Data Mining Technology in Detection and Identification of Bad Data in Power System

Honghai Wang(✉)

Anhui Sanlian College, Hefei 230601, Anhui, China
2858093479@qq.com

Abstract. Traditional bad data detection methods are estimated algorithms that require repeated state estimations. A large number of calculations may also cause "residual flooding" or "residual pollution" phenomena, which is the ideal state. The bad data can be detected and identified before the estimation, and the bad data detection and identification method based on association rule mining studied in this paper can solve these problems to a certain extent. This paper first analyzes the traditional bad data detection and identification methods and then leads to data mining technology. Second, it delves into the classic algorithm Apriori and improvement in association rules and studies the basic algorithm and improvement of periodic association rule mining. Application of improved algorithm. The current, active, and reactive power data of a certain line collected in the SCADA system of a dispatching center from May to September and five months were selected as sample data to finally verify the feasibility and effectiveness of the method.

Keywords: Data mining · Power system · Bad data · Detection and identification

1 Introduction

In order to meet the needs of the national economy, the scale of China's power grid is constantly expanding, and its structure and operation mode is becoming more and more complicated than before [1]. As the data acquisition and victory control system, the SCADA system has been widely used in power networks, the system may fail to measure or transmit data due to various force majeure factors during the process of measuring data or transmitting data. Abnormal, that is, bad data [2]. In order to improve the reliability of power system state estimation, and select and eliminate a small amount of bad data that occasionally appears in the SCADA system measurement sampling, many scholars at home and abroad have conducted in-depth research on bad data mining techniques. But looking at all kinds of methods, the accuracy, fastness, and comprehensiveness of the detection and identification of bad data are still big problems that plague electric power workers [3, 4].

At this stage, the protection and control system of the power grid has achieved a high degree of automation, which places higher requirements on the accuracy of the system data [5]. Obviously, once the data received by a substation automation system or dispatch

X. Yuan et al. (Eds.): ICUIA 2020, CCIS 1319, pp. 59–64, 2020.
https://doi.org/10.1007/978-981-33-4601-7_6

automation system is bad data, the impact of these error messages will interfere with the dispatcher's judgment and may cause the dispatcher to make wrong control decisions and even cause protection and control. The device malfunctioned, which seriously affected the safety of the power grid [6].

The focus of this paper is to obtain the association rules by mining historical data samples collected by the SCADA system when the topology and operating status of the power system network are unclear, to conduct research on the detection and identification of bad data before state estimation. It will provide a certain theoretical and practical basis for related fields, and contribute to the improvement of China's power system security.

2 Method

2.1 Data Mining

Data mining is not a random application of some existing or known analysis techniques to specific situations to solve specific problems [7, 8], but a way to solve problems and analyze problems. The whole process of data mining is shown in Fig. 1.

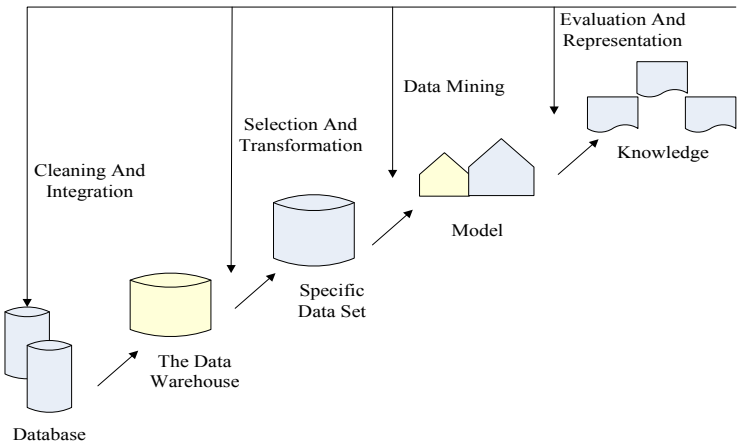

Fig. 1. Data mining process

2.2 Improvement of Association Rule Algorithm

This paper reduces the number of data subsets that need to be counted for the periodic support and proposes the CARM2 algorithm., Reducing the time complexity of the algorithm, the specific improvement steps are as follows:

Assume that the number of data subsets contained in period *(l, o)* is |*db (l, o)*|,

$$|db(l, o)| = |(n - o)/l| \tag{1}$$

The minimum periodic support is minCycle. If |*db (l, o)*| is divided into two parts of the data subset, they are:

$$|db(l, o)| \times (1 - \min Cycle) \tag{2}$$

$$|db(l, o)| \times \min Cycle - 1 \tag{3}$$

Then only the periodic support number of the association rules of the first part of the data subset can be counted, because the association rules of the second part of the data subset periodically do not meet the minimum periodicity support condition, so it cannot become periodic Association rules. On the other hand, assuming that the first m data subsets of |*db (l, o)*| have been calculated, if the periodic support number of an association rule in these m data subsets is less than:

$$m - |db(l, o)| \times (1 - \min Cycle) \tag{4}$$

Then this rule cannot become a periodic association rule.

Proof: Because the number of data subsets contained in period *(l, o)* is |*db(l, o)*|, assuming that the minimum periodic support specified by the user is *minCycle*, an association rule must become a periodic association rule. The number of periodic support must be at least $|db(l, o)| \times \min Cycle$. The first m data subsets of |*db(l, o)*| have been calculated, and |*db (l, o)*|-*m* data subsets remain, then in the first m data subsets that have been calculated This association rule appears at least:

$$|db(l, o)| \times \min Cycle - (|db(l, o)| - m) = m - |db(l, o)| \times (1 - \min Cycle) \tag{5}$$

Only then can this association rule become a periodic association rule and an improved CARM2 algorithm.

3 Experiment

3.1 Data

In this paper, the current, active, and reactive power data of a line collected in the SCADA system of a dispatch center from May to September and five months are used as sample data. Each daily active power, reactive power, and current data curve has 96 curves. Sampling point, the sampling interval is 15 min/time. It is known that the sample data used in this article has been trapped and cleaned up, and all are good data, and there is no missing in the middle.

3.2 Association Rule Mining

Because the selected historical data includes five months (150 days) of current and power distribution, the sampling interval is 15 min, and the time attribute is 96 timestamps per day, and the five months are divided into five periods of I-V according to the month, A total of 480 time units, the original database storage unit is shown in Table 1. Set the minimum support degree to 0.05, and perform periodic association rule mining on the data subset of each period to obtain the periodic frequent itemsets at each moment, and then to summarize the current and power distribution rules at that moment.

Table 1. Raw database storage unit

Sampling point on September 1st	Active	Reactive	Current
1	556.13	−32.47	653.93
2	548.02	−26.39	643.38
3	542.94	−26.39	648.07
4	549.03	−26.39	645.72
5	529.75	−31.46	623.46
6	503.36	−31.46	595.33
7	500.32	−30.45	592.99
8	505.39	−32.47	595.33
9	481.04	−35.52	569.55
10	489.15	−32.47	580.1

4 Discussion

4.1 Detection and Identification of Single Bad Data

Sample data of No. 2 is randomly selected, and the active power data of No. 10 sampling point is set as bad data. The active power data is increased by 10% based on the original

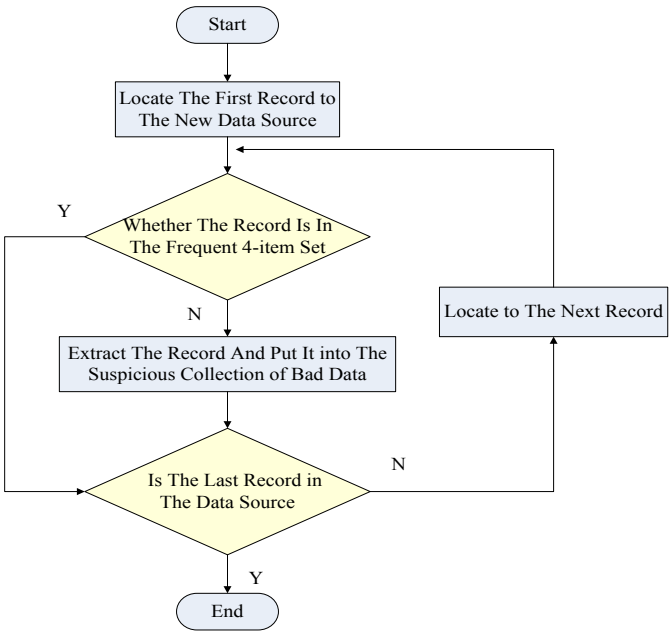

Fig. 2. Bad data detection process

normal data. It is known that the original data of this sampling point are: active power 489.15 MW, reactive power −32.47 MW, current 580.1A. The form after discretization is: T10, P7, Q6, 17. The point in active power increased by 10% to 538.06 MW, and the level changed from P7 to P3. All the modified sample data of No. 2 are discretized and stored as a new data source. The bad data detection process is shown in Fig. 2.

The test results are shown in Table 2. The results show that the data record (T10, P3, Q6, 17) was extracted into the suspicious collection of bad data. The record has a timestamp T10. It is obvious that the record can be identified. There is bad data in the 10th sampling record. The corresponding association rules obtained from the sample data mining in the previous section are as follows: T10 → P7, Q6, 17 (Sup = 0.17, Conf = 0.68), T10 → P7, Q6, 16 (Sup = 0.08, Conf = 0.32). It can be seen that there are only two cases of normal data at the 10th sampling point, so it can be determined that the active power of the record is bad data.

Table 2. Single bad data detection result

Suspect bad data set	Corresponding association rule
T10, P3, Q6, 17	T10 → P7, Q6, 17 (Sup = 0.17, Conf = 0.68)
	T10 → P7, Q6, 16 (Sup = 0.08, Conf = 0.32)

4.2 Detection and Identification of Multiple Bad Data

Sample data No. 18 was randomly selected, and the active power data at the 35th sampling point and the reactive power data at the 65th sampling point were set as bad data. One reduced the active power by 10% and the other reduced the reactive power by 20%. It is known that the original data of the 35th sampling point are: active power 527.72 MW, reactive power −52.77 MW, current 623.46A. After discretization: T35, P4, Q8, 14. Now reduce the active power by 10% to 474.95 MW, the grade becomes P8. The original data of sampling point 65 is known as: active power 553.09 MW, reactive power −33.49 MW, current 655.1A. After discretization: T65, P2, Q6, 12. Reduce the reactive power by 20% to −26.54 MW, and the grade becomes Q3. The test results are shown in Table 3. According to the table, the data records (T35, P8, Q8, 14) and (T65, P2, Q3, 12) were extracted into the suspicious set of bad data, and the 35th Bad data were present in the and 65th sampling records [9, 10].

Table 3. Multiple bad data detection result

Bad data set	Corresponding association rule
T35, P4, Q8, 14	T35 → P4, Q8, 14 (Sup = 0.25, Conf = 1)
T65, P2, Q3, 12	T65 → P2, Q6, 12 (Sup = 0.115, Conf = 0.46)
	T65 → P2, Q7, 12 (Sup = 0.135, Conf = 0.54)

5 Conclusion

In this paper, the association rules in data mining and the detection and identification of bad data in power systems are studied in-depth, and the association rules are introduced into the detection and identification of bad data. Detect and identify models to derive information with practical application value. The information obtained from the historical data of the power system using association rules helps to obtain the measured and predicted amount at each moment so that the decision has a scientific basis.

Acknowledgments. The Academic Funding Project for Outstanding Talents of Universities and Colleges (Professional) in Anhui Province in 2018 (Project Number: gxbjZD57).

References

1. Khan, Z., Razali, R.B., Daud, H.: Bad data detection in power system state estimation based on generalized likelihood ratio test. Int. J. Energy Stat. **04**(4), 1650016 (2016)
2. Deng, S., Zhou, A., Yue, D.: Distributed intrusion detection based on hybrid gene expression programming and cloud computing in cyber physical power system. IET Control Theory Appl. **11**(11), 1822–1829 (2017)
3. Jiang, X., Sheng, G.: Research and application of big data analysis of power equipment condition. High Volt. Eng. **44**(4), 1041–1050 (2018)
4. Zhou, W.: Research and application of data mining algorithm based on fuzzy neural network for nonlinear problems in large data environment. J. Comput. Theor. Nanosci. **13**(7), 4735–4738 (2016)
5. Falkenthal, M., Barzen, J., Breitenbücher, U.: Pattern research in the digital humanities: how data mining techniques support the identification of costume patterns. Comput. Sci. – Res. Dev. **32**(3–4), 1–11 (2016)
6. Fan, S.-K.S., Lin, S.-C., Tsai, P.-F.: Wafer fault detection and key step identification for semiconductor manufacturing using principal component analysis, AdaBoost and decision tree. J. Chin. Inst. Ind. Eng. **33**(3), 151–168 (2016)
7. Fatima, B., Ramzan, H., Asghar, S.: Session identification techniques used in web usage mining: a systematic mapping of scholarly literature. Online Inf. Rev. **40**(7), 1033–1053 (2016)
8. Yu, H., Du, Y., Ma, C.: Survey of compressed sensing technology for signal and data of power system. Yi Qi Yi Biao Xue Bao/Chin. J. Sci. Instr. **38**(8), 1943–1953 (2017)
9. Zhu, Y., Xing, N., Ji, Y.: Fault location algorithm of integrated data network for power system based on interactive active detection. Autom. Electr. Power Syst. **41**(4), 35–40 (2017)
10. Fernandes, E.R., Ghiocel, S.G., Chow, J.H.: Application of a phasor-only state estimator to a large power system using real PMU data. IEEE Trans. Power Syst. **32**(1), 1 (2016)