

一种基于条件熵的粗糙集连续属性离散化方法

陈静华^{1,2} 李小民¹

(军械工程学院光学与电子工程系¹,石家庄 050003;中国人民解放军 66489 部队²,北京 100095)

摘要 连续属性离散化是粗糙集应用研究的重点内容之一。基于条件熵可以反映属性依赖度的性质,将决策属性对条件属性的条件熵作为离散化标准,提出了一种粗糙集连续属性离散化方法,并通过实例证明了该方法的正确性。

关键词 粗糙集 离散化 属性依赖度 条件熵

中图法分类号 TP18; **文献标志码** A

粗糙集理论是由波兰学者 Z. Pawlak 于 1982 年提出的一种处理不确定和不精确性问题的数学工具^[1]。其主要思想是在保持原有知识分类能力的前提下,通过知识约简删除冗余信息,以提高分类效率。自粗糙集理论提出以来,成功应用于数据挖掘、模式识别、智能控制等众多领域,已成为国际和国内众多学者的研究热点之一。

Pawlak 提出的粗糙集理论不能直接用于处理连续属性,这在很大程度上限制了其应用范围,因此,连续属性离散化成为粗糙集理论应用研究的一个重要方面,基于信息熵的概念提出一种基于条件熵的连续属性离散化方法,并通过实例验证了该方法的正确性。

1 基本概念

1.1 粗糙集^[2]

定义 1 (决策表) 在粗糙集理论中,称四元组 $S = (U, A, V, f)$ 为一个决策表,其中, U 为对象的非空有限集合,称为论域, $A = C \cup D$ 为属性集合, $C = \{a \mid a \in A\}$ 为条件属性集合, a_i 为 C 的一个简单属性, $D = \{d \mid d \in D\}$ 为决策属性集合,且 $C \neq \emptyset$, $D \neq \emptyset$, $C \cap D = \emptyset$, $V = \cup V_a (a \in A)$ 为

属性值域, f 表示 $U \times A \rightarrow V$ 的一个映射,称为信息函数。当 $d = D$ 时,称决策表为单一决策表,一般地,决策表能够被等价地转化为单一决策表,现针对单一决策表进行研究。

定义 2 (不可分辨关系) 设 $S = (U, A, V, f)$ 为一个决策表, $P \subset A$, 定义不可分辨关系 $\text{IND}(P) = \{(x, y) \in U \times U \mid \forall a \in P, f(x, a) = f(y, a)\}$, $U/\text{IND}(P)$ 是一个等价关系,将论域 U 划分为 k 个等价类: $U/\text{IND}(P) = \{X_1, X_2, \dots, X_k\}$ 。

1.2 信息熵

设 P 和 Q 在论域 U 上导出的划分分别为 X 和 Y , 其中

$$X = U/\text{IND}(P) = \{X_1, X_2, \dots, X_n\};$$

$$Y = U/\text{IND}(Q) = \{Y_1, Y_2, \dots, Y_m\}.$$

可得到信息熵和条件熵的定义。

定义 3^[2] (信息熵) 给定知识 P 和它的概率分布,则称

$$H(P) = - \sum_{i=1}^n p(X_i) \lg_2 p(X_i) \quad (1)$$

为知识 P 的信息熵,其中 $p(X_i) = |X_i| / |U|$ 。

定义 4^[2] (条件熵) 给定知识 P 和 Q 以及它们各自的概率分布和条件概率分布,则称

$$H(Q \mid P) = - \sum_{i=1}^n p(X_i) \sum_{j=1}^m p(Y_j \mid X_i) \lg_2 p(Y_j \mid X_i) \quad (2)$$

为知识 Q 相对于 P 的条件熵,其中 $p(Y_j \mid X_i) = |Y_j \cap X_i| / |X_i|$

$Y_j \mid / \mid X_i \mid$ 。

如果 X 是一个有限集合,则不确定的 Hartley 度量定义为^[3] $H_0(P) = \log |X|$,已知知识 P 时,知识 Q 的正则条件熵为^[4]

$$H_0(Q \mid P) = - \sum_{i=1}^n p(X_i) \sum_{j=1}^m p(Y_j \mid X_i) \lg p(Y_j \mid X_i) / \lg_2 m$$

(3)

$H_0(Q \mid P)$ 反映了属性集 Q 关于 P 的信息依赖度^[5]:

- (1) $0 < H_0(Q \mid P) < 1$;
- (2) 属性集 Q 依赖于属性集当且仅当 $H_0(Q \mid P) = 0$;

- (3) 属性集 Q 独立于属性集 P 当且仅当 $H_0(Q \mid P) = 1$ 。

设 $S = (U, A, V, f)$ 为一个决策表, $A = C \cup D$, 则 $H_0(D \mid C)$ 反映了决策属性对条件属性的信息依赖度, 因为 $H_0(D \mid C) = H(D \mid C) / \lg_2 m$ 而对于一个决策表来说 $\lg_2 m$ 为确定值, 所以决策属性对条件属性的条件熵 $H(D \mid C)$ 同样反映了决策属性 D 对条件属性 C 的信息依赖度, $H(D \mid C)$ 越小说明决策属性 D 对条件属性 C 的依赖度越大, 决策属性对条件属性的依赖度是决策表整体分类能力的度量^[5], 本文以决策属性对条件属性的条件熵作为衡量离散化的标准。

2 基于条件熵的连续属性离散化方法

连续属性离散化实质上就是在连续属性的值域范围内插入若干断点, 将连续属性划分为若干个离散区间。如果划分得过细, 虽然能够提高属性依赖度, 使决策表分类能力增强, 但往往会增加复杂度, 不利于属性约简; 如果划分得较粗, 则可能导致决策表中不相容信息的增加^[6], 因此, 连续属性离散化应在保持原决策表决策属性对条件属性的条件熵不变的前提下寻找最优划分。

按区间等间隔法选取候选点。设 $S = (U, A, V, f)$ 为一个决策表, $A = C \cup D$, 某一条件属性 $a \in C$, 其值域 V_a 的最大值和最小值分别为

V_{\max} 、 V_{\min} , 预先设置的划分点数为 K_a , 则划分间隔为 $d = (V_{\max} - V_{\min}) / (K_a + 1)$ 。于是得到 a 的分割点集^[7]

$$C_a^{K_a} = \begin{cases} \varphi, & K_a = 0 \\ \{V_{\min} + id, \quad i = 1, 2, \dots, K_a\}, & K_a \geq 1, \end{cases}$$

按着预先设置的候选点数先对决策表进行初次离散化, 求出决策属性对所有条件属性的条件熵, 如果与原始决策属性对条件属性的条件熵不相等, 则减少断点数, 直到与决策属性对条件属性的原始条件熵相等。理论上要求离散化后的决策属性对条件属性的条件熵不变, 但实际上, 一般取一个很小的误差因数 β , 只要离散化后的条件熵与决策表的原始条件熵之间误差小于 β 即可。算法步骤如下。

- 1) 给定误差因数 β , $K = K_a$;
- 2) 令 $C' = \emptyset$, 对每个条件属性 $a_i \in C$ (i 为条件属性个数), 按(2)式计算决策属性对每个条件属性的条件熵, 取其中最小值作为决策表的原始条件熵, 即

$$H(D \mid C) = H(D \mid a') = \min(H(D \mid a_i))$$

并令 $C' = C' \cup \{a'\}$;

- 3) 对于 $i = 1, 2, \dots, n$, 且 $a_i \neq a'$ 重复执行:

- ① 给条件属性 a_i 的划分点数 K 赋初值 $K = K_a$, 得到 U 的一个划分;

- ② 令 $C' = C' \cup \{a_i\}$, 由(2)式计算条件熵 $H(D \mid C')$;

- ③ 判断 $|H(D \mid C') - H(D \mid C)| \leq \beta$ 是否成立, 若成立则 $i = i + 1$, 并转①, 否则转④;

- ④ $K = K - 1$, 对新的划分点数 K 用区间等间隔法重新划分 U , 计算 $H(D \mid C')$, 转③;

- 4) 对离散化后的属性值编码。

3 实例分析

采用文献[8]中如表 1 所示的汽轮机故障诊断实例进行实验, 该实例共有 11 个条件属性, 21 组数据, 从中抽取 15 组(2、3、4、5、6、8、9、10、12、13、14、15、17、19、20)作为训练样本, 其余 6 组作为测试数据, 用本文方法对其进行连续属性离散化, 取 $\beta =$

0.01 , $K_a = 3$, 经本文所给算法最终确定的各条件属性断点数 K 及离散化结果如表 2, 采用文献[9]的方法约简后得最小约简为 $\{S_2, S_4, S_8, S_9, S_{11}\}$, 根

据该最小约简得到以下 8 条规则:

$S_2([0, 0.3]) \text{ and } S_4([0.1, 0.55]) \text{ and } S_8([0, 0.45]) \text{ and } S_9([0, 0.225]) \text{ and } S_{11}([0, 0.5])$

表 1 汽轮机故障诊断实例

U	征兆											故障
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	
1	0.8	0	0.1	0.1	1	0	1	0.1	0.9	1	0	1
2	0.8	0	0.1	0.1	0.8	0	0.8	0.1	0.8	0.9	0	1
3	0.5	0	0.1	0.1	0.8	0	1	0.1	0.7	0.9	0	1
4	0.8	0	0.2	0.2	1	0	1	0.1	0.9	1	0	1
5	0.5	0	0.1	0.1	0.8	0	0.9	0.1	0.7	0.9	0	0
6	0.5	0.9	0	0.8	0	0.8	0.1	0.9	0.5	0.1	0.8	0
7	0.5	0.9	0	0.8	0	0.8	0.1	0.5	0.5	0.2	0.9	0
8	0.6	0.7	0	0.9	0	0.5	0.1	0.8	0.4	0.1	0.7	0
9	0.4	0.7	0	0.7	0	0.5	0.1	0.7	0.3	0.1	0.6	1
10	0.3	0.9	0	0.9	0	1	0	0.8	0.1	0	0.9	0
11	0.2	0.7	0	0.8	0	1	0	0.8	0	0	0.8	0
12	0.2	0.6	0	0.6	0	0.9	0	0.7	0	0	0.6	0
13	0.4	0.4	0.3	0.6	0	0	0.1	0.1	0.2	0	0.1	0
14	0.4	0.5	0.3	0.7	0.05	0	0	0.1	0.1	0	0.1	0
15	0.4	0.6	0.4	0.9	0	0.8	0	0.3	0.1	0	0.9	0
16	0.3	0.8	0.3	1	0	1	0	0.1	0	0	1	0
17	0.3	0.4	0.3	1	0	1	0	0.1	0	0	1	0
18	0.6	0.3	0.9	0.3	0.3	0	0	0	0	0	0.6	0
19	0.7	0.3	0.9	0.3	0.2	0	0	0	0	0	0.8	0
20	0.7	0.6	0.9	0.6	0	0	0.2	0.5	0.3	0	0.9	1
21	0.7	0.6	0.9	0.7	0	0	0.3	0.6	0.4	0	0.8	1

表 2 条件属性断点数 K 及离散化结果

条件属性	K	离散化区间			
		1	2	3	4
S_1	0	[0.2, 0.8]			
S_2	2	[0, 0.3)	[0.3, 0.6)	[0.6, 0.9]	
S_3	0	[0, 0.9]			
S_4	1	[0.1, 0.55)	[0.55, 1]		
S_5	3	[0, 0.25)	[0.25, 0.5)	[0.5, 0.75)	[0.75, 1]
S_6	0	[0, 1]			
S_7	3	[0, 0.25)	[0.25, 0.5)	[0.5, 0.75)	[0.75, 1]
S_8	1	[0, 0.45)	[0.45, 0.9]		
S_9	3	[0, 0.225)	[0.225, 0.45)	[0.45, 0.675)	[0.675, 0.9]
S_{10}	3	[0, 0.25)	[0.25, 0.5)	[0.5, 0.75)	[0.75, 1]
S_{11}	1	[0, 0.5)	[0.5, 1]		

表3 测试结果

测试样本	征兆					故障 D
	S_2	S_4	S_8	S_9	S_{11}	
1	[0,0.3)	[0.1,0.55)	[0,0.45]	[0.675,0.9]	[0,0.5)	1
7	[0.6,0.9]	[0.55,1]	[0.45,0.9]	[0.45,0.675)	[0.5,1]	0
11	[0.6,0.9]	[0.55,1]	[0.45,0.9]	[0,0.225)	[0.5,1]	0
16	[0.6,0.9]	[0.55,1]	[0,0.45)	[0,0.225)	[0.5,1]	0
18	[0.3,0.6)	[0.1,0.55)	[0,0.45)	[0,0.225)	[0.5,1]	0
21	[0.6,0.9]	[0.55,1]	[0.45,0.9]	[0.225,0.45)	[0.5,1]	1

$S_9([0.675,0.9]) \text{ and } S_{11}([0,0.5]) \rightarrow D(1)$
 $S_2([0.6,0.9]) \text{ and } S_4([0.55,1]) \text{ and } S_8([0.45,0.9]) \text{ and }$
 $S_9([0.45,0.675]) \text{ and } S_{11}([0.5,1]) \rightarrow D(0)$
 $S_2([0.6,0.9]) \text{ and } S_4([0.55,1]) \text{ and } S_8([0.45,0.9]) \text{ and }$
 $S_9([0,0.225]) \text{ and } S_{11}([0.5,1]) \rightarrow D(0)$
 $S_2([0.3,0.6]) \text{ and } S_4([0.55,1]) \text{ and } S_8([0,0.45]) \text{ and }$
 $S_9([0,0.225]) \text{ and } S_{11}([0,0.5]) \rightarrow D(0)$
 $S_2([0.6,0.9]) \text{ and } S_4([0.55,1]) \text{ and } S_8([0,0.45]) \text{ and }$
 $S_9([0,0.225]) \text{ and } S_{11}([0.5,1]) \rightarrow D(0)$
 $S_2([0.3,0.6]) \text{ and } S_4([0.1,0.55]) \text{ and } S_8([0,0.45]) \text{ and }$
 $S_9([0,0.225]) \text{ and } S_{11}([0.5,1]) \rightarrow D(0)$
 $S_2([0.6,0.9]) \text{ and } S_4([0.55,1]) \text{ and } S_8([0.45,0.9]) \text{ and }$
 $S_9([0.225,0.45]) \text{ and } S_{11}([0.5,1]) \rightarrow D(1)$

测试结果如表3所示,由实验结果可以看出按着本文所给算法对决策表离散化后所得最小约简比较简单,形成的决策规则数也不多,且能够进行正确分类,表明该方法具有一定有效性。

4 结论

决策属性对条件属性的依赖度是决策表整体分类能力的度量,基于决策属性对条件属性的条件熵可以反映决策属性对条件属性的依赖度理论,提出了一种基于条件熵的粗糙集连续属性离散化方法,该方法不涉及领域知识,约简效果较好,并通过实例分析验证了该方法的正确性。

参 考 文 献

- Pawlak Z. Rough Set. International Journal of Computer and Information Sciences, 1982; 11(5): 341—356
- 苗夺谦,李国道. 粗糙集理论、算法与应用. 北京: 清华大学出版社, 2008
- 陶志, 许宝栋, 汪定伟. 基于决策属性支持度的知识约简方法. 东北大学学报(自然科学版). 2002; 23(11): 1025—1028
- Pawiak Z, et al. Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies, 1998; 29: 81—95
- Hartley R V L. Transmission of information. The Bell Systems Technical Journal, 1928; 7(3): 535—563
- Wang J, Miao D Q. Analysis on attribute reduction strategies of rough set. Journal of Computer Science and Technology, 1998; 13(2): 189—192
- 陶志, 许宝栋, 汪定伟, 等. 一种基于粗糙集理论的连续属性离散化方法. 东北大学学报(自然科学版), 2003; 24(8): 747—750
- 杨叔子, 丁洪, 史铁林, 等. 基于知识的诊断推理. 南宁: 广西科学技术出版社, 1993
- 王柯. 基于粗糙集和支持向量机的智能故障诊断方法研究. 无锡: 江南大学, 2008

Fuzzy Simulated Intelligent Control Based on Inverted Pendulum System

REN Bing, LI Zhen-chen

(School of Electrical Engineering and Information Engineering

Lanzhou University of Technology, Lanzhou 730050, P. R. China)

[Abstract] As an absolute unstable controlled system, inverted pendulum is a high order, nonlinear, strong coupling plant. So it may prove the ability of a new control technique in theory. In order to realize the self-swing-up and stabilization control of the inverted pendulum system, fuzzy simulated intelligent control is proposed. Firstly, a human simulated intelligent controller is designed to swing up the inverted pendulum. Then a fuzzy controller is developed to keep the inverted pendulum stabilized near the equilibrium positon. Finally, the effectiveness of the proposed controller is demonstrated by simulation and the fuzzy simulated intelligent control has the ability to control the representative nonlinear instability system.

[Key words] inverted pendulum self-swing-up and stabilization fuzzy simulated intelligent control simulation

(上接第 3733 页)

A Method of Discretization of Continuous Attributes in Rough Sets Based on Conditional Entropy

CHEN Jing-hua^{1,2}, LI Xiao-min¹

(Department of Optics and Electronics Engineering¹, Ordnance Engineering College¹, Shijiazhuang, 050003, P. R. China; 66489

Unit of PLA², Beijing 100095, P. R. China)

[Abstract] The discretization of continuous attributes is one of the important contents in application study of rough sets. Based on the property that conditional entropy can reflect the attribute dependencies, taking conditional entropy of decision attributes to conditional attributes as the standard of discretization, a method of discretization of continuous attributes in rough sets is offered. And the method is proved right through actual example.

[Key words] rough sets discretization attribute dependencies conditional entropy