第 42 卷 第 11 期
Vol.42 No.11

山 东 大 学 学 报 (理 学 版)
Journal of Shandong University(Natural Science)

2007 年 11 月
Nov. 2007

文章编号:1671-9352(2007)11-0082-03

基于粗糙集理论偏序决策表知识获取方法研究

席慎思[1],洪晓光[1],孔磊[2],衣升起[1]

(1. 山东大学 计算机科学与技术学院,山东 济南 250061;2. 济宁市社会劳动保险处,山东 济宁 272100)

摘要:介绍了基于偏序关系的偏序决策表,研究了偏序决策表各条件分类和决策分类集合之间的关系,提出了从各分类中计算偏序决策表核及属性约简方法,通过实例,验证了这些方法的有效性。

关键词: 偏序关系;粗糙集;数据分析;核;知识约简

中图分类号:TP311 **文献标志码**:A

Knowledge acquisition from a partial order decision table based on rough sets theory

XI Shen-si[1], HONG Xiao-guang[1], KONG Lei[2], YI Sheng-qi[1]

(1. Department of Computer Science and Technology, Shandong University, Jinan 250061, Shandong, China;
2. Social Labor Insurance Department of Jining, Jining 272100, Shandong, China)

Abstract: A partial order decision table based on partial order relation was introduced. The relationship between conditional equivalence classes and decision equivalence classes was studied in a partial order decision table, and methods of core and attribute reduction were expatiated from these classes. Finally, an example was given to illustrate the efficiency of these methods.

Key words: partial order relation; rough sets; data analysis; core; knowledge reduction

0 引言

粗集理论是由波兰科学家 Pawlak. Z[1] 于 1991 年提出的研究不完整数据、不精确知识的表达、学习、归纳方法。这一理论从新的视角出发对知识进行了定义,它把知识看作是关于论域的划分,并引入代数学中的等价关系来讨论知识,它为智能信息处理提供了有效的处理技术,目前已经在人工智能、知识获取、模式识别、分类等方面得到了成功的应用。20 世纪 90 年代至今无论在 Rough 集理论体系完善还是实际应用方面都有了很大的发展, Yao. Y. Y 等[2]在经典 Rough 集理论的等价关系模型基础上又提出了一些非等价关系模型。知识约简是在保持知识库分类能力不变的前提下,删除不相关或者不重要的知识,使得在大量的数据信息中能够挖掘出简洁的、有价值的模式来辅助决策。在粗集理论中,知识约简和规则提取算法[3]是 NP 问题。

本文在决策表中按每个属性值排序对象排序,并挖掘整体排序的规则。为此,在决策表上引进了偏序关系(自反性、反对称性、传递性)得到偏序关系表,并在此基础上进行数据分析,决策规则简化,对各等价类集合之间的关系进行了研究,提出了新的核及属性约简计算算法。

1 相关基本概念

定义 1.1 一个决策表信息系统 $S = <U, R, V, F>$,其中 U 是对象的集合,也称为论域,$R = C \cup D$是属性集合,$C \cap D = \emptyset$, $D \neq \emptyset$,子集 C 和 D 分别称为条件属性集和决策属性集,V 是属性值的集合,$f: U \times R \to V$ 是一个信息函数,它指定了 U 中

收稿日期:2007-04-30
基金项目:国家自然科学基金资助项目(60673130);教育部科学技术研究重点资助项目(03102);省重大科技专项资助项目(2004GG4201022);山东省自然基金资助项目(Y2004G07, Y2006G29);山东省中青年科学家奖励基金资助项目(2005BS01002); 山东省科技攻关计划资助项目(2005GG3201088);山东省科学技术发展计划国际合作资助项目(2006GG2201052)
作者简介:席慎思(1982-),男,硕士研究生,数据库方向. Email:ssrunner@163.com

每个对象 x 的属性值。

定义 1.2[4] $\mathrm{Par}T(T,\{\leqslant_a \mid a\in C\cup D\})$是一个偏序的决策表。其中 T 为决策表，$\leqslant_a$ 是 V_a 上的一个偏序关系。

对象某属性 a 的值上的排序可得出对象之间的排序，$x,y\in U$：$I_a(x)\leqslant_a,\ I_a(y)\Leftrightarrow x\leqslant_a y$

其中$\leqslant_a$ 表示由属性 a 引出的 U 中对象之间的偏序关系。在属性 a 上，对象 x 排在对象 y 之前当且仅当在该属性上，x 的值排在 y 的值之前。

一般地 $B\subseteq C\subseteq A$，$\forall a\in B$：$I_a(x)\leqslant_a$，$I_a(y)\Leftrightarrow x\leqslant_a y$也就是说，在 B 中所有属性 x 的值都排在 y 的值之前，此时认为 x 排在 y 之前。

特别，当 $B=C$ 且 $B=\mathrm{POS}_c(D)$时，$\forall a\in B$，$I_a(x)\leqslant_a,\ I_a(y)\Leftrightarrow x\leqslant_a y$，称为偏序决策表 $\mathrm{Par}T$ 中对象的整体排序。

定义 1.3 给定偏序决策表 $\mathrm{Par}T[X]_c=U\{y\mid x\leqslant_c y,x,y\in U\}$称为偏序关系下 x 的等价类。

定义 1.4 $B\subseteq C$，$X\subset U$，$\mathrm{IND}(B)=\{<x,y>\in U\times U\mid \forall a\in B,a(x)=a(y)\}$为 B 上的不分明关系，则 $\mathrm{IND}_B^-(x)=\{x\mid [x]_B\cap X\neq\varnothing,x\in U\}$，$\mathrm{IND}_{B-}(x)=\{x\mid [x]_B\subseteq X,x\in U\}$分别称为在偏序关系下，集合 X 的上，下近似。

定义 1.5 属性子集 B 是相关的，$\exists b\in B$，对 $\forall x\in U,[x]_B=[x]_{B-\{b\}}$，否则称 B 是独立的。属性子集 $B\subseteq C$ 称为 C 的约简，如果 B 是独立且对 $\forall x\in U$，$[x]_B=[x]_C$ 所有约简的交集称为 C 的核。

2 数据分析[5-7]

本文对一个实例进行分析，某公司员工信息表见表1，表1中 E 表示教育程度，A 表示年龄，S 表示薪水，D 表示公司是否继续雇佣的意愿。$U=\{$甲,乙,丙,丁,戊,己$\}$，$A=C\cup D$，$C=\{E,A,S\}$，$D=\{D\}$。

表1 员工信息表

Table 1 Employee information table

员工	属性			
	E	A	S	D
甲	高	小	低	有意
乙	低	大	适当	无意
丙	中	中等	适当	无意
丁	中	中等	低	有意
戊	高	小	适当	有意
己	低	大	低	无意

(1) 属性排序

$\leqslant_E$：低$\leqslant_E$ 中$\leqslant_E$ 高；$\leqslant_A$：大$\leqslant_A$ 中等$\leqslant_A$ 小$\leqslant_S$：适当$\leqslant_S$ 低；$\leqslant_D$：无意$\leqslant_D$ 有意。

(2) 按属性，对员工排序

$\leqslant_{(E)}$：{乙，己} $\leqslant_{(E)}$ {丙，丁} $\leqslant_{(E)}$ {甲，戊}；$\leqslant_{(A)}$：{乙，己} $\leqslant_{(A)}$ {丙，丁} $\leqslant_{(A)}$ {甲，戊}；$\leqslant_{(S)}$：{乙，丙，戊} $\leqslant_{(S)}$ {甲，丁，己}；$\leqslant_{(D)}$：{乙，丙，己} $\leqslant_{(D)}$ {甲，丁，戊}。

(3) 对于条件属性集$\{E,A,S\}$对员工进行排序记 $C=\{E,A,S\}$。

$\leqslant_{(E,A,S)}$：甲 $\leqslant_{(C)}$ 甲，乙 $\leqslant_{(C)}$ 乙，丙 $\leqslant_{(C)}$ 丙，丁$\leqslant_{(C)}$丁，戊$\leqslant_{(C)}$戊，己$\leqslant_{(C)}$己。

乙$\leqslant_{(C)}$甲，乙$\leqslant_{(C)}$丙，乙$\leqslant_{(C)}$丁，乙$\leqslant_{(C)}$戊

丙$\leqslant_{(C)}$甲，丙$\leqslant_{(C)}$丁，丙$\leqslant_{(C)}$戊。

丁$\leqslant_{(C)}$ 甲，戊$\leqslant_{(C)}$ 甲，己 $\leqslant_{(C)}$ 丙，己 $\leqslant_{(C)}$ 戊，己$\leqslant_{(C)}$甲。

就 $\leqslant_{\{E,A,S\}}$ 的偏序关系，本文给出哈希图（见图1）。

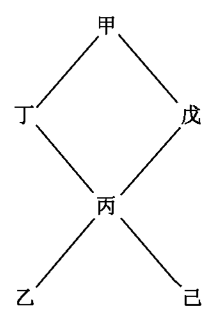

图1 哈希图

Fig.1 Hash map

(4) 令 $B=\{E,A\}\subseteq A$

$\mathrm{IND}(B)=\{<$甲,戊$>$，$<$乙,己$>$，$<$丙,丁$>\}$，取 $X=\{4,5\}\subseteq U$，$\mathrm{IND}_B^-(X)=\varnothing$，$\mathrm{IND}_B^-=\{$甲，丙，丁，戊$\}$。

(5) $C=\{E,A,S\}$，$U/C=\{<$甲,甲$><$乙,乙$><$丙,丙$><$丁,丁$><$戊,戊$><$己,己$>\}$

$B_1=\{E,A\}$，$U/B=\{<$甲,戊$><$乙,己$><$丙,丁$>\}\neq U/C$，所以 B 不是 C 的约简。

$B_2=\{E,S\}$，

$U/B_2=\{<$甲,甲$><$乙,乙$><$丙,丙$><$丁,丁$><$戊,戊$><$己,己$>\}=U/C$，另可验证 B_2 是独立的，所以 B_2 是 C 的约简。

$B_3=\{A,S\}$，

$U/B_3=\{<$甲,甲$><$乙,乙$><$丙,丙$><$丁,丁$><$戊,戊$><$己,己$>\}=U/C$，另可验证 B_3 是独立的，所以 B_3 是 C 的约简。

$B_2 \cap B_3 = \{S\}$,则$\{S\}$为 C 的核。

3 偏序决策表约简及核生成方法

设决策系统 S 条件属性 $A = \{a_1, a_2, \cdots a_n\}$,决策属性 $D = \{d\}$,$U/a_l = \{X_1, X_2, \cdots X_m\}(l = 1,2,\cdots n)$,$U/d = \{D_1, D_2, D_r\}$。

3.1 S 的核计算

设 $E_i = \{\{x, y\} \mid ([x]_{a_i} = [y]_{a_i}) \wedge ([x]_d = [y]_d = \emptyset), x, y \in U, i = 1,2,\cdots n\}$,很明显,$E_i$ 由属性 a_i 在 U 上导出的划分$\{X_1, X_2, \cdots X_m\}$,$|X_i| \geqslant 2$ 的等价类所生成,记 $E = \{E_i, E_2, \cdots E_n\}$。

命题 1　如果存在不分明关系 $\mathrm{IND}(a_k)$,使$[x]_{a_k} \neq [y]_{a_k}$,而在 $A \setminus a_k$ 属性 a_i 上,有$[x]_{a_i} = [y]_{a_i}(i = 1,2,\cdots n, i \neq k)$成立,则条件属性 a_k 是不可约的。

设$\{x, y\} \in E_i(i = 1,2,\cdots r)$则有以下推论成立:

推论　若 $r = n - 1$,则$\{x, y\} \notin E_i$所对应的属性 a_i 为不可约。

有了以上讨论,下面说明生成核的算法的步骤:

算法 1　决策表 S 的核计算

输入:决策表 S 由各条件属性、决策属性在论域 U 导出的划分,$r: = 1$。

输出:决策表 S 的核 $\mathrm{CORE}(S)$。

(1) 计算 E_i 并生成集合 $E = \{E_1, E_2, \cdots E_n\}$;

(2) 如果 $E \neq \emptyset$,则 $\mathrm{CORE}(S) = \emptyset$,转(7);

(3) 取$\{x, y\} \in E_i$,$E_i \leftarrow E_i^- \{x, y\}$,重复:

如果$\{x, y\} \in E_j(E_j \in E \setminus E_i)$,则 $r: = r + 1$,$E_j \leftarrow E_j - \{x, y\}$,$F \leftarrow F \cup \{a_j\}$,($a_j$ 为 E_j 对应属性)

(4) 如果 $r: = n - 1$,则 $\mathrm{CORE}(S) \leftarrow \mathrm{CORE}(S) \cup \{A \setminus F\}$;

(5) 如果 $E_i \neq \emptyset$,$r: = 1$,转(3);

(6) 如果 $E_i \neq \emptyset(E_j \in E \setminus E_i)$,$E_i \leftarrow E_j$,$r: = 1$,转(3);

(7) 结束。

3.2 决策表的约简

设 $G = \{\{x, y\} \mid \{[x]_{\mathrm{core}(s)} = [y]_{\mathrm{core}(s)} \wedge ([x]_d \cap [y]_d = \emptyset)\}, x, y \in U\}$,对任一$\{x, y\} \in G$,必存在一条件属性集合:$T = \{a_i \mid [x]_{a_i} = [y]_{a_i} = \emptyset, a_i \in A \setminus \mathrm{CORE}(S)\}$,设 $B_1, B_2, \cdots B_t$ 分别为 G 中每$\{x, y\}$对满足 T 的集合,记 $C = \{B_1, B_2, \cdots B_t\}$,并设 $B = \wedge_{B_j \in C}(\vee_{a_j \in B_j} a_i)(j = 1,2,\cdots t, a_i \in A \setminus \mathrm{CORE}(S))$,有以下命题成立:

命题 2　$B \wedge \mathrm{CORE}(S)$的值,是决策表 S 的约简。

算法 2　决策表 S 的约简

输入:决策表 S 由各条件属性、决策属性在论域 U 导出的划分及 $\mathrm{CORE}(S)$。

输出:决策表 S 的约简 $\mathrm{REDU}(S)$。

(1) 计算 G,如果 $G = \emptyset$,转(8);

(2) 取 G 元素$\{x, y\}$,$G \leftarrow G \setminus \{x, y\}$;

(3) 对 $A \setminus \mathrm{CORE}(S)$的每一条属性 a,重复;
如果$[x]_{a_i} \cap [y]_{a_i} = \emptyset$,则 $W \leftarrow W \cup \{a_i\}$;

(4) $W \leftarrow W$ 中属性作"$\vee$"运算;

(5) $B \leftarrow B \wedge W$;

(6) 如果 $G \neq \emptyset$,则 $W = \emptyset$,转(3);

(7) $\mathrm{REDU}(S) = \mathrm{CORE}(S) \wedge B$,转(9);

(8) $\mathrm{REDU}(S) = \mathrm{CORE}(S)$;

(9) 结束。

例:考虑表 1 所示的偏序关系,为简单起见,数字化表 1:

属性 A:低-1,中-2,高-3; 属性 E:大-1,中等-2,小-3;属性 S:低-1,适当-2; 属性 D:无意-1,有意-2。

条件属性 A, E, S,决策属性 D 对论域 U 的划分如下:

$U \setminus E = \{\{1,5\}\{2,6\}\{3,4\}\}$, $U \setminus A = \{\{1,5\}\{2,6\}\{3,4\}\}$, $U \setminus S = \{\{1,4,6\}\{2,3,5\}\}$, $U \setminus D = \{\{1,4,5\}\{2,3,6\}\}$

(1) 计算决策表 S 的核

首先由以上划分计算 E:$E_1 = \{\{3,4\}\}$,$E_2 = \{\{3,4\}\}$,$E_3 = \{\{1,6\},\{4,6\},\{2,5\},\{3,5\}\}$。

在集合 E 中,$\{3,4\}$属于 E_1, E_2,由算法 1 可知,E_3 对应的属性 S 为核属性,即 $\mathrm{CORE}(S) = \{S\}$。

(2) 计算决策表 S 的约简

由 $U \setminus S = \{\{1,4,6\}\{2,3,5\}\}$得:$G = \{\{1,6\},\{4,6\},\{2,5\},\{3,5\}\}$,$B_1 = \{E, A\}$,$B_2 = \{E, A\}$,$B_3 = \{E, A\}$,$B_4 = \{E, A\}$,于是 $C = \{B_1, B_2, B_3\}$,由算法 2,计算得 $B = E \vee A$。

同时计算得到决策表 S 的两个约简:

$\mathrm{REDU}(S)_1 = \{E, S\}$,$\mathrm{REDU}(S)_2 = \{A, S\}$。

4 结语

本文对决策系统引进了属性 (下转第 88 页)

(上接第 84 页) 偏序、对象偏序的概念,并给出了偏序一致决策表数值分析、决策规则约简等概念,详细阐述了核及属性约简计算算法,对不一致情况没有讨论,由各属性排序推导整体排序的关系是我们下一步工作的重点。

参考文献:

[1] PAWALK Z. Rough Sets-theoretical aspects of reasoning about data[M]. Dordrecht: Klumer Acadimic Publishers, 1991.
[2] YAO Y Y, ZHONG N. Potential applications of granular computing in knowledge discovery and data mining[C]// Proceedings of World Multiconference on Systemics, Cybernetics and Informatics[S.l.]. Computer Science and Engineering, 1999: 573-580.
[3] Skowron A. Rough sets in KDD[R]. Beijing: Special Invited Speaking, WCC 2000 in Beijing, 2000.
[4] 马垣.非经典关系数据库理论[M].北京:清华大学出版社,2005:177-202.
[5] 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.
[6] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社, 2001.
[7] 刘清. Rough 集及 Rough 推理[M].北京:科学出版社, 2001.

(编辑:孙培芹)