DOI: 10.16383/j.aas.2001.03.002
第 27卷 第 3期 自 动 化 学 报 Vol. 27, No. 3
2001年 5月 ACTA AUTOMATICA SINICA May, 2001

Rough Set理论中连续属性的离散化方法[1)]

苗夺谦
(山西大学数学系 太原 030006)
(E-mail: dqmiao@mail.sxu.edu.cn)

摘 要 Rough Set(RS)理论是一种新的处理不精确、不完全与不相容知识的数学工具. 传统的 RS理论只能对数据库中的离散属性进行处理,而绝大多数现实的数据库既包含了离散属性,又包含了连续属性. 文中针对传统 RS理论的这一缺陷,利用决策表相容性的反馈信息,提出了一种领域独立的基于动态层次聚类的连续属性离散化算法. 该方法为 RS理论处理离散与连续属性提供了一种统一的框架,从而极大地拓广了 RS理论的应用范围. 通过一些例子将本算法与现有方法进行了比较分析,得到了令人鼓舞的结果.

关键词 Rough Set理论,动态层次聚类,决策表,离散化,相容性.

A NEW METHOD OF DISCRETIZATION OF CONTINUOUS ATTRIBUTES IN ROUGH SETS

MIAO Duo-Qian
(*Department of Mathematics, Shanxi University, Taiyuan* 030006)
(E-mail: dqmiao@mail.sxu.edu.cn)

Abstract Rough set theory is a new mathematical tool to deal with imprecise, incomplete and inconsistent data. The traditional rough set theory can only deal with the discrete attributes in database. However, most real-life databases consist of not only discrete attributes but also continuous attributes. In order to overcome the limitation of the traditional rough sets, using feedback information from decision table consistency we propose a new method of discretization of continuous attributes based on dynamic layer cluster. A unified framework of the rough set theory to deal with discrete and continuous attributes is suggested, which extends the scope of application of rough sets. The results of comparison between this method and some existing algorithms of discretization are encouraging.

Key words Rough set theory, dynamic layer cluster, decision table, discretization, consistency.

1) 国家自然科学基金 (69805004)和山西省青年基金 (981017)资助项目. 本文被评为 98 中国智能自动化学术会议大会优秀论文.
收稿日期 1999-02-11 收修改稿日期 2000-01-28

1 引言

一般来说,数据库中的属性可以分为两种类型. 一种是连续(定量)属性,表示了被描述对象的某些可测性质,其值取自某个连续的区间,如温度、长度等;另一种是离散(定性)属性,这种属性的值是用语言或少量离散值来表示的,如性别、颜色等. 在绝大多数情况下,同一个数据库中既包含了连续属性,又包含了离散属性.

Pawlak提出的 Rough Set(RS)是一种新的处理不精确、不完全与不相容知识的数学理论[1]. 该理论为处理离散属性提供了一种很好的工具,但遗憾的是它不能直接处理连续属性. 这一缺陷大大限制了 RS理论的应用范围. 因此,将 RS理论拓广到能够处理连续属性,这既是 RS理论发展的要求,也是实际应用的需要.

在目前已有的文献中,归纳起来有三种处理连续属性的方法[2]. 为了便于做对比分析,现分别介绍如下.

1) S方法. Slowinski在研究一个医疗诊断决策表的 rough分类时[3],遇到了连续属性的问题. 要利用 RS理论处理这类数据,就必须将它们转换成定性词汇,象"低"、"中"、"高"和"很高"等. 在医疗诊断的实践中,这种转换通常是根据专家的经验标准来完成的. 然后,对定性词汇用数字 0, 1, 2,… 进行编码. Slowinski是利用领域知识进行连续属性离散化的.

2) H方法. Hu把连续属性的离散化看作是面向属性的泛化问题[4]. 泛化是通过该属性的概念树进行的,即如果一些值在概念树中存在着高层概念,那么用相应的高层概念代替对象中的那些值,就得到该连续属性的离散化. 概念树是由领域专家事先提供的,如果对该属性没有提供高层概念,那么该属性就不能通过概念树提升得到离散化. 这时,该算法将这个属性删除.

3) L方法. Lenarcik把原信息系统看成是随机信息系统[5]. 在此基础上,定义了离散(划分)质量的期望值. 设 $U_1, U_2, \cdots, U_k$ 是论域 U 的一个划分,该划分的质量记为 $Y(U_1, \cdots, U_k)$, $E(Y(U_1, \cdots, U_k))$ 表示 $Y(U_1, \cdots, U_k)$ 的数学期望值. 算法以属性的一个初始离散化(一般是将区间等间隔分割)开始. 在以后的每一步中,总是删除使得 $E(Y(U_1, \cdots, U_k))$ 增量最大的分点,直到删除每个剩余分点时,划分质量的值不再增加,算法停止. 这种方法类似于线性规划中参数的后向删除,导出的是次优结果.

本文利用决策表相容性的反馈信息,提出了一种领域独立的基于动态层次聚类的连续属性离散化算法. 只要对距离函数及阈值适当定义,那么,Pawlak关于离散属性等价类的定义便是本算法的特例. 也就是说,本方法为 RS理论处理离散与连续属性提供了一种统一的框架.

2 基于动态层次聚类的连续属性离散化算法

设 $T=\langle U, C\cup D, V, f\rangle$ 是一个决策表,这里 $U=\{x_1, \cdots, x_n\}$ 为论域; C, D 分别为条件与决策属性集; $V=\bigcup_{a\in C\cup D} V_a$, V_a 表示属性 a 的值域; $f: U\times (C\cup D)\to V$ 是一个信息函数,即对 $\forall x\in U, a\in C\cup D$,有 $f(x, a)\in V_a$. 令 $a\in C$ 为一连续属性, $V_c=[a, b]$.

本节将给出一种基于动态层次聚类的离散化算法. 在给出详细算法之前,首先对算法作些定性分析. 我们认为,论域 U 中的对象 x_i 在某连续属性 c 上的取值 $c(x_i)$ 可以看作是随机采集的一组数据. 因此,对这组数据可以根据某种相似度进行聚类分析,从而得到关于 U 的一种划分. 对于一个决策表而言,如果条件属性的划分较粗,则可能导致划分后的决策表不相容;如果划分较细,则可能使划分后的决策表中仍然含有很多冗余信息,使得约简率较低. 对于连续型决策表,由于各属性取很多不同的值,一般来说,该种决策表应是相容的. 因此,在不损失信息的前提下,离散后的决策表也应保持其相容性. 所以,我们对连续属性离散化的目标是,在保证划分后决策表相容性的前提下,寻找使得约简效率最高的划分.

所谓层次聚类算法,就是根据某种聚类准则 (如误差平方和准则)将 n 个样本逐步分成 k 类 $(k < n)$. 对于聚类分析来说,当相似性测度确定之后,影响聚类大小的就是阈值. 由上面的定性分析知道,可以通过划分后决策表的相容性反馈信息,来逐步调整阈值,从而得到连续属性的理想划分.

因为本文对决策表的属性是一一处理的. 所以,样本都是一维变量,它们之间的距离就定义成欧氏距离. 把论域 U 在某连续属性上的取值作为样本集,依据上面定义的距离及给定一个阈值 W,则通过层次聚类算法就得到关于该连续属性的一种划分.

设 $T=\langle U, C, D, V, f\rangle$ 是一个决策表,其中 $U=\{x_1,\cdots,x_n\}$, $C=\{c_1,c_2,\cdots,c_m\}$. 令属性 c 的不相容度为 T,则 T 由下式定义:

$$T=\frac{\mathrm{card}\,C}{\mathrm{card}\,U}, \tag{1}$$

其中 C = {只考虑条件属性 c 时, U 中不相容的对象},符号 $\mathrm{card}E$ 表示集合 E 的基数.

因为决策表中属性之间出现不相容的情况可看作是统计独立的,所以,整个决策表的不相容度 T_T 可表示为 $T_T=\prod_{i=1}^{m} T$. 如果作一次近似处理,即把 T 看作是近似相等的,记为 $\hat{T}$,则得到

$$T_T=\hat{T}^m. \tag{2}$$

从 (2)式可近似估计出 $\hat{T}$,即

$$\hat{T}\approx\sqrt[m]{T_T}. \tag{3}$$

虽然在理论上要求离散化后的决策表是相容的,即 $T_T=0$. 但在工程实际中,一般取 T_T 为一个充分小的数就够了,比如取 T_T 为万分之一. 事实上,决策表中不同属性之间的相容度是存在一定差别的. 因此,在实际处理中要求每个属性的不相容度 T 落在 $\hat{T}$ 的某一范围内就行了,即下式成立

$$|T-\hat{T}|\leqslant U, \tag{4}$$

其中 U 为预先给定的误差.

离散化算法.

输入: 一个决策表 $T=\langle U, C, D, V, f\rangle$,其中 $U=\{x_1,\cdots,x_n\}$, $C=\{c_1,c_2,\cdots,c_m\}$, c_i 均为连续属性, $i=1,2,\cdots,m$, D 为决策属性.

输出: 离散化的决策表.

1) 给 T_T, U 和 W 赋初值;

2) 由(3)式计算不相容度的估计值 $\hat{T}$;

3) 对于 $i=1,2,\cdots,m$,重复

a) 对属性 c_i 以及初始阈值 W,通过层次聚类法可得属性 c_i 关于 U的一种划分;

b) 由(1)式计算该属性的不相容度 T_i;

c)判断 $|T_i-\hat{T}|\leqslant U$是否成立? 若成立 $i=i+1$;否则,转 d);

d)若 $T_i>\hat{T}+U$,则 $W=(1-\Delta)W$,其中 Δ 表示每次调整阈值的步长;若 $T_i<\hat{T}-U$,则 $W=(1+\Delta)W$,并转 a);

4)对离散后的属性值用 0, 1, 2,…进行编码.

注意,虽然上述算法是针对决策表处理的,但它仍然适用于信息系统(无决策属性的数据库)的情况. 对信息系统的处理,只需在上述算法中,令 $D=C$即可.

3　Pawlak关于离散属性的划分是本算法的特例

在 Pawlak提出的 RS理论中,离散属性对于论域 U的划分是通过等价类来表示的. 本节要说明,如果对离散属性的值之间定义适当的距离,并选取一个适当的阈值 W,那么通过上节的算法就能得到该属性关于 U的与 Pawlak定义完全相同的划分.

定理 1. 设 $T=\langle U,C\cup D,V,f\rangle$是一个决策表,其中 $U=\{x_1,\cdots,x_n\}$,$a\in C$为一离散属性. 对属性 a的值定义距离函数如下:

$$d(a(x_i),a(x_j))=\begin{cases}1, & \text{若 } a(x_i)\neq a(x_j),\\ 0, & \text{否则.}\end{cases}$$

算法中的阈值 W可取介于 0与 1之间的任意实数,则本算法关于属性 a对 U的划分与 Pawlak定义的划分完全相同.

上述定理说明,Pawlak关于离散属性定义的划分可看作是本算法的一种特例. 从这种意义上讲,本文提出的算法为处理离散与连续属性的划分问题提供了一种统一的框架,极大地拓广了 RS理论的应用范围.

4　与相关工作的比较分析

4.1　与 L方法的比较

Lenarcik讨论了混凝土抗冻性决策表的离散化问题[5]. 该决策表中条件属性 $a1,a2,a3,a4$和 $a5$的值是刻划凝聚的 5个物理性质的检验结果,它们均为连续属性;最后一个属性(class)是决策属性,其值 1表示抗冻,0表示不抗冻.

Lenarcik把决策表看作是随机信息系统进行处理,利用他们给出的算法得到离散结果,如表 1所示.

表 1　用 L方法离散后的决策表

U	Codes of condition attributes			Class
	$a1$	$a2$	$a4$	
1	0	1	0	1
2	0	0	1	1
3	0	0	0	1
4	0	1	0	1
5	0	1	0	1
6	0	0	0	1
7	0	0	0	1
8	0	0	0	1
9	1	1	1	0
10	1	1	0	0
11	1	1	1	0
12	1	1	1	0
13	1	1	1	0
14	0	1	1	0
15	1	1	0	0
16	0	0	0	0

注意,该方法离散后的决策表中的条件属性由原来的 5个变为 3个. 其理由是属性 $a3$和 $a5$通过离散化处

理后各自的等价类只有一类,即它们各自在 U的所有对象上的取值相同,所以对决策没有影响,故被删除.

利用本文所给的方法对文献 [5]的决策表进行处理,此时取 $T=0.0001$, $U=0.2$,得到离散化后的结果如表 2所示.

通过分析可知,表 2中的条件属性集相对于决策而言是相依的,即该表中存在着冗余信息.用 RS理论对表 2进行约简,得表 3. 该表就是用本算法对原决策表离散化的最终结果.

表 2 用本算法离散后的决策表

U	离散化后的条件属性					Class
	$a1$	$a2$	$a3$	$a4$	$a5$	
1	0	0	1	0	0	1
2	0	0	0	1	1	1
3	0	0	1	0	0	1
4	0	0	0	0	0	1
5	0	1	0	0	0	1
6	0	0	0	0	0	1
7	0	0	0	0	0	1
8	0	0	0	0	0	1
9	0	1	0	0	1	0
10	1	1	0	0	0	0
11	1	1	0	1	1	0
12	1	1	0	1	1	0
13	1	1	0	1	0	0
14	0	0	0	1	0	0
15	1	1	0	0	0	0
16	0	0	0	0	1	0

表 3 用本算法离散后的最终决策表

U	约简后的条件属性				Class
	$a1$	$a2$	$a4$	$a5$	
1	0	0	0	0	1
2	0	0	1	1	1
3	0	0	0	0	1
4	0	0	0	0	1
5	0	1	0	0	1
6	0	0	0	0	1
7	0	0	0	0	1
8	0	0	0	0	1
9	0	1	0	1	0
10	1	1	0	0	0
11	1	1	1	1	0
12	1	1	1	1	0
13	1	1	1	0	0
14	0	0	1	0	0
15	1	1	0	0	0
16	0	0	0	1	0

对表 3和表 1进行对比分析,可得如下结论:

1)虽然条件属性 $a3$在两种方法下都是冗余的,但是,本算法的处理更为合理. 这是因为属性 $a3$的取值为区间 [1.15, 16],其跨度较大,L 方法将其离散化为一个等价类是不太合理的;在本算法下,$a3$离散后的等价类为两类 (0: [1.15, 9.4]; 1: [12, 16]),然后依据 RS的约简理论判断其为冗余;

2)虽然表 1比表 3中的条件属性少一个 $a5$,但是表 3是相容的,而表 1是不相容的. 一般认为,连续型决策表的离散化应保持其相容性不变. 由此看来,用 L方法离散化可能导致决策表信息的丢失.

4.2 与 S方法的比较

考虑由 122个接受过 HSV治疗的胃溃疡病人构成的决策表[3]. 表中每个病人由 11个条件属性 (除属性 1和 4外,都是连续属性)和一个决策属性描述. 决策属性是对手术效果的评价,分为四类: 1) Excellent, 2) Very good, 3) Satisfactory和 4) Unsatisfactory.

Slowinski是利用下面的领域知识 (表 4)对该决策表离散化的.

表 4　关于胃溃疡的领域知识

No.	Attribute(units)	Domain(code)				
		0	1	2	3	4
1	Sex	Δ	Δ			
2	Age(years)	≤ 35	> 35			
3	Duration of disease(years)	≤ 0.5	(0.5, 3]	> 3		
4	Complication of ulcer	none	acute	multiple	perforation	pyloric
5	HCL concentration(mmol HCL/100ML)	≤ 2	(2, 4]	> 4		
6	Volume of gastric juice per 1h (ml)	≤ 70	(70, 150]	> 150		
7	Volume of residual gastric juice (ml)	≤ 50	(50, 100]	> 100		
8	Basic acid output(BAO) [mmol HCL/h]	≤ 2	(2, 3]	> 3		
9	HCL concentration(mmol HCL/100ml)	≤ 10	(10, 15]	> 15		
10	Volume of gastric juice per 1h (ml)	≤ 100	(100, 250]	> 250		
11	Maximal acid output (mmol HCL/h)	≤ 15	(15, 25]	(25, 40]	> 40	

我们利用本算法对该决策表进行了处理[6]. 为了与 S方法对比,本文没有给出关于该决策表离散后的结果,而是给出了由本算法得到的对每个属性离散化的依据,见表 5所示.

表 5　用本算法得到的属性编码依据

No.	Attribute(units)	Domain(code)				
		0	1	2	3	4
1	Sex	0	1			
2	Age(years)	[21, 33]	[34, 60]	[63, 71]		
3	Duration of disease(years)	[0, 4]	[5, 20]	[22, 32]		
4	Complication of ulcer	0	1	2	3	4
5	HCL concentration(mmol HCL/100ML)	[1, 7.5]	[7.6, 16.8]	[17.4, 26.1]		
6	Volume of gastric juice per 1h (ml)	[15, 88]	[94, 180]	[182, 360]	[401, 525]	
7	Volume of residual gastric juice (ml)	[2, 55]	[60, 124]	[134, 254]		
8	Basic acid output(BAO) [mmol HCL/h]	[0.48, 11]	[11.2, 17.7]	[18.6, 26.8]	[36.1, 39.1]	
9	HCL concentration(mmol HCL/100ml)	[1.6, 9.1]	[11, 15.5]	[22, 34]	[38.7, 42.3]	
10	Volume of gastric juice per 1h (ml)	[21, 72]	[90, 286]	[307, 430]	[459, 627]	
11	Maximal acid output (mmol HCL/h)	[2.1, 49.8]	[52.5, 151.4]			

通过对表 4和表 5的对比分析,可得如下结论.

1) 属性 1和 4在原决策表中为离散属性,本算法对它们也进行了处理,并得到了与 Pawlak的定义完全相同的划分. 这也验证了本文第 3节的结论,即 Pawlak 关于离散属性的划分可以看作是本算法的一种特例.

2) 本算法对于属性 2, 3, 6, 7, 9和 10的处理比较好,其结果基本上与专家的划分标准相吻合. 这主要表现在本算法得到的这些属性的某些离散分点与专家关于它们的切分点很接近,因此对这些属性来说,只需将其某些离散区间合并 (需要少量的专家知识),就得到与表 4相符的结果.

3) 本算法对于属性 5, 8和 11的处理不很理想. 这主要是因为这些属性的取值分布在比较广的区间内,但它们的切分点却集中在区间的某一端点附近. 单纯的聚类分析是通过相似度来实现的,它很难有效地处理这种情况. 本算法中的相容性反馈信息,对解决这类问题有一定的作用.

5 结论

目前,大多数对连续属性的离散化方法采用的是领域知识,因而这类方法不具有普遍适应性. 本文利用决策表相容性的反馈信息,提出了一种领域独立的连续属性的离散化算法. 由于本算法只利用了决策表中数据的统计信息,不涉及决策表的领域知识,因此它是一种普遍适应的算法. 通过一些例子,将本算法与目前已有的几种方法做了比较分析,得到了令人鼓舞的结果. 本文的另一个特点是提出了一种处理离散与连续属性的统一框架,极大地拓广了 RS理论的应用范围.

参 考 文 献

1 Pawlak Z. Rough Sets—— Theoretical Aspects of Reasoning About Data. Kluwer: Kluwer Academic Pub., 1991.

2 苗夺谦. Rough Set理论及其在机器学习中的应用研究 [博士学位论文]. 北京: 中国科学院自动化研究所, 1997

3 Slowinski R. Rough Classification of HSV Patients. Intelligent Decision Support. Kluwer: Roman Slowinski, 1992, 77~ 94

4 Hu X H, Cercone N. Learning in relational databases: A rough set approach. *Inter. J. of Computational Intelligence*, 1995, **11**(2): 323~ 338

5 Lenarcik A, Piasta Z. Discretization of Condition Attributes Space. Intelligent Decision Support. Kluwer: Roman Slowinski, 1992, 373~ 389

6 WANG Jue, MIAO Duo-Qian. Analysis on attribute reduction strategies of rough set. *J. of Computer Science and Technology*, 1998, **13**(2): 189~ 192

苗夺谦 1964年出生,博士,教授. 主要研究领域为粗糙集理论、人工智能与模式识别.