文章编号: 1001- 2486(2010) 01- 0090- 05

# 基于条件熵的不完备信息系统属性约简算法*

滕书华, 周石琳, 孙即祥, 李智勇
(国防科技大学 电子科学与工程学院, 湖南 长沙 410073)

摘 要: 在相容关系下定义了三种不完备条件熵——$H'$ 条件熵、$E'$ 条件熵和 $I'$ 条件熵, 并对它们的性质进行了分析比较, 研究发现, $H'$ 条件熵和 $I'$ 条件熵不适用于相容关系下信息观点的约简。利用 $E'$ 条件熵刻画信息系统中属性的相对重要性, 设计了一种新的基于信息论观点的启发式约简算法, 它统一了完备信息系统与非完备信息系统中的约简方法。通过实例说明, 该算法能得到决策表的相对约简。

关键词: 粗糙集; 不完备信息系统; 属性约简; 条件熵
中图分类号: TP181 文献标识码: A

# Attribute Reduction Algorithm Based on Conditional Entropy under Incomplete Information System

TENG Shu-hua, ZHOU Shi-lin, SUN Ji-xiang, Li Zhi-yong
(College of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China)

**Abstract**: Knowledge reduction is an important issue in data mining. This paper focuses on the problem of attribute reduction in incomplete decision tables. Three types of incomplete conditional entropy are introduced based on tolerance relation, such as $H'$ conditional entropy, $E'$ conditional entropy, and $I'$ conditional entropy, which are proved to be an extension of the concept of conditional entropy in incomplete decision tables. Compared with $H'$ and $I'$ conditional entropy, $E'$ conditional entropy decreases monotonously with the amount of attributes. Based on $E'$ conditional entropy, a new reduced definition is presented, which integrates the complete and incomplete information systems into the corresponding reduced algorithm. Finally, the experimental result shows that this algorithm can find the reduct of decision tables.

**Key words**: rough set; incomplete information system; attribute reduction; conditional entropy

属性约简是粗糙集理论核心, 但已证明寻找信息系统最小属性约简是 NP-hard 问题。实际应用中往往采用启发式约简算法搜索最优或次优约简。目前有基于代数观点[1] 和基于信息观点[2-5] 等启发式约简算法。在基于信息量的约简算法中, 定义了多种条件熵来度量信息系统的协调程度: 王国胤等[5] 对粗糙集理论的代数观点和信息观点进行比较, 并把 Shannon 条件熵(本文中称为 $H$ 条件熵) 作为启发信息设计了一种启发式约简算法; 梁吉业等[6] 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(本文称之为 $E$ 条件熵), 文献[7] 以此条件熵为启发知识设计了一种启发式约简算法; 文献[8] 提出了条件信息量(本文称之为 $I$ 条件熵) 的概念, 并据此给出了一种新的启发式约简算法。

以上基于三种不同条件熵的启发式约简算法都是针对完备信息系统的, 但现实中信息的不完备(对象的属性值缺损) 现象广泛存在, 对不完备信息系统的研究受到广泛的关注[9-11]。特别是有时我们并不知道要处理的信息系统是否完备, 并且在数据分析处理和约简过程中经常要把不完备信息系统转换为完备信息系统, 希望对不完备信息系统的约简算法也适合完备信息系统。因此构造完备和不完备信息系统的统一约简算法具有重要的现实意义, 但相关的启发式约简算法的研究还比较匮乏。本文首先给出三种条件熵在相容关系下的表达式, 并对其性质进行了分析, 研究发现 $E$ 条件熵的不完备形式可应用于不完备系统的信息观点约简, 而其他两种条件熵的不完备形式则并不适用于不完备系统的信息观

* 收稿日期: 2009- 06- 09
基金项目: 国家自然科学基金资助项目(40901216)
作者简介: 滕书华(1979—), 男, 博士生。

点约简。最后利用 $E$ 条件熵的不完备形式构造了一种完备和不完备信息系统的统一约简算法。

## 1 完备信息系统中的三种条件熵及其性质

起源于经典热力学的熵, 用来度量系统的无序程度。而 Shannon 熵则被广泛应用于不确定性度量。在信息系统中, 许多学者[5,7-8]引进 Shannon 熵或其变形来度量知识的信息粒度, 进而反映知识的不确定性。下面给出现有文献中基于等价关系的三种条件熵的定义:

完备信息系统 $S=(U,A)$ 中, $Q, P\subseteq A$, 令等价关系 IND($P$) 和 IND($Q$) 对应的划分为 $U/P=\{X_1, X_2, \cdots, X_m\}$ 和 $U/Q=\{Y_1, Y_2, \cdots, Y_n\}$, 则:

(1) 知识 $Q$ 关于知识 $P$ 的 $H$ 条件熵[5] 为:

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

其中, $p(X_i)=|X_i|/|U|$, $p(Y_j\mid X_i)=|Y_j\cap X_i|/|X_i|$, 符号 $|X|$ 表示集合 $X$ 的势。

(2) 知识 $Q$ 关于知识 $P$ 的 $E$ 条件熵[7] 为:

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

(3) 知识 $Q$ 关于知识 $P$ 的 $I$ 条件熵[12] 为:

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

从式(1)~(3)可看出, $H$ 条件熵中的对数表达式 $\log_2 x$ 若替换为 $x-1$, 则可得 $I$ 条件熵, 而 $I$ 条件熵表达式再乘以 $p(X_i)$, 则可得 $E$ 条件熵。因 $\log_2 x\leqslant x-1$, 且 $p(X_i)\leqslant 1$, 从而可得三种条件熵的大小关系为 $H(Q\mid P)\geqslant I(Q\mid P)\geqslant E(Q\mid P)$。此外, 三种条件熵还有以下性质:

性质 1(单调性) 完备决策表 $S=(U,C,D)$ 中, 若 $P\subseteq C$, 则有: (1)[5] $H(D\mid C)\leqslant H(D\mid P)$; (2)[7] $E(D\mid C)\leqslant E(D\mid P)$; (3)[8] $I(D\mid C)\leqslant I(D\mid P)$。

性质 1 表明三种条件熵都随条件信息粒度的变小而变小, 即在决策粒度不变的条件下, 条件信息粒度越小, 则系统协调程度越高, 因此三种条件熵都很好地度量了系统的协调程度, 这种协调程度的度量能力使得三种条件熵都能够应用于完备信息系统的属性约简[5,7-8]。

## 2 不完备条件熵及其性质

上节给出的三种条件熵都是基于等价关系的, 不适合处理不完备信息系统。梁吉业等[13]扩展了等价关系下的两种信息熵, 在相容关系下给出了两种不完备信息熵, 即

$$H'(P)=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log_2\frac{|S_P(u_i)|}{|U|},\quad E'(P)=1-\sum_{k=1}^{|U|}\frac{|S_P(u_k)|}{|U|^2}$$

并证明了相容关系下的信息熵 $H'(P)$ 和 $E'(P)$ 分别是等价关系下的信息熵 $H(P)$ 和 $E(P)$ 的扩展, 但作者并没有给出相容关系下对应的条件熵。我们基于相容关系定义三种条件熵, 并证明相容关系下的三种条件熵分别是等价关系下三种条件熵的扩展。

不完备信息系统 $S=(U,A)$ 中, $Q, P\subseteq A$, 根据相容关系下的信息熵给出相容关系下三种不完备条件熵的定义。

定义 1 知识 $Q$ 关于知识 $P$ 的 $H'$ 条件熵为:

$$H'(Q\mid P)=H'(Q\cup P)-H'(P)=-\sum_{k=1}^{|U|}\frac{1}{|U|}\log_2\frac{|S_Q(u_k)\cap S_P(u_k)|}{|S_P(u_k)|}\tag{1}$$

定义 2 知识 $Q$ 关于知识 $P$ 的 $E'$ 条件熵为:

$$E'(Q\mid P)=E'(Q\cup P)-E'(P)=\sum_{k=1}^{|U|}\frac{|S_P(u_k)|-|S_Q(u_k)\cap S_P(u_k)|}{|U|^2}\tag{2}$$

仿照等价关系下 $I$ 条件熵表达式与 $H$ 条件熵和 $E$ 条件熵表达式之间的关系, 直接给出相容关系下

不完备 $I'$ 条件熵的定义:

定义3 知识 $Q$ 关于知识 $P$ 的 $I'$ 条件熵为:

$$I'(Q|P)=\sum_{k=1}^{|U|}\frac{1}{|U|}\left[1-\frac{|S_Q(u_k)\cap S_P(u_k)|}{|S_P(u_k)|}\right] \tag{3}$$

从三种不完备条件熵的表达式可知,式(1)中的对数表达式 $\log_2 x$ 若替换为 $x-1$,则可得到式(3),而式(3)乘以 $|S_P(u_k)|/|U|$ 则可得式(2),显然三种不完备条件熵表达式间的关系和完备条件熵一样,并且也有 $H'(Q|P)\geqslant I'(Q|P)\geqslant E'(Q|P)$。下面给出三种不完备条件熵的其他性质:

定理1(极值性) (1) $0\leqslant E'(Q|P)\leqslant 1-\frac{1}{|U|}$; (2) $0\leqslant I'(D|P)\leqslant 1-\frac{1}{|U|}$; (3) $0\leqslant H'(Q|P)\leqslant \log_2|U|$。

定理1给出了三种条件熵极值,三个不等式中右等号成立的条件为:当且仅当对 $\forall x\in U$,有 $S_P(x)=U$ 且 $S_Q(x)=\{x\}$;左等号成立的条件为:当且仅当 $U/SIM(P)\subseteq U/SIM(Q)$。

定理2 如果信息系统 $S=(U,A)$ 是完备的,$Q,P\subseteq A$,则(1) $H'(Q|P)=H(Q|P)$;(2) $E'(Q|P)=E(Q|P)$;(3) $I'(Q|P)=I(Q|P)$。

证明 由于等价关系是相容关系的一种特殊情况,因此如果 $S$ 是完备信息系统,则等价关系下的等价类和相容关系下的相容类有如下关系:$S_P(u_k)=[u_k]_P$。令 $U/Q=\{Y_1,Y_2,\cdots,Y_n\}$ 和 $U/P=\{X_1,X_2,\cdots,X_m\}$,则对 $\forall u_k\in X_l, X_l\subseteq U/P$,都有 $S_P(u_k)=X_l$,对 $\forall u_k\in Y_t, Y_t\subseteq U/Q$,都有 $S_Q(u_k)=Y_t$。因此有:

$$\begin{aligned}
(1)H'(Q|P)&=-\frac{1}{|U|}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{u_k\in Y_j\cap X_i}\log_2\frac{|S_Q(u_k)\cap S_P(u_k)|}{|S_P(u_k)|}\\
&=-\frac{1}{|U|}\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_j\cap X_i|\log_2\frac{|Y_j\cap X_i|}{|X_i|}\\
&=-\sum_{i=1}^{m}p(X_i)\sum_{j=1}^{n}p(Y_j|X_i)\log_2 p(Y_j|X_i)=H(Q|P)\\
(2)E'(Q|P)&=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{u_k\in Y_j\cap X_i}\frac{|S_P(u_k)|-|S_Q(u_k)\cap S_P(u_k)|}{|U|^2}\\
&=\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_j\cap X_i|\times\frac{|X_i|-|Y_j\cap X_i|}{|U|^2}\\
&=\sum_{i=1}^{m}p(X_i)^2\sum_{j=1}^{n}p(Y_j|X_i)[1-p(Y_j|X_i)]=E(Q|P)\\
(3)I'(Q|P)&=\frac{1}{|U|}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{u_k\in Y_j\cap X_i}\frac{|S_P(u_k)|-|S_Q(u_k)\cap S_P(u_k)|}{|S_P(u_k)|}\\
&=\frac{1}{|U|}\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_j\cap X_i|\times\frac{|X_i|-|Y_j\cap X_i|}{|X_i|}\\
&=-\sum_{i=1}^{m}p(X_i)\sum_{j=1}^{n}p(Y_j|X_i)[1-p(Y_j|X_i)]=I(Q|P)
\end{aligned}$$

定理2说明三种不完备条件熵是等价关系下的三种条件熵在相容关系下的扩展,即在完备信息系统中不完备条件熵和完备条件熵是等价的。

定理3(单调性) 不完备决策表 $S=(U,C,D)$ 中,若 $P\subseteq C$,则 $E'(D|C)\leqslant E'(D|P)$,等号成立的条件为对 $\forall u_k\in U,\{S_P(u_k)-S_C(u_k)\}\subseteq S_D(u_k)$。

证明 因为 $P\subseteq C$,则对 $\forall u_k\in U$,有 $S_C(u_k)\subseteq S_P(u_k)$,且 $S_D(u_k)\cap S_C(u_k)\subseteq S_D(u_k)\cap S_P(u_k)$,即 $|S_C(u_k)|\leqslant|S_P(u_k)|$, $|S_D(u_k)\cap S_C(u_k)|\leqslant|S_D(u_k)\cap S_P(u_k)|$,所以

$$E'(D|P)-E'(D|C)=\sum_{k=1}^{|U|}\frac{|S_P(u_k)|-|S_C(u_k)|-[|S_D(u_k)\cap S_P(u_k)|-|S_D(u_k)\cap S_C(u_k)|]}{|U|^2}$$

$$= \sum_{k=1}^{|U|} \frac{|S_P(u_k)| - |S_C(u_k)| - |S_D(u_k) \cap [S_P(u_k) - S_C(u_k)]|}{|U|^2} \geqslant 0$$

当且仅当对 $\forall u_k \in U, \{S_P(u_k) - S_C(u_k)\} \subseteq S_D(u_k)$ 时取等号。

定理 3 说明 $E'$ 条件熵和 $E$ 条件熵类似, 也具备单调性, 即能够很好地反映信息系统的协调程度, 但尽管在等价关系下 $H$ 条件熵和 $I$ 条件熵都是单调的, 可是在相容关系下 $H'$ 条件熵和 $I'$ 条件熵却不具备单调性, 不能用来度量信息系统的协调程度。

## 3 基于 $E'$ 条件熵的前向添加约简算法

本节把完备信息系统中信息观点约简定义扩展到相容关系下的不完备信息系统, 构造一种基于 $E'$ 条件熵的前向添加启发式约简算法, 从而统一完备和不完备信息系统的属性约简。

### 3.1 不完备信息系统中基于信息观点的属性重要性度量及相对约简

定义 4 不完备决策表 $S=(U, C, D)$ 中, $Q \in C$, 对任意 $q \in C - Q$ 的属性重要性为 $SGF(q, Q, D) = E'(D|Q) - E'(D|Q \cup \{q\})$。

定义 4 表明属性 $q$ 关于属性集 $Q$ 的重要性由 $Q$ 中添加 $q$ 后引起的信息量的变化大小来度量, 由定理 3 可知, 向属性集 $Q$ 中不断添加属性 $q$ 得到的条件信息量 $E'(D|Q \cup \{a\})$ 的变化规律呈递减性。因此 $SGF(q, Q, D)$ 值越大, 说明在已知 $Q$ 的条件下属性 $q$ 对于决策 $D$ 越重要。

定义 5 不完备决策表 $S=(U, C, D)$ 中, $Q \in C$, $Q$ 为 $C$ 的 $D$ 约简当且仅当 $\forall q \in Q$, $E'(D|Q) \neq E'(D|Q - \{q\})$ 且 $E'(D|Q) = E'(D|C)$。

由定义 4 和定义 5 可知, 相容关系下的不完备信息系统中基于信息观点的属性重要性度量和相对约简的定义与完备信息系统中的定义[5]一致, 但由于 $E'$ 条件熵对完备和不完备信息系统都适用, 因此本文的约简算法对完备和不完备信息系统的属性约简都适用。

### 3.2 约简算法步骤

相容关系下基于 $E'$ 条件熵前向添加约简算法($E'$-FARCE) 具体步骤如下:

输入: 完备或不完备决策表 $S=(U, C, D)$; 输出: 该决策表的一个相对约简 $Q$。

Step 1 令 $Q=\Phi$, $T=C$, 求 $E'(D; C)$;

Step 2 计算 $SGF(a_k, Q, D) = \max_{a_i \in T} SGF(a_i, Q, D)$, $1 \leqslant i \leqslant |T|$, 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 $Q = Q \cup a_k$, $T = T - \{a_k\}$;

Step 3 若 $E'(D|Q) = E'(D|C)$, 则转 Step 4, 否则转 Step 2;

Step 4 最终的 $Q$ 即为所求约简。

## 4 实 验

对于完备信息系统的约简, 由于 $E'$ 条件熵和 $E$ 条件熵等价, 因此可参考文献[7]中的基于 $E$ 条件熵的完备信息系统约简算法。下面仅在不完备信息系统中验证本文算法和结论。

表 1[9] 描述了一个关于小汽车的不完备决策表, Price, Mileage, Size 和 Max-Speed 是条件属性, 分别用字母 $P$, $M$, $S$ 和 $X$ 来代替, $d$ 是决策属性。从表 1 可知, $U=\{1, 2, 3, 4, 5, 6\}$, $C=\{P, M, S, X\}$, $D=\{d\}$, $U/SIM(D)=\{S_D(1), S_D(2), \cdots, S_D(6)\}$; 条件属性集 $C$ 对论域的分类为 $U/SIM(C)=\{S_C(1), S_C(2), \cdots, S_C(6)\}$, 其中, $S_C(1)=\{1\}$, $S_C(2)=\{2, 6\}$, $S_C(3)=\{3\}$, $S_C(4)=\{4, 5\}$, $S_C(5)=\{5, 4, 6\}$, $S_C(6)=\{6, 2, 5\}$; 令 $O=\{S, X\}$, 则 $O$ 对论域的分类为 $U/SIM(O)=\{S_O(1), S_O(2), \cdots, S_O(6)\}$, 其中, $S_O(1)=S_O(2)=\{1, 2, 6\}$, $S_O(3)=\{3\}$, $S_O(4)=S_O(5)=\{4, 5, 6\}$, $S_O(6)=\{6, 1, 2, 4, 5\}$。因此, $E'(D|O) - E'(D|C) = 0$。显然, 对 $\forall u_k \in U$, $\{S_O(u_k) - S_C(u_k)\} \subseteq S_D(u_k)$, 满足定理 3 中取等条件, 所以上式等于 0, 但对于 $H'$ 条件熵和 $I'$ 条件熵, 则有 $H'(D|O) - H'(D|C) = \frac{1}{6}\log_2 \frac{5}{8} < 0$, $I'(D|O) - I'(D$

$|C)=-\frac{1}{20}<0$, 说明对于 $O\subseteq C, H'(D|O)<H'(D|C)$, $I'(D|O)<I'(D|C)$, 因此, $H'$ 条件熵和 $I'$ 条件熵不满足单调性, 验证了定理3的正确性。由于 $E'$-FARCE 是建立在条件熵单调性基础上, 因此在不完备决策表中 $H'$ 条件熵和 $I'$ 条件熵(不具备单调性)不能度量信息系统协调程度, 从而不能用于信息观点的约简。

表1 非完备汽车表

Tab. 1 Incomplete car table

| Car | Price | Mileage | Size | Max-speed | $d$ |
|---|---|---|---|---|---|
| 1 | high | low | full | low | good |
| 2 | low | – | full | low | good |
| 3 | – | – | compact | low | poor |
| 4 | high | – | full | high | good |
| 5 | – | – | full | high | excellent |
| 6 | low | high | full | – | good |

利用 $E'$-FARCE 算法求表1的约简过程如下:

Step 1 令 $Q=\phi, T=C$, 求 $E'(D|C)=0.11$;

Step 2 计算 $SGF(a_k, Q, D)=\max_{a_i\in T} SGF(a_i, Q, D)$, $SGF(M, D)=SGF(P, D)=0$, $SGF(X, D)=0.22$, $SGF(S, D)=0.28$, 选出属性 $a_k=S$ 作为属性约简的第一个属性, 令 $Q=Q\cup a_k, T=T-S=\{P, M, X\}$, 由于 $E'(D|S)>E'(D|C)$, 转 Step 2′;

Step 2′ $SGF(P, Q, D)=SGF(M, Q, D)=0, SGF(X, Q, D)=0.11$, 选出属性 $a_k=X$ 加入约简集合 $Q=Q\cup X=\{S, X\}$。因为 $E'(D|Q)=0.11=E'(D|C)$, 从而算法终止, 得到最终约简 $Q=\{S, X\}$。算法得到的结果与文献[9]用区分函数得到的结果一致。通过上面的实例分析可以看出, 算法 $E'$-FARCE 能找到不完备信息系统的最小约简。

## 5 结论

现有的粗集理论中基于信息观点的约简都是基于完备信息系统的。本文在相容关系下定义了三种不完备条件熵($H'$ 条件熵、$E'$ 条件熵和 $I'$ 条件熵), 并证明了在完备系统中三种不完备条件熵与完备条件熵等价。$E'$ 条件熵可用于不完备系统信息观点的约简, 而 $H'$ 条件熵和 $I'$ 条件熵(不具备单调性)则不能用于不完备系统约简。以 $E'$ 条件熵为启发知识提出了完备和不完备信息系统统一约简算法, 实例说明该算法能找到决策表的相对约简。

## 参考文献:


[1] 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.

[2] 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.

[3] 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.

[4] Hu Q H, Yu D R, Xie Z X. Information Preserving Hybrid Data Reduction Based on Fuzzy-rough Techniques[J]. Pattern Recognition Letter, 2006, 27(5): 414–423.

[5] 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.

[6] Liang J Y, Chin K, 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.

[7] 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.

[8] 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.

[9] Kryszkiewicz M. Rough Set Approach to Incomplete Information Systems[J]. Information Sciences: An International Journal, 1998, 112(1–4): 39–49.

[10] 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展, 2002, 39(10): 1238–1243.

[11] Liang J Y, Xu Z B. The Algorithm on Knowledge Reduction in Incomplete Information Systems[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(1): 95–103.

[12] 钱进, 叶飞跃, 孟祥萍, 等. 一种基于新的条件信息量的属性约简算法[J]. 系统工程与电子技术, 2007, 29(12): 2154–2157.

[13] Liang J, Qian Y. Information Granules and Entropy Theory in Information Systems[J]. Science in China (Series F), 2008, 51(10): 1427–1444.