一种 Rough 集相对约简的计算方法*)

裴小兵 王元珍
（华中科技大学计算机学院数据库与多媒体研究所 武汉 430074）

摘 要 本文引入邻域分明合取项集概念，讨论了邻域分明合取项集的计算方法。在此基础上，利用约简集 $RED_Q(U-\{x_0\},P)$ 计算约简集 $RED_Q(U,P)$ 的思想，给出了相对约简的判定定理，从而提出了一种相对约简的计算方法。由于该方法不用计算分明矩阵的中间环节，节省了空间和时间，提高了运行效率。实验结果表明，该约简算法在效率上较现有的约简算法有一定提高。

关键词 Rough 集，相对约简，分明矩阵

A Calculation Method for Relative Reductions of Rough Sets

PEI Xiao-Bing WANG Yuan-Zhen
(Institute of Database and Multimedia, Department of Computer Science, Huazhong University of Science & Technology, Wuhan 430074)

Abstract In this paper, a new concept of neighbor discernible $\wedge$-clauses set is introduced, a calculation method for neighbor discernible $\wedge$-clauses set is given, and a judgement theorem for relative reduction is obtained, calculating the relative reductions $RED_Q(U,P)$ based on the relative reductiond $RED_Q(U-\{x_0\},P)$, from which an calcuation method for all relative reductions of rough set is proposed. Since it doesn't need to generate the medial link of discernible matrix, so it can spare space and time and raise the run efficiency of the calculation method. The experimental results show that the algorithm is efficient in comparison with that of the existing algorithms.

Keywords Rough set, Relative reduction, Discernible matrix

Rough 集理论[1]是 20 世纪 80 年代由波兰 Pawlak 教授提出的，它是处理模糊性和不确定性知识的一种新型数学工具，在数据挖掘、决策支持系统等领域得到了应用[2,3]。

相对约简是 Rough 集理论的核心内容之一，计算所有的相对约简已经被证明是 NP 完全问题[4]。高效的相对约简算法是 Rough 集应用于知识发现的基础，目前尚不存在一种非常有效的方法。因此，寻求快速的相对约简算法仍是 Rough 集理论研究的主要方向之一。目前，最常用的计算相对约简的方法是基于分明矩阵法[2,3,5]。但是，基于分明矩阵的约简算法效率很低，为指数复杂度。因此，当数据量很大时，这些约简算法的可行性将面临巨大的挑战。

现有相对约简算法的低效性在一定程度上限制了 Rongh 集理论的广泛应用。因此，提高 Rough 集相对约简算法的效率具有重要的意义。本文通过引入邻域分明合取项集概念，讨论了邻域分明合取项集的计算方法。在此基础上，利用相对约简 $RED_Q(U-\{x_0\},P)$ 计算相对约简 $RED_Q(U,P)$ 的思想，给出了相对约简的判定定理，从而提出了相对约简的一种计算方法。由于该方法不用计算分明矩阵及其对应的最小简化的析取范式的中间环节，节省了空间和时间，提高了运行效率。实验结果表明，该约简算法在效率上较现有的约简算法有一定提高。

1 Rough 集的基本概念

定义 1 一个信息系统 S 可以表示为：$S=\langle U,A,V,f\rangle$，其中 U 为对象的集合，即论域；A 是属性集合，$V=\bigcup_{a\in A}V_a$，V_a 表示属性 a 的值域；$f:U\times A\to V$ 是一信息函数，它指 U 中每一个对象 x 的属性值，即 $x\in U,a_k\in A$，有 $f(x,a_k)\in V_a$。

对于任意 $B\subseteq A$，记 $ind(B)=\{(x,y):f(x,a_k)=f(y,a_k),\forall a_k\in B\}$，
则 $ind(B)$ 都是 U 上的等价关系，称为由 B 决定的不可区分关系。它们产生的 U 上的划分分别为：$U/ind(B)=\{[x]_B:x\in U\}$，其中 $[x]_B=\{y:(x,y)\in ind(B)\}$，是 x 关于 B 的等价类[2,3]。

定义 2 信息系统 $S=\langle U,A,V,f\rangle$，P、Q 为 U 上的两个等价关系簇，Q 的 P 正域记为 $POS_p(U,Q)$，定义为：

$$POS_p(U,Q)=\bigcup_{X\in U/ind(Q)}P_*(X)$$

若 P 的 Q 独立子集 $S(S\subseteq P)$ 有 $POS_S(U,Q)=POS_P(U,Q)$，则称 S 为 P 的 Q 约简。记 P 的所有 Q 约简关系簇为 $RED_Q(U,P)$[2,3]。

定义 3 信息系统 $S=\langle U,A,V,f\rangle$，若 $B=\{b_1,\cdots,b_u\}\subseteq A$ 是一属性集合，对象 $x\in U$ 在属性集 B 上的信息向量：$inf_B(x)=\{f(x,b_1),f(x,b_2),\cdots,f(x,b_u)\}$。

显然，信息向量能够描述对象集 U 中的任一对象，它具有如下特点：$inf_B(x)=inf_B(y)\Leftrightarrow(x,y)\in ind(B)$。

对象集合 U 在属性集合 A 上的信息向量集：$inf_A(U)=\{inf_A(x)\mid x\in U\}$[2]。

2 相对约简的判定定理

定义 4 信息系统 $S=\langle U,A,V,f\rangle$，$R\subseteq 2^A$，其中 2^A 为 A 的幂集合，$N=\{s'_i\mid s'_i$ 是 $s_i(s_i\in R)$ 中各元素的合取$\}$，则 N 的消简是指满足如下条件的 P：1) $P\subseteq R$；2) $\forall p\in P,q\in P$ 都有 $p\not\subset q$ 且 $q\not\subset p$；3) 对任意的 $r\in R$ 都存在 $p\in P$，使得 $p\subseteq r$。

实例：取 $R=\{\{a\wedge c\},\{a\},\{b\wedge c\},\{d\},\{d\wedge e\}\}$，则 R 的消简为 $\{\{a\},\{b\wedge c\},\{d\}\}$。

*）本文得到科技部电子政务关键技术及应用系统研究项目的资助，项目编号：2001BA110B01。**裴小兵** 博士研究生，主要研究方向为数据挖掘、数据库、网络安全；**王元珍** 教授，博士生导师，主要研究方向为现代数据库理论及实现技术。

计算R的消简的时间复杂度为$O(|A|m^2)$，其中m表示R中元素的个数。

定义5 信息系统$S=\langle U,A,V,f\rangle$，$x_0\in U$，$V\subseteq U$，$V=\{x_1,\cdots,x_{|V|}\}$，P、Q为U上的两个等价关系簇，$R\subseteq 2^P$，定义$\{x_0\}$与V在属性集R下的邻域分明合取项集$N(x_0,V,R)$为由合取范式$CNF(x_0,V,R)$的最小简化的析取范式所有合取项构成的集合，其中，合取范式$CNF(x_0,V,R)$是一个有m一元变量$a_1,\cdots,a_m(a_i\in P,i=1,2,\cdots,m)$的布尔函数，它是$(\vee s'_i)\wedge c_j$的合取，$s'_i$是$s_i(s_i\in R)$中各元素的合取，$c_j$是集合$\{a\in P|f(x_0,a)\neq f(x_j,a),x_j\in V\}$中各元素的析取。

实例：考虑表1的信息系统，对象集$U=\{\#1,\#2,\#3\}$，属性集$P=\{a,b,c\}$，属性集$Q=\{d\}$：

表1 信息系统

U	a	b	c	d
1	2	2	0	1
2	0	0	0	0
3	1	0	1	0

取$R=\{\{ab\},\{ac\}\}$，$x_0=\#1$，$V=\{\#2,\#3\}$，则

$CNF(x_0,V,R)=((a\wedge b)\vee(a\wedge c))\wedge(a\vee b)\wedge(a\vee b\vee c)=(a\wedge b)\vee(a\wedge c)$

所以，$\{x_0\}$与V在集合R下的邻域分明合取项集$N(x_0,V,R)=\{a\wedge b,a\wedge c\}$。

算法1 计算邻域分明合取项集。

输入：信息系统$S=\langle U,A,V,f\rangle$，$P\subseteq A$，$R\subseteq 2^P$，$x_0\in U$，$V\subseteq U$；

输出：$\{x_0\}$与V在集合R下的邻域分明合取项集$N(x_0,V,R)$。

1 $N=\{s'_i|s'_i$是$s_i(s_i\in R)$中各元素的合取$\}$；$N'=\Phi$；

2 对所有的$x\in V$，重复以下步骤：

2.1 对所有的$\eta\in N$，重复以下步骤：

$N'=N'\cup\{\{\eta\wedge a|a\in P$且$\inf_a(x)\neq\inf_a(x_0)\}\}$

2.2 对$N=N'$的削简；$N'=\Phi$；

3 输出邻域分明合取项集$N(x_0,V,R)=N$。

算法1的时间复杂度主要由第2步决定，所以算法1的时间复杂度为$T=O(|A|(K|A|)^2|V|)=O(|A|^3K^2|V|)$，其中$K=Max\{Card(N_0(x_0,V',R)),V'\subseteq V\}$。

命题1 信息系统$S=\langle U,A,V,f\rangle$，P、Q为U上的两个等价关系簇，$x_0\in POS_P(U,Q)$，$\gamma\in RED_Q(U-\{x_0\},P)$，则$[x_0]_\gamma-\{x_0\}\subseteq U-POS_P(U,Q)$或$[x_0]_\gamma-\{x_0\}\subseteq POS_P(U,Q)$且$|([x_0]_\gamma-\{x_0\})/ind(Q)|=1$。

证明：假设$([x_0]_\gamma-\{x_0\})\cap POS_P(U,Q)\neq\Phi$且$([x_0]_\gamma-\{x_0\})\cap U-POS_P(U,Q)\neq\Phi$，则存在$y_1\in POS_P(U-\{x_0\},Q)$，$y_2\in U-\{x_0\}-POS_P(U-\{x_0\},Q)$，使得$\inf_\gamma(y_1)=\inf_\gamma(y_2)$，所以$y_1$、$y_2\in POS_\gamma(U-\{x_0\},Q)$，或$y_1$、$y_2\in U-\{x_0\}-POS_\gamma(U-\{x_0\},Q)$。又$\gamma\in RED_Q(U-\{x_0\},P)$，所以$POS_P(U-\{x_0\},Q)=POS_\gamma(U-\{x_0\},Q)$，矛盾。所以，$[x_0]_\gamma-\{x_0\}\subseteq U-POS_P(U,Q)$，或$[x_0]_\gamma-\{x_0\}\subseteq POS_p(U,Q)$。

若$[x_0]_\gamma-\{x_0\}\subseteq POS_P(U,Q)$，假设$|([x_0]_\gamma-\{x_0\})/ind(Q)|\neq 1$，则存在$y_1\in[x_0]_\gamma-\{x_0\}$，$y_2\in[x_0]_\gamma-\{x_0\}$，使得$\inf_Q(y_1)\neq\inf_Q(y_2)$，所以不存在$Y\in U-\{x_0\}/ind(Q)$，使得$[x_0]_\gamma-\{x_0\}\subseteq Y$，即$[x_0]_\gamma-\{x_0\}\not\subset POS_P(U-\{x_0\},Q)$，矛盾。所以，$|([x_0]_\gamma-\{x_0\})/ind(Q)|=1$。

定理1 信息系统$S=\langle U,A,V,f\rangle$，P、Q为U上的两个等价关系簇，$x_0\in POS_P(U,Q)$，$T=\{\gamma|\gamma\in N(x_0,V,\gamma')$，其中$\gamma'\in RED_Q(U-\{x_0\},P)$，$V=\{x|\inf_{\gamma'}(x)=\inf_{\gamma'}(x_0),X\notin P_*([x_0]_Q)\}\}$，则在$U$下的$P$的$Q$约简集$RED_Q(U,P)$等于$T$的消简。

证明：记集合T的消简为B。对$\forall\beta\in RED_Q(U,P)$，由于$x_0\in POS_P(U,Q)$，则$POS_P(U-\{x_0\},Q)=POS_\beta(U-\{X_0\},Q)$，因此存在$\gamma'\in RED_Q(U-\{x_0\},P)$，使得$\gamma'\subseteq\beta$。若$\gamma'=\beta$，则$\beta=B$；若$\gamma'\subset\beta$，由于$\gamma'\in RED_Q(U-\{x_0\}$，则存在$y\in U-\{x_0\}$，使得$\inf_{\gamma'}(x_0)=\inf_{\gamma'}(y)$且$x\notin P_*([x_0]_Q)$。

设$V=\{x|\inf_{\gamma'}(x)=\inf_{\gamma'}(x_0),x\notin P_*([x_0]_Q)\}$，由命题1，可假定$V\subseteq POS_P(U,Q)$。由于$\beta\in RED_Q(U,P)$，所以存在$\gamma\in N(x_0,V,\gamma')$，使得$\gamma\subseteq\beta$。又$POS_P(U,Q)=POS_\gamma(U,Q)$，所以$\gamma=\beta$，即$\beta\in B$。

另一方面，对$\forall\lambda\in B$，则存在$\gamma\in RED_Q(U-\{x_0\},P)$，$a_j\in c_j=\{a\in P|f(x_0,a)\neq f(x_j,a),x_j\in W\}$，使得$\lambda=\gamma\wedge a_1\wedge\cdots\wedge a_{|W|}$，其中$W=\{x|\inf_\gamma(x)=\inf_\gamma(x_0),x\notin P_*([x_0]_Q)\}$。所以，$POS_\lambda(U,Q)=POS_P(U,Q)$。假设存在$b\in\lambda$，使得$POS_{\lambda-\{b\}}(U,Q)=POS_P(U,Q)$，从而存在$\eta\in RED_Q(U-\{x_0\},P)$，$a'_j\in c_j=\{a\in P|f(x_0,a)\neq f(y_j,a),y_j\in W\}$，使得$\eta\wedge a'_1\wedge\cdots\wedge a'_{|W|}\subseteq\lambda-\{b\}\subset\lambda$，这与$\lambda\in B$矛盾。所以，$\lambda\in RED_Q(U,P)$。

定理1给出了通过约简$RED_Q(U-\{X_0\},P)$计算约简$RED_Q(U,P)$的判定定理，这为我们计算对约简提供了理论基础。

3 相对约简的计算方法

根据第2节中的结论，我们可以通过已知的约简$RED_Q(U-\{X_0\},P)$计算邻域分明合取项集，从而得到属性约简的计算方法。具体算法如下：

算法2 计算所有的相对约简。

输入：信息系统$S=\langle U,A,V,f\rangle$，P、Q为U上的两个等价关系簇，$P=\{a_1,\cdots,a_P\}$；

输出：信息系统S的相对约简集$RED_Q(U,P)$。

1 $RED=\{\{a_1\},\cdots,\{a_P\}\}$；$V=\Phi$；$M=\Phi$；

2 计算$POS_P(U,Q)$；

3 对所有$x\in POS_P(U,Q)$，执行以下步骤：

3.1 对所有$\gamma\in RED$，重复执行以下步骤：

3.1.1 计算$V_{1\gamma}=\{y|\inf_\gamma(x)=\inf_\gamma(y),y\in U-POS_P(U,Q)\}$；

3.1.2 计算$V_{2\gamma}=\{y|\inf_\gamma(x)=\inf_\gamma(y)$且$\inf_Q(x)\neq\inf_Q(y),y\in M\}$；

3.1.3 $V=V\cup V_{1\gamma}\cup V_{2\gamma}$；

3.2 计算$\{x\}$与V集合RED下的邻域分明合取项集$N(x,V,RED)$；

3.3 $RED=N(x,V,RED)$；

3.4 $M=M\cup\{x\}$；$V=\Phi$；转到3；

4 输出相对约简集$RED_Q(U,P)=RED$。

第2步的时间复杂度为$O((|P|+|Q|)|U|^2)$；步3.1的时间复杂度为$O(|K||U|)$。根据算法1的分析得步3.2的时间复杂度为$O(|A|^3K^2|U|)$，所以第3步的时间复杂度为$O(|A|^3K^2|U|^2)$。因此，算法1的时间复杂度$T=O(|A|^3K^2|U|^2)$，其中$K=Max\{Card(N(x,V,R)),V\subseteq U-\{x\},R\subseteq 2^P$，其中$2^P$为$P$的幂集合$\}$。在$K$是关于$|U|$与$|A|$的多项式时，本文给出的约简算法的时间复杂度为多项式的，文[3,5]中相应算法的时间复杂度为指数复杂度$O(2^{|A|}|U|^2|A|)$。因此，当K是关于$|U|$与$|A|$的多项式时，本文给出的约简算法的时间复杂度$O(|A|^3K^2|U|^2)$，低于文[3,5]中相应

算法的时间复杂度。

4 实验结果及分析

我们选用UCI机器学习数据库中的7个数据库在PC机(Intel-Pentium 2GHz,256MB RAM,Win2000 Professional,Microsoft Access)上进行实验,分别采用文[5](简称算法a)和本文中的约简算法2进行属性约简。实验结果如表2所示。

从表2可以看出,算法2在效率上相对于算法a有一定提高,这同该算法需要计算分明矩阵及其对应的最小简化的析取范式有直接关系。

表2 约简算法比较

数据库名称	实例数	属性集合P中属性个数	属性集合D中属性个数	算法a 执行时间(s)	算法2 执行时间(s)
Postoperative Patient	90	8	1	0.406	0.032
Monk's Problems (1)	432	6	1	0.156	0.033
Monk's Problems (2)	432	6	1	47.906	0.703
Monk's Problems (3)	432	6	1	0.141	0.046
Hayes-Roth Database	132	5	1	0.984	0.076
Teaching Assistant Evaluation	151	5	1	1.578	0.078
BUPA liver disorders	345	6	1	19.672	3.031
Balance-scale database	625	4	1	29.469	0.546
Car Evaluation Database	1728	6	1	2912.501	8.481

结论 高效的约简算法是Rough集应用于知识发现的基础。因此,寻求快速的Rough集相对约简算法具有重要的意义。本文介绍了一种相对约简的计算方法,采用已知的约简$RED_Q(U-\{x_0\},P)$计算约简$RED_Q(U,P)$的思想,提出了一种相对约简的计算方法。由于该方法不用计算分明矩阵的中间环节,节省了空间和时间,提高了运行效率。理论分析和实验结果表明,该约简算法在效率上较现有的算法有显著提高。

参考文献

1 Pawlak Z. Rough sets. International Journal of Computer and Information Science,1982,11(5):341~356

2 刘清. Rough集及Rough推理. 北京:科学出版社,2001

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

4 Wong S K M,Ziarko W. On optimal decision rules in decision tables. Bulletin of Polish Academy of Sciences,1985(33):693~696

5 Skowron A,Rauszer C. The discernibility matrics and function in information system. In: Slowinski R,ed. Intelligent Decision Support Handbook of Application and Advances of the Rough sets Theory. Dordreecht:Kluwer Academic Publishers,1991. 331~362

(上接第157页)

$Q,q_0,F,\delta)$,构造终止状态的单字母DNA有限自动机$M=(\Sigma,Q,q_0,F,\delta,P)$,其中:

$\rho=\{(a,a)|a\in\Sigma\}$;

$Q'=Q\cup\{q_f\},q_f\notin Q$;

$$F(q,a)=\begin{cases}\{q_f\},\delta(q,a)\cap F\neq\phi\\ \phi,\delta(q,a)\cap F=\phi\end{cases},q\in Q,a\in\Sigma;$$

$$\delta'(q,\binom{a}{\varepsilon})=\delta(q,a)\cup F(q,a),q\in Q,a\in\Sigma;$$

$$\delta'(q_f,\binom{\varepsilon}{a})=\{q_f\},a\in\Sigma;$$

$\delta'(q,\binom{x}{y})=\phi$,除上面之外的其它任何情况。

识别双链序列$\begin{bmatrix}w_1\\w_2\end{bmatrix}$,$M'$首先识别上链,这个识别过程同普通的有限自动机$M$一样,只是到了识别上链的最后一步时,有限自动机$M$到了终止状态,而$M'$进入了状态$q_f$,接着从左向右依照$\delta'(q_f,\binom{\varepsilon}{a})=\{q_f\}$识别双链序列的下链,这样就可以完成整个识别过程。 □

小结 本文介绍了根据DNA分子的粘接特性抽象出的DNA分子粘接计算模型的文法结构及其计算方法,证明了不同DNA粘接计算模型的计算能力,给出了DNA有限自动机的结构,进一步从生物学的角度证明了DNA有限自动机与正规文法的等价性。

参考文献

1 Adleman L M. Molecular computation of solutions to combinational problems [J]. Science, 1994, 266(5187):1021~1023

2 Adleman L M. On Constructing a Molecular Computer [J]. In: DNA based computers, American Mathematical Society, 1996 (27):1~21

3 Kari L,et al. DNA computing, sticker system, and universality. Acta Informatica, 1998,35(5):401~420

4 Kari L,et al. At the Crossroads of DNA Computing and Formal Languages: Characterizing Recursively Enumerable Languages by Insertion-Deletion Systems [A],In: Proc. of 3rd DIMACS Workshop on DNA- Based Computers, Philadel-phia,1997. 318~333

5 Paun G H,Rozenberg G. sticker systems. Theoretical Computer Sci,1998,205

6 邹海明,等. 形式语言、自动机和语法分析[M]. 武昌:华中工学院出版社,1985

7 Sipser M. 计算理论导引[M]. 北京:机械工业出版社,2000

8 Roweis S,Erik W, et al. A sticker based model for DNA computation [J]. Journal of Computational Biology,1998,5(4):615~629

9 Cox J C,Cohen D S,Ellington A D. The complexities of DNA computation [J]. Trends in Biotechnology,1999,17(4):151~154

10 Gatterdam R W. Splicing systems and regularity [J]. Intl. Journal of Computer Mathematics,1998(31):63~67

11 Garzon M H,et al. Biomolecular Computing and Programming [J]. IEEE Trans. on Evolutionary Computation,1999,3(3):236~250

12 Sakakibara Y. DNA computers: A new computing paradigm [J]. Journal of Photopolymer Science and Technology, 1998, 11(4):681~686

13 Martin-Vide C, et al. University results for finite H systems and for Watson-Crick finite automata. Computing with Biomolecules, Springer, Berlin,1998. 200~220