

97, 24(5)

自然语言

不确定性

粗糙理论

粗糙集理论

计算机科学 1997 Vol. 24 No. 5

人类自然语言不确定性及粗糙集理论的表示

Uncertainty of Human Natural Language and Representation Based on Rough Set Theory

魏长华

(华中师范大学计算机科学系 武汉 430070)(中科院自动化所模式识别国家重点实验室)

摘 要 In this paper the uncertainty of human natural language is researched and the six features of uncertainty are presented. It is given that representation of uncertainty knowledge based upon rough set theory and the rough set theory in comparison with D-S evidence theory.

关键词 Human natural language. Uncertainty knowledge representation, Rough set theory, D-S evidence theory

在现实世界中,人们对事物的推理、判断、预测和决策等智力行为大多数是在问题领域的信息不完全、不确定、不精确或者模糊的条件下进行的。为了完成这些智力行为,人们往往要对未知的信息进行猜测、估计,对采集的数据进行加工和处理,对获得的证据进行分析、比较、分辨、保留或废弃。为此,AI工作者陆续提出了许多行之有效的方法来模拟人类的这种智力行为,比如,信念网、主观贝叶斯、模糊集合论、D-S 证据理论、认证论等方法。在此基础上,近年来又提出了非概率可能性证据、基于 Dempster 规则假设的估价网络中的条件独立、概率网络法、模拟不确定性法^[1]。这些方法无疑对不确定性知识处理的研究起到促进作用。然而,该领域的研究工作正处于发展阶段,还有许多理论问题和实际问题亟待解决,尤其是针对不同的应用领域,如何采用更有有效的办法进行解决,是 AI 工作者面临研究的课题。

本文首先讨论了人类自然语言不确定性的几个特征。在此基础上,基于集合论论述了粗糙集理论的数学基础,同时给出了基于粗糙集理论的不确定知识的表示方法,并对粗糙集理论与 D-S 证据理论进行了比

较分析。

1 人类自然语言的不确定性

人类使用自然语言的目的是为了通信。人类自然语言的不确定性反映了我们社会的复杂性和丰富性。认知心理学和行为决策研究者关心人类自然语言的不确定性,关心人们如何使用和理解语言的不确定性以表示和推理,而知识工程师则更关心智能系统如何识别不确定性多界面(multiface)的性质和研究人类自然语言表示的形式。事实上,我们并不要求智能系统具有信息完全精确的特性,而且系统中信息的不确定性与人们的偏爱也是不相容的。因此,我们要研究在系统中如何用一种方法来度量人们所作出的决策、判断和与外界通信。

我们考察人类使用自然语言(口头语言和书面语言)和图形进行的社会活动,不难发现人类自然语言表示不确定性往往用数量、频率、概率和程度等几种方式来描述事件从模糊到精确的变化,以期表示对某一命题的信念程度。例如,人们经常使用的确定性语言词汇有:

程 度 型		数 量 型	频 率 型	概 率 型
considerably	exactly	all/every/each	always	certain(ly)
a great deal	precisely	most	not numerable	definite(ly)
(very)much	just	a majority	usual(ly)	undoubtedly
(quite) a lot	virtually	many/much	normal(ly)	probably/probale
rather	practically	a lot	general(ly)	likely
somewhat	more or less	enough	regular(ly)	perhaps
a bit	almost	some	often	possibly/possible
a little	nearly	a number	frequent(ly)	maybe
slightly	approximately	several	sometimes	unlikely
scarcely	about	a minority	occasional(ly)	improbable
hardly		a few/a little	rate(ly)	doubtful
only just		few/little	seldom	almost impossible
A 组	B 组	C 组	D 组	E 组

我们可以将这些词汇按语义范畴分类成数量型不确定(C组)、频率型不确定(D组)、概率型不确定(E组)和程度型不确定(A,B组)。比如,在C组词汇中,事物的发生可以从no/none/not/any所表示‘没有’到由all/every/each所表示‘全部、每一个’,从数量上描述从模糊不确定到精确不确定的变化。在D组词汇中,never表示事件从来不发生,而always的语义是总会发生,在这两者之间的词汇表示了事件发生的频率由小到大变化的过程。D组中的词汇描述了事件发生的概率,从almost impossible表示事件发生的概率几乎不可能到certain(ly)表示事件肯定要发生的概率,其中间过渡词汇doubtful,improbable,unlikely,maybe...definite(ly),在概率上描述了事件发生的可能性从模糊不确定性到精确不确定性的变化。与C、D、E组不同,A和B组中的词汇刻画了程度型的不确定性。

根据不确定世界的现实和上述对不确定性自然语言的讨论,我们不难发现人类使用自然语言表示事物的不确定性具有下面的几个特点:

(1)不确定性语言是可以度量的。人们在利用自然语言进行通信时,无论对于信息的发送者还是接收者,信息不确定性的传递都是以某种定量的形式来表示其测度,即使这种量是模糊量,或者是概率,或者是其他的量,其不确定性都具有一定的不确定度。

(2)事物的不确定性具有一定的不确定性区间。人们通常总是尽可能地利用某种方式来表达事物不确定性的上限和下限,构成一个不确定性区间,而且一旦条件允许就力求缩小这个区间,如果问题需要的话,这个区间有时可能会缩小成一个局部点。

(3)人们使用自然语言表示事物的不确定性和进行不确定性推理时,虽然信息本身具有随机性,但是人们的表示和推理并不是随机的,更不是随意的,而是遵循一定的认知规律和行为法则。

(4)在处理不确定性问题时,人们往往力求掌握更多处理问题所需的信息。这是因为当信息来自多个方面时,人们可以将问题的不确定性在多维空间上进行比较、权衡、调整、叠加和综合,从而形成一个综合性的信息。这个综合性信息的不确定性权值是来自于各个方面的信息的权值成正比的,也就是说掌握的信息越多,不确定性区间就越小,也就越有利于对问题的处理。

(5)一条信息的不确定性度量与其他的不确定性信息无关。这就是说信息传递的范围、传递的途

径、传递的形式和信息的权值均是各自独立的,并不因其他信息的变化而改变,它仅取决于自身的上下文和通信的方向。

(6)自然语言的不确定性与语境密切相关。因为人类自然语言的通信都是在一定的语言环境、文化环境和现实环境中进行的,所以信息的发送者和接收者对于语言的不确定性语义的理解都要受到这些环境的制约,一旦这些环境发生了变化其语义也就可能随之变化。

2 粗集理论的基本概念

粗集理论(rough set theory,简称为RST)是由Pawlak(1982)首先提出来的^[3],后来经包括他本人在内的许多学者进一步的研究,使得该理论得到了发展。RST特别适合于不要求精确数值结果的不确定性问题。这种方法具有较强的数学基础、方法简单、较强的鲁棒性和计算量小等优点。粗集理论可以叙述为:

定义1 令 $a_i(i=1,2,\dots,n)$ 是对象的一个基本概念或属性,则 a_i 的有限集合记为 $A=\{a_1,a_2,\dots,a_n\}$,称为一个概念集。

定义2 由概念集A中的元素所描述的被操作的对象 $e_j(j=1,2,\dots,m)$ 构成的集合,记为 $E=\{e_1,e_2,\dots,e_m\}$,称为一个对象集。

例如,描述一手非点数扑克牌的概念集,为: $A=\{\text{Spade, Head, Diamond, Club, J, Q, K, Joker}\}$;而 $E=\{\text{Spade J, Spade K, Head Q, Diamond J, Diamond Q, Diamond K, Club Q, Joker}\}$ 可以构成一个对象集。

如果 $e_j(j=1,2,\dots,m)$ 的描述仅基于某个 $a_k(k=1,2,\dots,n)$,那么会产生某些 e_j 是不可辨别的,因为辨别它们的属性不在概念集A中。比如,我们不能辨别Spade 3与Spade 5这两张牌,因为在概念集A中没有点数牌的描述。

定理1 当有限集A有定义,就存在有一个等价关系 $\approx$,比如 $e_j \approx e_k$ 。这时, e_j 和 e_k 对于所给定的A彼此是不可辨别的,对于任意一个 i, a_i 是 e_i 的一个属性,当且仅当它是 e_k 的一个属性。

由定理1可知,存在有一个关于E的分类:

$$P=\{p_1,p_2,\dots,p_r\}$$

其中 $\cup p_i=E$ 且 $p_i \cap p_j=\emptyset (i \neq j=1,2,\dots,r)$,称 p_i 是一个等价类(其中符号 $\cup$ 和 $\cap$ 分别为并和交运算)。上例中,有关E的分类是:

$$P=\{\{\text{Spade J, Spade K}\}, \{\text{Head Q}\}, \{\text{Diamond}$$

$\{J, \text{Diamond } Q, \text{Diamond } K\}, \{\text{Club } Q, \text{Club } K\}, \{\text{Joker}\}$

定义 3 令 $t \in T = \{t_1, t_2, \dots, t_p\}$ 是一对象, t 的属性 t_i 有关系 $t_i \subseteq A$, 则 t_i 必由属性的分类集 P 的项 e_i 来描述。于是有:

$$t^c(P, E) = \{e : e \in P, P_i \subset t\}$$

$$t^p(P, E) = \{e : e \in P, P_i \cap t \neq \emptyset\}$$

则称 $t^c(P, E)$ 是 E 和 P 上 t 的核, 而 $t^p(P, E)$ 是 E 和 P 上 t 的包络。

由定义 3 可知, 对于 t 的核, E 中所有等价对象集的所有属性 T 都具有, 而对于 t 的包络 E 中所有等价对象集至少有一个属性为 T 所具有。

在上例中, 若 $t = \{\text{Spade}, \text{Joker}\}$, 则

$$t^c(P, E) = \{\text{Joker}\}$$

$$t^p(P, E) = \{\text{Spade } J, \text{Spade } K, \text{Joker}\}$$

定义 4 由核 $t^c(P, E)$ 和包络 $t^p(P, E)$ 组成的对偶 $[t^c(P, E), t^p(P, E)]$ 称为一个粗集。

定义 5 如果一个等价对象集 E 中仅含有它的包络 $t^p(P, E)$, 而不含它的核 $t^c(P, E)$, 则称这样的集合为 T 的边界 (boundary), 记为 $t^b(P, E)$ 。即

$$t^b(P, E) = t^p(P, E) - t^c(P, E)$$

定义 6 E 中不包含 t 描述的所有等价对象的集合称为无关集, 记为 $t^i(P, E)$ 。即

$$t^i(P, E) = P - t^p(P, E)$$

令由被分类为 P 的 E 所定义的所有粗集集合为 R , 有对偶 $R = [R^c, R^p]$ 和 $R' = [R'^c, R'^p] \in R$ 。显然, 下列集合关系式是成立的:

$$(R \cup R')^c \supseteq R^c \cup R'^c \quad (R \cup R')^p = R^p \cup R'^p$$

$$(R \cap R')^c = R^c \cap R'^c \quad (R \cap R')^p \subseteq R^p \cap R'^p$$

$$(\neg R)^c = \sim(R^c) \quad (\sim R)^p = \sim(R^p)$$

$$(R \rightarrow R')^c \supseteq \sim R^c \cup R'^c \quad (R \rightarrow R')^p = \sim R^p \cup R'^p$$

3 基于粗集理论的不确定性知识表示

3.1 基于 RST 的不确定性知识表示

不确定性问题的一般处理方法是属于简化方法, 这些方法针对事件的具体情况给出一个主观的具有一定精确度的观察值。然而, 现实世界中的事件之所以是不确定的, 表现为它们不能够明显地确定是真还是假。事实上, 这些值并不是简单地依赖于问题的外部表象, 而是取决于事实之间的内部联系和事实与推理之间的关系。这就是不确定事件区别于确定事件的重要因素。在不确定性系统中, 因为事实中蕴涵的关系并不总是明显地表现出来, 有时被认为是一个重要的事实也不一定在与此相关的关系中

显露, 这就导致了事实之间关系的不确定性。

例如, 在医疗诊断系统中, 若症候 S_1 不是唯一地由疾病 D 引起的, 那么可能说, $S_1 \rightarrow D$ 的置信度为 CF 。这就是说, 疾病 D 的症候还可能有 $S_2, S_3, \dots$ 。在一般情况下, 我们得不到 D 到底是否由 S_1 引起的, 还是由 $S_2, S_3, \dots$ 引起的确切结论。然而, RST 可以很好地模拟这种现象。在 RST 中, 其基本依据是系统中事实所描述基本概念 (属性) 的集合。由前面讨论可知, 这里存在有映射, 比如粗集中的概念集与观察事实之间的映射, 概念集与系统辨别的假设集之间的映射等。

RST 应用于不确定知识处理的基本思想是将知识系统中的知识项看成对象项 E , 且对 E 经过一定的操作产生核 t^c 和包络 t^p , 从而形成该问题的一个粗集, 构成不确定性区间。

考虑知识系统中对象项的集合 $E = \{e_1, e_2, \dots, e_m\}$ 。例如, 有疾病诊断系统的症候作为对象集, 即 $E = \{\text{头痛}, \text{发烧}, \text{非柯氏斑}, \text{疹子}\}$, 存在有假设集 $T = \{t_1, t_2, \dots, t_p\}$, 比如 $T = \{\text{麻疹}, \text{结核}, \text{百日咳}\}$ 。显然, 这个假设集是通过观察, 参照症候对象集 E 而得到的。例如有诊断规则 $t_k \rightarrow e_i \wedge e_j$, 它的一个可能解释是“麻疹蕴涵发烧和非柯氏斑”, 即有 $t_1 \rightarrow e_2 \wedge e_3$ 。

再考虑概念集 $A = \{a_1, a_2, \dots, a_n\}$, 对于一个知识系统而言, 概念 $a_i (i=1, 2, \dots, n)$ 是领域知识中的基本属性, 来自于领域中的每条信息都存在有与 a_i 的确切关系。然而, 它们并不都一对一地映射到 e_i 和 t_k , 其中有些 a_i 是未知的, 有些已知的 a_i 与 e_j 相关联, 存在有 e_i 上的映射, 同时也存在有 t_k 上的映射。例如, 关于“麻疹”的症候规则也许是:

$$a_4 \wedge a_5 \rightarrow a_1 \wedge a_2 \wedge a_3 \quad (1)$$

这时, 由于观察到的事实在这些概念 $a_i (i=1, 2, 3, 4, 5)$ 上没有明显的映射关系, 医生仅根据所观察到的事实陈述病症, 所以这个陈述仅仅只有一种可能。这就是说 $t_k \rightarrow e_i \wedge e_j$, 其中 $e_i = e_2, e_j = a_3 \wedge a_5, t_k = a_4$, 那么有:

$$a_4 \rightarrow a_1 \wedge a_3 \wedge a_5 \quad (2)$$

显然, (1) 式与 (2) 不同。这说明应用观察事实描述系统, 一般来说仅仅有可能逼近所隐藏的概念。RST 用于不确定性表示的优点在于它可以跟踪这一逼近的过程。因为粗集中 e_i 和 t_k 对于 a_i 有关联, 如果 e_i 和 t_k 是由 a_i 得到的话, 那么就可以写出关于它们的逻辑表达式, 相应也就可以写出基于领域专家知识未知集 A 的 e_i 和 t_k 之间的逻辑关系。

A 和 E 上的对象 t 的大小可以用它的核 t^c 和包

格 t^* 的基来度量,即

- 若 $t^* = t^*$, 那么 A 由 (A, E) 精确定义
- 若 $t^* \neq t^*$ 且 $t^* \neq \emptyset$, 那么 A 由 (A, E) 的粗集定义
- 若 $t^* = \emptyset$, 那么 A 无 (A, E) 的下限定义, 即粗集无核
- 若 $t^* = E$, 那么 A 无 (A, E) 的上限定义, 即粗集无包络
- 若 $t^* = \emptyset$ 且 $t^* = E$, 那么 A 完全不能由 (A, E) 定义, 即既无核又无包络

3.2 RST 与 D-S 证据理论

从描述事物的不确定性的基本思想来看, 粗集理论与 D-S 证据理论十分相似。Skowron 和 G. Busse 指出粗集理论和 D-S 证据理论存在有一种封闭关系。在 D-S 证据理论中, 使用两个量作为不确定性的测度, 即由信任函数和似然函数构成一个置信区间。信任函数通过概率的形式描述了不确定事件已知对象的概率, 表示信任的程度, 它作为置信区间的下限; 而似然函数通过似然率描述了不确定事件未知对象的概率, 表示不否定的信任程度, 它作为置信区间的上限。

1) D-S 证据理论的基本概念。设 Ω 为变量 x 的所有可能值的穷举集合, 且设 Ω 中的各元素是互斥的, 我们称 Ω 为辨别框 (Frame of descerment)。设 Ω 中的元素个数为 N 则 Ω 的幂集合 2^Ω 元素个数为 2^N , 每个幂集合的元素对应于关于 x 取值的命题。

定义 7 对任一属于 Ω 的子集 A (命题), 令有一个相应数 $m \in [0, 1]$, 且满足:

$$m(\emptyset) = 0 \quad \sum_{A \subseteq \Omega} m(A) = 1$$

则称函数 m 为上 2^Ω 的基本概率赋值函数 bpa (Basic probability assignment)。于是有:

i) 若 $A \subseteq \Omega$ 且 $A \neq \emptyset$, 则 $m(A)$ 表示对 A 的精确信任程度。

ii) 若 $A = \Omega$, 则 $m(A)$ 表示这个数不知如何分配。

定义 8 若 $A \subseteq \Omega$, 且 $m(A) \neq \emptyset$, 称 A 为 m 的一个焦元 (Focal element)。

定义 9 命题的信任函数 Bel (Belief function): $2^\Omega \rightarrow [0, 1]$ 为:

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B) \quad A \subseteq \Omega$$

定义 10 命题的似然函数 Pl (plausibility function): $2^\Omega \rightarrow [0, 1]$ 为:

$$\text{Pl}(A) = 1 - \text{Bel}(A) = \sum_{B \cap A \neq \emptyset} m(B), A \subseteq \Omega,$$

其中 $A = \Omega - A$ 。

• 4 •

信任函数与似然函数有关系: $\text{Pl}(A) \geq \text{Bel}(A)$

$\text{Pl}(A)$ 和 $\text{Bel}(A)$ 分别构成了命题的上限函数和下限函数, 记为 $A[\text{Bel}(A), \text{Pl}(A)]$ 。

2) RST 与 D-S 证据理论的关系

i) 研究的对象。RST 的研究对象是由事物属性所描述的对象, 而 D-S 证据理论则往往是命题。

ii) 不确定性的测度。RST 的不确定性测度是由核和包络构成的区间, 而 D-S 证据理论是由信任函数和似然函数构成的区间。

iii) 不确定性的典型值。D-S 证据理论置信区间 $A[\text{Bel}(A), \text{Pl}(A)]$ 的典型值有:

$A[0, 1]$: 对 A 一无所知

$A[1, 1]$: A 为真

$A[0, 0]$: A 为假

$A[\text{Bel}(A), \text{Pl}(A)]$: 同时对 A 和 A 部分信任

$A[\text{Bel}(A), 1]$: 对 A 的部分信任度为 $\text{Bel}(A)$

$A[0, \text{Pl}(A)]$: 对 A 部分信任度为 $\text{Pl}(A)$

在 RST 中, 对于用命题逻辑表示的不确定性问题, 命题逻辑中命题的粗集值也可以用核 t^* 和包络 t^* 来度量。令 P 是命题变量, $P \in L(P)$ 是命题语言, 则语句 $P \in L(P)$ 的粗集测度 $R(P)$ 有:

若 $R(P) = [A, A]$, 那么 P 为真

若 $R(P) = \{Y, A\}$, 那么 P 为粗真

若 $R(P) = [\emptyset, A]$, 那么 P 是未知值

若 $R(P) = [\emptyset, X]$, 那么 P 为粗假

若 $R(P) = [\emptyset, \emptyset]$, 那么 P 为假

由上述的分析, RST 应用于不确定性表示的步骤是:

(1) 根据智能系统的具体情况, 构造描述领域知识的概念集 $A = \{a_1, a_2, \dots, a_n\}$, $a_i \in A$ 。

(2) 由系统的事实库形成对象集 $E = \{e_1, e_2, \dots, e_m\}$, $e_i \in A$ 。

(3) 基于分类集 P 和对象集 E , 由假设集 T 计算核 $t^*(P, E)$ 和包络 $t^*(P, E)$ 。

(4) 将核和包络组成对偶, 每个对偶 $[t^*(P, E), t^*(P, E)]$ 就是一个粗集。它给出了事件中概念 (属性) 不确定性的上限和下限所形成的估价区间。

参考文献

- [1] Simon Parsons, Uncertainty in artificial intelligence, The Knowledge Engineering Review, 9(1)1994
- [2] 马大猷, 语言信息和语言通信, 北京: 知识出版社, 1987
- [3] Thomas S. Wallstem 等, A review of human linguistic probability processing: General principles and empirical evidence, The Knowledge Engineering Review, 10(1)1995