DOI:10.3969/J. ISSN. 1672-6693. 2009. 03. 018

基于等价关系的信息熵及概率分配函数*

赵晓雨，周润珍
(重庆文理学院 应用技术师范学院，重庆 402160)

摘要：Pawlak 在 1982 年提出的粗糙(Rough)集是基于等价关系的理论，粗糙集的发展推动了人们对等价关系的研究。等价关系上的信息熵具有最为简单、规范的性质。本文研究基于等价关系上的信息熵及概率分配函数，讨论基于等价关系上的信息熵的基本性质，为等价关系的信息熵的各种应用提供理论基础，比如等价关系的信息熵在信息系统的约简方面可能发挥重要作用。文章主要从两方面进行论证：①等价关系的粗细对信息熵的影响，这点通过 8 个命题来说明；②等价关系与证据理论之间的联系。证据理论主要是通过概率分配函数、信任函数及似然函数来表述，从某种意义上说粗集理论继承和发展了证据理论。另外，本文的讨论均在有限论域 $U=\{u_1,u_2,\cdots,u_{|U|}\}$ 上进行，用具体的例子来说明抽象的数学命题，使之更容易理解。

关键词：等价关系；信息熵；粗糙集；入侵检测
中图分类号：TP301　　文献标识码：A　　文章编号：1672-6693(2009)03-0075-04

熵是 1865 年由 Clausius 最先提出，并用来量度热力学过程不可逆程度的；随后 Boltzmann 说明了熵的统计意义，把熵作为物质系统内部无序程度的量度。1948 年 Shannon 把熵的概念引入信息论中，作为随机事件不确定性的量度。Pawlak[1] 在 1982 年给出了粗糙(Rough)集的概念后，人们发现它在数据挖掘和知识发现中的许多应用，因此近年来该理论及其应用得到迅速发展[2-6]。由于等价关系是粗糙集理论赖以建立的基础，等价关系的研究得到空前的深入。例如在信息表的约简方面，由于信息表的每个属性对应于一个等价关系，因此信息表的约简实际上就是除去冗余的等价关系及等价类的约简。为此，人们想出了各种各样的方法，除常用的集合论方法外，还把 Shannon 定义的概率信息熵[7] 的方法搬到等价关系上[8-9]，使得信息表的约简与相应的等价关系的信息熵发生密切联系[10-11]。等价关系的信息熵在科学技术的许多方面都有重要应用。例如，最近人们把信息熵的方法应用于计算机的入侵检测(Intrusion detection)[12]，用信息熵来衡量或发现审计数据的规律性。本文主要研究等价关系的信息熵的基本性质，为等价关系的信息熵的各种应用提供理论基础，比如等价关系的信息熵在信息系统的约简方面可能发挥重要作用。

本文的讨论均在有限论域 $U=\{u_1,u_2,\cdots,u_{|U|}\}$ 上进行，若 R 为 U 上的等价关系，R 确定了 U 的一个划分，划分的结果称为商空间(或称商集)，记为 $U/R=\{X_1,X_2,\cdots,X_n\}$，借助 Shannon 定义的概率信息熵，$U$ 上的等价关系 R 的信息熵定义为

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

其中 $P(X)=|X_i|/|U|$

这里对数的底数一般取为 2(也可以是其它大于 1 的数，其中自然对数的底数 e 对数学处理较为方便)，当底数取为 2 时，信息熵的单位为比特(bite)，与概率的信息熵一样，等价关系信息熵的最大值为 $\log|U|$(对应于恒等关系)，最小值为 0(对应于全关系)。由于等价关系确定的分类给出后概率计算简单，等价关系的信息熵较概率的信息熵直观，计算也相对简单。

1 等价关系的粗细对信息熵的影响

等价关系的粗细指的是等价关系的包含。设 P，Q 为 U 上的两个等价关系，如果 $P\subseteq Q$，则称 P 较 Q 细或称 Q 较 P 粗，若 P 较 Q 细，则 P 相应的分类是 Q 的分类的细化，即可以理解为按分类标准 P 的分类

* 收稿日期：2009-04-13
资助项目：重庆市教育委员会科学技术研究项目(No. KJ081207)
作者简介：赵晓雨，男，副教授，硕士，研究方向为人工智能、数据挖掘。

结果是按分类标准 Q 的分类结果的进一步细分。若有一个等价关系的序列 $P_0 \subset P_1 \subset \cdots \subset P_n$,其分类的谱系图恰好是一棵 n 层树,所有的叶节点构成集合 U。容易看出有以下结果。

命题1　设 P,Q 为 U 上的两个等价关系,如果 $P \subset Q$,则 $H_U(Q) < H_U(P)$。

证明　设 $U/Q = \{X_1, X_2, \cdots, X_n\}$,因为 P 相应的分类是 Q 的分类的细化,即 P 的分类可以理解为 Q 的类逐步细分,而每次细分都是把 Q 的某个类细分为两个类,因此,为简单起见,可设 $A \cup B = X_1$,$A \cap B = \varnothing$,且 $U/P = \{A, B, X_2, \cdots, X_n\}$,注意到当 $0 < x, y < 1$ 时,有

$$(x+y)\log(x+y) > x\log x + y\log y$$

$$H_U(Q) - H_U(P) = -p(X_1)\log(p(X_1)) + p(A)\log(p(A)) + p(B)\log(p(B)) < 0$$

即 $H_U(Q) < H_U(P)$　证毕

命题1说明,划分越细,相应的熵越大。这对于入侵检测的异常检测方法而言,一个审计数据集中的每一个记录都代表一个类,熵越小,不同记录的个数就越少,审计数据就越具有规律性,因此利用熵值较小的数据集建立的异常检测模型具有较好的检测性能。

理论上,Shannon 在讨论信息熵的过程中,较多使用函数 $-x\log(x)$ 在区间上$(0,\infty)$的上凸性,因为信息熵的许多性质都依赖于函数 $-x\log(x)$ 的性质。Shannon 的条件信息熵也可以搬到两个等价关系上,用于研究两个等价关系之间的相互作用,设 P、Q 为 U 上的两个等价关系,$U/P = \{X_1, X_2, \cdots, X_n\}$,$U/Q = \{Y_1, Y_2, \cdots, Y_m\}$,条件信息熵定义为

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

其中 $p(Y_j \mid X_i) = |Y_j \cap X_i| / |X_i|$。

对于入侵检测的异常检测来说,通常使用一个训练数据集来建立模型,并将建立的模型应用于测试数据集,条件熵可以反映训练集与测试集之间的差异。

需要注意的是,若 P 为 U 上的等价关系,相应的商空间为 $U/P = \{X_1, X_2, \cdots, X_n\}$,则 Q 在 X_i 上的限制是也 X_i 上的等价关系,从而确定 X_i 的分类,商空间为 $X_i/Q = \{X_i \cap Y_1, X_i \cap Y_2, \cdots, X_i \cap Y_m\}$(可能有某些 $X_i \cap Y_j = \varnothing$),从条件信息熵的定义容易看出,条件信息熵 $H_U(Q \mid P)$ 实际上是 X_i 上的信息熵 $H_X(Q)$ 的加权算术平均值。即有

$$H_U(Q \mid P) = \sum_{i=1}^{n} p(X_i) H_{X_i}(Q)$$

条件信息熵有以下基本性质:

命题2[7]　设 P,Q 为 U 上的两个等价关系,则 $H_U(Q \mid P) + H_U(P) = H_U(P \cap Q)$。

命题3　设 P,Q,R 均为 U 上的等价关系,如果 $P \subset Q$,则 $H_U(Q \mid R) < H_U(P \mid R)$。

证明　由命题1及条件信息熵定义即得。证毕

命题4　设 P,Q 为 U 上的两个等价关系,则 $H_U(Q \mid P) = 0$ 当且仅当 $P \subseteq Q$。

证明　只需注意到 $H_U(Q \mid P) = 0$ 当且仅当 $H_{X_i}(Q) = 0$,即按照 Q 的标准,$X_i (i = 1, 2, \cdots, n)$ 只能分为一类,Q 为 X_i 上的全关系,这样 P 较 Q 细,因此 $P \subseteq Q$。　证毕

命题5　设 P,Q 为 U 上的两个等价关系,则 $H_U(Q \mid P) \leqslant H_U(Q)$。

证明　利用函数 $-x\log(x)$ 在区间$(0,\infty)$上的上凸性,见文献[7]。　证毕

一个很自然的问题是,如果 P,Q 为 U 上的两个等价关系,且 $P \subset Q$,是否一定有 $H_U(Q \mid P) < H_U(Q)$?笔者给出否定的答案,见下例:设 $= \{1, 2, 3, 4\}$,$U/P = \{\{1, 2\}, \{3, 4\}\}$,$U/Q = \{\{1, 3\}, \{2, 4\}\}$,则 $H_U(Q \mid P) = H_U(Q)$。

那么何时有 $H_U(Q \mid P) = H_U(Q)$?一般地,有如下命题。

命题6　设 P,Q 为 U 上的两个等价关系,$U/P = \{X_1, X_2, \cdots, X_n\}$,$U/Q = \{Y_1, Y_2, \cdots, Y_m\}$,若 $|X_i \cap Y_j| / |U| = (|X_i| / |U|)(|Y_j| / |U|)$,$(i = 1, 2, \cdots, n, j = 1, 2, \cdots, m)$ 则 $H_U(P \cap Q) = H_U(P) + H_U(Q)$ 及 $H_U(Q \mid P) = H_U(Q)$。

证明　这时只需注意 $U/(P \cap Q) = \{X_i \cap Y_j \mid 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$ 及 $p(X_i \cap Y_j) = p(X_i)p(Y_j)$,即得 $H_U(P \cap Q) = H_U(P) + H_U(Q)$,再由命题2得另一等式。　证毕

命题7　设 P,Q,R 均为 U 上的等价关系,如果 $P \subseteq Q$,则 $H_U(R \mid Q) \geqslant H_U(R \mid P)$。

证明　因为加细可以逐步实现,为简单起见,可设 $U/Q = \{X_1, X_2, \cdots, X_n\}$,$U/P = \{A, B, X_2, \cdots, X_n\}$ 且 $A \cup B = X_1$,$A \cap B = \varnothing$,$U/R = \{Y_1, Y_2, \cdots, Y_m\}$,由条件信息熵的定义及命题3有

$$H_U(R \mid Q) - H_U(R \mid P) = p(X_1)H_{X_1}(R) - p(A)H_A(R) - p(B)H_B(R) = \frac{|X_1|}{|U|}(H_{X_1}(R) - H_{X_1}(R \mid P)) \geqslant 0$$　证毕

命题 7 表明，若 R 不变，P 较 Q 细，则相应的条件信息熵 $H_U(R|P)$ 较 $H_U(R|Q)$ 小，下面笔者给出一个例子说明存在 P,Q，且 P 确较 Q 细，但 $H_U(R|P) = H_U(R|Q)$。

设 $U = \{1,2,3,4,5\}$，$U/P = \{\{1,2\},\{3,4\},\{5\}\}$，$U/Q = \{\{1,2,3,4\},\{5\}\}$ 及 $U/R = \{\{1,3,5\},\{2,4\}\}$ 则 $H_U(R|P) = H_U(R|Q)$。

命题 8　设 P,Q,R 均为 U 上的等价关系，如果 $P \subseteq Q$，则 $H_U(P) - H_U(Q) \geqslant H_U(P|R) - H_U(Q|R)$。

证明　由命题 2，$H_U(R|P) + H_U(P) = H_U(P \cap Q) = H_U(P|R) + H_U$® 及 $H_U(R|Q) + H_U(Q) = H_U(Q|R) + H_U$®，两式相减并注意到命题 6 知命题 8 成立。　　证毕

2 等价关系与证据理论的关系

证据理论（Theory of evidence）首先由 Dempster 提出，并由 Shafer 进一步完善[13]，可以用来处理由不知道引起的不确定性问题的理论模型，产生于上世纪 60 年代，它是概率论的进一步扩充，证据理论主要是通过概率分配函数、信任函数及似然函数来表述的，由莫比乌斯（Mobius）变换知道概率分配函数 m、信任函数 Bel 及似然函数 Pl 3 者之间可以相互唯一确定，容易看出它们与粗集的等价类、下近似及上近似有极为类似的关系，从某种意义上说粗集理论继承和发展了证据理论。

设 U 为有限论域，幂集 $P(U)$ 到区间 $[0,1]$ 的满足 $m(\varnothing) = 0$ 及 $\sum_{A \subseteq U} m(A) = 1$ 的函数 m 称为 $P(U)$ 上的概率分配函数，设 R 为 U 上的等价关系，R 确定的商集为 U/R，定义

$$m(X) = \begin{cases} |X|/|U|, X \in U/R \\ 0, X \notin U/R \end{cases}$$

容易验证 m 是 $P(U)$ 上的一个概率分配函数，这样 U 上的每个等价关系按照上述方法对应 U 上的一个概率分配函数 m，设 P,Q 为 U 上的两个等价关系，且 $U/P = \{X_1, X_2, \cdots, X_n\}$ 及 $U/Q = \{Y_1, Y_2, \cdots, Y_m\}$，若 P 对应的概率分配函数为 m_1，Q 对应的概率分配函数为 m_2，$U/(P \cap Q)$ 对应的概率分配函数为 m，很自然的问题是概率分配函数 m 与 m_1, m_2 的关系如何？

命题 9　设 P,Q 为 U 上的两个等价关系，$U/P = \{X_1, X_2, \cdots, X_n\}$，$U/Q = \{Y_1, Y_2, \cdots, Y_m\}$，设 P 对应的概率分配函数为 m_1，Q 对应的概率分配函数为 m_2，若满足 $|X_i \cap Y_j|/|U| = (|X_i|/|U|)(|Y_j|/|U|)$，$(i = 1,2,\cdots,n, j = 1,2,\cdots,m)$，则 $U/(P \cap Q)$ 对应的概率分配函数 m 为 m_1 与 m_2 的正交和 $m_1 \oplus m_2$。

证明　按照概率分配函数正交和的定义，$(m_1 \oplus m_2)(A) = k^{-1} \sum_{X_i \cap Y_j = A} m_1(X_i) m_2(Y_j)$，由于

$$|X_i \cap Y_j|/|U| = (|X_i|/|U|)(|Y_j|/|U|)$$

因此 $X_i \cap Y_j \neq \varnothing$，故

$$k = \sum_{X_i \cap Y_j \neq \varnothing} m_1(X_i) m_2(Y_j) = \sum_{i=1}^{n} m_1(X_i) \sum_{j=1}^{m} m_2(Y_j) = \sum_{i=1}^{n} (|X_i|/|U|) \sum_{j=1}^{m} (|Y_j|/|U|) = 1$$

再次利用 $|X_i \cap Y_j|/|U| = (|X_i|/|U|)(|Y_j|/|U|)$，得到

$$(m_1 \oplus m_2)(A) = \sum_{X_i \cap Y_j = A} m_1(X_i) m_2(Y_j) = \begin{cases} |X_i \cap Y_j|/|U| & \text{当 } A = X_i \cap Y_j \\ 0, & \text{其他} \end{cases}$$

即 $U/(P \cap Q)$ 对应的概率分配函数 m 为 m_1 与 m_2 的正交和 $m_1 \oplus m_2$。　　证毕

$|X_i \cap Y_j|/|U| = (|X_i|/|U|)(|Y_j|/|U|)$ 是一个很苛刻的条件，因此对于一般的等价关系 P,Q 来说 $m = m_1 \oplus m_2$ 不成立。

信息熵可以用于信息表的研究，信息表可以抽象为四元组 $T = \langle U,A,V,f \rangle$，其中 U 是论域，A 是属性集合，V 是属性值的集合，f 是信息函数；信息表可以区分为无决策信息表和有决策信息表，在信息表中若 $A = C \cup D$，C 是条件属性集合，D 是决策属性集合，则称信息表 $T = \langle U, C \cup D, V, f \rangle$ 为决策表。

设 P 为 U 上的一族等价关系，记 $ind(P) = \bigcap_{r \in P} r$，则 $r \in P$ 是 P 中的冗余知识当且仅当 $ind(P) = ind(P - \{\})$，它用条件信息熵来表述为命题 10。

命题 10　设 $T = \langle U,A,V,f \rangle$ 为信息表，$r \in A$，则下列 3 款等价：

1）r 是 A 的冗余属性；

2）$H_U(ind(A)) = H_U(ind(A - \{\}))$；

3）$H_U(r | ind(A - \{\})) = 0$。

证明　$r \in A$ 是 A 中的冗余属性当且仅当 $ind(A) = ind(A - \{\})$，即有 $H_U(ind(A)) = H_U(ind(A - \{\}))$，及 $r \supseteq ind(A - \{\})$，由命题 4 即得 $H_U(r | ind(A - \{\})) = 0$。

命题 10 中 2）说明，冗余属性不能提供任何信息。若 $T = \langle U, C \cup D, V, f \rangle$ 为决策表，决策表是否一致取决于是否有 $ind(C) \subseteq ind(D)$，用条件信息熵来表述就是看 $H_U(ind(C) | ind(D))$ 是否为零。对

于一致决策表有类似于命题10的结论。

3 结论

由于粗糙集理论在数据挖掘方面的应用而倍受关注,粗糙集理论的应用研究得到了长足发展。把信息熵用于粗糙集的研究,因而定义了等价关系上的信息熵。约简是粗糙集用于数据分析的重要概念。在大数据集下约简的有效计算问题是粗糙集应用于实际系统时必然面临的问题。为解决这些问题,现在许多的研究工作集中在寻求有效的约简算法和对经典粗糙集理论的扩展上。有许多文献讨论了概率的信息熵,而本文阐述了等价关系的粗细对信息熵的影响,笔者所讨论的基于等价关系的信息熵比概率的信息熵更直观,并能在信息系统的约简方面发挥重要作用。同时通过对等价关系与证据理论之间的关系研究,笔者对Pawlak粗糙集有了更深入的理解。另外,本文用具体的例子来说明抽象的数学命题,使之更容易理解。

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982,11:341-356.

[2] Yao Y Y,Lingras P J. Interpretations of belief functions in the theory of rough sets[J]. Information Sciences,1998,104:81-106.

[3] 刘清. Rough集及Rough推理[M]. 北京:科学出版社,2001.

[4] Intan R,Mukaidono M. Generalization of rough sets and its applications in information system[J]. Intelligent Data Analysis,2002,6:323-339.

[5] 雷晓蔚. 粗集理论的矩阵方法[J]. 计算机工程与应用,2006,42(17):73-75.

[6] 朱小飞,卓丽霞,彭建华. 一种基于分布特征的连续属性离散化方法[J]. 西南师范大学学报(自然科学版),2006,31(2):107-110.

[7] 朱雪龙. 应用信息论基础[M]. 北京:清华大学出版社,2001.

[8] 苗夺谦,王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报,1999,10(2):113-116.

[9] 王国胤. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766.

[10] 李玉榕. 粗糙集理论中不确定性的粗糙信息熵表示[J]. 计算机科学,2002,29(5):101-103.

[11] 赵晓雨. 粗集模型的特征函数表示[J]. 重庆师范大学学报(自然科学版),2007,24(4):54-57.

[12] 朱树人,李伟琴. 入侵检测技术研究[J]. 计算机工程与设计,2001,22(4):13-17.

[13] Shafe G. A mathematical theory of evidence[M]. Princeton:Princeton University Press,1976.

The Information Entropy and Probability Assignment Based on Equivalent Relations

ZHAO Xiao-yu, ZHOU Run-zhen

(College of Applied Science and Technology, Chongqing University of Arts and Sciences, Chongqing 402160, China)

Abstract: The rough set theory, proposed by Pawlak in 1982, is based on equivalence relation. Recently, the equivalence relation becomes more significant because of rough sets. Information entropy of an equivalence relation has many regular and interesting properties. In this paper we study the basic properties of information entropy and probability of assignment based on equivalent relation. This study gives a theoretical basis for the applications of information entropy based on equivalent relation. For example, on a reduction in information systems, the information entropy based on equivalent relation has an important effect. This paper demonstrates two main lemmas. Firstly, we focus on the influence of the inclusion of equivalent relation on the information entropy, which is explained through 8 propositions. Secondly, we also investigate the relation between the Shafer's evidence theory and equivalent relation. The evidence theory is mainly stated through the probability distribution function, and the function of trust and likelihood. In a sense, the rough set theory is inherited and has developed the evidence theory . Moreover, the discussion in this article is limited in the domain $U=\{u_1, u_2, \cdots, u_{|U|}\}$, by using specific examples to illustrate abstract mathematical proposition, in order to make it more easier to understand.

Key words: equivalent relation; information entropy; rough set; intrusion detection

(责任编辑　游中胜)