基于概念格的异构数据知识发现方法

牛娇娇 范 敏 李金海 殷允强
(昆明理工大学理学院 昆明 650500)

摘 要 基于概念格的知识发现方法已被广泛关注,同时也吸引了众多学者的研究兴趣,特别是决策形式背景的知识发现,近年来取得了一些重要的研究成果。然而,现有的知识发现方法在面临大数据环境时,缺乏可行性与有效性。考虑到异构性是大数据的主要数据特征之一,针对异构数据,研究了基于概念格的知识发现方法。具体地,提出了异构形式背景及其概念格,通过异构形式背景定义了异构决策形式背景,进一步在异构决策形式背景上讨论了规则提取问题,并给出了挖掘非冗余决策规则的有效算法。

关键词 概念格,异构形式背景,异构决策形式背景,知识发现
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.012

Knowledge Discovery Method for Heterogeneous Data Based on Concept Lattice

NIU Jiao-jiao FAN Min LI Jin-hai YIN Yun-qiang
(Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China)

Abstract Recently, much attention has been paid to concept-lattice-based knowledge discovery methods. In the meanwhile, this topic has attracted many research interests from the communities of formal concept analysis and rough set theory. Especially, in recent years, some substantial progresses have been made on studying formal decision contexts. However, the existing knowledge discovery methods are lack of feasibility and effectiveness when they are applied to big data. Considering that heterogeneity is one of the main characteristics of big data, this paper investigated concept-lattice-based knowledge discovery methods for heterogeneous data. Specifically, the notion of a heterogeneous formal context was proposed as well as its corresponding concept lattice, heterogeneous formal contexts were further employed to define heterogeneous formal decision contexts, and rule acquisition was discussed. Moreover, an algorithm of mining non-redundant decision rules from a heterogeneous formal decision context was explored.

Keywords Concept lattice, Heterogeneous formal context, Heterogeneous formal decision context, Knowledge discovery

1 引言

形式概念分析由德国 Wille 教授于 20 世纪 80 年代初提出[1],国内大概在 20 世纪 90 年代末才开始研究[2],并常称其为概念格理论。虽然国内概念格的研究起步稍晚,但是经过近 20 年的努力探讨,已逐渐形成了一些独具特色的研究方向,比如概念格属性约简[3]、概念学习[4-5]、属性拓扑图[6]、三支概念分析[7]、决策形式背景分析[8-12]。

实际上,决策形式背景是通过一对形式背景构造的[13],这样做的主要目的是通过在形式背景上引入决策属性来进行具体的决策分析。截至目前,决策形式背景的研究已取得初步成果。比如,曲开社等[14]将蕴含推广到决策形式背景,定义了决策蕴含,并讨论了其非冗余问题;魏玲等[8]在条件概念格和决策概念格之间引入两种序关系以分析其强弱协调性,并在此基础上讨论属性约简问题;邵明文等[9]则从最大规则提取的角度探讨了协调决策形式背景的属性约简;李金海等[10-11]在决策形式背景中定义了非冗余决策规则,并研究了保持非冗余决策规则不变的属性约简,其主要特点是该方法适用于任意决策形式背景;此外,吴伟志等[12]为了克服挖掘一般决策规则耗时太长的问题,提出了粒规则的概念,并给出了挖掘粒规则的有效算法。

尽管决策形式背景研究已取得了一些满意的成果,但是其也面临着一系列的挑战。比如,在应对大数据环境时缺乏可行性与有效性,因为现有的方法都是假设数据是同构的,即属性取值构成的偏序两两相同,然而大数据环境下的数据往往是异构的,即属性取值构成的偏序不尽相同,甚至完全不同[15]。在这种情况下,现有的处理方法将失效,无法给决策形式背景提供决策分析支持。

受此启发,本文基于概念格来研究异构决策形式背景的知识发现问题。首先,定义异构形式背景,并在此基础上提出异构决策形式背景;然后,分析异构形式背景的概念格构造,并讨论异构决策形式背景的规则提取问题;最后,给出挖掘非

到稿日期:2016-07-08 返修日期:2016-09-22 本文受国家自然科学基金(61305057,61562050,61573173)资助。
牛娇娇(1992—),女,硕士生,主要研究方向为粒计算与概念格;范 敏(1975—),女,博士,副教授,主要研究方向为粗糙集与模糊集,E-mail:fmkmust@163.com(通信作者);李金海(1984—),男,博士,副教授,主要研究方向为粗糙集、概念格与粒计算;殷允强(1980—),男,博士,教授,主要研究方向为数据优化与处理。

冗余决策规则的有效算法，并通过实例说明其可行性。

2　异构形式背景和异构决策形式背景

一个形式背景可以通过三元组(U,A,I)来形式化表示，其中$U=\{x_1,x_2,\cdots,x_n\}$为非空有限对象集，$A=\{a_1,a_2,\cdots,a_m\}$为非空有限属性集，I为笛卡尔积$U\times A$上的二元关系，$(x,a)\in I$表示对象x拥有属性a。

实际上，形式背景是一个特殊的信息系统。若将对象是否拥有属性分别记为数字 1 和 0，则形式背景就可以理解为二值信息系统。进一步地，还可以在取值之间引入序关系$\leqslant$，即$0\leqslant 1$。不仅如此，还可以定义“取小”运算，如表 1 所列。

表 1　经典二值“取小”运算

$\wedge$	0	1
0	0	0
1	0	1

不难发现，形式背景的每个属性在所有对象下的取值均为 1 和 0，序或“取小”运算也均相同。因此，形式背景可以看作序同构的数据集，简称同构形式背景。

然而，现实应用中遇到的形式背景的属性取值形成的序不尽相同，甚至完全不同。本文将这类数据集称为异构形式背景，其严格定义描述如下。

定义 1　一个异构形式背景可以表示成四元组(U,A,H,I)，其中$U=\{x_1,x_2,\cdots,x_n\}$是对象集，$A=\{a_1,a_2,\cdots,a_m\}$是属性集，A中属性的取值构成的序两两不同，H是对象在属性下所有取值组成的集合，I是$U\times A\times H$上的关系，$(x,a,h)\in I$表示对象x在属性a下取值为h。

例 1　表 2 列出了一个异构形式背景(U,A,H,I)，$U=\{x_1,x_2,x_3\}$，$A=\{a,b,c\}$。其中，a为名义属性，b为区间属性，c为不完备属性。

表 2　异构形式背景(U,A,H,I)

U	a	b	c
x_1	高	[1,2]	1
x_2	中	[2,3]	*
x_3	低	[3,4]	0

分别定义属性a,b,c在对象下的取值的“取小”运算，如表 3－表 5 所列。

表 3　属性a取值的“取小”运算

$\wedge_a$	高	中	低
高	高	中	低
中	中	中	低
低	低	低	低

表 4　属性b取值的“取小”运算

$\wedge_b$	[1,2]	[2,3]	[3,4]
[1,2]	[1,2]	[2,2]	[,]
[2,3]	[2,2]	[2,3]	[3,3]
[3,4]	[,]	[3,3]	[3,4]

表 5　属性c取值的“取小”运算

$\wedge_c$	1	*	0
1	1	*	0
*	*	*	0
0	0	0	0

利用上述“取小”运算，可以诱导出相应的序关系：

$s\leqslant t\Leftrightarrow s\wedge t=s$

定义 2　一个异构决策形式背景可以表示为七元组(U,A,H_A,I,D,H_D,J)，其中(U,A,H_A,I)和(U,D,H_D,J)为异构形式背景且$A\cap D=\varnothing$。通常地，A和D分别为异构决策形式背景的条件属性集和决策属性集。

例 2　表 6 列出了一个异构决策形式背景(U,A,H_A,D,H_D,J)，其中，$U=\{x_1,x_2,x_3,x_4,x_5\}$，$A=\{a,b,c\}$，$D=\{d_1,d_2\}$，$a$为不完备属性，$b$为名义属性，$c$为格值属性，$d_1$为模糊值属性，$d_2$为区间值属性。

表 6　异构决策形式背景(U,A,H_A,I,D,H_D,J)

U	a	b	c	d_1	d_2
x_1	1	高	l_5	0.9	[6,9]
x_2	*	中	l_1	0.7	[2,5]
x_3	0	低	l_3	0.5	[3,5]
x_4	1	高	l_2	0.6	[4,7]
x_5	0	中	l_4	0.8	[6,8]

属性a,b,d_2在对象下的取值的序关系类似于例 1 的定义，下面只给出属性c，属性d_1在对象下的取值的序关系，如表 7－表 8 所列。

表 7　属性c取值的“取小”运算

$\wedge_l$	l_5	l_1	l_3	l_2	l_4
l_5	l_5	l_1	l_3	l_2	l_4
l_1	l_1	l_1	l_1	l_1	l_1
l_3	l_3	l_1	l_3	l_2	l_3
l_2	l_2	l_1	l_2	l_2	l_2
l_4	l_4	l_1	l_3	l_2	l_4

表 8　属性d_1取值的“取小”运算

$\wedge_{d_1}$	0.9	0.8	0.7	0.6	0.5
0.9	0.9	0.8	0.7	0.6	0.5
0.8	0.8	0.8	0.7	0.6	0.5
0.7	0.7	0.7	0.7	0.6	0.5
0.6	0.6	0.6	0.6	0.6	0.5
0.5	0.5	0.5	0.5	0.5	0.5

最后，通过$s\leqslant t\Leftrightarrow s\wedge t=s$即可得到属性$c$，属性$d_1$在对象下的取值的序关系。

以上仅讨论“取小”运算与属性在对象下取值的序关系，实际上“取大”运算也可以类似地定义，且这些运算是封闭的。

3　异构形式背景的概念格

为了方便问题描述，先引入一些符号。对于任意$x\in U$，$a\in A$，用$I(x,a)$表示对象x在属性a下的取值h，即$I(x,a)=h\Leftrightarrow(x,a,h)\in I$。在此基础上，用$I(x,\cdot)$表示对象$x$在各个属性下的取值构成的向量。假设$X=\{x_1,x_2,\cdots,x_k\}$，$A=\{a_1,a_2,\cdots,a_m\}$，记

$$\bigwedge_{x\in X}I(x,a)=I(x_1,a)\wedge I(x_2,a)\wedge\cdots\wedge I(x_k,a)$$

$$\bigvee_{x\in X}I(x,a)=I(x_1,a)\vee I(x_2,a)\vee\cdots\vee I(x_k,a)$$

$$\bigcap_{x\in X}I(x,\cdot)=\left(\bigwedge_{x\in X}{}_1I(x,a_1),\bigwedge_{x\in X}{}_2I(x,a_2),\cdots,\bigwedge_{x\in X}{}_mI(x,a_m)\right)$$

$$\bigcup_{x\in X}I(x,\cdot)=\left(\bigvee_{x\in X}{}_1I(x,a_1),\bigvee_{x\in X}{}_2I(x,a_2),\cdots,\bigvee_{x\in X}{}_mI(x,a_m)\right)$$

设π为$I(X,\cdot)(X\subseteq U)$的集合，其中$I(X,\cdot)$表示对象

集 X 在各个属性下取值的向量，即 $I(X,\cdot)=\bigcap_{x\in X}I(x,\cdot)$，那么 $\pi=\{I(X,\cdot)\mid X\subseteq U\}$。约定 π 上的序关系：

$I(X_1,\cdot)\leqslant I(X_2,\cdot)\Leftrightarrow I(X_1,a_i)\leqslant_i I(X_2,a_i)(\vee a_i\in A)$

其中，$I(X,a)=\bigwedge_{x\in X}I(x,a)$。由于属性集 A 中各个属性的取值形成的序是异构的，因此 $\leqslant_i$ 两两不同。

例 3　以表 6 中的异构条件形式背景为例，取 $X=\{x_1,x_2,x_3\}$，$A=\{a,b,c\}$，则

$$\begin{aligned}\bigwedge_{x\in X}I(x,a)&=(I(x_1,a)\wedge_1 I(x_2,a))\wedge_1 I(x_3,a)\\&=(1\wedge_1 0)\wedge_1 *=0\wedge_1 *=0\\\bigvee_{x\in X}I(x,a)&=(I(x_1,a)\vee_1 I(x_2,a))\vee_1 I(x_3,a)\\&=(1\vee_1 0)\vee_1 *=1\vee_1 *=1\\\bigcap_{x\in X}I(x,\cdot)&=(\bigwedge_{x\in X}{}_1 I(x,a_1),\bigwedge_{x\in X}{}_2 I(x,a_2),\bigwedge_{x\in X}{}_3 I(x,a_3))\\&=((1\wedge_1 0)\wedge_1 *,(\text{高}\wedge_2\text{中})\wedge_2\text{低},(l_5\wedge_3 l_1)\wedge_3 l_3)\\&=(0\wedge_1 *,\text{中}\wedge_2\text{低},l_1\wedge_3 l_3)\\&=(0,\text{低},l_1)\\\bigcup_{x\in X}I(x,\cdot)&=(\bigvee_{x\in X}{}_1 I(x,a_1),\bigvee_{x\in X}{}_2 I(x,a_2),\bigvee_{x\in X}{}_3 I(x,a_3))\\&=((1\vee_1 0)\vee_1 *,(\text{高}\vee_2\text{中})\vee_2\text{低},(l_5\vee_3 l_1)\vee_3 l_3)\\&=(1\vee_1 *,\text{高}\vee_2\text{低},l_5\vee_3 l_3)\\&=(1,\text{高},l_5)\end{aligned}$$

定义 3　设四元组 (U,A,H_A,I) 为异构形式背景，在对象集 $X\subseteq U$ 和向量 $B\in\pi$ 上定义如下运算：

$$X^{\triangleright}=\bigcap\{I(x,\cdot)\mid\forall x\in X\}\tag{1}$$

$$B^{\triangleleft}=\{x\in U\mid B\leqslant I(x,\cdot)\}\tag{2}$$

定理 1　设 (U,A,H_A,I) 为异构形式背景，$P(U)$ 为 U 的幂集，$X,X_1,X_2\in P(U)$，$B,B_1,B_2\in\pi$，则

1) $X_1\subseteq X_2\Rightarrow X_2^{\triangleright}\leqslant X_1^{\triangleright}$，$B_1\leqslant B_2\Rightarrow B_2^{\triangleleft}\subseteq B_1^{\triangleleft}$；

2) $X\subseteq X^{\triangleright\triangleleft}$，$B\leqslant B^{\triangleleft\triangleright}$；

3) $X^{\triangleright}=X^{\triangleright\triangleleft\triangleright}$，$B^{\triangleleft}=B^{\triangleleft\triangleright\triangleleft}$；

4) $(X_1\bigcup X_2)^{\triangleright}=X_1^{\triangleright}\wedge X_2^{\triangleright}$，$(B_1\vee B_2)^{\triangleleft}=B_1^{\triangleleft}\bigcap B_2^{\triangleleft}$。

证明：1) 因为 $X_1\subseteq X_2$，所以对于任意 $a_i\in A$，有 $\bigwedge_{x\in X_2}{}_i I(x,a_i)\leqslant\bigwedge_{x\in X_1}{}_i I(x,a_i)$ 成立，故

$$(\bigwedge_{x\in X_2}{}_1 I(x,a_1),\bigwedge_{x\in X_2}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_2}{}_m I(x,a_m))\leqslant(\bigwedge_{x\in X_1}{}_1 I(x,a_1),\bigwedge_{x\in X_1}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_1}{}_m I(x,a_m))$$

即 $\bigcap_{x\in X_2}I(x,\cdot)\leqslant\bigcap_{x\in X_1}I(x,\cdot)$，也即 $X_2^{\triangleright}\leqslant X_1^{\triangleright}$。

同理，已知 $B_1\leqslant B_2$，那么 $B_2\leqslant I(x,\cdot)\Rightarrow B_1\leqslant I(x,\cdot)$ 成立，由此可得 $\{x\in U\mid B_2\leqslant I(x,\cdot)\}\subseteq\{x\in U\mid B_1\leqslant I(x,\cdot)$，即 $B_2^{\triangleleft}\subseteq B_1^{\triangleleft}$。

2) 由于 $X^{\triangleright\triangleleft}=\{x\in U\mid X^{\triangleright}\leqslant I(x,\cdot)\}=\{x\in U\mid\bigcap_{y\in X}I(y,\cdot)\leqslant I(x,\cdot)\}$，而任意 $x\in X$，必满足 $\bigcap_{y\in X}I(y,\cdot)\leqslant I(x,\cdot)$，这就证得了 $X\subseteq X^{\triangleright\triangleleft}$。

类似地，也可以证得 $B\leqslant B^{\triangleleft\triangleright}$。

3) 若把 $X^{\triangleright}$ 看作一个整体，则由 2) 可得 $X^{\triangleright}\leqslant X^{\triangleright\triangleleft\triangleright}$；同样，由 2) 可知 $X\subseteq X^{\triangleright\triangleleft}$，再结合 1) 有 $X^{\triangleright\triangleleft\triangleright}\leqslant X^{\triangleright}$。综上所述，即可证得 $X^{\triangleright}=X^{\triangleright\triangleleft\triangleright}$。

类似地，也可以证得 $B^{\triangleleft}=B^{\triangleleft\triangleright\triangleleft}$。

4) 因为

$$\begin{aligned}X_1^{\triangleright}&=\bigcap_{x\in X_1}I(x,\cdot)\\&=(\bigwedge_{x\in X_1}{}_1 I(x,a_1),\bigwedge_{x\in X_1}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_1}{}_m I(x,a_m))\\X_2^{\triangleright}&=\bigcap_{x\in X_2}I(x,\cdot)\\&=(\bigwedge_{x\in X_2}{}_1 I(x,a_1),\bigwedge_{x\in X_2}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_2}{}_m I(x,a_m))\end{aligned}$$

所以

$$\begin{aligned}X_1^{\triangleright}\wedge X_2^{\triangleright}&=(\bigwedge_{x\in X_1}{}_1 I(x,a_1),\bigwedge_{x\in X_1}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_1}{}_m I(x,a_m))\wedge(\bigwedge_{x\in X_2}{}_1 I(x,a_1),\bigwedge_{x\in X_2}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_2}{}_m I(x,a_m))\\&=(\bigwedge_{x\in X_1\cup X_2}{}_1 I(x,a_1),\bigwedge_{x\in X_1\cup X_2}{}_2 I(x,a_2),\cdots,\bigwedge_{x\in X_1\cup X_2}{}_m I(x,a_m))\\&=\bigcap_{x\in X_1\cup X_2}I(x,\cdot)=(X_1\bigcup X_2)^{\triangleright}\end{aligned}$$

类似地，也可以证得 $(B_1\vee B_2)^{\triangleleft}=B_1^{\triangleleft}\bigcap B_2^{\triangleleft}$。

定义 4[1]　给定两个偏序集 $(S,\leqslant_S)$ 和 $(T,\leqslant_T)$，称 $f:S\to T$ 和 $g:T\to S$ 是偏序集之间的伽罗瓦连接，如果它们满足以下性质：

1) $a\leqslant_S b\Rightarrow f(b)\leqslant_T f(a)$；

2) $c\leqslant_T d\Rightarrow g(d)\leqslant_S g(c)$；

3) $a\leqslant_S g(f(a))$，$c\leqslant_T f(g(c))$。

定理 2　设 (U,A,H_A,I) 为异构形式背景，$\triangleright:P(U)\to\pi$ 和 $\triangleleft:\pi\to P(U)$ 如定义 3 所示，则映射序对 $(\triangleright,\triangleleft)$ 构成偏序集 $(P(U),\subseteq)$ 和 $(\pi,\leqslant)$ 之间的伽罗瓦连接。

证明：由定理 1 和定义 4 显然成立。

定义 5　设 (U,A,H_A,I) 为异构形式背景，$X\in P(U)$，$B\in\pi$，若 $X^{\triangleright}=B$ 且 $B^{\triangleleft}=X$，则称 (X,B) 为异构形式背景的概念，其中 X 为概念的外延，B 为概念的内涵。

为了方便，记异构形式背景的所有概念组成的集合为 $L(U,A,H_A,I)$。定义概念之间的偏序关系如下：

$$(X_1,B_1)\leqslant(X_2,B_2)\Leftrightarrow X_1\subseteq X_2\Leftrightarrow B_1\geqslant B_2$$

定义下确界和上确界如下：

$$(X_1,B_1)\wedge(X_2,B_2)=(X_1\bigcap X_2,(B_1\vee B_2)^{\triangleleft\triangleright})$$

$$(X_1,B_1)\vee(X_2,B_2)=((X_1\bigcup X_2)^{\triangleright\triangleleft},B_1\wedge B_2)$$

至此，$L(U,A,H_A,I)$ 构成完备格，称为异构形式背景的概念格。

异构形式背景的概念格 $L(U,A,H_A,I)$ 与 Wille 概念格一样，都满足反序伽罗瓦连接，因此它们的构造原理是类似的。算法 1 给出构造 $L(U,A,H_A,I)$ 的步骤。

算法 1　异构形式背景的概念格构造算法

输入：一个异构形式背景(U,A,H,I)

输出：形式背景(U,A,H,I)的所有概念 Γ

Step1　初始化：Γ←{({x_1},I(x_1,·))}，i=2；

Step2　S=Γ；

Step3　从 S 中任选一个元素(X,B)；

Step4　如果 B∧I(x_i,·)=B，Γ←Γ\{(X,B)}，Γ=Γ∪{(X∪{x_i},B)}，转 Step6，否则转 Step5；

Step5　如果对于任意 $x_j\in\{x_1,x_2,\cdots,x_{i-1}\}\backslash X$，B∧I($x_j$,·)≤I($x_j$,·)都不成立，则 Γ←Γ∪{(X∪{$x_i$},B∧I($x_i$,·)}；

Step6　如果 S\(X,B)=∅，则 S=S\(X,B)，并返回 Step3；

Step7　如果对于任意 j<i，I(x_i，·)≤I(x_j，·)都不成立，则 Γ←Γ∪{({x_i}，I(x_i，·))}；

Step8　如果 i<|U|，则 i=i+1，并返回 Step2；

Step9　如果 $\bigcap_{(X,B)\in\Gamma} X=\varnothing$ 则 $\Gamma\leftarrow\Gamma\cup\{(\varnothing,\bigvee_{x\in U} I(x,\cdot))\}$；

Step10　输出 Γ。

类似于经典概念格的构造算法，算法 1 的时间复杂度也是指数级的。

4　异构决策形式背景的规则提取

为了便于讨论，下面给出异构形式背景的子背景。

定义 6　设(U,A,H_A,I)为异构形式背景，$S\subseteq A$，I_S 是将 I 限制在 $U\times S\times H_S$ 上得到的关系，称(U,S,H_S,I_S)为异构形式背景(U,A,H_A,I)的子背景，并记其相应的概念格为 $L(U,S,H_S,I_S)$。

显然，异构形式背景(U,A,H_A,I)上定义的概念诱导算子的所有性质，对其子背景上的概念诱导算子也是同样适用的。

定义 7　设(U,A,H_A,I,D,H_D,J)为异构决策形式背景，$S\subseteq A$，称(U,S,H_S,I_S)和(U,D,H_D,J)构成的异构决策形式背景(U,S,H_S,I_S,D,H_D,J)为异构决策形式背景(U,A,H_A,I,D,H_D,J)的子背景。

定义 8　设(U,S,H_S,I_S,D,H_D,J)为异构决策形式背景(U,A,H_A,I,D,H_D,J)的子背景，对于概念$(X,B)\in L(U,S,H_S,I_S)$，$(Y,C)\in L(U,D,H_D,J)$，如果 $X\subseteq Y$，则称 $B\rightarrow C$ 为概念(X,B)和(Y,C)产生的决策规则，其中 B,C 分别称为决策规则 $B\rightarrow C$ 的条件和结论。

为了方便，记 Ω_S 为 $L(U,S,H_S,I_S)$和 $L(U,D,H_D,J)$之间的概念形成的所有规则的集合。

定义 9　设(U,S,H_S,I_S,D,H_D,J)为异构决策形式背景(U,A,H_A,I,D,H_D,J)的子背景，$B'\rightarrow C'$，$B''\rightarrow C''\in\Omega_S$，称 $B''\rightarrow C''$相对于 $B'\rightarrow C'$是冗余的，如果 $B'\leqslant B''$且 $C''\leqslant C'$，记作 $B'\rightarrow C'\Rightarrow B''\rightarrow C''$。

定义 10　设(U,S,H_S,I_S,D,H_D,J)为异构决策形式背景(U,A,H_A,I,D,H_D,J)的子背景，称 $B'\rightarrow C'\in\Omega_S$ 在 Ω_S 中是冗余的，如果存在 $B''\rightarrow C''\in\Omega_S$，使得 $B''\rightarrow C''\Rightarrow B'\rightarrow C'$；否则，称 $B'\rightarrow C'$在 Ω_S 是非冗余的。

单从数据分析的角度来看，非冗余规则是人们挖掘决策规则的主要关注对象，因为冗余规则是可以被蕴含的。下面讨论如何从异构决策形式背景中挖掘非冗余规则。

设(U,S,H_S,I_S,D,H_D,J)为异构决策形式背景(U,A,H_A,I,D,H_D,J)的子背景，记

$$L_U(U,S,H_S,I_S)=\{X\mid(X,B)\in L(U,S,H_S,I_S)\}$$

$$L_U(U,D,H_D,J)=\{Y\mid(Y,C)\in L(U,D,H_D,J)\}$$

在此基础上，定义二元关系 λ_S,μ_S：对于$(X,Y)\in L_U(U,S,H_S,I_S)\times L_U(U,D,H_D,J)$，有

$$\lambda_S(X,Y)=\begin{cases}1, & \text{如果 } X\subseteq Y,\forall X'\in L_U(U,S,H_S,I_S),\\ & X\subset X'\Rightarrow X'\not\subset Y\\ 0, & \text{其他}\end{cases}\tag{3}$$

$$\mu_S(X,Y)=\begin{cases}1, & \text{如果 } X\subseteq Y,\forall Y'\in L_U(U,D,H_D,J),\\ & Y'\subset Y\Rightarrow X\not\subset Y'\\ 0, & \text{其他}\end{cases}\tag{4}$$

定理 3　设(U,S,H_S,I_S,D,H_D,J)为异构决策形式背景(U,A,H_A,I,D,H_D,J)的子背景，则 $B\rightarrow C\in\Omega_S$ 在 Ω_S 中是非冗余的，当且仅当 $\lambda_S(B^{\triangleleft},C^{\triangleleft_D})=1$ 且 $\mu_S(B^{\triangleleft},C^{\triangleleft_D})=1$，其中 $\triangleleft_D$ 表示映射 $\triangleleft$ 定义在异构形式背景(U,D,H_D,J)上。

证明：(必要性)由定义 10 可知，若 $B\rightarrow C\in\Omega_S$ 是非冗余的，则在 Ω_S 中不存在另一 $B'\rightarrow C'$使得 $B'\rightarrow C'\Rightarrow B\rightarrow C$，故 $\lambda_S(B^{\triangleleft},C^{\triangleleft_D})=1$且 $\mu_S(B^{\triangleleft},C^{\triangleleft_D})=1$。

(充分性)如果 $\lambda_S(B^{\triangleleft},C^{\triangleleft_D})=1$ 且 $\mu_S(B^{\triangleleft},C^{\triangleleft_D})=1$，那么由式(3)和式(4)可知 Ω_S 中不存在 $B'\rightarrow C'$使得 $B'\rightarrow C'\Rightarrow B\rightarrow C$，即 $B\rightarrow C$ 是非冗余的。

根据定理 3 中的结论，下面提出挖掘异构决策形式背景所有非冗余决策规则的算法。

算法 2　异构决策形式背景的非冗余决策规则提取算法

输入：一个异构决策形式背景(U,A,H_A,I,D,H_D,J)

输出：(U,A,H_A,I,D,H_D,J)的所有非冗余规则

Step1　初始化：Ω=∅；

Step2　调用算法 1 构造异构形式背景(U,A,H_A,I)和(U,D,H_D,J)的概念格，分别记为 L(U,A,H_A,I)，L(U,D,H_D,J)；

Step3　对于 L(U,A,H_A,I)×L(U,D,H_D,J)中的每个元素((X,B)，(Y,C))，如果 λ_A(X,Y)=1 且 μ_A(X,Y)=1，则 Ω←{B→C}；

Step4　输出 Ω。

由于算法 2 调用了算法 1，因此其时间复杂度也是指数级的。

例 4　根据表 6 中的异构决策形式背景继续进行分析，不难看出异构形式背景(U,A,H_A,I)有 9 个概念，异构形式背景(U,D,H_D,J)有 12 个概念，分别如图 1 和图 2 所示。

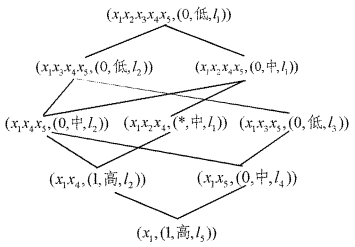

图 1　$L(U,A,H_A,I)$

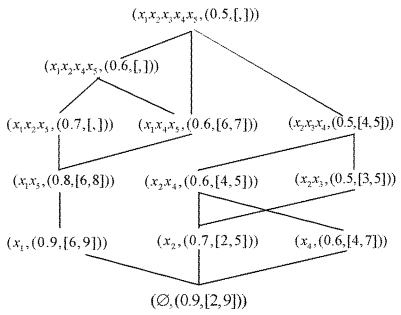

图 2　$L(U,D,H_D,J)$

由图 1 和图 2 可得如下决策规则：

(0,低,l_1)→(0.5,[,])，(0,低,l_2)→(0.5,[,])

(1,高,l_2)→(0.5,[,]),(0,中,l_1)→(0.5,[,])
(0,低,l_3)→(0.5,[,]),(*,中,l_1)→(0.5,[,])
(1,高,l_5)→(0.5,[,]),(0,中,l_2)→(0.5,[,])
(0,中,l_4)→(0.5,[,]),(*,中,l_1)→(0.6,[,])
(0,中,l_2)→(0.6,[,]),(0,中,l_1)→(0.6,[,])
(1,高,l_2)→(0.6,[,]),(0,中,l_4)→(0.6,[,])
(1,高,l_5)→(0.6,[,]),(0,中,l_4)→(0.7,[,])
(1,高,l_5)→(0.7,[,]),(0,中,l_2)→(0.6,[6,7])
(1,高,l_2)→(0.6,[6,7]),(1,高,l_2)→(0.6,[6,7])
(0,中,l_4)→(0.8,[6,8]),(1,高,l_4)→(0.8,[6,8])
(1,高,l_5)→(0.9,[6,9])

根据定义 10,进一步得到如下非冗余决策规则:

(0,低,l_1)→(0.5,[,])
(0,中,l_1)→(0.6,[,])
(0,中,l_4)→(0.7,[,])
(0,中,l_2)→(0.6,[6,7])
(0,中,l_4)→(0.8,[7,8])
(1,高,l_5)→(0.9,[6,9])

结束语 本文就决策形式背景异构数据的情况讨论了知识发现问题,主要给出了异构(决策)形式背景的定义,研究了概念格构造,给出了决策规则,得到了非冗余规则挖掘算法。

异构数据的知识发现是一个非常重要的研究课题,虽然本文提出了一些可行的分析方法,但是它们的时间复杂度均是指数级的,这意味着它们面对大规模数据时效率不高,因此继续探讨更加有效的知识发现方法是今后的工作方向。

参 考 文 献

[1] WILLE R. Restructuring lattice theory: an approach based on hierarchies of concepts[M]// Rival I, ed. Ordered Sets. Dordrecht-Boston: Reidel, 1982: 445-470.

[2] HU K Y, LU Y C, SHI C Y. Advances in concept lattice and its application[J]. Journal of Tsinghua University (Science and Technology), 2000, 40(9): 77-81. (in Chinese)
胡可云,陆玉昌,石纯一. 概念格及其应用进展[J]. 清华大学学报(自然科学版),2000,40(9):77-81.

[3] ZHANG W X, WEI L, QI J J. Attribute reduction theory and approach to concept lattice[J]. Science China Series F—Information Sciences, 2005, 35(6): 628-639. (in Chinese)
张文修,魏玲,祁建军. 概念格的属性约简理论与方法[J]. 中国科学(F 辑):信息科学,2005,35(6):628-639.

[4] LI J H, MEI C L, XU W H, et al. Concept learning via granular computing: A cognitive viewpoint[J]. Information Sciences, 2015, 298: 447-467.

[5] XU W H, LI W T. Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets[J]. IEEE Transactions on Cybernetics, 2016, 46(2): 366-379.

[6] ZHANG T, REN H L, HONG W X, et al. The visualizing calculation of formal concept that based on the attribute topologies[J]. Acta Electronica Sinica, 2014, 42(5): 925-932. (in Chinese)
张涛,任宏雷,洪文学,等. 基于属性拓扑的可视化形式概念计算[J]. 电子学报,2014,42(5):925-932.

[7] QI J J, WEI L, YAO Y Y. Three-way formal concept analysis[M]// Rough Sets and Knowledge Technology. 2014: 732-741.

[8] WEI L, QI J J, ZHANG W X. Attribute reduction theory of concept lattice based on decision formal contexts[J]. Science China Series F—Information Sciences, 2008, 38(2): 195-208. (in Chinese)
魏玲,祁建军,张文修. 决策形式背景的概念格属性约简[J]. 中国科学(F 辑):信息科学,2008,38(2):195-208.

[9] SHAO M W, LEUNG Y, WU W Z. Rule acquisition and complexity reduction in formal decision contexts[J]. International Journal of Approximate Reasoning, 2014, 55(1): 259-274.

[10] LI J H, MEI C L, LV Y J. A heuristic knowledge-reduction method for decision formal contexts[J]. Computers and Mathematics with Applications, 2011, 61(4): 1096-1106.

[11] LI J H, MEI C L, CHERUKURI A K, et al. On rule acquisition in decision formal contexts[J]. International Journal of Machine Learning and Cybernetics, 2013, 4(6): 721-731.

[12] WU W Z, LEUNG Y, MI J S. Granular computing and knowledge reduction in formal contexts[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(10): 1461-1474.

[13] 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.

[14] QU K S, ZHAI Y H, LIANG J Y, et al. Study of decision implications based on formal concept analysis[J]. International Journal of General Systems, 2007, 36(2): 147-156.

[15] ZHI H L. Extended model of formal concept analysis oriented for heterogeneous data analysis[J]. Acta Electronica Sinica, 2013, 41(12): 2451-2455. (in Chinese)
智慧来. 面向异构数据分析的形式概念分析的扩展模型[J]. 电子学报,2013,41(12):2451-2455.

(上接第 48 页)

[9] LINGRAS P, CHEN M, MIAO D. Rough cluster quality index based on decision theory[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(7): 1014-1026.

[10] YU H, LIU Z, WANG G. Automatically determining the number of clusters using decision-theoretic rough set[C]// International Conference on Rough Sets and Knowledge Technology. Banff, Canada, 2011: 504-513.

[11] MARTIN D, FOWLKES C, TAL D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]// Eighth IEEE International Conference on Computer Vision, 2001(ICCV 2001). IEEE, 2001: 416-423.

[12] YU Z, AU O C, ZOU R, et al. An adaptive unsupervised approach toward pixel clustering and color image segmentation[J]. Pattern Recognition, 2010, 43(5): 1889-1906.

[13] BEZDEK J C. Cluster validity with fuzzy sets[J]. Journal of Cybernetics, 1974, 3(3): 58-73.

[14] BEZDEK J C. Mathematical models for systematics and taxonomy[C]// Proceedings of Eighth International Conference on Numerical Taxonomy. 1975: 143-166.