文章编号: 1001-0920 (2016) 06-0983-07 DOI: 10.13195/j.kzyjc.2015.0478

基于边界域和知识粒度的粗糙集不确定性度量

黄国顺, 文 翰

(佛山科学技术学院 理学院，广东 佛山 528000)

摘 要: 为了克服现有作积形式不确定性度量方法的缺陷, 基于边界域提出一种用改进粗糙度和知识粒度求和形式的粗糙不确定性度量公式. 与现有方法相比, 它同时考虑了由边界域和知识粗糙性产生的不确定性, 从理论上证明了集成后的不确定性度量值确实比单个影响因素产生的不确定性度量值大, 是一种更加合理的不确定性度量方法. 将该方法推广到基于严凸函数知识粒度情形, 得到一类度量粗糙集不确定性度量方法, 并研究了随划分变细时, 粗糙度、改进粗糙度与所提出方法之间的关系. 最后设计了一组算例对它们进行比较, 比较结果表明, 所提出的方法对划分变细更加敏感.

关键词: 不确定性度量；边界域；粗糙度；知识粒度

中图分类号: TP18 **文献标志码:** A

Uncertainty measures of rough sets based on boundary region and knowledge granularity

HUANG Guo-shun, *WEN Han*

(Science School，Foshan University，Foshan 528000，China．Correspondent：HUANG Guo-shun，E-mail：fshgs_72@163.com)

Abstract: To overcome the defect of product form methods, an uncertainty measure integrating modified rough measure and knowledge granularity is proposed, which can simultaneously describe the uncertainties resulted from the boundary region and knowledge roughness. It is theoretically proved that more values are got after integrating than that produced by only one effective factor, so the proposed method is more reasonable than the existing methods. Then the similar methods are generalized to the knowledge granularity reduced from strictly concave functions. The relationship among the rough measure, the modified rough measure and the proposed method is discussed with respective to the refinement of partition. Finally, a group of examples are designed to compare them. The results show that the proposed method is more sensitive to the refinement of partition.

Keywords: uncertainty measure；boundary region；rough measure；knowledge granularity

0 引 言

随着研究的深入, 人们越发意识到粗糙集不确定性度量在粗糙集理论、粒度计算理论中的重要性, 目前它已广泛应用于属性约简、三支决策和粒度计算等领域[1-3].

粗糙集不确定性度量的研究方法主要分为3类, 即基于纯粗糙集的方法、基于香农信息熵及其变种方法和基于模糊熵函数方法[4]. 在这 3 种方法中, 提出最早的是基于粗糙度的方法[5-6], 由于它的某些缺陷 (如不具有单调性), 使得对这类方法的研究较少. 目前研究较多的是基于信息熵的方法, 其中代表性成果主要有 Beaubouef 等[7]提出的粗糙熵方法、Duntsch 等[8]提出的度量粗糙集预测不确定性方法、Wierman[9]提出的度量知识划分粒度的不确定性方法以及后期多种信息熵变种方法[10-13]. 模糊熵也是一种被研究得较多的方法, 它本质上是先将粗糙集转化为模糊集, 再利用刻画模糊性的方法来度量其不确定性[14-15], 但 Wei 等[16]的研究成果表明, 并不是每一个模糊熵方法都是有效的. 相反, 基于纯粗糙集的方法没有得到应有的重视, 目前基于纯粗糙集的不确定性度量研究主要取得了以下两方面的成果: 1) 对 Pawlak[5]提出的粗糙度进行改进, 提出改进粗糙度, 如 Yao[17]所做的工

收稿日期: 2015-04-17；**修回日期:** 2015-07-23.

基金项目: 广东省自然科学基金项目(2015A030313636)；广东省普通高校特色创新类项目(2014KTSCX152).

作者简介: 黄国顺(1972−), 男, 副教授, 博士, 从事粗糙集、粒度计算和不确定性度量等研究；文翰(1977−), 男, 讲师, 博士, 从事 WEB 挖掘的研究.

作; 2) 将粗糙度和知识粒度作积, 以克服粗糙度不具有单调性的缺点, 比较代表性的研究成果主要有 Beaubouef[7]的工作、Xu 等[18]的工作和 Liang 等[19-20]的工作. 但第2类方法仍会带来至少两方面的问题: 一是反应过度问题, 即只要知识划分变细, 所计算到的不确定度量值就会严格变小, 即使与待描述集合无关的知识颗粒的细分也会严格变小, 这不符合人们的直觉和知识变化的规律; 二是如果将知识粒度取值标准化后限制在区间 [0,1] 上, 则采用作积形式求出来的集成不确定性度量值反而比单个影响因素产生的不确定性值还要小, 这是不合理的.

本文主要针对基于粗糙度的不确定性度量方法展开研究, 首先对文献 [4] 提出的不确定性度量定义进行扩展, 提出一种基于分布同构的粗糙集不确定性度量定义, 并基于边界域提出一种用改进粗糙度和知识粒度求和形式的不确定性度量方法; 然后将该方法推广到基于严凸函数知识粒度上, 得到一类基于边界域和严凸函数知识粒度的不确定性度量方法; 最后研究了本文方法的若干性质以及它与粗糙度、改进粗糙度之间的关系, 并设计一组算例对它们进行比较分析. 与现有方法相比, 本文所提出的方法同时考虑由边界域和知识粗糙性两个影响因素产生的不确定性, 且集成后的不确定性度量值比单个影响因素产生的不确定性度量值大, 从而克服了现有作积方法的弊端.

1 相关基础知识

信息系统是个四元组, 记作 IS $= \langle U, V, f, A\rangle$. 其中: U 是一组对象的非空有限集, 称为论域; A 是有限的属性集; $V = \bigcup\limits_{a\in A} V_a$, V_a 是属性 a 的值域; $f: U\times A \to V$ 是信息函数. 对 U 上的任意属性集 $P\subseteq A$ 定义不可分辨关系 $\mathrm{ind}(P) = \{(x, y) \in U^2 | \forall a \in P, f(x, a) = f(y, a)\}$, 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 中的包含对象 x 的元素 $[x]_P = \{y | \forall a \in P, f(x, a) = f(y, a)\}$ 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, $C \bigcap D = \varnothing$, $V = \bigcup\limits_{a\in C} V_a$, $f: U \times (C \bigcup D) \to V$, 则称之为决策信息系统, 记作DIS.

给定 IS $= \langle U, V, f, A\rangle$, $\forall P \subseteq A, X \subseteq U$, 称 $\underline{P}X = \{[x]_P \in U/P : [x]_P \subseteq X\}$ 为 X 关于 P 的下近似集, $\overline{P}X = \{[x]_P \in U/P : [x]_P \bigcap X \neq \varnothing\}$ 为 X 关于 P 的上近似集. 显然, 粗糙集 X 的上、下近似集将论域分割成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 $\mathrm{BND}_P(X) = \overline{P}X - \underline{P}X$, 负域 $\mathrm{NEG}_P(X) = U - \overline{P}X$.

设 IS $= \langle U, V, f, A\rangle$, $P, Q \subseteq A$, 当且仅当对任意的非空 $X_i \in U/P$, 存在 $Q_j \in U/Q$, 使得 $X_i \subseteq Q_j$, 记作 $U/P \preceq U/Q$, 称 Q 粗于 P (或 P 细于 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), 记作 $U/P \prec U/Q$.

特别地, 如果 P 为恒等关系 ω, 即 $U/\omega = \{\{x\} | x \in U\}$, 则此时划分粒度最细; 如果 P 为全域关系 δ, 即 $U/\delta = \{U\}$, 则此时划分粒度最粗.

如果对于任意 $x \in U$, 有 $[x]_P = [x]_Q$, 则称划分 U/P 和 U/Q 相等, 记作 $U/P = U/Q$, 简记为 $P \approx Q$.

定理1 设 IS $= \langle U, V, f, A\rangle$, $P, Q \subseteq A$, $X \subseteq U$, 如果 $U/P \prec U/Q$, 则 $\underline{Q}(X) \subseteq \underline{P}(X) \subseteq X \subseteq \overline{P}(X) \subseteq \overline{Q}(X)$.

定理2 设 IS $= \langle U, V, f, A\rangle$, $P, Q \subseteq A$, $X \subseteq U$, 如果 $U/P \prec U/Q$, 则 $\mathrm{POS}_Q(X) \subseteq \mathrm{POS}_P(X)$, $\mathrm{BND}_P(X) \subseteq \mathrm{BND}_Q(X)$.

Wierman[9]在讨论知识粒度不确定性度量时, 提出了一种大小同构的概念, 即 $\forall P, Q \subseteq A$, 若存在双射 $h: U/P \to U/Q$, 使得对于任意的 $X_i \in U/P$, 有 $|X_i| = |h(X_i)|$, 则称划分 U/P 和 U/Q 大小同构.

苗夺谦等[21]首次提出计算知识划分的知识粒度计算公式, 定义如下.

定义1 设 IS $= \langle U, V, f, A\rangle$, $P \subseteq A$ 且 $U/P = \{X_1, X_2, \cdots, X_m\}$, 则 U/P 的知识粒度定义为

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

定理3 设 IS $= \langle U, V, f, A\rangle$, $P, Q \subseteq A$, $X \subseteq U$, 如果 $U/P \prec U/Q$, 则 $\mathrm{GK}(P) < \mathrm{GK}(Q)$.

2 现有基于粗糙度的不确定性度量方法

为了刻画出未知概念 X 在已知知识划分 U/P 下的可刻画完备程度, Pawlak[4-5]首先给出集合 X 在划分 U/P 的近似精度 (accuracy measure) 计算公式

$$\alpha_P(X) = \frac{|\underline{P}(X)|}{|\overline{P}X|}. \tag{2}$$

其中: $X \neq \varnothing$, $|\cdot|$ 代表集合的基数. 对于空集 $\varnothing$, 定义 $\alpha_P(\varnothing) = 1$. 显然, $0 \leqslant \alpha_P(X) \leqslant 1$. 根据近似精度, 定义相应的粗糙度为

$$\rho_P(X) = 1 - \alpha_P(X) = \frac{|\mathrm{BND}_P(X)|}{|\overline{P}X|}. \tag{3}$$

然而, 随着知识划分变细, $\rho_P(X)$ 不一定会严格变小, 这不符合人们的直觉, 因此国内外学者对它进行了改进. 其中一种方法是对粗糙度本身进行改进, 由于粗糙度是边界域基数与上近似集基数之比, 缺乏对负域信息变化的刻画能力, 导致有时即使知识划分变细, 粗糙度也依然不变, 具体反例如下[4].

例1 假设 $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $X = \{x_2,$

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

显然, 有 $U/P_2 \prec U/P_1$, 且 $|\mathrm{BND}P_2(X)| < |\mathrm{BND}P_1(X)|$, 但 $\rho_{P_1}(X) = \rho_{P_2}(X) = 1$. 与 U/P_1 相比, U/P_2 虽然没有进一步分离出正区域中的知识颗粒, 但仍然能够得到部分信息, 即 $\{x_3, x_4\}$ 一定不在 X 中. 因此, 可以考虑提出一种同时刻画正域、负域变化的粗糙集不确定性度量方法.

$$\alpha'_P(X) = \frac{|\underline{P}(X) \bigcup \underline{P}(X^C)|}{|\overline{P}(X) \bigcup \overline{P}(X^C)|} = \frac{|\underline{P}(X)| + |\underline{P}(X^C)|}{|U|} = \frac{|\mathrm{POS}_P(X)| + |\mathrm{NEG}_P(X)|}{|U|}, \tag{4}$$

$$\rho'_P(X) = 1 - \alpha'_P(X) = \frac{|\mathrm{BND}_P(X)|}{|U|}, \tag{5}$$

其中 $\rho'_P(X)$ 是 Yao 在文献 [16] 中提出的公式.

另一种对粗糙度的改进方法是将它与知识粒度作积, 其中 Liang 等[19-20]提出的方法最具代表性.

定义 2　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $P \subseteq A$, 则 X 关于划分 U/P 的粗糙熵定义为[19]

$$R_P(X) = \rho_P(X)\mathrm{GK}(P). \tag{6}$$

粗糙熵 $R_P(X)$ 能够有效克服粗糙度的一些缺陷, 它会随知识粒度划分变细而严格变小, 但它同时产生了另外的问题. 例如, 只要知识划分变细, 粗糙熵就会严格变小, 即使与待描述集合无关的知识颗粒的细分也会严格变小, 这不符合人们的直觉.

为了从理论上阐述粗糙集不确定性度量问题, 胡军等[22]给出了几个粗糙集不确定性度量的约束准则. Wei 等[16]基于模糊集和模糊熵研究了粗糙集的不确定性度量问题, 给出一种基于边界域和模糊集形式的公理化定义, 但该公式是基于模糊集形式, 很难与粗糙集的粒度计算形式相适应; 黄国顺等[4]对粗糙集的不确定性语义进行分析, 提出了一种新的基于边界域的粗糙集不确定性度量公理化定义, 同时给出两种基于条件概率的不确定性度量方法. 虽然文献 [4] 中提到大小同构的两边界域 (有相同的块数和颗粒结构) 具有相等的不确定性度量值, 但在具体定义不变性时没有反映这种大小同构的性质.

定义 3　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $P \subseteq A$, $X \subseteq U$, 记 $\pi(U)$ 是 U 上所有划分集合, $P(U)$ 是 U 的幂集. 若存在 $\pi(U) \times P(U)$ 到实数集的映射函数 $\mathrm{ER}(X|P) : \pi(U) \times P(U) \to R^1$ 满足以下条件, 则称 $\mathrm{ER}(X|P)$ 为粗糙集 X 在知识划分 U/P 下的不确定性度量.

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

2) 不变性. 若 $\mathrm{BND}_P(X)/P = \mathrm{BND}_Q(X)/Q$, 则 $\mathrm{ER}(X|P) = \mathrm{ER}(X|Q)$.

3) 单调性. 若 $U/P \prec U/Q$, 则 $\mathrm{ER}(X|P) \leqslant \mathrm{ER}(X|Q)$; 若 $U/P \prec U/Q$ 且 $|\mathrm{BND}_P(X)| < |\mathrm{BND}_Q(X)|$, 则 $\mathrm{ER}(X|P) < \mathrm{ER}(X|Q)$.

定义 3 强调, 当且仅当 $\mathrm{BND}_P(X) = \varnothing$ 时, 即如果 X 在划分 U/P 下是确定集, 则它在划分 U/P 中没有不确定性, 其值为 0 且达到最小, 从而避免了不具有不确定性但其度量值不为 0 的情形发生. 单调性方面, 强调只有当边界域分裂出确定区域 (正域或负域), 其不确定性度量值才会严格变小, 同时体现了知识粗糙性对不确定性度量的影响. 一般划分越细 (即越不粗糙), 不确定性度量值越小, 这是合理的. 文献 [4] 提到大小同构的两边界域 (有相同的块数和颗粒结构) 具有相等的不确定性度量值, 但定义 3 的不变性没有反映这种大小同构性质. 当然, 如果两边界域划分相同, 则它们一定是大小同构的, 但两大小同构边界域的不确定性度量值有可能是不同的. 仍以文献 [4] 的不确定性度量 $\mathrm{ER}_1(X|P)$ 举例说明如下.

定义 4　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[4]:

$$\mathrm{ER}_1(X|P) = -\sum_{i=1}^{m} \frac{|X_i \bigcap X|}{|U|} \log_2 \frac{|X_i \bigcap X|}{|X_i|}. \tag{7}$$

例 2　假设 $X = \{x_1, x_2, x_3, x_5, x_6, x_8\}$, $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$. $U/P_1 = \{\{x_1, x_2\}, \{x_3, x_4, x_5, x_7\}, \{x_6, x_8, x_9, x_{10}\}\}$, $U/P_2 = \{\{x_1, x_2\}, \{x_3, x_4, x_5, x_6\}, \{x_7, x_8, x_9, x_{10}\}\}$.

显然, 划分 U/P_1 和 U/P_2 是大小同构的, 但 $\mathrm{ER}_1(X|P_1) \neq \mathrm{ER}_1(X|P_2)$.

3　基于边界域和知识粒度的度量方法

对于给定的粗糙集 X, 它在划分 U/P 下的不确定性度量值是一个相对的概念, 其大小不但与划分 U/P 有关, 而且与 X 在划分 U/P 中的分布结构有关, 因此仅用划分的大小同构去刻画一个粗糙集的不确定性大小是不够的, 必须带上它的分布结构信息. 为此提出一种分布同构概念.

定义 5　设 $\mathrm{IS} = \langle U, V, f, A\rangle, \forall P, Q \subseteq A, X \subseteq U$, 若存在双射 $h : U/P \to U/Q$, 使得对于任意的 $X_i \in U/P$, 有 $|X_i| = |h(X_i)|$, 且 $|X_i \bigcap X| = |h(X_i) \bigcap X|$, 则称 X 在划分 U/P 和 U/Q 下具有相同的分布, 简称为分布同构, 记作 $U/P \cong U/Q$.

特别地, 如果在边界域满足定义 5, 则称为边界分布同构, 记作 $\mathrm{BND}_P(X)/P \cong \mathrm{BND}_Q(X)/Q$.

定义 6　设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $P, Q \subseteq A$, $X \subseteq U$, 记 $\pi(U)$ 是 U 上所有划分集合, $P(U)$ 是 U 的幂集. 若存在 $\pi(U) \times P(U)$ 到实数集的映射函数 $\mathrm{ER}(X|P)$: $\pi(U) \times P(U) \to R^1$ 满足以下条件, 则称 $\mathrm{ER}(X|P)$ 为

粗糙集 X 在知识划分 U/P 下的不确定性度量.

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

2) 不变性. 若 $\mathrm{BND}_P(X)/P \cong \mathrm{BND}_Q(X)/Q$, 则 $\mathrm{ER}(X|P) = \mathrm{ER}(X|Q)$.

3) 单调性. 若 $U/P \prec U/Q$, 则 $\mathrm{ER}(X|P) \leqslant \mathrm{ER}(X|Q)$; 若 $U/P \prec U/Q$ 且 $|\mathrm{BND}_P(X)| < |\mathrm{BND}_Q(X)|$, 则 $\mathrm{ER}(X|P) < \mathrm{ER}(X|Q)$.

例 3 (例 2 续) 进一步假设 $U/P_3 = \{\{x_1, x_2\}, \{x_3, x_6, x_9, x_{10}\}, \{x_4, x_5, x_7, x_8\}\}$.

显然, $U/P_1 \neq U/P_2 \neq U/P_3$, $\mathrm{BND}_{P_1}(X)/P_1 \neq \mathrm{BND}_{P_3}(X)/P_3$, 但 $U/P_1 \cong U/P_3$, $\mathrm{BND}_{P_1}(X)/P_1 \cong \mathrm{BND}_{P_3}(X)/P_3$. 根据定义 6, X 在划分 U/P_1 和 U/P_3 下具有相同的不确定性值.

定理 4 设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $P \subseteq A$, $X \subseteq U$, 则 $\rho'_P(X)$、$\mathrm{ER}_1(X|P)$ 是满足定义 6 的不确定性度量.

证明 由定理 2 可知, $\rho'_P(X)$ 是满足定义 6 的不确定性度量. 由定义 4 和定义 5 可知, $\mathrm{ER}_1(X|P)$ 也是满足定义 6 的不确定性度量. □

例 4 (例 1 续) $\rho'_{P_1}(X) = 1$, $\rho'_{P_2}(X) = 2/3$. 与 U/P_1 相比, 在 U/P_2 中得到了部分信息, 即 $\{x_3, x_4\}$ 一定不在 X 中, 所以 $\rho'_{P_2}(X) < \rho'_{P_1}(X)$, 这是合理的.

然而, $\rho'_P(X)$ 只刻画出了边界域的不确定性, 并没有刻画出不可区分关系所产生的不确定性.

例 5 假设 $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $X = \{x_2, x_3, x_5\}$, $U/P_1 = \{\{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}\}$, $U/P_2 = \{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, x_6\}\}$.

此时显然有 $\rho'_{P_1}(X) = \rho'_{P_2}(X) = 1$, 它并没有刻画出 U/P_2 是 U/P_1 细分的信息. 因此, 还必须对 $\rho'_P(X)$ 改进, 将边界域产生的不确定性和知识粗糙性产生的不确定性集成起来, 可考虑如下公式:

$$\mathrm{ER_{GK}}(X|P) = \frac{|\mathrm{BND}_P(X)|}{|U|} + \sum_{X_i \in \mathrm{BND}_P(X)/P} \frac{|X_i|^2}{|U|^2} - \frac{1}{\lambda}\frac{|\mathrm{BND}_P(X)|}{|U|} \times \sum_{X_i \in \mathrm{BND}_P(X)/P} \frac{|X_i|^2}{|U|^2}, \tag{8}$$

其中 $\lambda > 1$, 而 $\dfrac{1}{\lambda}$ 可以视为边界域和知识粒度叠加影响因子.

定理 5 设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $P \subseteq A$, $X \subseteq U$, 则 $\mathrm{ER_{GK}}(X|P)$ 是满足定义 6 的不确定性度量.

证明 1) 由于

$$\mathrm{ER_{GK}}(X|P) = 1 - \Big(1 - \frac{|\mathrm{BND}_P(X)|}{|U|}\Big)\Big(1 - \sum_{X_i \in \mathrm{BND}_P(X)/P} \frac{|X_i|^2}{|U|^2}\Big) + \Big(1 - \frac{1}{\lambda}\Big)\frac{|\mathrm{BND}_P(X)|}{|U|} \times \sum_{X_i \in \mathrm{BND}_P(X)/P} \frac{|X_i|^2}{|U|^2}. \tag{9}$$

又因为

$$\frac{|\mathrm{BND}_P(X)|}{|U|} \leqslant 1, \quad \sum_{X_i \in \mathrm{BND}_P(X)/P} \frac{|X_i|^2}{|U|^2} \leqslant 1,$$

显然有 $\mathrm{ER_{GK}}(X|P) \geqslant 0$, 且当且仅当 $\mathrm{BND}_P(X) = \varnothing$ 时, $\mathrm{ER_{GK}}(X|P) = 0$.

2) 如果 $\mathrm{BND}_P(X)/P \cong \mathrm{BND}_Q(X)/Q$, 则 $|\mathrm{BND}_P(X)| = |\mathrm{BND}_Q(X)|$, 且

$$\sum_{X_i \in \mathrm{BND}_P(X)/P} \frac{|X_i|^2}{|U|^2} = \sum_{Y_k \in \mathrm{BND}_Q(X)/Q} \frac{|Y_k|^2}{|U|^2},$$

因此有结论 $\mathrm{ER_{GK}}(X|P) = \mathrm{ER_{GK}}(X|Q)$.

3) 如果 $U/P \prec U/Q$, 根据定理 2 有 $\mathrm{BND}_P(X) \subseteq \mathrm{BND}_Q(X)$, $|\mathrm{BND}_P(X)| \leqslant |\mathrm{BND}_Q(X)|$, 且

$$\sum_{X_i \in \mathrm{BND}_P(X)/P} |X_i|^2/|U|^2 < \sum_{Y_j \in \mathrm{BND}_Q(X)/Q} |Y_j|^2/|U|^2,$$

根据式 (9) 有, $\mathrm{ER_{GK}}(X|P) \leqslant \mathrm{ER_{GK}}(X|Q)$; 如果 $U/P \prec U/Q$, 且 $|\mathrm{BND}_P(X)| < |\mathrm{BND}_Q(X)|$, 则 $\mathrm{ER_{GK}}(X|P) < \mathrm{ER_{GK}}(X|Q)$. □

由定理 5 的证明过程 3) 可知, 当 $U/\delta = \{U\}$ 时, $\mathrm{ER_{GK}}(X|P)$ 达到最大值 $\mathrm{ER_{GK}}(X|\delta) = 2 - \dfrac{1}{\lambda}$.

Wierman[9] 在 1998 年提出了一种知识粒度公理化定义; 梁吉业等[20] 对知识粒度的公理化展开系统研究, 讨论了信息粒度与信息熵之间的关系; Zhu[23] 提出了改进公理化定义; Yao 等[24] 在 2012 年提出了一种基于平均值的信息粒度度量公式; 黄国顺等[25] 提出了一种基于严凸函数导出的知识粒度度量方法. 考虑到定义 1 是一种特殊的知识粒度, 结合式 (8), 将式 (8) 推广到基于严凸函数知识粒度上.

定义 7 设 $\mathrm{IS} = \langle U, V, f, A\rangle$, 若 $\forall P \subseteq A$ 有实数 $G(P)$ 与之对应且满足以下条件, 则称 G 为信息系统 IS 上的知识粒度[25].

1) 不变性. $\forall P, Q \subseteq A$, 若存在双射 $h: U/P \to U/Q$, $\forall X_i \in U/P$, 有 $|X_i| = |h(X_i)|$, 则 $G(P) = G(Q)$.

2) 有界性. $\forall P \subseteq A$, 若 $U/P = \{\{u_i\} | u_i \in U\}$, 则 $G(P) = 0$; 若 $U/P = \{U\}$, 则 $G(P) = 1$.

3) 严格单调性. $\forall P, Q \subseteq A$, 若 $U/P \prec U/Q$, 则 $G(P) < G(Q)$.

定理 6[25] 设 $\mathrm{IS} = \langle U, V, f, A\rangle$, $|U| > 1$, $P \subseteq A$, $U/P = \{X_1, X_2, \cdots, X_m\}$. 若 $\phi(x)$ 是 $[1, +\infty)$ 上的非负严凸函数, 且 $\phi(1) = 0$, $\phi(|U|) = 1$, 则 $G(P) = \sum_{i=1}^{m} \phi(|X_i|)$ 是满足定义 7 的知识粒度函数.

满足定理 6 的非负凸函数很多, 例如 $G_1(P) =$

$\sum\limits_{i=1}^{m}\dfrac{C^2_{|X_i|}}{C^2_{|U|}}$. $G_1(P)=\dfrac{|U|}{|U|-1}\mathrm{GK}(P)-\dfrac{|U|}{|U|-1}$ 是知识粒度 $\mathrm{GK}(P)$ 的一个线性变换, 且具有相同的单调性.

定理 7　设 $\mathrm{IS}=\langle U,V,f,A\rangle$, $P\subseteq A$, $X\subseteq U$. $\phi(x)$ 是满足定理 6 中的非负严凸函数, $G(P)=\sum\limits_{i=1}^{m}\phi(|X_i|)$, $\lambda>1$, 则

$$\mathrm{ER}_\varphi(X|P)=\frac{|\mathrm{BND}_P(X)|}{|U|}+\sum_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)-\frac{1}{\lambda}\frac{|\mathrm{BND}_P(X)|}{|U|}\times\sum_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)\tag{10}$$

是满足定义 6 的不确定性度量.

证明　根据定义 7, 对于任意的 $U/P\prec U/\delta$, 有

$$\frac{|\mathrm{BND}_P(X)|}{|U|}\leqslant 1,$$

$$\sum_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)\leqslant\varphi(|U|)=1,$$

又因为

$$\mathrm{ER}_\varphi(X|P)=1-\left(1-\frac{|\mathrm{BND}_P(X)|}{|U|}\right)\left(1-\sum_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)\right)+\left(1-\frac{1}{\lambda}\right)\frac{|\mathrm{BND}_P(X)|}{|U|}\times\sum_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|),$$

所以 $\mathrm{ER}_\varphi(X|P)\geqslant 0$. 其他证明过程与定理 5 证明类似, 此处略. □

推论 1　$\mathrm{ER}_\varphi(X|P)\geqslant\dfrac{|\mathrm{BND}_P(X)|}{|U|}$, $\mathrm{ER}_\varphi(X|P)\geqslant\sum\limits_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)$.

证明　由于

$$\mathrm{ER}_\varphi(X|P)=\frac{|\mathrm{BND}_P(X)|}{|U|}+\left(1-\frac{|\mathrm{BND}_P(X)|}{\lambda|U|}\right)\times\sum_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|),$$

且 $\lambda>1$, $\dfrac{|\mathrm{BND}_P(X)|}{|U|}\leqslant 1$, 易知

$$\mathrm{ER}_\varphi(X|P)\geqslant\frac{|\mathrm{BND}_P(X)|}{|U|}.$$

类似地, $\mathrm{ER}_\varphi(X|P)\geqslant\sum\limits_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)$. □

推论 1 表明, 由公式 $\mathrm{ER}_\varphi(X|P)$ 集成的不确定性度量值比单个影响因素产生的不确定性 $|\mathrm{BND}_P(X)|/|U|$ 和 $\sum\limits_{X_i\in\mathrm{BND}_P(X)/P}\varphi(|X_i|)$ 的值都要大, 因此公式 (10) 是合理的.

推论 2　假设 $U/P\prec U/Q$, $X\subseteq U$.

1) 如果 $\rho_P(X)<\rho_Q(X)$, 则 $\rho'_P(X)<\rho'_Q(X)$;

2) 如果 $\rho'_P(X)<\rho'_Q(X)$, 则 $\mathrm{ER}_\varphi(X|P)<\mathrm{ER}_\varphi(X|Q)$.

证明　1) 根据定理 2, 由 $U/P\prec U/Q$, 易知 $\mathrm{BND}_P(X)\subseteq\mathrm{BND}_Q(X)$. 如果 $\rho'_P(X)<\rho'_Q(X)$ 不成立, 则有 $|\mathrm{BND}_P(X)|=|\mathrm{BND}_Q(X)|$, 又因为 $\rho_P(X)<\rho_Q(X)$, 从而有 $|\overline{P}(X)|>|\overline{Q}(X)|$, 与定理 1 的结论矛盾.

2) 根据式 (5)、定义 6 和定理 7 即有结论. □

推论 2 的结论表明, 在知识划分变细的过程中, 粗糙度是反应最不敏感的一个不确定性度量, 但如果一旦粗糙度 $\rho_P(X)$ 变小, 则改进粗糙度 $\rho'_P(X)$ 一定变小, $\mathrm{ER}_\varphi(X|P)$ 也随之变小. 反之, 如果对划分变细最敏感的不确定性度量 $\mathrm{ER}_\varphi(X|P)$ 不变, 则 $\rho'_P(X)$ 和 $\rho_P(X)$ 都保持不变.

4　算例分析

下面设计一组算例来比较各种不确定性度量方法的异同, 比较的对象包括粗糙度 $\rho_P(X)$[5]、改进粗糙度 $\rho'_P(X)$[17]、将粗糙度和知识粒度作积方法的代表 $R_P(X)$[19]、基于条件概率的不确定性度量方法的代表 $\mathrm{ER}_1(X|P)$[4]、模糊熵方法 $e^k_{03}(X|P)$[16]和本文方法的代表 $\mathrm{ER}_{\mathrm{GK}}(X|P)$ (其中 $\lambda=2$). 依次将它们简记为 ρ、ρ'、R、ER_1、e^k_{03} 和 $\mathrm{ER}_{\mathrm{GK}}$. 设计的算例主要基于以下几点考虑:

1) 当正域有知识颗粒细分时, 它是否严格变小;

2) 当负域有知识颗粒细分时, 它是否严格变小;

3) 当边界域分离出正域中的颗粒时, 它是否变小;

4) 当边界域分离出负域中的知识颗粒时, 它是否变小;

5) 当边界域不变, 但边界域细分后分布同构时, 它们是否保持不确定性度量值相等.

具体算例如下.

例 6　假设

$X=\{x_1,x_2,x_3,x_5,x_6,x_8\}$,

$U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10},x_{11},x_{12}\}$,

$U/P_1=\{\{x_1,x_2\},\{x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\},\{x_{11},x_{12}\}\}$,

$U/P_2=\{\{x_1\},\{x_2\},\{x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\},\{x_{11},x_{12}\}\}$,

$U/P_3=\{\{x_1,x_2\},\{x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\},\{x_{11}\},\{x_{12}\}\}$,

$U/P_4=\{\{x_1,x_2\},\{x_3,x_6\},\{x_4,x_5,x_7,x_8,x_9,x_{10}\},\{x_{11},x_{12}\}\}$,

$U/P_5 =$

$\{\{x_1, x_2\}, \{x_3, x_4, x_5, x_6, x_7, x_8\}, \{x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$,

$U/P_6 =$

$\{\{x_1, x_2\}, \{x_3, x_5, x_6, x_8\}, \{x_4, x_7, x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$,

$U/P_7 =$

$\{\{x_1, x_2\}, \{x_3, x_4, x_5, x_6\}, \{x_7, x_8, x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$,

$U/P_8 =$

$\{\{x_1, x_2\}, \{x_3, x_5, x_8, x_9\}, \{x_4, x_6, x_7, x_{10}\}, \{x_{11}, x_{12}\}\}$,

$U/P_9 =$

$\{\{x_1, x_2\}, \{x_3, x_6, x_9, x_{10}\}, \{x_4, x_5, x_7, x_8\}, \{x_{11}, x_{12}\}\}$.

表 1 不同不确定性度量比较结果

	ρ	ρ'	R	ER_1	e_{03}^k	$\mathrm{ER}_{\mathrm{GK}}$
P_1	0.800	0.667	0.400	0.333	0.667	0.963
P_2	0.800	0.667	0.389	0.333	0.667	0.963
P_3	0.800	0.667	0.389	0.333	0.667	0.963
P_4	0.600	0.500	0.200	0.264	0.459	0.688
P_5	0.750	0.500	0.250	0.195	0.459	0.688
P_6	0	0	0	0	0	0
P_7	0.800	0.667	0.222	0.270	0.541	0.815
P_8	0.800	0.667	0.222	0.270	0.541	0.815
P_9	0.800	0.667	0.222	0.333	0.667	0.815

显然, $U/P_i \prec U/P_1$, $i = 2, 3, \cdots, 9$, 且 $\mathrm{BND}_{P_7}(X)/P_7 \cong \mathrm{BND}_{P_8}(X)/P_8$, X 在不同划分下的不确定性度量值如表 1 所示. 尽管它们的大小不一, 但还是有如下几个特点.

1) 只要 $\mathrm{BND}_{P_i}(X) = \varnothing$, X 的不确定性度量值都为 0, 表 1 的 U/P_6 即是这样.

2) 因为 $\mathrm{BND}_{P_7}(X)/P_7 \cong \mathrm{BND}_{P_8}(X)/P_8$, 所以 $\mathrm{ER}_{\mathrm{GK}}(X|P_7) = \mathrm{ER}_{\mathrm{GK}}(X|P_8)$, 从而 $\rho'_{P_7}(X) = \rho'_{P_8}(X)$, $\rho_{P_7}(X) = \rho_{P_8}(X)$. 另外, 在本例中, 其他不确定性度量也相等, 如 $\mathrm{ER}_1(X|P_7) = \mathrm{ER}_1(X|P_8)$.

3) 当只有正域或负域发生细分时, 除了 $R_P(X)$ 之外, 其他各种度量值保持不变, 如本例的 U/P_2 和 U/P_3 即是这样. 而 $R_P(X)$ 只要划分变细, 就会严格变小, 包括与 X 无关的负域细分, 它也会严格变小, 如 U/P_3, 这是 $R_P(X)$ 不合理的地方.

4) 当边界域分离出正域或负域中的知识颗粒时, 不确定性度量一般会变小, 本例中 U/P_4 或 U/P_5 就是这样. $\rho_P(X)$ 在本例中也严格变小, 这是因为 $\mathrm{BND}_P(X) \neq U$. 但在例 1 中, 由于边界域等于整个论域, 导致边界域即使分离出负域中的知识颗粒, 它仍然保持不变, 这是 $\rho_P(X)$ 不合理的地方.

5) 当边界域细分前后条件概率不变时, $\mathrm{ER}_{\mathrm{GK}}(X|P)$ 会严格变小, 而基于条件概率的不确定性度量却保持不变, 例如本例中 $U/P_9 \prec U/P_1$, $\mathrm{ER}_{\mathrm{GK}}(X|P_9) < \mathrm{ER}_{\mathrm{GK}}(X|P_1)$, 但 $\mathrm{ER}_1(X|P_9) = \mathrm{ER}_1(X|P_1)$, 这就是它们不一样的地方. 当边界域细分前后条件概率发生改变时, $\mathrm{ER}_{\mathrm{GK}}(X|P)$ 和 $\mathrm{ER}_1(X|P)$ 都会严格变小, 如 U/P_8 就是这样. 而 $\rho'_P(X)$ 只依赖于边界域基数, 不管边界域如何细分, 只要边界域基数不变, $\rho'_P(X)$ 保持不变, 那么它就没法度量边界域知识粗糙性所产生的不确定性信息, $\mathrm{ER}_1(X|P)$ 能部分度量边界域知识粗糙性产生的不确定性信息, 而 $\mathrm{ER}_{\mathrm{GK}}(X|P)$ 只要边界域有细分, 都会严格变小.

6) 一般地, $\mathrm{ER}_{\mathrm{GK}}(X|P)$ 比 $\rho'_P(X)$ 和 $\sum\limits_{X_i \in \mathrm{BND}_P(X)/P} \dfrac{|X_i|^2}{|U|^2}$ 的值要大, 这是由于它集成了由边界域和知识粗糙性产生的两种不确定性, 而 $R_P(X)$ 采用作积的形式虽然集成了产生不确定性的两种影响因素, 但最后得到的值比集成前还小, 这是它不合理之处.

5 结 论

本文主要从纯粗糙集的角度研究了粗糙集的不确定性度量问题, 提出了一种基于边界域和知识粒度的不确定性度量方法. 与现有方法不同, 它将产生粗糙集不确定性的两种因素集成在一起, 所得计算结果比单个影响因素产生的不确定性度量值大, 因此是一种更加合理的不确定性度量方法. 研究结果发现, 由于本文不确定性度量方法是基于边界域的, 正域与负域的划分变细对不确定性度量没有影响, 只有当边界变细, 它才会变小. 与现有的 Pawlak 粗糙度和改进粗糙度相比, 本文建议的方法对划分变细更加敏感, 只要本文建议的方法计算结果不变, 那么粗糙度和改进粗糙度都保持不变, 反之则不一定成立. 下一步, 将对满足不确定性度量定义的纯粗糙集方法、信息熵方法与模糊熵结合起来研究.

参考文献(References)

[1] Sun L, Xu J C, Tian Y. Feature selection using rough entropy-based uncertainty measures in incomplete decision systems[J]. Knowledge-based Systems, 2012, 36(1): 206-216.

[2] Liang J Y, Qian Y H. Information granules and entropy theory in information systems[J]. Science in China Series F: Information Sciences, 2008, 51(10): 1427-1444.

[3] Deng X F, Yao Y Y. A multifaceted analysis of probabilistic three-way decisions[J]. Fundamenta Informaticae, 2014, 132(3): 291-313.

[4] 黄国顺, 曾凡智, 文翰. 基于条件概率的粗糙集不确定性度量[J]. 控制与决策, 2015, 30(6): 1099-1105.
(Huang G S, Zeng F Z, Wen H. Uncertainty measures of rough set based on conditional possibility[J]. Control and

Decision, 2015, 30(6): 1099-1105.)

[5] Pawlak Z. Rough sets[J]. Int J of Computer & Information Sciences, 1982, 11(5): 341-356.

[6] Pawlak Z. Rough sets: Theoretical aspects of reasoning about data[M]. Boston: Kluwer Academic, 1991: 16-17.

[7] 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.

[8] Düntsch I, Gediga G. Uncertainty measures of rough set prediction[J]. Artificial intelligence, 1998, 106(1): 109-137.

[9] Wierman M J. Measuring uncertainty in rough set theory[J]. Int J of General System, 1999, 28(4/5): 283-297.

[10] 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.

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

[12] Qian Y H, 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.

[13] 腾书华, 鲁敏, 杨阿峰, 等. 基于一般二元关系的粗糙集加权不确定性度量[J]. 计算机学报, 2014, 37(3): 649-665.

(Teng S H, Lu M, Yanf A F, et al. A weighted uncertainty measure of rough sets based on general binary relation[J]. Chinese J of Computers, 2014, 37(3): 649-665.)

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

[15] 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.)

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

[17] Yao Y Y. Notes on rough set approximations and associated measures[J]. J of Zhejiang Ocean University: Natural Science, 2010, 29(5): 399-410.

[18] Xu B W, Zhou Y M, Lu H M. An improved accuracy measure for rough sets[J]. J of Computer and System Sciences, 2005, 71(2): 163-173.

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

[20] 梁吉业, 钱宇华. 信息系统中的信息粒与熵理论[J]. 中国科学(E), 2008, 38(12): 2048-2065.

(Liang J Y, Qian Y H. Information granularity and entropy theory in information system[J]. Science in China(E), 2008, 38(12): 2048-2065.)

[21] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.

(Miao D Q, Fan S D. The calculation of knowledge granulation and its application[J]. Systems Engineering — Theory & practice, 2012, 22(1): 48-56.)

[22] 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.)

[23] Zhu P. An improved axiomatic definition of information granulation[J]. Fundamenta Informaticae, 2012, 120(1): 93-109.

[24] Yao Y Y, Zhao Liquan. A Measurement theory view on the granularity of partitions[J]. Information Sciences, 2012, 213(1): 1-13.

[25] 黄国顺, 曾凡智, 陈广义, 等. 基于严凸函数的知识粒度与相对粒度[J]. 模式识别与人工智能, 2013, 26(10): 897-908.

(Huang G S, Zeng F Z, Chen G Y, et al. knowledge granularity and relative granularity based on strictly convex function[J]. Pattern Recognition and Artificial Intelligence, 2013, 26(10): 897-908.)

(责任编辑：齐　霁)