知识的不确定性度量和基于信息熵的模糊粗糙度

郭 庆
(昆明理工大学 理学院,云南 昆明 650093)

摘要:基于另一角度——信息熵给出了模糊集在 λ - 水平下的模糊粗糙度的概念,从而更深刻地反映知识的不确定性及集合的不确定性.

关键词:信息熵;粗糙度;模糊粗糙集;不确定性
中图分类号:G206 **文献标识码**:A **文章编号**:1007 - 855X(2003)02 - 0142 - 03

Uncertainty of Knowledge and the Rough Measure of Fuzzy Rough Set Based on the Information Entropy

GOU Qing
(Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: Based on another angle - information entropy, the rough measure of fuzzy rough set under level is given to deeply reflect the uncertainty of knowledge and fuzzy rough set.

Key words: information entropy; rough measure; fuzzy rough set; uncertainty

0 引 言

粗糙集理论的知识的不确定性主要由两个原因引起的:一个原因是来自与在给定近似空间的粗糙集的边界,即我们定义的粗糙度,粗糙集的知识的不确定性引起的另一个原因是来自与论域上的二元关系及其产生的知识模块,即近似空间本身,这种不确定性称为概念的不确定性.处理概念的不确定性的方法通常用信息熵来刻画,知识的粗糙性与信息熵的关系比较密切,知识的粗糙性实质上是其所含信息的多少的更深层次的刻画.[1]

首先,度量信息源 X 的不确定性通常我们用信息熵来表示.

定义 1.1[1,4,5] 设 U 是论域,$X_1, X_2, \cdots, X_n$ 是 U 上的一个划分,其上有概率分布:

$$X = \begin{Bmatrix} X_1, X_2, \cdots, X_n \\ p_1, p_2, \cdots, p_n \end{Bmatrix}$$

称 $H(X) = -\sum_{i=1}^{n} p_i \log p_i$ 为信息源 X 的信息熵.

定义 1.2[4] 设 U 是论域,$K = (U, P)$ 和 $K_1 = (U, Q)$ 是关于 U 的两个知识库,如果 $U/\mathrm{ind}(P) \subseteq U/\mathrm{ind}(Q)$,则称知识 P 比知识 Q 较细,或 Q 比较 P 粗,记作 $P \prec Q$.

定理 1.1[4] 设 U 是论域,$K = (U, P)$ 和 $K_1 = (U, Q)$ 是关于 U 的两个知识库,并且 $P \prec Q$,则 $H(P) \geqslant H(Q)$.

尽管粗糙度能反映概念 X 的不确定性,但是他们并没有提供给我们那些完全属于 X 的下近似区域(正域)里面与不可分辨关系的知识粒度有关的知识的不确定性.所以我们可以在定义粗糙集的粗糙度时有必要把信息熵考虑进来.[1]

例[1] 设 U 是论域,$U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$,设集合 $X = \{x_1, x_2, x_3, x_4, x_5, x_8\}$,对如下等价关系 R, S, T 有:

$$U/P = \{(x_1, x_2, x_3, x_4), (x_5, x_6, x_7), (x_8, x_9)\};$$

收稿日期:2002 - 09 - 26.
作者简介:郭庆(1979 ~),男,硕士研究生;主要研究方向:模糊系统与粗集理论.

$$U/S = \{(x_1,x_2),(x_3,x_4),(x_5,x_6,x_7),(x_8,x_9)\};$$
$$U/T = \{(x_1),(x_2),(x_3),(x_4),(x_5,x_6,x_7),(x_8,x_9)\};$$

则　$\underline{R}X = \{x_1,x_2,x_3,x_4\}, \overline{R}X = U;$

$\underline{S}X = \{x_1,x_2\} \cup \{x_3,x_4\}, \overline{S}X = U;$

$\underline{T}X = \{x_1\} \cup \{x_2\} \cup \{x_3\} \cup \{x_4\}, \overline{T}X = U.$

显然,X 关于三个近似空间$(U,R),(U,S)(U,T)$ 的上下近似都是相同的

$$\overline{aprX} = \{x_1,x_2,x_3,x_4\}, \overline{aprX} = U$$

当然集合 X 的关于近似空间$(U,R),(U,S)(U,T)$ 的近似精度是相同 $\eta(X) = \frac{4}{9}$,经计算 X 关于R,S,T 的正则条件熵也是相同的,事实上,与划分 R 相比,划分 S 与T 被重新划分的部分恰好是X 关于三个近似空间的下近似部分,但很明显 R 的不确定性大于知识S 的不确定性,知识 S 的不确定性大于知识T 的不确定性,为此我们引进信息熵来重新定义知识的粗糙度.

定义 1.4[1]　设(U,R) 是近似空间,集合 X 关于 R 的粗糙度定义为:

$$E_r(X) = -\rho_R \cdot \sum_{i=1}^{n} p(X_i)\log p_i$$

这里 $p_i = \frac{1}{|X_i|}$,且 $X = \{X_1,X_2,\cdots,X_n\}$.

根据这个粗糙度的定义,对文章开头的例子,我们有:

$$E_R(X) = -\frac{5}{9} \cdot [\frac{4}{9}\log\frac{1}{4} + \frac{3}{9}\log\frac{1}{3} + \frac{2}{9}\log\frac{1}{2}] = 0.274$$

$$E_S(X) = -\frac{5}{9}[\frac{2}{9} \cdot 3 \cdot \log\frac{1}{2} + \frac{3}{9}\log\frac{1}{3}] = 0.20$$

$$E_T(X) = -\frac{5}{9}[4 \cdot \frac{1}{9} \cdot \log 1 + \frac{3}{9}\log\frac{1}{3} + \frac{2}{9}\log\frac{1}{2}] = 0.048$$

所以尽管 X 关于每个等价关系的上,下近似集是相等的,但划分越细,粗糙度越小.

我们说,一个模糊集合的截集把模糊集与经典集联系了起来,故我们想到对与于模糊集的 λ - 截集也会存在上述的情况,这里我们把信息熵定义的粗糙度引入当 A 是模糊集的情况.

定义 1.5[1]　设 $\forall\lambda \in [0,1]$,A 是任意的一个模糊集,则 A 关于λ 的截集定义为

$$A_\lambda(x) = \{x \mid A(x) \geqslant \lambda\}$$

定义 1.6[1]　设(U,R) 是Pawlak 近似空间,即 R 是论域U 上的一个等价关系.若 A 是U 上的一个模糊集合,则 A 关于(U,R) 的一对上下近似定义为一对模糊集合,其隶属函数

$$\underline{A}(x) = \inf\{A(y) \mid y \in [x]_R\}, \overline{A}(X) = \sup\{A(y) \mid y \in [x]_R\}(x \in U)$$

定义 1.7[2,3]　设 $A \in F(U)$,定义以下两个模糊集合 $\underline{R}A(x), \overline{R}A(x): U \to [0,1]$

$\underline{R}A(x) = \underline{R}A(X_i), \overline{R}A(x) = \overline{R}A(X_i)$.这里 $x \in X_i, i = 1,2,\cdots,n$.

定理 1.2　设 $\forall\lambda \in [0,1]$,U 是论域,R 是U 上的等价关系,且 $U/R = \{X_1,X_2,\cdots,X_n\}$,当 U 中的元素是离散的且有限时,我们有 $\underline{A}_\lambda(x) = (\underline{A(x)})_\lambda, \overline{A}_\lambda(x) = (\overline{A(x)})_\lambda$.

证明　$\underline{A}_\lambda = \{x \in A_\lambda \mid [x]_R \subseteq A_\lambda\} = \{x \in X_i \mid A(x) \geqslant \lambda\} = \{x \in X_i \mid \inf A(x) \geqslant \lambda\} = \{x \in \underline{A}(x) \mid \underline{A}(x) \geqslant \lambda\} = (\underline{A}(x))_\lambda$,所以,同理证明后一半.

由此我们会想到,对某一 λ - 水平下,模糊集合 A 正如前面讲到的对某些等价关系也会产生相同的上下近似.例如对于前面的三个等价关系,设一个模糊集合

$A = \{0.4/x_1,0.6/x_2,0.4/x_3,0.6/x_4,0.2/x_5,0.4/x_6,0.6/x_7,0.8/x_8,0.9/x_9\}$,当 $\lambda = 0.4$ 时 $A_\lambda = \{x_1,x_2,x_3,x_4,x_6,x_7,x_8,x_9\}$

$$R_1(\underline{A}_\lambda)(x) = \{x_1,x_2,x_3,x_4,x_8,x_9\}, R_1(\overline{A}_\lambda)(x) = U$$
$$R_2(\underline{A}_\lambda)(x) = \{x_1,x_2,x_3,x_4,x_8,x_9\}, R_2(\overline{A}_\lambda)(x) = U$$
$$R_3(\underline{A}_\lambda)(x) = \{x_1,x_2,x_3,x_4,x_8,x_9\}, R_3(\overline{A}_\lambda)(x) = U.$$

(下转第 146 页)

Al^{3+}分别在 pH = 5.0 和温度 20℃时,观察出现沉淀的情况,结果如表 1.

表 1 几种金属离子对人血清蛋白产生沉淀的影响

金属离子	Na^+	K^+	Zn^{2+}	Cu^{2+}	Mg^{2+}	Ca^{2+}	Co^{2+}	Fe^{3+}	Al^{3+}
金属离子浓度/mmol·L^{-1}	500	500	50	1.1	5	5	2.5	0.8	0.4
金属离子半径(Å)	1.33	0.95	0.74	0.72	0.65	0.99	0.74	0.64	0.5
沉淀现象	无	无	微絮	混浊	微絮	微絮	混浊	微絮	微絮

从表 1 看出,所测试金属使人血清蛋白产生沉淀情况,Al^{3+}最敏感 Fe^{3+}次之.说明金属离子诱导血清蛋白产生沉淀的能力与金属离子的半径和带的正电荷数有关,Al^{3+}的高正电荷数与小的离子半径使血清蛋白对其影响最敏感.也与所测试的其他金属离子比较,Al^{3+}对血清蛋白粘度的影响最敏感是一致的.

3 结 论

(1) 人血清蛋白对 Al^{3+}影响十分敏感,特别在 pH≤5 弱酸及酸性环境,Al^{3+}对人血清蛋白液的粘度下降和沉淀的产生均有显著影响,中性环境则影响较小,因而对人体的危害也小[2].

(2) 不同的金属离子与血清蛋白的亲和力不同,使血清蛋白液粘度的变化也不一样,和 Al^{3+}类似测试的其他二价及三价金属离子也对血清蛋白液粘度有影响.但这些二价及三价金属离子为人体必须的微量元素,能与蛋白质结合成人体的必须成份,微量的这些元素的影响对人体有益无害[3].而 Al^{3+}为非人体必须元素,与蛋白质结合后,引起蛋白质粘度下降过多,则成为影响人体的有害因素,成为病变因子.

参考文献:

[1] 孙祥瑞.必须微量元素的营养、生理及临床意义[M].合肥:安徽科技出版社,1982,14.
[2] 区耀华,周昕.金属酶的化学[J].大学化学,1987,(3):11.
[3] 王夔等.生物无机化学[M].北京:清华大学出版社,1988.3~7.

(上接第 143 页)

所以模糊集合在水平 λ 下,关于近似空间的粗糙度是相等的,即 $\rho^{\lambda}_{R_i}(A) = 1/3$,故我们定义模糊集合 A 在 λ 水平下关于近似空间(U,R)的粗糙度定义为

$$E_r^{\lambda}(A) = -\rho^{\lambda}_{R_i}(A)\cdot\sum_{i=1}^{n}p(X_i)\log p_i$$

这里 $q_i = 1/|X_i|$.

根据上述定义,我们有

定理 1.3 设 U 是论域,$R_1,R_2,\cdots,R_n$ 是 U 上的 n 个普通等价关系,若对 $\forall\lambda\in[0,1],A\in F(U)$,有 $E^{\lambda}_{R_i}(A) > E^{\lambda}_{R_j}$,则称 R_j 比 R_i 划分更细,记为 $R_j \prec R_i$.

所以,前面的情况对于模糊集合 $A,\lambda = 0.4$,我们有

$$E^{\lambda}_{R_1}(A) = -\frac{1}{3}\left[\frac{4}{9}\log\frac{1}{4}+\frac{3}{9}\log\frac{1}{3}+\frac{2}{9}\log\frac{1}{2}\right] = 0.165$$

$$E^{\lambda}_{R_2}(A) = 0.12$$

$$E^{\lambda}_{R_3}(A) = 0.075$$

所以 $R_3 \prec R_2 \prec R_1$,即 R_3 最细.对于当论域 U 是连续的情况,我们可以先把其离散化,再进行处理,这也是我们今后的研究方向之一.

参考文献:

[1] 张文修,等.粗糙集理论与方法[M].北京:科学出版社,2001.
[2] Banerjee M, Pal S K. Roughness of a Fuzzy set[J].Information Science, 1996,93:235~246.
[3] 程佚,莫智文.模糊粗糙集及粗糙模糊集的模糊度[J].模糊系统与数学,2001,15:15~17.
[4] 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,(1):35~40.
[5] 王国胤.Rough 集理论与知识获取[M].西安:西安交通大学出版社,2001.5.