基于知识粒度的不完备决策表的属性约简的矩阵算法

张清国　郑雪峰

(北京科技大学信息工程学院　北京 100083)

摘　要　基于不完备决策表的属性约简定义有多种，现研究基于知识粒度的属性约简。研究发现，差别矩阵是一种较好的设计属性约简算法的方法。为此，定义了一种粒度差别矩阵和基于该差别矩阵的属性约简，并证明了该差别矩阵的属性约简定义与基于知识粒度的属性约简定义等价。在此基础上，设计了一个新的基于信息量的不完备决策表的属性约简算法，其时间复杂度得以降低。

关键词　粗糙集，属性约简，不完备决策表，知识粒度，算法复杂度

中图法分类号　TP182　　**文献标识码**　A

Discernibility Matrix Algorithm of Attribute Reduction Based on Knowledge Granulaion in Incomplete Decision Table

ZHANG Qing-guo　ZHENG Xue-feng

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract　There are many attribute reduction definitions. We researched attribute reduction based on knowledge granuation in incomplete decision table. The discernibility matrix method is a good way to design attribute reduction algorithm. So we gave the definition of discernibility matrix of granulation and the corresponding definition of attribute reduction. At the same time, we proved that the definition is the same as the definition about attribute reduction based on knowledge granulation in incomplete decision table. On this condition, we used the above discernibility matrix of granulation to design an efficient algorithm of attribute reduction based on knowledge granulation in incomplete. Its time complexity is reduced.

Keywords　Rough set, Attribution reduction, Incomplete decision table, Knowledge granulation, Algorithm complexity

1　引言

Pawlak 教授[1,2]提出粗糙集理论以来，它已在许多领域中得到了较为成功的应用。属性约简是粗糙集理论中重要的研究内容。在经典粗糙集理论中，已给出了许多基于决策表的属性约简算法，这些算法的研究一方面丰富了粗糙集理论的研究内容，另一方面为其它一些应用提供了算法基础。然而，等价关系要求太严格。在现实生活中，许多数据难以满足这一要求，特别是针对一些数据不全或者缺失的数据集，更不能满足等价关系。为解决这些数据不全或者缺失的数据集，Kryszkiewicz[3,4]和 Stefanowski[5]教授给出了一种基于相容关系的知识提取方法。还有一些学者将等价关系扩充为相容关系[2-6]。为从不完备的决策表中提取有用的知识，王国胤教授[6]给出了一种基于限制容差关系的扩张的粗糙集模型。Tzung-Pei 教授[7]给出了一种直接从不完备决策表中提取知识的方法。文献[8-10] 给出了一种基于最大相容块的知识获取算法。

然而要从不完备的决策表中有效地提取知识，往往第一步要做的是将不完备的决策表进行属性约简，特别针对大型不完备决策表。由于考虑问题的出发点不同，基于不完备的决策表的属性约简定义也有多种，一些学者也提出了多种相应的不同属性约简算法。文献[11,12]给出了基于不完备信息系统(没有决策属性)的属性约简算法。文献[13-15]以知识粒度为基础，给出了知识粒度的不完备决策表的属性约简算法。特别是文献[13]以知识粒度为启发信息，设计了一个基于知识粒度的不完备决策表的属性约简算法，其时间复杂度均为 $O(|C|^3|U|^2)$。文献[16]中给出了一种基于差别矩阵的属性约简算法。在以往决策表属性约简算法中，差别矩阵是一种常用的设计方法。基于这一思想，定义了基于粒度的差别矩阵以及相应的属性约简，并证明了该定义与基于知识粒度的不完备决策表的属性约简定义等价。在此基础上，设计了一个新的基于知识粒度的不完备决策表的属性约简算法，其时间复杂度降为 $\max\{O(|C|^2|U_{pos}||U|), O(K|U||C|)\}$。最后用一实例说明了新算法的有效性。

2　基本知识

一个决策表定义为 $S=(U,C,D,V,f)$，其中 $U=\{x_1, x_2,\cdots,x_n\}$是对象集，$C=\{c_1,c_2,\cdots,c_r\}$是条件属性集，$D$ 是决

到稿日期：2011-03-02　返修日期：2011-05-27　本文受国家科技基础条件平台项目(2005DKA43600)，国家自然科学基金项目(60674054)资助。

张清国(1966－)，博士生，高级工程师，主要研究领域为数据挖掘、入侵检测，E-mail：qingguo_zhang@sina.com；郑雪峰(1951－)，男，教授，博士生导师，主要研究领域为信息技术安全、网络安全等。

策属性集，$C\cap D=\emptyset$；$V=\bigcup_{a\in C\cup D}V_a$，$V_a$ 是属性 a 的值域，$f:U\times(C\cup D)\to V$ 是信息函数，即 $\forall a\in C\cup D, x\in U, f(x,a)\in V_a$。

不知道决策表的某些条件属性的取值，即其值丢失或部分已知时，称这样的决策表为不完备的决策表。不过，此时决策表的决策表属性值都是精确的。

对于不完备决策表 $S=(U,C,D,V,f)$，任意条件属性子集 $B\subseteq C$，用 $*$ 定义其遗失值。不完备决策表上的相容关系定义 T 如下。

定义 1[3] 不完备决策表 $S=(U,C,D,V,f)$ 上的相容关系定义如下：

$\forall_{(x,y\in U)}T_B(x,y)\Leftrightarrow\forall_{c\in B}(f(x,c)=f(y,c)\vee f(x,c)=*\vee f(y,c)=*)$

显然，T 是自反的、对称的，但不是传递的。对象 x 关于条件属性集 B 的相容类 $S_B(x)$ 定义为 $S_B(x)=\{y|y\in U\wedge T_B(x,y)\}$。

定义 2[13] 在不完备决策表 $S=(U,C,D,V,f)$ 中，知识 $B\subseteq C$ 的知识粒度定义为

$$I(B)=\frac{1}{|U|^2}\sum_{i=1}^{n}|S_B(x_i)|$$

式中，$U=\{x_1,x_2,\cdots,x_n\}$，$|X|$ 表示集合 X 的基数。显然有 $I(\emptyset)=0$。

定义 3[13] 在不完备决策表 $S=(U,C,D,V,f)$ 中，知识 $B(B\subseteq C)$ 是 C 关于 D 的一个相对粒度的属性约简，当且仅当 B 满足条件：(1) $I(B)=I(C)$；(2) $\forall b\in B\Rightarrow I((B-\{b\}))\neq I(C)$。

定义 4 设在一个不完备决策表 $S=(U,C,D,V,f)$ 中，对 $\forall B\subseteq(C\cup D)$，定义论域 U 的一个覆盖 $U//S_B$：$U//S_B=\{S_B(x)|x\in U\}$。

定义 5 设 $S=(U,C,D,V,f)$ 是一个不完备决策表，$\forall P,Q\subseteq(C\cup D)$，设 $Q\subseteq P$，记 $U//S_P=\{S_1^P,S_2^P,\cdots,S_t^P\}$，$U//S_Q=\{S_1^Q,S_2^Q,\cdots,S_t^Q\}$，若 $\forall S_i^P\in U//S_P\Rightarrow\exists S_j^Q\in U//S_Q$ 使 $S_i^P\subseteq S_j^Q$，则称 $U//S_p$ 为 $U//S_Q$ 的加细，记为 $U//S_p\preceq U//S_Q$。

定义 6[3] 设在一个不完备决策表 $S=(U,C,D,V,f)$ 中，$X\subseteq U$，$\forall B\subseteq C\cup D$，记 $\underline{B}(X)=\{x\in U|S_B(x)\subseteq X\}$ 为 X 关于 B 的下近似集，$\overline{B}(X)=\{x\in U|S_B(x)\cap X\neq\emptyset\}$ 为 X 关于 B 的上近似集。

定义 7 设在一个不完备决策表 $S=(U,C,D,V,f)$ 中，$\forall B\subseteq C$，$U/D=\{D_1,D_2,\cdots,D_k\}$ 表示由决策属性集 D 对论域 U 的划分，称 $POS_C(D)=\bigcup_{D_i\in U/D}C\_(D_i)$ 为 C 关于 D 的正区域，简记为 U_{pos}，$U_{neg}=U-U_{pos}$。

定义 8 设在一个不完备决策表 $S=(U,C,D,V,f)$ 中，$U=U_{pos}\cup U_{neg}$，定义粒度差别矩阵 $M=(m(i,j))$，其元素定义如下：

$$m(i,j)=\begin{cases}\{c_k|c_k\in C,f(x_i,c_k)\neq *\wedge f(x_j,c_k)\neq\\ \quad *\wedge f(x_i,c_k)\neq f(x_j,c_k),f(x_i,D)\neq f(x_j,D)\\ \quad 且\ x_i\ 和\ x_j\ 一个在\ U_{pos}，一个在\ U_{neg}\ 中；\\ f(x_i,c_k)\neq *\wedge f(x_j,c_k)\neq *\wedge f(x_i,c_k)\neq\\ \quad f(x_j,c_k)且\ x_i,x_j\ 在\ U_{pos}\ 中\}\\ \emptyset，否则\end{cases}$$

式中，$k=1,2,\cdots,r$。

定义 9 设 $M=(m(i,j))$ 为不完备决策表 $S=(U,C,D,V,f)$ 的粒度差别矩阵，$\forall B\subseteq C$，若 B 满足 (1) $\forall\emptyset\neq m(i,j)\in M$，有 $B\cap m(i,j)\neq\emptyset$；(2) $\forall a\in B$，$B'=B-\{a\}$ 均不满足 (1)，则称 B 是 C 关于 D 的一个属性约简（基于粒度差别矩阵的属性约简的定义）。

3 相关定理

定理 1 设在一个不完备决策表 $S=(U,C,D,V,f)$ 中，$\forall P,Q\subseteq(C\cup D)$，设 $Q\subseteq P$，则有 $U//S_p\preceq U//S_Q$。

证明：设 $U//S_p=\{S_1^P,S_2^P,\cdots,S_t^P\}$，$U//S_Q=\{S_1^Q,S_2^Q,\cdots,S_s^Q\}$，任取 $S_i^P\in U//S_P$，由于 $Q\subseteq P$，则有 $S_i^P=S_P(x)=\{y|\forall a\in P,f(x,a)=f(y,a)\vee f(x,a)=*\vee f(y,a)=*\}\subseteq S_j^Q=S_Q(x)=\{y|\forall a\in Q,f(x,a)=f(y,a)\vee f(x,a)=*\vee f(y,a)=*\}$。

由 S_i^P 的任意性知，$U//S_p\preceq U//S_Q$。

定理 2[13] 不完备决策表 $S=(U,C,D,V,f)$ 中，$\forall P,Q\subseteq(C\cup D)$，若 $\forall x\in U$，$S_P(x)=S_Q(x)$，则有 $I(P)=I(Q)$。

定理 3 设 $M=(m(i,j))$ 为不完备决策表 $S=(U,C,D,V,f)$ 的粒度差别矩阵，$\forall B\subseteq C$，若 B 满足 $\forall\emptyset\neq m(i,j)\in M$，有 $B\cap m(i,j)\neq\emptyset$，则有 $I(B)=I(C)$。

证明：假设有 $I(B)\neq I(C)$，则至少存在 $x_i\in U'$，使得 $|S_B(x_i)|\neq|S_C(x_i)|$，则一定存在 $x_j\in S_B(x_i)\wedge x_j\notin S_C(x_i)$，且 $c_k\in C-B$，使得 $f(x_i,c_k)\neq *\wedge f(x_j,c_k)\neq *\wedge f(x_i,c_k)\neq f(x_j,c_k)$。当 $x_i,x_j\in U_{pos}$ 时，由定义 8 有 $c_k\in m(i,j)\Rightarrow m(i,j)\neq\emptyset$。另一方面，由于 $x_j\in S_B(x_i)$，故 $B\cap m(i,j)=\emptyset$，这与条件矛盾，故假设不成立，从而有 $I(B|D)=I(C|D)$。当 x_i,x_j 只有一个在 U_{pos}，设 $x_i\in U_{pos}$，则有 $x_j\in U_{neg}$。当 $f(x_i,D)\neq f(x_j,D)$，则由定义 8 有 $c_k\in m(i,j)\Rightarrow m(i,j)\neq\emptyset$。另外，由于 $x_j\in S_B(x_i)$，故 $B\cap m(i,j)=\emptyset$，这与条件矛盾。若 $f(x_i,D)=f(x_j,D)$，由于有 $x_j\in U_{neg}$，故一定存在 x_r，使得 $f(x_j,D)\neq f(x_r,D)$，从而有 $f(x_i,D)=f(x_j,D)\neq f(x_r,D)$，则由定义 8 有 $c_k\in m(i,j)\Rightarrow m(i,j)\neq\emptyset$。另一方面，由于 $x_j\in S_B(x_i)$，故 $B\cap m(i,j)=\emptyset$，这与条件矛盾。综上所述，有 $I(B)=I(C)$。

定理 4 设 $M=(m(i,j))$ 为不完备决策表 $S=(U,C,D,V,f)$ 的差别矩阵，$B\subseteq C$，若 $I(B)=I(C)$，则有 $\forall\emptyset\neq m(i,j)\in M$，$B\cap m(i,j)\neq\emptyset$。

证明：假设存在 $\emptyset\neq m(i,j)\in M$，有 $B\cap m(i,j)=\emptyset$，则存在属性 $c_h\in C\wedge c_h\notin B$，使得 $c_h\in m(i,j)$，故有 $S_c(x_i)\neq S_c(x_j)$。由于 $B\cap m(i,j)=\emptyset$，则由定义 8 可知 $S_B(x_i)=S_B(x_j)$。另一方面，有 $S_C(x_i)\subseteq S_B(x_i)\wedge T_C(x_j)\subseteq S_B(x_j)$，故有 $S_C(x_i)\cup S_C(x_j)\subseteq S_B(x_i)\cup S_B(x_j)=S_B(x_i)$，从而有 $S_C(x_i)\neq S_B(x_i)$，进而 $I(B)\neq I(C)$，这与条件矛盾，故假设不成立。综上所述，命题成立。

定理 5 基于相对粒度的属性约简定义与基于粒度差别矩阵的属性约简定义等价。

证明：在不完备决策表 $S=(U,C,D,V,f)$ 中，设基于相对粒度的所有属性约简的集合为 $RGReduc$，基于粒度差别矩阵的所有属性约简的集合为 $GDisReduc$。任取 $P\in GDisReduc$，则有 $\forall\emptyset\neq m(i,j)\in M$，有 $P\cap m(i,j)\neq\emptyset$，由定理 2 知，$I(P|D)=I(C|D)$。由于 $P\in GDisReduc$，则 $\forall c_h\in P$，有 $B=P-\{c_h\}$，至少存在 $\forall\emptyset\neq m(i_0,j_0)\in M$，有 $P\cap m(i_0,j_0)=\emptyset$，由定理 3 的逆否命题知，$\forall c_h\in P$ 有 $I(P-\{c_h\}|D)\neq I(C|D)$。故有 $P\in RGReduc$。由 P 的任意性知 $GDisReduc\subseteq RGReduc$。

任取 $P \in RGReduc$，则有 $I(P|D)=I(C|D)$，由定理 3 有 $\forall \emptyset \neq m(i,j) \in M$，且 $P \cap m(i,j) \neq \emptyset$。由 $P \in RGReduc$，则有 $\forall c_h \in P, POS_{P-\{c_h\}}(D) \neq POS_C(D)$，由定理 2 的逆否命题知，由 $P-\{c_h\}$ 不满足定义 9 的条件(1)，有 $P \in GDisReduc$。由 P 的任意性知 $GDisReduc \supseteq RGReduc$。

综上所述，命题成立。

定理 4 说明基于相对粒度的不完备决策表属性约简定义与基于粒度差别矩阵的属性约简定义是等价的，这样就可以将基于相对粒度的属性约简建立在粒度差别矩阵上，这一定理为下一步设计新的快速算法提供了理论依据。

为给出新的属性约简算法，首先分析文献[13]中的属性约简算法，并称之为老算法。为给出粒度差别矩阵算法，先给出如下的计算正区域的算法。

4 计算正区域的方法

求正区域的计算时间主要花在计算容差类 $T_C(x)(x \in U)$ 上。一般来说，求容差类 $T_C(x)$ 的算法是：对对象集 U 中的对象进行两两比较，比较它们在 C 中的每个属性是否满足容差类的定义。若满足，则属于同一个容差类。或者对对象集 U 中的每个对象，根据其 C 的取值判断是否属于现有的容差类。在最坏的情况下，以上两种方法在每个条件属性下都需要 $O(|U|)^2$ 次比较，故最坏的时间复杂度为 $O(|C||U|^2)$[5-11]。现根据定理 1 和容差关系的对称性，将计算容差类 $T_C(x)$ 的算法时间复杂度降为 $O(K|U||C|)$，其中 $K=\max\{|T_C(x_i)|, x_i \in U\}$。因为 $T_C(x_i) \subseteq U$，所以 $O(K) \leqslant O(|U|)$。显然，该时间复杂度比一般的算法时间复杂度 $O(|C||U|^2)$ 要低。算法描述如下。

算法 1 计算容差类 $T_{c_i}(x)(x \in U, c_i \in C)$

输入：不完备决策表 $S=(U,C,D,V,f)$，$U=\{x_1,x_2,\cdots,x_n\}$；

输出：容差类 $T_{c_i}(x)$

(1) 对条件属性 c_i 统计 $f(x_j,c_i)(j=1,2,\cdots,n)$ 的最大值、最小值和有无缺省值，分别记为 M_i、m_i 和 c_i^*；//$c_i^*=1$ 表示有缺省值，$c_i^*=0$ 表示无缺省值。

(2) "分配"：如果 $c_i^*=1$，建立 $A=M_i-m_i+2$ 个空队列，否则建立 $A=M_i-m_i+1$ 个空队列，并令 $front_k$ 和 end_k($k=0,1,2,\cdots,M_i-m_i$ 或 M_i-m_i+1) 分别为第 k 个队列的头指针和尾指针，$front_*$ 和 end_* 分别为最后一个队列(＊队列)的头指针和尾指针(若 $c_i^*=1$)。将链表中的对象分配到对应的 $f(x,c_i)-m_i$ 个队列中，或分配到最后队列(＊队列)中去。

(3) "收集"：表头指针指向第一个非空队列的头指针，修改每一个非空队列的尾指针，令其指向下一个非空队列的队头对象，将 A 个队列重新组成一个链表。

(4) 设由(3)得到的链表中的对象序列为 $x_1', x_2', \cdots, x_n'$；

$t=1; E_1=\{x_1'\}$；
for($j=2; j<=n; j++$)
if($f(x_j',c_i)=f(x_{j-1}',c_i)$)
{ $E_t=E_t \cup \{x_j'\}$; }
else
{ $t=t+1$; $E_t=\{x_j'\}$}。

(5) 如果 $c_i^*=1$，则做如下处理：

for($v=1; v<t; v++$)
{$T_v=E_v \cup E_t$;
$T_t=E_t \cup E_v(v=1,2,\cdots,t-1)$;}
否则
for($v=1; v<=t; v++$)
{ $T_v=E_v$; }。

算法 2 计算正区域算法

输入：不完备决策表 $S=(U,C,D,V,f)$，$U=\{x_1,x_2,\cdots,x_n\}$，$C=\{c_1,c_2,\cdots,c_r\}$，$T_{c_i}(x)$

输出：$T_C(x)$，U_{pos}，U_{neg}。

(1) $T_C(x)=\emptyset, U_{pos}=\emptyset, U_{neg}=\emptyset, C=\{c_i\}$；

(2) for ($j=1; j<r; j++$) // $j \neq i$
if($f(x,c_j)==*$)
$T_{C\cup\{c_j\}}(x)=T_C(x)$;
else $\forall y \in T_C(x)$
if($f(y,c_j)==* \vee f(y,c_j)==f(x,c_j)$)
$T_{C\cup\{c_j\}}(x)=T_{C\cup\{c_j\}}(x) \cup \{y\}$

(3) $j=j+1$；若 $j>r$，转向(4)；否则 $C=C\cup\{c_j\}$，转向(2)。

(4) 终止算法，输出 $T_C(x)$。

如果 $T_C(x)$ 容差类中的所有对象在决策属性上取值均相同，则取出 x 并入 U_{pos} 中，否则取出 x 并入 U_{neg} 中。

算法的复杂度分析：算法 1 第(1)步的时间复杂度为 $O(|U|)$，第(2)步的时间复杂度为 $O(|U|)$，第(3)步的时间复杂度为 $O(|A|)<O(|U|)$，第(4)步的时间复杂度为 $O(|U|)$，第(5)步的时间复杂度为 $O(|t|)<O(|U|)$，所以算法 1 的时间复杂度为 $O(|U|)$。

算法 2 的时间复杂度为 $O(K|U||C|)$，其中 $K=\max\{|T_C(x_i)|, x_i \in U\}$，因为 $T_C(x_i) \subseteq U$，所以 $O(K) \leqslant O(|U|)$。算法 2 计算正区域的时间复杂度比一般算法的时间复杂度 $O(|U|^2|C|)$ 要低。

5 属性约简算法

算法 3 基于正区域的不完备决策表的属性约简算法

输入：不完备决策表 $S=(U,C,D,V,f)$，$U=\{x_1,x_2,\cdots,x_n\}$，$C=\{c_1,c_2,\cdots,c_n\}$

输出：不完备决策表的属性约简 $reduce(C)$

(1) 由算法 2 求出 $U_{pos}=\{y_1,y_2,\cdots,y_s\}$，$U_{neg}=\{z_1,z_2,\cdots,z_t\}$；

(2) $reduce(C)=\emptyset; T=\emptyset$;

(3) for($i=1; i<s; i++$)
for($j=i+1; j<s+1; j++$)
if($f(y_i,D) \neq f(y_j,D)$)
{$m(i,j)=\emptyset$;
for($k=1; k<r+1; k++$)
if($f(y_i,c_k) \neq * \wedge f(y_j,c_k) \neq * \wedge f(y_i,c_k) \neq f(y_j,c_k)$)
$m(i,j)=m(i,j) \cup \{c_k\}$;
}

(4) for($i=1; i<s+1; i++$)
for($j=1; j<t+1; j++$)
{$m(i,j)=\emptyset$;
for($k=1; k<r+1; k++$)
if($f(y_i,c_k) \neq * \wedge f(z_j,c_k) \neq * \wedge f(y_i,c_k) \neq f(z_j,c_k)$
$\wedge f(y_i,D) \neq f(z_j,D)$)
$m(i,j)=m(i,j) \cup \{c_k\}$;
}

(5) while($M=m(i,j)$! $=\emptyset$)
{①对矩阵的每一元素判断：若该元素只有一个属性 a，则将该属性 a 并入 $reduce(C)$，并将包含属性 a 所有矩阵元素去掉；
②在 $C-T-reduce(C)$ 中任取一属性 b 并入 T，在所有包含属性 b 的矩阵元素中去掉属性 b；
}

算法的复杂度分析：由算法 1 知，算法 3 第(1)步的时间复杂度为 $O(K|U||C|)$，第(3)步和第(4)步的时间复杂度为 $O(|C||U_{pos}|(|U_{pos}|+|U_{neg}|))=O(|C||U_{pos}||U|)$。第(5)

(下转第 243 页)

sion making[J]. International Journal of Approximate Reasoning,2008,48(1):246-262
[9] Lin Y,Lee P,Ting H. Dynamic multi-attribute decision making model with grey number evaluations[J]. Expert Systems with Applications,2008,35(4):1638-1644
[10] 苏志欣,王理,夏国平. 区间数动态多属性决策的 VIKOR 扩展方法[J]. 控制与决策,2010,25(6):836-840
[11] Opricovic S. Multicriteria optimization of civil engineering systems[D]. Belgrade:Faculty of Civil,1998
[12] Opricovic S, Tzeng G. Compromise solution by MCDM methods:A comparative analysis of VIKOR and TOPSIS[J]. European Journal of Operational Research,2004,156(2):445-455
[13] Opricovic S,Tzeng G. Extended VIKOR method in comparison with outranking methods[J]. European Journal of Operational Research,2007,178(1):514-529
[14] 徐泽水. 直觉模糊信息集成理论及应用[M]. 北京:科学出版社,2008
[15] Tong L, Chen C, Wang C. Optimization of multi-response processes using the VIKOR method[J]. International Journal of Advanced Manufacturing Technology,2007,31(11):1049-1057

(上接第 211 页)

步每一次循环的最坏时间复杂度为 $O(|C||U_{pos}||U|)$,最多循环 $|C|-1$ 次,故第(5)步最坏的时间复杂度为 $O(|C|^2|U_{pos}||U|)$。故新属性约简算法的最坏时间复杂度为 $\max\{O(|C|^2|U_{pos}||U|),O(K|U||C|)\}$,其中 $K=\max\{|T_C(x_i)|,x_i\in U\}$。故求新算法的最坏空间复杂度为 $\max\{O(|C||U_{pos}||U|),O|U|\}$。

为更好地说明新算法的有效性,以文献[3]中的不完备决策表 1 为例进行分析说明。

表 1 不完备决策表

car	price	mileage	size	Max-speed	conclusion
x1	high	high	full	low	good
x2	low	*	full	low	good
x3	*	*	compact	high	poor
x4	high	*	full	high	good
x5	*	*	full	high	excel
x6	low	high	full	*	good

对不完备决策表 1 的 6 个对象,通过算法 1 和算法 2 计算可得

$T_C(x_i)(i=1,2,\cdots,6)$的结果如下:

$T_C(x_1)=\{x_1\}$;$T_C(x_2)=\{x_2,x_6\}$

$T_C(x_3)=\{x_3\}$;$T_C(x_4)=\{x_4,x_5\}$

$T_C(x_5)=\{x_5,x_4,x_6\}$

$T_C(x_6)=\{x_6,x_2,x_5\}$

$U/D=\{\{x_1,x_2,x_4,x_6\},\{x_3\},\{x_5\}\}$

所以,$U_{pos}=\{x_1,x_2,x_3\}$,$U_{neg}=\{x_4,x_5,x_6\}$。

由算法 3 的第(3)步和第(4)步得到的差别矩阵见表 2。

表 2 表 1 的粒度差别矩阵

	X1	X2	X3
X1	Ø	{price}	{size, max-speed}
X2	{price}	Ø	{size, max-speed}
X3	{size, max-speed}	{size, max-speed}	Ø
X4	Ø	Ø	{size}
X5	{max-speed}	{max-speed}	{size}
X6	Ø	Ø	{size}

由算法的第(5)步可知,由第一次循环的第①步得到 $reduce(C)=\{size,price,\max\text{-}speed\}$,将包含在 $reduce(C)$ 中的矩阵元素去掉后,矩阵为空,算法结束。所以属性约简为$\{size,price,\max-speed\}$,该约简与文献[13]中相同。

结束语 在决策表中,差别矩阵是一种常用的属性约简设计方法。经研究,给出了基于粒度的差别矩阵及其相应的属性约简定义,同时证明了该定义与基于短程粒度的不完备决策表属性约简算法定义等价。在此基础上,给出了知识粒度的不完备决策表的属性约简的差别矩阵算法。

参考文献

[1] Pawlak Z,Grzymala-Busse J,Slowinski R. Rough sets[J]. Communications of the ACM,1995,8(1):89- 95
[2] Pawlak Z. Rough set theory and its application to data analysis [J]. Cybernetics and Systems,1998,9(4):661-668
[3] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Science,1998,112(1):39-49
[4] Kryszkiewicz M. Rules in incomplete information systems[J]. Information Sciences,1999,113(2):271-292
[5] Stefanowski J, Tsoukias A. Incomplete information tables and rough classification[J]. Computational Intelligence, 2001, 17(4):545-566
[6] 王国胤. 不完备信息系统的粗糙集扩展[J]. 计算机研究与发展,2002,39(8):1238-1243
[7] Tzung-Pei H. Learning rules from incomplete training examples by rough sets[J]. Expert Systems with Application, 2002, 22(2):285-293
[8] Lenng Y,Li D Y. Maximal consistent block technique for rule acquisition in incomplete information systems [J]. Information Sciences,2003,153(1):95-106
[9] Wu W Z, Mi J S, Zhang W X. A new rough set approach to knowledge discovery in incomplete information systems[J]. ICMLC,2003:1713-1718
[10] Lenng Y,Wu W Z,Zhang W X. Knowledge acquisition in incomplete information systems: a rough set approach[J]. European Journal of Operational Research,2006,168(2):164-180
[11] Liang J Y,Xu Z B. An algorithm for knowledge in incomplete information systems[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems,2002,10(1):95-103
[12] 黄兵,周献中,张蓉蓉. 基于信息量的不完备信息系统属性约简[J]. 系统工程理论与实践,2005,25(4):55-60
[13] 李秀红,史开泉. 一种基于知识粒度的不完备信息系统的属性约简算法[J]. 计算机科学,2006,33(11):169-171
[14] 徐久成,史进玲,孙林. 一种基于相对粒度的决策表约简算法[J]. 计算机科学,2009,36(3):205 207
[15] 史先红,史进玲. 一种基于相对粒度的不完备决策表约简算法[J]. 河南师范大学学报:自然科学版,2010,38(4):51-54
[16] Huang B,He X,Zhou X Z. Rough Computational methods based on tolerance matrix[J]. Acta Automatica Sinica,2004,30(2):363-370