第 35卷 第 3期 昆明理工大学学报 (理工版) http://www.kustjournal.com/ Vol 35 No.3
2010年 6月 Journal of Kunming University of Science and Technology (Science and Technology) Jun 2010

doi: 10.3969/j.issn.1007-855x.2010.03.016

不确定系统的信息熵与知识化简

曾黄麟[1], 许立志[1], 梁 宁[2]

(1.四川理工学院 自动化与电子信息学院, 四川 自贡 643000; 2.四川理工学院 机械工程学院, 四川 自贡 643000)

摘要: 根据一个不确定系统的条件属性对决策结果产生的依赖关系, 建立了基于信息熵和知识重要性对不确定系统的知识冗余分析的联系, 得到了基于信息熵和基于 Rough集理论对不确定系统的知识冗余分析概念的统一和方法的一致性重要结论, 从信息熵的角度以统计概率的方法提出了不确定系统的知识冗余分析新的知识简化方法, 应用实例验证了本文理论和方法的可行性和有效性.

关键词: 不确定系统; Rough集; 信息熵; 知识约简

中图分类号: TP18 文献标识码: A 文章编号: 1007-855X(2010)03-0083-06

Analysis on Information Entropy and Knowledge Simplification of an Uncertain System

ZENG Huang-lin[1], XV Li-zhi[1], LIANG Ning[2]

(1. Institute of Automation and Electronic Information, Sichuan University of Science and Engineering, Zigong Sichuan 643000, China;
2. Institute of Machinery Engineering, Sichuan University of Science and Engineering, Zigong Sichuan 643000, China)

Abstract: The association of information entropy and knowledge significance with knowledge redundancy in an uncertain system is set up in this paper due to the decision dependence on condition attributes. It is concluded that the analyses based on information entropy and on rough sets are consistent in the concepts and methods. An approach to analyzing knowledge redundancy and knowledge simplification of an uncertain system is presented on information entropy. The theoretical analysis and the proposed method are validated by some examples.

Key words: uncertain system; rough sets; information entropy; knowledge simplification

0 引 言

不确定系统是一种随机事件系统, 不确定系统的信息处理常常从信息熵的角度以统计概率的方法分析, 但对于不确定系统中的知识冗余和知识简化往往很少从信息熵的角度讨论, 大多数研究者都是从 Rough集理论的角度进行研究, 通过知识冗余分析进行知识简化[1-5]. 对于信息的不确定性, 信息熵和 Rough集理论概念之间是否存在某些联系? 对于不确定系统中的知识冗余和知识简化, 能否建立两种方法之间的关系与一致性? 这些问题是智能信息处理研究领域十分感兴趣的问题, 但目前尚未得到明确的答案[8-12].

本文将从信息论中信息熵的角度出发, 对于信息的不确定性, 通过讨论不确定系统的条件属性对决策结果产生的依赖关系, 建立信息熵和 Rough集理论中知识的重要性之间的联系, 研究基于信息熵和基于 Rough集理论对不确定系统的知识冗余分析方法的一致性, 讨论基于信息熵和基于 Rough集理论对于不确定系统的知识冗余和知识简化方法的统一, 从信息熵的角度以统计概率的分析方法, 提出对于不确定系统知识冗余分析新的知识简化方法.

1 基本理论

一个不确定系统常表达为一个采集的数据信息系统 $S=(U, A, V, f)$, 其中: U是非空有限论域, 任意的

收稿日期: 2009-02-23. 基金项目: 四川省生物技术白酒重点实验室课题 (NJ2008-5); 四川省教育厅重大培育项目 (09ZX002).

作者简介: 曾黄麟 (1955-), 博士, 教授, 电子科技大学博士生导师. 主要研究方向: 人工神经网络, 粗集理论, 模糊逻辑, 智能信息处理, 模式识别等. E-mail: zhl@suse.edu.cn

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

对象 $x \in U$, $A = C \cup D$ 是属性集, C 和 D 分别是条件属性集和决策属性. 若一个不确定系统采集的数据中条件属性为 R_i, $i = 1, 2, \cdots, n$, $V = \bigcup_{R_i \in A} (V_{R_i})$ 是属性值的集合, 对于任意属性 $R_i \in A$, 存在 $f: U \times A \rightarrow V$ 是一个信息函数.

定义 1 当不可分辨关系是物种由属性集 C 表达时, 属性集 $R \subset C$, 任意的对象 $x, y \in U$, 当且仅当 $f(x, R_i) = f(y, R_i)$ 时, $R_i \in R$, x, y 是不可分辨的. 可表达为:

$$ind(R) = \{(x, y) \subset U: \text{所有的 } R_i \in R, f(x, R_i) = f(y, R_i)\} \tag{1}$$

且有

$$[x]_{ind(R)} = \bigcap [x]_C, R \subset C \tag{2}$$

定义 2[6-7] 设子集 $X, Y \subset U$, 若根据决策属性 D, X 和 Y 不可分辨时, 称其为 $ind(D)$, 表示为 $U \mid ind(D)$. 根据条件属性 C 划分的等价类 X_j, U 中可能归入基于决策属性 D 划分的等价类 Y 的元素的集合, 表示为 D 的上近似集, 定义为

$$P^{-}(D) = \bigcup \{X_i \subseteq U \mid ind(C): X_j \cap Y \neq 0\} \tag{3}$$

根据条件属性 C 划分的等价类 X_j, U 中一定能归入基于决策属性 D 划分的等价类 Y 的元素的集合, 表示为 D 的下近似集, 定义为

$$P_{-}(D) = \bigcup \{X_j \in U \mid ind(C): X_j \subseteq Y\} \tag{4}$$

对于 $S = (U, A, V, f)$ 是一个决策信息系统, 定义 D 的 C 正域, 表示为 $POS_C(D)$, 即:

$$POS_C(D) = \bigcup P_{-}(D) \tag{5}$$

系统的知识可以用决策属性 D 表达, 也可以由条件属性 C 描述, 根据数据库中的函数之间的依赖性关系, 知识的重要性定义为:

$$\alpha_R = \frac{|U| - |R^{-}(X) - R_{-}(X)|}{|U|} \tag{6}$$

α_R 表达了一种知识的重要性, 例如它可以表达利用系统参数 R 描述 X 中对象的隶属度情况, 它可以代表该特征 (或特征集) 对于分类的有效程度.

定义 3 设 $S = (U, A, V, f)$ 是一个不确定系统, 若属性 R_i 存在 n 种取值状态, 假定一个可能事件集合, 其事件出现的概率则为 $p(X_1), p(X_2), \cdots, p(X_n)$, 利用 Shannon提出的熵的概念作为不确定信息的统计测度, 则定义属性 R_i 的熵为:

$$H(R_i) = -\sum_{i=1}^{n} p(X_i)\log(p(X_i)) \tag{7}$$

定义 4 设 $S = (U, A, V, f)$ 是一个不确定系统, 系统决策结果被划分为 m $(m > 1)$ 个状态 (类别), 若属性 R_i 在它的第 i 个状态时的概率为 $p(x_i)$, 把第 j 类样品点落到第 i 段的频率记作 $p(Y_j / X_i)$, 基于属性 R_i 的系统熵的期望值, 则条件属性 R_i 对决策系统产生的不确定信息的统计测度, 则条件熵定义为:

$$H(D/R_i) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j / X_i)\log(p(Y_j / X_i)) \tag{8}$$

2 不确定信息熵和知识的重要性

信息熵从统计测度的角度研究了不确定信息中条件属性对决策结果产生的依赖关系, Rough集理论从不可分辨的角度研究了不确定信息中条件属性对决策结果产生的依赖关系, 对于信息的不确定性, 迄今尚未建立信息熵和 Rough集理论中知识的重要性之间的联系, 尚未见讨论对于不确定系统的信息熵与 Rough集理论分析方法两种方法之间的一致性问题. 因此, 我们将从条件属性对决策结果产生的依赖关系去研究信息的不确定性和两种分析方法之间的关系与一致性.

命题 1 对于一个不可分辨关系的决策信息系统 $S = (U, A, V, f)$, 设 U 是一个论域, P, Q 是 U 上的两个属性集合, 且 $P \subseteq Q$, 存在下列结论:

若 $ind(Q) = ind(P)$, 则 $H(Q) = H(P)$ (9)

证明 设 P, Q 在 U 上导出划分的等价类分别为 X, Y $(X = \{X_1, X_2, \cdots, X_N\}, Y = \{Y_1, Y_2, \cdots, Y_M\})$, 则 P, Q 在 U 的子集组成的代数上的概率分布为:

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$[X:p]=\begin{bmatrix} X_1 & X_2 & \cdots & X_N \\ p(X_1) & p(X_2) & \cdots & p(X_N) \end{bmatrix},$$

$$[Y:p]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_M \\ p(Y_1) & p(Y_2) & \cdots & p(Y_M) \end{bmatrix},$$

其中: $p(X_i)=\frac{|X_i|}{|U|}$, $i=1,2,\cdots,N$; $p(Y_j)=\frac{|Y_j|}{|U|}$, $j=1,2,\cdots,M$.

根据式(7)有:

$H(P)=-\sum_{i=1}^{N}p(X_i)\log(p(X_i))$, $H(Q)=-\sum_{i=1}^{M}p(Y_j)\log(p(Y_j))$, 若有 $ind(Q)=ind(P)$, 则对于任意的 $X_k(X_k\in\{X_1,X_2,\cdots,X_N\})$ 存在唯一的 $Y_l(Y_l\in\{Y_1,Y_2,\cdots,Y_M\})$ 与之对应. 其中 $k\in[1,N]$, $l\in[1,M]$.

即: $X_k(X_k\in\{X_1,X_2,\cdots,X_N\})$ 与 $Y_l(Y_l\in\{Y_1,Y_2,\cdots,Y_M\})$ 存在一一对应关系.

因而有: $p(X_k)=\frac{|X_k|}{|U|}=p(Y_l)=\frac{|Y_l|}{|U|}$;

所以 $-\sum_{i=1}^{N}p(X_i)\log(p(X_i))=-\sum_{i=1}^{M}p(Y_j)\log(p(Y_j))$;

即 $H(P)=H(Q)$.

命题 2 对于一个以等概率分类出现的决策信息系统 $S=(U,A,V,f)$, 在属于边界 $R_i^-(W_k)-R_{i\_}(W_k)$ 的特征 R_i 的每一等价类中的对象相等地分布在 m 个分类中时, 对于任意分类 W_k, 特征 R_i 减小系统的信息熵的不确定性贡献为:

$$\beta_{R_i}=H-H(R_i)=\frac{|U|+|R_i^-(W_k)-R_{i\_}(W_k)|}{|U|}=a_{R_i} \tag{10}$$

证明 一个以等概率分类出现的决策信息系统, $p(W_k)=\frac{1}{m}$, $k=1,2,\cdots,m$. 则这时系统的初始熵为:

$$H=-\sum_{k=1}^{m}p(W_k)\log_m p(W_k)=-\sum_{k=1}^{m}\frac{1}{m}\log_m\left(\frac{1}{m}\right)=1$$

令 F 代表特征的等价类族, 我们把 F 划分成两个不相联的子集, 即:

$F=\{R_1,R_2,\cdots,R_n\}\cup\{s_1,s_2,\cdots,s_k\}$

这里每一个 R_i 仅含有相同分类的对象. 根据特征 R_j 对系统产生的不确定信息的统计测度的定义 $H(R_i)=-\sum_{j=1}^{L}\sum_{k=1}^{m}p(b_j)p(W_k\mid b_j)\log_m p(W_k\mid b_j)$, 可得 R_i 对系统产生的不确定信息的统计测度 $H(R_i)=0$, $1\leqslant j\leqslant L$.

由条件假定第 i 个分类出现在 s_j 中的概率为 $=\frac{1}{m}$, $i=1,2,\cdots,m$, 根据特征对系统产生的不确定信息的统计测度的定义, 有 s_j 对系统产生的不确定信息的统计测度 $H(s_j)=1$, $1\leqslant j\leqslant k$. 则

$$H(R_i)=\sum_{i=1}^{n}\frac{|R_i|H(R_i)}{|U|}+\sum_{j=1}^{k}\frac{|s_j|H(s_j)}{|U|}=\sum_{j=1}^{k}\frac{|s_j|}{|U|} \tag{11}$$

根据粗集分类的概念, 因此集合 $\bigcup_{j=1}^{k}s_j$ 与 $BN_{R_i}(W_k)=-R_i^-(W_k)-R_{i\_}(W_k)$ 相同, 则

$$\sum_{j=1}^{k}|s_j|=\sum_{j=1}^{k}|R_i^-(W_k)-R_{i\_}(W_k)|,\quad 1\leqslant k\leqslant m \tag{12}$$

故 $\beta_{R_i}=H-H(R_i)=\frac{|U|-|R_i^-(W_k)-R_{i\_}(W_k)|}{|U|}=a_{R_i}$

结论(9), (10)和(11)揭示了对于不确定系统的信息熵与 Rough集理论分析方法的统一, 建立了两种方法之间的关系与一致性.

推论 1 $S=(U,A,V,f)$是一个决策信息系统, 条件属性 R_i 减小系统的信息熵的不确定性的 $H(R_i)$ 越大, 对影响系统决策的确定信息越小, 则条件属性 R_i 分类能力越弱; 反之亦然.

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

推论的结论是显然的,证明略.

3 基于信息熵和 $Rough$ 集理论新的冗余知识简化方法

设 $S=(U,A,V,f)$ 是一个不确定系统, $A=C\cup D$ 是属性集, C 和 D 分别是条件属性集和决策属性集, $R_i\in C$, 根据粗集理论[1-2], 如果存在

$$POS_C(D)=POS_{c-R_i}(D) \tag{13}$$

称 R_i 为 C 中相对于 D 可省略的,否则 R_i 为 C 中相对于 D 不可省略的.

命题 1 设 U 是一个论域, C,D 是 U 上的两个属性集合, D 是决策属性集. 存在

$$H(D\mid C)=H(D\cup C)-H(C) \tag{14}$$

证明 设 C,D 在 U 上导出划分的等价类分别为 X,Y $(X=\{X_1,X_2,\cdots,X_N\},Y=\{Y_1,Y_2,\cdots,Y_M\})$, 则 C,D 在 U 的子集组成的代数上的概率分布为:

$$[X:p]=\begin{bmatrix} X_1 & X_2 & \cdots & X_N \\ p(X_1) & p(X_2) & \cdots & p(X_N)\end{bmatrix},$$

$$[Y:p]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_M \\ p(Y_1) & p(Y_2) & \cdots & p(Y_M)\end{bmatrix},$$

其中: $p(X_i)=\dfrac{|X_i|}{|U|}$, $i=1,2,\cdots,N$; $p(Y_j)=\dfrac{|Y_j|}{|U|}$, $j=1,2,\cdots,M$; $p(\dfrac{Y_j}{X_i})=\dfrac{|Y_j\cap X_i|}{|X_i|}$.

根据式(7)和式(8)有:

$$H(C)=-\sum_{i=1}^{N}p(X_i)\log(p(X_i)),$$

$$H(D)=-\sum_{j=1}^{M}p(Y_j)\log(p(Y_j)),$$

$$H(\frac{D}{C})=-\sum_{i=1}^{n}p(X_i)\sum_{i=1}^{m}p(\frac{Y_j}{X_i})\log(p(\frac{Y_j}{X_i}))$$

令 $A=C\cup D$, 则 A 在 U 上导出的划分的等价类为 Z, 即对 X(或对 Y)的细分.

$$Z_k=\begin{cases} Y_j\cap X_i, & Y_j\cap X_i\neq\phi \\ \text{忽略}, & Y_j\cap X_i=\phi\end{cases},\quad k=1,2,\cdots,t;$$

则 $p(Z_k)=\dfrac{|Y_j\cap X_i|}{|U|}$.

根据式(7)有

$$H(A)=H(C\cup D)=-\sum_{i=1}^{t}p(Z_k)\log(p(Z_k))=-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|Y_j\cap X_i|}{|U|}\log(\frac{|Y_j\cap X_i|}{|U|})$$

因为 $H(C)=-\sum_{i=1}^{N}p(X_i)\log(p(X_i))=-\sum_{i=1}^{N}\dfrac{|X_i|}{|U|}\log(\dfrac{|X_i|}{|U|})$

所以

$$\begin{aligned} H(A)-H(C)&=H(D\cup C)-H(C)\\ &=-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|Y_j\cap X_i|}{|U|}\log(\frac{|Y_j\cap X_i|}{|U|})-\sum_{i=1}^{N}p(X_i)\log(p(X_i))\\ &=-\sum_{i=1}^{N}p(X_i)\sum_{j=1}^{M}\frac{|Y_j\cap X_i|}{|X_i|}\log(\frac{|Y_j\cap X_i|}{|U|})-\sum_{i=1}^{N}p(X_i)\log(\frac{|X_i|}{|U|})\\ &=-\sum_{i=1}^{N}p(X_i)\{[\sum_{j=1}^{M}\frac{|Y_j\cap X_i|}{|X_i|}\log(\frac{|Y_j\cap X_i|}{|U|})]-\log(\frac{|X_i|}{|U|})\}\end{aligned}$$

其中: $\log(\dfrac{|X_i|}{|U|})=\log(\dfrac{|X_i|}{|U|})\cdot\dfrac{\sum_{j=1}^{M}|Y_j\cap X_i|}{|X_i|}$,

$H(A)-H(C)=H(D\cup C)-H(C)$

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$= - \sum_{i=1}^{N} p(X_i) \{ [\sum_{j=1}^{M} \frac{|Y_j \cap X_i|}{|X_i|} \log(\frac{|Y_j \cap X_i|}{|U|})] - \log(\frac{|X_i|}{|U|}) \cdot \sum_{j=1}^{M} \frac{|Y_j \cap X_i|}{|X_i|} \}$$

$$= - \sum_{i=1}^{N} p(X_i) \sum_{j=1}^{M} \frac{|Y_j \cap X_i|}{|X_i|} \log(\frac{|Y_j \cap X_i|}{|X_i|})$$

$$= - \sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(\frac{Y_j}{X_i}) \log(p(\frac{Y_j}{X_i}))$$

$$= H(\frac{D}{C})$$

证明完毕.

命题 2 设 $S = (U, A, V, f)$ 是一个不确定系统, $A = C \cup D$ 是属性集, C 和 D 分别是条件属性集和决策属性集, $R_i \in C$, 如果存在

$$H(D \mid C) = H(D \mid C - R_i) \tag{15}$$

称 R_i 为 C 中相对于 D 可省略的, 否则 R_i 为 C 中相对于 D 不可省略的.

证明 首先令 $U \mid ind(C) = \{X_1, X_2, \cdots, X_n\}$, $U \mid ind(D) = \{Y_1, Y_2, \cdots, Y_m\}$.

因为论域 U 是在 C 上相对于 D 一致的, 即 $POSc(D) = U$. 所以 $U \mid ind(C)$ 是 $U \mid ind(D)$ 的细分, 有 $U \mid ind(C + D) = U \mid ind(C) = \{X_1, X_2, \cdots, X_n\}$,

$$H(\frac{D}{C}) = - \sum_{i=1}^{k} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log(p(Y_j \mid X_i)) = 0$$

必要性: 假设属性 R_i 是 C 中相对决策属性 D 不必要的, 则 $POS_{C-R_i}(D) = POS_C(D) = U$, 所以 $U \mid ind(C - R_i)$ 是 $U \mid ind(D)$ 的细分.

令 $U \mid ind(C - R_i + D) = U \mid ind(C - R_i) = \{Z_1, Z_2, \cdots, Z_K\}$, 则

$$H(D \mid C - R_i) = - \sum_{i=1}^{k} p(Z_i) \sum_{j=1}^{m} p(Y_j \mid Z_i) \log(p(Y_j \mid Z_i)) = 0,$$

故 $H(D \mid C) = H(D \mid C - R_i)$.

充分性: 假设 $POS_{C-R_i}(D) \neq U = POS_C(D)$

令 $U \mid ind(C - R_i) = \{Z_1, Z_2, \cdots, Z_k\}$, 则至少存在 Z_i, $Z_i \in U \mid ind(C - R_i)$, 且 Y_{i1}, $Y_{i1} \in U \mid ind(D)$ 和 Y_{i2}, $Y_{i2} \in U \mid ind(D)$, $Y_{i1} \neq Y_{i2}$, 使得 $Z_i \cap Y_{i1} \neq 0$, 且 $Z_i \cap Y_{i2} \neq 0$,

因此 $H(\frac{D}{C} - R_i) = - \sum_{i=1}^{k} p(Z_i) \sum_{j=1}^{m} p(Y_j \mid Z_i) \log(p(Y_j \mid Z_i)) \succ 0$

这与 $H(\frac{D}{C}) = H(D \mid C - R_i) = 0$ 相矛盾, 假设 $POS_{C-R_i}(D) \neq U$ 不成立.

故有 $POS_{C-R_i}(D) = POS_C(D) = U$

根据式 (14), 故 R_i 为 C 中相对于 D 可省略的.

证明完毕.

根据命题 2, 若属性 R_i, $R_j \in C$, $X_1, X_2, \cdots, X_n$ 是基于 $\frac{U}{R_i}$ 等价类, $Y_1, Y_2, \cdots, Y_m$ 是基于 $\frac{U}{R_j}$ 的等价类, 则相应的信息熵满足:

$$H(\frac{D}{C} - \{R_i\}) = H(\frac{D}{C} - \{R_j\}) \tag{16}$$

并且对于任何一个 X_k 都有与之相对应的 Y_l, 使得

$$POS_{C-R_j}(X_k) = POS_{C-R_j}(Y_l) \tag{17}$$

其中: $X_k \in X_1, X_2, \cdots, X_n$, $Y_l \in Y_1, Y_2, \cdots, Y_m$.

则: 属性 R_i 和属性 R_j 粗等价, 即知识 R_i, R_j 对于系统来说至少有一个为冗余属性.

这里从信息熵的角度以统计概率的方法解决了对于不确定系统中的知识冗余和知识简化问题.

下面我们从 Rough 集理论和信息熵的角度出发, 考察一个知识系统 (数据见表 1)[9] 的知识冗余和知

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

识约简,来验证本文的理论和方法的可行性和有效性.

这里,条件属性集 $C = \{R_1, R_2, R_3\}$,决策属性为 $D = D$.

因为 $U \mid C = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$,

$U \mid D = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$,

$U \mid C = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$,

有 $U \mid (C - R_2) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$,

$U \mid (C - R_3) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$.

条件属性 R_2 或 R_3 对于系统来说,至少有一个为冗余的,可省略其中的一个属性来简化系统.

表 1 一个知识系统的数据表

Tab. 1 A data table of a knowledge system

U	R_1	R_2	R_3	D
x_1	1	1	C	1
x_2	2	1	C	2
x_3	3	2	B	3
x_4	1	2	B	4
x_5	2	3	A	5
x_6	3	3	A	6

根据 $H(D \mid C - R_2) = -\sum_{i=1}^{6} p(x_i) \sum_{j=1}^{6} p(Y_j \mid x_i) \log(p(Y_j \mid x_i)) = 0$

$$H(D \mid C - R_3) = -\sum_{i=1}^{6} p(x_i) \sum_{j=1}^{6} p(Y_j \mid x_i) \log(p(Y_j \mid x_i)) = 0$$

且有 $POS_{C-R_2}(X_k) = POS_{C-R_3}(Y_l)$.

条件属性 R_2 或 R_3 对于系统来说,至少有一个为冗余的,可省略其中的一个属性来简化系统.

上述分析表明,我们提出的基于信息熵和基于 Rough集理论方法来分析知识冗余和知识简化两种方法得到的结果是一致的.

4 结 论

对于一个不确定系统,可以从统计测度的角度研究系统的不确定信息,也可以基于 Rough集理论研究系统的不确定信息.根据一个不确定系统的条件属性对决策结果产生的依赖关系,我们对知识冗余和知识简化进行深入分析,研究了基于信息熵和基于 Rough集理论知识冗余和知识简化两种方法之间的关系与一致性,得出:

1) 首次建立了不确定系统知识冗余分析和知识简化方法的统一;

2) 从信息熵的角度,以统计概率的分析方法提出了不确定系统的知识冗余分析新的知识简化方法.

参考文献:

[1] Pawlak Z. Some issues on rough sets[C]//Swiniarski B, Szczuka R W, eds. Transactions on Rough Sets. Springer-Verlag Berlin, 2004: 1-58.

[2] 曾黄麟. 智能计算——关于粗集、模糊、人工神经网络理论及其应用[M]. 重庆: 重庆大学出版社, 2004: 1-60.

[3] Qing Shen. Attribute Reduction of Multi-valued Information System Based on Conditional Information Entropy[C]//Proceedings of IEEE International Conference on Granular Computing. China, Aug. 2008: 562-566.

[4] Huanglin Zeng, Xiaohui Zeng. Redundant Data Processing Based on Rough-Fuzzy Approach[J]. Rough Sets and Knowledge Technology, Springer-Verlag Berlin Heidelberg, Germany, 2006: 156-161.

[5] Huanglin Zeng, Xiaohui Zeng. Studies on consistence of the knowledge simplification of an uncertain system[C]//Proceeding of 8th International Conference on Cognitive Informatics, HongKong, 2009: 162-166.

[6] Huanglin Zeng, Xiaohui Zeng. Reasoning Decision Rules of an Uncertain System[C]//The Fourth International Conference on Rough Sets and Knowledge Technology, Proceeding of RSKT 2009, Australia, Springer-Verlag Berlin Heidelberg, 2009: 634-642.

[7] Huanglin Zeng, Yan Huang, Xiaohui Zeng. A New Approach of Attribute Reduction Based on Ant Colony Optimization[C]//Proceeding of Fifth International Conference on Natural Computation, 2009(3): 3-7.

[8] 叶全明, 胡学钢. 一种基于属性重要性的属性约简启发式算法[J]. 计算机科学, 2008, 35(8A): 28-29.

[9] 张海云, 梁吉业. 一种基于划分的决策表属性约简算法[J]. 计算机科学, 2008, 35(8A): 18-21.

[10] 靳孝方, 祝峰, 等. 改进的基于属性重要度的决策表属性约简算法[J]. 计算机科学, 2009, 36(8A): 4-7.

[11] 李磊军, 米据生. 信息系统属性约简的比较研究[J]. 计算机科学, 2009, 36(8A): 42-44.

[12] 刘静, 米据生. 概率信息系统的属性约简[J]. 计算机科学, 2009, 36(8A): 45-48.

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net