第 37 卷 第 7 期
Vol.37 No.7

计 算 机 工 程
Computer Engineering

2011 年 4 月
April 2011

• 软件技术与数据库 •　　文章编号：1000—3428(2011)07—0056—03　　文献标识码：A　　中图分类号：TP301.6

粗糙集和信息熵的属性约简算法及其应用

吴尚智，苟平章

(西北师范大学数学与信息科学学院，兰州 730070)

摘　要：阐述粗糙集理论和信息熵的基本概念，并为寻找属性约简的有效方法，提出一种基于粗糙集和信息熵的属性约简算法。在决策表中添加某个属性引起的互信息变化的大小，以反映该属性的重要性，并求相对约简。研究表明，该算法不仅能得到最优的决策规则，而且能够减少信息系统所需的搜索空间，得到更优的属性约简效果。

关键词：粗糙集理论；信息熵；属性约简；信息系统

Attribute Reduction Algorithm on Rough Set and Information Entropy and Its Application

WU Shang-zhi, GOU Ping-zhang

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China)

【Abstract】 This paper expounds the basic conceptions of the rough set theory and information entropy. In order to find the effective approach of attribute reduction, an algorithm of attribute reduction based on rough set and information entropy is put forward. In decision table, a size of mutual information caused by an attribute reflects on the attribute significance, and gets the relative reduction. The studies show that the algorithm not only can get the optimal decision rules, but also can greatly decrease search space that the information system requires, and get more perfect attribute reduction effect.

【Key words】 rough set theory; information entropy; attribute reduction; information system

DOI: 10.3969/j.issn.1000-3428.2011.07.019

1　概述

粗糙集理论是一种处理不精确、不完整、不确定性数据的数学工具[1-4]。由于该理论在数据的决策与分析、模式识别、机器学习与知识发现等方面的成功应用，逐渐引起了世界各国学者的广泛关注。知识约简是粗糙集理论的核心内容之一，其主要思想是在保持分类能力不变的前提下，消除信息系统(决策表)中不必要的知识，导出最终的决策或分类规则。

本文提出了基于粗糙集和信息熵的属性约简算法，以决策表的相对核为起点，在决策表中添加某个属性引起的互信息变化的大小反映该属性的重要程度，即属性集 $R=R\cup\{x_i\}$。通过实例分析对提出的算法的有效性和可行性进行验证。

2　粗糙集理论

在粗糙集理论[3-8]中，知识是用信息系统(即属性-值对表)来表示的。在一般情况下，表中的列标记不同的属性；行标记论域的对象。如果将信息系统中的属性进一步分为条件属性和决策属性，则称该信息系统为决策表。决策表的约简又称为知识的相对约简，其最终结果是将决策表中的知识化成少量的决策规则。

定义 1 U 上的一族划分称为关于 U 的一个知识库。一个知识库就是一个关系系统 $K=(U, R)$，其中，U 是非空有限集；R 为 U 上等价关系的一个族集。U/R 表示 R 的所有等价类(或者 U 上的分类)构成的集合，$[x]_R$ 表示的是包含元素 $x\in U$ 的 R 等价类。

定义 2 若 $P\subseteq R$，且 $P\neq\phi$，则 P 中所有等价关系的交集也是一个等价关系，称为 P 上的不可分辨关系，记为 $ind(P)$，且有 $[x]_{ind(P)}=\bigcap_{R\in P}[x]_R$。

$U/ind(P)$(即等价关系 $ind(P)$的所有等价类)表示与等价关系族 P 相关的知识，称为 K 中关于 U 的 P 基本知识(P 基本集)。为简单起见，用 U/P 代替 $U/ind(P)$，$ind(P)$的等价类称为知识 P 的基本概念或基本范畴。

定义 3 设集合 $X\subseteq U$，$R\in ind(K)$，定义 2 个子集：

$$\underline{R}X=\bigcup\{Y\in U/R\mid Y\subseteq X\},\ \overline{R}X=\bigcup\{Y\in U/R\mid Y\cap X\neq\Phi\}$$

分别称它们为 X 的 R 下近似集和 R 上近似集。集合 $bn_R(X)=\overline{R}X-\underline{R}X$ 称为 X 的 R 边界域；$pos_R(X)=\underline{R}X$ 称为 X 的 R 正域；$neg_R(X)=U-\overline{R}X$ 称为 X 的 R 负域。

根据定义 3 可知，$\underline{R}X$ 或 $pos_R(X)$是指由那些知识 R 判断肯定属于 X 的 U 中元素组成的集合；$\overline{R}X$ 是指由那些知识 R 判断可能属于 X 的 U 中元素组成的集合；$bn_R(X)$是由那些知识 R 既不能判断肯定属于 X 又不能判断肯定属于~X 的 U 中元素组成的集合；$neg_R(X)$由那些知识 R 判断肯定不属于 X 的 U 中元素组成的集合。

定义 4 形式上一个四元组 $S=(U,A,V,f)$是一个知识表达系统，其中，U 为论域；A 为属性集；$V=\cup V_a(a\in A)$，V_a 是属性 a 的值域；f 为信息函数：$U\times A\rightarrow V$。

通常知识表达系统也称为信息系统，通常也用 $S=(U,A)$代替 $S=(U,A,V,f)$。

容易看出，一个属性对应一个等价关系，因此一个表就

基金项目：国家自然科学基金资助项目(71061013)；甘肃省自然科学基金资助项目(1010RJZA011)

作者简介：吴尚智(1965－)，男，副教授、硕士，主研方向：算法设计与分析，粗糙集；苟平章，副教授

收稿日期：2010-10-12　**E-mail**：wusz@nwnu.edu.cn

可以看成是定义的一族等价关系，即知识库。前文讨论的问题都可以用属性及属性值引入的分类来表示，知识约简可以转化为属性约简。

定义 5 设 $S=(U,A,V,f)$是一个知识表达系统，$A=C\cup D$, $C\cap D=\varphi$，C 称为条件属性集，D 称为决策属性集。具有条件属性集和决策属性集的知识表达系统称为决策表。

定义 6 设 U 是一个论域，P、Q 为定义在 U 上的 2 个等价关系族，且 $Q\subseteq P$，如果满足：(1)$ind(P)=ind(Q)$；(2)Q 是独立的，则称 Q 是 P 的一个约简。

定义 7 设 U 是一个论域，P 为定义在 U 上的一个等价关系族，P 中所有必要关系组成的集合，称为族集 P 的核，记作 $core(P)$。

通过以上的定义以及简单的推导，得到如下定理：

定理 1 $core(P)=\cap red(P)$，其中的 $red(P)$表示族集 P 的所有约简。

从定理可以看出，核这个概念的使用有 2 个方面：

(1)可以作为所有约简的计算基础，因为由核的定义知道，核包含在所有的约简之中，并且计算可以直接进行。

(2)核可以被解释为知识最重要部分的集合，在进行知识约简时是不能够去掉它的。

定义 8 设 P 和 Q 是 U 中的等价关系，Q 的 P 正域记为 $pos_P(Q)$，即 $pos_P(Q)=\bigcup_{x\in U/Q}\underline{P}X$，$Q$ 的 P 正域是 U 中所有根据分类 U/P 的信息可以准确地划分到关系 Q 的等价类中去的对象集合。

定义 9 设 P 和 Q 是 U 中的等价关系族，$R\in P$，如果 $pos_{ind(P)}(ind(Q))=pos_{ind(P-\{R\})}(ind(Q))$，则称 R 为 P 中 Q 不必要的；否则 R 为 P 中 Q 必要的。有时也会用 $pos_P(Q)$代替 $pos_{ind(P)}(ind(Q))$。如果 P 中的每个 R 都是 Q 必要的，则称 P 为 Q 独立的，否则就称为依赖的。

定义 10 设 $S\subseteq P$，S 为 P 的 Q 约简当且仅当 S 是 P 的 Q 独立子集族，且 $pos_S(Q)=po_P(Q)$，P 的 Q 约简称为相对约简。

P 中所有 Q 必要的初等关系构成的集合称为 P 的 Q 核，简称相对核，记为 $Core_Q(P)$。

定义 11 令 $K=(U,R)$是一个知识库，$P,Q\subseteq R$，当 $k=\gamma_P(Q)=|pos_P(Q)|/|U|$时，称知识 Q 是 $k(0\leqslant k\leqslant 1)$度依赖于知识 P 的，记作 $P\Rightarrow_k Q$。

(1)当 $k=1$ 时，称知识 Q 完全依赖于知识 P；

(2)当 $0<k<1$ 时，称知识 Q 部分(粗糙)依赖于知识 P；

(3)当 $k=0$ 时，称知识 Q 完全独立于知识 P。

可以将此解释为将对象分类的能力。确切地说，若 $k=1$，则论域的所有元素都能够用知识 P 来分类于 U/Q 的概念之中。若 $k\neq 1$，则仅仅是论域中属于正区域的那些元素能够用知识 P 来分类于 U/Q 的概念之中。若 $k=0$，则论域所有元素都不能用知识 P 来分类于 U/Q 的概念之中。因此，系数 $\gamma_P(Q)$ 可以看作是 Q 与 P 间的依赖程度。

3 基于信息熵的属性约简算法

把知识看作是关于论域的划分，从而使得对知识能够进行严密的分析和处理。

将对粗糙集理论中的知识做新的理解，如果建立知识和信息之间的关系，从信息熵的角度考察属性约简，可以获得高效的属性约简算法。

3.1 知识与信息熵的关系

将对粗糙集理论中的知识做新的理解，建立知识和信息熵的关系[9]。

定义 12 设 U 是一个论域，P 和 Q 为论域 U 上的 2 个等价关系族(即知识)，$U/ind(P)=\{X_1,X_2,\cdots,X_n\}$，$U/ind(Q)=\{Y_1,Y_2,\cdots,Y_m\}$，则 P、Q 在 U 上的子集的概论分布定义如下：

$$[X;p]=\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ p(x_1) & p(x_2) & \cdots & p(x_n)\end{bmatrix}$$

$$[Y;p]=\begin{bmatrix} y_1 & y_2 & \cdots & y_m \\ p(y_1) & p(y_2) & \cdots & p(y_m)\end{bmatrix}$$

其中，$p(x_i)=\dfrac{|x_i|}{|U|}$，$i=1,2,\cdots,n$；$p(y_j)=\dfrac{|y_j|}{|U|}$，$j=1,2,\cdots,m$；符号$|E|$表示集合 E 的基数。

有了知识概论分布的定义后，根据信息论，知识 P 的信息熵 $H(P)$为：

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

知识 P 相对于知识 Q 的条件熵 $H(Q|P)$为：

$$H(Q\mid P)=-\sum_{i=1}^{n}p(x_i)\sum_{j=1}^{m}p(y_j\mid x_i)\log p(y_j\mid x_i)$$

知识 P 与 Q 的互信息 $I(P;Q)$为：

$$I(P;Q)=H(Q)-H(Q\mid P)$$

粗糙集理论与信息熵的关系：熵度量了事件的不确定性，即信源提供的平均信息量的大小；条件熵 $H(Q\mid P)$ 度量了在事件 P 发生的前提下，事件 Q 仍存在的不确定性；互信息 $I(P;Q)$ 代表了包含在事件 P 中关于事件 Q 的信息，即互信息度量了一个信源从另一个信源获取的信息量的大小[10]。

3.2 主要概念与计算的信息关系

设 U 是一个论域，P 和 Q 为 U 上的 2 个等价关系族(即知识)，从信息的角度对粗糙集理论中的主要概念和信息进行了表达，并证明了知识约简在信息与代数 2 种不同表示下是等价的。

定理 2 设 U 是一个论域，P 和 Q 为 U 上的 2 个等价关系族，若 $ind(P)=ind(Q)$，则 $H(P)=H(Q)$。

证明：因为 $ind(P)=ind(Q)$，所以 2 个等价关系对论域 U 的划分相同，具有相同的概论分布，于是 $H(P)=H(Q)$。

定理说明，2 个代数表示下等价的知识库具有相同的信息量，但其逆不一定成立，但是有：

定理 3 设 U 是一个论域，P 和 Q 为 U 上的 2 个等价关系族，若 $P\subseteq Q$，且 $H(P)=H(Q)$，则 $ind(P)=ind(Q)$。

定理表明，当 2 个知识库存在包含关系时，由知识信息量的相等可以得到它们在代数表示下是等价的。

定理 4 设 U 是一个论域，P 和 Q 为 U 上的 2 个等价关系族，一个关系 $R\in P$ 在 P 中是不必要的(多余的)，其充分必要条件是 $H(R|P-\{R\})=0$。

定理表明，在代数表示下，不必要的知识在知识库中没有提供新的信息，可以删除；反之亦然。

推论 $R\in P$ 在 P 中的充分必要条件是：$H(R|P-\{R\})>0$。

定理 5 等价关系族 P 独立(不依赖)的充分必要条件是：对任意 $R\in P$，都有 $H(R|P-\{R\})>0$。

定理 6 设 U 是一个论域，P 和 Q 为 U 上的 2 个等价关系族，$Q\subseteq P$ 是 P 的一个约简的充分必要条件是下列 2 个条件成立：(1)$H(P)=H(Q)$；(2)对任意的 $q\in Q$，有 $H(q|Q-\{q\})>0$。

定理表明，对于知识约简而言，知识的信息表示与代数表示完全等价；但信息表示比代数表示更加直观，而且在信息表示的基础上，能够导出高效的知识约简算法。

3.3 算法原理及描述

在决策表中，关心的是哪些条件属性对于决策更重要，

就要考虑条件属性和决策属性之间的互信息[11]。因此，在决策表中添加某个属性引起的互信息变化的大小反映该属性的重要程度。

设 $T=(U,C\cup D)$为一个决策表，且 $R\subset C$，那么，在 R 中添加一个属性 $a\in C$ 之后的互信息的增量为：$I(R\cup\{a\};D)-I(R;D)=H(D|R)-H(D|R\cup\{a\})$。这里，$I(x,y)$表示 x 与 y 的互信息；$H(y|x)$表示在 x 下 y 的条件熵。如果该增量越大，说明在已知属性 R 的条件下，添加属性 a 后，对决策的影响就越大，表明属性 a 对决策 D 就越重要。

定义 13 设 $T=(U,C\cup D)$为一个决策表，且 $R\subset C$。对于任意属性 $a\in C-R$ 的重要性定义为：

$$SGF(a,R,D)=H(D|R)-H(D|R\cup\{a\})$$

若 $R=\phi$，则 $SGF(a,R,D)$变为：

$$SGF(a,R,D)=H(D)-H(D|a)=I(a;D)$$

即为属性 a 与决策 D 的互信息。

$SGF(a,R,D)$的值越大，说明在已知属性 R 的条件下，添加的属性 a 后，对决策的影响就越大，将表明属性 a 对决策 D 就越重要。

由于对于确定的决策属性 D 和已知属性集合 R，$H(D)$和 $H(D|R)$是确定的，因此在已知 R 的条件下，最重要的条件属性 a 可以描述为：$a\in C-R,\ \forall b\in C-R,\ H(D|R\cup\{a\})\leqslant H(D|R\cup\{b\})$。则算法就将 $SGF(a,R,D)$作为寻求属性约简的重要性信息，减少搜索空间。

根据粗糙集理论可知，任何决策表的相对核包含在所有的相对约简之中，所以，相对核可以作为求属性约简的起点。再由定理 6 知道，互信息相等就是寻找相对约简的终止条件。

基于互信息的相对约简算法，是以自底向上的方式求相对约简。它以决策表的相对核为起点，依据上面定义的属性的重要性，逐次选择最重要的属性添加到相对核中，直到终止条件满足。其算法描述如下：

输入 一个决策表 $T=(U,C\cup D)$，其中，U 为论域；C 和 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，即 C 的 D 约简。

Step1 计算决策表 T 的条件属性集 C 与决策属性集 D 的互信息 $I(C;D)$。

Step2 计算 C 相对于 D 的核 $Core=Core_D(C)$。

一般来说，$I(Core;D)<I(C;D)$；若 $Core=\Phi$，则 $I(Core;D)=0$。

Step3 令 $R=Core$，对条件属性集 $C-R$，重复：

(1)对每个属性 $a\in C-R$，计算条件互信息 $I(a;D|R)$。

(2)选择使条件互信息 $I(a;D|R)$最大的属性，记作 a，如果同时有多个属性达到最大值，选择与 R 属性组合数最少的属性记作 a，并且使 $R=R\cup\{a\}$。

(3)若 $I(R;D)=I(C;D)$，算法终止；否则转(1)。

Step4 最后得到的 R 就是 C 相对于 D 的一个相对约简。

算法的复杂性分析主要是由决策表中的属性组合引起的。设 $card(C)=N$，对于上述算法，如果忽略对象数对计算时间的影响，在最坏的情况下，当核为空集时，每选择一次属性后就要重新计算一次互信息，每次所考虑的属性数依次为 $N,N-1,\cdots,1$，则总的次数为 $N+(N-1)+\cdots+1=N(N+1)/2$。故该算法能在 $O(N^2)$的时间内找到满意的约简。

3.4 实例分析

利用文献[12]中的 CTR(Car Test Results)数据库作为测试的决策表，如表 1 所示。其中，$U=\{u_1,u_2,\cdots,u_{21}\}$；条件属性集 $C=\{x_1,x_2,\cdots,x_9\}$；决策属性 $D=\{y\}$。

表 1 测试决策表

U	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
u_1	c	6	y	E	m	h	h	a	m	m
u_2	c	6	n	E	m	m	h	ma	m	m
u_3	c	6	n	E	m	h	h	ma	m	m
u_4	c	4	y	E	m	h	h	ma	l	h
u_5	c	6	n	E	m	m	m	ma	m	m
u_6	c	6	n	B	m	m	m	a	he	lo
u_7	c	6	n	E	m	h	h	ma	he	lo
u_8	s	4	n	B	sm	h	lo	ma	l	h
u_9	c	4	n	B	sm	h	lo	ma	m	m
u_{10}	c	4	n	B	sm	h	m	a	m	m
u_{11}	s	4	n	E	sm	h	lo	ma	l	h
u_{12}	s	4	n	E	m	m	m	ma	m	h
u_{13}	c	4	n	B	m	m	m	ma	m	m
u_{14}	s	4	y	E	sm	h	h	ma	m	h
u_{15}	s	4	n	B	sm	m	lo	ma	m	h
u_{16}	c	4	y	E	m	m	h	ma	m	m
u_{17}	c	6	n	E	m	m	h	a	m	m
u_{18}	c	4	n	E	m	m	h	a	m	m
u_{19}	s	4	n	E	sm	h	m	ma	m	h
u_{20}	c	4	n	E	sm	h	m	ma	m	h
u_{21}	c	4	n	B	sm	h	m	ma	m	m

属性是：x_1——size over length；x_2——number of cylinders；x_3——presence of a turbocharger；x_4——type of fuel system；x_5——engine displacement；x_6——compression ratio；x_7—power；x_8——type of transmission；x_9——weight；y——mileage。

属性值如下：c——compact；s——subcompact；sm——small；y——yes；n——no；E——EFI；B——2-BBL；m——medium；ma——manual；h——high；he——heavy；l——light；lo——low；a——auto。

在计算约简之前，为便于处理，需要对决策表中的数据进行适当的变换，将字符型属性的取值变换为数字型，并将属性的取值离散化。处理的方式如下(为简便起见，只对决策表中非数值型的属性进行处理，不对本身取值就是离散化的属性再进行变换)：(1)c=2, s=1；(2)；(3)y=1, n=0；(4)E=1, B=0；(5)m=2, sm=1；(6)h=2, m=1；(7)lo=1, h=3, m=2；(8)a=1, ma=0；(9)m=2, l=1, he=3。

决策属性的变换：h=0, m=1, lo=2。

经过计算得：

$U/D=\{\{1, 2, 3, 5, 9, 10, 13, 16, 17, 18, 21\}, \{4, 8, 11, 12, 14, 15, 19, 20\}, \{6, 7\}\}$

$U/C=\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\}, \{11\}, \{12\}, \{13\}, \{14\}, \{15\}, \{16\}, \{17\}, \{18\}, \{19\}, \{20\}, \{21\}\}$

$pos_C(D)=\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21\}$

$\gamma_c(D)=|pos_C(D)|/|U|=1$

令 $C'=C-\{x_i\}$，经过计算得到：只有当 x_i 等于 x_4 或 x_9 时，$\gamma_{c'}(D)=19/21$；除了 x_4 或 x_9 外，其余属性的依赖度都为 $\gamma_{c'}(D)=1$。于是得到决策表的相对核为 $Core=Core_D(C)=\{x_4,x_9\}$。接下来，对该决策表计算得到 $I(C;D)=0.5143$，$I(Core;D)=0.3952$，令 $R=Core$，对每个属性 $a\in C-R$，计算条件互信息 $I(a;D|R)$，得到表 2。

表 2 条件互信息 $I(a;D|R)$的值

属性 a	x_1	x_2	x_3	x_5	x_6	x_7	x_8
$I(a;D\|R)$	0.087 2	0.053 3	0.029 6	0.053 3	0.082 1	0.072 5	0.021 0

选取使互信息最大的属性为 x_1，所以 $R=R\cup\{x_1\}=\{x_1,x_4,x_9\}$，并且新的 $I(R;D)=I(Core;D)+I(a;D|R)=0.3952+0.0874=0.4826$。重复这一个过程，直到 $I(R;D)=I(C;D)$，此时得到的属性集 $R=\{x_1,x_4,x_5,x_9\}$为该决策表的一个相对约简。

用基于可辨识矩阵属性约简的一般算法可以得到此决策系统的全部属性约简有 7 个，它们分别是：$\{x_1,x_4,x_5,x_9\}$, $\{x_1,$

(下转第 61 页)

```
                fv = false;
            i=i+1;
          }
        }
4   return fv;
```

应答范围查询时可以参照商覆盖立方体的相关算法，先缩小需搜索的记录范围，再将一个范围查询分解为多个点查询，从而得到结果。

5　实验结果与分析

在一台采用 INTEL Core 2 Duo 2.66 GHz CPU 和 DDRII 8 GB 内存的 Dell T100 服务器上，本文采用 weather 数据集[7]进行了实验，取文件 SEP85L.DAT 中的数据形成基本表，并选取不同的维属性组合，用整数对各个维值编码，生成了相应的商覆盖立方体和浓缩商覆盖立方体，实验结果如图 1 所示。实验结果说明，浓缩商覆盖立方体能在商覆盖立方体的基础上，进一步过滤部分上界格，从而只需保存较少数量的元组。另外，注意到被过滤掉的元组中不含为*的维值，被保留的元组中多含有为*的维值，而在物理存储时常用 NULL 表示数据仓库中的*值，因此，可以推知，较之于浓缩商覆盖立方体与商覆盖立方体的元组数量比例，实际数据文件的体积比例将会更小。

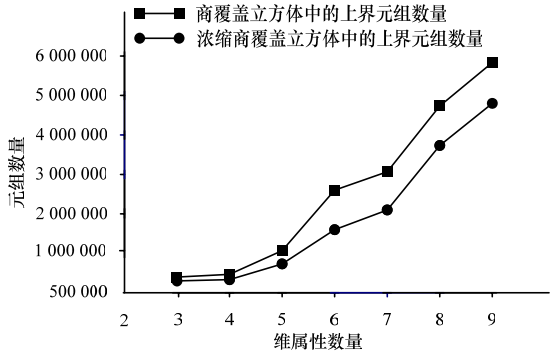

图 1　浓缩商覆盖立方体与商覆盖立方体中元组数量比较

实验结果说明，浓缩商覆盖立方体中的元组数量小于相应商覆盖立方体中的元组数量，因而数据文件的体积可以更小，如在采用 6 个维属性的情况下，浓缩商覆盖立方体中的元组数目仅为采用商覆盖立方体时的 62%。

6　结束语

本文研究了进一步压缩商覆盖立方体的方法，通过在商覆盖立方体中去除其与基本表重复存储的部分冗余数据，而缩小其体积。在下一步工作中，将研究如何将本文方法应用于其他压缩格式的数据立方体。

参考文献

[1] Jim G, Adam B, Andrew L, et al. Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-tab, and Sub-totals[C]//Proceedings of International Conference on Data Engineering. New Orleans, USA: [s. n.], 1996: 152-159.

[2] Wang Wei, Lu Hongjun, Feng Jianlin, et al. Condensed Cube: An Efficient Approach to Reducing Data Cube Size[C]//Proceedings of ICDE'02. San Jose, USA: [s. n.], 2002: 155-165.

[3] Lakshmanan L V S, Pei Jian, Zhao Yan. QC-Trees: An Efficient Summary Structure for Semantic OLAP[C]//Proceedings of SIGMOD'03. San Diego, USA: [s. n.], 2003: 64-75.

[4] 陈富强，奚建清. 一种新的商覆盖立方体算法[J]. 计算机工程与应用, 2008, 44(17): 151-152.

[5] Xi Jianqing, Chen Fuqiang, Zhang Pingjian. A New Bitmap Index and A New Data Cube Compression Technology[C]//Proceedings of International Conference on Computational Science and Its Applications. Wuhan, China: [s. n.], 2008: 1218-1228.

[6] 师智斌，黄厚宽，刘鸿敏. 一种保持语义的压缩数据立方体结构[J]. 计算机工程, 2008, 34(13): 37-39.

[7] Hahn C. Edited Synoptic Cloud Reports from Ships and Land Stations over the Globe[EB/OL]. (2003-10-20). http://cdiac.ornl.gov/ftp/ndp026b.

编辑　索书志

(上接第 58 页)

$x_2, x_4, x_6, x_9\}$, $\{x_1, x_4, x_6, x_7, x_9\}$, $\{x_2, x_3, x_4, x_6, x_8, x_9\}$, $\{x_2, x_3, x_4, x_6, x_7, x_9\}$, $\{x_1, x_2, x_3, x_4, x_7, x_9\}$, $\{x_1, x_2, x_3, x_4, x_8, x_9\}$，长度分别为 4、5、6。

因此，通过基于互信息的相对约简算法得到表 1 的决策系统的最优属性约简。

4　结束语

本文把知识看作是关于论域的划分，从而使得对知识能够进行严密的分析和处理。对粗糙集理论中的知识做新的理解，如果建立知识和信息之间的关系，从信息熵的角度考察属性约简，提出一种基于粗糙集和信息熵的高效的属性约简算法。通过理论证明和实际数据验证了该算法的有效性和可行性。该算法在实际问题中要得到更好的应用和实现，还需要更进一步的研究。

参考文献

[1] Pawlak Z. Rough Sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341-365.

[2] Pawlak Z. Rough Sets and Intelligent Data Analysis[J]. Information Sciences, 2002, 147(1-4):1-12.

[3] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. [S. l.]: Springer, 1991.

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

[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] 王加阳，高　灿. 改进的基于差别矩阵的属性约简算法[J]. 计算机工程, 2009, 35(3): 26-28.

[7] 任小康，吴尚智. 基于可辨识矩阵的属性频率约简算法[J]. 兰州大学学报, 2007, 43(1): 138-140.

[8] 吴尚智. 基于粗糙集的一种属性值约简算法及其应用[J]. 计算机应用与软件, 2009, 26(2): 263-265.

[9] Wang Jue, Miao Duoqian. Analysis on Attribute Reduction Strategies of Rough Set[J]. Journal of Computer Science and Technology, 1998, 13(2): 189-193.

[10] Guan J W, Bell D A. Matrix Computation for Information Systems[J]. Information Sciences, 2001, 131(14): 129-156.

[11] 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.

[12] Wong S K M, Ziarko W. On Optimal Decision Rules in Decision Tables[J]. Bulletin of Polish Academy of Science, 1985, 33(11/12): 693- 696.

编辑　顾逸斐