一般二元关系下信息系统知识的粒度描述

徐伟华[1],刘士虎[1],张文修[2]
XU Weihua[1],LIU Shihu[1],ZHANG Wenxiu[2]
1.重庆理工大学 数学与统计学院,重庆 400054
2.西安交通大学 理学院,西安 710049
1.School of Mathematics and Statistics,Chongqing University of Technology,Chongqing 400054,China
2.School of Science,Xi'an Jiaotong University,Xi'an 710049,China

XU Weihua,LIU Shihu,ZHANG Wenxiu.Granularity representation of knowledge in information system based on general binary-relation.Computer Engineering and Applications,2011,47(18):40-44.

Abstract: Probability space is established by introducing a portion function in information system based on general binary-relation,this avoids the limitations that systems based on general binary-relation can't be seen as a partition but only a cover. Moreover, granularity representation of knowledge is proposed in information system based on general binary-relation, and some important properties are obtained.It can be proved that the algebra representation of knowledge is equal to the granularity representation of knowledge in information system.Finally,the validity is examined by two examples,and the proposed framework paves the road for the further study of knowledge representation in information system.

Key words: General binary-relation;Granularity;Information system;Knowledge representation;Rough set

摘 要: 在一般二元关系下信息系统中通过引入关系划分函数,建立了系统的概率空间,从而避免了此类系统论域只能形成覆盖而不能构成划分的局限性。进一步给出了一般二元关系下信息系统的粒度描述,得到了其重要性质,并证明了此类系统中知识的代数表示与粒度描述是完全等价的。最后通过实例验证了该粒度描述的有效性,为信息系统知识表示的进一步研究奠定了一定的理论基础。

关键词: 一般二元关系;粒度;信息系统;知识表示;粗糙集
DOI:10.3778/j.issn.1002-8331.2011.18.012 **文章编号:** 1002-8331(2011)18-0040-05 **文献标识码:** A **中图分类号:** TP18

1 引言

1982年波兰数学家Pawlak Z.提出的粗糙集理论[1],是一种新的处理模糊和不确定性知识的软计算工具。它将知识看作是对论域的分类,把分类理解为在特定空间上的(等价)关系。其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。由于思想独特,方法新颖,被广泛应用于人工智能、决策分析、模式识别与数据挖掘等各种领域。

信息系统作为一个具有对象和属性(条件属性和目标属性)关系的数据库,对知识的最终表达是通过属性给出的。但是由于数据表的规模性和多样性(定值性、定量性、离散性、连续性、缺省性、集合值等),使得知识表达的对象与属性的关系不能被直接观察到。这就要求必须依赖于一定的数学方法与计算工具,获得相对易于理解的知识表达模式。

Pawlak Z.从代数学的等价关系和集合的运算角度,提出了知识的代数表示。该方法对知识分类能力强弱的区分,完全是通过由知识导出的划分块的多少来衡量的。但是对信息系统中知识的代数表示法,使得存在概念与运算的直观性较差,以及本质难以理解等一系列问题。基于此,苗夺谦从一个全新的角度考虑了知识,将知识看成是论域的子集组成的 $\sigma-$ 代数上的随机变量,以信息论为基础,提出了知识的信息表示[2-3],并证明了知识约简在信息与代数两种表示下是等价的[4]这一重要结论。虽然这种表示从更深的层次揭示了知识的本质,可是在对知识的表示形式以及计算的复杂度上都还有所欠缺。对此,冯琴荣[5-6]等从数学期望的角度,对知识的表示重新做了研究,给出了知识分类能力的一种量化表示方法,即知识的划分粒度表示法。这在知识的表示形式和计算的复杂程度上,均有所改善。然而,现实生活中存在许多系统均不以等价关系为基础,这给信息系统的知识表示带来了诸多不便。

本文在一般的二元关系下,讨论了信息系统中知识的粒度[7]描述。首先在定义信息系统中一般二元关系的基础上,通过引入划分函数,将一般二元关系基于划分函数导出的知识划分看成是论域 U 的子集组成的 $\sigma-$ 代数上的随机变量。以此为依据,通过给出随机变量的概率分布,定义了知识的粒度值计算表达式,它是划分粒大小的概率平均。进一步结合信息系统中对知识粒度描述的定义,讨论了其重要性质,最后证

基金项目: 重庆市教委科技项目(No.KJ090612);重庆市九龙坡区科技项目(No.2008Q98)。
作者简介: 徐伟华(1979—),男,博士,副教授,硕士生导师,主要研究领域:模糊集,粗糙集,人工智能的数学基础;刘士虎(1984—),男,硕士研究生;张文修(1940—),男,教授,博士生导师。E-mail:datongxuweihua@126.com
收稿日期: 2009-12-23;**修回日期:** 2010-03-24

明了知识的粒度描述与代数表示是完全等价的这一重要结论。

2 基本概念

为方便论述,首先给出一般关系下信息系统中代数表示的主要概念。

定义2.1[8] 设 $I=(U,A,F)$ 为一信息系统,或者称之为数据库系统,其中 U 为对象集,即 $U=(u_1,u_2,\cdots,u_n)$, U 中的每个 $u_i(i\leq n)$ 称为一个对象;而 A 为属性集,即 $A=(a_1,a_2,\cdots,a_m)$, A 中的每个 $a_j(j\leq m)$ 称为一个属性, F为U和A之间的关系集,即

$$F=\{f_j|f_j:U\to V_{a_j},a_j\in A,(1\leq j\leq m)\}$$

其中, V_{a_j} 为属性 a_j 的值域。

定义2.2[8] 设 $I=(U,A,F)$ 为一信息系统,对于 $B\subseteq A$,令

$$R_B=\{(u,v)\in U\times U|uRv,\forall a_i\in B\}$$

称 R_B 为 U 上关于 B 的一个二元关系,此时称信息系统 I 为一般关系下的信息系统。下文若未特别申明,信息系统均指一般关系下的信息系统。

另若记: $[u_i]_{R_B}=\{u_j\in U|(u_i,u_j)\in R_B\}$, $U/R_B=\{[u_i]_{R_B}|u_i\in U\}$, 则称 $[u_i]_{R_B}$ 为 u_i 关于 R_B 的邻域, U/R_B 为该信息系统中对象集 U 在关系 R 下关于 B 的一个分类。

为了更好地反映信息系统中知识的关系,给出以下的定义。

定义2.3[9] 设 $I=(U,A,F)$ 为一信息系统,且 $B,C\subseteq A$ 。

(1)若对任意的 $u\in U$,有 $[u]_{R_B}=[u]_{R_C}$,则称分类 U/R_B 等于分类 U/R_C ,记作 $U/R_B=U/R_C$ 。

(2)若对任意的 $u\in U$,有 $[u]_{R_B}\subseteq[u]_{R_C}$,则称分类 U/R_B 细于分类 U/R_C ,记作 $U/R_B\subseteq U/R_C$ 。

(3)若对任意的 $u\in U$,有 $[u]_{R_B}\subseteq[u]_{R_C}$,而对某些 $v\in U$, 有 $[v]_{R_B}\neq[v]_{R_C}$,称分类 U/R_B 真细于分类 U/R_C ,并记作 $U/R_B\subsetneq U/R_C$ 。

由以上定义,可以得到以下相应的性质。

命题2.1[6] 设 $I=(U,A,F)$ 为一信息系统,且 $B,C\subseteq A$ 。

(1)若 $R_C=R_B$,即知识 R_C 等价于知识 R_B ,则对任意的 $u,v\in U$,有 $uR_Cv\Leftrightarrow uR_Bv$ 。

(2)若 $R_C\geqslant R_B$,即知识 R_B 细于知识 R_C ,则对任意的 $u,v\in U$,有 $uR_Bv\Rightarrow uR_Cv$ 。

(3)若 $R_B\prec R_C$,即知识 R_B 真细于知识 R_C ,则 $R_C\geqslant R_B$, 而且 $R_B\neq R_C$ 。

定义2.4[10] 设 $I=(U,A,F)$ 为一信息系统,对于 $B\subseteq A,b\in B$, 若 $U/R_B=U/R_{B-\{b\}}$,则称 $b\in B$ 是不必要的,否则称 $b\in B$ 是必要的。

定义2.5[10] 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A$,对于每个 $b\in B$ 是必要的,则称 R_B 是独立的,否则称 R_B 是依赖的,或不独立的。

定义2.6[10] 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A$,若进一步满足:(1) R_B 是独立的;(2) $U/R_B=U/R_A$ 。则称 B 是 A 的一个约简,记为 $red(R_B)$ 。

定义2.7[10] 设 $I=(U,A,F)$ 为一信息系统,将 R_A 中所有必要属性组成的集合称之为 R_A 的核,记为 $core(R_A)$,即:

$$core(R_A)=\bigcap_{B\subseteq A}red(R_B)$$

3 知识的粒度描述

文中,认为论域上的任意二元关系,可以看成是定义在其导出的基于划分函数的粒上的随机变量,因此可以定义随机变量的概率分布及其数学期望。由于数学期望是随机变量的重要数字特征,它是随机变量的(概率)平均值。文中,该值等于由一般二元关系在划分函数的基础上定义的粒的平均长度,本文称之为粒度。

由上文知道, $U/R_B=\{[u]_{R_B}|u\in U\}$ 在一般关系下构成覆盖,但未必构成划分。这样,就无法直接以 $U/R_B=\{[u]_{R_B}|u\in U\}$ 中的知识块为随机变量,建立概率空间。为了解决这一问题,首先给出关系划分函数的定义。

定义3.1[10] 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A$,定义关系划分函数如下:

$$j_{R_B}:2^U\to 2^U$$

$$j_{R_B}(X)=\{u\in U|[u]_{R_B}=X\},X\in 2^U$$

显然,若 $j_{R_B}(X)\neq\varnothing$ 当且仅当 x 属于 u 关于 R_B 的邻域簇中, $u\in j_{R_B}(X)\Leftrightarrow[u]_{R_B}=X$ 。由于论域 U 中每一个对象的邻域是唯一的,因此显然有以下性质成立。

定理3.1 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A$,则算子具有以下性质:

(1) $\bigcup_{X\subseteq U}j_{R_B}(X)=U$;

(2)当 $X_i\neq X_j$ 时, $j_{R_B}(X)\cap j_{R_B}(X)=\varnothing$, $X_i,X_j\in 2^U$ 。

证明 由定义3.1直接得证。

注:由于集合族 $\{j_{R_B}(X)|X\in 2^U\}$ 中有很多是空集,故一般不再构成论域 U 的划分,只构成 U 的覆盖。

推论3.1 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A$,则集合簇 $\{j_{R_B}(X)\neq\varnothing|X\in 2^U\}$ 构成了论域 U 上的一个划分。

定义3.2 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A,R_B$ 导出的论域 U 基于算子 j 的划分为

$$\mathcal{J}_{R_B}=\{j_{R_B}(X_1),j_{R_B}(X_2),\cdots,j_{R_B}(X_m)\}$$

称 $j_{R_B}(X_i)\,(i=1,2,\cdots,m)$ 为信息系统 I 的划分粒, $|j_{R_B}(X_i)|$ 为其长度。

定义3.3 设 $I=(U,A,F)$ 为一信息系统,基于 $\mathcal{J}_{R_B}$ 所建立的概率空间上划分粒 $j_{R_B}(X_i)\,(i=1,2,\cdots,m)$ 的概率分布定义为

$$(\mathcal{J}_{R_B},p)=\begin{pmatrix}|j_{R_B}(X_1)| & |j_{R_B}(X_2)| & \cdots & |j_{R_B}(X_m)|\\ p(j_{R_B}(X_1)) & p(j_{R_B}(X_2)) & \cdots & p(j_{R_B}(X_m))\end{pmatrix}$$

其中 $p(j(X_i))=\dfrac{|j(X_i)|}{|U|},i=1,2,\cdots,m$ 。

定理3.2 设 $I=(U,A,F)$ 为一个信息系统, $B,C\subseteq A$,则

$$\mathcal{J}_{R_B}=\{j_{R_B}(X_1),j_{R_B}(X_2),\cdots,j_{R_B}(X_m)\}$$

$$\mathcal{J}_{R_C}=\{j_{R_C}(Y_1),j_{R_C}(Y_2),\cdots,j_{R_C}(Y_n)\}$$

划分粒 $j_{R_B}(X_i),j_{R_C}(Y_i)\,(1\leq i\leq m,1\leq j\leq n)$ 的联合概率分布为

$$(\mathcal{J}_{R_B\cup R_C},p)=\begin{pmatrix}|j_{R_B}(X_1)\cap j_{R_C}(Y_1)| & |j_{R_B}(X_1)\cap j_{R_C}(Y_2)| & \cdots & |j_{R_B}(X_m)\cap j_{R_C}(Y_n)|\\ p(j_{R_B}(X_1)\cap j_{R_C}(Y_1)) & p(j_{R_B}(X_2)\cap j_{R_C}(Y_2)) & \cdots & p(j_{R_B}(X_m)\cap j_{R_C}(Y_n))\end{pmatrix}$$

其中 $p(j_{R_B}(X_i)\cap j_{R_C}(Y_j))=\dfrac{|j_{R_B}(X_i)\cap j_{R_C}(Y_j)|}{|U|}$, $i=1,2,\cdots,m$, $j=1,2,\cdots,n$。

证明 由定义3.2直接得证。

定义3.4 设 $I=(U,A,F)$ 为一个信息系统,知识 R_B 在论域 U 上的划分为 $\mathcal{J}_{R_B}$。称

$$\sum_{i=1}^{m}\left|j_{R_B}(X_i)\right|\cdot p(j_{R_B}(X_i))$$

为知识 R_B 的粒度,记为 $E(R_B)$。

定理3.3 设 $I=(U,A,F)$ 为一个信息系统, $B,C\subseteq A$,则下列命题成立:

(1) $E(R_B)=\sum_{i=1}^{m}\dfrac{|j_{R_B}(X_i)|^2}{|U|}$;

(2) $E(R_{B\cap C})=\sum_{i,j}\dfrac{|j_{R_B}(X_i)\cap j_{R_C}(Y_j)|^2}{|U|}$。

证明 由定理3.2和定义3.4直接得证。

由上文可知:知识的粒度描述是对知识导出的划分中各划分粒"(概率)平均"长度的一种度量。针对某一给定的二元关系 R,当其分类能力越强,即划分的类越多时,所对应的粒度值越小;反之不一定成立。即对某给定的二元关系,其分类所对应的粒度值越小,不一定能说明该二元关系的分类能力就越强,为此,有下面的定义。

定义3.5 设 $I=(U,A,F)$ 为一信息系统, $B,C\subseteq A$,则有下述定义:

(1)若对任意的 i,j,都有 $j_{R_B}(X_i)=j_{R_C}(Z_j)$ 成立,则称划分 $\mathcal{J}_{R_B}$ 等于划分 $\mathcal{J}_{R_C}$,记作 $\mathcal{J}_{R_B}\equiv\mathcal{J}_{R_C}$。

(2)若对任意的 i,j,都有 $j_{R_B}(X_i)\subsetneq j_{R_C}(Z_j)$ 成立,则称划分 $\mathcal{J}_{R_B}$ 强细于划分 $\mathcal{J}_{R_C}$,记作 $\mathcal{J}_{R_B}\sqsubset\mathcal{J}_{R_C}$。

(3)若对任意的 i,j,都有 $j_{R_B}(X_i)\subseteq j_{R_C}(Z_j)$ 成立,则称划分 $\mathcal{J}_{R_B}$ 细于划分 $\mathcal{J}_{R_C}$,记作: $\mathcal{J}_{R_B}\sqsubseteq\mathcal{J}_{R_C}$。

(4)若存在 i_0,j_0,使得 $j_{R_B}(X_{i_0})\subseteq j_{R_C}(Z_{j_0})$ 不成立,且 $\gamma(R_B)\leq\frac{1}{2}|\mathcal{J}_{R_B}|$,则称 $\mathcal{J}_{R_B}$ 弱细于 $\mathcal{J}_{R_C}$,记作 $\mathcal{J}_{R_B}\sqsubset_{\sim}\mathcal{J}_{R_C}$。其中

$$\gamma(R_B)=\sum_{i,j}\chi_{R_B}(\mathcal{J}_{R_B},\mathcal{J}_{R_C})$$

$$\chi_{R_B}(\mathcal{J}_{R_B},\mathcal{J}_{R_C})=\begin{cases}1 & j_{R_B}(X_i)\subseteq j_{R_C}(Z_j)\\ 0 & \text{否则}\end{cases}$$

(5)若 $|\mathcal{J}_{R_B}|=|\mathcal{J}_{R_C}|$,且 $\gamma(R_B)=\gamma(R_C)$,则称划分 $\mathcal{J}_{R_B}$ 弱等于划分 $\mathcal{J}_{R_C}$,记作: $\mathcal{J}_{R_B}\approx\mathcal{J}_{R_C}$。

例3.1 某高校学生有关科目学分的信息系统,如表1。

表1 某高校学生科目学分数据表

学生	科目1	科目2	科目3	科目4	科目5
u_1	2	1	3	3	2
u_2	3	2	1	4	3
u_3	2	1	3	1	2
u_4	2	2	3	2	1
u_5	1	1	4	3	1

若取 $B=\{$科目1,科目2,科目3$\}$ 即可产生该信息系统上的一个优势关系 R_B, $[u_i]^{\geqslant}_{R_B}=\{u_j\in U|f_l(u_i)\geqslant f_l(u_j)\ (\forall a_l\in B)\}$。

由上述定义有

$$[u_1]^{\geqslant}_{R_B}=\{u_1,u_3\},[u_2]^{\geqslant}_{R_B}=\{u_2\},[u_3]^{\geqslant}_{R_B}=\{u_1,u_3\}$$

$$[u_4]^{\geqslant}_{R_B}=\{u_1,u_3,u_4,u_5\},[u_5]^{\geqslant}_{R_B}=\{u_5\}$$

则基于算子 j 产生的划分为 $\mathcal{J}_{R_B}=\{j_{R_B}(X_1),j_{R_B}(X_2),j_{R_B}(X_3),j_{R_B}(X_4)\}$, $X_1=\{u_1,u_3,u_5\}$, $X_2=\{u_2\}$, $X_3=\{u_1,u_3,u_4,u_5\}$, $X_4=\{u_5\}$。且在该划分所建立的概率空间上的概率分布为

$$(\mathcal{J}_{R_B},p)=\begin{pmatrix}|j_{R_B}(X_1)| & |j_{R_B}(X_2)| & |j_{R_B}(X_3)| & |j_{R_B}(X_4)|\\ p(j_{R_B}(X_1)) & p(j_{R_B}(X_2)) & p(j_{R_B}(X_3)) & p(j_{R_B}(X_4))\end{pmatrix}=\begin{pmatrix}2 & 1 & 1 & 1\\ \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\end{pmatrix}$$

其粒度值为 $E(R_B)=1.2$。

若取 $C=\{$科目1,科目2,科目4$\}$ 即可产生该信息系统上的一个优势关系 R_C: $[u_i]^{\geqslant}_{R_C}=\{u_j\in U|f_l(u_i)\geqslant f_l(u_j)\ (\forall a_l\in C)\}$。

由上述定义有

$$[u_1]^{\geqslant}_{R_C}=\{u_1,u_3,u_5\},[u_2]^{\geqslant}_{R_C}=\{u_1,u_2,u_3,u_4,u_5\},[u_3]^{\geqslant}_{R_C}=\{u_3\},$$

$$[u_4]^{\geqslant}_{R_C}=\{u_3,u_4\},[u_5]^{\geqslant}_{R_C}=\{u_5\}。$$

则基于算子 j 产生的划分为 $\mathcal{J}_{R_C}=\{j_{R_C}(X_1),j_{R_C}(X_2),j_{R_C}(X_3),j_{R_C}(X_4),j_{R_C}(X_5)\}$, $X_1=\{u_1,u_3,u_5\}$, $X_2=\{u_1,u_2,u_3,u_4,u_5\}$, $X_3=\{u_3\}$, $X_4=\{u_3,u_4\}$, $X_5=\{u_5\}$。且在该划分所建立的概率空间上的概率分布为

$$(\mathcal{J}_{R_C},p)=\begin{pmatrix}|j_{R_C}(X_1)| & |j_{R_C}(X_2)| & |j_{R_C}(X_3)| & |j_{R_C}(X_4)| & |j_{R_C}(X_5)|\\ p(j_{R_C}(X_1)) & p(j_{R_C}(X_2)) & p(j_{R_C}(X_3)) & p(j_{R_C}(X_4)) & p(j_{R_C}(X_5))\end{pmatrix}=\begin{pmatrix}1 & 1 & 1 & 1 & 1\\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\end{pmatrix}$$

其粒度值为 $E(R_C)=1$。

此时 $E(R_C)<E(R_B)$,即知识 R_C 对论域所做的划分较知识 R_B 所做的划分细,这与从表1中可以明显看出来的事实完全一致。

若记 $D=C\cap B$ 时,进而有

$$[u_1]^{\geqslant}_{R_D}=\{u_1,u_3,u_5\},[u_2]^{\geqslant}_{R_D}=\{u_1,u_2,u_3,u_4,u_5\}\ \ [u_3]^{\geqslant}_{R_D}=\{u_1,u_3,u_5\},$$

$$[u_4]^{\geqslant}_{R_D}=\{u_1,u_3,u_4,u_5\},\ \ [u_5]^{\geqslant}_{R_D}=\{u_5\},$$

基于算子 j 产生的划分为 $\mathcal{J}_{R_D}=\{j_{R_D}(X_1),j_{R_D}(X_2),j_{R_D}(X_3),j_{R_D}(X_4)\}$, $X_1=\{u_1,u_3,u_5\}$, $X_2=\{u_1,u_2,u_3,u_4,u_5\}$, $X_3=\{u_1,u_3,u_4,u_5\}$, $X_4=\{u_5\}$。且在该划分所建立的概率空间上的概率分布为

$$(\mathcal{J}_{R_D},p)=\begin{pmatrix}|j_{R_D}(X_1)| & |j_{R_D}(X_2)| & |j_{R_D}(X_3)| & |j_{R_D}(X_4)|\\ p(j_{R_D}(X_1)) & p(j_{R_D}(X_2)) & p(j_{R_D}(X_3)) & p(j_{R_D}(X_4))\end{pmatrix}=\begin{pmatrix}2 & 1 & 1 & 1\\ \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\end{pmatrix}$$

其联合粒度值为 $E(R_D)=1.2$。

下面讨论粒度的重要性质。

定理3.4 设 $I=(U,A,F)$ 为一信息系统, $B\subseteq A$,且由 B 导出的基于算子 j 的划分为

$$\mathcal{J}_{R_B}=\{j_{R_B}(X_1),j_{R_B}(X_2),\cdots,j_{R_B}(X_m)\}$$

则有

$$1\le\frac{|U|}{m}\le E(R_B)\le|U|$$

证明 不妨取 $|j_{R_B}(X_i)|=k_i$, $|U|=n$,则根据定义3.4有

$$E(R_B)=\sum_{i=1}^{m}\frac{|j_{R_B}(X_i)|^2}{|U|}=\sum_{i=1}^{m}\frac{k_i^2}{n}。$$

由于 $\sum_{i=1}^{m}k_i=n$,所以有:

$$n^2=\sum_{i=1}^{m}k_i^2+2\sum_{1\le s<t\le n}k_sk_t\le m(\sum_{i=1}^{m}k_i^2)$$

即 $\sum_{i=1}^{m}k_i^2\ge\frac{n^2}{m}$,

故 $E(R_B)\ge\frac{|U|}{m}\ge1$

因此,当对任意的 i,j, $k_i=k_j$,即每个划分粒含有相同的对象个数时,粒度值取到最小值。特别地,当对任意的 i, $k_i=1$,即每个粒中只有一个对象时,粒度值取到最小值1。另因为

$$\sum_{i=1}^{m}k_i^2\le(\sum_{i=1}^{m}k_i)^2=n^2,$$

所以 $E(R_B)\le n=|U|$,且等号成立当且仅当只有一个划分粒,即 U 中的任何对象在关系 R 下均无法区分的情况。

上述定理表明,当分类能力越强时,粒度越小。特别地,当 $m=|U|$ 时,即每个对象的领域中只含有自身时,分类能力达到最强,粒度显然取到最小,这与直观理解是完全相符合的。

定理3.5 设 $I=(U,A,F)$ 为一信息系统, $B,C\subseteq A$,下列命题成立:

(1)若 $\mathcal{J}_{R_B}\sqsubseteq J_{R_C}$,则 $E(R_B)\le E(R_C)$;

(2)若 $\mathcal{J}_{R_B}\sqsubset J_{R_C}$,则 $E(R_B)<E(R_C)$。

证明 (1)若 $\mathcal{J}_{R_B}\sqsubseteq\mathcal{J}_{R_C}$,则由定义3.5知,对任意的 $j_{R_B}(X_i)$,存在 $j_{R_C}(Z_J)$,使得 $j_{R_B}(X_i)\subseteq j_{R_C}(Z_i)$ 成立,由粒度的定义可得 $E(R_B)\le E(R_C)$。

(2)若 $\mathcal{J}_{R_B}\sqsubset\mathcal{J}_{R_C}$,则由定义3.5知,对任意的 $j_{R_B}(X_i)$,存在 $j_{R_C}(Z_J)$,使得 $j_{R_B}(X_i)\subseteq j_{R_C}(Z_i)$ 成立,且存在 i_0,j_0,使得 $j_{R_B}(X_{i_0})\ne j_{R_C}(Z_{j_0})$ 成立,故: $E(R_B)<E(R_C)$。

该性质表明,知识越细,则它的粒度值越小;相反知识约粗,它的粒度值越大。

推论3.2 设 $I=(U,A,F)$ 为一信息系统, $B,C\subseteq A$,则 $E(R_{B\cap C})\le\min(E(R_B),E(R_C))$。

定理3.6 设 $I=(U,A,F)$ 为一信息系统, $B,C\subseteq A$,且知识 R_B,R_C 导出的论域 U 上的划分分别为

$$\mathcal{J}_{R_B}=\{j_{R_B}(X_1),j_{R_B}(X_2),\cdots,j_{R_B}(X_m)\}$$

$$\mathcal{J}_{R_C}=\{j_{R_C}(Z_1),j_{R_C}(Z_2),\cdots,j_{R_C}(Z_{m-1}),j_{R_C}(Z_m),j_{R_C}(Z_{m+1})\}$$

若 $\mathcal{J}_{R_B},\mathcal{J}_{S_B}$ 满足:

(1) $j_{R_B}(X_i)=j_{R_C}(Z_i)$, $i=1,2,\cdots,m$;

(2) $j_{R_B}(X_m)=j_{R_C}(Z_m)\cup j_{R_C}(Z_{m+1})$;

(3) $j_{R_C}(Z_m)\cap j_{R_C}(Z_{m+1})=\varnothing$。

则有下式成立:

$$E(R_B)=E(R_C)+2\frac{|j_{R_C}(Z_m)||j_{R_C}(Z_{m+1})|}{|U|}。$$

证明 由定理3.5即证。

例3.2 由例3.1中可知:

$$1<\frac{4}{5}<E(R_B)=\frac{7}{5}<5$$

$$1=\frac{5}{5}=E(R_C)<5$$

而且进一步可得

$$1=E(R_{B\cap C})\le\min(E(R_B),E(R_C))=1$$

由于 $\mathcal{J}_{R_B}=\{\{u_1,u_3\},\{u_2\},\{u_4\},\{u_5\}\}$ 且 $\mathcal{J}_{R_C}=\{\{u_1\},\{u_2\},\{u_3\},\{u_4\},\{u_5\}\}$,故 $\mathcal{J}_{R_C}\sqsubset\mathcal{J}_{R_B}$。这与例3.1中的结果 $E(R_C)<E(R_B)$ 完全一致。另一方面又可验证表1所给的信息系统的划分粒满足定理3.6的条件,故有

$$E(R_B)=E(R_C)+2\times\frac{1\times1}{5}=\frac{7}{5}$$

4 知识的粒度描述与代数表示的关系

前面分析了信息系统中知识与粒度之间的关系,可以看出用粒度能定量的衡量知识分类能力的强弱。事实上就信息系统而言,知识的代数表示与粒度描述是等价的,下面给出详细的论述。

定理4.1 设 $I=(U,A,F)$ 为一信息系统, $B,C\subseteq A$,若 $U/R_B=U/R_C$,则 $E(R_B)=E(R_C)$。

证明 由于 $U/R_B=U/R_C$,所以 R_B,R_C 在论域上导出的划分相同,即 $\mathcal{J}_{R_B}=\mathcal{J}_{R_C}$,它们在建立的概率空间上的概率分布也相同。由粒度的计算公式可以立即得到 $E(R_B)=E(R_C)$。

该定理表明,两个代数表示等价的知识,具有相等的粒度,亦即具有相同的分类能力。但是,该定理的逆定理不一定成立。

定理4.2 设 $I=(U,A,F)$ 为一信息系统, $B,C\subseteq A$,若 $R_B\subseteq R_C$ 且有 $E(R_B)=E(R_C)$,则 $U/R_B=U/R_C$。

证明 不妨记

$$\mathcal{J}_{R_B}=\{j_{R_B}(X_1),j_{R_B}(X_2),\cdots,j_{R_B}(X_m)\}$$

$$\mathcal{J}_{R_C\backslash R_B}=\{j_{R_C\backslash R_B}(Y_1),j_{R_C\backslash R_B}(Y_2),\cdots,j_{R_C\backslash R_B}(Y_n)\}$$

根据定义3.4以及定理3.3知

$$E(R_B)=\sum_{i=1}^{m}\frac{|j_{R_B}(X_i)|^2}{|U|}$$

$$E(R_C)=E(R_B\cup(R_C\backslash R_B))=\sum_{i,j}\frac{|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_j)|^2}{|U|}$$

由 $E(R_B)=E(R_C)$ 可得

$$|j_{R_B}(X_i)|^2=\sum_{j=1}^{n}|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_j)|^2。$$

即 $\sum_{j=1}^{n}|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_j)|)^2=\sum_{j=1}^{n}|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_j)|^2$

把上式展开,经计算可知

$$\sum_{j<k}|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_j)|\cdot|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_k)|=0$$

也就是对每个 $j<k$ 有:

$$|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_j)|\cdot|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_k)|=0$$

又由于 $j_{R_C\backslash R_B}(Y_j),j_{R_C\backslash R_B}(Y_k)$ 取遍所有的划分块,而 $\bigcup_{j=1}^{n}j_{R_C\backslash R_B}(Y_j)=U$, $j_{R_B}(X_i)\subseteq U$,故存在 j_0,k_0,使得

$$|j_{R_B}(X_i)\cap j_{R_C\backslash R_B}(Y_{J_0})|=0$$

$|j_{R_B}(X_i)\cap j_{R_C\setminus R_B}(Y_{k_0})|\neq 0$

对于 $|j_{R_B}(X_i)\cap j_{R_C\setminus R_B}(Y_{k_0})|\neq 0$ 而言，若 $j_{R_B}(X_i)\subset j_{R_C\setminus R_B}(Y_{k_0})$，则存在正整数 $l_0\neq k_0$，使得

$|j_{R_B}(X_i)\cap j_{R_C-R_B}(Y_{l_0})|\neq 0$

事实上，这种情况是不会出现的，否则由 $|j_{R_B}(X_i)\cap j_{R_C-R_B}(Y_{l_0})|\cdot|j_{R_B}(X_i)\cap j_{R_C\setminus R_B}(Y_{k_0})|\neq 0$ 产生矛盾。因此对任意的 $j_{R_B}(X_i)$，存在 $j_{R_C\setminus R_B}(Y_j)$，使得 $j_{R_B}(X_i)\subset j_{R_C\setminus R_B}(Y_j)$。由知识的依赖性知 $U/R_B=U/R_C$。

定理4.2表明，若两个知识之间存在包含关系的时候，可以由粒度相等推出在代数表示下仍是相等的。

定理4.3 设 $I=(U,A,F)$ 为一信息系统，$B\subseteq A$，则 $b\in B$ 是不必要的当且仅当 $E(R_{B-\{b\}})=E(R_B)$。

证明 必要性可由定理4.1得证，充分性可由定理4.2得证。由定理4.3可知，给属性子集添加一个不必要的属性不会改变粒度的大小。

推论4.1 设 $I=(U,A,F)$ 为一信息系统，$B\subseteq A$，$b\in B$ 是不必要的当且仅当 $E(R_{B-\{b\}})>E(R_B)$。

定理4.4 设 $I=(U,A,F)$ 为一信息系统，$B\subseteq A$，知识 R_B 是独立的当且仅当对任意的 $b\in B$，有 $E(R_{B-\{b\}})>E(R_B)$。

证明 由独立的定义和上述推论可以得证。

定理4.5 设 $I=(U,A,F)$ 为一信息系统，$B\subseteq A$ 是关于知识 R_B 的一个约简的充分必要条件为：

(1) $E(R_A)=E(R_B)$；

(2)对任意的 $b\in B$，$E(R_{B-\{b\}})>E(R_B)$ 成立。

证明 由定理4.2和定理4.4即可得证。

上述这些定理表明，本文给出的粒度描述与代数表示是完全等价的。

5 结论

Pawlak Z.提出了对信息系统中知识的代数表示法；苗夺谦提出了知识的信息表示并证明了知识约简在信息与代数两种表示下是等价的这一重要结论；冯琴荣等人从数学期望的角度对知识的表示，给了一种新的方法，即粒度表示法。但是他们都是以等价关系为基础，对知识产生划分，而在现实中有好多问题都是不以等价关系为基础的。为此，在一般二元关系信息系统的基础上，通过引入划分函数，将一般二元关系基于划分函数导出的知识划分看成是论域 U 的子集组成的 $\sigma-$代数上的随机变量。以此为依据，通过给出随机变量的概率分布，定义了知识的粒度值计算表达式，它是划分粒大小的概率平均。进一步结合信息系统中对知识粒度描述的定义，讨论了其重要性质，最后证明了知识的粒度描述与代数表示是完全等价的这一重要结论。本文的方法以及得到的相关结论给信息系统中知识描述的进一步研究，奠定了一定的理论基础。

参考文献：

[1] Pawlak Z.Rough sets:theoretical aspects of reasoning about data[M].Boston:Kluwer Academic Publishers,1991.

[2] 苗多谦，胡桂荣.知识约简的一种启发式算法[J].计算机研究与发展，1999，36(6)：681-684.

[3] 苗多谦，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.

[4] 苗多谦.Rough Set理论及其在机器学习中的应用研究[D].北京：中国科学院自动化研究所，1997.

[5] 冯琴荣.粗糙集的期望表示[J].山西师范大学学报：自然科学版，2007，21(2)：24-29.

[6] 冯琴荣，苗多谦.知识的划分粒度表示法[J].模式识别与人工智能，2009，22(1)：64-69.

[7] 张文修，徐伟华.基于粒计算的认知模型[J].工程数学学报，2007，24(6)：957-971.

[8] 张文修，梁怡，吴伟志.信息系统与知识发现[M].北京：科学出版社，2003.

[9] 徐伟华，张晓燕.序信息系统中基于粗糙熵的不确定性度量[J].工程数学学报，2009，26(2)：283-289.

[10] 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.

(上接31页)

表3 3种算法结果比较

机器号	遗传算法SPT调度序列	蚁群算法SPT调度序列	CPOS优化EDD调度序列
M_1	051,081,041,091,043,053,023	031,041,081,053	051,041,031,023,053
M_2	012,031	051,091,012,023,043	091,081,012,043
M_3	021,011,082,061,032	021,011,032,082,061	021,011,082,032,061
M_4	101,042,083,054	071,022,042,054	071,022,042,054
M_5	071,072,052,092,022	101,072,092,083	101,072,092,083
M_6	013	013,052	052,013
生产周期	48	45	41

的调度。针对调度算法，本文研究了粒子群优化算法的早熟收敛的原因，将混沌思想引入粒子群算法，提出了一种基于Logistic映射的混沌粒子群优化(CPSO)算法。在此基础上，进行了相关的仿真实验，结果表明该优化算法不但具有很强的全局搜索能力和较快的收敛速度，而且能有效避免粒子群优化算法的早熟收敛问题。

参考文献：

[1] 吕晓慧，马金平，阮家港.一种混合蚁群算法在JSP问题中的应用研究[J].科学技术与工程，2008，8(23)：6361-6364.

[2] 秦娜，乐晓波，刘武.基于Petri网模型的JSP粒子群优化调度[J].计算机应用，2008，28(8)：2166-2169.

[3] 李小青，张文祥.混沌粒子群算法及其在优化设计中的应用[J].计算机系统应用，2009(4)：171-174.

[4] 陶泽，隋天中.基于Petri网和GASA的双资源JSP动态优化调度[J].东北大学学报：自然科学版，2008，34(2)：46-49.

[5] 刘军民，高岳林.混沌粒子群优化算法[J].计算机应用，2008，28(2)：322-325.

[6] 袁崇义.Petri网原理与应用[M].北京：电子工业出版社，2005：58-64.