计算机科学 2002Vol. 29№. 10

Pawlak 粗糙集模型的随机集表示及合成

The Random Set Expression of Pawlak Rough Set Model and its Composition

陈德刚　张文修

（西安交通大学理学院信息与系统工程研究所　西安 710049）

Abstract　In this paper we express the Pawlak rough set model by random set, we also study the composition of two Pawlak rough set models.

Keywords　Information system, Rough set, Random set

1. 引言和预备知识

粗糙集作为一种处理不精确、不确定与不完全数据的新的数学理论，最初是由波兰数学家 Z. Pawlak[1]于 1982 年提出的。在 Pawlak 粗糙集模型中，论域中的对象或元素可以用可利用的信息（或知识库中的知识）来描述。当两个不同的对象具有相同的描述时称这两个元素是不可区分的。所有具有相同描述的元素构成了一个等价类，所有等价类构成了这个论域的一个划分，任意给定论域的一个子集，人们不一定能用知识库中的知识来精确地描述，这时就用关于这个集合的一对上、下近似来描述。粗糙集理论的主要思想是利用已知的知识库，将不精确或不确定的知识用已知知识库中的知识来（近似）刻画。

设 U 是非空有限论域，R 是 U 上的二元等价关系，序对 $A=(U,R)$ 称为近似空间，$X\subseteq U$，X 关于 $A=(U,R)$ 的下近似集 $\underline{apr}_R X$ 和上近似集 $\overline{apr}_R X$ 定义为：

$$\underline{apr}_R X=\bigcup\{[x]:[x]\subseteq X\}=\{x\in U:[x]\subseteq X\}$$

$$\overline{apr}_R x=\bigcup\{[x]:[x]\cap X\neq\phi\}=\{x\in U:[x]\cap X\neq\phi\}$$

若 $\underline{apr}_R X=\overline{apr}_R X$，则称 X 是可定义的，否则称 X 是不可定义的或粗糙的。

粗糙集理论中的知识表达方式一般采用信息系统的形式，它可以表示为四元有序组 $K=(U,A,V,\rho)$，其中 U 是论域，A 是属性全体，$V=\bigcup_{a\in A}V_a$，V_a 是属性 a 的值域，$\rho:U\times A\to V$ 是一个信息函数，$\rho_x:A\to V$，$x\in U$，反映了对象 x 在 K 中的完全信息，其中 $\rho_x(a)=\rho(x,a)$。

对于这样的信息系统，每个属性子集就定义了论域上的一个等价关系，即 $\forall B\subseteq A$，定义 $R_B:xR_By\Leftrightarrow\rho_x(b)=\rho_y(b)$，$\forall b\in B$。

在实际应用中，信息函数 ρ 往往是由专家给出的，在 U，A，V 相同的前提下，不同的专家给出的信息函数 ρ 可能不尽一致，因此考虑如何把不同的专家给出的信息系统进行合成就是十分必要的，在实际应用中有着重要的意义。本文试图在这一方面进行一些探讨。

2. 信息系统中 Pawlak 粗糙集模型的随机集表示

设 (Ω,Φ)，$(2^W,A)$ 是两个可测空间，集值函数 F 称为 $\Phi-A$ 随机集，若 F 是 $\Phi-A$ 可测的，即对于 $\forall A\in A$，有 $\{\omega\in\Omega:F(\omega)\in A\}\in\Phi$。当 Ω 和 W 是两个非空有限集合，若考虑 $\Phi=2^\Omega$，$A=\sigma(2^W)$，即 A 是 2^W 的幂集时，则集值函数 $F:\Omega\to 2^W$ 是一个随机集。

设 $K=(U,A,V,\rho)$ 是一个信息系统，由于多个属性的信息系统可以通过定义不可区分关系归结为单一属性的信息系统[5]，因此只需考虑单一属性时的情况，这时 K 可以表示为 $K=(U,a,V_a,\rho)$。如果定义 $R_a:xR_ay\Leftrightarrow\rho(x)=\rho(y)$，则 R_a 是 U 上的一个等价关系。

定义 $F:(V_a,2^{V_a})\to(2^U,\sigma(2^U))$，$F(p)=\rho^{-1}(p)$，$p\in V_a$，则 F 是一个随机集。对于 $\forall X\subseteq U$，X 在 V_a 中关于 F 的下近似集和上近似集定义为：

$$\underline{apr}_F X=\{p\in V_a:F(p)\subseteq X\}$$

$$\overline{apr}_F X=\{p\in V_a:F(p)\cap X\neq\varnothing\}$$

$\underline{apr}_F X$ 可以解释为 K 中肯定支持 X 的属性值（知识）的全体，$\overline{apr}_F X$ 可以解释为 K 中可能支持 X 的属性值（知识）的全体。这样定义的 X 在 V_a 中的下、上近似集与由 R_a 定义的 X 在 U 中的下、上近似集是完全一致的，有如下定理。

定理 1　$\underline{apr}_{R_a}X=\rho^{-1}(\underline{apr}_F X)$，$\overline{apr}_{R_a}X=\rho^{-1}(\overline{apr}_F X)$。

证明：

$$\begin{aligned}\underline{apr}_{R_a}X&=\bigcup\{[x]_{R_a}:[x]_{R_a}\subseteq X\}\\&=\bigcup\{F(p):F(p)\subseteq X\}=\bigcup\{\rho^{-1}(p):F(p)\subseteq X\}\\&=\rho^{-1}(\underline{apr}_F X)。\end{aligned}$$

同理可证 $\overline{apr}_{R_a}X=\rho^{-1}(\overline{apr}_F X)$。

易见 $\underline{apr}_{R_a}X=\overline{apr}_{R_a}X\Leftrightarrow\underline{apr}_F X=\overline{apr}_F X$，因此用随机集 F 定义的 X 的下、上近似集与用等价关系 R_a 定义的 X 的下、上近似集对 X 的描述是完全一致的，前者可以看成是后者的随机集表示。

对任意的一个随机集 $F:(V_a,2^{V_a})\to(2^U,\sigma(2^U))$，如果它满足 $\{F(p):p\in V_a\}$ 是 U 的一个划分，则它定义了一个单值信息函数 $\rho:U\to V_a$，$\rho(x)=p$，$x\in F(p)$，此时信息系统 (U,a,V_a,ρ) 也可以用 (U,a,V_a,F) 来等价描述。如果 $\{F(p):p\in V_a\}$ 不是 U 的一个划分，则它可以定义一个集值信息函数 $\rho(x)=\{p:x\in F(p)\}$，对这种不完备的信息系统在文[2]中同样用随机集的方法进行研究。

用 F 定义的下、上近似集有如下性质。

定理 2

(1) $\underline{apr}_F X=(\overline{apr}_F X^C)^C$；$\overline{apr}_F X=(\underline{apr}_F X^C)^C$，$X\subseteq U$；

(2) $\underline{apr}_F U=V_a$，$\overline{apr}_F\phi=\phi$；

陈德刚　博士后。张文修　教授，博士生导师。

(3)$\underline{apr}_F(X\cap Y)=\underline{apr}_F X\cap\underline{apr}_F Y$, $\overline{apr}_F(X\cup Y)=\overline{apr}_F X\cup\overline{apr}_F Y$, $X,Y\subseteq U$;

(4)$\underline{apr}_F X\subseteq\overline{apr}_F X$;

(5)$\underline{apr}_F(X\cup Y)\supseteq\underline{apr}_F X\cup\underline{apr}_F Y$, $\overline{apr}_F(X\cap Y)\subseteq\overline{apr}_F X\cap\overline{apr}_F Y$, $X,Y\subseteq U$;

(6)若 $X\subseteq Y\subseteq U$,则$\underline{apr}_F X\subseteq\underline{apr}_F Y$, $\overline{apr}_F X\subseteq\overline{apr}_F Y$。

对 $\forall A\subseteq V_a$,定义 $P(A)=|\rho^{-1}(A)|/|U|$,则 P 是$(V_a, 2^{V_a})$上的概率测度,即$(V_a,2^{V_a},P)$是一个概率空间,若 F 为如上定义的随机集,定义 $m(B)=P(F^{-1}(B)),B\in 2^U$,则 m 是概率分配函数[2]。若记 $B(A)=P(\underline{apr}_F A)$, $L(A)=P(\overline{apr}_F A)$, $A\in 2^U$,则 $B(A)$和 $L(A)$分别是 2^U 上的信任函数和似然函数[2],并且有 $B(A)=\sum_{B\subseteq A}m(B)$, $L(A)=\sum_{B\cap A\neq\phi}m(B)$, A 的信任区间为$[B(A),L(A)]$。在实际应用中,往往是先由专家给出概率分配函数(解释为专家的一种评价),再得到信任函数和似然函数。

3. 信息系统中 Pawlak 粗糙集模型的合成

设 $K=(U,a,V_a,\rho)$是一个信息系统,则由前所述可知存在一个等价关系 R_a 与之对应,它可以对论域 U 进行划分,产生一个 Pawlak 粗糙集模型。但不同的信息系统可以产生相同的等价关系,看下例。

例 1 设 $U=\{x_1,x_2,\cdots,x_6\}$, a 是某一个属性, $V_a=\{p_1,p_2,p_3\}$,信息函数

$\rho_1:\rho_1(x_1)=\rho_1(x_2)=p_1,\rho_1(x_3)=\rho_1(x_4)=p_2,\rho_1(x_5)=\rho_1(x_6)=p_3$,

$\rho_2:\rho_2(x_1)=\rho_2(x_2)=p_2,\rho_2(x_3)=\rho_2(x_4)=p_3,\rho_2(x_5)=\rho_2(x_6)=p_1$,

则 $K_1=(U,V_a,a,\rho_1)$和 $K_2=(U,V_a,a,\rho_2)$是两个信息系统,$K_1\neq K_2$,但易见它们所对应的等价关系是相同的。

之所以会出现上述情况,就是因为在由信息函数 ρ 定义等价关系 R_a 时,把属性值的区别没有考虑在内。在研究单个信息系统时,这样做既方便同时也容易解释。但是在考虑多个信息系统的合成时,如果不考虑属性值的区别,就会产生不可调和的矛盾。用随机集定义的粗糙集模型把属性值的区别考虑在内,这也是我们用随机集定义粗糙集模型的主要原因。

设 $K_1=(U,a,V_a,\rho_1)$, $K_2=(U,a,V_a,\rho_2)$是两个信息系统, F_1,F_2 分别是由 ρ_1,ρ_2 定义的随机集,令 $F:V_a\to 2^U$, $F(p)=F_1(p)\cup F_2(p),p\in V_a$,则 F 是一个随机集。对$\forall X\subseteq U$,可由 F 定义 X 的一对下、上近似集

$\underline{apr}_F X=\{p:F(p)\subseteq X\}$, $\overline{apr}_F X=\{p:F(p)\cap X\neq\phi\}$,

称为由 F_1,F_2 定义的 X 的下、上近似的合成。由 F,F_1,F_2 定义的 X 的下、上近似有如下关系。

定理 3 1)$\underline{apr}_F X=\underline{apr}_{F_1}X\cap\underline{apr}_{F_2}X$;

2)$\overline{apr}_F X=\overline{apr}_{F_1}X\cup\overline{apr}_{F_2}X$。

证明:

1)$p\in\underline{apr}_F X\Leftrightarrow F(p)\subseteq X\Leftrightarrow F_1(p)\subseteq X,F_2(p)\subseteq X\Leftrightarrow p\in\underline{apr}_{F_1}X\cap\underline{apr}_{F_2}X$。

2)$p\in\overline{apr}_F X\Leftrightarrow F(p)\cap X\neq\phi\Leftrightarrow F_1(p)\cap X\neq\phi$ 或 $F_2(p)\cap X\neq\phi\Leftrightarrow p\in\overline{apr}_{F_1}X\cup\overline{apr}_{F_2}X$。

易证由 F 定义的下、上近似满足定理 2。

由 F 定义的 X 的下近似可以解释为在两个信息系统(知识库)中都支持 X 的属性值(知识)的集合,X 的上近似可以解释为至少有一个信息系统(知识库)认为有可能支持 X 的属性值(知识)的集合。合成后的 X 的下近似比原来的两个都要小,而 X 的上近似比原来的两个都要大,这也比较符合实际情况。如果$\underline{apr}_{F_1}X\neq\phi$, $\underline{apr}_{F_2}X\neq\phi$,但$\underline{apr}_F=\phi$,则说明两个知识库对概念 X 的认识差别很大,甚至相反。如果 $K_1=K_2$,则 $F=F_1=F_2$,此时即是两个信息系统完全相同,故对概念的描述合成前与合成后也是一致的。

一般来说,若 $F_1\neq F_2$,则$\{F(p):p\in V_a\}$不构成 U 的一个划分,F 就不能定义一个信息函数,也就不能定义一个等价关系。即尽管信息系统 K_1,K_2 分别都是完备的,但它们的合成就不一定是完备的,在实际例子中接触到的也往往是不完备的信息系统。因此,如果用 $K=(U,a,V_a,F)$来表示不完备信息系统,则更具有广泛的意义,此时信息系统 $K=(U,a,V_a,F)$可以看成是 $K_1=(U,a,V_a,F_1)$与 $K_2=(U,a,V_a,F_2)$的合成。这样的合成也易于从 K_1,K_2 中提取共同的知识,并且适合两个以上信息系统的合成。

易见$\bigcup\{F(p):p\in V_a\}=U$,因此可以由 F 定义一个集值的信息函数 $\rho(x)=\{p:x\in F(p)\}$。事实上,由 F 定义的 ρ 即是文[2]中研究的随机集。同样给出一个集值信息函数 ρ,可以定义一个随机集 $F(p)=\{x:p\in\rho(x)\}$。文[2]中和本文用随机集定义下、上近似算子的方法虽相似,但研究的内容本质上是不同的。文[2]中研究的是 V_a 的子集在 U 中的近似,而本文讨论的是 U 的子集在 V_a 中的近似,当然可以把两者结合起来研究。

由 F_1,F_2 合成的随机集 F 也可以定义 $X\subseteq U$ 在 U 中的下、上近似集如下:$\underline{apr}_{F_U}X=\bigcup\{F(p):F(p)\subseteq X\}$, $\overline{apr}_{F_U}X=\bigcup\{F_i(p):F_i(p)\cap X\neq\phi,i=1,2\}$,易证它们有如下性质。

定理 4 1)$\underline{apr}_{F_U}X\subseteq X\subseteq\overline{apr}_{F_U}X$;

2)$\underline{apr}_{F_U}\phi=\overline{apr}_{F_U}\phi=\phi$, $\underline{apr}_{F_U}U=\overline{apr}_{F_U}U=U$;

3)$\overline{apr}_{F_U}(X\cup Y)=\overline{apr}_{F_U}X\cup\overline{apr}_{F_U}Y$;

4)$\underline{apr}_{F_U}(X\cap Y)=\underline{apr}_{F_U}X\cap\underline{apr}_{F_U}Y$;

5)$X\subseteq Y\Rightarrow\underline{apr}_{F_U}X\subseteq\underline{apr}_{F_U}Y$, $\overline{apr}_{F_U}X\subseteq\overline{apr}_{F_U}Y$;

6)$\underline{apr}_{F_U}(X\cup Y)\supseteq\underline{apr}_{F_U}X\cup\underline{apr}_{F_U}Y$;

7)$\overline{apr}_{F_U}(X\cap Y)\subseteq\overline{apr}_{F_U}X\cap\overline{apr}_{F_U}Y$;

8)$\underline{apr}_{F_U}(\underline{apr}_{F_U}X)=\underline{apr}_{F_U}X$, $\overline{apr}_{F_U}(\overline{apr}_{F_U}X)=\overline{apr}_{F_U}X$;

一般来说,$\underline{apr}_{F_U}X$ 和 $\overline{apr}_{F_U}X$ 不再具有对偶性,但是 $\overline{apr}_F X$ 和 $\underline{apr}_F X$ 具有对偶性,这也是本文用随机集来合成信息系统的一个原因。

设 m_1,m_2 分别是前面定义的 K_1,K_2 上的概率分配函数,则由 Dempster-Shafer 合成公式[3]m_1,m_2 可合成一个新的概率分配函数,$m:m(\phi)=0$, $m(A)=\frac{1}{N}\sum_{E\cap F=A}m_1(E)m_2(F)$, $A\neq\phi$,其中 $N=\sum_{E\cap F\neq\phi}m_1(E)m_2(F)$($N>0$ 为显然,故 m_1,m_2 可合成),由 m 可以定义 2^U 上的信任函数和似然函数。

参 考 文 献

1 Pawlak Z. Rough sets. International Journal of Computer Information Sciences, 1982, 11: 341～359

2 张文修,吴伟志. 基于随机集的粗糙集模型(I). 西安交通大学学报, 2000, 34(12): 15～19

3 张文修,梁怡. 不确定性推理原理. 西安交通大学出版社, 1996

4 张文修,吴伟志. 粗糙集理论介绍和研究综述. 模糊系统与数学, 2000, 4: 1～12

5 张文修,吴伟志,梁吉业,李德玉. 粗糙集理论与方法. 科学出版社, 2001