文章编号:1000 - 582X(2006)10 - 0127 - 04

不完备信息系统下的一种概率属性约简算法*

白俊卿,刘琼荪
(重庆大学 数理学院,重庆 400030)

摘　要:不完备信息系统中的属性约简一直是粗糙集研究领域的一个难点.在不完备信息系统中,用以往的属性约简算法得到的约简结果中,某一属性要么属于该约简,要么不属于该约简.但在实际生活中,当属性以比较大的概率可区分两对象时,这就表明该属性可能以某一概率属于约简结果.基于这种想法,文中通过构造概率区分矩阵,并在此基础上给出相应的区分函数,提出了一种概率属性约简算法,分析了算法正确性.从该算法得到的约简,可以看出各属性属于约简的可能性.最后用实例表明该算法是有效和可行的.

关键词:容差关系;概率核;概率区分矩阵;区分函数
中图分类号:O23　　**文献标识码**:A

粗糙集理论是 Z. Paklaw[1] 教授于 1982 年提出来的一种处理不确定性知识的数学工具.经典的粗糙集理论以等价关系为基础,以完备的信息系统为对象,然而现实生活中由于能力、技术、资金等原因我们无法获取一些信息或者要获取这些信息代价太大或有时由意外原因使得数据库破损数据遗失,从而我们得到的信息系统几乎都是不完备的,因此在经典的粗糙集理论基础上对不完备信息系统进行研究非常必要.针对不完备信息系统,人们将等价关系扩充为容差关系,相似关系,然而这 2 种关系使得知识粒度过大,概念表示不清楚,不少人提出了各种粗糙集模型[2-4].然而不管是那种模型,对于属性约简结果来说,某一属性要么属于该约简要么不属于该约简.而在不完备信息系统里,当两对象中至少有一对象在某一属性上为缺省值时,按上述两关系而言,该属性无法区分两对象.但往往实际生活中,该属性以比较大的概率可区分两对象,这就表明该属性以某一概率属于属性约简.基于这种想法提出了一种概率属性约简,并给出算法.

1　基本概念

1.1　不完备信息系统

设 $S=(U,A)$ 是信息系统,其中 U 是对象的非空有限集合,A 是属性的非空有限集合,对于每个 $a\in A$ 有 $a:\ U\to V_a$,其中 V_a 称为 a 的值域.对于一个对象,一些属性值可能是缺省的,通常给定一个区分值(即空值)给这些属性.如果至少有一个属性 $a\in A$ 使得 V_a 含有空值,则称 S 为一个不完备信息系统,此处用 $*$ 表示空值.

1.2　容差关系[5]

设不完备信息系统 $S=(U,A)$,其中 U 是对象的非空有限集合,$A=C\cup D$,$C\cap D=\phi$,C 是条件属性,D 是决策属性,对 $\forall x,y\in U$,$\forall B\subseteq A$,定义容差关系:

$$T=\left\{\begin{matrix}(x,y)\in U\times U\mid \forall b\in B, b(x)=*\ or\\ b(y)=*\ or\quad b(x)=b(y)\end{matrix}\right\}.$$

对象 $x\in U$ 的容差类定义为:

$$T(x)=\{y\in U\mid (x,y)\in T\},$$

U 上的概念 X 的下上近似分别为:

$$\underline{T}=\{x\in U\mid T(x)\subseteq X\};$$

$$\overline{T}=\{x\in U\mid T(x)\cap X\neq\phi\}.$$

1.3　概率核

以 $0<p<1$ 的概率属于属性约简的核称为概率核.

1.4　概率区分矩阵

给定阈值 λ(用户所要求的最低可区分概率,一般取 $\lambda>\frac{1}{2}$),$p(x)=P(c(x)=z\mid d(x)=r)\geq\lambda$,表示对象在决策值为 r 时在属性 c 上取值为 z 的条件概率.如果 $c(x)=*$ 且 $P(c(x)=z\mid d(x)=r)\geq\lambda$,则令 $c(x)=$

* 收稿日期:2006 - 06 - 05
作者简介:白俊卿(1979 -),女,山西交城人,重庆大学硕士研究生,主要从事智能计算与应用的研究.

z,此时不完备信息系统中的某些值已得到修改,得到系统 S',在该系统里构造区分矩阵

$$M'(i,j)=\begin{Bmatrix}\{c_k \mid c_k\in(B_v(x_i)\cap B_v(x_j))\\ \wedge\, c_k(x_i)\neq c_k(x_j)\}\ ,\\ 当\ d(x_i)\neq d(x_j)\ \wedge\\ B_v(x_i)\cap B_v(x_j)\neq\phi\\ \wedge\ (\exists c_k\in(B_v(x_i)\cap B(x_j))\\ \wedge\, c_k(x_i)\neq c_k(x_j))\\ 0,其他\end{Bmatrix}$$

$$B_v(x)=\{c_k\in C\mid c_k\neq *\}$$

在区分矩阵 $M'(i,j)$ 里,如果 $M'(i,j)\neq 0$,对每一个 $c_k\in M'(i,j)$,若 $c_k(x_i)=z$ 是以概率 $\lambda\leqslant p=P(c_k(x_i)=z\mid d(x_i)=r)<1$ 修正后的值,并且 $c_k(x_j)$ 未做修正,则 c_k 用 pc_k 代替;同理若 $c_k(x_i)$ 未修正,$c_k(x_j)$ 是修正后的值 $\lambda\leqslant p_0=P(c_k(x_j)=z_0\mid d(x_j)=r_0)<1$,则 c_k 用 p_0c_k 代替;若 $c_k(x_i)$ 和 $c_k(x_j)$ 都是修正值,则取 $p_0=P(c_k(x_i)=z\mid d(x_i)=r)P(c_k(x_j)=z_0\mid d(x_j)=r_0)$,如果 $p_0<\lambda$,则令 $p_0=0$,c_k 用 0 来代替,如果 $p_0\geqslant\lambda$ 则 c_k 用 p_0c_k 代替.

此时得到的矩阵称为概率区分矩阵记为 $M(i,j)$. 通常在属性约简过程中概率区分矩阵用一个行为 $\frac{n(n-1)}{2}$,列为 m 的表来表示,其中 n 为对象的个数,m 为条件属性个数,其元素为

$$M_1((x_i,x_j),k)=\begin{Bmatrix}1,M(i,j)\neq 0\ \wedge\ c_k\in M(i,j)\\ p_0,M(i,j)\neq 0\ \wedge\ p_0c_k\in M(i,j)\\ 0,其它\end{Bmatrix}$$

2 利用区分函数进行属性约简

在概率区分矩阵中,若 $M(i,j)\neq 0$,则令 $\sum M(i,j)$ 是包含在 $M(i,j)$ 中属性对应变量的析取,其中把 pc_k 看作整体定义,如果 $p_1<p_2$ 则 $p_1c_k\wedge p_2c_k=p_2c_k$,$p_1c_{k1}\wedge p_2c_{k2}=p_1c_{k1}\wedge p_2c_{k2}$,$p_1c_k\vee p_2c_k=p_2c_k$,如果 $p_1=p_2$ 则 $p_1c_k\wedge p_2c_k=p_2c_k$,$p_1c_k\vee p_2c_k=p_2c_k$,$p_1c_{k1}\wedge p_2c_{k2}=p_1c_{k1}\wedge p_2c_{k2}$,$p_1c_{k1}\vee p_2c_{k2}=p_1c_{k1}\vee p_2c_{k2}$;若 $M(i,j)=0$,令 $\sum M(i,j)=1$,这里 $k1\neq k2$. 则决策表的区分函数为:$\nabla^*=\prod\limits_{(x_i,x_j)\in(U,U)}\sum M(i,j)$,决策表中对象 x_i 的区分函数为:$\nabla^*=\prod\limits_{x_j\in U}\sum M(i,j)$.

3 基于概率区分矩阵的属性约简算法

如果希望在区分两对象的属性中取概率较大的,即若属性 a 以概率 p_1 可以区分对象 x_1 和 x_2,属性 b 以概率 p_2 可以区分对象 x_1 和 x_2,且 $p_1>p_2$ 则取 a 为约简中的属性,此时采用算法 1;如果实际中只要求约简中所包含的属性至少以某一概率可以区分对象,此时采用算法 2. 这里当 $\lambda=1$ 时得到的约简即为原信息系统的约简.

3.1 算法 1

输入 λ,S'(为修正后的不完备系统),输出 B 为约简结果.

1)先对决策表进行简化,简化后的决策表不含有相同的对象,用 U_1 表示,相同的对象用 N 作了标记.

2)求出系统的概率区分矩阵并获得表 $M_1((x_i,x_j),k)$.

3)对 $1\leqslant i\leqslant n-1$,$i<j\leqslant n$ 若 $d(x_i)\neq d(x_j)$,则先求得 $\max=\max\limits_{1\leqslant k\leqslant m}\{m_1((x_i,x_j),k)\}$,$m$ 为条件属性的个数,然后对每个 $k(1\leqslant k\leqslant m)$,判断,如果 $M_1((x_i,x_j),k)<max$,则令 $M_1((x_i,x_j),k)=0$.

4)求出 $M_1((x_i,x_j),k)$ 的列和,找出列和中的值大于 0 且小于或等于 1 的元素所在行 $\{r_1,r_2,\cdots,r_k\}$,找出这些行中的值大于 0 且小于或等于 1 的元素所在的列 $\{l_1,l_2,\cdots,l_k\}$,并记下相应的元素值 $\{y_1,y_2,\cdots,y_k\}$. 若同一列 l_m 对应多个不同的 y_m 值,则取这些值中最大的赋给 $B(l_m)$,若唯一直接赋给 $B(l_m)$. (注:因为 l_m 与 l_n 有可能相等即指的是同一列,所以在同一列上可能出现多个 y_m). 这样得到的 $B(l_m)$ 中,若 $B(l_m)=1$ 表明属性 l_m 就是通常所说的核,否则就是一个概率核.

5)令 $N=B$.

6)置 $B(l_m)=1$ 的列 l_m 中非零元素所在行的每一元素为 0,得到 $M_2((x_i,x_j),k)$.

7)找出 $M_2((x_i,x_j),k)$ 的行和中的最大值所在的列,若存在多列任选一列,找出该列中的最大元素 z,令 $B(l)=z$,置该列中所有非 0 元素所在行的元素为 0,得到 $M_3((x_i,x_j),k)$ 赋给 $M_2((x_i,x_j),k)$.

8)如果 $M_2((x_i,x_j),k)$ 为非零矩阵,返回步骤 7,直到 $M_2((x_i,x_j),k)$ 为零矩阵.

9)结束,得到 B. 若 $B(k)=0$ 表示第 k 个属性不属于约简;若 $B(k)=1$ 表示第 k 个属性完全属于属性约简;若 $0<B(k)<1$ 表示第 k 个属性以概率 $B(k)$ 属于约简.

3.2 算法 2

该算法只在算法 1 的基础上稍做修改,输入 λ,S'(为修正后的不完备系统),输出 B 为约简结果

1)与算法 1 中的步骤 1 相同.

2)与算法 1 中的步骤 2 相同.

3)求出 $M_1((x_i,x_j),k)$ 的列和,找出列和中的值大于 0 且小于或等于 1 的元素所在行 $\{r_1,r_2,\cdots,r_k\}$,找出这些行中的值大于 0 且小于或等于 1 的元素所在

的列$\{l_1,l_2,\cdots,l_k\}$,并使$B(l_m)=1$.

4)置列$l_m(1\leqslant m\leqslant k)$中非零元素所在行的每一元素为0,得到$\boldsymbol{M}_2((x_i,x_j),k)$.

5)找出$\boldsymbol{M}_2((x_i,x_j),k)$的行和中最大值所在的列,若存在多列任选一列,令$B(l)=1$,置该列中所有非0元素所在行的元素为0,得到$\boldsymbol{M}_3((x_i,x_j),k)$赋给$\boldsymbol{M}_2((x_i,x_j),k)$.

6)如果$\boldsymbol{M}_2((x_i,x_j),k)$为非零矩阵,返回步骤5,直到$\boldsymbol{M}_2((x_i,x_j),k)$为零矩阵.

7)输出B,结束.

3.3 算法分析

如果对原系统不做任何修改或置$\lambda=1$,则从文献[6]知该算法可得到约简结果.当对原系统作了某些修正时,且置$\lambda<1$时,表$\boldsymbol{M}_1((x_i,x_j),k)$中不只包含有0和1这两元素值,也包含了大于$\lambda$而小于1的数值,因而有可能出现一些属性为概率核.在算法1中,通过步骤4可以得到约简中的概率核和通常核,步骤5用N保留了概率核和通常核.因为有可能一个属性既完全属于原系统的属性约简结果,又有可能是修正系统后的概率核.步骤6和7保证了在这一种情况下该属性仍是完全属于约简的.通过比较N和B知如果$N(i)=B(i)$且$0<N(i)<1$,则表明属性既是概率核同时以某一概率属于约简的,可以按照该概率来决定是否要采集对象在该属性上的值.若$N(i)=B(i)=1$表明该属性一定是核,采样或预测时肯定得先考虑该属性,若$N(i)\neq B(i)$且$0<N(i)<1$,$B(i)=1$表明该属性属于约简,而且很有可能是核,所以采样或预测时先考虑通常属性核,次之考虑该属性,然后才是约简中的其它属性.

3.4 实　例

给定不完备信息系统$S=\{U,A\}$,其中$U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$ $A=\{a,b,c,d,e\}$,$\{a,b,c,d\}$是条件属性集合,$\{e\}$是决策属性,见表1,表2为修正后的系统S'.

表1　系统S

u	a	b	c	d	e
x_1	1	1	1	1	1
x_2	2	*	1	1	1
x_3	*	*	2	2	2
x_4	1	*	1	2	1
x_5	*	*	1	2	3
x_6	2	1	1	*	1

说明：*表示缺省值

表2　系统S'

u	a	b	c	d	e
x_1	1	1	1	1	1
x_2	2	*	1	1	1
x_3	*	*	2	2	2
x_4	1	*	1	2	1
x_5	1-8/9	*	1	2	3
x_6	2	1	1	1-7/9	1

说明：*表示缺省值；{1-8/9}表示对象5在属性a上取值为1的概率为8/9.

原系统S的属性约简为$\{c,d\}$,在系统S'里取$\lambda=0.7$按照算法1得到的约简为$\{8/9a,c,d\}$,在这里属性$\{a\}$是一个概率核,它以8/9的概率属于约简,按照这个结果我们就可以以8/9的概率来决定是否要采集对象在属性$\{a\}$上的取值.按照算法2得到的约简为$\{c,d\}$.比较这两结果可以看出能用属性a区分的对象可以用c或d以不低于$\lambda=0.7$的概率来区分.

4　$p(x)=P(c(x)=z|d(x)=r)$的获取方法

1)按照文中的思想,最主要是通过专家或以往经验给出.

2)用属性在其它所有对象的取值次数最多的值来填充该缺失的属性值,概率就取其出现的频率.或定义一个相似标准,在所有与该对象相似的对象中找出频率最高的值来替代,概率就取其频率.

3)也可通过其它合理的方法给出.

5　与其它算法的比较

文献[7]和文献[8]以及该文都是将容差关系作了限制,即对两对象在某属性上相似作了限制.从约简结果来说,文献[7]和文献[8]得到的约简中的属性都是以同一概率属于约简的,而从该文得到的约简可以看到属性在满足给定阈值的条件下属于约简的最大概率.

6　结　论

文中给出了概率区分矩阵的概念,在此基础上提出了概率属性约简算法.其最大的特点是能够从该算法得到的约简中判断出某一属性属于约简的概率,实例表明该算法是有效的.

参考文献:

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

[2] 张宏宇,梁吉业. 不完备信息系统下的变精度粗糙集模型及其知识约简算法[J]. 计算机科学,2003,30(4):153 - 155.

[3] 黄兵,周献中. 不完备信息系统中基于联系度的粗集模型拓展[J]. 系统工程理论与实践,2004,(1):88 - 92.

[4] 程玉胜,胡学钢,江效克. 不完备信息系统的属性约简方法研究[J]. 计算机工程与应用,2004,(1):68 - 70.

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

[6] 潘丹,曾安,郑启伦. 一种基于粗糙集理论的智能故障诊断新方法[J]. 计算机工程与应用,2003,(14):48 - 50.

[7] 李滢雪,吴顺祥. 一种基于容差关系的辨识矩阵属性约简法[J]. 厦门大学学报,2005,44(增):281 - 283.

[8] 张宏宇,梁吉业. 不完备信息系统下的变精度粗糙集模型及其知识约简算法[J]. 计算机科学,2003,30(4):153 - 155.

Probability Algorism for Attributes Reduction in Incomplete Information System

BAI Jun-qing, . LIU Qiong-sun

(College of Mathematics and Physics, Chongqing University, Chongqing, 400030, China)

Abstract: Attributes reduction based on rough set theory is an important but difficult task under incomplete information system. For the attributes reduction which is gained by the old attributes reduction algorisms , the attribute belongs to it or not . Nevertheless in the practice when there is the probability that the attribute can discern two objects ,this shows the attribute may belong to the attribute reduction. Probability discernibility matrix is defined and corresponding discernibility function is given. Then a probability algorism for attributes reduction is proposed and an example shows the algorism is effective. The probability that the attribute belong to the reduction can be know from the reduction which is gained by the algorism.

Key words: tolerance relation; probability core; probability discernibility matrix; discernibility function

(编辑 姚 飞)