Hindawi
Complexity
Volume 2018, Article ID 9702304, 12 pages
https://doi.org/10.1155/2018/9702304

Research Article

Mining Outlier Data in Mobile Internet-Based Large Real-Time Databases

Xin Liu, Yanju Zhou, and Xiaohong Chen

School of Business, Central South University of China, Changsha 410083, China

Correspondence should be addressed to Yanju Zhou; zyj4258@sina.com

Received 8 May 2017; Accepted 7 November 2017; Published 10 January 2018

Academic Editor: Eulalia Martínez

Mining outlier data guarantees access security and data scheduling of parallel databases and maintains high-performance operation of real-time databases. Traditional mining methods generate abundant interference data with reduced accuracy, efficiency, and stability, causing severe deficiencies. This paper proposes a new mining outlier data method, which is used to analyze real-time data features, obtain magnitude spectra models of outlier data, establish a decisional-tree information chain transmission model for outlier data in mobile Internet, obtain the information flow of internal outlier data in the information chain of a large real-time database, and cluster data. Upon local characteristic time scale parameters of information flow, the phase position features of the outlier data before filtering are obtained; the decision-tree outlier-classification feature-filtering algorithm is adopted to acquire signals for analysis and instant amplitude and to achieve the phase-frequency characteristics of outlier data. Wavelet transform threshold denoising is combined with signal denoising to analyze data offset, to correct formed detection filter model, and to realize outlier data mining. The simulation suggests that the method detects the characteristic outlier data feature response distribution, reduces response time, iteration frequency, and mining error rate, improves mining adaptation and coverage, and shows good mining outcomes.

1. Introduction

With the rapid development of broadband wireless access (BWA) technologies and mobile terminals, the currently emerging mobile Internet integrates both mobile communication and Internet access. The term "mobile Internet" is a generic term that refers to implementations of activities that combine Internet technologies, platforms, commercial patterns, and applications with mobile communication technologies. As the mobile Internet has developed, applying mobile technology to a large real-time database can significantly improve the database's operational efficiency. However, many potential safety hazards can occur during operation of large real-time databases under mobile Internet conditions. Therefore, to avoid security vulnerabilities, investigating how to effectively monitor and mine outlier data in mobile Internet-based large real-time databases has become hot research topics [1–3].

Traditionally, mining methods that integrate clustering of mapping perturbation searches and the fuzzy C mean value have been adopted for outlier data mining in large mobile Internet-based real-time databases. However, these methods neglect the complexities of large real-time databases in mobile Internet conditions and limit the efficiency with which outlier data can be detected in such databases [4]. A cosine function-based improved logistic model was presented in literature [5] that analyzed chaotic time series using nonlinear time series analysis, constructed an online method called fuzzy least square support vector machine (FLS-SVM), and realized outlier data mining in a mobile Internet-based large real-time database. However, time scale impacts must be further considered for this method, and its confidence probability needs to be improved to increase the accuracy of data mining. Moreover, the required computation process is too complex to be adopted in practice [5]. A decision-tree outlier-feature classification-based method for mining outlier data in a mobile Internet-based large real-time database was presented in literature [6]; however, this method tends to be affected by substantial amounts of interference data, which compromises the confidence coefficient

and accuracy of the mining algorithm [6]. Other proposed methods for mining outlier data in a mobile Internet-based large real-time database include a support vector machine (SVM) based method [7], a firefly algorithm-based method [8], a clustering and rapid computation-based method [9], an association-rules detection-based method [10], and a knowledge granularity-based method [11]. However, all the algorithms based on these traditional methods have shortcomings (e.g., poor outlier data mining accuracy, large errors, and excessive execution times) that require further improvement [12–15].

Aiming to solve the shortcomings of these traditional methods, a decision-tree outlier-classification feature-filtering detection-based method is presented in this paper to mine outlier data in a mobile Internet-based large real-time database. This method's advantages are that it can construct a magnitude spectra model of outlier data by analyzing real-time data features. Based on this model, the preliminary signals of outlier data are analyzed to obtain an information chain in which outlier data exist. The method improves the efficiency of subsequent outlier detection. A transmission model for a decision-tree information chain using internal outlier data under mobile Internet conditions is constructed to precisely cluster mobile and scattered data and obtain an information chain with outlier data, which is then introduced into the database in an orderly fashion and, therefore, supports the improved precision of subsequent mining. An information flow with internal outlier data in an information chain is obtained in a large real-time database and undergoes clustering to further improve its precision, reduce the complexity of mining computation, and lay the foundation for a later division of the confidence interval. Later, using the improved mining method, the phase position characteristic of the prefiltered outlier data is obtained in outlier data-contained information flows, and a decision-tree outlier-classification-characteristic filtering algorithm is used to obtain an analytic signal and a constant amplitude. Then, the decision-tree filtering method is used to screen out useless high- and low-frequency components and to achieve the phase-frequency characteristics of outlier data. WT threshold denoising is combined with signal denoising, which greatly reduces the response time, iteration frequency, and error rate of mining, improves mining adaptation and precision, and corrects the detection filter model. The method effectively improves the precise coverage and probability of outlier data mining and achieves outlier data mining in a mobile Internet-based large real-time database.

2. Constructing an Information Transmission Model for a Decision-Tree Information Chain with Outlier Data and Acquiring Information Flows

2.1. Constructing an Information Transmission Model for a Decision-Tree Information Chain with Outlier Data under Mobile Internet Conditions. Assume that $A(t)$ represents the information amplitude of the outlier data and that f_0 refers to the initial data frequency under mobile Internet conditions. Then, the model of the outlier data amplitude spectrum under mobile Internet conditions is shown below:

$$s(t) = A(t)\,\mathrm{rect}\left(\frac{t}{T}\right)\exp\left[j\left(2\pi f_0 t + \pi k t^2\right)\right], \quad (1)$$

where $\mathrm{rect}(t/T)$ represents the average time for detection of outlier data under mobile Internet conditions, j is a constant, and $k = B/T$ refers to the frequency modulation slope of the outlier data information under mobile Internet conditions. Here, B refers to the distribution bandwidth of the outlier data under mobile Internet conditions, and T refers to the refresh cycle of data under mobile Internet conditions.

The decision-tree cross-term information chain is adopted for data under the mobile Internet using the following expression:

$$f(t) = f_0 + kt. \quad (2)$$

In the above formula, f represents global frequency of the information and shows a linear relationship with the slope of information frequency modulation in the scope of time.

Formulas (1) and (2) are used to screen outlier data-carrying information chains under the mobile Internet, remove interference information chains, conduct dynamic analyses, and process the outlier data-carried information chains [16–18].

Under mobile Internet conditions, assume that $a(t)$ and $\theta(t)$ represent the bandwidth and time duration of outlier data, respectively, $x(t)$ represents the band internal frequency spectrum of outlier data [19], and $y(t)$ represents the intrinsic mode function of outlier data under mobile Internet. Then, formula (3) is used to establish a transmission model for the outlier data-carried decision-tree information chains [20].

$$z(t) = x(t) + iy(t) + a(t)\,e^{i\theta(t)}. \quad (3)$$

In the preceding formula, i represents a constant. During mobile Internet operations, the instantaneous frequency of outlier data shows a linear correlation with time within the same pulse width [21, 22]. The magnitude spectra of the outlier data described with the model of formula (1) are reflected in the information chain table required for the topological structure of formula (2), and information chains with outlier data are transmitted via formula (3). The transmission model for decision-tree information chains with outlier data under mobile Internet is shown in Figure 1.

2.2. Acquiring Outlier-Data-Carrying Information Flows in a Large Real-Time Database. In Section 2.1, based on acquiring and transmitting the outlier data-containing information flows under mobile Internet, the outlier data-containing information flows are acquired for the information chains transmitted to a large real-time database. The specific steps are as follows.

First, a time domain and spatial domain analysis are conducted for the information chains transmitted to the large

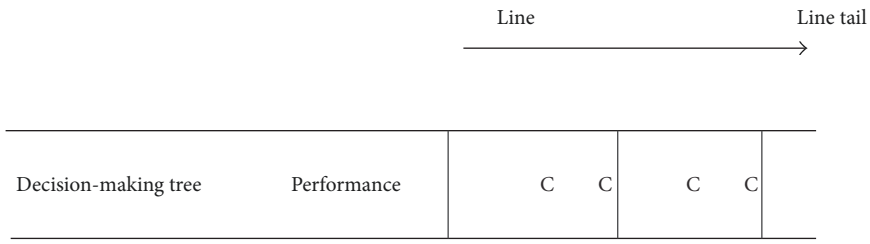


Figure 1: Transmission model for outlier-data-carrying decision-tree information chain under mobile Internet conditions.

real-time database. Formula (4) is used to obtain chirp-signal amplitude-frequency [23] features of the outlier data-containing information flows in the information chains:

$$|s(f)| = A\sqrt{\frac{1}{2k}}\left\{\left[c(v_1)+c(v_2)\right]^2+\left[s(v_1)+s(v_2)\right]^2\right\}, \tag{4}$$

where $c(v_1) + c(v_2)$ and $s(v_1) + s(v_2)$ represent the attribute value range and integrity of outlier data-contained information flows in the information chains, respectively, while A and k, respectively, represent the number of elements and the element weights of outlier data in the information flows.

Assume that f represents the global frequency of the outlier data-contained information flows of the information chains in the large real-time database. Then, the global frequency feature can be expressed as follows:

$$\varphi_l(f) = -\frac{\pi(f-f_0)^2}{k} + \arctan\frac{s(v_1)+s(v_2)}{c(v_1)+c(v_2)}. \tag{5}$$

Second, based on acquiring the global frequency feature of the information flows, the C4.5 decision-tree model [24] is introduced for information flow abnormal classification feature modeling. In addition, the chirp-signal amplitude-frequency features presented in formula (4) are used for the signal analytical model for information flows. Then, formula (6) is utilized to obtain the enveloping feature of outlier data-contained information flows in the information chains of the large real-time database:

$$a(t) = \sqrt{x^2(t)+y^2(t)}, \quad \theta(t) = \arctan\frac{y(t)}{x(t)}. \tag{6}$$

In the above formula, $\theta(t)$ represents the high-frequency component of the decision tree and $a(t)$ and $\theta(t)$ are the interference feature amplitude and phrase information of the large real-time database, respectively. Formula (7) is then used to obtain the interfering frequency feature formed by outlier data-containing information flows in the information chains:

$$f(t) = \frac{1}{2\pi}\times\frac{d\theta(t)}{dt}. \tag{7}$$

The preceding formula describes the check-bit generated by outlier data-containing information flows in the information chains in the large real-time database that represent the data interference frequency. On this basis, formula (8) is used to calculate the weight probability of outlier data-containing information flows in the information chains:

$$w_{ij} = \beta \times w\left(e_p k_q\right) \quad (\beta > 1). \tag{8}$$

Third, after preliminary data screening, dynamic mining is conducted for outlier data-carried information flows. The interference ratio is calculated for every decision-making root node *SIR*, and the information content of the time domain waveform for a fixed frequency band is obtained as follows:

$$r_1 = x(t) - c_1. \tag{9}$$

Finally, based on the processing steps above, the outlier data-containing information flows in the information chains of the large real-time database are obtained, forming a basis for the design of a decision-tree outlier-feature-classification algorithm-based outlier data mining method.

2.3. Clustering Outlier Data-Containing Information Flows in a Large Real-Time Database. After acquiring the outlier data-containing information flows as presented in the preceding sections, the information flows are clustered, and the MOC algorithm [25] optimizes the evaluation criteria for the two clusters using the Xie-Beni (XB) [26] index and the FCM objective function [27] J_{FCM}, respectively, and effectively processes noise contained in the information flow set as well overlapping information flow aggregates. Therefore, the MOC algorithm is adopted to cluster the outlier data-containing information flows.

The (XB) index measures in-class compactness and interclass separation after the division of the outlier data-containing information flow clusters and is defined as the ratio of the internal compactness of the outlier data-containing information flow aggregate J_{F_COMP} to the minimal distance among information flow aggregates J_{F_SEP} [28]:

$$\mathrm{XB} = \frac{J_{F\_\mathrm{COMP}}}{N\times J_{F\_\mathrm{SEP}}}. \tag{10}$$

Then, the optimized FCM objective function J_{FCM} is expressed as follows:

$$J_{\mathrm{FCM}} = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^m d^2\left(x_j, x_i\right). \tag{11}$$

During the iterative process of the MOC algorithm, outlier data-containing information flows are set first, before setting the weighting coefficient of every outlier data aggregate, W, and the cluster center, V. Then, the fuzzy membership (u_{ij}) of the sample to every cluster center is calculated according to the following formula:

$$u_{ij} = \frac{\left(d_{ij}\right)^{-1/m-1}}{\sum_{i=1}^{C}\left(d_{ij}\right)^{-1/\tau}}, \quad i = 1, \ldots, C, \ j = 1, \ldots, N. \tag{12}$$

In formula (12), m denotes the fuzzy clustering index, and τ refers to the weighted index. The cluster center v_{ik} of MOC algorithm is calculated as follows:

$$v_{ik} = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_{jk}}{\sum_{j=1}^{N} u_{ij}^{m}}. \tag{13}$$

Note that the preceding formula is used to calculate the cluster center of every outlier information flow aggregate during the clustering process. When the variance of the cluster center is below 10^{-3}, all cluster centers are reinitialized to obtain new result of MOC algorithm cluster center w_{ik}, as shown in the following formula:

$$w_{ik} = \frac{\sum_{j=1}^{N} u_{ij}^{m}\left(x_{jk} - v_{ik}\right)^{2}}{\sum_{k=1}^{D}\sum_{j=1}^{N} u_{ij}^{m}\left(x_{jk} - v_{ik}\right)^{2}}. \tag{14}$$

In conclusion, the MOC algorithm is used to cluster outlier data-containing information flows in a large real-time database. the clustered information requires further data mining.

3. Improvement of Outlier Data Mining in a Mobile Internet-Based Large Real-Time Database

The procedures discussed in the previous section improve the method for outlier data mining in a mobile Internet-based large real-time database. The decision-tree outlier-classification feature-based filter algorithm is adopted in this paper. Essentially, decision-tree outlier-classification feature-filtering is a continuous process that moves from high-frequency filtering to low-frequency filtering [28–30]. According to the outlier data-containing information flows in the large real-time database, the outlier data phase features before filtering at the local feature time scale parameter are obtained as follows:

$$\begin{aligned} s(v) &= \int_{0}^{v} \sin\left(\frac{\pi}{2}x^{2}\right) dx, \\ c(v) &= \int_{0}^{v} \cos\left(\frac{\pi}{2}x^{2}\right) dx. \end{aligned} \tag{15}$$

In the above formula, any outlier data signal $x(t)$ in the system can be expressed as

$$x(t) = R\left(a(t)\, e^{i\theta(t)}\right) = a(t)\cos\theta(t). \tag{16}$$

For decision-tree classification feature-filtering, the above formula is adopted as a real signal orthogonal term. The corresponding analytic signal and instantaneous amplitude are obtained as follows:

$$\begin{aligned} v_1 &= \sqrt{BT}\frac{1 + 2\left(f - f_0\right)/B}{\sqrt{2}}, \\ v_2 &= \sqrt{BT}\frac{1 - 2\left(f - f_0\right)/B}{\sqrt{2}}. \end{aligned} \tag{17}$$

In the above formulas, $c(v_1)$ and $s(v_1)$ are orthogonal integrals, while $c(v_2)$ and $s(v_2)$ represent Fresnel integrals of the outlier data signals in the information flows of the large real-time database. Decision-tree filtering is used to remove several high-frequency components before the outlier data and several low-frequency fixed characteristic components after the outlier data. The phrase-frequency feature is obtained as follows:

$$\varphi_l(f) = -\frac{\pi\left(f - f_0\right)^{2}}{k} + \arctan\frac{s(v_1) + s(v_2)}{c(v_1) + c(v_2)}. \tag{18}$$

The introduced decision-tree outlier-classification feature-filtering algorithm is adopted in this study combined with the threshold noise reduction method for wavelet transforms to reduce signal noise. The offset degree is analyzed. First, the feature state response distribution of outlier data contained by the information flows of the large real-time database is analyzed [31, 32]. Then, a threshold is set. A wavelet coefficient larger than the threshold is assumed to be generated by an outlier data signal. Finally, formula (19) is utilized to obtain the detection filtering model for the outlier data:

$$\widehat{W} = \begin{cases} \operatorname{sgn}(W)\left(|W| - \alpha Ts\right) & |W| \geq Ts \\ 0 & |W| < Ts. \end{cases} \tag{19}$$

In addition, formula (20) is used to modify the detection filtering model for the outlier data, which finally achieves outlier data mining in the information flows of a mobile Internet-based large real-time database:

$$\mu_{ij}^{(k+1)} = \left[\sum_{k=1}^{c}\left(\frac{\left\|\sigma_j - \widehat{W}_i^{(k)}\right\|}{\left\|\sigma_j - \widehat{W}_l^{(k)}\right\|}\right)^{2/(\alpha-1)}\right]^{-1}. \tag{20}$$

In the preceding formula, α represents an accommodation coefficient, W is the wavelet decomposition detail coefficient of the natural vibration mode feature of the outlier data in the information flows of the large real-time database, and its value range is $0 \leq \alpha \leq 1$. To improve the confidence coefficient and accuracy of the monitoring model, reduce the false alarm probability, and prevent missed reports, the design requires optimization to improve the probability confidence range of the outlier data mining. Because outlier data in the information flows of the large real-time database show a certain abruptness, their offset degree and average are calculated, the decision-tree classification feature-filtering

Table 1: Parameters used to classify outlier data with the decision-tree in large real-time databases.

Title	Explanation	Default
Binary splits	The binary tree method is used to divide noun attributes.	False
Confidence factor	Prune confidence factors (factors smaller than the given value are pruned from the subtree).	0.25
MinNumObj	The number of instantiation that will be pruned from leaf nodes.	2
NumFolds	This value is used to reduce the error-pruning data flow; the remaining data are used to construct the tree.	3
ReducedErrorPruning	Prune with the error-reduction method.	False
Seed	Prune with error-reduction method and transplant subtree seeds of random data.	1
Unpruned	Determines whether result tree has been pruned.	False

Table 2: Parameters of the simulation experiments.

CPU amount of a single node	4
Single CPU	Core i5 3.11 GHZ
Internal storage of a single node	4 G
Operating system	Windows 7
Hardware	500 G
Switched network	200 M optical network

treatment is applied, and the wavelet inverse transformation is utilized to remove interference generated by noise. The length of outlier data signal is assumed to be k, and σ_j is the standard deviation at the jth layer of noise contained by each intrinsic mode function. According to the above analysis, the decision-tree outlier-classification feature-filter algorithm-based method for outlier data mining is improved. The mining accuracy, mining efficiency, mining stability, and time efficiency of the improved outlier data mining method are further verified through a simulation experiment.

4. Simulation Experiment and Performance Test

4.1. Simulation Experiment. To evaluate the efficiency of the improved method for outlier data mining in a mobile Internet-based large real-time database, a simulation experiment was conducted. The experiment was based on Matlab simulation software. A large real-time database model was constructed. The selected signal model of data under mobile Internet conditions was a group of LFM signal frequency spectrum with a frequency band of 2~10 KHz and a duration of 4 ms. The $\delta(t)$ similarities were 1.60 dB, 3.52 dB, 5.38 dB, and 6.79 dB, respectively. The proposed improved method was adopted for data mining. For the constructed decision-tree outlier-classification feature model, the parameter settings for the modeling process and the simulation experiment are shown in Tables 1 and 2, respectively.

According to the parameters in Tables 1 and 2, the mining model for outlier data in a mobile Internet-based large real-time database was constructed as the basis for the data mining experiment.

In the experiment based on the unimproved method, according to the offset degree analysis of the model for

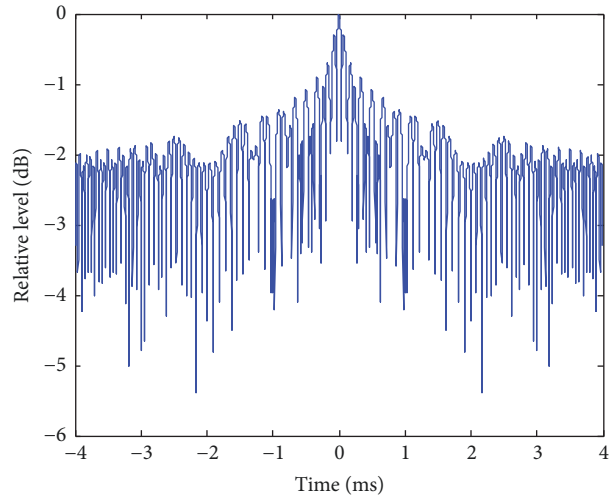


Figure 2: Outlier data signal model obtained using the traditional method.

outlier data in a mobile Internet-based database, a decision-tree outlier-feature classification method was adopted for outlier data mining. The steps of the traditional method are as follows. First, a decision-tree-based information chain transmission model for outlier data in a mobile Internet-based large real-time database is constructed. Then, the decision-tree classification feature algorithm is used to mine the outlier data. Finally, the model for outlier data signals in the large real-time database is shown in Figure 2.

Interference term datasets were shown when mining outlier data, and there was substantial interference noise in the wide-ranging subspaces of the outlier data series, resulting in a low confidence coefficient of the mining algorithm. It was difficult to construct a model for outlier data mining in a mobile Internet-based large real-time database; consequently, the conditions shown in Figure 2 appeared because the outlier data is submerged in the normal data and can barely be mined.

Therefore, after obtaining the transmission model for the outlier data-containing information chains, denoising and dynamic mining of outlier data-contained information flows and clustering of information flow are required for information chains transmitted into the large real-time database. Only after the denoising and dynamic mining processes are

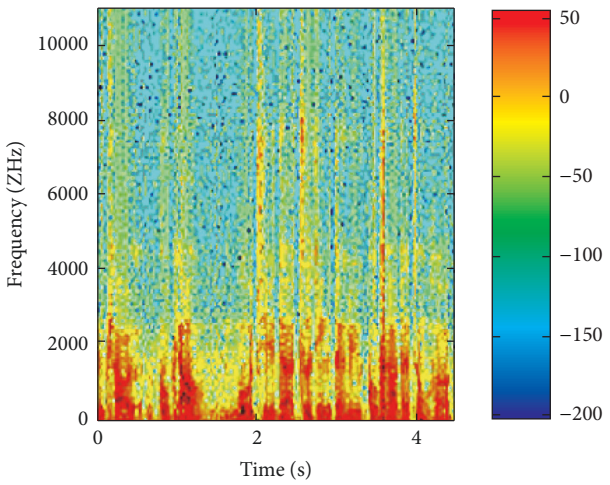


Figure 3: Time domain sequence waveform of the original outlier data information flows.

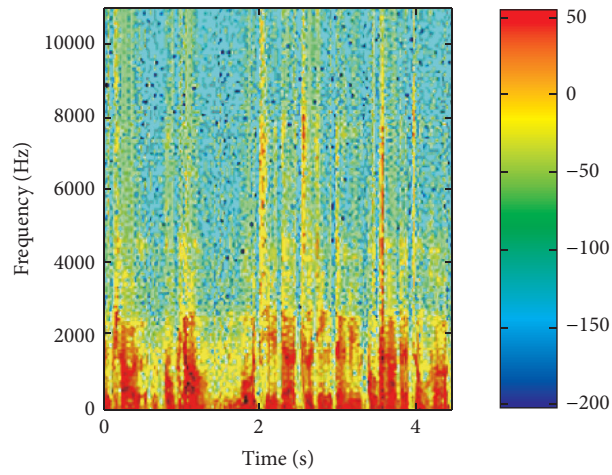


Figure 4: Time domain sequence waveform of outlier data-containing information flows obtained by the proposed method.

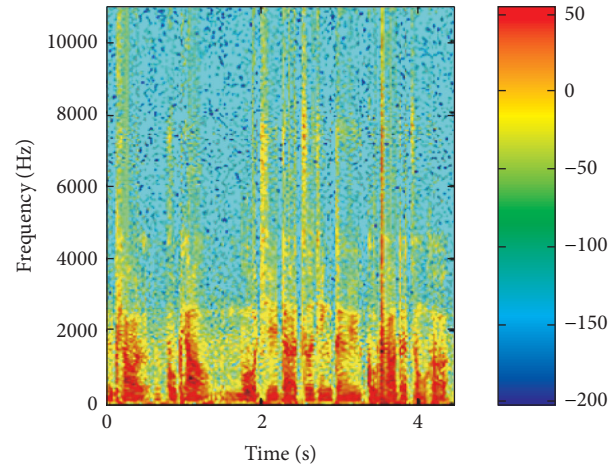


Figure 5: Time Domain sequence waveform of outlier data-containing information flows obtained by the CCPSWNIDA method.

complete, can effective and precise outlier data mining be realized.

Using the proposed method, outlier data-containing information flows in information chains of a large real-time database are obtained first. Then, the chirp-signal amplitude-frequency features are utilized to obtain the global frequency features and to model the data outlier-classification features and analyze its signals. On this basis, the enveloping features formed by outlier data-containing information flows are obtained from the information chains of the large real-time database, and their interference frequencies are acquired. Next, the probability weights are calculated, and after preliminary big-data screening, dynamic mining for the outlier data-containing information flows in the information chains obtains the information content of the time domain waveform for certain frequency bands.

The transmission process of outlier data-contained information flows in information chains in a large real-time database was sampled; and the input and output were set as Data In or Data Out. The sampling frequency for the global outlier data-contained information flows was 12.58 Hz, the sampling interval was 32.4 s, and a total of 1,024 sampling points were output stably [33]. The time domain sequence formed by the outlier data information flows was tested first. The time domain sequence waveform of the original outlier data information flows is shown in Figure 3.

To test the effects of the improved method, the CCPSWNIDA method (Message request is in proportion to message response. With the function of monitoring abnormal conditions, once the percentage of outlier variation is found to be beyond a normal range, the outlier will be alarmed so as to realize the purpose of invasion detection.) (Ni, 2014) [34] was adopted for comparison. The time domain sequence waveform of the outlier data-containing information flows obtained from our method and the CCPSWNIDA method are shown in Figures 4 and 5, respectively.

As shown in Figures 4 and 5, compared with the time domain sequence waveform of the original outlier data information flows, the time domain sequence waveform of the outlier data-containing information flows obtained by the CCPSWNIDA method exhibits high errors and a relatively disordered waveform, while the time domain sequence waveform of outlier data-containing information flows obtained by the method presented in this paper is identical to the time domain sequence waveform of the original outlier data information flows, and it presents a relatively clear waveform. This result indicates that the improved method for acquiring outlier data-containing information flows in the information chains of a mobile Internet-based large real-time database shows certain superiority.

Later, the MOC algorithm was adopted to cluster the outlier data-containing information flows obtained from the presented method, and the clustering extraction result is shown in Figure 6. As Figure 6 shows, the MOC algorithm adaptation effectively reflects the condition features presented by the outlier data-containing information flows, which improves the performance of subsequent outlier data mining.

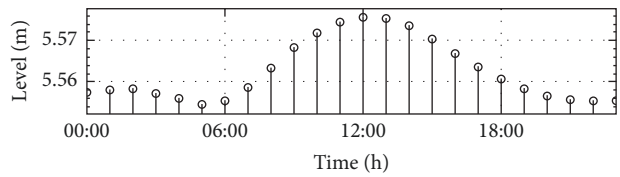


Figure 6: Directional clustering results of outlier data-contained information flows under MOC algorithm.

Finally, outlier data mining was conducted for the outlier data-containing information flows using a top-down pattern. Further experiments were conducted using the method presented in this paper, which was obtained according to the model for outlier data-containing information chain transmission under mobile Internet and the model for acquiring outlier data-containing information flows in the information chains of a large real-time database.

First, the phase position features of the prefiltered outlier data were obtained. Then, corresponding analytical signals and instant amplitude values were obtained according to the decision-tree outlier-classification feature-filtering, and the decision-tree filtering method was adopted to remove several high-frequency components before the outlier data and several low-frequency components after the outlier data. The phase-frequency features were then obtained. On this basis, the decision-tree outlier-classification feature-filtering algorithm was introduced and combined with the WT threshold denoising method to obtain the final model for outlier data detection and filtering. This process confirmed the generation of outlier data signals and obtained the feature state response distribution of the outlier data, as shown in Figure 7. A traditional multithreading dynamic data scheduling method was adopted as a comparison method. The feature state response distribution of outlier data based on the traditional method is shown in Figure 8.

In Figure 8, representing the outlier data state response detected by the traditional method-based multithreading dynamic scheduling information flows, the corresponding features are unclear; the method's performance is relatively poor, and substantial interference that compromises the accuracy of detected outlier data exists. In contrast, the improved method presented in this paper presents good outlier data state response features, good response performance, and less spatial interference, resulting in improved outlier data detection accuracy.

For the outlier data feature states, the times and errors of multithreading dynamic scheduling of information flows were compared between the traditional method and the proposed method. The maximum number of detected samples was set to 500, and the time and errors of outlier data feature state responses were compared between the two methods under different sample amounts. The experimental results are shown in Tables 3 and 4.

Tables 2 and 3 indicate that the average time required by the proposed method to detect outlier data state responses from dynamic scheduled information flows is 2.83 s, shorter than the traditional method, which required 7.86 s. The difference amounts to a reduction of 63.99% compared to

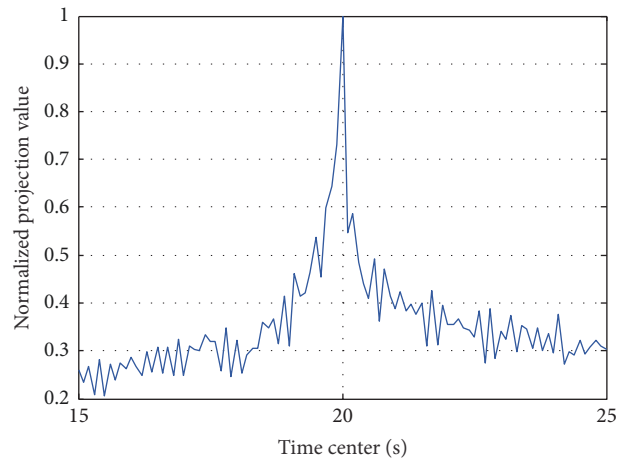


Figure 7: Curve of outlier data feature state response distribution detected with dynamic scheduling of information flows (the proposed method).

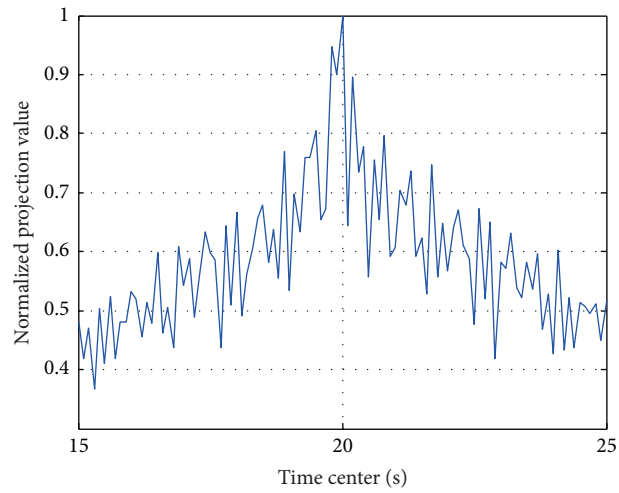


Figure 8: Curve of outlier data feature state response distribution detected with dynamic scheduling of information flows (the traditional method).

the traditional method, suggesting that the proposed method has higher response efficiency. The average error rates of the proposed method and the traditional method are 0.0025 and 0.066, respectively; the improved method reduces the average error rate by 96.2% compared to the traditional method. This result suggests that the proposed method guarantees detection efficiency and reduces detection error of the system, showing the absolute superiority of its detection effect.

Later, the number of iterations and the fitness of the outlier data state response detected by dynamic scheduling information flows were compared between the traditional and the proposed methods. for this comparison, there were 10 trials, and the number of iterations and the fitness were compared between the two methods under different sample amounts. The results are shown in Tables 5 and 6.

An analysis of Tables 5 and 6 reveals that the numbers of iterations of the outlier data feature state responses required by different methods are different. The number of iterations represents the performance of each method to acquire an

Table 3: Times and errors of outlier data feature state responses detected with the proposed method.

Detection sample amount	Proposed method	
	Detection time/s	Error/%
50	2.7	0.003
100	3.0	0.002
150	2.4	0.002
200	2.3	0.003
250	3.4	0.004
300	3.3	0.003
350	2.6	0.002
400	2.7	0.002
450	3.1	0.002
500	2.8	0.002

Table 4: Times and errors of outlier date feature state responses detected with the traditional method.

Detection sample amount	Traditional method	
	Detection time/s	Error/%
50	6.9	0.1
100	6.4	0.08
150	7.4	0.09
200	7.8	0.07
250	8.0	0.06
300	10.2	0.07
350	6.7	0.04
400	7.3	0.06
450	9.7	0.05
500	8.2	0.04

Table 5: Number of iterations and fitness of outlier data feature-state responses detected with the traditional method.

Test amount	Traditional method	
	Iterations	Fitness value
1	72	58
2	83	89
3	91	116
4	61	159
5	99	82
6	73	85
7	88	64
8	83	73
9	77	151
10	84	82
Average	81.1	95.9

optimal solution. The average number of iterations required by the traditional method is 81.1, while the average number of iterations required by the proposed method is 19.4, merely 23.9% of the traditional method. This result indicates that the proposed method detects outlier data rapidly and obtains an optimal solution quickly, showing certain superiority; it also explains why the time required for outlier data dynamic scheduling is shorter with the proposed method than with the traditional multithreading feature-information extraction-based method. With respect to fitness value, the traditional method fitness values range from 58 to 159 (with an average of 95.9), while the proposed method's fitness values range from 62 to 69 (with an average of 65.1), only 67.8% of the traditional method. This result indicates that, viewed from the average perspective, the scheduling sequence fitness of the proposed method is smaller and requires less total time for task execution.

Table 6: Number of iterations and fitness of outlier data feature-state responses detected with the proposed method.

Test amount	Proposed method	
	Iterations	Fitness value
1	26	62
2	17	67
3	19	67
4	18	65
5	16	64
6	21	65
7	19	63
8	17	65
9	17	64
10	24	69
Average	19.4	65.1

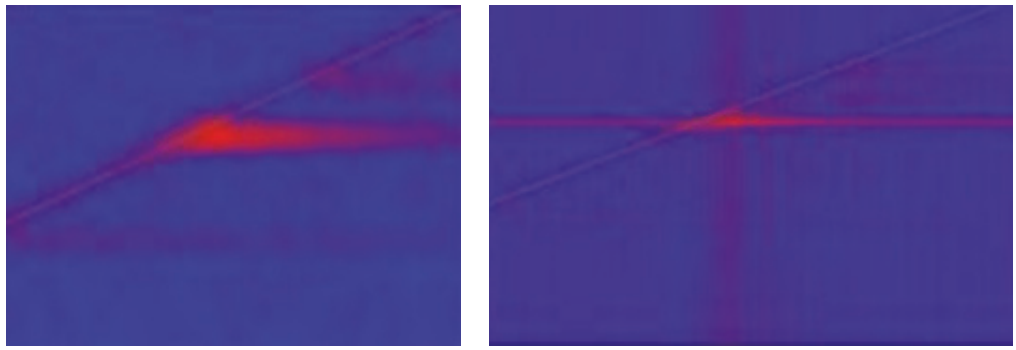
According to the significant features of outlier data feature response detected by the improved method-based dynamic scheduling information flows, an outlier data detection filtering model was obtained for subsequent experiments.

By further correcting the model of outlier data detection filtering, the probability confidence coefficient range was improved for outlier data mining, the offset degree and average of outlier data in an information flow were calculated, and the inverse wavelet transform was utilized to remove noise interference to finally realize outlier data mining in the information flow. To compare outcomes, the outlier data confidence coefficient and offset degree were adopted as the measurement indexes. The simulation result of the outlier data cloud picture was obtained, as shown in Figure 9. As the figure shows, the proposed method effectively improves the precise coverage and probability of outlier data mining and shows promise for applications.

The improved method presented in the paper was extended to achieve precise mining of outlier data under one-time transmission in outlier data-contained information flows in information chains of a mobile Internet-based large real-time database.

The results of the outlier data mining are shown in Figure 10.

The waveform in Figure 10 shows that the outlier data mining based on the improved method removes interdata noise and shows relatively high impact response features, full information content, and superior performances, indicating



(a) Simulation of outlier data mining based on the proposed method

(b) Simulation of outlier data mining based on the traditional method

Figure 9: Cloud picture comparison of outlier data mining in information flows of a large real-time database.

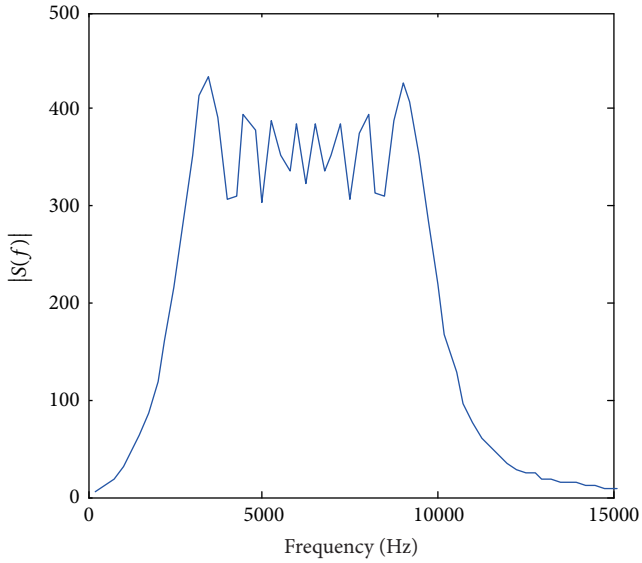


Figure 10: Results of outlier data mining in a mobile internet-based large real-time database.

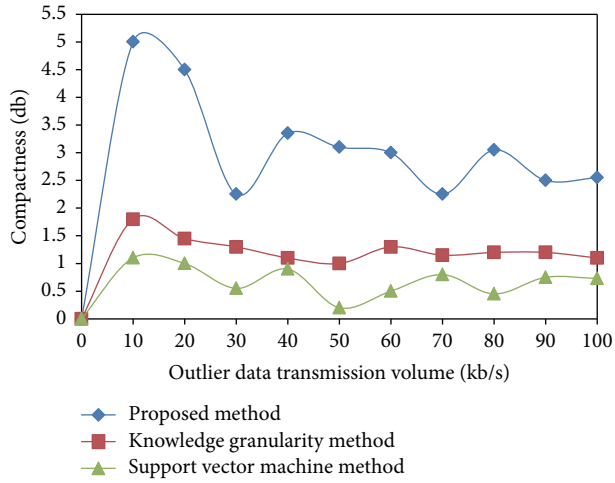


Figure 11: Comparison of outlier data mining compactness.

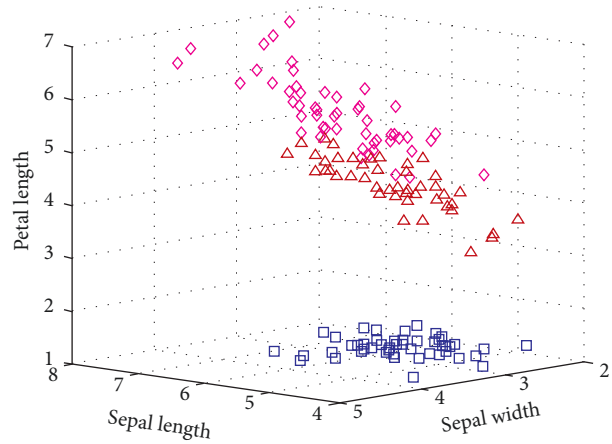


Figure 12: Clustering results of outlier data.

that the method performed well when applied to outlier data mining in a mobile Internet-based large real-time database.

4.2. Performance Test. To verify the effectiveness of the outlier data mining method presented in this paper, compactness and clustering results of outlier data mining were adopted as the basic performance evaluation indexes of outlier data mining, and the improved method was compared with the traditional knowledge granularity method and support vector machine (SVM) method. Figure 11 shows the compactness analysis results. Then, the outlier data mining results were compared among the different methods. Figure 12 shows the clustering analysis results.

Figure 11 shows that under the same outlier data transmission volume, outlier data mining in a mobile Internet-based large real-time database using the proposed method shows mining compactness far higher than the compactness achieved by the traditional knowledge granularity and support vector machine methods. As the outlier data transmission volume increases, the mining compactness of the proposed method remains higher than that of the traditional methods, suggesting that the proposed method has a substantial advantage. In Figure 11, the average compactness of the proposed method is approximately 3.155 dB, while the average compactness values of the traditional knowledge granularity and support vector machine methods are 1.26 dB and

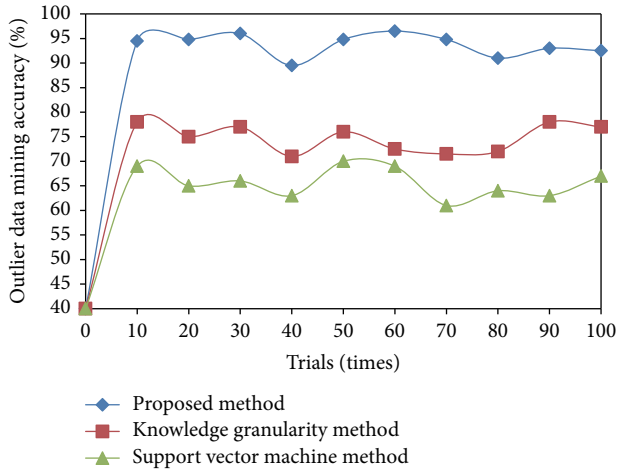


Figure 13: Comparison of outlier data mining accuracy.

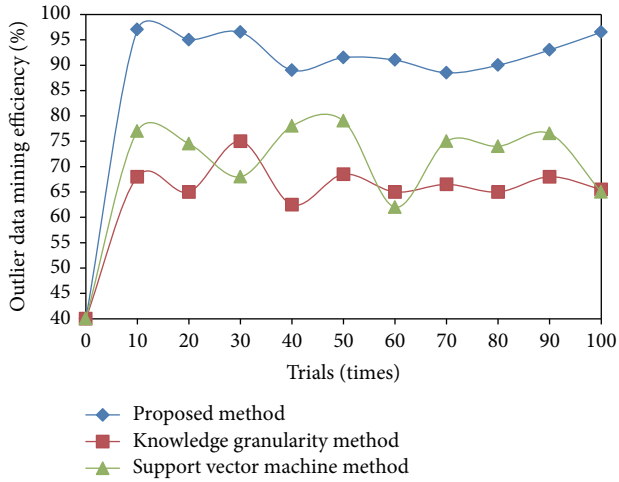


Figure 14: Comparison of outlier data mining efficiency.

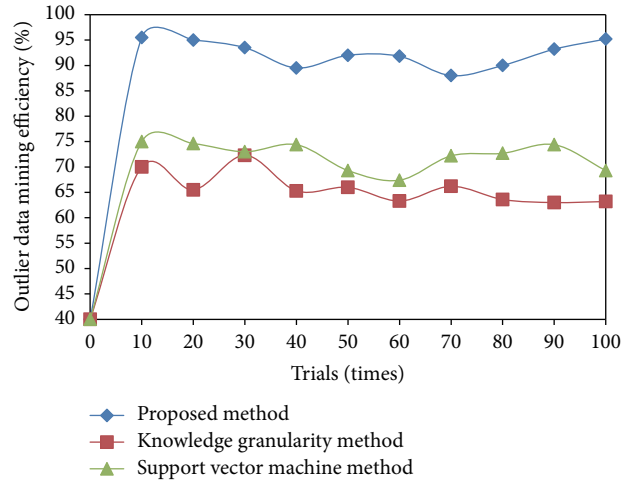


Figure 15: Comparison of outlier data mining stability.

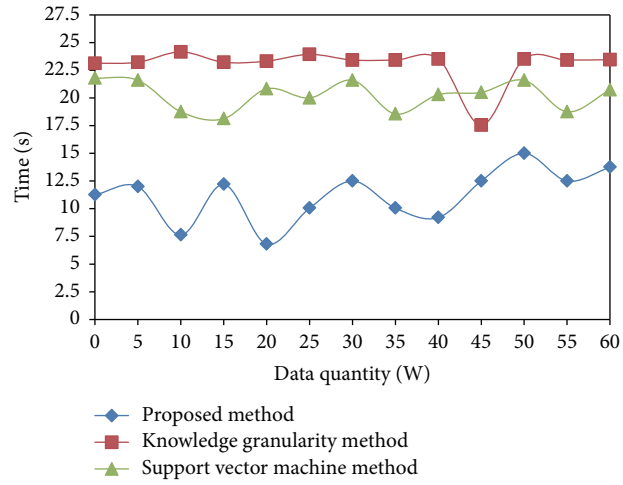


Figure 16: Comparison of outlier data mining time.

0.698 dB, respectively. The proposed method shows increases of 150% and 352% over the two traditional methods.

In Figure 12, according to different measurement scales of data, the pink, the red, and the blue represent different outlier data of different classifications, orders, and values. When using the proposed method for mining of outlier data in a mobile Internet-based large real-time database, the three data types are well-differentiated and clustered. Although some pink outlier data is mixed in with the red outlier data, the amount is low and acceptable. Therefore, the above results indicate that mining of outlier data in a mobile Internet-based large real-time database with the proposed method achieves good outlier data clustering outcomes.

To further verify the high efficiency of the proposed outlier data mining method, outlier data mining accuracy, mining efficiency, mining result stability, and required mining time were adopted as indexes to evaluate the global performances of outlier data mining methods. The traditional knowledge granularity and support vector machine methods were adopted for comparisons. The results are shown in Figures 13–16.

As shown in Figure 13, under the same number of trials and with uncertain data volumes, the outlier data mining in a mobile Internet-based large real-time database using the proposed method achieves a higher mining accuracy than do the traditional knowledge granularity and support vector machine methods. As the number of trials increases, the advantage of the proposed method becomes more obvious. Viewed from a global level, when using the proposed method for outlier data mining, the average mining accuracy is approximately 97.74%, while the average mining accuracies of the traditional knowledge granularity and support vector machine methods are approximately 78.8% and 69.7%, respectively. The proposed method achieved accuracy improvements of 24.03% and 40.22% over the compared methods, respectively. This result demonstrates that the proposed method presented in the paper is absolutely dominant in outlier data mining accuracy.

It can be viewed by analyzing Figure 14 that, under the condition of the same experiment amount and uncertain data amount, the outliner data mining in a mobile Internet-based large real-time database with the method presented

in the paper shows higher mining efficiency than the traditional knowledge granularity method and the support vector machine method. With the increase of experiment amount, the advantage of method presented in the paper is more obvious. Viewed from a global level, when using the method presented in the paper for outlier data mining, the average mining efficiency is about 96.8%, while the average mining efficiency of the traditional knowledge granularity method and that of the support vector machine method are about 70.9% and 76.9%. The method presented in the paper shows improvement of 36.53% and 25.87% over the two methods. It is demonstrated that the method presented in the paper shows an absolute predominance for outlier data mining efficiency.

As shown in Figure 15, under the same number of trials and with uncertain data volumes, the outlier data mining in a mobile Internet-based large real-time database using the proposed method shows higher mining stability than the traditional knowledge granularity method and the support vector machine method. As the number of trials increases, the advantage of the proposed method becomes more obvious. Viewed from a global level, when using the proposed method for outlier data mining, the average mining stability is approximately 96.37%, while the average mining stability of the traditional knowledge granularity method and that of the support vector machine method are approximately 69.84% and 76.23%, respectively. The proposed method shows a stability improvement of 37.98% and 26.42% over the compared methods, respectively. This result demonstrates that the proposed method is predominant for outlier data mining stability.

As shown in Figure 16, with uncertain data volumes, the outlier data mining in a mobile Internet-based large real-time database using the proposed method results in a shorter mining time than the time required by the traditional knowledge granularity and support vector machine methods. As the amount of data increases, the advantage of the proposed method becomes more obvious. Viewed from a global level, when using the proposed method for outlier data mining, although there is relatively high fluctuation, the average mining time for the improved method-based outlier data mining is approximately 15.84 s, while the average mining times of the traditional knowledge granularity and support vector machine methods are approximately 31.23 s and 27.61 s, respectively. The proposed method improves the mining time by 15.39 s and 11.77 s over the compared methods, respectively. This result demonstrates that the proposed method is predominant for outlier data mining time.

5. Conclusion

Outlier data mining in a mobile Internet-based large real-time database using traditional methods is prone to generating mass interference data and reducing the accuracy, efficiency, and stability of data mining, all of which are severe deficiencies. To address these issues, this paper presents a decision-tree outlier-classification feature-filtering detection-based method for outlier data mining in a mobile Internet-based large real-time database. The method is used to analyze features of real-time data, obtain magnitude spectra models of outlier data, and conduct preliminary analysis on signals of outlier data. A decisional-tree information chain transmission model for outlier data existing under mobile Internet conditions is established to precisely cluster mobile and scattered data, obtain outlier data-containing information chains, and transmit them into the real-time database in an orderly manner. Outlier data-containing information flows are obtained from the information chains in the large real-time database and then clustered to improve the precision of subsequent mining, significantly reduce the complexity of mining computation, and lay a foundation for the subsequent division of the confidence interval. Consequently, the mining method is subsequently improved by acquiring the phase position features of prefiltered outlier data in outlier data-containing information flows. The decision-tree outlier-classification feature-filtering algorithm is adopted to obtain analytical signals and instant amplitude, and the decision-tree filtering method is used to remove useless high- and low-frequency components and to obtain the phase-frequency features of the outlier data. The WT threshold denoising method is integrated to perform signal denoising. Then, the data offset degree is analyzed. The results of simulation experiments indicate that the method significantly reduces mining response time, the number of required iterations, and the error rate and improves the fitness of precise mining. The formed detection filtering model is corrected and the simulation experiments indicate that the precise coverage and probability of outlier data mining are effectively improved. This approach achieved outlier data mining in a mobile Internet-based large real-time database and yielded favorable outlier data mining effects. With respect to performance, the proposed method was compared with the traditional knowledge granularity method and the support vector machine method regarding compactness, accuracy, efficiency, stability, and time of outlier data mining as well as clustering. The results showed the absolute superiority of the proposed method for outlier data mining.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research described in this paper was substantially supported by Grants from the National Natural Science Foundation of China (no. 71471178) and the State Key Program of National Natural Science Foundation of China (no. 71431006) and Projects of International Cooperation and Exchanges NSFC (no. 71210003) and the National Innovation Research Group Science Foundation for China (no. 70921001).

References

[1] Y. Latif, C. Cadena, and J. Neira, "Robust loop closing over time for pose graph SLAM," *International Journal of Robotics Research*, vol. 32, no. 14, pp. 1611–1626, 2013.

[2] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS Quarterly: Management Information Systems*, vol. 36, no. 4, pp. 1165–1188, 2012.

[3] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[4] H. Zheng, Y. Wang, Z. Xiong, L. I. Kunming, Z. Chong, and F. Yin, "Parallel mining on label-constraint proximity pattern," *Computer Engineering Applications*, vol. 51, no. 9, pp. 135–141, 2015.

[5] Y.-Z. Zhang, J. Xiao, X.-C. Yun, and F.-Y. Wang, "DDoS attacks detection and control mechanisms," *Journal of Software*, vol. 23, no. 8, pp. 2058–2072, 2012.

[6] W. U. Chun-Qiong, "Network intrusion detection model based on feature selection," *Computer Simulation*, vol. 29, no. 6, pp. 234–237, 2012 (Chinese).

[7] O. N. Almasi and M. Rouhani, "Fast and de-noise support vector machine training method based on fuzzy clustering method for large real world datasets," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 24, no. 1, pp. 219–233, 2016.

[8] Y. Zeng, Z. Zhang, and A. Kusiak, "Predictive modeling and optimization of a multi-zone HVAC system with data mining and firefly algorithms," *Energy*, vol. 86, pp. 393–402, 2015.

[9] Z. Zhang, J. Liu, and Y. Xu, "Rapid computation of structural static extreme response based on reduced basis method," *Chinese Journal of Computational Mechanics*, vol. 32, no. 1, pp. 94–98, 2015.

[10] A. Purarjomandlangrudi, A. H. Ghapanchi, and M. Esmalifalak, "A data mining approach for fault diagnosis: an application of anomaly detection algorithm," *Measurement*, vol. 55, no. 3, pp. 343–352, 2014.

[11] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1690–1700, 2014.

[12] M. A. A. Mamun, M. A. Hannan, A. Hussain, and H. Basri, "Theoretical model and implementation of a real time intelligent bin status monitoring system using rule based decision algorithms," *Expert Systems with Applications*, vol. 48, pp. 76–88, 2016.

[13] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining," *Automation in Construction*, vol. 74, no. 7, pp. 39–54, 2017.

[14] L. Tran, L. Fan, and C. Shahabi, "Distance-based outlier detection in data streams," *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 1089–1100, 2016.

[15] B. S. M. Hosseini, B. Amiri, M. Mirzabagheri, and Y. Shi, "A new intrusion detection approach using pso based multiple criteria linear programming," *Procedia Computer Science*, vol. 55, no. 4, pp. 231–237, 2015.

[16] J. Liu and H. F. Deng, "Outlier detection on uncertain data based on local information," *Knowledge-Based Systems*, vol. 51, no. 1, pp. 60–71, 2013.

[17] N. N. R. Ranga Suri, M. Narasimha Murty, and G. Athithan, "Detecting outliers in categorical data through rough clustering," *Natural Computing*, vol. 15, no. 3, pp. 385–394, 2016.

[18] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[19] R. B. Torbert, C. T. Russell, W. Magnes et al., *The FIELDS instrument suite on MMS: Scientific objectives, measurements, and data products*, vol. 199, Space Science Reviews, 2016.

[20] E. M. Schwartz, E. T. Bradlow, and P. S. Fader, "Model selection using database characteristics: Developing a classification tree for longitudinal incidence data," *Marketing Science*, vol. 33, no. 2, pp. 188–205, 2014.

[21] N. A. Khan and B. Boashash, "Multi-component instantaneous frequency estimation using locally adaptive directional time frequency distributions," *International Journal of Adaptive Control and Signal Processing*, vol. 30, no. 3, pp. 429–442, 2015.

[22] N. Ali Khan, S. Ali, and M. Jansson, "Direction of arrival estimation using adaptive directional time-frequency distributions," *Multidimensional Systems and Signal Processing*, vol. 45, no. 6, pp. 1–19, 2016.

[23] M. A. B. Othman, J. Belz, and B. Farhang-Boroujeny, "Performance analysis of matched filter bank for detection of linear frequency modulated chirp signals," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 1, pp. 41–54, 2017.

[24] W. Dai and W. Ji, "A mapreduce implementation of C4. 5 decision tree algorithm," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 49–60, 2014.

[25] S. Stimpson, B. Collins, and B. Kochunas, "Improvement of transport-corrected scattering stability and performance using a Jacobi inscatter algorithm for 2D-MOC," *Annals of Nuclear Energy*, vol. 105, no. 6, pp. 1–10, 2017.

[26] Z. Zhang, H. Fang, and H. Wang, "A new mi-based visualization aided validation index for mining big longitudinal web trial data," *IEEE Access*, vol. 4, no. 5, pp. 2272–2280, 2016.

[27] K. Zhou and S. Yang, "Exploring the uniform effect of fcm clustering: a data distribution perspective," *Knowledge-Based Systems*, vol. 96, no. 3, pp. 76–83, 2016.

[28] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: methods, systems and tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[29] M. A. Ambusaidi, X. He, and P. Nanda, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 2986–2998, 2016.

[30] R. Latif, H. Abbas, S. Latif, and A. Masood, "EVFDT: an enhanced very fast decision tree algorithm for detecting distributed denial of service attack in cloud-assisted wireless body area network," *Mobile Information Systems*, vol. 2015, Article ID 260594, 13 pages, 2015.

[31] S. Patidar, R. B. Pachori, and N. Garg, "Automatic diagnosis of septal defects based on tunable-Q wavelet transform of cardiac sound signals," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3315–3326, 2015.

[32] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," *Knowledge and Information Systems*, vol. 34, no. 1, pp. 23–54, 2013.

[33] R. Laxhammar and G. Falkman, "Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories," *Annals of Mathematics and Artificial Intelligence*, vol. 74, no. 1-2, pp. 67–94, 2015.

[34] H.-X. Ni, "Based sliding window cloud computing platform of network intrusion detection algorithm," in *Proceedings of the 14th IEEE International Conference on Computer and Information Technology, CIT 2014*, pp. 927–930, China, September 2014.