一种基于二进制表示的快速求核算法

胡帅鹏[1] 张清华[1,2] 姚龙洋[1]

(重庆邮电大学计算智能重庆市重点实验室 重庆 400065)[1] (重庆邮电大学理学院 重庆 400065)[2]

摘 要 在基于粗糙集的知识发现过程中,计算条件属性对论域的划分U/C和求解属性核是尤为关键的步骤。一般需要逐个比较对象的所有条件属性值才能得出结果。提出一种基于二进制表示的方法,只需比较对象的属性值的"和"。该方法先求得所有条件属性值的"和",仅对该"和"进行一次比较,再通过判断该"和"是否重复,就能得出U/C,理论分析得到该算法的复杂度为$O(|C||U|)$;然后把计算U/C的思想应用于求解属性核,提出了一种新的快速计算属性核的高效算法。理论分析表明,无论信息系统是否一致,该算法的复杂度均可达到$O(|C||U|)$。随后通过一个实例阐明了算法的具体步骤,最后通过实验验证了算法的正确性和高效性。

关键词 粗糙集,属性核,二进制表示,信息系统,高效算法

中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.013

Effective Algorithm for Computing Attribute Core Based on Binary Representation

HU Shuai-peng[1] ZHANG Qing-hua[1,2] YAO Long-yang[1]

(Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[1]

(School of Science, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[2]

Abstract Computing partitions of the domain (U/C) based on condition attributes and searching for attribute core are the most critical and time-consuming computation during the process of knowledge discovery based on rough sets. Generally, them through comparing every attribute value of each object. In this paper, based on the binary representation, the "sum" of all the conditional attribute was got firstly. Comparing the "sums" once , we can obtain U/C through judging whether the "sums" are repeated or not, and its time complexity is $O(|C||U|)$. Then this method for computing U/C was used to design a new efficient algorithm for quickly computing attribute core, and its time complexity is $O(|C||U|)$ whether information system is consistent or not. An example was used to illustrate the detail steps of the proposed algorithms. Finally, experimental results show that the new algorithms are not only exact but also efficient.

Keywords Rough set, Attribute core, Binary representation, Information systems, Efficient algorithm

粗糙集理论是Pawlak教授于1982年提出的一种能够定量分析处理不精确、不一致、不完整信息与知识的数学工具[1,2]。属性约简是粗糙集理论中的重要研究内容,研究者们已经从多方面提出了许多关于属性约简的算法和应用[3-7]。在启发式的属性约简算法中[8-12],属性核的求解是其关键步骤。因此,高效的属性核算法是非常重要的研究课题,特别是针对复杂数据信息系统,建立高效的论域划分算法和核属性算法成为人们利用粗糙集获取知识效率的重要手段。

为此,根据Skowron提出的差别矩阵[13],Hu提出了最早的求核算法[14],该算法的时间复杂度为$O(|C||U|^2)$。叶东毅[15]举例指出了Hu的求核算法存在的问题,并且提出了新的差别矩阵及求核方法,算法的时间复杂度仍为$O(|C||U|^2)$。王国胤[16]对求解属性核问题进行了深入的探讨,分析了代数观和信息观求解属性核的差异性,得出Hu的方法由于不一致对象的比较导致了出错,有效补充和完善了文献[15]。赵军、王国胤等[17]基于决策表的一致性,定义"关键属性"的概念,在此基础上,提出一种新的计算属性核的算法,其时间复杂度为$O(|C||U|\log|U|)$,但它要求决策表是一致的,无法用于不一致的信息表。杨明、孙志挥[18]提出了一种改进的差别矩阵和求核方法,该算法的时间复杂度为$O(|C||U|\log|U|)+O(|C||U||U'|)$,其中$U'$是简化差别矩阵的论域。它为设计效率更高的基于正区域的属性核算法提供了一种新的思路,即求取简化的信息表,此时的关键步骤为求出条件属性对论域的划分,即U/C。刘少辉等[19]提出了一种时间复杂为$O(|C||U|\log|U|)$的快速排序计算U/C的方法。在此基础上,徐章艳等[20]提出了一种利用基数排序快速求U/C的方法,其时间复杂度降为$O(|C||U|)$。文献[21,22]在文献[20]的基础上,利用计算U/C的方法,分别提出了基于差别矩阵和正区域的快速求核算法,使其时间复杂度降低为$\max(O(|C||U|),O(|C||U/C||U'_{pos}|))$,但此类算法仅在处理不一致信息系统的情况下效果比较显著。

在目前的求U/C的算法中,时间复杂度较好的是文献

到稿日期:2015-11-24 返修日期:2016-04-22 本文受国家自然科学基金项目(61472056),重庆邮电大学科研训练计划项目(A2014-45)资助。

胡帅鹏(1989-),男,硕士生,主要研究方向为粗糙集与粒计算,E-mail:hushpe@163.com(通信作者);张清华(1974-),男,博士,教授,硕士生导师,主要研究方向为智能信息处理、粗糙集与粒计算等;姚龙洋(1989-),男,硕士生,主要研究方向为粗糙集与粒计算。

[20]提出的基数排序算法。文献[21,22]在文献[20]的基础上提出求属性核算法，对不一致的信息系统的效果较好，时间复杂度为 $\max(O(|C||U|),O(|C||U/C||U'_{pos}|))$。如果信息表是一致的，时间复杂度会增长到 $O(|C||U|\log|U|)$，甚至 $O(|C||U|^2)$。为此，本文首先提出了一种基于二进制表示的快速计算 U/C 的算法。首先把信息表的条件属性值转化为首位为1、其他位为0的二进制数；然后对每个对象的条件属性求“和”，该“和”称为对象的属性二进制数值和，并在文中证明了该和的唯一性；最后根据该性质，求出 U/C；同时区分出所有不重复的数值“和”及其所对应的对象集合，该集合为简化信息系统 U'。理论分析表明，其时间复杂度为 $O(|C||U|)$。根据求取条件属性论域的划分 U/C 的方法，把简化的信息系统转化为互补信息系统。在此基础上提出了新的快速计算属性核的算法。理论分析表明，该步骤的时间复杂度为 $O(|C||U'|)(|U'|\leqslant|U|)$，且满足无论信息系统是否一致，该算法的总体时间复杂度均为 $O(|C||U|)$。同时，用实例阐明了算法的具体过程和步骤。最后，通过实验验证了本文所提出的求属性核算法的正确性和高效性。

1 基本概念

为方便描述问题，本节首先介绍粗糙集和属性核的基本概念和相关结论。

定义 1[23] 决策表信息系统 S 可以表示为 $S=(U,A,V,f)$。其中，U 是对象全集，也称为论域，它是非空有限的对象集合；$A=C\cup D$ 是属性全集，子集 C 和 D 分别称为条件属性集和决策属性集，且 $C\cap D=\varnothing$；V 是属性值的集合，即属性集 A 的值域；$f:U\times A\to V$，f 为一个信息函数，表示对每个实例对象的每个属性赋予一个信息值。

定义 2[23] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，属性子集 $R\subseteq A$ 决定了一个不可分辨二元关系（不分明关系）：$IND(R)=\{(x,y)|(x,y)\in U^2,\forall_{b\in R}(b(x)=b(y))\}$。

不可分辨二元关系 $IND(R)$ 形成了 U 的划分，子集 R 对 U 的划分为 $U/R=\{X|X\subseteq U\wedge\forall_{x\in X,y\in X,b\in R}(b(x)=b(y))\}$。

定义 3[23] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，对于每个子集 $X\subseteq U$ 和属性集合 $R\subseteq A$，X 关于 R 的上近似集 $\overline{R}(X)$ 和下近似集 $\underline{R}(X)$ 分别定义如下：

$$\overline{R}(X)=\bigcup\{Y_i|Y_i\in U/R\wedge Y_i\cap X\neq\varnothing\}$$

$$\underline{R}(X)=\bigcup\{Y_i|Y_i\in U/R\wedge Y_i\subseteq X\}$$

其中 $i=1,2,3,\cdots,U/R$ 是不分明关系 R 对 U 的划分。

定义 4[23] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$R\subseteq C$ 是条件属性集，$U/D=\{[x]_D|x\in U\}$ 表示决策属性集 D 对论域 U 的划分，称 $POS_R(D)=\bigcup_{x\in U/D}\underline{R}(x)$ 为条件属性 R 关于决策属性 D 的正区域。其中 $[x]_D$ 表示对象 x 的等价类。

定义 5[13] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域，$D=\{d\}$ 是决策属性集，令差别矩阵 $M=\{m_{ij}\}$，其元素定义如下：

$$m_{ij}=\begin{cases}\{c_k|c_k\in C,f(x_i,c_k)\neq f(x_j,c_k)\}, & f(x_i,d)\neq f(x_j,d)\\ \varnothing, & \text{otherwise}\end{cases}$$

其中，$k=1,2,3,\cdots$。

定义 6 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$R\subseteq C$ 是条件属性子集，$U/R=\{[x]_R|x\in U\}$ 是条件属性子集 R 对论域 U 的一个划分。对于任意对象 x 且 $x\in U$，令 $l(x)=card\{f(x,d)|x\in[x]_R\}$，称 $l(x)$ 为对象 x 在划分 U/R 下决策值的个数。

这里，如果 $S=(U,A,V,f)$ 是一致信息系统，则 $l(x)=1$；如果 $S=(U,A,V,f)$ 是不一致信息系统，则 $l(x)\geqslant1$。

定义 7[15] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域，$M=\{m_{ij}\}$ 是差别矩阵，令新差别矩阵 $NM=\{m'_{ij}\}$，其元素定义如下：

$$m'_{ij}=\begin{cases}m_{ij}, & \min(l(x_i),l(x_j))=1\\ \varnothing, & \text{otherwise}\end{cases}$$

定义 8[22] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$U/C=\{[x_1']_c,[x_2']_c,\cdots,[x_m']_c\}$，令 $S'=\{U',A,V,f\}$，S' 为简化信息系统，其中 $U'=\{x_1',x_2',\cdots,x_m'\}$ 是简化信息系统的论域。

定义 9[22] 设 $S'=(U',A,V,f)$ 是简化决策表信息系统，论域为 $U'=\{x_1',x_2',\cdots,x_m'\}$ 且 $U'_{pos}=\{x_i'|l(x_i')=1\wedge x_i'\in U'\}$，称 U'_{pos} 是简化决策表信息系统的正域，$U'_{neg}=U'-U'_{pos}$ 为其负域。

定义 10[22] 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$a(a\in C)$ 是一个条件属性，若 $POS_{C-\{a\}}(D)\neq POS_C(D)$，则称属性 a 在 C 中是不可缺少的；C 中所有不可缺少的属性的集合称为信息系统的属性核，记为 $Core(C)$。

定理 1[15] 设 $S'=(U',A,V,f)$ 是一个简化信息系统，当且仅当某个 m''_{ij} 为单个属性时，该属性是 $Core(C)$ 的元素，即如果 $SM(C)=\{m'_{ij}|m'_{ij}\in NM\wedge m'_{ij}$ 是单个属性$\}$ 成立，则 $SM(C)=Core(C)$。

证明：文献[15]中已证明。

2 基于二进制表示的快速计算 U/C 的算法

为了更清楚地介绍本文的算法，首先提出相关的定义并给出相关的定理，然后给出该算法的具体步骤。

2.1 相关定义和原理

基于二进制表示的快速计算 U/C 的算法的主要步骤：1）赋值，即在信息系统中，对于不同的条件属性，根据其值的不同按顺序从小到大进行赋值；2）求和，即对赋值后的信息系统求取所有对象的属性和；3）判断，根据属性和与决策值的不同，判断是否属于同一个划分。下面给出详细介绍：给定信息系统 $S=(U,A,V,f)$，$U=\{x_1,x_2,\cdots,x_n\}$，$C=\{c_1,c_2,\cdots,c_m\}$，记 $C'=\{c_1',c_2',\cdots,c_m'\}$ 表示条件属性对应的不同属性值的个数。

定义 11（二进制赋值） 有条件属性 $c_k\in C(k=1,2,\cdots,m)$。总体的赋值方法是随着 k 的增加，t 从1到 c_k' 递增。即第一个条件属性 c_1 对应的第 $t(t\leqslant c_1')$ 个属性值赋值的二进制数为：二进制数1B左移 $t-1$ 位；第 $k(2\leqslant k\leqslant m)$ 个条件属性 c_k 对应的第 $t(t\leqslant c_2')$ 个属性值赋值的二进制数为：二进制数1B左移 $(\sum_{i=1}^{k-1}c_i')-(k-1)+t-1$ 位。并记赋值后新形成的决策信息系统为 $S_1=(U,A,V_1,f_1)$。

实际上，赋值的二进制数与十进制数是相通的：第 c_k 个条件属性对应的全部的 c_k' 个不重复的属性值的赋值结果为：$2^{(\sum_{i=0}^{k-1}c_i')-(k-1)+0}$，$2^{(\sum_{i=0}^{k-1}c_i')-(k-1)+1}$，…，$2^{(\sum_{i=0}^{k-1}c_i')-(k-1)+c_k'-1}=2^{(\sum_{i=0}^{k}c_i')-k}$，其中 $c_0'=0$。

例1 设 $S=(U,A,V,f)$ 是一个决策表信息系统，如表1

所列。它有 a,b,c 3 个条件属性，它们对应的不重复的属性值的个数分别为 3,3,2。此时，$c_1=a, c_2=b, c_3=c$，且 $c_1'=3$，$c_2'=3, c_3'=2$。

表 1 信息系统 S

U	a	b	c	D
x_1	1	1	1	Y
x_2	2	2	2	N
x_3	3	3	2	Y
x_4	1	1	1	N

根据定义得：条件属性 $c_1=a(k=1)$ 的 3 个不同的属性值分别赋值为 1B(1B 左移 (1－1))，10B，100B，十进制表示为 2^0，2^1 与 2^2；条件属性 $c_2=b(k=2)$ 的 3 个不同的属性值分别赋值为 100B，1000B(1B 左移(3)－(2－1)＋(2－1))，10000B，十进制表示为 2^2，2^3 与 2^4；条件属性 $c_3=c(k=3)$ 的两个不同的属性值赋值为 10000B 和 100000B(1B 左移(3＋3)－(3－1)＋(2－1))，十进制表示为 2^4 与 2^5。

赋值后的信息系统如表 2 所列。

表 2 赋值后的信息系统 S_1

U	a	b	c	D
x_1	2^0	2^2	2^4	Y
x_2	2^1	2^3	2^5	N
x_3	2^2	2^4	2^5	Y
x_4	2^0	2^2	2^4	N

定理 2 在决策表信息系统 $S=(U,A,V,f)$ 中，按定义 11 赋值后的信息系统为 $S_1=(U,A,V_1,f_1)$，则必有 $f_1(x_i,c_a)\leqslant f_1(x_i,c_b)$，其中 $0\leqslant a\leqslant b\leqslant m$。

证明：由定义即得。

定义 12(对象二进制数值和) 在由定义 11 得到的信息系统 $S_1=(U,A,V_1,f_1)$ 中，令 $f_1(x_i,c_k)$ 为对象 x_i 在对应条件属性 c_k 下的值，记 $sum_i=\sum_{k=1}^{m}f_1(x_i,c_k)$，称 sum_i 为对象 x_i 的二进制数值和，其中 $x_i\in U, c_k\in C$。

表 2 中，赋值后的信息系统 S_1 求和后的结果如表 3 所列。

表 3 信息系统 S_1 求和

U	Sum	D
x_1	$2^0+2^2+2^4=21$	Y
x_2	$2^1+2^3+2^5=42$	N
x_3	$2^2+2^4+2^5=52$	Y
x_4	$2^0+2^2+2^4=21$	N

定理 3(对象二进制数值和唯一) 给定决策信息系统 $S=(U,A,V,f)$，$\forall x_i,x_j(x_i\in U,x_j\in U)$，对象 x_i 与 x_j 对应的对象二进制数值和 $sum_i=sum_j$ 成立的充要条件是它们的条件属性值完全相同。

证明：设 $S=(U,A,V,f)$ 是一个决策表信息系统，赋值后新形成的信息系统 $S_1=(U,A,V_1,f_1)$，记 x_i 对应的二进制数值分别为 $2^{a_1},2^{a_2},\cdots,2^{a_m}$；$x_j$ 对应的二进制数值分别为 $2^{b_1},2^{b_2},\cdots,2^{b_m}$。对象 x_i 与 x_j 对应的条件属性值完全相同等价于 $\{(2^{a_1},2^{b_1}),(2^{a_2},2^{b_2}),\cdots,(2^{a_m},2^{b_m})\}$ 中所有值对均相等。

由定义 12 知，对象 x_i 的二进制数值和是 $sum_i=2^{a_1}+2^{a_2}+\cdots+2^{a_m}$，对象 x_j 二进制数值和是 $sum_j=2^{b_1}+2^{b_2}+\cdots+2^{b_m}$，且由定理 2 可知，$2^{a_1}\leqslant 2^{a_2}\leqslant\cdots\leqslant 2^{a_m}$ 成立，$2^{b_1}\leqslant 2^{b_2}\leqslant\cdots\leqslant 2^{b_m}$ 成立。

充分性：当 $\{(2^{a_1},2^{b_1}),(2^{a_2},2^{b_2}),\cdots,(2^{a_m},2^{b_m})\}$ 中所有值对都相等时，即有 $2^{a_1}=2^{b_1},2^{a_2}=2^{b_2},\cdots,2^{a_m}=2^{b_m}$ 成立，所以 $2^{a_1}+2^{a_2}+\cdots+2^{a_m}=2^{b_1}+2^{b_2}+\cdots+2^{b_m}$，故 $sum_i=sum_j$ 成立。

必要性：$sum_i=sum_j\Leftrightarrow 2^{a_1}+2^{a_2}+\cdots+2^{a_m}=2^{b_1}+2^{b_2}+\cdots+2^{b_m}$

当 $2^{a_1}\neq 2^{b_1}$ 时，不妨设 $2^{a_1}<2^{b_1}$，因为 $2^{a_1}\leqslant 2^{a_2}\leqslant\cdots\leqslant 2^{a_m}$，且 $2^{b_1}\leqslant 2^{b_2}\leqslant\cdots\leqslant 2^{b_m}$，所以方程两边同时除以 2^{a_1} 后得：左边为奇数，右边为偶数。与题设矛盾，故必有 $2^{a_1}=2^{b_1}$。此时，两边分别减去 2^{a_1} 和 2^{b_1}，等式仍然成立。

同理，可以得出 $2^{a_2}=2^{b_2},2^{a_3}=2^{b_3},\cdots,2^{a_m}=2^{b_m}$。

综上所述，定理得证。

实际上，由定理 3 可知，对信息系统的属性值用二进制形式表示后的条件属性值满足首位为 1，其他位都为 0。任取 k 个这种二进制数，它们的和可以计算得出。反过来，如果把这个和分解成 k 个同种类型的二进制数，它的分解方法是唯一的。

2.2 基于二进制的 U/C 算法及其具体步骤

算法 1 BinaryPartition()

输入：决策表 $S=\{U,A,V,f\}$，$U=\{u_1,u_2,\cdots,u_n\}$，$A=C\cup D$，且 $C=\{c_1,c_2,\cdots,c_m\}$

输出：C 对应的划分 U/C

step1 根据定义 11，对整个信息系统进行赋值：

```
for(j=0,num=0;j<m;++j)
for(i=0;i<n;++i)
    f1(ui,cj)=2^num,++num;
```

step2 根据定义 12，对赋值后的信息表进行求和：

```
for(i=0;i<n;++i)
  for(j=0;j<m;++j)
    sumi+=f1(ui,cj);
```

step3 把相同的和归于同一个划分：

```
for(i=0;i<n;++i)
```

统计 sum_i 的值以及将对应的论域序列号存入向量中；

根据得到的序列号，分别标记为 $[u_1']_C,[u_2']_C,\cdots,[u_s']_C$，即有 $U/C=\{[u_1']_C,[u_2']_C,\cdots,[u_s']_C\}$，$U'=\{u_1',u_2',\cdots,u_s'\}$。

step4 初始化 $U'_{pos}=\varnothing$；$U'_{neg}=\varnothing$；

```
for(i=0;i<s;++i)
    for(j=0;j<|[ui']C|;++j)
        得到 U'pos与 U'neg
```

如果 $[u_i']_C$ 中的对象的决策值都相等，取 $[u_i']_C$ 中的第一个对象放入 U'_{pos}；如果 $[u_i']_C$ 中的对象的决策值不相等，取 $[u_i']_C$ 中的第一个对象放入 U'_{neg}；且 $U'=U'_{pos}\cup U'_{neg}$。

算法 1 的 step1 和 step3 的时间复杂度均为 $O(|C||U|)$；step2 只需要遍历整个信息表，并对每行求和，所以时间复杂度为 $O(|C||U|)$；step4 需要遍历整个论域，所以时间复杂度为 $O(|U|)$。从而算法 1 的时间复杂度可以达到 $O(|C||U|)$。

3 基于二进制的快速求核算法

一个条件属性是否为核属性，主要是判断除去该条件属性后其他条件属性是否可以区分决策信息系统。基于二进制的快速求核算法的基本原理是：在赋值求和后的信息系统中，如果除去某个条件属性后，考虑剩下的所有其他条件属性的二进制数值和是否可以区分信息系统，若不能区分，则该条件属性就是核属性。

3.1 基于二进制的快速求核算法的基本概念

定义 13(补信息系统) 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$S_1'=(U',A,V_1,f_1)$ 是简化的二进制赋值后的

信息系统，令 $f_2(x_i,c_k)=sum_i-f_1(x_i,c_k)$，其中 $x_i\in U,c_k\in C$，称 $S_2'=(U',A,V_2,f_2)$ 是 S_1' 的补信息系统。

在 S_1' 的补信息系统 $S_2'=(U',A,V_2,f_2)$ 中，$f_2(x_i,c_t)$ 记录的是除去条件属性 c_k 外，其他条件属性在对象 x_i 中的二进制数值和。

定理 4 设 $S=(U,A,V,f)$ 是一个决策表信息系统，$S'=(U',A,V,f)$ 是其简化的信息系统，$S_1'=(U',A,V_1,f_1)$ 是简化的二进制赋值后的信息系统，$S_2'=(U',A,V_2,f_2)$ 是 S_1' 的补信息系统。当 $f_2(x_i,c_k)=f_2(x_j,c_k)(x_i\in U',x_j\in U',c_k\in C)$ 且同时满足 $f_2(x_i,d)\neq f_2(x_j,d)$ 时，则 m'_{ij} 为单个属性值。

其中，$m'_{ij}=\begin{cases}\{m_{ij}\}, & \min(l(x_i),l(x_j))=1\\ \varnothing, & \text{otherwise}\end{cases}$

证明：由算法 1 的 step4 知，在简化的信息系统 $S'=(U',A,V,f)$ 中，U'/C 中所有划分集合包含的都是单个元素。即在 U' 中，不存在条件属性值完全相同的两个论域。

1) $l(x_i)=l(x_j)=1$，其中 $x_i\in U',x_j\in U'$。由此得出在信息系统 S_1' 中，对象 x_i 与 x_j 对应的属性值没有完全相同。

2) 当 $f_2(x_i,c_k)=f_2(x_j,c_k)$ 成立时，由定义 13 知，在 x_i 与 x_j 中除去条件属性 c_k 外，其他条件属性的属性值完全相同，且有 $f_2(x_i,d)\neq f_2(x_j,d)$，其中 $D=\{d\}$，所以在信息系统 S_2' 中，除了条件属性 c_k 外，剩下的条件属性值全相同，因此 m'_{ij} 为单个属性。

综合 1) 与 2)，定理得证。

定理 4 表明，如果条件属性 c_k 是核属性，则其必须满足条件 $f_2(x_i,c_k)=f_2(x_j,c_k)$ 且 $\min(l(x_i),l(x_j))=1$。

定理 5 设 $S=(U,A,V,f)$ 是一个信息系统，$S'=(U',A,V,f)$ 是简化的信息系统，$\forall x_i\in U'$，如果 $x_i\in U'_{pos}$，则 $l(x_i)=1$；如果 $x_i\in U'_{neg}$，则 $l(x_i)>1$。

证明：由算法 1 的 step4 易知定理成立。

定理 5 表明，在信息系统 $S=(U,A,V,f)$ 中，$\min(l(x_i),l(x_j))=1$ 的充要条件是 $x_i\in U'_{pos}$ 或 $x_j\in U'_{pos}$ 成立，即对象 x_i 与 x_j 至少有一个从 U'_{pos} 中取出。

3.2 基于二进制的快速求核算法的具体步骤

给定信息系统 $S=(U,A,V,f)$，由定理 1 知，S 的核属性满足定义 7 给出的新差别矩阵值是单个属性的集合；由定理 5 知，$\min(l(x_i),l(x_j))=1$ 等价于 $x_i\in U'_{pos}$ 或者 $x_j\in U'_{pos}$ 成立；由定理 4 知，m'_{ij} 为单个属性值等价于 $f_2(x_i,c_t)=f_2(x_j,c_t)$，且 $f_2(x_i,d)\neq f_2(x_j,d)$，其中 $D=\{d\}$。

下面给出基于二进制的快速求核算法。

算法 2 BinaryCore()

输入：决策表 S={U,A,V,f}，U={$u_1,u_2,\cdots,u_n$}，A=C∪D，且 C={$c_1,c_2,\cdots,c_m$}

输出：决策表的核 core(C)

step1 由算法 1 求出 U'={$x_1',x_2',\cdots,x'_{ps+ng}$}，$U'_{pos}$={$x_1,x_2,\cdots,x_{ps}$}，$U'_{neg}$={$y_1,y_2,\cdots,y_{ng}$}，core(C)=∅；

step2 for(i=0;i<ps+ng;++i)

for(j=0;j<m;++j)

$f_2(x_i',c_j)=sum_i-f_1(x_i',c_j)$；

step3 for(j=0;j<m;++j)

step3.1 for(i=0;i<ps;++i)

查找 $f_2(x_i,c_j)$ 的重复项，并判断对应的决策值是否没出现过，如果出现过且相对应的决策值与 $f_2(x_i,D)$ 不等，则 core(C)=core(C)∪{c_j}；

step3.2 for(i=0;i<ng;++i)

查找 $f_2(x_i,c_j)$ 的重复项，如果出现过，则 core(C)=core(C)∪{c_j}；

step4 输出 core(C)。

算法 2 的 step2 需要遍历信息表 S'，因此时间复杂度为 $O(|C||U'|)$；step3 的时间复杂度为 $O(|C|)$；step3.1 需要遍历 U'_{pos} 查找，所以时间复杂度为 $O(|U'_{pos}|)$；step3.2 需要遍历 U'_{neg} 查找，所以时间复杂度为 $O(|U'_{neg}|)$；故 step3 时间复杂度为 $O(|C|)\cdot O(|U'_{pos}|+|U'_{neg}|)=O(|C||U'|)$。由于算法 1 的时间复杂度为 $O(|C||U|)$，因此算法 2 的时间复杂度为 $O(|C||U|)+O(|C||U'|)$。当为一致性信息表时，由于 $S=S'$，$|U|=|U'|$，此时时间复杂度最大为 $O(|C||U|)$。

4 实例分析

为了更好地说明算法的有效性，以表 3 所列的决策信息系统为例，详细地阐述算法 1-BinaryPartition() 与算法 2-BinaryCore() 的具体步骤。

4.1 算法 1-BinaryPartition()

表 4 所列的信息系统论如下，其中论域 $|U|=15$，条件属性 $C=\{a,b,c,d\}$，决策属性为 D。

表 4 信息系统 S

U	a	b	c	d	D
x_1	2	1	2	1	0
x_2	2	2	2	1	1
x_3	2	1	2	1	0
x_4	2	3	2	3	0
x_5	2	2	2	1	1
x_6	3	1	2	1	0
x_7	1	2	3	2	2
x_8	4	3	1	2	3
x_9	3	1	2	1	1
x_{10}	1	2	3	2	2
x_{11}	3	1	2	1	1
x_{12}	4	3	1	2	3
x_{13}	4	3	4	2	1
x_{14}	1	2	3	2	3
x_{15}	4	3	4	2	2

经过算法 1 Step1 的二进制赋值后，新的决策信息系统 S_1 如表 5 所列。

表 5 赋值后的信息系统 S_1

U	a	b	c	d	D
x_1	2^0	2^3	2^5	2^8	0
x_2	2^0	2^4	2^5	2^8	1
x_3	2^0	2^3	2^5	2^8	0
x_4	2^0	2^5	2^5	2^9	0
x_5	2^0	2^4	2^5	2^8	1
x_6	2^1	2^3	2^5	2^8	0
x_7	2^2	2^4	2^7	2^{10}	2
x_8	2^3	2^5	2^6	2^{10}	3
x_9	2^1	2^3	2^5	2^8	1
x_{10}	2^2	2^4	2^7	2^{10}	2
x_{11}	2^1	2^3	2^5	2^8	1
x_{12}	2^3	2^5	2^6	2^{10}	3
x_{13}	2^3	2^5	2^8	2^{10}	1
x_{14}	2^2	2^4	2^7	2^{10}	3
x_{15}	2^3	2^5	2^8	2^{10}	2

经过算法 1 Step2 对每个对象的二进制数值求和，得到如表 6 所列的对应关系。

表 6　对象的二进制数值和

U	sum	U	sum	U	sum
x_1	297	x_6	298	x_{11}	298
x_2	305	x_7	1108	x_{12}	1192
x_3	297	x_8	1192	x_{13}	1320
x_4	577	x_9	298	x_{14}	1108
x_5	305	x_{10}	1108	x_{15}	1320

算法 1 Step3 查找相同的对象二进制数值和(*sum* 值)的对象形成的集合，即为 U/C;$U/C=\{\{x_1,x_3\},\{x_2,x_5\},\{x_4\},\{x_6,x_9,x_{11}\},\{x_7,x_{10},x_{14}\},\{x_8,x_{12}\},\{x_{13},x_{15}\}\}$，所以 $U'=\{x_1,x_2,x_4,x_6,x_7,x_8,x_{13}\}$。

经过算法 1 Step5 比较 U/C 中决策值是否唯一得出：$U'_{pos}=\{x_1,x_2,x_4,x_8\}$，$U'_{neg}=\{x_6,x_7,x_{13}\}$。

4.2　算法 2-BinaryCore()

首先，由算法 1 得出 U'_{pos} 与 U'_{neg}；其次，在此基础上，根据算法 2 的 step2，可以得出简化的补信息系统，如表 7 所列。

算法 2 Step3：

第一次指向 a：在 step3. 1 中，没有相等的属性值；在 step3. 2 中，x_1 与 x_6 指向的属性值相等，所以 $core(C)=\{a\}$；

第二次指向 b：在 step3. 1 中，x_1 与 x_2 的属性值相等，并且决策值 $0\neq1$，所以 $core(C)=\{a,b\}$；

第三次指向 c：在 step3. 1 中，没有相等的属性值；在 step3. 2 中，x_8 与 x_{13} 指向的属性值相等，所以 $core(C)=\{a,b,c\}$；

第四次指向 d：在 step3. 1 中，没有相等的属性值；在 step3. 2 中，没有相等的属性值。

在算法 2 的 step4 中输出 $core(C)=\{a,b,c\}$。

表 7　简化的补信息系统 S_2'

U	a	b	c	d	D
x_1	296	289	265	41	0
x_2	304	289	273	49	1
x_4	576	545	545	65	0
x_6	296	290	266	42	0
x_7	1104	1092	1044	84	2
x_8	1184	1160	1064	168	3
x_{13}	1312	1288	1064	296	1

5　实验

为了验证本文算法，在 Pentium(R) CPU 3.00GHz，RAM 为 4GB，Visual studio 2013 的 C++ 环境下，分别对文献[20]的基数排序算法和本文算法 1 中的算法进行比较，以及对文献[22]和本文算法 2 中的求核算法进行比较。

选择了 UCI 机器学习数据库中的 6 个数据集进行测验，测试结果如表 8 所列(其中，$|U|$ 表示对象个数，$|C|$ 表示条件属性个数，$|U'|$ 表示简化决策表的对象个数，$|core|$ 表示核属性的个数)。

表 8　对比试验结果

数据集	|U|	|C|	一致性	|U'|	|Core|	执行时间(ms)			
						文献[20]算法	本文算法 1	文献[22]算法	本文算法 2
Zoo	101	17	否	59	8	95	**49**	101	**59**
Spect heart	187	22	否	169	11	208	**116**	187	**172**
Qualitative	250	6	是	103	0	103	**54**	250	**67**
Balance Scale	625	4	是	625	4	189	**121**	625	**139**
Tic-Tac-Toe	958	9	是	958	0	521	**303**	958	**492**
Contraceptive	1473	9	否	1358	9	842	**453**	1473	**549**

依次将表 8 中的 6 个数据集记为 A,B,C,D,E,F，它们的执行时间的折线图如图 1 所示。

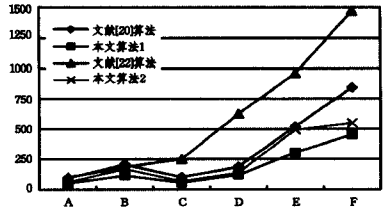

图 1　对比实验结果

分析表 8 和图 1 发现，文献[20]中算法求取条件属性的划分所需的时间明显高于本文算法 1 的时间，其原因为文献[20]算法 1 的 Step1 中，需要统计属性的最大与最小值，而本文算法不需统计；文献[22]中算法的时间开销普遍多于算法 2 的时间，而且在信息表是一致或者不一致的对象较少的情况下，随着数据集的增大，时间开销越大，这也符合我们的理论分析。

结束语　粗糙集是一种有效的处理不确定性知识发现的方法。在知识发现的过程中，主要是通过对不分明关系(U/C)的分类以及分类对于目标的近似来实现知识发现的。在求解 U/C 时，一般需要逐个比较所有条件属性的值，再据此判断对象是否可以划分到同一个论域当中，效率较低。本文在研究信息系统的基础上，提出了一种基于二进制的快速求解 U/C 的算法——BinaryPartition()。首先，对所有的条件属性值进行二进制赋值，然后对每个对象的二进制值进行求和，最后仅需比较一个属性，即对象和是否相同，得出求取论域的划分。该方法在时间效率上优于传统的逐个比较所有条件属性值的方法，为求解不分明关系提供了一种新的思路。同时，根据本文提出的计算 U/C 的思想，把求解条件属性是否为核属性的问题转化为求解条件属性值的和是否相等的问题，设计出了一种新的快速求属性核的算法，并通过实验验证了该算法的有效性和高效性。在接下来的研究中，将以此为基础来实现粗糙集的属性约简和规则获取。

参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 38(11): 341-356

[2] Pawlak Z. Rough set theory and its application to data analysis [J]. Cybernetics and Systems, 2010, 29(7): 661-688

[3] Qian Jin, Miao Duo-qian, Zhang Zhe-hua, et al. Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation[J]. International Journal of Approximate Reasoning, 2011, 52(2): 212-230

[4] Zhang Yi-rong, Xian Ming, Xiao Shun-ping, et al. An Anomaly Intrusion Detection Technique of Support Vector Machine Based on Rough Set Attribute Reduction[J]. Computer Science, 2006, 33(6): 64-68(in Chinese)
张义荣，鲜明，肖顺平，等. 一种基于粗糙集属性约简的支持向量异常入侵检测方法[J]. 计算机科学，2006，33(6)：64-68

(下转第 107 页)

Conference on Uncertainty in Artificial Intelligence. AUAI Press,2009:27-34

[12] Yao L, Mimno D, McCallum A. Efficient methods for topic model inference on streaming document collections[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2009:937-946

[13] Hoffman M, Bach F R, Blei D M. Online learning for latent dirichlet allocation[C]// Advances in Neural Information Processing Systems. 2010:856-864

[14] Zeng J, Liu Z Q, Cao X Q. Fast Online EM for Big Topic Modding[J]. IEEE Transactions on Knowledge & Data Engineering,2016,28(3):675-688

[15] Ye Y, Gong S, Liu C, et al. Online belief propagation algorithm for probabilistic latent semantic analysis[J]. Frontiers of Computer Science,2013,7(4):526-535

[16] Asuncion A, Welling M, Smyth P, et al. On smoothing and inference for topic models[C]// Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press,2009:27-34

[17] Braun M, McAuliffe J. Variational inference for large-scale models of discrete choice[J]. Journal of the American Statistical Association,2010,105(489):324-335

[18] Wallach H M, Mimno D M, Mccallum A. Rethinking LDA: why priors matter[J]. Advances in Neural Information Processing Systems,2009(23):1973-1981

[19] Gao Yang, Yang Lu, Liu Xiao-sheng, et al. Study of Semantic Understanding by LDA[J]. Computer Science,2015,42(8):279-282(in Chinese)
高阳,杨璐,刘晓升,等. LDA 语义理解研究[J]. 计算机科学,2015,42(8):279-282

(上接第 83 页)

[5] Zhang Ying-chun, Su Bo-hong, Cao Juan. Study on application of attributive reduction based on Rough sets in Data Mining[J]. Computer Science,2013,40(8):223-226(in Chinese)
张颖淳,苏伯洪,曹娟. 基于粗糙集的属性约简在数据挖掘中的应用研[J]. 计算机科学,2013,40(8):223-226

[6] Li Ming, Deng Shao-bo, Feng Sheng-zhong, et al. Fast assignment reduction in inconsistent incomplete decision systems[J]. Journal of Systems Engineering & Electronics,2014,25(1):83-94

[7] Zhang Qin-hua, Guo Yong-hong, Xiao Yu. Attribute Reduction Based on Approximation Set of Rough Set[J]. Journal of Computational Information Systems,2014,10(16):6859-6866

[8] Wang Yong-sheng, Zheng Xue-feng, Suo Yan-feng. Dynamic algorithm for computer attribute reduction based on information granularity[J]. Computer Science,2015,42(4):213-216(in Chinese)
王永生,郑雪峰,锁延锋. 一种基于信息粒度的动态属性约简求解算法[J]. 计算机科学,2015,42(4):213-216

[9] Yang Xi-bei, Yan Xu, Xu Su-ping, et al. New Heuristic Attribute Reduction Algorithm Based on Sample Selection[J]. Computer Science,2016,43(1):49-52(in Chinese)
杨习贝,颜旭,徐苏平,等. 基于样本选择的启发式属性约简方法研究[J]. 计算机科学,2016,43(1):49-52

[10] R I, CT G, Enriquez S, et al. Attributing reductions in coral calcification to the saturation state of aragonite, comments on the effects of persistent natural acidification[J]. Proceedings of the National Academy of Sciences of the United States of America,2014,111(3):E300-E301

[11] Wang Chang-zhong, He Qiang, Chen De-gang, et al. A novel method for attribute reduction of covering decision systems[J]. Information Sciences,2014,254(5):181-196

[12] Qian Jin, Lv Ping, Yue Xiao-dong, et al. Hierarchical attribute reduction algorithms for big data using MapReduce[J]. Knowledge-Based Systems,2015,73(12):18-31

[13] Skowron A, Rauszer C. The Discernibility Matrices and Functions in Information Systems[J]. Theory & Decision Library,2012,11:331-362

[14] Hu Xiao-hua, Cercone N. Learning in Relational Data-bases: A Rough Set Approach[J]. Computational Intelligence, 1995, 11(2):323-338

[15] Ye Dong-yi, Chen Zhao-jiong. A new discernibility matrix and the computation of a core[J]. Chinese Journal of Electronic,2002,30(7):1086-1088(in Chinese)
叶东毅,陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报,2002,30(7):1086-1088

[16] Wang Guo-yin. Calculation methods for core attributes of decision table [J]. Chinese Journal of Computers,2003,26(5):611-615(in Chinese)
王国胤. 决策表核属性的计算方法[J]. 计算机学报,2003,26(5):611-615

[17] Zhao Jun, Wang Guo-yin, Wu Zhong-fu, et al. An efficient approach to compute the feature core[J]. Mini-Micro Systems,2003,24(11):1950-1953(in Chinese)
赵军,王国胤,吴中福,等. 一种高效的属性核计算方法[J]. 小型微型计算机系统,2003,24(11):1950-1953

[18] Yang Ming, Sun Zhi-hui. Improvement of discernibility matrix and the computation of core[J]. Journal of Fudan University,2004,43(5):865-868(in Chinese)
杨明,孙志挥. 改进的差别矩阵及其求核方法[J]. 复旦学报(自然科学版),2004,43(5):865-868

[19] Liu Shao-hui, Sheng Qiu-jian, Shi Zhong-zhi. A New Method for Fast Computing Positive Region[J]. Journal of Computer Research and Development,2003,40(5):637-642(in Chinese)
刘少辉,盛秋戬,史忠植. 一种新的快速计算正区域的方法[J]. 计算机研究与发展,2003,40(5):637-642

[20] Xu Zhang-yan, Liu Zuo-peng, Yang Bing-ru, et al. A quick attribute reduction algorithm with complexity of max $\{O(|C||U|), O(|C|^2|U/C|)\}$ [J]. Chinese Journal of Computers,2006,29(3):391-399(in Chinese)
徐章艳,刘作鹏,杨炳儒,等. 一个复杂度为 max$\{O(|C||U|), O(|C|^2|U/C|)\}$的快速属性约简算法[J]. 计算机学报,2006,29(3):391-399

[21] Xu Zhang-yan, Yang Bing-ru, Song Wei. Quick Computing Core Algorithm Based on Discernibility Matrix[J]. Computer Engineer and Applications,2006,42(6):4-6(in Chinese)
徐章艳,杨炳儒,宋威. 一个基于差别矩阵的快速求核算法[J]. 计算机工程与应用,2006,42(6):4-6

[22] Xu Zhang-yan, Yang Bing-ru, Cai Wei-dong, et al. Quick algorithm for computing core based on the positive region [J]. Systems Engineering and Electronics,2006,28(12):1902-1905(in Chinese)
徐章艳,杨炳儒,蔡卫东,等. 一个基于正区域的快速求核算法[J]. 系统工程与电子技术,2006,28(12):1902-1905

[23] Wang Guo-yin. Rough set theory and knowledge acquisition [M]. Xi'an: Xi'an Jiao Tong University Press, 2011(in Chinese)
王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001