

一种基于知识粒度的决策系统属性约简算法

廖洪建

(柳州师范高等专科学校 数学与计算机科学系, 广西 柳州 545004)

摘 要: 本文针对决策系统给出了一种新的知识粒度模型, 并给出了知识粒度下的核属性定义和分析了知识粒度模型的属性约简与正区域模型的属性约简的等价性. 在此基础上利用知识粒度的重要性作为启发式信息构造了决策系统的属性约简算法. 算法能够快速获取决策系统的属性约简, 算例分析进一步说明了算法的可靠性.

关键词: 知识粒度; 决策系统; 属性约简; 正区域模型

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2013)02-0087-04

Attribute Reduction Algorithm of Decision-Making System Based on Knowledge Granularity

LIAO Hong-jian

(Department of Mathematics and Computer Science, Liuzhou Teachers College, Liuzhou 545004, China)

Abstract: In this paper, new measurement method for granularity knowledge model based on decision-making system, and it is proved that attribute reduction based on granularity knowledge model is equivalent to that based on positive region model in decision-making system, and then attribute reduction algorithm of decision-making system based on knowledge granularity is proposed, the algorithm quickly calculates attribute reduction in decision-making system, the example also show that the algorithm is effective and reliable.

Key words: knowledge granularity; decision system; attribute reduction; positive region model

1 引言

粗糙集理论为处理不确定、不精确和不完全信息的知识分类和规则挖掘提供一种更符合人类认知机理的数学工具, 它无需任何附加信息就直接对数据信息处理, 突破例如模糊集、概率统计和证据理论等的技术瓶颈, 该理论已应用在诸多领域中^[1-2]. 粒度计算是智能信息处理的新方法^[3-4]; 而粗糙集理论作为粒度计算的重要方法之一, 现已成为人工智能和机器学习领域的重要研究内容.

在粒度计算理论中, 可将许多复杂的信息通过粒度细化的方法, 把复杂的信息根据数据本身的属性特征和功能特点将其划分成使人类更容易理解的知识块, 也称知识粒度. 知识粒度是粗糙集理论中度量属性约简的重要方法之一, 目前许多学者利用知

识粒度研究属性约简, 提出了一些有效的算法, 其中有基于粒度计算的特征选择算法^[4], 在多值信息系统下利用粒度计算方法设计多级概念获取算法^[5], 在粒度空间下构造属性约简的启发式矩阵算法^[6], 这些算法为知识粒度模型下的属性约简提供了有效的解决途径; 上述算法都是针对信息系统而设计的, 对于决策系统的研究少见报道, 而实际领域中的信息系统很多都带决策属性, 称之为决策系统, 如何利用知识粒度设计出针对决策系统的属性约简算法是一个尚需解决的重要课题.

本文提出一种基于知识粒度的决策系统属性约简算法. 首先针对决策系统给出了一种新的知识粒度模型, 同时给出了知识粒度下的核属性定义, 并分析了知识粒度模型的属性约简与正区域模型的属性约简的等价性. 然后利用知识粒度的重要性作为启

收稿日期: 2012-05-25; **修回日期:** 2012-06-19

基金项目: 国家自然科学基金项目(60963008)

发式信息构造了决策系统的属性约简算法. 由于算法使用粒度细化的思想使得算法的时空复杂度得到降低, 同时算例分析也表明了该算法的可靠性.

2 粗糙集的相关概念

本小节仅介绍本文所涉及的相关概念, 其他详细概念可查阅文献[1-2].

定义 1^[2] 四元组 $IS = (U, A, V, f)$ 为信息系统, 其中 U 为有限非空对象集, 亦称论域; A 为有限非空属性集; V 表示属性集 A 的值域, 即 V_a 表示属性 a 的值域; $f: U \times A \rightarrow V$ 为信息函数, 即有信息系统中的每个对象的所有属性均赋予相应的值. 若 $A = C \cup D$ 且 $C \cap D = \emptyset$, 此时 C 为条件属性集, D 为决策属性集, 则该信息系统叫决策系统 $S = (U, C, D, V, f)$.

定义 2^[2] 五元组 $S = (U, C, D, V, f)$ 为决策系统, 对于任意的属性集 $c_i \in P \subset C$, 则 P 对应的不可辨别关系 $Ind(P) = \{(x, y) \in U \times U \mid c_i \in P, f(x, c_i) = f(y, c_i)\}$, 同时可知决策系统在属性集上 P 的划分为: $U/Ind(P) = \{P_1, P_2, \dots, P_v\}$, 每个划分即为一个等价类.

定义 3^[2] 五元组 $S = (U, C, D, V, f)$ 为决策系统, 对于任意的属性集 $P \subset C$ 对应的划分为 $U/Ind(P) = \{P_1, P_2, \dots, P_v\}$, 决策属性 D 对应的划分为 $U/Ind(D) = \{D_1, D_2, \dots, D_w\}$, 属性集 P 相对于决策属性 D 的正区域为 $POS_P(D) = \bigcup_{D_i \in U/D} P(D_i)$, 负区域 $NEG_P(D) = U - POS_P(D)$.

定义 4^[2] 五元组 $S = (U, C, D, V, f)$ 为决策系统, $Core$ 为正区域模型下的核属性集, 则对于 $\forall c_i \in Core$, 满足条件 $POS_{C-\{c_i\}}(D) \neq POS_C(D)$.

定义 5^[2] 五元组 $S = (U, C, D, V, f)$ 为决策系统, $Red(C)$ 为正区域模型的属性约简, 则 $Red(C)$ 满足条件: (1) $POS_{Red(C)}(D) = POS_C(D)$; (2) 对于 $\forall c_i \in Red(C)$, 使得 $POS_{Red(C)-\{c_i\}}(D) \neq POS_C(D)$.

3 知识粒度与等价性定理

定义 6 五元组 $S = (U, C, D, V, f)$ 为决策系统, 对于任意的属性集 $\forall R \subset C$, 则条件属性 R 相对于决策属性 D 的知识粒度为:

$$Gr(R) = \sum_{X_i \subseteq POS_R(D)} \frac{|X_i|^2}{|U|^2}$$

其中, X_i 为在属性集上的划分(等价类)且属于正区域的对象集; 在粒度空间理论中, 称等价类为知识粒度.

定义 7 五元组 $S = (U, C, D, V, f)$ 为决策系统, $GCore$ 为知识粒度模型下的核属性集, 则对于 $\forall c_i \in GCore$, 必满足条件 $Gr(C - \{c_i\}) \neq Gr(C)$.

定义 8 五元组 $S = (U, C, D, V, f)$ 为决策系统, $GRed(C)$ 为知识粒度模型下的属性约简, 则 $GRed(C)$ 必满足条件: (1) $Gr(GRed(C)) = Gr(C)$; (2) 对于 $\forall c_i \in GRed(C)$, 使得 $Gr(GRed(C) - \{c_i\}) \neq Gr(C)$.

定义 9 五元组 $S = (U, C, D, V, f)$ 为决策系统, 对于 $\forall R \subset C$, 条件属性 c_i 的知识粒度重要性 $SGF(c_i, R, D)$ 定义为: $SGF(c_i, R, D) = Gr(R \cup \{c_i\}) - Gr(R)$.

定理 1 在决策系统 $S = (U, C, D, V, f)$ 中, 知识粒度模型下的属性约简 $GRed(C)$ 等价于正区域模型下的属性约简 $Red(C)$.

证明: 对于决策系统 $S = (U, C, D, V, f)$, 设决策系统在条件属性 C 上的划分为 $U/C = \{A_1, A_2, \dots, A_n\}$ 及决策属性 D 上的划分 $U/D = \{B_1, B_2, \dots, B_m\}$, 条件属性 C 相对于决策属性 D 的正区域为 $POS_C(D) = \{X_1, X_2, \dots, X_r\} \subseteq \{A_1, A_2, \dots, A_n\}$; 若 $c_i \in GRed(C)$, 则可知属性 c_i 必将用于划分决策系统的正区域 $POS_C(D)$, 又由于知识粒度的属性约简满足 $Gr(GRed(C)) = Gr(C)$, 在正区域模型的属性约简满足 $POS_{Red(C)}(D) = POS_{Red(C)}(D)$, 故可知这两个条件在度量方法是等价的. 同时, 对于 $c_i \in GRed(C)$, 若删除属性 c_i , 则条件属性 $GRed(C) - c_i$ 所对应的知识粒度将发生改变, 即满足 $Gr(GRed(C) - \{c_i\}) \neq Gr(C)$, 再由知识粒度的定义可知, 此时决策系统的正区域 $POS_{Red(C)}(D)$ 也必定改变, 即有 $POS_C(D) \neq POS_{Red(C)-\{c_i\}}(D)$ 成立, 故可知这两个条件在度量上也是等价的. 综上分析, 定理成立.

4 基于知识粒度的属性约简算法

传统的算法都基于信息系统上给出知识粒度模型, 这对于决策系统而言并不适用, 为此给出了针对决策系统的知识粒度模型, 并有效利用了决策系统的核属性和知识粒度的重要性度量, 在此基础上构造了一种基于知识粒度的决策系统属性约简算法.

算法 基于知识粒度的决策系统属性约简算法
输入: 决策系统 $S = (U, C, D, V, f)$

输出: 属性约简 $GRed(C)$

Step1. 对决策系统 $S = (U, C, D, V, f)$ 中的对象计算出在条件属性 C 上的划分及决策属性 D 上的

划分 Step2. 根据 U/C 和 U/D 的划分, 计算其正区域 $POS_C(D)$, 负区域 $NEG_C(D)$;

Step3. 计算条件属性 C 相对于决策属性 D 的知识粒度 $Gr(C)$;

Step4. 对于 $\forall c_i \in C$, 分别计算 $C - \{c_i\}$ 相对决策属性的知识粒度 $Gr(C - \{c_i\})$; 如果 $Gr(C - \{c_i\}) \neq Gr(C)$, 则属性为核属性 c_i , 将其存入属性约简 $GRed(C)$; 此时若 $Gr(Red(C)) = Gr(C)$, 则输出属性约简, 算法结束; 否则 $R = GRed(C)$;

Step5. 对于 $\forall c_i \in C - R$, 依次计算其对应的划分 $U/\{R \cup c_i\}$ 和知识粒度 $Gr(R \cup \{c_i\})$, 并根据知识粒度的属性重要性度量 $SGF(c_i, R, D) = Gr(R \cup \{c_i\}) - Gr(R)$, 选择其中的最大值所对应的属性 c_k 存入属性集 R 中, 即 $R = R \cup \{c_k\}$;

Step6. 如果 $Gr(R) = Gr(C)$, 则输出属性约简 $GRed(C) = R$, 算法结束; 否则, 算法继续执行 Step5.

算法的时间和空间复杂度:

算法 Step1 的时间复杂度为 $O(|U|^2 |C|) + O(|U|^2 |D|)$, 由于一般情况下 $|C| \gg |D|$, 则 Step1 的时间复杂度为 $O(|U|^2 |C|)$; Step2 的时间复杂度为 $O(|U|)$; Step3 的时间复杂度为 $O(|U|^2 |C|)$; Step4 的时间复杂度为 $O(|U|^2 |C - 1|)$; Step5 的时间复杂度为 $O(|U| |C|^2)$; 故算法的时间复杂度为 $O(|U|^2 |C|)$; 同时可知算法的空间复杂度为 $O(|U| |C|)$. 由于基于知识粒度的信息系统的属性约简算法复杂度为 $O(|U|^2 |C|^3)$, 为此本文算法的时间复杂度得到显著改善, 是一种有效的属性约简算法.

5 算例分析

为进一步验证本文所提算法是有效可靠的, 下面以表 1 中的决策系统为例来详细分析本文算法的执行过程. 其中条件属性集 $C = \{a, b, c, e, f\}$, 决策属性集 $D = \{d\}$.

表 1 决策系统

U	a	b	c	e	f	d
1	1	0	0	0	1	0
2	0	0	1	0	1	0
3	0	1	1	1	0	1
4	0	0	1	0	1	0
5	0	1	1	1	0	1
6	1	0	0	0	0	1
7	1	0	0	0	0	1
8	0	1	1	0	1	0
9	0	1	1	0	1	1

根据算法 Step1 可计算得, 条件属性的划分为 $U/\{C\} = \{\{1\}, \{2, 4\}, \{3, 5\}, \{6, 7\}, \{8, 9\}\}$, 决策属性的划分 $U/\{d\} = \{\{1, 2, 4, 8\}, \{3, 5, 6, 7, 9\}\}$; 根据算法 Step2 可知, 决策系统的正区域 $POS_C(D) = \{\{1\}, \{2, 4\}, \{3, 5\}, \{6, 7\}\}$, 负区域 $NEG_C(D) = \{\{8, 9\}\}$; 根据算法 Step3, 计算条件属性 C 相对于决策属性 d 的知识粒度为:

$$Gr(C) = \sum_{X_i \subseteq POS_{\{C-\{a\}\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{13}{81}$$

根据算法 Step4, 依次计算如下知识粒度:

由于 $U/\{C - \{a\}\} = \{\{1\}, \{2, 4\}, \{3, 5\}, \{6, 7\}, \{8, 9\}\}$, 则对应的知识粒度为:

$$Gr(C - \{a\}) = \sum_{X_i \subseteq POS_{\{C-\{a\}\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{13}{81}$$

由于 $U/\{C - \{b\}\} = \{\{1\}, \{2, 4\}, \{3, 5\}, \{6, 7\}, \{8, 9\}\}$, 则对应的知识粒度为:

$$Gr(C - \{b\}) = \sum_{X_i \subseteq POS_{\{C-\{b\}\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{13}{81}$$

由于 $U/\{C - \{c\}\} = \{\{1\}, \{2, 4\}, \{3, 5\}, \{6, 7\}, \{8, 9\}\}$, 则对应的知识粒度为:

$$Gr(C - \{c\}) = \sum_{X_i \subseteq POS_{\{C-\{c\}\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{13}{81}$$

由于 $U/\{C - \{e\}\} = \{\{1\}, \{2, 4\}, \{3, 5\}, \{6, 7\}, \{8, 9\}\}$, 则对应的知识粒度为:

$$Gr(C - \{e\}) = \sum_{X_i \subseteq POS_{\{C-\{e\}\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{13}{81}$$

由于 $U/\{C - \{f\}\} = \{\{1, 6, 7\}, \{2, 4\}, \{3, 5\}, \{8, 9\}\}$, 则对应的知识粒度为:

$$Gr(C - \{f\}) = \sum_{X_i \subseteq POS_{\{C-\{f\}\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{8}{81}$$

由于 $Gr(C) \neq Gr(C - \{f\})$, 根据定义 7 的核属性定义可知, 决策系统的核属性为 $\{f\}$. 为此将属性 $\{f\}$ 存入属性约简 $GRed(C) = \{f\}$, 由于此时 $Gr(Red(C)) \neq Gr(C)$, 则根据算法的 Step5, 需选择 $C - \{f\}$ 中的属性分别计算其对应的知识粒度.

由于 $U/\{f, a\} = \{\{1\}, \{2, 4, 8, 9\}, \{3, 5\}, \{6, 7\}\}$, 则对应的知识粒度为:

$$Gr(\{f, a\}) = \sum_{X_i \subseteq POS_{\{f, a\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{9}{81}$$

由于 $U/\{f, b\} = \{\{1, 2, 4\}, \{3, 5\}, \{6, 7\}, \{8, 9\}\}$, 则对应的知识粒度为:

$$Gr(\{f, b\}) = \sum_{X_i \subseteq POS_{\{f, b\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{8}{81}$$

由于 $U/\{f, c\} = \{\{1\}, \{2, 4, 8, 9\}, \{3, 5\}, \{6, 7\}\}$, 则对应的知识粒度为:

$$Gr(\{f, c\}) = \sum_{X_i \subseteq POS_{\{f, c\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{9}{81}$$

由于 $U/\{f, e\} = \{\{1, 2, 4, 8, 9\}, \{3, 5\}, \{6, 7\}\}$, 则对应的知识粒度为:

$$Gr(\{f, e\}) = \sum_{X_i \subseteq POS_{\{f, e\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{8}{81}$$

选择使 $SGF(c_i, R, D) = Gr(R \cup \{c_i\}) - Gr(R)$ 最大的属性, 通过计算可知属性 $\{a, c\}$ 满足条件, 则随机选择属性 a 存入属性约简中, 此时 $R = \{f, a\}$ 重复执行 Step5, 从 $\{b, c, e\}$ 中选择属性依次计算其对应的知识粒度:

$$Gr(\{f, a, b\}) = \sum_{X_i \subseteq POS_{\{f, a, b\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{13}{81}$$

$$Gr(\{f, a, c\}) = \sum_{X_i \subseteq POS_{\{f, a, c\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{9}{81}$$

$$Gr(\{f, a, e\}) = \sum_{X_i \subseteq POS_{\{f, a, e\}}(D)} \frac{|X_i|^2}{|U|^2} = \frac{9}{81}$$

选择使 $SGF(c_i, R, D) = Gr(R \cup \{c_i\}) - Gr(R)$ 最大的属性, 通过计算知属性 $\{b\}$ 满足条件, 则选择属性 b 存入属性约简中, 此时 $R = \{f, a, b\}$, 由于 $Gr(R) = Gr(C)$, 根据算法 Step6 可知, 决策系统的属性约简已计算出, 即属性约简 $GRed(C) = \{f, a, b\}$, 算法结束.

通过上述实例的详细分析发现, 本文的算法利用知识粒度的度量方法和粒度细化的思想能够快速求解出决策系统的属性约简.

6 结束语

本文针对决策系统给出了一种条件属性相对于决策属性的知识粒度模型. 同时给出了该模型下的核属性定义, 并阐述了知识粒度模型的属性约简与

正区域模型的属性约简的等价性分析, 并利用知识粒度的重要性作为启发式信息构造了决策系统的属性约简算法, 这为利用粒度计算求解代数观下的属性约简提供一种新的有效途径.

参考文献:

- [1] 胡平, 彭纪奎. 粗糙集—遗传神经网络在挖掘机故障诊断中的应用研究[J]. 微电子学与计算机, 2011, 28(3): 55—58.
- [2] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2003.
- [3] Yao Y Y, Ning Z. Granular computing using information table[M]. Rough Set Theory and Granular Computing. Heidelberg: Physica—Verlag, 2000.
- [4] 孙丽君, 苗夺谦. 基于粒度计算的特征选择方法[J]. 计算机科学, 2008, 35(4): 14—15.
- [5] 邱桃荣, 刘清, 黄厚宽. 多值信息系统中基于粒计算的多级概念获取算法[J]. 模式识别与人工智能, 2009, 22(1): 22—27.
- [6] 钟珞, 梅磊, 郭翠翠, 等. 粒矩阵属性约简启发式算法[J]. 小型微型计算机系统, 2011, 32(3): 516—520.

作者简介:



廖洪建 男, (1977—), 硕士, 讲师. 研究方向为粗糙集理论及应用、数据挖掘.