

粗糙集近似与信息粒度^{*})徐优红¹ 竺定宏²

(浙江海洋学院数理与信息学院 舟山 316004) (浙江国际海运职业技术学院海洋旅游学院 舟山 316012)

摘要 用粗糙集近似描述了三类常见信息系统(即 Pawlak 信息系统、不完备信息系统、不完备模糊信息系统)中对象的基本信息粒度。通过信息系统中对象属性值关于对象属性近似空间的上近似可以得到与对象具有相同或相似信息的对象集,即利用对象属性值关于对象属性近似空间的上近似将对象属性值信息变换成为对象的基本信息粒度。所得结论对信息系统中基本信息粒度的物理意义有了比较清楚和更加合理的解释。

关键词 粗糙集,信息系统,近似算子,模糊集,信息粒度

Rough Set Approximations and Information Granules

XU You-Hong¹ ZHU Ding-Hong²

(School of Mathematics, Physics, and Information Science, Zhejiang Ocean University, Zhoushan 316004)

(Ocean Tourism School, Zhejiang International Maritime College, Zhoushan 316012)

Abstract In this paper, we prove that basic information granules in Pawlak information systems, incomplete information systems, and incomplete fuzzy information systems can be respectively described as three types of rough set approximations. By using the upper approximations of attribute value of an object with respect to the object-attribute approximation space, we can obtain an object set in which all the objects have the same or similar information, that is, we can transform the object-attribute value into basic information granule for each object. The obtained results give clear and reasonable explanation of the physical meaning of basic information granules.

Keywords Rough sets, Information systems, Approximation operators, Fuzzy sets, Information granules

1 引言

粗糙集在智能信息处理中的应用通过信息系统来表示,这种信息系统也称对象-属性值表,对象的属性取值反映了该对象在系统中的完全信息。

在 Pawlak 信息系统^[1,2](即经典的完备信息系统)中,对于每一个属性子集,由对象的属性取值可以得到对象集合上的不可分辨关系,从而得到和每一个对象有相同信息的等价类。在不完备信息系统中^[3,4],对于每一个属性子集,由对象的属性取值可以定义对象集合上的相似关系,从而得到和每一个对象可能有相同信息的相似类。在模糊信息系统中,对于每一个属性子集,由对象的属性取值可以定义对象集合上的模糊相似关系,从而得到和每一个对象可能有相同信息的模糊相似类^[5]。这些等价类或相似类构成了对象在信息系统中的基本信息粒度,利用这些基本信息粒度可以获得每一个对象集合的下近似和上近似。当信息系统成为一个决策表时,利用每一个决策类关于条件属性集的下近似和上近似可以分别挖掘出蕴含在系统中的确定性决策规则和可能性决策规则^[2,6]。

显然,对象在信息系统中的基本信息粒度是通过定义在对象集合(即论域)上二元关系得到的。由于集合的下近似和上近似有比较明确的物理意义,若对象在信息系统中的基本信息粒度可以表示为某一个集合的下近似或上近似,则信息系统中对象的基本信息粒度可以通过粗糙集近似得到比较清楚的解释。本文我们说明 Pawlak 信息系统中的等价类、不完备信息系统中的相似类、不完备模糊信息系统中的模糊相似类这三类基本信息粒度都可以通过对象属性值关于对象属性近似空间的上近似来描述。

2 粗糙集与近似算子

本文恒设 U 和 W 都是有限非空论域, $P(U)$ 和 $F(U)$ 分别表示 U 中的经典子集全体和模糊子集全体。

定义 1 设 R 是从 U 到 W 上的一个二元经典关系,称三元组 (U, W, R) 为广义近似空间。 $\forall X \subseteq W$, A 关于近似空间 (U, W, R) 的下近似 $\underline{R}(X)$ 和上近似 $\bar{R}(X)$ 分别定义如下:

$$\underline{R}(A) = \{x \in U : R_S(x) \subseteq X\}$$

$$\bar{R}(A) = \{x \in U : R_S(x) \cap X \neq \emptyset\}$$

其中 $R_S(x) = \{y \in U : (x, y) \in R\}$ 。

序对 $(\underline{R}(X), \bar{R}(X))$ 称为广义经典粗糙集,算子 $\underline{R}$ 与 $\bar{R} : P(W) \rightarrow P(U)$ 分别称为(广义)粗糙下、上近似算子。

注:显然,若 $U=W$,且 R 是 U 上的等价关系,则当 $X \subseteq U$ 时,有

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

$$\begin{aligned} \bar{R}(X) &= \{x \in U : [x]_R \cap X \neq \emptyset\} \\ &= \bigcup \{[x]_R : [x]_R \cap X \neq \emptyset\} \end{aligned}$$

故广义近似算子是 Pawlak 经典近似算子的推广形式。

定义 2 设 (U, W, R) 为广义近似空间, $\forall X \in F(W)$, 模糊集 X 关于近似空间 (U, W, R) 的上近似 $\bar{R}(X)$ 与下近似 $\underline{R}(X)$ 是 U 上的一对模糊子集,其隶属函数定义如下:

$$\bar{R}(X)(x) = \bigvee_{y \in R_S(x)} X(y), x \in U$$

$$\underline{R}(X)(x) = \bigwedge_{y \in R_S(x)} X(y), x \in U$$

序对 $(\underline{R}(X), \bar{R}(X))$ 称为粗糙模糊集, $\underline{R}$ 与 $\bar{R} : F(W) \rightarrow F(U)$ 分别称为粗糙模糊下近似算子和粗糙模糊上近似算子。

定义 3 设 R 是从 U 到 W 上的一个模糊二元关系,称三元组 (U, W, R) 为模糊近似空间, $\forall X \in F(W)$, X 关于 $(U, W,$

^{*}) 本文获得国家自然科学基金(No. 60673096)资助。

R 的上近似 $\bar{R}(X)$ 与下近似 $\underline{R}(X)$ 是 U 上的一对模糊子集,其隶属函数定义如下: $\forall x \in U$

$$\bar{R}(X)(x) = \bigvee_{y \in W} [R(x, y) \wedge X(y)]$$

$$\underline{R}(X)(x) = \bigwedge_{y \in W} [(1 - R(x, y)) \vee X(y)]$$

序对 $(\underline{R}(X), \bar{R}(X))$ 称为 X 关于 (U, W, R) 的模糊粗糙集, $\underline{R}$ 与 $\bar{R}: F(W) \rightarrow F(U)$ 分别称为模糊粗糙下近似算子和模糊粗糙上近似算子。

3 Pawlak 信息系统中的信息粒度与粗糙近似

一个Pawlak信息系统 S 是一个四元组 (U, AT, V, F) ,其中 U 是非空有限对象集称为论域, AT 是非空有限属性集, $V = \prod_{a \in AT} V_a$, V_a 是属性 a 的值域, $F = \{f_a: a \in AT\}$, $\forall a \in AT$, $f_a: U \rightarrow V_a$ 是信息函数,使得 $f_a(x) = a(x) \in V_a, x \in U$ 。

每一个属性子集 $A \subseteq AT$ 可以确定一个二元关系 $IND(A)$ 如下

$$IND(A) = \{(x, y) : \forall a \in A, f_a(x) = f_a(y)\}$$

易见, $IND(A)$ 是 U 上的一个等价关系,称为不可分辨关系,它将 U 划分成为等价类 $U/IND(A)$ 。记 $[x]_A$ 为对象 x 关于属性集 A 的等价类,即

$$[x]_A = \{y \in U : (x, y) \in IND(A)\}$$

由二元关系 $IND(A)$ 的定义可知,两个对象是不可分辨的当且仅当对于每个属性 $a \in A$ 取相同的值,故 $[x]_A$ 中的对象在系统中的描述是不可分辨的。

对于任意子集 $X \subseteq U$, X 可能不能由 $IND(A)$ 的等价类来精确描述,但可以由 X 关于 $IND(A)$ 的下近似和上近似来描述:

$$\underline{A}(X) = \{x \in U : [x]_A \subseteq X\}$$

$$= \bigcup \{[x]_A : [x]_A \subseteq X\}$$

$$\bar{A}(X) = \{x \in U : [x]_A \cap X \neq \emptyset\}$$

$$= \bigcup \{[x]_A : [x]_A \cap X \neq \emptyset\}$$

下近似 $\underline{A}(X)$ 是由等价类肯定属于 X 的那些对象全体,上近似 $\bar{A}(X)$ 是由等价类可能属于 X 的那些对象全体,序对 $(\underline{A}(X), \bar{A}(X))$ 称为 X 关于近似空间 $(U, IND(A))$ 的Pawlak粗糙集。

对于一个完备信息系统 (U, AT, V, F) , $B \subseteq AT$,记 $V_B = \prod_{a \in B} V_a$,并令

$$f_B(x) = \prod_{a \in B} f_a(x), x \in U$$

则 (U, V_B, f_B) 构成了一个近似空间,称为对象属性近似空间。 $\forall M \subseteq V_B$,则可以定义 M 关于近似空间 (U, V_B, f_B) 的下近似和上近似:

$$\underline{f}_B(M) = \{y \in U : \{f_B(y)\} \subseteq M\}$$

$$\bar{f}_B(M) = \{y \in U : \{f_B(y)\} \cap M \neq \emptyset\}$$

则 $\underline{f}_B$ 和 $\bar{f}_B$ 都是从 $P(V_B)$ 到 $P(U)$ 的映射。易见,

$$\underline{f}_B(M) = \{y \in U : \{f_a(y)\} \subseteq M|_{V_a}, \forall a \in B\}$$

$$\bar{f}_B(M) = \{y \in U : \{f_a(y)\} \cap M|_{V_a} \neq \emptyset, \forall a \in B\}$$

其中 $M|_{V_a}$ 是 M 在 V_a 上的投影(或限制)。由于 (U, AT, V, F) 是完备信息系统,因此, $\forall y \in U, \{f_B(y)\}$ 都是单点(向量)集,故 $\underline{f}_B(M) = \bar{f}_B(M)$ 。特别地,当 $M = \{f_B(x)\}$ 时, $\{f_B(x)\}$ 也是单点集,因此,

$$\{f_B(y)\} \subseteq \{f_B(x)\} \Leftrightarrow f_B(y) = f_B(x)$$

故

$$\underline{f}_B(\{f_B(x)\}) = \bar{f}_B(\{f_B(x)\}) = \{y \in U : f_a(y) = f_a(x), \forall a \in B\} = [x]_B$$

特别地

$$\underline{f}_a(\{f_a(x)\}) = \bar{f}_a(\{f_a(x)\})$$

$$= \{y \in U : f_a(y) = f_a(x)\} = [x]_a$$

因此

$$\begin{aligned} \underline{f}_B(\{f_B(x)\}) &= \bigcap_{a \in B} \underline{f}_a(\{f_a(x)\}) = \bar{f}_B(\{f_B(x)\}) \\ &= \bigcap_{a \in B} \bar{f}_a(\{f_a(x)\}) = \bigcap_{a \in B} [x]_a = [x]_B \end{aligned}$$

上式表明信息系统中每一个对象的不可分辨类(等价类)可以通过计算该对象在相应属性近似空间的上近似或下近似得到,即通过计算每一个对象-属性值的上近似或下近似将对象-属性信息变换成为对象空间上的基本信息粒度(等价类)。

4 不完备信息系统中的信息粒度与粗糙近似

当信息系统中某些对象的某些属性取值不知道,即缺省的,或部分知道(如肯定不取某某值),则这样的信息系统称为不完备信息系统^[3,4],我们仍记为 $S = (U, AT, V, F)$,其中 $F = \{F_a: a \in AT\}$ 为信息函数全体。在这种情形下,系统的每个信息函数 F_a 可以看成是从 U 到 $P(V_a)$ 的集值映射,当 $F_a(x)$ 是缺省值时,我们用属性的所有可能取值的集合(即值域)来代替 $F_a(x)$,即 $F_a(x) = V_a$,而当 $F_a(x)$ 是部分知道时,例如 $F_a(x)$ 不取 $m, n \in V_a$,则我们用 $V_a \setminus \{m, n\}$ 来代替 $F_a(x)$ 。

对于不完备信息系统中每一个属性子集 $A \subseteq AT$,我们定义相似关系 $SIM(A)$ 如下:

$$SIM(A) = \{(x, y) \in U \times U : F_a(x) \cap F_a(y) \neq \emptyset, \forall a \in A\}$$

相似关系是自反和对称的,但一般不满足传递性,因此是相容关系。令 $S_A(x)$ 表示不完备信息系统 S 中关于属性集 A 与对象 x 相似的那些对象全体,即

$$S_A(x) = \{y \in U : (x, y) \in SIM(A)\}$$

称 $S_A(x)$ 为系统 S 中对象 x 关于属性集 A 的相似类,记 U/A 为系统 S 中关于属性集 A 的所有相似类全体,即

$$U/A = \{S_A(x) : x \in U\}$$

则 U/A 为不完备信息系统 S 中对象关于属性集 A 的基本信息粒度。

对于一个不完备信息系统 (U, AT, V, F) ,设 $B \subseteq AT, M \subseteq V_B = \prod_{a \in B} V_a$,记 $F_B(x) = \prod_{a \in B} F_a(x), x \in U$,定义

$$\underline{F}_B(M) = \{y \in U : F_B(y) \subseteq M\}$$

$$\bar{F}_B(M) = \{y \in U : F_B(y) \cap M \neq \emptyset\}$$

则 $\underline{F}_B$ 和 $\bar{F}_B$ 都是从 $P(V_B)$ 到 $P(U)$ 的映射。易见,

$$\underline{F}_B(M) = \{y \in U : F_a(y) \subseteq M|_{V_a}, \forall a \in B\}$$

$$\bar{F}_B(M) = \{y \in U : F_a(y) \cap M|_{V_a} \neq \emptyset, \forall a \in B\}$$

特别地,当 $M = F_B(x)$ 时,

$$\bar{F}_B(F_B(x)) = \{y \in U : F_B(y) \cap F_B(x) \neq \emptyset\}$$

$$= \{y \in U : F_a(y) \cap F_a(x) \neq \emptyset, \forall a \in B\} = S_B(x)$$

上式表明不完备信息系统中每一个对象的相似类可以通过计算该对象在相应属性近似空间的上近似得到,即通过计算每一个对象属性值的上近似将对象-属性信息变换成为对象空间上的基本信息粒度(相似类)。

当 $B = \{a\}$ 为单个属性时,则

$$\bar{F}_a(F_a(x)) = S_a(x), x \in U$$

容易验证,

$$\bar{F}_B(F_B(x)) = \bigcap_{a \in B} \bar{F}_a(F_a(x)) = \bigcap_{a \in B} S_a(x), x \in U$$

5 不完备模糊信息系统中的信息粒度与粗糙近似

在一个不完备信息系统中,每一个对象关于属性 $a \in AT$ 的取值是 V_a 的一个子集,如果对于每一个对象 $x \in U$ 和每一个属性 $a \in AT$, x 关于属性 a 的取值 $F_a(x)$ 是 V_a 的一个正则

模糊集,则我们称这样的信息系统为不完备模糊信息系统,仍记为 $S=(U,AT,V,F)$,此时, V_a 上的模糊子集 $F_a(x) \in F(V_a)$ 可以看成是 V_a 上的一个可能性分布,值 $F_a(x)(v), v \in V_a$, 表示对象 x 在属性 a 上取值为 v 的可能性程度。这里我们实际上作了这样一个假设: $\forall x \in U$ 和 $\forall a \in AT, V_a$ 中至少有一个值的取值是完备可能的,即存在 $v \in V_a$ 使得 $F_a(x)(v)=1$ 。

不完备模糊信息系统可以看成是完备信息系统和不完备信息系统的推广形式。一个完备信息系统是一个不完备模糊信息系统的极端情形,即每一个对象关于每一个属性的信息都是完备的,也就是说对于某个 $v \in V_a$ 有 $a(x)(v)=1$,而对于其它的 $s \neq v$ 有 $a(x)(s)=0$;缺省值不完备信息系统是不完备模糊信息系统的另一种极端情形,即对于某些缺省值我们认为对它的信息认识是完全不知道的,也就是说,对于任意 $v \in V_a$ 有 $a(x)(v)=1$ 。

例1 表1给出了一个不完备模糊信息系统 $S=(U, AT, V, F)$,其中

$$U=\{x_1, x_2, \dots, x_8\}, AT=\{a_1, a_2, a_3\},$$

$$V_{a_1}=\{H, L, N\}, V_{a_2}=\{r, s, t\}, V_{a_3}=\{m, n, p\}$$

表1 一个不完备模糊信息系统

U	a_1	a_2	a_3
x_1	1/H	1/s	1/m+0.1/p
x_2	1/H+0.3/N	1/s+0.4/t	1/m+0.5/n
x_3	1/H	1/r+1/s+1/t	1/m+1/n+1/p
x_4	1/N+1/L	1/t	1/n
x_5	1/N	1/r+0.6/s	0.7/m+1/p
x_6	0.1/H+0.4/N+1/L	1/r+1/s	1/m+1/p
x_7	1/H+1/N+1/L	1/t	1/n+0.2/p
x_8	0.8/N+1/L	0.7/s+1/t	0.8/m+1/n

属性值 $F_{a_1}(x_7), F_{a_2}(x_3), F_{a_3}(x_3)$ 均为缺省值, $F_{a_1}(x_4)$ 不取“H”值, $F_{a_2}(x_6)$ 不取“t”值, $F_{a_3}(x_6)$ 不取“n”值*。值 $F_{a_1}(x_1), F_{a_1}(x_3), F_{a_1}(x_5), F_{a_2}(x_1), F_{a_2}(x_4), F_{a_2}(x_7), F_{a_3}(x_4)$ 都是完备的。

设 $S=(U, AT, V, F)$ 是一个不完备模糊信息系统,对于 $a \in AT, F_a: U \rightarrow F(V_a)$ 是一个模糊集值函数,这样的模糊集值函数可以导出从 U 到 V_a 的一个模糊二元关系 R_a : $\forall (x, y) \in U \times V_a$,

$$R_a(x, y) = F_a(x)(y)$$

因此,由模糊集值函数 F_a 可以导出一个模糊近似空间 $(U, V_a, F_a), \forall M \in F(V_a)$,由定义3可得 M 关于近似空间 (U, V_a, F_a) 的上近似 $\overline{F_a}(M): \forall u \in U$,

$$\overline{F_a}(M)(u) = \bigvee_{v \in V_a} [F_a(u)(v) \wedge M(v)]$$

则 $\overline{F_a}$ 是从 $F(V_a)$ 到 $F(U)$ 的模糊算子。特别地,对于 $x, y \in U$,当 $M=F_a(x)$ 时,记

$$S_a(x, y) = \overline{F_a}(F_a(x))(y) = \bigvee_{v \in V_a} [F_a(y)(v) \wedge F_a(x)(v)]$$

则 S_a 是 U 上的一个模糊二元关系,由于对于任意 $x \in U, F_a(x)$ 是正则的,所以

$$S_a(x, x) = \bigvee_{v \in V_a} [F_a(x)(v) \wedge F_a(x)(v)] = 1$$

故 S_a 是模糊自反关系,显然 S_a 是对称的,因此, S_a 是模糊相似关系。记

$$S_A(x) = \overline{F_a}(F_a(x)), x \in U$$

则 $S_A(x) \in F(U)$ 是对象 x 在不完备模糊信息系统中关于属性 a 的模糊相似类。于是,我们通过计算每一个对象属性值的模糊上近似将对象-属性信息变换成为对象空间上的模糊信息(模糊相似类)。记

$$S_A(x) = \bigcap_{a \in A} S_a(x), x \in U, A \subseteq AT$$

则 $S_A(x) \in F(U)$ 是对象 x 在不完备模糊信息系统中关于属性集 A 的模糊相似类,它构成了对象 x 在系统中关于属性 A 的基本信息粒度。

例2 在例1中,表1给出了对象属性间的模糊关系,可以验证:

$$S_{AT}(x_1) = \bigcap_{a \in AT} S_a(x_1) = \bigcap_{a \in AT} \overline{F_a}(F_a(x_1)) = 1/x_1 + 1/x_2 + 1/x_3 + 0/x_4 + 0/x_5 + 0.1/x_6 + 0/x_7 + 0/x_8$$

同理,我们可得到其它对象关于属性集 AT 的信息粒度(模糊相似类)。

结束语 本文说明了粗糙集方法为信息系统的知识表示提供了一个合适的工具。通过信息系统中对象属性值关于对象属性近似空间的上近似可以得到与该对象具有相同或相似信息的对象集,即利用对象属性值关于对象属性近似空间的上近似将对象属性值信息变换成为对象的基本信息粒度。

参考文献

- Pawlak Z. Rough sets. International Journal of Computer and Information Sciences, 1982, 11: 341~356
- Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Boston: Kluwer Academic Publishers, 1991
- Kryszkiewicz M. Rough set approach to incomplete information systems. Information Sciences, 1998, 112: 39~49
- Kryszkiewicz M. Rules in incomplete information systems. Information Sciences, 1999, 113: 271~292
- Wu Wei-Zhi, Zhang Wen-Xiu, Li Huai-Zu. Knowledge acquisition in incomplete fuzzy information systems via rough set approach. Expert Systems, 2003, 20(5): 280~286
- 张文修, 梁怡, 吴伟志. 信息系统与知识发现. 北京: 科学出版社, 2003
- Keil F C. Concepts, Kinds, and Cognitive Development. MIT Press, 1989
- Sowa J F. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole, 2000
- Yao Y Y, Zeng Y, Zhong N. Supporting literature exploration with granular knowledge structure. LNAI 4482, Springer, 2007. 182~189
- Yao Y Y. Three perspectives of granular computing. 南昌工程学院学报, 2006, 25(2): 16~21
- Langseth H, Aamodt A, Winnem O M. Learning retrieval knowledge from data. In: Proc. of the IJCAI99, 1999
- Tenenbaum J. Knowledge Representation: Spaces, Trees, Features. Course Lecture notes on Computational Cognitive Science, MIT Open Courseware, Fall 2004
- Collins A M, Quillian M R. Retrieval time from semantic Memory. Journal of Verbal Learning and Verbal Behavior, 1969. 240~248
- Yao Y Y. A partition model of granular computing. LNCS Transactions on Rough Sets, 2004, 1: 232~253
- 马建敏, 张文修, 姚一豫. 粗糙集与粒度计算. 粗糙集与概念格, 西安交通大学出版社, 2006

(上接第196页)

- Ricardo B Y, Berthier R N. Modern Information Retrieval. Addison Wesley, 1999
- Fensel D, van Harmelen F. Unifying reasoning and search to web scale. IEEE Internet Computing, 2007, 11(2): 94~96
- Berners L T, Hall W, Hendler J A, et al. A Framework for Web Science. Foundations and Trends in Web Science, 2006, 1(1)
- 周宁, 张玉峰, 张李义. 信息可视化与知识检索. 科学出版社, 2005
- Benjamins V R, Fensel D. Community is knowledge! in (KA)2. In: Proc. of KAW98, Banff, Canada, 1998
- Martin P, Eklund P W. Knowledge retrieval and the World Wide Web. IEEE Intelligent Systems, 2000, 15(3): 18~25
- Travers M. A visual representation for knowledge structures. In: Proc. of the 2nd annual ACM conference on Hypertext and Hypermedia, United States, 1989. 147~158
- Quillian M R. Semantic Memory. In: Semantic information processing. MIT Press, 1968. 216~270
- Martin P, Alpay L. Conceptual structures and structured documents. In: Proc. of Int Conf on Conceptual Structures, 1996. 145~159