一种基于相对粒度的决策表约简算法

徐久成 史进玲 孙 林

(河南师范大学计算机与信息技术学院 新乡 453007)

摘 要 在知识粒度的基础上,针对决策表提出了相对粒度和属性相对重要性的概念,证明了知识的相对粒度随着知识粒度的增大而单调增加的变化规律,在此基础上提出了一种基于相对粒度的启发式约简算法,以弥补基于正区域的约简方法处理不一致决策表时存在的不足。通过理论分析和实例验证表明,该算法是有效的,且其时间复杂度相对较低。

关键词 决策表,知识粒度,相对粒度,属性约简

Attribute Reduction Algorithm Based on Relative Granularity in Decision Tables

XU Jiu-cheng SHI Jin-ling SUN Lin

(College of Computer & Information Technology, Henan Normal University, Xinxiang 453007, China)

Abstract A relative attribute significance of decision tables, based on the theory of knowledge granularity, was defined by introducing the concept of relative granularity, and the relative granularity's monotonous increasing property with the increase of knowledge granularity was proved, then a heuristic reduction algorithm based on relative granularity was proposed. The algorithm which eliminates the limitation of reduction algorithms based on positive region in dealing with inconsistent decision table, by analyzing essential theory and application examples, is proved as effective, and it's time complexity is relatively low.

Keywords Decision tables, Knowledge granularity, Relative granularity, Attribute reduction

1 引言

知识约简是粗糙集理论[1]的核心内容之一,然而寻求所有约简或最小约简已被证明是NP-Hard问题[2],因此,寻求高效快速的属性约简仍是粗糙集理论研究的核心领域之一。近年来,有许多学者对属性约简进行了深入研究,并取得了大量成果[3-6]。然而这些经典的粗糙集约简方法在处理不一致决策表时仍存在一定的不足[7]。文献[8]从知识粒度的角度出发,提出了一种启发式属性约简算法,该算法直观有效,但没有全面考虑决策表的情形。针对以上问题,本文将知识粒度概念引入决策表中,定义了知识的相对粒度,并基于相对粒度随知识粒度增大而单调增加的变化趋势,提出了一种有效的决策表启发式属性约简算法。

2 主要概念

定义 2.1 [1](信息系统) 设四元组 $S=(U,R,V,f)$ 是一个信息系统,其中 U 表示对象的非空有限集合,也称为论域;R 表示属性的非空有限集合,$R=C\cup D, C\cap D=\Phi$,其中 C 表示条件属性集,D 表示决策属性集;$V=\bigcup\{V_r \mid r\in R\}$,$V_r$ 为属性 r 的值域;$f:U\times R\to V$ 是一个信息函数,它为每个对象的每个属性赋予一个信息值,即 $\forall r\in R, x\in U$,有 $f(x,r)\in V_r$。特别地,当 $D\neq\Phi$ 时称信息系统 S 为决策系统。

定义 2.2 [1] 在信息系统 $S=(U,R,V,f)$ 中,任意属性子集 $P\subseteq R$ 决定了一个二元不可区分关系:$IND(P)=\{(x,y)\in U\times U \mid \forall p\in R, f(x,p)=f(y,p)\}$。关系 $IND(P)$ 确定了一个划分,用 $U/IND(P)$ 表示,简记为 U/P。U/P 中的任何元素 $[x]_P=\{y \mid f(x,a)=f(y,a), \forall a\in P\}$ 称为等价类。

性质 2.1 [9] 设 P,Q 为论域 U 上的两个属性集合,则有 $U/(P\cup Q)=U/P\cap U/Q$。

性质 2.2 [10] 在信息系统 $S=(U,R,V,f)$ 中,$P\subseteq R$, $a\in R\backslash P$,则有:

$U/(P\cup\{a\})=\bigcup\{X/\{a\} \mid \forall X\in U/P\}$。

性质 2.2 表明划分 $U/(P\cup\{a\})$ 可以由属性 a 对划分 U/P 中的每个等价块进行再划分得到。

定义 2.3 [8] 设四元组 $S=(U,R,V,f)$ 是一个信息系统,$U/R=\{X_1,X_2,\cdots,X_n\}$,则 R 的知识粒度定义为:$GD(R)=\sum_{i=1}^{n}|X_i|^2/|U|^2$(其中 $|X|$ 表示集合 X 的基数)。

3 知识的相对粒度及其属性重要性度量

定义 3.1 设 U 为一个论域,P 和 Q 为定义在 U 上的两个等价关系簇,则知识 Q 关于知识 P 的相对粒度定义为:$GD(Q\mid P)=GD(P)-GD(P\cup Q)$。

相对粒度 $GD(Q\mid P)$ 反映了知识 Q 相对于知识 P 在论域 U 上的分辨能力,即 $GD(Q\mid P)$ 越小,表明 Q 相对于 P 对 U 中

到稿日期:2008-05-30 本文得到国家自然科学基金项目(69803014, 60173058),河南省高校新世纪优秀人才支持计划(2006HANCET-19)资助。

徐久成 男,博士,教授,主要研究方向为粗糙集理论、粒计算、数据挖掘等;史进玲 女,硕士研究生,主要研究方向为粗糙集理论与粒计算。

对象的分辨能力就越弱；反之，若 $GD(Q|P)$ 越大，则表明 Q 相对于 P 对 U 中对象的分辨能力就越强。

定理 3.1 设 P_1, P_2, Q 为论域 U 上的等价关系簇，且 $U/P_1=\{X_1, X_2, \cdots, X_n\}$，$U/P_2=\{X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n, X_i \cup X_j\}$，$U/Q=\{Y_1, Y_2, \cdots, Y_m\}$，其中 U/P_2 是将 U/P_1 中的任意两个等价块（X_i 和 X_j）合并而得到的划分，则有：

$GD(Q|P_1) \leqslant GD(Q|P_2)$。

证明：由相对粒度的定义 3.1 和性质 2.1 得，

$$GD(Q|P_1)=GD(P_1)-GD(P_1 \cup Q)=\sum_{k=1}^{n}|X_k|^2/|U|^2-\sum_{k=1}^{n}\sum_{l=1}^{m}|X_k \cap Y_l|^2/|U|^2$$

同理，$GD(Q|P_2)=GD(P_2)-GD(P_2 \cup Q)=\sum_{k=1}^{i-1}|X_k|^2/|U|^2+\sum_{k=i+1}^{j-1}|X_k|^2/|U|^2+\sum_{k=j+1}^{n}|X_k|^2/|U|^2+|X_i \cup X_j|^2/|U|^2-\sum_{k=1}^{i-1}\sum_{l=1}^{m}|X_k \cap Y_l|^2/|U|^2-\sum_{k=i+1}^{j-1}\sum_{l=1}^{m}|X_k \cap Y_l|^2/|U|^2-\sum_{k=j+1}^{n}\sum_{l=1}^{m}|X_k \cap Y_l|^2/|U|^2-\sum_{l=1}^{m}|(X_i \cup X_j) \cap Y_l|^2/|U|^2$。

因此有，$GD(Q|P_1)-GD(Q|P_2)=|X_i|^2/|U|^2+|X_j|^2/|U|^2-|X_i \cup X_j|^2/|U|^2-\{\sum_{l=1}^{m}|X_i \cap Y_l|^2/|U|^2+\sum_{l=1}^{m}|X_j \cap Y_l|^2/|U|^2-\sum_{l=1}^{m}|(x_i \cup x_j) \cap Y_l)^2/|U|^2\}$。

由 $X_i \cap X_j=\Phi$，且集合的交运算对并运算满足分配律，则有：

$$|X_i \cup X_j|^2=(|X_i|+|X_j|)^2$$

$$|(X_i \cup X_j \cap Y_l|^2=|(X_i \cap Y_l) \cup (X_j \cap Y_l)|^2=(|X_i \cap Y_l|+|X_j \cap Y_l|)^2$$

故 $GD(Q|P_1)-GD(Q|P_2)=-2|X_i||X_j|/|U|^2+\sum_{l=1}^{m}2|X_i \cap Y_l||X_j \cap Y_l|/|U|^2$。

令 $|X_i|=x$，$|X_j|=y$，$|X_i \cap Y_l|=a_l x$，$|X_j \cap Y_l|=b_l y$，$l=1,2,\cdots,m$，这里 $x, y \geqslant 0$，$0 \leqslant a_l, b_l \leqslant 1$。

则 $GD(Q|P_1)-GD(Q|P_2)=-2xy/|U|^2+\sum_{l=1}^{m}2a_l b_l xy/|U|^2=-2xy/|U|^2(1-\sum_{l=1}^{m}a_l b_l)$。

由等价块及划分的性质得，$\sum_{l=1}^{m}a_l=1$，$\sum_{l=1}^{m}b_l=1$，故 $\sum_{l=1}^{m}(a_l+b_l)=2$，则有：

$$0 \leqslant \sum_{l=1}^{m}2\sqrt{a_l b_l} \leqslant \sum_{l=1}^{m}(a_l+b_l)=2$$

$$\Rightarrow 0 \leqslant \sum_{l=1}^{m}\sqrt{a_l b_l} \leqslant 1 \Rightarrow 0 \leqslant (\sum_{l=1}^{m}\sqrt{a_l b_l})^2 \leqslant 1$$

$$\Rightarrow 0 \leqslant \sum_{l=1}^{m}a_l b_l \leqslant (\sum_{l=1}^{m}\sqrt{a_l b_l})^2 \leqslant 1$$

$$\Rightarrow (1-\sum_{l=1}^{m}a_l b_l) \geqslant 0$$

从而有，$GD(Q|P_1)-GD(Q|P_2) \leqslant 0$，故 $GD(Q|P_1) \leqslant GD(Q|P_2)$。

定理 3.1 表明，在对 P_1 中任意两个等价块合并后，相对粒度单调增加。特别地，当 $X_i \cup X_j \subseteq Y_l$ 时，有 $GD(Q|P_1)=GD(Q|P_2)$。由此，我们可以得到下面的推论。

推论 3.1 在决策表 $S=(U,C,D,V,f)$ 中，对 $\forall a_i \in C$，$i=1,2,\cdots,m(m=|C|)$，则有：

$GD(D|\{a_1\}) \geqslant GD(D|\{a_1\} \cup \{a_2\}) \geqslant \cdots \geqslant GD(D|\{a_1\} \cup \{a_2\} \cup \cdots \cup \{a_m\})=GD(D|C)$

定理 3.2 在决策表 $S=(U,C,D,V,f)$ 中，对 $\forall a \in C$，若 $GD(D|C)=GD(D|C \setminus \{a\})$，则称 a 是 C 相对于决策属性 D 所不必要的，否则称 a 是必要的。

定义 3.2（属性重要度$_1$） 在决策表 $S=(U,C,D,V,f)$ 中，对于 $\forall a \in C$，定义属性 a 在 C 中相对于决策属性集 D 的重要性为：$Sig(a,C,D)=GD(D|C \setminus \{a\})-GD(D|C)$。

性质 3.1 属性 a 为 C 中相对于 D 所必要的属性，当且仅当 $Sig(a,C,D)>0$。

性质 3.2 $Core_C(D)=\{a \in C | Sig(a,C,D)>0\}$。

定义 3.3 设属性集 $P \subseteq C$，若 $\forall a \in P$，有 $Sig(a,P,D)>0$，则称 P 为独立的。

定理 3.3 在决策表 $S=(U,C,D,V,f)$ 中，$P \subseteq C$，若 $GD(D|P)=GD(D|C)$，且 P 独立，则称 P 为 C 的 D 相对约简。

定理 3.4 若决策表 $S=(U,C,D,V,f)$ 是一致的，即 $Pos_C(D)=U$。若 $P \subseteq C$，则有：

$Pos_P(D)=Pos_C(D) \Leftrightarrow GD(D|C)=GD(D|P)$。

证明：由定理 3.1 容易得证。

定理 3.4 表明，在一致决策表中，约简的相对粒度描述方法等价于约简的正区域描述方法。然而，在不一致决策表中，即 $Pos_C(D) \neq U$ 时，定理 3.4 的结论不成立，这时我们可以用约简的相对粒度描述方法弥补约简的正区域描述方法的局限性。下面以表 1 所列的不一致决策表为例来说明，其中条件属性集 $C=\{a,b,c\}$，决策属性集 $D=\{d\}$。

表 1 不一致决策表 1

U	a	b	c	d
x_1	0	3	3	1
x_2	1	3	3	2
x_3	1	3	3	3
x_4	1	3	0	2
x_5	1	3	0	3

我们根据文献[6]中基于正区域的约简方法求得表 1 的约简结果为$\{a\}$，然而，相对粒度 $GD(D|C) \neq GD(D|\{a\})$。

由分析可知，文献[6]中基于正区域的约简方法仅考虑决策表中是否产生新的不一致对象，而没有考虑决策表中原有不一致对象属于各决策类的隶属度是否发生变化，即文献[6]中基于正区域的约简方法仅考虑决策表中确定性规则的可信度是否发生变化，然而在实际决策应用中，决策规则的对象覆盖度也是衡量决策表分类能力的重要指标[7]。因此，文献[6]中基于正区域的约简方法在处理不一致决策表时存在一定的不足。然而，基于相对粒度的约简方法考虑约简后决策表的相对粒度是否发生变化，由定理 3.1 可知，引起相对粒度变化的因素有决策表是否出现新的不一致对象，以及原有不确定规则的可信度和对象覆盖度是否发生变化，因此本文提出的基于相对粒度的约简方法可以弥补基于正区域的约简方法的不足，有助于从决策表中获取最优或次优属性约简。

由于在以核为起点的属性约简过程中，我们往往通过一种度量方法不断地向属性核中增加属性来求取约简。为此，我们下面给出了属性重要性的另一种定义形式。

定义 3.4（属性重要度$_2$） 在决策表 $S=(U,C,D,V,f)$ 中，$C_0 \subseteq C$，定义 $\forall a \in C \setminus C_0$ 关于属性集 C_0 对 D 的重要性为：

$Sig(a,C_0,D)=GD(D|C_0)-GD(D|C_0 \cup \{a\})$

定义 3.4 表明属性 a 关于属性集 C_0 对 D 的重要性是由 C_0 中添加属性 a 后所引起的相对粒度变化的大小来度量的，即 $GD(D|C_0 \cup \{a\})$ 越小，则 $Sig(a,C_0,D)$ 的值就越大，表明属性 $a \in C \setminus C_0$ 关于属性集 C_0 对 D 就越重要。

由于核是任何约简的的交集，核是唯一的，并且由定义 3.2 容易求出决策表的相对核，因此，我们可以把核作为求最小约简的起点，令 $C_0 = Core$，通过不断地选取 $Sig(a,C_0,D)$ 最大的属性 a，即 $GD(D|C_0 \cup \{a\})$ 最小的属性 a 添加到 C_0 中，直到相对粒度 $GD(D|C_0) = GD(D|C)$。

4 基于相对粒度的属性约简算法

由上述理论可知，以相对粒度为启发式知识的属性约简算法可能需要反复计算 $GD(D|C_0 \cup \{a\})$，因此为了降低该算法的时间复杂度，本文算法利用文献[11]中计算划分的方法，并结合性质 2.2 的再划分方法，以渐增式方法计算相对粒度 $GD(D|C_0 \cup \{a\})$，从而减少了许多不必要的计算。下面给出以核为起点，基于相对粒度的启发式约简算法。

4.1 基于相对粒度的属性约简算法

输入：决策表 $S=(U,C,D,V,f)$，C 为条件属性集，D 为决策属性集。

输出：决策表的一个最小相对属性约简。

Step 1　计算 C 相对于 D 的核 $Core$，令 $C_0 = Core$；

Step 2　if $(GD(D|C) = GD(D|C_0))$ then go Step 8；

Step 3　对每个 $a_i \in C \setminus C_0$，执行 Step4－Step5；

Step 4　根据划分 U/C_0 计算 $U/\{C_0 \cup \{a\}\}$，$U/\{C_0 \cup \{a\} \cup D\}$；

Step 5　计算 $GD(D|C_0 \cup \{a\})$；

Step 6　选择使 $GD(D|C_0 \cup a_i)$ 最小的属性 a_i（若满足条件的属性有多个，则选择使 $GD(C_0 \cup a_i)$ 最小的属性 a_i，若使 $GD(C_0 \cup a_i)$ 最小的属性也不止一个，则从中任选一个属性）作为扩展属性，令 $C_0 = C_0 \cup a_i$；

Step 7　if $(GD(D|C) \neq GD(D|C_0))$ then go Step3；

Step 8　输出 C_0 为最小相对约简，算法结束。

由于计算划分 U/C 的时间复杂度为 $O(|C||U|)$，经分析本文算法可知，该算法总的时间复杂度为 $O(|C|^2|U|)+O((|C|+|C-1|+\cdots+1)(|D|+1+1)|U|)$。由于在通常情况下，决策系统中决策属性仅包含一个属性，因此本文算法总的时间复杂度为 $(|C|^2|U|)$，与文献[12]中算法的时间复杂度相同。

4.2 实例分析与比较

表 2 列出了一个不一致决策表[12]，其中 $U=\{x_1,x_2,\cdots,x_{10}\}$，条件属性集 $C=\{a_1,a_2,\cdots,a_5\}$，决策属性集 $D=\{d\}$。

利用本文算法求出表 2 的属性约简结果为 $\{a_1,a_2,a_5\}$，与文献[12]中基于包含度的不一致决策表约简算法得到的约简结果相一致。而文献[6]中基于正区域的约简算法获取属性约简为 $\{a_1,a_3,a_4,a_5\}$，而不是最小约简 $\{a_1,a_2,a_5\}$[12]。由比较可知，本文从知识粒度的角度出发，以相对粒度为启发式信息，提出的属性约简算法能有效地获取最优或次优相对约简。

表 2　不一致决策表 2

U	a_1	a_2	a_3	a_4	a_5	d
x_1	1	1	1	1	0	1
x_2	1	0	0	0	1	0
x_3	0	0	1	0	0	0
x_4	1	0	0	0	1	1
x_5	1	1	0	1	0	1
x_6	0	0	1	0	1	0
x_7	1	0	0	0	0	0
x_8	0	1	0	0	0	0
x_9	0	0	1	0	0	1
x_{10}	1	0	0	0	0	1

结束语　本文基于知识粒度的定义，在决策表中引入了知识的相对粒度和相对重要度的概念。为克服基于正区域的约简方法在处理不一致决策表时存在的不足，提出了一种基于相对粒度的启发式属性约简算法。理论分析和实例验证表明，该算法的时间复杂度相对较低，为从决策表中求取最小相对约简提供了一种有效方法。

参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Sciences, 1982, 11(5): 341-356

[2] Wang S K M, Ziarko W. On optimal decision rules in decision tables [J]. Bullet in of Polish accademy of Science, 1985, 333: 693-696

[3] Wang Guo-yin, Yu Hong, Yang Da-chun. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759-766

[4] Han Jian-chao, Hu Xiao-hua, Lin T Y. A New Computation Model for RoughSet Theory Based on Database Systems[J]. RSFDGrC, 2003, 2639: 114-121

[5] Xiao Jian-mei, Zhang Teng-fei. New Rough Set Approach to Knowledge Reduction in Decision Table[C]//Proceedings of the International Conference on Machine Learning and Cybernetics. Shanghai, 2004

[6] 刘少辉，盛秋戬，等. Rough 集高效算法的研究[J]. 计算机学报，2003, 26(5): 524-529

[7] 蒋思宇，卢炎生. 两种新的决策表属性约简概念[J]. 小型微型计算机系统，2006, 27(3): 512-515

[8] 苗夺谦，范世栋. 知识粒度的计算及其应用[J]. 系统工程理论与实践，2002(1): 48-56

[9] 孙林，徐久成，等. 基于新的条件熵的决策树规则提取方法[J]. 计算机应用，2007, 27(4): 884-887

[10] 高学东，丁军. 一种新的信息系统属性约简算法[J]. 系统工程理论与实践，2007(1): 131-136

[11] 徐章艳，刘作鹏，等. 一个复杂度为 max(O(|C||U|), O(|C|²|U|))的快速属性约简算法[J]. 计算机学报，2006, 29(3): 391-399

[12] 孙林，徐久成，等. 基于包含度的不一致决策表约简新方法[J]. 计算机工程与应用，2007, 43 (24): 166-168