Intrusion Detection with Tree-Based Data Mining Classification Techniques by Using KDD

Mirza Khudadad(✉) and Zhiqiu Huang

College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics,
Nanjing 210016, Jiangsu, China
mirza_khudadad@hotmail.com, zqhuang@nuaa.edu.cn

Abstract. In the recent time a huge number of public and commercial service is used through internet so that the vulnerabilities of current security systems have become the most important issue in the society and threats from hackers have also increased. Many researchers feel intrusion detection systems can be a fundamental line of defense. Intrusion Detection System (IDS) is used against network attacks for protecting computer networks. On another hand, data mining techniques can also contribute to intrusion detection. The intrusion detection has two fundamental classes, Anomaly based and Misuse based. One of the biggest problem with the anomaly base intrusion detection is detecting a high numbers of false alarms. In this paper a solution is provided to increase the attack recognition rate and a minimal false alarm generation is achieved with the study of different Tree-based data mining techniques. KDD cup dataset is used for research purpose by using WEKA tool.

Keywords: Data mining · Intrusion detection system · Decision Tree J48
Hoeffding Tree · Rep Tree · Random Forest · Random Tree · KDD dataset

1 Introduction

With the passage of time, internet security is gaining a huge importance in the recent times. Data security has been suffering from numerous groups of attacks which are emerging as hazardous for trust of user and organization's repute now. So it's a need of time to propose the most effective and accurate detecting model for network data protection. The intrusion detection (ID) on computer networks, is a form of security management systems and hence intrusion detection system is implemented for knowing about computer attacks by examining different logs and records of data. The key role of a network IDS is a passive as it only works by gathering, identifying, logging and alerting IDS systems. It uses different attempts to identify intrusions that misuses as well as abuses the computer network system by malicious users in addition with some IDS monitors only a single computer while others have ability to monitor several connected computers on a network. There are two types of attacks as network based and host based. In host based an attack attempts to access a restricted service or resource from a single computer. While network based attack restricts legitimate users

X. Gu et al. (Eds.): MLICOM 2017, Part II, LNICST 227, pp. 294–303, 2018.
https://doi.org/10.1007/978-3-319-73447-7_33

from accessing several services of network by capturing network resources and its services as this can be achieved by sending a large number of network traffics. In network based attack, network traffic detection can be analyzed from the intrusion encounter by leading to two subcategories of anomaly inquiry systems. The 1st one is described with specification and set of rules. The 2nd one is based on learning and training the normal behavioral system. So like IDS, it is usually used for rule base intrusion detection in which rules are written manually for identification of known attacks. Other type is behavior based IDS and the benefit of this approach is to identify attempts, to exploit new and unforeseen vulnerabilities. One of the major problems of anomaly based IDS is detection of high false alarm and here in this paper this issue is solved by applying the different data mining Tree based algorithms as well as by finding the most appropriate algorithm that could give the best results on comparing to other algorithms.

2 Literature Survey

Anderson (1980) [1], 1st time presented his ideas related to IDS in his technical report as he accomplished the computer audit transformable mechanism and became able to provide a list of risks and warnings for techniques of computer safety. This discovery provided analytical way of applicability on user's behavior for disclosing those intruders who had an illegal system access. Therefore, in 1987 Dorothy gave the paradigm on intrusion recognition as Denning and Neumann both were the starters of intrusion exploration domain. With this they found the framework of intrusion-detection expert system and that was called IDES (Intrusion Detection Expert System) [2] as it was originated in 1985's paper of requirements and model on IDES – a real-time intrusion detection system [3]. Hoge and Austin provided a detailed investigation on anomaly disclosure by the help of machine learning and numerical processes [4]. Both of them recommended a study of latest operations for exceptional detections. Moreover Markou and Singh [5] granted a wide range of inspections for intrusion detection by employing ANN as well as arithmetical structure. Patcha and Park [6] further extended the research of various anomaly techniques concerning cyber intrusion detection. A lot of books and research materials were again observed for intrusion and irregularities of observation (Hawkins 1980, Barnett, Lewis 1994, Bakar, et al.) [7–9] and various anomaly detection systems are like NIDES (Next generation Intrusion Detection Expert System) [10], ALAD (Application Layer Anomaly Detector) [11] and PHAD (Packet Header Anomaly Detector) [12] for generating mathematical proven shape to an ordinary network data flow with warning generation technique was discovered on finding deviation in a normal model. After all, many of them used network packet header's feature extraction as ALAD and NIDES used the source, TCP connection state, port address and destination IP.

Zhang et al. [13] showed network survey related to techniques and methods of anomaly detection. Peng et al. (2007) [14] made exhaustive survey of techniques for detecting DoS and distributed DoS attacks. Wu and Banzhaf [15] analyzed the main methods of CI, including soft computing, swarm intelligence, artificial immune systems, evolutionary computation, fuzzy systems along with artificial neural networks.

Dong et al. [15] conferred the mechanism in accordance to them and proved to be a more credible on its comparison to Markov and K. mean Graph-based Sequence Learning Algorithm (GSLA) included construction, normal profile data pre-processing in addition with session marking. Within GSLA, an average figure was created by a session-learning lineup and it was defined to determine an anomaly period. Udzir [16] invented a Signature-Based Anomaly Detection Scheme (SADS) that could be enforced to study packet header behavior patterns with more precisely and promptly. Integrating data mining classifiers such as Naive Bayes and Random Forest could also be utilized in decreasing fake bugs for shortening the time of processing too. As a part of analysts likewise preferred the concept, selection of features to recognize intrusion. Liu et al. [17] described feature selection as a useful way for dimension downsizing injunction with a compulsory step in effectual data mining applications and its direct advantages include: sample building with better clear models, making data mining efficient and helping in preparing clear understandable data. Harbola [18] also used featured adoption procedure to advance accuracy. Its main aim was to deliver the broad conclusion of feature selection design for NSL-KDD intrusion identification dataset.

3 Intrusion Detection System

Intrusion can be said as an illegal attempt to get access to any network or system. The system regarding intrusion detection is developed to expose this kind of mistrustful activity on a network or device. The IDS examines hardware, software or a union of both to check the network flow for the hunt of intrusions. An intrusion detection system (IDS) reviews entire out going, in coming network activity and identifies doubtful patterns.

3.1 Type of IDS

Intrusion detection system can also be subcategorized under two main divisions as Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS).

3.1.1 Host-Based Intrusion Detection System

Host based intrusion detection (HIDS) defines the detection of intrusion which happens to a single host system. It is an application of a software which is installed on a system for the sake of its protection against intruders. HIDS is an operating system dependent so need some prior outlining ahead of its execution by having a capability of buffer overflow for attack's examination.

3.1.2 Network-Based Intrusion Detection System

Thus, network-based intrusion detection system (NIDS) has concern with network traffic control to secure a system from threats from network-based intrusion. All inbound packets and searches for any suspicious patterns are processed by NIDS. It is operating system independent and it appoints advanced protection to deal with denial of service (DoS) attacks.

3.2 Type of Attacks Detected by IDS

Four categories of IDS detected attacks (Table 1).

Table 1. IDS Detected Attacks

Denial of Service (DoS): Attack or an attempt which makes a network resource inaccessible to its expected legal users such as services suspension of a host connected to the internet
User to Root (U2R): Attack where an attacker attempts to get an unauthorized access of a targeted system
Remote to User Attack (R2L): Where an attacker tries to control a remote machine by guessing its password
Probing Attack (Probe): Where an attacker examines the machine to get useful information

4 KddCup'99 Dataset

For research objective the standard sets of data were published in KDD CUP 1999 [19]. IDS used it to assess various feature selection methods. This set of data has 41 features and 42nd feature shows the connection as 'Normal' or an attack nature. Here 4 main forms (DoS, Probe, R2L and U2R) cover this set of data which has altogether 24 kinds of attacks that have already been discussed.

Most of the datasets were repeated out of 5 million instances as just 10% KddCup'99 dataset was tried for training and verification of a suggested framework. There were 494021 instances in 10% of KddCup'99 dataset so 396743 instances were assumed to be in any one type of attack and remaining 97278 instances were declared as 'Normal' instances.

4.1 Preprocessing

In recommended model to reduce the performance evaluation complexity of 42nd feature of KddCup'99 dataset is defined in five leading sections in the pre-processing class labelled module. These labels i.e. DoS, Probe, R2L, U2R and Normal are considered as five subclasses which are formed in an action of pre-processing.

4.2 Splitting into Test and Train Dataset

The training and testing sets are two autonomous sets of the given data so testing set contains 44% of the dataset and other 66% of the data is assigned to training set. The derived model of accuracy is determined by the testing set as advised framework is concluded by the training set. After dividing it into two sets training set has 326054 instances and testing sets have 167967 instances.

4.3 Four Distinct Types of Attacks Used in Experimental Dataset

Categories of Attacks & Associated Tags (Table 2).

Table 2. Attack Categories & Associated Tags

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

Samples of KDD'99 Intrusion Detection Datasets (Table 3).

Table 3. Intrusion Detection Datasets Samples

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

5 Results and Experiment

We performed the experiment with KDD cup dataset by using 10% [20] train and test dataset (using WEKA).

5.1 Experiment Setup

Experiment performed under following hardware and software.

- Hardware: Intel core i5, 1.8 GHz processor with 4 GB Ram.
- Software: Microsoft Windows 10, WEKA 3.7.

5.2 Using Train Dataset

Experiment performed under the above mentioned hardware and software system specifications (Tables 4, 5, 6 and 7).

Table 4. Classifiers & Instances using Train 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

Table 5. Classifiers & DoS, PROBE Class Attacks using Train Dataset

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

Table 6. Classifiers & R2L, U2R Class Attacks using Train Dataset

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

Table 7. Classifiers & Normal Class Attacks using Train Dataset

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

5.3 Using Test Dataset

See Tables 8, 9, 10, and 11.

Table 8. Classifiers & Instances 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**
Random Tree	98.0371	1.9629
RepTree	98.0262	1.9738

Table 9. Classifiers & DoS, PROBE Class Attacks using Test Dataset

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

Table 10. Classifiers & R2L, U2R Class Attacks using Test Dataset

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

Table 11. Classifiers & Normal Class Attacks using Test Dataset

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

6 Result and Analysis

Percentage of results using Test Set.

The above table shows the results of test dataset that proves J48 classifier performs well in U2R, R2L and Normal categories (Fig. 1). In DoS and PROBE, Random Forest (RF) has a minor difference (Table 12).

Percentage of result using Train set.

In the above table it is analyzed that more than 90% attack detection is done by all classifiers in DoS, PORBE as well as in R2L but the Normal category has more than 99% of attack detection results as only in U2R attack its ratio is less than 75% and it's just because of having fewer attacks in training dataset (Fig 2). By comparing to other classifiers it is proven Random Forest performs slightly better in DoS, U2R, R2L but J48 works better only in PROBE (Table 13).

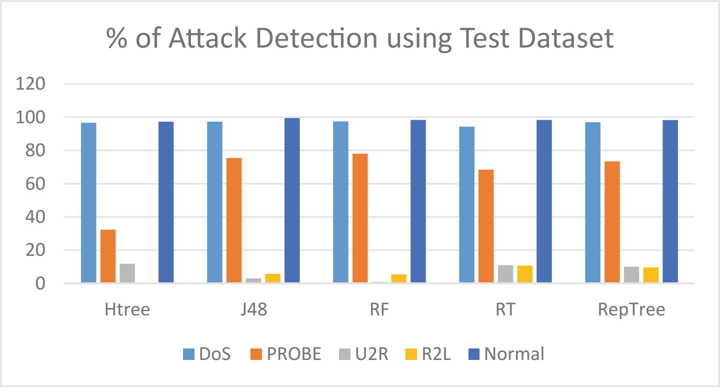

Fig. 1. Attack Detection Analysis with Test Dataset

Table 12. % of Results using Test Set

	DoS	PROBE	U2R	R2L	Normal
Htree	96.67048	32.35717715	11.84210526	0.012354	97.2884657
J48	97.31785	75.42006721	3.070175439	5.843474	99.485089
RF	**97.42401**	**77.98847816**	0.877192982	5.49756	98.3034344
RT	94.3342	68.45895343	10.96491228	10.71098	98.3298401
RepTree	96.97676	73.45175228	10.0877193	9.691766	98.2456719

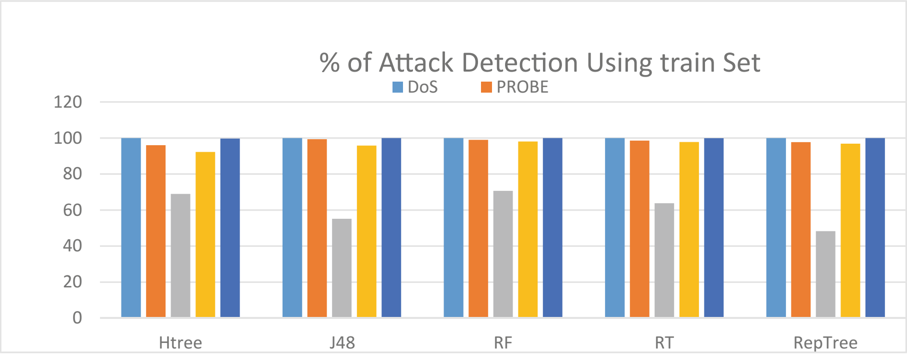

Fig. 2. Attack Detection Analysis with Train Dataset

Table 13. % of Results using Train Set

	DoS	PROBE	U2R	R2L	Normal
Htree	99.97140515	96.03117	68.9655172	92.26667	99.66706
J48	99.99493896	99.36693	55.1724138	95.82222	99.96172
RF	**99.9994939**	99.0017	**70.6896552**	**98.04444**	**99.97685**
RT	99.9936737	98.56343	63.7931034	97.77778	99.93412
RepTree	99.99114319	97.68688	48.2758621	96.88889	99.95638

7 Conclusion and Future Work

The classification techniques like Hoeffding tree, J48, Random Forest, Random Tree and RepTree of tree based data mining algorithms were practiced to study intrusion detection dataset of KDD Cup1999 by using WEKA 3.9 tool. In general results show using 10 fold cross validation, Random forest is the best for train set and J48 is the best for test dataset by considering their comparative classification accuracy.

Achieving high detection rate along with the lowest false alarm ratio is the biggest challenge to intrusion detection so not even a single classifier is efficient enough to give high veracity of decreasing false alarm percentage. Finally, to improve overall attack detection performance two or more classifiers can be combined.

References

1. https://www.sans.org/reading-room/whitepapers/detection/history-evolution-intrusion-detection344
2. Denning, D.E.: An intrusion-detection model. IEEE Trans. Softw. Eng. **SE-13**(2), 222–232 (1987)
3. Denning, D.E., Neumann, P.E.: Requirements and model for IDES-A real-time intrusion detection system. Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA, USA (1985)
4. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. J. Artif. Intell. Rev. **22**, 85–126 (2004)
5. Markou, M., Singh, S.: Novelty detection: a review-part 1: statistical approaches
6. Patcha, A., Park, J.: An overview of anomaly detection techniques. Existing solutions and latest technological trends
7. Bakar, Z., Mohemad, R., Ahmad, A., Deris, M.: A comparative study for outlier detection techniques in data mining
8. Hawkins, D.: Identification of Outliers. Monographs on Applied Probability and Statistics. Springer, Heidelberg (1980). https://doi.org/10.1007/978-94-015-3994-4
9. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, Hoboken (1994)
10. Javits, H., Valdes, A.: "The NIDES statistical component" Description and justification. Technical report, SRI International, Computer Science Laboratory (1993)
11. Mahoney, M.: Network traffic anomaly detection based on packet bytes. In: Proceedings of ACMSAC (2003)
12. Mahoney, M., Chan, P.K.: Learning non stationary models of normal network traffic for detecting novel attacks. In: Proceedings of SIGKDD (2002)
13. Zhang, W., Yang, Q., Geng, Y.: A survey of anomaly detection methods in networks. In: Proceedings of International Symposium on Computer Network and Multimedia Technology, pp. 1–3, January 2009
14. Wu, S.X., Banzhaf, W.: The use of computational intelligence in intrusion detection systems: a review (2010)
15. Dong, Y., Hsu, S., Rajput, S., Wu, B.: Experimental analysis of application level intrusion detection algorithms. Int. J. Secur. Netw. **5**, 198–205 (2010)
16. Yassin, W., Udzir, N., Abdullah, A.: Signature-based anomaly intrusion detection using integrated data mining classifiers. In: International Symposium on Biometrics and Security Technologies (ISBAST) (2014)

17. Liu, H., Motoda, H., Setiono, R.: Feature selection: an ever evolving frontier in data mining (2010)
18. Harbola, A., Harbola, J.: Improved intrusion detection in DDOS applying feature selection using rank & score of attributes in KDD-99 data set (2014)
19. Tavallaee, M., Baghe, E.: A detailed analysis of the KDD cup 99 data set (2009)
20. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html2