Load Quality Analysis and Forecasting for Power Data Set on Cloud Platform

Jixiang Gan, Qi Liu(✉), and Jing Zhang

School of Computer and Software, Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
{20201249454,qi.liu,20191221030}@nuist.edu.cn

Abstract. In the era of big data, The prediction management system combined with cloud computing platform can start from massive structured, semi-structured and unstructured data, which has a positive impact on improving the compliance quality analysis and prediction of power data sets. This paper focuses on the characteristics of all kinds of data sets needed in the research of power demand side business process of cloud platform at home and abroad, and analyzes, compares and summarizes all kinds of data sets. First, this paper analyzes the problems existing in various common data sets, and expounds the methods to improve the quality of data sets from two aspects of data cleaning and data preprocessing. Secondly, the LSTM prediction model and ARIMA prediction model are used to predict and analyze the collected power data to judge whether the data set has obvious defects in advance. Finally, through the experimental comparison of the two models, a more efficient prediction model is analyzed.

Keywords: Data quality · Load forecasting · Data cleaning · Cloud platform

1 Introduction

Under the background of "energy Internet plus new electricity reform", the power technology represented by smart grid is constantly integrating with information technology represented by cloud computing and big data, changing the way of production and operation of power enterprises [1, 2]. Through the combination of "cloud platform + big data". On the one hand, the established big data management system can not only coordinate the data management within the enterprise, but also make use of the existing data processing technology to clean and preprocess the relevant data, ensure the quality of data, and improve the efficiency of data utilization. On the other hand, cloud platform technology can be used to share various data and improve data utilization [3].

Cloud is what we usually call network or Internet, and cloud platform is a space for data storage and operation (such as network disk, micro disk, etc.) with the help of network or Internet. It is abbreviated as CMP (cloud management platform), which is a third-party platform derived from the development of cloud computing. Using the convenient and fast characteristics of the cloud platform, the staff can use their own files in

M. R. Khosravi et al. (Eds.): CloudComp 2021, LNICST 430, pp. 3–16, 2022.
https://doi.org/10.1007/978-3-030-99191-3_1

different network environments or different computers, providing convenient channels and space for enterprises. Since the development trend of cloud platform application technology in China [4]. The data quality in the big data management system established by "cloud platform + big data" has become a key and difficult point. Improving the quality of process data can improve the value of data in use, reduce the cost and risk in the later stage of the project, and improve the accuracy of real-time information analysis [5]. At the same time, the cloud platform can further concentrate on information processing, transmission, storage and other links, improve the speed, and make it more convenient to access resources. Therefore, the factors affecting data quality in cloud platform big data have become one of the concerns of major enterprises to improve big data management system.

With the continuous development of smart grid construction, as an important way of power consumption information acquisition and control, power consumption information acquisition system produces more and more big data. In the face of massive power consumption data, how to ensure the quality of data and how to improve the quality of relevant power data in the cloud platform big data management system has become the current research hotspot [6]. For example, in the process of power data collection, through the use of intelligent terminal devices such as smart meters, the operation data of the whole power system can be collected, so as to realize the authenticity and accuracy in the data collection stage, and then the collected power big data can be processed and analyzed systematically, so as to realize the timeliness of data. Finally, combined with big data analysis and power system model, it can diagnose, optimize and forecast the operation of power grid, and provide guarantee for safe, reliable, economic and efficient operation of power grid [7–9]. Therefore, to achieve reasonable business process management, we need to have enough perfect data sets to improve the quality of process data.

In order to obtain accurate and perfect data sets, in the power business process management of cloud platform big data management system, each scholar collected and established different power data sets. In "Non-intrusive load monitoring and load disaggregation using transient data analysis" by Sachin Kumar Jain in 2018, the author collected the power consumption data of 20 types and 8 different appliances in the campus through the sensor acquisition device, with the sampling frequency of 12800 Hz and the duration of about 5 h each time, After noise removal and unified data format preprocessing, it is used as the experimental data set [10]. Meanwhile, in 2020, Seongbae Kong and others developed a set of power data acquisition system in "home appliance load disaggregation using cepstrum-smoothing based method", which is used to obtain the characteristic signals of electrical appliances [11]. But whether it is sensor acquisition device or power data acquisition system, compared with the existing public data set, although the data they collected is more targeted than the public data set, their sampling accuracy and duration are not perfect, and for the existing monitoring model, it is not as universal as the current public data set.

2 Analysis and Improvement of Power Data

2.1 Classification and Comparison of Common Power Data Sets

With the improvement of smart meters and the popularity of data management system, a large number of power data acquisition, storage and management have been realized. At the same time, with the mature application of big data analysis, machine learning and other technologies, the information value contained in smart meters can be better mined. The authenticity, accuracy and integrity of power data set are guaranteed. At present, the more popular power data sets mainly include REDD data set, BLUED data set and UK-DALE data set [12]. The detailed information table of REDD data set, BLUED data set and UK-DALE data set is shown in Table 1.

REDD Data Set. REDD data set contains data of six American families in different time periods. REDD data set provides both high and low frequency household power data. The main advantage of REDD data set is that its data volume is rich enough to adapt to the large network training model [13]. REDD data set is relatively clean compared with other data sets, and good results can be achieved without post-processing [14].

BLUED Data Set. The BLUED data set contains data collected by an American family in about 8 days. The BLUED data set provides high-frequency household electric power data with a frequency of 12 kHz. The data set also provides an event list when the household appliance changes state [15]. The advantage of BLUED data set is that it provides equipment power data with time axis, which is conducive to use in evaluation algorithm to confirm decomposition results [16]. The main disadvantage of BLUED data set is the high redundancy of high frequency data.

UK-DALE Data Set. The UK-DALE data set contains two to four years' electricity data of six UK households. UK-DALE data set provides both high-frequency household power data with sampling frequency of 16 kHz and an independent low-frequency household power data [17]. The main advantage of UK-DALE data set is that it can provide a long duration of data and meet the needs of monitoring at different times of the year. The main disadvantage of UK-DALE dataset is that due to the large number of data samples in UK-DALE dataset, the data classification is unbalanced. Finally, the accuracy of the experiment is affected.

The analysis shows that REDD data set, BLUED data set and UK-DALE data set have obvious differences in collection area, data duration, collection frequency and data volume. According to these different attributes, the application scenarios of different data sets are also different. However, due to the reasonable collection time and diverse collection frequency of REDD data set, REDD data set has gradually become the standard data set for benchmark test of decomposition prediction algorithm [18]. At the same time, REDD data set, blue data set and UK-DALE data set play an important role in the implementation and management of power business process.

2.2 The Function of Electric Power Data Set

A reasonable power data set can not only analyze in the power data system of cloud platform and formulate a reasonable management process, but also predict and plan the

Table 1. Detailed information of REDD data set, BLUED data set and UK-DALE data set

Dataset name	Acquisition frequency	Collection area	Duration	Collection of house types	Number of equipment
REDD dataset	16500 Hz/1 Hz	America	A few months	Residential district	6 houses
BLUED dataset	12000 Hz	America	8 days	Residential district	1 house
UK-DALE dataset	16000 Hz/1 Hz	Britain	2–4 years	Residential district	6 houses

future schedule, resource allocation and process management. In the whole life cycle of business process management, reasonable data sets play different roles in different stages. In the business process evaluation stage, the appropriate business process is defined by analyzing the data set. In the business process design and analysis stage, based on the actual data set, using the defined business process, the business process content is modeled and the modeling model is verified and analyzed. In the business process configuration stage, the real data set is used to configure the analyzed business process and deploy the related process. In the business process implementation stage, the configured business process is properly implemented and monitored with effective data set [19–23]. Business process management flow chart, as shown in Fig. 1. Therefore, in order to improve the efficiency of business process management, it is essential to improve the data quality of relevant data.

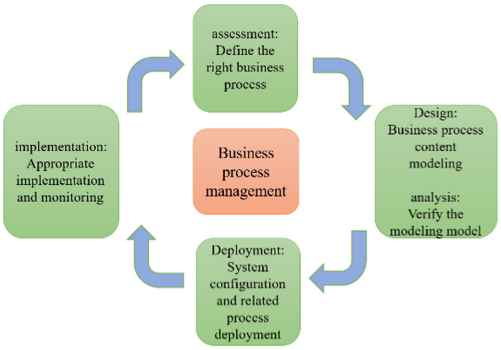

Fig. 1. Business process management flow chart

2.3 Methods of Improving Data Quality

After big data construction, most enterprises have achieved good data construction results and accumulated considerable amount of data. However, it is a great risk to use these data directly. Therefore, it is necessary to analyze the data before processing to improve the data quality [24]. There are many ways to improve the quality of data, and the

academia generally starts to consider data cleaning and data preprocessing [25]. From the perspective of data cleaning, improving the quality of data mainly from three perspectives: box method, clustering method and regression method. From the perspective of data preprocessing, there are two ways to improve data quality: standardization and normalization.

Data Cleaning. There are three methods for data cleaning, which are box division, clustering and regression [26].

(1) *Separate Box Method*. Due to the complex characteristics of data quality management, the data to be processed is put into the box according to certain rules, and then the data in each box is tested. also, the data is processed by means of the actual situation of each box in the data.
(2) *Clustering*. It is to group abstract objects into different sets, find unexpected outliers in the collection, which are noise. This allows you to detect noise directly and then clear it.
(3) *Regression Method*. Data regression method is to use the data of the function to draw the image, and then smooth the image. There are two regression methods, one is single linear regression, the other is multiple linear regression. Single linear regression is to find out the best line of two attributes, which can predict another attribute from one. Multilinear regression is to find many attributes, and then fit the data to a multi-dimensional surface, which can eliminate the noise.

Data Preprocessing. Data preprocessing is mainly divided into normalization and standardization. Normalization is to change the number into a decimal between (0, 1), which is mainly proposed for the convenience of data processing. It is more rapid and convenient to map the data to the range of 0–1. Standardization is to scale data to a specific range. Standardized data is more conducive to the use of the properties of standard normal distribution. The common methods of standardization and normalization are z-sorce (zero mean) method and min max method [27].

(1) *Z-sorce*. Zero mean normalization is also called standard deviation normalization. The mean value of processed data is 0 and the standard deviation is 1. The formula is as follows: Where, $\bar{x}$ is the mean value of the original data, σ is the standard deviation of the original data.

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{1}$$

(2) *Min-Max*. Min max normalization, also known as discrete normalization, is a linear transformation of the original data, mapping the data values between [0, 1]. The formula is as follows:

$$x^* = \frac{x - min}{max - min} \tag{2}$$

Discrete standardization retains the relationship existing in the original data, and is the simplest method to eliminate the influence of dimension and data value range. The

disadvantage of this method is that if the values are concentrated and a certain value is large, the normalized values are close to 0.

3 Data Prediction

In practical application, it is very important to ensure the data quality of the application system and to supervise and predict the data quality of the future period. The premise of the guarantee is the accurate prediction and evaluation of the time series data. According to the historical series data of the time series data, the series value of the future period can be predicted. We can learn from the common time series prediction methods and models. ARIMA model (autoregressive integrated moving average model) and LSTM model (long short term memory) have good case results in forecasting [28–30]. This paper also uses the above models to forecast and analyze the load power data, and reviews whether the data set meets the data requirements of the application system through the predicted data, Judge whether the quality of follow-up data is qualified.

3.1 ARIMA Model

Arima is a typical time series model, which consists of three parts: AR model (autoregressive model), MA model (moving average model), and difference method. Therefore, ARIMA (p, d, q) is called autoregressive moving average model [31].

Autoregressive Model (AR). The autoregressive model describes the relationship between the current value and the historical value, and forecasts itself with the historical time data of the variable itself. In the autoregressive model, the order P should be determined first, which means that the current value can be predicted by the historical value of several periods. The formula of p-order autoregressive model is defined as:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \in_t \tag{3}$$

Moving Average Model (MA). The moving average model focuses on the accumulation of error terms in the autoregression model. The formula of q-order autoregression process is defined as follows:

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \in_{t-i} + \in_t \tag{4}$$

By combining autoregressive model, moving average model and difference method, ARIMA (p, d, q) is obtained, where D is the order of data difference.

Determine the Parameters of ARIMA Model. The identification and order determination of the model are mainly to determine the three parameters p, d and q. the order D of the difference is generally obtained by observing the graph. The determination of P and Q is mainly obtained by autocorrelation function ACF (auto correlation function) and partial autocorrelation function PACF (partial auto correlation function).

Autocorrelation Function (ACF). Correlation: compared with the ordered random variable sequence, the autocorrelation function reflects the correlation of its own data in the same sequence in different time series. The formula of autocorrelation function is as follows:

$$ACF(k) = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \tag{5}$$

Autocorrelation function $ACF(k) = \rho_k$ The value range of K is $[-1, 1]$. The definition of this value range is as follows, -1 is negative correlation, $+1$ is positive correlation and 0 is no correlation.

Partial Autocorrelation Function (PACF). The partial autocorrelation function is used to measure the effect of temporary adjustment of all other short lag terms ($y_t - 1, y_t - 2, \ldots,$ $y_t - k - 1$). In time series, K time units (y_t and $y_t - k$). It is the linear correlation between the time series observations and the expected past observations given the intermediate observations.

ARIMA Model Overall Process

- Check the data for stability, and if it is not stable, carry out differential processing
- White noise test of stationary post sequence
- ACF function and PACF function are used to determine the order of sequence meeting the requirements
- Test the accuracy of the model
- Using model prediction to forecast and analyze data

3.2 LSTM Model

LSTM network is a special type of RNN, which can learn long-term dependent information. Of course, LSTM and baseline RNN are not very different in structure, but they use different functions to calculate the "hidden" state. LSTM can avoid the problem of long-term dependence, and the central idea is cell state. The network consists of various gates. Through the sigmoid function and point multiplication operation to complete the creation of a door. RNN is a repetitive single neural network layer, while the repetitive module in LSTM contains four interactive layers, three sigmoid layers and one tanh layer, and interacts in a very special way [32–34]. The network structure of LSTM is shown in Fig. 2.

Forgetting Gate. LSTM decides which part of the original cell state to delete. The decision is made through a structure called the forgetting gate. The forgetting gate reads the last output h_{t-1} and the current input x_t. Do a sigmoid nonlinear mapping, and then output a vector f_t. The value of each dimension of the vector is between 0 and 1. 0 means to give up completely and 1 means to keep. Finally, it is related to cell state C_{t-1}. The screening formula of forgetting gate is as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{6}$$

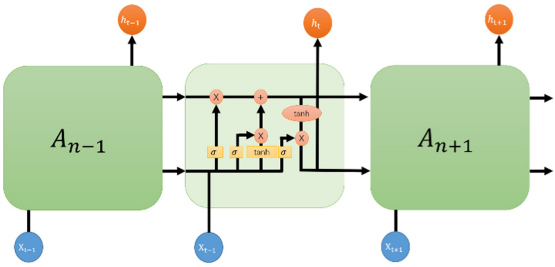

Fig. 2. Network structure of LSTM

Among them, σ Is the sigmoid function, W_f is the parameter of linear relation coefficient, b_f is the bias parameter.

Input Gate. In order to determine the required information of the cell state, firstly, the input information needs to be filtered by sigmoid function, and then the new update state is created by tanh function $\widetilde{C}_t$. The update formula of the input door is:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{7}$$

$$\widetilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{8}$$

Cellular State. Cell state is to update the old cell state, discard part of the old cell information and add new information. C_{t-1} to C_t.New state C_t is from the old state C_{t-1} and f produced by forgetting gate f_t to determine the information to be updated, and then add the product of the parameters generated by the input gate: $i_t * \widetilde{C}_t$. The change formula of cell state is as follows:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{9}$$

Output Gate. The final output is achieved by fitting sigmoid function and tanh function. A sigmoid function to determine which part of the cell's state will be output. Then, the cell state is processed through the tanh layer (a value between -1 and 1 is obtained) and multiplied by the output of the sigmoid function to determine the part of the output. The related formula of output gate is as follows:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{10}$$

$$h_t = o_t * \tanh(C_t) \tag{11}$$

3.3 Experimental Results and Evaluation

Dataset Description. The data set used to verify the data quality is the power data of a room in October, and the sampling rate is one hour. ARIMA model and LSTM model were established respectively. The two models established in this paper are implemented in Python language, and the pandas, numpy, statsmodel and Matplotlib libraries are downloaded for use. Table 2 shows the partial power data of a single room in October 2020.

Table 2. Partial power data of single room in October 2020.

Time	Power consumption (KWH)	Time	Power consumption (KWH)	Time	Power consumption (KWH)
2020100100	670718	2020100110	608067	2020100120	568008
2020100101	648975	2020100111	639951	2020100121	354354
2020100102	635599	2020100112	663172	2020100122	435257
2020100103	536269	2020100113	669666	2020100123	360796
2020100104	356870	2020100114	674347	2020100200	674078
2020100105	395430	2020100115	683242	2020100201	652474
2020100106	420034	2020100116	692810	2020100202	643138
2020100107	477297	2020100117	711472	2020100203	536794
2020100108	528439	2020100118	757839	2020100204	383872
2020100109	573939	2020100119	783315	2020100205	389773

Time Series Power Data Prediction Based on ARIMA Model. After the ADF stationarity test and judging whether the sequence is a stationary sequence, the p-value of the sequence is $6.356402463960776 * 10^{(-14)} < 0.05$, and the stationary sequence is obtained. The white noise test of the stationary sequence shows that $0.0027 < 0.05$ is a non white noise sequence. The order of ARIMA model is determined by observing the autocorrelation diagram and partial autocorrelation diagram of the stationary sequence, as shown in Fig. 3. Finally, the values of P and Q are 2 and 4 respectively, which are determined as ARIMA (2, 1, 4) according to the rule of thumb.

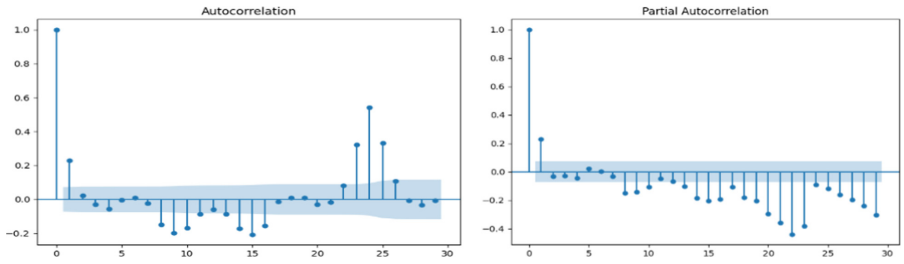

Fig. 3. Autocorrelation graph and partial autocorrelation graph

Through data validation, this paper establishes ARIMA (2, 1, 4) time series model as a model of load forecasting power data. After the model is determined, the data in the next 250 days are predicted. The prediction results are shown in Fig. 4.

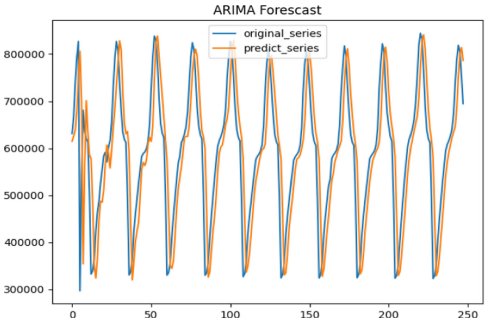

Fig. 4. Data prediction results of ARIMA model in the next 250 days

Time Series Power Data Prediction Based on LSTM Model. Due to the small amount of data, the number of iterations epochs is set to 50, batch_size is 1. In order to further prevent over fitting, dropout technique is used in model training [35]. The loss rate of the experimental model is set to 0.2, that is, 20% of the nodes are discarded in each round of weight update. The loss function is MSE, the activation function is sigmoid, and the optimizer parameter is Adam. The input data also use ADF stationary test and white noise test. Finally, the data of the next 250 days are also predicted, and the prediction results are shown in Fig. 5.

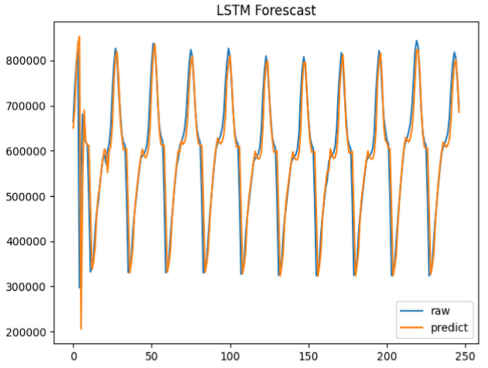

Fig. 5. Data prediction results of LSTM model in the next 250 days

Comparison of Prediction Results of Different Models. Through the comparison of the prediction results in Fig. 4 and Fig. 5, we can intuitively see the effect comparison of the two models, in which the abscissa is hour and the ordinate is kilowatt hour. It can be seen from Fig. 4 that the lag of ARIMA model is serious, and the prediction offset is large in the first 50 h, while the LSTM model shown in Fig. 5 has no serious lag, and the offset is small in the first 50 h. This also illustrates the outstanding performance of LSTM model in single prediction model.

This paper also uses two representative evaluation error standards to evaluate the accuracy of prediction, namely root mean square error (RMSE) and mean absolute deviation (MAE). Their mathematical expressions are as follows:

$$MSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (F_i - R_i)^2} \tag{12}$$

$$MAE = \sum_{i=1}^{n} \frac{|F_i - R_i|}{n} \tag{13}$$

Among them, F_i is the i-th predicted value, R_i is the ith real value, and N is the sequence length (number of sample points). The smaller the RMSE and Mae, the smaller the prediction error and the higher the accuracy of the model, the better the fitting ability of the model. Table 3 shows the RMSE and Mae values of the prediction results of each model. It can be seen from Table 3 that LSTM model is lower than ARIMA model in all indexes, and the difference is large. LSTM model performs better than ARIMA model in time series prediction.

Therefore, LSTM model has the highest prediction accuracy and the best performance, and can be better applied to the prediction and analysis of data quality.

Table 3. Comparison of prediction results of different models

Model	RMSE	MAE
ARIMA	0.96154	0.74402
LSTM	0.59511	0.32526

4 Conclusion

The load data quality analysis of cloud platform big data can be divided into process evaluation, process design and analysis, process deployment and process implementation. This paper starts with the process implementation, analyzes the process data quality related problems in process implementation management and the ways to improve the data quality. Taking the power business process execution management as an example, the paper analyzes the shortcomings of the existing power data sets: (1) In terms of data scale, the mainstream data sets can provide large-scale household power data, but in large-scale data, the data classification is very unbalanced, which will seriously lead to the accuracy of the experiment. (2) In terms of data frequency, high frequency data is prone to redundant data, while low frequency data has a long sampling period and data collection is slow. (3) In terms of data duration and distribution, they are different, and the big data sets are not unified. These aspects seriously affect the quality of process data in the implementation and management of power process. In view of this kind of problem, and summarized the methods to improve the data quality. From the load data itself, the paper respectively improves the final data quality from two methods: data cleaning and data preprocessing. Finally, there are common methods of data prediction, and compare

them on the data set collected by ourselves. The advantages and disadvantages of the prediction model are compared through experiments. The results of future prediction are analyzed, and the advantages and disadvantages of data sets are compared. Thus, the quality problems of the dataset can be predicted in advance and the business risks caused by the data quality will be reduced. In cloud platform data management, data quality is an important part. Therefore, how to efficiently process the data sets to be used in business processes and how to improve the data quality of these data still need to be explored and discovered constantly.

Acknowledgements. This work has received funding from the Key Laboratory Foundation of National Defence Technology under Grant 61424010208, National Natural Science Foundation of China (No. 41911530242 and 41975142), 5150 Spring Specialists (05492018012 and 05762018039), Major Program of the National Social Science Fund of China (Grant No. 17ZDA092), 333 High-Level Talent Cultivation Project of Jiangsu Province (BRA2018332), Royal Society of Edinburgh, UK and China Natural Science Foundation Council (RSE Reference: 62967_Liu_2018_2) under their Joint International Projects funding scheme and basic Research Programs (Natural Science Foundation) of Jiangsu Province (BK20191398 and BK20180794).

References

1. Hengjing, H., Wei, Z., Songling, H., et al.: Research on the application of cloud computing in power user electric energy data acquisition system. Electr. Measur. Instrum. **53**(1), 1–7 (2016)
2. Junwei, C., Zhongda, Y., Yangyang, M., et al.: Survey of big data analysis technology for energy internet. South. Power Syst. Technol. **9**(11), 1–2 (2015)
3. Yao, Y.: The construction of comprehensive budget management in Colleges and Universities under the environment of "big data + cloud platform" – Taking D University as an example. Friends Acc. **01**, 119–124 (2020)
4. Xiao, B., Wang, Z., Liu, Q., Liu, X.: SMK-means: an improved mini batch K-means algorithm based on Mapreduce with big data. Comput. Mater. Continua **56**(3), 365–379 (2018)
5. Neubauer, T., Stummer, C.: Extending business process management to determine efficient IT investments. In: Proceedings of the 2007 ACM symposium on Applied computing (SAC 2007), pp. 1250–1256. Association for Computing Machinery, New York (2007)
6. Huang, A.Q., Crow, M.L., Heydt, G.T., et al.: The future renewable electric energy delivery and management (FREEDM) system: the energy internet. Proc. IEEE **99**(1), 133–148 (2010)
7. Wang, Y., Chen, Q.X., Hong, T., Kang, C.Q., et al.: Review of smart meter data analytics: applications, methodologies, and challenges. IEEE Trans. Smart Grid **10**(3), 3125–3148 (2019)
8. Wang, K., Xu, C., Zhang, Y., Guo, S., Zomaya, A.Y., et al.: Robust big data analytics for electricity price forecasting in the smart grid. IEEE Trans. Big Data **5**(1), 34–45 (2017)
9. Tao, W., Xiaolei, W., Rui, Y., et al.: Research on power energy big data acquisition and application based on big data cloud platform. Electron. World **15**, 155–156 (2020)
10. Zhang, J., Liu, Q., Chen, L., Tian, Y., Wang, J., et al.: Non-intrusive load management based on distributed edge and secure key agreement. Wirel. Commun. Mob. Comput. (WCMC) (2021)
11. Liu, Q., Kamoto, K.M., Liu, X., Sun, M., Linge, N., et al.: Low-complexity non-intrusive load monitoring using unsupervised learning and generalized appliance models. IEEE Trans. Consum. Electron. **65**(1), 28–37 (2019)

12. Dash, S., Sodhi, R., Sodhi, B., et al.: An appliance load disaggregation scheme using automatic state detection enabled enhanced integer-programming. IEEE Trans. Ind. Inf. **17**, 1176–1185 (2020)
13. Kolter, Z.J., Redd, J.M.J., et al.: A public data set for energy disaggregation research. In: Proceedings of the in Workshop on Data Mining Applications in Sustainability (SIGKDD), pp. 59–62, San Diego, CA, USA (2007)
14. Liu, Q., Lu, M., Liu, X., Linge, N., et al.: Non-intrusive load monitoring and its challenges in a NILM system framework. Int. J. High Perform. Comput. Netw. **14**(1), 102–111 (2019)
15. Anderson, K., Ocneanu, A., Benitez, D., Carlson, D., Rowe, A., Berges, M.: BLUED: a fully labeled public dataset for event-based non-intrusive load monitoring research. In: Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD), Beijing, China (2012). Young, M.: The Technical Writer's Handbook. University Science, Mill Valley (1989)
16. Liu, Q., Li, S., Liu, X., Linge, N.: A method for electric load data verification and repair in a home energy management environment. Int. J. Embed. Syst. **10**(3), 248–256 (2018). https://doi.org/10.1504/IJES.2018.091788
17. Kelly, J., Knottenbelt, W.: The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci. Data **2**, 1–14 (2015). https://doi.org/10.1038/sdata.2015.7.150007
18. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A., et al.: Fundamentals of Business Process Management (2013)
19. Mathias, W.: Business Process Management. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28616-2
20. Jeston, J., Nelis, J., et al.: Business Process Management: Practical Guidelines to Successful Implementation. Routledge, London (2008)
21. Becker, J., Kugeler, M., Rosemann, M., et al.: Process management: a guide for the design of business processes. Springer Publishing Company, Heidelberg (2011). https://doi.org/10.1007/978-1-4302-3645-0_15
22. Rosemann, M., Brocke, J., et al.: The six core elements of business process management. Handbook on Business Process Management, vol. 1, pp. 107–122 (2010). https://doi.org/10.1007/978-3-642-00416-2
23. Ahmad, T., Looy, A.V., et al.: Business process management and digital innovations: a systematic literature review. Sustainability **12**(17), 6827 (2020). https://doi.org/10.3390/su12176827
24. Pipino, L.L., Lee, Y.W., Wang, R.Y., et al.: Data quality assessment. Commun. ACM **45**(4), 211–218 (2002). https://doi.org/10.1145/505248.506010
25. Jingyu, H., Lizhen, X., Yisheng, D., et al.: Review of data quality research. Comput. Sci. **02**, 1–5+12 (2008)
26. Ilyas, I.F., Chu, X.: Data Cleaning. Association for Computing Machinery, New York (2019)
27. Gupta, V., Hewett, R.: Adaptive normalization in streaming data. In: Proceedings of the 2019 3rd International Conference on Big Data Research (ICBDR 2019), pp. 12–17. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3372454.3372466
28. Shiguang, P., Xianhui, G., et al.: Prediction of China's soybean import volume and import volume based on ARIMA and GM (1, 1) models. Soybean Sci. **39**(4), 626–632 (2020)
29. Yurou, C.: Short term prediction of China's core CPI based on ARIMA model. Time Honor. Brand Mark. **7**, 37–38 (2020)
30. Zijian, H., Yuanhua, L., et al.: Application of long-term and short-term memory model in stock price trend prediction. Prod. Res. **1**, 36–39 (2020)
31. Box, G.E.P., Pierce, D.A., et al.: Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J. Am. Stat. Assoc. **65**(332), 1509–1526 (1970)

32. D'Informatique, D.E., Ese, N., Esent, P., et al.: Long short-term memory in recurrent neural networks. EPFL (2001)
33. Yamak, P.T., Li, Y., Gadosey, P.K.: A comparison between ARIMA, LSTM, and GRU for time series forecasting. In: ACAI 2019: 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence 92019)
34. Wang Xin, W., Ji, L., et al.: Fault time series prediction based on LSTM recurrent neural network. J. Beijing Univ. Aeronaut. Astronaut. **44**(4), 772–784 (2018)
35. Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)