International Journal of Hybrid Information Technology
Vol. 9, No.12 (2016), pp. 201-208
http://dx.doi.org/10.14257/ijhit.2016.9.12.18

Intrusion Detection by Using Hybrid of Decision Tree And K-Nearest Neighbor

Bilal Ahmad[1], Wang Jian[1] and Muhammad Shafiq[2]

[1]Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.
[2]Electronic Engineering Department, Sir Syed University of Engineering & Technology, Karachi, Pakistan.
ahmad@nuaa.edu.cn.

Abstract

In the modern age of information technology security of valuable asset become much important issue. Intrusion detection system plays a most important role in this area. It protects the system by attacks or threats by unauthorized access or person. The previous study has identified the need for more enhancements in the research of intrusion detection. This study gives the outline for intrusion detection and proposed a hybrid classification based method based on Decision Tree and K-Nearest Neighbor. This experiment perform on the bases of cross-10 fold validation techniques on the basis of decision tree and KNN classifiers and proposed hybrid classifier by using KDD cup dataset. Experimental result shows that the proposed idea gives good result as compared to individual base algorithms

***Keywords**: Intrusion detection,, decision tree, k nearest neighbor, Network security*

1. Introduction

Intrusion detection system is security mechanisms which use to detect abnormal behavior in the network. By the high progress in information technology the use of computer network highly increase, network security become more important. Landwehr, Bull, McDermott & Choi, 1994) show all network have security issues which can be difficult to handle and also expensive for production. Therefor the idea of intrusion detection in network security is more important. This is also important that intrusion detection not only to detect intrusion but also monitors the attempts to break security. (Chen, Hsu &Shen, 2005)[1] Describe the old security methods like authentication, programming errors, encryption or firewall can be first line of protection. Weak password cannot be prevent unauthorized access or use, sometime firewall are defenseless to mistakes in formation and irregular to unclear or undecided security rules. So intrusion detection is needed as an extra strength for system protection and privacy. (Heady, Luger, Maccabe, &Servilla, 1990)[2] (Sundaram, 1996) [3] intrusion detection get to collect and analysis number of key point from system and network to find any unusual behavior against the policies of the system. The mixture of hardware and software for detection intrusion is IDS (Zhao, Chen & Lou, 2011) [4]. There are two main tactics for intrusion detection. Misuse and anomaly detection. Misuse detection contains the pattern which already defines in the database. This kind of detection also recognized as signature base detection. On the other hand anomaly detection consider as anomalous. This means for establish normal activity profile we can have theory to flag all system states to establish outline as intrusion strike. These two type of detection have own strength and weakness. The former can notice known attacks and have great correctness vs pattern matching on known signatures, but drawback is that they cannot detect unknown attack. The other techniques anomaly can detect unknown attacks by statistical analysis but drawback is that it has high number of false alarm issue. In the propose methods we design hybrid algorithm which have ability to low number of false alarm and also increase detection accuracy by detecting known and unknown attacks. In the experiment we use dataset NSL-KDD 1999. A standard dataset. It was formed from KDDCUP 1999. Through the experiment we examine four types of attack DOS, U2R, R2L and Probe.

Intrusion detection system made analysis the logs or information collected by sensors, and back to a mixture of input the sensors to system admin. Admin gets the solution or treatment handled by IDS. In these days data mining comes like a vital tool for analysis the input of sensors. Following figure shows the setup of IDS to make protection server from internal and external network.

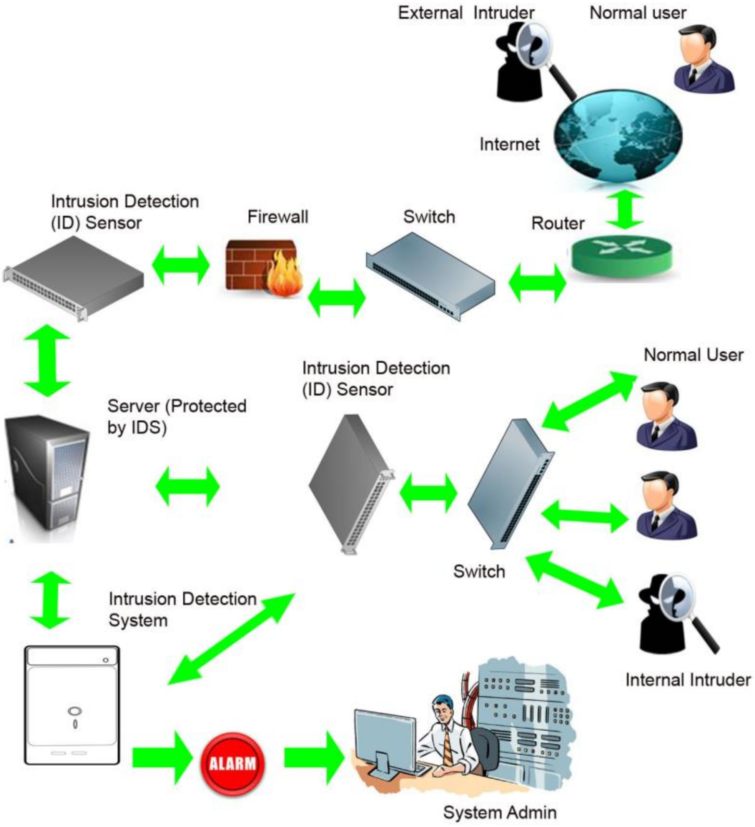

External Intruder
Normal user
Internet
Intrusion Detection (ID) Sensor
Firewall
Switch
Router
Server (Protected by IDS)
Intrusion Detection (ID) Sensor
Normal User
Switch
Intrusion Detection System
Internal Intruder
ALARM
System Admin

Figure 1. A Typical Setup of Intrusion Detection

Intrusion detection must have detection rate for attack of 100% also with false positive 0%. Actually, practically it’s very hard to cover it. The very important parameters or limitations included in the efficiency estimation of intrusion detection shows following in Table 1.

Table 1. Limitations for Presentation Approximation of IDS

Parameters	Definition
True Positive (*TP*) or Detection Rate (*DR*)	Attack occur and alarm raised
False Positive (*FP*)	No attack but alarm raised
True Negative (*TN*)	No attack and no alarm
False Negative (*FN*)	Attack occur but no alarm

2. Literature Review

Early 1980, the idea of intrusion detection started by Anderson's paper (Anderson, 1980) [5] he produced a risk classification idea that made a security scanning scrutiny system depends on detection irregularities in user behavior. In 1986, Dr. Denning introduce many ideas for IDS built on "statics", "Markov chains", "Time series", *etc* (Denning, 1987) [6]. In the starting 1980's stand-ford research institute develop an IDS expert system that observe user behavior and doubtful events (Denning & Neumann, 1985) In 1988, anomaly detection statistical based IDS proposed by Haystack (Smaha & Haystack, 1988) [7], which had capabilities both user & collection/group based anomaly detection tactics. In (1996, Forrest) [8] planned a similarity among the human protected scheme and intrusion detection that include monitoring program system call sequences to make a normal profile. In 2000 (valdes and Skinner) [9] made an anomaly based IDS that used NB network to execute intrusion detection on traffic bursts

In 2003 (Kruegel, Mutz) [10] developed multisensory fusion scope used NB classifier for organization and defeat of false alarms that comes from different IDS devices were gathered to produce single alarm. In 2003(Yeung & Ding, 2003) [11] developed an "anomaly-based intrusion detection" by the use of "Markov model" that figure the sample probability of practical sequence using the onward or backward algorithm for classifying anomalous. In 2008, (P. Garcı a-Teodoroa) [12] provide survey of the famous and well-known anomaly-based intrusion detection approaches, available platforms, schemes under grow and outline the main experiments to be allocated with for the big measure deployment of anomaly-based intrusion detectors, with special stress on assessment issues. In 2010 (Anna Sperotto, Gregor Schaffrath) [13] give a survey of research in the field of flow based intrusion detection. Secondly discuss the suitable platform, system under construction and research areas for given topic. Finally discuss the main challenges for implementation in high scale anomaly based IDS

In the same year, 2010 (Shelly Xiaonan Wu) [14] provide review about an outline of the research progress in applied computational Intelligence approaches to the problem of intrusion detection. The research acquired essential approaches of CI, ANN, fuzzy systems, artificial security systems, evolutionary computation, swarm intelligence, and soft computing. In 2016 (Anna L. Buczak) [15] provide a survey in which Data mining methods is explained, debate on challenges by usage ML/DM for internet security and some ideas on when to use a given method are provided.

3. Dataset Description

Since KDD99 1999 (Yimin, 2004) [16] most use and popular dataset for the assessment of anomaly detection approaches. The dataset make on the bases on data capture in "DARPA-98 evaluation program" (KDD, 99). "DARPA-98" is near about 4 GB of compact binary tcp-dump data of seven weeks of network flow. The dataset contain two million connection records for two weeks network flow for test dataset. KDD training dataset have near about "4,900,00" single connection paths every of which contains 40 features and label as normal or abnormal, which can specify as following attack types

Denial of Service Attack (DOS): It is a malicious attempt in which the attacker makes server or network resource unavailable or much busy to hold the requests. User to Root Attack

User to Root (U2R): It is a type of exploit in which attacker has a local normal user account access the system. But an attacker takes the advantage of present vulnerabilities in the system like sniffing passwords, and gets the super user privilege access.

Remote to Local Attack (R2L): is a type of attack in which an attacker machine is able to send a packet remotely but does not have an account on the victim machine. So by taking advantage of any vulnerabilities on the target machine, the attacker gets access to the victim machine.

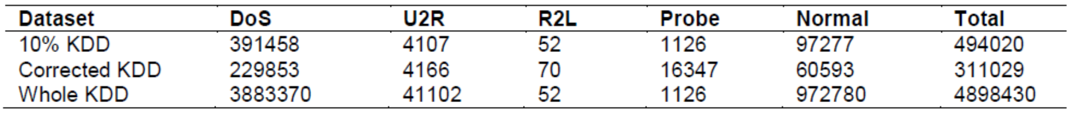
Probing Attack: It is a very basic and initial step of misusing any system. The attacker scans a machine to find out the weakness or vulnerabilities in the network to exploit the victim machine.

3.1. Corrected KDD Dataset

The KDDCup99 dataset have high number of useless records which cause the algorithms to be unfair and so stop them, in corrected KDD dataset all jobless or useless removed so in this way classifier may not be unfair

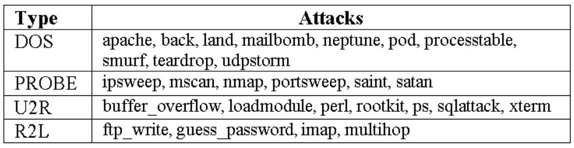
3.2. 10% KDD Dataset

The KDDCup99 have much huge in size so rarely it useed completely for train or test phase. Mostly 10% of dataset use. Following figure the total of instances dataset and the number of specific attacks present in each of the variant is given.



Dataset	DoS	U2R	R2L	Probe	Normal	Total
10% KDD	391458	4107	52	1126	97277	494020
Corrected KDD	229853	4166	70	16347	60593	311029
Whole KDD	3883370	41102	52	1126	972780	4898430

Figure 2. Attacks Division in Dataset



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

Figure 3. Attack Types with Their Corresponding Type

4. Decision Tree and K-Nearest Neighbor

Decision tree known as classification methods from data mining. The classification method inductively educated to made model from categorized group of record .Decision tree categories the given records by the use of their attributes. DT starts made by the group of already classified data.

DT use the training record which labeled in standing of the attributes. The big problem here to determine the attribute which can be good divider for the records into many classes. ID3 uses the info academic method for this problem. Info theory use the thought of “Entropy”, which determine to make data clear. The rate of “Entropy” is little when the class division is not event, it happen when all data item have one class. The rate of “Entropy” become higher when class division become more even, then happen when data item have more classes. “Information Gain” based on unity of every attribute in categorizing the data item. For classification unknown object started from root of DT and tracts the branch given by the result of every test until a leaf node come. Decision tree starts from the implementation of many algorithms some are ID3 and after extends into C4.5. It avoid over fitting records it handles nonstop attributes and able to get a suitable attribute selection. It carries the training data with missing values and improve computation performance. DT have ability to get the organization problem from intrusion detection by learning the model from the dataset and determine the new data item from one of class mentioned by dataset. DT uses “misuse intrusion detection” as can learn a training data based on model and can guess the upcoming data as one of the threat cast or normal as on leaned model. DT can carry large dataset also can handle big data flows in network so high performance of DT made it useful in intrusion detection. DT make easily describable model that useful for admin for inspection.

The k Nearest Neighbor is a part from data mining society. This algorithm is determined by use of the distance which calculated by using the difference in the distances between each of the topographies of the instance and

It's surrounding.

5. Ensemble Mathods

Ensemble techniques are used to combine many theories, mostly this idea used to find strong learner method among weak learner. Ada boost Boosting bagging are methods which are famous in this area and used to make less over-fitting drawback comes from machine learning.

Bagging or bootstrap aggregation is only coolest but extremely positive set of theory for improving the glitches in classification. For example, algorithms as like decision tree algorithms can be variable , especially when the label of a point less change and training can behavior into many different trees . This theory is mostly used for DT algorithms, but also useful with other classification algorithms as like NB, K-NN, induction rule, *etc*.

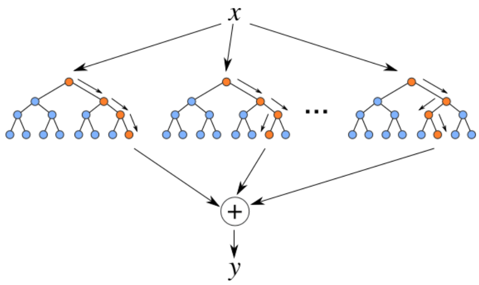
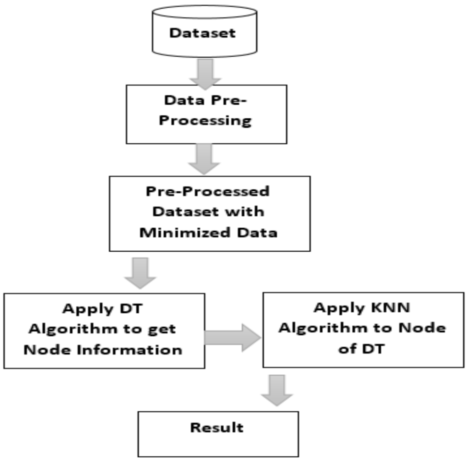


Figure 4. Decision Tree in Bagging

This method (Figure.4) create many datasets by “bootstarping”, create one decision trees for each dataset. Combine each tree by averaging (or voting) final decision, on average batter result then individual result

6. Proposed Hybrid Algorithm


Dataset
Data Pre-Processing
Pre-Processed Dataset with Minimized Data
Apply DT Algorithm to get Node Information
Apply KNN Algorithm to Node of DT
Result

Fignure 5. Proposed Algorithm

To achieve the efficiency of the algorithms (Figure 5), both was trained on the KDD dataset with "10-fold cross validation" test mode. To test and determine the method we use "10-fold cross validation". In this procedure the dataset is separated in 10 subset. Every time one of the ten subsets is use as a test-set and the other k-1 subset form the train set. Performance analysis measured through all 10 trials this provides a batter signal how well the method/classifier can work on unseen data. Hybrid DT-KNN perform better than the separate decision tree and KNN for normal class. For Probe and Normal classes it perform better than a separate decision tree and KNN method.

Table 2. Results

Attack Type	Classifiers Performance		
	Decision Tree	KNN	DT-KNN
Normal	100%	99.97%	100%
U2R	75.00%	86.54%	86.54%
PROBE	99.49%	99.44%	99.50%
R2L	98.05%	98.76%	98.76%
DOS	96.83%	98.32%	99%

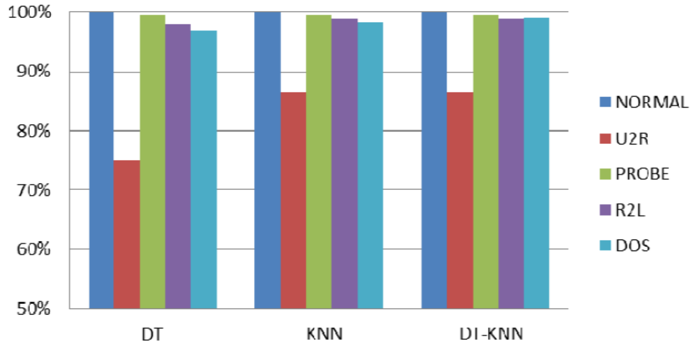


Figure 6. Graph for the Accuracies of the Three Classifiers (DT, KNN and DT-KNN)

After the above results (Figure.6), we can achieve that while the node info collected from the decision tree did improve the performance of KNN, on the entire hybrid DT-KNN model did not have the expected 100 percent progress in many classes of attacks, but it succeeds the individual decision tree and K Nearest Neighbor classification algorithm. Figure above defines the performance of the correctly classified instances of each algorithm. After classification of KDD test dataset, it is shown that the hybrid DT-KNN algorithm shows the higher detection accuracy.

5. Conclusion and Future Works

As from experiment on KDD Dataset, the given hybrid classifier could get an accuracy of 100% with a false positive are of 0%. Comparison with other NIDSs that also useful KDD cup dataset, this classifier present good performance in U2R, R2L, DoS and Probe attacks, through this was not good for U2R and R2L attacks unless as it gives equal detection rate as KNN but it takes long time which can be ignored. However, in term of accuracy, the given classifier could get the best progress at 100%. This practical was performed on 10% KDD cup dataset. Some new attack instances in dataset which never appeared in training can also detected by this system

References

[1] G. Stein, B. Chen, A. S. Wu and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection", In Proceedings of the 43rd annual Southeast regional conference, vol. 2, **(2005)**, pp. 136–141.

[2] R. Heady, G. Lugar, A. McCabe and M. Servilla, "The architecture of a Network Level Intrusion detection system", Technical Report, CS90-20, Computer Science Dept., University Of New Mexico, Albuguerque, NM87131-USA, **(1990)**.

[3] A. Sundaram, "An introduction to intrusion detection. Magazine Crossroads", Special issue on computer security Crossroads Homepage archive, vol. 2, issue 4, **(1996)**, pp. 3-7.

[4] J. Zhao, M. Chen and Q. Lou, "Research of intrusion detection system based on neural networks", Proceedings of IEEE 3rd international Conference on Communication Software and Networks, Xi'an, China, **(2011)**, pp. 174-178.

[5] J. P. Anderson, "Computer security threat monitoring and surveillance", Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, **(1980)**.

[6] D. E. Denning, "An intrusion detection model", IEEE Transaction on Software Engineering, SE-, vol. 13, no. 2, **(1987)**, pp. 222-232.

[7] S.E. Smaha and Haystack, "An intrusion detection system", in Proc. of the IEEE Fourth Aerospace Computer Security Applications Conference, Orlando, FL, **(1988)**, pp. 37-44.

[8] S. Forrest, S.A. Hofmeyr, A. Somayaji and T.A. Longstaff, "A sense of self for Unix processes", in Proc. of the IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, **(1996)**, pp. 120-128.

[9] A. and K. Skinner, "Adaptive model-based monitoring for cyber-attack detection", in Recent Advances in Intrusion Detection Toulouse, France, **(2000)**, pp. 80-92.

[10] C. Kruegel, D. Mutz, W. Robertson and F. Valeur, "Bayesian event classification for intrusion detection", in Proc. of the 19th Annual Computer Security Applications Conference, Las Vegas, **(2003)**.

[11] D.-Y. Yeung and Y. Ding, "Host-based intrusion detection using dynamic and static behavioral models", Pattern Recognition, **(2003)**, pp. 229–243

[12] P. Garcı´a-Teodoroa, "Anomaly-based network intrusion detection: Techniques, systems and challenges", Journal Computers and Security archive, vol. 28, issue. 1-2, **(2008)**, pp. 18-28.

[13] A. Sperotto, "An Overview of IP Flow-Based Intrusion Detection", IEEE Communications Surveys & Tutorials, vol. 12, no. 3, **(2010)**

[14] S. X. Wu, "Applied Soft Computing", vol. 10, issue 1, **(2010)**, pp1–35.

[15] A. L. Buczak, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE Communications Surveys & Tutorials, vol. 18, no. 2, **(2016)**.

[16] Y. Wu, "High-dimensional Pattern Analysis in Multimedia Information Retrieval and Bioinformatics" , Doctoral Thesis, State University of New York, **(2004)**.