

基于粗糙集的不确定知识表示方法

Uncertainty Knowledge Representation Method Based on Rough Sets

胡涛 吕炳朝 陈光福

(电子科技大学自动化系 成都610054)

Abstract A property set model for knowledge representation is proposed by discussing knowledge structure and representation mechanism, and defining property set and atomic concept based on rough set theory. Uncertainty knowledge can be represented by upper and lower approximation, and certainty knowledge can be also represented exactly by using the model. The model presents a new idea to uncertainty knowledge representation.

Keywords Rough set, Knowledge representation, Atomic concept, Property set model

1 引言

知识表示是人工智能(AI)研究中最关键的分支之一。传统的知识表示模型(如 AQ11^[1], ID3^[2]等),对知识描述是确定的、清晰的,即:被描述的对象具有或不具有某种属性是明确的。然而,在现实世界中,人们常常面对的是在领域信息不完整、不确定、不精确的前提下,完成对事物的认识、分析、推理、判断、预测和决策。这种智能行为往往要求人们对未知的信息进行估计、推测;对不完整数据进行分析、处理;对已知的证据进行分辨、扬弃。为此,人们提出了许多有效的方法来模拟人类的这种智能行为,如 Dempster-Shafer 证据理论、模糊集理论、主观 Bayesian 方法等。这些方法无疑对不确定性知识处理的研究起到了促进作用。然而,该领域的研究工作尚处于发展阶段,还有许多理论和实际问题亟待解决,特别是随着计算机技术的高速发展,知识的含义已被赋予了許多新的意义,因而如何对不完整数据进行分析、推理,发现数据间的关系,提取有用特征,简化信息处理,研究不精确、不确定知识的表示、学习、归纳等方法已成为人工智能工作者面临的重要研究课题。

粗糙集理论^[3,4](Rough set theory)是由 Z. Pawlak 于是1982年提出的一种刻画不完整性和不确定性的数学理论,它从新的角度对知识进行了定义,把知识看作是關於论域的划分,从而认为知识是有粒度(granularity)的,知识的不精确性是由于组成论域知识的颗粒太大引起的。粗糙集理论的主要特点之一是它仅利用数据本身所提供的信息,不需要任何附加信息或先验知识,比如 Dempster-Shafer 证据理论中的

基本概率赋值,模糊集理论中的隶属度,统计学中的概率分布等,而这些信息有时并不容易得到。因此,粗糙集理论相对于许多其它处理不确定知识方法,更具有客观性,它是一种能有效地分析和处理不确定知识的新的数学工具。

本文基于粗糙集理论的基本概念和理论基础,定义了用于知识表示的特征集和原子概念,并在此基础上提出了一种基于粗糙集理论的知识表示特征集模型,最后,用这种特征集模型对确定/不确定知识的表示进行了讨论。

2 粗糙集理论的基本概念和理论基础

2.1 知识、划分与等价关系

“知识”这个概念在不同的范畴内有多种不同的含义。在粗糙集理论中,知识被认为是一种对对象进行分类的能力。

定义1 设 $U \neq \emptyset$ 是感兴趣的对象组成的有限集合,称为论域。任何子集 $X \subseteq U$,称为 U 中的一个概念或范畴。空集也认为是一个概念。则 U 中的一族概念,就称为关于 U 的知识。

定义2 设 $U \neq \emptyset$ 是论域, $C = \{X_1, X_2, \dots, X_n\}$,使得 $X_i \subseteq U, X_i \neq \emptyset, X_i \cap X_j = \emptyset$, 对 $i \neq j, i, j = 1, 2, \dots, n$, 且 $\bigcup_{i=1}^n X_i = U$, 则称 C 为一个划分(partition)。 X_i 称为划分 C 的一个等价类。 U 上的一族划分,称为关于 U 的一个知识库。 U 上的一个划分与其上的一个等价关系是等价的。每一个等价关系描述的是论域 U 上的某一个属性,即属性亦可看作一个等价关系。

2.2 信息表、不可分辨关系和基本集

一个知识表示系统 S 可表示为 $S = \langle U, At, Val, f \rangle$, U 是论域, At 是属性集, $Val = \bigcup_{A \in At} Val_A$ 是属性值的集合, $f: U \times At \rightarrow Val$ 是一个信息函数, 它指定 U 中每一对象的属性值, 由这样的“属性-值”对就构成了一张表, 称之为信息表 (information table)。有时, 针对某类问题, $At = C \cup D$, C 和 D 分别称为条件属性和结果属性, 这样就构成了一种特殊的信息表——决策表。

不可分辨关系 (indiscernibility relation) 是粗糙集理论中一个十分重要的概念, 它是一族等价关系集合的最细划分, 该关系中的每一个等价类不能由原等价关系族的任一等价类再细分, 称之为基本集合。不可分辨关系揭示了知识的颗粒状结构, 基本集就是组成论域知识的颗粒。

2.3 粗糙集的下、上近似及边界区

给定一个有限的非空集 U 称为论域, R 为 U 上的一族等价关系, R 将 U 划分为互不相交的基本等价类, 二元对 $K = (U, R)$ 构成一个近似空间。

定义3 设 $U \neq \emptyset$ 是论域, $X \subseteq U$, ω 为 U 中的一个对象, $[\omega]_R$ 表示所有与 ω 不可分辨的对象组成的集合, 即由 ω 决定的等价类。当集合 X 能表示成基本等价类组成的并集时, 则称集合 X 是可精确定义的; 否则, 集合 X 只能通过逼近的方式来刻画。

集合 X 关于 R 的下近似:

$$R_-(X) = \{\omega \in U; [\omega]_R \subseteq X\} \quad (1)$$

集合 X 关于 R 的上近似:

$$R^+(X) = \{\omega \in U; [\omega]_R \cap X \neq \emptyset\} \quad (2)$$

集合 X 的边界区:

$$Bn(X) = R^+(X) - R_-(X) \quad (3)$$

$R_-(X)$ 实际上是那些根据已有知识判断肯定属于 X 的对象所组成的最大集合, 亦称 X 的正域, 记作 $POS(X)$ 。由根据已有知识判断肯定不属于 X 的对象所组成的集合称为 X 的负域, 记作 $NEG(X)$ 。

$R^+(X)$ 是那些可能属于 X 的对象所组成的最小集合。显然, $R^+(X) + NEG(X) = U$ 。

$Bn(X)$ 为集合 X 的上、下近似之差。如 $Bn(X) = \emptyset$, 则称 X 关于 R 是清晰的 (crisp); 反之, 如 $Bn(X) \neq \emptyset$, 则称集合 X 是关于 R 的粗糙集。

3 知识表示特征集模型

知识表示系统的基本成分是研究对象的集合, 关于这些对象的知识是通过指定对象的属性及其属性值来描述的。为此, 我们把这种描述对象的“属性-值”对看成是对象所具有的某种特征。于是有以下的定义:

定义4 设 $U \neq \emptyset$ 是论域, $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ 是对象集, $\Omega \subseteq U$, Ω 中任何对象所具有的特征所组成的集合, 称为特征集 (property set): $T = \{\tau_1, \tau_2, \dots, \tau_n\}$ 。这

些特征可由如下的“属性-值”对表示: $\tau = (a, v)$, $a \in At$, $v \in Val$ 。由特征集构成的表, 称之为特征表。

对于任一对象 $\omega \in \Omega$, 如 ω 可由 τ 描述, 则这种关系可表示为: $(\omega, \tau) \in \Gamma$, Γ 表示特征关系。于是, 可以得到:

$$[\omega] = \{\tau \in T; (\omega, \tau) \in \Gamma\} \quad (4)$$

定义5 在论域 U 中, 已知一特征集 T , 对于每个 $\tau \in T$ 满足下式:

$$[\tau] = \{\omega \in U; (\omega, \tau) \in \Gamma\} \quad (5)$$

则称 $[\tau]$ 为 τ 特征集。

同理, 非 τ 特征集 $[\hat{\tau}]$ 可定义如下:

$$[\hat{\tau}] = \{\omega \in U; (\omega, \tau) \notin \Gamma\} = U - [\tau] \quad (6)$$

利用 $[\tau]$, $[\hat{\tau}]$, 可构造如下原子集:

$$\begin{aligned} [\alpha_0] &= [\hat{\tau}_1] \wedge [\hat{\tau}_2] \wedge \dots \wedge [\hat{\tau}_n] = [\hat{\tau}_1] \cap [\hat{\tau}_2] \cap \dots \cap [\hat{\tau}_n] \\ [\alpha_1] &= [\tau_1] \wedge [\hat{\tau}_2] \wedge \dots \wedge [\hat{\tau}_n] = [\tau_1] \cap [\hat{\tau}_2] \cap \dots \cap [\hat{\tau}_n] \\ [\alpha_2] &= [\hat{\tau}_1] \wedge [\tau_2] \wedge \dots \wedge [\hat{\tau}_n] = [\hat{\tau}_1] \cap [\tau_2] \cap \dots \cap [\hat{\tau}_n] \\ &\vdots \\ [\alpha_N] &= [\tau_1] \wedge [\tau_2] \wedge \dots \wedge [\tau_n] = [\tau_1] \cap [\tau_2] \cap \dots \cap [\tau_n] \end{aligned} \quad (7)$$

$$N = 2^n - 1$$

这些原子集形成了对论域 U 的一种划分。

为简便起见, 这里, 用 $\tau, \hat{\tau}, \alpha$ 来分别标记 $[\tau]$, $[\hat{\tau}]$, $[\alpha]$, 同时省略合取符 ($\wedge$)。于是得到:

$$\begin{aligned} \alpha_0 &= \hat{\tau}_1 \hat{\tau}_2 \dots \hat{\tau}_n \\ \alpha_1 &= \tau_1 \hat{\tau}_2 \dots \hat{\tau}_n \\ \alpha_2 &= \hat{\tau}_1 \tau_2 \dots \hat{\tau}_n \\ &\vdots \\ \alpha_N &= \tau_1 \tau_2 \dots \tau_n \\ N &= 2^n - 1 \end{aligned} \quad (8)$$

我们称 $\alpha_0, \alpha_1, \dots, \alpha_N$ 为原子概念, 所有原子概念之集称为概念空间 C 。事实上, 每个原子概念都是一种布尔表示的等价类, 在同一等价类中所有对象都是不可区分的, 这些等价类共同形成了对论域 U 的划分。原子概念是信息系统知识表示的最基本范畴, 或者说是构成知识的最小“颗粒”。很显然, “颗粒”越细, 知识表示越清晰; 反之, “颗粒”越粗, 知识表示越粗糙。

根据粗糙集理论对于知识的定义, 我们可以用原子集/原子概念来对知识 $A \subseteq U$ 进行描述和表达: 若 A 可由一族原子概念精确描述, 即 A 可表示成原子集的并集时, 则 A 是确定性知识, 是可精确定义的; 反之, 则 A 是不确定性知识, 只能粗糙定义。显然, 由于原子集对 U 的划分不可能是无限的, 知识表达的最小单元——“颗粒”肯定存在, 因此, 概念空间的表达能力是有限的, 在一定前提下, 可用原子集对知识进行精确表示, 得到确定性知识; 而更多的情况下, 我们不得不面对不确定性知识表达这种严峻而现实的问题。对于不能用原子集精确表示的不确定性知识, 我们采用逼近的方式来描述, 即使用两个精确集——上近似集和下

近似集来表示。知识 A 的下近似、上近似分别表示如下:

$$A_- = \bigvee_{\alpha_i \in A} \alpha_i \quad (9)$$

$$A^+ = \bigvee_{\alpha_i \in A^+} \alpha_i \quad (10)$$

由此,我们可得如下结论:

(1) 当 $A_- = A^+$, 则称 A 可用原子集精确定义, 即 A 是确定性知识;

(2) 当 $A_- \neq A^+$, 则称 A 不能用原子集精确定义, 即 A 是不确定性知识, 只能采用逼近的方式来描述, 即使用两个精确集——上近似集和下近似集来表示。对不确定性知识 A 的描述又可分为四种情况:

(a) 当 $A_- \neq \emptyset$, 且 $A^+ \neq U$, 则称 A 为粗糙可定义的;

(b) 当 $A_- = \emptyset$, 且 $A^+ \neq U$, 则称 A 为内不可定义的;

(c) 当 $A_- \neq \emptyset$, 且 $A^+ = U$, 则称 A 为外不可定义的;

(d) 当 $A_- = \emptyset$, 且 $A^+ = U$, 则称 A 为完全不可定义的;

4 例子与讨论

4.1 例子

一个知识表达系统具有基本特征集: $T = \{\tau_1, \tau_2, \tau_3\}$, 该特征集定义了原子概念:

$$\alpha_0 = \tau_1 \wedge \tau_2 \wedge \tau_3$$

$$\alpha_1 = \tau_1 \wedge \tau_2$$

$$\alpha_2 = \tau_1 \wedge \tau_3$$

$$\alpha_3 = \tau_2 \wedge \tau_3$$

$$\alpha_4 = \tau_1$$

$$\alpha_5 = \tau_2$$

$$\alpha_6 = \tau_3$$

$$\alpha_7 = \tau_1 \wedge \tau_2 \wedge \tau_3$$

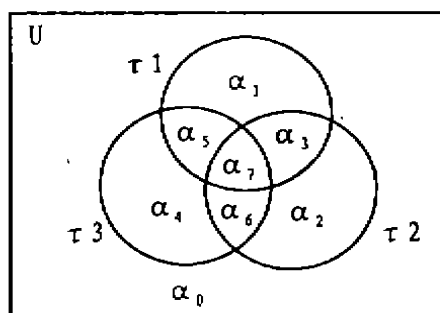


图1 原子概念 $\{\alpha_0, \alpha_1, \dots, \alpha_7\}$

形成了对论域 U 的划分

原子概念 $\{\alpha_0, \alpha_1, \dots, \alpha_7\}$ 形成了对论域 U 的划分。如图1所示。

若有知识 A, 知识 B (如图2, 图3所示阴影部分, 我们可用原子概念 $\{\alpha_0, \alpha_1, \dots, \alpha_7\}$ 对其进行描述和表示。首先, 由式(9)、(10)分别计算知识 A, 知识 B 的下、上

近似:

$$A_- = \alpha_3, \quad A^+ = \alpha_0 \vee \alpha_1 \vee \alpha_2 \vee \alpha_3 \vee \alpha_7$$

$$B_- = \alpha_5 \vee \alpha_7, \quad B^+ = \alpha_5 \vee \alpha_7$$

由于 $A_- \neq A^+$, 即不能用原子概念 $\{\alpha_0, \alpha_1, \dots, \alpha_7\}$ 精确地描述和表达知识 A, 所以 A 是不确定性知识; 又因 $A_- \neq \emptyset$, 且 $A^+ \neq U$, 则称不确定知识 A 为粗糙可定义的。

而 $B_- = B^+$, 即知识 B 可用原子概念 $\{\alpha_0, \alpha_1, \dots, \alpha_7\}$ 精确地描述和表达, 故 B 称为确定性知识, 它可以被精确定义和表示。

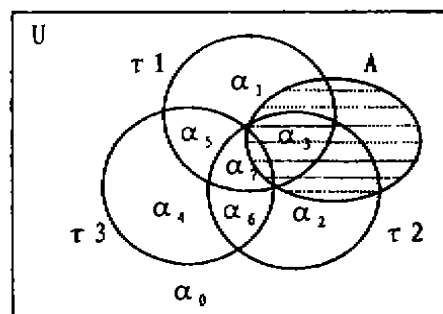


图2 不确定知识 A 的粗糙定义

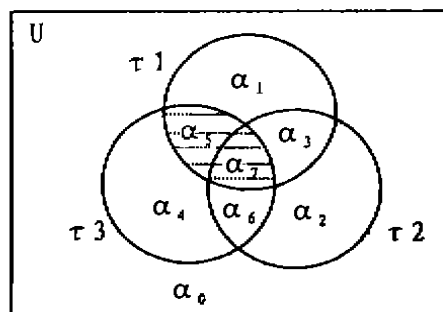


图3 确定知识 B 的精确定义

4.2 讨论

知识的精确表达是相对的, 是在一定范畴、前提条件下的“精确”。如图3, 若重新选取特征集 $T' = \{\tau_1, \tau_2, \tau_3\}$, 从而得到新的原子概念 $\{\alpha'_0, \alpha'_1, \dots, \alpha'_7\}$, 则 $B_- \neq B^+$, B 就变成了不确定知识, 即, 用新的原子概念已不能精确地描述和表达知识 B 了; 另一种情况, 当特征集大小发生变化, 即原子概念的规模发生变化时, 原确定知识可能变成了不确定知识, 而有些原不确定知识随着原子概念规模的不断增大有可能逐渐变成了确定知识。

总之, 从辩证的角度来讲, 知识的精确、确定表示

(下转封四)

(上接第 92 页)

是相对的,而知识表示的粗糙性、不确定性则是“绝对”的。那么,如何改进方法,提高知识表示的准确性呢?可从以下两方面着手:

(1)增大特征集,扩大原子概念的规模,使知识表达的最小“颗粒”更加细化。从理论上讲,随着知识表达“颗粒”的无限细化,则知识表示将趋向“完全清晰、准确”。

(2)在原子概念规模不变的前提下,改变原子的内涵,即从一个新的视角来观察事物,提取特征,这往往有利于认清事物的本质,改善知识表示的精确性、准确性。

结论 本文基于粗糙集的基本理论,分析了知识的构成及表达机理,定义了用于知识表示的特征集和原子概念,并在此基础上提出了一种基于粗糙集理论的知识表示特征集模型。这特征集模型对确定性知识可进行精确描述和表达;而对于不确定性知识则可采用逼近的方式来描述,即使用两个精确集——上近似集和下近似集来表示。该模型揭示了知识的颗粒状结构。本文最后对两种能有效改善知识表示精确性的方法(扩大原子概念规模和改变原子内涵)进行了讨论。知识表示是人工智能研究中最重要的问题之一,而

不确定知识表示则是知识表示中最为复杂的部分,如何利用“混合”方法来处理不确定知识,这是一个具有前景非常好的研究方向,也是我们下一步将要进行的工作。

参考文献

- 1 Michalski R S, Larson J B. Selection of the most representative training examples and incremental generation of VLI hypothesis: the underlying methodology and the description of programs ESEL and AQ11. [Rept. No. 867]. Dept. of Computer Science, University of Illinois, 1978
- 2 Quinlan J R. Induction of decision tree. Machine Learning, 1986, 1: 81~106
- 3 Pawlak Z. Rough set. Intl. J. of Information and Computer Science, 1982(11): 341~365
- 4 Pawlak Z. Rough set-theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers, 1991
- 5 Davis R, et al. What is a knowledge representation? AI Magazine, 1993, 14(1)
- 6 Parsons S. Uncertainty in artificial intelligence. The Knowledge Engineering Review, 1994, 9(1)

计算机科学

(1974年1月创刊)

第27卷第3期(月刊)

2000年3月25日出版

中国标准刊号: ISSN 1002-137X
CN 51-1239/TP

定价: 9.00元 国外定价: 5美元

邮发代号: 78-68

发行范围: 国内外公开

主管单位: 国家科学技术部

主办单位: 国家科技部西南信息中心

编辑出版: 《计算机科学》杂志社

重庆市渝中区胜利路132号 邮政编码: 400013

电话: (023) 63500828 传真: (023) 63502473

主 编: 朱宗元

印刷者: 国家科技部西南信息中心印刷厂

总发行处: 重 庆 市 邮 政 局

订购处: 全 国 各 地 邮 政 局

国外总发行: 中国国际图书贸易总公司(北京 399 信箱)

国外代号: 6210M