项 X,Y 的事务在数据库中出现的频度满足用户的要求；渐近性，表示模糊项 X,Y 的关联程度，其值可由 Lukasiewicz 蕴涵算子 $I(x,y)=\min\{1,1-x+y\}$ 计算[7]，其中 x,y 为数据库中某一事务相对模糊项 X 和 Y 的隶属度；确定性，表示模糊项 X 对 Y 的确定性，两者之间的确定性程度可由 Kleene-Dienes 蕴涵算子 $I(x,y)=\max\{1-x,y\}$ 或 Reichenbach 蕴涵算子 $I(x,y)=1-x+xy$ 确定[7].

文献[12]指出模糊关联规则 $X\Rightarrow Y$ 将数据库 D_f 划分为支持模糊关联规则的事务集合 S_+，不支持规则的事务集合 S_-，以及规则无关的事务集合 $S_\pm$，D_f 中的事务 t 对每一集合的隶属度按下式计算：

$$\begin{aligned}S_+(t)&=X(t)\otimes Y(t),\\ S_-(t)&=1-(X(t)\mapsto Y(t)),\\ S_\pm(t)&=1-X(t).\end{aligned} \quad (3)$$

其中：$\otimes$ 表示 t- 模算子；$\mapsto$ 表示模糊蕴涵算子；$X(t),Y(t)$ 分别表示 t 对模糊项 X,Y 的隶属度. 并且上式满足

$$S_+(t)+S_-(t)+S_\pm(t)=1. \quad (4)$$

符合式(4) 的 t- 模算子与其所对应的模糊蕴涵算子如表 1[12] 所示.

表 1　t- 模算子与其对应的模糊蕴涵算子

T 三角模算子	模糊蕴涵算子
$T(x,y)=\min(x,y)$	$I(x,y)=\min(1,1-x+y)$
$T(x,y)=x\times y$	$I(x,y)=1-x(1-y)$
$T(x,y)=\max(x+y-1,0)$	$I(x,y)=\max(1-x,y)$

由模糊关联规则的语义与表 1 可知，在模糊频繁项的挖掘过程中，如采用 $\min(\alpha,\beta)$ 算子及其对应的蕴涵算子计算模糊频繁项的支持度和规则的蕴涵度，可产生具有渐近性逻辑语义的模糊关联规则. 采用 $\alpha\beta$ 或 $\max(\alpha+\beta-1,0)$ 算子及其蕴涵算子，可产生具有确定性语义的模糊关联规则.

4　基于 TD-FP-growth 的模糊关联规则挖掘算法

文献[11]提出的 TD-FP-growth 算法对 FP-tree 树头表中的项采取自顶向下的处理方式，在挖掘过程中不需生成条件模式库和条件 FP-tree 树. 本文对该算法进行扩充，与文献[9]的区别为：由表 1 中的 3 种 t- 模算子及蕴涵算子计算模糊频繁项的支持度及其对应关联规则的蕴涵度，产生的规则具有不同的逻辑语义，递归计算模糊频繁项支持度的正确性由 t- 模算子满足的交换律和结合律来保证；以数据库中事务的惟一标识为键值，散列存储每个事务相对 FP-tree 中每个结点的模糊项的隶属度；保持 FP-tree 头表结构不变，树内结点除模糊项的支持度计数改为指向该结点对应的 hash 区指针外，其他结构不变. 算法分为以下 3 个步骤：

1) 构造模糊 FP-tree. 通过扫描一次数据库得到 1- 模糊频繁项集，并将该项集按支持度计数的递减次序排序构成列表 L. 再次扫描数据库，数据库中每个事务的模糊项按列表 L 中的顺序插入模糊 FP-tree. 同时，构建每个结点对应的 hash 区，hash

表中的每个表项包括事务标识和隶属度值. 采用 hash 函数:$h(x)=\lfloor m\times((x\times 0.618)\bmod 1)\rfloor$,其中 m 为 hash 空间大小,取 2 的 n 次幂,n 的大小取决于数据集大小,采用链接法解决 hash 冲突.

2) 模糊频繁项集的产生. 对模糊 FP-tree 头表中的每个模糊项 I 按自顶向下的顺序扫描处理,每次通过遍历模糊 FP-tree 和递归构建该模糊项 I 对应的子头表,都产生所有以模糊项 I 结尾的模糊频繁项. 为此,模糊 FP-tree 中的每个结点还需存储遍历路径上的由遍历开始结点与当前结点构成的模糊项所共享事务的隶属度值的临时 hash 区.

3) 通过模糊频繁项集产生满足最小蕴涵度的强模糊关联规则,规则的蕴涵度由文献[7]提出的 $\mathrm{Dimp}(X\Rightarrow Y)=1+\mathrm{Dsupp}(X\cup Y)-\mathrm{Dsupp}(X)$ 进行计算,以避免为求蕴涵度而扫描数据库. 具体算法描述如下:

输入:事务数据库 D_f,最小支持度 min_sup,最小蕴涵度 min_imp;

输出:模糊频繁项集 L,强模糊关联规则集 R.

Step1 Proc CreateFFP-tree()

{ 扫描数据库生成支持度计数降序排序的 1- 模糊频繁项集列表 L;

For each item I in L

将 item I 及其支持度计数降序插入 FP-tree 的头表 H;

For each transaction T in D_f

{ currentnode = root;

//root 为 FP-tree 的根结点

for each item I in T

//T 中的 item I 要按列表 L 的次序排序

{ if(I 属于列表 L 并且事务 T 对 I 的隶属度值大于 0) then

{ if currentnode 没有标记为 I 的子结点 cNode then

{ 创建 currentnode 的子结点 childnode,将事务 T 对 I 的隶属度通过事务标识 hash 存储在子结点 childnode 的 hash 区 membershipD 中;

将 childnode 插入与头表 H 中的 item I 的链表尾部;

currentnode = childnode;

}

Else

将事务 T 对 I 的隶属度值 hash 存储在 currentnode 的子结点 cnode 的 hash 区 membershipD 和 tempmembershipD 中;

currentnode = currentnode 的子结点 cnode;

}

}}.

Step2 Proc Minefrequentitemsets(X,H)

//X 为模糊频繁项,该过程第 1 次调用时 $X=\varnothing$,H 为算法中 Step1 构建的 FP-tree 的头表.

{ For each item I in H // 自顶向下的顺序处理 H 中的 item I

If $H(I)>=$ min_sup 并且 X 中没有与 I 相同的数量属性生成的模糊项 then

//$H(I)$ 表示模糊项 I 的支持度计数

{$L=L\cup IX$;

// 将模糊频繁项 IX 加入集合 L

调用过程 Buildsubtable(I) 创建新的头表 H_1;

Minefrequentitemsets(IX,H_1);

}

}

Proc Buildsubtable (I)

{For each 结点 n in I. sidelink

// I. sidelink 表示 FP-tree 中模糊项名称为 I 的结点的集合

{$c=n$;// 将结点 n 赋给 c

while($c\rightarrow$ parent! = root)

//root 为 FFP-tree 的根结点

{$b=c->$ parent;

Temp = triangular(n. tempmembershipD, b. membershipD);

// Temp 记录了当前遍历路径上数据库中事务对结点 n 和 b 构成的模糊频繁项的隶属度值和事务标识,由顺序链表实现,triangular 是表 1 中的任何一种 t- 模算子.

If 头表 H_1 中没有结点 b 中的模糊项 b. item then

{将 b. item 插入头表 H_1,b. item 的支持度计数由 Temp 计算;

将 b. tempmembershipD 中存储的所有隶属度值置 0,将 Temp 中的隶属度值 hash 存储在 b. tempmembershipD 中;

将结点 b 插入头表 H_1 表项 b. item 的链表尾,并将结点 b 指向 FP-tree 中下一个 b. item 结点的指针置空;

}

Else

{由 Temp 更新头表 H_1 中模糊项 b. item 的支持度计数;

If 结点 b 不在头表 H_1 表项 b.item 的链表中 then

{b.tempmembershipD 中存储的所有隶属度值置 0，将 Temp 中的隶属度值 hash 存储在 b.tempmembershipD 中；}

Else

将 Temp 中的隶属度值 hash 存储在 b.tempmembershipD 中；

}

$c = c$ —> parent；

}

}

}.

Step3　Proc Minefuzzyrule(L)

// 算法输入为除 1- 模糊频繁项的模糊频繁项集 L

{For each $1 \in L$

For each $1' \subset L$

If Dimp($1' \Rightarrow 1-1'$) > min_imp then

$R = R \cup (1' \Rightarrow 1-1')$

// 集合 R 强模糊关联规则集

}.

5　算法复杂度分析与实验结果

5.1　算法复杂度分析

设模糊数据项的个数为 m，模糊 1- 频繁项的个数为 k，数据库中事务数量为 n，算法产生的模糊频繁项的最大长度为 q，则产生的模糊频繁项的个数最多为 $|C| = c_k^1 + c_k^2 + \cdots + c_k^q$. Apriori 算法扫描数据集的时间复杂度最坏情况为 $O(|C|mn)$，而本文算法扫描数据集的时间复杂度为 $O(mn)$，其时间开销主要在于对内存 FP-tree 和 hash 区的扫描. 对 hash 区扫描的代价与数据集大小无关，其时间代价是 $O(1)$. 本文算法的空间开销有 3 部分：FP-tree，hash 区和递归生成的头表. 其中 FP-tree 和 hash 区所占内存大小由对原始数据集的压缩比决定. 对于递归生成的头表考虑最坏情形，初始头表中的前 $q-1$ 项由递归生成的表项总和为 $2^{q-1}-q+1$，初始头表从第 q 至 k 表项，每一表项的递归最大深度为 $q-1$，递归生成的总表项为 $(k-q+1)\times(2^{q-1}-1)$，加上初始头表的 k 个表项，头表的空间复杂度最坏为 $O((k-q+2)\times 2^{q-1})$. Apriori 算法的空间开销为存储迭代过程中产生的候选模糊频繁项集，其最坏复杂度为 $O(|C|)$.

5.2　实验结果

实验数据来自中国气象科学数据共享服务网中国地面国际交换站气候资料日值数据集. 本文从该数据集中整理出北京、天津两台站 1951 ～ 2007 年的月平均气压、平均气温、平均相对湿度、降水量、平均风速、日照时数共 6 个数量属性，1332 条记录. 实验环境：Intel(R) P4，CPU 2.80GHz，512RAM，Win2000，Visual C++6.0. 每个数量属性通过模糊聚类划分为高、中、低 3 个等级. 表 2 列出了部分数量属性聚类中心和模糊划分区间.

表 2　部分数量属性的模糊聚类结果

等级	平均气压		平均气温		平均相对湿度	
	聚类中心	划分区间	聚类中心	划分区间	聚类中心	划分区间
高	10243	[10162, 10342]	23.84	[17.6,29.6]	74.29	[64,86]
中	10143	[10052, 10223]	12.68	[5.5,19.1]	58.98	[48,70]
低	10030	[9951, 10121]	−0.8	[−7.6,9.8]	43.61	[22,52]

表 3 ～ 表 5 列出了采用 3 种 t- 模算子，最小支持度为 0.2，最小蕴涵度为 0.95，挖掘出的强模糊关联规则. 由结果可知，不同的 t- 模算子挖掘出的具有相同模糊项的强关联规则的蕴涵度相差不大. 但在相同的最小支持度和蕴涵度下，挖掘得到的强模糊关联规则集不同. 挖掘出的模糊关联规则具有不同的逻辑语义. 例如：表 3 中的规则气压高 => 降水量低，可解释为：气压越高降水量越低；而表 4，表 5 中同样的规则可解释为：气压越高，降水量低的确定性程度越大. 其中表 3 中的 3 条规则：湿度中 => 降水量低，风速中，湿度中 => 降水量低，日照时数低 => 降水量低，没有在表 4，表 5 中出现. 因此，给定最小支持度和蕴涵度，由相同的模糊项组成的关联规则在渐近性逻辑语义下成立，而在确定性逻辑语义下不成立.

本文算法在不同支持度下与基于 apriori 模糊频繁项挖掘算法运行比较如图1所示，图中的时间

表 3　min(α,β) 算子挖掘结果

模糊关联规则	蕴涵度
湿度中 => 降水量低	0.96
风速中，湿度中 => 降水量低	0.98
气压高 => 降水量低	1
气温低 => 降水量低	1
气压高，气温低 => 降水量低	1
气压低 => 气温高	0.97
气压中 => 降水量低	0.97
日照时数低 => 降水量低	0.95
湿度低 => 降水量低	1
气温中 => 降水量低	0.98

表 4 $\alpha\beta$ 算子挖掘结果

模糊关联规则	蕴涵度
气压高 =＞ 降水量低	0.99
气温低 =＞ 降水量低	1
气压高,气温低 =＞ 降水量低	1
气压低 =＞ 气温高	0.96
气压中 =＞ 降水量低	0.96
湿度低 =＞ 降水量低	0.99
气温中 =＞ 降水量低	0.98

表 5 $\max(\alpha+\beta-1,0)$ 算子挖掘结果

模糊关联规则	蕴涵度
气压高 =＞ 降水量低	0.99
气温低 =＞ 降水量低	1
气压高,气温低 =＞ 降水量低	1
气压低 =＞ 气温高	0.95
湿度低 =＞ 降水量低	0.99
气温中 =＞ 降水量低	0.97

为挖掘模糊频繁项的所需时间. FP-tree中的每个结点的hash地址空间大小m取128. 不难看出,本文算法相对基于 apriori 的模糊频繁项挖掘算法的在时间方面的有效性.

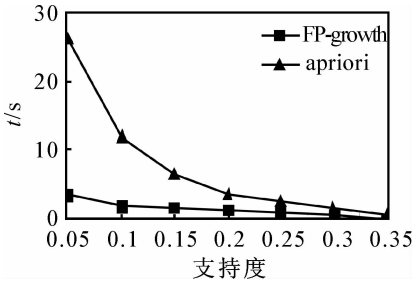

图 1 本文算法与 apriori 模型模糊频繁项挖掘运行时间比较

6 结 论

本文提出了一种基于 TD-FP-growth 的模糊关联规则挖掘算法. 实验表明,对于给定的数据集,在相同的最小支持度和蕴涵度下,同一模糊关联规则在渐近性逻辑语义下成立,而在确定性逻辑语义下不成立. 因此,从逻辑的角度考虑模糊关联规则的挖掘是必要的. 通过散列存储数据库中每个事务对FP-tree 中每个结点表示的模糊项的隶属度减少了计算模糊支持度的时间开销. 实验结果表明了本文算法在时间上的有效性. 本文算法在模糊分类系统中的应用以及不同语义的规则对分类准确率的影响将是下一步研究工作.

参考文献(References)

[1] Gyenesei A. A fuzzy approach for mining quantitative association rules [R]. Finland: Truku Center for Computer Science, 2000.

[2] 张雷, 李人厚. 基于免疫原理的模糊关联规则挖掘算法[J]. 控制与决策, 2008, 23(8): 957-960.
(Zhang L, Li R H. Algorithm for mining fuzzy association rules based on immune principles[J]. Control and Decision, 2008, 23(8): 957-960.)

[3] 高雅, 马琳, 戴齐. 模糊关联规则的挖掘算法[J]. 西南交通大学学报, 2005, 40(1): 26-29.
(Gao Y, Ma L, Dai Q. Algorithms of mining fuzzy association rules [J]. J of Southwest Jiaotong University, 2005, 40(1): 26-29.)

[4] 陆建江, 宋自林,钱祖平. 挖掘语言值关联规则[J]. 软件学报, 2001, 12(4): 607-611.
(Lu J J, Song Z L, Qian Z P. Mining linguistic value association rule[J]. J of Software, 2001, 12(4): 607-611.)

[5] Miguel Delgado, Nicolas Marin, Daniel San che2, et al. Fuzzy association rules: General model and applications [J]. IEEE Trans on Fuzzy Systems, 2003, 11(2): 214-225.

[6] Eyke Hüllermeier. Fuzzy methods in machine learning and data mining: Status and prospects[J]. Fuzzy Sets and Systems, 2005, 156(3): 387-406.

[7] Didier Dubois, Henri Prade, Thomas Sudkamp. On the representation, measurement, and discovery of fuzzy associations [J]. IEEE Trans on Fuzzy Systems, 2005, 13(2): 250-262.

[8] Chen G Q, Yan Peng, Kerre Etienne E. Computationally efficient mining for fuzzy implication-based rules in quantitative databases[J]. Int J of General Systems, 2004, 33(2): 163-182.

[9] Lopez F J, Blanco A, Garcia F, et al. Fuzzy association rules for biological data analysis: A case study on yeast [J]. BMC Bioinformatics, 2008, 9(1): 107.

[10] Mathieu Serrurier, Didier Dubois, Henri Prad, et al. Learning fuzzy rules with their implication operators [J]. Data and Knowledge Engineering, 2007, 60(1): 71-89.

[11] Wang Ke, Tang Liu, Han Jiawei, et al. Top down FP-growth for association rule mining[C]. Proc of the 6th Pacific Area Conf on Knowledge Discovery and Data Mining. Taipei, 2002: 334-340.

[12] Didier Dubois, Eyke Hullermerier, Henri Prade. A systematic approach to the assessment of fuzzy association rules [J]. Data Mining and Knowledge Discovery, 2006, 13(2): 167-192.