International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 10, Issue 4, April 2021)

Anomaly Based Network Intrusion Detection System

Anuj kumar Gautam[1], Nitin Choudhary[2]
[1,2]*CSE, Kopal Institute of Science & Technology, Bhopal, India*

***Abstract*— As malware attacks are increasing rapidly in numbers and severity over the past few years, intrusion detection system (IDS) is becoming a critical component to secure the network. Due to escalating counts of safekeeping audits dynamic properties of intruders behaviors enhancing performance of IDS becomes a vital problem that is receiving maximum attention from the research community. Intrusion poses a security risk in a close network environment. In this proposed work, a new intrusion detection method based on Principal Component Analysis (PCA) and Random Forest with low overhead and high efficiency is presented. System call data and command sequences data works as information sources to validate the proposed method. The recurrence of individual system calls and commands in a data block are computed and then data column vectors which represent the traces and blocks of the data are formed as data input. Principal Component Analysis is actually used to lower down the high dimensional data vectors and distance between a vector and its projection onto the subspace reduced for anomaly detection. Experimental results show that the proposed method is promising in terms of detection accuracy, approximity, computational expense and implementation for real-time intrusion detection.**

***Keywords*— CBR, CART, IDS, NSL-KDD, PCA.**

I. INTRODUCTION

The field of network intrusion detection has received increasing attention in recent years. One reason for this is the high growth of the Internet and the large number of networked systems that exist in all types of organizations. The increase in the number of connected machines has lead to an increase in unauthorized activity, not only from external attackers, but also from internal attackers, such as disgruntled employees and normal folks abusing their privileges for self gain. Security is a big issue for all networks in current environment. Malwares have done many more successful attempts to bring down large company networks and web services. Sevral techniques are implemented to secure the network infrastructure and successful communication over the Internet, among them the use of firewalls, encryption, and virtual private networks.

Network intrusion detection is an advancement to such techniques. Network intrusion detection methods started appearing in the last few years.

Using network intrusion detection methods, you can collect and use information from known types of attacks and find out if someone is trying to attack your network or particular hosts. The information collected in this way can be used to harden your network security, as well as for legal purposes.

One of the main threat in the security management of high-speed networks is the detection of abnormality in network traffic.

A secure network must provide the following:

- *Confidentiality:* Data transferred through the network should be accessible only to authorized.
- *Integrity:* Integrity of data should be maintain from the moment they are transmitted to the moment they are actually received. No corruption or loss of data is accepted either from random events or malicious activity.
- *Availability:* The network should be supple to Denial of Service attacks.

II. BACKGROUND OF INTRUSION DETECTION SYSTEM

Network intrusion detection is the continuous process of keeping an eye on all the events occurring in the network and analyzing them for those of upcoming incidents. Although there are many malicious activities in nature, but not all are malicious. for example, a person mistakenly typed the address of a computer and fortuitously attempt to connect to a different system without authorization. It is the process of diagnosing an individuals those who are using computer network resources without authorization to prevent authorized users from accessing network resources. Intruders can easily attack the systems through internet or from inside the specified computer network system. This highlights the two different types of network systems; host based intrusion detection system and network based intrusion detection system. A security system that is proficient of detecting inside abuses in the computer network is called as a host based intrusion detection system. A network based intrusion detection system is competent of recognize unauthorized uses or attempts of the computer network from outside the system.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 10, Issue 4, April 2021)

There are several forms of network intrusions:

- *Denial-of-service Attack* - This is a serious type of attack that results to a damage of worth million of dollars over past few years. While a significant problem, Denial-of-service attacks are usually quite simple. They typically associate an attacker disabling or rendering inaccessible a network-based information resource.
- *Guessing rlogin Attack* - In this kind of attack intruders tries to find out the password that provides security to the computer network to gain access.
- *Scanning Attacks* - The intruder perform scanning through different ports of the suffered system to find some weak points from where they can launch other attacks.

Ideally, IDS should have an attack detection rate of 100% along with false positive (FP) of 0%. Nonetheless, in practice this is hard to achieve. The most vital parameters involved in the performance estimation of intrusion detection system are shown in Table 2.1

Table 2.1:
Different Types of attacks in NSL-KDDCup Dataset

Attack Classes	Attacks
Denial of Service (DoS)	POD, LAND, TEARDROP, BACK, SMURF, NEPTUNE
Remote to User (R2L)	GUESS, FTP_WRITE _PASSWD, MULTIHOP, IMAP, SPY, PHF,
WAREZCLIENT, WAREZMASTER	
User to Root (U2R)	BUFFER_OVERFLOW, PERL, LOADMODULE, ROOTKIT
Probing	IPSWEEP, NMAP, PORTSWEEP, SATAN

Table 2.2:
Parameters for performance estimation of intrusion detection system

Parameters	Definition
True Positive or Detection Rate	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

Detection rate (DR) and false positive (FP) are used to estimate the performance of intrusion detection system [17], which is given as bellow:

DR = Total Attacks Detected/ Total Attacks
FP = Total misclassified process/ Total normal process

III. Common Detection Methodologies

The primary classes of detection methodologies are:

a. Signature or Misuse based detection

Signature-based detection is the easiest detection technique because it just matches the recent unit of activity, such as a packets or a login entries, to a file of signatures using string comparison action that are going to be performed. Signature-based detection technologies have understanding of various network or application protocols and cannot trail and recognize the state of complex communications

b. Anomaly detection

An IDS using anomaly-based detection has profiles that symbolize the normal activities of such things as users, hosts, set-up connections, or applications. The profiles are urbanized by monitoring the individuality of typical activity over a period of moment. A network strength reflects the Web activity comprises an average of 13% of bandwidth at the Internet border during typical workday hours. The IDS then uses arithmetical methods to compare the characteristics of current activity to thresholds correlated to the profile.

IV. Review Of Literature

Bayesian reasoning is considered here a general phrase for a range of techniques that exploit Bayes theorem to deal with uncertainty. In short, Mitchell [20] provides the following definition "Bayesian reasoning provides a probabilistic approach to inference. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data."

In recent years, Bayesian networks have been utilized in the decision process of hybrid systems [11]. Kruegel et al. [11] argue that most hybrid systems obtain high false alarm rates due to simplistic approaches to combining the outputs of the techniques in the decision phase.

Naïve bayes is a simplified version of Bayesian networks, which offer machine learning capabilities. Naïve bayes does assume that all the features in the data are independent of each other [20]. Nevertheless, Naïve bayes (utilized as a classifier) has been successfully applied to network based intrusion detection by several researchers.

Ben Amor et al. [21] conducted an empirical investigation on the KDD Cup '99 data set, comparing the performance of Naïve bayes and a Decision Tree.

Gharibian and Ghorbani [23] compare the performance of two probabilistic techniques, Naïve bayes and a Gaussian classifier, and two predictive techniques, a decision tree and random forest. They analyse the performance of the techniques on three different training sets of the 10% KDD Cup '99 data set (all tested on the original test set).

One of the main benefits of clustering is unsupervised learning

There are several applications of clustering techniques to network based anomaly detection [12]. These studies make two common assumptions about the data: (1) that the vast majority of the data is normal and (2) that the intrusions are statistically different from normal data [2].

Portnoy et al. [24] propose a version of single linkage clustering, which only takes one pass of the data to create the clusters. The algorithms start with no clusters. For each instance in the training set, the Euclidean distance to existing clusters is calculated to determine the closest cluster (if any). If this distance is within a predefined threshold, the instance is assigned to that cluster. Otherwise, a new cluster is created with the instance as its centroid. Thereafter, according to assumption one, above, the largest clusters are labeled as 'normal'. The remaining, small, clusters are labeled as 'intrusion'. This obtained true positive rates of approximately 50% with 2% false positives.

A supervised clustering and classification technique has been proposed by Ye and Li [10], which aims to learn both normal and intrusive behavior. Initially, two clusters are created, one for normal data and one for intrusive data.

Spinosa et al. [28] propose utilising both supervised and unsupervised clustering for network based intrusion detection. First, they perform supervised learning of normal traffic.

V. Implementation & Result Analysis

Network intrusion detection is the mechanism of monitoring the events taking place in a computer system or network and doing their analysis whether they have any intrusion or not [1]. Not similar to misuse detection, which rings an alarm when a known attack signature is matched, anomaly detection identifies activities that deviate from the normal behavior of the monitored system and consequently detects the possibility of any attack [14]. This work proposes to design anomaly-based intrusion feasible enough.

NIDS Framework Based On PCA Via Random Forest

This proposed work intends to filter out unnecessary information and considerably decrease number of computer resources, both memory and CPU time required to detect attacks. PCA (principal component analysis) transform is deployed to shrink the features and trained random forest is used to identify kinds of new attacks. Test and comparison are done on NSL-KDD dataset. It is a latest enhancement of KDDcup99 and has a few benefits over KDDcup99, the experiments with NSL-KDD data demonstrate that our proposed model gives better and robust representation of data as it was capable to shrink features showing 80.4% data reduction, around 40% reduction in training time and 70% drop in testing time is achieved. Our proposed method not only reduces the number of the input features and time but also do not effect classification detection rate.

The result indicates the superiority of algorithm.

In our experiment, we used NSL-KDD data set. It has solved some of the inherent problems of the KDDCup'99[5]. It is considered as standard benchmark for network intrusion detection evaluation [8]. The instruction dataset of NSL-KDD related to KDDCup'99 consist of roughly 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or attack type ,with exactly one specific attack type . Empirical analysis shows that technique of reducing features feature can also reduce the size of dataset. The time and space complexities of most classifiers used are exponential task of their key vector size [15].

The process of Normalization is utilized for data preprocessing, where the characteristics of data are scaled to remain within a small particular range such as -1.0 to 1.0 or 0.0 to 1.0. If using neural network back propagation algorithm for classification, normalizing the input values for every attribute measured in the training samples will facilitate the learning phase.

Since this algorithm is designed to be common, it must be capable of creating clusters given a dataset from an arbitrary distribution.

Feature diminution applies a mapping of the multidimensional gap into a space of lower dimensions [12]. Feature extraction comprises of construction of features, space dimensionality reduction, sparse representations, and feature selection. The feature space having decreased dimensionalities Contributes in the real sense to classification that cuts pre-processing costs and minimizes the effects of the 'peaking phenomenon' in classification [13]. Thereby improving the overall performance of classifier based on network intrusion detection systems.

The algorithm named Principal component Analysis (PCA) is deployed for dimensionality reduction. PCA is a mathematical technique that changes a number of possibly correlated variables into a new set of uncorrelated variables called principal components. The first principal component stands for the highest variability in the dataset. Therefore, the remaining can be overlooked with minimal loss of the information value during the dimension reduction process.

The various results of our experiments are given above. we show the results of our experiment by taking different input values such as in table I. we have taken 9642 test sample and 11850 training samples, 200 steps to train the network, 25 hidden layers, learning rate 0.5, momentum is 1, results of table 5.3.2 shows the classification in 5 classes (Normal, DOS , Probe ,U2R , R2L). We can see from this table the accuracy achieved before and after the features reduction, the training time is reduced by 40%, testing time is reduced to 78.5 %.

Table I
Correctly And Incorrectly Classified Instances Of Different Nsl Dataset

Dataset Name	Classified Instances (In Number)		Classified Instances (In %)		Total Number Of Instances	No Of Attributes
	Correctly	Incorrectly	Correctly	Incorrectly		
KDD TEST+ WITH PCA	21914	630	97.20	2.79	22544	6
KDD TEST+ WITHOUT PCA	22233	311	98.62	1.37	22544	42
KDD TEST-21 WITH PCA	9675	2175	81.64	18.35	11850	2
KDD TEST-21 WITHOUT PCA	11553	297	97.49	2.50	11850	42
KDD TRAIN+ WITH PCA	125596	377	99.70	0.30	125973	11
KDD TRAIN+ WITHOUT PCA	125661	270	99.75	0.21	125973	42
KDD TRAIN+_20 PERSENT WITH PCA	24975	217	99.13	0.86	25192	5
KDD TRAIN+_20 PERSENT WITHOUT PCA	25148	64	99.74	0.25	25192	42

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 10, Issue 4, April 2021)

A "Confusion Matrix" is sometimes used to represent the result of testing, as shown in Table II. The benefit of via this template is that it not only tells us how numerous get misclassified but as well what misclassifications occurred. It has been classified as normal and anomaly based. As discussed above we are using four NSL-KDDCup data set KDDTest+, KDDTest_20, KDDTrain+, KDDTrain+_20. A confusion matrix is generated when PCA via Random Forest is been applied on the given four datasets.

Table II
Confusion Matrix

DATASET NAME	CLASSIFIED AS□	A=NORMAL	B=ANOMALY
KDD TEST+ WITH PCA	A= NORMAL	9389	322
	B=ANOMALY	308	12525
KDD TEST+ WITHOUT PCA	A= NORMAL	9564	147
	B=ANOMALY	16	12669
KDD TEST-21 WITH PCA	A= NORMAL	1067	1085
	B=ANOMALY	1090	8608
KDD TEST-21 WITHOUT PCA	A= NORMAL	2004	148
	B=ANOMALY	149	9549
KDD TRAIN+ WITH PCA	A= NORMAL	67157	186
	B=ANOMALY	191	58439
KDD TRAIN+ WITHOUT PCA	A= NORMAL	67203	117
	B=ANOMALY	153	58458
KDD TRAIN+_20 PERSENT WITH PCA	A= NORMAL	13373	76
	B=ANOMALY	141	11602
KDD TRAIN+_20 PERSENT WITHOUT PCA	A= NORMAL	13440	9
	B=ANOMALY	55	11688

VI. Conclusion And Future Work

Our research work based on network intrusion detection system, we found that Most of the existing IDs use all 41 features in the network to estimate and seems for intrusive guide some of these features are surplus and extraneous. The drawbacks of this system is unbearable time consuming and undignified detection process which effects the performance of ID system.

To decipher this difficulty we proposed an algorithm based on PCA (Principal Component Analysis) and RF (Random Forest) that uses key Component Analysis as a Features reduction algorithm. The goal of PCA is to reduce the dimensionality of the data while retaining as much as probable for the dissimilarity present in the original dataset and trained artificial neural network to identify any kind of new attacks .Tests and comparison are done on NSL-KDD CUP dataset. The test data contains 4 kinds of unusual attacks in totaling to standard system call.

Our investigational results showed that the proposed model gives better and robust illustration of data as it was able to reduce features resulting in a 80.4% data reduction and just about 35%-40% reduction in testing time and 75%-80% reduction in testing time ,classification accurateness achieved in detecting new attacks. Meantime it is drastically reducing in numbers, both memory and CPU time, mandatory to detect an attack. This shows that our proposed algorithm is trustworthy and reliable in network intrusion detection.

Currently in some cases the detection rate reduces when we apply the dimension reduction techniques. In future, we will continue on our research of improving detection performance of both normal and malicious activities.

REFERENCES

[1] Power, R. (2002), Computer Security Issues & Trends. Vol. 8, No. 1, 2002. pg4.

[2] Denning D E (1987), An Intrusion-Detection Model, In IEEE Transaction on Software Engineering, Vol. Se-13, No. 2, February 1987, 222-232.

[3] Lee, W, Stolfo S and Mok K (2000), Adaptive Intrusion Detection: A Data Mining Approach, In Artificial Intelligence Review, Kluwer Academic Publishers, 14(6): 533 - 567, December 2000.

[4] Satinder Singh, Guljeet Kaur (2007), Unsupervised Anomaly Detection In Network Intrusion Detection Using Clusters, Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007) RIMT-IET, Mandi Gobindgarh. March 23, 2007.

[5] Eric Bloedorn , Alan D. Christiansen , William Hill , Clement Skorupka , Lisa M. Talbot , Jonathan Tivel (2001), Data Mining for Network Intrusion Detection: How to Get Started, CiteSeer, 2001

[6] L. Portnoy (2000) Intrusion Detection with Unlabeled Data Using Clustering, Undergraduate Thesis, Columbia University, 2000.

[7] Theodoros Lappas and Konstantinos Pelechrinis(2006), Data Mining Techniques for(Network)IntrusionDetectionSystems, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.2533&rep=rep1&type=pdf.

[8] Dewan Md. Farid, Nouria Harbi, Suman Ahmmed, Md. Zahidur Rahman, and Chowdhury Mofizur Rahman (2010), Mining Network Data for Intrusion Detectionthrough Naïve Bayesian with Clustering , World Academy of Science, Engineering and Technology 66 2010

[9] The KDD Archive. KDD99 cup dataset, 1999.http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[10] X. Li and N. Ye. 2005. A supervised clustering algorithm for computer intrusion detection. Knowledge and Information Systems, 8, 498-509. ISSN 0219-1377.

[11] Kruegel C., Mutz D., Robertson W., Valeur F. Bayesian event classification for intrusion detection. In: Proceedings of the 19th Annual Computer Security Applications Conference; 2003.

[12] Portnoy L., Eskin E., Stolfo S.J. Intrusion detection with unlabeled data using clustering. In: Proceedings of The ACM Workshop on Data Mining Applied to Security; 2001.

[13] Paxson V., “Bro: A System for Detecting Network Intruders in Real-Time”, Computer Networks, 31(23-24), pp. 2435-2463, 14 Dec. 1999.

[14] D.Barbara, J.Couto, S.Jajodia, and N.Wu, "ADAM: A test bed for exploring the use of data mining in intrusion detection”, SIGMOD, vol30, no.4, pp 15-24, 2001.

[15] P.Domingos, and M.J. Pizzani, "On the optimality of the simple Bayesian classifier under zero-one loss”, m/c learning, Vol.29, no2-3, pp 103-130, 1997.

[16] F. Provost, and T. Fawcett, “Robust classification for imprecise environment,” Machine Learning, vol. 42/3, 2001, pp. 203-231.

[17] Athanasios Papoulis and S. Unnikrishna Pillai, 2002. "Probability, Random Variables and stochastic Processes ", Book

[18] P. Kabiri and A.A. Ghorbani. September 2005. Research on Intrusion Detection and Response: A Survey. International Journal of Network Security, 1, 84-102.

[19] A. Patcha and J-M. Park. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Comput. Netw., 51, 3448-3470. ISSN 1389-1286.

[20] T.M. Mitchell. Machine Learning. McGraw-Hill, 1997. ISBN: 0-07-115467-1.