粗糙集理论中知识粗糙性与信息熵关系的讨论*

苗夺谦 王 珏

(中国科学院自动化研究所 北京 100080)

摘 要

粗糙集理论把知识看作是具有粒度的，引入了知识粗糙性的概念．本文主要讨论知识粗糙性与信息熵之间的关系，证明了熵与互信息对于由知识粗糙性定义的偏序“较细”都是单调下降的．通过反例说明，一般情况下，其逆关系是不成立的．同时给出了逆关系成立的条件．揭示了知识粗糙性实质上是其所含信息多少的更深层次上的刻划．

关 键 词 粗糙集理论，知识粗糙性，信息熵，互信息

中图法分类号 TP18

1 引 言

粗糙集理论 (Rough set theory) 是由 Z. Pawlak 于 1992 年提出的[2]．这一理论为处理具有不精确和不完全信息的分类问题提供了一种新的框架．归纳起来，它具有如下特点：(1) 从新的视角对知识进行了定义．把知识看作是关于论域的划分，从而认为知识是具有粒度 (granularity) 的．(2) 认为知识的不精确性是由知识粒度太大引起的．(3) 为处理数据 (特别是带噪声、不精确或不完全数据) 分类问题提供了一套严密的数学工具，使得对知识能够进行严密的分析和操作．

由于以上这些特点，目前，粗糙集理论已广泛应用于机器学习、故障诊断、控制算法获取、过程控制以及关系数据库中的知识获取等各种应用领域[2,4]，并取得了很大成功．

粗糙集理论的主要思想就是在保持分类能力不变的前提下，通过知识约简，导出概念的分类规则．该理论通过不可区分关系和集包含关系定义了知识粗糙性，但其本质含义不易被人理解．众所周知，ID3[3] 也是一种从数据中导出分类规则 (决策树表示) 的方法．它是以熵作为划分度量准则的，而熵具有很好的解释性．那么，基于不可区分关系和集包含关系定义的知识粗糙性与信息熵之间有没有关系呢？它们的关系又如何呢？

本文主要讨论知识粗糙性与信息熵之间的关系，证明了熵及互信息对于定义在知识上的偏序“较细”是单调下降的．通过反例，说明它们之间的逆关系一般是不成立的．同时对逆关系成立的条件进行了讨论．本文揭示了知识粗糙性实质上是其所含信息多少的更深层次上的刻划．从而，对知识粗糙性给出了信息论的解释，使人们更容易理解其本质．

* 收稿日期：1996-09-23

2　知识的信息熵与互信息的定义

2.1 知识粗糙性

为了对知识进行严密的分析和有效地操作，粗糙集理论对知识作了形式化的定义．把知识看作是关于论域 (感兴趣的对象集合) 的划分，从而使得知识具有了粒度．

设 $U=\phi$ 是感兴趣的对象组成的有限集合，称为论域．任何子集 $X\subseteq U$, 称为 U 中的一个概念． U 中的一族概念，称为关于 U 的知识．该理论主要是对在 U 上能形成划分的那些知识感兴趣．一个划分 C 定义为： $C=\{X_1,X_2,\cdots,X_n\}$, 使得 $X_i\subseteq U$, $X_i\neq\phi$, $X_i\cap X_j=\phi$, 对 $i\neq j$, $i,j=1,2,\cdots,n$. 且 $\bigcup_{i=1}^{n}X_i=U$.

U 上的一族划分，称为关于 U 的一个知识库．可以证明， U 上的一个划分与其上的一个等价关系是等价的 [2]．但等价关系在数学上容易处理，所以关于 U 的一个知识库可以理解为一个关系系统 $K=(U,R)$, 其中 U 为论域， R 为 U 上的一族等价关系．若 $P\subseteq R$ 且 $P\neq\phi$, 则 $\bigcap P$ 也是一个等价关系，记作 $IND(P)$. 并称为 P 上的一个不可区分关系．符号 $U/IND(P)$ 表示不可区分关系 $IND(P)$ 在 U 上导出的划分．

设 $K=(U,P)$ 和 $K_l=(U,Q)$ 是两个知识库．如果 $U/IND(P)\subseteq U/IND(Q)$, 则称知识 P 比知识 Q 较细 (finer), 或 Q 比 P 较粗 (coarser), 记作 $P\prec Q$. 该关系可以看作是知识库 K 上的一个偏序．这种通过等价关系和集包含关系定义的知识粗糙性，其含义是不易理解的 [1]．要进一步了解粗糙集理论，可参看文献 [2].

注意： $U/IND(P)\subseteq U/IND(Q)$ 是指对任意的 $A\in U/IND(P)$, 总存在 $B\in U/IND(Q)$, 使得 $A\subseteq B$ 成立．

2.2 知识的信息熵与互信息

设 U 为一个论域， P、Q 为 U 上的两个等价关系．我们认为 U 上任一等价关系 (即划分) 可以看作是定义在 U 上的子集组成的 σ- 代数上的一个随机变量．其概率分布通过如下方法来确定．

设 P、Q 在 U 上导出的划分分别为 X、Y:

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

则 P、Q 在 U 的子集组成的 σ- 代数上定义的概率分布分别为

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

和

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

其中， $p(X_i)=\dfrac{\mathrm{card}X_i}{\mathrm{card}U}$, $i=1,2,\cdots,n$. $p(Y_j)=\dfrac{\mathrm{card}Y_j}{\mathrm{card}U}$, $j=1,2,\cdots,m$. P 与 Q 的联合概率分布为

$$[XY;p]=\begin{bmatrix}X_1\cap Y_1 & \cdots & X_i\cap Y_j & \cdots & X_n\cap Y_m\\ p(X_1Y_1) & \cdots & p(X_iY_j) & \cdots & p(X_nY_m)\end{bmatrix}.$$

其中

$$p(X_iY_j)=\frac{\mathrm{card}(X_i\cap Y_j)}{\mathrm{card}U},\quad i=1,2,\cdots,n;\quad j=1,2,\cdots,m.$$

符号 $\mathrm{card}E$ 表示集合 E 的基数.

有了知识概率分布的定义之后，根据信息论[5]可以定义知识的熵与条件熵的概念. 知识 P 的熵 $H(P)$ 为

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

知识 Q 相对于知识 P 的条件熵 $H(Q|P)$ 为

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

知识 P 与 Q 的互信息 $I(P;Q)$ 为

$$I(P;Q)=H(Q)-H(Q|P)\ .$$

熵度量了信源提供的平均信息量的大小. 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.

3 知识粗糙性与信息熵的关系

本节主要讨论知识粗糙性与信息熵之间的关系，证明熵及互信息对于定义在知识上的偏序“较细”来说都是单调下降的. 通过反例说明，其逆关系是不成立的.

3.1 主要结论

定理 1 设 U 为一个论域. $K=(U,P)$ 和 $K_1=(U,Q)$ 是关于 U 的两个知识库，并且 $P\prec Q$. 则 $H(P)\geq H(Q)$.

证明：设知识 P 与 Q 的概率分布分别为

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

和

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

因为 $P\prec Q$，即 $U/IND(P)\subseteq U/IND(Q)$；以及 $\{X_i,\ i=1,2,\cdots,n\}$ 与 $\{Y_i,\ i=1,2,\cdots,m\}$ 都是 U 的划分，所以存在 $\{1,2,\cdots,n\}$ 的子集 E_i，$E_i\cap E_j=\phi, i\neq j$，$i,j=1,2,\cdots,m$，使得

$$Y_i=\bigcup_{j\in E_i}X_j\ ,\qquad j=1,2,\cdots,m\ .$$

从而有

$$p(Y_i)=\sum_{j\in E_i}p(X_j)\ ,\qquad i=1,2,\cdots,m\ .$$

由熵函数的递增性[5], 有

$$\begin{aligned} H(P) = & H[p(X_1), p(X_2), \cdots, p(X_n)] \\ = & H\left[\sum_{j\in E_1} p(X_j), \sum_{j\in E_2} p(X_j), \cdots, \sum_{j\in E_m} p(X_j)\right] \\ & + \sum_{i=1}^{m} \sum_{l=1}^{|E_i|} \sum_{k=l} p(X_k) H\left[\frac{\sum\limits_{j\in E_i, j\neq i} p(X_j)}{\sum\limits_{j\in E_i} p(X_j)}, \frac{\sum\limits_{j=i} p(X_j)}{\sum\limits_{j\in E_i} p(X_j)}\right] . \end{aligned} \tag{1}$$

因为

$$H(Q) = H[p(Y_1), p(Y_2), \cdots, p(Y_m)] = H\left[\sum_{j\in E_1} p(X_j), \sum_{j\in E_2} p(X_j), \cdots, \sum_{j\in E_m} p(X_j)\right] ,$$

从概率及熵函数的非负性可知，式 (1)≥ 0, 所以， $H(P) \geq H(Q)$.

下面的定理给出了知识粗糙性与互信息之间的关系.

定理 2　设 U 为一个论域. $K = (U, P)$ 和 $K_1 = (U, Q)$ 是关于 U 的两个知识库， D 是 U 上的决策. 若 $P \prec Q$, 则 $I(P; D) \geq I(Q; D)$.

证明：设知识 P、Q 及决策 D 定义的概率分布分别为

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

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

和

$$Z = \begin{bmatrix} d_1 & d_2 & \cdots & d_r \\ p(d_1) & p(d_2) & \cdots & p(d_r) \end{bmatrix} .$$

因为 $P \prec Q$, 以及 $\{X_i, i = 1, 2, \cdots, n\}$ 与 $\{Y_i, i = 1, 2, \cdots, m\}$ 分别为 U 的划分, 所以存在 $\{1, 2, \cdots, n\}$ 的子集 E_i， $E_i \cap E_j = \phi$， $i \neq j$， $i, j = 1, 2, \cdots, m$, 使得

$$Y_i = \bigcup_{j\in E_i} X_j , \qquad i = 1, 2, \cdots, m .$$

从而有

$$p(Y_i) = \sum_{j\in E_i} p(X_j) , \qquad i = 1, 2, \cdots, m .$$

$$\sum_{j=1}^{n} p(X_j) = \sum_{i=1}^{m} \sum_{j\in E_i} p(X_j) . \tag{2}$$

由 (2) 式及 Shannon 辅助定理 [5] 有

$$\begin{aligned} H(D|P) &= -\sum_{j=1}^{n} P(X_j) \sum_{i=1}^{r} p(d_i|X_j) \log p(d_i|X_j) \\ &\le -\sum_{k=1}^{m} \sum_{j \in E_k} p(X_j) \sum_{i=1}^{r} p(d_i|X_j) \log p(d_i|Y_k) \\ &\le -\sum_{k=1}^{m} p(Y_k) \sum_{i=1}^{r} p(d_i|Y_k) \log p(d_i|Y_k) \\ &= H(D|Q) . \end{aligned}$$

因为

$$\begin{aligned} I(P;D) &= H(D) - H(D|P) , \\ I(Q;D) &= H(D) - H(D|Q) . \end{aligned}$$

所以

$$I(P;D) \ge I(Q;D) .$$

一个知识库实际上可以看成一种知识，所以把定理 1 和定理 2 限制在两种知识上时，其结论仍然成立.

3.2 反例

我们自然会问上一小节给出的结论反过来是否也成立呢？若成立，那就意味着知识粗糙性和信息熵之间存在一一对应关系. 从而可以简化粗糙集理论中知识约简的复杂性. 但遗憾的是，其逆关系一般是不成立的.

例子: 考虑如下信息系统 (表 1)[3]

表 1 一个信息系统

U	Condition attributes(C)				Decision attribute(D)
	Outlook(a_1)	Temperature(a_2)	Humidity(a_3)	Windy(a_4)	Class
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	P
4	rain	mild	high	false	P
7	overcast	cool	normal	true	P
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
10	rain	mild	normal	false	P
11	sunny	mild	normal	true	P
12	overcast	mild	high	true	P
13	overcast	hot	normal	false	P
14	rain	mild	high	true	N

令 $P=$ “ Outlook ”，　$Q=$ “ Temperature ”，　$D=$ “ Class ”. 则从表 1 可算出：

$$H(P)=-\sum_{i=1}^{3}p_i\log p_i=-\left[\frac{5}{14}\log\frac{5}{14}+\frac{4}{14}\log\frac{4}{14}+\frac{5}{14}\log\frac{5}{14}\right]=1.577\,,$$

$$H(Q)=-\sum_{i=1}^{3}p_i\log p_i=-\left[\frac{4}{14}\log\frac{4}{14}+\frac{6}{14}\log\frac{6}{14}+\frac{4}{14}\log\frac{4}{14}\right]=1.556\,,$$

且

$$I(P;D)=H(D)-H(D|P)=0.940-0.694=0.246\,,$$
$$I(Q;D)=H(D)-H(D|Q)=0.940-0.911=0.029\,,$$

即有

$$H(P)\geq H(Q);\quad I(P;D)\geq I(Q;D)\,.$$

但是

$$U/IND(P)=\{\{1,2,8,9,11\},\{3,7,12,13\},\{4,5,6.10,14\}\}\,,$$
$$U/IND(Q)=\{\{1,2,3,13\},\{4,8,10,11,12,14\},\{5.6,7,9\}\}\,,$$

显然，$U/IND(P)\not\subset U/IND(Q)$. 这就说明，定理 1 和 2 的逆关系是不成立的.

4　逆关系成立的条件

通过分析发现，要使逆关系成立，不仅要考虑信息量的大小，而且还需考虑两种知识的相依关系.

定理 3　设 U 为一个论域.　$K=(U,P)$ 和 $K_1=(U,Q)$ 是关于 U 的两个知识库，如果 $H(P)>H(Q)$ 且 $H(Q|P)=0$, 则 $P\prec Q$.

证明：因为 $H(P)>H(Q)$, 则下式一定不成立，

$$U/IND(P)\supseteq U/IND(Q)\,,\tag{3}$$

(若式 (3) 成立，则由定理 1 可知，$H(P)\leq H(Q)$, 矛盾!). 这就推出

$$U/IND(P)\supseteq U/IND(Q)\,,\tag{4}$$

或

$$U/IND(P)\not\subset U/IND(Q)\tag{5}$$

成立. 令

$$U/IND(P)=\{A_1,A_2,\cdots,A_n\}\,,$$
$$U/IND(Q)=\{B_1,B_2,\cdots,B_m\}\,,$$

若式 (5) 成立，则至少存在一个 A_k, 使得对任意的 B_j, 满足

$$A_k\not\subset B_j,\qquad j=1,2,\cdots,m\,.$$

不妨设与 A_k 的交不空的集合为 $B_{j_1},B_{j_2},\cdots,B_{j_k}$, 则有

$$0<p(B_{j_l}|A_k)<1\,,\qquad l=1,2,\cdots,k\,.$$

从而，

$$H(Q|P) = -\sum_{i=1}^{n} p(A_i) \sum_{j=1}^{m} p(B_j|A_i) \log p(B_j|A_i) \geq -p(A_k) \sum_{l=1}^{k} p(B_{jl}|A_k) \log p(B_{jl}|A_k) > 0 ,$$

这与条件 $H(Q|P) = 0$ 矛盾！故只有式 (4) 成立，即

$$U/IND(P) \subseteq U/IND(Q) .$$

同理，可以证明下面的定理成立.

定理 4 设 U 为一个论域，$K = (U, P)$ 和 $K_1 = (U, Q)$ 是关于 U 的两个知识库，D 为 U 上的决策. 如果 $I(P; D) \geq I(Q; D)$ 且 $H(Q|P) = 0$, 则 $P \prec Q$.

定理 3 和 4 说明：在一定条件下，知识粗糙性与信息熵 (互信息) 之间存在着一一对应关系.

5 结 论

粗糙集理论是一种处理不精确和不完全知识的工具. 它通过不可区分关系与集包含关系定义了知识的粗糙性，但其本质含义不易被人所理解. 本文建立了知识粗糙性与信息熵之间的关系，证明了熵与互信息对于定义在知识上的偏序“较细”都是单调下降的. 通过反例说明，它们的逆关系是不成立的. 同时给出了逆关系成立的条件. 从而，揭示了知识粗糙性实质上是其所含信息多少的更深层次上的刻划. 为知识粗糙性提供了一种信息解释，使人们更容易理解其本质.

参 考 文 献

[1] Kent R. E. Rough Concept Analysis. Proceedings of the International Workshop on Rough Set and Knowledge Discovery, Canada, 1993, 248—255

[2] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991

[3] Quinlan J R. Induction of Decision Trees. Machine Learning, 1986, 1: 81—106

[4] Ziarko W. Special Issue on Rough Set and Knowledge Discovery. International Journal of Computational Intelligence, 1995, 11(2)

[5] 姜 丹，钱玉美. 信息理论与编码. 中国科学技术出版社，1992

ON THE RELATIONSHIPS BETWEEN INFORMATION ENTROPY AND ROUGHNESS OF KNOWLEDGE IN ROUGH SET THEORY

Miao Duoqian, Wang Jue

(*AI Lab of Inst of Automation, Chinese Academy of Sciences, Beijing* 100080)

ABSTRACT

Rough Set theory is a kind of new tool for dealing with imprecise knowledge. In this paper, relationships between roughness of knowledge and information entropy are mainly discussed. We prove that entropy and mutual information are decreasing for the partial order finer on knowledge. Through negative examples, we show that the inverse relationships between them are not valid. The conditions that satisfy the inverse relationships are also given. In fact, roughness of knowledge does deeply give a description of its information.

Key Words Rough Set Theory, Roughness of Knowledge, Information Entropy, Mutual Information