Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(1), 2014, 96-101

ISSN 2347 – 7393

Original Research Article

PATTERN BASED SECURITY USING MACHINE LEARNING TECHNIQUES

Anagha Khati, Auzita Irani, Naba Inamdar, Rashmi Soni

BE Comp,Modern Education Society's College of Engineering, Pune.

Abstract-
Network security has become a very important aspect in today's internet enabled systems. As the internet keeps developing, its use on public networks, the number of security attacks as well as their severity has shown a significant increase. An Intrusion Detection System can provide an additional layer of security to these systems and ensure the protection of data. The goal of an intrusion detection system is to identify those entities that attempt to destabilize the security controls that are already in place. The field of machine learning is gaining rising attention in the development of these intrusion detection systems. Machine learning techniques that are used for solving intrusion detection problems can be broadly classified into three broad categories: Supervised, Un-supervised and semi-supervised. The supervised learning method exhibits good classification accuracy for those attacks that are aready known. But this method requires a large amount of training data. In the real world, the availability of labelled data is not only time consuming but also costly. The emerging field of semi-supervised learning offers a promising direction for further research. So in this project we propose a semi-supervised approach for a pattern based IDS to improve performance and to reduce the false alarm rate. The experimentation is performed on KDD CUP99 dataset.

Keywords: KDD CUP99, intrusion detection, semi-supervised learning, unsupervised learning, supervised learning

For Correspondence:
anagha.khati@gmail.com
Received on: February 2014
Accepted after revision: February 2014
Downloaded from: www.johronline.com

INTRODUCTION

As the Internet continues to enjoy widespread use in our day to day lives, a variety of attacks, such as Denial of Service (DoS), Probing, Spam and so on, emerge on a regular

basis. IDS is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station. Intrusion detection system (IDS) is an effective method to guard against malicious attacks and has been widely used. Intrusion detection and prevention systems are primarily focused on identifying possible incidents, logging information about them, and reporting attempts. It attracts many research interests and we have attempted to add to this large research effort. Our project is an attempt to reduce false alarm rate for Intrusion Detection System by using Machine Learning algorithm. We aim to design IDS by using machine learning which can meet the demands of Reducing False Alarm Rate with higher detection rate. Many types of IDS already exist in the world which provides assistance at different stages of project development. But a problem commonly observed is the high False Alarm Rate. Our software should be able to assist the developers in this department greatly along with the individual stage support.

RELATED WORK AND MOTIVATION

Semi-supervised learning methods use unlabeled data to either modify or reprioritize premises obtained from labeled data alone. Recently, learning with labeled and unlabeled data, also known as semi- supervised learning has drawn much attention. It aims to attain good classification performance with the assistance of unlabelled data in the presence of the small sample problem, and a few promising results have been reported. Therefore, instead of training the model with only labeled data, we incorporate the unlabelled data before active learning starts. "A Comprehensive Analysis and study in Intrusion Detection System using Data Mining Techniques", G.V. Nadiammai, S.Krishnaveni, M. Hemalatha [9] have been referred in order to understand the use different data mining techniques in order to implement an intrusion detection system. In this paper, they are provided with a summary of the current research directions in detecting such attacks using collaborative intrusion detection systems (CIDSs). "A survey on intrusion detection techniques", Sandip Sonawane, Shailendra Pardeshi and Ganesh Prasad [4] are presented with three types of intrusion detection based on the source of detection – host based, network based and hybrid intrusion detection and also focuses on intrusion detection techniques that is, misuse detection and anomaly detection techniques, supervised and unsupervised based learning based on the various approaches

.

Table I
Feature categories in kddcup99 dataset

Categories	Features
TCP basic features(1~9)	duration, protocol type, service, flag, src_bytes, dst_bytes, land, wrong fragment, urgent
TCP content features(10~22)	hot,num_failed_logins, logged_in,num_compromised, root_shell, su_attempted,num_root, num_file_creations, num_shells,num_access_files, num_outbound_cmds,is_hot_login, is_guest_login

TCP Traffic features(23~31)	count, srv_count, serror_rate, srv_seror_rate,rerror_rate, rv_rerror_rate, same_srv_rate,diff_srv_rate, srv_diff_host_rate
Host-Based Network Traffic(32~41)	dst_host_count, dst_host_srv_count,dst_host_same_srv_rate, dst_host_diff_srv_rate,dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate,dst_host_srv_serror_rate, dst_host_srv_serror_rate,dst_host_rerror_rate, dst_host_srv_rerror_rate

PROBLEM DEFINITION

There are various types of learning methods such as supervised, semi-supervised and unsupervised learning. Supervised learning refers to training the system with labeled data and un-supervised learning refers to training the system with unlabeled data. While semi-supervised learning denotes to training the system with labeled as well as unlabeled data. The supervised learning method exhibits good classification accuracy for known attacks. But the main problem with it is that it requires a large amount of training data. In the real world, the availability of labeled data is time consuming and costly. An emerging field of semi- supervised learning offers a promising direction for further research. When the unlabeled data is used in combination with a small amount of labeled data it can produce substantial improvement in learning precision. Hence we are making use of semi-supervised learning. We will employ the use of the KDD CUP 1999 data set for our project. It was devised at MIT's Lincoln Lab and developed for IDS evaluations by DARPA. It represents the activities at US Air Force local area network (LAN), which have normal traffic and malicious activities that were injected in the datasets. In spite of several drawbacks, it has served as a dependable benchmark data set for many researches on network based intrusion detection algorithms.

TYPES OF ATTACKS

Various types of attacks are possible on a system and to construct an efficient intrusion detection system we have to take into account all of them. A DOS attack is a denial of service attack. a denial-of-service (DoS) or distributed denial-of-service (DDoS) attack is an attempt to make a machine or network resource unavailable to its intended users. U2R attack refers to an unauthorized access to the local super user (root) privileges. The R2L attack is one in which there is unauthorized access to the local super user (root) privileges. Another type of attack is the probe attack.

CLASSIFICATION OF ATTACKS

The KDD Cup 99 data set contains 23 different attack types. Their names are shown in Table II and its features are grouped as Basic Features, Traffic Features and Content Features.

Basic features contains all the attributes of the TCP/IP connection and leads to delay in detection.

Traffic features are evaluated according to

the window interval and two features as same host and same service.

(i) Same host feature:- It examines the number of connections in the past 2 seconds from the same destination host.

The probability of connections will be done in a specific time interval.

(ii) Same service feature:- It inspects the number of connections in a particular time interval that possesses same service.

Content features: Dos & probe attack have frequent intrusion sequential patterns compared to R2L & U2R. These two attacks include many connections to several hosts at a particular time period whereas R2L and U2R achieve only a single connection. In order to detect these types of attacks, domain knowledge is important to access the data portion of the TCP packets.

Table II
Name of the attacks classified under four groups

Denial of Service	Back, land, neptune, pod, smurf, teardrop
Probes	Satan, ipsweep, nmap, port sweep
Remote to local	ftp_write, imap, guess_passwd, phf, spy, warezclient, multihop, warezmaster
User to root	Buffer_overflow, load module, Perl, root kit

CLASSIFIER AND ALGORITHM SELECTION

Decision tree induction is the learning of decision trees from class labeled training tuples. A decision tree is a flow chart like tree structure where each internal node denotes a test on an attribute. Each branch represents an outcome of the test and each leaf node holds a class label. Weka is an open source data-mining tool that supports only supervised and clustering algorithm. One can integrate java code implementations of semi-supervised algorithms in Weka. The semi-supervised J48 algorithm comes under the decision tree classifier. We have selected this algorithm as it has the highest accuracy. One of the main benefits of the J48 classifier is that is relatively quick to train, and should finish almost immediately on a small data set. In this algorithm classification model is formed according to the training set and later the unlabeled set can be labeled according to the model. Our implementation will be capable of various options like accepting the training and testing set. The tree can be printed in pruned as well as un-pruned form.

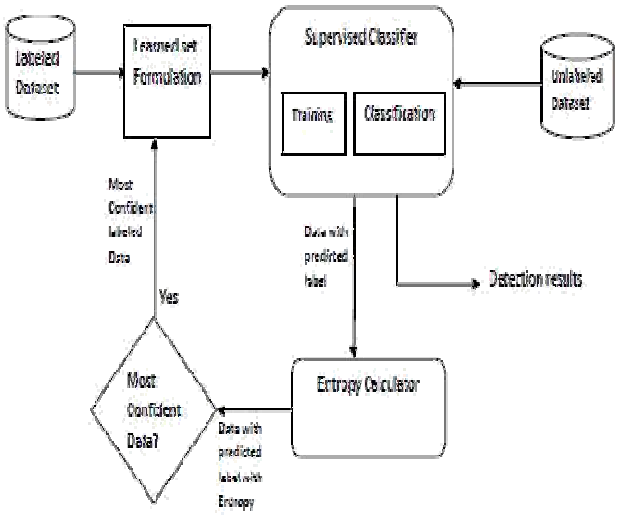

FIG. I Architecture for semi-supervised learning

PROPOSED ARCHITECTURE

We propose an architecture that will have a number of different modules. The first module will be a training module. We will provide labeled datasets to the system. The next module will be a testing module. In this module unlabeled data will be used in order to test the trained system. The mixed module will contain both labeled as well as unlabeled data. An algorithm will be used for classification. Entropy calculation will be used to select the most confident data and a semi-supervised module will be created. The most confident data will be selected by choosing a threshold value of the entropy values. In this way both labeled as well as unlabeled data will be used in order to create the semi-supervised learning system.

ENTROPY CALCULATION

Entropy needs to be calculated for each data tuple to identify the most confident data. The confident data obtained is then combined with train set and filtering is done. Entropy of a tuple D is given by,

$$E(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

Where d is a data packet, m is the number of attributes and Pi is the probability of the i^{th} attribute. According to these entropies of each packet, the most confident data will be chosen which will be decided according to a threshold value. This data will then be added to the training set hence enhancing it.

CONCLUSION

We have proposed an algorithm for semi-supervised learning using decision tree classifier. The strength of our proposed algorithm lies in its ability to improve the performance of any given base classifier in the presence of unlabeled samples. Our algorithm will be capable of giving higher accuracy than already available algorithms and will also ensure that there will be a lower false alarm rate.

FUTURE WORK

This system can be put to use as a real-time application in networking. We can also use the system for different data-sets format. It also finds applications in detecting new attacks. We can use more machine learning algorithms to improve accuracy. Based on this work, a recurring scan cloud can be created to quickly monitor which services and programs run on a machine allowing for an even more precise rule set. So instead of a sensor capturing all traffic on a network, the client machines will monitor their own traffic. The clients will automatically report data to centralized

monitoring station.

REFERENCES

1. Sharmila Kishor Wagh, Vinod K Pachghare and Satish R Kolhe. "Survey on Intrusion Detection System using Machine Learning Techniques." *International Journal of Computer Applications* 78(16): 30-37, September 2013. Published by Foundation of Computer Science, New York, USA
2. Sharmila Kishor Wagh, Gaurav Nilwarna, S. R. Kolhe, "A Comprehensive Analysis and Study in Intrusion Detection System Using KNN Algorithm", the 6th multidisciplinary workshop on Artificial Intelligence 2012 (MIWAI 2012), organized at Ho Chi Minh city, Vietnam.
3. V. K. Pachghare, V. K. Khatavkar, and Dr. Parag Kulkarni, "Pattern based network security using semi-supervised learning," *IJINS*, vol. 1, no. 3, pp.228-234, August 2012.
4. Sandip Sonawane, Shailendra Pardeshi, Ganesh Prasad, "A survey on intrusion detection techniques", World journal of science and technology, vol. 2, pp.127-133, 21st April 2012.
5. Jiawei Han, Micheline Kamber, Jian Pei "Data mining: Concepts and Techniques" Elsevier publishing.
6. Dorothy E. Denning, "An Intrusion-Detection Model," IEEE transactions on software engineering, vol. SE-13, no. 2, pp.222-232, Feb 1987
7. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Marcia-Fernandez, E. Vazquez "Anomaly-based network intrusion detection: Techniques, systems and challenges," Elsevier publishing, Computer and security vol.28, pp.18-28, August 2008.
8. C. Kemp, T. Griffiths, S. Stromsten, J. Tenenbaum, "Semi-supervised learning with trees" Advances in Neural Information Processing System, 2003
9. G. V. Nadiammai, S.Krishnaveni, M. Hemalatha, "A Comprehensive Analysis and study in Intrusion Detection System using Data Mining Techniques"