ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.10, October 2008, pp.2597−2610
DOI: 10.3724/SP.J.1001.2008.02597

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

可信关联规则及其基于极大团的挖掘算法*

肖 波[1,2+], 徐前方[2], 蔺志青[1], 郭 军[2], 李春光[2]

[1](北京邮电大学 电信工程学院,北京 100876)

[2](北京邮电大学 信息工程学院,北京 100876)

Credible Association Rule and Its Mining Algorithm Based on Maximum Clique

XIAO Bo[1,2+], XU Qian-Fang[2], LIN Zhi-Qing[1], GUO Jun[2], LI Chun-Guang[2]

[1](School of Telecommunication Engineering, Beijing University of Posts & Telecommunications, Beijing 100876, China)

[2](School of Information Engineering, Beijing University of Posts & Telecommunications, Beijing 100876, China)

+ Corresponding author: E-mail: xiaobo@bupt.edu.cn

Xiao B, Xu QF, Lin ZQ, Guo J, Li CG. Credible association rule and its mining algorithm based on maximum clique. *Journal of Software*, 2008,19(10):2597−2610. http://www.jos.org.cn/1000-9825/19/2597.htm

Abstract: Existing association-rule mining algorithms mainly rely on the support-based pruning strategy to prune its combinatorial search space. This strategy is not quite effective in the process of mining potentially interesting low-support patterns. To solve this problem, the paper presents a novel concept of association pattern called credible association rule (CAR), in which each item has the same support level. The confidence directly reflects the credible degree of the rule instead of the traditional support. This paper also proposes a MaxCliqueMining algorithm which creates 2-item credible sets by adjacency matrix and then generates all rules based on maximum clique. Some propositions are verified and which show the properties of CAR and the feasibility and validity of the algorithm. Experimental results on the alarm dataset and Pumsb dataset demonstrate the effectiveness and accuracy of this method for finding CAR.

Key words: credible association rule; maximum clique; data mining; adjacency matrix; alarm correlation

摘 要: 目前的关联规则挖掘算法主要依靠基于支持度的剪切策略来减小组合搜索空间.如果挖掘潜在的令人感兴趣的低支持度模式,这种策略并非有效.为此,提出一种新的关联模式——可信关联规则(credible association rule,简称 CAR),规则中每个项目的支持度处于同一数量级,规则的置信度直接反映其可信程度,从而可以不必再考虑传统的支持度.同时,提出 MaxCliqueMining 算法,该算法采用邻接矩阵产生 2-项可信集,进而利用极大团思想产生所有可信关联规则.提出并证明了几个相关命题以说明这种规则的特点及算法的可行性和有效性.在告警数据集及 Pumsb 数据集上的实验表明,该算法挖掘 CAR 具有较高的效率和准确性.

关键词: 可信关联规则;极大团;数据挖掘;邻接矩阵;告警关联

中图法分类号: TP311 **文献标识码**: A

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z417 (国家高技术研究发展计划(863)); the 111 Project of China under Grant No.B08004 (高等学校学科创新引智计划)

Received 2007-08-08; Accepted 2008-01-10

1 问题背景

关联规则(association rule)挖掘技术是数据挖掘研究的重要内容之一,旨在从大量数据中提取人们未知却又潜在有用的规则.Agrawal 等人在 1993 年首先提出从交易数据库发现项目间关联规则的相关问题,并给出了基于频繁集的 Apriori 算法[1,2].该算法以递归统计为基础,以最小支持度为依据剪切生成频繁集.此后出现的各种挖掘方法大都依靠这一策略来减小组合搜索空间,以加快产生频繁集,对于每个频繁集再使用最小置信度参数评价规则的有效性[3,4].关联规则挖掘也被应用到有序事件的频繁模式发现中,如告警关联分析(alarm correlation analysis),采用的方法依然是支持度-置信度框架[5,6].

我们在对某省电信公司 GPRS 网管系统告警数据库[6]进行告警关联分析时发现,只有一小部分告警发生非常频繁,而绝大多数告警发生次数非常少.图 1 给出了所有告警的支持度分布.将告警按支持度分布分为 G_1,G_2,G_3,G_4,G_5 五组,表 1 显示了每一分组包含告警项的数量,其中有 87.9%的告警支持度小于 1%,显然,数据集具有杂乱的支持度分布(skewed support distribution).如果仍然采用上述支持度-置信度框架对这种数据集进行挖掘,则无法选择合适的最小支持度参数.若将最小支持度设置得较高,将会遗漏支持度较低但令人感兴趣的规则.这类关联规则可能置信度较高,对于人们预防网络重大故障具有重要意义.若将最小支持度设置得较低,则挖掘结果会含有大量虚假规则,这些规则所包含项目的支持度处于不同数量级,可能对用户没有实际意义.

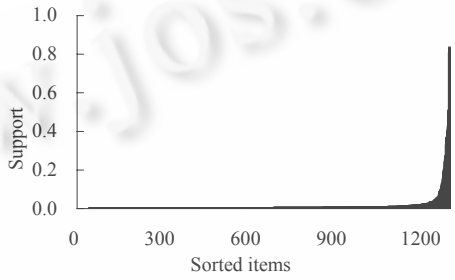


Fig.1 The support distribution of alarms in alarm database

图 1 告警数据库中告警项的支持度分布

Table 1 Groups of alarms with different support range

表 1 按照支持度分布将告警分组

Group	G_1	G_2	G_3	G_4	G_5
Support (%)	<0.1	0.1~1	1~10	10~50	>50
Items number	249	895	128	27	2
Ratio (%)	19.1	68.8	9.8	2.1	0.2

事实上,这类问题已经引起了人们的注意.Omiecinski 于 2003 年曾提出 all-confidence 兴趣度度量方法[7],Xiong 在此基础上又提出 *h*-confidence 兴趣度度量[8],这些新的度量方法都旨在减少虚假规则的产生.将它们结合到频繁项集的产生过程,可以极大地压缩生成的候选项集数量,并能挖掘出强亲密度关联模式.但这些方法大都仍然基于 Apriori 算法,不但要多次扫描数据库,而且需要判别每个候选项的兴趣度,因此时间性能偏低.我们也曾提出一种基于相关度统计的关联规则挖掘算法[6],可以同时挖掘出频繁项和非频繁项的关联规则,但需要进行大量的相关度计算.

本文提出一种新的关联模式——可信关联规则(credible association rule,简称 CAR),其定义不同于传统关联规则.可信关联规则要求所包含的每个项目的支持度处于同一数量级,而不关心整个规则支持度的大小.利用重新定义的置信度可以反映规则的可信程度,可以不再考虑传统支持度.针对可信关联规则的挖掘,本文提出 MaxCliqueMining 算法,该算法采用邻接矩阵产生 2-项可信集,进而利用极大团(maximum clique)思想产生所有可信关联规则,从而避免多次扫描数据库.同时,本文还提出并证明几个相关命题来说明这种规则的特点及算法的可行性和有效性.实验结果表明,在网络告警数据库中的确存在大量可信关联规则,并且 MaxCliqueMining 算

法针对杂乱分布数据集挖掘这类规则时具有较高的效率和准确性.

2 相关定义

设 $I=\{i_1,i_2,...,i_m\}$是 m 个不同项目的集合,给定一个事务数据库 $D=\{T_1,T_2,\ldots,T_n\}$,其中每个交易 T_i 是 I 中一组项目的集合,$T_i\subset I$.传统的关联规则定义[1]是一个形如 $X\to Y$ 的蕴涵式,其中 $X\subset I,Y\subset I$,而且 $X\cap Y=\varnothing$.如果 D 中 s% 的事务同时包含 X 和 Y,则定义关联规则 $X\to Y$ 的支持度为 s%,记 $Support(X\to Y)=s$%.若 D 中 c%的事务在包含 X 时也包含 Y,则定义关联规则 $X\to Y$ 的置信度为 c%,记 $Confidence(X\to Y)=P(Y|X)=c$%.对比传统关联规则的定义,在此提出可信关联规则的概念.

定义 1(可信关联规则). 设 $I=\{i_1,i_2,\ldots,i_m\}$是 m 个不同项目的集合,给定一个事务数据库 $D=\{T_1,T_2,\ldots,T_n\}$,其中每个交易 T_i 是 I 中一组项目的集合,$T_i\subset I$.若存在 k 个项目 $x_1,\ldots,x_k$,对于$\forall i,j\in\{1,\ldots,k\}(i\neq j)$,有$(x_i\to x_j)\wedge(\neg x_i\to\neg x_j)$,则称由这 k 个项目构成 k-项可信关联规则,记为 $R(x_1,...,x_k)$.

其中 $(x_i\to x_j)\wedge(\neg x_i\to\neg x_j)$ 为两个蕴涵式的合取式,将其进一步简化:$(x_i\to x_j)\wedge(\neg x_i\to\neg x_j)\Leftrightarrow(x_i\to x_j)\wedge(x_j\to x_i)\Leftrightarrow x_i\leftrightarrow x_j$.其物理含义为:若 x_i 出现,则 x_j 出现;若 x_i 不出现,则 x_j 不出现,即 x_i 与 x_j 是同现(co-occurrence)的.因此,可信关联规则的实质即规则中的任意两项均满足同现关系,由此表达的关联程度更强.

特别地,对于 2-项可信关联规则 $R(x_1,x_2)$,也可以直接表示为 $x_1\leftrightarrow x_2$.

定义 2(可信关联规则的支持度). 与传统关联规则支持度的定义类似,如果 D 中 s%的事务同时包含项目 $x_1,\ldots,x_k$,则可信关联规则 $R(x_1,\ldots,x_k)$的支持度为 s%,记 $S_{x_1\ldots x_k}=s$%.

为了准确度量可信关联规则的关联程度,在此给出置信度的新定义.

定义 3(可信关联规则的置信度). 如果 D 中 a%的事务包含项目 x_1,b%的事务包含项目 x_2,s%的事务同时包含 x_1 和 x_2,则 2-项可信关联规则 $x_1\leftrightarrow x_2$ 的置信度定义为 $\dfrac{s}{a+b-s}$,记为 $C_{x_1x_2}$.不难发现,新定义的置信度具有确定的物理意义:$x_1\leftrightarrow x_2$ 的置信度为当 x_1 出现或 x_2 出现时,x_1x_2 同时出现的概率,即 $C_{x_1x_2}=P(x_1\cap x_2|x_1\cup x_2)=\dfrac{|x_1\cap x_2|}{|x_1\cup x_2|}$.这里,$x_1\cap x_2$ 表示 x_1x_2 共同出现,$x_1\cup x_2$ 表示 x_1 出现或 x_2 出现,$|\cdot|$ 表示出现的次数,下同.同样地,对于 k-项可信关联规则 $R(x_1,\ldots,x_k)$,其置信度定义为 $C_{x_1\ldots x_k}=\dfrac{\left|\bigcap_{i=1}^{k}x_i\right|}{\left|\bigcup_{i=1}^{k}x_i\right|}$.设置 $minconf$,若对于$\forall i,j\in\{1,\ldots,k\}(i\neq j)$,都有 $C_{x_ix_j}\geq minconf$,则称 $minconf$ 为规则的二项集最小置信度.

命题 1. 设可信关联规则 $x_1\leftrightarrow x_2$ 的置信度为 $C_{x_1x_2}$,D 中 x_1 的支持度为 $sup(x_1)$,x_2 的支持度为 $sup(x_2)$,则有 $C_{x_1x_2}\leq any\left(\dfrac{sup(x_1)}{sup(x_2)},\dfrac{sup(x_2)}{sup(x_1)}\right)\leq\dfrac{1}{C_{x_1x_2}}$.

证明:设 D 中同时包含 x_1 和 x_2 的事务占总事务的比例为 z,即 $z=\dfrac{|x_1\cap x_2|}{|D|}$.由可信关联规则的置信度定义及容斥原理可知: $C_{x_1x_2}=\dfrac{|x_1\cap x_2|}{|x_1\cup x_2|}=\dfrac{z}{sup(x_1)+sup(x_2)-z}$,整理该式可得:

$$sup(x_1)+sup(x_2)=z\left(1+\frac{1}{C_{x_1x_2}}\right)\tag{1}$$

又 $sup(x_1)+sup(x_2)=\min\{sup(x_1),sup(x_2)\}+\max\{sup(x_1),sup(x_2)\}$,$0\leq z\leq\min\{sup(x_1),sup(x_2)\}$,代入公式(1):

$\min\{sup(x_1),sup(x_2)\}+\max\{sup(x_1),sup(x_2)\}\leq\min\{sup(x_1),sup(x_2)\}\left(1+\dfrac{1}{C_{x_1x_2}}\right)$,整理可得:

$\frac{\min\{sup(x_1),sup(x_2)\}}{\max\{sup(x_1),sup(x_2)\}} \geq C_{x_1x_2}$,因此有 $C_{x_1x_2} \leq \frac{\min\{sup(x_1),sup(x_2)\}}{\max\{sup(x_1),sup(x_2)\}} \leq 1 \leq \frac{\max\{sup(x_1),sup(x_2)\}}{\min\{sup(x_1),sup(x_2)\}} \leq \frac{1}{C_{x_1x_2}}$.

即: $C_{x_1x_2} \leq any\left(\frac{sup(x_1)}{sup(x_2)}, \frac{sup(x_2)}{sup(x_1)}\right) \leq \frac{1}{C_{x_1x_2}}$. □

命题 1 表明,当可信关联规则 $x_1 \leftrightarrow x_2$ 的置信度较高时,x_1 与 x_2 的支持度相差不大,处于同一数量级.由此得出,在进行挖掘时,可以不再考虑利用支持度产生频繁项集,而只利用可信关联规则的置信度即可挖掘出高可信的关联规则.例如,若可信关联规则 $x_1 \leftrightarrow x_2$ 的置信度 $C_{x_1x_2}$=0.8,则有 $0.8 \leq \frac{sup(x_1)}{sup(x_2)} \leq 1.25$(显然,$0.8 \leq \frac{sup(x_2)}{sup(x_1)} \leq 1.25$ 也成立).

对比传统关联规则中置信度的定义,有如下结论:

命题 2. 设传统关联规则 $x_1 \rightarrow x_2$ 的置信度为 C_1,$x_2 \rightarrow x_1$ 的置信度为 C_2,可信关联规则 $x_1 \leftrightarrow x_2$ 的置信度为 $C_{x_1x_2}$,则有 $\frac{1}{C_1}+\frac{1}{C_2}=\frac{1}{C_{x_1x_2}}+1$.

证明:由可信关联规则置信度的定义:

$$C_{x_1x_2}=P(x_1 \cap x_2 | x_1 \cup x_2)=\frac{|x_1 \cap x_2|}{|x_1 \cup x_2|}=\frac{s}{a+b-s}=\frac{1}{\frac{1}{\frac{s}{a}}+\frac{1}{\frac{s}{b}}-1}=\frac{1}{\frac{1}{P(x_2|x_1)}+\frac{1}{P(x_1|x_2)}-1}=\frac{1}{\frac{1}{C_1}+\frac{1}{C_2}-1},$$

整理得 $\frac{1}{C_1}+\frac{1}{C_2}=\frac{1}{C_{x_1x_2}}+1$,命题得证. □

命题 2 揭示了可信关联规则的置信度与传统两个互为前后件的关联规则的置信度都有关,因此表达的可信程度更强.

3 用邻接矩阵求 2-项可信集

将每个项目(item)看作图中的点,每个顶点的权值代表包含该项目的事务个数.若某两个项目曾共同出现在若干事务中,则将它们对应的点连为边.边的权值为两项共同出现的事务个数.在此定义 2-项集邻接矩阵来表示该图.

定义 4(2-项集邻接矩阵). 设数据库中所有项目的集合为 $I=\{i_1,i_2,\ldots,i_n\}$,所有事务的集合为 $T=\{T_1,T_2,\ldots,T_m\}$.设 $A=(a_{ij})$为满足下列条件的 $n \times n$ 的矩阵:(1) 如果 T 中有且仅有 k 个事务包含项目集$\{v_i,v_j\}(i \neq j)$,则矩阵中的元素 $a_{ij}=k$;(2) 如果 T 中有且仅有 k 个事务包含项目 v_i,则矩阵中的元素 $a_{ii}=k$.称该矩阵是 2-项集邻接矩阵,记为 D_2.

2-项集邻接矩阵记录了 1-项集和 2-项集在数据库中出现的次数.其中对角线的元素 a_{ii} 记录项目 i 在事务集中出现的次数,矩阵中的元素 $a_{ij}(i \neq j)$记录项目集$\{v_i,v_j\}$在事务集中出现的次数.根据定义 3,可信关联规则 $v_i \leftrightarrow v_j$ 的置信度 $C_{v_iv_j}$ 可以通过 D_2 中的 a_{ii},a_{ij},a_{jj} 计算得到,即有: $C_{v_iv_j}=\frac{a_{ij}}{a_{ii}+a_{jj}-a_{ij}}$.

定义 5(2-项可信集邻接矩阵). 对于 2-项集邻接矩阵 D_2 中每一个非零元素 $a_{ij}(i \neq j)$,设置 $minconf$ 为二项集最小置信度,如果 $C_{v_iv_j} < minconf$,则令 $a_{ij}=0$,否则 a_{ij} 保持不变,称该矩阵是 2-项可信集邻接矩阵,记为 D_{c2}.

定义 6(2-项可信集). 2-项可信集邻接矩阵 D_{c2} 中的每一个非零元素 $a_{ij}(i \neq j)$,设置 $minsup$ 为一项集最小支持度,若 $a_{ii} \geq minsup \times |D|$ 且 $a_{jj} \geq minsup \times |D|$,则 a_{ij} 对应的项目集为$\{v_i,v_j\}$,称为 2-项可信集.

2-项可信集去除了支持度小于 $minsup$ 的项.对于每个 2-项可信集的支持度,有如下结论:

命题 3. 设 2-项可信集 v_iv_j 的支持度为 sup_2,则有:$sup_2 \geq minsup\left(\frac{2minconf}{1+minconf}\right)$.

证明:显然, $C_{v_iv_j}=\frac{a_{ij}}{a_{ii}+a_{jj}-a_{ij}}\geq minconf$,整理可得

$$a_{ij}\geq(a_{ii}+a_{jj})\left(\frac{minconf}{1+minconf}\right) \tag{2}$$

又 $a_{ij}=sup_2\times|D|$,$a_{ii}\geq minsup\times|D|$,$a_{jj}\geq minsup\times|D|$,代入式(2)可得:$sup_2\geq minsup\left(\frac{2minconf}{1+minconf}\right)$. □

命题 3 指出,虽然 2-项可信集仅对一项集作了最小支持度的约束,但通过一项集最小支持度和二项集最小可信度已隐含对 2-项集也作了约束.对于具有杂乱支持度分布的数据集,由于无法选取合适的支持度,可以不考虑支持度的影响,直接设置 *minsup* 为 0,显然,2-项可信集仍然可以求得.

算法 1 描述了求事务数据库中所有 2-项可信集的方法.

算法 1. *GetC2Matrix*().

输入:数据库 *D*,二项集最小置信度 *minconf*,一项集最小支持度 *minsup*,所有项目个数 *n*.

输出:2-项可信集的集合 CS_2.

符号:t_k,数据库的第 *k* 条事务;v_i,事务中的项;a_{ii},矩阵 D_{c2} 对角线上的元素;a_{ij},矩阵 D_{c2} 第 *i* 行 *j* 列上的元素.

步骤: (1) for *i*=1 to *n* do {for *j*=1 to *n* do a_{ij}=0; }/* Initialization D_{c2}*/

(2) for all t_k in *D* do { for $\forall v_i\in t_k$ do $a_{ii}=a_{ii}+1$;

(3) for $\forall v_i\in t_k$ and $\forall v_j\in t_k$ and $i\neq j$ do $a_{ij}=a_{ij}+1$;}

(4) for *i*=1 to *n* do {for *j*=1 to *n* do {if $a_{ij}/(a_{ii}+a_{jj}-a_{ij})<minconf$ then a_{ij}=0;}}

(5) $CS_2=\varnothing$;

(6) for *i*=1 to *n* do {

(7) for *j*=*i*+1 to *n* do

(8) if $a_{ii}\geq minsup$ and $a_{jj}\geq minsup$ and $a_{ij}>0$ then

(9) $CS_2=CS_2\cup\{v_i,v_j\}$}/*得到 2-项可信集的集合*/

2-项可信集的邻接矩阵保留了 2-项集邻接矩阵中满足置信度条件的元素,使矩阵的稀疏度提高,减少了后续的计算量.通过算法 1 不难发现,*minsup* 或 *minconf* 设置得越低,得到的 2-项可信集就越多.

例 1:设表 2 是某个数据库中的事务记录,则由表 2 产生的 2-项集邻接矩阵如表 3 所示,对应的无向图如图 2 所示.若设置二项集最小置信度 *minconf*=0.5,一项集最小支持度 *minsup*=0,根据算法 1,很多边由于不满足条件而去除,产生的 2-项可信集邻接矩阵见表 4,对应的无向图及所有 2-项可信集如图 3 所示.

Table 2 List of items in database

表 2 数据库中的项目

TID	List of item ID
001	*ABCDEFGI*
002	*ABCGHI*
003	*CEFGHIJ*
004	*CDEGI*
005	*ABCEGI*
006	*FIJ*
007	*CEGHI*
008	*DI*

Table 3 Adjacency matrix of 2-item sets

表 3 2-项集邻接矩阵

Item	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
A	3	3	3	1	2	1	3	1	3	0
B	3	3	3	1	2	1	3	1	3	0
C	3	3	6	2	5	2	6	3	6	1
D	1	1	2	3	2	1	2	0	3	0
E	2	2	5	2	5	2	5	2	5	1
F	1	1	2	1	2	3	2	1	3	2
G	3	3	6	2	5	2	6	3	6	1
H	1	1	3	0	2	1	3	3	3	1
I	3	3	6	3	5	3	6	3	8	2
J	0	0	1	0	1	2	1	1	2	2

Table 4 The adjacency matrix of 2-item credible sets, *minconf*=0.5

表 4 2-项可信集邻接矩阵,*minconf*=0.5

Item	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
A	3	3	3	0	0	0	3	0	0	0
B	3	3	3	0	0	0	3	0	0	0
C	3	3	6	0	5	0	6	3	6	0
D	0	0	0	3	0	0	0	0	0	0
E	0	0	5	0	5	0	5	0	5	0
F	0	0	0	0	0	3	0	0	0	2
G	3	3	6	0	5	0	6	3	6	0
H	0	0	3	0	0	0	3	3	0	0
I	0	0	6	0	5	0	6	0	8	0
J	0	0	0	0	0	2	0	0	0	2

通过例 1 不难发现,最小置信度越大,满足置信度条件的邻接边越少,邻接矩阵越稀疏,产生的 2-项可信集越

少.2-项可信集实际就是 2-项可信关联规则,通过算法 1 仅仅能够得到 2-项可信关联规则,若存在含有 k 个项目的集合,其中任意 2 个项目均为 2-项可信集,则这些 2-项可信集可以继续合并产生 k-项集(k≥3).

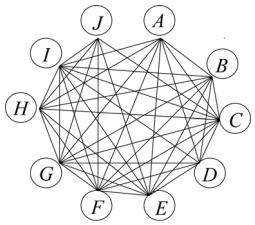
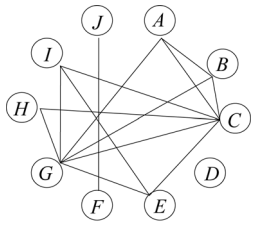


Fig.2 Undirected graph of 2-item sets

图 2 2-项集无向图



2-Item credible set:
AB,AC,AG,BC,
BG,CE,CG,CH,
CI,EG,EI,FJ,
GH,GI

Fig.3 The graph of 2-item credible set, *minconf*=0.5

图 3 2-项可信集对应的图,*minconf*=0.5

4 由 k-项可信集生成(k+1)-项可信集

在此先给出几个命题来说明由 k-项可信集生成(k+1)-项可信集的可行性.

命题 4. 设 $X=\{x_1,\ldots,x_n\}$为 n-项可信集,n≥3,则其任意 k-项子集也是可信集(2≤k≤n−1).

证明:先用反证法证明任意(n−1)-项子集是可信集.已知 X 为 n-项可信集,设 $C_{x_1\ldots x_n}\geq minconf$.

假定 X 的某 n−1 项子集$\{x_1,\ldots,x_{p-1},x_{p+1},\ldots,x_n\}$不可信(1≤$p$≤$n$),即有 $C_{x_1\ldots x_{p-1}x_{p+1}\ldots x_n}=\dfrac{\left|\bigcap\limits_{\substack{i=1\\ i\neq p}}^{n}x_i\right|}{\left|\bigcup\limits_{\substack{i=1\\ i\neq p}}^{n}x_i\right|}<minconf$.

又 $\left|\bigcap\limits_{\substack{i=1\\ i\neq p}}^{n}x_i\right|\geq\left|\bigcap\limits_{i=1}^{n}x_i\right|$, $\left|\bigcup\limits_{\substack{i=1\\ i\neq p}}^{n}x_i\right|\leq\left|\bigcup\limits_{i=1}^{n}x_i\right|$,因此有 $C_{x_1,\ldots,x_n}=\dfrac{\left|\bigcap\limits_{i=1}^{n}x_i\right|}{\left|\bigcup\limits_{i=1}^{n}x_i\right|}\leq\dfrac{\left|\bigcap\limits_{\substack{i=1\\ i\neq p}}^{n}x_i\right|}{\left|\bigcup\limits_{\substack{i=1\\ i\neq p}}^{n}x_i\right|}<minconf$,与 $C_{x_1\ldots x_n}\geq$min$conf$ 矛盾.

依此类推,可证明其任意 k-项子集也是可信集.命题得证. □

命题 4 指出,对于 n-项可信集,其任意子集均为可信集.在产生可信关联规则时,就可以用一条 n-项可信关联规则来代表所有其子集对应的关联规则,不但使关联规则得到压缩,也使关联规则更加准确.而由算法 1 得到的仅仅是 2-项可信集,利用 2-项可信集得到 3-项集,其置信度将会下降.命题 5 给出了由 2-项可信集生成 3-项集时 3-项集置信度上下界与 2-项可信集最小置信度的关系.

命题 5. 设 x_1,x_2,x_3 为项目集 I 中的 3 个项目,并满足$\{x_1,x_2\},\{x_2,x_3\},\{x_1,x_3\}\in CS_2$,设 $C_{2\min}=\min\{C_{x_1x_2},C_{x_2x_3},C_{x_1x_3}\}$,则由 x_1,x_2,x_3 构成的可信关联规则的置信度 $C_{x_1x_2x_3}$ 满足:

(1) $C_{x_1x_2x_3}\leq C_{2\min}$;

(2) $C_{x_1x_2x_3}\geq\max\{0,1.5C_{2\min}-0.5\}$.

证明:(1) 易知项目 x_1,x_2 共同出现的次数不小于 x_1,x_2,x_3 共同出现的次数,即有:

$$|x_1\cap x_2|\geq|x_1\cap x_2\cap x_3|\geq 0 \tag{3}$$

又 x_1 或 x_2 出现的次数不大于 x_1 或 x_2 或 x_3 出现的次数,即有:

$$0\leq|x_1\cup x_2|\leq|x_1\cup x_2\cup x_3| \tag{4}$$

由式(3)、式(4)有:$\dfrac{|x_1\cap x_2|}{|x_1\cup x_2|}\geq\dfrac{|x_1\cap x_2\cap x_3|}{|x_1\cup x_2\cup x_3|}$,即:$C_{x_1x_2}\geq C_{x_1x_2x_3}$.

同理可得 $C_{x_2x_3}\geq C_{x_1x_2x_3}$, $C_{x_1x_3}\geq C_{x_1x_2x_3}$,因此有 $C_{x_1x_2x_3}\leq C_{2\min}$.(1)得证.

(2) 易知:

$$C_{2\min}\le C_{x_1x_2}=\frac{|x_1\cap x_2|}{|x_1\cup x_2|}=\frac{|x_1\cap x_2\cap x_3|+|x_1\cap x_2\cap \overline{x_3}|}{|x_1\cup x_2\cup x_3|-|\overline{x_1\cup x_2}\cap x_3|}\le\frac{|x_1\cap x_2\cap x_3|+|x_1\cap x_2\cap \overline{x_3}|+|\overline{x_1\cup x_2}\cap x_3|}{|x_1\cup x_2\cup x_3|}$$

$$=\frac{|x_1\cap x_2\cap x_3|+|x_1\cap x_2\cap \overline{x_3}|+|\overline{x_1}\cap\overline{x_2}\cap x_3|}{|x_1\cup x_2\cup x_3|} \tag{5}$$

同理,

$$C_{2\min}\le C_{x_2x_3}\le\frac{|x_1\cap x_2\cap x_3|+|\overline{x_1}\cap x_2\cap x_3|+|x_1\cap\overline{x_2}\cap\overline{x_3}|}{|x_1\cup x_2\cup x_3|} \tag{6}$$

$$C_{2\min}\le C_{x_1x_3}\le\frac{|x_1\cap x_2\cap x_3|+|x_1\cap\overline{x_2}\cap x_3|+|\overline{x_1}\cap x_2\cap\overline{x_3}|}{|x_1\cup x_2\cup x_3|} \tag{7}$$

将式(5)~式(7)左右分别相加:

$$3\,C_{2\min}\le\frac{3|x_1\cap x_2\cap x_3|+|\overline{x_1}\cap x_2\cap x_3|+|x_1\cap\overline{x_2}\cap x_3|+|x_1\cap x_2\cap\overline{x_3}|+|\overline{x_1}\cap\overline{x_2}\cap x_3|+|x_1\cap\overline{x_2}\cap\overline{x_3}|+|\overline{x_1}\cap x_2\cap\overline{x_3}|}{|x_1\cup x_2\cup x_3|}$$

$$=\frac{2|x_1\cap x_2\cap x_3|+|x_1\cup x_2\cup x_3|}{|x_1\cup x_2\cup x_3|}=2\,C_{x_1x_2x_3}+1.$$

即有: $C_{x_1x_2x_3}\ge 1.5C_{2\min}-0.5$.又 $C_{x_1x_2x_3}\ge 0$,因此 $C_{x_1x_2x_3}\ge\max\{0,1.5C_{2\min}-0.5\}$.(2)得证. □

命题 5 给出了可信 3-项关联规则的置信度上下界与其 2-项子集最小置信度的关系,如图 4 所示.从图中也可以看出,当 2-项子集最小置信度较大时,可信 3-项关联规则的置信度的范围在实际应用中也可被用户接受.例如,若 *minconf*=0.8,根据算法 1,易知所有 $C_{2\min}$ 均大于等于 *minconf*,则产生的 3-项集置信度 C_3 的最小值为 0.7.

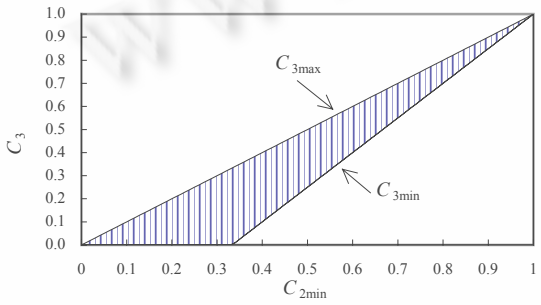


Fig.4 The relationship between the range of C_3 and $C_{2\min}$

图 4 C_3 的取值范围与 $C_{2\min}$ 的关系

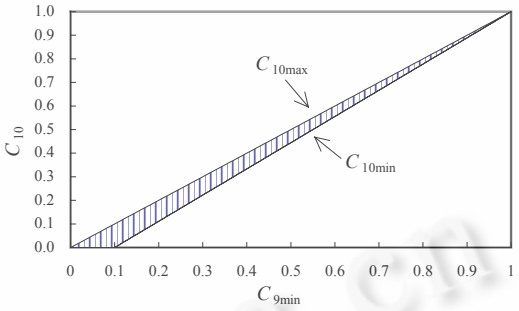


Fig.5 The relationship between the range of C_{10} and $C_{9\min}$

图 5 C_{10} 的取值范围与 $C_{9\min}$ 的关系

下面的推论给出了由 *k*-项可信集生成(*k*+1)-项集时(*k*+1)-项集置信度的取值上下界与 *k*-项可信集最小置信度的关系.

命题 5 推论. 设 $x_1,x_2,\ldots,x_k,x_{k+1}$ 为项目集 *I* 中的 *k*+1 个项目,并满足 $\{x_2,\ldots,x_{k+1}\},\{x_1,x_3,\ldots,x_{k+1}\},\ldots,\{x_1,\ldots,x_{k-1},x_{k+1}\},\{x_1,\ldots,x_k\}\in CS_k$,设 $C_{k\min}=\min\{C_{x_2\ldots x_{k+1}},C_{x_1x_3\ldots x_{k+1}},\ldots,C_{x_1\ldots x_{k-1}x_{k+1}},C_{x_1\ldots x_k}\}$,则 $C_{x_1\ldots x_{k+1}}$ 满足:

(1) $C_{x_1\ldots x_{k+1}}\le C_{k\min}$;

(2) $C_{x_1\ldots x_{k+1}}\ge\max\left\{0,\frac{k+1}{k}C_{k\min}-\frac{1}{k}\right\}$;

(3) $C_{x_1\ldots x_{k+1}}\ge\max\left\{0,1-\frac{k+1}{2}(1-C_{2\min})\right\}$.

证明:用类似命题 5 的方法容易证明(1)和(2),(3)可由(2)递推得到,证明在此略. □

命题 5 推论指出,虽然(k+1)-项集置信度不大于最小 k-项集置信度,但其取值下界为 k-项集最小置信度的一次函数,并且随着 k 的增大,取值下界越来越趋近于 k-项集的置信度,图 5 给出了 10-项集置信度的取值范围与 9-项集最小置信度的关系.这样,如果(k+1)-项可信集的所有 k-项子集的置信度都满足条件,我们可以近似认为该(k+1)-项可信集的置信度也满足条件,实际上,置信度变小.在实际应用中,开始时的二项集最小置信度 *minconf* 可以设置得稍大一些.

通过以上分析可知,由 k-项可信集生成(k+1)-项可信集是可行的,为了讨论其生产方法,本文引入有序序列来表示可信集,下面给出一系列定义.

定义 7(有序可信集). 对于可信集 $X=\{x_1,x_2,\ldots,x_n\}$,如果对于$\forall l,k$ 且 $1\le l<k\le n$,有 $x_l<x_k$ 成立,则称可信集 X 为有序可信集.

本文约定用有序串 $x_1x_2...x_n$ 表示 X.

不难发现,算法 1 产生的 2-项可信集为有序可信集.本文所讨论的所有可信集均为有序可信集,因此不再特殊指明.

例 2:若以字母的 ASCII 作为其值,则可信 4-项集 *ABCD* 为有序可信集,而 *ACBD* 不是有序可信集.

定义 8($\prec$关系). 设 $X_1=\{x_{11},\ldots,x_{1n}\}$,$X_2=\{x_{21},\ldots,x_{2n}\}$均为 n-项有序可信集,二元关系 $\prec$ 称为 $\prec$ 关系,规定为 $X_1\prec X_2$,当且仅当$\exists k\in\{1,\ldots,n\}$,对于$\forall i\in\{1,\ldots,k-1\}$使得 $x_{1i}=x_{2i}$ 成立,并且 $x_{1k}<x_{2k}$. $\prec$ 关系为偏序关系.

例 3:设 *ABCD*,*ABCE* 为有序可信集,则满足关系 $ABCD\prec ABCE$.

定义 9(n-项可信集的有序集合). 对于由 n-项可信集 $X_i=\{x_{i1},\ldots,x_{in}\}$构成的集合 $CS_n(1\le i\le|CS_n|)$,如果对于$\forall l,k$ 且 $1\le l<k\le|CS_n|$,有 $X_1\prec X_k$ 成立,则称集合 CS_n 为 n-项可信集的有序集合.

同样,算法 1 产生的 2-项可信集的集合为有序集合.本文所讨论的所有可信集的集合也均为有序集合.

例 4:设可信 4-项集的集合 $CS_4=\{ABCD,ABCE,ABDE,ACDE,BCDE\}$,则 CS_4 为有序集合.因为 $ABCD\prec ABCE\prec ABDE\prec ACDE\prec BCDE$.

通过以上讨论得知,CS_k 为有序集合,因此可以方便地进行顺序读取和查找集合中的元素.由此,我们可以采用如下思想产生(k+1)-项集:首先由 k-项可信集的集合 CS_k 中顺序抽取 2 个 k-项集.若它们前 k–1 项都相同,并且这 2 个 k-项集的最后一项构成的 2-项集在 CS_2 中存在,则将前 k–1 项和 2 个 k-项集的最后一项构成候选(k+1)-项集,然后判别该候选集的其他 k-项子集是否都可信,若可信,则产生该(k+1)-项集,并将其所有 k-项子集进行标记.全部抽取完毕,若 k-项可信集的集合中仍存在没有标记的项集,则它们为 k-项关联规则,不能再进一步进行合并.

算法 2 给出了由 k-项可信集生成(k+1)-项可信集的方法.算法中的 $Sub_k(temp)$函数用于求得(k+1)-项集 *temp* 的所有的 k-项子集构成的集合,例如,设 5-项集 $temp=ABCDE$,则 $Sub_4(temp)$得到的 *temp* 所有 4-项子集构成的集合为$\{ABCD,ABCE,ABDE,ACDE,BCDE\}$.算法中的 *head* 和 *end* 操作定义如下:对于 k-项集 X,$X.head$ 表示 X 的前 k–1 项,$X.end$ 表示 X 的最后一项.如 $X=ABCDE$,则 $X.head=ABCD$,$X.end=E$.

算法 2. GetCS$_{k+1}$Set().

输入:k-项可信集的有序集合 CS_k,2-项可信集的有序集合 CS_2.

输出:k 项可信关联规则集 R_k,(k+1)-项可信集的有序集合 CS_{k+1}.

符号:$CS_k(j)$,k-项可信集的集合 CS_k 中第 j 个 k-项集;$Sub_k(temp)$,(k+1)-项集 *temp* 的所有的 k-项子集构成的集合;$X.head$,k-项集 X 的前 k–1 项;$X.end$,k-项集 X 的最后一项.

步骤: (1) $R_k=CS_k$;$CS_{k+1}=\varnothing$;

(2) for i=1 to $|CS_k|$ do {

(3) for $j=i$+1 to $|CS_k|$ do {

(4) if $CS_k(i).head=CS_k(j).head$ and $\{CS_k(i).end,CS_k(j).end\}\in CS_2$ then {

(5) $temp=\{CS_k(i),CS_k(j).end\}$; /*构造$k$+1项集*/

(6) tag=0; /*设置标记为0,若tag一直为0,表示$temp$的所有k项子集均存在*/

(7) if k≥3 then{/*若k≥3,还需要判别k+1项集除$CS_k(i)$和$CS_k(j)$的其他k项子集是否存在*/

(8) for all t in $Sub_k(temp)\backslash\{CS_k(i),CS_k(j)\}$ do

(9) if $t\notin CS_k$ then {tag=1;break;}} /*若某k项子集不存在,则查找子集终止*/

(10) if tag=0 then {/*若标记未修改,说明所有k项子集都存在*/

(11) $CS_{k+1}=CS_{k+1}\cup\{temp\}$; /*将新构造$k$+1项集$temp$加入到集合中*/

(12) $R_k=R_k\backslash Sub_k(temp)$;/*在$R_k$中去除$temp$所有的$k$项子集*/}}

(13) else break;}}

(14) output R_k;

(15) output CS_{k+1};

例 5:针对例 1 中数据,设置 $minconf$=0.5,由算法 1 得到 2-项可信集的集合 CS_2,根据算法 2,可以得到 R_2 和 CS_3,结果见表 5,在表中也给出了得到的每个 3-项集的置信度.

Table 5 2-item credible association rules and all 3-item sets by algorithm 2

表5 算法2得到的2-项关联规则及3-项集

CS_2	R_2	CS_3	CS_3 confidence
AB	FJ	ABC	0.5
AC		ABG	0.5
AG		ACG	0.5
BC		BCG	0.5
BG		CEG	0.8
CE		CEI	0.57
CG		CGH	0.5
CH		CGI	0.75
CI		EGI	0.57
EG			
EI			
FJ			
GH			
GI			

5 基于极大团的可信关联规则挖掘算法 MaxCliqueMining

5.1 算法描述

分析算法 2 不难发现,产生(k+1)-项集的条件是它的所有 k-项子集都存在.若某个 k-项子集不存在,则该(k+1)-项集不能产生,也就不能再继续聚合成更高项集.这个过程实际上是一个发现所有极大团[9]的过程.

定义 10(极大团). 对于图 $G=(V,E)$,$\exists V'\subseteq V$,如果给定点集 V'导出的子图 $G'=(V',E')$是完全图,则称 G'为图 G 中的团;如果 $\neg\exists v\in V$且$v\notin V'$,使得点集 $V'\cup\{v\}$导出的子图是完全图,则称 G'为图中的极大团.

通过引入团和极大团的概念,可知算法 2 实际上是在已知所有 k 项点团的条件下,计算所有的 k+1 项点团.将算法 1 和算法 2 结合,即得到基于极大团的可信关联规则挖掘算法 MaxCliqueMining,如算法 3 描述.

算法 3. MaxCliqueMining().

输入:数据库 D,二项集最小置信度 $minconf$,一项集最小支持度 $minsup$.

输出:所有的可信关联规则集.

步骤:(1) 使用算法 1 产生 2-项可信集集合 CS_2;

(2) k=2

(3) while($CS_k\neq\varnothing$){

(4) 使用算法 2 由 k-项集集合 CS_k 生成(k+1)-项可信集集合 CS_{k+1} 和 k 项关联规则集 R_k;

(5) $k=k$+1;}

(6) 返回$\cup R_i$;

MaxCliqueMining 算法仅在产生邻接矩阵时通过算法 1 扫描一遍数据库,因此提高了挖掘的时间性能.

例 6:针对例 1 中数据,设置 *minconf*=0.5,*minsup*=0.根据算法 3,首先得到 2-项可信集的集合 CS_2,然后循环可顺序得到 R_2,CS_3,R_3,CS_4,R_4 及 CS_5,结果见表 6.最后得到的可信关联规则为 $R_2 \cup R_3 \cup R_4$,表 7 统计了每条关联规则的支持度和置信度.显然,这些规则的支持度差别很大,而所有的关联规则置信度都不小于 *minconf*.

Table 6 The mining results of algorithm 3

表6 算法3的挖掘结果

CS_2	R_2	CS_3	R_3	CS_4	R_4	CS_5
AB	*FJ*	*ABC*	*CGH*	*ABCG*	*ABCG*	∅
AC		*ABG*		*CEGI*	*CEGI*	
AG		*ACG*				
BC		*BCG*				
BG		*CEG*				
CE		*CEI*				
CG		*CGH*				
CH		*CGI*				
CI		*EGI*				
EG						
EI						
FJ						
GH						
GI						

Table 7 The confidences of rules

表 7 关联规则置信度

CARs	Support	Confidence
FJ	0.25	0.67
CGH	0.375	0.50
ABCG	0.375	0.50
CEGI	0.625	0.625

分析例 6 的结果,不难发现产生的关联规则集即为将所有 2-项集为边的图分解为极大团的结果,如图 6 所示,分解后的每个子图均为原图的一个极大团,且原图中所有的极大团均为算法 3 产生的结果集,下面将给出相关命题.

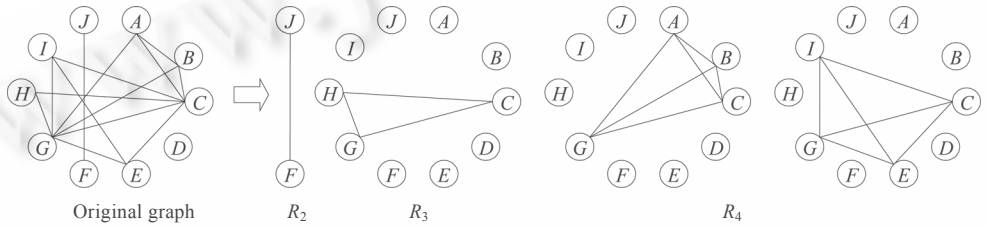


Fig.6 The relationship between mining results and maximum cliques

图 6 挖掘结果与极大团的关系

命题 6. 给定任意图 $G=(V,E)$,若将 E 看作 2-项可信集,则图 G 包含的所有极大团与算法 3 产生的所有可信关联规则是一一对应的.

证明:显然,G 中包含的每个极大团都是唯一的,所有极大团构成的集合也是唯一的.设图 G 中存在$(k+1)$顶点极大团 S,则从 k+1 个顶点任取 k 顶点构成的团必为 S 所包含,算法 2 中,由步骤(6)~步骤(11)可知,S 必然包含在 CS_{k+1} 中,又 S 为极大团,不可能存在 k+2 顶点团 S'且 $S' \supset S$,即算法 2 的步骤(12)不可能将 S 从 R_{k+1} 中去除,因此 R_{k+1} 中必然有 S;另一方面,若 R_{k+1} 中含有可信关联规则 S,由算法 2 的步骤(11)、步骤(12)可知,S 是团,并且是极大团.命题得证. □

命题 6 揭示了算法 3 挖掘结果的本质,同时也给我们一点启示,即可选择其他优化方法求解极大团集合作为可信关联规则集,以提高挖掘效率.

5.2 算法性能分析

对于算法 1,仅需要扫描数据库一次.在生成 2-项可信集邻接矩阵及 2-项可信集时,需对对称矩阵进行遍历.设 n 为数据库中所有事务的数目,m 为所有项目的数目,则算法 1 的时间复杂度为 $O(n+m^2/2)$.

对于算法 2,在生成 k+1 项集时,抽取的每个 k 项集都要查询每个后续项集的前 k–1 项是否与自己的前 k–1 项相同.设 k 项集共有 p 个,前 k–1 项相同的 k 项集平均有 q 个,则时间复杂度为 $O(pq)$,$q<<p$.另外,产生 k+1 项集后需要查找其他 k 项子集是否存在,采用优化的查找算法时间复杂度可控制在 $O(1)$~$O(\log_2 p)$之间,比较前者此

项可忽略.因此,算法 2 的最优时间复杂度近似可认为 $O(pq)$.

对于算法 3,考虑实际应用中一条关联规则包含的项数一般不会很多,因此算法中循环次数也不会很多.但是,每次循环产生的 k+1 项集的数量 p 却变化很大.例如,设原始图存在一个 n 顶点极大团,仅产生该团包含的 2-项集到 n-项集,则每个项集的数量分别为 $C_n^2,C_n^3,...,C_n^{\lfloor \frac{n}{2} \rfloor},...,C_n^n$,不难发现项集数最大的为 $\left\lfloor \frac{n}{2} \right\rfloor$-项集,项集数量为 $C_n^{\lfloor \frac{n}{2} \rfloor}$.考虑实际应用中关联规则包含项数不会很大,算法 3 是适用的.对于某些特殊应用,如果关联规则包含的项目数较大,则需要对算法 3 进行优化,命题 6 的结论指出了挖掘可信关联规则的本质,因此可采用优化方法来挖掘极大团集合,这里不再阐述.

6 实验结果分析

6.1 数据集及实验环境描述

实验中分别采用网管告警数据集[6]和 Pumsb 数据集[8,10]进行测试.

网管告警数据集是某省电信公司 GPRS 网管系统告警数据库中连续两周的原始告警数据(15 万条记录).每条告警包含的信息主要有告警发生时间(event time)、发生告警的设备标识(element ID)以及告警内容(event title)等信息.表 8 给出了部分告警示例.

Table 8 The samples of original alarms

表 8 原始告警示例

Event time	Element ID	Event title
2006-2-6 0:00:00.000	1033478163	NM ROUTE ASR SUPERVISION
2006-2-6 0:00:26.000	1552978014	Scf Free
2006-2-6 0:01:44.000	985664880	Standby link connection failure—FMIC
2006-2-6 0:02:22.000	384783557	Message ID of speech file not exist
2006-2-6 0:02:32.000	1492754060	Tx VSWR antenna fault—FMIC
2006-2-6 0:03:19.000	1249452812	Optocoupler 5—FMIC
2006-2-6 0:03:31.000	2128863748	Database configuration and hardware mismatch—FMIC
…	…	…

在进行挖掘之前,首先将每条告警记录中发生告警的设备标识和告警内容组合成唯一的告警标识号,用该标识号来唯一地标识每一个告警,这样,原始告警记录就可转换为只含有告警发生时间和告警标识号的数据.将每个告警认为是一个项目,则一个时间窗 T_w 内的所有告警认为是一个事务的所有项目[5],每个时间窗与前一个时间窗的起始时间间隔称为滑动步长 S.这样,整个告警数据库可转换为包含若干个事务的数据集.实验设置告警时间窗 T_w 设为 3min,滑动步长 S 设为 1.5min.转换之后数据集共包含 13 440 个事务和 1 301 个项,其支持度分布在图 1 和表 1 中已有描述.

Pumsb 数据集是一个人口普查数据集,由 49 046 个事务和 2 113 个项组成,每个事务的长度为 74.该数据集的支持度也具有杂乱分布,见表 9.

Table 9 Groups of items for Pumsb dataset with different support range

表9 按照支持度分布将Pumsb数据项分组

Group	G_1	G_2	G_3	G_4	G_5
Support (%)	<0.1	0.1~1	1~10	10~50	>50
Items number	1099	636	243	83	52
Ratio (%)	52.0	30.1	11.5	3.9	2.5

实验环境如下:测试机器 CPU 为 P4 2.80GHz,内存为 1GB,使用 Windows XP 操作系统.设置一项集最小支持度 $minsup$=0,分别设置二项集最小置信度 $minconf$ 为 0.5,0.6,0.7,0.8,0.9,对各个数据集进行测试.

6.2 MaxCliqueMining算法挖掘结果分析

图 7 针对不同的 *minconf* 给出了 *k*-项可信关联规则集 R_k 与 *k* 的关系.对于每个数据集,*k* 越大,相应的关联规则数越少.*minconf* 越小,则最长规则的长度越大.在 *minconf*=0.5 时两个数据集得到的最长规则长度分别为 10 和 34.实验结果也说明,在两种数据集中的确存在着大量的可信关联规则.

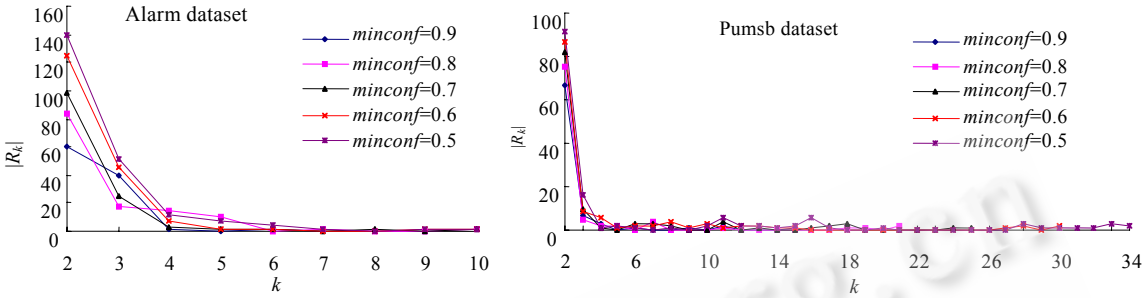


Fig.7 Number of *k*-item association rules

图 7 *k*-项可信关联规则的数量统计

将不同 *minconf* 取值下的挖掘结果进行支持度和置信度统计,见表 10.从表中可以看出,挖掘结果在不同支持度级别下均可挖掘出可信关联规则,从而验证了 MaxCliqueMining 算法在完全忽略传统支持度的条件下仍然可以得到合理的挖掘结果,避免了由于无法选择合适的支持度而产生过少或虚假的规则.另一方面,分析挖掘结果中各个关联规则的置信度,绝大多数都大于二项集最小置信度 *minconf*.置信度小于 *minconf* 的规则并不多,且这些规则的置信度与 *minconf* 相差得也并不大,这一点在命题 5 及其推论中已经指出,因此挖掘结果可以被接受.例如,在 *minconf*=0.8 时,置信度小于 *minconf* 的规则对于告警数据有 4 条,对于 Pumsb 数据有 9 条,分别占所有规则的 3.8%和 9.7%.

Table 10 Statistics of support and confidence of the rules under different *minconf*

表 10 不同 *minconf* 取值下所有规则的支持度和置信度统计

Minconf	Alarm dataset					Pumsb dataset				
	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
Total number of all rules	220	183	131	106	74	147	127	114	93	82
Number of rules with *support*≤0.1%	64	49	34	28	24	61	50	46	44	41
Number of rules with *support*∈(0.1%,1%]	139	120	84	67	39	20	20	17	10	7
Number of rules with *support*∈(1%,10%]	14	11	11	9	10	21	18	16	14	11
Number of rules with *support*>10%	3	3	2	2	1	45	39	35	25	23
Number of rules with *confidence*<*minconf*	24	11	5	4	0	39	26	20	9	6
Minimum confidence among all rules	0.35	0.44	0.52	0.6	0.9	0.1	0.29	0.43	0.56	0.71

6.3 MaxCliqueMining算法与其他算法的比较

针对 Pumsb 数据集,将 MaxCliqueMining 算法与 *h*-confidence 方法[8]和相关统计算法[6]在挖掘出的规则数及耗时上进行比较.实验中 3 种方法的 *minsup* 均设置为 0.

3 种算法都是针对强亲密度关联规则进行挖掘的,图 8 比较了各种算法在不同 *minconf* 时挖掘的规则数目.*h*-confidence 方法得到的规则数较多,这是因为它对规则亲密度的限定能力过弱,结果中会含有虚假规则.而相关统计算法对规则亲密度的限定能力过强,结果中丢失了很多有趣的规则.MaxCliqueMining 算法得到的规则数介于两者之间.另外,将各种算法产生的规则进行比较发现,MaxCliqueMining 算法产生的规则基本上包含了 *h*-confidence 方法和相关统计算法产生规则的交集.这也验证了该算法具有较高的准确性.

图 9 给出了各种算法的耗时比较.MaxCliqueMining 算法平均时间开销最小,这是由于算法产生 *k*+1 项集时不需要再扫描数据库,之后也不再作任何兴趣度判别,因此提升了处理速度.相关统计算法由于要计算任意两

项目之间的相关度,耗时稍大.而 *h*-confidence 方法是基于 Apriori 算法的,不但需要多次扫描数据库,在产生候选项集后还需要计算每个项集的 *h*-confidence 值,因此耗时最大.随着 *minconf* 的减小,每种算法的时间开销都有所增加.

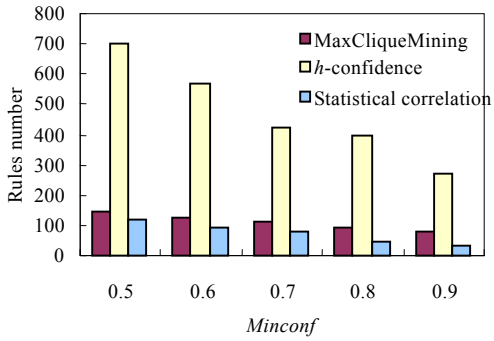


Fig.8 Rules number comparison

图 8 规则数比较

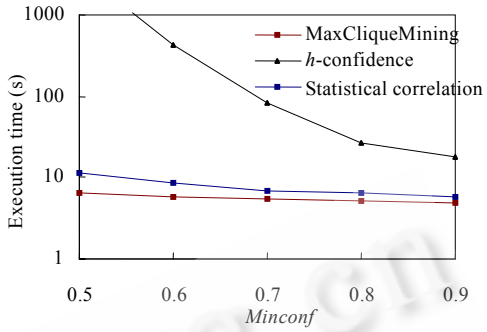


Fig.9 Execution time comparison

图 9 时间开销比较

7 结束语

针对具有杂乱支持度分布的数据集,用户无法选取合适的支持度进行关联规则挖掘,本文引入一种新的关联模式——可信关联规则,规则中每个项目的支持度处于同一数量级别,规则的置信度直接反映其可信程度,从而不再考虑传统的支持度.针对可信关联规则的挖掘,本文提出了 MaxCliqueMining 算法.该算法采用邻接矩阵产生 2-项可信集,进而利用极大团思想产生所有可信关联规则,而不需要对数据库进行多次扫描,从而使时间性能得以提高.文中还提出并证明几个相关命题来说明这种规则的特点及算法的可行性和有效性.实验结果表明,MaxCliqueMining 算法在杂乱支持度分布的数据集中挖掘可信关联规则具有较高的效率和准确性.

后续研究将围绕以下工作展开:(1) 进一步分析可能存在可信关联规则的各种数据集的特点;(2) 探讨可信关联规则挖掘算法的优化.

References:

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S, eds. Proc. of the ACM SIGMOD Conf. on Management of Data (SIGMOD'93). New York: ACM Press, 1993. 207−216.

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Bocca JB, Jarke M, Zaniolo C, eds. Proc. of the 20th Int'l Conf. on Very Large Data Bases. Santigo: Morgan Kaufman Publishers, 1994. 478−499.

[3] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Chen WD, Naughton J, Bernstein PA, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2000). New York: ACM Press, 2000. 1−12.

[4] Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining—A general survey and comparison. SigKDD Explorations, 2000,2(1):58−64,

[5] Mannila H, Toivonen H, Verkamo A. Discovery of frequent episodes in event sequences. In: Katharina M, Hanna K eds. Data Mining and Knowledge Discovery 1. Netherlands: Kluwer Academic Publishers, 1997. 259−289.

[6] Xu QF, Xiao B, Guo J. A mining algorithm with alarm association rules based on statistical correlation. Journal of Beijing University of Posts and Telecommunications, 2007,30(1):66−70 (in Chinese with English abstract).

[7] Omiecinski ER. Alternative interest measures for mining associations in databases. IEEE Trans. on Knowledge and Data Engineering (TKDE), 2003,15(1):57−69.

[8] Xiong H, Tan PN, Kumar V. Mining strong affinity association patterns in data sets with skewed support distribution. In: Wu XD, Tuzhilin A, Shavlik J, eds. Proc. of the ICDM 2003. Melbourne: IEEE Computer Society, 2003. 387−394.

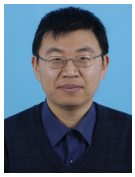
[9] Chen AL, Tang CJ, Tao HC, Yuan CA, Xie FJ. An improved algorithm based on maximum clique and FP-tree for mining association rules. Journal of Software, 2004,15(8):1198−1207 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1198.htm

[10] Pumsb dataset. http://fimi.cs.helsinki.fi/data/

附中文参考文献:

[6] 徐前方,肖波,郭军.一种基于相关度统计的告警关联规则挖掘算法.北京邮电大学学报,2007,30(1):66−70.

[9] 陈安龙,唐常杰,陶宏才,元昌安,谢方军.基于极大团和 FP-Tree 的挖掘关联规则的改进算法.软件学报,2004,15(8):1198−1207. http://www.jos.org.cn/1000-9825/15/1198.htm



肖波(1975－),男,山东高密人,博士生,讲师,主要研究领域为数据挖掘.

郭军(1959－),男,博士,教授,博士生导师,主要研究领域为模式识别,网络搜索,数据挖掘.

徐前方(1975－),女,博士,讲师,主要研究领域为数据挖掘,网络管理.

李春光(1979－),男,博士,讲师,主要研究领域为数据挖掘,模式识别.

蔺志青(1959－),女,教授,主要研究领域为智能信息处理,计算机网络及应用.