一种基于云模型的决策表连续属性离散化方法

李兴生 (1)　　李德毅 (2)

(1) 解放军理工大学通信工程学院　南京 210016　itisnothing@163.com

(2) 中国电子系统工程研究所　　北京 100039　ziqin@public2.bta.net.cn

摘 要　传统 Rough 集理论只能处理离散属性，所以在对决策表进行处理之前，必须对决策表中的连续属性进行离散化。本文提出了一种基于云模型的、领域独立的决策表连续属性离散化方法，尤其适合大数据量的情形。该方法首先根据数据的实际分布，利用云变换将连续属性的定义域划分为多个基于云的定性概念；然后利用决策表不确定性程度的反馈信息合并相邻的定性概念。这种离散化方法是一种软划分，更加符合实际的数据分布和人的思维方式；另外通过合并相邻的定性概念，能够有效提高信息系统[1]中信息的粒度，从而提高所挖掘规则的统计意义和预测强度。

关 键 词　云模型，云变换，Rough 集，决策表，不确定性

A NEW METHOD BASED ON CLOUD MODEL FOR DISCRETIZATION OF CONTINUOUS ATTRIBUTES IN ROUGH SETS

Xingsheng Li (1)　　Deyi Li (2)

(1) (PLA University of Science and Technology, Nanjing 210016)

(2) (Institute of Electronic System Engineering, Beijing 100039)

Abstract　Because traditional Rough Sets theory can only deal with discrete attributes, continuous attributes must be converted into discrete attributes before coping with decision table. In this paper, a new method of discretization of continuous attributes in decision table based on cloud model is introduced, especially suitable for a large amount of data processing. This method makes use of cloud transform to partition the domain of every continuous attribute into many concepts represented by cloud models, and merges neighboring concepts according to feedback information from decision table uncertainty. It can not only reflect the real distribution of data, but also efficiently increase the information granularity of information system.

Keywords　cloud model, cloud transform, Rough Sets, decision table, uncertainty

1、引言

Rough 集理论是一种新型的处理模糊和不确定知识的数学工具。目前已在模式识别、机器学习、知识发现和决策支持等方面获得了广泛的应用。传统 Rough 集理论只能处理离散属性，所以在对决策表进行处理之前，必须对决策表含有的连续属性进行离散化。

针对决策表连续属性离散化问题，目前国内外学者已提出一些方法，如等距离划分方法、等频率划分方法等，但这些方法需要事先人为给定划分的维数。另外一些方法不需要事先给定参数，如布尔逻辑和 Rough 集理论相结合的算法[1]、1RD(One Rule Discretizer)离散化算法[2]等，这些算法都可归结为利用选取的断点对连续属性构成的空间进行划分，得到有限个区域，使得每个区域中的对象的决策属性值相同；但这些算法在本质上对属性空间是硬划分，而且上述有些算法的时间复杂度和空间复杂度都很高，在数据量大的情况下不可取。

文献[3]已证明连续属性的最优划分问题是 NP 的。本文提出了一种基于云模型的、领域独立的决策表连续属性离散化方法，它是一种启发式方法。该方法利用了云变换和决策表不确定性程度的反馈信息，将连续属性的定义域划分为多个基于云的定性概念。这种离散化方

国家自然科学基金（69975024）资助项目

法是一种软划分，划分的边界是模糊的，不确定的，更加符合实际的数据分布和人的思维方[2]式；另外通过合并相邻的定性概念，能够有效提高信息系统中信息的粒度，从而提高所挖掘规则的统计意义和预测强度。

2、 基本概念

2.1 云模型

云模型强调概念是认知的基元，数据是形成概念的要素。云是用自然语言值表示的某个定性概念与其定量表示的数据之间的不确定性转换模型[4]。云由许许多多云滴组成，每一个云滴就是这个定性概念映射到数域空间的一个点，即一次具体实现。这种实现带有不确定性，模型同时给出这个点能够代表该定性概念的确定程度。某一个云滴也许是无足轻重的，但云的整体形状反映了定性概念的重要特性。云模型的详细描述和具体算法可参考文献[7]和文献[8]。

云的数字特征用期望值 Ex 、熵 En 、超熵 He 来表征，这些特征反映了定性知识的定量特性（如图 1 所示的一维正态云模型），它们的具体含义如下：

期望 Ex ：在数域空间中最能代表这个定性概念的数值，反映了云滴群的重心位置。

熵 En ：反映了在数域空间中可被这个定性概念接受的范围，即模糊度，是定性概念亦此亦彼性的度量；另一方面，还反映了在数域空间中的点能够代表这个定性概念的概率，表示定性概念的云滴出现的随机性。总之，熵揭示了模糊性和随机性的关联性。

超熵 He ：是熵的不确定性度量，即熵的熵，反映了在数域空间中代表该语言值的所有点的不确定度的凝聚性，即云滴的凝聚度。

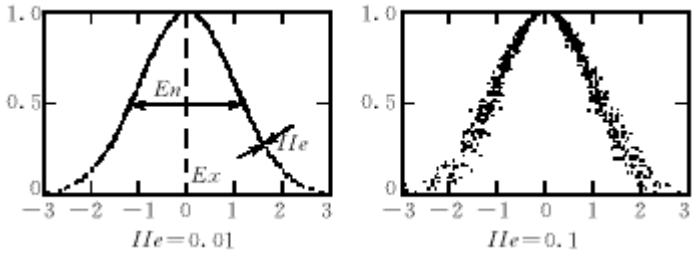

图 1 一维正态云模型

2.2 云变换

云变换是从某一论域的实际数据分布中恢复其概念描述的过程。参考文献[5]和[6]给出了云变换的定义，所谓云变换就是把一个任意不规则的数据分布，根据某种原则进行数学变换，使之成为若干个大小不同的云的叠加，即 $g(x) \approx \sum_{j=1}^{m} c_j * f_j(x)$ $(0<g(x)-\sum_{j=1}^{m} c_j * f_j(x)<e)$，其中 g(x)为数据的分布函数，$f_j(x)$为基于云的概率密度期望函数 $y = e^{\frac{-(x-Ex_j)^2}{2(En_j)^2}}$ ，c_j 为系数，m 为叠加的云的个数，e 为用户定义的可允许的最小误差。叠加的云越多，误差越小。

文献[6]依据如下两个启发性原理：1）论域中的元素对定性概念的隶属程度是一统计属性，具有随机性；2）并且高频率元素对定性概念的贡献大于低频率元素对定性概念的贡献。给出了基于峰值法云变换的概念划分算法。文献[5]中的算法在寻找合适的 En_j 时，是通过

---

[2]

计算云模型的期望曲线 $y = e^{\frac{-(x-Ex)^2}{2(En)^2}}$ 在 Ex_j-5*En_j 到 Ex_j+5*En_j 范围内与 $g'(x)$进行拟合，当拟合后的误差小于允许的误差范围e后，即认为找到了合适的 En_j。但在实际操作中 En_j 的确定比较烦琐。本文采用了一种启发式方法确定 En_j，避免了盲目性，具体方法如下：在迭代的第 j 步中，确定了 Ex_j 后，可以根据实际情况在 Ex_j 周围选取若干个代表点，利用这些代表点通过逆向云算法[4]获得 En_j 的估计值，再通过拟合后的误差对 En_j 进行调整，直到拟合后的误差小于允许的误差范围e后，就认为得到了合适的 En_j。这样就得到了改进的峰值云变换算法，在下一节里将应用改进后的算法对决策表中的连续属性进行处理，生成多个以正态云表征的定性概念。

3、 基于云模型的决策表连续属性离散化算法

3.1 决策表连续属性离散化问题的一般性描述

设 $S =< U, R, V, f >$ 为一个决策表，其中 $R = C \cup D$ 为属性集合，子集 C 和 D 分别为条件属性集和决策属性集；$U = \{x_1, x_2, \cdots, x_n\}$ 为论域；$V = \bigcup_{r \in R} V_r$ 是属性值的集合，表示属性 $r \in R$ 的值域；$f : U \times R \to V$ 是一个信息函数，它指定了 U 中每一个对象 x 的属性值。

对于具有连续属性的决策表，因为连续属性可取很多不同的值，一般来说这种决策表是相容的。令 $r \in C$ 为一连续属性，对其的离散化可以看作是根据论域 U 中的每一个对象在连续属性 r 上的取值，依据某种准则对论域 U 进行的一种划分。对于决策表来说，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，又使得划分后的决策表中含有很多冗余信息，不利于数据约简[9]。由于在许多实际应用中存在干扰和噪声，造成事物之间的边界是模糊的，对数据的划分过程中难免会引入不确定性，但如何适当地引入不确定性，使对连续数据的划分能够与实际的数据分布相符合，是一般的硬划分方法很难做到的，本文利用云模型和决策表不确定性程度的反馈信息较好地解决了这个问题。

决策表的不确定性度量反应了决策表中包含冲突样本的情况，不确定性程度的增加可以提高系统的适应性。根据文献[10]所给出的决策表的不确定性度量，我们给出某一条件属性所决定的分类，相对于决策属性分类的确定性程度定义：

定义 1：设 $S =< U, R, V, f >$ 为一个决策表，$R = C \cup D$，$C = \{c_1, c_2, \cdots, c_m\}$ 为条件属性集，D 为决策属性集，某一条件属性 $c_k \in C$ 所决定的分类为 $E_j \in U / IND(c_k)$ $(j = 1, \cdots, v)$，v 为条件属性 $c_k \in C$ 所决定的分类的数量，$\{X_1, X_2, \cdots, X_n\}$ 是 U 上由决策属性集对 U 的一个划分，则对于任意分类 $E_j \in U / IND(c_k)$ $(j = 1, \cdots, v)$，其相对于决策属性分类的确定性程度为：

$$m_{c_K}(E_j) = \max(\{|\ E_j \cap X_i\ | / |\ E_j\ | : X_i \in U \mid IND(D)\}) \qquad (1)$$

因此，某一条件属性 c_k 确定的分类所造成的决策表的不确定性程度可定义为：

$$m_{uncer}(S/c_K)=1-\sum_{j=1}^{v}\frac{|E_j|}{|U|}m_{c_K}(E_j) \qquad (2)$$

由定义 1 可以很容易地得出如下定理：

定理 1：将主成分（即划分类中多数对象所具有的类别）相同的两个划分类 E_j 和 $E_i \in U/IND(c_k)$ $(j,i=1,\cdots,v)$ 合并成为一个新的划分类后，条件属性 c_k 确定的分类所造成的决策表不确定性程度保持不变。

证明略。

我们对连续属性离散化的目标是，在保证划分后的决策表不确定性程度不变的前提下，寻找使得约简效率最高的划分。本文对决策表中的条件属性是一一处理的。首先利用云变换将某一连续属性的定义域划分为多个基于云的定性概念，由于云模型软划分的特性，在定性概念边界的数据存在亦此亦彼性，所以自然地引入了不确定性；然后根据定理 1，利用决策表不确定性程度的反馈信息合并相邻的定性概念，以得到约简效率最高的划分。

应用云模型对连续属性进行划分，可能会造成离散化后的决策表不相容，这里再对这种划分结果的合理性做进一步说明。虽然一般的决策表离散化算法要求离散化后的决策表是相容的，但在工程实际中，为了使最后获得的决策规则具有更好的适应性和鲁棒性，一般都要适当地引入一定的不确定性，这也符合现实世界中事物存在的本来状态。概念是认知的基元，在现实世界中广泛存在着不精确概念，如“年轻人”、“成绩优秀”、“性能可靠”等等，这些定性概念的外延是不确定的，具有模糊性和随机性。连续属性的离散化本质上就是对处在某一论域中的对象的泛化，也就是一个从低级概念层次向高级概念层次爬升的过程，在这个过程中，由于不精确概念的外延是不确定的，必然会引入不确定性。本文提出的离散化方法利用了云模型这个定性定量转换模型，较好地处理了模糊性和随机性，用云变换方法对连续属性实现了软划分，虽然引入了一定的不确定性，但划分的结果更为合理。

文献[5]和[6]也都利用了云变换的方法，如果将它们的划分算法直接应用于 Rough 集中一般会导致离散化的效果并不突出，关键在于它们不是专门针对 Rough 集理论提出的，没有考虑到 Rough 集对决策表的特殊要求。本文提出的离散化方法利用了决策表不确定性程度的反馈信息，对通过云变换获取的定性概念在不影响决策表不确定性程度的前提下进行合并，减少了离散化后决策表中所含有的冗余信息，并且提高了信息系统中信息的粒度，从而提高了所挖掘规则的统计意义和预测强度。在实际应用中，我们将不确定性程度变化的阈值 a 设为一个较小的数，如千分之一，若合并两个相邻的概念后，相对于合并前不确定性程度的改变小于设定的阈值 a，则认为合并这两个概念是合理的，否则不能合并。

3.2 离散化算法

由前面的分析，我们给出了一个基于云模型的决策表连续属性离散化算法：

输入：决策表 $S=<U,R,V,f>$，其中 $R=C\cup D$ 为属性集合，论域 $U=\{x_1,x_2,\cdots,x_n\}$，条件属性集 $C=\{c_1,c_2,\cdots,c_m\}$，$c_k(k=1,2,\cdots,m)$ 均为连续属性，决策属性集为 D。

输出：离散化的决策表。

步骤 1. 给误差阈值 ε 和不确定性程度变化阈值 a 赋初值；

步骤 2. 循环：k 从 1 到m，执行：

① 生成第 k 个连续属性的数据分布概率密度函数 $f(x)$；

② 利用 2.2 节中改进的峰值云变换算法生成第 k 个连续属性的以云模型表征的定性概念的集合: $CLOUDS(k) = CloudTransform(f(x), e)$；

③ 选择距离最近的两个云模型：$(A, B) = SelectMinDis(CLOUDS(k))$；

④ 采用软或操作进行概念合并：$C = SoftOr(A, B)$；

⑤ 由 3.1 节中的（2）式，分别计算概念合并前和概念合并后该属性相对于决策属性分类的不确定性程度 m_{uncer} 和 m'_{uncer}；

⑥ 若 $|m'_{uncer} - m_{uncer}| \leq a$，$CLOUDS(k) = CLOUDS(k) - \{A, B\} + \{C\}$，并转③；否则，将连续属性映射到概念集 $CLOUDS(k)$ 中的相应概念上，并对概念集用 1，2，3,…进行编码，实现对连续属性 c_k 的离散化，$k = k + 1$，并转①；

步骤 3．结束。

算法的几点说明：

（1）连续属性的数据分布概率密度函数 $f(x)$ 的生成可以通过将原始数据直方图化获得，要注意的是，这里进行的直方图化处理并不是对连续属性的离散化，只是为了获得数据的统计曲线，从而利用云变换获得一组以云模型表征的定性概念的集合。数据量越大，统计曲线越能反映真实数据分布，所以本文所提的算法适合于大数据量的情况。

（2）峰值云变换算法中用到了误差阈值 ε，它的值越大，划分的概念也就越多，详细的讨论可参考文献[6]。

（3）两个云模型的距离定义：设 $A_1(Ex_1, En_1, He_1)$ 和 $A_2(Ex_2, En_2, He_2)$ 是某论域 U 上的两个基本云模型，它们之间的距离定义为：$d(A_1, A_2) = \dfrac{|Ex_1 - Ex_2|}{En_1 + En_2}$。

（4）对相邻两个云模型进行软或操作可获得一个较高层的新概念，我们采用了文献[5]中的方法。设 $A_1(Ex_1, En_1, He_1)$ 和 $A_2(Ex_2, En_2, He_2)$ 是某论域 U 上的两个相邻的基本云模型，如果 $Ex_1 \leq Ex_2$，那么 A_1 和 A_2 进行软或得到新的云模型 $A_3(Ex_3, En_3, He_3)$：

$$A_3 = A_1 \cup A_2 \leftrightarrow \quad Ex_3 = \frac{Ex_1 + Ex_2}{2} + \frac{En_2 - En_1}{4}$$

$$En_3 = \frac{Ex_2 - Ex_1}{4} + \frac{En_1 + En_2}{2}$$

$$He_3 = \max(He_1, He_2)$$

（5）算法步骤 2 中的第⑥小步中需要将连续属性值映射到概念集$CLOUDS(k)$中的相应概念上，可以先求出每个具体的属性值隶属于各概念的隶属度，并将其划分到隶属度最大的那个概念中去。由于云的边界是模糊和不确定的，同样的数值会得到不同的隶属度，所以在两个概念相交的区域，数值相同的元素可以被划分到不同的概念中，这就实现了对数据的软划分。

（6）算法复杂性：我们仅考察对一个连续属性进行离散化时的时间复杂度。在步骤 2 中，第①步需要扫描一遍决策表，执行时间为$O(n)$；第②步的执行时间为$O(m*w)$，其中m为划分的概念个数，w为该属性在第①步中划分的直方图个数，一般来说m和w都远小于记录个数n；第③到⑥步执行时间为$O((2j+1)*n)$，其中j为概念合并的次数，j远小于n，后面的那个$O(n)$为第⑥步的执行时间。因此对一个连续属性进行离散化的时间复杂度为$O(2(j+1)*n)+O(m*w) \approx O(2(j+1)*n)$，则对整个决策表进行离散化的时间复杂度为$O(2(j+1)*n*k)$，其中$k$为决策表条件属性个数。而由文献[3]中的 MD-heuristic 算法的时间复杂度则为$O(n^3k)$，可见本文提出的启发式算法在算法复杂度上有一定优势。

4、 实例分析

为了验证算法的有效性，本文对雷达辐射源特征模拟数据库进行实验分析。选择射频、重复频率和脉冲宽度三个特征参数构成雷达特征向量，它们构成了决策表的连续条件属性，雷达的类别号为决策属性。雷达数据的模拟生成是根据雷达的标称参数利用文献[11]的方法生成的，出于实验的目的，只选择了三种雷达，每种分别生成 5000 条记录构成模拟数据库。生成的部分雷达目标数据见表 1（见第 7 页）。

以条件属性射频为例，应用本文提出的算法对其进行处理，过程可参见图 2 至图 5（见第 7 页）。利用图 5 中的定性概念集对射频属性进行软化分，从而实现了连续属性的离散化。对其它连续属性进行相似处理，即可实现决策表的离散化。表 1 中部分雷达目标数据的离散化结果可参见表 2（见第 7 页）。由表 2 中的结果可知，经过离散化后的决策表引入了不确定性，由于干扰或其它原因，这种不确定性是客观存在的。用云模型对决策表连续属性进行软化分，能够较好地反映数据的实际分布情况，并且恰当地引入了不确定性，使所挖掘的规则具有更强的适应性。由图 4 和图 5 可以看出，通过合并相邻的定性概念，增强了离散化的效果，能够有效提高信息系统中信息的粒度，从而提高所挖掘规则的统计意义和预测强度。对所有连续属性离散化后，就可以进行约简[12][13]和挖掘决策表缺省规则[14] [15]，具体过程和结果略。

5、 结论

尽管 Rough 集理论对模糊和不完全知识的处理比较出色，但其对原始模糊数据的处理

能力比较弱，而基于云模型的定性定量转换方法作为 Rough 集的预处理手段是比较适合的。本文提出了一种适合大数据量的、基于云模型的、领域独立的决策表连续属性离散化方法，实践证明是有效的。如何与云模型更好的结合，以增强 Rough 集方法处理问题的能力，是我们进一步研究的方向。

表 1、 部分雷达目标数据

序号	射频 (MHZ)	重频 (HZ)	脉宽 (μs)	类别
1	3365	1830	7.0	3
2	1890	2310	24	1
3	2470	2730	21.4	2
4	2200	2450	22.5	2
5	1899	2340	26	1
6	3400	1900	6.0	3

表 2、 表 1 中数据的离散化结果

序号	射频	重频	脉宽	类别
1	3	1	1	3
2	1	4	2	1
3	2	6	2	2
4	1	4	2	2
5	1	4	3	1
6	3	2	1	3

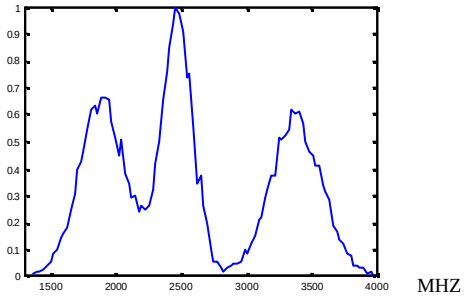

图 2 射频原始数据分布

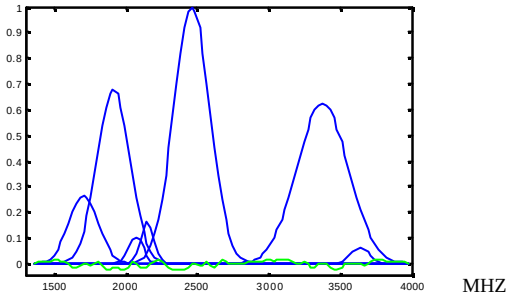

图 3 拟合云模型（实线）和残差（虚线）

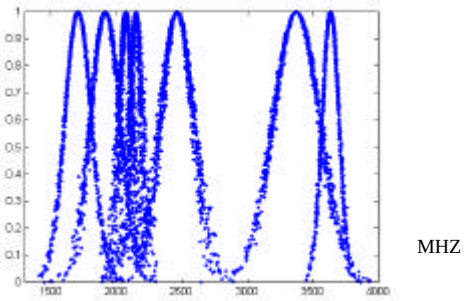

图 4 云模型划分的初步结果

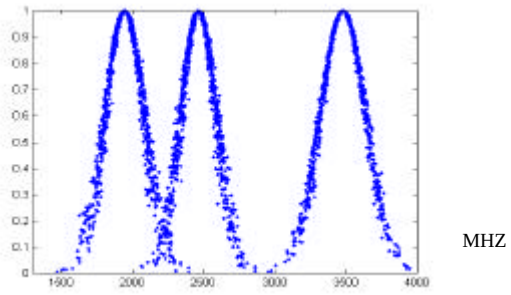

图 5 云模型划分的最后结果

参 考 文 献

1 Nguyen S H, Skowron A, Quantization of Real Value Attributes—Rough Set and Boolean Reasoning Approach. Proc of the Second Joint Conference on Information Sciences, 1995,34~37

2 Holte R C, Very Simple Classification Rules Performs Well on Most Commonly Used Datasets, Machine Learing 11, 1993, 63~90

3 Nguyen S H, Discretization of Real Value Attributes: A boolean reasoning approach, PHD dissertation, Warsaw University, 1997.

4 李德毅，王晔，吕辉军，“知识发现机理研究”，《中国人工智能进展 2001》，2001,314~325

5 蒋嵘，李德毅，范建华，“数值型数据的泛概念树的自动生成方法”，计算机学报，2000,23(5):470~476

6 杜鹢，李德毅，“基于云模型的概念划分及其在关联采掘上的应用”，软件学报，2001,12(2):196~203

7 李德毅，孟海军，史雪梅，“隶属云和隶属云发生器”，计算机研究与发展，1995,32（6）：1520.
8 范建华，“基于云理论的数据开采技术及其在指挥自动化系统中的应用”，解放军理工大学博士学位论文，1999.
9 苗夺谦，“Rough Set 理论中连续属性的离散化方法”，自动化学报，2001,27(3):296~302
10 王国胤，“Rough 集理论与知识获取”，西安交通大学出版社，2001。
11 杜鹢，李德毅，“一种测试数据挖掘算法的数据源生成方法”，计算机研究与发展，2000,37(7):776~782
12 Hu X H, Cercone N, Learning in Relational Databases: A Rough Set Approach. Inter. J. of Computational Intelligence. 1995,11(2),323~338
13 王珏，“基于差别矩阵属性频率的约简算法”，中国科学院自动化研究所技术报告，1996。
14 Mollestad T, Skowron A. A Rough Set framework for data mining of propositional default rules. In: The 9th International Symposium on Methodologies for Intelligent Systems, ISMIS’96, Poland, 1996.
15 尹旭日，陈世福，“一种基于 Rough 集的缺省规则挖掘算法”，计算机研究与发展，2000,37(12):1441~1445