POLISH MARITIME RESEARCH Special Issue 2017 S3 (95) 2017 Vol. 24; pp. 221-227
10.1515/pomr-2017-0126

RESEARCH ON BIG DATA ATTRIBUTE SELECTION METHOD IN SUBMARINE OPTICAL FIBER NETWORK FAULT DIAGNOSIS DATABASE

Ganlang Chen
School of Software, South China Normal University, Foshan, 528225, China

ABSTRACT

At present, in the fault diagnosis database of submarine optical fiber network, the attribute selection of large data is completed by detecting the attributes of the data, the accuracy of large data attribute selection cannot be guaranteed. In this paper, a large data attribute selection method based on support vector machines (SVM) for fault diagnosis database of submarine optical fiber network is proposed. Mining large data in the database of optical fiber network fault diagnosis, and calculate its attribute weight, attribute classification is completed according to attribute weight, so as to complete attribute selection of large data. Experimental results prove that ,the proposed method can improve the accuracy of large data attribute selection in fault diagnosis database of submarine optical fiber network, and has high use value.

Keywords: submarine optical fiber network, fault diagnosis database; big data attribute selection

INTRODUCTION

With the development of computer and Internet technology, the computer network is booming. It brings people convenience while also the network virus [1-2] affecting network security. According to the real-time performance of submarine optical fiber network fault diagnosis[3-5], a big data attribute selection method based on rough set of submarine optical fiber network fault diagnosis database is proposed[6].. The current candidate reduction is chosen to be the big data reduction in the submarine optical fiber network fault diagnosis database, so as to complete its attribute selection[7]. This method has become the focus of discussion of relevant experts and scholars, and its research has gradually entered the scope of experts and scholars. With the deepening of the research content, lots of research results have been produced .

In literature [8], a big data attribute selection method in submarine optical fiber network fault diagnosis database based on decision tree local time scale feature extraction is proposed. The drawback of this method is that the selection of attributes is quit slow. Literature [9] proposed a big data attribute selection method for network fault diagnosis database. This method has a small range of applications, it may increase the load for big data attribute selection.

To solve above problems, this paper proposes a method of big data attribute selection based on support vector machine in submarine optical fiber network fault diagnosis database. First, the decision tree method is used to mine and calculate the big data in the submarine optical fiber network fault diagnosis database, and the attribute of the big data is obtained. Then, the big data attribute classification is completed through a subset of assessment, stop criteria and result validity verification generated by big data attribute subset in submarine optical fiber network fault diagnosis database. According to the similarity degree of data attribute

space, the calculation method of attribute similarity and weight is obtained. The loss function is analyzed to improve the feature selection algorithm of big data attribute and calculate the weight of big data attribute. The gradient rise method is used to solve the saddle point, and furthermore to realize the large data attribute selection in the submarine optical fiber network fault diagnosis database. Experiments show that the proposed method can effectively improve the accuracy of big data attribute selection in submarine optical fiber network fault diagnosis database, reduce the calculation process, energy and time consumption, and has good practical value.

RESEARCH ON BIG DATA ATTRIBUTE SELECTION METHOD IN SUBMARINE OPTICAL FIBER NETWORK FAULT DIAGNOSIS DATABASE

A. COLLECTION AND ANALYSIS OF BIG DATA ATTRIBUTE SELECTION METHOD

(a) Collection of big data attribute selection method

It is necessary to mine data in the submarine optical fiber network fault diagnosis database and then calculate its attributes to realize big data attribute analysis[10-12].By using the tree structure to show the result of data mining, the method is simple and intuitive[13-14], and therefore it is suitable for this paper. The specific process is shown in Figure 1.

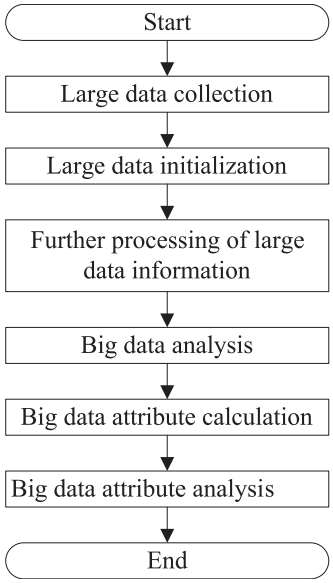

Fig. 1.Dig data attribute analysis process in submarine optical fiber network fault diagnosis database

U is the big data set in submarine optical fiber network fault diagnosis database, F_1 and F_2 are two big data attributes on node N of the decision tree. The information gain of F_1 is greater than that of F_2, so the big data attribute F_1 on node N is selected as a classification attribute.

Assume E_1 and E_2 are the information entropy of F_1 and F_2 respectively, we get

$$gain(F_1) \geq gain(F_2) \Leftrightarrow I(p,n) - E(F_1) \geq I(p,n) - E(F_2) \Leftrightarrow E(F_1) \leq E(F_2) \Leftrightarrow E_1 \leq E_2 \quad (1)$$

wherein, $gain(F_1)$, $gain(F_2)$ are the increased coefficients of F_1 and F_2, p and n are information entropy coefficient of F_1 and F_2.

Let M be a line recording the reduction of big data in submarine optical fiber network fault diagnosis database, which belongs to the range of attribute j on node N. When the record is not reduced, the information entropy of the node attribute can be described as

$$E = \sum_{i=1}^{m} \frac{p_i + n_i}{p+n} \left(-\frac{p_i}{p_i + n_i} \log\left(\frac{p_i}{p_i + n_i} \right) - \frac{n_i}{p_i + n_i} \log\left(\frac{n_i}{p_i + n_i} \right) \right) = \frac{\varepsilon}{p+n} \quad (2)$$

In equation (2), m represents the range count of big data attribute in given submarine optical fiber network fault diagnosis database, n_i and p_i are the information entropy of big data attribute i value segment in the database, and ε represents the value segment of a big data attribute in a given database.

Reduce the centralized record of the dig data in submarine optical fiber network fault diagnosis database[15-16], we get attribute information entropy of the big data attribute node in new database as following:

$$\varepsilon = -\sum_{i=1}^{m} \left(p_i \log\left(\frac{p_i}{p_i + n_i} \right) + n_i \log\left(\frac{n_i}{p_i + n_i} \right) \right), \text{且} \varepsilon \geq 0 \quad (3)$$

$$A = p_j \log\left(\frac{p_j}{p_j + n_j} \right) + n_j \log\left(\frac{n_j}{p_j + n_j} \right) - (p_j - 1) \log\left(\frac{p_j - 1}{p_j + n_j - 1} \right) - n_j \log\left(\frac{n_j}{p_j + n_j - 1} \right) \quad (4)$$

Let $x = p_j$, $y = n_j$, we get equation (5)

$$\Delta E = E' - E = \frac{\varepsilon + A}{p+n-1} - \frac{\varepsilon}{p+n} = \frac{A}{p+n-1} + \left(\frac{\varepsilon}{p+n-1} - \frac{\varepsilon}{p+n} \right) = \frac{A}{p+n-1} + \frac{\varepsilon}{(p+n-1)(p+n)} \quad (5)$$

Wherein, $A(x,y)$ represents a function of x, y, the big data attribute variables, in the database.

(b) Analysis of big data attribute selection method in submarine optical fiber network fault diagnosis database

The big data attribute selection error cloud formula is expressed as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - o_i \right|^2 \quad (6)$$

In the above function, E represents the sum of squared errors for all big data attributes, p represents an object of the

dataset, o_i is mean of class C_i, C_i is the submarine optical fiber network fault diagnosis database, and indicates the number of data object in class C_i. Use formula (7) to calculate the distance from each p in the data set to k cluster center:

$$dist(x,y)=\sqrt{\sum_{i=1}^{d}(x_i-y_i)^2} \tag{7}$$

And then the property extraction is completed through the attribute similarity. $H_i, H_j \in R^D$ are two object spaces, where R^D represents a submarine optical fiber network fault diagnosis database, $d(H_i,H_j)$ represents the distance between two object spaces, and $d(H_{ik},H_{jk})$ represents the spatial distance of the k-th dimension of the two object spaces.

$$d(H_i,H_j)=\max_{k=1,2,\cdots,D}\{d(H_{ik},H_{jk})\} \tag{8}$$

B. SELECTION BIG DATA ATTRIBUTE IN SUBMARINE OPTICAL FIBER NETWORK FAULT DIAGNOSIS DATABASE BASED ON SUPPORT VECTOR MACHINE

According to the above discussion, the loss function of the big data attribute in the submarine optical fiber network fault diagnosis database is:

$$V\left(y_i f(x_i)\right)=L(y_i)\cdot(f(x_i)-y_i)_+ \tag{9}$$

In the above equation:

$(f(x_i)-y_i)_+=((f_1(x_i)-y_{i1})_+,\cdots,(f_m(x_i)-y_{im})_+)$. Class mark y_i is encoded to $y_i=(y_{i1},\cdots,y_{im})$, an m-dimensional vector. Assume that the corresponding large data attribute diagnostic type is j, the j-th component of y_i is 1, and the remaining components are denoted by $-1/(m-1)$. $L(y_i)$ is also an m-dimensional vector, with 0 as its j-th component and 1 as its remaining components.

In order to ensure that each attribute belongs to only a certain category, $f_c(x)$ need to meet the conditions:

$$\sum_{c=1}^{m} f_c(x)=0 \tag{10}$$

Since the above condition is satisfied for any of the data attributes x, it can be converted into

$$\sum_{c=1}^{m}\beta_c=0,\sum_{c=1}^{m}\beta_{0c}=0 \tag{11}$$

we get that SVM-based supervised big data attribute feature selection algorithm is equivalent to optimization problem:

$$\min_{\beta_c,\beta_{0c},c=1,\cdots,m} C\sum_{c=1}^{m}\sum_{i=1,y_i\neq c}^{n_c}\left(\beta_{0c}+x_i\beta_c+\frac{1}{m-1}\right)+\lambda_1\sum_{c=1}^{m}\left\|\beta_c\right\|_1+\frac{\lambda_2}{2}\sum_{c=1}^{m}\left\|\beta_c\right\|_2^2$$
$$s.t.\begin{cases}\sum_{c=1}^{m}\beta_c=0\\ \sum_{c=1}^{m}\beta_{0c}=0\end{cases} \tag{12}$$

Wherein, C is the penalty parameter of the big data attribute in the submarine optical fiber network fault diagnosis database, λ_1 and λ_2 are adjustment parameters [17, 18]. n_c indicates the number of data that does not belong to the c-th big data attribute. By solving the above equation, we get the weight β_{ci} of each attribute in the c-th big data, $i=1,2,\cdots,p$ which is also an important measure of the i-th characteristic of the data attribute, so as to complete the determination of the big data attribute in the submarine optical fiber network fault diagnosis database.

The normalization of big data attribute eigne values is:

$$\sum_{c=1}^{m}\left|\hat{\beta}_{cj}-\beta_{ck}\right|^2\le\frac{2M}{\lambda_2}\sum_{c=1}^{m}\left|\beta_{cj}-\beta_{ck}\right|\sqrt{2n_c(1-\rho_c)} \tag{13}$$

ρ_c represents the correlation coefficient between feature j and feature k that does not belong to the big data attribute in the c-th submarine optical fiber network fault diagnosis database.

In order to solve the problem of the saddle point, the gradient rise method is used to solve the dual problem:

$$\max_{u_c,v_c,p,q}\min_{\beta_c,\beta_{0c},a_c,t_c} L(\beta_c,\beta_{0c},a_c,t_c,u_c,v_c,p,q) \tag{14}$$

In the above discussion, the big data attribute selection algorithm is improved by calculating the loss function, the weight of the big data attribute is calculated, and the saddle point is solved by the gradient rising method, so as to realize the selection process of big data attribute in submarine optical fiber network fault diagnosis database based on support vector machine. The process is shown in Figure 2.

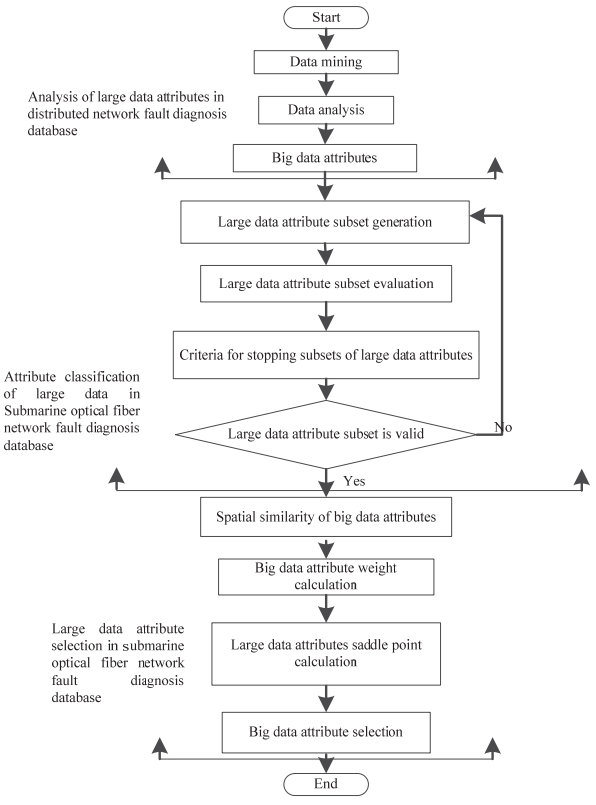

Fig. 2.Big data attribute selection process in submarine optical fiber network fault diagnosis database

EXPERIMENTAL RESULTS AND ANALYSIS

In order to prove the validity of the big data attribute selection method in the submarine optical fiber network fault diagnosis database based on support vector machine, we use MATLAB 2008a as the platform and Intel P4 2G processor to perform the simulation experiment[19-23].

In this paper, we use three data sets in the network database to experiment, analyze the data attributes of three different experimental data sets, and compare the time consuming of three data sets.

In the first data set, the method proposed in this paper is compared with the data attribute selection method proposed in [8] and [9], and the comparison result is shown below.

First, the time consumed (min), calculated through formula (15),in the three methods for big data attribute selection is compared[24].

$$T = \frac{E\log L(\gamma)}{\sum_{i=1}^{m} f(x)} \cdot \sigma \tag{15}$$

σ represents the response time parameter when the big data attribute is selected, and the average response time of the big data attribute selection is obtained according to the above three methods. The comparison results are shown in Table 1.

Tab. 1.Time-consuming comparison of three methods

Numbers of experiments / times	The proposed method/min	The method proposed in literature [8]/ min	The method proposed in literature [9]/ min
50	18	25	23
100	35	47	44
150	50	71	66
200	67	89	85
250	83	112	109
350	115	159	148
500	129	218	213

According to the formula (16), the average time-consuming comparison of the three methods in the second data set is calculated. In order to ensure the accuracy of the experiment, 500 experiments were carried out, with 50 experimental data as a set of data, so as to complete the average time calculation, the time unit is seconds (s), the formula is:

$$T_0 = \frac{T - T'}{50} \tag{16}$$

In the above formula, T' represents the time spent in other work in the experiment. Through the calculation, we get the average time comparison of the three methods for big data attribute selection. The comparison results are shown in Table 2.

Tab. 2.Average time-consuming comparison of three methods for large data attribute selection

Numbers of experiments / times	The proposed method/min	The method proposed in literature [8]/ min	The method proposed in literature [9]/ min
50	14.8	22.8	19.1
100	14.4	22.1	17.8
150	14.1	21.9	21.6
200	15.1	22.0	17.9
250	15.2	22.5	17.6
300	15.1	22.9	20.9
350	15.5	22.6	19.5
400	15.2	21.8	19.7
450	15.3	22.3	20.2
500	15.0	22.1	20.8

Then, we compared the average time-consuming of three method in the third data set, and got the results shown in Figure 3.

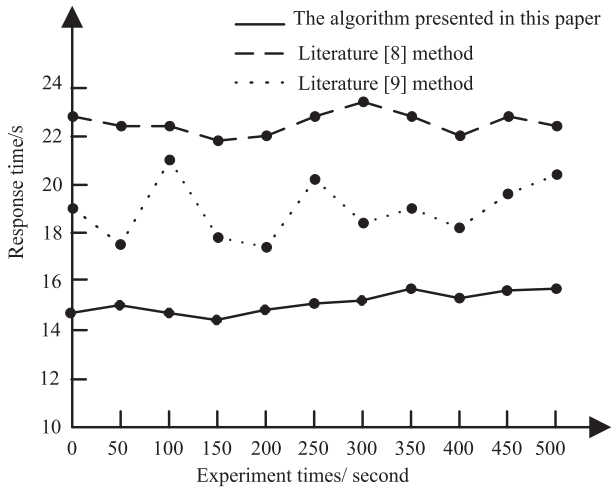

Fig. 3.The average time-consuming comparison of the three methods

In the figure above, the starting point of the line indicating the theoretically time consuming of the three methods. It can be seen that difference between actual and theoretical time-consuming of the proposed method is less than that of the literature [8] and the literature [9]. The average time-consuming polyline of the proposed method is close to a straight line and the fluctuation is small, which indicates that the proposed method is stable in the big data attribute selection.

Then compare the energy consumption of three methods, we assume N as the energy consumption unit,

$$N = \sum\sum_{i=1}^{m} \beta_c \cdot \varepsilon \tag{17}$$

According to the above formula, the energy consumption of the three methods for big data attribute selection is compared. The results are shown in Table 3.

Tab. 3. Energy consumption comparison of the three methods for big data attribute selection

Time/h	The proposed method /N	The method proposed in literature [8]/N	The method proposed in literature [9]/N
5	31	47	61
10	58	92	117
15	86	126	179
20	113	174	236
25	139	226	292
30	167	268	348
40	194	313	463
50	227	359	562

In order to better display the results, we converted Table 3 into the following line chart. The energy consumption comparison results of the three methods are shown in Figure 4.

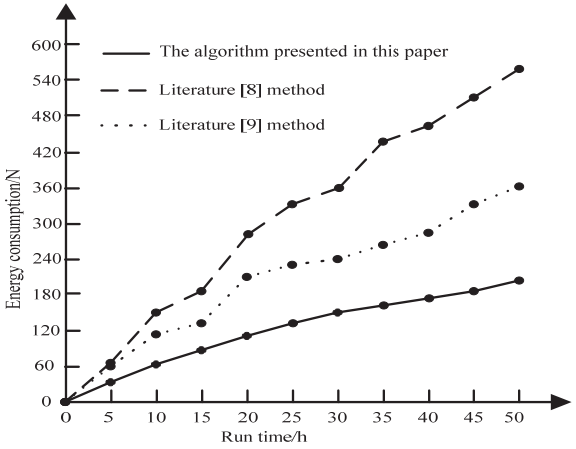

Fig. 4 The energy consumption comparison results of the three methods

It can be seen that the proposed method can effectively reduce the energy consumption in big data attribute selection process. The energy consumption fluctuation of the proposed method in big data attribute selection is smaller than that of literature [8] and the literature [9], which indicates that the proposed method is stable in the big data attribute selection.

At last, we compared the accuracy of three methods in big data attribute selection process. The experiment used three methods to select the data attributes of any seven databases in the network.Since the number of attributes in the database is large, accuracy indicates that the number of big data attributes can be selected correctly, and error indicates that the number of big data attributes can not be selected accurately. The results are shown in Table 4.

Tab. 4.The accuracy of three method for big data attribute selection

Number of attributes	The proposed method		The method proposed in literature [8]		The method proposed in literature [9]	
	Accuracy	Error	Accuracy	Error	Accuracy	Error
132	123	9	109	23	113	19
210	203	7	187	23	196	14
218	207	11	201	17	289	29
345	339	6	319	26	327	18
426	413	13	401	25	407	19
457	443	14	426	31	431	26
543	522	21	507	36	516	27

The accuracy ratio is the ratio of the exact quantity to the total quantity. The error rate is the ratio of the number of errors to the total quantity. The formula is as follows (18).

$$\begin{cases} \eta = \dfrac{\text{Exact number}}{\text{Total quantity}} \times 100\% \\ \lambda = \dfrac{\text{Error number}}{\text{Total quantity}} \times 100\% \end{cases} \tag{18}$$

In the formula, η and λ indicate the accuracy and error rate. Using the above table information, the accuracy of the three methods for big data attribute selection are compared, the results are shown in Figure 5, Figure 6.

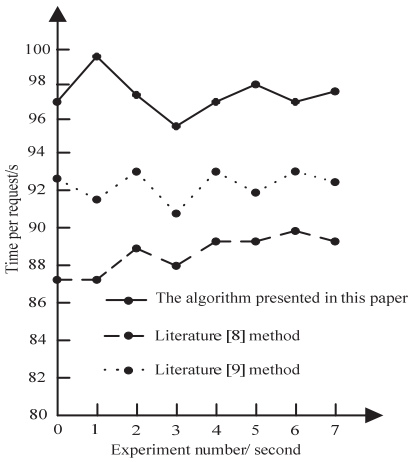

Fig. 5 Comparison of accuracy of three methods

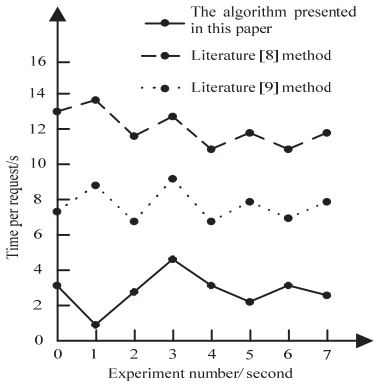

Fig. 6.Comparison of error rate of three methods

Through the above figure we can see that the method proposed in this article has the highest accuracy and the lowest error rate.

In summary, the method proposed in this paper can effectively reduce the energy consumption and cost of big data attribute selection in the submarine optical fiber network fault diagnosis database, improve the efficiency of big data attribute selection in the submarine optical fiber network fault diagnosis database, ensure the real-time of network fault diagnosis, and has great practical value.

CONCLUSION

The choice of big data attribute in submarine optical fiber network fault diagnosis database is the basis of data mining and processing of submarine optical fiber network fault diagnosis database. Improve the time consumption of big data attribute selection is conducive to ensure the real-time fault diagnosis, thus improving the capability of submarine optical fiber network fault diagnosis. In this paper, the big data attribute method based on SVM in submarine optical fiber network fault diagnosis database can effectively reduce the time taken for fault diagnosis and improve the efficiency of fault diagnosis, and has good practical value.

REFERENCES

1. Karabadji N E I, Seridi H, Khelf I, et al. Improved decision tree construction based on attribute selection and data sampling for fault diagnosis in rotating machines. Engineering Applications of Artificial Intelligence, 2014, 35(35):71-83.

2. Zhang Q H, Qin A, Shu L, et al. Vibration sensor based intelligent fault diagnosis system for large machine unit in petrochemical industry. International Journal of Distributed Sensor Networks, 2015, 2015(3):1376-1381.

3. Jin S, Cui W, Jin Z, et al. AF-DHNN: Fuzzy Clustering and Inference-Based Node Fault Diagnosis Method for Fire Detection.. Sensors, 2015, 15(7):17366-17396.

4. Panda M, Khilar P M. Distributed self fault diagnosis algorithm for large scale wireless sensor networks using modified three sigma edit test. Ad Hoc Networks, 2015, 25(PA):170-184.

5. Zhang Q H, Hu Q, Sun G, et al. Concurrent Fault Diagnosis for Rotating Machinery Based on Vibration Sensors. International Journal of Distributed Sensor Networks, 2015, 2013(1):59-72.

6. Reppa V, Polycarpou M M, Panayiotou C G. Distributed Sensor Fault Diagnosis for a Network of Interconnected Cyberphysical Systems. IEEE Transactions on Control of Network Systems, 2015, 2(1):11-23.

7. Islam R, Khan S A, Kim J M. Discriminant Feature Distribution Analysis-Based Hybrid Feature Selection for Online Bearing Fault Diagnosis in Induction Motors. Journal of Sensors, 2016, 2016(2):1-16.

8. LAn-qiang, Liu Z, Yin C Q, et al. A Fault Diagnosis Method Forwavelet Packet and Neural Network-Based Submarine Cables. Study on Optical Communications, 2016, 42(2):16-22..

9. Gao Y, Yang C, Tian S, et al. Entropy Based Test Point Evaluation and Selection Method for Analog Circuit Fault Diagnosis. Mathematical Problems in Engineering, 2014, 2014(6):1-16.

10. Lei Y, Jia F, Lin J, et al. An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. IEEE Transactions on Industrial Electronics, 2016, 63(5):3137-3147.

11. Wang S, Sun X, Li C. Wind Turbine Gearbox Fault Diagnosis Method Based on Riemannian Manifold. Mathematical Problems in Engineering, 2015, 2014(4):1-10.

12. Jin X, Chow T W S, Sun Y, et al. Kuiper test and autoregressive model-based approach for wireless sensor network fault diagnosis. Wireless Networks, 2015, 21(3):829-839.

13. Kelkar S, Kamal R. Adaptive Fault Diagnosis Algorithm for Controller Area Network. IEEE Transactions on Industrial Electronics, 2014, 61(10):5527-5537.

14. Unal M, Onat M, Demetgul M, et al. Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network. Measurement, 2014, 58:187–196.

15. Lu Chong, Xu Hui, Yang Yongchun. Research and application of . decision tree classification algorithm based on electronic design engineering, 2016, 24 (18): 1-3.

16. Gao, W. and W. Wang, The fifth geometric-arithmetic index of bridge graph and carbon nanocones. Journal of Difference Equations and Applications, 2017. 23(1-2SI): p. 100-109.

17. Gao, W., et al., Distance learning techniques for ontology similarity measuring and ontology mapping. Cluster Computing-The Journal of Networks Software Tools and Applications, 2017. 20(2SI): p. 959-968.

18. Xue C, Jing L I, Wang H, et al. Effects of Target and Distractor Saturations on the Cognitive Performance of an Integrated Display Interface. Chinese Journal of Mechanical Engineering, 2015, 28(1):208-216.

19. Halim H,Abdullah R, Nor M J M,Aziz H A, Rahman N A. Comparison Between Measured Traffic Noise in Klang Valley, Malaysia And Existing Prediction Models. Engineering Heritage Journal, 2017, 1(2):10–14.

20. Ebrahimi N, Gharibreza M,Hosseini M, Ashraf M A. Experimental study on the impact of vegetation coverage on flow roughness coefficient and trapping of sediment. Geology, Ecology, and Landscapes, 2017, 1(3): 167-172.

21. Adugna O,Alemu A. Evaluation of brush wood with stone check dam on gully rehabilitation. Journal CleanWAS, 2017, 1(2): 10-13.

22. Guoming L,Yanmin C,Guowe Y, Xiaoping Y.Research on Data Management Model of National Defense Mobilization Potential Based on Geo-Spatial Framework. Malaysian Journal Geosciences, 2017, 1(2): 10-12.

23. Simon N, Roslee R, Lai G T. Temporal Landslide Susceptibility Assessment Using Landslide Density Technique. Geological Behavior, 2017, 1(2):10–13.

24. Isemael Y Y. Molecular, Histological and biochemical effects of tea seed cake on hepatic and renal functions of Oreochromis niloticus. Acta Scientifica Malaysia, 2017, 1(1): 13-15.

CONTACT WITH THE AUTHOR

Ganlang Chen

e-mail: chenganlang629@163.com

School of Software
South China Normal University
Foshan 528225
CHINA