广西科学院学报
Journal of Guangxi Academy of Sciences

2009,25(4):241～244
Vol. 25,No. 4 November 2009

基于信息熵属性约简的矩阵方法*

Matrix Method for Attribute Reduction Based on Information Entropy

蒙 韧[1],徐章艳[2]

MENG Ren[1],XU Zhang-yan[2]

(1. 广西师范大学财务处,广西桂林 541004;2. 广西师范大学计算机系,广西桂林 541004)

(1. Finance Office,Guangxi Normal University,Guilin,Guangxi,541004,China;2. Department of Computer,Guangxi Normal University,Guilin,Guangxi,541004,China)

摘要:给出基于信息熵的属性约简的矩阵表示,由此设计一个新的基于信息熵的属性约简算法,并用实例验证算法的可行性.该算法比较直观,容易理解,而且所占用的辅助空间少.

关键词:粗糙集理论 属性约简 信息熵 矩阵

中图法分类号:TP18 **文献标识码**:A **文章编号**:1002-7378(2009)04-0241-04

Abstract: The matrix expression of attribute reduction based on information entropy is given. Then a new algorithm of attribute reduction based on information entropy is designed. And the feasibility of algorithm is verified with an example. The algorithm is relatively straightforward, easy to understand, and use less auxiliary space.

Key words: rough set theory, attribute reduction, information entropy, matrix

在粗糙集理论[1]中,属性约简是重要的研究内容之一.目前有许多属性约简定义,例如基于正区域的属性约简[1],基于 Skowron 差别矩阵的属性约简[2]和基于信息熵的属性约简[3].文献[4～6]的研究表明,上述 3 种属性约简是彼此不等价的.目前,多数学者用启发式方法和差别矩阵方法设计属性约简算法。文献[7,8]用矩阵方法给出一个基于信息系统的属性约简算法,文献[9～12]用布尔矩阵的方法设计不同的基于正区域的属性约简算法,文献[13,14]用矩阵的方法设计基于正区域的规则提取算法.很少有学者用矩阵的方法设计基于信息熵属性约简算法。为此,本文先给出一个新矩阵和基于该矩阵的属性约简定义,然后证明新的属性约简定义与基于信息熵的属性约简定义等价(即给出了基于信息熵属性约简的矩阵表示),最后,以矩阵为基础,定义了属性的重要性,并利用属性的重要性设计一个新的基于信息熵的属性约简算法.

收稿日期:2009-10-08

作者简介:蒙 韧(1973-),男,工程师,主要从事粗糙集理论及其应用与数据挖掘研究。

*广西教育厅科研基金项目(200807MS015),广西师范大学博士科研基金项目资助。

1 信息熵属性约简的矩阵表示

定义 1[1] 五元组 $S=(U,C,D,V,f)$ 是一个决策表,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 表示对象的非空有限集合,称为论域;C 表示条件属性的非空有限集;D 表示决策属性的非空有限集且 $C\cap D=\varnothing$;$V=\bigcup_{a\in C\cup D}V_a$,$V_a$ 是属性 a 的值域,$f:U\times C\cup D\rightarrow V$ 是一个信息函数,它对一个对象的每一个属性赋予一个信息值,即 $\forall a\in C\cup D,x\in U$,有 $f(x,a)\in V_a$;每一个属性子集 $P\subseteq(C\cup D)$ 决定了一个二元不可区分关系 $IND(P)$:

$$IND(P)=\{(x,y)\in U\times U|\forall a\in P,f(x,a)=f(y,a)\}.$$

关系 $IND(P)$ 构成 U 的一个划分,用 $U/IND(P)$ 表示,简记为 U/P,U/P 中的任何符合 $[x]_P=\{y|\forall a\in P,f(x,a)=f(y,a)\}$ 的元素称为等价类.

定义 2[3] 在决策表 $S=(U,C,D,V,f)$ 中,可以认为 U 上任一属性集合 $B\subseteq C\cup D$(知识,等价关系簇,$U/B=\{B_1,B_2,\cdots,B_t\}$)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3[3] 在决策表 $S=(U,C,D,V,f)$ 中，决策属性集 $D(U/D=\{D_1,D_2,\cdots,D_k\})$ 相对于条件属性集 $C(U/C=\{C_1,C_2,\cdots,C_m\})$ 的条件熵 $H(D|C)$ 定义为

$$H(D|C)=-\sum_{i=1}^{m}p(C_i)\sum_{j=1}^{k}p(D_j|C_i)\log p(D_j|C_i),$$

其中 $p(D_j|C_i)=|D_j\cap C_i|/|C_i|$.

定义4[3] 在决策表 $S=(U,C,D,V,f)$ 中，对 $\forall\ b\in B\subseteq C$，若 $H(D|B)=H(D|(B-\{b\}))$，则称 b 为 B 中相对于 D 是可省的(不必要的)；否则称 b 为 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5[3] 在决策表 $S=(U,C,D,V,f)$ 中，若 $\forall\ B\subseteq C,H(D|B)=H(D|C)$ 且 B 相对于 D 是独立的，则称 B 是 C 相对于 D 的一个基于信息熵的属性约简.

定义6 在决策表 $S=(U,C,D,V,f)$ 中，$\forall\ B\subseteq C,\forall\ x\in U$，定义

$$\mu_B(x)=\left(\frac{|D_1\cap[x]_B|}{|[x]_B|},\ \frac{|D_2\cap[x]_B|}{|[x]_B|},\cdots,\frac{|D_k\cap[x]_B|}{|[x]_B|}\right),$$

称 $\mu_B(x)$ 为 B 关于 D 的概率分配函数.

定义7 在决策表 $S=(U,C,D,V,f)$ 中，条件属性集 $B(B\subseteq C)$ 的信息熵矩阵 $M_B=(r_{ij}^B)_{n\times n}$ 的定义为

$$r_{ij}^B=\begin{cases}1 & \forall\ b\in B,f(x_i,b)=f(x_j,b),\mu_C(x_i)\neq\mu_C(x_j),\\ 0 & \exists\ b\in B,f(x_i,b)\neq f(x_j,b),\mu_C(x_i)\neq\mu_C(x_j),\\ 1 & \mu_C(x_i)=\mu_C(x_j).\end{cases}$$

定义8 设 $M_1=(r_{ij}^1)_{n\times n}$ 和 $M_2=(r_{ij}^2)_{n\times n}$ 是2个布尔矩阵，$M_1\leqslant M_2\Leftrightarrow\forall\ r_{ij}^1\in M_1,r_{ij}^2\in M_2\Rightarrow r_{ij}^1\leqslant r_{ij}^2$；$M_1=M_2\Leftrightarrow\forall\ r_{ij}^1\in M_1,r_{ij}^2\in M_2\Rightarrow r_{ij}^1=r_{ij}^2$.

定义9 设 $M_1=(r_{ij}^1)_{n\times n}$ 和 $M_2=(r_{ij}^2)_{n\times n}$ 是2个布尔矩阵，则 $M_3=(r_{ij}^3)_{n\times n}=M_1\cap M_2=(r_{ij}^1\wedge r_{ij}^2)_{n\times n}$；$M_4=(r_{ij}^4)_{n\times n}=M_1-M_2=(r_{ij}^1-r_{ij}^2)_{n\times n}$.

性质1 在决策表 $S=(U,C,D,V,f)$ 中，若条件属性集 $B(B\subseteq C)$ 的矩阵为 $M_B=(r_{ij}^B)_{n\times n}$，任意属性集 $\{b\}\subseteq B$ 的矩阵为 $M_{\{b\}}=(r_{ij}^{\{b\}})_{n\times n}$，则有 $M_B=\bigcap\limits_{b\in B}M_{\{b\}}$.

证明 由条件属性集的矩阵定义即可证明性质1.

性质2 在决策表 $S=(U,C,D,V,f)$ 中，条件属性集 $A\subseteq B\subseteq C$，则有 $M_A\geqslant M_B$.

定义10 在决策表 $S=(U,C,D,V,f)$ 中，对任意 $B(B\subseteq C)$，若 $M_B=M_C$ 且 $\forall\ b\in B$ 均有 $M_{B-\{b\}}\neq M_C$，则称 B 是 C 关于 D 的一个基于矩阵 M_C 的属性约简.

定义11 在决策表 $S=(U,C,D,V,f)$ 中，任意 $P,Q\subseteq(C\cup D)$，记 $U/P=\{P_1,P_2,\cdots,P_t\},U/Q=\{Q_1,Q_2,\cdots,Q_S\}$，若 $\forall\ Pi\in U/P\Rightarrow\exists\ Q_j\in U/Q$ 使 $Pi\subseteq Q_j$，则称 U/P 为 U/Q 的加细，记为 $U/P\leqslant U/Q$.

定理1[5] 在决策表 $S=(U,C,D,V,f)$ 中，$\forall\ Q\subseteq P\subseteq(C\cup D)$，则有 $U/P\leqslant U/Q$.

定理2[5] 在决策表 $S=(U,C,D,V,f)$ 中，$\forall\ B\subseteq C$，则 $H(D|B)=H(D|C)$ 的充分必要条件是 $\forall\ x\in U\Rightarrow\mu_B(x)=\mu_C(x)$.

定理3 在决策表 $S=(U,C,D,V,f)$ 中，对任意 $B(B\subseteq C)$，设条件属性集 B 的信息熵矩阵为 $M_B=(r_{ij}^B)_{n\times n}$，若 $H(D|B)=H(D|C)$，则有 $M_B=M_C=(r_{ij}^C)$.

证明 设 $H(D|B)=H(D|C)$，由定理2知，$\forall\ x\in U\Rightarrow\mu_B(x)=\mu_C(x)$．假设 $M_B\neq M_C$，则至少存在 $r_{i_0j_0}^B\neq r_{i_0j_0}^C$，故有 $\mu_C(x_{i_0})\neq\mu_C(x_{j_0})$(因为 $r_{i_0j_0}^B$，$r_{i_0j_0}^C$ 中有一个其值为0)．由于 $B\subseteq C$，由定理1知，$[x_{i_0}]_B\supseteq[x_{i_0}]_C,[x_{j_0}]_B\supseteq[x_{j_0}]_C$，从而有 $r_{i_0j_0}^B=1,r_{i_0j_0}^C=0$．由于 $r_{i_0j_0}^B=1$，故有 $[x_{i_0}]_B=[x_{j_0}]_B$，从而有 $\mu_B(x_{i_0})=\mu_B(x_{j_0})$．由于 $r_{i_0j_0}^C=0$，故有 $[x_{i_0}]_C\neq[x_{j_0}]_C$，从而有 $\mu_C(x_{i_0})\neq\mu_C(x_{j_0})$．又由于 $\forall\ x\in U\Rightarrow\mu_B(x)=\mu_C(x)$，故有 $\mu_B(x_{i_0})=\mu_C(x_{i_0})$ 和 $\mu_B(x_{j_0})=\mu_C(x_{j_0})$，而 $\mu_B(x_{i_0})=\mu_B(x_{j_0})$，故 $\mu_C(x_{i_0})=\mu_C(x_{j_0})$．这与 $\mu_C(x_{i_0})\neq\mu_C(x_{j_0})$ 矛盾．故假设不成立，从而定理3成立，即有 $M_B=M_C$.

定理4 在决策表 $S=(U,C,D,V,f)$ 中，对任意 $B(B\subseteq C)$，设条件属性集 B 的信息熵矩阵为 $M_B=(r_{ij}^B)_{n\times n}$，若 $M_B=M_C$，则有 $H(D|B)=H(D|C)$.

证明 当 $M_B=M_C$ 时，假设 $H(D|B)=H(D|C)$ 不成立，由定理2知，至少存在 $x_{i_0}\in U\Rightarrow\mu_B(x_{i_0})\neq\mu_C(x_{i_0})$．由分布函数的定义知 $[x_{i_0}]_B$

$\neq [x_{j_0}]_C$. 又由定理 1 知$[x_{i_0}]_B \supseteq [x_{j_0}]_C$,故一定存在$[x_{j_0}]_C \neq [x_{i_0}]_C$,使得$[x_{i_0}]_B = [x_{j_0}]_B \supseteq [x_{j_0}]_C$. 由于$[x_{j_0}]_C \neq [x_{i_0}]_C$,故有 $\mu_C(x_{i_0}) \neq \mu_C(x_{j_0})$. 由于$[x_{j_0}]_C = [x_{i_0}]_C$,故至少存在一属性$a \in C$,使得$f(x_{i_0},a) \neq f(x_{j_0},a)$,由定义 7 知 $r^C_{i_0j_0} = 0$. 再由于 $\mu_C(x_{i_0}) \neq \mu_C(x_{j_0})$ 和$[x_{i_0}]_B = [x_{j_0}]_B$,由定义 7 知 $r^B_{i_0j_0} = 1$,从而有 $r^B_{i_0j_0} \neq r^C_{i_0j_0}$. 这与$M_B = M_C$矛盾. 故假设不成立,从而 $H(D|B) = H(D|C)$.

定理 5　在决策表$S = (U,C,D,V,f)$中,对任意$B(B \subseteq C)$,设B是C关于D的一个基于信息熵的属性约简,则B也是C关于D的一个基于矩阵M_C的属性约简.

证明　设B是C关于D的一个基于信息熵的属性约简,则有(1)$H(D|B) = H(D|C)$;(2)$\forall\, b \in B, H(D|(B - \{b\})) \neq H(D|C)$. 当 $H(D|B) = H(D|C)$ 时,由定理 3 知 $M_B = M_C$. 当 $\forall\, b \in B$, $H(D|(B - \{b\})) \neq H(D|C)$,时,由定理 4 的逆否命题知 $M_{B-\{b\}} \neq M_C$. 由基于矩阵的属性约简定义,B 也是 C 关于 D 的一个基于矩阵 M_C 的属性约简.

定理 6　在决策表$S = (U,C,D,V,f)$中,对任意$B(B \subseteq C)$,设 B 是 C 关于 D 的一个基于矩阵 M_C 的属性约简,则 B 也是 C 关于 D 的一个基于信息熵的属性约简.

证明　设B是C关于D的一个基于矩阵M_C的属性约简,则有 $M_B = M_C$ 且 $\forall\, b \in B$ 均有 $M_{B-\{b\}} \neq M_C$. 当 $M_B = M_C$ 时,由定理 4 有 $H(D|B) = H(D|C)$. 当 $M_{B-\{b\}} \neq M_C$ 时,由定理 3 的逆否命题有 $H(D|(B - \{b\})) \neq H(D|C)$,即 $\forall\, b \in B$ 有 $H(D(B - \{b\})) \neq H(D|C)$. 由基于信息熵的属性约简定义知,$B$ 是 C 关于 D 的一个基于信息熵的属性约简.

定理 7　决策表$S = (U,C,D,V,f)$中,基于矩阵 M_C 的属性约简与基于信息熵的属性约简等价.

证明　由定理 4 和定理 5 即得定理 7.

2　基于矩阵的属性约简算法

定义 12　在决策表 $S = (U,C,D,V,f)$ 中,属性 $a \in C - B$ 相对于 $B(B \subseteq C)$ 的重要性定义为 $Sig_B(a) = |M_B - M_{B\cup\{a\}}|$,其中 $|M|$ 表示矩阵M中的非零元个数.

基于矩阵的属性约简算法:

输入:$S = (U,C,D,V,f)$, $U = \{x_1,x_2,\cdots,x_n\}$, $C = \{c_1,c_2,\cdots,c_r\}$;

输出:属性约简 R.

步骤 1　计算决策表的矩阵,记为 M_C;

步骤 2　令 $R = \varnothing$;

步骤 3　对任一 $a \in C - R$,计算 $Sig_R(a)$;

步骤 4　令 $S_R(r) = \max\limits_{a\in C-R} Sig_R(a)$;若 $Sig_R(r) = 0$,输出属性约简 R 并终止算法;否则令 $R = R \cup \{r\}$;转向步骤 3;

步骤 5　输出属性约简 R 并终止算法.

定理 8　算法输出是基于信息熵的属性约简.

证明　假设$M_R \neq M_C$,由性质 2 知,$M_R > M_C$,故至少存在$a \in C - R$使得$M_{R\cup\{a\}} = M_R \cap M_{\{a\}} < M_R$,否则$\forall\, b \in C - R$均有$M_{R\cup\{a\}} = M_R \cap M_{\{a\}} = M_R$,则 $M_C = M_{R\cup(C-R)} = M_R \cap (\bigcap\limits_{b\in C-R} M_{\{b\}}) = \bigcap\limits_{b\in C-R} (M_R \cap M_{\{b\}}) = \bigcap\limits_{b\in C-R} M_R = M_R$. 故有 $Sig_R(r) = \max\limits_{c\in C-R} Sig_R(c) \geqslant Sig_R(a) = |M_B - M_{B\cup\{a\}}| > 0$. 这与算法终止时 $Sig_R(r) = 0$ 矛盾. 故假设不成立,从而算法终止时有 $M_R = M_C$. 由定理 7 知,R 是决策表的一个基于信息熵的属性约简.

3　算法实例验证

以表 1 为例,说明基于矩阵的属性约简算法的可行性.

表 1　决策表

	a	b	c	d	D	$\mu_C(\cdot)$
$X1$	2	1	2	1	0	(1,0,0)
$X2$	2	2	2	1	1	(0,1,0)
$X3$	2	1	2	1	0	(1,0,0)
$X4$	2	2	2	1	1	(0,1,0)
$X5$	3	1	2	1	0	(1,0,0)
$X6$	1	2	3	2	2	(0,0,1)

根据算法的第 1 步,我们得到

$$M_C = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M_{\{a\}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M_{\{b\}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix},$$

$$M_{\{c\}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M_{\{d\}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

由算法的第3步可知(注意:$M_{\varnothing} = (1)_{6\times 6}$):$Sig_{\varnothing}(a) = 14$,$Sig_{\varnothing}(b) = 18$,$Sig_{\varnothing}(d) = 10$.再根据算法的第4步得:选择属性$b$,$R = \{b\}$,算法转入第3步.

$$M_{\{b,a\}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M_{\{b,c\}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M_{\{b,d\}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

又由算法的第3步知:$Sig_{\{b\}}(a) = 4$,$Sig_{\{b\}}(c) = 4$,$Sig_{\{b\}}(d) = 4$.再根据算法的第4步得:选择属性a,$R = \{b,a\}$,算法转入第3步.由算法的第3步知:$Sig_{\{b,c\}}(c) = Sig_{\{b,d\}}(d) = 0$,算法终止,输出的属性约简是$R = \{b,a\}$.

实例验证结果表明,算法是有效的,而且比较直观,容易理解,所占用的辅助空间比差别矩阵方法设计的算法少.但是我们在实例验算中发现了重复的计算,因此下一步工作要设计基于矩阵的属性约简的高效算法来避免重复计算.

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science,1982,11(5):341-356.

[2] Hu X H,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):759-766.

[4] 邓大勇,黄厚宽,李向军.不一致决策系统中属性约简的比较[J].电子学报,2007,35(2):252-255.

[5] 徐章艳,杨炳儒,宋威,等.几种不同属性约简的比较研究[J].小型微型计算机系统,2008,29(5):848-853.

[6] Wang G Y. Rough reduction in algebra view and information view[J]. International Journal of Intelligence Systems,2003,18(5):679-688.

[7] Guan J W,Bell D Z. Matrix computation for information systems[J]. Information Science,2001,131(1):129-156.

[8] 李金海,吕跃进.一种基于关系矩阵的信息系统属性约简算法[J].计算机工程与应用,2008,44(9):147-149.

[9] 杨勇,牛彩云,李廉.基于矩阵取小乘法的粗糙集[J].计算机工程与应用,2008,44(6):48-50.

[10] 雷晓蔚.粗集理论的矩阵方法[J].计算机工程与应用,2006,42(17):73-75.

[11] 闫德勤,王杨.基于关联矩阵的属性约简算法[J].计算机工程与应用,2005,41(20):181-183.

[12] 任艳玲,朱明放.基于粗糙集的属性约简的矩阵方法[J].陕西理工学院学报,2006,22(3):76-80.

[13] 耿志强,朱群雄.基于粗糙集信息等价矩阵的关联规则挖掘[J].计算机与应用化学,2007,24(12):1663-1667.

[14] 罗来鹏,刘二根,王广超.基于矩阵的最简决策规则获取[J].计算机工程,2008,34(19):41-43.

(责任编辑:尹 闯)