European Journal of Computer Science and Information Technology

Vol.5, No.6, pp.11-18, December 2017

Published by European Centre for Research Training and Development UK (www.eajournals.org)

INTRUSION DETECTION WITH TREE-BASED DATA MINING CLASSIFICATION TECHNIQUES BY USING KDD DATASET

Bilal Ahmad

Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing China

ABSTRACT: *In the recent time a huge number of public and commercial services are used through via Internet, so that security of systems becomes most important issue in the society and threats from hackers also increased. So many researcher feels intrusion detection systems can be fundamental line of defense. Intrusion Detection System (IDS) used against attacks for protected to the Computer networks. On another hand, data mining techniques can also contribute to intrusion detection. Intrusion detection can be classified into two classes: Anomaly based and Misuse based. One of the biggest problem with the anomaly base intrusion detection is detecting the number of high false alarm ratio. In this paper solution will be provided to increase attack recognition rate with the minimum false alarm with the study of different tree-based data mining techniques. KDD cup dataset used for research purpose with WEKA tool.*

KEYWORDS: Data Mining; Intrusion Detection System; Decision Tree j48; Hoeffding Tree; Rep Tree; Random Forest; Random Tree; KDD dataset

INTRODUCTION

The security of the internet is becoming more and more serious in recent years. We have been suffered from many kind of attacks are appearing. Therefor it's very necessary to purpose an effective and accurate detection model to protection. Intrusion detection (ID) is a type of security management system for computers and networks. Intrusion detection system used to detect computer attacks by examining different logs or data records. The role of a network IDS is passive, only gathering, identifying, logging and alerting IDS system use to attempts to identify intrusions which are misuses or abuses of computer system r network by malicious user. Some IDSs monitors a single computer while other monitor several computers connected by a network. There are two types of attack network base and host base. In host base attack attempts to access restricted service or resource from single computer. While network base attack restrict legitimate user from access various network service by thought occupying network resource and services. This can be done by sending large no of amount of network traffic. In network base attack detection network traffic can be analyzed from intrusion detection basically two type of anomaly detection system. First one is based on specification or set of rules. The second one base upon learning or training the normal behavior of system. Snort like IDS usually use for rule base intrusion detection in which rule are written manually for identification known attacks. Other type is behavior base IDS. Advantages of behavior-based approaches are that they can detect attempts to exploit new and unforeseen vulnerabilities. One of the major problem anomaly based IDS is detection of high false alarm to solve this issue by applying different data mining tree based algorithms and find which algorithm gives us most best result as compared to other algorithms

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

LITERATURE SURVEY

The idea of IDS was first came from the technical report from **Anderson (1980) [1].** He drive the computer audit mechanism which should me transformed and able to provide risk and threats for computer safety techniques. This idea should provide statistical methods which can apply on user behavior and detect intruders who can access the system illegally. In **1987 Dorothy** suggested a prototype for intrusion detection **Dorothy E. Denning and Peter Neumann (1987**) were early pioneers in the Intrusion Detection arena. They had provided the framework for an intrusion-detection expert system, which was called IDES (Intrusion Detection Expert System) **[2]** based off of the **1985** paper Requirements and model for IDES – A real-time intrusion detection system **[3]**. **Hoge and Austin (2004)** provides detail survey of anomaly detection using machine learning and statistical methods **[4]**. They introduce a survey of contemporary techniques for outlier detection. **Markou and Singh [5]** also presented extensive reviews for intrusion detection using ANN and statistical methods. Patcha and park **[6]** also presented survey of various anomaly techniques cyber intrusion detection. Many books and article also written based on Outliers and intrusion detection (**Douglas M. Hawkins 1980, V. Barnett, and T. Lewis 1994, Z. Bakar, R. Mohemad, A. Ahmad, M. Deris [7,8,9]** Many anomaly detection system such as NIDES (Next generation Intrusion detection system expert system) **[10]** ALAD (application layer anomaly detector)[11], PHAD (packet header anomaly detector) [12] generate statistical model for normal network traffic and alarm generates if some deviation found in normal model. Most of then use feature extraction from network packet header. For example NIDES and ALAD use source, destination IP, port address and TCP connection state.

Zhang, Yang, and Geng (2009) [13] presented survey of network anomaly detection methods and techniques. **Peng, Leckie and Ramamohanarao (2007) [14]** make exhaustive survey of techniques for detecting DOS and distributed DoS attacks. **Wu and Banzhaf (2010) [15]**The area of this review will include main methods of CI, including artificial neural networks, fuzzy systems, evolutionary computation, artificial immune systems, swarm intelligence, and soft computing. **Dong, Hsu, Rajput (2010) [16]** presented the method which is according to them is more authentic then Markov and K. means. Graph-based Sequence-Learning Algorithm (GSLA) includes data pre-processing, normal profile construction and session marking. In GSLA, the normal profile is built through a session-learning method, which is used to determine an anomaly session. **Warusia, Udzir (2014) [17]** purpose novel Signature-Based Anomaly Detection Scheme (SADS) which could be applied to study packet headers behavior patterns more precisely and promptly is proposed. Integrating data mining classifiers such as Naive Bayes and Random Forest can be utilized to decrease false alarms and reducing processing time. Some researchers also use feature selection techniques for intrusion detection. **Liu, Motoda and Setiono (2010**) **[18]** describe Feature selection is an effective technique for dimension reduction and an essential step in successful data mining applications. Its direct benefits include: building simpler and more comprehensible models, improving data mining performance, and helping prepare, clean, and understand data. **Harbola, Jyoti (2014) [19]** also use feature selection techniques to improve accuracy. The main objective of this analysis is to deliver the broad analysis feature selection methods for NSL-KDD intrusion detection dataset.

Intrusion Detection System:

Intrusions can be said as the illegal attempt for getting access on a system or network. Intrusion detection is the system to detect this kind of suspicious activity on the network or a device. The IDS is consider as hardware or software or combination of both that can monitoring of the network flow for the search of intrusions. An intrusion detection system (IDS) reviews all out going in going network activity and identifies doubtful patterns.

Type of IDS:

Intrusion detection system can be classified into Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS).

Host-based Intrusion Detection System:

Host based intrusion detection (HIDS) refers to intrusion detection that takes place on a single host system. It is a software application which is installed onto a system in order to protect it from intruders. HIDS are operating system dependent so require some prior planning before implementation and are efficient in detecting buffer overflow attacks.

Network-based Intrusion Detection System:

A network-based intrusion detection system (NIDS) is used to monitor network traffic to protect a system from network-based threats. A NIDS reads all inbound packets and searches for any suspicious patterns. It is operating system independent and it provides better security against denial of service (DOS) attacks

Type of Attacks Detected by IDS:

Following are the four types of attacks on ground being detected by IDS:

Denial of Service Attack/attempt to make a network resource unavailable to its intended users such as suspend services of a host connected to the Internet.

User to Root (U2R) Attack where an attacker attempts to get unauthorized access of target system

Remote to User Attack (R2L) where attacker try to control of remote machine by guessing password

Probing Attack (Probe) where attacker scene/examine the machine to get useful information

KddCup'99 Dataset:

The KDD CUP 1999 [20] standard datasets are published for research purpose. It is used in order to assess different feature selection method for Intrusion detection system. The data set consists of 41 features and a separate feature (42^{nd} feature) that labels the connection as 'normal' or a type of attack. The data set contains a total of 24 attack types that fall into 4 major categories (DoS, Probe, R2L and U2R) that are already discussed.

For the training and testing of the proposed framework the 10% of the KddCup'99 dataset is used as the full KddCup'99 dataset consists of 5 million instances many of them are redundant.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

The 10% of the KddCup'99 dataset consists of 494021 instances. In which 97278 are 'Normal' instances and remaining 396743 are belongs to any one type of attack.

Preprocessing:

In the preprocessing module the class label presents in the 42 feature of KddCup'99 dataset is recast into five major categories for the sake of decreasing complexity of performance evaluation of the proposed model. In the result of preprocessing major classes are formed as the class label i.e. DoS, Probe, R2L, U2R and Normal.

Split into Training and testing:

In this phase the given data are randomly partitioned into two independent sets, a training set and a test set. The 66% of the data is allocated to the training set and the remaining 44% of the dataset is allocated to the testing set. The training set is used to derive the proposed framework while the test set is used to assess the accuracy of the derived model. After divided into two sets Training set have 326054 instances and testing set have 167967 instances.

Different types of attacks in experimental dataset which are classified into four categories are shown below

Attack types with their corresponding categories

Type	Attacks
DOS	apache, back, land, mailbomb, neptune, pod, processtable, smurf, teardrop, udpstorm
PROBE	ipsweep, mscan, nmap, portsweep, saint, satan
U2R	buffer_overflow, loadmodule, perl, rootkit, ps, sqlattack, xterm
R2L	ftp_write, guess_password, imap, multihop

KDD '99 Intrusion Detection Datasets in terms of number of samples

Type	Train	Test
DOS	391458	229853
PROBE	4107	4166
U2R	1126	16347
R2L	52	70
NORMAL	97278	60591

RESULTS AND EXPERIMENT:

We performed the experiment with KDD cup dataset using 10% **[21]** train and test dataset (using WEKA)

A. Experiment Setup

- Experiment performed under following hardware and software
- Hardware: Intel core i5 1.8 Ghz processor with 4 GB Ram.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

- Software: Microsoft windows 10, WEKA 3.7

Section ONE

B. Using Training Dataset

Classifiers	**Classified Instances**	
	Correctly	Incorrectly
Hoeffding Tree	99.472	0.527
J48	99.963	0.036
Random Forest	**99.983**	**0.017**
Random Tree	99.963	0.036
RepTree	99.950	0.496

Classifiers	**DOS**		**PROBE**	
	Correct	False +V	Correct	False +V
Hoeffding Tree	390637	821	2987	1120
J48	391435	23	4076	31
Random Forest	**391455**	**3**	**4079**	**26**
Random Tree	391442	16	4071	36
Rep Tree	391420	38	4012	95

Classifiers	**R2L**		**U2R**	
	Correct	False +V	Correct	False +V
Hoeffding Tree	711	415	13	39
J48	1076	50	25	27
Random Forest	**1105**	**21**	**36**	**16**
Random Tree	1091	35	36	16
Rep Tree	1099	27	25	48

Classifiers	**Normal**	
	Correct	False +V
Hoeffding Tree	97069	209
J48	97229	39
Random Forest	**97262**	**16**
Random Tree	97202	76
Rep Tree	97220	58

C. Using Test Dataset

Classifiers	**Classified Instances**	
	Correctly	Incorrectly
Hoeffding Tree	97.0501	2.9499
J48	98.0416	1.9584
Random Forest	**98.0818**	**1.9182**

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

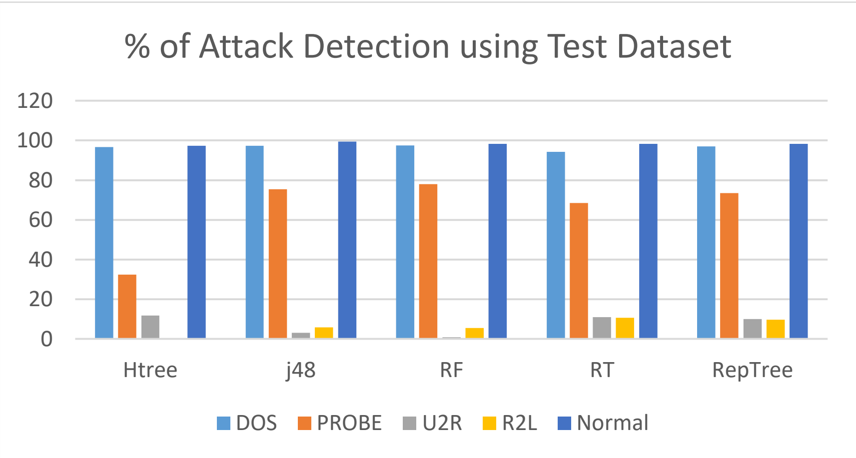
Random Tree	98.0371	1.9629
RepTree	98.0262	1.9738

Classifiers	**DOS**		**PROBE**	
	Correct	False +V	Correct	False +V
Hoeffding Tree	229407	446	3792	374
J48	229825	28	4098	68
Random Forest	**229835**	**18**	**4122**	**44**
Random Tree	229823	30	4099	67
Rep Tree	229817	36	4071	95

Classifiers	**R2L**		**U2R**	
	Correct	False +V	Correct	False +V
Hoeffding Tree	12923	3424	52	18
J48	13518	2829	32	38
Random Forest	**13553**	**2794**	52	18
Random Tree	13540	2807	49	21
Rep Tree	13458	2889	50	20

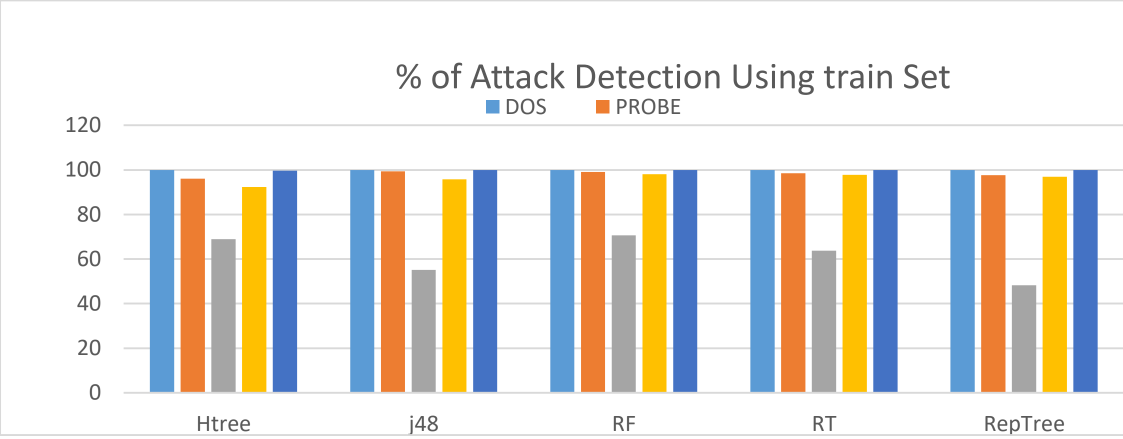
Classifiers	**Normal**	
	Correct	False +V
Hoeffding Tree	55678	4913
J48	57463	3128
Random Forest	**57499**	**3092**
Random Tree	57411	3180
Rep Tree	57492	3099

RESULT AND ANALYSIS



The above table shows the result of test dataset it is clear that j48 classifier perform well in U2R R2L and normal categories and DOS PROBE it is slightly behind Random Forest

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)



In the above table we can see all classifier achieve more the 90% attack detection in DOS, PORBE, R2L and more the 99% in normal category only U2R attack ratio less the 75 % this is due to U2R type attack are very less in training dataset. Compare to other classifier Random Forest perform slightly Batter in DOS, U2R, R2L but j48 perform batter in PROBE

CONCLUSION AND FEATURE WORK

Tree based data mining classification techniques such as Hoeffding tree, j48, Random Forest, Random Tree, RepTree were use in this study on intrusion detection dataset KDD Cup1999 by use WEKA 3.9 tool. In general result show using 10 fold cross validation Random forest best for Train set and J48 best for test dataset considering their comparative classification accuracy.

The big challenge in intrusion detection is to achieve high detection rate and low false alarm. Any single classifier is not sufficient to achieve high accuracy and low false positive or negative. Their for more than one classifier can be combined to improve overall performance of attack detection

REFERENCES

A Comparative Study for Outlier Detection Techniques in Data Mining by: Z. Bakar, R. Mohemad, A. Ahmad, M. Deris

A Survey of Outlier Detection Methodologies. Victoria J. Hodge and Jim Austin Dept. of Computer Science, University of York, York, UK

A. Patcha and J. Park "An overview of anomaly detection techniques", Existing solutions and latest technological trends.

Aditya Harbola, Jyoti Harbola "Improved Intrusion Detection in DDoS Applying feature selection Using Rank & Score of Attributesin KDD-99 data set" 2014

D. Hawkins, "Identification of outliers" Monographs on Applied Probability and Statistics, 1980

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Denning, Dorothy E., "An Intrusion-Detection Model," IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. SE-13, NO. 2, FEBRUARY 1987, 222-232

Denning, Dorothy E., and Neumann, Peter E, "Requirements and model for IDES-A real-time intrusion detection system," Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA, USA, 1985.

Dong, Hsu, Rajput Experimental analysis of application-level intrusion detection algorithms, International Journal of Security and Networks archive, March 2010

H. Javits and A. Valdes. "The NIDES statistical component" Description and justification. Technical report, SRI International, Computer Science Laboratory, 1993.

detection 344

Huan Liu, Hiroshi Motoda, Rudy Setiono "Feature Selection: An Ever Evolving Frontier in Data Mining" 2010

M. Mahoney "Network Traffic Anomaly Detection Based on Packet Bytes" Proc. ACMSAC 2003.

M. Mahoney, P. K. Chan. "Learning Non stationary Models of Normal Network Traffic for Detecting Novel Attacks", Proc. SIGKDD 2002

M. Markou, and S. Singh "Novelty detection: a review-part 1: statistical approaches

Mahbod Tavallaee, Ebrahim Baghe A Detailed Analysis of the KDD CUP 99 Data Set, 2009

S. X. Wu and W. Banzhaf "The use of computational intelligence in intrusion detection systems: A review", 2010

V. Barnett, and T. Lewis "Outliers in statistical data." John Wiley and sons, 1994.

W. Yassin, N. Udzir, A. Abdullah "Signature-Based Anomaly Intrusion Detection using Integrated Data Mining Classifiers" International Symposium on Biometrics and Security Technologies (ISBAST) 2014

W. Zhang, Q. Yang, and Y. Geng "A Survey of Anomaly Detection Methods in Networks," in Proc. International Symposium on Computer Network and Multimedia Technology, 1–3, January 2009.

Y. Dong, S. Hsu, S. Rajput, and B. Wu, "Experimental Analysis of Application Level Intrusion Detection Algorithms", International J. Security and Networks 2010

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

https://www.sans.org/reading-room/whitepapers/detection/history-evolution-intrusion-

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)