ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.9, September 2007, pp.2174−2182
DOI: 10.1360/jos182174

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

形式概念分析对粗糙集理论的表示及扩展*

曲开社+, 翟岩慧, 梁吉业, 李德玉

(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,山西 太原 030006)

Representation and Extension of Rough Set Theory Based on Formal Concept Analysis

QU Kai-She+, ZHAI Yan-Hui, LIANG Ji-Ye, LI De-Yu

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

+ Corresponding author: Phn: +86-351-7011566, Fax: +86-0351-7018176, E-mail: quks@sxu.edu.cn

Qu KS, Zhai YH, Liang JY, Li DY. Representation and extension of rough set theory based on formal concept analysis. *Journal of Software*, 2007,18(9):2174−2182. http://www.jos.org.cn/1000-9825/18/2174.htm

Abstract: This paper aims to establish the relationship between formal concept analysis and rough set theory. The following results are obtained: (1) a derivative formal context of an information system can be induced by the notion of nominal scale and the technique of plain scaling in formal concept analysis; (2) some core notions in rough set theory such as partition, upper and lower approximations, independence, dependence and reduct can be reinterpreted in derivative formal contexts. In addition, the limitation of rough set theory to data processing is analyzed. The results presented in this paper provide a basis for the synthesis of formal concept analysis and rough set theory.

Key words: rough set; formal concept analysis; nominal scale; plain scaling; concept lattice

摘 要: 侧重于建立形式概念分析与粗糙集之间融合的理论基础.利用形式概念分析中名义梯级背景(nominal scale)的概念,对信息系统进行平面梯级(plain scaling)得到了衍生的形式背景.证明了粗糙集理论中的划分、上下近似、独立、依赖、约简等核心概念都可以在相应的衍生背景中进行表示.揭示了粗糙集理论在分析处理数据时的局限性,指出了利用梯级的方法可以扩展粗糙集理论.

关键词: 粗糙集;形式概念分析;名义梯级背景;平面梯级;概念格

中图法分类号: TP18 **文献标识码**: A

粗糙集理论是20世纪80年代由波兰学者Pawlak提出的一个用于分析数据的数学理论[1,2].由于粗糙集理论能够分析处理不精确、不一致和不完备信息[3,4],因此作为一种具有极大潜力和有效的知识获取工具而受到人工智能工作者的广泛关注.

形式概念分析(formal concept analysis)是德国学者Wille[5]提出的一种从形式背景(formal context)建立概念

* Supported by the National Natural Science Foundation of China under Grant Nos.60573074, 70471003 (国家自然科学基金); the Natural Science Foundation of Shanxi Province of China under Grant No.2007011040 (山西省自然科学基金); the Foundation of Doctoral Program Research of the Ministry of Education of China under Grant No.20050108004 (高等学校博士学科点专项科研基金)

Received 2005-11-23; Accepted 2006-07-05

格来进行数据分析和规则提取的强有力工具,已被广泛加以研究[6-9],并应用到机器学习[10]、软件工程[11]和信息获取[12]等领域.

对形式概念分析和粗糙集理论进行结合研究,目前已有一些成果.Kent[13]和 Yao[14]将粗糙集理论中上下近似的思想引入到形式概念分析中,分别讨论了形式概念分析中的几种近似算子.王志海等人[15]研究了在形式概念分析中如何实现粗糙集合的基本运算,并运用这些运算来求得函数依赖.张文修教授[16]将粗糙集理论中属性约简和辨识矩阵的概念引入到形式概念分析中,实现了形式背景中冗余知识的约简.我们[17]将包含度和偏序集的概念引入到形式概念分析中,对形式概念分析中的一些基本概念分别用包含度和偏序集加以表示.然而,这些文献主要侧重于将粗糙集的概念引入到形式概念分析中,并没有对形式概念分析和粗糙集理论进行融合研究.本文侧重于建立粗糙集和形式概念分析之间融合的理论基础.本文利用形式概念分析中的名义梯级背景[6](nominal scale)和平面梯级[6](plain scaling)的概念,论证了粗糙集理论中的上下近似、属性依赖等核心概念都可以在相应的衍生背景中进行表示,并指出了利用梯级的概念可以对粗糙集理论进行扩展.

本文第 1 节和第 2 节简要介绍粗糙集理论和形式概念分析的基本概念.第 3 节利用形式概念分析对粗糙集理论的概念加以表示.第 4 节讨论粗糙集中对象集合的上近似与形式概念分析中对象集合的外延之间的联系.第 5 节简要分析粗糙集理论的局限性,指出利用形式概念分析对粗糙集进行扩展的重要性及可行性.第 6 节对全文进行总结.

1 粗糙集理论

定义 1[4].信息系统 S 是一个四元组 $S=(G,M,W,I)$.这里,G 称为论域,它的元素称为对象;M 称为属性集,它的元素称为属性;G 与 M 是有限的非空集合;$W=\bigcup_{m\in M}W_m$ 称为属性值域,W_m 为属性 m 的值域;I 为 $G\times M$ 到 W 的一个映射,对任意的 $g\in G,m\in M,I(g,m)\in W_m$,通常称 I 为信息函数或描述函数.

一个简单的信息系统表示见表 1[2],其中对象集为病人集合,属性集为症状集合.

Table 1 An information system

表 1 信息系统

Patient	Headache	Muscle-Pain	Temperature	Flu
$p1$	No	Yes	High	Yes
$p2$	Yes	No	High	Yes
$p3$	Yes	Yes	Very high	Yes
$p4$	No	Yes	Normal	No
$p5$	Yes	No	High	No
$p6$	No	Yes	Very high	Yes

设 $S=(G,M,W,I)$为一个信息系统,$B\subseteq M$ 为一个属性子集,$g,h\in G$,等价关系

$$IND(B)=\{(g,h)|\forall m\in B,I(g,m)=I(h,m)\}$$

称为 B 不可分辨关系.令

$$[g]_B=\{h\in G|(g,h)\in IND(B)\}.$$

我们称$[g]_B$为由 g 决定的 B 等价类,此时,$G/B=\{[g]_B|g\in G\}$为论域 G 的一个划分.

定义 2[4]. 设 $S=(G,M,W,I)$为一个信息系统,$A\subseteq G$ 为一个对象子集,$B\subseteq M$ 为一属性子集,A 关于 B 的下近似定义为

$$\underline{B}(A)=\{g\in G|[g]_B\subseteq A\}.$$

A 关于 B 的上近似定义为

$$\overline{B}(A)=\{g\in G|[g]_B\cap A\neq\varnothing\}.$$

定义 3[4]. 设 $S=(G,M,W,I)$为一个信息系统,$B,C\subseteq M$.如果 $m\in B$ 且 $IND(B)=IND(B-\{m\})$,则称 m 在 B 中是不必要的;否则称 m 在 B 中是必要的.进一步地,如果每一个 $m\in B$ 都为 B 中必要的,则称 B 是独立的;否则,称 B 是依赖的.如果 $C\subseteq B$,C 是独立的且 $IND(B)=IND(C)$,则称 C 是 B 的一个约简.称属性 C 依赖于属性 B,当且仅当 $IND(B)\subseteq IND(C)$.

2 形式概念分析

2.1 基本概念

定义 4[6]. 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 $g\in G,m\in M,(g,m)\in I$ 或者 gIm 表示"对象 g 具有属性 m".

一个形式背景见表 2.在这个形式背景中,对象集 G 为病人集合,属性集 M 为症状和症状值组成的序偶集合.对于症状,我们使用了缩写,H 表示 *Headache*,M 表示 *Muscle-pain*,T 表示 *Temperature*,F 表示 *Flu*.这样,对于对象 $g\in G$ 和属性 $m=(q,w)\in M,(g,m)\in I$ 表示病人 g 在症状 q 上的值为 w.例如,病人 $p1$ 具有属性(H,no),表示病人 $p1$ 在症状 H 上的取值为 *no*,即该病人不头疼,其余情况类推.

Table 2 A formal context

表 2 形式背景

	(H,yes)	(H,no)	(M,yes)	(M,no)	(T,high)	(T,very high)	(T,normal)	(F,yes)	(F,no)
p1	0	1	1	0	1	0	0	1	0
p2	1	0	0	1	1	0	0	1	0
p3	1	0	1	0	0	1	0	1	0
p4	0	1	1	0	0	0	1	0	1
p5	1	0	0	1	1	0	0	0	1
p6	0	1	1	0	0	1	0	1	0

定义 5[6]. 设 $K=(G,M,I)$为一个形式背景.对于集合 $A\subseteq G$,记

$$A^I=\{m\in M|(g,m)\in I,\forall g\in A\}.$$

相应地,对于集合 $B\subseteq M$,记

$$B^I=\{g\in G|(g,m)\in I,\forall m\in B\}.$$

定义 6[6]. 设 $K=(G,M,I)$为一个形式背景,$A\subseteq G,B\subseteq M$,称 $X=(A,B)$为 K 的一个概念,如果 $A^I=B$ 且 $B^I=A$.此时,称 A 为 X 的外延,B 为 X 的内涵.我们用 $\boldsymbol{B}(K)$记 K 的所有概念组成的集合.

定义 7[6]. 设 $K=(G,M,I)$为一个形式背景,$C_1=(A_1,B_1),C_2=(A_2,B_2)$是 K 的两个概念,规定

$$X_1\leq X_2\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_1\supseteq B_2).$$

显然,关系"$\leq$"是集合 $\boldsymbol{B}(K)$上的一个偏序,它可诱导出 $\boldsymbol{B}(K)$上的一个格结构,可以证明,它是一个完备格,此完备格称为形式背景 K 的概念格,在没有歧义的情况下,仍然记为 $\boldsymbol{B}(K)$.

定义 8[6]. 设 $K=(G,M,I)$为一个形式背景,$B_1,B_2\subseteq M$.任给 $g\in B_1^I$,若对任意 $m\in B_2$,有$(g,m)\in I$ 成立,则称属性集 B_1 蕴涵属性集 B_2,记为 $B_1\rightarrow B_2$,称为 K 上的蕴涵.

2.2 多值背景、梯级背景及梯级

定义 9[6]. 一个多值背景是四元组(G,M,W,I),其中,G 为对象集,M 为属性集,W 为属性值域,$I\subseteq G\times M\times W$ 为三元关系序偶,且当$(g,m,w)\in I,(g,m,v)\in I$ 时有 $w=v$ 成立.

显然,多值背景和信息系统具有相同的含义,因此,我们对它们并不作区别,在文中统一使用信息系统这一术语.

我们知道,形式概念分析仅对值域为 0 和 1 的形式背景进行研究,而通常的信息系统即(多值背景)的值域具有多值性,因此,要使用形式概念分析对信息系统进行研究,需要对信息系统进行转换,这就需要使用梯级背景对信息系统进行梯级.具体地说,首先对信息系统的每个属性 m 根据合适的意义或额外的知识得到相应的梯级背景,接着对信息系统利用梯级背景进行梯级就可以得到一个衍生的形式背景[6].这个衍生背景就是信息系统在形式概念分析中的相应背景,这样,我们就可以利用形式概念分析的方法对此背景进行处理.在这个过程中,梯级背景起到了一个中介作用,并不是最终的形式背景.

梯级背景的形式化定义如下:

定义 10[6]. 设(G,M,W,I)为一个信息系统,$m\in M$.m 的一个梯级背景(scale)是一个形式背景 $S_m=(G_m,M_m,I_m)$,

满足

$$m(G)=\{m(g)|g\in G\}\subseteq G_m.$$

由于在定义 10 中仅要求 $m(G)\subseteq G_m$,没有对 G_m 和 M_m 作更多的限制,所以,我们通常需要对相应的梯级背景加以限制,使得与实际情况更相符合.

定义 11[6]. 设(G,M,W,I)为一个信息系统,$m\in M$.m 的一个名义梯级背景(nominal scale)是一个形式背景$S_m=(G_m,M_m,I_m)$,它满足:

1) $G_m=M_m=m(G)=\{m(g)|g\in G\}$,

2) 对 $g\in G_m,n\in M_m$,有 $gI_mn\Leftrightarrow g=n$.

对于表 1 中的属性 *Temperature*,其名义梯级背景见表 3,其他属性的名义梯级背景见表 4.

Table 3 The nominal scale of attribute *Temperature*

表 3 属性 *Temperature* 的名义梯级背景

	Normal	High	Very high
Normal	1	0	0
High	0	1	0
Very high	0	0	1

Table 4 The nominal scale of other attributes

表 4 其余属性的名义梯级背景

	Yes	No
Yes	1	0
No	0	1

利用属性的梯级背景可以将一个信息系统通过梯级转换为形式背景.一般说来,梯级的方式也是比较随意的,也就是说,我们可以按照信息系统的特点及相应的领域知识对信息系统进行梯级.本文仅考虑一种梯级,即平面梯级.

定义 12[6]. 设(G,M,W,I)为一信息系统,$S_m=(G_m,M_m,I_m)$为梯级背景,其中,$m\in M$.记

$$\dot{M}_m=\{m\}\times M_m=\{(m,n)\mid n\in M_m\}.$$

一个由平面梯级衍生的背景(derived context with respect to plain scaling)为一个形式背景(G,N,J),其中,$N=\bigcup_{m\in M}\dot{M}_m$,对于 $g\in G,(m,n)\in\dot{M}_m$,二元关系 J 由下式确定:

$$gJ(m,n)\Leftrightarrow I(m,g)=w \text{ 且 } wI_mn.$$

直观上,一个由平面梯级衍生的背景就是保持原信息系统中的对象不变,由梯级背景 S_m 的梯级属性代替多值属性 m 而得到的一个形式背景,而结合信息系统和梯级背景得到衍生背景的过程便是平面梯级.对表 1 进行平面梯级后衍生的背景见表 2,其中所用到的名义梯级背景见表 3 和表 4.

3 形式概念分析对粗糙集的表示

本节中,我们将用形式概念分析对粗糙集理论中的一些概念加以表示.下文中,若非特别说明,我们对信息系统进行平面梯级所用到的梯级背景均是名义梯级背景.

定理 1(等价类表示). 设(G,M,W,I)为一个信息系统,(G,N,J)为相应的平面梯级衍生背景,$A\subseteq G,B\subseteq M,B_N$ 为信息系统中属性集 B 在衍生背景中相应属性集,即

$$B_N=\{(m,n)|m\in B,n\in M_m\}.$$

对于 $g\in G,[g]_B$ 表示在信息系统中由 g 决定的 B 等价类,在衍生的形式背景中,令

$$\overline{[g]}_B=\{h\in G\mid g^J\cap B_N=h^J\cap B_N\},$$

其中,g^J 和 h^J 表示衍生背景中对象 g 和对象 h 相对应的内涵,则有 $\overline{[g]}_B=[g]_B$.

证明:首先证明 $[g]_B\subseteq\overline{[g]}_B$.设 $h\in[g]_B$,由 $[g]_B$ 的定义可知$(g,h)\in IND(B)$,再由不可分辨关系的定义,对任意的 $m\in B$,我们有 $I(g,m)=I(h,m)$.

令 $I(g,m)=I(h,m)=w\in W_m$.在名义梯级背景 S_m 中,对任意的 $n\in M_m$,我们有 $I(g,m)=w$ 且 wI_mn,当且仅当 $I(h,m)=w$ 且 wI_mn,即 $gJ(m,n)\Leftrightarrow hJ(m,n)$.由 m 和 n 的任意性及 B_N 的定义可知,对任意的 $b\in B_N$ 有 $gJb\Leftrightarrow hJb$,因此,

$g^J \cap B_N = h^J \cap B_N$,即 $h \in \overline{[g]}_B$.

其次,我们证明 $\overline{[g]}_B \subseteq [g]_B$.设 $h \in \overline{[g]}_B$,对任意的 $m \in B$ 及 $n \in M_m$,由 $\overline{[g]}_B$ 和 B_N 的定义有 $g^J \cap B_N = h^J \cap B_N$,即 $gJ(m,n) \Leftrightarrow hJ(m,n)$.因此,存在 $w_1, w_2 \in m(G) = G_m$,使得

$$I(g,m)=w_1 \text{ 且 } w_1 I_m n \Leftrightarrow I(h,m)=w_2 \text{ 且 } w_2 I_m n$$

成立.由于 $w_1, w_2 \in G_m, n \in M_m$,从而由定义 11 有,$w_1 = n \Leftrightarrow w_1 I_m n \Leftrightarrow w_2 I_m n \Leftrightarrow w_2 = n$,即 $w_1 = w_2 = n$,这意味着对任意 $m \in B, I(g,m) = I(h,m)$.因而有 $(g,h) \in IND(B)$,从而 $h \in [g]_B$.

因为 $[g]_B \subseteq \overline{[g]}_B$ 且 $\overline{[g]}_B \subseteq [g]_B$,所以 $\overline{[g]}_B = [g]_B$. □

例 1:在信息系统表 1 和衍生的名义梯级背景表 2 中,取 $B=\{Muscle\text{-}pain, Temperature\}$,通过计算得到

$$IND(B)=\{\{p1\},\{p2,p5\},\{p3,p6\},\{p4\}\},$$

$$B_N=\{(M,yes),(M,no),(T,high),(T,very\ high),(T,normal)\},$$

$$\overline{[p1]}_B=\{p1\},\overline{[p2]}_B=\{p2,p5\},\overline{[p3]}_B=\{p3,p6\},\overline{[p4]}_B=\{p4\}.$$

由 $IND(B)$可以看出,对任意的 $g \in G$,都有 $\overline{[g]}_B = [g]_B$.

另外,由定理 1 可以看出,对任意的 $g \in G$,在衍生背景可以得到相应的 $\overline{[g]}_B$,再由 $\overline{[g]}_B = [g]_B$,从而可以得到一个与 G/B 完全相同的划分,即 $G/B = \{[g]_B \mid g \in G\} = \{\overline{[g]}_B \mid g \in G\}$,这表明粗糙集中的等价划分可以通过衍生背景来实现.

定理 2(上下近似表示). 设 (G,M,W,I)为一个信息系统,(G,N,J)为相应的平面梯级衍生背景,$A \subseteq G, B \subseteq M$, $\overline{B}(A)$ 和 $\underline{B}(A)$ 分别为信息系统中 A 关于 B 的上近似和下近似,则

$$\overline{B}(A)=\{g \in G \mid \overline{[g]}_B \cap A \neq \varnothing\},$$

$$\underline{B}(A)=\{g \in G \mid \overline{[g]}_B \subseteq A\}.$$

证明:由定理 1 知 $\overline{[g]}_B = [g]_B$,再由定义 2 可知结论成立. □

上下近似在粗糙集中是最基本的也是最重要的概念,许多工作都离不开上下近似.例如,粗糙集中的近似精度是基于上下近似,它将集合的不精确性定量化.由等价关系 $B \subseteq M$ 定义的集合 $A \subseteq G$ 的近似精度为

$$\alpha_B(A)=\frac{\underline{B}(A)}{\overline{B}(A)}.$$

基于定理 2,我们可以对近似精度在衍生背景中进行如下表示

$$\alpha_B(A)=\frac{\{g \in G \mid \overline{[g]}_B \subseteq A\}}{\{g \in G \mid \overline{[g]}_B \cap A \neq \varnothing\}}.$$

在粗糙集中,还有其他一些重要的概念[4]也是基于上下近似的,如集合的上粗相等、下粗相等、粗相等,粗糙集的 4 个拓扑特征,即粗糙可定义、内不可定义、外不可定义和全不可定义,这些概念都可以由定理 2 直接推导出来.

定理 3. 设 (G,M,W,I)为一个信息系统,(G,N,J)为相应的平面梯级衍生背景,$B \subseteq M, m \in B$,则

1) m 在 B 中是不必要的,当且仅当对任意的 $g \in G, \overline{[g]}_B = \overline{[g]}_{B-\{m\}}$;

2) B 是独立的,当且仅当对任意的 $m \in B$,存在 $g \in G$,使得 $\overline{[g]}_{B-\{m\}} \subsetneq \overline{[g]}_B$.

证明:1) 在粗糙集中,m 在 B 中是不必要的,当且仅当 $IND(B)=IND(B-\{m\})$,即对任意的 $g \in G$ 有 $[g]_B=[g]_{B-\{m\}}$,由定理 1 知 $\overline{[g]}_B = [g]_B$ 且 $\overline{[g]}_{B-\{m\}} = [g]_{B-\{m\}}$,因此有 $\overline{[g]}_B = \overline{[g]}_{B-\{m\}}$,故有 1)成立.

2) B 是独立的,当且仅对当任意的 $m \in B$ 都为 B 中必要的,即对任意的 $m \in B, IND(B) \neq IND(B-\{m\})$,由于 $IND(B-\{m\}) \supseteq IND(B)$,所以 $IND(B-\{m\}) \subsetneq IND(B)$,此时,必然存在 $g \in G$ 且 $[g]_{B-\{m\}} \subsetneq [g]_B$,即 $\overline{[g]}_{B-\{m\}} \subsetneq \overline{[g]}_B$. □

定理 4(约简表示). 设 (G,M,W,I)为一个信息系统,(G,N,J)为相应的平面梯级衍生背景,$C \subseteq B \subseteq M$,则 C 是 B 的一个约简当且仅当

1) 对任意的 $c \in C$,存在 $g \in G$,使得 $\overline{[g]}_{C-\{c\}} \subsetneq \overline{[g]}_C$;

2) 对任意的 $g\in G,\overline{[g]}_B=\overline{[g]}_C$.

证明:由粗糙集约简的定义,C 是 B 的一个约简,当且仅当 C 是独立的且 $IND(B)=IND(C)$.由定理 3 的 2)可知,C 是独立的,当且仅当对任意的 $c\in C$,存在 $g\in G$,使得 $\overline{[g]}_{C-\{c\}}\subsetneq\overline{[g]}_C$,即 C 是独立的,当且仅当条件 1)成立.又 $IND(B)=IND(C)$,当且仅当对任意的 $g\in G,[g]_B=[g]_C$,即 $\overline{[g]}_B=\overline{[g]}_C$,因此 $IND(B)=IND(C)$,当且仅当条件 2)成立.于是定理成立. □

例 2:在信息系统表 1 中,取 $B=\{Headache,Muscle\text{-}pain,Temperature\}$,$C=\{Headache,Temperature\}$,计算可得

$$IND(B)=IND(C)=\{\{p1\},\{p3\},\{p4\},\{p2,p5\},\{p6\}\},$$
$$IND(C-\{Temperature\})=\{\{p1,p4,p6\},\{p2,p3,p5\}\}\neq IND(C),$$
$$IND(C-\{Headache\})=\{\{p1,p2,p5\},\{p4\},\{p3,p6\}\}\neq IND(C).$$

从而,C 是 B 的一个约简.

在衍生的名义梯级背景表 2 中,计算得到

$$B_N=\{(H,yes),(H,no),(M,yes),(M,no),(T,high),(T,very\ high),(T,normal)\},$$
$$C_N=\{(H,yes),(H,no),(T,high),(T,very\ high),(T,normal)\}.$$

相应于定理 4 中的条件 2),有

$$\overline{[p1]}_B=\overline{[p1]}_C=\{p1\},\overline{[p2]}_B=\overline{[p2]}_C=\{p2,p5\},\overline{[p3]}_B=\overline{[p3]}_C=\{p3\},$$
$$\overline{[p4]}_B=\overline{[p4]}_C=\{p4\},\overline{[p6]}_B=\overline{[p6]}_C=\{p6\}.$$

对于 $\{Headache\}\in C$,存在 $p1$,

$$\overline{[p1]}_{C-\{Headache\}}=\{p1,p2,p5\}\subsetneq\{p1\}=\overline{[p1]}_C.$$

对于 $\{Temperature\}\in C$,存在 $p1$,

$$\overline{[p1]}_{C-\{Temperature\}}=\{p1,p4,p6\}\subsetneq\{p1\}=\overline{[p1]}_C.$$

即 C 满足定理 4 的条件 1),从而验证了 C 为 B 的一个约简.

定理 5(依赖表示). 设 (G,M,W,I) 为一个信息系统,(G,N,J) 为相应的平面梯级衍生背景,$B,C\in M$,则 C 依赖于 B,当且仅当对任意的 $g\in G$,存在 $h\in G$,使得 $\overline{[g]}_B\subseteq\overline{[h]}_C$.

证明:由于 C 依赖于 B,当且仅当 $IND(B)\subseteq IND(C)$,即对任意 $g\in G$,存在 $h\in G$,使得 $[g]_B\subseteq[h]_C$.再由定理 1 有 $\overline{[g]}_B=[g]_B$ 且 $\overline{[h]}_C=[h]_C$,从而 $\overline{[g]}_B\subseteq\overline{[h]}_C$,结论成立. □

例 3:在信息系统表 1 中,取 $B=\{Headache,Temperature\}$,$C=\{Muscle\text{-}pain\}$,计算可得

$$IND(B)=\{\{p1\},\{p3\},\{p4\},\{p2,p5\},\{p6\}\},IND(C)=\{\{p1,p3,p4,p6\},\{p2,p5\}\},$$

由 $IND(B)\subseteq IND(C)$ 可知,C 依赖于 B.

在衍生的名义梯级背景表 2 中,可得

$$B_N=\{(H,yes),(H,no),(T,high),(T,very\ high),(T,normal)\},$$
$$C_N=\{(M,yes),(M,no)\}.$$

进而有

$$\overline{[p1]}_B\subseteq\overline{[p1]}_C,\overline{[p2]}_B\subseteq\overline{[p2]}_C,\overline{[p3]}_B\subseteq\overline{[p3]}_C,\overline{[p4]}_B\subseteq\overline{[p4]}_C,\overline{[p6]}_B\subseteq\overline{[p6]}_C.$$

从而验证了 C 依赖于 B.

4 上近似与外延之间的联系

设 (G,M,W,I) 为一个信息系统,(G,N,J) 为相应的平面梯级衍生背景,在 (G,M,W,I) 中对象集 $A\subseteq G$ 包含于它的上近似 $\overline{M}(A)$ 中,即 $A\subseteq\overline{M}(A)$.而在 (G,N,J) 中,对象集 A 包含在它的外延 A^{JJ} 中,即 $A\subseteq A^{JJ}$.那么,$\overline{M}(A)$ 和 A^{JJ} 之间有什么关系呢?一般说来,我们有下面的定理.

定理 6. 设 (G,M,W,I) 为一个信息系统,(G,N,J) 为相应的平面梯级衍生背景,$A\subseteq G$,$\overline{M}(A)$ 为 A 在 (G,M,W,I) 中

相对于 M 的上近似,A^{JJ} 为 A 在衍生背景(G,N,J)中相对应的外延,则 $\overline{M}(A)\subseteq A^{JJ}$.

证明:若 $g\in\overline{M}(A)$,由上近似的定义有,$[g]_M\cap A\neq\varnothing$,此时,存在 $g_1\in[g]_M$ 且 $g_1\in A$.由 $[g]_M=\overline{[g]}_M$,有 $g_1\in\overline{[g]}_M$,而 $\overline{[g]}_M=\{h\in G\mid g^J\cap M_N=h^J\cap M_N\}=\{h\in G\mid g^J\cap N=h^J\cap N\}=\{h\in G\mid g^J=h^J\}$,所以有 $g^J=g_1^J$.在衍生背景中,这意味着 $g^{JJ}=g_1^{JJ}$.又因为 $g_1\in A$,因此 $g_1^{JJ}\subseteq A^{JJ}$,从而 $g\in g^{JJ}=g_1^{JJ}\subseteq A^{JJ}$,即 $g\in A^{JJ}$,所以 $\overline{M}(A)\subseteq A^{JJ}$. □

需要说明的是,上述定理结论中,逆包含一般情况下并不一定成立,例如集合$\{p3,p4\}$,

$$\overline{M}(\{p3,p4\})=\{p3,p4\},\{p3,p4\}^{JJ}=\{p1,p3,p4,p6\}.$$

显然,$\overline{M}(\{p3,p4\})\subseteq\{p3,p4\}^{JJ}$,但逆包含不成立.

5 形式概念分析对粗糙集的扩展

我们知道,在使用粗糙集进行分析和处理数据时,并不需要额外的信息或知识,这是粗糙集的优点,但也是它的不足之处.也就是说,粗糙集在处理某些需要额外信息或知识的数据时,就会显得无能为力.例如表 1 中,对于属性 *Temperature*,在默认的情况下,粗糙集理论会认为属性值 *normal*,*high*,*very high* 之间是无关的,也就是说,属性 *Temperature* 的 3 个属性值之间无蕴涵关系,从表 3 我们可以看出这一点.结合表 1 和表 2,我们还可以看出,如果一个病人(比如病人 *p*3)的 *Temperature* 是 *very high*,即这个病人的体温很高,那么这个对象的 *Temperature* 必然不是 *high*,即这个对象的体温不高.虽然有时我们会认为 *high* 和 *very high* 是无关的,然而更多的时候,我们可能认为 *high* 和 *very high* 之间是相关的,即我们可能认为 *very high* 也是 *high*,见表 5.

Table 5 Another scale of attribute *Temperature*

表 5 属性 *Temperature* 的另一梯级背景

	Normal	*High*	*Very high*
Normal	1	0	0
High	0	1	0
Very high	0	1	1

将表 4 和表 5 作为梯级背景,以平面梯级的方式对表 1 进行梯级,得到的衍生背景见表 6.

Table 6 Formal context

表 6 形式背景

Patient	(*H*,yes)	(*H*,no)	(*M*,yes)	(*M*,no)	(*T*,*high*)	(*T*,*very high*)	(*T*,*normal*)	(*F*,yes)	(*F*,no)
*p*1	0	1	1	0	1	0	0	1	0
*p*2	1	0	0	1	1	0	0	1	0
*p*3	1	0	1	0	1	1	0	1	0
*p*4	0	1	1	0	0	0	1	0	1
*p*5	1	0	0	1	1	0	0	0	1
*p*6	0	1	1	0	1	1	0	1	0

从表 5 和表 6 我们可以看出:若一个病人的体温正常(normal),那么,此病人的体温不高(没有属性 high),也不是非常高(没有属性 very high);若一个病人的体温高(high),那么,此病人的体温不正常(没有属性 normal),但不是非常高(没有属性 very high);若一个病人的体温非常高(very high),那么,此病人的体温不正常(没有属性 normal),而且体温高(有属性 high).

从形式概念分析的角度来看,(*H*,yes)→(*T*,*high*)并不是形式背景表 2 的一条蕴涵.也就是说,并不是所有的病人头痛时都会导致体温高,这是因为病人 *p*3 头痛(具有属性(*H*,yes))但体温却不高(不具有属性(*T*,*high*)).然而(*H*,yes)→(*T*,*high*)却是形式背景表 6 的一条蕴涵,从表 6 我们可以清楚地看出这一点.从实际的意义看,因为 *p*3 头痛且其体温非常高,因而可以说,*p*3 体温也是高的,所以蕴涵(*H*,yes)→(*T*,*high*)与实际是相符合的.

显然,从形式背景表 2 可以得到信息系统表 1.同样,相应于表 6,我们可以得到表 7.

由表 7 可以看出,它已不再是一个信息系统,原因是病人 *p*3 和 *p*6 在属性 *Temperature* 下的值已经变为一个集合,粗糙集理论中并没有相应的处理方法,而形式概念分析却可以从相应的表 6 中提取出蕴涵.

由上面的例子可以看出,粗糙集理论的局限在于被处理的信息系统值域中的各值之间必须是无蕴涵关系

的.我们前面的表示理论也说明了这一点.在上面的例子中,由于我们选择了不同的梯级背景,而得到了不同的衍生背景,进而必然会产生不同的结论.因而我们可以利用这一点来对粗糙集理论进行扩展,一般情况下,我们可以利用常识或额外的信息或知识来得到梯级背景,这样就会产生相应的衍生梯级背景.此时,如果我们从衍生背景中提取规则,则得到的规则就更为细致,进而可能获取一些额外的知识,这一点在粗糙集中是无法做到的.

事实上,由表 7 我们可以看出,只有对全部的属性采用名义梯级背景得到的衍生背景才可以还原为信息系统,而采用其他梯级背景[6]得到的衍生背景将不能还原为信息系统.利用这一点,我们可以对粗糙集理论中的信息系统概念进行扩展,而研究扩展意义下的粗糙集模型也将是一项有意义的工作.

Table 7 The corresponding table of Table 6

表 7 表 6 的相应表格

Patient	Headache	Muscle-Pain	Temperature	Flu
$p1$	No	Yes	*High*	Yes
$p2$	Yes	No	*High*	Yes
$p3$	Yes	Yes	{*very high, high*}	Yes
$p4$	No	Yes	*Normal*	No
$p5$	Yes	No	*High*	No
$p6$	No	Yes	{*very high, high*}	Yes

6 结 论

对粗糙集和形式概念分析进行融合是人工智能领域的研究热点.本文利用形式概念分析中名义梯级背景和梯级的概念,证明了粗糙集中的上下近似、属性依赖等核心概念都可以在衍生的平面梯级背景中进行表示,这也为两者的融合提供了一个理论平台.用形式概念分析对信息系统进行研究的优点在于,对不同的领域可以灵活地选择梯级背景和梯级方式,其缺点是梯级后的属性较多,而这个问题可以采用文献[16]中提出的属性约简理论进行处理.

在第 5 节中,我们还揭示粗糙集在分析处理数据时的局限性,对于不同的领域和不同的问题,我们可以依据实际情况及一些额外的信息选择不同的梯级背景和梯级方式,由此获取的知识可能会更具有应用价值,我们将进一步对此进行研究.

References:

[1] Pawlak Z. Rough sets. Int'l Journal of Computer and Information Sciences, 1982,11(5):341−356.

[2] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordecht: Kluwer Academic Publishers, 1991.

[3] Li DY, Zhang B, Yee L. On knowledge reduction in inconsistent decision information systems. Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004,12(5):651−672.

[4] Liang JY, Li DY. Uncertainty and Knowledge Acquisition in Information Systems. Beijing: Science Press, 2005 (in Chinese).

[5] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts. In: Rival I, ed. Ordered Sets. Dordrecht: Reidel. 1982. 445−470.

[6] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations. Berlin: Springer-Verlag, 1999.

[7] Qu KS, Liang JY, Wang JH, Shi ZZ. The algebraic properties of concept lattice. Journal of Systems Science and Information, Research Information Ltd UK, 2004,2(2):271−277.

[8] Xie ZP, Liu ZT. A fast incremental algorithm for building concept lattice. Chinese Journal of Computers, 2002,25(5):490−495 (in Chinese with English abstract).

[9] Liang JY, Wang JH. An algorithm for extracting rule-generating sets based on concept lattice. Journal of Computer Research and Development, 2004,41(8):1339−1344 (in Chinese with English abstract).

[10] Zupa B, Bohance M. Learning by discovering concept hierarchies. Artificial Intelligence, 1999,109(1−2):211−242.

[11] Dekel U. Revealing Java class structure with concept lattices [MS. Thesis]. Technion—Israel Institute of Technology, 2003.

[12] Valtchev P, Missaoui R, Godin R, Meridji M. Generating frequent itemsets incrementally: Two novel approaches based on Galois lattice theory. Journal of Experimental and Theoretical Artificial Intelligence, 2002,14(2−3):115−142.

[13] Kent RE. Rough concept analysis. In: Ziarko WP, ed. Rough Sets and Fuzzy Sets Knowledge Discovery (RSKD'93). London: Springer-Verlag, 1994. 248-255.

[14] Yao YY. Concept lattices in rough set theory. In: Dick S, Kurgan L, Pedrycz W, Reformat M, eds. Proc. of the 2004 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2004). IEEE, 2004. 796-801.

[15] Wang ZH, Hu KY, Liu ZT, Zhang DC, Huang HK. Rough set operations and functional dependence generation based on concept lattice. Journal of Tsinghua University, 1998,38(S2):1-4 (in Chinese with English abstract).

[16] Zhang WX, Wei L, Qi JJ. Attribute reduction in concept lattice. Science in China (Series E), 2005,35(6):628-639 (in Chinese with English abstract).

[17] Qu KS, Zhai YH. Posets, inclusion degree theory and FCA. Chinese Journal of Computers, 2006,29(2):219-226 (in Chinese with English abstract).

附中文参考文献:

[4] 梁吉业,李德玉.信息系统中的不确定性与知识获取.北京:科学出版社,2005.

[8] 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.

[9] 梁吉业,王俊红.基于概念格的规则产生集挖掘算法.计算机研究与发展,2004,41(8):1339-1344.

[15] 王志海,胡可云,刘宗田,张莫成,黄厚宽.概念格上的粗糙集合运算与函数依赖生成.清华大学学报,1998,38(S2):1-4.

[16] 张文修,魏玲,祁建军.概念格的属性约简理论与方法.中国科学(E辑),2005,35(6):628-639.

[17] 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.

曲开社(1954－),男,山西运城人,副教授,主要研究领域为人工智能,粗糙集,形式概念分析.

翟岩慧(1981－),男,硕士生,主要研究领域为粗糙集,形式概念分析.

梁吉业(1962－),男,教授,博士生导师,CCF高级会员主要研究领域为人工智能,粗糙集理论及应用.

李德玉(1965－),男,教授,博士生导师,CCF高级会员,主要研究领域为人工智能,粗糙集理论及应用.