IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 29, 2018, accepted October 13, 2018, date of publication October 22, 2018,
date of current version November 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877423

An Anomaly Recognition and Autonomic Optimization Method to User's Sequence Behaviors for D2D Communications in MCC

**RUIJUAN ZHENG, JUNLONG ZHU, MINGCHUAN ZHANG, QINGTAO WU,
RUOSHUI LIU, (Member, IEEE), KANG LIU, AND JING CHEN**
College of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China

Corresponding authors: Ruijuan Zheng (zhengruijuan@haust.edu.cn) and Qingtao Wu (wqt8921@haust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602155, Grant U1604155, Grant 61871430, and Grant 61370221, in part by the Henan Science and Technology Innovation Project under Grant 174100510010, and in part by the Industry University Research Project of Henan Province under Grant 172107000005.

ABSTRACT Mobile cloud computing uses cloud computing to deliver applications to mobile devices. These applications can be delivered among different devices with different operating systems, computing tasks, and data storage capabilities, adopting D2D communication mode. However, because the application delivery process covers three entities, namely user-environmental-service, trusted problems are ubiquitous. Therefore, before cloud provides substantive services, how to identify the trusted degree of user identity and its behaviors for D2D communications is the core problem. First, from the perspective of user trustability, this paper proposed an analysis method to user abnormal behaviors for D2D communications in mobile cloud environment. In this method, user behavior is normalized to a "user sequence operation" identity fragment with the same length, offset, and amplitude. The hierarchical matching method and the blacklist mechanism are used to determine whether the user behavior for D2D communications is beyond the scope of trusted tolerance. Second, considering that the user sequence behavior step is a complex graph structure with continuous dynamic growth, this paper proposed a pattern growth method based on maximum and right-most path extension for autonomous optimization. At last, the experimental results showed that the classification accuracy under the KDD CUP99 data set and real network environment was 94.8% and 90.2%, respectively, which was 5.3% and 6.9% higher than the traditional methods. In addition, it can be seen from the experimental results that this scheme could significantly improve the recognition speed.

INDEX TERMS MCC, D2D communications, user's sequence behaviors, on-line identification.

I. INTRODUCTION

Recently, the exploitation of cloud resources for augmenting mobile devices leads to the emergence of a new research area called MCC [1], [2]. As a product of the combination of mobile Internet, cloud computing and D2D communications, MCC has developed rapidly in recent years and spawned numerous new information services and application modes, which have attracted extensive attention from the academic and industrial circles. According to the working principle of multi-layer collaboration of its "User-Environment-Service" for D2D communications, MCC has a high degree of complexity. The system complexity of mobile cloud service access device (User) is far beyond the traditional terminal equipment. Its storage capacity, computing capacity and battery capacity are limited, and the acquired resources are not rich enough and easier to be lost. The hosted network is converted from Internet to mobile overlay Internet, so its network connection (Environment) has become more complicated. The complexity of mobile cloud service platform (Service), the diversity of resources and the characteristics of multi-terminal sharing make the delivery, storage, access, update and destruction of cloud services for D2D communications more likely to face security and reliability problems. As the demand of mobile terminal users is dynamic and has a strong personal color, the trust of each member and link for D2D communications has become a barrier to the development and application of mobile cloud services.

In trusted mobile cloud services architecture, the issue of user trustability in mobile cloud services [3] is equivalent to trustability of user identity and the time-sequence characteristics of user operations. That is, through the comparison of users' behavior sequence, the conclusion of users' behavior sequence for D2D communications is obtained: normal or abnormal. When analyzing users' abnormal behavior, whether it is the mining of users' normal sequential behavior [4] or the real-time discrimination of users' sequential behavior for D2D communications, at the beginning of the set construction, it is impossible to avoid the situation that the mining or judgment result caused by incomplete operation behavior is somewhat different from the actual situation, thus resulting in ''hard division''. In addition, with the deepening of cloud service access and user/server interaction, the branching and interference of ''user sequence operation'' identity fragments for D2D communications will increase. Therefore, it is necessary to research the method of eliminating the interference factors of dynamic ''user sequence operation'' identity fragments in the dimensions of user identity identification [5], sequence feature determination and operational feature comparison, so as to accurately match the sequence behavior graph with the normal sequence behavior set.

Users for D2D communications have a certain randomness and suddenness for various cloud service requests, but in the long run, the request operation behavior of specific users for cloud service basically tends to be stable, which brings the possibility for the accurate mining of normal behavior pattern. There are two key problems in the process of identifying users' abnormal behaviors for D2D communications: How to obtain an ordered set of ''user sequence operation'' before the trustability determination and recognition of user identity and its behavior? And how to complete the real-time discrimination and autonomous optimization of the ordered set of ''user sequence operation'' so as to realize the recognition and control of the operation and action behaviors implemented by users? In view of the above problems, this paper proposes an anomaly recognition and autonomic optimization method for user's sequence behaviors for D2D communications in mobile cloud environment, which makes following contributions. The main contributions are as follows:

- The user behavior ordered tree for D2D communications is constructed by the normal behavior pattern mining method, and the most recently unused replacement method is adopted to save the storage space and improve the recognition speed.
- The preprocessed sample data is compared with the normal sequence behavior database of users for D2D communications, and the hierarchical matching method is used to judge whether the user behavior for D2D communications is beyond the range of trusted tolerance, which improves the detection accuracy. In addition, the blacklist mechanism is adopted to motivate abnormal users to obtain services, and the blacklist is set at three levels: level 1, level 2 and level 3. The second-chance mechanism is implemented after the behaviors beyond the scope of trusted tolerance are added to the blacklist.
- By using the method of autonomic optimization of users' sequential behavior for D2D communications, a complete space for the growth of user behavior pattern is constructed, and less users' sequence behavior steps are used to realize the early determination of users' sequence behavior.
- The experimental results verify that our method is suitable for identifying abnormal behaviors of mobile devices for D2D communications in the mobile cloud environment, laying a foundation for the mobile cloud platform to provide efficient and personalized services. The experimental results based on simulation data and real mobile cloud service environment show that the method has a better effect on the evaluation indexes such as detection speed, detection rate, missing rate, time complexity and so on to some extent, which verifies the feasibility and effectiveness of the method.

The rest of this paper is organized as follows: Section II describes the research status of user behavior recognition for D2D communications. Section III presents the method of user behavior patterns mining and on-line identification, and expounds principle of autonomous optimization of user sequence behaviors based on pattern growth. Section IV provides the experiment, which includes the experiment environment and results analysis. Section V makes a conclusion and presents some future works.

II. RELATED WORK

The evolving MCC paradigm enables mobile users to offload their computing tasks to cloud servers [6]. In recent years, the research on user behavior identification has gained a lot of attention [7]–[10]. The most common method is calculating the distance between sample properties. Literature [11] used Euclidean distance to measure the similarity of attributes, which could measure not only 2-dimensional linear space, but also d-dimensional linear space. Literature [12] used the lexical similarity k-means algorithm based on fuzzy logic Euclidean distance. Literature [13] re-expressed the Euclidean distance according to datas probability density function, and got a probabilistic metric to compute the dissimilarity between two uncertain series.

Some references used model-based methods, such as literature [14] designed a multi-level network detection method based on Support Vector Machine (SVM) model. Literature [15] introduced Bayesian model for unsupervised learning with multiple types of data. Literature [16] used novel unsupervised method for abnormal behavior detection, which adopted Multi-scale Markov random field and considered both local and global contextual information. Literature [14] selected data mining technology to automatically extract normal pattern and abnormal ones from a large set of network data and distinguish them from each other. Combining with the abnormal recognition and misuse recognition, [18] proposed a collaborative analysis method of user abnormal behavior based on reputation voting. Reference [19] proposed an innovative and efficient framework, called KCE, to dynamically detect network structure and manage communication resources, leveraging the insightful knowledge obtained from D2D communications among mobile users.

How to design a low-latency and accurate approach for user behavior anomaly detection over data streams has become a great challenge [20]. Reference [21] proposes a dynamic trust relationship aware data privacy protection (DTRPP) mechanism for mobile crowd-sensing, which can protects the data privacy effectively and has better performance on the average delay. A new anomaly detection model based on Principal Component Analysis (PCA) is proposed in [22]. Reference [23] applied PCA subspace anomaly detection method for the detection of anomalous behavior instances in a web server. Reference [24] introduced a behavior analysis method that learns its context and detects multiple types of insider threats from raw logs and network traffic in real-time.

Outlier detection is a very important concept in abnormal analysis. Literature [25] proposed a pattern-based outlier detection method to identify abnormal attributes in software project data, after discovering the reliable frequent patterns that reflect typical characteristics of the software project data, outliers and their abnormal attributes are detected by matching the software project data with those patterns. Literature [26] proposed a new outlier detection method inspired by spectral clustering, and combined with k-Nearest Neighbor (kNN) and spectral clustering techniques to obtain the abnormal data as outliers by using the information of eigenvalues and eigenvectors statistically in the feature space, which was well scalable to modern large datasets. Literature [27] used cross decision support, knowledge discovery and data mining to allow the performance of outlier detection tasks with an almost-linear complexity. A hybrid technique for user activities outliers detection is introduced in [28], consisting of a two-stage integration of principal component analysis and fuzzy rule-based systems. Reference [29] present a novel approach to the security of the corporate mobile clients, in particular when they operate in the offline mode, in which, the protection of the sensitive data is provided by the combination of cryptographic means and analytics methods to detect malicious user behavior.

The above results have laid a good theoretical and methodological foundation for related research. In the future mobile cloud environment, these literatures play an important role in the process of mobile cloud users communicating with the mobile cloud and completing a series of operations. Nevertheless, there is a lack of research on user abnormal behaviors for D2D communications under MCC network environment. Our research aims to identify the abnormal sequence behavior of mobile terminal users for D2D communications, and predict the service steps, service effects, and diffusion influencing scope. These are the main differences between our work and previous research.

III. USER BEHAVIOR PATTERNS MINING AND ON-LINE IDENTIFICATION METHOD FOR D2D COMMUNICATIONS

A. USER BEHAVIOR PATTERNS MINING AND ON-LINE IDENTIFICATION MODEL

Definition 1: User Sequence Operation. It can define and specialize user's sequential behavior for D2D communications, which is the basis of abnormal user behavior identification. The sequence behavior steps of user U_i are expressed as $S_{ui} =< UID|S_1 \ldots S_n|OP_1 \ldots OP_m >$, where UID is the user identity, S_i is the the i-th sequence, and OP_j expresses the j-th operation step.

Definition 2: Uniqueness of user sequence behavior for D2D communications. A user sequence behavior for D2D communications with n length is an ordered collection, consisting of record value, depth, label and record time, denoted as $X = (< X_1, d(X_1), l_1, t_1 >, < X_2, d(X_2), l_2, t_2 >, \ldots, < X_n, d(X_n), l_n, t_n >)$. In general, the sampling interval time of user sequence $\Delta t = t_i - t_{i-1}$ is equal, which could be viewed as $t_1 = 0$, $\Delta t = 1$. Therefore, the sequence behavior can be denoted as $X = (< X_1, d(X_1), l_1 >, < X_2, d(X_2), l_2 >, \ldots, < X_n, d(X_n), l_n >)$.

The model of user behavior patterns mining and on-line identification for D2D communications is described as follows. Firstly, the formalized definition and description method of "user sequence operation" identification is researched. By taking advantage of the dependency relationship and timing characteristics between the nodes of the normal sequence behavior, the incremental mining of the sequence behavior pattern is realized, and a complete set of the normal user behavior pattern for D2D communications is constructed. Secondly, comparing normal user behavior database with the preprocessed data samples for D2D communications, hierarchical matching method is used to judge whether user behaviors are beyond the trusted tolerance scope, and then identified behaviors are made corresponding rewards and punishments according by blacklist technology with the second chance mechanism. Secondly, the database of users' normal sequence behavior for D2D communications is compared with the preprocessed data samples, and the method based on hierarchical matching is adopted to judge whether the user behavior is beyond the scope of trusted tolerance. The identified user behavior is rewarded or punished according to the blacklist technology of the second-chance mechanism. Finally, based on the extensible feature of user sequence behaviors, through the analysis of the function flow and data flow of nodes, the model growth method of maximum and right-most path extension is adopted to research service steps and service effects that may radiate from abnormal time series behaviors, and feedback is formed on the mining of users' normal sequence behaviors and the real-time identification of users' sequence behaviors for D2D communications, so as to optimize the mining and identification effects. Based on the mean value and threshold of abnormal behaviors, the nearest neighbor algorithm is used to classify the identified abnormal behaviors.

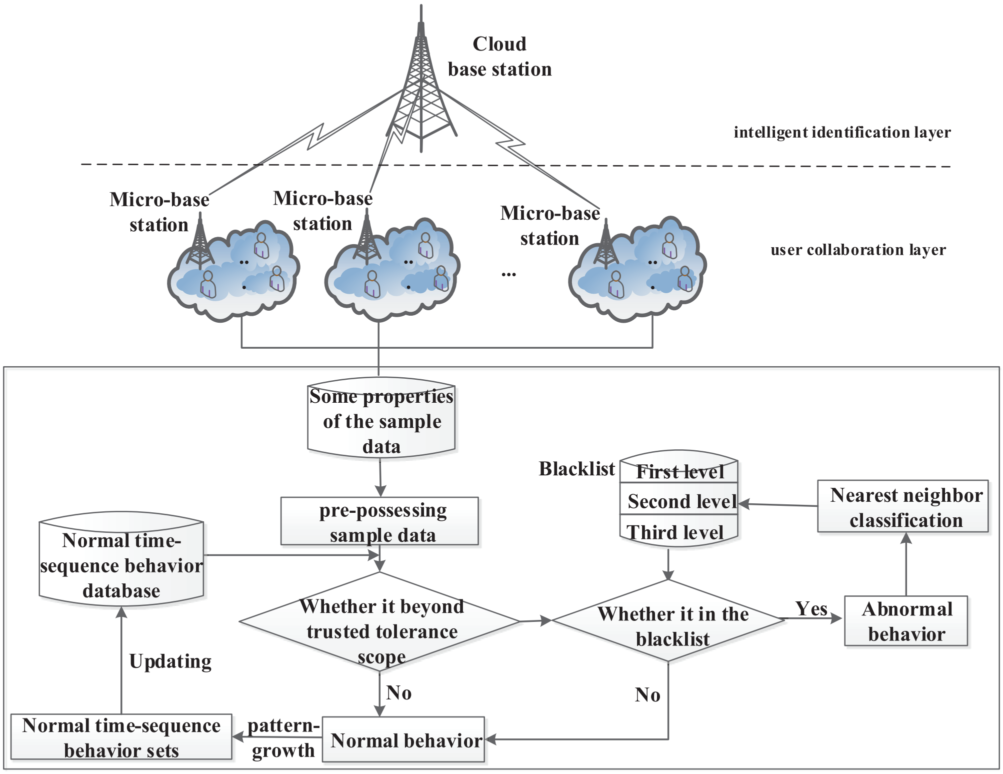


FIGURE 1. User behavior patterns mining and on-line identification model for D2D communications.

The user behavior pattern mining and online identification model for D2D communications is shown in Figure 1. From the perspective of user trustability, the mobile cloud environment is divided into user collaboration layer and intelligent identification layer, and the base stations of corresponding layers are micro-base station and cloud base station respectively. The micro-base station is to process the user behavior information and carry out computing identification at the user collaboration layer, while the cloud base station is to conduct unified computing processing of the user behavior information transmitted to the cloud computing center. The cloud base station can realize resource sharing and dynamic allocation, achieve the goal of low cost and high broadband with collaborative and virtual technologies, and provide efficient and personalized cloud services for legitimate users in real time.

This paper mainly researches from three aspects for D2D communications, which are the normal sequence behavior pattern mining of mobile cloud users, the real-time identification method and the autonomic optimization method. The details are as follows.

B. NORMAL BEHAVIOR MODEL MINING METHOD WITH TIME-INTERVAL CONSTRAINTS

Since it is meaningless to compare users' sequence behaviors with different offsets and amplitudes [30], it is necessary to preprocess users' behavioral data before mining, that is, to normalize users' sequence behaviors [31]. The number of mobile cloud users for D2D communications is very large and approximately satisfies the normal distribution. The mean and variance are $\mu(x)$ and $\rho(x)$ respectively. X is standardized $X = (< X'_1, d(X'_1), l'_1 >, < X'_2, d(X'_2), l'_2 >, \ldots, < X'_n, d(X'_n), l'_n >)$, where $x'_i = (x_i - \mu(x))/\rho(x)$.

The standardized user behavior is extracted and an ordered tree is constructed. The data is extracted from log file and grouped by UID, and the time constraint threshold is set as δ. According to equation (1), if user i is related to user j, the data is pruned or the behavior data within the time period is merged; otherwise, it should build a new ordered tree.

$$\begin{cases} |t_i - t_j| < \delta, & \mathit{others} \\ |t_i - t_j| > \delta, & a_i \mathit{\ and\ } a_j \mathit{\ are\ related} \end{cases} \tag{1}$$

For example, a related behavior data extracted from the Kieker website log is as follows: $1;1275046487445543183;
0;cn.com.jdlssoft.etax.web.filter.QxkzFilter.isIgnoreURL
(java.lang.String);NULL;655836695735828483;
12750464874 45118229;1275046487445536923;SPJ02;3;2.

The description of each parameter in monitoring records is shown in Table 1. The *operation* in Table 1 includes package name, class name, method name and parameter type of the method. The build of the call dependency graph needs to first build the run path to get the method call relationship.

The run path is the execution of a series of methods in the system caused by a user request for D2D communications,

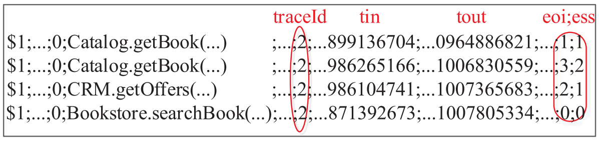
TABLE 1. Description of monitoring parameters.

Parameters	Description
autoId	Metadata records/Operation execution records
experimentId	Experiment identification
operation	Method name with parameters
sessionId	Distinguish different user requests
traceId	Identify running paths
tin Start	execution time of the method
tout	End execution time of the method
vmname	Default is host name
eoi	Execution order index
ess	The size of execution stack

namely the collection of method calls and returns. In order to distinguish different run paths, **traceId** is adopted as the run path identifier in the monitoring data. All monitoring records for the same run path have the same **traceId**. Based on the **traceId** in the monitoring record, all records with the same **traceId** can be extracted from the monitoring data. Where, the value of **eoi** always increases with the method call; The value of the **ess** increases or decreases with the method call, reflecting the depth of the method call.

In Bookstore class of Kieker, the run paths of **traceId**, **eoi** and **ess** are built as follows.

1) In monitoring data, the same running path records have the same traceId, therefore, we can extract all records of the same path according to traceId. The processed data related to running path are shown in Figure 2.

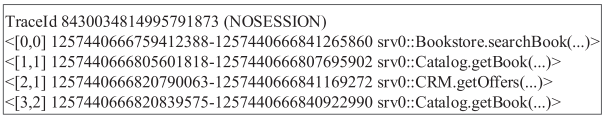


```
                                   traceId    tin          tout        eoi;ess
$1;...;0;Catalog.getBook(...)        ;...;2;...899136704;...0964886821;...;1;1
$1;...;0;Catalog.getBook(...)        ;...;2;...986265166;...1006830559;...;3;2
$1;...;0;CRM.getOffers(...)          ;...;2;...986104741;...1007365683;...;2;1
$1;...;0;Bookstore.searchBook(...);...;2;...871392673;...1007805334;...;0;0
```

FIGURE 2. Extracting monitoring records by traceId.

2) In the same run path, the value of **eoi** increases gradually from 0, so the monitoring records are sorted by **eoi**. The sorted records are shown in Figure 3. As can be seen from Figure 3, after all monitoring records with the same **traceId** are sorted by the value of **eoi**, the value of **ess** increases or decreases with the method call or return. Here, **ess** represents the size of the execution stack in the system's run path, so its value changes as follows.

- After method *a* calls method *b*, the **ess** of *b* is equal to the **ess** of *a* plus 1.
- After method *b* returns method *a*, the **ess** of *a* is equal to the **ess** of *b* minus 1.



```
TraceId 8430034814995791873 (NOSESSION)
<[0,0] 1257440666759412388-1257440666841265860 srv0::Bookstore.searchBook(...)>
<[1,1] 1257440666805601818-1257440666807695902 srv0::Catalog.getBook(...)>
<[2,1] 1257440666820790063-1257440666841169272 srv0::CRM.getOffers(...)>
<[3,2] 1257440666820839575-1257440666840922990 srv0::Catalog.getBook(...)>
```

FIGURE 3. Monitoring data sorted by eoi.

3) According to the data correlation, the time interval constraint is set as 14 days. Branch and merge all monitoring records with the same **traceId** within the time interval constraint, and code each node in dictionary order, as shown in Figure 4.

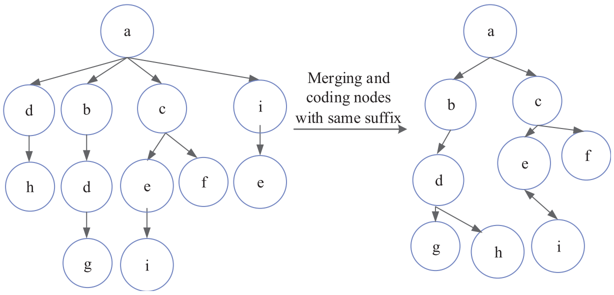


FIGURE 4. Encoding tree.

The mining process of users' normal sequence behavior for D2D communications is shown in Figure 5.

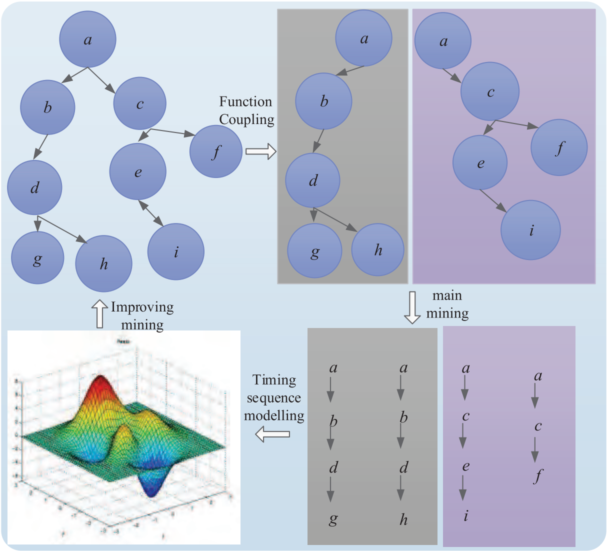


FIGURE 5. Mining process of normal sequential behavior for D2D communications.

C. MULTILEVEL BLACKLIST

The blacklist in this article stores the device name and device identifier IMEI (International Mobile Equipment Identity). IMEI is a 15-digit "electronic serial number", which corresponds to each mobile device and is the global unique code. The device name and device identifier are used to ensure that the mobile device used in the system is not misappropriated or illegal.

The two mechanisms to manage the blacklist are automatic management and manual management. At the beginning, the system will obtain a list of abnormal users similar to the blacklist through automatic detection or manual labeling. In order to reduce the false alarm rate, the blacklist is classified according to the deviation degree from normal behaviors for D2D communications, which are the first-level risk user, the second-level abnormal user and the third-level malicious user respectively. When the deviation degree $H \in [0, 0.4)$, the user for D2D communications is the first-level risk user, that is, the false normal user. The system does not take any measures, but keeps the event record and monitors its behavior in real time. When $H \in [0.4, 0.8)$, the user is a second-level abnormal user, and can only conduct a series of operations to access non-sensitive data such as browsing the website. The system keeps the event record and protects the file system. When $H \in [0.8, 1)$, the user is a third-level malicious user, and the system directly terminates all requests and isolates him from the network. The purpose of using the blacklist technology is to ensure that the mobile cloud service system can guarantee the basic functions of the system to the maximum extent and satisfy the user's service request to the maximum extent for D2D communications with security risks and threats.

D. REAL-TIME IDENTIFICATION OF USERS' SEQUENCE BEHAVIOR BASED ON HIERARCHICAL MATCHING METHOD FOR D2D COMMUNICATIONS

Through the mining of user behavior pattern for D2D communications, the sequential behavior of each user within the time interval constraint is grouped and coded, and the hierarchical matching method is used for real-time discrimination. The hierarchical matching in this paper is divided into two stages: fine-grained matching and coarse-grained matching. Fine-grained matching refers to similarity matching algorithm to determine whether the behavior is less than the set threshold value. If it is less than the threshold value, it is considered as a normal user. The user is allowed to carry out series of operations, and the cloud system should provide timely and efficient cloud service. Otherwise, coarse-grained matching is carried out for this behavior, that is, to judge whether its deviation degree is within the trusted tolerance range of the user; if so, it is false normal behavior. Then, the identified user behavior is compared with the blacklist, and the behavior is updated to the blacklist and corresponding response is made.

1) FINE-GRAINED MATCHING

In this paper, only a few important attributes are extracted for similarity calculation, and the users' sequence behavior with different offsets and amplitudes is normalized. In this way, the user's sequence behavior has the same length and is free from noise such as offset and amplitude, so it is suitable to use Euclidean distance for similarity measurement. In order to increase the discrimination accuracy, weighted Euclidean distance is used in this paper, and the matching is as follows.

$$D(i,j) = \sqrt{\sum_{m=1}^{n} (\tau_t \cdot \omega_{norm(i)}^{t} \cdot (i_t - j_t)^2)} \leq \varepsilon \tag{2}$$

$$\tau_t = \frac{|t_{current} - t_b|^{-1}}{\sum_{k=1}^{n} |t_{current} - t_k|^{-1}} \tag{3}$$

$$\omega_{norm(i)}^{t} = \frac{H(X^i)}{\sum_{i=1}^{n} H(X^i)} = \frac{\sum_{j=1}^{r} p(x_j^i) log_2 p(x_j^i)}{\sum_{i=1}^{n} \sum_{j=1}^{r} p(x_j^i) log_2 p(x_j^i)} \tag{4}$$

Where i and j are two sequence behaviors with both lengths of n, and τ_t indicates the time decay factor in t-th time period. ε denotes the threshold, and the better threshold is determined to be 0.64 through 10-fold cross validation and gradient descent method. t_b represents the time when the user for D2D communications has the b-th interaction with the cloud service resource. $\omega_{norm(i)}^{t}$ denotes the weight factors of each attribute of node i after normalization based on information entropy, and $\omega_{norm(i)}^{t} \in [0, 1]$. x_j^i represents the j-th attribute variate of user i, r is the attribute dimension of each user, and $p(x_j^i)$ represents the probability of each user behavior emerging in the total user behaviors, where $\sum_{j=1}^{r} p(x_j^i) = 1$, $0 \leq p(x_j^i \leq 1$ $(i = 1, 2, \ldots, r)$, $j = 1, \ldots, n$. That is, the longer the time, the smaller the weight; The closer to the current time, the greater the weight. If the formula (2) is satisfied, it is normal sequence behavior; otherwise, it is abnormal sequence behavior and coarse-grained matching is carried out.

2) COARSE-GRAINED MATCHING

Construct the random variable $u = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$. The procedure for calculating the confidence interval of each parameter item of the user's sequence behavior for D2D communications is as follows. The sample variance of each parameter is calculated as equation (5).

$$D(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{5}$$

where X_i is the sample value of each parameter term above, and $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

The standard deviation is calculated according to equation (6).

$$S(x) = \sqrt{D(x)} \tag{6}$$

The upper and lower bound of each parameter is set according to the standard deviation.

$$(\overline{X} - S(x), \overline{X} + S(x)) \tag{7}$$

If users' behaviors are identified online simply by using the trusted tolerance range specified in formula (7), the evaluation indexes such as detection rate and false alarm rate are not ideal. In order to reduce the false alarm rate, the tolerance factor δ is introduced to change the size of the confidence interval. Here, the better tolerance factor is determined by 10-fold cross validation and gradient descent method. The overall trusted tolerance range of user i is calculated as follows.

$$TH = (\overline{X} - S(x) \cdot \delta, \overline{X} + S(x) \cdot \delta) \tag{8}$$

The distribution function of user *i* at the current moment is shown in equation (9).

$$f_{X_i} = \frac{1}{\sqrt{2\pi}S(x)} e^{-\frac{(X_i-\overline{X})}{2D(x)}} \tag{9}$$

The deviation of user *i* is calculated as follows.

$$H = P\left\{|X_i - \overline{X}| \leq f(\frac{TH}{2})\right\} \tag{10}$$

Judge whether the deviation degree of user *i* at the current moment falls into the range of formula (10), and if it falls into the range of trusted tolerance, it is considered as normal user behavior for D2D communications; Otherwise it's abnormal behavior.

E. AUTONOMOUS OPTIMIZATION OF USER SEQUENCE BEHAVIORS BASED ON PATTERN GROWTH

Real-time and accurate identification of users' sequence behavior for D2D communications is an important prerequisite to determine whether the requested service can be normally provided. Although some irrelevant interference subgraphs can be filtered out in the process of hierarchical matching, so as to improve the real-time performance of the identification process as much as possible, the accuracy and effectiveness of real-time identification will still be affected by the fact that the timing steps involved in users' sequence behavior are complex graph structures with continuous and dynamic growth. Therefore, based on the extensible features of user behaviors, this paper has analyzed the function flow and data flow of nodes after real-time identification of user's sequence behavior. We use pattern-growth method based on maximum rightmost path extension to research the skipping and diffusing process of user abnormal steps. At last, the form of normal and abnormal node subgraph is used to predict service steps, service effect and scope of diffusion effect, which might be affected by abnormal behavior. The prediction results will affect the determination of users' sequence behavior for D2D communications from three aspects. 1) Fewer user behavior steps are used to identify abnormal behavior in advance. 2) It generates feedback on the completed process of user behaviors identification and user normal sequence behavior sets to optimize the identification process and update normal sequence behavior sets. 3) The nearest neighbor algorithm is used to classify user abnormal behaviors and update them to blacklist.

Based on the dependency relationship between node $S^{(l)}$ and $S^{(l+1)}$ in user abnormal sequence behavior fragment, to determine whether node $S^{(l)}$ has effect on the calling node $S^(l-1)$ and the called node $S^{(l+1)}$. Firstly, starting from the first exception node $S^{(l)}$, change the parameter value passed in by the calling node $S^{(l-1)}$ or its return values used by other node $S^{(l+1)}, \ldots, S^{(m)}$. Then the functional flow is used to analysis whether the implementation of node $S^{(l)}$ will affect the outside world. If its implementation is relatively closed, the exception of node $S^{(l)}$ will not be propagated. Otherwise, the node $S^{(l+1)}, \ldots, S^{(m)}$ that may be affected by $S^{(l)}$ will be further analyzed by analyzing the flow direction of its return value. Then, we proceed from the node $S^{(l+1)}, \ldots, S^{(m)}$ that may be affected to carry out the diffusion analysis.

The process of analyzing function flow and data flow is shown as Figure 6. The abnormal and non-abnormal diffusion node sets are formed as $A_s = \{S^{(1)}, S^{(3)}, S^{(l+3)}\}$ and $N_s = \{S^{(2)}, S^{(l)}, S^{(l+1)}, S^{(l+2)}\}$ respectively.

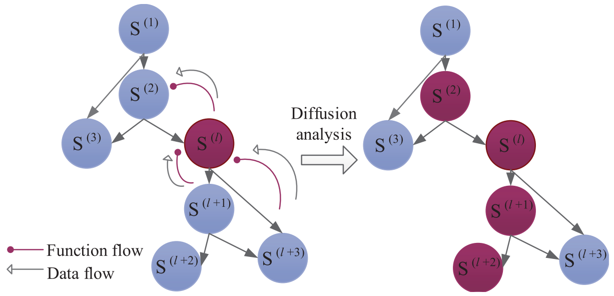


FIGURE 6. Analysis to node function flow and data flow.

In the real world, the scale of data in the mobile cloud environment grows exponentially every day, and the database keeps changing with time. For this kind of massive dynamic database, if every update of the database requires a new mining of the entire database, it will lead to inefficiency and waste of large amounts of resources. Therefore, the pattern growth method is adopted to mine the updated incremental data, and the right-most path extension method [32] can improve the completeness of the updated database and reduce the redundancy. However, this method produces a large number of candidate subsets for large data sets. In this paper, the maximum right-most path extension method is adopted, that is, only the last node that is traversed in depth priority is extended.

Using this method, a complete space for user behavior pattern growth for D2D communications is constructed to generate normal node subgraph sets and abnormal influence node subgraph sets (as shown in Figure 7). These subgraph sets are used to update user normal behavior sets and optimize the judging process of abnormal behaviors for D2D communications. The anomalous diffusion node set A_s and non-anomalous diffusion node set N_s, in diffusion result of

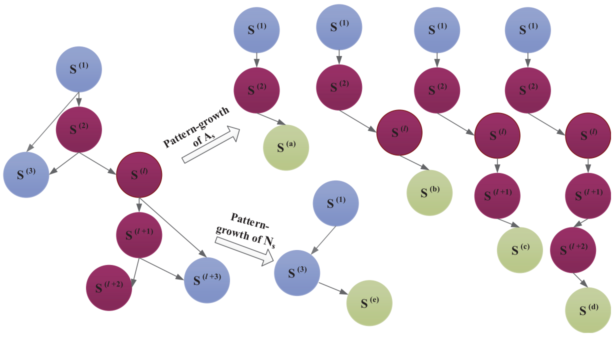


FIGURE 7. Maximum rightmost path extension schematic.

abnormal nodes, are extended by the depth-first algorithm respectively.

To classify user abnormal behaviors for D2D communications, that is, to cluster the abnormal nodes, there is a problem of high overhead. Therefore, the calculated mean value and confidence interval are used to represent the abnormal sequence behavior of users. It is assumed that there are n nodes and m nearest neighbor nodes, and the overall average absolute error is calculated as follows.

$$\varpi = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}|x_{ij} - x'_{ij}| \tag{11}$$

$$x'_{ij} = \frac{|x_{ij} - \overline{x_{ij}}|}{\sum_{i=1}^{n}\sum_{j=1}^{m}|x_{ij} - \overline{x_{ij}}|} \tag{12}$$

where x'_{ij} is the predicted value, and x_{ij} is the actual value.

If equation (13) is satisfied, the abnormal behavior is classified into the same group.

$$\frac{x_{ij} - \varpi}{\varpi} < \zeta \tag{13}$$

The autonomous optimization method of user behavior for D2D communications is realized by calculating the similarity between the user node and its neighbor node, which can predict the operation effect that the user behavior may bring. In the classification result set, the neighbor node sets of node i are the nodes in the circle whose center is i, and radius is R. The similarity between node i and j is defined as the number of paths with path length in l range. It is shown as equation (14).

$$s(i,j) = \sum_{k=2}^{l}\frac{1}{k-1}\cdot\frac{1}{t_m - t_s}\cdot\frac{|path^{k}_{i,j}|}{\Pi^{k}_{p=2}(n-p)} \tag{14}$$

Where n is the number of nodes, l is the path length between node i and j, in which, a node can only appear once in a path, that is, it does not contain a circular path. $\frac{1}{k-1}$ and $\frac{1}{t_e - t_s}$ are decay factors, and the path weight depends on its length l and time t. The weight will be smaller when the path is longer and the time is older, otherwise, the weight will be greater. $|path^{k}_{i,j}|$ is the set of all paths that can go from i to j with length k. Assume that each node in the tree could be connected to others, and $\Pi^{k}_{p=2}(n-p)$ is the set of all paths from i to j. The similarity is calculated by $|path^{k}_{i,j}|/\Pi^{k}_{p=2}(n-p)$, and it is in (0, 1], which conforms to the predicted results. If the two nodes are highly similar, the value of s(i,j) is close to 1; otherwise, if the two nodes are not similar, the value of $s(i,j)$ is close to 0. Arrange the similarity in descending order $(s_1, s_2, \ldots, s_n)$, and take max $(s_1, s_2, \ldots, s_u)$ as the best prediction node.

F. ALGORITHM OF USER BEHAVIOR PATTERNS MINING AND ONLINE IDENTIFICATION FOR D2D COMMUNICATIONS

The User behavior patterns mining and online identification for D2D communications mainly include the following steps:

1) Normalize user's sequence behaviors: some important attributes of behavior data are extracted to calculate the mean $\mu(x)$ and variance $\rho(x)$ of the user's behavior X, which will be standardized as $X = (< X'_1, d(X'_1), l'_1 >, < X'_2, d(X'_2), l'_2 >, \ldots, < X'_n, d(X'_n), l'_n >)$.
2) Construct the ordered tree: the preprocessed user sequence behavior data are grouped and encoded, using the most recent unused replacement algorithm.
3) Identify abnormal user behavior in real time: The ordered tree is fine-grained matched with user normal behavior database for D2D communications, and the weighted Euclidean distance is used to calculate similarity. 10 fold-cross validation and gradient descent method determine the better threshold. If the value of the behavior is less than the threshold, it is a normal one, and the algorithm goes to step 5, otherwise step 4.
4) Identify more deeply: by random variable $u = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ to calculate the standard deviation of the user's sequence behavior, and then calculate the tolerance range of trustability $TH = (\overline{X}-S(x)\cdot\delta, \overline{X}+S(x)\cdot\delta)$ and the deviation $H = P\left\{|X_i - \overline{X}| \le f(\frac{TH}{2})\right\}$. It is a normal behavior if the deviation is within the range, otherwise, it is an abnormal one.
5) Judge the level of blacklist: to judge the blacklist level of the identified user behavior according to the second chance mechanism. Users in different levels will obtain the corresponding services. On the basis of guaranteeing the basic function of the system, the service request of the user can be satisfied to a maximum extent.
6) Analyze the functional flow and data flow: based on the dependence between node $S^{(l)}$ and $S^{(l+1)}$ in the segment of the abnormal user's sequence behavior for D2D communications, the function flow and data flow analysis of the recognized abnormal behaviors are analyzed by changing the parameter values passed in by the node $S^{(l-1)}$ that calls it or the return values used by other nodes $S^{(l+1)}, \ldots, S^{m}$.
7) Construct behavior pattern space: the pattern-growth method based on maximum and rightmost path extension is used to construct the complete user behavior patterns growth space $A_s = \{S^{(1)}, S^{(3)}, S^{(l+3)}\}$ and $N_s = \{S^{(2)}, S^{(l)}, S^{(l+1)}, S^{(l+2)}\}$ for D2D communications, to update users normal time-sequence behavior sets and optimize the process of identifying abnormal behavior.
8) Get the best predictor: the total mean absolute error ϖ of the recognized abnormal behavior is calculated to get the best predictor max$(s_1, s_2, \ldots, s_u)$, for identifying the abnormal types next time.

It is shown in Alogrithm1.

IV. EXPERIMENT AND ANALYSIS

In order to verify the recognition effect of the proposed algorithm in the real system, some simulations and experiments are carried out. KDD CUP99 dataset containing

Algorithm 1 The Process of Algorithm of User Behavior Patterns Mining

```
Input: $X$.
Output: The values of users' behavior.
1:  Normalize the $X$ as $X = (<X'_1, d(X'_1), l'_1>, <X'_2, d(X'_2), l'_2>, \ldots, <X'_n, d(X'_n), l'_n>)$.
2:  Using the replacement algorithm to construct the ordered tree and calculate $\varpi$.
3:  if $\varpi < \varepsilon$ then
4:      goto step 7.
5:  end if
6:  Calculate equations 7, 8 and 10 by random variable $u = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Then, using equation 7 to judge the normality of use' behavior.
7:  Using the second chance mechanism to judge the level of blacklist.
8:  : Analysis the functional flow and data flow by the dependence between node $S^{(l)}$ and $S^{(l+1)}$.
9:  Update the behavior pattern space by user behavior patterns growth space $A_s = \{S^{(1)}, S^{(3)}, S^{(l+3)}\}$ and $N_s = \{S^{(2)}, S^{(l)}, S^{(l+1)}, S^{(l+2)}\}$ based on the pattern-growth method.
10: Calculate the total mean absolute error $\varpi$ and choose the best predictor $\max(s_1, s_2, \ldots, s_u)$.
```

TABLE 2. The experimental environment.

Parameters	Values
CPU	3.4GHz Intel(R) Core(TM) i3-2130 CPU
RAM	4GB
HDD	498GB/7200r/min
OS	Windows7
IDE	MatLab R2012b

approximately 5 million data records is used as the source data. Each record contains 1 logo and 41 properties, which have been identified as normal or specific attacks. The real environment is an online shop demo that has been instrumented with Kieker, where each record contains 11 properties. This paper uses MatLab to simulate the algorithm, which is translated into MatLab frame for verification. The experiment runs on ordinary PC, and the configuration follows as Table 2.

A. ABNORMAL ANALYSIS EVALUATION INDEX

The indexes are calculated as follows.

$$I_{DR} = \frac{e}{g} \tag{15}$$

where e is the number of detected attack samples, and g is the total number of attack samples.

$$I_{AR} = \frac{c}{a} \tag{16}$$

where c is the detected sample number in all abnormal samples, and a is the total number of abnormal samples.

$$I_{FNR} = \frac{h}{k} \tag{17}$$

where h is the detected abnormal samples mistaken as normal samples, and k is the total abnormal sample number.

$$I_{FAR} = \frac{b}{z} \tag{18}$$

where b is number of normal behaviors mistaken as abnormal behaviors, and z is total normal samples number.

B. EXPERIMENTAL PROCEDURE

Because of the space limitation, this part only illustrates the well-known KDD CUP99 dataset.

1) NUMERICAL ENCODING

KDD CUP99 dataset has completed data collection, partial data preprocessing and feature extraction. However, protocol type of the 2^{nd} dimension, service type of the 3^{rd} dimension and state flag of the 4^{th} dimension in the 41-dimensional characteristic variable are all non-numeric forms, which cannot be recognized by the algorithm in this paper, so it must be numerically processed.

The rule of numerical encoding is to count the data in 2^{nd} dimension, 3^{rd} dimension and 4^{th} dimension, and replace the original contents with serial Numbers.

- **Protocol type**
  There are three types of protocol that are encoded as shown in Table 3.

TABLE 3. Protocol type encoding.

Protocol type	Encoding
Icmp	001
tcp	010
udp	100

- **Service type**
  A total of 70 types of services appear, coded as shown in Table 4.
- **Status flag**
  There are a total of 11 state flags, coded as shown in Table 5.
- **Standardize behavior data attribute value**
  Assume that there are n behavior data. The attribute value of each behavior data X is normalized to be distributed in [0,1], as shown in equation (19).

$$X' = \frac{X - min}{max - min} \tag{19}$$

where X' represents normalized data, min and max represent the minimum and maximum value of the specific attributes in n behavior data respectively.

TABLE 4. Service type encoding.

Status type	Encoding	Status type	Encoding
aol	00000000000000000000000000000000000 00000000000000000000000000000000001	Netbios_dgm	00000000000000000000000000000000001 00000000000000000000000000000000000
auth	00000000000000000000000000000000000 00000000000000000000000000000000010	netbios_ns	00000000000000000000000000000000010 00000000000000000000000000000000000
bgp	00000000000000000000000000000000000 00000000000000000000000000000000100	Netbios_ssn	00000000000000000000000000000000100 00000000000000000000000000000000000
courier	00000000000000000000000000000000000 00000000000000000000000000000001000	netstat	00000000000000000000000000000001000 00000000000000000000000000000000000
csnet_ns	00000000000000000000000000000000000 00000000000000000000000000000010000	nnsp	00000000000000000000000000000010000 00000000000000000000000000000000000
ctf	00000000000000000000000000000000000 00000000000000000000000000000100000	nntp	00000000000000000000000000000100000 00000000000000000000000000000000000
daytime	00000000000000000000000000000000000 00000000000000000000000000001000000	ntp_u	00000000000000000000000000001000000 00000000000000000000000000000000000
discard	00000000000000000000000000000000000 00000000000000000000000000010000000	other	00000000000000000000000000010000000 00000000000000000000000000000000000
domain	00000000000000000000000000000000000 00000000000000000000000000100000000	pm_dump	00000000000000000000000000100000000 00000000000000000000000000000000000
domain_u	00000000000000000000000000000000000 00000000000000000000000001000000000	pop_2	00000000000000000000000001000000000 00000000000000000000000000000000000
echo	00000000000000000000000000000000000 00000000000000000000000010000000000	pop_3	00000000000000000000000010000000000 00000000000000000000000000000000000
eco_i	00000000000000000000000000000000000 00000000000000000000000100000000000	printer	00000000000000000000000100000000000 00000000000000000000000000000000000
ecr_i	00000000000000000000000000000000000 00000000000000000000001000000000000	private	00000000000000000000001000000000000 00000000000000000000000000000000000
efs	00000000000000000000000000000000000 00000000000000000000010000000000000	red_i	00000000000000000000010000000000000 00000000000000000000000000000000000
exec	00000000000000000000000000000000000 00000000000000000000100000000000000	remote_job	00000000000000000000100000000000000 00000000000000000000000000000000000
finger	00000000000000000000000000000000000 00000000000000000001000000000000000	rje	00000000000000000001000000000000000 00000000000000000000000000000000000
ftp	00000000000000000000000000000000000 00000000000000000010000000000000000	shell	00000000000000000010000000000000000 00000000000000000000000000000000000
ftp_data	00000000000000000000000000000000000 00000000000000000100000000000000000	smtp	00000000000000000100000000000000000 00000000000000000000000000000000000
gopher	00000000000000000000000000000000000 00000000000000001000000000000000000	sql_net	00000000000000001000000000000000000 00000000000000000000000000000000000
harvest	00000000000000000000000000000000000 00000000000000010000000000000000000	ssh	00000000000000010000000000000000000 00000000000000000000000000000000000
hostnames	00000000000000000000000000000000000 00000000000000100000000000000000000	sunrpc	00000000000000100000000000000000000 00000000000000000000000000000000000
http	00000000000000000000000000000000000 00000000000001000000000000000000000	supdup	00000000000001000000000000000000000 00000000000000000000000000000000000
http_2784	00000000000000000000000000000000000 00000000000010000000000000000000000	systat	00000000000010000000000000000000000 00000000000000000000000000000000000
http_443	00000000000000000000000000000000000 00000000000100000000000000000000000	telnet	00000000000100000000000000000000000 00000000000000000000000000000000000
http_8001	00000000000000000000000000000000000 00000000001000000000000000000000000	tftp_u	00000000001000000000000000000000000 00000000000000000000000000000000000
imap4	00000000000000000000000000000000000 00000000010000000000000000000000000	tim_i	00000000010000000000000000000000000 00000000000000000000000000000000000
IRC	00000000000000000000000000000000000 00000000100000000000000000000000000	time	00000000100000000000000000000000000 00000000000000000000000000000000000
iso_tsap	00000000000000000000000000000000000 00000001000000000000000000000000000	urh_i	00000001000000000000000000000000000 00000000000000000000000000000000000
klogin	00000000000000000000000000000000000 00000010000000000000000000000000000	urp_i	00000010000000000000000000000000000 00000000000000000000000000000000000
kshell	00000000000000000000000000000000000 00000100000000000000000000000000000	uucp	00000100000000000000000000000000000 00000000000000000000000000000000000
ldap	00000000000000000000000000000000000 00001000000000000000000000000000000	uucp_path	00001000000000000000000000000000000 00000000000000000000000000000000000
link	00000000000000000000000000000000000 00010000000000000000000000000000000	vmnet	00010000000000000000000000000000000 00000000000000000000000000000000000
login	00000000000000000000000000000000000 00100000000000000000000000000000000	whois	00100000000000000000000000000000000 00000000000000000000000000000000000
mtp	00000000000000000000000000000000000 01000000000000000000000000000000000	X11	01000000000000000000000000000000000 00000000000000000000000000000000000
name	00000000000000000000000000000000000 10000000000000000000000000000000000	Z39_50	10000000000000000000000000000000000 00000000000000000000000000000000000

TABLE 5. State flag encoding.

Status type	Encoding	Status type	Encoding
OTH	00000000001	S1	00001000000
REJ	00000000010	S2	00010000000
RSTO	00000000100	S3	00100000000
RSTOS0	00000001000	SF	01000000000
RSTR	00000010000	SH	10000000000
S0	00000100000		

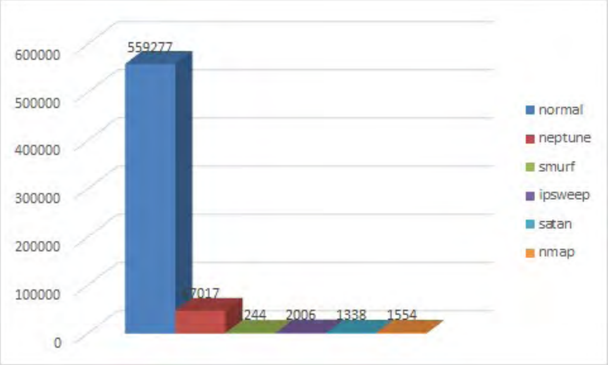


FIGURE 8. The flowchart of learning sample distribution.

2) CONSTRUCT TEST SAMPLE

The file of kddcup.data.gz provided by KDD99 dataset has millions of data. After deleting repeated data, 614,451 valid data were finally selected as learning samples, and their distribution is shown in Figure 8.

3) FEATURES SELECTION

The number of samples is large, and there are also many features and attributes of each sample, which will inevitably increase the calculation amount, leading to too long training time, so it is impossible to conduct data mining and analysis directly. Therefore, statistical variances and rough set theory are adopted for attribute reduction. At last, 6 basic features that have a great impact on the type of abnormal behavior are selected. See Table 6 for detailed description of the features.

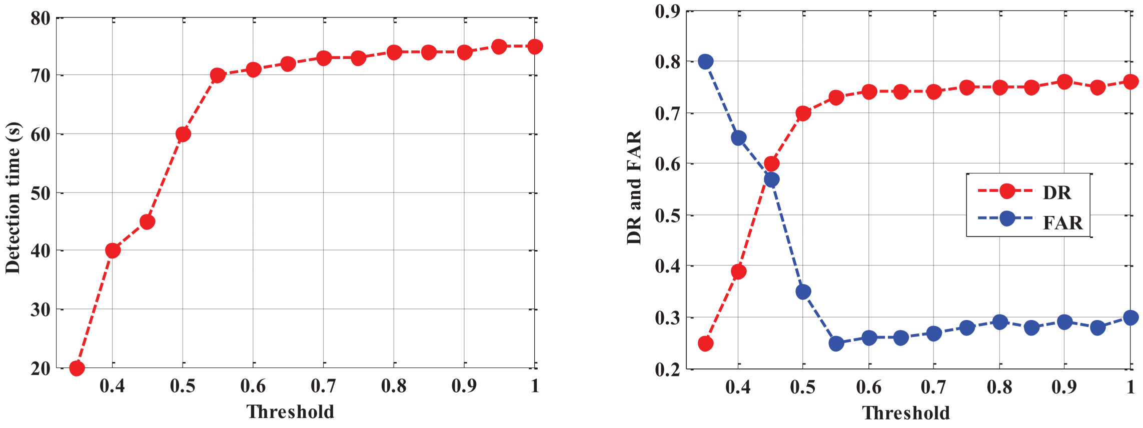
C. SIMULATION EXPERIMENT AND ANALYSIS

As we can see from Figure 9, with the increase of threshold value, before 0.65, there is almost a linear growth, and after 0.65, there is a slow growth. This is because as the threshold increases, the more stringent the requirement to determine the malicious node, the more time it takes. According to the simulation results, after weighing the relationship among response time, detection rate and false alarm rate, the optimal confidence is set as 0.82.

The following simulations compare five indicators, Detection Speed (DS), Detection Rate (DR), Accuracy Rate (AR), False Alarm Rate (FAR) and False Negative Rate (FNR). Due to the unbalanced distribution of data types in the data set of KDD99, which is inconsistent with the distribution of data types in the real network, the experiment of this system randomly sampled the unbalanced data in the training sample of KDD99 to build the training sample. See Table 7. All of following simulation results are the average of ten random simulation experiment results.

TABLE 6. Table of basic features.

Sequence number	Feature name	Feature type	Feature description
1	Duration	continuous	The connecting record time(s)
2	protocol_type	symbolic	Protocol type such as tcpčňudp, etc.
3	service	symbolic	Service type of destination node such as http
4	flag	symbolic	The connection is wrong or right state
5	src_bytes	continuous	The number of bytes sending from source node to end node
6	dst_bytes	continuous	The number of bytes sending from end node to source node



(a)

(b)

FIGURE 9. The determination of optimum threshold value. (a) Testing time corresponding to different confidence. (b) The average DR and FAR corresponding to different confidence.

TABLE 7. The simulation data.

Data set	Sum	Abnormal data
Set 1	100000	5%
Set 2	100000	20%
Set 3	200000	5%
Set 4	200000	20%
Set 5	300000	10%
Set 6	300000	30%
Set 7	400000	15%
Set 8	400000	40%
Set 9	500000	10%
Set 10	500000	25%

As the abnormal behaviors of cloud users for D2D communications are analyzed in the mobile cloud environment, so we only contrast evaluation indexes with larger test samples. Figure 10 shows DS comparison among user abnormal behavior analysis method based on Bayesian clustering (Bayes), user abnormal behavior analysis method based on BP Neural Network (BPNN) and the proposed Patterns Mining and On-line Identification method (PMOI). As shown in Figure 10, DS of the proposed method is better and more stable than that of Bayes and BPNN. Because Bayes has many parameters, the training of network structure is too tedious, while the BPNN algorithm needs to model the sample in advance and train a large amount of data. In addition, both of them need to analyze all the attributes of the sample, some of which are very small and meaningless to DR's calculation. However, the algorithm proposed in this paper only analyzes part of the attributes. Based on the preprocessing result of the data, it uses a simple Euclidean distance to make abnormal judgment. Moreover, because hierarchical matching method is used in the proposed method to analyze user abnormal behaviors for D2D communications, its complexity is lower, which improves DS.

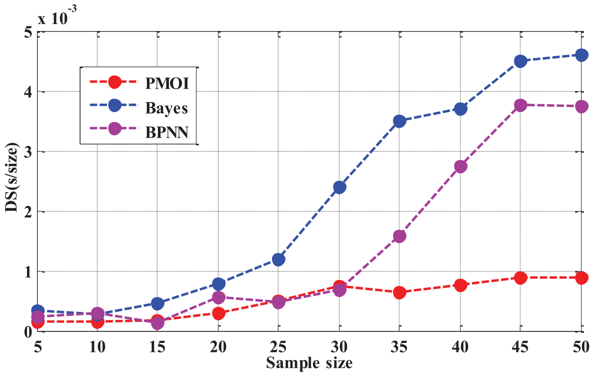


FIGURE 10. The comparison of DS.

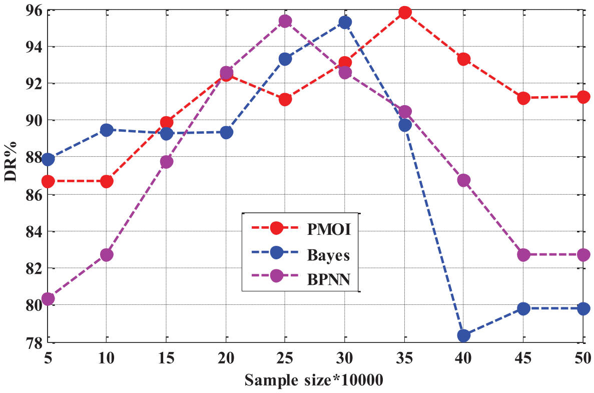


FIGURE 11. The comparison of DR.

An algorithm with higher detection rate can more accurately analyze abnormal behavior, intercept attack behavior, and effectively protect user personal data for D2D communications. As can be seen from figure 11, the overall DR of PMOI algorithm is relatively high and stable, because this algorithm does not require complex network topologies and does not need to select many parameters. However, it is difficult to select the element parameters and network topology in BPNN algorithm. Bayes algorithm solves the posterior probability according to the prior probability. If some important parameters change, Bayes will not play a key role. Here, DR of PMOI is lower than that of the other two algorithms when test samples are in 50000–150000, because there are unknown abnormal types. However, the algorithm in this paper has the ability of autonomous optimization and the detection rate gradually returns to normal, which is better than other two algorithms. But when the sample is in 200000–300000, new unknown attack types appear in test sample, and attack types in database are gradually saturated. So with the increasing of testing samples, abnormal types reach saturation gradually, which shows obvious advantages. Hierarchical matching and pattern-growth method are used in autonomous optimization, which results in better DR.

In Figure 12, AR of the proposed algorithm is lower when testing samples are about 50000–150000, because there are unknown attack types in testing samples. In this paper, the weighted Euclidean distance is used to carry out abnormal analysis, because the user's attributes for D2D communications have been preprocessed, and the weighted Euclidean distance is enough to meet their needs. In order to further guarantee its AR, this paper further identifies the abnormal

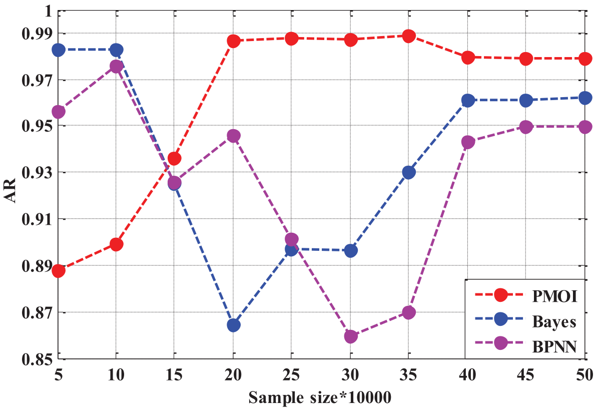


FIGURE 12. The comparison of AR.

behaviors identified for the first time. Moreover, AR returns to normal because of its ability of self-optimization. Overall, with the increasing of sample size, the proposed algorithm has higher AR than the other two algorithms, and it becomes gradually smooth.

The contrasts of FAR and FNR are shown in Figure 13 and Figure 14 respectively. The proposed algorithm in this paper is superior to the other two algorithms on the whole, because the BPNN algorithm is sensitive to noise data, while the Bayes algorithm often fails to establish the hypothesis between attributes, and because prior probability is needed to know, there is error rate in classification decision. In which, FAR of this algorithm is higher than that of the other two when testing samples are about 500001̃00000 and 1500002̃50000. On the other hand, its FNR is lower. The proposed algorithm has better scalability, self-adaptability and higher recognition ability.

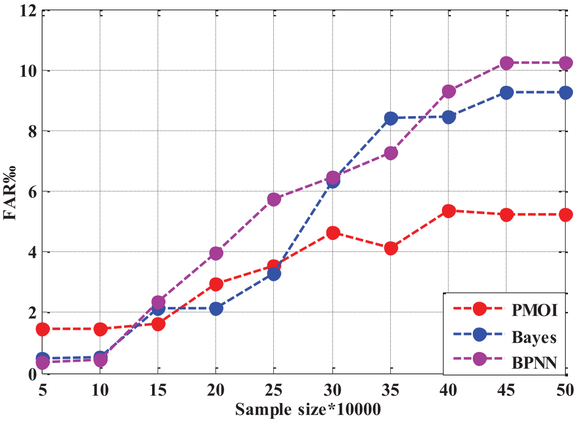


FIGURE 13. The comparison of FAR.

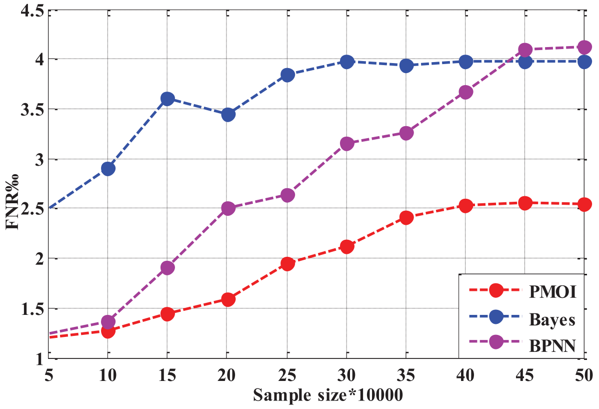


FIGURE 14. The comparison of FNR.

In KDD99 dataset, the five kinds of attack type such as perl, ftp_write, phf, multihop and spy take up a very small proportion, and the recognition of it only reduces the test speed and makes no sense. Therefore, these 5 types of samples are removed in the classification of this paper.

Figure 15 is a comparison between the predicted classification results and the actual classification results of the algorithm in this paper. We can see from it that DR of classification is as high as 94.8%.

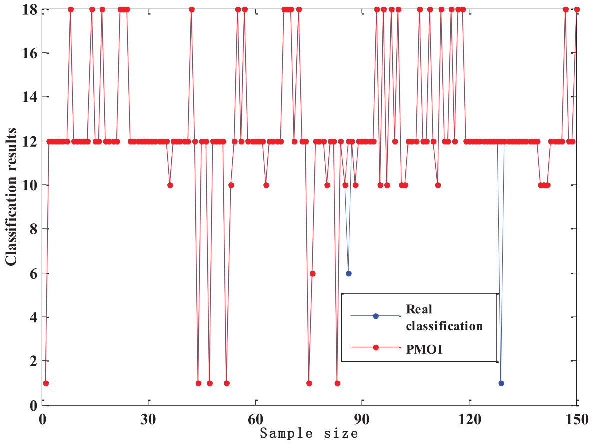


FIGURE 15. The comparison of forecast classification result and real classification result.

The following are the experimental results of applying the algorithm proposed in this paper to the real network environment of Kieker. Table 8 shows that our algorithm has an ideal application results in real environment.

Figure 16 is the comparison of forecast classified result and real classified result. From it, we can see that AR of classification is up to 90.2%. Figure 17 is the comparison of recognition rate of different testing data in real network environment. From it, we can see that the proposed algorithm has better ability to identify unknown exception types, and has a more satisfying classification result for unknown types.

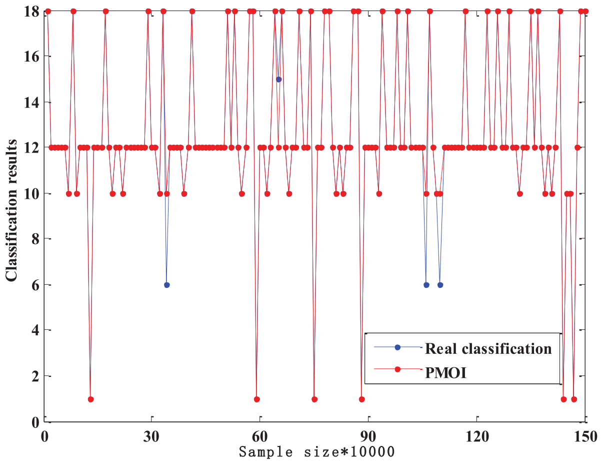


FIGURE 16. The comparison of forecast classification result and real classification result.

The experimental results show that the proposed algorithm improves AR and ensures DS, and FAR and FNA are improved also. As we can see from above simulation results, the algorithm has a good stability, which could identify abnormal behaviors effectively, with better scalability and self-adaptability. After several experimental verifications, the simulation results are consistent with above results.

TABLE 8. The experimental results in real environment.

Indicators \ Samples	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
DR(%)	86.73	84.73	95.86	96.48	93.12	92.12	90.86	89.35	84.2	87.26
AR(%)	96.29	96.29	90.55	80.42	83.72	84.62	89.01	91.13	90.09	92.22
FAR(%)	0.46	0.46	7.02	15.94	12.54	11.64	8.15	5.37	6.24	5.24
FNR(%)	0.37	0.36	0.31	0.49	0.38	0.28	0.41	0.36	0.39	0.34
Needing time(s)	0.032	0.032	0.027	0.037	0.037	0.042	0.045	0.047	0.050	0.053

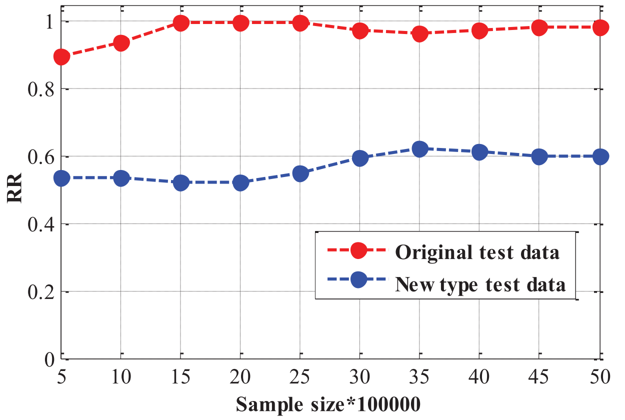


FIGURE 17. The comparison of recognition rate of different testing data.

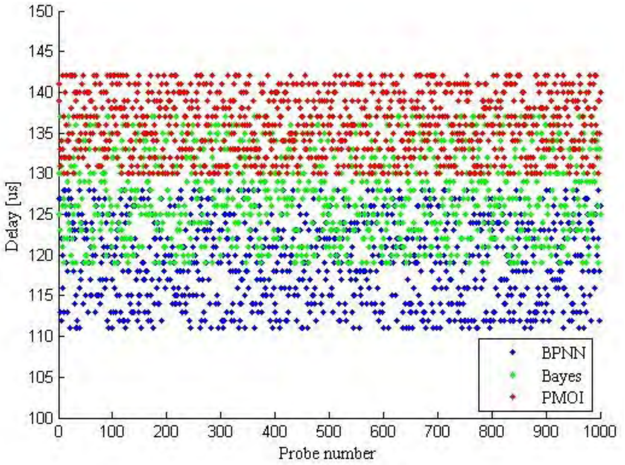


FIGURE 18. Comparison of delay.

Figure 18 shows the delay of the three algorithms. The PMOI algorithm has the lowest delay, because it has the lowest time complexity and better stability. It can carry out a deeper identification to the behavior identified as the abnormal one for the first time. Figure 19 shows the values of the delay collected during the tests without and with PMOI. The measured average delay without PMOI is 24 1.2 us, which shows in Figure 19. The measured average delay with PMOI is 136 5 us. It can be seen that the algorithm in this paper improves the accuracy to a certain extent, and reduces the rate of missing report and false reporting. This algorithm could better identify abnormal sequence behaviors of mobile terminal user for D2D communications and predict service

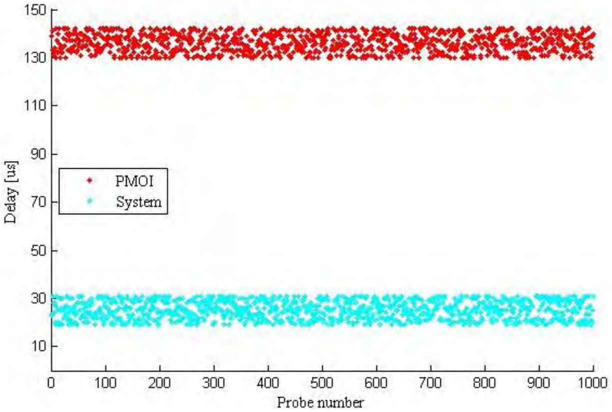


FIGURE 19. Distribution of probes in RC tests.

effect and the scope of the diffusion effect, which may be affected by abnormal behaviors.

V. CONCLUSIONS

With the further development of MCC and the further mature of D2D communications, MCC services develop rapidly in the world, becoming a new hotspot of mobile Internet services. While improving resource utilization, MCC brings great security risks and challenges to information security. It is of great significance to carry out research on the abnormal analysis and related key technologies of user behavior for D2D communications in MCC, helping to improve the security technology system and guarantee the security of MCC. A novel analysis method of user abnormal behavior based on autonomous optimization for D2D communications is proposed to identify user abnormal behaviors for D2D communications in real-time. The normalization is used to standardize identification fragment of "user sequence operation" which approximately satisfies normal distribution, then the method of hierarchical matching is used to real-time identify user behaviors for D2D communications by judging whether user behaviors for D2D communications are beyond the trusted tolerance scope. And pattern-growth method based on maximum rightmost path extension is used to construct complete user behavior patterns growth space, aiming at autonomously optimizing the process of identifying abnormal behaviors. The experimental results show that the algorithm proposed in this paper has good performance. It can improve recognition speed and recognition accuracy to a certain extent, and reduce FAR and FNR. It lays a solid foundation for the identification of the user identity and the trustability of its behavior for D2D communications before the cloud services enter the substantive service process. In addition, this scheme has better identifiable ability for the known abnormal types, which has poor recognition for unknown abnormal types. In future work, we should pay attention to improve the ability to analyzing unknown abnormal behaviors, research how to further improve the algorithm stability, and reduce FNR and FAR.

COMPETING INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCES

[1] Q. Fan and L. Liu, "A survey of challenging issues and approaches in mobile cloud computing," in *Proc. Int. Conf. Parallel Distrib. Comput., Appl. Technol. (PDCAT)*, 2016, pp. 87–90.

[2] N. Chalaemwongwan and W. Kurutach, "Mobile cloud computing: A survey and propose solution framework," in *Proc. Int. Conf. Elect. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, 2016, pp. 1–4.

[3] R. Zheng, M. Zhang, Q. Wu, W. Wei, and C. Yang, "A^{3srC}: Autonomic assessment approach to IOT security risk based on multidimensional normal cloud," *J. Internet Technol.*, vol. 16, no. 7, pp. 1271–1282, 2015.

[4] S. Naulaerts *et al.*, "A primer to frequent itemset mining for bioinformatics," *Briefings Bioinf.*, vol. 16, no. 2, pp. 216–231, 2015.

[5] K. M. Kumar and A. R. M. Reddy, "A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method," *Pattern Recognit.*, vol. 58, pp. 39–48, Oct. 2016.

[6] H. Shah-Mansouri, V. W. Wong, and R. Schober, "Joint optimal pricing and task scheduling in mobile cloud computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5218–5232, Aug. 2017.

[7] T. Bhatia and A. K. Verma, "Data security in mobile cloud computing paradigm: A survey, taxonomy and open research issues," *J. Supercomput.*, vol. 73, no. 6, pp. 2558–2631, 2017.

[8] J. Zhang, Z. Zhang, and H. Guo, "Towards secure data distribution systems in mobile cloud computing," *IEEE Trans. Mobile Comput.*, vol. 16, no. 11, pp. 3222–3235, Nov. 2017.

[9] B. Zhou and R. Buyya, "Augmentation techniques for mobile cloud computing: A taxonomy, survey, and future directions," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 13:1–13:38, 2018.

[10] G. Feng *et al.*, "Optimal transmission for scalable video coded streaming in cellular wireless networks with the cooperation of local peer-to-peer network," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 4, pp. 1–13, 2017. [Online]. Available: http://journals.sagepub.com/doi/pdf/10.1177/1550147717701436

[11] F. Bavaud, "Euclidean distances, soft and spectral clustering on weighted graphs," in *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Discovery Databases (ECML PKDD)*, Barcelona, Spain, Sep. 2010, pp. 103–118.

[12] H. Ayeldeen, A. E. Hassanien, and A. A. Fahmy, "Lexical similarity using fuzzy Euclidean distance," in *Proc. IEEE 2nd Int. Conf. Eng. Technol. (ICET)*, Cairo, Egypt, Apr. 2014, pp. 1–6.

[13] Q.-Y. Yan, S.-X. Xia, and K.-W. Feng, "Probabilistic distance based abnormal pattern detection in uncertain series data," *Knowl.-Based Syst.*, vol. 36, pp. 182–190, Dec. 2012.

[14] S.-Y. Ji, B.-K. Jeong, S. Choi, and D. H. Jeong, "A multi-level intrusion detection method for abnormal network behaviors," *J. Netw. Comput. Appl.*, vol. 62, pp. 9–17, Feb. 2016.

[15] P. Agius, Y. Ying, and C. Campbell, "Bayesian unsupervised learning with multiple data types," *Statist. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–27, 2009.

[16] L. Qin, Y. Ye, L. Su, and Q. Huang, "Abnormal event detection based on multi-scale Markov random field," in *Proc. 1st Chin. Conf. Comput. Vis. (CCCV)*, Xi'an, China, Sep. 2015, pp. 376–386.

[17] X. Ni, D. He, S. Chan, and F. Ahmad, "Network anomaly detection using unsupervised feature selection and density peak clustering," in *Proc. 14th Int. Conf. Appl. Cryptogr. Netw. Secur. (ACNS)*, Guildford, U.K., Jun. 2016, pp. 212–227.

[18] R. Zheng, J. Chen, M. Zhang, Q. Wu, J. Zhu, and H. Wang, "A collaborative analysis method of user abnormal behavior based on reputation voting in cloud environment," *Future Gener. Comput. Syst.*, vol. 83, pp. 60–74, Jun. 2018.

[19] R. Wang, J. Yan, D. Wu, H. Wang, and Q. Yang, "Knowledge-centric edge computing based on virtualized D2D communication systems," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 32–38, May 2018.

[20] Y. Zhou, Y. Wang, and X. Ma, "A user behavior anomaly detection approach based on sequence mining over data streams," in *Proc. Int. Conf. Parallel Distrib. Comput., Appl. Technol. (PDCAT)*, 2017, pp. 376–381.

[21] D. Wu, S. Si, S. Wu, and R. Wang, "Dynamic trust relationships aware data privacy protection in mobile crowd-sensing," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2958–2970, Aug. 2018.

[22] M. Bi, J. Xu, M. Wang, and F. Zhou, "Anomaly detection model of user behavior based on principal component analysis," *J. Ambient Intell. Humanized Comput.*, vol. 7, no. 4, pp. 547–554, 2016.

[23] M. M. Najafabadi, T. M. Khoshgoftaar, C. Calvert, and C. Kemp, "User behavior anomaly detection for application layer DDoS attacks," in *Proc. IEEE 18th Int. Conf. Inf. Reuse Integr. (IRI)*, San Diego, CA, USA, Aug. 2017, pp. 154–161.

[24] A. Zargar, A. Nowroozi, and R. Jalili, "XABA: A zero-knowledge anomaly-based behavioral analysis method to detect insider threats," in *Proc. Int. Iranian Soc. Cryptol. Conf. Inf. Secur. Cryptol. (ISCISC)*, 2016, pp. 26–31.

[25] K. Yoon and D.-H. Bae, "A pattern-based outlier detection method identifying abnormal attributes in software project data," *Inf. Softw. Technol.*, vol. 52, no. 2, pp. 137–151, 2010.

[26] Y. Wang, X. Wang, and X. Wang, "A spectral clustering based outlier detection technique," in *Proc. 12th Int. Conf. Mach. Learn. Data Mining Pattern Recognit. (MLDM)*, New York, NY, USA, Jul. 2016, pp. 15–27.

[27] F. Maciá-Pérez, J. V. Berna-Martinez, A. F. Oliva, and M. A. A. Ortega, "Algorithm for the detection of outliers based on the theory of rough sets," *Decision Support Syst.*, vol. 75, pp. 63–75, Jul. 2015.

[28] S. M. Mahmoud, A. Lotfi, and C. Langensiepen, "User activities outliers detection; integration of statistical and computational intelligence techniques," *Comput. Intell.*, vol. 32, no. 1, pp. 49–71, 2016.

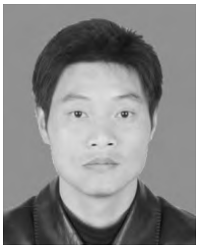
[29] T. Galibus *et al.*, "Offline mode for corporate mobile client security architecture," *Mobile Netw. Appl.*, vol. 22, no. 4, pp. 743–759, 2017.

[30] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: A survey and empirical demonstration," *Data Mining Knowl. Discovery*, vol. 7, no. 4, pp. 349–371, 2003.

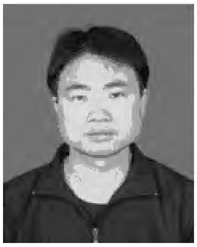
[31] G. Neelima and S. Rodda, "Predicting user behavior through sessions using the Web log mining," in *Proc. IEEE Int. Conf. Adv. Hum. Mach. Interact. (HMI)*, Bengaluru, India, Apr. 2016, pp. 1–5.

[32] M. H. Chehreghani and M. Bruynooghe, "Mining rooted ordered trees under subtree homeomorphism," *Data Mining Knowl. Discovery*, vol. 30, no. 5, pp. 1249–1272, 2016.

RUIJUAN ZHENG received the Dr.Eng. degree in computer application from Harbin Engineering University, Harbin, China, in 2008. She has been a Professor with the Henan University of Science and Technology since 2018. Her research interests include mobile cloud computing, bio-inspired networks, Internet of Things, future Internet, and computer security.



JUNLONG ZHU was born in Shaoyang, Hunan, China, in 1982. He received the Ph.D. degree from the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. He is currently with the Henan University of Science and Technology. His research interests include large-scale optimization theory and its applications.



MINGCHUAN ZHANG was born in Luoyang, Henan, China, in 1977. He received the Dr.Eng. degree in computer application from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014. He has been an Associate Professor with the Henan University of Science and Technology since 2005. His research interests include ad hoc network, Internet of Things, cognitive network, and future Internet technology.

QINGTAO WU was born in Jiujiang, Jiangxi, China, in 1975. He received the Dr.Eng. degree in computer application from the East China University of Science and Technology, Shanghai, China, in 2006. He has been a Professor with the Henan University of Science and Technology since 2006. His research interests include component technology, computer security, and future Internet security. He holds a membership of China Computer Federation.

RUOSHUI LIU (M'07) was born in Zhengzhou, Hunan, China, in 1983. He received the M.Eng. degree (Hons.) from the University of York in 2007 and the Ph.D. degree in computer science from the University of Cambridge in 2012. His research interests include wireless sensor networks, smart health, machine learning, and cloud computing. He is a member of the ACM.

KANG LIU was born in Nanyang, Henan, China, in 1992. He is currently pursuing the degree in computer application technology with the Henan University of Science and Technology, Henan. His research interests include mobile Internet and dependability management.

JING CHEN was born in Luoyang, Henan, China, in 1990. She is currently pursuing the degree in computer application technology with the Henan University of Science and Technology, Henan. Her research interests include mobile Internet and dependability management.