第 14卷第 4期 模 糊 系 统 与 数 学 Vol. 14, No. 4
2000年 12月 Fuzzy Systems and Mathematics Dec., 2000

文章编号: 1001-7402(2000) 04-0001-12

粗糙集理论介绍和研究综述*

张文修,吴伟志

(西安交通大学 理学院 信息与系统科学研究所,陕西 西安 710049)

摘 要: 介绍粗糙集理论的基本原理,粗糙集理论中的知识表示以及目前粗糙集理论的研究状况。

关键词: 粗糙集;模糊集;近似空间;近似算子;信息系统

中图分类号: TP18 文献标识码: A

1 引言

粗糙集作为一种处理不精确、不确定与不完全数据的新的数学理论,最初是由波兰数学家 Z. Pawlak[24]于 1982年提出的。由于最初关于粗糙集理论的研究大部分是用波兰语发表的,因此当时没有引起国际计算机学界和数学界的重视,研究地域也仅局限在东欧一些国家,直到 20世纪 80年代末才逐渐引起各国学者的注意。近几年来,由于它在机器学习与知识发现[6, 25]、数据挖掘[4, 18]、决策支持与分析[19, 29, 30]等方面的广泛应用,研究逐渐趋热。1992年,第一届关于粗糙集理论国际学术会议在波兰召开。1995年,ACM Communication将其列为新浮现的计算机科学的研究课题。1998年,国际信息科学杂志(Information Sciences)还为粗糙集理论的研究出了一期专辑。

粗糙集理论是建立在分类机制的基础上的,它将分类理解为在特定空间上的等价关系,而等价关系构成了对该空间的划分。粗糙集理论将知识理解为对数据的划分,每一被划分的集合称为概念。粗糙集理论的主要思想是利用已知的知识库,将不精确或不确定的知识用已知的知识库中的知识来(近似)刻画。该理论与其他处理不确定和不精确问题理论的最显著的区别是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,所以对问题的不确定性的描述或处理可以说是比较客观的,由于这个理论未能包含处理不精确或不确定原始数据的机制,所以这个理论与概率论,模糊数学和证据理论等其他处理不确定或不精确问题的理论有很强的互补性。

本文首先介绍粗糙集理论的基本概念,粗糙集理论中的知识表示,然后分析了该理论

---

* 本文系编辑与出版工作委员会、教育与普及工作委员会联合特约专稿。
收稿日期: 2000-06-20

作者简介: 张文修(1940-),男,山西翼城人,西安交通大学研究生院院长,教授,博士生导师,中国数学会常务理事,陕西省数学会理事长,研究方向: 应用概率与集值随机过程,计算机智能推理与计算机仿真,遗传算法等;吴伟志(1964-),男,浙江舟山人,浙江海洋学院数学系副主任,副教授,西安交通大学理学院信息与系统科学研究所博士研究生,研究方向: 粗糙集理论,集值分析与随机集等。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://w

的研究现状.

2 粗糙集理论的基本概念

设 U 是非空有限论域, R 是 U 上的二元等价关系, R 称为不可分辨关系,序对 $A=(U,R)$ 称为近似空间. $\forall (x,y)\in U\times U$,若 $(x,y)\in R$,则称对象 x 与 y 在近似空间 A 中是不可分辨的. U/R 是 U 上由 R 生成的等价类全体,它构成了 U 的一个划分. 可以证明, U 上划分可以与 U 上的二元等价关系之间建立一一对应. U/R 中的集合称为基本集或原子集. 若将 U 中的集合称为概念或表示知识,则 $A=(U,R)$ 称为知识库,原子集表示基本概念或知识模块. 任意有限的基本集的并和空集均称为可定义集,否则称为不可定义的. 可定义集也称为精确集,它可以在知识库中被精确地定义或描述,可表示已知的知识. 可以验证所有可定义集全体可构成 U 上的一个拓扑.

粗糙集理论的成功应用必须基于对其理论所包含的各种概念的比较清晰的理解.

对于论域 U 上任意一个子集 X, X 不一定能用知识库中的知识来精确地描述,即 X 可能为不可定义集,这时就用 X 关于 A 的一对下近似 $\underline{apr}X$ 和上近似 $\overline{apr}X$ 来"近似"地描述,其定义如下:

$$\underline{apr}X=\bigcup\{[x]\mid [x]\subseteq X\}=\{x\in U\mid [x]\subseteq X\}$$

$$\overline{apr}X=\bigcup\{[x]\mid [x]\cap X\neq\varnothing\}=\{x\in U\mid [x]\cap X\neq\varnothing\}$$

其中 $[x]$ 是 x 所在的 R 等价类.

下近似 $\underline{apr}$ 也称作 X 关于 A 的正域,记作 $POS(X)$,它可以解释为由那些根据现有知识判断出肯定属于 X 的对象所组成的最大集合,上近似 $\overline{apr}$ 可以解释为由那些根据现有知识判断出可能属于 X 的对象所组成的最小集合. $U\backslash\overline{apr}X$ 称作 X 关于 A 的负域,记作 $NEG(X)$,可以解释为由那些根据现有知识判断出肯定不属于 X 的对象所组成的集合. $\overline{apr}\backslash\underline{apr}X$ 称作 X 的边界(域),记作 $BN(X)$,它可以解释为由那些根据现有知识判断出可能属于 X 但不能完全肯定是否一定属于 X 的对象中所组成的集合.

从上面的定义可以看出,下近似 $\underline{apr}$ 是 A 中含在 X 中的最大可定义集,而上近似 $\overline{apr}$ 是 A 中包含 X 的最小可定义集. 因此, X 是可定义的当且仅当 $\underline{apr}=\overline{apr}$; X 是不可定义的当且仅当 $\underline{apr}\neq\overline{apr}$,这时称 X 是粗糙集. 称 $(2^U,\cap,\cup,\sim,\underline{apr},\overline{apr})$ 为粗糙集代数系统,其中 $\sim$ 表示集合补.

X 关于 A 的近似质量定义为

$$r_A(X)=\frac{|\underline{apr}X|}{|U|}$$

其中 $|X|$ 表示集合 X 的基数. 近似质量反映了知识 X 中肯定在知识库中的部分在现有知识中的百分比.

X 关于 A 的粗糙性测度定义为

$$d_A(X)=1-\frac{|\underline{apr}X|}{|\overline{apr}X|}$$

显然, $0\leqslant d_A(X)\leqslant 1$, X 是可定义的当且仅当 $d_A(X)=0$, X 是粗糙的当且仅当 $d_A(X)>0$. 粗糙性测度反映了知识的不完全程度.

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://w

称 $T_A(X)=\frac{|\underline{apr}X|}{|\overline{apr}X|}$为 X 关于 A的近似精度,近似精度反映了根据现有知识对 X的了解程度。

粗糙集理论还对于集合类关于近似空间定义了下近似和上近似。设 $F=\{X_1,X_2,\cdots,X_n\}$是由 U的子集所构成的集类,则 F 关于近似空间 A的下近似 $\underline{apr}F$和上近似 $\overline{apr}F$定义为

$$\underline{apr}F=\{\underline{apr}X_1,\underline{apr}X_2,\cdots,\underline{apr}X_n\}$$
$$\overline{apr}F=\{\overline{apr}X_1,\overline{apr}X_2,\cdots,\overline{apr}X_n\}$$

F 关于 A的近似精度 $T_A(F)$和近似质量 $r_A(F)$分别定义为

$$T_A(F)=\frac{\sum_{i=1}^{n}|\underline{apr}X_i|}{\sum_{i=1}^{n}|\overline{apr}X_i|}$$
$$r_A(F)=\frac{\sum_{i=1}^{n}|\underline{apr}X_i|}{|U|}$$

当 F也是 U的划分时,F关于 A的近似在判别一个决策表是否是协调的和规则提取中有重要应用。

3　粗糙集理论中的知识表示

粗糙集理论中的知识表达方式一般采用信息表或称为信息系统的形式,它可以表示为四元有序组 $K=(U,A,V,d)$,其中

- U是对象的全体,即论域;
- A是属性全体;
- $V=\bigcup_{a\in A}V_a$,V_a 是属性的值域;
- $d_x: U\times A\to V$是一个信息函数,$d_x: A\to V$,$x\in U$,反映了对象 x在 K中的完全信息,其中 $d_x(a)=d(x,a)$。

对于这样的信息系统,每个属性子集就定义了论域上的一个等价关系,即 $\forall B\subseteq A$,定义 R_B:

$$xR_By\Leftrightarrow d_x(b)=d_y(b),\ \forall b\in B$$

由此可见,信息系统类似于关系数据库模型的表达方式。有时属性集 A还分为条件属性 C和决策(结论)属性 D,这时的信息系统称为决策表,常记为 $(U,C\cup D,V,d)$。无决策的数据分析和有决策的数据分析是粗糙集理论在数据分析中的两个主要应用。

粗糙集理论给出了对知识(或数据)的约简和求核的方法从而提供了从信息系统中分析多余属性的能力。

设 $K=(U,AT,V,d)$是一个信息系统,我们记由属性集 $B\subseteq AT$所导出的等价关系为 R_B。$\forall a\in AT$,若 $R_{AT}=R_{AT\setminus\{a\}}$,则称属性 a是多余的;若在系统中没有多余属性,则称 AT是独立的;子集 $B\subseteq AT$称为是 AT的约简,常记作 $red(AT)$,若 $R_B=R_{AT}$且 B中没有多余

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://v

属性; AT 的所有约简的交集称为 AT 的核,记作 $core(AT)$. 一般属性的约简不唯一而核是唯一的.

例 3.1(无决策情形) $S=(U,A,V,d)$,其中 $U=\{x_1,x_2,\cdots,x_8\}$,属性集 $A=\{c_1,c_2,c_3,c_4\}$, $V_1=V_2=V_3=\{1,2,3\}$, $V_4=\{1,2\}$,信息函数 d 见表 3.1.

表 3.1 一个信息系统

U	c_1	c_2	c_3	c_4
x_1	1	1	1	1
x_2	1	2	2	1
x_3	1	1	1	1
x_4	1	2	2	1
x_5	2	2	1	1
x_6	2	2	1	1
x_7	3	3	3	2
x_8	3	3	3	2

显然

$$U/c_1=\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6\},\{x_7,x_8\}\}$$
$$U/c_2=\{\{x_1,x_3\},\{x_2,x_4,x_5,x_6\},\{x_7,x_8\}\}$$
$$U/c_3=\{\{x_1,x_3,x_5,x_6\},\{x_2,x_4\},\{x_7,x_8\}\}$$
$$U/c_4=\{\{x_1,x_2,x_3,x_4,x_5,x_6\},\{x_7,x_8\}\}$$
$$U/C=\{\{x_1,x_3\},\{x_2,x_4\},\{x_5,x_6\},\{x_7,x_8\}\}$$

将对象及其信息压缩后得表 3.2.

表 3.2 压缩后的信息表

U/C	c_1	c_2	c_3	c_4
$\{x_1,x_3\}$	1	1	1	1
$\{x_2,x_4\}$	1	2	2	1
$\{x_5,x_6\}$	2	2	1	1
$\{x_7,x_8\}$	3	3	3	2

表 3.2 可以简明地表示为表 3.3.

可以验证,信息表 3.1(或表 3.2)中属性 c_4 是多余属性,而且可以计算此信息表有三个最简属性约简, $\{c_1,c_2\}$, $\{c_1,c_3\}$ 和 $\{c_2,c_3\}$,从而可得信息系统的三个最简约简表 3.4 表 3.5 和表 3.6.

表 3.3 最简压缩表

c_1	c_2	c_3	c_4
1	1	1	1
1	2	2	1
2	2	1	1
3	3	3	2

表 3.4 约简表

c_1	c_2
1	1
1	2
2	2
3	3

表 3.5 约简表

c_1	c_3
1	1
1	2
2	1
3	3

表 3.6 约简表

c_2	c_3
1	1
2	2
2	1
3	3

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://v

粗糙集理论除了给出了对知识(或数据)的约简和求核的方法外,还提供了从决策表中抽取规则的能力,机器学习和从数据库中的机器发现就是基于这个能力。这个方法就可以做到在保持决策一致的条件下将多余属性删除。

在一个决策表$(U, C\cup D, V, f)$中,若$\forall X\in U/D_1$, X关于由C_1导出的近似空间的下近似和上近似相等,即$\underline{apr}_{C_1}X=\overline{apr}_{C_1}X$,则称条件属性子集$C_1\subseteq C$关于决策属性$D_1\subseteq D$是协调的,这时也称决策表$(U, C\cup D_1, V, f)$是协调的,否则为不协调。如果用包含度理论[48]来解释,则决策表$(U, C\cup D_1, V, f)$是协调的当且仅当包含度

$$D(D_1/C_1)=1$$

其中$D(D_1/C_1)=\dfrac{|\underline{apr}_{C_1}(U/D_1)|}{|\overline{apr}_{C_1}(U/D_1)|}$。

从协调的决策表中可以抽出确定性规则;而从不协调的决策表中只能抽出不确定性的规则或可能性规则有时也称为广义决策规则,这是因为在不协调的系统中存在着矛盾的事例。

决策表中的决策规则一般可以表示为形式

$$\wedge\ (c, v)\rightarrow\vee\ (d, w)$$

其中$c\in C, v\in V_c, w\in V_d$. $\wedge\ (c, v)$称为规则的条件部分,而$\vee\ (d, w)$称为规则的决策部分。决策规则即使是最优的也不一定唯一。

在决策表中抽取规则的一般方法为:

(1) 在决策表中将信息相同(即具有相同描述)的对象及其信息删除只留其中一个得到压缩后的信息表,即删除多余事例;

(2) 删除多余的属性;

(3) 对每一个对象及其信息中将多余的属性值删除;

(4) 求出最小约简;

(5) 根据最小约简,求出逻辑规则。

例 **3. 2**(有决策情形)　表 3. 7给出的信息系统是一个决策表,其中$C=\{c_1, c_2, c_3, c_4\}$是条件属性,$D=\{d_1, d_2\}$是决策属性。

表 **3. 7**　一个决策表

U	c_1	c_2	c_3	c_4	d_1	d_2
x_1	1	1	1	1	1	1
x_2	1	2	2	1	2	2
x_3	1	1	1	1	1	3
x_4	1	2	2	1	2	4
x_5	2	2	1	1	3	5
x_6	2	2	1	1	3	5
x_7	3	3	3	2	4	5
x_8	3	3	3	2	4	5

对于由例 3. 2给出的决策子表$(U, C\cup\{d_1\}, V, f)$和$(U, C\cup\{d_2\}, V, f)$,我们可分别得到它们的一个约简表 3. 8和表 3. 9(一般不唯一)。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://v

表 3.8 约简表

c_1	c_3	d_1
1	1	1
1	2	2
2	1	3
3	3	4

表 3.9 约简表

c_1	c_2	d_2
1	1	1
1	2	2
1	1	3
1	2	4
2	2	5
3	3	5

可以验证,例 3.2给出的决策表中,子表 $(U, C\cup\{d_1\}, V, d)$是协调的,而子表 $(U, C\cup\{d_2\}, V, d)$是不协调的。表 3.8是协调的,并且可以得到决策表 $(U, C\cup\{d_1\}, V, d)$的四条最优决策规则,并且这四条规则都是确定的:

$$r_1:\ (c_1,1)\wedge(c_3,1)\rightarrow(d_1,1)$$
$$r_2:\ (c_1,1)\wedge(c_3,2)\rightarrow(d_1,2)$$
$$r_3:\ (c_1,2)\rightarrow(d_1,3)$$
$$r_4:\ (c_1,3)\rightarrow(d_1,4)$$

表 3.9是不协调的,也可以得到决策表 $(U, C\cup\{d_2\}, V, d)$的四条最优决策规则,但这四条规则中 r'_1和 r'_2是不确定的,而只有 r'_3和 r'_4是确定的:

$$r'_1:\ (c_1,1)\wedge(c_2,1)\rightarrow(d_2,1)\vee(d_2,3)$$
$$r'_2:\ (c_1,1)\wedge(c_2,2)\rightarrow(d_1,2)\vee(d_2,4)$$
$$r'_3:\ (c_1,2)\rightarrow(d_2,5)$$
$$r'_4:\ (c_1,3)\rightarrow(d_2,5)$$

4 粗糙集的理论研究

粗糙集理论的研究由于其历史较短,所以至今为止,对粗糙集的概念的定义还没有完全统一,一种就是原始的 Pawlak[24]意义下的,也有由上,下近似构成的一对集合来命名的[5],还有以下近似和上近似构成的区间集(集合类)来定义的[37],定义观点的不同往往带来研究的侧重面的不同。目前,对粗糙集理论的研究主要集中在:粗糙集的模型的推广,问题的不确定性的研究,与其他处理不确定性,模糊性问题的数学理论的关系与互补,纯粹的数学理论方面的研究,粗糙集的算法研究和人工智能其他方向关系的研究等。这些研究有的是受应用的推动而产生,有的是纯理论的。

4.1 粗糙集模型的推广

Pawlak 粗糙集模型的推广一直是粗糙集理论研究的主流方向,目前主要有两种方法:(1)构造性方法;(2)代数性(公理化)方法。

(1)构造性方法是对原始 Pawlak 粗糙集模型的一般推广,其主要思路是从给定的近似空间出发去研究粗糙集和近似算子。它是以论域上的二元关系或布尔子代数作为基本要素的,然后导出粗糙集代数系统 $(2^U, \sim, \cup, \cap, \underline{apr}, \overline{apr})$。这种方法所研究的问题往往来源于实际,所建立的模型有很强的应用价值,其主要缺点是不易深刻了解近似算子的代

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://

数结构。

在 Pawlak粗糙集模型中有三个最基本的要素: 一个论域 U, U上的一个二元等价关系 R(或划分) (它们构成了近似空间) ,一个被近似描述的 (经典)集合 X,也称为专家概念。这样,推广的形式主要也有三个方向,即从论域方向、从关系方向 (包括近似空间)和从集合方向。

· 从论域方向推广的目前只有一种,就是双论域的情形[40],当然这时的二元关系就变成为两个论域迪卡尔乘积的一个子集。对于将论域推广到多个的情形来研究粗糙集理论的文献我们没有见到。

· 关系的推广: 一种是将论域上的二元等价关系推广成为任意的二元关系得到了一般关系下的粗糙集模型[39];另一种是将对象 x 所在的等价类看成是 x 的一个领域,从而推广导出了基于领域算子的粗糙集模型[37];也有将由关系导出的划分推广成为一般的布尔子代数的,以此出发去定义粗糙集和近似算子的[23];更一般的有将普通关系推广成模糊关系或模糊划分[3, 10, 15, 20]而获得模糊粗糙集模型。

· 将集合和近似空间进行推广。这一类的推广是与其他处理不确定,不精确或模糊的知识 (如概率论,模糊数学,信息论,证据理论等)结合起来进行研究的。

当知识库中的知识是由于随机原因或经统计得到的,即知识库中的知识很可能是不确定的,很多学者提出了统计 (或概率)粗糙集模型[28, 35, 41],变精度粗糙集模型[43]实质上也可以归入这类模型,寻求具有最小风险的 Bayes决策问题也可转化为这类模型[35]。这一类模型在数据分析的增量式机器学习中有重要应用[47]。目前见到的此类模型中,近似空间中二元关系大都是等价关系,对于非等价关系给出的情形文章的尚没见到,我们提出了基于随机集的粗糙集模型[50]作为一种尝试,既是对基于领域算子的粗糙集模型的推广,又适用于双论域情形,同时也是对统计粗糙集模型的推广。我们认为在统计粗糙集模型和变精度粗糙集模型中,近似逼近好坏的本质是张文修提出的包含度[48]的大小,因此我们认为粗糙集理论与包含度理论的关系非常密切。

当知识库中的知识模块都是清晰概念,而被描述的概念是一个模糊概念,人们建立了粗糙模糊集模型[1]来解决此类问题的近似推理。当知识库中的知识模块也是模糊的,有些学者就提出了模糊粗糙集模型[10, 11, 23]。对于知识库中的知识模块既是模糊知识又是随机得到的至今未见论及,但现实问题肯定存在的,因此也是值得研究的。

(2)代数方法也称为公理化方法有时也称为算子方法,这种方法不是以二元关系为基本要素,它的基本要素是一对满足某些公理的一元 (集合)近似算子 L, H: $2^U \to 2^U$,即粗糙代数系统 $(2^U, \sim, \cup, \cap, L, H)$中近似算子是事先给定的。这种方法研究的明显优点是能够深刻地了解近似算子的代数结构,其缺点是应用性不够强。

近似算子的某些公理能保证有一些特殊类型的二元关系的存在,使这些关系能够通过构造性方法产生给定的算子;反过来,由二元关系通过构造性方法导出的近似算子一定满足某些公理,使这些公理通过代数方法产生给定的二元关系。

公理化方法的研究一开始只局限于 Pawlak粗糙代数系统,即公理与二元等价关系相对应情形,后逐渐发展到一般关系下的粗糙集系统[38, 42]。至今为止,关于公理化方法的粗糙集理论研究大多局限于经典集情形,对于模糊集情形虽有讨论[20],但比较少。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://v

4.2 不确定性问题的理论研究

粗糙集理论中知识的不确定性主要由两个原因产生的: 一个原因是直接来自于论域上的二元关系及其产生的知识模块,即近似空间本身,如果二元等价关系产生的每一个等价类中只有一个元素,那么等价关系产生的划分不含有任何信息。划分越粗,每一个知识模块越大,知识库中的知识越粗糙,相对于近似空间的概念和知识就越不确定,这时处理知识的不确定性的方法往往用香农信息熵来刻画,知识的粗糙性与信息熵的关系比较密切,知识的粗糙性实质上是其所含信息多少的更深层次的刻画[45]。单从这个角度来看,粗糙集理论与信息论的关系就比较密切,不少学者在这方面做了研究工作[2, 45, 46]。

粗糙集理论中知识不确定性的另一个原因来自于给定论域里粗糙近似的边界,当边界为空集时知识是完全确定的,边界越大知识就越粗糙或越模糊。至今,粗糙集理论刻画概念 X 的不确定性用正则条件熵 $H_0(X^* \mid R^*)$(其中 $X^* = \{X, \sim X\}$是由 X 产生的一个划分,$R^* = \{X_1, X_2, \cdots, X_n\}$是由 R 产生的一个划分)[28]和粗糙性测度 $d_4(X)$来实现的。但是这两个度量并没有提供那些完全属于 X 的下近似的区域里面与不可分辨关系的知识粒度有关的不确定性,于是有人引进了粗糙熵 $E_r(X)$的概念来刻画概念 X 的不确定性[2]。

寻求一个合适的度量来刻画知识的不确定性也是粗糙集理论研究的一个重要方向。

4.3 与其他处理不确定性方法的理论的研究

在粗糙集理论与其他处理模糊性或不确定性方法的理论研究中,主要集中在它与概率统计、模糊数学、D-S证据理论和信息论的相互渗透与补充。

在信息系统中,知识库的知识的类型一般有两类: 一类库中所有对象的描述是完全已知的, Pawlak粗糙集模型和一般二元关系下的粗糙集模型就是属于这一种;另一类库中的对象的描述只有部分是已知的,即知识库中的知识是不确定的,它只能通过训练样本所提供的信息来刻画概念,为了使从训练样本获得的规则符合整个论域的对象,在抽取样本时应符合统计规律性,粗糙集理论不管这一类工作,因此概率统计作为研究自然界,人类社会及技术过程中大量随机现象的规律性的一门学科,它与粗糙集理论的结合就显得非常自然。

粗糙集理论用粗糙隶属函数来刻画知识的模糊性。对于只建立在一般二元关系 R 下的近似空间 $A = (U, R)$,粗糙隶属函数为

$$\mu_X(x) = \frac{|X \cap R_s(x)|}{|R_s(x)|}$$

其中当 R 是等价关系时, $R_s(x) = [x]$,在概率近似空间下,粗糙隶属函数为

$$\mu_X(x) = \frac{P(X \cap R_s(x))}{P(R_s(x))}$$

粗糙隶属函数一般不是 Zadeh意义下的隶属函数[38, 41]。

模糊集和粗糙集理论在处理不确定性和不精确性问题方面都推广了经典集合论。虽有一定的相容性和相似性,然而它们的侧重面不同。从知识的"粒度"的描述上来看,模糊集是通过对象关于集合的隶属程度来近似描述的,而粗糙集是通过一个集合关于某个可利用的知识库的一对上、下近似来描述的;从集合对象间的关系来看,模糊集强调的是集

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://v

合边界的病态定义上的,即边界的不分明性,而粗糙集强调的是对象间的不可分辨性;从研究的对象来看,模糊集研究的是属于同一类的不同对象间的隶属关系,重在隶属程度,而粗糙集研究的是不同类中的对象组成的集合关系,重在分类。虽然模糊集的隶属函数和粗糙集的粗糙隶属函数都反映了概念的模糊性,直观上有一定的相似性,但是模糊集的隶属函数大多是专家凭经验给出的,因此往往带有很强烈的主观意志,而粗糙集的粗糙隶属函数的计算是从被分析的数据中直接获得的,非常客观。也正因为如此,将粗糙集理论和模糊集理论进行某些"整合"后去描述知识的不确定性和不精确性比它们各自描述知识的不确定性和不精确性可望显示出更强的功能。目前所见的模糊粗糙集模型[10, 11, 23]是其中的一些成功范例。

由建立在二元连通关系下的上近似和下近似的概率分别是可能性测度和必然性测度与通常的正则模糊集产生的可能性测度和必然性测度所反映的观点是不同的,因为前者是直接从给定的数据上导出的,用它来处理知识的模糊性相当客观,这与基本的模糊集的隶属度是凭经验或由领域专家带有明显的主观意志给出的有本质的差异。

粗糙集理论与 D- S证据理论在处理不确定性的问题方面其产生和研究的方法是不同的,但却有某种相容性,粗糙集理论是为开发规则的机器自动生成而提出的,而 D- S理论主要用于证据推理。粗糙集理论用概念的一对上,下近似对其进行描述,而 D- S证据理论是用一对信任函数和似然函数在给定证据下对假设进行估计和评价。粗糙集理论中的下近似和上近似的概率恰好分别是信任函数和似然函数[40, 50],然而生成信任函数和似然函数的基本概率分配函数(即 mass函数)方法是不同的,前者来自于系统中数据本身,比较客观,而后者往往来自于专家的经验,带有很强的主观性。粗糙集理论与与 D- S证据理论有很强的互补性。

4.4 算法研究

粗糙集理论中有效算法研究是粗糙集在人工智能方向上研究的一个主要方向。目前,粗糙集理论中有效算法研究主要集中在导出规则的增量式算法,约简的启发式算法,粗糙集基本并行算法[44],以及与粗糙集有关的神经网络与遗传算法等[13]。这些研究的成功应用有的已经获得了商业价值。

4.5 与其他数学理论的联系

对粗糙集理论的研究的不断深入,与其他数学分支的联系也更加紧密。例如,从算子的观点看粗糙集理论,与之关系较紧的有拓扑空间、数理逻辑、模态逻辑、格与布尔代数、算子代数等;从构造性和集合的观点来看,它与概率论、模糊数学、证据理论、图论、信息论等联系较为密切。粗糙集理论研究不但需要以这些理论作为基础,同时也相应地带动这些理论的发展。例如从算子的角度来看,粗糙集代数系统$(2^U, \cup, \cap, \sim, \underline{apr}, \overline{apr})$是普通布尔代数系统$(2^U, \cup, \cap, \sim)$加上两个一元集合算子 $\underline{apr}$ 和 $\overline{apr}$ 的推广。由于逻辑是计算机推理的基础,基于粗糙集的逻辑的研究也是粗糙集理论研究的比较活跃的一个方向。例如粗糙集代数系统$(2^U, \cup, \cap, \sim, \underline{apr}, \overline{apr})$中的五个集合算子恰好对应模态逻辑的五个算子,因此基于粗糙集的模态逻辑的研究显得特别活跃,各种模型的粗糙集代数系统恰好对应于各种模态逻辑系统[33, 38, 39],二者的结合有重要的应用,基于这种联系粗糙集理论能丰富模态逻辑理论,反之亦然。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://v

至今为止,就我们所知粗糙集理论研究与应用只限于对数据给出的知识问题的处理,对于由文本和连续图像问题的处理我们尚未见到,由于随机集理论在图像处理中已获得了成功的应用,因此我们认为对粗糙集理论与随机集理论的结合的进一步研究有望使它在图像处理上获得成功。

目前,纯粹的数学理论与粗糙集理论结合起来进行研究已有文章出现,并不断有新的数学概念出现,如"粗糙逻辑"[17,22]、"粗糙理想"、"粗糙半群"[16]等等。我们认为,随着粗糙结构与代数结构,拓扑结构,序结构等各种结构的不断整合,必将不断涌现出新的富有生机的数学分支。

5 结束语

粗糙集理论是波兰数学家 Z. Pawlak于 1982年提出的,由于种种原因一直没有被人们重视,直到 1990年前后,由于它在数据的决策与分析、机器学习、模式识别等计算机领域的成功应用,才逐渐被人们所重视。该理论与概率论,模糊数学,信息论和证据理论等其他处理不确定性和不精确性问题的理论有很强的互补性。粗糙集理论在人工智能的应用上主要有两大类:一类是无决策的分析,内容主要包括数据压缩、约简、聚类与机器发现等;另一类是有决策的分析,内容主要包括决策分析、规则提取等,当然也涉及对原始数据的预处理,如数据压缩与约简等。作为处理不确定性和不精确性问题的的一种数学工具,目前,粗糙集理论已在医学、生物学、化学、材料学、地理学和金融等其他学科得到了成功的应用。粗糙集的理论研究目前正成为信息科学的一个热点,有关这个理论与应用研究被 SCI收录的文章近几年来急剧上升,从被 SCI收录文章的作者的数字分布来看,我国在这方面的研究与世界先进水平有一定的差距。王珏等[44]曾于 1996年对粗糙集理论与应用作过综述,我们将这个理论目前的研究状况介绍给数学工作者,希望我国有更多的感兴趣的数学同行加入到这一领域的研究行列中来,以提高我国在这一领域的整体研究水平。

参考文献:

[1] Banerjee M, Pal S K. Roughness of a fuzzy set[J]. Information Sciences, 1996, 93: 235- 246.

[2] Beaubouef T, Petry F, Arora G. Information- theoretic measures of uncertainty for rough sets and rough relational databases[J]. Information Sciences, 1998, 109: 185- 195.

[3] Bodjanova S. Approximation of fuzzy concepts in decision making[J]. Fuzzy Sets and Systems, 1997, 85: 23- 29.

[4] Chan C C. A rough set approach to attribute generalization in data mining[J]. Journal of Information Sciences, 1998, 107: 169- 176.

[5] Chanas S, Kuchta D. Further remarks on the relation between rough and fuzzy sets[J]. Fuzzy Sets and Systems, 1992, 47: 391- 394.

[6] Chmielewski M R, Grzymala-Busse J W. Global discretization of continuous attributes as preprocessing for machine learning[J]. International Journal of Approximate Reasoning, 1996, 15: 319- 331.

[7] Dimitras A I, Slowinski R, Susmaga R, Zopounidis C. Business failure prediction using rough sets[J]. European Journal of Operational Research, 1999, 114: 263- 280.

[8] Coker D. Fuzzy rough sets are intuitionistic L-fuzzy sets[J]. Fuzzy Sets and Systems, 96: 381- 383.

[9] Dontsch I. Statistical evaluation of rough set dependency analysis[J]. Int. J. Human-Computer

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://

Studies, 1997, 46: 589- 604.

[10] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. Int. J. General Systems, 1990, 17: 191- 209.

[11] Dubois D, Prade H. Twofold fuzzy sets and rough sets- some issues in knowledge representation [J]. Fuzzy Sets and Systems, 1987, 23: 3- 18.

[12] Duntsch I. A logic for rough sets[J]. Theoretical Computer Sciences, 1997, 179: 427- 436.

[13] Jagielska I, Matthews C, Whitfort T. An investigation into the application of neural networks, fuzzy logic, genetic algorithms, and rough sets to automated knowledge acquisition for classification problems[J]. Neurocomputing, 1999, 24: 37- 54.

[14] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112: 39- 49.

[15] Kuncheva L I. Fuzzy rough sets: Application to feature selection[J]. Fuzzy Sets and Systems, 1992, 51: 147- 153.

[16] Kuroki N. Rough ideals in semigroups[J]. Information Sciences, 1997, 100: 139- 163.

[17] Lin T Y. A rough logic formalism for fuzzy controllers: A hard and soft computing view[J]. International Journal of Approximate Reasoning, 1996, 15: 395- 414.

[18] Lingras P J, Yao Y Y. Data mining using extensions of the rough set model[J]. Journal of the American Society for Information Science, 1998, 49(5): 415- 422.

[19] McSherry D. Knowledge discovery by inspection[J]. Decision Support Systems, 1997, 21: 43- 47.

[20] Morsi N N, Yakout M M. Axiomatics for fuzzy rough sets[J]. Fuzzy Sets and Systems, 1998, 100: 327- 342.

[21] Nakamura A, Gao J M. A logic for fuzzy data analysis[J]. Fuzzy Sets and Systems, 1991, 39: 127- 132.

[22] Nakamura A. A rough logic based on incomplete information and its application[J]. International Journal of Approximate Reasoning, 1996, 15: 367- 378.

[23] Nanda S, Majumdar S. Fuzzy rough sets[J]. Fuzzy Sets and Systems, 1992, 45: 157- 160.

[24] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341- 356.

[25] Pawlak Z. Rough sets: Theoretical Aspects of Reasoning about Data[A]. Boston: Kluwer Academic Publishers, 1991.

[26] Pawlak Z. Rough classification[J]. Int. J. Man-Machine Studies, 1984, 20: 469- 483.

[27] Pawlak Z. Rough set theory and its applications to data analysis[J]. Cybernetics and Systems: An International Journal, 1998, 29: 661- 688.

[28] Pawlak Z, Wong S K M, Ziarko W. Rough sets: probabilistic versus deterministic approach[J]. International Journal of Man-Machine Studies, 1988, 29: 81- 95.

[29] Pawlak Z. Rough set approach to Knowledge- based decision support[J]. European Journal of Operational Research, 1997, 99: 48- 57.

[30] Pomerol J C. Artificial intelligence and human decision making[J]. European Journal of Operational Research, 1997, 99: 3- 25.

[31] Salonen H, Nurmi G. A note on rough sets and common knowledge events[J]. European Journal of Operational Research, 1999, 112: 692- 695.

[32] Tsumoto S. Automated extraction of medical expert system rules from clinical databases basesd on

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://v

rough set theory[J]. Information Sciences, 1998, 112: 67- 84.
[33] Vakarelov D. A modal logic for similarity relations in Pawlak knowledge representation systems [J]. Fundamenta Informaticae, 1991, 15: 61- 79.
[34] Webster L, Chen J G, Tan S S, Watson C, Korwin A De. Validation of authentic reasoning expert systems[J]. Information Sciences, 1999, 117: 19- 46.
[35] Wong S K M, Ziarko W. Comparison of the probabilistic approximate classification and the fuzzy set model[J]. Fuzzy Sets and Systems, 1987, 21: 357- 362.
[36] Wygralak M. Rough sets and fuzzy sets- some remarks on interrelations[J]. Fuzzy Sets and Systems, 1989, 29: 241- 243.
[37] Yao Y Y. Relational interpretations of neighborhood operators and rough set approximation[J]. Information Sciences, 1998, 111: 239- 259.
[38] Yao Y Y. Two views of the theory of rough sets in finite universes[J]. International Journal of Approximate Reasoning, 1996, 15: 291- 317.
[39] Yao Y Y, Lin T Y. Gerneralization of rough sets using modal logic[J]. Intelligent Automation and Soft Computing, 1996, 2: 103- 120.
[40] Yao Y Y, Lingras P. Interpretations of belief functions in the theory of rough sets[J]. Information Sciences, 1998, 104: 81- 106.
[41] Yao Y Y. A decision theoretic framework for approximating concepts[J]. International Journal of Man-Machine Studies, 1992, 37: 793- 809.
[42] Yao Y Y. Constructive and algebraic methods of the theory of rough sets[J]. Information Sciences, 1998, 109: 21- 47.
[43] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46: 39- 59.
[44] 王珏,苗夺谦,周育健.关于 Rough Set理论与应用的综述[J].模式识别与人工智能, 1996, 9: 337 - 344.
[45] 苗夺谦,王珏.粗糙集理论中知识粗糙性与信息熵关系的讨论[J].模式识别与人工智能, 1998, 11: 34- 40.
[46] 苗夺谦,王珏.粗糙集理论中概念与运算的信息表示[J].软件学报, 1999, 10: 113- 116.
[47] 李莉.基于可变精度粗集模型的增量式归纳学习[J].计算机科学, 1999, 26: 55- 58.
[48] 张文修,梁怡.不确定推理原理[M].西安交通大学出版社, 1996.
[49] 施恩伟.粗糙集中不可分辨关系的某些性质[J].科学通报(英文辑), 1990, 35: 338- 341.
[50] 张文修,吴志伟.基于随机集的粗糙集模型(I)[J].西安交通大学学报, 2000, 34(12): 15- 19.

An Introduction and a Survey for the Studies of Rough Set Theory

ZHANG Wen-xiu, WU Wei-zhi
(Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract This paper introduces the basic ideas of rough set theory and the knowledge representation in rough set theory, then reviews the recent studies for this theory.

Key words Rough Sets; Fuzzy Sets; Approximation Spaces; Approximation Operators; Information System

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://v