连续属性离散化的 Bayesian-Chi2 算法

刘 磊,闫德勤,桑 雨
LIU Lei,YAN De-qin,SANG Yu

辽宁师范大学 计算机系,辽宁 大连 116029
College of Computer Science,Liaoning Normal University,Dalian,Liaoning 116029,China
E-mali:lei820515@sohu.com

LIU Lei,YAN De-qin,SANG Yu.Bayesian-Chi2 algorithm for discretization of real value attributes.Computer Engineering and Applications,2008,44(18):39-40.

Abstract: Discretization is an effective technique to deal with continuous attributes for machine learning and data mining.Reasonability of a discretization process determines the accuracy of expression and extraction for information.Dealing with the discretization of real value attributes,Chi2 algorithm can get a good result of the conflict-free data but do not well in inconsistency and incomplete data.This paper makes full use of the Bayseian model which allows for the wrong classification in nature and improved the Chi2 algorithm.The improved algorithm is not only more suitable for inconsistency and incomplete data,but also make the interval merging more reasonable.The experimental results have proven the validity of the new algorithm.

Key words: discretization of real value attributes;Chi2 algorithm;Bayseian

摘 要:连续属性离散化在机器学习和数据挖掘领域中有着重要的作用。连续属性离散化方法是否合理决定着对信息的表达和提取的准确性。Chi2 算法在对连续属性进行离散化处理时,无冲突的数据能够得到较好的结果,但是,对不协调和不完全的数据实验结果不是很理想。利用了 Bayseian 模型允许一定程度错误分类存在的性质,对 Chi2 算法进行了改进。改进后的 Chi2 算法不仅更适合不协调和不完全的数据,还使得区间的合并更加合理。实验结果证明了算法的有效性。

关键词:连续属性离散化;Chi2 算法;贝叶斯
DOI:10.3778/j.issn.1002-8331.2008.18.012 **文章编号:**1002-8331(2008)18-0039-02 **文献标识码:**A **中图分类号:**TP18

1 前言

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任务。在规则提取、特征分类等很多算法中,特别是应用粗糙集理论方法进行数据挖掘的研究和应用中,离散化是处理连续属性的一个有效的技术。现今的很多分类算法都是用来处理离散的数据或者二元的数据,可是,在实际应用中,有很多数据都是连续数据,算法无法直接处理,这就要求对属性所取的连续值进行离散化处理,将其变为离散的符号量。优良的离散化,应使划分尽可能简约,又尽可能多地保留有样本数据代表的对象的固有特性[1]。

离散化在于选定划分断点,依据选择方式、有无类别属性以及是否基于全部样本数据确定断点等, 离散化算法可分为:合并和分割、有监督和无监督、全部和局部等类型。最常用的是由 Kerber[2]提出的 ChiMerge 方法,应用统计学中的皮尔逊统计量 χ^2 来判别当前断点是否应该被去掉, 即与该断点相邻的两个间隔是否该合并。Liu 等人[3]在此基础上提出了 Chi2 系列算法,通过设定不一致率及自由度的变化,以及对断点合并标准的变化,不断的改进了 ChiMerge 算法。

早期的 Chi 算法是局部的、有监督的合并算法,在判断样本取值点两边的间隔与类别的相关性是否显著时,需先设定合适的显著性水平。改进后的 Chi2 算法用不一致率检验离散化程度, 优于显著性水平设定的 Chi 算法。但 Chi2 算法中是以 Pawlak RS 模型为基础选取的不一致率,在数据处理中,抛弃了许多还是有用的信息,本文中在此算法的不一致率选取中引用了 Bayseian 模型。Bayseian 模型是在基本粗糙集模型的基础上允许一定程度的错误分类率存在,一方面完善了近似空间的概念,另一方面也有利于粗糙集理论从认为不相关的数据中发现相关的数据。在数据挖掘中 Bayseian 模型的主要任务是解决属性无函数或不确定关系的数据分类问题,为处理由于噪声所引起的数据不一致性问题上提供了很好的方法。将 Bayseian 模型与 Chi2 算法相结合,在判断相邻区间是否合并时,新模型的标准不仅更适合不协调和不完全的数据,还使得区间的合并更加合理。实验结果验证了算法的有效性。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60372071);辽宁省教育厅高等学校科学研究基金(No.2004C031);辽宁师范大学校基金。

作者简介:刘磊(1982-),男,硕士研究生,主要研究领域为数据挖掘和模式识别;闫德勤(1962-),男,博士,教授,主要研究领域为模式识别、数字水印、信息安全、数据挖掘等;桑雨(1982-),男,硕士研究生。

收稿日期:2007-09-19 **修回日期:**2007-11-30

2 Chi2 算法

对于连续属性由于取值的有限性，其值间自然形成间隔，又称区间。取值点称为节点。如当连续属性取值为$\{b_1,b_2,\cdots,b_n\}$时，就形成了区间$[b_1,b_2],(b_2,b_3],\cdots,(b_{n-1},b_n]$。$b_i(i=1,2,\cdots,n)$为节点，简称为 i 节点。在 Chi2 及相关算法中，连续属性离散化就是按一定的方式根据一定的准则合并区间（即取消中间节点，简称合并中间节点）。

对于连续属性离散化，在 Chi2 及相关算法中，从 ChiMerge 算法到 Extended Chi2 算法都用到统计量 χ^2。χ^2 的计算方法为：

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})2}{E_{ij}} \tag{1}$$

其中，k 为类别数，A_{ij} 为 i 区间中 j 类样本的个数，$E_{ij}=R_i\times C_j/N$。$R_i=\sum_{j=1}^{k}A_{ij}$ 为 i 区间中样本数，$C_j=\sum_{i=1}^{2}A_{ij}$ 为 j 类样本的个数，$N=\sum_{i=1}^{2}R_i$ 为总样本个数。

在公式（1）中，C_j/N 是 j 类样本在总体中占的比例（可看作概率），$E_{ij}=R_i\times C_j/N$ 则是在这样的比例（概率）下 i 区间中应有 j 类样本的个数。因此公式（1）的统计量 χ^2 反映的是节点相邻的两个区间中 j 类样本分布的均匀程度。χ^2 越小，说明越均匀，节点就越不重要，应该被合并。

由数理统计中的皮尔逊定理[4]知，统计量 χ^2 的渐进分布是自由度为 $k-1$ 的 χ^2 分布，即 $\chi^2_{(k-1)}$ 分布。当给定显著水平 α 时，可确定相应的临界值 χ^2_α。在 ChiMerge 算法中，给定固定的显著水平，根据节点对应的统计量 χ^2 与临界值之差的大小决定节点是否被合并（对应差值最大的节点被合并）。在 Chi2 算法中，不是使用固定显著水平，而是使显著水平 α 不断下降，从而临界值 χ^2_α 不断增大，合并使差 $D=\chi^2_\alpha-\chi^2$ 最大的节点。这种处理方法大大地提高了计算效率和信息的利用率。然而，Chi2 算法应用 $\chi^2_{(k-1)}$ 分布时，使用的是固定自由度：$v=k-1$，其中 k 是系统的总类别数。事实上，自由度的选取应该依据节点相关联的类别数，而不应该是系统的总类别数，这是 $\chi^2_{(k-1)}$ 分布的要求。因此 Modified Chi2 算法[3]对此进行了修正。

当选取 k 为节点对应的类别数而不是系统的总类别数时，不同节点对应 $\chi^2_{(k-1)}$ 分布的自由度就会不同，因此在相同的显著水平 α 下临界值 χ^2_α 就不同。那么用 $D=\chi^2_\alpha-\chi^2$ 的大小来确定节点是否被合并就没有根据，或者说没有比较节点重要性的共同标准。为解决这个问题，Extended Chi2 算法[5]进行了改进：用 $D/\sqrt{2v}$ 代替 Modified Chi2 算法中的 $D=\chi^2_\alpha-\chi^2$ 进行计算。

3 Bayesian 粗糙集模型及 Bayesian-Chi2 算法

原始的 RS 模型（常称为 Pawlak RS 模型）是建立在二元等价关系的基础上的，但由于实际问题的需要，RS 模型的应用受到了限制。另一方面，RS 模型是基于可利用信息的完全性的，因而忽视了可利用信息的不完全性和可能存在的统计和随机信息，这类模型对于不协调的决策表的规则提取往往显得无能为力，因此，Slezk[6]等提出了一种 Bayseian 粗糙集模型，该模型以概率表示为基础，给出了正域、负域荷边界的定义方法。

设等价关系 R 在论域上 U 的等价类为$\overline{R}=\{E_1,E_2,\cdots,E_k\}$。$X$ 为论域 U 上的集合，$P(X)$是以论域 U 为样本空间，集合 X 为样本点的概率表示。令 $K(X)=\max\{P(X),1-P(X)\}$关于集合 X 的正域，负域和边界定义为：

$$POS(X)=\cup\{E\in\overline{R}:P(X|E)>K(X)\}$$

$$NEG(X)=\cup\{E\in\overline{R}:P(X|E)<1-K(X)\}$$

$$BND(X)=\cup\{E\in\overline{R}:1-K(X)<P(X|E)<K(X)\}$$

有近似精度

$$\eta=POS(X)/POS(X)+NEG(X)$$

不一致率

$$\delta=1-\eta=NEG(X)/POS(X)+NEG(X)=\frac{\cup\{E\in\overline{R}:P(X|E)>K(X)\}}{\cup\{E\in\overline{R}:P(X|E)>1-K(X)\}} \tag{2}$$

利用 Bayesian 模型解决属性间无函数或不确定关系的数据分类的优点，对 Extended Chi2 算法对数据不一致性容忍度低的不足进行了改进，增强了产生规则的鲁棒性，提高了算法的预测精度。改进后的 Bayesian Chi2 算法如下：

（1）初始化。令 $\alpha=0.5$。用公式（2）计算系统不一致率 δ_0。

（2）对所有属性排序数据并根据公式（1）计算 χ^2。再计算差异 D，寻找 D 最大的断点进行合并。计算合并后的不一致率 δ，检查 $\delta<\delta_0$ 是否成立，若是，继续（2），否则，进至下一步。

（3）令 $\alpha_0=\alpha$，对 α 降级，更新差异 D，若 α_0 不是最后一级，则返回（2），是则进至下一步。

（4）对每个属性 $c_1,c_2,\cdots,c_n$ 单独排序后，分别处理单个属性内差异：

①再当前属性内寻找 D 最大的断点进行合并。计算合并后的不一致率 δ，检查 $\delta<\delta_0$ 是否成立，若是，继续在这个属性内寻找符合条件的断点，否则进至②。

②对 α 降级，继续合并，若 α 到最后一级，则看当前属性是否为最后一个待处理的属性，若不是，则继续①，即开始处理新的属性，否则算法中止。

4 实验与结果

对 Extended Chi2 算法与改进后的 Bayesian Chi2 算法对 Glass、Breast 和 Wine 数据集分别进行离散化处理，进行了对比实验。数据集的数据信息如表 1 所示。

表 1 数据集

	条件属性	决策类	样本数
Iris	4	3	150
Glass	9	7	214
Breast	9	2	683
Wine	13	3	178

对离散化后的数据应用 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。对统计平均正确识别率、错误识别率和拒绝识别率，以及平均决策树节点个数（node）和提取出规则的平均个数（rule）进行对比，对比结果见表 2。其中 Extended 与 Bayesian 分别指 Extended Chi2 与Bayesian Chi2 算法。

与 Extended Chi2 算法相比，改进后的算法正确识别率有所上升，平均决策树节点个数和提取出规则的平均个数均有下降，显示了改进后的 Bayesian Chi2 算法的优势。

（下转 43 页）