DOI: 10.7667/PSPC150548

一种基于 Spark 和聚类分析的辨识电力系统不良数据新方法

孟建良，刘德超

（华北电力大学控制与计算机工程学院，河北 保定 071003）

摘要：随着电力系统智能化建设的不断深入和推进，电力系统数据呈现海量化、高维化的趋势。针对电力系统中的不良数据将导致电力系统状态估计结果的准确性降低，而传统聚类算法处理海量高维数据时单机计算资源不足，近年来较流行的 MapReduce 框架不能有效处理频繁迭代计算等问题，提出一种基于 Spark 的并行 K-means 算法辨识不良数据的新方法。以某一节点电力负荷数据为研究对象，运用基于 Spark 的并行 K-means 聚类算法提取出日负荷特征曲线，分别对输电网状态估计中的不良数据进行检测和辨识。选用 EUNITE 提供的真实电力负荷数据进行实验，结果表明此方法能有效提高状态估计结果的准确性，与基于 MapReduce 框架的方法相比，具有更好的加速比、扩展性，能更好地处理电力系统的海量数据。

关键词：Spark；聚类；K-means；电力系统；不良数据；负荷曲线分类

A new method for identifying bad data of power system based on Spark and clustering analysis

MENG Jianliang, LIU Dechao

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

Abstract: With the development of intelligent power system construction, power data shows a massive and multi dimensions trends. The bad data in power system reduces the accuracy of the estimation results in the state of the power system, computational resources of the traditional clustering algorithms dealing with massive high dimensional data with single machine are insufficient, and the MapReduce, more popular in recent years, cannot effectively deal with frequent iteration calculation problem. According to the above, this paper puts forward a new method of identifying bad data with parallel K-means algorithm based on Spark. To a certain node load data as the research object, the parallel K-means clustering algorithm based on Spark is used to extract daily load characteristic curve, to detect and identify bad data in state estimation of power transmission network respectively. Experiments are conducted with the data of the real load provided by EUNITE, the results show that this method can effectively improve the accuracy of state estimation, and compared with the method based on the MapReduce, it has better speed-up ratio, scalability, and can better process massive data in power system.

Key words: Spark; clustering; K-means; power system; bad data; load curve classification

0 引言

随着智能电网的迅速发展，电力系统的数据呈指数级增长，其结构和运行模式也越来越复杂，因此对系统运行的可靠性、安全性和稳定性也就提出了更高的要求[1-3]。电力系统状态估计是电力系统信息管理系统中一个重要的组成部分[4]。由于客观原因，除了正常的数据噪声，各信息采集单元所获取的测量数据不可避免会有不良数据。不良数据的存在会在不同程度上使电力系统状态估计结果失真，从而不能准确得到系统真实的运行状态，可能会引发未知的安全后果。因此，对不良数据进行检测和处理就显得尤为重要[5-6]。

传统不良数据检测方法取得了大量成果，但仍有不少问题未得到妥善解决。现今对电力数据进行分析和分类控制时，前期处理大多用的是神经网络法和聚类分析法等。文献[7]运用基于蚁群优化算法的负荷序列聚类分析，提高了对外部气象等因素的敏感性，对负荷曲线轮廓相似性具有更细致的聚类性能，但聚类时间较长；文献[8]将模糊聚类技术与

人工神经网络中的BP网络相结合，通过C均值模糊聚类方法实现不同用户日负荷曲线的分类；文献[9]提出一种基于传统K-means聚类算法并结合有效指数准则的不良数据检测和处理方法，但收敛速度慢且易陷入局部极小。为了提高处理海量数据的能力，文献[10]在Hadoop云平台下，建立并行局部加权线性回归模型，并采用最大熵建立坏数据分类模型。然而这些算法几乎都是通过大量的频繁迭代来实现，算法复杂度相当高。尽管传统串行算法可以对电力负荷数据进行聚类，但单机的计算资源依然无法满足算法在处理海量高维数据时大量的资源消耗；而基于MapReduce的算法能处理海量数据，却不能有效处理频繁迭代计算。随着电力系统智能化建设的不断深入，对不良数据的处理有了更高的要求，云计算的出现，为更准确地进行不良数据的检测与辨识提供了可能[11-12]。

围绕上述问题，对输电网状态估计中的不良数据进行识别和纠正，以提高状态估计的准确性。以某个节点的历史负荷数据为研究对象，在云集群环境下，利用基于 Spark 的并行 K-means 算法对该节点的负荷数据进行聚类，提取出日负荷特征曲线；通过与特征曲线对比，辨别和处理不良数据。通过在实验室搭建的 Hadoop 和 Spark 云集群，并采用真实电力负荷数据进行算例分析，验证基于 Spark 平台的方法得到的状态估计结果准确性优于基于传统 K-means 聚类的方法[9]，与传统 Hadoop 平台相比，具有更好的加速比、扩展性，能更好地满足处理电力系统海量数据的需求。

1 基于 Spark 改进的 K-means 并行算法

1.1 传统 K-means 算法

传统 K-means 算法[13-14]的基本思想：首先从 N 个数据对象中随机初始化 K 个聚类中心；对于剩下的其他对象，计算其与 K 个聚类中心的距离，分别将其分配给与其距离最近的类簇；然后再计算每个类簇新的聚类中心，即该类簇中所有对象的均值；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用簇内误差平方总和作为标准测度函数，其定义为

$$E=\sum_{i=1}^{K}\sum_{X\in C_i}\left|X-\overline{X_i}\right|^2 \tag{1}$$

其中：K 为簇的总数；$\overline{X}_i$ 为簇 C_i 的平均值。

1.2 基于 Spark 改进的 K-means 并行算法

1.2.1 Spark架构和弹性分布式数据集RDD

Spark 由加州大学伯克利分校 AMPLab 开发，由于引进了弹性分布式数据集(Resilient Distributed Dataset，RDD)[15]的概念，Spark 可在集群计算中将数据集分布式缓存在各节点内存中，省去大量的磁盘 IO 操作，从而大大缩短访问延迟。作为 Spark 架构的核心机制，RDD 是一种基于分布式内存的并行数据结构，它能将用户数据存储在内存，并控制分区划分以优化数据分布。数据存储在内存中，尤其对于需要多次迭代使用的数据，省去了多次载入到内存和存储到磁盘的过程，大大加快了处理速度。Spark 还支持 RDD 的显式缓存(cache)及持久化(persistence)存储。

Spark运行架构如图1所示，Spark应用在集群上以独立的执行器(executor)运行在不同节点，在主程序中以 SparkContext 对象来进行总体调度。SparkContext 可以与三类集群资源管理器(Standalone、Mesos或者YARN)相连接，集群资源管理器的作用为在不同Spark应用间分配资源。Spark在执行程序时，需要将应用代码发送给工作节点(worker node)的执行器去执行任务(task)，以尽可能实现数据的本地化计算。

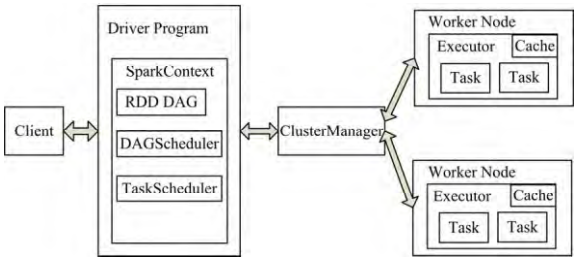

图 1 Spark运行架构

Fig. 1 Spark running architecture

1.2.2 改进 K-means 算法思想

传统 K-means 聚类算法属于聚类中一种基本的划分方法，具有简单、快速的优点。然而这种算法对初值的依赖性很强，初值选取的不同往往导致聚类结果相当不稳定。其次，当初始聚类中心选择不当时，算法极易陷入局部极小点；并且容易受“噪声”数据的影响。其复杂度由 $O(TKN)$表示，其中 K 是期望的聚类簇的个数，T 是迭代次数，N 是数据对象的个数；则其并不能适合处理海量数据。因此考虑用最大最小距离法来优化初始聚类中心。

当最大最小距离法处理的样本规模为 N，每次寻找新的聚类中心时，很明显要进行 N 次距离计算。若共找到 k 个聚类中心，则算法结束时共进行的计算次数为 N^{k-1}。最大最小距离法的计算量取决于 N 的规模，直接将最大最小距离法作用于原始数据集的执行效率很低。考虑到数据集合本身的规律性以

及算法的适用性，因此将其与抽样技术相结合。

初值优化流程图如图 2 所示。

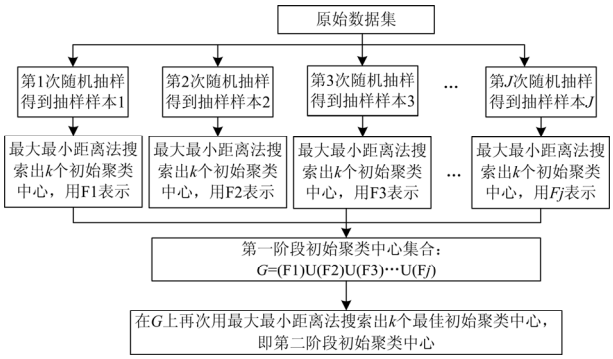

图 2 初值优化流程图

Fig. 2 Flow chart of initial value optimization

首先对原始数据集进行多次随机抽样，然后基于 Spark 运用两阶段最大最小距离法以产生最佳初始聚类中心，最后再用基于 Spark 的并行 K-means 算法进行聚类。因此，此算法的处理流程为：多次随机抽样、最大最小距离法搜索出最佳初始聚类中心、K-means 迭代处理。该算法通过优化 K-means 算法中初始聚类中心, 获得更准确的负荷特征曲线；利用 Spark 并行计算框架实现并行化，克服无法处理海量电力数据的问题，最终实现精确高效的电力负荷曲线分类。

1.2.3 基于 Spark 的改进 K-means 算法并行化实现

利用 Spark 并行实现 K-means，总体上也是采用“ map ”“ reduce ”的思想，即在每次迭代中，先用“ map ”计算所有样本和中心点距离并归类，再用“ reduce ”分类求均值算得新的中心点。然而与 Hadoop 的 MapReduce[16]最大的不同是，Spark 对所有中心点的所有次迭代运算都是在内存中对 RDD 计算完成，中间不需要与磁盘交互，而 Hadoop 的这个过程则要与磁盘有 n (迭代次数×分类数)次的交互。基于 Spark 的改进 K-means 算法实现如图 3 所示。

基于 Spark 的 K-means 算法并行化实现分两部分。第一部分，首先读取 HDFS 的文件(已经预处理过的文件)并创建新的 RDD，并在本地执行 Cache 操作缓存 RDD 数据。之后多次随机抽样产生 J 个抽样样本, 在 Map 过程利用最大最小距离法在本地产生若干初始聚类中心集合，然后在 Reduce 过程将这些初始聚类中心集合汇总，再次调用最大最小距离法得到最佳初始聚类中心集合。第二部分，通过 Map 操作执行局部数据的聚类，Reduce 操作执行汇总局部数据的聚类，计算全局的聚簇。聚类算法的并行化执行是由Spark内核调度完成, 内核会根据工作节点数目，自动将数据集及执行任务分配到不同节点, 继而多个计算节点会并行执行聚类计算。

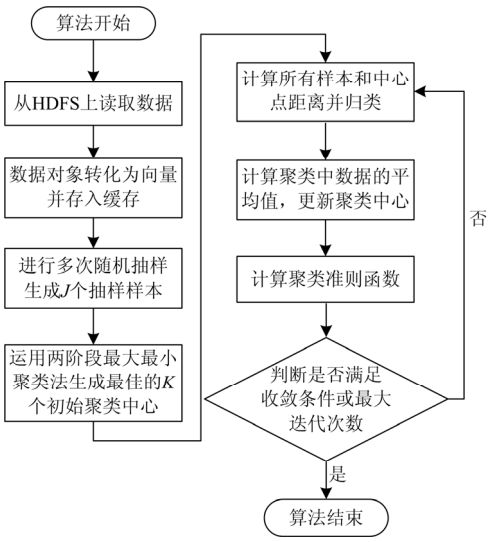

图 3 基于 Spark 的改进 K-means 算法流程图

Fig. 3 Flow chart of improved K-means algorithm based on Spark

与Hadoop相比，新一代并行计算架构Spark的最大优势是以RDD内存计算为核心，即将迭代计算的数据块定义为RDD，以分区(Partitions)的形式分布存储在不同节点的内存中，再由位于这些节点的Tasks针对本地内存Partitions重复完成迭代计算即可，中间完全无需和磁盘进行交互。

2　基于并行 K-means 聚类的负荷特征曲线提取

以一个节点一天 24 小时所测量的负荷数据为纵坐标，以该天各个测量时刻点为横坐标，得出该天的负荷曲线。不良负荷数据在这里特指某个或多个时刻点的负荷值偏离正常值过多。要辨别一条曲线上的某个数据是否为不良数据，需要一个正常数据作为参考，这个参考标准就是负荷特征曲线。相邻几天内的负荷曲线是类似的，下面就是根据曲线的相似性来检测和处理不良数据。为了方便论述，这里定义几个概念。

定义 1 一天中连续 m 个时间点上测量的负荷值连成的曲线称作负荷曲线，记为 $K_i=(x_{i1}, x_{i2}, \cdots, x_{im})$，$ik$ 为第 k 个测量时间点，x_{ik} 为第 k 个测量时间点的负荷值，$k=1, 2, \cdots, m$。则一条负荷曲线即为本文算法中的一个样本。

负荷曲线的相似性是辨别和处理不良数据的关键。以直角坐标为参考，纵向相似性特指相邻几天

内的负荷曲线的形状是类似的，这里以曲线间的距离来表征。

定义 2 负荷曲线 X_i 和 X_j 的距离

$$D_{ij}=\max\left\{\left|x_{ik}-x_{jk}\right|\right\},k=1,2,\cdots,m$$

即两条曲线的距离就是两条曲线上各个测量点上的负荷值差中的最大值。距离 D_{ij} 越小，则曲线 X_i 和曲线 X_j 的相似度就越高，反之则相似度越小。相似精度在一定范围内的曲线归为一个曲线类，也就相当于用本文算法聚类后的一个结果类；这里也把这个范围叫做相似精度。

定义 3 记曲线类 C 为($X_1, X_2, \cdots$)，其中，X_i ($x_{i1}, x_{i2}, \cdots, x_{in}$)。曲线类 C 的相似精度为

$$E(C)=\max_{\substack{i=1,2,\cdots,m\\ j=1,3,\cdots,m}}\max_{k=1,2,\cdots,n}\left\{\left|x_{ik}-x_{jk}\right|\right\}$$

最后给出质心的概念，也即负荷特征曲线。

定义 4 曲线类 C 的质心为

$$\overline{C}=\left(\overline{Y_1},\overline{Y_2},\cdots,\overline{Y_n}\right)$$

$$\overline{Y_k}=\frac{1}{m}\sum_{j=1}^{m}x_{jk},1\le k\le n$$

定义 5 定义曲线 X_j 和曲线类 C 的距离为曲线 X_j 到质心 $\overline{C}$ 的距离

$$D_{jC}=D_{j\overline{C}}=\max_{k=1,2,\cdots,n}\left\{\left|x_{jk}-\overline{Y_k}\right|\right\}$$

负荷特征曲线的提取本质上就是求取各个曲线类的质心。不良数据的产生是偶然的，在所有的数据中所占比例极小，因此它对质心的求取影响也是极小的。要辨别出不良数据首先得辨别出不良数据所在的负荷曲线。正常负荷曲线模式也即负荷特征曲线，要将不正常的负荷曲线提取出来，可以转化为求取到负荷特征曲线也即质心的距离大于某个阈值的负荷曲线。本质上来说，这就是数据挖掘中的聚类分析问题。

聚类分析可以将负荷曲线集分成若干个曲线类。根据定义 1，负荷曲线的横坐标由各个测量时间序列组成，X_i=($x_{i1}, x_{i2}, \cdots, x_{im}$)，每个时间点 x_{ik} 就是一个属性。显然，这是一个 m 维的样本。将所有这样的样本作为本文算法的输入，设定阈值相似精度。正常天气情况下，工作日，周末和节假日的用电负荷显然是不同的。对于配电网来说，即便是同样的日期，不同用户(如居民用电，企业用户和商业用电)的负荷曲线显然也是不同的，因此本文算法中的 K 取值肯定是大于等于 2 的。因为数据来源和篇幅限制，本文只讨论输电网状态估计中的不良数据处理。

3 基于负荷特征曲线的不良数据处理

假设某个曲线类 C 提取出来的负荷特征曲线为 X_t，待检测负荷曲线为 X_i=($x_{i1}, x_{i2}, \cdots, x_{im}$)。从模式识别的角度，辨别含有不良数据的负荷曲线就是计算待检测负荷曲线 X_i 与负荷特征曲线 X_t 的距离 D_{ti}，观察其是否在设定的 D_{ti} 内。如果距离在 D_{ti} 内，则该待检测负荷曲线便属于正常负荷曲线模式。否则，该待测负荷曲线即为非正常负荷曲线。假设待检测负荷曲线 X_i 中的负荷值与负荷特征曲线相应位置的负荷值的差值超过预定范围，则可确定该时间点即为不良数据的具体位置。

设 X_i=($x_{i1}, x_{i2}, \cdots, x_{im}$)，$m$ 为一天中的采样时刻点数。对于采样时间点 k，X_t 和 X_i 对应的负荷值分别为 x_{tk} 和 x_{ik}，则 X_i 相对于负荷特征曲线 X_t 的负荷变化率为 $\delta(k)=(x_{ik}-x_{tk})/x_{ik}\times100\%$。根据运行该算法时所设定的阈值计算出该曲线类 C 历史上该点的负荷变化率的范围，若 $\delta(k)$ 在这个范围内，则为该点为正常数据，反之为不良数据，并且 k 点也是该不良数据的具体位置。

判定某一个时刻点的数据为不良数据后，可以根据提取出来的特征曲线进行不良数据的修正。由于从曲线集中提取出来的特征曲线不止一条，在修正之前必须正确找到对应的特征曲线。每条特征曲线就是一个质心，每个质心对应一个曲线类，只要找到离待检测负荷曲线正常数据点距离最近的质心，该质心就是所对应的特征曲线。再以该特征曲线为基准进行修正，具体的修正公式为

$$X_{\mathrm{C}}(i)=X_{\mathrm{t}}(i)\times\left[\frac{X_{\mathrm{d}}(p-1)}{X_{\mathrm{t}}(p-1)}+\frac{X_{\mathrm{d}}(q+1)}{X_{\mathrm{t}}(q+1)}\right]/2 \quad (2)$$

$$i=p, p+1, \cdots, q$$

式中：X_{d} 为待检测负荷曲线；X_{c} 为修复好的负荷曲线；X_{t} 为特征曲线；p 到 q 是 X_{d} 上的不良数据。该方法主要是利用负荷曲线的横向相似性，将特征曲线对应位置的值平移嫁接到待检测曲线上。

4 实验与算例分析

4.1 实验环境

实验平台配置为 10 个服务器节点，每个节点均为双核、4 GB 内存的 PC；其中一台作为 master，其他 9 台作为 slaves；每个节点操作系统均为 Linux Ubuntu12.04 desktop；Hadoop 版本为 2.2.0，Java 开发包为 JDK1.6 版本，Hadoop 程序使用 java 编写；Spark 版本为 1.0.2，scala 版本为 2.9.3，Spark 程序由 scala 编写。

电力负荷数据采集自SCADA系统, 由于客观原因，各信息采集单元所获取的测量数据不可避免会有不良数据，且具有偶然性、分布不确定性。验证该方法检测和辨识不良数据的实用性，算例分析数据集选用欧洲智能技术网络(European Network on Intelligent Technologies，EUNITE)组织的中期电力负荷预测竞赛提供的某地区1997、1998 年真实负荷数据[17]。以其中1997年1月至12月每天24点的实测负荷数据为研究对象, 一共365天的负荷数据, 其日负荷曲线如图4所示。

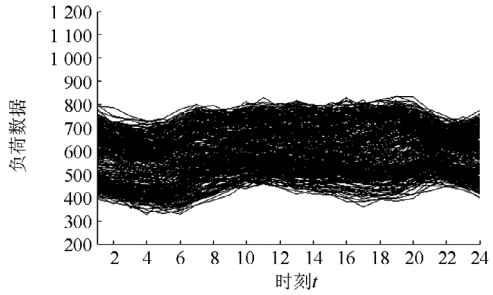

图4 日负荷曲线

Fig. 4 Daily load curve

实验分别在Hadoop和Spark集群平台上进行测试，共进行了2类实验：

(1) 基于Spark平台的状态估计结果准确性测试；

(2) 基于Hadoop和Spark平台的并行k-means算法加速比、扩展率测试。

4.2 算例分析

4.2.1 状态估计结果准确性

本实验将基于Spark的并行K-means算法与传统K-means算法[7]进行比较, 测试本文算法的状态估计准确性及收敛速度。

为了测试该方法能否对出现在同一日连续时段内的多个不良数据进行准确辨识，以上面数据集为研究对象，人为设置一些不良数据点。将3月10日的第12、13、14点原始数据652、643、638分别增加60%的误差，变为1 043.2、1 028.8、l 020.8，并对含有这三个不良数据的数据集分别用两种方法进行聚类, 这样就得到两组特征曲线。则3月10日对应的日负荷曲线与两组日负荷特征曲线分别如图5、图6所示。

其中粗线表示2月10日负荷曲线对应的特征曲线。第12、13、14点数据在两种方法下的负荷变化率分别为：{40.46%、39.09%、39.44%}、{40.01%、38.65%、39.18%}，这几个变化率均不在正常范围内，则被认定为是不良数据。应用式(2)对这些不良

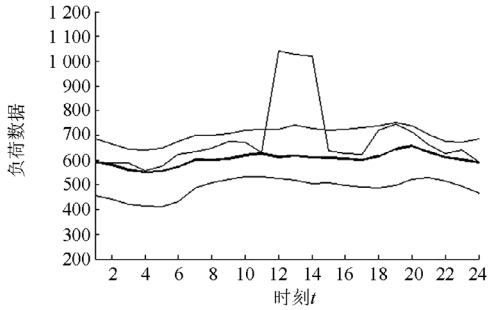

图 5 传统K-means算法下日负荷特征曲线

Fig. 5 Daily load characteristic curve with the traditional K-means algorithm

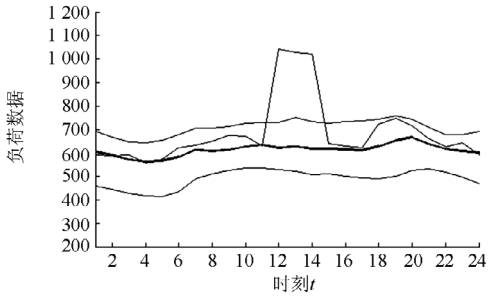

图 6 基于Spark的并行K-means算法下日负荷特征曲线

Fig. 6 Daily load characteristic curve with parallel K-means algorithm based on Spark

数据进行修正，两种方法下修正后的数据与其实际值的误差百分比及两种方法的收敛速度见表1。

表 1 基于Spark的并行K-means算法和传统K-means算法两种情况下的误差百分比及收敛速度

Table 1 Error percentage and convergence rate of two methods

算法	实际值	修正值	误差百分比%	迭代次数
传统K-means算法	652	629.12	3.51	15
	643	634.67	1.30	
	638	626.15	1.86	
基于Spark的并行K-means算法	652	633.29	2.87	7
	643	638.72	0.66	
	638	628.24	1.53	

测试两种方法的收敛速度，即各自完成聚类需要的迭代次数。由表分析可知，基于Spark的并行K-means算法下修正后的数据和实际数据更接近，误差百分比更小，迭代次数更少，状态估计结果的准确性及收敛速度优于基于传统K-means聚类的方法。因此可知，本文方法为输电网状态估计提供了相对精度高的量测值，降低了不良数据的影响，加快了收敛速度, 确保了电力系统安全运行的可靠性。

4.2.2 加速比、扩展性

加速比是指通过并行计算使运行时间减少所

获得的性能提升，它是衡量并行计算性能的一个重要指标，其计算公式为$S_d=T_s/T_d$，其中T_s表示串行算法(即在单节点上)计算所消耗的时间，T_d表示并行算法(即在d个相同节点上)计算所消耗的时间。加速比越大，表明并行计算消耗的相对时间越少，并行效率和性能提升越高。将EUNITE 提供的负荷数据样本人工扩充为原数据集的1 000倍、2 000倍、4 000倍不同大小的数据集，分别在单机环境、Hadoop和Spark云集群节点数为2、4、6、8、10的平台上运行，从而完成加速比和扩展率的对比。

由图7可知：随着云集群节点数增加和数据集增大，基于Hadoop和Spark平台的加速比越高，且基于Spark平台的加速比要优于Hadoop平台的。分析可知，当数据量足够大，单机无法处理的时候，集群并行化能有效地提高算法的计算速度。在实际应用中，尤其对于大数据集时，并行计算的效果越明显，即满足电力系统海量高维数据的负荷分类的性能需求。

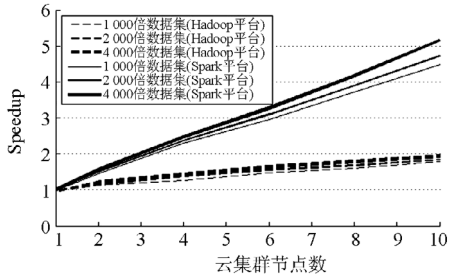

图 7 Hadoop和Spark平台下的加速比

Fig. 7 Speedup on Hadoop and Spark platform

扩展比表示并行算法执行过程中集群的利用率情况，其公式为$J=S_d/d$，其中S_d表示算法的加速比，d表示计算节点数。若可扩展比越高，则平台和并行算法的扩展性越好。

由图 8 可知，随着数据集增大，并行算法的扩展比曲线下降速率相对趋缓，且随着节点数增加整体趋于平稳。这说明在 Spark 平台下，随着数据量的增大和节点数量的增多，其扩展比逐渐趋于稳定，所以基于 Spark 的并行 K-means 算法有较好的可扩展性，能够应付电力数据规模的不断扩大，保证了程序的高可靠性。

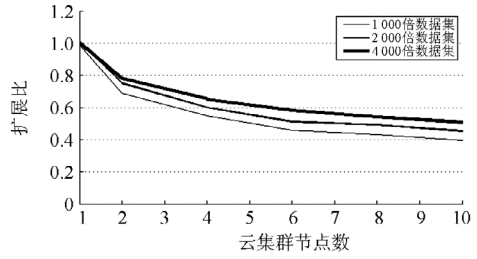

图 8 Spark环境下的扩展比

Fig. 8 Scaleup on Spark

5 结论

本文提出了基于 Spark 和聚类分析的辨识不良数据的新方法，将抽样技术和最大最小距离法引入到传统 K-means 算法中，克服了收敛速度慢且易陷入局部极小等问题；并结合 Spark 并行计算模型，解决了海量高维数据的计算量问题。通过对电力负荷数据的算例分析和实验，表明该方法效果良好，提高了电力系统状态估计结果的准确性及收敛速度，且具有更好的加速比和扩展性，满足了电力系统处理海量高维数据的需求，在保证电力系统状态估计准确性方面具有十分重要的应用价值。

参考文献

[1] 张东霞，苗新，刘丽平，等. 智能电网大数据技术发展研究[J]. 中国电机工程学报, 2015, 35(1): 2-12.
ZHANG Dongxia, MIAO Xin, LIU Liping, et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE, 2015, 35(1): 2-12.

[2] 王建华，张国钢，耿英三，等. 智能电器最新技术研究及应用发展前景[J]. 电工技术学报, 2015, 30(9): 1-11.
WANG Jianhua, ZHANG Guogang, GENG Yingsan, et al. The latest technology research and application prospects of the intelligent electrical apparatus[J]. Transactions of China Electrotechnical Society, 2015, 30(9): 1-11.

[3] 高志远，姚建国，郭昆亚，等. 智能电网对智慧城市的支撑作用研究[J]. 电力系统保护与控制，2015，43(11): 148-153.
GAO Zhiyuan, YAO Jianguo, GUO Kunya, et al. Study on the supporting role of smart grid to the construction of smart city[J]. Power System Protection and Control, 2015, 43(11): 148-153.

[4] 王韶，江卓翰. 基于奇异值分解和等效电流量测变换的电力系统状态估计[J]. 电力系统保护与控制，2012，40(12): 111-115.
WANG Shao, JIANG Zhuohan. Power system state estimation based on singular value decomposition and equivalent current measurement transformation[J]. Power System Protection and Control, 2012, 40(12): 111-115.

[5] 朱倩雯，叶林，赵永宁，等. 风电场输出功率异常数据识别与重构方法研究[J]. 电力系统保护与控制，2015，43(3): 38-45.
ZHU Qianwen, YE Lin, ZHAO Yongning, et al. Methods

for elimination and reconstruction of abnormal power data in wind farms[J]. Power System Protection and Control, 2015, 43(3): 38-45.

[6] 王兴志, 严正, 沈沉, 等. 基于在线核学习的电网不良数据检测与辨识方法[J]. 电力系统保护与控制, 2012, 40(1): 50-55.
WANG Xingzhi, YAN Zheng, SHEN Chen, et al. Power grid bad data detection and identification based on online kernel learning method[J]. Power System Protection and Control, 2012, 40(1): 50-55.

[7] 孙雅明, 王晨力, 张智晟, 等. 基于蚁群优化算法的电力系统负荷序列的聚类分析[J]. 中国电机工程学报, 2005, 25(18): 40-45.
SUN Yaming, WANG Chenli, ZHANG Zhicheng, et al. Clustering analysis of power system load series based on ANT colony optimization algorithm[J]. Proceedings of the CSEE, 2005, 25(18): 40-45.

[8] 黎祚, 周步祥, 林楠, 等. 基于模糊聚类与改进BP算法的日负荷特性曲线分类与短期负荷预测[J]. 电力系统保护与控制, 2012, 40(3): 56-60.
LI Zuo, ZHOU Buxiang, LIN Nan, et al. Classification of daily load characteristics curve and forecasting of short-term load based on fuzzy clustering and improved BP algorithm[J]. Power System Protection and Control, 2012, 40(3): 56-60.

[9] 刘莉, 王刚, 翟登辉, 等. k-means 聚类算法在负荷曲线分类中的应用[J]. 电力系统保护与控制, 2011, 39(23): 65-68.
LIU Li, WANG Gang, ZHAI Denghui, et al. k-means clustering algorithm in load curve classification[J]. Power System Protection and Control, 2011, 39(23): 65-68.

[10] 张素香, 赵丙镇, 王风雨, 等. 海量数据下的电力负荷短期预测[J]. 中国电机工程学报, 2015, 35(1): 37-42.
ZHANG Suxiang, ZHAO Bingzhen, WANG Fengyu, et al. Short-term power load forecasting based on big data[J]. Proceedings of the CSEE, 2015, 35(1): 37-42.

[11] 张逸, 林焱, 吴丹岳, 等. 电能质量监测系统研究现状及发展趋势[J]. 电力系统保护与控制, 2015, 43(2): 138-147.
ZHANG Yi, LIN Yan, WU Danyue, et al. Current status and development trend of power quality monitoring system[J]. Power System Protection and Control, 2015, 43(2): 138-147.

[12] 宋亚奇, 周国亮, 朱永利, 等. 云平台下并行总体经验模态分解局部放电信号去噪方法[J]. 电工技术学报, 2015, 30(18): 213-222.
SONG Yaqi, ZHOU Guoliang, ZHU Yongli, et al. Research on parallel ensemble empirical mode decomposition denoising method for partial discharge signals based on cloud platform[J]. Transactions of China Electrotechnical Society, 2015, 30(18): 213-222.

[13] HAN Jiawei, KAMBER M. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2000.

[14] 王丽婕, 冬雷, 高爽. 基于多位置NWP与主成分分析的风电功率短期预测[J]. 电工技术学报, 2015, 30(5): 79-84.
WANG Lijie, DONG Lei, GAO Shuang. Wind power short-term prediction based on principal component analysis of nwp of multiple locations[J]. Transactions of China Electrotechnical Society, 2015, 30(5): 79-84.

[15] 高彦杰. Spark 大数据处理技术、应用与性能优化[M]. 北京: 机械工业出版社, 2014.

[16] 李建江, 崔健, 王聃, 等. MapReduce 并行编程模型研究综述[J]. 电子学报, 2011, 39(11): 2635-2642.
LI Jianjiang, CUI Jian, WANG Dan, et al. Summary of MapReduce parallel programming model[J]. Journal of Electronics, 2011, 39(11): 2635-2642.

[17] EUNITE (Europe Network on Intelligent Technologies for Smart Adaptive Systems). World-wide competition within the EUNITE network[EB/OL]. [2001]. http://neuron.tuke.sk/competition/.

收稿日期：2015-04-05；　修回日期：2015-07-29

作者简介：

孟建良(1956-)，男，教授，硕士研究生导师，研究方向为电力信息化、人工智能及应用；

刘德超(1988-)，男，硕士研究生，研究方向为电力信息化、云计算及数据挖掘。E-mail: 568702182@qq.com

(编辑　姜新丽)