

基于信息熵理论的连续属性离散化方法

高建国, 崔业勤

(廊坊师范学院 计算机系, 河北 廊坊 065000)

摘 要: 很多数据挖掘和机器学习方法仅仅依赖于离散值的属性, 这样必须离散连续的属性. 文中提出一种基于信息熵理论的数据离散化方法(IED), 利用信息熵的思想衡量离散区间是否类似, 同时考虑离散区间大小对离散化结果的影响, 该方法综合考虑了离散区间与类之间的独立性. 实验结果表明, IED 显著地提高了 Naïve-bayes 分类学习精度.

关键词: 离散化; 数据挖掘; 信息熵

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2011)07-0187-03

A New Discretization Method for Continuous Attributes Based on Information Entropy

GAO Jian-guo, CUI Ye-qin

(Department of Computer, Langfang Teachers' College, Langfang 065000, China)

Abstract: Most data mining and induction learning methods only rely on discrete attributes. So, continuous attributes must be discretized. This paper presents a new data discretization method for continuous attributes based on information entropy, namely IED. It measures the similarity of intervals by using information entropy and considers the effect of the discrete interval size on discretization results. This method synthetically takes into account the independence between the merged intervals and target class. Experimental results show that IED can yield more classification accuracy by implementing Naïve-bayes.

Key words: discretization; data mining; information entropy

1 引言

随着数据库系统中信息量的增加, 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、机器学习等算法(如 C4.5 和 Naïve-bayes)无法处理连续属性值的数据, 他们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, 不仅可以简化数据, 而且能够提高算法学习的速度和精度.

目前, 已经有很多离散化方法被提出, 其类型^[1]主要可分以下几类:

(1) 自底向上(bottom-up)和自顶向下(top-

down): Bottom-up 离散化方法起始以每个属性值为一个区间, 然后迭代地合并相邻区间. Top-down 离散化方法起始把整个属性的值域视为一个区间, 递归地向该区间中添加断点;

(2) 有监督和无监督: 有监督离散化方法考虑标签信息, 如 MDLP^[2]、CAIM^[3]和 Chi merge^[4]等算法. 前面两种方法均没有考虑多变量下的不一致问题, 导致了较多的信息丢失, 使得学习精度不准确; Chi merge 算法使用卡方统计来确定当前相邻区间是否被合并, 并采用显著性水平值逐渐降低的方法检验系统的不一致率, 确定离散化进程是否终止. 然而, Chi merge 在衡量区间差异时没有考虑区间大小和区间类别数对离散化结果的影响, 另外, 它定义的不一致衡量标准不能很好地权衡信息丢失与

离散的区间数.

无监督离散化方法则不考虑类的信息. 传统的 top-down 无监督离散化方法包括 EQW 和 EQF^[1], 这两个算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;

(3) 全局和局部: 全局离散化采用整体样本空间产生离散化方案, 是一种静态的方法; 局部离散化方法仅仅将部分样本用于离散化中, 是一种动态的方法.

文中提出一种基于信息熵和概率理论的数据离散化方法 (IED). 实验结果表明, IED 方法可以得到较高的 Naïve-bayes 分类学习精度.

2 IED 离散化方法

文中着重于 bottom-up 离散化形式的方法, 其实质是在最小化信息丢失的情况下, 根据一定的区间合并准则, 消除断点、合并相邻区间. 对于每个离散化任务, 要求数据集包含 N 个样本, 每个样本包含 m 个条件属性和一个决策类标签. 起初, 连续属性相邻两个值的均值被视为一个断点, 两个断点构成一个区间. 决策类与离散区间构成二维信息表 (见表 1). 在表 1 中, S 为决策类别数, A_{ij} 为 i 区间 j 类样本数, $M_{i+} = \sum_{j=1}^S A_{ij}$ 为 i 区间样本数, $M_{+j} = \sum_{i=1}^2 A_{ij}$ 为相邻两区间中 j 类样本数, $M = \sum_{i=1}^2 M_{i+}$ 为相邻两区间总的样本数.

表 1 二维表信息

相邻两区间	决策类				行求和
	C_1	C_2	$\dots$	C_S	
$(d_0, d_1]$	A_{11}	A_{12}	$\dots$	A_{1S}	M_{1+}
$(d_1, d_2]$	A_{21}	A_{22}	$\dots$	A_{2S}	M_{2+}
列求和	M_{+1}	M_{+2}	$\dots$	M_{+S}	M

Bottom-up 离散化方法的目标是选用一种有效的区间合并标准, 迭代地合并相邻区间, 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. 文中提出一种基于信息熵和概率理论的数据离散化方法 (IED), 综合考虑了各对合并区间与目标类之间的独立性, 即离散区间与目标类越独立, 它们越应该合并.

信息熵可以衡量随机变量的不确定性, 它反映了随机变量对应类分布的特性, 熵值越大, 不确定性越大, 反之亦然; 当每个类含有等数量样本时, 熵取

最大值 $\log S$, 当区间中仅有一个类时, 熵取最小值.

定义 1 设 X 是取有限值的随机变量 $P_i = P\{X = x_i\}, i = 1, 2, \dots, n$, 则 X 的熵定义为:

$$H(X) = - \sum_{i=1}^n p_i \cdot \log_2(1/p_i) \quad (1)$$

对于两个相邻区间 I_1 和 I_2 , 其信息熵可被定义为:

$$H(I_i) = - \sum_{j=1}^S \frac{A_{ij}}{M} \log \frac{A_{ij}}{M} \quad (2)$$

如果合并两个区间 $I = I_1 \cup I_2$, 根据相同的规则有:

$$H(I) = - \sum_{j=1}^S \frac{A_{1j} + A_{2j}}{M} \log \frac{A_{1j} + A_{2j}}{M} \quad (3)$$

如果独立地对待每一个区间, 可以得到相邻两区间的总体熵, 即带有权重的每个区间熵的和

$$H(I_1, I_2) = \sum_{i=1}^2 \frac{M_{i+}}{M} H(I_i) \quad (4)$$

对于一个连续属性的各对相邻区间, 它们对应的类分布是不同的, 类分布最相似的区间应该首先被合并. 事实上, 从信息通信的角度考虑, 区间在合并前与合并后需要转换信息量, 转换的信息量越小, 说明两个区间对应的类分布越相似, 它们应该被合并, 反之亦然. 另外, 由于相邻两区间的样本数为 M , 需要转换 M 次, 因此, 用 $M \cdot (H(I) - H(I_1, I_2))$ 作为区间相似性的衡量标准.

为了更好地衡量离散区间与目标类之间的独立性, 仅仅考虑类分布的相似性是不够的, 还需要考虑离散区间的大小和区间中类别数对离散化结果的影响, 进而会影响到分类器的学习精度. 下面, 通过概率的方法获得两个因素的衡量标准. 对于任意连续属性, 每对区间 (I_1 和 I_2) 的样本数是不同的, 可视作为变量 $\{M_i\}$, 则 $p(\{M_{i+}\})$ 代表两个区间样本数的集合可能性, 即:

$$\begin{aligned} p(\{M_{i+}\}) &= p(\{M_{2+}\})p(\{M_{1+} | M_{2+}\}) + \\ & p(\{M_{1+}\})p(\{M_{2+} | M_{1+}\}) \\ &= \frac{1}{M \cdot M_{1+}} + \frac{1}{M \cdot M_{2+}} \\ &= \frac{2}{\prod_{i=1}^2 M_{i+}}, \quad i \in \{1, 2\} \end{aligned}$$

上式取负对数, 将概率的最大化转化为最小化形式:

$$-\log(p(\{M_{i+}\})) = \sum_{i=1}^2 \log M_{i+} - 1 \quad (5)$$

由于每个区间中的类别数越小, 类则有越相似的可能性, 即区间样本数和类数越少, 他们越应该被

合并;因此根据式(5),可以采用 $S_i \log M_{i+}$ 作为区间合并标准的一部分来评价两因素对离散化结果的影响. 基于上面的分析,提出一种基于信息熵和概率理论的区间合并标准(IMC):

$$\text{IMC} = M \cdot (H(I) - H(I_1, I_2)) + \sum_{i=1}^2 S_i \log M_{i+} \quad (6)$$

其中, S_i 指 i 区间中类别数, $i \in \{1, 2\}$. IMC 代表了离散区间与目标类之间的独立性衡量,其值越小,二者越独立,区间越应该被合并,反之亦然. 文中采用粗糙集理论中的一致性水平来控制数据的信息丢失^[5]. 下面给出 IED 算法具体步骤.

IED 算法具体步骤如下:

输入: N 个样本的数据集, m 个连续属性, S 个类.

输出: 离散后的数据集, 每个属性 k_i 有个区间.

① 计算原始数据的一致性水平 $\text{Inc}_{\text{original}}$.

② 对每一个连续属性,将属性值从小到大排序 $\{[d_0, d_1], (d_1, d_2], \dots, (d_{l-1}, d_l]\}$. 初始,相同值的集合视为一个区间;

③ 计算所有属性相邻区间的合并标准值 IMC, 合并最小 IMC 值的两个区间;

④ 计算当前数据的一致性水平 $\text{Inc}_{\text{discretizal}}$, 如果 $\text{Inc}_{\text{original}} - \text{Inc}_{\text{discretizal}} < \xi$ (ξ 为数据可容忍的信息丢失率), 返回 ③; 否则, 停止离散化.

接下来分析 IED 算法的时间复杂性, 计算一致性水平的时间为 $O(N^2)$; 对连续属性值排序的时间为 $O(N \log N)$; 计算区间合并标准的时间为 $O(S)$, 则计算所有属性相邻区间的合并标准为 $O(mNS)$. 因此, 该算法总的时间复杂性为:

$$O(N^2) + O(N \log N) + O(mNS) = O(N^2)$$

3 实验结果

为了评价 IED 的性能, 采用了 UCI 机器学习数据库^[6]中的 15 个数据集(见表 2), 该数据集是数据挖掘等实验所常用的数据, 包括两个大的数据集 Page-blocks 和 Letter.

将文中提出的 IED 方法与下列方法比较:

① EQF: 传统的无监督离散化方法^[1];

② MDLP: 基于熵的最小描述长度离散化方法^[2];

③ CAIM: 经典的 top-down 离散化方法^[3];

④ Chi merge: 经典的 bottom-up 离散化方法^[4].

15 个数据集分别用上面的离散化方法进行离散数据, 使用 Weka 数据挖掘工具进行实验, 采用 Naïve-bayes 分类器对离散后的数据进行分类预测. 采用 10 折交叉验证的方法, 将数据集分成 10 等份, 分别将其其中 9 份作为训练集, 剩下一份作为测试集, 重复 10 次取平均值, 对平均学习精度统计进行对比(见表 3).

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
Auto	5	2	3	392
Breast	9	0	2	683
Ionosphere	34	0	2	351
Glass	9	0	7	214
Wine	13	0	3	178
Machine	7	0	8	209
Heart	5	8	2	296
Sonar	60	0	2	208
Vehicle	18	0	4	846
Vowel	10	3	6	990
Bupa	6	0	2	345
Yeast	6	2	10	1484
Page-blocks	10	0	5	5473
Letter	16	0	26	20000

表 3 Naïve-bayes 平均学习精度

数据集	离散化方法(%)				
	IED	CAIM	Chi merge	MDLP	EQF
Iris	96.2	92.6	94.6	93.3	92.6
Auto	83.5	75.8	81.4	78.6	73.1
Breast	96.3	96.7	96.3	96.0	93.8
Ionosphere	94.6	92.0	92.9	86.8	82.6
Glass	77.1	76.9	70.9	73.2	57.9
Wine	95.4	92.3	94.1	83.3	90.7
Machine	88.6	83.3	85.6	80.4	72.0
Heart	81.9	76.8	83.2	74.8	69.6
Sonar	64.8	56.7	58.5	59.9	54.6
Vehicle	72.9	67.6	71.6	68.6	63.4
Vowel	97.0	95.6	97.8	97.5	95.0
Bupa	69.8	62.6	65.8	62.2	57.8
Yeast	93.7	89.5	93.1	90.8	85.1
Page-blocks	96.7	95.6	95.6	95.1	94.8
Letter	79.2	77.8	76.5	76.9	68.3

可以看出, 除了 Breast、Heart、Vowel 数据集, 文中提出的 IED 离散化方法的 Naïve-bayes 平均学习精度均有所上升, 这点正是离散化方法期望得到的结果, 充分显示了 IED 算法的优势. 对于 CAIM、MDLP、EQF 三种离散化方法均没有引入不一致衡量标准, 换句话说, 它们没有对数据的有效性

(下转第 194 页)

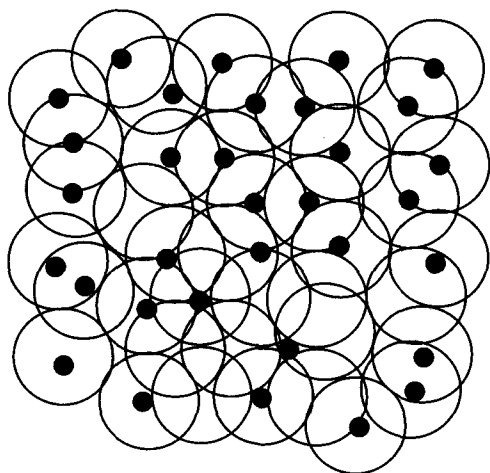


图6 多传感器网络节点部署

当网络中某个节点失效后,可自适应启动睡眠节点,重新建立三角形通信区域,保证可靠、完全、实时和无缝采集、传输数据;并给出所需最小传感器节点和簇首节点的数目。仿真实验表明,该机制是一种高效节能且能够完全覆盖给定监测区域的节点部署方案。

参考文献:

- [1] Wang L, Shen H, Chen zhe et al. Voronoi tessellation based rapid coverage decision algorithm for wireless sensor networks[J]. Computer Science, 2007, 46(11): 495-502.

(上接第 189 页)

进行控制,在离散化过程中丢失了大量的信息,所以分类预测的精度比 Chi merge 和 IED 方法平均低很多。

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。近年来尽管有多种连续属性离散化方法提出,但具有一定影响和一定理论意义的方法并不多。文中提出一种基于信息熵和概率理论的数据离散化方法,该方法综合地考虑了各对合并区间与目标类之间的独立性,能够更合理准确地离散化。

参考文献:

- [1] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous feature[C]// Machine learning: Proc. 12th Int'l Conf. Austin, Texas: Morgan Kaufmann Publishers, 1995: 194-202.
- [2] Fayyad U, Irani K. Multi-interval discretization of

- [2] Meguerdichian S. Coverage problem in wireless Ad Hoc sensor networks[C]// Proc of the 20th Annual Conference of the IEEE Computer and Communications Societies. Anchorage, AK, USA: IEEE, 2001 (3): 1380-1387.
- [3] YE F Zhang Honghai, Lu Songwu, et. al. A randomized energy conservation protocol for resilient sensor networks[J]. Wireless Networks, 2006, 12 (5): 637-652.
- [4] 袁辉勇, 阙清贤, 羊四清. 传感器网络中基于能耗均衡的节点优化部署[J]. 计算机仿真, 2010(8): 100-102.
- [5] 汪学清, 等. 无线传感器网络中基于网格的覆盖问题研究[J]. 计算机科学, 2006, 33(11): 38-39.
- [6] 屈斌, 胡访宇. 高效节能的无线传感器网络路由协议研究[J]. 计算机仿真, 2008, 25(5): 113-116.
- [7] 周永华, 李鹏, 毛宗源. 一种新的混合杂交方法及其在约束优化中的应用[J]. 计算机工程与应用, 2006: 48-50.

作者简介:

王改香 女, (1970-), 硕士, 讲师. 研究方向为计算机网络技术。

高太平 男, (1970-), 教授, 硕士生导师. 研究方向为图论、网络优化、算法设计与分析。

continuous-valued attributes for classification learning [C]// Proc Thirteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 1993. 1022-1027, 1993.

- [3] Kurgan L A, Cios K J, CAIM discretization algorithm [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153.
- [4] Kerber R. ChiMerge: discretization of numeric attributes[C]// Proceedings Ninth National Conference on Artificial Intelligence. [S. l.]: AAAI Press, 1992: 123-128.
- [5] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11 (5): 341-356.
- [6] Hettich S. Bay S D. The UCI KDD archive [DB/OL] [2010-12-01]. <http://kdd.ics.uci.edu/>, 1999.

作者简介:

高建国 男, (1970-), 硕士, 副教授, 研究方向为数据挖掘、模式识别、数据库及数据库应用开发。

崔业勤 女, (1973-), 硕士, 副教授, 研究方向为人工智能、数据挖掘、数据压缩与多媒体应用。