ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Intrusion Detection System using Data Mining

Abhishek Sawant[1], Jyoti Yadav[2], Avneet Kaur Arora[3], Janhavi Deo[4], Nutan Dhange[5]

Student, Information Technology, Atharva College of Engineering, Mumbai, India[1,2,3,4]

Professor, Information Technology, Atharva College of Engineering, Mumbai, India[5]

Abstract: With the tremendous growth of the usage of computers over network and development in application running on various platform captures the attention toward network security[1]. Intrusion detection system has become an important component of a network infrastructure protection mechanism. The Intrusion Detection System (IDS) plays a vital role in detecting anomalies and attacks in the network [5]. In this work, data mining concept is integrated with an IDS to identify the relevant, hidden data of interest for the user effectively and with less execution time. In proposed system, we first preprocess dataset (KDD 99 cup). Then we study different types of decision tree algorithms (C4.5 and its extension) of data mining for the task of detecting intrusions and compare their relative performances. Based on this study, it can be concluded that even extended C4.5 is complex but decision tree obtained is the most suitable with high true positive (correct detection of attacks) and low false positive (Incorrect detection) with high accuracy.

Keywords: Intrusion detection system, KDD 99 cup, Data Mining, Decision Tree Algorithms, C4.5 and its extensions

I. INTRODUCTION

Nowadays, many organizations and companies use Internet services as their communication and marketplace to do business such as at EBay and Amazon.com website. Together with the growth of computer network activities, the growing rate of network attacks has been advancing, impacting to the availability, confidentiality, and integrity of critical information data. Therefore a network system must use one or more security tools such as firewall, antivirus, IDS and Honey Pot to prevent important data from criminal enterprises. A network system using a firewall only is not enough to prevent networks from all attack types. The firewall cannot defense the network against intrusion attempts during the opening port. Hence a Real-Time Intrusion Detection System (RT-IDS), shown in Fig 1, is a prevention tool that gives an alarm signal to the computer user or network administrator for antagonistic activity on the opening session, by inspecting hazardous network activities[6].

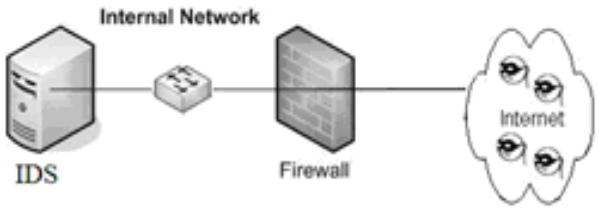

Fig.1 Intrusion detection system environment

A. AIM

An Intrusion Detection System is an important security feature to detect the threats to a vulnerable network. In the proposed system we use efficient decision tree techniques to filter the data set and train the IDS to treat new data approaching the system. The system will use data mining algorithm like C4.5 and its extension.

Also the different techniques involved to detect the intrusion are compared with the help of parameters like accuracy, time taken to identify the attack, number of false positives, number of false negatives etc in graphical format.

B. OBJECTIVE

With hacker attacks against well-known businesses and organizations on the rise, network security has made headlines. Of course, there are many attacks that do not make headlines and are not reported due to a loss of credibility or embarrassment. Then there are the attacks that are not even detected. The Defense Information Services Agency (DISA) states that up to 98% of attacks go unnoticed. These revelations have caused many businesses to rethink or to start thinking about the security of their own networks. Maximum Processing computation and more time consuming task has always been a limit in processing huge network intrusion data.

C. SCOPE

IDS using data mining helps in protecting a network from malicious attacks from outsiders as well as insiders. It also helps to identify the type of attack. It will help in Intrusion Detection in various platforms like college network, corporate network etc.

II. RELATED WORK

A. INTRUSION DETECTION SYSTEM

An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network. There are network based (NIDS), host based (HIDS) intrusion detection systems, signature based and Anomaly based. There are IDS that simply monitor and alert and there are IDS that perform an action or actions in response to a detected threat. We'll cover each of these briefly.

a. NIDS:

Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. This type is susceptible a bottleneck that would impair the overall speed of the network.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

b. HIDS:

Host Intrusion Detection Systems are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator of suspicious activity is detected.

c. Signature Based:

A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

d. Anomaly Based:

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline identifies what is normal for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator.

B. DATA MINING

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

a. The Scope of Data Mining:

Data mining derives its name from the similarities between searching for valuable business information in a large and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automated prediction of trends and behaviors: Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data. It also provides various models that help in forecasting.

Automated discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions.

C. DATA SET

Software to detect network intrusions protects a computer network from unauthorized users, including perhaps insiders. The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided. The 1999 KDD intrusion detection contest uses a version of this dataset. 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 various attributes in dataset are:

Duration	Protocol Type	Service	Flag
Scr bytes	Des bytes	Land	Wrong fragment
Urgent	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 host login	In guest login	Count	Srv count
Serror rate	Svr serror rate	Rerror rate	Srv rerror rate
Same srv rate	Diff srv rate	Srv diff host rate	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 rerror rate
Dst host srv rerror rate	Normal or attack		

Table.1 Attributes of KDD Data Set

D. C4.5 ALGORITHM

C4.5 is an extension of ID3 which generates a decision tree. The decision tree generated by C4.5 can be used classification.

a. Method:

(1) Create a node N;
(2) If tuples in D are all of the same class, C then
(3) Return N as a leaf nod labeled with the class C;
(4) If attribute_list is empty then
(5) Return N as a leaf node labeled with the majority class in D; //majority voting

DOI 10.17148/IJARCCE.2015.4253

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

(6) Apply attribute_seletion_method (D, arrtibute_list) to find the "best" splitting_criterion;
(7) Label node N with splitting_criterion;
(8) If splitting_attribute is discrete-valued and Multiway splits allowed then // not restricted to binary trees
(9) attribute_list→attribute_list - splitting_attribute; //remove splitting_attribute
(10) for each outcome j of splitting_criterion // partition the tuples and grow sub-tees for each partition
(11) Let Dj be the set of a data tuples in D satisfying outcome j; // a partition
(12) If Dj is empty then
(13) Attach a leaf labeled with the majority class in D to node N;
(15) Else attach the node returned by Geneate_decision_tree (Dj, attribute list) to node N;
(16) Return N;

E. IMPROVEMENTS FROM ID3 ALGORITHM

C4.5 made a number of improvements to ID3. Some of these are: Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations. Handling attributes with differing costs. Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree.C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors: A possibility to use continuous data. Using unknown (missing) values which have been marked by "?".Possibility to use attributes with different weights. Pruning the tree after being created

.

III. PROPOSED ARCHITECTURE

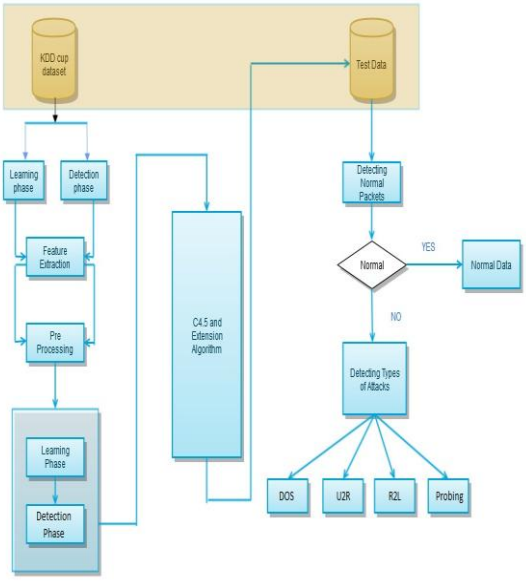

Fig.2 Proposed System

KDD (Knowledge Discovery Dataset) dataset is data collected from US Air Force LAN. It is labeled data which consists of 38 different attacks classified in 4 categories. Various attributes of dataset are studied and required features are extracted. The data is pre-processed i.e., null values are replaced with some default values, empty spaces are filled with appropriate values. The algorithm used is C4.5 and its extension, the algorithm is trained with the 10% labeled dataset. Labeled dataset is used for training the algorithm and is called training phase. The algorithm is then tested with unlabeled data which contains various types of attacks. . The algorithm creates a decision tree which helps to identify type of attack. If a particular tuple does not contain any attack then it is treated as normal data. If that particular tuple contains attacks then it specifies which type of attack it is. In an active system the network is monitored, it monitors the packets. If the packets are normal then it is allowed to pass if it detects attack it identifies the type of attack and alerts the user.

IV. FUTURE WORK

The existing C4.5 algorithm will be enhanced so that it performs better and gives more accurate results. Extended C4.5 will have better and improved method to calculate information or entropy. The extended C4.5 will be compared with some other algorithm and the results will be compared to check how efficiently C4.5 is working. The KDD dataset will be experimented with WEKA tool to get graphical output which helps to understand working of decision tree in a better way.

V. CONCLUSION

There are many successful applications of data mining in the field of IDS, but there are several possibilities for improvement in existing technology. In this paper we have proposed an intrusion detection system which detects 4 different types of attacks. The algorithm used is C4.5 which is a decision tree algorithm that helps to identify various types of attack. Initially the algorithm is trained with labeled KDD dataset which is then experimented with unlabeled dataset. Weka tool is used for giving graphical output which helps to understand the output in a better way.

ACKNOWLEDGEMENT

It stands to reason that the completion of main project needs the support of many people. We take this opportunity to express our boundless thanks and commitment to each and everyone, who helped us in successful completion of our main project. We are happy to acknowledge the help of all the individuals to fulfill our attempt.
First and foremost we wish to express wholehearted in debtedness to God Almighty for his gracious constant care and magnanimity showered blissfully over us during this endeavor.
We are thankful to Prof. Neelima Pathak, Head of Department, Information Technology, Atharva College of Engineering, for providing and availing us of all the required facilities to prepare this paper. We express our

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

heartfelt gratitude to Prof. Sumita Chandak, Lecturer in Information Technology for working as our project co-coordinator, who corrected us and gave valuable suggestions, Prof. Nutan Dhange, for working as project supervisor, who guided us and helped us overcome our mistakes and difficulties.

Gratitude is extended to all teaching and non teaching staffs of Department of Information Technology, Atharva College of Engineering for their cooperation to complete this paper.

We are also thankful to our parents who constantly supported us. Gratitude may be extended to all well-wishers and our friends who supported us to complete this paper in time.

REFERENCES

[1] Effective approach toward Intrusion Detection System using data mining techniques by G.V. Nadiammai, M. Hemalatha in Egyptian Informatics Journal (2014) 15, 37-50

[2] Data Mining in Education for Students Academic Performance: A Systematic Review by Er.Anurag Jindal, Er. Williamjeet Singh in ISSN 2277-3061

[3] Mining With Noise Knowledge: Error-Aware Data Mining by Xindong Wu and Xingquan Zhu in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS,VOL. 38, NO. 4, JULY 2008

[4] Combined Mining: Discovering Informative Knowledge in Complex Data by Longbing Cao, Senior Member, IEEE, Huaifeng Zhang, Member, IEEE, Yanchang Zhao, Member, IEEE, Dan Luo, and Chengqi Zhang, Senior Member, IEEE in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL.41, NO. 3, JUNE 2011

[5] Intrusion Detection System Using Data Mining Technique: Support Vector Machine by Yogita B. Bhavsar, Kalyani C.Waghmare in International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3,March 2013)

[6] Network Intrusion Detection Using Improved Decision Tree Algorithm by K.V.R. Swamy, K.S. Vijaya Lakshmi in (IJCSIT) International Journal of Computer Science and Information Technologies,Vol. 3 (5) , 2012,4971 – 4975

[7] Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm by Swati Paliwal, Assistant Professor, Dept. of C.S.E., Sharda University, Gr.Noida, Ravindra Gupta, Assistant Professor Dept. of C.S.E., SSSIST, Sehore, India in International Journal of Computer Applications (0975 – 8887) Volume 60– No.19, December 2012

[8] Ham Kember. Data mining concepts and techniques. Studying various algorithms.