Title: Botnet Analysis Using Ensemble Classifier

Author: Anchit Bijalwan Nanak Chand Emmanuel Shubhakar
Pilli C. Rama Krishna

PII: S2213-0209(16)30142-2
DOI: http://dx.doi.org/doi:10.1016/j.pisc.2016.05.008
Reference: PISC 288

To appear in:

Received date: 20-2-2016
Revised date: 15-5-2016
Accepted date: 31-5-2016

Botnet Analysis Using Ensemble Classifier

Anchit Bijalwan[1], Nanak Chand[2], Emmanuel Shubhakar Pilli[3], C. Rama Krishna[2]
Department of Computer Science and Engineering
[1]*Uttarakhand Technical University, Dehradun, India, anchit.bijalwan@gmail.com*
[2]*National Institute of Technical Teacher's Training and Research, Chandigarh, India, nanak.cse@nittrchd.ac.in, rkc_97@yahoo.com*
[3] *Malviya National Institute of Technology, Jaipur, India, espilli.cse@mnit.ac.in*

Abstract

This paper analyzes the botnet traffic using Ensemble of classifier algorithm to find out bot evidence. We used ISCX dataset for training and testing purpose. We extracted the features of both training and testing datasets. After extracting the features of this dataset, we bifurcated these features into two classes, normal traffic & botnet traffic and provide labeling. Thereafter using modern data mining tool, we have applied ensemble of classifier algorithm. Our experimental results show that the performance for finding bot evidence using ensemble of classifiers is better than single classifier. Ensemble based classifiers perform better than single classifier by either combining powers of multiple algorithms or introducing diversification to the same classifier by varying input in bot analysis. Our results are showing that by using voting method of ensemble based classifier accuracy is increased upto 96.41% from 93.37%.

Keywords: Botnet, Ensemble of classifier, machine learning.

I INTRODUCTION

Botnet has become a common phenomena on Internet. It is a collection of infected machine. In other word it is kind of army of infected bots targeted at spreading malicious activity and expansion of bot army. The Botmaster controls and communicates through C&C channels. IRC is most commonly and widely utilized channel.

It is very difficult to observe the user, who is becoming the part of Botnet or about to be infected by stealthy malware. Bot can forward secret information to the adversaries. Bot herder can command the bot to spread Denial of Services attack (DoS), phishing activity, key logging, forwarding the spam and click fraud etc. In order to combat from these kind of attacks researchers needs to focus on the features of the Botnet .

It is necessary to analyze the Botnet traffic for this purpose, which is used to analyze and record

Botnet has become a common phenomena on Internet. It is a collection of infected machine. In other word it is kind of army of infected bots targeted at spreading malicious activity and expansion of bot army. The Botmaster controls and communicates through C&C channels. IRC is most commonly and widely utilized channel.

In this paper we have taken ISCX training dataset and testing dataset to analyze the Botnet traffic using Ensemble of classifier algorithm to find out evidence about Bot. We extract the features of both training and testing dataset. After extracting the features of this dataset, we bifurcate this features into two classes i.e. on normal traffic & Botnet traffic and provide labeling. Thereafter using data mining tool, we applied ensemble of classifier algorithm.

Our experimental analysis result shows the performance for findings bot evidence using ensemble of classifier is better than single classifier.

Ensemble based classifiers perform better than single classifier by combining multiple algorithms in bot analysis.

This paper categorizes as section II shows the review of literature of previous work on machine learning technique for the Botnet analysis. Section III defines the ensemble classifier framework section IV exhibits the experiments and result and finally concludes the paper in section V.

II REVIEW OF LITERATURE

Livadas et al. [1] identified the Botnet traffic using machine learning technique. For this purpose he segregated the whole traffic into IRC and non IRC traffic. After segregation he differentiated the IRC traffic & real traffic and compare this analysis with J48, naïve Bayes & Bayesian network classifiers.

Beigi et al. [2] focuses on statistical network flow features rather than packet content is unable to differentiate between botnet IRC traffic and benign traffic. Author shows the loophole on previous methods such as principle component analysis (PCA), correlation feature selection (CFS), minimum redundancy maximum relevance (mRMR) and improper evaluation of features set on testbed datasets. He built a dataset which incorporate different variety of botnet of different protocol in realistic environment.

Saad et al. [3] Proposed a new approach (detecting P2P Bot before launch the attack) to characterize and detect through network traffic behavior. Using machine learning technique he extracted, analyzed the set of C&C traffic behavior &its characteristics. He differentiated among five machine learning technique i.e. Super vector machine (SVM), artificial neural network (ANN), Nearest neighbors classifier (NNC), Gaussian based classifier (GBC) and Naïve bayes classifier (NBC).

Rokach et al. [4] divided ensemble model into dependent and independent method. In dependent method the most well versed model instance is Boosting which is known as resampling and combining. It is used to improve the performance of week classification on distributed training data. Through iterative process AdaBoost is well known ensemble algorithm to improve simple boosting algorithm. In independent well known method is Bagging and Wagging.

III ENSEMBLER CLASSIFIER FRAMEWORK

Initially authors used bagging method for Ensembler learning. Meta algorithm of model averaging was built for classification initially. It used the multiple training set. by utilizing bootstrap, it used many version of training set. It is ensemble meta algorithm of machine learning which is made for improving the accuracy of machine learning algorithm for both regression and classification. Each version of data set utilizes for training of different model. Through averaging and voting output of the model combined and then create a single output in case of regression and classification. KNN is a simple classifier for basic recognition problem. It is slow for real time prediction but good for basic problem, use training data itself for classification. Decision tree learning is a predictive modeling maps observation for a data. This approach uses in various data mining and machine learning. It describes the classification tree which shows the trees structure, leaves, class labels and branches. ADABoost is powerful classifiers work effectively both in basic and complex recognition problem. It combine all weak and inaccurate classifier and make one ensemble classifier. AdaBoost classifier is train by classification data structure. Figure 1 is describing about the flow diagram of ensemble based classifier.

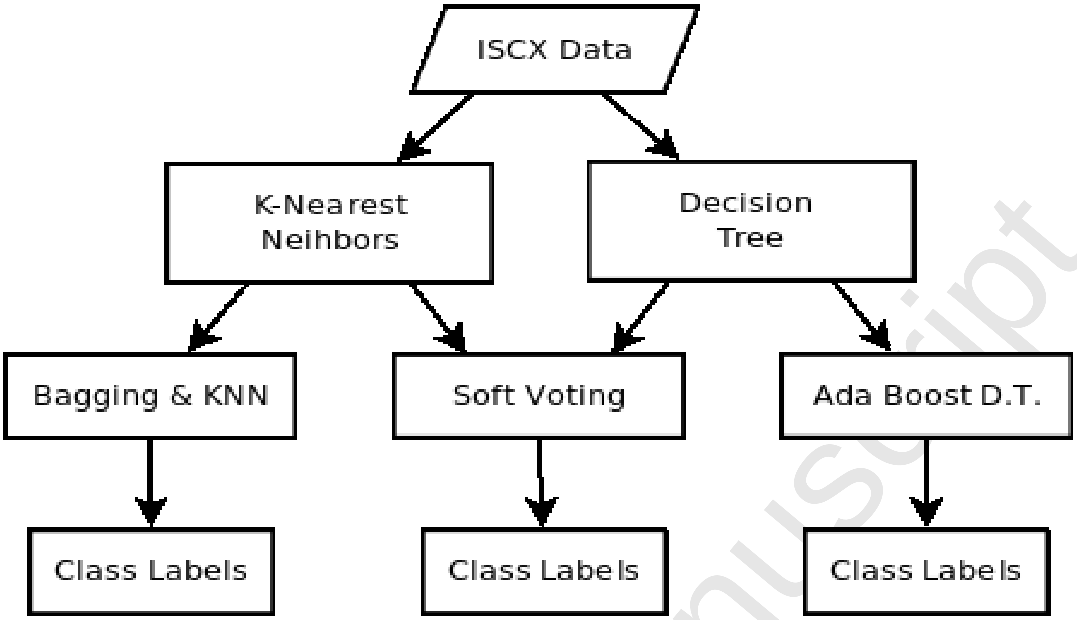

Figure No:1 Ensemble Classifier Methods

IV Experiments And Results

In Our experiments, we have used ISCX Botnet dataset [5] from which we have extracted 42 attributes, provided labels to every instances and splitted it into training and testing datasets. After that using Scikit-Learn (a python library) we have applied machine learning and ensemble algorithms to this dataset. We have used bagging, AdaBoost, Soft-Voting method of ensemble based classifier. We have compared the performance of each classifier based on their accuracy to predict classes of unknown instances. Table No:1 is describing the comparison of different classifiers. As observed from Table No:1 performance of bagging KNN i.e. 95.69% is better than KNN i.e. 93.87% because it reduces variance in input data and acoids overfitting based classifier is better than single classifier and highest accuracy i.e. 96.41%. Ada-Boost Decision tree also increases accuracy from 93.37% to 94.78% improving learning process of decision tree. And highest accuracy is achieved by using soft-voting rule because it merges the power of two algorithms and give more weightage to the decision of better performing algorithm. The output of single classifier does not give perfect Bot findings. this paper shows the performance of Bot evidence using Ensemble of classifier is better than the single classifier.

Classifier	Accuracv
KNN	93.87%
Decision Tree	93.37%
Bagging with KNN	95.69%
Ada-Boost with Decision Tree	94.78%
Soft Voting of KNN & Decision Tree	96.41%

Table No 1 Performance Comparison Table of Classifiers

V CONCLUSION

This ensemble based classifier always performs better performance because it is made up by combining multiple algorithm in Bot analysis. this paper extracts the features of both training and testing dataset. It segregated this features into classes i.e. on normal traffic & botnet traffic and provide labeling. Thereafter using data mining tool, we applied ensemble of classifier algorithm. We analyze the Botnet traffic Using ISCX training dataset and testing dataset to

References

[1] C. Livadas, R. Walsh, D. Lapsley, and W. T. Strayer, "Usilng machine learning technliques to identify botnet traffic," in *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, 2006, pp. 967-974.

[2] E. B. Beigi, H. H. Jazi, N. Stakhanova, and A. A. Ghorbani, "Towards effective feature selection in machine learning-based botnet detection approaches," in *Communications and Network Security (CNS), 2014 IEEE Conference on*, 2014, pp. 247-255.

[3] S. Saad, I. Traore, A. Ghorbani, B. Sayed, D. Zhao, W. Lu, J. Felix, and P. Hakimian, "Detecting P2P botnets through network behavior analysis and machine learning," in *Privacy, Security and Trust (PST), 2011 Ninth Annual International Conference on*, 2011, pp. 174-180.

[4] L. Rokach, "Ensemble-based classifiers,"*Artificial Intelligence Review*, vol. 33, pp. 1-39, 2010.

[5] Beigi, Elaheh Biglar, et al. "Towards effective feature selection in machine learning-based botnet detection approaches."*Communications and Network Security (CNS), 2014*

IEEE Conference on. IEEE,
2014. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6997492&tag=1