一种近似等频离散化方法

蒋盛益， 李 霞， 郑 琪
(广东外语外贸大学信息学院，广东 广州 510006)

[摘 要] 在数据挖掘和机器学习研究中，许多算法以离散值为处理对象，常常需要对连续属性进行离散化. 由于正态分布的广泛性，本文提出一种基于正态分布的近似等频离散化方法. 该方法实现简单，关于数据集大小具有线性时间复杂度，适用于大规模数据集. 在许多数据集上与文献中多个离散化方法进行了对比测试，实验结果表明，提出的无指导的离散化方法是有效、可行的.

[关键词] 正态分布； 离散化； 等频方法
[中图分类号] TP311.13 [文献标识码] A [文章编号] 1000-9965(2009)01-0031-04

Approximate equal frequency discretization method

JIANG Sheng-yi, LI Xia, ZHENG Qi
(College of Information, Guangdong University of Foreign Studies, Guangzhou 510006, China)

[**Abstract**] Many algorithms for data mining and machine learning require that training examples contain only discrete attributes. In order to use these algorithms when some attributes have numeric attributes, the numeric attributes must be converted into discrete attributes. Because of the extensiveness of normal distribution, an approximate equal frequency discretization method which based on normal distribution is presented. The method is simple to implementation. Time complexity of the presented discretization method is nearly linear with the size of dataset and can be used to large dataset. The experimental results on real datasets show that the discretization method is effective and practicable.

[**Key words**] normal distribution; discretization; equal frequency method

所谓离散化是指将连续属性的值域划分为若干子区间，每个子区间对应一个离散值，最后将原始数据更新为离散值. 连续属性的离散化是许多数据挖掘和机器学习算法的重要预处理步骤，有效的离散化方法不仅可以减少系统对存储空间的实际需求，提高数据挖掘、机器学习算法的效率，而且在离散化的数据集上获取的知识往往具有更简洁的表达形式，更易于理解和使用. 研究表明，求取连续属性值的最优断点集合是一个NP完全问题[1]，离散化结果的优劣不仅与离散化算法本身有关，而且还和需要离散的数据分布及划分点的数目有关，同样的离散化方法应用于不同的数据集上，结果的差异也可能很大. 离散化方法的优劣只有通过离散化后的数据在后续的学习过程运用中的情况才能体现出来，因此离散化的好坏还与随后所采用的归纳算法有关.

[收稿日期] 2008-09-25
[基金项目] 国家自然科学基金项目(60673191)；广东省高等学校自然科学研究重点项目(06Z012)；
广东外语外贸大学科研创新团队项目(GW2006-TA-005)
[作者简介] 蒋盛益(1963-)，男，教授，博士，研究方向：数据挖掘，网络安全

针对连续属性的离散化,典型的方法有等宽划分方法(EW)、等频率划分方法(EF)、统计检验方法[2-6]、信息熵方法[7-11]及基于聚类的方法[12-13]等,这些算法都可归结为利用选取的分点对连续属性构成的空间进行划分,得到有限个区域,并用符号对每个区域进行编码. 这些算法根据离散化处理时是否以类别信息做参考,而分为有监督离散化和无监督离散化算法. 研究表明[14],由于没有利用蕴含在数据集中的类别属性值,对于分类问题而言,无监督离散化算法的效果要略逊于有监督离散化算法. 但是对于一些不存在类别属性的数据集而言,就无法用有监督离散化算法. 本文研究无监督离散化方法.

1 近似等频算法

1.1 等频算法描述

等频算法是无监督离散化算法,是将数值属性的值均匀地划分到若干区间. 如果属性在整个取值区间内共有 N 个点,划分区间为 k 个,那么每个区间含有 N/k 个点. 一种等频离散化算法实现过程如下:

设有 N 个点的序列:$\{x_i \mid i=1,2,\cdots,N\}$,将其升序排列,结果仍用 $\{x_i \mid i=1,2,\cdots,N\}$ 表示,则 k 个划分点为:$b_i = x_{i*N/k}(i=1,2,\cdots,k-1)$,得到 k 个划分区间为:$(b_i,b_{i+1}](i=0,1,\cdots,k-1)$,最后将每个区间中的点用一个离散值代替.

这种实现策略的时间复杂度为 $O(N\log N)$,等频算法简单易于实现,但忽视了样本分布信息,在有些情况下难以将区间的边界设置在最合适的位置上,而导致性能不佳. 正态分布具有很好的特性,也是许多统计方法的理论基础,自然科学与行为科学中的许多统计量在大样本时近似地服从正态分布,比如同一种生物不同个体的身长、体重等指标、同一种种子不同颗粒的重量等. 因此,本文借助正态分布提出一种近似等频离散化方法.

1.2 近似等频算法描述

近似等频离散化算法 AEFD (approximate equal frequency discretization method)是基于数据近似服从正态分布的假设,对连续属性进行离散化. 若一个变量服从正态分布,则其观测值落在一个区间的频率与变量在一个区间取值的概率应该相同,利用正态分布变量的分位点将取值区间划分为若干区间,使每个区间的取值概率相同,进而得到离散区间.

假设将属性取值区间划分为 k 个区间:$(b_i,b_{i+1}](i=0,1,\cdots,k-1)$,这里 $b_0=-\infty$,$b_k=\infty$,使正态分布在每个区间取值的概率均为 $1/k$,离散化后,以一个符号表示区间 $(b_i,b_{i+1}](i=0,1,\cdots,k-1)$ 中的每一个属性值. 近似等频离散化算法由三步组成,具体步骤描述如下:

Step1:计算划分点,确定初始划分区间.

按照 $b_i=\bar{x}+Z_{\alpha_i}\cdot\sigma(i=1,2,\cdots,k-1)$ 计算划分点,这里 $\bar{x}$、σ 分别为属性值的平均值和标准差,Z_{α_i} 为标准正态分布 $\xi\sim N(0,1)$ 的分位点 $P(\xi\leqslant Z_{\alpha_i})=\alpha_i=\frac{i}{k}(i=1,2,\cdots,k-1)$.

利用得到的划分点,得到初始的 k 个划分区间 $(b_i,b_{i+1}](i=0,1,\cdots,k-1)$.

Step2:合并划分区间,将包含记录频率很低的区间合并到最接近的区间.

统计各个区间 $(b_i,b_{i+1}](i=0,1,2,\cdots,k-1)$ 包含的记录频数及包含记录的均值,从右往左搜索,当区间 $(b_i,b_{i+1}](i=0,1,2,\cdots,k-1)$ 包含的记录频率小于 $1/(3k)$,即不到应有频率的三分之一时,将该区间合并到最接近的区间,并修改相应区间的记录频数及包含记录的均值. 这个过程重复到没有区间合并为止,最后得到划分区间.

Step3:将连续属性值根据划分区间转换成离散值.

注:①如果知道数据的分布,Step1 改为指定分布的分位点,效果会更好;②为提高算法效率 Step2、Step3 两步可以采用二分法思想以确定每个记录所在区间;③对于多个数值属性可以同步离散化.

1.3 算法时间复杂度

Step1 的运算量是固定的,时间复杂度为 $O(1)$;Step 2,Step 3 都只需要简单扫描一趟数据集,即可知道每个区间所包含的对象,时间复杂度为 $O(N*m)$,这里 N、m 分别为数据集大小、需要离散化的属性个数. 因此整个算法需要扫描数据集两遍,总的时间复杂度为 $O(N*m)$,可用于大规模数据集的离散化. 近似等频方法离散化算法的时间复杂度低于等频算法和基于混合概率模型的方法.

2 实验结果分析

为检验提出的离散化方法的性能,选取 UCI 中的 15 个数据集[15]及一个实际工资数据集 Salary 进行离散化处理,选择的数据集的特征如表 1 所示,KDDcup99 数据集太大,从中随机选取 20 000 记录进行处理. 实验使用 Weka 软件[16]平台中 C4.5、等

频离散化方法和有监督离散化方法(即 Fayyad & Irani 的 MDL 方法).

表 1　实验数据集汇总

数据集	离散/连续属性个数	数据集大小	类别数
austra	8/6	690	2
Breast	0/9	699	2
Credit	9/6	690	2
diabetes	0/8	768	2
german	13/7	1 000	2
Glass	0/9	214	6
Heart	7/6	270	2
horse-colic	19/7	368	2
ionosphere	0/34	351	2
Iris	0/4	150	3
KDDcup99	7/34	494 020	23
Labor	0/8	57	2
Letter-recognition	0/16	20 000	26
Liver	0/6	345	2
Pima	0/8	768	2
Segment-test	0/19	2 310	7
Wine	0/13	178	3
Salary	1	80	4

Salary 数据集是某单位 80 名员工税后工资的数据,仅包含职称和税后工资两个属性,数据总体分布情况如表 2.

表 2　Salary 数据集中数据的分布

职称	人数	税后工资/元
教授	8	5 266 ~4 858
副教授	22	4 272 ~3 644
讲师	40	3 438 ~2 885
助教	10	2 702 ~2 420

对税后工资采用近似等频离散化算法对其进行离散化处理,取 $k=9$,最终得到 6 个划分点,分别是 2 466,2 797,3 035,3 239,3 619,4 392,离散化后得到 7 个区间. 易见离散化后教授及副教授分别对应 1 个区间,而讲师对应 3 个区间,助教对应 2 个区间.

对税后工资采用等频离散化算法对其进行离散化处理,得到 7 个划分点,分别是 2 793,2 913,2 974,3 099,3 541,3 990,4 155,离散化后得到 8 个区间,区间(4 155,5 266]包含教授、副教授两类对象,可见离散效果不如近似等频离散化算法.

对税后工资采用 MDL 方法对其进行离散化处理,得到 3 个划分点,分别是 2 793, 3 541,4 565,离散化后得到 4 个区间. 易见离散后 4 类不同职称分别对应不同的离散区间,达到了非常理想的状况.

对选取的 16 个数据集,分别采用近似等频方法(AEFD)、等频方法(EF)、MDL 方法、及 EMD[11]、FD[11]、HD[11]方法进行离散化处理,并对离散化前后的数据采用 C4.5 的 10 次交叉验证方法进行分类学习,并与相关文献中的算法的分类精度进行对比,测试结果见表 3.

表 3　在 C4.5 上的分类精度比较(1)　%

数据集	无监督离散化方法			有监督离散化方法			
	离散化前	AEFD	EF	MDL	EMD	FD	HD
Breast	94.56	96.70	94.99	95.71	96.60	91.50	95.80
german	70.50	72.90	71.70	72.10	70.60	71.80	73.10
Glass	72.90	63.55	51.40	74.77	68.60	69.20	70.10
Heart	76.67	78.15	73.33	81.48	80.20	78.30	75.10
horse-colic	67.93	66.84	67.12	66.30	85.30	81.50	82.70
Iris	94.00	94.67	92	93.33	94.50	95.60	96.30
6 个数据集平均	79.43	78.80	75.09	80.62	82.63	81.32	82.18
austra	84.93	84.20	84.93	85.22			
Credit	85.22	85.22	84.64	87.10			
diabetes	72.40	74.22	74.61	78.13			
ionosphere	91.45	86.86	87.75	89.17			
KDDcup99_20000	99.95	99.94	99.93	99.95			
Labor	73.68	73.68	64.91	80.70			
Letter recognition	88.20	81.14	77.60	78.76			
Liver	66.67	62.79	56.81	63.19			
Pima	73.96	75.91	74.35	77.73			
Wine 1	93.26	91.01	78.65	94.38			
Salary	93.75	100	97.50	100			
16 个数据集总平均	82.38	81.63	78.37	83.41			

按训练集与测试集各占 50%,使用 C4.5 分类器测试,AEFD、EF、MDL 及基于信息熵的粗糙集连续属性离散化算法(用 A1 表示)[10]在选取的 6 个数据集上的分类精度对比结果如表 4.

表 4　在 C4.5 上的分类精度比较(2)　%

数据集	无监督离散化方法			有监督离散化方法	
	离散化前	AEFD	EF	MDL	A1
austra	83.48	83.48	83.77	83.48	83.80
Glass	66.36	64.49	46.73	67.29	70
Heart	75.56	77.78	77.78	79.26	73.40
horse-colic	65.76	64.04	69.02	69.02	76.80
Iris	96.00	93.33	94.67	98.67	96
Pima	71.61	74.48	73.44	73.70	71.10
平均值	76.46	76.27	74.24	78.57	78.52

按训练集与测试集分别占60%、40%,使用Naive-Bayes分类测试,AEFD、MDL及基于混合概率模型的无监督离散化算法(用A2表示)[5]在选取的4个数据集上的分类精度对比结果如表5.

表5 在Naive-Bayes分类器上的分类精度比较 %

数据集	离散化前	AEFD	A2	MDL
Breast	96.07	97.50	90	96.43
diabetes	77.92	72.73	74.06	78.90
Glass	48.84	61.63	61.63	68.60
Iris	93.33	90.00	87	93.33
平均值	79.04	80.47	78.09	84.32

实验结果表明,近似等频方法较之等频离散化方法的性能有所改善,实际的划分点个数有所减少,而分类精度有所提高. 近似等频离散化方法与离散化前的平均分类精度非常接近,而优于等频方法和基于混合概率模型的无监督离散化算法. 较有监督的离散化算法的平均分类精度低1.2%~5.3%,这与文献[13]的结论类似.

3 结论

近似等频离散化算法是以正态分布理论为基础,根据现实数据在大样本情况近似服从正态分布的特性,对数据进行离散化. 算法思想朴素、没有复杂的原理,容易理解,实现简单. 算法不需要进行大量的分析和计算,具有线性时间复杂度. 从实验结果分析比较来看,虽然近似等频离散化算法思想朴素,但性能仍然很有效. 类似于等频离散化方法,近似等频算法也没有充分考虑样本分布信息,在有些情况下也难以将区间的边界设置在最合适的位置上. 进一步的研究工作将寻找有效识别一维数据不同分布密度区间的方法,发现自然的离散区间以更有效地实现离散化.

致谢:感谢陈宁同学实现了本文部分算法.

[参考文献]

[1] NGUYEN H S, SKOWRON A. Quantization of Real Values Attributes Rough Set and Boolean Reasoning Approaches. Proc. of the 2th joint Annual Conf on Information Sci[C]. USA Wrightsville Beach: NC, 1995: 34-37.

[2] KERBER R. Discretization of Numeric Attributes. The 9th International Conference on Artificial Intelligence [C]. AAAI Press/The MITS Press, Cambridge, MA, 1992: 123-128.

[3] LIU H, SETIONO R. Feature selection via discretization [J]. IEEE Transactions on Know ledge and Data Engineering, 1997, 9(4): 642-645.

[4] TAY E H, SHEN L. A modified Chi2 algorithm for discretization[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 666-670.

[5] 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.

[6] KONONENKO I. Inductive and bayesian learning in medical diagnosis[J]. Applied Artificial Intelligence, 1993, 7: 317-337.

[7] FAYYAD U M, IRANI K B. Multi-interval discretization of continuous valued attributes for classification learning. Proc of the 13th International Joint Conference on Artificial Intelligence[C]. Chambéry France: Morgan Kauman, 1993: 1022-1029.

[8] CLARKE E J, BRATON B A. Entropy and MDL discretization of continuous variables for Bayesian belief networks [J]. International Journal of Intelligence Systems, 2000, 15: 61-92.

[9] CHIU D K Y, CHENG B, WONG A K C. Information synthesis based on hierarchical maximum entropy discretization[J]. Journal of Experimental and Theoretical Artificial Intelligence, 1990, 2: 117-129.

[10] 谢 宏,程浩忠,牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报. 2005, 28(9): 1570-1574.

[11] LEE Chang-Hwan. A Hellinger-based discretization method for numeric attributes in classification learning [J]. Knowledge-Based Systems. 2007, 20(4): 419-425.

[12] 李兴生,李德毅. 一种基于密度分布函数聚类的属性离散化方法[J]. 系统仿真学报,2003.6:804-806.

[13] 席 静,欧阳为民. 基于聚类的连续属性最佳离散化算法[J]. 小型微型计算机系统,2000,21(10):1025-1027.

[14] DOUGHERTY J R, KOHAVI, SAHAMI M. Supervised and Unsupervised Discretization of Continuous Features. Machine Learning. Proc of 12th International Conference [C]. Morgan Kaufmann, 1995: 194-202.

[15] MERZ C J, MERPHY P. UCI repository of machine learning databases[EB/OL]. URL: http://www.ics.uci.edu/~mlearn/MLRRepository.html.

[16] Ian H. Witten and eibe frank. Weka[EB/OL]. http://www.cs.waikato.ac.nz/ml/weka/

[责任编辑:黄建军]