2-stage Feature Selection for Intrusion Detection Systems by Using a Multi-Objective Genetic Algorithm

CHANG GENG[†] MASAHARU MUNETOMO[†]

Abstract: In this paper, we employ feature selection strategy iteratively to achieve a fine-grained level botnet detection. And a multi-objective genetic algorithm is utilized in searching the optimal features to take the trade-off relation between botnet detection accuracy and computation cost in IDSs into consideration. The machine learning algorithm C4.5 decision trees algorithm is used to evaluate the feature subset by classifying the botnet attack behavior from benign traffic. The experiments on KDDCup99 show that the proposed feature selection strategy can effectively select the optimal feature set and achieve a good detection accuracy.

Keywords: feature selection, multi-objective genetic algorithm, IDSs

1. Introduction

Nowadays it is very important to maintain a high-level security to ensure safe and trusted communication of information between various organizations because of the growth of botnet activity in cyberspace. Intrusion Detection Systems (IDSs) could separate the botnet behavior from benign traffic. Because the raw data from audit log usually includes amounts of features in intrusion detection system, which sometime leads detection procedure to cost too much time. Feature selection, as a process selecting a subset of features from the candidate features and deleting the irrelevant attribute, is expected to help to reduce the computational cost and improve the accuracy and generalizability of a predictive model.

In this paper, to improve the trade-off relation between botnet detection accuracy and computation cost used in IDSs, we propose a 2-stage feature selection to gain the optimal feature set by using a multi-objective genetic algorithm. The best optimal features used models is expected to achieve a good detection accuracy with lower computational cost. We apply C4.5 decision trees to evaluate the selected feature set to determine which feature leads a better classification performance. To verify our proposal, we run the experiments on KDDCup99. The result shows that the proposed 2-stage feature selection can effectively select the optimal feature set and achieve a good detection accuracy.

2. Main Point

2.1 Related work

As a pre-processing step, only a few of related work with algorithms based on connections with time intervals to detect botnets have put the focus on feature selection.

Beigi et al. [1] presented a proposal that they categorized all features into four groups: Byte-based, Packet-based, Time, and Behavior-based. Group exclusion and feature inclusion by using greedy algorithm is to evaluate all the features and choose the final feature set. ISOT dataset, ISCX 2012 IDS dataset and botnet traffic generated by the malware capture facility were merged in one unified dataset to conduct an experiment. Even though their final feature achieved a high detection rate of 99% on a dataset including limited number of botnets, 75% shown on a much more diverse botnet traces. Alejandre et al. [2] utilized a Genetic Algorithm (GA) as an optimizer algorithm to select the feature set and a classifier C4.5 to evaluate the potential set of features generated by the GA, delivering just the detection rate obtained. Their experiment respectively achieved 99.44% and 96.52% True Positive Rate(TPR) for ISOT and ISCX dataset. Even though they achieved a promising result, the result was still questionable because of the limited diversity of the botnet. And the computational cost has not been considered in their study.

2.2 2-stage Feature Selection Proposal

To deal with trade-off relation between detection accuracy and computation cost in IDSs, we proposed a 2-stage feature selection proposal.

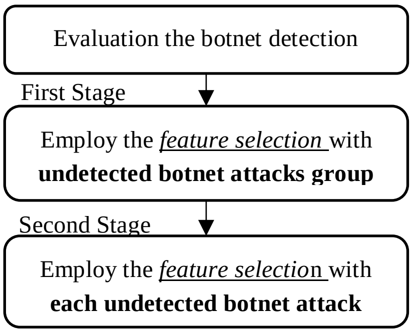
Evaluation the botnet detection
First Stage
Employ the feature selection with undetected botnet attacks group
Second Stage
Employ the feature selection with each undetected botnet attack

Figure 1. 2-stage Feature Selection

In our proposal, as Figure 1 shows, firstly, we evaluate detection accuracy with full features. Then, we pick up the undetected botnet attacks as a group. After that, we employ the two-stage feature selection strategy. We employ the feature selection to generate an optimal feature subset to improve the TPR of the undetected botnet group at first stage. After that, we evaluate the botnet detection accuracy with selected features. At the second stage, we pick up the undetected botnet attacks in the first stage and employ the feature selection strategy again to generate an optimal feature subset for each specified botnet and evaluate them. Finally, the best optimal feature set for each botnet is gained.

From figure 2, we show our feature selection, which regards NSGA-II (Non-dominated Sorted Genetic Algorithm) as a search

[†] Hokkaido University.

method to choose the optimal feature. Compared with [2] obtaining the feature set just with highest true positive detection rate(TPR), our proposal takes true positive detection rate(TPR) and the cardinality of the subset the into consideration to obtain the optimal feature set. And the obtain TPR of botnet attack is evaluated by C4.5, an algorithm used to generate a decision tree developed by Ross Quinlan.

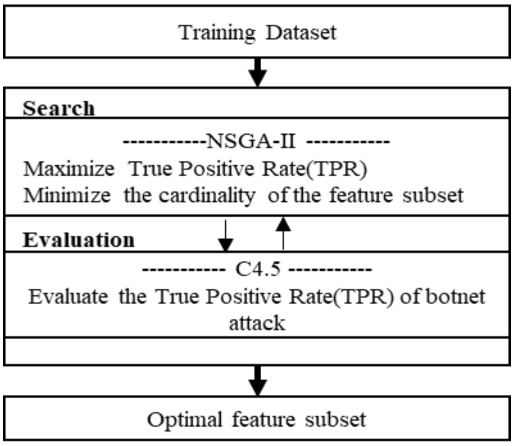
Training Dataset
Search
-----------NSGA-II -----------
Maximize True Positive Rate(TPR)
Minimize the cardinality of the feature subset
Evaluation
----------- C4.5 -----------
Evaluate the True Positive Rate(TPR) of botnet attack
Optimal feature subset

Figure 2. Feature selection

2.3 Experiments

To apply our proposal, all experiments have been executed on a 2 processors machine Intel(R) Core(TM) i5-5200U CPU@2.20GHz 2.19GHz, with 8 GB RAM, running window 10.

- Analyze the botnet detection

To analyze the evaluation of botnet detection, firstly we run classification on the full KDDCup99 dataset (743MB) with C4.5(J48 in Weka), without feature selection. Although the botnet was well detected (TPR:99.99%) in total, there were still 7 types of botnet attacks (belong to R2L and U2R attacks) could not be detected (lower than 50% TPR). They are embedded in the data portions of packets, and normally involve only a single connection. It took 863.94 seconds to build a model.

- 2-stage Feature Selection Proposal

We run feature selection with C4.5 based wrapper method and NSGA-II based multi-objective evolutionary search strategy on Weka to achieve a fine-grained botnet detection.

First Stage

To improve the left 7 types of botnet detection rate and lower the computational, firstly, we run feature selection with 10% KDD Cup99 dataset (77MB) to get the optimal set.

The parameters set in experiment is as below:

Number of generations: 10

Population size: 100

Fitness function 1: the average TPR of 7 types of botnet attack

Fitness function 2: the number of selected features

After experiment, we could see the trade-off relation between botnet detection accuracy and computational cost from figure 3. And protocol_type, service, src_bytes, lnum_root, same_srv_rate, dst_host_same_srv_rate, total 6 features were chosen.

Then, we used the full KDDCup99 dataset with C4.5(J48 in Weka) to analyze the evaluation of botnet detection. Usually, the more features we used in the classification task, the more time would be taken in building a mode. So, the number of features in figure 3 indicated the computational cost in IDSs. The figure 3 reflected the trade-off relation between botnet detection accuracy and the computational cost.

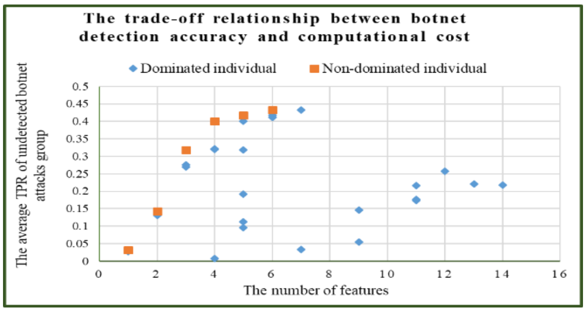

Figure 3. the trade-off relation between botnet detection accuracy and the number of features

Second Stage

According to the first stage experiment, the average TPR of left 7 types of botnet detection was improved. In this part, to gain a fine-grained level botnet detection, we focused on the botnet detection of each specified botnet. The parameters set in this experiment were almost same as the first stage, just only changing the Fitness function 1 into the TPR of one specified botnet attack, loadmoudle, rootkit, ftp_writite, multihop, or imap. The optimal features for each botnet attack were selected.

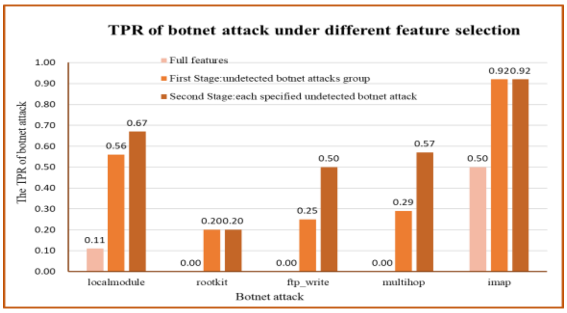

Figure 4. The TPR of botnet attack under different feature selection

Finally, we evaluated detection ability for each botnet attack by using their selected optimal features. We also compare the TPR of each botnet attack under different feature selection method. The results were shown in figure 4.

3. Concluding Remarks

In the future, we intend to adjust the parameters in multi-objective genetic algorithm to achieve a better optimization between detection accuracy and computational cost in intrusion detection systems.

Reference

[1] E. Beigi, H. Jazi, N. Stakhanova and A. Ghorbani, Towards Effective Feature Selection in Machine Learning-Based Botnet Detection Approaches, in IEEE Conference on Communications and Network Security (CNS), pp. 247-255. (2014)

[2] Alejandre FV, Cortés NC, Anaya EA. Feature selection to detect botnets using machine learning algorithms. In Electronics, Communications and Computers (CONIELECOMP), 2017 International Conference pp. 1-7. (2017)