Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(1), 2014, 126-129johronline

ISSN 2347 – 7393

Original Research Article

AN ANOMALOUS REAL-TIME INTRUSION DETECTION SYSTEM USING MACHINE LEARNING ALGORITHM

Deepak C Mahajan*, Sanket A Sancheti, Prasad R Jadhav, Atul M Nikhade, Sharmila K Wagh

BE Comp, MESCOE, Pune-01

Abstract-
With the growth of the users of internet, the number of security threats are also increasing. These can be dealt with by deploying the intrusion detection system. It can be useful in effective detection of intruded access. In this paper, we aim to create a Real Time Intrusion Detection System (RTIDS) that would facilitate in detecting attacks against computer systems and networks. This RTIDS is designed to detect system attacks and classify system activities into normal and abnormal form with respect to their behavior. Machine learning techniques which have an important role in detecting intrusions have been applied to our RTIDS. This paper also helps to clarify the system design of an Intrusion Detection System (IDS) to reduce false alarm rate and improve accuracy to detect intrusion.

Keywords- Real Time Intrusion Detection System (RTIDS), Machine Learning Techniques, Anomaly Detection, False Alarm Rate (FAR).

Introduction

Many kinds of systems over the Internet such as online shopping, Internet banking, foreign exchanges, and online auctions etc, have been developed. Thus due to the open society of the Internet, the security of computer system and its data is always at risk. Therefore, the security of computer networks has been in the focus of research for years. The organizations have come to realize that information & network security technology has become very important in protecting information. This extensive growth of Internet has prompted network intrusion detection to become a critical component of protection mechanisms. We can define network intrusion detection as identifying a set of malicious actions that

For Correspondence:
deepakcmahajan92@gmail.com
Received on: February 2014
Accepted after revision: February 2014

threaten the integrity, availability, and confidentiality of a network resource. Industries face new security challenges frequently, which they can tackle through the IDS.

Intrusion detection mechanism is divided into the two categories: misuse detection and anomaly detection. Misuse detection searches for specific patterns or sequences of programs and user behaviors that match well-known intrusion scenarios, i.e. they can detect many or all known attack pattern, but the weakness of misuse based intrusion detection systems is that the incapability of identifying new types of attacks or variations of known attacks. On the other hand, anomaly detection develops models of normal network behavior and then new intrusions are detected by evaluating significant deviations from the normal behavior, i.e. the normal behavior of system or network traffic are represented and, for any behavior which varies over a pre-defined threshold, an anomalous activity is identified. Thus the main advantage of anomaly detection is that it may detect novel intrusions that have not been observed yet.

In anomaly based IDS, the number of false positives generated are higher than that of those based on signatures. An important issue in anomaly based system is how these systems should be trained, i.e., how to define what is a normal behavior of the system or network environment (which of the features are relevant) and how to represent the behavior computationally.

Nowadays Machine Learning Intrusion Detection System has been giving high accuracy and good detection of novel attacks. Intrusion detection system (IDS) is a network security technique attempting to detect various attacks. Machine learning is concerned with the design and the development of algorithms and methods that allows computer systems to autonomously infer and integrate knowledge to continuously improve them to finish its tasks efficiently and effectively.

Motivation

Real-time Intrusion Detection and Classification

In this paper, the author has proposed a Real-Time Intrusion Detection System (RT-IDS) using Decision tree technique to classify an online network data that is pre processed to have only 13 features. The number of features affects to the RT -IDS detection speed and resource consumption. In addition RT-IDS can classify normal network activities and main attack types consisting of Probe and Denial of Service (DoS). Hence, it helps to decrease time to diagnose and defense each network attack.

Survey on Intrusion Detection System using Machine Learning Techniques

In this paper authors has represented an overview of the machine learning techniques which are being utilized for the attack detection in Intrusion Detection System and effective design system of effective IDS. In today's environment security of the information in computer based systems is a major concern to researchers. The work of IDS and various methods which has been a major focus of the information security related research. Though machine learning is a vast and advanced field still it is relatively not so mature and not optimized for IDS.

RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks

The RT-UNNID system is system, which is capable of intelligence, using unsupervised neural network, the real time IDS is build. Unsupervised neural nets can improve their analysis of new data without retraining. In previous their work, they have evaluated Adaptive Resonance Theory (ART) and Self-Organizing Map (SOM) neural networks using offline data. In this paper, they presented a real-time solution using unsupervised neural nets to detect known and new attacks in network traffic. They have evaluated approaches using 27 types of attack, and observed 97% precision using ART nets, and

95% precision using SOM nets.

Evaluating machine learning algorithms for detecting network intrusions

In this paper they mainly focused on detecting network intrusions. They employ ensemble algorithms in modeling network intrusion detection systems, to improve detection performance. In this they described the methods employed in their proposed framework and given how to apply these methods to build an efficient intrusion detection system model. They are given overview of the framework and ensemble learning methods like Ad a Boost, Random Forest, Naïve Bayes. They presented Naïve Bayes algorithm also in order to compare their earlier results on the proposed method, to find its suitability in building an efficient network intrusion detection model.

Problem Description

Our paper An Anomalous Real-Time Intrusion Detection System Using Machine Learning Algorithm is an attempt to detect the attack in the online system and reduce false alarm rate for Intrusion Detection System by using Machine Learning algorithm. We aim to design a RTIDS by using machine learning which can meet the demands of Reducing False Alarm Rate with higher detection rate in real time system.

This attempt is to mainly reduce the false alarm rate (FAR) as compare to the legacy system as discussed before. With the aim to detect the novel attacks which are not known to the security system.

Solution and Experiment

This intrusion detection system can analyze the captured packets and detect whether it would be an intrusion or not. The main assumption for this experiment is used is that the nature of the abnormal or anomalous traffic is different from that of the normal network traffic. From the view of architecture, the diagram of system includes several modules, which has shown in Figure 1.

7. Packet Monitor :

This module captures packets from the network to serve for the data source for the IDS.

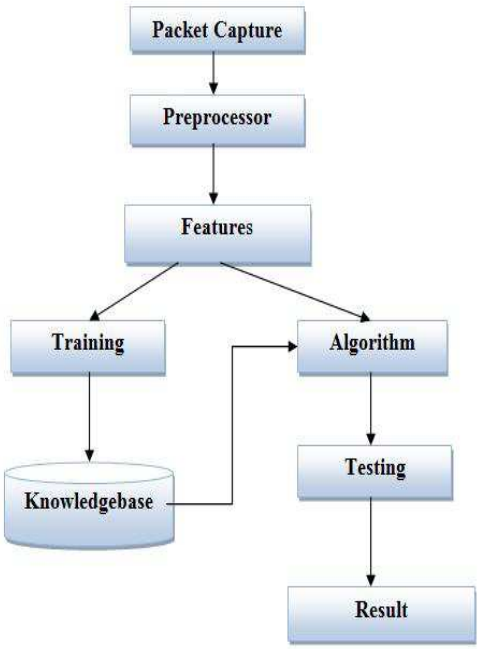

Fig. 1 Basic Architecture of Anomalous Real Time Intrusion Detection System

Pre-processor :
In pre-processing phase, dataset is collected and processed for use as input to the system.
Feature Extractor : This module extracts feature vector from the packets and submits the feature vector to the classifier module.
Algorithm : The function of this module is to analyze the packet stream and to draw a conclusion whether the given packet is attack or normal.
Result :While detecting the intrusion, this module will send a warning message to the user.
Knowledgebase : This module serves for the training samples of the classifier phase. The intrusion samples can be perfected under user interference, so that the capability of the detection can be improved.
The detailed explanation of the experiment is given as follows : A. Preliminary Setups:
For the experimental purpose we have created the attack by attack generating tool such as LOIC. This tool generates the DOS attack from the source machine to the destination machine. Denial of Service (DOS) attack is an attempt to make a computer system or network resource unavailable to its expected users. Here at destination machine we are doing the packet capturing.

B. Packet Capturing:

At the starting packet monitoring phase, extraction of packet features is done with the help of packet sniffing tools such as Wireshark, Caspa. We have used Wireshark packet capturing tool to capture network packets such as IP, ICMP, TCP, UDP.

Wireshark is a network analyzer tool which is used to read packets from the network, then decodes, and presents them in the easily understandable format. The most important aspect of Wireshark is that it is open source, actively maintained, and free. These packets captured from the wireshark tool are saved in the text file format for the pre processing phase.

C. Pre processing :

This phase of the pre processing deals with the file conversion i.e. the packets captured from that of the wireshark tool in the text format are not understandable. Thus, this file is changed to the another format with the help of the CSV converter it is java code which convert the packets in the CSV format which can be easy to understand and use by the code for the further packet analysis purpose.

D. Feature Extraction:

The captured packet of the wireshark tool has mainly 24 attributes which can be further used for analysis. This attributes can be of various types such as categorical or continuous. The nature of this attributes helped us to determine the applicability of anomaly detection techniques. Considering overall scenario some features/attributes does not play any specific role while observing the anomalous behavior such as source address, destination address, time etc. So that at this phase the main aim is to select the features such that these are contributing the analysis of the networks behavior here features like type, version and header length, Differentiated service field, Total length, flags, checksum etc.

E. Classification

In this phase the data which is received from the previous phase for analyzing whether the packet is normal packet or attack packet. The selected feature values and the algorithm will help classify the packet into their familiar groups. It consists of Training Phase and Testing Phase

a. Training Phase:

The training phase consist answer class provision along with packet features, that will help to formulate ways to decide mapping domains. We can change is ways or rules depending as per the future training. Every algorithm has its own strategy of

classification. Here we are using the decision tree algorithm for this.

b. Testing Phase

The Testing Phase, consist of the captured network data which is then given to the intrusion detection system to sample whether the given packet is normal or attack. This training phase process is performed by providing the input as packets without the specification of the answer class.

F. Machine Learning Algorithm:

The motivation to apply machine learning techniques for intrusion detection is to build the model based on the training data set automatically. This data set contains the collection of data and its instances, each of can be described using a set of attributes (features) and the associated labels.

Here for the training purpose we are using the DECISION TREE algorithm which will create the tree specifying the trained input to the testing phase. Decision tree learning algorithm is being very successfully used in the experts Systems to capture the knowledge. Our expected task is performed in these systems is by using the inductive methods to provided values of attributes of an unknown object to achieve appropriate classification as per the decision tree rules. It is extended the domain to the values as numeric values and discrete values rather than the Boolean values.

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.

A decision tree classifies instances by traversing from the root node to leaf node. We are starting from root node of tree, while testing the attribute specified by root node, then by moving towards the down to the tree branch according to its attribute value in the given set. The same process is again repeated the sub-tree levels. Thus, resulting generated decision tree gives the concept which appeals as it renders the self-evident classification process.

A decision tree built from the fixed ‘n’ attributes. The leaf nodes of the decision tree contain answer class whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch can be a possible value of the attribute. This Decision Tree uses information gain and help to decide which attribute goes into a decision node. This is the advantage of a decision tree program, rather than a knowledge engineer, which elicits the knowledge from an expert.

a) Data Description

The sample data used by Decision Tree has following requirements:

- Attribute value description - The same attributes must describe each example and have a fixed number of values.
- Predefined classes – The attributes must be already defined, i.e. they are not learned by algorithm.
- Discrete classes - Classes must be represented precisely. Continuous classes broken to the vague categories such as we can say metal being "hard, quite hard, flexible, soft, quite soft" are suspect.
- Sufficient examples - As inductive generalization is used there must be enough test cases so as to distinguish the valid patterns from chance occurrences.

b) Attribute Selection

To know the best attribute this algorithm has a statistical property, which is called as information gain, is used. We can measure gain as how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. Here to define gain, we first borrow an idea from information theory called entropy. Entropy is used to measure the amount of information in an attribute.

Given a collection S of C outcomes

$$( ) = \sum - ( ) \log_2 ( )$$

Where, ( ) is the proportion of S belonging to class I. ∑ is over C.
Here, S is an entire sample set.

Example

If S is a collection of 20 examples with 11 YES and 9 NO examples then

Entropy(S) = - (11/20) Log2 (11/20) - (9/20) Log2 (9/20) = 0.4895

Notice entropy is 0 if all members of S belong to the same class (the data is perfectly classified).
The range of entropy is 0 ("perfectly classified") to 1 ("totally random").
Gain(S, A) is information gain of example set S on attribute A is defined as

$$Gain(S, A) = ( ) - ((| |/| |) * ( ))$$

Where:
S is each value v of all possible values of attribute A
Sv = subset of S for which attribute A has value v
|*Sv*| = number of elements in *Sv*
|*S*| = number of elements in *S.*

On the basis of value of gain for each attribute the attribute selection takes place. The attribute which is having more gain value is select as the decision node and the further selection of child nodes is also done in the same way. By analyzing all the attributes the decision tree is formed and this is considered as a result of the training phase. This tree is generated from the set of training packet sets. This generated tree trains our IDS for attack detection. After this the packets which are to be tested are given as an input to the decision tree and on the basis of attributes of the packets in the file the decision for them is generated whether the packet is attack or not. This phase is called as testing phase in the project.

CONCLUSION

After studying the existing work done on the intrusion detection, in this paper we have presented an Anomalous Real-Time Intrusion Detection System Using Machine Learning Algorithm as Decision tree. This can be successfully implemented for the detection of attacks in real time intrusion detection system. As machine learning is a vast and advanced field and is relatively immature and not so optimized for IDS. Thus, security of information in computer based systems can be achieved with the help of this system maximally with the detection of the novel attacks by decision tree machine learning algorithm.

FUTURE WORK

In recent some of the years, the challenges which lie ahead of us in intrusion detection system are huge, on which following future work is to be done.

I. The very large amount of the data can be capture with the less processing time.
II. Accuracy can be increase when amount of testing data is in large amount.
III. Attribute selection process can be done at time of capturing of packets only.
IV. As we are capturing data and then processing it manually it can be done automatically such way that system will capture packets itself.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, "*Data mining concepts"*

[2] S. K. Wagh, V. K. Pachghare, S. K. Kolhe, "*A survey on intrusion detection system using machine learning techniques*", IJCA(0975-8887) Vol. 78.

[3] Wei Peng, Juhua, Chen, Haiping Zhou, "

An implementation of ID3-Decision tree machine learning algorithm " Project of Comp 9417.

[4] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn, Chalermpol Charnsripinyo, "*Real time intrusion detection [6]*

and classification".

[5] Morteza Amini, Rasool Jalili, Hamid Reza Shahriari, "*RT-UNNID: A Practical solution to real time network based intrusion detection using unsupervised neural network*".