

一种基于密度的分布式聚类改进算法

刘 恒¹, 苏 静¹, 唐咸艳², 卢佳佳³, 梁志胜⁴, 洪月华⁵

(1 广西医科大学 信息与管理学院, 广西 南宁 530021; 2 广西医科大学 公共卫生学院, 广西 南宁 530021;

3 广西医科大学 研究生院, 广西 南宁 530021; 4 广西医科大学 信息中心, 广西 南宁 530021;

5 广西经济管理干部学院 计算机系, 广西 南宁 530007)

摘 要: 基于密度分布式聚类算法(DBDC)在分布式聚类运算中运用广泛,但具有较高的算法时间复杂度. 本文提出了一种基于代表点交互的全局聚类方法,利用数据网格运算方法将数据对象映射到空间网格,改进基于密度的分布式聚类算法,优化原算法的空间搜索过程,从而改进生成本地聚类的效率. 同时在全局聚类层面可利用中心点作为代表点来降低聚类误差. 通过实验,表明该方法能够改进基于密度的分布式聚类算法,较传统分布式聚类算法提高了准确性度并降低了算法时间复杂度.

关键词: 聚类; 分布式; 数据挖掘; 代表点

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2018)05-0128-05

A Density-based Improvement for Distributed Clustering Algorithm

LIU Heng¹, SU Jing¹, TANG Xian-yan², LU Jia-jia³,
LIANG Zhi-sheng⁴, HONG Yue-hua⁵

(1 School of Information and Management, Guangxi Medical University, Nanning 530021, China;

2 School of Public Health, Guangxi Medical University, Nanning 530021, China;

3 Graduate School, Guangxi Medical University, Nanning 530021, China;

4 Information Center, Guangxi Medical University, Nanning 530021, China; 5 Department of Computer Science, Guangxi Cadres University of Economics and Management, Nanning 530007, China)

Abstract: The density based distributed clustering algorithm (DBDC) has been employed widely in process of distributed clustering. A major drawback of DBDC is the high time complexity. This work presented a method based on representative point for global clustering. A data grid mapping method which mapped data object to space grid, which improved DBDC and enhance the space search process and efficiency of implementation in local clustering level. Clustering error can be reduced by a global clustering method based on representative point. Using verification experiment with real data set from open source software, results showed this algorithm can improve original DBDC in two aspects: accuracy and reducing time complexity.

Key words: clustering; distributed clustering; data mining; representative point

1 引言

为了聚类分布式数据,传统的聚类分析方法将这些分布式数据集统一存储至中心站点,然后在中心

心站点中对本地数据进行整合,用集中聚类方法进行聚类分析获得最终的集簇化分析结果^[1]. 然而,当分布式系统增长,数据增量和数据安全问题使前述解法不再具有可行性^[2]. 例如 K 均值聚类算法作为

收稿日期: 2017-07-12; 修回日期: 2017-08-24

基金项目: 国家自然科学基金(81502890); 国家社会科学基金(15XTQ010); 广西高校科学技术研究项目(KY2015YB351、KY2015LX045); 广西研究生教育创新计划项目(JGY2015037)

一种传统的聚类方法,其算法简单速度快,但不能识别非球形的簇^[3,4].因此,人们提出分布式聚类方法用以分析在分布式环境中的大量分布式数据,同时解决传统聚类化方法无法在分布式环境下使用的问题,实现大范围数据并行运算,同时确保数据安全,让数据分析手段不局限于数据体量,安全性,私密性和其他约束.分布式聚类是目前聚类分析的必然趋势^[5].

通过改进基于密度的分布式聚类算法(Density-based distributed clustering, DBDC)和可伸缩的基于密度的分布聚类(Scalable density-based distributed clustering, SDBDC),基于密度空间聚类噪声算法(Density-based spatial clustering of applications with noise, DBSCAN)可实现在分布式数据集下的针对分布式数据集的聚类分析^[6].DBSCAN利用邻域内的相邻点数来度量所在空间密度,可以在聚类数未知的前提下,找出形状不规则的聚类,而集成了DBDC和SDBDC算法的优点,DBSCAN算法可寻找任意聚类的形状和聚类过程,找到噪声点数据并消减噪声点数据的影响,而其敏感度不受数据输入的顺序的影响.但两者结合后,在本地聚类和全局聚类在使用DBSCAN算法两次后,会导致整个聚类过程的时间复杂度过高.另外该方法必须多次选择后,选定全局聚类邻域半径.而邻域半径值的选定会影响全局聚类结果,如果过高,会导致合并聚类存在失败的风险,而如果该值过低,将不容易发现本应聚合在一起的类别.因此选用合适的全局聚类邻域半径值是制约该方法的最大问题.

针对上述问题,我们提出了一种改进的基于密度的分布式聚类方法.该方法利用数据网格映射的方法将数据对象首先在本地层面映射到空间网格,然后通过减少DBSCAN算法的检索空间来实现改进本地聚类的效率.在全局层面,采用代表点交叉方法,在本地模型中提取代表点的特性,利用代表点中的中位点来减少聚类错误.该方法不需要在全局聚类层面输入聚类参数.

2 相关工作

现有的聚类算法可以分为划分、层次、基于密度和基于网格几种^{[7][8]},正如前文所述,只要对邻近区域密度超过特定阈值,可归于一个类的聚类算法由于其高效和快速的特点,在数据挖掘实践中得到广泛运用^[9].基于密度的分布聚类算法DBDC如在分布式环境中结合DBSCAN,可以实现对于分布式数

据的聚类分析.图1描述了DBDC对分布式数据处理的完整过程,其将整个聚类过程分为四步:本地聚类,本地模型选取,全局模型选取以及标签化本地数据.

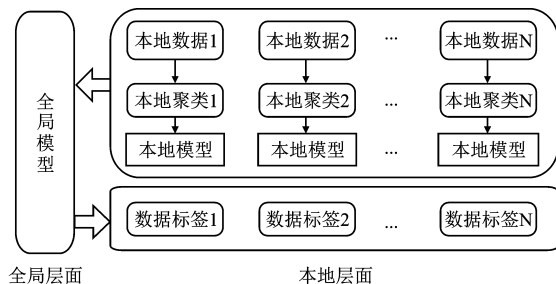


图1 DBDC对分布式数据处理过程示意图

如图1所示,DBDC利用DBSCAN算法将每个分布式本地站点的数据进行并行聚类,然后将由本地模型选取的本地代表点发送到中央站点.中央站点中的全局模型再利用DBSCAN算法来在分布式环境下实现对所有数据的实时聚类并逐一用标签标识每个本地数据.

DBDC和相应的改进算法继承了DBSCAN的优点,但也存在如下缺点:

(1)DBDC在本地聚类和全局聚类层面运用了两次DBSCAN算法,从而导致整个聚类过程的时间复杂度上升.

(2)没有合适的方法正确迅速的选择合适的全局聚类邻域半径,所以本地模型的选择有很大的随意性.

(3)本地噪声数据被忽略会影响整个聚类过程的精准度.

3 基于密度的分布式聚类改进算法

3.1 算法分析

DBSCAN聚类算法寻找数据对象的邻域,而找到每一个数据对象的邻域需要遍历整个数据集.然而许多距离较远的数据可以直接不予考虑.如果引入一种方法将数据对象首先映射到空间网格,再通过基于密度的DBSCAN聚类算法在网格地图中获得核心对象,这样就只需要遍历同一和相邻网格中的数据点,而避免了遍历所有数据集的问题,如此一来可以减少整个算法的时间复杂度和提升算法效率^[10].

3.2 算法描述

本文引入基于最小数据集数据映射方法,令特征点 z 和两个空间方向矢量 α_1 和 α_2 .首先引入由

特征点 z 和 $\partial 1$ 和 $\partial 2$ 定义的空间平面 P_z ,然后将数据对象利用矢量投影的方法映射到平面 P_z . 最终网格化空间平面后,可以得到空间网格上对应的数据点集. 利用基于密度的 DBSCAN 聚类算法,最终该方法可以极大程度降低整个算法的时间复杂度并提升效率.

整个方法的关键是网格的划分,该方法中网格边界点界定算法伪码如下,该算法复杂度为 $O(n \log n)$:

算法 1: 边界探测算法

```
input : 聚类 C, 平衡矢量集  $b_{pi}$ , 其中参数 V 为空间网格邻域半径
output: 聚类 C 的边界点  $B_C$ 
1  $B_C \leftarrow C$ ;
2 for every point  $p \in B_C$  do
3 for every point  $q \in N^C(p)$  do
4 if  $(q - p)b \geq \cos(v)$  then
5 delete  $p$  from  $B_C$ ;
6 Break;
7 return  $B_C$ ;
```

改进后 DBSCAN 的算法分两步,第一步先按前述算法进行网格划分,将数据点映射到网格中,第二步将数据集中的点均标为核心点、边界点或噪声点,通过边界点和核心点交互关系交叉检验相邻网格关系,同时将任意两个距离小于领域半径 $Eps(\epsilon)$ 的核心点收录到同一个聚类. 任何与核心点距离相近的边界点也放到与之相同的聚类中. 设定义核心点的阈值 $MinPts$ 为 M . 领域中的数据点在选取过程中不断对代表点所在 Eps 区域的点进行区域查询,如果当前代表点也是核心点,则也加入到子集中,通过这种方式迭代找到一个完整的聚类. 其改进的具体算法伪码如下:

算法 2:改进算法

```
Step 1 初始化
1 给定参数  $\epsilon$  和  $M$ 
2 生成  $N_\epsilon(i), i=1,2,\dots,N$ .
3 令  $k=1; m_i=0, i=1,2,\dots,N$ .  $//m_i$  为标记数组,如为核心点,则属于第  $j$  个聚类,如为噪音点,则为 -1
4 令  $I=\{1,2,\dots,N\}$ .
Step2 生成聚类标记数组
While( $I \neq \emptyset$ )
{从  $I$  中任取一个元素  $i$ ,并令  $I_i = I \setminus \{i\}$ .
If( $m_i=0$ )//即  $i$  号节点还没被处理过
{
初始化  $T_i = N_\epsilon(i)$ 
If  $|T_i| < M$ ,则令  $m_i = -1$ .  $//i$  为噪音点
```

```
Else if  $|T_i| \geq M //i$  为核心点
 $m_i = k$ .  $//i$  归入  $k$  个聚类
while( $T_i \neq \emptyset$ )
{从  $T_i$  中任取一个元素  $j$ ,并令  $T_i = T_i \setminus \{j\}$ 
If  $m_j=0$  或  $-1$ , 则令  $m_j = k$ .
If  $N_\epsilon(j) \geq M //j$  为核心点,  $T_i = T_i \cup N_\epsilon(j)$ 
}
 $k = k + 1 //$ 第  $k$  个聚类完成,开始下一个聚类
}
```

图 2 是加入改进前(a)后(b)的 DBSCAN 算法对实验样本数据的聚类结果,可见通过合理设置领域半径 $Eps(\epsilon)$ 和阈值 $MinPts(M)$,针对相同数据点集,改进后的 DBSCAN 算法较传统的 DBSCAN 算法能区别实际中两个聚类,而不会单纯通过密度将部分核心点误判为噪声点.

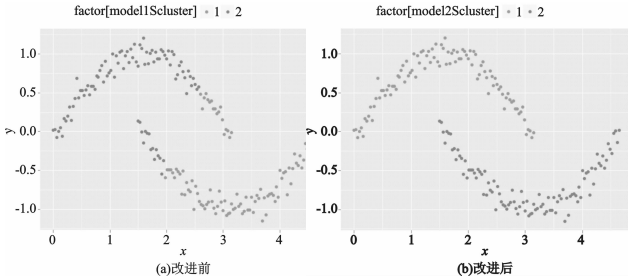


图 2 改进前后的 DBSCAN 算法示意图

4 实验验证和效率分析

4.1 准确度验证

为了评估聚类划分的准确度和对分布式数据进行聚类分析,采用集中聚集结果作为参考对象. 设存在数据对象 x , 在分布式聚类分析中归属于聚类 C_d ,而在集中聚类分析中又属于聚类 C_c , 对其进行准确度评估过程如下:

$$dar = \frac{\sum_{i=1}^n p(x_i)}{n} \tag{1}$$

聚类准确度是评估聚类和分布式聚类算法优劣的重要指标,其中分布式聚类结果较其他结果更加直观. 本文我们采用 UCI 机器学习库的 Wall-Following Robot 导航数据[11],该数据由 SCITOS G5 机器导航系统,该数据集通过 24 个超声波传感器沿顺时针方向在特定房间进行收集,该数据有实例 5 456 个,包含属性 24 个,其数据集特征通过多变量顺序记录.

DBSCAN 算法利用代表点生成本地聚类模型. 针对不同量级的本地数据集,利用表 1 所示不同的

本地聚类模型对各类型点的需求,对各个数据片段进行多重迭代代入改进算法求聚类.

表 1 本地聚类模型结果

数据量级	1 000	2 000	3 000	4 000	5 000
平均核心点数	35	49	60	64	70
平均噪声点数	34	41	45	52	63
平均聚类数	3.8	4	4	4	4

在分布聚类阶段,改进算法可针对本地聚类代表点,直接应用而无需调整聚类参数.而当进行全局聚类时,可利用代表点的中位点减少聚类产生时的误差.最终我们选择 Eps 为 0.7,MinPts 的值为 10,可提高改进型算法的可比较性,减少噪声点的分布.在比较不同站点和数据量级下,原有算法和改进算法的准确度用下述方法进行:

(1)比较两种算法在不同站点的准确度:在不同站点,我们将所有数据分成了一定数量的相关子集,分别用两种算法经过数轮迭代可获得聚类结果.

(2)比较两种不同算法在不同数据量级下的准确度:通过选择三个本地站点下三种数据量级:600,1 200,1 800. 其结果如表 2 所示,可见改进算法较远 DBDC 算法在每个数据量级下聚类准确度均有显著提高.

表 2 聚类准确度统计表

数据量集	聚类准确度	
	改进算法%	DBDC%
600	95.8	94.1
1 200	95.6	93.9
1 800	95.1	93.2

4.2 效率分析

在本地聚类过程中 DBDC 算法利用 DBSCAN 算法进行本地聚类,该算法的时间复杂度为 $O(n^2)$,加入空间索引 $R^* - Tree$ 的支持后,算法时间复杂度变为 $O(n\log(n))$,但需要注意的是该时间复杂度并未考虑索引的建立损耗时间.本文所提改进方法首先利用 DBSCAN 算法将数据完全映射到网格,此时时间复杂度为 $O(n)$,基于网格的 DBSCAN 算法将对搜索空间进行降维.此时时间复杂度为:

$$t_{gc} = \sum_{i=1}^n iY_i$$

(2)

式中, i 为数据集中的数据对象; Y_i 为依附在网格上数据对象 i 的个数.则改进后的算法针对本地聚类阶段时间复杂度为:

$$t_c = n + \sum_{i=1}^n iY_i$$

(3)

在全局聚类阶段,本文所提改进算法和 DBDC 算法均利用到了本地站点典型数据.其数据集为

$$n_g = \sum_{j=1}^m |LocalModel_j|$$

(4)

式中, m 为本地站点数量; $|LocalModel_j| = \sum_{i=1}^t |Scor_G; t|$ 为站点 j 的本地聚类数量.在全局聚类阶段,继续采用 DBSCAN 算法来获取全局聚类结果,其时间复杂度为 $O(n_g^2)$.由上述式子可见本文所述方法利用融合不同本地站点的本地聚类方法实现了全局聚类.在全局聚类阶段,该算法的时间复杂度为:

$$n_g = \sum_{i=1}^{m-1} \sum_{k=1}^m |LocalModel_i| |LocalModel_k|$$

(5)

聚类的生成利用了宽度优先搜索算法,对代表点的考察穿插在宽度有限搜索的过程中,整个算法的时间复杂度为 $O(n + ik2)$,和 i 和 k 的值有关, i 和 k 值越小时间复杂度越低.由此可见,基于网格的 DBSCAN 算法较采用 DBDC 算法的时间复杂度得到显著降低.

5 结束语

本文描述了一种基于密度的改进型聚类分布式算法,通过引入数据网格映射方法,将数据直接映射到空间网格,减少了对于本地聚类算法搜索空间的时间消耗,从而弥补了 DBDC 算法时间复杂度过高的问题.通过将网格内核心点同边界点交互行为模式的提取,提高了对于聚类选取的准确度.从而减少了全局聚类算法的输入参数.最终,通过实验证明了改进后的算法在聚类过程上可以有效的减少进程冗余,并实现快速准确的聚类划分.

参考文献:

[1] Li Zeng, Ling Li, Lian Duan. Distributed data mining research: a survey [J]. Information Technology and Management, 2012, 13(4): 403-409.

[2] Ruoming Jin, Gagan AgraWI. Shared memory parallelization of data mining algorithms techniques [J]. Programming Interface and Performance. IEEE Transaction on knowledge and Data Mining, 2005, 17(1): 72- 73.

[3] 吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,

2008,19(1):48-61.

[4] 洪月华. 蜂群 K-means 聚类算法改进研究[J]. 科技通报,2016,32(4):170-173.

[5] Yongli Li, Peng Sun, Minghui Sun, Rui Mao, et. al, An improved distributed clustering algorithm based on density [C], the proceeding of 8th International Conference on Intelligent Networks and Intelligent Systems, 2015, pp133-136.

[6] 张枫,邱保志. 基于网格的高效 DBSCAN 算法[J], 计算机工程与应用,2007,43(17):167-169.

[7] 刘淑芬,孟冬雪,王晓燕. 基于网格单元的 DBSCAN 算法[J], 吉林大学学报(工学版), 2014, 44 (4): 1135-1139.

[8] 邓左祥,李春贵. 一种有效的多关系聚类算法[J]. 微电子学与计算机,2016,33(4):33-41.

[9] 赵慧,刘希玉,崔海青. 网格聚类算法[J]. 计算机技术与发展,2010,20(9):83-85.

[10] G. Luo, X. Luo, TF Gooch, et al, A Parallel DBSCAN Algorithm Based on Spark, IEEE International Conferences on Big Data & Cloud Computing, 2016; 548-553.

[11] UCL, Wall-following robot navigation data data set [EB/OL] ,2017, 引用日期:2017/12/12 <http://archive.ics.uci.edu/ml/datasets/Wall-Following+Robot+Navigation+Data>.

作者简介:

刘 恒 女,(1979-),硕士,副教授. 研究方向为数据挖掘、数据可视化和信息安全.

苏 静 女,(1983-),硕士,讲师. 研究方向为数据挖掘和人工智能.

唐咸艳 男,(1982-),博士,副教授. 研究方向为数据挖掘、时空流行病与时空统计学.

卢佳佳 男,(1979-),男,硕士,工程师. 研究方向为数据挖掘和信息安全.

梁志胜 男,(1979-),男,硕士,高级工程师. 研究方向为数据挖掘和信息安全.

洪月华(通讯作者) 女,(1973-),硕士,教授. 研究方向为大数据挖掘和人工智能.

E-mail: huibian2005@163. com.

(上接第 127 页)

参考文献:

[1] 曹军梅. 一种用于 MP3 文件版权保护的鲁棒水印算法[J]. 现代电子技术,2016,39(4):6-8.

[2] 刘丽,赵学民,彭代渊等. 适用于广播监视的视频水印协议[J]. 计算机学报,2014,37(11):2389-2394.

[3] 梁强,邱志宏,张爱科. 基于水印技术的隐蔽通信算法设计[J]. 计算机工程与科学,2010,32(8):32-35.

[4] 任小康,范丽,白勇峰,等. 一种基于 DWT-SVD 的彩色图像水印嵌入方法[J]. 微电子学与计算机,2011, 28(1):113-115.

[5] 余香敏,焦占亚. 一种基于 DWT 与 SVD 的数字图像水印算法[J]. 微电子学与计算机,2008, 25 (10): 199-202.

[6] 李林静,郑月斋. 基于 DWT-SVD 和 Arnold 变换图像

盲水印算法[J]. 计算机与数字工程,2015,43(11): 2035-2039.

[7] YANG X S. Nature-inspired metaheuristic algorithm [M]. Frome:Luniver Press,2008:81-96.

作者简介:

王世辉 男,(1992-),硕士研究生. 研究方向为图像处理技术. E-mail:shw2011zc@163. com.

王仪明 男,(1965-),博士,教授,博士生导师. 研究方向为印刷机械检测与故障诊断技术.

武淑琴 女,(1973-),硕士,副教授. 研究方向为印刷机械检测与故障诊断技术.

田 贝 女,(1992-),硕士研究生. 研究方向为 VOCs 处理技术.