

混合的大规模数据库中数值型数据聚类算法研究

何育朋

(广东财经大学 华商学院, 广东 广州 511300)

摘 要: 大规模数据库中的海量数据多具有混合属性, 即数值型数据与其他类型的数据混合于一体、数据量庞大、不易区分。传统算法往往忽视多种属性间的关联性, 算法复杂、聚类速度慢、效果差。对此提出一种基于划分聚类的混合大规模数据库中数值型数据聚类算法。首先为降低传统算法的高复杂度, 要从大规模的数据库中合理抽取多个小数据集, 小数据集中要包含数据库中全部的自然簇; 依据相似度定义构建小数据集的相似度矩阵, 并分别进行数值型数据及其他类型数据的相似度计算; 最后对抽样聚类的结果进行整合、均值更新和划分, 实现混合的大规模数据库中数值型数据的聚类。仿真实验表明, 提出的算法计算速度快、运算量相对较小、误差率低, 能够得到更理想的聚类效果, 适用于大规模的数据聚类处理。

关键词: 大规模; 数值型; 数据聚类

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2017)02-0119-04

Research on Numerical Data Clustering Algorithm in Hybrid Large Scale Database

HE Yu-peng

(Huashang College, Guangdong University of Finance and Economics, Guangzhou 511300, China)

Abstract: The mass data in the large scale database has mixed attributes, namely, the mixed data of symbolic data and numerical data, and the quantity of data is complex and difficult to distinguish. Traditional algorithms often ignore the correlation between the two attributes, the calculation is complex, the clustering speed is slow, the effect is poor. A numerical study of mixed database clustering in large-scale data clustering algorithm based on the traditional algorithm, firstly in order to reduce the high complexity, from reasonably extracting large-scale databases of multiple small data sets, all natural clusters contain database on small data sets; similarity matrix similarity is defined to construct small data set based on the then, the similarity data symbols and numerical data calculation; integration, the final result of the sample clustering updated mean and classification, clustering of numeric data mixed in large databases. Simulation results show that the proposed algorithm can get a better clustering result, and is suitable for large scale data clustering processing. The algorithm has a fast calculation speed, a relatively small amount of computation, and a low error rate.

Key words: large scale; numerical model; data clustering

1 引言

随着计算机科学及物联网技术的爆炸式发展^[1-2], 数据库的概念已经突破了人们传统意义上对其概念的理解, 并呈现出混合性的特征^[3-4]. 对于巨量的数据集需要新的处理方法和模式来提升洞察

力和决策力, 并对其流程进行优化, 以适应海量、多样化及高增长率的数据信息产业的发展^[5-6]. 目前, 物联网和云计算仍面临着数据量过大、计算高度复杂的难题, 因此对混合大规模数据库进行有效的数据挖掘, 对于大数据和云计算环境下物联网发展具有重要的现实意义.

数据聚类算法作为一门静态数据分析技术,是数据挖掘的一种重要方法^[7-8].数据聚类技术是在经验条件缺失的条件下,来揭示混合大规模数据库中,各种复杂数据之间的内在关系,将有这相同属性特征的数据聚合成簇,使簇内数据统一性或相似性最大化,而使不同簇之间的相似性尽可能最小化,实现数据的科学分组和归类^[9-10].数值型数据是数量表示和运算的最基础数据类型,对混合数据库中数值型数据的聚类处理更为重要^[11-12].大数据的环境下,数值型数据的聚类处理在图像分析、数据挖掘、模式识别及机器学习等诸多高科技领域有着十分广泛的用途.传统基于网格的算法^[13-14]在高维数据运算过程中耗时过长,运算量大;基于稀疏编码的分类算法^[15]分类误差率高、聚类效果差.

由此,本文提出了一种基于划分聚类思想的混合大规模数据库中数值型数据聚类算法研究.

2 一种混合大规模数据库中数值型数据聚类算法研究

2.1 混合大规模数据库中数值型数据集的抽样

大规模数据库中由于数据量特别巨大,而且混有多种数据类型.为降低数值型数据聚类算法的复杂程度和聚类算法的误差,本文从大规模的数据库中合理抽取多个包含数据库中全部自然簇的小规模数据集,进而对数据集进行基于划分的聚类处理.

随机抽取的小规模数据集,要具有一定的合理性,与混合大规模数据集一样含有的自然簇数量,设包含的自然簇总数为 n' ,采用小规模数据集抽样方法进行聚类,能够有效降低混合大规模数据库的复杂程度.为获取比较理想的抽样样本,本文设计了一种样本估算方法,设抽取的样本为 S ,指定数据抽取比例为 $\zeta, \zeta \in [0, 1]$;抽取样本自然簇的规模为 n ,从自然簇中抽取 $\zeta \times n$ 个数据的概率为 τ ,那么抽取的样本数据可以描述为

$$S = \zeta \times n + \frac{n}{n_i} \times \log \frac{1}{\tau} + \frac{n}{n_i} \sqrt{\log \left(\frac{1}{\tau} \right)^2 + 2\zeta \times n \times \log \frac{1}{\tau}} \quad (1)$$

假设自然簇规模达到一层以上,即自然簇大小、类别数小而且均等,那么抽取更小的规模就可以满足要求;指定数据抽取比例为 ζ 取值很小情况下,抽取数据的规模也可以更小.

为提高数据集抽样的便捷性和合理性,对混合

大规模数据集 A 进行等规模抽样,设小数据集的规模为 A_i ,抽样次数为 m ,整个抽样过程的控制条件为

$$\begin{aligned} A_i \cap A_j &= \emptyset \\ m_i &= m_j \\ i &= 1, 2, 3, \dots, m, i \neq j \end{aligned} \quad (2)$$

令式(2)中的各个抽样数据集无交集,抽样过程的示意图,如图1所示.

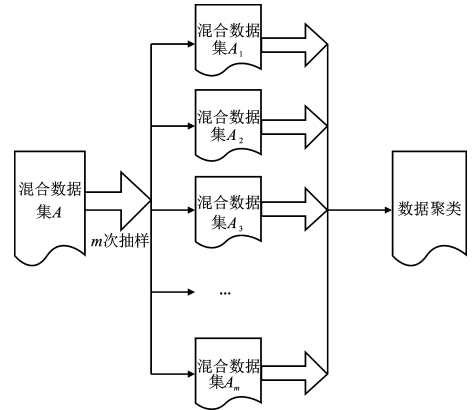


图1 数据抽样聚类过程的示意图

因为每一个抽样的数据集规模相对较小,因此凝聚层次的聚类过程会很快,进而可以提高数值型数据的聚类速度,并提高聚类的精度.

2.2 数值型数据相似度矩阵的构建及计算

完成对混合大规模数据库的抽样,进一步针对小规模的数据集中的数值型数据进行聚类处理.将小规模数据集中的数值型数据和其他类型的数据进行剥离,为简化算法,假定除了数值型数据之外,剩余仅有一种符号型数据.

对混合数据集中的数值型数据和符号型数据分别进行提取,形成独立数据集,再依据各列数据来构建相似度矩阵,如图2所示.

分别对数值型数据和符号型数据的相似度进行计算.可以利用高斯函数来构建数值型数据的矩阵,设 $T_{i,j}$ 为数值型数据矩阵, d 为数据点间的欧式距离, λ 为高斯函数特征参数,那么 $T_{i,j}$ 可以表示为

$$T_{i,j} = \exp\left(-\frac{d}{2\lambda^2}\right), i, j = 1, 2, \dots, n \quad (3)$$

而符号型数据属性可以定义为

$$T'(x_i, x_j) = \begin{cases} 0, & x_i \neq x_j \\ 1, & x_i = x_j \end{cases}, i, j = 1, 2, \dots, n \quad (4)$$

那么数值型矩阵 $T_{i,j}$ 和符号型数据 $T'_{i,j}$ 可以分别表示为

$$T_{i,j} = \begin{bmatrix} 1.000 & 0.331 & 0.475 & 0.358 \\ 0.331 & 1.000 & 0.331 & 0.135 \\ 0.475 & 0.331 & 1.000 & 0.216 \\ 0.358 & 0.135 & 0.216 & 1.000 \end{bmatrix} \quad (5)$$

$$T'_{i,j} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \quad (6)$$

依据数值型矩阵 $T_{i,j}$ 和符号型数据 $T'_{i,j}$, 构建的相似度矩阵图, 分别如图 2 和图 3 所示.

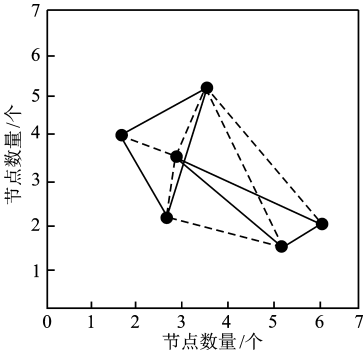


图 2 数值型数据矩阵图

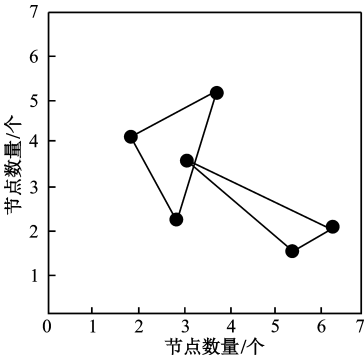


图 3 符号型数据矩阵图

本文依据不同的相似度矩阵来分别构建的不同的数据类型矩阵图, 将两种矩阵相结合:

$$T'' = T_{i,j} + \sum_{i=1}^n P_i \times T'_{i,j} \quad i, j = 1, 2, \dots, n \quad (7)$$

式中, P_i 为相似度最后的权重, C_i 为聚类前后节点所组图形的公共边数, 则 P_i 也可以表示为

$$P_i = \frac{C_i}{\sum_{i=1}^n C_i} \quad (8)$$

基于相似度定义, 构建小数据集的相似度矩阵,

完成数值型数据的相似度计算后, 进一步还要对小数据集的相似度结果进行整合、均值更新和划分, 完成混合的大规模数据库中数值型数据的聚类.

2.3 混合大规模数据库中数值型数据聚类的实现

对抽样小数据集的相似度结果进行整合, 计算数据集自然簇的均值 x_i , 进而对小数据集中的数值型数据进行聚类处理:

$$x_i = \frac{1}{n_{i,j=1}} \sum_{j=1}^n x_{ij} \quad (9)$$

以均值 x_i 代表该自然簇, 得到新的数据集 C , 采用 SINGLE 算法对新数据集 C 进行聚类, 对 k 个簇进行合并, 计算自然簇质心的初步位置.

由于抽样数据集属于局部信息, 自然簇质心的真实位置相对于初始位置, 会发生一定的偏离. 因此必须对自然簇质心的初始位置进行偏差修正, 以获取精确度的质心的位置. 按照与新数据集的 k 个初始质心距离远近的关系, 数据集分配至距离最近的簇中, 其中最小距离的分类原则为

$$\tau = \operatorname{argmax}(-|\Delta x - x_i|^2) \quad (10)$$

式中, τ 是确定的类别, Δx 为未知数据集样本属性, x_i 为数据属性均值, 数据属性的均值更新方法为

$$x'_i = \frac{x_i \times n_i + \Delta x}{1 + n_i} \quad (11)$$

随着剩余数据的逐个加入, 均值的位置也不断地被更新, 初始簇的规模逐渐扩大, 一点点地向自然簇的真实中心移动, 均值的更新过程如图 4 所示.

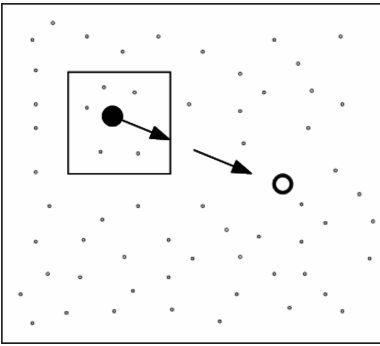


图 4 质心移动、均值更新简图

根据已经确定的自然簇质心对数据集进行划分, 该步骤实现了混合数据库中数值型数据的最终类别划分. 数据划分在没有特别要求的前提下, 根据数据距离簇质心的远近确定类别的归属, 通过对小数据集的划分, 实现对混合大规模数据库中的数值型数据的聚类处理.

本文提出了一种基于划分聚类的混合大规模数据库中数值型数据聚类算法研究, 从大规模的数据

库中合理抽取多个小数据集,并依据相似度定义构建小数据集的相似度矩阵,对小数据集中的数值型数据的相似度进行计算,最后对抽样聚类的结果进行整合、均值更新和划分,以完成混合的大规模数据库中数值型数据的聚类.

3 实验结果与分析

本文分别采用基于网格算法和本文提出的划分聚类算法进行仿真实验,验证两种不同算法下,混合大规模数据库中数值型数据的聚类效果及控制误差.

3.1 大规模数据库中数值型数据聚类效果对比

在聚类效果对比方面,如图 5 和图 6 所示,传统基于网格的算法,对数值型数据和符号型数据的区分效果模糊,聚类效果差.

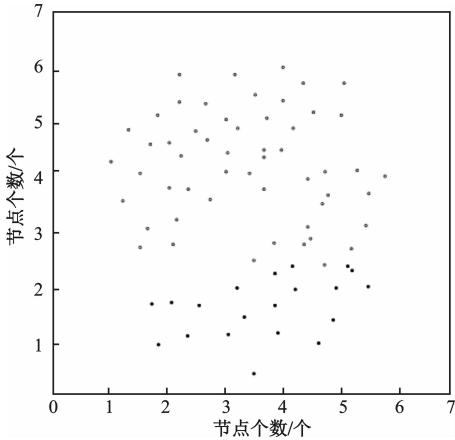


图 5 基于网格算法下数值型数据聚类效果

而本文提出的基于划分聚类的算法,能够清晰对数值型数据和符号型数据进行有效区分,聚类效果表现良好.

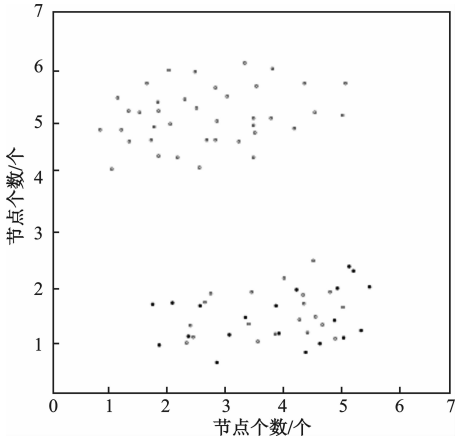


图 6 本文算法下数值型数据聚类效果

3.2 大规模数据库中数值型数据聚类误差控制曲线对比

在聚类误差率控制表现方法,本文提出的方法同样具有优势,图 7 所示为两种算法下聚类误差率控制曲线对比,本文算法具有更低的控制误差,而传统的基于网格的算法在初始阶段的误差较高.

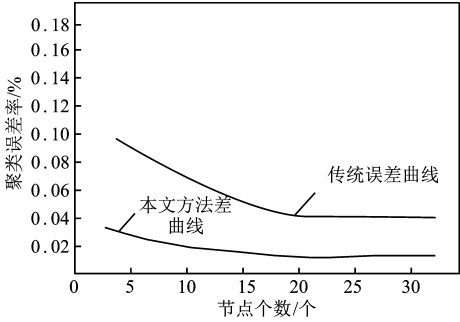


图 7 两种算法下数值型数据聚类误差曲线对比

仿真实验表明,本文提出的基于划分聚类思想的混合大规模数据库中的数值型数据聚类算法,能够有效地对数据库中的数值型数据进行聚类处理,在最终的聚类效果和误差控制方面,优于比传统基于网格的算法.

4 结束语

数据聚类能够在经验条件缺失等诸多不利条件下,揭示混合大规模数据库中,各种复杂数据之间的内在关系,在物联网和大数据时代已经到来的今天,该算法有着重要的意义.本文提出的基于划分聚类的算法能够有效地实现对数据库中的数值型数据的聚类处理,在聚类效果和误差控制方法有良好的表现.

参考文献:

[1] 邹静,罗媛. 探究计算机物联网技术在多领域的应用[J]. 科学中国人,2014,30(22):46-46.

[2] 李彬,王永亮,谢洁婷. 物联网技术在计算机监控系统中的应用[J]. 物联网技术,2014,29(7):43-44.

[3] 陈晓,赵晶玲. 大数据处理中混合型聚类算法的研究与实现[J]. 信息安全,2015,32(4):45-49.

[4] 熊杰. 基于混合模型的肿瘤组学数据聚类研究[D]. 长沙:湖南师范大学,2016.

[5] 高献卫,师智斌. 基于 Hadoop 二阶段并行模糊 c-Means 数据聚类算法[J]. 计算机测量与控制,2015,23(3):842-846.

[6] 张宇. 基于极值特征的雷达侦察数据 BIRCH 聚类方法[J]. 电子设计工程,2016,24(9):15-18.

而言,对应曲线的斜率较小,说明运算量的增大导致的运算时间的变化并不明显.

表 1 运算量增加带来的加速比变化

角度个数	单机运算 时间/s	快速算法 运算时间/s	加速比
2	83.53	85.75	0.98
3	169.32	93.74	1.82
4	274.63	102.41	2.69
5	387.87	111.15	3.49
6	532.29	136.68	3.91

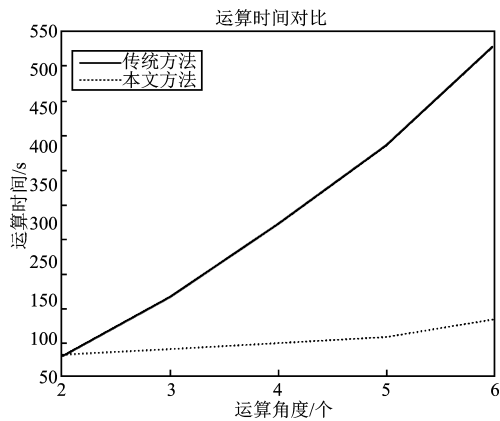


图 9 两种方法运算时间对比

5 结束语

针对传统的点云数据配准所需计算时间过长的
问题,本文在计算机集群的基础上提出了快速点云
匹配算法.实验表明,相比于单机的点云配准方法,
本文提出的快速匹配算法不仅能够获取正确的配准
效果,而且随着匹配角度数量的增加,能够获取到越
来越大的加速比,有效缩短点云配准所需的运算
时间.

参考文献:

[1] 王冬,周凯. 点云配准在大型曲面工件定位中的应用[J]. 计算机应用研究,2015,32(8):2347-2349.

[2] 刘同海,滕光辉,张盛南,等. 基于点云数据的猪体曲面三维重建与应用[J]. 农业机械学报,2014,45(6):291-295.

[3] 卫维恕,罗笑南,李峥. 基于水平集的扫描点云简化算法及在人体模型上的应用[J]. 计算机研究与发展,2008,45(9):1605-1611.

[4] 蔡来良,吴侃,张舒. 点云平面拟合在三维激光扫描仪变形监测中的应用[J]. 测绘科学,2010,35(5):231-232.

[5] 李波,王增辉,曹福毅. 基于 5 点数据手套的虚拟手精确控制方法[J]. 沈阳工程学院学报(自然科学版),2015,11(2):155-158.

[6] 王勇,唐靖,饶勤菲,等. 采用点云重心距离进行边界检测的点云数据配准[J]. 小型微型计算机系统,2015,36(9):2096-2101.

[7] 武敬民,韩燮,李定主,等. 基于法向夹角与 Hausdorff 距离的点云精简方法研究[J]. 微电子学与计算机,2014,31(4):52-55.

[8] 郑立华,麦春艳,廖崴,等. 基于 Kinect 的苹果树三维点云配准[J]. 农业机械学报,2016,47(5):1-7.

[9] 余文利,周明全,税午阳,等. 基于曲率的点云自动配准方法[J]. 系统仿真学报,2015,27(10):2374-2379.

[10] 李波,赵怀慈,花海洋,等. 基于集群和 GPGPU 技术的三维地形数据生成方法研究[J]. 微电子学与计算机,2014,31(10):1-4.

作者简介:

李 波 男,(1980-),博士.研究方向为虚拟现实与仿真技术、高性能计算机. E-mail:leebo@sia. cn.

李 智 男,(1961-),博士.研究方向为仿真技术.

王祥凤 女,(1981-),硕士.研究方向为仿真技术.

(上接第 122 页)

[7] 王虹,孙红. 基于混合聚类算法的客户细分策略研究[J]. 电子科技,2016,29(1):29-32.

[8] 冯进攻. 大规模数据聚类分析方法研究[D]. 哈尔滨:哈尔滨工程大学,2015.

[9] 陈晋音,何辉豪,杨东勇. 一种面向混合属性数据流的基于密度的聚类算法研究[J]. 小型微型计算机系统,2016,37(1):43-47.

[10] 孙浩军,闪光辉,高玉龙,等. 一种高维混合属性数据聚类算法[J]. 计算机工程与应用,2015,51(8):128-133.

[11] 余文利,余建军,方建文. 混合属性数据 k-prototypes 聚类算法[J]. 计算机系统应用,2015,24(6):168-172.

[13] 张树森. 聚类分析算法研究及其在数据密集型计算环境下的实现[D]. 淄博:山东理工大学,2015.

[14] 张宇. 基于极值特征的雷达侦察数据 BIRCH 聚类方法[J]. 电子设计工程,2016,24(9):15-18.

[15] 段宝彬,韩立新,谢进. 基于堆叠稀疏自编码的模糊 C-均值聚类算法[J]. 计算机工程与应用,2015,51(4):154-157.

作者简介:

何育朋 男,(1980-),硕士,讲师.研究方向为计算机应用技术. E-mail:530272504@qq. com.