**Computer Science and Application** 计算机科学与应用, 2017, 7(4), 388-397
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74047

# Uncertainty Measures for Continuous-Valued Information Systems

**Xin Xu**

School of Computer and Control Engineering, Yantai University, Yantai Shandong
Email: 1845691121@qq.com

Received: Apr. 16th, 2017; accepted: Apr. 26th, 2017; published: Apr. 30th, 2017

## Abstract

**Approach to uncertainty measures is one of the hottest topics in area of artificial intelligence, which attracts attention from many researchers. Relevant research results have been applied in data mining, decision analysis, pattern recognition and artificial intelligence. In this paper, the uncertainty measures of continuous-valued information systems have been investigated systematically by using binary relation and entropy. Based on the approximation accuracy, knowledge granulation and information entropy in Pawlak rough set theory, we propose the rough accuracy, knowledge granulation and knowledge entropy in continuous-valued information systems. We also have the comparative study about three measures in the paper. The proposed uncertainty measures for continuous-valued information systems could provide the theoretical foundation for knowledge reduction and representation in continuous information systems.**

## Keywords

**Continuous-Valued Information Systems, Rough Sets, Uncertainty Measure, Entropy**

# 连续值信息系统的不确定性度量

**许　鑫**

烟台大学计算机与控制工程学院，山东　烟台
Email: 1845691121@qq.com

收稿日期：2017年4月16日；录用日期：2017年4月26日；发布日期：2017年4月30日

## 摘　要

**不确定性的度量方法是人工智能研究的重要课题之一，受到国内外专家学者的广泛关注，相关研究成果已经成功的应用于数据挖掘，决策分析，模式识别与人工智能领域中。通过二元关系与熵，对连续值信**

文章引用: 许鑫. 连续值信息系统的不确定性度量[J]. 计算机科学与应用, 2017, 7(4): 388-397.
https://doi.org/10.12677/csa.2017.74047

**息系统中的不确定性度量进行了系统研究。基于经典Pawlak粗糙集理论中的近似精度、知识粒度与信息熵，提出了连续值信息系统的粗糙度、知识粒度与知识熵，并对三种度量方式进行了比较分析。三种不确定性度量方式的提出，为连续值信息系统知识约简与表示的研究提供了理论基础。**

## 关键词

**连续值信息系统，粗糙集，不确定性度量，熵**



## 1. 引言

由于客观世界中广泛存在不确定性，不确定性问题的研究工作一直备受关注[1]-[8]。某些应用背景中，知识库中的数据以连续值形式呈现，例如某季节的气温范围、某公司中员工的年龄区间、以及工业批量产品制造的误差范围。连续值决策系统的知识库中描述的知识的属性并非同等重要，其中有些属性是冗余的。冗余的属性不利于对知识的泛化与规则的提取，同时也加大了系统存储的负担，造成了资源的浪费。

度量不确定性的程度称为不确定性度量，最早的度量方法是由Kolmogorov于1933年提出的概率论。随着通信技术的发展，Shannon于1948年提出信息熵的概念，解决了对信息量的度量问题。近些年来，随着粗糙集理论研究的不断深入，粗糙集理论中的不确定性研究成为热点问题。国际上，Düntsch等[1]利用Shannon熵提出了三个模型选择准则，这些标准用于指导人们如何挑选最优条件属性集合来描述一个决策值，并用于刻画粗糙集预测的质量。Wierman [2]从公理化角度出发，给出了一种不确定性度量，称为粒度度量，在五条公理约束下，可以证明其提出的粒度度量与Shannon熵具有相同的形式。Beaubouef [3]应用Shannon熵分别研究了粗糙集中概念的粗糙度和关系数据库中的粗糙度。国内，1997年，苗夺谦[9]将信息熵的概念引入粗糙集理论研究中。文献[10]提出了知识的粗糙性、信息熵与互信息等概念，并讨论了知识粗糙性与信息熵之间的关系。2002年，苗夺谦在文献[11]中通过等价关系的基数定义了知识粒度与分辨度的概念。作为两个概念的应用，分别介绍了重要度在求最小约简、协调度在构造决策树方面的应用。王国胤[12]提出了代数观点与信息观点下的粗糙约简。黄兵等人于2004年在文献[13]中基于一般二元关系提出了信息系统的广义粗糙熵概念。2006年，Liang等在文献[14]中将完备信息系统中的知识粒度、分辨度与信息熵等概念扩展到不完备信息系统中，分别提出了知识粒度、信息熵与粗糙熵等概念，用来度量不完备信息系统中的不确定性。冯琴荣在文献[15] [16]中提出了基于数学期望的知识粒度定义，将每个知识粒看作是一个一维对象，粒的基数看作是其长度，从而针对信息系统定义了知识粒度的测度，该测度定义为划分中粒长度的数学期望。类似于不完备信息系统中的不确定性度量[17]，Xu等人于2009年在文献[18]中提出了序信息系统中的知识粒度、知识熵与知识的不确定性度量。2011年，王国胤等在文献[19]中综述了知识不确定性问题的粒计算模型，从粒计算模型的角度分析了模糊集、粗糙集以及商空间理论模型中的不确定性问题，并对知识不确定性问题的研究工作进行了讨论和总结，对有待研究的重要问题进行了展望。基于扩展的条件信息熵，Dai等[20]在2013年对区间值信息系统中的不确定性度量的进行了相关研究。

经典粗糙集中完备信息系统的不确定性度量有三类：1)信息系统的粗糙度；2)知识粒度(或分辨度)；3)信息熵。通过推广完备信息系统中的不确定性度量，本文在第 3 小节提出了三种连续值信息系统中的不确定性度量：1)连续值信息系统中的粗糙度；2)连续值信息系统中的知识粒度(分辨度)；3)连续值信息系统中的知识熵。在第 4 小节对三种度量方法进行了比较。

## 2. 连续值信息系统中的粒表示

Guan 等在文献[21]中提出了连续值信息系统，采用相似度量方法 $r\left(x_i,x_j\right)=1-\max\left\{\left|c_{ik}\left(x_i\right)-c_{ik}\left(x_i\right)\right|\,|\,c_{ik}\in B\right\}$ 来度量对象 $x_i$ 与 $x_j$ 的相似程度。由相似性度量方法，可给出连续值信息系统中的相似关系为：$R_\beta^B=\left\{\left(x_i,x_j\right)\in U\times U,r_{i,j}^B\left(B\right)=1\right\}$。通过相似关系 $R_\beta^B$，给出连续值信息系统中的两种粒化方式：1)相似类(与对象 $x_i$ 满足相似关系 $R_\beta^B$ 的对象集合)：$T_\beta^B\left(x_i\right)=\left\{x_j\in U\,|\,r\left(x_i,x_j\right)\in R_\beta^B\right\}$；2)极大相容类(满足类中任意两个对象均满足相似关系 $R_\beta^B$ 的最大集合)：$K_\beta^B\left(x_i\right)$，且 $T_\beta^B\left(x_i\right)=\bigcup\left\{K_\beta^B\left(x_i\right)\in CCR_\beta^B\left(x_i\right)\right\}$。将极大相容类作为连续值信息系统中的基本信息粒，给出粗糙上下近似分别为：

$$\overline{ER}_\beta^B(X)=\left\{x\,|\,\exists K\in CCR_\beta^B(x):K\cap X\neq\varnothing\right\};\quad \underline{AR}_\beta^B(X)=\left\{x\,|\,\forall K\in CCR_\beta^B(x):K\subseteq X\right\}。$$

对连续值信息系统更加详细的粒化描述与性质，请参见文献[21]。限于本文篇幅，这里不再一一累述。

## 3. 连续值信息系统的不确定性度量

### 3.1. 连续值信息系统的粗糙度

Pawlak 教授研究近似精度、近似质量时提出了粗糙度的概念。粗糙集的粗糙度通过集合的上、下近似来定义，它充分反映了由于集合边界域的存在所引起的不确定性。类似地，连续值信息系统中的粗糙度用来度量连续值信息系统中的不确定性。

**定义 1** 给定的连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，对任意的 $X\subseteq U$，$B\subseteq C$，$\beta\in[0.5,1]$，集合 $X$ 的粗糙近似精度为：

$$\lambda_\beta^B\left(X\right)=\frac{\left|\underline{AR}_\beta^B\left(X\right)\right|}{\left|\overline{ER}_\beta^B\left(X\right)\right|}=\frac{\left|\underline{AR}_\beta^B\left(X\right)\right|}{|U|-\left|\underline{AR}_\beta^B\left(X\right)\right|}。$$

其中：

$\overline{ER}_\beta^B\left(X\right)=\left\{x\,|\,\exists K\in CCR_\beta^B(x):K\cap X\neq\varnothing\right\}$ 是集合 $X$ 的上近似算子，

$\underline{AR}_\beta^B\left(X\right)=\left\{x\,|\,\forall K\in CCR_\beta^B(x):K\subseteq X\right\}$ 是集合 $X$ 的下近似算子，

$|S|$ 表示集合 $S$ 的基数。

定义连续值信息系统中的粗糙度为：

$$\theta_\beta^B\left(X\right)=1-\lambda_\beta^B\left(X\right)=1-\frac{\left|\underline{AR}_\beta^B\left(X\right)\right|}{\left|\overline{ER}_\beta^B\left(X\right)\right|}=\frac{\left|BR_\beta^B\left(X\right)\right|}{\left|\overline{ER}_\beta^B\left(X\right)\right|}。$$

其中，$BR_\beta^B\left(X\right)=\overline{ER}_\beta^B\left(X\right)-\underline{AR}_\beta^B\left(X\right)$。

粗糙度 $\theta_\beta^B\left(X\right)$ 有如下性质：

1) $\theta_\beta^B\left(X\right)$ 值与连续值信息系统的不确定性成正比，即，$\theta_\beta^B\left(X\right)$ 越大，不确定性越大；反之，$\theta_\beta^B\left(X\right)$ 越小，不确定性越小；

2) $\theta_\beta^B\left(X\right)=0$ 时，$\overline{ER}_\beta^B\left(X\right)=\underline{AR}_\beta^B\left(X\right)$，集合 $X$ 是精确集，不确定度为 0；$0<\theta_\beta^B\left(X\right)<1$ 时，集合

$X$ 是不确定集，不确定度的值 $\theta_\beta^B(X)\in(0,1)$； $\theta_\beta^B(X)=1$ 时， $\underline{AR}_\beta^B(X)=0$，集合 $X$ 是完全不确定集，不确定性为最大值 1。

### 3.2. 连续值信息系统中的知识粒度

为方便引入连续值信息系统中的知识粒度，先给出一些基本定义。

设连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， $A,B\subseteq C$，有

$$T_\beta(A)=\left\{T_\beta^A(x_1),T_\beta^A(x_2),\cdots,T_\beta^A\left(x_{|U|}\right)\right\}，\ T_\beta(B)=\left\{T_\beta^B(x_1),T_\beta^B(x_2),\cdots,T_\beta^B\left(x_{|U|}\right)\right\}。$$

定义连续值信息系统中的二元关系“$\preceq$”，“$\approx$”与“$\prec$”如下：

$T_\beta(A)\preceq T_\beta(B)\Leftrightarrow$ 若对于任意的 $i\in\{1,2,\cdots,|U|\}$，有 $T_\beta^A(x_i)\subseteq T_\beta^B(x_i)$，其中 $T_\beta^A(x_i)\in T_\beta(A)$，$T_\beta^B(x_i)\in T_\beta(B)$。简记为 $A\preceq B$。

$T_\beta(A)\approx T_\beta(B)\Leftrightarrow$ 若对于任意的 $i\in\{1,2,\cdots,|U|\}$，有 $T_\beta^A(x_i)=T_\beta^B(x_i)$，其中 $T_\beta^A(x_i)\in T_\beta(A)$，$T_\beta^B(x_i)\in T_\beta(B)$。简记为 $A\approx B$；

$T_\beta(A)\prec T_\beta(B)\Leftrightarrow T_\beta(A)\preceq T_\beta(B)$ 且 $T_\beta(A)\neq T_\beta(B)$，简记为 $A\prec B$。

若 $A\prec B$，称属性集 $B$ 对应的论域分类比属性集 $A$ 对应的分类粗，或称属性集 $A$ 对应的论域分类比属性集 $B$ 对应的分类细；若 $A\approx B$，称属性集 $B$ 对应的论域分类与属性集 $A$ 对应的分类相等。

**定理 1** 连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$，记 $T=\{T_\beta(A)\mid A\subseteq C\}$，则 $(T,\preceq)$ 是一个偏序集。

**证明**

令 $L,M,N\subseteq C$，有

$$T_\beta(L)=\left\{T_\beta^L(x_1),T_\beta^L(x_2),\cdots,T_\beta^L\left(x_{|U|}\right)\right\},$$

$$T_\beta(M)=\left\{T_\beta^M(x_1),T_\beta^M(x_2),\cdots,T_\beta^M\left(x_{|U|}\right)\right\},$$

$$T_\beta(N)=\left\{T_\beta^N(x_1),T_\beta^N(x_2),\cdots,T_\beta^N\left(x_{|U|}\right)\right\}。$$

1) 对任意的 $x_i\in U$，有 $T_\beta^L(x_i)=T_\beta^L(x_i)$ 成立，因此 $L\preceq L$。

2) 假设 $M\preceq N$ 和 $N\preceq M$。由上面的定义，可得：

$M\preceq N\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使得 $T_\beta^M(x_i)\subseteq T_\beta^N(x_i)$，其中，$T_\beta^M(x_i)\in T_\beta(M)$，$T_\beta^N(x_i)\in T_\beta(N)$；

$N\preceq M\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使得 $T_\beta^N(x_i)\subseteq T_\beta^M(x_i)$，其中，$T_\beta^N(x_i)\in T_\beta(N)$，$T_\beta^M(x_i)\in T_\beta(M)$。

因此，有 $T_\beta^N(x_i)\subseteq T_\beta^M(x_i)\subseteq T_\beta^N(x_i)$，即 $T_\beta^M(x_i)=T_\beta^N(x_i)$。所以，对于任意的 $x_i$，都有 $T_\beta^M(x_i)=T_\beta^N(x_i)$，即 $M\approx N$。

3) 假设 $L\preceq M$ 和 $M\preceq N$。由上面的定义，可得：

$L\preceq M\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使 $T_\beta^L(x_i)\subseteq T_\beta^M(x_i)$，其中 $T_\beta^L(x_i)\in T_\beta(L)$，$T_\beta^M(x_i)\in T_\beta(M)$；

$M\preceq N\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使 $T_\beta^M(x_i)\subseteq T_\beta^N(x_i)$，其中 $T_\beta^M(x_i)\in T_\beta(M)$，$T_\beta^N(x_i)\in T_\beta(N)$。

因此，对于任意的 $i\in\{1,2,\cdots,|U|\}$，有 $T_\beta^L(x_i)\subseteq T_\beta^N(x_i)$，即 $T_\beta^L(x_i)\subseteq T_\beta^N(x_i)$，所以 $L\preceq N$。

考虑到上述三点，$(T,\preceq)$ 是偏序集。

在本章中，将运用这种偏序关系对连续值信息系统中的不确定性进行研究。

Yao 等在文献[22]中给出了信息系统中粒度的一般性定义，为构建和比较知识粒度提供了有利条件。

**定义 2** [22]设信息系统为 $IS=(U,AT,V,f)$，对任意的 $A\subseteq AT$，有 $GD(A)$ 满足：

1) **非负性**：$GD(A)\geq 0$；

2) **恒等性**：对于 $\forall A,B\subseteq AT$，若 $A\approx B$ 时，有

$$GD(A)=GD(B)；$$

3) **单调性**：对于 $\forall A,B\subseteq AT$，若 $A\prec B$ 时，有

$$GD(A)<GD(B)。$$

则称 $GD(A)$ 为信息系统 $IS=(U,AT,V,f)$ 上关于属性集 $A$ 的知识粒度。算子“$\approx$”与“$\prec$”是信息系统 $IS$ 中的粒度偏序关系。在粗糙集理论中，不同的知识粒度实质上是对信息细化的不同层次的平均度量。

为度量完备信息系统中的知识不确定性，苗夺谦等在文献[11]中首先给出了完备信息系统中知识粒度的定义：

**定义 3** 设 $IS=(U,AT,V,f)$ 是一个完备信息系统，$U/IND(A)=\{X_1,X_2,\cdots,X_m\}$，则 $IS$ 关于属性 $A$ 的知识粒度(Knowledge Granularity)定义为

$$GD(A)=\frac{1}{|U|^2}\sum_{i=1}^{m}|X_i|^2。$$

其中，$\sum_{i=1}^{m}|X_i|^2$ 是由 $\bigcup_{i=1}^{m}(X_i\times X_i)$ 决定的等价关系中元素数目。

**定理 2** [23]完备信息系统中的知识粒度 $GD(A)$ 是定义 2 意义下的一个知识粒度。

考虑定义 3，设 $X(x_i)$ 是论域 $U$ 中对象 $x_i$ 的等价类，给出完备信息系统中知识粒度的另外一种表示为：

$$GD(A)=\frac{1}{|U|^2}\sum_{i=1}^{m}|X_i|^2=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|X(x_i)|=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}。$$

$X(x_i)$ 是关于对象 $x_i\in U$ 的等价类，$X(x_i)$ 的基是论域 $U$ 中与对象 $x_i$ 满足等价关系的对象数目。$|U|\times|U|$ 是论域中对象间全部的关系数，且 $GD(A)\in[1/|U|,1]$。

1) 当论域的等价类划分最细，即每个划分仅包含单个元素时，论域中对象最易分辨：

$$GD(A)=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}=\sum_{i=1}^{|U|}\frac{1}{|U|\times|U|}=\frac{1}{|U|}。$$

2) 当论域的等价类划分最粗，即论域的划分是整个论域，论域中对象完全不可分辨：

$$GD(A)=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}=\sum_{i=1}^{|U|}\frac{|U|}{|U|\times|U|}=1。$$

$GD(A)=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}$ 可看作论域划分产生的对象关系和与论域中所有对象总和 $|U|\times|U|$ 的比，所以，易将完备信息系统中基于等价类的知识粒度扩展到连续值信息系统中的知识粒度为：

**定义 4** 连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$，对任意的 $B\subseteq C$，$CIS$ 中属性集 $B$ 的知识粒度定义为：

$$CKG_\beta(B)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|\cup K_\beta^B|=\sum_{i=1}^{|U|}\frac{|\cup K_\beta^B|}{|U|\times|U|}=\sum_{i=1}^{|U|}\frac{|T_\beta^B(x_i)|}{|U|\times|U|}。$$

其中， $K_\beta^B \in T_\beta^B(x_i)$， $\left|\cup K_\beta^B\right|$ 是 $T_\beta^B(x_i)$ 中所有元素并的基，且 $\left|\cup K_\beta^B\right| = \left|T_\beta^B(x_i)\right|$。

考虑 $KG_\beta(B)$ 的取值情况，给出如下三个定理：

**定理 3 (极小值)**连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$， $R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中相对于属性集 $B$ 的粒度最小值是 $1/|U|$，当且仅当 $R_\beta^B = \breve{R}_\beta^B$，其中 $\breve{R}_\beta^B$ 为恒等相似关系，有

$$U/\breve{R}_\beta^B = \left\{T_\beta^B(x_i) = (x_i) : x_i \in U\right\} = \left\{\{x_1\}, \{x_2\}, \cdots, \left\{x_{|U|}\right\}\right\}。$$

**证略。**

**定理 4 (极大值)**连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$， $R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中相对于属性集 $B$ 的粒度最大值是 1，当且仅当 $R_\beta^B = \hat{R}_\beta^B$，其中 $\hat{R}_\beta^B$ 为全域相似关系，有

$$U/\hat{R}_\beta^B = \left\{T_\beta^B(x_i) = U : x_i \in U\right\} = \{U, U, \cdots, U\}。$$

**证略。**

**定理 5 (边界性)**连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$， $R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中相对于属性 $B$ 的粒度存在边界为：

$$1/|U| \le CKG_\beta(B) \le 1，$$

其中： $KG_\beta(B) = 1/|U|$，当且仅当 $R_\beta^B = \breve{R}_\beta^B$；

$KG_\beta(B) = 1$，当且仅当 $R_\beta^B = \hat{R}_\beta^B$。

**证略。**

**定理 6** 连续值信息系统 $CIS$ 中的知识粒度 $CKG_\beta(B)$ 是定义 2 意义下的信息粒度。

**证明：**

1) 显然， $KG_\beta(B) \ge 0$。

2) 令 $A, B \subseteq C$，则连续值信息系统 $CIS$ 下的论域分类可表示为

$$T_\beta(A) = \left\{T_\beta^A(x_1), T_\beta^A(x_2), \cdots, T_\beta^A\left(x_{|U|}\right)\right\}，\ T_\beta(B) = \left\{T_\beta^B(x_1), T_\beta^B(x_2), \cdots, T_\beta^B\left(x_{|U|}\right)\right\}。$$

若 $A \approx B$，则对于任意的 $i \in \{1, 2, \cdots, |U|\}$，都有 $T_\beta^A(x_i) = T_\beta^B(x_i)$，即 $\left|T_\beta^A(x_i)\right| = \left|T_\beta^B(x_i)\right|$。因此，有

$$CKG_\beta(A) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^A(x_i)\right| = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^B(x_i)\right| = CKG_\beta(B)。$$

3) 令 $A, B \subseteq C$，且 $A \prec B$，则对于任意的对象 $u_i \in U$，有 $T_\beta^A(x_i) \subseteq T_\beta^B(x_i)$。所以

$$CKG_\beta(A) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^A(x_i)\right| < \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^B(x_i)\right| = CKG_\beta(B)。$$

由此可得， $CKG_\beta(B)$ 是定义 2 意义下的一个知识粒度。

连续值信息系统中的知识粒度可以表示知识分辨能力， $CKG_\beta(B)$ 越小，分辨能力越强； $CKG_\beta(B)$ 越大，分辨能力越小。为更符合认知与方便计算，提出分辨度的定义为：

**定义 5** 连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$，对任意的 $B \subseteq C$，连续值信息系统 $CIS$ 的分辨度(Discernibility)定义为：

$$CDS_\beta(B) = 1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|\cup K_\beta^B\right| = \sum_{i=1}^{|U|}\frac{1}{|U|}\left(1 - \frac{\left|T_\beta^B(x_i)\right|}{|U|}\right)。$$

其中， $K_\beta^B \in T_\beta^B(x_i)$， $\left|\cup K_\beta^B\right|$ 是 $T_\beta^B(x_i)$ 中所有元素并的基，且 $\left|\cup K_\beta^B\right| = \left|T_\beta^B(x_i)\right|$。

**定理 7 (极小值)**连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，$R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中相对于属性集 $B$ 的分辨度最小值是 0，当且仅当 $R_\beta^B=\widehat{R}_\beta^B$，其中 $\widehat{R}_\beta^B$ 为全域相似关系，有

$$U/\widehat{R}_\beta^B=\left\{T_\beta^B\left(x_i\right)=U:x_i\in U\right\}=\{U,U,\cdots,U\}。$$

**定理 8 (极大值)**连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，$R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中相对于属性集 $B$ 的分辨度最大值是 $1-1/|U|$，当且仅当 $R_\beta^B=\breve{R}_\beta^B$，其中 $\breve{R}_\beta^B$ 为恒等相似关系，有

$$U/\breve{R}_\beta^B=\left\{T_\beta^B\left(x_i\right)=\{x_i\}:x_i\in U\right\}=\left\{\{x_1\},\{x_2\},\cdots,\left\{x_{|U|}\right\}\right\}。$$

**定理 9 (边界性)**连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，$R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中相对于属性 $B$ 的分辨度存在边界为：

$$0\le CDS_\beta\left(B\right)\le 1-1/|U|，$$

其中，$CDS_\beta\left(B\right)=0$，当且仅当 $R_\beta^B=\widehat{R}_\beta^B$；

$CDS_\beta\left(B\right)=1-1/|U|$，当且仅当 $R_\beta^B=\breve{R}_\beta^B$。

连续值信息系统中的分辨度可以更加直观的表示知识的分辨能力：$CDS_\beta\left(B\right)$ 越大，分辨能力越强；$CDS_\beta\left(B\right)$ 越小，分辨能力越小。这符合人们的直观理解也便于计算。

### 3.3. 连续值信息系统的知识熵

粗糙集理论研究中，论域 $U$ 上的一个等价关系(即划分)可以看作是定义在 $U$ 的子集组成的 $\sigma$-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 6** [9]设 $U$ 为论域，$P$ 为 $U$ 上的等价关系，$P$ 在 $U$ 上导出的划分为 $\pi$，其中 $\pi=\{X_1,X_2,\cdots,X_n\}$，则 $P$ 在 $U$ 的子集组成的 $\sigma$-代数上定义的概率分布为

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

其中，$p\left(X_i\right)=\dfrac{|X_i|}{|U|}$，$i=1,2,\cdots,n$。

有了知识概率分布的定义后，根据信息论就可以定义知识熵的概念[9]。

**定义 7** [9]设 $P$ 是知识库 $K=(U,\mathbf{R})$ 中的知识，$U/P=\{X_1,X_2,\cdots,X_n\}$，$P\in\mathbf{R}$，定义知识 $P$ 的熵 $H\left(P\right)$ 为

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

对于 $H\left(P\right)$ 取值范围，有

1) 当等价关系是恒等关系时，

$$H\left(P\right)=-\sum_{i=1}^{n}p\left(X_i\right)\log p\left(X_i\right)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{1}{|U|}=\log|U|。$$

2) 当等价关系是全域关系时，

$$H\left(P\right)=-\sum_{i=1}^{n}p\left(X_i\right)\log p\left(X_i\right)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}=-\frac{|U|}{|U|}\log\frac{|U|}{|U|}=0。$$

故，$0\le H\left(P\right)\le\log|U|$。

等价关系下，若将每一个对象 $x_i$ 单独看待，它所在的等价类 $X(x_i)$ 看作邻域，那么对象 $x_i$ 所在等价类的基即为论域中与对象 $x_i$ 满足等价关系的对象数目。等价类 $X(x_i)$ 的基越大，粒度越大， $X(x_i)$ 中与对象 $x_i$ 满足等价关系的对象数目就越多，分辨能力就越弱；反之亦然。通过将邻域从等价类扩展到相似类，将相似类 $T_\beta^B(x_i)$ 中的元素个数看作与对象 $x_i$ 满足相似关系的对象集合。$T_\beta^B(x_i)$ 的基越大，与对象 $u_i$ 满足相似相容关系的对象数目越多，分辨能力越弱[24]。基于这种考虑，同时保持熵的良好性质(易计算性，单调性)，给出一种连续值信息系统的不确定性度量如下：

**定义 8** 连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， $R_\beta^B$ 是相似关系。$U/R_\beta^B=\{T_\beta^B(x_i):x_i\in U\}$，给出连续值信息系统 $CIS$ 的知识熵 $CKE_\beta(B)$ 为：

$$CKE_\beta(B)=\sum_{i=1}^{|U|}\frac{1}{|U|}\cdot\log_2\left|T_\beta^B(x_i)\right|。$$

**定理 10** (**极小值**)连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， $R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中 $CKE_\beta(B)$ 的最小值是 0，当且仅当 $R_\beta^B=\breve{R}_\beta^B$，其中 $\breve{R}_\beta^B$ 为恒等相似关系，有

$$U/\breve{R}_\beta^B=\{T_\beta^B(x_i)=\{x_i\}:x_i\in U\}=\{\{x_1\},\{x_2\},\cdots,\{x_{|U|}\}\}。$$

**定理 11** (**极大值**)连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， $R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中 $CKE_\beta(B)$ 的最大值是 $\log_2|U|$，当且仅当 $R_\beta^B=\hat{R}_\beta^B$，其中 $\hat{R}_\beta^B$ 为全域相似关系，有

$$U/\hat{R}_\beta^B=\{T_\beta^B(x_i)=U:x_i\in U\}=\{U,U,\cdots,U\}。$$

**定理 12** (**边界性**)连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， $R_\beta^B$ 是相似关系。连续值信息系统 $CIS$ 中 $CKE_\beta(B)$ 的边界为：

$$0\le CKE_\beta(B)\le\log_2|U|,$$

其中， $CKE_\beta(B)=0$，当且仅当 $R_\beta^B=\breve{R}_\beta^B$；

$CKE_\beta(B)=\log_2|U|$，当且仅当 $R_\beta^B=\hat{R}_\beta^B$。

定义 8 给出的 $CKE_\beta(B)$ 保持了熵良好的单调性与计算的方便性。所以，可把 $CKE_\beta(B)$ 作为一种连续值信息系统中的不确定性度量。

## 4. 三种不确定性度量的比较

下面讨论连续值信息系统中 $CKG_\beta(B)$， $CDS_\beta(B)$ 及 $CKE_\beta(B)$ 之间的关系，如表 1 所示。当 $R_\beta^B$ 由最粗的全域关系变为最细的恒等关系时，粒度 $KG_\beta(B)$ 由 1 减小到 $1/|U|$，分辨度 $CDS_\beta(B)$ 由 0 增大到 $1-1/|U|$， $CKE_\beta(B)$ 由 $\log_2|U|$ 减少到 0。故从全域关系变为恒等关系，粒度会越来越细，分辨度越来越高，知识熵越来越低。

**Table 1.** The comparison of three kinds of measure
**表 1.** 三种度量方式的比较

| $T_\beta^B$ | $\breve{T}_\beta^B$ | $\hat{T}_\beta^B$ | 取值范围 |
|---|---|---|---|
| $CKG_\beta(B)$ | $1/\|U\|$ | 1 | $1/\|U\|\le CKG_\beta(B)\le 1$ |
| $CDS_\beta(B)$ | $1-1/\|U\|$ | 0 | $0\le CDS_\beta(B)\le 1-1/\|U\|$ |
| $CKE_\beta(B)$ | 0 | $\log_2\|U\|$ | $0\le CKE_\beta(B)\le\log_2\|U\|$ |

## 5. 结束语

粗糙集理论中知识不确定性主要来源于两方面：不可分辨关系与粗糙上下近似。当边界域中的上、下近似不相等时，边界域不为空，即存在不确定性。本文基于经典粗糙集中的粗糙度、知识粒度与信息熵提出了连续值信息系统中的三种不确定性度量方法：粗糙度；知识粒度(分辨度)与知识熵。给出了三种度量的取值范围，并对三种度量间的关系进行了分析研究。基于本文提出的不确定性度量方法，接下来将在连续值信息系统中的属性约简方面开展进一步的研究工作。

## 参考文献 (References)


[1] Düntsch, I. and Gediga, G. (1998) Uncertainty Measures of Rough Set Prediction. *Artificial Intelligence*, **106**, 109-137. https://doi.org/10.1016/S0004-3702(98)00091-5

[2] Wierman, M. (1999) Measuring Uncertainty in Rough Set Theory. *International Journal of General Systems*, **28**, 283-297. https://doi.org/10.1080/03081079908935239

[3] Beaubouef, T., Petry, F.E. and Arora, G. (1998) Information-Theoretic Measures of Uncertainty for Rough Sets and Rough Relational Databases. *Information Sciences*, **109**, 185-195. https://doi.org/10.1016/S0020-0255(98)00019-X

[4] Qian, Y.H., Liang, J.Y. and Wang, F. (2009) A New Method for Measuring the Uncertainty in Incomplete Information Systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **17**, 855-880. https://doi.org/10.1142/S0218488509006303

[5] Liang, J.Y. and Shi, Z.Z. (2004) The Information Entropy, Rough Entropy and Knowledge Granulation in Rough Set Theory. I*nternational Journal of General Systems*, **12**, 37-46. https://doi.org/10.1142/s0218488504002631

[6] Yao, Y.Y. and Zhao, L.Q. (2012) A Measurement Theory View on the Granularity of Partitions. *Information Sciences*, **213**, 1-13. https://doi.org/10.1016/j.ins.2012.05.021

[7] 刘财辉, 苗夺谦, 岳晓冬, 赵才荣. 知识不确定性度量及其关系研究[J]. 计算机科学, 2015, 41(3): 66-70.

[8] 张楠, 苗夺谦, 岳晓冬. 区间值信息系统的知识约简[J]. 计算机研究与发展, 2010, 47(8): 1362-1371.

[9] 苗夺谦. Rough Set 理论及其在机器学习中的应用研究[D]: [博士学位论文]. 北京: 中国科学院自动化研究所, 1997.

[10] 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.

[11] 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.

[12] Wang, G. (2003) Rough Reduction in Algebra View and Information View. *International Journal of Intelligent Systems*, **18**, 679-688. https://doi.org/10.1002/int.10109

[13] 黄兵, 周献中, 史迎春. 基于一般二元关系的知识粗糙熵与粗集粗糙熵[J]. 系统工程理论与实践, 2004, 24(1): 93-96.

[14] Liang, J., Shi, Z., Li, D. and Wierman, M. (2006) Information Entropy, Rough Entropy and Knowledge Granularity in Incomplete Information Systems. *International Journal of General Systems*, **35**, 641-654. https://doi.org/10.1080/03081070600687668

[15] Feng, Q.R., Miao, D.Q., Zhou, J. and Cheng, Y. (2010) A Novel Measure of Knowledge Granularity in Rough Sets. *International Journal of Granular Computing, Rough Sets and Intelligent Systems*, **1**, 233-251. https://doi.org/10.1504/IJGCRSIS.2010.029580

[16] 冯琴荣. 基于多维数据模型的粒计算方法研究[D]: [博士学位论文]. 上海: 同济大学, 2009.

[17] 王国胤, 苗夺谦, 吴伟志, 梁吉业. 不确定信息的粗糙集表示和处理[J]. 重庆邮电大学学报(自然科学版), 2010, 22(5): 541-544.

[18] Xu, W.H., Zhang, X.Y. and Zhang, W.X. (2009) Knowledge Granulation, Knowledge Entropy and Knowledge Uncertainty Measure in Ordered Information Systems. *Applied Soft Computing*, **9**, 1244-1251. https://doi.org/10.1016/j.asoc.2009.03.007

[19] 王国胤, 张清华, 马希骜, 杨青山. 知识不确定性问题的粒计算模型[J]. 软件学报, 2011, 22(4): 676-694.

[20] Dai, J.H., Wang, W.T., Xu, Q. and Tian, H.W. (2012) Uncertainty Measurement for Interval-Valued Decision Systems Based on Extended Conditional Entropy. *Knowledge-Based Systems*, **27**, 443-450. https://doi.org/10.1016/j.knosys.2011.10.013

[21] Guan, Y.Y., Wang, H.K., Wang, Y. and Yang, F. (2009) Attribute Reduction and Optimal Decision Rules Acquisition

for Continuous Valued Information Systems. *Information Sciences*, **179**, 2974-2984. https://doi.org/10.1016/j.ins.2009.04.017

[22] Yao, Y.Y. and Zhao, L.Q. (2012) A Measurement Theory View on the Granularity of Partitions. *Information Sciences*, **213**, 1-13. https://doi.org/10.1016/j.ins.2012.05.021

[23] 张楠. 区间值信息系统与知识空间的粒计算方法研究[D]: [博士学位论文]. 上海: 同济大学, 2012.

[24] 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.