Triangle Area Method Based Multivariate Correlation Analysis to Detect Denial of Service Attack using Packet Marking Trace Back

K.Sujithra [1], V.Vinoth Kumar [2]

[1] M.E CSE, Dept of CSE, Kalaignar Karunanidhi Institute of Technology, Coimbatore

[2] Asst Prof, Dept of CSE, Kalaignar Karunanidhi Institute of Technology, Coimbatore

Abstract

In present years a major problem on today internet user is Denial of service (DOS) attack. In order to overcome from this attack, we proposed a technique called Multivariate Correlation Analysis (MCA). This can be used to extract the geometrical correlation between legitimate and the unknown attack of the network traffic features. Triangle Area method is used to enhance and speedup the process of Multivariate Correlation Analysis. This makes our solution capable of detecting the known and unknown attack. In many instance, Dos attack can be prevented, if the source IP address is traced back to its origin. Recently IP Trace Back mechanism called Packet Marking have been Proposed for achieving trace backs of Dos attack. In this paper, we show that probabilistic packet marking – due to its efficiency it suffers under spoofing of marking field in the IP header by the attacker which can slow down the trace back by the victim. The proposed system can be evaluated using KDD Cup Data set.

Index Terms: Multivariate Correlation Analysis, Denial of Service, Triangle Area Method, Packet Marking, Probabilistic Packet Marking, KDD

1. INTRODUCTION

The most critical problem of today internet user is Denial of Service (Dos) attack. Dos are an attempt to make resources unavailable to the intended users. A Dos attack generally consists of efforts to temporarily or indefinitely suspend or delay services of a connected to the internet.

To maintain the consistency and the availability of network services, research community has put a lot of efforts to the development of intrusion detection techniques. Benefiting from the principal of detection, which monitors and flags any network activity presenting any significant deviation from their normal profiles as a suspicious and show more advanced in detecting zero day intrusion.

Therefore recent works in Dos attack mainly focus on anomaly-based techniques and various detection techniques have been proposed. However, some of these proposed techniques often suffer high false positive

ISSN: 2278 – 1323

rate since the dependencies and correlation of the features are intrinsically neglected.

A serious problem to fight these DoS attacks is that attackers use incorrect or spoofed IP addresses in the attack packets and hence disguise the real origin of the attacks. Due to the stateless nature of the Internet, it is a difficult problem to determine the source of these spoofed IP packets, which is called the IP trace back problem.

While many IP trace back techniques have been proposed, they all have shortcomings that limit their usability in practice. The IP trace back method permits the routers to encode particular data on the attack packets. The routers encode the data based on some programming probability.

The victims can build a set of paths which are navigated or traversed by the attack packets when they receive marked packets in sufficient number. As a result of this, the location of attacker can be recognized by the help of victim.

2. OVERVIEW

Correlations between any two distinct features within each single network traffic record are through this analysis. It estimates the relationship between two variables and also plays an important role in Dos attack. It is based on two ways

- Euclidean Distance Map
- Triangle Area Map

In this the existing system is covariance matrix method. To find the correlation between sequential samples we go for this approach. These approaches improve detection accuracy, and it is in danger to attacks that linearly change all monitored features. This approach can only label a group of observed samples or traffics but not individual in the group.

To discover the relationship among the feature within the observed data objects with Euclidean Distance Map (EDM). Euclidean distance and the extracted valuable correlative information, application of the multivariate correlation analysis make Dos attack detection more effective and efficient. It achieves high detection accuracy while retaining a low false positive rate. Moreover, the benefits from a principal of anomaly detection our Dos attack detection approach is independent on prior knowledge of attack and is capable of detecting both known and the unknown Dos attack.

The proposed Dos detection system architecture is given in this section,. In this we discussed about framework and sample – by- sample detection.

A. Framework

The framework consists of three steps

Step 1: Monitoring and analyzing network to reduce the malicious activities only on relevant inbound traffic. To provide a best protection for a targeted internal network.

Step 2: In this step to extract the correlation between two distinct features within each traffic record. The distinct features are come from step 1 or “feature normalization module”.

All the extracted correlation are stored in a place called “Triangle area Map”(TAM), are then used to replace the

ISSN: 2278 – 1323

original records or normalized feature record to represent the traffic record. Its differentiate between legitimate and illegitimate traffic records.

Step 3: The anomaly based detection mechanism is adopted in decision making. Decision making involves two phases

- Training phase
- Test phase

Normal profile generation is work in "Training phase" to generate a profile for individual traffic record and the generated normal profile are stored in a database. In test phase "tested profile generation" are used to build profiles for individual observed traffic records. Then at last the tested profiles are handed over to "Attack Detection" it compares tested profile with stored normal profiles. This module distinguishes the Dos attack from legitimate traffic.

B. Sample-by-Sample Detection

The group based detection technique maintained a high probability in classifying a group of sequential network traffic samples than the sample-by-sample detection mechanism. This proof was based on assumption that the samples in a tested groups are belongs to same distribution. It is difficult to predicate traffic which is belongs to same group.

To overcome the above problem we can classifying the group individually .This benefits are not found in group based mechanisms.

3. RELATED WORK

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 $\boldsymbol{g}$ the training records are

$$X^{normal} = \{x_1^{normal}\ x_2^{normal}\ \ldots x_g^{normal}\ \}$$

B. Mahalanobis Distance

Measuring the distance between a point P and distribution D. it is a multi dimensional generalization of the idea of measuring many standard variations away P is from the mean D. This is zero if P is at the mean of D, and grows as p moves away from the mean

$$D(x) = \sqrt{(x-\mu)}tS - 1(x-\mu)$$

C. *Threshold selection*

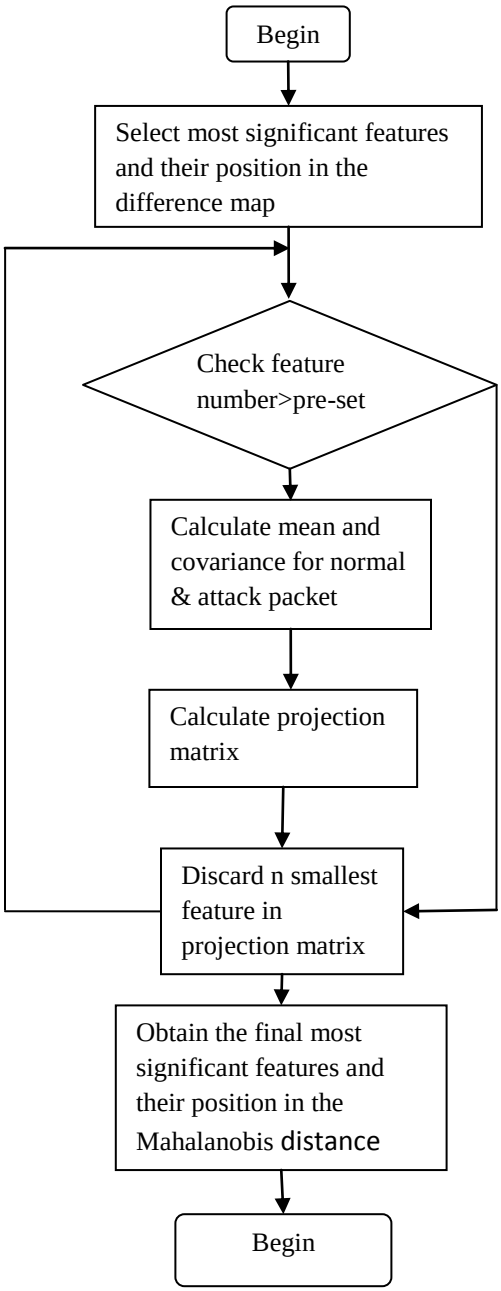
It is used to separate attack traffic from the legitimate one

$$Threshold = \mu + \sigma * \alpha$$

D. *Attack Detection*

To detect Dos attacks, the lower triangle of TAM of an observed record needs to be generated using the future triangle- area-based MCA approach

ISSN: 2278 – 1323



4. IMPLEMENTATION

A. *The Triangle Area Based Nearest Approach*

The proposed approach (TANN) is composed of three stages, which are clustering centers extraction, the new data formation by the triangle area, and K -NN training and testing based on the new data. The idea behind TANN is to extend the centroid-based and nearest neighbor classification approaches. Specifically, we assume that

All the centroid over a given dataset has their bias capabilities for distinguishing both similar and dissimilar classes. That is, the distance between an unknown data and its nearest centroid and other distances between this unknown data and other centroid can be all considered for classification. Therefore, in the feature space, an unknown data with any two centroid can result in a triangle area, thus TANN is proposed and intended to be able to improve classification performances over the centroid-based and nearest neighbor approaches. In particular, using the triangle area as the feature space for classification is the novelty of this paper.

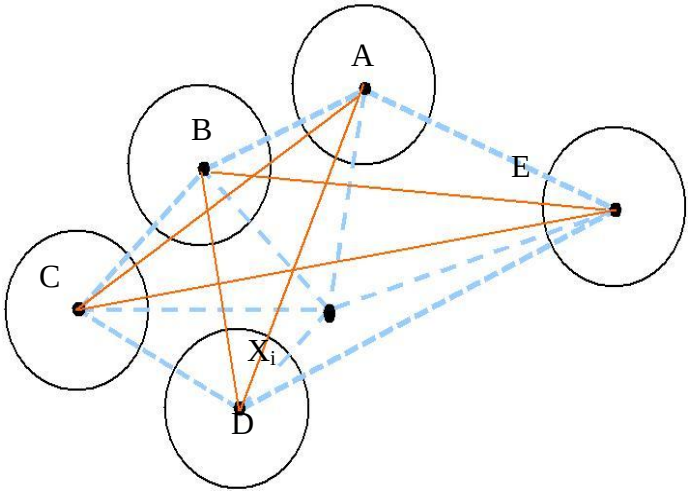


Fig. 1 An example of forming triangle areas

New Data Formation by Triangle Areas

To calculate a triangle area in the feature space, three data points need to be provided. In this stage, two cluster centers obtained by *k*-means and one data point from the dataset are used to form a triangle

ISSN: 2278 – 1323

area. Fig. 1 shows an example of the five cluster centers (*A*, *B*, *C*, *D*, and *E*) and one data point (*Xi*). Subsequently, ten triangle areas are obtained to form a new feature vector for the data point (*Xi*). That is, *XiAB*, *XiAC*, *XiAD*, *XiAE*, *XiBC*, *XiBD*, *XiBE*, *XiCD*, *XiCE*, and *XiDE*.

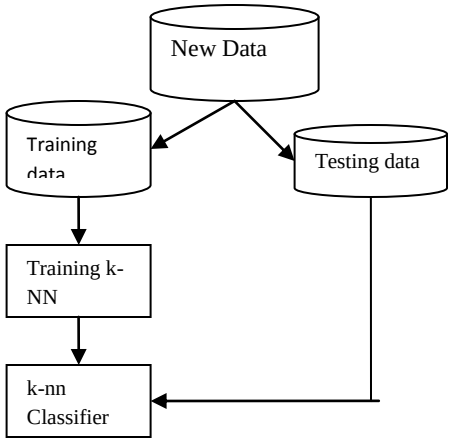


Fig 2 Training & testing KDD-N

B. *Resolved Probabilistic Packet Marking*

The algorithm of PPM is implemented very specially using two different procedures:

- Graph reconstruction procedure
- Packet marking procedure

It is defined to be the most famous packet identification techniques. In this methods, the packets are marked with the router's IP address from which they traversed or the path edges from which the packet is being transmitted.

Marking the packets with the router's address is the best approach when compared to the two alternatives provided here, where if a packet is lost of affected with any attack, the source router address can be fetched and send back to the actual router. Now the router checks the packets and re-transmits the packet to the actual destination.

With this implementation, an accuracy of 95% can be achieved to identify the actual attack path.

Second approach considered in probabilistic packet marking is edge marking and here the IP address of two nodes are required to mark the packets. This approach is much complicated when compared to marking the IP address of the router, where much state information of the packet is required in the former case.

The attacked graph edges are encoded by the packets in random. The encoded information is used to construct the new graph. The graph obtained newly should be as that of the older attacked graph.

The PPM algorithm constructs the new graph where as the graph attacked consists of the set of paths where the packets are traversed.

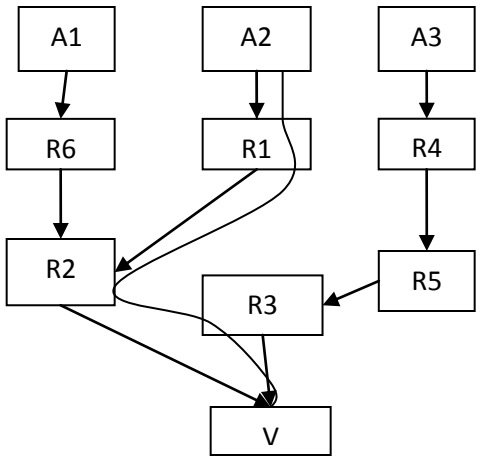


Fig. 3 The attacked graph containing the

ISSN: 2278 – 1323

path attacked

The view of network can be defined as a directed graph having G = (V, E), here E represents the edges set, and V represents the nodes set. The single host that is under attack can be considered as V. The origin of all the potential attacks is at A_i which is represented as a leaf in tree that is being embedded at V, and there are routers in the path namely R_i that are present among A_i as well as V. The routers ordered list that is between A_i and V having the packets traversed is considered as the "attack path" which is represented in the figure1 with a dotted line in the above figure1 it is (R_1, R_2, R_3). The number of routers that are present in between the R_i and V in a path is considered as the "distance" which is represented in the figure for the path R_3, R_1, R_2. Those packets that are utilized in the attacks of DoS are considered as the 'attack packets'.

Marking procedure to be followed at router 'R':

for each packet w
assume that x be a random number from [0..1]a
if $x < p_m$ then
write 0 into w.distance and R into w. start
else
f w.distance = 0 then
write R into w.end
increment w.distance

Fig. 4 Packet marking procedure

Resolved Marking Procedure to be followed at router 'R'

For each packet Pkt
$t \leftarrow t-1$
if $t_0 > t_h$
$\leftarrow t_0 - t$
else
$h \leftarrow 1$; $t \leftarrow t_0$
let r be a random number in (0,1)
if r<= 1/h
write 0 into pkt distance and R into pkt start
else
if pkt distance = 0 then
write R into pkt end
increment pkt distance

Fig.5 Resolved Packet Marking Techniques

Here, t is the TTL value being marked, tp is the maximum path length and h is the maximum remaining distance that packet would traverse.

The design of packet procedure is as follows:

MARKING FIELD	DATA

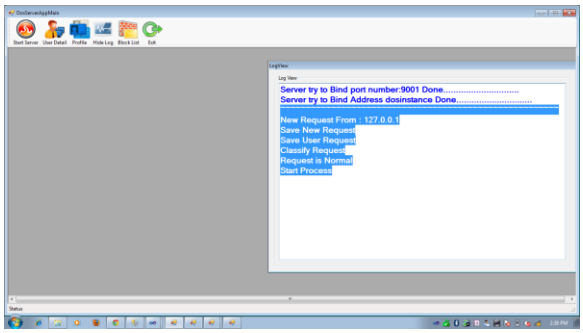
TTL	START	END	DISTANCE	DATA

Fig. 6. Design of packet marking

5. RESULT AND DISCUSSION

In the proposed algorithm the marking probability depends on TTL (Time to Live). The average number of marked packets required for a correct graph reconstruction against different values of marking probability has been decreased by the Resolved PPM when compared with PPM algorithm. The time required to reconstruct the graph also decreases compared to PPM algorithm.

ISSN: 2278 – 1323



The above figure said that the whether the request is attack or not. It gives the detail about from which ip address the request are came, and save the new request and verify its known attack or unknown attack.

6. CONCLUSION

The problem in our paper however, can be solved by utilizing statistical normalization technique to eliminate the bias from the data. This technique extracts the geometrical correlations hidden in individual pairs of two distinct features within each network traffic record, and offer extra true characterization for network traffic behaviors. Evaluation can be conducted using KDD data set to give a effective performance. The results have discovered that when working with non-normalized data, our detection system achieves maximum 95.20 percent detection accuracy although it does not work well in identifying Land, Neptune, and Teardrop attack records. The proposed system achieves equal or better performance.

To be part of the future work, we will further test our DoS attack detection system using real-world data and employ more sophisticated classification techniques to further alleviate the false-positive rate.

7. REFERENCES

[1] V. Paxson, "Bro: A System for Detecting Network Intruders in Real-Time," 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 and Security, vol. 28, pp. 18-28, 2009.

[3] D.E. Denning, "An Intrusion-Detection Model," IEEE Trans.Software Eng., vol. TSE-13, no. 2, pp. 222-232, Feb. 1987.

[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," IEEE Trans. Systems, Man, and Cybernetics Part B, vol. 38, no. 2, pp. 577-583, Apr. 2008.

[7] C. Yu, H. Kai, and K. Wei-Shinn, "Collaborative Detection of DDoS Attacks over Multiple Network Domains," IEEE Trans. Parallel and Distributed Systems, vol. 18, no. 12, pp. 1649-1662, Dec.
2007.

[8] S.T. Sarasamma, Q.A. Zhu, and J. Huff, "Hierarchical Kohonenen Net for Anomaly Detection in Network Security," IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 35, no. 2,
pp. 302-312, Apr. 2005.

ISSN: 2278 – 1323

[9] S. Yu, W. Zhou, W. Jia, S. Guo, Y. Xiang, and F. Tang,"Discriminating DDoS Attacks from Flash Crowds Using Flow Correlation Coefficient," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp. 1073-1080, June 2012.

[10] S. Jin, D.S. Yeung, and X. Wang, "Network Intrusion Detection in Covariance Feature Space," Pattern Recognition, vol. 40, pp. 2185- 2197, 2007.

[11] C.F. Tsai and C.Y. Lin, "A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection," Pattern Recognition, vol. 43, pp. 222-229, 2010.

[12] A. Jamdagni, Z. Tan, X. He, P. Nanda, and R.P. Liu, "RePIDS: AMulti Tier Real-Time Payload- Based Intrusion Detection System," Computer Networks, vol. 57, pp. 811-824, 2013.

[13] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R.P. Liu, "Denial-of- Service Attack Detection Based on Multivariate Correlation Analysis," Proc. Conf. Neural Information Processing, pp. 756-765, 2011.

[14] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R.P. Liu, "Triangle- Area-Based Multivariate Correlation Analysis for Effective Denialof- Service Attack Detection," Proc. IEEE 11th Int'l Conf. Trust,
Security and Privacy in Computing and Comm., pp. 33-40, 2012.

[15] S.J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P.K. Chan, "Cost- Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project," Proc. DARPA Information Survivability Conf. and
Exposition (DISCEX '00), vol. 2, pp. 130-144, 2000.

[16] G.V. Moustakides, "Quickest Detection of Abrupt Changes for a Class of Random Processes," IEEE Trans. Information Theory, vol. 44, no. 5, pp. 1965-1968, Sept. 1998.

[17] swathy vodithala, S.Nagaraju, V. Chandra Shekara Rao"A resolved IP Traceback through Probabilistic Packet Marking Algorithm"IJCST, Volume 2, issue 7, October 2011.

ISSN: 2278 – 1323