PIE:实值属性离散化方法及应用

李 杰[1,2],王 欢[2]
(1.中国科学院研究生院,北京 100040;
2.北华航天工业学院 计算机科学与工程系,河北 廊坊 065000)

摘 要:提出一种基于概率与信息熵理论的实值属性离散化方法,综合考虑了各对合并区间之间的差异性;该方法利用信息熵衡量相邻区间的相似性,同时考虑离散区间大小和区间类别数对学习精度的影响,并通过概率的方法得到了这两个因素的衡量标准。仿真结果表明,新方法对 See5/C5.0 分类器有较好的分类学习能力,并在肿瘤诊断中得到了很好的应用。

关键词:离散化;数据挖掘;概率;信息熵
中图分类号:TP18 文献标识码:A 文章编号:1674-7720(2011)15-0068-03

PIE:discretization method for real attributes and its application

Li Jie[1,2], Wang Huan[2]
(1.Graduate University of Chinese Academy of Sciences, Beijing 100040, China;
2.Department of Computer Science and Engineering, North China Institute of Aerospace Engineering, Langfang 065000, China)

Abstract: This paper presents a discretization method for real attributes based on probability and information entropy, namely PIE, which synthetically considers the variance among the merged intervals. This method measures the similarity of adjacent two intervals by using information entropy and takes into account the effect of the discrete interval size and class number of each interval on learning accuracy, and the measurement of two factors is achieved with probabilistic means. Simulation results show that PIE can yield more classification and learning accuracy by running See5/C5.0 classifier and has better application on tumor diagnosis.

Key words: discretization; data mining; probability; information entropy

连续属性离散化是数据挖掘和机器学习的重要预处理步骤,直接影响到机器学习的效果。在分类算法中,对训练样本集进行离散化具有两重意义:一方面可以有效降低学习算法的复杂度,加快学习速度,提高学习精度;另一方面可以简化、归纳获得的知识,提高分类结果的可理解性。很多离散化方法的提出,主要分为以下两种类型[1]:(1)自底向上和自顶向下的离散化方法。自底向上离散化方法是以每个属性值为一个区间,然后迭代地合并相邻区间;自顶向下离散化方法是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)有监督和无监督离散化方法。有监督方法使用决策类信息进行离散化,如 Ent-MDLP[2]、CAIM[3]和 Chi2-based[4-5]等算法。Ent-MDLP 使用熵的理论来评价候选断点,选择使得整体熵值最小的断点作为最终断点,并且通过最小描述长度原则来确定离散区间数;CAIM 是一种自顶向下离散化方法,该方法依据类与属性间的关联度,提出一种启发式离散化标准,计算当前状态的标准值来判别当前断点是否应该被加入断点集合中。自底向上的 Chi2-based 离散化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,Chi2-based 方法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,可能会导致学习精度的降低;而无监督离散化方法则不考虑类的信息。传统的无监督离散化方法包括 EWD (Equal Width Discretization)和 EFD(Equal Frequency Discretization),这两个算法实现简单且计算消耗低,但结果往往难以满足预计的要求。

本文提出一种基于概率与信息熵理论的实值属性离散化方法 PIE (Probability and Information Entropy),综

合考虑了各对合并区间之间的差异性，利用信息熵衡量相邻区间的相似性，同时考虑离散区间大小和区间类别数对分类能力的影响，并通过概率的方法得到了这两个因素的衡量指标。实验结果表明，PIE 显著地提高了 See5/C5.0 分类器分类学习精度，并在乳腺肿瘤诊断中得到了很好的应用。

1 PIE 离散化

离散化问题描述如下：对于 m 个连续属性的数据集，样本点个数为 N，决策类别数为 S，数据集中任意一个连续属性为 a，可以将连续属性的值域离散成 I 个区间：

$$P:\{[d_0,d_1],[d_1,d_2],\cdots,[d_{I-1},d_I]\}$$

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，$\{d_0,d_1,d_2,\cdots,d_{I-1},d_I\}$ 为离散过程中的断点集合。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。

本文主要针对自底向上离散化形式的方法，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。对于每个自底向上离散化任务而言，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。定义 A_{ij} 为 i 区间 j 类样本数 $(i\in\{1,2\},1\leqslant j\leqslant S)$，$M_{i+}=\sum_j^S A_{ij}$ 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。

自底向上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。本文提出一种基于概率与信息熵理论的实值属性离散化方法 PIE，综合考虑各对合并区间之间的差异性；利用信息熵衡量相邻区间的相似性，同时考虑离散区间大小和区间类别数对分类能力的影响。

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。对于两个相邻区间 I_1 和 I_2，其信息熵可被定义为：

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

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：

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

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。由于相邻两区间的样本数为 M，需要转换 M 次，因此，用 $M\times[H(I)-H(I_1,I_2)]$ 作为区间相似性的衡量标准。

为了更好地衡量各对合并区间之间的差异性，仅考虑类分布的相似性是不够的，还需要考虑离散区间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。通过概率的方法可获得两个因素的衡量标准，对于任意连续属性，每一对相邻区间（I_1 和 I_2）的样本数是不同的，可视为变量 $\{M_i\}$，则 $p(\{M_{i+}\})$ 代表两个区间样本数的集合可能性，即：

$$p(\{M_{i+}\})=p(\{M_{2+}\})p(\{M_{1+}|M_{2+}\})+p(\{M_{1+}\})p(\{M_{2+}|M_{1+}\})=\frac{1}{M\times M_{1+}}+\frac{1}{M\times M_{2+}}=\frac{2}{\prod_{i=1}^{2}\times M_{i+}},i\in\{1,2\}$$

式中取负对数，将概率的最大化转化为最小化形式：

$$-\log(p(\{M_{i+}\}))=\sum_{i=1}^{2}\log M_{i+}-1 \tag{3}$$

由于每个区间中的类别数越小，类分布可能越相似，即区间样本数和类数越少，越应该被合并。因此根据式(3)，采用 $S_i\log M_{i+}$ 作为区间合并标准的重要部分来评价两个因素对离散化结果的影响。基于此，基于概率与熵的区间合并标准 pie 被定义为：

$$pie=M(H(I)-H(I_1,I_2))+\sum_{i=1}^{2}S_i\log M_{i+} \tag{4}$$

其中，S_i 代表 i 区间中类别数，$i\in\{1,2\}$。pie 代表了离散区间之间的差异性衡量，其值越小，区间越应该被合并，反之亦然。PIE 采用粗糙集中的近似精度[6]来控制数据的信息丢失。PIE 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，每个属性有 t_i 个区间。

(1)计算原始数据的近似精度 Lc_{original}；

(2)对每一个连续属性值从小到大排序。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对的合并标准值 pie，合并最小 pie 值的两个区间；

(4) 计算当前数据的一致性水平 $Lc_{\text{discretized}}$，如果 $Lc_{\text{original}}-Lc_{\text{discretized}}<\theta$（$\theta$ 为数据可容忍的信息丢失率），返回步骤(3)；否则，停止离散化。

对 PIE 算法的时间复杂性进行分析：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 $O(N\log_2 N)$；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 $O(N^2)+O(N\log_2 N)+O(mNS)-O(N^2)$。

2 仿真结果

2.1 UCI 数据集实验结果

为了评价 PIE 的性能，采用了 UCI 机器学习数据

库[7]中的 10 个数据集，见表 1 所示。该数据集是数据挖掘等实验常用的数据，其中包括两个大的数据集 Page-blocks 和 Letter。PIE 方法与以下几种方法进行了比较：传统的无监督离散化方法 EFD；基于熵的最小描述长度离散化方法 Ent-MDLP；流行的自顶向下离散化方法 CAIM；经典的自底向上离散化方法 Chi2。

表 1 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
Glass	9	0	7	214
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
Yeast	6	2	10	1 484
Page-blocks	10	0	5	5 473

10 个数据集分别采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 See5 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，分别将其中 9 份作为训练集，剩下 1 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 2 所示。

表 2 See5 平均学习精度/%

数据集	离散化方法				
	PIE	CAIM	Chi2	Ent-MDLP	EFD
Iris	96.2	92.6	94.6	93.3	92.6
Auto	83.5	75.8	81.4	78.6	73.1
Glass	77.1	76.9	70.9	73.2	57.9
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
Yeast	93.7	89.5	93.1	90.8	85.1
Page-blocks	96.7	95.6	95.6	95.1	94.8

从表 2 中可以看出，除了 Heart 和 Vowel 数据集，本文提出的 PIE 离散化方法的 See5 平均学习精度均有所上升，这正是离散化方法期望得到的结果，由此充分显示了 PIE 算法的优势。而对于 CAIM、Ent-MDLP 和 EFD 三种离散化方法均则未引入不一致衡量标准，即它们没有对数据的有效性进行控制，在离散化过程中丢失了大量的信息，导致分类预测的精度比 Chi2 和 PIE 方法平均低很多。

2.2 PIE 在乳腺肿瘤诊断上的效用

乳腺肿瘤诊断的实验数据来自于 UCI 机器学习数据库中的 Breast Cancer Wisconsin 数据集，将 Breast Cancer Wisconsin 删掉属性值不全的病例样本，剩下 683 个病例样本，病理检测有 9 项 (Clump Thickness、Uniformity of Cell Size、Uniformity of Cell Shape、Marginal Adhension、Single Epithelial Cell Size、Bare Nuclei、Bland Chromatin、Normal Nucleoli、Mitoses)，即 9 个属性，每个属性取值范围[1,10]，病情状况分为两类：一类表示肿瘤为恶性，另一类表示肿瘤为良性。这样，每个样本有 9 个连续条件属性，1 个决策属性，选取样本的 80%作为训练集，20%作为测试集。

将 Breast Cancer Wisconsin 用本文所提出的 PIE 算法进行离散化，然后分别使用 See5 和 PIE+See5 对离散前和离散后的数据进行分类预测，结果见表 3。

表 3 BCW 病例数据集实验结果对比

方法	测试准确度/%	离散后剩余属性个数	离散后不同样本占样本总数比例/%
See5	92.55	--	--
PIE+See5	94.89	6	45.97

从表 3 中可以明显看出，未经过离散化处理的 BCW 病例数据集进行 See5 分类预测的测试准确度为 92.55%，而 PIE+See5 方法的测试准确度为 99.27%，比未被离散化的进行 See5 预测精度高出 6.72%，相当于每 1 000 个患者中就多出约 67 个患者可以被准确地诊断出肿瘤为良性或是恶性，对患者及时治疗有很大帮助。

在 BCW 数据被离散化后，其病理指标被删去了三项：Uniformity of Cell Shape (细胞形状均匀度)、Bland Chromatin (平淡的染色质)、Mitoses，可以只考虑其他六项，简化了信息系统，减轻了医生的工作量。另外，利用 PIE+See5 方法离散后不同样本占样本总数比例只有 44.36%，删除冗余的病例样本后，只剩余了 303 个病例样本，从而使原来的病例样本空间在横向和纵向上都得到了降维，可以得到更加稳固的训练模型，在医学数据挖掘中具有良好的发展前景。

连续属性离散化方法的研究对数据挖掘与机器学习领域的研究与应用具有重要的作用。本文提出一种基于概率与信息熵理论的实值属性离散化方法，综合考虑了各对合并区间之间的差异性，能够更合理准确地离散化，该方法为该领域提供了新思路，具有一定应用价值意义。

参考文献

[1] DOUGHERTY J, KOHAVI R, SAHAMI M. Supervised and unsupervised discretization of continuous feature [C]. Proceedings of the 12th International Conference of Machine learning. San Francisco: Morgan Kaufmann, 1995.

[2] FAYYAD U, IRANI K. Multi-interval discretization of continuous-valued attributes for classification learning [C]. Proceedings of the 13th International Joint Conference on

Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 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] LIU H, SETIONO R. Feature selection via discretization[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(4): 642-645.

[5] CHAO T S, JYH H H. An extended chi2 algorithm for discretization of real value attributes [J]. IEEE Transactions Knowledge and Data Engineering, 2005,17(3):437-441.

[6] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982,11(5):341-356.

[7] HETTICH S, BAY S D. The UCI KDD Archive [DB/OL]. http://kdd.ics.uci.edu/, 1999.

(收稿日期:2011-04-06)

作者简介:

李杰,女,1982 年生,讲师,主要研究方向:数据挖掘、模式识别及数据库应用。

王欢,男,1985 年生,硕士研究生,主要研究方向:智能系统、数据挖掘和计算机网络。