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

Zhao-Yang Qu1, Yong-Wen Wang2,2, Chong Wang3, Nan Qu4, and Jia Yan5

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, Hohhot 010020, 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. 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

1 This work was supported by the National Natural Science Foundation of China (Grant NO.51277023).
2 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.
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 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. [3] proposed a new algorithm for detection and repair of inconsistent data based on Hadoop framework. This method requires human intervention, and because of complex data relationship models, it is difficult to designate the Function Dependencies for each relationship. [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.

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. (2)Data continuity can be compromised when abnormal data is deleted by usual in the abnormal data processing.

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 cluster sample acquisition, anomaly data identification, anomaly data modification, revised data storage.
2.1 Normal Cluster Sample Acquisition Algorithm based on improved CURE

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. However, the following problems exist in the CURE cluster algorithm when the outliers are deleted: (1) It is difficult to define a class of very slow growth in clustering [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. Normal cluster sample Acquisition algorithm based on improved CURE algorithm as follows:

```
int sampleSize = calcuteSampleSize();
List randomPointSet = selectRandomPoints(sampleSize);
List[] partitionedPointSet = partitionPointSet(randomPointSet);
List[] subClusters = clusterSubPartitions(partitionedPointSet);
float[] factory = calculateReducingFactory(subClusters);
float AD = calculateAD(factory);
float level = calculateOutlinerLevel(subClusters);
if factory[i] > level*AD then
    eliminateOutliersFirstStage(subClusters);
end if
list clusters = clusterAll(subClusters);
labelRemainingDataPoints(clusters);
```

where `calculateAD(factory)` is used to calculate the outlier degree decision value, `calculateOutlinerLevel(subClusters)` is used to calculate the outlier level.

2.2 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.

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.

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.

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

Step 1. Calculate the distance between point \( t_i \) in set \( T \) and point \( b_j \) in boundary sample set, let \( d_i = \text{distance}(t_i, b_j) \), where \( T=\{t_m\} \), m is the total amount of the test data, and \( D=\{d_1, d_2, ..., d_n\} \) be the set of all distances.

Step 2. Find the minimum distance \( d_{\min} \) from \( D \), let \( l_1 = d_{\min} = \min(d_i) \), the boundary point is \( b_{\min} \).

Step 3. Find the farthest point \( b_j \) from boundary node \( b_i \) in boundary sample, and then calculate the distance between point \( t_i \) in set \( T \) and point \( b_j \), let \( l_2 = \text{distance}(t_i, b_j) \).

Step 4. If \( l_1 \geq l_2 \), \( t_i \) is the anomaly data; if \( l_1 < l_2 \) then calculate the distance between point \( t_i \) and center of mass, let \( d_{im} \) is the new distance value. If \( d_{im} > r_s \) then \( t_i \) is the anomaly data, where \( r_s \) is radius of sample.

Step 5. Location the anomaly data in the test sample.

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.

2.3 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.

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\( \bar{x} = \frac{1}{n} \sum_{i=0}^{n} w_i x_i \theta_p x_i \), Where \( \frac{1}{n} \) 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 \( \frac{1}{n} \). So we can use weighting moving mean value to modify abnormal data, the modified value is defined as

\[
\bar{x} = \sum_{i=0}^{n} \lambda (1-\lambda)^i x_{i-1}.
\]
3 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 Slave1-Slave5, the configuration of each node is shown in Table 1. 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>
<thead>
<tr>
<th>Table 1. The configuration of each server node.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server type</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Blade</td>
</tr>
</tbody>
</table>

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.

Experiment 1. 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 Table 2. 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 rate 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>
<thead>
<tr>
<th>Table 2. Comparison of outlier detection algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm Name</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>K-means</td>
</tr>
<tr>
<td>Apriori</td>
</tr>
<tr>
<td>Original CURE</td>
</tr>
<tr>
<td>Improved CURE</td>
</tr>
</tbody>
</table>

Experiment 2. 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 Table 3, 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

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Actual Error Data Number</th>
<th>Detected Error Data Number</th>
<th>Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1G</td>
<td>28</td>
<td>29</td>
<td>96.551%</td>
</tr>
<tr>
<td>2G</td>
<td>60</td>
<td>65</td>
<td>92.308%</td>
</tr>
<tr>
<td>3G</td>
<td>93</td>
<td>102</td>
<td>91.177%</td>
</tr>
<tr>
<td>4G</td>
<td>116</td>
<td>118</td>
<td>98.305%</td>
</tr>
<tr>
<td>5G</td>
<td>148</td>
<td>152</td>
<td>97.368%</td>
</tr>
</tbody>
</table>

Experiment 3. 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 Table 4. 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

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Stand-alone Data Cleaning/s</th>
<th>Data Cleaning based on Spark/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1G</td>
<td>5808</td>
<td>45</td>
</tr>
<tr>
<td>2G</td>
<td>20160</td>
<td>108</td>
</tr>
<tr>
<td>3G</td>
<td>37912</td>
<td>247</td>
</tr>
<tr>
<td>4G</td>
<td>62570</td>
<td>357</td>
</tr>
<tr>
<td>5G</td>
<td>95067</td>
<td>432</td>
</tr>
</tbody>
</table>

4 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

   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.


