文章编号: 1001-0920 (2015) 06-1099-07 DOI: 10.13195/j.kzyjc.2014.0331

基于条件概率的粗糙集不确定性度量

黄国顺[a], 曾凡智[b], 文 翰[a]

(佛山科学技术学院 a. 理学院，b. 电子信息工程学院，广东 佛山 528000)

摘 要: 通过语义分析, 提出一种修正的粗糙集不确定性度量公理化定义. 首先, 对该定义的数学特征进行分析, 提出两种基于条件概率的粗糙集不确定性度量方法; 然后, 证明它们满足所提出的公理化定义, 并导出相应的知识不确定性度量, 发现其中一个是现有条件信息熵, 另一个与确定性度量形成互补关系. 设计算例对各种不确定性度量进行比较分析, 验证了所提出的度量公式与不确定性语义保持一致.

关键词: 不确定性度量；公理化定义；边界域；条件概率

中图分类号: TP18 **文献标志码:** A

Uncertainty measures of rough set based on conditional possibility

HUANG Guo-shun[a], *ZENG Fan-zhi*[b], *WEN Han*[a]

(a. Science School，b. Electronics and Information Engineering School，Foshan University，Foshan 528000，China. Correspondent：HUANG Guo-shun，E-mail：fshgs_72@163.com)

Abstract: By a semantic analysis to the uncertainty measure of rough set, an improved axiomatic definition of the uncertainty measure for the rough set is proposed. Firstly, based on the analysis of mathematical characters of axiomatic definition, two new uncertainty measures based on conditional possibility are proposed. Then, it is proved that they are the uncertainty measures under the axiomatic definition, and the corresponding uncertainty measuring formulas of knowledge are derived, respectively. It is found that one of them is just the existing conditional information entropy, the other has a complement relationship with the certainty measure. An example is given to compare the uncertainty measures, which illustrates that the proposed formulas are consistent with the semantics of uncertainty for the rough set.

Keywords: uncertainty measure；axiomatic definition；boundary region；conditional possibility

0 引 言

随着学者们对粗糙集理论研究的深入, 越来越发现不确定性度量问题在粗糙集理论研究中的作用, 关于不确定性度量问题的研究已经成为粗糙集理论研究中的热点.

目前对于粗糙集不确定性度量的研究主要有3种方法: 基于纯粗糙集方法; 基于信息理论方法; 基于模糊熵方法. 在第1种方法中, Pawlak[1]利用粗糙集的上、下近似集, 用粗糙度来度量完备信息系统中粗糙集的不确定, Dai 等[2]针对不完备信息系统提出了基于粗糙度的不确定性度量方法. 由于粗糙度只依赖于正区域和边界域, 与负域无关, 导致有时即使边界域分离出负域中的知识颗粒, 但其值保持不变, 这不符合人们的直觉. Beaubouef 等[3-4]通过将粗糙度与知识粒度做积, 定义了一种粗糙熵来度量粗糙集的不确定性, 但这类方法仍会产生一些问题, 即与待描述集合无关的知识颗粒的细分会导致知识粗糙熵变小, 这同样不合理. 在第2种方法中, 李健等[5]给出了完备信息系统的基于信息熵的不确定性度量方法. 针对不完备信息系统, 多位学者给出了基于信息熵及其变形公式的不确定性度量方法, 如 Liang 等[6-7]给出的信息熵和粗糙熵方法, Bianucci 等[8-9]提出了 co-entropy, Qian 等[10]提出了混合熵等. Dai 等[11-12]基于信息熵方法研究了集值信息系统和区间值决策系统的不确定性度量问题; Chakrabary 等[13]首先基于模糊度研究了粗糙集的不确定性度量问题; 王国胤等[14]研究了在不同知识粒度下的粗糙集不确定性度量问题, 提出一种基于信息熵的模糊度度量方法; Wei 等[15]系统地研究了模糊熵与粗糙熵之间的区别和联系, 从中筛选出一些适用于粗糙集不确定性度量的模糊熵方法. 以上方

收稿日期: 2014-03-11；**修回日期:** 2014-09-10.

基金项目: 广东省高等学校科技创新重点项目(2014KTSCX152).

作者简介: 黄国顺(1972−), 男, 副教授, 博士, 从事粗糙集、粒度计算的研究；曾凡智(1965−), 男, 教授, 博士生导师, 从事数据库理论、数据挖掘等研究.

法均存在非常明显的缺陷: 首先, 模糊熵本身存在诸多限制, 例如隶属函数在0.5处达到最大值, 隶属函数关于0.5对称等; 其次, 不是每一个模糊熵方法都能用来度量粗糙集的不确定性问题, 这恰好说明模糊熵与粗糙集不确定性的语义不相一致, 而有些模糊熵能用于刻画粗糙集的不确定性度量问题, 只能说明两者之间有交集.

胡军等[16]给出了粗糙集不确定性度量必须满足的一些约束准则, 其中非负性准则条件过于宽松, 导致没有不确定性的粗糙集(即精确集)的不确定性度量可能不为0, 与人们的直觉和习惯相冲突. 由于粗糙集的不确定性主要来自边界域, 程玉胜等[17]基于边界域讨论了粗糙集的不确定性度量问题; Wei 等[15]从边界域出发, 给出一种基于边界域的粗糙熵公理化定义, 但该定义是基于模糊集提出来的, 难以与完备信息系统的粒度计算模式相适应. 本文首先加强粗糙集不确定性度量准则边界条件, 强调当且仅当边界域为空时其不确定性取值才为0; 然后在粒度计算模式下, 给出粗糙集不确定性度量的单调性约束条件, 并对这些准则的数学含义进行分析, 提出两种基于条件概率的粗糙集不确定性度量方法, 证明它们满足本文给出的不确定性度量准则, 并导出相应的知识不确定性度量.

1 相关基本概念

定义 1[1] 设信息系统

$$\mathrm{IS}=\langle U,V,f,A\rangle.$$

其中: U 为一组对象的非空有限集合, 称为论域; A 为有限的属性集; $V=\bigcup\limits_{a\in A}V_a$, V_a 为属性 a 的值域; $f:U\times A\to V$ 为信息函数. 对 U 上的任意属性集 $P\subseteq A$, 定义不可分辨关系为

$$\mathrm{ind}(P)=\{(x,y)\in U^2|\forall\, a\in P, f(x,a)=f(y,a)\}.$$

关系 $\mathrm{ind}(P)$ 构成 U 的一个划分, 记作 $U/\mathrm{ind}(P)$, 简记为 U/P. U/P 中的任一元素 $[x]_P=\{y|\forall\, a\in P, f(x,a)=f(y,a)\}$ 称为等价类. 若 $A=C\bigcup D$, C 为有限的条件属性集, D 为有限的决策属性集, $C\bigcap D=\varnothing$, $V=\bigcup\limits_{a\in C}V_a$, $f:U\times(C\bigcup D)\to V$, 则称 IS 为决策信息系统, 记作 DIS.

定义 2[1] 设 $\mathrm{IS}=\langle U,V,f,A\rangle$, 对于 $\forall P\subseteq A$, $X\subseteq U$, 称 $\underline{P}X=\{x\in U:[x]_P\subseteq X\}$ 为 X 关于 P 的下近似集, $\overline{P}X=\{x\in U:[x]_P\bigcap X\neq\varnothing\}$ 为 X 关于 P 的上近似集.

记粗糙集 X 在知识 U/P 中的边界域为 $\mathrm{BN}_P(X)=\overline{P}X-\underline{P}X$. 显然, 粗糙集 X 的上、下近似将论域分割成3个区域, 即正区域、边界域和负域. 其中: 正区域 $\mathrm{POS}_P(X)=\underline{P}X=\bigcup\{X_i|X_i\in U/P\bigwedge X_i\subseteq X\}$, 边界域 $\mathrm{BN}_P(X)=\overline{P}X-\underline{P}X$, 负域 $\mathrm{NEG}_P(X)=U-\overline{P}X$.

为了刻画出粗糙集的不确定性问题, Pawlak[1]给出集合 X 关于知识 U/P 的粗糙度为

$$\rho_P(X)=\frac{|\mathrm{BN}_P(X)|}{|\overline{P}X|}. \tag{1}$$

显然, 粗糙度只依赖于正区域和边界域, 与负域无关, 这导致它对粗糙集的不确定性度量不够灵敏, Beanbouef 等[3-4]各自通过将粗糙度与知识粒度做积, 得到一类基于知识粒度的粗糙集不确定性度量公式.

定义 3[4] 设 $\mathrm{IS}=\langle U,V,f,A\rangle$, P、$Q\subseteq A$. 当且仅当对于任意的非空 $X_i\in U/P$, 存在 $Q_j\in U/Q$, 使得 $X_i\subseteq Q_j$ 时, 称 Q 粗于 P (或 P 细于 Q), 并记作 $U/P\preceq U/Q$. 如果 $U/P\preceq U/Q$, 且存在某个 $X_{i_0}\in U/P, Q_{j_0}\in U/Q$, 使得 $X_{i_0}\subset Q_{j_0}$, 则称 Q 严格粗于 P (或 P 严格细于 Q), 并记作 $U/P\prec U/Q$.

定义 4[4] 设 $\mathrm{IS}=\langle U,V,f,A\rangle$, $P\subseteq A$, 则 X 关于知识 U/P 的粗糙熵定义为

$$R_P(X)=\rho_P(X)\mathrm{GK}(P), \tag{2}$$

其中 $\mathrm{GK}(P)=\sum\limits_{i=1}^{m}|X_i|^2/|U|^2$.

定义 5[18] 设 $\mathrm{IS}=\langle U,V,f,A\rangle$, $P\subseteq A$, 记 $U/P=\{X_1,X_2,\cdots,X_m\}$, 则 P 在论域 U 上信息熵定义为

$$H(P)=-\sum_{X_i\in U/P}p(X_i)\log_2 p(X_i), \tag{3}$$

其中 $p(X_i)=|X_i|/|U|$.

定义 6[18] 给定 $\mathrm{DIS}=\langle U,V,f,C\bigcup D\rangle$, 设 $P\subseteq C$, $U/P=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{D_1,D_2,\cdots,D_r\}$, 则 D 相对 P 的条件熵定义为

$$H(D|P)=-\sum_{X_i\in U/P}p(X_i)\sum_{D_j\in U/D}p(D_j|X_i)\log_2 p(D_j|X_i). \tag{4}$$

其中: $p(X_i)=|X_i|/|U|$; $p(D_j|X_i)=|X_i\bigcap D_j|/|X_i|$; $i=1,2,\cdots,m$; $j=1,2,\cdots,r$.

2 粗糙集的不确定性度量

粗糙集 X 在知识空间 U/P 下的不确定性度量, 应该满足如下几个约束条件:

1) 只要 X 在知识空间 U/P 中是确定集, 它的不确定性度量值要求达到最小值0, 即使 U/P 再进一步细分其值也不会变小;

2) 与 X 无关的部分知识即使进一步细分, 还是与 X 无关, 不会影响它的不确定性度量值;

3) 不确定性度量只受边界域的影响, 若两个边界域同构(有相同的块数和颗粒结构), 则对应的不确定性度量值应该相等;

4) 随着边界域划分变细, 不确定性度量值变小, 如果边界域划分变细过程中分离出正区域或负域中的知识颗粒, 则不确定性度量值会严格变小.

根据以上语义分析, 提出如下不确定性度量公理化定义.

定义 7　设 $\mathrm{IS}=\langle U,V,f,A\rangle$, $P\subseteq A$, $X\subseteq U$, $\Pi(U)$ 是 U 上所有划分的全体, $P(U)$ 是 U 的幂集, 若存在 $\Pi(U)\times P(U)$ 到实数集的映射函数 $E_r(X|P):\Pi(U)\times P(U)\to R^1$ 满足如下条件, 则称 $E_r(X|P)$ 为粗糙集 X 在知识 U/P 下的不确定性度量:

1) 非负性. $E_r(X|P)\geqslant 0$, 当且仅当 $\mathrm{BN}_P(X)=\varnothing$ 时, 有

$$E_r(X|P)=0.$$

2) 不变性. 若 $\mathrm{BN}_{P_1}(X)/P_1=\mathrm{BN}_{P_2}(X)/P_2$, 则

$$E_r(X|P_1)=E_r(X|P_2).$$

3) 单调性. 若 $U/P_1\prec U/P_2$, 则

$$E_r(X|P_1)\leqslant E_r(X|P_2);$$

若 $U/P_1\prec U/P_2$, 且 $|\mathrm{BN}_{P_1}(X)|<|\mathrm{BN}_{P_2}(X)|$, 则

$$E_r(X|P_1)<E_r(X|P_2).$$

与文献 [16] 的定义相比, 定义 7 加强了非负性的约束条件, 强调当且仅当 $\mathrm{BN}_P(X)=\varnothing$ 时, X 在知识空间 U/P 中的不确定性度量值为 0 且达到最小, 从而避免了精确集的不确定性度量值不为 0 的情形发生. 不变性则强调不确定性只来自边界域, 同构的边界域具有相同的不确定性度量值, 从而说明不确定性度量只与知识颗粒的基数有关. 单调性强调不确定性度量会随着知识划分变细而变小, 同时强调当边界域分裂出确定区域(正区域或负域)知识颗粒时, 其不确定性度量值会严格变小.

由于粗糙度 $\rho_P(X)$ 的计算与负域无关, 有时即使分离出与 X 无关的颗粒, 但其值保持不变, 对不确定性的变化情况反应不灵敏; 粗糙熵 $R_P(X)$ 则反应过度, 即只要知识粒度变细, 哪怕与 X 无关的知识粒度变细都会导致粗糙熵变小, 这是不合理的, 具体反例如下.

例 1　设

$$U=\{x_1,x_2,x_3,x_4,x_5,x_6\},$$
$$X=\{x_2,x_5\},$$
$$U/P_1=\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6\}\},$$
$$U/P_2=\{\{x_1,x_2\},\{x_3,x_4\},\{x_5,x_6\}\}.$$

显然, $U/P_2\prec U/P_1$ 且 $|\mathrm{BN}_{P_2}(X)|<|\mathrm{BN}_{P_1}(X)|$, 但 $\rho_{P_1}(X)=\rho_{P_2}(X)=1$, 违反了定义 7 的单调性条件.

例 2　设

$$U=\{x_1,x_2,x_3,x_4,x_5,x_6\},$$
$$X=\{x_1,x_2,x_3\},$$
$$U/P_3=\{\{x_1,x_2\},\{x_3,x_4\},\{x_5,x_6\}\},$$
$$U/P_4=\{\{x_1,x_2\},\{x_3,x_4\},\{x_5\},\{x_6\}\}.$$

$\mathrm{BN}_{P_3}(X)/P_3=\mathrm{BN}_{P_4}(X)/P_4$, 但 $R_{P_4}(X)<R_{P_3}(X)$, 违反了定义 7 的不变性条件.

下面对定义 7 的数学含义进行分析.

条件 1) 中, $\mathrm{BN}_P(X)=\varnothing$, 即对任意的 $x\in U$, 有 $p(X|[x]_P)=1$ 或 $p(X|[x]_P)=0$.

条件 2) 表明, 不确定性度量只与边界域及其划分有关, 如果两边界域相同, 且具有相同的划分, 则它们的不确定性度量值相同, 此时对上述两边界域中的任意 x, X 关于 $[x]_P$ 的条件概率 $p(X|[x]_P)$ 相同.

条件 3) 给出的是知识划分细分对不确定性的影响. 如果细分的只是确定域(即正区域或负域), 即使它被进一步细分, 也不会影响边界域(包括大小和结构), 从而保证不确定性度量值不变, 此时各知识颗粒的条件概率在细分前后保持不变; 如果细分的是边界域, 且分离出一些正区域或负域中的知识颗粒, 此时 $|\mathrm{BN}_{P_1}(X)|<|\mathrm{BN}_{P_2}(X)|$, 意味着存在元素 $x_0\in\mathrm{BN}_{P_2}(X)$, 使得 $p(X|[x_0]_{P_1})\neq p(X|[x_0]_{P_2})$, 此时细分前后条件概率发生了变化, 不确定性度量变小. 因此可以发现, 不确定性度量与边界域细分前后的条件概率变化紧密相关, 可考虑将不确定性度量看作是条件概率 $p(X|[x]_P)$ 的某种非负函数, 同时要求它满足如下边界条件:

1) 当且仅当 $p(X|[x]_P)=1$ 或 $p(X|[x]_P)=0$ 时, $E_r(X|P)=0$;

2) 随知识划分变细而变小, 随边界域基数变小而严格变小.

基于上述分析, 可构造出基于条件概率的累加型粗糙集不确定性度量方法.

定义 8　设 $\mathrm{IS}=\langle U,V,f,A\rangle$, $P\subseteq A$, $X\subseteq U$, $U/P=\{X_1,X_2,\cdots,X_m\}$, 定义如下公式:

$$E_{r_1}(X|P)=-\sum_{i=1}^{m}\frac{|X_i\bigcap X|}{|U|}\log_2\frac{|X_i\bigcap X|}{|X_i|}. \quad (5)$$

定理 1　$E_{r_1}(X|P)$ 是定义 7 下的不确定性度量.

证明　1) 非负性. 显然有 $E_{r_1}(X|P)\geqslant 0$, 且当且仅当对任意 $X_i\in U/P$, 有 $p(X|X_i)=0$ 或 $p(X|X_i)=1$ 时, $E_{r_1}(X|P)=0$. 分如下两种情况讨论:

① 若 $X=\varnothing$, 则显然有 $\mathrm{BN}_P(X)=\varnothing$;

② 若 $X\neq\varnothing$, 则存在 $X_{i_0}\in U/P$, 使得 $X_{i_0}\bigcap X\neq\varnothing$, 从而必有 $p(X|X_{i_0})=1$, 这意味着只要 $X_{i_0}\bigcap X\neq\varnothing$, 就有 $X_{i_0}\subseteq X$, 从而 $\overline{P}X=\underline{P}X$, 即 $\mathrm{BN}_P(X)=\varnothing$.

反之, 若 $\mathrm{BN}_P(X)=\varnothing$, 则显然有 $E_{r_1}(X|P)=0$.

2) 不变性显然成立.

3) 单调性. 如果$U/P_1 \prec U/P_2$, 则$E_{r_1}(X|P_1) \leqslant E_{r_1}(X|P_2)$.

假设$U/P_1 = \{X_1, X_2, \cdots, X_m\}$, $U/P_2 = \{Y_1, Y_2, \cdots, Y_n\}$, 由于$U/P_1$是$U/P_2$的一个细分, 记$T[Y_j] = \{X_i|X_i \subseteq Y_j\}$, $j = 1, 2, \cdots, n$. 设$T[Y_j] = \{X_{j_1}, X_{j_2}, \cdots, X_{j_l}\}$, 则

$$\sum_{k=1}^{l} \frac{p(X_{j_k})}{p(Y_j)} p(X|X_{j_k}) = \frac{|Y_j \bigcap X|}{|Y_j|} = p(X|Y_j).$$

根据$x\log_2 x$的凸性及Jensen不等式, 有

$$E_{r_1}(X|P_1) =$$
$$-\sum_{i=1}^{m} p(X_i)p(X|X_i)\log_2 p(X|X_i) =$$
$$-\sum_{j=1}^{n} p(Y_j) \sum_{k=1}^{l} \frac{p(X_{j_k})}{p(Y_j)} p(X|X_{j_k}) \log_2 p(X|X_{j_k}) \leqslant$$
$$-\sum_{j=1}^{n} p(Y_j) \Big(\sum_{k=1}^{l} \frac{p(X_{j_k})}{p(Y_j)} p(X|X_{j_k}) \cdot$$
$$\log_2 \Big(\sum_{k=1}^{l} \frac{p(X_{j_k})}{p(Y_j)} p(X|X_{j_k}) \Big) \Big) =$$
$$-\sum_{j=1}^{n} p(Y_j)p(X|Y_j)\log_2 p(X|Y_j) =$$
$$-\sum_{j=1}^{n} \frac{|Y_j \bigcap X|}{|U|} \log_2 \frac{|Y_j \bigcap X|}{|Y_j|} =$$
$$E_{r_1}(X|P_2),$$

且等号成立的充要条件是

$$p(X|X_{j_1}) = p(X|X_{j_2}) = \cdots = p(X|X_{j_l}) = p(X|Y_j).$$

若$U/P_1 \prec U/P_2$, 且$|\mathrm{BN}_{P_1}(X)| < |\mathrm{BN}_{P_2}(X)|$, 则存在元素$x_0 \in \mathrm{BN}_{P_2}(X)$, 使得$p(X|[x_0]_{P_1}) \neq p(X|[x_0]_{P_2})$, 根据前述证明过程, 有

$$E_{r_1}(X|P_1) < E_{r_1}(X|P_2).$$

结论成立. □

例3 利用$E_{r_1}(X|P)$分别计算例1和例2, 结果符合定义7的要求, 其中

$$E_{r_1}(X|P_1) = \frac{1}{2} > \frac{1}{6}\log_2 6 = E_{r_1}(X|P_2),$$
$$E_{r_1}(X|P_3) = E_{r_1}(X|P_4) = \frac{1}{6}.$$

下面给出另一种不确定性度量方法.

定义9 设$\mathrm{IS} = \langle U, V, f, A\rangle$, $P \subseteq A$, $X \subseteq U$, $U/P = \{X_1, X_2, \cdots, X_m\}$, 定义如下公式:

$$E_{r_2}(X|P) = \sum_{i=1}^{m} \frac{|X_i \bigcap X|}{|U|} \Big(1 - \frac{|X_i \bigcap X|}{|X_i|}\Big). \tag{6}$$

定理2 $E_{r_2}(X|P)$是定义7下的不确定性度量.

证明 1) 非负性. 显然有$E_{r_2}(X|P) \geqslant 0$. 下面证明当且仅当$BN_P(X) = \varnothing$时, $E_{r_2}(X|P) = 0$.

若$E_{r_2}(X|P) = 0$, 则对于任意的$X_i \in U/P$, 有

$$\frac{|X_i \bigcap X|}{|U|} \Big(1 - \frac{|X_i \bigcap X|}{|X_i|}\Big) = 0.$$

若$X \neq \varnothing$, 则存在$X_{i_0} \in U/P$, 使得$X_{i_0} \bigcap X \neq \varnothing$, 从而$p(X|X_{i_0}) = 1$, 即$X_{i_0} \subseteq X$, 有$\overline{P}X = \underline{P}X$, 所以$\mathrm{BN}_P(X) = \varnothing$. 若$X = \varnothing$, 则显然有$\mathrm{BN}_P(X) = \varnothing$. 因此不论什么情况, 只要$\frac{|X_i \bigcap X|}{|U|}\Big(1 - \frac{|X_i \bigcap X|}{|X_i|}\Big) = 0$, 就有$\mathrm{BN}_P(X) = \varnothing$.

反之, 若$\mathrm{BN}_P(X) = \varnothing$, 则显然有$E_{r_2}(X|P) = 0$.

2) 不变性显然成立.

3) 单调性.

$$E_{r_2}(X|P) = \frac{|X|}{|U|} - \sum_{i=1}^{m} \frac{|X_i|}{|U|} \Big(\frac{|X_i \bigcap X|}{|X_i|}\Big)^2.$$

假设$U/P_1 = \{X_1, X_2, \cdots, X_m\}$, $U/P_2 = \{Y_1, Y_2, \cdots, Y_n\}$, 因$U/P_1$是$U/P_2$的一个细分, 记$T[Y_j] = \{X_i|X_i \subseteq Y_j\}$, $j = 1, 2, \cdots, n$. 设$T[Y_j] = \{X_{j_1}, X_{j_2}, \cdots, X_{j_l}\}$, 根据$x^2$的凸性及Jensen不等式, 仿照定理1的证明过程, 类似有结论$E_{r_2}(X|P_1) \leqslant E_{r_2}(X|P_2)$成立, 且等号成立的充要条件是

$$p(X|X_{j_1}) = p(X|X_{j_2}) = \cdots = p(X|X_{j_l}) = p(X|Y_j).$$

当$U/P_1 \prec U/P_2$, 且$|\mathrm{BN}_{P_1}(X)| < |\mathrm{BN}_{P_2}(X)|$时, 类似地有

$$E_{r_2}(X|P_1) < E_{r_2}(X|P_2).$$

结论成立. □

例4 利用$E_{r_2}(X|P)$分别计算例1和例2, 结果符合定义7的要求, 其中

$$E_{r_2}(X|P_1) = \frac{5}{24} > \frac{7}{36} = E_{r_2}(X|P_2),$$
$$E_{r_2}(X|P_3) = E_{r_2}(X|P_4) = \frac{1}{12}.$$

3 知识的不确定性度量

下面讨论知识的不确定性度量问题, 为此先给出知识的边界域和正区域定义.

定义10 设$\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则知识$U/Q$在知识$U/P$中的边界域$\mathrm{BN}_P(U/Q)$和正区域$\mathrm{POS}_P(U/Q)$分别定义为

$$\mathrm{BN}_P(U/Q) = \bigcup_{j=1}^{n} \mathrm{BN}_P(Y_j), \tag{7}$$

$$\mathrm{POS}_P(U/Q) = \bigcup_{j=1}^{n} \underline{P}Y_j. \tag{8}$$

定理3 给定$\mathrm{IS} = \langle U, V, f, A\rangle$, $Q \subseteq A$. 如果$U/P_1 \prec U/P_2$, 则$\mathrm{BN}_{P_1}(U/Q) \subseteq \mathrm{BN}_{P_2}(U/Q)$.

证明 因为$U/P_1 \prec U/P_2$, 所以对任意的$x \in U$, 有$[x]_{P_1} \subseteq [x]_{P_2}$, 从而对任意的$Y_j \in U/Q$, 有$\mathrm{BN}_{P_1}($

$Y_j) \subseteq \mathrm{BN}_{P_2}(Y_j)$, 结论成立. □

由粗糙集的不确定性度量可导出相应的知识不确定性度量.

定义 11　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则知识 U/Q 在知识 U/P 中的不确定性度量定义为

$$E_{r_1}(Q|P) = -\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|X_i \bigcap Y_j|}{|U|}\log_2\frac{|X_i \bigcap Y_j|}{|X_i|}. \quad (9)$$

定理 4　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则有:

1) 当且仅当 $\mathrm{BN}_P(U/Q) = \varnothing$ 时, $E_{r_1}(Q|P) = 0$.

2) 如果 $\mathrm{BN}_{P_1}(U/Q)/P_1 = \mathrm{BN}_{P_2}(U/Q)/P_2$, 则 $E_{r_1}(Q|P_1) = E_{r_1}(Q|P_2)$.

3) 若 $U/P_1 \prec U/P_2$, 则 $E_{r_1}(Q|P_1) \leqslant E_{r_1}(Q|P_2)$; 若 $U/P_1 \prec U/P_2$, 且 $|\mathrm{BN}_{P_1}(U/Q)| < |\mathrm{BN}_{P_2}(U/Q)|$, 则 $E_{r_1}(Q|P_1) < E_{r_1}(Q|P_2)$.

证明略 (根据定理 1 和定义 11 可知结论成立).

由于 $\mathrm{BN}_P(U/Q) = U - \mathrm{POS}_P(U/Q)$, 从而当且仅当 $\mathrm{POS}_P(U/Q) = U$ 时, $E_{r_1}(Q|P) = 0$, 即 $U/P \prec U/Q$. 这意味着 U/P 能完全刻画 U/Q, U/Q 在知识 U/P 中没有不确定性.

定理 4 表明, 知识的不确定性度量随划分变细而变小, 随边界域基数的变小而严格变小.

定理 5　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则有

$$E_{r_1}(Q|P) = H(Q|P).$$

证明

$$E_{r_1}(Q|P) = -\sum_{i=1}^{m}\frac{|X_i|}{|U|}\sum_{j=1}^{n}\frac{|X_i \bigcap Y_j|}{|X_i|}\log_2\frac{|X_i \bigcap Y_j|}{|X_i|} = H(Q|P),$$

结论成立. □

定理 5 表明, 条件熵 $H(Q|P)$ 刻画的是知识 U/Q 在知识 U/P 中的不确定性度量 $E_{r_1}(Q|P)$. 从而揭示出条件信息熵的不确定性本质.

定义 12　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则知识 U/Q 在知识 U/P 中的不确定性度量定义为

$$E_{r_2}(Q|P) = \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|X_i \bigcap Y_j|}{|U|}\Big(1 - \frac{|X_i \bigcap Y_j|}{|X_i|}\Big). \quad (10)$$

定理 6　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则有:

1) 当且仅当 $\mathrm{BN}_P(U/Q) = \varnothing$ 时, $E_{r_2}(Q|P) = 0$.

2) 如果 $\mathrm{BN}_{P_1}(U/Q)/P_1 = \mathrm{BN}_{P_2}(U/Q)/P_2$, 则 $E_{r_2}(Q|P_1) = E_{r_2}(Q|P_2)$.

3) 若 $U/P_1 \prec U/P_2$, 则 $E_{r_2}(Q|P_1) \leqslant E_{r_2}(Q|P_2)$; 若 $U/P_1 \prec U/P_2$, 且 $|\mathrm{BN}_{P_1}(U/Q)| < |\mathrm{BN}_{P_2}(U/Q)|$, 则 $E_{r_2}(Q|P_1) < E_{r_2}(Q|P_2)$.

证明略 (根据定理 2 和定义 12 可知结论成立).

定理 7　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, P、$Q \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Y_1, Y_2, \cdots, Y_n\}$, 则有

$$E_{r_2}(Q|P) = 1 - \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|X_i \bigcap Y_j|^2}{|U||X_i|}.$$

证明

$$E_{r_2}(Q|P) = \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|X_i \bigcap Y_j|}{|U|} - \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|X_i \bigcap Y_j|^2}{|U||X_i|} = 1 - \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|X_i \bigcap Y_j|^2}{|U||X_i|}.$$

结论成立. □

Qian 等[19]提出了一种决策信息系统的确定性度量公式, 定义如下.

定义 13　给定 $\mathrm{DIS} = \langle U, V, f, C \bigcup D\rangle$, 设 $P \subseteq C$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/D = \{D_1, D_2, \cdots, D_r\}$, 则 DIS 的确定性度量 $\alpha(\mathrm{DIS})$ 定义为

$$\alpha(\mathrm{DIS}) = \sum_{j=1}^{r}\sum_{i=1}^{m}\frac{|X_i \bigcap D_j|^2}{|U||X_i|}. \quad (11)$$

特别地, 对于决策信息系统, 取 $Q = D$, 则

$$E_{r_2}(D|P) = 1 - \alpha(\mathrm{DIS}).$$

这表明 $E_{r_2}(D|P)$ 与 $\alpha(\mathrm{DIS})$ 构成互补关系, 它们的和为常数 1.

4　不确定性度量比较分析

文献 [3] 给出了一种基于粗糙度和知识粒度乘积形式的粗糙熵定义.

定义 14　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $P \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$, $X \subseteq U$, 则 X 在 U/P 中的粗糙熵定义为

$$E_P(X) = -\rho_A(X)\sum_{i=1}^{m}\frac{|X_i|}{n}\log_2\frac{1}{|X_i|}, \quad (12)$$

其中 $n = |U|$.

文献 [15] 给出了多种基于模糊熵的粗糙集不确定性度量公式. 限于篇幅, 摘录其中一个与本文方法进行比较.

设 $R^+ = [0, +\infty)$, $F(U)$ 表示论域 U 上的所有模糊集集合, 对于任意 $\tilde{X} \in F(U)$, 记

$$\tilde{X} = \frac{\mu_{\tilde{X}(x_1)}}{x_1} + \frac{\mu_{\tilde{X}(x_2)}}{x_2} + \cdots + \frac{\mu_{\tilde{X}(x_n)}}{x_n},$$

$\tilde{X}^c$ 表示 $\tilde{X}$ 的补集, 即 $\mu_{\tilde{X}^c}(x) = 1 - \mu_{\tilde{X}}(x)$.

Fan 等[20]提出了一种 σ-熵, 归一化后变成如下形式:

$$e_{02}^{\alpha,\beta}(\tilde{X})=\frac{1}{n}\sum_{i=1}^{n}f^{\alpha,\beta}(\mu_{\tilde{X}}(x_i)),\tag{13}$$

其中

$$f^{\alpha,\beta}(x)=\frac{1}{(1-\alpha)\beta}[(x^{\alpha}+(1-x)^{\alpha})^{\beta}-1],$$

$$\alpha>0,\ \alpha\neq 1,\ \beta\neq 0,\ x\in[0,1].$$

知识粒度的细分有正区域的细分、负域的细分和边界域的细分, 而边界域的细分又分为条件概率是否改变的两种情形. 因此本文设计如下算例, 其中: U/P_2 只细分了 U/P_1 的正区域, U/P_3 只细分了 U/P_1 的负域, U/P_4 和 U/P_5 只细分了 U/P_1 的边界域, U/P_4 细分前后条件概率发生变化, U/P_5 细分前后条件概率保持不变. 具体见下例.

例 5 假设

$$X=\{x_1,x_2,x_5,x_7,x_8\},$$

$$U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\},$$

$$U/P_1=$$

$$\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6\},\{x_7,x_8\},\{x_9,x_{10}\}\},$$

$$U/P_2=$$

$$\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6\},\{x_7\},\{x_8\},\{x_9,x_{10}\}\},$$

$$U/P_3=$$

$$\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6\},\{x_7,x_8\},\{x_9\},\{x_{10}\}\},$$

$$U/P_4=$$

$$\{\{x_3\},\{x_1,x_2,x_4\},\{x_5,x_6\},\{x_7,x_8\},\{x_9,x_{10}\}\},$$

$$U/P_5=$$

$$\{\{x_1,x_3\},\{x_2,x_4\},\{x_5,x_6\},\{x_7,x_8\},\{x_9,x_{10}\}\}.$$

显然, $U/P_2\prec U/P_1$, $U/P_3\prec U/P_1$, $U/P_4\prec U/P_1$, $U/P_5\prec U/P_1$.

各种不确定性度量公式的计算结果如表 1 所示, 其中 $e_{02}^{\alpha,\beta}(\tilde{X})$ 取参数 $\alpha=3,\beta=0.01$.

表 1 不同不确定性度量计算结果的比较

	P_1	P_2	P_3	P_4	P_5
$\rho_{P_i}(X)$	0.75	0.75	0.75	0.714	0.75
$R_{P_i}(X)$	0.21	0.195	0.195	0.157	0.15
$E_{P_i}(X)$	1.05	0.9	0.9	0.768	0.75
$E_{r_1}(X\|P_i)$	0.3	0.3	0.3	0.217	0.3
$E_{r_2}(X\|P_i)$	0.15	0.15	0.15	0.117	0.15
$e_{02}^{\alpha,\beta}(\tilde{X}_{P_i})$	0.413	0.413	0.413	0.3016	0.413

尽管上述计算结果大小不一, 但有以下几个特点:

1) 当正区域或负域细分时, $\rho_P(X)$、$E_{r_1}(X|P)$、$E_{r_2}(X|P)$ 和 $e_{02}^{\alpha,\beta}(\tilde{X}_P)$ 都保持不变, 这是合理的; 而粗糙熵 $R_P(X)$ 和 $E_P(X)$ 却会严格变小, 这是不合理的. 例如, 虽然 $U/P_3\prec U/P_1$, 但它只是将 U/P_1 中与 X 无关的知识颗粒 $\{x_9,x_{10}\}$ 细分成 $\{x_9\},\{x_{10}\}$, 按照前文的分析, 不确定性度量值保持不变才合理, 然而此时 $R_{P_3}(X)<R_{P_1}(X)$, $E_{P_3}(X)<E_{P_1}(X)$.

2) 当边界域细分时, 如果细分前后条件概率保持不变, 则 $\rho_P(X)$、$E_{r_1}(X|P)$、$E_{r_2}(X|P)$ 和 $e_{02}^{\alpha,\beta}(\tilde{X}_P)$ 保持不变, 而粗糙熵 $R_P(X)$ 和 $E_P(X)$ 会严格变小; 如果细分前后的条件概率发生了变化, 则 $R_P(X)$、$E_P(X)$、$E_{r_1}(X|P)$ 和 $E_{r_2}(X|P)$ 都会严格变小. 对于 $\rho_P(X)$ 和 $e_{02}^{\alpha,\beta}(\tilde{X}_P)$, 在本例中也变小, 如本例的 U/P_4 就是这种情形, 但在例 1 中, 当边界域分离出负域中的知识颗粒时, 条件概率发生了变化, $\rho_P(X)$ 值却保持不变, 这是不合理的; 而对于 $e_{02}^{\alpha,\beta}(\tilde{X}_P)$, 当边界域分离出负域中的知识颗粒时, 细分前后的条件概率发生了变化, 它的值却反而变大, 具体反例见下例.

例 6 设 $U=\{x_1,x_2,\cdots,x_{20}\}$, $X=\{x_1,x_2\}$, $U/P_1=\{\{x_1,x_2,\cdots,x_{10}\},\{x_{11},x_{12},\cdots,x_{20}\}\}$, $U/P_2=\{\{x_1,x_2,\cdots,x_{20}\}\}$, 取 $\alpha=3$, $\beta=0.01$, 则

$$e_{02}^{3,0.01}(\tilde{X}_{P_1})=0.162\,948,$$

$$e_{02}^{3,0.01}(\tilde{X}_{P_2})=0.157\,108.$$

虽然 $U/P_1\prec U/P_2$, 且 $|\mathrm{BN}_{P_1}(X)|<|\mathrm{BN}_{P_2}(X)|$, 但 $e_{02}^{3,0.01}(\tilde{X}_{P_1})>e_{02}^{3,0.01}(\tilde{X}_{P_2})$, 违背了文献 [15] 的定义和本文的单调性原则.

综合例 1、例 5 和例 6 的计算结果可知, 当边界域分离出无关的知识颗粒时, $\rho_P(X)$ 和 $e_{02}^{\alpha,\beta}(\tilde{X}_P)$ 的计算结果不与不确定性语义相一致, 而粗糙熵 $R_P(X)$ 和 $E_P(X)$ 只要有知识划分的细分就会严格变小, 也不与不确定性语义相一致, 只有 $E_{r_1}(X|P)$ 和 $E_{r_2}(X|P)$ 与本文提出的粗糙集不确定性语义保持一致.

5 结 论

本文先对粗糙集的不确定性度量准则进行语义分析, 强调当且仅当边界域为空时, 其不确定性度量值达到最小值 0, 这样可以避免精确集的不确定性度量不为 0. 提出了一种修正的不确定性度量定义, 进一步的数学含义分析表明, 不确定性度量与条件概率紧密相关, 将粗糙集不确定性度量看成满足边界条件的关于条件概率的非负函数, 并在此基础上提出两种基于条件概率的粗糙集和知识的不确定性度量公式. 研究结果表明, 本文提出的一个知识不确定性度量公式与王国胤等[18]提出的条件信息熵等同, 从而揭示了条件信息熵的不确定性度量本质; 另一个知识不确定性度量公式与钱宇华等[19]提出的确定性度量形成互补关系. 设计的两个算例比较了多种不确定性度量公式的优劣, 发现只有本文提出的两种度量方法与粗糙集不确定性语义保持一致, 其他方法都存在不同程度的不足. 下一步的研究重点是基于模糊熵的粗糙集不确定度量方法, 进一步从理论上研究模糊熵与粗糙集不确定性度量问题之间的内在联系.

参考文献(References)

[1] Pawlak Z. Rough set: Theoretical aspects of reasoning about data[M]. Boston: Kluwer Academic, 1991: 16-56.

[2] Dai J H, Xu Q. Approximations and uncertainty measures in incomplete information systems[J]. Information Sciences, 2012, 198(1): 62-80.

[3] Beaubouef T, Petry F E, Arora G. Information-theoretic measures of uncertainty for rough sets and rough relational databases[J]. Information Sciences, 1998, 109(1/2/3/4): 185-195.

[4] Liang Jiye, Wang Junhong, Qian Yuhua. A new measure of uncertainty based on knowledge granulation for rough sets[J]. Information Sciences, 2009, 179(4): 458-470.

[5] 李健, 史开泉. 基于条件粗糙熵的粗集不确定性度量[J]. 系统工程与电子技术, 2008, 30(3): 473-476.
(Li J, Shi K Q. Uncertainty measurement of rough sets based on conditional entropy[J]. Systems Engineering and Electronics, 2008, 30(3): 473-476.)

[6] Liang J Y, Shi Z Z. The information entropy, rough entropy and knowledge granulation in rough set theory[J]. Int J of Uncertainty, Fuzziness and Knowledge-based Systems, 2004, 12(1): 37-46.

[7] Liang J Y, Shi Z Z, Wierman M J. Information entropy, rough entropy and knowledge granulation in incomplete information systems[J]. Int J of General Systems, 2006, 35(6): 641-654.

[8] Bianucci D, Cattaneo G, Ciucci D. Entropies and co-entropies of coverings with application to incomplete information systems[J]. Fundamenta Informatiace, 2007, 75(1/2/3/4): 77-105.

[9] Zhu Ping, Wen Qiaoyan. Entropy and co-entropy of a covering approximation space[J]. Int J of Approximate Reasoning, 2012, 53(4): 528-540.

[10] Qian Y, Liang J. Combination entropy and combination granulation in rough set theory[J]. Int J of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 16(2): 179-193.

[11] Dai J H, Wang W T, Xu Q, et al. Uncertainty measurement for interval-valued decision systems based on extended conditional entropy[J]. Knowledge-based Systems, 2012, 27(1): 443-450.

[12] Dai Jianhua, Tian Haowei. Entropy measures and granularity measures for set-valued information systems[J]. Information Sciences, 2013, 240(1): 72-82.

[13] Chakrabary K, Biswas R, Nanda S. Fuzziness in rough sets[J]. Fuzzy Sets and Systems, 2000, 110(2): 247-251.

[14] 王国胤, 张清华. 不同知识粒度下粗糙集的不确定性研究[J]. 计算机学报, 2008, 31(9): 1588-1598.
(Wang G Y, Zhang Q H. Uncertainty of rough sets in different knowledge granularities[J]. Chinese J of Computers, 2008, 31(9): 1588-1598.)

[15] Wei Wei, Liang Jiye, Qian Yuhua, et al. Can fuzzy entropies be effective measures for evaluating the roughness of a rough set[J]. Information Sciences, 2013, 232(1): 143-166.

[16] 胡军, 王国胤. 粗糙集的不确定度量准则[J]. 模式识别与人工智能, 2010, 23(5): 606-615.
(Hu J, Wang G Y. Uncertainty measure rule sets on rough sets[J]. Pattern Recongtion and Artificial Intelligence, 2010, 23(5): 606-615.)

[17] 程玉胜, 张佑生, 胡学钢. 基于边界域的知识粗糙熵与粗集粗糙熵[J]. 系统仿真学报, 2007, 19(9): 2008-2011.
(Cheng Y S, Zhang Y S, Hu X G. Entropy of knowledge and rough set based on boundary region[J]. J of System Simulation, 2007, 19(9): 2008-2011.)

[18] Wang G Y, Zhao J, An J J, et al. A comparative study of algebra viewpoint and information viewpoint in attribute reduction[J]. Fundamenta Informaticae, 2005, 68(3): 289-301.

[19] Qian Y H, Liang J Y, Li D Y, et al. Measures for evaluating the decision performance of a decision table in rough set theory[J]. Information Sciences, 2008, 178(1): 181-202.

[20] Fan J L, Ma Y L. Some new fuzzy entropy formulas[J]. Fuzzy Sets and Systems, 2002, 128(2): 277-284.

(责任编辑：齐　霁)