Anomaly detection in network traffic

Shweta Bhartia
F16-IG-3002
Indiana University
Bloomington, IN
sbhartia@umail.iu.edu

Pramod Sripada
F16-IG-3018
Indiana University
Bloomington, IN
ksripada@umail.iu.edu

Ramkaushik Nallani
F16-IG-3015
Indiana University
Bloomington, IN
rnallani@umail.iu.edu

ABSTRACT

Biggest threat in twenty first century is Cyber war. Hackers are trying to intrude into network and hack the systems present in that system thereby stealing the sensitive information. This urges the necessity of detecting anomalous or malicious activities in network so that proper actions can be taken. Anomaly detection in network traffic is challenging task. The anomaly detection classifiers or the anomaly detection models built should be sensitive in not misclassifying normal usage as threats. Too many misclassifications of normal usage results in the ignorance of all the notifications which puts entire system at stake. This paper speaks about data mining techniques devised to detect anomalies in network traffic. Supervised learning methods such as classification can be used in detecting known threats, i.e. the threats which have been experienced in historic data. Anomalies fall under "unknown-unknown" category. Hence supervised techniques cannot be used to find the unknown unknown. Anomaly detection in network traffic is challenging task. Data mining techniques make it possible to search large amount of data for characteristic rules and patterns. In this paper, we build an anomaly classifier using classification methods such as Decision Trees and Random Forest Classifiers, discuss its challenges and then present an anomaly detection algorithm which will be built on K means clustering. Anomaly classifier has the ability to classify earlier patterns but the application of K means technique on training dataset results in the formation of clusters. Corresponding cluster centroids are used as patterns for detection of anomalies in new monitoring data. Proper care has to be taken while building the models which should not be overfitting as small changes in the testing data can render the model useless. We precisely described about the data mining methodology devised for anomaly detection in network traffic in this paper.

General Terms

Network Intrusion, Computer Security, Network Security, Data Mining, Decision Trees, Random Forest Classifier, K-means clustering, Apache Spark, HDFS, MLlib

1. INTRODUCTION

Advancement in technology and advent of Internet of things has resulted in the huge generation of data. Though storage cost of data is technically reducing day by day, extracting information from vast datasets has really turned into real time challenging problem. Data mining, which strategically amalgamates two different fields namely computer science and statistics is helping in extracting information from the given large sets of raw data. Data scientists find hidden patterns by applying these statistical techniques and forecast future of client's business. Data mining techniques are attracted as they can be applied to any type of dataset to learn about correlations. Cyber-attacks are growing at very fast pace these days. Most of the attacks attempt to flood a computer with network traffic to gain unauthorized access to any computer. However, if the exploit behavior follows any patterns, they can be detected very easily either by writing simple business rules or through supervised techniques. For instance, if hacker is trying to access services being running in system, he tries accessing ports repeatedly. Hence if a rule is written which throws notification if large number of distinct ports were accessed in short time, attack can be caught. This falls under know known category. But above mentioned techniques cannot be used for detecting unknown unknowns. Detecting anomalies is the main part in detecting network intrusions. These are connections that aren't known to be attacks, but do not resemble connections observed in the past. Hence known supervised methodology can be used only for classifying known attacks and assist the organizations in building rule based systems. Unsupervised learning techniques can be used in detecting unknown attacks, hence this would be a better technique to be used in identifying network intrusion. Intrusion detection systems process large amount of monitoring data. Some of the most vital Intrusion detection systems are network based Intrusion Detection Systems and Host based Intrusion detection systems. Network based IDS detects harmful packets or packet flows by exploring network monitoring data while host based IDS detects suspicious activities by exploring log files on computer. Advancement of research and development activities in data mining area in late 1990s resulted in the development of better methods for network and host based intrusion detection. Data mining techniques can be used for detecting anomaly and attack detection by monitoring network data. Anomaly detection can be done in two methods: 1) Classification techniques and 2) Cluster-

ing techniques ML methods are used to train the model to determine what could be the notion of normality. Using classification techniques, certain rules are defined to identify abnormal records. When any of these rules are triggered for a certain connection, it is flagged abnormal. ML based techniques are adaptive, dynamic and requires less human intervention, so they are preferred. Anomaly detection is very intolerant to errors as falsely classifying normal records as attacks or falsely ignoring the attacks renders the system totally useless. Time spent by network security engineer on fixing the anomalies detected is linearly dependent on number of anomalies detected. Hence model should be designed in such a way that anomalies have to be filtered out in best possible way. In this paper, we present an approach for anomaly detection using Decision Tree technique, Random Forest Classifier and K-Means clustering. Reuters and network monitors which use either Cisco Netflow protocol or IPFIX protocol are the major sources of data related to computer networks. All the data collected constitute flow records. Flow records can be very easily collected as flow monitoring techniques are already deployed in computer networks for administration purposes. Network data mining extracts valuable information from monitoring data which helps in determining dominant characteristics and identifying outliers within the data records. Data mining techniques can also be used to develop rules which are typical for special kind of traffic. The Dataset used for this project is from the 1998 DARPA Intrusion Detection Evaluation Program which was prepared and managed by MIT Lincoln Labs. Lincoln Labs set up an environment to acquire nine weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks. The raw training data is about five million connection records. Similarly, the test data is around two million connection records. A connection is a sequence of TCP packets starting and ending at some well-defined times, between which data flows to and from a source IP address to a target IP address under some well-defined protocol. Each connection is labeled as either normal, or as an attack, with exactly one specific attack type. The task is more realistic as it includes specific attack types not in the training data. This makes the task more realistic. The datasets contain a total of 24 training attack types, with an additional 14 types in the test data only. The dataset is in comma separated value format. We will be performing feature reduction, feature extraction and feature scaling on the dataset for improved performance. [5] [3] [2] [1] [4]

2. TIMELINE

1. Oct3 - Oct8: Use Case and data understanding. Loading the data into HDFS
2. Oct9: Oct16: Perform exploratory data analysis
3. Oct17: Oct22: Perform feature scaling and feature selection
4. Oct23- Nov 5: Building machine learning models
5. Nov 6- Nov 13: Validating and tuning machine learning models
6. Nov14- Nov 21: Visualizing the results in D3.js
7. Nov 22- Nov 30: Documenting the results and final paper submission

3. REFERENCES

[1] Anomaly-based intrusion detection system. Web Page.

[2] Network behavior anomaly detection. Web Page.

[3] R. R. Amuthan Prabakar Muniyandi, R. Rajeswari. Network anomaly detection by cascading k-means clustering and c4.5 decision tree algorithm. Volume 30, 2012:174–182.

[4] W. L. A. P. Salvatore J. Stolfo, Wei Fan and P. K. Chan. kddcup task. Web Page.

[5] M. Wazid. Hybrid anomaly detection using k-means clustering in wireless sensor networks. pdf.