RSL:基于Rough Set的表示语言*

周育健 王 珏

(中国科学院自动化研究所 北京 100080)

摘要 本文给出了一种基于Rough Set理论的表示语言—RSL,该语言包括面向应用与面向研究两部分.应用部分主要服务于对信息表进行分析与处理的用户,研究部分则主要是为研究Rough Set及利用Rough Set理论构造更复杂算法的研究者所设计.鉴于Rough Set理论中求最小约简的过程是NP完全问题,为了使RSL表示语言可以分析与处理规模更大的信息表,本文还为RSL表示语言设计了一个新的对求取最小约简而言的领域独立的近似算法.

关键词 Rough Set,决策分析;不确定表示,表示语言.

中图法分类号 TP18

Rough Set(以下简称RS)理论是波兰科学家Z. Pawlak在1982年提出的一种数学理论[1],主要用来进行数据分析,尤其是对不精确和不确定的数据进行分析.[2]该理论中所提出的核、约简、上下近似等概念,为数据分析、决策分析等研究提供了新的数学工具.

在这10余年中,RS理论一方面在自身的理论体系方面进行了不断的拓展和完善,另一方面它已渗透到人工智能各分支,如近似推理、数字逻辑分析和约简、控制算法获取、机器学习、机器发现及模式识别等研究领域,并使该理论在解决实际问题中获得了愈来愈广泛的应用,如利用RS理论作医疗诊断分析、从大型数据库中发掘知识、进行数据处理及在药物学、商业、银行、市场研究、工程设计、气象学、震动分析、冲突分析、图象处理、语音识别、在线系统分析、决策分析等领域的应用.美国宇航局的Johnson空间实验中心已利用LERS系统(基于RS的事例学习系统)作为工具来研制决策专家系统.

随着该理论的日益成熟和应用的日益广泛,建立与设计一种关于RS的表示语言就显得日益重要,文献[3]也提到过这一需求.目前,尽管RS理论已得到广泛地应用,但一般均是将其作为某类系统(例如,学习系统)的一个组成部分,而没有着力于建立一种基于这个理论的独立的表示语言.因此,我们从应用和研究两方面考虑,构造了一种架构于Common Lisp上的关于RS的表示语言——RSL.它在结构上分为面向研究和面向应用2个部分,前者所提供的直接数据处理工具,是为分析数据的用户所设计;后者则提供了多种灵活的底层概念函数,以供理论研究者进行构造性使用.同时,该语言与Common Lisp完全兼容,用户

---

* 本文研究得到国家863高科技项目基金资助.作者周育健,女,1971年生,硕士生,主要研究领域为人工智能.王珏,1948年生,研究员,主要研究领域为人工智能.

本文通讯联系人:周育健,北京 100080,中国科学院自动化研究所

本文1996-09-11收到修改稿

可在此语言的基础上架构更复杂的语言和算法. 现已证明RS理论中求全部约简的过程是NP完全问题[4],为了使RSL表示语言可以分析与处理规模更大的信息表,我们为RSL表示语言设计了一个新的求取尽可能最小的约简的近似算法,并且该算法是领域独立的,因而利于形成一个独立的语言.

本文侧重于利用RS理论进行不确定表示及数据约简,对于RS理论中的决策逻辑部分暂不考虑. 另本文将有关RS的基本概念放于附录中,供读者参考.

1 RSL语言的描述1——面向应用的部分

面向应用部分的语言主要是针对AI中的一些应用,如对大型数据库的约简、求核、求最小约简等. 因此,语言的面向应用部分主要包括如下几部分:表的定义、约简、核、近似最小约简以及对信息表进行分析的一些辅助工具.

为描述方便,在这一部分语言描述中,令table代表待处理的决策表,condition-attributes代表条件属性的集合,decision-attributes代表决策属性的集合.

1.1 决策表的定义

要对数据进行分析,在RSL语言中首先得将数据定义为一张决策表. 所有的应用分析都将在此表上进行. RSL语言提供一个宏命令:

(diftable name condition-attributes decision-attributes & rest objects)

来进行这个转换,其中name是用户给定的信息表的名字,该宏命令输出一张数组形式的决策表.

这是从键盘输入来形成决策表的方式,为了方便用户,RSL语言另外提供了从数据文件定义决策表的方式,并提供了多种不同的文件格式供用户选择.

1.2 决策表的约简

在数据分析中对数据的化简一直是一个非常重要的问题,如气象预测时大量的数据都必须进行化简之后才可进行有效的分析和利用. 因此对决策表求约简在RSL语言应用中是非常重要的一部分内容. RSL语言提供函数:

(value-reduction-of-decision-table table condition-attributes decision-attributes)

来对决策表中每个例子的值进行约简. 该函数输出决策表中所有例子的值约简的集合. 因为该函数利用组合方法进行求解,故只适于小规模的数据,在本文后面还将给出一个近似算法来计算大规模数据的值约简.

另外在应用中也可能对所有的特征列求约简,RSL语言中也提供了相应的函数来求其约简. 对一般的信息表(不区分决策属性与条件属性的信息表)也提供了一套求列的约简及值约简的函数.

1.3 决策表的核

决策表中的核反映了决策表中的一些最必要的属性值,因此在决策分析中核起着非常重要的作用. 如在银行的报表分析中,哪些核值将在决策中引起更充分的注意. RSL语言提供函数:

(values-core-of-decision-table table condition-attributes decision-attributes)

来对决策表中每个例子求取值核. 该函数输出与决策表同维的值核表. 其中元素的取值为原

值(如该元素为值核)或者*(如该元素非值核).

另外对于应用中可能遇到的求特征列的核,RSL语言也提供了相应的函数.对于一般信息表也提供了一套求值核与特征列核的函数.

1.4 辅助应用语言

为了实现上述基本功能,该语言还提供了一系列的辅助函数供用户选择.这一部分辅助函数主要包括表的整理与编辑语句和一些基本的显示操作.其中包括表的整理(删除表中重复的行及删除重复的列)和表的编辑(向表中添加一列、添加一个例子及修改表中的一个值).基本的显示操作包括一些输出功能和按特定要求重写表等.

2 最小约简的近似算法

在RS理论中求得值约简后即可自动抽取出规则,但每一个例子都可能存在多个值约简,求取所有例子的所有值约简的组合是一个NP完全问题[4],因而在实际的应用中遇到大的数据就将无法求取所有值约简.同时,在实际应用中,一般并不需要求取所有的值约简的组合,而只是关心最小的值约简组合(即值约简后不同的例子数最少),但最小值约简的完备性在理论上还无法证明,因而采用近似算法.文献[4]中给出了以属性的重要性作为附加信息来求取关于决策表的属性的最小约简的近似算法,本文则给出了以特殊定义的属性的值的重要性为附加信息的关于最小值约简的近似算法.

令R表示一些特征的集合,$\gamma(R)_j$表示决策表中包含第j个例子的特征集合R的等价类中成员的个数,D为决策属性集合,定义j例子中的特征集合R的分类一致性系数为:

$$Q(R)_j=\frac{\gamma(R\cup D)_j}{\gamma(R)_j}$$

令a为条件属性集合中某个条件属性,RED为已选定作为约简的特征集合,$a\notin RED$,定义j例子中的属性a的取值关于集合RED的可合并度为:

$$COMB(RED,a,D)_j=\frac{\gamma(RED\cup\{a\}\cup D)_j}{\gamma(RED\cup D)_j}$$

定义第j个例子中的属性a的取值关于集合RED的敏感度为:

$$SENS(RED,a,D)_j=\frac{(\gamma(RED)_j-\gamma(RED\cup D)_j)-(\gamma(RED\cup D)_j-\gamma(RED\cup D\cup\{a\})_j)}{\gamma(RED)_j-\gamma(RED\cup D)_j}$$

定义第j个例子中的属性a的取值关于集合RED的约简的重要性为:

$$SIGNIF(RED,a,D)_j=COMB(RED,a,D)_j+SENS(RED,a,D)_j$$

该重要性即作为关于最小值约简近似算法中选择属性值的附加信息.RSL提供了基于这种原理的关于最小值约简的近似算法:

从$j=1$开始循环直至最后一个例子:

1. 计算j例子的所有条件属性P的分类一致性系数$Q(P)j$.

2. 计算j例子的值核集合C.无核时C为空集合.

3. 令$RED=C$.

(a) 如果$Q(RED)_j=Q(P)_j$,则j例子的值约简集合即为RED,转入下一个例子的求解,令$j=j+1$,返回第1步.否则进行步骤(b).

(b) 计算所有$a\in A$(除RED以外的其余条件属性)的属性值重要性:$SIGNIF(RED,a,D)_j$

(c) 取出 *SIGNIF* 取值最大的属性 a，令 $RED=RED\cup\{a\}$

(d) 返回步骤(a).

RSL 语言提供了函数：

(find-min-value-reduction-of-decision-table-by-signif table condition decision)

用来求取近似的最小值约简.

为了检验该方法的有效性，本文利用上述的 3 个指标分别设计了 3 种方法对美国国会选举及豌豆疾病 2 例(例子的详细介绍见第 4 节)进行了实验. 方法 1 采用 COMB 信息作为附加信息，方法 2 采用 SENS 作为附加信息，方法 3 采用 SIGNIF 作为附加信息. 实验结果见表 1，从表中可看到方法 3 给出了例子数尽量少的约简. 另外对文献[1]中的一些小例子，方法 3 都求出了最小值约简.

表 1

	例子 1			例子 2	
值约简后的例子数	80	58	50	142	90
约简后例子平均长度	5.2	2.8	3.5	2.98	4.92
约简后剩余比率	4.3%	2.3%	2.5%	3.94%	4.12%

上面定义中的敏感度表示了该属性值对区分不同等价类的能力，与文献[5]中的属性的重要性定义类似，但在列的约简中不必考虑规则(即约简后例子)的合并，因而只用敏感程度即可满足要求，而值约简中须考虑例子的合并，因而单纯使用敏感度将不够，故本文的计算附加信息中加上了属性值的可合并性因素，即当属性值的敏感性和可合并性都好时，该属性值越有可能入选最小值约简集合. 从实验结果看(见表 1)，使用属性的重要性度量比单纯使用敏感度或可合并度的效果要好，即最后形成的规则的个数最少. 因而这种定义下的计算附加信息是有效的. 另外，因它与领域无关(即无需领域知识)，因此可作为 RSL 语言的一部分.

3 RSL 语言描述 2——面向研究的部分

这一部分语言主要是描述 RS 理论的一些基本概念，如等价类、不可区分关系、集合的粗糙近似、成员的近似关系、近似程度等. 对于作理论研究的研究者可利用 RSL 语言来构造更上层的算法及用其来验证 RS 理论中的一些性质，如核是所有约简的交集等. 为描述方便，在这一部分语言描述中令 relation 是一个等价关系，令 setx，sety 表示 2 个集合.

3.1 等价类

等价类是 RS 理论中的一个重要概念，它描述一个等价关系中的不可区分的概念集合. RSL 语言提供函数：

(def-equivalence-class table attribute attribute-value)

从 table 表中取 attribute 的值为 attribute-value 的等价类.

3.2 不可区分关系

RS 理论中的一个重要概念是不可区分关系，它描述了一个等价关系集合的最细划分. RSL 语言提供函数：

(indp relation-set)

求取等价关系集合的不可区分关系,其中 relation-set 是等价关系的集合,该函数返回一个不可区分关系.

3.3 关于集合的粗糙近似(Approximation of set)

在 AI 中常常对于一个待求解的问题需估计其用现有知识求解时的准确程度. RS 理论的一个非常显著的优点是能够完全采用确定的方法进行不确定性分析即有关集合的一系列粗糙性概念. 这也是 RS 理论在 AI 中得到日益广泛应用的一个原因.

这部分语言主要是对给定的概念集合求取其各种近似,典型语句如集合的粗糙下近似:

(rough-lower-approximation relation setx)

该函数返回集合 setx 关于等价关系 relation 的粗糙下近似集合.

类似的还有集合的粗糙上近似(rough-upper-approximation);集合的粗糙正区域(rough-positive-region);集合的粗糙负区域(rough-negative-region);集合的粗糙边界区域(rough-borderline-region)等.

为了用户的应用方便,RSL 语言也给出了有关分类的粗糙近似、成员的近似属于关系等一系列函数. 对于不精确性的拓扑特征又特别给出 4 种判断函数,即集合可粗糙定义(roughly-r-definable)、集合外不可粗糙定义(x-externally-r-undefinable)、集合内不可粗糙定义(x-internally-r-undefinable)、集合完全不可粗糙定义(x-totally-r-undefinable)以及集合间的粗糙包含关系、2 个集合的粗糙相等及分类中的类似的概念,如分类的粗糙近似的数字特征、分类的约简,核,独立性、分类的并集的约简,核,独立性等概念,RSL 语言中都提供了相应的函数.

4 应用该语言的例子

例 1:对医生诊断表的分析与处理

例子选自文献[3]. 在如下的医生诊断表 2 中 Headache,Muscle_pain,Temperature 为条件特征,Flu 为决策特征.

表 2　医生诊断原始数据表

U	Headache	Muscle_pain	Temperature	Flu
e1	yes	yes	normal	no
e2	yes	yes	high	yes
e3	yes	yes	very_high	yes
e4	no	yes	normal	no
e5	no	no	high	no
e6	no	yes	very_high	yes

对该表进行列约简的过程为:

(setq table1 (make-table "physician.dat"))

(setq table2 find-min-reduction-of-columns-of-decision-table-by-signif table1 "(Headache Muscle_pain Temperature)" (Flu))

执行之后可得到一张约简后的决策表,如表 3.

表3 经过特征约简后的医生决策表

U	Headache	Temperature	Flu
e1	yes	normal	no
e2	yes	high	yes
e3	yes	very_high	yes
e4	no	normal	no
e5	no	high	no
e6	no	very_high	yes

这时属性Muscle_pain被删除,这说明,根据实例集合,这个属性对决策无意义.再对删减后的表求取所有的值约简,

(value-reducts-of-decision-table table2 "(Headache Temperature)" (Flu))

得到如表4的所有的值约简表,即所有潜在的规则.

表4 所有的值约简表

U	Headache	Temperature	Flu
e1	*	normal	no
e2	yes	high	yes
e3	*	very_high	yes
e4	*	normal	no
e5	no	high	no
e6	*	very_high	yes

利用最小值约简函数,(find-min-value-reduction-of-decision-table-by-sensitive table2 "(Headache Temperature)" (Flu))

则可获得一组更为简洁的诊断规则,如表5.

表5 最小值约简表

U	Headache	Temperature	Flu
e1	*	normal	no
e2	yes	high	yes
e3	*	very_high	yes
e5	no	high	no
e6	*	very_high	yes

以下2个例子的数据原来都是用来作机器学习的,本文用来验证RS的有效性.

例2:1984年美国国会投票记录数据约简

该数据来自the UCI Repository of Machine Learning Databases and Domain Theories,由Jeff Schlimmer在1987年提供,出自美国国会的年鉴季刊(CQA,Vol. XL).该例子集给出了1984年美国参议院的投票记录,共435个记录,其中有267个民主党派的记录和168个共和党派的记录,每个记录包含16个条件特征和1个分类特征.Jeff Schlimmer用该例子集作概念获取的研究.本文利用RS技术对特征进行最小约简,将特征数约简为9个,再进行值的最小约简获得57条规则,每条规则平均包含3.5个特征值,总的数据约简量达97.5%以上.

例3:豌豆疾病数据约简

数据来源同例子 2,由 Ming Tan & Jeff Schlimmer 于 1988 年提供,R. S. Michalski 与 R.L. Chilausky 曾经用该数据研究被告知的学习与从例子学习. 该数据包含 307 个事例,每个事例包含 35 个特征和 1 个类别属性,共 19 个类别. 利用 RS 技术对特征进行最小约简,将特征数约简为 25 个,再进行值的最小约简获得 90 条规则,每条规则平均包含 4.92 个特征值,总的数据约简量达 95.9%以上.

5 总结与讨论

本文描述了一种基于 RS 理论的表示语言,为了使这个语言可以被使用于处理大型问题,我们为这个语言设计了一个基于属性值的重要性的关于最小值约简的近似算法,尽管我们还未能证明这个算法的完备性,但是,试验证明这个算法是十分有效的.

由于该语言比较简洁而实用,因此,用户可以在很短的时间内学会使用这个语言. 目前,我们已使用这个语言进行了几个比较大的问题的数据分析(实例集的规模少则几百个,多则上千个). 因为该语言目前还只是一个实验性的研究语言,故现在的版本是基于 LISP 语言的,如在应用中有进一步的需要,将其转换为 C 语言的形式将不是一件难事.

参考文献

1 Pawlak Z. Rough sets. Int. J. Comput. Inf. Sci., 1982,**11**:341~356.

2 Pawlak Z. Vagueness and uncertainty: a rough set prospective. Int. J. Comput. Intellig., May 1995,**11(2)**:227~232.

3 Pawlak Z. Grzymala-Busse J, Slowinski R *et al*. Rough sets. Int. J. Communications of the ACM, November 1995,**38(11)**:89~95.

4 Ziarko W. The discovery, analysis, and representation of data dependencies in databases. In: Piatetsky-Shapiro G, Frawley W J eds. Knowledge Discovery in Databases, AAAI/MIT Press, Cambrige, MA, 1990. 213~228.

5 Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks. Int. J. Comput. Intellig., May 1995,**11(2)**:339~347.

附录 有关 Rough Set 的基本概念

为便于读者了解这个语言,这里简单介绍 RS 理论的基本概念[1]:

· 知识、划分与等价关系(Knowledge, Classification and Equivalence Relation):RS 理论是一种新的知识表示和处理理论. 在该理论中知识被视为一种划分能力的表现. 其形式化定义如下:设 $U\neq\varnothing$ 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. 则 U 中的一族概念,就称为关于 U 的知识. 一个划分 C 定义为:$c=\{X_1,X_2,\ldots,X_n\}$,使得 $X_i\subseteq U, X_i\neq\varnothing, X_i\cap X_j=\varnothing$,对 $i\neq j, i,j=1,2,\ldots,n$. 且 $\bigcup_{i=1}^{n}X_i=U$. 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. 可以证明,U 上的一个划分与其上的一个等价关系是等价的. 因而关于 U 的一个知识库也可以理解为一个关系系统,其中 U 为论域,R 为 U 上的一族等价关系.

· 等价关系与属性(Equivalence Relation and Attributes):每一个等价关系描述的是领域 U 上的某一个属性,即属性就可看作一个等价关系.

· 信息表:在 RS 理论中假定现实世界中的信息是用一张表来表达,并称之为信息表,它是用属性值对构成的表,列为属性,行为例子. 对有些问题,属性将分为条件属性和决策属性,构成特殊的信息表——决

策表.

· 不可区分关系(Indiscernibility Relation):RS 理论中的一个重要概念是不可区分关系,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(Elementary set).

· 对知识不确定性的描述:任意给定的一个领域 U 中的有限集合在 RS 理论中可描述其相对于特定概念或者知识的不确定性,即定义它相对于一个特定等价关系的粗糙近似. 粗糙近似包括概念的下近似和上近似. 概念的下近似是指概念可由等价关系定义的最小集合,概念的上近似是指概念可由等价关系定义的最大集合. 同样对于分类可作类似的近似定义.

· 核及约简(Core and Reduction):核在信息表中代表着一些不变的信息,又分为列的核及值核. 列的核是由所有的独立的特征列组成的集合. 独立的特征列是指去掉该列后其余列所构成的不可区分关系将不同于原来所有列的不可区分关系. 值核是对每一个例子中的所有特征的值求核. 列约简是所有特征列集合的一个子集,该子集形成的不可区分关系与原集合的不可区分关系相同,并且该子集中的每一个特征列都是独立的,即任意一个约简代表着一个无冗余的完备信息. 值约简则是代表了每一个例子的无冗余的完备信息.

· 最小约简(Minimal Reduction):所有列约简中包含最少特征的列约简与所有例子的值约简的组合中包含最少例子数的值约简组合.

RSL: A REPRESENTATION LANGUAGE BASED ON ROUGH SET THEORY

ZHOU Yujian WANG Jue

(*Institute of Automation The Chinese Academy of Sciences Beijing* 100080)

Abstract This paper presents a representation language based on Rough Set theory, called RSL. This language has two parts: one is for application and the other for theory research. The application part is designed mainly for information analysis, such as data analyses and decision making. The research part tries to provide a tool for researchers on theory or on constructing more complicate algorithms. Finding the smallest reduction has been proved to be an NP-complete problem, a domain-independent approximate algorithm is presented in this paper. It makes the RSL more suitable to deal with large information tables.

Key words Rough set, decision analysis, uncertainty representation, representation language.

Class number TP18