第34卷 第10期 计 算 机 工 程 2008年5月
Vol.34 No.10 Computer Engineering May 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)10—0166—02 文献标识码：A 中图分类号：TP18

基于粗糙集的确定性控制规则决策模型

邓九英[1,2]，毛宗源[1]，杜启亮[1]，谭光兴[1]
(1. 华南理工大学自动化科学与工程学院，广州 510641；2. 广东教育学院计算机科学系，广州 510303)

摘 要：粗糙集属性分区数的变化会影响属性重要性和属性对决策属性的支持度。该文对知识表示系统的数据相关性进行分析，综合考虑系统的泛化能力，提出能生成确定性控制规则的决策模型，给出决策模型中属性分区数求取以及属性相对约减产生的判据与算法实现。实验结果表明，该算法简洁有效，验证了决策模型的准确性与实用性。

关键词：决策模型；确定性控制规则；属性分区；属性重要性；支持度

Decision Model of Certain Control Rules Based on Rough Set

DENG Jiu-ying[1,2], MAO Zong-yuan[1], DU Qi-liang[1], TAN Guang-xing[1]
(1. College of Automatic Science & Engineering, South China University of Technology, Guangzhou 510641;
2. Department of Computer Science, Guangdong Institute of Education, Guangzhou 510303)

【Abstract】 Different attributes partitions affect about attribute significance and attributes support degree to decision attribute. This paper analyzes the relationship among data in knowledge represent system based on rough set. Generalization of system is combined to consider. A decision model is presented for generating certain control rules. Judgment and algorithm are introduced to compute attributes partitions in decision model and find relative attributes reduction in decision model. Experimental results show conciseness and efficiency of algorithms and preciseness and utility of decision model.

【Key words】 decision model; certain control rules; attribute partition; attribute significance; support degree

1 概述

粗糙集方法是人工智能(AI)和认知学的基础，也是机器学习、知识获取、决策分析和从数据库中发现知识等领域的基本工具，在决策支持系统和数据挖掘方面粗糙集显得格外重要。在粗糙集方法用于数据分析时，数据模式特征用近似的方法表示，或等价于从数据中归纳出的决策规则[1]。科研人员围绕决策规则做了大量的研究工作，包括用粗糙集评价重要规则的方法[2]、选择具有高特征的属性约减子集算法和获取具有泛化能力的简洁规则策略[3]等。

通过知识获取生成决策规则的质量，与规则应用到实践中的控制结果的好坏有直接关系[4]，对生成与选择决策规则质量的评判依据包括几方面，其中主要的一点是规则的准确性；另外规则的简洁性直接关系到规则的实用性。由于属性的分区数(量化值的个数)变化会影响属性的重要性和属性对决策属性的支持度，因此本文提出一种构建确定性决策规则的模型。

2 决策表

2.1 定义与性质

决策表可用来定义知识表示系统(Knowledge Representation System, KRS)，设$S=(U, A)$为知识表示系统，$C,D\subset A$是2个属性子集，分别为条件属性与决策属性；KRS也可表示为$S=(U, A, C, D)$。关系$IND(C)$和$IND(D)$的等价类分别称为条件等价类与决策等价类[5]。其中，对于$B\in A$，$IND(B)=\{(x,y)\in U\times U:$ 对任意$a\in B, a(x)=a(y)\}$。

对任意$x\in U$，存在函数：$A\to V$，以及对任意$a\in C\cup D$，有$des_x(a)=a(x)$；则函数des_x称作(S中的)决策规则。限于对C的des_x(表示为$des_x|C$)和限于对D的des_x(表示为$des_x|D$)分别称为des_x的条件与决策(或动作)。

如果对任意$y\neq x$，$des_x|C=des_y|C$蕴涵$des_x|D=des_y|D$，那么决策规则des_x是(在S中)一致的；否则，决策规则是不一致的。如果所有的决策规则是一致的，那么决策表是一致的；否则，决策表是不一致的。如果决策表是不一致的，只要删除x(或des_x)就可把决策表变为一致。

2.2 决策属性的支持度

为了计算某个属性(或属性集)的重要性，常见的方法是：从表中去除这个属性，看属性去除前后分类的变化情况，如果属性删除后的分类变化大，就说明这个属性的重要性高；反之，这个属性的重要性低。

定义 1 假设一给定知识库$S=(U, R)$，子集$X\subseteq U$和等价关系$R\in IND(K)$(缩写为$R\in K$)，R-下近似集表示为：$\underline{R}X=\cup\{Y\in U/R, Y\subseteq X\}$ or $\underline{R}X=\{x\in U, [x]_R\subseteq X\}$。

另一种表示形式为：$POS_R(X)=\underline{R}X$，X的R-正域。

定义 2 决策属性$y\in D$，决策子集$W\in U/y$，决策属性y关于条件属性$a\in C$的支持子集定义为

$$S_a(y)=\cup_{W\in U/y}POS(U/a)=\cup_{W\in U/y}(\cup_{V\in U/a,\ V\subseteq W}V) \quad (1)$$

y关于a的支持度表示为$spt_a(y)=|\cup_{W\in U/y}POS(U/a)|/|U|$。

令$W\subseteq U$为U的子集，决策表$(U, C\cup D)$，W关于条件属性

作者简介：邓九英(1962－)，女，副教授，主研方向：人工智能，数据挖掘技术，仿真技术；毛宗源，教授、博士生导师；杜启亮、谭光兴，博士研究生

收稿日期：2007-06-28 **E-mail：**djy1111@126.com

集 $X\subseteq C$ 的支持子集可表示为 $S_X(W)=\cup_{V\in U/X,V\subseteq W}V$。$spt_X(W)=|\cup_{W\in U/X}POS(U/X)|/|U|$称为$W$关于$X$的支持度。

3 核属性与决策规则

设决策表$S=(U, A, V, f)$, $A=C\cup D$, $C\cap D=\phi$。令X_i和Y_j分别代表U/C与U/D中的等价类，des_{X_i}是等价类X_i的描述，即等价类X_i对于各条件属性值的特定取值；des_{Y_j}是等价类Y_j的描述，即等价类Y_j对于各决策属性值的特定取值[6]。

定义 3 决策规则定义为r_{ij}：$des_{X_i}\rightarrow des_{Y_j}$，$Y_j\cap X_i\neq\phi$。规则的确定性因子：

$$\mu(X_i,Y_j)=|Y_j\cap X_i|/|X_i|,\ 0<\mu(X_i,Y_j)\leq 1 \tag{2}$$

如果$\mu(X_i,Y_j)=1$，则r_{ij}是确定的；如果$0<\mu(X_i,Y_j)<1$，则r_{ij}是不确定的。在得出决策规则之前，决策表需要进行属性约减。

假设知识表示系统$S=(U, A, V, f)$, $n=|U|$。系统S的区分矩阵是一个$n\times n$维的矩阵，其中的元素为$a(x, y)=\{a\in A,\ |f(x, a)\neq f(y, a)\}$。如果$a(x, y)=\{a_1, a_2,\cdots, a_k\}\neq\phi$，用符号$\Sigma a(x, y)$表示布尔函数$a_1\vee a_2\vee\cdots\vee a_k$；如果$a(x, y)=\phi$，则取值为布尔常量 1。区分函数$\Delta$定义为

$$\Delta=\prod_{(x, y)\in U\times U}\sum a(x, y) \tag{3}$$

如果 $B\subseteq A$ 是一个满足条件 $B\cap a(x, y)\neq\phi$, $\forall a(x, y)\neq\phi$的极小子集，则 B 就是 A 的一个约减。

核属性是由区分函数中单个因子组成的集合，表示如下：

$$core(A)=\{a\in A,\ |a(x, y)=\{a\},\ x, y\in U\} \tag{4}$$

4 确定性决策规则模型

为了使得到的决策规则简短而又准确，构建 KR-system 决策模型的主要任务包括 2 个方面：(1)确定属性的分区数；(2)进行属性约减，在满足相对约减子集对决策属性的支持度趋于 1(或等于 1)的条件下，使相对约减子集中的元素个数较少。

由于属性的量化分区数的不同会改变属性的重要性参数，如果要生成确定性决策规则，应确定使属性对决策属性的支持度为用户给定数值的临界(最小)分区数[4]，属性分区数的取值应大于这个临界值。从系统的泛化能力考虑，属性约减子集中的元素个数应大于某基数(视系统的具体情况来定)。本文在此基础上提出 2 个判据和相应的 2 个算法，分别用于确定决策模型中的属性分区数和属性约减子集。

假定知识表示系统不包含冗余与不相容数据，属性取值为连续量，对具有固定明确状态值的属性，不适合用以下的判据与算法。

判据 1 属性分区的确定

(1)选取每个条件属性$a_i\in C$的分区数临界值，$divnum(a_i)$；

(2)设定条件属性C的分区数初值，$divC=\max\{a_i, |a_i\in C\}$；

(3)设定决策属性 D 的分区数初值，$divD$；

(4)循环执行，$divC+1\rightarrow divC$, $divD+1\rightarrow divD$，属性分区量化取值后，求 $Core$ 中的元素个数 sum，直至 sum 大于给定基数；

(5)循环执行，$divC+1\rightarrow divC$, $divD+1\rightarrow divD$，属性分区量化取值后，计算区分函数，直至区分函数中只包含核属性因子；

(6)后推计算，$divC-1\rightarrow divC$, $divD-1\rightarrow divD$，属性分区量化取值后，计算区分函数，直至区分函数中出现了除核属性外的其他属性因子；

(7)停止。

得出的 $divC$ 和 $divD$ 分别是系统 S 中条件属性与决策属性的分区数。

判据 2 属性相对约减子集的确定

(1)设定属性约减子集 X 为属性核，计算 X 对决策属性的支持度 $spt(X)$；

(2)在几个约减方案中选择a_j，并计算a_j+X对决策属性的支持度$spt(a_j+X)$，使$spt(a_j+X)-spt(X)$为最大的a_j加入X；

(3)重复(2)，直至 $spt(X)=1$；

(4)停止。

在 KRS 具有约减属性及其明确的属性数字量后，应用粗糙集理论的决策算法，可从 KRS 中抽取决策规则，再对决策规则进行约减，得出实用的精简决策规则。

5 决策模型的算法实现

5.1 决策模型的相应算法

假定在系统$S=(U, A, V, f)$的数据预处理阶段，完成了每个属性$a_i\in C$对决策属性有极大支持度的最小属性分区数$divnum(a_i)$的计算，运用算法 1 可计算出确定性规则的最小属性分区数，运用算法 2 可计算出系统的属性约减的合理子集。

引入一些符号表示参数：$n=Card(C)$, $m=Card(Core)$, $Ns=|U|$, $sum=3$(泛化能力基数), $Card(D)=1$。

算法 1 计算属性分区数(AAP)

AAP 1: 赋初值, divnum(i), i=1,2,⋯,n;
AAP 2: divC=Max{divnum(i), i=1,2,⋯,n};
AAP 3: 赋初值, divD(比如 4), sum(比如 3), m=1;
AAP 4: Do while m≤sum
{divC+1→divC, divD+1→divD, 计算 Core, Card(Core)→m};
AAP 5: Do while m<>Card(Δ)
{divC+1→divC, divD+1→divD, 计算Δ与 Core, Card(Core)→m};
AAP 6: Do while m=Card(Δ)
{divC−1→divC, divD−1→divD, 计算Δ与 Core, Card(Core)→m};
AAP 7: End

将系统 S 按照分区数 $divC$ 和 $divD$ 对属性进行量化取值。

算法 2 求属性相对约减子集(AADS)

AADS 1: Core→X, 计算 spt(X);
AADS 2: K=Card(Δ)−Card(X);
AADS 3: Do while spt(X)<>1 and K<>0
{对所有{Δ−X}a_j, j=1,2,⋯,K, X+a_j→X′, 找出spt(X′)−spt(X)值最大的a_j;
X+a_j→X; 计算spt(X)};
AADS 4: End

这时，只包含 X 属性子集的系统 S，可用于归纳出决策规则。

5.2 实验结果

选取广州某化工厂锌钡白煅烧过程实测数据，采样时间为 2004-2-3 12:35 PM至 2004-2-7 10:33 AM，每 10 min采样一次数据，采样数据一共分为 19 个属性。剔除异常与无关数据，保留了 13 个属性与 200 组数据，其中，条件属性用符号表示：C_1为窑头温度(ytwd)，C_2为干燥温度(gzwd)，C_3为排风温度(pfwd)，C_4为#1 进料电流(jldl1)，C_5为#1 进料频率反馈(jlfreq1)，C_6为#2 进料电流(jldl2)，C_7为#2 进料频率反馈(jlfreq2)，C_8为#3 进料电流(jldl3)，C_9为#3 进料频率反馈(jlfreq3)，C_{10}为#4 进料电流(jldl4)，C_{11}为#4 进料频率反馈(jlfreq4)，C_{12}为立德粉粉种(fenkind)；决策属性，d为煅烧温度(dswd)。

决策属性 d 的分区数初始值为 5，经过数据预处理，得出条件属性对决策属性支持度给定数值时的临界分区数如表 1 所示。 (下转第 170 页)