知识粗糙性和条件信息熵的关系

陈凤娟[1],闫德勤[2]
CHEN Feng-juan[1],YAN De-qin[2]

1.辽宁对外经贸学院 信息技术系,辽宁 大连 116052
2.辽宁师范大学 计算机与信息技术学院,辽宁 大连 116029
1.Department of Information Technology,Liaoning University of International Business and Economic,Dalian,Liaoning 116052,China
2.Department of Computer Science,Liaoning Normal University,Dalian,Liaoning 116029,China
E-mail:123chenfengjuan123@163.com

CHEN Feng-juan,YAN De-qin.Relationship between roughness of knowledge and conditional information entropy.Computer Engineering and Applications,2009,45(21):160-162.

Abstract: At present,there are two views in rough set theory.They are algebra view and information view.The roughness of knowledge embodies the granularity of knowledge in algebra view.Two concepts are defined in information view,which are information entropy of knowledge and conditional information entropy of knowledge.Theorem has been proved that information entropy and roughness of knowledge have corresponding relationships.It establishes the link between algebra view and information view,but they are not one-to-one relationships.By re-prove the theorem on the relationship between information entropy and roughness of knowledge,this paper finds that it is conditional information entropy which exists one-to-one relationship with roughness of knowledge.Then several related theorems are given and proved.

Key words: roughness of knowledge;information entropy;conditional information entropy

摘 要: 目前,粗糙集理论存在着两种观点,它们分别是代数观和信息观。在代数观点中,知识粗糙性体现了知识的粒度;而在信息观中,定义了知识的信息熵和条件信息熵。已经有定理证明了信息熵与知识的粗糙性存在对应关系,它建立了代数观和信息观之间的联系,但是这种关系却不是一一对应的。该文通过重新证明知识粗糙性和信息熵的对应关系定理,找到与知识粗糙性存在一一对应关系的是条件信息熵,并给出相关定理及其证明。

关键词: 知识粗糙性;信息熵;条件信息熵
DOI: 10.3778/j.issn.1002-8331.2009.21.047 **文章编号:** 1002-8331(2009)21-0160-03 **文献标识码:** A **中图分类号:** TP18

1 引言

粗糙集(Rough Set)理论是由 Z.Paw lak 于 1982 年提出的,该理论是继概率论、模糊集、证据理论之后的又一个处理不确定性的数学工具,它能有效地处理具有模糊、不精确或不完全信息的分类问题[1]。粗糙集理论作为一种新的软计算方法,近年来越来越受到重视,是当前国际上人工智能理论及其应用领域中的研究热点之一[2-4]。

粗糙集的主要思想是在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。这通常被称为粗糙集理论的代数观点, 苗等提出了粗糙集的信息论观点,并分析了知识粗糙性与信息熵的关系,证明了信息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不成立,并说明了使逆关系成立的条件[5]。分析了知识粗糙性与信息熵的关系,重新证明了该定理,在分析该定理的逆关系时,发现其逆关系中的条件信息熵才是使逆关系成立的关键条件,由此提出了新的知识粗糙性与条件信息熵的对应关系定理及其逆定理,并给出了其证明过程。这两个新定理揭示了知识粗糙性与条件信息熵的一一对应关系,说明了知识的粗细和知识的相依关系是完全对应的。

2 基本概念

首先给出代数观中的相关概念[6]。

定义 1 一个近似空间(approximate space)(或知识库)定义为一个关系系统(或二元组)$K=(U,R)$。其中 $U\neq\Phi$(Φ 为空集) 是一个被称为全域或论域的所有要讨论的个体的集合,R 是 U 上等价关系的一个族集。

定义 2 设 $P\subseteq R$, 且 $P\neq\Phi$,P 中所有等价关系的交集称为 P 上的一种不分明关系(indiscernibility relation)(或称不可区分关系),记作 $IND(P)$,即

$$[x]_{IND(P)}=\bigcap_{R\in P}[x]_R$$

基金项目: 辽宁省教育厅高等学校科学研究基金(No.2008344);大连市科技局科技计划项目(No.2007A10GX117)。
作者简介: 陈凤娟(1979-),女,讲师,主要研究领域为数据挖掘,粗糙集理论;闫德勤(1962-),男,博士,教授,主要研究领域为模式识别,知识发现,粗糙集理论等。
收稿日期: 2009-05-04 **修回日期:** 2009-06-05

$[x]_R$表示的是包含元素$x\in U$的R等价类。$IND(P)$也是等价关系且是唯一的。

定义 3 设$K=(U,P)$和$K_1=(U,Q)$是两个知识库,如果$IND(P)=IND(Q)$,则称K和K_1(或P和Q)是等价的,记作$K\backsimeq K_1$(或$P\backsimeq Q$)。

对于论域U上的两个等价关系集合P,Q,如果有$U/IND(P)\subseteq U/IND(Q)$,则称知识$P$比知识$Q$较细,或知识$Q$比知识$P$较粗,记作$P\prec Q$。其中,$U/IND(P)\subseteq U/IND(Q)$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。

下面是基于信息论的观点定义的知识的信息熵与条件信息熵[2]。

设U为一个论域,P,Q为U上两个等价关系(即知识)。把U上任一等价关系看作是定义在U上的子集组成的σ代数上的一个随机变量,其概率分布可通过如下方法来确定。

定义 4 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,$p(X_i)=|X_i|/|U|,i=1,2,\cdots,n;p(Y_j)=|Y_j|/|U|,j=1,2,\cdots,m$;符号$|E|$表示集合$E$的基数。

有了知识概率分布的定义之后,根据信息论可以定义知识的信息熵与条件信息熵的概念。

定义 5 知识P的熵$H(P)$定义为$H(P)=-\sum_{i=1}^{n}p(X_i)\log p(X_i)$。

定义 6 知识Q相对于知识P的条件熵$H(Q|P)$定义为:

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

其中,$p(Y_j|X_i)=|Y_j\cap X_i|/|X_i|,i=1,2,\cdots,n;j=1,2,\cdots,m$。

3 知识粗糙性与信息熵的关系

苗在文献[5]中给出了知识粗糙性和信息熵之间的关系定理及其证明,在该证明过程中使用了熵函数的递增性及相关性质,不易看出知识的信息熵的定义与代数观点的定义之间的内部联系,下面重新证明这个定理。

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,若知识P比知识Q较细,则$H(P)\geqslant H(Q)$。

证明 设$U/IND(P)=\{X_1,X_2,\cdots,X_n\},U/IND(Q)=\{Y_1,Y_2,\cdots,Y_m\}$,则由已知条件知识$P$比知识$Q$较细,可知对于任意的$X_i$总存在$X_i\subseteq Y_j$,其中$1\leqslant i\leqslant n,1\leqslant j\leqslant m$,又因为$X$和$Y$都是$U$的划分,所以任意的$Y_j$由大于等于一个的$X_i$组成。

由知识的熵的定义有:

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

$$H(Q)=-\sum_{j=1}^{m}p(Y_j)\log(p(Y_j))$$

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

$[p(Y_1)\log(p(Y_1))+p(Y_2)\log(p(Y_2))+\cdots+p(Y_m)\log(p(Y_m))]-[p(X_1)\log(p(X_1))+p(X_2)\log(p(X_2))+\cdots+p(X_n)\log(p(X_n))]$

(1)先分析Y_j由一个X_i组成的情况,即$Y=X$,知识P和知识Q对U的划分完全相等。

对每个X和Y来说,都有$p(Y)\log(p(Y))=p(X)\log(p(X))$,此时,$H(P)-H(Q)=0$,即$H(P)=H(Q)$。

(2)再分析任意Y_j由大于一个的X_i组成的情况。

假设$Y_1=X_1\cup X_2,|X_1|=a,|X_2|=b$,则$|Y_1|=|X_1|+|X_2|=a+b$,则对应的$p(Y_1)\log(p(Y_1))-[p(X_1)\log(p(X_1))+p(X_2)\log(p(X_2))]=(a+b)\log(a+b)-(a\log a+b\log b)=a\log\frac{a+b}{a}+b\log\frac{a+b}{b}$

$\because a>0,b>0$

$\therefore p(Y_1)\log(p(Y_1))-[p(X_1)\log(p(X_1))+p(X_2)\log(p(X_2))]>0$

$\therefore p(Y_1)\log(p(Y_1))>p(X_1)\log(p(X_1))+p(X_2)\log(p(X_2))$

由假设$Y_1=X_1\cup X_2$,可以得到结论:当Y是由两个X组成时,Y的熵大于两个X的熵的和,其中Y和X有任意性。

由归纳推理可得上面的Y可以推广到多个X的组合的情况,原因如下:

因为上式是由多个$(a\log(a+b)-a\log a)$结构组成的,所以该式的本质是比较$(a\log(a+b)-a\log a)=a\log[(a+b)/a]$的符号是否为正,当$Y$是由大于2个的$X$组合而成时,$(a\log(a+b)-a\log a)$的结构变成$(a\log(a+b+c+\cdots+n)-a\log a)=a\log[(a+b+c+\cdots+n)/a]$该式仍然大于0,所以整个式子都大于0。

所以,当Y是由大于两个X组成时,Y的熵大于多个X的熵的和,其中Y和X有任意性。即当Y是由$\geqslant 2$个的X组成时,$p(Y)\log(p(Y))>\sum_{i=1}^{n}p(X_i)\log(p(X_i))$,所以$H(P)-H(Q)>0$。

(3)最后分析部分Y_j由等于一个的X_i组成,部分Y_j由大于一个的X_i组成的情况。

对于Y_j由等于一个的X_i组成的情况,适用(1)中的分析,即$H(P)=H(Q)$。

对于Y_j由大于一个的X_i组成的情况,适用(2)中的分析,即$H(P)>H(Q)$。

由上面的情况(1)、(2)和(3)可知,对于$K=(U,P)$和$K_1=(U,Q)$这两个关于U的知识库,如果有P比Q细,则必有$H(P)\geqslant H(Q)$成立。当P和Q对U的划分完全相同时等号成立。

该定理通过两个划分块的子集的包含关系与信息熵公式的对应关系证明了信息熵与知识粗糙性的关系,从证明过程中可以清楚地看到代数观中划分与信息观中的信息熵之间的联系。

文献[5]中又举反例说明该定理的逆不成立,要使逆关系成立,不仅要考虑信息量的大小,还要考虑两种知识的相依关系,并给出定理2及其证明。

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,若$H(P)>H(Q)$且$H(Q|P)=0$,则$P\prec Q$。

在对定理进行分析及证明的过程中,发现真正使得该定理成立的条件是$H(Q|P)=0$,而条件$H(P)>H(Q)$在该定理中不起关键作用,甚至可以省略。所以,修改定理2为下面新的定理3,该定理可以看出知识粗糙性和条件信息熵的关系。

4 知识粗糙性与条件信息熵的关系

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,若 $H(Q|P)=0$,则 $P\prec Q$。

已知:P 和 Q 是对 U 的划分,$H(Q|P)=0$

求证:$P\prec Q$

证明 设 $U/IND(P)=\{X_1,X_2,\cdots,X_n\}$,$U/IND(Q)=\{Y_1,Y_2,\cdots,Y_m\}$

$$H(Q|P)=0\Leftrightarrow-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))=0\Leftrightarrow$$

$$\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))=0$$

$\because$ 对于所有的 X_i,都有 $P(X_i)=|X_i|/|U|>0$

$\therefore$ 要使上面的求和结果为零,则必有每个 $\sum_{j=1}^{m}p(Y_j|X_i)\times\log(p(Y_j|X_i))=0$ 成立,而这个求和式子中,$1\geqslant p(Y_j|X_i)=|Y_j\cap X_i|/|X_i|\geqslant 0$,$\log(p(Y_j|X_i))\leqslant 0$,$p(Y_j|X_i)\log(p(Y_j|X_i))\leqslant 0$,也就是该和式中,每一项都同号($\leqslant 0$),要使 $\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))=0$ 成立,则每个求和项都得等于零。使每个求和项都等于零只有两种可能:

$p(Y_j|X_i)=0$ 或 $\log(p(Y_j|X_i))=0$;

$p(Y_j|X_i)=\Rightarrow|Y_j\cap X_i|/|X_i|=0\Rightarrow|Y_j\cap X_i|=0\Rightarrow Y_j$ 与 X_i 的交集为空;

$\log(p(Y_j|X_i))=0\Rightarrow p(Y_j|X_i)=1\Rightarrow|Y_j\cap X_i|=X_i\Rightarrow Y_j$ 与的交集为 X_i。

所以知识 P 和知识 Q 对于论域 U 的划分块 X 和 Y,存在下面的关系:任意的 Y 与任意的 X 的交集要么为空,要么为 X,也就是对任意的 $X\in U/IND(P)$,总存在 $Y\in U/IND(Q)$ 使得 $X\subseteq Y$ 成立,即 $P\prec Q$。

由定理 3 的证明过程可以看到,该定理的逆也成立。

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,若 $P\prec Q$,则 $H(Q|P)=0$。

已知:P 和 Q 是对 U 的划分,$P\prec Q$。

求证:$H(Q|P)=0$。

证明 设 $U/IND(P)=\{X_1,X_2,\cdots,X_n\}$,$U/IND(Q)=\{Y_1,Y_2,\cdots,Y_m\}$

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

因为 P 和 Q 是对 U 的划分,且 $P\prec Q$,所以对于任意的划分块 X 和 Y,存在下面的关系,即任意的 Y 与任意的 X 的交集要么为空,要么为 X,这样就有 $p(Y_j|X_i)=0$ 和 $\log(p(Y_j|X_i))=0$ 二者必有其一成立,则 $\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))=0$ 成立,所以 $H(Q|P)=0$ 成立。

由定理 1 和定理 2 可以看到知识粗糙性和信息熵之间有一定的联系,而这种联系却不是一一对应的,只有在特定的条件下,二者才存在一一对应关系。而由定理 3 和定理 4 的证明过程及其结论可以看出,知识粗糙性和条件信息熵之间的关系是一一对应关系,只要知识 P 比知识 Q 较细,则 $H(Q|P)=0$ 一定成立,反之,若有 $H(Q|P)=0$,则必有知识 P 比知识 Q 较细。这两个定理揭示了知识粗糙性实质上是两种知识的相依关系的更深层次上的刻画,同时也使基于信息观的条件信息熵的定义更容易理解。

5 结论

粗糙集代数观通过不可区分关系与集合包含关系定义了知识的粗糙性,而信息观通过信息熵定义了知识的熵和条件熵。该文建立了知识粗糙性与条件信息熵的关系,证明了二者存在一一对应的关系,建立了代数观和信息观之间的桥梁,为知识约简的理论研究和实际应用提供基础。

参考文献:

[1] Pawlak Z.Rough set:Theoretical aspects of reasoning about data [M].Dordrecht,Netherlands:Kluwer Academic Publishers,1991.

[2] Lingras P J,Yao Y Y.Data mining using extensions of the rough set model[J].Journal of the American Society for Information Science,1998,49(5):415-422.

[3] Tsumoto S.Automated discovery of positive and negative knowledge in clinical databases based on rough set model[J].IEEE EMB Magazine,2000,19(4):56-62.

[4] Wojcik Z M.Detecting spots for nasa space programs using rough sets[C]//Proc of the 2nd International Conference on Rough Sets and Current Trends in Computing,RSCTC'2000,Canada,2000:531-537.

[5] 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(1):34-40.

[6] 曾黄麟.粗集理论及其应用[M].修订版.重庆:重庆大学出版社,1998.

(上接 87 页)

[10] Al-Riyami S S,Paterson K G.Certificateless public key cryptography[C]//LNCS 2894:Advance in Cryptography Asiacrypt 2003.Berlin:Springer Verlag,2003:452-473.

[11] Cheng Zhao-hui,Chen Li-qun,Ling Li,et al.General and efficient certificateless public key encryption constructions[C]//Takagi T.LNCS 4575:Pairing 2007.Berlin Heidelberg:Springer Verlag,2007:83-107.

[12] Dent A W.A survey of certificateless encryption schemes and security models.Cryptology ePrint Archive,Report 2006/211.2006.

[13] Huang Xin-yi,Mu Yi,Susilo W,et al.Certificateless signature revisited[C]//LNCS 4586:ACISP 2007.Berlin:Springer Verlag,2007:308-322.

[14] Zhang Z,Wong D S,Xu J,et al.Certificateless public-key signature:Security model and efficient construction[C]//LNCS 3989:Applied Cryptography and Network Security 2006.Berlin:Springer Verlag,2006:293-308.

[15] Barbosa M,Farshim P.Certificateless signcryption.Cryptography Eprint Archive,Report 2008/143.URL http://eprint.iacr.org/2008/143.2008.