维普资讯 http://www.cqvip.com

决策表属性约简的相对信息量表示

胡逢彬[1],桂现才[2]
(1.湛江师范学院 商学院,广东 湛江 524048;
2.湛江师范学院 数学与计算科学学院,广东 湛江 524048)

摘 要:知识约简是粗糙集理论研究的主要内容之一,该文在信息系统中引入了知识的相对信息量的概念。对一致决策表,证明了其属性约简在代数表示下和相对信息量表示下是等价的,但对不一致决策表,举例说明其属性约简的代数表示不能用相对信息量来等价表示。由此可见,相对信息量表示比代数表示直观,但不能完全代替代数表示方法。

关键词:粗糙集;决策表;相对信息量;互信息;属性约简
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2006)07-0039-03

Relative Information Quantity Representation for Attribute Reduction of Decision Tables

HU Feng-bin[1], GUI Xian-cai[2]
(1. Business School, Zhanjiang Normal College, Zhanjiang 524048, China;
2. Mathematics and Computational Science School, Zhanjiang Normal College, Zhanjiang 524048, China)

Abstract: Reduction of knowledge is one of the important topics in the research on the rough set theory. In this paper, the concept of the relative information quantity of decision table is introduced in the information system. In a consistent decision table, the equivalence properties between algebraic representation and relative information quantity representation of attribute reduction are proved. Through examples, it shows that attribute reduction of an inconsistent decision table cannot entirely be represented by relative information quantity. This shows that relative information quantity representation is more visual than algebraic representation, but it can't displace absolutely each other.

Key words: rough set; decision tables; relative information quantity; mutual information; attribute reduction

0 引 言

粗糙集(Rough Set)[1,2]理论是波兰数学家 Z. Pawlak 于1982年提出的一种处理模糊、不精确的分类问题的新型数学工具。其主要思想是:在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。

粗糙集的所有概念和运算都是通过代数学的等价关系和集合运算来定义的,但这种表示的直观性较差。近年来,一些学者从信息论的角度对粗糙集理论进行研究,提出了粗糙集理论的信息论观点[3,4]。在信息观中,主要是从 shannon 熵和条件熵入手,文献[5,6]从信息量和条件信息量出发,讨论了属性约简方法。文中对条件信息量的概念进行改造,给出了相对信息量和互信息的概念,建立了知识与相对信息量的关系,同样证明了对于一致决策表,其属性约简在代数表示和相对信息量表示下是等价的。但对不一致决策表,两者表示结果是不一致的。

收稿日期:2006-02-15
基金项目:湛江师范学院科研基金(W0428)
作者简介:胡逢彬(1968-),男,江西泰和人,讲师,硕士,从事计算机基础教学和粗糙集理论研究。

1 基本概念

本节仅介绍一下与文中有关的主要概念,其他概念可参考文献[1,2]。

定义1 四元组 $S=\langle U,A,V,f\rangle$ 是一个信息系统,其中 U 表示对象的非空有限集合,称为论域;A 表示属性的非空有限集合;$V=\bigcup_{a\in A}V_a$,V 为属性值的集合,V_a 表示属性 $a\in A$ 的值域;$f:U\times A\to V$ 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 给定信息系统 $S=\langle U,A,V,f\rangle$,对于每个子集 $R\subseteq A$,可定义等价关系 IND(R),称为不可分辨关系,定义为:$\mathrm{IND}(R)=\{(x,y)\in U\times v \mid \forall a\in R, f(x,a)=f(y,a)\}$

对于每个子集 $X\subseteq U$ 和不可分辨关系 $R\subseteq A$,X 的下近似集和上近似集可以分别定义为:$R_-(X)=\bigcup\{Y\in U \mid \mathrm{IND}(R) \mid Y\subseteq X\}$;$R^-(X)=\bigcup\{Y\in U \mid \mathrm{IND}(R) \mid Y\cap X\neq\varnothing\}$。

定义3 设 $S=\langle U,A,V,f\rangle$ 是一个信息系统,若 $A=C\cup D$,且 $D\cap C=\varnothing$,C 和 D 分别为条件属性集和决策属性集,则信息系统称为决策表,下文把决策表简记为 $T=<U,C\cup D,V,f>$。

维普资讯 http://www.cqvip.com

定义 4 对决策表 $T = < U, C \cup D >$，若 $U/\mathrm{IND}(C) \subseteq U/\mathrm{IND}(D)$，则称决策表是一致的(相容的)，否则称决策表是不一致的(不相容的)。

在一致决策表中，当对象在条件属性集上取值相同时，决策属性值也必定相同；而在不一致决策表中，至少存在两个对象，在条件属性集上取值相同，但它们的决策值却不相等。

定义 5 设 $T = < U, C \cup D, V, f >$ 是决策表，如果 $\mathrm{POS}_C(D) = \mathrm{POS}_{C-\{a\}}(D)$，则称属性 a 是关于 D 可省的，否则称属性 a 是关于 D 不可省的。其中：$\mathrm{POS}_B(D) = \bigcup_{x \in U/\mathrm{IND}(D)} B_-(X)$ 是 D 关于 B 的正域。

定义 6 如果决策表 $T = < U, C \cup D, V, f >$ 中每个条件属性 $a \in C$ 都是关于 D 不可省的，则称条件属性集 C 是关于 D 独立的，否则称 C 是关于 D 依赖的。

定义 7 决策表 $T = < U, C \cup D, V, f >$ 中条件属性集 C 的一个子集 B 是关于 D 独立的，并且 $\mathrm{POS}_B(D) = \mathrm{POS}_C(D)$，则称 B 是 C 的一个 D - 约简。

2 知识的信息量和相对信息量

定义 8 给定信息系统 $S = \langle U, A, V, f \rangle$，$P \subseteq A$，$Q \subseteq A$，$U/\mathrm{IND}(P) = \{X_1, X_2, \cdots, X_n\}$，$U/\mathrm{IND}(Q) = \{Y_1, Y_2, \cdots, Y_m\}$，知识(属性集合)$P$ 的信息量定义为：

$$E(P) = \sum_{i=1}^{n} \frac{|X_i|}{|U|} \frac{|X_i^c|}{|U|} = \sum_{i=1}^{n} \frac{|X_i|}{|U|}\left(1 - \frac{|X_i|}{|U|}\right)$$

知识(属性集合)Q 相对于知识(属性集合)P 的条件信息量 $E(Q \mid P)$ 定义为：

$$E(Q \mid P) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|U|} \frac{|Y_j^c - X_i^c|}{|U|} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|U|} \frac{|X_i - Y_j|}{|U|}$$

P 与 Q 的互信息定义为：

$$E(Q;P) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|U|} \frac{|X_i^c \cap Y_j^c|}{|U|} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|U|} \frac{|U - (X_i \cup Y_j)|}{|U|}$$

其中 $|X|$ 表示集合 X 的基数，$R_i^c = U - R_i$，$|R_i|/|U|$ 表示等价类 R_i 在论域 U 上的可能性(概率)，$|R_c^i|/|U|$ 表示 R_i 的余集在论域 U 上的可能性，也即不属于 R_i 的概率。有关 $E(R)$ 的性质以及如何利用 $E(R)$ 对知识进行属性约简，可以参考文献[5]。

定义中用“相对信息量”概念，而不用“条件信息量”，是为了与文献[6]的“条件信息量”区别，此外式子中没有了条件概率的意义。

由于 $Y_j = \bigcup_{i=1}^{n}(Y_j \cap X_i)$；$Y_j^c = (Y_j^c - X_i^c) \cup (X_i^c \cap Y_j^c)$；可容易得到下面的定理：

定理 1[7] 给定信息系统 $S = \langle U, A, V, f \rangle$，$P \subseteq A$，$Q \subseteq A$，$U/\mathrm{IND}(P) = \{X_1, X_2, \cdots, X_n\}$，$U/\mathrm{IND}(Q) = \{Y_1, Y_2, \cdots, Y_m\}$，则有：

(1) $E(Q) = E(Q;P) + E(Q \mid P)$

(2) $E(P) = E(P;Q) + E(P \mid Q)$

(3) $E(Q;P) = E(P;Q)$

3 一致决策表的等价表示

定理 2 决策表 $T = (U, A, C, D)$ 是一致的，其充分必要条件是 $E(D \mid C) = 0$。

证明：设 $U/\mathrm{IND}(C) = \{X_1, X_2, \cdots, X_n\}$，$U/\mathrm{IND}(D) = \{Y_1, Y_2, \cdots, Y_m\}$

(必要性) 由于 T 是一致的，即 $U/\mathrm{IND}(C) \subseteq U/\mathrm{IND}(D)$，任取 $X_i \in U/\mathrm{IND}(C)$，存在 $Y_j \in U/\mathrm{IND}(D)$，使得 $X_i \subseteq Y_j$，又因为 $\{Y_1, Y_2, \cdots, Y_m\}$ 是 U 上的一个划分，所以对任意 $k(1 \leqslant K \leqslant m)$，且 $k \neq j$，必有 $Y_k \cap X_i = \varnothing$，总之，对任意 i, j，或者 $Y_j \cap X_i = X_i$ 或者 $Y_j \cap X_i = \varnothing$ 成立；若 $Y_j \cap X_i = X_i$，则 $|X_i - Y_j| = 0$；若 $Y_j \cap X_i = \varnothing$，则 $|Y_j \cap X_i| = 0$，由 $E(D \mid C)$ 的定义可知，$E(D \mid C) = 0$。

(充分性) 设：$E(D \mid C) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|U|} \cdot \frac{|X_i - Y_j|}{|U|} = 0$

则对任意 $i, j (i = 1,2,\cdots,n; j = 1,2,\cdots,m)$，有 $|Y_j \cap X_i| = 0$，或者 $|X_i - Y_j| = 0$ 成立，则 $Y_j \cap X_i = \varnothing$ 或 $Y_j \cap X_i = X_i$ 成立(任意 i, j)，所以，对任意 $X_i \in U/\mathrm{IND}(C)$，则一定存在 $Y_j \in U/\mathrm{IND}(D)$，使得 $X_i \subseteq Y_j$，由定义 3 知决策表 T 是一致的。

本定理说明，决策表是否一致可由决策属性相对于条件属性的相对信息量的值来判断。

推论 1 决策表 $T = (U, A, C, D)$ 是一致的，其充分必要条件为 $E(D;C) = E(D)$

由定理 1 以及 $E(D;C) = E(D) - E(D \mid C)$，则结论成立。

推论 2 对决策表 $T = (U, A, C, D)$，以下条件等价：

决策表是不一致的；$E(D \mid C) > 0$；$E(D;C) \neq E(D)$。

定理 3 设 $T = (U, A, C, D)$ 是一致的决策表，以下条件等价：

(1) 条件属性 a 是关于 D 可省的；

(2) $E(D \mid C - \{a\}) = E(D \mid C) = 0$；

(3) $E(D;C - \{a\}) = E(D;C)$

证明：(下面证明(1)与(2)等价，(1)与(3)的等价证明类似。)

决策表 $T = (U, A, C, D)$ 是一致的且条件属性 a 是关于 D 可省的 $\Leftrightarrow$

决策表 $T_1 = (U, A - \{a\}, C - \{a\}, D)$ 是一致的 $\Leftrightarrow E(D \mid C - \{a\}) = 0 = E(D \mid C)$。

本定理说明，一致的决策表中相关可省的属性没有提

维普资讯 http://www.cqvip.com

供新的信息,反之亦然。

推论 3　设 $T=(U,A,C,D)$ 是一致的决策表,以下条件等价:

(1) 条件属性 a 是关于 D 不可省的;

(2) $E(D\mid C-\{a\})>0$;

(3) $E(D;C-\{a\})\neq E(D;C)$。

推论 4　设 $T=(U,A,C,D)$ 是一致的决策表,以下条件等价:

(1) 条件属性集 C 是关于 D 独立的;

(2) 对任意 $a\in C$,有 $E(D\mid C-\{a\})>0$;

(3) 对任意 $a\in C$,有 $E(D;C-\{a\})\neq E(D;C)$。

定理 4　设 $T=(U,A,C,D)$ 是一致的决策表,以下条件等价:

(1) 条件属性集 C 的一个子集 B 是 C 的一个 D - 约简;

(2) 对任意 $a\in B$,有 $E(D\mid B-\{a\})>0$,且 $E(D\mid B)=0$;

(3) 对任意 $a\in B$,有 $E(D;B-\{a\})\neq E(D;B)$,且 $E(D;B)=E(D;C)$。

证明:下面证明(1)与(2)等价,(1)与(3)的等价证明类似。

由约简的定义知,B 是 C 的一个 D - 约简的充分必要条件是:(1)B 是关于 D 独立的;(2)决策表 $T_1=(U,A_1,B,D)$ 是一致的,其中 $A_1=B+D$。

而这两个条件又等价于:(1) 对任意属性 $a\in B$,有 $E(D\mid B-\{a\})>0$;(2)$E(D\mid B)=0$。

以上定理说明,对于一致决策表,其属性约简用相对信息量来表示和原来的代数表示是等价的,但用相对信息量表示比代数表示更直观。而对于不一致的决策表,情况如何?考察下面的例子。

例 1　不一致决策表 $T=(U,A,C,D)$ 如表 1 所示,条件属性 $C=\{a,b,e\}$,决策属性 $D=\{d\}$。

表 1　决策表 $T=(U,A,C,D)$

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

$U/\{a\}=U/\mathrm{IND}(a)=\{\{1,2,5,8\},\{3,6,7\},\{4\}\}$;

$U/\{b\}=U/\mathrm{IND}(b)=\{\{1,2,4,5,7,8\},\{3\},\{6\}\}$;

$U/\{e\}=U/\mathrm{IND}(e)=\{\{1,5\},\{2,7,8\},\{3,4,6\}\}$;

$U/\{d\}=U/\mathrm{IND}(D)=\{Y_1,Y_2,Y_3,Y_4,Y_5\}=\{\{1\},\{2,7\},\{3,6\},\{4\},\{5,8\}\}$;

$U/\{a,b\}=U/\mathrm{IND}(C-\{e\})=\{Z_1,Z_2,Z_3,Z_4,Z_5\}=\{\{1,2,5,8\},\{3\},\{4\},\{6\},\{7\}\}$;

$U/\{a,b,e\}=U/\mathrm{IND}(C)=\{X_1,X_2,X_3,X_4,X_5,X_6\}=\{\{1,5\},\{2,8\},\{3\},\{4\},\{6\},\{7\}\}$;

由于 $\mathrm{POS}_C(D)=\{3,4,6,7\}=\mathrm{POS}_{C-\{e\}}(D)$,由代数观点可知,属性 e 是关于 D 可省的,但是:

$$E(D\mid C)=\sum_{i=1}^{6}\sum_{j=1}^{5}\frac{|X_i\cap Y_j|}{|U|}\frac{|X_i-Y_j|}{|U|}=\frac{1}{8}\times\frac{1}{8}(2+2+0+0+0+0+0)=\frac{2}{32}$$

而 $E(D\mid C-\{c\})=\sum_{i=1}^{5}\sum_{j=1}^{5}\frac{|Z_i\cap Y_j|}{|U|}\frac{|Z_i-Y_j|}{|U|}=\frac{1}{8}\times\frac{1}{8}(10+0+0+0+0)=\frac{5}{32}$

这里 $E(D\mid C-\{e\})>E(D\mid C)$,虽然去掉的是关于 D 可省的属性,但还是使决策表的相对信息量发生了改变。可见,不一致决策表中属性是否相关可省与相对信息量是否改变并没有等价的关系。从例 1 可以看出,在去掉属性 c 之前,规则 1 与规则 5 冲突,规则 2 与规则 8 冲突;而在去掉属性 c 之后,规则 1,2,5,8 均相互冲突,显然这时的不确定信息增多,导致相对信息量的改变。而在粗糙集理论中,对这两种情况并不加以区别,从而导致了决策表属性约简的代数表示和相对信息量表示的不一致性。

4　结　论

目前,关于粗糙集理论的研究和应用越来越广泛,文中建立了决策表与相对信息量之间的关系,证明了一致决策表的属性约简在相对信息量表示和代数表示下是等价的,但是,对不一致决策表的属性约简却不能用相对信息量来等价表示。由此可见,虽然用相对信息量来表示粗糙集的部分理论可能比代数表示更加直观,能够导出高效的知识约简算法,但是它并不能完全取代粗糙集理论的代数表示方法。

参考文献:

[1] Pawlak Z. Rough set theoretical aspects of reasoning about date [M]. Warsaw, Poland: [s. n.], 1991. 72 - 79.

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

[3] 苗夺谦,王　珏. 粗糙集理论中概念和运算的信息表示[J]. 软件学报,1999(2):113 - 116.

[4] 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.

[5] 梁吉业,曲开社,徐宗本. 信息系统的属性约简[J]. 系统工程理论与实践,2001,21(12):76 - 80.

[6] 刘振华,刘三阳,王　珏. 基于信息量的一种属性约简算法[J]. 西安电子科技大学学报,2003,30(6):835 - 838.

[7] Liang JiYe, Chin K S, Dang ChuangYin, et al. A new method for measuring uncertainty and fuzziness in rough set theory [J]. International Journal of General System, 2002, 31(4): 330 - 342.