doi: 10.3969/j.issn.0490-6756.2018.05.014

数据挖掘中并行离散化数据准备优化

刘 云, 袁浩恒
(昆明理工大学信息工程与自动化学院, 昆明 650500)

摘 要: 在海量数据挖掘中,针对元数据的离散化数据准备处理能有效提高数据挖掘效率.本文提出了一种并行比较并获得最优离散化的数据准备算法(AOA),针对不同数据集,先进行数据集的特性检测以获得数据集分布特性,按照分布特性进行数据集的异常值检测和剔除,并行完成与分布特性适配的离散化方法处理,通过比较不同离散化方法的熵、方差指数、稳定性参数的最小欧氏距离,根据三个参数自动化比选,获得最优离散化的预处理成果.仿真表明,对不同样本数据库进行关联规则挖掘结果中,比较四种固定的离散化数据预处理方法,在使用 AOA 数据准备算法并行比选出最优的离散化来数据预处理后,在不同最小支持度阈值情况下,挖掘得到关联规则数都更少,因此效率得到提高.

关键词: 数据挖掘; 数据准备; 并行调用; 分布检测; 数据离散化
中图分类号: TN929.5 **文献标识码**: A **文章编号**: 0490-6756(2018)05-0993-07

Parallel discretization of data preparation optimization in data mining

LIU Yun, YUAN Hao-Heng
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: In data mining, the discretization of data can improve the efficiency of data mining effectively. In this paper, we propose a data preprocessing algorithm (AOA) to obtain the optimal discretization using parallel comparison. For different data sets, we first perform the feature detection of the data set to obtain the distribution characteristics of the data set. Then the outliers of the data set are detected according to the distribution characteristics. IN addition, the discretization results are obtained by comparing the minimum Euclidean distance of the entropy, the variance index and the stability parameter of the different discretization methods. In simulation experiment, we compare the AOA with four typical data discretization methods in different databases by running the association rule mining algorithm on the discretization data obtained using AOA and other four methods, respectively. The results show that, under different minimum support thresholds, the number of association rules extracted from the discretization data obtained using AOA is the least, indicating higher efficiency of AOA.

Keywords: Data mining; Data preparation; Parallel invocation; Distributed detection; Data discretization

1 引 言

并行体系结构[1]已经成为我们构想现代化架构的标准.由于处理器被设计提供平行指令,所以操作系统都为这一设计而构建,特别是多任务[2](包括多线程)管理.由于这一硬件条件支持,应用

收稿日期: 2017-10-24
基金项目: 国家自然科学基金(61262040)
作者简介: 刘云(1973-),男,云南昆明人,副教授,研究方向为无线通信研究. E-mail: liuyun@kmust.edu.cn
通讯作者: 袁浩恒. E-mail: 279660340@qq.com

程序可以分配到多个核心同时进行.因此,多核应用程序的运行速度更快,比给定要求处理器执行时间更少.因此,我们把这一并行体系思路融入到数据准备过程[3]中.

数据预处理[4],这一过程主要可以分为:数据清理,数据集成,数据变换,数据规约和数据离散化5个部分.因为更好的数据预处理,就会得到更好的挖掘结果.目前许多数据预处理方法,即使是少数优化的数据预处理文献中,只从某一方面进行预处理,并没有对数据离散化,或者只利用单一离散化方法进行数据准备.但在挖掘步骤中,对数据进行离散化将提高后续数据处理过程的速度和效果,对于数据准备来说有重要的意义.并且在离散化步骤中,没有任何离散化算法可以适用于任何环境下,在实际应用时,需要根据数据集的特点和学习环境等选择合适的离散化方法.

针对这一问题,本文提出一种基于数据清理与数据离散化的自动优化数据预处理方法(Auto Optimize Algorithm,AOA).AOA算法主要步骤为:先对数据集计算确定其分布情况,根据数据分布,先选择数据异常值检测方法来进行数据清理;再并行分别进行四种离散化方法离散;最后根据各方法离散前后定义的最小欧式距离自动选择调用最好离散化的方法来对数据进行离散化处理.使用AOA进行数据准备的数据集在进行APRIORI关联规则挖掘的过程中,对于不同最小支持度阈值的情况下,得到关联规则数更少,效率更高.

2 检测数据分布模型

对一给定数据集对其进行数据准备时,确定该数据集的分布函数[5](或者是密度函数)的形状是非常重要的.我们把函数的形状分为四类:(1) 多峰分布;(2) 对称或反对称分布;(3) 均匀分布;(4) 正态分布.我们按从(1)~(4)这一顺序依次进行分布检测,最终确定数据集分布.针对不同分布类型的数据集采用不同的数据准备方法会得到更好挖掘的效果.

我们下文的步骤中,对数据集做以下规定:X是一个关系数据库的数值列.x是任意的值;X_i是第i个值;N是X的大小;σ是其标准偏差;μ是平均值;S是中心倾向参数.X_1和X_N分别代表X的最小值和最大值,k指用户的参数,或由系统计算而得.

多峰分布:我们得到一个数据集,首先根据数据的统计数量来得到离散的样本密度图;然后使用曲线拟合[6](最小二乘法等)方法来构建数据集的密度函数;最后通过计算它的二阶导数为0的点数来计算峰的数量.只要峰的数量大于1,则该分布为多峰分布.

对称和反对称分布:在下一步中,我们检测分布是否为对称或反对称分布.我们通过使用偏度[7](Skewness)来进行检测.它是描述数据分布形态的统计量,其描述的是某总体取值分布的对称性.记为γ_3.

$$\gamma_3 = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] \tag{1}$$

式(1)中,如果$\gamma_3 = 0$,则分布是对称的.实际上,这个规则太精确了,我们放松这个条件,把0附近的极限都作为相对容忍区间,这允许我们判断一个分布是不是相对"反对称".该检测根据一个数据检测方法,设零假设[8]为:分布是对称的.构造统计量为

$$T_{\text{Skew}} = \frac{N}{6}(\gamma_3^2) \tag{2}$$

在零假设情况下,统计量T_{Skew}服从一个自由度的卡方分布.当$\alpha = 0.05$时,T_{Skew}大于3.8415,分布拒绝原假设,则该分布是反对称的.

均匀分布:仅当上两步中检测分布为单峰对称分布时,我们才执行这一步的检测,否则输出函数分布类型.在这一步中,我们使用标准化峰度来判定分布是否为均匀分布.记为γ_2.峰度(Kurtosis)是描述总体中所有取值分布形态陡缓程度的统计量.与正态分布相比,来测量分布的峰度或概率密度的分组概率.当γ_2(参见方程(3))接近于零,该分布具有一个标准化峰度.我们再次使用统计检验来确定分布是否有标准化峰度.设零假设为:分布有一个标准化峰度.

$$\gamma_2 = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] - 3 \tag{3}$$

构造统计量:

$$T_{\text{Kurto}} = \frac{N}{6}\left(\frac{\gamma_2^2}{4}\right) \tag{4}$$

在零假设的情况下,T_{Kurto}满足一个自由度的卡方分布.当$\alpha = 0.05$时,零假设的拒绝域$T_{\text{Kurto}} > 6.6349$,则该分布是均匀分布.

正态分布:仅当上三步中检测分布为单峰对称非均匀分布时,我们才执行这一步的检测.我们使用J-B检验[9]如下:JB统计量全称叫Jarque-Bera统计量,是用来检验一组样本是否能够认为来自正态总体的一种方法,它依据OLS残差,对大样本进

行检验(或称为渐进检验).

我们首先根据上文所述,计算出分布函数的偏度 γ_3 和峰度 $\gamma 2$. 对于正态分布变量,偏度为零,峰度为 3. 构建统计量 JB 为

$$JB = \frac{n}{6}(\gamma_3^2 + \frac{\gamma_2^2}{4}) \tag{5}$$

在正态分布的假设下,JB 统计量渐进地服从自由度为 2 的卡方分布. 若变量服从正态分布,则 γ_3 为零,γ_2 为 3,因而 JB 统计量的值为零;如果变量不是正态变量,则 JB 统计量将为一个逐渐增大值.

3　AOA 数据准备

3.1　异常值检测

在对数据集离散化之前,进行数据清洗[10],对异常值进行检测和剔除是很有必要的. 只有去除数据集上下界的"毛刺"值,才能使整个数据集更规范,在后续的离散化过程中分箱才能让数据集更规范,挖掘效率更高[11-13].

这里,我们只讨论以下两种异常值检测方法.

(1) 箱线图[14]:这种方法,创始于杜克,是根据四分位数之间的差 Q_1 和 Q_3. 它区分了两类极端值的下界(LB)和上界(UB),如下式.

$$\begin{cases} LB = Q_1 - k \times (Q_3 - Q_1) \\ UB = Q_3 + k \times (Q_3 - Q_1) \end{cases} \tag{6}$$

其中,第一四分位数(Q_1),又称"较小四分位数"或"下四分位数",等于该样本中所有数值由小到大排列后第 25% 的数字;第二四分位数(Q_2),又称"中位数",等于该样本中所有数值由小到大排列后第 50% 的数字;第三四分位数(Q_3),又称"较大四分位数"或"上四分位数",等于该样本中所有数值由小到大排列后第 75% 的数字. 第三四分位数与第一四分位数的差距又称四分位间距. 在上下界之外的数据将被视为异常值,并将其剔除.

(2)"格鲁布斯"测试[15]:格鲁布斯方法,为较低或较高的异常数据统计检验. 该测试是基于数据为正态分布,使用样品平均值和极值之间的差异来进行异常值检测与剔除.

将 x 按从小到大顺序依次排列从 X_1 到 X_N. 通常异常值即为 X_1 或 X_N. 检测所用的统计数据是

$$T = \max(\frac{X_N - \mu}{\sigma}, \frac{\mu - X_1}{\sigma}) \tag{7}$$

在显著水平 α 的情况下,检测值(X_1 或 X_N)不为异常值的拒绝域为

$$T > \frac{N-1}{\sqrt{n}} \sqrt{\frac{\beta}{n - 2\beta}} \tag{8}$$

其中,$\beta = t_{\alpha/(2n),n-2}$ 是服从 $\alpha/(2n)$ 的 $n-2$ 个自由度的 T 分布四分位数. 若检测值在拒绝域内则为异常值,对其进行剔除.

大多数现有的异常检测方法假设分布是正态的. 然而,我们发现在现实中,许多样本为不对称和多峰分布,如果使用的那些方法将会对数据挖掘步骤造成显著影响. 在这一情况下,本文采用对数据集进行 JB 检测,根据数据集是否为正态分布,分别采用线箱图或者格鲁布斯检测来进行异常值检测与剔除,如图 1 所示.

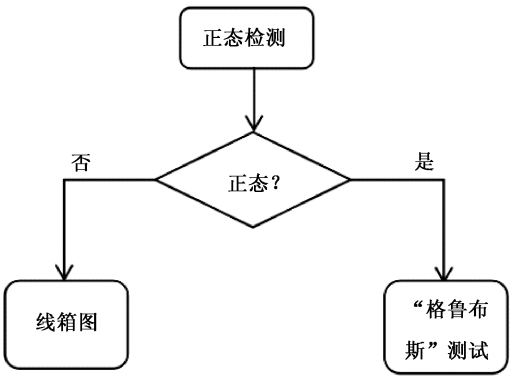

图 1　检测和消除异常值方法的选择过程

Fig. 1　Detects and eliminates the selection process for outliers

3.2　离散化方法

数据离散化[16]作为对数据集的预处理过程,其输出直接被用作随后进行的数据挖掘算法,如分类和预测算法的输入. 这些算法大多数是针对离散型数据的,对于连续型数据不适用;有些算法即使能够处理连续型数据,效果也不如处理离散型数据好. 在数据库系统中连续型受占多数,要更好地分析处理这些数据就有必要对这些数据进行离散化. 在本节中,我们主要讨论以下四种离散化方法.

(1) 等宽离散化法(EWD). 根据制定区间数目,将数值属性的值域划分 K 个区间,使每个区间宽度相等.

(2) 等频率费希尔-甄克思离散化(EFD-Jenks)等频离散化法,根据兹定于区间数目,按要求满足每个区间的对象数目相等.

(3) 基于平均值和标准偏差离散化(AVST). 根据均值和标准差以等宽的距离划分间隔.

(4) K-均值聚类(Kmeans)法[17]. K-均值聚类是一种应用广泛的算法,在用户指定了离散化产生的区间数目 K 后,K-均值法从 K 个区间中随机找出 K 个数据作为重点,然后根据重心的欧式距离,

对所有的对象聚类,然后重新计算各区间重心,并利用新的重心重新聚类,逐步循环直到所有区间的重点不再改变为止.

这些方法在已有论文中的已被广泛讨论.在前人的研究基础上,在图 2 中,我们总结给出了数据集分布与实用性匹配关系.

3.3 数据分布与离散化方法对应模型

不同的数据样本拥有不同的样本特性,对其进行离散化时,不同的离散化方法会得到不同的结果.为了能更有效的离散,并且针对不同离散化方法的适用情况.我们将上文所罗列的四种离散化方法和五种数据分布函数形状进行了适用性匹配.

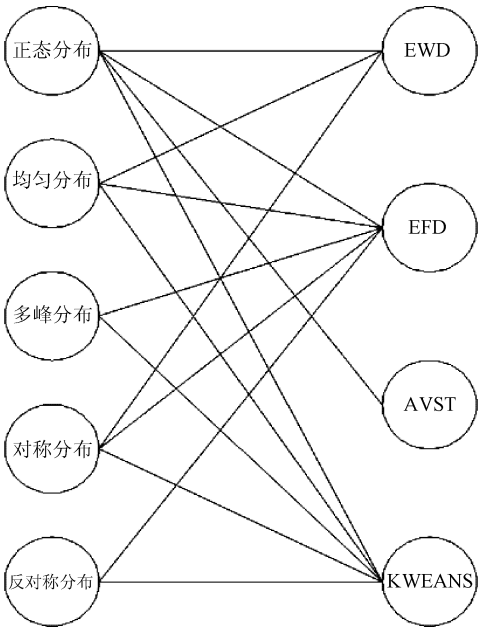

图 2 数据集分布与实用性匹配关系

Fig. 2 Data set distribution and practical match

为了取得更好的离散化效果.对于不同数据集的分布类型,我们只选用能够与之想匹配的离散化方法对其进行离散化处理.

3.4 多标准方法寻找最佳离散化方法

虽然已有很多离散化方法,但是没有一种离散化方法对任何数据集以及任何算法都是有效的,也没有一种离散化方法一定比其他方法产生更好的离散化结果.因为离散化本身就是一个非确定性问题,所以在使用时一定要根据数据集的特点、环境以及使用者个人的偏好理解等选择合适的离散化方法,以取得尽可能好的离散化效果.离散化必须保持最初的统计特性以保持间隔的均匀性,以及减少最终产生的数据的大小.所以如何进行离散化有许多目标相互矛盾.为此,我们选择了一个多标准分析法,来评估可应用的离散方法.我们用以下三个标准.

(1) 熵(H):熵测量的是间隔均匀性.熵越高,离散化在每个间隔中的元素数越多,如下式.

$$H = -\sum_{i=1}^{\mathrm{NbBins}} p_i \log_2(p_i) \tag{9}$$

其中,P_i是区间 i 内点的数量,i 是被划分为 N 个区间中的一个;NbBins 是区间数.当离散化使 NbBins 与区间内相元素数数量相同时;H 得到最大值.在这种情况下,H 为 $\log_2(\mathrm{NbBins})$.

(2) 方差指数(J):引用(Lindman,2012),测量组间方差与总方差的比.这个值越接近 1,离散化越均匀.

$$J = 1 - \frac{\text{Intra-intervals variance}}{\text{Total variance}} \tag{10}$$

(3) 稳定性(S):对应于离散化前后分布函数之间的最大距离.设 F_1 和 F_2 是离散化之前和之后的属性分布函数.

$$S = \sup_x(|F_1(x) - F_2(x)|) \tag{11}$$

我们的目标是在各项性能指标之间找到一种折中的解决办法.对这些方法进行自动评估.针对这一目的,我们采用聚合方法将多元素整合为单一元素判定标准.聚合方法在多标准分析中最广泛常用的方法.在这类方法中的关键是对每个标准加权和构建关联权重.它的缺点是增加了受限阈值,出于这些原因,我们选择距离理想目标的最小欧氏距离[18]的方法($H = \mathrm{LOG}_2(N_{\mathrm{bBins}})$,$J = 1$,$S = 0$).

定义 1 设 D 是一种任意的离散方法,V_D是一种使用多标准分析分割质量的测量方法.

$$V_D = \sqrt{(H_D - \log_2(N_{\mathrm{bBins}}))^2 + (J_D - 1)^2 + S_D^2} \tag{12}$$

下面的命题是本文的主要结果:它表明我们如何选择最好最合适的离散化方法.

命题 1 设 DM 是一组离散化方法;其中任意一个记为 D,当 V_D(见式(12))为最小值时,$\forall D \in \{\mathrm{DM}\}$,这一 V_D 所对应的离散化方法是最优的方法.

我们定义的 AOA(并行优化的数据的制备)方法,在 3.2 节中总结了四种方法针对每个判定属性,可否调用与评估情况.每一个可适用离散化方法被调用进行离散化,再计算各个方法的 V_D值,自动比较选取最小 V_D的方法为最优离散化法.AOA 算法如算法 1.

Algorithm 1: AOA: auto optimized algorithm

for data preparation

Input：X set of numeric values to discretize, DM set of discretization methods applicable

Output：Best set of bins for X

Parallel_Invoke For each method $D \in$ DM do Compute γ_2, γ_3 and perform Jarque-Bera test;

End

Parallel_Invoke For each method $D \in$ DM do remove D from DM if it does not satisfy the criteria given in Fig. 2;

End

Parallel_Invoke For each method $D \in$ DM do Discretize X according to D;

$V_D =$

$$\sqrt{(H_{\mathrm{D}} - \log_2(N_{\mathrm{bBins}}))^2 + (J_D - 1)^2 + S_D^2}$$

End

D=argmin($\{V_D \forall D \in \mathrm{DM}_{\cdot}\}$).

Return set of bins obtained line 8 according to D

举例 1

SX{ 59.04, 60.13, 60.93, 59.04, 60.13, 64.26, 70.34, 72.89, 74.42, 79.40, 80.46, 74.42 } 代表 12 个人的体重. 对这一数据集分别用四种上述方法进行离散化，并在下表中总结离散化方法的评估结果. 需要指出的是，为了举例需要，在这一例子中，没有经过 3.2 节中图 2 适用性选择步骤来筛选，而是对四种方法都进行离散化并分别计算权值评估结果，如表 1 所示.

表 1　四种离散化方法选择计算结果

Tab. 1　Four discretization methods Select the results

方法	熵(H)	方差指数(J)	稳定性(S)	最小欧氏距离(V_{DM})
EWD	1.5	0.972	0.25	0.313
EFD-Jenks	2	0.985	0.167	0.028
AVST	1.92	0.741	0.167	0.101
KMEANS	1.95	0.972	0.167	0.031

结果显示，DFD-jenks 和 Kmens 得到最小的 VD 值，即为最优方法. 而 EWD 和 AVST 最差. 因为该数据集为多峰分布，所以这一评估结果也与 3.2 节中总结的适用情况一致.

4　仿真分析

在本节中，我们通过对三个样本进行 APRIORI 算法进行关联规则挖掘. 我们样本 1 是随机生成的数据库. 样本 2 和样本 3 是对真实数据进行测量得到的数据库. 在表 2 中对样本情况进行了阐述.

表 2　样本属性

Tab. 2　Sample properties

样本	列数	元素数	类型
样本 1	10	468	生成
样本 2	1281	296	真实
样本 3	8	727	真实

实验在 4 核心(具有 2.8 GHz 处理器)和 12 Gb RAM 在 Windows 7 64 位操作系统下用 2010 C ++ Microsoft 进行.

实验为了比较不同离散度方法进行数据的准备情况，对三个样本数据库进行 APRIORI 算法关联规则挖掘. 设置最小置信度(Minconf)为 0.5，不同最小支持度(MinSup)阈值为变量. 随着最小支持阈值提高. 每个算法所得到挖掘关联规则数也随之减少. APRIORI 算法进行关联性规则挖掘，每一阶都得逐一扫描条件范围内的频繁项集. 最终挖掘得到关联规则数越少，则挖掘效率越高.

每个仿真图中反应了四种离散化方法(未进行数据清洗)离散后的挖掘效率及使用本文提出的 AOA 后的结果. 虽然离散化方法的选择规则与评价规则不为同一准则，但是 AOA 选比算法总是在四种之中选择最优离散化方法，故下文中只针对四种单一方法进行效率比较，AOA 给出最后结果.

对样本 1 进行挖掘时，随着变量最小支持度阈值逐渐增高，所有算法得到挖掘关联规则数也随之减少. 在全区间中对于四种离散化方法，EWD 离散化后挖掘效率最低，KMEANS 离散化后挖掘效率最高，如图 3 所示.

对样本 2 进行挖掘时，随着变量最小支持度阈值逐渐增高，所有算法得到挖掘关联规则数也随之减少. 在全区间中，EWD 离散化后挖掘效率最差，KMEANS 离散化后挖掘效率最高如图 4 所示.

对样本 3 进行挖掘时，实验表明用 AOA 算法进行数据准备，随着变量最小支持度阈值逐渐增高，所有算法得到挖掘关联规则数也随之减少. 在全区间中，EWD 挖掘效率最差，AVST 离散化后挖掘效率最高，如图 5 所示.

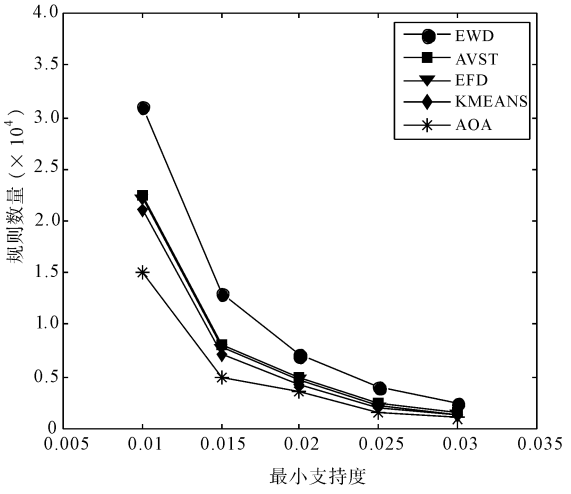

图 3 样本 1 的挖掘效率仿真结果

Fig. 3 Sample-based mining efficiency simulation results

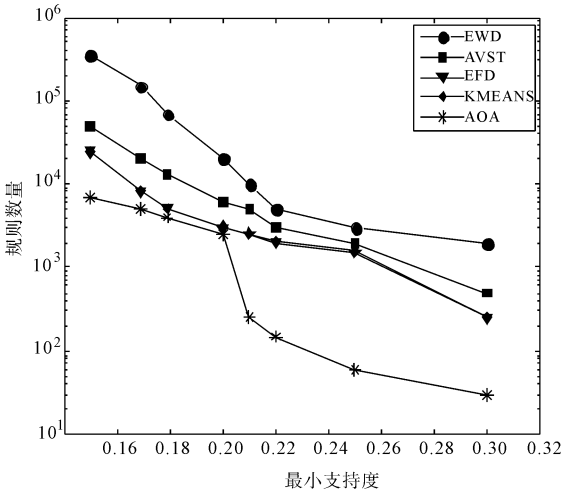

图 4 样本 2 的挖掘效率仿真结果

Fig. 4 Sample two mining efficiency simulation results

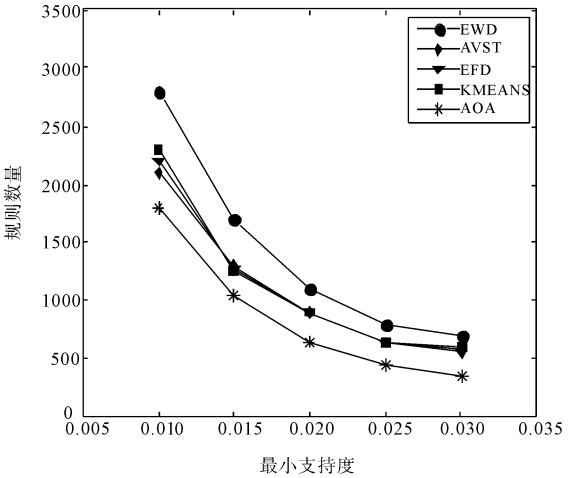

图 5 样本 3 的挖掘效率仿真结果

Fig. 5 Sample three simulation results of mining efficiency

实验表明,对于不同的情况,在本文所列举的四种离散化方法中,并没有某一种离散化方法离散后挖掘效率一直最高. 针对这一问题,我们提出并行调用的全局方法. 对应不同数据集,根据 AOA 先进行异常值剔除,再选择最优方法来进行数据准备,能得到最高挖掘效率.

5 结 论

在本文中,我们提出一种根据 AOA 数据准备方法的自动选择机. 该方法主要分为两步:(1)检测和消除异常值;(2)基于不同分布筛选离散化方法,多个离散化方法并行离散化后再根据三个评估指标,综合权值来判定选择最好的离散化方法. 试验使用真实和合成数据库进行评估,验证了我们的方法使用更合适的离散化方法确实可以减少挖掘得到关联规则数量. 对于未来的工作,我们的目标是:在我们的系统添加其他离散化方法(Khiops, Chimerge,法耶兹伊拉尼等);运用我们的方法与其他数据挖掘技术(决策树,SVM,神经网络)进行匹配试验;要使用其他数据库进行更多的实验.

参考文献:

[1] Casali A, Ernst C. Extracting correlated patterns on multicore architectures[M]// Availability, Reliability, and Security in Information Systems and HCI. Berlin: Springer Berlin Heidelberg, 2013.

[2] Beveridge J, Wiener R, 贝弗里奇 ,等. Win32 多线程程序设计[M]. 武汉:华中科技大学出版社, 2002.

[3] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27: 54.

[4] Casali A, Ernst C. Discovering correlated parameters in semiconductor manufacturing processes: a data mining approach[J]. IEEE Trans Semiconduct M, 2012, 25:118.

[5] Grubbs F E. Procedures for detecting outlying observations in samples[J]. Technometrics, 1969, 11: 1.

[6] 罗成汉, 刘小山. 曲线拟合法的 Matlab 实现[J]. 现代电子技术, 2003, 26: 16.

[7] 王学民. 偏度和峰度概念的认识误区[J]. 统计与决策, 2008, 2008: 145.

[8] 葛虹. 科学地选择零假设及显著性水平[J]. 统计科学与实践, 2005, 2005: 15.

[9] Thadewald T, Büning H. Jarque-Bera test and its competitors for testing normality: A power compar-

ison [J]. J　Appl Stat，2007，34：87.

[10] 王曰芬，章成志，张蓓蓓，等. 数据清洗研究综述[J]. 现代图书情报技术，2007，2：50.

[11] 侯燕，郭慧玲. 关联规则挖掘结合简化粒子群优化的哈希回溯追踪协议[J]. 重庆邮电大学学报：自然科学版，2016，28：239.

[12] 宋军，陈潇君. 转换时间数据流的加权 FPTree 挖掘算法[J]. 江苏大学学报：自然科学版，2017，38：330.

[13] 费贤举，李晓芳. 基于压缩感知理论的无线传感器网络数据融合算法[J]. 吉林大学学报：理学版，2016，54：575.

[14] 查如琴. 简谈几种“箱线图绘制”的描述[J]. 读与写：教育教学刊，2012，7：54.

[15] 王钟灵，吴霖生，姜妍妍. 如何选用格鲁布斯检验法的临界值[J]. 滁州学院学报，2002，4：80.

[16] Mitov I，Ivanova K，Markov K，*et al*. Comparison of discretization methods for preprocessing data for pyramidal growing network classification method [C]/. [s. l.]：[s. n.]，2009.

[17] Xu J H，Liu H. Web user clustering analysis based on KMeans algorithm[C]// International Conference on Information NETWORKING and Automation. [s. l.]：IEEE，2010.

[18] 李佳佳，何小海，吴晓红. 基于网格模板的最小欧式距离准则图像自动拼接技术[J]. 成都信息工程学院学报，2007，22：80.

引用本文格式：

中　文：刘云，袁浩恒. 数据挖掘中并行离散化数据准备优化[J]. 四川大学学报：自然科学版，2018，55：993.

英　文：Liu Y，Yuan H H. Parallel discretization of data preparation optimization in data mining[J]. J Sichuan Univ：Nat Sci Ed，2018，55：993.