

基于二进制可辨矩阵的知识粒度研究及应用

汪小燕^{1,2}, 王浩¹

(1. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009;

2. 安徽工业大学 计算机学院, 安徽 马鞍山 243002)

摘要:粗糙集理论是一种新型的处理模糊和不确定知识的数学工具, 其对知识的理解是认为知识与分类相关、知识是有粒度的。文中利用粗糙集理论中的二进制可辨矩阵讨论知识的粒度计算及其应用, 得到了二进制可辨矩阵若干定理及推论, 并提出计算知识的分辨率和属性重要度的新方法, 利用这些理论和公式, 可快速计算出知识的分辨率和属性重要度, 相对正域和负域等, 为以后的属性约简和规则提取打下基础。并给出这些方法的应用, 表明了文中提出的方法的有效性。

关键词:粗糙集; 二进制可辨矩阵; 粒度; 重要度

中图分类号: TP39; O159

文献标识码: A

文章编号: 1673-629X(2006)10-0091-03

Research and Application of Knowledge Granulation
Based on Binary Discernible MatrixWANG Xiao-yan^{1,2}, WANG Hao¹

(1. School of Computer and Information, Hefei University of Technology, Hefei 230009, China;

2. School of Computer, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: Rough set is a new instrument in math. It considers that knowledge is connected with kinds and knowledge has granulation. This paper introduces calculation and application of knowledge granulation by using binary discernable matrix. Several theorems and deduction of binary discernable matrix are gained by utilizing these concepts. This paper proposes a new computational method of knowledge resolution and the significance of attribute. By utilizing the results gained, they can be calculated quickly, for example: knowledge resolution, the significance of attribute, relation positive regions and relation negative regions, etc. Attribution reduction and rule obtaintion can be based on these theorems, too. The application of the methods in the paper demonstrates the effectiveness of the result obtained.

Key words: rough set; binary discernable matrix; granulation; significance

0 引言

由 Z. Pawlak 与他的同事们在 20 世纪 80 年代初发展起来的粗糙集理论已受到国内外学者越来越广泛的重视, 已逐渐成为一种重要的数据处理方法。它以一种全新的视觉审视知识, 认为知识与分类相关、知识是有粒度的。知识的这种颗粒状结构通过等价关系的等价类表示。正是由于知识的这种颗粒状, 导致了知识表示的粗糙性。它对世界的理解, 只是达到了等价关系的一个一个的等价类这种颗粒状的程度, 而对颗粒内的对象是无法分辨的。文中利用二进制可辨矩阵讨论知识的粒度的计算及其应用, 得到了二进制可辨矩阵若干定理及推论。同时提出计算知识的分辨率和属性重要度的新方法, 并给出这些方法的应用。

1 二进制可辨矩阵

定义 1^[5] 设决策表为: $T = (U, R, V, F)$, 其中 $U = \{u_1, u_2, \dots, u_n\}$, $R = C \cup D (C \cap D = \emptyset)$, $C = \{c_1, c_2, \dots, c_m\}$, $D = \{e\}$

决策表 T 相应的二进制可辨矩阵 MT 构造如下: 矩阵 MT 的每一列对应一个条件属性, 共有 m 列, 每一行对应论域中的一个对象对 (u_p, u_q) , 且 $d(u_p) \neq d(u_q)$, 即这一对对象属于不同的决策类, MT 至多有 $n(n-1)/2$ 行, 设矩阵中一元素 $m_{((p,q),i)}$ 所在行对应对象对 (u_p, u_q) , 所在列对应条件属性 c_i ,

$$m_{((p,q),i)} = \begin{cases} 1 & c_i(u_p) \neq c_i(u_q) \\ 0 & \text{否则} \end{cases}$$

设 $I = (U, A, V, f)$ 是一个信息系统, 则相应于 I 的二进制可辨矩阵 MI 可类似得到, 由于信息系统是条件属性与决策属性所处位置相同的决策表, 故 MI 的行应对应在不同分辨的关系 $IND(A)$ 下可分辨的对象对 (u_p, u_q) , MI 的列对应属性集 A 中的属性 a_i , 设 $MI = (m_{((p,q),i)})$, 则:

$$m_{((p,q),i)} = \begin{cases} 1 & a_i(u_p) \neq a_i(u_q) \\ 0 & \text{否则} \end{cases}$$

收稿日期: 2006-01-12

作者简介: 汪小燕(1974-), 女, 安徽桐城人, 讲师, 硕士研究生, 研究方向为数据挖掘、计算机数据库; 王浩, 教授, 博士, 主要研究领域为 Agent, 数据挖掘, 软件工程。

为缩小二进制可辨矩阵的规模,在表中只列出 $p < q$ 的行。

根据二进制可辨矩阵的定义,可得如下的推论:在二进制可辨矩阵中,若存在某一行不全为 0,则该行所对应的一个对象对 (u_p, u_q) 在属性集 R 下是可分辨的,否则该对象对 (u_p, u_q) 在属性集 R 下是不可分辨的。

2 基于分辨矩阵的知识粒度与计算

设 U 是论域, R 是属性集,由 R 导出的划分(等价类),记作 $U/R = (X_1, X_2, \dots, X_r)$,表示 R 的分类能力,称作知识 R 。利用分辨矩阵 $M(R)$ 可以判断对象的是否可分辨,借此可以计算知识 R 的分辨能力。

定义 2^[4] 称 $DIS(R)$ 是知识 R 的分辨度,而且

$$DIS(R) = |m_{ij}(R) \neq \emptyset| / |M(R)| \quad (1)$$

式中, $|m_{ij}(R) \neq \emptyset|$ 表示分辨矩阵中 $m_{ij}(R) \neq \emptyset$ 的元素个数, $|M(R)|$ 表示分辨矩阵所含元素的个数,即 $|M(R)| = |U|^2$,分辨度的大小反映了知识的分辨能力。

定义 3^[4] 称 $GRD(R)$ 是知识 R 的粒度,而且

$$GRD(R) = \sum_{i=1}^r |X_i|^2 / |U|^2 \quad (2)$$

式中 $|X_i|$ 表示等价类 X_i 的基数。显然,知识 R 的粒度与知识 R 的分辨度之间存在关系: $GRD(R) = 1 - DIS(R)$ 。

一般情况下,有 $1/|U| \leq GRD(R) \leq 1$ 。知识的粒度可以表示知识的分辨能力, $GRD(R)$ 越小,分辨能力越强。当 $(u, v) \in X_i$ 时,表明对象 u, v 在 R 下不可分辨,属于 R 的同一个等价类。否则,它们可分辨,属于不同的 R -等价类。因此, $GRD(R)$ 表示在 U 中随机选择两个对象,这两个对象 R -不可分辨的可能性大小。可能性越大,即 $GRD(R)$ 越大,表明 R 的分辨能力越弱,否则,越强^[3]。

设 $S = (U, R)$ 是信息系统, $A \subset R$ 是属性子集, $x \in R$ 是属性。考虑 x 对于 A 的重要度,即 A 中增加属性 x 之后知识的分辨度的提高程度,提高程度越大,认为 x 对于 A 越重要。

定义 4^[4] 设 $A \subset R$ 是属性子集, $x \in R$ 是属性,称 $Sig_A(x)$ 是 x 对于 A 的重要度,而且

$$Sig_A(x) = 1 - |m_{ij}(A \cup \{x\}) = \emptyset| / |m_{ij}(A) = \emptyset| \quad (3)$$

定义 5^[1] (一致性定义) 若信息决策系统中的两个对象,满足条件之一: (1) 条件属性的取值至少有一个不同; (2) 有相同的条件属性取值时,决策属性的取值是相同的,则称作这两个对象是一致的;否则称作不一致的。若决策表中任何一对对象都是一致的,称该决策表是一致的;否则称该决策表是不一致。

定义 6^[1] 设 $S = (U, C \cup D)$ 是一个信息决策系统, $k_C(D) = \text{card}(\text{POS}_C(D)) / \text{card}(U)$, 称 $k_C(D)$ 是决策属性集 D 对条件属性集的依赖度。其中 $\text{card}(S)$ 表示集合 S 的基数, $\text{POS}_C(D)$ 是 D 的 C -正域, $\text{NEG}_C(D) = U - \text{POS}_C(D)$ 称为 D 的 C -负域。

若 $S = (U, R, V, f)$ 是信息决策系统,则 $R = C \cup D$ ($C \cap D = \emptyset$), C 是条件属性集和 $D = \{e\}$ 是决策属性集,对应的分辨矩阵记作 $M(C, e)$ 。

定理 1^[2] 在分辨矩阵中,若存在某个 $m_{ij}(C, e) = \{e\}$, 则信息决策系统是不一致的;否则信息决策系统是一致的。

定理 2^[2] 在分辨矩阵中,若存在某一行 $m_{ij}(C, e) \neq \{e\}$, 则 $U_i \in \text{POS}_C(D)$ 。

3 基于二进制可辨矩阵的知识粒度与计算

基于分辨矩阵的知识的分辨度与知识的粒度定义,同样可得出二进制可辨矩阵的知识的分辨度与知识的粒度定义,设 $S = (U, R)$ 是信息系统。

定义 7 称 $DIS(R)$ 是知识 R 的分辨度,而且

$$DIS(R) = 2 |M_{((p,q),i)}(R) \neq 0| / |U|^2 \quad (4)$$

式中 $|M_{((p,q),i)}(R) \neq 0|$ 表示二进制可辨矩阵中所有不全为 0 行的对象对个数。二进制可辨矩阵的表中只列出 $p < q$ 的行,故式中出现 2。

关于二进制可辨矩阵知识 R 的粒度计算同定义 3。

定义 8 设 $A \subset R$ 是属性子集, $x \in R$ 是属性,称 $Sig_A(x)$ 是 x 对于 A 的重要度,而且

$$Sig_A(x) = 1 - |M_{(p,q)}(A \cup \{x\}) = 0| / (2 |M_{(p,q)}(A) = 0| + |U|) \quad (5)$$

式中: $|M_{(p,q)}(A \cup \{x\}) = 0|$ 表示 A 中增加属性 x 之后,二进制可辨矩阵中所有的全为 0 行的对象对个数, $|M_{(p,q)}(A) = 0|$ 表示 A 中未增加属性 x 时,二进制可辨矩阵中所有全为 0 行的对象对个数,在二进制可辨矩阵中不列出 (u_p, u_q) 对象对,假如列出,它所对应的行也全为 0,故式中出现 $|U|$ 。

若 $I = (U, R, V, f)$ 是信息决策系统,则 $R = C \cup D$ ($C \cap D = \emptyset$), C 是条件属性集, $D = \{e\}$ 是决策属性集。

定理 3 在二进制可辨矩阵中,若存在某一行全为 0,则信息决策系统是不一致的;否则信息决策系统是一致的。

证明:信息决策系统的二进制可辨矩阵的每一行对象对的决策属性是不相同的,若存在某一行全为 0,则说明两个对象的条件属性完全相同,而决策属性不相同,则信息决策系统是不一致的。

定理 4 在信息决策系统二进制可辨矩阵中,所有全为 0 的行所对应的对象对的并集构成 $\text{NEG}_C(D)$ 。

证明:在信息决策系统二进制可辨矩阵中,若某一行全为 0,则该行所对应的对象对的条件属性完全相同,而决策属性不相同,这两个对象,根据现有的知识,即条件属性,无法确定归入哪一个决策类,而 $\text{NEG}_C(D)$ 中的对象是所有不能确定一定归入哪一个决策类的元素的集合。故在信息决策系统二进制可辨矩阵中,所有全为 0 的行所对应的对象对的并集构成 $\text{NEG}_C(D)$ 。由定理 4 可以得到:

推论 1 设 $S = (U, C \cup D)$ 是一个信息决策系统, C 是条件属性集, $D = \{e\}$ 是决策属性集, 则二进制可辨矩阵中, 在 U 中除去所有全为 0 的行所对应的对象对的并集后, 余下的元素组成 $\text{POS}_C(D)$ 。

推论 2 在二进制可辨矩阵中, 若存在每一行都不全为 0, 则 $\text{POS}_C(D)$ 为 U 的全集。

4 粒度计算的应用

给定信息系统 $S = (U, R)$, 其中属性集 $C = \{a, b, c, d\}$, 如表 1 所示。表 2 为表 1 的二进制可辨矩阵。

表 1 $S = (U, R)$

U	a	b	c	d
1	1	0	1	1
2	1	0	1	0
3	0	0	1	1
4	0	0	1	0
5	1	1	1	1

表 2 表 1 的二进制可辨矩阵

	a	b	c	d		a	b	c	d
1,2	0	0	0	1	2,4	1	0	0	0
1,3	1	0	0	0	2,5	0	1	0	1
1,4	1	0	0	1	3,4	0	0	0	1
1,5	0	1	0	0	3,5	1	1	0	0
2,3	1	0	0	1	4,5	1	1	0	1

由定义 3 和 7 知: $\text{DIS}(R) = 2 \times 10/5^2 = 4/5$, $\text{GRD}(R) = 5/25 = 1/5$

由定义 8 知: $\text{Sig}_{R-\{a\}} = 1 - 5/(5 + 2 \times 2) = 4/9$, $\text{Sig}_{R-\{b\}} = 1 - 5/(5 + 2 \times 1) = 2/7$, $\text{Sig}_{R-\{c\}} = 1 - 5/(5 + 2 \times 0) = 0$, $\text{Sig}_{R-\{d\}} = 1 - 5/(5 + 2 \times 2) = 4/9$

若表 1 是信息决策系统 $S = (U, R)$, $R = C \cup D$, 其中属性集 $C = \{a, b, c\}$, $D = \{d\}$, 则其二进制可辨矩阵如表 3 所示。

由定理 3 知, 表 3 的第一、第五行全为 0, 则该信息决策系统是不一致的; 由定理 4 知, $\text{NEG}_{C(D)} = \{1, 2\} \cup \{3,$

$4\} = \{1, 2, 3, 4\}$; 由定理 4 的推论 1 知: $\text{POS}_C(D) = U - \{1, 2, 3, 4\} = \{5\}$, $k_C(D) = 1/5$, $\text{NEG}_{C-a}(D) = \{1, 2\} \cup \{1, 4\} \cup \{2, 3\} \cup \{3, 4\} = \{1, 2, 3, 4\}$, $\text{POS}_{C-a}(D) = U - \{1, 2, 3, 4\} = \{5\}$, $\text{NEG}_{C-b}(D) = \{1, 2\} \cup \{2, 5\} \cup \{3, 4\} = \{1, 2, 3, 4, 5\}$, $\text{POS}_{C-b}(D) = U - \{1, 2, 3, 4, 5\} = \emptyset$, $\text{NEG}_{C-c}(D) = \{1, 2\} \cup \{3, 4\} = \{1, 2, 3, 4\}$, $\text{POS}_{C-c}(D) = U - \{1, 2, 3, 4\} = \{5\}$ 。

表 3 决策表的二进制可辨矩阵

	a	b	c
1,2	0	0	0
1,4	1	0	0
2,3	1	0	0
2,5	0	1	0
3,4	0	0	0
4,5	1	1	0

5 结束语

文中是利用二进制可辨矩阵讨论知识的粒度的计算及其应用, 得到了二进制可辨矩阵若干定理及推论, 提出计算知识粒度和属性重要度的新方法, 并给出这些方法的应用。利用文中得出的二进制可辨矩阵若干定理及推论, 用于属性及属性值约简, 将另文给出。

参考文献:

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356.
- [2] Skowron A. Rough Sets in KDD[R]. Special Invited Speaking, WCC2000 in Beijing, 2000. 36-39.
- [3] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.
- [4] 郑书富. 分辨矩阵与知识粒度的应用[J]. 聊城大学学报, 2004, 10: 16-18.
- [5] 支天云, 苗夺谦. 二进制可辨矩阵的变换及高效属性约简算法[J]. 计算机科学, 2002, 29(2): 140-142.

(上接第 90 页)

场景中的物体运动也较大的图像, 该方法仍能在较大的压缩比下得到重建质量良好的视频图像, 这说明文中所提出的方法是有所改进的, 特别是对于运动变化较大的图像。

4 结束语

文中采用了三维小波变换对视频图像序列进行压缩编码, 它较好地改善了当图像场景中的物体进行快速运动时, 使得时间域的小波系数突然变大而使得压缩效率变低的不足。该算法可以快速、高效地压缩图像, 并且无运动补偿和运动估计, 降低了算法的复杂度。今后对该算法在量化矩阵的构造和阈值的选取上还可以进一步地进行研究和完善。

参考文献:

- [1] 杨春玲, 余英林. 基于三维小波变换嵌入式视频压缩算法的研究[J]. 电子学报, 2001(10): 1381-1383.
- [2] 张宗平, 刘贵忠, 侯兴松. 一种改进的三维小波视频编码[J]. 西安交通大学学报, 2001, 35(6): 595-599.
- [3] Watson A B, Yang G U, Solomon J A, et al. Visibility of wavelet quantization noise[J]. IEEE Tran Image Process, 1997, 6: 1164-1175.
- [4] Ferguson K L, Allinson N M. Psychophysically derived quantisation model for efficient DWT image coding[J]. IEE Proc - Vis Image signal Process, 2002, 149(1): 51-56.
- [5] 李春华, 白云飞. 一种基于三维小波变换的图像编码算法[J]. 仪器仪表学报, 2002, 23(增刊): 199-200.