一种基于粗糙集理论的连续属性离散化新算法 *

李 慧[1], 闫德勤[1], 韩 丽[2]
(1. 辽宁师范大学 计算机系, 辽宁 大连 116081; 2. 柴河林业局第一小学, 黑龙江 海林 157131)

摘 要: 粗糙集理论中要求离散化保持原有决策系统的不可分辨关系,但以往的一些算法在离散过程中会使近似精度控制在可以接受的范围,即允许一定的错分。针对此不足,在保证决策属性绝对不改变的情况下,提出一种新的区间拆分方法,更合理有效地对连续属性进行离散化。实验通过 C4.5 和支持向量机分别对离散化后的数据进行识别与分类预测,实验结果证明了算法的有效性。

关键词: 连续属性离散化; 粗糙集; 决策表; 离散区间; 数据挖掘
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0077-02
doi:10.3969/j.issn.1001-3695.2010.01.022

Novel algorithm for discretization of continuous attributes based on rough sets theory

LI Hui[1], YAN De-qin[1], HAN Li[2]
(1. *Dept. of Computer Science, Liaoning Normal University, Dalian Liaoning* 116081, *China*; 2. *The Primary School of Chaihe, Hailin Heilongjiang* 157131, *China*)

Abstract: The rough set required that discretization should be maintained indiscernibility of the original decision-making system, however, many algorithms before permitted approximate quality descended controlled certain scope. This paper proposed a novel method of splitting interval. The novel algorithm was more reasonable and effective to discretization of continual attribute, and assured not to change decision-making attributes. By using C4.5 and SVM, performed the experiments respectively with the results of discreted data. The experiment results show that the presented algorithm is effective.

Key words: discretization of continuous attributes; rough set; decision table; discretization interval; data mining

连续属性离散化是机器学习和数据挖掘研究和应用中的一个重要方面,在规则提取、特征分类等很多算法中,连续(实值)属性必须进行离散化。目前,离散化算法的类型可分为[1]合并(自底向上)与拆分(自顶向下)、有监督与无监督、全局与局部、静态与动态和直接式与增量式。大多数离散化算法都是基于统计学或基于信息熵的,如 extended-Chi2[2]、entropy-MDLC、CAIM 算法[3]等。离散化算法的关键在于如何获得最优划分,最大程度地保持信息表示的意义,减少信息损失。

粗糙集由波兰数学家 Pawlak[4] 于 1982 年提出,它是一种处理模糊不确定性知识表达、学习及归纳的数学理论,现已被广泛应用到人工智能、模式识别和数据挖掘等很多方面。在粗糙集理论中,离散化要求保持原有决策系统的不可分辨关系,由此,本文提出一种保证决策属性绝对不改变的离散化新算法。

1 粗糙集理论

设 $S=(U,A,V,F)$ 为一信息系统。其中,$U=\{x_1,x_2,\cdots,x_n\}$ 是论域;A 是属性集合;V 是属性取值集合;F 是 $U\times A\to V$ 的映射。若 $A=C\cup D$,$C\cap D=\varnothing$,C 称为条件属性集,D 称为决策属性集,则该信息系统称为决策表。

定义 1 $x,y\in U$,对于 $P\subseteq A$,θ_P 是 U 上的一个等价关系,如果满足 $x\theta_P y\Leftrightarrow(\forall p\in P)(f_p(x)=f_p(y))$,则称 θ_p 是 x,y 的一个不可分辨关系。

定义 2 设 U 为一个论域,P、Q 为 U 上的两个等价关系簇,Q 的 P 正域记为 $\mathrm{POS}_P(Q)$,定义为 $\mathrm{POS}_P(Q)=\bigcup_{X\in U/Q}P\_(X)$。

定义 3 设 $P\subseteq C$,对于划分 $\{Y_1,Y_2,\cdots,Y_K\}$ 的 P 的近似精度为 $\gamma_p=\sum_{i=1}^{k}\mathrm{card}(P\_Y_i)/\mathrm{card}(U)$。其中:$\mathrm{card}(\ )$ 表示集合的基数;γ_p 反映决策表分类的正确程度,描述了关于论域 U 的知识完备程度。

粗糙集理论中的一类重要研究课题——数据离散化问题,它属于粗糙集理论中的预处理问题(决策表的补齐和数据的离散化)之一,在粗糙集理论分析的其他环节(如属性约简和值约简)之前进行。

2 本文新算法的提出

在粗糙集理论中对离散化的要求是保持原有决策系统的不可分辨关系。以往的一些算法一般是在离散过程中会使近似精度降低范围控制在可以接受的范围,也就是说允许一定的错分。而近似精度是反映决策表分类的正确程度,描述关于论域的知识完备程度(定义 3),因此,降低近似精度会或多或少使信息有一定损失。针对此不足,本文提出了一种新的区间拆分方法,在保证决策属性绝对不改变的情况下,更合理、更有效

收稿日期: 2009-05-11; **修回日期**: 2009-06-20 **基金项目**: 国家自然科学基金资助项目(60372071);中国科学院自动化研究所复杂系统与智能科学重点实验室开放课题基金资助项目(20070101);辽宁省教育厅高等学校科学研究基金资助项目(2008344);大连市科技局科技计划资助项目(2007A10GX117)

作者简介:李慧(1980-),女,黑龙江海林人,硕士研究生,主要研究方向为数据挖掘和粗糙集理论(huili_913@yahoo.com.cn);闫德勤(1962-),男,山东成武人,教授,博士,主要研究方向为模式识别、数据挖掘和信息安全等;韩丽(1981-),女,黑龙江海林人,一级教师.

地对连续属性进行离散化。根本出发点始终遵循离散化算法的实质——在于如何获得最优划分,最大程度地保持信息表示的意义,减少信息损失。

本文在离散过程中,候选断点的选择分两个步骤:a)对首尾数据进行处理(这样做比传统断点选择时间复杂度有所下降);b)对中间样本进行处理,不过都是采用拆分区间方式完成。最终确定断点采取合并方式完成。离散化算法描述如下:

input: M 为实验数据集总样本数,S 为数据集的类的个数,以及连续属性 A

a)将条件属性 i 的实值按从小到大顺序排序。

b)找到首次出现决策属性不同的地方插入第一个断点,将数据集 S 划分成两个区间。例如共有 20 个样本,排序后前八个样本决策属性都为 1,第九个样本决策属性非 1,这时第一个断点就插入到八、九样本之间。

c)将条件属性 i 的实值大于某一个值后其决策属性不再发生变化的样本归为一个区间。例如共有 345 个样本,从 255 个样本之后直到 245 个样本决策属性都是 4,将这些样本一次性划分到一个区间,进行完此步骤,数据集 S 被划分成三个区间。

d)对中间区间所有样本进行以下处理:

(a)在步骤 a)排序不变的基础上,对条件属性实值相同而决策属性值不同的样本在小范围内重新调整顺序,遵循原则是同类相吸引,异类相排斥。

(b)从决策属性不同的地方依次插入断点,将中间区间进一步离散成若干更小的区间。

e)对整个数据集 S,将相邻区间决策属性相同的进行合并。

f)给所有区间赋区间号,产生最终的离散区间。

g)i + +;重复步骤 a) ~ f)。

h)如果 i 大于 A,离散化结束。

下面用含 13 个样本的数据集 U 举例以便理解整个离散过程,如表 1、2 所示。

表 1　未离散化的决策表

U	C_1	C_2	C_3	C_4	D	U	C_1	C_2	C_3	C_4	D
1	5.4	3	4.5	1.5	2	8	5	3.8	1.9	0.4	1
2	5.5	2.3	4.1	3	2	9	5.7	2.5	5	2	3
3	5.5	2.4	3.8	1.1	2	10	5.6	2.8	4.9	2	3
4	5.7	4.4	1.5	0.4	1	11	5.7	2.6	3.5	1	2
5	5.8	2.7	3.9	1.2	2	12	5.7	2.8	4.1	1.3	2
6	5.8	4.1	2	0.2	1	13	5.1	3.8	1.5	0.3	1
7	5.1	3.8	1.6	0.2	1						

表 2　属性 C_1 离散化后的决策表

U	C_1	C_2	C_3	C_4	D	U	C_1	C_2	C_3	C_4	D
1	2	3	4.5	1.5	2	8	1	3.8	1.9	0.4	1
2	2	2.3	4.1	3	2	9	3	2.5	5	2	3
3	2	2.4	3.8	1.1	2	10	3	2.8	4.9	2	3
4	5	4.4	1.5	0.4	1	11	4	2.6	3.5	1	2
5	6	2.7	3.9	1.2	2	12	4	2.8	4.1	1.3	2
6	5	4.1	2	0.2	1	13	1	3.8	1.5	0.3	1
7	1	3.8	1.6	0.2	1						

3　实验结果与分析

为验证本文算法的有效性,对 UCI 机器学习数据库中的三个数据集(表 3)分别用 extended Chi2 算法、CACC 算法[5]和本文算法作了对比实验,实验在VC ++6.0 环境下实现。

表 3　数据信息表

数据集	连续属性	离散属性	类别数	样本数
iris	4	0	3	150
glass	9	0	7	214
bupa	6	0	2	345

将实验数据集分别用上述三种算法进行离散,将离散后的数据分别应用 C4.5[6]方法构造决策树,随机选取 80% 作为训练集,其余 20% 作为测试集,统计平均正确识别率,对比结果如表 4 所示。同时,使用 SVM 对离散后的数据用一对多(1-v-r)多类分类方法进行分类[7,8],随机选取 80% 作为训练集,其余 20% 作为测试集。模型类型选为 C-SVC,核函数类型选为 RBF 函数,penalty C: 100,gamma:0.5。由于核函数依赖于输入样本向量的内积,大的属性值容易导致计算复杂、训练时间较长。为避免上述情况发生,将训练集和测试集的属性值归一化:

$$\bar{x}_i = 2\frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} - 1$$

归一化后 $x_i \in [-1, +1]$。此归一化函数为集成分类软件 DMBench 1.0 ahpha 中 SVM 分类模块自带函数。用 SVM 分类预测精度对比结果如表 5 所示。

表4　C4.5识别实验结果

指标	算法	数据集		
		iris	glass	bupa
平均正确识别率/%	extended Chi2	93.5	59.7	45.3
	CACC	93.3	67.4	65.2
	本算法	100	55.8	96.9

表5　SVM(1-v-r)分类预测实验结果

指标	算法	数据集		
		iris	glass	bupa
预测精度/%	extended Chi2	93.3	67.4	68.1
	CACC	96.7	69.8	58.0
	本算法	96.7	65.1	66.7

从表 4 可以明显看出,本文提出的算法平均正确识别率总体较 extended Chi2 和 CACC 算法效果好,iris 数据识别率高达 100%,bupa 数据识别效果也明显优于其他两种算法。可以预测本文算法对样本数大而类别数相对少的数据集会更加适用。表 5 结果显示,分类预测精度与 extended Chi2 对比,iris 数据精度有所提高,其他略有下降;与 CACC 算法对比,bupa 数据预测精度明显提高。

4　结束语

连续属性离散化是使用粗糙集理论和方法进行知识获取中的研究热点和难点。目前离散化算法缺乏统一的理论指导,本文抓住了离散化的实质,在保证决策属性绝对不改变的前提下选取断点,产生离散区间,候选断点的选取方式采用拆分式,计算时间复杂度低于合并式(时间复杂度分析参见文献[5],这里不作详细分析);而后确定断点采用合并方式,总体离散化过程简单,易于实现,实验结果达到了非常好的离散目的,为离散化问题提供了新思路。理论和实验结果证明了该算法的有效性。

参考文献:

[1] LIU Huan, FARHAD H, LIM T C, *et al*. Discretization: an enabling technique[J]. Data Mining and Knowledge Discovery, 2002,6(4):393-423.

[2] SU C T, HSU J H. An extended Chi2 algorithm for discretization of real value attributes[J]. IEEE Trans on Knowledge and Data Engineering, 2005,17(3):437-441.

[3] KURGAN L A, CIOS K J. CAIM discretization algorithm[J]. IEEE Trans on Knowledge and Data Engineering, 2004,16(2):145-153.

[4] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982,11(5):341-356.

[5] TAI C J, LEE C I, YANG Wei-pang. A discretization algorithm based on class-attributes contingency coefficient[J]. Information Sciences, 2008,178:714-731.

[6] QUINLAN J M. C4.5: programs for machine learning[J]. Machine Learning, 1993,16(3):235-240.

[7] HSU C W, LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks, 2002,13(2): 415-425.

[8] CRISTIANINI N,SHAWE-TAYLOR J. 支持向量机导论[M]. 李国正,王猛,曾华军,等译. 北京:电子工业出版社, 2000.