DOI: 10. 3969 / j. issn. 1673-4785. 201411011
网络出版地址: http: //www. cnki. net/kcms/detail/23. 1538. tp. 20150716. 0934. 003. html

一种基于内存计算的电力用户聚类分析方法

王德文,孙志伟
(华北电力大学 控制与计算机工程学院,河北 保定 071003)

摘 要: 随着智能电表与采集终端采集的用电数据迅猛增长,传统数据分析方法已经不能满足大数据环境下智能用电行为分析的需要。鉴于 K - means 算法具有计算效率高、容易并行化等特点,采用弹性分布式数据集与并行内存计算框架对其进行改进与并行化,减少作业的运行与输入输出操作时间,提高聚类分析的处理能力。对用电测量数据进行预处理构建实验数据集,实验结果表明本方法对电力用户聚类分析的准确率高于单机 K - means 方法,其处理速度和能力明显优于单机和基于 MapReduce 并行计算框架的聚类方法,并对数据的增长具有较好的适应性。

关键词: 大数据; 智能用电; 弹性分布式数据集; 内存计算; 聚类分析
中图分类号: TP18 **文献标志码:** A **文章编号:** 1673-4785(2015)04-0569-08

中文引用格式: 王德文,孙志伟. 一种基于内存计算的电力用户聚类分析方法[J]. 智能系统学报, 2015, 10(4): 569-576.
英文引用格式: WANG Dewen, SUN Zhiwei. A method for cluster analysis of electric power consumers based on in - memory computing[J]. CAAI Transactions on Intelligent Systems, 2015, 10(4): 569-576.

A method for cluster analysis of electric power consumers based on in-memory computing

WANG Dewen, SUN Zhiwei
(School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China)

Abstract: With the rapid growth of electricity consumption data collected by smart electric meters and data acquisition terminals, the traditional data analysis method cannot meet the demand of smart power consumption behavior analysis in the big data environment. Since K-means algorithm demonstrates high calculation efficiency, easy parallelization and other characteristics, a method for improving and parallelizing K-means with the resilient distributed data set and parallel in-memory computing framework is presented, the running time of job operation and I/O operations is reduced, and the ability of clustering analysis is improved. The experimental data set is built by preprocessed electricity consumption data. Eexperimental results show that the accuracy rate by this cluster analysis method for electric power users is obviously better than the single machine K-means algorithm. The processing speed and ability of this method are superior to the single machine and the clustering method based on MapReduce parallel computing framework, and this method has good adaptability for the growth of data.

Keywords: big data; smart electricity consumption; resilient distributed data set; in-memory computing; cluster analysis

电力用户行为分析是通过分析用电数据之间关联性和相似性,发现用户潜在的行为习惯,进行用户细分,对于引导用户的用电行为与节能改造具有重要意义[1-2]。随着智能用电的飞速发展,智能电表与采集终端得到广泛应用,已扩大到居民用户等各种电力场所,采集及处理的用电数据呈指数级增长、数据量巨大、结构类型繁多、交互性强,逐渐进入用电大数据时代[3]。传统的数据分析与处理方法存在计算能力不足、处理效率低的瓶颈,已不能完全满足大数据环境下智能用电数据快速分析的需求。

收稿日期: 2014-11-10. 网络出版日期: 2015-07-16.
基金项目: 国家自然科学基金资助项目(61074078); 中央高校基本科研业务费专项资金资助项目(12MS113).
通信作者: 孙志伟. E-mail: sunzw20120901@126. com.

聚类分析作为数据挖掘[4]中的一个重要分支，能够对数据进行全局分析，得出数据的分布特征，已经被用于电力用户行为分析领域。例如，文献[5]通过对电力用户负荷特性进行分析，在传统行业划分为基础上使用聚类算法对用户进行分类研究，但没有将用户的用电习惯考虑进去。文献[6]针对变电站负荷提出模糊C均值聚类方法，把变电站负荷分为工业、农业、市政等类别，结论认为该方法明显优于基于等价关系的聚类法。文献[7]将模糊聚类方法应用于电力销售领域，利用负荷曲线特征实现对电力用户分类，为售电企业制定合理的电价和有效实施负荷管理提供参考。上述传统聚类方法均没有考虑智能用电行为分析在大数据环境下对海量数据的可靠存储、高效管理与快速分析等方面所面临的挑战。

大数据分析侧重于通过分布式或并行算法提高现有数据挖掘方法对海量数据的处理效率。云计算具有高可靠性、海量数据处理、扩展性强以及设备利用率高等优点，已经成为大数据分析的基础支撑技术。业界已经采用云计算技术对智能用电数据的存储与分析进行了探索，并取得了一定成果。例如，文献[8]基于Hadoop并行计算框架将K-means算法并行化，对居民用电行为进行分析，但对K-means算法的一些参数的选取没有进行相关说明。文献[9]对K-means参数选择进行了改进，但是同样是利用Hadoop并行计算框架对K-means进行并行化。算法在计算过程中需要大量的迭代计算以及I/O操作，Hadoop并不适合处理具有大量迭代计算以及I/O操作的作业，Hadoop在执行过程会有大量的I/O操作，使I/O成为并行计算的瓶颈，严重降低并行计算的性能。并行内存计算框架Spark能够充分利用集群内存，进一步提升快速处理分析能力，为智能用电行为分析提供了一个全新的技术思路[10]。

本文提出一种基于内存计算的并行聚类分析方法(spark-Canopy-Kmeans，SCK)，利用Hadoop的分布式文件系统高效的存储能力[11-13]以及Spark强大的并行内存计算能力，对K-means算法参数选取的盲目性进行改进，并进行内存并行化，实现智能用电数据的准确与快速分析。在Spark集群中开展实验，与传统K-means聚类算法和基于MapReduce并行化的K-means算法(MR-Kmeans)进行对比实验。

1 基于并行内存计算的聚类算法分析

1.1 并行内存计算框架Spark

Spark是一个开源的分布式集群系统，用于大数据的快速处理分析。Spark克服了Hadoop在迭代计算上的不足，现已成为Apache的顶级项目。Spark提供了一种内存并行化计算框架，框架将作业所需数据读入内存，所需数据时直接从内存中查询，这样比基于磁盘的MapReduce访问数据的速度快，减少了作业的运行时间，也减少了I/O操作[10]。

Spark的计算任务特点是在多个计算应用中支持数据集合的共享和重用。为了实现计算过程中的数据集的重用，Spark设计了一个弹性分布式数据集合RDD(resilient distributed dataset)，它是一种类似于分布式内存的数据抽象结构。RDD数据集是一个只读的分区集合，可以在多个计算应用中共享，它不仅支持基于数据集的应用，还具有容错、局部计算调度和扩展性。RDD支持用户在执行查询时选择缓存数据集在内存中，便于下次计算的数据集重集，减少不必要的数据重复读写操作[14]。

Spark没有自己的文件系统，但可以使用Hadoop支持的文件系统作为输入源或者输出地。Spark作为MapReduce的内存计算的扩展已被广泛的应用于雅虎、Facebook，淘宝等互联网公司的海量数据处理分析中。Spark作业的执行过程如图1所示。

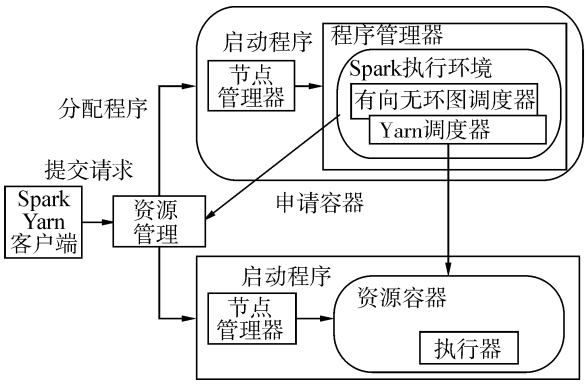


图1 Spark作业执行过程

Fig. 1 The execution process of Spark job

Spark作业的执行过程首先由客户端提交一个作业请求，通过验证之后向资源管理器提交作业，资源管理器将作业初始化并分配一个资源容器，在某个节点管理器中启动程序管理器，程序管理器主要负责对作业的分配，向资源管理器申请资源容器并与相应的节点点管理器进行交互运行作业任务[15-16]。

1.2 内存并行化聚类算法分析

1.2.1 聚类算法分析

Canopy算法是众多聚类算法中计算比较快速的算法，但其聚类精度较低，往往将其作为传统聚类算法的第一步，先对数据集进行粗聚类，然后对粗聚类的结果使用传统的聚类方法进行精细聚类。

K-means 算法主要由 2 步迭代操作构成：第 1 步是分类阶段，将数据集中的数据通过欧式距离划分到离自己最近的聚类中；第 2 步是更新阶段，计算新聚类中的质心以更新之前的质心[17]。上述 2 步迭代是完全独立的，适应并行化运行环境、实现简单、计算效率高。另外，K-means 算法已经被研究应用于用电行为分析领域，便于进行分析比较以验证本文工作成果，因此本文围绕 K-means 算法进行并行化分析、改进与实验对比。

1.2.2　Canopy 算法原理及并行化分析

Canopy 的算法过程首先会选择 2 个阈值 T_1 和 T_2（$T_1 > T_2$），然后从数据集中选择一个数据点作为第 1 个 Canopy 子集的中心点，随后计算各个数据点到此中心点的距离，根据之前设定的 T_1、T_2 阈值来决定隶属哪个 Canopy 子集。其算法步骤为：

1）设置初始距离阈值 T_1、T_2（$T_1 > T_2$），T_1、T_2 的设定原则可以根据实际需求进行多次实验选取也可以使用交叉验证选取。

2）从数据集中随机挑选一个数据点作为第 1 个 Canopy 子集的中心点，并从数据集中删除。

3）计算数据集中第 i 个数据点与 Canopy 子集中心点的粗糙距离 d。

4）判断 d 与 T_1、T_2 的关系。如果 $d < T_2$，将此数据点隶属于当前 Canopy 子集并从数据集中删除此数据点；如果 $d < T_1$，将此数据点隶属于当前 Canopy 子集但并不从数据集中删除此数据点；如果 $d > T_1$，将当前数据点作为一个新的 Canopy 子集中心点。

5）重复第 3）、4）步，直到数据集为空，算法结束。

Canopy 算法流程图如图 2 所示。

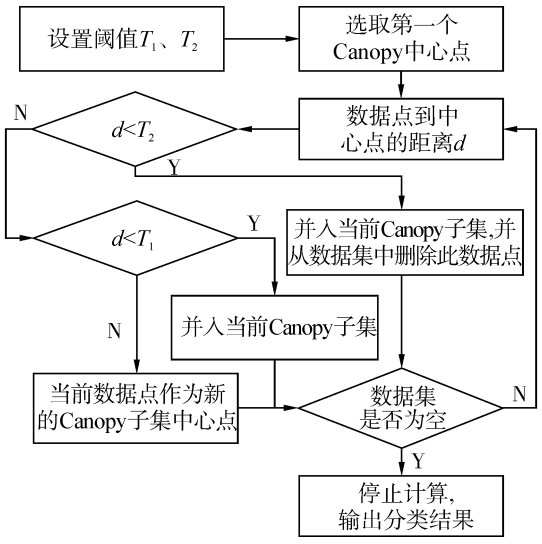


图 2　Canopy 算法流程

Fig. 2　Flowchart of Canopy algorithms

Canopy 算法把聚类过程分为 2 部分，第 1 部分使用一个简单快捷距离计算方法将数据集分为若干个重叠的 Canopy 子集，此过程中每个数据点之间没有联系，只是计算与 Canopy 子集的中心点的距离，可以把数据集分布在若干个计算节点上进行并行计算。第 2 部分为使用一个精准的距离计算方法计算出现在第 1 部分中的同一个 Canopy 子集中的数据与中心点的距离，同样也适合并行计算。

1.2.3　K-means 算法原理及并行化分析

K-means 算法是解决聚类问题的经典算法，其主要思想是从数据集 S 中选择 k 个点作为初始聚类的质心，接下来将数据中的每个点与距它最近的质心聚类[18-20]。K-means 执行流程图如图 3 所示。

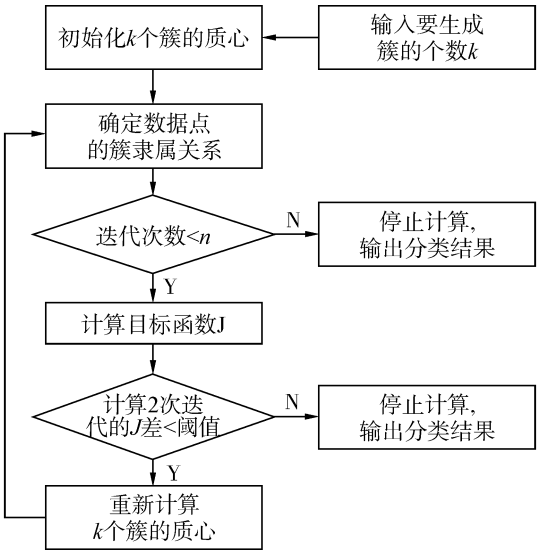


图 3　K-means 聚类流程图

Fig. 3　K-means clustering flowchart

其算法步骤如下：

1）对数据集 S 决定 k（$k < |S|$）的值，也就是对数据集 S 的分类个数。

2）在数据集 S 中选取 k 个数据点作为初始簇的质心 k_1，k_2，…，k_k。

3）对数据集 S 中第 i 个样本点 s_i 计算其与各个簇质心 k_j 的距离，将 s_i 分配给最近的簇质心。第 i 个样本点到第 j 个质心的距离

$$K_j(i) = \min\{\| s_i - k_j \|^2\} \tag{1}$$

式中：$i = 1, 2, \cdots, |s|$，$j = 1, 2, \cdots, k$，s_i 表示 s 中第 i 个样本点，k_j 表示第 j 个质心，公式中距离采用欧式距离。

4）判断是否满足迭代次数，满足则停止计算；否则采用误差平方函数计算目标函数：

$$J = \sum_{j=1}^{k} \sum_{i=1}^{|S|} \| s_i - k_i \|^2 \tag{2}$$

式中:k 为要聚类的个数,$|S|$ 为样本的个数,k_i 为第 j 个质心。

5) 计算 ΔJ,判断是否满足阈值,满足则停止计算。否则执行第6)步。

6) 对上步得到的新簇重新估算 k 个簇的质心,

$$k_j = \frac{1}{|K_j|}\sum_i s_i, i \in K_j, j = 1,2,\cdots,k \quad (3)$$

式中:s_i 表示数据集中的样本点,$|K_j|$ 表示第 j 个聚类中样本点的个数,k_j 则为新聚类的中心点。之后转到第3)步。

1.3 基于内存计算的聚类分析方法

K-means 算法虽然简单、容易理解和实现,但是仍有一些不足,如初始 k 值无法确定,需反复多次尝试寻求最优解 k;初始的聚类中心点无法确定,目前多是随机选取 k 个中心点,当面对海量的数据集时其迭代过程繁琐,运行时间较长等。

使用 Canopy 算法能够快速对数据进行粗聚类的特点,将原始数据集分为 p 个重叠的子集,则此时的 p 即为随后 K-means 算法中初始的 k 值,p 个重叠的子集的中心点为 K-means 算法中初始的 K-means 聚类中心点。其次,将此设计思路在并行内存计算框架 Spark 上实现。实现的具体步骤如下:

1) 从分布式文件系统上读取数据集生成 RDD。

2) 将原数据集通过 map 进行格式化,并执行 cache 操作,将数据读入内存。

3) 在各计算节点上读取本地数据进行计算与 Canopy 中心点的距离 d。

4) 判断距离 d 与 T_1、T_2 的关系。如果 $d < T_2$,将此数据点隶属于当前 Canopy 子集并从数据集中删除此数据点;如果 $d < T_1$,将此数据点隶属于当前 Canopy 子集但并不从数据集中删除此数据点;如果 $d > T_1$,将当前数据点作为一个新的 Canopy 子集中心点,并广播到全局的 Canopy 中心点集中。

5) 如果数据集为空时,将生成的 p 个 Canopy 子集进行 RDD 操作。否则转到第3)步。

6) 将上一步产生的 p 个 Canopy 中心点赋值给 K-means 中 k 个聚类的中心点,且 $k = p$。

7) 计算 Canopy 子集中每个数据点到中心点的距离,进行 K-means 聚类。

8) 对 RDD 执行 Reduce 操作将局部聚类合并成全局聚类,并计算新聚类中数据点的平均值,作为新聚类的中心点。

9) 对新中心点做 Map 操作,计算其所属的 Canopy 子集,计算新旧中心的平方差,更新聚类中心点。其算法流程如图4所示。

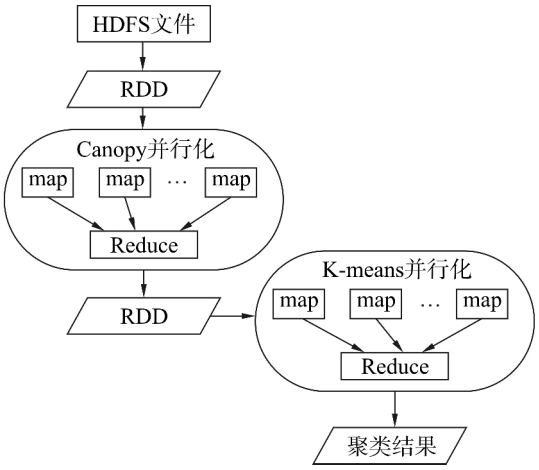


图4 Canopy 和 K-means 的内存并行化流程图

Fig. 4 The in-memory parallelization flowchart of Canopy and K-means

SCK 利用 Spark 的特性将 Canopy 粗聚类的数据放置在内存中,方便随后 K-means 聚类的时候可以多次重复使用,而不需要再次从分布式文件系统中读取,减少 IO 操作,提高访问速度。而且在 K-means 计算过程中只需要计算 Canopy 子集中的数据,而无需对整个数据集进行计算,减少了计算量,更加适合进行大数据处理。

2 实验与结果分析

2.1 智能用电系统架构

本文设计一个智能用电系统,安装在智能小区中,包括智能插座、智能开关、智能电表和相关传感器,其系统逻辑架构如图5所示。

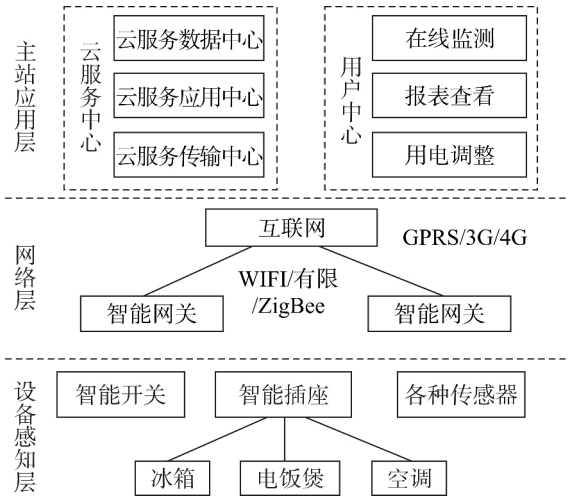


图5 智能用电系统架构图

Fig. 5 Architecture diagram of smart electricity consumption system

2.2　实验数据

1) 原始测量数据

本实验原始数据来源于居民用电的实际测量数据，数据的采集频率为 1 min，每户数据约 200 万条，采集内容包括用户标识、采集日期、采集时间、有功功率、电压、电流、智能插座 1 用电量、智能插座 2 用电量、智能插座 3 用电量等，如表 1 所示。

表 1　居民用电测量数据

Table 1　Measurement data of electric power consumption

字段属性	描述
用户标识	用户唯一标识
采集日期	格式为 2012/12/17
采集时间	格式为 20:27
有功功率/kW	平均每分钟有功功率
电压/V	电压
电流/A	平均每分钟电流
智能插座/W · h	冰箱、空调、洗衣机、微波炉等家用大功率电器的用电量

图 6 给出原始测量数据中某天有功功率曲线实例。

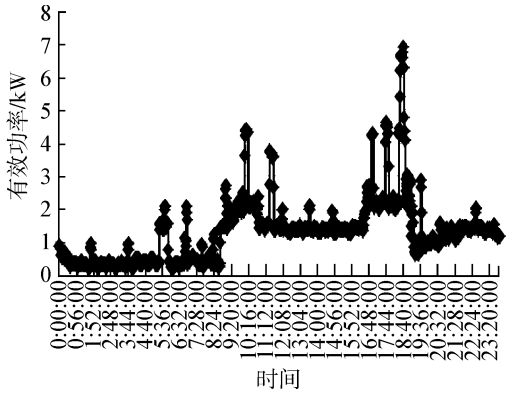


图 6　某天测量数据实例

Fig. 6　Example of measurement data in a day

2) 数据预处理与实验数据集构建

原始测量数据无法直接用于实验分析，需要对其进行预处理，按照实验目的构建实验数据集。

原始测量数据中的电压、电流在实验中无需使用，需要进行删除。原始测量数据中存在约 1.3% 的空缺值，需要对其进行删除，并增加常驻人口与居住面积等数据。原始数据采集频率为 1 min，实验所需的数据无须精确到分钟，将每天的数据进行合并，计算统计出每天用电量、峰电量、谷电量与平电量等，并进行单位转换。新构建的实验数据集包括用户标识、采集日期、每日用电量、峰电量、谷电量、平电量、常住人口与居住面积等，如表 2 所示。

峰电量为当日用电高峰期所用电量，例如 7:00 ~ 12:00，19:00: ~00:00，谷电量为当日用电低谷期所用电量，例如 00:00 ~ 7:00。平电量为当日用电不是高峰期和用电低谷期用的电量。图 7 给出实验数据集所构成的某户一周内用电量曲线。

表 2　实验数据集

Table 2　Experimental data sets

字段属性	描述
用户标识	用户唯一标识
采集日期	格式为 20121217
用电量/kW · h	每日用电量
峰电量/kW · h	每日峰电量
谷电量/kW · h	每日谷电量
平电量/kW · h	每日平电量
常住人口/人	家庭居住人口数
居住面积/m^2	住房实际使用面积单位

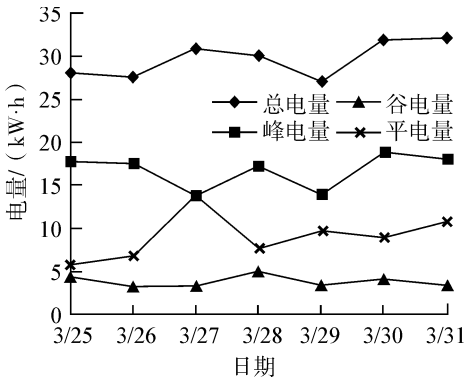


图 7　一周内用电量曲线图

Fig. 7　Electric power consumption graphs in a week

在下面的实验过程中，将从实验数据集中随机选取具有高耗能、中等耗能、低耗能典型特征的用户用电数据进行实验测试，进行多次的测试，取平均值为最终实验结果。实验数据集虽然没有达到大数据的规模，但可以用此实验数据进行算法正确性实验，并对实验数据集扩充进行内存并行化性能测试。

2.3　实验结果分析

1) 实验 1 结果分析

本实验采用 SCK 对采集到的海量智能用电数据进行聚类分析，其聚类结果的准确率达到了 90.7%，其中 9.3% 的用户聚类错误的原因为用户在某一天或者某一时刻改变了用电规律，造成采集的用电数据发生较大的波动，但也不排除用电数据在采集过程中或者传输过程中发生错误。其聚类结果如表 3 所示。由表 3 中的数据计算可得使用 SCK

的准确度为90.7%，高于单机K-means聚类算法的准确度86.37%，而且各个类别的单独聚类结果也普遍高于单机K-means算法结果。

2) 实验2结果分析

本实验采用SCK对采集到的海量智能用电数据进行聚类分析，并与单机K-means聚类算法进行效率对比。所采集的数据有限，在实验过程中需要人为不断增加数据规模(0.32、1.8、5.2、20.8 GB)，以考察数据集大小的变化与聚类时间和精度的关系。对比实验结果如图8所示。

表3 电力用户聚类分析结果

Table 3 Cluster analysis results for electricity users

类别	SCK 正确率/%	单机K-means 正确率/%
商业用户	100	80.9
上班族+老人+上学族	92	86
上班族+上学族	90.1	90.9
老人+上学族	85.2	63
老人	84.3	76.6
闲置房	98.7	95.2

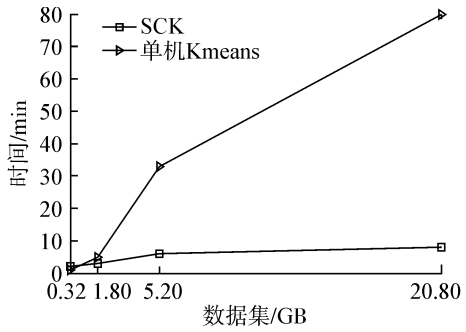


图8 SCK与单机K-means对比图

Fig. 8 Comparison chart of SCK and single machine K-means

图8显示了2种算法在不同数据集中的运行时间，SCK展现了比较好效率。由于SCK在计算初期需要进行一些额外的作业部署工作，在数据集较小时，部署时间所占的比例要大于作业计算的时间，所以当数据集较小时SCK没有单机K-means高效；但是随着数据集的扩大，SCK展现了优越的性能，而单机K-means所展现的性能已不能适合进行聚类分析。SCK通过分布式集群将大数据进行切分部署在不同的计算节点上，并通过将所需数据读入内存进行反复直接访问，有效减少了IO操作，缩短了数据访问时间，并且通过各个独立的处理机提升了数据并行计算的能力，因此能够对大数据进行高效聚类。

3) 实验3结果分析

本实验将SCK与MR-Kmeans算法进行效率对比实验。将不同大小的数据集分别采用SCK和MR-Kmeans算法进行聚类分析，其实验结果如图9所示。

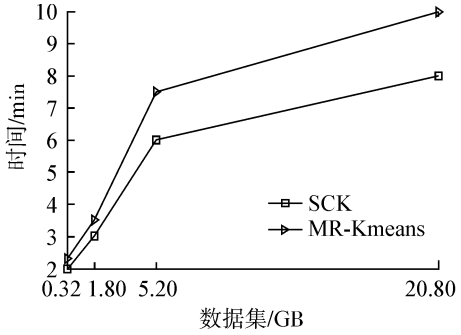


图9 SCK与MR-Kmeans对比图

Fig. 9 Comparison chart of SCK and MR-Kmeans

图9显示了2种改进的K-means算法的运行时间对比图，由图得知相同数据集下SCK运行时间比MR-Kmeans算法略快，随着数据集的增大两者的时间差也在增大，但是SCK时间增长比较缓慢，由此可以得出SCK更加适合处理大数据。

4) 实验四结果分析

K-means并行化后需要衡量算法并行性的好坏，本实验在不同集群大小上运行内存并行化的K-means算法，利用加速比来衡量并行性的好坏，加速比公式为

$$S = t/T \tag{4}$$

式中：t为单机运行的时间，T为集群运行的时间。

将不同大小的数据集分别运行在不同大小的集群中，其运行结果如图10所示。

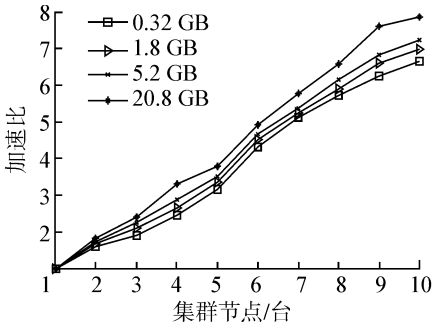


图10 SCK的加速比实验

Fig. 10 The speedup experiments of SCK

由图10可以看出SCK在不同数据量不同大小的分布式集群中显示了接近线性增长的趋势，并且在相同集群大小的情况下数据量越大加速比也越大，但是随着集群的增多加速比会减少，但总的来说随着集群数量的增多加速比会变大。

3　结束语

本文针对传统数据分析方法不能满足大数据环境下智能用电行为分析的问题，给出一种基于内存计算的聚类分析方法，利用并行内存计算框架 Spark 对 K-means 进行改进，实现对智能用电大数据的快速准确分析。实验结果表明，本方法比单机 K-means 和 MR-Kmeans 方法运算速度快并且容易扩展，可以提高聚类精度与处理效率，能够较好满足智能用电大数据分析处理的需要。

虽然实验环境中数据集的大小受到限制，但所进行的实验已模拟数据量的增加，实验结果具有参考价值。下一步工作准备对更大规模数据集进行并行计算分析，并将上述方法应用到智能电网大数据分析的其他领域。

参考文献：

[1] 王蓓蓓，李扬，高赐威. 智能电网框架下的需求侧管理展望与思考 [J]. 电力系统自动化，2009，33(20)：17-22.
WANG Beibei，LI Yang，GAO Ciwei. Demand side management outlook under smart grid infrastructure [J]. Automation of Electric Power Systems，2009，33(20)：17-22.

[2] 何永秀，王冰，熊威，等. 基于模糊综合评价的居民智能用电行为分析与互动机制设计 [J]. 电网技术，2012，36(10)：247-252.
HE Yongxiu，WANG Bing，XIONG Wei，et al. Analysis of residents' smart electricity consumption behavior based on fuzzy synthetic evaluation and the design of interactive mechanism [J]. Power System Technology，2012，36(10)：247-252.

[3] 宋亚奇，周国亮，朱永利. 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. Present status and challenges of big data processing in smart grid [J]. Power System Technology，2013，37(4)：927-935.

[4] 何清. 物联网与数据挖掘云服务 [J]. 智能系统学报，2012，7(3)：189-194.
HE Qing. The Internet of things and the data mining cloud service [J]. CAAI Transactions on Intelligent Systems，2012，7(3)：189-194.

[5] 冯晓蒲，张铁峰. 基于实际负荷曲线的电力用户分类技术研究 [J]. 电力科学与工程，2010，26(9)：18-22.
FENG Xiaopu，ZHANG Tiefeng. Research on electricity users classification technology based on actual load curve [J]. Electric Power Science and Engineering，2010，26(9)：18-22.

[6] 李培强，李欣然，陈辉华，等. 基于模糊聚类的电力负荷特性的分类与综合 [J]. 中国电机工程学报，2005，25(24)：73-78.
LI Peiqiang，LI Xinran，CHEN Huihua，et al. The characteristics classification and synthesis of power load based on fuzzy clustering [J]. Proceedings of the CSEE，2005，25(24)：73-78.

[7] 段铡，张彩庆，刘爱芳. 模糊聚类在电力用户分类中的应用 [J]. 电力需求侧管理，2005，7(5)：18-20.
DUAN Ru，ZHANG Caiqing，LIU Aifang. Application of fuzzy clustering method in classification of electricity customers [J]. Power DSM，2005，7(5)：18-20.

[8] 张素香，刘建明，赵丙镇，等. 基于云计算的居民用电行为分析模型研究 [J]. 电网技术，2013，37(6)：1542-1546.
ZHANG Suxiang，LIU Jianming，ZHAO Bingzhen，et al. Cloud computing-based analysis on residential electricity consumption behavior [J]. Power System Technology，2013，37(6)：1542-1546.

[9] 毛典辉. 基于 MapReduce 的 Canopy-Kmeans 改进算法 [J]. 计算机工程与应用，2012，48(27)：22-26.
MAO Dianhui. Improved Canopy-Kmeans algorithm based on MapReduce [J]. Computer Engineering and Applications，2012，48(27)：22-26.

[10] ZAHARIA M，CHOWDHURY M，FRANKLIN M J，et al. Spark：cluster computing with working sets [C] //Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. Berkeley，CA，USA：USENIX Association，2010.

[11] 赵薇，刘杰，叶丹. 基于组件的大数据分析服务平台 [J]. 计算机科学，2014，41(9)：75-79.
ZHAO Wei，LIU Jie，YE Dan. Module based big data analysis platform [J]. Computer Science，2014，41(9)：75-79.

[12] 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)：2715-2720.
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)：2715-2720.

[13] 程艳柳. 基于云计算的智能电网数据挖掘的研究 [D]. 保定：华北电力大学，2013：15-20.
CHENG Yanliu. Research on smart grid data mining based on cloud computing [D]. Baoding：North China Electric Power University，2013：15-20.

[14] ZAHARIA M，CHOWDHURY M，DAS T，et al. Resilient distributed datasets：a fault-tolerant abstraction for in-memory cluster computing [C] //Proceedings of the 9th

USENIX Conference on Networked Systems Design and Implementation. Berkeley, USA: USENIX Association, 2012: 1-14.

[15] LIN X Q, WANG P, WU B. Log analysis in cloud computing environment with Hadoop and Spark [C] //2013 5th IEEE International Conference on Broadband Network & Multimedia Technology (IC-BNMT). Guilin, China: IEEE, 2013: 273-276.

[16] GU L, LI H. Memory or time: performance evaluation for iterative operation on Hadoop and Spark [C]. 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC). Zhangjiajie, China: IEEE, 2013: 721-727.

[17] 海沫，张书云，马燕林. 分布式环境中聚类问题算法研究综述 [J]. 计算机应用研究，2013，30(9)：2561-2564.
HAI Mo, ZHANG Shuyun, MA Yanlin. Algorithm review of distributed clustering problem in distributed environments [J]. Application Research of Computers, 2013, 30(9): 2561-2564.

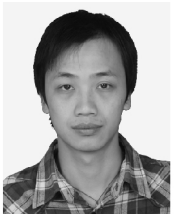
[18] 余晓山，吴扬扬. 基于 MapReduce 的文本层次聚类并行化 [J]. 计算机应用，2014，34(6)：1595-1599，1680.
YU Xiaoshan, WU Yangyang. Parallel text hierarchical clustering based on MapReduce [J]. Journal of Computer Applications, 2014, 34(6): 1595-1599, 1680.

[19] MCCALLUM A, NIGAM K, UNGAR L H. Efficient clustering of high-dimensional data sets with application to reference matching [C] //Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2000: 169-178.

[20] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient k-means clustering algorithm: Analysis and implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881-892.

作者简介：

王德文，男，1973 年生，副教授，主要研究方向为云计算、大数据分析。



孙志伟，男，1987 年生，硕士研究生，主要研究方向为云计算与大数据挖掘。

[责任编辑：刘畅]

第一届国际智能信息系统应用研讨会
First International Workshop on Applied Intelligent Information Systems (AIIS 2015)

Our society needs and expects more high-value services. Such "knowledge-intensive" services can only be delivered if the necessary organizational and technical requirements are fulfilled. In addition, the cost-benefit analysis from the service provider point of view needs to be positive. There is a large and rapidly increasing literature on how artificial intelligence might be used to developmore "intelligent" information systems. The proposed workshop will address all possible research in the Intelligent Information Systems.

The workshop will primarily address the following themes:

1) Information Storage and Retrieval;
2) Data Structures, Cryptology and Information Theory;
3) Artificial Intelligence (incl. Robotics);
4) IT in Business;
5) Document Preparation and Text Processing;
6) Industry Sectors;
7) Electronics;
8) IT & Software;
9) Telecommunications.

Website: http://www.icdim.org/iis.html