Available online at www.ijsrnsc.org

IJSRNSC

Volume-5, Issue-6, December 2017
Review Paper

Int. J. Sc. Res. in
Network Security
and Communication

E-ISSN:2321-3256

www.ijsrnsc.org

A Study of Intrusion Detection System using Efficient Data Mining Techniques

P. Rutravigneshwaran

Department of Computer Application, Rev. Jacob Memorial Christian College, Ambilikkai, India

**Corresponding Author: rutra20190@gmail.com*

Received: 16/Oct/2017, Revised: 29/Oct/2017, Accepted: 18/Nov/2017, Published: 31/Dec/2017

Abstract- IDS is a software consequence monitors the humiliation or behavior plus investigate any immoral operation suggest itself. Fantastic increase and tradition of internet raises concerns in relation to how to defend and communicate the digital in order in a safe approach. Nowadays, hackers use different types of attacks for getting the valuable information. IntheproposedFast Hierarchical Relevance Vector Machine (FHRVM), AnalyticalHierarchy ProcessMethod (AHP) isusedtoselect the inputweightsandhiddenbiases. Simulation has been carried out using Math works MATLAB R2012a. KDD Cup 1999 dataset istakenfor testingthe performanceoftheproposedworkandtheresults indicate that FHRVM has achieved higher detectionrate and lowfalse alarmrate thanthat ofexistingSVMalgorithm. This research evaluate the efficiency of machine learning methods in intrusion detection system, together with classification tree and support vector machine, with the expect of given that reference for establishing intrusion detection system in future. Compared with further interrelated works in data mining-based intrusion detectors accuracy, detection rate, false alarm rate. It moreover show improved act than KDD Winner, particularly used for two types of attacks namely, U2R type and R2L type. Comparison results of C4.5, SVM. we finds that C4.5 is superior to SVM in accuracy and detection; in accuracy for Probe,Dos and U2R attacks, C4.5 is also better than SVM and FHRVM; but in false alarm rate FHRVM is better. In this paper enhance that FHRVM is better than c4.5 and SVM for U2R attack & R2L attack.

***Keywords*:** Classificationtree, SVM, FHRVM, Internet attack, Intrusion detection system (IDS)

I. INTRODUCTION

The Internet has become a part of daily life and a vital role today. Internet has been used asan important component of industry models. For the industry users it can be used for both industry and customers applythe Internet application such as website and e-mail on industry performance. Therefore, information security of using Internet as themedia needs to be carefully concerned [1]. Intrusion detection is onemajor research problem for business and personal networks.As there are many risks of network attacks under the Internetenvironment, there are various system designed to blockthe system. IDS aid the network to resist outside attacks. Theprimary goal of IDS isused to provide a wall of defense to confront theattacks of computer systems on the network. It can be used ondistinguish between the dissimilarity between the types of malicious network communicationsand computer systems usage,

Whereas the conventional firewall cannot perform this task. Intrusion detection is based on the assumption that the behavior of intruders different from a legaluser.Nevertheless while internet bring concerning expediency and concurrent lines,consequently comes in sequence protection difficulty.Forexample: servers are attacked interior data and information is stolen. In the event of such cases, big losses in network may be caused in cash and industry credit. With the growth of Internet there have seen a tremendous increase in the number of attacks, the intrusion detectionsystem has become a main stream of information security. With the help of the firewallit can be used to provide some protection butthey do not provide full protection. The purpose of the intrusion detection system is to help computer systems todeal about attacks. There are two kinds of IDS that can be used to base on the types of operations used to detectintrusions [2]. Anomaly detection system creates a database of normal behavior and any deviations from the normalbehavior are occurred an alert is triggered regarding the occurrence of intrusions. Misuse Detection system stores thepredefined attack patterns in the database if a similar data and if similar situations occur it is classified as attack.Based on the source of data the intrusion detection system are classified to Host based IDS and Network based IDS.In network based IDS the individual packet flowing through the network are analyzed. The host based IDS analyzes the activities on the single computer or host[5].

At current, hackers are unnecessaryto have a wide knowledge of specialized knowledge, and yearlyinternet attack cases are increasing to a great extentCommon enterprises adopt firewall as the first line of defensefor safety in network to superviseaccessing behaviors of internet, and it owns restricted detectioncapability for internet attacks. Therefore, Intrusion Detection System,IDS is always applied to detect internet encapsulation, to improveprotectingcapability of internet safety. IDS appears like internet supervision and alarm device, to surveyand consider whether the internet attacks may suggest itself, timelysend alarm before risks are caused by attacks, execute correspondingresponse measures, and reduce rate of bigger losses.Moreover, some technologies are based on pattern check, withlow mis-judgment rate, but the pattern-based should be upgradedon a normal basis, such technology does not have enough detectioncapacity for unknown and renewed attack manners.

The research process is shown in Fig. 1.

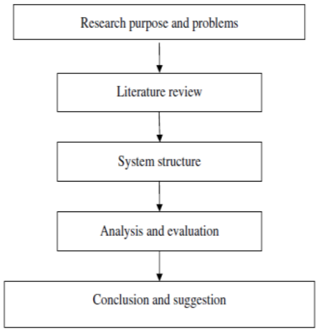

Fig. 1. Research flow.

Recently, many researches applied the technology of data mining and machine learning, which can analysis bulk data, and such technologies own better detection capacity for unknown attacks. Though some research achievements have been scored, there is a lot of development potential. Under such situation with most same conditions, how is the efficiency of different machine learning methods applied in intrusion detection. In addition the said manners, what methods are there? Therefore, the research intends to compare the efficiency of different machine learning methods applied in intrusion detection, include classification tree, support vector machine, and so on, with the hope of providing possible suggestion for improvement, as the reference for IDS.

II. REVIEW OF LITERATURE

A) AN INTRODUCTION ABOUT IDS:

An **intrusion detection system** (**IDS**) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to anorganizationlocation. Some systems may try to end an intrusion effort but this is neither required nor expected of a monitoring structure. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incident, sortingin order about them, and exposure attempts. In addition, organizations use IDPS for other purpose, such as identifytroubles with protection policies, document existing threats and deterring individuals from violating security policies.This method goes one step further than a firewall and gives you additional security to ensure that your data is safe and protected, by combining the protection of the IDPS with a firewall.Network IDS (NIDS) and Host-based IDS (HIDS) systems. The NIDS analyses the data through the network. It does this by analyzing the audit logs in the systemand intrusion detectionsystem with various patterns was put forward.Analyze anysignal arising from related safety problems, send alarms whensafety tribulationscrop up, and inform interconnected personnel units totake relevant measures to reduce possible risks[8].

Itsframework includes three parts:

1. Information collection: Data collection: the source of these collecteddata can be separated into host, network and application,according to the position.
2. Analysis engine: Analysis engine is able to analyze whether ornot there are symptom of any intrusion.
3. Response: Take actions after analysis, record analysis results,send real-time alarm.

B) Types of intrusion detection system:

There are two kinds of classification methodsfor intrusion detection system:

1. According to different data sources, intrusion detection systemincludes host-based IDS and network-based IDS.
2. According to different analysis methods, intrusion detectionsystem includes Misuse Detection and Anomaly Detection.

The advantage and disadvantage of theintrusion detectionsystems.

a) Host-based IDS: A host-based intrusion detection system (HIDS) is a system that monitors a computer system on which it is installed to detect an intrusion and/or misuse, and responds by logging the activity and notifying the designated authority. Whether internal or external, has circumvented the system's security policy.

C) Pros and cons:

A) It can judge whether or not the host is intruded moreaccurately: Because its data comes form system auditrecords and system logs of hosts, comparing with network-based intrusion detection system, it can moreaccurately judge network attacks or intrusion onhosts.

B) It can detect attacks under encrypted network environment:Because the data comes from system filesand transmitted encrypted data in network.

C). it does not need additional hardware: It just needsmonitoring system installed in specified hosts, withoutadditional hardware.

D). Higher cost: Monitoring systems must be installed ineach host; and because of different hosts, the auditfiles and log pattern are accordingly different, thusdifferent intrusion detection systems are required ineach host.

E) It may affect system efficiency of monitored hosts:Intrusion detection system in monitoring state mayoccupy system sources of hosts.

F) Network-based IDS:

Its data is mainly collectednetwork generic stream going through networksegments such as Internet packets.

G) Pros and Cons in IDS:

1. Low cost: Only network-based IDS are able toidentify allattacks in a Local Area Network, and the cost is just for the device.
2. It can detect attacks that cannot be done by hostbased IDS, such as: Dos, DDos.
3. The flux is large, and some packets may be lost, and itcannot detect all packets in network.
4. In large-scale network, it requires more rapidCPU and more memory space, to analyze bulkdata.

(H) Different Types of analysis method in IDS:

Misuse Detection:

Intrusion is well-defined attacks on known weak points of system. All Intrusion which object is to misuse system resources and break it, are fall in this categories. Misuse intruder can be detected by watching for certain action being performed on certain objects and also by doing the pattern matching on audit trail information. Its advantage is high detection rate andlow false alarm rate for known attacks.

Anomaly Detection: Anomaly-based signatures are typically geared to look for network traffic that deviates from what is seen normally. It is based on an supposition that intruder's behavior is different from normal users'behavior.

The Detection rate of the method is high, and itis more likely to detect un-known attacks, but mis-judgmentrate is also high [7].

Hybrid: The advantage of misuse detection is low misjudgmentrate, as well as low detection capacity forunknown attacks; comparatively, anomaly detectionowns the capacity of detecting unknown attacks, butwith high mis-judgment rate [8].

Current analysis method

The current analysis methods are mainly applied inintrusion detection system. Some of the types of current analysis methods are,State transition: State transition is applied to describe the relation of arisingevents, which is usually used for misuse detection.[10]

Statistical models: Statistics method is applied to construct normalbehavior mode, including: threshold measures, mean andstandard deviation, multivariate model, clustering and outlierdetection which is usually used for anomalydetection.

Neural network: In computer science and related fields, artificial neural networks are computational models inspired by animals' central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition.[11] They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network.

Bayesian network: Graph method is applied to express the relationamong variables[12] when performing detection, conditional probabilityis used to calculate proper detection value.

Rule-based: Behavior or mode is expressed by rule method, andthose conforming to rule can be judged to be attack behavior.[9] It’s commonly used for misuse detection.

Data mining (Machine learning) methods: It concludes Markovprocess model classification tree support vector machine,association rule, link analysis, sequence analysis, and so on.

Machine learning:

Machine learning is widely applied in various areas such as: Biological signature differentiation, search engine, medicaldiagnosis, and bond market analysis, some of the common machine learning technologies:

Bayesian decision theory

Multivariate methods

Clustering

Classification trees

Linear discrimination

Multilayer perceptions

Local models

Hidden Markov models

Reinforcement learning

I) *Classification tree*

Classification tree is a prediction mode in machine learning andit is also called Decision tree [4]. The most fundamental and commonalgorithm used for classification tree is FHRVM,C4.5 and SVM.

Three types of tree construction methods are Top-down treeconstruction, Bottom-up pruning,FHRVM belongto top-down tree construction.

III. SYSTEM STRUCTURE

System structure graph

The proceeding flow of the research is

KDD Cup 99 dataset:

The data applied in the research comes from KDD Cup 99dataset,which was initially used for The Third International KnowledgeDiscovery and Data Mining Tools Competition.It was proposed to assess the efficiency of intrusion detectionalgorithm. Therefore, the research also applies the dataset.[13]There are approximately 4,940,000 kinds of data in trainingdataset, 10% of which is provided; there are 3,110,291 kinds of datain test dataset, and there are totally 47 types of network connection characteristic.

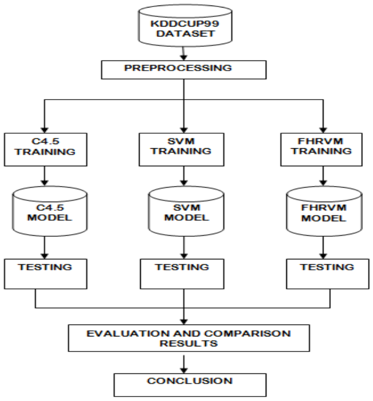

Fig.2

Data) in each kind of network connection record. And its propertycan be divided into three major types: Basis characteristic of networkconnection, characteristic of network connection content,network transmission characteristic; Data pattern include nominal,binary and numeric. There are 23 types of attacks contained in traininginformation, and 37 types of attacks contained in test information,14 types of attacks more than training information,[14] thus test informationcan be used to assess the detection capacity for unknown attacks.[3] The attacks contained in test information can be separatedinto the following major types:

Prob. Attackers usually apply probe catch information, to determine the targets and the type of operation system.

Dos (Denial of service.A distributed denial-of-service (Dodos) attack is one in which a multitude of compromised systems attack a single target, thereby causing denial of service for users of the targeted system. The attackusually occupies all system source of server, or occupies theband width and disables system resource and makes operation stop.

U2R (User gain root): In the attack, users take advantage of systemleak to get access to legal purview or administrator’s purview,such as: Buffer Overflow is among them.

R2L (Remote file access): The attack is to apply the advantage of server providing services, to get related safety setting or user’sencrypted files, such as: Unicode leak, SQL Injection, and so on.

Preprocess of data

The research intends to compare the efficiency of C4.5, SVMand FHRVMunder different circumstances, sample training dataset (10% kddcup.data_10_percent.gz) and testdataset. Based on the normal proportion, select each 10,000 groupof data where normal proportion is 10%, 20%, 30%. . . 90% in trainingdataset and test dataset.

Training and testing

Training stage of SVM also requires setting parameter helpfulto provide python program seeking optimization parameter.

IV. ANALYSIS AND EVALUATION

True positive (TP): The amount of attack detected when it iscorrectly identified.

True negative (TN): The amount of normal detected when it isincorrectly identified.
False positive (FP): The amount of attack detected when it iscorrectly rejected.
False negative (FN): The amount of normal detected correctly rejected.

When it isattacks can be detected byIDS itrequires high detectionrate and low false alarm rate.
The make inquiries compares accuracy, detection rate and false alarm rate, and lists the Assessmentoutcomeof various attacks.
Assessment of accuracy
Accuracy refers to the proportion of data classified an accuratetype in total data in the situation TP and TN, thus the accuracyis

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\%$$

B) Assessment of detection rate
Accuracy refers to the proportion of attack detected among allattack data in the TP, thus detection rate is

$$\text{Detection rate} = \frac{\text{TP}}{\text{TP+FN}} * 100\%$$

Measurement of false alarm rate
False alarm rate refers to the proportion that normal data is falselydetected as attack behavior FP, thusfalse alarm rate is

$$\text{False alarm rate} = \frac{\text{FP}}{\text{FP+TN}*1} * 100\%$$

D) Accuracy measurement between different attacks: Accuracy of various attacks refers to the quantity that the type of data is corrected classified.
There are four types comparedin the research. They are Probe, Dos, U2R, and R2L [6].

Dos attack: When the proportion of normal data is low, FHRVM is better.

U2R attack: IntegrallyFHRVM is better than C4.5 and SVM.
R2L attack: In proportion to the average among two methodsare related in accuracy. [15] In proportion to the average value, except that these two methodsare similar in accuracy in R2L attack, FHRVM is greater to C4.5 and SVMin accuracy otherwise.Evaluate to the average results got in the research is compare with theresults obtained through KDD Cup 99 winner.

V. CONCLUSION:

The probe compares accuracy, detection rate, false alarm rateand accuracy of other attacks below different proportion of normalinformation. KDD Cup 99 dataset is current standard dataset inintrusion detection. Compared with further interrelated works in data mining-based intrusion detectors accuracy, detection rate, false alarm rate. It moreover show improved act than KDD Winner, particularly used for two types of attacks namely, U2R type and R2L type Comparison results of C4.5, SVM. find that C4.5 is better-quality to SVM in accuracy and detection; in accuracy for Probe,Dos and U2R attacks, C4.5 is also better than SVM and FHRVM; but in falsealarm rate FHRVM is better. In this paper enhance that FHRVM is better than c4.5 and SVM for U2R attack & R2L attack.As a future work, numerous training algorithms are utilized to enhance its performance.

REFERENCES

[1] James P. Anderson, “Computer Security Threat Monitoring and Surveillance,” Technical report, James P. Anderson Co., Fort Washington, Pennsylvania. April 1980.

[2] Tomas Abraham, “IDDM: INTRUSION Detection using Data Mining Techniques”, Technical report DTSO

[3] MahbodTavallaee, IbrahimBagheri, Wei Lu, and Ali A. Ghorbanifar A Detailed Analysis of the KDD CUP 99 Data Set proceeding of the 2009 IEEE symposium on computational Intelligence in security and defense application .

[4] Xiao a Wang, Zhaohui Shi, Chongqing Wu and Wei Wang. An Improved Algorithm for Decision-Tree-Based SVM.IEEE-2006.

[5] Pang-Ming Tan, Michael Steinbach, Vidin Kumar. Introduction to data mining.Pearson Education.

[6] Liu, Y., Wang, Z., Fang, Y., &GU, H. Y. (2006). Dos intrusion detection based onincremental learning with support vector machines. Computer Engineering,
32(4), 179–186.

[7] P.Rutravigneshwaran “Intrusion Detection using Neutrosophic classifier” published in IJST Vol.2 Issue.13, Dec 2014.

[8] Peddabachigari, S., Abraham, A., Groans, C., & Thomas, J. (2007). Modeling intrusiondetection system using hybrid intelligent systems. Journal of Network andComputer Applications, 30(1), 114–132

[9] Nadiammai, Hemalatha, "Perspective analysis of machine learning algorithms for detecting network intrusions" 2012 Third International Conference onComputing Communication & Networking Technologies (ICCCNT), 2012 , Page(s): 1 - 7.

[10] Natesan,Rajesh, "Cascaded classifier approach based on Adaboost to increase detection rate of rare network attack categories" 2012 International Conference on Recent Trends In Information Technology (ICRTIT),2012 , Page(s): 417 - 422.

[11] Landwehr, Bull, McDermott, and Choi, “A taxonomy of computer program security flaws,” ACM Comput. Surv., vol. 26, no. 3, pp. 211–254, 1994.

[12] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Ocotber 2007.

13] Lippmann, Fried, Graf, Haines, Kendall, McClung, Weber, Webster, Wyschogrod, Cunningham, Zissman, “Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation,” discex, vol. 02,p. 1012, 2000.

[14] Stolfo, Fan, Lee, Prodromidis, Chan, “Costbased modeling for fraud and intrusion detection: Results from the jam project,” discex, vol. 02, p. 1130, 2000.

[15] Reda Elbasiony, Elsayed Sallam, Tarek Eltobely, Mahmoud Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means", Ain Shams Engineering Journal, Elsevier, 2013 4, 753–762.

Author Profiles

P.Rutravigneshwaran is working as an Assistant Professor in the Department of Computer Application, Rev. Jacob memorial Christian college, Dindigul, Tamilnadu, India and his research interests include Data Mining and Networking. He has published 08 research papers in reputed international journals and conferences and it’s also available online. His main research work focuses on Data mining Algorithms, Data Security.