关联规则挖掘的软集包含度方法

耿生玲[1,2],李永明[1],刘 震[3]
(1.陕西师范大学计算机科学学院,陕西西安 710069;2.青海师范大学计算机学院,青海西宁 810008;
3.日本长崎综合科技大学,日本长崎 851-0193)

摘 要: 本文在深入研究软集数据分析的基础上,将包含度引入软集数据关联规则挖掘中,利用包含度理论描述属性集之间的量化关系,给出软集上属性集间的包含度、关联规则和最大关联规则的概念,讨论包含度和可信度之间的联系.在此基础上给出利用包含度在事务数据软集中挖掘满足给定的支持度和可信度阈值的软关联规则方法,以及最大软关联规则的提取算法.理论证明和实例分析表明该关联规则挖掘方法是有效的,并通过实验对算法的性能进行了比较.

关键词: 软集;包含度;关联规则;软最大关联规则.
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2013) 04-0804-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.04.030

An Approach to Association Rules Mining Using Inclusion Degree of Soft Sets

GENG Sheng-ling[1,2], LI Yong-ming[1], LIU Zhen[3]
(1. *College of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi* 710069, *China*;
2. *School of Computer Science, Qinghai Normal University, Xining, Qinghai* 810008, *China*;
3. *Nagasaki Institute of Applied Science, Nagasaki* 851-0193, *Japan*)

Abstract: This paper aims to present an approach for mining regular association rules and maximal association rules using soft set and inclusion degree theory from transactional datasets. We first give the notions of inclusion degree, association rule and maximum association rules between attribute sets of soft set. Then we discuss the relationship between inclusion degree and confidence. Furthermore, we give an algorithm of soft maximal association rules mining using inclusion degree of soft set. The experiments show the algorithm improves greatly the performance of maximal association rules mining.

Key words: soft sets; inclusion degree; association rules; soft maximal association rules

1 引言

为了解决经济学、工程学、环境学和社会科学等各门学科中遇到的不确定的问题,相继产生了一些数学理论和工具,诸如概率论、模糊集理论、粗糙集等等.但是所有这些理论和结果都有其不完善的方面[1],其中一个主要的原因在于参数化工具的不足,由此 Molodtsov 提出了软集[1]的概念,软集参数的无约束性,使解决不确定性问题的应用更广,也简化了决策过程.P K Maji 等在文献[2]中给出详细的软集理论,之后又在文献[3]中给出软集在决策中的应用.P K Maji 等在文献[4]中给出模糊软集的相关定义和结论.F Feng 等在文献[5,6]给出了软集、粗集和模糊集进行比较,结合定义软粗集、粗软集和模糊粗软集等的相关概念,并指出粗集是软集的一种特殊情况.在文献[7~9]中给出软集及模糊软集等在决策、属性约简等方面的应用.研究表明软集理论在决策分析、模式识别和数据挖掘等领域具有很大的应用潜力.

包含度是一种描述不确定性关系的有效的度量方法[10,11],是对已有的不确定性推理方法(如概率推理方法、证据推理方法、模糊推理方法以及信息推理方法等)的概括,因而为不确定性推理提供了一个一般性原理.同时,它还便于进行信息的合成、传播和修正.特别地,在各种关系数据库中有着直接的应用.此外,包含度理论的研究也为有序结构的数学理论提供一种定量分析方法,在人工智能、专家系统和模糊集理论等领

收稿日期:2012-05-10;修回日期:2012-12-11
基金项目:973计划前期研究专项基金(No.2011CB311809);国家自然科学基金(No.11271237,No.6126104);教育部"春晖计划"合作科研项目基金(No.Z2011026);青海省应用基础研究基金(No.2011-Z-749)

域[12~14]有着重要的应用.

关联规则[14](Association Rules)是由 R Agrawal 等人提出的,是当前数据挖掘研究的主要模式之一,侧重于确定数据中不同领域之间的联系,找出满足给定支持度和可信度阈值的多个域之间的依赖关系.根据所挖掘的关联关系,可以从一个数据对象的信息来推断另一个数据对象的信息.关联规则挖掘的研究一直是数据挖掘领域的热点问题,近年来,文献[15,16]等提出了一些有效的关联规则挖掘方法.这些方法较多是针对事务数据库中的布尔值类型的关联规则的挖掘,对于数量型关联规则的挖掘问题主要也是转换为布尔型关联规则的挖掘问题来处理.在这类事务数据库中属性的值可以考虑为一个布尔值,即一个项目包含在一个事务中其值视为"1",否则视为"0".这样一个事务数据库就可以转换为一个布尔值信息系统 $S=(U,A,V_{\{0,1\}},f)$.

本文将包含度引入软集数据分析中,在软集的属性集之间定义了包含度,利用包含度理论描述了属性集之间的量化关系.同时,研究包含度和可信度之间的联系,利用包含度在事务数据软集中挖掘满足给定的支持度阈值和可信度阈值的关联规则的方法,给出了由软集包含度进行最大关联规则的提取算法,理论证明和实例验证了该算法是有效的.

2　基本概念

2.1　软集

定义 1[1]　设 U 是非空有限的对象集合,E 是参数集合,$P(U)$表示 U 的幂集,$A\subseteq E$,设 $F:A\to P(U)$为一个映射,则称$\partial=(F,A)$为 U 上的一个软集.

定义 2[2]　如果(F,A)和(G,B)是 U 上的两个软集,如果 $B\subseteq A$ 且对于$b\in B$,$G(b)\subseteq F(b)$,则称(G,B)是(F,A)的软子集,表示为$(G,B)\subseteq(F,A)$.

定义 3[2]　在 U 上的软集(F,A)中,$\forall\, a,b\in A$,存在 U 上的软集$(F_\neg,A)$,$(F_\wedge,A\times A)$,$(F_\vee,A\times A)$,定义:

(1) $F_\neg(a)=U-F(a)$

(2) $F_\wedge(a,b)=F(a)\cap F(b)$

(3) $F_\vee(a,b)=F(a)\cup F(b)$

假设 $I=\{i_1,i_2,\cdots,i_m\}$是项目集合,$T=\{t_1,t_2,\cdots,t_{|U|}\}$是一个事务数据库,其中 t_i 表示 T 的第 i 个事务,它是 I 中的一组布尔型项目的集合,$t_i\subseteq I$.这个事务数据库可以看成是一个布尔值的信息系统 $S=(U,A,V_{\{0,1\}},f)$.因此,一个事务数据库也可表示成一个软集 $S=(F,A)$.即

$$\left.\begin{array}{l} i_1\to a_1\\ i_2\to a_2\\ \cdots\\ i_m\to a_m \end{array}\right\}\Leftrightarrow I=\{i_1,i_2,\cdots,i_m\}\to A=\{a_1,a_2,\cdots,a_m\}$$

$$\left.\begin{array}{l} t_1\to u_1\\ t_2\to u_2\\ \cdots\\ t_{|U|}\to u_{|U|} \end{array}\right\}\Leftrightarrow T=\{t_1,t_2,\cdots,t_{|U|}\}\to U=\{u_1,u_2,\cdots,u_{|U|}\}.$$

2.2　包含度

定义 4[10]　设$(L,\leq)$为一偏序集.若对于任意的 $x,y\in L$,有实数 $D(y/x)$与之对应,且满足

(1)$0\leq D(y/x)\leq 1$

(2)若 $x\leq y$,$D(y/x)=1$;

(3)若 $x\leq y\leq z$,$D(x/z)\leq D(x/y)$;

(4)若 $x\leq y$,对任意的 $z\in L$,有 $D(x/z)\leq D(y/z)$,

则 D 称为偏序集$(L,\leq)$上的包含度.

事实上,包含度是对偏序关系的一种度量.上述定义中,(1)是对包含度的规范化;(2)表示包含度与经典包含的协调性,经典包含关系是包含度为 1 的特殊情况;(3)与(4)是包含度的单调性.

2.3　关联规则

关联规则[14]指的是一个形如 $A\Rightarrow B$ 的表达式,其中 A 和 B 是属性的集合,其直观含义是:软集中具有属性集 A 中的对象也可能具有属性集 B 中的属性.

定义 5　[17]　设(F,A)是 U 上的一个软集,$u\in U$.对象 u 的共现集定义为:

$$Co(u)=\{e\in A:f(u,a)=1\}\tag{1}$$

显然,$Co(u)=\{e\in A:f(e)=1\}$.

对于 $X\subseteq A$,$u\in U$,如果 $X\subseteq Co(u)$,那么可以说 u 支持属性集 X.

定义 6　[17]　(F,A)是 U 上的一个软集,$X\subseteq A$.属性集 X 的支持度 $sup(X)$定义为:

$$sup(X)=\frac{|\{u:X\subseteq Co(u)\}|}{|U|},\tag{2}$$

其中$|X|$是 X 的基数.即 $sup(X)$是在 U 中支持 X 的对象数.

对于(F,A),假设 $X,Y\subseteq A$,并且 $X\cap Y=\varnothing$,可以得到关联规则 $X\Rightarrow Y$,其中 X 和 Y 分别称为前项和后项.其支持度是:

$$sup(X\Rightarrow Y)=sup(X\cup Y)=\frac{|\{u:X\cup Y\subseteq Co(u)\}|}{|U|}.\tag{3}$$

可信度是:

$$conf(X\Rightarrow Y)=\frac{sup(X\cup Y)}{sup(X)}=\frac{|\{u:X\cup Y\subseteq Co(u)\}|}{|\{u:X\subseteq Co(u)\}|}. \tag{4}$$

在实际应用中,人们更多的是关心支持度和可信度足够大的规则.因此,可以设置一个支持度的阈值 α 和一个可信度的阈值β,这样人们可以只关心支持度大于 α 的节点和可信度大于β 的软集元素集.本文讨论可信度不为 1 的关联规则,即满足给定的支持度阈值和可信度阈值的关联规则.

3 软集包含度

定义 7 设(F,A)是 U 上的一个软集,$e_1,e_2\in V$.定义(F,A)上 $F(e_1)$和 $F(e_2)$的包含度为:

$$Clu(F(e_2)/F(e_1))=\frac{|u:e_2\cup e_1\subseteq Co(u)|}{|u:e_1\subseteq Co(u)|}=\frac{|F(e_2)\cap F(e_1)|}{|F(e_1)|}. \tag{5}$$

定理 1 设(F,A)是 U 上的一个软集,$U=(h_1,h_2,\cdots,h_n)$是对象集,$A=(e_1,e_2,\cdots,e_n)$是条件属性集,则 Clu 为软集(F,E)上的包含度.

证明 设 $F(e_1),F(e_2),F(e_3)\in(F,E)$,可以证明

(1)由于 $F(e_1)\cap F(e_2)\subseteq F(e_1)$,所以$|F(e_1)\cap F(e_2)|\leq|F(e_1)|$,从而$0\leq Clu(F(e_2)/F(e_1))\leq 1$.

(2)若$|F(e_1)|\leq|F(e_2)|$,则有 $F(e_1)\subseteq F(e_2)$,即$F(e_1)\cap F(e_2)=F(e_1)$,所以有

$$Clu(F(e_2)/F(e_1))=\frac{|F(e_2)\cap F(e_1)|}{|F(e_1)|}=\frac{|F(e_1)|}{|F(e_1)|}=1.$$

(3)若$|F(e_1)|\leq|F(e_2)|\leq|F(e_3)|$,则有 $F(e_1)\subseteq F(e_2)\subseteq F(e_3)$,所以有

$$Clu(F(e_1)/F(e_3))=\frac{|F(e_1)\cap F(e_3)|}{|F(e_3)|}\leq\frac{|F(e_1)|}{|F(e_2)|}=Clu(F(e_1)/F(e_2)).$$

(4)若$|F(e_1)|\leq|F(e_2)|$,对于$\forall F(e_3)\in(F,E)$,则有

$$Clu(F(e_1)/F(e_3))=\frac{|F(e_1)\cap F(e_3)|}{|F(e_3)|}\leq\frac{|F(e_2)\cap F(e_3)|}{|F(e_3)|}=Clu(F(e_2)/F(e_3)).$$

因此,Clu 为软集(F,A)上的包含度.证毕.

由定理 1 给出了软集的包含度,进而我们讨论以下几个结论.

引理 1 设(F,A)是 U 上的一个软集,$e_1,e_2\in A$,则 $Clu(F(e_2)/F(e_1))=1\Leftrightarrow F(e_1)\subseteq F(e_2)$.

证明 由定义 3 和定义 7 即可得证.

引理 2 设(F,A)是 U 上的一个软集,$X,Y\subseteq A$,$X\cap Y=\varnothing$.则(F,A)上 $F(X)$和 $F(Y)$的包含度为:

$$Clu(F(X)/F(Y))=\frac{|F(X)\cap F(Y)|}{|F(Y)|}. \tag{6}$$

其中 $F(X)=\bigcap_{e\in X}F(e)$,$F(Y)=\bigcap_{e\in Y}F(e)$.

证明 由定义 3 和定义 7 即可得证.

4 基于软集包含度的关联规则挖掘

4.1 软集包含度的关联规则挖掘

定理 2 设(F,A)是 U 上的一个软集,$X,Y\subseteq A$,$X\cap Y=\varnothing$.若 $Clu(F(Y)/F(X))=1$,则关联规则 $Y\Rightarrow X$ 的可信度 $conf(Y\Rightarrow X)$ 就为 $Clu(F(X)/F(Y))$.

证明 若 $Clu(F(Y)/F(X))=1$,引理 2 得到 $F(X)\subseteq F(Y)$.

所以 $Clu(F(X)/F(Y))=\frac{|F(X)\cap F(Y)|}{|F(Y)|}=\frac{|F(X)|}{|F(Y)|}$.

又因为由 $Clu(F(Y)/F(X))=\frac{|F(Y)\cap F(X)|}{|F(X)|}=1$,可以推出 $F(X)\subseteq F(Y)$,所以关联规则 $Y\Rightarrow X$ 的可信度 $Conf(Y\Rightarrow X)=\frac{|\{u:X\cup Y\subseteq Co(u)\}|}{|\{u:Y\subseteq Co(u)\}|}=\frac{|F(X)|}{|F(Y)|}$.

即 $Clu(F(X)/F(Y))=Conf(Y\Rightarrow X)$.所以当 $Clu(F(Y)/F(X))=1$时,关联规则 $Y\Rightarrow X$ 的可信度就为包含度 $Clu(F(X)/F(Y))$.证毕.

这说明当两个属性集之间的包含度达到一定的条件时,就可以用它们之间的包含度代替可信度来进行关联规则的提取.

定理 3 设(F,A)是 U 上的一个软集,$X,Y\subseteq A$,$X\cap Y=\varnothing$,给定支持度阈值 α 和可信度的阈值β,则 X,Y 两属性集满足下列条件:

(1) $Clu(F(Y)/F(X))=1$

(2) $|F(X)|/|G|\geq\alpha$ 或者$|F(X)|=\alpha\cdot|G|$

(3) $Clu(F(X)/F(Y))\geq\beta$

则可在两属性集上提取关联规则 $Y\Rightarrow X$.

证明 设(F,A)是 U 上的一个软集,$X,Y\subseteq A$,$X\cap Y=\varnothing$,满足上述条件,由(1)及引理 2 可知

$$Clu(F(Y)/F(X))=\frac{|F(Y)\cap F(X)|}{|F(X)|}=1\Leftrightarrow F(X)\subseteq F(Y),$$

这说明若满足条件(2),则说明关联规则 $Y\Rightarrow X$ 的支持度满足

$$sup(Y\Rightarrow X)=sup(X\cup Y)=\frac{|\{u:X\cup Y\subseteq Co(u)\}|}{|U|}\geq\alpha$$

若再满足条件(3),由推论 1 知 $Clu(F(X)/F(Y))\geq\beta\Leftrightarrow Conf(Y\Rightarrow X)\geq\beta$.因此提取关联规则 $Y\Rightarrow X$.证毕.

由此可以看出,利用软集元素间的包含度可以更

方便提取人们所关心的可信度大于 β 的关联规则.进一步考虑在支持属性集提取过程中存在的大量冗余问题,支持属性集的数量可达到 2^m(m 为属性的个数),若对所有的支持属性集都进行包含度的计算,几乎是不可能的,所以需要约减.

4.2 软集最大包含关联规则挖掘

定义 8 设(F,A)是 U 上的一个软集,$\bar{X}\subseteq A$,如果 $\bar{X}=Co(u)\cap A$,则属性集 $\bar{X}$ 被称为是对象 u 的最大支持属性集.

$\bar{X}\in P(A)$,由 U 上的对象和 $Co(u)$的所有最大支持属性集生成(F,A)的最大支持软集(F',X'),其中对于任意元素 $\bar{X}\in X'$,$F(\bar{X})=\{u:\bar{X}=Co(u)\cap A\}$.

定理 4 由(F,A)诱导出的最大支持软集(F',X')是支持软集(F'',X)的子软集,即 $X'\subseteq X$,并且对于任意的 $\bar{X}\in X'$,$F''(\bar{X})\subseteq F'(\bar{X})$.

证明 由定义即可证.

这说明最大支持属性集是对支持属性集的一个约减.

定义 9 设(F,A)是 U 上的一个软集,$\bar{X},\bar{Y}\subseteq A$ 是两个最大支持属性集,并且 $\bar{X}\cap\bar{Y}=\varnothing$,定义$(F,A)$上 $F(\bar{X})$和 $F(\bar{Y})$的包含度为:

$$MClu(F(\bar{X})/F(\bar{Y}))=\frac{|F''(\bar{Y})\cap F''(\bar{X})|}{|F''(\bar{Y})|}\qquad(7)$$

下面我们在 SCAR(S,C)算法的基础上给出在事务数据库软集上的基于软集包含度的最大关联规则提取算法.

算法　SCMAR(S,C)

输入:1.事务数据软集 $S=(F,A)$,

项目集:$U=\{u_1,u_2,\cdots,u_m\}$,

属性集:$A=\{A_1,A_2,\cdots,A_s\}$,

分类集:$A_i=\{e_j^i:j=1,2,\cdots,t\}$,$A=\bigcup_{i=1}^{m}A_i$.

2.项目的共现集:$C=\{Co(u_1),Co(u_2),\cdots,Co(u_m)\}$

输出:R 关联规则集.//R 存放软集的关联规则.

Setp1:针对每一个 $A_i(i=1,2,\cdots s)$,在共现集 C 上生成它的所有最大候选属性集 $X=\{X_j^i:X_j^i\subset P(A_i)\wedge X_j^i=Co(u_j)\cap A_i\}$,并形成相应的最大支持软子集$(F,X)$.每一个元素按包含项目的个数进行分类,即$|F(X)|=j$ 的集元素存放在数组 $B_i[j]$中,并令 $Size[i]:=\max(j)$.

Step2:初始化关联规则集 $R=\varnothing$.

Step3:对给定的支持度的阈值 α 和可信度的阈值β,令 $j=\lceil\alpha^*|U|\rceil$.

Step4:若 $B_i[j]\neq\varnothing$,则对 $B_i[j]$中的每个软属性集元素 $F(X)$,计算 $B_{i+1}[1],B_{i+1}[2],\cdots,B_{i+1}[size]$中的软属性集元素 $F(Y_p)$与 $F(X)$的包含度 $Clu(F(X)/F(Y_p))$.若 $Clu(F(X)/(Y_p))=1$,则进一步计算 $Clu(F(Y_p)/F(X))$;若 $Clu(F(Y_p)/F(X))\geq\beta$,则可以提取关联规则 $X\Rightarrow Y_p$,$R=R\cup\{X\Rightarrow Y_p\}$.

Step 5 $j:=j+1$,若 $j\leq size[j]$,重复 Step 4.

Step 6 $i:=i+1$,若 $j\leq size[i]$,重复 Step 4.

Step 7 输出 R.

分析 SCMAR(S,C)算法可知,假设属性集分类数为 $s=n$,算法 Step1 的时间复杂度达到 $O(n|U|))$,在最坏情况下为 $O(|A|\cdot|U|)$.Step4 复杂度为 $O(|B_i|\cdot|B_j|)$,在最坏情况下为 $O(|U|^2)$,所以最坏情况下的时间复杂度为 $O(\max(|A|\cdot|U|,|U|^2))$.

5 实例分析

下面我们利用文献[17~19]的实例来说明和分析我们所提出的基于包含度的软集关联规则提取算法 SCMAR.如表 1 所示的数据库由 10 个事务所组成.基于事务数据建立分类 $A=\{countries,topics\}$,其中 $countries=\{Canada,Iran,USA\}$,$topics=\{crude,ship,earn,jobs,cpi,sugar,tea,trade,acq\}$.将此事务数据库表示为软集 $S=(F,A)$,其中 $U=\{u_1,u_2,\cdots,u_{10}\}$,$A=\{A_1,A_2\}$,$A_1=\{Canada,Iran,USA\}$,$A_2=\{crude,ship,earn,jobs,cpi,sugar,tea,trade,acq\}$.

表 1　事务数据库

TID	Items
1	*Canada*, *Iran*, *USA*, *crude*, *ship*
2	*Canada*, *Iran*, *USA*, *crude*, *ship*
3	*USA*, *earn*
4	*USA*, *jobs*, *cpi*
5	*USA*, *jobs*, *cpi*
6	*USA*, *earn*, *cpi*
7	*Canada*, *sugar*, *tea*
8	*Canada*, *USA*, *trade*, *acq*
9	*Canada*, *USA*, *trade*, *acq*
10	*Canada*, *USA*, *earn*

表示为对应的软集:

$(F,E)=\{F(Canada)=\{u_1,u_2,u_7,u_8,u_9,u_{10}\},F(Iran)=\{u_1,u_2\},F(USA)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8,u_9,u_{10}\},F(crude)=\{u_1,u_2\},F(ship)=\{u_1,u_2\},F(earn)=\{u_3,u_{10}\},F(jobs)=\{u_4\},F(cpi)=\{u_3,u_{10}\},F(sugar)=\{u_7\},F(tea)=\{u_7\},F(trade)=\{u_8,u_9\},F(acq)==\{u_8,u_9\}\}$.

给定的支持度的阈值 $\alpha=0.2$,可信度的阈值 $\beta=0.50$.根据定义 5 得到如下的共现集:

$C=\{Co(u_1)=\{Canada,Iran,USA,crude,ship\},Co(u_2)=\{Canada,Iran,USA,crude,ship\},Co(u_3)=\{USA,earn\},Co(u_4)=\{USA,jobs,cpi\},Co(u_5)=\{USA,jobs,cpi\},Co(u_6)=\{USA,earn,cpi\},Co(u_7)=\{Canada,sugar,tea\},Co(u_8)=\{Canada,USA,trade,acq\},Co(u_9)=\{Canada,USA,trade,acq\},Co(u_{10})=\{Canada,USA,earn\}\}$.

(1)依据定义 8 由所有最大支持属性集生成(F,A_i)的最大支持软子集(F,X_i),并按包含项目的个数分

别存放在数组 $B_1[j]$和 $B_2[j]$中:

$B_1(1)=\{F(Canada)=\{u_7\}\}$

$B_1(2)=\{F(\{Canada,Iran,USA\})=\{u_1,u_2\}\}$

$B_1(3)=\{F(\{Canada,USA\})=\{u_8,u_9,u_{10}\}\}$

$B_1(4)=\{F(\{USA\})=\{u_3,u_4,u_5,u_6\}\}$.

$B_2(1)=\{F(\{earn,cpi\})=\{u_6\},F(\{sugar,tea\})=\{u_7\}\}$

$B_2(2)=\{F(\{earn\})=\{u_3,u_{10}\},F(\{crude,ship\})=\{u_1,u_2\},$
$F(\{jobs,cpi\})=\{u_4,u_5\},F(\{trade,acq\})=\{u_8,u_9\}\}$.

因为给定的 $\alpha=0.2$ 和 $\beta=0.50$,所以$\lceil\alpha*|U|\rceil=0.2\times 10=2$.则 $j=2\geq 2$.从 $B_1[2]$中取一个软属性 $F(\{Canada,Iran,USA\})=\{u_1,u_2\}$,依此与 B_2 中的每一个最大软属性集 $F(Y)$计算,判断 $Clu(F(\{Canada,Iran,USA\})/F(Y)=1$是否成立,其中只有 $Clu(F(\{Canada,Iran,USA\})/F(\{crude,ship\}))=1$成立,则进一步计算 $Clu(F(crude,ship)/F(Canada,Iran,USA))=|\{u_1,u_2\}|/|\{u_1,u_2\}|=1\geq 0.5$,所以提取关联规则 $Canada,Iran,USA\Rightarrow crude,ship$.

(2)同理,从 $B_1[3]$和 $B_1[4]$中取软属性集重复以上过程,最后我们得到该事务数据软集上所有满足要求的关联规则为:

$Canada,Iran,USA\Rightarrow crude,ship$,可信度为 100%;

$Canada,USA\Rightarrow trade,acq$,可信度为 66%;

$USA\Rightarrow jobs,cpi$,可信度为 50%.

由此可见,上述 SCMAR 算法与文献[17~19]中提取的关联规则结果相同.同时本文对文献[17]中的 SAI 算法与 SCMAR 算法在不同的事务数和参数个数情况下进行性能测试,比较结果如图 1 和图 2 所示.

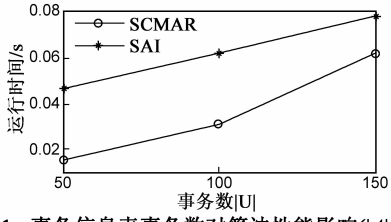

图1 事务信息表事务数对算法性能影响($|A|=5$)

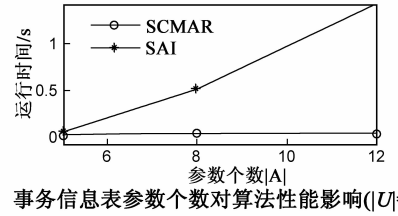

图2 事务信息表参数个数对算法性能影响($|U|=50$)

从图中可看出,基于软集包含度方法较大的提高了算法的效率,而且提取的关联规则无冗余.在实际应用中更容易处理多参数、大量数据的信息表.

6 结论

软集规则关联挖掘是在数据分析应用中的一个重要研究内容,对信息表数据的组织形式更有利于事务数据分析,软集的规则关联挖掘是布尔型关联规则挖掘.基于软集包含度理论的关联规则挖掘更有利于挖掘有意义的关联规则.我们将包含度引入软集数据关联规则挖掘中,利用包含度理论描述属性集之间的量化关系,给出软集上属性集间的包含度、关联规则和最大关联规则的概念,讨论了包含度和可信度之间的联系,给出最大软关联规则的提取算法,可以在事务数据软集中挖掘满足给定的支持度阈值和可信度阈值的关联规则.通过实例与粗集理论的有关算法作比较,表明该算法的方法是正确的,极大地约简了冗余,提高了算法的效率.在此基础上,今后我们将通过实践,进一步深入研究软集在决策分析中的更广泛应用.

参考文献

[1] D Molodtsov. Soft set theory-first results[J]. Computers and Mathematics with Applications, 1999, 37: 19 – 31.

[2] P K Maji, R Biswas, A R Roy. Soft set theory[J]. Computers and Mathematics with Applications, 2003, 45: 555 – 562.

[3] P K Maji, A R Roy. An application of Soft sets in a decision making problem[J]. Computers and Mathematics with Applications, 2002, 44: 1077 – 1083.

[4] P K Maji, A R Biswas, A R Roy. Fuzzy soft sets[J]. The Journal of Fuzzy Mathematics, 2001, 9(3): 589 – 602.

[5] Feng Feng, Jun Youngbae, Zhao Xianzhong. Soft semi-rings [J]. Computers and Mathematics with Applications, 2008, 56: 2621 – 2628.

[6] Feng Feng, Li Changxing, B Davvaz, M Irfan Ali. Soft sets combined with fuzzy sets and rough sets: a tentative approach [J]. Soft Computing, 2010, 14: 899 – 911.

[7] A R Roy, P K Maji. A fuzzy soft set theoretic approach to decision making problems[J]. Journal of Computational and Applied Mathematics, 2007, 203(3): 412 – 418.

[8] Yuncheng Jiang, Hai Liu, Yongtang, Qimai Chen. Semantic decision making using ontology-based soft sets[J]. Mathematical and Computer Modelling, 2011, 53(5): 1140 – 1149.

[9] 耿生玲,李永明,冯峰.软集决策信息系统的属性约简[J].小型微型计算机系统,2011,32(4):721 – 725.
Geng Shengling, Li Yongming, Feng Feng. Attributereduction of decision information system of soft sets[J]. Journal of Chinese Computer Systems, 2011, 32(4): 721 – 725. (in Chinese)

[10] 张文修,徐宗本,梁怡.包含度理论[J].模糊系统与数学,1996,10(4):1 – 9.
Zhang Wenxiu, Xu Zongben, Liang Yi. Inclusion degree theory [J]. Fuzzy Systems and Mathematics, 1996, 10(4): 1 – 9. (in

Chinese)

[11] 张文修,梁广锡,梁怡.包含度及其在人工智能中的应用[J].西安交通大学学报,1995,29(8):111 – 116.
Zhang Wenxiu, Liang Guangxi, Liang Yi. Including degree and its applications to artificial intelligence[J]. Journal of Xi'an Jiao Tong University, 1995, 29(8): 111 – 116. (in Chinese)

[12] 王云岚,李增智,屈科文.基于候选项集个数上阶的增量式关联规则更新算法[J].电子学报,2004,32(5):731 – 734.
WANG Yun-lan, LI Zeng-zhi, QU Ke-wen. A general incremental algorithm for mining as sociation rules[J]. Acta Electronica Sinica, 2004, 32(5): 731 – 734. (in Chinese)

[13] 范九伦,吴成茂.FCM 算法中隶属度的新解释及其应用[J].电子学报,2004,32(2):350 – 352.
FAN Jiu-lun, WU Cheng-mao. The new explanation of membership degree in FCM and its applications[J]. Acta Electronica Sinica, 2004, 32(2): 350 – 352. (in Chinese)

[14] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[A]. Proceedings of the ACM SIGMOD Conference on Management of Data[C]. New York: ACM, 1993. 207 – 216.

[15] 刘远超,王晓龙,徐志明,等.基于粗集理论的中文关键词短语构成规则挖掘[J].电子学报,2007,35(2):371 – 374.
LIU Yuan-chao, WANG Xiao-long, XU Zhi-ming, et al. Mining construction rules of Chinese keyphrase based on rough set theory[J]. Acta Electronica Sinica, 2007, 35(2): 371 – 374. (in Chinese)

[16] A H L Lim, C S Lee. Processing online analytics with classification and association rule mining[J]. Knowledge-Based Systems, 2010, 23(3): 248 – 255.

[17] T Herawan, M Mat Deris. A soft set approach for association rules mining[J]. Knowledge-Based Systems, 2011, 24(1): 186 – 195.

[18] Y Bi, T Anderson, S McClean. A rough set model with ontologies for discovering maximal association rules in document collections[J]. Knowledge-Based Systems, 2003, 16(5): 243 – 251.

[19] D A Bell, J W Guan, D Y Liu. Mining association rules with rough sets[J]. Studies in Computational Intelligence, 2005, 5: 163 – 184.

作者简介

耿生玲　女,1970 年 12 月出生,青海都兰人,博士生、教授,主要研究方向:计算理论、数据挖掘、图形处理.
E-mail: gengsl@qhnu.edu.cn

李永明　男,1966 年出生,陕西大荔人,博士,教授,博士生导师,研究方向:计算智能、量子逻辑、量子计算、模型检测.
E-mail: liyongm@snnu.edu.cn