DOI:10.11992/tis.201506029
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20151229.0837.020.html

基于知识粒度的不完备决策表的属性约简算法

乔丽娟[1,2],徐章艳[1,2],谢小军[1,2],朱金虎[1,2],陈晓飞[2],李娟[2]
(1.广西师范大学 广西多源信息挖掘与安全重点实验室,广西 桂林 541004; 2.广西师范大学 计算机科学与信息工程学院,广西 桂林 541004)

摘 要:知识粒度是属性约简的有效方法,但对于大型的决策表,计算知识粒度过于费时,算法效率不高。在引入粒度差别矩阵后,设计了一个计算粒度差别矩阵中条件属性出现频率的函数,有效地降低粒度差别矩阵的存储空间,根据此函数设计了一个高效属性约简算法。新算法使得时间复杂度与空间复杂度都降为 $O(K|C||U|)$(其中 $K=\max\{|Tc(x_i)|, x_i\in U\}$ 和 $O(|U|)$。最后通过实例仿真说明了此算法的高效性和可行性。

关键词:属性约简;知识粒度;不完全决策表;条件属性频率;差别矩阵;启发信息
中图分类号:TP18 **文献标志码**:A **文章编号**:1673-4785(2016)01-0129-07

中文引用格式:乔丽娟,徐章艳,谢小军,等.基于知识粒度的不完备决策表的属性约简算法[J]. 智能系统学报, 2016, 11(1): 129-135.
英文引用格式:QIAO Lijuan, XU Zhangyan, XIE Xiaojun, et al. Efficient attribute reduction algorithm for an incomplete decision table based on knowledge granulation[J]. CAAI Transactions on Intelligent Systems, 2016, 11(1): 129-135.

Efficient attribute reduction algorithm for an incomplete decision table based on knowledge granulation

QIAO Lijuan[1,2], XU Zhangyan[1,2], XIE Xiaojun[1,2], ZHU Jinhu[1,2], CHEN Xiaofei[2], LI Juan[2]
(1.Guangxi Key Laboratory of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, China; 2. College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)

Abstract: The use of knowledge granularity is an effective attribute reduction approach. But for a large decision table, computing knowledge granularity is so time-consuming that the algorithm is not efficient for practical use. After the introduction of the discernibility matrix of granularity, a function was designed for calculating the occurrence frequency of condition attributes in the matrix. In this paper, we design an efficient attribute reduction algorithm based on the granularity discernibility matrix. The new algorithm reduces the time and space complexities to $O(K|C||U|)$ ($K=\max\{|Tc(x_i)|, x_i\in U\}$) and $O(|U|)$, respectively. The results from our simulation example verify that the proposed algorithm is feasible and highly efficient.

Keywords: attribute reduction; knowledge granularity; incomplete decision table; condition attribute frequency; discernibility matrix; heuristic information

波兰的数学家 Pawlak 在 20 世纪 80 年代提出的粗糙集是一种新型的用来处理不完全、不精确与不相容的数学工具和理论[1-2]。经过了 30 多年的研究和发展,粗糙集理论已在知识发现、数据挖掘、模式识别等领域得到了大量应用[3-4]。属性约简作为粗糙集理论的重要研究内容,已被广大学者所研究,提出了围绕完备决策表的属性约简算法,但是现实生活中的数据往往存在误差,缺失及多源等特征。如何对不完备决策表进行直接处理,已成为粗糙集理论的一个研究热点[4]。近年来针对不完备决策表的研究也取得了显著的进步,已有学者提出很多有效的不完备决策表属性约简算法[5-11]。知识粒度[12-13]作为粗糙集理论中度量属性约简的重要方法之一,被广泛运用于不完备属性约简算法。文献

收稿日期:2015-06-16. 网络出版日期:2015-12-29.
基金项目:国家自然科学基金资助项目(61262004,61363034,60963008);广西自然科学基金资助项目(2011GXNSFA018163);大学生创新资助项目(201410602099).
通信作者:乔丽娟. E-mail:347671379@qq.com.

[5]以属性重要性为启发信息,设计了一个基于知识粒度的属性约简算法[5];文献[6]通过不断向核属性集中添加属性的方法,设计出一种基于相对知识粒度的不完备决策表属性约简算法[6];文献[7]定义了一个粒度差别矩阵,进而设计了基于知识粒度的不完备决策表的属性约简算法[7],其时间复杂度为 $\max\{O(|C|^2|U||U_{\text{pos}}|),O(|K||C||U|)\}$,其中 $K=\max\{|T_C(x_i)|,x_i\in U\}$,其空间复杂度为 $\max\{O(|C||U||U_{\text{pos}}|),O(|U|)\}$;文献[8]给出了一个计算条件属性频率的公式,设计一个基于知识粒度的属性约简算法[8];文献[9] 设计了一种基于对象矩阵的属性约简算法[9];文献[11]提出简化差别矩阵定义,设计了一种快速的属性约简算法[11];文献[12]中根据区分对象对集的思想,设计了基于正区域的属性约简算法[12];文献[13]根据粒计算的思想构建了粒矩阵,在此基础上,设计了属性约简算法。文献[14]在粒计算属性约简算法的基础上进行了改进,得到一个新的算法。上述算法大多因为要多次计算知识粒度,导致计算效率都不太理想,为此设计出基于知识粒度的高效属性约简算法具有非常重要的现实意义[5]。

差别矩阵作为粗糙集理论的重要技术之一,被广泛应用,但是求解差别矩阵费时,本文引入了基于粒度的差别矩阵,利用条件属性在区分对象时出现频率的属性约简思想,设计一个基于粒度差别矩阵计算属性频率的启发函数。

1 粗糙集基本概念

定义 1[3] 五元组 $S=(U,C,D,V,f)$ 是一个不完备决策表,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 表示对象的非空有限集合,称为论域;$C=\{c_1,c_2,\cdots,c_m\}$ 表示条件属性的非空有限集合;D 表示决策属性的非空有限集合,且 $C\cap D=\varnothing$;$V=\bigcup\limits_{a\in C\cup D}V_a$,$V_a$ 是属性 a 的值域;$f:U\times C\cup D\to V$ 是一个信息函数,它对一个对象的每一个属性赋予一个信息值,即 $\forall a\in C\cup D,x\in U$,有 $f(x,a)\in V_a$。

在五元组中,如果至少有一个属性 $a\in C$,使得 V_a 包含空值(用 * 表示),即至少有一个属性 $a\in U$,存在一个 $a\in U$,使得 $f(x,a)=*$,称之为不完备决策表。

定义 2[3] 在不完备决策表 $s=(U,C,D,V,f)$ 中,令 $B\subseteq C$,定义 U 上的容差关系 $T(B)$ 为 $T(B)=\{(x,y)\in U\times U|\forall b\in B\}$,$f(x,b)=f(y,b)\vee f(x,b)=*\vee f(y,b)=*\}$。用 $T_B(x)$ 表示在 B 中与 x 具有容差关系的全体对象集 $\{y\in U|(x,y)\in T(B)\}$。

定义 3[16] 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 $GD(B)=\dfrac{1}{|U|^2}\sum\limits_{i=1}^{n}|T_B(x_i)|$. 其中 $U=\{x_1,x_2,\cdots,x_n\}$,$|X|$ 表示集合 X 的基数. 显然有 $CD(\varnothing)=0$。

性质 1[16] 设 $S=(U,C,D,V,f)$ 是一个不完备信息系统,知识 $B\subseteq C$ 的知识粒度定义为 $GD(B)$,则 $1/|U|\leqslant GD(B)\leqslant 1$。

性质 2[16] 设 $S=(U,C,D,V,f)$ 是一个不完备信息系统,其中 $P,Q\subseteq C$,如果 $\forall i\in\{1,2,\cdots,|U|\}$ 有 $T_P(x_i)\subseteq T_Q(x_i)$,则 $GD(P)\leqslant GD(Q)$。

知识粒度可以描述知识的区分能力,知识粒度越小,其区分能力越强,反之区分能力越弱[5]。

定义 4[5] 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 的一个知识粒度的属性约简,当且仅当 B 满足条件:

1) $GD(B)=GD(C)$;

2) $\forall b\in B\Rightarrow GD((B-\{b\}))\neq GD(C)$。

定义 5[7] 在不完备决策表 $S=(U,C,D,V,f)$ 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 $\text{POS}_C(D)=\bigcup\limits_{D_i\in U/D}C\_(D_i)$ 为 C 关于 D 的正区域,设条件属性对论域的划分为 $U/C=\{[x_{i1}]_c,[x_{i2}]_c,\cdots,[x_{ik}]_c\}$,$U_{\text{pos}}=\{x_{ij}|[x_{ij}]_c\subseteq \text{POS}_C(D)\}$,$U_{\text{neg}}=U-U_{\text{pos}}$。

2 粒度差别矩阵相关概念

定义 6[11] 设在一个不完备决策表 $S=(U,C,D,V,f)$ 中,$U=U_{\text{pos}}\cup U_{\text{neg}}$,定义粒度差别矩阵 $\boldsymbol{M}=(m(i,j))$,其元素定义如下:

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

式中:$k=1,2,\cdots,r$。

定义 7[7] 设 $\boldsymbol{M}=(m(i,j))$ 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) $\forall\varnothing\neq m(i,j)\in\boldsymbol{M}$,有 $B\cap m(i,j)\neq\varnothing$;

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,此约简记为基于粒度差别矩阵的属性约简。

定理 1 在不完备决策表 $S=(U,C,D,V,f)$ 中,有 $R_C=\bigcup\limits_{a\in C}R_{|a|}$。

证明　由定义 1 知:命题显然成立。

定理 2[7]　基于知识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是等价的。

定理 2 说明基于知识粒度的属性约简可以转化到粒度差别矩阵上进行。

针对不完备决策表,文献[7]中给出了一个基于粒度差别矩阵的属性约简算法,其时间复杂度为 $\max\{O(|C|^2|U_{\text{pos}}||U|),O(K|U||C|)\}$。算法对粒度差别矩阵进行遍历,若只包含一个条件属性就将其放入属性约简中,并去掉差别矩阵中任何含有该条件属性的差别元素,直至差别矩阵为空。该算法虽然有效降低了时间复杂度,但是构造粒度差别矩阵仍然需要占用大量的空间,对于处理大型数据集仍然具有一定的难度。

经分析,算法中在粒度差别矩阵中出现的条件属性才是能区分对象的条件属性,由于构造粒度差别矩阵耗费空间,参考文献[16]的方法,设计一种计算粒度差别矩阵中含有的条件属性频率的函数,然后给出计算该函数的快速算法,无须构造粒度差别矩阵就可以将其中能有效区分对象的条件属性找出,以降低算法的时间和空间复杂度。

3　计算属性频率的启发函数

定理 3　在决策表 $S=(U,C,D,V,f)$ 中,$B\subseteq C$,$U/B=\{A_1,A_2,\cdots,A_l\}$,$A_i/\{a\}=\{A_{i1},A_{i2},\cdots,A_{ik}\}$,$A_{ij}=\text{pos}_i\cup\text{Neg}_j$,$U=U_{\text{pos}}\cup U_{\text{neg}}$,其中 $\text{pos}_i=A_{ij}\cap U_{\text{pos}}$,$\text{Neg}_j=A_{ij}\cap U_{\text{neg}}$,$\text{pos}_i/D=\{D_{i1},D_{i2},\cdots,D_{iD}\}$,$\text{Neg}_j/D=\{\bar{D}_{j1},\bar{D}_{j2},\cdots,\bar{D}_{jD}\}$。令 $s_i=|\text{pos}_i/D|=\sum\limits_{1\leqslant j\leqslant|D|}|D_{ij}|=|\text{pos}_i|$,则所有集合中属于正域的集合对 D 划分 pos_i/D 总和为 $S=\sum\limits_{1\leqslant i\leqslant k}S=\sum\limits_{1\leqslant i\leqslant k}|\text{pos}_i|$,所有集合中属于正域的所有集合对 D 划分 pos_i/D 中决策值相同集合总数为 $T_j=\sum\limits_{1\leqslant i\leqslant k}D_{ij}$。

根据定义 6,粒度差别矩阵中包含的条件属性可由两部分产生,设对象都在 U_{pos} 里产生的条件属性的个数为 N_1,则

$$N_1=\sum_{1\leqslant i<j\leqslant k}\text{pos}_i\text{pos}_j \tag{1}$$

两个对象一个在 U_{pos} 中,另一个在 U_{neg} 中,产生的条件属性频率为 N_2,则

$$N_2=\sum_{1\leqslant i\leqslant k,1\leqslant j\leqslant|D|}\bar{D}_{ij}(S-S_i-T_j+D_{ij}) \tag{2}$$

计算条件属性的频率函数 $|F_B(U,a)|$ 如下:

$$F_B(U,a)=\sum_{1\leqslant i\leqslant l}(2N_1+N_2),$$

即
$$F_B(U,a)=\sum_{1\leqslant i\leqslant l}\Big(2\sum_{1\leqslant i\leqslant j\leqslant k}\text{Pos}_i\text{Pos}_j+\sum_{1\leqslant i\leqslant k,1\leqslant j\leqslant|D|}\bar{D}_{ij}(S-S_i-T_j+D_{ij})\Big) \tag{3}$$

证明　由粒度差别矩阵的定义知,计算 $A_i/\{a\}=\{A_{i1},A_{i2},\cdots,A_{ik}\}$ 产生的条件属性频率,可分两部分计算,一种是对象都在 U_{pos} 中;另一种是一个在 U_{pos} 中,而另一个在 U_{neg} 中的。

1)若两个对象都在 U_{pos} 中,由划分的定义知,在同一个划分集合里的两个对象值相等,即只有不同划分集合里才有可能产生有效区分对象的条件属性。则只有不同划分集合的 U_{pos} 之间才能产生条件属性频率;若两个对象都在 U_{pos} 中,产生的条件属性频率为 $N_1=\sum\limits_{1\leqslant i<j\leqslant k}\text{pos}_i\text{pos}_j$,任意两个划分集合都可产生,因为在正域之间产生的差别矩阵的元素是对称的,故条件属性频率为 $2N_1$。

2)若一个对象在 U_{pos} 中,另一个对象在 U_{neg} 中,由划分的定义知,同属一个集合里的两个对象值相等,即只有不同划分集合里才有可能产生条件属性频率,且 U_{pos} 和 U_{neg} 之间要求决策值不同,故需要对每个划分集合里属于 U_{pos} 的集合对 D 划分,同时属于 U_{neg} 的集合也对 D 划分。所以,Neg_j/D 划分集合里每个集合与 pos_i/D 划分集合里对于决策属性在不同划分集合里就能产生条件属性频率。

为了方便叙述,假设将 $A_i\{b\}$ 所有集合中属于正域的所有集合对 D 划分 pos_i/D 存放在一个矩阵中,矩阵的行表示每一个非空集合对 D 的划分,矩阵的列表示决策值相同的集合,生成的矩阵为

$$\boldsymbol{D}=\begin{bmatrix}D_{11}&\cdots&D_{1j}&\cdots&D_{1|D|}\\D_{21}&\cdots&D_{2j}&\cdots&D_{2|D|}\\\vdots&&&&\vdots\\D_{i1}&\cdots&D_{ij}&\cdots&D_{i|D|}\\\vdots&&&&\vdots\\D_{k1}&\cdots&D_{kj}&\cdots&D_{k|D|}\end{bmatrix} \tag{4}$$

同理,将 $A_i\{b\}$ 所有集合中属于负域的所有集合对 D 划分 Neg_j/D 存放在另一个矩阵中,生成的矩阵为

$$\boldsymbol{D}=\begin{bmatrix}\bar{D}_{11}&\cdots&\bar{D}_{1j}&\cdots&\bar{D}_{1|D|}\\\bar{D}_{21}&\cdots&\bar{D}_{2j}&\cdots&\bar{D}_{2|D|}\\\vdots&&&&\vdots\\\bar{D}_{i1}&\cdots&\bar{D}_{ij}&\cdots&\bar{D}_{i|D|}\\\vdots&&&&\vdots\\\bar{D}_{k1}&\cdots&\bar{D}_{kj}&\cdots&\bar{D}_{k|D|}\end{bmatrix} \tag{5}$$

从这两个矩阵中可以看出,$\bar{D}_{ij}$ 只能与式(4)矩阵中与其不同行不同列的集合产生条件属性频率,为了求得所有条件属性频率且不重复计算,在式(4)矩阵中,定义任一行的和,即 $S_i=|\text{pos}_i/D|=$

$\sum_{1\leqslant j\leqslant|D|}|D_{ij}|=|\mathrm{pos}_i|$，则所有行的总和 $S=\sum_{1\leqslant i\leqslant k}S_i$。

定义任一列的和：$T_j=\sum_{1\leqslant i\leqslant k}D_{ij}$。

则若两个对象一个在 U_{pos} 中，另一个在 U_{neg} 中，产生的条件属性频率 $N_2=\sum_{1\leqslant i\leqslant|D|,1\leqslant j\leqslant k}\bar{D}(S-S_i-T_j+D_{ij})$。故 $F_B(U,a)=\sum_{1\leqslant i\leqslant l}(2N_1+N_2)$ 表示简化决策表中所有对象相对于条件属性集 B 产生的条件属性频率的总个数，证明完毕。

根据定义 6 可知，只有属性值不同且不为缺省值的才能包含条件属性，所以在本文的所有算法中，对象 U 对属性 a 的划分，将含有缺省值的放在划分的最后一个集合里，不予处理。

4 属性约简算法

首先，对不完备决策系统中的对象进行划分。

算法 1 论域 U 对属性 a 的划分

输入 不完备决策表 $S=(U,C,D,V,f)$，$C=\{a_1,a_2,\cdots,a_m\}$，$U=\{x_1,x_2,\cdots,x_{|U|}\}$

输出 $U/a=\{A_1,A_2,\cdots,A_t\}$

1) $t=1$；$A_t=\{x_i\}$；

2) for($j=2$；$j<|U|+1$；j++)。

若任一条件属性 $a_i\in C(i=1,2,\cdots,|C|)$ 均有 $f(x_i,a_i)=f(x_j,a_i)\neq *$，则 $A_t=A_t\cup\{x_j\}$；否则 $t=t+1$；$A_t=\{x_j\}$；(其中在此求划分时 * 单独放到一块)。

3) 输出 $U/a=\{A_1,A_2,\cdots,A_t\}$。

算法 1 中，1)、3) 时间复杂度忽略不计，2) 的时间复杂度为 $O(|U|)$，则算法 2 的时间复杂度是 $O(|U|)$，空间复杂度为 $O(|U|)$。

算法 2 求条件属性频率的函数

输入 $U/A=\{A_1,A_2,\cdots,A_t\}$，条件属性的最大值和最小值分别标记为 M_b,m_b；

输出 $U/(A\cup\{b\})$，条件属性频率函数 $|F_a(U,b)|$；

1) $|F_A(U,b)|=0$，$U/(A\cup\{b\})=\varnothing$；

2) 对 $\forall A_i=\{x_1,x_2,\cdots,x_j\}\in U/A$，以静态链表为存储空间，依次放入对象 $x_1,x_2,\cdots,x_j$；令表头指针指向 x_i；

①建立 M_b-m_b+2 空队列，令 front[k] 和 end[k] ($k=0,1,2,\cdots,M_b-m_b+1$) 分别为第 k 个队列的头指针和尾指针，将链表中的对象 $x\in A_i$ 按链表中的次序分配到第 $f(x,b)-m_b$ 个队列中去，将链表中的对象值为 * 的对象分配到 * 队列中。

②对除 * 队列的每个非空队列作如下处理：

a) 将非空队列中属于 U_{pos} 的对象放入 $\mathrm{pos}_i(i=0,1,2,\cdots,k)$ 中，属于 U_{neg} 的对象放入 $\mathrm{Neg}_i(i=1,2,\cdots,k)$ 中。并计算两个对象都在 U_{pos} 中产生的条件属性频率 N_1，则 $N_1=\sum_{1\leqslant i<j\leqslant k}\mathrm{pos}_i\mathrm{pos}_j$。

b) 计算每个非空队列中 $\mathrm{pos}_j/D=\{D_{j1},D_{j2},\cdots,D_{j|D|}\}$，$\mathrm{Neg}_j/D=\{D_{j1},D_{j2},\cdots,D_{j|D|}\}$，则在正域矩阵中 $S_i=|\mathrm{pos}_i/D|=\sum_{1\leqslant j\leqslant|D|}|D_{ij}|$，$S=\sum_{1\leqslant i\leqslant k}S_i$ 所有集合中属于正域的所有集合对 D 划分 pos_j/D 中决策值相同集合总数为 $T_j=\sum_{1\leqslant i\leqslant k}D_{ij}$。一个对象在 U_{pos} 中，一个在 U_{neg} 中，产生的条件属性总频率为 $N_2=\sum_{1\leqslant i\leqslant|D|,1\leqslant j\leqslant k}\bar{D}_{ij}(S-S_i-T_j+D_{ij})$，产生的条件属性总频率为 $|F_A(U,b)|=2N_1+N_2$；

3) 输出 $U/(A\cup\{b\})$，条件属性总频率数 $|F_A(U,b)|$。

算法时间空间复杂度分析：算法 2 中 1) 的时间复杂度忽略不计，2)①的时间复杂度为 $O(|A_i|)$，设 $\mathrm{pos}_i/\{b\}=\{A_{i1},A_{i2},\cdots,A_{ik}\}$，则 2)②$a$ 时间复杂度为 $O(A_{ij})$ ($j=1,2,\cdots,k$)，2)②b 时间复杂度为 $O(A_{ij})$，即 2)②时间复杂度为 $O(|A_i|)$，2) 时间复杂度 $O(|A_i|)+O(|A_2|)+\cdots+O(|A_t|)\leqslant O(|U|)$。故算法 2 的最坏时间复杂度为 $O(|U|)$，同理可得最坏空间复杂度为 $O(|U|)$。

算法 3 以条件属性的频率为启发信息的属性约简算法

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

输出 属性约简 Red(C)。

1) 由文献[11]求出容差类 $T_{ci}(x_i)$ ($x_i\in U$)，$U_{\mathrm{pos}},U_{\mathrm{neg}}$ 计算知识粒度 $|GD(c_i)|=\sum_{i=1}^{n}|T_{ci}(x_i)|/|U|^2$，令 $|K_i|=GD(c_i)$；

2) 将 K_i 按从小到大运用快速排序方法得到 $|K_{i1}|\leqslant|K_{i2}|\leqslant\cdots\leqslant|K_{im}|$，它们对应的属性为 $c_{i1},c_{i2},\cdots,c_{im}$ 令 Red(C)=$\{c_{i1}\}$；

3) for($k=2,k<m+1;k$++)

由算法 3 计算；$|F_{\mathrm{red}}(U,c_{i(k-1)})|$

if($|F_{\mathrm{Red}}(U,c_{i(k-1)})|\neq 0$)

$\mathrm{Red}(C)=\mathrm{Red}(C)\cup\{c_{i(k-1)}\}$；

4) 输出属性约简 Red(C)。

算法正确性分析：若 $|F_{\mathrm{Red}}(U,c_{i(k-1)})|=0$，即当前属性不能将两个对象区分开，则 $R_{\mathrm{Red}\cup|c_{ik}|}=R_{\mathrm{Red}}$，则由算法 3 知，当输出约简 Red($C$)时，有 $R_C=R_{\mathrm{Red}}$。由定理 2 知，算法 3 求出的属性约简就是基于知识粒度的属性约简。

算法时间复杂度分析：算法 3 的 1) 由文献[11]知时间复杂度为 $O(K|C||U|)$ (其中 $K=\max\{|T_c(x_i)|,x_i\in U\}$)，空间复杂度为 $O(|U|)$。

2)的时间复杂度为 $O(|C|)+O(|U|)$,空间复杂度为 $O(|U|)$(由算法 1 的复杂度分析可得)。3)的时间复杂度为 $O(|C||U|)$,空间复杂度为 $O(|U|)$。故算法 3 的时间复杂度为 $O(K|C||U|)$(其中 $K=\max\{|T_C(x_i)|,x_i\in U\}$,空间复杂度为$O(|U|)$。

5　实例分析

为了证明算法的可行性,以文献[16]中的不完备决策表 1 为例子进行相应说明。

表 1　不完备决策表

Table 1　The table of incomplete decision

car	price	mileage	size	max-speed	conclusion
x_1	high	high	full	low	good
x_2	low	*	full	low	good
x_3	*	*	compact	high	poor
x_4	high	*	full	high	good
x_5	*	*	full	high	excel
x_6	low	high	full	*	good

为方便计算,将属性值从左至右简记为 P、M、S、X,则该表的条件属性为 $C=\{P,M,S,X\}$。

由算法 3 1)求得各属性的知识粒度分别是:

$|K_1|=GD(P)=(4+4+6+4+6+4)/36=28/36$;

$|K_2|=GD(M)=(6+6+6+6+6+6)/36=36/36$;

$|K_3|=GD(S)=(5+5+5+5+5+1)/36=26/36$;

$|K_3|=GD(X)=(3+3+4+4+4+6)/36=24/36$;

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

由 2)排序$|K_4|\leqslant|K_3|\leqslant|K_1|\leqslant|K_2|$,他们对应的属性为 X、S、P、M,则有 $\text{Red}(C)=\{X\}$,$R_C=\varnothing$。

由 3)计算$|F_{\varnothing}(U,X)|=6$,计算的$|F_X(U,S)|=6$,计算的$|F_{\{X,S\}}(U,P)|=1$,计算的$|F_{\{X,S,P\}}\{U,M\}|=0$,算法结束,输出约简 $\text{Red}(C)=\{X,S,P\}$。

由算法 2 求$|F_{\text{Red}}(X)|$。

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

由算法 2,2)的 2)①对 $A_1=\{x_1,x_2,x_3,x_4,x_5,x_6\}$求得:

front[1]→x_1→x_2→end[1];

front[2]→x_3→x_4→x_5→end[2];

front[*]→x_6→end[*];

对第 1 个非空队列有 $\text{pos}_1=\{x_1,x_2\}$,$\text{Neg}_1=\varnothing$;

对第 2 个非空队列 $\text{pos}_2=\{x_3\}$,$\text{Neg}_2=\{x_4,x_5\}$,则 $N_1=\sum_{1\leqslant i\leqslant j\leqslant 2}\text{pos}_i\text{pos}_j=|\text{pos}_1|*|\text{pos}_2|=2*1=2$。

由算法 2,2)的②计算每个非空队列中的 pos_i/D。

$$D_{11}=\{x_1,x_2\},D_{12}=\varnothing,D_{13}=\varnothing,$$
$$D_{21}=\varnothing,D_{22}=\{x_3\},D_{23}=\varnothing,\text{则}$$
$$S_1=|\text{pos}_1/D|=\sum_{1\leqslant i\leqslant 3}|D_{i1}|=2+0+0=2$$
$$S_2=|\text{pos}_2/D|=\sum_{1\leqslant i\leqslant 3}|D_{i1}|=0+1+0=1$$
$$S=\sum_{1\leqslant i\leqslant k}S_i=S_1+S_2=2+1=3$$
$$T_1=\sum_{1\leqslant i\leqslant 3,1\leqslant j\leqslant 2}D_{ij}=2+0=2$$
$$T_2=\sum_{1\leqslant i\leqslant 3,1\leqslant j\leqslant 2}D_{ij}=0+1=1$$
$$T_3=\sum_{1\leqslant i\leqslant 3,1\leqslant j\leqslant 2}D_{ij}=0+0=0$$

每个非空队列中的 Neg_i/D:

$$\bar{D}_{11}=\varnothing,\bar{D}_{12}=\varnothing,\bar{D}_{13}=\varnothing,$$
$$\bar{D}_{21}=\{x_4\},\bar{D}_{22}=\varnothing,\bar{D}_{23}=\{x_5\}$$
$$\boldsymbol{D}=\begin{bmatrix}2&0&0\\0&1&0\end{bmatrix},\bar{\boldsymbol{D}}=\begin{bmatrix}0&0&0\\1&0&1\end{bmatrix}$$

$N_2=\sum_{1\leqslant i\leqslant 3,1\leqslant j\leqslant 2}\bar{D}_{ij}(S-S_i-T_j+D_{ij})=0*(3-2-2+2)+0*(3-2-1+1)+0*(3-2-0+0)+1*(3-1-2+0)+0*(3-1-1+1)+1*(3-1-0+0)=1*2=2$

对 $A_*=\{x_6\}$,因 A_* 不能区分对象,故无需计算。

故$|F_{\varnothing}(U,X)|=2N_1+N_2=2*2+2=6$,

求$|F_X(U,S)|$。

输入　$U/(X)=\{\{x_1,x_2\},\{x_3,x_4,x_5\}\}$

由算法 2 2)的①对 $A_1=\{x_1,x_2\}$求得 front[1]→x_1→x_2→end[1];

对其划分有 $\text{pos}_1=\{x_1,x_2\}$,$\text{Neg}_1=\varnothing$;

易知,$|F_X(U,S)|_1=0$,

对 $A_2=\{x_3,x_4,x_5\}$求得

front[1]→x_3→end[1];

front[2]→x_4→x_5→end[2];

对第 1 个非空队列有 $\text{pos}_1=\{x_3\}$,$\text{Neg}_1=\varnothing$;

对第 2 个非空队列 $\text{pos}_2=\varnothing$,$\text{Neg}_2=\{x_4,x_5\}$,

则 $N_1\sum_{1\leqslant i<j\leqslant 2}\text{pos}_i\text{pos}_j=1*0=0$,对决策属性划分后得

$$\boldsymbol{D}=\begin{bmatrix}0&1&0\\0&0&0\end{bmatrix},\bar{\boldsymbol{D}}=\begin{bmatrix}0&0&0\\1&0&1\end{bmatrix}$$

易知 $N_2=0+0+0+1*3+0+1*3=6$

$|F_X(U,S)|_2=2N_1+N_2=0+6=6$

$|F_X(U,S)|=|F_X(U,S)|_1+|F_X(U,S)|_2=6$

输入　$U/(X\cup(\{S\}))=\{\{x_1,x_2\},\{x_3\},\{x_4,x_5\}$由算法 2 的 2)①对 $A_1=\{x_1,x_2\}$求得

front[1]→x_1→end[1];

front[2]→x_2→end[2];

对第 1 个非空队列有 $\text{pos}_1=\{x_1\}$,$\text{Neg}_1=\varnothing$;

对第 2 个非空队列 $\text{pos}_2=\{x_2\}$,$\text{Neg}_2=\varnothing$。

则 $N_1=\sum_{1\leqslant i\leqslant j\leqslant 2}\text{pos}_i\text{pos}_j=1*1=1$ 易知 $N_2=0$,

故 $|F_{\{X,S\}}(U,P)|_1=1$。

对 $A_2=\{x_3\}$ 求

front[1]→x_3→end[1];

易知,$|F_{\{X,S\}}(U,P)|_2=0$

对 $A_3=\{x_4,x_5\}$ 求得

front[1]→x_4→end[1];

front[*]→x_5→end[*];

易知,$|F_{\{X,S\}}(U,P)|_3=0$,

$$|F_{\{X,S\}}(U,P)|=|F_{\{X,S\}}(U,P)|_1+|F_{\{X,S\}}(U,P)|_2+|F_{\{X,S\}}(U,P)|_3=1$$

输入 $U/(\{X,S\}\cup\{P\})=\{\{x_1\},\{x_2\},\{x_4\}\}$

求得 $|F_{\{X,S,P\}}(U,M)|=0$。

实例说明,该约简与文献[5]相同。新算法不仅通俗易懂,且在粒度差别矩阵的基础上大大减少存储空间,且大大提高了算法收敛的时间速度,即新算法是一个高效可行的属性约简算法。

6 实验对比

为了更好地说明新算法比其他同类算法更具有有效性和实用性,选用 UCI 机器学习数据库中的 6 个数据集:Credit、Car、Hepatitis、Soybean-large、Vote 和 Wine 进行实验。选取比较新的算法进行对比,考察新算法的高效性,分别与文献[17]、文献[18]、文献[11]进行对比,文献[17]是在差别矩阵的基础上提出的属性约简算法,文献[17]算法运行时间记为 t_1,文献[18]是基于冲突域的属性约简算法,算法运行时间记为 t_2,文献[11]算法运行时间记为 t_3,本文算法运行时间记为 t_{new},对比结果见表 2。为了增强实验结果的可靠性,本文所取的最终时间为 7 次实验结果的平均值。实验运行的环境为:CPU 为 AMD,2.00 GB内存,在 Visual Stdio2010 平台。

表 2 UCI 数据集信息

Table 2 The information of UCI data sets

数据集	完备	$\|C\|$	$\|U\|$
Car	是	6	1 700
Hepatitis	否	15	199
Vote	否	16	435
Credit	否	15	690
Soybean-large	否	35	351
Wine	是	14	178

表 2 中的数据集,$|C|$ 代表条件属性个数,$|U|$ 代表对象个数。

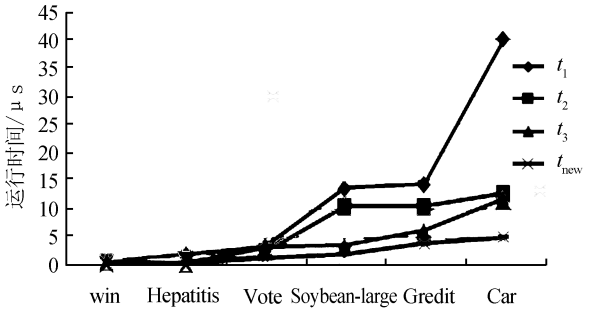


图 1 UCI 数据集对比

Fig.1 The comparison of UCI data sets

从表 2 中的实验数据可以看出,对于小的数据集({Hepatitis,15,199},{Wine,14,178})上,对比的 4 种算法的运行时间相差不大。但是对于较大的数据集,运行时间就相差很大,而且随着数据集的扩大,新算法的运行时间相对于其他 3 种算法的增长幅度小得多,表明新算法具有较好的可扩展性。

7 结束语

在决策表中,知识粒度是有效进行属性约简的方法,以往的属性约简算法由于计算知识粒度浪费了大量时间,算法效率不高。因此,本文设计一个基于知识粒度的计算条件属性频率的启发函数,以知识粒度为启发信息,提出新的属性约简算法,大大降低了算法的时间复杂度。在以后的研究中,可以将计算属性频率的思想利用到其他属性约简的方法中,如相容矩阵、差别矩阵等,也可进一步应用到规则获取中。

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 applications to data analysis[J]. Cybernetics and systems: an international, 1998, 29(7): 661-668.

[3] KRYSZKIEWICZ M. Rough set approach to incomplete information systems[J]. Information sciences, 1998, 112(1-4): 39-49.

[4] 钱文彬, 杨炳儒, 徐章艳, 等. 基于不完备决策表的容差类高效求解算法[J]. 小型微型计算机系统, 2013, 34(2): 345-350.
QIAN Wenbin, YANG Bingru, XU Zhangyan, et al. Efficient algorithm for computing tolerance classes of incomplete decision table[J]. Journal of Chinese computer systems, 2013, 34(2): 345-350.

[5] 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): 169-170, 199.

LI Xiuhong, SHI Kaiquan. A knowledge granulation-based algorithm for attribute reduction under incomplete information systems[J]. Computer science, 2006, 33(11): 169-170, 199.

[6]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): 51-53, 84.
SHI Xianhong, SHI Jinling. A reduction algorithm based on relative granularity in incomplete decision tables[J]. Journal of Henan normal university:natural science, 2010, 38(4): 51-53, 84.

[7]张清国, 郑雪峰. 基于知识粒度的不完备决策表的属性约简的矩阵算法[J]. 计算机科学, 2012, 39(2): 209-211, 243.
ZHANG Qingguo, ZHENG Xuefeng. Discernibility matrix algorithm of attribute reduction based on knowledge granulaion in incomplete decision table[J]. Computer science, 2012, 39(2): 209-211, 243.

[8]张伟, 徐章艳, 王晓宇. 一种结合概率启发信息和知识粒度的属性约简算法[J]. 计算机应用与软件, 2013, 30(7): 43-45, 50.
ZHANG Wei, XU Zhangyan, WANG Xiaoyu. An attribute reduction algorithm combining probability heuristic information and knowledge granularity[J]. Computer applications and software, 2013, 30(7): 43-45, 50.

[9]PAWLAK Z. Rough sets and intelligent data analysis[J]. Information sciences, 2002, 147(1-4): 1-12.

[10]王炜, 徐章艳, 李晓瑜. 不完备决策表中基于对象矩阵属性约简算法[J]. 计算机科学, 2012, 39(4): 201-204.
WANG Wei, XU Zhangyan, LI Xiaoyu. Attribute reduction algorithm based on object matrix in incomplete decision table[J]. Computer science, 2012, 39(4): 201-204.

[11]舒文豪, 徐章艳, 钱文彬, 等. 一种快速的不完备决策表属性约简算法[J]. 小型微型计算机系统, 2011, 32(9): 1867-1871.
SHU Wenhao, XU Zhangyan, QIAN Wenbin, et al. Quick attribution reduction algorithm based on incomplete decision table[J]. Journal of Chinese computer systems, 2011, 32(9): 1867-1871.

[12]韩智东, 王志良, 高静. 用差别矩阵思想设计的基于正区域的高效属性约简算法[J]. 小型微型计算机系统, 2011, 32(2): 299-304.
HAN Zhidong, WANG Zhiliang, GAO Jing. Efficient attribute reduction algorithm based on the idea of discernibility object pair set[J]. Journal of Chinese computer systems, 2011, 32(2): 299-304.

[13]钟珞, 梅磊, 郭翠翠, 等. 粒矩阵属性约简的启发式算法[J]. 小型微型计算机系统, 2011, 32(3): 516-520.
ZHONG Luo, MEI Lei, GUO Cuicui, et al. Heuristic algorithm for attribute reduction on granular matrix[J]. Journal of Chinese computer systems, 2011, 32(3): 516-520.

[14]唐孝, 舒兰. 基于粒计算的属性约简改进算法[J]. 计算机科学, 2014, 41(11A): 313-315, 346.
TANG Xiao, SHU Lan. Improved algorithm of attribute reduction based on granular computing[J]. Computer science, 2014, 41(11A): 313-315, 346.

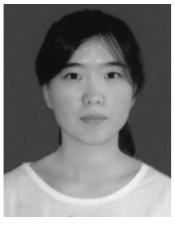
[15]张清国, 郑雪峰. 相容矩阵的高效属性约简算法[J]. 小型微型计算机系统, 2012, 33(9): 1944-1947.
ZHANG Qingguo, ZHENG Xuefeng. An efficiency attribute reduction algorithm of tolerance matrix[J]. Journal of Chiese computer systems, 2012, 33(9): 1944-1947.

[16]梁吉业, 李德玉. 信息系统中的不确定性与知识获取[M]. 北京: 科学出版社, 2005: 1-70.

[17]王炜, 徐章艳, 李晓瑜.不完备决策表中基于对象矩阵属性约简算法[J]. 计算机科学, 2012, 39(4): 201-204.
WANG Wei, XU Zhangyan, LI Xiaoyu. Attribute reduction algorithm based on object matrix in incomplete decision table[J]. Computer science, 2012, 39(4): 201-204.

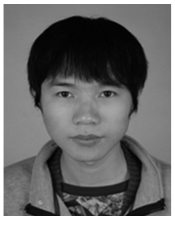
[18]周建华, 徐章艳, 章晨光. 一种基于冲突域的不完备决策表属性约简算法[J]. 计算机应用与软件, 2014, 31(3): 239-241, 255.
ZHOU Jianhua, XU Zhangyan, ZHANG Chenguang. An incomplete decision table attribute reduction algorithm based on conflict region[J]. Computer applications and software, 2014, 31(3): 239-241, 255.

作者简介:



乔丽娟,女,1988 年生,硕士研究生,主要研究方向为数据挖掘及粗糙集理论。

徐章艳,男,1972 年生,教授,博士,主要研究方向为数据挖掘、模糊集、粗糙集理论。主持国家自然科学基金项目 1 项,参与国家自然科学基金项目 2 项,主持省部级科研项目 1 项;厅局级项目 2 项;主持校级项目 2 项。发表学术论文被 SCI 检索 3 篇,被 EI 检索 5 篇。



谢小军,男,1990 年生,硕士研究生,主要研究方向为数据挖掘及粗糙集理论。