差别矩阵属性约简的信息观解释*)

徐章艳[1,2] 杨炳儒[1] 宋 威[1] 侯 伟[1]
(北京科技大学信息工程学院 北京 100083)[1] (广西师范大学计算机系 桂林 541004)[2]

摘 要 常见的属性约简定义有三种,即基于代数观的属性约简,基于信息观的属性约简和基于 HU 差别矩阵的属性约简。已有文献证明这三种属性约简彼此之间不等价。王国胤教授定义了一种新的决策表信息熵计算方法,在此方法基础上给出了粗糙集理论代数观的一种新的信息观解释。最近有学者提出了一种基于新信息熵的属性约简。经深入研究,我们证明了该属性约简与基于 HU 差别矩阵的属性约简是等价的,从而给出了基于 HU 差别矩阵的属性约简的信息观解释。

关键词 粗糙集,正区域,差别矩阵,信息熵,新信息熵

Illustrating the Attribute Reduction Based on HU's Discernibility Matrix with Information View

XU Zhang-Yan[1,2] YANG Bing-Ru[1] SONG Wei[1] HOU Wei[1]
(School of Information Engineering, Science and Technology University of Beijing, Beijng 100083)[1]
(Department of Computer, Guangxi Normal University, Guilin 541004)[2]

Abstract The attribute reduction definitions based on algebra view, based on information view and based on HU's discernibility matrix are familiar in rough set theory. It is proved that these three definitions of attribute reduction are not equivalent to each other. Professor Wang defined a new entropy method for decision table. And based on this method, a new information view that can comprehensively illustrate the algebra view is introduced. Recently, one attribute reduction definition based on the new information entropy is proposed by some researchers. In this paper, it is proved that this new attribute reduction definition based on the new information entropy is equivalent to that based on HU's discernibility matrix. And it is also given the illustration of the attribute reduction based on HU's discernibility matrix with information view.

Keywords Rough set, Positive region, Discernibility matrix, Information entropy, New information entropy

1 引言

在粗糙集理论[1]中,属性约简是重要研究内容之一。一些学者根据不同的应用目的,提出了不同的属性约简定义。其中最常见的有 Pawlak 教授提出的基于代数观的属性约简[1];胡小华教授提出的基于 HU 差别矩阵的属性约简[2];王国胤教授提出的基于信息观的属性约简[3,4]。文[4,5]研究了基于代数观的属性约简与基于信息观的属性约简之间的关系,得到如下的结论:在协调的决策表中,二者是等价的;但在不协调的决策表中,二者不等价。文[6~8]研究了基于代数观的属性约简和基于 HU 差别矩阵的属性约简的关系,得到如下的结论:在协调决策表中,二者是等价的;但在不协调的决策表中二者是不等价的,文[5]指出基于 HU 差别矩阵的属性约简与基于信息熵的属性约简在不协调决策表中也是不等价的。故在不协调决策表中,三者彼此不等价。

王国胤教授[9]定义了一种新的决策表信息熵计算方法,在此方法的基础上,给出了粗糙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。最近,梁吉业教授、罗平教授分别在文[10]和[11]中提出了一种基于新的信息熵的属性约简定义。经深入研究,我们证明该属性约简与基于 HU 差别矩阵的属性约简是等价的。因此,我们也找到了基于 HU 差别矩阵的属性约简的信息观解释。

2 基本知识

定义 1[1] 一个决策表定义为:$S=(U,C,D,V,f,d)$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 是论域;$C=\{c_1,c_2,\cdots,c_r\}$ 为条件属性集;D 为决策属性集;$f:U\times C\to V$ 和 $d:U\times D\to V$ 是信息函数,其中 $F=C\cup D$,$C\cap D=\varnothing$,$V=\cup V_a$,$a\in F$,V_a 表示属性 a 的值域。

定义 2[2] 设在决策表 $S=(U,C,D,V,f,d)$ 中,定义 HU 差别矩阵为 $M=(m_{ij})$,其元素定义如下:

$$m_{ij}=\begin{cases}\{c_k \mid c_k\in C, f(x_i,c_k)\neq f(x_j,c_k), d(x_i,D)\neq d(x_j,D)\}\\ \varnothing \quad 否则\end{cases}$$

其中 $k=1,2,\cdots,r;i,j=1,2,\cdots,n$。

定义 3[2] 在决策表 $S=(U,C,D,V,f,d)$ 中,设 $M=(m_{ij})$ 是 HU 差别矩阵,$\forall B\subseteq C$,若 B 满足:(1) $\forall \varnothing\neq m_{ij}\in M$,有 $B\cap m_{ij}\neq\varnothing$;(2) $\forall b\in B$,$B-\{b\}$ 不满足(1),则称 B 是 C 相对于 D 的基于 HU 差别矩阵的属性约简。记其所有的属性约简为 $HRED_C(D)$。

定义 4[10] 在决策表 $S=(U,C,D,V,f,d)$ 中,可以认为

*)基金项目:国家自然科学基金资助项目(No. 60675030),国家科技成果重点推广项目计划(2003EC000001)资助。**徐章艳** 博士研究生,研究方向为粗糙集理论及其应用与数据挖掘;**杨炳儒** 教授,博士生导师,研究方向为人工智能,数据挖掘和柔性建模;**宋 威** 博士研究生,研究方向为粗糙集理论及其应用与数据挖掘;**侯 伟** 博士研究生,研究方向为粗糙集理论及其应用与数据挖掘。

U 上任一属性集合 $B\subseteq C\cup D$(知识,等价关系族,$U/B=\{B_1,B_2,\cdots,B_t\}$)是定义在 U 上的子集组成的代数上的一个随机变量,其概率分布可通过如下方法来确定:

$$[B:p]=\begin{bmatrix} B_1 & B_2 & \cdots & B_t \\ p(B_1) & p(B_2) & \cdots & p(B_t)\end{bmatrix}$$

其中 $p(B_j)=|B_j|/|U|,j=1,2,\cdots,t$。

定义 5[10,11] 在决策表 $S=(U,C,D,V,f,d)$ 中,决策属性集 $D(U/D=\{D_1,D_2,\cdots,D_k\})$ 相对于条件属性集 $C(U/C=\{C_1,C_2,\cdots,C_m\})$ 的新的条件熵 $I'(D|C)$ 定义为:

$$I(D|C)=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{|D_j\cap C_i|}{|U|}\frac{|D_j^c\cap C_i|}{|U|}$$

其中 D_j^c 表示 D_j 的补集。

定义 6[10,11] 在决策表 $S=(U,C,D,V,f,d)$ 中,对 $\forall b\in B\subseteq C$,若 $I(D|B)=I(D|(B-\{b\}))$,则称 b 为 B 中相对于 D 是可省的(不必要的);否则 b 为称 B 中相对于 D 是不可省的(必要的)。对 $\forall B\subseteq C$,若其任一元素相对于 D 都是必要的,则称 B 相对于 D 是独立的。

定义 7[10,11] 在决策表 $S=(U,C,D,V,f,d)$ 中,若 $\forall B\subseteq C,I(D|B)=I(D|C)$ 且 B 相对于 D 是独立的,则称 B 是 C 相对于 D 的基于新的信息熵的属性约简。记其所有的属性约简为 $NewRED_C(D)$。

定义 8 在决策表 $S=(U,C,D,V,f,d)$ 中,$\forall p,Q\subseteq(C\cup D)$,记 $U/P=\{P_1,P_2,\cdots,P_h\},U/Q=\{Q_1,Q_2,\cdots,Q_s\}$,若 $\forall P_i\in U/P\Rightarrow\exists Q_j\in U/Q$ 使 $P_i\subseteq Q_j$,则称 U/P 为 U/Q 的加细,记为 $U/P\leqslant U/Q$。

3 基于新的信息熵与 HU 差别矩阵的属性约简的等价性

为证明两种属性约简的等价性,我们先引入如下的引理。

引理 1[11] 在决策表 $S=(U,C,D,V,f,d)$ 中,设 $U/B=\{B_1,B_2,\cdots,B_t\},U/D=\{D_1,D_2,\cdots,D_k\},U/B'=\{B_1,B_2,\cdots,B_{p-1},B_{p+1},\cdots,B_{q-1},B_{q+1},\cdots,B_t,B_p\cup B_q\}$,则有:

$I(D|B')\geqslant I(D|B)$

等号成立的充分必要条件是:$|D_j\cap B_p||D_j^c\cap B_q|=0\wedge|D_j\cap B_q||D_j^c\cap B_p|=0(j=1,2,\cdots,k)$。

定理 1 在决策表 $S=(U,C,D,V,f,d)$ 中,$\forall Q\subseteq P\subseteq(C\cup D)$,则有 $U/P\leqslant U/Q$。

证明:记 $U/P=\{P_1,P_2,\cdots,P_h\},U/Q=\{Q_1,Q_2,\cdots,Q_s\}$,任取 $P_i=[x]_P\in U/P$,由于 $Q\subseteq P$,则 $P_i=[x]_P=\{y|\forall a\in P,f(y,a)=f(x,a)\}\subseteq Q_j=[x]_Q=\{y|\forall a\in Q,f(y,a)=f(x,a)\}$。由 P_i 的任意性知:$U/P\leqslant U/Q$。

定理 2 在决策表 $S=(U,C,D,V,f,d)$ 中,$\forall A_1\subseteq A_2\subseteq C$,则有 $I(D|A_1)\geqslant I(D|A_2)$。等号成立的充要条件为:对任意 $X_i,X_j\in U/A_2,X_i\neq X_j$,若 $(X_i\cup X_j)\subseteq Y\in U/A_1$ 均有 $|D_r\cap X_i||D_r^c\cap X_j|=0$ 且 $|D_r\cap X_j||D_r^c\cap X_i|=0$。其中 $D_r\in U/D,r=1,2,\cdots,k$。

证明:由定理 1 和引理 1 即得。

定理 3 在决策表 $S=(U,C,D,V,f,d)$ 中,$\forall B\subseteq C$,若 $\forall\varnothing\neq m_{ij}\in M$ 使得 $B\cap m_{ij}\neq\varnothing$,则有 $I(D|B)=I(D|C)$。其中 M 是决策表的 H 差别矩阵。

证明:由于 $B\subseteq C$,故由定理 2 知 $I(D|B)\geqslant I(D|C)$。假设 $I(D|B)\neq I(D|C)$,则有 $C-B\neq\varnothing$;同时由定理 2 知,至少存在 $X_{i_0},X_{j_0}\in U/C,X_{i_0}\neq X_{j_0}$,且 $(X_{i_0}\cup X_{j_0})\subseteq Y_0\in U/B$,使得 $|D_r\cap X_{i_0}||D_r^c\cap X_{j_0}|=0$ 和 $|D_r\cap X_{j_0}||D_r^c\cap X_{j0}|=0(D_r\in U/D,r=1,2,\cdots,k)$ 至少有一个不成立。

(1) 当 $|D_r\cap X_{i_0}||D_r^c\cap X_{j_0}|\neq0$ 时,其中 $r=1,2,\cdots,k$,则至少存在 $D_p\in U/D$,使得 $|D_p\cap X_{i_0}||D_p^c\cap X_{j_0}|\neq0$,即一定存在 $x_{k_0}\in X_{i_0},x_{k_1}\in X_{j_0}$,使得 $d(x_{k_0}D)\neq d(x_{k_1},D)$(若不然,则有 $\forall x\in X_{i_0}y\in X_{j_0}$,均有 $d(x,D)=d(y,D)$,故 $(X_{i_0}\cup X_{j_0})\subseteq D_p\in U/D$,故 $|D_p\cap X_{i_0}||D_p^c\cap X_{j_0}|=0$,这不可能)。另一方面,由于 $(X_{i_0}\cup X_{j_0})\subseteq Y_0\in U/B$,故 $\forall b\in B$,均有 $f(x_{k_0},b)=f(x_{k_1},b)$;同时 $X_{i_0}\neq X_{j_0}$,故至少存在一个属性 $a\in C-B\neq\varnothing$,使得 $f(x_{k_0},a)\neq f(x_{k_1},a)$;另一方面由于 $d(x_{k_0},D)\neq d(x_{k_1},D)$,故有 $m_{k_0k_1}\neq\varnothing\wedge m_{k_0k_1}\cap B=\varnothing$,这与条件矛盾,故假设不成立,从而有 $I(D|B)=I(D|C)$。

(2) 当 $|D_r\cap X_{j_0}||D_r^c\cap X_{i_0}|\neq0$ 时,其中 $r=1,2,\cdots,k$,同理可证 $I(D|B)=I(D|C)$。

综上所述,若 $\forall\varnothing\neq m_{ij}\in M$ 使得 $B\cap m_{ij}\neq\varnothing$,则有 $I(D|B)=I(D|C)$。

定理 4 在 $S=(U,C,D,V,f,d)$,$\forall B\subseteq C$,$\forall b\in B$,如果 $\exists\varnothing\neq m_{i_0j_0}\in M$ 使得 $B-\{b\}\cap m_{i_0j_0}=\varnothing$,则 $H_{liang}(D|(B-\{b\}))>H_{liang}(D|C)$,其中 M 是决策表的 H 差别矩阵。

证明:由于 $B-\{b\}\subset C$,同定理 2 知有 $H_{liang}(D|(B-\{b\}))\geqslant H_{liang}(D|C)$。

另一方面,对于 $\forall m_{i_0j_0}\neq\varnothing$ 和 $\forall b\in B$,一定存在 $x_{i_0}\in[x_{i_0}]_C,x_{j_0}\in[x_{j_0}]_C([x_{i_0}]_C\neq[x_{j_0}]_C)$ 和属性 $a\in C$ 使得 $f(x_{i_0},a)\neq f(x_{j_0},a)$ 和 $d(x_{i_0},D)\neq d(x_{i_0},D)$ 成立。由于 $m_{i_0,j_0}\cap(B-\{b\})=\varnothing$,故对 $\forall q\in B-\{b\}$,则有 $f(x_{i_0},q)=f(x_{j_0},q)$。因此一定存在 $Y_0\in U/(B-\{b\})$ 使得 $([x_{i_0}]_C\cup[x_{j_0}]_C)\subseteq Y_0$ 成立。现在分析 $[x_{i_0}]_C,[x_{j_0},D]$ 与 U/D 的关系。令 $D_p=\{x|x\in U\wedge d(x,D)=d(x_{i_0},D)\},D_q=\{x|x\in U\wedge d(x,D)=d(x_{j_0},D)\}$,其中 $D_p,D_q\in U/D$。由于 $d(x_{i_0},D)\neq d(x_{j_0},D)$,故 $p\neq q$。

(1) 当 $[x_{i_0}]_C\subseteq D_p,[x_{j_0}]_C\subseteq D_q$,则有 $|D_p\cap[x_{i_0}]_C||D_p^c\cap[x_{j_0}]_C|=[x_{i_0}]_C||[x_{j_0}]_C|\neq0$。

(2) 当 $[x_{i_0}]_C\subseteq D_p$ 和 $[x_{j_0}]_C\not\subset D_q$,则有 $D_q\cap[x_{j_0}]_C\neq\varnothing$。故有

$|D_q\cap[x_{i_0}]_C||D_p^c\cap[x_{j_0}]_C|\geqslant|[x_{i_0}]_C||D_p^c\cap[x_{j_0}]_C|>0$。

(3) 当 $[x_{j_0}]_C\not\subset D_q$ 和 $[x_{i_0}]_C\not\subset D_p$,则有 $D_p\cap[x_{i_0}]_C\neq\varnothing$。故有

$|D_p\cap[x_{i_0}]_C||D_p^c\cap[x_{i_0}]_C|\geqslant|D_p\cap[x_{i_0}]_C||[x_{j_0}]_C|>0$。

(4) 当 $[x_{j_0}]_C\not\subset D_q$ 和 $[x_{i_0}]_C\not\subset D_p$,则有 $D_p\cap[x_{i_0}]_C\neq\varnothing$ 和 $D_q\cap[x_{j_0}]_C\neq\varnothing$。故有

$|D_p\cap[x_{i_0}]_C||D_p^c\cap[x_{j_0}]_C|\geqslant|D_p\cap[x_{i_0}]_C||D_q\cap[x_{j_0}]_C|>0$。

综上所述,对于 $([x_{i_0}]_C\cup[x_{j_0}]_C)\subseteq Y_0\in U/(b-\{b\})$,一定存在 $D_p,D_q\in U/P$ 使得 $|D_p\cap[x_{i_0}]_C||D_p^c\cap[x_{j_0}]_C|\neq0$。故由定理 2 知有 $H_{liang}(D|(B-\{b\}))\neq H_{liang}(D|C)$。即 $H_{liang}(D|(B-\{b\}))>H_{liang}(D|C)$。

定理 5 在决策表 $S=(U,C,D,V,f,d)$ 中,$\forall B\in HRED_C(D)$,则有 $B\in NewRED_C(D)$。

证明:由定义 3 和定义 7 以及定理 3 和定理 4 即得。

定理 6 在决策表 $S=(U,C,D,V,f,d)$ 中,$B\subseteq C$,若 $I(D|B)=I(D|C)$,则有 $\forall\varnothing\neq m_{ij}\in M$ 使得 $B\cap m_{ij}\neq\varnothing$。

证明:假设存在 $m_{i_0j_0}\neq\varnothing$,使得 $m_{i_0j_0}\cap B=\varnothing$。由于

$m_{i_0j_0}\neq\varnothing$，故有 $d(x_{i_0},D):d(x_{j_0},D)$ 且至少存在一属性 $a\in C-B$，使得 $f(x_{i_0},a)\neq f(x_{i_0},a)\neq f(x_{j_0},a)$，故有 $[x_{i_0}]_C\neq[x_{j_0}]_C$。另一方面有 $[x_{i_0}]_B=[x_{j_0}]_B$（这是因为 $m_{i_0j_0}\cap B=\varnothing$），即 $[x_{i_0}]_C\cup[x_{j_0}]_C\subseteq Y_0\in U/B$。又由于 $d(x_{i_0},D)\neq d(x_{j_0},D)$，故一定存在 $D_p\cap[x_{i_0}]_C\neq\varnothing$，$D_q\cap[x_{j_0}]_C\neq\varnothing$，$D_pD_q\in U/D$，$p\neq q$，故有 $|D_p\cap[x_{i_0}]_C||D_p^C\cap[x_{j_0}]_C|\neq 0$，由定理 2 知有 $I(D|B)>I(D|C)$，这与条件矛盾，故假设不成立，从而命题成立。

定理 7 在决策表 $S=(U,C,D,V,f,d)$ 中 $B\subseteq C$，$\forall b\in B$，若 $I(D|(B-\{b\}))>I(D|C)$，则一定存在 $\varnothing\neq m_{i_0j_0}\in M$，使得 $(B-\{b\})\cap m_{i_0j_0}=\varnothing$。

证明：假设结论不成立，即对 $\varnothing\neq m_{ij}\in M$，使得 $(B-\{b\})\cap m_{ij}\neq\varnothing$。由定理 3 的证明过程可知有 $I(D|(B-\{b\}))=I(D|C)$，这与条件矛盾，故假设不成立，从而命中成立。

定理 8 在决策表 $S=(U,C,D,V,f,d)$ 中，$\forall B\in NewRED_C(D)$，则有 $B\in HRED_C(D)$。

证明：由定理 6 和定理 7 即得。

定理 9 在决策表 $S=(U,C,D,V,f,d)$ 中，基于新的信息熵的属性约简与基于 HU 差别矩阵的属性约简是等价的。

证明：由定理 5 和定理 8 即得。

结论 王国胤教授定义了一种新的决策表信息熵计算方法，在此方法的基础上，给出了粗糙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。在本文中，我们证明了梁吉业教授和罗平教授提出的一种基于新的信息熵的属性约简与基于 HU 差别矩阵的属性约简是等价的。从而找到了基于 HU 差别矩阵的属性约简的信息观解释。这一研究思路揭示了属性约简中代数方法与信息熵方法的联系，在结合代数方法与信息熵方法优点的基础上，设计高效属性约简算法是我们进一步的研究方向。

参考文献

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

2 Hu Xiao Hua, Cercone N. Learning in relational databases: a rough set approach[J]. Computational Intelligence,1995,11(2): 323-337

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

4 Wang GuoYin. Rough reduction in algebra view and information view[J]. International Journal of Intelligent System, 2003(18): 679～688

5 王国胤. 决策表核的计算方法[J]. 计算机学报，2003，26(5)：611～615

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

7 杨明，孙志挥. 改进的差别矩阵及其求核方法[J]. 复旦学报(自然科学版)，2004，43(5)：865～868

8 胡健，徐章艳，杨炳儒. 关于"一种信息系统求核的新方法"的注记[J]. 计算机工程与应用，200，41(27)：51～52

9 龚勋，王国胤. 对代数观 Rough 集理论的信息观解释[J]. 计算机科学，2006，33(4)：151～154

10 Liang Jiye, Chin K S, Dang Chuangyin, Yam R C M. A new method for measuring uncertainty and fuzziness in rough set theory[J]. International Journal of General Systems, 2002, 31(4): 331～4342

11 Luo Ping, He Qing, Shi Zhongzhi. Theoretical study on a new information entropy and its use in attribute reduction [C]. ICCI, 2005. 73～79

（上接第 185 页）

增加也就使预测的准确度降低了。随着用户浏览序列的增加，分类和用户类别判定所出现的错误率就会越来越小，预测的准确率将会越来越高，这也就是当浏览序列大于 9 之后分类静态 Bayes 网预测模型的预测准确率一直高于其他模型的原因所在。同样随着用户浏览序列的增加，用户的浏览状态就越能体现用户浏览的特征，所以当用户浏览序列增加时三种模型的预测准确度都在上升。

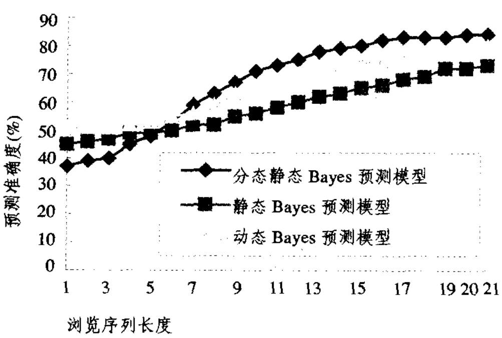

图 1　三种预测模型预测准确度的比较

总结 本文针对 Web 上用户的浏览序列这一时间序列问题所提出的三种预测模型。实验表明以上三种模型与传统的预测模型相比有以下的优点：

①三种模型都与 Bayes 网络相结合。充分利用了传统时间序列预测模型和 Bayes 网络的优点。时间序列预测问题的不确定性本质，表明该问题本身就是一个典型的不确定性处理问题。因此利用 Bayes 网对其进行处理和研究能够充分地利用 Bayes 网络的理论提高预测的准确度。

②实际应用中合理的选用这三种模型，可以达到降低模型的存储空间，充分地利用实验数据的优势，从而提高预测精度和时间复杂度。

当然，本文所提出的预测模型还有很多未完善的地方需要进一步的研究。比如在对条件概率的计算和降低条件概率表的存储度以及模型中所提到的三个搜索算法等都需要做更进一步的改进。

参考文献

1 (US) Box G E P, (UK) Jenkins G M. Times Series Analysis: Forecasting and Control[M]. 2nd edn. San Francisco: Holden-Day, 1976

2 West M, Harrison P J. Bayesian Forecasting and Dynamic Models[M]. New York: Springer-Verlag, 1997

3 张孝令，刘福升，张承进，等．贝叶斯动态模型及其预测[M]. 济南：山东科学技术出版社，1992

4 Shephard N. Statistical aspects of ARCH and stochastic volatility. In Time Series Models(eds D. R. Cox et al.). London. Chapman and Hall, 1996. 1～87

5 Fusion P J. Propagation, and Structuring in Belief Networks. Artificial Intelligence, 1985, 29: 241～288

6 Dohmen, K. Improved Bonferroni Inequalities with Applications: Inequalities and Identities of Inclusion-Exclusion Type. Berlin, Germany: Springer-Verlag, 2003

7 边肇祺，张学工，等. 模式识别(第二版). 北京：清华大学出版社，2000