International Journal of Database Theory and Application
Vol.9, No.3 (2016), pp. 137-150
http://dx.doi.org/10.14257/ijdta.2016.9.3.15

A Data Cleaning Model for Electric Power Big Data Based on Spark Framework[1]

Zhao-Yang Qu[1], Yong-Wen Wang[2,2], Chong Wang[3], Nan Qu[4] and Jia Yan[5]

[1, 2]School of Information Engineering of Northeast Dianli University, Jilin 132012, China[3]
Information &Telecommunication Branch Company, State Grid East Inner Mongolia Electric Power CO.LTD, 010020 Hohhot, China
[4]Repair Branch Company, State Grid Jiangsu Electric Power Company, 210000 Nanjing, China[5]
State Grid Jilin Electric power Supply Company, 130000 Changchun, China
qzywww@mail.nedu.edu.cn, danger_w@qq.com, wangchongky@163.com, {351178520, 3743405}@qq.com

Abstract

The data cleaning of electrical power big data can improve the correctness, the completeness, the consistency and the reliability of the data. Aiming at the difficulties of the extracting of the unified anomaly detection pattern and the low accuracy and continuity of the anomaly data correction in the process of the electrical power big data cleaning, the data cleaning model of the electrical power big data based on Spark is proposed. Firstly, the normal clusters and the corresponding boundary samples are obtained by the improved CURE clustering algorithm. Then, the anomaly data identification algorithm based on boundary samples is designed. Finally, the anomaly data modification is realized by using exponential weighting moving mean value. The high efficiency and accuracy is proved by the experiment of the data cleaning of the wind power generation monitoring data from the wind power station.

Keywords*: Electric power big data, Data cleaning, Anomaly identification, Anomaly modification*

1. Introduction

Along with the publication of "the white paper on the development of electric power in China" [1], has led the research boom of electric power big data within power industry. Accurate and reliable is essential to ensure the precision of big data analysis and process; therefore, the quality for power big data raised higher requirements. Data cleaning for electric power big data can effectively guarantee the correctness, the completeness, the consistency and the reliability of the data.

Big data of electric power has the characteristics of large quantity, high dimension, and various modes and so on. It is inevitable to have abnormal data in the process of electric power data acquisition, so it is necessary to do some amount of cleanup before data analysis. In the domestic and foreign, the research on data cleaning of electric power big data mainly has the clustering and correlation analysis [2], the conditional function dependence [3], the Markov model [4], the DS evidence theory [5]. Most of the data

[1] This work was supported by the National Natural Science Foundation of China (Grant NO.51277023). This work was Supported by the Key Projects of Science and Technology Plan of Jilin Province (NO.20140204049GX).
[2] Corresponding author email: danger_w@qq.com.

ISSN: 2005-4270 IJDTA

cleaning techniques need to rely on the data model itself to construct the abnormal data identification rules. To deal with abnormal data by deleting or filling of the mean value, it will destroy the continuity, integrity, accuracy of the data. Comprehensive domestic and foreign research, the difficulty of the data cleaning for electric power big data performance in the following: (1)It is not suitable to set up rules of abnormal data identification for big data, because those electric power big data not only have numerous types, various characteristics, different carriers and platform, but also have different resources structures and data quality.[3]proposed a new algorithm for detection and repair of inconsistent data based on Hadoop framework, but in the processing phase, the conditional function of relational schema is required to be given to the property set. This method requires human intervention, and because of complex data relationship models and data table in a wide variety of information, it is difficult to designate the Function Dependencies for each relationship. (2)In a data sample, the normal data sample is much, the abnormal data is little, and the different types of electric power big data is difficult to be identified by setting the threshold. (3)Data continuity can be compromised when abnormal data is deleted by usual in the abnormal data processing. (4)When abnormal data is reconstructed, we need to rely on external sources of data. [6] Proposed identifying abnormal data by quartile method and reconstructing missing data by using the neighboring wind farms outputs and multi-point cubic spline. However, identification method is easy to eliminate the non-recognition of normal data, missing data at the time of construction of the wind power relies on the nearby wind power data is likely to cause errors.

Aiming at difficulties of data cleaning for electric power big data, a data cleaning model for electric power big data based on Spark is proposed. Firstly, the normal clusters and the corresponding boundary samples are obtained by the improved CURE clustering algorithm. Then, the anomaly data identification algorithm based on boundary samples is designed. Finally, the anomaly data modification is realized by using exponential weighting moving mean value. Compared to some of the data cleaning model for electric power big data, the model proposed in this paper reduces human intervention, does not need to set the identification rules based on the data relationship model, the correction of anomaly data is based on the analysis of the data in the same time series. This model can eventually be able to clean up the anomaly data in the historical or real time data.

2. Data Cleaning Model for Electric Power Big Data Based on Spark Framework

Power big data cleaning is the process of correcting the anomaly data in the electric power big data. Using the Spark framework to build a power big data cleaning model is divided into the following stages: data preparation, normal sample acquisition, anomaly data identification, anomaly data modification, revised data storage. Data preparation is to transfer the data in the traditional relational database to a non-relational database which is suitable for big data processing, and then load them into the Resilient Distributed Datasets of Spark. The normal cluster sample which is used to realize the anomaly data identification based on boundary samples can be obtained by extracting a certain amount of samples of electrical power big data and applying the hierarchical clustering algorithm on them. The anomaly data identification algorithm is based on the boundary sample, we can finish the task of the abnormal data detection of the electrical power big data by the anomaly identification based on boundary samples. Abnormal data correction is to complete the repairing of anomaly data in electric power big data. Figure 1 gives a data cleaning model of electric power big data based on Spark framework, and the cleaning procedure is as follows:

(1) Data preparation is to store electric power big data into the distributed file system of Hadoop.

(2) Reading data from the distributed file system and performing the cache operation to generate the RDDs, the data can be read into memory.

(3)Using improved parallel CURE clustering algorithm to obtain the normal data samples.

(4)Select the boundary samples data from normal samples.

(5)Design anomaly data identification algorithm based on boundary samples, and detects the anomaly data in test samples.

(6) Mark where abnormal data is detected in a sample location.

(7) Modify anomaly data with exponentially weighted moving average model.

(8)Form revised data set and save.

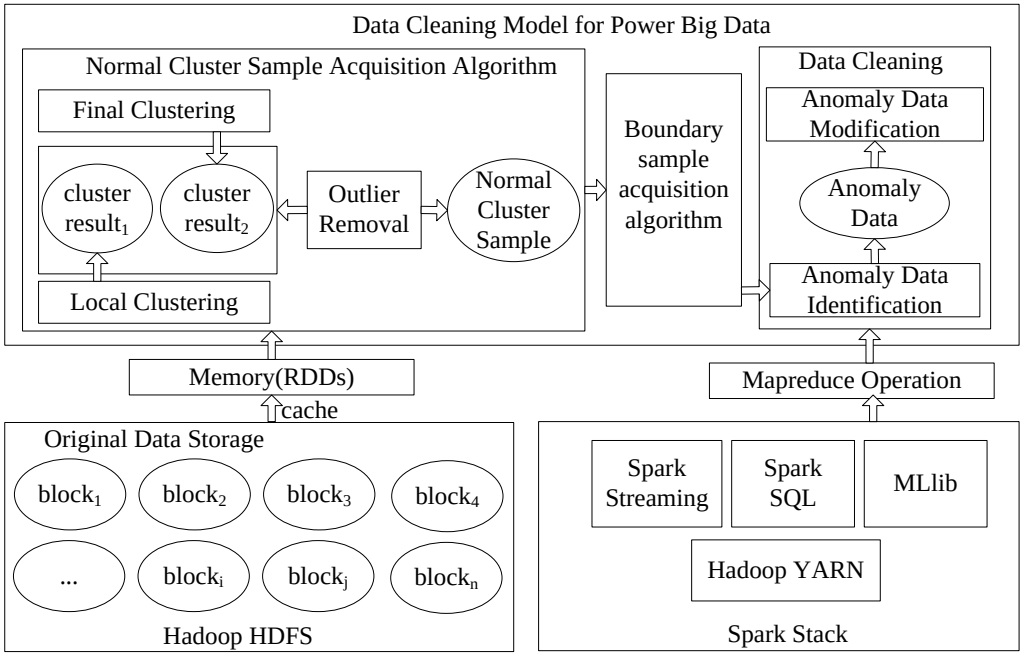

Figure 1. Data Cleaning Model for Power Big Data Based on Spark

In this paper, several key steps of data cleaning model for electric power big data will be discussed. Firstly, the normal cluster sample acquisition algorithm based on Improved CURE clustering is described in detail. Secondly, the selection process of the boundary sample is introduced, and the algorithm of outlier data identification is analyzed in detail. At last, the paper introduces the exponential weighted moving mean value to modify the anomaly data. At the end of this paper, the experimental analysis and verification are given, and the work of this paper is summarized and the future research directions are pointed out.

3. Normal Cluster Sample Acquisition Algorithm

Electric power data acquisition devices have data validation capabilities, so the data collected are most normal data, and fewer anomaly data. And electric power big data has wide variety types, it is unable to directly construct a single rule or set thresholds for anomaly data identification. And it is large calculation quantity and low efficiency to do abnormal identification on the acquired electrical power big data directly. Therefore, we can obtain the normal cluster sample from the data sample of electirc power big data, and based on the boundary sample set of the normal cluster, we can identify the history or real time data. This kind of abnormal identification algorithm does not depend on the data

attribute threshold and the mathematical model rules, and can improve detection efficiency.

CURE clustering algorithm by means of eliminating outliers reduce the impact on the clustering results, so we can carry out CURE clustering algorithm on the test sample to obtain the normal cluster sample. CURE cluster algorithm is used to delete the outliers in two stages: the first stage is to remove the outliers in the cluster growth is very slow; the second stage is at the end of the cluster, data node quantity of cluster significantly less as the outlier removed. However, the following problems exist in the CURE cluster algorithm when the outliers are deleted: (1) It is difficult to define the type of growth rate at the first stage [7], (2) because feature of local data distribution, local data in cluster is submerged after outliers removed [8].

For the problem existed in the CURE clustering algorithm to eliminate the outliers, in this paper, we use outlier detection to determine the outlier, which can effectively solve the problem to define the class of slow growth and local outlier submerged phenomenon. Related defined as follows:

Definition 1. clustering for each divided data block, and the obtained aggregate of data is expressed as $p_i(mp_i, w_i)$, where p_i represents the central point of the ith aggregate, mp_i represents the central point of the ith aggregate, w_i represents the weight value of each central point, is the number of data in each aggregate. Therefore each divided data block can use several p_i to represent, referred to as the representative point.

Definition 2. Assumed the set of representative point is P, the deviation distance of the central point of each representative point p_i to any point out of aggregate is expressed as degree of outlier：

$$d_i = \sqrt{\sum_{j=0}^{n}(x_{ji} - y_{ji})^2} \tag{1}$$

Using Euclidean distance to represent the deviation distance of one point, a point is further away from the central point of aggregate, the degree of outlier much bigger.

Definition 3. Assumed the degree of outlier set is D, and the decision value of degree of outlier is defined as:

$$AD = \frac{1}{m}\sum_{i=1}^{m} d_i \tag{2}$$

Definition 4. Assumed outlier parameter is δ, the minimum of outlier degree is $\min(d_i)$,

$$\delta = \frac{AD + (AD - \min(d_i))}{AD} \tag{3}$$

Definition 5. For any $d_i \in D$ in outlier degree set D, if $d_i > \delta * AD$, then the d_i corresponding representative point p_i is the outlier point, its data in aggregate is the outlier data.

The normal cluster sample Acquisition algorithm based on improved CURE algorithm is designed as following pseudo code.

//select random points


```
for i = 0 to samplesize do
   randomPoints.add(sample[random.next(i)]);
end for
//partition pointSet
for i = 0 to numberofpartitions do
   repeat
      partitionset[i].add(pointset.next());
   until pointIndex < pointset.size/numberofpartitons
end for
//cluster subpartitions
for i = 0 to partitionsetsize do
   initializecontainers();
   initalizepoints()
   buildKDTree()
   buildHeap()
   repeat
      newcluster = merge(mincluser,closestcluster)
      adjustheap(newcluster,mincluster,closetcluster)
   until headsize > clusterstobefound
   subcluster = clusterset.getallclusters()
end for
//calculate reducing factory
reducingfactory = calculatereducingfactorysubclass()
//calculate outlier degree decision value
AD = calculateAD(subclusters)
//calculate the outlier level
level = calculateoutlever(subclusters)
for i = 0 to subclusterssize do
   if factory[i] > level * AD then
      eliminateoutliersfirststage(subcluster)
   end if
end for
//clustering the m/q clusters
clusterset = mergeclustersforall()
labelRemainingDataPoints(clusters)
```

The basic idea of the normal cluster sample Acquisition algorithm: firstly, a random sample is drawn from data, and the selection of the sample should be representative; secondly, the samples are divided into several data sets of the same size, and then those data sets are clustered for the first time, to get the m/q clusters, and the degree of outlier determination value(AD) and outlier parameter(δ) of each point in the cluster are calculated, then delete outliers that are not satisfying $d_i > \delta * AD$; thirdly, perform the clustering on m/q clusters for the second time, meanwhile, delete the cluster in the data sample that are obviously small; at last, all the remaining data points are assigned to the nearest cluster, then we will get the normal cluster samples.

4. Algorithm for Anomaly Data Identification Based on Boundary Samples

In this paper, we propose a method for anomaly data identification based on boundary samples. First, we obtain the boundary sample set of normal clusters; then, according to the anomaly data identification algorithm to detect anomaly data; finally mark and record the location of the anomaly data.

Anomaly data identification is the process of detecting the anomaly data in the historical or real time data in the power big data, which is based on the boundary samples of normal clusters. The boundary sample of each normal cluster must have the characteristics as follows:

(1) furthest from the normal cluster quality;

(2) Scattered around the normal sample;

(3) represent the shape of normal samples.

The relevant mathematical proofs are given to ensure that the boundary samples can be scattered around the normal cluster samples.

Proof1: In a triangle, existing a apex located on the perpendicular bisector of the bottom edge, making the sum of distance between apex and the bottom vertex is shortest, as well as the perimeter is shortest. Figure2 gives the case of vertex of a triangle in perpendicular and not in the perpendicular.

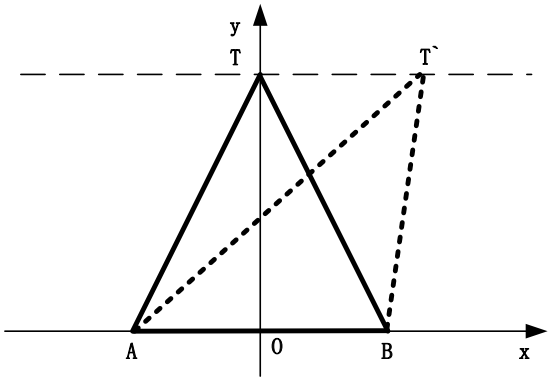

Figure 2. The Case of Vertex of a Triangle in Perpendicular and Not in the Perpendicular

We assume the height of triangle TAB is b, and the value of the bottom |AB|=2a, the vertex coordinate T is T(x , b). We need to prove that exiting a vertex T in perpendicular makes the minimum of |TA|+|TB|. According to the formula, we can get that:

$$\text{distance}_{(|TA|+|TB|)} = \sqrt{|x+a|^2+|b|^2} + \sqrt{|x-a|^2+|b|^2} \tag{4}$$

Formulas on both sides of the derivative can be obtained:

$$\text{distance}_{(|TA|+|TB|)}` = \frac{x+a}{\sqrt{|x+a|^2+|b|^2}} + \frac{x-a}{\sqrt{|x-a|^2+|b|^2}} \tag{5}$$

When x = 0, we can get the minimum value of distance(|TA|+|TB|), the value is $2\sqrt{|a|^2 + |b|^2}$。

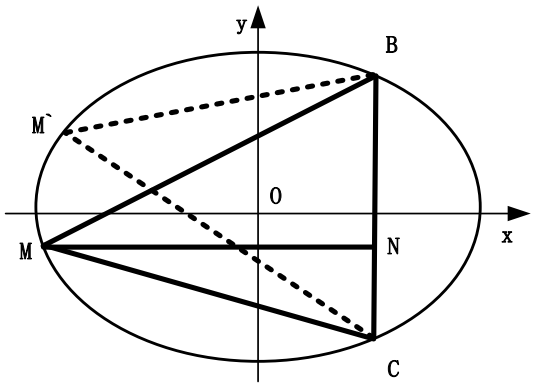

Figure 3. Schematic Diagram of Elliptical Dispersion Point

Proof 2: Existing a point on ellipse makes |MB|+|MC| is the largest, B and C are the point on the ellipse.

Figure 3 shows the distribution of the elliptical data nodes. We assume the value of elliptical long axis is 2m, the value of short axis is 2n,|BC|=2a, and the farthest point from BC belong to the straight line MN, |MN|=b, M` is the farthest point that makes |M`B|+|M`C| is the largest. According to the formula4, we can get that:

(6)

$$\text{distance}_{(|MB|+|MC|)} \geq 2\sqrt{|a|^2 + |b|^2}$$

(7)

$$\text{distance}_{(|M`B|+|M`C|)} > \text{distance}_{(|MB|+|MC|)}$$

(8)

$$\text{distance}_{(|M`B|-|M`C|)} < \text{distance}_{(|BC|)}$$

After operation process to formula 6 to 8, we can obtain the formula 9:

(9)

$$\text{distance}_{(|M`B|)} > \sqrt{a^2 + b^2} - a$$

By the formula 9 we can get that M` will not in $\sqrt{a^2 + b^2} - a$ areas of B and C, therefore, if you choose B, C as a representative point, M` can also be used as a representative point, making M`, B, C as a representative point enough dispersion.

In the selection of boundary samples, the characteristics of the boundary sample points should be kept. The details of boundary sample selection algorithm are explained below：

Step 1. Calculate the center for cluster k, center point = $(n_1 + n_2 + \cdots + n_m)/m$ and n_i is the node in cluster, m is the total number of cluster.

Step 2. The first boundary point is the farthest point from the center, and the second boundary point is the farthest point from the first sample point.

Step 3. The next selected boundary point is the point which is with the max sum of the distance from the previous two boundary points, the selection will not be terminated until the selected sample point can represent the cluster k.

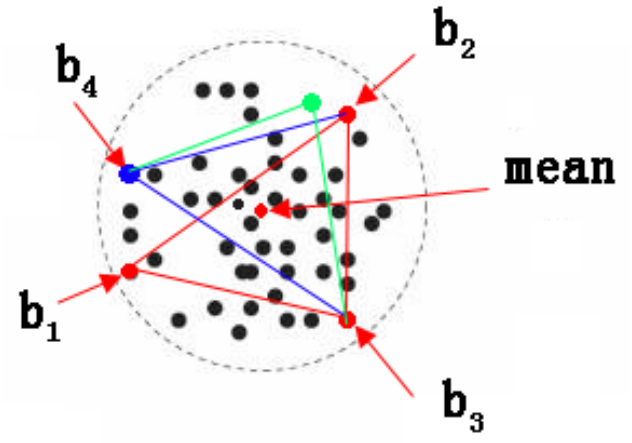

Figure 4. A Sample of Boundary Points Selection Process

The boundary points are scattered around normal sample cluster, which can represent the shape of cluster. Using the boundary sample of normal cluster to identify the anomaly data from test sample, can reduce the amount of calculation of anomaly identification algorithm.

The boundary sample set of normal cluster sample is B= {b1,b2,…, bn}, the recognition radius of normal samples is rs, the sample to be tested is T= {t1,t2,…,tm}.The boundary sample set is used as a plane detector, and the rule of anomaly identification is:

$$\begin{cases} t_i \in N, & l_1 \geq l_2 \\ t_i \in N, & d_m > r \\ t_i \in S, & d_m \leq r \end{cases}$$

Where l_1 is the minimum distance from the sample point to the boundary sample, l_2 is the maximum distance from points in the sample to be tested and boundary sample, is the distance from sample to be identified to centroid, r is the recognition radius for normal cluster samples.

The steps of anomaly data identification algorithm based on boundary sample described as follows:

Algorithm input: samples to be identified T={t_i, i∈[1,m], m is total number}, boundary sample set B={ b_j, j∈[1,n], n is total number}, the centroid m for normal cluster sample k_i ;

Algorithm output: anomaly sample Q;

Step 1. Calculate the distance from the sample point ti to the boundary sample point bj, formed a distance sequence dist, dist={d_1,d_2,…,d_n} ;

Step 2. Calculate the recognition radius r_s for normal cluster sample;

Step 3. Find the minimum value l1 from sequence dist, l1=min(d_i) ;

Step 4. Find the point b_k from boundary sample which is farthest from the point t_i in sample to be identification, calculate the distance l_2 from t_i to b_k, l_2=distance(t_i, b_k);

Step 5. If $l_1 \geq l_2$, t_i is the anomaly data, if $l_1 < l_2$, then execute step 6;

Step 6. Calculate the distance from t_i to centroid m of normal cluster sample k_i, dim=distance(t_i,m);

Step 7. If $d_{im} > r_s$ then t_i is the anomaly data;

Step 8. Location the anomaly data in the test sample;

Step 9. Repeat execute step 1 to 8 until all samples have to be identified;

Step 10. Output all anomaly data.

By using boundary sample anomaly identification algorithm, we do not have to set up the threshold for anomaly data identification, and can avoid the complexity of using data mode, and can improve the efficiency of the anomaly identification.

5. Anomaly Data Modification Based on Time Series Analysis

Electric power big data is the accumulation of data collection in a certain period of time. Because of its variety, with the time change generally presents three kinds of law: periodic variation, amplitude variation is small, slowly increasing type. The effect of anomaly data in the time series of power big data is shown in two forms: the first one is the additive outliers, such outliers affect only the moment sequence outliers occur on, but do not affect the value of the time series. The second one is to update the outliers, which not only affects the generation of the abnormal points at that time, but also affects all the measurements in a period of time.

The anomaly data should be analyzed and modified according to the feature and the forms of the anomaly data. When correcting the anomaly data of electrical power big data which is in the case of slow-increasing or attenuation, the reference data series will be selected as the [n, m] interval in the series which the anomaly data is in. when correcting the anomaly data of the electrical power big data which is periodic, the selected data series will be the data series of n-periods which includes the abnormal data and the moment of T which the anomaly data is in.

When the anomaly data is modified, the general method is to use the average number of the data sequence where abnormal data occur to replace the abnormal data. The modified value is

$$\bar{x} = {}^{1}/_{n} \sum_{i=0, x_i \epsilon \theta_p}^{n} x_i \tag{10}$$

Where ${}^{1}/_{n}$ is a weight given to x_i, n is the total number of data sequence. However, the effect of a sequence of values on the sequence values is attenuated, but not always ${}^{1}/_{n}$. So we can use weighting moving mean value to modify abnormal data, the modified value is defined as

$$\bar{x} = \sum_{i=0}^{n} \lambda(1-\lambda)^i x_{t-i} \tag{11}$$

6. Experiment and Result Analysis

In this paper, the “Spark on Yarn” cluster model is used to build a data cleaning model experiment environment for power big data. We implement this experiment platform using 6 servers running Ubuntu-12.04.1 operating system with Hadoop-2.6.0, Spark-1.3.1,

Scala-2.10.5, JDK-1.7.0_79 being installed. One of those nodes is used as Master, and other five nodes are lave1-Slave5, the configuration of each node is shown in Table1. The experimental platform is developed in the Idea Scala development environment, and the results are stored in HDFS of Hadoop.

In this paper, we use the wind power monitoring data of a wind farm as the research object of data cleaning. The size of this wind power monitoring data is 5GB collected from 5 wind power generators every second, recorded from February 1, 2012 through February 29, 2012.

Table 1. The Configuration of Each Server Node

Server type	Server number	Memory	CPU kernel number	Network card rate	Hard disk capacity	Processor type
Blade	6	6×128GB	6×16	1Gbit/s	6×1T	Intel Xeon 2.00 GHz

In this paper, we will test and verify the accuracy of anomaly identification and efficiency of anomaly modification of data cleaning model of power big data.

Experiment1. In order to compare the detection rate of outlier detection algorithm in normal sample, this paper tests several outlier detection algorithms, the experimental results are shown in Table2. Compared with the Apriori algorithm, the proposed algorithm in this paper has lower false detection rate in the case of similar detection rate. Lower false detection rata is good for improving the normal sample quality, and also can guarantee the accuracy of the anomaly identification algorithm based on boundary sample. And compared with the original CURE clustering algorithm, the improved CURE clustering algorithm is improved in both detection rate and false detection rate.

Table 2. Comparison of Outlier Detection Algorithm

Algorithm Name	Detection Rate (%)	False Detection Rate (%)
K-means	46.78~78.96	5.02~15.45
Apriori	84.3~87.8	8.1~17.4
Original CURE	81.09~85.10	3.47~-5.49
Improved CURE	86.65	2.54

Experiment2. In order to verify the correctness of the power big data anomaly identification algorithm to identify abnormal data, in this paper, we maintain the number of nodes in the cluster, and constantly changing the size of the test data. From Table3, the experimental result shows accuracy of identification can be reached above 90%, the algorithm can identify almost of anomaly data.

Table 3. The Accuracy of Anomaly Identification Algorithm of Power Big Data

Sample Number	Actual Error Data Number	Detected Error Data Number	Accuracy Rate
1G	28	29	96.551%
2G	60	65	92.308%
3G	93	102	91.177%
4G	116	118	98.305%
5G	148	152	97.368%

Experiment3. In order to verify the efficiency of data cleaning model for power big data, traditional data cleaning model and power big data model based on Spark are tested. Maintain cluster node data is fixed, and constantly adjust the amounts of the sample to be cleaned, and then record the cleaning time of the different amounts of data sample .Test results are shown in Table4. Eliminating overhead of task scheduling and network communication between nodes, the efficiency of data cleaning for power big data based on Spark is higher than that of traditional single machine data cleaning.

Table 4. Comparison of Single and Parallel Data Cleaning Time

Sample Number	Stand-alone Data Cleaning/s	Data Cleaning based on Spark/s
1G	5808	45
2G	20160	108
3G	37912	247
4G	62570	357
5G	95067	432

Experiment 4. First, test data sample size is fixed, and then 150000 data is randomly selected as the experimental test sample, the normal cluster sample number is 5, the number of boundary samples of each normal sample is 25, 35, 45, 55, 65. In the case of the number of test sample is fixed, adopting anomaly identification algorithm based on boundary sample for test sample to test the influence of the number of boundary samples on the results. The results are shown in Table5.

Table 5. The Influence of the Number of Boundary Samples on the Results of the Anomaly Identification Algorithm

Sample Number	Boundary Sample Number	Actual Error Data Number	Detected Error Data Number	Accuracy Rate
15000	15	102	151	32.450%
15000	25	102	143	28.671%
15000	35	102	136	25.000%
15000	45	102	126	19.048%
15000	55	102	110	7.272%

7. Conclusion

In this paper, some difficulties in the process of data cleaning are discussed, according to the characteristics of power big data and cleaning difficulties, a data cleaning model for power big data based on Spark is proposed. This data cleaning model has the following characteristics: (1) The anomaly data identification is not required for external source data. (2) A higher accuracy of outliers identification and correction. (3) A higher efficiency of dealing power big data by using Spark framework. However, the problem still exists in the selection of boundary sample, that when the number of optimal boundary samples is reached. The correction of abnormal data is established on the same time series, the accuracy of correction for anomaly data affected by the outliers in those time series. In order to solve the above problem, it is need to further explore and improve the data cleaning model for power big data.

References

[1] "China Institute of Electrical Engineering Information Committee", The White Paper on The Development of Electric Power in China, **(2013)**.
中国电机工程学会信息化专委会.中国电力大数据发展白皮书, **(2013)**.

[2] C. Lukasz, "Application of Clustering and Association Method in Data Cleaning", Proceedings of the International Multi Conference on Computer Science and Information Technology, IEEE Press, New York, **(2008)**, pp. 97-103.

[3] Z. Anzhen, M. Xueying, W. Hongzhi, L. Jianzhong and G. Hong, "Hadoop-Based Inconsistence Detection and Reparation Algorithms for Big Data", Journal of Frontiers of Computer Science and Technology, CWAJ Press, Beijing **(2014)**, pp. 1-12.
张安珍,门雪莹,王宏志,李建中,高宏, 大数据上基于 Hadoop 的不一致数据检测与修复算法[J],计算机科学与探索, vol. 12, **(2014)**, pp. 1-12.

[4] C. Shuang, S. Jinyu, D. Xingchun and C. Jianjun, "Estimation of Enumerative Missing Values Based on Relational Markov Model", Journal of Shanghai Jiao Tong University, SJTU Press, Shanghai, **(2013)**, pp. 1246--1250.
陈爽,宋金玉,刁兴春,曹建军,基于关系马尔可夫模型的枚举型缺失值估计[J],上海交通大学学报, vol. 43, no. 8, **(2013)**, pp. 1246-1250

[5] F. Jinhui, Y. Kun, Z. Jixian and L. Weiyi, "Data Cleaning Based on D-S Evidence Theory", Journal of Yunnan University(Natural Sciences Edition), YNDXXB Press, Kunming, **(2014)**, pp. 815--822.
樊金辉,岳昆,张骥先,刘惟一, 基于 D-S 证据理论的不确定数据清洗[J],云南大学学报(自然科学, vol. 36, no. 6, **(2014)**, pp. 815-822.

[6] Z. Qianwen, Y. Lin, Z. Yongning, L. Yansheng and S. Xuri, "Methods for elimination and reconstruction of abnormal power data in wind farms", Power System Protection and Control, vol. 43, no. 3, **(2015)**, pp. 38-45.
朱倩雯,赵永宁,郎燕生,宋旭日,风电场输出功率异常识别与重构方法研究[J],电力系统保护与控制, vol. 43, no. 3, **(2015)**, pp. 38-45.

[7] S. Jie, Z. Lei and Y. Jiwen, "Hierarchical clustering algorithm based on partition", Computer Engineering and Applications, vol. 43, no. 31, **(2007)**, pp. 175-177.
沈洁,赵雷,杨季文,李榕,一种基于划分的层次聚类算法[J] ,计算机工程与应用, vol. 43, no. 31, **(2007)**, pp. 175-177.

[8] Y. Fuping, W. Hongguo, D. Shuxia and N. Jiaxiang, "Two Stage Outliers Detection algorithm Based on Clustering Division", Application Research of Computers, vol. 30, no. 7, **(2013)**, pp. 1942-1945.
杨福萍,王洪国,董树霞,牛家洋,丁艳辉,基于聚类划分的两阶段离群点检测算法[J],计算机应用研究, vol. 30, no. 7, **(2013)**, pp. 1942-1945.

Authors

Zhao-Yang Qu, received his MSc and PhD degree in computer science from Dalian University of Technology, China in 1988 and North China Electric Power University, China in 2012, respectively. His research interests are in artificial intelligence, machine learning and data mining. He is now a professor of Northeast Dianli University.

Yong-Wen Wang, received his Bachelor degree in computer science from Northeast Dianli University, China in 2013. His research interests are machine learning and data mining. he is now a postgraduate of Northeast Dianli University.