第36卷 第17期 计 算 机 工 程 2010年9月
Vol.36 No.17 Computer Engineering September 2010
·软件技术与数据库· 文章编号：1000—3428(2010)17—0069—03 文献标识码：A 中图分类号：TP311

序信息系统中属性约简的启发式算法

徐伟华[1]，张晓燕[1]，钟坚敏[1]，张文修[2]
(1. 重庆理工大学数学与统计学院，重庆 400054；2. 西安交通大学理学院，西安 710049)

摘 要：基于序信息系统的知识粗糙熵，在系统中引入属性重要性的概念，利用该测度能度量序信息系统中属性集的不确定性，基于此，提出序信息系统中基于知识粗糙熵的启发式约简算法。通过实例对该方法的有效性进行检验，结果显示该算法可以作为一种有效的数据挖掘工具，为序信息系统的知识发现提供理论基础。

关键词：粗糙集；信息系统；优势关系；启发式算法

Heuristic Algorithm for Attributes Reduction in Ordered Information Systems

XU Wei-hua[1], ZHANG Xiao-yan[1], ZHONG Jian-min[1], ZHANG Wen-xiu[2]
(1. School of Mathematics and Statistics, Chongqing University of Technology, Chongqing 400054;
2. School of Sciences, Xi'an Jiaotong University, Xi'an 710049)

【Abstract】A definition of attribute significance is proposed based on knowledge rough entropy in ordered information systems, and important properties are obtained. It can be found that using the definition can measure uncertainty of an attribute set in the ordered information systems. A heuristic algorithm for attributes reduction is acquired in the systems. An example illustrates the validity of this algorithm, and results show that the algorithm is an efficient tool for data mining, and provides an important theoretical basis for knowledge discovery in ordered information systems.

【Key words】rough set; information system; dominance relation; heuristic algorithm

1 概述

粗糙集理论[1]是近年来发展起来的一种处理不精确性、不确定性和模糊知识的软计算工具，已被成功应用于人工智能、数据挖掘、模式识别与智能信息处理等领域，并越来越引起国际学术界的关注。经典粗糙集以完备信息系统为研究对象，以等价关系(满足自反性、对称性、传递性)为基础，通过等价关系将论域分成互不相交的等价类，划分越细，知识越丰富，信息越充分。

知识约简是粗糙集理论的核心问题之一。在实际的知识库中，描述知识的属性并不是同等重要的，甚至其中有些属性是冗余的。所谓知识约简，就是在保持知识库分类能力不变的条件下，删除其中不相关或不重要的属性。通过知识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。目前，许多学者通过不同的方法从不同的角度对知识约简做了深入的研究，并取得了很多成果[2-7]。

然而，这些研究主要是在等价关系下的信息系统中进行的，在实际问题中，有许多信息系统由于各种原因(如噪声、信息缺损等)是基于优势关系的，而且是不协调的。要想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就必须对系统进行知识约简。因而，对于优势关系下的不协调目标信息系统知识约简的研究意义重大[8-13]。为此，本文对这一问题进行探讨研究，通过研究序信息系统中知识与粗糙熵之间的关系，分析粗糙熵的属性重要性，提出序信息系统中基于知识粗糙熵的启发式约简算法。

2 粗糙集与序信息系统

信息系统有时也叫数据表或知识表示系统等，其主要是通过一张表来反映对象与属性之间的关系。下面先给出有关基本概念。

定义 1[7] 称一个三元组 $I^{\succ}=(U,A,F)$ 为一个信息系统，其中，U 是有限对象集，$U=\{x_1,x_2,\cdots,x_n\}$；A 是有限属性集，$A=\{a_1,a_2,\cdots,a_p\}$；F 是 U 与 A 的关系集，$F=\{f_k:U\to V_k,\ k\leqslant p\}$，$V_k$ 是 a_k 的有限值域；

在 Pawlak 近似空间意义下的信息系统，对每个属性集就决定了一个二元不可区分关系，即等价关系。然而，在实际生活中，有许多系统并不是基于等价关系的，有不少是基于优势关系的，即对每个属性值域有按照递增或者递减的一个偏序关系，如一个班级的各科成绩情况等问题。这时就需要建立基于优势关系下的信息系统，即序信息系统。

定义 2[7] 在一个信息系统中，如果在某个属性值域上建立了偏序关系，称这个属性为一个准则。当所有的属性都为准则时，该信息系统称为序信息系统。

设在信息系统 (U,A,F) 中属性 a 是一个准则，并且在 a 的值域上建立的偏序关系是“$\succcurlyeq_a$”。于是对于对象 x,y，说

基金项目：重庆市教委科学技术研究基金资助项目“优势关系下信息系统知识获取的方法研究”(KJ090612)；重庆市九龙坡区科学计划研究基金资助项目“粒计算理论及其在农作物疾病预防中的应用研究”(2008Q98)

作者简介：徐伟华(1979－)，男，副教授、博士，主研方向：粗糙集理论与应用，不确定性推理；张晓燕，讲师、硕士；钟坚敏，副教授；张文修，教授、博士生导师

收稿日期：2010-03-19 **E-mail：**chxuwh@gmail.com

$x \succcurlyeq_a y$ 表示 x 至少和 y 关于准则 a 是一样好的，或者说 x 优于 y。

不失一般性，本文取属性的值域为实数，即 $V_k \subseteq R$（R 表示实数集)。定义 $x \succcurlyeq_a y$ 为 $x \succcurlyeq_a y \Leftrightarrow f(x,a) \geqslant f(y,a)$。于是，对于属性集 $B \subseteq A$，$x \succcurlyeq_B y$ 是指 x 关于属性集 B 中的所有准则都优于 y。一般，序信息系统用 $I^{\succ} = (U, A, F)$ 来表示。

定义 3[7] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，对于 $B \subseteq A$，令 $R_B^{\succcurlyeq} = \{(x, y) \in U \times U : f_l(x) \geqslant f_l(y), \forall a_l \in B\}$，则 $R_B^{\succcurlyeq}$ 称为序信息系统 $I^{\succ} = (U, A, F)$ 的优势关系。

若记：

$$[x_i]_B^{\succcurlyeq} = \{x_j \in U \mid (x_j, x_i) \in R_B^{\succcurlyeq}\} = \{x_j \in U \mid f_l(x_j) \geqslant f_l(x_i),\ \forall a_l \in B\}$$

$$U / R_B^{\succcurlyeq} = \{[x_i]_B^{\succcurlyeq} \mid x_i \in U\}$$

则称 $[x_i]_B^{\succcurlyeq}$ 为对象 x_i 的优势类，$U / R_B^{\succcurlyeq}$ 为该序信息系统对象集关于属性集 B 的一个分类。

易见，优势关系有如下性质：

命题 1[7] 设 $I^{\succ} = (U, A, F)$ 为序信息系统，则如下命题成立：

(1) $R_B^{\succcurlyeq}$ 是自反的和传递的，未必是对称的，因而一般不再是等价关系；

(2)当 $B \subseteq A$ 时有：$R_A^{\succcurlyeq} \subseteq R_B^{\succcurlyeq}$；

(3)当 $B \subseteq A$ 时有：$[x_i]_A^{\succcurlyeq} \subseteq [x_i]_B^{\succcurlyeq}$；

(4)当 $x_j \in [x_i]_A^{\succcurlyeq}$ 时有：$[x_j]_B^{\succcurlyeq} \subseteq [x_i]_A^{\succcurlyeq}$；

(5) $|[x_j]_A^{\succcurlyeq}| \subseteq |[x_i]_A^{\succcurlyeq}|$ 当且仅当 $f(x_i, a) = f(x_j, a), (\forall a \in A)$；

(6)对于任意的 $x_i \in U$，有 $|[x_i]_A^{\succcurlyeq}| \geqslant 1$；

(7) $U / R_A^{\succcurlyeq}$ 形成了 U 的一个覆盖，即：对于任意的 $x \in U$，有 $[x_i]_A^{\succcurlyeq} \neq \phi$ 且 $\bigcup_{x \in U} [x]_A^{\succcurlyeq} = U$；

其中，$|\cdot|$ 表示集合的势。

为了更好地反映序信息系统中知识的关系，本文给出如下定义：

定义 4[7] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B, C \subseteq A$。

(1)如果对于任意的 $x \in U$ 有 $[x]_B^{\succcurlyeq} = [x]_C^{\succcurlyeq}$，则称分类 $U / R_B^{\succcurlyeq}$ 等于分类 $U / R_C^{\succcurlyeq}$，并记作：$U / R_B^{\succcurlyeq} = U / R_C^{\succcurlyeq}$。

(2)如果对于任意的 $x \in U$ 有 $[x]_B^{\succcurlyeq} \subseteq [x]_C^{\succcurlyeq}$，则称分类 $U / R_B^{\succcurlyeq}$ 细于分类 $U / R_C^{\succcurlyeq}$，并记作：$U / R_B^{\succcurlyeq} \subseteq U / R_C^{\succcurlyeq}$。

(3)如果对于任意的 $x \in U$ 有 $[x]_B^{\succcurlyeq} \subseteq [x]_C^{\succcurlyeq}$，且对于某些 $y \in U$ 有 $[y]_B^{\succcurlyeq} \neq [y]_C^{\succcurlyeq}$，则称分类 $U / R_B^{\succcurlyeq}$ 分类真细于 $U / R_C^{\succcurlyeq}$，并记作：$U / R_B^{\succcurlyeq} \subset U / R_C^{\succcurlyeq}$。

显然由命题 1 和上面的定义，对于序信息系统 $I^{\succ} = (U, A, F)$ 以及 $B \subseteq A$，可以立即得到 $U / R_A^{\succcurlyeq} \subseteq U / R_B^{\succcurlyeq}$。

与经典粗糙集类似，序信息系统也可以定义上下近似这一对算子。

对于任意 $X \subseteq U$，定义 X 关于优势关系下 $R_B^{\succcurlyeq}$ 的下近似和上近似分别定义为

$$\underline{R_B^{\succcurlyeq}}(X) = \{x_i \in U : [x_i]_B^{\succcurlyeq} \subseteq X\}$$

$$\overline{R_B^{\succcurlyeq}}(X) = \{x_i \in U : [x_i]_B^{\succcurlyeq} \cap X \neq \phi\}$$

优势关系下的上、下近似也满足类似于 Pawlak 近似空间中的许多性质，详细请参考文献[7]。

例 1[7] 表 1 给出了某旅游公司投资项目的序信息系统示例，其中，对象集 $U = (x_1, x_2, \cdots, x_6)$ 表示被考察项目集合；准则集 $A = (a_1, a_2, a_3)$ 表示项目指标集，分别指地域环境、人文环境以及交通环境。

表 1　某序信息系统的对象集与准则集

U	a_1	a_2	a_3
x_1	1	2	1
x_5	3	2	2
x_3	1	1	2
x_4	2	1	3
x_5	3	3	2
x_6	3	2	3

按照优势关系的定义有：

$[x_1]_A^{\succcurlyeq} = \{x_1, x_2, x_5, x_6\}$

$[x_2]_A^{\succcurlyeq} = \{x_2, x_5, x_6\}$

$[x_3]_A^{\succcurlyeq} = \{x_2, x_3, x_4, x_5, x_6\}$

$[x_4]_A^{\succcurlyeq} = \{x_4, x_6\}$

$[x_5]_A^{\succcurlyeq} = \{x_5\}$

$[x_6]_A^{\succcurlyeq} = \{x_6\}$

不难验证上述关于序信息系统的性质，本文不再熬述。

3　序信息系统中知识的粗糙熵与属性重要性

3.1　序信息系统中知识的粗糙熵

文献[13]给出了序信息系统中引入知识粗糙熵的概念，并建立了知识粗糙熵和知识粗糙性的关系。为了后文方便描述，本文列出有关内容。

定义 5[7] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B \subseteq A$，则知识 B 的粗糙熵定义为

$$E(B) = \sum_{i=1}^{|U|} \frac{|[x_i]_B^{\succcurlyeq}|}{|U|} \cdot \mathrm{lb} |[x_i]_B^{\succcurlyeq}|$$

例 2[13] 由于上面定义，本文可以计算例 1 的序信息系统中知识 $A = \{a_1, a_2, a_3\}$ 的粗糙熵为

$$E(A) = \frac{4}{6} \cdot \mathrm{lb}4 + \frac{3}{6} \cdot \mathrm{lb}3 + \frac{5}{6} \cdot \mathrm{lb}5 + \frac{2}{6} \cdot \mathrm{lb}2 + \frac{1}{6} \cdot \mathrm{lb}1 + \frac{1}{6} \cdot \mathrm{lb}1 = \frac{2}{3} \cdot 2 + \frac{1}{2} \cdot \mathrm{lb}3 + \frac{5}{6} \cdot \mathrm{lb}5 + \frac{1}{3} = 4.390\,49$$

由定义 5 易得：

命题 2[13] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B \subseteq A$，则如下命题成立：

(1) $E(B)$ 取最大值 $|U| \cdot \mathrm{lb}|U|$ 当且仅当 $U / R_B^{\succcurlyeq} = U$；

(2) $E(B)$ 取最小值 0 当且仅当 $U / R_B^{\succcurlyeq} = \{\{x_1\}, \{x_2\}, \cdots, \{x_{|U|}\}\}$。

由上面的关于知识的粗糙熵的性质，可知在知识 B 下不能区分论域中任意 2 个对象，那么知识 B 的粗糙性最大，如果在知识 B 下能够区分论域中任意的对象，那么知识 B 达到了最精确程度，这与直观解释是完全一致的。

定理 1[13] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B_1, B_2 \subseteq A$，若 $U / R_{B_1}^{\succcurlyeq} \subset U / R_{B_2}^{\succcurlyeq}$，则有 $E(B_1) < E(B_2)$。

推论 1[13] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B_1, B_2 \subseteq A$，若 $B_2 \subseteq B_1$，则有 $E(B_1) \leqslant E(B_2)$。

由上面定理可知序信息系统中知识的粗糙熵随着分辨能力的增强单调减少。

定理 2[13] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B_1, B_2 \subseteq A$。若 $U / R_{B_1}^{\succcurlyeq} = U / R_{B_2}^{\succcurlyeq}$，则有 $E(B_1) = E(B_2)$。

定理 3[13] 设 $I^{\succ} = (U, A, F)$ 为一序信息系统，且 $B_1, B_2 \subseteq A$。若 $U / R_{B_1}^{\succcurlyeq} \subseteq U / R_{B_2}^{\succcurlyeq}$，且 $E(B_1) = E(B_2)$，则有 $U / R_{B_1}^{\succcurlyeq} = U / R_{B_2}^{\succcurlyeq}$。

推论 2[13] 设 $I^{\succ}=(U,A,F)$ 为一序信息系统，且 $B_1,B_2\subseteq A$。若 $B_2\subseteq B_1$，且 $E(B_1)=E(B_2)$，则有 $U/R^{\succeq}_{B_1}=U/R^{\succeq}_{B_2}$。

由以上定理 3 可知，如果 2 个知识表示之间存在包含关系，而又他们的知识粗糙熵相同，那么这 2 个知识表示是一样的，即 2 个分类是完全相同的。

3.2 序信息系统中属性的重要性

本节给出序信息系统属性重要性的概念，它是下文启发式算法的基础。

定义 6 设 $I^{\succ}=(U,A,F)$ 为序信息系统，且 $a\in A$，称 $E(A/\{a\})$-$E(A)$ 为 a 在 A 中的绝对重要度，记作：$DS(a,A)$，即：

$$DS(a,A)=E(A\backslash\{a\})-E(A)$$

特别地，当 $A=\{a\}$，用 $DS(a)$ 表示 $DS(a,\{a\})$ 且有：

$$DS(a)=E(\Phi)-E(a)=DS(a)\ =E(\Phi)\ -E(a)=$$
$$|U|\mathrm{lb}|U|-E(\{a\})$$

由上述定义易得如下性质：

性质 1 $0\leqslant DS(a,A)\leqslant|U|\mathrm{lb}|U|$。

性质 2 属性 a 在 A 中是必要的，当且仅当 $DS(a,A)>0$。

性质 3 $Core(A)=\{a\in A\,|\,DS(a,A)>0\}$。

定义 7 设 $I^{\succ}=(U,A,F)$ 为序信息系统，$C\subseteq A$，对于任意 $a\in A\backslash C$，称 $E(C)-E(C\cup\{a\})$ 为 a 相对于 C 的相对重要度，记作：$DR(a,C)$，即 $DR(a,C)=E(C)-E(C\cup\{a\})$。

上述定义表明，属性 $a\in A\backslash C$ 关于属性集 C 的重要性 $DR(a,C)$ 的值越大，属性 $a\in A\backslash C$ 关于属性集 C 就越重要。因此，可把 $DR(a,C)$ 作为寻找最小属性约简的启发式信息，以减少搜索空间。

根据定义 7，如果 $DR(a,C)>DR(b,C)$，则有 $E(C\cup\{a\})<E(C\cup\{b\})$，因此，在实际运算中，可以把属性重要性最大的、满足 $DR(a,C)=\max\limits_{a'\in A-C}\{DR(a',C)\}$ 的属性 a 并入核中，即把满足 $E(C\cup\{a\})=\min\limits_{a'\in A-C}\{E(C\cup\{a'\})\}$ 的属性 a 并入核中。

定理 4 设 $I^{\succ}=(U,A,F)$ 为序信息系统，$P\subseteq A$，若 $E(P)=E(A)$，且对任意的 $a\in P$ 有 $DR(a,P)>0$，则 P 为的一个约简。

证明：由定义 7 易证定理成立。

由定理 4 可以求出属性集的核 $Core(A)$，由于核唯一并且为任何约简的子集，因此核可以作为最小约简的起点。根据定义 7 中属性的重要性，逐次选择最重要的属性添加到核中，直到其粗糙熵等于 $E(A)$，即在 $Core(A)$ 的基础上通过增加属性构成的最小约简。

4 基于粗糙熵属性约简的启发式算法

4.1 启发式算法

算法步骤如下：

输入 序信息系统 $I^{\succ}=(U,A,F)$，其中，U 为论域；A 为属性集。

输出 该序信息系统的约简。

Step 1 计算序信息系统中的粗糙熵 $E(A)$。

Step 2 计算属性集的核，$a\in A-C$，

$$DR(a,Red(A))=\max\{DR(a,Red(A))\}$$

Step 3 $Core(A)\rightarrow Red(A)$。

Step 4 计算粗糙熵 $E(Red(A))$，若 $E(Red(A))=E(A)$ 成立，则转 Step7，否则转 Step5。

Step 5 计算所有的 $a\in A-Red(A)$ 的值 $DR(a,Red(A))$，取 a_1 满足 $DR(a,Red(A))=\max\{DR(a,Red(A))\}$。

Step 6 $Red(A)\cup\{a_1\}\rightarrow Red(A)$，转 Step4。

Step 7 输出最小约简 $Red(A)$。

4.2 实例分析

对于表 1 给出的序信息系统 $I^{\succ}=(U,A,F)$，计算步骤如下：

Step1 计算 $E(A)$：

$$E(A)=\frac{4}{6}\cdot\mathrm{lb}4+\frac{3}{6}\cdot\mathrm{lb}3+\frac{5}{6}\cdot\mathrm{lb}5+\frac{2}{6}\cdot\mathrm{lb}2+\frac{1}{6}\cdot\mathrm{lb}1+\frac{1}{6}\cdot\mathrm{lb}1=4.390\,49$$

Step2 计算：

$$DS(\{a_1\},A)=0$$

$$DS(\{a_2\},A)=2.044\,1$$

$$DR(\{a_3\},A)=2.442\,5$$

因此，$Core(A)=\{a_2,a_3\}$。

Step3 由于 $E(RED(A))=E(\{a_2,a_3\})=E(A)$，因此 $\{a_2,a_3\}$ 为序信息系统的最小约简。

5 结束语

要从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，就必须对系统进行知识约简。本文从研究序信息系统中知识与粗糙熵之间的关系出发，在此基础上建立了基于知识粗糙熵的启发式约简算法。在今后的研究中，将把该算法应用于不协调序目标信息系统，为复杂决策系统的规则提取提供便利途径。

参考文献

[1] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. Boston, USA: Kluwer Academic Publishers, 1991.

[2] Pawlak Z. Rough Sets[J]. Communication of the ACM, 1995, 38(1): 89-95.

[3] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.

[4] 苗夺谦, 王　珏. 基于粗糙集的多变量决策树构造方法[J]. 软件学报, 1997, 8(6): 425-431.

[5] 米据生, 吴伟志, 张文修. 不协调目标信息系统知识约简的比较研究[J]. 模糊系统与数学, 2003, 17(3): 54-60.

[6] 张文修, 米据生, 吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报, 2003, 26(1): 12-18.

[7] 张文修, 梁　怡, 吴伟志. 信息系统与知识发现[M]. 北京: 科学出版社, 2003.

[8] Greco S, Matarazzo B, Slowinski R. Rough Approximation of Preference Relation by Dominance Relations[J]. European Journal of Operational Research, 1999, 117(1): 63-68.

[9] Xu Weihua, Zhang Wenxiu. Methods for Knowledge Reduction in Inconsistent Ordered Information Systems[J]. Journal of Applied Mathematics & Computing, 2008, 26(1/2): 313-323.

[10] 徐伟华, 张文修. 基于优势关系信息系统分配约简的矩阵算法[J]. 计算机工程, 2007, 33(14): 4-7.

[11] 徐伟华, 张文修. 基于优势关系下不协调目标信息系统的知识约简[J]. 计算机科学, 2006, 33(2): 182-184.

[12] 徐伟华, 张文修. 基于优势关系下不协调目标信息系统的分布约简[J]. 模糊系统与数学, 2007, 21(4): 124-131.

[13] 张晓燕, 徐伟华. 序信息系统的知识粗糙熵与粗集粗糙熵[J]. 计算机工程与应用, 2007, 43(27): 62-65.

编辑　顾姣健