刘晓悦,郭强.海量用电数据并行聚类分析[J].辽宁工程技术大学学报(自然科学版),2016,35(1):76-80. doi:10.11956/j.issn.1008-0562.2016.01.015
LIU Xiaoyue,GUO Qiang.Cloud computing based cluster analysis on data of power utilization[J].Journal of Liaoning Technical University(Natural Science),2016,35(1):76-80. doi:10.11956/j.issn.1008-0562.2016.01.015

海量用电数据并行聚类分析

刘晓悦，郭　强
（华北理工大学 电气工程学院，河北 唐山 063009）

摘　要：针对用电数据量大、用电数据挖掘效率低等问题，采用理论分析和实验的方法，进行用电数据并行分析构架的研究，研究了 Canopy 和 K-means 两种典型的聚类算法，提出一种新的聚类思路，使用 Canopy 先对用电数据进行粗略处理，得到聚类个数和聚类中心，再用 K-means 精确聚类，既利用了 K-means 算法简单、收敛速度快的优势，又使其不容易陷入局部最优.为达到处理海量数据的目的，把提出的算法部署到 MapReduce 框架上进行实验.研究结果表明：提出的算法在海量用电数据的处理方面高效可行，并且具有良好的加速比.

关键词：K-means 算法；Canopy 算法；云计算；MapReduce 框架；聚类
中图分类号：TM 734　　**文献标志码**：A　　**文章编号**：1008-0562(2016)01-0076-05

Cloud computing based cluster analysis on data of power utilization

LIU Xiaoyue, GUO Qiang
(College of Electrical Engineering, Huabei University of Science and Technology, Tangshan 063009, China)

Abstract: Aiming at the issues of huge amount of electricity data and low clustering efficiency in data mining, this paper adopted the method of theoretical analysis and experiment, analyzed the electricity data parallel study of the architecture and studied the Canopy and K - means two typical clustering algorithms. This study proposed a new clustering approach. The approach use the Canopy to rough handling of electricity data and get the cluster number and cluster center, then use K - means clustering precision. The approach both use the K - means the advantage of simple algorithm and fast convergence speed, and make it not easy to fall into local optimum. In order to reach the goal of dealing with huge amounts of data, the proposed algorithm was set on the MapReduce frame. The results show that the proposed algorithm is efficient and feasible in huge amounts of electricity data processing, and has a good speedup ratio.

Key words: K-means algorithm; canopy algorithm; cloud computing; MapReduce frame; cluster

0　引言

随着经济的快速发展，人民生活水平日益提高，用电需求也急剧增加，电力供应日益紧张.电能具有不易存储的特点，这就决定了电力生产必须“即产即销”，整个生产和消费同时完成，电力生产必须根据用电需求的变化作出相应的调整，所以用电数据分析一直都是一个至关重要的研究课题.在中国，供电企业用电信息采集覆盖的范围从仅覆盖重要的专线专变用户，逐渐扩大到包括各类专线专变用户、一般工商业户、低压居民等多种电力现场，接入各类采集终端及表计的规模也随之增加，每日要采集及处理的用电数据量呈指数级增长[1].这些海量数据中隐藏着用户的用电习惯等有用信息，寻找高效、准确的数据挖掘算法，成为用电领域亟待解决的问题.

聚类是将物理对象分为多个类或簇的过程，同一类中的对象尽可能相似，而不同类中的对象尽可能相异[2]，聚类分析可以高效的获得数据在全局范围的分布特征，聚类分析方法在用电数据分析中占主导地位.K-means 是一种经典的基于距离的聚类算法，具有收敛速度快，算法简单等特点，但需提前选取分类个数，如果分类个数选取不当，很容易陷入局部最优.Canopy 聚类算法虽然聚类精度不高，但不需提前

收稿日期：2015-01-16
作者简介：刘晓悦（1965-），女，河北 唐山人，博士，教授，主要从事大数据处理和数据挖掘方面的研究.
通讯作者：郭强（1989-），男，河北 邯郸人，硕士研究生，主要从事大数据处理和数据挖掘方面的研究.　本文编校：朱艳华

设定聚类个数，而且简单快速.结合 Canopy 和 k-means 两种聚类算法的特点，本文提出一种聚类思路：首先通过 Canopy 算法进行聚类，以确定聚类个数以及初始聚类中心，接着通过 K-means 算法进行迭代运算，收敛出最后的聚类结果.

以 MapReduce 分布式平台为代表的云计算，能同时使用多个处理器进行并行计算和分布式处理[3]，面对海量用电数据，把云技术应用到用电数据分析中，完成海量用电数据分析任务.Hadoop 是一种开源的分布式系统平台，具有扩展能力强、成本低、效率高以及可靠性好等特点，用户能轻松地构建一个高效的分布系统[4].本文结合 Hadoop 平台，实现了基于 Canopy 的 K-means 并行聚类算法，基于居民用电数据，在 Hadoop 集群上进行测试，验证了在用电数据分析方面的高效性和可行性.

1　基于云计算的用电数据分析构架

本文采用主/从构架实现用电数据分析，并基于并行挖掘算法实现海量用电数据的挖掘分析.图 1 为基于云计算的用电数据分析构架，主要分为两个模块：用电数据采集模块、云处理模块.用电数据采集模块完成对用户数据的采集，通过智能电表、传感器等设备完成不同用电设备、不同用电时间的数据采集，这些数据包括电能量、交采数据、工况数据、电能质量等用电数据.云处理模块主要在各类用电数据采集完以后完成数据的处理工作，云数据挖掘模型管理服务器会选择 1 种或多种适用此类数据的并行挖掘算法，将任务分解后分配给各个节点，完成用电数据的并行处理任务.

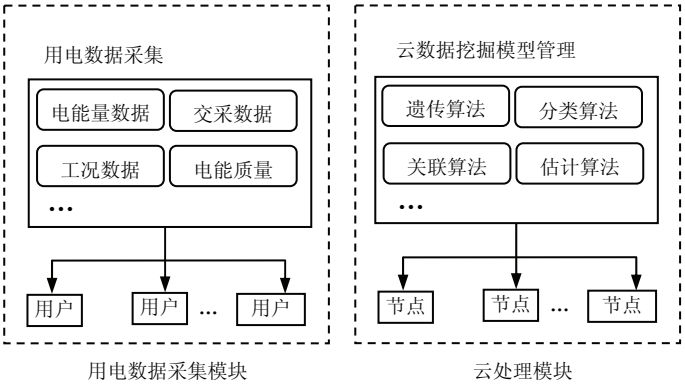

图 1　基于云计算的用电数据分析构架

Fig.1　based on analysis of power of cloud computing data structure

2　基于云计算的并行聚类算法

2.1　传统的K-means聚类算法和Canopy聚类算法

MacQueen提出的K-means算法是一种基于距离的聚类算法[5-6]，根据样本的距离来作为聚类的评价指标.该算法的核心思想是将n个数据划分为k个聚类，使得数据到聚类中心的平方和最小.传统的K-means聚类算法的流程：设定划分簇别为K，随机选取K个数据作为初始的聚类中心，计算除聚类中心以外的所有数据的欧式距离，把数据分配到相似度最高的聚类，然后以所有归到各个类中样本的平均值做为新的聚类中心，重新计算距离，更新聚类，直到平方误差函数小于给定的阀值.

设集合 $A=\{x_1,x_2,\cdots,x_m\}$，$x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$，则样本的欧式距离$d_{ij}$为

$$d_{ij}=\sqrt{(x_{i1}-x_{i2})^2+(x_{i2}-x_{i2})^2+\ldots+(x_{in}-x_{in})^2}\ .$$

设 $\bar{x}$ 为聚类 c_k 所有数据的均值，c为该聚类的中心，其平方误差函数为 $J_c=\sum_{i=1}^{k}\sum_{x\in c_k}\left|c-\bar{x}\right|^2$.

与传统的 K-means 聚类算法不同，Canopy 聚类算法最大的特点是不需要提前确定 K 值（即聚类个数），且聚类简单快速，因此有很大的实际应用价值，传统的 Canopy 聚类算法流程是：将数据集向量化得到一个 list 后放入内存，选择两个距离阈值：T_1 和 T_2，其中 $T_1 > T_2$，T_1 和 T_2 的值可以用交叉校验来确定，从 list 中任取一点 P，计算点 P 与所有 Canopy 之间的距离（如果当前不存在 Canopy，则把点 P 作为一个 Canopy），如果点 P 小于 T_1，则

将点 P 加入到这个 Canopy，如果小于 T_2，认为点 P 此时与这个聚类已经够近了，因此它不可以再做其它类的中心了，则需要把点 P 从 list 中删除，直到 list 为空结束[7].

2.2 基于Canopy的K-means并行聚类算法

本文使用的云计算平台 Hadoop 主要由 HDFS（分布式文件系统）和 MapReduce 计算模型组成[8]，HDFS 采用主/从构架，HDFS 集群是由一个 Namenode（管理节点）和若干个 Datanode（数据节点）组成，每个节点均是一台普通的计算机.MapReduce 是一种并行的编程模式，包括 Map 阶段和 Reduce 阶段，在 Map 阶段，各节点服务器以已经分割好的若干个数据块 split 作为输入，对每块中每条记录执行 map 函数，生成中间结果<key,value>键值对，在 Reduce 阶段，把 Map 生成的键值对作为输入，执行 Reduce 函数，最后生成新的键值对，最终把结果写入 HDFS 中[9-10].

并行的 Canopy 和 K-means 算法都是非常容易实现的.其基于 Canopy 的 K-means 聚类算法的并行实现也非常简单，本文使用 2 个 Map 函数、2 个 Reduce 函数和 1 个 Combine 函数.其具体的实现过程如下：

（1）初始化数据，以行形式存储用户用电信息，转化成<key,value>键值对形式供 Map 函数处理，key 是当前样本相对于输入数据文件起始点的偏移量，发送到 m 个待执行 Map 函数中；

（2）在 Canopy Map 阶段，逐行扫描数据，将输入的数据转化为向量，计算数据点与所有 Canopy 的距离（如果当前不存在 Canopy，则把该点作为一个 Canopy），并将距离与参数 T_1 与 T_2 进行比较，根据传统的 Canopy 聚类算法流程，生成 Canopy；

（3）在完成完 Canopy 函数后需要执行 Canopy Reduce 函数对 Map 阶段的输出结果进行汇总.将 Map 函数输出结果中 key 值相同的 value 汇总起来，传递给 Reduce 函数，然后 Reduce 函数对局部 Canopy 进行处理，处理过程与 Canopy 相同，最后得到各个聚类及其中心点，并将其输出到 HDFS 中.此时 Canopy 算法已经结束，可以得到聚类个数和聚类中心点.

（4）根据 Canopy 聚类的结果，设定初始聚类中心和聚类个数，执行 K-means Map 函数.

（5）K-means Map 函数把文件的每一行作为样本，以<key，value>键值对的形式表示，计算每个样本到各个聚类中心的距离，并选择距离最小的聚类中心，把该数据分配到此聚类中心，并把该数据样本标记为此聚类类别，形成<key,value>键值对的输出形式.

（6）执行完 K-means Map 函数后需执行 K-means Combine 函数对 Map 函数输出的中间结果进行本地规划处理，以减轻数据在各节点传输过程中的消耗.Combine 首先将各维度坐标值相加，得到局部聚类结果的累加和，并且得到总的样本个数.输出的<key,value>对中 key 为聚类类别，value 为样本总数和各维度坐标的累加值；

（7）最后 K-means Reduce 函数通过 Combine 函数得到的结果解析出样本个数和各维度坐标的累加值，然后对各维度坐标值分别对应相加，最后除以样本的个数，得到新的聚类中心坐标，并将其作为下一轮迭代使用.根据 Reduce 最终的结果，更新分布式文件系统 HDFS，并重新执行迭代，直到收敛为止.

其具体流程见图 2.

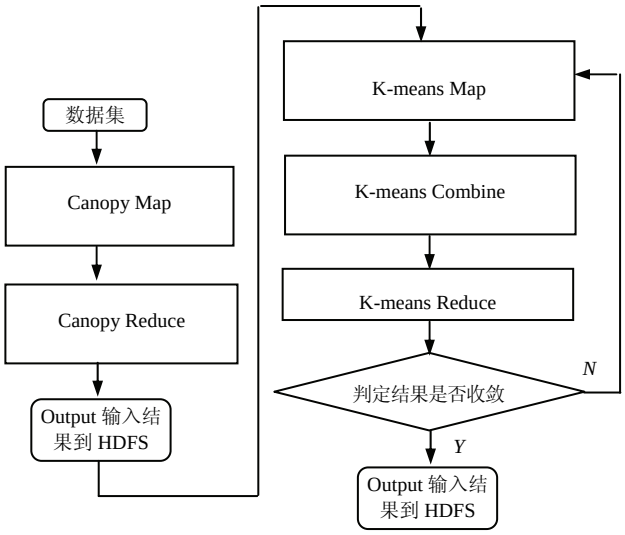

图 2 算法流程

Fig.2 algorithm flow chart

3 实验

3.1 实验环境

实验选用由6台PC搭建Hadoop云平台，4台PC为双核2.4 GHz，2 GB内存，2台为双核2.1 GHz，1 GB内存.Hadoop版本为1.2.1，每台计算机使用千兆网卡，通过交换机连接.

本文收集了唐山市某电网的 153 户居民用户的用电数据作为样本数据，采样频率为 10 min/次，采样时间为 24 h，每个用户每天采样 144 点数据.

3.2　单机对比实验

为验证本文所实现的算法适宜海量数据的处理，比较提出的并行聚类算法与传统聚类算法的处理效率差别，通过测试不同大小的数据集，查看聚类效率的变化.选取两台配置相同的计算机，其中一台安装了数据挖掘软件weka，可以实现本地串行聚类计算；另一台来自于云平台的一个节点.通过复制样本数据构造不同大小的5组数据集，对每组测试分别进行10次重复操作，取平均值为最终结果，结果见图3.

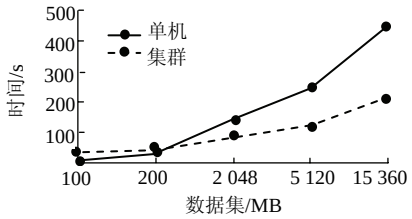

图 3　单机对比结果

Fig.3　single comparison results

从实验结果可以看出，当数据量较少时，单机情况下挖掘算法的效率明显优于并行挖掘算法.这是因为，聚类算法在处理小规模数据时耗时较短，并行挖掘算法在处理少量数据时，耗时主要花费在从节点任务的启动和任务分配上，但随着数据量的增多，单机处理效率急剧下降，并行挖掘算法表现出了明显的优势，说明提出的并行数据挖掘算法非常适合海量用电数据处理.

3.3　加速比性能实验

加速比是相同任务在一台计算机上执行的时间和n台计算机执行的时间之比.加速比可以反应集群执行的性能和效率.在实验中测试了处理样本大小为15 360 MB、2 048 MB和5 210 MB的加速比，记为数据1、数据2和数据3，实验结果见图4.

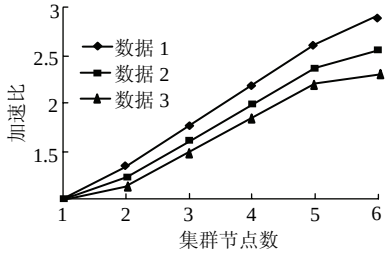

图 4　加速比性能实验结果

Fig.4　speedup performance experimental results

从图4中可以看出集群加速比接近线性，表明节点数的增加能有效提高集群的效率，且数据量越大，加速比性能越好，说明提出的并行聚类算法能够高效的处理海量数据，并且表现出良好的性能，这主要是因为两个Map和两个Reduce函数设计的比较合理，特别是在K-means处理阶段，Combine 函数的加入，使中间结果<key，value>键值对在节点间的通信对带宽的消耗大大降低.

3.4　用电数据聚类分析实验

结合本文提出的聚类算法和样本数据完成对居民用电数据聚类分析的任务，聚类结果见图5.从图5中可以看出，电力用户被分为了A类、B类和C类，把153户实际用电曲线分别取上述3类曲线进行对比，结果聚类精度达到了九成以上.从图5中可以看出，A类用户在夜间用电量很低，在8点和12点左右出现小的波峰，到晚上6点至22点达到用电高峰，推测此类用户可能为上班族较多的用户；B类用户在夜间用电量也很低且小于A类用户，在白天用电量开始上升，用电量基本稳定，推测此类用户可能为老人比较多的用户；C类用户一直处于高用电量状态，推测此类用户可能为商业用户.可以通过对3种用户类型用电量的对比制定相应的用电策略，指导最优用电.

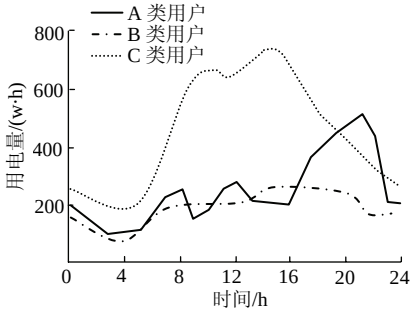

图 5　用电数据分析结果

Fig.5　electricity data analysis results

4　结论

（1）针对用电数据量大的难题，进行了基于云计算的用电数据分析构架研究，利用云计算技术达到处理海量用电数据的目的.

（2）利用Canopy聚类算法简单、快速和无需提前设定聚类个数的优点，提出一种新的聚类思路，即利用了K-means简单、快速的优点，又避免了因初

始聚类个数设置不当而造成局部最优的缺陷.

（3）利用云计算技术高效处理海量数据的特点，把提出的算法部署到了MapReduce框架上，以居民用电数据为基础，通过实验验证了算法的可行性、高效性、准确性.

参考文献：

[1] 赵莉,候兴哲,胡君,等.基于改进k-means算法的海量智能用电数据分析[J].电网技术,2014,38(10):2 715-2 720.

ZHAO Li,HOU Xingzhe,HU Jun,et al.Improved K-Means algorithm based analysis on massive data of intelligent power utilization[J].Power System Technology,2014,38(10):2 715-2 720.

[2] 郝洪星,朱玉全,陈耿,等.基于划分和层次的混合动态聚类算法[J].计算机应用研究,2011,28(1):51-53.

HAO Hongxing,ZHU Yuquan,CHEN Geng,et al.Hybrid dynamic clustering algorithm based on partition and hierarchical clustering[J]. Application Research of Computers,2011,28(1):51-53.

[3] 孙福权,张达伟,程勖,等.基于Hadoop企业私有云存储平台的构建[J].辽宁工程技术大学学报(自然科学版),2011,30(6):913-916.

SUN Fuquan,ZHANG Dawei,CHENG Xun,et al.Establishment of enterprise private cloud storage platform based on Hadoop[J].Journal of Liaoning Technical University(Natural Science),2011,30(6):913-916.

[4] 宋亚奇,周国亮,朱永利,等.智能电网大数据处理技术现状与挑战[J].电网技术,2013,37(4):927-935.

SONG Yaqi,ZHOU Guoliang,ZHU Yongli,et al.Present status and challenges of big data processing in smart grid[J].Power System Technology,2013,37(4):927-935.

[5] 谢雪莲,李兰友.基于云计算的并行K-means聚类算法研究[J].计算测量与控制,2014,22(5):1 510-1 512.

XIE Xuelian,LI Lanyou.Research on parallel K-means algorithm based on cloud computing platform[J].Computer Measure and Control,2014, 22(5):1 510-1 512.

[6] 穆瑞辉,苗国义.基于粒子群优化的模糊K-means目标分类算法[J].计算测量与控制,2013,21(5):1 266-1 268.

MU Ruihui,MIAO Guoyi.Algorism for goal classification based on particle swarm optimization and fuzzy K-means[J].Computer Measure and Control,2013,21(5):1 266-1 268.

[7] 钱彦江.大规模数据聚类技术研究与实现[D].成都:电子科技大学,2009.

QIAN Yanjiang.Research and implementation of large scale data clustering technology[D].Chengdu:University of Electronic Science and Technology of China,2009.

[8] 张石磊,武装.一种基于Hadoop云计算平台的聚类算法优化的研究[J].计算机科学,2012,39(10):115-118.

ZHANG Shilei,WU Zhuang.Clustering algorithm optimization research based on Hadoop[J].Computer Science,2012,39(10):115-118.

[9] 江务学,张璟,王志明.云计算及其架构模式[J].辽宁工程技术大学学报:自然科学版,2011,30(4):575-579.

JIANG Wuxue,ZHANG Jing,WANG Zhiming.Cloud computing and its architecture model[J].Journal of Liaoning Technical University:Natural Science,2011,30(4):575-579.

[10] 常润梅,孟利青,刘万军.电信企业云计算数据中心容量管理[J].辽宁工程技术大学学报:自然科学版,2013,32(8):1 112-1 117.

CHANG Runmei,MENG Liqing,LIU Wanjun.Telecom enterprise cloud computing data center capacity management[J].Journal of Liaoning Technical University:Natural Science,2013,32(8):1 112- 1 117.