Research on Algorithm and experimental verification of product quality data cleaning based on artificial intelligence

Liangwen Yue, Zhaojun Wang
Beijing Sunway World Science & Technology Co., Ltd; Beijing China, 100070

***Abstract*—Aiming at the limitation of the research on product quality data cleaning at home and abroad, based on the theory of BP artificial neural network in the subject of artificial intelligence, this paper constructs the product quality data cleaning model and algorithm with the improved BP artificial neural network of L-M algorithm, and takes the refrigerator product quality data as an example, carries out the experimental verification of product quality data cleaning with the designed model. The experimental results show that the product quality data cleaning model given in this study is a kind of universal, scientific and reasonable product quality data cleaning model and algorithm, which supports most of the product quality data cleaning. It not only enriches the theory of product quality data cleaning, but also can be applied to the practice of economic and social development. The model supports automatic, intelligent and high-speed cleaning of product quality data, and provides an important methodology for the research of National Quality Infrastructure (NQI) common technology.**

***Index Terms*—BP artificial neural network; L-M algorithm; national quality infrastructure; product quality; data cleaning**

I. INTRODUCTION

Quality power has always been China unremitting pursuit. Since the 19th National Congress of the Communist Party of China, China has determined the quality power as an important national strategy, and clearly put forward to improve quality and efficiency as a new engine of economic and social development, especially to strengthen the national quality and technological infrastructure. Accelerating the construction of National Quality Infrastructure (NQI), carrying out research on common technology of NQI, improving the integrated service level of NQI, and realizing one-stop service of measurement, standard, certification, accreditation and inspection and detection have become the urgent need of economic development and the endogenous demand of supply improvement. As an important part of the research on NQI common technology, the research on product quality data cleaning is not perfect. Therefore, systematic research on the theory, model, algorithm, tool and experimental verification method of product quality data cleaning can improve the theory and practice of product quality data cleaning, so as to promote the research on NQI common technology. Based on the theory of BP artificial neural network in the subject of artificial intelligence, this paper constructs the model and algorithm of product quality data cleaning with the improved BP artificial neural network of L-M algorithm, and carries out the experimental verification of product quality data cleaning with the designed model based on the refrigerator product quality data, and achieves the expected research results.

II. REVIEW OF RELATED ISSUES

A. Literature review of relevant issues

Generally speaking, the research results of nqi common technology, especially the model, algorithm, tool and method of product quality data cleaning are few. The literature review related to this study is as follows.

He Jun, Zhang Yunfei and Zhang Dehai[1]discussed the automatic combination method of data cleaning rule chain based on Petri net, and carried out empirical research. Zhang Quan, Chen Hui[2]Based on the theory and algorithm of minimum hash, constructed the cleaning method and model of duplicate data, and carried out simulation experiments. Chang Zheng, Lu Yong[3]built a massive data cleaning system based on regular expression, and carried out experimental verification. The experimental results verify that the system has good applicability and accuracy for common data processing problems in limited application scenarios. Wang Zhen and Lin Xin[4]build a data cleaning algorithm for probabilistic resource description framework, and verify the algorithm through experiments, which shows that the algorithm has a good effect. Zhu Huijuan et al[5]proposed a data cleaning method drdcm based on dynamic configurable rules, which supports complex logical operations among various types of rules and multiple dirty data repair behaviors. Zhang Peigen and Huang Shucheng[6]proposed a new algorithm to clean repetitive data based on Chinese word segmentation and synonym checking. Based on the optimization technology of task merging, Yang Donghua et al [7]constructed the optimization model and algorithm of big data cleaning process, and carried out experimental verification research. Lin Jun et al[8]proposed a method of data cleaning for transformer online monitoring based on association rule analysis and neural network, and carried

out experimental verification. David et al[9]constructed a set of data cleaning algorithm for recommendation classification and regression tasks, and carried out experimental verification. Huygues beaufond et al [10]constructed an automatic data cleaning algorithm for short-term load forecasting of distribution network, and carried out experimental verification.

B. Limitations of existing research

Generally speaking, there are few theories about nqi common technology research at home and abroad, especially the model, algorithm, tool and method of product quality data cleaning, which are in the initial and exploration stage as a whole. This paper discusses how to enrich and develop the research of product quality data cleaning.

III. PRODUCT QUALITY DATA CLEANING MODEL BASED ON ARTIFICIAL NEURAL NETWORK

A. BP artificial neural network improved by L-M algorithm

BP (back propagation) artificial neural network is a multilayer feedforward network trained by error back propagation algorithm, which is one of the most widely used neural network models at present. BP artificial neural network can learn and store a large number of input-output mapping relationships without revealing the mathematical equations describing the mapping relationship in advance. Its learning rule is to use the steepest descent method to adjust the weights and thresholds of the network through back propagation, so as to minimize the sum of squares of the network errors. The topological structure of BP neural network model includes input layer, hidden layer and output layer. The specific working principle of BP artificial neural network is to obtain the output value through the non-linear transformation from beginning to end. The condition of each neuron will affect the corresponding neurons in the next layer, so that the fastest speed of error can be reduced. Through continuous repeated learning and training, the error can reach the appropriate range, and the training can be stopped. If the expected value error value is larger than the obtained value, the error will be caused Transfer to the reverse propagation process, the direction is output layer - hidden layer - input layer three layers, through these two processes alternately implemented, at the same time, modify the threshold value and weight value of each layer of neurons, shrink the error until the output value forces the near-term expected value, the network training ends, thus completing the process of information acquisition and memory. In practical application, the traditional BP artificial neural network algorithm is not competent, so there are many improved algorithms. For example: using momentum method to improve BP algorithm, adaptive learning rate algorithm, momentum adaptive learning rate adjustment algorithm, L-M (Levenberg Marquardt) algorithm, etc. In this paper, L-M algorithm is used, which is much faster than the previous algorithms. However, for complex problems, this method needs a lot of storage space. However, with the rapid development of the new generation of table information technology, such as big data, cloud computing, artificial intelligence and so on, the rapid development of computer high-speed computing, mass storage technology provides a guarantee for the practicality of the improved BP artificial neural network method of L-M algorithm. The weight adjustment rate of L-M optimization method is:

$$\Delta w = (J^T J + \mu I)^{-1} \cdot J^T e \quad (1)$$

Among them, e represents the error vector; J represents the Jacobian matrix of the network error derivative to the weight; μ represents the scalar, when μ is large, the upper formula is close to the gradient method, when μ is small, the upper formula becomes the Gauss Newton method, in this method, μ is also self-adaptive.

To determine the parameters of BP neural network, first of all, preprocess the neural network. A three-layer BP neural network can be used to simulate the approximation. For continuous functions with all nonlinearity and in a specific interval, a three-layer BP neural network can be used to get a good approximation performance, and the training time becomes less. Therefore, this experiment requires BP neural network model to be divided into three levels: 600 times of training, global minimum error of 0.005, learning rate of 0.01, momentum factor of 0.1, dynamic parameter of 0.8, minimum training rate of 0.9. The network nodes include three input layers, seven hidden layers and three output layers, forming 3-7-3 In the BP neural network model, the hidden layer nodes should be compared according to the actual situation to determine the optimal network structure.

B. Product quality data cleaning model based on artificial neural network

This paper constructs a general product quality data cleaning model. The model and its indicators are aimed at the general products. The products refer to the products of all manufacturing industries, which is of universal significance. The product quality attribute index of this paper can be set according to the needs of users, that is to say, it can meet international standards, national standards, provincial and ministerial standards, industrial standards, enterprise standards, etc. In this paper, the refrigerator is taken as an example to test and verify, and the refrigerator is only one of them. Other products, such as clothing, mechanical products, furniture, cars, computers, color TV sets, mobile phones, etc., are also applicable. That is to say, the models and indicators are universal, and there are many examples. The refrigerator in this paper is just one of them. Because this paper is aimed at general products, the product quality attribute index designed in this paper is mainly the index that the public can directly feel, including product service life, product function perfection and product appearance novelty. The longer the service life of the product, the better the durability and stability of the product; the more perfect the product function, the stronger the ability of the product to meet the needs of

the user; the more novel the product appearance, the better the appearance quality of the product, and meet the aesthetic needs of the user. These three indicators (product life, product function improvement and product appearance novelty) are the most important indicators to describe product quality attributes, so this paper mainly uses the data of these three indicators for experimental verification. Taking the refrigerator as an example, it has a long service life, perfect product functions (intelligence, energy conservation and environmental protection, input and output functions, etc.), and good appearance. We can basically say that the refrigerator is of good quality. By combining the data of product quality (product service life, product function perfection and product appearance novelty) with the principle and algorithm of BP artificial neural network, and using the mapping ability of BP neural network, the experimental data can be analyzed, cleaned and predicted. Figure 1 is a schematic diagram of product quality data cleaning model based on artificial neural network. In the figure, product life is the abbreviation of product service life, product function is the abbreviation of product function perfection, and product appearance is the abbreviation of product appearance novelty.

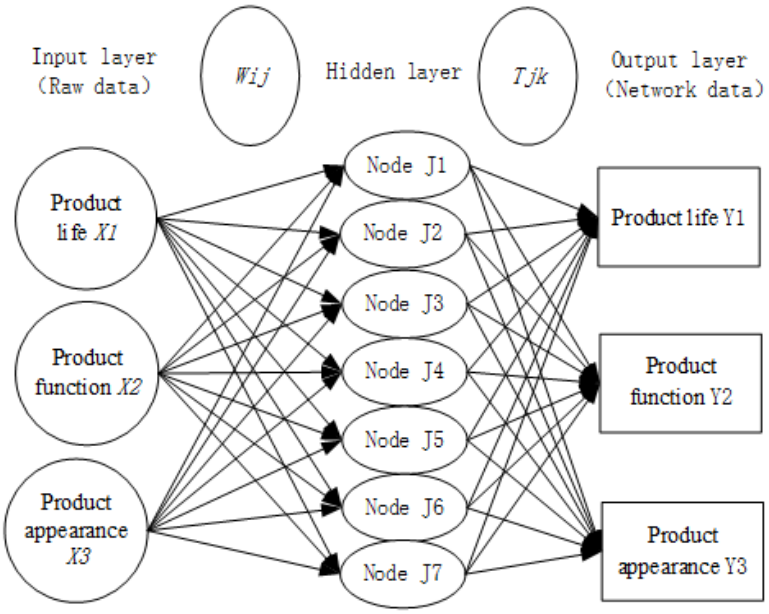

Figure 1 Product quality data cleaning model based on Improved BP artificial neural network

In Figure 1, X1, X2 and X3 represent the original data of the input layer. After being processed by the artificial neural network, the data of the output layer is obtained. Represented by Y1, Y2 and Y3, the output layer is the last layer of the artificial neural network, which has the maximum layer number of the network and is responsible for outputting the calculation results of the artificial neural network. According to the set error value, the data in accordance with the error range can be obtained, and the data not in accordance with the error range can be replaced by the network data, so the cleaned data can be obtained. Wij represents the network weight between input layer and hidden layer, Tjk represents the network weight between hidden layer and output layer.

In addition to the input layer and output layer, other layers are hidden layers. The output formula of hidden layer node is shown in formula (2), where f represents the transfer function and θ_i represents the threshold value of hidden neuron I.

$$J_i = f\left(\sum_{j=1}^{3} W_{ij}X_j - \theta_i\right), i = 1,2,3,4,5,6,7 \quad (2)$$

The calculation formula of the output layer is shown in equation (3), where θ_l is the threshold value of the output neuron L.

$$Y_l = f\left(\sum_{k=1}^{7} T_{jk}H_k - \theta_l\right), l = 1,2,3 \quad (3)$$

According to the number of training experience to determine the choice of hidden neurons, the real problem has a direct relationship with the number of input units and hidden layer units. The formula used is shown in equation (4), where m represents the number of input nodes, n represents the number of output nodes, and C represents the constant between 1 and 10.

$$L = (m + n^{1/2}) + C \quad (4)$$

IV. Data Processing in Models and Algorithms

3.1 Quantification of quality attribute indexes

Because it is difficult to express the attributes such as the degree of product function perfection and the degree of product appearance novelty with accurate numbers, we must first use an appropriate vocabulary set to

express them. The vocabulary set used in this study is {high, relatively high,general, low,very low} level five. Next, we need to make a quantitative transformation of the vocabulary set. Generally, the fuzzy set theory in fuzzy mathematics is used to transform. See Table 1 for the conversion control mode.

TABLE 1 VOCABULARY SET OF FUZZY SET NUMERICAL MEASUREMENT

Lexical set	Fuzzy set value $[\mu_{SA}, \nu_{SA}-\eta_{SA}]$
high	[0.900,0.100- η_{SA}]
relatively high	[0.700,0.300- η_{SA}]
general	[0.500,0.500- η_{SA}]
low	[0.300,0.700- η_{SA}]
very low	[0.100,0.900- η_{SA}]

We can design the vocabulary level and uncertainty η_{SA}, of quality attributes, and then get the corresponding fuzzy set numerical expression $[\mu_{SA}, \nu_{SA}-\eta_{SA}]$, according to table 1, which μ_{SA} means membership and ν_{SA} means non membership. Finally, according to equation (5), we can convert the quality and standard attributes of vocabulary expression into values.

$$\rho_{SA} = \mu_{SA} - \nu_{SA} \times \eta_{SA} \quad (5)$$

3.2 Standardization of quality attribute data

The meaning of each quality attribute is often different, and its magnitude and dimension are often different. Therefore, we need to use standardized functions for standardized transformation. See formula (6) and formula (7) for the conversion function.

$$Q_{SA_{ij}} = \begin{cases} \dfrac{q_{SA_{ij}} - q_{SA_j}^{\min}}{q_{SA_j}^{\max} - q_{SA_j}^{\min}} & q_{SA_j}^{\max} - q_{SA_j}^{\min} \neq 0 \\ 1 & q_{SA_j}^{\max} - q_{SA_j}^{\min} = 0 \end{cases} \quad (6)$$

$$Q_{SA_{ij}} = \begin{cases} \dfrac{q_{SA_j}^{\max} - q_{SA_{ij}}}{q_{SA_j}^{\max} - q_{SA_j}^{\min}} & q_{SA_j}^{\max} - q_{SA_j}^{\min} \neq 0 \\ 1 & q_{SA_j}^{\max} - q_{SA_j}^{\min} = 0 \end{cases} \quad (7)$$

Quality attributes can be divided into positive and negative significance indicators. Positive significance indicators refer to the greater the value, the better, such as product function improvement, product appearance novelty, etc., which can be measured by equation (6); negative significance indicators refer to the smaller the value, the better, such as the difficulty of implementing product quality standards, etc., which can be measured by equation (7). Where, $q_{SA_{ij}}$ represents the value of the jth attribute of the ith quality subject, $q_{SA_j}^{\max}$ represents the maximum value compared in the jth attribute of the quality subject, and $q_{SA_j}^{\min}$ represents the minimum value compared in the jth attribute of the quality subject.

V. EXPERIMENTAL VERIFICATION

According to the refrigerator product data, this paper has carried on the product quality data cleaning experiment verification with the designed model. This paper is based on MATLAB R2017b, and establishes a three-layer BP neural network algorithm model. Take product life, product function and product appearance as input points, and the number of hidden layers is 7 Through multiple training and reference, the output points are respectively product life, product function and product appearance. The output product life, product function and product appearance are repeatedly trained, and the data with high error are cleaned and analyzed. 120 groups of experimental data are used for network training using the functions of BP neural network toolbox, The number of network training is 600, and the learning rate is 0.01. After the network training, the network value is tested, and 60 groups of data in the experimental group are randomly selected for error analysis between the real value and the network value. Table 2 is the comparison table of 60 groups of product life original data and network data, table 3 is the comparison table of product function original data and network data, and table 4 is the comparison table of product appearance original data and network data.

TABLE 2 COMPARISON OF PRODUCT LIFE ORIGINAL DATA AND NEURAL NETWORK DATA

Raw data	network data	relative error	Raw data	network data	relative error	Raw data	network data	relative error
10.33	10.24	0.09	12.58	12.57	0.01	11.92	11.92	0.00
12.24	12.16	0.08	10.35	10.29	0.06	14.72	13.91	0.81
11.48	11.48	0.00	14.57	13.80	0.77	10.71	10.63	0.08
9.91	9.83	0.08	11.29	11.26	0.03	12.95	12.80	0.15
8.02	8.75	-0.73	9.49	9.43	0.06	9.81	9.65	0.16
13.35	13.26	0.09	14.93	14.51	0.42	11.14	11.07	0.07
14.81	14.29	0.52	8.37	8.93	-0.56	12.65	12.63	0.02
15.94	15.18	0.76	12.45	12.38	0.07	10.19	10.12	0.07
10.40	10.32	0.08	11.74	11.74	0.00	11.43	11.43	0.00

9.75	9.14	0.61	10.72	10.65	0.07	8.20	8.81	-0.61
11.37	11.35	0.02	12.07	12.03	0.04	13.57	13.50	0.07
12.85	12.74	0.11	9.84	9.70	0.14	11.75	11.75	0.00
13.63	13.58	0.05	10.93	10.82	0.11	12.71	12.64	0.07
11.24	11.20	0.04	11.58	11.58	0.00	11.86	11.86	0.00
10.97	10.91	0.06	12.35	12.26	0.09	11.08	10.97	0.11
9.58	9.05	0.53	10.90	10.81	0.09	14.69	14.08	0.61
14.72	14.16	0.56	9.67	9.58	0.09	11.63	11.63	0.00
12.71	12.65	0.06	11.41	11.39	0.02	13.08	12.97	0.11
10.68	10.59	0.09	9.62	9.51	0.11	9.92	9.21	0.71
13.20	13.14	0.06	10.83	10.72	0.11	11.51	11.51	0.00

TABLE 3 COMPARISON OF PRODUCT FUNCTION ORIGINAL DATA AND NEURAL NETWORK DATA

Raw data	network data	relative error	Raw data	network data	relative error	Raw data	network data	Raw data
0.895	0.872	0.023	0.875	0.861	0.014	0.917	0.917	0.000
0.742	0.718	0.024	0.817	0.822	-0.005	0.792	0.741	0.051
0.974	0.912	0.062	0.794	0.783	0.011	0.862	0.857	0.005
0.810	0.795	0.015	0.875	0.875	0.000	0.871	0.892	-0.021
0.643	0.652	-0.009	0.532	0.679	-0.147	0.897	0.803	0.094
0.747	0.736	0.011	0.923	0.847	0.076	0.852	0.849	0.003
0.914	0.892	0.022	0.361	0.483	-0.122	0.791	0.752	0.039
0.887	0.875	0.012	0.937	0.881	0.056	0.832	0.827	0.005
0.947	0.928	0.019	0.936	0.936	0.000	0.893	0.893	0.000
0.472	0.516	-0.044	0.854	0.853	0.001	0.375	0.492	-0.117
0.893	0.893	0.000	0.871	0.862	0.009	0.932	0.925	0.007
0.671	0.682	-0.011	0.642	0.638	0.004	0.941	0.941	0.000
0.759	0.772	-0.013	0.820	0.814	0.006	0.901	0.901	0.000
0.832	0.832	0.000	0.928	0.928	0.000	0.927	0.927	0.000
0.951	0.951	0.000	0.851	0.846	0.005	0.971	0.971	0.000
0.856	0.837	0.019	0.892	0.879	0.013	0.861	0.859	0.002
0.942	0.935	0.007	0.393	0.482	-0.089	0.958	0.958	0.000
0.865	0.852	0.013	0.835	0.819	0.016	0.847	0.839	0.008
0.953	0.932	0.021	0.542	0.539	0.003	0.591	0.574	0.017
0.853	0.847	0.006	0.695	0.681	0.014	0.915	0.915	0.000

TABLE 4 COMPARISON OF PRODUCT APPEARANCE ORIGINAL DATA AND NEURAL NETWORK DATA

Raw data	network data	relative error	Raw data	network data	relative error	Raw data	network data	Raw data
0.902	0.893	0.009	0.894	0.893	0.001	0.953	0.953	0.000
0.951	0.951	0.000	0.761	0.772	-0.011	0.915	0.904	0.011
0.948	0.948	0.000	0.893	0.885	0.008	0.763	0.752	0.011
0.872	0.857	0.015	0.917	0.903	0.014	0.890	0.875	0.015
0.415	0.482	-0.067	0.635	0.623	0.012	0.352	0.491	-0.139
0.782	0.763	0.019	0.891	0.883	0.008	0.923	0.923	0.000
0.759	0.745	0.014	0.428	0.493	-0.065	0.847	0.831	0.016
0.875	0.862	0.013	0.859	0.840	0.019	0.757	0.742	0.015
0.742	0.736	0.006	0.941	0.941	0.000	0.963	0.963	0.000
0.549	0.532	0.017	0.692	0.685	0.007	0.562	0.562	0.000
0.721	0.719	0.002	0.937	0.937	0.000	0.835	0.835	0.000
0.843	0.825	0.018	0.517	0.582	-0.065	0.932	0.932	0.000
0.874	0.862	0.012	0.652	0.639	0.013	0.859	0.857	0.002
0.882	0.875	0.007	0.908	0.908	0.000	0.928	0.928	0.000
0.773	0.762	0.011	0.865	0.852	0.013	0.846	0.843	0.003
0.671	0.671	0.000	0.915	0.915	0.000	0.892	0.875	0.017
0.852	0.843	0.009	0.728	0.715	0.013	0.916	0.916	0.000
0.764	0.753	0.011	0.914	0.892	0.022	0.832	0.824	0.008
0.697	0.693	0.004	0.637	0.625	0.012	0.335	0.487	-0.152
0.872	0.865	0.007	0.936	0.936	0.000	0.937	0.937	0.000

We can determine the final data cleaning result according to the set error value. The table lists the comparison and error of the original data and network data of product life, product function and product appearance. If the error of the real value and network value data is large (greater than the set error), the real value with large error shall be directly replaced by the network value output, so as to achieve the purpose of data cleaning. For example, for product life data, if the absolute error value of product life original data and network data is set to be ≤ 0.60, the data with error greater than 0.60 can be replaced by the network value, which is the cleaned data. See Table 5 for product life data after cleaning. For product function data, if the absolute error value of product function original data and network data is set to be ≤ 0.060, the data with the absolute error value greater than 0.060 can be replaced by the network value, which is the cleaned data. See Table 6 for the product function data after cleaning. For product appearance data, if the absolute error value of product appearance original data and network data is set to ≤ 0.060, the data with absolute error value greater than 0.060 can be replaced by the network value, which is the cleaned data. See Table 7 for product appearance data after cleaning.

TABLE 5 DATA OF PRODUCT LIFE AFTER CLEANING

10.33	12.24	11.48	9.91	8.75	13.35	14.81	15.18	10.40	9.14
11.37	12.85	13.63	11.24	10.97	9.58	14.72	12.71	10.68	13.20
12.58	10.35	13.80	11.29	9.49	14.93	8.37	12.45	11.74	10.72
12.07	9.84	10.93	11.58	12.35	10.90	9.67	11.41	9.62	10.83
11.92	13.91	10.71	12.95	9.81	11.14	12.65	10.19	11.43	8.81
13.57	11.75	12.71	11.86	11.08	14.08	11.63	13.08	9.21	11.51

TABLE 6 DATA OF PRODUCT FUNCTION AFTER CLEANING

0.895	0.742	0.912	0.810	0.643	0.747	0.914	0.887	0.947	0.472
0.893	0.671	0.759	0.832	0.951	0.856	0.942	0.865	0.953	0.853
0.875	0.817	0.794	0.875	0.679	0.847	0.483	0.937	0.936	0.854
0.871	0.642	0.820	0.928	0.851	0.892	0.482	0.835	0.542	0.695
0.917	0.792	0.862	0.871	0.803	0.852	0.791	0.832	0.893	0.492
0.932	0.941	0.901	0.927	0.971	0.861	0.958	0.847	0.591	0.915

TABLE 7 DATA OF PRODUCT APPEARANCE AFTER CLEANING

0.902	0.951	0.948	0.872	0.482	0.782	0.759	0.875	0.742	0.549
0.721	0.843	0.874	0.882	0.773	0.671	0.852	0.764	0.697	0.872
0.894	0.761	0.893	0.917	0.635	0.891	0.493	0.859	0.941	0.692
0.937	0.582	0.652	0.908	0.865	0.915	0.728	0.914	0.637	0.936
0.953	0.915	0.763	0.890	0.491	0.923	0.847	0.757	0.963	0.562
0.835	0.932	0.859	0.928	0.846	0.892	0.916	0.832	0.487	0.937

Experiments show that the cleaning model of product quality data constructed in this study has high cleaning efficiency. When cleaning the product quality data, the application of improved BP neural network to correct the abnormal data can better improve the training quality and enhance the accuracy of the abnormal data, so as to realize the effective cleaning of the data, and can make the follow-up products The reliability of quality assessment, analysis and prediction is further improved.

VI. CONCLUSION

Aiming at the limitation of the research on product quality data cleaning at home and abroad, based on the theory of BP artificial neural network in the subject of artificial intelligence, this paper constructs the product quality data cleaning model and algorithm with the improved BP artificial neural network of L-M algorithm, and takes the refrigerator product quality data as an example, carries out the experimental verification of

product quality data cleaning with the designed model. The experimental results show that the product quality data cleaning model given in this study is a universal, scientific and reasonable product quality data cleaning model and algorithm, which supports most of the product quality data cleaning, not only enriches the theory of product quality data cleaning, but also can be applied to the practice of economic and social development. The model supports automatic, intelligent and high-speed cleaning of product quality data, and provides an important methodological support for the research of National Quality Infrastructure (NQI) common technology.

ACKNOWLEDGEMENTS

This study is supported by the National Key Research and Development Project:"Internet +"NQI integrated services generic technology (No.2017YFF0209600), in particular,Projects 1:Research on the basic theory and general standards of NQI integrated services (No.2017YFF0209601),and Projects 3:Research on key application technologies of NQI integrated services (No.2017YFF0209603).

REFERENCE

[1] He Jun, Zhang Yunfei, Zhang Dehai. Automatic combination method of data cleaning rule chain based on Petri net. Computer Engineering, 2019, (12): 1-12.

[2] Zhang Quan, Chen Hui. Cleaning method of repeated data based on minimum hash. Communication technology, 2019, 52 (11): 2653-2658.

[3] Chang Zheng, Lu Yong. Massive data cleaning system based on regular expression. Computer application, 2019, 39 (10): 2942-2947.

[4] Wang Zhen, Lin Xin. Data cleaning for probability RDF database query. Journal of East China Normal University (NATURAL SCIENCE EDITION), 2018, (1): 76-90.

[5] Zhu Huijuan, Jiang Tonghai, Zhou Xi, Cheng Li, Zhao fan, Ma Bo. Data cleaning method based on dynamic configurable rules. Computer application, 2017, 37 (4): 1014-1020.

[6] Zhang Peigen, Huang Shucheng. A neighbor sorting algorithm for Chinese data cleaning. Computer application and software, 2018, 35 (8): 285-291.

[7] Yang Donghua, Li Ningning, Wang Hongzhi, Li Jianzhong, Gao Hong. Optimization of parallel big data cleaning process based on task merging. Journal of computer science, 2016, 39 (1): 97-108.

[8] Lin Jun, Yan Yingjie, Sheng Gefu, Jiang Xiuchen, Yang Yi, Chen Yufeng. Data cleaning of transformer online monitoring considering time series correlation. Grid technology, 2017, 41 (11): 3733-3741.

[9] David Camilo Corrales, Agapito Ledezma, Juan Carlos Corrales. A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks. Applied Soft Computing Journal, 2020, 90(1): 106-180.

[10] Huyghues-Beaufond Nathalie, Tindemans Simon, Falugi Paola, Sun Mingyang, Strbac Goran. Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. Applied Energy, 2020, 261(1): 1-17.