第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·软件技术与数据库· 文章编号：1000—3428(2007)02—0068—03 文献标识码：A 中图分类号：TP18

一种新的决策表约简方法

李 明，黄文涛，王 丽
（兰州理工大学计算机与通信学院，兰州 730050）

摘 要：信息熵理论已经被证明可以作为一种有效的属性约简的方法，是基于粗糙集理论研究的最新研究成果，该文揭示信息表与决策表之间的联系，从该联系出发，用信息熵理论对决策表进行约简，为寻找更高效的决策表约简算法奠定了基础。
关键词：粗糙集；信息表；决策表；信息熵；约简

A New Method of Decision Table Reduction

LI Ming, HUANG Wentao, WANG Li
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050)

【Abstract】 Information entropy theory proved an effective method of the attributes reduction. Based on the latest research of rough sets, this paper demonstrates the relation between the decision table and information table, proposes a new method using the information entropy to reduce decision table. It lays the foundation for finding the more efficient decision table reduction algorithm.
【Key words】 Rough set; Information table; Decision table; Information entropy; Reduction

粗糙集理论是由波兰科学家 Z.Pawlak 于 1982 年提出的[1]，它具有很强的定性分析能力，是一种能够有效地处理不精确、不连续、不完整数据的数学工具，经过 20 多年的发展，目前已经被广泛应用于粒度计算、聚类分析、软计算、信息检索、图像和信号识别、数据挖掘等领域。众所周知，知识库中的知识不是同等重要的，有些甚至是冗余，这些冗余知识会干扰决策者做出正确的决策，因此完全有必要对知识库中的知识进行约简。但遗憾的是，有人已经证明知识约简是 NP-hard 问题。因此，有许多学者提出各种不同的启发式约简算法[2~4]。这些约简算法大致可以分为 3 类：(1)基于代数理论的；(2)基于信息熵理论的：(3)基于区分矩阵和区分函数的。

粗糙集理论是以不可辨分关系为基础，通过引入上、下近似集，在集合运算上定义的，这通常被称为粗糙集理论的代数观点。此外还有一些学者从信息论的角度出发，提出了粗糙集理论的信息论观点。本文将主要研究粗糙集理论的信息论观点，从信息熵出发得出信息表的核和约简，并描述了信息表与决策表之间的一些内在联系。因为处理信息表的时间复杂度要比决策表低一个数量级，所以不影响算法的整体复杂度，但在客观上却降低了决策表的规模，所以本文运用条件信息熵和决策表与信息表之间的内在联系提出了决策表的约简算法，最后用实例说明算法的有效性。

1 粗糙集的基本概念

粗糙集理论从新的视角出发对知识进行了定义，将知识看作一种分类能力，从而使得对知识能够进行分析和处理，设U 是一个论域，可以认为U 上任一属性集合(知识、等价关系族)是定义在U 上的子集组成的σ-代数上的一个随机变量，其概率分布可以通过以下方法来定义。

定义 1 设 P,Q 在 U 上导出的划分分别为 $X,Y(X=\{X_1,X_2,...,X_n\}, Y=\{Y_1,Y_2,...,Y_m\})$，则 P,Q 在 U 上的子集组成的σ-代数上的概率分布为

$$[X:p]=\begin{bmatrix} X_1,\cdots,X_n \\ p(X_1),\cdots,p(X_n)\end{bmatrix},\quad [Y:p]=\begin{bmatrix} Y_1,\cdots,Y_m \\ p(Y_1),\cdots,p(Y_m)\end{bmatrix}$$

其中 $p(X_i)=|X_i|/|U|$，$i=1,2,...,n$；$p(Y_j)=|Y_j|/|U|$，$j=1,...,m$。

定义 2 知识(属性集合) P 的熵 $H(P)$ 定义为

$$H(p)=-\sum_{i=1}^{n} p(X_i)\lg p(X_i)$$

定义 3 知识（属性集合）$Q(U/IND(P)=\{Y_1,Y_2,...,Y_m\})$ 相对于知识（属性集合）$P(U/IND(P)=\{X_1,X_2,...,X_n\})$ 的条件熵：

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

其中：

$$p(Y_j/X_i)=|Y_j\cap X_i|/|X_i|,\quad i=1,\ldots,n,\quad j=1,\ldots,m$$

设由属性集合 P 和 $D=\{d\}$ 导出的相对论域 $U(|U|=n)$ 的划分分别为

$$U/IND(P)=\{X_1,...,X_t\},\ U/IND(\{d\})=\{Y_1,Y_2,...,Y_s\}$$

则有如下的定理[5]：

定理 1 设U 是论域，P,Q 是U 上的两个属性集合，若 $IND(Q)=IND(P)$，则 $H(P)=H(Q)$。

定理 2 设U 是一个论域，P,Q 是U 上的两个属性集

基金项目：甘肃省自然科学基金资助项目(3ZS042-B25-007)
作者简介：李 明(1959－)，男，教授，主研方向：智能信息处理；黄文涛、王 丽，硕士生
收稿日期：2006-03-20 **E-mail：**huangwent123@163.com

合，且 $P \subseteq Q$，如果 $H(P)=H(Q)$，则 $IND(P)=IND(Q)$。

对于约简而言，信息熵表示形式与代数表示形式是等价的，可以从信息熵的角度来研究属性约简问题[5]。首先我们给出一些相关的概念。

定义 4 设$<U,C\cup D,V,f>$为一个决策表，其中 $C\cap D=\Phi$，且 C 为条件属性集，D 为决策属性集，则称$<U,C,V,f>$为$<U,C\cup D,V,f>$所对应的信息表。

定义 5 若$<U,C\cup D,V,f>$为决策表，则称$<U,C\cup D,V,f>$的核为相对核，而其所对应的信息表的核为绝对核。

定理 3 若 $C_1=RED(C)$, 则 $POS_{C_1}D=POS_CD$。

证明：因为 $C_1=RED(C)$，所以有下式成立：$IND(C_1)=IND(C)$，从而有：$U/IND(C)=U/IND(C_1)$，因此：$POS_{C_1}D=POS_CD$。

定理 4 相对核是绝对核的子集。

证明：设 $C_1=\{C_{11},C_{12},\cdots,C_{1n}\}$ 是信息表的所有约简的集合，显然对 $\forall C_{1i}\in C_1,(i=1,2,\cdots,n)$，由定理 3 可得：$POS_{C_{1i}}D=POS_CD$。再设：$C_2=\{C_{21},C_{22},...,C_{2m}\}$ 是决策表的所有约简的集合，则对 $\forall C_{1i}\in C_1$，均 $\exists C_{2j}\in C_2$，使得 $C_{2j}\subseteq C_{1i}$，而由相对核的定义可知：

$$C_2'=\bigcap_{j=1}^{m}C_{2j}$$

由绝对核的定义可知：

$$C_1'=\bigcap_{i=1}^{n}C_{1i}$$

所以有 $C_2'\subseteq C_1'$。

该定理保证了以下“基于绝对核的约简算法”不会因为核值属性的缺省而失败。

定理 5 $a_i\in C$ 是可约的 $\Leftrightarrow H(C)=H(C-\{a_i\})$。

证明：“$\Rightarrow$” $a_i\in C$ 是可约的，则 $IND(C)=IND(C-\{a_i\})$，由定理 2 可得：$H(C)=H(C-\{a_i\})$。“$\Leftarrow$”，显然 $(C-\{a_i\})\subseteq C$，由定理 1 可得：$IND(C)=IND(C-\{a_i\})$，因此是可约的。

定理 6 设 U 是一个论域，某个等价类在 U 上的划分为 $A_1=\{X_1,X_2,...,X_n\}$，而 $A_2=\{X_1,...,X_{i-1},X_{i+1},...,X_{j-1},X_{j+1},...,X_n,X_i\cup X_j\}$ 是将划分 A_1 中的某两个等价块 X_i 和 X_j 合并得到的新划分，则 $H(A_1)>H(A_2)$。

证明：因为

$$H(A_1)=-\sum_{k=1}^{n}p(X_k)\lg p(X_k)$$

$$H(A_2)=-\sum_{k=1}^{n}p(X_k)\lg p(X_k)+p(X_i)\lg p(X_i)$$

$$+p(X_j)\lg p(X_j)-p(X_i\cup X_j)\lg p(X_i\cup X_j)$$

所以

$$\Delta H=H(A_2)-H(A_1)$$

$$=p(X_i)\lg p(X_i)+p(X_j)\lg p(X_j)$$

$$-p(X_i\cup X_j)\lg p(X_i\cup X_j),$$

因为 $X_i\cap X_j=\Phi$，所以 $p(X_i\cup X_j)=p(X_i)+p(X_j)$

$$\Delta H=p(X_i)[\lg p(X_i)-\lg p(X_i\cup X_j)]+p(X_j)$$

$$[\lg p(X_j)-\lg p(X_i\cup X_j)]$$

故 $\Delta H<0$，由此可以得出：$H(A_1)>H(A_2)$。

由定理 6 可得，如将信息表属性的分类合并，将导致信息熵的单调递减。

定理 7 $a_i\in C$ 是核值属性 $\Leftrightarrow H(C-a_i)<H(C)$。

证明：“$\Rightarrow$”a_i 为 A 的核值属性，所以 a_i 在 C 中的任何约简均不可缺少，若缺省，则会引起信息表某些分类的合并，由定理 6 可知：$H(C-\{a\}_i)<H(C)$。

“$\Leftarrow$”当 $H(C-\{a_i\})<H(C)$ 时，根据定理 5 可知：a_i 是不可约的。（反证法）假若 a_i 不是核值属性，则必定存在一个 C 的约简 Q，使得：$H(Q)=H(C)$ 且 $a_i\notin Q$，因为 $Q\subseteq C$ 所以可得：$IND(Q)=IND(C)$，因为 $Q\subseteq(C-\{a_i\})$，所以可以得到下列表达式：$IND(Q)\subseteq IND(C-\{a_i\})$，由定理 6 可知：$H(C-\{a_i\})\geq H(Q)$，这与题设矛盾，故假设不成立。

基于以上分析，我们提出求绝对核的算法，再由绝对核出发求出决策表的相对约简。

算法 1 求绝对核

输入 信息表 $S=<U,C,V,f>$，其中 U 是论域，C 是条件属性集。

输出 信息表的绝对核（$ab-core$）。

Step1 计算 $H(C)$

Step2 令 $ab-core=\Phi$

Step3 for $i=1:m$

if $H(C-\{a_i\})<H(C)$

then $ab-core=ab-core\cup\{a_i\}$

Step4 返回 $ab-core$。

2 决策表的约简算法

在决策表中，每一个属性都可能有不同的重要性，本文从条件信息熵的角度，给出每个属性的重要性度量，并以此为启发信息，从绝对核出发设计约简算法。为此我们引用了一个很重要的定理。

定理 8[2]：设论域为 U，某个等价关系在 U 上形成的划分：

$A_1=\{X_1,...,X_n\}$，$A_2=\{X_1,...,X_{i-1},...,X_{i+1},...,X_{j-1},...,X_{j+1},...,X_n,X_i\cup X_j\}$，而是将划分 A_1 中的某两个等价块 X_i 和 X_j 合并为 $X_i\cup X_j$ 得到的新划分，$B=\{Y_1,...,Y_m\}$，也是 U 上的一个划分，则 $H(B/A_2)\geq H(B/A_1)$。

该定理说明条件熵的值越大，条件属性相对于决策属性的不确定程度越高。粗糙集理论认为：知识是基于对象分类的能力。粒度越大，说明知识越粗糙。知识越粗糙一般越难得到确定的规则。基于以上分析，条件熵越大，说明条件属性相对于决策属性的不确定性越大，即越不重要。

设 B 为决策表$<U,C\cup D,V,f>$的一个条件属性子集，初始时令：$B=ab-core$，$P=\Phi$，对 $\forall a_i\in B$，属性重要性定义为 $SGF(a_i)=H(D/\{a_i\})$，对 $\forall a_i\in C-B$，属性重要性定义为 $SGF(a_j)=H(D/B\cup\{a_i\})$ 条件信息熵越小，说明属性越重要。

算法 2 基于绝对核的相对约简

输入　决策表 $S=<U,C\cup D,V,f>$

其中 C,D 分别为条件属性和决策属性

输出　决策表的一个约简 B

Step1 令 $B=ab-core$，计算 $POS_B D$ ；

Step2 如果 $POS_B D=POS_C D$，则转到 Step3，否则转到 Step5；

Step3 计算集合 B 中所有属性 a_j 的条件信息熵 $H(D/\{a_j\})$，并将其按 $H(D/\{a_j\})$ 递减顺序排列；

Step4 按排好的顺序对 B 中每一个属性执行下列操作：

Step4.1 计算 $POS_{(B-\{a_j\})}(D)$ ；

Step4.2 若 $POS_{(B-\{a_j\})}(D)=POS_B(D)$，则 a_j 在 B 中可以约简，$B=B-\{a_j\}$，否则 a_j 不能约简，B 保持不变；

Step5 在 $(C-B)$ 中选择使 $H(D/B\cup\{a_i\})$ 最小的属性 a_k (若同时有几个达到最小，则选择组合数最小的)，令 $B=B\cup\{a_k\}$，计算 $POS_B(D)$ ；

Step6 若 $POS_B(D)=POS_C(D)$，则输出 B，否则转到 Step5。

算法 3　二步约简法

输入　决策表 $S=<U,C\cup D,V,f>$，其中 C,D 分别为条件属性和决策属性

输出　决策表的一个约简 B

初始化 $B=C$

Step1 计算信息表中的信息熵 $H(C)$ ；

Step2 对信息表中的每一个属性执行下列操作：

Step2.1 计算 $H(C-\{a_i\})$ ；

Step2.2 若 $H(C-\{a_i\})=H(C)$，则 $B=B-\{a_i\}$，否则 B 不变；

Step3 计算决策表中 $POS_B(D)$ 及 B 中的每一个的属性 a_j 的条件熵 $H(D/\{a_j\})$，并将 a_j 按 $H(D/\{a_j\})$ 的降序排列(若有几个属性的条件熵相等，则将组合数多的排列在前)；

Step4 按 $H(D/\{a_j\})$ 的递减顺序，对 B 中的每一个属性执行下列操作：

Step4.1 计算 $POS_{B-\{a_j\}}(D)$ ；

Step4.2　若 $POS_{B-\{a_j\}}(D)=POS_B(D)$，则 a_j 在 B 中可以约简，$B=B-\{a_j\}$，否则 a_j 不能被约简，B 不变。

算法分析：

假设论域 U 中有 n 个对象，即 $|U|=n$，条件属性中有 m 个属性，即 $|C|=m$，在最坏的情况下算法的复杂度为 $O(n^3+mn^2)$，因为，一般情况 $m<<n$，所以算法的最终复杂度为 $O(n^3)$。

3 相关实验

为了考察以上算法的有效性，本文选择文献[6]中的一个决策表进行分析论域为 U，条件属性集 C ={size,cylinder,turbocharge,fuel,displacement,compression,power,transmission,weight}，决策属性集 D ={mileage}。

下面利用本文提出的两种算法对决策表进行约简。对于算法 2 先求绝对核（由算法 1 求得）$ab-core$ ={size,cylinder, fuel,compression,power,transmission,weight}。

(1)对决策表 $POS_C(D)=U$ ，计算得 $POS_{ab-core}(D)=U$ 。

(2)根据算法 2，应删除绝对核中的元素，计算重要性（见表 1）。

表 1　重要性

size	cylinder	fuel	compression
0.338	0.279	0.405	0.352
power	transmission	weight	/
0.367	0.342	0.207	/

排序得{fuel, power, compression, transmission, size, cylinder, weight}。

(3)最后求得一个约简为 B ={cylinder,fuel, compression, power, weight}。

对于算法 3 先求出信息表的一个约简 C' ={size, cylinder, fuel,displacement,compression,power,transmission,weight}。

(1)计算信息表约简中属性的重要性，并将其排序，结果见表 2。

(2)按照算法 3 对以上属性依次逐一进行计算得到一个约简 B' ={size,fuel,displacement,weight}。

表 2　排序结果

transmission	fuel	power	compression
0.687	0.398	0.366	0.353
displacement	cylinder	size	weight
0.333	0.284	0.237	0.205

基于粗糙集方法的机理，由以上属性组成的约简上的分类质量指标和原始条件属性集是一样的，而以上两种约简集的规模大约只有原表的一半，有效降低了原始样本集的规模。

4 结语

以上两种算法都是首先对信息表进行处理，然后再求关于决策表的约简。因为对信息表的处理时间复杂度比决策表要低一个数量级，所以不影响算法的最终时间复杂度，但是先对信息表进行处理，一般情况下可以在一定程度上减小决策表的规模，方便以后的处理，另外，当决策属性不止一个时，即要做出多个决策时，信息表是相对固定的，可以将其存储起来，供多次使用。

粗糙集理论为开发自动生成规则系统提供了一种有效的工具，它不需要任何先验知识，通过对决策表进行知识约简，从而导出决策规则。然而遗憾的是有人已证明，寻找一个决策表的最小约简是一个 NP-hard 问题。本文从信息论的角度出发，并结合代数观点，给出一种决策表相对约简的启发式算法。最后通过实例来说明了该算法的有效性。

参考文献

1 Pawlak Z. Rough Set——Theoretical Aspect of Reasoning About Date[M]. Kluwer Academic Pub., 1991.

2 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): 759-766.

3 Xiao Jianmei, Zhang Tengfei. New Rough Set Approach to Knowledge Reduction in Decision Table[C]. Proceedings of the International Conference on Machine Learning and Cybernetics, Shanghai, 2004-08.

4 Wang Jue, Wang Ju. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method[J]. J. Comput. Sci.& Technol., 2001, 16(6): 489-504.

5 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.

6 张文修, 吴伟志. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.