ATLANTIS PRESS

Advances in Computer Science Research, volume 92

International Conference on Mathematics, Big Data Analysis and Simulation and Modeling (MBDASM 2019)

Data Cleaning Algorithms for Power Information Communication Assets Data Based on Self-coding

Qiang Wang[1], Jianliang Zhang[2], Jian Wu[2], Feng Gao[3,*] and Xuewu Ren[3]

[1]State Grid Shanxi Electric Power Company, China

[2]State Grid Shanxi Electric Power Company Information and Communication Branch, China

[3]Beijing Qianrunhe Technology Co., Ltd, China

[*]Corresponding author

***Abstract*—With the continuous application of new information and communication technology in the asset management of power information and communication equipment, a large amount of data will be generated in all aspects of asset management. In data acquisition, due to sensor faults and other reasons, data anomalies will inevitably occur in the data set, which causes great trouble to the subsequent data analysis. In this paper, a data cleaning algorithm based on stack self-encoder (DCbS) is proposed, which can improve the ability of distinguishing and recovering outliers of data by saving the short-term correlation between data in sliding windows and residual analysis between noisy data and lossless data. Finally, the advantages of the algorithm are highlighted from two aspects: data recovery and outlier identification.**

Keywords—big data;data clean; self-coding network; residual analysis

I. INTRODUCTION

As a new generation of technology, the hidden value of big data will bring disruptive changes to many industries. Big data technology aims to extract its economic value from a large amount of data through rapid capture, collection and analysis methods. In recent years, the State Grid Corporation has also carried out various research projects on the application of large data in smart grid[1]-[4]. Large data sources in smart grid are wide, such as smart meter measurement data, equipment management, control and maintenance data, user load data, etc.. At present, the data volume in smart grid field is exponentially increasing, so the potential value of data in power field can be analyzed by using large data technology reasonably and efficiently, which can be used to guide power production and management[5]-[7]. For large data in the field of electric power, due to the particularity of the industry, there are certain requirements for the integrity of data. In recent years, a lot of research has been done on the cleaning of large data in the electric power industry[8].

In this paper, a data cleaning algorithm based on SDAE (DCbS) based on stack self-encoder is proposed, which combines the characteristics of high data quality required by power information and communication assets. DCbS algorithm can distinguish abnormal data, recover the reconstructed singular points and missing data, and train the model from the perspective of residual by introducing sliding window to preserve the short-term correlation between the data, so as to reduce the training data needed by the model to distinguish abnormal data points. In view of the abnormal operation of power information and communication assets, this method can effectively filter interference data

II. DATA CLEANING TECHNOLOGY FOR POWER INFORMATION AND COMMUNICATION ASSETS

Compared with the traditional big data technology, the large data technology for power information and communication assets needs higher efficiency and accuracy, so there is a higher requirement for data integrity. Before data processing, data need to be cleaned[9]-[11]. At present, the research on big data cleaning at home and abroad mainly includes clustering and association analysis, conditional function dependence, Markov model, etc. Big data cleaning technology relies on the data model itself to build anomaly data recognition rules, delete and clean the anomaly data[12]-[14].

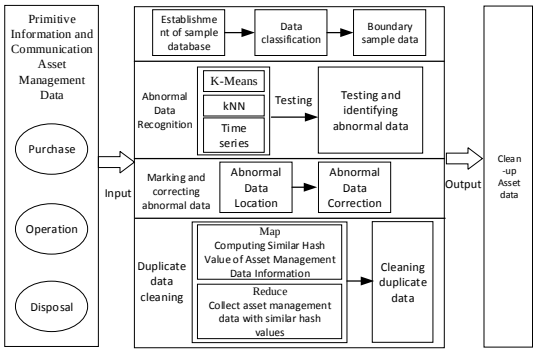

FIGURE I. BIG DATA CLEANING TECHNOLOGY OF POWER INFORMATION COMMUNICATION ASSETS

For large data cleaning of power information and communication assets, it includes four steps: establishing sample database, identifying abnormal data, marking and correcting abnormal data and cleaning duplicate data. The purpose of establishing the sample database is to identify the boundary sample data and use clustering algorithm to identify the abnormal data as the boundary. Exceptional data includes missing data and error data, data that can be corrected is corrected, and data that cannot be corrected is marked and

deleted. Data cleaning includes not only the correction and deletion of abnormal data, but also the cleaning of duplicate data[15]. The purpose of cleaning duplicate data is to improve the efficiency of subsequent processing and reduce data redundancy. The data cleaning algorithm based on stack self-encoder is studied in this paper. The stack self-encoder is trained by sample database to recognize abnormal data and predict missing data[16].

III. Data Cleaning Algorithm Based on Stack Self-encoder

A. Self-coding Network

The state parameter information of power information communication assets has non-linear correlation, and each monitoring parameter is time series data, which has short-term dependence characteristics. If the SDAE model is trained directly with the original monitoring data of power information and communication assets, the non-linear relationship among the parameters can be acquired, but the inherent short-term dependence of the monitoring parameters will be neglected, which will affect the accuracy of the unit condition monitoring results.

Therefore, in DCbS, the state parameter data of power information communication assets need to be processed first to obtain model input data with short-term dependence. The sliding window technology is used to process the state parameter data of power information communication assets, so that the sliding time window contains the state information of the current time and the previous time, and generates an augmented state data matrix taking into account the short-term dependence of the parameters. With the augmented state data matrix as input, the SDAE model is established and the non-linear correlation and short-term dependence of the parameters are obtained simultaneously. The specific steps of the sliding window processing method are as follows.

$\mathbf{X} = \{x_i^{(j)}\}$ is a data set of power information and communication assets status parameters, where I = 1,2,... N, J = 1, 2,... M, n is the number of monitoring variables, m is the number of sample data collected. The first component of data X represents the sample data of i monitoring variables of the unit, namely $\mathbf{X}_i = [x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m)}]$. Let the width of the sliding window be beta (that is, the sliding window contains beta time point data). The window moves one time point at a time. For the data X of M samples, there are m-beta+1 sliding window, and $\mathbf{S}_i^{(l)}$ (i=1,2,... (n) Data of monitoring parameters of *l* collected from sliding windows of i, then

$$\mathbf{S}_i^{(l)} = [x_i^l, x_i^{l+1}, ..., x_i^{l+d-1}]^T \tag{1}$$

The data collected from the i th sliding window are as follows:

$$\mathbf{S}^{(l)} = [\mathbf{S}_1^{(l)}, \mathbf{S}_2^{(l)}, ..., \mathbf{S}_n^{(l)}]^T \tag{2}$$

Therefore, the input data augmented state data matrix of SDAE model is obtained by sliding window processing of state parameter data of power information communication assets from formulas (1) and (2).

$$\mathbf{Y} = [\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(m-d+1)}] \tag{3}$$

In addition, when beta = 1, AAAAA, the augmented state data matrix is the original state parameter data. Using sliding window technology, the input data of SDAE increases from n-dimension of original data X to N-B dimension of augmented state data matrix Y, and the amount of sample data changes from m to m-b+1 accordingly. Obviously, compared with the original state parameter data X, the augmented state data matrix Y obtained by simultaneously obtaining the current and previous state information of the unit through the sliding window includes both the current state parameter information of the unit and the state information of the previous time.

In the training process, DAE is trained layer by layer unsupervised. In the pre-training process, errors are propagated backwards and network parameters are continuously fine-tuned and optimized. The optimization goal of traditional stack noise reduction self-encoder is to optimize the end-to-end between original data D and destroyed data d0. This optimization method needs more features and requires a large amount of computation. But from the point of view of residual, the original function mapping becomes D and d+n, n denotes the noise in the data, the target of DCbS algorithm is the mapping between D and n, and the error function is changed to:

$$R_H = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{2}\|n\|^2 + R'_H \tag{4}$$

In order to prevent the over-fitting of the model, the heavy attenuation term (regular term) is introduced to prevent the over-fitting of the model, and the weight attenuation term is used in the formula.

$$R'_H = \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\mathbf{W}_{ji}^{(l)}) \tag{5}$$

λ is the weight attenuation coefficient, λ = 3e-3; $\mathbf{W}_{ji}^{(l)}$ represents the weight parameters connecting the jth neuron in layer L and the I I neuron in layer L + 1; s1, S2 and S3 are the number of nodes in the first, second and third layers respectively, that is, the number of nodes in the input layer, hidden layer and output layer of the network.

B. Model Description

Because the abnormal state data of power information and communication assets mainly comes from two aspects: one is the abnormal data record, which is the object to be cleaned by DCbS algorithm, the other is the abnormal operation data caused by the abnormal state of assets. Therefore, in the training process, the sample error and error duration are taken as the criteria for judging the data type. The maximum error function is preset as R_{MAX} and the maximum error duration is T_{MAX}. The data whose error exceeds R_{MAX} and the error duration is less than T_{MAX} are judged as a kind of abnormal data. These data are abnormal data Point. The data with errors over R_{MAX} and long-term zero or a specific value are classified as two kinds of abnormal data. This data is missing. If the errors exceed RMAX and there are some variation rules in the data, three kinds of abnormal data are determined. This data is abnormal data of equipment status. DCbS algorithm cleaning target is one or two kinds of abnormal data, the specific flow is as follows:

Step 1: DCbS model is trained with normal data, and the initial model and parameters are obtained.

Step 2: Calculate the error loss from the initial model, estimate the nuclear probability density, and get the R_{MAX}^1 and T_{MAX}^1 of the initial model.

Step 3: The model is trained with the data set containing fault cases to get the fault data model, and the error loss function is calculated to get further R_{MAX}^2 and T_{MAX}^2 .

Step 4: Use the model to test the test data, identify the abnormal data types, repair and reconstruct the first and second types of abnormal data; For the third type of abnormal data, input the fault data model again to identify, identify the abnormal data, belong to the third type of abnormal data, do not go anywhere. The rest are repaired and reconstructed.

Step 5: Integrate the revised data obtained in step 4, iterate the repaired data of the fault data model into the repaired data of the initial model, and the reconstructed data of the lossless data and the singular value constitute the final valid data

IV. USING THE CASE

The example uses the state detection information of some substation switches from 2014 to 2015 as training and testing data. In the test data, noise is artificially added to the index of switch throughput to detect the performance of the algorithm. The initial model is obtained by selecting normal operation state training, and the fault data model is obtained by selecting the real information including abnormal operation state for training. Among them, the input layer of the model is 80, including three hidden layers, which are 70, 50 and 70, the learning cycle is 1000, and the number of training samples is 10000 groups.

TABLE I. DATA CLEANING RESULTS of DCbS

Number	True value（Mbps）	Pollution value（Mbps）	Correction value（Mbps）	Whether to amend	异常类型
1	18.85	-	19.43	-	-
2	35.63	-	33.64	-	-
3	43.82	-	41.75	-	-
4	20.64	-	19.54	-	-
5	27,87	-	29.01	-	-
6	63,28	-	62.19	-	-
7	59.45	0	56.82	Y	Data missing
8	34.56	0	37.82	Y	Data missing
9	0.54	-	1.45	N	Abnormal state
10	1.09	-	1.43	N	Abnormal state
11	0.03	-	0.72	N	Abnormal state
12	2.31	-	3.12	N	Abnormal state
13	39.26	0.12	31.53	Y	singular value
14	64.34	643.4	69.19	Y	singular value

Table 1 shows some data sets cleaned after adding noise. It can be seen that the data cleaned by DCbS algorithm is basically consistent with the true value, and has good recovery ability for polluted data. The deviation between the processed lossless data and the original data is kept within 5%. For the abnormal data of equipment status, it can also be well recognized. After processing the abnormal data of equipment status, it still keeps its characteristics, which is convenient for the follow-up processing. For the first and second types of data outliers, DCbS algorithm also maintains good recognition characteristics, and carries out prediction recovery. The data of change 7 and 8 are missing data, and the real data are artificially nullified. DCbS algorithm restores the device to the true value level, which is 3.26 difference from the maximum error of the true value. The data numbered 13 and 14 are artificial singularities, i.e. the data produces large deviations and lasts for a short time. DCbS algorithm is also ideal for the recovery of singular values. It can be seen that the DCbS algorithm is used to clean the large data of power information and communication assets, and the real data distribution characteristics of the cleaning results load are obtained.

V. CONCLUSION

In this paper, a data cleaning algorithm based on SDAE (DCbS) of stack self-encoder is proposed, which saves the short-term correlation between data through sliding window and residual analysis between noisy data and non-destructive

data, in order to reduce the training data needed by the model to identify abnormal data points. In view of the abnormal operation of power information and communication assets, this method can effectively filter interference data.

The experimental results show that the proposed algorithm improves the ability to distinguish and recover outliers. Finally, the advantages of the algorithm are highlighted from two aspects: data recovery and outlier recognition.

REFERENCES

[1] Li Xuelong,Gong Haigang.A survey on big data systems[J].SCIENTIA SINICA Technologica,2015,45(01):1-44.

[2] Mayerschönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think.[J]. Mathematics & Computer Education, 2014, 47(17):181-183

[3] Cheng Xueqi,Jin Xiaolong,Wang Yuanzhuo,et.al.Survey on big data system and analytic technology [J]. Journal of Software, 2014, 25(09): 1889-1908.

[4] 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):1-12(in Chinese).

[5] Stimmel C L. Big Data Analytics Strategies for the Smart Grid[M]. Auerbach Publications, 2014.

[6] Wang Dewen,Sun Zhiwei.Big Data Analysis and Parallel Load Forecasting of Electric Power User Side[J].Proceedings of the CSEE, 2015, 35 (03): 527-537.

[7] Zhao Teng,Zhang Yan,Zhang Dongxia.Application Technology of Big Data in Smart Distribution Grid and Its Prospect Analysis[J].Power System Technology,2014,38(12):3305-3312.

[8] Zhang D, Xin M, Liu L, et al. Research on Development Strategy for Smart Grid Big Data[J]. Proceedings of the Csee, 2015, 35(1):2-12.

[9] Qu Chaoyang, Zahng Yijing,Wang Yongwen,et al.Big energy data cleaning model for energy internet based on Spark framework [J]. Electrical Measurement & Instrumentation, 2018, 55 (02): 39-44.

[10] Dai Jiejie,Song Hui,Yang Yi,et al.Cleaning Method for Data of Power Transmission and Transformation Equipment Based on Stacked Denoising Autoencoders[J]. Automation of Electric Power Systems,2017,41(12):224-230.

[11] Gu Y, Jiang H, Zhang Y, et al. Knowledge discovery for smart grid operation, control, and situation awareness — a big data visualization platform[C]// North American Power Symposium. IEEE, 2016:1-6.

[12] Ren X, Dai G, Geng Z. Research on Graph of Power Assets Based on NEO4J[J]. Power System & Clean Energy, 2017.

[13] Xu F, Zheng H,Jiang H,et al. Cost-Effective Cloud Server Provisioning for Predictable Performance of Big Data Analytics[J]. IEEE Transactions on Parallel and Distributed Systems, 2018:1-1.

[14] Zhang Yadi,Wang Hongjie,Zhou Hong. The design and application research of visualization system on data assets in power grid system base[J] Electrical Measurement & Instrumentation,2018,55(07):41-46.

[15] Wang X, Yang L T, Xie X, et al. A Cloud-Edge Computing Framework for Cyber-Physical-Social Services[J]. IEEE Communications Magazine, 2017, 55(11):80-85.

[16] Lin Y T , Huang S J . The Design of a Software Engineering Lifecycle Process for Big Data Projects[J]. IT Professional, 2018, 20(1):45-52.