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

2009,25(2):89~91
Vol. 25,No. 2 May 2009

基于信息熵的信息系统及决策表的属性约简
Attribute Reduction Based on Information Entropy in Information System and Decision Table

王 静,屈玲玲,孙宗剑
WANG Jing,QU Ling-ling,SUN Zong-jian

(河北科技大学理学院,河北石家庄 050018)
(College of Sciences,Hebei University of Science and Technology,Shijiazhuang,Hebei,050018,China)

摘要:利用新的信息熵,给出信息系统信息熵的约简方法及决策表的相对信息熵约简判定定理,证明分布协调集一定是相对信息熵的协调集.

关键词:信息熵 属性约简 代数协调集
中图法分类号:O159,TP18 **文献标识码**:A **文章编号**:1002-7378(2009)02-0089-03

Abstract: The attribute reduction method of information entropy in information system is obtained from new information entropy and the determination theorem of reducing relation information entropy is in decision table. At the same time, the distribution coordination set is proved to be the coordination set of relative information entropy.

Key words: information entropy, attribute reduction, algebra coordination set

属性约简是信息系统知识发现的核心问题之一,也是粗糙集理论研究的重要课题.文献[1]研究多种属性约简之间的关系,并利用辨识矩阵和辨识函数计算出所有代数约简和分布约简的方法.对于信息系统,本文提出信息熵约简概念,并证明代数意义下的属性约简等价于信息熵意义下的属性约简;对于决策表,又提出相对信息熵约简概念,给出这种约简的判定定理,证明分布协调集一定是相对信息熵协调集.本文恒假定 U 上的概率分布为:$P(X)=\frac{|X|}{|U|}$,$\forall\ X\subseteq U$.

1 信息系统的属性约简

定义 1[2] 设(U,R)为 Pawlak 近似空间,P 为 U 上的概率.记 $U/R=\{C_1,C_2,\cdots,C_s\}$ 是由等价关系 R 生成的 U 上的划分,则 R 的信息熵定义为划分 U/R 的熵,即

$$E(R)=\sum_{i=1}^{s}P(C_i)(1-P(C_i)). \tag{1}$$

定义 2 设 $\pi_1=\{A_1,A_2,\cdots,A_n\}$ 和 $\pi_2=\{B_1,B_2,\cdots,B_m\}$ 是集合 U 的两个划分,则 π_1 在 π_2 条件下的条件熵定义为

$$E(\pi_1|\pi_2)=\sum_{j=1}^{m}P(B_j)\sum_{i=1}^{n}P(A_i|B_j)(1-P(A_i|B_j)).$$

设(U,A)是信息系统,$B\subseteq A$,记 $R_B=\{(x,y)\in U\times U:b(x)=b(y),\forall\ b\in B\}$,则 R_B 是 U 上的等价关系.于是 $U/B=\{[x]_B\}$ 是 U 的一个划分,其中 $[x]_B=\{y\in U:(x,y)\in R_B\}$ 是 x 所在的 R_B 的等价类.

定义 3 设(U,A)是信息系统,$B\subseteq A$,

(i)[3] 如果 $R_B=R_A$,则称 B 是(U,A)的代数协调集;若 B 是代数协调集,但 B 的任何真子集都不是代数协调集,则称 B 是(U,A)的代数约简;

(ii) 若 $E(R_B)=E(R_A)$,则称 B 是(U,A)的信息熵协调集;若 B 是信息熵协调集,但 B 的任何真子集都不是信息熵协调集,则称 B 是(U,A)的信息熵约简.其中 $E(R_A)$ 是由(1)式定义的信息熵.

收稿日期:2008-04-25
作者简介:王 静(1974-),女,讲师,主要从事人工智能的数学基础研究。

命题1 设 π_1,π_2,π_3 是集合 U 的3个划分，如果 $\pi_1 \prec \pi_2$，则 (1)$E(\pi_1) \leqslant E(\pi_2)$；(2)$E(\pi_1|\pi_3) \leqslant E(\pi_2|\pi_3)$；(3)$E(\pi_3|\pi_1) \leqslant E(\pi_3|\pi_2)$.

定理1 设(U,A)是信息系统，$B \subseteq A$. 则 B 是 (U,A) 的代数约简当且仅当 B 是(U,A)的信息熵约简.

证明 只需证明 B 是(U,A)的代数协调集当且仅当 B 是(U,A)的信息熵协调集.

设 B 是(U,A)的代数协调集，则 $R_B = R_A$，于是 $U/R_B = U/R_A$. 这样由定义1易知 $E(R_B) = E(R_A)$，即 B 是(U,A)的信息熵协调集.

若 B 是(U,A)的信息熵协调集，则 $E(R_B) = E(R_A)$. 记 $R/A = \{C_1,C_2,\cdots,C_m\}$，$R/B = \{E_1,E_2,\cdots,E_m\}$.

若 $R_A \neq R_B$，由于 $R_A \subseteq R_B$，故 $m > n$ 且存在$\{1,2,\cdots,m\}$的一个划分 $J = \{I(1),I(2),\cdots,I(n)\}$，使得 $E_i = \bigcup\limits_{j\in I(i)} C_j$. 由命题的证明知 $E(R_B) = 1 - \sum\limits_{i=1}^{n}(\sum\limits_{j\in I(i)}^{n} P(C_j))^2$. 又因为 $m > n$，存在 $I(i_0) \in J$ 使得 $|I(i_0)| > 1$. 于是 $(\sum\limits_{j\in I(i_0)} P(C_j))^2 > (\sum\limits_{j\in I(i_0)} P(C_j))^2$ 而 $(\sum\limits_{j\in I(i_0),i\neq i_0} P(C_j))^2 \geqslant (\sum\limits_{j\in I(i_0),i\neq i_0} P(C_j))^2$，因此 $E(R_B) = 1 - \sum\limits_{i=1}^{n}(\sum\limits_{j\in I(i)}^{n} P(C_j))^2 < 1 - \sum\limits_{i=1}^{n}\sum\limits_{j\in I(i)}^{n}(P(C_j))^2 = E(R_A)$. 产生矛盾. 故 $R_A = R_B$，即 B 是(U,A)的代数协调集.

2 决策表的属性约简

设(U,A)是信息系统，$B \subseteq A$，$\forall X \subseteq U$，X 关于(U,R)的一对下近似 $\underline{R}(X)$ 和上近似 $\overline{R}(X)$ 定义为[4]：$\underline{R}(X) = \{x \in U:[x]_R \subseteq X\} = \bigcup\{[x]_R:[x]_R \subseteq X\}$，$\overline{R}(X) = \{x \in U:[x]_R \cap X \neq \varnothing\} = \bigcup\{[x]_R:[x]_R \cap X \neq \varnothing\}$，$\underline{R}$ 和 $\overline{R}$ 称为 Pawlak 近似算子，也称为由(U,R)导出的近似算子. 有关近似算子的性质可参阅文献[5].

对于 $P,Q \subseteq A$，Q 的 P 正域定义为 $POS_P(Q) = \sum\limits_{X\in U/Q} \underline{P}(X)$. 设$(U,A \cup D)$是决策表，其中 A 是条件属性集，D 是决策属性集. 如果 $POS_A(D) = U$，则称$(U,A \cup D)$为协调的决策表，否则称为不协调.

对于决策表 $S = (U,A \cup D)$，$B \subseteq A$，如果 $POS_A(D) = POS_B(D)$，则称 B 是 S 的相对协调集，若 B 是相对协调集，但 B 的任何真子集都不是相对协调集，则称 B 是 S 的相对约简. 文献[6]给出利用辨识矩阵和辨识函数计算所有相对约简的方法.

记 $U/A = \{C_1,C_2,\cdots,C_m\}$，$U/B = \{E_1,E_2,\cdots,E_n\}$，$U/D = \{D_1,D_2,\cdots,D_r\}$，$\mu_B(x) = \{D(D_1/[x]_B),D(D_2/[x]_B),\cdots,D(D_r/[x]_B)\}$，其中 $D(D_i/[x]_B) = \dfrac{|D_i \cap [x]_B|}{|[x]_B|}$ 是 $[x]_B$ 在 D 中的包含度.

为简单起见，在定义4中，把 U/A 关于 U/A 的条件熵 $E(U/D|U/A)$ 简记为 $E(D|A)$.

定义4 设 $S = (U,A/D)$ 是决策表，$B \subseteq A$，

(1)[1] 如果 $\forall x \in U$，有 $\mu_A(x) = \mu_B(x)$，则称 B 是 S 的分布协调集；若 B 是分布协调集，但 B 的任何真子集都不是分布协调集，则称 B 是 S 的分布约简.

(2) 如果 $E(D|B) = E(D|A)$，则称 B 是 S 的相对信息熵协调集；若 B 是相对信息熵协调集，但 B 的任何真子集都不是相对信息熵协调集，则称 B 是 S 的相对信息熵约简.

定理2 设 $S = (U,A/D)$ 是决策表，$B \subseteq A$. 则 B 是 S 的相对信息熵协调集，当且仅当 $\forall\ 1 \leqslant l \leqslant r$，$\forall\ 1 \leqslant j \leqslant n$，$|D_l \cap E_j|(1 - \dfrac{|D_l \cap E_j|}{E_j}) = \sum\limits_{t=T_j} |D_l \cup C_t|(1 - \dfrac{|D_l \cup C_t|}{|C_t|})$，其中 $E_j = \bigcup\limits_{t\in T_j} C_t$，$1 \leqslant j \leqslant n$.

证明 “⇐”. 容易计算

$$E(D|A) = \sum_{i=1}^{m}\frac{|C_i|}{|U|}\sum_{l=1}^{r}\frac{|D_l \cap C_i|}{|C_i|}(1 - \frac{|D_l \cap C_i|}{|C_i|}) = \frac{1}{|U|}\sum_{i=1}^{m}\sum_{l=1}^{r}|D_l \cap C_i|(1 - \frac{|D_l \cap C_i|}{|C_i|}),$$

$$E(D|B) = \frac{1}{|U|}\sum_{j=1}^{n}\sum_{l=1}^{r}|D_l \cap E_j|(1 - \frac{|D_l \cap E_j|}{|E_j|}) = \frac{1}{|U|}\sum_{j=1}^{n}\sum_{l=1}^{r}\sum_{t\in T_j}|D_l \cap C_t|(1 - \frac{|D_l \cap C_t|}{|C_t|}) = \frac{1}{|U|}\sum_{l=1}^{r}\sum_{i=1}^{m}|D_l \cap C_i|(1 - \frac{|D_l \cap C_i|}{|C_i|}),$$

所以 $E(D|A) = E(D|B)$，即 B 是 S 的相对信息熵协调集.

“⇒”. 由于

$$|D_l \cap E_j|(1 - \frac{|D_l \cap E_j|}{|E_j|}) - \sum_{t\in T_j}|D_l \cap C_t|(1 - \frac{|D_l \cap C_t|}{|C_t|}) = \sum_{t\in T_j}|D_l \cap C_i|(1 - \frac{|D_l \cap E_j|}{|E_j|})(1 - \frac{|D_l \cap C_t|}{|C_t|}) \geqslant 0,$$

所以

$$|D_l \cap E_j|(1-\frac{|D_l \cap E_j|}{|E_j|}) \geqslant \sum_{t\in T_j}|D_l \cap C_t|(1-\frac{|D_l \cap C_t|}{|C_t|}).$$

若存在 $1\leqslant l\leqslant r$,或 $1\leqslant j\leqslant n$ 使得上述不等式的严格不等号成立,则

$$\sum_{l=1}^{r}\sum_{j=1}^{n}|D_l \cap E_j|(1-\frac{|D_l \cap E_j|}{|E_j|}) > \sum_{l=1}^{r}\sum_{j=1}^{n}\sum_{t\in T_j}|D_l \cap C_t|(1-\frac{|D_l \cap C_t|}{|C_t|}).$$

这样 $E(D|B)>E(D|A)$,与 B 是 S 的相对信息熵协调集矛盾.

定理 2 给出相对信息熵协调集判定定理,由此不难得到计算 S 的相对信息熵约简的方法.

定理 3　设 $S=(U,A/D)$ 是决策表,$B\subseteq A$. 若 B 是 S 的分布协调集,那么 B 也是 S 的相对信息熵协调集.

证明　$\forall x\in U$,记 $J_{x,B}=\{[y]_A:[y]_A\subseteq[x]_B\}$,则 $J_{x,B}$ 是 $[x]_B$ 的划分. 由于 B 是 S 的分布协调集,故 $\forall\ 1\leqslant l\leqslant r,\forall\ [y]_A\in J_{x,B}$,$\frac{|D_l \cap [y]_A|}{|[y]_A|}=\frac{|D_l \cap [y]_B|}{|[y]_B|}=\frac{|D_l \cap [x]_B|}{|[x]_B|}$,于是

$$|D_l \cap [x]_B|(1-\frac{|D_l \cap [x]_B|}{|[x]_B|})=\sum_{[y]_A\in J_{x,B}}|D_l \cap [y]_A|(1-\frac{|D_l \cap [x]_B|}{|[x]_B|})=\sum_{[y]_A\in J_{x,B}}|D_l \cap [y]_A|(1-\frac{|D_l \cap [y]_A|}{|[y]_A|}).$$

由定理 2 知 B 是 S 的相对信息熵协调集.

参考文献:

[1] Zhang W X,Mi J S,Wu W Z. Approaches to knowledge reductions in inconsistent systems [J]. International Journal of Intelligent Systems,2003(18):989-1000.

[2] Kosko B. Fuzzy entropy and conditioning[J]. Information Sciences,1986,40:165-174.

[3] 张文修,梁怡,吴伟志. 信息系统与知识发现[M]. 北京:科学出版社,2003.

[4] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982,11:341-356.

[5] 张文修,吴伟志,梁吉业. 粗糙集理论与方法[M]. 北京:科学出版社,2001.

[6] Skowron A,Rauszer C. The discernibility matrices and functions in information systems [M]//Slowi R. Intelligent decision support. Handbook of Applications and Advances of Rough Sets Theory. Dordrecht: Kluwer Academic Publishers,1992:331-362.

(责任编辑:尹　闯)

(上接第 85 页)

例 3.1　考虑

$$A=\begin{pmatrix}1 & \frac{1}{2} & \frac{1}{2}\\ 0 & 1 & 1\\ 0 & 1 & -1\end{pmatrix}.$$

矩阵 A 的特征值为 $-\sqrt{2}$,1,$\sqrt{2}$. 估计区间的结果如表 1 所示.

表 1　估计区间结果

定理	特征值包含区域
盖尔圆定理[7]	[-2,2]
Cassini 卵形域[8]	$[-\sqrt{2},2]$
引理 1.1[4]	[-3,2]
引理 1.2[2]	$[-1-\sqrt{2},2]$
定理 3.3	[-2,-1],1,[1,1.5564]

由表 1 结果容易看出定理 3.3 的结果优于其他的定理.

参考文献:

[1] Fieldler M,Ptak V. On matrices with non-positive off-diagonal elements and positive principal minors[J]. J Czech Math,1962,87:382-400.

[2] Peña J M. On an alternative to gerschgorin circles and ovals of cassini[J]. Numer Math,2003,95:337-345.

[3] Li Houbiao,Huang Tingzhu,Li Hong. On some subclasses of P- matrices [J]. Numerical Linear Algebra Appl,2007,14:391-405.

[4] Peña J M. A class of P -matrices with applications to the localization of the eigenvalues of a real matrix[J]. SIAM J Matrix Anal Appl,2001,22:1027-1037.

[5] Berman A,Plemmons R J. Nonnegative matrices in the mathematical sciences[M]. New York:Academic Press, 1979.

[6] 安国斌,郭希娟. 双对角占优与非奇 M- 矩阵的判定[J]. 应用数学与计算数学学报,2002,14(2):93-96.

[7] Horn R A, Johnson C R. Matrix analysis [M]. Cambridge:
Cambridge Univ Press,1985.

[8] Brualdi R A. Matrices,eigenvalues and directed grtaphs[J]. Lin Multilin Alg,1982,11:143-165.

(责任编辑:尹　闯)