

基于 CACC 的连续数据离散化改进算法

刘小龙¹, 江虹¹, 吴丹²

(1. 西南科技大学信息工程学院, 四川 绵阳 621010; 2. 四川航天职业技术学院, 成都 610100)

摘要: 针对粗糙集及主要机器学习算法一般都无法高效处理连续数据的问题, 提出一种基于 CACC 的连续数据离散化的改进算法。该算法采用 CACC 标准选取断点, 通过增加数据不一致率约束条件, 从而减少数据丢失信息量。仿真结果表明, CACC 改进算法与 Modified Chi2、Extent-Chi2、CAIM、CACC 算法相比, 并通过 C4.5 和 SVM 算法验证, 数据识别率和精度可提高近 8%。

关键词: 粗糙集; 离散化; 重要属性; 不一致率; CACC 改进算法; 精度

Improved Algorithm Based on CACC for Discretization of Continuous Data

LIU Xiao-long¹, JIANG Hong¹, WU Dan²

(1. School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China;

2. Sichuan Aerospace Polytechnic, Chengdu 610100, China)

【Abstract】 Aiming at the problem that rough set and the main machine learning algorithms can not efficiently handle continuous data, this paper presents an improved CACC algorithm for discretization of the continuous data. This algorithm adopts the CACC standard to select breakpoints to increase constraints on data inconsistency, thereby reducing the amount of information loss. Simulation results show that the algorithm outperforms the corresponding algorithms, such as Modified Chi2, Extent-Chi2, CAIM, CACC, through the C4.5 and SVM algorithm validation, the maximum amplitude of data recognition rate and accuracy is increased by 8%.

【Key words】 rough set; discretization; important attributes; inconsistency; CACC improved algorithm; accuracy

DOI: 10.3969/j.issn.1000-3428.2013.04.012

1 概述

粗糙集理论^[1]建立在论域不可分辨关系之上, 它不依赖于所需处理数据之外任何先验信息, 是继概率论、模糊理论后一种定量分析处理不精确、不一致、不完整信息与知识的数学工具。近年来, 粗糙集理论在数据挖掘、人工智能、模式识别和决策分析等领域取得了很多成功的应用, 在诸如故障诊断、医疗诊断、信号分类分析等领域有巨大应用前景。

经典 Pawlak 模型中的不分明关系是一种等价关系, 对于名义型及离散型数据有较好的描述及分辨能力。但实际处理的数据大多是连续型属性值, 因此, 将连续数据离散化是粗糙集的关键研究内容之一。连续属性值离散化本质是将具有相关性较大的数据划分在一个区间中, 用离散的数值或字母代替。由于离散过程中不可避免地会对原始数

据产生信息损失, 离散化应保证在原始数据信息量丢失最小的前提下, 划分出较小的区间个数。

连续值属性离散化算法具有不同划分, 如动态和静态、监督和非监督、全局和局部、自顶向下和自底向上等。典型离散化算法, 如等宽及等频^[2]非监督离散算法, 需预先设定离散区间个数, 离散效果丢失信息大, 分类精度低; 基于信息熵的 D2、Ent-MDLP^[3]算法, 计算复杂, 离散时间较长; 基于统计 χ^2 分布的 ChiMerge、Extended Chi2^[4]算法, 使用卡方统计来判断区间是否被合并, 采用显著性水平值逐渐降低的方法检验不一致率, 判定离散是否终止; 基于类别和属性关联程度的 CADD^[3]、CAIM^[3]、CACC^[5]等算法, 这类算法不断选取能够使类和属性相互依赖程度最大的断点, 需保证离散区间不小于类别数。

现有离散算法大多对单个属性进行离散, 离散中未考虑数据整体信息, 因此离散过程复杂, 离散结果信息丢失

基金项目: 国家自然科学基金资助项目“认知无线电智能学习与决策关键技术研究”(61072138)

作者简介: 刘小龙(1987—), 男, 硕士研究生, 主研方向: 智能算法, 数据挖掘; 江虹, 教授、博士; 吴丹, 助教

收稿日期: 2012-07-04 **修回日期:** 2012-08-27 **E-mail:** qjuxueerlai@yahoo.cn

较大。本文提出一种改进的 CACC 连续数据离散化算法, 使得整体数据信息丢失少, 类和属性依赖可达到最大。

2 粗糙集相关概念

粗糙集核心思想是用上下近似来描述事物不确定性。经典粗糙集中若能用基本等价粒度来表示整个集合, 则该集合能被精确分辨; 否则需用上下近似来表示该集合的粗糙度。

2.1 信息系统

设 $S=(U, A, V, f)$ 为一个信息系统^[6]。 $U=\{U_1, U_2, \dots, U_{|U|}\}$ 为论域样本集合; $A=\{a_1, a_2, \dots, a_{|A|}\}$ 为属性集合。若将 A 中的属性分为不相交条件属性集 C 和决策属性集 J , 即 $A=C \cup J, C \cap J = \emptyset$, 则 S 也称为决策表。 $V=\cup V_a$, 其中, $a \in A, V_a$ 为属性 a 值域; $f: U \times A \rightarrow V$ 为信息函数, 对 $a \in A, x \in U, f(x, a) \in V_a$, 确定 U 中每一样本的属性值。

2.2 不可分辨关系

对任一子集 $P \neq \emptyset$, 其在 U 上的不可分辨关系 I 定义为: $I(P)=\{(x, y) \in U \times U: f(x, q)=f(y, q) \forall q \in P\}$, 若 $(x, y) \in I, x$ 和 y 则不可分辨, 也称基本集。

2.3 下、上近似

设 $X \subseteq U, B \subseteq A, X$ 关于 B 的下近似^[7]表示为 B 基本集被 X 包含的并: $\underline{BX}=\{U[x_i]_{(B)} \subseteq X\}$ 。 X 关于 B 的上近似^[7]表示为 B 基本集与 X 有相交的并: $\overline{BX}=\{x_i \in U \mid [x_i]_{(B)} \cap X \neq \emptyset\}$ 。

3 数据离散化算法

χ_a^2 是初始设定阈值, ChiMerge^[8]离散算法如式(1)计算置信度对区间合并:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^S \frac{(h_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

其中, S 是决策类别; h_{ij} 是 i 区间的 j 类样本数目; $E_{ij}=R_i \times C_j / N, R_i$ 为 i 区间样本数, C_j 为相邻两区间 j 类样本数, N 为总样本数。Chi2 算法动态设定阈值 t 为 $\chi_a^2 - \chi^2$, Modified Chi2 算法自由度 ν 由当前断点相邻的 2 个区间中决策类的个数决定。改进的 Extended Chi2 对阈值 t 除以 $\sqrt{2\nu}$ 动态提高临界值。该类算法基于统计, 理论性强, 但计算复杂, 需查表比较其置信范围。

设数据集样本数为 N , 类别为 S, a 为数据集中的单个连续属性, 利用离散方案 $D: \{[d_0, d_1] [d_1, d_2] \dots [d_{k-1}, d_k] \dots [d_{n-1}, d_n]\}$ 将连续属性 a 的值离散到 n 个区间中, 并使属性 a 的每一个样本只划分在一个区间中。 $\{[d_0, d_1, \dots, d_k, \dots, d_n]\}$ 为离散过程中的所有离散断点, 在不同离散方案下, 基于类别 S 和离散方案 D 得到二维量化矩阵表, 如表 1 所示。 d_0, d_n 为属性 a 的最值。 h_{ik} 为第 k 区间中第 j 类样本数, N_{i+} 为属性 a 中第 i 类样本数, N_{+k} 为属性 a 中第 k 区间中的样本数, $i=1, 2, \dots, m; k=1, 2, \dots, n$ 。

表 1 基于类与属性的二维量化矩阵

决策类别	区间						每类样本个数
	$[d_0, d_1]$	$[d_1, d_2]$	\dots	$[d_{k-1}, d_k]$	\dots	$[d_{n-1}, d_n]$	
S_1	h_{11}	h_{12}	\dots	h_{1k}	\dots	h_{1n}	N_{1+}
S_2	h_{21}	h_{22}	\dots	h_{2k}	\dots	h_{2n}	N_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
S_i	h_{i1}	h_{i2}	\dots	h_{ik}	\vdots	h_{in}	N_{i+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
S_m	h_{m1}	h_{m2}	\dots	h_{mk}	\dots	h_{mn}	N_{s+}
每个区间样本个数	N_{+1}	N_{+2}	\dots	N_{+k}	\dots	N_{+n}	N

二维量化矩阵信息表 Shannon 熵定义为:

$$H(S, D | a) = \sum_{i=1}^m \sum_{k=1}^n p_{ik} \text{lb} \frac{1}{p_{ik}} \quad (2)$$

其中, $p_{ik}=h_{ik}/N$ 为第 i 类样本在第 k 区间中出现的概率。信息熵度量系统的不确定度, 不确定性越大, 熵越大, 把它分辨清楚所需的信息量也越大。一个系统越是有序, 信息熵就越低。

离散方案 D 在属性 a 上得到的区间和类别 S 的互信息定义为:

$$MI(S, D | a) = \sum_{i=1}^m \sum_{k=1}^n p_{ik} \text{lb} \frac{p_{ik}}{p_{i+} p_{+k}} \quad (3)$$

其中, $p_{i+}=N_{i+}/N$ 为第 i 类样本出现的概率; $p_{+k}=N_{+k}/N$ 是第 k 区间样本出现的概率。

文献[3]提出了 CADD 算法, 其离散化的评价标准定义为 CAIR, 如式(4)所示:

$$\max \text{CAIR} = MI(S, D | a) / H(S, D | a) \quad (4)$$

该标准体现了类与属性之间的相互依赖程度, 其值越大, 表明离散方案的区间中样本不确定越小, 类与属性之间的依赖度越大, 离散方案越好。但 CADD 算法需预先指定区间数目, 可能选择较差离散点, 训练时间较长, 离散不充分。

基于 CAIR 的理念, Kurgan 等提出 CAIM 算法, 其离散化的评价标准定义如式(5)所示:

$$\max \text{CAIM} = \sum_{k=1}^n \frac{\max x_k^2}{N_{+k}} / n \quad (5)$$

其中, $\max x_k$ 表示 k 区间中最大样本数 h_{ik} ; n 是离散区间数。CAIM 不用预先设定区间数目, 自适应选取最小区间和最多类样本的最佳组合, 最终使 CAIM 值达到最大。但该算法过于偏重区间中样本数最多的类, 容易导致离散区间过少, 离散不足。

文献[5]提出了 CACC 算法, 该算法的离散化评价标准如式(6)所示:

$$\max \text{CACC} = \sqrt{\frac{y}{y+N}} \quad (6)$$

其中, $y = N \left[\left(\sum_{i=1}^m \sum_{k=1}^n \frac{h_{ik}^2}{N_{i+} N_{+k}} \right) - 1 \right] / \log(n)$ 。该算法考虑了

数据在属性中的整体分布，并使用 $\log(n)$ 代替 CAIM 中 n ，能加快离散化的进程，防止出现离散过度，该算法对单个属性能获取较好断点。

4 CACC 改进算法

由于 CACC 标准将数据类和属性相关程度最大化，仅对单个属性实现区间离散结果最优，选取断点没有描述数据整体信息，没有考虑与原始数据的不一致率，会造成原始数据信息量的丢失。本文算法对 CACC 进行改进，实现断点的较优选择，并且使原始数据信息丢失量尽量小，在 CACC 算法基础上，定义重要属性、不一致率来约束。

连续数据属性为 a ，决策类别为 S ，样本个数为 N ，定义重要属性 sig 为定理 1，不一致率 f 为定理 2。

定理 1 重要属性

if $S(i+1)-S(i) \neq 0$ 且 $i < N, K++$ ，则 $sig(a)=K$ 。

由于 CACC 算法中处理单个属性 a 时，数据按升序排列选取断点，升序数据对应决策类别 S ，类别变化频次 K 高，说明将该属性分辨清楚需要断点越多，因此该属性越重要。

定理 2 不一致率

$(x,y) \in U_i \times U_i: x,y \in I(U_i), \text{if } J(x) \neq J(y)$

$G = \text{sum}(\text{length}(I(U_i)))$ ，则 $f = G/N$ 。

其中， $I(U_i)$ 为论域的第 i 个不可分辨的集合； f 表示属性值相同但决策类别不一致的样本 G 占样本数 N 的比率。

4.1 CACC 改进算法

算法流程描述如图 1 所示。

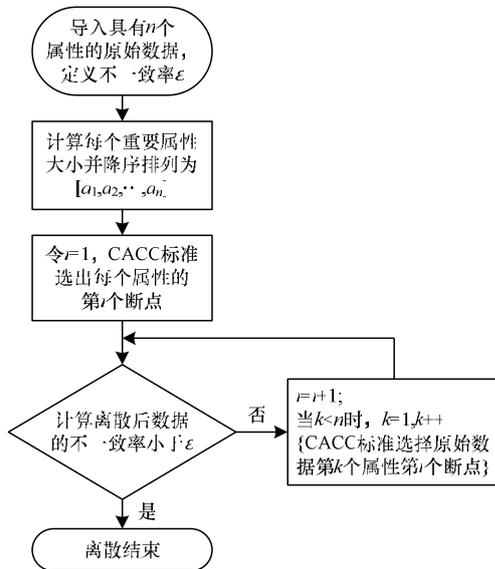


图 1 CACC 改进算法流程

CACC 改进算法描述如下：

输入 实验数据集样本数 N ，决策类别个数 S ，连续属性 a

(1)按定理 1 计算重要属性，降序排列属性为 $[a_1, a_2, \dots, a_n]$ ，设定不一致率 $\epsilon = 1 \times 10^{-2}$ 。

(2)设初始属性 $k=1$ ，划分断点数 $i=1$ ，对样本数据在属性 a_k 上的值升序排列得到 D' ，获得样本实例在属性 a_k 上的最小值 d_0 和最大值 d_n 。

(3)计算排序后 D' 集中两两相邻元素的中点，组成断点候选集 L 。

(4)初始化断点集 L' 为 $[d_0, d_n]$ ，全局最大值 CACCvalue 值为 0。

(5)将断点候选集 L 中不属于断点集 L' 的断点值逐个添加到 L' 中；并根据 CACC 评判标准计算对应离散化方案的 CACCvalue 值。

(6)选择 CACCvalue 值最大的断点，保存到 L' 中。

(7)所有属性选取第 1 个断点后，根据定理 2 计算数据不一致率；如果结果不满足设置条件， $i=i+1$ ，当 $k < n$ ，继续执行步骤(6)，选取第 k 个属性的 i 个断点， $k=k+1$ ；满足设置条件则停止。

4.2 实验结果

本文对数据进行归一化^[9]预处理，将数据转化到区间 $[0,1]$ 中，减小计算复杂度，如式(7)所示：

$$\tilde{x}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{7}$$

实验处理数据从 UCI^[10]中下载，如表 2 所示。

表 2 实验使用数据

数据集	样本数	连续属性	离散属性	类别数	离散前区间总个数
Iris	150	4	0	3	60
Breast	683	9	0	2	81
Wine	178	13	0	3	723
Auto	392	5	2	3	417
Bupa	345	6	0	2	262
Machine	209	7	0	8	197
Pima	768	8	0	2	865
Glass	214	9	0	6	702

本文定义重要属性与文献[11]对比结果如表 3 所示，重要属性度量按降序排列。

表 3 重要属性验证结果

数据集	文献[11]处理结果	定理 1 处理结果
Iris	{3,4,1,2}	{3,4,2,1}
Breast	{3,8,4,9,2,6,5,7,1}	{8,9,6,3,4,1,7,2,5}
Wine	{7,13,12,10,1,11,6,2,8,4,9,5,3}	{7,10,13,12,1,11,2,6,4,9,5,8,3}
Auto	{2,3,1,4,5}	{2,1,4,5,3}
Bupa	{5,6,1,2,3,4}	{1,6,4,5,2,3}
Machine	{4,5,6,7,3,2,1}	{2,4,5,3,6,1,7}
Pima	{8,1,3,5,7,6,2,4}	{8,1,3,4,5,6,7,2}
Glass	{1,2,6,5,9,4,8,7,3}	{9,3,5,1,6,8,7,4,2}

根据数据冗余原则，属性重要度大的不同属性组合能够代替所有属性表示的数据信息量。因此结果表明，定理 1 重要属性评价标准对本文算法是有效的。本文算法与 Modified Chi2、Extended Chi2、CAIM、CACC 对上述数据处理后得到相应断点，离散区间个数如表 4 所示。基于统

计的 Modified Chi2、Extended Chi2 算法得到数据的离散区间最多, 离散区间最少的主要集中在 CAIM 与 CACC, 如表中加底线数据所示。CACC 与 CACC 改进算法获取的断点区间数相差不大, 更靠近 CAIM 算法的最小离散区间。

表4 离散算法离散区间个数

数据集	Modified Chi2	Extended Chi2	CAIM	CACC	Improved CACC
Iris	30	30	<u>12</u>	<u>12</u>	17
Breast	33	33	<u>18</u>	<u>18</u>	26
Wine	53	53	<u>39</u>	41	41
Auto	78	76	<u>15</u>	54	72
Bupa	78	78	<u>12</u>	<u>12</u>	51
Machine	48	<u>45</u>	49	47	47
Pima	99	99	24	<u>22</u>	45
Glass	83	64	<u>54</u>	74	57
区间数均值	62.75	59.75	<u>27.875</u>	35	44.500

离散结果使用 C4.5 和 SVM 算法进行精度验证, 结果如表 5、表 6 所示。

表5 C4.5 正确识别率

数据集	Modified Chi2	Extended Chi2	CAIM	CACC	Improved CACC
Iris	0.916 7	0.916 7	0.918 3	0.936 7	<u>0.938 3</u>
Breast	0.925 5	0.925 5	0.925 5	<u>0.932 7</u>	<u>0.932 7</u>
Wine	0.802 8	0.802 8	0.844 4	0.934 3	<u>0.966 7</u>
Auto	0.762 7	0.771 5	0.790 5	0.815 7	<u>0.893 0</u>
Bupa	0.452 9	0.452 9	0.452 9	0.456 7	<u>0.531 9</u>
Machine	0.773 8	0.773 8	0.790 5	0.776 2	<u>0.793 3</u>
Pima	0.618 2	0.618 2	0.618 2	<u>0.649 7</u>	<u>0.649 7</u>
Glass	0.405 8	0.511 6	0.536 0	0.564 6	<u>0.587 2</u>
平均识别率	0.707 3	0.721 6	0.734 5	0.758 3	<u>0.786 6</u>

表6 SVM 分类预测精度

数据集	Modified Chi2	Extended Chi2	CAIM	CACC	improved CACC
Iris	0.933	0.933	0.933	0.978	0.967
Breast	<u>0.985</u>	<u>0.985</u>	<u>0.985</u>	0.956	<u>0.985</u>
Wine	0.972	0.972	0.933	<u>0.981</u>	0.972
Auto	<u>0.759</u>	0.696	0.747	0.754	<u>0.759</u>
Bupa	0.681	0.681	0.681	0.683	<u>0.710</u>
Machine	0.690	0.690	0.881	0.825	<u>0.976</u>
Pima	0.747	0.747	0.747	0.747	<u>0.753</u>
Glass	0.674	0.674	0.558	<u>0.785</u>	0.705
平均精度	0.805 1	0.797 3	0.808 1	0.838 6	<u>0.853 4</u>

验证结果显示 Improved CACC 算法的正确识别率在 C4.5 算法数据识别率是最高的, 数据 Pima 的精度在 CACC 算法下得到同一个最高值, 平均识别率比最低的 Modified Chi2 高出 7.93%; 在使用 SVM 算法中选择 RBF 核函数, 核函数中 gamma 和惩罚因子 c 用网格搜索法选取; 采用五折交叉验证(80%数据为训练集, 20%为测试集)数据精度; 结果表明, Iris、Wine、Glass3 组数据精度在 CACC 算法达到最高, 其余数据精度在改进后的 Improved CACC 算法下达到最高, 精度比最低的 Extended Chi2 高出 5.61%。由此

可见, Improved CACC 算法划分的离散区间虽比 CAIM 稍多, 但其精度在几种算法中相对最好。

5 结束语

随着数字化时代的发展, 人们需要快速地从海量数字信息中获取有效信息。在考虑整体数据信息的前提下, 为有效减少数据信息的丢失, 本文提出了一种改进的 CACC 离散化算法。仿真结果表明, 该算法在获取适当断点的同时能保证原始的数据信息, 与 Modified Chi2、Extended Chi2、CAIM、CACC 等算法相比, 数据的识别率和预测精度都有较大的提高, 取得了较好的效果。在此基础上, 今后将针对大数据的在线处理在实时性和精度方面获取优化组合进行研究。

参考文献

- [1] 苗夺谦, 姚一豫, 王国胤, 等. 不确定性与粒计算[M]. 北京: 科学出版社, 2011.
- [2] 李慧. 基于粗糙集理论的连续属性离散化算法研究[D]. 大连: 辽宁师范大学, 2010.
- [3] 杨萍, 杨天社, 李济生, 等. 一种基于类别属性关联程度最大化离散算法[J]. 控制与决策, 2011, 26(4): 592-597.
- [4] Su Chaoton. An Extended Chi2 Algorithm for Discretization of Real Value Attributes[J]. IEEE Trans. on Knowledge and Data Engineering, 2005, 17(3): 437-441.
- [5] Tsai Cheng-Jung, Lee Chien-I. A Discretization Algorithm Based on Class-attribute Coefficient[J]. Information Sciences, 2008, 178(3): 714-731.
- [6] 肖大伟, 王国胤, 胡峰. 一种基于粗糙集理论的快速并行属性约简算法[J]. 计算机科学, 2009, 36(3): 208-211.
- [7] 江虹, 伍春, 包玉军, 等. 基于粗糙集的认知无线网络跨层学习[J]. 电子学报, 2012, 40(1): 155-161.
- [8] 赵长雷. 数据挖掘中属性离散化方法研究[D]. 大连: 大连理工大学, 2010.
- [9] 董卓, 朱永利, 胡资斌. 基于遗传规划和数据归一化的变压器故障诊断[J]. 电力科学与工程, 2011, 27(9): 31-34.
- [10] Li Min, Deng Shaobo. An Effective Discretization Based on Class-attribute Coherence Maximization[J]. Pattern Recognition Letters, 2011, 32(5): 1962-1973.
- [11] Hu Qinghua. Information-preserving Hybrid Data Reduction Based on Fuzzy-rough Techniques[J]. Pattern Recognition Letters, 2006, 27(5): 414-423.