第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)04—0077—02 文献标识码：A 中图分类号：TP18

一种改进的 CAIM 算法

李 慧，闫德勤，张迎春
(辽宁师范大学计算机与信息技术学院，大连 116081)

摘 要：在 CAIM 算法中，离散判别式仅考虑了区间中最多的类与属性间的依赖度，使离散化过度而导致结果不精确。基于此，提出对 CAIM 的改进算法，该算法考虑到按属性重要性从小到大顺序进行离散，同时根据粗糙集理论提出条件属性可分辨率概念，与近似精度同时控制信息表最终的离散程度，有效解决了离散化过度问题。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和分类预测，结果证明了该算法的有效性。

关键词：连续属性离散化；粗糙集；属性可分辨率

Modified Algorithm of CAIM

LI Hui, YAN De-qin, ZHANG Ying-chun
(School of Computer and Information Technology, Liaoning Normal University, Dalian 116081)

【Abstract】 In Class-Attribute Interdependency Maximization(CAIM) algorithm, discretization criterion only accounts for the trend of maximizing the number of values belonging to a leading class within each interval. The disadvantage makes CAIM generate irrational discrete results and further leads to the decrease of predictive accuracy of a classifier. This paper proposes a modified algorithm of CAIM. With the algorithm, the importance of attributes is adopted in discretization process, and a concept of attribute discernibility rate is proposed based on rough set. Both attribute discernibility rate and approximate quality are used for discretization intervals, which effectively resolve the problem of over-discretization. By using C4.5 and SVM, experiments are performed respectively with the results of discreted data, which show that the presented algorithm is effective.

【Key words】 discretization of continuous attributes; rough set; attribute discernibility rate

1 概述

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。离散化是把连续属性的取值范围或取值区间划分为若干个数目不太多的小区间，其中每个小区间对应着一个离散的符号。大多数离散化算法是基于统计学或基于信息熵的，如 Entropy-MDLC, Extended-Chi2 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。针对文献[1]提出基于信息理论的类-属性间最大相互依赖(Class-Attribute Interdependency Maximization, CAIM)的连续属性离散化算法存在的不足，本文提出对 CAIM 的改进算法。

2 粗糙集理论

2.1 基本概念

设 $S=(U,A,V,F)$ 为一个信息系统，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 $A=C\cup D$，$C\cap D=\phi$，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 2 个等价关系簇，Q 的 P 正域记为 $POS_P(Q)$，定义为

$$POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}(X)$$

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

$$\gamma_P=\sum_{i=1}^{k}card(\underline{P}Y_i)/card(U)$$

其中，$card(\ )$ 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。

定义 4 设 $S=(U,A,V,F)$ 是一个决策表，条件属性子集 $B\subset C$，任意条件属性 $a\in C$ 相对于条件属性集合 B 对决策属性集合 D 依赖程度的属性重要度定义为

$$sgf(a,B,D)=\gamma_{B+\{a\}}-\gamma_B$$

2.2 条件属性可分辨率概念

粗糙集理论中一个重要的观点——将知识与区分事物能力对应起来，即知识就是区分事物的能力。在论域中，若任意 2 个对象都能被区分，那么其含有的知识最大；若所有对象都能被划为一个等价类(定义 1)，那么其含有的知识最少。文献[2]对知识进行量化，证明了量化的合理性，以量化后的区分能力即知识量作为启发函数指导属性约简。本文受到文献[2]的启发，根据知识量的含义，提出了条件属性可分辨率概念。过去的离散化算法区间是否被合并或拆分起最终评定作用的是粗糙集的经典标准模型近似精度(定义 3)，本文在近

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基金资助项目(2007A10GX117)
作者简介：李 慧(1980－)，女，硕士研究生，主研方向：数据挖掘，粗糙集理论；闫德勤，教授、博士；张迎春，硕士研究生
收稿日期：2009-08-10 **E-mail：**huili_913@yahoo.com.cn

似精度起控制作用的同时，又考虑了条件属性可分辨率是否下降，这无疑是为离散化提供了双保险，有效控制离散化过度。下面给出条件属性可分辨率推导过程及公式：

如果论域 U 含有 N 个对象，某属性集合将论域分成 m 个等价类，每个等价类含有对象(元素)个数分别为 $n_1,n_2,\cdots,n_m$，那么该属性集合具有的知识量 $W(n_1,n_2,\cdots,n_m)=W(1,1)\times\sum_{1\le i<j\le m}n_i\times n_j$ [2]，其中，$W(1,1)$ 是常数，本文取值为 2。

如果论域 U 含有 N 个对象，若任意 2 个对象都能被区分，那么近似精度等于 1，将会有 $\frac{1}{2}N(N-1)$ 个可分辨对，即含有 N 个对象的信息表的可分辨对的个数达到最大值：$\frac{1}{2}N(N-1)$，将最大可分辨对数乘以 $W(1,1)$ 就是信息表所含最大知识量。

条件属性可分辨率是信息表某一个条件属性集所含有的知识量占整个信息表最大知识量的百分比。

用 Discernibility rate 的首字母 *Dr* 表示条件属性可分辨率。则有条件属性可分辨率代数表达式：

$$Dr=\frac{W(1,1)\times\sum_{1<i<j<m}n_i\times n_j}{W(1,1)\times\frac{1}{2}N(N-1)}=\frac{2\sum_{1\le i<j\le m}n_i\times n_j}{N^2-N}\quad Dr\in[0,1] \tag{1}$$

3 CAIM 算法存在的不足及新算法

3.1 CAIM 算法

2004 年，Lukasz A Kurgan 和 Krzysztof J.Cios 提出了基于信息理论的类-属性间最大相互依赖(Class-Attribute Interdependency Maximization, CAIM)的连续属性离散算法。该算法的离散判别式如下：

$$CAIM(C,D\mid A)=\frac{\sum_{r=1}^{n}\frac{\max_r^2}{M_{+r}}}{n} \tag{2}$$

其中，n 为目前区间的数目；$\max_r$ 为第 r 个区间中最大的类别数。该标准试图使类与属性之间的相互依赖程度最大化，达到最小化区间数目的目的。此算法有 2 个缺点：(1)它仅仅考虑了区间中最多的类与属性间的依赖度；(2)最终产生的离散方案中区间的数目非常接近类的个数，用 Age dataset(见表 1)为例，通过该算法将数据集离散为 3 个区间 [3.00,10.50], (10.50, 61.50)和(61.50, 71.00]，显然这个结果并不好，理想区间应该是 5 个：样本 1 到 3，4 到 6，7 到 9，10 到 12 和 13 到 15。因此很容易得出结论：CAIM 算法会使得离散化过度而导致结果不精确。

表 1　Age dataset

ID	Age	class	ID	Age	class
1	3	Care	9	46	Work
2	5	Care	10	51	Edu
3	6	Care	11	56	Edu
4	15	Edu	12	57	Edu
5	17	Edu	13	66	Care
6	21	Edu	14	70	Care
7	35	Work	15	71	Care
8	45	Work			

3.2 本文算法

在信息系统中，重要属性(定义 4)对决策划分的影响很大，对决策属性而言比较重要。针对 CAIM 算法离散化过度导致结果不精确的不足，本文对 CAIM 算法的改进之处如下：(1)算法考虑到离散化时所有条件属性的顺序问题，以往的离散化算法一般都按数据集中条件属性的自然排序进行，事实上离散的顺序对离散后的信息表示有一定影响，本文依据属性重要性从小到大进行离散，属性重要性依据标准采用文献[3]中的 CAIR 表达式(见式(3))；(2)用本文提出的条件属性可分辨率与近似精度共同控制离散进程，在保证近似精度不下降的基础上，把条件属性可分辨率的下降程度控制在合理范围内，这样使得区间的拆分更为合理，最大程度地保持信息表示的意义，减少信息损失。

$$CAIR(C,D\mid A)=\frac{I(C,D\mid A)}{H(C,D\mid A)} \tag{3}$$

其中，$I(C,D\mid A)=\sum_{i=1}^{s}\sum_{r=1}^{n}p_{ir}\cdot\mathrm{lb}\left(p_{ir}/(p_{i+}p_{+r})\right)$ 是互信息表达式；$H(C,D\mid A)=\sum_{i=1}^{s}\sum_{r=1}^{n}p_{ir}\cdot\mathrm{lb}(1/p_{ir})$ 是 Shannon 熵表达式；*CAIR*(*C*,*D*|*A*)的值越大，表明属性 *A* 越重要，离散程度相对应小些；反之亦然。

离散化算法描述如下：

输入 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

步骤：

(1)根据式(3)计算 every A_i 的 CAIR 值；

(2)按属性的 CAIR 值从小到大顺序排序；

(3)选择所有可能的断点为初始候选断点，然后确定候选断点，产生离散区间；//此过程均保留 CAIM 算法中的标准[1]

(4)if (inconsistency()>0&& $Dr(A_i)<O\_Dr(A_i)*C$))

进一步拆分成更小的区间；

// $Dr(A_i)$表示离散化后属性 A_i 的可分辨率；$O_Dr(A_i)$表示属性 //A_i 初始的可分辨率；C 的取值根据属性重要性不同略有变化，本文 //实验中取的 $C\in[0.95,0.98]$

(5)output 新的离散化的属性 A_i 的 k 个区间；

(6)为同在一个区间中的属性值赋相同符号；

(7)离散化结束。

4 实验结果与分析

为说明改进算法的有效性，用本文提出的算法、CAIR 算法和 CAIM 算法分别对 UCI 机器学习数据库中的 5 个数据集(见表 2)进行离散化，实验在 VC++6.0 环境下实现。

表 2　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
breast	9	0	2	683
wine	13	0	3	178
sonar	60	0	2	208
vehicle	18	0	4	846

将离散后的数据应用 C4.5 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。对统计正确识别率进行对比(见表 3)。

表 3　C4.5 正确识别率　(%)

算法	数据集				
	iris	breast	wine	sonar	vehicle
CAIR	83.3	90.8	78.8	64.3	42.0
CAIM	90.0	93.4	91.7	69.0	46.5
本文算法	93.7	92.2	86.1	71.7	57.6

同时，使用 SVM 对离散数据用“一对多(1-V-r)”多类分类方法进行分类[4]，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，因此大的属性值容易导致计算复杂，训练时间较长，为了避免上述情况发生，将训练集和测试集的属性值归一化：

(下转第 81 页)