多数据库中负关联规则挖掘算法

尚世菊[1],董祥军[1],李 杰[2]
SHANG Shi-ju[1],DONG Xiang-jun[1],LI Jie[2]

1.山东轻工业学院 信息科学与技术学院,济南 250353
2.广西大学 数学与信息科学学院,南宁 530004
1.School of Information Science and Technology,Shandong Institute of Light Industry,Jinan 250353,China
2.College of Mathematics and Information Science,Guangxi University,Nanning 530004,China
E-mail:shiju82@163.com

SHANG Shi-ju,DONG Xiang-jun,LI Jie.Algorithms for mining negative association rules in multi-database.Computer Engineering and Applications,2009,45(24):150-152.

Abstract: Nowadays the techniques of negative association rules mining focus on mono-database.With the rapid development of database technologies,multi-database mining is becoming more and more important.Knowledge conflicts within databases may occur when mining both the positive and negative association rules simultaneously.This paper proposes the model of synthesizing association rules and synthesis correlation to resolve conflicts on base of previous work on multi-database mining.The experimental results demonstrate that the algorithm is correct and effective.

Key words: negative association rules;data mining;multi-database

摘 要:现行的负关联规则挖掘主要是对于单一数据库的挖掘,但随着数据库技术的发展,多数据库挖掘越来越重要。当同时挖掘多数据库中的正负关联规则时,有可能会引起知识冲突问题,在前人对多数据库挖掘的基础上采用了一种关联规则合成模式,并利用相关性解决了知识冲突问题,最后用实验证明了该方法的正确性。

关键词:负关联规则;数据挖掘;多数据库
DOI:10.3778/j.issn.1002-8331.2009.24.044 **文章编号:**1002-8331(2009)24-0150-03 **文献标识码:**A **中图分类号:**TP311

1 前言

数据挖掘是一个从大量数据中发现并提取隐藏在其中的、以前未知的、潜在的有用信息和知识的过程。关联规则是数据挖掘研究的主要领域之一。通常所说的关联规则是形如$A\Rightarrow B$的正关联规则,而对于$A\Rightarrow\neg B$(或者$\neg A\Rightarrow B$、$\neg A\Rightarrow\neg B$)的负关联规则研究较少。但负关联规则却包含了非常有价值的信息,在决策分析特别是竞争分析和投资分析中有重要作用。

现行的负关联规则挖掘算法主要是针对单一数据库进行挖掘。随着数据库技术的不断发展,多数据库系统已经应用到现实生活中,决策者需要分析分布在不同分支的多个数据库,从而做出更加合理有效的决策,这就涉及多数据库挖掘问题。多数据库挖掘可以定义为从多个不同类的数据库中发现新颖的有用的模式的过程。可以分为三个步骤:第一,对多数据库进行分类。第二,挖掘每个数据库的知识,即局部模式分析。第三,把同类数据库挖掘到的知识进行合成。

针对多数据库挖掘的第三个步骤进行了详细阐述,提出了一种知识合成的模式,在文献[1]中提出了一种用权值来合成多数据源的合成模式,但该方法具有一定的局限性,它只适用于数据库大小相等的情况,而且只考虑了正关联规则的挖掘。该文采用一种新的合成多数据库中关联规则的方法,该方法适用于任意大小的数据库,而且利用合成相关性来解决在同时挖掘正负关联规则时所遇到的知识冲突问题,并设计了一个改进算法,最后用实验证明了该方法的正确性。

2 负关联规则挖掘相关技术

2.1 问题描述

设$I=\{i_1,i_2,\cdots,i_m\}$是由m个不同属性(项目)组成的集合,i_k($k=1,2,\cdots,m$)称为项(item)。事务数据库D是事务T(transaction)的集合,其事务数记作$|D|$,其中T是项的集合,并且$T\subseteq I$。对应每一个事务有唯一的标识,记作TID。设X是一个I中项的集合(项集),如果$X\subseteq T$,那么称事务T包含X。若X包含的项的个数为k($1\leqslant k\leqslant m$),则称X为k-项集。一条负关联规则就是一个形如$A\Rightarrow\neg B$(或$\neg A\Rightarrow B$、$\neg A\Rightarrow\neg B$)的蕴涵式,其中$A,B\subseteq T$而且$A\cap B=\Phi$。给定支持度$supp$和置信度$conf$,如果

基金项目:山东省自然科学基金(the Natural Science Foundation of Shandong Province of China under Grant No.Y2007G25);山东省优秀中青年科学家奖励基金项目(No.2006BS01017)。

作者简介:尚世菊(1982-),女,硕士研究生,研究方向:数据挖掘;董祥军(1968-),男,教授,硕士生导师,主要研究领域:数据挖掘与数据库技术;李杰(1980-),男,硕士研究生,研究方向:最优化理论与方法。

收稿日期:2008-05-08 **修回日期:**2008-08-05

D 中有(100×$supp$)%的事务包含 A 但不包含 B,则负关联规则 $A\Rightarrow\neg B$ 的支持度为 $supp$,如果包含 A 的事务中有(100×$conf$)%的事务不包含 B,则负关联规则 $A\Rightarrow\neg B$ 在 D 中的置信度为 $conf$。挖掘关联规则问题就是找出频繁项集中满足用户给定的最小置信度($minconf$)的关联规则。即,如果 $A\Rightarrow\neg B$ 是一条有效的关联规则,则必须满足:(1)$supp(A\Rightarrow\neg B)\geqslant minsupp$;(2)$conf(A\Rightarrow\neg B)\geqslant minconf$。该问题可以分解成如下两个子问题:(1)产生所有支持度大于最小支持度的项集;(2)对于每个频繁项集,产生所有比最小置信度大的规则。

2.2　负关联规则的挖掘技术

Brin 等人 1997 年首次中提到了两个频繁项集间的负相关[2],Savasere 等人在文献[3]中阐述了强关联规则,文献[4]提出了一种基于兴趣度的正负关联规则挖掘算法,文献[5]提出了一种基于支持度、置信度、关联系数的正负关联规则挖掘算法。文献[6]提出了一种 PNARC 模型,该模型采用相关性检验的方法,不仅能够同时挖掘出频繁项集中的正负关联规则,而且能够检测并删除相互矛盾的规则。文献[7]中给出一种基于多置信度和 χ^2 检验的挖掘正负关联规则的方法。文献[8]将相关强度和最小支持度结合起来,提出了一种新的度量量 VRRCC 并提出了一种 PNAR_MLMS 算法,能正确地从频繁项集和非频繁项集中挖掘出正负关联规则。

3　多数据库中负关联规则挖掘技术

在科学研究及应用中,通常用权值的方法来分析和合成不同数据库的信息,这里利用文献[9]中设定数据库的权值的方法来合成各数据库中的关联规则。文献[1]认为各个子公司相对于总公司具有平等的投票权,但是从商业理念来说,各个子公司的投票权是不一样的。例如在大城市的某个子公司的销售额是小城镇子公司销售额的 10 倍,则大城市的子公司应该具有较大的决策权。假设若一个数据库的交易事务数越多,则表明该公司的销售额也越高,相应地权值也就越大,因此数据库的权值与交易事务数是成正比的,这与现实生活中的应用也是一致的。

3.1　数据库及规则的权值

设 $D_1,D_2,\cdots,D_m$ 为 m 个不同的数据库,$S_1,S_2,\cdots,S_m$ 分别为各同类数据源的关联规则集,$S=\{S_1,S_2,\cdots,S_m\}$为总关联规则集,$R_1,R_2,R_j,\cdots,R_n$ 为总规则集 S 中具体的关联规则,其中 $j=1,2,3,\cdots,n$,$Num(D_i)$为数据库 D_i 的事务数,则数据库 D_i 的权值为:

$$\omega_{D_i}=\frac{Num(D_i)}{\sum_{i=1}^{m}Num(D_i)}$$

规则 R_j 权值为包含该规则的数据库的权值之和,规范化的规则的权值为:

$$\omega_{R_j}=\frac{\sum_{i=1\,R_j\subset S_i}^{m}\omega_{D_i}}{\sum_{j=1}^{n}\sum_{i=1\,R_j\subset S_i}^{m}\omega_{D_i}}$$

3.2　去除知识冲突精简规则集

在对多数据库进行负关联规则挖掘时,一个数据库中的规则可能会和其他数据库中的规则产生冲突,例如,若 D_1 中有规则 $A\Rightarrow B$,D_2 中有规则 $A\Rightarrow\neg B$,这样就产生了矛盾。为了得到正确的关联规则,采用相关性来进行判断,文献[6]利用相关性来检测矛盾规则只是用于单一数据库,这里则将该方法运用到多数据库中。(1)$A\Rightarrow B$,(2)$A\Rightarrow\neg B$,(3)$\neg A\Rightarrow B$,(4)$\neg A\Rightarrow\neg B$,很明显(1)(4)与(2)(3)是相互矛盾的,若数据库中存在这种矛盾的规则就要对项集 A、B 合成后的相关性进行判断:

$$corr_{\omega(A,B)}=\frac{supp_\omega(AB)}{supp_\omega(A)supp_\omega(B)}$$

其中 $supp_\omega(AB)$、$supp_\omega(A)$、$supp_\omega(B)$是频繁项集合成后的支持度。

(1)如果 $corr_\omega(A,B)>1$,仅挖掘规则 $A\Rightarrow B$ 和$\neg A\Rightarrow\neg B$;

(2)如果 $corr_\omega(A,B)<1$,挖掘规则 $A\Rightarrow\neg B$ 和$\neg A\Rightarrow B$;

(3)如果 $corr_\omega(A,B)=1$,不挖掘规则。

这样进行判断之后,数据库中的矛盾的规则将被去除掉。

去掉矛盾的规则后,各子公司挖掘出的规则结果提交给总公司时所产生的数据集还是非常大的,因此在对关联规则合成前应该先对其进行预处理。设定一个有效投票率 min.$\gamma effective$,将规则的权值与之相比较,将那些权值小于该阈值的无太大意义的噪声规则删除。

3.3　合成负关联规则

设 $D_1,D_2,\cdots,D_m$ 为 m 个不同的数据库,规则集 S_i 是数据库 $D_i(i=1,2,\cdots,m)$中的关联规则集,$\omega_{D_1},\omega_{D_2},\cdots,\omega_{D_m}$ 分别是数据库 $D_1,D_2,\cdots,D_m$ 的权值,对于特定的关联规则 $A\Rightarrow B$,$A\Rightarrow\neg B$(或$\neg A\Rightarrow B$,$\neg A\Rightarrow\neg B$)合成模式为:

$$supp_\omega(A\Rightarrow B)=\omega_{D_1}\times supp_1(A\Rightarrow B)+\omega_{D_2}\times supp_2(A\Rightarrow B)+\cdots+\omega_{D_m}\times supp_m(A\Rightarrow B)$$

$$supp_\omega(A\Rightarrow\neg B)=\omega_{D_1}\times supp_1(A\Rightarrow\neg B)+\omega_{D_2}\times supp_2(A\Rightarrow\neg B)+\cdots+\omega_{D_m}\times supp_m(A\Rightarrow\neg B)$$

$$conf_\omega(A\Rightarrow B)=\frac{supp_\omega(A\cup B)}{supp_\omega(A)}$$

$$conf_\omega(A\Rightarrow\neg B)=\frac{supp_\omega(A\cup\neg B)}{supp_\omega(A)}$$

根据合成模式可以得到各规则的支持度和置信度,这样就可以把有趣的关联规则提交给总公司决策之用。

4　算法设计

算法 $RuleSelection(S)$

输入:$min.\gamma effective$,最小有效投票率;S,个数为 N 的规则集;ω_{D_i},数据库 D_i 的权值,$i=1,2,\cdots,m$。

输出:S,缩减了的规则集。

(1) If 规则集 S 中存在矛盾的规则 do

$corr_{\omega(A,B)}=\frac{supp_\omega(AB)}{supp_\omega(A)supp_\omega(B)}$

If $corr_\omega(A,B)>1$

$S\leftarrow S-\{\neg A\Rightarrow B,A\Rightarrow\neg B\}$

If $corr_\omega(A,B)<1$

$S\leftarrow S-\{A\Rightarrow B,\neg A\Rightarrow\neg B\}$

(2) For 对于在规则集 S 中每一个规则 R do

$\omega_{R_j}\leftarrow\frac{\sum_{i=1\,R_j\subset S_i}^{m}\omega_{D_i}}{\sum_{j=1}^{n}\sum_{i=1\,R_j\subset S_i}^{m}\omega_{D_i}}$

If $\omega_{R_j}<min.\gamma effective$

$S\leftarrow S-\{R_j\}$;

end for;

(3) output S;

通过规则选取后,规则集的数量就减少了,然后再利用数据库的权值来合成关联规则。

算法 Rule Synthesizing

输入:$S_1,S_2,\cdots,S_m$ 规则集;*minsupp*,支持度阈值;*minconf*,置信度阈值。

输出:合成后的关联规则。

(1) $S\leftarrow\{S_1\cup S_2\cup\cdots\cup S_m\}$;

(2) Call Rule Selection(S);

(3) for 规则集中的每条规则 $A\Rightarrow B$ do

$$supp_\omega(A\Rightarrow B)=\omega_{D_1}\times supp_1(A\Rightarrow B)+\omega_{D_2}\times supp_2(A\Rightarrow B)+\cdots+\omega_{D_m}\times supp_m(A\Rightarrow B)$$

$$conf_\omega(A\Rightarrow B)=\frac{supp_\omega(A\cup B)}{supp_\omega(A)}$$

(4)按支持度的高低排列规则集 S 中的关联规则 R

(5)输出 S 中支持度和置信度大于等于阈值的关联规则 R

5 数值实验及分析

为了更好地说明算法,举个简单的例子(表 1)。

表 1 实验数值

数据库	事务数	规则	*SuppL*	*SuppR*	*Supp*	*Conf*
D_1	0.50	$A\Rightarrow B$	0.3	0.5	0.2	0.667
		$B\Rightarrow\neg C$	0.5	0.4	0.3	0.600
		$C\Rightarrow D$	0.6	0.3	0.3	0.500
		$E\Rightarrow F$	0.8	0.2	0.2	0.250
D_2	0.25	$A\Rightarrow B$	0.5	0.3	0.3	0.600
		$C\Rightarrow\neg D$	0.4	0.3	0.2	0.500
		$E\Rightarrow\neg F$	0.8	0.4	0.6	0.750
D_3	0.25	$A\Rightarrow B$	0.4	0.3	0.2	0.500
		$B\Rightarrow\neg C$	0.3	0.3	0.2	0.667
		$C\Rightarrow\neg D$	0.7	0.4	0.3	0.429
		$E\Rightarrow F$	0.7	0.5	0.4	0.571

从表中可以得到数据库的权值为 ω_{D_1}=事务数/总的事务数=0.5/(0.5+0.25+0.25)=0.5,同理可得 ω_{D_2}=0.25,ω_{D_3}=0.25。规则集 $S=\{R_1:A\Rightarrow B,R_2:B\Rightarrow\neg C,R_3:C\Rightarrow D,R_4:C\Rightarrow\neg D,R_5:E\Rightarrow F,R_6:E\Rightarrow\neg F\}$,其中 R_3 与 R_4,R_5 与 R_6 是相互矛盾的,所以应该利用合成相关性来确定哪个才是正确的规则。

$$corr_\omega(C,D)=\frac{supp_\omega(CD)}{supp_\omega(C)supp_\omega(D)}=\frac{(0.5\times0.3+0.25\times0.2+0.25\times0.4)}{[(0.5\times0.6+0.25\times0.4+0.25\times0.7)(0.5\times0.3+0.25\times0.7+0.25\times0.6)]}=1.098\ 6>1$$

同理可以计算 $corr_\omega(E,F)=0.86<1$,所以 $C\Rightarrow D$ 和 $E\Rightarrow\neg F$ 是正确的规则,将规则 $C\Rightarrow\neg D$,$E\Rightarrow F$ 从规则集中删除。现在的规则集为 $S=\{R_1:A\Rightarrow B,R_2:B\Rightarrow\neg C,R_3:C\Rightarrow D,R_6:E\Rightarrow\neg F\}$。

求得各个规则的权值为:

$\omega_{R_1}=0.5+0.25+0.25=1$,$\omega_{R_2}=0.5+0.25=0.75$,$\omega_{R_3}=0.5$,$\omega_{R_6}=0.25$,$\omega_R=1+0.75+0.5+0.25=2.5$

规范化的权值为:$\omega_{R_1}=0.4$,$\omega_{R_2}=0.3$,$\omega_{R_3}=0.2$,$\omega_{R_6}=0.1$。

设定 $min.\gamma effective=0.2$,则规则 R_6 将从规则集中删除,只需计算 R_1、R_2、R_3 合成后的支持度和置信度。

R_1:$supp(A\Rightarrow B)=0.5\times0.2+0.25\times0.3+0.25\times0.2=0.225$

$$conf_\omega(A\Rightarrow B)=\frac{supp_\omega(A\cup B)}{supp_\omega(A)}=\frac{0.225}{(0.5\times0.3+0.25\times0.5+0.25\times0.4)}=0.6$$

同理得到 R_2 的支持度和置信度为 0.2、0.375,R_3 的支持度和置信度为 0.3、0.5。

假如设定的支持度和置信度阈值分别为 0.2、0.3,则最后提交给总公司的关联规则为 R_1、R_2、R_3。

为了更好地说明该算法的正确性,利用 3 个合成数据库验证了该算法。这 3 个数据库具有的事务数分别为 400、600、1 000 个,每个数据库具有的属性值为 98、96、99,每行的属性平均数分别为 4、5、5。利用文献[6]中的单一数据库的挖掘方法和该文采用的方法得到了按支持度和置信度排序的相同的前 20 条关联规则,这证明了该方法的正确性。

6 结论

研究了当同时挖掘多数据库中的正负关联规则时带来的知识冲突问题及解决办法,并采用了一种利用数据库的事务数来确定数据库的权值,并利用该权值合成规则的模式。实验证明了该方法的正确性。

参考文献:

[1] 唐懿芳,牛力,张师超.多数据源关联规则挖掘算法研究[J].广西师范大学学报:自然科学版,2002,20(4):27-31.

[2] Brin S,Motwani R,Silverstein C.Beyond market:Generalizing association rules to correlations[C]//Processing of the ACM SIGM OD Conference 1997.New York:ACM Press,1997:265-276.

[3] Savasere A,Omiecinski E,Navathe S.Mining for strong negative associations in a large database of customer transaction[C]//Proceedings of the IEEE 14th Int Conference on Data Engineering.Los Alamitos:IEEE-CS,1998:494-502.

[4] Wu Xin-dong,Zhang Cheng-qi,Zhang Shi-chao.Mining both positive and negative association rules[C]//Proceedings of the 19th International Conference on Machine Learning (ICML-2002).San Francisco:Morgan Kaufmann Publishers,2002:658-665.

[5] Antonie M-L,Zaiane O.Mining positive and negative association rules[C]//LNCS 3202,Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD04),Pisa,Italy .Berlin Heidelberg :Springer-Verlag,2004:27-38.

[6] 董祥军,王淑静,宋瀚涛,等.负关联规则的研究[J].北京理工大学学报,2004,24(11):978-981.

[7] Dong X,Sun F,Han X,et al.Study of positive and negative association rules based on multi-confidence and Chi-squared test[C]//LNCS 4093.Berlin Heidelberg:Springer-Verlag,2006:100-109.

[8] Dong X,Niu Z,Shi X,et al.Mining both positive and negative association rules from frequent and infrequent itemsets [C]//LNCS 4632:3rd International Conference on Advanced Data Mining and Applications,ADMA 2007.Berlin Heidelberg:Springer-Verlag,2007:122-133.

[9] Ramkumar T,Srinivasan R.Modified algorithms for synthesizing high-frequency rules from different data sources[J].Knowledge and Information Systems,2008,17(3):313-334.