Computing and Informatics, Vol. 43, 2024, 797–818, doi: 10.31577/cai_2024_4_797

DATA MINING ALGORITHM FOR WEB LEARNING RESOURCE INFORMATION FLOW LOSS BASED ON WEIGHTED DEPTH FOREST

Shuling Zhou

Institute of Artificial Intelligence
Hefei College of Finance and Economics
Hefei 230601, China
e-mail: zsl020213@163.com

Abstract. When processing the lost data of web learning resource information flow, the noise in the data signal cannot be eliminated, resulting in inaccurate detection of the lost data of web learning resource information flow in the later stage. Therefore, a data mining algorithm is proposed based on weighted depth forest for web learning resource information flow loss. Based on building a brand-driven Web data acquisition model to collect data, this method uses clustering analysis technology to extract the lost data feature information of web learning resource information flow. It carries out wavelet threshold denoising on it. According to the characteristics of lost data, the lost data mining of web learning resource information flow is completed. Experimental results show that the proposed algorithm has a low error rate, high accuracy, high labour intensity, high efficiency and high performance.

Keywords: Weighted depth forest, cluster analysis, wavelet threshold denoising, data mining algorithm, data acquisition

1 INTRODUCTION

With the rapid development of the Internet, more and more enterprises and government departments have websites to provide Web services, such as product introduction and information release. Web has become the basis of e-commerce. With the continuous increase of Web data, there is a loss of Web learning resource data,

which will directly impact Web services [1]. Web learning resource drain Data mining is applying data mining technology to the Web. It is a comprehensive technology involving data mining, web, informatics, computer linguistics, and many other disciplines. Mining Web learning resource loss can improve Web data classification ability [2]. Research on data mining algorithms of Web learning resource loss is significant in Internet information processing.

Sun et al. [3] proposed a prediction method for bandwidth demand and quality of service (QoS) of Web applications based on network simulation. This method provides a modelling framework and formal description suitable for Web services, adopts a simplified parallel load model, extracts model parameters from Web application access logs using automated data mining methods, and uses network simulation tools to establish a system model to simulate complex network transmission processes, which can predict bandwidth requirements and QoS changes under different load intensities. However, the algorithm must be more explicit in dividing the lost data of web learning resource information flow and has a high error rate. Frequent itemset mining, such as in Liang et al. [4], is significant in many data mining applications, such as web log mining and trend analysis. However, if the data is sensitive (for example, web browsing history), direct publishing of frequent itemsets and their support may violate user privacy. Therefore, under differential privacy, this method ensures that the calculated output is insensitive to any single tuple by adding noise, thus protecting the user's privacy. However, this algorithm takes a long time to calculate the lost data of web learning resource information flow and has low labour intensity and efficiency. Yang and Wang [5] proposed data stream processing technology for Web learning resources based on data mining. Firstly, it analyses the research progress of Web learning resource flow processing and the factors that cause the poor performance of Web learning resource flow processing. Then, the breadth-first algorithm is used to search web pages and collect data streams of web learning resources. The Bayesian network algorithm mines association rules of data streams to find the most valuable information on the web pages. The elastic scalability mechanism is used to process the data streams of learning resources. Through looking for common if-then structures in the information and utilising the supporting or reliability criterion to pinpoint the highest significant associations, connection results are created. However, the algorithm extracts web learning resources information stream with considerable noise, low accuracy and low performance.

To solve the problems in the above algorithms, a data mining algorithm for web learning resource information flow loss based on weighted depth forest is proposed. The specific steps are as follows:

1. Build a Xinpai-driven Web data acquisition model to collect Web data signals.
2. Using clustering analysis technology to extract the lost data feature information of web learning resource information flow.
3. Wavelet threshold denoising method is used to process data signals.

4. The lost data of web learning resource information flow is detected by weighted depth forest.
5. Experiments and discussions verify the overall effectiveness of the proposed algorithm by classifying the lost data of web learning resource information flow, the clarity of extracted data signals, and data mining time.
6. Conclusion.

2 DATA PREPROCESSING

2.1 Xinpai Driven Web Data Collection Model

The lost data mining algorithm of web learning resource information flow based on weighted depth forest takes the essential elements of Petri net as the theoretical basis, applies it to system modelling and designs a brand-driven web data collection model [6].

2.1.1 Introduction to Petri Net

Dr. Carl Adam Petri first proposed the concept of Petri nets in 1962. 1980 he held the first international symposium on Petri nets theory and application. Since then, the annual symposium has continuously enriched and improved the theory and related applications of Petri nets. Petri net is a diagrammatic computing paradigm for modelling simultaneous networks. It has been mostly utilised to simulate synthetic organisations, including industrial services and network technologies. A Petri Net is a network paradigm for the sheer necessity of machines demonstrating parallelism in their operations. The network can be separated, the two network kinds comprising locations depicted as spheres and movements represented as bars. The graph's lines are oriented or go from locations to changes or vice versa. Petri nets study the behaviour characteristics of system models, including reachability between labels, reachability of states, the activity of transitions, reversibility of initial states, persistence between transitions, boundedness of positions, synchronisation distance between events, fairness, etc. [7]. According to a persistence paradigm, there will be no alterations among the present and the predicted period. Hence, the potential value of a response variable will be determined under this premise. The four components of a Petri net are locations, transformations, connections, or symbols. Locations are depicted visually through circles, transformations by squares, boundaries by directional lines, and symbols by tiny solid (full) groups. The Petri net model is analysed using the incidence matrix, state equation, reachability tree, invariants and analysis reduction rules [8]. Any modifications we make to the platform will not change an immutable standard template, a law that applies consistently when two alternatives are semantically equivalent; a reducing criterion states this. We define computing as a reworking procedure in which simple data processing principles change a phrase. As a mathematical and graphical modelling tool, Petri Net provides an integrated modelling, analysis and control environment, which facilitates

the design of the system. The physical process of web space environment data collection is shown in Figure 1. However, the input has been attained from data source information to control the web space environment data acquisition. From this, the tasks are generated in real time and recycled properly to handle the collection tasks.

2.1.2 Model Design

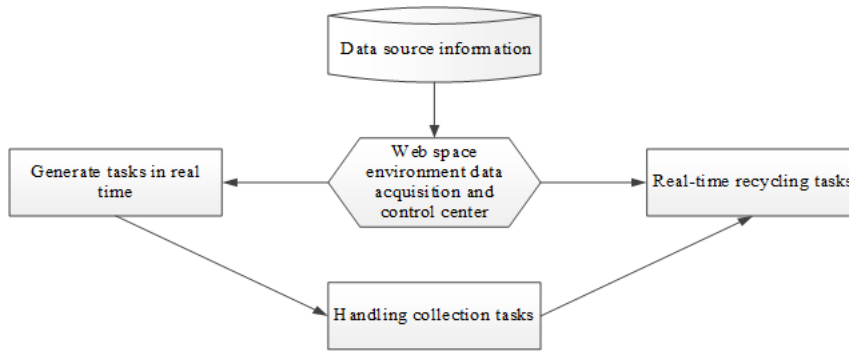

Figure 1. Physical process of web space environment data collection

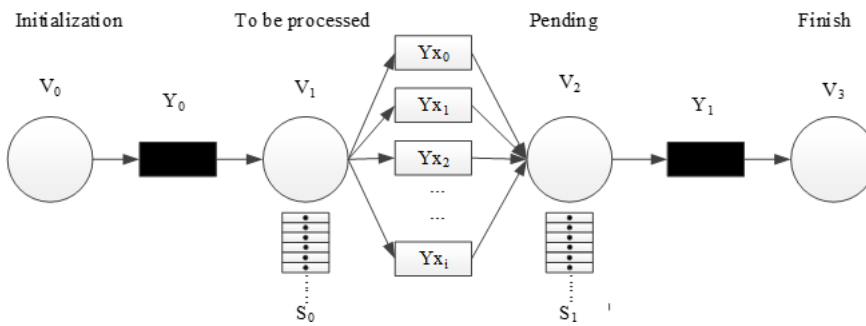

Figure 2. Xinpai driven web data collection model

2.1.3 Model Definition

The formal definition of the driven web data collection model [9] is divided into three parts: static description, state description and execution process, as shown in Figure 1. These three parts jointly define the collaboration relationship and execution order between activities and depict the data collection process of the model entirely.

Definition 1. A data collection process Q is composed of multiple groups, and its expression is as follows:

$$Q = (H, B, P, E, S, V, I, Yx, G), \tag{1}$$

where the set of global variables used for control and communication in the whole process is represented as H; the value range of related variables is B; variable type (i.e. $\forall h \in H$, $P(h) \in B$) $H \to B$ represents P; for $\forall e \in E$, the credit set of a credit entity is represented as E; the data structure of dynamic storage E, namely dynamic queue set, is represented as S; the letter box set is represented as V, which is divided into front and rear letter box sets. The letter box contains H, E or S; the activity set is represented as Y, and $\forall y \in Y$ represents an activity entity, which is specifically composed of attributes such as name, ID, type, function, and state function. Concurrent activities are represented as Yx, $\forall Yx \in Y$ and Yx activities have independent resources, do not interfere with each other, and are developed concurrently; the contact set is represented as G, which defines the contact $G \in (V \times Y) \cup (T \times V)$ between the activity sets of the tray collection.

Definition 2. $\forall v \in V$, the expressions of the front mailbox deck ${}^{\bullet}v$ and the rear mailbox deck $v^{\bullet}$ are as follows:

$$\begin{cases} {}^{\bullet}v = \{y \in T, (v, y) \in G\}, \\ v^{\bullet} = \{y \in T, (y, v) \in G\}. \end{cases} \tag{2}$$

Definition 3. The expression of the function is shown in Formula (3):

$$\Phi_i\left(h_1, h_2, \ldots, h_n, e_i\right) = u_i, \tag{3}$$

where $h_i, u_i \in H, e_i \in E$. The function's primary function is to realize the specific business processing involved in the data collection process. Depending on the techniques used to acquire them, information can be divided into four basic categories: observable, empirical, simulation, and generated. The nature of this study information gathered could impact how you handle that information. $\forall y \in Y$ has one or more nested function functions corresponding to it.

Definition 4. The expression of the state function is shown in Formula (4):

$$\Theta_i\left(e_i\right) = d_i, \tag{4}$$

where $e_i, d_i \in E$. When an activity is completed, use the state function to update the state attribute of the letter and put it into the post letter box according to the rules. State functions are numbers that rely on a product's condition, such as its quantity or kind, heat, or stress. State functions are independent of how the form was created or achieved. Original information gathering and deductive approach are the two categories of sampling techniques. When the function function is completed, start the state function.

State description describes the state of different periods during data collection in the model.

Initialization status: The token box is initialized at V_0, and the global variable set H, mapping P, range set B, and the token set E_5 included in it need to be initialized. The initialization configuration is completed with the user's direct participation.

Pending status: The pending signboard box is V_1, which mainly involves the status of each card in the dynamic queue S_0 in the signboard box belongs to the pending stage. Whenever the back is in the last array place, that is, if the backlog is entirely full (MaxSize − 1). Therefore, no additional components can be added if the backlog contains some free spaces.

To be completed status: The card box to be completed is V_2, which means that the task involved in each card in the dynamic queue S_1 has been completed, and the card is waiting to be recovered.

End status: The end signal box is V_3, which means that the data collection of the whole stage has been completed.

Execution process describes the multi-group between the letter box and the activity set in the model according to Figure 2.

- V_0: The user initializes the global variable H, value domain B, mapping P, and E to be initialized, especially the flexible configuration of E.
- Y_0: It is the starting behaviour of the data collection task with a timestamp. The method of obtaining information involves documenting details on behaviors. The DBMS developed the timestamp as a special identification to show when a session began. Timestamp rates are generally given in the sequence that events are sent through the network. It uses the structure of the phase loop diagram to expand the loop traversal. The timestamp of the token is the only sign of starting. The purpose of the function function at this stage is to complete the creation of the token.
- V_1: The dynamic queue S_0 stores the set of cards expanded by the Y_0 status function and waits for the cards to be transferred to Yx.
- Y_x: Its boundary must be controlled to avoid the explosion of concurrent active nodes. It is defined as $Z = \{Yx_0, Yx_1, \ldots, Yx_i; i < Z\}$. Among them, the user defines the Z value. The function of Yx activity focuses on implementing the signboard task, which is also the core process of completing data collection.
- V_2: The copy of the trust card in S_0 is the S_1's trust card. The difference between the two is that the S_1's trust card is a set of cards processed by the Y_x's state function, and the trust card has expired and is waiting to be recycled.
- Y_2 refers to the end behavior of the data collection task with a time stamp. It uses the structure of a phase loop graph to expand loop traversal. It is the only sign for S_1 to start when the queue is empty.

When the S_1 queue is empty, no operation is performed, and when it is not empty, no task recycling operation is performed.

- V_3 represents the end state of the entire web learning resource information flow data collection model.

2.2 Cluster Analysis

Agglomerative clustering begins by collecting individual nodes into groups. One type of Hierarchy technique is the divided approach, which begins grouping with the dataset provided before breaking it into divisions. With the rapid growth of network information and the emergence of Web information extraction technology, clustering technology [10] has become an important content in the field of data mining and is widely used in real life. A study's statistical significance can be lowered by incomplete information, which can also lead to skewed estimations and false findings. The lost data mining algorithm of web learning resource information flow based on weighted depth forest uses clustering analysis technology to extract the lost data feature information of web learning resource information flow. By finding people who share characteristics, cluster analysis helps businesses comprehend their clients and can help them improve when they interact with them.

2.2.1 Feature Selection

In the clustering process, the most important step is feature selection [11, 12, 13]. The feature weight can reflect the importance of each feature. To handle huge datasets, we require grouping methods that are easily customizable. Capacity to control various qualities procedures ought to apply to every type of information, including quantitative, category, and intermission (mathematical) information. Eliminating consonants from English phrases is done using the Porter stemming algorithm. Automated ending removal is a helpful process, particularly in knowledge discovery. A manuscript is often parameterized of letters or concepts in an IR context. In the process of preprocessing the lost data of web learning resource information flow, the tfc algorithm and porter stemming algorithm are used to select features and extract stemming. Stemming, which items to endings, prefixed, or the foundations of commonly defined as phrases, is the practice of removing a term to its root. It is significant in language comprehension (NLU) or natural language synthesis (NLP). Every procedure done on original data to prepare it for a further computational operation is referred to as data preprocessing, which is a part of the information preparation. Preprocessing the data is essential to ensure its great service. Data preprocessing is broken down into four steps to simplify things: information extraction, data aggregation, discretization, and data conversion. The tfc algorithm is used to calculate the feature weight c_{ij} of feature j in record i. The

expression is shown in formula (5):

$$c_{ij} = \frac{g_{ij} \times \lg\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{k=1}^{M}\left[g_{ik} \times \lg\left(\frac{N}{n_k}\right)\right]}}, \tag{5}$$

where the frequency of feature j in record i is g_{ij}, and the total number of records in the data set lost by web learning resource information flow is N. The number of all features is M, and the number of records of feature j and feature k are n_j and n_k, respectively. To measure the similarity between clusters [14], the matrix for storing clustering feature vectors is shown in Equation (6):

$$\text{Content Units} \Rightarrow \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \Rightarrow \begin{bmatrix} c_{11} & \cdots & c_{1M} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{MM} \end{bmatrix}, \tag{6}$$

where each line represents a vector corresponding to a record. Suppose that N records are composed of M features, and the feature weight value corresponding to the j^{th} feature of c_i is expressed as c_{ij}. Therefore, the eigenvector corresponding to c_i is also c_i.

2.2.2 Similarity Measurement

Suppose that two clusters are V_i and V_j, and their corresponding eigenvectors are $G_i = (c_{i1}, c_{i2}, \ldots, c_{im})$ and $G_j = (c_{j1}, c_{j2}, \ldots c_{jn})$ respectively, and the Euclidean distance [15] between V_i and V_j is expressed as Formula (7):

$$\text{Dist}\,(G_i, G_j) = \sqrt{\sum_{k=1}^{m} |c_{ik} - c_{jk}|^2}, \quad i \neq j. \tag{7}$$

According to Formula (7), their distance reaches the maximum when V_i and V_j have no identical characteristics. When the characteristics and weights between V_i and V_j are the same, the distance between them is the minimum value, namely $c_{ik} = c_{jk}$ $(i \neq j)$. According to Formula (5) and Formula (7), the maximum and minimum

values between clusters are Formula (8):

$$\begin{aligned}\int \max \operatorname{Dist}(G_i, G_j) &= \sqrt{\sum_{k=1}^{m} |c_{ik} - c_{jk}|^2} \\ &= \sqrt{\sum_{k=1}^{m} c_{ik}^2 + \sum_{k=1}^{m} c_{jk}^2} \\ &\left\{ \sqrt{\sum_{k=1}^{m} \left(\frac{g_{ik} \times \lg\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{l=1}^{m} \left[g_{il} \times \lg\left(\frac{N}{n_l}\right) \right]^2}} \right)^2} + \sum_{k=1}^{m} \left(\frac{g_{jk} \times \lg\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{l=1}^{m} \left[g_{jl} \times \lg\left(\frac{N}{n_l}\right) \right]^2}} \right)^2 \right\} \quad (8) \\ &= \sqrt{1+1} = \sqrt{2} \\ &\text{subject to } i \neq j \\ \min \operatorname{Dist}(G_i, G_j) &= \sqrt{\sum_{k=1}^{m} |c_{ik} - c_{jk}|^2} = 0 \\ &\text{subject to } i \neq j\end{aligned}$$

The range of distance between two clusters is $[0, \sqrt{2}]$. Therefore, the similarity calculation between clusters is normalized, and the similarity between V_i and V_j is shown in Equation (9):

$$\operatorname{sim}(G_i, G_j) = 1 - \sqrt{\frac{\sum_{k=1}^{m} |c_{ik} - c_{jk}|^2}{2}}, \quad i \neq j. \tag{9}$$

According to Formula (9), the range of distance between two clusters is $[0, 1]$. According to Formula (8) and Formula (9), when the distance between two clusters reaches the maximum $\sqrt{2}$, the minimum similarity between them is 0. The data mining algorithm for information flow loss of web learning resources based on weighted depth forest uses iterative calculation of the distance between clusters, and combines the clusters with the greatest similarity. The conditions for combining clusters are shown in Formula (10):

$$\operatorname{Cluster}(G_i, G_j) \Leftrightarrow \max\{\operatorname{sim}(G_i, G_j) \mid i \neq j\} \text{ and } \operatorname{sim}(G_i, G_j) \geq \vartheta. \tag{10}$$

The similarity between V_i and V_j clusters is $\operatorname{sim}(G_i, G_j)$, and the threshold value that the similarity needs to meet is ϑ. When the similarity between two clusters is greater than or equal to the threshold value, ϑ combines them into a cluster.

2.2.3 Comparison Variables

In the similarity measurement in the previous section, a threshold is defined to determine whether clusters are merged. However, selecting a threshold has always been a very challenging problem in clustering research. Generally, the threshold cannot be fixed, and it needs to be set according to the needs of different clusters and the dataset's characteristics. Based on the weighted depth forest [16], the data mining algorithm for information flow loss of web learning resources proposes the concept iteration of comparison variable CV to judge whether the two nearest clusters should be merged.

Assuming that the distance between V_i and V_j is the closest, the feature vector set of all records of cluster V_i is represented by $S = \{s_1, \ldots, s_k, \ldots, s_n\}$. The expression of the calculation method to determine whether the clustering V_i and V_j can be combined is shown in Formula (11):

$$CV\left(V_i, V_j\right) = \begin{cases} 1, & \text{if } \prod_{k=1}^{n}\left(1 - \operatorname{sim}\left(s_k, V_j\right)\right)^{\mu} \leq \max\left\{\operatorname{sim}\left(s_k, V_j\right) \mid s_k \in V_j\right\}, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where the similarity between the eigenvector of k recorded in cluster V_i and the eigenvector of V_j is $\operatorname{sim}\left(s_k, V_j\right)$, and $1 - \operatorname{sim}\left(s_k, V_j\right)$ represents the degree of dissimilarity between the eigenvector of k recorded in cluster V_i and the central vector of cluster V_j, that is, the degree of dissimilarity.

Compare the maximum value between $\prod_{k=1}^{n}\left(1 - \operatorname{sim}\left(s_k, V_j\right)\right)^{\mu}$ and $\operatorname{sim}\left(s_k, V_j\right)$. If the latter is large, it means that the similarity between k and V_j of the record closest to V_j in V_i is greater than or equal to the difference between all the records of V_i and V_j. At this time, $CV\left(V_i, V_j\right) = 1$ can be combined. If the former is relatively large, it means that the similarity between k and V_j of the record closest to V_j in V_i is less than that of all records between V_i and V_j. In this case, $CV\left(V_i, V_j\right) = 0$ cannot be combined. μ refers to the factor regulating clustering stringency. The higher the μ value, the lower the clustering stringency. The proof process is as follows.

Proof. If $\mu \to 0$, $\prod_{k=1}^{n}\left(1 - \operatorname{sim}\left(s_k, V_j\right)\right)^{\mu} \to 1$. In this case, $\prod_{k=1}^{n}\left(1 - \operatorname{sim}\left(s_k, V_j\right)\right)^{\mu} \geq \max\left\{\operatorname{sim}\left(s_k, V_j\right) \mid s_k \in V_j\right\}, CV\left(V_i, V_j\right) = 0$ exists for any cluster V_i and V_j. At this time, the cluster is the most strict, and all records can be used as separate clusters. On the contrary, if $\mu \to +\infty$ then $\prod_{k=1}^{n}\left(1 - \operatorname{sim}\left(s_k, V_j\right)\right)^{\mu} \to 0$ In this case, $\prod_{k=1}^{n}\left(1 - \operatorname{sim}\left(s_k, V_j\right)\right)^{\mu} < \max\left\{\operatorname{sim}\left(s_k, V_j\right) \mid s_k \in V_j\right\}$, $CV\left(V_i, V_j\right) = 1$ exists for any cluster V_i and V_j. At this time, the strictness of the cluster is the lowest, and all records can be clustered to obtain a whole cluster family. □

Different μ should be selected for different data sets to improve the quality of clustering, and the conditions for merging clustering should be redefined. The expression is shown in Formula (12):

$$\operatorname{Cluster}\left(V_i, V_j\right) \Leftrightarrow \left\{\max\left(G_i, G_j\right) \mid i \neq j\right\} \text{ and } CV\left(V_i, V_j\right) = 1. \tag{12}$$

3 DATA MINING ALGORITHM FOR WEB LEARNING RESOURCE INFORMATION FLOW LOSS BASED ON WEIGHTED DEPTH FOREST

3.1 Denoising of Lost Data of Web Learning Resource Information Flow

In general, signals and noises have different characteristics at different scales. For this reason, relevant scholars have proposed different filtering and denoising methods, mainly divided into Bayesian methods [17] and non Bayesian methods. Bayesian methods are subdivided into three categories: denoising methods based on modulus maximum principle, spatial scale correlation denoising methods and wavelet threshold denoising methods [18, 19]. Among the three wavelet denoising methods of signals, wavelet threshold denoising method is the first proposed typical nonparametric noise suppression method, and is the simplest and best method to achieve. The proposed algorithm uses two-dimensional wavelet threshold denoising to denoise the lost data of web learning resource information flow, which can more directly and accurately remove the noise in the data. It breaks down a data (a visual) into several spectral analysis at distinct in appearance levels (that is, multiresolution). This makes it possible to disclose an object's time and spectral properties concurrently. A potent method for eliminating distortion from diverse data is the wavelet transform (WT). Additional suppression might well be obtained by integrating WT with other noise-reduction strategies. Stochastic Variable Compression (SVD), like WT, is a powerful method for noise removal.

Let the loss data signal of a two-dimensional noisy web learning resource information flow be $g(x, y)$, which can be expressed as the $M \times N$ matrix. The expression is shown in Formula (13):

$$\boldsymbol{g}(x, y) = h(x, y) + \delta(x, y), \quad x = 1, 2, \ldots, M; y = 1, 2, \ldots, N, \tag{13}$$

where the noise component is $\delta(x, y)$, independent of $g(x, y)$, and the denoised signal component is expressed as $h(x, y)$.

The purpose of denoising is to find the signal closest to the real signal $h(x, y)$ according to $\mathbf{g}(x, y)$, that is, to obtain the estimated $\mathbf{g}(x, y)$ of $\mathbf{g}(x, y)$, to minimize its mean square error (MSE), as shown in Equation (14):

$$MSE = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{x=1}^{N} (\ \mathrm{g}(x, y) - \hat{\mathrm{g}}(x, y))^2. \tag{14}$$

In order to restore the real signal $h(x, y)$ from the lost data signal $\mathrm{g}(x, y)$ of the noisy web learning resource information flow, the wavelet transform is used as a tool to realize denoising. By processing the decomposition coefficient of the wavelet, the signal and noise can be separated according to the different characteristics of the data signal and noise, overcoming the limitations of traditional methods when processing non-stationary signals [20, 21].

After discrete sampling of the lost data signal $g(x, y)$ of the two-dimensional noisy web learning resource information flow, obtain the $M \times N$ point discrete signals $g(i, j)$, $i = 1, 2, \ldots, M; j = 1, 2, \ldots, N$. In the wavelet threshold, expand the orthogonal wavelet transform of Equation (13) to obtain Equation (15):

$$U[i, j] = C[i, j] + B[i, j], \quad i = 1, 2, \ldots, M; j = 1, 2, \ldots, N. \tag{15}$$

The noisy wavelet coefficient is $U[i, j]$ and the denoising wavelet coefficient is $C[i, j]$. The matrix representation of the corresponding signal: $g = \{g(i, j)\}_{ij}$, $h = \{h(i, j)\}_{ij}$, $\delta = \{\delta(i, j)\}_{ij}$, g, h, δ wavelet transform coefficient matrices are: $U = Eh$, $C = Eg$, $B = E\delta$, where E is a two-dimensional orthogonal wavelet transform operator. After processing the original signal $g(x, y) = h(x, y) + \delta(x, y)$ with two-dimensional discrete wavelet transform, according to the linear property of the wavelet transform the wavelet coefficients corresponding to signal $h(x, y)$ and noise $\delta(x, y)$ constitute the decomposed wavelet coefficients. Generally, these wavelet coefficients are divided into subbands according to different scales and directions, as shown in Figure 3.

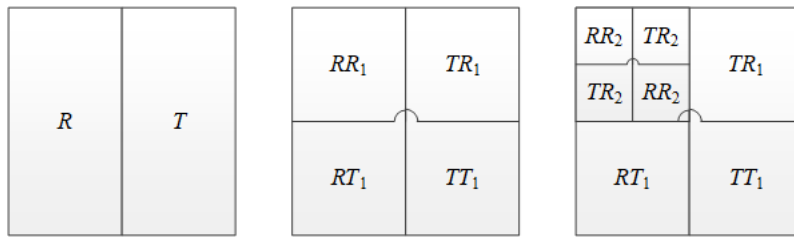

Figure 3. Structure of 2-D wavelet decomposition coefficient

In Figure 3, the high frequency is T, the low frequency is R; the subband RR is the low-frequency approximation coefficient, and the subband TT_j, TR_j, RT_j, $j = 1, 2, \ldots, J$ is the high-frequency detail coefficient, where the decomposition scale is j, the maximum decomposition scale is J, and the subband size on the j scale is $\frac{N}{2^j} \times \frac{N}{2^j}$.

Figure 3 shows the process of sub-band segmentation in the data decomposition of web learning resource information flow loss. The total number of coefficients after decomposition equals the number of original data. After the first decomposition layer, the original data is overwritten and replaced by a low-frequency subband R and a high-frequency subband T. After the first decomposition layer, R is decomposed into RR and RT, and T is decomposed into TR and TT. The decomposition of the next layer retains RT, TR and TT, and only continues to decompose RR into four subbands. The subscripts distinguish the subbands of different levels in the figure. The core of wavelet denoising uses different wavelet bases to transform the original signal into the wavelet threshold [22, 23, 24]. It is based on the various effects and distributions of signal $h(x, y)$ and noise $\delta(x, y)$ on (i, j) wavelet coefficients

of the location points, and it is used to judge the contribution of wavelet coefficients in signal and noise energy.

The idea of wavelet threshold denoising method is to decompose the lost data signal of web learning resource information flow by wavelet decomposition, select the appropriate threshold Y among the wavelet coefficients of different scales, apply the rule to process the coefficients whose modulus is greater than or less than this threshold, obtain the estimated wavelet coefficients, and then reconstruct them so that the noise in the lost data signal of web learning resource information flow can be effectively controlled.

3.2 Weighted Depth Forest Construction Method

For the deep forest algorithm [25], when constructing the cascaded forest module, the feature selection is very random, which is likely to reduce the accuracy of the forest and affect the algorithm's performance. The weighted deep forest-based web learning resource information flow loss data mining algorithm proposes a construction method for weighted deep forest. The following definitions are given to better describe the method.

Definition 5. Set the lost data set $F = \{Z_1, Z_2, \ldots, Z_m\}$, the category is $V = \{V1, V2, \ldots, Vn\}$, and the prediction probability matrix of the forest is shown in Equation (16):

$$Q = \begin{bmatrix} Q_1 & Q_{12} & \ldots & Q_{1n} \\ Q_{21} & Q_{22} & \ldots & Q_{2n} \\ \vdots & \vdots & & \vdots \\ Q_{m1} & Q_{m2} & \ldots & Q_{mn} \end{bmatrix}. \tag{16}$$

In the formula, the probability that the i data is divided into the j data is Q_{ij}. The subscript of the column of the maximum value in each row of the forest prediction probability matrix is considered as the final prediction category of this data, which is marked as 1 in the matrix and the rest as 0, as shown in Formula(17):

$$Q = \begin{bmatrix} 0 & 0 & \ldots & 1 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 \end{bmatrix}. \tag{17}$$

Then the accuracy of the forest in the algorithm is expressed as Equation (18):

$$A = \frac{\sum_{i=0}^{m} \sum_{j=1}^{n} W[i][j] \cap Q[i][j]}{m}. \tag{18}$$

In the formula, the total number of data is m, the number of categories is n, the actual category matrix of the lost data samples of the web learning resource

information flow is $W[i][j]$, and the prediction probability matrix is $Q[i][j]$. The accuracy of prediction for each forest is obtained through cross-operation.

Definition 6. Assuming forest $\mathrm{T} = \{1, 2, \ldots, t\}$, the definition of forest weight factor λ is shown in Formula (19):

$$\lambda = \frac{\lg_2 A_i}{\sum_{i=1}^{r} \lg_2 A_i}. \tag{19}$$

In the formula, the accuracy of the i^{th} forest is A_i, then the class prediction probability matrix of the i^{th} forest is expressed as Formula (20):

$$Q_{(i)} = Q \times \lambda. \tag{20}$$

When constructing the weighted depth forest, the cascade forest module should be analyzed first. The next is the cascading forests, which, combined with random forests, develop increasingly exclusionary interpretations while supervised by incoming interpretations at every level. The random forest and completely random forest used in the algorithm are random selection features. For a much more precise forecast, Random Forest produces numerous decision forests that are then combined. It is based on the idea that several statistically independent systems (the different decision trees) work significantly stronger together than they do separately. After Definition 5, the forest prediction probability matrix can be obtained, and the forest accuracy can be calculated. Then, the weight factor [26, 27] in Definition 6 is introduced to weigh the forest probability results. The higher the accuracy is, the higher the weight factor is, and the greater the forest weight is. On the contrary, the smaller the weight is. This step can effectively reduce the impact of irrelevant attributes in the randomly selected root node of the feature on the algorithm performance. The weighted probability vector is used as the input to expand the next forest cascade layer to continue training, and the above processing process is repeated until the maximum number of cascaded layers is reached.

3.3 Data Loss Detection of Web Learning Resource Information Flow

To better apply the weighted depth forest algorithm [28] to the lost data mining of web learning resource information flow, it is necessary to redefine the isolation factor as follows:

Definition 7. If the formal background is binary $\{F, V\}$, the record set is $F = \{Z_y | y = 1 - n\}$, and definition of the isolation factor β is as shown in Formula (21).

In the formula, the distance between the y point and the first data point is $f(Z_1, Z_1)$, the class density of the i type data is $\sigma(V_i)$ and the definition of the threshold function $G(\tau)$ of the isolation factor β is shown in Formula (21).

$$G(\tau) = \frac{1}{\tau} \times \log_{10} \tau, \tag{21}$$

where τ represents the number of categories $\tau > 0$ in the dataset, and the threshold function $G(\tau)$ is a monotone decreasing function. τ increases and $G(\tau)$ decreases.

After the weighted depth forest construction is completed, the distribution of the lost data of the web-learning resource information flow is different, and the data is divided into classes with different densities. According to Definition 7, the class density of the lost data of various web learning resource information flows can be obtained. The more data there is, the smaller the class density. The local isolation factor β is calculated by Formula (21). The larger the β, the greater the probability of losing the web learning resource information flow. The above formula can also obtain the isolation factor function threshold $G(\tau)$ in different data sets. The data whose β value is greater than $G(\tau)$ is the lost data of web learning resource information flow.

3.4 Algorithm Description

The lost data mining algorithm of web learning resource information flow based on weighted depth forest is divided into two parts. The first part is to build a weighted depth forest, and the second is to detect the lost data of web learning resource information flow. In the first part, scan the data through the sliding window [29], divide the data into multiple different dimensions, enter the random forest and the utterly random forest for training, and combine the obtained class probability vectors as the input of the cascaded weighted forest module. In the cascaded weighted forest part, the input data enters the forest integration part for training, including two random forests and two completely random forests. To prevent the occurrence of overfitting, the K-fold cross-validation method is used to train them, and the Definition 5 method is used to obtain the forest prediction probability matrix. Nevertheless, all portions of the information can be utilized as a test dataset whenever K-fold cross-testing is applied. This enables them to assess the system's effectiveness by using all of the information from our tiny sample for both ongoing training. The prediction result is the subscript corresponding to the maximum value of each row in the matrix. Then, the accuracy of the forest is calculated, and the weight factor in Definition 6 is added to weight the forest. The weight factor of the forest is proportional to the weight value. The larger the weight factor, the greater the weight value. Combine the weighted probability vectors of each forest with the probability results of the granularity scanning layer, take them as the input of the next forest cascade layer, calculate the error rate of this layer, and recurse the above steps until the maximum number of layers is reached, or the error rate of this layer increases, and the iteration stops.

After the weighted forest is constructed, the isolation factor β and the threshold function $G(\tau)$ of the dataset are calculated by defining 7 pairs of sample data of each type in the dataset. The $\beta > G(\tau)$ data is the lost data of the web learning resource information flow.

4 EXPERIMENT AND DISCUSSION

In order to verify the overall effectiveness of the weighted depth forest based web learning resource information flow loss data mining algorithm, it is necessary to carry out relevant tests. The classification of lost data in web learning resource information flow has a direct impact on its mining. Figure 4 shows the classification and division results of lost data in web learning resource information flow using the proposed algorithm, reference [3] algorithm, reference [4] algorithm, and reference [5] algorithm.

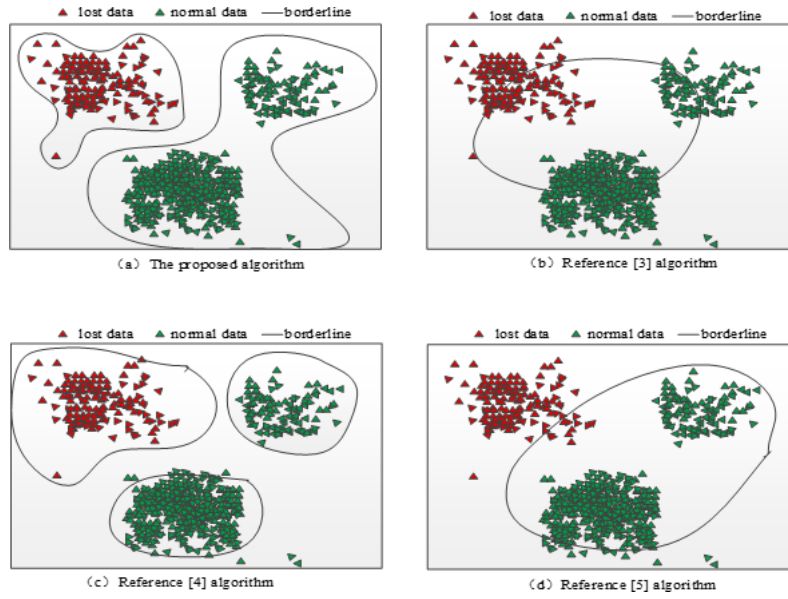

Figure 4. Classification results of different algorithms

It can be seen from Figure 4 that the classification and division results of the lost data of the web learning resource information flow by different algorithms are completely different. The proposed algorithm has the best classification and division effect on the lost data of the web learning resource information flow. Other algorithms have problems such as too broad data division, unclassified identification of lost data, and incomplete classification and division of data, resulting in a high error rate. The proposed algorithm does not have the above problems; because the proposed algorithm carries out relevant pre-processing on the lost data of web learning resource information flow and then cluster analysis on it, and classifies the standard data and the lost data separately, it can efficiently complete the classification and division of the lost data of web learning resource information flow.

Figures 5, 6, 7 and 8 show the extraction results of the web learning resource information flow loss data signal using the proposed algorithm, the reference [3]

algorithm, the reference [4] algorithm, and the reference [5] algorithm, taking the clarity of the extracted data signal as an indicator.

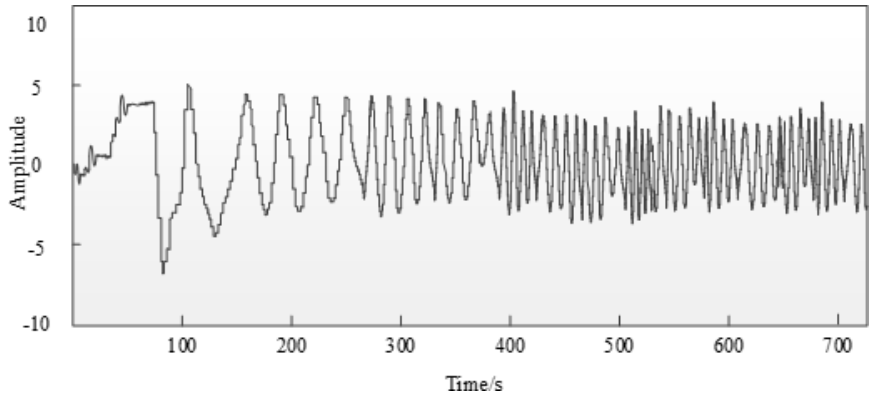

Figure 5. Data signal extracted by the proposed algorithm

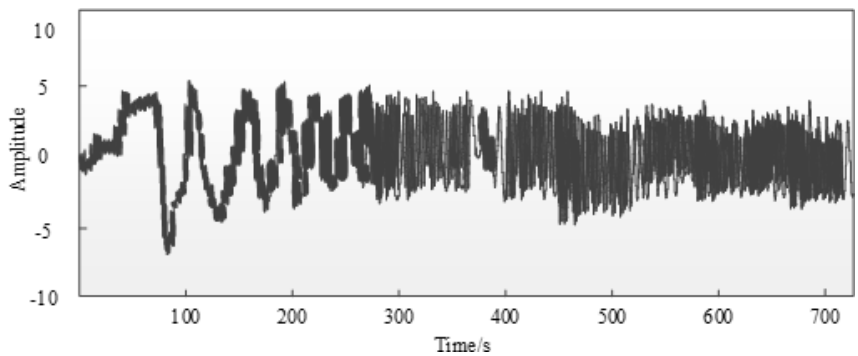

Figure 6. Data signal extracted by reference [3] algorithm

It can be seen from Figures 5, 6, 7 and 8 that the data loss signal of the web learning resource information flow extracted by the proposed algorithm is the most accurate. The data loss signal of the web learning resource information flow extracted by the reference [3] algorithm, the reference [4] algorithm and the reference [5] algorithm all have different levels of noise, especially the noise of the reference [4] algorithm is the most obvious, indicating that the proposed algorithm has high accuracy and high performance.

Table 1 shows the operation time of the proposed algorithm, reference [3] algorithm, reference [4] algorithm, and reference [5] algorithm, taking the operation time of web learning resource information flow loss data mining as an indicator.

According to the above table, with the increase in the number of experiments, the operation time of the four algorithms for data mining of web learning resource

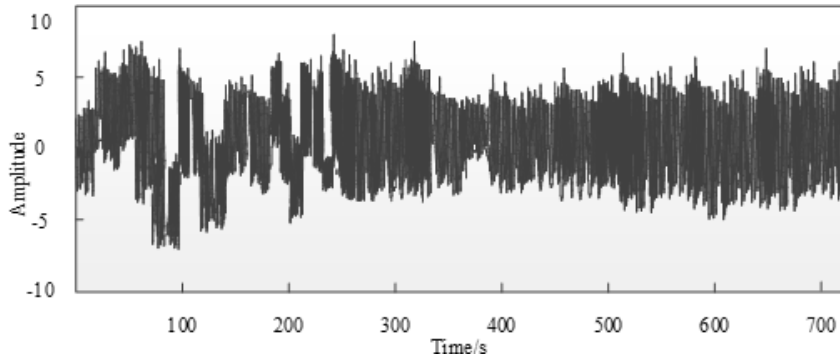

Figure 7. Data signal extracted by reference [4] algorithm

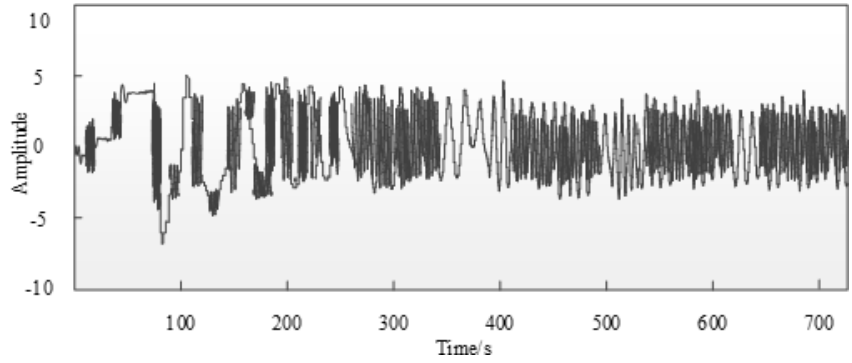

Figure 8. Data signal extracted by reference [5] algorithm

information flow loss also increases. The operation time of the proposed algorithm is the lowest, with the highest value of 11.64 s. The operation time of the other three algorithms is about 10 times that of the proposed algorithm. In particular, the operation time of the algorithm in reference [5] is up to 355.49 s, which is about 30 times that of the proposed algorithm, indicating that the proposed algorithm is labour-intensive and efficient.

Number of Experiments	Running Time [s]			
	Proposed Algorithm	Reference [3] Algorithm	Reference [4] Algorithm	Reference [5] Algorithm
1	8.76	206.88	90.77	268.14
2	9.41	224.32	95.82	299.63
3	10.37	257.81	99.34	304.82
4	11.64	276.27	106.19	355.49

Table 1. Operation time of data mining for information flow loss of web learning resources

5 CONCLUSION

Aiming at the problems existing in the current data mining algorithm for web learning resource information flow loss, a data mining algorithm is proposed based on a weighted depth forest. First, a brand-driven web data collection model is constructed to collect data; second, data features are extracted, wavelet threshold denoising is carried out for data, and finally, a weighted depth forest is constructed to detect web learning resource information flow loss data, complete the data mining of web learning resource information flow loss. It provides a new direction for internet information processing in web learning resource information flow loss data mining.

Declarations

Funding:

1. General Project of Outstanding Young Talents Support Plan in Colleges and Universities (gxyq2022294);
2. Key Research Project of Natural Science in Colleges and Universities of Anhui Province (KJ2021A1593).

Conflict of interest: The authors have no conflict of interest.

Data availability: All data generated or analysed during this study are included in the manuscript.

Code availability: Not applicable.

Author's contributions: Shuling Zhou contributed to the design and methodology of this study, the assessment of the outcomes, and the manuscript's writing.

REFERENCES

[1] Namouchi, S.—Farah, I. R.: Graph-Based Classification and Urban Modeling of Laser Scanning and Imagery: Toward 3D Smart Web Services. Remote Sensing, Vol. 14, 2022, No. 1, Art. No. 114, doi: 10.3390/rs14010114.

[2] Luo, C.—Tan, Z.—Min, G.—Gan, J.—Shi, W.—Tian, Z.: A Novel Web Attack Detection System for Internet of Things via Ensemble Classification. IEEE Transactions on Industrial Informatics, Vol. 17, 2021, No. 8, pp. 5810–5818, doi: 10.1109/TII.2020.3038761.

[3] Sun, T.—Hu, J.—Huang, J.—Fan, Y.: Bandwidth Resource Prediction and Management of Web Applications Hosted on Cloud. Journal of Computer Applications, Vol. 40, 2020, No. 1, pp. 181–187, doi: 10.11772/j.issn.1001-9081.2019050903 (in Chinese).

[4] Liang, W.—Chen, H.—Zhang, J.—Zhao, D.—Li, C.: An Effective Scheme for Top-k Frequent Itemset Mining Under Differential Privacy Conditions. Science China

Information Sciences, Vol. 63, 2020, No. 5, Art. No. 159101, doi: 10.1007/s11432-018-9849-y.

[5] Yang, L.—Wang, C.: Web Learning Resource Data Stream Processing Technology Based on Data Mining. Modern Electronics Technique, Vol. 45, 2022, No. 13, pp. 62–66, doi: 10.16652/j.issn.1004-373x.2022.13.012 (in Chinese).

[6] Zaitsev, D. A.—Shmeleva, T. R.—Pröll, B.: Spatial Specification of Hypertorus Interconnect by Infinite and Reenterable Coloured Petri Nets. International Journal of Parallel, Emergent and Distributed Systems, Vol. 37, 2022, No. 1, pp. 1–21, doi: 10.1080/17445760.2021.1952580.

[7] Iqbal, S.—Shahid, A.—Roman, M.—Khan, Z.—Al-Otaibi, S.—Yu, L.: TK-FIM: Top-K Frequent Itemset Mining Technique Based on Equivalence Classes. PeerJ Computer Science, Vol. 7, 2021, Art. No. e385, doi: 10.7717/peerj-cs.385.

[8] Mei, Y.—Zeng, Z.—Ye, J.: A Computing Model: The Closed-Loop Optimal Control for Large-Scale One-of-a-Kind Production Based on Multilevel Hierarchical PERT-Petri Net. IEEE Transactions on Engineering Management, Vol. 68, 2021, No. 6, pp. 1637–1649, doi: 10.1109/TEM.2020.3035230.

[9] Pajany, M.—Zayaraz, G.: A Robust Lightweight Data Security Model for Cloud Data Access and Storage. International Journal of Information Technology and Web Engineering (IJITWE), Vol. 16, 2021, No. 3, pp. 39–53, doi: 10.4018/IJITWE.2021070103.

[10] Liang, W.—Li, K. C.—Long, J.—Kui, X.—Zomaya, A. Y.: An Industrial Network Intrusion Detection Algorithm Based on Multifeature Data Clustering Optimization Model. IEEE Transactions on Industrial Informatics, Vol. 16, 2020, No. 3, pp. 2063–2071, doi: 10.1109/TII.2019.2946791.

[11] Malarvizhi, K.—Amshakala, K.: Feature Linkage Weight Based Feature Reduction Using Fuzzy Clustering Method. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, Vol. 40, 2021, No. 3, pp. 4563–4572, doi: 10.3233/JIFS-201395.

[12] Sornalakshmi, M.—Balamurali, S.—Venkatesulu, M.—Navaneetha Krishnan, M.—Ramasamy, L. K.—Kadry, S.—Manogaran, G.—Hsu, C. H.—Muthu, B. A.: Hybrid Method for Mining Rules Based on Enhanced Apriori Algorithm with Sequential Minimal Optimization in Healthcare Industry. Neural Computing and Applications, Vol. 34, 2022, pp. 10597–10610, doi: 10.1007/s00521-020-04862-2.

[13] Ahmed, U.—Lin, J. C. W.—Srivastava, G.—Aleem, M.: A Load Balance Multi-Scheduling Model for OpenCL Kernel Tasks in an Integrated Cluster. Soft Computing, Vol. 25, 2021, No. 1, pp. 407–420, doi: 10.1007/s00500-020-05152-8.

[14] Feng, J.—Yao, Y.: An Optimization Clustering Algorithm Based on Multi-Population Genetic Simulated Annealing Algorithm. Computer Simulation, Vol. 37, 2020, No. 9, pp. 226–230 (in Chinese).

[15] Prajapat, R.—Yadav, R. N.—Misra, R.: Energy-Efficient k-Hop Clustering in Cognitive Radio Sensor Network for Internet of Things. IEEE Internet of Things Journal, Vol. 8, 2021, No. 17, Art. No. 13593–13607, doi: 10.1109/JIOT.2021.3065691.

[16] Liu, J.—Wu, D.—Wang, Z.—Jin, X.—Dong, F.—Jiang, L.—Cai, C.: Auto-

matic Sleep Staging Algorithm Based on Random Forest and Hidden Markov Model. Computer Modeling in Engineering & Sciences, Vol. 123, 2020, No. 1, pp. 401–426, doi: 10.32604/cmes.2020.08731.

[17] Josephine, S.—Murugan, S.: Noise Removal from Brain MRI Images Using Adaptive Bayesian Shrinkage. Journal of Computational and Theoretical Nanoscience, Vol. 17, 2020, No. 4, Art. No. 1818–1825, doi: 10.1166/jctn.2020.8446.

[18] Kumar, B. S.—Sivaparthipan, C. B.—Kalaikumaran, T.—Karthik, S.: A Case Study of Customer Relationship Management Using Data Mining Techniques. International Journal of Technological Exploration and Learning, Vol. 2, 2013, No. 6, pp. 275–80.

[19] Liu, Y.—Lu, X.—Bei, G.—Jiang, Z.: Improved Wavelet Packet Denoising Algorithm Using Fuzzy Threshold and Correlation Analysis for Chaotic Signals. Transactions of the Institute of Measurement and Control, Vol. 43, 2021, No. 6, pp. 1394–1403, doi: 10.1177/0142331220979229.

[20] Sharma, V.: A Review on Vibration-Based Fault Diagnosis Techniques for Wind Turbine Gearboxes Operating Under Nonstationary Conditions. Journal of the Institution of Engineers (India), Series C, Vol. 102, 2021, No. 2, pp. 507–523, doi: 10.1007/s40032-021-00666-y.

[21] Bozhokin, S. V.—Barantsev, K. A.—Litvinov, A. N.: Method of Translation Transfer for Estimation of Stability of a Nonstationary Quantum Frequency Standard. Technical Physics, Vol. 66, 2021, No. 1, pp. 28–33, doi: 10.1134/S1063784221010035.

[22] Si, Y.—Zhang, Z.—Kong, L.—Zheng, J.: Condition Monitoring of Deep-Hole Drilling Process Based on Improved Empirical Wavelet De-Noising and High Multiple Frequency Components of Rotation Frequency. The International Journal of Advanced Manufacturing Technology, Vol. 114, 2021, No. 7-8, pp. 2201–2214, doi: 10.1007/s00170-021-06965-z.

[23] Lu, H.—Sivaparthipan, C. B.—Antonidoss, A.: Improvement of Association Algorithm and Its Application in Audit Data Mining. Journal of Interconnection Networks, Vol. 22, 2022, No. Supp03, Art. No. 2144002, doi: 10.1142/S0219265921440023.

[24] Kompella, K. C. D.—Rayapudi, S. R.—Rongala, N. S.: Investigation of Bearing Faults in Three Phase Induction Motor Using Wavelet De-Noising with Improved Wiener Filtering. International Journal of Power and Energy Conversion (IJPEC), Vol. 12, 2021, No. 2, pp. 115–136, doi: 10.1504/IJPEC.2021.114484.

[25] Tang, J.—Xia, H.—Qiao, J.—Guo, Z.: Soft Measurement of Dioxin Emission Concentration Based on Deep Forest Regression Algorithm. International Journal of System Control and Information Processing, Vol. 3, 2021, No. 3, pp. 208–228, doi: 10.1504/IJSCIP.2021.117695.

[26] Bukenberger, J. P.—Webster, M. D.: Approximate Latent Factor Algorithm for Scenario Selection and Weighting in Transmission Expansion Planning. IEEE Transactions on Power Systems, Vol. 35, 2020, No. 2, pp. 1099–1108, doi: 10.1109/TPWRS.2019.2942925.

[27] Wang, F.—Xie, H.—Chen, Q.—Davari, S. A.—Rodríguez, J.—Kennel, R.: Parallel Predictive Torque Control for Induction Machines Without Weighting Fac-

tors. IEEE Transactions on Power Electronics, Vol. 35, 2020, No. 2, pp. 1779–17889, doi: 10.1109/TPEL.2019.2922312.

[28] Jiang, J.—Li, W.—Wen, Z.—Bie, Y.—Schwarz, H.—Zhang, C.: Series Arc Fault Detection Based on Random Forest and Deep Neural Network. IEEE Sensors Journal, Vol. 21, 2021, No. 15, pp. 17171–17179, doi: 10.1109/JSEN.2021.3082294.

[29] Zhu, M.—Mitchell, D. G. M.—Lentmaier, M.—Costello, D. J.—Bai, B.: Error Propagation Mitigation in Sliding Window Decoding of Braided Convolutional Codes. IEEE Transactions on Communications, Vol. 68, 2020, No. 11, pp. 6683–6698, doi: 10.1109/TCOMM.2020.3015945.

Shuling Zhou is working as Associate Professor. She graduated from the Northeast Agricultural University in 2003, majoring in information and computing science. In 2016, she was awarded her Master's degree from the Hefei University of Technology, majoring in computer technology. Now she is working in the Hefei College of Finance and Economics. Her research fields include data mining and software engineering. She has published nine academic papers. Meanwhile, she presided over and participated in six collaborative research projects.