基于粗集和模糊集的启发式规则约简算法研究

赵惠文1,2,　朱根标1,　张凤鸣1

(1. 空军工程大学工程学院，陕西西安 710038；2. 空军工程大学理学院，陕西西安 710051)

摘要：提出了一种新型的决策规则约简方法。基于均匀划分和正态分布隶属度函数，对决策表的连续属性进行模糊化，用欧氏距离贴近度来构建相似矩阵，并提出了一种论域的模糊划分算法；依据粗糙集隶属度进行属性约简的基础上，给出了一种决策规则约简算法，从而达到发掘知识并简化知识的目的。

关键词：模糊集；粗糙集；模糊相似关系；属性约简；规则约简

中图分类号：O159　　文献标识码：A　　文章编号：1009-3516(2005)05-0088-03

粗糙集理论[1]为一种处理不完整和不确定的信息提供了一种新型的数学工具，它具有知识提取完全由数据驱动，不需人为假设的优点。但就目前的应用而言，粗糙集理论对连续属性处理能力非常有限。Pawlak提出粗集和模糊集是互为补充的，而不是互相排斥的[2]。Dubois等进一步指出两种理论是处理两种不确定性的不同的数学方法[3]，并提出了模糊粗糙集和粗糙模糊集的数学模型。模糊集可以用来表示集合中子类边界的模糊性，是由隶属度函数刻画的[4~5]。因此，粗糙集和模糊集结合起来生成模糊－粗糙集模型可以更好的描述信息系统，尤其是具有连续属性的信息系统。

1　模糊－粗糙集模型

1.1　连续属性模糊化

一个决策表 T 可以表示为四元组 $\langle U,A,V,f\rangle$ 。其中 $U\neq\varnothing$ 为论域；$A=C\cup D$，C 和 D 分别为条件属性集和决策属性集，且 $C\cap D=\varnothing$；V 为属性的值域集，$V=\bigcup\limits_{a\in A}V_a$，$V_a$ 为属性 a 的值域；f 为信息函数，$f:U\times A\rightarrow V$。假设有一连续属性 a ，则 a 的值域可以表示为 $V=\{V_a(u):a\in A,u\in U\}$ 。根据值域的大小和属性值的分布，可以把 a 模糊化为 k 个语义变量 $Y_i(i=1,2,\cdots,k)$ 。

1.2　基于模糊相似关系的论域划分模型

把粗糙集的不可分辩关系推广为模糊相似关系，引入相似系数 r_{ij}，本文利用基于欧氏距离的贴近度计算模糊相似系数。

定义1　$\tilde{A}=(u\tilde{A}(u_1),\cdots,u\tilde{A}(u_n),),\tilde{B}(u\tilde{B}(u_1),\cdots,u\tilde{B}(u_n),)$ 为两个 n 维向量，则 $\sigma(\tilde{A},\tilde{B})=1-\sqrt{\sum\limits_{i=1}^{n}|u\tilde{A}(u_i)-u\tilde{B}(u_i)|^2}$ 称为基于欧氏距离的贴近度。

定义2　已知 $\tilde{R}\in R_{n\times n}$ 称为模糊相似矩阵，引入一置信水平 λ，经过如下操作，得到模糊矩阵 $\tilde{R}_\lambda$，$\tilde{R}_\lambda$ 成为 λ 下的普通相似关系矩阵。

$$r_{ij}=\begin{cases}1 & r_{ij}>\lambda\\ 1/2 & r_{ij}=\lambda\\ 0 & r_{ij}<\lambda\end{cases}\qquad i,j=1,2,\cdots,n$$

经过上述分析，我们可以得到算法1：如果对于某个 j 有 $a_\mu(u_i)=a_\mu(v_i)$ ，则令 $u_i\in V_j$；否则，建立一个新

收稿日期：2005－01－15
作者简介：赵惠文(1964－)，女，陕西户县人，讲师，硕士生，主要从事数学、信息系统数据分析及知识不确定性研究.

类,$s+1\to s$,$V_s=\{u_i\}$ 。当算法结束时,划分 $U/\mathrm{IND}(R)=(V_1,V_2\cdots V_s)$ 。其具体步骤如下。

Step 1:[初始化] 设置 $1\to i,1\to j,1\to s$。$V_1=\{U_1\}$。

Step 2:判断[$i=|U|$?] 如果 $i=|U|$,那么划分完成,得到 $U/\mathrm{IND}(R)=\{V_1,V_2,\cdots V_s\}$。如果 $i<|U|$,那么转向 step 3。

Step 3:[增加一个 i]$i+1\to i$,$1\to j$,转向 step 4。

Step 4:[判断$j=s$?] 如果$j=s$,那么建立一个新类 $s+1\to s,V_s=\{u_i\}$,然后转向 step 2 并输入一个对象 t。如果$j<s$,那么转向 step 5。

Step 5:[增长一个 j]$j+1\to j$,转向 step 6。

Step 6:[判断 $a(u_j)=a(V_j)$?] 如果 $a_\mu(u_i)=a_\mu(v_i)$,那么 $u_i\in V_j$,转向 step 2。否则,转向 step 4 并检查下一个 V_j。

根据算法 1,可以计算出属性 $a_i\in A$ 在置信水平 r_i 下对论域 U 的划分 $U/\mathrm{IND}(\tilde{R}_{\lambda_i}^{a_i})$,其中不同的属性可以采用不同的置信水平对论域进行划分,而属性集 A 对论域的划分可以表示为 $U/\mathrm{IND}(\tilde{R}_{\lambda_i}^{a_i})=\otimes\{U/\mathrm{IND}(\tilde{R}_{\lambda_i}^{a_i}):a_i\in A,\lambda_i\in\lambda\}$。其中:$A$ 和 λ 分别为属性集与对应的置信水平集;⊗算子定义如下:

$$A\otimes B=\{X\cap\gamma:\forall X\in A,\forall\gamma\in B,X\cap\gamma\neq\phi\}\ 。$$

定义 3　有一决策表 $T=\langle U,C\cup D,V,f\rangle$,设 $U/\mathrm{IND}(\tilde{R}_\lambda^C)=X=\{X_1,X_2,\cdots,X_k\}$,$\gamma=\{\gamma_1,\gamma_2,\cdots\gamma_r\}$ 则 γ 的 $\tilde{R}_\lambda^A$ 正域可以表示为 $\mathrm{POS}_C^\gamma(\gamma)=\cup\{\tilde{R}_\lambda^C(\gamma_i):\gamma_i\in\gamma_i\}$。

1.3　基于模糊相似关系的属性约简

定义 4　有一决策表 $T=\langle U,C\cup D,V,f\rangle$,集合 $C'\subseteq C$ 是 C 的一个最小约简,如果 C'满足如下条件:①$\mathrm{POS}_C^\lambda(\gamma)=\mathrm{POS}_C^\lambda(\gamma)$;②不存在 $C''\subset C'$,使得 $\mathrm{POS}_{C''}^\lambda(\gamma)=\mathrm{POS}_{C'}^\lambda(\gamma)$ 。

定义 5　对象 x 在属性集合 R 下,对集合 Y 的粗糙隶属函数如下 $\mu_X(x)=\dfrac{\mathrm{card}(Y\cap[x]_R)}{\mathrm{card}([x]_R)}$。

根据定义 5 设计以下层次属性约简算法,具体步骤如下:

Step 1:$R=\varnothing$;

Step 2:对每个属性 a_i,计算其粗糙集隶属度 $\mu_X(x)$;

Step 3:选择使 $\mu_X(x)$最大的属性 a_i,且 $R\Leftarrow R\cup\{a_i\}$;

Step 4:如果 $\mathrm{POS}_R^\lambda(\gamma)=\mathrm{POS}_C^\lambda(\gamma)$,转 Step5 跳出,否则转 Step2 继续;

Step5:输出 R。

2　一种决策规则约简算法(FRDRA)

根据上面的模糊 - 粗糙集模型,提出一种决策规则约简算法,其算法步骤如下:

Step 1:利用均匀划分法确定 k 个模糊中心 m_i,并采用正态分布隶属度函数对连续属性进行模糊化;

Step 2:对每个属性采用欧氏距离模糊贴近度计算其模糊相似矩阵;

Step 3:根据算法 1 计算模糊相似关系 $\tilde{R}_\lambda^a$ 对整个论域的划分 $U/\mathrm{IND}\tilde{R}_\lambda^a$;

Step 4:根据算法 2 计算决策表的属性约简;

Step 5:求条件属性相对于决策属性的属性核,删除冗余属性,得到条件属性最小简化,删除重复实例;

Step 6:对每个实例求其属性值的值核,并删除多余的属性值,得到其最小属性值简化;

Step 7:删除决策表中重复实例,归纳出决策规则。

3　实例

为了验证此方法的有效性,选择如表 1 所示的关系型数据库。

表 1　属性决策表

序号	*D*	*E/a*	*S*	序号	*D*	*E/a*	*S*
1	Ph. D.	7.2	6.3	12	Master	3.6	4.1
2	Master	2.0	3.7	13	Master	10	6.8
3	Bachelor	7.0	4.0	14	Ph. D.	5.0	5.7
4	Ph. D.	1.2	4.7	15	Bachelor	5.0	3.6
5	Master	7.5	5.3	16	Master	6.2	5.0
6	Bachelor	1.5	2.6	17	Bachelor	0.5	2.3
7	Bachelor	2.3	2.9	18	Master	7.2	2.5
8	Ph. D.	2.0	5.0	19	Master	6.5	5.1
9	Ph. D.	3.8	5.4	20	Ph. D.	7.8	6.5
10	Bachelor	3.5	3.5	21	Master	8.1	6.4
11	Master	3.5	4.0	22	Ph. D.	8.5	7.0

对{E,S}两个连续属性进行模糊化,利用 Kohonen 网络自组织映射算法分别有 5 个模糊划分的中心 m_i,构造图 1 所示的正态隶属度函数。利用基于欧氏距离的模糊贴近度分别对 3 个属性建立模糊相似关系,引入置信水平,把模糊相似矩阵转化为一般相似矩阵,然后利用算法 1 分别计算模糊相似关系 $\tilde{R}_{\lambda}^{|a|}$,对整个论域 U 的划分 $U/\mathrm{IND}(\tilde{R}_{\lambda}^{|a|})$。

如果 $\lambda_D=1.0,\lambda_E=0.75,\lambda_s=0.8$,则计算结果如下:

$$U/\mathrm{IND}(\tilde{R}_{1.0}^{D})=\{\{1,4,8,9,14,22\},\{2,5,11,12,13,16,18,19,21\},\{3,6,7,10,15,17\}\}$$

$$U/\mathrm{IND}(\tilde{R}_{0.75}^{E})=\{\{2,4,6,7,8,9,11,12,17\},\{14,15\},\{1,3,5,13,16,18,19,20,21,22\}\}$$

$$U/\mathrm{IND}(\tilde{R}_{0.8}^{S})=\{\{1,13,20,21,22\},\{6,7,17\},\{2,3,4,5,8,9,10,11,12,14,15,16,18,19\}\}$$

利用算法 2 进行属性约简,然后利用本文提出的决策规则约简算法,得到如下规则:

1) IF D = Ph. D. AND 6.2≤E≤10 THEN 6.3≤S≤7.0;

2) IF D = Bachelor AND 0.5≤E≤3.8 THEN 2.3≤S≤2.9;

3) IF D = Master THEN 3.5≤S≤5.7;

4) IF D = Bachelor AND 6.2≤E≤10 THEN 3.5≤S≤5.7;

5) IF D = Ph. D. AND 0.5≤E≤3.8 THEN 3.5≤S≤5.7;

6) IF E = 5 THEN 3.5≤S≤5.7。

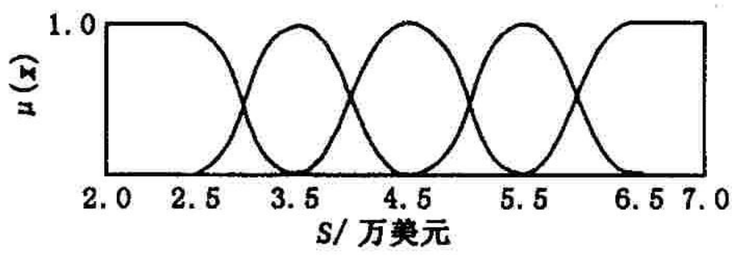

图 1 模糊隶属度函数

从上面的计算可以看出,FRDRA 算法根据 λ 取值的不同,可以得到包含不同规则数目的知识库;文献[5]中提出一种模糊决策树算法 FCLS,得到的规则中存在冗余属性。两种算法的比较见表 2,其中,$\lambda_D=1.0,\lambda_E=0.75,\lambda_s=0.8$。

从表 2 可以看出,一方面,在 FRDRA 中,随着 λ 值的增大,其覆盖率和正确率上升;另一方面,当正确率和覆盖率都为 100% 时, FRDRA 得到的规则数目显然比 FCLS 得到的要少。因此,利用 FRDRA 可以大大减少获取领域知识的数量,更有效地以较少的规则提供领域知识,这对减少知识库的存储空间,降低知识系统中组合爆炸的可能性具有很重要的意义。

表 2 FRDRA 算法与 FCLS 的比较

FCLS	FRDRA			
	$\lambda_E=0.6$	$\lambda_E=0.75$	$\lambda_E=0.8$	$\lambda_E=0.95$
规则数	17	4	6	7	12
覆盖率	100	84	100	100	100
正确率	100	64	77	91	100

4 结论

基于 Rough Set 的机器学习理论是数据挖掘的一个重要手段。信息系统中的连续属性经模糊化处理后,再用 Rough Set 进行属性约简和决策规则的约简,可以提高机器学习与知识发现能力和效率的重要途径。属性约简和决策规则的约简对决策支持系统中规则的提取有着重大意义。决策规则要求的是简明、基于事实的,而大量的关系数据库中的交易数据对决策者来说是很难做出决策判断的,通过属性约简和决策规则约简,在具有同等分类能力下使决策规则极大简化。

参考文献:

[1] Pawlak Z. Ai and Intelligent Industrial Applications: the Rough Set Perspective [J]. Cybernetics and Systems: An international Journal, 2003, 31(4): 227 - 252.

[2] Pawlak Z. Rough Sets and Fuzzy Sets [J]. Fuzzy Sets and Systems, 1985, 17(1): 88 - 102.

[3] Dubois D, Prade H. Rough Fuzzy and Fuzzy Rough Sets [J]. Int J General Systems, 1990, 17:191 - 208.

[4] Kankana C. Fuzziness in Rough Set [J]. Fuzzy Sets and Systems, 2000, 110(2): 247 - 251.

[5] Chen S M, Yeh M S. Generating Fuzzy Rules From Relational Database Systems for Estimating Null Values [J]. Cybernetics and System: An International Journal, 1997,28(8): 695 - 723.

(编辑:姚树峰)

(下转第 94 页)