基于改进的并行*K*-Means算法的电力负荷聚类研究

许元斌[1],李国辉[2,3],郭 昆[2,3],郭松荣[2,3],林 炜[2,3]
XU Yuanbin[1], LI Guohui[2,3], GUO Kun[2,3], GUO Songrong[2,3], LIN Wei[2,3]

1.国网信通亿力科技有限责任公司,福州 350001
2.福州大学 数学与计算机科学学院,福州 350116
3.福建省网络计算与智能信息处理重点实验室,福州 350116
1.State Grid Electic Power Company, Fuzhou 350001, China
2.College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
3.Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, China

XU Yuanbin, LI Guohui, GUO Kun, et al. Research on parallel clustering of power load based on improved *K*-Means algorithm. Computer Engineering and Applications, 2017, 53(17):260-265.

Abstract: The electrical power enterprise usually based on power load data, uses the traditional *K*-Means algorithm to classify the customers, but the biggest drawback of this method must be specified by the user manual clustering number of clusters. It proposes a method combining Canopy algorithm and *K*-Means algorithm based on load clustering, without the need to manually specify the number of clusters, the automatic division of the customer. First of all, it collects users' electricity data, uses the parallel computing framework MapReduce to preprocess the original data. Then, it uses Canopy and *K*-Means algorithm to establish the clustering model of automatic load. Finally, in the real consumption data on the empirical analysis, by using the Silhouette index to evaluate, it shows that the proposed method is more stable and convenient, and has wider applicability.

Key words: load clustering; parallel computing; Canopy; *K*-Means

摘 要: 电力企业通常根据电力负荷数据,采用传统的*K*-Means算法对客户进行划分,而这种方法最大的缺陷就是必须由用户手动指定聚类簇数。提出了一种将Canopy算法和*K*-Means算法结合应用于负荷聚类的方法,无需手动指定聚类簇数。收集到的用户历史用电数据,使用并行计算框架MapReduce对原始数据进行预处理。应用Canopy和*K*-Means算法建立自动负荷聚类模型。在真实用电数据上进行实证分析,通过使用Silhouette指标对结果进行评估,证明提出的方法更加稳定和具有广泛的适用性。

关键词: 负荷聚类;并行计算;Canopy;*K*-Means
文献标志码: A **中图分类号:** TP311 doi:10.3778/j.issn.1002-8331.1603-0110

1 引言

随着我国的经济的蓬勃发展,电力系统的结构变得越来越复杂,导致了电力负荷规模愈加庞大。如何提升电网的安全性、稳定性以及如何提高经济效益已成为电力企业关注的问题。在电力企业的生产经营和管理中,对电力系统负荷数据有效划分是重中之重的工作环节,也是基本工作环节之一[1]。建立符合实际的动态负荷模型对电力系统规划、设计和运行等诸方面均有十分重要现实意义。对电力负荷数据的分类可以使用数据挖掘技术中的聚类分析。聚类是将物理或抽象对象的集合分

基金项目: 国家自然科学基金(No.61300104);福建省科技创新平台建设(No.2009J1007);福建省自然科学基金(No.2013J01230);福建省高校杰出青年科学基金(No.JA12016);福建省高等学校新世纪优秀人才支持计划(No.JA13021)。

作者简介: 许元斌(1970—),男,高级工程师,研究领域为电力行业信息系统自动化;李国辉(1992—),男,研究领域为大数据挖掘,E-mail:823896856@qq.com;郭昆(1979—),男,博士,副教授,研究领域为大数据挖掘;郭松荣(1991—),男,硕士,研究领域为大数据挖掘;林炜(1992—),男,硕士,研究领域为大数据挖掘。

收稿日期: 2016-03-08 **修回日期:** 2016-06-23 **文章编号:** 1002-8331(2017)17-0260-06
CNKI**网络优先出版:** 2016-12-07, http://www.cnki.net/kcms/detail/11.2127.TP.20161207.0935.002.html

成相似的对象类的过程,这些对象与同一集合内的对象有较相近的特性,而与不同集合中的数据对象有较大的差异[2]。*K*-Means算法是聚类分析中的一种算法,在数据密集且区别明显时,效果特别好,因此,具有较好的实际应用。但算法本身一个比较大的缺陷就是需要手动指定 K 值,使得算法在某些情况下的应用受到限制[3-4]。

国内外许多专家学者也对电力负荷聚类进行了大量的研究。刘建华等人为建立合适的变电站负荷模型,将聚类方法引入负荷特性分析,提出了一种基于ACO-PAM的综合聚类算法[5]。张红斌等人提出了应用KOHONEN神经网络解决负荷动特性聚类的方法[6]。陈星莺等人通过构建特性指标和综合相似系数并计算,将种类繁多的用户聚为有限几类,提出了一种基于需求响应的电力负荷聚类方法[7]。黄麒元等人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的分类[8]。

当前,并行计算应用于各个领域中,效果显著,它是一门综合性的计算机学科,它包括硬件技术,也包括算法、语言、程序设计等软件方面的问题[9]。郑友华将并行计算技术应用于海量遥感图像数据处理和震害提取中,以大幅度提高遥感震害分析处理的速度和精度[10]。吴立新等人通过相应的并行计算策略,进而研发了面向新型硬件构架的新一代GIS的基础地理并行计算算法库和中间件,并已集成到国产高性能GIS平台——HiGIS中[11]。熊玮等人考虑到大型电力系统分析计算耗时过多,提出一种适用于电力系统分析计算的多核并行技术[12]。

聚类分析在生物学、农学、电力等领域的应用日趋广泛。本文讨论聚类在电力负荷中的应用。通过对用户用电负荷数据进行聚类,将用户分成几个大类,针对每类用户分别建模,从多侧面多角度来描述负荷的行为。可以对不同负荷类别的用户进行划分,将其转变为决策型信息,根据不同的用户提供不同的服务,帮助电网企业的市场营销决策并提高其客户服务水平。

本文针对*K*-Means算法需要用户手动指定 K 值的缺点以及并行计算带来的优势,提出了一种负荷聚类的方法,该方法通过结合Canopy算法和*K*-Means算法来建立模型。与传统的*K*-Means算法相比,有自动进行聚类的优点,并且提高了运行速度。本文其他部分安排如下:介绍了聚类相关算法Canopy和*K*-Means算法。介绍了负荷聚类的处理流程。分析说明了实验数据以及实验结果。对本文所做的内容进行总结并对未来提出展望。

2　聚类相关技术

2.1　Canopy简介

Canopy算法使用一种代价比较低的相似性度量方法,快速粗略地把数据划分为若干个重叠子集,每个子集可以看成一个簇[13]。算法过程如下:

(1)把所有数据点加入一个数据集,设置两个距离阈值 $T1$, $T2$ 的值,其中 $T1>T2$ 。

(2)从数据集合中任意取一个数据点,构建初始Canopy,并从该数据集中删除该数据点。

(3)从其余的数据集合中任意取一个数据点,分别计算它到所有Canopy中心的距离,如果该点与某个Canopy中心点的距离在 $T1$ 以内,将这个点加入到该Canopy中。如果该点与某个Canopy中心点的距离小于 $T2$,将这个点加入到该Canopy中,并从数据集合中删除。

(4)重复步骤(3),直至数据集为空。

其中, $T1$ 和 $T2$ 的值可以根据交叉验证法确定。

2.2　*K*-Means算法简介

K-Means算法的核心思想是把 n 个数据点划分为 K 个聚类,使得聚类中心到每个聚类中的观测点的距离和最小[14]。算法过程如下:

(1)随机选择 K 个数据点作为初始聚类中心。

(2)对于其他的数据点,分别计算其到 K 个聚类中心的距离,并将该对象划分到离它最近的聚类中心所在的类中。

(3)分配完全部数据点之后,重新计算每个类的聚类中心。

(4)重复步骤(2),(3),直到各个聚类中心不再发生改变。

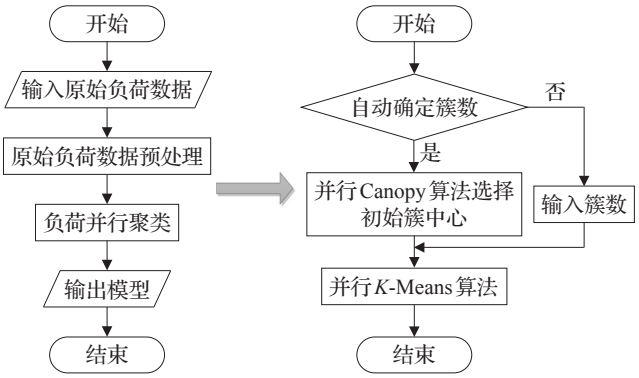
K-Means算法是个非常简洁而且高效的算法,但是它也存在着一些不足。首先,由于实际中往往不知道聚类的簇数 K ,因此算法要求必须事先给定 K 值。其次,算法对噪声和离群点数据比较敏感,而实际数据往往包含一些孤立点,它们会严重影响最后的聚类效果。最后,对初始聚类中心的选择比较敏感,随机选择不同的初始值将得出不同的效果。*K*-Means聚类算法的时间复杂度是 $O(nkt)$,其中 n 代表数据集中对象的数量, t 代表着算法迭代的次数, k 代表着簇的数目。

3　并行负荷聚类

当数据规模较大时,相对于传统的串行处理方式,并行处理能够大大提高程序运行的效率。Hadoop提供了MapReduce并行计算框架和可用于存储海量数据的分布式文件系统HDFS,通过聚类算法对负荷数据进行聚类,流程如图1所示。

3.1　数据预处理

数据挖掘是在大量的数据中挖掘出有用模式的过程,数据源的质量直接影响到了挖掘的效果。因此,在进行数据挖掘之前必须要对数据进行预处理。数据预处理是数据挖掘的重要步骤,它主要包括数据过滤、数据填充、特征规范化等步骤[15]。通过使用MapReduce并行计算模型对负荷数据进行预处理,大大缩短了运行时间。



(a)总体流程　(b)并行聚类流程

图1　并行负荷聚类流程

(1)数据过滤

原始数据中通常会存在某些数据缺失或错漏的情况,因此,需要对数据进行处理,使算法输入更加有意义的数据[16]。常用处理缺失值的方法主要有两种,一种是对含有空缺值的记录进行删除操作,但是这会使记录数变少,往往导致数据不连续,出现一些潜在有用的数据被抹去。因此,对于缺失值个数较多的记录,通常采用直接过滤的方法,将其进行剔除。另一种比较可行的方法是采用某种策略对缺失值进行填充操作。

在原始电力负荷数据中,存在同一个用户编号(同一个user_no值)对应多条记录的情况,处理方法是将同一个用户的所有记录中的同一时刻负荷值进行累加,然后将它们合并成一条记录。另外,针对记录中相邻的多个负荷值,如果出现连续为空值的情况,就设置一个最大允许为连续为空的特征数参数nullCount,当连续为空的负荷特征的数量超过nullCount时就将删除该记录。

(2)数据填充

与上述类似,如果连续的缺失值个数不是很多,通常会使用某种策略对其进行填充。常见的处理方式包括:填充为统一的默认值、填充为特征的统计量(如均值、最小值、中位数等)、删除包含异常值的记录等[17],具体采用哪种方式需要根据问题的的实际情况确定使用哪种可行值进行填充。

本文对缺失值采用处理方法是:填充为相邻负荷的均值。若相邻负荷也是空值,则继续向前向后查找非空负荷值。如果直至第1项负荷(或最末1项负荷)仍为空值,则默认其为0后再计算均值。对于异常值,负荷值允许为负数,因此不作处理。

(3)特征提取

为了减少预测算法中的数据处理量,提高算法的预测效率,还需要对数据做特征选择。特征选择的过程主要包括统计分析方法以及特征选择方法,由于数据中包含许多无意义的值,如果不进行处理则会影响聚类的建模结果。

在负荷数据表中,忽略存在大量名称、时间等对分析作用不大的特征,在数据分析时不予考虑。由于Canopy算法和*K*-Means算法要对负荷数据中的每个数值点进行聚类,需要生成一份完全为负荷值的数据才能作为聚类算法输入。因此,直接提取原始负荷数据中的负荷值。

(4)特征规范化

原始数据不同特征的值域可能存在较大差异。例如原始数据属性的某些特征之间的属性的单位可能不一致,如果直接在原始数据上分析,数值大的特征将湮没数值小的特征,使值域较小的特征无法得到有效利用,使结果发生变化和差错。因此,为了避免这些问题,应该先对原始数据做规范化或标准化操作[18]。常见的规范化方法包括。

① 区间规范化:根据"(原始值 – 最小值)/(最大值 – 最小值)"将数值归一化到[0,1]区间,计算如式(1)。

$$v_i=\frac{v_a-min_A}{max_A-min_A} \tag{1}$$

其中,v_a 为属性 A 的某个值,min_A ,max_A 为属性 A 的最小值和最大值,v_i 为规范化后的值。

② 最大值规范化:根据"原始值/最大值"将数值归一化到[0,1]区间,计算如式(2)。

$$v_i=\frac{v_a}{max_A} \tag{2}$$

其中,v_a 为属性 A 的某个值,max_A 为属性 A 的最大值,v_i 为规范化后的值。

③ 标准规范化:根据"(原始值 – 均值)/标准差"将数值归一化,不一定在[0,1]区间上,计算如式(3)。

$$v_i=\frac{v_a-\bar{A}}{S} \tag{3}$$

其中,v_a 为属性 A 的某个值,$\overline{A}$ 为属性 A 的平均值,S 为属性 A 的标准差,v_i 为规范化后的值。

本文采用的方案是区间规范化。所有的负荷值属性的数据值都被规范在[0,1]区间。需要注意的是,如果某一行中的负荷值全为0,就不进行特征规范化,保持原始的0值。如果某一行的负荷值完全相同,其值就为该数的 n 等分之一。

3.2　并行负荷聚类过程

对电力负荷数据进行负荷聚类主要分为以下几个步骤。

(1)输入原始负荷数据:原始负荷数据原本是保存在本地上,使用分布式计算框架时,为了提高存储的效率和读写的速度,首先将其保存到Hadoop分布式文件系统HDFS上。

(2)数据预处理:由于原始负荷数据中存在一些噪声数据、缺失数据和不一致的数据等,为了保证数据的准确性、完整性、一致性、时效性、可信性,需要对其进行预处理。常见的处理方法包含缺失值处理,数据规约,特征规范化,特征选择等方法。

① 数据过滤,填充作业

Map阶段:

输入:key=记录行偏移,value=记录行。

输出:key=user_no,value=其余特征值。

Reduce阶段:

输入:key=user_no,value=其余特征值。

输出:key=user_no,value=其余特征值。

② 特征提取作业

Map阶段:

输入:key=user_no,value=其余特征值。

输出:key=user_no,value=负荷特征值。

③ 特征规范化作业

Map阶段:

输入:key=user_no,value=负荷特征值。

输出:key=NULL,value=特征规范化后的负荷字段值。

(3)并行聚类:对预处理之后的负荷数据进行聚类,本文采用自动聚类和手动聚类两种方式。如果采用自动聚类,先通过Canopy算法确定簇数 K ,再使用K-Means算法对数据进行聚类。如果采用手动聚类,则直接使用K-Means算法进行聚类。

(4)输出聚类结果:输出类中心及聚类结果,将结果保存到Hadoop分布式文件系统上。

4　实验与分析

4.1　数据集

实验数据来自某电力企业在某城市的用户用电负荷数据,主要包括用户编号,用户名称,行业名称,用户类型,用电类别名称,一天24个时刻的负荷值等特征。各数据的说明如表1所示。

表1　原始数据表特征说明

特征名	特征说明
user_no	用户编号
user_name	用户名称
trade_no	所属行业
trade_name	行业名称
user_type	用户类型
electro_type	用电类别
electrotype_name	用电类别名称
stat_cycle	统计时间
load0 … load23	时刻0至时刻23的电力负荷值
Updatetime	更新时间

4.2　聚类评估指标Silhouette

在聚类的评估中,有基于外部数据的评价,也有单纯的基于聚类本身的评价,其基本思想就是:在同一类中,各个数据点越近越好,并且和类外的数据点越远越好;前者称为内聚因子,后者称为离散因子[19]。

采用计算聚类结果的Silhouette指标的方式评估聚类质量。该指标的计算公式(4)所示。

$$S=\frac{1}{NC}\sum_{i=1}^{NC}\left(\frac{1}{n_i}\sum_{x\in c_i}^{NC}\frac{b(x)-a(x)}{\max(b(x),a(x))}\right)$$

$$a(x)=\frac{1}{n_i-1}\sum_{x,y\in c_i,x\neq y}d(x,y)$$

$$b(x)=\min_{j,j\neq i}\left(\frac{1}{n_j}\sum_{y\in C_i}d(x,y)\right) \tag{4}$$

其中, $a(x)$ 为内聚因子, $b(x)$ 为离散因子。Silhouette指标越大表示聚类质量越好,其最大值对应的类数作为最优的聚类个数。

4.3　实验结果及分析

(1)不同簇数的实验结果

使用Canopy算法结合K-Means算法和单独使用K-Means算法对实验数据进行实验,选取了三组不同 K 值进行实验。为了避免传统K-Means聚类方法存在聚类中心初始值难以确定,聚类结果不稳定的问题,首先通过Canopy算法来粗糙地计算Canopy的数量即为簇数 K 的估计值。经过多次实验,确定 $T1$ 和 $T2$,选取簇数为2,5,7。再通过确定的 K 值,使用K-Means算法进行聚类。由于电力负荷记录较多,图中只显示负荷聚类后的类中心,负荷值是经过归一化后的数值,类中心结果如图2至图7所示。

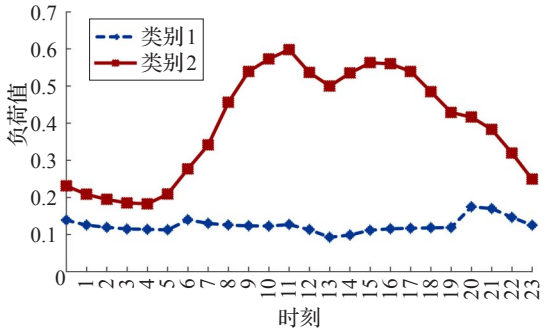


图2　自动聚类($K=2$)

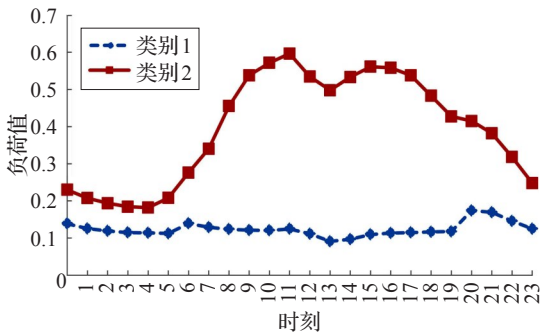


图3　手动聚类($K=2$)

从上面的实验结果分析图可以看出:

① 不同类别对应的不同簇之间的负荷特征确实有显著不同,而且对于类别数较少的情况,聚类算法得到

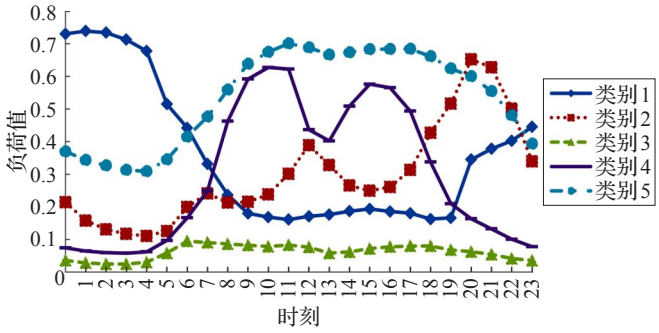


图4　自动聚类($K=5$)

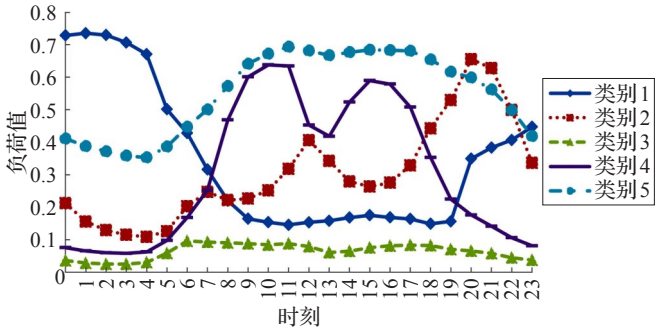


图5　手动聚类($K=5$)

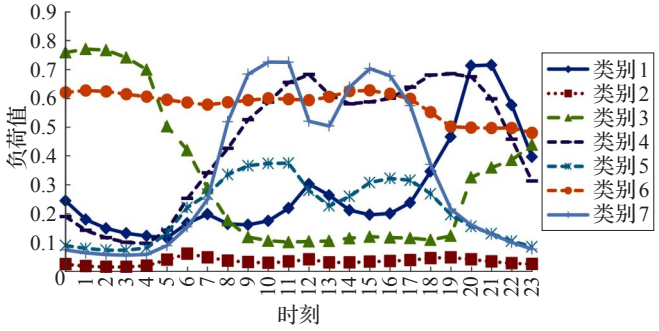


图6　自动聚类($K=7$)

的结果的区分度效果更加明显。其中,当 $K=2$ 时,聚类效果最好,区别度更加明显。

② 当K=2,5,7,自动聚类和手动聚类的类中心结果相近。说明两者聚类的效果较接近。

③ 参数 $T2$ 对聚类结果有显著影响。当 $T2$ 值较小时,由于太接近某个Canopy,而被删除的对象较少,因此生成的初始簇较多。相反地,当 $T2$ 值较大时,较多对象会被删除,因此生成的初始簇较少。

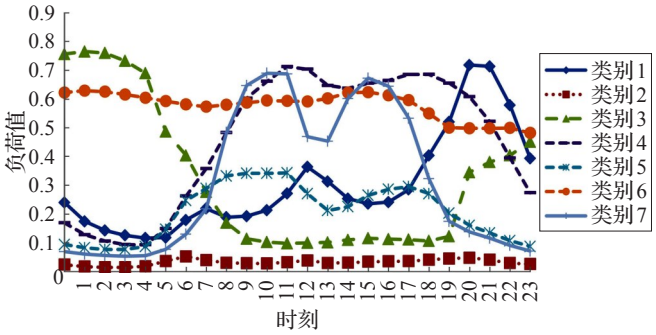


图7　手动聚类($K=7$)

比较两者方法,由于不需要手动指定簇数,因此,Canopy结合*K*-Means算法比手动指定簇数的*K*-Means算法更加方便。

(2)运行时间实验结果

使用Canopy算法结合*K*-Means算法和单独使用*K*-Means算法进行实验,计算所需运行时间。分别进行五次实验。实验结果如图8至图10所示。

从图8至图10的实验结果分析图可以看出:首先,在直接使用*K*-Means算法时,由于类中心是随机产生的,它会影响*K*-Means的迭代次数,因此,不同实验的算法运行时间波动幅度较大。而通过在运行*K*-Means算法前先运行Canopy算法粗略选择类中心后,可以看出*K*-Means算法的的运行时间相对稳定。因此,Canopy算法*K*-Means算法的结合能够得到比单独使用*K*-Means算法更稳定的结果。其次,从 K 值变化对运行时间的影响上看,当 K 值增加时,算法的运行时间呈线性增加,这也符合前面对*K*-Means算法和Canopy算法的时间复杂度与 K 值成正比的讨论。

(3)聚类指标实验结果

由于聚类属于无监督学习,在先前无法得到所属类别信息时,无法判断聚类效果的好坏。采用上述的Silhouette指标,分别对Canopy结合*K*-Means和单独使用*K*-Means的聚类结果进行评估。实验取*K*=2,5,7。评估结果如表2所示。

比较Canopy结合*K*-Means算法和单独使用*K*-Means算法的聚类评估指标系数,重复进行5次实验,对结果分别取平均值。当聚类簇数分别为2和5时,两者指标

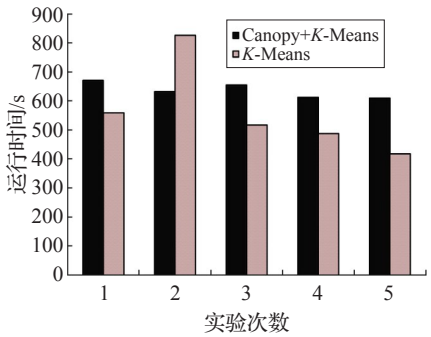


图8　运行时间($K=2$)

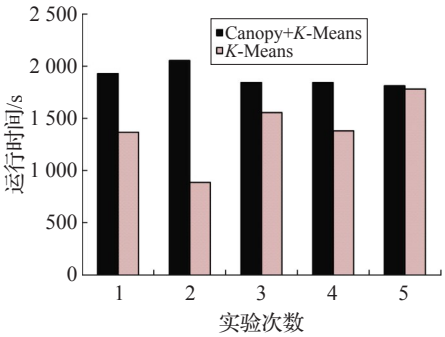


图9　运行时间($K=5$)

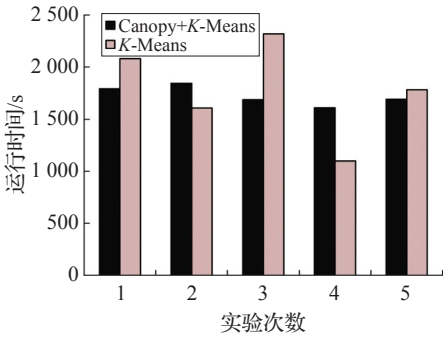


图10　运行时间($K=7$)

表2　Silhouette聚类评估结果

算法	簇数	Silhouette指标值
Canopy+*K*-Means	2	0.341 129 27
	5	0.308 772 53
	7	0.279 651 50
K-Means	2	0.341 733 34
	5	0.308 733 37
	7	0.287 906 23

值非常相近，说明两个算法的在聚类质量基本相同。在聚类质量相同的条件下，使用Canopy结合*K*-Means算法由于不需要手动指定聚类簇数，可以进行生成最合适的簇，更加符合现实中电力大数据聚类应用的需求。

5　结束语

本文提出了一种基于改进的*K*-Means算法的电力负荷聚类方法。首先，对负荷数据进行去重并过滤连续多个空值的记录，再对其进行数据填充和数据特征规范化等预处理操作，采用Canopy和*K*-Means算法结合对预处理后的负荷数据进行聚类，在真实的企业提供的用户用电数据集上进行了实验，此外，通过与传统的*K*-Means算法的对比，验证了提出方法更加稳定和方便，进一步提高了研究结果的效率和可信度。通过本文提出的方法，对大量的用户负荷数据进行分析，并对不同类别的用户进行划分，将其转变为决策型信息，这些决策型信息能够辅助电网企业的市场营销决策并提高其客户服务水平。下一步将就算法如何实现实时聚类，以及进一步提高求解效率等问题展开深入研究。

参考文献：

[1] 黄麒元，王致杰，朱俊，等.基于模糊聚类的电力系统负荷分类分析[J].电力学报，2015(3)：200-205.

[2] Han Jiawei.数据挖掘概念与技术[M].北京：机械工业出版社，2006：251-303.

[3] 周爱武，于亚飞.*K*-Means聚类算法的研究[J].计算机技术与发展，2011，21(2)：62-65.

[4] 白雪峰，蒋国栋.基于改进*K*-means聚类算法的负荷建模及应用[J].电力自动化设备，2010，30(7)：80-83.

[5] 刘建华，王进，杨洪春，等.基于ACO-PAM综合算法的电力负荷聚类分析[J].电力科学与技术学报，2012，26(4)：94-99.

[6] 张红斌，贺仁睦，刘应梅.基于KOHONEN神经网络的电力系统负荷动特性聚类与综合[J].中国电机工程学报，2003，23(5)：1-5.

[7] 陈星莺，王刚，姚建国，等.一种基于需求响应的电力负荷聚类方法：中国，CN104240144A[P].2014.

[8] 黄麒元，王致杰，朱俊，等.基于模糊聚类的电力系统负荷分类分析[J].电力学报，2015(3)：200-205.

[9] 刘赫男，罗霄，高晓东.并行计算的现状与发展[J].煤，2001，10(1)：56-57.

[10] 郑友华.并行计算技术在遥感震害分析处理中的应用研究[D].北京：中国地震局地震预测研究所，2010.

[11] 吴立新，杨宜舟，秦承志，等.面向新型硬件构架的新一代GIS基础并行算法研究[J].地理与地理信息科学，2013，29(4)：1-8.

[12] 熊玮，夏文龙，余晓鸿，等.多核并行计算技术在电力系统短路计算中的应用[J].电力系统自动化，2011，35(8)：49-52.

[13] 余长俊，张燃.云环境下基于Canopy聚类的FCM算法研究[J].计算机科学，2014(B11)：316-319.

[14] 王千，王成，冯振元，等.*K*-means聚类算法研究综述[J].电子设计工程，2012，20(7)：21-24.

[15] 袁梅宇.数据挖掘与机器学习——WEKA应用技术与实践[M].北京：清华大学出版社，2014.

[16] 纪良浩，王国胤，杨勇.基于协作过滤的Web日志数据预处理研究[J].重庆邮电大学学报，2006(5)：646-649.

[17] 朱晓峰.缺失值填充的若干问题研究[D].南宁：广西师范大学，2007.

[18] 王娟，慈林林，姚康泽.特征选择方法综述[J].计算机工程与科学，2005(12).

[19] Wang L，Tan T，Ning H，et al.Silhouette analysis-based gait recognition for human identification[J].IEEE Transactions on Pattern Analysis & Machine Intelligence，2004，25(12)：1505-1518.