

统计指标离散化方法及应用

陈 浩

(天津科技大学 电子信息与自动化学院, 天津 300222)

摘 要: 连续数据离散化能够提高数据挖掘算法的分类能力. 文中提出一种基于统计指标的连续属性离散化方法, 凭借相关系数衡量类与属性间的关联度, 获取最优区间列表. 引入变精度粗糙集模型, 有效地控制数据由离散化导致的信息丢失. 该方法在声纳传感器数据识别以及其它领域上进行了应用. 实验结果表明, 该方法在 J48 决策树上有很好的分类能力.

关键词: 离散化; 数据挖掘; 关联度; 变精度粗糙集; 决策树

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2011)11-0106-04

A Discretization Method Based on Statistical Indicator

CHEN Hao

(School of Electronic Information and Automation, Tianjin University of Technology, Tianjin 300222, China)

Abstract: Discretization of continuous data can improve classification ability of data mining algorithms. This paper proposes a discretization method based on statistical indicator. It measures the interdependence between class label and attributes with the aim to find optimal interval lists by means of correlation coefficient. It also introduces variable precision rough set model to effectively control information loss generated by discretization. The presented method is applied to sonar sensor data recognition and other fields. Experimental results show that this method significantly improves the classification ability on J48 decision tree

Key words: discretization; data mining; interdependence; variable precision rough set (VPRS); decision tree

1 引言

随着数据量的急速增加, 数据挖掘等领域已成为当今研究热点. 然而, 现实世界中获取的数据往往是连续型的, 并且大多数的数据挖掘、归纳学习算法无法处理连续值的数据, 仅仅依赖于离散值的属性. 因此, 连续属性必须进行离散化, 即在最小化信息丢失的情况下将连续属性值域转换成小数目有限的区间, 这样不仅可以简化数据, 而且能够使得算法学习更快更精确.

离散化方法主要被分为有监督和无监督两种类型: 有监督方法存在决策类的信息, 如基于熵的离散化(Entropy-based)^[1]、基于类-属性关联度的离散化(CAI-based)^[2-3]和基于卡方统计独立性的离

散化(Chi2-based)^[4-5]等算法; 无监督方法不考虑类的信息, 传统的无监督离散化方法包括等宽离散化(Equal Width Discretization, EWD)和等频离散化(Equal Frequency Discretization, EFD)^[6]. 这两种方法的特点是实现简单且计算消耗低, 但结果往往难以满足预计的要求.

文中提出一种基于统计指标的数据离散化方法(Statistical Indicator Method, SIM), 凭借相关系数衡量类与属性间的关联度, 获取最优区间列表. 纳入变精度粗糙集模型, 有效地控制数据由离散化导致的信息丢失. 所提方法在声纳传感器数据识别以及其它领域上进行了应用. 实验结果表明, 文中提出方法产生了很好的离散化结果, 减少了信息丢失, 显著地提高了 J48 决策树的分类学习精度.

2 SIM 离散化方法

2.1 离散化问题描述

对于给定数据集,样本数为 N ,类别数为 S , A 为数据集中任意一个连续属性,则存在离散化方案 D 可以将连续属性 A 的值域离散成 I 个区间:

$$D: \{[d_0, d_1], (d_1, d_2], \dots, (d_{I-1}, d_I]\}$$

其中, d_0 是连续属性 A 的最小值, d_I 是 A 的最大值,属性 A 的值按升序进行排列, $\{d_0, d_1, d_2, \dots, d_{I-1}, d_I\}$ 为离散化过程中的断点集合。

属性 A 的每个值都可以划分到离散的 I 个区间的某一个区间中,离散化方案 D 作用在属性 A 上得到的区间和类别 S 构成二维的信息表(见表1)。在表1中, N 为样本总数, N_{ij} 为 i 区间中 j 类的样本数, N_{i+} 为 i 区间的样本数, N_{+j} 为 j 类的样本数, $i = 1, 2, \dots, I; j = 1, 2, \dots, S$ 。

表1 关系信息表

决策类	区间				列求和
	$[d_0, d_1]$	$(d_1, d_2]$	$\dots$	$(d_{I-1}, d_I]$	
C_1	N_{11}	N_{21}	$\dots$	N_{I1}	N_{+1}
C_2	N_{12}	N_{22}	$\dots$	N_{I2}	N_{+2}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
C_S	N_{1S}	N_{2S}	$\dots$	N_{IS}	N_{+S}
行求和	N_{1+}	N_{2+}	$\dots$	N_{I+}	N

2.2 λ 相关系数标准

在统计学中,相关是指两个或多个变量之间存在着一种相互依赖关系,通常是在两个变量之间。利用一个统计值来衡量变量之间的关系的的过程被称为相关分析,其中,这个统计值就是所谓的相关系数^[7]。根据用于测量变量层次的不同,相关系数可分为多种不同的类型。比如用于测定两个定类数据相关性的 λ 系数,用于测量两个定序数据相关性的 Gamma 系数等。文中采用 λ 相关系数。

λ_{xy} 相关系数适用于非对称性的两个定类变量,即需要区分自变量和因变量,通常以 x 代表自变量, y 代表因变量。其计算公式为:

$$\lambda_{xy} = \frac{\sum m_y - M_y}{N - M_y} \quad (1)$$

其中, m_y 指变量 x 中每个值对应变量 y 中每个值出现样本次数的最大值, M_y 指变量 y 中每个值出现样本次数的最大值, N 指出现样本的总数。

相关系数的绝对值越大,表示相关程度越大。本文将 λ 相关系数引进到离散化中,提出一种基于 λ 相

关系数的离散化标准:

$$\text{sim} = \frac{\sum_{i=1}^I \max_{i+} - \max_c}{N - \max_c} \quad (2)$$

其中, I 为当前被离散的区间数, $\max_{i+}$ 为 i 区间中最多类样本数, $\max_c$ 为数据集中最多类样本数。

sim 反映了类-属性间的相关性, sim 值越大,

类与属性越相关。如果 $\sum_{i=1}^I$ 值较大,说明 I 个区间的最多类样本数占它们各自区间样本数的比例都很大。然而,如果 I 个区间各自对应最多的类均属于同一类,则 $\max_c$ 也很大,导致 sim 很小,类与属性不相关,也就是说,断点不应该把相同类的属性值分开;如果它们各自对应最多的类属于不同的类,则 $\max_c$ 很小, sim 很大,类与属性相关度高,断点应该被加入,即被选断点应该把不同类的属性值分开。因此通过上面的分析,所提出的 λ 相关系数的离散化标准是非常合理的。举一个简单的例子加以说明,假设一个属性被离散成2个区间 $[d_0, d_1]$ 和 $(d_1, d_2]$, 且有2个类(见表2),根据公式(2), $\max_{1+} = 29$, $\max_{2+} = 59$, $\max_c = 74$, $N = 111$, 因此, λ 相关系数的离散化标准值:

$$\begin{aligned} \text{sim} &= \frac{(\max_{1+} + \max_{2+}) - \max_c}{N - \max_c} \\ &= \frac{(29 + 59) - 74}{111 - 74} \approx 0.38 \end{aligned}$$

从表2中可以直观看出,如果总的样本数 N 不变, $\max_{1+}$ 和 $\max_{2+}$ 均增加,且两个区间对应的最多类不是同一类,则 sim 值将变大,类与属性越相关,断点应该被加入。

表2 简单的类与属性二维表

类标签	属性区间		列求和
	$[d_0, d_1]$	$(d_1, d_2]$	
C_1	29	8	37
C_2	15	59	74
行求和	44	67	111

2.3 信息丢失控制规则及 SIM 算法

好的离散化结果不仅需要一个好的离散化标准,还需要一个有效的信息丢失控制规则。先前很多离散化方法主要集中在单个连续属性的分类问题上,没有考虑在多变量下被离散的数据产生的信息丢失,从而降低了学习精度。文中利用变精度粗糙集模型^[8](VPRS)中的 β -近似精度控制离散化产生的信息丢失,允许数据存在一定的分类错误,可以很

好地权衡离散区间数与分类错误。

变精度粗糙集是对传统粗糙集理论的一种推广,它通过引入一个精度参数 β 来处理部分分类问题, β 值代表一定比例的样本点被分到同一类情况下条件概率的界限。

设 $S = (U, A, V, F)$ 为一信息系统, 其中 $U = (x_1, x_2, \dots, x_n)$ 是论域, A 是属性集合, V 是属性取值集合, F 是 $U \times A \rightarrow V$ 的映射. 若 $A = C \cup D$, $C \cap D = \emptyset$, 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. $x, y \in U$, 对于 $P \subseteq A$, Q_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 1 设 $X \subseteq U$ 为论域的一个子集, $P \subseteq C$, 集合 X 的关于 P 的 β -下近似为:

$$C_\beta(D) = \bigcup_{1-CP(Y|x_i) \leq \beta} \{x_i \in [x]_P\} \quad (3)$$

其中, $[x]_P$ 表示 U 中在等价关系 P 下的等价类元素构成的集合, $CP(Y|x_i) = |Y \cap x_i| / |x_i|$, $Y \in [x]_D$, $|\cdot|$ 指集合的基数. β -下近似代表样本点关于决策属性的正确分类程度。

定义 2 设 $S = (U, A, V, F)$ 是一决策表, 条件属于 C 的 β -近似精度定义如下:

$$\gamma_\beta = \frac{|C_\beta(D)|}{|U|} \quad (4)$$

其中, $C_\beta(D)$ 代表样本点关于决策属性的正确分类程度, γ_β 指正确分类率, 允许 $1-\beta$ 分类误差, 即 β -近似精度。

用 γ_β 控制数据的信息丢失, 权衡离散区间数与分类错误. 在离散化过程中, 如果原始数据的 β -近似精度 $\gamma_{\beta\text{-original}}$ 比离散数据的 β -近似精度 $\gamma_{\beta\text{-discretized}}$ 大于 ξ (ξ 为数据可容忍的信息丢失率), 即 $\gamma_{\beta\text{-original}} - \gamma_{\beta\text{-discretized}} > \xi$, 则根据 sim 标准添加断点, 直到 $\gamma_{\beta\text{-original}} - \gamma_{\beta\text{-discretized}} \leq \xi$ 停止离散化。

sim 算法:

- (1) 初始原始数据 β -近似精度 $\gamma_{\beta\text{-original}}$.
- (2) 对每个连续属性, 将属性值从小到大排序.
- (3) 计算每个候选断点添加到该区间后的 sim 标准值, 选择使得 sim 值最大的断点作为结果断点.
- (4) 计算当前数据的 β -近似精度 $\gamma_{\beta\text{-discretized}}$, 如果 $\gamma_{\beta\text{-original}} - \gamma_{\beta\text{-discretized}} > \xi$, 跳到(3); 否则, 停止离散化.

时间复杂性分析: 计算 β -近似精度的时间为 $O(N^2)$; 对连续属性值排序的时间为 $O(N \log N)$; 计算 sim 离散化标准的时间为 $O(I)$, 则计算所有断

点的 sim 标准的时间为 $O(N \cdot I)$. 该算法时间复杂性为 $O(N^2) + O(N \log N) + O(N \cdot I) = O(N^2)$.

3 实验结果与分析

3.1 在声纳传感器数据识别上的效用

声纳传感器从不同的角度和不同的条件下收集声纳信号的 208 个样本的 Sonar 数据集^[9], 跨越 90 度的圆柱体和 180 度的岩石, 其中 111 个样本为金属圆柱体反弹的声纳信号; 97 个样本为圆柱形的岩石反弹的声纳信号. 每个样本包含 60 个特征, 每个特征值的范围在 0.0 到 1.0 之间. 每个数值代表在一个特定频段的能量. 除了传感器采集的数据外, 不需要任何其它信息。

分别采用 Entropy^[1]、EQW^[6]、Chi2^[4]、CACC^[5]和文中提出 SIM 方法将 Sonar 数据集进行离散化, 目的是洞察提出离散化方法的有效性. 用 J48 分类器将离散化的结果进行分类识别预测, 对平均学习精度统计进行对比, 如表 3 所示。

表 3 J48 识别结果

方法	识别精度(%)	离散后不同样本 占样本总数比例(%)
J48	76.8	—
Entropy+J48	80.6	89.4
EQW+J48	76.2	87.9
Chi2+J48	88.3	90.6
CACC+J48	89.6	89.4
SIM+J48	90.7	85.3

从表 3 可以明显看出, 对于 Sonar 数据集而言, SIM 方法离散后得到的不同样本数占样本总数的比例是最少的. 与其他离散化方法相比, SIM 可以得到相对简化的数据, 并且不丢失原始数据的有效信息, 大大简化了数据维度, 降低了存储要求. 而对于声纳信号的识别来说, 基于 SIM 离散化算法的 J48 分类识别精度是最高的, 这正是数据挖掘技术期望得到的结果, 充分显示了提出技术的优势。

3.2 其他领域数据的性能评价

对 UCI 机器学习数据库^[9]中的 9 个数据集(见表 4)用上述离散化方法进行离散化. 应用 J48 方法构造决策树对离散后的数据进行分类预测. 对离散化产生的不一致数和平均学习精度进行对比(见表 5、表 6). 注: 在实验中, EQW 离散化方法的分箱数为 10; 并且取分类错误率 $\beta=0.05$; 不一致数指属性值完全相同但决策类不同的样本数目的总和。

表 4 数据信息表

数据集	连续属性数	类别数	样本数
Labor	8	2	57
Iris	4	3	150
Wine	13	3	178
Glass	9	7	214
Ionosphere	34	2	351
Pima	8	2	768
Vehicle	18	4	846
Satimage	36	6	6435
Adult	6	2	48842

表 5 不一致数比较

数据集	离散化方法				
	EQW	Chi2	Entropy	CACC	SIM
Labor	5	8	10	13	3
Iris	3	9	14	0	7
Wine	9	7	3	11	9
Glass	17	21	12	7	10
Ionosphere	21	23	43	46	15
Pima	104	67	81	54	41
Vehicle	65	23	42	31	39
Satimage	432	284	351	371	279
Adult	4212	1753	2049	2453	1079

表 6 J48 平均学习精度(%)

数据集	离散化方法				
	EQW	Chi2	Entropy	CACC	SIM
Labor	71.1	90.9	85.6	84.2	90.1
Iris	92.8	94.0	93.6	94.6	96.2
Wine	91.3	94.5	92.6	94.9	95.9
Glass	79.3	93.1	88.4	87.9	93.5
Ionosphere	83.7	93.4	91.3	92.5	93.7
Pima	73.4	82.1	84.2	80.7	87.6
Vehicle	69.7	77.9	73.2	76.4	78.4
Satimage	78.3	85.6	87.1	84.5	89.7
Adult	81.8	86.1	83.5	82.3	86.9

由表 5 可以看出, SIM 离散化方法所产生的数据不一致数平均最小. 对于 CACC 方法, 由于它所产生的离散区间数往往接近于类数, 因此, 对类少的数据会产生较多的不一致(如 Labor、Iris 和 Adult), 而对类多的数据则产生较少的不一致(如 Glass、Vehicle 和 Satimage). 对于 EQW 方法, 固定了每个属性的离散区间数, 因此, 对样本多、类少的数据将会产生较多的不一致(如 Pima、Vehicle、Satimage 和 Adult), 而对样本少、类多的数据将产生较少的不一致(如 Labor 和 Iris). Entropy 方法没有考虑多变量下的信息丢失, 所以也出现了或多或少的不一致数. 尽管 Chi2 方法考虑了多变量下的不一

致问题, 但它定义的不一致衡量标准不够严格, 牺牲了部分的信息丢失. 在表 6 中, 经过 SIM 离散化后, J48 有较好的分类学习能力, 充分显示了 SIM 离散化算法的有效性.

4 结束语

连续属性离散化方法的研究对于计算机应用的多方面研究与应用有着重要作用. 文中提出一种统计指标离散化方法, 可以找到有效的区间列表, 减少信息丢失, 提高决策树的分类学习能力, 并得到了很好的应用.

参考文献:

[1] Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning[C] // Proc. Thirteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 1993: 1022-1027.

[2] Kurgan L A, Cios K J. CAIM discretization algorithm [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153.

[3] Tai C. J, Lee C I, Yang W P. A discretization algorithm based on class-attributes contingency coefficient[J]. Information Sciences, 2008 (178): 714-731.

[4] Liu H, Setiono R. Feature selection via discretization [J]. IEEE transactions on knowledge and data engineering, 1997, 9(4): 642-645.

[5] Su C T, Hsu J H. An extended Chi2 algorithm for discretization of real value attributes[J]. IEEE transactions on knowledge and data engineering, 2005, 17(3): 437-441.

[6] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous feature [C] // Machine learning; Proc. 12th Int'l Conf. Lake Tahoe, CA. 1995: 194-202.

[7] 梁前德. 统计学[M]. 北京: 高等教育出版社, 2004.

[8] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Science, 1993 (46): 39-59.

[9] Hettich S, Bay S. D. The UCI KDD archive [EB/OL]. [2011-01-20]. <http://kdd.ics.uci.edu/>.

作者简介:

陈 浩 男, (1956-), 副教授. 研究方向为电子技术应用、数据挖掘.