第 37 卷 第 24 期
Vol.37 No.24
计 算 机 工 程
Computer Engineering
2011 年 12 月
December 2011
• 软件技术与数据库 • 文章编号：1000—3428(2011)24—0052—03 文献标识码：A 中图分类号：TP312

基于优势关系的启发式属性约简算法

廖 帆[1]，滕书华[2]，邵世雷[1]
(1. 中国人民解放军理工大学通信工程学院，南京 210007；2. 国防科学技术大学电子科学与工程学院，长沙 410073)

摘 要：根据优势原理，提出一种具有明确粗糙集理论含义的指标——优势度，用于度量序目标信息系统的协调程度。在证明优势度粒化单调性的基础上，给出属性集重要性度量函数，提出一种基于优势度的序目标信息系统启发式约简算法。该算法与经典粗糙集理论约简有相同的理论基础，易于理解。应用结果表明，该算法适用于优势关系下目标信息系统的知识发现。

关键词：粗糙集；属性约简；优势关系；不协调信息系统；属性重要性

Heuristic Algorithm for Attribute Reduction Based on Dominance Relation

LIAO Fan[1], TENG Shu-hua[2], SHAO Shi-lei[1]
(1. Institute of Communications Engineering, PLA University of Science & Technology, Nanjing 210007, China;
2. Institute of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China)

【Abstract】A new uncertainty measure, such as dominance degree is proposed in ordered objective information systems based on dominance principle, and an explicit theoretical meaning of rough set is given to the dominance degree which can be used to measure the inconsistence of objective information system. The granulation monotonicity of dominance degree is proved, based on which a new measure of attribution importance is designed. An heuristic reduct algorithm in objective information system is provided based on dominance relation. An example illustrates the validity of this algorithm, and results show that the algorithm has the same theoretical foundation with classical reduct algorithm in rough set theory, and it is easily understood. The algorithm provides an important theoretical basis for knowledge discovery in ordered objective information systems.

【Key words】rough set; attribute reduction; dominance relation; inconsistent information system; attribute importance

DOI: 10.3969/j.issn.1000-3428.2011.24.017

1 概述

粗糙集理论是一种新的处理不精确、不相容和不完全数据的数学工具。属性约简是粗糙集理论的核心问题之一。通过属性约简去掉不必要的属性，可以使知识表示简化，又不丢失基本信息。现有的属性约简算法[1-3]主要是在等价关系下求得的，在实际中有许多信息系统是基于优势关系的，而且是不协调的。要从这种复杂的基于优势关系的不协调信息系统中获取简洁的不确定性命题，必须对此类系统进行属性约简。优势关系下不协调信息系统的约简具有重要的实际应用价值[4]。但目前针对优势关系下不协调信息系统的约简研究还比较匮乏[5-7]。

文献[8]将 Shannon 熵引入序信息系统作为知识的粗糙熵，用于度量序信息系统中属性集的不确定性，进而建立了基于优势关系的信息系统约简算法，但粗糙熵在粗糙集理论中缺乏直观的解释和明确含义，且此约简算法并不适用于目标信息系统。本文根据优势关系下目标信息系统中知识的分类能力，提出一种具有明确粗糙集理论含义的、用于度量序目标信息系统协调程度的测度，在此基础上给出一种序目标信息系统的启发式约简算法。

2 粗糙集基本概念

粗糙集基本概念详见文献[9]，以下仅对本文相关概念进行描述。

定义 1 四元组 $S=(U,C,D,V,f)$ 是一个目标信息系统，其中，$U=\{u_1,u_2,\cdots,u_{|U|}\}$ 是对象的非空有限集合，称为论域；$C=\{a_1,a_2,\cdots,a_{|C|}\}$ 是有限条件属性集，也称为知识；D 是有限目标属性集；$V=\bigcup_{a_j\in C}V_{a_j}$，$V_{a_j}$ 表示属性 a_j 的值域；$f:U\times A\to V$ 是信息函数，对于 $\forall u_i\in U$ 和 $\forall a_j\in C$，$f(u_i,a_j)\in V_{a_j}$。

定义 2 在目标信息系统 $S=(U,C,D,V,f)$ 中，若在某个属性值域上建立了偏序关系，则称这个属性为一个准则。当所有的属性都为准则时，该系统称为序目标信息系统，通常简写为 $S=(U,C,D)$。

在序目标信息系统 $S=(U,C,D)$ 中，在 $a_k\in C$ 的值域上建立的偏序关系假设为“$\prec$”。对于对象 $u_i,u_j\in U$，$u_i\prec_{a_k}u_j$ 表示 u_j 至少与 u_i 关于准则 a_k 是一样好的。不失一般性，取属性的值域为实数，即 $V_{a_k}\subseteq\mathbf{R}$。因此，$u_i\prec_{a_k}u_j\Leftrightarrow f(u_i,a_k)\prec f(u_j,a_k)$。对于属性集 $P\subseteq C$，$u_i\prec_P u_j$ 表示 u_j 关于属性集 P 中的所有准则都优于 u_i。

定义 3 在序目标信息系统 $S=(U,C,D)$ 中，$P\subseteq C\cup D$ 对应的优势关系 $R_P^{\prec}$ 定义为：

$$R_P^{\prec}=\left\{\left(u_i,u_j\right)\in U\times U\,\middle|\,\forall a_k\in P,f\left(u_i,a_k\right)\prec f\left(u_j,a_k\right)\right\}$$

显然，优势关系是自反和传递的，但未必是对称的。利

基金项目：国家自然科学基金资助项目(40901216)
作者简介：廖 帆(1986－)，男，硕士研究生，主研方向：数据挖掘，约简算法；滕书华，博士；邵世雷，副教授
收稿日期：2011-05-16 **E-mail：**liaoif12@gmail.com

用优势关系得到的优势信息粒度记为：

$$R_P^{\succ}(u_i)=\left\{u_j \middle| u_j\in U\wedge(u_i,u_j)\in R_P^{\succ}\right\}$$

称 $R_P^{\succ}(u_i)$ 为对象 u_i 的优势类。$U/R_P^{\succ}=\{R_P^{\succ}(u_i)|u_i\in U\}$ 表示属性集 P 对该序信息系统论域的一个分类。

定义 4 在序目标信息系统 $S=(U,C,D)$ 中，若 $R_C^{\succ}\subseteq R_D^{\succ}$，则称该序目标信息系统是协调的，否则是不协调的。

定义 5 在序目标信息系统 $S=(U,C,D)$ 中，$P,Q\subseteq C$。

(1)如果对 $\forall u_i\in U$，都有 $R_P^{\succ}(u_i)=R_Q^{\succ}(u_i)$，则称分类 $U/R_P^{\succ}$ 等于分类 $U/R_Q^{\succ}$，并记作 $U/R_P^{\succ}=U/R_Q^{\succ}$。

(2)如果对 $\forall u_i\in U$，都有 $R_P^{\succ}(u_i)\subseteq R_Q^{\succ}(u_i)$，则称分类 $U/R_P^{\succ}$ 细于分类 $U/R_Q^{\succ}$，并记作 $U/R_P^{\succ}\subseteq U/R_Q^{\succ}$。

3 优势度的定义及性质

文献[5]指出，在序目标信息系统 $S=(U,C,D)$ 中，$u_i\in U$，关于条件属性集 C 比 u_i 优的对象的关于目标属性 D 也应该比 u_i 优，并称这一原理为优势原理。据此给出如下定义：

定义 6 对于 $\forall u_i,u_j\in U$，若对象 u_j 关于条件属性集 C 比 u_i 优，关于目标属性 D 也比 u_i 优，则称 (u_i,u_j) 为序信息系统中的协调有序对；否则，若对象 u_j 关于 C 比 u_i 优，而对象 u_i 关于 D 比 u_j 优，则称 (u_i,u_j) 为序信息系统中的不协调有序对。

定义 4 和定义 6 表明，若序信息系统中不存在不协调有序对，则此序信息系统是协调的。显然，序信息系统中不协调有序对个数越多，系统不协调程度越强。据此给出序信息系统不协调程度的度量。

定义 7 在序目标信息系统 $S=(U,C,D)$ 中，$P\subseteq C$。知识 D 相对于知识 P 的优势度 $DOM(R_D^{\succ}/R_P^{\succ})$ 定义为：

$$DOM(R_D^{\succ}/R_P^{\succ})=\left|R_P^{\succ}-R_D^{\succ}\cap R_P^{\succ}\right| \tag{1}$$

其中，$R_D^{\succ}\cap R_P^{\succ}$ 是知识 P 下序目标信息系统中协调有序对集合；$R_P^{\succ}-R_D^{\succ}\cap R_P^{\succ}$ 则为序目标信息系统中不协调有序对集合。因此，优势度描述了在知识 P 下序目标信息系统的不协调程度，$DOM(R_D^{\succ}/R_P^{\succ})$ 越大，序目标信息系统包含的不协调有序对个数越多，序目标信息系统的不协调程度越高。

定理 1 在序目标系统 $S=(U,C,D)$ 中，$P,Q\subseteq C$。如分类 $U/R_P^{\succ}$ 细于分类 $U/R_Q^{\succ}$，则 $DOM(R_D^{\succ}/R_Q^{\succ})\geqslant DOM(R_D^{\succ}/R_P^{\succ})$。

证明：由 $R_D^{\succ}\cap R_P^{\succ}\subseteq R_P^{\succ}$、$R_D^{\succ}\cap R_Q^{\succ}\subseteq R_Q^{\succ}$ 及定义 7 得：

$$DOM(R_D^{\succ}/R_Q^{\succ})-DOM(R_D^{\succ}/R_P^{\succ})=\left|R_Q^{\succ}\right|-\left|R_D^{\succ}\cap R_Q^{\succ}\right|-\left|R_P^{\succ}\right|+\left|R_D^{\succ}\cap R_P^{\succ}\right|$$

因为分类 $U/R_P^{\succ}$ 细于分类 $U/R_Q^{\succ}$，即对于 $\forall u_i\in U$，都有 $R_P^{\succ}(u_i)\subseteq R_Q^{\succ}(u_i)$，所以 $R_P^{\succ}\subseteq R_Q^{\succ}$，因此有：

$$DOM(R_D^{\succ}/R_Q^{\succ})-DOM(R_D^{\succ}/R_P^{\succ})=\left|R_Q^{\succ}-R_P^{\succ}\right|-\left|R_D^{\succ}\cap R_Q^{\succ}-R_D^{\succ}\cap R_P^{\succ}\right|$$

因为 $R_D^{\succ}\cap R_Q^{\succ}-R_D^{\succ}\cap R_P^{\succ}=R_D^{\succ}\cap(R_Q^{\succ}-R_P^{\succ})\subseteq R_Q^{\succ}-R_P^{\succ}$，所以 $DOM(R_D^{\succ}/R_Q^{\succ})\geqslant DOM(R_D^{\succ}/R_P^{\succ})$ 成立。

定理 1 表明，属性集 P 分类越细，目标属性 D 和条件属性集 P 共同确定的协调有序对数目越多，即 $DOM(R_D^{\succ}/R_P^{\succ})$ 越小，P 和 D 之间的不协调程度越低。因此，优势度描述了条件属性集和目标属性集之间的不协调程度。

定理 2 在序目标信息系统 $S=(U,C,D)$ 中，$P\subseteq C$，优势度的表达式为：

$$DOM(R_D^{\succ}/R_P^{\succ})=\sum_{i=1}^{|U|}\left|R_P^{\succ}(u_i)\right|-\sum_{j=1}^{|U|}\left|R_{P\cup D}^{\succ}(u_j)\right| \tag{2}$$

证明：由定义 3 可知 $R_{P\cup D}^{\succ}=R_P^{\succ}\cap R_D^{\succ}$ 且 $\left|R_P^{\succ}\right|=\sum_{i=1}^{|U|}R_P^{\succ}(u_i)$，又因为 $R_D^{\succ}\cap R_P^{\succ}\subseteq R_P^{\succ}$，$\left|R_P^{\succ}-R_P^{\succ}\cap R_D^{\succ}\right|=\left|R_P^{\succ}\right|-\left|R_P^{\succ}\cap R_D^{\succ}\right|$，所以式(2)成立。

注：如果 $P=\varnothing$，则令 $\left|R_P^{\succ}(u_i)\right|=|U|$。

4 基于优势度的属性约简

下面先给出序目标信息系统中属性重要性测度及相对约简概念，然后给出基于优势度的启发式属性约简算法步骤。

推论 在序目标信息系统 $S=(U,C,D)$ 中，$Q\subseteq P\subseteq C$，则 $DOM(R_D^{\succ}/R_Q^{\succ})\geqslant DOM(R_D^{\succ}/R_P^{\succ})$。

推论表明，$DOM(R_D^{\succ}/R_Q^{\succ})$ 随着条件属性集 Q 中元素个数的增加单调下降，利用该粒化单调性，可在优势关系下构建基于前向添加搜索策略的约简算法。下面给出优势关系下属性重要性测度的定义。

定义 8 在序目标信息系统 $S=(U,C,D)$ 中，$Q\subseteq C$。对于 $\forall a_i\in(C-Q)$，属性 a_i 相对于属性集 Q 的重要性测度 $SIG(a_i,Q,D)$ 定义为：

$$SIG(a_i,Q,D)=DOM(R_D^{\succ}/R_Q^{\succ})-DOM(R_D^{\succ}/R_{Q\cup\{a_i\}}^{\succ}) \tag{3}$$

在定义 8 中，$SIG(a_i,Q,D)$ 描述了向属性集 Q 中添加属性 a_i 后不协调有序对的减少量。$SIG(a_i,Q,D)$ 越大，说明 a_i 相对于属性集 Q 越重要。在前向添加约简过程中，可选择使 $SIG(a_i,Q,D)$ 最大的条件属性作为约简元素。据此，给出优势关系下基于优势度的属性约简的定义。

定义 9 在序目标信息系统 $S=(U,C,D)$ 中，$Q\subseteq C$。Q 是 C 相对于 D 的一个相对约简，当且仅当其满足：

(1) $DOM(R_D^{\succ}/R_Q^{\succ})=DOM(R_D^{\succ}/R_C^{\succ})$；

(2) $\forall q_i\in Q$，$DOM(R_D^{\succ}/R_{Q-\{q_i\}}^{\succ})>DOM(R_D^{\succ}/R_C^{\succ})$。

由定义 9 可知，基于优势度的属性约简目标就是寻找保持序目标信息系统中协调有序对个数不变的最小条件属性集合。

基于优势度的属性约简算法(Attribute Reduction Algorithm Based on the Dominance Degree, ARA-DD)的步骤如下：

输入 序目标信息系统 $S=(U,C,D)$

输出 相对约简 Q

Step1 计算 $DOM(R_D^{\succ}/R_C^{\succ})$。

Step2 令 $Q=\varnothing$，Q 为已选择属性的集合。

Step3 对 $\forall a_i\in(C-Q)$，计算 $SIG(a_i,Q,D)$。

Step4 选择满足 $SIG(a_k,Q,D)=\max\{SIG(a_i,Q,D),a_i\in(C-Q)\}$ 的属性 a_k 加入到约简集合中，若这样的属性不止一个，则选择 $\left|R_{\{a_i\}}^{\succ}\right|$ 最大的属性为 a_k，令 $Q=Q\cup\{a_k\}$。

Step5 如果 $DOM(R_D^{\succ}/R_Q^{\succ})=DOM(R_D^{\succ}/R_C^{\succ})$，则 Q 即为所求的约简；否则，转 Step3。

5 实例分析

下面通过实例来验证 ARA-DD 算法处理序目标信息系统的有效性。

序目标信息系统 $S=(U,C,D)$ 如表 1 所示，其中，$U=$是 $\{u_1,u_2,\cdots,u_{11}\}$；$C=\{a_1,a_2,\cdots,a_6\}$；$D=\{d\}$。目标属性把论域分为 3 类：

$Cl_0 < Cl_1 < Cl_2$, $Cl_0 = \{u_6, u_7, u_8, u_{11}\}$

$Cl_1 = \{u_2, u_4, u_9, u_{10}\}$

$Cl_2 = \{u_1, u_3, u_5\}$

表 1 序目标信息系统

U	A						
	a_1	a_2	a_3	a_4	a_5	a_6	d
u_1	2	3	3	3	3	3	3
u_2	2	3	2	1	2	2	2
u_3	3	3	3	3	3	3	3
u_4	3	2	1	2	2	2	2
u_5	3	2	3	3	3	3	3
u_6	1	2	2	1	1	1	1
u_7	1	2	1	1	2	2	1
u_8	1	1	1	2	1	1	1
u_9	2	1	2	1	2	2	2
u_{10}	2	1	1	2	1	1	2
u_{11}	1	1	1	1	1	1	1

由算法 ARA-DD 求表 1 的约简过程如下：

Step1 求得 $DOM(R_D^{\prec}/R_C^{\prec}) = 0$。

Step2 令 $Q = \varnothing$。

Step3 计算每个属性的重要性：

$SIG(a_1,Q,D) = 34$, $SIG(a_2,Q,D) = 24$, $SIG(a_3,Q,D) = 33$.

$SIG(a_4,Q,D) = 28$, $SIG(a_5,Q,D) = 33$, $SIG(a_6,Q,D) = 33$

Step4 求得 $a_k = a_1$，令 $Q = \{a_1\}$。

Step5 因 $\left|DOM(R_D^{\prec}/R_Q^{\prec})\right| - \left|DOM(R_D^{\prec}/R_C^{\prec})\right| = 6$。

Step6 计算 $\{a_2, a_3, \cdots, a_6\}$ 中每个属性的重要性，可得如下结果：$SIG(a_2,Q,D) = 4$，$SIG(a_3,Q,D) = 6$，$SIG(a_4,Q,D) = 6$，$SIG(a_5,Q,D) = 6$，$SIG(a_6,Q,D) = 6$。

Step7 由于属性 $a_3 \sim a_6$ 的重要性测度同时取得了最大值，因此要做如下比较：$\left|R_{\{a_3\}}^{\prec}\right| = \left|R_{\{a_5\}}^{\prec}\right| = \left|R_{\{a_6\}}^{\prec}\right| = 81 < \left|R_{\{a_4\}}^{\prec}\right| = 82$，显然 $\left|R_{\{a_4\}}^{\prec}\right|$ 最大，令 $a_k = a_4$，$Q = \{a_1, a_4\}$。

Step8 由于 $DOM(R_D^{\prec}/R_Q^{\prec}) = DOM(R_D^{\prec}/R_C^{\prec})$，算法停止，因此 $Q = \{a_1, a_4\}$ 即为所求的相对约简。

6 结束语

本文根据优势原理提出了一种具有明确粗糙集理论含义的序目标信息系统协调程度的度量——优势度，它根据属性分类能力很好地描述了条件属性集和目标属性集之间的不协调程度。根据优势度的粒化单调性，引入一种新的属性重要性测度，提出了一种基于优势度的序目标信息系统启发式约简算法，该算法与经典粗糙集约简的理论相一致，即都保持了原有目标信息系统的协调程度不变。实例表明该算法对目标信息系统的约简是行之有效的。

目前，学者们从理论上基于不同角度定义了多种序信息系统约简算法，但在实际应用中的效果并不理想。由于本文的约简算法与经典粗糙集有相同的理论基础，因此下一步工作是利用本文方法解决序目标信息系统的实际问题。

参考文献

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

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

[3] 滕书华，魏荣华，孙即祥，等. 基于不可区分度的启发式快速完备约简算法[J]. 计算机科学, 2009, 36(8): 196-200.

[4] Greco S, Matarazzo B, Slowinski R. Rough Approximation of a Preference Relation by Dominance Relations[J]. European Journal of Operational Research, 1999, 117(1): 63-83.

[5] Greco S, Matarazzo B, Slowinski R. Rough Approximation by Dominance Relations[J]. International Journal of Intelligent Systems, 2002, 17(2): 153-171.

[6] 陈 娟，王国胤，胡 军. 优势关系下不协调信息系统的正域约简[J]. 计算机科学, 2008, 35(3): 216-218, 227.

[7] 袁修久，何华灿. 优势关系下的相容约简和下近似约简[J]. 西北工业大学学报, 2006, 24(5): 604-608.

[8] 徐伟华，张晓燕，钟坚敏，等. 序信息系统中属性约简的启发式算法[J]. 计算机工程, 2010, 36(17): 69-71.

[9] 张文修，姚一豫，梁 怡. 粗糙集与概念格[M]. 西安：西安交通大学出版社, 2006.

编辑 张正兴

(上接第 51 页)

参考文献

[1] Özsu M T, Valduriez P. Principles of Distributed Database Systems[M]. [S. l.]: Pearson Education, 2002.

[2] Zubi Z S. On Distributed Database Security Aspects[C]//Proc. of 2009 International Conference on Multimedia Computing and Systems. Quarzazate, Morocco: IEEE Press, 2009: 231-235.

[3] Dekker M A C, Crampton J, Etalle S. RBAC Administration in Distributed Syatems[C]//Proc. of the 13th ACM Symposium on Access Control Models and Technologies. Estes Park, Colorado, USA: ACM Press, 2008: 93-101.

[4] Park J, Sandhu R. The UCONABC Usage Control Model[J]. ACM Transactions on Information and System Security, 2004, 7(1): 128-174.

[5] Zhang Xinwen. Formal Model and Analysis of Usage Control[D]. Fairfax, Virginia, USA: George Mason University, 2006.

[6] Pretschner A, Hilty M, Basin D. Distributed Usage Control[J]. Communications of the ACM, 2006, 49(9): 39-44.

[7] Zhang Xinwen, Seifert J P, Sandhu R. Security Enforcement Model for Distributed Usage Control[C]//Proc. of 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing. Taichung, China: IEEE Computer Society, 2008: 10-18.

[8] 桂劲松，陈志刚，胡玉平. 服务网格授权决策的 UCONA 模型[J]. 计算机工程, 2009, 35(2): 70-73.

编辑 金胡考