广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(4):235~237
Vol. 23,No. 4 November 2007

基于启发式信息熵的粗集数值属性离散化算法

Discretization of Numerical Attributes in Rough Set Theory Based on Information Entropy with Heuristics Information

李春贵,王 萌,原庆能
LI Chun-gui,WANG Meng,YUAN Qing-neng

(广西工学院计算机系,广西柳州 545006)
(Department of Computer, Guangxi University of Technology, Liuzhou, Guangxi, 545006, China)

摘要:在一致性假设前提下,以数据集的统计性质作为启发式知识,从候选离散点集中选择离散点,根据数据集的期望值和方差来确定搜索最优离散点的区域,提出一种新的基于信息熵粗集数值属性离散化算法,并采用UCI国际标准数据集来验证新算法.新算法与已报道的算法所得到的离散断点集完全一致,决策表的离散化结果也相同,但时间代价不同,新算法比其计算效率提高40%~50%.

关键词:信息熵 粗糙集 数值属性 离散化 统计性质
中图法分类号:TP301.6;TP18 **文献标识码**:A **文章编号**:1002-7378(2007)04-0235-03

Abstract: According to the consistency assumption in machine learning, the heuristics information of the data set statistic properties is used to select the discretization points from the candidate point set, in more detail, the mean and variance of data set are used to ascertain the region for searching optimal discretization points. A novel algorithm of numerical attributes discretization based on information entropy is proposed. The testing experiment with the UCI data sets has been performed. The results of experiment show that the discretization point set selected by using the new algorithm is the same as those by using the existing algorithm, and so does the results of decision tables discretization, but the time cost is different, the computing time of the new algorithm has saved about 40%~50% compared to the existing algorithm.

Key words: information entropy, rough set, numerical attribute, discretization, statistic property

粗集(又称为粗糙集)理论是波兰数学家Z. Pawlak[1]提出的一种分析数据的数学理论,是一种基于符号的机器学习方法,主要用来处理不确定和不精确的信息.其特点是不需要预先给定某些特征和属性的数量描述,而是直接从给定问题的描述集合出发,找出该问题的内在规律,其基本思想更接近现实情况.目前该理论已经在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到了较为成功的应用.

在应用粗集理论时,一般要求实际数据构成的决策表中各个属性必须用离散值表达(这样的属性称为名字属性).如果某些属性的值域为连续时(这样的属性称为数值属性),则处理前必须经过离散化,并且要考虑到粗集理论的特殊性.

在不同的应用领域中粗集数值属性离散化都有自己独特的要求,没有统一的、通用的离散化方法.根据不同的应用情况,人们提出了很多新的离散化方法[2~5],这些方法一般可分为两大类:第一类的思想是离散化时不改变决策表的相容性,即对连续属性离散化应该在保证决策表相容性不变的条件下选择最少的断点(分点);第二类的思想则不把相容性是否改变作为指标而仅考虑数据本身的规律,这样做有可能得到较少的离散化断点集合.文献[6]的作

收稿日期:2007-08-15

作者简介:李春贵(1968-),男,副教授,博士,主要从事机器学习与智能系统研究。

*广西自然科学基金项目(桂科自0481016),广西教育厅2006年科研基金资助项目(149)资助。

者提出了一种基于信息熵的粗集连续属性离散化算法，该算法考虑了粗集理论的特殊性，在进行知识获取的实验中，效果比较好. 由于数值属性的值域是连续的，这样在处理在大样本数据集时，应用文献[6]的方法，虽然在离散化时，考虑了粗集理论的相容度，但是候选断点非常多，需要大量的信息熵的计算，计算复杂度比较高. 本文在一致性假设前提下，以数据集的统计性质作为启发式知识，从候选离散点集中选择离散点，根据数据集的期望值和方差来确定搜索最优离散点的区域，提出了一种改进的基于信息熵的数值属性离散算法，并采用UCI国际标准数据集验证了算法的有效性.

1 相关定义及理论[1]

定义1 一个近似空间(或知识库)定义为一个关系系统(或二元组) $K=(U,R)$，其中 U 是一个被称为全域或论域的所有要讨论的个体的非空集合，R 是 U 上等价关系的一个族集.

定义2 给定知识库 $K=(U,A)$，$P\subseteq A$，$P\neq\varnothing$，P 中所有等价关系的交集也是一种等价关系，称为 P 上的不可分辨关系，记为 $IND(P)$，即

$$[X]_{IND(P)}=\bigcap_{R\in P}[X]_R.$$

定义3 给定知识库 $K=(U,A)$，对于每个子集 $X\subset P$ 和一个等价关系 $R\in ind(K)$，称 $R\_(X)=\{x|x\in U,[X]_R\subseteq X\}$ 为 X 的 R 下近似集；称 $R^-(X)=\{x|x\in U,[X]_RIX\neq\varnothing\}$ 为 X 的 R 上近似集；称 $BN_R(X)=R^-(X)-R\_(X)$ 为 X 的 R 边界域.

定义4 知识表达系统定义为 $S=(U,C,D,f)$，其中 U 是数据集中对象的集合，$C\cup D=A$ 是属性的集合，子集 C 和 D 分别称为条件属性和决策属性，$V=\bigcup_{r\in A}V_r$ 是属性值的集合，V_r 表示 $r\in A$ 的属性值域，$f:U\times A\to V$ 是一个信息函数，它指定 U 中每一个对象的属性值.

设 Φ 是由决策表各个条件属性按属性值相等确定的等价关系簇，Φ 中等价关系的交仍然是一个等价关系，用 Q 表示. 用 P 表示由决策属性按属性值相等确定的等价关系. 设由 Q 确定的等价类子集簇为 $\{X_1,X_2,\cdots,X_s\}$，则可给出如下定义5.

定义5 决策表的相容度定义为 $d_p(Q)=\sum_{k=1}^{s}|P\_(X_k)|/|X|$.

设 $S=(U,A,V,f)$ 是一个决策表，只有一个决策属性 d，且为名字属性(取离散值)，其值域为 $V_d=\{1,2,\cdots,k\}$，表示有 k 个类属. 对条件数值属性 $a\in C$，设值域为 $V_a=[l_a,h_a]$，其中一组点：$l_a<c_1^a<c_2^a<\cdots<c_{n_a}^a<h_a$，则区间 $[l_a,h_a]$ 可划分为

$$V_a=[l_a,c_1^a)\cup[c_1^a,c_2^a)\cup\cdots\cup[c_{n_a}^a,h_a].$$

这样就将属性 a 的取值分为 n_a+1 个等价类，这里每一个 $c_i^a(i=1,2,\cdots,n_a)$ 就称为离散点，离散化的目的就是对所有的数值属性都找到适宜的离散点集. 令 $\tilde{f}(x,a)=0\Leftrightarrow f(x,a)\in[l_a,c_1^a)$，$\tilde{f}(x,a)=i\Leftrightarrow f(x,a)\in[c_i^a,c_{i+1}^a)$，$i=1,2,\cdots,n_a$，则可得到一个新的决策表 $(U,A,\tilde{V},\tilde{f})$，用新决策表代替原决策表进行属性约简.

2 算法

算法首先确定第1个最优离散点，根据期望值和方差的物理意义，第1个最优的离散点在以期望为中心，以方差为半径的区域内，第2个离散点在邻接第1个查找区域的直径为两倍方差的区域内查找，依此类推，并交叉地以方差为中心的左右两边区域来查找离散点.

对条件数值属性 $a\in C$，设值域为 $V_a=[l_a,h_a]$，使用的记号同文献[6]，算法叙述如下.

输入：样本数据集(决策表) U，条件属性 a，决策属性 d.

输出：离散化断点集合 P.

步骤1：$P=\varnothing$；$L=\{U\}$；$H=H(U)$；$X=U$，计算数据集 X 的期望值 m 和方差 δ；

步骤2：计算候选断点集 B；

步骤3：以 δ 为半径，m 为中心，事先划分区间 $[l_a,h_a]$，即 $V_a=\cdots\cup[m-3\delta,m-\delta)\cup[m-\delta,m+\delta)\cup[m+\delta,m+3\delta)\cup\cdots$，其中，中间每个子区间的长度均为 2δ，靠近区间端点的两个子区间的长度有可能与其它子区间不等($<2\delta$)，$[l_a,h_a]$ 划分为 k 个区间.

步骤4：对每一个 B 中的 $c\in[m-\delta,m+\delta]$，计算 $H(c,L)$，选择使 $H(c,L)$ 最小的断点 $c_{\min}$，若 $H<\min\{H(c,L)\}$，则结束；否则，$H=H(c_{\min},L)$，$c_{\min}$ 加入到 P 中，$c_{\min}$ 把等价类 X 划分为子集 X_l 和 X_r，从 L 中去掉 X，把等价类 X_l 和 X_r 加入.

步骤5：继续计算断点集 P.

1° $i=1$；

2° 对每一个 B 中的 $c\in[m-(2i+1)\delta,m-(2i-1)\delta]$ 或 $c\in[m+(2i-1)\delta,m+(2i+1)\delta]$(交叉进行)，计算 $H(c,L)$；

3° 选择使 $H(c,L)$ 最小的断点 c_{min},若 $H < H(c_{min},L)$,结束;否则,

4° 把 c_{min} 加到 P 中,$H = H(c_{min},L)$,对所有的 $X \in L$,c_{min} 把等价类 X 划分为子集 X_l 和 X_r,那么,从 L 中去掉 X,把等价类 X_l 和 X_r 加入到 L 中.

5° 如果 L 中各个等价类中的实例都具有相同的决策,结束;否则,$i = i + 1$,转到 2°.

可以看出,新算法改进了文献[6]的算法的离散点选择和计算过程.由于使用了期望值和方差作为启发式知识,新算法仅对属于某一区间段的候选断点进行搜索,计算复杂度也有所减小,在候选离散点均匀分布的情况下,新算法的复杂度是文献[6]的算法的 $2\delta/(h_a - l_a)$.

3 模拟实验

用含有数值型属性的数据集进行离散化实验来检验新算法的有效性和性能,数据来源于 UCI 标准测试数据集[7],所有算法用 Matlab 7 来实现,测试计算机的基本配置为 CPU P4 2.66 GHz,内存 512M.在实验中,两种算法所得的离散断点集完全一致,决策表的离散化结果相同,时间代价不同.断点计算结果见表 1.从表 1 结果可以看出,新算法比文献[6]的算法计算效率提高了 40%~50%.

表 1 断点计算结果

数据集	样本数	属性数	数值属性数	断点数	计算时间(s)*		比值 t_1/t_2
					t_1	t_2	
iris	150	5	4	6	1.344	2.188	0.614
clev	297	14	5	9	5.625	11.000	0.511
glass	214	11	10	11	15.532	25.531	0.608

* t_1 为本文的新算法,t_2 为文献[6]的算法.

4 结束语

把粗集理论应用于含数值属性的决策表时,要事先对数值属性进行离散化,在离散化过程中,是否保持决策表的相容性是一项重要的评价指标,本文研究的基于信息熵的粗集数值属性的离散化方法,不改变决策表的相容性.在已有算法的基础上,通过引入样本数据的统计性质作为启发式知识,把原来的全局搜索改为局部搜索,提出改进的算法并进行实验。实验结果表明,新算法的计算效率有较明显的提高.从本质上来说,新算法仍然具有多项式级的复杂度,在解决大规模数据时,计算复杂度还很高.粗集数值属性的离散化问题,仍有待进一步的研究和探讨.

参考文献:

[1] PAWLAK Z. Rough sets[J]. Communications of ACM, 1995,38(11):89-95.

[2] DAI JIANHUA,LI YUANXIANG. Study on discretization based on rough set theory:proceedings of the first International Conference on Machine Learning and Cybernetics[C]. Beijing:IEEE Press, 2002:1371-1373.

[3] 李兴生.一种基于云模型的决策表连续属性离散化方法[J].模式识别与人工智能,2003,16(3):33-38.

[4] 曾建武,张建明,王树青.基于人工鱼群算法的离散化方法[J].模式识别与人工智能,2006,19(5):611-616.

[5] 张永,丁洪昌.连续属性离散化的 MaxDiff 方法[J].计算机工程与应用,2007,43(19):80-82.

[6] 谢宏,程浩忠,牛东晓.基于信息熵的粗糙集连续属性离散化算法[J].计算机学报,2005,28(9):1570-1573.

[7] ASUNCION A,NEWMAN D J. UCI machine learning repository [DB/OL]. [2007-07-15]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

(责任编辑:尹 闯)

淡化海水可能不利于农作物生长

缺水现象在世界各地愈来愈普遍,海水淡化是解决缺水的主要方向之一。不过,海水淡化不仅要考虑食用的问题,也必须考虑到灌溉上的用途。在灌溉时使用淡化海水,需要考虑如何才能有利于农作物生长。以色列是全球使用淡化海水浇灌农作物最多的国家之一。以色列研究人员最近发表报告说,由于淡化海水中矿物质含量通常较低,使用淡化海水灌溉可能不利于某些农作物生长,例如番茄和有些花卉就会受到影响。研究报告指出,用来灌溉的淡化海水如果缺钙或镁,会影响植物的生长。如果植物是种在沙土或无土的状况下,使用淡化海水灌溉对植物的影响会更大,因为植物无法从土壤中吸取所需的矿物质。

(据科学网)