Hindawi
Security and Communication Networks
Volume 2018, Article ID 6459326, 14 pages
https://doi.org/10.1155/2018/6459326

Research Article

Flow Correlation Degree Optimization Driven Random Forest for Detecting DDoS Attacks in Cloud Computing

Jieren Cheng,[1,2,3] **Mengyang Li**,[1,2] **Xiangyan Tang**,[1,2] **Victor S. Sheng**,[4] **Yifu Liu**,[1,2] **and Wei Guo**[1,2]

[1]*Key Laboratory of Internet Information Retrieval of Hainan Province, Hainan University, Haikou 570228, China*
[2]*College of Information Science and Technology, Hainan University, Haikou 570228, China*
[3]*State Key Laboratory of Marine Resource Utilization in South China Sea, Haikou 570228, China*
[4]*Department of Computer Science, University of Central Arkansas, Conway, AR 72035, USA*

Correspondence should be addressed to Mengyang Li; 1098743772@qq.com

Received 23 August 2018; Accepted 1 November 2018; Published 19 November 2018

Guest Editor: Lianyong Qi

Distributed denial-of-service (DDoS) has caused major damage to cloud computing, and the false- and missing-alarm rates of existing DDoS attack-detection methods are relatively high in cloud environment. In this paper, we propose a DDoS attack-detection method with enhanced random forest (RF) optimized by genetic algorithm based on flow correlation degree (FCD) feature. We define the FCD feature according to the asymmetric and semidirectivity interaction characteristics and use the two-tuples FCD feature consisting of packet-statistical degree (PSD) and semidirectivity interaction abnormality (SDIA) to describe the features of attack flow and normal flow. Then we use a genetic algorithm based on the FCD feature sequences to optimize two key parameters of the decision tree in the RF: the maximum number of decision trees and the maximum depth of every single decision tree. We apply the trained RF model with optimized parameters to generate the classifier to be used for DDoS attack-detection. The experiment shows that the proposed method can effectively detect DDoS attacks in cloud environment with a higher accuracy rate and lower false- and missing-alarm rates compared to existing DDoS attack-detection methods.

1. Introduction

Cloud computing is a powerful technology to perform massive-scale and complex computing in which a huge amount of storage, data, and services is available over the Internet. Cloud services are distributed in nature so they can be sharable by millions of users, so that the cloud environment has to face numerous security challenges; in particular, distributed denial-of-service (DDoS) is one of the most prominent security attack in cloud computing. In recent years, DDoS attacks are on rise in frequency and severity in cloud computing and have become a growing problem because automated tools have been continuously improved and botnets of computers can be easily rented and organized to launch attacks by less sophisticated attackers [1, 2].

A DDoS trend and analysis report [3] shows that the average global enterprise encounters 237 DDoS attacks each month, which is equivalent to eight attacks per day. The main purpose of attackers is to force enterprise system servers unavailable or steal sensitive data. At the same time, the average number of DDoS incidents that global companies have experienced every month (Q3 2017) has increased by 35%. The scale and harm of DDoS attacks are increasing by leaps and bounds. Various forms of flooding and vulnerability attacks still affect and destroy networks and services. What is more, the Internet of things (IoT), industry 4.0, smart cities, and novel artificial-intelligence (AI) applications that require devices to be connected to cloud platforms provide an increasing wide range of potential botnet zombies, and the issue of controlling these botnets to launch DDoS attacks has become increasingly severe and important in cloud computing environment. Research in this area is important and significant.

Through the above analysis, we can understand the necessity of a DDoS attack-detection method. This paper seeks a better feature for attack-detection and a relatively

accurate and stable random forest (RF) attack-detection model by experiments and analysis. The organization of this paper is as follows. Section 2 introduces related work. In Section 3, we analyze attack characteristics and flow correlation degree (FCD) features. Section 4 introduces a random forest detection model based on genetic algorithm optimization. Section 5 introduces our experiments and their results. We provide our conclusions and ideas for future work in Section 6.

2. Related Research

Much research has been dedicated to DDoS attack-detection technology. Soft computing or artificial-intelligence methods are widely used in attack-detection [4]. Depending on the analysis method, DDoS detection methods can be classified into the three types of misuse, anomaly-based, and hybrid detection.

(1) Feature-based detection is also known as misuse, pattern, knowledge-based, and rule-based detection. This approach captures the required behavior from available datasets (such as protocol provisions and network-traffic events) and collects information about various attacks and system risks. This type of method uses the signature or mode of an attack, and such information as the index of the source IP address, destination IP address, and key of the port and packet payload in the IP packet. It matches incoming traffic to a stored pattern to identify an attack instance. IDES and INBOUNDS [5] are both signature-based detection methods. In recent years, new research has been conducted. Zhou et al. [6] proposed a DDoS attack-detection method which distinguished the constant attacks and the pulsing attacks from normal traffic by using the expectation of packet size. However, this method relies excessively on packet size and cannot adapt to multiple attack scenarios. Dodig et al. [7] proposed a new data structure based on a novel Dual Counting Bloom Filter to reduce detection errors for matching packages and theoretically analyzed the detection probability of determining the error rate and the requirement of increasing memory.

(2) Detection methods based on anomalies (also known as outliers and performance-based) can detect new types of attacks and unknown or emerging (undefined) attacks. When the difference between observed and expected behavior exceeds a predefined threshold, the detection system will generate an alarm. This method uses statistical methods, data mining, artificial intelligence, information theory, K-nearest neighbor, and other methods to identify anomalies in network traffic. Bhuyan et al. [8] proposed a scheme for DDoS flooding attack-detection and IP traceback by measuring the metric difference between the lightweight extended entropy of normal flow and attack flow. Latif et al. [9] proposed an enhanced decision tree algorithm based on a lightweight iterative pruning technique to detect DDoS attack and evaluated the performance of the proposed algorithm from classification accuracy, time, and space complexity, but the algorithm displays some defects in robustness due to flaws in decision tree classifier.

(3) Hybrid-based DDoS attack-detection combines two or more of the above strategies. A hybrid model can analyze common system behavior and inappropriate attacker behavior to improve the monitoring capabilities of the detection system. If hybrid system has both detection technology based on anomalies and features, the hybrid system can handle familiar and anonymous attacks and has characteristics of two detection methods, such as a high detection rate and low false-alarm rate [10]. Feature-based systems use anomaly-based techniques to detect attackers who try to change the attack patterns in the stored signature database. In recent years, some researchers have conducted extensive research on hybrid detection techniques. Gu et al. [11] presented a semisupervised clustering detection method using multiple features to solve the problems of large amount of unlabeled data in supervised learning, low detection accuracy and slow convergence speed of unsupervised learning. Liu et al. [12] proposed a DDoS attack-detection method based on conditional random fields, in which two sets of traffic feature conditional entropy (TFCE) and behavior profile deviate degree (BPDD) were depicted the characteristics of DDoS attacks. However, the training convergence speed of this method is slow. Bojović et al. [13] proposed a DDoS attack-detection method based on an exponential moving average algorithm. However, this method cannot detect attacks well when the packet forwarding rate of attack traffic is small.

Recent DDoS attack-detection methods have tended to be hybrid methods using a combination of multimode and multipart detection in the expectation of better performance. At the same time, the advent of the cloud computing era has seen increased security analysis and strategic research in these related realms. For example, research on providing reliable, stable, efficient, and secure services as well as data to the users of cloud computing [14–21], research on security strategies and privacy protection on the IoT [22–29], research on efficient cryptography to improve system security [30–32], and research on data processing, feature extraction, and information protection by machine learning method [33, 34] are all continuously deepened. There is also more research related to machine learning and integrated learning, combining attack features or optimization algorithms with time-series, ensemble learning, and deep-learning methods for network security analysis and traffic analysis. Intrusion-detection and attack-detection can improve detection results and speed. Cheng et al. [35] proposed a prediction approach based on abnormal network flow feature sequence to solve the problems of long response time and large computing resources of a DDoS attack detector in the big-data environment. However, this method requires relatively high stability for time-series data. Jia et al. [36] proposed a hybrid heterogeneous multiclassifier ensemble learning method to detect DDoS attacks, and constructed a heuristic detection system based on singular value decomposition, but the computational efficiency of this system may be low.

In general, the false- and missing-alarm rates of existing DDoS attack-detection methods are still relatively high in a cloud computing environment. In response to the problem, this paper analyzes network traffic, proposes a flow correlation degree feature, applies a random forest detection model,

optimizes its parameters to accurately and effectively detect DDoS attacks, and conducts research on attack characteristics and algorithm detection-performance optimization.

3. DDoS Attack Feature Extraction

3.1. DDoS Attack Feature Analysis. DDoS attack features have an important impact on attack-detection results. A feature that can effectively and steadily reflect DDoS attacks has a significant improvement in detection. Generally, DDoS attack features are extracted by describing the current network state through certain parameters or by observing changes in network parameter values, such as IP addresses, ports, payloads, and sizes of IP packets. The following two points are drawn from the consideration of cloud computing environments, as well as a great deal of research on feature extraction of DDoS attacks [37–39].

(1) The net source address and destination address, source address and destination port, and destination port and destination address all have a "many to one" relationship resulting in attacks that present the characteristics of flow asymmetry. Currently, many flooding attacks rely on botnets to attack target hosts or networks, forming a many to one attack mode to expand the scope of attacks and increase the harm of attacks, which can restrict or even paralyze them. At the same time, attacks can be more targeted, resulting in a certain service in the target network that cannot be used normally. Furthermore, system resources are attacked on multiple ports, so that multiple services cannot be used normally. Attacks can present a large amount of flow asymmetry.

(2) The network flows in direct or reflected DDoS attacks have higher semidirectivity interaction. In addition to flooding attacks, for an open shared-resource platform that lacks source IP address authentication or authentication capability of the packet source, the attacker uses packet source IP spoofing to attack. Using existing tools, numerous fake IP data packets are sent to the target network or host, causing abnormal or degraded network service. Most of the normal traffic at the monitoring point will respond to the destination and destination-to-source addresses. A large number of attacks will seriously affect the interaction. Therefore, the source IP address cannot receive a valid reply from the destination IP address. That is, the attack will greatly increase semidirectivity interaction of the network. Therefore, based on flow asymmetry and semidirectivity interaction characteristics, we propose the following feature extraction process.

3.2. Feature Extraction Rules. Assume that, within a unit time T, the net flow F is $< (t_1, s_1, d_1, dp_1), \ldots, (t_i, s_i, d_i, dp_i), \ldots, (t_n, s_n, d_n, dp_n) >$. Among them, $i = 1, 2, \ldots, n$, t_i, s_i, d_i, dp_i, represent the time of the i-th packet, source IP address, destination IP address, and destination port number. To classify these n packets, we use the following rules:

(1) Packets with the same source and destination IP addresses are grouped in the same category. All data with the source IP address A_m and the destination IP address A_n are marked. The packet formation class is $SDIP(A_m, A_n)$. For the above formed classes, execute the following deletion rule.

If there are different destination IP addresses A_n and A_k, ensure that the classes $SDIP\ (A_m,\ A_n)$ and $SDIP\ (A_m, A_k)$ are not empty, and delete all the classes whose source IP address is A_m.

Assume that the last remaining classes are $RSD_1, \ldots, RSD_m$, which define the packet-statistical degree (PSD) of the network flow F as

$$PSD_F = \sum_{i=1}^{m} W\left(RSD_i\right). \tag{1}$$

where $W(RSD_i) = \alpha Port(RSD_i) + (1 - \alpha) Packet(RSD_i)$, $(0 < \theta < 1)$, $Port(RSD_i)$ is the number of different port numbers of class RSD_i, $Packet(RSD_i)$ is the number of packets in the class of RSD_i, and α is the weighted value. In general, $\alpha = 0.5$.

(2) Classifying the n packets, separate data packets from the same source and destination IP addresses in the same class. $SIPC\ (A_m)$ represents the class of data packets with source IP address A_m. $DIPC\ (A_n)$ represents the class of data packets with destination IP address A_n.

If the source IP address A_m of class $SIPC\ (A_m)$ causes $DIPC\ (A_m)$ to be NULL, we define all of the data packets as source semidirectivity interaction flow and mark them as $SOH(A_m)$. This respects the property of source semidirectivity interaction, and we mark the different port numbers as $Port(SOH(A_m))$.

According to the above definition of source semidirectivity interaction, we obtain all the source semidirectivity interaction flow SOHs, expressed as $SOH_1, \ldots, SOH_s$.

Classifying the flow of SOH, we place the SOHs with the same destination IP in the same class marked as $SDH\ (Mton_m, A_m), m = 1, 2, \ldots, l$, l represents the amount of the destination IP address in SOH flow. The number of SOH flows with different source IP addresses and the same destination IP address is marked as $Mton_m$.

Suppose $Mton_m \geq M (M \geq 2$, where a greater value of M signifies a better effect of removing normal flow interference. To improve the coverage of attack-detection, we define $M = 2$. If we have SDH class asSDH_1, SDH_2,..., SDH_k, the number of destination port numbers in a class is expressed as $Port(SDH_i), i = 1, 2, \ldots, k$.

Semidirectivity interaction abnormality (SDIA) of the network flow F is defined as

$$SDIA_F = \frac{1}{f(k)} \left(\sum_{i=1}^{k} \left(Mton_i + weight\left(Port\left(SDH_i\right)\right)\right) - k \right). \tag{2}$$

Here, $f(x) = \{x, x > 1; 1, x \leq 1\}$, $weight(x) = \{x, x/\Delta t > \theta_1; 0, x/\Delta t \leq \theta_1\}$, Δt is the sampling-time period, θ_1 is weighted thresholds for the number of different destination ports, and $\theta_1 = \max(Port(SDH_i))\ /\Delta t, i = 1, 2, \ldots, k$. One can also specify a threshold based on experience.

(3)Combined with the feature extraction rule of (1) and (2), in a unit time T, two features of PSD and SDIA are

calculated and extracted, respectively, and a two-tuple feature is structured from these two features of PSD and SDIA to generate the network flow correlation degree (FCD) feature of the network flow F; we compute

$$FCD_F = \left(PSD_F, SDIA_F\right). \tag{3}$$

Normal network flow and DDoS attack flow in large data environment have the characteristics of high capacity, diversity, and burst, but FCD feature can still reflect the essential difference between normal flow and attack flow. First, the two parts in FCD feature are both extracted based on the asymmetry of DDoS attacks, and the FCD eigenvalues in attack cases are significantly larger than those in normal cases and last longer. Second, PSD features extraction is the weighted statistical features of the source IP address and port of the network flows of the "many to one" and "one to one" session mode, which eliminates the interference the network flows of "one-to-multi" session mode and reflects the correlation between attack flow and normal flow in the network more clearly. However, what the SDIA feature extracts is the weighted statistical information of the one-way flows of the "many to one" session mode in the network flow, which can more accurately describe the dramatic increase of the one-way flow when the network is attacked by DDoS attack. The combination of these two pieces of statistical information can accurately describe the phenomenon that attack flows converge at the injured end and directly affect the change of normal traffic and that a partially converged attack flow is mixed with a large amount of normal flow. This feature can present the higher source address distribution, destination address concentration, source destination IP address asymmetry, and high-traffic bursts for DDoS attacks in cloud computing environment, which provides more accurate, timely, and complete information about the network before and after the attack.

4. Implementation of DDoS Attack-Detection Method Based on Random Forest and FCD

4.1. FCD Feature Sequence Extraction. According to the rule described above in Section 3.2, the data of net flow are sampled by time interval, and the values of PSD and SDIA in each sampling-time are calculated and integrated into a two-element combination. After N samples, FCD time-series sample M is obtained, $M(N, \Delta t) = \{FCD_i, i = 1, 2, \ldots, N\}$, where N is the sequence length. With the accumulation of sampling-time Δt, the sequence is a time-characteristic sequence with a time length of N. Based on the FCD feature sequence extracted above, we can construct a RF classifier to detect DDoS attacks.

4.2. Random Forest. Random forest is a classification method of integrated learning. In the training process, it can use a resampling technique (bootstrap method) in which each sample returned from the original training data is randomly selected from the same number of samples, consisting of a new training dataset, and multiple decision trees are independently generated. In each decision tree, according to some evaluation criteria like the information entropy and Gini coefficient, the selection of the best test from the new training dataset is used as the decision point to carry on the split test, and then the result of the single decision tree is produced; the final decision is formed by calculating the mode of classification results of all decision trees. A formal description is given below.

Suppose the whole RF classifier is $R(x)$; decision tree i is denoted as $t(x)$, $R(x) = \{t_i(x), i \in [0, n\_estimators]\}$, where $n\_estimators$ represents the number of decision trees in the RF, x is the input training sample to be classified, and $\text{sign}(x) \in S$ is the tag value of x, in which S is the set of labeled categories, the output of the $t_i(x)$ is a certain value in S, and the output of the $R(x)$ is the mode of the estimated value of $\{t_i(x), i \in [0, n\_estimators]\}$. In the use of RF for testing, x is the value of the new training dataset randomly generated by resampling technology in the FCD feature training set; there are only two kinds of labels in DDoS attack-detection, which represent abnormal and normal. Therefore, $S = \{-1, 1\}$, and $\text{sign}(x)$ can only take the value -1 or 1 to represent the attack sample labels and normal sample labels, respectively.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2. \tag{4}$$

In this paper, the Gini coefficient is selected as the quantitative evaluation criterion of the single-decision tree division, as specified in formula (4). In this equation, D represents the sample space of n samples and k categories, and p_i represents the proportion of the i samples of the entire sample. When used in a specific experiment, D is a sample space constructed for the set of feature datasets for training, where $k = 2$, and n is the size of the training sample. The Gini coefficient represents the impurity of the training model. The smaller the value, the lower the purity and the better the characteristics. In addition, the main reason for the use of Gini coefficients as splitting indices of decision trees in the RF is that the coefficient cannot only reflect the proportion of all categories of samples and different types of sample proportion changes but can also make their values meet between $(0, 1)$, to facilitate the processing analysis.

According to the information above, the RF-detection model is constructed based on the FCD feature sequence. In the construction process, a genetic algorithm is selected to optimize and determine the number of decision trees and the maximum depth of the single decision tree in the RF. The process is introduced in Section 4.3.

4.3. Genetic Algorithm Optimization of Random Forest. The genetic algorithm is based on Darwin's biological theory of evolution. We search for the optimal solution by simulating the process of natural evolution in a certain range of solution sets. The solution set most in accordance with the "survival of the fittest" principle as in generational evolution is the approximate optimal solution. As a global optimization probability algorithm, a genetic algorithm can guarantee effectiveness in a large dataset using a heuristic method, and it can search the optimal solutions of all problems in any sense of form and function in a global sense. Therefore, the

range of key parameters in RF can be determined based on empirical values and mathematical derivation. In a relatively simple way, a genetic algorithm is used to select more reliable detection parameters.

In the process of constructing RFs, the parameters in a forest, such as the number of producing subdecision trees, the number of random attributes, and the maximum depth of trees, will affect the final classification results. Whether the number of subdecision trees selected is appropriate for the training results of a RF has a critical impact. Too small a number will lead to inadequate training, which cannot produce good results, while too large a number will lead to a long construction time and overly complex RF. A single-decision-tree depth also has a great impact on the training results and training time. The appropriate depth can guarantee the subtree of the leaf node to have a more reasonable classification, and it also reduces the training time. Therefore, we choose two key parameters, the number of estimators (*n_estimators*) and the maximum depth (*max_depth*) of the subtree as the parameters to be optimized. The process is as followed.

(1) Choose the parameter-selection strategy and fitness function. Parameter selection includes the determination of the population size, the number of iterations, selection, crossover, and mutation probability. Fitness function is the basis of genetic variation of individuals and population evolution in genetic algorithm. Here, considering the impact of constructing RF and finding optimal parameters on the time of constructing classification model, the following parameter values and the ranges of parameters to be optimized are determined. Set that the initial population size is 10, the number of iterations is 30, the range of *n_estimators* in RF is (2, 30), the range of *max_depth* of the subdecision tree is (2, 8), and the mutation rate and cross rate are default. Considering the generality and reliability of the fitness function value, the average value of the area (area under curve, AUC) under the ROC curve in the cross validation of the training sample is selected as the fitness function value. The greater the value, the more conducive to the inheritance and evolution of the individual.

(2) Encode and initialize the population. The binary encoding method is used for coding. From a given set of two positive integer parameter ranges, the parameter combination (*n_estimators*, *max_depth*) is randomly selected and encoded as chromosome X = {*n_estimators*, *max_depth*}. The initial population G is randomly initialized by multiple individuals resulting from the crossover and mutation of the chromosome X. Binary coding of chromosomes can increase the likelihood of mutation and crossover, thus providing more diverse solutions.

(3) Evaluate the fitness value. According to the fitness function value mentioned above in (1), the fitness value of each individual population can be calculated, as shown in formula (5), in which K represents the fold number of cross validation, AUC is the area under the calculated ROC curve when the training sample is tested as a test sample in cross validation, and when this value is greater, the fitness value is better. Then the fitness values of each individual are calculated. By comparing the fitness values of individuals, those with the best fitness value are selected to generate the initial individuals of the next generation of the population, so as to carry out subsequent crossover and mutation operations.

$$Fitness = \frac{1}{K}\sum_{i=1}^{K} AUC_i. \tag{5}$$

(4) Judge terminating conditions. In the process of continuous iteration, it is judged whether the fitness meets the established standard. If it is not satisfied, then step 3 is repeatedly performed until the termination condition is reached. At this time, we select the individual with the largest fitness value in the population and extract the corresponding decimal values of binary-coded chromosome X in the individual as the optimal parameters of RF for training.

(5) Apply the optimal parameters. The optimal *n_estimators* and *max_depth* values are selected as the parameters of the RF for training, the RF classifier is trained based on the training set of FCD feature sequence and this two optimal parameters, and the DDoS attack-detection model based on genetic algorithm optimization and RF is constructed.

By optimizing the parameters above and constructing RF model, we can obtain an RF-detection model optimized by the genetic algorithm, which is more accurate than the general RF-detection model. Considering the heuristic searching ability of genetic algorithm, the combination of genetic algorithm and RF can effectively improve the classification ability of RF, so as to detect DDoS attacks more accurately and effectively.

4.4. Random Forest Detection Based on Genetic Algorithm Optimization. According to the above description in Section 4, we optimized the parameters based on the FCD feature, trained the RF classifier and obtained the genetic algorithm-optimized random forest (GAORF) based on the FCD feature sequence. In this paper, the DDoS attack-detection method with the model generated by FCD feature sequence and GAORF algorithm is referred to as FGRF attack-detection method. The process of the application of the method in this paper is shown in Figure 1.

An attack can be identified according to the model of the FGRF detection method trained to characterize the network state. The model actually solves the problem of binary classification in machine learning. The detection task can only identify an attack or not. Assuming that the detection model detects that net flow does not have feature anomalies during a certain period of time under normal conditions, we set the detection model output flag to 1. Under attack conditions, the FCD feature value will rise obviously with the time change, which is gradually higher than the normal value, and then we set the output flag of the detection model to another value, and we set it to -1 in this paper. These two settings can characterize whether the network is attacked or not. As the FGRF detection method is used to detect the real DDoS attack, after the FCD value of the net flow is entered into the model, the output flag returned by the model can reflect whether the network is attacked.

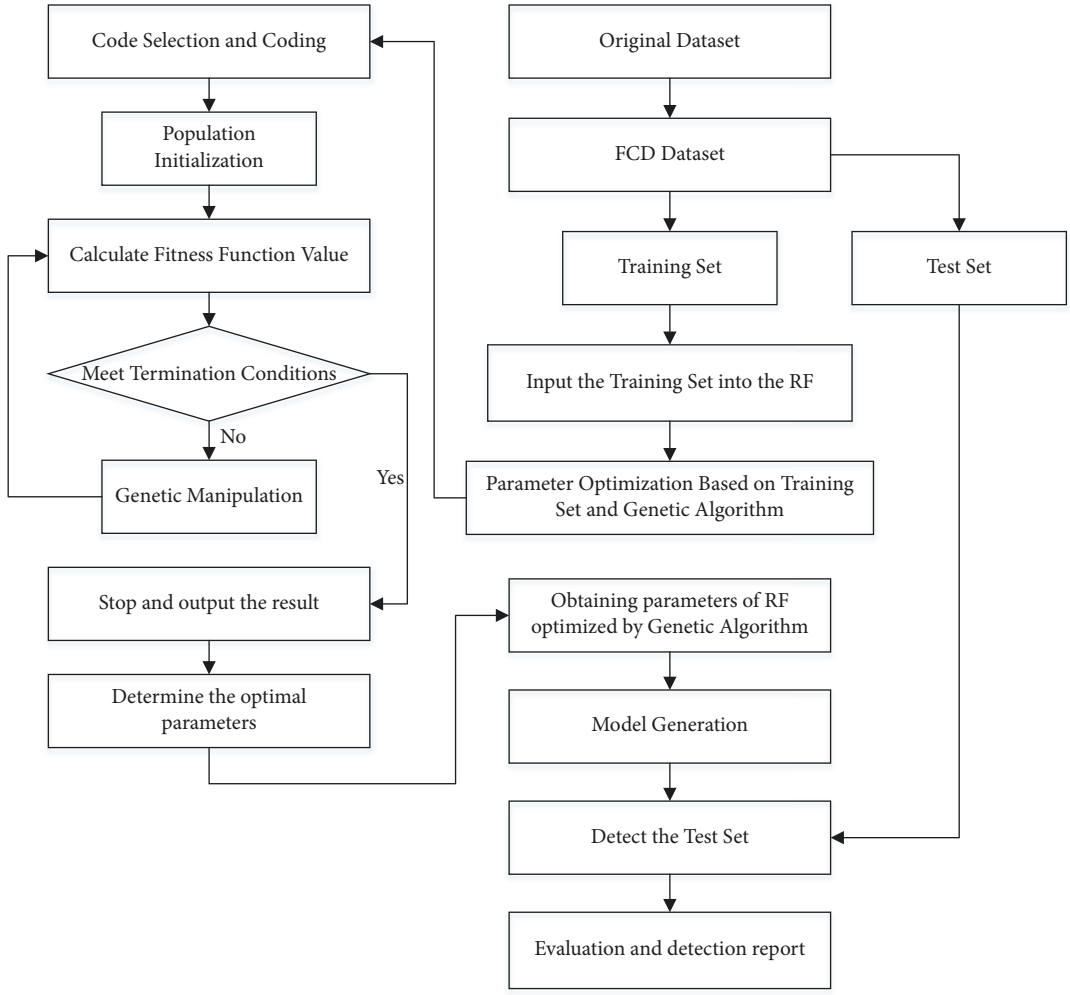

Figure 1: The process of FGRF DDoS attack-detection method.

The analysis in Section 3.2 of this paper shows that FCD feature sequence can better reflect the different state characteristics of normal flow and DDoS attack flow in cloud computing environment. Multiple decision trees are integrated in the RF, the bootstrap method is used to reduce the size of the single-decision-tree training sample set, and a more reasonable classification result is selected using the voting mechanism. The combination of these mechanisms in RF can improve the accuracy of detecting high-capacity traffic information in DDoS attacks under cloud computing environment. Moreover, the method based on genetic algorithm to optimize RF parameters effectively improves the classification ability of RF. Therefore, the FGRF attack-detection method proposed in this paper can effectively detect DDoS attacks under cloud computing environment.

5. Experiment

5.1. Data Set and Evaluation Criteria. The experimental hardware had 8G memory and an i7 processor. The experiment was carried out on a Windows 10 64-bit system running Python 3.5.2 |Anaconda 4.2.0 (64 bits).

The experiment was based on the dataset of the CAIDA DDoS attack in 2007 [40]. It contained data on an anonymous DDoS attack that lasted for about an hour on August 4, 2007. This type of attack attempts to prevent access to target servers by consuming computing resources on servers and all the bandwidth of connecting servers to Internet networks. The total size of the dataset was 21 GB, accounting for about an hour (20:50:08 UTC -21:56:16 UTC). The attack started at about 21:13 and caused the rapid growth of the network load (in a few minutes) from about 200,000 bits/sec to 80 MB/sec. The attack traffic was divided into five-minute files and stored in PCAP format.

To judge the validity of attack-detection, some evaluation criteria were used to fully illustrate the performance of the test, including the accuracy rate, missing-alarm rate (MR), and false-alarm rate (FR). Suppose TP is the number of normal samples marked correctly, TN is the number of attack samples that are correctly marked, FN is the number of attack samples marked in error, and FP is the number of normal samples marked in error.

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}. \quad (6)$$

$$FR = \frac{FP}{TP + FP}. \quad (7)$$

$$MR = \frac{FN}{TN + FN}. \quad (8)$$

The accuracy rate is the proportion of the correctly identified samples in all samples; the false-alarm rate is the proportion of samples judged to be attacked in the normal sample, and the missing-alarm rate is the proportion of the sample that is not successfully identified. Then TN/TN + FN is the detection rate. Through the environment, datasets, and evaluation criteria described above, experiments are carried out according to the process described in Section 4. FCD feature sequences are extracted from the data sets described in Section 5.1, and all normal samples are labeled as 1 and all attack samples are labeled as -1 according to Section 4.4. Training and test sets are selected from the FCD feature sequences. The parameters of RF are optimized by genetic algorithm based on training set, and the model of the FGRF attack-detection method is established, and the performance of classification model is verified by test set. SVM algorithm is more classic and has better classification results because of its use of the mechanism of hyperplane classification. In order to better illustrate the good performance of the FGRF attack-detection method proposed in this paper, the model is compared with several detection models generated by a variety of SVMs which is trained based on the FCD feature sequences. The scikit-learn [41] toolkit was used to complete the implementation of RF and GAORF. The LIBSVM [42] toolkit was used to complete the contrast test in SVMs. The experimental process and its results are introduced in Section 5.2.

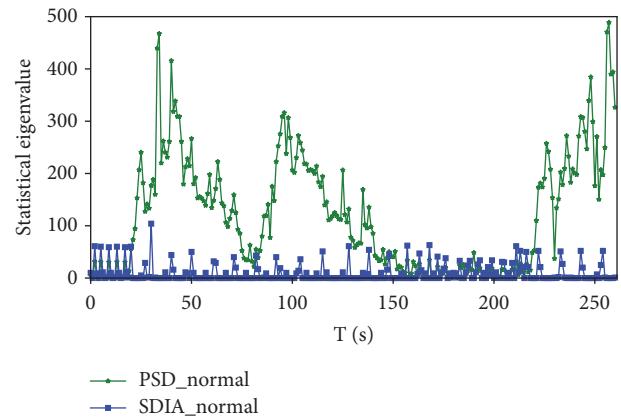

Figure 2: Comparison of PSD and SDIA features in normal flow.

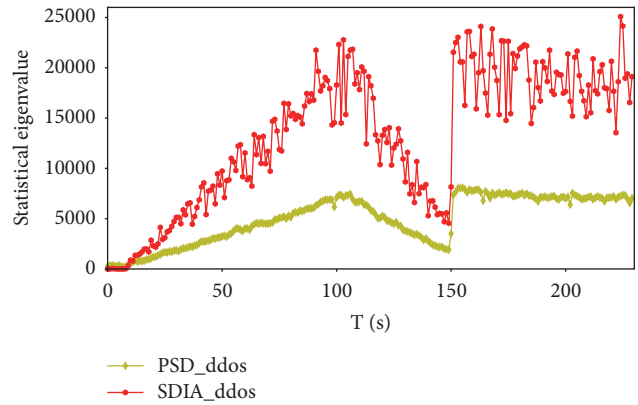

Figure 3: Comparison of PSD and SDIA features in attack flow.

5.2. Experimental Data Analysis. We obtained a normal data sample from ddostrace.20070804_134936.pcap and an attack data sample from ddostrace.20070804_141436.pcap in the DDoS Attack 2007 dataset. According to the feature extraction rules in Section 3.2 and the feature sequence extraction method in Section 4.1, FCD feature sequence was extracted from normal and attack samples. For convenience of calculation and processing, we set $\Delta t = 1$s as the sampling interval. The parameters of PSD and SDIA in FCD feature were set according to Section 3 and the FCD value time-series sample M as shown in Figures 2 and 3.

As shown in Figure 2, in the normal flow, the sequence of the PSD eigenvalues shows a stronger volatility, and the highest feature value can reach about 500, while the sequence of the SDIA eigenvalues is relatively stable and their values are floating within the range of 150. The PSD feature statistics are the characteristic information of the network flows of "one to one" and "many to one" session mode, and because of network congestion, similar network flows are more common in normal flow, so the values of PSD features will fluctuate in a certain range, which can better reflect the abnormal changes of normal flow state caused by attack flow than SDIA features. The SDIA feature statistics are the characteristic information of the one-way flows of the "multi to one" session mode. In the normal network, the one-way flows are relatively less than the bidirectional flows, so the sequence of SDIA eigenvalues is more stable.

As shown in Figure 3, both PSD eigenvalues and SDIA eigenvalues increase with the increase of DDoS attack flow, but SDIA eigenvalues are relatively higher than PSD eigenvalues at the same time, The SDIA eigenvalue reaches a peak value of about 25000, while the PSD eigenvalue reaches a peak value of about 8000. Obviously, the change of the SDIA feature between them is more obvious. The one-way flows of the "many to one" session mode in the network will increase rapidly caused by DDoS attacks; both PSD and SDIA feature have weighted the information of the one-way flows of the "many to one" session mode, so their values will increase and can reflect the attack state to a certain extent. In addition, the two eigenvalues in Figure 3 show a sudden decrease and then continue to increase, which is caused by the decrease of the one-way flows of the "many to one" session mode in the network caused by such factors as the delay of attack at that time. The SDIA feature, which is different from PSD feature in the weighted calculation method, more centrally describes the related information about the one-way flows of the "many to one" session mode, so it can describe attack flow more accurately than PSD feature. It can well reflect the semidirectivity interaction of large-area network flow caused by DDoS attacks.

The combination of PSD feature and SDIA feature is the FCD feature proposed in this paper. This feature can integrate the advantages of the two features, not only can describe the attack flow well, but also can reflect the abnormal changes of the normal flow state caused by the attack flow, so it can better identify the attack.

In the process of experiment, training and testing samples were selected first. To facilitate integration, calculation, and processing, 200 FCD features were selected as test datasets, which include 100 normal flow features and 100 attack flow features, respectively. In the rest of the features of FCD, 100 normal flow features and 120 attack flow features are selected as training samples. To study the negative sample, which is the attack sample, we made an appropriate increase under the restriction of the existing characteristic dataset, so as to obtain better training.

After selecting the training set and test set from the whole feature set, the data samples are normalized, and the genetic algorithm is used to optimize the RF model trained by the training set. Due to the small number of samples, it is still necessary to ensure reasonable and effective testing. Therefore, $K = 2$ in formula (5) are used when evaluating the fitness value in Section 4.3.

In the experiment of optimizing the parameters, the number of training samples is small, and the initial population size is large. Considering the good classification performance of the RF algorithm itself, a good parameter-solution set will be found quickly within the specified number of iterations. In addition, the properties of the genetic algorithm in random search of the optimal parameter-solution set in the prescribed range also increase the possibility of producing better results. Therefore, the combination of genetic algorithm and RF algorithm can find the approximate optimal solution of these parameters to a large extent in the global scope.

In the end, after the 30 iterations we set in Section 4.3, a relatively high-quality parameter-solution set was determined based on the training sample set, that is, the value of the two optimal parameters of the number of subtrees and the maximum depth of subtree. These two parameter values were brought into the RF model for training, and a classification model was generated for detection. Finally, the results described in Section 5.1 were used to judge the test results. To make the results more effective and reliable, we conducted comparative experiments. The experimental results are introduced in Section 5.3.

5.3. Experiments and Results. To verify the detection capabilities of our proposed FCD feature combined with the detection model constructed by the RF algorithm and the genetic algorithm, we performed comparison experiments, and the specific steps and the results of the comparison experiment are as follows:

(1) In accordance with the description in Section 5.2, the training data set and test data set are selected. Here, the test data set is kept unchanged, and the following two change operations were performed on the normal sample and the attack sample in the training set to perform comparison experiments: the number of fixed attack training samples was 120, and the number of normal training samples was increased to 100 on the basis of 10 normal training samples, in order to simulate the change of normal flow in network caused by the delay of DDoS attack and other factors; the number of fixed normal training samples was 100, and the number of attack training samples was increased to 120 on the basis of 10 attack training samples to simulate the situation that the normal network is gradually starting to be attacked by DDoS attacks, resulting in a gradual increase in the attack flow. The different training samples were applied to train each model to detect the same test set, and the final test results were obtained.

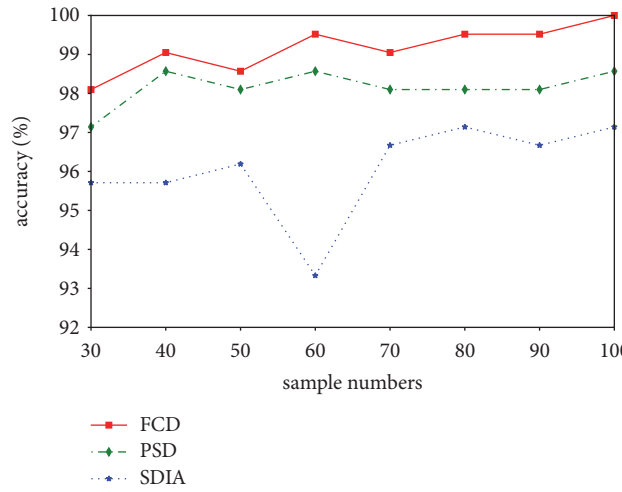

Figure 4: Accuracy comparison results of three statistical features with changing numbers of normal training samples (%).

(2) In order to further verify the good performance of the FCD feature proposed in this paper for DDoS detection, the feature FCD was compared and analyzed with the PSD and SDIA features during the experimental operation (1) in Section 5.3. The PSD, SDIA, and FCD features are extracted from the same training samples, and three classifiers are generated based on three features training RF model respectively, and then the same test set is used to test the three classifiers in order to compare the ability of the three features to distinguish between normal flow and attack flow. With the number of fixed attack training samples, Figure 4 shows the accuracy rates obtained by changing the number of normal training samples, and Figure 5 shows the false- and missing-alarm rates obtained by changing the number of normal training samples. With the number of fixed normal training samples, Figure 6 shows the accuracy rates obtained by changing the number of attack training samples, and Figure 7 shows the false- and missing-alarm rates obtained by changing the number of attack training samples. Among them, FCD_MR, PSD_MR, and SDIA_MR are missing-alarm rates based on FCD feature, PSD feature, and SDIA feature, respectively. FCD_FR, PSD_FR, and SDIA_FR are false-alarm rates based on FCD feature, PSD feature, and SDIA feature, respectively.

As shown in Figures 4 and 5, all three features can better identify attack, among which the FCD feature is the best. Seen from the aspect of accuracy, with the increase of

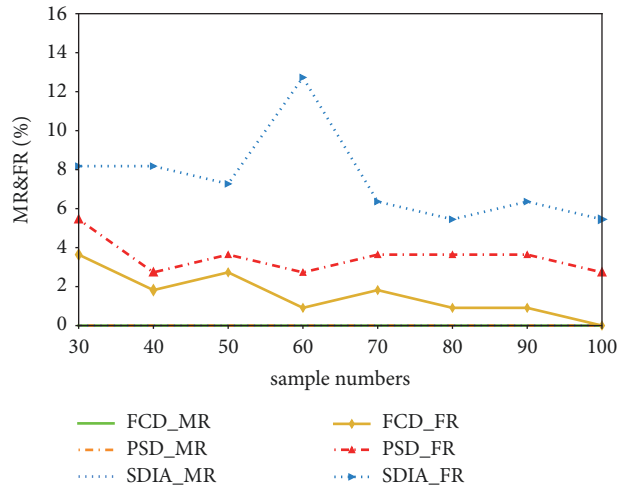

Figure 5: False-alarm rate and missing-alarm rate comparison results of three statistical features with changing numbers of normal training samples (%).

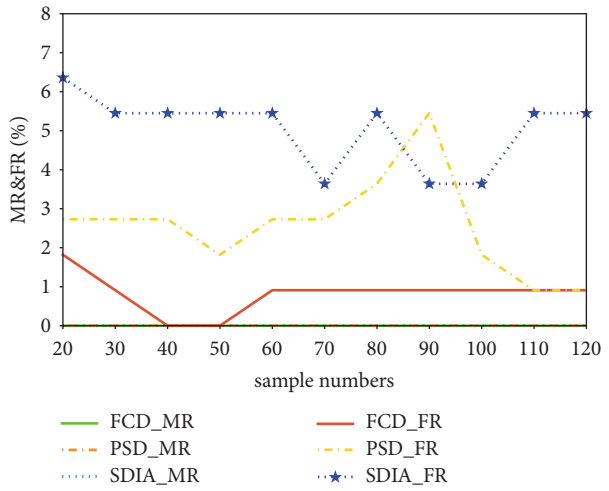

Figure 7: False-alarm rate and missing-alarm rate comparison results of three kinds of statistical feature with changing numbers of attack training samples (%).

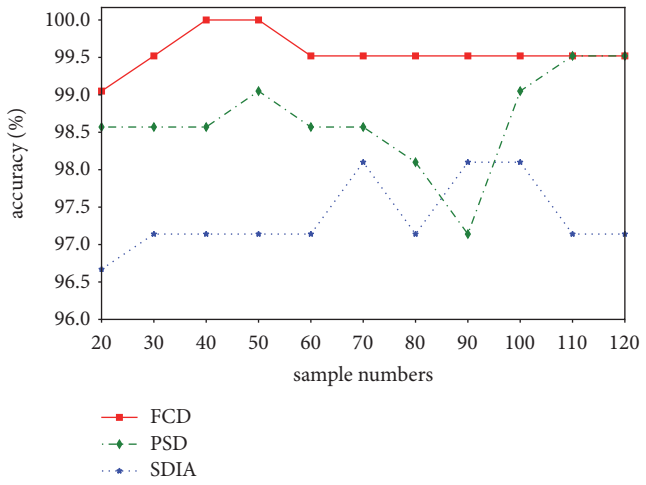

Figure 6: Accuracy comparison results of three statistical features with changing numbers of attack training samples (%).

normal samples, the accuracy rate based on FCD feature is the highest, which keeps above 98%, and increases to nearly 100%. The accuracy of PSD feature is also increased, but it is about 1% lower than that of FCD feature. As for the SDIA feature, the accuracy is kept below 97%. From the aspect of false- and missing-alarm rates, as the number of normal flow increases, FCD_MR, PSD_MR, and SDIA_MR are all zero. FCD_FR also tends to zero, although PSD_FR is down steadily to about 3%, and SDIA_FR decreases to 2% in fluctuation; they were still higher relative to the combined feature. Among them, when the number of normal samples is 60, the accuracy rate of SDIA feature decreased to about 93% and SDIA_FR increased to 13%. The accuracy rate of PSD features is about 5% higher than that of SDIA features, while the accuracy rate of FCD remains above 99%, and PSD_FR is about 10% lower than that of SDIA_FR, and FCD_FR is less than 2%. The PSD feature is the statistics of the network flows of the "many to one" and "one to one" session mode, including normal flow, its value will change with the increase of normal flow, that is, the PSD feature can better reflect the abnormal changes of normal flow state caused by attack flow, so PSD features maintain higher accuracy rate and lower false-alarm rate than the SDIA feature. The SDIA feature is the statistics of the one-way flows of the "many to one" session mode. It can describe attack characteristics more centrally, but cannot describe subtle changes of normal flow state better. Therefore, when the number of normal training samples is 60, the false-alarm rate of the detection suddenly increases, thus reducing the accuracy. FCD features contain two statistical information provided by PSD and SDIA features, so the accuracy rate of FCD features is higher, and the missing- and false-alarm rate are lower. Compared with FCD and SDIA features, FCD features can better identify DDoS attacks.

Figures 6 and 7 show that the FCD-based RF-detection method can maintain higher accuracy rate with low false- and missing-alarm rates compared to that based on PSD and SDIA features. When the attack flow increases, the detection based on FCD features has a high accuracy of up to 99% and low false- and missing-alarm rates below 2%. PSD_FR and SDIA_FR both fluctuate over 1%, resulting in low accuracy. FCD_MR, PSD_MR, and SDIA_MR, are all zero. When the number of attack samples increases to 90, the false-alarm rate of PSD features suddenly increases to more than 5%, which is about 1% higher than that of SDIA features, and its accuracy rate decreases to less than that of SDIA features. In this case, the accuracy rate of FCD feature still maintains accuracy above 99% and false-alarm rate about 1%. Among the above data analysis results, the detection results of the three characteristics are mainly reflected in the trend of false-alarm rate. The PSD feature can well reflect the abnormal changes of normal flow state caused by attack flow, so when the proportion of normal flow in the network is still large and attack flow changes little, PSD_FR is generally lower than SDIA_FR. However, the PSD and SDIA eigenvalues are

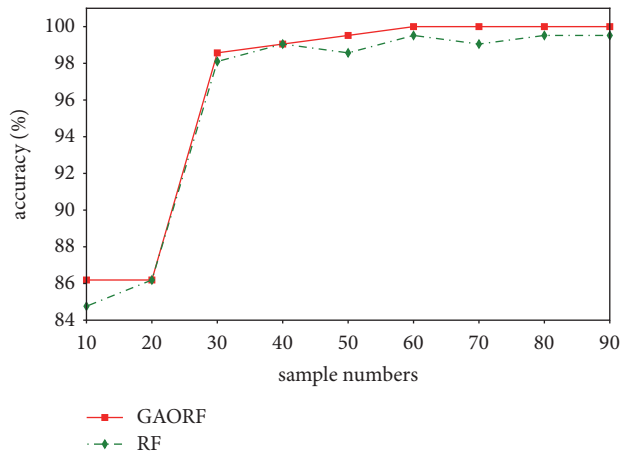

Figure 8: Comparison of accuracy rate between optimization and common model detection with changing numbers of normal training samples (%).

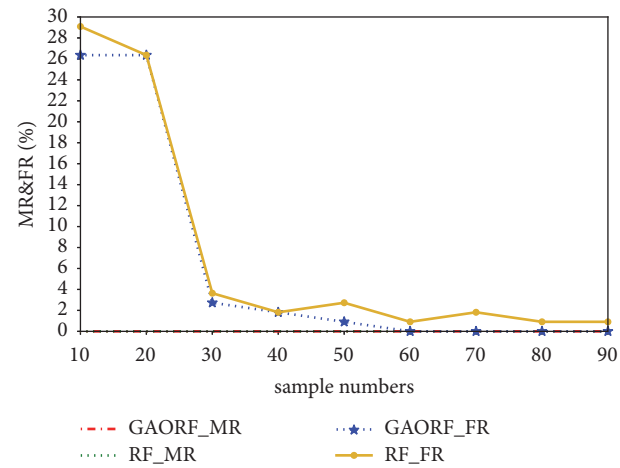

Figure 9: Comparison of false-alarm rate and missing-alarm rate between optimization and common model detection with changing numbers of normal training samples (%).

generally small in normal flow, the early attack traffic is generally small and the impact on normal flow is also small, so the PSD and SDIA eigenvalues change little in the early attack and are more likely to cause false- and missing-alarm rates. The SDIA feature is the statistics of the one-way flows of the "many to one" session mode, which can describe attack characteristics more centrally, Therefore, SDIA_FR will be significantly reduced when the early attack traffic is small or the attack is delayed, which results in a situation similar to that when the number of attack samples is 90. As for the FCD feature, it contains the information provided by the above two features, so the feature has better detection results and can better identify DDoS attacks.

(3) In order to further verify the validity of the genetic algorithm in optimizing the RF classification model, a comparison experiment was made between the RF classifier with parameter optimization by genetic algorithm and the RF classifier without parameter optimization based on the FCD feature sequences during the experimental operation (1) in Section 5.3. Two classifiers are generated based on FCD feature sequence training RF model and GAORF model respectively, and then the same test set is used to test the two classifiers in order to compare the classification ability of RF model and GAORF model.

Figure 8 shows the accuracy rates from the number of fixed attack training samples and the number of varied normal training samples. Figure 9 shows the false- and missing-alarm rate from the number of fixed attack training samples and the number of varied normal training samples. Figure 10 shows a comparison of accuracy rates from changing the number of attack training samples and fixing the number of normal training samples. Figure 11 shows a comparison of the false- and missing-alarm rates from changing the number of attack training samples and fixing the number of normal training samples. Here, GAORF_MR and RF_MR are the missing-alarm rate of GAORF detection and RF detection, respectively. GAORF_FR and RF_FR are the false-alarm rate of GAORF detection and RF detection, respectively.

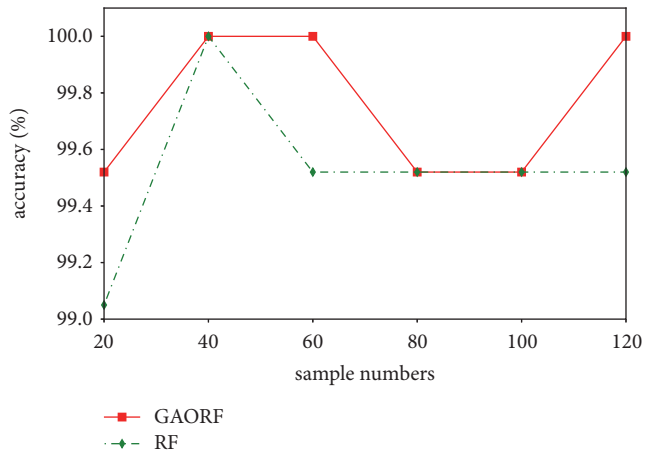

Figure 10: Comparison of accuracy rate between optimization and common model detection with changing numbers of attack training samples (%).

Combined with Figures 8 and 9, it can be seen that the accuracy rates of RF-detection model and GAORF detection model based on FCD feature sequences increases to a certain extent, and the false-alarm rates decrease gradually when the attack training sample is invariable and the normal training sample is increasing. The accuracy rate of GAORF detection model is about 2% higher and the false-alarm rate is about 2% lower. Because the heuristic parameter searching method of genetic algorithm can find better training parameters for RF classifier based on the correlation between normal flow and DDoS attack flow, which is shown by PSD features contained in FCD features, the classification performance of GAORF detection model is improved. It is worth considering that the parameter optimization process will also be constrained by the number of normal training samples, but the genetic algorithm can still find better training parameters for RF-detection model, so that can maintain the original better detection results.

Table 1: Comparison results of four algorithm detection evaluation criteria with changing numbers of normal training samples.

		Sample numbers			
		30	50	70	90
GAORF (%)	accuracy	98.57	99.52	100	100
	MR	0	0	0	0
	FR	2.72	0.91	0.0	0.0
nu-SVM (%)	accuracy	93.33	85.24	99.05	100
	MR	0	0	0	0
	FR	12.72	28.18	1.81	0
C-SVM (%)	accuracy	91.90	100	100	100
	MR	0	0	0	0
	FR	15.45	0	0	0
one-class-SVM (%)	accuracy	37.62	38.10	40.95	45.71
	MR	21.00	21.00	21.00	21.00
	FR	100	99.09	93.64	84.55

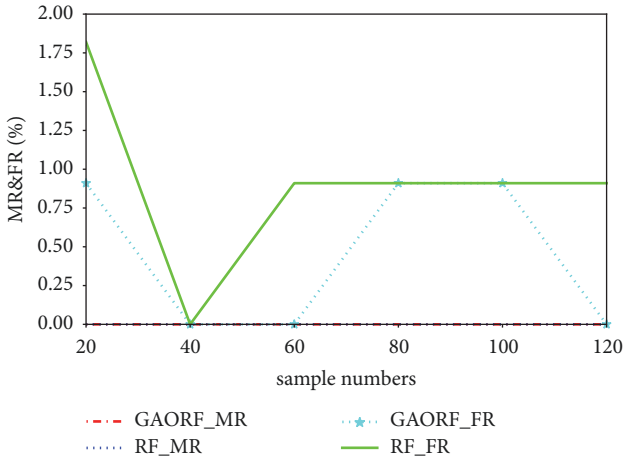

Figure 11: Comparison of false-alarm rate and missing-alarm rate between optimization and common model detection with changing numbers of attack training samples (%).

As shown in Figures 10 and 11, when the normal training samples remain unchanged and the attack training samples increase, the GAORF detection model has no missing-alarm and the false-alarm rate is about 1% lower than the RF-detection model; thus the overall accuracy rate is about 1%. Because the genetic algorithm can optimize the GAORF detection model based on the asymmetry and semidirectivity interaction characteristics of the attack flow described by the SDIA feature included in the FCD feature, the classification performance of the RF-detection model can be improved. Because the attack traffic in the early stage of DDoS attack has little influence on normal flow, the value of PSD and SDIA features in the FCD features in the early stage of DDoS attack is lower, thus affecting the detection results of the model. Genetic algorithm can still find better training parameters for the RF-detection model, so as to maintain better detection results. To sum up, using genetic algorithm to optimize the parameters of RF-detection model can effectively improve accuracy rate and reduce the false-alarm rate of DDoS attack detection.

(4) To further verify the good performance of the FGRF attack-detection method proposed in this paper, the GAORF classification model was compared with nu-SVM, C-SVM, and one-class-SVM classification models based on the FCD feature sequence during the experimental operation (1) in Section 5.3. Considering that SVM is a supervised learning algorithm with good classification performance and is widely used in previous research for DDoS attack detection, furthermore, nu-SVM, C-SVM, and one-class-SVM among SVM algorithms show stronger mode identification and classification ability; thus we chose these three SVM algorithms as the comparison algorithms. The FCD feature sequence was trained in the GAORF and three classical SVM classification methods, respectively, and then the same test set is used to test the four classifiers. We fixed the number of training samples in the attack flow and changed the number of training samples in the normal flow. The results are shown in Table 1. We fixed the number of training samples in the normal flow and changed the number of training samples in the attack flow. The results are shown in Table 2.

From Table 1, we can see that FCD combined with the GAORF detection method has higher accuracy and lower false-alarm and missing-alarm rates compared with three traditional SVM detection methods, especially when the number of normal training samples is relatively small. When the attack training samples remain unchanged, with the increase of normal training samples, the accuracy of GAORF detection model remains above 98% and the false-alarm rate remains below 3%. On the one hand, RF has a good and stable classification performance, which can be used to mine and utilize FCD features to represent the abnormal changes of normal flow state caused by attack. On the other hand, genetic algorithm optimizes RF parameters and improves RF classification ability by learning normal training sample set, so the classification effect of GAORF classification model is the best. The false-alarm rate of nu-SVM detection model fluctuates greatly, and the accuracy ranges from 85% to 100%.

Table 2: Comparison results of four algorithm detection evaluation criteria with changing numbers of attack training samples.

		Sample numbers			
		30	60	90	120
GAORF (%)	accuracy	100	100	100	100
	MR	0	0	0	0
	FR	0	0	0	0
nu-SVM (%)	accuracy	90.0	98.1	99.05	100
	MR	0	0	0	0
	FR	19.09	3.63	1.82	0
C-SVM (%)	accuracy	97.14	100	100	100
	MR	6.0	0	0	0
	FR	0	0	0	0
one-class-SVM (%)	accuracy	65.0	65.0	65.0	65.0
	MR	0	0	0	0
	FR	70.0	70.0	70.0	70.0

The training set contains some data with lower attack eigenvalues in the early stage of the attack, and these eigenvalues are similar to the normal flow eigenvalues; it is difficult to distinguish normal samples in the classification hyperplane of nu-SVM model, thus affecting the detection results. We can see that C-SVM detection model has no classification error when the number of normal training samples is more than 50, while the false-alarm rate is about 15% when the number of normal samples is 30. As the penalty coefficient of C-SVM does not change due to the excessive increase of normal training samples, the model shows good stability. However, when the number of normal samples is small, the model is difficult to obtain the optimal classification hyperplane, resulting in a sudden increase in false-alarm rate. For the one-class-SVM detection model, the detection of this model keeps the accuracy rate under 50%, the higher false-alarm rate and the false-alarm rate. The reason is that one-class-SVM can only train normal training samples to generate classification model, which makes it more difficult to recognize attacks. Therefore, it is difficult to achieve a more ideal classification effect.

As shown in Table 2, when the number of normal training samples is constant and the number of attack training samples increases, the GAORF detection method does not have a classification error, showing a better performance compared with the SVM detection methods. On the one hand, RF itself has good and stable classification performance and can better mine and utilize FCD features to characterize the characteristics of attack flow; on the other hand, genetic algorithm optimizes RF parameters by learning attack training sample set and improves the classification ability of RF so the classification effect of GAORF classification model in the four classification models is still best. The nu-SVM detection model has a better detection effect when the attack training samples increase, but its detection result is much worse than that of GAORF model when the attack training samples are few. In the early stage of attack, the attack eigenvalues are small, which can easily affect the location of the optimal hyperplane of SVM classification model, affect the recognition of normal flow, and increase the false-alarm rate. As for the C-SVM detection model, when the attack training sample is 30, the accuracy rate is 97.14% and the missing-alarm rate is 6%. With the increase of attack training samples, the classification performance becomes better. This is still the result of different fitting degree of the attack training samples, but the overall performance is still worse than GAORF. In addition, the one-class SVM detection model can only train normal training samples, thus increasing the number of attack samples, and it does not change the classification results. However, the accuracy rate of one-class-SVM attack-detection model based on FCD itself remains below 50% and missing-rate and false-alarm rate are higher, and its performance is much worse than that of the FGRF attack-detection method.

The comprehensive Tables 1 and 2 show that the GAORF classification model has stronger learning classification ability and robustness than the various classic SVM classification models for the constant change of normal samples and attack samples. Especially in the cloud computing environment, the sample feature dimension and the scale of datasets are increasing. Compared with the SVM classification model, RF can better adapt to the requirements of cloud computing. At the same time, facing the difficulty of finding the effective parameters for the detection model in cloud computing, the genetic algorithm provides a simple and effective search method, which can find the relative ideal parameters for the attack detection in a larger data range and the higher-dimension data sets. According to the characteristics of the FCD features, the characteristics of the two algorithms of GA and RF, and the experimental results, it is known that the FGRF detection method can detect attacks effectively, reduce the false- and missing-alarm rates and have good robustness. This detection method has better adaptability to DDoS attack-detection in a cloud computing environment.

6. Conclusion

In this paper, we proposed a DDoS attack-detection method based on FCD-RF, which can enhance the accuracy of

DDoS attack-detection in a cloud computing environment. We designed a feature-tuple with the statistical features of PSD and SDIA, which can describe the features of attack flow and normal flow, i.e., the FCD feature. This feature can reflect the asymmetric and semidirectivity interaction characteristics of the attack flow. The classification model was trained by the FCD feature sequence using the optimized RF based on a genetic algorithm. It could increase the accuracy rate of DDoS attack-detection and reduce the false- and missing-alarm rates. The experiment demonstrates that the detection model based on FCD and optimized RF can achieve higher accuracy and lower false- and missing-alarm rates with relatively good adaptability and robustness in a cloud computing environment.

A possible goal for our future research would be to consider multilayer mitigation and defense using profound resources in cloud computing.

Data Availability

The CAIDA UCSD "DDoS Attack 2007" Dataset used to support the findings of this study were supplied by the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) under license and so cannot be made freely available. Requests for access to these data should be made to the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT), registration and authorization on the official website of https://www.impactcybertrust.org/. After registration, datasets can be downloaded on the CAIDA official website of http://www.caida.org/data/passive/ddos-20070804_dataset.xml.

Conflicts of Interest

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

Authors' Contributions

For the six authors in this manuscript, Jieren Cheng and Mengyang Li completed the main tasks of conducting experiments, writing and revising manuscripts. Xiangyan Tang and Victor S Sheng revised and perfected English grammar and language expression. Yifu Liu modified the format of the manuscript. Wei Guo perfected and standardized the format of the references of this manuscript.

Acknowledgments

This work was supported by the Hainan Provincial Natural Science Foundation of China [2018CXTD333, 617048]; the National Natural Science Foundation of China [61762033, 61702539]; Hainan University Doctor Start Fund Project [kyqd1328]; and Hainan University Youth Fund Project [qnjj1444].

References

[1] S. Behal and K. Kumar, "Characterization and comparison of DDoS attack tools and traffic generators - a review," *International Journal of Network Security*, vol. 19, no. 3, pp. 383–393, 2017.

[2] A. Pras, J. J. Santanna, J. Steinberger et al., "DDoS 3.0 - How Terrorists Bring Down the Internet," in *Proceedings of the 18th International GI/ITG Conference on Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*, pp. 1–4, Springer International Publishing, 2016.

[3] *CORERO DDOS TRENDS REPORT (Q2 - Q3 2017)*, Corero Network Security, 2017.

[4] N. Singh, A. Hans, K. Kumar, and M. P. Singh Birdi, "Comprehensive Study of Various Techniques for Detecting DDoS Attacks in Cloud Environment," *International Journal of Grid and Distributed Computing*, vol. 8, no. 3, pp. 119–126, 2015.

[5] T. Xia, G. Qu, S. Hariri et al., "An efficient network intrusion detection method based on information theory and genetic algorithm," in *Proceedings of the 24th IEEE International Performance, Computing, and Communications Conference*, pp. 11–17, 2005.

[6] L. Zhou, M. Liao, C. Yuan, and H. Zhang, "Low-Rate DDoS Attack Detection Using Expectation of Packet Size," *Security and Communication Networks*, vol. 2017, Article ID 3691629, 14 pages, 2017.

[7] I. Dodig, V. Sruk, and D. Cafuta, "Reducing false rate packet recognition using Dual Counting Bloom Filter," *Telecommunication Systems*, vol. 68, no. 1, pp. 67–78, 2018.

[8] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "E-LDAT: a lightweight system for DDoS flooding attack detection and IP traceback using extended entropy metric," *Security and Communication Networks*, vol. 9, no. 16, pp. 3251–3270, 2016.

[9] 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.

[10] T.-M. Choi, H. K. Chan, and X. Yue, "Recent Development in Big Data Analytics for Business Operations and Risk Management," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 81–92, 2017.

[11] Y. Gu, Y. Wang, Z. Yang, F. Xiong, and Y. Gao, "Multiple-features-based semisupervised clustering ddos detection method," *Mathematical Problems in Engineering*, vol. 2017, Article ID 5202836, 10 pages, 2017.

[12] Y. Liu, Z.-P. Cai, P. Zhong, J.-P. Yin, and J.-R. Cheng, "Detection approach of DDoS attacks based on conditional random fields," *Journal of Software*, vol. 22, no. 8, pp. 1897–1910, 2011.

[13] P. D. Bojović, B. Ilija, O. Stanislav et al., "A Practical Approach to Detection of DDoS Attacks Using a Hybrid Detection Method," *Computers and Electrical Engineering*, 2017.

[14] J. Zhan, X. Fan, L. Cai, Y. Gao, and J. Zhuang, "TPTVer: A trusted third party based trusted verifier for multi-layered outsourced big data system in cloud environment," *China Communications*, vol. 15, no. 2, pp. 122–137, 2018.

[15] J. Shen, Z. Gui, S. Ji, J. Shen, H. Tan, and Y. Tang, "Cloud-aided lightweight certificateless authentication protocol with anonymity for wireless body area networks," *Journal of Network and Computer Applications*, vol. 106, pp. 117–123, 2018.

[16] L. Qi, X. Zhang, W. Dou, and Q. Ni, "A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2616–2624, 2017.

[17] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang, "Energy-efficient resource allocation for d2d communications underlaying cloud-ran-based lte-a networks," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 428–438, 2016.

[18] W. Lin, S. Xu, L. He, and J. Li, "Multi-resource scheduling and power simulation for cloud computing," *Information Sciences*, vol. 397-398, pp. 168–186, 2017.

[19] Y. Xu, L. Qi, W. Dou, and J. Yu, "Privacy-Preserving and Scalable Service Recommendation Based on SimHash in a Distributed Cloud Environment," *Complexity*, vol. 2017, Article ID 3437854, 9 pages, 2017.

[20] P. Li, J. Li, Z. Huang et al., "Multi-key privacy-preserving deep learning in cloud computing," *Future Generation Computer Systems*, vol. 74, pp. 76–85, 2017.

[21] Z. Huang, S. Liu, X. Mao, K. Chen, and J. Li, "Insight of the protection for data security under selective opening attacks," *Information Sciences*, vol. 412-413, pp. 223–241, 2017.

[22] L. Qi, X. Zhang, W. Dou, C. Hu, C. Yang, and J. Chen, "A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment," *Future Generation Computer Systems*, vol. 88, pp. 636–643, 2018.

[23] W. Jiang, G. Wang, M. Z. A. Bhuiyan, and J. Wu, "Understanding graph-based trust evaluation in online social networks: Methodologies and challenges," *ACM Computing Surveys*, vol. 49, no. 1, 2016.

[24] C. Yan, X. Cui, L. Qi, X. Xu, and X. Zhang, "Privacy-Aware Data Publishing and Integration for Collaborative Service Recommendation," *IEEE Access*, vol. 6, pp. 43021–43028, 2018.

[25] E. Luo, Q. Liu, and G. Wang, "Hierarchical Multi-Authority and Attribute-Based Encryption Friend Discovery Scheme in Mobile Social Networks," *IEEE Communications Letters*, vol. 20, no. 9, pp. 1772–1775, 2016.

[26] T. Peng, Q. Liu, D. Meng, and G. Wang, "Collaborative trajectory privacy preserving scheme in location-based services," *Information Sciences*, vol. 387, pp. 165–179, 2017.

[27] W. Gong, L. Qi, and Y. Xu, "Privacy-Aware Multidimensional Mobile Service Quality Prediction and Recommendation in Distributed Fog Environment," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 3075849, 8 pages, 2018.

[28] J. Kaur and K. Kaur, "A fuzzy approach for an IoT-based automated employee performance appraisal," *Computers, Materials and Continua*, vol. 53, no. 1, pp. 24–38, 2017.

[29] X. Zhang, Y. Tan, C. Liang, Y. Li, and J. Li, "A Covert Channel Over VoLTE via Adjusting Silence Periods," *IEEE Access*, vol. 6, pp. 9292–9302, 2018.

[30] J. Xu, L. Wei, Y. Zhang, A. Wang, F. Zhou, and C. Gao, "Dynamic Fully Homomorphic encryption-based Merkle Tree for lightweight streaming authenticated data structures," *Journal of Network and Computer Applications*, vol. 107, pp. 113–124, 2018.

[31] Q. Lin, H. Yan, Z. Huang, W. Chen, J. Shen, and Y. Tang, "An ID-based linearly homomorphic signature scheme and its application in blockchain," *IEEE Access*, vol. 6, no. 1, pp. 20632–20640, 2018.

[32] Q. Lin, J. Li, Z. Huang, W. Chen, and J. Shen, "A short linearly homomorphic proxy signature scheme," *IEEE Access*, vol. 6, pp. 12966–12972, 2018.

[33] T. Li, J. Li, Z. Liu, P. Li, and C. Jia, "Differentially private Naive Bayes learning over multiple data sources," *Information Sciences*, vol. 444, pp. 89–104, 2018.

[34] C. Gao, Q. Cheng, P. He, W. Susilo, and J. Li, "Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack," *Information Sciences*, vol. 444, pp. 72–88, 2018.

[35] J. R. Cheng, R. M. Xu, and X. Y. Tang, "An Abnormal Network Flow Feature Sequence Prediction Approach for DDoS Attacks Detection in Big Data Environment," *Computers, Materials & Continua*, vol. 55, no. 1, pp. 95–119, 2018.

[36] B. Jia, X. Huang, R. Liu, and Y. Ma, "A DDos attack detection method based on hybrid heterogeneous multiclassifier ensemble learning," *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 4975343, 9 pages, 2017.

[37] S. Yu, Y. Tian, S. Guo, and D. O. Wu, "Can we beat DDoS attacks in clouds?" *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 9, pp. 2245–2254, 2014.

[38] J. R. Cheng, J. Yin, Y. Liu et al., "Detecting distributed denial of service attack based on address correlation value," *Journal of Computer Research and Development*, vol. 46, no. 8, pp. 1334–1340, 2009.

[39] J. Cheng, X. Tang, and J. Yin, "A change-point DDoS attack detection method based on half interaction anomaly degree," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 10, no. 1, pp. 38–54, 2017.

[40] The Cooperative Association for Internet Data Analysis, *The Caida Ucsd 'DDoS Attack 2007' Dataset*, 2007.

[41] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[42] C. C. Chang and C. J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp. 1–27, 2011.