ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2017,
Vol. 10, No. (2):
Pgs. 491-496

Intrusion Detection System Based on Data Mining Techniques

ABHINAV KUMRA, W. JEBERSON and KLINSEGA JEBERSON

Department of Computer Science and Information Technology, SHUATS

http://dx.doi.org/10.13005/ojcst/10.02.33

(Received: May 15, 2017; Accepted: June 01, 2017)

ABSTRACT

Network security is one of the most important non-functional requirements in a system. Over the years, many software solutions have been developed to enhance network security. Intrusion Detection System (IDS) we have provided an overview of different types of intrusion Detection Systems, the advantages and disadvantages of the same. The need for IDS in a system environment and the generic blocks in IDS is also mentioned. The examples are as follows: (1) Misuse intrusion detection system that uses state transition analysis approach, (2) Anomaly based system that uses payload modeling and (3) Hybrid model that combines the best practices of Misuse and Anomaly based intrusion systems.

Keywords: Intrusion Detection System, Web log files, J48 decision tree.

INTRODUCTION

An intrusion detection system (IDS) is a software tool that monitors network and/or system activities for malicious activities or policy violations and produces reports to a Management Station. Some systems may attempt to stop an intrusion attempt but this is neither required nor expected of a monitoring system. Intrusion detection is primarily focused on identifying possible incidents, logging information about them, and reporting attempts. In addition, organizations use IDS for other purposes, such as identifying problems with security policies, documenting existing threats, and deterring individuals from violating security policies. IDS (Intrusion Detection system) have become a necessary addition to the security infrastructure of nearly every organization.

Classification of IDS

Intrusion detection system can be broadly classified based on two parameters as: Analysis method used to identify intrusion, which is classified into Misuse IDS and Anomaly IDS. Source of data that is another method, which is classified into Host based IDS and Network based IDS.

Misuse IDS

Misuse based IDS is a very prominent system and is widely used in industries. Most of

the organizations that develop anti-virus solutions base their design methodology on Misuse IDS. The system is constructed based on the signature of all-known attacks. Rules and signatures define abnormal and unsafe behavior. It analyzes the traffic flow over a network and

System

matches against known signatures. Once a known signature is encountered the IDS triggers an alarm. With the advancement in latest technologies, the number of signatures also increases. This demands for constant upgrade and modification of new attack signatures from the vendors and paying more to vendors for their support. s

Anomaly IDS

Anomaly IDS is built by studying the behavior of the system over a period of time in order to construct activity profiles that represent normal use of the system. The anomaly IDS computes the similarity of the traffic in the system with the profiles to detect intrusions. The biggest advantage of this model is that new attacks can be identified by the system as it will be a deviation from normal behavior.

Host Based IDS

When the source of data for IDS comes from a single host (System), then it is classified as Host based IDS. They are generally used to monitor user activity and useful to track intrusions caused when an authorized user tries to access confidential information.

Network Based IDS

The source of data for these types of IDS is obtained by listening to all nodes in a network. Attacks from illegitimate user can be identified using a network based IDS. Commercial IDSs are always a combination of the two types mentioned above.

Application

Applications of intrusion detection by data mining are as follows:

- The goal of intrusion detection is to detect security violations in information systems. Intrusion detection is a passive approach to security as it monitors information systems and raises alarms when security violations are detected.
- Risk Assessment and Fraud area also uses the data-mining concept for identifying the inappropriate or unusual behavior etc.
- Customer Retention in the form of identification of patterns of defection and prediction of likely defections is possible through data mining.

Background and Related Works

R. Heady et. al, (1990) in their study they gained understanding of computer attacks in order to identify intrusion and security threats. Soo *et al.* (1997) proposed Direct Hashing and Pruning [DHP] algorithm, an effective hash based technique for

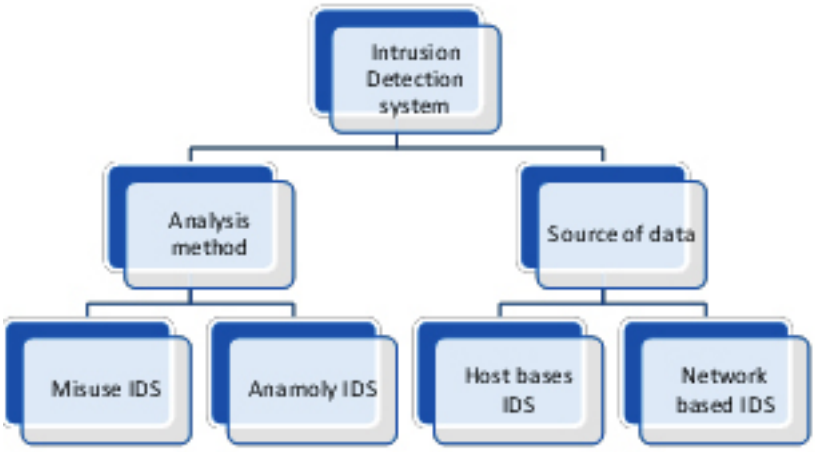

Fig. 1: Classification of Intrusion Detection System

mining the association rules. This algorithm employs effective pruning techniques to progressively reduce the transaction database Size. DHP utilizes a hashing technique to filter the ineffective candidate frequent 2 item sets. DHP also avoids database scans in some passes as to reduce the disk I/O cost involved.

Wang *et al.* (2002) presented PRICES, an efficient algorithm for mining association rules. Their approach reduces large item set generation time, which dictates most of the time in generating candidates by scanning the database only once. Another algorithm called Matrix Algorithm developed by Yuan and Huang (2005) generates a matrix which entries 1 or 0 by passing over the cruel database only once. The frequent candidate sets are then obtained from the resulting matrix. Association rules are then mined from the frequent candidate sets.

METHODS AND MATERIALS

Simulation Tools

Hardware Requirements

The minimum hardware required to execute the complete application are mentioned. According to their utility Hardware plays a vital role for the application. It provides the entire interface required and if we increase the configuration of the end system in which application is loaded the definitely the application is executed very fast and responds to user very quickly.

Hard disk : Minimum 2 GB
Ram : Minimum 2GB
Processor : Intel(R) Pentium(R) CPU B950, 2.10 GHz processor

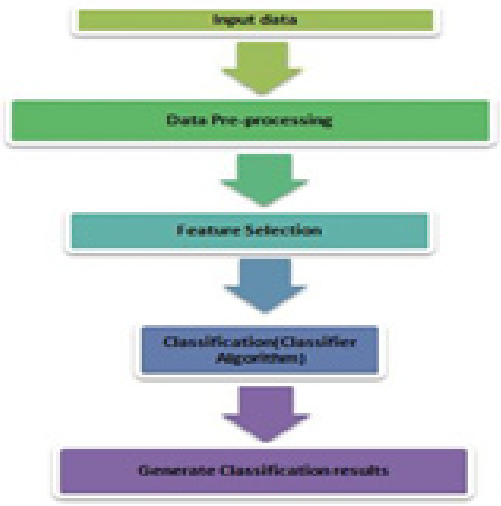

Fig. 2: Process involved in Intrusion through Data Mining

Software Requirements

The software applies all Software Engineering Concepts. Software is information transformed producing, managing, acquiring, modifying, displaying or transmitting information that can be as simple bit or a complex multimedia presentation.

JAVA

The test was carried out using the JAVA program. JAVA-Java is a set of computer software and specifications developed by Sun Microsystems, which was later acquired by the Oracle Corporation that provides a system for developing application software and deploying it in a cross-platform computing environment. Java is used in a wide variety of computing platforms from embedded devices and mobile phones to enterprise servers and supercomputers. While they are less common than standalone Java applications, Java applets run in secure, sand boxed environments to provide many features of native applications.

Netbeans IDE

NetBeans IDE lets you quickly and easily develop Java desktop, mobile, and web applications, as well as HTML5 applications with HTML, JavaScript, and CSS. The IDE also provides a great set of tools for PHP and C/C++ developers. It is free and open source and has a large community

Fig. 3: Live network interface by wireshark

of users and developers around the world. An IDE is much more than a text editor

APACHE TOMCAT SERVER

Apache Tomcat (or Jakarta Tomcat or simply Tomcat) is an open source Servlet container developed by the Apache Software Foundation (ASF). Tomcat implements the Java Servlet and the Java Server Pages (JSP) specifications. It basically makes our Java Web applications to run on host and server based system and it is configured on local host port 8080. It generally runs JSP, Servlet etc. There is a built in web container called Catalina in the tomcat bin directory. It loads all http related request and has privilege to instantiate the GET and POST method's object. It also uses cynote i.e. an http connector through network layer of the computer. All the execution is managed by JSP engine.

Wireshark (A network analyzer tool)

Wireshark is a network packet analyzer. A network packet analyzer will try to capture network packets and tries to display that packet data as detailed as possible. You could think of a network packet analyzer as a measuring device used to examine what's going on inside a network cable, just like a voltmeter is used by an electrician to examine what's going on inside an electric cable (but at a higher level, of course). In the past, such tools were either very expensive, proprietary, or both. However, with the advent of Wireshark, all that has changed.

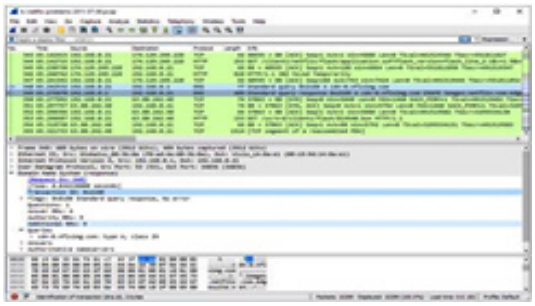

Fig. 4: Wireshark Capturing Packet Data

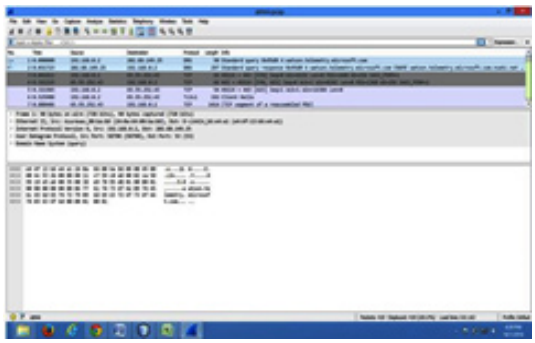

Fig. 5: Packets Captured Through Wireshark

Input Data

The first step is data input which can be done by fetching data through Wireshark tool by clicking the capture button in a live network. Wireshark is perhaps one of the best open source packet analyzers available today though it acts as a sniffer. The standard packet capture format is .pcap which is fetched by saving the tcpdump files. The way the packets are stored is illustrated by the below figures so that it can easily be understood.

Advantages of WIRESHARK

The Wireshark capture engine provides the following features:

- Capture from different kinds of network hardware such as Ethernet or 802.11.

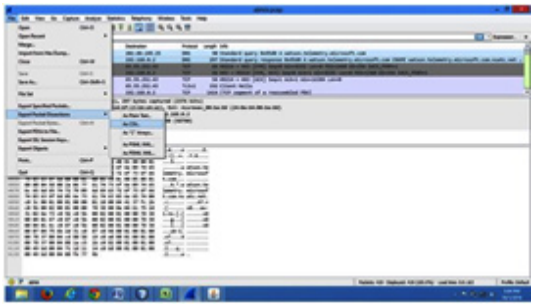

Fig. 6: Export Packet Dissections as CSV file

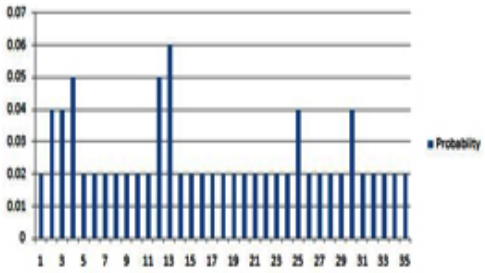

Fig. 7: Intrusion probability in the messages captured from packet

Table 1: Comparing Accuracy of different functions

Kernel Type	Accuracy
RBF Kernel	98.5749%
Polynomial Kernel	98.4281%
Naïve bayes	99.9748%

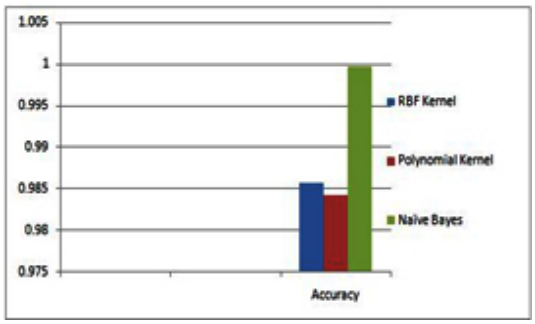

Fig. 8: Comparing Accuracy of different functions

- Stop the capture on different triggers such as the amount of captured data, elapsed time, or the number of packets.
- Simultaneously show decoded packets while Wireshark is capturing.
- Simultaneously capture from multiple network interface.

Save packets in multiple files while doing a long term capture, optionally rotating through a fixed number of files The above diagram shows the wireshark tool simultaneously capturing packets in the network. It's a command-line based capture tool. Displays filters (also called post-filters) only filter the view of what you are seeing. All packets in the capture still exist in the trace.

The captured packet then is saved to a .pcap format in order to conversion to .csv(comma separated value) for further classification. The packets which are saved in .pcap format is then converted into comma separated values so that it can be easily viewed in word document/notepad file for its feature selection.

RESULTS

The results of the study entitled "Intrusion Detection through data mining technique" was carried out during were discussed in this chapter. The findings have been illustrated with the help of tables, graphs and pictures and were perceived essential to clarify the results. Initially the attribute relation file converted to two types of classification firstly the J48 pruned tree and then naïve bayes classifier.

The graph denotes the probability of intrusion to be occurring in the class. The lower the value of class less will be intrusion probability and if it's higher, probability of intrusion will be higher. Thus the mean intrusion obtained by calculating is differing from the earlier researches done through different kernel functions.

By analysis, the Naïve Bayes obtained better results than previously RBF Kernel and Polynomial functions were used for intrusion detection. Thus machine learning results obtained from Naïve bayes and cross validation technique gives better accuracy. The results show that applied pre-processing data and relevant feature selection using information gain for reducing feature of dataset are very important to increase the classification accuracy.

CONCLUSION

The proposed method was triumphantly tested on the data log files and the database. The results of the proposed testimony are produce more accurate and irrelevant sets of patterns and the discovery time is less than other approach. As a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation to this research work. In order to alleviate this problem so as to reduce the false positives, active platform or event based classification may be thought of using Bayesian network.

REFERENCES

1. Aurobindo Sundaram, 1996 *An Introduction to Intrusion Detection,* Crossroads, Volume 2, Issue 4, Pages: 3 – 7,

2. C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers, Cambridge, MA, 2000. MIT Press.

3. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin**(**2003**),** A Practical Guide to Support Vector Classification Department of Computer Science National Taiwan University, Taipei 106, Taiwan.

4. Chunhua Gu and Xueqin Zhang (2009)" A Rough Set and SVM Based Intrusion Detection Classifier", Second International Workshop on Computer Science and Engineering. http://ilta.ebiz.uapps.net/ProductFiles/productfiles/672/wireshark.ppt

5. James P. Anderson(April 1980)"Computer Security Threat Monitoring and Surveillance" Technical report, James P. Anderson Co., Fort Washington, Pennsylvania.

6. MrutyunjayaPanda1, Manas RanjanPatra (May 2009) ''Evaluating Machine Learning Algorithms for Detecting Network Intrusions", Internation- al Journal of Recent Trends in Engineering, Vol. 1, No. 1

7. R. Heady, G. Luger, A. Maccabe, and M. Servilla (August 1990) *The Architecture of a Network Level Intrusion Detection System.* Technical report, Department of Computer Science, University of New Mexico,.

8. RafatRana S.H. Rizvi *A Review on Intrusion Detection System* Professor Computer Science and Engineering H.V.P.M's C.O.E.T Amravati, India Computer Science and Engineering H.V.P.M's C.O.E.T Amravati, India

9. *Ranjit R Keole A Review on Intrusion Detection System Professor Information Technology* India Computer Science and Engineering H.V.P.M's C.O.E.T Amravati, India

10. Sandeep Kumar**(** August 1995) Classification and Detection of Computer Intrusions. Ph.D. Dissertation,.

11. Upendra,(2013)''An Efficient Feature Reduction Comparison of Machine Learning Algorithms for Intrusion Detection System", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) , Volume 2, Issue 1, January – February.

12. Wenke Lee and Salvatore J. Stolfo, (Nov 2000) "A Framework for constructing features and models for intrusion detection systems",ACM transactions on Information and system security (TISSEC), vol.3, Issue 4.

13. Y. Hu, B. Panda, "A Data Mining Approach for Database Intrusion Detection", Proceedings of the ACM Symposium on Applied Computing, pp. 711-716 (2004)

14. Yogita B. Bhavsar, Kalyani C.Waghmare March 2013 Intrusion Detection System Using Data Mining Technique www.ijetae.com Support Vector Machine

15. Yogita B.Bhavsar. Intrusion Detection System Using Data Mining Technique: Support Vector