基于知识含量的粗糙集不确定性度量*)

刘纪芹[1] 史开泉[2]

(山东财政学院统计与数理学院 济南 250014) (山东大学数学与系统科学学院 济南 250100)[2]

摘 要 粗糙集的不确定性度量是粗糙集理论中一项重要的数值特征,而 Z. Pawlak 提出的粗糙集的不确定性度量,即传统的近似精度与粗糙度具有局限性。考虑导致粗集粗糙性的原因,将传统的粗糙度与知识的含量测度结合起来,提出了一种新的粗糙集不确定性的度量方法,讨论了这一度量的特性,通过实例说明这一新的度量方法的合理性及计算的简便性。

关键词 粗糙集,不确定性,知识含量测度,近似精度,粗糙度

An Uncertainty Measure for Rough Sets Based on Knowledge Capacity

LIU Ji-Qin[1] SHI Kai-Quan[2]

(Department of Statistics and Mathematics, Shandong Finance Institute, Jinan 250014)[1]

(School of Mathematics and System Sciences, Shandong University, Jinan 250100)[2]

Abstract The measure of uncertainty is an important numerical characterization in rough sets theory. The measure of uncertainty proposed by Z. Pawlak has its limitations. Considering the causes of uncertainty in rough sets, we propose a new uncertainty measure by combining traditional roughness with knowledge capacity measure. Some good properties of the new uncertainty measure are further discussed. In the end, an example shows the rationality and simplicity of the new uncertainty measure.

Keywords Rough sets, Uncertainty, Knowledge capacity measure, Accuracy, Roughness

1 引言

波兰数学家 Z. Pawlak 提出的粗糙集理论[1]是处理不完全和不精确信息的一种新的数学工具[2],这一理论已引起各国学者的广泛关注[3,4]。在粗糙集理论中,粗糙集的不确定性度量是一项重要的数值特征,很多学者对此进行了讨论[5~8]。Z. Pawlak 给出了用集合 X 的上、下近似来刻画粗糙集不确定性的两个数值特征[8,9],即近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$。尽管近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$ 能反映集合 X 的不确定性,但是它们并没有提供给我们那些完全属于 X 的下近似区域和负域里面与不可分辨关系 R 的知识粒度有关的知识的不确定性。因为在不同的近似空间中,当知识的不确定性明显不同时,用 $\rho_R(X)$ 可以得到相同的粗糙度,所以用近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$ 描述粗糙集的不确定性存在不足之处。

由文[5,9]知,粗糙集理论中,粗糙集的不确定性主要由两个原因引起:一个原因来自于给定近似空间的粗糙集的边界,当边界为空集时知识是完全确定的,边界越大知识越粗糙,这种不确定性称为系统的不确定性。粗糙集理论处理这类不确定性是通过引进近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$ 来实现的。另一个原因是直接来自于论域上二元关系对论域中对象进行分类的能力,体现在论域中对象的不可分辨性,即知识的粒度。知识粒度越大,知识越粗糙,相对于近似空间中的概念和知识就越不确定,这种不确定性称为概念的不确定性。考虑这两种原因引起的粗糙集的不确定性,文[5]利用信息理论中熵的概念定义了粗糙集的不确定性度量。文[6]利用过剩熵(excess entropy)给出了粗糙集的不确定性度量,但计算起来不算简便。本文利用知识含量[10]概念,并对此进行改进,利用这一改进的概念给出了一种新的粗集粗糙性的度量方法。

2 粗糙集中主要概念简述

为了便于讨论,下面简要给出粗糙集中几个主要概念[5,6,9]。

设论域 U 为有限集,R 是 U 上的一族等价关系,$K=(U,R)$ 为一知识库(或近似空间),$[x]_R$ 为 U 上的 R-等价类。

定义 2.1 设 $K=(U,R)$ 为一知识库,$P\subseteq R$,且 $P\neq\phi$,则 $\bigcap P$ 也是一种等价关系,称其为 P 上的不可分辨关系,记为 $ind(P)$,且有

$$[x]_{ind(P)}=\bigcap_{r\in P}[x]_r$$

$U/ind(P)$ 即为等价关系 $ind(P)$ 的所有等价类,为简单起见,我们用 U/P 代替 $U/ind(P)$。

定义 2.2 两个子集

$$\underline{R}X=\{x\in U|[x]_R\subseteq X\},$$

$$\overline{R}X=\{x\in U|[x]_R\cap X\neq\phi\},$$

分别称为 X 的 R-下近似集和 R-上近似集。

由上、下近似可得 X 的 R-正域为:$pos_R(X)=\underline{R}X$;X 的 R-负域为:$neg_R(X)=U-\overline{R}X$;X 的 R-边界域为:$bn_R(X)=\overline{R}X-\underline{R}X$。

*)国家自然科学基金项目(60364001),山东省自然科学基金(Y2004A04)。刘纪芹 副教授,博士研究生,研究方向:粗系统理论与应用。史开泉 教授,博士生导师,研究方向:粗系统理论与应用。

集合的不精确性是由于边界域的存在而引起的，集合的边界域越大，其精确性越低，由此引入描述集合 x 的不确定性的如下概念。

定义 2.3 称

$$\alpha_R(X)=\frac{|\underline{RX}|}{|\overline{RX}|}$$

为集合 X 的 R-近似精度，其中 $|X|$ 表示 X 的基数，$\overline{RX}\neq\phi$，若 $\overline{RX}=\phi$，则定义 $\alpha_R(X)=1$，称

$$\rho_R(X)=1-\alpha_R(X)$$

为集合 X 的 R-粗糙度。

用近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$ 来描述粗糙集的不确定性具有局限性。下面通过例子来理解这种局限性。

例 1 设论域 $U=\{1,2,3,4,5,6,7,8,9,10\}$，集合 $X=\{1,2,3,4,6,7\}$，R_1,R_2,R_3 为 U 上的三个等价关系，其导出的划分分别为

$U/R_1=\{\{1,2,3,4\},\{5,6,7\},\{8,9,10\}\}$，

$U/R_2=\{\{1,2\},\{3,4\},\{5,6,7\},\{8,9\},\{10\}\}$，

$U/R_3=\{\{1\},\{2\},\{3\},\{4\},\{5,6,7\},\{8\},\{9\},\{10\}\}$。

则 X 的 R_1,R_2,R_3 的下近似与上近似分别为

$\underline{R_1}X=\underline{R_2}X=\underline{R_3}X=\{1,2,3,4\}$，$\overline{R_1}X=\overline{R_2}X=\overline{R_3}X=\{1,2,3,4,5,6,7\}$。

则 X 关于 R_1,R_2,R_3 的近似精度相同，即为

$$\alpha_{R_1}(X)=\alpha_{R_2}(X)=\alpha_{R_3}(X)=4/7。$$

X 关于 R_1,R_2,R_3 的粗糙度相同，即为

$$\rho_{R_1}(X)=\rho_{R_2}(X)=\rho_{R_3}(X)=3/7。$$

显然，$ind(R_3)\subset ind(R_2)\subset ind(R_1)$，即知识 R_3 比 R_2 精细，知识 R_2 比 R_1 精细，但它们的近似精度与粗糙度却相同。因此，传统的近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$ 具有局限性，它们只能部分反映 X 的不确定性，不能完全反映 X 的不确定性。这需要对传统的近似精度 $\alpha_R(X)$ 与粗糙度 $\rho_R(X)$ 进行改进，应同时考虑知识 R 的含量。

3 知识含量的度量

由引言中的叙述我们得知，引起粗糙集不确定性的两个原因之一是论域上二元等价关系对论域中对象进行分类的能力，即知识的粒度。由此，讨论知识含量的度量。

设 $K=(U,R)$ 为一知识库，$U/R=\{X_1,X_2,\cdots,X_n\}$。文[10]给出了知识 R 的知识含量概念，但当 $|X_i|=1,i=1,2,\cdots,n$ 时，这是所有对 U 的划分中最为精细的一个，此时知识含量的测度应达最大值 1，但利用文[10]中定义的知识含量 $I(R)=1-\frac{1}{|U|^2}\sum_{i=1}^{n}|X_i|^2$ 可知该种情况下的知识含量测度不是 1。故对此进行改进，给出如下概念。

定义 3.1 设 $K=(U,R)$ 为一知识库，$U/R=\{X_1,X_2,\cdots,X_n\}$。称

$$I(R)=\begin{cases}1, & |X_i|=1,i=1,2,\cdots,n\\ 1-\frac{1}{|U|^2}\sum_{i=1}^{n}|X_i|^2, & \text{其它}\end{cases}$$

为知识 R 含量的测度，其中 $|X_i|$ 表示集合 X_i 的基数。

由定义 3.1 可得如下知识含量测度的性质。

定理 3.1 设 $K=(U,R)$ 为任一知识库，对知识 R，有 $0\subseteq I(R)\subseteq 1$。特别地，当 $U/R=\{U\}$ 时，$I(R)=0$。当 $U/R=\{X_1,X_2,\cdots,X_n\}$，$|X_i|=1,i=1,2,\cdots,n$ 时，$I(R)=1$。

由定义 3.1 知显然成立，证明略。

定理 3.2 设 $(U,R_1),(U,R_2)$ 为两个知识库，若 $U/R_1=U/R_2$，则 $I(R_1)=I(R_2)$。

由定义 3.1 易证，略。

定理 3.3 设 $(U,R_1),(U,R_2)$，为两个知识库，若 $U/R_1\subset U/R_2$，则 $I(R_1)>I(R_2)$。

证明：设 $U/R_1=\{X_1,X_2,\cdots,X_n\}$，$U/R_2=\{Y_1,Y_2,\cdots,Y_n\}$，因为 $U/R_1\subset U/R_2$，则 $m>n$，而且存在 $\{1,2,\cdots,m\}$ 的一个划分 $\{D_1,D_2,\cdots,D_n\}$，使得 $Y_i=\bigcup_{j\in D_i}X_j$，且 $|Y_i|=\sum_{j\in D_i}|X_j|$，故

$$\sum_{i=1}^{n}\frac{|Y_i|^2}{|U|^2}=\sum_{i=1}^{n}\frac{\sum_{j\in D_i}|X_j|^2}{|U|^2}>\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2},$$

$$1-\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2}>1-\sum_{i=1}^{n}\frac{|Y_i|^2}{|U|^2}$$

由定义 3.1 得，$I(R_1)>I(R_2)$。

定理 3.4 设 $K=(U,R)$ 为一知识库，R 将论域 U 分为两个等价类，即 $U/R=\{X_1,X_2\}$。当 $|X_1|=\begin{cases}|U|/2, & |U|\text{为偶数},\\ (|U|+1)/2, & |U|\text{为奇数}。\end{cases}$ 时，$I(R)$ 达到最大，即平均划分的知识库知识含量测度最大。

证明：设 $|U|=n>1$，$|X_1|=n_1$，则 $|X_2|=n-n_1$，故

$$\sum_{i=1}^{2}\frac{|X_i|^2}{|U|^2}=\frac{n_1^2+(n-n_1)^2}{n_2}=\frac{2n_1^2-2nn_1+n_2}{n^2},$$

由微积分中最值的求法知，当 $n_1=\begin{cases}n/2, & n\text{为偶数},\\ (n+1)/2, & n\text{为奇数}。\end{cases}$ 时，$\sum_{i=1}^{2}\frac{|X_i|^2}{|U|^2}$ 最小，由定义 3.1 知，此时 $I(R)$ 达到最大，即当 $|X_1|=\begin{cases}|U|/2, & |U|\text{为偶数},\\ (|U|+1)/2, & |U|\text{为奇数}。\end{cases}$ 时，$I(R)$ 达到最大。

定义 3.2[11] 设 $K=(U,R)$ 为一知识库，$U/R=\{X_1,X_2,\cdots,X_n\}$，$1\leqslant n\leqslant|U|$，若知识库 K 满足以下条件，则称其为 n 划分平均知识库：

(1) 若 $|U|\bmod n=0$，有 $|X_i|=|U|/n,(1\leqslant i\leqslant n)$；

(2) 若 $|U|\bmod n=s(1\leqslant s\leqslant n)$ 时，将 U/R 排序后，使得对于 $1\leqslant i\leqslant s$，有 $X_i=\left\lfloor\frac{|U|}{n}\right\rfloor+1$，对于 $s<i\leqslant n$，有 $X_i=\left\lfloor\frac{|U|}{n}\right\rfloor$，其中 $\left\lfloor\frac{|U|}{n}\right\rfloor$ 表示数值 $\frac{|U|}{n}$ 的整数部分。

定理 3.5 当知识库 $K=(U,R)$ 为 n 划分平均知识库时，$I(R)$ 达到最大。

证明类似于定理 3.4，略。

4 一种新的粗糙集不确定性度量

考虑到传统的近似精度与粗糙度的局限性及影响粗糙集不确定性的两种原因，下面给出一种新的粗糙集不确定性度量。

定义 4.1 设 $K=(U,R)$ 为一知识库，$X\subseteq U$，$U/R=\{X_1,X_2,\cdots,X_n\}$，称

$$Rough_R(X)=\rho_R(X)(1-I(R))$$

为 X 的 R-改进粗糙度。即

$$Rough_R(X)=\rho_R(X)(1-I(R))=\begin{cases}0, & |X_i|=1,i=1,2,\cdots,n,\\ (1-\frac{|\underline{RX}|}{|\overline{RX}|})\cdot\frac{1}{|U|^2}\sum_{i=1}^{n}|X_i|^2, & \text{其它}\end{cases}$$

称

$D_R(X)=1-Rough_R(X)$

为 X 的 R-改进近似精度，即

$$D_R(X)=1-Rough_R(X)=1-\rho_R(X)(1-I(R))$$
$$=I(R)+\alpha_R(X)-I(R)\alpha_R(X)。$$

对于这种改进的近似精度与粗糙度，有下面一些性质：

定理 4.1 设 $K=(U,R)$ 为一知识库，$X\subseteq U$，$U/R=\{X_1,X_2,\cdots,X_n\}$，则

(1)$0\leqslant Rough_R(X)\leqslant 1$，

(2)$0\leqslant D_R(X)\leqslant 1$。

证明：由定义 2.3，定义 4.1 及定理 3.1 知显然成立，略。

推论 (1)$\phi\subset X\subset U$，且 $U/R=\{U\}$，则 $Rough_R(X)=1$，$D_R(X)=0$。

(2)$\phi\subset X\subset U$，且 $U/R=\{X_1,X_2,\cdots,X_{|U|}\}$，则 $Rough_R(X)=0$，$D_R(X)=1$。

定理 4.2 设 (U,R_1)，(U,R_2) 为两个知识库，$X\subseteq U$。若 $U/R_1=U/R_2$，则

(1)$Rough_{R_1}(X)=Rough_{R_2}(X)$，

(2)$D_{R_1}(X)=D_{R_2}(X)$。

由定义 4.1 及定理 3.2 知显然成立，证明略。

定理 4.3 设 (U,R_1)，(U,R_2) 为两个知识库，$X\subseteq U$。若 $U/R_1\subseteq U/R_2$，则

(1)$Rough_{R_1}(X)=Rough_{R_2}(X)$，

(2)$D_{R_1}(X)=D_{R_2}(X)$。

证明：(1) 因为 $U/R_1=U/R_2$，则 $[x]_{R_1}\subseteq[x]_{R_2}$，设

$$x\in\underline{R_2}(X)=\{x|x\in U,[x]_{R_2}\subseteq X\}\Rightarrow x\in\{x|x\in U,[x]_{R_1}\subseteq X\}=\underline{R_1}(X),$$

故 $\underline{R_2}(X)\subseteq\underline{R_1}(X)$，$|\underline{R_2}(X)\leqslant|\underline{R_1}(X)$。

设

$$x\in\overline{R_1}(X)=\{x|x\in U,[x]_{R_1}\cap X\neq\phi\}\Rightarrow x\in\{x|x\in U,[x]_{R_2}\cap X\neq\phi\}=\overline{R_2}(X),$$

故 $\overline{R_1}(X)\subseteq\overline{R_2}(X)$，$|\overline{R_1}(X)|\leqslant|\overline{R_2}(X)|$。

从而有

$$1-\frac{|\underline{R_1}(X)|}{|\overline{R_1}(X)|}\leqslant 1-\frac{|\underline{R_2}(X)|}{|\overline{R_2}(X)|},$$

$$\rho_{R_1}(X)\leqslant\rho_{R_2}(X)。$$

又由定理 3.3 知，$I_{R_1}(X)\geqslant I_{R_2}(X)$，则 $1-I_{R_1}(X)\leqslant 1-I_{R_2}(X)$。由定义 4.1，有

$Rough_{R_1}(X)\leqslant Rough_{R_2}(X)$。

(2)由(1)及定义 4.1 知，显然有 $D_{R_1}(X)\geqslant D_{R_2}(X)$。

定理 4.4 设 $K=(U,R)$ 为一知识库，$\phi\subset X\subseteq U$，$\phi\subset Y\subseteq U$，

(1)若 $\underline{R}X=\underline{R}Y$，则

(a)$Rough_R(X\cap Y)\leqslant\min\{Rough_R(X),Rough_R(Y)\}$，

(b)$D_R(X\cap Y)\geqslant\max\{D_R(X),D_R(Y)\}$。

(2)若 $\overline{R}X=\overline{R}Y$，则

(a)$Rough_R(X\cup Y)\leqslant\min\{Rough_R(X),Rough_R(Y)\}$，

(b)$D_R(X\cup Y)\geqslant\max\{D_R(X),D_R(Y)\}$。

证明：(1)因为 $\underline{R}(X\cap Y)=\underline{R}X\cap\underline{R}Y$，又 $\underline{R}X=\underline{R}Y$，则 $|\underline{R}(X\cap Y)|=|\underline{R}X|=|\underline{R}Y|$。又 $\overline{R}(X\cap Y)\subseteq\overline{R}X\cap\overline{R}Y\subseteq\overline{R}X$，故 $|\overline{R}(X\cap Y)|\leqslant|\overline{R}Y|$。同理，$|\overline{R}(X\cap Y)|\leqslant|\overline{R}Y|$，则

$$\rho_R(X\cap Y)=1-\frac{|\underline{R}(X\cap Y)|}{|\overline{R}(X\cap Y)|}\leqslant 1-\frac{|\underline{R}X|}{|\overline{R}X|}=\rho_R(X)。$$

同理，$\rho_R(X\cap Y)\leqslant\rho_R(Y)$，故

$\rho_R(X\cap Y)\leqslant\min\{\rho_R(X),\rho_R(Y)\}$。

又 $1-I(R)\geqslant 0$，由定义 4.1 知

$Rough_R(X\cap Y)\leqslant\min\{Rough_R(X),Rough_R(Y)\}$，

$D_R(X\cap Y)\geqslant\max\{D_R(X),D_R(Y)\}$。

(2)同理可证，略。

由此可见，定义 4.1 给出的新的粗糙集不确定性度量具有文[6]中的性质。

5 应用举例

例 2 考虑第 2 节中例 1，由定义 3.1 知，$I(R_1)=0.66$，$I(R_2)=0.78$，$I(R_3)=0.84$，故得 X 的 R_1，R_2，R_3-改进粗糙度分别为 $Rough_{R_1}(X)=0.146$，$Rough_{R_2}(X)=0.094$，$Rough_{R_3}(X)=0.069$。R_1，R_2，R_3-改进近似精度分别为 $D_{R_1}(X)=0.854$，$D_{R_2}(X)=0.906$，$D_{R_3}(X)=0.931$，故有

$Rough_{R_1}(X)>Rough_{R_3}(X)>Rough_{R_2}(X)$，

$D_{R_1}(X)<D_{R_2}(X)<D_{R_3}(X)$。

因此，分类更精细的知识其粗糙度更小，近似精度更大。对比例 1、例 2 可知，定义 4.1 给出的改进粗糙度和近似精度解决了经典粗糙集粗糙度与近似精度的局限性，而且计算简便。

结论 由于粗糙集的不确定性是由两个原因，即系统的不确定性和概念的不确定性引起的[5,9]，而 Pawlak 定义的粗糙集近似精度与粗糙度只考虑了系统的不确定性，不能确切描述某些粗糙集的不确定性，因此，这种定义存在局限性。鉴于此，本文给出了描述粗糙集不确定性的新的度量方法，即将传统的粗糙度与知识的含量测度结合起来。最后通过实例说明了这种新的度量方法克服了经典粗糙集粗糙度与近似精度的局限性，且计算简便。

参 考 文 献

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

2 Felix R, Ushio T. Rule induction from inconsistent and incomplete data using rough sets [A]. In:Proceedings of the IEEE International Conference on Systems, Man and Cybernetics [C], 1999,5(10):154～158

3 Ziarko W. Variable precision rough set model [J]. Journal of Computer and Information Sciences, 1993, 46 (1):39～59

4 Zhang Huaguang, Liang Hongli, Liu Derong . Two new operators in rough set theory with applications to fuzzy sets [J]. Information Sciences, 2004(166):147～165

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

6 Xu Baowen, Zhou Yuming, Lu Hongmin. An improved accuracy measure for rough sets [J]. Journal of Computer and Information Sciences, 2005(71):163～173

7 苗夺谦，王钰. 粗集理论中概念与运算信息表示. 软件学报，1999,10(2): 34～37

8 Pawlak Z. Rough sets: Theoretical Aspects of Reasoning about Data [M]. Kluwer Academic Publishers, Norwell, MA,1991

9 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

10 Liang Jiye, Li Deyu. Information measures of roughness of knowledge and significance of attribute in rough set theory [J]. Journal of Engineering Mathematics, 2000,17(Supp):106～108

11 王瑜，胡运发，张凯. 基于粗集理论的知识含量度量研究[J]. 计算机研究与发展，2004,41(9):1500～1506