# 基于粗集的最小规则集提取算法研究

鲍松堂

(五邑大学 信息学院,广东 江门 529020)

**摘 要**：粗集理论是在数据分析中对于具有不精确、模糊和不确定性进行分析、处理的一种数学理论。从该理论的基础原理出发,运用支持子集相对于决策的分类能力,提出一种最小规则集的提取算法,并给出例子分析算法过程,表明其有效性。

**关键词**：粗糙集;支持子集;最小规则集

**中图分类号**：TP312 **文献标识码**：A

## Algorithm to minimal induction of decision rules based rough set

BAO Song Tang

(School of Informatics, Wuyi University, Jiangmen 529020, China)

**Abstract**: Rough sets theory is a tool for data mining, deal with the vagueness and granularity and uncertainty. After presenting their basic properties, this paper extends concepts of support subset in classification of objects, puts forward an algorithm to minimal induction of decision rules. And experimentation makes clear the validity of this new algorithm by using examples.

**Key words**: rough sets; support subset; minimal rules sets

粗集理论是由波兰华沙理工大学 PAWLAK Z 教授[1-2]于 1982 年提出的,主要研究不完整数据、不精确知识的表达、学习、归纳等方法。从新的视角对知识进行了定义,将知识看作是关于论域的划分,并引入代数中的等价关系来讨论知识,为智能信息处理提供了有效的处理技术。目前已经在人工智能、机器学习与知识发现、模型识别、分类、故障诊断等方面得到了较成功的应用。

属性约简和规则提取是粗集研究的重要内容。基于粗集方法的规则抽取过程是规则简化的过程,以这样的方法决策可使用条件属性的最小集合来确定。由于冗余属性往往会降低数据挖掘结果的精度和解释能力,属性约简是为了去除信息表中的冗余条件属性,并为得到一个较好的规则集做准备。由于目前算法所生成的规则过多(包含许多无用规则),不利于决策。参考文献[4]介绍了一种基于粗集的最小规则集提取算法,但其无法导出包含所有实例的有效性规则。参考文献[5]是一种改进的规则集提取算法,然而算法过程繁琐,在添加原子时太过单一。所以本文借用参考文献[3]中支持子集的选取方法选出规则,并且在此基础上提出了新的最小规则集提取算法。

## 1 准备知识

设 $U$ 为非空的论域,$R$ 是 $U$ 上的等价关系。参考文献[6]中将 $R$ 称为不可区分关系,因而在 $U$ 上产生一个分类 $U/R=\{Y_1,Y_2,\cdots,Y_m\}$,$Y_1,Y_2,\cdots,Y_m$ 是通过等价关系 $R$ 产生的等价关系类,也是关系 $R$ 上的元素集。

对于任何 $X\subseteq U$,通过关系 $R$ 的元素集和上、下近似来描述 $X$。

下近似 $\underline{R}(X)=\cup\{Y_i\in U/R\,|\,Y_i\subseteq X\}$,上近似 $\overline{R}(X)=\cup\{Y_i\in U/R\,|\,Y_i\cap X\neq\phi\}$。

对于决策表 $S=(U,C,D,f,V)$,$A=C\cup D$,对于每个 $u\in U$,定义一个函数 $r:\theta\rightarrow\varphi$。$r$ 称为决策表 $S$ 中的决策规则,$\theta$ 和 $\varphi$ 分别为决策规则 $\theta\rightarrow\varphi$ 的因和果。定义原子条件集 $M$,表示为 $M=\{(a,v)|\,\forall a\in C,\forall v\in V_a\}$。用 $C$ 来表示单一的原子条件,$\forall C\in M$。则 $\theta$ 可以表示为多个 $C$ 的交集,$\varphi$ 为对应的决策取值。

2 个属性 $a,b\in U$,需要计算论域 $U$ 的下面分类 $U/ab$:2 个对象 $u,v\in U$ 在同一类当且仅当 $a(u)=a(v)$ 且 $b(u)=b(v)$。对于属性集 $X\subseteq A$,按下面定义论域 $U$ 的分类:2 个对象 $a,b\in U$ 在同一类当且仅当对每个 $a\in X$ 有 $a(u)=a(v)$。

令 $W\subseteq U$ 是 $U$ 的子集，对于条件属性集 $X\subseteq C$，定义 $W$ 的下近似为 $\underline{W}(X)=\cup_{V\in U/X,V\subseteq W}V$；子集 $\underline{W}(X)$ 称为 $W$ 关于 $X$ 的支持子集，$spt_X(W)=|\underline{W}(X)|/|U|$ 称为 $W$ 关于 $X$ 的支持度；定义 $W$ 的上近似为 $\overline{W}(X)=\cup_{V\in U/X,V\cap W\neq\phi}V$。

## 2 最小规则集提取算法

如果对于属性 $a\in U$ 有 $\underline{W}(a)\neq\phi$，则必有 1 个或多个 $a=v$，使其包含的实例是 $W$ 的一个子集，则有规则的因 $a=v$ 可导出 $W$ 对应的 $\varphi$ 的决策取值。单个的原子条件用 $C$ 表示。同理如果支持子集 $\underline{W}(X)\neq\phi$，则由这个子集就可以导出决策，其中决策的因是条件属性集 $X\subseteq C$ 中符合支持子集条件的原子条件集。故而有下面的最小规则集提取算法。

输入：输入决策表 $S=(U,C,D,f,V)$，$U=\{u_1,u_2,\cdots,u_n\}$，$C=\{a_1,a_2,\cdots,a_m\}$ 是条件属性集，$D$ 是决策属性集，$U/D=\{Y_1,Y_2,\cdots,Y_k\}$。

输出：决策表 $S$ 的最小规则集。决策类 $Y_1,Y_2,\cdots,Y_k$ 对应的决策属性 $d$ 的属性值分别为 $v_1,v_2,\cdots,v_k$；$R$ 为规则集，$C$ 表示原子条件，$[C]$ 表示决策表中该原子条件所覆盖的实例集合。

```
i=1,j=1,U'=U
While(j≤k) Do
  {if [Y_j]≠φ,
    {While(i≤m) Do
      { ∀C_1,C_2,…,C_i∈M
        令 β=[C_1]∩[C_2]∩…∩[C_i]∩U'⊆Y_j,
        选取 1 组元素最多的|β|(如果元素最多的不止 1 组，则选取最先出现的进行计算)。
        If β≠φ,
          {产生 1 条规则 r:C_1,C_2,…,C_i→d=v_j,
           R=R∪r,
           U'=U'-β,
           Y_j=Y_j-β}
        Else i=i+1
      }
    }
  Else j=j+1
}
```

## 3 实例分析

决策表如表 1 所示，条件属性集 $C=\{a_1,a_2,a_3,a_4,a_5\}$，决策属性集 $D=\{d\}$。

表 1 决策表

| $U$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 2 | 3 | 1 | 3 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 3 | 1 | 1 | 3 |
| $a_2$ | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| $a_3$ | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 1 | 2 | 3 | 2 | 1 | 3 | 3 | 2 | 3 |
| $a_4$ | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 1 |
| $a_5$ | 3 | 3 | 2 | 1 | 1 | 2 | 2 | 3 | 1 | 2 | 2 | 3 | 1 | 2 | 2 | 1 |
| $d$ | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |

算法在实例中的运行过程如下：

$Y_1$:$\{(a_1,2)\}=\{1,8\}$；$\{(a_3,1)\}=\{1,3,6,8,12\}$选中

$\{(a_1,1),(a_4,1)\}=\{7,14\}$选中；$\{(a_2,2),(a_4,2)\}=\{15\}$；

$\{(a_2,2),(a_5,2)\}=\{14,15\}$；$\{(a_4,1),(a_5,2)\}=\{7,14\}$

$\{(a_2,2),(a_4,2)\}=\{15\}$选中；$\{(a_2,2),(a_5,2)\}=\{15\}$；

$\{(a_3,2),(a_4,2)\}=\{10\}$；$\{(a_3,2),(a_5,2)\}=\{10\}$

$\{(a_3,3),(a_4,2)\}=\{10\}$选中；$\{(a_3,3),(a_5,2)\}=\{10\}$

$Y_2$:$\{(a_1,3)\}=\{2,4,9,13,16\}$选中

$\{(a_1,1)\}=\{5,11\}$选中

出的规则为：

$\{(a_3,1)\}\rightarrow d=1$，$\{(a_1,1),(a_4,1)\}\rightarrow d=1$，$\{(a_2,2),(a_4,2)\}\rightarrow d=1$，$\{(a_3,3),(a_4,2)\}\rightarrow d=1$，$\{(a_1,3)\}\rightarrow d=2$，$\{(a_1,1)\}\rightarrow d=2$

如在算法中加入输出规则覆盖的实例和支持度，与上述规则对应的实例和支持度则分别为：

{覆盖实例：1，3，6，8，12。支持度：31.25%}{覆盖实例：7，14。支持度：12.5%}{覆盖实例：15。支持度：6.25%}{覆盖实例：10。支持度：6.25%}{覆盖实例：2，4，9，13，16。支持度：31.25%}{覆盖实例：5，11。支持度：12.5%}

本文通过分析粗集中支持子集的计算，结合最小规则集的提取过程，提出一种新的最小规则集提取算法。算法相对参考文献[4-5]，过程简单，规则提取完毕后不用再进行约简，通过实例证明了，在其协调决策系统中最小规则提取运行的有效性。

## 参考文献

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

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

[3] 张文修.粗糙集理论与方法[M].北京：科学出版社，2000.

[4] STEFANOWSKI J. On rough sets based approaches to induction of decision rules [A]. Rough sets in knowledge discovery[C]. Heidelbery: Physica Verlag. 1998:500-529.

[5] 吴顺祥.基于粗集理论的一种规则提取方法[J].厦门大学学报，2004(9):64-66.

[6] PAWLAK Z. Rough sets: Theoretical aspects of reasoning about data[M]. Boston: Kluwer Academic Publishers,1991.

(收稿日期：2009-09-01)

作者简介：

鲍松堂，男，1977 年生，硕士研究生，主要研究方向：嵌入式软件，粗集理论。