

基于信息熵的决策表约简

曹付元,梁吉业,钱宇华

(山西大学 计算机与信息技术学院, 山西 太原 030006)

(cfy@sxu.edu.cn)

摘 要:从信息论的角度,对决策表中属性重要性的大小进行度量,并在此基础上,提出了一种基于互信息大小的知识约简算法,实例表明能够有效得到决策表的近似最小约简。

关键词:Rough 集;知识约简;互信息;信息熵

中图分类号: TP311.13 **文献标识码:** A

Decision table reduction based on information entropy

CAO Fu-yuan, LIANG Ji-ye, QIAN Yu-hua

(School of Computer & Information Technology, Shanxi University, Taiyuan Shanxi 030006, China)

Abstract: An algorithm based on mutual information for knowledge reduction was proposed in information system, in which knowledge reduction was defined from the view of information. An illustrative example showed the application potential of the algorithm. The experiment results show that this algorithm can find the minimal reduction for decision table.

Key words: rough set; knowledge reduction; mutual information; information entropy

0 引言

Rough 集理论 1982 年由 Z. Pawlak^[1] 提出,它是一种处理模糊和不确定知识(属性)的工具。目前,它在人工智能、数据挖掘、机器学习、智能控制、故障诊断等领域有广泛的应用。

知识约简是 Rough 集理论的核心内容之一。众所周知,知识库中的知识(属性)并不是同等重要的,甚至其中某些知识是冗余的。所谓知识约简,就是在保持知识库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。

目前,对 Rough 集理论的研究基于两种观点。一是代数观点,即以不可分辨关系为基础,通过引入上近似集和下近似集,在集合运算上定义。二是信息论观点,即从信息论的角度对 Rough 集理论进行研究。

信息熵是信息论的核心内容,它由 Shannon^[2] 于 1948 年提出。文献[3]等扩展了 Shannon 的信息熵,能够有效地对 Rough 集模糊性进行度量,增益函数使其拥有了补集的本质。本文从文献[3]定义的信息熵出发,对决策表中属性的重要性进行了有效的度量。并在此基础上,提出了一种有效的决策表约简算法。

1 扩充的 Rough 集理论的信息论观点

定义 1^[3] (信息熵) 设 $K = (U, R)$ 是一近似空间, R 是 U 上的划分(等价关系),则 Rough 集的信息熵定义为 $E(R) = \sum_{i=1}^m \frac{|R_i| |R_i^c|}{|U| |U|}$, 其中 $R_i^c = U - R_i$, $|R_i| / |U|$ 表示等价类 R_i 在论域 U 的可能性, $|R_i^c| / |U|$ 表示 R_i 的补集在论域 U 的可能性。

定义 2^[3] (条件熵,互信息) 设 U 是论域, $K_1 = (U, P)$

和 $K_2 = (U, Q)$ 是关于 U 的两个知识库,其中 $P = \{P_1, P_2, \dots, P_m\}$, $Q = \{Q_1, Q_2, \dots, Q_n\}$, 则 P 相对于 Q 的条件熵定义

为 $E(Q|P) = \sum_{i=1}^n \sum_{j=1}^m \frac{|Q_i \cap P_j| |Q_i^c - P_j^c|}{|U| |U|}$, P 与 Q 的互信

息定义为 $E(Q:P) = \sum_{i=1}^n \sum_{j=1}^m \frac{|Q_i \cap P_j| |Q_i^c \cap P_j^c|}{|U| |U|}$ 。

定理 1^[3] 设 U 是论域, $K_1 = (U, P)$ 和 $K_2 = (U, Q)$ 是关于 U 的两个知识库,则有 $E(Q:P) = E(Q) - E(Q|P)$ 。

定义 3 (决策表的互信息) 设 $T = \langle U, C \cup D, V, f \rangle$, $P \subseteq C$, 则定义 P 对 D 的互信息为 $E(Q:P) = \sum_{i=1}^n \sum_{j=1}^m \frac{|d_i \cap P_j| |d_i^c \cap P_j^c|}{|U| |U|}$, 其中 $P = \{P_1, P_2, \dots, P_m\}$, $D = \{d_1, d_2, \dots, d_n\}$ 。

定理 2 设 $T = \langle U, C \cup D, V, f \rangle$, P 和 Q 分别为 U 上的两个属性集合。若 $IND(P) = IND(Q)$, 则有 $E(D:P) = E(D:Q)$ 。

证明: 因为 $IND(P) = IND(Q)$, 所以 P 和 Q 在 U 上得到的划分是相同的, 故有 $E(D:P) = E(D:Q)$ 。

注: (偏序) 设 U 为论域, P 和 Q 是 U 上的划分(等价关系), 定义 $P \leq Q \Leftrightarrow \forall P_i \in P, \exists Q_j \in Q \rightarrow P_i \subseteq Q_j$, 表明 P 对 U 的划分是 Q 对 U 划分的加细。

定理 3^[3] 设 U 为论域, $K_1 = (U, P)$ 和 $K_2 = (U, Q)$ 是关于 U 的两个知识库, D 是 U 的决策属性。如果 $P < Q$, 则有 $E(D:P) \geq E(D:Q)$ 。

定理 4 设 (U, A) 是一个信息系统, $P \subseteq A$, $a \in A - P$, 则 $P \cup \{a\} \leq P$ 。

证明: 根据偏序的定义, 随着属性的增加, 属性对对象的

划分在加细。

定理 5 设 $T = \langle U, C \cup D, V, f \rangle$, $a_i \in C, i = 1, 2, 3, \dots, m (m = |C|)$, 则有 $E(D; \{a_1\}) \leq E(D; \{a_1\} \cup \{a_2\}) \leq \dots \leq E(D; \{a_1\} \cup \dots \{a_i\} \cup \dots \{a_m\}) = E(D; C)$ 。

证明:由定理 3 和定理 4 可证明定理 5。

对于决策表的相对约简,我们有如下定理。

定理 6 设 $T = \langle U, C \cup D, V, f \rangle$, 且论域 U 是在 C 相对于 D 是一致的, 则 C 中的一个属性 a 相对于决策属性 D 是不必要的, 其充分必要条件为 $E(D; C) = E(D; C - \{a\})$ 。

定理 7 设 $T = \langle U, C \cup D, V, f \rangle$, 且论域 U 是在 C 相对于 D 是一致的。则 C 相对于决策属性 D 是对立的, 其充分必要条件为 $E(D; C) \neq E(D; C - \{a\})$ 。

定理 8 设 $T = \langle U, C \cup D, V, f \rangle$, 且论域 U 是在 C 相对于 D 是一致的, 则 $P \subseteq C$ 是 C 相对于决策属性 D 的一个约简的充分必要条件为 $E(D; C) = E(D; P)$, 且 P 相对于决策属性 D 是独立的。

定义 4 (属性重要性的信息论观点) 设 $T = \langle U, C \cup D, V, f \rangle$, $P \subset C$, 则对于任意属性 $a \in C - A$ 的重要性定义为 $SGF(a, A, D) = E(D | P) - E(D | P \cup \{a\})$ 。

定义 5 (属性重要性的代数观点)^[4] F 是属性集 D 导出的分类, C 是条件属性集, $P \subset C$, 则对于任意属性 $a \in C - A$ 的重要性定义为 $SGF(a, A, D) = r_{P \cup \{a\}}(F) - r_A(F)$ 。

$SGF(a, A, D)$ 的值越大, 说明在 A 的条件下, 属性 a 对于决策 D 越重要。

下面证明属性重要性的信息定义包含了其代数定义。

定理 9 如果 $E(D | P \cup \{a\}) = E(D; P)$, 则 $POS_{A \cup \{a\}}(F) = POS_A(F)$ 。

证明:由定理 5 可知, 随着条件属性的增加, 将导致互信息的上升, 只有在加细后对于决策类的隶属度相等的情况下, 才可能不导致互信息的变化。

其次, 划分 $U | Ind(P)$ 是可以通过将划分 $U | Ind(P \cup \{a\})$ 中的部分等价类合并得到的, 如果 $E(D | P \cup \{a\}) = E(D; P)$, 则所有被合并在一起的等价类对于决策类的隶属度均相等。因此在合并后, 每个条件属性分类中的等价类对于各个决策属性分类的隶属度不会发生变化。所以 $POS_{A \cup \{a\}}(F) = POS_A(F)$ 。

2 基于互信息的信息约简算法

由 Rough 集理论知道, 任何决策表的相对核是唯一的, 而且它包含在所有的相对约简之中, 所以相对核可以作为求最小知识约简的起点。基于互信息的信息约简算法 MIBARK * 以决策属性 D 与条件属性 a 的 $E(D; \{a\})$ 的大小作为条件属性 a 对于决策的参考重要度, $E(D; \{a\})$ 的值越大, 属性 a 对于决策的参考重要度越大。算法的起点是初始条件属性集, 采用逐步删除属性来达到约简的目的。

算法 MIBARK *

输入: 一个决策表 $T = \langle U, C \cup D, V, f \rangle$, 其中, U 为论域, C, D 分别为条件属性集和决策属性集。

输出: 该决策表的一个相对约简 B 。

Step1 计算决策表 T 中决策属性 D 与条件属性 C 的互信息 $E(D; C)$ 。

Step2 计算决策属性相对于每个条件属性的互信息 $E(D; \{a_i\}) (a_i \in C)$, 将 a_i 按 $E(D; \{a_i\})$ 升序排列。

Step3 令 $B = C$, 按 $E(D; \{a_i\})$ 递增的顺序对每个 a_i 重复

下列操作:

Step3.1 计算决策属性集相对于条件属性集 B 在删掉 a_i 后的互信息 $E(D; C - \{a_i\})$;

Step3.2 如果 $E(D; C) = E(D; C - \{a_i\})$, 则属性 a_i 应约简, $B = B - \{a_i\}$; 否则, 属性 a_i 不能被约简, B 不变。

3 例子

考虑表 1 给出的决策表。这里 $C = \{a, b, c\}, D = \{d\}$ 。

表 1 决策表

U	a	b	c	d	U	a	b	c	d
1	2	2	0	1	4	0	0	0	0
2	1	2	0	0	5	1	0	1	0
3	1	2	0	1	6	2	0	1	1

由表 1, 我们有:

$U/ind(a) = \{[1, 6], [2, 3, 5], [4]\}$,

$U/ind(b) = \{[1, 2, 3], [4, 5, 6]\}$,

$U/ind(c) = \{[1, 2, 3, 4], [5, 6]\}$,

$U/ind(d) = \{[1, 3, 6], [2, 4, 5]\}$

根据算法 MIBARK *:

Step1 $E(D; C) = 0.22$;

Step2 $E(D; \{c\}) = 0.1, E(D; \{b\}) = 0.14,$

$E(D; \{a\}) = 0.2$;

Step3 由 $E(D; C - \{c\}) = E(D; C)$, 所以 c 可删除; 则 $\{a, b\}$ 为表 1 的一个约简。

4 结语

Rough 集理论为开发自动规则生成系统提供了一种工具。它通过对决策表进行知识约简, 从而导出其决策规则。由于知识约简的不唯一性, 使得知识约简的优劣直接影响着决策规则的繁简。人们期望得到关于决策表的最简洁的规则, 这就需要计算决策表的最小约简。然而, 遗憾的是已经证明找出一个决策表的最小约简是 NP-hard 难题。

本文从信息论的观点定义了决策表中条件属性对决策属性的重要性, 提出了一种基于互信息的信息约简算法, 实例表明该算法能够得到决策表的近似最小约简。

参考文献:

- [1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11: 341-356.
- [2] SHANNON CE. The mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3/4): 373-423.
- [3] LIANG JY, CHIN KS, DANG CY, et al. A new method for measuring uncertainty and fuzziness in rough set theory[J]. International Journal of General Systems, 2002, 31(4): 331-342.
- [4] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 1-8.
- [5] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(9): 681-684.
- [6] 徐燕, 怀进鹏, 王兆其. 基于区分能力大小的启发式约简算法及其应用[J]. 计算机学报, 2003, 26(1): 97-103.
- [7] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
- [8] LIANG JY, XU ZB. The algorithm on knowledge reduction in incomplete information system[J]. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 10(1): 95-103.