International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 3 187 - 191

Machine Learning based Traffic Classification using Statistical Analysis

Abirami Sivaprasad
Assistant Professor
IT-SAKEC, Mumbai,India.
abi.lecturer@gmail.com

Neha Ghawalkar
IT-SAKEC
Mumbai,India.
neha.ghawalkar2018@gmail.com

Srushti Hodge
IT-SAKEC
Mumbai,India.
srushtihodge@gmail.com

Maitri Sanghavi
IT-SAKEC
Mumbai,India.
maitris04@gmail.com

Vidhya Shinde
IT-SAKEC
Mumbai,India.
vidhyass.shinde@gmail.com

***Abstract*—** In this paper, Automated system is built which contains processing of captured packets from the network. Machine learning algorithms are used to build a traffic classifier which will classify the packets as malicious or non-malicious. Previously, many traditional ways were used to classify the network packets using tools, but this approach contains machine learning approach, which is an open field to explore and has provided outstanding results till now. The main aim is to perform traffic monitoring, analyze it and govern the intruders. The CTU-13 is a dataset of botnet traffic which is used to develop traffic classification system based on the features of the captured packets on the network. This type of classification will assist the IT administrators to determine the unknown attacks which are broadening in the IT industry.

Keywords*—Data Mining,Machine learning,Networksecurity,IDS,Attacks, Malicious, Classification.*

*****

I. INTRODUCTION

As we all know in the current modern network the size of the captured network data is growing exponentially, so there is a greater need to apply the classification algorithms to the collected data set which helps in determining the set of malicious and normal traffic. This type of classification is important for the purpose of network monitoring systems and security incidents[6]. Later well assigned port numbers were used for the purpose of identification of network traffic. For example port 80 is used for HTTP communication and port 25 for SMTP communication. But in currentwith fastgrowing internet, applications are using dynamic changed port numbers which is making the port based traffic classification a tedious job. After port based classification of network traffic, payload based inspection came into play. This classification can achieve good accuracy once the payload can be accessed and inspected properly. In spite of good accuracy the payload based classification has its own limitations in terms of slowness and resource consumptions. In research community, [1] some authors proposed automatic mechanisms for derivation of payload features and proved some promising results,[2] but these approaches still have their own limitations. The methodologies discussed in it depend and require large amount of memory and processing time. But if we inspect only initial few bytes of the payload than it requires less amount of memory and processing time[3]. With the change in technology the size of the network data is increasing day by day, now the researchers have been using machine learning techniques based on the features to classify data. Machine learning based algorithms create the classification model by using the large data set and calculated features. [4] Moreover, the statistical properties based features of the network traffic is also becoming important for machine learning based classifications such as packet length statistics for a network traffic flow, for example the minimum, mean, maximum, standard deviation of the packet sizes. With the consideration of the Machine learning(ML) based techniques and based on these calculated features statistics, a good traffic classifier can be developed[5]. While ML classifiers have shown good efficiency and promising accuracy, accuracy is often lower than that of payload-based classifiers (for traffic for which payload signatures exist).

*A.*AIM OF THE PROJECT

The aim of the proposed work is to perform the traffic monitoring based on machine learning techniques, analyze the network attack logs to determine the intruders and build the traffic classifier for the determination of malicious and normal traffic from the built data set. Live data packets capturing through DOS attack will benefit the analysis of networkbased classification.

*B.*OBJECTIVE OF THE PROJECT

1.Automated network data capturing and logging mechanism.
2.Feature extraction of captured data.
3.Data pre-processing engine to extract relevant & attack data features.
4.Data Analysis and classification based on “R” tool.
5.Performance measurement and result analysis.

*C.*SCOPE OF THE PROJECT

Previous work proves that machine learning techniques can be successfully applied to network traffic analysis. The research work can be extended by implementing various machine learning algorithms and hybrid algorithms for intrusion detection.

II. LITERATURE SURVEY

WernhuarTarng, Cheng-Kang Chou and Kuo-Liang Ou performed detection of P2P botnet viruses in the infection stage and report to network managers to avoid further infection [13]. The system adopted real-time flow identification techniques to detect traffic flows produced by P2P application programs and botnet viruses. The experimental results showed that the accuracy of Bayes Classifier was 95.78% and that of NN Classifier was 98.71% in detecting P2P botnet viruses and suspected flows to achieve the goal of infection control.

The supervised machine learning model was built to differentiate between benign P2P applications and P2P botnets that could also detect unknown P2P botnet traffic with high accuracy. The three classifiers such as Decision trees, Random forests, and Bayesian network were used to consistently detect P2P botnets with a recall ranging between 88% to 95% and achieved a low false positive rate of 0.2% to 0.3%.

The authors, PijushBarthakur, ManojDahal and Mrinal KantiGhose presented a comparative analysis of machine-learning based classification of botnet command & control(C&C) traffic for proactive detection of Peer-to-Peer (P2P) botnets[11]. In this paper three models like Decision Tree (C4.5), Bayesian Network and Linear Support Vector Machines were used to detect botnet and their performance results were compared. The proposed algorithm produces better accuracy than the original decision tree classifier.

The research community has begun looking for IP traffic classification techniques that do not rely on `well known' TCP or UDP port numbers,orinterpreting the contents of packet payloads [4]. New work is emerging on the use of statistical traffic characteristics to assist in the identification and classification process. This survey paper looks at emerging research into the application of Machine Learning (ML) techniques to IP traffic classification - an inter-disciplinary blend of IP networking and data mining techniques.A number of key requirements for the employment of ML-based traffic classifiers in operational IP networks, and qualitatively critique the extent to which the reviewed works meet these requirements have also been discussed.

R.Kannana and V.Ramani developed a system for botnet detection to identify a botnet activity in a network, based on traffic behavior analysis and flow intervals [10]. The approach was to classify packets based on source IP, destination IP, number of packet, etc., using decision tree classification technique in machine learning. The attribute selection was mainly based on packet attribute and does not consider the data part. The feasibility of the work was to detect botnet activity without having seen a complete network flow by classifying behavior based on time intervals.

The authors, Rajesh Kumar and Tajinder Kaur generated an automated system that contained packet capturing, processing of multiple attack logs, labeling of network traffic based on low level features and delivered a traffic classifier which have classified the normal and malicious traffic [15]. The attack data was collected through honeypot system and normal user browser. The classification algorithm was applied and the model has been built.

III. PROBLEM DEFINITION

A. Existing System

A major threat to the network is the presence of botnet and it is the predominant threat on the Internet today [10]. Botnets are designed to operate in the background, often without any visible evidence of their existence. Over the past decade botnets are heavily used for all kinds of computer crimes such as phishing, distributing pirated media and software, identity theft, adware, stealing information and computer resource and so on. Majority of these attacks are focused on making money through illegal means. Hence finding ways to counter botnets is a challenge of great importance. Botnet detection typically demonstrate uniformity of traffic behavior, present unique communications behavior, and that these behaviors may be characterized and classified using a set of attributes which distinguishes them from normal traffic. Therefore, botnet traffic detection is a significant task in any communication network environment.

Disadvantages of existing system

1.Models that can be not user friendly understood
2.Low accuracy in detecting many kinds of known malicious.

B. Proposed System

In this proposed system two modules are include administrator and user. In first module administrator upload the data non malicious and malicious data in database use Wireshark analyzer and honeypot security mechanism and save data set in pcap format.

In second module user first registration website and login in website then user can enter the input system this data formatted, Pre-processing of data building Weka and Building classifier in R.

System check whether the packet/traffic malicious or not and Sent alert to user.

Advantages of Proposed system

Supervised classification techniques

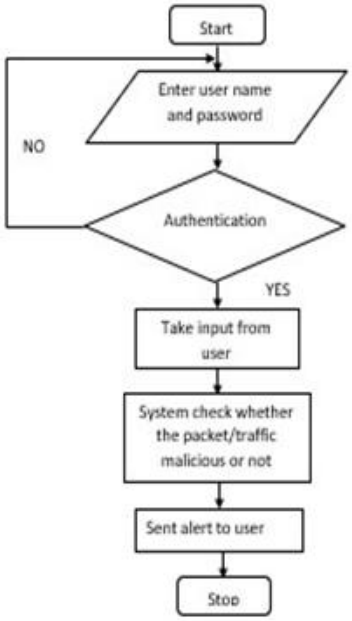
- Models that can be easily understood

- High accuracy in detecting many kinds of known malicious.

Flow chart for website:

Semi-supervised classification techniques

- Models that can be easily understood
- Normal behavior can be accurately learned.



IV. DESIGN & IMPLEMENTATION

System flowchart

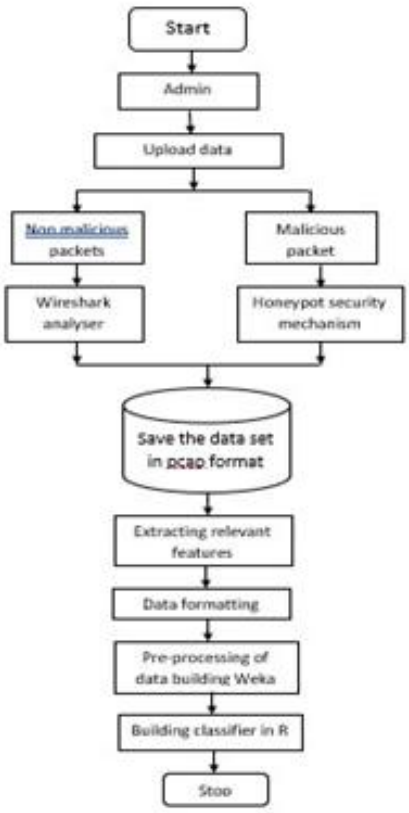


Fig.1 System flow chart

V. ADVANTAGES

Supervised classification techniques

- Models that can be easily understood
- High accuracy in detecting many kinds of known malicious.

Semi-supervised classification techniques

- Models that can be easily understood
- Normal behavior can be accurately learned.

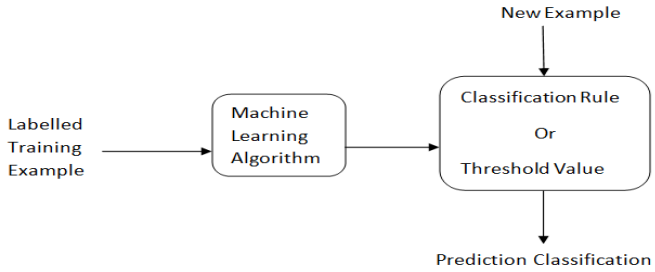
VI. APPLICATION

Using the same scenario of system, we can implement following mobile applications*:*

- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining

VII. Machine Learning Algorithm

Machine learning studies how to automatically discover to make accurate predictions based on past observations. This type of algorithm we provided Knowledge with result for learning purpose and then we providing knowledge algorithm give result on past observation. At the time of learning algorithm makes some rules or threshold value for each classifier. There are many algorithm are used as machine like Decision tree, Naïve Bayesian, Support Vector Machine etc.



Naïve Bayesian algorithm:

Naive Bayesian classifier is a simple classification scheme, which estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label c. The conditional independence assumption can be formally stated as follows:

$$P(A \mid C = c) = \prod_{i=1}^{n} P(A_i \mid C = c)$$

Where each attribute set A = {A1,A2,….,An}consists of n attribute values. With the conditional independence assumption, instead of computing the class conditional probability for every grouping of A, only estimate the conditional probability of each Ai, given C. The latter approach is more practical because it does not require a very large training set to obtain a good estimate of the probability. To classify a test example, the naïve Bayesian classifier computes the posterior probability for each class C.

$$P(C \mid A) = \frac{P(C)\prod_{i=1}^{n} P(A_i \mid C)}{P(A)}$$

The naïve Bayesian classifier has several advantages. It is easy to use, and unlike other classification approaches, only one time scan of the training data is required. The naïve Bayesian classifier can easily handle missing attribute values by simply omitting the probability when calculating the likelihoods of membership in each class.

Support Vector Machines:
Support Vector Machines have been proposed as a novel technique for intrusion detection. A Support Vector Machine (SVM) maps input (real-valued) feature vectors into a higher dimensional feature space through some nonlinear mapping. SVMs are powerful tools for providing solutions to classification, regression and density estimation types of problems. These are developing on the principle of structural risk minimization. Structural risk minimization seeks to find a hypothesis for which one can find lowest probability of error. The structural risk minimization can be achieved by finding the hyper plane with maximum separable margin for the data. Computing the hyper plane to separate the data points i.e. training a SVM leads to quadratic optimization problems. SVM uses a feature called kernel to solve this problems. Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; some of them are Polynomial, radial basis function, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training classifier, which selects support vectors along the surface of this function. SVMs classify data by via these support vectors, which are member of the set of training inputs that outline a hyper plane in feature space. The implementation of SVM intrusion detection system has two phases: training and testing. The main advantage of this method is speed of the SVMs, as the capability of detecting intrusions in real-time is very important. SVMs can learn a larger set of patterns and be able to better scale, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification. The main disadvantage is SVM can only handle binary-class classification whereas intrusion detection requires multi-class classification.

VIII. CONCLUSION AND FUTURE WORK

In the modern internet world of malwares, the traffic classification is not easy, current techniques for intrusion detection have their limitations. There are various tools available for traffic classification. The main aim of the system is to provide optimized output. In the recent years, the field of machine learning has shown significant results which can be used to defend the cybercrime. Here, in this research implementation, we propose “Machine Learning Based Traffic Classification using Statistical Analysis” which implements the network traffic monitoring and classifies the system using machine learning algorithms. We presented some literature and background study to finalize the problem statement. During the implementation, we designed a system whichincludes the various modules such as packet capturing by applying a dos attack, data processing, feature extraction, data labeling and classifier. Each module of the system has its own significance. We also presented the network architecture of the implementation. The data set is created by capturing the packets from applying a dos attack as well as using dataset CTU-13. In future to extend the research to incorporate the traffic related to all the protocols by capturing internet traffic on large networks for longer durations. Data set as one of important entity in the field of data mining for classification, by creating the state of the art dataset, more complex algorithm can be applied. Basic research of which indicate the importance of data mining in network security was shown here which can be further extended for more complex algorithms and statistical mining. We also propose to make the system as real time classifier that will perform the real time classifications of the internet traffic.

IX. ACKNOWLEDGMENTS

I would like to sincerely thank Mrs.AbiramiSivaprasad(Assistant Professor) for her contribution and help in writing this paper.

X. BIBLIOGRAPHY

[1]. P. Haffner, S. Sen., O. Spatscheck, and D. Wang, “ACAS: Automated Construction of Application Signatures,” in ACM SIGCOMM MineNetWorkshop, (Philadelphia, PA, USA: ACM), Aug. 2005.

[2]. J. Ma, K. Levchenko, C. Kreibich, S. Savage, G. M. Voelker, “Unexpected Means of Protocol Inference,” in 6th ACM SIGCOMM Conference on Internet Measurement (IMC), pp. 313–326, 2006.

[3]. A. Finamore, M. Mellia, M. Meo, and D. Rossi, “KISS: Stochastic Packet Inspection Classifier for UDP Traffic,” IEEE/ACM Transactions on Networking, vol. 18, pp. 1505–1515, Oct. 2010.

[4]. T. T. T. Nguyen and G. Armitage, “A Survey of Techniques for Internet Traffic Classification using Machine Learning,” IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56–76, 2008.

[5]. T. T. T. Nguyen, G. Armitage, P. Branch, and S. Zander, “Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic,” IEEE/ACM Transactions on Networking, vol. 20, no. 6, pp. 1880–1894, 2012.

[6]. [6]http://www.iana.org/assignments/service-namesportnumbers/service-names-portnumbers.Xml accessed on 4/04/13.

[7]. Nguyen, T.T.T.; Armitage, G., "A survey of techniques for internet traffic classification using machine learning," Communications Surveys & Tutorials, IEEE , vol.10, no.4, pp.56,76, Fourth Quarter 2008doi: 10.1109/SURV.2008.080406

[8]. http://www.ncftp.com/ncftpd/doc/misc/ephemeral_ports.html accessed on 04/04/13.

[9]. M. Roughan, S. Sen., O. Spats check and N. Duffield“lassof-service mapping for QoS: A statistical signature-based approach to IP traffic classification”Proc. ACM/SIGCOMM Internet Measurement Conference (IMC) 2004, 2004.

[10]. R.Kannana and V.Ramani,“Flow based analysis to identify Botnet infected systems”,2014.

[11]. PijushBarthakur, ManojDahal and Mrinal KantiGhose“An efficient machine learning based classification scheme for detecting distributed command and control traffic of P2P botnets”, 2013.

[12]. A. Madhukar and C. Williamson“A longitudinal study of P2P traffic classification”14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer andTelecommunication Systems, 2006.

[13]. S. Sen., O. Spatscheck and D. Wang“Accurate, scalable in network identification of P2P traffic using application signatures”WWW2004, 2004.

[14]. Callado, A.; Kamienski, C.; Szabo, G.; Gero, B.; Kelner, J.; Fernandes, S.; Sadok, D., "A Survey on Internet Traffic Identification," Communications Surveys & Tutorials, IEEE , vol.11, no.3, pp.37,52, 3rd Quarter 2009.

[15]. OrialMula-Valls “A practical retraining mechanism for network traffic classification in operational environments” June 2011.

[16]. Z. Shi Principles of Machine Learning 1992, International Academic Publishers.