基于二进制的粗糙集基本运算研究*

Research of Rough Set Basic-Operations Based on Binary

李天志[1,2]，梁家荣[2]，范　平[2]，徐凤生[1]

LI Tian-zhi[1,2]，LIANG Jia-rong[2]，FAN Ping[2]，XU Feng-sheng[1]

（1. 德州学院计算机系，山东德州　253023；2. 广西大学计算机与电子信息学院，广西南宁 530004）

（1. Department of Computer Science and Technology，Dezhou University，Dezhou，Shandong，253023，China；2. College of Computer and Electronic Information，Guangxi University，Nanning，Guangxi，530004，China）

摘要：通过讨论二进制与粗糙集之间的内在联系，提出基于二进制的粗糙集运算理论，并借助二进制的位运算操作，给出粗糙集的上近似集和下近似集、等价类的交、集合的基数的运算算法. 该算法比传统的粗糙集运算算法的运算速度更快，效率更高. 该算法为扩展粗糙集的应用提供了理论基础.

关键词：粗糙集　算法　上近似集　下近似集　基数

中图法分类号：TP311.131　　**文献标识码**：A　　**文章编号**：1005-9164(2006)02-0109-04

Abstract：A novel idea of binary-based rough set operation is presented through analyzing the internal relation between binary and rough set. And several rough set operation algorithms are also offered in this paper，such as upper approximation set，lower approximation set，intersection，cardinal number and so on. These algorithms are more efficient and effective compared to the traditional methods for operating rough set. It provides the theoretical foundation to expand application of rough set.

Key words：rough set，algorithm，upper approximation set，lower approximation set，cardinal number

Rough 理论[1,2]是由波兰华沙理工大学 Pawlak 教授于 20 世纪 80 年代初提出的一种研究不完整、不确定知识和数据的表示、学习、归纳的理论方法，是一种有别于模糊集的、新型的数据分析技术. 它以等价关系、上近似、下近似等作为基本概念. Rough 理论为发现不精确和不确定知识中的重要数据结构以及复杂对象的分类提供了强有力的理论基础，已广泛应用于数据挖掘[3,4]、人工智能[5]和分类[6]等领域.

给定一个知识库，具体求出其粗糙集是一件不易的工作，传统的粗糙集运算操作主要是利用字符串匹配、查询、插入等来实现，其运算速度慢、效率低，而且对空间要求较高，大大制约了粗糙集理论在实际中的应用. 如何快速有效地利用计算机对粗糙集运算进行处理，一直是学术界关心的课题. 我们在总结前期关于二进制应用研究的基础上[7]，通过讨论二进制与粗糙集之间的内在联系，提出基于二进制的粗糙集运算理论，并借助二进制的位操作对粗糙集进行运算，给出几种基于二进制的粗糙集运算的算法.

1　粗糙集理论基础[8]

设 R 是 U 上的一个等价关系，U/R 表示 R 的所有等价类（或者 U 上的分类）构成的集合，$[x]_R$ 表示包含元素 $x \in U$ 的 R 等价类.

定义 1　论域 U 是有限集，R 是 U 的等价关系簇，则 $K = (U,R)$ 称为知识库.

定义 2　设知识库 $K = (U,R)$，有 $P \subseteq R$ 且 $P \neq \Phi$，则 $\cap P$（P 中全部等价关系的交集）也是一种等价关系，称其为 P 上的不可区分关系，记为 $IND(P)$. $IND(P)$ 中的每个集合称为基本范畴.

若 2 个对象分别处于 $IND(P)$ 的不同划分中，那么它们可以为现有的知识所区分；若两个对象处于同一个划分中，则它们不能为现有的知识所区分. 令 $X \subseteq U$，R 为 U 上的一个等价关系.

收稿日期：2005-12-30

作者简介：李天志（1977-），男，山东德州人，讲师，主要从事数据挖掘、数据结构和算法研究。

* 国家自然科学基金（批准号：60564001）和广西“十百千人才工程”专项基金（2003207）联合资助。

定义 3 当 X 能表达成某些 R 基本范畴的并时，称 X 是 R 可定义的，否则称 R 是不可定义的. R 可定义集也称为 R 精确集，R 不可定义集也称为 R 粗糙集.

定义 4 粗糙集的下近似集：$\underline{R}X = \cup \{Y | Y \in U/R$ 且 $Y \subseteq X\}$.

定义 5 粗糙集的上近似集：$\overline{R}X = \cup \{Y | Y \in U/R$ 且 $Y \cap X \neq \Phi\}$.

定义 6 集合 $BN_R(X) = \overline{R}X - \underline{R}X$ 称为 X 的 R 边界域.

定义 7 集合 $POS_R(X) = \underline{R}X$ 称为 R 的正域.

定义 8 集合 $NEG_R(X) = U - \overline{R}X$ 称为 X 的 R 负域.

定义 9 $\alpha_R(X) = \frac{|\underline{R}X|}{|\overline{R}X|}$，称为由等价关系 R 定义的集合 X 的近似精确度，其中 $X \neq \Phi$，$|X|$ 表示集合 X 的基数.

由上面定义知，$POS_R(X)$ 表示由那些根据知识 R 判断肯定属于 X 的 U 中元素组成的集合；$NEG_R(X)$ 表示由根据知识 R 判断肯定不属于 X 的 U 中元素组成的集合；$BN_R(X)$ 表示由那些根据知识 R 既不能判断肯定属于 X 又不能肯定不属于 X 的 U 中元素组成的集合；精度 $\alpha_R(X)$ 用来反映了解集合 X 的知识的完全程度.

2 二进制与粗糙集的内在联系

二进制通过 0、1 的组合，可以表示任意的数据，而且基于二进制的运算，在计算机中运算速度最快.

我们知道，n 位二进制数所表示的数据个数与具有 n 个元素的集合的所有子集的数目都是 2^n，因此，可以用二进制数来表示集合的子集. 为此，可以规定：对于集合 $U = \{i_0, i_1, \cdots, i_{n-1}\}$，其子集 A 与一个 n 位二进制数 $P = p_{n-1}p_{n-2}\cdots p_1p_0$ 相对应，其中：p_k 为 0 表示 A 中没有元素 i_k，p_k 为 1 表示 A 中有元素 $i_k (k = 0, 1, \cdots, n-1)$.

例 1 对于集合 $U = \{a, b, c\}$，其子集与二进制数的对应关系为表 1 所示的对应关系.

定义 10 集合 U 的子集 A 所对应的二进制数的十进制数值称为子集 A 的下标，记为 $[A]$.

集合 U 的 2 个子集的交与并可以通过其对应子集下标分别进行与运算、或运算而得到. 例如：$A = \{a, b\}$，$B = \{a, c\}$，则 $[A] = (011)_2$，$[B] = (101)_2$，因为 $[A]\&[B] = (011)_2\&(101)_2 = (001)_2$，$[A]|[B] = (011)_2|(101)_2 = (111)_2$，所以 $A \cap B = \{a\}$，$A \cup B = \{a, b, c\}$.

表 1 集合 $l = \{a, b, c\}$ 的子集与二进制数的对应关系

Table 1 Relation between binary and the subsets of $l = \{a, b, c\}$

序号 Number	子集 Subset	二进制数 Binary
0	Φ	000
1	$\{a\}$	001
2	$\{b\}$	010
3	$\{a, b\}$	011
4	$\{c\}$	100
5	$\{a, c\}$	101
6	$\{b, c\}$	110
7	$\{a, b, c\}$	111

集合 U 的 2 个子集的包含关系也可以通过其子集下标来判断. 我们有，$A \subseteq B$ 当且仅当 $[A]\&[B] = [B]$.

对知识库 $K = (U, R)$，关于 R 的等价类，可以用等价类对应的子集下标来表示.

例 2 给定知识库 $K = (U, R)$ 和一个等价关系 $R \in IND(K)$，其中 $U = \{x_0, x_1, \cdots, x_{10}\}$，且有 R 的下列等价类：$E_0 = \{x_0, x_1\}$，$E_1 = \{x_2, x_6, x_9\}$，$E_2 = \{x_3, x_5\}$，$E_3 = \{x_4, x_8\}$，$E_4 = \{x_7, x_{10}\}$.

利用子集下标表示的等价类分别为：

$[E_0] = (00000000011)_2 = 3$

$[E_1] = (001001000100)_2 = 580$

$[E_2] = (000000101000)_2 = 40$

$[E_3] = (000100010000)_2 = 272$

$[E_4] = (010010000000)_2 = 1152$

关于集合的其它运算也可以用其子集下标的运算来实现，在此不再列述. 所有这些，为我们定义基于二进制的粗糙集运算提供了理论依据.

3 基于二进制的粗糙集基本运算

利用二进制的位运算可以快速、高效地实现求集合 X 的上近似集、下近似集和近似精度等. 为此，需要定义相关的数据结构如下.

设常量 N 为知识库中的元素个数；

```
typedef struct byte //定义知识库中等价类或子集下标
{
  unsigned bit:N;
}Byte;
typedef struct roughest //等价类结点
 {
   int n;   //等价类的个数
   Byte subset[N];   //等价类对应的子集
```

```
}Roughset;
```

3.1 粗糙集的上近似集

由粗糙集的上近似集公式 $\overline{R}X=\cup\{Y|Y\in U/R$ 且 $Y\cap X\neq\Phi\}$ 可知，$\overline{R}X$ 为 R 的所有等价类中与 X 的交非空的等价类的并. 而集合的交可由其子集下标进行与运算得到，若该值为 0，则说明交集为空集. 于是，将与集合 X 交为非空的集合的子集下标进行或运算即可得到 $\overline{R}X$ 对应的子集下标. 求上近似集的具体算法如下.

```
Byte upperapprox(Roughset R,Byte X,Roughset &M)  //M 记录上近似集计算信息，返回值为上近似集的对应子集下标
{ int i;
  Byte u; //u 记录上近似集下标
  u.bit=0;
  for(i=0;i<R.n;i++)
  {
    if((R.subset[i].bit&X.bit)>0)
      { M.subset[M.n++].bit=R.subset[i].bit;//记录归并的等价类下标
        u.bit=u.bit|R.subset[i].bit;
      }
  }
  return u;
}
```

3.2 粗糙集的下近似集

由粗糙集的下近似集公式 $\underline{R}X=\cup\{Y\in U/R|Y\subseteq X\}$ 可知，$\underline{R}X$ 是 R 的所有等价类中包含于 X 的等价类的并. 集合的包含关系可由其子集下标的运算而得到，即 $A\subseteq B$ 当且仅当 $[A]\cup[B]=[B]$. 求下近似集的具体算法如下.

```
Byte lowerapprox(Roughset R,Byte X,Roughset &M)//返回值为下近似集对应的子集下标
{ int i;
  Byte u;
  u.bit=0;
  M.n=0;
  for(i=0;i<R.n;i++)
  {
    if ((R.subset[i].bit|X.bit)==X.bit)
    {
      M.subset[M.n++].bit=R.subset[i].bit;
      u.bit=u.bit|R.subset[i].bit;
    }
  }
  return u;
}
```

3.3 等价类的交

设 (U,A) 是一个信息系统，$U/\{a\}$、$U/\{b\}$、$U/\{a,b\}$ 分别为由属性 $\{a\}$，$\{b\}$ 及 $\{a,b\}$ 得到的等价类集合. 由等价划分定义知，$U/\{a,b\}$ 可由 $U/\{a\}$ 与 $U/\{b\}$ 中元素求交而得到. 于是，由 $U/\{a\}$、$U/\{b\}$ 求得 $U/\{a,b\}$. 具体算法如下.

```
Roughset intersection(Roughset V1,Roughset V2)
{
  Roughset V;//记录运算结果
  Byte t;
  int i,j;
  V.n=0;
  for(i=0;i<V1.n;i++)
  for(j=0;j<V2.n;j++)
  {
   t.bit= (V1.subset[i].bit)&(V2.subset[j].bit);
  if(t.bit>0)//交集不为空
  V.subset[V.n++]=t;
  }
  return V;
}
```

3.4 集合的基数

在计算近似精确度 $\alpha_R(X)$ 以及其它很多运算时，需要计算集合的基数，即元素个数. 由集合对应的子集下标可以非常容易地计算其基数，即子集下标二进制表示中“1”的个数，该值可以利用移位运算来得到. 具体算法如下.

```
int elementnumber(Byte element) //本算法稍加改动可以实现打印集合中所有元素
{int sum;
  sum=0;
  while(element.bit>0)
    {if((element.bit&1)>0)//当前值的最后一位为 1
      sum++;
      element.bit=element.bit>>1;
    }
  return sum;
}
```

关于粗糙集中其它运算的算法，可以仿照上述算

法给出.

3.5 粗糙集运算举例

例 3 在例 2 中，设 $X=\{x_0,x_3,x_4,x_5,x_8,x_{10}\}$，按照前面的算法计算可得：

$[X]=(10100111001)2=1337$

$[\overline{RX}]=\text{upperapprox}(R,X,M)=1467$

$[\underline{RX}]=\text{lowerapprox}(R,X,M)=312$

$[BN_R(X)]=1467-312=1155$

$[NEG_R(X)]=2047-1467=580$

$$\alpha_R(X)=\frac{|\underline{RX}|}{|\overline{RX}|}=\frac{4}{8}.$$

由此可以看出，基于二进制的粗糙集基本运算方便，快捷.

4 结束语

通过讨论二进制与粗糙集之间的内在联系，我们提出了基于二进制的粗糙集运算理论，并借助二进制的位运算操作，给出了几种基于二进制的粗糙集运算的算法，其算法的时间复杂度主要与等价类的个数有关. 由于算法中没有等价类元素的比较且主要是位运算，所以，与传统的算法相比，其运算速度更快，效率更高. 该算法的思想，也可以用于计算决策属性的支持度、依赖度等，为扩展粗糙集的应用提供了理论基础. 该算法思想已在数据挖掘的分类、决策分析中得到应用，取得了较好的效果. 基于二进制的粗糙集集运算算法在其它领域的应用有待于进一步研究.

参考文献：

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[2] PAWLAK Z. Rough classification[J]. Int J Man-Machine Studies, 1984, 20(5): 469-483.

[3] 程岩. 数据库中挖掘决策偏好信息的粗糙集方法研究[J]. 计算机工程, 2003, 29(6): 14-16.

[4] 赛煜. 一种基于粗糙集理论的最简规则挖掘方法[J]. 计算机工程, 2003, 29(20): 77-79.

[5] 王兵, 陈善本, 林涛. 粗糙集在智能系统知识维护中的应用[J]. 机器人, 2001, 23(7): 667-670.

[6] 陈文林, 郝丽娜, 徐心和. 粗糙集-神经网络-专家系统混合系统及其应用[J]. 计算机工程, 2003, 29(9): 147-178.

[7] 徐凤生, 李天志. 命题逻辑中的数字表示[J]. 德州学院学报, 2004, 20(2): 46-48

[8] 张文修. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.

(责任编辑：邓大玉)

新疫苗降低宫颈癌发病率

宫颈癌发病率仅次于乳腺癌，居全世界女性恶性肿瘤的第二位。但与其他许多恶性肿瘤不同的是，宫颈癌的病因现在已经明确是人乳头瘤病毒(HPV)感染。目前已经确定的 HPV 类型大约有 110 余种之多，其中尤以 HPV16 和 HPV18 为高危类型。HPV 通过性生活传播，感染后通常没有症状。在大多数国家，HPV 感染很常见，感染的高峰年龄在 18～28 岁。大部分妇女 HPV 感染期比较短，一般在 8～10 个月左右便可消失，但仍有大约 10%～15%的 35 岁以上的妇女有持续感染的情况。这些持续感染 HPV 的妇女则成为宫颈癌的高危人群。

对于 HPV 感染引起的宫颈癌可以通过接种疫苗来预防。黎巴嫩 Dartmouth 医学院 Harper 等研究人员的一项研究结果表明，一种新型的 2 价 HPV 16 和(或)18 L1 病毒样颗粒 AS04 疫苗具有类似于病毒的良好的抗原性和免疫原性，可以刺激机体产生中和抗体，而不含 DNA，不会在机体内引发因为接种疫苗而导致的病毒感染，在接种四五年后仍对宫颈损伤具有较高的防护作用，接种这种疫苗有助于减少宫颈癌的发病率。但是也有研究人员认为，有关这种新型宫颈癌疫苗在大样本人群中的预防效果还有待研究。

(据《科学时报》)