

一种基于模糊聚类的离散化方法

王伟¹, 高亮¹, 吴涛^{1,2}

(1. 安徽大学 数学与计算科学学院, 安徽 合肥 230039;

2. 安徽大学 智能计算与信号处理教育部重点实验室, 安徽 合肥 230039)

摘要:由于粗糙集只能对离散属性进行处理,因而连续属性的离散化也就成了粗糙集的主要问题之一。提出了一种从模糊聚类出发的离散化方法,并给出了一个判别函数,由该函数从聚类结果中选择最优的一个解,因而是一种自寻优的求解过程,避免了人为划分分类数的主观影响。最后进行了实验比较,证实了该方法的有效性和合理性。

关键词:模糊聚类;离散化;粗糙集;连续属性

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2008)03-0053-03

Discretization of Continuous Attributes Based on Fuzzy Cluster

WANG Wei¹, GAO Liang¹, WU Tao^{1,2}

(1. Sch. of Mathematics and Computational Science of Anhui Univ., Hefei 230039, China;

2. Ministry of Education Key Lab. of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China)

Abstract: Because traditional rough set theory can only deal with the discrete attributes in database. So, the discretization of continuous attributes is one of the main problems in rough sets. In this paper, a method of discretization by fuzzy cluster is offered and a criterion function is given, it can select the best solution from the cluster, so it is a superior solution process and subjective influence is avoided. Finally, some experiments are applied to compare with others. The result proved this method is effective.

Key words: fuzzy cluster; discretization; rough set; continuous attributes

0 引言

粗糙集是由波兰科学家 Z. Pawlak 于 1982 年提出的一种理论^[1],该理论对处理不确定性和模糊的知识有很好的效果。目前已广泛运用到人工智能、数据挖掘等各个领域^[2]。在一个决策表中,各个属性可能为离散的,但也可能有连续的。但粗糙集只能对离散的属性进行操作,因此如何将连续属性离散化,成为困扰粗糙集的一个难题。并且离散的结果对最后的分析有很大的影响,如果离散的好,则结果将比较好地反映实际情况,反之,则会得出荒谬的结果。

目前,对连续属性离散化的研究中,已经提出了多种方法,比如,根据决策表的相容度^[3]、微粒群算法^[4]等各个方面进行的离散化方法。文中尝试从模糊集这

一概念出发,构造各样本间的相似矩阵,以截关系将其聚类,并用一个判别函数,在不同层次的聚类中选择一个最优的聚类结果。

1 离散化问题描述

对一决策表 $S = (U, C \cup D)$, 其中 U 为论域, C 为条件属性, D 为决策属性。 $\forall a \in C$, a 为连续属性, 令 $V_a = (\min a(u), \max a(u))$, $u \in U$ 。 $\forall c \in V_a$, 将 V_a 分成 $(\min a(u), c]$ 和 $(c, \max a(u))$ 两个区间, 称 c 为属性 a 的一个断点。假设 a 被 k 个断点分割, 令 $P_a = \{(c_0, c_1), [c_1, c_2), \dots, [c_k, c_{k+1})\}$, 共有 $k+1$ 个区间, 其中 $c_0 = \min a(u)$, $c_{k+1} = \max a(u)$ 。若 $a(u) \in [c_{i-1}, c_i)$, 令 $a(u) = i$, $\forall u \in U$, $i = 1, 2, \dots, k+1$, 则 a 即被离散化。

由此可以看出,如何离散化连续属性 a , 关键在于怎样选择断点, 因此离散化问题也被归结于怎样选择断点对条件属性构成的空间进行划分的问题。但从另一角度考虑, 给一组断点将属性 a 离散化, 也就是找一个等价关系 R 将 a 进行划分, 对于如何找断点, 也就对应于怎样选择等价关系 R 。文中正是从这一点出发, 来

收稿日期: 2007-06-05

基金项目: 国家自然科学基金(60475017, 60675031); 安徽省自然科学基金(050420208); 安徽省高等学校省级自然科学基金项目(2006 KJ244B); 安徽大学学术创新团队和安徽大学人才队伍建设经费

作者简介: 王伟(1984-), 男, 河南信阳人, 硕士研究生, 研究方向为智能计算与信息处理; 吴涛, 博士, 副教授, 硕士生导师, 主要从事机器学习、智能计算及其应用的研究。

寻找一个合适的等价关系 R 将 a 离散。下面首先引入几个概念^[5,6]。

2 基本概念

定义 1: 设 $a \in C$ 为条件属性, $y \in D$ 为决策属性, 称 $S_a(y)$ 为 y 的支持子集, 其中

$$S_a(y) = \bigcup_{w \in U/y} (\bigcup_{v \in U/a, v \subseteq w} V)$$

称 $\text{Spt}_a(y)$ 为 y 关于 a 的支持度(以下简写 S)。其中

$$\text{Spt}_a(y) = |S_a(y)| / |U| \quad (1)$$

显然, 支持度越大, 即 a 被划分的越细, 则 a 对 y 的分类能力就越强, 从而支持度可以反映一条件属性对决策表的分类能力。

定义 2: 设 \underline{R} 为 $U \times U$ 上的模糊关系, 若

$$(1) \forall x \in U, \underline{R}(x, x) = 1$$

$$(2) \forall x, y \in U, \underline{R}(x, y) = \underline{R}(y, x)$$

$$(3) \forall x, y, z \in U, \underline{R}(x, z) \geq \sup_y (\min(\underline{R}(x, y), \underline{R}(y, z)))$$

则称 \underline{R} 为 U 上的模糊等价关系。若 \underline{R} 只满足(1)、(2), 则称 \underline{R} 为 U 上的模糊相似关系。

定义 3: 设 \underline{R} 为 U 上的模糊关系, $\forall \lambda \in [0, 1]$, 称 $R_\lambda = \{(u, v) \mid \underline{R}(u, v) \geq \lambda\}$ 为 \underline{R} 的 λ -截关系。

很显然, 若 \underline{R} 为模糊等价关系, 则 R_λ 就是 U 上普通的等价关系。

3 模糊聚类

简单地说, 模糊聚类就是给出各个样本之间的模糊相似矩阵, 使得 x_i, x_j 在 λ 水平下, 将 $\underline{R}(x_i, x_j) \geq \lambda$ 的归为一类, 即 x_i, x_j 的相似程度不低于 $\lambda, x_i, x_j \in U$ 。因此, 模糊聚类法大致可分为三类^[6]: 一类是基于模糊等价关系的传递闭包法; 一类是基于模糊相似关系的直接聚类法, 包括最大树法和编网法; 第三类是基于模糊 C -划分聚类法。在聚类结果上, 前两种方法是一致的, 可根据情况而选。对于第三种方法, 要事先指定划分的类数 C , 然后依据所给规则, 找到最优聚类。对于一些属性, C 可由业内专家给出, 但对于同一个问题, 不同专家给出的类数可能不尽相同, 因而带有很大主观性在内, 这也是 C -聚类的一个缺点。文中方法是给出一判别函数, 在保证属性分类能力的前提下, 自动寻找最优的划分结果, 从而避免主观因素的影响。

4 基于模糊聚类的离散化方法

对属性进行离散化, 至少要满足两个准则: (1) 离散化后的结果要尽可能地反映原始数据的信息, 即尽

量减少信息的流失, 但又不增加冗余的信息; (2) 聚类的效果要好, 即类内的各点要尽可能地在一起, 而类间则要分开明显。第一个条件, 可用条件属性的支持度来进行控制, 以反映分类的效果, 第二个条件, 引入 F -统计量进行控制, 即

$$F = \frac{\sum_{i=1}^c \frac{n_i |\bar{x}^i - \bar{x}|^2}{c-1}}{\sum_{i=1}^c \sum_{j=1}^{n_i} \frac{|x_j^i - \bar{x}^i|^2}{n-c}} \quad (2)$$

其中 n_i 为第 i 类的个数, c 为类数, \bar{x}^i 为第 i 类的中心, \bar{x} 为全体元素的中心, n 为样本个数。

由上式可以看出 F 的分子表征类与类之间的距离, 而分母表征类内素的距离, 因而 F 的值越大, 则分类越合理, 因而可由 F 来反映第二个条件。这样对于离散的结果可由(1)、(2)两式共同控制, 只须对二者分配一下权重。

该文对连续属性是逐一进行离散化的^[3]。对于连续属性 a , 首先得出相似矩阵 $M = (r_{ij})_{n \times n}$, 其中相似函数有多种, 可灵活选之。文中用的是绝对指数函数, 即 $r_{ij} = \exp(-|x_i - x_j|)$, 对于矩阵 M , 其所有不同元素组成的集为 X , 对 $\lambda \in X$, 依次作截关系 R_λ , 并用编网法得到聚类结果(容易得出, λ 越大, 分类越细; λ 越小, 分类越粗, 当 $\lambda = 0$ 时, 分类最粗, $\lambda = 1$ 时, 分类最细, 因而用编网法得到的聚类, 对不同的 λ 构成一分层递阶结构)。并计算(1)式和(2)式的值, 得到二者的综合 $G_\lambda = \alpha F + \beta S$, 其中 α, β 为二者的权重系数。对于所有的 $\lambda \in X$, 取 G_λ 最大的一个作为聚类结果。具体算法如下:

(1) 首先给 α, β 赋值。

(2) 对连续属性 a , 令 $B = \emptyset, G = 0$, 得出其相似矩阵 $M_{n \times n}$, 并得出 M 中各不相同的值组成的集合 X 。

(3) 依次对 $\lambda \in X$, 得到截关系 R_λ , 用编网法得到聚类结果 $B1$, 并计算出 $G_\lambda = \alpha F + \beta S$ 的值。

(4) 若 $G_\lambda - G < 0$, 则转入(3), 进行下一 λ 的计算; 若 $G_\lambda - G > 0$, 令 $B = B1, G = G_\lambda$, 再转入(3)。

(5) 最后得到聚类结果 B , 分别用 1, 2, 3, ... 来编码各类中的个体。

(6) 转到(2), 对下一个连续属性进行离散化。

5 实验

(1) 文中采用文献[7]中的数据作对比实验, 该数据描述的是影响转子平衡的各项数据, 共有 11 个条件属性, 1 个决策属性。用本法得到的离散化结果见表 1 ($\alpha = 0.4, \beta = 0.6$), 算法实现环境为 matlab 7.0。文中随机选取 70% 的样本作为训练集, 30% 作为测试

集,用 Rose2 软件对训练集进行最小规则提取,然后用所得规则对测试集进行分类。文中共做了 10 次实验,结果见表 2。可以看出 10 次平均分类正确率为 0.917,高于文献[7]的 0.833,说明用本法得到的离散化结果具有优越性,也说明了该法的有效性。

表1 离散化后的数据

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	d
1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	1	2	1	1	1	1	1
3	2	1	1	1	2	1	1	1	2	1	1	1
4	1	1	2	2	1	1	1	1	1	1	1	1
5	2	1	1	1	2	1	1	1	2	1	1	0
6	2	2	3	3	3	2	3	2	3	2	2	0
7	2	2	3	3	3	2	3	3	3	3	2	0
8	2	3	3	3	3	3	3	2	3	2	3	0
9	2	3	3	4	3	3	3	4	4	2	3	0
10	3	2	3	3	3	2	4	2	5	4	2	0
11	3	3	3	3	3	2	4	2	6	4	2	0
12	3	3	3	4	3	2	4	4	6	4	3	0
13	2	3	2	4	3	1	3	1	4	4	4	0
14	2	3	2	4	3	1	4	1	5	4	4	0
15	2	3	4	3	3	2	4	5	5	4	2	0
16	3	2	2	3	3	2	4	1	6	4	2	0
17	3	3	2	3	3	2	4	1	6	4	2	0
18	2	3	5	2	4	1	4	6	6	4	3	0
19	2	3	5	2	5	1	4	6	6	4	2	0
20	2	2	5	4	3	1	5	3	4	4	2	1
21	2	2	5	4	3	1	5	3	3	4	2	1

表2 实验结果

实验次数	1	2	3	4	5	6	7	8	9	10
分类正确率	1	0.833	0.833	1	0.833	1	0.833	0.833	1	1

(2)文中还对文献[8]中的数据进行了试验,数据中共有 6 个条件属性($a_1, a_2, b_1, b_2, c_1, c_2$),分别表示故障中垂直和水平的方差、峭度和偏斜度。决策属性为故障名称(松动,喘动,流体激励)。文献[8]对属性离散的方法是自适应特征映射网络法(SOM),最后得到的故障特征 $T = \{c_1, c_2, b_1\}$,从而偏斜度为最能反映故障的特征。若用文中的离散化方法对该数据进行离散,并用文献[8]的故障提取法得到的故障特征为 $T_1 = \{c_1, c_2, b_2\}$,与 T 相比,只是故障中峭度的垂直和水平有所不同,但最能反映故障特征的还是偏斜度。在实际中三故障的重心是有明显变化的,流体激励较

稳定,松动较大,喘动介于二者之间。所以偏斜度最能反映该三故障的特征。文中得到的结果也与实际一致,所以文中的离散化方法能很好地保持和反映数据的本质特征。

6 结语与展望

文中从等价关系这一角度出发对连续属性进行离散化,并不需要额外的知识,只根据数据本身的结构而得出结果,因而是一种普遍适用的算法,而且不需要指定类数,避免了主观因素的影响。从计算复杂度上看,因模糊聚类要逐一扫描相似矩阵,其复杂度要高于等距、等频离散法,但这两种离散法根本没有考虑数据本身的结构,因而与该两种方法相比,模糊离散法更具合理性。

在实际情况中,对于步骤(2)中的 X ,并不需要逐一进行扫描,在实验中发现,把 X 从小到大排序后,对 X 的前一部分值得到的聚类结果都是一类,可以不用重复扫描这些值,可以采用二分法来进行筛选,这将大大提高程序运行效率,节省时间。

参考文献:

- [1] Pawlak Z. Rough sets theory and its application to data analysis[J]. Cybernetics and Systems, 1998, 29(9): 661-688.
- [2] Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning about Data [M]. Dordrecht: Kluwer Academic Publishers, 1991.
- [3] 苗夺谦. Roughsets 理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302.
- [4] 张腾飞, 王锡淮, 肖键梅. 基于微粒群优化的连续属性离散化算法[J]. 计算机工程, 2006, 32(3): 44-46.
- [5] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001: 1-40.
- [6] 胡宝清. 模糊理论基础[M]. 武汉: 武汉大学出版社, 2004: 148-175.
- [7] 陶志, 许宝栋. 一种基于粗糙集理论的连续属性离散化方法[J]. 东北大学学报, 2003, 24(8): 747-750.
- [8] 郭小芸, 马小平. 基于粗糙集的故障诊断特征提取[J]. 计算机工程与应用, 2007, 43(1): 221-224.

(上接第 52 页)

接入各种新设备,使维护真正智能化。

参考文献:

- [1] 石柱, 张子义. 基于 CDT 规约的通用 RTU 测试系统研制[J]. 吉林电力, 2004, 27(2): 30-31.
- [2] 鞠阳, 陈锦涛. 用 VB 设计 CDT 循环规约[J]. 江西电力,

2005, 27(4): 17-19.

- [3] 刘艺. DELPHI 模式编程[M]. 北京: 机械工业出版社, 2004: 27-45.
- [4] 罗贤缙, 孟建良. 电力领域构件存储及检索方法的研究与实现[J]. 微机发展, 2005, 15(5): 89-90.
- [5] 胡磊, 张曙光, 戈晓斐. 代码统一设计的若干方法及比较[J]. 计算机应用研究, 2004, 21(3): 21-22.