对代数观 Rough 集理论的信息观解释*)

龚 勋[1,2] 王国胤[2]

(西南交通大学计算机与通信工程学院 成都 610031)[1]

(重庆邮电学院计算机科学与技术研究所 重庆 400065)[2]

摘 要 Rough 集理论的代数观点和信息论观点在不相容决策表中的不等价性导致了这两种观点得出的结论不一致。我们研究了使这两种观点等价的条件,定义一种新的决策表信息熵计算方法,在此方法的基础上给出了 Routh 集理论代数观的一种新的信息观解释,并证明了这种新的信息观与代数观是等价的。新的信息观定义为寻找高效的知识约简算法奠定了基础。

关键词 Rough 集,决策表,分割,信息观,代数观

Illustrating the Algebra View of Rough Set Theory with Information View

GONG Xun[1,2] WANG Guo-Yin[2]

(School of Computer Science and Communication Engineering, Southwest Jiaotong University, Chengdu 610031)[1]

(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065)[2]

Abstract The inequality between algebra view and information view of rough set theory leads to the different result from the inconsistent decision table. Firstly, the differences between the two kinds of the views are studied. Moreover, a new entropy calculation method for decision table is proposed in this paper. Based on this method, a new information view that could comprehensively illustrate the algebra view is introduced, and more efficient algorithms of knowledge reduction will be developed on the base of the new information view which is proved equals to the algebra view.

Keywords Rough set, Decision table, Division, Information view, Algebra view

1 引言

由波兰学者 Pawlak 教授提出的粗糙集理论是分析和处理不一致、不完整、不精确信息系统的有力工具[1],近年来在机器学习、数据挖掘、智能控制等多个领域得到了广泛的应用。Rough 集理论存在代数观和信息观两种观点,分别从不同的角度描述 Rough 集理论。代数观的核心思想是在保持分类能力不变的前提下,通过知识约简导出规则,也就是通过不可辨识关系和集合包含关系定义知识的粗糙性。而信息观则是从信息论的角度用信息熵来讨论知识的粗糙性。

Rough 集理论的信息观和代数观在不相容决策表中是不等价的[2~5],这种不等价根源于决策表中的不相容信息。具体地说,信息观考虑了决策表中不相容部分相互间的信息,而代数观却忽略了这部分信息[6]。因此,要让信息观同代数观一致,或者让代数观同信息观一致,就需要对这部分信息进行取舍。

本文考虑用信息观来解释代数观,即让信息观同代数观一致。正因为代数观方法忽略了决策表中不相容部分的影响,为了得到与代数观一致的结论,我们提出一种基于正域的重组分割方法把决策表中不相容子块分开,从而避免了在不相容子块合并时对信息熵的影响;为了建立子决策表与原始决策表之间的联系,我们用一种新的决策表信息熵计算方法重新定义了属性重要性、决策表属性必要性、属性约简、核属性,并证明了这些定义与代数观是一致的。

2 基本概念

为了便于叙述,这里先对文中涉及到的 Rough 集理论相关概念进行简单介绍。

2.1 信息表知识表达系统概念

定义 1(信息系统)[5] 信息系统 S 可以表示为 $S=\langle U,R,V,f\rangle$,U 是对象集合,称为论域;$R=C\cup D$,C 是条件属性集合,D 是决策属性集合;$V=\bigcup_{r=R}V_r$,是属性值的集合;$f:U\times R\to V$ 是一个信息函数,它指定 U 中每一个对象 x 的属性值。

定义 2(决策表)[5] 决策表是一个信息系统 $S=\langle U,R,V,f\rangle$,$R=C\cup D$ 是属性集合,子集 C 和 D 分别称为条件属性集和决策属性集,$D\neq\varnothing$。

2.2 代数观相关概念

定义 3(上、下近似集)[5] 给定决策表信息系统 $S=\langle U,R,V,f\rangle$,对于每个子集 $X\subseteq U$ 和不分明关系 B,X 的下近似集和上近似集可以分别定义为:$B_-(X)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\subseteq X)\}$,$B^-(X)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\cap X\neq\phi)\}$,其中,$U\mid IND(B)=\{X\mid(X\subseteq U\wedge\forall_{x\in X,y\in X,b\in B}(b(x)=b(y)))\}$是不分明关系 B 对 U 的划分。

定义 4(分类质量)[5] 设集合族 $F=\{X_1,X_2,\cdots,X_n\}$($U=\bigcup_{i=1}^{n}X_i$)是论域 U 上定义的知识,B 是一个属性子集,定义 B 对 F 近似分类的质量为 $r_B(F)$ 为:$r_B(F)=\sum_{i=1}^{n}|B_-(X_i)|/|U|$。

定义 5(属性重要性)[5] F 是属性集 D 导出的分类,C 是条件属性集合,D 是决策属性集合,且 $A\subset C$,则对于任意属性 $a\in C-A$ 的重要性 $SGF(a,A,D)$ 定义为:$SGF(a,A,D)=r_{A\cup\{a\}}(F)-r_A(F)$。

这表示当我们在属性集 A 中增加属性 a 对 F 近似分类的质量的影响。

定义 6(属性必要性)[5] 设 U 为一个论域,P、Q 为 U 上

*)资助项目:国家自然科学基金(No. 60373111)、教育部新世纪优秀人才支持计划、重庆市自然科学基金重点项目、重庆市教委科技计划项目(No. 040505)。龚 勋 博士生,主要研究领域为人工智能、模式识别等;王国胤 博士,博士生导师,主要研究领域包括 Rough 集理论、神经网络、机器学习、数据挖掘等。

两个属性集合，若 $POS_P(Q)=POS_{P-\{r\}}(Q)$，则称 r 为 P 中相对于 Q 可省略的(不必要的)；否则，称 r 为 P 中相对于 Q 不可省略的(必要的)。

定义 7(独立性)[5] 设 U 为一个论域，P、Q 为 U 上的两个属性集合，若 P 中的每一 r 都是 P 中 Q 不可省略的，则称 P 为(相对于)Q 独立的。

定义 8(约简)[5] 设 U 为一个论域，P、Q 为 U 上的两个属性集合，若 P 的 Q 独立子集 $S(S\subset P)$ 有 $POS_S(Q)=POS_P(Q)$，则称 S 为 P 的 Q 约简。

可以记 P 的所有 Q 约简属性集为 $RED_Q(P)$。

定义 9(核属性)[5] 设 U 为一个论域，P、Q 为 U 上的两个属性集合，P 的所有 Q 不可省略原始属性集合称为 P 的 Q 核，记为 $CORE_Q(P)$。

利用属性必要性及核属性定义，可以得到以下命题：

命题 1[6] 属性 $a\in CORE(C)$ 当且仅当 a 是必要属性。

2.3 信息观相关概念

定义 10(知识概率分布)[5] 设 P、Q 在 U 上导出的划分分别为 X、$Y(X=\{X_1,X_2,\cdots,X_n\},Y=\{Y_1,Y_2,\cdots,Y_m\})$，则 P、Q 在 U 的子集组成的 σ 代数上的概率分布为：

$$[X:P]=\begin{bmatrix}X_1 & X_2 & \cdots & X_n\\ p(X_1) & p(X_2) & \cdots & p(X_n)\end{bmatrix},$$

$$[Y:p]=\begin{bmatrix}Y_1 & Y_2 & \cdots & Y_m\\ p(Y_1) & p(Y_2) & \cdots & p(Y_m)\end{bmatrix}$$

其中 $p(X_i)=|X_i|/|U|,i=1,2,\cdots,n,p(Y_j)=|Y_j|/|U|,j=1,2,\cdots,m$。

有了知识的概率分布定义后，根据信息论就可以定义知识的熵与条件熵的概念。

定义 11(信息熵)[5] 知识(属性集合)P 的熵 $H(P)$ 定义为：$H(P)=-\sum_{i=1}^{n}p(X_i)\log(p(X_i))$。

定义 12(条件熵)[5] 知识(属性集合)$Q(U|IND(Q)=\{Y_1,Y_2,\cdots,Y_m\})$ 相对于知识(属性集合)P $(U|IND(P)=\{X_1,X_2,\cdots,X_n\})$ 的条件熵 $H(Q|P)$ 定义为：

$$H(Q|P)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))$$

其中，$P(Y_j|X_i)=|Y_j\cap X_i|/|X_i|,i=1,2,\cdots,n,j=1,2,\cdots,m$。

我们对定义 12 中的条件熵公式进行分解，记 $DH(Q,\{X_i\})=-\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))$，故 $H(Q|P)$ 可以记为：$H(Q|P)=\sum_{i=1}^{n}p(X_i)DH(Q,\{X_i\})$。

3 基于正域的决策表重组分割方法

决策表中不相容部分信息相互间产生的条件熵是导致代数观与信息观不一致的根源。我们通过把存在不相容信息的决策表进行基于正域的重组分割得到一系列子决策表，把不相容等价类分开，从而在每个子决策表中避免了不相容信息对条件熵计算的影响。

下面，首先给出正域、负域在信息观下的定义。

3.1 信息观下的正域与负域

定义 13(信息观下的正域) 设 U 为一个论域，P、Q 是 U 上的两个属性集合，设 P、Q 分别对 U 导出的划分为 $U|IND(P)=\{X_1,X_2,\cdots,X_n\},U|IND(Q)=\{Y_1,Y_2,\cdots,Y_m\}$，$Q$ 的 P 正域记为 $IPOS_P(Q)$，正域定义为：$IPOS_P(Q)=\bigcup\{X_i|p(X_i)DH(Q,\{X_i\})=0\},1\leqslant i\leqslant n$。

定义 14(信息观下的负域) 设 U 为一个论域，P、Q 是 U 上的两个属性集合，设 P、Q 分别对 U 导出的划分为 Q 的 P 负域记为 $INEG_P(Q)$，负域定义为：$INEG_P(Q)=U\backslash IPOS_P(Q)$。

命题 2 $IPOS_P(Q)=POS_P(Q)$

证明：(1)取 $X_i\in U|IND(P),1\leqslant i\leqslant n$，且 $p(X_i)DH(Q,\{X_i\})=0$，即 $-p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))=0$。假设 $X_i\not\subset POS_P(Q)$，那么一定存在 $Y_j,Y_k\in U|IND(Q)$，使得 $Y_j\cap X_i\neq\varnothing,Y_j\cap X_i\neq X_i,Y_k\cap X_i\neq\varnothing,Y_k\cap X_i\neq X_i$，故 $0<p(Y_j|X_i)<1,0<p(Y_k|X_i)<1$，所以 $-p(Y_j|X_j)\log(p(Y_j|X_i))>0,-p(Y_k|X_i)\log(p(Y_k|X_i))>0$。

又根据函数 $-p(x)\log(p(x))$ 的非负性可知，$-p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))>0$，与条件相矛盾，故 $X_i\subseteq POS_P(Q)$。

(2)对任意 $X_i\in U|IND(P),1\leqslant i\leqslant n$，$X_i\subseteq POS_P(Q)$，故存在 $Y_j\in U|IND(Q),1\leqslant j\leqslant m$，使得 $X_i\subseteq Y_j$，即 $p(Y_j|X_i)=1$，而对于任意 $k\in[1,m]$，且 $k\neq j$，都有 $X_i\cap Y_k=\varnothing$，即 $p(Y_k|X_i)=0$。所以，

$$p(X_i)DH(Q,\{X_i\})=-p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))=(-p(X_i)p(Y_j|X_i)\log(p(Y_j|X_i)))+(-p(X_i)\sum_{k=1,k\neq j}^{m}p(Y_k|X_i)\log(p(Y_k|X_i)))=0$$

根据 1,2 可知，$IPOS_P(Q)=POS_P(Q)$。证毕。

命题 3 $INEG_P(Q)=NEG_P(Q)$

证明：根据 $POS_P(Q)$、$NEG_P(Q)$ 的定义及命题 2，本命题显然成立。

3.2 基于正域的重组分割

利用信息观下正域、负域的概念，下面给出决策表基于正域的重组分割方法描述。

S

U	a	b	c	d
1	1	0	1	0
2	0	1	0	1
3	0	0	0	0
4	0	0	0	1
5	0	0	0	1
6	0	0	0	1
7	0	0	1	1
8	0	0	1	0
9	0	0	1	1

图 1 决策表 S

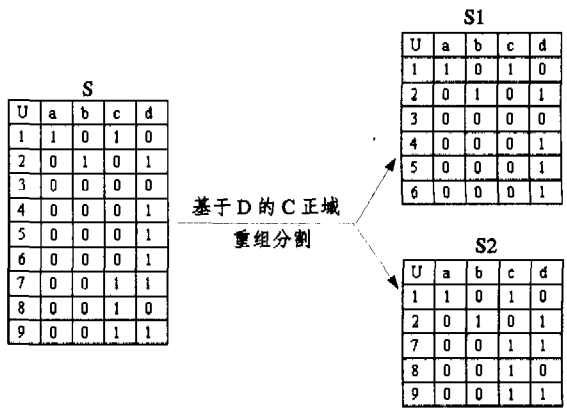

图 2 对决策表 S 进行基于 D 的 C 正域重组分割

决策表 $S=\langle U,R,V,f\rangle$，P、Q 是论域 U 上的两个属性集合，记 $IPOS_P(Q)|IND(P)=\{X'_1,X'_2,\cdots,X'_s\}$，$INEG_P(Q)|IND(P)=\{X''_1,X''_2,\cdots,X''_t\}$，令 $U_i=X''_i\cup IPOS_P(Q)$，得

到 t 个子决策表，$S_i=\langle U_i,R,V,f\rangle$，其中，$1\leqslant i\leqslant t$，这种分割方法称为对决策表 S 基于 Q 的 P 正域重组分割。

特别地，当 $P=C,Q=D$ 时，称为决策表 S 基于 D 的 C 正域重组分割。

例 1 对图 1 的决策表 S 进行基于 D 的 C 正域的重组分割，其中条件属性 $C=\{a,b,c\}$，决策属性 $D=\{d\}$。

易知 $IPOS_P(D)=\{1,2\}$，$INEG_P(D)=\{3,4,5,6,7,8,9\}$。

条件属性 C 对负域的划分为：$INEG_P(D)\mid IND(C)=\{X''_1,X''_2\}=\{\{3\ 4\ 5\ 6\},\{7\ 8\ 9\}\}$。

所以，$U_1=X''_1\cup IPOS_P(Q)=\{1,2,3,4,5,6\}$，$U_2=X''_2\cup IPOS_P(Q)=\{1,2,7,8,9\}$。分割结果如图 2 所示。

4 信息观新定义

对决策表进行基于正域的重组分割后，如何从子决策表得到原始决策表相关信息，是一个十分关键的问题。本节提出信息观下条件熵的一种新的计算方法来建立二者的联系，并以此为基础对信息观下的相关概念进行重新定义。

4.1 一种新的决策表条件熵定义

设 $S=\langle U,R,V,f\rangle$ 是一个决策表，P、Q 是 U 上的两个属性集合，对 S 进行基于 Q 的 P 正域重组分割，得到 t 个子决策表 $S_i=\langle U_i,R,V,f\rangle$，$(1\leqslant i\leqslant t)$。令 $IPOS_P(Q)\mid IND(P)=\{X'_1,X'_2,\cdots,X'_s\}$，$INEG_P(Q)\mid IND(P)=\{X''_1,X''_2,\cdots,X''_t\}$，则 $U_i=X''_i\cup IPOS_P(Q)$。

在子决策表 S_i 上，P、Q 是 U_i 上两个属性集合，设 P、Q 分别对 U_i 导出的划分为 $U_i\mid IND(P)=\{X_1,X_2,\cdots,X_n\}$，$U_i\mid IND(Q)=\{Y_1,Y_2,\cdots,Y_m\}$。$S_i$ 上知识 Q 相对知识 P 的条件熵记为 $H_{S_i}(Q|P)$，$H_{S_i}(Q|P)=-\sum_{j=1}^{n}p(X_j)\sum_{k=1}^{m}p(Y_k|X_j)\log(p(Y_k|X_j))$。同定义 12，对 $H_{S_i}(Q|P)$ 分解，记为：$H_{S_i}(Q|P)=\sum_{j=1}^{n}p(X_j)DH_i(Q,\{X_j\})$，其中 $DH_i(Q,\{X_j\})=-\sum_{k=1}^{m}p(Y_k|X_j)\log(p(Y_k|X_j))$，$1\leqslant j\leqslant n,1\leqslant i\leqslant t$。

命题 4 对决策表 S 进行基于 D 的 P 正域重组分割，则 $H_{S_i}(Q|P)=p(X''_i)DH_i(D,\{X''_i\})$。

证明：根据信息观下正域的定义可知正域的条件熵为 0，显然上式成立。

命题 4 说明，在子系统 S_i 中，条件熵 $H_{S_i}(Q|P)$ 完全是由不相容分块产生的。

定义 15（决策表新的条件熵定义） 对决策表 S 进行基于 Q 的 P 正域重组分割，则 Q 相对 P 的条件熵记为 $H_S(Q|P)$，定义为：$H_S(Q|P)=\sum_{i=1}^{t}H_{S_i}(Q|P)$。

定义 16（属性重要性） 设 $S=\langle U,R,V,f\rangle$ 是一个决策表系统，U 是论域，$R=C\cup D$，C 是条件属性集合，D 是决策属性集合，对于任意属性 a 相对 $C-\{a\}$ 的重要性 $SGF_S(a,C-\{a\},D)$ 定义为：对决策表 S 进行基于 Q 的 P 正域重组分割，$SGF_S(a,C-\{a\},D)=H_S(D|C-\{a\})-H_S(D|C)$。

4.2 基于新决策表条件熵的一些判定定理

引理 1 对决策表 S 进行基于 D 的 P 正域重组分割，生成 t 个子决策表，属性集 $P\subseteq C$，$\forall r\in P$，则有：$H_S(D|P)\leqslant H_S(D|P-\{r\})$

证明：由于合并划分块，决策表的条件熵呈递增性[2]，由此可知：$H_{S_i}(D|P)\leqslant H_{S_i}(D|P-\{r\})$，其中 $1\leqslant i\leqslant t$。

故由定义 15 可知，$H_S(D|P)\leqslant H_S(D|P-\{r\})$。证毕。

引理 1 说明随着条件属性的减少，基于正域重组分割的决策表条件熵单调上升。

根据上一节决策表新的条件熵定义，可以得到以下几个判定定理。

定理 1（属性必要性） 设 $S=\langle U,R,V,f\rangle$ 是一个决策表，U 是论域，$R=C\cup D$，C 是条件属性集合，D 是决策属性集合，条件属性集合 $P\subseteq C$，对于 P 中任意属性 r 相对于决策 D 是不必要的，其充分必要条件是：对决策表 S 进行基于 D 的 P 正域重组分割，$H_S(D|P)=H_S(D|P-\{r\})$。

在证明定理 1 之前，首先引入文[4]中的一个定理：

定理 2[4] 如果 $H(D|A\cup\{a\})=H(D|A)$，则 $POS_{A\cup\{a\}}(F)=POS_A(F)$。

下面来证明定理 1。

证明：对于相容决策表，$H_S(D|P)=H(D|P)$，文[2]已经证明。下面仅对不相容决策进行证明。

令 $IPOS_P(D)\mid IND(P)=\{X'_1,X'_2,\cdots,X'_s\}$，$INEG_P(D)\mid IND(P)=\{X''_1,X''_2,\cdots,X''_t\}$，对 S 进行基于 D 的 P 正域重组分割，得到 t 个子系统 $S_i=\langle U_i,R,V,f\rangle(1\leqslant i\leqslant t)$，其中，$U_i=X''_i\cup IPOS_P(D)$。

（1. 充分性）$H_S(D|P)=H_S(D|P-\{r\})$

由引理 1 条件熵的递增性可知：

$H_S(D|P)=H_S(D|P-\{r\})\Leftrightarrow H_{S_i}(D|P)=H_{S_i}(D|P-\{r\})$，其中 $1\leqslant i\leqslant t$。

由定理 2，显然 $POS_{P-\{r\}}(D)=POS_P(D)$，从而在代数观下属性 r 相对于决策 D 是不必要的。

（2. 必要性）r 相对于决策 D 是不必要的，在代数观的定义下，即：$POS_{P-\{r\}}(D)=POS_P(D)$。

对任意 S_i，由命题 4 可知，

$$H_{S_i}(D|P)=p(X''_i)DH_i(D,\{X''_i\}) \tag{1}$$

设 $IPOS_P(D)\mid IND(P-\{r\})=\{Y_1,Y_2,\cdots,Y_{s'}\}$，易知 $X''_i/(P-\{r\})=X''_i/P=X''_i$。

故 $H_{S_i}(D|P-\{r\})=p(X''_i)DH_i(D,\{X''_i\})+\sum_{j=1}^{s'}p(Y_j)DH_i(D,\{Y_j\})$

因为 $POS_{P-\{r\}}(D)=POS_P(D)$，故 $IPOS_{P-\{r\}}(D)=IPOS_P(D)$。由正域的信息观定义可知，$\sum_{j=1}^{s'}p(Y_j)DH_i(D,\{Y_j\})=0$，故

$$H_{S_i}(D|P-\{r\})=H_{S_i}(D|P)=p(X''_i)DH_i(D,\{X''_i\}) \tag{2}$$

由(1)、(2)式$\Rightarrow H_S(D|P)=H_S(D|P-\{r\})$。

综合(1)，(2)可知，P 中任意属性 r 相对于决策 D 是不必要的，其充分必要条件是 $H_S(D|P)=H_S(D|P-\{r\})$。证毕。

定理 3（独立性） 对决策表 S 进行基于 D 的 P 正域重组分割，其中 $P\subseteq C$，是条件属性集合，P 是相对决策 D 独立的，充分必要条件是：$\forall r\in P$，$H_S(D|P)\neq H_S(D|P-\{r\})$ 成立。

证明：对相容决策表，$H_S(D|P)=H(D|P)$，$H_S(D|P-\{r\})=H(D|P-\{r\})$，文[2]已经证明。

对不相容决策表，根据定理 1 的结论可知，上述对一般决策系统的属性相对独立性的信息观描述与代数观的描述是一致的。

定理 4（约简） 对决策表 S 进行基于 D 的 P 正域重组分割，其中 $P\subseteq C$，是条件属性集合，则 $Q\subseteq P$ 是 P 相对于决策 D 约简的充分必要条件是：

（1）$H_S(D|Q)=H_S(D|P)$，

（2）Q 相对决策 D 独立。

证明：对相容决策表，$H_S(D|Q)=H(D|Q)$，$H_S(D|P)=H(D|P)$，文[2]已经证明。

对不相容决策表，由定理1可知，$H_S(D|Q)=H_S(D|P)\Leftrightarrow POS_Q(D)=POS_P(D)$。又根据定理3可知，代数观与信息观对属性集的相对独立性的描述是一致的，因而对上述一般决策表的属性约简的信息观描述与代数观的描述是一致的。

定理5(信息观、代数观属性重要性关系) 设$S=\langle U,R,V,f\rangle$是一个决策表系统，U是论域，$R=C\cup D$，C是条件属性集合，D是决策属性集合，F是属性D导出的分类，对于任意属性$a\in C$，$r_C(F)=r_{C-\{a\}}(F)$的充分必要条件是：$H_S(D|C-\{a\})=H_S(D|C)$。

证明：对相容决策表，文[4]已经证明。

对不相容决策表，

$$r_C(F)-r_{C-\{a\}}(F)=\frac{|POS_C(D)|-|POS_{C-\{a\}}(D)|}{|U|} \quad (3)$$

由定理1可知，

$$H_S(D|C-\{a\})=H_s(D|C)\Leftrightarrow POS_C(D)=POS_{C-\{a\}}(D) \quad (4)$$

由(3)、(4)式可知，$H_s(D|C-\{a\})=H_S(D|C)\Leftrightarrow r_C(F)=r_{C-\{a\}}(F)$。证毕。

定理5说明，一个条件属性在代数观定义下的重要性为0，当且仅当其在新信息观下定义的重要性为0。

例2 计算图1的决策表S条件属性c相对集合$\{a,b\}$的属性重要性，其中条件属性$C=\{a,b,c\}$，决策属性$D=\{d\}$。

文[1]中已经计算得到属性c在原始信息观定义下的重要性$SGF(c,\{a,b\},\{d\})=\frac{1}{9}\log(\frac{823543}{800000})$，而代数观下为$SGF(c,\{a,b\},\{d\})=0$，因而得到二者不一致的结论。

下面，根据定义16新的信息熵下属性重要性的定义可以得到：

$H_{S_1}(D|C)=0.540852$，

$H_{S_2}(D|C)=0.550978$

$H_{S_1}(D|C-\{c\})=0.540852$，

$H_{S_2}(D|C-\{c\})=0.550978$

$\Rightarrow H_S(D|C)=H_{S_1}(D|C)+H_{S_2}(D|C)=1.09183$，

$\Rightarrow H_S(D|C-\{c\})=H_{S_1}(D|C-\{c\})+H_{S_2}(D|C-\{c\})=1.09183$

$\Rightarrow SGF_S(c,C-\{c\},D)=H_S(D|C-\{c\})-H_S(D|C)=0$，与代数观下的结果一致。

下面给出核属性的判定定理。

定理6(核属性) 对决策表S进行基于D的P正域重组分割，其中$P\subseteq C$，是条件属性集合，则$r\in P$是核属性的充分必要条件是：$H_S(D|P)\neq H_S(D|P-\{r\})$。

证明：根据定义1可知，$H_S(D|P)\neq H_S(D|P-\{r\})$当且仅当$r$是必要属性，根据命题1，$r$是核属性。证毕。

例3 求图3的决策表S的核属性，其中条件属性$C=\{C1,C2,C3\}$，决策属性$D=\{d\}$。

S

U	C1	C2	C3	d
x1	1	0	1	1
x2	1	0	1	0
x3	1	0	1	2
x4	0	0	1	1
x5	0	0	1	0
x6	1	1	1	1

图3 决策表S

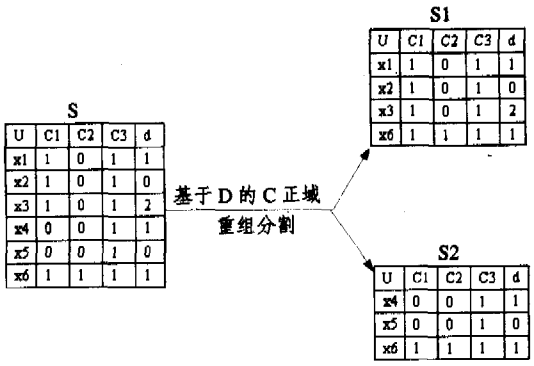

图4 求核属性例子

在文[2]中已经得到如下结论：在代数观下，核属性为$CORE_D(C)=\{C_2\}$；在原始信息观下，$CORE_D(C)=\{C_1,C_2\}$，因而得到二者不一致的结论。

用定理6，首先对决策表S进行基于D的C正域重组分割，得到S_1，S_2两个分块：

$H_{S_1}(D|C)=1.188722$

$H_{S_2}(D|C)=0.666667$

$H_{S_1}(D|C-\{c1\})=1.188722$

$H_{S_2}(D|C-\{c1\})=0.66667$

$H_{S_1}(D|C-\{c2\})=1.5$

$H_{S_2}(D|C-\{c2\})=0.66667$

$H_{S_1}(D|C-\{c3\})=1.188722$

$H_{S_2}(D|C-\{c3\})=0.66667$

$\Rightarrow H_S(D|C)=H_S(D|C-\{c1\})=H_S(D|C-\{c3\})\neq H_s(D|C-\{c2\})$

因而$CORE_D(C)=\{C2\}$，与代数观方法一致。

结论 在不相容决策系统中，Rough集理论的信息观与代数观下的结果存在不一致，这种不一致产生的根本原因是不相容信息对条件熵的计算产生了影响。本文给出了一种对决策表重组分割的方法，由于这种分割方法不会丢失信息，并保留了原决策表的所有信息，故在此种分割基础上提出的信息观的新定义与代数观是一致的，是对代数观的信息观下的解释，文中给出了相关证明。基于新的信息观的定义，可以相应提出属性核、属性重要性、属性约简等算法，对代数观算法进行补充，这将是我们下一步的工作。

参考文献

1 Pawlak Z. Rough Sets [J]. International Journal of Computer and Information Sciences, 1982,11(5): 341～356

2 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简. 计算机学报，2002,25(7):759～766

3 王国胤. 决策表核属性的计算方法. 计算机学报，2003，26(5):611～615

4 王国胤. Rough集理论代数与信息论观点的关系研究. 世界科技研究与发展,2002,24(5):20～26

5 王国胤. Rough集理论与知识获取. 西安：西安交通大学出版社，2001

6 叶东毅，陈昭炯. 一个新的差别矩阵及其求核方法. 电子学报，2002，30(7)：1086～1088

7 苗夺谦，王珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论. 模式识别与人工智能，1998,11(1):34～40

8 苗夺谦，王珏. 粗糙集理论中概念与运算的信息表示. 软件学报，1999,10(2):113～116