基于粒计算的不完备信息系统的规则提取方法

于海燕[1,2],王道平[1],张 霞[2,3]
YU Hai-yan[1,2],WANG Dao-ping[1],ZHANG Xia[2,3]

1.北京科技大学 经济管理学院,北京 100083
2.河北经贸大学 计算机中心,石家庄 050061
3.北京科技大学 信息工程学院,北京 100083
1.School of Economics and Management,University of Science and Technology Beijing,Beijing 100083,China
2.Computer Center,Hebei University of Economics and Business,Shijiazhuang 050061,China
3.School of Information and Engineering,University of Science and Technology Beijing,Beijing 100083,China
E-mail:yuhyan68@163.com

YU Hai-yan,WANG Dao-ping,ZHANG Xia.Method for rule acquisition in incomplete information systems based on granular computing.Computer Engineering and Applications,2009,45(8):143-145.

Abstract: An algorithm of rule acquisition from incomplete information systems is proposed based on granular computing.The information table is decomposed to hierarchy.Certain rule will be acquired according to certain information in every hierarchy.Because the null value is uncertain,it is regarded not equal to any certain value in processing.
Key words: incomplete information systems;granular computing;knowledge acquisition

摘 要:根据粒计算思想提出了一种从不完备决策表中分层提取确定规则的方法,将决策表进行分解,然后按决策表提供的确定信息分层提取相应的确定规则,在处理过程中认为空值提供的信息是不可靠的,所以与确定值严格加以区分,该方法充分利用不完备信息系统中的确定信息,得到长度不等的确定规则。
关键词:不完备信息系统;粒计算;知识获取
DOI:10.3778/j.issn.1002-8331.2009.08.044 **文章编号**:1002-8331(2009)08-0143-03 **文献标识码**:A **中图分类号**:TP18

1 简介

粗糙集理论是一门处理不精确、不确定信息的数学理论,已经被成功应用于机器学习、人工智能、模式识别、智能信息处理等领域,尤其在处理大数据量,消除冗余信息等方面,RS 理论有着良好的效果。

最初的 RS 理论是面向完备信息系统的,对于不完备信息系统的处理目前主要有两种方法,一种是将不完备信息系统转化为完备信息系统,如去掉包含空值的对象或按某种度量方法将空值补齐,另一种是直接处理不完备信息系统,主要是对经典 Rough 集理论中相关概念在不完备信息系统下进行适当扩充,Krysckiewcz 提出了容差关系;Stefanowki 等人提出了非对称相似关系和量化容差关系;王国胤在 Krysckiewcz 和 Stefanowki 提出的容差关系和非对称相似关系基础上提出了介于两者之间的限制容差关系。基于这些扩充关系有些学者提出了相应的从不完备信息系统中进行规则提取的方法。

这些对不完备信息系统的处理方法都对原系统加入了人为的测算、估计,或多或少改变了原系统所提供的信息。本文提出的对不完备信息系统的规则提取方法中认为空值本身是不确定信息,所以认为空值和任何值都不相等,严格加以区分,对决策表根据粒计算思想提取确定规则。

粒计算是信息处理的一种新的概念和计算范式,覆盖了所有有关粒度的理论、方法、技术和工具的研究,现已成为人工智能领域研究的热点之一。所谓的信息粒,是指人类在解决处理和存储信息的有限能力上的一种反映,即人类在解决和处理大量复杂信息问题时,由于人类的能力有限,需把大量复杂信息按其各自的特征和性能将其划分成数个较简单的信息块,以方便处理,每个如此划分的信息块就被认为是一个粒[1]。Rough 集理论中用任意二元关系将全集 U 分类被称为粒化,被粒化得到的类被称为粒,用 Rough 集方法去处理这些粒被称为粒计算,而且是当前国际上最热门、最重要的研究学科,也是当前国际上最具有挑战性的处理信息的信息化技术。在全集 U 上经等价或不可区分关系的粒化,得到的粒是一种特殊的粒,即它们之间是独立、互不相交的[2]。本文基于粒计算思想将不完备决策表进行分解,充分利用决策表中的确定信息,分层提取确定规则。

2 基本概念

定义 1(决策表信息系统)[3] 一个决策表信息系统(简称决策表)$S=<U,R,V,f>$,其中,U 是对象的非空有限集合,也称为论域,$R=C\cup D$ 表示属性的非空有限集合,子集 C 和 D 分别称为

作者简介:于海燕(1968-),女,博士生,副教授,主要研究方向:智能信息处理;王道平,教授,博导;张霞,博士生,讲师。
收稿日期:2008-01-23 **修回日期**:2008-04-18

条件属性集和决策属性集，且 $C\cap D=\varnothing$；$V=\bigcup_{a\in A}V_a$，V_a 表示属性 a 的值域；f 表示 $U\times A\to V$ 的一个信息函数，它为每个对象在每个属性上赋予一个信息值，即 $\forall a\in A, x\in U, f(x,a)\in V_a$。若存在一个 $x\in U, a\in C, f(x,a)$ 未知（记作：$f(x,a)=*$），则称信息系统是不完备的；否则称信息系统是完备的。

定义 2（不可区分关系）一个决策表信息系统（简称决策表）$S=<U,R,V,f>$，对于每个属性子集 $P\subset R$，定义一个不可区分（Indiscernibility）关系 $IND(P)$，即

$$IND(P)=\{(x,y)|(x,y)\in U\times U, \forall p\in P, f_p(x)=f_p(y)\}$$

关系 $IND(P)$ 是一个等价关系，且构成了 R 的一个划分。

定义 3（条件分类和决策分类）[3] 给定不完备信息决策表 $S=<U,R,V,f>$，$R=C\cup D$，C 和 D 分别为决策表的条件属性和决策属性，$U|IND(C)$ 和 $U|IND(D)$ 分别为论域 U 在属性集 C 和 D 上形成的划分，条件分类定义为 $X_i\in U|IND(C)(i=1,\cdots,m,m$ 为分类的个数）；决策分类定义为 $X_j\in U|IND(D)(j=1,\cdots,n,n$ 为决策分类的个数）。

定义 4（公式）定义公式如下[4]：

（1）(a,v) 或写为 $a_v, a\in R, v\in V_a$，表示属性 a 的取值为 v 是原子公式；原子公式是公式。

（2）如果 A 和 B 是公式，那么 $\neg A, A\wedge B, A\vee B, (A), A\to B$ 都是公式。

（3）只有按定义（1）和（2）所组成的式子是公式。

定义 5（决策规则）公式 $A\to B$ 的逻辑含义称为规则前件，B 称为规则后件，它们表达一种因果关系。其中，公式 A 中所包含的原子公式中只有决策表中的条件属性，B 中所包含的原子公式中只有决策表中的决策属性。

定义 6（规则长度）给定决策表 $S=<U,R,V,f>$，$R=C\cup\{d\}$ 子集 C 和 $\{d\}$ 分别称为条件属性集和决策属性集，公式 $(a_1,v_1)\wedge(a_2,v_2)\wedge\cdots\wedge(a_n,v_n)$ 为基本公式，其中 $v_i\in V_{a_i}, \{a_1,a_2,\cdots,a_n\}\in C$，如果 A 是基本公式且 $B=(d,d_i)$，$A\to B$ 为决策规则。A 中原子公式的个数称为决策规则的长度。

定义 7（可信度）对于决策表 $S=<U,R,V,f>$，$R=C\cup D$ 是属性集合，子集 C 和 D 分别为条件属性集和决策属性集，决策规则 $A\to B$ 的可信度 $CF(A\to B)$ 定义为：

$$CF(A\to B)=\frac{|X\cap Y|}{|X|}$$

其中，集合 X 为条件属性值满足公式 A 的样本的集合，集合 Y 为决策属性值满足公式 B 的样本的集合。

3 规则提取算法

因为不完备决策表中含有空值，不能用传统的粗糙集规则提取方法，本文采用粒计算的思想对决策表进行分解，分层提取出不完备决策表中的确定规则。

首先，将决策表中条件属性按属性值中是否包含“*”值分为两部分，假设共有 m 个属性包含“*”值，则共分成 m 层，每一层由完备信息部分与不完备信息部分的属性子集中的 1 个属性、2 个属性、……、m 个属性组合，分别构成提取规则的属性子集。将决策表分解以后，在每一层里实际是按每个对象的部分属性值产生规则，在这部分信息里，如果某一对象含有不确定信息“*”值，则该对象被认为这部分属性不含有足够产生规则的确定信息，不参与分类，在按其他属性划分等价类时，该对象所含的信息如果是确定信息，它将按相应信息进行分类，用这样的方法，对包含“*”值的对象，只利用了它的确定信息部分。在每一层中按粗糙集等价类划分方法划分等价类后，提取可信度为 1 的决策规则。

用这样的方法，按不完备决策表中确定数据所提供的信息分层提取规则，挖掘出规则长度不等的确定规则。

3.1 算法 1

输入：约简后的不完备决策表 $S=(U,A=(C\cup D))$。

输出：规则集 RS。

步骤 1 令 $RS=\{\varnothing\}, CS=\{\varnothing\}, DS=\{\varnothing\}, m=1$。

步骤 2 计算按决策属性的划分 $X_j\in U|IND(D), j=1,\cdots,|U|IND(D)|$。

步骤 3 将决策表条件属性中按属性值是否包含“*”值分为两部分 A 和 B，A 为不包含空值的属性子集，B 为包含空值的属性子集 $B=\{a_1,a_2,\cdots,a_n\}$，其中 n 为包含空值的属性个数。

步骤 4 计算第 m 层的分解，$m=1,\cdots,n$，n 为包含“*”值属性的个数。$W^m=A\cup P^m$，其中 $P^m\in 2^B$ 且 m 表示 P 中的元素个数，即按属性子集 W^m 划分等价类 $X_i^m\in|U|IND(W^m)|, i=1,\cdots,|U|IND(W^m)|$。

划分等价类过程中如果对象相应的属性值包含“*”值，该对象不被划分到任何分类。

步骤 5 计算可信度：$CF(X_i^m\to X_j)=\frac{|X_i^m\cap X_j|}{|X_i^m|}$，其中 $i=1,\cdots,|U|IND(W^m)|, j=1,\cdots,|U|IND(D)|$。如果 $CF(X_i^m\to X_j)=1$，则 $\{RS=RS\cup\{X_i^m\to X_j\}\}$。

步骤 6 如果 $m=n$，算法停止，否则 $m=m+1$，转步骤 4。

3.2 实例分析

给定一个不完备决策信息表，如表 1。下面按算法 1 提取表 1 中的规则。

表 1 不完备决策信息表

	c	d	e	f
x_1	2	2	1	2
x_2	2	3	1	1
x_3	*	2	2	2
x_4	3	1	2	2
x_5	3	1	2	2
x_6	2	2	*	1
x_7	3	2	1	3
x_8	*	1	1	3

表中不包含 * 值的条件属性只有属性 $A=\{d\}$，包含“*”值的条件属性为 $B=\{c,e\}$。因为 B 中包含两个元素，即 $n=2$，所以共分解为两层。

第 1 层分解由属性 d 和属性 c、e 分别组合，按属性子集 $\{c,d\}$、$\{d,e\}$ 划分等价类，按决策属性 $\{f\}$ 划分等价类，结果如表 2。

表 2 第 1 层等价类划分结果

$(a_1,v_1)\wedge\cdots\wedge(a_n,v_n)$	$IND(C)$	(d,d_i)	$IND(D)$
$(c,2)\wedge(d,2)$	$\{x_1,x_6\}$	$(f,1)$	$\{x_2,x_6\}$
$(c,2)\wedge(d,3)$	$\{x_2\}$	$(f,2)$	$\{x_1,x_3,x_4,x_5\}$
$(c,3)\wedge(d,1)$	$\{x_4,x_5\}$	$(f,3)$	$\{x_7,x_8\}$
$(c,3)\wedge(d,2)$	$\{x_7\}$		
$(d,2)\wedge(e,1)$	$\{x_1,x_7\}$		
$(d,3)\wedge(e,1)$	$\{x_2\}$		
$(d,2)\wedge(e,2)$	$\{x_3\}$		
$(d,1)\wedge(e,2)$	$\{x_4,x_5\}$		
$(d,1)\wedge(e,1)$	$\{x_8\}$		

下面举例说明如何处理"*"值,如在按属性子集$\{c,d\}$划分等价类时,因为对象x_3和x_8包含空值,所以没有被划分到任何等价类中,也就是说这两个对象的这两个属性值没有提供足够的确定信息,所以在这里不参与计算,而在按属性d和e划分等价类时,对象x_3和x_8的属性d和属性e的值不是"*"值,所以按其属性值被划分到相应的等价类里。

对划分结果按定义7计算可信度,结果如表3。如果可信度为1,则提取规则。

表3　规则可信度

$(a_1,v_1)\wedge\cdots\wedge(a_n,v_n)$	$(f,1)$	$(f,2)$	$(f,3)$
$(c,2)\wedge(d,2)$	1/2	1/2	0
$(c,2)\wedge(d,3)$	1	0	0
$(c,3)\wedge(d,1)$	0	1	0
$(c,3)\wedge(d,2)$	0	0	1
$(d,2)\wedge(e,1)$	0	1/2	1/2
$(d,3)\wedge(e,1)$	1	0	0
$(d,2)\wedge(e,2)$	0	1	0
$(d,1)\wedge(e,2)$	0	1	0
$(d,1)\wedge(e,1)$	0	0	1

在第1层中得到如下规则:

$(c,2)\wedge(d,3)\rightarrow(f,1)$

$(c,3)\wedge(d,1)\rightarrow(f,2)$

$(c,3)\wedge(d,2)\rightarrow(f,3)$

$(d,3)\wedge(e,1)\rightarrow(f,1)$

$(d,2)\wedge(e,2)\rightarrow(f,2)$

$(d,1)\wedge(e,2)\rightarrow(f,2)$

$(d,1)\wedge(e,1)\rightarrow(f,3)$

第2层按属性d和任意两个包含"*"值的属性组成子集划分等价类,因为在该表中只有两个包含空值的属性c和e,所以按属性子集$\{c,d,e\}$划分等价类结果如表4。

表4　第2层等价类划分结果

$(a_1,v_1)\wedge\cdots\wedge(a_n,v_n)$	$IND(C)$	(d,d_i)	$IND(D)$
$(c,2)\wedge(d,2)\wedge(e,1)$	$\{x_1\}$	$(f,1)$	$\{x_2,x_6\}$
$(c,2)\wedge(d,3)\wedge(e,1)$	$\{x_2\}$	$(f,2)$	$\{x_1,x_3,x_4,x_5\}$
$(c,3)\wedge(d,1)\wedge(e,2)$	$\{x_4,x_5\}$	$(f,3)$	$\{x_7,x_8\}$
$(c,3)\wedge(d,2)\wedge(e,1)$	$\{x_7\}$		

对第2层划分结果计算可信度,结果如表5。

在第2层中得到规则如下:

表5　第2层规则可信度

$(a_1,v_1)\wedge\cdots\wedge(a_n,v_n)$	$(f,1)$	$(f,2)$	$(f,3)$
$(c,2)\wedge(d,2)\wedge(e,1)$	0	1	0
$(c,2)\wedge(d,3)\wedge(e,1)$	1	0	0
$(c,3)\wedge(d,1)\wedge(e,2)$	0	1	0
$(c,3)\wedge(d,2)\wedge(e,1)$	0	0	1

$(c,2)\wedge(d,2)\wedge(e,1)\rightarrow(f,2)$

$(c,2)\wedge(d,3)\wedge(e,1)\rightarrow(f,1)$

$(c,3)\wedge(d,1)\wedge(e,2)\rightarrow(f,2)$

$(c,3)\wedge(d,2)\wedge(e,1)\rightarrow(f,3)$

对于决策信息表1用该方法共得到了11条确定规则。

4　结论

已有不完备信息系统的研究方法都或多或少地改变了原信息,本文提出的不完备信息系统的规则提取方法认为空值是不确定信息,将空值与确定值严格区分,在规则提取中认为空值和任何确定值都不相等,对决策表进行分解,分层提取出确定规则,充分利用了决策表所提供的确切信息,得到了规则长度不等的确定规则。

参考文献:

[1] Yao Y Y.Granular computing:Basic issues and possible solutions[C]//Paul P.Proceedings of the 5th Joint Conference on Information Sciences.USA:Elsevier Publishing Company,2000:186-189.

[2] 刘清,孙辉.从Rough集的发展前景看粒计算的研究趋势[J].南昌工学院学报,2006,25(5):1-10.

[3] 王国胤,何晓.一种不确定性条件下的自主式知识学习模型[J].软件学报,2003,14(6):1096-1102.

[4] Wang G Y.Rough set theory and knowledge acquisition[M].Xi'an:Press of Xi'an Jiaotong University,2001.

[5] Gan Quan,Wang Guo-yin,Hu Jun.A self-learning model based on granular computing[C]//2006 IEEE International Conference on Granular Computing,Atlanta,2006:530-533.

[6] An J J,Wang G Y,Wu Y,et al.A rule generation algorithm based on granular computing[C]//Proc IEEE Int Conf on Granular Computing,Beijing,2005:102-108.

[7] Yao Y Y.On modeling data mining with granular computing[C]//Proc COMPSAC 2001,Chicago,USA,2001:638-643.

(上接122页)

可以看到,本文的加密效果很好,没有改变小波系数的数值,而且运算速度快,加密造成的膨胀量小。

6　结论

本文将模运算运用于小波域来加密图像,同时与混沌模板序列相结合,对图像的小波系数进行置乱和置换变换,实现了小波变换域的高强度加密。本文提出的算法不仅能获得很好的加密效果,而且相比于其他的置乱算法,可以很大程度地节约计算时间,减小计算的复杂度。同时采用混沌加密模板,比传统的混沌序列有更高的加密强度,密钥简单但是密钥空间却很大,不易于被破解。仿真实验表明,该算法的实用性很强,在不改变小波系数值的前提下,运算速度快,加密造成的膨胀量小,且能够应用于一些压缩编码方法中。

参考文献:

[1] 李昌刚,韩正之,张浩然.图像加密技术综述[J].计算机研究与发展,2002(10):1317-1324.

[2] 尹显东,姚军,唐丹,等.基于小波变换域的图像加密技术研究[J].信息与电子工程,2005(3):1-5.

[3] 刘家胜,黄贤武,朱灿焰,等.基于模运算与混沌映射的图像加密算法的研究[J].微电子学与计算机,2006(12):206-212.

[4] 平亮,孙军,周军.一种基于JPEG2000标准的数字图像加密算法[J].视频技术应用与工程,2006(7):87-90.

[5] Norcen R,Andreas U.Selective encryption of the JPEG2000 bitstream[C]//Proc IFIP International Federation for Information.Chicago:[s.n.],1999.

[6] Lian Shi-guo,Wang Zhi-quan.Comparison of several wavelet coefficient confusion methods applied in multimedia encryption[C]//2003 International Conference on Computer Networks and Mobile Computing.[S.l.]:IEEE Computer Society,2003.