一种电力感知数据的离群点检测方案

李 寒[1,2],余 斌[1,2],佟 宁[3],王鑫浩[1,2]
(1. 北方工业大学 计算机学院,北京 100144;
2. 大规模流数据集成与分析技术北京市重点实验室,北京 100144;
3. 大连交通大学 软件学院,辽宁 大连 116052)

摘 要:鉴于离群点引发的数据质量问题给电力应用造成的不良影响,对电力感知数据的特征进行了分析,并基于电力感知数据的时间特征和异常检测技术的易用性需求,提出一种电力感知数据的离群点检测方案。该方案由异常检测服务框架和离群点检测方法构成。异常检测服务框架借鉴 Web 服务的思想,基于大数据技术,能够支持电力感知数据的存储和计算,并且以服务的形式提供电力感知数据的异常检测能力。离群点检测方法是基于聚类算法和考虑时间属性的数据分段方法来检测电力感知数据中的离群点异常。通过实验验证了该方法的可行性和有效性,结果表明该方法能够有效识别具有时间相关性和连续性的电力感知数据中存在的离群点,且在数据规模增大时,具有良好的并行性和可扩展性。

关键词:电力感知数据;离群点检测;聚类;数据分类;服务
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2020)02-0153-06
doi:10.3969/j.issn.1673-629X.2020.02.030

An Electric Power Sensor Data Oriented Outlier Detection Solution

LI Han[1,2], YU Bin[1,2], TONG Ning[3], WANG Xin-hao[1,2]
(1. School of Computer Science, North China University of Technology, Beijing 100144, China;
2. Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing 100144, China;
3. School of Software, Dalian Jiaotong University, Dalian 116052, China)

Abstract: In view of the adverse effects of data quality problems caused by outliers on power applications, the characteristics of power sensor data are analyzed. Based on the temporal characteristics of power sensor data and the usability of anomaly detection technology, an electric power sensor data oriented outlier detection solution is proposed, which consists of an anomaly detection service framework and an outlier detection method. The anomaly detection service framework refers to the idea of Web service, and based on big data technology it can support the storage and calculation of power sensing data, and provide anomaly detection capability of power sensing data in the form of service. The outlier detection method is accomplished on the basis of clustering algorithm and a temporal characteristics related data segmentation method to detect outlier anomalies in power perception data. The feasibility and effectiveness of the proposed method are verified by experiment. The results show that this method can effectively identify outliers in power sensing data which are time-related and time-continuous, and has great parallelism and scalability when the data scale increases.

Key words: electric power sensor data; outlier detection; clustering; data classification; service

0 引 言

随着电力相关物联网技术的发展,大量反映电网实际运行状况的电力感知数据持续产生且不断积累。基于这些电力感知数据,将有机会提供更精确、更智能及更综合的电力服务。所以如何利用好电力感知数据已成为电力工程领域一个新兴且关键的问题[1]。然而,由于干扰源的影响和数据采集及网络传输异常的存在,电力感知数据的质量很难保证[1]。因此,检测并消除数据源中的异常在学术界和工业界受到普遍关注。

异常数据种类多样,离群点是一类典型且主要的数据异常[2]。以电力为代表的工业控制领域大多通过设置阈值发现离群点。尽管这种基于阈值的方法简便易行,但却不能发现未超出阈值的异常数据。文中将离群点定义为超出阈值,或未超出阈值但与相邻数据

收稿日期:2019-02-28 **修回日期**:2019-06-28 **网络出版时间**:2019-09-25
基金项目:北京市教育委员会科技计划一般项目(SQKM201810009004);国家自然科学基金(61702014)
作者简介:李 寒(1981-),女,博士,讲师,CCF 会员(44705M),研究方向为大数据分析、数据质量管理。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1520.020.html

存在明显差异的数据。近年来,虽然提出了一些离群点检测方法,但存在方法难以应用于大规模数据或不适合电力感知数据的情况。离群点检测是运用各类数据处理模型和技术发现数据资源中的异常数据的过程,是发现数据异常,提升数据质量的前提和必要环节。离群点检测本身是一种能力,如果一个组织将其离群点检测能力提供给其他组织或个人,这就是离群点检测服务。在大数据环境下,数据资源不仅体量巨大而且种类繁多,对数据处理能力的要求也在不断提高。在这种情况下,只有少部分机构具备独立处理大数据的能力,对于不具备上述能力的机构,则需要对外寻求数据处理服务。因此,同大数据资源一样,离群点检测成为一类必不可少的数据处理服务。针对发现异常电力感知数据和以服务形式提供异常数据检测能力的需求,文中提出一种电力感知数据的离群点检测解决方案,包含异常检测服务框架和离群点检测方法。

1 相关工作

近年来,为提高数据质量,提出一系列数据异常处理技术,如缺失数据填补、对象重复检测、离群点检测、逻辑错误检测和不一致数据检测等[3]。由于离群点会对后续数据处理和分析带来严重负面影响,因此离群点检测被认为是数据质量保障环节最重要的问题之一[2]。

1887 年,F. Edgeworth 发表了关于不一致实验数据的研究成果,从此开启了离群点检测研究的序幕[3]。通常,离群点可划分为五类,分别是基于统计的离群点[4]、基于聚类的离群点[5]、基于分类的离群点[5]、基于距离的离群点[6]和基于密度的离群点[7]。近年来,随着数据资源的重要性提升,针对离群点检测的研究开始增多。2010 年,江峰等提出一种基于边界和距离的离群点检测方法[8]。该方法针对不确定和不完整数据,基于粗糙集理论和基于聚类的离群点检测方法实现检测。在临床诊断数据集上的实验验证了该方法的有效性,但该方法在其他领域的应用效果尚有待验证。2012 年,Z. Yao 等提出一种基于临近图和 PageRank 算法的离群点检测方法[9]。该方法使用离群分数标记数据的离群程度,具有较低的时间复杂度,对高维数据的离群点检测效果较好。2015 年,G. Tang 等提出一种多维情景的离群点检测方法[10]。该方法首先对数据分类,并将类别作为数据领域,再利用群闭包理论检测情景离群点,是条件离群点检测方面的新尝试。

此外,一些针对大规模数据的离群点检测方法也开始受到关注。2015 年,Y. Diao 等提出一种基于动态离群点检测的大数据在线清洗算法[11]。该算法基于 Hadoop 平台实现,能大幅提高实时数据预处理的效率,但未与领域特征相结合。2016 年,王习特等提出一种高效的分布式离群点检测算法(BOD)[12]。该算法将数据分块处理,通过均衡化每个节点的工作负载,能有效提升离群点检测的效率并控制网络开销。

由于领域数据具有特殊数据特性,有必要开展领域相关的离群点检测研究。在电力领域,结合领域知识的离群点检测方法还十分有限。2015 年,程超等提出一种基于离群点算法和用电信息采集系统反窃电研究。该研究将离群点算法与电力应用相结合,探索了离群点与电力业务之间的相关性。然而,该研究侧重窃电分析,离群点算法仅是窃电分析的一个操作步骤[13]。

为更便捷地对外提供大数据处理能力,数据相关服务也开始受到关注。2014 年,张志强等研究了数据可视化服务,提出一种基于 B/S 架构的雾霾专题数据可视化服务系统[14],能够支持雾霾数据的实时更新、统计、显示功能。2017 年,夏虹等探讨了面向工业的开放数据服务平台[15]。提出了一种面向工业的开发数据服务平台的体系结构和工作流程,但未阐述具体技术和方法。2018 年,佟杰等对海洋测绘数据服务保障系统展开研究[16],尝试将海洋测绘数据以服务的形式对外发布。该研究尚处于起步阶段,有待深入探讨和应用验证。总之,目前关于数据处理服务的研究还十分有限,且不存在针对异常数据检测服务的相关研究。

综上所述,现有的离群点检测方法大多将注意力集中在数据值上,而忽略了领域相关且能反映数据特征的数据属性。此外,共享面向大数据的异常数据检测能力的需求已存在。因此,文中提出了一种考虑电力感知数据的时间属性的离群点检测方法,并设计了一种电力感知数据异常检测服务框架。

2 电力感知数据的异常检测服务框架

绝大多数电力数据是由各种电力传感器产生的。这些电力传感器属于不同的电力设备,分布广泛。在中国,电网规模很大,并且正在逐年扩大。随着电网规模的扩大,电力感知数据量迅速增加。然而,并非所有机构都具备大规模的电力感知数据的处理能力。因此,为了有效和方便地对外提供电力感知数据异常检测的能力,有必要探讨电力感知数据的异常检测能力的使用模式。

借鉴 Web 服务的思想,为了满足易用的特点,并同时考虑可扩展性,文中融合 Web 服务的思想和大数据技术,设计了一种电力感知数据的异常检测服务框架,如图 1 所示。在该框架中,处理对象是由各类电力传感设备产生的电力感知数据,输出为带有异常数据标记的电力感知数据。该框架由四个主要层次构成,自顶向下包括应用层、服务层、计算层和存储。

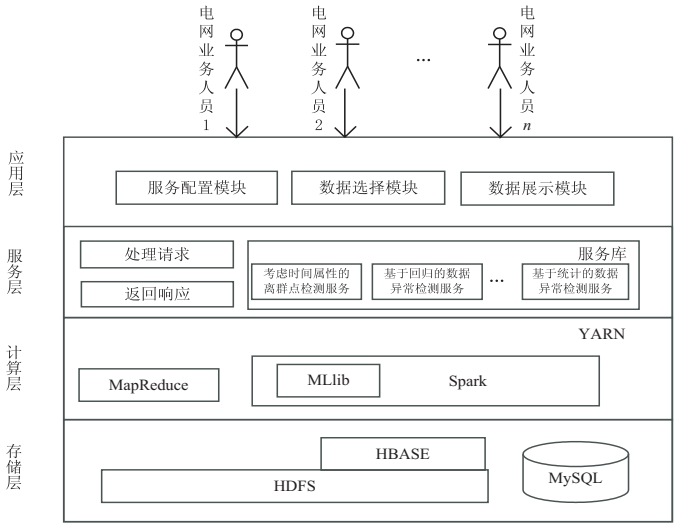

图 1　电力感知数据的异常检测服务框架

框架各层次描述如下:

(1)应用层。

应用层是用户与系统直接交互的窗口,也是整个系统核心功能的入口。电网业务人员在服务配置模块中可选择所需的异常检测服务并对其相关参数进行配置。数据选择模块支持待检测的数据的选择,这些数据存储在存储层中,用户选择后就会将请求发往服务层进行进一步的异常检测。数据展示模块则负责返回部分异常数据检测结果,供用户查看。

(2)服务层。

服务层用于托管应用服务。首先,服务层会接收来自应用层的请求,处理请求中的配置和数据。然后,进入服务库,使用相对应的异常数据检测服务在计算层对数据实施异常检测处理。文中提出的考虑时间属性的离群点检测方法将作为服务在服务库中提供。最后,部分异常数据检测的结果将打包返回给应用层的数据展示模块,方便用户定位异常数据。

(3)计算层。

计算层是一个混合的计算环境,用于执行异常数据检测方法。基于 YARN,MapReduce 和 Spark 是主要的分布式计算组件。其中,MapReduce 在批处理中具有良好的性能,Spark 则提供快速的内存处理,且 Spark 的机器学习库 MLlib 包含许多算法和实用工具,如分类、决策树、推荐、聚类等,能够有效支撑异常数据检测方法的实现。文中提出的离群点检测方法使用了 MLlib 库的聚类算法。

(4)数据层。

数据层位于最底层,该层采用一个集成的存储环境保存数据,包括关系数据库(MySQL)、NoSQL 数据库(Hbase)和分布式文件系统(HDFS)。MySQL 用于保存计算结果和从原始电力感知数据中解析获得的所有结构化数据。HDFS 用于电力感知数据的保存。Hbase 则以电力感知数据的时间和空间属性为依据保存电力感知数据。

3　电力感知数据的离群点检测方法

3.1　电力感知数据的特征

电力感知数据是持续产生的,影响电力感知数据的因素通常不具有突变性,因此,电力感知数据具有明显的时间相关性和连续性。当影响因素发生突变时,即可能产生离群点,如异常断电、设备故障等。这些离群点不由时间因素引起,但却与正常数据同样具有时间属性。图 2 为光伏电能质量数据中的无功功率随着时间的变化趋势,其中离群点用圆形圈出。这些离群值不超过阈值,但每个离群点和它的邻居点之间存在明显的偏差。根据检查记录,这些异常值是由传输异

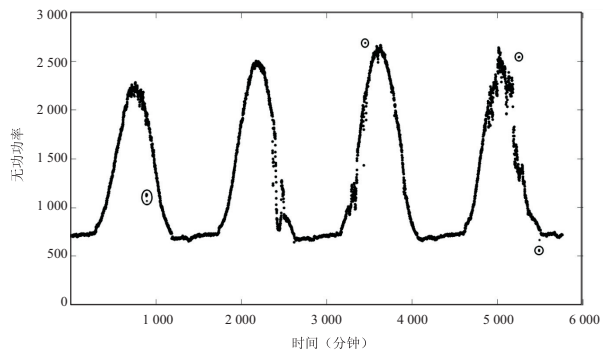

图 2　无功功率随时间变化的趋势

常或特殊的环境因素引起的。因此,考虑到电力感知数据的时间相关性,提出了一种考虑时间属性的离群点检测方法,用于发现远离邻近点的离群点。

3.2　考虑时间属性的离群点检测方法

由于电力感知数据具有明显的时间相关性和连续性特征,且影响因素突变较少,数据常呈规律性变化且邻近数据之间的偏差不大。鉴于现有电力数据聚类及分类算法未考虑数据的时间特性的不足,文中将时间

属性引入离群点发现,提出一种考虑时间属性的离群点检测方法。该方法先采用基于数据值的 k-means 聚类获取数据值中心,再利用基于时间属性的数据分段识别违背时间连续性和相关性的电力感知数据离群点。

3.2.1 考虑时间属性的数据分类

考虑时间属性的数据分类由 k-means 聚类和数据分段两个阶段构成,其流程如图 3 所示。

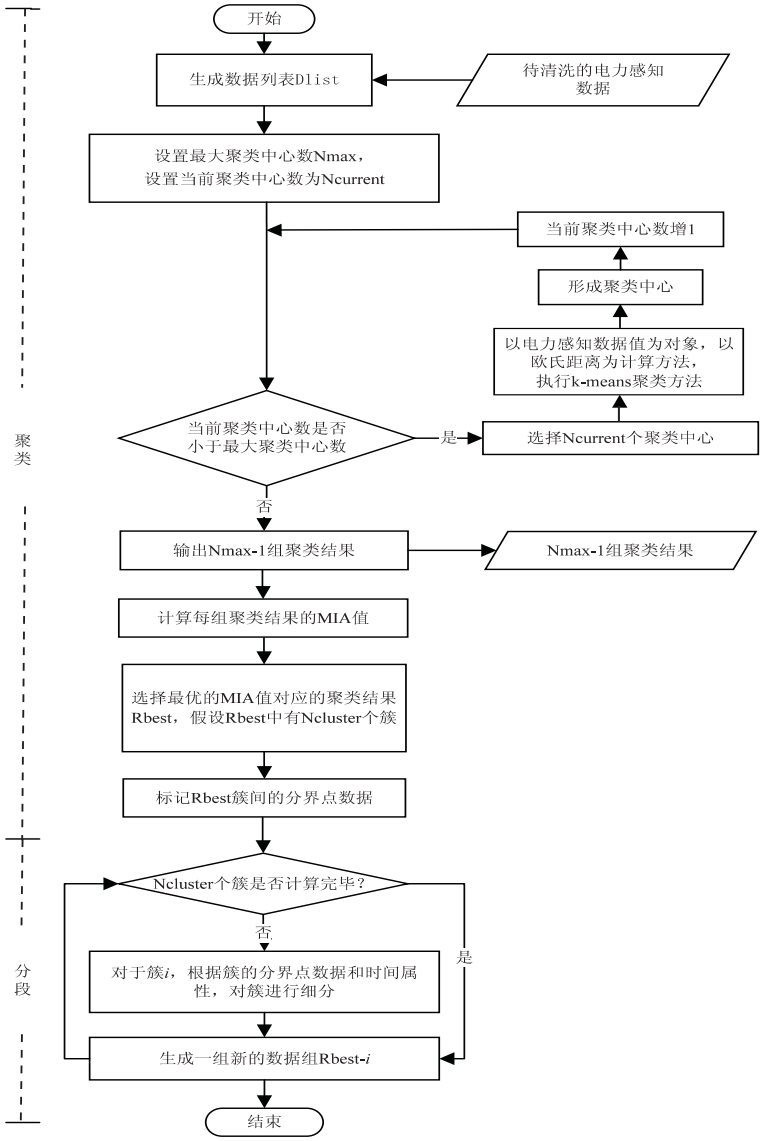

图 3 考虑时间属性的数据分类流程

在第一阶段,由于时间属性是均匀分布且连续的,不具有聚类条件,k-means 聚类以数据值为聚类对象,将生成若干组具有多个数值聚类中心的簇,并标记簇之间的分界点数据。数据分类的输入数据是 k-means 聚类结果中的最佳者。

如图 3 所示,由于 k-means 聚类的聚类个数需要预先确定,提出的方法将依据设置的最大聚类个数 $N_{\max}$,从 2 个聚类个数开始执行 $N_{\max}-1$ 次聚类。为了从 $N_{\max}-1$ 个聚类结果中选取最佳者,采用 MIA 指数(mean index adequacy)评估聚类结果的质量[17]。MIA 被描述为每个簇中心和属于相应簇的所有元素之间的平均距离。MIA 值越小表明簇内元素的紧密度越高,聚类结果越好。MIA 指数计算方法如式 1 和式 2 所示。

$$D_{\mathrm{MIA}}=\sqrt{\frac{1}{k}\sum_{k=1}^{k}d^2(x_c,x_k)} \tag{1}$$

$$d(x_c,x_k)=\sqrt{\frac{1}{n_k}\sum_{n=1}^{n_k}d^2(x_c,C_k^n)} \tag{2}$$

其中,k 表示簇的数目,C_k 表示第 k 簇,C_k^n 表示簇 C_k 的第 n 个元素,n_k 表示 C_k 的元素数,x_c 表示 C_k 的聚类中心,$d(x_c,C_k^n)$ 表示 x_c 和 C_k^n 之间的距离。在第二阶段,数据分段以数据的时间属性为处理对象,将进一步细分 k-means 的聚类结果,获得更多的数据组。具体而言,数据分段将根据簇之间的分界点数据,以及数据之间的时间连续性,对最佳聚类结果中的簇进行细分。获取的数据组将具有时间上的连续性,且数据组

中的数据基本隶属于最佳聚类结果中的同一个簇。

3.2.2　离群点识别

离群点识别负责发现数据组中与邻近数据存在较大偏差的异常数据,图 4 为离群点识别流程。

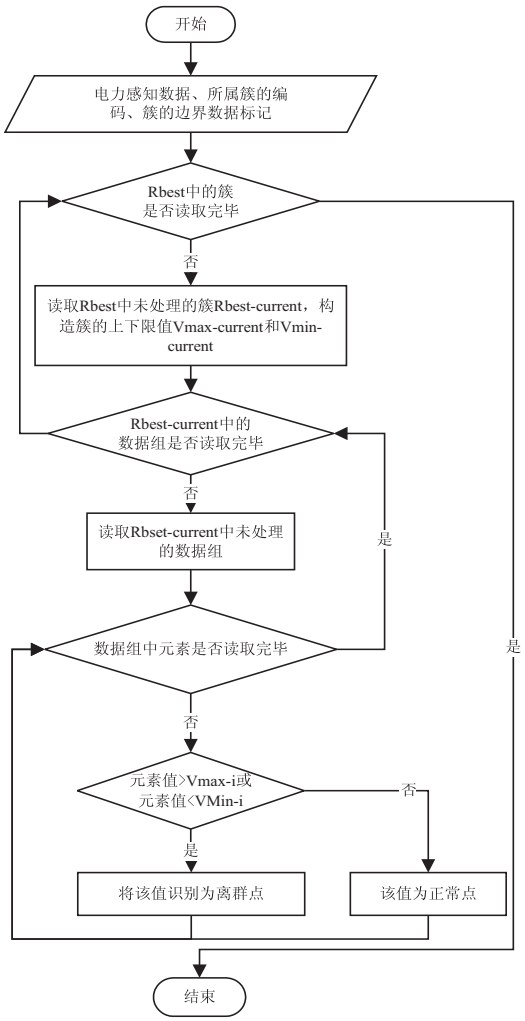

图 4　离群点识别流程

首先,为最佳聚类结果中的每一个簇生成上下限值。然后,判定经数据分类方法处理后获得的数据组中是否存在超出所属簇上下限的数据值,并将该数据标记为离群点。

4　实验和结果分析

电力感知数据具有大规模的特性,通常存储于大数据平台中。以电能质量数据为例,全国近 1 万个监测点,各监测点每 3 s 采集 2 千余指标数据,每天的数据累积量高达 2.75 T。因此,文中提出的电力感知数据异常检测服务框架和离群点检测方法基于大数据技术实现。具体的,由于 Spark 不仅具有海量数据的处理能力,还具有提供丰富算法的机器学习库 MLlib。该方法以 MLlib 库的聚类算法为基础实现,实验环境为四台虚拟机构成的并行集群,虚拟机的硬件配置为 8 核,32 G 内存,500 G 硬盘,软件平台为 Spark1.6.0。该方法针对具有时间连续性的电力感知数据展开,已在充电桩数据和谐波监测数据上进行验证。采用充电桩的三相基波电流为数据集,通过设置不同的错误率以支持不同的实验,错误率指离群点在数据中的占比。鉴于电力感知数据的多样性,该方法还有待应用于更丰富的电力感知数据集。

4.1　聚类结果的选择

该方法采用 MIA 评估聚类结果的质量,图 5 所示为不同聚类结果的 MIA 值。由图 5 可知,随聚类个数(K 值)增大,MIA 值有减小趋势。然而,却不能仅以 MIA 值为依据选取较大的 K 值,因为当聚类的数量过大时,每个簇的数据量会相应降低,从而影响后续的计算和分析。如图 5 所示,当 K 值由 3 变作 4 时 MIA 值存在明显的变小趋势,而随着 K 值的增加,此趋势逐渐减缓,所以 K 值取 4,即数据分为四类。

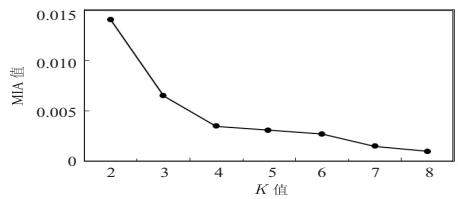

图 5　不同聚类结果的 MIA 值

4.2　召回率

召回率指被检测到的离群点占实际离群点的比例。实验将错误率分别设置为 1%、3%、5%、7% 和 9%,图 6 所示为不同错误率情况下的召回率。如图 6 所示,当错误率从 1% 增加到 10% 时,召回率的值略有下降。鉴于召回率的最小值仍接近 80%,文中方法能够发现大多数离群值。与基于聚类和基于数据分段的离群点检测方法相比,该方法采用一次聚类叠加数据分类的方式,弥补了聚类算法仅实现数据归类却无法识别离群点,以及数据分类无法估计正常数据范围的不足,能够有效识别具有时间连续性的离群点。

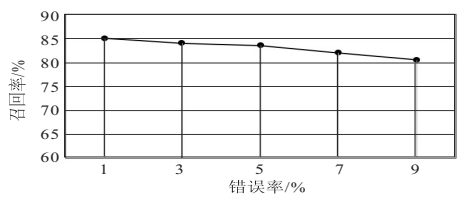

图 6　不同错误率情况下的召回率

4.3　并行加速比

并行加速比指在单个机器上的运行时间与在并行集群上的运行时间的比率,主要用于评价并行系统的性能或并行算法的并行度。实验将错误率设置为 5%,数据量分别为 100 MB、200 MB、300 MB 和 400 MB,并分别在一台虚拟机和一个包含四个虚拟机的并行集群上进行实验。图 7 所示为数据集规模与并行

加速比之间的关系。

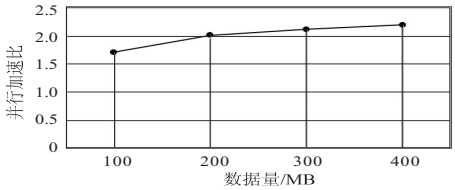

图 7 数据集规模与并行加速比之间的关系

如图 7 所示,随着数据量的增大,并行加速比逐步提升,说明该方法具有较好的并行性能。

4.4 可扩展性

实验通过分析集群规模对并行加速比的影响反映该方法的可扩展性。实验将错误率设置为 5%,数据量分别为 200 MB 和 400 MB。图 8 所示为集群规模对并行加速比的影响,位于下方的虚线为 200 M 数据集的并行加速比,位于上方的实线为 400 M 数据集的并行加速比。

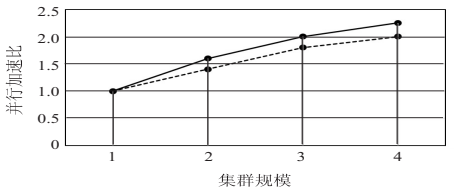

图 8 集群规模与并行加速比之间的关系

如图 8 所示,该方法近似线性加速,说明其具有良好的可扩展性。上述实验验证了该方法具有良好的并行加速比和扩展性。此外,与基于模型、基于距离和基于密度的离群点检测算法相比,该方法无需预知数据分布,参数简单,且复杂度较低。

5 结束语

为了提高电力感知数据的数据质量,提出了一种电力感知数据的离群点检测解决方案,包括异常数据检测服务框架和考虑时间属性的离群点检测方法。给出的框架由存储层、计算层、服务层和应用层构成,能够提供电力感知数据的异常数据检测服务。提出的考虑电力感知数据的时间属性的离群点检测方法,基于电力感知数据的时间特性,基于 k-means 聚类和时间相关的数据分段方法实现,弥补了聚类算法仅实现数据归类却无法识别离群点,以及数据分类无法估计正常数据范围的不足,能够有效识别具有时间连续性的离群点。实验结果表明,该方法具有良好的离群点检出率,并有良好的并行性能和可扩展性。接下来,将进一步开展更丰富的实验验证,以及探索基于电力感知数据其他属性的离群点检测方法。

参考文献:

[1] 中国电力工程师信息委员会. 中国电力发展白皮书[R]. 北京:中国电力出版社,2013.

[2] SWAPNA S,NIRANJAN P,SRINIVAS B,et al. Data cleaning for data quality[C]//3rd international conference on computing for sustainable global development (INDIACom). New Delhi:IEEE,2016:344-348.

[3] EDGEWORTH F. On discordant observation[J]. Philosophical Magazine,1887,23(5):364-375.

[4] CHEN Shuyan,WANG Wei,ZUYLEN H. A comparison of outlier detection algorithms for ITS data[J]. Expert System with Applications,2010,37(2):1169-1178.

[5] HAN J. Data mining:concepts and techniques[M]. San Francisco:Morgan Kaufmann,2005.

[6] AGGARWAL C C, YU P S. Outlier detection for high dimensional data[C]//Proceedings of the ACM SIGMOD international conference on management of data. Santa Barbara,California,USA:ACM,2011:37-46.

[7] PAPADIMITRIOU S, KITAGAWA H, GIBBONS P B,et al. Loci:fast outlier detection using the local correlation integral[C]//Proceedings 19th international conference on data engineering (Cat. No. 03CH37405). Bangalore, India:IEEE,2003:315-326.

[8] 江 峰,杜军威,眭跃飞,等. 基于边界和距离的离群点检测[J]. 电子学报,2010,38(3):700-705.

[9] YAO Z, MARK P, RABBAT M. Anomaly detection using proximity graph and PageRank algorithm[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(4): 1288-1300.

[10] TANG G,BAILEY J,PEI J,et al. Mining multidimensional contextual outlier from categorical relation data[J]. Intelligent Data Analysis,2015,19(5):1171-1192.

[11] DIAO Y,LIU K,MENG X,et al. A big data online cleaning algorithm based on dynamic outlier detection[C]//International conference on cyber-enabled distributed computing and knowledge discovery. Xi'an:IEEE,2015:230-234.

[12] 王习特,申德荣,白 梅,等. BOD:一种高效的分布式离群点检测算法[J]. 计算机学报,2016,39(1):36-51.

[13] 程 超,张汉敬,景志敏,等. 基于离群点算法和用电信息采集系统的反窃电研究[J]. 电力系统保护与控制,2015,43(17):69-74.

[14] 张志强,何文春,朱 江,等. 基于 B/S 架构的雾霾专题数据可视化服务系统设计与实现[J]. 计算机应用,2014,34(S2):140-142.

[15] 夏 虹,郭 超,陈彦萍,等. 面向工业的开放数据服务平台研究[J]. 微处理机,2017,38(3):88-92.

[16] 佟 杰,吕 蓬,李 磊. 海洋测绘数据服务保障系统的设计研究[J]. 测绘地理信息,2018,43(1):28-31.

[17] CHICCO G,NAPOLI R,POSTOLACHE P,et al. Customer characterization options for improving the tariff offer[J]. IEEE Transactions on Power Systems,2003,18(1):381-387.