一种新的关联规则挖掘方法

彭 珍[1,2],裴丽丽[3],杨炳儒[1]
PENG Zhen[1,2],PEI Li-li[3],YANG Bing-ru[1]

1.北京科技大学 信息工程学院 知识工程研究所,北京 100083
2.华北科技学院 计算机系,北京 101601
3.唐山工业职业技术学院,河北 唐山 063020
1.School of Information Engineering,Beijing University of Science and Technology,Beijing 100083,China
2.Department of Computer,North China Institute of Science and Technology,Beijing 101601,China
3.Tangshan Industrial Vocation-Technical College,Tangshan,Heibei 063020,China
E-mail:yx_dpzc@yahoo.com.cn

PENG Zhen,PEI Li-li,YANG Bing-ru.One new association rules mining approach.Computer Engineering and Applications,2009,45(27):127-129.

Abstract: Mining association rules is one of the important tasks in data mining.With the aim to further improve the cognitive feature and the performance of association rules mining algorithm,the paper proposes one new idea of association rules mining and one RBFCM-based association rules mining algorithm,which uses rule based fuzzy cognitive map to represent knowledge and to be accessible fuzzy inference to each association rule mined as so to reduce the frequency of interaction with the database. And the experiment demonstrates that the approach effectively increases the effectiveness of association rules mining and the intelligence compared with the Apriori algorithm.

Key words: data mining;frequent itemsets;association rules;Rule Based Fuzzy Cognitive Map(RBFCM);accessible inference

摘 要:关联规则挖掘是数据挖掘的主要任务之一。为了进一步提高关联规则挖掘算法的认知特性和运算效果,提出了一种新的关联规则挖掘思想并由此构造了一种基于规则模糊认知图的关联规则挖掘算法。该算法使用规则模糊认知图进行知识表示,对每个挖掘到的关联规则进行可达模糊推理,从而减少了与数据库交互的次数。实验证明该方法与 Apriori 的关联规则算法相比,提高了关联规则挖掘的效率,增强了智能化程度。

关键词:数据挖掘;频繁项集;关联规则;规则模糊认知图;可达推理
DOI:10.3778/j.issn.1002-8331.2009.27.038 **文章编号:**1002-8331(2009)27-0127-03 **文献标识码:**A **中图分类号:**TP182

1 引言

关联规则挖掘[1-2]是数据挖掘中最活跃的研究方法之一。一般地,给定一个事物数据库,关联规则挖掘问题就是通过指定的最小支持度和最小可信度来寻找强关联规则的过程。关联规则挖掘可以划分为两个子问题:一是发现频繁项集,即按照从小到大的顺序查找满足最小支持度的所有项目子集;二是生成关联规则,即在频繁项集中找到满足最小可信度的关联规则。而在关联规则挖掘中最关键且最耗时的是频繁项集挖掘算法。对经典的频繁项集挖掘算法比由 Apriori[3]与 FP-growth[4]等的改进在一定程度上提高了关联规则挖掘的效率,但未从根本上对关联规则挖掘的效率和智能化程度等方面有较大的完善。

该文提出了一种新的关联规则挖掘方法,它改变了传统关联规则挖掘模式,其设计思想是基于 RBFCM[5-6]的知识表示和推理机制,具体而言是按照从大到小的顺序查找频繁项集,每找到一个频繁项集就能够自主地模糊推理出其他的可达关联规则。这种方法大大减少了与事务数据库交互的次数,提高了关联规则挖掘的效率和智能性。

2 RBFCM 知识表示和推理机制

2.1 RBFCM 的知识表示方法

模糊认知图(Fuzzy Cognitive Map,FCM)[7-8]是一种软计算工具,它的概念及概念间的关系是模糊变量,知识存储在概念结点及概念结点间的关系中。但是,FCM 中每个概念及概念间的关系只具有一个模糊成员函数,只适应概念间的简单模糊关系,不能处理概念间的"and"关系。而规则模糊认知图(Rule Based Fuzzy Cognitive Maps,RBFCM)能够解决这些问题,它的知识表示和推理能力更强。该文提出了一种用于关联规则挖掘的 RBFCM,定义如下:

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60675030,No.60875029)。
作者简介:彭珍(1981-),女,博士研究生,讲师,CCF 会员,主要研究领域为数据挖掘;杨炳儒(1943-),教授,博士生导师,主要研究领域为知识发现与智能系统、柔性建模与集成技术。
收稿日期:2009-02-17 **修回日期:**2009-03-27

定义 1 RBFCM 的拓扑结构 U 是一个四元组即 $U=(C,E,W,S)$，其中 $C=\{C_1,C_2,\cdots,C_n\}$表示 RBFCM 的所有素概念结点以及关系结点所涉及到的合概念结点(包括指向关系结点的所有概念结点的与、关系结点指向的所有概念结点的与)；$E=\{<C_i,C_j>|C_i,C_j\in C\}$表示概念结点之间的有向弧，即关系结点涉及到的有向弧；$\boldsymbol{W}=\{W_{ij}|W_{ij}$ 是有向弧$<C_i,C_j>$的权值$\}$；$S=\{S_i|S_i$ 是 C_i 的状态值$\}$。

如图 1 所示，每个 RBFCM 中每个素概念结点代表的是数据库中的每个属性，这些结点存放于 Nodes 集合中(包括所有的素结点和已存在的合结点)。Nodes 对应一个 S，其中的 S_j 是概念结点 C_j 对应的状态值，它等于 $\sigma(C_j)/N$，其中 $\sigma(C_j)$表示在数据集中 C_j 为真的记录数，N 是数据记录的总数。

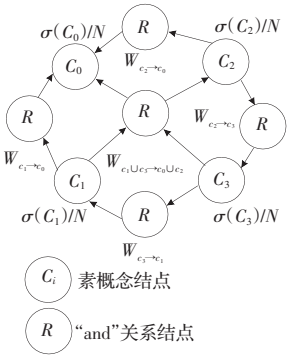

图 1　一个规则模糊认知图模型

	C_0	C_1	C_2	C_3	$C_0\cup C_1$	$C_0\cup C_2$	$C_0\cup C_3$	$C_1\cup C_3$	…
C_0	−1	0	0	0	0	0	0	0	…
C_1	1	−1	0	0	0	0	0	0	…
C_2	1	0	−1	1	0	0	0	0	…
C_3	0	1	0	−1	0	0	0	0	…
$C_0\cup C_1$	−1	−1	0	0	−1	0	0	0	…
$C_0\cup C_2$	−1	0	−1	0	0	−1	0	0	…
$C_0\cup C_3$	−1	0	0	−1	0	0	−1	0	…
$C_1\cup C_3$	0	−1	0	−1	0	1	0	−1	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

图 2　规则模糊认知图的关联矩阵 W

RBFCM 中每个关系结点都对应一个边权值 w，表示为结点 C_i 与 C_j 之间存在着"if C_i then C_j"的概率关系规则，其中 C_i 为指向该关系结点的各个概念结点的"与"，C_j 为该关系结点指向的各个概念结点的"与"，关系结点的个数即为规则的个数。

定理 1 任意结点 C_j，若存在 $C_i\subset C_j$ 且 $C_j\subset C_k$，那么 C_i 是 C_j 结点的子结点；C_k 是 C_j 结点的父结点，并且 C_j 结点的父结点的状态值都小于等于其子结点的状态值，该结点的状态值处于其子结点中的最小状态值与其父结点中的最大状态值之间。

证明 假设有任意结点 C_j，那么对于其任意子结点 $C_i\subset C_j$，其任意父结点 $C_k\supset C_j$，$\delta(C_j)$是指 C_j 属性值在数据库中成立的数量。所以 $\delta(C_j)\geqslant\delta(C_j)$且 $\delta(C_j)\geqslant\delta(C_k)$成立，那么任意父结点的状态值 $\delta(C_k)/N$ 必定小于等于任意子结点的状态值 $\delta(C_i)/N$，且 C_j 结点的状态值必定处于父结点中最大状态值与子结点中最小状态值之间。

每个 RBFCM 都对应一个 $\boldsymbol{W}$ 关联矩阵，其中每个权值 w 表示为 1、0 或−1。对于必定成立的规则比如 $C_i\rightarrow C_i$、$C_i\cup C_i\rightarrow C_i$，其权值 w 记为−1，表明它们属于冗余规则，必定成立且无需参与运算。$\sup(C_i\rightarrow C_j)$表示的是规则 $C_i\rightarrow C_j$ 的支持度，可信度 con $(C_i\rightarrow C_j)$则为 $\sup(C_i\rightarrow C_j)/S_i$，两者说明了规则"if C_i then C_j"成立的模糊程度。对于其他的规则，如果支持度 $\sup(C_i\rightarrow C_j)$和可信度 $\sup(C_i\rightarrow C_j)/S_i$ 分别大于等于最小支持度 minsup 和最小可信度 mincon，规则 $C_i\rightarrow C_j$ 为关联规则，w 则记为 1；否则规则 $C_i\rightarrow C_j$ 为非关联规则，记为 0。

2.2　RBFCM 的推理机制

RBFCM 的推理机制包括推论 1~推论 3，它们能够对关联矩阵中存在关联规则推出其可达规则，缩小了关联规则挖掘的范围。

推论 1 对合结点 A 存在 $A_1\subset A$，若 A 的状态值大于等于最小支持度并且大于等于 A_1 的状态值与最小可信度的乘积，则规则 $A_1\rightarrow(A-A_1)$、$A_1\rightarrow A$ 必定成立。

证明 任意合结点 A 都可看作 $A_1\cup(A-A_1)$，对 $A_1\rightarrow(A-A_1)$而言，其支持度 $\sigma(A_1\cup(A-A_1))/N=\sigma(A)/N\geqslant$minsup，并且其可信度 $\sigma(A_1\cup(A-A_1))/\sigma(A_1)\geqslant$mincon。因此，规则 $A_1\rightarrow(A-A_1)$必定成立。对规则 $A_1\rightarrow A$，同理可证。

推论 2 若规则 $A\rightarrow B$ 成立，其中后件 B 是合结点，那么对任意 $B_1\subset$B，规则 $A\rightarrow B_1$ 必定成立。

证明 已知规则 $A\rightarrow B$ 成立，所以 $\sigma(A\cup B)/N\geqslant$minsup，$\sigma(A\cup B)/\sigma(\mathrm{A})\geqslant$mincon 必定成立。对规则 $A\rightarrow B_1$ 而言，对任意 $B_1\subset$B，其支持度 $\sigma(A\cup B_1)/N\geqslant\sigma(A\cup B)/N$；其可信度 $\sigma(A\cup B_1)/\sigma(A)\geqslant\sigma(A\cup B)/\sigma(A)$。所以 $\sigma(A\cup B_1)/N\geqslant$minsup 与 $\sigma(A\cup B_1)/\sigma(A)\geqslant$mincon 必然成立，则 $A\rightarrow B_1$ 必定成立。

推论 3 对结点 A 存在 $A_1\subseteq A$，则 $A\rightarrow A_1$ 必定成立。

证明 略(证明思路同上)。

3　RBFCM 关联规则挖掘算法

RBFCM 关联规则挖掘算法的思想是从最大的合结点入手，对每挖掘出的规则，通过 RBFCM 关联矩阵的可达推理算法可以自主地发现相关规则，从而减少了挖掘的时间，提高的挖掘效率，增加了挖掘的智能性。

3.1　可达推理算法

可达推理算法主要来源于 RBFCM 的三个推论(推论 1~3)，可达矩阵的初始化采用推论 3 的思想，对应算法 Initial；可达矩阵的推理算法采用推论 1 和推论 2 的思想，算法 AcsInference 实现。

算法 1 可达推理初始化算法 Initial

输入：关联矩阵 $\boldsymbol{W}$；

输出：更新权值为−1 的关联矩阵 $\boldsymbol{W}$

Initial($\boldsymbol{W}$)

(1) for(l=Nodes.Length();l>0;l−−)

(2)　for(r=0;r<l;r++)

(3)　　if(Nodes.IsSubNode(C_l,C_r)==1)

(4)　　W.SetValue(r,1,−1);

算法扫描关联矩阵 $\boldsymbol{W}$ 的整个左下三角，符合推论 3 的规则，函数 IsSubNode 判定规则后件是否是前件的子结点，若是将其值置为−1。

算法 2 可达推理算法 AcsInference

输入：关联矩阵 $\boldsymbol{W}$，结点标识 x；

输出：更新权值为 1 的关联矩阵 $\boldsymbol{W}$；

AcsInference(W,x)

(1) for(;r=Nodes.GetSubID(C_x);)

(2) {S.SetState(r);

```
(3)  con=S.GetState(x)/S.GetState(r);
(4)  if(con>=mincon)
(5)  {W.SetValue(r,x,1);
(6)   l=Nodes.GetSubID(C_x,C_r);
(7)   W.SetValue(r,l,1);
(8)   if(Nodes.IsCoNode(C_l))
(9)      for each C_k⊂C_l
(10)   {k=Nodes.GetID(C_k);
(11)     S.SetState(k);
(12)     W.SetValue(r,k,1);}
(13)  }
(14) }
```

算法的前七步完成了推论 1,对推论 1 产生的规则再利用推论 2 去进一步地推理。其中,Nodes 和 S 分别表示 RBFCM 的概念结点队列对象和概念状态队列的全局对象。Nodes 是按照先素结点后合结点由小到大的顺序排列,它包含所有的素结点(对应属性)和合结点,其中的每个结点都有一个编号 id。而 S 最初只包含从挖掘的所有素结点状态值。

函数 GetID 实现的是返回参数结点的标号。GetSubID 有两个重载函数,一个参数的用于获得该参数结点的子结点标号;两个参数的取第一个参数结点中除去第二个参数结点的子结点标号。如果合结点状态值未知,函数 SetState 则根据定理 1 按照取中原则,取该结点的当前所有子结点的最小状态值与所有父结点的最大状态值的平均值,它是一个模糊值,但基本上能够达到正确挖掘的目的。

3.2　关联规则挖掘算法

关联规则挖掘方法建立在可达矩阵推理的基础之上,在一定挖掘的基础上通过推理得到更多的关联规则,减少了挖掘的次数,从而提高了挖掘的效率。

算法 3　关联规则挖掘算法 RBFCM_Miner

输入:数据库对象 DB,初始关联矩阵 $\boldsymbol{W}$

输出:关联矩阵 $\boldsymbol{W}$ 中的关联规则

```
RBFCM_Miner(DB,W)
(1) Nodes.Create(DB);
(2) W.Create(Nodes);
(3) S.Create(RBFCM_FI(DB));
(4) Initial(W);
(5) for(;W.CoNode(x)==1;)
(6)  {support=RBFCM_FI(x,DB);
(7)   if(support>=minsup)
(8)     AcsInference(W,x);
(9) }
```

该算法首先根据事务数据库中的属性创建 Nodes 队列,并据此创建初始权值均为 0 的关联矩阵 $\boldsymbol{W}$,接着调用 RBFCM_FI 从事务数据库中挖掘所有素结点(属性)的状态值创建 S 队列。然后用算法 1-Initial 对关联矩阵初始化操作。

函数 CoNode 用于在关联矩阵 $\boldsymbol{W}$ 中确定最需挖掘的结点,即在 $\boldsymbol{W}$ 中未重新更新的且包含素结点最多的合结点。而后调用函数 RBFCM_FI 的另一重载函数从事务数据库中挖掘出给定结点的支持度。对找到的频繁项集利用算法 2-AcsInference 对关联规则进行可达推理。这样再一次去确定最需挖掘结点时,通过可达矩阵的推理缩小搜索的范围,从而减少了挖掘的消耗。

由对算法 3-RBFCM_Miner 的分析可见,第一个函数 RBFCM_FI(DB)仅仅挖掘素结点的状态值,它与数据库中属性个数有关而与支持度无关,且仅需扫描数据库一次。第二个函数 RBFCM_FI(x,DB)在最好的情况下,也就是说挖掘出的 support 满足最小支持度,就可以达到更新关联矩阵 $\boldsymbol{W}$ 所有权值的目的,也就是说也仅需扫描数据库一次;但在最坏的情况,也就是说挖掘出的 support 都不满足最小支持度,那么需要与数据库交互的次数是$\lceil N/3\rceil+(N-4)\times2+1$。

4　实验与结果

实验的数据集是 Chess,在主频为 1.60 GHz 的 Window XP 环境下,采用 VC++6.0 开发环境分别就 RBFCM_Miner 和基于 Apriori 的关联规则挖掘算法在不同支持度下的运行时间进行了比较,如图 3 所示。

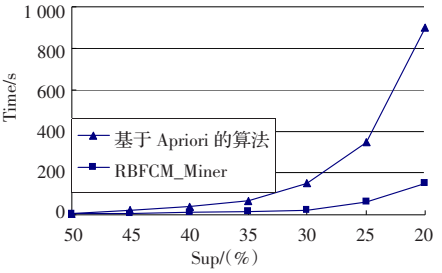

图 3　实验运行耗时比较结果

因为关联规则挖掘算法 RBFCM_ Miner 中存在的模糊推理,在不同的支持度条件下对关联规则准确度(正确的关联规则数/总关联规则)进行了实验,将推理出的关联规则与实际挖掘出的关联规则进行了对比,如表 1 所示,结果显示准确度在 90%以上。

表 1　算法 RBFCM_Miner 的准确度

支持度	准确度/(%)
0.50	90.73
0.45	91.09
0.40	91.46
0.35	92.85
0.30	95.12

5　结论

提出了一种新的关联规则挖掘思想及其算法实现,它采用从大到小的顺序进行频繁项集挖掘,并对每个得到的频繁项集进行立刻处理,利用了规则模糊认知图的模糊推理的特性,根据已知规则自主地发现相关关联规则,这种方法与典型的关联规则挖掘算法相比大大提高了进行效率,这足以弥补在准确性上存在的误差。实验证明了它的可行性。

参考文献:

[1] Jiawei Han,Kamber M.数据挖掘概念与技术[M].范明,孟小峰,译.2 版.北京:机械工业出版社,2007.

[2] Agrawal R,Imielinski T,Swami A.Mining associations between sets of items in massive databases[C]//Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD'93),Washington,DC,1993:207-216.

(下转 132 页)