文章编号: 1001 - 9081(2005)03 - 0637 - 02

一种基于熵的连续属性离散化算法

贺 跃[1],郑建军[2],朱 蕾[1]
(1.北京理工大学 信息科学技术学院,北京 100081; 2.北京理工大学 管理与经济学院,北京 100081)
(zjj76983@bit.edu.cn)

摘 要:连续属性离散化的关键在于合理确定离散化划分点的个数和位置。为了提高无监督离散化的效率,给出一种基于熵的连续属性离散化方法。该方法利用连续属性的信息量(熵)的特性,通过对连续属性变量的自身划分,最小化信息熵的减少和区间数,并寻求熵的损失与适度的区间数之间的最佳平衡,以便得到优化的离散值。实验表明该算法是行之有效的。

关键词:熵; 连续属性; 离散化; 分类
中图分类号: TP311.13 **文献标识码**: A

An entropy-based algorithm for discretization of continuous variables

HE Yue[1], ZHENG Jian-jun[2], ZHU Lei[1]
(1. *School of Information Science and Technology, Beijing Institute of Technology, Beijing* 100081, *China*;
2. *School of Management and Economics, Beijing Institute of Technology, Beijing* 100081, *China*)

Abstract: It is very important to ascertain rationally the number and positions of split points for discretization of continuous variables. To improve the efficiency of unsupervised discretization, an entropy-based algorithm was proposed for discretization of continuous variables. It made use of the characteristics of the information content(entropy) of a continuous variable, and partitioned the continuous variable by itself for minimizing both the loss of entropy and the number of partitions, in order to find the best balance between the information loss and a low number of partitions, so then obtained an optimal discretization result. The experiments show this approach effective.

Key words: entropy; continuous variable; discretization; classification

连续属性的离散化是数据挖掘和机器学习的重要预处理步骤,直接关系到学习的效果。

在分类算法中,对训练样本集进行离散化预处理,具有双重意义。一方面,可以有效降低学习算法的复杂度,加快学习速度,甚至提高学习/分类精度;另一方面,还可以简化、归纳获得的知识,提高分类结果的可理解性。正因为如此,离散化问题得到了较为广泛和深入的研究[1]。

根据是否利用类信息,连续属性的离散化方法可以分为有监督和无监督两种。与有监督离散化不同,无监督离散化可以处理不存在类别属性的数据集。

等宽和等频区间法是常见的无监督离散化算法,虽然都易于实现,但因为忽视了样本分布信息,因而难以将区间边界设置在最合适的位置上,从而使得它们的性能在大多数情况下无法令人满意[2]。

鉴于无监督离散化与数据聚类在目标上的近似,无监督离散化过程中也常采用 K-means等聚类分析算法。但对于连续属性的离散化而言,采用欧几里德距离作为区间划分的依据,尚缺乏理论根据[3],同时,由于需要依靠用户来指定区间数目,从而不能自动确定最合适的离散区间数。

此外,还有 Fayyad和 Irani[2]提出的基于熵的方法,以及 Hong[4]的基于矩阵的离散化方法等。

总的来看,目前的离散化算法还缺乏统一的理论指导,存在的主要问题[5]是候选分割点的选择带有主观色彩,某些离散算法的效率也值得考虑。

为了提高无监督离散化的效率,给出一种基于熵的连续属性离散化方法,该方法利用连续属性的信息量(熵)的特性,通过对连续属性变量的自身划分,最小化信息熵的减少和区间数,并寻求熵的损失与适度的区间数之间的最佳平衡,以便得到优化的离散值。实验表明了该方法的有效性。

1 连续属性的熵特性

连续属性离散化的直观含义是:首先为被离散的连续属性选定离散值数目,寻找一些划分点把连续属性的连续取值范围划分成一些子区间,每个子区间对应于一个离散值,这样就可以用一些离散的取值点来表示这个连续属性的整个取值范围。

对于数据库中任意一个连续属性,将它的取值范围划分为若干区间,每个区间至少包含一个样本。m 个样本至多分成 m 个区间($O(m)$)。这样,可将连续属性变量转换成具有 $O(m)$ 个值的离散变量。

当频率(或概率)分布拥有最大的属性值个数时,熵(或信息)被最大化[6]。既然连续属性分布中的每个不重复属性值对应地有一个离散区间,那么数据库样本就没有信息(或熵)的损失。

定义 1 一个离散随机变量 X 的熵定义为:

$$H(X) = -\sum_{x\in X} p(x)\log_2 p(x) \tag{1}$$

上式也可记为 $H(p)$。

收稿日期: 2004 - 08 - 19;**修订日期**: 2004 - 12 - 06

作者简介:贺跃(1950 -),女,辽宁沈阳人,副教授,主要研究方向:数据挖掘、智能图像处理; 郑建军(1969 -),男,河北宣化人,博士,主要研究方向:数据挖掘、信息系统分析与设计、智能决策支持系统; 朱蕾(1980 -),女,河北平山人,硕士研究生,主要研究方向:数据挖掘.

凸函数的定义[6]如下。

定义 2 函数 $f(x)$ 在区间 (a, b) 上是凸的,如果对于 (a, b) 内的每个 $x_1, x_2, 0 \leq \lambda \leq 1$,有:

$$f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2) \quad (2)$$

同时定义:函数 f 是凹的,如果 $-f$ 是凸的。可以证明,$H(p)$ 是 p 的凹函数。

在数据库中,设一个连续随机变量 X 的每个不重复的值可被一个分离区间表示,区间数为 $k, 1 \leq k \leq m$, m 是样本数;设每个区间中,X 的概率是 $p(i), 1 \leq i \leq k$;设 k 个离散区间的分布的熵是 $H(p_k)$。据此,并在上述定义的基础上,可以证明[7]:如果两个毗邻区间 $i, i+1$ 被合并,使得 $H(p_k) - H(p_{k-1})$ 最小,那么,被合并区间的概率 $p(i) + p(i+1)$ 是单调非减的。

另外,在数据库中,如果 X 是一个具有 k 个不重复值 $(1 \leq k \leq m)$ 的连续随机变量,且每个不重复值对应一个区间,那么,$H(p_k)$ 是 k 上的一个凹函数,此时,减少区间数的选择可以用来最小化 $H(p_k)$ 的改变量[7]。

2 一种基于熵的连续属性离散化方法

连续属性离散化的关键在于合理确定离散化划分点的个数和位置[8]。其离散化结果应满足下列两点:1)离散化后的空间维数尽量小,也就是每一个离散化后的属性值的种类尽量少;2)属性值被离散化后的信息丢失尽量少。

这里考虑一种基于熵的连续属性离散化方法,即对于数据库中的每个连续属性,先将它的取值范围划分为若干区间,每个区间对应一个不重复值;然后选择两个毗邻区间进行合并,使合并前后的熵差最小。

在上述过程中,如果合并前后具有最小熵差的毗邻区间超过了一对,那么随机合并其中一对。重复这一合并步骤,直到满足条件后停止,同时存储定义区间的划分点。

确定最佳停止点的判断如下:

由上一节可知,熵 $H(p)$ 作为凹函数,是随区间数 k 的增加而单调增加的,并且当 k 逼近最大值时,熵的增加率总是在减小[7]。同时,在熵的函数曲线上,熵取最大值的点(不妨设为 v_1 点),对应的区间数 k 也取最大值。

因此,如果从熵的函数曲线的起始点(设为 v_2)到 v_1 点画一条直线段 l,那么该凹函数曲线上的所有点都会落在 l 的上方。根据凹函数曲线的特性可以判断,在曲线上距 l 最远的一点(设为 v_0)处,即函数曲线的拐弯处,k 的变化变得大于熵的变化,此时,达到了熵损失和适度的区间数之间的最佳平衡。这样,v_0 点即可作为合并毗邻区间的停止点。

过函数曲线上的任意一点作到 l 的垂直线段 h,可得:

$$h = (k_{\max} - 1) H(p) - H_{\max}(p) \quad (k - 2) \quad (3)$$

式中,$k_{\max}$ 表示最大区间数,即不重复属性值的个数,$H_{\max}(p)$ 表示对应的最大熵值。

显然,在 v_0 点处,h 取最大值,据此即可求出 v_0 点及其对应的区间数。

综上所述,可得基于熵的连续属性离散化算法(EADC):

输入 数据库 D

输出 D 中所有连续属性的离散值和划分点

```
for(每个连续属性){
  对于 k个不重复属性值,计算概率,保存划分点,并利用(1)式
  计算熵 H(p_k);
  令 Loop = TRUE; k_0 = k; C_k = 0;
  while(Loop){
    选择 2个毗邻区间进行合并,使合并前后的熵差最小,并重
    置划分点,保存合并后的熵;
    计算 C_{k-1} = (k_0 - 1) * H(p_{k-1}) - H(p_{k_0}) * (k - 2);
    if (C_{k-1} > C_k) {k--;重新计算区间概率;}
    else {Loop = FALSE; k--;保存划分点;}
}}
输出所有连续属性的离散值和划分点
```

3 实验结果

为了说明上述算法的有效性,采用算法 EADC 对 UCI 机器学习数据库[9]中的 4个只含连续属性的样本集进行无监督离散化预处理,并进行分类学习。这些样本集如下所示:

1) Breast(Breast-cancer-wisconsin)。该数据集含有 699个实例、9个条件属性和 2个类。

2) Glass(Glass identification database)。该数据集包含 214个样本,9个条件属性和 6个类。

3) Iris(Iris plant database)。在分类问题中被广泛用于测试,包含 150个样本,4个条件属性和 3个类。

4) Thyroid-disease(New-thyroid)。包含 215个样本,5个条件属性和 3个类。

针对每个数据集,实验过程分三步进行:1) 将整个数据集离散化;2)把离散化后的数据集中的样本随机分为两组,70%作为训练集,剩下的 30%作为测试集;3)利用训练集对朴素贝叶斯分类器进行训练,再用测试集对训练后的贝叶斯分类器进行分类精度测试。按照这种方法,针对每个数据集,各进行了 5次实验。

例如,用算法 EADC 对数据集 Breast-cancer-wisconsin 进行离散化,以其中的属性 Bland Chromatin 为例,该属性共有 10个不重复样本值,其对应概率如图 1所示。

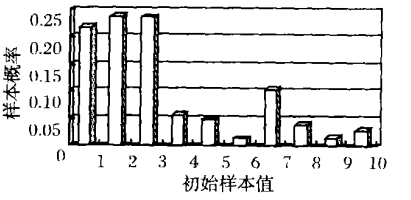

图 1 属性 Bland Chromatin 的样本概率

经算法 EADC 离散后,可得 3个划分点:1.5,2.5 和 3.5,而离散化所得 4个区间 (0,1.5], (1.5,2.5], (2.5,3.5] 和 (3.5,∞),可分别取离散值为 0,1,2 和 3。相应概率如图 2所示。

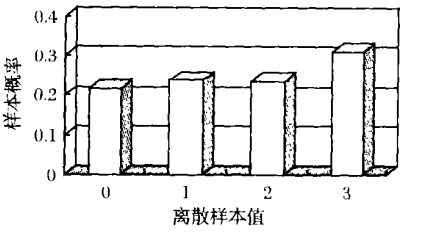

图 2 属性 Bland Chromatin 离散化后的样本概率

采用 EADC 对上述数据集进行预处理和朴素贝叶斯分类。为了便于比较,同时还采用 K-means 算法(将诸属性离散化后的区间数都取为 5)以及传统的基于熵的离散化方法[2](CEADC)对上述数据集进行预处理和朴素贝叶斯分类。

由表 2可以看出,对于上述 4个数据集而言,多数情况下,经算法 EADC 离散化处理后,贝叶斯分类器的分类精度都

(下转第 651页)

表 1　挖掘出的几条规则

规则前件	规则后件	规则支持度 (%)	置信度 (%)
E D	E	4. 41	100
D E	E	3. 03	100
F D	E	3. 03	100
E A	E	5. 30	100
E I	E	4. 55	100

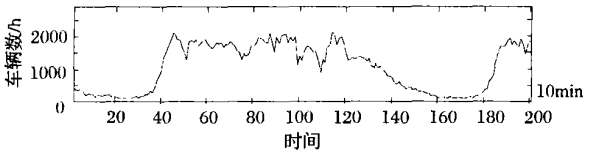

图 2　交通流量图

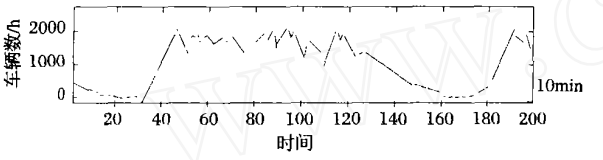

图 3　线性分段后获得的离散的符号序列图

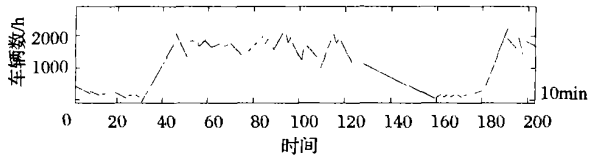

图 4　由聚类中心还原所得的时间序列分段表示

我们用 2002年 11月 11日至 17日的朝阳门桥北的交通流量数据进行了验证。情况如下:

规则 (ED⇒E)前件匹配到 9次,后件预测结果预测对了 7次,有 2次错误,预测的后验有效度为 77. 8%。

规则 (DE⇒E)前件匹配 7次,后件预测结果预测对了 3次,4次错误,后验有效度为 42. 9%。

规则 (FD⇒E)前件匹配 4次,后件预测结果预测对了 3次,1次错误,后验有效度为 75%。

规则 (EA⇒E)前件匹配 15次,后件预测结果预测对了 11次,4次错误,后验有效度为 74. 4%。

规则 (EI⇒E)前件匹配 5次,后件预测结果预测对了 5次,后验有效度为 100%。

由此我们可以看到,通过这种挖掘方法发掘出几条有意义的规则,可以对未来的趋势作出一定的分析推测。

4　结语

本文在原有的状态推演数据挖掘框架之下,提出了一种对交通流量数据进行数据挖掘的方法,并且给出了挖掘的流程,而后在实验环节对这种方法的有效性进行了验证。

在线性分段化的环节,对分段算法进行了一定的改进,使得分段误差有明显降低。

挖掘出的规则是否有效在很大的程度上取决于每一步的参数选择,包括分段数、平均误差指数以及聚类的簇个数等,这个问题有待于进一步解决。

参考文献:

[1] 李作敏,黄中祥,张亚平. 高速公路交通流分形特性分析 [J]. 中国公路学报, 2000, 13 (3).

[2] 张保稳. 时间序列数据挖掘研究 [D]. 西北工业大学, 2002.

[3] DAS G, LIN K, MANNILA H, *et al*. Rule Discovery from Time Series[M]. KDD, 1998. 16 - 22.

[4] 李斌,谭立湘,章劲松, 等. 面向数据挖掘的时间序列符号化方法研究 [J]. 电路与系统学报, 2000, 5 (2): 9 - 14.

[5] HAN JW, KAMBER M. 数据挖掘概念与技术 [M]. 范明,孟小峰,译. 北京:机械工业出版社, 2001.

[6] KAUFMAN L, ROUSSEEUW PJ. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wiley & Sons, 1990.

[7] AGRAWAL R, SRIKANT R. Mining Sequential Patterns: Generalizations and performance improvements[Z]. IBM Almaden Research Center, 1996.

(上接第 638页)

比用 K-means算法和 CEADC处理后的精度有所提高,只有在 Glass中, EADC处理后的精度较差,这表明 EADC是有效的。

表 1　Breast-cancer-wisconsin采用 EADC的离散化结果

属性名称	不重复样本数	划分点个数	划分点
Bland Chromatin	10	3	1. 5, 2. 5, 3. 5
Bare Nuclei	10	3	1. 5, 4. 5, 9. 5
Clump Thickness	10	3	1. 5, 3. 5, 5. 5
Marginal Adhesion	10	3	1. 5, 5. 5
Mitoses	9	2	1. 5, 3. 5
Normal Nucleoli	10	2	1. 5, 7. 5
Unif. of Cell Size	10	2	1. 5, 5. 5
Unif. of Cell Shape	10	3	1. 5, 3. 5, 7. 5
Single Epithelial Cell Size	10	3	1. 5, 2. 5, 4. 5

表 2　3种方法在分类结果上的比较

数据集	K-means	CEADC	EADC
Breast-Cancer	0. 889	0. 762	0. 986
Glass	0. 593	0. 487	0. 450
Iris	0. 800	0. 908	0. 976
Thyroid-disease	0. 751	0. 632	0. 839

需要指出的是,算法 EADC可以用于多种需要离散化连续属性的知识发现问题中,但由于它的计算量较大,因而适合于小规模数据集或规模缩减后的数据库。

参考文献:

[1] KURGAN LA, CIOS KJ. CAIM discretization algorithm[J]. IEEE Transactions on Knowledge and Data Engeering, 2004, 16(2): 145 - 153.

[2] USAMA FM, KEKI IB. Multi-interval discretization of continuous-valued attributes for classification learning[A]. Proceedings of the 13th International Joint Conference on Artificial Intelligence [C]. San Mateo, CA: Morgan Kaufmann, 1993, 2. 1022 - 1027.

[3] 李刚, 童频. 基于混合概率模型的无监督离散化算法 [J]. 计算机学报, 2002, 25 (2): 158 - 164.

[4] HONG SJ. Use of contextual information for feature ranking and discretization[J]. IEEE Transactions on Knowledge and Data Engeering, 1997, 9(5): 718 - 730.

[5] 苗夺谦. Rough Set理论中连续属性的离散化方法 [J]. 自动化学报, 2001, 27(3): 296 - 302.

[6] COVER TM, THOMAS JA. Elements of information theory[M]. New York: John Wiley & Sons, 1991.

[7] CLARKE EJ, BARTON BA. Entropy and MDL discretization of continuous variables for Bayesian belief networks[J]. International Journal of Intelligence Systems, 2000, 15(1): 61 - 92.

[8] ISHIBUCHI H, YAMAMOTO T. Deriving fuzzy discretization from interval discretization [A]. The IEEE International Conference on Fuzzy Systems[C], 2003. 749 - 754.

[9] BLAKE CL, MERZ CJ. UCI repository of machine learning databases [DB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.