一种不完备信息系统的属性约简算法研究①

An Algorithm for Attribute Reduction in Incomplete Information System

赵亚鹏　丁以中　（上海海事大学 交通运输学院 上海 200135）

摘　要：通过分析不完备信息系统的粗糙集模型，引入了一种基于约束相似关系的二进制可辨矩阵的构造方法，该方法不仅适用于一致性的不完备信息系统，也适用于不一致性的不完备信息系统，并提出了一种基于下近似二进制可辨矩阵直接求取不完备信息系统的属性核和属性约简的算法，实验表明算法简单有效。

关键词：粗糙集 二进制可辨矩阵 不完备信息系统 属性约简 约束相似关系

在实际工作中，由于数据测量的误差、数据登记时的疏忽和数据获取限制等原因，常使得大多数信息系统是不完备的，对于不完备信息系统，经典的粗糙集理论存在局限性，因为基于经典不分明关系的 Rough 集理论所处理的信息系统必须是完备的，即每个对象在所有属性集上的取值是已知的，不能处理不完备信息系统的。因此，近年来，从不完备信息系统提取知识逐渐成了粗糙集理论和应用研究的一个热点，对不完备信息系统的研究也取得了许多成果[1,2]，但大多数算法较复杂，计算较繁琐。

文献[3]对可辨矩阵进行了改进，用二进制 0 和 1 作为分明矩阵中的元素，提出了二进制可辨矩阵，由于采用了二进制的表达形式，其计算比施行等价类计算要快得多，灵活得多，而且更加简单直观。文献[3,4]已对利用二进制可辨矩阵求取完备信息系统的属性约简做了一些探索，不过是基于完备信息系统的。

1　基本概念

四元组 $S=\{U,R,V,F\}$ 是一个信息系统，其中 U 为非空对象集，称为论域，$R=C\cup D$，$C\cap D=\varnothing$，C 为条件属性集，D 为决策属性集；$\forall a\in R, V=\bigcup_{a\in R}V_a$，$V_a$ 是属性 a 的值域；$f: U\times R\rightarrow V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即 $\forall a\in R; x\in U, f(x,a)\in V_a$ 若至少有一个属性使得含有空值，用 $*$ 表示空值，则称 S 为不完备信息系统；否则称 S 为完备信息系统。

定义 1. 决策表 $S=\{U,C,D,V,f\}$，$\forall x_i, x_i\in U$，$\forall b\in C$，设二进制可辨矩阵 M 中的任一元素 $u((i,j),k)$ 所在行对应对象对 (x_i,x_j)，所在列对应条件属性 b_k，则决策表 S 相应的二进制可辨矩阵 M 定义为：

$$u((i,j),k)=\begin{cases}1 & b_k(x_i)\neq b_k(x_j)\\ 0 & b_k(x_i)=b_k(x_j)\end{cases}$$

定义 2. 在不完备信息系统 S 中，$\forall x,y\in U$，属性子集 $B\subseteq C$，定义相似关系 SIM 为：

$$SIM(B)=\left\{(x,y)\in U\times U\,\middle|\,\forall b\in B(b(x)=*\vee b(x)=b(y))\right\}$$

显然，在 SIM 关系中，一些对象明显有大量相同的已知属性值，直观上就可以判断为相似，但因不满足相似关系而被划分在不同的相似类中，实际上，这些对象相同的可能性较大，这就与实际情况不相符合。因此，文献[5]提出了约束相似关系扩充模型，实例说明其明显优于现有的 Rough 集扩充模型，详见文献[5]。

定义 3. 在不完备信息系统 S 中，$\forall x,y\in U$，属性子集 $B\subseteq C$，定义约束相似关系 CSIM 为：

$$CSIM(B)=\{(x,y)\in U\times U\,|\,\forall b\in B, b(x)=*\\ \vee(P(x)\cap P(y)\neq\varnothing)\wedge(b\in(P(x)\cap P(y))\rightarrow b(x)=b(y)\}$$

其中，$P(x)=\left\{b\,\middle|\,\forall b\in B, b(x)\neq *\right\}$

定义 4. 在不完备信息系统 S 中，$\forall x,y\in U$，约束相似

① 基金项目：上海市教委重点项目（06ZZ43）；上海市重点学科建设项目（T0602）

于 x 的对象集合 $R_B(x)$，x 与之约束相似的对象集合 R_B^{-1} 定义为：

$R_B(x)=\{y\in U:CSIM(y,x)\}$；

$R_B^{-1}(x)=\{y\in U:CSIM(x,y)\}$

定义 5：在不完备信息系统 S 中，基于关系 CSIM，对象集合 $X\subseteq U$ 关于属性子集 $B\subseteq C$ 的下近似（X_{BS}）、上近似（X^{BS}）定义为：

$X_{BS}=\{x:x\in U,R_B^{-1}(x)\subseteq X\}$

$X^{BS}=\{y\in R_B(x),x\in X\}$

考虑到不完备信息系统存在的不一致性。这里引用一个记号：对于 $\forall x_i,\in U$，记：

$\underline{\delta}(x_i)=card(f(x_i,D):x_i\in R_B^{-1}(x_i))$

$\overline{\delta}(x_i)=card(f(x_i,D):x_i\in R_B(x_i))$

为了可以方便地求出不完备信息系统的属性核和属性约简，在约束相似关系下，对二进制可辨矩阵进行了扩充，构造了约束相似关系下的上、下近似二进制可辨矩阵。

2 基于约束相似关系的上、下近似二进制可辨矩阵

定义 7. 不完备信息系统 $S=\{U,R,V,F\}$，$R=C\cup D$，条件属性子集 C，$\forall b\in C$，决策属性集 $D=\{d\}$，定义约束相似关系下的下近似二进制可辨矩阵为：

$$U_C=(u(i,j),k)=\begin{cases}1:\{((x_i,x_j)\in U\times U\wedge\forall b\in C\wedge((b(x_i)\neq *\\ \wedge b(x_j)=*)\vee b(x_i)\neq b(x_j)))\wedge(\forall d\in D\\ \wedge(d(x_i)\neq d(x_j)))\wedge(\min(\underline{\delta}(x_i),\underline{\delta}(x_j))=1)\}\\ 0\qquad\qquad\qquad\qquad\text{否 则}\end{cases}$$

定义 8. 不完备信息系统，$S=\{U,R,V,F\}$，$R=C\cup D$，条件属性子集 C，$\forall b\in C$ 决策属性集 $D=\{d\}$，定义约束相似关系下的上近似二进制可辨矩阵为：

$$U^C=(u(i,j),k)=\begin{cases}1:\{((x_i,x_j)\in U\times U\wedge\forall b\in C\wedge((b(x_i)=*\\ \wedge b(x_j)\neq *)\vee b(x_i)\neq b(x_j)))\wedge\ \forall d\in D\\ \wedge(d(x_i)\neq d(x_j)))\wedge(\min(\overline{\delta}(x_i),\overline{\delta}(x_j))=1)\}\\ 0\qquad\qquad\qquad\qquad\text{否 则}\end{cases}$$

从上、下近似二进制可辨矩阵的构造可以看出，在矩阵中，某行的元素为 1 表示此元素 1 所对应的条件属性 c_k 在原决策表中能够区分对象 x_i 和 x_j。$\min(\underline{\delta}(x_i),\underline{\delta}(x_j))=1$ 将不一致的对象认为在现有关系和属性条件下不可区分的。这里定义的上、下近似二进制可辨矩阵等价于文献[4]中二进制可辨矩阵，所不同是这里的二进制可辨矩阵适用于不完备信息系统，并且在这种二进制可辨矩阵基础上提出的算法完全适用于不一致性不完备信息系统。

上述构造的上、下近似二进制可辨矩阵具有如下的性质：

定理 1. 在基于约束相似关系的上、下近似二进制可辨矩阵中，当某行的 1 的个数为 1 时，此 1 所对应的属性为原决策表的核属性。

证明：根据约束相似关系构造的上、下近似二进制可辨矩阵知，每行中所含 1 的个数表示的是可以区分该行所对应的对象对的属性个数，每列中 1 的个数表示该列所对应的属性可以区分对象对的个数，因此，行中 1 的个数越少，表示能区分该行对应的对象对的属性越少，说明这些属性越重要。当某行的 1 的个数为 1 时，表明此 1 所对应的属性是唯一能区分该行所对应的对象对的属性，说明该属性是必要的，是核属性。

由此可得到如下命题。

命题 1：在上、下近似二进制可辨矩阵中，1 的个数为 1 的所有行中相应的 1 所对应的属性组成的集合即是原决策表的相对核。

定理 2. 在不完备决策表 S 中，$B\subset C$ 是 C 在约束相似关系下的一个 D 约简，当且仅当由 B 中属性所构成的上、下近似二进制可辨矩阵 M_B 中不全为 0 的行数等于 M_C 中不全 0 的行数，且 $\forall B'\subset B$ 使得 B′不满足此条件。

3 属性约简算法

根据上述上、下近似二进制可辨矩阵定理，这里给出一个简单的求取不完备信息系统属性约简算法，算法以核集为基础，逐步选择属性比较重要的属性加入该集合，直至上、下近似二进制矩阵为空集。由于上近似二进制可辨矩阵和下近似二进制可辨矩阵计算是相同的，这里只讨论基于下近似二进制可辨矩阵的属性约简算法。

算法：基于约束相似关系下的下近似二进制可辨矩阵属性约简算法

输入:不完备决策表 $S=\{U,C,D,V,F\}$;

输出:属性(相对)约简 Reduct;

Step1:构造不完备决策表 S 的下近似二进制可辨矩阵 M;

Step2:计算 M 中的各行和各列中 1 的数目,结果分别存入到一个新的列和新的行中;

Step3:求出 1 的数目为 1 的各行 $\{r_1,r_2,\cdots,r_w\}$ 及 1 所对应的属性 $\{b_1,b_2\cdots,b_w\}$,则决策表的核属性 $Core_D(C)=\{b_1,b_2,\cdots b_w\}$,并令 $Reduct=Core_D(C)$;

Step4:从 M 中消去各行 $\{r_1,r_2,\cdots r_w\}$ 及 1 所对应的各列 $\{b_1,b_2,\cdots,b_w\}$,得到新矩阵 M_i;

Step5:求出 M_i 中 1 的个数最大的一列及其属性 b_i,令 $Reduct=Reduct\cup\{b_i\}$,否则,计算这些列中 1 所对应的各行 1 的数目,取最小的属性 b_j,如果仍然相同,任意选取,并令 $Reduct=Reduct\cup\{b_j\}$;

Step7:从 M_i 中消去属性 b_i、b_j 所对应的列及其中 1 所对应的行,把新矩阵赋给 M_i;

Step8:判断若 M_i 为空集,则终止,输出属性(相对)约简 Reduct;否则转 Step5。

4 实例

为了便于说明,以决策表 1 为例求取属性相对约简,见表 1。根据上述定义,可构造下近似二进制可辨矩阵如表 2。

表 1　决策表

R / U	c_1	c_2	c_3	c_4	c_5	d
X_1	3	2	3	3	2	1
x_2	3	2	3	*	*	1
x_3	3	3	*	*	*	0
x_4	3	*	*	*	*	0
x_5	3	*	*	2	3	1
x_6	2	3	2	2	3	0
x_7	2	3	2	*	*	0
x_8	2	2	*	*	*	1
x_9	2	*	*	*	*	1
x_10	2	*	3	*	2	0

表 2　下近似二进制可辨矩阵

(x_ix_j)	c_1	c_2	c_3	c_4	c_5
(1,4)	0	1	1	1	1
(1,10)	1	1	0	1	0
(2,4)	1	1	0	0	0
(2,6)	1	1	1	0	0
(3,2)	0	1	0	0	0
(3,9)	1	1	0	0	0
(5,4)	0	0	1	1	0
(5,6)	1	0	0	0	0
(5,10)	1	0	0	1	1
(7,5)	1	1	1	0	0
(7,9)	0	1	1	0	0
(10,2)	1	0	0	0	1
(10,5)	1	0	1	0	1
(10,9)	0	0	1	0	1

从表 2 可以看出,第 5、8 行 1 的数目为 1,可直接得出不完备决策表的核属性为这两行的 1 所对应的属性构成的集合 $\{c_1,c_2\}$,利用上述算法可以很方便地求出不完备决策表的属性约简为 $\{c_1,c_2,c_5\}$ 和 (c_1,c_2,c_3,c_4)。

5 结论

不完备信息系统的属性求核和属性约简是粗糙集理论和应用研究的热点之一,本文在约束相似关系模型下,构造了上、下近似二进制可辨矩阵,可方便直接地求取属性核和属性相对约简,实验结果表明所给算法简单有效。约束相似关系下如何有效地定义属性重要度、属性的增量获取以及更有效地求取属性约简还有待进一步的研究

参考文献

1 Grzymala – Busse J W. Data with missing attribute values: generalization of indiscernibility relation and rule induction. Transactions on Rough Sets I, 2004, 78 – 95.

2 张腾飞,王锡淮,肖健梅. 不完备信息系统的一种属性相对约简算法. 计算机工程, 2007, 33(9): 184 – 185.

(下转第 5 页)

(上接第 48 页)

3 Felix R, Ushio T. Rough sets - based machine learning using a binary discernibility matrix. IPMM99 Published,1999:299 - 305.

4 王锡淮,张腾飞,肖健梅. 基于二进制可辨矩阵的决策规则约简算法. 计算机工程与应用,2007,43(27):178 - 180.

5 尹旭日,商琳. 不完备信息系统中 Rough 集的扩充模型. 南京大学学报(自然科学),2006,42(4):337 - 341.