DOI:10.19595/j.cnki.1000-6753.tces.2013.s2.063

状态监测大数据存储及聚类划分研究

周国亮[1,2] 宋亚奇[1] 王桂兰[1] 朱永利[1]
（1. 华北电力大学控制与计算机工程学院 保定 071003
2. 国网冀北电力有限公司技能培训中心 保定 071051）

摘要 随着输变电设备状态监测的广度和深度不断加强，收集的监测数据越来越多，逐渐形成了智能电网状态监测大数据。然而，如何有效的存储和分析状态监测大数据是大数据在状态监测领域应用的关键问题之一。基于云计算平台并考虑状态监测数据的特点，将监测数据海量小文件组合成大的序列文件，并压缩存储，从而提高存储和处理效率。针对状态监测大数据价值密度低的特点，首先利用分形理论对监测数据降维处理，提取时域和频域特征量，并使用密度聚类算法 DBSCAN(Density-Based Spatial Clustering of Applications with Noise) 对样本数据聚类划分，提取不同聚类的特征数据；然后结合云平台的数据处理能力设计 MapReduce 并行算法，实现状态监测大数据的聚类划分，从而有助于从大数据中发现有价值的特征量；最后，利用大数据聚类结果更新和丰富样本特征数据。实验结果表明该方法可以有效存储状态监测大数据并对其聚类划分，对提高设备的状态评估及故障诊断水平具有一定辅助作用。

关键词：状态监测 大数据 云存储 聚类
中图分类号：TM769

Research of Condition Monitoring Big Data Storage and Clustering

Zhou Guoliang[1,2] *Song Yaqi*[1] *Wang Guilan*[1] *Zhu Yongli*[1]
（1. North China Electric Power University Baoding 071003 China
2. State Grid Jibei Electric Power Company Limited Skill Training Center Baoding 071051 China）

Abstract In recent years, with the power transmission equipment condition monitoring continuously strengthen in the breadth and depth, more and more monitoring data were collected, and gradually formed big data of smart grid condition monitoring. However, how to effectively store and analyze condition monitoring big data is a key issue of big data applying in power transmission equipment condition monitoring fields. Based on cloud computing platform and considering characteristics of condition monitoring data, monitoring data will be combined a large mass of small files into sequence file and compressed storage, thereby improving the efficiency of storage and processing. For condition monitoring big data low-density characteristics, firstly fractal theory was used for dimensionality reduction monitoring data, and computing time domain and frequency domain feature quantity, in addition utilizing DBSCAN algorithm clustering the sample data and acquiring different cluster feature data; then integrating with cloud computing platform parallel data processing capability designed MapReduce algorithm for clustering condition monitoring big data, and contributing found valuable feature quantity from big data; finally, using cluster result of big data updated and enriched sample data. Experimental results show that this method can efficiently store and cluster condition monitoring big data, for improving the status of device evaluation and fault diagnosis has a certain role.

Keywords：Condition monitoring, big data, cloud storage, cluster

中央高校基本科研业务费专项资金(13MS103) 和河北省高等学校科学技术研究(Z2011306) 资助项目。
收稿日期 2013-10-20 改稿日期 2013-11-25

1 引言

近年来，随着智能电网建设的不断推进及对电力系统安全稳定运行的要求越来越高，对输变电设备状态监测的广度和深度不断扩大，并逐步实现了设备的实时在线监测，同时状态监测向高采样率、连续稳态记录和大存储的趋势发展，逐渐形成了智能电网状态监测大数据。比如美国田纳西河流域管理局（Tennessee Valley Authority，TVA）每年会收集大约 15TB 的 PMU 状态监测数据，而且随着时间推移和更多 PMU 设备的加入，数据量还会进一步增加。

当前状态监测数据的存储主要采用企业级关系数据库，按分钟级准实时数据的采集速度考虑数据存储的要求，并将超过一定时限的历史数据删除；要求采集装置具有一定缓存和数据处理能力，上传的是经过加工处理的“熟数据”，而不是原始的“生数据”。这种方式主要存在三个问题：

（1）存储“熟数据”而不是“生数据”，虽然可以减少网络传输和数据库存储的数据量，但不可避免的损失“生数据”中隐藏的重要信息，不能反映真实的状态监测情况，不利于监测特征量的识别，也不利于后续的设备状态诊断及优化；

（2）关系数据库由于固有的数据处理及存储模式，可扩展性差、成本高、数据处理能力有限，面对输变电设备的状态监测大数据，关系数据库显得无能为力，更无法满足状态监测大数据分析处理的要求；

（3）删除历史数据不仅损失了历史数据中有价值的信息，而且也不能实现设备全生命周期状态监测。

基于上述分析，可以采用云平台存储状态监测大数据。但状态监测数据主要由海量小文件组成，比如绝缘子泄漏电流数据大约在几十 KB 到几百 KB 之间，而云平台最小数据分块是 64MB，云平台上海量小文件存在存储代价高和处理效率低下等问题；另外，状态监测数据中包含海量重复的平稳信号数据，常规存储会浪费海量空间，需要采用压缩存储。

大数据一般具有“3V”特征，即规模大（Volume）、类型多（Variety）和价值密度低（Value）。而价值密度低的问题在状态监测大数据中尤为突出。大部分监测数据是设备正常运行的数据，而有价值的、能有效反映设备状态的特征数据较少，并且被淹没在大数据中。利用分形理论计算监测数据的时域和频域分形维数特征量，基于特征量采用密度聚类算法（DBSCAN）对实测数据进行聚类划分，从而发现各种聚类的特征数据，形成样本。以样本数据为参考，结合云平台计算能力，对存储在云平台中的状态监测大数据设计 MapReduce 并行程序进行聚类，分析特征数据，并动态丰富和更新样本数据。从而有助于从状态监测大数据发现有价值的特征量。

2 相关工作

2.1 智能电网中的大数据

目前，大数据已成为各行各业普遍关注的问题，将来的智能电网在各个环节都会生成大数据，通过全景、实时大数据分析技术也有助于从数据分析的角度提高电力系统安全稳定运行等级[1]。

随着大数据处理技术（云计算）在互联网领域获得了广泛应用，一些学者考虑在智能电网数据处理中引入云计算技术[2-4]。文献[2]针对智能电网状态监测的特点，结合 Hadoop 云计算技术，提出了智能电网状态监测云计算平台解决方案，并探讨了云计算中的虚拟化、分布式存储与并行计算编程模型等问题，实现智能电网大数据的可靠存储与并行处理。但该文献只是提出了一个初步设想，海量后续工作还有待展开。文献[3]提出了电力系统仿真云计算中心的系统架构，包括几个层次：基础设施云、数据管理云、仿真计算云等。文献[4]探讨了未来智能电网控制中心面临的挑战，提出物联网和云计算技术结合是新型控制中心的技术支撑，但并没有考虑实际应用中的性能问题，相关研究工作有待进一步深入。

2.2 基于数据驱动的状态监测技术

近年来随着工业化与信息化的融合，海量反映系统运行过程和运行状态的数据被记录下来，分析挖掘这些海量数据有利于提高系统运行的安全性和可靠性，而基于数据驱动的分类预测、状态评估是其中的一项重要研究内容[5,6]。数据驱动技术通过对海量在线和离线数据进行分析，在无需知道系统精确模型的情况下，实现对数据分类处理，从而评估系统或设备工作状态。基于数据驱动的评价技术在电力系统中也获得了广泛应用，并取得了可喜的研究成果[7-9]。文献[7]基于电站海量历史数据信息，提出了以蒸汽热能品质的稳定性能综合得分描述燃烧的稳定程度的方法。文献[8]研究了对行波暂态高速

采集的海量录波数据通过半监督聚类技术进行划分，实现对多通道海量录波数据中故障数据集的有效筛选。文献[9]讨论了利用水轮发电机组振动故障数据，提出基于模糊聚类的分析诊断方法。基于数据驱动的状态监测技术可以为状态监测大数据分析提供有价值的参考。

2.3　基于分形理论的特征量提取技术

一般情况下，状态监测收集到的是海量时序数据，需要提取特征量，实现降维处理。而基于分形理论的分形维数特征量，可以较好的表征这类数据的特性[10,11]。分形理论描述了系统的粗糙、破碎、不规则及复杂性，在自然科学、社会科学、思维科学等各个领域作为一种新的数学概念被广泛应用于处理、分析具有复杂细节特征的自然现象，它的价值在于在有序和混沌之间提供了一种中间可能性，用来衡量一个几何形状的不规则程度。分形维数可以有效的表征时序数据的特性，并且在状态监测领域获得了成功应用。文献[10]对不同污秽度下绝缘子分形特征研究表明泄漏电流的分形维数可以预测绝缘子表面污秽的轻重。文献[11]探讨了绝缘子污秽度与分形维数之间的关系，并证明了泄漏电流分形维数的变化规律能有效预测污秽放电发展趋势及污闪的发生。而基于分形维数特征量的异常数据发现技术较少文献涉及。

3　状态监测大数据存储技术

3.1　海量小文件组合存储

在状态监测过程中，从设备传送过来的数据一般都是记录几个工频周期的小文件，大小一般在几十 KB 到几百 KB 之间。这些小文件所需存储空间远远小于分布式文件系统中定义的最小存储单元块 Block64MB 的大小。海量小文件会给云平台的扩展性和性能带来严重问题。

首先，在分布式文件系统中任何 Block、文件主要以对象的形式存储在内存中，每个对象约占 150 字节，如果有 1 千万个小文件，每个文件占用一个 Block，则在采用默认 3 份备份复制策略情况下，命名节点（NameNode）需要 3GB 内存空间。从而造成 NameNode 内存容量制约了集群的可扩展性。

其次，访问海量小文件代价远远高于访问几个大文件，分布式文件系统主要是为访问离线大文件开发的，如果访问海量小文件，需要不断的在多个数据节点之间跳转，从而严重影响性能。

最后，每个小文件需要作为一个 Block 来处理，因此每个小文件由一个任务槽（Slot）单独处理，而频繁的任务启动和释放将耗费海量时间，从而影响数据分析的效率。

针对云计算平台上大规模小文件处理问题，可以采取如下两种解决方案：设计一种打包工具，每隔一段时间将小文件打包为一个大的文件，这种方式包括 HadoopArchive、序列文件（SequenceFile）和 CombineFileInputFormat 等；从系统层面解决 HDFS 小文件，文献[12,13]提出在原有 HDFS 基础上添加一个小文件处理模块，先将海量小文件组合成一个大文件，然后为小文件建立索引，以便进行快速存取和访问。当用户上传一个文件时，判断该文件是否属于小文件，如果是，则交给小文件处理模块处理，否则，使用通用文件模块处理。

针对状态监测大数据的特点，采用将小文件组合成大的序列文件方法。序列文件是一种二进制文件，数据以<key, value>的形式存储在文件中。将每个小文件的文件名作为 key，文件内容作为 value，从而实现海量小文件的组合存储。过程大致如图 1 所示。

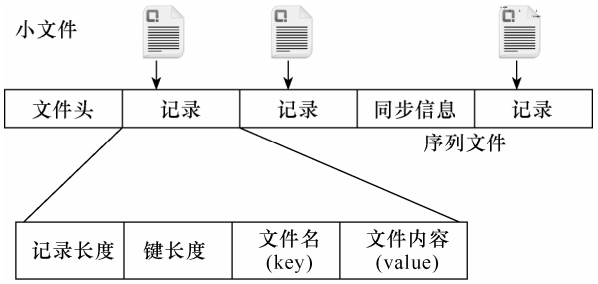

图 1　小文件组合存储为序列文件

Fig.1　Combination of small files as sequence file

3.2　序列文件压缩存储

在状态监测大数据中含有海量重复或相似度很高的监测数据，如果采用普通存储会严重浪费存储空间。因此，需要压缩存储。通过数据压缩技术可以有效减少存储文件所需的空间，提高存储效率；同时也加快了数据在网络上传输和从磁盘读取写入的速度。

目前，云平台上采用的常见压缩算法包括 LZO、GZ、bzip2 等，Hadoop 可以根据扩展名自动选择解码器解压。云平台下压缩技术对用户是透明的，通过配置文件设置好以后，MapReduce 任务在执行时能够自动将压缩文件解压。

云平台下常见压缩技术及特点见表 1。

表 1 云平台下各种存储格式及特点

Tab.1 Various compression techniques in cloud platform

压缩格式	工具	算法	扩展名	多文件	可分割性
GZ	gzip	DEFLATE	.gz	不	不
bzip2	bzip2	bzip2	.bz2	不	是
LZO	lzop	LZO	.lzo	不	是

其中，LZO 算法具有支持分块、并行处理和压缩解压速度快等特点，因此对合并后的序列文件采用 LZO 压缩算法进行压缩。

4 基于分形维数的密度聚类划分

4.1 基于时域特征和频域特征的分形维数

分形学理论是由波兰科学家在 1967 年提出的[14]，用来研究整体与部分的相似度，起初主要应用在集合拓扑领域，后来普遍应用到自然科学和工程领域，比如刻画路面的不平度[15]、机械设备振动情况[16]等。分形理论的另一个优点是具有较强的抗噪能力，对噪声数据不敏感。实际环境中收集的状态监测数据由于工作环境的复杂性经常包含各种噪声数据，而分形理论可以有效的处理此类数据。分形维数是分形理论中的主要参数，他可以定量描述分形集的复杂性。

分形维数可以采用盒计数法计算，但盒计数法计算复杂度高，当应用于海量时间序列数据时，计算量庞大。因此，文献[17]给出了对于数字化离散时间序列数据的分形维数计算公式。设信号的时间序列为：$x(t_1)$、$x(t_2)$、…、$x(t_N)$、$x(t_{N+1})$，其中 N 为偶数。令

$$d(\Delta)=\sum_{i=1}^{N}\left|x(t_i)-x(t_{i+1})\right| \tag{1}$$

$$d(2\Delta)=\sum_{i=1}^{N/2}\begin{pmatrix}\max\left\{x(t_{2i-1}),x(t_{2i}),x(t_{2i+1})\right\}\\-\min\left\{x(t_{2i-1}),x(t_{2i}),x(t_{2i+1})\right\}\end{pmatrix} \tag{2}$$

因此，通过盒计数法计算得到的时域分形维数 D_t 可以表示为：

$$D_t=1+\log_2\frac{d(\Delta)}{d(2\Delta)} \tag{3}$$

基于时间序列（时域）的分形维数表征了序列的波动程度，也就是序列中任意一段与序列整体之间的相似度；分形维数越高，表明监测数据具有较大波动，反之表明数据比较平稳。

状态监测数据在频域中也包含了丰富的信息，比如泄漏电流数据三次和五次谐波在污闪前会迅速升高[18]。而基于功率谱密度的分形维数值表征了状态监测数据的频域特征。功率普分形维数越大，表明信号的波动大，信号中相邻点之间的相关性弱，信号包含的高频成份多。

在计算功率普估计中，AR 模型具有良好的特性，其计算公式如下：

$$P_x\left(\mathrm{e}^{\mathrm{j}w}\right)=\sigma^2/\left|1+\sum_{k=l}^{p}a^k\,\mathrm{e}^{-\mathrm{j}\omega k}\right|^2 \tag{4}$$

式中，σ^2 是激励白噪声的方差；$P_x\left(\mathrm{e}^{\mathrm{j}w}\right)$为功率谱密度；$a^k$为模型参数。

首先，AR 模型功率谱估计需要通过 levinson_dubin 递推算法由 Yule-Walker 方程求得 AR 的参数: σ^2、a^1、a^2、…、a^k。

然后，在双对数功率谱 $\log p(\omega)-\log\omega$ 图上描述各个点，利用最小二乘法拟合直线，设所拟合的直线斜率为 K，则基于现代功率普估计的频域分形维数 D_h 可以表示为

$$D_h=(5+K)/2 \tag{5}$$

4.2 分形维数的计算过程及聚类效果

选取在某下雨天监测的一段状态监测数据（泄漏电流）为例计算其时域和频域的分形维数，采集数据的时间为 2012 年 7 月 4 日晚上。采集数据的波形图如下图 2 所示。

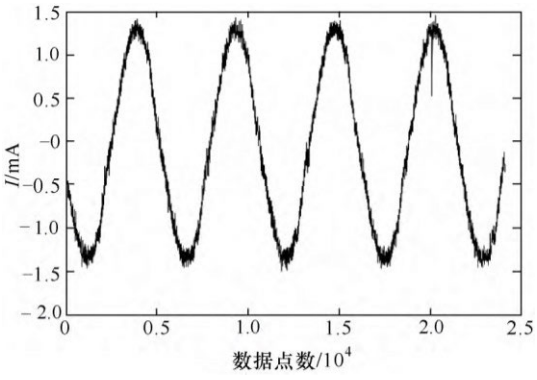

图 2 实测绝缘子泄漏电流波形数据

Fig.2 Measured insulator leakage current waveform data

依据式（1）～式（3），可以计算得到基于盒计数法的时域特征分形维数：D_t=1.332 2。

计算频域分形维数的过程如下，首先计算得到功率谱估计，结果如图 3 所示。根据功率谱估计，得到双对数功率谱 $\log p(\omega)-\log\omega$ 曲线（图 4）。

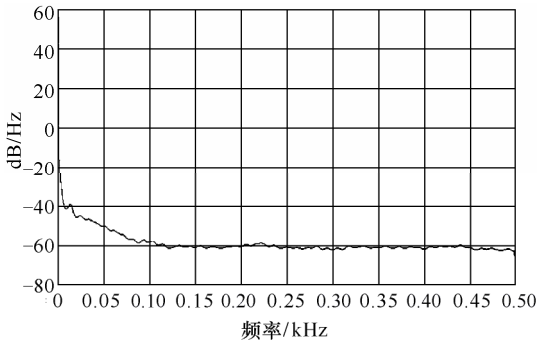

图 3　现代功率谱估计曲线

Fig.3　Modern power spectrum estimation curve

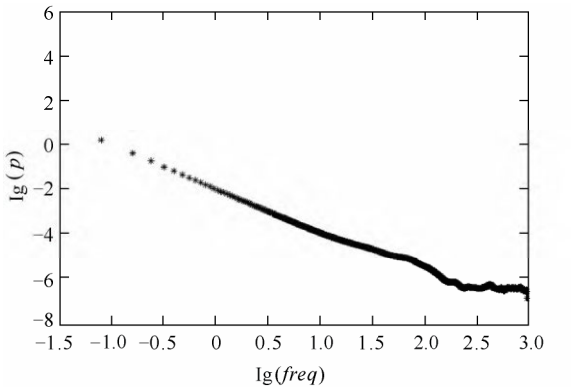

图 4　双对数曲线图

Fig.4　Double logarithmic graph

利用最小二乘法拟合曲线的斜率为–3.015 6，进而得到功率谱估计的频域特征分形维数：D_h=(5–3.015 6)/2=0.992 2。

通过对不同天气状况下收集的实测数据计算其时域和频域分形维数，结果如下表 1 所示，这里主要列举了三种常见天气情况的计算值。

表 2　实测绝缘子泄漏电流数据的分形维数值

Tab.2 Fractal dimension value of measured insulator leakage current data

	D_t	D_h	Others
雨天	1.332 2	0.992 2	…
	1.339 4	1.008 4	…
	1.332 3	1.078 0	…
闷热	1.428 9	0.287 6	…
	1.424 9	0.232 7	…
	1.416 2	0.311 3	…
凉快	1.408 0	0.081 2	…
	1.426 9	0.024 7	…
	1.406 0	0.071 3	…

根据计算得到的特征量，以 D_t 为横轴，D_h 为纵轴绘制的散点图如图 5 所示。通过观察图形，可以发现不同天气情况下的数据具有很强的内聚性，形成不同天气情况下的聚类。通过上图大致可以发现实测状态监测数据的聚类情况，然而当面对海量数据时，需要借助聚类算法来有效的处理数据，而不是直观的观察。关于监测数据聚类的划分应具有如下原则：由于数据在空间中的几何形状是任意的，因此聚类算法要求可以发现任意形状的聚类；异常数据不影响聚类划分的效果，聚类算法可以处理噪声数据和较强的抗干扰能力。基于以上两点，选择基于密度的聚类算法 DBSCAN。

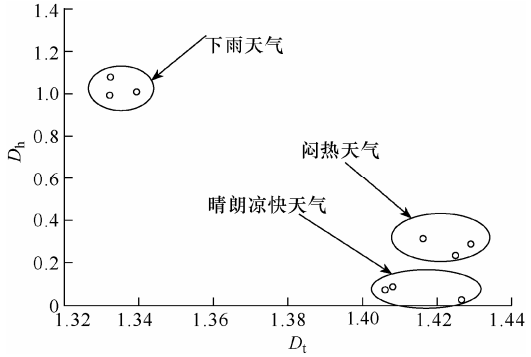

图 5　散点图

Fig.5　Scatter graph

4.3　密度聚类算法 DBSCAN

DBSCAN 是一个比较有代表性的基于密度的聚类算法。它将聚类定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为聚类，并可在有噪声的空间数据库中发现任意形状的聚类[19]，具有较强的抗干扰能力和较少人工干预的优势。DBSCAN 算法在电力系统中也获得了应用，文献[20]研究了基于变密度聚类的居民阶梯分段电量制定方法。

DBSCAN 算法计算过程大致如下：从数据集 D 中选取任意点 p，并查找 D 中满足 Eps 范围内，所有密度可达的点。如果点的个数＞$MinPts$，则 p 是核心点，则根据该算法可以找到一个关于参数 Eps 和 $MinPts$ 的聚类。如果 p 是一个边界点，即 p 的半径为 Eps 的邻域中包含的对象数＜$MinPts$，则没有对象从 p 密度可达，p 被暂时地标注为异常点。然后以同样的方式处理数据集 D 中的下一个未标注的点，直到所有数据点处理完毕。对收集的每段绝缘子状态监测数据，构建特征量：$x=[D_t,D_h,\cdots]$，从而形成数据集 D。为了提高区分度，可以进一步增加其他特征量。

在 DBSCAN 算法中需要确定参数 $MinPts$ 和 Eps，对于 $MinPts$ 可以采用经验估计法，如果点数

较少，可以取 3，如果点数足够多，可以取一个较大的值。

对 *Eps* 可以采用如下公式预估计[19]

$$Eps=\left(\frac{\prod_{i=1}^{n}(\max(x_i)-\min(x_i))\times minPts\times\prod(0.5\times n+1)}{m\times\sqrt{\pi^n}}\right)^{\frac{1}{n}} \tag{6}$$

式中，m 表示特征向量的长度；n 表示数据集的大小。

通过利用分形维数特征量和密度聚类算法，可以有效实现对状态监测数据的划分，从而有助于对数据分类处理，也有利于发现异常特征数据。表 3 记录了部分实测数据的密度聚类划分结果，其中 ClusterID 表示数据点属于哪个聚类，–1 表示不属于任何聚类的异常点；Type 表示是核心点、边界点和异常点，1 表示核心点，0 表示边界点，–1 表示异常点。

表 3 实测绝缘子泄漏电流数据的聚类结果

Tab.3 Clustering results of measured insulator leakage current data

D_t	D_h	ClusterID	Type
1.332 2	0.992 2	1	1
1.339 4	1.008 4	1	1
1.332 3	1.078 0	1	1
1.428 9	0.287 6	2	1
1.424 9	0.232 7	2	1
1.416 2	0.311 3	2	1
1.408	0.081 2	3	1
1.426 9	0.024 7	3	1
1.406	0.071 3	3	1
1.500 0	1.400 0	–1	–1
1.336 5	1.002 3	1	1
1.347 8	0.996 8	1	1
1.327 6	1.002	1	1
1.359 2	0.998 7	1	1
1.330 0	1.010 0	1	1
1.425 0	0.256	2	1
1.400 0	0.300 1	2	1
1.445	0.298	2	1
1.409	0.229 9	2	1
1.418	0.071 2	2	1
1.436 9	0.034 7	3	1
1.446 0	0.061 3	3	1
1.480 0	1.220 0	–1	–1
1.460 0	1.330 0	–1	–1

5 大数据聚类算法及仿真试验

5.1 状态监测大数据的聚类划分算法

对于小规模数据集，DBSCAN 算法可以有效的完成聚类划分，但该算法具有较高的时间复杂度 $O(n^2)$，即使采用空间索引技术（*R* 树等）时间复杂度仍为 $O(n*\log(n))$。这样当使用 DBSCAN 算法处理状态监测大数据时，性能很难保障；并且，将该算法用 MapReduce 技术重写也具有较高的难度。

因此，可以借鉴半监督聚类算法思想，同时结合云平台的计算能力和 DBSCAN 算法的有效性，设计基于 MapReduce 的并行算法，利用已知样本数据的聚类情况，对未知的大数据样本进行划分，从而提高聚类算法的执行效率。算法基本流程如下：

（1）将已标记好的聚类样本数据分发到云平台的各个节点，此步骤可以利用 Hadoop 的分布式内存技术实现；

（2）每个节点执行 Map 操作计算每个监测数据条目的分形维数特征量，为了避免多次读取数据，一次计算多个特征量；

（3）根据计算生成的特征量，与样本中数据进行比较，如果与某一节点距离小于 *Eps*，则该数据条目标记为相同的聚类，如果节点中数据与样本中所有数据距离大于 *Eps*，则标记为异常数据；

（4）对所有产生的异常点数据进行分析过滤，这些数据中可能含有对状态监测有较高价值的特征数据；并应用 DBSCAN 算法进行聚类划分，分析结果；根据聚类结果更新样本数据。

（5）随着监测数据的增加，重复步骤（1）～（4）。其中步骤（1）～（3）以增量计算的方式完成。

算法的执行过程，如下图 6 所示：

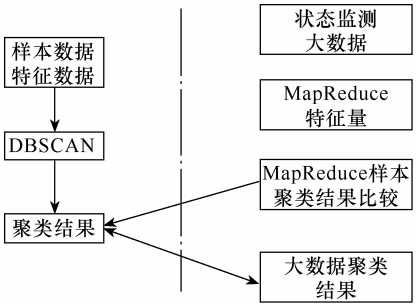

图 6 MapReduce 密度聚类算法

Fig.6 MapReduce based density cluster algorithm

5.2 实验环境及测试结果

实验室在输变电设备状态监测领域收集了丰富的实测数据，并搭建了包含 6 个节点的 Hadoop 测

试云平台，这为开展状态监测大数据存储和处理工作提供了技术和数据保障。以实验室在真实环境中采集的监测数据为样本数据，并以样本数据为基础，生成模拟的状态监测大数据，分别测试数据的存储效率和聚类划分效果，实验效果证明了本方法的可以实现大数据的聚类划分，从而有助于从大数据中发现有价值的特征量。

6　结论

随着大数据在电力系统中的逐渐形成，利用大数据处理分析技术并结合电力系统特点，有助于从数据分析方面解决电力系统中面临的某些问题，比如新能源的布局和接入等。本文初步探讨了在状态监测领域大数据的存储及处理技术，并借助大数据提高聚类划分的效果，丰富状态特征数据。随着状态监测大数据的不断增长，应用领域逐渐扩大，基于大数据分析的设备状态评估、故障预测也将成为可能，从而有利于提高现有电力系统状态监测的水平。

参考文献

[1] 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.

[2] 王德文，宋亚奇，朱永利. 基于云计算的智能电网信息平台[J]. 电力系统自动化, 2010(22): 7-12.
Wang Dewen, Song Yaqi, Zhu Yongli. Information platform of smart grid based on cloud computing [J]. Automation of Electric Power Systems, 2010(22): 7-12.

[3] 沐连顺，崔立忠，安宁. 电力系统云计算中心的研究与实践[J]. 电网技术, 2011, 35 (6): 170-175.
Mu Lianshun, Cui Lizhong, An Ning. Research and practice of cloud computing center for power system[J]. Power System Technology, 2011, 35(6): 170-175.

[4] 王广辉，李保卫，胡泽春，等. 未来智能电网控制中心面临的挑战和形态演变[J]. 电网技术, 2011, 35 (8): 1-5.
Wang Guanghui, Li Baowei, Hu Zechun, et al. Challenges and future evolution of control center under smart grid environment. Power System Technology, 2011, 35(8): 1-5.

[5] 李晗，萧德云. 基于数据驱动的故障诊断方法综述[J]. 控制与决策, 2011, 26(1): 1-16.
Li Han, Xiao Deyun. Survey on data driven fault diagnosis methods[J]. Control and Decision, 2011, 26(1): 1-16.

[6] 王燕，申元霞，陶春梅. 面向领域的数据驱动自主式知识获取模型及实现[J]. 重庆邮电大学学报：自然科学版, 2009, 21(4): 502-506.
Wang Yan, Shen Yuanxia, Tao Chunmei. Domain oriented data-driven knowledge acquisition model and its implementation[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2009, 21(4): 502-506.

[7] 刘吉臻，杨光军，谭文，等. 基于数据驱动的电站燃烧稳定度综合评价[J]. 中国电机工程学报，2007, 27(35): 1-6.
Liu Jizhen, Yang Guangjun, Tan Wen, et al. Synthetic evaluation on the degree of combustion stability in power station based on data-driven[J]. Proceedings of the CSEE, 2007, 27(35): 1-6.

[8] 张广斌，束洪春，于继来. 利用广义电流模量的行波实测数据半监督聚类筛选[J]. 中国电机工程学报, 2012, 32(10): 150-159.
Zhang Guangbin, Shu Hongchun, Yu Jilai. Travelling Wave Field Data Contingency Screening Based on Semi-supervised Clustering Using Generalized Current Modal Components. Proceedings of the CSEE, 2012, 32(10): 150-159.

[9] 陈铁华，陈启卷. 模糊聚类分析在水电机组故障诊断中的应用[J]. 中国电机工程学报，2002，22(3): 43-47.
Chen Tiehua, Chen Qijuan. Fuzzy clustering analysis based vibration fault diagnosis of hydroelectric generating unit[J]. Proceedings of the CSEE, 2002, 22(3): 43-47.

[10] 陈伟根，夏青，罗兵，等. 用于绝缘子污秽度预测的泄漏电流分形特征[J]. 高电压技术，2011，37(5): 1136-1141.
Chen Weigen, Xia Qing, Luo Bing, et al. Fractal characteristic of leakage current for insulators contamination degree prediction[J]. High Voltage Engineering, 2011, 37(5): 1136-1141.

[11] 陈伟根，夏青，孙才新，等. 绝缘子放电区段划分及污秽预测的泄漏电流分形维数研究[J]. 中国电机工程学报, 2011, 31(13): 121-127.
Chen Weigen, Xia Qing, Sun Caixin, et al. Research on fractal dimension of leakage current for discharge zones dividing and contamination forecasting of insulators[J]. Proceedings of the CSEE, 2011, 31(13): 121-127.

[12] Xuhui Liu, Jizhong Han, Yunqin Zhong. Implementing WebGIS on Hadoop: A case study of improving small file I/O performance on HDFS[J]. Cluster, 2009: 1-8.

[13] Bo Dong, Jie Qiu, Qinghua Zheng, et al. A Novel Approach to Improving the Efficiency of Storing and Accessing Small Files on Hadoop: A Case Study by PowerPoint Files[C]. In Proceedings of IEEE SCC' 2010: 65~72.

[14] Mandelbrot B B. The fractal geometry of nature[M]. San Francisco: Freeman, 1998.

[15] 关凯书，赵虹，王志文. 路表不平的分形特征[J]. 农业机械学报, 2000, 31(6): 21-24.
Guan Kaishu, Zhao Hong, Wang Zhiwen, Fractal Behavior of a Rugged Road Surface[J]. Transactions of the Chinese Society for Agricultural Machinery, 2000, 31(6): 21-24.

[16] 吴振升，王玮，杨学昌，等. 基于分形理论的高压断路器机械振动信号处理[J]. 高电压技术，2005, 31(6): 19-21.
Wu Zhengsheng, Wang Wei, Yang Xuechang, et al. Processing of Mechanical Vibration Signals of High-Voltage Circuit Breakers Based on Fractal Theory[J]. High Voltage Engineering. 2005, 31(6): 19-21.

[17] 吕铁军，郭双兵，肖先赐. 调制信号的分形特征研究[J]. 中国科学 E 辑, 2001, 31(6): 508-513.
Lv Tiejun, Guo Shuangbing, Xiao Xianci. Research of fractal features of the modulated signal. Science in China, Ser. E, 2001, 31(6): 508-513.

[18] 姚陈果，李璟延，米彦，等. 绝缘子安全区泄漏电流频谱特征提取及污秽状态预测[J]. 中国电机工程学报, 2007, 27(30): 1-8.
Yao Chenguo, Li Jingyan, Mi Yan, et al. Abstracting Frequency Spectrum Characteristics of Insulators Leakage Current in Safety Zone to Forecast the Contamination Condition [J]. Proceedings of the CSEE, 2007, 27 (30): 1-8.

[19] M. Ester, H. -P. Kriegel, J. Sander, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Data Mining and Knowledge Discovery, 1996: 226-231.

[20] 黄海涛，张粒子，乔慧婷，等. 基于变密度聚类的居民阶梯分段电量制定方法[J]. 电网技术，2010, 34(11): 111-116.
Huang Haitao, Zhang Lizi, Qiao Huiting, et al. A Method to Determine Step-Shaped Electricity Consumption Levels for Residential Area Based on Variable-Density Clustering[J]. Power System Technology, 2010, 34(11): 111-116.

作者简介

周国亮 男，1978 年生，博士，副教授，主要研究方向为云计算和电力大数据处理分析技术。

宋亚奇 男，1979 年生，博士研究生，主要研究方向为电力信息智能处理和云计算。