第44卷第4期 南京航空航天大学学报 Vol. 44 No. 4
2012年8月 Journal of Nanjing University of Aeronautics & Astronautics Aug. 2012

基于全知熵的模式集成不确定性度量模型

胡文彬[1,2] 张 宏[1] 李千目[1]

(1.南京理工大学计算机科学与技术学院,南京,210094;
2.淮海工学院计算机工程学院,连云港,222005)

摘要:不确定性是模式集成的一个固有性质,不确定性度量对模式集成具有重大影响。本文提出一种度量模型,在该模型中模式对象及其属性清洗模块使该模型免受规模影响。根据模式集成多属性分阶段决策的特点,本文基于粗糙集理论的全知熵不确定率进行各阶段的不确定性度量,并把过程模型的不确定性度量引入到总体不确定性的度量中,最后给出了合成多不确定率的方法。实例分析证实所设计模型是可行、有效的。

关键词:不确定性度量;粗糙集理论;模式集成;全知熵;过程模型

中图分类号:TP18;TP31 **文献标识码**:A **文章编号**:1005-2615(2012)04-0575-05

Uncertainty Measure Model of Schema Integration Based on All Known Entropy

Hu Wenbin[1,2], *Zhang Hong*[1], *Li Qianmu*[1]

(1. College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, 210094, China;
2. School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, 222005, China)

Abstract: Uncertainty is intrinsic in schema integration. An uncertainty measure model of schema integration system (SIS) is presented. Schema object cleanout module and its attribute cleanout module can make the model measure uncertainty of large-scale SIS. Schema integration is a process with multi-attribute and decision-making. Uncertainty ratio based on all known entropy of rough set is adopted for measuring uncertainty of submodels of SIS. Uncertainty measure of process model is used in whole uncertainty measure. The method for synthesizing the uncertainty ratio is provided. Experimental results show that the presented model is feasible and effectual.

Key words: uncertainty measure; rough set theory; schema integration; all known entropy; process model

因诸多不确定性因素的存在,使不确定性成为模式集成的一个固有性质,它极大地影响着模式集成的准确实现和性能优化。由于模式集成的不确定性是使数据集成存在不确定性的一个方面[1],若能够很好地度量其不确定性,对于数据集成的成本估算、规模处理、策略选择以及性能提高都颇具决策参考价值。由于在制定决策过程中不确定性是不可避免的,不确定性度量在不确定性推理中扮演着重要角色[2]。

粗糙集(Rough set,RS)是由Pawlak[3]提出的一种度量和处理不确定信息的重要概念,相关的知识对不确定性的度量提供了一系列严密的分析与

基金项目:国家自然科学基金(60903027)资助项目;江苏省自然科学重大研究(BK2011023)资助项目;江苏省自然科学基金(BK2011370)资助项目。

收稿日期:2011-11-16;**修订日期**:2012-01-06

通讯作者:胡文彬,女,博士研究生,副教授,1976年生,E-mail:hwb1008@163.com。

操作,它无需提供问题所需处理的数据集合之外的任何先验信息。长期以来人们比较注重研究RS本身不确定性的度量[4-5]和基于RS的不确定性度量方法[6-7],这些大都是针对单一集合、单一决策系统和单粒度集合不确定性进行度量的方法,而少有对复杂信息系统的不确定性度量进行研究。在度量决策系统不确定性方面,人们基于RS已经提出多种度量方式:(1)基于近似质量的系统不确定率[8];(2)基于条件等价类确定性的度量方式[9];(3)基于信息熵及变精度粗糙集模型的度量方式[10];(4)基于全知熵的度量方式[8]。方式(1)完全建立在"相对正域"这一基本的RS概念上,其度量结果可能会夸大系统的不确定性。方式(2)类似于方式(1),只是扩充系统的确定部分。方式(3)容易受系统规模等因素影响,度量结果受参数β控制。方式(4)中"基于全知熵"概念的不确定性度量方式较方式(1~3)的性能好,对系统的不确定性比较敏感,能够较为准确地反映系统不确定性的变化规律。

诸多度量方式中均未详细讨论影响系统不确定性的因素对度量结果的影响程度,且都是针对决策信息系统中的决策规则而提出的。本文通过分析模式集成中产生不确定性的原因和其分段决策的特点,分段度量其不确定性;提出了一个能够有效度量模式集成不确定性的模型,运用全知熵不确定率的系统不确定性度量方法进行不确定性度量;引入过程模型不确定性度量,综合考虑了产生各种不确定性的因素对度量的影响;给出了多段不确定率合成的方法;模式对象及其属性清洗可降低度量规模,使得所提出的不确定性度量模型能够适应大规模模式集成的不确定性度量(Uncertain measure of schema integration, UMSI)。对UMSI的研究可为度量其他复杂系统的不确定性提供借鉴。

1 模式集成中的不确定性

1.1 多属性分段决策的模式集成

模式匹配和模式合并是模式集成的两个核心问题[11],在模式集成系统(Schema integration system, SIS)中进行模式集成,它是一种多属性分段决策过程,SIS的不确定性度量即是UMSI。UMSI包含模式匹配和模式合并的不确定性度量两部分。

有关系统的定义如下:

定义1(模式集成系统(SIS))。SIS=(S, SMA, SME, IS, f, A)是多属性分段决策系统,其中:$S=\{s_1, s_2, \cdots, s_n\}$为源模式集合;SMA为模式匹配,SME为模式合并,IS为集成模式集,SMA的输出为SME的输入;f:(SMA, SME)$\rightarrow IS$是从模式匹配和模式合并到集成模式的映射函数;$A=C\cup D$, $C\cap D=\Phi$, $C=C_1\cup C_2$表示条件属性集,C_1和C_2分别是模式匹配和模式合并的条件属性集,$D=D_1\cup D_2$表示决策属性集,D_1和D_2分别是模式匹配和模式合并的决策属性集。

SIS是由模式匹配决策子系统(Schema matching decision subsystem, SMaDS)和模式合并决策子系统(Schema merging decision subsystem, SMeDS)组成的多属性分段决策系统。

根据以上定义可知SIS是一决策系统,因此UMSI可采用基于RS的系统不确定性度量方法进行度量。

1.2 模式集成中不确定性的表现

SIS利用模式匹配和模式合并技术从多个原始数据源模式构造统一的集成模式。模式集成中内在不确定性的具体表现[12]有:(1)不同数据源使用多种术语(词汇)表示同一概念;(2)同一概念在不同的数据源中表达不同的含义;(3)相同名字在不同模式中表示不同的概念;(4)各数据源使用不同的结构来表示相同(或相似)的信息;(5)各数据源中的模式之间存在着各种联系。由于现在大多数数据仍然保存于关系数据库中,因此本文主要研究关系模式集成中的不确定性度量。

从以上表现来看,对象与对象之间的粗糙性导致了对象与概念之间的关系具有不确定性,因此,基于RS的方法可很好地进行模集成的不确定性度量。

1.3 模式集成的不确定性定义

定义2(SIS的不确定性)。存在模式集成系统SIS=(S, SMA, SME, IS, f, A),其中SMA实例与运用匹配结果进行的SME实例均为确定的,则SIS为确定的,否则SIS为不确定的。

定义3(不确定模式匹配(USMA)[12])。不确定模式匹配定义为一个四元组$<O, C_1, D_1, UM>$,其中:O为模式对象集;C_1为匹配条件属性集;D_1为决策属性集;$UM=\{um_1, um_2, \cdots, um_n\}$为模式对象间的不确定匹配关系集合,$C_1\times D_1\in UM$。

定义4(不确定模式合并(USME))。不确定模式合并定义为一个四元组$<O, C_2, D_2, UD>$,其中:O为模式对象集;C_2为匹配条件属性集;D_2为决策属性集;UD为集成模式对象的不确定度集合$\{ud_1, ud_2, \cdots, ud_n\}$。

性质　模式集成系统中的不确定性具有传递性。

在SIS 中，对输入的处理具有时序特征，因此在模式匹配处理中产生的不确定性将会保持到模式合并阶段，而且这种传递性会影响不确定性的度量。

定义5（系统不确定性度量）。S 为非空集合，F 为 S 上的代数，F 中的每个元素 E 都被赋予一个数 μ，如果满足：

(1) $\mu(S)=1$；

(2) 当 $e_1 \subseteq e_2$ 时 $(e_1, e_2 \in E)$，有 $\mu(e_1) \leqslant \mu(e_2)$；

(3) 对于任意的 E，有 $\mu(E)+\mu(E^C)=1$；

(4) 对于每个可数序列 $\{e_i\}$，有 $\mu\{\bigcup_{i=1}^{\infty} e_i\} \leqslant \{\bigcup_{i=1}^{\infty} \mu(e_i)\}$，那么 μ 就被称为系统的不确定性度量。

2　模式集成的不确定性度量模型

2.1　不确定性度量模型

模式集成的不确定性度量模型（Uncertainty measures model of schema integration，UMMSI）由模式对象清洗子模型（Schema object cleanout submodel，SOC）、模式匹配不确定性度量子模型（Uncertainty measures of schema matching submodel，UMSMat）、模式合并不确定性度量子模型（Uncertainty measures of schema merging submodel，UMSMer）和总体不确定性度量子模型（Whole uncertainty measures submodel，WUM）组成，模型结构如图1 所示。

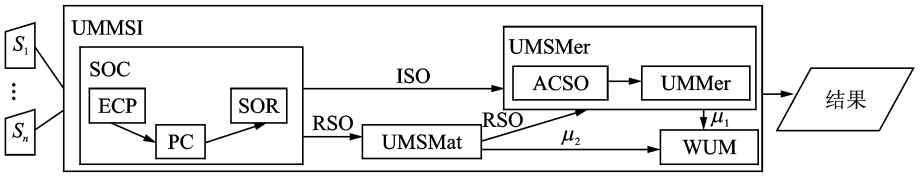

图1　模式集成的不确定性度量模型

图 1 中：模式 $S_i(i=1,\cdots,n)$ 为输入，UMMSI 的输出是模式集成的总体不确定率；SOC 的主要任务是对输入模式所包含的模式对象进行确定部分和不确定部分的划分；UMSMat 用于度量模式匹配部分的不确定性；UMSMer 用于度量模式合并部分的不确定性；WUM 的任务是根据前面子模型计算得到 μ_1，μ_2 和过程不确定率计算模式集成的总体不确定率。

UMSI 的过程是根据运行时管理者的决策或过程数据有条件执行，因而其过程中存在不确定性是毫无疑问的。使用Petrinets 表示的UMSI 执行的过程模型（Process model，PM）如图 2 所示。

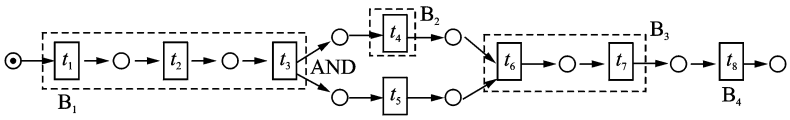

图2　UMMSI 的过程模型

B_1 为SOC 的执行过程，B_2 为UMSMat 的执行过程，B_3 为 UMSMer 的执行过程，B_4 为 WUM 的执行过程。

2.2　模式对象清洗子模型

SOC 对输入模式所包含的模式对象进行等价类划分（Equivalence class partition，ECP）后，进行正域计算（Positiveness calculation，PC）和模式对象的约简（Schema object reduction，SOR），分离出来的独立模式对象（Independent schema object，ISO）是系统的确定部分，直接进入模式合并阶段，其他模式对象即为模式对象的约简（Reduction of schema object，RSO），RSO 即是需进行匹配的模式对象集，其中存在不确定部分。

2.3　模式匹配不确定性度量子模型

UMSMat 对 SMaDS 的不确定性进行度量，SMaDS 能够表达其系统内的条件属性知识 C_1 和决策属性知识 D_1，这个系统的不确定性结构和程度可以表达全部相关信息，因此使用较为合理的基于全知熵的度量方式[9]度量系统的不确定性。信息熵是信息理论中用于分析不确定程度的一种重要度量，它从统计学角度得到描述一个给定问题所需的最小信息量，从而以所需信息量的多少来衡量不确定性的程度[8]。根据全知熵不确定率的定义[8]，模式匹配决策子系统SMaDS 的全知熵不确定率为

$$\mu_1 = 1 - \frac{(H_{\mathrm{all}}(C_1 \rightarrow D_1) - H(D_1))}{(\log(|O_1|) - H(D_1))}$$

其中：$H_{all}(C_1 \rightarrow D_1)=H(C_1)+H(D_1|C_1)$为全知熵；$H(C_1)$为$C_1$在$O$上的信息熵；$H(D_1|C_1)$为条件熵；$H(D_1)$为$D_1$在$O$上的信息熵。

模式匹配的不确定性主要体现在模式对象名、语义关系、属性名、属性类型、关键字约束和数据实例6个方面，在进行不确定性度量时将这6个方面作为SMaDS的条件属性C_1的集合元素，决策属性是匹配结果。

2.4 模式合并不确定性度量子模型

UMSMer对SMeDS的不确定性进行度量，首先进行模式对象属性清洗(Attribute cleanout of schema object, ACSO)，再由UMMer根据ACSO的处理结果进行模式对象合并。SMeDS同SMaDS一样表达了全部系统相关信息，因此可使用较为有效的基于信息熵的度量方式度量系统的不确定性。

定义6(模式对象属性清洗(ACSO))。OA是模式对象属性关系集，ST是属性分类器，ST={AEP, APC, SOAR}。具体定义如下：

ACSO:={OA, ST}

AEP：是模式对象属性等价类划分，论域U为模式对象属性集合。

APC：根据U关于各等价关系的划分分别计算决策属性相对于各等价关系的相对正域。

SOAR：根据计算的相对正域判断出可约去部分与不可约去部分(独立部分)。

根据全知熵不确定率的定义，模式合并决策子系统SMeDS的全知熵不确定率为

$$\mu_2 = 1 - \frac{(H_{all}(C_2 \rightarrow D_2) - H(D_2))}{(\log(|O_2|) - H(D_2))}$$

其中：$H_{all}(C_2 \rightarrow D_2)=H(C_2)+H(D_2|C_2)$为全知熵，$H(C_2)$为$C_2$在$O$上的信息熵，$H(D_2|C_2)$为条件熵，$H(D_2)$为$D_2$在$O$上的信息熵。

经过模式对象及其属性清洗后，要合并的模式对象中模式对象属性和数据实例两个方面存在不确定性，在进行不确定性度量时SMeDS的条件属性C_2={模式对象属性，数据实例}，决策属性是合并结果。

2.5 总体不确定性度量子模型

WUM的计算内容由μ_1，μ_2和UMMSI的过程不确定率3部分组成。SIS中的不确定性主要由SMaDS，SMeDS和执行过程中的不确定性决定，每个部分对SIS的影响比重是不同的，在计算总体不确定率时，权重的分配至关重要，将直接影响SIS的不确定性度量。

SIS的过程不确定性度量公式如下

$$U(PM) = -\sum_{k=1}^{M} P(BS_k)\log_2 P(BS_k) + \sum_{g=1}^{N} P(B_g)U(B_g)$$

其中：M为可能执行任务的总数，BS_k为第k个可能被执行的任务，$P(BS_k)$为执行概率，N为过程中任务块的总数量，B_g为过程中第g个任务，$P(B_g)$为为了完成整个过程而由所有M个可执行任务执行的B_g的概率。

由图2可知，在SIS中PM由B_1，B_2，B_3和B_4组成，根据文献[6]中对过程的分解原理，有

$$U(PM) = U(B_1) + U(B_2) + U(t_5) + U(B_3) + U(B_4)$$

其中：由于B_1，B_3和B_4均为顺序过程，t_5是个必须执行的任务，因此$U(B_1)=0$，$U(B_3)=0$，$U(B_4)=0$，$U(t_5)=0$，又$U(B_2)=U(t_4)$，因此$U(PM)=U(t_4)$。

PM不确定率μ_{PM}的计算公式如下

$$\mu_{PM} = \frac{U(PM)}{N}$$

其中：N为执行过程的总数。

B_2的过程是个较为复杂的多属性匹配过程，其过程不确定性的度量可根据文献[6]中所提供的过程模型不确定性度量方法给出。

计算SIS的整体不确定率的公式如下

$$\mu_{whole} = \omega_1\mu_1 + \omega_2\mu_2 + \omega_3\mu_{PM}$$

其中：$\omega_1 = \frac{\mu_1}{\mu_1+\mu_2+\mu_{PM}}$，$\omega_2 = \frac{\mu_2}{\mu_1+\mu_2+\mu_{PM}}$和$\omega_3 = \frac{\mu_{PM}}{\mu_1+\mu_2+\mu_{PM}}$分别为SMaDS，SMeDS和过程不确定率对SIS不确定性影响的权重。

3 实例与分析

3.1 实 例

实验数据为2005～2009年的VFP等级考试数据(取每年的其中一次考试数据)，数据情况见表1。较小集成不确定性的度量情况中模式数量较少，很容易对提出的模型和方法进行相关验证。因此，本实验中针对较大规模模式集成不确定性的度量情况对文中提出的模型和方法进行验证，数据中有5个模式，共包含56个模式对象。实验参数见表2。在模式匹配的不确定性度量中，经过模式对象清洗后，对象个数减至52个，条件属性C_1={模式对象名、语义关系、属性名、属性类型、关键字约束、数据实例}，属性值域$V_{C1}=V_{D1}$={0(不匹配)，1(匹配)，2(不确定)}，决策属性D_1={匹配结果}。在模式合并的不确定性度量中，经过模式对象清洗后，合并的对象个数减至49个，经过模式对象属性清洗后合并规模明显降低，条件属性C_2={模式对象属性，

数据实例}，属性值域$V_{C2}=V_{D2}=\{0$(合并)，1(不确定)$\}$，决策属性$D_2=\{$合并结果$\}$。模式集成的不确定性度量数据见表 3。

表 1　实验数据情况

序号	1	2	3	4	5
年份	2005sp	2006au	2007au	2008sp	2009au
$\|O\|$(对象个数)	17	17	7	8	7

注：sp 表示春节；au 表示秋节。

表 2　模式匹配不确定性度量和模式合并不确定性度量实验参数

模式匹配			C_1		D_1	
$\|S\|$	$\|U_P\|$	$\|O_1\|$	$\|C_1\|$	V_{C1}	$\|D_1\|$	V_{D1}
5	56	52	6	{0,1,2}	1	{0,1,2}
模式合并			C_2		D_2	
$\|S\|$	$\|U_H\|$	$\|O_2\|$	$\|C_2\|$	V_{C2}	$\|D_2\|$	V_{D2}
5	56	49	2	{0,1}	1	{0,1}

表 3　模式集成的不确定率

模式	SMaDS	SMeDS	SIS	SIS
不确定率	$\mu_1=0.43$	$\mu_2=0.29$	$\mu_{PM}=0.10$	$\mu_{\text{whole}}=0.34$

注：表中 $\mu_{PM}=0.10$ 是假定值，需根据具体情况进行计算。

3.2　分　　析

从表 2 中的数据可以看出 UMMSI 中的 SOC 和 UMSMer 中的 ACSO 使得匹配规模和合并规模明显减小，从而提高了整个不确定性度量过程的执行效率。表 3 中的数据验证了 UMMSI 能够正确地度量 SIS 的不确定性，且 SMaDS 的不确定率要比 SMeDS 的不确定率影响大。综上实验和分析证明基于全知熵的 UMMSI 是可行、有效的。

4　结束语

本文通过分析模式集成中产生不确定性的原因表现以及 SIS 的特点，给出相关的系统定义和不确定性度量定义。由于 RS 在处理不确定性问题方面具有很大的优势，提出了基于 RS 全知熵的模式集成系统的不确定性度量方法和模式集成系统的确定性度量模型 UMMSI。UMMSI 中采用了模式对象及其属性清洗算法，降低了处理大规模模式集成的复杂度，使模式集成系统的不确定性度量成为可能；给出的不确定性度量方法中融合了过程模型的不确定率，综合考虑了度量过程中所出现的不确定性。本文给出的方法和模型可供其他复杂系统的不确定性度量借鉴，今后还需在方法和模型性能上进行细致的研究，进一步深入研究模式集成系统的合理拆分和总不确定性的合成方法，使不确定性度量更为准确、合理。

参考文献：

[1] Dong X L, Halevy A Y, Yu C. Data integration with uncertainty [J]. The VLDB Journal, 2009, 18 (2): 469-500.

[2] Yu Daren, Hu Qinghua, Wu Congxin. Uncertainty measures for fuzzy relations and their applications [J]. Applied Soft Computing, 2007, 7 (3): 1135-1143.

[3] Pawlak Z, Skowron A. Rudiments of rough sets [J]. Information Sciences, 2007, 177(1):3-27.

[4] Qian Yuhua, Liang Jiye, Li Deyu, et al. Measures for evaluating the decision performance of a decision table in rough set theory [J]. Information Sciences, 2008, 178(1): 181-202.

[5] Liu Guilong. Rough set theory based on two universal sets and its applications [J]. Knowledge-Based Systems, 2010, 23(2):110-115.

[6] Jung Jaeyoon, Chin Changho, Cardoso J. An entropy-based uncertainty measure of process models [J]. Information Processing Letters, 2011, 111(3): 135-141.

[7] Magro M C, Pinceti P. A confirmation technique for predictive maintenance using the rough set theory [J]. Computing & Industrial Engineering, 2009, 56 (4):1319-1327.

[8] 赵军，周应华. 基于粗集理论的系统不确定性度量方式研究[J]. 小型微型计算机系统，2010, 31(2): 354-359.
Zhao Jun, Zhou Yinghua. Study on system uncertainty measure based on rough set theory [J]. Journal of Chinese Computer Systems, 2010, 31 (2): 354-359.

[9] Wang Guoying, He Xiao. A self-learning model under uncertain condition [J]. Journal of Software, 2003, 14 (6): 1096-1102.

[10] Chen Xianghui, Zhu Shanjun, Ji Yindong. Rule uncertainty measurements based on entropy and variable rough set [J]. Journal of Tsing hua University (Science & Technology Edition), 2001, 47(3): 109-112.

[11] Magnani M, Rizopoulos N, Brien M C, et al. Conceptual Modeling-ER 2005 [M]. Heidelberg: Springer Berlin, 2005: 31-46.

[12] 胡文彬，李千目，张 宏. 基于领域知识的不确定性关系模式集成 [J]. 南京理工大学学报(自然科学版), 2010, 34(4): 409-414.
Hu Wenbin, Li Qianmu, Zhang Hong. Uncertainty retional schema integration based on domain knowledge [J]. Journal of Nanjing University of Science and Technology (Natural Science), 2010, 34 (4): 409-414.