

形式概念分析的概念之间包含度理论

曲开社 梁亮 梁吉业 李德玉 陈红星

(山西大学计算机与信息技术学院计算智能与中文信息处理省部共建教育部重点实验室 太原 030006)

摘要 在形式概念分析中的概念之间引进了包含度理论,利用包含度理论描述了概念之间的量化关系,对概念格中概念之间的亚概念和超概念进行了刻画。同时,采用包含度在概念格中进行关联规则的提取,给出了由包含度进行无冗余关联规则的提取算法,并由实例验证了该算法是有效的。

关键词 概念格,包含度,关联规则

Inclusion Degree Theory among Concepts of Formal Concept Analysis

QU Kai-she LIANG Liang LIANG Ji-ye LI De-yu CHEN Hong-xing

(Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract Inclusion degree theory was introduced into the concepts of formal concept analysis. In this theory, quantitative relation among concepts was described and the sub-concept and super-concept in concept lattice was depicted. Meanwhile, by using inclusion degree association rules were extracted from the concept lattice, and an algorithm of extracting non-redundant association rules was presented. Finally, an example was given to show that the algorithm is effective.

Keywords Concept lattice, Inclusion degree, Association rules

1 引言

形式概念分析^[1] (Formal Concept Analysis) 是德国学者 Wille 于 1982 年作为一种数学理论首先提出的,现已被广泛研究^[2-6],并应用到机器学习^[7]、软件工程^[8]和信息获取^[9]等领域,是一种强有力的数据分析工具。

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域^[12-14]有着重要的应用。

关联规则^[15] (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。

文献[14]把包含度的概念引入到粗糙数据分析中,得到了一些较好的结果。文献[16]把包含度的概念引入到形式概念分析中,在形式背景 $K=(G, M, I)$ 的对象集 G 、属性集 M 和 $G \cup M$ 上分别定义了包含度,并利用包含度理论对形式概

念分析中的内涵和外延以及蕴涵规则进行了表示,为从定量分析角度研究形式概念分析提供了依据。本文在概念格的概念之间引入了包含度,利用包含度理论描述了概念之间的量化关系,刻画了概念格中概念之间的亚概念和超概念之间的关系,这有利于对概念进行进一步的量化研究。同时,研究了包含度和可信度之间的联系,并利用包含度在概念格中进行关联规则的提取,给出了由包含度进行无冗余关联规则的提取算法,并由实例验证了该算法是有效的。

2 基本概念

2.1 形式概念分析

在形式概念分析中,数据是用形式背景来表示的,下面我们给出它的形式化定义^[1]。

定义 1 一个形式背景 K 是一个三元组: $K=(G, M, I)$, 其中 G 为所有对象的集合, M 为所有属性的集合, $I \subseteq G \times M$ 为 G 和 M 中元素之间的关系集合。对于 $g \in G, m \in M, (g, m) \in I$ 表示“对象 g 具有属性 m ”。

定义 2 设 $K=(G, M, I)$ 为一形式背景。对于集合 $A \subseteq G$, 记

$$A' = \{m \in M \mid (g, m) \in I, \forall g \in A\}$$

相应地,对于集合 $B \subseteq M$, 记:

$$B' = \{g \in G \mid (g, m) \in I, \forall m \in B\}$$

到稿日期:2008-03-04 本课题得到国家自然科学基金(60773133, 70471003, 60573074), 山西省自然科学基金(2007011040)资助。

曲开社(1954-),男,教授,主要研究方向为人工智能、粗糙集与概念格等, E-mail: quks@sxu.edu.cn; 梁亮(1984-),男,硕士研究生,研究方向为概念格; 梁吉业(1962-),男,教授,博士生导师,主要研究方向为人工智能等; 李德玉(1965-),男,教授,博士生导师,主要研究方向为人工智能等; 陈红星(1962-),男,讲师,主要研究方向为人工智能等。

为方便起见,我们用 g' 表示 $\{g\}'$, 用 m' 表示 $\{m\}'$ 。

定义 3 设 $K=(G, M, I)$ 为一形式背景, $A \subseteq G, B \subseteq M$, 称 $C=(A, B)$ 为 K 的一个概念, 如果 $A'=B$ 且 $B'=A$ 。此时称 A 为 C 的外延, B 为 C 的内涵。我们用 $L(K)$ 记 K 的所有概念组成的集合。

定义 4 设 $K=(G, M, I)$ 为一形式背景, $C_1=(A_1, B_1)$, $C_2=(A_2, B_2)$ 是 K 的两个概念, 规定 $C_1 \leq C_2 \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_1 \supseteq B_2)$ 。此时, C_2 称为 C_1 的超概念(super-concept), C_1 称为 C_2 的亚概念(sub-concept)。

2.2 包含度

定义 5^[10] 设 $(L, \leq)$ 为一偏序集。若对于任意的 $x, y \in L$, 有数 $D(y/x)$ 与之对应, 且满足

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

$$(2) \text{若 } x \leq y, D(y/x) = 1;$$

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

(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, 经典包含仅是包含度为 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^[17]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: 形式背景中具有属性集 A 中的属性的对象也可能具有属性集 B 中的属性。

对于概念格中的亚概念-超概念节点对 (C_1, C_2) , 假设 $C_1=(A, B_1 \cup B)$, $C_2=(A_2 \cup A, B)$, 并且 $A \cap A_2 = \emptyset$, $B \cap B_1 = \emptyset$, 可以得到关联规则 $B \Rightarrow B_1$, 其支持度是

$$\text{supp}(B \Rightarrow B_1) = \frac{|A|}{|G|} = \frac{|\text{extent}(C_1)|}{|G|}$$

可信度是

$$\text{conf}(B \Rightarrow B_1) = \frac{|A|}{|A_2 \cup A|} = \frac{|\text{extent}(C_1)|}{|\text{extent}(C_2)|}$$

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 θ 和一个可信度的阈值 ϕ , 这样人们可以只关心支持度大于 θ 的节点和可信度大于 ϕ 的节点对。本文只讨论可信度不为 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。

定理 1^[17] 对于符合阈值条件的节点序列 $C_1 > C_2 > C_3$, C_1 和 C_2 上提取的规则及 C_2 和 C_3 上提取的规则都是由 C_1 和 C_3 上提取的规则推导出来的。

3 概念上的包含度

定义 6 设 $K=(G, M, I)$ 为一形式背景, $L(K)$ 为该形式背景上的所有概念的集合, $C_1=(A_1, B_1)$, $C_2=(A_2, B_2)$ 为形式背景上的两个概念, 定义 K 上概念集合 $L(K)$ 的包含度为 $Cu: L(K) \times L(K) \rightarrow [0, 1]$ 为

$$Cu(C_2/C_1) = \frac{|A_2 \cap A_1|}{|A_1|}$$

定理 2 设 $K=(G, M, I)$ 为一形式背景, $L(K)$ 为该形式背景上的所有概念的集合, 则 Cu 为概念集合 $L(K)$ 上的包含度。

证明: 设

$$C_1=(A_1, B_1), C_2=(A_2, B_2), C_3=(A_3, B_3) \in L(K)$$

(1) 由于 $A_2 \cap A_1 \subseteq A_1$, 所以 $|A_2 \cap A_1| \leq |A_1|$, 从而 $0 \leq Cu(C_2/C_1) \leq 1$ 。

(2) 若 $C_1 \leq C_2$, 则有 $A_1 \subseteq A_2$, 即 $A_2 \cap A_1 = A_1$, 所以有

$$Cu(C_2/C_1) = \frac{|A_2 \cap A_1|}{|A_1|} = \frac{|A_1|}{|A_1|} = 1$$

(3) 若 $C_1 \leq C_2 \leq C_3$, 则有 $A_1 \subseteq A_2 \subseteq A_3$, 所以有

$$|A_1 \cap A_3| = |A_1| = |A_1 \cap A_2| \text{ 和 } |A_2| \leq |A_3|$$

从而有

$$\begin{aligned} Cu(C_1/C_3) &= \frac{|A_1 \cap A_3|}{|A_3|} = \frac{|A_1|}{|A_3|} \leq \frac{|A_1|}{|A_2|} \\ &= \frac{|A_1 \cap A_2|}{|A_2|} = Cu(C_1/C_2) \end{aligned}$$

(4) 若 $C_1 \leq C_2, \forall C_3 \in L(K)$, 则有 $A_1 \subseteq A_2 \Rightarrow A_1 \cap A_3 \subseteq A_2 \cap A_3 \Rightarrow |A_1 \cap A_3| \leq |A_2 \cap A_3|$ 从而

$$Cu(C_1/C_3) = \frac{|A_1 \cap A_3|}{|A_3|} \leq \frac{|A_2 \cap A_3|}{|A_3|} = Cu(C_2/C_3)$$

从而 Cu 为概念集合 $L(K)$ 上的包含度。证毕。

下面我们将利用概念之间的包含度来讨论概念格上的亚概念和超概念之间的关系。

引理 1 设 $K=(G, M, I)$ 为一形式背景, $C_1=(A_1, B_1)$, $C_2=(A_2, B_2)$ 为形式背景上的两个概念, 则 $Cu(C_2/C_1) = 0 \Leftrightarrow A_1 \cap A_2 = \emptyset$ 。

证明: 显然,

$$\begin{aligned} Cu(C_2/C_1) &= \frac{|A_2 \cap A_1|}{|A_1|} = 0 \Leftrightarrow |A_1 \cap A_2| = 0 \Leftrightarrow A_1 \cap A_2 \\ &= \emptyset \end{aligned}$$

证毕。

引理 2 设 $K=(G, M, I)$ 为一形式背景, $C_1=(A_1, B_1)$, $C_2=(A_2, B_2)$ 为形式背景上的两个概念, 则 $Cu(C_2/C_1) = 1 \Leftrightarrow A_1 \subseteq A_2$ 。

证明: 显然,

$$\begin{aligned} Cu(C_2/C_1) &= \frac{|A_2 \cap A_1|}{|A_1|} = 1 \Leftrightarrow |A_2 \cap A_1| = |A_1| \Leftrightarrow \\ A_2 \cap A_1 &= A_1 \Leftrightarrow A_1 \subseteq A_2 \end{aligned}$$

证毕。

定理 3 设 $K=(G, M, I)$ 为一形式背景, $C_1=(A_1, B_1)$, $C_2=(A_2, B_2)$ 为形式背景上的两个概念, 若 $Cu(C_2/C_1) = 0$, 则对 C_1 的任意亚概念 $C_3=(A_3, B_3)$, 有 $Cu(C_2/C_3) = 0$ 。

证明: 若 $Cu(C_2/C_1) = 0$, 由引理 1 知 $A_1 \cap A_2 = \emptyset$, 又因为 C_3 是 C_1 的亚概念, 所以有 $C_3 \leq C_1$ 。因此有 $A_3 \subseteq A_1$, 从而

$$A_1 \cap A_2 = \emptyset \Rightarrow A_3 \cap A_2 = \emptyset \Rightarrow |A_3 \cap A_2| = 0$$

$$\text{即有 } Cu(C_2/C_3) = \frac{|A_2 \cap A_3|}{|A_3|} = 0.$$

证毕。

定理 4 设 $K=(G, M, I)$ 为一形式背景, $C_1=(A_1, B_1)$, $C_2=(A_2, B_2)$ 为形式背景上的两个概念, 若 $Cu(C_2/C_1) = 0$, 则对 C_2 的任意亚概念 $C_3=(A_3, B_3)$, 有 $Cu(C_3/C_1) = 0$ 。

证明: 若 $Cu(C_2/C_1) = 0$, 由引理 1 知 $A_1 \cap A_2 = \emptyset$, 又因为 C_3 是 C_2 的亚概念, 所以有 $C_3 \leq C_2 \Leftrightarrow A_3 \subseteq A_2$, 因此由 $A_1 \cap A_2 = \emptyset$ 可推出 $A_1 \cap A_3 = \emptyset$ 即 $|A_1 \cap A_3| = 0$, 故有

$$Cu(C_3/C_1) = \frac{|A_1 \cap A_3|}{|A_1|} = 0$$

证毕。

定理3和定理4说明,若概念 C_2 完全不包含 C_1 ,则 C_2 也完全不包含 C_1 的任意亚概念,并且 C_2 的任意亚概念也完全不包含 C_1 。

定理5 设 $K=(G,M,I)$ 为一形式背景, $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$ 为形式背景上的两个概念,若 $Cu(C_2/C_1)=1$,则对 C_1 的任意亚概念 $C_3=(A_3,B_3)$,有 $Cu(C_2/C_3)=1$ 。

证明:若 $Cu(C_2/C_1)=1$,则由引理2知 $A_1 \subseteq A_2$,又因为 C_3 是 C_1 的亚概念,所以有 $C_3 \subseteq C_1$,即 $A_3 \subseteq A_1$,从而 $A_3 \subseteq A_2$ 。故有

$$Cu(C_2/C_3) = \frac{|A_2 \cap A_3|}{|A_3|} = \frac{|A_3|}{|A_3|} = 1$$

证毕。

定理6 设 $K=(G,M,I)$ 为一形式背景, $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$ 为形式背景上的两个概念,若 $Cu(C_2/C_1)=1$,则对 C_2 的任意超概念 $C_3=(A_3,B_3)$,有 $Cu(C_3/C_1)=1$ 。

证明:若 $Cu(C_2/C_1)=1$,由引理2知 $A_1 \subseteq A_2$,又因为 C_3 是 C_2 的超概念,所以有 $C_3 \supseteq C_2$,从而有 $A_2 \subseteq A_3$,所以可以推出 $A_1 \subseteq A_3$,所以

$$Cu(C_3/C_1) = \frac{|A_1 \cap A_3|}{|A_1|} = 1$$

证毕。

定理5和定理6说明了若 C_2 包含 C_1 ,则 C_2 也包含 C_1 的任意亚概念,并且 C_2 的任意超概念也包含 C_1 。

4 基于包含度的关联规则提取

4.1 运用包含度提取关联规则

我们发现,当两个概念之间的包含度满足一定的条件时,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。因此,我们有以下结果:

定理7 设 $K=(G,M,I)$ 为一形式背景, $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$ 为形式背景上的两个概念,若 $Cu(C_2/C_1)=1$,则关联规则 $B_2 \Rightarrow B_1 - B_2$ 的可信度为包含度 $Cu(C_1/C_2)$ 。

证明:若 $Cu(C_2/C_1)=1$,由引理2得 $A_1 \subseteq A_2$,所以

$$Cu(C_1/C_2) = \frac{|A_1 \cap A_2|}{|A_2|} = \frac{|A_1|}{|A_2|}$$

又因为由 $Cu(C_2/C_1) = \frac{|A_2 \cap A_1|}{|A_1|} = 1$ 可以推出 $A_1 \subseteq A_2$,即 $C_2=(A_2,B_2)$ 是 $C_1=(A_1,B_1)$ 的超概念,所以关联规则 $B_2 \Rightarrow B_1 - B_2$ 的可信度为

$$\text{conf}(B_2 \Rightarrow B_1 - B_2) = \frac{|\text{extent}(C_1)|}{|\text{extent}(C_2)|} = \frac{|A_1|}{|A_2|}$$

所以, $Cu(C_1/C_2) = \text{conf}(B_2 \Rightarrow B_1 - B_2)$ 。所以当 $Cu(C_2/C_1)=1$ 时,关联规则 $B_2 \Rightarrow B_1 - B_2$ 的可信度就为包含度 $Cu(C_1/C_2)$ 。证毕。

由定理7,有以下推论成立。

推论1 在形式背景 $K=(G,M,I)$ 上, $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$ 为形式背景上的两个概念,若 $Cu(C_2/C_1)=1$,则关联规则 $B_2 \Rightarrow B_1 - B_2$ 的可信度 $\text{conf}(B_2 \Rightarrow B_1 - B_2)$ 大于等于阈值 ϕ 当且仅当包含度 $Cu(C_1/C_2)$ 大于等于阈值 ϕ 。

根据推论1,就可以用概念之间的包含度代替可信度来进行亚概念-超概念节点对的关联规则提取。

定理8 在形式背景 $K=(G,M,I)$ 上, $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$ 为形式背景 K 上的两个概念,给定支持度阈值 θ 和

可信度的阈值 ϕ ,若概念 C_1 和 C_2 满足下列条件

$$1) Cu(C_2/C_1)=1$$

$$2) |A_1|/|G| \geq \theta \text{ 或者说 } |A_1| \geq \theta \cdot |G|$$

$$3) Cu(C_1/C_2) \geq \phi$$

则可在节点对 (C_1, C_2) 上提取关联规则

$$B_2 \Rightarrow B_1 - B_2$$

证明:设概念 C_1 和 C_2 满足上述条件,由1)及引理2可知

$$Cu(C_2/C_1) = \frac{|A_2 \cap A_1|}{|A_1|} = 1 \Leftrightarrow A_1 \subseteq A_2$$

这说明 (C_1, C_2) 是亚概念-超概念节点对,且 $C_1 \subseteq C_2$ 。进一步,若满足条件2),则说明关联规则 $B_2 \Rightarrow B_1 - B_2$ 的支持度满足 $\text{supp}(B_2 \Rightarrow B_1 - B_2) = |A_1|/|G| \geq \theta$ 。最后,若再满足条件3),由推论1知

$$Cu(C_1/C_2) \geq \phi \Leftrightarrow \text{conf}(B_2 \Rightarrow B_1 - B_2) \geq \phi$$

这就满足了在亚概念-超概念节点对上提取关联规则的所有条件,因此可在节点对 (C_1, C_2) 上提取关联规则

$$B_2 \Rightarrow B_1 - B_2$$

证毕。

由定理8我们可以看到,利用概念间的包含度可以更方便地提取人们所关心的可信度大于 ϕ 的关联规则。

利用定理8,可以给出一种在形式背景 $K=(G,M,I)$ 生成关联规则的算法:

算法1 基于包含度的关联规则提取算法

输入:形式背景 $K=(G,M,I)$

输出:关联规则集 R

1)利用文献[18]中的方法在形式背景 $K=(G,M,I)$ 上生成所有的概念;

2)把所有的概念按其外延包含元素的个数进行分类,即外延元素个数为 i 的概念存放在数组 $B[i]$ 中,并令 $\text{size} = \max(i)$;

3)初始化规则集 $R = \emptyset$;

4)对给定的支持度的阈值 θ 和可信度的阈值 ϕ ,令 $i = \lceil \theta \cdot |G| \rceil$;

5)若 $B[i] \neq \emptyset$,则对 $B[i]$ 中的每个概念 C ,计算 $B[i+1]$, $B[i+2]$, ..., $B[\text{size}]$ 中的概念 C_p 与 C 的包含度 $Cu(C_p/C)$ 。若 $Cu(C_p/C)=1$,则进一步计算 $Cu(C/C_p)$;若 $Cu(C/C_p) \geq \phi$,则就可以提取关联规则

$$B_p \Rightarrow B - B_p, R = R \cup \{B_p \Rightarrow B - B_p\}$$

6) $i = i + 1$,若 $i \leq \text{size}$,重复过程5);

7)输出 R 。

需要指出的是,采用以上方法提取的关联规则是有大量冗余的,所以就需要约简。下面将给出提取无冗余关联规则的算法。

4.2 无冗余关联规则的提取算法

由定理1,对于任意的符合阈值条件的节点序列,只要在序列两端的节点对上提取规则就可以了。

为了从形式背景上提取出无冗余的关联规则,特给出以下定义:

定义7 设 $K=(G,M,I)$ 为一形式背景, $L(K)$ 为该形式背景上的所有概念的集合, $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$ 为形式背景上的两个概念,并且 $C_1 \neq C_2$,对于给定的可信度的阈值

ϕ , 如果 $Cu(C_2/C_1)=1$ 且 $Cu(C_1/C_2) \geq \phi$, 并且不存在 $C_3=(A_3, B_3) \in L(K)$ 且 $C_3 \neq C_1 \neq C_2$, 使得 $Cu(C_3/C_2)=1$ 且 $Cu(C_1/C_3) \geq \phi$, 则称 C_2 为 C_1 的 ϕ 最大关联概念。集合 $MXA_{C_1}=\{C|C=(A, B) \in L(K) \wedge C \text{ 是 } C_1 \text{ 的 } \phi \text{ 最大关联概念}\}$ 称为 C_1 的 ϕ 最大关联概念集。

定义 8 设 $K=(G, M, I)$ 为一形式背景, $L(K)$ 为该形式背景上的所有概念的集合, $C_1=(A_1, B_1), C_2=(A_2, B_2)$ 为形式背景上的两个概念, 并且 $C_1 \neq C_2$, 对于给定的可信度的阈值 ϕ , 如果 $Cu(C_2/C_1)=1$ 且 $Cu(C_1/C_2) \geq \phi$, 并且不存在 $C_3=(A_3, B_3) \in L(K)$ 且 $C_3 \neq C_1 \neq C_2$, 使得 $Cu(C_1/C_3)=1$ 且 $Cu(C_3/C_2) \geq \phi$, 则称 C_1 为 C_2 的 ϕ 最小关联概念。集合 $MNA_{C_2}=\{C|C=(A, B) \in L(K) \wedge C \text{ 是 } C_2 \text{ 的 } \phi \text{ 最小关联概念}\}$ 称为 C_2 的 ϕ 最小关联概念集。

根据定义 7、定义 8 和定理 1, 可以得出一种在形式背景 $K=(G, M, I)$ 上生成无冗余的关联规则的算法:

算法 2 基于包含度的无冗余关联规则的提取算法

输入: 形式背景 $K=(G, M, I)$

输出: 关联规则集 R

算法:

1) 利用文献[18]中的方法在形式背景 $K=(G, M, I)$ 上生成所有的概念并把所有的概念按其外延包含元素的个数进行分类, 即外延元素个数为 i 的概念存放在数组 $B[i]$ 中, 并令 $size:=\max(i)$;

2) 初始化规则集 $R=\emptyset$;

3) 对给定的支持度的阈值 θ 和可信度的阈值 ϕ , 令 $i=\lceil \theta \cdot |G| \rceil$;

4) 若 $B[i] \neq \emptyset$, 则对 $B[i]$ 中的每个概念 $C_j=(A_j, B_j)$, 寻找其 ϕ 最大关联概念集 MXA_{C_j} 和其 ϕ 最小关联概念集 MNA_{C_j} 。

① 若 $MXA_{C_j} \neq \emptyset$ 且 $MNA_{C_j} \neq \emptyset$, 则

$\forall C_x=(A_x, B_x) \in MXA_{C_j}$

$\forall C_n=(A_n, B_n) \in MNA_{C_j}$

若 $Cu(C_n/C_x) \geq \phi$, 就可以提取关联规则 $B_x \Rightarrow B_n - B_x$ 。 $R=R \cup \{B_x \Rightarrow B_n - B_x\}$;

② 若 $MXA_{C_j} \neq \emptyset$ 且 $MNA_{C_j} = \emptyset$, 则 $\forall C_x=(A_x, B_x) \in MXA_{C_j}$, 就可以提取关联规则 $B_x \Rightarrow B_j - B_x$ 。 $R=R \cup \{B_x \Rightarrow B_j - B_x\}$ 。

5) 令 $i:=i+1$, 若 $i \leq size$, 重复过程 4);

6) 输出 R 。

由定理 1, 可知利用包含度在概念格中提取关联规则集是无冗余的。

4.3 例子

下面用一个例子来说明运用包含度来提取无冗余关联规则的方法。

例 1, 2, 3, 4, 5 五个顾客买 5 种书的情况如表 1 所列, 表中有 $\times$ 的地方说明该顾客买了该书。对于表 1 中的形式背景, 设定支持度阈值 $\theta=0.4$ 和可信度阈值 $\phi=0.6$, 计算得:

表 1 例子中的形式背景

	A	B	C	D	E
1	$\times$		$\times$	$\times$	
2		$\times$	$\times$		$\times$
3	$\times$	$\times$	$\times$		$\times$
4		$\times$			$\times$
5	$\times$	$\times$	$\times$		$\times$

1) 利用文献[18]中的方法在该形式背景上生成所有的概念并按其各自外延包含元素的个数分别存放在数组 $B[i]$ 中后的结果为

$B[0]=\{(\emptyset, \{ABCDE\})\}$

$B[1]=\{(\{1\}, \{ACD\})\}$

$B[2]=\{(\{35\}, \{ABCE\})\}$

$B[3]=\{(\{235\}, \{BCE\}), (\{135\}, \{AC\})\}$

$B[4]=\{(\{1235\}, \{C\}), (\{2345\}, \{BE\})\}$

$B[5]=\{(\{12345\}, \emptyset)\}$

2) 初始化规则集 R 为空。

3) 因为 $\theta=0.4, |G|=5$ 和 $\phi=0.6$, 所以 $\theta \cdot |G|=0.4 \times 5=2$ 。则 $i=2 \geq 2$, 所以

① 从 $B[2]$ 中取一个概念 $C=(\{35\}, \{ABCE\})$, 求得它的 ϕ 最大关联概念集 $MXA_C=\{(\{235\}, \{BCE\}), (\{135\}, \{AC\})\}$ 和 ϕ 最小关联概念集 $MNA_C=\emptyset$ 。则可得关联规则 $BCE \Rightarrow A, AC \Rightarrow BE$ 。由于初始时 R 为空, 所以把这两条规则加入 R 中, 得 $R=\{BCE \Rightarrow A, AC \Rightarrow BE\}$

② 然后, 从 $B[3]$ 中取一个概念 $C_1=(\{235\}, \{BCE\})$, 求得其 ϕ 最大关联概念集 $MXA_{C_1}=\{(\{12345\}, \emptyset)\}$ 和 ϕ 最小关联概念集 $MNA_{C_1}=\{(\{35\}, \{ABCE\})\}$, 由于当 $C_x=(\{12345\}, \emptyset)$ 和 $C_n=(\{35\}, \{ABCE\})$ 时, $Cu(C_n/C_x)=0.4 < \phi$, 所以不在这对节点上提取规则。

接下来从 $B[3]$ 中取概念 $C_2=(\{135\}, \{AC\})$, 求得其 ϕ 最大关联概念集 $MXA_{C_2}=\{(\{12345\}, \emptyset)\}$ 和 ϕ 最小关联概念集 $MNA_{C_2}=\{(\{35\}, \{ABCE\})\}$, 由于当 $C_x=(\{12345\}, \emptyset)$ 和 $C_n=(\{35\}, \{ABCE\})$ 时, $Cu(C_n/C_x)=0.4 < \phi$, 所以不在这对节点上提取规则。

③ 然后, 从 $B[4]$ 中取一个概念 $C_1=(\{1235\}, \{C\})$, 求得其 ϕ 最大关联概念集 $MXA_{C_1}=\{(\{12345\}, \emptyset)\}$ 和 ϕ 最小关联概念集 $MNA_{C_1}=\{(\{135\}, \{AC\})\}$, 由于当 $C_x=(\{12345\}, \emptyset)$ 和 $C_n=(\{135\}, \{AC\})$ 时, $Cu(C_n/C_x)=0.6 \geq \phi$, 所以可在这对节点上提取关联规则 $\emptyset \Rightarrow AC$ 。把该规则加入 R 中, 得 $R=R \cup \{\emptyset \Rightarrow AC\}=\{BCE \Rightarrow A, AC \Rightarrow BE, \emptyset \Rightarrow AC\}$ 。

接下来从 $B[4]$ 中取概念 $C_2=(\{2345\}, \{BE\})$, 求得其 ϕ 最大关联概念集 $MXA_{C_2}=\{(\{12345\}, \emptyset)\}$ 和 ϕ 最小关联概念集 $MNA_{C_2}=\{(\{235\}, \{BCE\})\}$, 由于当 $C_x=(\{12345\}, \emptyset)$ 和 $C_n=(\{235\}, \{BCE\})$ 时, $Cu(C_n/C_x)=0.6 \geq \phi$, 所以可在这对节点上提取关联规则 $\emptyset \Rightarrow BCE$ 。把该规则加入 R 中, 得

$R=R \cup \{\emptyset \Rightarrow BCE\}=\{BCE \Rightarrow A, AC \Rightarrow BE, \emptyset \Rightarrow AC, \emptyset \Rightarrow BCE\}$

④ 最后从 $B[5]$ 中取概念 $C=(\{12345\}, \emptyset)$, 由于求得其 ϕ 最大关联概念集 $MXA_C=\emptyset$, 所以没有规则可以提取。

最后, 得出该形式背景上的所有关联规则为 $R=\{BCE \Rightarrow A, AC \Rightarrow BE, \emptyset \Rightarrow AC, \emptyset \Rightarrow BCE\}$ 。这些关联规则的含义是: 购买 BCE 三本书的顾客也有可能购买 A 书; 购买 AC 两本书的顾客也有可能购买 BE 两本书; 所有的人都有可能购买 AC 两本书; 所有的人都有可能购买 BCE 三本书。

结束语 本文把包含度引入到概念格的概念之间的量化, 讨论了概念的亚概念和超概念之间的量化关系。另外, 本文利用包含度给出了在概念上提取关联规则算法, 该算法可

(下转第 219 页)

参考文献

- [1] 丁玮,齐东旭. 数字图像变换及信息隐藏技术[J]. 计算机学报, 1998,21(9):838-843
- [2] Shamir A. How to share a secret[J]. Communications of ACM, 1979,22(11):612-613
- [3] Naor M, Shamir A. Visual cryptography // Proc. of Eurocrypt '94. 1994;1-12
- [4] Cao Zhenfu. A threshold key escrow based on public key cryptosystem[J]. Science in China (Series E), 2001,44(4):441-448
- [5] Shannon C E. Communication theory of secrecy systems[J]. Bell System Technical Journal, 1949,28(4):656-715
- [6] Salam F, Marsden J, Varaiya P. Arnold diffusion in the swing equations of a power system[J]. IEEE Trans. Circuits and Systems, 1984,31(8):29-42
- [7] 李昌刚,韩正之,张浩然. 图像加密技术综述. 计算机研究与发展[J], 2002,39(10):1317-1324
- [8] Matthews R. On the derivation of a Chaotic encryption algorithm[J]. Cryptologia, 1989,13(1):29-42
- [9] Hedieu D, Ogorzalek M J. Identifiability and identification of chaotic systems based on adaptive synchronization[J]. IEEE Trans. on Circuits and Systems, 1997,44(10):948-962

- [10] Yang T, Yang Linbao, Yang Chunmei. Application of neural networks to unmasking chaotic secure communication[J]. Physica D, 1998,124:248-257
- [11] Wolfram S. Cryptography with cellular automata[J]. Advances in Cryptology, 1985:429-432
- [12] Nandi S, Kar B K, Pal P. Chaudhuri, Theory and application of cellular automata in cryptography[J]. IEEE Trans. on Computer, 1994,43(12):1346-1356
- [13] Seredynski F, Bouvry P, Zomaya A Y. cellular automata computations and secret key cryptography[J]. Parallel computing, 2004,30(5):753-766
- [14] Guan S-U, Zhang Shu, Maric Therese Quieta. 2-D CA Variation With Asymmetric Neighborhood for Pseudorandom Number Generation[J]. IEEE Trans. on computer-aided design of integrated circuits and systems, 2004,23(3):378-388
- [15] 张传武,沈野樵,彭启宗. 细胞自动机反向迭代加密技术研究[J]. 计算机学报, 2004,27(1):125-129
- [16] Gutowitz H, Victor J D, Knight B W. Local structure for cellular automata[J]. Physica D, 1987,28:18-48
- [17] Marsaglia G. Diehard. [OL]. <http://stat.fsu.edu/~geo/diehard.html>, 1998
- [18] ENT test suit. [OL]. <http://www.fourmilab.ch/random>, 1998

(上接第 213 页)

以提取无冗余的关联规则集。同时由算法看到,关联规则是利用概念之间的量化关系提出的,所以该算法更适用于在大型形式背景上进行关联规则的提取。

参考文献

- [1] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts // Rival I, ed. Ordered Sets. Dordrecht; Reidel, 1982;445-470
- [2] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations. Berlin; Springer-Verlag, 1999
- [3] Yao Y Y. Concept lattices in rough set theory // Proceedings of 2004 Annual Meeting of North American Fuzzy Information Processing society. Canada, 2004;796-801
- [4] Qu K S, Liang J Y, Wang J H, et al. The algebraic properties of concept lattice. Journal of Systems Science and Information, 2004,2(2):271-277
- [5] 谢志鹏,刘宗田. 概念格的快速渐进式构造算法. 计算机学报, 2002,25(5):490-495
- [6] 曲开社,翟岩慧,梁吉业,等. 形式概念分析对粗糙集理论的表示及扩展. 软件学报, 2007,18(9):2174-2182
- [7] Zupa B, Bohance M. Learning by discovering concept hierarchies. Artificial Intelligence, 1999,109(1-2):211-242
- [8] Tonella. Using a concept lattice of decomposition slices for program understanding and impact analysis. IEEE Transactions on

Software Engineering, 2003,29(6):495-509

- [9] Qu Kai-She, Zhai Yan-Hui, Liang Ji-Ye, et al. Study of decision implications based on formal concept analysis. International Journal of General Systems, 2007,36(2):147-156
- [10] 张文修,徐宗本,梁怡,等. 包含度理论. 模糊系统与数学, 1996,10(4):1-9
- [11] 张文修,梁怡. 不确定性推理原理. 西安:西安交通大学出版社, 1996
- [12] 梁怡,张文修. 模糊规则的谐调和矛盾规则的排除方法. 计算机学报, 1997,20(10):947-952
- [13] 张文修,梁广锡,梁怡. 包含度及其在人工智能中的应用. 西安交通大学学报, 1995,29(8):111-116
- [14] 梁吉业,徐宗本,李月香. 包含度与粗糙集数据分析中的度量. 计算机学报, 2001,24(5):544-547
- [15] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases // Proceedings of the ACM SIGMOD Conference on Management of Data. Washington D. C, 1993;207-216
- [16] 曲开社,翟岩慧. 偏序集、包含度与形式概念分析. 计算机学报, 2006,29(2):219-226
- [17] 苗夺谦,王国胤,刘清,等. 粒计算:过去、现在与展望. 科学出版社, 2007
- [18] 翟岩慧,曲开社,曹桃云. 基于矩阵秩的概念格生成算法. 电脑开发与应用, 2006,19(5):11-12