属性约简准则与约简信息损失的研究

邓大勇[1,2,3],薛欢欢[1],苗夺谦[3],卢克文[1]
(1. 浙江师范大学数理与信息工程学院,浙江金华 321004;2. 浙江师范大学行知学院,浙江金华 321004;
3. 同济大学电子与信息工程学院,上海 201804)

摘 要: 属性约简是粗糙集的重要研究内容,信息熵是度量信息量的方法.在研究绝对约简和几种相对约简的基础上,归纳出属性约简的一般准则.定义了基于条件属性信息熵的属性约简和基于联合熵的属性约简,研究了几种属性约简与绝对约简之间的关系.定义了基于条件属性信息熵的约简信息损失,澄清了属性约简不损失信息的含糊观念,指出了属性约简只是在约简准则意义下不损失信息,在信息熵意义下可能损失信息.为进一步研究粗糙集、粒计算中属性约简与分类夯实了信息论基础.

关键词: 粗糙集;属性约简;信息熵;联合熵;信息损失
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2017)02-0401-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.02.019

Study on Criteria of Attribute Reduction and Information Loss of Attribute Reduction

DENG Da-yong[1,2,3], XUE Huan-huan[1], MIAO Duo-qian[3], LU Ke-wen[1]
(1. *College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
2. *Xingzhi College, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
3. *School of Electronics and Information Engineering, Tongji University, Shanghai* 201804, *China*)

Abstract: Attribute reduction is one of important topics in rough set theory, and information entropy is an index of measuring the amount of information. After investigating absolute attribute reduct and several kinds of relatively attribute reducts, a general criterion of reducts is induced in rough set theory. With this criterion of reducts, attribute reduct based on information entropy and attribute reduct based on joint entropy are defined. The relationships among attribute reducts and absolute attribute reduct are investigated. Moreover, information loss based on information entropy for attribute reducts is defined, which can measure information loss after attribute reduction has been conducted. The old concepts that attribute reduction can not lose information are improved, and attribute reduction and classification can be further investigated from information loss and information entropy.

Key words: rough sets; attribute reduction; information entropy; joint entropy; information loss

1 引言

粒计算[1,2]是人类智能处理问题的思维方式,也是处理不确定性问题的方法.粒计算的主要方法有模糊集[3]、粗糙集[4~7]、商空间[8]和云模型[9]等.粗糙集理论[4~7]是一种处理不精确、不完全、含糊数据的有效数学工具,是数据挖掘和分类的重要方法.

粗糙集最重要的应用在于不确定性分析和属性约简.研究者们提出了上、下近似[4~7]、隶属度[10]、信息熵[11]、条件熵[12,13]、粗糙熵、模糊熵[14,15]等不确定性度量来刻画和描述数据的不确定性,其中很大一部分不确定性指标被用来作为条件属性约简的准则,例如:条件熵、互信息等,由此衍生出了各种各样关于属性约简的研究[11~23].几乎所有的条件属性约简都笼统地宣称保持分类不变或信息不变,人们也非常认同这个观点.但是条件属性约简是否真的保持信息不变、信息无损

收稿日期:2016-03-21;修改日期:2016-05-06;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61572442, No. 61203247, No. 61273304, No. 61573259, No. 61472166);浙江省自然科学基金(No. LY15F020012);浙江省自然科学青年基金(No. Q13F020006)

失？能够作为属性约简准则的指标满足什么条件？其他形式的信息熵是否可以作为属性约简的准则？这些问题对粗糙集、粒计算，乃至数据挖掘、人工智能来说，都非常重要.

对于条件属性约简的信息损失，长期以来存在的误区和盲点，以及属性约简的一些本源问题，本文结合粒计算、粗糙集、信息论的观点，以绝对约简、基于正区域的相对约简、基于属性依赖度的相对约简、基于互信息的相对约简、基于条件熵的相对约简为例，归纳出属性约简准则所满足的条件. 定义了基于条件属性信息熵的属性约简和基于联合熵的属性约简，并分析其性质，证明了基于条件属性信息熵的属性约简等价于绝对约简以及在一致的决策表中基于联合熵的属性约简等价于绝对约简. 以条件属性信息熵为信息量的度量指标定义了属性约简的信息损失，指出了各种类型的属性约简仅仅不存在该约简准则下的信息损失，但是可能存在条件属性信息熵意义下的信息损失，从而澄清了人们长期以来存在的误区和盲点.

2 基础知识

本节简单介绍粗糙集[4~6]与信息熵[11~15,24]的相关基本知识.

2.1 粗糙集

信息系统 $IS=(U,A)$ 中，U 是论域，A 是论域 U 上的条件属性集. 对于任意条件属性 $a\in A$ 都存在函数 a：$U\to V_\alpha$，V_α 为属性 a 的值域. U 中每个元素称为个体、对象或行.

对于任意 $B\subseteq A$ 和任何 $x\in U$ 都对应着如下的信息函数：

$$Inf_B(x)=\{(a,a(x)):a\in B\}.$$

B - 不分明关系(或称为不可区分关系)定义为

$$IND(B)=\{(x,y):Inf_B(x)=Inf_B(y)\}.$$

任何满足 $IND(B)$ 的 2 个元素 x,y 都不能由 B 的任何子集区分，$[x]_B$ 表示由 x 引导的 $IND(B)$ 等价类.

信息系统 $IS=(U,A)$ 中，$B\subseteq A$，$X\subseteq U$. 上、下近似与边界区域的个体表示为：

$$\overline{B}(X)=\overline{B}(IS,X)=\{x\in U:[x]_B\cap X\neq\varnothing\},$$

$$\underline{B}(X)=\underline{B}(IS,X)=\{x\in U:[x]_B\subseteq X\},$$

$$BN(X)=\overline{B}(IS,X)-\underline{B}(IS,X).$$

上、下近似及边界区域的信息粒表示为：

$$\overline{B}(X)=\overline{B}(IS,X)=\cup\{[x]_B\subseteq U:[x]_B\cap X\neq\varnothing\},$$

$$\underline{B}(X)=\underline{B}(IS,X)=\cup\{[x]_B\subseteq U:[x]_B\subseteq X\},$$

$$BN(X)=\overline{B}(IS,X)-\underline{B}(IS,X).$$

在决策系统 $DS=(U,A,d)$ 中，$\{d\}\cap A=\varnothing$，决策属性 d 把论域 U 划分为块，$U/\{d\}=\{Y_1,Y_2,\cdots Y_M\}$，其中 $Y_i(i=1,2,\cdots M)$ 是等价类. 决策系统 $DS=(U,A,d)$ 的正区域定义为

$$POS_A(d)=\bigcup_{Y_i\in U/\{d\}}\underline{A}(Y_i).$$

定义 1 在决策系统 $DS=(U,A,d)$ 中，称决策属性 d 以程度 $h(0\leqslant h\leqslant 1)$ 依赖条件属性集 A，其中，

$$h=\gamma(DS,A,\{d\})=\frac{|POS_A(d)|}{|U|},$$

符号 $|\cdot|$ 表示集合的势.

在决策系统 $DS=(U,A,d)$ 中，$\partial(x)=\{(d,d(y)):y\in[x]_A\wedge x\in U\}$. 若对于任意的 $x\in u$ 都有 $|\partial(x)|=1$，则决策系统 $DS=(U,A,d)$ 称为一致的，否则称为不一致.

2.2 信息熵

给定一个决策系统 $DS=(U,A,d)$，设 A 和 $\{d\}$ 在论域 U 上导出的划分分别为 X 和 Y，其中 $X=U/A=\{X_1,X_2,\cdots X_N\}$，$Y=U/\{d\}=\{Y_1,Y_2,\cdots Y_M\}$，$p(X_i)=\frac{|X_i|}{|U|}$，$p(Y_j)=\frac{|Y_j|}{|U|}$，$p(X_i,Y_j)=\frac{|X_i\cap Y_j|}{|U|}$，$p(Y_j|X_i)=\frac{|X_i\cap Y_j|}{|X_i|}$，$i=1,2,\cdots N$，$j=1,2,\cdots M$. A 和 $\{d\}$ 的信息熵分别定义为：

$$H(DS,A)=-\sum_{i=1}^{N}p(X_i)lbp(X_i),$$

$$H(DS,\{d\})=-\sum_{j=1}^{M}p(Y_j)lbp(Y_j).$$

$\{d\}$ 相对于 A 的条件熵定义为：

$$H(DS,\{d\}|A)=-\sum_{i=1}^{N}p(X_i)\sum_{j=1}^{M}p(Y_j|X_i)lbp(Y_j|X_i).$$

$\{d\}$ 相对于 A 的互信息定义为：

$$I(DS,\{d\}|A)=H(DS,\{d\})-H(DS,\{d\}|A).$$

$\{d\}$ 与 A 的联合熵定义为：

$$H(DS,A,\{d\})=-\sum_{i=1}^{N}\sum_{j=1}^{M}p(X_i,Y_j)lbp(X_i,Y_j).$$

3 属性约简

本节我们讨论粗糙集理论中的绝对约简与 4 种相对约简.

绝对约简定义如下：

定义 2[6,25,26] 给定信息系统 $IS=(U,A)$，$B\subseteq A$ 称为 DS 的绝对约简，iff $B\subseteq A$ 满足下列条件：

(1)对于任意的 $x\in U$，都有 $[x]_A=[x]_B$；

(2)对于任意的 $S\subset B$，存在 $x\in U$ 使得 $[x]_S\neq[x]_A$.

基于正区域的相对约简定义为：

定义 3[4~7] 给定决策系统 $DS=(U,A,d)$，$B\subseteq A$ 是 DS 的基于正区域的相对约简，iff $B\subseteq A$ 满足下列

条件:

(1)$POS_B(d)=POS_A(d)$;

(2)对于任意 $S\subset B$,都有 $POS_S(d)\neq POS_A(d)$.

定义 2,定义 3 从结构角度分别定义了绝对约简和基于正区域的相对约简. 定义 3 还可以用基于属性依赖度的相对约简等价地表示为:

定义 4[4~7]　给定一个决策系统 $DS=(U,A,d)$,$B\subseteq A$ 称作决策系统 DS 的一个基于属性依赖度的相对约简,iff $B\subseteq A$ 满足如下两个条件:

(1)$\gamma(DS,B,\{d\})=\gamma(DS,A,\{d\})$;

(2)对任意的 $S\subset B$,均有 $\gamma(DS,A,\{d\})\neq\gamma(DS,S,\{d\})$.

基于互信息的相对约简定义为:

定义 5[11]　给定一个决策系统 $DS=(U,A,d)$,$B\subseteq A$ 称为决策系统 DS 的一个基于互信息的相对约简,iff $B\subseteq A$ 满足下列条件:

(1)$I(DS,\{d\}\mid B)=I(DS,\{d\}\mid A)$;

(2)对任意 $S\subset B$,都有 $I(DS,\{d\}\mid S)\neq I(DS,\{d\}\mid A)$.

基于条件熵的相对约简定义为:

定义 6[12]　给定一个决策系统 $DS=(U,A,d)$,$B\subseteq A$ 称为决策系统 DS 的一个基于条件熵的相对约简,iff $B\subseteq A$ 满足下列条件:

(1)$H(DS,\{d\}\mid A)=H(DS,\{d\}\mid B)$;

(2)对任意 $S\subset B$,都有 $H(DS,\{d\}\mid S)\neq H(DS,\{d\}\mid A)$.

之所以正区域、属性依赖度等能成为属性约简的准则,是因为它们具有如下性质[4~7,10~13]:

(1)给定一个信息系统 $IS=(U,A)$,$B_1\subset B_2\subseteq A$,则对于任意的 $x\in U$,都有$[x]_{B_2}\subseteq[x]_{B_1}$.

(2)给定一个决策系统 $DS=(U,A,d)$,$B_1\subset B_2\subseteq A$,则有 $POS_{B_1}(d)\subseteq POS_{B_2}(d)$.

(3)给定一个决策系统 $DS=(U,A,d)$,$B_1\subset B_2\subseteq A$,则有 $\gamma(DS,B_1,\{d\})\leqslant\gamma(DS,B_2,\{d\})$.

(4)给定一个决策系统 $DS=(U,A,d)$,$B_1\subset B_2\subseteq A$,则有 $I(DS,\{d\}\mid B_1)\leqslant I(DS,\{d\}\mid B_2)$.

(5)给定一个决策系统 $DS=(U,A,d)$,$B_1\subset B_2\subseteq A$,则有,$H(DS,\{d\}\mid B_2)\leqslant H(DS,\{d\}\mid B_1)$.

可见,无论是等价类、正区域、属性依赖度,还是信息熵和条件熵,它们相对于条件属性来说都具有单调性. 也就是说,对于条件属性来说,某种度量只要满足一定的单调性,就可以成为条件属性的约简准则. 即:

定律 1　在决策系统 $DS=(U,A,d)$ 中定义的函数 $F(B)$($B\subseteq A$)只要满足单调不增或单调不减,则 $F(B)$ 可成为条件属性约简准则.

4　基于条件属性信息熵和联合熵的属性约简

对于信息熵而言,互信息和条件熵已经成为条件属性的约简准则. 很容易证明,基于互信息的相对约简也是基于条件熵的相对约简,反之亦然.

那么,条件属性的信息熵和联合熵是否可以成为条件属性的约简准则呢?

定理 1　给定一个决策系统 $DS=(U,A,d)$,$B_1\subset B_2\subseteq A$,则有 $H(DS,B_1)\leqslant H(DS,B_2)$.

证明:对于 $U/B_1=\{X_1,X_2\cdots X_N\}$ 中的每一块,它要么等于 $U/B_2=\{X_1',X_2'\cdots X_M'\}$ 中的某一块,要么是某几块的并. 不失一般性,有 $\frac{|(X_{l,1}'\cup X_{l,2}'\cup\cdots X_{l,n}')|}{|U|}=\frac{|X_k|}{|U|}$,其中 $X_k=X_{l,1}'\cup X_{l,2}'\cup\cdots X_{l,n}'$,根据函数 $H(X)=-p(X)lbp(X)$ 的性质,易得 $H(DS,B_1)\leqslant H(DS,B_2)$. 证毕.

定理 2　给定一个决策系统 $DS=(U,A,d)$,$B_1\subset B_2\subseteq A$,则有 $H(DS,B_1,\{d\})\leqslant H(DS,B_2,\{d\})$.

证明:在决策系统 $DS=(U,A,d)$ 中,如果把决策属性 d 当成条件属性,则联合熵变成条件属性的信息熵. 根据定理 1,当 $B_1\subset B_2\subseteq A$ 时有 $H(DS,B_1,\{d\})\leqslant H(DS,B_2,\{d\})$. 证毕.

根据定律 1 和定理 1,可以定义基于条件属性信息熵的条件属性约简如下:

定义 7　给定一个决策系统 $DS=(U,A,d)$,$B\subseteq A$ 称为决策系统 DS 的一个基于条件属性信息熵的属性约简,iff $B\subseteq A$ 满足下列条件:

(1)$H(DS,B)=H(DS,A)$;

(2)对任意 $S\subset B$,都有 $H(DS,S)\neq H(DS,A)$.

根据定律 1 和定理 2,可以定义基于联合熵的条件属性约简如下:

定义 8　给定一个决策系统 $DS=(U,A,d)$,$B\subseteq A$ 称为决策系统 DS 的一个基于联合熵的属性约简,iff $B\subseteq A$ 满足下列条件:

(1)$H(DS,B,\{d\})=H(DS,A,\{d\})$;

(2)对任意 $S\subset B$,都有 $H(DS,S,\{d\})\neq H(DS,A,\{d\})$.

定理 3　给定一个决策系统 $DS=(U,A,d)$,$B\subseteq A$ 是 DS 的一个基于条件属性信息熵的属性约简,iff $B\subseteq A$ 是信息系统 $IS=(U,A)$ 的绝对约简.

证明:(1)⇐:显然成立.

(2)⇒:反设结论不成立,即存在一个 $S\subset B$,使得对于任意的 $x\in U$,$[x]_S=[x]_A$,即 $S\subset B$ 是信息系统 $IS=(U,A)$ 的绝对约简,根据证明(1)有 $H(DS,S)=H(DS,A)$,与 $B\subseteq A$ 是 DS 的一个基于条件属性信息熵的属性约简矛盾.

综合(1)、(2),定理 3 成立. 证毕.

定理 4 在决策系统 $DS=(U,A,d)$ 中，$B\subseteq A$ 是信息系统 $IS=(U,A)$ 的绝对约简，则有 $H(DS,B,\{d\})=H(DS,A,\{d\})$.

证明：根据相关定义，立得. 证毕.

注：绝对约简不一定是基于联合熵属性约简的超集，主要是因为基于联合熵属性约简中决策属性也参与了条件属性的分类，而绝对约简中决策属性并没有参与条件属性的分类.

定理 5 在一致的决策系统 $DS=(U,A,d)$ 中，$B\subseteq A$ 是 DS 的一个基于联合熵的属性约简，iff $B\subseteq A$ 是信息系统 $IS=(U,A)$ 的绝对约简.

证明：在一致的决策系统 $DS=(U,A,d)$ 中，$U/A=U/(A\cup\{d\})$，从而有 $H(DS,A)=H(DS,A,\{d\})$，再根据定理 3，得定理 5 成立. 证毕.

根据定理 5，在一致的决策系统中，基于联合熵的属性约简既是相对约简又是绝对约简.

根据文献[25,26]的相关结论以及定理 3、定理 4 有：

定理 6 给定一个决策系统 $DS=(U,A,d)$，$B\subseteq A$ 是信息系统 $IS=(U,A)$ 的绝对约简，则：

(1) $POS_B(d)=POS_A(d)$；

(2) $\gamma(DS,B,\{d\})=\gamma(DS,A,\{d\})$；

(3) $H(DS,B)=H(DS,A)$；

(4) $I(DS,\{d\}|B)=I(DS,\{d\}|A)$；

(5) $H(DS,\{d\}|B)=H(DS,\{d\}|A)$；

(6) $H(DS,B,\{d\})=H(DS,A,\{d\})$.

定理 6 显示，信息系统 $IS=(U,A)$ 的绝对约简（或基于条件属性信息熵的属性约简）是基于正区域的相对约简、基于属性依赖度的相对约简、基于互信息相对约简的超集，同样它也是其它形式相对约简的超集，但不一定是基于联合熵属性约简的超集. 于是，

命题 1 在决策系统 $DS=(U,A,d)$ 中，绝对约简是所有相对约简的超集.

定理 7 在决策系统 $DS=(U,A,d)$ 中，若 $B_1\subseteq A$ 是基于联合熵的属性约简，$B_2\subseteq A$ 是绝对约简，则有 $H(DS,B_1)\leqslant H(DS,B_2)$.

证明：因为 $B_1\subseteq A$，根据定理 1，有 $H(DS,B_1)\leqslant H(DS,A)$. 再根据定理 3，有 $H(DS,B_2)=H(DS,A)$. 所以 $H(DS,B_1)\leqslant H(DS,B_2)$. 证毕.

虽然绝对约简不一定是基于联合熵属性约简的超集，但是绝对约简的信息熵不小于基于联合熵属性约简的信息熵.

5 属性约简的信息损失

虽然各种形式的条件属性约简都宣称能保持信息不变，即没有信息损失. 但是它们往往只能保证约简准则意义下的信息不变，比如，基于属性依赖度的相对约简，仅仅是保证约简前后的属性依赖度不变，并不能保证其他约简准则下的信息不变.

在属性约简的过程中，变化的是条件属性，而决策属性不发生变化. 根据定理 1－定理 6 以及命题 1，我们用条件属性的信息熵作为决策系统中信息量的度量指标，则属性约简的信息损失定义如下：

定义 9 设 $DS=(U,A,d)$ 是一个决策系统，$B\subseteq A$ 是它的一个约简，则 $B\subseteq A$ 约简的信息损失定义如下：

$$\Delta(B)=H(DS,A)-H(DS,B);$$

$B\subseteq A$ 约简的信息损失率定义如下：

$$s(B)=\frac{\Delta(B)}{H(DS,A)}\times 100\%=1-\frac{H(DS,B)}{H(DS,A)}\times 100\%.$$

于是，我们有如下命题：

命题 2 决策系统 $DS=(U,A,d)$ 中绝对约简（或基于条件属性信息熵的约简）的信息损失为 0.

证明：根据绝对约简的定义、定理 3 及定义 9 立得.

命题 3 在一致的决策系统 $DS=(U,A,d)$ 中基于联合熵属性约简 $B\subseteq A$ 的信息损失为 0.

证明：根据定理 5，$B\subseteq A$ 是绝对约简，再根据命题 2，立得.

引理 1[25,26] 决策系统 $DS=(U,A,d)$ 中基于互信息的相对约简是相应基于属性依赖度相对约简的超集.

命题 4 决策系统 $DS=(U,A,d)$ 中基于互信息（或条件熵）的相对约简的信息损失不大于相应的基于属性依赖度相对约简的信息损失.

证明：设 $B\subseteq A$ 是决策系统 $DS=(U,A,d)$ 的基于互信息的相对约简，根据引理 1，存在 $B_0\subseteq B\subseteq A$ 使得 B_0 是决策系统 $DS=(U,A,d)$ 的基于属性依赖度的相对约简，根据定理 1 有 $H(DS,B_0)\leqslant H(DS,B)$，从而有 $\Delta(B_0)\geqslant\Delta(B)$，证毕.

命题 5 在一致的决策系统 $DS=(U,A,d)$ 中基于互信息（或条件熵）的相对约简的信息损失等于相应的基于属性依赖度相对约简（或基于正区域的相对约简）的信息损失.

证明：根据文献[25,26]中相关定理，在一致的决策系统 $DS=(U,A,d)$ 中基于互信息（或条件熵）的相对约简等于相应的基于属性依赖度相对约简（或基于正区域的相对约简），所以它们的信息损失相等. 证毕.

下面我们通过两个例题来计算属性约简及其信息损失，其中第一个例题是一个一致决策表，第二个例题是一个不一致决策表.

例 1 $DS_1=(U,A,d)$ 是一个决策系统，如表 1 所示. U 为论域，$A=\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$ 是条件属性，d 是决策属性.

表 1　决策系统 $DS_1=(U,A,d)$

U	α_1	α_2	α_3	α_4	d
e_1	0	1	0	1	0
e_2	1	1	0	0	1
e_3	1	1	0	1	1
e_4	0	1	1	1	0
e_5	0	2	1	0	1
e_6	1	2	0	1	0

$$U/A=\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5\},\{e_6\}\},$$
$$U/\{d\}=\{\{e_1,e_4,e_6\},\{e_2,e_3,e_5\}\}$$
$$H(DS_1,A)=-\sum p(X)lbp(X)$$
$$=-6\times\frac{1}{6}\times lb\frac{1}{6}=lb6=2.585.$$

(1)绝对约简与基于联合熵的属性约简:

先求绝对约简. 容易知道,所有的条件属性都不可约,所以 A 的绝对约简为$\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$.

根据命题 2,绝对约简(或基于条件属性信息熵的约简)没有信息损失.

因为决策表 $DS_1=(U,A,d)$是一致的,根据定理 5,基于联合熵的属性约简也等于$\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$,没有信息损失.

(2)相对约简:

(a)基于正区域(或基于属性依赖度)的相对约简:

易得到,基于正区域的相对约简 $B=\{\alpha_1,\alpha_2\}$,$U/B=\{\{e_1,e_4\},\{e_2,e_3\},\{e_5\},\{e_6\}\}$,$H(DS_1,B)=-\sum p(X)lbp(X)=-2\times\frac{2}{6}lb\frac{2}{6}-2\times\frac{1}{6}lb\frac{1}{6}=1.918$.

故,B 的信息损失量为:

$$\Delta(B)=H(DS_1,A)-H(DS_1,B)$$
$$=2.585-1.918=0.667.$$

B 信息损失率为:

$$s(B)=\frac{\Delta(B)}{H(DS_1,A)}\times100\%=25.80\%.$$

(b)基于互信息(或条件熵)的约简:

决策系统 DS_1 是一个一致的决策系统,根据文献[25,26]中的相关结论,B 也是基于互信息(或条件熵)的约简,它的信息损失与信息损失率同上.

例 2　一个关于玩具销售的决策系统 $DS_2=(U,A,d)$,如表 2 所示. 其中 $A=\{a,b,c\}$. a 是“颜色”,b 是“形状”,c 是“尺寸”. $V_a=\{0,1,2\}$,0 是黄,1 是红,2 是蓝. $V_b=\{0,1\}$,0 是圆形,1 是方形. 决策属性 d 是“销售”. $V_d=\{0,1\}$,0 是不畅销,1 是畅销. $U=\{1,2,3,4,5,6\}$是 6 种玩具,具体如表 2 所示:

表 2　决策系统 $DS_2=(U,A,d)$

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

$$U/A=\{\{1,6\},\{2\},\{3\},\{4\},\{5\}\},$$
$$U/\{d\}=\{\{1,2,3\},\{4,5,6\}\},$$
$$U/(A\cup\{d\})=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}\neq U/A.$$

因此,该决策表 DS_2 是非一致的决策表.

决策表 DS_2 的条件属性信息熵为:

$$H(DS_2,A)=-\sum p(X)lbp(X)$$
$$=-\frac{2}{6}lb\frac{2}{6}-4\times\frac{1}{6}lb\frac{1}{6}=2.252.$$

(1)绝对约简:

易得绝对约简为$\{a,b\},\{a,c\}$. 绝对约简的信 息损失为 0.

(2)基于联合熵的属性约简:

容易得到基于联合熵的属性约简为$\{a,b\},\{a,c\}$和 $B_0=\{b,c\}$. 基于联合熵的约简并不一定是绝对约简的子集. 前 2 个基于联合熵的属性约简信息损失与信息损失率都为 0.

B_0 的信息熵为:

$$H(DS_2,B_0)=-\sum p(X)lbp(X)$$
$$=-(2\times\frac{2}{6}lb\frac{2}{6}+2\times\frac{1}{6}lb\frac{1}{6})$$
$$=1.918.$$

B_0 的信息损失为:

$$\Delta(B_0)=H(DS_2,A)-H(DS_2,B_0)$$
$$=2.252-1.918=0.334.$$

B_0 的信息损失率为:

$$s(B_0)=\frac{\Delta(B_0)}{H(DS_2,A)}\times100\%=14.83\%.$$

(3)相对约简:

容易知道,$\{a,b\},\{a,c\}$既是决策系统 DS_2 基于正区域的属性约简,也是基于条件熵的属性约简,同时还是绝对约简,所以,信息损失与信息损失率都为 0.

6　结论与展望

在研究绝对约简和几种相对约简的基础上,本文归纳出粗糙集属性约简的一般准则. 运用这个准则,定

义了基于条件属性信息熵的属性约简和基于联合熵的属性约简.基于条件属性信息熵的属性约简等价于绝对约简.根据属性约简的性质和信息熵的性质,定义了属性约简的信息损失,澄清了属性约简不损失信息的含糊观念,为粗糙集、粒计算中属性约简和分类研究夯实了信息论基础.

进一步研究为其他类型约简的信息损失、属性约简信息损失的性质、属性约简的信息损失与分类准确率之间的关系.

参考文献

[1] Hobbs J R. Granularity[A]. Proc of the Ninth In-ternational Joint Conference on Artificial Intelligece[C]. Los Angeles,1985. 432 – 435.

[2] Lin T Y. Granular Computing[M]. Announcementof the BASIC Special Interest Group on Granular Computing,1997.

[3] Zadel L A. Fuzzy sets[J]. Information and Control,1965,8(3):338 – 353.

[4] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982,11(5):341 – 356.

[5] Pawlak Z. Rough Sets —Theoretical Aspect of Reasoning about Data[M]. Dordrecht :Kluwer Academic Publishers,1991.

[6] 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.
Wang Guoyin. Rough Set Theory and Knowledge Acquisition[M]. Xi' an: Xi' an Jiaotong University Press, 2001. (in Chinese)

[7] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation :An accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence,2010,174:597 – 618.

[8] 张钹,张铃. 问题求解理论及应用[M]. 北京:清华大学出版社,1990.
Zhang Bo, Zhang Ling. Theories and Applications for Problem Solving[M]. Beijing: Tsinghua University Press, 1990. (in Chinese).

[9] 李德毅,孟海军,史雪梅. 隶属云和隶属云发生器[J]. 计算机研究与发展,1995,32(6):16 – 18.
Li Deyi, Meng Haijun, Shi Xuemei. Membership clouds and Membership cloud generators[J]. Journal of Computer Research and Development, 1995, 32(6): 16 – 18. (in Chinese)

[10] Pawlak Z,. Skowron A. Rough membership functions [A]. Adances in the Dempster Shafer Theory of Evidence[M]. John Wiley, New York, 1994. 251 – 271.

[11] 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(6):681 – 684.
Miao Duoqian, Hu Guirong. A heuristic algorithm for reduction of knowledg[J]. Journal of Computer Research and Development, 1999, 36(6): 681 – 684. (in Chinese)

[12] 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.
Wang Guoyin, Yu Hong, Yang Dachun. Decision table reduction on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759 – 766. (in Chinese)

[13] 杨明. 决策表中基于条件信息熵的近似约简[J]. 电子学报,2007,35(11):2156 – 2160.
Yang Ming. Approximate reduction based on conditional information entropy in decision tables[J]. Acta Eletronica Sinica, 2007, 35(11): 2156 – 2160. (in Chinese)

[14] Liang J Y, Chin K. S, Dang C. Y. A new method for measuring uncertainty and fuzziness in rough set theory [J]. International Journal of General Systems, 2002, 31(4):331 – 342.

[15] 梁吉业,李德玉. 信息系统中的不确定性与知识获取[M]. 北京:科学出版社,2005.
Liang Jiye, Li Deyu. Uncertainty and Knowledge Acquisition in Information Systems[M]. Beijing: Science Press, 2005. (in Chinese)

[16] 张文修,魏玲,祁建军. 概念格的属性约简理论与方法[J]. 中国科学 E 辑 信息科学,2005,35(6):628 – 639.
Zhang Wenxiu, Wei Ling, Qi Jianjun. Theory and method of attribute reduction for concept lattices[J]. Science in China Ser. E Information Sciences, 2005, 35(6): 628 – 639. (in Chinese)

[17] 胡清华,于达仁,谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报,2008,19(3):640 – 649.
Hu Qinghua, Yu Daren, Xie Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008, 19(3): 640 – 649. (in Chinese)

[18] Jia Xiuyi, Li Weiwei, Shang Lin, et al. An optimization view-point on decision-theoretic rough set model[A]. LNCS6954:Proceedings of the 6th International Conference on Rough Sets and Knowledge Technology(RSKT11)Banff, Canada, Oct9-12, 2011[C]. Berlin, Heidelberg: Springer-Verlag, 2011. 457 – 465.

[19] 滕书华. 基于粗糙集理论的不确定性度量和属性约简方法研究[D]. 国防科学技术大学,2010.
Teng Shuhua. Study on methods for uncertainty measure and attribute reduction based on rough set theory[D]. National University of Defense Technology, 2010. (in Chinese)

[20] 张文修,仇国芳,吴伟志. 粗糙集属性约简的一般理论[J]. 中国科学 E 辑 信息科学,2005,35(12):1304 – 1313.
Zhang Wenxiu, Chou Guofang, Wu Weizhi. Generalized

theory of attribute reduction in rough sets[J]. Science in China Ser. E Information Sciences,2005,35(12):1304 – 1313. (in Chinese)

[21] 徐久成,孙林 . 一种新的基于决策熵的决策表约简方法[J]. 重庆邮电大学学报(自然科学版) ,2009,21(4):479 – 483.
Xu Jiuchen,Sun Lin. New reduction method based on decision information entropy in decision table[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Editon),2009,21(4):479 – 483. (in Chinese)

[22] 周杰. 概率粗糙集模型的知识获取算法研究[D]. 同济大学,2011.
Zhou Jie. Research on knowledge acquisition algorithms in probabilistic rough set models[D]. Tongji University, 2011. (in Chinese)

[23] 高灿. 基于粗糙集理论的机器学习方法研究[D]. 同济大学,2013.
Gao Can. Rough sets ~ based machine learning research [D]. Tongji University,2013. (in Chinese)

[24] 叶中行. 信息论基础[M]. 高等教育出版社,2007.
Ye Zhongxing. Information Theory Foundation[M]. Higher Education Press,2007. (in Chinese)

[25] 邓大勇,黄厚宽,李向军. 不一致决策系统中约简之间的比较[J]. 电子学报,2007,35(2):252 – 255.
Deng Dayong, Huang Houkuan, Li Xiangjun. Comparison of various types of reductions in inconsistent system[J]. Acta Eletronica Sinica,2007,35(2):252 – 255. (in Chinese)

[26] 邓大勇. 基于粗糙集的数据约简及粗糙集扩展模型的研究[D]. 北京交通大学,2007.
Deng Dayong. Research on data reduction based on rough sets and extension of rough set models[D]. Beijing Jiaotong University,2007. (in Chinese)

作者简介

邓大勇(通信作者)　男,1968 年出生,副教授,博士,现为浙江师范大学行知学院教师,主要研究方向为粗糙集、粒计算、数据挖掘等.
E-mail:dayongd@163. com

薛欢欢　女,1990 年出生于河南商丘. 现为浙江师范大学数理信息工程学院计算机科学与技术专业硕士研究生. 主要研究方向为粗糙集、数据挖掘.
E-mail:1530043379@qq. com

苗夺谦　男,1964 年出生,教授,博士,博士生导师,现为同济大学电信工程学院教师,主要研究方向为粗糙集、粒计算、数据挖掘、计算智能、图像处理等.
E-mail:dqmiao@tongji. edu. cn

卢克文　男,1992 年出生于安徽明光. 现为浙江师范大学数理与信息工程学院计算机科学与技术专业硕士研究生. 主要研究方向为粗糙集、数据挖掘.
E-mail:709882771@qq. com