Decision Tree Applied For Detecting Intrusion

Poonam Gupta
M.Tech. C.S.E. Scholar
CVRIST,Bilaspur,India

S. R. Tandan
Asst. Professor,Dept.of C.S.E
CVRIST,Bilaspur,India

Rohit Miri
Asst. Professor,Dept.of C.S.E
CVRIST,Bilaspur,India

Abstract

Information is most important asset for organization and they require proper management and protection. Nowadays computer attack has become very common. Although there are many existing mechanisms for Intrusion detection, but the major issues is the security and accuracy of the system. In this paper we investigate and evaluate the decision tree data mining techniques as an intrusion detection mechanism. Our research shows that Decision trees gives better overall performance.

Keywords: Inrusion detection,Decision tree.

1. INTRODUCTION

Due to increased number of internet users there is a problem due to intrusion which may damage data and information stored in computer server or data base server. So we need a filter which is able to filter malicious data and normal data.

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. The intrusion detection and other security technologies such as cryptography, authentication and firewalls has gained in importance in last few years[1].

There are two types of intrusion detection techniques: Misuse and Anomaly. Misuse detectors analyze system activity, looking for events or sets of events that match a predefined pattern of events that describe a known attack. As the patterns corresponding to known attacks are called signatures, misuse detection is sometimes called "signature-based detection." Anomaly detectors identify abnormal unusual behaviour (anomalies) on a host or network. They function on the assumption that attacks are different from "normal" (legitimate) activity and can therefore be detected by systems that identify these differences[2].

In this paper we have suggested data mining approach to intrusion detection. This paper mainly focuses on the signature based intrusion detection systems and presents a way to identify patterns of harmful attacks by training the system on a database and testing the same. In order to support the training and testing the NSL-KDD dataset is used, which consists of different types of network connections labeled with the category. A model with high accuracy will be tried to develop .Model will be trained and tested on the normal and known attacks.

2. RELATED WORK

Currently building an effective IDS is an enormous knowledge engineering task. System builders relay on their intuition and experience to select the statistical measures for anomaly detection. Experts first analyze and categorize attack scenarios and system vulnerabilities, and hand-code the corresponding rules and patterns for misuse detection. Because of the manual and adhoc nature of the development process, current IDSs have limited extensibility and adaptability. Many IDSs only handle one particular audit data source, and their updates are expensive and slow[3][4].

Heba Ezzat Ibrahim et al.[5] proposed a multi-Layer intrusion detection. There experimental results showed that the proposed multi-layer model using C5 decision tree achieves higher classification rate accuracy, using feature selection by Gain Ratio, and less false alarm rate than MLP and naïve Bayes. Using Gain Ratio enhances the accuracy of U2R and R2L for the three machine learning techniques (C5, MLP and Naïve Bayes) significantly. MLP has high classification rate when using the whole 41 features in Dos and Probe layers. Limitation is this paper propagates errors as to simulate the real system and results be more accurate and real.

Sandhya Peddabachigari et.al[11] In this paper they investigate and evaluate the decision tree data mining techniques as an intrusion detection mechanism and we compare it with Support Vector Machines (SVM). Intrusion detection with Decision trees and SVM were tested with benchmark 1998 DARPA Intrusion Detection dataset. Their research shows that Decision trees gives better overall performance than the SVM.

K.Nageswara rao et al.[6] evaluated the influence of attribute pre-selection using Statistical techniques on real-world kddcup99 data set. Experimental result shows that accuracy of the C4.5 classifier could be improved with the robust pre-selection approach when compare to traditional feature selection techniques But the only limitation in this research paper is implementing correct attribute selection measure in C4.5 decision tree algorithm.

Ala' Yaseen et al.[7] This paper concludes many clustering techniques that were previously proposed to solve the inherent IDS problems. Where, the clustering techniques involved in three general aspects namely: data preprocessing, anomaly detection, and data projection/alarm filtering. Eventually, recommendations for future researches followed by the conclusion are depicted at the end of this paper.

Reema Patel et al[10] a discussion of the future technologies and methodologies which promise to enhance the ability of computer systems to detect intrusion is provided and current research challenges are pointed out in the field of intrusion detection system.

M. Sathya Narayana et al[12] in this paper they proposes system uses a classification-based approach to summarize the characteristic features of a node together with a path to sentence generator todescribe these features in natural language.

Mahmood[13] The goal of this paper is to provide a survey of some works that employ data mining techniques for intrusion detection and to address some technical issues. They proposed a new a idea in this paper that will view intrusion detection from a data warehouse perspective and integrate data mining and on-line analytical processing (OLAP) for intrusion detection purposes. One of the major limitations of the systems is that they lack adaptability to changing behavior patterns. Some technical issues were discussed which are critical in developing a true adaptive, real-time intrusion detection system.

.

3. SYSTEM IMPLEMENTATION

Proposed research work introduces a framework to develop a classifier based on data mining techniques . In this framework NSL-KDD[8] dataset is given to Preprocessing stage which classify in C4.5 algorithm and reduce irreverent features from the data set so that data with less number of feature will require to feed to the classifier and will provide efficiency to the classifier. Machine learning tools WEKA are used to analyze the performance of datasets. This approach involve several steps-

Step 1. Preprocess the datasets.

- Load data
- Analyze attributes.

Step 2. Classify the datasets.

- Select Test Options e.g:
  – Use Training Set
  – Percent Split,
  – Cross Validation
- Run classifiers
- View results

3.1 System Architecture

There are many existing mechanisms for Intrusion detection system, but the major issues is the security and accuracy of the system . To improve the problem of accuracy and the efficiency of the system,a very common classification approach i.e. decision tree is used. Proposed research work introduces a framework to develop a classifier based on data mining techniques as shown in fig.1:

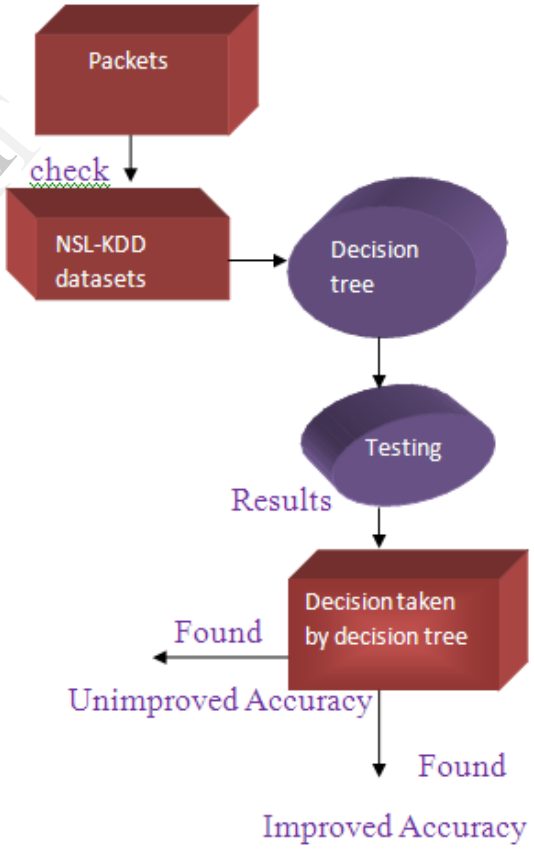


Figure1: System Architecture

4. EXPERIMENTAL METHODOLOGY

The experimental methodology followed in this research includes data sets and classification technique i.e. C4.5 algorithm . The description of these methodologies are described below.

4.1 Data Description

NSL-KDD is a dataset suggested to solve some of inherent problems of KDD99 datasets which are mentioned. Although this new version of KDD data set still suffers from some of problems and may not be a representatives of existing real network because of lack of public dataset for network based intrusion detection system[8].

The training dataset consists of 25,192 records and contains 42 attributes and its class is labeled as either normaly or anomaly, in which anomaly is with exactly one specific attack type. The attacks types are grouped into four categories-

(1) DOS: Denial of service – e.g. syn flooding

(2) Probing: Surveillance and other probing, e.g. port scanning.

(3) U2R: unauthorized access to local super user (root) privileges, e.g. buffer overflow attacks.

(4) R2L: unauthorized access from a remote machine, e.g. password guessing[9].

4.2 C4.5 Algorithm

Just like Classification and Regression Tree, the C4.5 algorithms recursively visits each node, selecting the optimal split, until no further splits are possible. The steps of C4.5 algorithm for growing a decision tree is given below

1. Choose attribute for root node by using attribute selection measure Gain Ratio
2. Create branch for each value of that attribute.
3. Split cases according to branches.
4. Repeat process for each branch until all cases in the branch have the same class or all attributes are processed[6].

5. RESULTS

For training the system a part of the 20% of NSL-KDD dataset is considered which consists of 25, 192 records of the network connection out of which 13449 records are of normal non-malicious category, 01 connections of land, 8282 connections of neptune, 181 connections of warezclient, 710 connections of ipsweep,188 connections of teardrop,587 connections of portsweep,38 connections of pod,10 connections of guess_passwd,301 connections of nmap,691 connections of satan,529 connections of smurf,2 connections of multihop,196 connections of back,1 connections of ftp_write ,5 connections of imap,2 connections of phf,4 connections of rootkit,7connections of warezmaster.

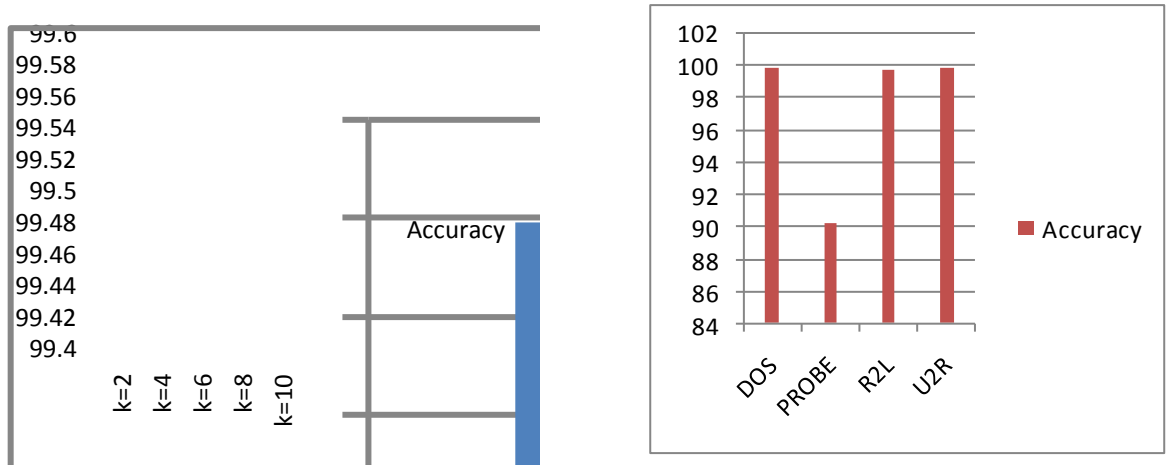
Once the system has been trained, it can be tested for it's performance. The data sets include whole training set itself, cross validation is applied on the training set, splitting the training dataset and providing a completely different test dataset. Based on the records of the different datasets results are obtained separately for the system as shown in the Table

Table 1:Testing the system by cross validation datasets-However,in this experiment k=8 have highest accuracy but in ROC k=10 have highest accuracy as shown in table 1

Datasets used for testing	Correctly classified instances	Incorrectly classified instances	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
K=2	99.468%	0.531%	0.995	0.005	0.995	0.995	0.995	0.995
K=4	99.579%	0.420%	0.996	0.004	0.996	0.996	0.996	0.997
K=6	99.555%	0.444%	0.996	0.004	0.996	0.996	0.996	0.997
K=8	99.579%	0.420%	0.996	0.004	0.996	0.996	0.996	0.997
K=10	99.559%	0.440%	0.996	0.004	0.996	0.996	0.996	0.998

Table 2:Testing the system for all the attack system

Attack Types	Correctly classified instances	Incorrectly classified instances
DOS	99.898%	0.101%
PROBE	90.207%	9.792%
R2L	99.714%	0.285%
U2R	99.918%	0.0817%



(a) (b)

Figure 2:(a)A bar chart for cross validation and accuracy (b) attacks and accuracy

Table 3: Testing the system by splitting datasets on different percentage

Percentage split on training datasets	Correctly classified instances	Incorrectly classified instances	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
50%	99.444%	0.555%	0.994	0.006	0.994	0.994	0.997	0.997
60%	99.444%	0.555%	0.994	0.006	0.994	0.994	0.994	0.996
70%	99.484%	0.516%	0.995	0.005	0.995	0.995	0.995	0.998
80%	99.662%	0.337%	0.997	0.004	0.997	0.997	0.997	0.999

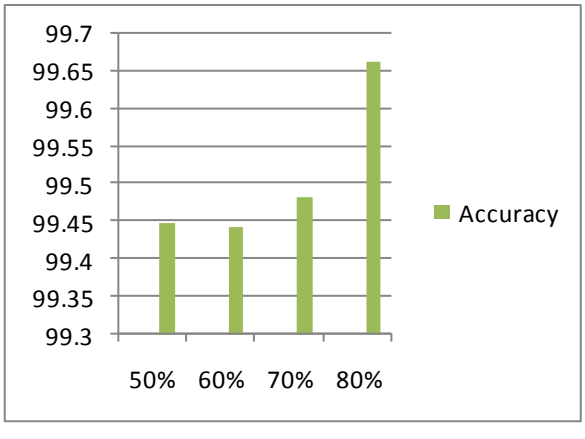


Figure 3:A bar chart for percent splitting of training datasets and accuracy

Figure 4: shows the decision tree that is constructed after the system is trained.The number of leaves used to build the tree is 329, and the size of the tree is 383.

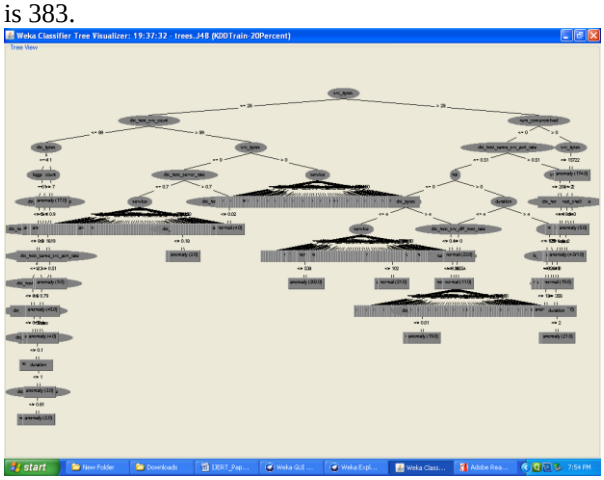


Figure 4: Visualization of decision tree

6. CONCLUSION AND FUTURE ENHANCEMENT

In this research we have implemented techniques for intrusion detection which gives better performance. In this research we have investigated in signature based intrusion detection which detect only known attacks. However this is one of the major drawback of this system that it can't detect unknown and attacks.

The future enhancement of this system is, it removes its drawback by implementing a system that detect both unknown and known attack.

7. ACKNOWLEDGEMENT

The author would like to thanks Mr. S.R.Tandan and Mr.Rohit Miri of the Department of Computer Science at Dr. C.V. Raman University for valuable suggestions. Appreciation also goes to the Department of Computer Science and the Dr. C.V. Raman University of Science and Technology for the support.

8. REFERENCES

[1]E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.Govinda Rajulu,"A Study of Intrusion Detection in Data Mining", Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011, July 6 - 8, 2011, London, U.K.

[2]Rebecca Bace and Peter Mell,"*Intrusion Detection Systems"*, NIST Special Publication on Intrusion Detection Systems.

[3]Anusha Jayasimhan,Jayant Gadge," Identifying Intrusion Patterns using a Decision Tree", *International Journal of Computer Applications* (0975 – 8887) Volume 45– No.9, May 2012.

[4] Lee,Salvatore J. Stolfo," A framework for constructing features and models for intrusion detection systems," *ACM Transactions on Information and System Security*, Vol. 3, No. 4, November 2000, Pages 227–261.

[5]Heba Ezzat Ibrahim,Sherif M. Badr, Mohamed A. Shaheen," Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", *International Journal of Computer Applications (0975 – 8887)*,Volume 56– No.7, October 2012.

[6] K.Nageswara rao, D.RajyaLakshmi, T.Venkateswara Rao," Robust Statistical Outlier based Feature Selection Technique for Network Intrusion Detection", *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307*, Volume-2, Issue-1, March 2012.

[7] Ala' Yaseen Ibrahim Shakhatreh , Kamalrulnizam Abu Bakar ,"A Review of Clustering Techniques Based on Machine learning Approach in Intrusion Detection Systems", *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 5, No 3, September 2011.

[8] "NSL-KDD data set for network-based intrusion detection systems ", Available on: http://nsl.cs.unb.ca/NSL-KDD.

[9]AdetunmbiA.Olusola,Adeola S.Oladele.,Daramola O.Abosede,"Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010, San Francisco, USA.

[10] Reema Patel, Amit Thakkar, Amit Ganatra,"A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", *International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307*, Volume-2, Issue-1, March 2012.

[11] Sandhya Peddabachigari, Ajith Abraham*, Johnson Thomas,"Intrusion Detection Systems Using Decision Trees and Support Vector Machines".

[12] M. Sathya Narayana, B. V. V. S. Prasad, A. Srividhya and K. Pandu Ranga Reddy", Data Mining Machine Learning Techniques – A Study on Abnormal Anomaly Detection System",*International Journal of Computer Science and Telecommunications* [Volume 2, Issue 6, September 2011].

[13] Mahmood Hossain," Data Mining Approaches For Intrusion Detection: Issues And Research Directions",Department of Computer Science, Mississippi State University, MS 39762, USA