Denial of Service Detection System using KDD Cup Dataset

Amith B N
Dept. CSE
TJIT, Bangalore

Suma R
Asst. professor, Dept CSE
TJIT, Bangalore

***Abstract:*-Denial of service attack is a threat for computing systems such as web server, database server, and cloud computing server. In order to protect these systems a denial of service attack detection methods are used. The multivariate correlation analysis is one of the denial of service detection method, in which the network traffic is considered by taking the geometrical correlation between network traffic. The multivariate correlation analysis also provides the anomaly based detection in attacks. This helps to identify both known and unknown denial of service attacks by identifying the pattern of network traffic. In order to increase the speed up of multivariate correlation analysis the triangle area based technique is used. In this the detection system is evaluated using the KDD cup data set for both normalized and non-normalized data.**

***Keyword: Network traffic, correlation*.**

I INTRODUCTION

Denial of service (DoS) attack is an attack in which systems are unavailable to the intended users and it is one type of aggressive behavior for online service. The symptoms for denial of service attacks are, slow network performance, unavailability of websites, inability to access websites, disconnection of internet connection. The DoS attack strictly degrade the availability of a victim, which can be host, a router or an entire network. They impose intensive computation task to the victim by exploiting its system vulnerability or flooding huge amount of useless packet. The victim can be forced out service from a few minutes to even several days. This cause serious damage to the services running on the victim. The effective detection of DoS is essential to protection of online service. The detection system, monitor the network traffic transmitted over the network. This ensures protected online services by monitoring attacks and ensures that the server can be dedicated themselves to provide the quality of service with minimum delay in response.The network based detection system can be classified into two categories, Misuse based detection system, and Anomaly based detection system

Misuse based detection system, detect the attack by monitoring network activates, gathering the information and compare with the large data base of signature. It will always look only for the documented attack. The misuse attack detection will work only when the database is strong. Anomaly based detection system, the administrator defines normal state of the network or the baseline. If the state of network changes as defined by the administrator, there will be an intruder attack. At this moment the intruder will be detected and will be blocked. The anomaly based detection technique shows more promising in detecting intrusion that exploit previously unknown system vulnerability. It is not constrained in the network security, due to the fact that the profiles of legitimate behaviors are developed based on techniques such as statistical analysis, machine learning, and data mining. It will suffer from high false positive rate. The false positive rate is due to negligence of correlation between the attribute. In order to overcome from this, a flow correlation coefficient among suspicious flows are considered. A covariance matrix based approach was designed to mine the correlation for sequential samples. This approach improves the detection accuracy and can only label an entire group of observed samples as legitimate traffic but not the individuals in the group. To deal with the above problem, a triangle based approach to generate better distinguish between the features. This approach have dependency on prior knowledge of the malicious behavior. To overcome from this a geometrical structure based analysis technique, where mahalanobis distance was used to extract the correlation between the selected packet payload features. The multivariate explores relationship between the variables. The word multivariate simply means involving many variables instead of one or two. The multivariate correlation summarize the strength of the linear relationship between each pair of variables using the correlation. It identifies dependencies, outliers and clustering using scatter plotmatrix. This MCA approach employs triangle area for extracting the correlative information between the features within an observed data object.The Denial of service attack detection system employs the principles of multivariate correlation analysis and anomaly based detection. The detection system with capabilities of accurate characterization for traffic behaviors and detection of known and unknown attacks respectively. A triangle area technique is developedto speed up andto enhance the process of MCA. A statistical normalization technique is used to eliminate the bias from raw data. The denial of service detection system is evaluated using KDD Cup 99 dataset.Since 1999, KDD'99 has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared and is built based on the data captured real time network traffic. It is about 4 gigabytes of compressed data of 7 weeks of network traffic, which can be processed into large connection records,. The three weeks of test data have around large connection records. KDD training dataset consists of approximately 4,900,00 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. Denial of Service Attack is an attack in which the attacker makes some computing or memory resource too busy.

II. RELATED WORK

In paper [1] the author discussed about, with growing Internet connectivity comes growing opportunities for attackers to illicitly access computers over the network. The problem of detecting s attacks is termed network intrusion. System for detecting network intruder in the real time by monitoring network in which the intruder traffic transmit. The system design provides the high speed monitoring, real time notifications, policy and extensibility.

In paper [2] authors discussed about, the anomaly based detection is based on defining the network behavior. The network behavior is accordance with the predefined behavior, then it is accepted or triggers the event in the anomaly detection. The accepted behavior is learned by the network administrator. The IDS engine must able to process the protocol, through this protocol analysis it helps to increase the rule set provides the less false positive alarm. The major drawback of detection system is defining the rule set. The rule defining process is affected by the various protocols. Detection to occur correctly, the detailed knowledge about accepted network behavior need to be developed by the administration.

In paper [3] the author discussed about, a model of real time intruder detection expert system capable of detecting break in, penetration, and other form of computer abuse are described. In this model is based on the security violation can be detected by monitoring a system audits records for abnormal pattern of system usage. The break-in will happens when someone break into system might generate an abnormally high rate of failure with respect to a single account.

In paper [4] authors discussed about, distributed denial of service attack (DDOS) DDoS attack detection model based on data mining algorithm. Cluster algorithm and Apriori association algorithm used to extracts network traffic and network packet protocol status. The threshold is set for detection model. DDOS brings a very serious threat to send to the scalability of the internet.

In paper [5] authors discussed about,Fuzzy logic is dealing with reasoning that is approximate rather than precise. The fuzzy data mining technique is used to extract the pattern that represents normal behavior for intrusion detection.

In paper [6] authors discussed about, Network intrusion detection aims at distinguishing the attacks on the internet from normal se of the internet. Due to the variety of network behavior and the rapid development of attack fashion, it is necessary to develop fast machine-learning-based intrusion detection algorithms with high detection rates and low false –alarm rates.

In paper [7] authors discussed about, the parametric methods to detect network anomalies using only aggregate traffic statistics in contrast to other works requiring flow separation, even when the anomaly is small fraction of the total traffic. By adopting simple statistical models for anomalous and background traffic in the time-domain, can estimate model parameters in real-time thus obviating the need for a long training phase or manual parameter tuning. The bivariate parameter detection mechanism uses a sequential probability ratio test, for control over the false positive rate while examining the trade-off between detection time and the strength of anomaly. Additionally it sees both traffic-rate and yielding a bivariate model that eliminates most false positives.

In paper [8] authors discussed about, the collaborative detection detects flooding attacks at traffic flow level.The system is suitable for t implementation over the core networks operated by Internet service providers (ISP). At the early stage of a DDoS attack, some traffic fluctuations are detectable at Internet routers or at gateways of edge networks. A distributed change-point detection (DCD) architecture using c (CAT). The idea is to detect abrupt traffic changes across multiple network domains at the earliest time.

In paper [9] the authors discussed about, a hybrid learning model based on the triangle area based nearest neighbors (TANN) in order to detect attacks more effectively. In TANN, thek-means clustering is firstly used to obtain cluster centers corresponding to the attack classes.

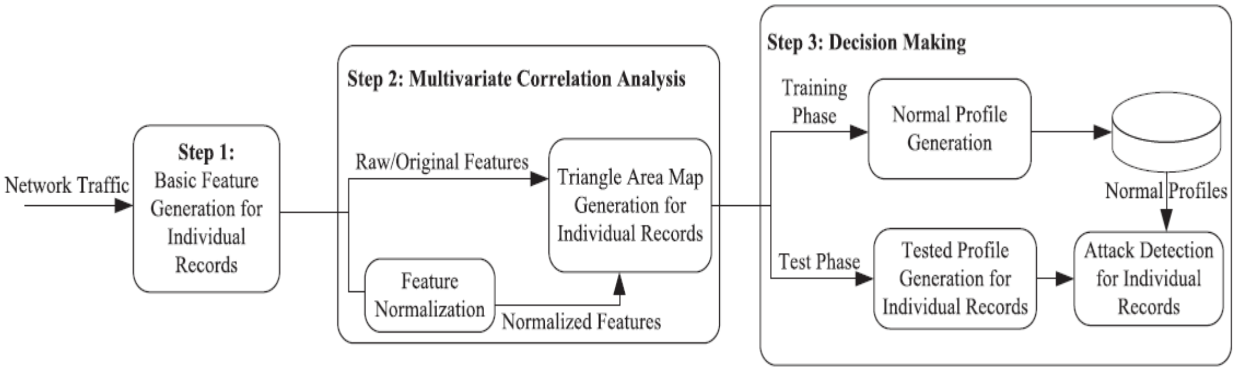
Network Traffic
Step 1:
Basic Feature Generation for Individual Records
Step 2: Multivariate Correlation Analysis
Raw/Original Features
Triangle Area Map Generation for Individual Records
Feature Normalization
Normalized Features
Step 3: Decision Making
Training Phase
Normal Profile Generation
Normal Profiles
Test Phase
Tested Profile Generation for Individual Records
Attack Detection for Individual Records

Fig 1: DOS Detection System

respectively. Then, the triangle area by two cluster centerswith one data from the given dataset is calculated and formed a new feature signature of the data. Finally, the*k*-NN classifier issued to classify similar attacks based on the new feature represented by triangle areas. The triangle area based nearest neighbors approach is used to detect the attacks more effectively. TANN can effectively detect intrusion attacks and provide higher accuracy and detection rates, and the lower false alarm rate.

In paper [10] authors discussed about, Data preprocessing, attribute normalization is essential to detection performance. In many anomaly detection methods do not normalize attributes before training and detection. Few methods consider to normalize the attributes but the question of which normalization method is more effective still remains.

III. SYSTEM ARCHITECTURE

As shown in fig1 DOS detection system includes the following explanationIn step 1, the basic features are generated from the network traffic for individual records. So that the internal network and the server are safe from the network threat. This provides the information for detector to provide protection which is the best fit for the targeted internal network because legitimate traffic profiles used by the detectors.

In step 2, multivariate correlation analysis, in which the triangle area map generation module is used to extract the correlations between two distinct features within each traffic record coming from the feature normalization moduleand the raw features. The occurrence of network intrusions cause changes to correlations so that the changes can be used to identify the intrusion activities.

The triangle area maps are used to store the correlations, namely triangle areas, are then used to replace the original basic features or the normalized features to represent the traffic records. This helps higher discriminative information to differentiate between legitimate and illegitimate traffic records.

In step 3,decision making, the anomaly based detection mechanism is adopted. It facilitates the detection of any DoS attacks without requiring any attack relevant knowledge. Furthermore, the labor-intensive attack analysis and the frequent update of the attack signature database in the case of misuse-based detection are avoided.

Two phases are involved in decision making.

- Training phase
- Test phase

In training phase,the normal profile generation module is operated in the training phase to generate profiles for various types of legitimate traffic records, and the generated normal profiles are stored in a database.

In test phase,tested profile generation module is used in the test phase to build profiles for individual observed traffic records.

The tested profiles are handed over to the attack detection module, which compares the individual tested profiles with the respective stored normal profiles. A threshold-based classifier is employed in the attack detection module to distinguish DoS attacks from legitimate traffic.

IV. SAMPLE-BY-SAMPLE DETECTION

Systematically proved that the group-baseddetection mechanism maintained a higher probability inclassifying a group of sequential network traffic samplesthan the sample-by-sample detection mechanism.Whereas, the proof was based on an assumption thatthe samples in a tested group were all from the samedistribution (class). This restricts the applications of thegroup-based detection to limited scenarios, because attacksoccur unpredictably in general and it is difficult toobtain a group of sequential samples only from the samedistribution.To remove this restriction, system in this paperinvestigates traffic samples individually. This offers benefitsthat are not found in the group-based detectionmechanism.

Attacks can be detected ina prompt manner in comparison with the group-baseddetection mechanism.Intrusive traffic samples canbe labeled individually. The probability of correctlyclassifying a sample into its population is higherthan the one achieved using the group based detectionmechanism in a general network scenario.

V. MULTIVARIATE CORRELATION ANALYSIS

DoS attack traffic behaves differently from the legitimate network traffic, and the behavior of network traffic is reflected by its statistical properties. To describe these statistical properties, present a novel Multivariate Correlation Analysis (MCA).This MCA approach employs triangle area for extracting the correlative information between the features within an observed data object (i.e., a traffic record).where X1,X2,X3.......Xn are the attributes.

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

The covariance between the two attributes will be given by

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

where μ is the mean, which is given by

$$\mu_i = \mathrm{E}(X_i)$$

MCA approach supplies with the following benefits to data analysis.

First, it does not require the knowledge of historictraffic in performing analysis.

Second, unlike the Covariance matrix approaches which isvulnerable to linear change of all features, triangle-area-based MCA withstands the problem.

Third,It provides characterization for individual network trafficrecords rather than model network traffic behavior ofa group of network traffic records. This results in lowerlatency in decision making and enable sample-by-sampledetection.

Fourth, the correlations between distinct pairsof features are revealed through the geometrical structureanalysis. Changes of these structures may occurwhen anomaly behaviors appear in the network. Thisprovides an important signal to trigger an alert.

VI. DETECTION SYSTEM

Detection Mechanism include threshold based anomaly detector, their normal profiles are generated using purely legitimate network traffic records and it is used for future comparisons with new incoming investigated traffic record.

A. Normal profile Generation

The triangle area based MCA approach is applied to analyze the record. Assume that there is a set of k the training records are

$$X_{\text{normal}}=\{x_{1\text{normal}},x_{2\text{normal},}\ \ldots,x_{k\text{normal}}\}$$

A.Mahalanobis Distance

Mahalanobis distance (MD) used to extract the correlations between the selected packet payload features**.** It works with network packet payloads.Mahalanobis distance is adopted to measure the dissimilarity between traffic records**.** Attack detection based on Mahalanobis distance.

$$D(x) = \sqrt{(x - \mu)^t S - (x\text{-}\mu)}$$

B.Threshold Selection

In this module is todistinguish DoS attacks from legitimate traffic**.** The threshold given is used to differentiate attack traffic from the legitimate one.Normal profile is greater than the threshold, it will be considered as an attack.It is powered by the triangle-areabasedMCA technique and the anomaly-based detection technique.

B.Attack detection

Attack detection system that uses multivariate correlation analysis (MCA) for accurate network trafficIt characterization by extracting the geometrical correlations between network traffic features**.**It compares the individual tested profiles with the respective stored normal profiles.

VII. CONCLUSION

A MCA-based DoS attack detection system which is powered by the triangle-area based MCA technique and the anomaly-based detection technique. The former technique extracts the geometrical correlations hidden in individual pairs of two distinct features within each network traffic record, and offers more accurate characterization for network traffic behaviors. The latter technique facilitates our system to be able to distinguish both known and unknown DoS attacks from legitimate network traffic. The future work is applying classification mechanisms and try to work on the real world data.

REFERENCES

[1] V. Paxson, “Bro: A System for Detecting Network Intruders in Realtime,”Computer Networks, vol. 31, pp. 2435-2463, 1999.

[2] P. Garca-Teodoro, J. Daz-Verdejo, G. Maci-Fernndez, and E. Vzquez, “Anomaly-based Network Intrusion Detection: Techniques, Systems and Challenges,” Computers & Security, vol. 28, pp. 18-28, 2009.

[3] D. E. Denning, “An Intrusion-detection Model,” IEEE Transactions on Software Engineering, pp. 222-232, 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] A. Tajbakhsh, M. Rahmati, and A. Mirzaei, “Intrusion detection using fuzzy association rules,” Applied Soft Computing, vol. 9, no. 2, pp. 462-469, 2009

[6]W. Hu, W. Hu, and S. Maybank, “AdaBoost-Based Algorithm for Network Intrusion Detection,” Trans. Sys. Man Cyber. Part B, vol. 38, no. 2, pp. 577-583, 2008.

[7] G. Thatte, U. Mitra, and J. Heidemann, “Parametric Methods for Anomaly Detection in Aggregate Traffic,” Networking, IEEE/ACM Transactions on, vol. 19, no. 2, pp. 512-525, 2011.

[8] C. Yu, H. Kai, and K. Wei-Shinn, “Collaborative Detection of DDoS Attacks over Multiple Network Domains,” Parallel and Distributed Systems, IEEE Transactions on, vol. 18, pp. 1649-1662, 2007.

[9] C. F. Tsai and C. Y. Lin, “A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection,” Pattern Recognition, vol. 43, pp. 222-229, 2013.

[10] W. Wang, X. Zhang, S. Gombault, and S. J. Knapskog, “Attribute Normalization in Network Intrusion Detection,” The 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN), 2011, pp. 448-453.