基于容差关系的加权知识粗糙熵*)

邱卫根
(广东工业大学计算机学院 广州 510090)

摘 要 粗集理论为从信息论角度研究知识粗糙熵和属性约简问题提供了一种重要的途径和方法。本文提出了基于容差关系下的不完备信息系统加权的知识熵和条件熵概念,将等价关系下的粗糙熵自然地推广到不完备信息系统的容差关系情形。本文的结果为在一般二元关系下的知识获取提供了理论依据。

关键词 容差关系,等价关系,加权粗糙熵,属性约简

Weighted Rough Entropy Based on Tolerance Relation

QIU Wei-Gen
(Computer Faculty of Guangdong University of Technology, Guangzhou 510090)

Abstract The rough theory provides an important approach for the research of knowledge entropy and attribute reduction from information theory pointveiw. In this paper, the concepts of weighted rough entropy and conditional entropy of incomplete information system based on tolerance relation are presented, which is the genealization of rough entropy conept of complete information systems based on equivalence relations, and some properties are proved in the mean time. The results of this paper provide theory basis for knowledge acquisition in information systems based on general binary relation.

Keywords Tolerance relation, Equivalence relation, Rough entropy, Attribute reduction

1 引言

粗集理论认为知识是具有粒度的,因此对知识的认识也可能是不精确的,为此引入了上、下近似运算来逼近表示知识论域的概念。这种通过代数学的等价关系和集合包含关系及运算来定义、描述知识的粗糙性,存在两个方面的问题:其一,直观性差,其本质不容易被理解,其中的约简问题是一个 NP 完全问题,需要更为高效率的约简算法来研究知识粗糙性;其二,经典的 Pawlak 粗集理论是基于等价关系的,过于苛刻,现实中更多面对的是非完备信息系统,例如部分数据不真实,或属性取空值。苗夺谦[1]等人提出了知识粗糙集的信息表示方法,研究了知识粗糙性与信息熵之间的关系,对于不完备信息系统,人们提出了容差关系[2]、非对称相似关系和量化容差关系[3]、限制容差关系[4]。梁吉业[5]、黄兵[6]等人在相容关系下通过引入信息熵,建立了知识粗糙熵和粗集粗糙熵的概念。

本文从信息论角度仔细地研究了完备信息系统知识熵及其属性约简问题。通过引入权系数概念,建立了基于容差关系的非完备信息系统的知识的加权信息熵和条件熵概念,证明了该熵概念是等价关系下信息熵的推广。本文利用粗糙理论合理地解释了不完备信息系统下知识粗糙性的本质,为进一步在非完备信息系统中的知识获取提供了理论依据。

2 完备信息系统的信息熵

信息系统可以被看作四元组(U,A,V,f),或记为(U,A)。其中U是对象的集合,也称为论域,符号$|U|$表示集合U的基数;A是对象所有属性的集合,在决策分析中经常被分成条件属性子集C和决策属性子集D;V是对象属性取值域,f是$U\times A\to V$的映射;不完备的信息系统中,存在一个特殊值,即空值,记为$*$,表示取值未知,不能比较。

完备信息系统(U,A),任意属性子集$B\subseteq A$,定义U上等价关系:

$$R(B)=\{(x,y)\mid x\in U\wedge y\in U\wedge(\forall b\in B\Rightarrow f(x,b)=f(y,b))\}$$

该关系将在论域U上导出划分$U/IND(B)$,称该划分为关于U的一个知识。

定义 1 称属性集B是信息系统(U,A)的一个属性约简,当且仅当$R(B)=R(A)$,且对$\forall C\subset B, R(C)\supset R(B)$。

设$P,Q\subseteq A$,为U上的两个等价关系,在U上导出的划分(即知识)X,Y分别为:

$$X=\{X_1,X_2,\cdots,X_n\},Y=\{Y_1,Y_2,\cdots,Y_m\}$$

将知识X,Y看成U的幂集组成的σ-代数上的随机变量,其概率分布可以定义为:

$$[X;p]=\begin{bmatrix}X_1 & X_2 & \cdots & X_n\\ p(X_1) & p(X_2) & \cdots & p(X_n)\end{bmatrix},$$

$$[Y;p]=\begin{bmatrix}Y_1 & Y_2 & \cdots & Y_m\\ p(Y_1) & p(Y_2) & \cdots & p(Y_m)\end{bmatrix}$$

其中$p(X_i)=\frac{|X_i|}{|U|}, i=1,2,\cdots,n; p(Y_j)=\frac{|Y_j|}{|U|}, j=1,2,\cdots,m$。

推论 1[1] 对于完备信息系统(U,A),$P,Q\subseteq A$,X,Y分别P,Q导出的知识,则

1) $R(P)=\bigcap_{b\in P}R(\{b\})$

*)国家自然科学基金(60474072)、广东省自然科学基金(04009465)资助项目。**邱卫根** 博士,副教授,目前研究兴趣为粗糙集理论与应用、形式概念格模型和数据挖掘。

2) $P\subset Q\Rightarrow R(P)\supseteq R(Q)$

3) $P\subset Q\Rightarrow$对任意 $X_i\in X$，存在 $Y_{i1},Y_{i2},\cdots,Y_{ik}\in Y$，使得 $X=Y_{i1}\cup Y_{i2}\cup\cdots\cup Y_{ik}$。

根据信息论原理，可以用知识熵度量信源提供的平均信息量的大小，定义知识 P 的熵 $H(P)$ 和知识 Q 相对于知识 P 的条件熵 $H(Q|P)$ 分别为：

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

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

对上述知识熵和条件熵作如下简单推导，其中 $H_R(P)$ 定义为知识 P 的知识熵，用于描述 P 的不精确(粗糙)程度，显然 $H(P)+H_R(P)=\log|U|$。

$$\begin{aligned}H(Q|P)&=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log p(Y_j|X_i)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\sum_{j=1}^{m}\frac{|Y_j\cap X_i|}{|X_i|}\log\frac{|Y_j\cap X_i|}{|X_i|}\\&=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\sum_{j=1}^{m}\frac{|Y_j\cap X_i|}{|X_i|}\left(\log\frac{|Y_j\cap X_i|}{|U|}-\log\frac{|X_i|}{|U|}\right)\\&=\sum_{i=1}^{n}\frac{|X_i|}{|U|}\sum_{j=1}^{m}\frac{|Y_j\cap X_i|}{|X_i|}\log\frac{|X_i|}{|U|}-\sum_{j=1}^{n}\frac{|X_i|}{|U|}\sum_{i=1}^{m}\frac{|Y_j\cap X_i|}{|X_i|}\log\frac{|Y_j\cap X_i|}{|U|}\\&=\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_j\cap X_i|}{|U|}\log\frac{|Y_j\cap X_i|}{|U|}=H(P\cup Q)-H(P)\end{aligned}$$

$$\begin{aligned}H(P)&=-\sum_{i=1}^{n}p(X_i)\log(X_i)=\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log|U|-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\\&\log|X_i|=\log|U|-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log|X_i|=\log|U|-\\&H_R(P)\end{aligned}$$

推论 2[6] 完备信息系统 (U,A)，$P,Q\subseteq A$，(U,P) 和 (U,Q) 分别是 P,Q 导出的知识库，则

1) $H(P)$ 取最大值 $\log|U|$ 当且仅当 $X_i=\{x_i\}$，$1\leqslant i\leqslant k=|U|$；

2) $H(P)$ 取最小值 0，当且仅当 $X_1=U$；即 U 中所有元在 P 下均是不可分辨的；

3) $P\subset Q\Rightarrow H(P|Q)=0$；

4) $P\subset Q\Rightarrow H(P)\leqslant H(Q)$；

5) $H(P|Q)=0\Leftrightarrow H_R(P)=H_R(Q)$；

6) $H(P)\leqslant H(Q)\Leftrightarrow H_R(P)\geqslant H_R(Q)$。

定理 1 属性集 B 是信息系统 (U,A) 的一个属性约简，当且仅当 $H(B)=H(A)$，且对任意 $C\subset B$，$H(C)<H(B)$。

证明：假设属性集 B 是系统 (U,A) 的一个属性约简，则 $R(A)=R(B)$，且对于任何 $C\subset B$，$R(C)>R(B)$。因此由 A 导出的划分与由 B 导出的划分完全相同，即由等价关系 A、B 定义了完全相同的知识，这样 $H(A)=H(B)$。对于任何 $C\subset B$，则 $B=C\cup C_1$，$C_1\neq\phi$，$C\cap C_1=\phi$。由于 $R(C)\neq R(B)$，即 $R(C)\supset R(B)$。这表明在 C 导致的划分块中至少有一个块 X_i，在 B 导致的划分中存在 $Y_{i1},Y_{i2},\cdots,Y_{ik}$，$k>1$，使得 $X_i=Y_{i1}\cup Y_{i2}\cup\cdots\cup Y_{ik}$。由条件熵的计算公式，$H(C_1|C)>0$，而 $H(C_1|C)=H(C\cup C_1)-H(C)=H(B)-H(C)>0$，即 $H(C)<H(B)$。

另一方面，假设 $H(B)=H(A)$，且对任意 $C\subset B$，$H(C)<H(B)$。由于 $B\subset A$，又 $H(B)=H(A)$，则必有 $R(A)=R(B)$。否则 $R(A)\subset R(B)$，这将导致 $H(B)<H(A)$。对任意 $C\subset B$，如果 $R(C)>R(B)$ 不成立，则必有 $R(C)=R(B)$，将产生 $H(C)=H(B)$ 的矛盾，因此 $R(C)>R(B)$ 必定成立。

3 不完备信息系统的加权粗糙熵

由于不完备信息系统中存在属性取空值，不能建立等价关系，引入相容关系或相似关系，类之间存在重叠，划分变成覆盖。如果仍然利用分块大小来衡量知识的信息量或粗糙性，将变得不合理。不完备信息系统 (U,A)，$\forall B\subseteq A$，定义 U 上关系：

$$R(B)=\{(x,y)|x\in U\wedge y\in U\wedge(\forall b\in B\Rightarrow(f(x,b)=*\vee f(x,b)=f(y,b)))\}$$

显然上述关系具有自反性、传递性，但不对称。若 $(x,y)\in R(B)$，y 比 x 提供了更多的信息，x 的描述包含在 y 的描述里。$\forall x\in U$，定义集合 $R_B(x)$、$R_B^{-1}(x)$ 如下，一般称 $R_B(x)$ 是 x 的最小描述集或邻域，x 是 $R_B(x)$ 的特征元。

$$R_B(x)=\{y|\ y\in U\wedge(x,y)\in R(B)\}$$

$$R_B^{-1}(x)=\{y|y\in U\wedge(y,x)\in R(B)\}$$

如果将完备信息系统中的每个等价类看作是其中每个对象的邻域，则每个对象出现的次数就是其所在等价类的大小。类中所有对象是平等的，从信息论角度来看是没有区别的，包含完全等价的信息。在非完备信息系统中，等价划分变成了覆盖，覆盖块仍可以看成其对象的邻域，但每个对象所处的地位是不平等的，包含的信息是不等价的。因此对于完备信息系统，有下面的结论。

推论 3 设 (U,A) 是完备信息系统，$P,Q\subseteq A$，则

1) 集合簇 $\{R_B(x)|x\in U\}$ 构成 U 上的一个覆盖；

2) 若 $y\in R_B(x)$，则 $R_B(y)\subseteq R_B(x)$；

3) 假设属性子集 $B\subseteq A$，则 $R(B)=\bigcap_{b\in B}R(\{b\})$；

4) $P\subset Q\Leftrightarrow$对任意 $x\in U$，必有 $R_Q(x)\subseteq R_P(x)$。

定义 2 不完备信息系统 (U,A)，称属性子集 $B\subseteq A$ 是 A 的属性约简，当且仅当 $R(A)=R(B)$，且对 $\forall C\subset B$，$R(C)\supseteq R(B)$，$R(C)\neq R(B)$。

例 1 假设不完备信息系统[5] $S=(U,A,V,f)$ 如下表所示，其中 $U=\{a_1,a_2,\cdots,a_{12}\}$，$A=C\cup\{d\}$，$C=\{c_1,c_2,c_3,a_4\}$，$V_C=\{0,1,2,3\}$，$V_{\{d\}}=\{\psi,\phi\}$。

A	c_1	c_2	c_3	a_4	d
a_1	3	2	1	0	ψ
a_2	2	3	2	0	ψ
a_3	2	3	2	0	ϕ
a_4	*	2	*	1	ψ
a_5	*	2	*	1	ϕ
a_6	2	3	2	1	ϕ
a_7	3	*	*	3	ψ
a_8	*	0	0	*	ϕ
a_9	3	2	1	3	ϕ
a_{10}	1	*	*	*	ψ
a_{11}	*	2	*	*	ϕ
a_{12}	3	2	1	*	ψ

则 $R_C(a_1)=\{a_1\}$，$R_C(a_2)=R_C(a_3)=\{a_2,a_3\}$，$R_C(a_4)=R_C(a_5)=\{a_4,a_5\}$，$R_C(a_{12})=\{a_1,a_9,a_{12}\}$，$R_C(a_6)=\{a_6\}$，$R_C(a_7)=\{a_7,a_9\}$，$R_C(a_8)=\{a_8\}$，$R_C(a_9)=\{a_9\}$，$R_C(a_{10})=$

$\{a_{10}\}$，$R_C(a_{11})=\{a_1,a_4,a_5,a_9,a_{11},a_{12}\}$。

在非对称相似关系 $R(B)$ 中，论域 U 中每个对象都有且只有一个邻域 $R_B(x)$，共有 $|U|$ 个邻域。$R_B(x)$ 中对 x 的描述是最不完备的，信息最少，不确定性最大。$R_B^{-1}(x)$ 表示对象 x 出现的次数，每个邻域出现的概率是 $1/|U|$，则对象 x 出现的概率是 $p(x)=R_B^{-1}(x)/|U|$。但每一次出现包含不相等的信息量，为此引入 x_i 关于信息熵的平均权数 ω_i，则

定义 3 不完备信息系统 (U,A)，$U=\{x_1,x_2,\cdots,x_{|U|}\}$，$B\subseteq A$，$R_B$ 是如前面定义的非对称相似关系，则知识 (U,B) 的信息熵如下式定义。

$$H^*(B)=-\sum_{i=1}^{|U|}\omega_i\frac{|R_B^{-1}(x_i)|}{|U|}\log\frac{|R_B^{-1}(x_i)|}{|U|}$$

其中 $R_B^{-1}(x_i)$、$R_B(x_i)$ 如前面定义，ω_i 是 x_i 的信息熵权数，一般取为 $1/R_B^{-1}(x_i)$。

定义 4 非完备信息系统 (U,A)，P、$Q\subseteq A$，信息熵权数 $\omega_i=1/R_B^{-1}(x_i)$，知识库 (U,P) 和 (U,Q)，则知识 Q 相对于知识 P 的条件熵 $H^*(Q|P)$ 定义为：

$$H^*(Q/P)=\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|R_P^{-1}(x_i)|}{|R_{Q\cup P}^{-1}(x_i)|}$$

定理 2 设 R_B 是知识 (U,B) 的等价关系，则 $H^*(B)=H(B)$，$H^*(Q|P)=H(Q|P)$。

证明：设 R_B 是知识 (U,B) 下的等价关系，则 U 被划分成 k 个不同的等价类：

$$U/IND(B)=\{X_1,X_2,\cdots,X_k\}$$

则 $\forall x_i\in U$，$R_B^{-1}(x_i)=R_B(x_i)$，设 $R_B(x_i)$ 就是 x_i 所在的等价类 X_i，$\omega_i=1/|X_i|$。

$$\begin{aligned}H^*(B)&=-\sum_{i=1}^{|U|}\omega_i\frac{|R_B^{-1}(x_i)|}{|U|}\log\frac{|R_B^{-1}(x_i)|}{|U|}\\&=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{R_B^{-1}(x_i)}{|U|}=-\sum_{i=1}^{|U|}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}\\&=H(B)\end{aligned}$$

$$\begin{aligned}H(Q|P)&=\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}-\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\log\frac{|Y_j\cap X_i|}{|U|}\\&=\sum_{i=1}^{n}\frac{1}{|U|}\sum_{x\in X_i}\log\frac{|X_i|}{|U|}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{|U|}\sum_{x\in Y_j\cap X_i}\log\frac{|Y_j\cap X_i|}{|U|}\\&=\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{|R_P^{-1}(x_i)|}{|U|}-\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{|R_{P\cup Q}^{-1}(x_i)|}{|U|}\\&=\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|R_P^{-1}(x_i)|}{|R_{P\cup Q}^{-1}(x_i)|}=H^*(Q|P)\end{aligned}$$

定理 3 不完备信息系统 (U,A)，P、$Q\subseteq A$，知识库 (U,P) 和 (U,Q)，则

1) $Q=Red(P\cup Q)\Leftrightarrow H^*(P|Q)=0$；

2) $P\subset Q\Rightarrow H^*(P)\geqslant H^*(Q)$。

证明：

1) 由于 $Q=Red(P\cup Q)$，则 $R(Q)=R(P\cup Q)$，

$R_Q^{-1}(x)=\{y|\ y\in U\wedge(y,x)\in R(Q)\}=\{y|\ y\in U\wedge(y,x)\in R(P\cup Q)\}=R_{P\cup Q}^{-1}(x)$。

即 $\log\frac{|R_Q^{-1}(x_i)|}{|R_{Q\cup P}^{-1}(x_i)|}=0$，故 $H^*(P|Q)=0$。反过来，若 $H^*(P|Q)=0$，由定义式，$\forall x_i\in U$，$R_Q^{-1}(x_i)=R_{P\cup Q}^{-1}(x_i)$，则 $R(Q)=R(P\cup Q)$，$Q=Red(P\cup Q)$。

2) 由于 $P\subset Q$，则 $\forall x\in U$，

$R_Q^{-1}(x)=\{y|\ y\in U\wedge(y,x)\in R(Q)\}\subseteq\{y|\ y\in U\wedge(y,x)\in R(P)\}=R_P^{-1}(x)$

$$\begin{aligned}H^*(Q)&=-\sum_{i=1}^{|U|}\omega_i\frac{|R_Q^{-1}(x_i)|}{|U|}\log\frac{|R_Q^{-1}(x_i)|}{|U|}\\&=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{|R_Q^{-1}(x_i)|}{|U|}\leqslant-\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{|R_P^{-1}(x_i)|}{|U|}=H^*(P)\quad i=1,2,\cdots,n\end{aligned}$$

定理 4 属性集合 B 是不完备信息系统 (U,A) 的属性约简，当且仅当 $H^*(B)=H^*(A)$，且对任意 $C\subset B$，$H^*(C)<H^*(B)$。

证明：假设属性集 B 是系统 (U,A) 的一个属性约简，则 $R(A)=R(B)$，且对于任何 $C\subset B$，$R(C)>R(B)$，即对任意 $x_i\in U$，$R_B^{-1}(x_i)=R_A^{-1}(x_i)$，故 $H^*(B)=H^*(A)$。$R(C)>R(B)$ 意味着对任意 $x_i\in U$，$R_B^{-1}(x_i)\subseteq R_C^{-1}(x_i)$，由熵的定义式，必有 $H^*(C)<H^*(B)$。

另一方面，假设 $H^*(B)=H^*(A)$，且对任意 $C\subset B$，$H^*(C)<H^*(B)$。由于 $B\subset A$，又 $H^*(B)=H^*(A)$，则对任意 $x_i\in U$，必有 $R_B^{-1}(x_i)=R_A^{-1}(x_i)$，即 $R(A)=R(B)$。对任意 $C\subset B$，如果 $R(C)>R(B)$ 不成立，则必将有 $R(C)=R(B)$，这将产生 $H^*(C)=H^*(B)$ 的矛盾。因此 $R(C)>R(B)$ 必定成立。

结论 本文从信息论角度仔细地研究了完备信息系统的信息熵及其属性约简问题，并在分析了非完备信息系统的特殊性之后，通过引入权系数，建立了基于容差关系的加权信息熵和条件熵概念，将信息熵概念从等价关系逻辑地推广到容差关系。本文的结果合理地解释了不完备信息系统知识的粗糙性本质，为进一步在不完备信息系统中获得知识提供了理论依据。

参 考 文 献

1 苗夺谦，王珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论[J]. 模式识别与人工智能，1998，11(3)：34～40

2 Krysckiewicz M. Rough set approach to incomplete information system [J]. Information Sciences, 1998, 112: 39～49

3 Stefamowski J, Tsoukeas A. On the extension of rough sets under incomplete information [J]. International Journal of Interlligent System, 1999, 16(1): 29～38

4 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

5 Liang ji-ye, Qu kai-she. Information measures of roughness of knowledge reduction in incomplete information systems [J]. Journal of Systems Science and Systems Engineering, 2001, 10(4): 418～424

6 黄兵，周献中，史迎春. 基于一般二元关系的知识粗糙熵与粗集粗糙熵. 系统工程理论与实践，2004(1)：93～96