DOI:10.11992/tis.201603043
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160513.0925.026.html

基于相容模糊概念的规则提取方法

胡小康,王俊红
(山西大学 计算机与信息技术学院,山西 太原 030006)

摘 要:概念格是具有严格数学模型的数据分析与规则提取的一种有效工具,大部分情况下是在完备的精确形式背景即二值背景下进行研究,然而在现实生活中遇到的大多数情况是不完备的模糊形式背景,不完备模糊形式背景中包含许多不确定的信息,其上的知识表示与完备形式背景下的知识表示既有区别又有联系。为了研究两者的内在联系,本文定义了相似模糊概念和相容模糊概念,构建了相似模糊概念格和建立了在不完备模糊形式背景下相容模糊概念之间的偏序关系,进而设计出面向不完备模糊形式背景下的关联规则挖掘算法.最后通过实验验证了该方法的有效性和可行性。

关键词:形式背景;概念格;相似模糊概念;相容模糊概念;知识获取;关联规则;偏序关系;相容关系
中图分类号:TP18 **文献标志码**:A **文章编号**:1673-4785(2016)03-0352-07

中文引用格式:胡小康,王俊红.基于相容模糊概念的规则提取方法[J]. 智能系统学报, 2016, 11(3): 352-358.
英文引用格式:**HU Xiaokang, WANG Junhong. Research on rule extraction method based on compatibility fuzzy concept[J]. CAAI transactions on intelligent systems**, 2016,11(3): 352-358.

Research on rule extraction method based on compatibility fuzzy concept

HU Xiaokang, WANG Junhong
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: The concept lattice is an effective data analysis and rule extraction tool with a strict mathematical model. In most instances, studies are carried out in a complete formal context, i.e., a two-value context. However, in real life, an incomplete fuzzy formal context is frequently experienced. Incomplete fuzzy contexts contain a lot of uncertain information. There are both distinctions and relationships that can be identified between the forms of knowledge representation in the incomplete fuzzy formal and complete formal contexts. To study their internal relationship, in this paper, we define approximate fuzzy and compatible fuzzy concepts, establish an approximate fuzzy concept lattice, and identify a partial ordering relationship between compatible fuzzy concepts in an incomplete fuzzy formal context. We extend the design of an association rules mining algorithm to address the background of the incomplete fuzzy formal context, and conduct an experiment to demonstrate the feasibility and effectiveness of the proposed method.

Keywords: formal context; concept lattice; approximate fuzzy concept; compatible fuzzy concept; knowledge representation; association rules; partial ordering relation; compatible relation

概念格也称为 Galois 格,又叫做形式概念分析,由德国的 Wille[1] 在 20 世纪 80 年代提出。概念格的每个节点是一个形式概念,概念格结构模型是形式概念分析中的核心结构,它描述了对象和属性之间的关系。概念格是研究知识表示和推理的理论,它具有严格的数学模型,已经在机器学习、数据挖掘、软件工程等领域[2-6]得到广泛的应用。

通常我们研究的形式背景是完备的,也就是对象和属性之间的关系是已知的,但是在实际应用中,大多数信息是模糊[7]的、复杂的。更糟糕的是在现

收稿日期:2016-03-19. **网络出版日期**:2016-05-13.
基金项目:国家自然科学基金项目(61202018,61303008,61305057).
通信作者:王俊红.E-mail:wjhwjh@sxu.edu.cn.

实生活中由于人的认知能力以及机器的局限性,人们经常不能准确地判断对象和属性之间的关系,使得获取的形式背景经常存在数据缺失,从而得到形式背景是不完备的模糊形式背景,这对于知识获取产生了很大障碍。因此不完备模糊形式背景的研究获得了广泛的关注。

粗糙集理论中的信息系统就是形式概念分析中的形式背景,对于不完备信息系统[8],粗糙集已通过相容关系、非对称相似关系等进行了一些研究。在形式概念分析中 Liu J 等在文献[9]中将多值形式背景转变为单值形式背景后,通过把不完备属性在不同对象上的不同取值进行扩展,从而得到了完备的形式背景来进行概念的提取。Dubois D 等在文献[10]提出了利用概率论来解决不完备形式背景的方法。Krupka M 等在文献[11]中定义了不完备的模糊形式背景,然后提出了在不完备模糊形式背景下构建概念格的方法。Djouadi Y 等在文献[12]中将不完备模糊形式背景中的隶属度值均采用区间值来表示,将不完备模糊形式背景转化为区间值模糊形式背景(interval-valued fuzzy formal concept, IVFF),在此基础上提出了基于区间值形式背景的概念格构建方法。Li J H 等在文献[13]中提出了在不完备形式背景下构建相似概念格的方法,此外基于相似概念格还研究了在不完备的决策形式背景下获取规则的方法。上述研究中,无论是将不完备形式背景转化为区间值形式背景,还是对不完备属性进行扩展来构造概念格的方法,仅仅适用于形式背景中数据量缺失较少的情况。当形式背景中数据缺失量较大时,所构造的概念格中包含有大量不确定的信息,这对知识获取造成了很大影响。

本文为了减少形式背景中数据缺失量对知识获取的影响,提出并定义了相似模糊概念和相容模糊概念并给出了相容模糊概念的构建方法,建立了相容模糊概念之间的偏序关系,进而设计面向模糊不完备信息的关联规则挖掘算法。

1　基本概念

1.1　形式概念分析

定义 1[1]　一个形式背景 $K=(G,M,I)$ 是一个三元组 ,其中 G 是对象集合,M 是属性集合, I 是 G 与 M 之间的一个二元关系 gIm 或 $(g,m)\in I$,表示对象 g 具有属性 m。在形式背景中定义式(1)和式(2):

$$F(A)=\{m\in M\mid gIm,\forall g\in A\},A\subseteq G \quad (1)$$

$$G(B)=\{g\in G\mid gIm,\forall m\in B\},B\subseteq M \quad (2)$$

形式背景能用一个二维表表示,如表 1,其中对象集 $G=\{x_1,x_2,x_3,x_4\}$,属性集 $M=\{a,b,c,d\}$ 。表中 1 表示某对象拥有某属性,0 表示某对象不拥有某属性,例如 x_1 有 a 属性,没有 b 属性。

表 1　形式背景

Table 1　A formal context

G	a	b	c	d
X_1	1	0	0	1
X_2	0	1	1	0
X_3	1	0	1	0
X_4	0	0	1	0

定义 2[1]　形式背景 $K=(G,M,I)$ 中一对二元组 (A,B) 称为形式概念,当且仅当 $F(A)=B$ 与 $G(B)=A$ 同时成立($A\subseteq G,B\subseteq M$),其中 A 叫做形式概念的外延,B 叫做形式概念的内涵。

假定 (A_1,B_1) 与 (A_2,B_2) 是形式背景 (G,M,I) 下的两个概念,这两个概念间可以建立起偏序关系 $(A_1,B_1)\leqslant(A_2,B_2)\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_2\subseteq B_1)$ 。领先次序意味着 (A_1,B_1) 是 (A_2,B_2) 的子概念。根据概念间的偏序关系生成格的 Hasse 图见图 1。下面是在形式背景 K 下生成的概念:

1) $\{\varnothing,(a,b,c,d)\}$;
2) $\{(x_1),(a,d)\}$;
3) $\{(x_3),(a,c)\}$;
4) $\{(x_2),(b,c)\}$;
5) $\{(x_1,x_3),(a)\}$;
6) $\{(x_2,x_3,x_4),(c)\}$;
7) $\{(x_1,x_2,x_3,x_4),\varnothing\}$。

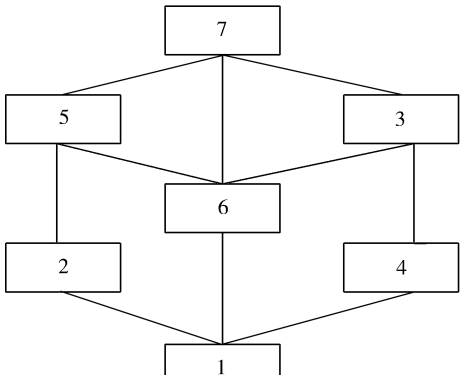


图 1　表 1 对应概念格的 Hasse 图

Fig.1　Hasse diagram of table 1

1.2 模糊形式概念

定义 3[14] 一个模糊形式背景是一个三元组(G',M',I'),其中G'是对象的有限集,M'是属性有限集,I'是$G'\times M'$的模糊集合。$(g,m)\in I'$有一个隶属度值$u(g,m)\in[0,1]$。

定义 4[14] 给定一个模糊形式背景$K'=\{G',M',I'=\varphi(G'\times M')\}$和一个置信度阈值$T=[t_1,t_2]$,在形式背景中定义式(3)与式(4):

$$FA(A)=\{m\in M'\mid \forall g\in A:t_1\leqslant u(g,m)\leqslant t_2\} \tag{3}$$

式中$A\subseteq G'$。

$$FO(B)=\{g\in G'\mid \forall m\in B:t_1\leqslant u(g,m)\leqslant t_2\} \tag{4}$$

式中$B\subseteq M'$。

模糊形式背景(G',M',I')同置信度阈值T下的一个二元组(A,B)($A\subseteq G',B\subseteq M'$)是模糊形式概念,当且仅当$FA(A)=B$与$FO(B)=A$同时成立。$A$、$B$分别叫做模糊形式背景的模糊外延和模糊内涵。

定义 5[14] (A_1,B_1)和(A_2,B_2)是形式背景(G',M',I')的两个模糊概念。(A_1,B_1)是(A_2,B_2)的子概念,记作$(A_1,B_1)\leqslant(A_2,B_2)$,当且仅当$A_1\subseteq A_2(\Leftrightarrow B_2\subseteq B_1)$。

目前所研究的形式背景是完备的,换句话讲,此时对象或者具有某属性,或者不具有某属性,他们之间的关系是确定的。数据缺失现象在生活中普遍存在。例如,对一些突发事件,并没有该事件的完整记录;再如病人突发疾病,而不能对病人进行全面检查,然后来制定相应的治疗方案。下面给出一个例子来说明,表 2 是医生诊断表,即为不完备模糊形式背景,其中o_1、o_2、o_3、o_4表示病人编号,组成对象集G'。a、b、c、d、e、f表示病人的症状,其代表为头痛、血压、恶心、食欲不振、咳嗽、乏力,组成属性集M'。用$*$来表示缺失数据,但是这些数据是客观存在的。

表 2 初始模糊形式背景

Table 2 The initial fuzzy formal context

	a	b	c	d	e	f
o_1	0.8	0.1	0.61	0.6	0.8	*
o_2	0.9	0.85	*	0.2	0.7	0.9
o_3	0.21	*	0.87	*	0.6	0.6
o_4	0.6	*	0.30	*	0.5	0

一个置信度阈值 T 设置在区间$[t_1,t_2]$中。通过设置置信度阈值可以消除一些不在这个值之内的关系,对于t_1与t_2的值用户可以根据需要来设定。例如在表 2 中设定模糊形式背景的置信度阈值为 T=[0.5,1],对于表中(o_1,b)的隶属度值为 0.1,认为病人o_1的血压没有问题,可以不考虑。

表 3 置信度阈值为 *T* 的模糊形式背景

Table 3 Confidence thresholds for *T* fuzzy formal context

	a	b	c	d	e	f
o_1	0.8	0	0.61	0.6	0.8	*
o_2	0.9	0.85	*	0	0.7	0.9
o_3	0	*	0.87	*	0.6	0.6
o_4	0.6	*	0	*	0.5	0

2 相似模糊概念与相容模糊概念

在形式概念分析中,对不完备形式背景进行完备化处理,一般可采用以下 3 种方法。

1)删除法。删除法即删除形式背景中缺失数据的一列或者一行,也就是删除一个对象或者删除一个属性。这类方法操作起来比较简单,但是在删除过程中会导致原先存在的数据缺失,可能会造成获取的知识不准确。

2)填补法。填补法就是对不完备形式背景中缺失的数据填充为 1 或者 0,使之补全为一个完备的形式背景。这类方法比较简单,但是容易造成获取的知识错误,因为好多缺失信息都是人为地填充 0 或者 1。

3)扩展属性法[15]。扩展属性法即把原有不完备形式背景下的属性集合中的属性分为完备和不完备属性两部分,然后将不完备属性在不同对象的不同取值进行扩展,从而把不完备形式背景补充完整。此方法的好处是既不会增加知识也不会缺失知识,但是增加了知识获取的时间和空间复杂度。

定义 6 在不完备模糊形式背景$K_c=(G',M',I_M)$中,对于集合$A\in G'$,记作:

$$\underline{FA(A)}=\{m\in M'\mid \forall g\in A:t_1\leqslant u(g,m)\leqslant t_2 \text{ 或 } u(g,m)=*\} \tag{5}$$

式中$A\in G'$。

$$\underline{FO(B)}=\{g\in G\mid \forall m\in B:t_1\leqslant u(g,m)\leqslant t_2 \text{ 或 } u(g,m)=*\} \tag{6}$$

式中$B\subseteq M'$。

如果$\underline{FA(A)}=B$且$\underline{FO(B)}=A$称(A,B)为模糊形式背景K_c下的一个相似模糊概念,$g\in A$时u_g为(A,B)中对象g的隶属度值,其表示如式(7):

$$u_g = \min(u(g,m)) \; m \in B \tag{7}$$

特别地当 $u(g,m) = *$ 时可用补全法补全 g 与 m 的隶属度值。在不完备形式背景下所有相似模糊概念构成的集合可表示为 $w(K_c)$。

在相似概念(A,B)中,如果含有大量缺失数据,则涉及(A,B)的任何应用都是不可靠的,即它不仅降低了知识获取的有效性,反而会使不确定性进一步扩散。下面在相似模糊概念的基础上提出了相容模糊概念,通过设置参数(α,β)可满足不同用户的需求,(α,β)就叫做相容参数。

定义 7　在不完备模糊形式背景 $K_c = (G',M',I_M)$ 中,设 $A \subseteq G', B \subseteq M', (A,B) \in w(K_c)$, $0 \leqslant \alpha,\beta \leqslant 1$,记作:

$$\chi(A,B) = \{a \in B \mid |(A,B)_a| \geqslant \alpha \times |A|\} \tag{8}$$

$$\varphi(A,B) = \{x \in A \mid |(A,B)^x| \geqslant \alpha \times |B|\} \tag{9}$$

$$\gamma(A,B) = \frac{1}{|A| + |B|}(|\chi(A,B)| + |\varphi(A,B)|) \tag{10}$$

式中:$(A,B)_a = \{x \in A \mid t_1 \leqslant u(x,a) \leqslant t_2\}$, $(A,B)^x = \{a \in B \mid t_1 \leqslant u(x,a) \leqslant t_2\}$,在相似模糊概念$(A,B)$中,$\chi(A,B)$表示属性 $a(\forall a \in B)$ 在 K_c 中满足集合$(A,B)_a$ 中元素数量大于 $\alpha \times |A|$的属性集合。$\varphi(A,B)$表示为对象 $x(\forall x \in A)$ 在 K_c 满足集合 $(A,B)^x$ 中元素数量大于 $\alpha \times |B|$的对象集合。

如果 $\gamma(A,B) \geqslant \beta$ 则称(A,B)是相容模糊概念,定义 $\bar{u}$ 为(A,B)的隶属度值,$\bar{u}$ 可以表示为

$$\bar{u} = \frac{\sum_{g \in A} u_g}{|A|} \tag{11}$$

在不完备模糊形式背景 K_c 中,基于参数(α,β)的所有相容概念构成的集合为 $w_\beta^\alpha(K_c)$。不完备模糊形式背景 K_c 用补全法转化为完备的模糊形式背景,其上获得的相似模糊概念 $w(K_c)$ 中有许多填充的信息,通过参数 α 与外延中对象数量与内涵中属性数量的乘积,即 $\alpha \times |A|$与 $\alpha \times |B|$可以去掉填充值较大概念。

定义 8　如果在一个相容模糊概念中有 $u(g,m) = *$ 则这个相容模糊概念称为粗糙概念,反之称为精确概念。

定理 1　在不完备模糊形式背景 $K_c = (G',M',I_M)$ 中,如果(A,B)是粗糙概念,那么这个相容模糊概念的子概念至少存在一个概念,其粗糙度为

$$\frac{|u(g,m)_{(\forall g \in A)} = *|}{|A| \times |B|} \tag{12}$$

证明　假设(A_1,B_1)是(A,B)的一个子概念,(A,B)是粗糙相容模糊概念,即存在 $u(g,m) = *$,根据概念之间的继承关系可知 $g \in A_1, m \in B_1$。

相似模糊概念与相容模糊概念既有区别也有联系,在经典的不完备形式背景中"补全法"将缺失数据补充为 1,而在不完备的模糊形式背景中,相似模糊概念是将不完备模糊形式背景中的缺失数据补充为 0.5 得到的。而相容模糊概念是对相似模糊概念的扩展,它是在相似模糊概念基础上通过设置参数(α,β),去除一些数据量缺失较大的相似模糊概念而得到的。根据定义 6 和定义 7 以及传统的概念获取算法[16],可以得出相似模糊概念和相容模糊概念的构造算法,具体算法步骤参考算法 1 与算法 2。

算法 1　在不完备形式背景 K_c 中,相似模糊概念的构造算法。

输入　不完备模糊形式背景 $K_c = (G',M',I_M)$, $w(K_c)$为空集。

输出　相似模糊概念 $w(K_c)$。

1)先对不完备模糊形式背景进行处理,如果 $u(g,m)$小于置信度阈值 T,则 $u(g,m)$为 0,然后将 K_c 中的空缺数据 $*$ 全部填充为 0.5,即用补全法把不完备形式背景转化为完备形式背景。

2)获得第一个概念$(FO(M),M)$设置概念的隶属度值并加入 $w(K_c)$中。

3)遍历对象 g,其中 $g \in G$,如果遍历完成转到 6),反之转到 4)。

4)遍历近似模糊概念(A,B),其中$(A,B) \in w(K_c)$,如果遍历完成转到 3),否则转到 5)。

5)求出 B 与 $FA(g)$交集 I,如果获得的交集 I 不是已获得 $w(K_c)$的内涵,则计算出$(FO(I),I)$隶属度值并加入 $w(K_c)$中然后回到 4)。

6)输出 $w(K_c)$,算法结束。

算法 2　在不完备形式背景 K_c 中,相容模糊概念的获取算法。

输入　不完备模糊形式背景 $K_c = (G',M',I_M)$, 相似模糊概念 $w(K_c)$, $w_\beta^\alpha(K_c)$为空集。

输出　相容模糊概念 $w_\beta^\alpha(K_c)$。

1)任取 $w(K_c)$里的相似模糊概念并计算出 $\chi(A,B)$、$\varphi(A,B)$与 $\gamma(A,B)$。如果 $\gamma(A,B) > \beta$,计算出(A,B)的隶属度值 $\bar{u}$ 并加入到相容模糊概念 $w_\beta^\alpha(K_c)$中。

2)如果相似模糊概念都被进行计算过,则输出相容模糊概念转到 3),反之再进行 1)。

3)输出 $w_\beta^\alpha(K_c)$,算法结束。

3 基于相容模糊概念的规则提取

关联规则数据挖掘中最活跃的研究方法之一[17-21]。规则就是形如“如果…那么…(If…Then…)”前者为条件,后者为结果。典型的关联规则发现问题是对超市中的购物篮数据进行分析,例如最著名的案例就是啤酒与尿布。

对于关联规则 $A\Rightarrow B$ 的支持度是 $\mathrm{sup}p(A\Rightarrow B)=|FO(A\cup B)|/|U|$,置信度为 $\mathrm{conf}(A\Rightarrow B)>\beta$ 关联规则 $A\Rightarrow B$ 被称为关于 (ω,τ) 关联规则,当 $\mathrm{sup}p(A\Rightarrow B)>\omega$ 时把 $A\cup B$ 称为频繁的。

在不完备的模糊背景下,规则提取是一件比较困难的工作,在之前的工作中已经获得了相似模糊概念和相容模糊概念,然后根据算法 3 构造好相似模糊概念格,但是由于相似模糊概念中有许多不确定性信息,所以构造的相似模糊概念格也是不准确的。通过对相似模糊概念的筛选,最后得到了较为准确的相容模糊概念,可以在构造好的相似模糊概念格基础上得到相容模糊概念的之间的偏序关系,从而可以提取出可信度较高的关联规则,具体算法参考算法 4。

算法 3 相似模糊概念格的构造算法。

输入 不完备模糊形式背景 $K_c=(G',M',I_M)$,相似模糊概念 $w(K_c)$。

输出 相似模糊概念格。

1)遍历相似模糊概念 (A,B),其中 $(A,B)\in w(K_c)$,并且设置 (A,B) 的 count 为 0。如果遍历完成则转 4),否则转 2)。

2)遍历属性 m,其中 $m\in M$,并求得 A 与 $FO(m)$ 交集 I。如果遍历结束转到 1),否则转到 3)。

3)在 $w(K_c)$ 找出相似模糊概念 (A_1,B_1) 使得 $A_1=I$,并把概念 (A_1,B_1) 的 count 值加 1。假如 $|B_1|-|B|$ 等于 (A_1,B_1) 的 count 值,则增加边在 (A_1,B_1) 与 (A,B),反之转 2)。

4)输出相似模糊概念格,算法结束。

算法 4 不完备形式背景下规则提取的算法。

输入 不完备模糊形式背景 $K_c=(G',M',I_M)$,相容模糊概念 $w_\beta^\alpha(K_c)$。

输出 关联规则 $\sum$。

1)对相似模糊概念格进行处理,除去相容模糊概念之外的概念,更新父节点和子节点。

2)如果概念 $C_1=(A_1,B_1)$ 与 $C_2=(A_2,B_2)$ 满足 C_2 是 C_1 的父节点,且满足 C_1 与 C_2 的隶属度大于给定的阈值 η,即 $\dfrac{\min(\bar{u}(A_1,B_1),\bar{u}(A_2,B_2))}{\max(\bar{u}(A_1,B_1),\bar{u}(A_2,B_2))}>\eta$,则可以将得到的规则 $B_2=B_1-B_2$ 加入 $\sum$ 中,其可信度是 $|A_1|/|A_2|$。

3)如果对于节点 $C=(A,B)$ 有多个双亲节点,则任取两个双亲节点 $C_1=(A_1,B_1)$ 与 $C_2=(A_2,B_2)$,如果满足条件 $\dfrac{\min(\bar{u}(A_1,B_1),\bar{u}(A_2,B_2))}{\max(\bar{u}(A_1,B_1),\bar{u}(A_2,B_2))}>\eta$,则可以提取到的规则 $B_1\Rightarrow B_2$ 与 $B_2\Rightarrow B_1$,并将其加入 $\sum$ 中,支持度分别为 $|A|/|A_1|$ 与 $|A|/|A_2|$。

4)输出 $\sum$。

在得到提取规则 $\sum$ 后,可以给其支持度阈值 ω 与置信度阈值 τ。然后根据需要从提取的规则中筛选出符合要求的规则。

4 示例展示

现在举例来展示规则提取的过程,在表 3 中讨论的置信度阈值是 $T=[0.5,1]$,通过算法 1,可以得出在表 3 的不完备模糊背景下形成的相似模糊概念为

1) $\{\varnothing,(a,b,c,d,e,f)\}$;
2) $\{(0.6/o_1),(a,c,d,e,f)\}$;
3) $\{(0.5/o_2),(a,b,c,e,f)\}$;
4) $\{(0.61/o_1,0.5/o_2),(a,c,e,f)\}$;
5) $\{(0.5/o_3),(b,c,d,e,f)\}$;
6) $\{(0.6/o_1,0.5/o_3),(c,d,e,f)\}$;
7) $\{(0.5/o_2,0.5/o_3),(b,c,e,f)\}$;
8) $\{(0.61/o_1,0.5/o_2,0.6/o_3),(c,e,f)\}$.;
9) $\{(0.5/o_4),(a,b,d,e)\}$;
10) $\{(0.6/o_1,0.5/o_4),(a,d,e)\}$;
11) $\{(0.7/o_2,0.5/o_4),(a,b,e)\}$;
12) $\{(0.8/o_1,0.7/o_2,0.5/o_4),(a,e)\}$;
13) $\{(0.5/o_3,0.5/o_4),(b,d,e)\}$;
14) $\{(0.6/o_1,0.5/o_3,0.5/o_4),(d,e)\}$;
15) $\{(0.7/o_2,0.5/o_3,0.5/o_4),(b,e)\}$;
16) $\{(o_1,o_2,o_3,o_4),(0/a,0/b,0/c,0/d,0.5/e,0/f)\}$。

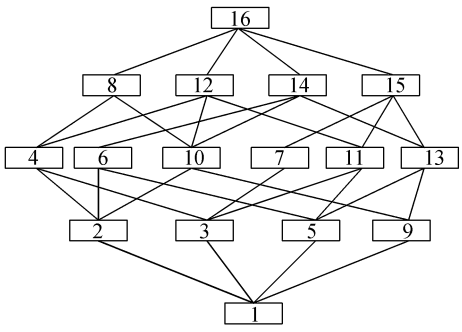


图 2 相似模糊概念的相似模糊概念格

Fig.2 Approximat fuzzy concept lattice

图 2 是相似模糊概念格对应的Hasse 图。上述

已经得出不完备形式背景下(表 3)的相似模糊概念,然后在此基础上根据算法 3 构建出相似模糊概念格,通过设置相容参数($\alpha=0.6,\beta=0.8$)来获得所需要的可靠的相容模糊概念 $w_{0.8}^{0.6}(K_c)$:

1) $\{0.6/(o_1),(a,c,d,e,f)\}$;

2) $\{0.5/(o_2),(a,b,c,e,f)\}$;

3) $\{0.57/(o_1,o_2,o_3),(c,e,f)\}$;

4) $\{0.55/(o_1,o_4),(a,d,e)\}$;

5) $\{0.6/(o_2,o_4),(a,d,e)\}$;

6) $\{0.667/(o_1,o_2,o_4),(a,e)\}$;

7) $\{0.083/(o_1,o_2,o_3,o_4),(c)\}$。

根据算法 4 可以得出阈值 τ 为 0.9,置信度阈值为 0.5 的关联规则:

1) $\{a,d,e\}\Rightarrow\{c,f\}$ 置信度为 0.5。

解释:如果头疼、食欲不振、咳嗽则恶心、乏力。

2) $\{a,e\}\Rightarrow\{b\}$ 置信度为 0.67。

解释:如果头疼、咳嗽则血压会高。

5　实验结果与分析

本文基于相容模糊概念的关联规则提取可分为在不完备模糊形式背景中相容模糊概念的构造过程和根据相容模糊概念的偏序关系进行提取规则的过程。本文算法在 Win7 环境下用 MATLAB 来实现,并在 UCI 数据库的 water 数据集(526 个对象,38 个属性)进行实验。实验主要对 2 个指标进行测量:第 1 个是在不完备模糊形式背景下相似模糊概念数目与对象数目以及相容模糊概念与对象数目之间的关系;第 2 个是提取的关联规则数目与对象数目之间的关系。在本实验中针对不完备模糊形式背景,设定相容模糊参数为($\alpha=0.8,\beta=0.9$),关联规则的置信度阈值为 0.8。在不完备形式背景下相似模糊概念与相容模糊概念的数量关系可由图 3 体现,图 4 则体现了对象数目与关联规则数量之间的关系。图 3 与图 4 都在相容模糊参数与属性数量都不变的情况下,取 water 数据集中的前 200 个对象进行测量。图 3 与图 4 中横坐标都表示对象的数量,初始为 0,分别一次递增 50 与 10 进行测试。图 3 纵坐标表示由不完备形式背景形成概念的数量,图 4 纵坐标表示由相容模糊概念获得的关联规则的数量。通过图 3、4 的实验结果可以观察到,在不完备模糊形式背景中随着对象数目的增多,通过本算法获得知识准确性与传统的方法相比具有一定的优势。

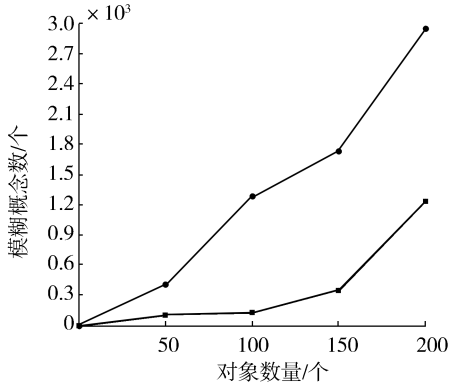


图 3　对象与概念个数关系

Fig.3　Relationship between object and concept

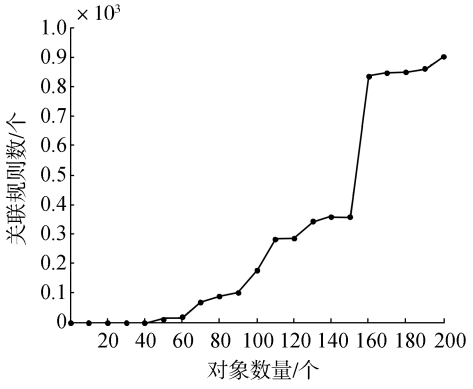


图 4　对象与关联规则个数关系

Fig.4　Relationship between object and association rule

6　结束语

目前在不完备模糊形式背景下的研究越来越多。本文结合在传统的不完备形式背景下获取概念的方法,提出了在不完备模糊形式背景中提取出相容模糊概念,并根据相似模糊概念格提取出相容规则的方法。实验表明,该方法可以有效的降低形式背景中因数据缺失和数据的模糊性对获取知识准确性带来的的影响。未来的工作还需要改进和细化文中的一些算法,例如如何在知识库分类能力保持不变的情况下删除不相关的冗余属性;如何把模糊概念格与粗糙集理论有效结合以解决不确定规则提取中的规则冗余性等问题。

参考文献:

[1] WILLE R. Restructuring lattice theory: an approach based on hierarchies of concepts[M]//RIVAL I. Ordered sets. Netherlands: Springer, 1982: 445-470.

[2] POELMANS J, IGNATOV D I, KUZNETSOV S O, et al. Formal concept analysis in knowledge processing: a survey on applications[J]. Expert systems with applications,

2013, 40(16): 6538-6560.

[3] MINEAU G W, GODIN R. Automatic structuring of knowledge bases by conceptual clustering[J]. IEEE transactions on knowledge and data engineering, 1995, 7(5): 824-829.

[4] COLE R, EKLUND P W. Scalability in formal concept analysis[J]. Computational intelligence, 1999, 15(1): 11-27.

[5] CARPINETO C, ROMANO G. A lattice conceptual clustering system and its application to browsing retrieval[J]. Machine learning, 1996, 24(2): 95-122.

[6] MA Jianmin, ZHANG Wenxiu. Axiomatic characterizations of dual concept lattices[J]. International journal of approximate reasoning, 2013, 54(5): 690-697.

[7] 胡明涵, 张莉, 任飞亮. 模糊形式概念分析与模糊概念格[J]. 东北大学学报:自然科学版, 2007, 28(9): 1274-1277.
HU Minghan, ZHANG Li, REN Feiliang. Fuzzy formal concept analysis and fuzzy concept lattices[J]. Journal of northeastern university : natural science, 2007, 28(9): 1274-1277.

[8] GRZYMALA-BUSSE J W. Rough set approach to incomplete data[C]//Proceedings of the 7th international conference on artificial intelligence and soft computing-ICAISC 2004. Berlin Heidelberg, Germany, 2004: 50-55.

[9] LIU Jun, YAO Xiaoqiu. Formal concept analysis of incomplete information system[C]//Proceedings of the 7th international conference on fuzzy systems and knowledge discovery. Yantai, China, 2010, 5: 2016-2020.

[10] DJOUADI Y, DUBOIS D, PRADE H. Possibility theory and formal concept analysis: Context decomposition and uncertainty handling[C]//Proceedings of the 13th international conference on information processing and management of uncertainty. Berlin Heidelberg, Germany, 2010: 260-269.

[11] KRUPKA M. Fuzzy concept lattices with incomplete knowledge[C]//Proceedings of the 14th international conference on information processing and management of uncertainty in knowledge-based systems. Berlin Heidelberg, Germany, 2012: 171-180.

[12] DJOUADI Y, PRADE H. Interval-valued fuzzy formal concept analysis[C]//Proceedings of the 18th international symposium. Berlin Heidelberg, Germany, 2009: 592-601.

[13] LI Jinhai, MEI Changlin, LV Yuejin. Incomplete decision contexts: approximate concept construction, rule acquisition and knowledge reduction[J]. International journal of approximate reasoning, 2013, 54(1): 149-165.

[14] LAI Hongliang, ZHANG Dexue. Concept lattices of fuzzy contexts: formal concept analysis vs. rough set theory[J]. International journal of approximate reasoning, 2009, 50(5): 695-707.

[15] 何淑贤, 王育红, 翟岩慧, 等. 不完备形式背景及其完备化方法[J]. 山西大学学报:自然科学版, 2006, 29(4): 364-367.
HE Shuxian, WANG Yuhong, ZHAI Yanhui, et al. Incomplete formal context and the completion approach[J]. Journal of Shanxi university : natural science edition, 2006, 29(4): 364-367.

[16] 谢志鹏, 刘宗田. 概念格的快速渐进式构造算法[J]. 计算机学报, 2002, 25(5): 490-496.
XIE Zhipeng, LIU Zongtian. A fast incremental algorithm for building concept lattice[J]. Chinese journal of computers, 2002, 25(5): 490-496.

[17] 梁吉业, 王俊红. 基于概念格的规则产生集挖掘算法[J]. 计算机研究与发展, 2004, 41(8): 1339-1344.
LIANG Jiye, WANG Junhong. An algorithm for extracting rule-generating sets based on concept lattice[J]. Journal of computer research and development, 2004, 41(8): 1339-1344.

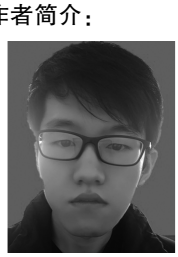
[18] LEKHA A, SRIKRISHNA C V, VINOD V. Fuzzy association rule mining[J]. Journal of computer science, 2015, 11(1): 71-74.

[19] LAKHAL L, STUMME G. Efficient mining of association rules based on formal concept analysis[M]//GANTER B, STUMME G, WILLE R. Formal concept analysis. Berlin Heidelberg: Springer-Verlag, 2005: 180-195.

[20] KUMAR CH A, DIAS S M, VIEIRA N J. Knowledge reduction in formal contexts using non-negative matrix factorization[J]. Mathematics and computers in simulation, 2015, 109: 46-63.

[21] 王志海, 胡可云, 胡学纲, 等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报, 1991, 22(1): 66-70.
WANG Zhihai, HU Keyun, HU Xuegang, et al. General and incremental algorithms of rule extraction based on concept lattice[J]. Chinese journal of computers, 1991, 22(1): 66-70.

作者简介:



胡小康, 男, 1991 年生, 硕士研究生, 主要研究方向为形式概念分析、数据挖掘。

王俊红, 女, 1979 年生, 副教授, 主要研究方向形式概念分析、粗糙集和数据挖掘。主持或参与多项国家 863 计划、国家自然科学基金和省部级等科研项目。发表学术论文 10 余篇。