基于粗糙集理论的属性约简与决策方法

何圣姿
抚州幼儿师范高等专科学校

摘　要：粗糙集理论是处理不完备、不确定知识的新方法，它可以通过对知识进行有效约简从而导出决策方法，论文结合实例展示基于粗糙集理论的属性约简与决策方法。

关键词：属性约简；区分函数；重要程度；决策方法
【DOI】10.12252/j.issn.2096-627X.2022.07.210

引言

信息爆炸时代产生了海量数据信息，其中不乏很多无效、干扰信息，各行各业若能从海量数据信息中提取有效信息，进行合理决策，对生产力的快速发展起到推动作用。当今，数据挖掘的方法有很多，而粗糙集理论是处理不完备、不确定知识的新方法，它可以通过对知识进行有效约简从而导出决策方法。

一、预备知识

定义1[1, 3]：四元组$S=(U,A,V,f)$是一个知识表达系统，其中，U：论域；A：属性的非空有限集合；V：属性的值域；$f:U\times A\to V$是一个信息函数。当A包含条件属性集（C）与决策属性集（D）时，称S为决策表。

定义2[1, 3]：$S=(U,A,V,f)$，$R\subseteq A$，不可区分等价关系为：

$$U/R=\{(x_i,x_j)\in U\times U:\forall a\in R,f(x_i,a)=f(x_j,a)\}。$$

定义3[1, 3]：$S=(U,A,V,f)$，$R\subseteq A$，$X\subseteq U$，X的下近似集为：

$\underline{R}X=\cup\{Y\in U/R|Y\subseteq X\}$；X的上近似集为：$\overline{\mathrm{R}}X=\cup\{Y\in U/R|Y\cap X\neq\phi\}$；

X的R边界域为：$bn_R(X)=\overline{\mathrm{R}}X-\underline{R}X$；X的R正域为：$pos_R(X)=\underline{R}X$；X的R负域为：$neg_R(X)=U-\overline{R}X$。

当$bn_R(X)=\phi$时，X为R可定义集，否则，X为R可粗糙集。

定义4[1, 3]：令C，D为U中的等价关系，D的C正域为：$pos_C(D)=\underset{X\in U/D}{\cup}\underline{C}X$；若$pos_{(C-\{a\})}(D)=pos_C(D)$，则称$a$为C中D不必要的，否则，$a$为C中D必要的。

定义5[1, 3]：设$a\subseteq C$，a为C的D约简（相对约简）当且仅当a是C的D独立子族且$pos_a(D)=pos_C(D)$。

定义6[1, 3]：$S=(U,A,V,f)$，$|U|=n$，S的区分矩阵为一个$n\times n$矩阵，其中任一元素为：

$a(x,y)=\{a\in A|f(x,a)\neq f(y,a)\}$，即$a(x,y)$为区别对象$x,y$的所有属性集合。区分函数为：

$$\Delta=\prod_{(x,y)\in U\times U}\sum a(x,y)。$$

定义7[1, 3]：决策表(U,A,V,F) $A=C\cup D$，决策属性D对条件属性C的依赖程度为：

$$r(C,D)=|\mathrm{pos}_C D|/|U|$$

定义8[1, 3]：决策表(U,A,V,F) $A=C\cup D$，属性a的重要性定义为：

$SGF(a,C,D)=r(C,D)-r(C-\{a\},D)$。

定义9[1, 3]：决策规则：$r_{ij}:des(X_i)\to des(Y_j),Y_j\cap X_i\neq\phi$，规则的确定因子为：

$$\mu_{ij}(X_i,Y_j)=|Y_j\cap X_i|/|X_i|,0<\mu_{ij}(X_i,Y_j)\leq 1,$$

当$\mu_{ij}(X_i,Y_j)=1$时，r_{ij}是确定的，当$0<\mu_{ij}(X_i,Y_j)<1$时，r_{ij}是不确定的。

二、应用实例

以上预备知识是粗糙集理论的核心内容，熟练掌握预备知识的本质便能应用粗糙集理论进行属性约简、聚类分析、数据挖掘等，但预备知识抽象难懂，下面结合实际决策表辅以解释并展示基于粗糙集理论的属性约简与决策方法。

例　表1给出了一个关于病人：$e_1,e_2,e_3,e_4,e_5,e_6,e_7,e_8$的决策表，其中展示了病人一些主要病症：“头痛”，“肌肉痛”，“体温”及诊断

结果。则

论域：$U=\{e_1,e_2,e_3,e_4,e_5,e_6,e_7,e_8\}$，

条件属性：$C=\{$头痛，肌肉痛，体温$\}=\{a,b,c\}$，

决策属性：$D=\{$流感$\}$.

U关于属性的不可区分关系如下：

表1

条件属性				决策属性
病人	头痛（a）	肌肉痛（b）	体温（c）	流感（D）
e_1	是	是	正常	否
e_2	是	是	高	是
e_3	是	是	很高	是
e_4	否	是	正常	否
e_5	否	否	高	否
e_6	否	是	很高	是
e_7	否	否	高	是
e_8	否	是	很高	否

U按“肌肉痛”否进行划分：$U/\{b\}=\{\{e_1,e_2,e_3,e_4,e_6,e_8\},\{e_5,e_7\}\}$；

U按“体温”情况进行划分：$U/\{c\}=\{\{e_1,e_4\},\{e_2,e_5,e_7\},\{e_3,e_6,e_8\}\}$；

U按条件属性进行划分：$U/C=\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5,e_7\},\{e_6,e_8\}\}$；

U按决策属性进行划分：$U/D=\{\{e_1,e_4,e_5,e_8\},\{e_2,e_3,e_6,e_7\}\}$。

条件属性中忽视某一子属性后对U继续划分：

$U/C-\{a\}=\{\{e_1,e_4\},\{e_2\},\{e_3,e_6,e_8\},\{e_5,e_7\}\}$，

$U/C-\{b\}=\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5,e_7\},\{e_6,e_8\}\}$，

$U/C-\{c\}=\{\{e_1,e_2,e_3\},\{e_4,e_6,e_8\},\{e_5,e_7\}\}$。

下面讨论属性的约简情况：

$pos_C(D)=\{e_1\}\cup\{e_2\}\cup\{e_3\}\cup\{e_4\}=\{e_1,e_2,e_3,e_4\}$，

$pos_{(C-\{a\})}(D)=\{e_1\}\cup\{e_4\}\cup\{e_2\}=\{e_1,e_2,e_4\}$

$\neq pos_C(D)$，

故a为C中D必要的。

$pos_{(C-\{b\})}(D)=\{e_1\}\cup\{e_2\}\cup\{e_3\}\cup\{e_4\}=\{e_1,e_2,e_3,e_4\}$

$=pos_C(D)$，

故b为C中D不必要的。

U按“头痛”否进行划分：$U/\{a\}=\{\{e_1,e_2,e_3\},\{e_4,e_5,e_6,e_7,e_8\}\}$，（$\{e_1,e_2,e_3\}$病症都是“头痛”，故划为一个不可区分类中，$\{e_4,e_5,e_6,e_7,e_8\}$病症都是不“头痛”，划为一个不可区分类中）；

表2

条件属性			决策属性
病人	头痛（a）	体温（c）	流感（d）
e_1	是	正常	否
e_2	是	高	是
e_3	是	很高	是
e_4	否	正常	否
e_5	否	高	否
e_6	否	很高	是
e_7	否	高	是
e_8	否	很高	否

$pos_{(C-\{c\})}(D)=\phi\neq pos_C(D)$，

故c为C中D必要的。

因此，C的D约简为$C-\{b\}=\{a,c\}=\{$头痛,体温$\}$，也就是说表1的属性约简表即表2。

我们还可以通过区分矩阵得出属性约简，首先对决策表1做如下约定：

否，1-是，2-正常，3-高，4-很高，则决策表1可用表3来描述：

表3

U	a	b	c	d
e_1	1	1	2	0
e_2	1	1	3	1
e_3	1	1	4	1
e_4	0	1	2	0
e_5	0	0	3	0
e_6	0	1	4	1
e_7	0	0	3	1
e_8	0	1	4	0

由于区分关系是相互的，所以，区分矩阵是对称的，因此，只需要计算矩阵的一半元素。元素本身之间没有区分关系，只需计算元素e_i,e_j，$i\neq j$之间的区分关系，

e_1与e_2具有相同的属性a，b，属性c不同，所以其对应的矩阵元素为c；e_1与e_6具有相同的属性b，属性a，c不同，所以其对应的矩阵元素为a，c；以此类推，区分矩阵如表4所示。

表4

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
e_1								
e_2	c							
e_3	c	c						
e_4	a	a，c	a，c					
e_5	a，b，c	a，b	a，b，c	b，c				
e_6	a，c	a，c	a	c	b，c			
e_7	a，b，c	a，b	a，b，c	b，c		b，c		
e_8	a，c	a，c	a	c	b，c		b，c	

故其区分函数为：

$$\Delta = ca(a \vee b \vee c)(a \vee c)(a \vee b)(b \vee c) = ac$$

因此，表1的属性约简为$\{a,c\}=\{$头痛,体温$\}$。也就是说，“肌肉痛”对于判定“流感”否没有任何帮助，只需关注属性：“头痛”与“体温”。

下面考虑这两个属性的重要程度。

$r(C,D)=|\mathrm{pos}_C D|/|U|=4/8=1/2$ ， $r(C-\{a\},D)=|\mathrm{pos}_{C-\{a\}}D|/|U|=3/8$，

$r(C-\{c\},D)=|\mathrm{pos}_{C-\{c\}}D|/|U|=0/8=0$，

$SGF(a,C,D)=r(C,D)-r(C-\{a\},D)=1/8$，

$SGF(c,C,D)=r(C,D)-r(C-\{c\},D)=1/2$，

因此，在决策表1中，“体温”最重要，其次是“头痛”，而“肌肉痛”是判断“流感”的冗余信息。

接下来讨论属性约简后的决策规则。

$$U/\{a,c\}=U/C-\{b\}=\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5,e_7\},\{e_6,e_8\}\}$$
$$=\{X_1,X_2,X_3,X_4,X_5,X_6\},$$

$$U/D=\{\{e_1,e_4,e_5,e_8\},\{e_2,e_3,e_6,e_7\}\}$$
$$=\{Y_1,Y_2\}$$

$\mu_{ij}(X_i,Y_j)=|Y_j\cap X_i|/|X_i|=1$，i=1，2，3，4，j=1，2，所以，确定性决策规则有：r_{12}，r_{21}，r_{31}，r_{42}。

$\mu_{ij}(X_i,Y_j)=|Y_j\cap X_i|/|X_i|=1/2$，i=5，6，j=1，2，所以，不确定性决策规则有：r_{51}，r_{52}，r_{61}，r_{62}，它们的确定性因子均为1/2。

即可以归纳出以下几条规则，揭示“头痛”、“体温”情况与是否“流感”之间的关系：

RULE1：IF（头痛=是），（体温=高）OR（头痛=是），（体温=很高）THEN（一定是流感）

RULE2：IF（头痛=是），（体温=正常）OR（头痛=否），（体温=正常）THEN（一定不是流感）

RULE3：IF（头痛=否），（体温=高）OR（头痛=否），（体温=很高）THEN（可能是流感）

也就是说，可以根据RULE1和RULE2来决策病人是否“流感”，否则不能得出绝对正确的决策。

结语

以上实例展示基于粗糙集理论的属性约简与决策方法，依照以上原理，可以对任何决策表进行属性约简并得到有效决策，但当信息系统比较庞大时，依靠人工进行数据处理显然不能满足生产需求，因此，需要学者继续研究其相应算法进行高效约简与决策。

参考文献

[1]Pawlak Z.Rough sets[J].International Journal of Computer Information Science，1982，11（5）：341-356.

[2]Pawlak Z.Rough Sets：Theoretical Aspects of Reasoning about Data[M].Boston：Kluwer Academic Publishers，1991.

[3]张文修，吴伟志，等.粗糙集理论与方法[M].北京：科学出版社，2001.

[4]姚晟，吴照玉，陈菊，等.基于决策理论粗糙集的一种新属性约简方法[J].微电子学与计算机，2019，36（5）：76-81.

作者简介：何圣姿（1985-），女，江西乐安，硕士，讲师，研究方向：模糊数学与粗糙集理论。

基金项目：江西省教育厅科学技术研究项目：基于粗糙模糊理论的聚类分析及其应用（项目编号：GJJ213803）。