Rough 逻辑及其在数据约简中的应用*

刘 清, 刘少辉, 郑 非

(南昌大学 计算机科学与工程系,江西 南昌 330029)

E-mail: qliu@263.net

摘要: 讨论了被定义在邻域值决策表上的Rough逻辑及其公式的真值,它在数据约简中的应用比Pawlak定义的决策表上的决策逻辑更加广泛.目前常用的数据约简方法有Pawlak的数据分析和Skowron的分明矩阵法.前者是非形式的,不易机械化;而后者虽说直观、易理解,但还要求生成一个分明矩阵的中间环节,从而造成时空上的不必要的开销.采取一边从邻域值决策表关于属性值邻域是分明的属性并构成邻域分明合取范式,一边做这种逻辑公式的等价变换直接得到邻域值决策表的诸多约简.由于不用生成分明矩阵的中间环节,这样便节省了空间和时间,提高了运行效率.对此,对拥有6个属性(4个条件和两个决策属性)以及102个个体的一致决策表或邻域值决策表进行处理并生成了约简的决策规则.用两种不同方法在PII 233/64M的微机上用DELPHI3.0分别对它们进行约简并得到相同的结果,采用一边从表中提取公式一边做约简的方法,所用时间约1分54秒;而用分明矩阵法却耗去1分55秒.由于增加了一个数组(分明矩阵),便增加了空间复杂度$O(m\times n^2)$,其中m为属性数,n为个体数,随着属性数和个体数的增加,所占的空间和时间也将急剧增加.可见,从空间和时间消耗上来看,这两种方法的优劣是十分明显的.

关键词: Rough逻辑;邻域值决策表;Rough逻辑公式演算;数据约简

中图法分类号: TP181 **文献标识码**: A

Rough集的创始人Z. Pawlak提出了基于决策表上的决策逻辑[1],它实质上是古典二值逻辑的特殊情况.但创立Rough集的背景是为了解决早在1904年谓词逻辑的创始人G. Frege提出的边界线上的含糊元素的计算问题,也就是说,逻辑上的绝对真和假是罕见的,而介于真和假之间的程度值是大量的,因此需要创立一种解决边界线上的含糊元素计算问题的理论.由于经等价关系定义的Rough集理论实现了Frege的计算真和假之间程度值的思想,所以Rough集理论的创立是为了改造经典二值逻辑,使之能计算出真和假之间程度值的多值逻辑.本文试图拓广这种决策逻辑,使其真值不局限于二值而是多值,为此可以计算真和假之间的程度值,实现Frege的边界线区域元素计算问题;另一方面,决策表上的属性值扩充为邻域值,因此也叫邻域值决策表.

1 Rough逻辑

Rough逻辑中的原子公式为$(a,\langle v,r\rangle)$,a是属性,$\langle v,r\rangle$是关于a在邻域值决策表上以v为中心,r为半径的属性值的邻域.为简便起见,我们记$\langle v,r\rangle$为$n(v)$.将这样的原子公式$(a,n(v))$和通常的逻辑联结词组合起来就得到这种逻辑的复合公式.

定义1. 设Rough逻辑语言RLL(rough logic language)是由属性集A、属性值集$V=\bigcup_{a\in A}V_a$以及通常的逻辑联结词和圆括号集及其如下递归定义的合式公式(wffs)组成:

* **收稿日期**: 1999-10-22; **修改日期**: 2000-01-21

基金项目: 国家自然科学基金资助项目(69773001);江西省自然科学基金资助项目(9911027)

作者简介: **刘清**(1938—),男,江西南昌人,教授,主要研究领域为人工智能,Rough集理论及其近似推理;**刘少辉**(1977—),男,江西余干人,硕士生,主要研究领域为人工智能;**郑非**(1975—),女,江西九江人,硕士生,主要研究领域为人工智能.

(1) 在 RLL 中一切形如$(a,n(v))$的原子公式,都是 RLL 中的 wff,其中$a\in A,n(v)\subseteq V_a$;

(2) 设φ,ψ是 RLL 中的 wffs,则$\sim\varphi,\varphi\vee\psi,\varphi\wedge\psi,\varphi\rightarrow\psi$和$\varphi\leftrightarrow\psi$也是 RLL 中的 wffs;

(3) 凡经有限次重复引用(1)和(2)而得到的公式都被认为是 RLL 中的 wffs.

对 RLL 中公式的解释以及在给出解释下对公式中出现的谓词变量、命题变量、函数项以及个体变量同时给予解释和赋值,记成$T_{IRuR}(\varphi)=K(|\varphi|)/K(U)$,其中$T_{IRuR}$被称作联合赋值符号[2],$K(S)$表示集合$S$的基数,$||$是命题变量或公式到个体集合的意义函数.

定义 2. 设$[\zeta_*,\zeta^*]$是算子集,其中ζ_*和ζ^*分别表示$|\varphi|$的下近似和上近似对U上的总元数之比[2],$[0,1]$是公式的真值集,如果$(\zeta_*+\zeta^*)/2\geqslant 0.5$,且$T_{IRuR}(\varphi)\geqslant\zeta^*$,则称$\varphi$是$[\zeta_*,\zeta^*]$有效,记成 OI-有效.反之,如果$(\zeta_*+\zeta^*)/2<0.5$,且$T_{IRuR}(\varphi)<\zeta^*$,则称$\varphi$是关于$[\zeta_*,\zeta^*]$不一致的,记成 OI-不一致.

属性的邻域值涉及 Frechet(V)空间概念[3].它非常一般,在其上建立的邻域系统不要求任何公理.设v是V中的一个点,v的邻域是V的一个子集,一般来说,这个子集可以包含v,也可以不包含v.这就是邻域闭包和内点,所以它能被说明是 F-拓扑.Lin 和 Liu 于 1996 年发表了关于闭包引导的上近似算子H和内点引导的下近似算子L的一阶 Rough 逻辑的文章[4].我们之所以引入邻域系统是因为它对每个实体都可以得到一个相容邻域(自反,对称,而没有传递),而一个相容邻域比较适合讨论个体关于属性的邻域值.RLL 中公式的语义用一个五元组

$$M=\langle U,A,\mathrm{OI},T_{IRuR},||\rangle$$

模型确定,其中U是非空个体域;A是非空属性集;OI 是算子区间$[\zeta_*,\zeta^*]$;T_{IruR}被称作联合赋值函数;$||$是公式到个体集合的意义函数.给定一个模型M,我们说公式φ是被模型中U上个体x关于属性的邻域值满足,这意味着在模型M上,使得$|\varphi|\subseteq U$,$(\zeta_*+\zeta^*)/2\geqslant 0.5$以及$T_{IRuR}(\varphi)\geqslant\zeta^*$.形式地写成$M\models_{NS}\varphi$,其中$NS$是邻域决策表,表示$\varphi$在$M$上关于$NS$的邻域值取真符号,本文中简写成$M\models\varphi$.

定义 3. 这种逻辑的公式关于逻辑联结词的意义被定义如下:

(1) $M\models(a,n(v))\cong|(a,n(v))|=\{x\in U:a(x)=v'\in n(v)\subseteq V\wedge T_{IRuR}((a,n(v)))\geqslant\zeta^*\wedge(\zeta_*+\zeta^*)/2\geqslant 0.5\}$;

(2) $M\models\sim\varphi\cong\sim M\models\varphi$;

(3) $M\models\varphi\vee\psi\cong M\models\varphi\vee M\models\psi$;

(4) $M\models\varphi\wedge\psi\cong M\models\varphi\wedge M\models\psi$;

(5) $M\models\varphi\rightarrow\psi\cong M\models\sim\varphi\vee M\models\psi$;

(6) $M\models\varphi\leftrightarrow\psi\cong M\models\varphi\rightarrow\psi\wedge M\models\psi\rightarrow\varphi$.

其中$\cong$是相似于的意思,因为从个体关于属性的邻域值上它是可满足的,所以不用等价于.显然,满足φ的$x\in U$是一个集合:$|\varphi|=\{x\in U:x\models\psi\}$.

命题

(1) $M\models\varphi\cong K(|\varphi|)/K(U)\geqslant\zeta^*\wedge(\zeta_*+\zeta^*)/2\geqslant 0.5\wedge|\varphi|\subseteq U$;

(2) $M\models\sim\varphi\cong K(|\sim\varphi|)/K(U)<\zeta_*\wedge(\zeta_*+\zeta^*)/2<0.5\wedge|\sim\varphi|\subseteq U$;

(3) $M\models\varphi\rightarrow\psi\cong|\varphi|\subseteq|\psi|$;

(4) $M\models\varphi\leftrightarrow\psi\cong|\varphi|=|\psi|$.

2 Rough 逻辑在数据约简中的应用

这种 Rough 逻辑公式的演算所引用的公理、定理和推理规则与古典逻辑非常相似,然而假言

推理规则被应用于Frechet(V)空间邻域上的推理,其推理符号用"$\approx$"表示,也就是说,一个个体x满足公式φ,是指x关于属性的值落在这个邻域内,被写成$a(x)=v'\in n(v)$.所以,假言推理规则MP(modus ponens)被修改成NMP(near modus ponens):如果$\approx\varphi\rightarrow\psi$和$\approx\varphi$,则$\approx\psi$,意味着任何个体关于属性的邻域值满足公式$\varphi$和$\varphi\rightarrow\psi$,则$\psi$也被这个个体关于属性的邻域值满足.

例:让我们考虑下面的邻域决策表(表1).

Table 1 NS=(U,A)

表1 NS=(U,A)

U/A	a	b	c	d	e
1	5	4	0	1	0
2	3	4	0	2	1
3	3	4	0	2	2
4	0	2	0	1	2
5	3	2	1	2	2
6	5	2	0	1	0

表1中的属性值v表示一个以v为中心的属性值邻域$n(v)$的缩写.设$B=\{a,b\}\subseteq\{a,b,c,d,e\}=A$,则$(a,n(5))\wedge(b,n(4))\wedge\ldots\wedge(a,n(5))\wedge(b,n(2))$是B-邻域基本公式;$(a,n(5))\wedge(b,n(4))\vee(a,n(3))\wedge(b,n(4))\vee\ldots\vee(a,n(5))\wedge(b,n(2))$被称作B邻域基本公式的析取范式,缩写成$\vee(B)$.

定义4. 设$NS=(U,A)$是属性邻域值决策表,$u_i,u_j\in U$是任意两个不同的个体,关于属性$a\in A$,有$a(u_i)\neq a(u_j)$,也就是说,$a(u_i)=v'_i\wedge v'_i\in n(v_i)\wedge v'_i\notin n(v_j)\wedge v'_j\notin n(v_i)\wedge v'_j=a(u_j)\wedge v'_j\in n(v_j)$,它们是分明的.如此的属性$a$的全体构成的合取范式被称作邻域分明合取范式;邻域决策表上的个体关于属性值是邻域分明的全体属性构成的邻域分明合取范式被称作邻域分明全合取范式.例如,

$((a,\langle n(5),n(3)\rangle)_{12}\vee(d,\langle n(1),n(2)\rangle)_{12}\vee(e,\langle n(0),n(1)\rangle)_{12})\wedge((a,\langle n(5),n(3)\rangle)_{13}\vee(d,\langle n(1),n(2)\rangle)_{13}\vee(e,\langle n(0),n(2)\rangle)_{13})\wedge((a,\langle n(5),n(0)\rangle)_{14}\vee(b,\langle n(4),n(2)\rangle)_{14}\vee(e,\langle n(0),n(2)\rangle)_{14})\wedge((a,\langle n(5),n(3)\rangle)_{15}\vee(b,\langle n(4),n(2)\rangle)_{15}\vee(c,\langle n(0),n(1)\rangle)_{15}\vee(d,\langle n(1),n(2)\rangle)_{15})\vee(e,\langle n(0),n(2)\rangle)_{15}\wedge(b,\langle n(4),n(2)\rangle)_{16}$

是NS中的一个邻域分明合取范式.

邻域决策表$NS=(U,A)$上的个体关于属性值的邻域是分明的全体属性构成的邻域分明全合取范式,并经引用吸收律和合并同类项等逻辑演算而得到一个简化的邻域分明合取范式.如,

$$(e,\langle n(1),n(2)\rangle)_{23}\wedge(b,\langle n(4),n(2)\rangle)_{16}\wedge((a,\langle n(3),n(5)\rangle)_{56}\vee(c,\langle n(1),n(0)\rangle)_{56}\vee(d,\langle n(2),n(1)\rangle)_{56})$$

是表1中提取的邻域分明全合取范式并经逻辑演算后得到的一个约简.如果$\wedge$对$\vee$作分配律运算并展开后,其每个析取项就是邻域值决策表的约简之一,所以有如下的定理.

定理1. 从邻域决策表提取的邻域值分明合取范式并经约简后,如果出现单个属性为一合取项,则这些属性构成的集合是核集.而不包含核的多于一个属性的合取项,其中每一个可能都是可省的.

通过表1可以生成一个邻域分明的合取范式,而且可以一边生成一边化简,直到所有邻域值可分明的相关属性都构成了邻域分明全合取范式为止,然后将化简后的合取范式等价地变换成一个析取范式.这个算法描述如下:

Procedure ColRed (W);

```
Begin
  W←{};
  For i←1 to n−1 do
   Begin
     For j←i+1 to n do
       W←W∪{a:a(i)≠a(j),直到A中所有属性都检查完};
     W←W   (*Simplification via absorbable law and other logical operations*);
   End;
  W←W (*via distributive law operations of ∧ to ∨*);
End.
```

以表1为例执行该算法将是如下过程：

u_1与u_2,u_3,u_4,u_5,u_6关于属性值的邻域是分明的邻域分明合取范式：

$$(a\vee d\vee e)\wedge(a\vee d\vee e)\wedge(a\vee b\vee e)\wedge(a\vee b\vee c\vee d\vee e)\wedge b=(a\vee d\vee e)\wedge b. \tag{1}$$

u_2与u_3,u_4,u_5,u_6关于属性的邻域值是分明的邻域分明合取范式并加上式(1)：

$$(a\vee d\vee e)\wedge b\vee e\wedge(a\vee b\vee d\vee e)\wedge(b\vee c\vee e)\wedge(a\vee b\vee d\vee e)=b\wedge e. \tag{2}$$

u_3与u_4,u_5,u_6关于属性的邻域值是分明的邻域分明合取范式并加上式(2)：

$$b\wedge e\wedge(a\vee b\vee d)\wedge(b\vee c)\wedge(a\vee b\vee d\vee e)=b\wedge e. \tag{3}$$

u_4与u_5,u_6关于属性的邻域值是分明的邻域分明合取范式并加上式(3)：

$$b\wedge e\wedge(a\vee c\vee d)\wedge(a\vee e)=b\wedge e\wedge(a\vee c\vee d). \tag{4}$$

u_5与u_6关于属性的邻域值是分明的邻域分明合取范式并加上式(4)：

$$b\wedge e\wedge(a\vee c\vee d)\wedge(a\vee c\vee d\vee e)=b\wedge e\wedge(a\vee c\vee d). \tag{5}$$

对式(5)施行∨对∧的分配律运算，得到如下的邻域分明析取范式：

$$(a\wedge b\wedge e)\vee(b\wedge c\wedge e)\vee(b\wedge d\wedge e),$$

由此得到3个约简：$Red(a,b,e)$，$Red(b,c,e)$和$Red(b,d,e)$.

定理2. 设$\varphi\rightarrow\psi$是邻域值决策表上的一条决策规则，属性值的邻域$n(v)\subseteq V$是可约简的，当且仅当$(\models\approx\varphi\rightarrow\psi)\rightarrow(\models\approx\varphi-\{(a,n(v))\}\rightarrow\psi)$，其中$\varphi$和$\psi$均为RLL中的公式.

定理2实际上是邻域值决策表约简的定义[1]，可作为判断一属性值的邻域$n(v)\in V$是否可约简的准则.

3 结束语

这种Rough逻辑演算的数据约简方法与文献[5]在逻辑演算的形式上似乎是相似的，但含义却不同. 属性的邻域值决策表上的个体关于属性的邻域值可以是整型数的集合，也可以是实型数的集合. 因此，Rough逻辑演算要求的条件比较宽松，是一种非单调的近似推理，所以用$\approx$作为公式φ取真符号，而不是文献[5]中所采用的经典逻辑演算. 它要求的条件比较强，是一种绝对精确的单调推理. 所以，本文提供的Rough逻辑演算方法的数据约简既是文献[5]中意义上的拓广，又是文献[5]中操作上的简化. 这在时间和空间上都得到了精简，从而加快了运行速度，提高了效率.

References:

[1] Pawlak, Z. Rough Sets, Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Acadimic Publishers Group, 1991. 72~80.

[2] Liu, Q. The OI-resolution of operator rough logic. LNAI 1424, Berlin: Springer-Verlag, 1998. 434~439.
[3] Lin, T. Y., Liu, Q., Yao, Y. Y. Logic systems for approximate reasoning: via rough sets and topology. LNAI 869, Verlin: Springer-Verlag, 1994. 65~74.
[4] Lin, T. Y., Liu, Q. First-Order rough logic I: approximate reasoning via rough sets. Fundamenta Informaticae, 1996, 27 (2,3): 137~154.
[5] Skowron, A., Suraj, Z. Discovery of concurrent data models from experimental data tables: a rough set approach. Ics Research Report, Institute of Computer Science, Warsaw University of Technology, 1995.

Rough Logic and Its Applications in Data Reduction *

LIU Qing, LIU Shao-hui, ZHENG Fei

(*Department of Computer Science and Engineering, Nanchang University, Nanchang* 330029, *China*)
E-mail: qliu@263.net

Abstract: The rough logic defined on neighbor-valued decision tables and its truth values of the formulas are discussed in this paper. It is more general than the decision logic defined by Pawlak in the applications of data reduction. At present, the methods used are often Pawlak's data analysis and Skowron's discernible matrix methods. The former is informal, no ease mechanization. The latter is intuitive, easy to understand, but it requires to generate a medial link of discernible matrix, to make unnecessary expenses on time and space. Therefore, in the paper, one side extracts the attributes of attribute neighbor-valued discernible from the neighbor-valued decision table and discernible Conjunctive Normal Form is constructed. The other side simplifies the formula to use absorbable laws and other calculus of logical formulas. It obtains directly all reductions in the neighbor-valued decision table. Since it doesn't need to generate the medial link of discernible matrix, so it can spare space and time, and raise the efficiency of the program run. Thus, reduction of the tables is handled to possess 6 attributes (4 conditional attributes and 2 decision attributes) and 102 objects to use two methods respectively, and to obtain the same results. It uses one side to extract formulas from the tables, and the other side to reduce the formulas in DELPHI 3.0 on P II 233/64 M. The time of program running is about 1 minute 54 seconds; while time of spending is about 1 minute 55 seconds to use the discernible matrix method. Due to the increase of an array (discernible matrix), its space degree of complexity is $O(m \times n^2)$, where m is the number of attributes, n is the number of objects. So, the space and time occupied will also increase rapidly along with the increment of attributes and objects. The strong points and shortcomings of two methods are quite clear from space and time used.

Key words: rough logic; neighbor-valued decision table; calculus of logical formula; data reduction

* Received October 22, 1999; accepted January 21, 2000
Supported by the National Natural Science Foundation of China under Grant No. 69773001; the Natural Science Foundation of Jiangxi Province of China