Anti-D Chain: A Lightweight DDoS Attack Detection Scheme Based on Heterogeneous Ensemble Learning in Blockchain

Bin Jia*, Yongquan Liang

College of Computer Science & Engineering, Shandong University of Science and Technology, Qingdao 266590, China
* The corresponding author, email: jiabin@sdust.edu.cn

Abstract: With rapid development of blockchain technology, blockchain and its security theory research and practical application have become crucial. At present, a new DDoS attack has arisen, and it is the DDoS attack in blockchain network. The attack is harmful for blockchain technology and many application scenarios. However, the traditional and existing DDoS attack detection and defense means mainly come from the centralized tactics and solution. Aiming at the above problem, the paper proposes the virtual reality parallel anti-DDoS chain design philosophy and distributed anti-D Chain detection framework based on hybrid ensemble learning. Here, AdaBoost and Random Forest are used as our ensemble learning strategy, and some different lightweight classifiers are integrated into the same ensemble learning algorithm, such as CART and ID3. Our detection framework in blockchain scene has much stronger generalization performance, universality and complementarity to identify accurately the onslaught features for DDoS attack in P2P network. Extensive experimental results confirm that our distributed heterogeneous anti-D chain detection method has better performance in six important indicators (such as Precision, Recall, F-Score, True Positive Rate, False Positive Rate, and ROC curve).

Keywords: DDoS attack detection; parallel blockchain technology; ensemble learning; AdaBoost; random forest

I. INTRODUCTION

Currently, the Internet interconnects billions of computers and smart device, and provides a global communication, storage and computation resources. Internet security is facing great challenges in many fields that include politics, economics, military, and social life. Security problem has been regarded as the dominated bottleneck of the development of Internet of Thing, Big Data, Cloud Computing, Artificial Intelligence, and Software Defined Networking (SDN) yet. Confidentiality, integrity, and availability are three important security issues of networks [1]. For availability of Internet service, the Distributed Denial of Service (DDoS) attack is one of the most significant threats. In spite of some forceful safeguards, the attack is still the most frequent and the most devastating one. With the rapid development of blockchain technology, the attack is bound to endanger blockchain network and many business applications.

Received: Feb. 22, 2020
Revised: May 15, 2020
Editor: Shangguang Wang

In this work, we have formulated an anti-D chain framework and designed a detection method based on heterogeneous ensemble learning with the built-in lightweight hybrid classifiers and virtual reality parallel blockchain tactics in order to detect DDoS attack in the blockchain scene.

1.1 Distributed denial of service attack

The DDoS attack is a coordinated attack, and is generated by using many compromised hosts [2]. It is by expanding one-to-one relationship attack mode to many-to-one relationship attack model [3]. The intensity of the attack has become stronger in recent years. Usually, prominent websites are the prime victims of such attacks, for instance, Twitter, Spotify, and Amazon suffered interruptions in their services for almost two hours on Oct. 21, 2016 owe to DDoS attack [4]. According to the sampling monitoring from National Internet Emergency Center (CNCERT or CNCERT/CC), China still suffered serious DDoS attack, and attack peak flow has been keeping rising in 2017.

At present, a new DDoS attack has arisen, and it is the DDoS attack in P2P network. The attacker utilizes the P2P file shared system which is widely used. It makes a mass of normal users visit targeted object. Moreover, the DDoS attack may disconnect some nodes each other in anonymity P2P network from blockchain. Therefore, it has attracted the attention of security researchers both in academia and industry. They proposed successively many different methods to detect and mitigate against the DDoS threat. Despite these efforts, DDoS attacks still keep some huge hazards.

1.2 Blockchain and parallel blockchain technology

The traditional and centralized relational database system is managed and maintained by a center organization. The center organization commands and updates all the data. However, in the multi-agency collaboration mode, the centralized database system has lots of trust issues.

Nowadays the development and popularization of blockchain technology have shown explosive growth. Such extreme speed has exceeded the expectation of many experts and scholars [5]. The blockchain technology attracted a lot of attentions due to its decentered, shared, anonymous, traceable, and tamper-resistant data structure. And the immutable feature for blockchain is determined by its irreversibility. Blockchain has led to the conversion from the information Internet to the value Internet [6], and applied to the trustless crowd-intelligence ecosystem in Industrial Internet of Things [7].

However, from a view of academics, the existing blockchain technology is short of virtual reality synchronization power. In addition, the computing experiment, prediction level and analyzing ability are inadequate. For this reason, the Ref. [8] proposed the parallel blockchain. The parallel blockchain is the theoretical method to solve effectively some problems related to blockchain modeling, experiment and decision-making. The parallel blockchain is based on the parallel intelligence theory and ACP (Artificial systems + Computational experiments + Parallel execution) approach. It constructs artificial blockchain system by describing formally the static characteristics and dynamic behaviors of the blockchain ecosystem core elements. The parallel blockchain technology based on ACP is a “Description to Prediction to Guidance” parallel intelligence technology. It is an inevitable development trend for Internet and intelligence industry drove by the blockchain technology.

1.3 Network security of blockchain

According to the hierarchy model of computer network, the existing blockchain network security solution mainly includes as follows. In network layer, the P2P architecture is used in the blockchain network. The security routing, anonymous mechanism, and trust model are research foci in P2P network of blockchain. In transport layer, the data transmission encryption and transmission reliability are the key scientific problems. In application layer, the detection and defense of known and unknown network attacks and privacy protection are crucial to blockchain business application. Therefore, the actual blockchain technology and blockchain network are in essence linked data structure, distributed computing framework and network model, finance systems, and the modeling and application of society and economics. However, the simulated attack experiment and intelligent detection ability for various application scenarios are defective in blockchain security system. Meanwhile, the abilities of intelligent analysis and making decision are devoid to guide the reality by virtualizing and the automation by artificializing.

We hope that blockchain is not like Internet, since the Internet security technology lagged far behind the Internet technology itself. In Big Data Era, the data is the most valuable asset for corporations and governments. The user data would be explosively increased as well. In blockchain business network, once user data is linked into chain, the data will be incrementally written. When Internet services deny providing suddenly, it will lead to form "isolated data island" in physical distributed storage systems, and can hardly realize data updating and information sharing.

1.4 Restrain DDoS attack used by parallel blockchain tactics

Because nowadays computer and information systems receive a mass of junk information and files by centralized server, which leads to DDoS attack. However, a decentralized platform allows users to rent out their bandwidth. It reduces immensely the possibility of DDoS success by the aggregated bandwidth.

In addition, due to the decentralized nature of blockchain, and because each node has complete blockchain information and it can verify the data validity of other nodes, the DDoS detection and defense by parallel blockchain technology is feasible and necessary when a DDoS attack occurs. For example, when a node is destroyed (unless the attacker controls 51% nodes of the network) and the others continue to provide services, attack detection need to be carried through, and attack traffic and data package ought to be filtrated in time.

For that reason, DDoS attack can be restrained by parallel blockchain technology based on ACP approach. And an intelligent parallel blockchain detection and defense system based on the parallel intelligence theory and ACP approach can effectively resist DDoS attack. It redeploys user network, shared bandwidth, and other resources to withstand collaboratively DDoS attack. It speeds up website loading by deploying Caching Device Network (CDN).

To the best of our knowledge, this paper proposes a set of method and technique based on the virtual reality parallel blockchain tactics and heterogeneous ensemble learning for the first time and attempts to apply them in DDoS attack detection of blockchain application scenario. Compared with the previous solutions, our proposed model and framework have the following advantages.

(i) We propose an anti-D chain scheme and design a detection method based on heterogeneous ensemble learning with the built-in lightweight hybrid classifiers and parallel blockchain technology in order to detect DDoS attack in the real blockchain scenario.

(ii) Our distributed heterogeneous anti-D chain detection framework has much stronger generalization performance, universality and complementarity to accurately recognize the DDoS attack traffic in blockchain network.

(iii) Our detection method has better performance in six important indicators (such as Precision, Recall, F-Score, True Positive Rate, False Positive Rate, and ROC curve) in artificial blockchain network. And the artificial blockchain can also effectively guide the DDoS attack defense in actual blockchain.

The remainder of this paper is organized as follows. Section II introduces the related work in DDoS attack detection and security research in blockchain in recent years. Section III describes the outline of ensemble learning, and the theoretical approaches and mathematics models of AdaBoost and Random Forest. What is more, a parallel blockchain scheme based on heterogeneous ensemble learning method is given. In Section IV, we design a novel lightweight parallel anti-D chain architecture to DDoS attack detection. Section V discusses the experimental details, and gives the experimental results and analyses. In Section VI, we summarize this paper.

II. Related Work

2.1 DDoS attack detection

The recent research work for DDoS attack detection years mainly includes the followings.

In 2016, Jia et al. [9] focused on how to distinguish the attack traffic from normal data flows in Big Data and brought forward a novel real-time DDoS attack detection mechanism based on Multivariate Dimensionality Reduction Analysis (MDRA) algorithm. In this mechanism, the dimensionalities of multiple characteristic variable in a network traffic record by Principal Component Analysis (PCA) first are reduced. Next, the correlation of lower dimensional variables is analyzed. Finally, the attack traffic can be differentiated from the normal traffic by MDRA and Mahalanobis Distance (MD). In 2017, Somani et al. [10] raised a new "Scale Inside-out" approach which reduces the "Resource Utilization Factor" to a minimal value for quick absorption of the attack. The novel approach sacrifices victim service resources and provides those resources to mitigation service in addition to other co-located services to ensure resource availability during the attack.

Although the above-mentioned researchers have achieved the updated developments and research fruits, DDoS attack detection in blockchain application scenarios is still a hot research field in academia and industry. Compared with the traditional methods, our approach based on a lightweight heterogeneous ensemble learning tactics and parallel blockchain technology is good at detecting DDoS attack traffic in many blockchain application scenarios.

2.2 Security in blockchain

Blockchain has been touted as a technology that can provide a robust and strong cybersecurity solution and high level of privacy protection [11]. The security research in blockchain in recent years mainly includes the following three aspects.

(i) Consensus mechanism. In order to ensure the data consistency and the transaction dependability in the blockchain system, the decentralized consensus mechanism was used. In 2016, Yuan et al. [12] introduces the classic distributed consistency algorithm, as the milestone research efforts and the key conclusion of distributed consensus algorithm. They also propose a novel model and classification approach of blockchain consensus algorithm and summarize the consensus algorithms and their performance measures by using an evolutionary tree.

(ii) Majority attack. In order to keep off double-spending problem in blockchain products based on PoW consensus mechanism (e.g. Bitcoin) and to protect blockchain security, the majority attack must be kept away. The 51% attack is a typical threat to any consensus protocol [13]. After the 51% computing power of bitcoin network is under the thumb, the identified blocks will be recalculated by using more computing power. However, the attack can modify its own trading record, and cannot alter any other trading record. It may hold back the blocks to affirm segmental or entire transaction.

(iii) Privacy Preserving. Nowadays, privacy preserving has been definitely becoming one of the most crucial issues and one of the most popular research topics at Internet of Thing (IoT) and blockchain, where data encryption is to achieve the important target. In 2018, Wang et al. [14] presented a privacy preserving blockchain incentive mechanism in the crowdsensing application. A cryptocurrency built on blockchain is used as a secure incentive way. Some high-quality contributors will get their payments that are recorded in transaction blocks. The miners would verify the transaction according to the sensing data assessment criteria published by the server.

It is acknowledged that the blockchain technology and its applications (e.g. Bitcoin [15] and Ethereum [16]) are no longer novelty, and the security of blockchain network and its blockchain-based application have become explosively popular. Although there are many researches on the security of blockchain recently, they are extremely rare for DDoS attack detection and defense mechanism on blockchain network and the security enhancement means in blockchain scenarios.

III. PRELIMINARIES

3.1 Outline of ensemble learning

Ensemble learning is a hot technology that can reinforce generalization performance of machine learning system. It generates multiple individual learners or classifiers and bring them together with a certain strategy in order to complete a specific learning or classification task. The individual learners or classifiers may not only be homogeneous, but also be heterogenous. The individual learners use the same type of algorithm in homogeneous ensemble, and they are known as base learners. The individual learners use the different types of algorithm in heterogenous ensemble, and they are known as component learners. The key research in ensemble learning is to design and to combine the outstanding and diverse individual learners. According to the generation approach of individual learner, the existing ensemble learning was divided into the serialization method and the parallelization method [17].

3.2 AdaBoost

Boosting is a family of algorithms that promote weak learner to strong learner. There are strong dependencies between every two individual learners in Boosting. It is an iterative serialization ensemble learning method, and it repetitively trains every base learner in the training sample again and again. The next iteration can offset the deficiency of existing model. The N base learners would be combined by weighting until the number of base learners reaches to the pre-set N.

Adaptive boosting (AdaBoost) [18] is the most perfect example of Boosting family. For

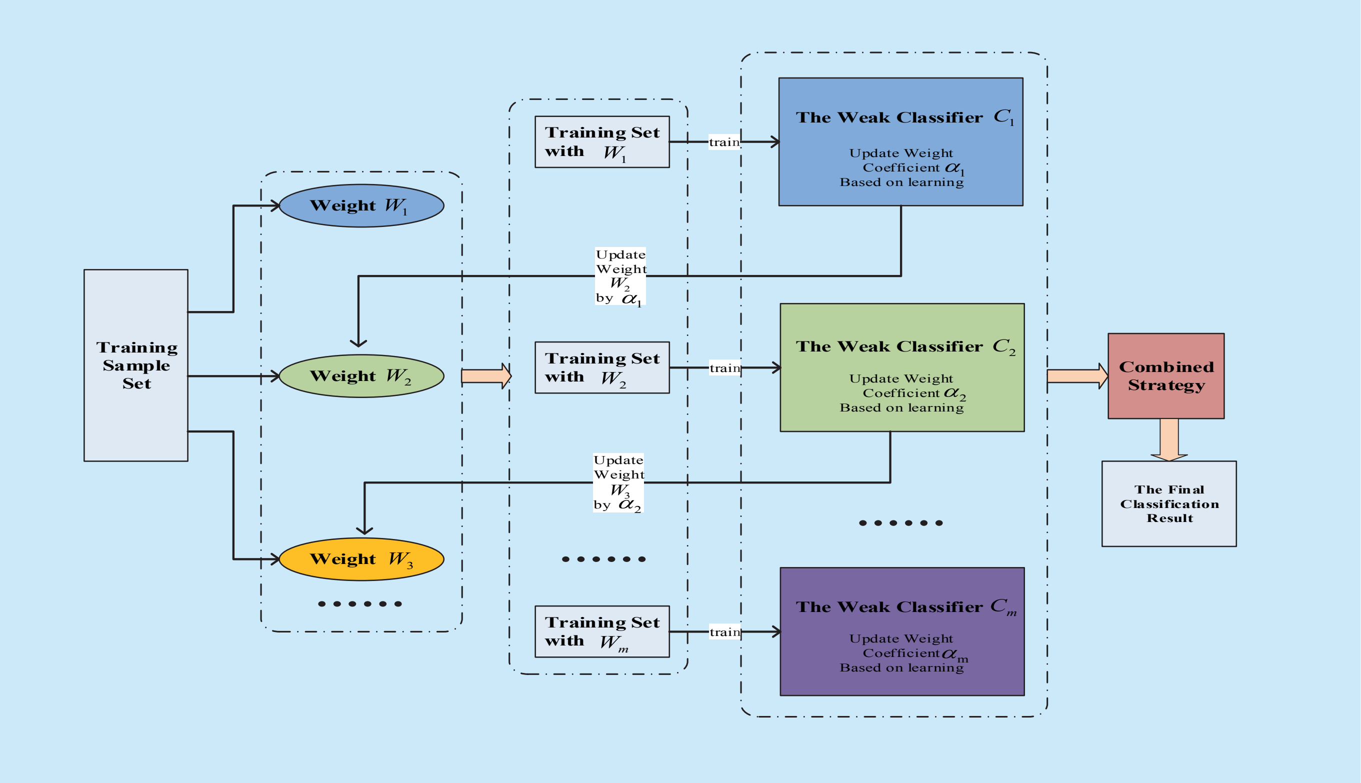
Training Sample Set
Weight W_1
Weight W_2
Weight W_3
......
Training Set with W_1
Training Set with W_2
......
Training Set with W_m
train
train
train
Update Weight W_2 by α_1
Update Weight W_3 by α_2
The Weak Classifier C_1
Update Weight Coefficient α_1 Based on learning
The Weak Classifier C_2
Update Weight Coefficient α_2 Based on learning
......
The Weak Classifier C_m
Update Weight Coefficient α_m Based on learning
Combined Strategy
The Final Classification Result

Fig. 1. *AdaBoost algorithm implementation schematic.*

AdaBoost algorithm, the sample weights of misclassification in learner will be enhanced in previous iteration, the sample weights of correct classification will be lowered. The distribution of sample is changed in every iteration, the sampling cannot be repeated. The algorithm uses weighted majority voting strategy. The weights show performance of every weak learner. The core of this strategy is to increase the weights of weak learner whose classification error rate is smaller, and to decrease the weights of weak learner whose classification error rate is higher. It has the lower generalization error, the higher classification accuracy, and the smaller probability to overfit. AdaBoost is sensitive to noise and outlier. Many classification algorithms (e.g. Bayesian, CART, C4.5, Decision Stump, ID3, SVM and so on) are used for weak learner in AdaBoost.

The Figure 1 shows the implementation schematic of AdaBoost algorithm.

Definition 1. The weighted classification error is that the classification error of sample weight for weak learner is acquired when the weight of classifying sample contrasted the errors over the entire data set. It is specifically shown as follows.

$$\varepsilon_i = \sum W_i(k) I[L_i(k) \neq y_k], \tag{1}$$

where ε_i is the weighted classification error of the i^{th} weak learner. $W_i(k)$ is the weight of sample k after i repetitions training. $L_i(k)$ is the result that the sample k is classified by the i^{th} weak learner. y_k is the label of sample k. $I[L_i(k) \neq y_k]$ is an indicator function, and it is expressed as follows.

$$I[L_i(k) \neq y_k] = \begin{cases} 0, & L_i(k) = y_k \\ 1, & L_i(k) \neq y_k \end{cases}. \tag{2}$$

According to the above formula, we can find that the weighted classification error is related to the error that the weak learner acts over the sample and the weight of each sample in the training set. Therefore, AdaBoost associates with the weak learner by the weighted classification error.

Definition 2. The relation between the weight α_i of weak learner and the weighted classification error itself is shown as follows.

$$\alpha_i = \frac{1}{2} \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right), \tag{3}$$

where the range of α_i is [0,1]. The weight is in inverse ratio to its weighted classification error. It is that the weight of the weak learner is larger when the weighted classification error is smaller, and it is vice versa.

Definition 3. The formula of the weight update of training set is expressed as follows.

$$W_{i+1}(k) = \frac{W_i(k)}{sum(W_i)} \times \begin{cases} e^{-a_i}, L_i(k) = y_k \\ e^{a_i}, L_i(k) \neq y_k \end{cases}, \tag{4}$$

where $W_{i+1}(k)$ is the result that the weight of sample k is updated by calculating weight α_i of weak learner L_i. If the classification on the sample k by using the learner is correct, the sample weight ought to be lower. Otherwise, the sample weight should be increased.

The model of AdaBoost algorithm is divided into the training phase and the classification phase.

(i) Training phase. Firstly, the weight W_i of each sample is allocated in training set. The individual weak learner is trained. Secondly, the weighted classification error ε_i and the weight α_i of weak learner L_i are computed respectively. Last but not least, the weight of the sample is updated, and the corresponding weight of each weak learner is got.

(ii) Classification phase. Firstly, the weighted average value of each weak learner is calculated. Then, the final classification result is got according to some decision rule.

3.3 Random forest

Bagging [19] is a famous representative of parallelization ensemble learning method. There are weak dependencies between every two individual learners in Bagging. It selects randomly training sample data by repositioning based on bootstrap sampling. It is similar with AdaBoost that Bagging does not limit to weak learner, however, the most common weak learners are the decision tree and neural network. In addition, the integration strategy usually adopts the simple voting. Because the bagging algorithm trains model by sampling every time, its generalization ability is stronger. It is effective to reduce model variance.

Random Forest [20] is an extension version of Bagging. Firstly, Random Forest is constructed by the base learners based on decision tree. Secondly, it selects randomly sample characteristic attribute except for training sample data. The Random Forest algorithm need not to prune in the process of construction and growth of every decision tree, so the randomness of sample characteristic attribute choice is added. The characteristic attribute randomness not only improves classification accuracy, but also reduces the correlation coefficient between every two decision trees in forest. Because the growth process of all trees is different, the other random factor is increased. The diversity of base learner in Random Forest comes from sample disturbance and attribute disturbance, which makes the generalization performance of ensemble to be promoted with enhancing the difference degree between every two individual learners. The predicted value is from the majority vote for classification, and from the average for regressions [21].

The implementation principle of Random Forest algorithm is showed in Figure 2.

Definition 4. The edge function is shown as follows.

$$\begin{aligned} mg(x,y) = {} & av_k\{I[h_k(X)=Y]\} \\ & -\max_{j\neq Y} av_k\{I[h_k(X)=j]\}, \end{aligned} \tag{5}$$

where I () is an indicator function. Y shows the correct classification vector, and j denotes incorrect classification vector. av_k is an average number.

Definition 5. The generalization error is shown as follows.

$$PE = P_{X,Y}[mg(X,Y)<0], \tag{6}$$

where the subscript X and Y show the definition space of probability. What is more, the upper bound range of generalization error is as follows.

$$PE \leqslant \frac{\bar{\rho}(1-S^2)}{S^2}. \tag{7}$$

Here $\bar{\rho}$ is the average value of all relevancies between every two decision trees, and s denotes the mean intensity of decision trees.

Equation (7) shows that the generalization performance of Random Forest is better when the classification ability of individual decision tree is stronger. However, the classification ability of decision tree is stronger when the

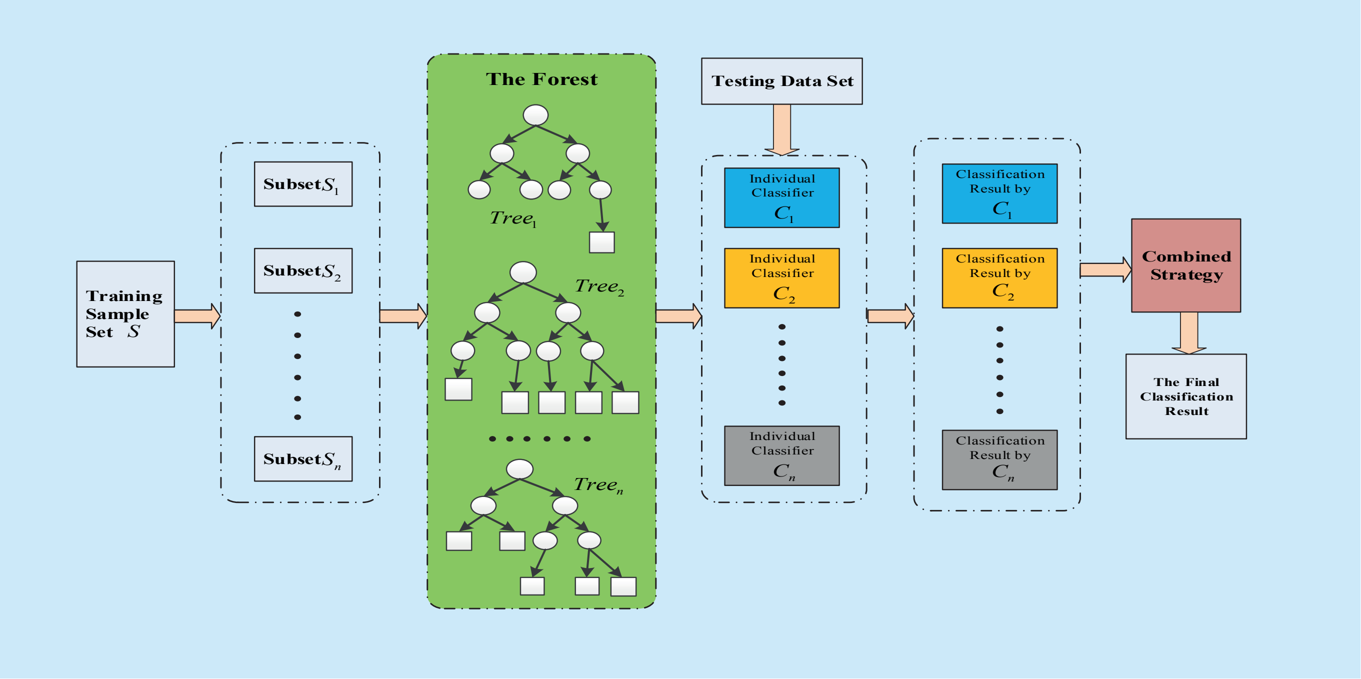
The Forest
Testing Data Set
Training Sample Set S
SubsetS_1
SubsetS_2
SubsetS_n
$Tree_1$
$Tree_2$
$Tree_n$
Individual Classifier C_1
Individual Classifier C_2
Individual Classifier C_n
Classification Result by C_1
Classification Result by C_2
Classification Result by C_n
Combined Strategy
The Final Classification Result

Fig. 2. *Random forest algorithm implementation principle.*

correlation between every two individual decision trees is smaller.

Theorem 1. The Convergence Theorem of Random Forest [22]. By the data structure of decision tree, (4) and (5) of the above, and in accordance with the law of large numbers, it can be drawn that the generalization errors of all decision trees in Random Forest should be converged to the following limit.

$$\lim_{n\to\infty} PE = P_{X,Y}\{P_{\Theta}[k(X,\Theta)=Y] \\ -\max_{j\neq Y} P_{\Theta}[k(X,\Theta)=j]<0\}, \tag{8}$$

where n is the number of decision tree in Random Forest.

Compared with AdaBoost, because of two random choices in Random Forest, the dependence of Random Forest on the individual learner is weaker, and it is not susceptible to the interference to noise data.

IV. Parallel Anti-DDoS Chain Architecture

4.1 Virtual Reality parallel anti-DDoS chain design philosophy based on hybrid lightweight ensemble learners

The parallel blockchain is a theoretical method that can solve effectively blockchain modeling, experiment, and related problem decision, and is deep combination between original research paradigm of parallel intelligent and emerging blockchain technology [23]. On the basis of theoretical analysis in ensemble learning, it can be drawn that the error of the integrated system will be smaller when the error correlation between every two individual learners is smaller [24]. So, the ensemble learning model based on homogeneous individual learners is already difficult to meet the requirement of higher integration performance.

Based on the above theories, a novel parallel blockchain scheme based on hybrid lightweight ensemble learning method is proposed, and it is applied to DDoS attack detection and resistance in blockchain network here. Our research scheme is as follows. The different ensemble learning algorithms are deployed to the different blockchains respectively, and the different sub-classifier algorithms in the same ensemble learning method are applied to the different blocks in the same chain respectively. The detecting and resisting units in different blocks are relatively independent with each other.

The heterogeneous multi-classifiers ensemble learning model design need to follow two key principles [25]. Firstly, the strong correlation between every two individual classifiers should be avoided as much as possible. In others words, the individual classifiers should be different. The differences include diversity, orthogonality, and complementarity [26]. Secondly, the optimal combination strategy of individual learner should be chosen.

In our method, some different lightweight classifiers are integrated into the same ensemble learning algorithm, such as CART and ID3. Then the different algorithms are randomly deployed in parallel blockchains respectively. They coordinate complementarily to detect and resist DDoS attack in the chains. The ensemble learning algorithm that uses different lightweight classifiers is able to improve immensely the generalization performance, universality and complementarity to accurately identify the onslaught features to launch an attack.

The parallel anti-DDoS chain design philosophy uses virtual and realistic interaction strategy. The strategy is that the artificial blockchain and actual blockchain interact by computing, experimenting, and evaluating. The artificial blockchain can constantly optimize the actual blockchain. Otherwise, the actual blockchain can continually guide the artificial blockchain as well.

4.2 Distributed heterogeneous anti-D chain framework

Here, we creatively propose and design a novel distributed heterogeneous anti-DDoS chain architecture, as shown in Figure 3. The framework is based on hybrid lightweight ensemble learning method and uses Virtual Reality (VR) parallel tactics.

(i) In the first blockchain, the different lightweight classifiers are alternately deployed in different blocks of the same chain, such as CART and ID3. They coordinate complementarily to detect and resist DDoS attack by combining strategy and thought of AdaBoost. In the second blockchain, similarly, the different lightweight classifiers are alternately deployed in different blocks of the same chain, and they synchronously detect and resist DDoS attack by combining strategy and thought of Random Forest. For the other blockchains, we can choose to use the other ensemble learning algorithms.

(ii) The artificial blockchain based on VR parallel tactics is structured in an experimental setting, and it is connected with the actual blockchain by virtual and realistic interaction means. The interaction includes two aspects. Firstly, the fast and precise intelligent detection for DDoS attack traffic in artificial blockchain can effectively guide the DDoS attack defense in actual blockchain. Secondly, the competence deficit for late-model and unknown DDoS attack traffic detection and defense can prompt to improve stronger detection capability as well.

V. Experimental Results and Analysis

In the section, firstly, we introduce the experimental data set, the experimental platform, and the data pretreatment method to serve our experiment. Then, the experimental results are got. Finally, we analyze and evaluate performance of our detection framework and method in six important indicators (such as Precision, Recall, F-Score, True Positive Rate, False Positive Rate, and ROC curve). Next, we will describe the experiment in detail.

5.1 Experimental Platform and Data Preprocessing

In this paper, we use a high-performance server as our experimental platform, and the virtualization method is employed in the server. A small cluster is built on the Hadoop big data platform. Here, the DataNode of Hadoop Distributed File System (HDFS) is used as the block node in the blockchain, MapReduce is the parallel processing framework of Hadoop. On the platform, we set up the artificial blockchains and simulate the Anti-DDoS Chain detection scheme in the actual blockchain.

In addition, we use the version 2016 of the famous Knowledge Discovery and Data Mining (KDD) Cup 99 data set [27]. In our

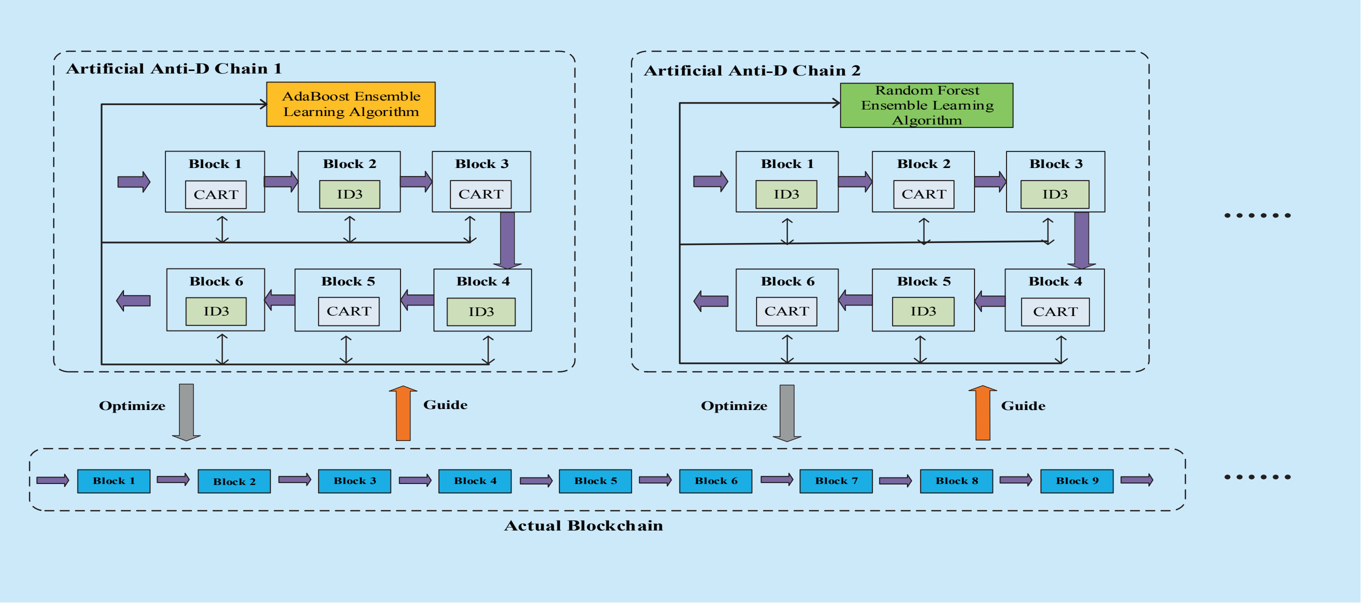

Fig. 3. *Distributed heterogeneous anti-DDoS chain framework based on ensemble learning and VR parallel.*

experiment, kddcup.data_10_percent.gz and corrected.gz in the data set are used as the training set and texting set respectively. The data set covers four main categories of attack, i.e. DoS, R2L, U2R, and Probing. Here, we use these records labeled as "normal" in the above training set to construct our benchmark data and employ the above testing set to verify our detection method. The DoS category is divided into four types, and they are smurf, neptune, back, and others. The data pretreatment procedure is shown as follows.

Firstly, for every network traffic record, it includes the information that has been separated into 41 features plus 1 label [28]. We need to get all numeric data for 41 features of every record. However, there are three nonnumeric features in all features, and these are protocol type, service, and flag. They should be transformed into numeric type. Type conversion is achieved by the pd.Categorical module in pandas tool of NumPy. Secondly, the feature value in numeric type needs to be standardized as well. Here, we use the StandardScaler module in scikit-learn of Python as our data normalization method.

5.2 Evaluation index and experimental results analysis

Some evaluation indexes to evaluate one detection method are indispensable. In this paper, the six important indicators are used, and they are Precision, Recall, F-Score, True Positive Rate (TPR), False Positive Rate (FPR), and Receiver Operating Characteristic (ROC) curve. Precision is the measurement of accuracy, and it denotes the proportion of true positive samples in all predicted positive samples. Recall is the measurement of coverage, and it measures how many the positive samples are exactly predicted. F-Score is a harmonic average between Precision and Recall, and the range of F-Score is between 0 and 1. TPR is the proportion of the predicted positive, and true positive samples in all positive samples. FPR is the proportion of the predicted positive, but true negative samples in all negative samples. Every point on the ROC curve reflects the sensitivity to the same signal stimulus. The curve is obtained by setting different thresholds, and there is a tradeoff between the TPR and FPR. In the ROC space, the abscissa is FPR, and the ordinate is TPR. The formulae of the first five indicators are defined as follows.

$$\text{Precision}=\frac{TP}{TP+FP}, \tag{9}$$

$$\text{Recall}=\frac{TP}{TP+FN}, \tag{10}$$

$$\text{F-Score}=(1+\beta^2)\times\frac{Precision\times Recall}{\beta^2\times Precision+Recall}. \tag{11}$$

Here β is used to balance the weights between Precision and Recall in F-Score computation. When,

(i) β=1, it shows that Precision is as important as Recall.

(ii) β<1, it expresses that Precision is more important than Recall.

(iii) β>1, it denotes that Recall is more important than Precision.

Here, we set β=1, therefore,

$$\text{F1-Score}=\frac{2\times Precision\times Recall}{Precision+Recall}, \tag{12}$$

namely,

$$\text{F1-Score}=\frac{2\times TP}{2\times TP+FP+FN}. \tag{13}$$

So F1-Score takes comprehensively into account the two results (i.e. Precision and Recall).

$$\text{TPR=Recall}=\frac{TP}{TP+FN}, \tag{14}$$

$$\text{FPR}=\frac{FP}{FP+TN}. \tag{15}$$

Where,

-- *TP* (True Positive): It is the number that positive samples are correctly classified as positive samples;

-- *FP* (False Positive): It is the number that negative samples are incorrectly classified as positive samples;

-- *TN* (True Negative): It is the number that negative samples are correctly classified as negative samples;

-- ***FN*** (False Negative): It is the number that positive samples are incorrectly classified as negative samples.

The matrix is called as the confusion matrix [29] based on the above counts. The matrix is listed in Table 1 as follows.

According to the above six evaluation indicators, we would measure the performance of our detection method. It should be specially explained that the two voting strategies (hard voting and soft voting) are used to preferably validate our method. In hard voting, the final result is determined on the basis of majority rule. In soft voting, in the light of the average value of probabilities that all models forecast to be a certain category, the type with the highest probability is the final predicted result. In addition, a heuristic judgment based on experience is used as our weight selection approach.

In Figure 4 and Figure 5, the heat maps of confusion matrix for hard voting and soft voting are given respectively.

In Figure 6 and Figure 7, the Precision, Recall, and F1-score for hard voting and soft voting are shown respectively. Here, we refer to the past experience threshold value in ensemble learning model, and we finally select six threshold values in every 50 intervals from 50 to 300 to evaluate the performance of our method.

In Figure 8 and Figure 9, the ROC Curve for hard voting and soft voting are exhibited respectively.

In conclusion, it is easy to find that our detection method has the outstanding performance in above six indicators (such as Precision, Recall, F-Score, True Positive Rate, False Positive Rate, and ROC curve) in artificial blockchain network. And the artificial blockchain can also effectively guide the DDoS attack defense in actual blockchain.

Table I. *Confusion matrix.*

	Predicted Positive	Predicted Negative	Total
True Positive	TP	FN	P
True Negative	FP	TN	N
Total	P’	N’	P+N (P’+N’)

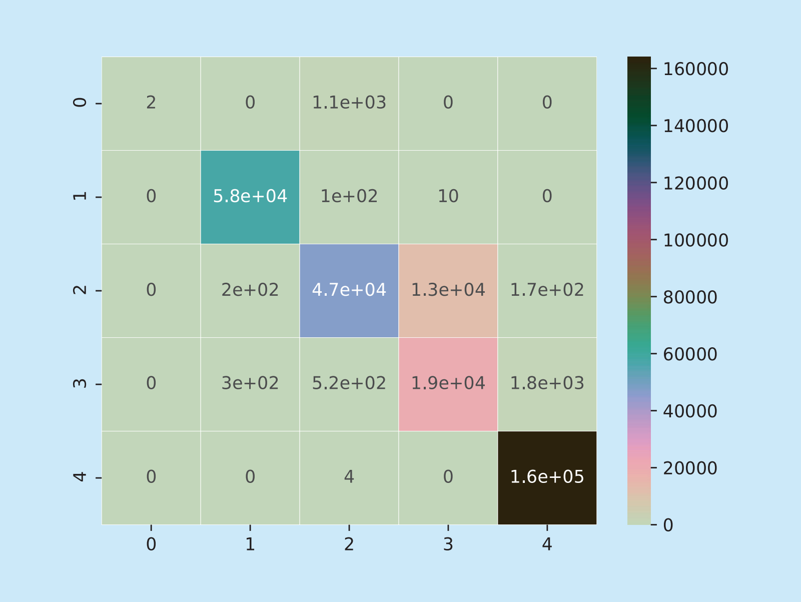

Fig. 4. *Heat map of confusion matrix for hard voting.*

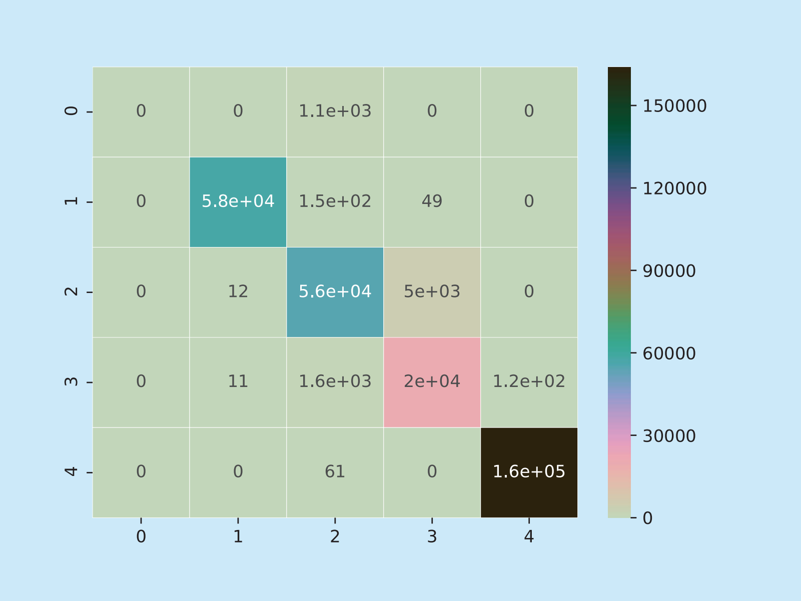

Fig. 5. *Heat map of confusion matrix for soft voting.*

VI. Conclusion

In this work, we have formulated an anti-D chain framework and designed a detection method based on heterogeneous ensemble learning with the built-in lightweight hybrid classifiers and virtual reality parallel blockchain tactics in order to detect DDoS attack in

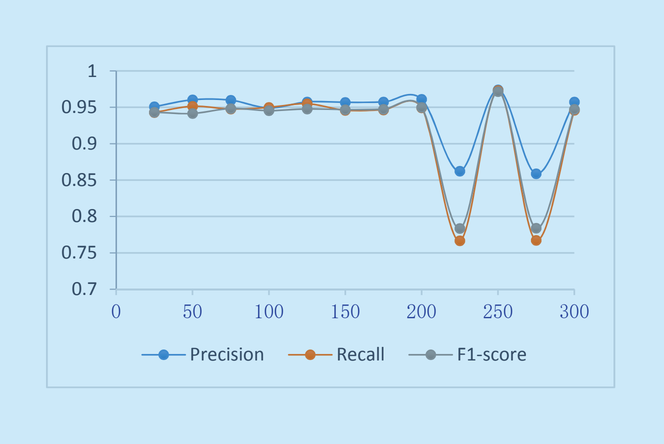

Fig. 6. *Precision, recall, and F1-score for hard voting.*

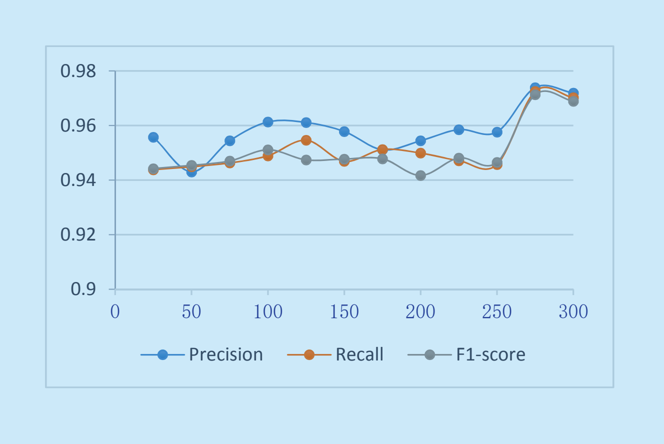

Fig. 7. *Precision, recall, and F1-score for soft voting.*

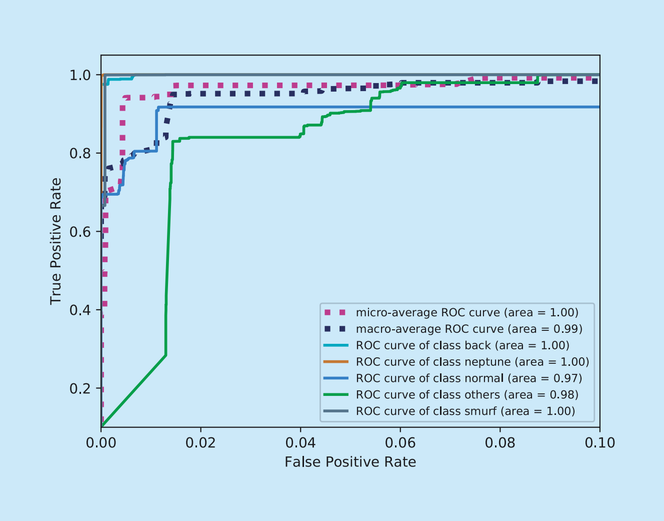

Fig. 8. *ROC curve for hard voting.*

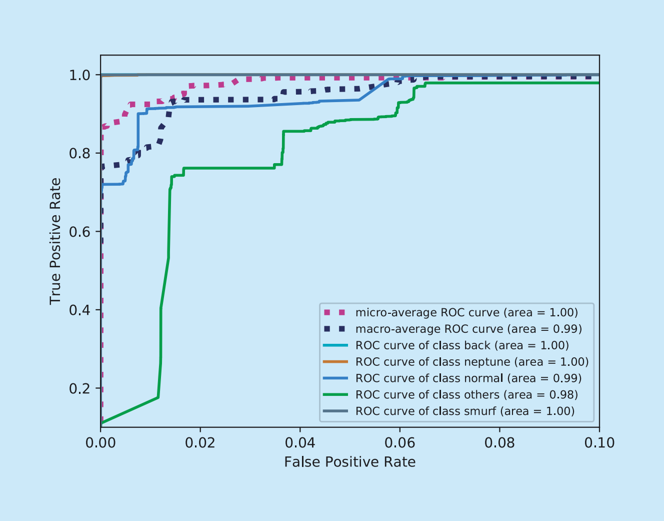

Fig. 9. *ROC curve for soft voting.*

the blockchain scene. Compared with traditional and existing DDoS attack detection and defense means based on centralized tactics and solution, our distributed heterogeneous anti-D chain detection framework has much stronger generalization performance, universality and complementarity to accurately recognize the DDoS attack features in blockchain network, and our detection method has the better performance in six important indicators (such as Precision, Recall, F-Score, True Positive Rate, False Positive Rate, and ROC curve) in artificial blockchain network. Meanwhile the artificial blockchain can also effectively guide the DDoS attack defense in actual blockchain.

ACKNOWLEDGEMENT

This work has been performed in the Project "Cloud Interaction Technology and Service Platform for Mine Internet of things" supported by National Key Research and Development Program of China (2017YFC0804406), and partly supported by the Project "Massive DDoS Attack Traffic Detection Technology Research based on Big Data and Cloud Environment" supported by Scientific Research Foundation of Shandong University of Science and Technology for Recruited Talents (0104060511314).

References

[1] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation computer systems*, vol. 28, no. 3, pp. 583-592, 2012.

[2] N. Hoque, D.K. Bhattacharyya, and J.K. Kalita, "Botnet in DDoS attacks: trends and challenges," *IEEE communication surveys & tutorials*, vol. 17, no. 4, pp. 2242-2270, 2015.

[3] Q. Wei, Z. Wu, K. Ren, and Q. Wang, "An openflow user-switch remapping approach for DDoS defense," *KSII Transactions on Internet and information systems*, vol. 10, no. 9, pp. 4529-4548, 2016.

[4] S. Behal and K. Kumar, "Detection of DDoS attacks and flash events using information theory matrics-An empirical investigation," *Computer Communications*, vol. 103, pp. 18-28, 2017.

[5] R.Y. Chen, "A traceability chain algorithm for artificial neural networks using T–S fuzzy cognitive maps in blockchain," *Future Generation Computer Systems*, vol. 80, pp. 198-210, 2018.

[6] J.J. Sikorski, J. Haughton, and M. Kraft, "Blockchain technology in the chemical industry: machine-to-machine electricity market," *Appl. Energy*, vol. 195, pp. 234-246, 2017.

[7] J. Xu, S. Wang, B. K. Bhargava, and F. Yang. "A Blockchain-enabled Trustless Crowd-Intelligence Ecosystem on Mobile Edge Computing," IEEE Transactions on Industrial Informatics, vol. 15, no. 6, pp. 3538-3547, 2019.

[8] Y. Yuan, and F.Y. Wang, "Parallel blockchain: concept, methods and issues," *Acta Automatica Sinica*, vol. 43, no. 10, pp. 1703-1712, 2017.

[9] B. Jia, Y. Ma, X. Huang, Z. Lin, and Y. Sun, "A novel real-Time DDoS attack detection mechanism based on MDRA algorithm in big data," *Mathematical Problems in Engineering*, pp. 1-10, Sep. 2016.

[10] G. Somani, M.S. Gaur, D. Sanghi, M. Conti, and M. Rajarajan, "Scale Inside-Out: Rapid Mitigation of Cloud DDoS Attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 6, pp. 959-973, 2017.

[11] N. Kshetri, "Blockchain's roles in strengthening cybersecurity and protecting privacy," *Telecommunications Policy*, vol. 41, pp. 1027-1038, 2017.

[12] Y. Yong, X. Ni, S. Zeng, and F.Y. Wang, "Blockchain Consensus Algorithms: The State of the Art and Future Trends," *Acta Automatica Sinica*, vol. 44, no. 11, pp. 2011-2022, 2016.

[13] E. Deirmentzoglou, G. Papakyriakopoulos, and C. Patsakis, "A Survey on Long-Range Attacks for Proof of Stake Protocols," *IEEE Access*, vol. 7, pp. 28712-28725, 2019.

[14] J. Wang, M. Li, Y. He, H. Li, K. Xiao, and C. Wang, "A Blockchain Based Privacy-Preserving Incentive Mechanism in Crowdsensing Applications," *IEEE Access*, vol. 6, pp. 17545-17556, 2018.

[15] F. Tschorsch and B. Scheuermann, "Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2084-2123, 2016.

[16] T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhang, "Understanding Ethereum via Graph Analysis," *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, IEEE, pp. 1484-1492, 2018.

[17] Z. Zhou, "Machine Learning," *Tsinghua University Press*, pp. 171–173, 2016.

[18] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *In European conference on computational learning theory*, Springer, Berlin, Heidelberg, pp. 23-37, 1995.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[20] L. Breiman, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1-39, 2010.

[21] P.A.A. Resende and A.C. Drummond, "A Survey of Random Forest Based Methods for Intrusion Detection Systems," *ACM Computing Surveys*, vol. 51, no. 3, pp. 48:1-48:36, 2018.

[22] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[23] F.Y. Wang, X. Wang, L. Li, and L. Li, "Steps toward parallel intelligence," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 4, pp. 345-348, 2016.

[24] F. Palmieri, S. Ricciardi, U. Fiore, M. Ficco, and A. Castiglione, "Energy-oriented Denial of Service Attacks: An Emerging Menace for Large Cloud Infrastructures," *The Journal of Supercomputing*, vol. 71, no. 5, pp. 1620–1641, 2015.

[25] Z.L. Fu and X.H. Zhao, "Dynamic Combination Method of Classifiers and Ensemble Learning Algorithms based on Classifiers Combination," *Journal of Sichuan University (Engineering Science Edition)*, vol. 43, no. 2, pp. 58-65, 2011.

[26] L.I. Kuncheva, "That Elusive Diversity in Classifier Ensembles," *In Proceedings of the 1st Iberian Conference on Pattern Recognition and Image Analysis (ibPRIA'03)*, Springer, Mallorca, Spain, pp 1126-1138, 2003.

[27] M. Tavallaee, E. Bagheri, W. Lu, A.A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE International Conference on Computational Intelligence for Security & Defense Applications*, IEEE, pp. 1-6, 2009.

[28] C. Bae, W.C. Yeh, M.A. Shukran, Y.Y. chung, and T.J. Hsieh, "A novel anomaly-network intrusion detection system using ABC algorithms," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 12, pp. 8231-8248, 2012.

[29] C. Guo, Y. Zhou, Y. Ping, Z. Zhang, G. Liu, and Y. Yang, "A distance sum-based hybrid method for

intrusion detection," *Applied Intelligence*, vol. 40, no. 1, pp. 178-188, 2014.

Biographies

Bin Jia (M'19), is a Lecturer in the College of Computer Science & Engineering of Shandong University of Science and Technology, Qingdao, China. He received a M.E. degree in Computer Technology from Beijing University of Posts and Telecommunications, and a Ph.D. in Software Engineering from Beijing University of Posts and Telecommunications. His current research is in blockchain technology and its security, 5G mobile communication security, system security in Industrial Internet of Things, and Hacker psychology and social engineering. Email: jiabin@sdust.edu.cn

Yongquan Liang, is a Professor in the College of Computer Science & Engineering of Shandong University of Science and Technology, Qingdao, China. He received a M.S. degree in Computer Software from Beijing University of Aeronautics and Astronautics, and a Ph.D. in Computer Software and Theory from Institute of Computing Technology, Chinese Academy of Sciences. His current research is in artificial intelligence, cloud computing, and big data analytics and decision making. Email: lyq@sdust.edu.cn