Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 1467051, 10 pages
http://dx.doi.org/10.1155/2016/1467051

Research Article

A Novel Real-Time DDoS Attack Detection Mechanism Based on MDRA Algorithm in Big Data

Bin Jia,[1,2,3] Yan Ma,[1] Xiaohong Huang,[1] Zhaowen Lin,[1,2,3] and Yi Sun[2,3,4]

[1]*Information and Network Center, Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, Shijiazhuang 050081, China*
[3]*National Engineering Laboratory for Mobile Network Security (No. [2013] 2685), Beijing 100876, China*
[4]*Network and Information Center, Institute of Network Technology and Institute of Sensing Technology and Business, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Correspondence should be addressed to Bin Jia; jb_qd2010@bupt.edu.cn

Received 25 March 2016; Revised 25 July 2016; Accepted 10 August 2016

Academic Editor: Nazrul Islam

In the wake of the rapid development and wide application of information technology and Internet, our society has come into the information explosion era. Meanwhile, it brings in new and severe challenges to the field of network attack behavior detection due to the explosive growth and high complexity of network traffic. Therefore, an effective and efficient detection mechanism that can detect attack behavior from large scale of network traffic plays an important role. In this paper, we focus on how to distinguish the attack traffic from normal data flows in Big Data and propose a novel real-time DDoS attack detection mechanism based on Multivariate Dimensionality Reduction Analysis (MDRA). In this mechanism, we first reduce the dimensionality of multiple characteristic variables in a network traffic record by Principal Component Analysis (PCA). Then, we analyze the correlation of the lower dimensional variables. Finally, the attack traffic can be differentiated from the normal traffic by MDRA and Mahalanobis distance (MD). Compared with previous research methods, our experimental results show that higher precision rate is achieved and it approximates to 100% in True Negative Rate (TNR) for detection; CPU computing time is one-eightieth and memory resource consumption is one-third of the previous detection method based on Multivariate Correlation Analysis (MCA); computing complexity is constant.

1. Introduction

The Denial of Service (DoS) attack is one of the most popular attacks on the Internet. It is implemented by forcing a kidnapped computer to launch or consuming its resources, such as CPU cycle, memory, and network bandwidth. When the DoS attack is generated by a great variety of distributed computers, it is called Distributed Denial of Service (DDoS). DDoS has become one of the main challenges to cyber security today.

DDoS attack is launched by some remote-controlled Zombies. It prevents legitimate users from accessing some specific network services or paralyzes the victims' own services by occupying computer resources or network bandwidth partly or completely. If there are more abnormal traffic data packets and more kidnapped Zombies hosts, more damage occurs in the network. If the number of Zombies hosts is large enough, it even can disrupt the whole network environment and all servers fleetly.

In the summer of 1999, the Computer Incident Advisory Capability (CIAC) reported the first DDoS attack incident [1]. Since then, DDoS has become the mostly convenient and effective attack means frequently used by hackers. In 2000, it is the answer told by Internet sites (e.g., Microsoft, Yahoo, and Amazon) that cannot be accessed for a long time, because of severe DDoS attack.

DDoS attacks are mainly classified into three categories based on different attacked subjects. The first kind is called

Netflow-DDoS attack and there are many typical instances such as DNS amplification attack, SNMP amplification attack, UDP Flood, and ICMP Flood. The second one is connection-DDoS attack. SYN Flood and TCP Flood are the most influential attack cases. Besides, there is a kind of DDoS attack based on application such as HTTP Get Flood and SSL Flood. In this paper, we focus on how to detect the Netflow-DDoS and connection-DDoS attacks.

In spite of all the effort from industry to academia, DDoS attack is still an open problem. In recent years, technique and level of DDoS attack are ceaselessly advancing with the improvement of capability for attack detection. With the emergence of Big Data technology, it is particularly much more difficult than ever before to prevent the network from various DDoS attacks. The continuously growing network traffic makes it impossible to detect network attack behavior from such large scale of network traffic based on previous detection methods.

In this paper, we address the abovementioned challenges and propose a novel method for real-time DDoS attack detection based on Multivariate Dimensionality Reduction Analysis (MDRA) algorithm, which combines Principal Component Analysis (PCA) and Multivariate Correlation Analysis (MCA). Compared with the previous solutions, our proposed algorithm has the following advantages:

(i) Higher precision rate approximates to 100% in True Negative Rate (TNR).

(ii) CPU computing time is one-eightieth of the previous detection method based on MCA.

(iii) Memory resource consumption is one-third of the previous detection method based on MCA.

(iv) Computing complexity is constant.

To the best of our knowledge, this paper proposes the theoretical method for the first time and attempts to apply it in the field of DDoS attack detection.

The remainder of this paper is organized as follows. Section 2 introduces the related work in DDoS attack detection and analyzes related shortcomings. Section 3 describes the theoretical approach to our detection mechanism. What is more, we design the attack detection framework based on MDRA. Section 4 discusses the experimental details and gives the experimental results and analyses. In Section 5, we summarize this paper.

2. Related Work

Although there is a development history of almost 20 years for it, DDoS attack detection is still a hot field of research in industry and academia. And its corresponding method and technique have to keep up with the times along with complexity and diversity of DDoS attack means. Previous work mainly includes the following.

In 2004, Kim et al. [2] proposed a combined data mining approach for the DDoS attack detection of the various types, which studied the automatic feature selection module and the classifier generation module. Because the analysis of per data flow is indispensable to DDoS attack detection, they used the data based on Netflow as the gathering data. In 2007, Scherrer et al. [3] focused on how to extract DDoS attack features and how to detect and filter DDoS attack packets by a number of known characteristics. In 2008, Lee et al. [4] designed a method for proactive detection of DDoS attack by exploiting its architecture and selecting different variables based on attack features; then, they performed cluster analysis for proactive detection of attack. In 2010, Nguyen and Choi [5] introduced a method for preliminary detection of DDoS attacks by classifying the network conditions. They selected some variables based on the key features. What is more, they applied the k-nearest neighbor (k-NN) method to classify the network conditions into each phase of DDoS attack. In addition, Tsai and Lin [6] told us a new method to detect the DDoS attack called "Triangle Area Based Nearest Approach." By using this approach, the accuracy and the False Positive Rate (FPR) were improved. In 2012, Bhange et al. [7] presented the idea about the DDoS attack and its impact on network traffic. This paper studied DDoS attack by analyzing the distribution of network traffic in order to distinguish anomaly traffic from the normal network behavior. In 2014, Tan et al. [8] brought forth a more sophisticated DoS attack detection approach using MCA. Following the emerging method, their paper proposed a new detection system based on MCA to protect online services against DoS attacks. In the same year, Luo et al. [9] developed a mathematical model for estimating the combined impact of DDoS attack pattern and network environment on attack effect by originally capturing the adjustment behaviors of victim TCPs congestion window.

DDoS attack can be detected by statistical analysis, data mining, and machine learning. However, some existing detection methods and techniques still suffer from low precision and TNR, or some of them cannot actively detect DDoS attacks. The previous detection methods and techniques already cannot meet the requirements of the Big Data era in particular because of their low detection efficiency, high resource consumption, and high computing complexity. In this paper, we propose a novel detection mechanism based on MDRA to show how to detect DDoS attack traffic effectively and in real time.

3. Detection Mechanism

Figure 1 shows the overview of our real-time DDoS detection framework. We first collect network traffic data sample from Internet and then input them into data acquisition system, which is composed of data cleaning, data store, and data anonymization module. Next, the processed traffic data are fed into traffic feature Big Data system. The traffic features in this system have two functions. The first one is applied to Online Attack Detection, and the other one is used for Offline Traffic Analysis based on Knowledge Base. Here, the results of Offline Traffic Analysis provide the feature recognition for Online Attack Detection. Last but not least, current network is adjusted on the basis of routing policy offered by the results of Online Attack Detection.

In this section, our novel method is separated into three components, that is, traffic feature dimensionality reduction, traffic feature correlation analysis, and attack

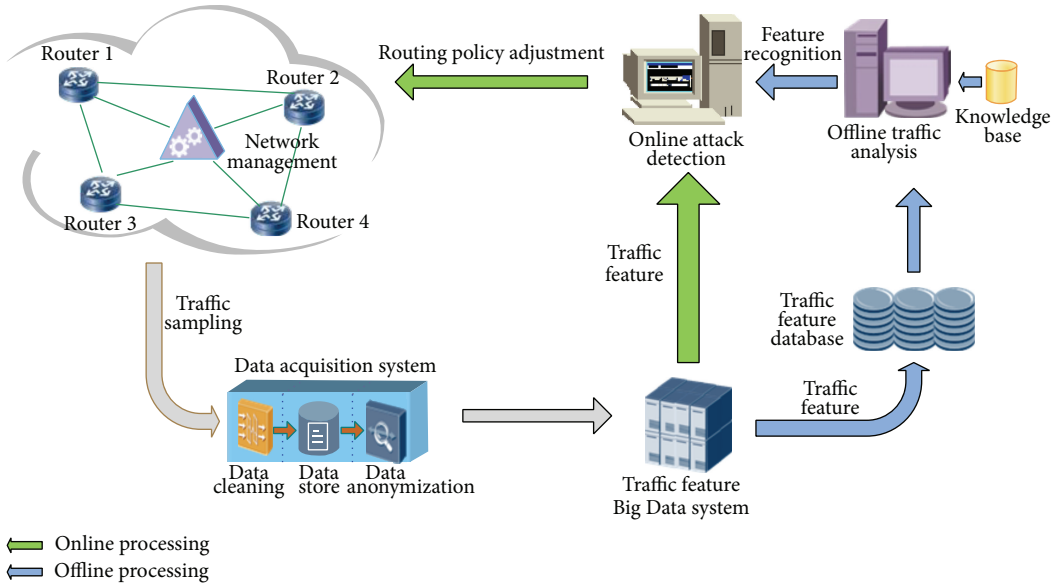

Figure 1: Overview of real-time DDoS attack detection framework.

detection framework based on MDRA and threshold. These components are introduced in following subsections.

3.1. Traffic Feature Dimensionality Reduction. A network traffic record encompasses a wide variety of high dimensional features. However, some of these high dimensional features are redundant or noisy. They may influence the effectiveness and efficiency of attack detection. In order to eliminate data redundancy and data noise, we introduce a dimensionality reduction technique into our detection method. The PCA method is used to extract less dimensional and more representative features. The projections on the remaining dimensionalities are called the principal components [10]. One advantage of PCA is its data-driven design by keeping the principal components of feature data and eliminating the correlated and measured feature data. Currently, PCA has been widely applied in the domain of intrusion detection [11] (such as [12, 13]) and the other fields (such as [14]).

In the PCA method, some original dependent random variables are transformed into new random variables whose components are uncorrelated by orthogonal transformation. The covariance matrix that is composed of original random variables is transformed into a diagonal matrix in the form of algebra. The original coordinate system is transformed into a new orthogonal coordinate system that points to multiple orthogonal directions in the form of geometry.

PCA is able to obtain P principal components. The first principal component is the linear combination for the maximum variance. If the first principal component is not enough to represent information of the original variables, we select the second linear combination. In order to effectively reflect the original information, the existing information for the first principal component needs not to appear in the second principal component. By this analogy, all subsequent principal components can be constructed. We assume that a network traffic record sample set X includes n samples and the dimension of each sample is d. That is to say, $X = \{X_1, X_2, \ldots, X_n\}$ and $X_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in R^d$, $i = 1, 2, \ldots, n$. The representation of sample matrix S is $S \in R^{n \times d}$. Then, the covariance matrix of sample matrix is calculated by the following formula:

$$C = \frac{S^T S}{n-1}, \quad C \in R^{d \times d}. \tag{1}$$

Next, the covariance matrix C needs to be diagonalizable. Here, the matrix C is a symmetric matrix, and the purpose of symmetric matrices diagonalization is to find an orthogonal matrix P; let

$$P^T C P = \Lambda \cdot P, \quad \Lambda \in R^{d \times d}. \tag{2}$$

Assuming that we get the corresponding dimensions for the first p ($p < d$) biggest eigenvalues, a new diagonal matrix Λ_1 ($\Lambda_1 \in R^{p \times p}$) is set up according to the p eigenvalues. The corresponding p eigenvalues constitute a new eigenvector matrix P_1 ($P_1 \in R^{d \times p}$). Actually, these eigenvalues in P_1 constitute a new coordinate system in low dimension space, and those are the principal components.

Assuming that the sample matrix after PCA dimensionality reduction is S_1, according to the purpose of PCA, the covariance between every two dimensions basically is zero in S_1. In other words, the covariance matrix of S_1 is Λ_1. It is to satisfy the following condition:

$$\frac{S_1^T S_1}{n-1} = \Lambda_1. \tag{3}$$

We can get the following formula by (2):

$$\begin{aligned} &P^T C P = \Lambda \Longrightarrow \\ &P_1^T C P_1 = \Lambda_1. \end{aligned} \tag{4}$$

Equation (4) is put into (2), and we get

$$\frac{S_1^T S_1}{n-1} = \Lambda_1 = P_1^T C P_1 = P_1^T \left(\frac{S^T S}{n-1}\right) P_1 = \frac{(SP_1)^T (SP_1)}{n-1} \Longrightarrow \quad (5)$$
$$S_1 = SP_1, \quad S_1 \in R^{n\times p}.$$

Because the covariance matrix of S_1 is a diagonal matrix, it means that the components are basically independent between every two different dimensions. The process of PCA has been done.

3.2. Traffic Feature Correlation Analysis. From the view of the correlation based on statistical theory, DDoS attack traffic features reflect different statistical properties versus legitimate network traffic features. Here, we apply MCA [8, 15, 16]. This approach is based on a triangle area technique and Mahalanobis distance (MD). The triangle area technique is able to extract geometrical correlative information between every two features in an acquired network traffic record. And MD is capable of similarity measurement between every two traffic records. The analysis is presented as follows.

Assume that there is a captured network traffic record data set: $X = \{x_1, x_2, \ldots, x_n\}$. Here, $x_i^T = [f_1^i, f_2^i, \ldots, f_m^i]$, $1 \le i \le n$, where x_i represents the ith traffic record and f_j^i indicates the jth feature in the ith record. For example, f_j^i and f_k^i are a couple of features in x_i. The area of a triangle $T_{j,k}^i$ is shown as

$$T_{j,k}^i = \frac{\left(\left|f_j^i\right| \times \left|f_k^i\right|\right)}{2}, \quad (6)$$

where $1 \le i \le n$, $1 \le j, k \le m$, and $j \ne k$. Figure 2 shows the area of a triangle.

On the basis of (6), we get the area of the triangle for every two distinct features in x_i. By that analogy, the areas of these corresponding triangles between every two distinct features for each and every network traffic record of all are acquired. And a Triangle Area Matrix (TAM) has been set up. When j is equal to k, the value of $T_{j,k}^i$ is zero. So the values of these elements on the main diagonal of the matrix are zero. Because $T_{j,k}^i$ and $T_{k,j}^i$ represent the same triangle area, the values of the two are equal.

As a consequence, we draw the following conclusion: TAM^i is a symmetric matrix, and the elements of its main diagonal are zero. Here, the low triangle of TAM is chosen to convert into another vector $\mathrm{TAM}_{\mathrm{low}}^i$, and it is shown as follows:

$$\mathrm{TAM}_{\mathrm{low}}^i = \left[T_{2,1}^i \;\; T_{3,1}^i \;\; \cdots \;\; T_{m,1}^i \;\; T_{3,2}^i \;\; T_{4,2}^i \;\; \cdots \;\; T_{m,2}^i \;\; \cdots \;\; T_{m,m-1}^i\right]^T. \quad (7)$$

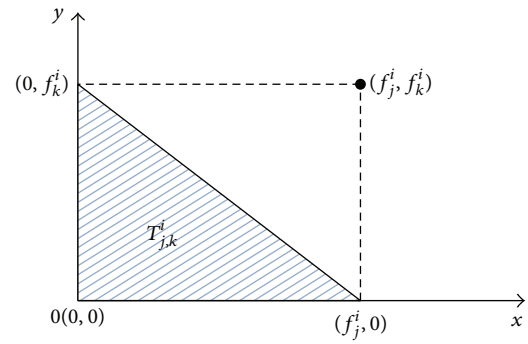

Figure 2: Sketch for the area of a triangle.

DDoS attack is detected by the application to inherent MCA of traffic features in the network environment of Big Data. The geometrical correlation between every two pairs of traffic features has changed when anomaly behaviors of DDoS attack appear on the Internet. This approach provides an important warning signal.

3.3. Attack Detection Framework. In this section, we first establish benchmark data by covariance matrix and MD. Secondly, the attack traffic detection based on MD and the selected threshold is implemented. Last but not least, we present the MDRA DDoS attack detection algorithm.

3.3.1. Benchmark Data Formation by Covariance Matrix and MD. The benchmark data is established based on normal network traffic records. It is used to compare with the fresh incoming traffic records. The inferior benchmark data can lead to the erroneous estimate that an incoming traffic record is regarded as a legitimate record.

Assume that there are t normal training traffic feature records: $X^{\mathrm{nor}} = \{x_1^{\mathrm{nor}}, x_2^{\mathrm{nor}}, \ldots, x_t^{\mathrm{nor}}\}$. We need to do two things.

(i) Computing the Covariance Matrices between the Areas of Every Two Triangles. The MCA method is applied to benchmark data formation. The acquired lower triangles are denoted as follows: $X_{\mathrm{TAM}_{\mathrm{lower}}}^{\mathrm{nor}} = \{\mathrm{TAM}_{\mathrm{lower}}^{\mathrm{nor1}}, \mathrm{TAM}_{\mathrm{lower}}^{\mathrm{nor2}}, \ldots, \mathrm{TAM}_{\mathrm{lower}}^{\mathrm{nort}}\}$. Then, we compute the covariance matrices between the areas of every two triangles; that is,

$$C_T = \begin{bmatrix} \sigma_{T_{2,1}^{\mathrm{nor}}, T_{2,1}^{\mathrm{nor}}} & \sigma_{T_{2,1}^{\mathrm{nor}}, T_{3,1}^{\mathrm{nor}}} & \cdots & \sigma_{T_{2,1}^{\mathrm{nor}}, T_{m,m-1}^{\mathrm{nor}}} \\ \sigma_{T_{3,1}^{\mathrm{nor}}, T_{2,1}^{\mathrm{nor}}} & \sigma_{T_{3,1}^{\mathrm{nor}}, T_{3,1}^{\mathrm{nor}}} & \cdots & \sigma_{T_{3,1}^{\mathrm{nor}}, T_{m,m-1}^{\mathrm{nor}}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T_{m,m-1}^{\mathrm{nor}}, T_{2,1}^{\mathrm{nor}}} & \sigma_{T_{m,m-1}^{\mathrm{nor}}, T_{3,1}^{\mathrm{nor}}} & \cdots & \sigma_{T_{m,m-1}^{\mathrm{nor}}, T_{m,m-1}^{\mathrm{nor}}} \end{bmatrix}. \quad (8)$$

In this formula, the covariance between every two arbitrary elements in $\mathrm{TAM}_{\mathrm{lower}}$ is defined as follows:

$$\sigma_{(T_{j,k}^{\mathrm{nor}}, T_{u,v}^{\mathrm{nor}})} = \frac{1}{t-1} \sum_{i=1}^{t} \left(T_{j,k}^{\mathrm{nor},i} - \mu_{T_{j,k}^{\mathrm{nor}}}\right)\left(T_{u,v}^{\mathrm{nor},i} - \mu_{T_{u,v}^{\mathrm{nor}}}\right), \quad (9)$$

where the mean of the (j,k)th elements and the mean of the (u,v)th elements of TAMs for t normal training traffic records are, respectively, defined as

$$\mu_{T_{j,k}^{\text{nor}}} = \frac{1}{t}\sum_{i=1}^{t} T_{j,k}^{\text{nor},i}, \tag{10}$$

$$\mu_{T_{u,v}^{\text{nor}}} = \frac{1}{t}\sum_{i=1}^{t} T_{u,v}^{\text{nor},i}. \tag{11}$$

(ii) Computing the MD between Every Two TAMs of Traffic Records. The covariance distance of data is signified by MD. MD is an effective approach to compute the similarity of the two unknown sample sets. The difference between MD and Euclidean Distance (ED) is that the relations between all kinds of characters are considered and that MD is not relevant to the scale of the measurement.

The MD between the normal training records and their expectation and the MD between the fresh captured traffic record and the expectation of normal training records are shown by the following formulas:

$$\text{MD}^{\text{nor}i} = \sqrt{\left(\text{TAM}_{\text{lower}}^{\text{nor}i} - \overline{\text{TAM}_{\text{lower}}^{\text{nor}}}\right)^{T} \text{cov}^{-1} \left(\text{TAM}_{\text{lower}}^{\text{nor}i} - \overline{\text{TAM}_{\text{lower}}^{\text{nor}}}\right)}, \tag{12}$$

$$\text{MD}^{\text{fresh}} = \sqrt{\left(\text{TAM}_{\text{lower}}^{\text{fresh}} - \overline{\text{TAM}_{\text{lower}}^{\text{nor}}}\right)^{T} \text{cov}^{-1} \left(\text{TAM}_{\text{lower}}^{\text{fresh}} - \overline{\text{TAM}_{\text{lower}}^{\text{nor}}}\right)}. \tag{13}$$

Moreover, the expectation of $\text{TAM}_{\text{lower}}^{\text{nor}}$ for the t normal training records is shown as follows:

$$\overline{\text{TAM}_{\text{lower}}^{\text{nor}}} = \frac{1}{t}\sum_{i=1}^{t} \text{TAM}_{\text{lower}}^{\text{nor}i}. \tag{14}$$

3.3.2. Attack Detection Standard Based on MD and Threshold. For DDoS attack detection, we set a threshold value to distinguish DDoS anomaly traffic from the normal traffic feature. Next, we give a formula [8] about the threshold value:

$$\text{Threshold} = \mu + \sigma * \alpha, \tag{15}$$

where μ was shown by (10) or (11) and σ is shown as follows:

$$\sigma = \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\left(\text{MD}^{\text{nor}i} - \overline{\text{MD}^{\text{nor}}}\right)^2}, \tag{16}$$

$$\overline{\text{MD}^{\text{nor}}} = \frac{1}{t}\sum_{i=1}^{t} \text{MD}^{\text{nor}i}. \tag{17}$$

In order to conform to the normal distribution [8], the range of the σ value is set from 1 to 3 with the increment of 0.2 in this paper. Then, the standard of DDoS attack detection is obtained. An attack behavior is considered when the MD between a fresh acquired traffic record and the expectation of normal training records is greater than the threshold.

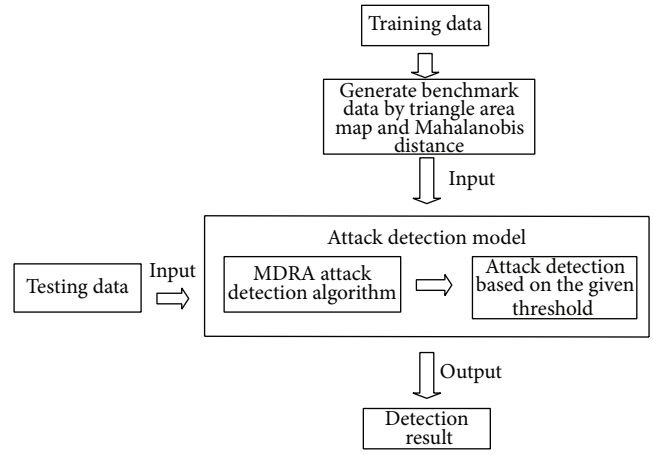

Figure 3: Flowchart of attack detection.

3.3.3. A MDRA DDoS Attack Detection Algorithm. Tan et al. [8] proposed the algorithm to use for normal profile generation based on triangle-area and MCA and to use for attack detection based on MD. By evaluation and comparison with some state-of-the-art approaches, it is easy to find that the previous attack detection method and its system have some advantages in detection performance, Detection Rate (DR) and accuracy rate. However, in Big Data of cyberspace security, especially when the network attack behaviors of large traffic are growing increasingly, detection efficiency, resource consumption, and computing complexity need be taken adequately into account for attack detection. For the reasons given above, we propose the MDRA algorithm to detect efficiently the network anomaly traffic. Algorithm 1 depicts the procedures of the algorithm for DDoS attack detection metric based on MDRA in detail.

4. Experiments

In this section, we discuss how to apply our algorithm in detecting efficiently the DDoS attack traffic. The flowchart of attack detection is shown in Figure 3.

Firstly, we present the data set used in our experiments and the data pretreatment approach to serve our experiments. Then, the experimental results are got to evaluate the algorithm performance. Finally, we make comparisons with the previous unoptimized approach in terms of time cost, resource consumption, and computing complexity.

The computer environment to run our experiments is shown in Table 1.

Next, we describe our experiments in detail.

4.1. Data Set and Pretreatment. In this paper, we use the famous Knowledge Discovery and Data Mining (KDD) Cup 1999 data set [17–21] as our novel algorithm verification. We have to admit that this data set has some shortages, but it is still uniquely public and relatively credible labeled benchmark data set so far. This data set has been widely applied to researching and evaluating network intrusion detection methods [22, 23].

(1) Input a set of training data of normal network traffic records $X^{\text{nor}} = \{x_1^{\text{nor}}, x_2^{\text{nor}}, \ldots, x_t^{\text{nor}}\}$, where, $x_i^{\text{nor}} = [f_1^i, f_2^i, \ldots, f_m^i], 1 \le i \le n$.
(2) Extract the principal components of X^{nor} to reach 70% for the accumulative contribution rate based on PCA, and obtain the principal component data set $X^{P\text{nor}}$.
(3) Calculate $\text{TAM}_{\text{lower}}^{P\text{nor}i}$ and $\overline{\text{TAM}_{\text{lower}}^{P\text{nor}}}$ of $X^{P\text{nor}}$.
(4) Calculate the covariance matrices between the areas of every two triangles $T^{P\text{nor}}$ in $X^{P\text{nor}}$.
(5) **for** $i = 1$ to t **do**
(6) Input $\text{TAM}_{\text{lower}}^{P\text{nor}i}$ and $\overline{\text{TAM}_{\text{lower}}^{P\text{nor}}}$
(7) Calculate $\text{MD}^{P\text{nor}i}$ between $\text{TAM}_{\text{lower}}^{P\text{nor}i}$ and $\overline{\text{TAM}_{\text{lower}}^{P\text{nor}}}$
(8) Output $\text{MD}^{P\text{nor}i}$
(9) **end for**
(10) Calculate μ by $\text{MD}^{P\text{nor}i}$.
(11) Calculate σ by $\text{MD}^{P\text{nor}i}$ and μ.
(12) Input a fresh incoming traffic record x^{fresh}.
(13) Reduce the dimensions of the features for x^{fresh} based on PCA, then get the records which include the principal components $x^{P\text{fresh}}$.
(14) Calculate $\text{TAM}_{\text{lower}}^{P\text{fresh}}$ of $x^{P\text{fresh}}$.
(15) Calculate $\text{MD}^{P\text{fresh}}$ between $\text{TAM}_{\text{lower}}^{P\text{fresh}}$ and $\overline{\text{TAM}_{\text{lower}}^{P\text{nor}}}$.
(16) Input the threshold value α.
(17) **If** $(\mu - \sigma * \alpha) \le \text{MD}^{P\text{fresh}} \le (\mu + \sigma * \alpha)$ **then**
(18) **return** Normal
(19) **else**
(20) **return** Attack
(21) **end if**

Algorithm 1: Algorithm for DDoS attack detection based on MDRA.

Table 1: Computer environment to run our experiments.

CPU	Memory	Hard disk	OS	MATLAB
Intel® Xeon® CPU E5-2640 v2 @2.00 GHz 2.00 GHz (2 processors)	32 GB	2 TB	Windows Server 2008 R2 Enterprise	R2013a (8.1.0.604) 64-bit (win64)

Table 2: Data sets used in our experiments.

Category	Training data set (10%)	Testing data set (corrected)
Normal	97278	60593
DoS	391458	229853

KDD CUP 1999 data set comprises about five million network records and provides a training subset of 10 percent of the network records and a testing subset. It covers four main categories of attack, that is, DoS, R2L, U2R, and Probing. Here, we use these records labeled as "normal" in the abovementioned training subset to construct our benchmark data and employ this testing subset "corrected" to verify the validity and efficiency of our algorithm. In this paper, we choose DoS network attack as our algorithm evaluation and comparison with the previous approaches. The data sets used in our experiments are shown in Table 2. The data pretreatment procedure is shown as follows.

Firstly, for each network traffic record, it includes the information that has been separated into 41 features plus 1 class label [24] in this data set. In our experiments, we need to get all numeric data for 41 features of every record. However, there are 3 nonnumeric features in all features, and these are protocol_type, service, and flag. They must be transformed into numeric type. The type conversion is achieved according to Table 3, where we emphatically analyze the pretreatment process with reference to the feature "service." The analysis process is as follows.

There are 70 kinds of network service types in the "service" feature; however, some of them rarely appear or never appear. For these features, we can ignore them completely. Among the 494021 records in the training subset of 10 percent, we find that the top three network service types, respectively, are ecr_i, private, and http by counting and sorting, and their ratios, respectively, are 56.96%, 22.45%, and 13.01%. The sum of all the other types accounts merely for 7.58%. The ratios of the top four types in "service" feature are shown in Table 4.

Secondly, among the 41 features of these records labeled as "normal" in the training subset of 10 percent, there are three invalid features (i.e., wrong_fragment, num_outbound_cmds, and is_hot_login) by PCA. This is because all the values of the three features are zero. Therefore, we get rid of the three features in our experiments.

Last but not least, we extract the principal components according to the rate of accumulative contribution based on PCA algorithm. As a general rule, we set the value of the rate of accumulative contribution to be equal to or to be greater than 50% to extract important features from the chosen data set [6]. In order to obtain the more important principal components, the value of the rate of accumulative contribution is set to 70% in our experiments. These principal components extracted in the 41 features are listed in Table 5.

Table 3: Type conversion for numbers 2, 3, and 4 of 41 features.

Number	Feature name	Type setting 1	Type setting 2	Type setting 3	Type setting 4
2	protocol_type	TCP = 1	UDP = 2	ICMP = 3	/
3	service	ecr_i = 1	private = 2	http = 3	others = 0
4	flag	SF = 1	others = 0	/	/

Table 4: Top four types in "service" feature.

Type name	The percentage
ecr_i	56.96%
private	22.45%
http	13.01%
others	7.58%

Table 5: The principal components extracted in 41 features.

Number	Feature name of principal component
1	duration
2	protocol_type
3	service
4	flag
5	src_bytes
6	dst_bytes
7	land
8	urgent
9	hot
10	num_failed_logins
11	logged_in
12	num_compromised

4.2. Experimental Results. Our experiments aim at showing exhaustive and comparable results between the DDoS attack detection method based on MCA and the method based on MDRA. These results prove that the latter is superior to the former.

In order to estimate the advantage of our method, it is indispensable to establish some evaluating indications. Here, we present four formulae to evaluate our algorithm, and they are Precision, TNR, FPR, and DR [11]. The formulae are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{18}$$

$$\text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \tag{19}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{20}$$

$$\text{DR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{21}$$

where

(i) TP (True Positive) is the number of attacks correctly classified as attacks;

(ii) FP (False Positive) is the number of normal records incorrectly classified as attacks;

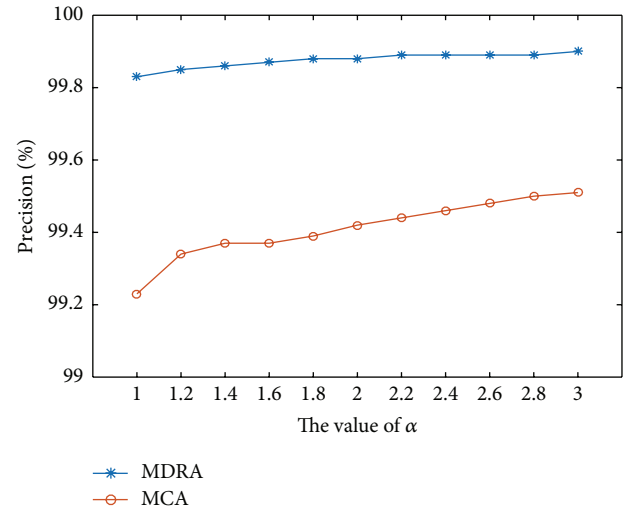

Figure 4: Precision for comparing detection methods based on MDRA and MCA.

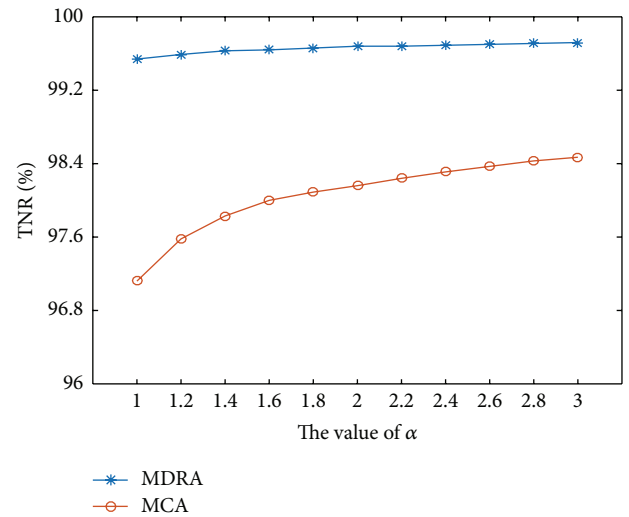

Figure 5: TNR for comparing detection methods based on MDRA and MCA.

(iii) TN (True Negative) is the number of normal records correctly classified as normal records;

(iv) FN (False Negative) is the number of attacks incorrectly classified as normal records.

Table 6 shows all results of TP, FP, TN, and FN for α that is set from 1 to 3 with the increment of 0.2 when we use DDoS attack detection methods based on MDRA and MCA.

Here, the detection results of precision and TNR with the different α values are shown in Figures 4 and 5.

Table 6: Results of TP, FP, TN and FN based on MDRA and MCA.

α	Indicators based on MDRA				Indicators based on MCA			
	TP	FP	TN	FN	TP	FP	TN	FN
$\alpha = 1$	166299	278	60315	63554	223587	1743	58850	6266
$\alpha = 1.2$	166299	249	60344	63554	221873	1469	59124	7980
$\alpha = 1.4$	166292	227	60366	63561	206504	1313	59280	23349
$\alpha = 1.6$	166289	217	60376	63564	191190	1214	59379	38663
$\alpha = 1.8$	166289	204	60389	63564	190394	1159	59434	39459
$\alpha = 2$	166289	194	60399	63564	190342	1115	59478	39511
$\alpha = 2.2$	166289	191	60402	63564	190311	1065	59528	39542
$\alpha = 2.4$	166289	188	60405	63564	190277	1027	59566	39576
$\alpha = 2.6$	166282	180	60413	63571	190254	988	59605	39599
$\alpha = 2.8$	166282	176	60417	63571	190230	953	59640	39623
$\alpha = 3$	166282	172	60421	63571	190199	927	59666	39654

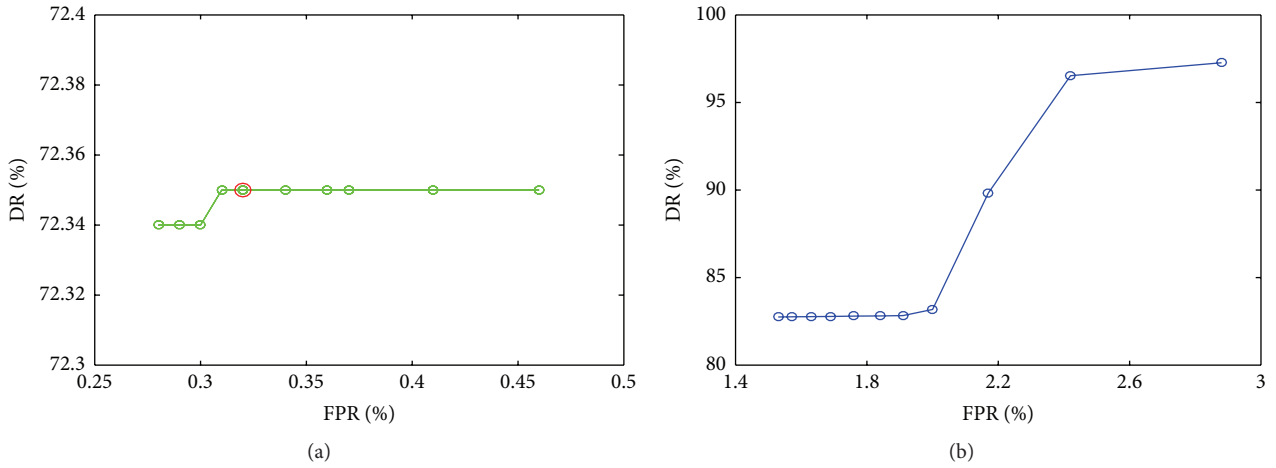

Figure 6: (a) ROC for analyzing detection method based on MDRA. (b) ROC for analyzing detection method based on MCA.

In Figure 4, it is not hard to find that when the value of α gradually increases from 1 to 3 with the increment of 0.2, the precision of attack detection method based on MDRA is superior to the counterpart based on MCA, and the former is about 0.4 to 0.6 percent higher than the latter.

In Figure 5, similarly, we find that the TNR of our detection method is completely superior to another one with the progressive increment of α, and the former is about 1.2 to 2.4 percent higher than the latter.

In addition, the relationship between DR and FPR is frequently used to evaluate the detection performance by the Receiver Operating Characteristic (ROC) curve. The ROC curve is obtained by setting different thresholds, and there is a tradeoff between the DR and FPR [25]. The ROC curves of the comparisons about the two detection methods are shown in Figure 6. In Figures 6(a) and 6(b), the two ROC curves that are used to analyze attack detection performance based on our method and another one show the growing tendency. In Figure 6(a), the ROC curve of our method climbs gradually from 72.34% to 72.35% for DR, and it reflects that the change of DR with different α values is fairly small. Likewise, in Figure 6(b), this change is relatively large, and the ROC curve jumps dramatically from 83.18% to 89.84%. However, in Big Data, we pay more attention to instantaneity, time cost, resource consumption, and computational complexity of attack detection. Therefore, a shade of discrepancy of DR could be ignored. At this point, our method has the vast majority of advantages in comparison to other methods. The discussion about this topic will be opened up in the next section.

4.3. Results Comparisons in terms of Time Cost and Resource Consumption. Here, we emphatically analyze time cost and memory resource consumption based on MDRA and MCA.

On the one hand, our detection mechanism is superior to another one based on triangle-area and MCA proposed by Tan et al. in time cost. In our experimental environment, we employ this server which has two CPUs and where every CPU has 16 cores. When we ran the abovementioned experimental data, one of two CPUs opened and 16 cores of this CPU would gradually load to its full capacity. At the moment, the comparing results in CPU time of running the experimental data based on our detection method and the other one are shown in Figure 7. However, in the same

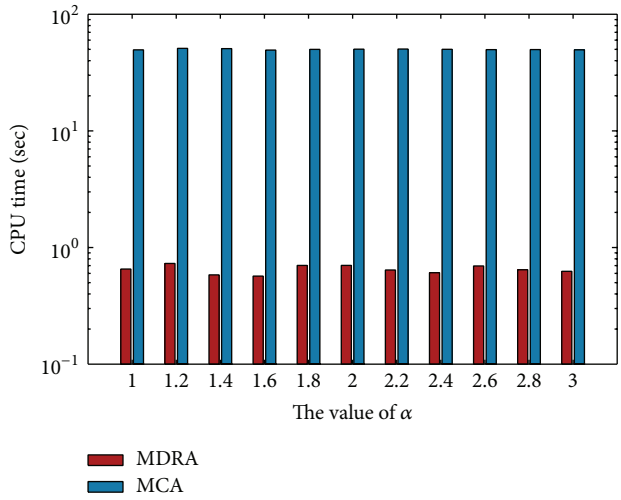

Figure 7: Comparing results in CPU time based on MDRA and MCA.

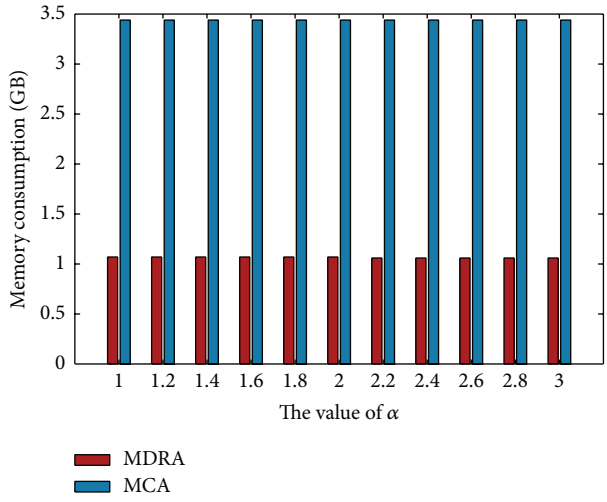

Figure 8: Comparing results in memory consumption based on MDRA and MCA.

experimental environment, the CPU time of the detection method proposed by Tan et al. is about 80 times as long as ours, and our CPU time is less than one second.

On the other hand, in terms of memory consumption, our detection mechanism is also a cut above the rest of the method proposed by Tan et al. This is because the memory occupied by our detection method in the experiments takes up less than 1 GB; however, another one needs memory space of more than 3 GB. In the same experimental environment, the occupied memory space in the detection method proposed by Tan et al. is more than 3 times as long as ours. The comparing results in memory consumption of running the experimental data are shown in Figure 8.

To sum up, our detection method can be perfectly applied in real-time DDoS attack detection under the environment of vast amount of network traffic in Big Data.

4.4. Computing Complexity Analysis. In this section, we analyze the computing complexity of our detection method. Because the previous method based on MCA has the computing complexity of $O(m^2)$ and m is a fixed number, the overall computing complexity is equal to $O(1)$ [8]. However, our detection mechanism based on MDRA uses the similar computational principle. What is more, the fixed feature dimensionality m after reducing dimensionality in our method is one-third of the previous method based on MCA. Hence, the computing complexity of our method is also equal to $O(1)$. At this point, our detection mechanism is equal to or is better than the other methods in [6, 8, 16].

5. Conclusion

In this paper, we present a real-time DDoS attack detection mechanism based on the MDRA algorithm in Big Data. Compared with previous methods, the experimental results demonstrate that our solution has the better effectiveness and efficiency to distinguish attack traffic from vast amount of normal network traffic on the aspects of precision rate, TNR, time cost, memory resource consumption, and computing complexity.

Competing Interests

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

Acknowledgments

This work was supported by the International Science and Technology Cooperation Project of China (2013DFE13130).

References

[1] P. J. Criscuolo, *Distributed Denial of Service: Trin00, Tribe Flood Network, Tribe Flood Network 2000, and Stacheldraht CIAC-2319*, Lawrence Livermore National Laboratory, 2000.

[2] M. Kim, H. Na, and K. Chae, "A combined data mining approach for DDoS attack detection," in *Information Networking. Networking Technologies for Broadband and Mobile Networks*, vol. 3090 of *Lecture Notes in Computer Science*, pp. 943–950, Springer, Berlin, Germany, 2004.

[3] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry, "Non-Gaussian and long memory statistical characterizations for Internet traffic with anomalies," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, no. 1, pp. 56–70, 2007.

[4] K. Lee, J. Kim, K. H. Kwon, Y. Han, and S. Kim, "DDoS attack detection method using cluster analysis," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1659–1665, 2008.

[5] H. V. Nguyen and Y. Choi, "Proactive detection of DDoS attacks utilizing k-NN classifier in an anti-DDoS framework," *World Academy of Science, Engineering and Technology, International Science Index*, vol. 4, no. 3, pp. 247–252, 2010.

[6] C.-F. Tsai and C.-Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern Recognition*, vol. 43, no. 1, pp. 222–229, 2010.

[7] A. Bhange, A. Syad, and S. Singh Thakur, "DDoS attacks impact on network traffic and its detection approach," *International Journal of Computer Applications*, vol. 40, no. 11, pp. 36–40, 2012.

[8] Z. Y. Tan, A. Jamdagni, X. J. He, P. Nanda, and R. P. Liu, "A system for denial-of-service attack detection based on multivariate correlation analysis," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 2, pp. 447–456, 2014.

[9] J. Luo, X. Yang, J. Wang, J. Xu, J. Sun, and K. Long, "On a mathematical model for low-rate shrew DDoS," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1069–1083, 2014.

[10] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[11] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[12] G. Liu, Z. Yi, and S. Yang, "A hierarchical intrusion detection model based on the PCA neural networks," *Neurocomputing*, vol. 70, no. 7–9, pp. 1561–1568, 2007.

[13] Y. Kanda, K. Fukuda, and T. Sugawara, "Evaluation of anomaly detection based on sketch and PCA," in *Proceedings of the 53rd IEEE Global Telecommunications Conference (GLOBECOM '10)*, pp. 1–5, IEEE, Miami, Fla, USA, December 2010.

[14] Y. Zhang and L. Wu, "An MR brain images classifier via principal component analysis and kernel support vector machine," *Progress in Electromagnetics Research*, vol. 130, pp. 369–388, 2012.

[15] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "Denial-of-service attack detection based on multivariate correlation analysis," in *Neural Information Processing*, pp. 756–765, Springer, Berlin, Germany, 2011.

[16] Z. Tan, A. Jamdagni, X. J. He, P. Nanda, and R. P. Liu, "Triangle-area-based multivariate correlation analysis for effective denial-of-service attack detection," in *Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom '12)*, pp. 33–40, IEEE, Liverpool, UK, June 2012.

[17] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, "Cost-based modeling for fraud and intrusion detection: results from the JAM project," in *Proceedings of the DARPA information survivability conference and exposition (DISCEX '00)*, pp. 130–144, Hilton Head, SC, USA, 2000.

[18] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA off-line intrusion detection evaluation," *Computer Networks*, vol. 34, no. 4, pp. 579–595, 2000.

[19] J. McHugh, "Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 262–294, 2000.

[20] S. Stofo, *The Third International Knowledge Discovery and Data Mining Tools Competition*, The University of California, 2002, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[21] S. Mukkamala, A. H. Sung, and A. Abraham, "Intrusion detection using an ensemble of intelligent paradigms," *Journal of Network and Computer Applications*, vol. 28, no. 2, pp. 167–182, 2005.

[22] K.-C. Khor, C.-Y. Ting, and S. Phon-Amnuaisuk, "A cascaded classifier approach for improving detection rates on rare attack categories in network intrusion detection," *Applied Intelligence*, vol. 36, no. 2, pp. 320–329, 2012.

[23] P. Prasenna, A. V. T. Raghav Ramana, R. Krishna Kumar, and A. Devanbu, "Network programming and mining classifier for intrusion detection using probability classification," in *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME '12)*, pp. 204–209, IEEE, Salem, Tamilnadu, March 2012.

[24] C. Bae, W.-C. Yeh, M. A. M. 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.

[25] W. Wang, X. Zhang, S. Gombault, and S. J. Knapskog, "Attribute normalization in network intrusion detection," in *Proceedings of the 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN '09)*, pp. 448–453, December 2009.