Intrusion Detection System Using PCA And Random Forest

[1]Nishant J Seth, [2]Trish Bandhekar, [3]Sonu Yadav

[1,2,3]Department of Information Technology, Fr Conceicao Rodrigues College of Engineering, Mumbai, India.

Abstract: **As technology has been evolving at a rapid speed, it has led to more vulnerable data and that has increased the number of unauthorized access attempts. To overcome this, an Intrusion Detection System is used. An IDS is used to detect any such malicious attempts. Current generation of IDS' cannot detect complex attacks and take way too long to detect when using high dimensional data. Other problems of IDS include the high false alarm and low detection. To solve these drawbacks, we propose a system that uses machine learning based technique to identify these malicious packets in a low amount of time. We use a technique known as Principal Component Analysis (PCA) to reduce our high dimensional dataset into a lower dimensional dataset while still keeping our accuracy up without too much loss of data. Random Forest is used as a classification algorithm to detect our packets. An accuracy of 0.996 has been obtained. This experiment was conducted on the UNSW-NB15 dataset.**

***Keywords* - Intrusion Detection System (IDS), Principal Component Analysis (PCA), Random Forest, Classification.**

I. INTRODUCTION

Computer and network security has gained importance as there has been an increase in the number of attacks targeting the confidentiality, the integrity, and the availability of the data. Intrusions are targeting an individual's or an organization's network to steal their valuable data. Many schemes and efforts have been done to detect the intrusions to the data. Intrusion detection systems are one of them which aim is to detect intrusions [1].Intrusion detection systems are classified into two categories, Network based intrusion detection system (NIDS) and Host based intrusion detection system (HIDS), which are based on the data source [2]. The data source is used to accommodate the audit data for IDS. By analysing that audit data, the IDS triggers an alarm as it detects an intrusion or an attack. Host-based intrusion detection system (HIDS) is setup on a single system also called as a target system, which presumes apt to attack. HIDS wields system log files in order to detect any attack by analysing the changes in these log files. Since HIDS is deployed on the target system and depends on the target operating system, any shortcomings in the operating system will co-operate with attacker to dodge the HIDS. Compromising of the user/host system can cause the undermining of HIDS as well. Compromising host might compromise HIDS as well, for having some bugs in the running operating system. On the other hand, the Network-based intrusion detection system (NIDS) is deployed on the network segment in order to detect any malicious network data activity that tries to infiltrate into the network of the organization. Network based intrusion detection system (NIDS) is used for detecting any infiltrate malicious network packet by installing it at the network segment. That is why NIDS is transparent to the other systems connected to the internet. Core of the intrusion detection system is the detection method which has helped in order to detect the intrusions. Two types of intrusion detection methods are signature based detection method and anomaly based detection method and they are used in the intrusion detection systems [3].

Nowadays, the resulting data of any organization has become its most precious asset on every scale and everything an organization does involves using that data in some way or another. Methods of conducting business have changed over time as the world has become more fiercely connected, and the increase in this connectivity has provided an access to the varied resources of data; moreover, it has provided an access path to the data from virtually anywhere in the network. The internet connectivity is no longer an option for most organizations; on the other hand, consequently, running a business environment that is secure, is one of the primary concerns of an IT department that exist in an organization.

II. LITERATURE SURVEY

There are three approaches currently that can be used for Intrusion Detection:

Machine Learning Approach:- Machine learning is the study of algorithms that improvise and improve their performance with experience by learning and are meant to computerize exercises; the machine follows necessary steps consummately furthermore in an organized way. It is a type of artificial intelligence that provides computers with the ability to learn without being programmed. This Paper covers different prediction techniques that are utilized for examination, which include Fuzzy Logic [4], K-nearest Neighbour (k-NN) [5], Support Vector Machine [6], Decision Trees [7] and K-means Clustering [8].

Data Mining Approach:- Data Mining based Intrusion Detection System IDS techniques generally falls into one of the following two categories; misuse detection and anomaly detection techniques. In the misuse detection technique, each instance in a data set has been labelled as 'normal' or 'intrusion' and the learning algorithm is usually trained over the labelled data. Anomaly detection is then used for building models of the normal behaviour, and automatically detects any infrequent deviation from that behaviour, flagging the latter one as a suspect.

Classification: is used in mapping a data item into one of the several predefined categories.
Link analysis: determines the relation between the fields in the database. Furthermore, sequence analysis models all the necessary sequential patterns too. Now, Misuse Detection [9], Anomaly detection [10], Classification model with association rules algorithm [11], Link Analysis, Sequence analysis [11] are the examinable techniques for the data mining approach.

Statistical Model Approach:- Statistical approaches are based on modelling the data on universal statistical properties and by using this information, it's able to estimate whether the test samples come from the same distribution of data or different distributions of data. The techniques that have been tested differ in terms of their complexity. Generalized Anomaly and Fault Threshold system [12], The Kolmogorov-Smirnov Test [13], Clustering analysis [14] are the various prediction techniques utilized for examination.

III. Methodology

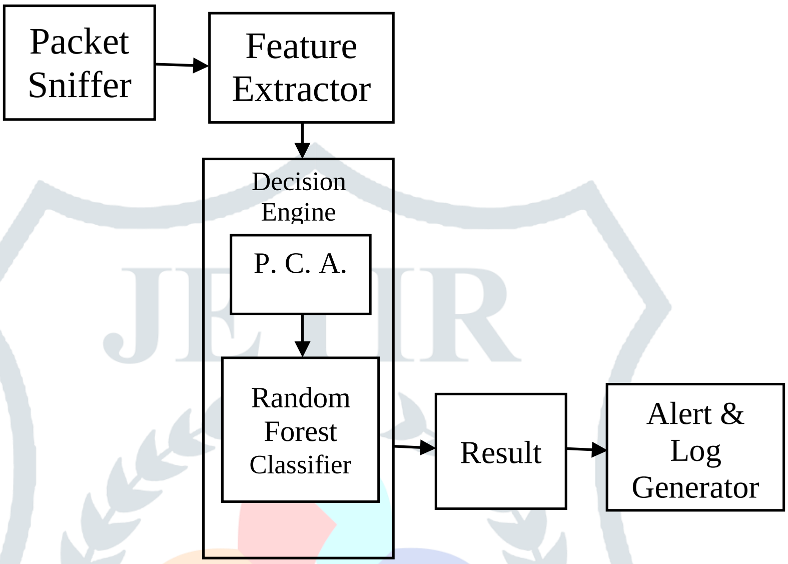

Figure (1) The System Flow Design

Principal Component Analysis PCA is a feature extracting technique that is used to generate new features which are a linear combination of the initial features. It is the first step in the decision engine as seen in fig. (1). PCA maps each instance of an experimental dataset present in a d-dimensional space to a k-dimensional subspace such that, k d. The set of k new dimensions are generated and are called the Principal Components (PC). Each principal component is directed towards a maximum variance despite the variance already accounted for in all its preceding components. Subsequently, the first component covers the maximum variance and each component that follows it covers a lesser value of that variance [15]. Some basic knowledge of PCA is briefly described in the next.

Assume variable $\{x_t\}$ where t = 1, 2... N are stochastic n-dimensional input data recorded with a mean (μ). It is defined by the following Equation:

$$\mu = \frac{1}{N}\sum_{t=1}^{N} x_t \qquad (1)$$

The covariance matrix of x_t is defined by the following:

$$C = \frac{1}{N}\sum_{t=1}^{N}(x_t - \mu).(x_t - \mu)^T \qquad (2)$$

PCA solves the following eigenvalue problem of covariance matrix C as:

$$Cv_i = \lambda_i v_i \qquad (3)$$

Where λ_i (i = 1, 2... n) are the eigenvalues and v_i(i = 1, 2... n) are the corresponding eigenvectors.

To represent data records with low dimensional vectors, we only need to compute the m eigenvectors (called principal directions) corresponding to those m largest eigenvalues (m<n). It is also well known that the variance of the projections of the input data onto the principal direction is greater than that of any other directions. Let,

$$\varphi = [\nu_1, \nu_2, \ldots\ldots, \nu_m], \Lambda = \mathrm{diag}[\lambda_1, \lambda_2, \ldots\ldots, \lambda_m] \qquad (4)$$

Then,

$$C\Phi = \Phi\Lambda \qquad (5)$$

The parameter ν denote to the approximation precision of the m largest eigenvectors so that the following relation holds.

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq \nu \qquad (6)$$

Based on (5) and (6) the number of eigenvectors can be selected and given a precision parameter v, the low dimensional feature vector of a new input data x is determined by,

$$x_f = \Phi^T \mathrm{x}$$

Random forest algorithm is applied for the purpose of building multiple decision trees and merging them together in order to get an accurate and a stable prediction. The dimensional data from the PCA is moved to the Random Forest Classifier in the decision engine as seen from figure (1). A very big advantage of the random forest is, that it can be used for both classification and regression problems simultaneously. Random Forest also has nearly the same hyper parameters as in a decision tree or a bagging classifier.

Random Forest includes an additional randomness to the model, while growing the trees. While it searches for the most important feature, it splits the node and searches for the best feature among the random subset of features. This results in a varied diversity that generally results in a better functioning model.

IV. Dataset

UNSW-NB15 is the dataset that is used to identify the significant features and reduce the number of features to a smaller number in the UNSW-NB15 dataset. Therefore, a subset of the significant features in detecting intrusion can be proposed by using machine learning techniques. These features can be then used in the design of the Intrusion Detection Systems (IDS) that are working towards automation of the anomaly detection with a less overhead.

The UNSW-NB15 dataset [16] was published in the year 2015 which includes nine different modern attack types (compared to the 14 attack types in KDD'99 dataset) and the wide varieties of real normal activities as well as 49 features inclusive of the class label consisting a total of 2,540,044 records. These features are then categorised into six groups called the Flow Features, Basic Features, Content Features, Time Features, Additional Generated Features and Labelled Features. The Additional Generated Features are further categorised into two subgroups called General Purpose Features and Connection Features. Features numbering from 36-40 are known as General Purpose Features. Features numbering from 41-47 are known as connection features. Further, the attacks of the UNSW-NB15 dataset are categorised into 9 types known as the Reconnaissance, Shellcode, Exploit, Fuzzers, Worm, DoS, Backdoor, Analysis and Generic. The UNSW-NB15 dataset has been divided into two Training datasets (#82, 332 records) and a Testing dataset (#175, 341 records) including all attack types and the normal traffic records. Both the Training and the Testing datasets have 45 features each respectively. It is important to note that the first feature (i.e. id) was not mentioned in the full UNSW-NB15 dataset features and also the features scrip, sport, d-stip, s-time and l-time have been missing in the Training and Testing dataset.

V. Experimental Analysis

Figure (1) shows the overall description of the proposed intrusion detection system. We first analysed the dataset i.e. UNSW-NB15. We now perform Principal Component Analysis on this modified dataset. After giving it a multiple different component values, we find that reducing the dataset to 14 principal components gives us a high accuracy without much loss of data.

We then create train and test splits on this modified dataset and then pass it to our classifier

This training dataset is then passed onto our classifier. Our classifier of choice is Random Forest. It aims to make the trees de correlated and prune the trees by setting a stopping criterion on the nodes; we could either use 'gini' or 'entropy'. We chose the latter as we wanted information gain.

Using train test split, we get an accuracy of 0.9961 from our classifier.

VI. Results and Discussion

Various Machine Learning algorithms were tried but Random Forest gave a very high accuracy. We use the train and test split in this experiment. Dataset provides data for following attacks: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. We get an accuracy of 0.9961 witha False Positive Rate(FPR) equal to 0.0523. The whole experiment was conducted on a machine with the following specifications: Processor : Intel(r) Core(TM) i5 8250U CPU @ 1.60 GHz, Installed Memory 8 GB (RAM) and 64 bit System type. The classification is done in Jupyter Notebook along with sk-learn library.

VII. Conclusion and Future Scope

It is observed that along with PCA, the time taken to detect an anomaly is significantly less than when done without PCA. Furthermore, Tree algorithms outperform in terms of speed and accuracy when compared to other algorithms. Lastly, a self-adaptive IDS with the above given approach would be implemented which will incorporate new records dynamically which will increase the accuracy further and reduce the training time.

References

[1] S. Willium, “*Network Security and Communication*”, IEEE Transaction, Vol.31, Issue.4, pp.123-141, 2012.

[2] R. Solanki, “*Principle of Data Mining*”, McGraw-Hill Publication, India, pp. 386-398, 1998.

[3] M. Mohammad, “*Performance Impact of Addressing Modes on Encryption Algorithms*”, In the Proceedings of the 2001 IEEE International Conference on Computer Design (ICCD 2001), Indore, USA, pp.542-545, 2001.

[4] A. A. Aburommanand M. B -I. Reaz, “Evolution of Intrusion Detection System Based on Machine Learning Methods”, Australian Journal of Basic and Applied Sciences, 7(7): 799-8 13, 2013.

[5] S. Manocha, and M. A. Girolami, “An empirical analysis of the probabilistic K-nearest neighbour classifier,” Pattern Recognition Letters, 28, 1818- 1824.2007.

[6] S. J. Horng, M. Y. Su, Y. H. Chen, T. W. Kao, R. J. Chen, J. L. Lai and C. D. Kara, "A novel intrusion detection system based on hierarchical clustering and support vector machines," Expert Systems with Applications, 38(1): 306-313. 2011.

[7] S. Peddabachigari, A.Abraham, C.Gransen and J.
Thomas, "Modeling intrusion detection system using hybrid intelligent systems." Journal of Network and Computer Applications, 30(1) : 114-132, 2007.

[8] A. Ahmad and L. Dey, “A k-mean clustering algorithm for mixed numeric and categorical data,” Data & Knowledge Engineering, vol. 63, no. 2, pp. 503–527, 2007.

[9] D.E. Denning, An Intrusion Detection Model, IEEE Transactions on Software Engineering, SE-13:222-232, 1987.

[10] H.S. Javitz, and A. Valdes, The NIDES Statistical Component: Description and Justification, Technical Report, Computer Science Laboratory, SRI International, 1993

[11] W. Lee and S. Stolfo. Data mining approaches for intrusion detection. In Proceedings of the 7th USENIXSecurity Symposium, San Antonio, TX, 1998.

[12] C. Manikopoulos et al., “Generalized Anomaly Detection in Next Generation Internet: Architecture and Evaluation,” submitted for publication, 2002.

[13] B. D. Joao et al., “Statistical Traffic Modeling for Network Intrusion Detection,” Proc. 8th Int’l. Symp. Modeling, Analysis Sim. Comp. Telecommun. Sys., Aug. 2000, pp. 466–73

[14] Verwoerd, Theuns, and Ray Hunt. "Intrusion detection techniques and approaches." *Computer communications* 25, no. 15 (2002): 1356-1365.

[15] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2, no. 1-3 (1987): 37-52.

[16] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." In *2015 military communications and information systems conference (MilCIS)*, pp. 1-6. IEEE, 2015.