基于信息观点的约简算法比较

滕书华 孙即祥 周石琳 李智勇
(国防科学技术大学电子科学与工程学院 长沙 410073)

摘 要 在相关文献的基础上,给出了三种不同条件熵的性质,进而研究了基于三种条件熵的不同搜索策略的约简的算法之间的关系,包括对一致和不一致决策表约简的执行时间、约简质量和分类性能的比较。理论分析和实验结果表明三种条件熵不同性质导致了三种条件熵的约简算法存在各自的优缺点。上述工作为用户根据实际需要选择合适的约简算法提供了有益的参考。

关键词 粗糙集,完备,约简,条件熵,不一致决策表
中图法分类号 TP18 **文献标识码** A

Comparison with Attribute Reduction Algorithms in Information View

TENG Shu-hua SUN Ji-xiang ZHOU Shi-lin LI Zhi-yong
(School of Electronical Science and Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract *Many types of conditional entropy reduction have been proposed and applied in information systems. It is thus important to clarify the relationships among the existing types of conditional entropy reduction. In this paper, the properties of three different conditional entropies were compared and analyzed, then we investigated the relations among three conditional entropy reduction based on different selection strategies, including reduction quality, run time and classification accuracy in cosistent and inconsistent decision table. The experimental result shows that three conditional entropy reducts have different advantages and disadvantages. The above work can give a valuable reference for application.*

Keywords Rough set, Complete, Reduction, Conditional entropy, Inconsistent decision table

1 引言

特征选择和属性约简是机器学习和知识发现研究的一个热点,受到了广泛关注。作为处理不精确、不相容和不完全数据的新的数学工具——粗糙集理论[1],在属性约简方面取得了大量成果[2]。已证明寻找信息系统最小约简是NP-hard问题,实际应用中往往采用启发式算法搜索最优或次优约简,如基于正区域[3-5]、基于属性频率[6]和基于信息观点[7-11]等启发式约简算法。在基于信息观点的启发式约简算法中构造了多种条件熵作为特征评价函数:文献[11]把Shannon条件熵作为评价函数设计了一种启发式约简算法;文献[12]扩展了Shannon熵,提出了一种具有补的性质的信息熵,并给出了相应的条件熵;文献[13]以此条件熵为评价函数设计了一种启发式约简算法;文献[14]把决策属性集相对条件属性集的条件信息量作为评价函数,给出了一种新的启发式约简算法。基于三种不同条件熵的约简算法都取得了较理想的约简结果,但是三种不同条件熵的约简结果是不同的,有必要澄清三种条件熵约简算法间的关系:如约简率哪个最优,分类效果哪个最好?本文在对三种条件熵性质及关系探讨的基础上,对基于不同搜索策略的三种条件熵约简算法在一致和不一致决策表上的约简结果及分类性能进行了研究比较,得到了一些有意义的结论。

2 粗糙集的基本知识

四元组 $S=(U,A,V,f)$ 是一个信息系统,其中 $U=\{u_1,u_2,\cdots,u_{|U|}\}$ 是论域,$|U|$ 表示集合 U 的势;A 是所有属性的集合;$V=\bigcup_{a_j\in A}V_j$,V_j 表示属性 $a_j\in A$ 的值域;$f:U\times A\to V$ 是信息函数,使得 $\forall a_j\in A,u_i\in U,f(u_i,a_j)\in V_j$;如果 $A=C\cup D$ 且 $C\cap D=\emptyset$,则称 $S=(U,C,D,V,f)$ 为决策表,其中 C 为条件属性集,D 为决策属性集;$S=(U,A,V,f)$ 通常简写为 $S=(U,A)$。

完备信息系统 $S=(U,A)$ 中,$Q,P\subseteq A$,有以下定义:

(1) 属性集 $P\subseteq A$ 的不可区分关系 $IND(P)$ 为:

$$IND(P)=\{(u_i,u_j)\in U\times U\mid \forall a\in P,f(u_i,a)=f(u_j,a)\}$$

显然不可区分关系是一种等价关系。$IND(P)$ 在 U 上导出的划分记为 U/P。$[u_i]_P$ 表示包含 u_i 的 P 等价类,表达式为 $[u_i]_P=\{u_j\in U\mid \forall a\in P,f(u_i,a)=f(u_j,a)\}$。

(2) 若 $Q\subseteq P$,则对 $\forall X_i\in U/P$,$\exists Y_j\in U/Q$ 使得 $X_i\subseteq Y_j$,这意味着 P 划分比 Q 精细或 Q 划分比 P 粗糙。

(3) $U/P=\omega=\{\{u_i\}\mid u_i\in U\}$,则 P 为恒等关系;若 $U/P=\delta=\{U\}$,则 P 为全域关系。

(4) 决策表 $S=(U,C,D)$ 中,如果 $U/C\subseteq U/D$,则称 S 为

到稿日期:2010-02-15 返修日期:2010-04-29 本文受自然科学基金项目(40901216)资助。
滕书华(1979-),男,博士生,主要研究方向为人工智能、数据挖掘,E-mail:tengshuhua1979@sohu.com;孙即祥(1946-),男,教授,博士生导师,主要研究方向为计算机视觉、模式识别、图像处理。

一致决策表，否则为不一致决策表。

下面给出基于信息观点的属性重要性和相对约简定义：

(5) 决策表 $S=(U,C,D)$ 中，$Q\in C$，则对任意 $q\in\{C-Q\}$ 的属性重要性为：

$$SGF(q,Q,D)=H(D|Q)-H(D|Q\cup\{q\})$$

式中，$H(D|Q)$ 是条件熵。

(6) 决策表 $S=(U,C,D)$ 中，$Q\in C$，Q 为 C 的 D 约简当且仅当 $\forall q\in Q, H(D|Q)\neq H(D|Q-\{q\})$ 且 $H(D|Q)=H(D|C)$。

3 三种条件熵的定义与性质

完备信息系统 $S=(U,A)$ 中，$Q,P\subseteq A$，令 $U/P=\{X_1,X_2,\cdots,X_m\}$，$U/Q=\{Y_1,Y_2,\cdots,Y_n\}$，下面给出三种条件熵的相关定义。

定义 1[11] 知识 P 对应的信息熵 $H(P)$ 为：

$$H(P)=-\sum_{i=1}^{m}p(X_i)\log_2 p(X_i) \tag{1}$$

知识 Q 关于知识 P 的 H 条件熵为：

$$H(Q|P)=-\sum_{i=1}^{m}p(X_i)\sum_{j=1}^{n}p(Y_j|X_i)\log_2 p(Y_j|X_i) \tag{2}$$

式中，$p(X_i)=\frac{|X_i|}{|U|}$，$p(Y_j|X_i)=\frac{|Y_j\cap X_i|}{|X_i|}$。

定义 2[12] 知识 P 对应的信息熵 $E(P)$ 为：

$$E(P)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\left(1-\frac{|X_i|}{|U|}\right) \tag{3}$$

知识 Q 关于知识 P 的 E 条件熵为：

$$E(Q|P)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|Y_j^c-X_i^c|}{|U|}$$

式中，X_i^c 表示 X_i 的补集，即 $X_i^c=U-X_i$。E 条件熵表达式 $E(Q|P)$ 也可以写为：

$$E(Q|P)=\sum_{i=1}^{m}p(X_i)^2\sum_{j=1}^{n}p(Y_j|X_i)[1-p(Y_j|X_i)] \tag{4}$$

定义 3[14] 知识 Q 关于知识 P 的 I 条件熵为：

$$I(Q|P)=\sum_{i=1}^{m}p(X_i)\sum_{j=1}^{n}p(Y_j|X_i)[1-p(Y_j|X_i)] \tag{5}$$

从以上定义可以看出，式(1)和式(2)中的对数表达式 $\log_2 x$ 若替换为 $x-1$，则可分别得到式(3)和式(5)，而式(5)乘以 $p(X_i)$ 则可得式(4)，因此三种条件熵中 I 条件熵计算最简单，H 条件熵(需计算对数)次之，E 条件熵计算相对复杂。又由 $\log_2 x\leqslant x-1$，且 $p(X_i)\leqslant 1$，从而可得三种条件熵的大小关系如下。

定理 1 $H(Q|P)\geqslant I(Q|P)\geqslant E(Q|P)$

定理 1 给出了三种条件熵的大小排序，则需指出的是，如果 $P_1,P_2\subseteq A$，且 $H(Q|P_1)\geqslant H(Q|P_2)$，则 $I(Q|P_1)\geqslant I(Q|P_2)$ 或 $E(Q|P_1)\geqslant E(Q|P_2)$ 并不一定成立。同样，如果 $I(Q|P_1)\geqslant I(Q|P_2)$，则 $E(Q|P_1)\geqslant E(Q|P_2)$ 或 $H(Q|P_1)\geqslant H(Q|P_2)$ 也不一定成立。举例说明如下：

例 1 令 $U/Q=\{Y_1,Y_2,Y_3\}$，$\{|Y_1|,|Y_2|,|Y_3|\}=\{13,10,9\}$，$U/P_1=\{X_1',X_2',X_3'\}$，$\{|X_1'|,|X_2'|,|X_3'|\}=\{26,3,3\}$，$U/P_2=\{X_1,X_2,X_3\}$，$\{|X_1|,|X_2|,|X_3|\}=\{16,13,3\}$，$|Y_1\cap X_1'|=13$，$|Y_2\cap X_1'|=4$，$|Y_3\cap X_1'|=9$，$|Y_1\cap X_2'|=0$，$|Y_2\cap X_2'|=3$，$|Y_3\cap X_2'|=0$，$|Y_1\cap X_3'|=0$，$|Y_2\cap X_3'|=3$，$|Y_3\cap X_3'|=0$，$|Y_1\cap X_1|=6$，$|Y_2\cap X_1|=9$，$|Y_3\cap X_1|=1$，$|Y_1\cap X_2|=5$，$|Y_2\cap X_2|=0$，$|Y_3\cap X_2|=8$，$|Y_1\cap X_3|=2$，$|Y_2\cap X_3|=1$，$|Y_3\cap X_3|=0$，则

$$H(Q|P_1)=\frac{-26}{32}\times\left(\frac{13}{26}\times\log_2\frac{13}{26}+\frac{9}{26}\times\log_2\frac{9}{26}+\frac{4}{26}\times\log_2\frac{4}{26}\right)=1.17$$

$$E(Q|P_1)=\frac{26^2}{32^2}\times\left(\frac{13}{26}\times\frac{13}{26}+\frac{9}{26}\times\frac{17}{26}+\frac{4}{26}\times\frac{22}{26}\right)=0.40$$

$$I(Q|P_1)=\frac{26}{32}\times\left(\frac{13}{26}\times\frac{13}{26}+\frac{9}{26}\times\frac{17}{26}+\frac{4}{26}\times\frac{22}{26}\right)=0.49$$

同理 $H(Q|P_2)=1.10$，$E(Q|P_2)=0.22$，$I(Q|P_2)=0.50$。显然有 $H(Q|P_1)>I(Q|P_1)>E(Q|P_1)$，$H(Q|P_2)>I(Q|P_2)>E(Q|P_2)$。尽管 $H(Q|P_1)>H(Q|P_2)$，但 $I(Q|P_1)<I(Q|P_2)$，而 $E(Q|P_1)>E(Q|P_2)$。

以上这种关系使得由三种条件熵评估得出的属性重要度排序是不同的，从而导致基于三种条件熵的前向添加约简算法的约简结果存在较大差别，参看 4.4.1 节实验部分的表 4。

决策表 $S=(U,C,D)$ 中，$Q\subseteq C$，令 $U/C=\{X_1,X_2,\cdots,X_m\}$，$U/D=\{Y_1,Y_2,\cdots,Y_n\}$，下面给出三种条件熵的相关性质。

定理 2 决策表 $S=(U,C,D)$ 是一致的，当且仅当 $H(D|C)$(或 $E(D|C)$ 或 $I(D|C)$)等于 0。

证明：令 $U/C=\{X_1,X_2,\cdots,X_m\}$，$U/D=\{Y_1,Y_2,\cdots,Y_n\}$。在一致决策表中 $U/C\subseteq U/D$，即 $|X_i\cap Y_j|/|X_i|$ 等于 1 或 0，因此 $H(D|C)=E(D|C)=I(D|C)=0$；

反之，如果 $H(D|C)$，$E(D|C)$ 或 $I(D|C)$ 中有一个等于 0，由三种条件熵的表达式可知对 $\forall i\in\{1,2,\cdots,m\}$，$\forall j\in\{1,2,\cdots,m\}$ 都有 $|X_i\cap Y_j|/|X_i|$ 等于 1 或 0，即 $X_i\subseteq Y_j$ 或 $X_i\cap Y_j=\varnothing$，因此 $U/C\subseteq U/D$，决策表为一致决策表。

定理 2 说明在判断决策表是否一致上三种条件熵是等价的。由定理 2 可得如下推论。

推论 1 决策表 $S=(U,C,D)$ 中，有如下等价关系：$H(D|C)=0\Leftrightarrow E(D|C)=0\Leftrightarrow I(D|C)=0$。

推论 2 一致决策表 $S=(U,C,D)$ 中，对 $\forall C_i\in C$，存在如下等价关系：$H(D|C)=H(D|C-C_i)\Leftrightarrow E(D|C)=E(D|C-C_i)\Leftrightarrow I(D|C)=I(D|C-C_i)$。

推论 1 和推论 2 给出了三种条件熵之间的两种等价关系。

下面给出三种条件熵的单调性质。

性质 1(单调性) 决策表 $S=(U,C,D)$，若 $Q\subseteq C$ 且 $U/Q=\{X_1,X_2,\cdots,X_{k-1},X_{k+1},\cdots,X_{l-1},X_{l+1},\cdots,X_m\}$ 是将划分 U/C 中的等价类 X_k 和 X_l 合并为 $X_k\cup X_l$ 后产生的新划分，则

(1)[11] $H(D|C)\leqslant H(D|Q)$，等号成立的条件为 $\forall j\in\{1,2,\cdots,n\}$，$\frac{|X_k\cap Y_j|}{|X_k|}=\frac{|X_l\cap Y_j|}{|X_l|}$；

(2) $E(D|C)\leqslant E(D|Q)$，等号成立的条件为 $\frac{\{X_k\cup X_l\}}{D}=\{X_k\cup X_l\}$；

(3)[14] $I(D|C)\leqslant I(D|Q)$，等号成立的条件为 $\forall j\in\{1,2,\cdots,n\}$，$\frac{|X_k\cap Y_j|}{|X_k|}=\frac{|X_l\cap Y_j|}{|X_l|}$。

证明：下面我们仅给出(2)的证明，(1)和(3)的证明可分别参看文献[11,14]。

$$\begin{aligned}E(D|Q)-E(D|C)&=\frac{1}{|U|^2}\sum_{j=1}^{n}[(|Y_j\cap X_k|+|Y_j\cap X_l|)\times(|Y_j^c\cap X_k|+|Y_j^c\cap X_l|)-|Y_j\cap X_k|\times|Y_j^c\cap X_k|-|Y_j\cap X_l|\times|Y_j^c\cap X_l|]\end{aligned}$$

$$=\frac{1}{|U|^2}\sum_{j=1}^{n}(|Y_j\cap X_k|\times|Y_j^c\cap X_l|+|Y_j\cap X_l|\times|Y_j^c\cap X_k|)\geqslant 0$$

当且仅当 X_k 和 X_l 同时属于 Y_j 时上式等于 0，即 $\frac{\{X_k\cup X_l\}}{D}=\{X_k\cup X_l\}$ 时上式等号成立。

性质 1 说明如果将决策表条件属性集的分类进行合并，将可能导致三种条件信息量的增加，只有在发生合并的两个分类对于决策类的隶属度(概率)均相等的情况下，才不导致 H 条件信息量和 I 条件信息量的变化，而对于 E 条件熵，只有在发生合并的两个分类都属于同一个决策类的条件下才不导致 E 条件信息量的变化。不难看出，

$$\{X_k\cup X_l\}/D=\{X_k\cup X_l\}\Rightarrow\frac{|X_k\cap Y_j|}{|X_k|}=\frac{|X_l\cap Y_j|}{|X_l|}$$

即 E 条件熵取等条件比 H 条件熵和 I 条件熵更为严格，而 H 条件熵和 I 条件熵取等条件相同。从而我们得到如下定理。

定理 3 决策表 $S=(U,C,D)$ 中，$Q\subseteq C$，则

(1) $H(D|C)=H(D|Q)\Leftrightarrow I(D|C)=I(D|Q)$；

(2) $E(D|C)=E(D|Q)\Rightarrow H(D|C)=H(D|Q)$；

(3) $E(D|C)=E(D|Q)\Rightarrow I(D|C)=I(D|Q)$。

定理 3 的证明由性质 1 可得。定理 3 对一致和不一致决策表都适用，但(2)和(3)的逆对于不一致决策表不成立。由性质 1 还可得以下推论。

推论 3 决策表 $S=(U,C,D)$ 中，如果 $a_i\subset C,i\in\{1,2,\cdots,|C|\}$，则有

(1) $H(D|\{a_1\})\geqslant H(D|\{a_1\}\cup\{a_2\})\geqslant\cdots\geqslant H(D|\{a_1\}\cup\{a_2\}\cup\cdots\cup\{a_{|C|}\})$；

(2) $E(D|\{a_1\})\geqslant E(D|\{a_1\}\cup\{a_2\})\geqslant\cdots\geqslant E(D|\{a_1\}\cup\{a_2\}\cup\cdots\cup\{a_{|C|}\})$；

(3) $I(D|\{a_1\})\geqslant I(D|\{a_1\}\cup\{a_2\})\geqslant\cdots\geqslant I(D|\{a_1\}\cup\{a_2\}\cup\cdots\cup\{a_{|C|}\})$。

推论 3 说明决策属性集相对条件属性集的三种条件信息量的变化规律呈现非严格单调性。

4 基于三种条件熵的启发式约简算法比较

设计启发式约简算法需要解决两个关键问题：属性质量评价指标和最优子集搜索策略。本节根据信息观点的属性约简定义，分别对一致和不一致决策表中基于三种不同条件熵(属性质量评价指标)的前向添加和后向删除策略(搜索策略)的启发式约简算法的约简结果及其分类性能进行比较。

4.1 启发式约简算法步骤

在启发式约简算法中，不同的搜索策略将会产生不同的约简结果。前向搜索算法能够确保重要的属性首先被加入到约简中，从而不损失重要的特征；后向删除算法却难以保证这个结果，最终可能保留大量区分能力很弱而不是少量区分能力很强的特征[15]。下面给出基于 H 条件熵的前向添加和后向删除策略的约简算法步骤：

输入：给定决策表 $S=(U,C,D)$；

输出：该决策表的一个相对约简 Q。

(1) 基于 H 条件熵前向添加约简算法 H-FARCE：

Step1 令 $Q=\emptyset,T=C$，求 $H(D|C)$；

Step2 计算 $SGF(a_k,Q,D)=\max\limits_{a_i\in T}SGF(a_i,Q,D),1\leqslant i\leqslant|T|$，若有多个属性都达到最大值，则从中选取一个与 Q 组和数最少的属性作为 a_k。令 $Q=Q\cup\{a_k\},T=T-\{a_k\}$；

Step3 若 $H(D|Q)=H(D|C)$，转 Step4，否则转 Step2；

Step4 Q 即为所求约简。

(2) 基于 H 条件熵的后向删除约简算法 H-BACE：

Step1 令 $Q=C=\{C_1,C_2,\cdots,C_m\},i=m$，求 $H(D|C)$；

Step2 计算 $H(D|Q-C_i)$，如果 $H(D|Q-C_i)=H(D|C)$，令 $Q=Q-C_i$；

Step3 若 $i>0$，令 $i=i-1$，转 Step2，否则 Q 即为所求约简。

算法 H-FARCE 和 H-BACE 中的 H 条件熵若换为 E 条件熵或 I 条件熵，则可分别得到 E-FARCE，E-BACE 和 I-FARCE，I-BACE 算法。

4.2 三种条件熵的后向删除约简算法间的关系

由三种条件熵性质间的关系可得基于三种条件熵后向删除约简算法之间的关系定理如下。

定理 4 算法 H-BACE，E-BACE 和 I-BACE 对一致决策表 $S=(U,C,D)$ 的约简结果相同。

定理 4 的证明由定理 2、推论 2 和条件熵后向删除约简算法步骤可得。定理 4 给出了一致决策表中三种条件熵的后向删除算法约简结果之间的关系。在不一致决策表中，有以下关系定理。

定理 5 对于不一致决策表 $S=(U,C,D)$，算法 E-BACE 的约简结果包含 H-BACE 和 I-BACE 的约简结果，而算法 H-BACE 和 I-BACE 约简结果相同。

定理 5 的证明由定理 3 和条件熵后向删除约简算法步骤可得。

定理 4 和定理 5 给出了基于三种条件熵的后向删除约简算法对一致和不一致决策表约简结果之间的关系，但对于三种条件熵的前向添加约简算法的约简结果之间并不存在这种关系，见实验分析。

4.3 算法复杂度分析

由式(2)、式(4)和式(5)可知计算三种条件熵的时间复杂度相同，当 $U/D=U/C=\omega$ 时有最差时间复杂度为 $O(|U|^2)$。由约简算法的步骤可知，三种条件熵的前向添加算法最差时间复杂度为 $O(|C|^2|U|^2)$，后向删除约简算法最差时间复杂度为 $O(C|U|^2)$。

4.4 试验结果与分析

本文选用表 1[16]和表 2 两个不一致决策表(其中 a,b,c,e,f 为条件属性，d 为决策属性)以及 UCI 机器学习数据库(http://www.ics.uci.edu/~mlearn/MLRe-pository.html)中的 5 个具有离散属性的一致决策表实例(见表 3)在 WindowsXP，CPU2.4GHz，RAM-512MB，Matlab 上进行了编程实验。

表 1 Data1

U	a	b	c	e	f	d
1	0	0	0	0	1	0
2	0	1	1	1	0	1
3	1	1	0	1	1	1
4	0	1	1	1	0	0
5	0	0	1	0	1	0
6	1	1	0	1	0	1
7	0	1	1	1	1	1
8	1	1	1	0	1	1
9	1	1	0	1	1	0
10	0	1	1	1	1	0

表 2 Data2

U	a	b	c	d
1	3	3	3	3
2	1	3	1	3
3	3	3	1	3
4	3	1	3	3
5	3	1	3	1
6	2	3	3	3
7	2	1	3	3
8	2	1	3	1
9	2	1	1	1

4.4.1 约简结果比较

表3、表4和表5分别给出了三种条件熵的后向删除和前向添加约简算法的约简结果，其中 n 为对象数，m 为约简前条件属性数，r 和 t 为三种约简算法对应的约简后的条件属性数和约简时间(单位为秒)。表中加黑结果表示最快的约简算法所用的时间，带下划线的约简结果表示三种约简算法中该约简个数最少。从表3可以看出基于三种条件熵的后向删除算法约简结果有以下特点：

(1) 对于一致决策表，三种约简算法的约简结果完全相同(因此表中只列出了约简后属性个数)，验证了定理4的正确性；

(2) 对于不一致决策表，*H*-BACE，*E*-BACE 和 *I*-BACE 算法对 Data1 的约简结果相同，而对 Data2 的约简结果为：[2,3]，[2,1,3]和[2,3]，显然两个不一致决策表的 *E*-BACE 约简结果包含 *H*-BACE 和 *I*-BACE 的约简结果，而 *H*-BACE 和 *I*-BACE 的约简结果相同，验证了定理5的正确性；

(3) 比较三种算法约简时间可知，尽管三种条件熵的后向删除约简算法时间复杂度相同，但三种条件熵表达式中，*I* 条件熵计算最为简单，*H* 条件熵其次，*E* 条件熵计算相对复杂，从而使得 *I*-BACE 约简速度最快，*H*-BACE 其次，最慢的是 *E*-BACE，但对于大数据集，*E*-BACE 约简速度快于 *H*-BACE，如数据 Chess，因此使得约简平均时间 *E*-BACE 要快于 *H*-BACE 算法。这是因为 *H* 条件熵需计算对数。

表3 基于三种条件熵的后向删除策略约简结果比较

data	n	m	H-BACE		E-BACE		I-BACE	
			r	t	r	t	r	t
Data1	10	5	4	0.0166	4	0.0216	4	**0.0161**
Data2	9	3	2	0.0102	3	0.0110	2	**0.0101**
Lung Cancer	32	56	6	1.2926	6	1.3053	6	**1.2853**
Zoo	101	16	6	0.7696	6	0.7751	6	**0.7689**
Large Soybean	307	35	10	24.3538	10	24.5007	10	**24.3179**
Voting records	435	16	10	2.3655	10	2.3833	10	**2.3642**
Chess	3196	36	29	91.9399	29	90.0632	29	**89.3195**
Average			10.83	17.2497	10.83	17.006	10.83	**16.8689**

表4和表5给出了三种条件熵前向添加约简算法的约简结果，表5中用属性的排序号代表属性，可看出三种约简算法的约简结果有以下特点：

(1) 对于不一致决策表，三种条件熵约简结果之间不存在确定的包含关系(如表5中的 Data1 和 Data2)。值得注意的是文献[16]通过对 Data1 的约简发现基于代数观点和 *H* 条件熵的约简算法均不能得到最小约简([1,4,5])，因此指出这两种方法均不能很好地描述属性的重要性。在表5中 *I*-FARCE 和 *H*-FARCE 算法均没得到最小约简，但 *E*-FARCE 算法得到了最小约简；

(2)对于一致决策表，三种条件熵的约简结果基本不同，即使约简结果相同，但属性重要性排序也不一样，如表5所列，与例1的分析一致；

(3) 从表4看出，对于一致决策表，在搜索最小或次优约简方面，*E*-FARCE 算法最优，*H*-FARCE 算法次之，*I*-FARCE 算法约简结果属性个数最多；

(4) 表4中，三种约简算法的执行时间与约简结果属性个数的多少有关，约简结果属性越少，算法执行越快(如 Data1 和 Voting records)；但在约简结果属性个数相同时，*I*-FARCE 算法最快，*E*-FARCE 算法最慢(如 Zoo)，甚至在约简属性个数较多时，*I*-FARCE 算法仍快于 *E*-FARCE 算法(如 Large Soybean 和 Chess)。三种条件熵的前向添加策略的约简算法(见表4)明显慢于后向删除策略的约简算法(见表3)，体现了两种算法复杂度的不同。

表4 基于三种条件熵的前向添加策略约简结果比较

data	n	m	H-FARCE		E-FARCE		I-FARCE	
			r	t	r	t	r	t
Data1	10	5	4	0.0262	3	**0.0251**	4	0.0261
Data2	9	3	2	0.0118	3	0.0223	2	**0.0117**
Lung Cancer	32	56	4	**1.4058**	4	1.5612	5	1.8337
Zoo	101	16	5	1.3627	5	1.3793	5	**1.3619**
Large Soybean	307	35	10	106.4591	9	106.2489	10	**102.5915**
Voting records	435	16	9	**5.5454**	10	7.1252	11	7.0450
Chess	3196	36	30	534.9759	29	599.9648	30	**533.0899**
Average			9.14	92.8271	9.00	102.3324	9.57	**92.2820**

表5 基于三种条件熵的前向添加策略约简结果

data	n	m	H-FARCE	E-FARCE	I-FARCE
Data1	10	5	[2,4,1,5]	[1,4,5]	[2,4,1,5]
Data2	9	3	[2,3]	[2,1,3]	[2,3]
Lung Cancer	32	56	[20,6,3,53]	[6,3,12,15]	[40,6,20,4,53]
Zoo	101	16	[13,4,6,8,3]	[13,3,6,8,4]	[13,4,6,8,3]
Large Soybean	307	35	[29,15,1,22,7,6,4,8,9,16]	[1,7,10,22,15,6,4,2,16]	[29,22,1,15,4,7,6,8,9,16]
Voting records	435	16	[4,11,3,13,16,2,1,15,9]	[4,11,3,12,16,13,2,1,15,9]	[4,11,9,3,7,16,15,1,10,2,13]
Chess	3192	36	C-H	C-E	C-I

C-H=[21,10,33,32,6,35,15,1,34,7,16,23,17,4,2,30,5,27,3,9,20,25,31,12,13,24,18,28,26,36]

C-E=[33,10,35,21,6,15,24,18,9,13,5,11,26,23,34,1,36,16,30,20,17,7,27,4,25,31,3,12,28]

C-I=[21,33,10,32,35,6,15,1,34,7,16,4,30,23,17,2,5,27,3,9,20,25,31,28,24,18,12,13,26,36]

4.4.2 约简结果分类性能比较

数据约简程度是比较不同约简算法的一个指标，但对于分类问题而言，更重要的是选择的属性不能显著降低分类能力。一个优秀约简算法要在保持或提高分类能力的基础上，尽可能减少分类建模所需的属性数。为了检验选中属性的分类能力，对表5中5个 UCI 约简后的数据在选中的属性子空间中用 CART 和 RBF 支持向量机分别建立分类模型，基于10折交叉验证方法得到分类精度随三种条件熵前向添加约简算法选中的属性个数变化的情况，图1中(1)－(10)分别为用 CART 算法和 RBF-SVM 算法对5个约简数据集的分类结果。从图1可以看出：

(1) 除了图1(5)外，三种约简算法的约简结果整体分类精度均高于或等于原数据分类精度，且除了图1(8)外，算法 *I*-FARCE 约简结果整体分类精度最高，而算法 *E*-FARCE 约简结果整体分类精度最低；

(2) 随着约简算法选中属性数量的增加，分类精度首先提高，达到一个峰值后将会下降或保持不变。这表明峰值之后选中的特征并没有带来分类性能的改进，反而恶化了分类。因此三种条件熵约简算法均不同程度存在属性选择的过拟合

现象；

(3)从表4可知，*I*-FARCE算法选中的属性最多，但由图1中的(1)、(2)、(5)－(10)可知，*I*-FARCE算法约简结果的最后一个属性并没有使分类率增加，反而保持或降低了分类率；

(4) 总体上，*I*-FARCE算法的分类效果最好，*H*-FARCE算法次之，*E*-FARCE算法分类效果较差。

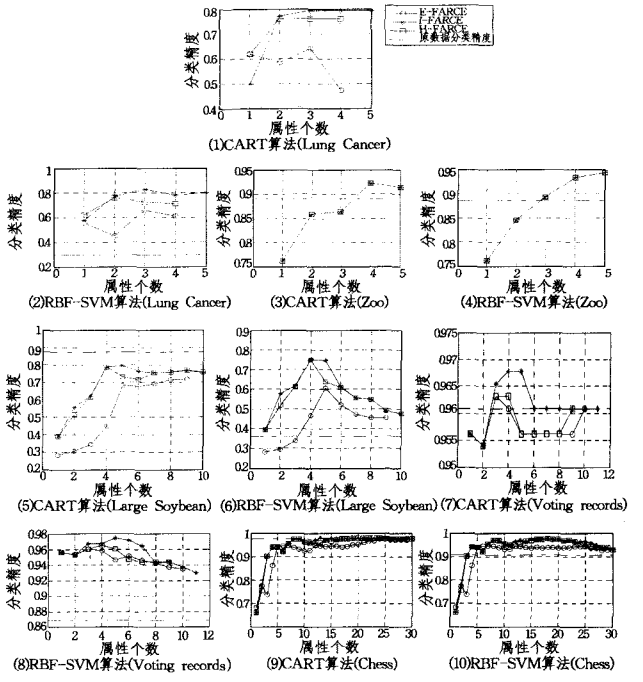

图1 分类精度与选中特征个数之间的关系

结束语 属性约简是粗糙集理论的核心内容。人们为解决求约简的NP问题，提出了各种启发式算法。本文对现有的基于三种条件熵的启发式约简算法进行了比较研究，并通过实验分析得出了以下结论：

(1) 基于三种条件熵的后向删除约简算法对一致决策表的约简结果相同，对于不一致决策表，*E*-BACE的约简结果包含*H*-BACE和*I*-BACE的约简结果，而*H*-BACE和*I*-BACE的约简结果相同；

(2) 基于三种条件熵的前向添加约简算法对一致和不一致决策表的约简结果基本不同，即使约简结果相同，但属性重要性排序也不一样；

(3) 对于一致决策表，在搜索最小或次优约简方面，*E*-FARCE算法最优，*H*-FARCE算法次之，*I*-FARCE算法约简结果属性个数最多；

(4) 在算法复杂度上，基于三种条件熵的后向删除约简算法复杂度为$O(C|U|^2)$，优于基于三种条件熵的前向添加约简算法复杂度$O(|C|^2|U|^2)$；

(5) 在对一致决策表的分类性能上，*I*-FARCE算法的分类效果最好，*H*-FARCE算法次之，*E*-FARCE算法分类效果较差，且三种约简算法均不同程度地存在属性选择的过拟合现象。

以上结论是对现有信息观点约简算法的有益补充，掌握这些算法的优缺点，对用户根据实际需要选择恰当的约简算法具有重要的参考价值和指导意义。实际应用中，不完备信息系统广泛存在[17]，因此构建基于三种条件熵的不完备信息系统的约简算法将在另文中给出。

参 考 文 献

[1] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. London: Kluwer Academic Publisher, 1991

[2] Thangavel K, Pethalakshmi A. Dimensionality reduction based on rough set theory: A review[J]. Appl Soft Comput, 2009, 9(1): 1-12

[3] 吴子特，叶东毅. 一种可伸缩的快速属性约简算法[J]. 模式识别与人工智能，2009，22(2)：234-239

[4] 刘少辉，盛秋戬，吴斌，等. Rough集高效算法的研究[J]. 计算机学报，2003，26(5)：524-529

[5] 徐章艳，刘作鹏，杨炳儒，等. 一个复杂度为max(O(|C||U|)，O(|C|²|U/C|))的快速属性约简算法[J]. 计算机学报，2006，29(3)：391-399

[6] Wang Jue, Wang Ju. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method[J]. J of Computer Science and Technology, 2001, 16(6): 489-504

[7] 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

[8] Hu Q H, Yu D R, Xie Z X. Information-preserving hybrid data reduction based on fuzzy-rough techniques[J]. Pattern Recognit Lett, 2006, 27(5): 414-423

[9] Li Fei, Yin Yunqiang. Approaches to knowledge reduction of covering decision systems based on information theory [J]. Information Sciences, 2009, 179(11): 1694-1704

[10] 滕书华，魏荣华，孙即祥，等. 基于不可区分度的启发式快速完备约简算法[J]. 计算机科学，2009，36(8)：196-200

[11] 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

[12] Liang J Y, Chin K S, Dang C Y, et al. A new method for measuring uncertainty and fuzziness in rough set theory[J]. Int J Gen Syst, 2002, 31(4): 331-342

[13] 祁立，刘玉树. 基于条件信息量的快速粗集约简算法[J]. 北京理工大学学报，2007，27(7)：604-608

[14] 刘振华，刘三阳，王珏. 基于信息量的一种属性约简算法[J]. 西安电子科技大学学报，2003，30(6)：835-838

[15] 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

[16] 刘启和，李凡，闵帆，等. 一种基于新的条件信息熵的高效知识约简算法[J]. 控制与决策，2005，20(8)：878-882

[17] 滕书华，周石琳，孙即祥，等. 基于条件熵的不完备信息系统属性约简算法[J]. 国防科技大学学报，2010，32(1)：90-94