文章编号: 1001-0920(2019)05-1011-07 DOI: 10.13195/j.kzyjc.2017.1537

基于概念格的不完备信息系统最简规则提取算法

陈泽华[1†], 宋 波[2], 闫继雄[2], 柴 晶[2]

(1. 太原理工大学 大数据学院，太原 030024；2. 太原理工大学 信息工程学院，太原 030024)

摘 要: 概念格是以概念为元素的偏序集,通常可以对形式背景描述的完备信息系统进行分析和处理,然而在多数情况下信息系统是不完备的,粗糙集理论是一种刻画不完整、不确定性问题的有效方法. 针对此问题,从粗糙集的角度出发,基于概念格理论定义一种描述不完备信息系统的增广形式背景,在此基础上,定义并讨论极概念和极概念格及其相关性质,进而提出增广形式背景的极概念生成算法. 为了获得更加简洁的决策规则,同时提出一种新的无冗余属性的决策规则获取算法. 通过实例计算和UCI数据集的对比实验,表明了所提出算法的可行性和有效性,特别地,当信息系统完备时极概念将退化为经典的概念.

关键词: 概念格；粗糙集；不完备信息系统；极概念；极概念格；规则提取

中图分类号: TP273 **文献标志码:** A

Concise rule extraction algorithm of incomplete information system based on concept lattice

CHEN Ze-hua[1†], *SONG Bo*[2], *YAN Ji-xiong*[2], *CHAI Jing*[2]

(1. College of Data Science，Taiyuan University of Technology，Taiyuan 030024，China；2. College of Information Engineering，Taiyuan University of Technology，Taiyuan 030024，China)

Abstract: Concept lattice is a partial order set of concept elements, which can be used to analyze and process the complete information system described by a formal context. However, in most cases, the information system is incomplete. The rough set theory is an effective mathematical tool to deal with incompleteness and uncertainty. From the perspective of rough sets, an augmented formal context is defined to describe an incomplete information system. Then, the extreme concept and the extreme concept lattice are respectively defined on the basis of the concept lattice theory, and their properties are discussed. Furthermore, an extreme concept generating algorithm from the augmented formal context is proposed. In order to obtain a more concise decision rule, a new decision rule obtaining algorithm without redundant attributes is also proposed. Finally, calculation example and comparition of UCI dataset verify the feasibility and effectiveness of the proposed algorithm, particularly, when the incomplete information system becomes complete, the extreme concept degenerates into the classical concept.

Keywords: concept lattice；rough set；incomplete information system；extreme concept；extreme concept lattice；rule acquisition

0 引 言

经典粗糙集理论是波兰数学家Pawlak[1]提出的处理不确定性问题的数学方法,近似算子[2]是研究粗糙集的有效工具. 完备信息系统的知识约简是其研究的核心问题之一,而不完备信息系统更具一般性,因此受到越来越多学者的关注. Kryszkiewicz[3]在不减少原数据信息的基础上提出了不完备信息系统的知识约简方法. 文献[4-5]通过定义分辨函数获取不完备信息系统的决策规则. Qian等[6]针对不一致不完备决策表提出了一种保持决策类上、下近似分布的上、下近似约简的属性约简方法. 杨习贝等[7]提出了可变精度分类的拓展模型,直接从不完备信息系统中获取确定与可能性规则. 程玉胜等[8]提出了针对不完备信息系统规则提取的快速矩阵算法,将条件属性相容矩阵和决策属性分配决策矩阵合并,提高了算法的执行效率. 文献[9]提出了一种有效的计算属性缺失块方法,该方法为属性约简、规则提取提供了新的思路. Li等[10]基于区间集模型思想,从不完备信息

收稿日期: 2017-11-14；**修回日期:** 2018-01-09.

基金项目: 国家自然科学基金项目(61402319, 61403273)；山西省自然科学基金项目(2014021022-4).

作者简介: 陈泽华(1974−), 女, 教授, 博士, 从事粒计算、智能信息处理、工业大数据等研究；宋波(1991−), 男, 硕士生, 从事粒计算、形式概念分析的研究.

†通讯作者. E-mail: zehuachen@163.com.

系统中获取规则. 纪霞等[11]通过定义属性分辨度,进行不完备信息系统的规则提取,效率相比于LEM2算法更高,规则更简洁.

形式概念分析,也称为概念格,是德国数学家Wille[12]于1982年提出的一种在形式背景中进行数据分析和规则提取的强有力工具. 形式概念分析与粗糙集二者之间有很多关联,国内外学者对于二者之间的关系作了深入研究[13-15]. 文献[16]首次提出决策形式背景的概念,为条件属性与决策属性之间的规则推理提供了新的研究方向. 杨凯等[17]在文献[16]的基础上,针对完备决策表给出了基于概念格的多层属性约简算法,提出相容概念和内涵亏值,为概念格中属性约简提供了一种新的方法,但并没有进行规则提取方面的工作. Li等[18-19]针对完备与不完备决策形式背景,分别构建概念格与近似概念格获取决策规则,再通过分辨矩阵以及布尔函数去除冗余属性,最终得到较为简洁的决策规则,但冗余规则的判定较为繁琐. 文献[20]将不完备信息系统转换为一种特殊的形式背景,基于相容概念和相容概念格获取规则,但相容概念的求取过程比较复杂,获取的决策规则中存在冗余. 李金海等[21]将遗传算法运用到形式背景的规则提取中,尽管获取的规则更加的紧凑,但该算法并不能保证得到最小约简. 翟岩慧等[22]提出了决策蕴含规范基,能够较好地抑制冗余决策的生成,但是无法有效地计算决策前提,并且在某些情况下无法确定生成规则的准确性. 邵明文等[23]提出了针对多值形式背景的属性算法,但是复杂度高,且没有生成规则集.

针对不完备信息系统,本文首先将其扩展为增广形式背景,在此基础上提出极概念与极概念格并讨论极概念的性质和生成算法. 为了克服传统概念格规则提取算法中存在冗余规则的问题,同时提出一种新的获取非冗余决策规则的算法. 最后,通过实例计算和UCI数据集实验测试,表明了所提出算法的正确性和有效性.

1 预备知识

下面首先介绍形式概念分析的相关概念和性质.

定义1 [24] 形式背景用$K=(U,A,I)$表示,其中U表示对象集,A表示属性集,$I\subseteq U\times A$表示对象与属性之间的二元关系. 对于$\forall u\in U,a\in A,(u,a)\in I$表示对象$u$具有属性$a$,$(u,a)\notin I$表示对象$u$不具有属性$a$.

定义2 [24] 在形式背景$K=(U,A,I)$中,定义概念的内涵和外延:对于$\forall P\subseteq U,Q\subseteq A$,定义以下关系:

$$P^{\uparrow}=\{a\in A|\forall u\in P,(u,a)\in I\},\tag{1}$$

$$Q^{\downarrow}=\{u\in U|\forall a\in Q,(u,a)\in I\}.\tag{2}$$

则$P^{\uparrow}$定义为集合P的内涵,表示同时具有对象P的最大属性集合;$Q^{\downarrow}$定义为集合Q的外延,表示同时具有属性Q的最大对象集合.

定义3 [25] 在形式背景$K=(U,A,I)$中,对于$\forall P\subseteq U,Q\subseteq A$,若满足$P^{\uparrow}=Q$且$Q^{\downarrow}=P$,则二元组$(P,Q)$称为一个形式概念,其中$P$称为此概念的外延,$Q$称为此概念的内涵. 形式背景$K$中全体概念用$\mathfrak{B}(U,A,I)$表示.

性质1 [12] 形式背景$K=(U,A,I)$,对于$\forall X_1,X_2,X\subseteq U,\forall B_1,B_2,B\subseteq A$,有如下基本性质:

1) $X_1\subseteq X_2\Rightarrow X_2^{\uparrow}\subseteq X_1^{\uparrow},B_1\subseteq B_2\Rightarrow B_2^{\uparrow}\subseteq B_1^{\uparrow}$;

2) $X^{\uparrow}=X^{\uparrow\downarrow\uparrow},B^{\downarrow}=B^{\downarrow\uparrow\downarrow}$;

3) $(X_1\bigcup X_2)^{\uparrow}=X_1^{\uparrow}\bigcap X_2^{\uparrow},(B_1\bigcup B_2)^{\downarrow}=B_1^{\downarrow}\bigcap B_2^{\downarrow}$.

对于$\forall\,(X_1,B_1),(X_2,B_2)\in\mathfrak{B}(U,A,I)$,有

$$(X_1,B_1)\bigwedge(X_2,B_2)=(X_1\bigcap X_2,(B_1\bigcup B_2)^{\downarrow\uparrow}),\tag{3}$$

$$(X_1,B_1)\bigvee(X_2,B_2)=((X_1\bigcup X_2)^{\uparrow\downarrow},B_1\bigcap B_2).\tag{4}$$

2 极概念

下面介绍极概念的相关定义和性质.

2.1 增广形式背景

在容差关系模型[3]中,将信息系统中缺失的属性值用“*”表示,*可以是任意值,本文仅考虑遗漏型语义.

定义4 [3] 四元组$\mathrm{IS}=\{U,\mathrm{AT},V,f\}$称为一个信息系统,$U$表示非空有限对象集,AT表示论域$U$上的非空有限属性集,$V=\bigcup\limits_{a\in\mathrm{AT}}V_a$表示所有属性的取值集合,$V_a$表示属性$a$的取值集合. 若$\exists\ x\in U,a\in\mathrm{AT},\mathrm{s.t.}\ f(x,a)=*$,则称四元组IS为不完备信息系统IIS.

定义5 设不完备信息系统为$\mathrm{IIS}=\{U,\mathrm{AT},V,f\}$,若$\exists x\in U,a\in\mathrm{AT},\mathrm{s.t.}f(x,a)=*$,则对于$\forall x,y\in U,a\in\mathrm{AT}$,有

$$(x\ \text{and}\ y)Ia\Leftrightarrow f(x,a)=f(y,a)\ \text{or}\ f(x,a)=*\ \text{or}\ f(y,a)=*.\tag{5}$$

根据式(5),将不完备信息系统转换成的形式背景称为增广形式背景.

定义6 增广形式背景表示为$k=(U,A,D,\{\times,*\},I,J)$. 其中:$U$为对象集,$A$为条件属性集,$I\subseteq U\times A$为对象与条件属性之间的二元关系,$D$为决策属性集,$J\subseteq U\times D$为对象与决策属性之间的二元关系. 对于$\forall u\in U,a\in A,(u,a)=\times$表示对象$u$具有属性$a$,$(u,a)=*$表示对象$u$可能具有属性$a$.

下面通过例1具体说明如何利用定义5将不完

备信息系统转换为增广形式背景.

例1　不完备信息系统[3]IIS $=\{U,\mathrm{AT},V,f\}$,其中论域$U=\{x_1,x_2,\cdots,x_6\}$,条件属性集$A=\{a,b,c,d\}$,a 为 Price,b 为 Mileage,c 为 Size,d 为 Max-speed. 决策属性集$D=\{y\}$,其中y表示Y. 不完备信息系统如表1所示.

表 1　不完备信息系统

Car	Price	Mileage	Size	Max-speed	Y
x_1	high	low	full	low	good
x_2	low	*	full	low	good
x_3	*	*	compact	high	poor
x_4	high	*	full	high	good
x_5	*	*	full	high	excellent
x_6	low	high	full	*	good

由定义5将不完备信息系统转变为增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,如表2所示. 其中论域$U=\{x_1,x_2,\cdots,x_6\}$,条件属性集$A=\{a_1,a_2,b_1,b_2,c_1,c_2,d_1,d_2\}$,决策属性集$D=\{y_1,y_2,y_3\}$.

表 2　增广形式背景

U	a_1	a_2	b_1	b_2	c_1	c_2	d_1	d_2	y_1	y_2	y_3
x_1	×		×		×		×		×		
x_2		×	*	*	×		×		×		
x_3	*	*	*	*		×		×		×	
x_4	×		*	*	×			×	×		
x_5	*	*	*	*	×			×			×
x_6		×		×	×		*	*	×		

2.2　极概念的定义

定义7　不完备信息系统所形成的增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,由定义5,对于$\forall X\subseteq U$,定义

$$\mathrm{LF}(X)=\{a\in A|\forall x\in X,u(x,a)=\times\},\tag{6}$$

$$\mathrm{LG}(X)=\{x\in U|\forall a\in\mathrm{LF}(X),u(x,a)=\times\bigvee u(x,a)=*\}.\tag{7}$$

其中:$(u,a)=\times$表示对象u具有属性a,$(u,a)=*$表示对象u可能具有属性a;$\mathrm{LF}(X)$表示所有对象X所具有的确定的最大属性集合;$\mathrm{LG}(X)$表示对于$\mathrm{LF}(X)$中的全体属性集所具有的可能对象集合,显然有$X\subseteq\mathrm{LG}(X)$. 对象对$(X,\mathrm{LG}(X))\in 2^U\times 2^U$决定了全体$\mathrm{LF}(X)$所具有的对象范围.

定义8　设不完备信息系统所形成的增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,对于$(X,\mathrm{LG}(X))\in 2^U\times 2^U$,算子$\triangle$: $2^U\times 2^U\to 2^A$和算子∇: $2^A\to 2^U\times 2^U$分别表示为

$$\mathrm{LF}(X)^{\nabla}=(X,\mathrm{LG}(X)),\tag{8}$$

$$(X,\mathrm{LG}(X))^{\triangle}=\{a\in A|u(x,a)=\times\}.\tag{9}$$

称$((X,\mathrm{LG}(X)),\mathrm{LF}(X))$为不完备信息系统所形成的增广形式背景的极概念,增广形式背景$\boldsymbol{k}$中条件属性生成的全体极概念用$\mathfrak{B}(U,A,\{\times,*\},I)$表示. 根据定义3,增广形式背景$\boldsymbol{k}$中的决策属性生成的全体决策概念用$\mathfrak{B}(U,D,\{\times\},J)$表示.

由定义8求得的极概念与经典概念(根据定义3)的区别在于:极概念的外延是近似的,内涵是精确的;经典概念中的外延和内涵都是精确的.

定义9　增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,对于$\forall((X,Y),B)\in\mathfrak{B}(U,A,\{\times,*\},I)$和$(M,C)\in\mathfrak{B}(U,D,\{\times\},J)$,若$Y\subseteq M$且$X$、$Y$、$B$和$M$、$C$非空,则称$((X,Y),B)$蕴含$(M,C)$,记为$((X,Y),B)\to(M,C)$.

定义10　增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,若存在蕴含规则$((X,Y),B)\to(M,C)$,则定义极概念$((X,Y),B)$的启发算子He值为

$$\mathrm{He}=|Y|,\tag{10}$$

其中$|Y|$为Y中元素的个数. 通过启发式算子得到的规则依然含有冗余属性.

2.3　极概念的性质

在介绍极概念的性质前,先给出如下相关定义.

定义11　增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,对于$\forall(X_1,Y_1),(X_2,Y_2)\in 2^U\times 2^U$,若$X_1\subseteq X_2$且$Y_1\subseteq Y_2$,则称$(X_1,Y_1)\subseteq(X_2,Y_2)$.

性质2　增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,对于$\forall(X_1,Y_1),(X_2,Y_2)\in 2^U\times 2^U$且$B_1,B_2\in 2^A$,有如下基本性质:

1) $(X_1,Y_1)\subseteq(X_2,Y_2)\Rightarrow(X_2,Y_2)^{\triangle}\subseteq(X_1,Y_1)^{\triangle}$,$B_1\subseteq B_2\Rightarrow B_2^{\nabla}\subseteq B_1^{\nabla}$;

2) $(X_1,Y_1)^{\triangle}=(X_1,Y_1)^{\triangle\nabla\triangle}$,$B_1^{\nabla}=B_1^{\nabla\triangle\nabla}$;

3) $((X_1,Y_1)\bigcup(X_2,Y_2))^{\triangle}=(X_1,Y_1)^{\triangle}\bigcap(X_2,Y_2)^{\triangle}$,$(B_1\bigcup B_2)^{\nabla}=B_1^{\nabla}\bigcap B_2^{\nabla}$.

证明　1) 若$\exists(X_1,Y_1)\subseteq(X_2,Y_2)$,设$(X_1,Y_1)^{\triangle}=B_1$,$(X_2,Y_2)^{\triangle}=B_2$,则由定义7和定义8可得$B_2\subseteq B_1$,故$(X_2,Y_2)^{\triangle}\subseteq(X_1,Y_1)^{\triangle}$,同理可证$B_1\subseteq B_2\Rightarrow B_2^{\nabla}\subseteq B_1^{\nabla}$.

2) 由式(6)和(7),设$(X_1,Y_1)^{\triangle}=B_1$,又由式(8)和(9),必然存在唯一的$(X_2,Y_2)$使得$B_1^{\nabla}=(X_2,Y_2)$(若$\exists(X_3,Y_3)$使得$B_1^{\nabla}=(X_3,Y_3)$,则与定义6相矛盾),最后根据定义7和定义8,易知$(X_1,Y_1)=(X_2,Y_2)$,综上有$(X_1,Y_1)^{\triangle\nabla}=B_1$,$B_1^{\nabla}=(X_2,Y_2)=(X_1,Y_1)$,$(X_1,Y_1)^{\triangle}=(X_1,Y_1)^{\triangle\nabla\triangle}$,$B_1^{\nabla}=B_1^{\nabla\triangle\nabla}$.

3) $((X_1,Y_1)\bigcup(X_2,Y_2))^{\triangle}=(X_1\bigcup X_2,Y_1\bigcup Y_2)^{\triangle}$,不妨设$a\in(X_1\bigcup X_2,Y_1\bigcup Y_2)^{\triangle}$,对于$\forall(x,y)\in(X_1\bigcup X_2,Y_1\bigcup Y_2)$,有$((x,y),a)=\times$ or $*$. 对于$\forall(x,y)\in(X_1,Y_1)$,或$\forall(x,y)\in(X_2,Y_2)$,有$((x,y),a)=\times$ or $*$.

则对于$\forall(X_1,Y_1)^{\vartriangle},(X_2,Y_2)^{\vartriangle}$,均有$a\in(X_1,Y_1)^{\vartriangle},a\in(X_2,Y_2)^{\vartriangle}$,从而有$a\in(X_1,Y_1)^{\vartriangle}\bigcap(X_2,Y_2)^{\vartriangle}$,$((X_1,Y_1)\bigcup(X_2,Y_2))^{\vartriangle}=(X_1,Y_1)^{\vartriangle}\bigcap(X_2,Y_2)^{\vartriangle}$. 同理可证$(B_1\bigcup B_2)^{\triangledown}=B_1^{\triangledown}\bigcap B_2^{\triangledown}$. □

对于$\forall((X_1,Y_1),B_1)$、$((X_2,Y_2),B_2)\in\mathfrak{B}(U,A,\{\times,*\},I)$,记

$$((X_1,Y_1),B_1)\leqslant((X_2,Y_2),B_2)\Leftrightarrow$$
$$(X_1,Y_1)\subseteq(X_2,Y_2)\Leftrightarrow B_1\supseteq B_2. \tag{11}$$

其中:“$\leqslant$”为$\mathfrak{B}(U,A,\{\times,*\},I)$上的偏序关系,$((X_1,Y_1),B_1)$称为$((X_2,Y_2),B_2)$的亚极概念,$((X_2,Y_2),B_2)$称为$((X_1,Y_1),B_1)$的超极概念.

若有$((X_1,Y_1),B_1)\leqslant((X_2,Y_2),B_2)$,且不存在$((X_3,Y_3),B_3)$,$((X_1,Y_1),B_1)\neq((X_3,Y_3),B_3)\neq((X_2,Y_2),B_2)$,使得$((X_1,Y_1),B_1)\leqslant((X_3,Y_3),B_3)\leqslant((X_2,Y_2),B_2)$,则称$((X_2,Y_2),B_2)$是$((X_1,Y_1),B_1)$的父极概念,$((X_1,Y_1),B_1)$是$((X_2,Y_2),B_2)$的子极概念.

增广形式背景中,父、子极概念体现了极概念之间隐含的信息. 若$((X_1,Y_1),B_1)$是$((X_2,Y_2),B_2)$的子极概念,$((X_2,Y_2),B_2)$称为$((X_1,Y_1),B_1)$的父极概念,则$((X_1,Y_1),B_1)$具有较少的外延、更多的内涵,更为具体,而$((X_2,Y_2),B_2)$具有较多的外延、更少的内涵,更为抽象.

推论1 增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,若$(X_q,Y_q)\in2^U\times2^U$ $(q\in Q)$且$B_t\in2^A$ $(t\in T)$,其中Q和T是两个指标集,则有

$$(\bigcup_{q\in Q}(X_q,Y_q))^{\vartriangle}=\bigcap_{q\in Q}(X_q,Y_q)^{\vartriangle}, \tag{12}$$
$$(\bigcup_{t\in T}B_t)^{\triangledown}=\bigcap_{t\in T}(B_t)^{\triangledown}. \tag{13}$$

定理1 增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,$\mathfrak{B}(U,A,\{\times,*\},I)$是一个完备格,极概念$\{((X_t,Y_t),B_t)|\,t\in T\}$的上、下确界分别为

$$\bigwedge_{t\in T}((X_t,Y_t),B_t)=(\bigcap_{t\in T}(X_t,Y_t),(\bigcup_{t\in T}B_t)^{\triangledown\vartriangle}), \tag{14}$$
$$\bigvee_{t\in T}((X_t,Y_t),B_t)=((\bigcup_{t\in T}(X_t,Y_t))^{\vartriangle\triangledown},\bigcap_{t\in T}B_t). \tag{15}$$

证明 对于$\forall((X_1,Y_1),B_1)$、$((X_2,Y_2),B_2)\in\mathfrak{B}(U,A,\{\times,*\},I)$,根据性质2的3),不妨设

$$((X_1,Y_1),B_1)\bigwedge((X_2,Y_2),B_2)=$$
$$((X_1,Y_1)\bigcap(X_2,Y_2),(B_1\bigcup B_2)^{\triangledown\vartriangle\triangledown}),$$
$$((X_1,Y_1),B_1)\bigvee((X_2,Y_2),B_2)=$$
$$(((X_1,Y_1)\bigcup(X_2,Y_2))^{\vartriangle\triangledown\vartriangle},B_1\bigcap B_2).$$

根据定义6和定义7易知,$((X_1,Y_1)\bigcap(X_2,Y_2),(B_1\bigcup B_2)^{\triangledown\vartriangle\triangledown})$与$(((X_1,Y_1)\bigcup(X_2,Y_2))^{\vartriangle\triangledown\vartriangle},B_1\bigcap B_2)$均为极概念,故$\mathfrak{B}(U,A,\{\times,*\},I)$是一个格. 根据推论1易知,$\mathfrak{B}(U,A,\{\times,*\},I)$是一个完备格. □

3 极概念构造算法

下面介绍增广形式背景中极概念生成算法(算法1),其算法步骤如下.

算法1 增广形式背景的极概念构造.

输入:增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$;

输出:所有极概念e.

Step 1:对于$\forall$单个属性$a\in A$,计算$((X,F(X)),G(X))$,得到第1层极概念.

Step 2:设上层求得的极概念个数为m,将上层极概念$e_1\sim e_{m-1}$的外延$(X,F(X))$作为行,极概念$e_2\sim e_m$的外延$(X,F(X))$作为列,行列相交,得到一个维数为$(m-1)\times(m-1)$的下三角矩阵.

Step 3:记录所有产生的外延中X非空的新外延,并同时计算内涵(定义8),即为下层极概念.

Step 4:重复Step 2、Step 3,直至不再产生新的极概念.

Step 5:输出所有极概念,算法结束.

本文采用例2详细说明算法1的计算过程.

例2 增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$,如表2所示. 根据算法1,得到极概念的外延$(X,\mathrm{LG}(X))\in2^U\times2^U$过程如下:为了便于描述,非空外延(内涵)集只用其元素序列表示,如$\{5,6\}$和$\{c,d\}$简记为$\{56\}$和$\{cd\}$. 第1层极概念的外延为$(14,145),(26,256),(1,1),(6,256),(12\,456,12\,456),(3,3),(12,126),(345,3\,456)$;第2层极概念的外延为$(2,26),(4,45),(45,456)$.

第2层极概念的外延求取过程如表3所示,黑色表示上层极概念中已经存在的外延. 例2得到的所有极概念所对应的极概念格如图1所示.

表3 第2层极概念的外延求取过程

	$(14,145)$	$(26,256)$	$(1,1)$	$(6,256)$	$(12\,456,12\,456)$	$(3,3)$	$(12,126)$
$(26,256)$	∅						
$(1,1)$	**(1,1)**	∅					
$(6,256)$	∅	**(6,256)**	∅				
$(12\,456,12\,456)$	**(14,145)**	**(26,256)**	**(1,1)**	$(6,256)$			
$(3,3)$	∅	∅	∅	∅	∅		
$(12,126)$	**(1,1)**	$(2,26)$	**(1,1)**	∅	**(12,126)**	∅	
$(345,3\,456)$	$(4,45)$	∅	∅	∅	$(45,456)$	**(3,3)**	∅

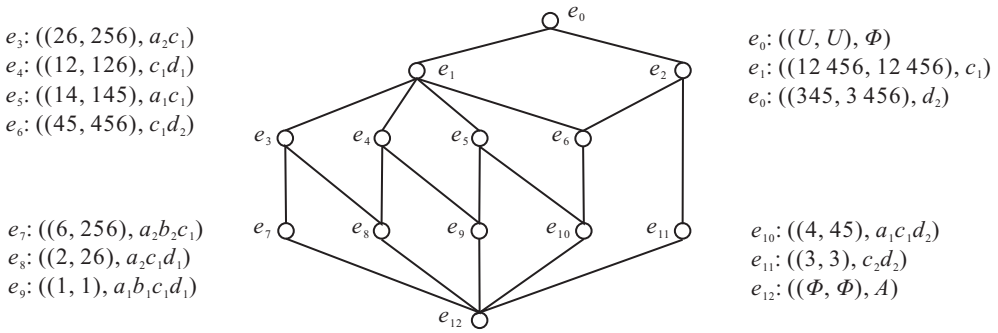

图 1　$\mathfrak{B}(U, A, \{\times, *\}, I)$

4　基于极概念的规则提取算法

4.1　算法描述

基于极概念的性质,本文提出一种不完备决策表的非冗余决策规则提取算法(算法2),其算法步骤如下所示.

算法 2　不完备决策表所形成的增广形式背景的非冗余决策规则提取.

输入:不完备决策表;

输出:最简决策规则.

Step 1:将不完备决策表转化为增广形式背景.

Step 2:求出增广形式背景的所有极概念(根据算法1)和决策概念(根据定义3).

Step 3:寻找所有极概念$((X, \mathrm{LG}(X)), \mathrm{LF}(X))$中满足$\mathrm{LG}(X)$是决策概念外延子集的极概念,并计算He.

Step 4:按照He从大到小对极概念进行排序,并存放在规则表中.

Step 5:识别规则表中的新规则,并记录每类决策中未识别的论域元素.

Step 6:若记录的每类决策中未识别的论域元素都为空,则输出所有的一致规则;否则,将每两类决策中未识别的论域元素合并,形成不一致决策.

Step 7:重复Step 4 ~ Step 6,直至覆盖论域,输出所有的不一致规则.

Step 8:对于上述得到的每一条规则,若依次去除某属性后,依然能够得到相同的决策类,则删除,否则不删除.

Step 9:输出所有的决策规则,算法结束.

算法2主要分为3个步骤:第1步将不完备决策表转化为增广形式背景,并生成所有的极概念和决策概念;第2步是一致规则和不一致规则的提取过程;第3步去除冗余属性,获取最简决策规则. 特别地,当决策信息系统完备时,文中的极概念将退化为定义3中经典的形式概念,故对于决策表的规则提取,本文算法2更具一般性.

4.2　实例分析

下面利用例3具体说明算法2.

例 3　由表2根据定义3求出决策属性的外延,即$y_1^{\downarrow} = \{1\,246\}, y_2^{\downarrow} = \{3\}, y_3^{\downarrow} = \{5\}$. 根据算法2,在$\mathfrak{B}(U, A, \{\times, *\}, I)$中找出极概念中$\mathrm{LG}(X)$为$y_1^{\downarrow}$、$y_2^{\downarrow}$和$y_3^{\downarrow}$子集的所有概念,根据$\mathrm{LG}(X)$中元素个数计算He并排序,得到表4和表5,其中$y_3^{\downarrow}$无子集,黑色规则为已经区分的规则,故不再获取.

$r_1: c_1d_1 \to y_1, r_2: c_2d_2 \to y_2$,易知$r_1, r_2$覆盖论域$\{1\,236\}$,未覆盖论域4、5,且构成不一致决策. 故将论域4、5合并$\{45\}$,找出极概念中$\mathrm{LG}(X)$为$\{45\}$的非空子集的所有概念,根据$\mathrm{LG}(X)$计算He并排序,得到表6.

$r_3: a_1c_1d_2 \to y_1 \vee y_3$,至此规则$r_1, r_2, r_3$已经覆盖论域.

表 4　y_1的规则

极概念	$\mathrm{LG}(X)$	He	规则
$((12, 126), c_1d_1)$	126	3	$c_1d_1 \to y_1$
$((2, 26), a_2c_1d_1)$	26	2	$\boldsymbol{a_2c_1d_1 \to y_1}$
$((1, 1), a_1b_1c_1d_1)$	1	1	$\boldsymbol{a_1b_1c_1d_1 \to y_1}$

表 5　y_2的规则

极概念	$\mathrm{LG}(X)$	He	规则
$((3, 3), c_2d_2)$	3	1	$c_2d_2 \to y_2$

表 6　不一致规则表

极概念	$\mathrm{LG}(X)$	He	规则
$((4, 45), a_1c_1d_2)$	45	2	$a_1c_1d_2 \to y_1 \vee y_3$

由例3得到决策规则r_1, r_2, r_3,根据算法2对r_1, r_2, r_3进行去冗余. 对于规则$r_1: c_1d_1 \to y_1$,属性c_1的$c_1^{\downarrow} = \{12\,456\} \not\subset \{1\,246\}$,属性$d_1$的$d_1^{\downarrow} = \{126\} \subseteq \{1\,246\}$,故规则$r_1$简化为$r_1: d_1 \to y_1$;对于规则$r_2: c_2d_2 \to y_2$,属性$c_2$的$c_2^{\downarrow} = \{3\} \subseteq \{3\}$,故规则$r_2: c_2 \to y_2$;对于规则$r_3: a_1c_1d_2 \to y_1 \bigvee y_3$,由于$r_3$是不一致决策,又$y_1 \bigvee y_3^{\downarrow} = \{12\,456\}$,属性$a_1$的$a_1^{\downarrow} = \{1\,345\} \not\subset \{12\,456\}$,属性$c_1$的$c_1^{\downarrow} = \{12\,456\} \subseteq \{12\,456\}$,故规则$r_3$简化为$r_3: c_1 \to y_1$

$\bigvee y_3$. 最后,由表1得到的最简决策规则如表7所示,与文献[3]中的计算结果一致.

表7 Car的规则

$r_1\ d_1 \to y_1$	即: $\text{Max}-\text{speed}=\text{low}\to Y=\text{good}$
$r_2\ c_2 \to y_2$	即: $\text{Size}=\text{compact}\to Y=\text{poor}$
$r_3\ c_1 \to y_1 \vee y_3$	即: $\text{Size}=\text{full}\to Y=\text{good}\vee\text{excellent}$

4.3 复杂性分析

设不完备信息系统IIS $=\{U,\text{AT},V,f\}$,根据定义6转换为增广形式背景$\boldsymbol{k}=(U,A,D,\{\times,*\},I,J)$. 设不完备信息系统IIS中条件属性值缺失个数为$M$,根据算法1求得增广形式背景中的所有极概念,在增广形式背景$\boldsymbol{k}$中满足$(U,A)\in I$的元素个数为$|U||A|-|M|$,因此每层复杂度基数是$O((|U||A|-|M|)^2)$. 由于总的层数是$\log_2(|U||A|-|M|)$,故算法1的复杂度为

$$O(|U||A|-|M|)+O((|U||A|-|M|)^2\log_2(|U||A|-|M|)).$$

设不完备信息系统IIS中决策类有W类,在最坏情况,即全是不一致决策下,规则提取复杂度为$O(W(W-1)/2)=O(W^2)$. 因此,整个算法2的复杂度为

$$O(|U||A|-|M|)+O((|U||A|-|M|)^2\log_2(|U||A|-|M|))+O(W^2).$$

4.4 实验测试

下面通过几组数据集进行测试,验证本文算法的正确性和有效性. 选用UCI数据集,经过Rosetta软件对连续型数据集进行离散化处理. 对数据集随机去除3.5%、4.7%、9.8%和13.5%的属性值,得到不完备数据集. 从数据集中随机抽取50%作为训练样本,然后应用本文算法2、基于属性分辨度的规则提取算法(记作算法a)[11]、基于属性序的规则提取算法(记作算法b)[26]、基于相容概念的规则提取算法(记作算法c)[20]和基于近似概念的规则提取算法(记作算法d)[19],分别进行规则获取. 利用得到的规则对整个数据集进行识别,最终得到规则个数,规则长度和识别率. 实验对比结果如表8所示.

表8 实验结果对比

数据集	缺失率/%	规则个数					规则长度					识别率				
		算法2	算法a	算法b	算法c	算法d	算法2	算法a	算法b	算法c	算法d	算法2	算法a	算法b	算法c	算法d
Iris	3.5	7	6	12	14	7	**13**	15	28	35	15	**94.3**	93.7	94.8	92.8	93.9
Iris	4.7	**7**	8	22	15	7	**15**	18	77	40	17	92.7	92.8	89.3	90.2	93.1
Iris	9.8	**8**	8	27	25	8	**18**	20	82	64	18	**92.6**	92.5	88.8	87.5	92.6
Iris	13.5	**10**	12	30	31	12	**21**	25	85	75	23	**91.5**	91.1	90.8	90.7	91.4
Glass	3.5	42	45	50	58	40	**152**	155	172	208	160	86.9	87.6	86.8	86.3	88.1
Glass	4.7	46	50	81	67	45	**70**	97	149	127	85	84.8	85.2	82.1	83.9	84.7
Glass	9.8	**49**	52	87	90	50	**180**	203	278	295	214	83.5	83.9	81.7	83.2	83.6
Glass	13.5	**52**	63	92	98	54	**214**	248	312	335	230	**82.7**	81.6	82.3	81.5	82.5
Voting	3.5	20	23	18	22	18	**63**	70	97	79	70	95.3	97.8	94.7	92.7	97.9
Voting	4.7	21	20	19	20	20	**66**	75	68	68	72	94.2	95.9	92.8	92.4	96.0
Voting	9.8	27	23	22	30	25	**112**	138	101	128	120	**94.8**	93.5	92.9	91.9	94.2
Voting	13.5	**35**	48	44	49	40	**152**	204	189	234	180	**93.7**	92.1	93.2	91.5	93.4

由表8可知:在同一数据集缺失率增加的情况下,算法2、算法a、算法b、算法c和算法d在规则个数方面均有所增加. 但是在规则长度上,算法2要优于算法a、算法b、算法c和算法d;在识别率方面,算法2的识别率与算法a、算法d的识别率相当,高于算法b、算法c的识别率. 算法2用较少的规则获得了较高的分类效果,其原因在于,规则个数、规则长度、识别率不仅与缺失属性值的个数有关,还与缺失属性值在整个数据集中的分布有很大关系. 算法2的优点体现在从相容关系出发,定义了极概念,然后从所有的极概念中获取规则,并针对每一条规则进行去冗余化,从而保证了所提取规则的无冗余,也保证了每条规则的长度最小;从覆盖论域的角度考虑,每条规则的长度最小化,使得算法2具有较高的识别率.

5 结 论

本文通过定义增广形式背景,将形式概念分析方法引入到不完备信息系统的数据分析中,便于从形式概念分析的角度去分析、解决粗糙集中的问题,更进一步地理解形式概念分析与粗糙集之间的关系. 提出了极概念与极概念格,并讨论了极概念的性质,在极概念基础上提出了一种增广形式背景的极概念生成算法以及无冗余规则提取算法. 本文从实例计算和对比实验两个方面表明了算法的正确性和有效性. 当不完备信息系统退化为完备信息系统时,极概念将退化为传统的概念. 本算法的不足之处是,在获取不一致规则时仅考虑了两类决策构成不一致的情形;对于大型决策信息系统,由于极概念个数会急剧增加,从而会降低算法的效率. 如何利用形式概念分

析工具进行不完备信息系统的快速知识获取是下一步的研究方向,相关工作仍在继续.

参考文献(References)

[1] Pawlak Z. Rough set[J]. Int J of Computer and Information Sciences, 1982, 11(5): 341-352.

[2] Wu W Z, Zhang W X. Constructive and axiomatic approaches of fuzzy approximation operators[J]. Information Sciences, 2004, 159(3): 233-254.

[3] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences An Int J, 1998, 112(1-4): 39-49.

[4] Leung Y, Li D Y. Maximal consistent block technique for rule acquisition in incomplete information systems[J]. Information Sciences, 2003, 153(1): 85-106.

[5] Li R P, Zhang D D, Zhao Y S, et al. Rule extraction from incomplete decision tables[C]. Wase Int Conf on Information Engineering. Taiyuan, 2009: 639-642.

[6] Qian Y H, Liang J Y, Li D Y, et al. Approximation reduction in inconsistent incomplete decision tables[J]. Knowledge-Based Systems, 2010, 23(5): 427-433.

[7] 杨习贝,杨静宇,於东军,等. 不完备信息系统中的可变精度分类粗糙集模型[J]. 系统工程理论与实践, 2008, 28(5): 116-121.
(Yang X B, Yang J Y, Yu D J, et al. Rough set model based on variable parameter classification in incomplete information systems[J]. Systems Engineering — Theory & Practice, 2008, 28(5): 116-121.)

[8] 程玉胜,张佑生,胡学钢. 不完备决策系统中规则提取的快速矩阵算法[J]. 系统仿真学报, 2008, 20(15): 4036-4040.
(Cheng Y S, Zhang Y S, Hu X G. Fast matrix computation algorithm for rules extraction in incomplete decision systems[J]. J of System Simulation, 2008, 20(15): 4036-4040.)

[9] Meng Z Q, Gan Q L, Shi Z Z, et al. On efficient methods of computing attribute-value blocks in incomplete decision systems[J]. Knowledge-Based Systems, 2016: 113: 171-185.

[10] Li H X, Wang M H, Zhou X Z, et al. An interval set model for learning rules from incomplete information table[J]. Int J of Approximate Reasoning, 2012, 53(1): 24-37.

[11] 纪霞,李龙澍. 基于属性分辨度的最大相容块规则提取算法[J]. 控制与决策, 2013, 28(12): 1837-1842.
(Ji X, Li L S. Algorithm for rules acquisition from maximal consistent blocks based on attribute discernibility[J]. Control and Decision, 2013, 28(12): 1837-1842.)

[12] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts[M]. Ordered Sets: Springer Netherlands, 1982: 445-470.

[13] Yao Y Y. Concept lattices in rough set theory[C]. Proc of the Annual Meeting of the North American Fuzzy Information Processing Society. Banff, 2004: 796-801.

[14] Wang H, Zhang W X. Relationships between concept lattice and rough set[J]. Artificial Intelligence and Soft Computing, 2006, 4029(3): 538-547.

[15] Wei L, Qi J J. Relation between concept lattice reduction and rough set reduction[J]. Knowledge-Based Systems, 2010, 23(8): 934-938.

[16] 张文修,仇国芳. 基于粗糙集的不确定性决策[M]. 北京: 清华大学出版社, 2005: 445-470.
(Zhang W X, Qiu G F. Uncertain decision making based on rough sets[M]. Bering: Tsinghua University Press, 2005: 445-470.)

[17] 杨凯,马垣. 基于概念格的多层属性约简方法[J]. 模式识别与人工智能, 2012, 25(6): 922-927.
(Yang K, Ma Y. Multilevel attribute reduction methods based on concept lattice[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(6): 922-927.)

[18] Li J H, Mei C L, Lyu Y J. Knowledge reduction in decision formal contexts[J]. Knowledge-Based Systems, 2011, 24(5): 709-715.

[19] Li J H, Mei C L, Lyu Y J. Incomplete decision contexts: approximate concept construction, rule acquisition and knowledge reduction[J]. Int J of Approximate Reasoning, 2013, 54(1): 149-165.

[20] 李想,王素格,李德玉,等. 形式概念分析在不完备信息系统中的知识获取[J]. 计算机科学, 2014, 41(7): 250-253.
(Li X, Wang S G, Li D Y, et al. Knowledge acquisition in incomplete information system based on formal concept analysis[J]. Computer Science, 2014, 41(7): 250-253.)

[21] 李金海,梅长林,张红英,等. 基于遗传算法的决策形式背景的属性约简方法及其在决策分析中的应用[J]. 小型微型计算机系统, 2015, 36(8): 1803-1808.
(Li J H, Mei C L, Zhang H Y, et al. Attribute reduction method for formal decision contexts based on genetic algorithm and its application to decision-making analysis[J]. J of Chinese Computer Systems, 2015, 36(8): 1803-1808.)

[22] 翟岩慧,李德玉,曲开社. 决策蕴涵规范基[J]. 电子学报, 2015, 43(1): 18-23.
(Zhai Y H , Li D Y, Qu K S. Canonical basis for decision implications [J]. Acta Electronica Sinica, 2015, 43(1): 18-23.)

[23] Shao M W, Li K W. Attribute reduction in generalized one-sided formal contexts[J]. Information Sciences, 2016, 378: 317-327.

[24] Ganter B. Formal concept analysis: Mathematical foundations[M]. New York: Springer-Verlag, 1999: 17-243.

[25] Ganter B. Two basic algorithms in concept analysis[C]. Int Conf on Formal Concept Analysis. Agadir, 2010: 312-340.

[26] Guan L H, Hu F, Han F Q. A rule induction algorithm in incomplete decision table based on attribute order[J]. J of Intelligent and Fuzzy Systems, 2016, 30(2): 961-969.

(责任编辑: 郑晓蕾)