基于语言场理论的连续属性离散化方法及实现*)

周 颖 杨炳儒
(北京科技大学信息工程学院 北京100083)

The Discretization Method and its Realization of Continuous Attribute Based on Language Field Theory

ZHOU Ying YANG Bing-Ru
(College of Information Engineering, Beijing University, Beijing 100083)

Abstract The paper introduces author's job on realization of continuous attribute discretization based on language field theory that Prof. Yang put forward. It applies a new algorithm of seeking border values and its incremental one rather than seeking boundary ones that is a difficulty. The theory of the algorithm is self-contained, and its realization is simple. And the paper introduces simply four thoughts about defining language values and then discretizing for non-numerical value that author already realized in KDD*.

Keywords Data mining, Language field, Continuous attribute, Discretization

1. 引言

在机器学习和KDD(Knowledge Discovery in Database)研究中,大多数算法都是以离散值为处理对象的[1]。因此,常常需要对连续值属性进行离散化。目前,人们已经提出了很多离散化算法[2],如等长度区间法、等频率区间法、基于信息熵(C4.5)的二元分割方法和各种聚类分析方法[3],等等。不同的离散化算法,没有一个绝对的性能评价标准。在众多的离散化方法中,每种方法都有它的适用场合。

语言场与语言值结构是杨炳儒教授早年提出的一种新的知识表示方法,这一表示方法经过了严格的、系列性的定义、定理证明[4,5],并在因果关系定性推理[6]、专家知识归纳获取、Fuzzy集成算法[7]、关联规则的挖掘算法——Maradbcm算法[8]、因果关联规则的评价以及知识发现的自动评价中得到了广泛的应用。本文用一种比较简单的方法实现了基于语言场与语言值结构知识表示的连续属性离散化,这种方法对于加速这一新知识表示方法的工程应用是十分有益的。

2. 基本概念描述

2.1 语言场理论[4,5,9]

令给定数据库D上的所有属性集合$A=\{a_1,a_2,\cdots,a_m\}$,其中,a_i也称为语言变量,而每个属性又可以由不同的程度词来描述属性的状态,如对第一个属性a_i可以表示为$a_i=\{a_{i1},a_{i2},\cdots,a_{ik}\}$,其中,$a_{ij}$也称为语言值,$a_{ij}$的i表示第i个属性,j表示该属性的第j个程度词,如对温度而言,"很高"、"高"等都是程度词,也即语言值。属性程度词是把某一属性和它的一个程度词放在一起(即语言变量+语言值),表示该属性的某种状态,例如,"温度很高"是一个属性程度词。

语言变量和语言值以及基础变量之间的关系可以用下图1来表示。

根据图1所列关系,给出如下相关定义:

定义1 在语言变量相应的基础变量论域中,各个被划分的交叉区间的中点连同ε-邻域(ε通常为允许误差值)内的点,称为标准样本(点),其取值邻域称为标准值;其余诸点均称为非标准样本(点),其取值称为非标准值。它们分别构成标准样本空间与非标准样本空间,并统称为一般样本空间。

定义2 $C=\langle D,I,N,\leqslant_N\rangle$,若满足下列条件:

D为R上交叉闭区间的集合(基础变量论域);

$N\neq\Phi$为语言值的有限集;

$\leqslant_N$为N上的全序关系;

$I:N\to D$为标准值映射,满足保序性,则称C为语言场。

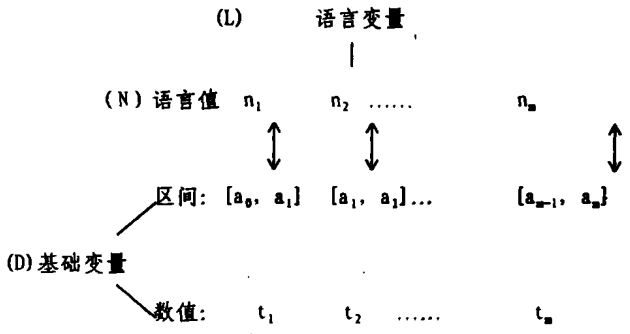

图1 语言变量和语言值以及基础变量之间的关系

定义3 对于语言场$C=\langle D,I,N,\leqslant_N\rangle$,称$F=\langle C,W,K\rangle$为C的语言值结构,如果满足以下条件:

(1)C满足定义2;

(2)K为自然数;

(3)$W:N\to R^k$满足:

$\forall n_1,n_2\in N(n_1\leqslant_N n_2\to W(n_1)\leqslant_{dic}W(n_2))$,

$\forall n_1,n_2\in N(n_1\neq n_2\to W(n_1)\neq W(n_2))$。

其中,$\leqslant_{dic}$为$[0,1]^k$上的字典序,即$(a^1,\cdots,a^k)\leqslant_{dic}(b^1,\cdots,b^k)$当且仅当存在h,使得当$0\leqslant j<h$时$a^j=b^j$,$a^h\leqslant b^h$。

定理1[4](扩张定理) 设C_1,C_2为两个语言场,C_1是C_2的扩张的充要条件是,C_1与C_2是同型语言场(即$|N_1|=|N_2|$)。

*)国家自然科学基金重点项目(69835001),北京市自然科学基金(4022008)。周 颖 博士生,主要研究方向为知识发现。杨炳儒 教授(首席一级),博士生导师,主要研究方向为知识发现与智能系统;柔性建模与集成技术。

定理2[4]（同构定理） 设F为C的语言值结构，则F与F的double扩展在加权Hamming距离下同构。

针对Fuzzy语言变量，可得到相应于上述Fuzzy语言场的定义与定理。

2.2 离散化算法的主要思想

对于一个指定的连续属性（即语言变量）的划分，首先专家（或用户）确定描述该连续属性（语言变量）所需语言值的个数，语言值的标准值和误差半径，阈值上、下限以及对应该语言值的隶属度，其中的误差半径就是用来描述定义1中的ε-邻域。

例如对于连续属性（语言变量）“温度”，用5个语言值“很低”、“低”、“一般”、“高”、“很高”来描述，对应的标准样本点分别为 $a_1=10, a_2=30, a_3=50, a_4=70, a_5=90$，单位是“度”，误差半径分别为 $r_1=2, r_2=2, r_3=2, r_4=2, r_5=2$，单位“度”，对应语言值的隶属度只需由专家或缺省给定其中一个，其余通过计算得到，如“低”为 $A_2=[1, 0.8, 0.5, 0.2, 0]$；专家给定如 $\mu_{high}=1-\exp[-(0.5/(1-x))^{2.5}\ x\in[-1,1]$，“很高” $\mu_{very\ high}=(\mu_{high})^2\ x\in[-1,1]$ 等。阈值上、下限为基础变量论域的上、下限。

对于一个具体的数值u来说，其映射到哪个语言值，可分两种情况（如下图2所示）。

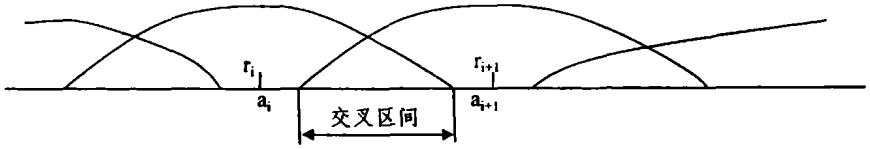

图2 交叉区间示意图

一种情况：如果u落在非交叉区间，则该实际的值映射到该语言值，如上例中69度落在 a_4 的非交叉区间，则它映射到语言值“高”；

另一种情况：如果u落在 a_i 和 a_{i+1} 之间的交叉区间，则利用如下的插值公式：

$$U=A_i\left(1-\frac{|u-a_i|}{l_i}\right)+A_{邻}\frac{|u-a_i|}{l_i}$$

来求取u的非标准向量U，其中 a_i 为第i个区间标准样本点，l_i 为第i个区间长度，A_i 为第i个区间标准向量，$A_{邻}$ 为依u落点所在的相邻区间标准向量，可能是 A_{i+1}，或是 A_{i-1}）。然后根据U与 A_i、A_{i+1} 的测度，或U与 A_i、A_{i-1} 的测度，决定取 A_i 或 A_{i+1} 或 A_{i-1}（取测度较小的），其测度可以用海明距来计算。令非标准向量U的各个分量为 $u_1, u_2, \cdots, u_n$；标准向量 A_i 的各个分量为 $b_1, b_2, \cdots, b_n$，则U与的 A_i 海明距离为：

$$d(U,A)=\frac{1}{n}\times\sum_{i=1}^{n}|u_i-b_i|$$

有了以上的算法就可以得到语言值所映射的区间，其关键是求临界点，然后再对真实数据库进行处理，转换为挖掘数据库。

2.3 离散化算法实现中遇到的问题

现有的语言场理论体系中已经证明了临界点在理论上是存在的，但从上文的介绍中可以看出临界点是一个无限逼近的值，也就是说它是无法用一个实际的值来表达的。所以本文提出一种离散化算法的实现方法，它回避了求取临界点的解决思路，利用数据库中的值求区间边界值，这种方法实现简单，易于理解，从而易于应用。

3. 基于语言场理论的连续属性离散化算法DCL的实现方案

3.1 DCL算法的实现

DCL算法由三个子模块组成：1. 定义语言变量；2. 定义语言值，根据语言值的定义求取区间边界值；3. 根据区间边界值对连续属性离散化。其中2的求取区间边界值是DCL算法的核心，在本节中对它进行介绍，见图3。

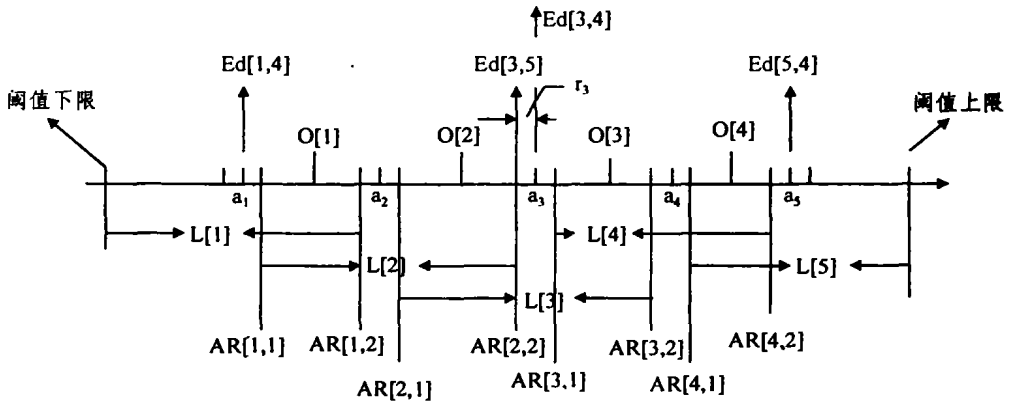

图3 求取区间边界值实现示意图

由于数据量增大时，很可能出现超出原有数据库取值范围的数据值，因此阈值上、下限由领域专家给出，是基础数据论域的上、下限，而不是属性值的上、下限。假设用户定义的语言值个数为5。a_1, a_2, a_3, a_4, a_5 即标准样本点，图3中所示 r_3 即 a_3 的误差半径。如果属性在[阈值下限，AR[1,1]]；[AR[i,2]，AR[i+1,1]]，i=1,2,3；或[AR[4,2]，阈值上限]内取值，即标准样本，其余为非标准样本。所求的边界值形式地表示成：第i个交叉区间的下限是LI[i,1]，上限是LI[i,2]。下面给出

DCL 算法实现中需要用到的两个性质。

性质1 假设语言值个数为5，边界点满足如下关系：LI[1,1]= 阈值下限，LI[5,2]=阈值上限，并且 LI[i,2]<LI[i+1,1]。

性质2 如果没有属性在区间内取值则 LI[i,2]=AR[i,1];LI[i+1,1]=AR[i,2]。

所以，所求边界值为 LI[i,2],LI[i+1,1],i=1…4共 8个值，在一个 DCL 算法的一个 repeat 循环中同时求两个值，即 LI[i,2]和 LI[i+1,1],i=1…4。算法中用 TABTV 表示去除重复值的升序排列的属性值表，用 d_1表示 U 与 a_i 的海明距离，用 d_2表示 U 与 a_{i+1}的海明距离。

3.2 DCL 算法流程图(图4)

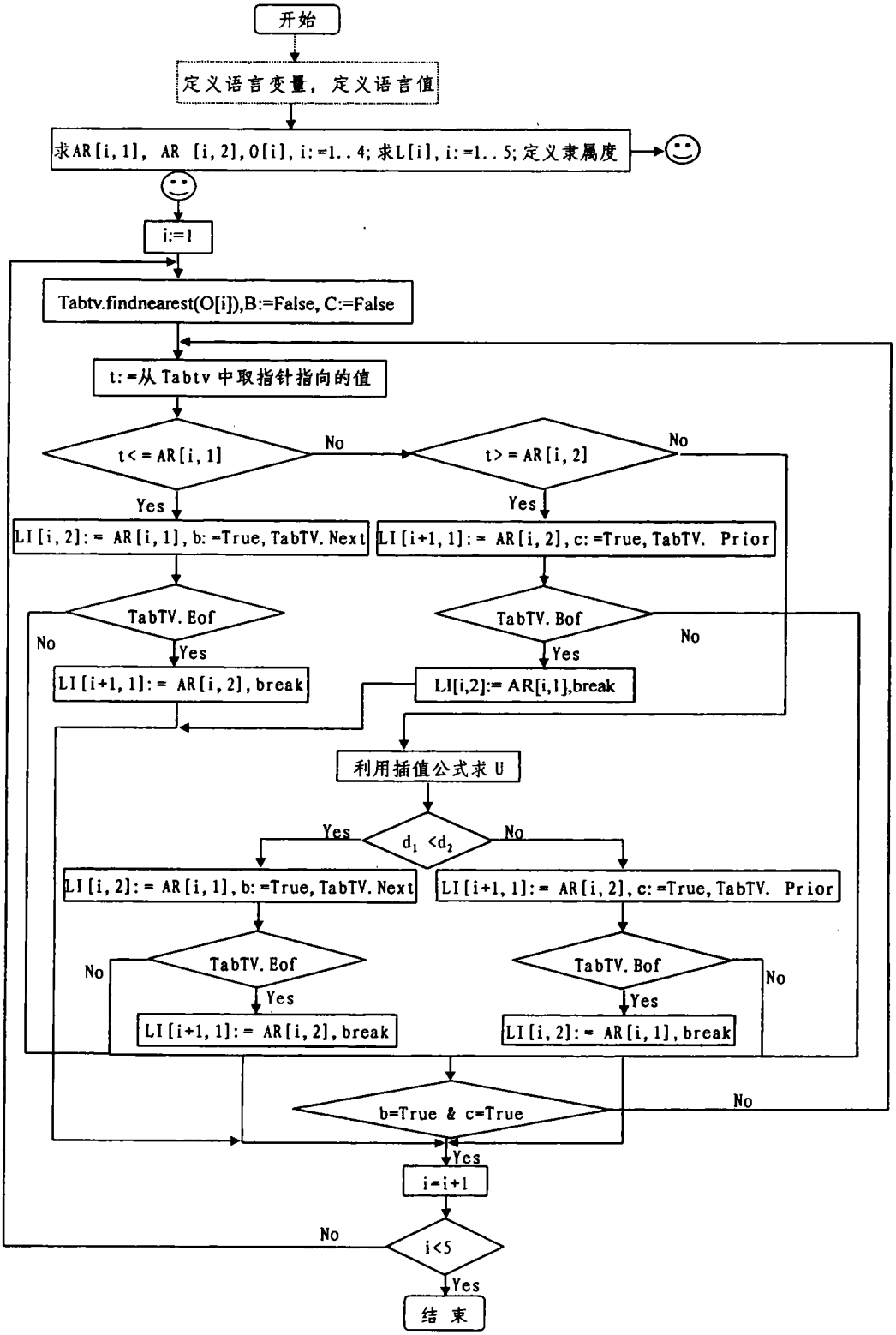

图4 DCL 算法流程图

3.3 DCL 算法实现界面(图5)

图5是为连续的数值型属性定义语言值的程序界面，DCL 算法在单击确定后运行。

3.4 离散化的增量算法

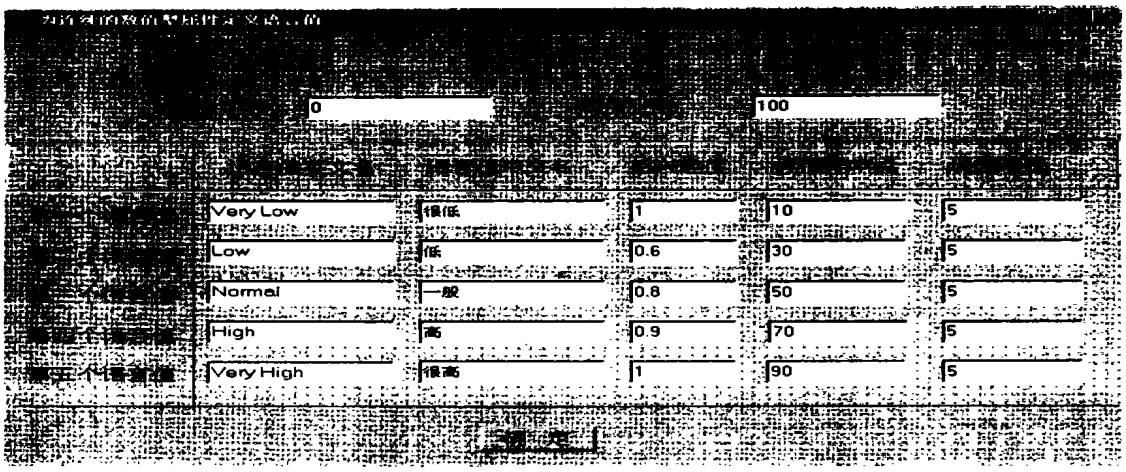

图5 DCL 算法实现界面

数据库是动态变化的，当数据量不断扩张时，只需判断在每个左区间和其相邻右区间中是否有新增属性值，如有，则直接调用原程序，判断新的边界点即可。

因为虽然临界点无法求出，但可以想见，它应该离区间中点相对较近，所以去除了在其周围的重复值的属性值中寻找边界点，程序的运行速度应该很快，不成问题；如果遇到数据量以 M 字节甚至 G 字节增长，并且，关键是新出现的属性值在去除了重复值后，其数据量仍然大到影响程序运行速度时，只需对上述算法稍加改进：即将选取属性值时的移动步长根据具体数值分布情况改为大于1的值，如 a（提供窗口供用户选择），然后就新值进行判断，下一步的移动与前次相同，则步长为 a；方向不同则步长为[a/2]，需保证[a/2]>=1。这样就会大大减少无谓的判断，保证了程序的快速运行。

这一算法已经应用到了 KDD* 的开发中，形成了完善的软件产品。（详见北京科技大学信息工程学院知识工程研究所的网页 www.ustb.edu.cn/kdd/default.htm）

作者在 KDD* 的属性值离散化中所做的比较有特色的工作还包括对非数值型属性提出了四种不同的定义语言值的方法，然后再根据用户对语言值的定义进行离散化，从而基本上满足了用户对数据的不同层次上的离散化需求。这四种方法分别为："一一对应"——即将每一个属性值分别定义为语言值；"特征分类"——属性值中由于部分字段相同，而具有了相同特征；"布列分类"——某些属性值中不具有相同字段，但仍具有某种没有在属性值中表达出来的共同特征；"选择对应"——如果属性值过多，数据量又很大，不能将语言值的数量控制在一个理想的范围内，此时应用选择对应方法定义语言值，可以只选出用户感兴趣的那些值分为有限的几类，程序会在离散时在用户未选到的属性值上记0，表示用户对其不感兴趣，并不影响其他属性的后续处理。

结论 本文介绍了作者在基于语言场理论进行连续属性离散化的具体实现上所做的工作，提出了回避求取临界值这一实现中的难点问题，而改求边界值的算法。这一算法具有理论完备，实现简单的特征，同时在实现中不仅避免了极小数据量下的边界值选取的混乱情况，而且适合大数据量的离散化，不会因为数据量的增大而过大增加程序运行的负担。由于它考虑了数据动态变化中的连续属性离散化的问题，提出了完整的解决方案，故适合于相关领域的工程应用。

参 考 文 献

1 Chen M S, Han J, Yu P S. Data Mining: An overview from a database perspective[J]. IEEE Transaction on Knowledge and Data Engineering, 1996, 8(6): 866～883

2 Shan N, Hamilton H J, Ziarko W, Cercone N. Discretization of continuous valued attributes in classification systems[C]. In: Proc. of the 4th intl. Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, Tokyo, 1996. 74～81

3 席静，欧阳为民. 基于聚类的连续值属性最佳离散化算法. 小型微型计算机系统，2000，21(10)：1025～1027

4 Yang Bingru. Language Field and Its Application, ICICS'97, 1997

5 Yang Bingru. FIM and CASE for Evaluation of Hazard level Based on Fuzzy Language Field. Fuzzy Sets and Systems, North-Holland, 1998, 95(1): 83～89

6 杨炳儒，王忠民. 基于综合语言场的因果关系定性推理模型及其应用. 模式识别与人工智能，1996，9(1)：31～36

7 Yang Bingru. A Type of Language Field Integrated Algorithm Used for Analysis and Control of Complicated System. Journal of System Engineering and Electronics, 1998, 9(1): 66～76

8 孙海洪，杨炳儒. 语言场理论在发掘关联规则中的应用. 计算机科学，2000，27(增刊)

9 孙海洪. KDD 算法和启发型协调器的理论研究及其应用：[博士学位论文]. 北京科技大学，2001