doi: 10.7690/bgzdh.2026.05.006

基于并行计算的混合数据多约束挖掘算法仿真

叶　舟，李　晶
（浙江财经大学文华创新学院，杭州 310018）

摘要：为提高大量混合数据中目标数据的挖掘效率，提出基于并行计算的混合数据多约束挖掘算法。通过小波阈值法对混合数据展开去噪处理；提出基于 Cublas 库中矩阵乘法函数的距离并行算法，以获取混合数据间的距离；对去噪后的混合数据展开正负关联约束，并基于约束条件结合数据间距离对混合数据展开聚类，根据聚类结果完成同类数据挖掘。实验结果表明，该方法的数据处理效果好、数据挖掘性能高。

关键词：并行计算；混合数据挖掘；多约束；小波阈值
中图分类号：TP301.6　　**文献标志码：**A

Simulation of Multi-constraint Mining Algorithm for Hybrid Data Based on Parallel Computing

Ye Zhou, Li Jing
(*Wenhua School*, *Zhejiang University of Finance and Economics*, *Hangzhou* 310018, *China*)

Abstract: In order to improve the mining efficiency of target data in a large number of mixed data, a multi-constraint mining algorithm for mixed data based on parallel computing is proposed. The wavelet threshold method is used to denoise the mixed data. A distance parallel algorithm based on the matrix multiplication function in Cublas library is proposed to obtain the distance between the mixed data. Positive and negative association constraints are extended to the denoised mixed data, and the mixed data are clustered based on the constraints and the distance between the data, and the similar data mining is completed according to the clustering results. The experimental results show that the method has good data processing effect and high data mining performance.

Keywords: parallel computing; hybrid data mining; multi-constraint; wavelet thresholding

0　引言

数据挖掘技术能够从大量不同种类的数据中获取到隐藏的重要数据和知识，被应用的领域极其广泛且包含的方法多种多样[1]。信息技术的迅猛发展，使得网络数据数量和种类剧增，产生了大量混合数据，对数据的采集、处理以及挖掘都造成了极大的困难。目前的混合数据挖掘算法存在数据处理效果差、挖掘到的数据关联度低、挖掘效率低等问题。并行挖掘数据的方法能够通过并行结构来大大提高数据挖掘的速度、降低挖掘成本[2]，因此，对混合数据的并行挖掘方法展开研究。

黄文秀等[3]确定数据的密集分布区域并对其展开裁剪和改进，使数据分布均匀；利用传统 k 最邻近算法获取数据的权重。基于权重得到数据分类结果，并按照分类结果展开挖掘。该方法的数据处理效果差，影响了后续的数据挖掘结果。周燕等[4]利用具有分类和索引功能的效用链表来构建数据索引结构，以实现全部项集的搜索。将不符合条件的效用链表对应的内存展开回收再分配处理。利用提前结束策略提高数据挖掘效率。该方法挖掘的数据关联度较低，数据挖掘精度低。蒋华等[5]通过兴趣度约束和支持度自适应策略获取特征量强关联规则。采取局部离群点检测算法去除 K-means 聚类离群点。根据最大最小距离获得聚类中心和 K 值，以此完成异常与非异常数据分类，按照分类展开数据挖掘。该方法的数据挖掘速度慢，数据挖掘效率低。

为了解决上述方法中存在的问题，提出基于并行计算的混合数据多约束挖掘算法仿真。

1　混合数据的噪声抑制

由于采集到的混合数据中含有大量的噪声信号，会直接影响数据的分类和挖掘精度，因此需要利用小波阈值[6]对混合数据展开噪声的抑制。

设定 f 是混合数据中含有噪声信号的数据：

$$f=d+q。\tag{1}$$

式中：d 为数据中的有效信号；q 为噪声信号。对数据 f 展开小波阈值去噪[7]。

收稿日期：2024-12-05；**修回日期：**2025-01-06
基金项目：浙江省高等教育“十三五”第二批教学改革研究项目(jg20190301)
第一作者：叶　舟(1975—)，男，浙江人，硕士。

1) 对混合数据 f 中的所有信号展开小波变换[8]处理，得到不同尺度下的小波变换系数 c。

2) 确定小波阈值。设定 Y 为小波阈值：

$$Y=\varsigma\sqrt{\lg M}\ 。\tag{2}$$

式中：ς 为混合数据中噪声信号的标准差；M 为噪声信号的长度。在对混合数据的实际去噪过程中，噪声信号通常分布在最高分辨率层中；因此，噪声标准差通常可以通过 ξ=MAD/0.674 5 来计算，式中的 MAD 为最高频小波系数对应的绝对值中间值。

3) 对小波变换系数 c 展开非线性阈值处理。基于低频的小波变换系数更能反映信号的整体情况，选择保留全部噪声信号经小波变换后的最低频系数，将剩下的小波系数通过硬阈值和软阈值结合的方法展开处理。硬阈值 $\sigma_Y(c)$为：

$$\sigma_Y(c)=\begin{cases}c & |c|>Y\\ 0 & |c|\leqslant Y\end{cases}。\tag{3}$$

式中$|c|$为剩下的小波系数对应的绝对值。

软阈值 $\sigma_Y(c)'$的定义为：

$$\sigma_Y(c)'=\begin{cases}\mathrm{sign}(c)(|c|-Y) & |c|>Y\\ 0 & |c|\leqslant Y\end{cases}。\tag{4}$$

在软阈值处理中，同样将$|c|$与 Y 展开比较，并将绝对值大于阈值的 c 值取绝对值和阈值的差值，正负符号保持取原 c 值的符号，绝对值小于阈值的 c 值同样取 0。通过硬阈值处理能够最大限度地保留混合数据中的有效信号，而通过软阈值处理能够使混合数据信号的光滑程度较高。

2　基于并行计算的多约束挖掘算法

2.1　基于 Cublas 库的数据距离并行计算

Cublas Library 属于一种基于 GPU[9]的通用函数库，CublasSgemm 是 Cublas 库中的一种矩阵乘法函数，能利用 GPU 中的多个处理器核心来完成数据间距离的加速计算，提高数据挖掘的效率。

矩阵乘法函数[10]CublasSgemm 的调用参数列表为 CublasSgemm(trans$a$, trans$b$, e, w, l, *alpha, *$\boldsymbol{S}$, lds, *$\boldsymbol{R}$, ldr, *beta, *$\boldsymbol{V}$, ldv)，其中 trans$a$ 和 trans$b$ 指矩阵 $\boldsymbol{S}$ 与 $\boldsymbol{R}$ 的转置，当 trans$a$ 或 trans$b$ 的值为 Cublas_OP_N 时，对应的矩阵 $\boldsymbol{S}$ 或 $\boldsymbol{R}$ 不展开转置操作；当其中一个值为 Cublas_OP_T 时，对应的矩阵 $\boldsymbol{S}$ 或 $\boldsymbol{R}$ 展开转置操作。e 为矩阵 $\boldsymbol{S}$ 与 $\boldsymbol{V}$ 的行数，w 为矩阵 $\boldsymbol{R}$ 与 $\boldsymbol{V}$ 的列数，l 是 $\boldsymbol{S}$ 的列数和 $\boldsymbol{R}$ 的行数。Cublas 库中的全部矩阵都是基于优先形式展开保存，lds、ldr 和 ldv 分别为矩阵 $\boldsymbol{S}$、$\boldsymbol{R}$ 和 $\boldsymbol{V}$ 的行数，混合数据点矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{R}$ 的维度分别是 e*l 和 l*w，聚类中心矩阵 $\boldsymbol{V}$ 的维度为 e*w。*alpha 和*beta 为参数和地址指针，*$\boldsymbol{S}$、*$\boldsymbol{R}$ 与*$\boldsymbol{V}$ 表示对应的矩阵在 GPU 中的储存地址；因此，在使用矩阵乘法函数之前需要确定混合数据矩阵、聚类中心矩阵和对应的指针，且通知距离矩阵的地址。矩阵乘法函数的计算过程为：

$$\Re=\mathrm{alpha}^{*}OP(\boldsymbol{S})^{*}OP(\boldsymbol{R})+\mathrm{beta}^{*}\boldsymbol{V}\ 。\tag{5}$$

为了符合上述运算格式，将欧几里得距离公式展开平方计算，则基于 K−means 算法[11]的混合数据间距离可转换为：

$$|c-u|^2/\Re=\mathrm{alpha}(c*u)+\mathrm{beta}(c^2+u^2)/\Re\ 。\tag{6}$$

基于此，在建立 CUDA 核函数时，假设存在 w 个混合数据；且各个混合数据的维度都是 g，聚类中心的数量为 l；混合数据点矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{R}$ 分别为上式中的 w*g 和 u，其中 $\boldsymbol{S}$ 是一个 w*g 维的矩阵，$\boldsymbol{R}$ 是一个 l*g 维的矩阵，基于矩阵的优先保存规则，在实际计算时 $\boldsymbol{S}$ 的维度是 g*l，$\boldsymbol{R}$ 的维度是 g*l，为了获取所有混合数据点到每个聚类中心[12]的距离矩阵，要对 w*g 展开转置操作，此时需要将 transa 设定为 Cublas_OP_T，将 trans$b$ 设定为 Cublas_OP_N。

由于混合数据矩阵 $\boldsymbol{S}$ 的行数和列数分别是 w 和 g，$\boldsymbol{R}$ 的列数是 l；因此，将 CublasSgemm 函数的参数列表中的 e 更换为 w，w 换为 l，l 换为 g。在函数开始调用前需要确定聚类中心矩阵 $\boldsymbol{V}$，即 c^2+u^2 的值，c^2 值能够通过开启 w 个线程展开并行计算，各个线程负责获取对应的一个混合数据向量 g 个维度上的平方和；u^2 值通过开启 l 个线程展开计算，且各个线程负责对应的一个聚类中心 g 个维度上的平方和，对 2 个平方和展开加法计算即可获得聚类中心矩阵 $\boldsymbol{V}$，即 c^2+u^2 的值。则利用 CublasSgemm 函数对混合数据与聚类中心矩阵之间的距离展开并行计算的过程为：

$$\begin{aligned}&\mathrm{alpha}*OP(\boldsymbol{S})*OP(\boldsymbol{R})+\mathrm{beta}*\boldsymbol{V}=\\ &\mathrm{alpha}(c*u)+\mathrm{beta}(c^2+u^2)\begin{bmatrix}s_{11}&s_{12}&\cdots&s_{1g}\\ \vdots&\vdots&&\vdots\\ s_{w1}&s_{w2}&\cdots&s_{wg}\end{bmatrix}*\begin{bmatrix}r_{11}&r_{12}&\cdots&r_{1l}\\ \vdots&\vdots&&\vdots\\ r_{g1}&r_{g2}&\cdots&r_{gl}\end{bmatrix}+1*\begin{bmatrix}v_{11}&v_{12}&\cdots&v_{1l}\\ \vdots&\vdots&&\vdots\\ v_{w1}&v_{w2}&\cdots&v_{wl}\end{bmatrix}。\end{aligned}\tag{7}$$

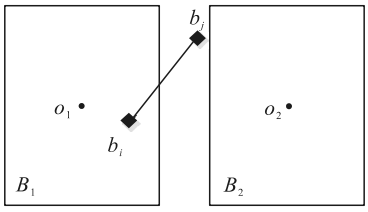
通过上述计算能够获得维度为 $w*l$ 的结果矩阵 $\boldsymbol{V}$，$\boldsymbol{V}$ 中的各行指混合数据分别和 l 个聚类中心的距离。

2.2　多约束下的混合数据挖掘

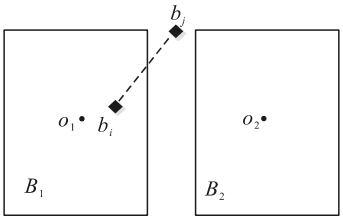
为提高混合数据挖掘的质量，基于并行计算的混合数据多约束挖掘算法仿真基于正关联和负关联 2 种约束关系(即成对约束)[13]展开混合数据中同类数据的挖掘。

正关联约束是指混合数据中 2 个数据为同类型数据；负关联约束是指混合数据中的 2 个数据不为同一类数据。现定义 Must-Link 为正关联约束，Cannot-Link 为负关联约束，则在同类数据的挖掘过程中，每个数据必须符合 Must-Link 和 Cannot-Link 的要求。

现设定混合数据中数据集的聚类数量为 L，B_1, B_2, …, B_L 为混合数据集划分后的区域，o_1, o_2, …, o_L 为各区域的中心点，b_i 和 b_j 表示混合数据中的任意 2 个数据，混合数据挖掘过程中的正负关联约束如图 1 所示。



(a) 正关联约束



(b) 负关联约束

图 1　正负关联约束情况

利用距离并行计算法将 b_i 归类到距离最近的类 B_1 中，对于数据 b_j，若 $i=j$，则$(b_i, b_j)\in$Must-Link，即 b_i 与 b_j 属于如图 1(a)所示的正关联约束，此时即使 b_j 与 o_2 间的距离更短，但基于正关联约束的要求，b_j 必须与 b_i 归为一类；若 $i\neq j$，则 $(b_i, b_j)\in$ Cannot-Link，即 b_i 与 b_j 属于如图 1(b)所示的负关联约束，此时 b_j 不为 B_1 类，依据距离并行计算将 b_j 分配到除 B_1 外的距离最近的类别中。

基于正关联约束和负关联约束的并行计算数据挖掘[14-15]过程为：

1) 在去噪后的混合数据中任意选取 L 个数据当做初始聚类中心点，即 $B_0=b_1$, …, $B_{l-1}=b_l$。

2) 任意选择若干个混合数据来产生正关联约束集合 ML 和负关联约束集合 CL。

3) 对于混合数据中的剩余待归类数据 b_j，如果 $(b_i, b_j)\in$Must-Link，且 b_i 属于 B_i 类，则判定 b_j 属于 B_i 类；如果$(b_i, b_j)\in$Cannot-Link，且 b_i 属于 B_i 类，则判定 b_j 属于非 B_i 类的与 b_j 距离最近的类 B_j；否则根据距离并行计算法将 b_j 直接归类到最近的类 B_j 中。

4) 每次数据归类迭代结束后，将各个类的中心点展开更新。

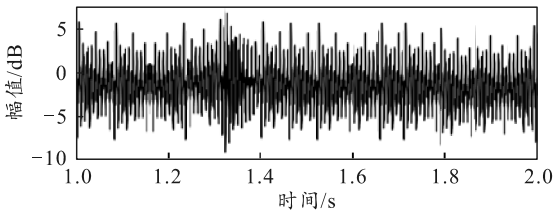
5) 不断重复 3)-4)步，直到所有数据归类完毕。

按照上述归类结果即可完成混和数据中同类数据的挖掘。

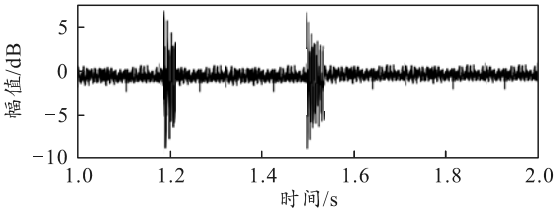
3　实验与分析

为了验证基于并行计算的混合数据多约束挖掘算法仿真的整体有效性，需要对其展开测试。

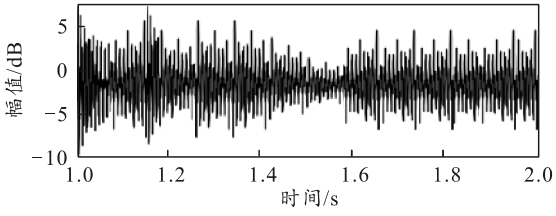
1) 数据处理效果。由于混合数据中存在的大量噪声数据会干扰数据的分类挖掘，因此在数据挖掘之前，需要对混合数据展开预处理，数据处理效果越好，同类数据挖掘的精度越高。现将一组混合数据截取其采集时间 1～2 s 的数据信号，并利用基于并行计算的混合数据多约束挖掘算法、文献[3]算法和文献[4]算法对数据信号展开处理，其处理效果如图 2 所示。



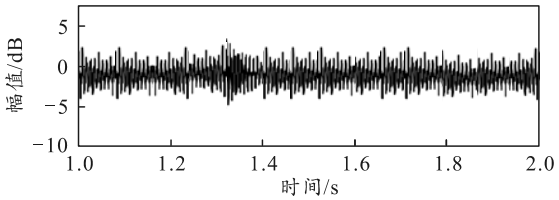
(a) 原始信号



(b) 所提方法



(c) 文献[3]方法



(d) 文献[4]方法

图 2　混合数据信号处理效果

根据图 2 可知：所提方法能够有效清除图 2(a)中的噪声信号，并保留有效信号。文献[3]算法对噪声信号的抑制效果较差。文献[4]算法虽然抑制了其中的噪声信号，但有效信号也被抑制；因此，所提方法的混合数据信号处理效果远高于文献[3]和文献[4]算法的数据信号处理效果。

2) 数据关联度。通常情况下，同类数据间具有较高的关联度，对于混合数据的挖掘目标是将同类数据归类并展开挖掘，为比较基于并行计算的混合数据多约束挖掘算法、文献[3]算法和文献[4]算法的同类数据挖掘精度，现利用上述 3 种方法对同一组混合数据中的 B_1 类数据展开挖掘，挖掘到的数据间关联度随挖掘数量的变化情况如表 1 所示。

表 1　数据间关联度

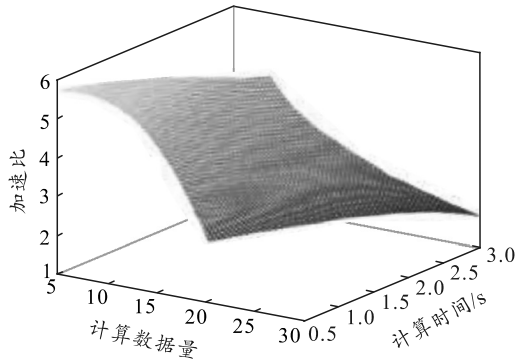
数据个数	所提方法	文献[3]算法	文献[4]算法
10	0.999	0.922	0.936
20	0.999	0.918	0.931
30	0.998	0.914	0.924
40	0.998	0.907	0.918
50	0.997	0.903	0.906
60	0.996	0.898	0.899
70	0.995	0.895	0.892
80	0.995	0.891	0.885
90	0.994	0.886	0.879
100	0.994	0.881	0.873

由表 1 可推出：随着挖掘数据数量的增加，3 种算法的数据间关联度均有所下降，但所提方法的关联度值最接近于 1 且下降幅度最小，而文献[3]算法和文献[4]算法的数据关联度较低且下降幅度较大，表明所提方法的稳定性强于文献[3]和文献[4]算法，且同类数据的挖掘精度较高。

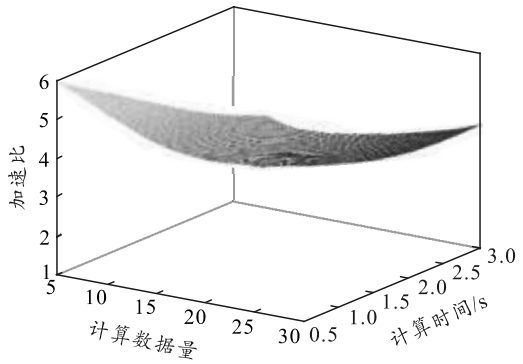
3) 加速比。在混合数据的挖掘实验中，加速比是衡量并行计算性能的重要指标，加速比越大，表明并行计算的性能越好，即数据挖掘的效率越高。加速比 D 的定义为：

$$D=Y_D/Y_A。\tag{8}$$

式中：Y_D 为对一个数据展开计算所需要的时间；Y_A 为 A 个性能一致的数据展开并行计算所需的时间。为了进一步判断基于并行计算的混合数据多约束挖掘算法的数据挖掘效率，现利用 5 台处理器分别以串行方式和并行方式对同一组混合数据展开距离计算，不同方式的距离计算加速比随计算数据量和时间的变化情况如图 3 所示。



(a) 串行方式



(b) 所提方法

图 3　加速比

分析图 3 可得：随着计算数据量和计算时间的增加，串行方式和所提并行方式的距离计算加速比都有所下降，但当计算同样数量的数据时，并行算法的加速比明显高于串行算法，且所提方法的加速比整体下降幅度较小，因此，所提方法的并行计算性能较好，即数据挖掘的效率更高。

4　结束语

笔者提出一种基于并行计算的混合数据多约束挖掘算法。验证结果表明：该方法能有效提高数据处理效果和挖掘效率，挖掘到的数据间关联度极高。

参考文献：

[1] 鲁娟利, 姜建国, 李鹏伟. 基于Web数据挖掘的水肥一体机功能参数实现[J]. 农机化研究, 2022, 44(5): 204−207.

[2] 李成严, 辛雪, 赵帅, 等. Sp-IEclat: 一种大数据并行关联规则挖掘算法[J]. 哈尔滨理工大学学报, 2021, 26(4): 109−118.

[3] 黄文秀, 唐超尘, 神显豪, 等. 改进的 k 最邻近算法在海量数据挖掘中的应用[J]. 济南大学学报(自然科学版), 2021, 35(1): 24−28.

（下转第 36 页）