IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 6.078

(Volume 6, Issue 3)

Available online at: www.ijariit.com

Bi-clustering and classification-based detection for DDoS attacks

Santoshi Sahu
santoshisahu323@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Mamidi Sushma Venkata Anisha
anishamamidi03@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Rayudu Venusri Teja
venusri1635@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Sai Smruti Rout
smrutirout7281998@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

ABSTRACT

There are several Machine Learning (ML) techniques have been adopted for detecting DDoS attacks, But the attacks still became major threat. The various existing systems worked on supervised and unsupervised ML-based approaches. Various supervised ML approaches considers both labeled and unlabeled network traffic datasets to detect DDoS attacks. Whereas, unsupervised ML approaches depends on incoming network traffic data to the attacks. Both approaches analyses using large amount of network traffic data with very low accuracy and high false positive rates. In this presented paper, we propose semi-supervised Machine Learning approach for DDoS detection based on various algorithms orderly, Entropy estimation, Bi-clustering approach, and Random Trees decision making algorithm. The unsupervised part allows to remove the irrelevant traffic data for DDoS detection which allows to decrease false positive rates and increases efficiency. Whereas, the supervised part allows to reduce the false positive rates from the unsupervised part and to accurately classify the DDoS traffic data. Various experiments were conducted to evaluate the proposed approach using public NSL-KDD dataset. An accuracy of 98.66% is achieved for respectively NSL-KDD dataset, with respective to the false positive rate of 0.31%.

Keywords— *DDoS attacks, Entropy estimation, Bi-clustering, Feature selection, Randomized Trees Classification*

1. INTRODUCTION

There are two categories of DDoS attack - Direct DDoS attack and Reflection-based DDoS. In the Direct DDoS attack, the ddos attacker uses different zombie virus systems such as botnets to make floods with large number of packets to the victim system(figure1). Whereas, Reflection based DDoS attack, the attacker uses the same botnets to take control over a set of hosts with command and control technique called Reflectors. These reflectors forward a massive amount of traffic to the victim. Attackers uses Cloud Computing and Internet of Things to generate a large amount of network traffic more than 665 Gb/s [1,2]. Analyzing this large network traffic at once is inefficient and hard to compute and leads the intrusion detection systems to fail.

Artificial Intelligence, Pattern Recognition, Machine Learning (ML), Information Theory are used data mining techniques for intrusion detection [3,7,9]. Data mining and ML techniques requires selection, preprocessing, transformation, mining, and interpretation [3]. selection, mining and interpretation are important for selecting relevant data, noisy data filter and intrusions detection [3]. Supervised ML approaches uses labeled network traffic datasets to construct the model. But supervised approaches face two difficulties, the labeled network traffic datasets are costly to compute. Without continuous update the supervised machine learning algorithms unable to predict the new attacks. The presence of large amount of normal data in the incoming network traffic is noisy and degrades the performances of supervised ML classifiers. Unsupervised approaches have no labeled dataset to construct the detection model. The attacked and the normal traffics are classified based on the analysis of their distribution characteristics. The main drawback is the high false positive rates and the curse of dimensionality prevents unsupervised approaches to detect attacks accurately [9].

The semi-supervised ML approaches has advantages of both supervised and unsupervised approaches by the ability to work on unlabeled and labeled datasets and allows to increase accuracy and to decrease false positive rates. However, this has drawbacks of both approaches. In order to overcome the drawbacks of supervised and unsupervised, we propose an online sequential semi-supervised ML approach for detecting DDoS attacks. A time sliding window algorithm is used to calculate the entropy of the network header of the incoming network packets. When the entropy exceeds its threshold, the Bi-clustering algorithm splits the network traffic into three clusters. Then we consider information gain based on entropy of

features from header which are selected for preprocessing and classification using an Extremized random Trees algorithm [5]. The unsupervised approaches are entropy estimation, Bi-clustering and information gain ratio and the Extra-Trees ensemble classifiers is the supervised approach. The unsupervised approach ensures to reduce the irrelevant and noisy normal traffic, hence increasing accuracy and reducing false positive rates of the supervised part. Whereas, the supervised part reduces the false positive rates and classifies the DDoS traffic accurately. To better evaluate the performance of the proposed approach NSL-KDD dataset [6]. The results are satisfactory when compared with the state-of-the-art DDoS detection methods.
approach.

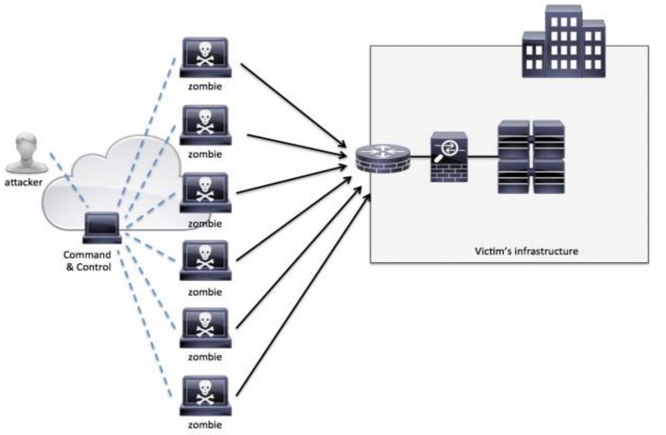

Fig. 1: distributed denial of service attacks

2. NSL-KDD DATASET

The NSL-KDD dataset has four types of attacks such as DoS, Probe, U2R and R2L. It has 40 features divided into three groups: Traffic features, Basic features and Content features. This dataset contains a total of 148,510 records in both training and testing sets. We selected this dataset for benchmarking and it contains an important percentage of DoS attack traffic. Also, it overcomes records redundancy and duplication.

3. WORKING

This section explains the working of the proposed approach and the methods for detecting the DDoS attack. It consists five major steps: pre-processing, network traffic Entropy estimation, Bi-clustering, information gain ratio computation and extra trees classification. Our methodology to detect DDoS attack is illustrated in Figure 2.

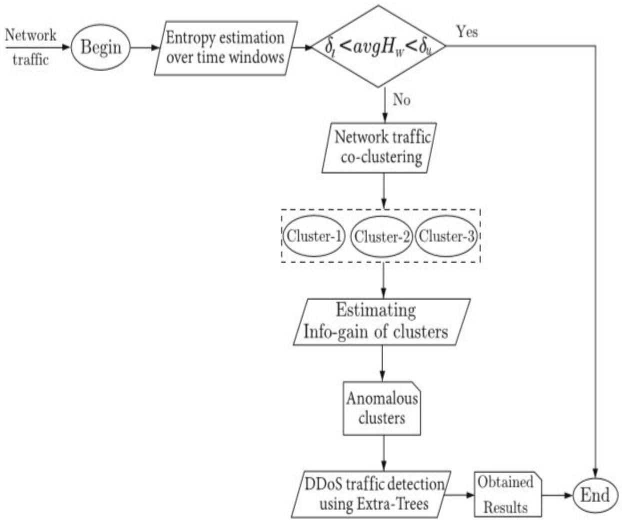
Network traffic
Begin
Entropy estimation over time windows
$\delta_l < avgH_w < \delta_u$
Yes
No
Network traffic co-clustering
Cluster-1
Cluster-2
Cluster-3
Estimating Info-gain of clusters
Anomalous clusters
DDoS traffic detection using Extra-Trees
Obtained Results
End

Fig. 2: The proposed approach

3.1 Entropy Estimation

The entropy is a measure of the randomness of a distribution [12]. The analysis of the network traffic flow entropy over time sliding windows reduces high dimensionality of the network traffic distribution to a single measure [13]. The flow size distribution (FSD) features, the source/destination packets count and the bytes count, are used to estimate the entropy. As, zombies send a large number of packets to the server which generates a large amount of network traffic. Hence, estimating entropy of the FSD features allows to detect changes in the network flow. Shannon formula is defined as:

$$H(X)= -\sum p(x_i)\cdot\log(p(x_i))$$

Where X represents xi(i =1,⋯,n)are the frequencies of items of X received and n is the total number of items of X. In order to determine the clusters, we have to estimate the information gain ratio based on the average entropy of FSD features.

3.2 BI-Clustering Algorithm

Bi-clustering algorithm performs a simultaneous clustering of rows and columns of a data matrix based on a specific feature [4,15]. It forms clusters of rows and columns that represent sub-matrices of the original data matrix with desired properties. Clustering simultaneously of a data matrix yields three major benefits:

(a) Dimensionality reduction, as each cluster is divided based on features.
(b) Representation of data in a more compressed form with original data preservation.
(c) Significant reduction of the clustering complexity. The Bi-clustering computational complexity is O(nkl+mkl) Where m is the number of rows and n is the number of columns, l is the number of column clusters and k is the number of clusters.

3.3 Data Preprocessing

A set of similar features are selected and the incoming network traffic data is to be normalized.

3.3.1 Feature Selection: Each cluster created by the Bi-clustering algorithm is the subset of the original features. we want to classify the two anomalous clusters such as DDOS features and normal features produced by the Bi-clustering algorithm, we combine their corresponding feature subsets into relevant subsets. This allows us to preserve information of both anomalous clusters. This is beneficial because the attackers continually update their tools and changes behaviors.

3.3.2 Data Normalization: Attribute values in a network traffic dataset are not uniformly distributed. We have to distribute each attribute values uniformly before starting process. We use Min-Max normalization method.

4. EXTRA-TREES CLASSFICATION

Ensemble-based Random Forest and Extra-Trees are useful to overcome the problems of the univariate trees. Hence, ensemble-based trees are used for classification and regression problems [16]. Extremely randomized trees is a tree-based ensemble method for supervised regression and classification problems. It can randomize both cut-point choice and attribute during splitting a tree node. The Extra-Trees algorithm follows classical top down approach to build an ensemble of unpruned decision trees or regression trees. The differences from other tree ensemble methods are splitting nodes by choosing cut-points randomly to grow the trees. The explicit randomization of the cut-point and attribute reduces the variance than the randomization schemes used by other tree-based methods. Overall, Extra-Trees achieves good variance for classification problems. From the computational point of view, the complexity in the order of O(MNKlogN). Where, N is the number of samples, M represents the number of trees and K is the number of variables drawn from each node. The entire proposed algorithm is discussed here. This algorithm has accuracy and computational efficiency.

5. RESULT DISCUSSION

5.1 Entropy Estimation

The estimated entropy of FSD features of NSL-KDD dataset is shown in Figure3. The upper and lower thresholds of time series are represented by dashed lines. These thresholds are estimated by using the max and min entropy of dataset. If time reaches lower or upper threshold the incoming data will be clustered. Estimating entropy is important to reduce preprocessing data and to classify by network anomalies.

Algorithm of semi-supervised:

```
Input: traffic data
Output: classification results
Ddos(data_w)
S={spkts, dpkts }
//source, destination packets
while traffic_data do
     foreach w
         avgH=avgentr(data_w,S)
         if(avgH≠upp_lim,low_lim) then
                 clusters
              for c in clusters
                   infogain(data_w,c)=avgentr(data_w,S)-
                          avgentr(c,S)*(len(c)/len(data_w)
              end
              //div c with infogain(data_w,c) forms
               anomalous clusters
               //preprocess the anomalous clusters
               Extratrees(anomalous cluters)
        end
    end
end
```

5.2 Determination of accuracy of Bi-clustering and extra trees algorithm

To determine optimal features, we use 10-fold cross validation. For each parameter we apply 10-fold cross validation to obtain best performances. To obtain optimal parameters of proposed Bi-clustering algorithm, we use10-fold cross validation on NSL-KDD dataset. The criterion used to obtain the best parameters of the Bi-clustering algorithm is Mean Purity metric. A 100% high purity is noticed for dataset for elements of Bi-clustering.

5.3 BI-Clustering of the NSL-KDD Dataset

The performance of clustering results of NSL-KDD dataset is shown in figure 4. The purity is above 90% which explains about formed clusters. This figure shows high precision rates at 1, 5- and 7-time windows. We can see high FPR due to high normal data which is removed using classification phase.

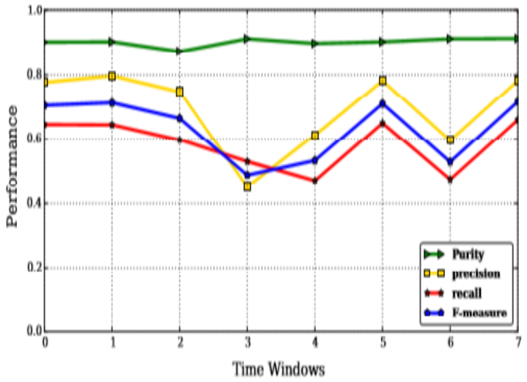

Fig. 4: Average Bi-clustering performance with respect to time window

5.4 Bi-Clustering and information Gain Ratio for data reduction

Bi-clustering algorithm is applied in the proposed method to divide the incoming traffic packets into three clusters. One cluster contains only attack traffic, second one contains normal data and third one contains both attack and normal traffic. Bi-clustering achieves high clustering performances high Purity and precision for dataset in splitting. The high Purity is achieved during time-based windows says there are different classes in the clusters divided based on normal or attack. Information gain ratio algorithm distinguishes clusters. DDoS attack clusters and normal clusters are further preprocessed and classified. The cluster with normal traffic is again examined to reduce important amount of network traffic data. The data reduction of normal data in NSL-KDD dataset is shown in Figure 4.

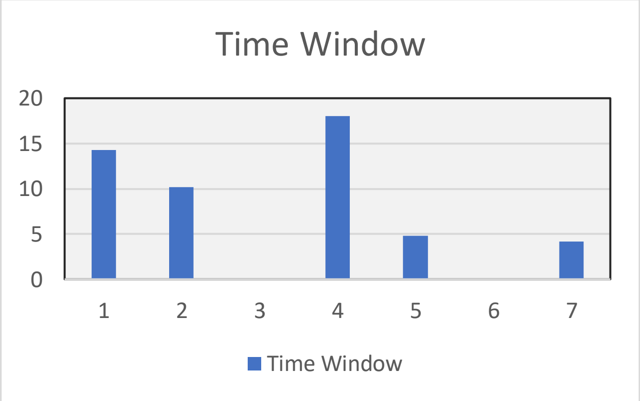

Fig. 4: The reduction of normal data at each time using Bi-clustering and Information Gain

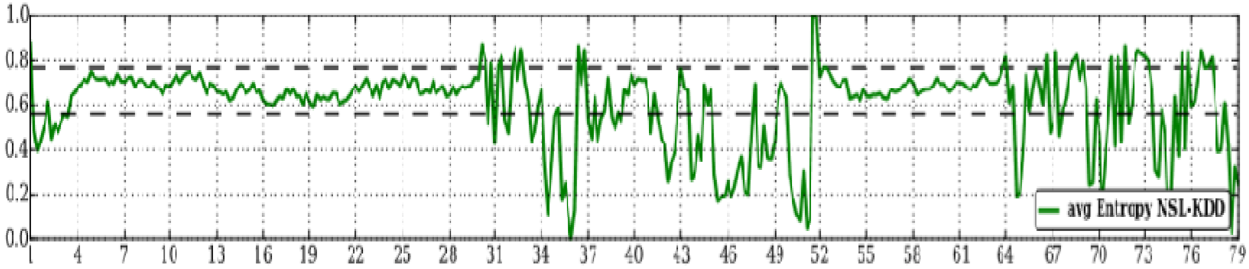

Fig. 3: Entropy estimation of FSD features

5.5 Extra-Trees Performance

Extra-trees are used to remove normal data obtained from clustering. It achieved 85% accuracy and 0.33 FPR with NSL-KDD dataset. Extra achieves low false positive rates which is the important advantage of using this classification algorithm.

6. CONCLUSION

We proposed a system to detect DDOS attacks using semi-supervised learning based on Entropy estimation, Bi-clustering, Extra Trees classification algorithms. The entropy estimator estimates the network traffic data entropy with time sliding window. When the entropy exceeds its threshold, the received network traffic is split into three clusters using the Bi-clustering algorithm. The network traffic data clusters are considered as anomalous and selected for preprocessing and classification using an Extra-Trees classification algorithm.

We conducted various experiments to get the performance of proposed approach using NSL-KDD dataset. This proposed approach shows good performances with the NSL-KDD dataset, it is important to evaluate its performances in real world applications. For future work, we are planning to perform real world deployment of the proposed approach and evaluate it against several DDoS attacks.

7. REFERENCES

[1] Wikipedia (2016) 2016 DYN cyberattack. https://en.wikipedia.org/ wiki/2016 DYN cyberattack.

[2] DDOS attack using Marai botnet on dyn servers using DNS information on the internet. https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet.

[3] Han J, Pei J, Kamber M (2006) What is data mining. Data mining: concepts and techniques. Morgan Kaufinann

[4] Berkhin P (2006) Clustering techniques in Grouping multidimensional data.

[5] Geurts P, Ernst D, Wehenkel L (2006) Extra-trees clasification for decision making. Machn Learn 63(1)

[6] Tavallaee M, Bagheri E, Lu W, Ghorbani A-A (2009) A detailed analysis of the KDD cup 99 data set in second IEEE conference on computational intelligence for security applications 2009

[7] Ayat S, Boroujerdi AS (2013) A Neuro-fuzzy classification for detection of ddos attack in 3rd international conference held on computer science and network technology (ICCSNT). IEEE.

[8] Ahmed M, Mahmood AN (2015)Network traffic pattern analysis using clustering based on collective anomalies detection.

[9] Saied A, T Radzik (2016) Detection of labeled and unlabeled DDOS attacks using artificial neural networks.

[10] Boro D, Bhattacharyya a dynamic protocol of specific defense for high traffic network distributed denial of service flooding attacks, Microsystems Technol.

[11] Nicolau M, McDermott J et al (2016) A hybrid autoencoder and density estimation model for anomaly detection. In: International M. Idhammad et al. conference on parallel problem solving from nature.

[12] Lakhina A, Crovella M, Diot C (2005) Data Mining of anomalies using traffic feature classification in ACM SIGCOMM computer communication.

[13] Liu T, Wang Z, Wang H, Lu K (2014) An entropy-based method for attack detection in large scale network. Int J Comput Commun Control

[14] Papalexakis EE, Beutel A, Steenkiste P (2014) Network anomaly detection using Bi-clustering. In: Encyclopedia of social network analysis and mining.

[15] Ahmed M, Mahmood (2014) Information theoretic Bi-clustering based anomaly detection for network traffic pattern analysis in International conference on security.

[16] Ahmad A (2014) Decision tree ensembles based on kernel features.

[17] McKinney W (2014) Pandas, python data analysis library.