VOTING-BASED ENSEMBLE MODEL FOR NETWORK ANOMALY DETECTION

This article is summary of paper "VOTING-BASED ENSEMBLE MODEL FOR NETWORK ANOMALY DETECTION"

Cyber-attacks against Internet users and the network are fast increasing in tandem with the increasing demands on the Internet. As a result, network security and network anomaly detection (NAD) become a pressing concern that must be addressed.

What is NAD?

NAD aims to capture potential abnormal behaviors by observing traffic data over a period of time.

The dataset used for this problem is not published but contains the below mentioned attacks:

Attack Analysis:

DDOS-Smurf: In this type of attack all the destination IP are ended with 255.

Probing-IP Sweep: Is a type of attack wherein it consists of an interesting pattern in the "destination IP" feature; most of them only differ from one another in the last octet but be the same in the first three octets.

Probing-Port Sweep: This attack is similar to Probing-IP Sweep where the source IP is same but destination IP differs only in the last octet.

Probing-Nmap Sweep: Here all the destination IP starts with "172.24" in the first two octets.

Feature Engineering:

Misconception 1: IP address column is a useful feature.

If the collected traffic in both training and testing belongs to same network then IP address turns out to be a useful feature i.e., when PCA is applied on testing dataset to visualize the distribution of the data it is found that they are not similar to the training dataset.

So, assuming IP's to be important feature was wrong and it was eliminated.

Misconception 2: Ports column can be useful feature.

Although some applications or protocols correspond to a common port, ports can be a random number in many cases. So, ports column was also eliminated.

Crafting new columns from the existing columns in the dataset:

inner src, inner dst: It is observed that destination IPs of Probing-Nmap start with "172.24". Thus, it is assumed that the features of Inner IPs may help to detect Probing-Nmap. Therefore, the IP is considered

as inner IP if it belongs to the IPv4 address ranges reserved for a private network.

dst Ip end with 255: The destination IPs of DDOS-smurf data are all ended with 255, it is added as binary feature in the input set.

is sweep: For Probing-Port sweep and Probing-IP sweep, the destination IP often differs only in the last octet from the consecutive data with the same label in the neighboring. Therefore, the difference of "src" and "dst" between each row is calculated. If "src" is the same as the previous row and "dst" is only different in the last octet. Then "is sweep" is assigned to be 1.

Modeling: Modeling is done using ensemble method with XGBoost and Deep Neural Network as classification algorithms.

What is Ensemble method?

It is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

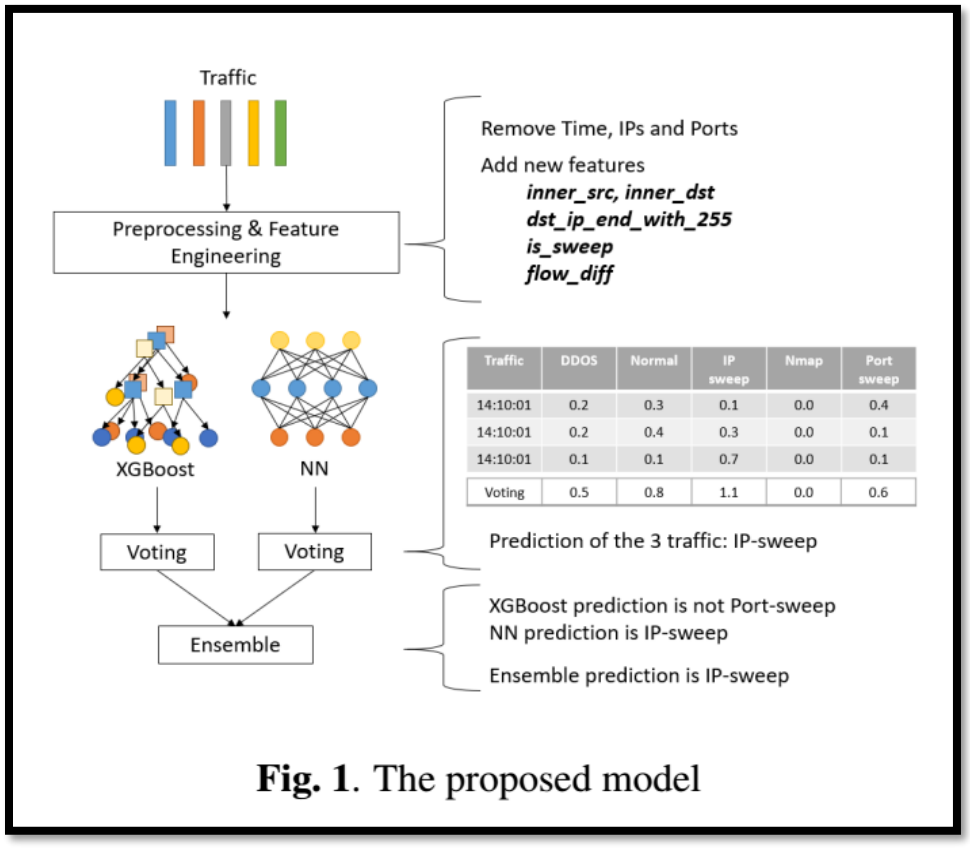

Fig. 1. The proposed model

Authors of the paper uses XGBoost and Deep neural network for predicting the anomaly.

What is XGBoost Classifier?

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

What is Deep neural network?

Deep neural network represents the type of machine learning when the system uses many layers of nodes to derive high-level functions from input information. It means transforming the data into a more creative and abstract component.

Result Visualization:

In XGBoost evaluation, the F1 score of Probing-IP sweep was 80%, while in neural networks evaluation, the F1 score of Probing-IP sweep was 85%. Therefore, we use neural networks to predict Probing-IP sweep and use XGBoost to predict other kinds of attacks to generate the final result.

Voting method in post processing:

Firstly, a fixed time splitting method by defining the data are consecutive is whether they are in the same hour or minutes. That is, if two traffic happened in the same hour and with the same source IP, we view them as consecutive data. They would get involved in the same voting session.

For the best efforts, a combined voting method that uses “same minute” for DDOS-smurf and “same hour” for all the others is composed.

Furthermore, a dynamic time splitting method by defining two traffic in the same group if their source IPs are the same and the time difference between two traffic is less than one hour. The result shows the improvement in predicting testing data.

Conclusion:

Performance across model with proposed solution and feature engineering:

The base model used in this experiment is XGBoost with voting post-processing. Below table shows the F1 score for each label across different combination of feature engineering. The result shows that performance further boosts in predicting DDOS-smurf, IP sweep, and Port sweep by adding additional features “is sweep” and “dst ip end with 255”. For adding “is sweep”, the experiment result shows that the performance improves not only in predicting IP sweep and Port sweep but also in predicting Nmap sweep. For adding “dst ip end with 255”, the model could filter out DDOS-smurf attacks by this feature since it could strongly represent the DDOS-smurf behavior.

	voting	voting + is_sweep	voting + is_sweep + dst_end_with_ip_255
DDOS	0.0546	0.1677	0.9918
Normal	0.9960	0.9973	0.9973
IP sweep	0.8081	0.8755	0.8670
Nmap sweep	0.3626	0.6352	0.6352
Port sweep	0.9173	0.9201	0.9191
Criteria	0.6408	0.6873	0.8083

Medium article link: https://medium.com/@poojashree.ns/voting-based-ensemble-model-for-network-anomaly-detection-eb0841b931f2