相容RS理论中的属性约简算法*)

王 珏 刘三阳 刘振华
(西安电子科技大学理学院 西安710071)

Tolerant Rough Set Based Attribute Reduction Algorithm

WANG Jue LIU San-Yang LIU Zhen-Hua
(School of Science, Xidian University, Xi'an 710071)

Abstract In this paper, a novel definition of entropy is introduced. It is used for measuring the uncertainty of roughness of knowledge in tolerant rough sets. In addition, we prove that the entropy of knowledge decreases monotonously as the granularity of information becomes smaller. Then, a new reduction algorithm based on entropy is developed. Simulation results show that the algorithm can find the minimal reduction in most cases.

Keywords Tolerant RS theory, Tolerance relation, Information granularity, Information entropy, Reduction of attribute

1. 引言

Skowron 等[1]提出的相容RS理论是经典的RS理论的推广，它用相容关系(tolerance)代替原来的不可分辨关系(indiscernibility)，可以发现属性值之间的相似性、滤除属性值之间的微小偏差，提高系统决策的鲁棒性和决策效率。Duntsch、Gediga 等从信息论的角度建立了粗糙集理论中的知识与信息熵的关系[2]，而在相容RS理论中，按相容关系划分得到的相容类一般不构成对论域的划分，而是构成了对论域的覆盖，即把知识看作是关于论域的覆盖，这时如果再用传统的信息熵来度量知识的粗糙性，显然是不合理的。本文引入了一种新的信息熵的定义，建立了相容粗糙集理论中的知识与信息熵的关系，即信息熵是随着信息粒度的减小而单调递减的，在此基础上提出了相容RS理论中的属性约简算法，分析了算法的复杂度。实验分析表明该算法能获得信息系统的最小约简。

2. 相容粗糙集的基本概念[3]

定义1 $\mathscr{R}_A=\{R_a\subseteq V_a\times V_a : a\in A\}$是定义在属性集A上的相容关系(tolerance relation)的集合，当且仅当相容关系 R_a 满足：

(1)自反性 $\forall v\in V_a, vR_av$

(2)对称性 $v_1R_av_2\Rightarrow v_2R_av_1$

因此有结论：x 和 y 是关于 a 相似的，当且仅当 $a(x)R_aa(y)$。

定义2 x 和 y 是关于 A 相似的，当且仅当$\forall a\in A$，有 $a(x)R_aa(y)$，记为 $x\mathscr{R}_Ay$。

定理1 $\mathscr{R}_A$ 是一个相容关系，$\mathscr{R}_A=\bigcap_{a\in A}R_a$。

定义3 $\forall x\in U$，x 关于 A 的相容类 $S_A(x)=\{y\in U: x\mathscr{R}_Ay\}$。即 $S_A(x)$ 是与 x 相似的所有对象构成的集合。

另外，对于各种不同属性的属性值之间的相容关系需要有一定的度量函数 $\lambda_a: U\times U\to[0,1]$ 进行度量，以确定两个数值之间的相似程度，对于各个不同的 $a\in A$，λ_a 可以根据属性的各自特征给出不同的定义[4]，以更加符合实际情况。

定理2 v_i 与 v_j 关于属性 a 相似，当且仅当 $\lambda_a(v_i,v_j)\geqslant t(a)$，其中 $t(a)$ 为属性 a 的各属性值的相似阈值。

定理3 v_i 与 v_j 关于属性 $B\subseteq A$ 集合相似，当且仅当对 $\forall a\in B$，有 $\lambda_a(v_i,v_j)\geqslant t(a)$，其中 $t(a)$ 为属性 a 的各属性值的相似阈值。

所以，可以根据属性的不同相似度定义以及不同的相似度阈值，构造一个相容粗糙集模型中的相容关系。论域 U 按这个相容关系 $\mathscr{R}_A$ 来划分，得到了 $|U|$ 个相容类，即 $U/\mathscr{R}_A=\{S_A(x)\mid x\in U\}$，$U/\mathscr{R}_A$，中的相容类一般不构成对论域的划分，它们构成了对论域的覆盖，即 $\bigcup U/\mathscr{R}_A=U$。这时如果我们还用传统的信息熵来度量相容粗糙集模型中知识的粗糙性，显然是不合理的。下面笔者引入一种新的信息熵定义，从信息论的角度建立了相容RS理论中的知识与信息熵的关系，即信息熵是随着信息粒度的减小而单调递减的。

3. 相容RS理论中的知识粗糙性度量

定义4 信息系统 $S=\langle U,A,U,f\rangle$，$B\subseteq A$，则知识(属性集合)B 的信息熵定义为：

$$H(B)=-\sum_{i=1}^{n}\frac{|S_B(x_i)|}{|U|}\log\frac{1}{S_B(x_i)}$$

其中 $U=(x_1,x_2,\cdots,x_n\}$，$|U|$ 是集合 U 的势，$\log x=\log_2^x$。

定理4 $S=\langle U,A,U,f\rangle$，$B,C\subseteq A$，如果 $U/\mathscr{R}_B\subseteq U/\mathscr{R}_C$，那么 $H(B)\leqslant H(C)$。

证明：由 $U/\mathscr{R}_B\subseteq U/\mathscr{R}_C$，可以得出，对$\forall x_i\in U$，$S_B(x_i)\subseteq S_C(x_i)$，所以有 $|S_B(x_i)|\leqslant|S_C(x_i)|$，因此 $\sum_{i=1}^{n}|S_B(x_i)|\log|S_B(x_i)|\leqslant\sum_{i=1}^{n}|S_C(x_i)|\log|S_C(x_i)|$，进而推得 $H(B)\leqslant H(C)$。

这个定理表明，相容粗糙集理论中知识的信息熵是随着知识划分的信息颗粒的减小而单调递减的。

定理5 $S=\langle U,A,V,f\rangle$，$B,C\subseteq A$，如果 $U/\mathscr{R}_B\subseteq U/\mathscr{R}_C$ 而且 $H(B)=H(C)$，那么 $U/\mathscr{R}_B=U/\mathscr{R}_C$。

证明：因为 $H(B)=H(C)$，所以

$$\sum_{i=1}^{n}|S_B(x_i)|\log|S_B(x_i)|=\sum_{i=1}^{n}|S_C(x_i)|\log|S_C(x_i)| \quad (*)$$

由 $U/\mathscr{R}_B\subseteq U/\mathscr{R}_C$，易知对$\forall x_i\in U, S_B(x_i)\subseteq S_C(x_i)$ (#)

从而$1\leqslant|S_B(x_i)|\leqslant|S_C(x_i)|$，所以有

$|S_B(x_i)|\log|S_B(x_i)|\leqslant|S_C(x_i)|\log|S_C(x_i)|$

*)基金项目：国家自然科学基金项目(69972036)；陕西省自然科学基金项目(2000SL03)。王 珏 硕士生，主要研究兴趣：Rough集理论及其应用。

又由式(*),得到$|S_B(x_i)|\log|S_B(x_i)|=|S_C(x_i)|\log|S_C(x_i)|$,从而对$\forall x_i\in U$,$|S_B(x_i)|=|S_C(x_i)|$,再由式(#),得到$S_B(x_i)=S_C(x_i)$,即$U/\mathscr{R}_B=U/\mathscr{R}_C$。由上面的定理,可以得出如下推论:

推论1 $S=\langle U,A,V,f\rangle$,$B,C\subseteq A$,如果$B\subseteq C$,那么$H(C)\leqslant H(B)$

推论2 $S=\langle U,A,V,f\rangle$,其中U为论域,A为属性集,$B=core(A)$,C为信息系统的一个约简,则有$H(B)\geqslant H(B\cup\{a_1\})\geqslant\cdots\geqslant H(B\cup\{a_1\}\cdots\{a_i\}\cup\cdots)\geqslant\cdots\geqslant H(C)$。其中$a_i\in A-B$。

推论3 $S=\langle U,A,V,f\rangle$,$C\subseteq A$而且$H(C)=H(A)$,那么$U/\mathscr{R}_C=U/\mathscr{R}_A$。

推论4 $S=\langle U,A,V,f\rangle$,其中U为论域,A为属性集,C为信息系统的一个约简,如果$H(C)=H(A)$,那么$U/\mathscr{R}_C=U/\mathscr{R}_A$。

推论2表明,如果属性约简以信息表的核为起点,那么在相容RS的属性约简过程中,信息熵的变化规律是单调递减的。推论4表明我们可以以约简后信息系统的信息熵等于初始信息系统的信息熵作为算法的终止条件,在此理论基础上我们就可以构造下面的相容粗糙集中的属性约简算法。

4. 基于信息熵的属性约简算法

在给出算法之前,先给出属性重要性的信息熵定义:

定义5 设$S=\langle U,A,V,f\rangle$是一个信息系统,属性$a\in A$在A中的属性重要性定义为:$SIG_{A-\{a\}}(a)=H(A-\{a\})-H(A)$。

定义表明属性a在A中的重要性可以用A中去掉a后引起的信息熵的变化来度量。

定义6 设$S=\langle U,A,V,f\rangle$是一个信息系统,$C\subseteq A$,任意一个属性$a\in A-C$关于属性集C的重要性定义为:$SIG_C(a)=SIG_{C\cup\{a\}-\{a\}}(a)=H(C)-H(C\cup\{a\})$。

性质1 属性$a\in A$在A中是必要的当且仅当$SIG_{A-\{a\}}(a)>0$。

某个属性的重要性越大,说明它关于属性集就越重要,因此我们将属性重要性作为约简算法的启发式因子,来构造相容RS理论中基于熵的属性约简算法。

基于信息熵的属性约简算法:

输入:一个信息系统$S=\langle U,A,V,f\rangle$,其中U为论域,A为属性集。

输出:该信息表的一个约简。

步骤1:计算信息系统的信息熵$H(A)$。

步骤2:计算属性集的核$core(A)$。计算每个属性$a\in A$在A中的重要性$SIG_{A-\{a\}}(a)$,且令$core(A)=\phi$。若$SIG_{A-\{a\}}(a)$不为0,则$core(A):=core(A)\cup\{a\}$,最后得到属性集$A$的核;若$H(core(A))=H(A)$,则终止,此时,$core(A)$就是$A$的最小约简;否则,转步骤3。

步骤3:令$B=core(AS)$,对属性集$A-B$;

(1)对$A-B$中的每个属性a,计算属性重要度$SIG_B(a)$;

(2)选择属性a使其满足$SIG_B(a)=\max_{a'\in A-B}SIG_B(a')$,$B:=B\cup\{a\}$(若同时存在多个属性达到最大值,则从中选取一个与的属性值组合数最少的属性作为a)。

(3)若$H(B)=H(A)$,则停止。否则,转步骤1。

这时,B为A的一个近似最小约简。

算法复杂度分析:

由于寻找最小约简是NP-hard问题[6],其复杂性主要是由信息表中的属性组合和数据量引起的。计算$core(A)$需计算$|A|$次$SIG_{A-\{a\}}(a)$,而计算一次$SIG_{A-\{a\}}(a)$的时间复杂性为$o(|A|\cdot|U|^2)$;计算约简需计算$SIG_B(a)$的次数为$|A|\cdot(|A|+1)/2=o(|A|^2)$。所以整个算法的时间复杂度为$o(|A|^3\cdot|U|^2)$。

5. 实例分析

给定信息系统$S=\langle U,A,V,f\rangle$,其中$U=\{1,2,3,4,5\}$,$A=(a,b,c,d)$,t为相似度阈值。如表1所示。

表1

	a	b	c	d
1	12	22	1	30
2	13	3	3	16
3	5	5	4	2
4	6	4	2	4
5	14	6	25	18
t	0.75	0.80	0.85	0.90

在这里,我们采用文[4]中的关于λ的相似度量方法进行计算。由定理2.2可以得到如下相似类:

$U/\mathscr{R}_a=\{\{1,2,5\}\{2,1,5\}\{3,4\}\{4,3\}\{5,1,2\}\}$。

$U/\mathscr{R}_b=\{\{1\}\{2,3,4,5\}\{3,2,4,5\}\{4,3,2,5\}\{5,2,3,4\}\}$。

$U/\mathscr{R}_c=\{\{1,2,3,4\}\{2,1,3,4\}\{3,1,2,4\}\{4,1,2,3\}\{5\}\}$。

$U/\mathscr{R}_d=\{\{1\}\{2,5\}\{3,4\}\{4,3\}\{5,2\}\}$。

步骤1:计算$H(A)$,由定理2.1,$U/\mathscr{R}_A=\{S_A(1),S_A(2),S_A(3),S_A(4),S_A(5)\}$。其中

$S_A(1)=\{1\}$,$S_A(2)=\{2\}$,$S_A(3)=\{3\}$,$S_A(4)=\{4,3\}$,$S_A(5)=\{5\}$,

所以

$$H(A)=-(0+0+0+\frac{2}{5}\log\frac{1}{2}+\frac{2}{5}\log\frac{1}{2})=\frac{4}{5}。$$

步骤2:计算每个属性的重要性,因为

$SIG_{A-\{a\}}(a)=SIG_{A-\{b\}}(b)=SIG_{A-\{d\}}(d)=0$,$SIG_{A-\{c\}}(c)=H(A-\{c\})-H(A)=4/5$

所以

$core(A)=\{c\}$,$H(core(A))=H(\{c\})=32/5$。

易见$H(core(A))\neq H(A)$,转步骤3。

步骤3:令$B=core(A)$,对$A/B=\{a,b,d\}$重复:

(1)$SIG_B(a)=4.8$,$SIG_B(b)=3.82$,$SIG_B(d)=5.6$;

(2)$SIG_B(d)=\max\{SIG_B(a):a\in A/B\}$,$B:=B\cup\{d\}=\{c,d\}$;

(3)此时$H(B)=H(A)=4/5$,停止。

因此,输出信息系统的最小约简$B=\{c,d\}$。

结束语 本文给出了相容RS理论中的一种新的信息熵定义,从信息论的角度建立起了相容RS理论中的知识与信息熵的关系,证明了信息熵是随着信息粒度的减小而单调递减的,在此基础上提出了相容RS理论中的属性约简算法,分析了算法的复杂度。实验分析表明了该算法的有效性。

参考文献

1 Skouraon A, Stepaniak J. Generalized approximation spaces. In: Li T Y, ed. Conf. Proc. of the Third Intl. Workshop on Rough Sets and Soft Computing (RSSC'94) San Jose, Califonia, USA, 1994.156～163

2 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(1):34～40

3 Kim D. Data classification based on tolerant rough set[J]. Pattern Recognization, 2001, 34:1613～1624

4 马志锋,邢汉承,郑晓妹,樊恂毅.基于不分明与相似关系的Rough集的超图描述[J].计算机科学,1999,26(9):35～39

5 Liang Jiye, Xu Zongben. Uncertainty Measures of Roughness of Knowledge and Rough sets in Incomplete Information Systems. In: Proc. of the third World Congress on Intelligent Control and Automation. Hefei, P. R. China, 2000

6 Wang S K M, Ziarko W. On optimal decision rules in decision tables[J]. Bulletin of Polish Academy of Science, 1985, 33:676～693