

文章编号: 1005—0523(2006)04—0093—03

基于 Rough 熵的决策表约简算法研究

宋 岚, 黄兆华, 王 洪

(华东交通大学 信息工程学院, 江西 南昌 330013)

摘要: 知识约简是 Rough 集理论研究中的核心内容之一, 已证明求决策表的最小约简是 NP-hard 问题, 解决办法一般是采用启发式算法. 本文分析了信息熵在 Rough 集理论中的应用, 利用 Rough 熵定义了决策表中条件属性的相对重要度, 讨论了其性质, 在此基础上提出了一种用于计算决策表中的相对约简的启发式算法.

关 键 词: Rough 集; Rough 熵; 知识约简

中图分类号: TP391.72

文献标识码: A

0 引言

Rough 集理论是八十年代初由波兰科学家 Z. Pawlak 首先提出的一个分析数据的数学理论. 知识约简是 Rough Set 理论的核心内容之一. 所谓知识约简, 就是在保持知识库的分类或决策能力不变的条件下, 删除其中不相关或不重要的知识. 因此, 人们期望找到具有最少属性的约简, 即最小约简. 目前, 人们做了许多的工作, 提出了多种属性约简算法. 如基于正区域的属性约简算法^[1], 基于区分矩阵的属性约简算法^[2]和基于信息熵的属性约简算法^[3].

然而, 遗憾的是 Wong. S. K. M 和 Ziarko. W 已经证明找出一个决策表的最小约简是 NP-hard 问题, 在人工智能中, 解决这类问题的一般方法是采用启发式算法^[4]. 本文首先从信息角度对决策表中的属性的重要性进行了定义, 提出了基于 Rough 熵的知识相对约简算法.

1 基本概念

在 Rough Set 理论中, 依据是否考虑决策属性

而将知识库分别表示为信息系统或决策表的形式.

1.1 知识的不确定性度量

设 $S = (U, A)$ 是一个信息系统, $P, Q \subseteq A$, 令 $U/P = \{X_1, X_2, \dots, X_n\}$, $U/Q = \{Y_1, Y_2, \dots, Y_m\}$.

记 $P(X_i) = \frac{|X_i|}{|U|}$ 知识熵的定义为^[5,8]:

$$H(P) = - \sum_{i=1}^n P(X_i) \log P(X_i)$$

记 $P(Y_j|X_i) = \frac{|Y_j \cap X_i|}{|X_i|}$. 知识 Q 相对于知识 P 的条件熵定义^[6,7]为:

$$H(Q|P) = - \sum_{i=1}^n P(X_i) \sum_{j=1}^m P(X_i) P(Y_j|X_i) \log P(Y_j|X_i)$$

P 和 Q 的联合熵为: $H(P, Q) = H(P|Q) + H(Q)$.

P 和 Q 的互信息为: $I(P, Q) = H(P) - H(P|Q) = H(Q) - H(Q|P) = H(P) + H(Q) - H(P, Q)$.

定理 1^[8] 设 $S = (U, A)$ 是一信息系统, $P, Q \subseteq A$, 若 $U/IND(Q) \subseteq U/IND(P)$, 则 $H(P) \leq H(Q)$.

定理表明知识熵随着信息粒度的变小(通过更

收稿日期: 2005—10—08

基金项目: 江西省自然科学基金项目(编号为 0411035).

作者简介: 宋岚(1978—), 女, 江西万年人, 华东交大在读研究生, 研究方向: 人工智能.

中国知网 <https://www.cnki.net>

细的划分)而单调增加.

如果 X 是一个有限集合, 则不确定的 Hartley 度量定义为:

$$H_0(X) = \log |X|$$

知识熵与不确定性的 Hartley 度量之间有如下关系:

$$H(P) = - \sum_{i=1}^n \frac{|X_i|}{|U|} \log \frac{|X_i|}{|U|} = H_0(U) - \sum_{i=1}^n \frac{|X_i|}{|U|} H_0(X_i)$$

这样, 知识熵是集合 U 的 Hartley 度量减去划分中元素的权 Hartley 度量之和.

已知知识 P 时知识 Q 的正则条件熵为:

$$H_0(Q|P) = - \sum_{i=1}^n P(X_i) \sum_{j=1}^m P(Y_j|X_i) \log P(Y_j|X_i) / \log m$$

$H_0(Q|P)$ 有如下性质:

定理 2^[7] 设 $S(U, A)$, $P, Q \subseteq A$, 则:

(1) $0 \leq H_0(Q|P) \leq 1$.

(2) 属性集 Q 函数依赖于属性集 P 当且当 $H_0(Q|P) = 0$.

(3) 属性集 Q 函数完全独立于属性集 P 当且当 $H_0(Q|P) = 1$.

1.2 Rough 熵与属性重要度的概念

设 $P \subseteq A$, 定义知识 P 的 Rough 熵为:

$$E(P) = - \sum_{i=1}^n Q_i \log P_i$$

其中, $P_i = \frac{1}{|D(S_p(x_i))|}$, $Q_i = \frac{|D(S_p(x_i))|}{|U|}$, $|$ 表示其中包含的集合的基数.

设 $P \subseteq C$, 定义如下关系: $SIM(P) = \{(x, y) \in U \times U | \forall a \in P, f(x, a) = f(y, a)\}$.

x 的 p —相似类表示可能与 x 关于 p 有相似关系的最大的对象集合: $S_p(x) = \{y \in U | (x, y) \in SIM(P)\}$

$U/SIM(P) = \{S_p(x) | x \in U\}$ 中的元素称为相似类或者信息粒, 并构成 U 的一个覆盖.

用 $D(S_p(x_i))$ 表示 x_i 的相似类的对象在决策属性上的所有可能的取值.

性质 1(单调性) $T = (U, C \cup D, V, F)$, 设 P 和 Q 是决策表的条件属性 C 的两个子集, 若 $U/(SIM(P)) \subset U/(SIM(Q))$, 则 $E(P) < E(Q)$.

证明 由 $U/(SIM(P)) \subset U/(SIM(Q))$ 得 $D(S_p(x_i)) \subset D(S_q(x_i))$, 设 $D(S_p(x_i)) = \{P_1, P_2, \dots, P_n\}$, $D(S_q(x_i)) = \{Q_1, Q_2, \dots, Q_n\}$, 若 $\forall P_i \in P, \exists Q_i \in Q$, 都有 $P_i \subseteq Q_i$, 显然 $m > n$, 且存在

$\{1, 2, \dots, m\}$ 的一个划分 $B = \{B_1, B_2, \dots, B_n\}$, 使得 $Q_j = \sum_{i \in B_j} P_i, j = 1, 2, \dots, n$, 有 $E(Q) = - \sum_{j=1}^n$

$$\frac{|D(S_q(x_i))|}{|U|} \log \frac{1}{|D(S_q(x_i))|} \\ = \frac{1}{|U|} \sum_{j=1}^n |D(S_q(x_i))| \log |D(S_q(x_i))| \\ = \frac{1}{|U|} \sum_{j=1}^n \sum_{i \in B_j} |D(S_p(x_i))| \log \sum_{i \in B_j} |D(S_p(x_i))|$$

又因为 $\sum_{i \in B_j} |D(S_p(x_i))| \log \sum_{i \in B_j} |D(S_p(x_i))| \geq \sum_{i \in B_j} |D(S_p(x_i))| \log |D(S_p(x_i))|$, 且 $m > n$, 则至少存在一个 $B_k \in B$, 使得 $|B_k| \geq 1$, 有

$$\sum_{i \in B_j} |D(S_p(x_i))| \log \sum_{i \in B_k} |D(S_p(x_i))| \geq \sum_{i \in B_k} |D(S_p(x_i))| \log |D(S_p(x_i))|$$

故

$$E(Q) > \frac{1}{|U|} \sum_{j=1}^n \sum_{i \in B_j} |D(S_p(x_i))| \log |D(S_p(x_i))| \\ = \frac{1}{|U|} \sum_{j=1}^n |D(S_p(x_i))| \log |D(S_p(x_i))| = E(P)$$

即 $E(P) < E(Q)$, 结论得证.

上面性质说明, 若一个决策表的对象数与条件属性不变, 而决策属性的属性值划分得越细, 则其决策的信息熵就越小. 即 Rough 集的 Rough 熵随着信息粒度的变小(通过更细的划分)而单调减少.

性质 2 $T = (U, C \cup D, V, F)$ 为一决策表, $P \subseteq C$, 有 $0 \leq (E(P) \leq \log |U|)$.

证明 设 $P \subseteq C, i = 1 \dots n$, 若 X_i 的不可分辨类在 D 上的划分子集达到最多的情况时, 即 $|D(S_p(X_i))| = 1$, $E(P)$ 达到它的最小值 0.

若 X_i 的不可分辨类在 D 上的划分子集达到最少的情况时, 即 $|D(S_p(X_i))| = |U|$, $E(P)$ 达到最大值 $\log |U|$. 根据 $E(P)$ 的单调性, 知 $0 \leq E(P) \leq \log |U|$, 结论得证.

根据 Rough 熵可以得到如下定义:

定义 1 设 $T = (U, C \cup D, V, F)$ 是一个决策表, C 是条件属性集合, D 是决策属性集合, $a \in C, A = C \cup D$. 可以定义条件属性 a 在条件属性集 C 中的重要度为:

$$SIG_{A-\{a\}}(a) = E(A - \{a\}) - E(A)$$

性质 3 (1) a 在 A 中不可约等价于 $SIG_{A-\{a\}}(a) > 0$.

(2) $core(A) = \{a(A) | SIG_{A-\{a\}}(a) > 0\}$.

定义 2 $T = (U, C \cup D, V, F)$ 为一决策表, 其中 C 为条件属性, D 为决策属性, $P \subseteq C, a \in C - P$, 可以定义条件属性 a 相对于条件属性集 C 中的

重要度为:

$$SIG_{P(a)} = E(P) - E(P \cup \{a\})$$
$$a \in C - P \text{ 是不可约去的, 当且仅当 } SIG_P(a) > 0.$$

因为 $SIG_P(a) = SIG_{P \cup \{a\} - \{a\}}(a)$, 根据定义 1 有: $SIG_{P \cup \{a\} - \{a\}}(a) = E(P) - E(P \cup \{a\})$.

2 算法

利用上面给出的 Rough 熵的属性重要性的度量, 下面提出一种启发式算法, 基本思想是先计算决策表的核然后再对非核属性进行约简. 一算法

输入: 一个决策表 $T = \{U, C \cup D, V, F\}$, 其中, U 为论域, C, D 分别为条件属性集和决策属性集.

输出: 决策表的一个相对约简.

步骤 1: 计算 C 的 Rough 熵.

步骤 2: 计算条件属性集 C 相对决策属性集 D 的核属性集 $B, Att = C - B$.

步骤 3: 如果 $E(B) \neq E(C)$, 则

Begin

(1) 对每个属性 $a_j \in Att$, 计算决策属性 D 相对于条件属性集 $B \cup \{a_j\}$ 的 Rough 熵 $E(B \cup \{a_j\})$.

(2) 在 Att 中选择具有 $E(B \cup \{a_j\})$ 最小值的属性(如果有多个, 则从中选取一个与 B 的属性值组合最少的属性). $Att = Att - \{a\}, B = B \cup \{a\}$.

(3) 若 $E(B) = E(C)$ 则转步骤 4, 否则转 1.

(4) 输出决策表的一个约简 B .

End.

该算法以决策表核属性集为起点, 逐次选择使 $E(B \cup \{a_j\})$ 最小的非核条件属性 a 添加到核属性集中去, 直到满足终止条件 $E(A) = E(C)$, 得到约简结果. 设在最坏的情况下, 每次考虑的属性数依次为 $n, n-1, \dots, 1$ (n 为决策表的条件属性数), 故总次数为 $n + (n-1) + \dots + 1 = n(n+1)/2$. 因此,

在最坏的情况下, 该算法能够在 $O(n^2)$ 时间复杂性内找到满意的约简.

3 结束语

本文从信息角度出发, 综述并分析了信息熵在 Rough 集理论中的应用, 利用信息熵对决策表中属性的重要性给出度量, 提出了一种启发式算法, 但仍有一些问题需进一步研究, 如利用信息熵建立知识不确定性和模糊性之间的联系, 利用信息熵度量不完全决策表的不确定性, 利用信息熵构造不完备决策表知识约简的启发式算法等.

参考文献:

[1] HU XH, CERONE N. Learning in relational databases: A rough set approach[J]. International Journal of Computational Intelligence, 1995, 11(2):323-338.

[2] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information system[A]. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory[C]. Dordrecht: Kluwer Academic Publishers, 1992. 331-338.

[3] BEAUBOUEF T, PETRY F, ARORA G. Information theoretic measures of uncertainty for rough sets and rough relational databases[J]. information Sciences, 1998. 109:185-195.

[4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[A]. 西安交通大学学报, 2004, 38(6).

[5] 王国胤, 于洪等. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7).

[6] DOMINIK SLEZAK, JAKUB WROBLEWSKI. Order Based Genetic Algorithms for the Search of Approximate Entropy Reducts [J]. Springer - Verlag Berlin Heidelberg , 2003, pp. 308 - 311.

[7] PAWLAK Z. et al. Rough sets: probabilistic versus deterministic approach[J]. International Journal of Man - Machine Studies, 1998. 29:81-95.

[8] GEDIGA G, et al. Uncertainty measures of rough set prediction [J]. Artificial Intelligence. 1998. 106:109-137.

The Algorithm to Attribute Reduction Based on Rough Entropy in Decision Table

SONG Lan, HUANG Zhao-hua, WANG Hong

(School of Information Eng., East China Jiaotong Univ., Nanchang 330013, China)

Abstract: The attribute reduction is a core problem of the rough set theory, It has been proven that computing the optimal reduction of decision table is a NP-hard problem. In the paper here, the application of rough entropy in rough sets theory is analysed, the uncertainty measure of the importance of attribute in decision table is proposed; then, a heuristic algorithm based on rough entropy for reduction of knowledge is proposed.

Key words: rough sets; rough entropy; knowledge reduction