Intrusion Detection Systems based on Machine Learning

Oumaima Chentoufi[1] and Khalid Chougdali[2]
[1] National School of Applied Sciences of Kenitra, Ibn Tofail University, Kenitra, Morocco
[2]Department of Computing, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

Keywords: Intrusion detection system, machine learning, classifiers, Principal Component analyses.

Abstract: This paper contains an introduction to intrusion detection systems known as IDS. There are two types of techniques to detect an intrusion, misuse detection and anomaly detection; both can be used in a complementary way to increase the system's efficiency is used for EMERALD, JiNao... It was determined that using machine learning for IDS is an efficient way to detect attacks, and this paper will provide information about machine learning and its classifiers.

1 INTRODUCTION

Over the years, the number of attacks on different networks has increased exponentially, and even though companies do their best to protect their network, the intruders always find a loophole, a fault in the system that they analyse and use to their advantage in order to attack a system. The constant growth in technology leads to the appearance of new and different security techniques, and that is done either by encrypting the data, using firewalls, or obtaining an Intrusion Detection System, or many other techniques to protect the network.

Even though the word "intrusion" is typically used to describe an attack that has been successful from the victims' prospect, an IDS detects attacks no matter their status, and companies usually use detection systems not only to protect their network but also to be informed about the status of the network in order to correct its vulnerabilities. The primary purpose of the intrusion detection systems is to detect if an attack has taken place, whether it is successful or not, and this system can be in the form of a software or hardware device; it can also be both of them. We should make the difference between an IDS and IPS, the first one being a passive system that analyses and locates abnormal activities in our system, the second one (Intrusion Prevention Systems) is an active system that analyses the activities, anticipates and prevents the attack based on the configuration. Intrusion detection can be manual, which means using a resource to analyse every activity in our network or automatic. After detecting an attack, the IDS must save all the information to help later while generating an alert.

We can identify multiple types of IDS, there are Network-based IDS (NIDS) where the IDS interpret and analyse the packets circulating in the network, Host-based IDS (HIDS) where the IDS analyse exclusively the information concerning this host, Network-Node IDS (NNIDS), and this type of IDS works like NIDS except that it concerns only the packets intended for a node of the network. There are also Application-based IDS (ABIDS) where the IDS controls the interactions between the user and a program. Finally, the Hybrid IDS (hybrid intrusion detection system) is a system that analyses information coming from the machines and the network; combining both allows us to have better attack detection.

The purpose of this paper is to explain how we can make intrusion detection systems more efficient. In order to do that, we must find a solution to reduce the dimensionality of the data while still having the most relevant features.

2 STATE OF THE ART

Intrusion Detection System (IDS) is one of many efficient ways to protect a network; they are based on misuse detection or anomaly detection; it can also use both to have a more efficient system. The first one is a type of IDS that contains a database of different signatures of known attacks and tries to identify an attack by comparing the signatures with

Chentoufi, O. and Chougdali, K.
Intrusion Detection Systems based on Machine Learning.
DOI: 10.5220/0010734300003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 355-359
ISBN: 978-989-758-559-3

the collected data. This database should be updated every time; if an attack signature is not in that database, we cannot identify the attack later on, and if not, the number of false negatives increases. The second one is Anomaly detection, and this type of IDS analyses the system and defines the "normal" behaviour; if there is any change in the performance, it identifies it as an attack, even if it is not, which means that the number of false positives increases.

Intrusion detection systems are composed of three principal components.

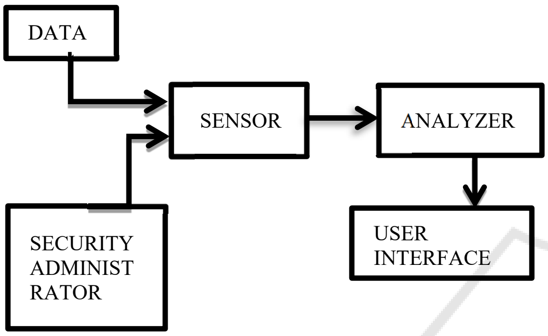

Figure 1: Schematic model of an intrusion detection system

The first component being the **SENSOR**; this part of the IDS is responsible for collecting the data, which is later on analysed by the **ANALYSER**; it receives all the collected data and defines if an attack occurred. The third component is the **USER INTERFACE**; it displays the data to determine the system's behaviour.

In addition to these components, IDS may be propped by a honeypot that stays visible to the intruder; it is like a trap that notifies the defenders of an attempt to access the honeypot by an unauthorized user. We can use it to identify and explore the flaws in our system (only if an attack occurs) while also reducing the false alerts.

The positioning of an ID system must be optimized and efficient. The IDS can be installed right before the firewall. This way, the IDS can identify all the attacks, the downside of this is that the analysis is too complicated. The IDS can also be installed in the DMZ, and this way, we can identify the attacks that were not detected by the firewall. The third option is to have it right after the firewall, and it will filter any attacks since most of them are done from the inside. Moreover, for the usage of multiple IDS, there are two approaches, the first one being the centralized approach which simplifies the implementation and the second one decentralized approach it helps to reduce the dependence on a single ID system.

After implanting an IDS, there should be some type of evaluation to identify if the system is working. We can identify two techniques of evaluating an IDS, the first one being TEST EVALUATION; this test uses not only sequential intrusions from a single attack but also simultaneous intrusions from several sessions. This test aims for three significant objectives: Stress testing, Intrusion identification, and resource use. The second technique is the ANALYTICAL EVALUATION; this evaluation allows us to master the model by defining the methods. It consists of: Classifying attacks based on the observed characteristics by the IDS, collecting and analysing the data and determining if a particular type of attack can be detected by the IDS.

We can define two types of vulnerabilities to take advantage of to create attacks, Network vulnerabilities, and application vulnerabilities, the first one is due to a fault in the protocol or its implementation; it allows the intruder to have information about the system, while the second one is due to a software fault.

We can define four general types of attacks:
DOS (denial of service): The IDS is saturated by data, to avoid this, it is necessary to filter and correctly stock the data.
R2L: Unauthorized access from a remote machine.
U2R: Unauthorized access to local superuser (root).
Probing: Port scanning.

Furthermore, like any other attack, the intruder must detect the IDS to attack it. This system can either be online, which means it detects the attack simultaneously, or offline, which means the execution is done periodically and we can only see the results.

3 MACHINE LEARNING

Machine learning includes building a model from data through an algorithm. The implementation of machine learning is based on five big steps:

1/ Obtaining the data
2/ Model realization
3/ Learning phase
4/ Validation phase
5/ Execution phase.

We can identify two types of learning, supervised and unsupervised learning. For supervised learning, the machine learns under guidance. We are giving the input data while knowing exactly how the output should be. We can identify two types of data, the first one being Numerical data which is the most used one, and the second one the categorical data which contains characters rather than numbers.

For supervised learning, we have Classification or regression, which will be later on explained. For unsupervised learning, we do not supervise the model; in other words we let the model work on its own to discover the information. It uses machine learning algorithms that conclude unlabelled data. For this type of learning, we have clustering, it finds patterns and groupings from unlabelled data.

Unsupervised learning has more difficult algorithms than supervised learning, which is logical since we know little to no information about the outcome. With unsupervised learning, we are looking to perform dimensionality reduction.

As already mentioned, classification and regression are part of supervised learning. Regression is about continuous values, mapping the input to some real number as an output.

Classification is the process of taking an input and mapping it to an output, which will be some discreet label. The main goal of classification is to identify the category or class the new data will fall under. There are six types of classification algorithms:

KNN (K-nearest neighbours), also known as lazy learners, is the laziest algorithm in machine learning; there is little or no prior knowledge about the distribution of data.

Decision Tree: Starts with a single node, which branches into possible outcomes forming additional nodes that lead into other possibilities, this gives it a tree-like shape. They can be used to map out an algorithm that predicts the best choice mathematically.

Random Forest: Uses many decision trees; each one is different from the other. When we get new data, we take the majority vote of the ensemble to get the result.

Naïve Bayes: Works on the principle of Bayes Theorem and finds the probability of an event occurring given the probability of another event that has already happened.

Support Vector Machine (SVM): The Maximal-Margin classifier is a hypothetical classifier. In SVM, a hyperplane is selected to best separate the points in the input variable space by their class.[5]

Logistic Regression: Input values are combined linearly using weights or coefficient values to predict an output different from linear regression because the output being modelled is binary instead of continuous.

We can take as an example Intrusion detection systems using a hybrid system DSSVM, meaning the use of both SVM (Support Vector Machine) and the distance sum, this approach has been discussed by Chun Guo, Yajian Zhou, Yuan Ping, Zhongkun Zhang, Guole Liu and Yixian Yang in the paper A distance sum-based hybrid method for intrusion detection. DSSVM is based on integrating two techniques; the first is used to optimize the learning performance, and the second to predict. For the implementation part, they used the KDD'99 dataset to demonstrate that the detection rate using K-NN is lower than SVM, and they are both lower than the detection rate of DSSVM. We can also quote the paper DDoS Attack Detection Based On Neural Network written by Jin Li, Yong Liu and Lin Gu, where they proposed a DDOS detection method using Learning Vector Quantisation Neural Network (LVQ NN) where they achieved a 99.732% recognition rate for host anomaly detection and compared it to BP Neural network with 89.9% recognition rate. They have set two types of results; the first category was normal and the second one attack. They identified five implementing phases: Data Set collect system, pre-processing Data Set, Determining the LVQ NN, training System, and testing system. The results were obtained by redoing the same thing ten times for both LVQ neural network and BP neural network to improve the authenticity of the results.

4 PROPOSED APPROACH

The Principal Component Analyses (PCA) is used to reduce the number of variables. The main idea of PCA is to identify patterns in a data set so it can be transformed into another data set with lower dimensions without losing any vital information; the same is done by altering the variables, also known as Principal Components (PCs) and are orthogonal. They are ordered so that the retention of variation present in the original variables decreases as we move down the order. So, in general, PCA is a tool used to reduce features and to lower dimensions while retraining most of the information and finding patterns in the data of high dimensions. PCAs' key advantages are their low voice sensibility, reduce the need for capacity and memory.

Let us consider a data set of $X= [x_1,x_2,\ldots,x_n]$, where d is the dimensionality of data and n is the number of training samples. The covariance matrix is defined as follow:

$$\sum_x = HH^T \quad (1)$$

Where H is defined by:

$$H = \frac{1}{n}[x1 - \mu, x2 - \mu, \ldots., xn - \mu] \quad (2)$$

And μ is the centroid of training data and defined by:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} xi \quad (3)$$

The covariance matrix $\sum_x$ is a symmetric matrix. If the dimensionality is large (d>> n), with n the number of training samples, then the computation of EVD (eigenvalue decomposition) of $\sum_x$ becomes a slow procedure. In this case, we can use either the SVD based PCA or QR based PCA.

SVD stands for Singular Value Decomposition, and it is a technique for dimension reduction using multiple mathematical transformations to classify the data into simpler linearly independent components and minimize the reconstruction errors. So, for the SVD based PCA, the decomposition will be applied to the matrix H.

Furthermore, based on what has been discussed by Alok Sharma, Kuldip K. Paliwal, Seiya Imoto and Satoru Miyano, on the paper Principal component analysis using QR decomposition, QR based PCA uses the rectangular matrix (where d>>n) to carry out the eigenvalue decomposition (EVD) of the covariance matrix in a numerically stable manner. The rectangular matrix H defining the covariance matrix can be decomposed into orthogonal matrix Q_1 and upper triangle matrix R_1 using QR decomposition, which will give:

$$H = Q_1R_1 \quad (4)$$

Substituting it in the first equation of the covariance matrix, they will have:

$$\sum_x = Q_1R_1R_1^TQ_1^T \quad (5)$$

Moreover, by using SVD, the matrix R_1^T will be written as follow:

$$R_1^T = U_1D_1V^T \quad (6)$$

The final equation for the covariance matrix is:

$$\sum_x = Q_1VD_1U_1^TU_1D_1V^TQ_1^T \quad (7)$$

Or $$\sum_x = Q_1VD_1^2V^TQ_1^T \quad (8)$$

Based on their implementation and simulation parts, it has been concluded that using QR based PCA is computationally more efficient and faster than the SVD based PCA.

The number of attacks has grown exponentially, which means that we must always be up to date in our defending mechanisms. Using intrusion detection systems is one of many efficient ways to protect a network, and machine learning is becoming the most training field of this century; it is starting to redefine the way we live. So, using IDS and combining it with machine learning will make a more efficient system. Moreover, we can combine a classification method and principal component Analyses. The QR based PCA will be used to optimize the data while the classifier will be used as a model for the prediction of the output, i.e., to define if it is a true negative (TN), false negative (FN), false positive (FP), or false negative (FN), which means it will identify if there was an actual attack or just a false alarm. We will have two sets of data, and from the training data we will extract the features using QR based PCA to compare then with the information from the test data by using a classifier to define the output.

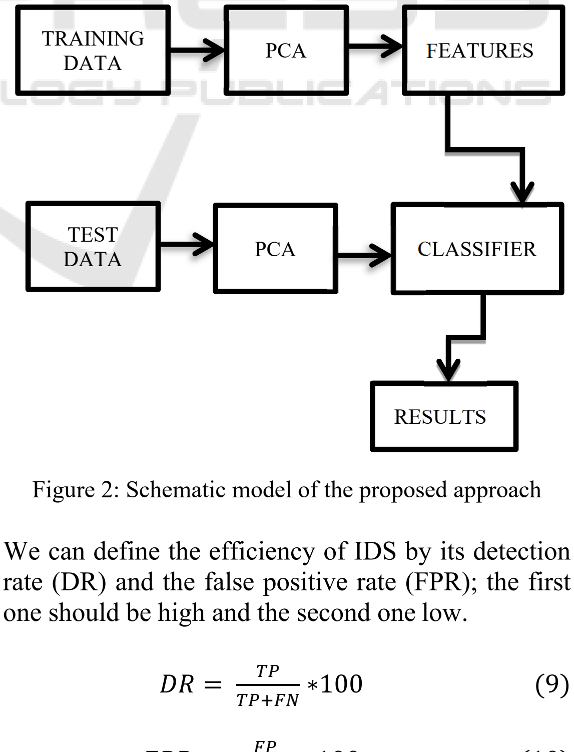

Figure 2: Schematic model of the proposed approach

We can define the efficiency of IDS by its detection rate (DR) and the false positive rate (FPR); the first one should be high and the second one low.

$$DR = \frac{TP}{TP+FN} * 100 \quad (9)$$

$$FPR = \frac{FP}{FP+TN} * 100 \quad (10)$$

5 CONCLUSION

This paper introduces the intrusion detection systems explaining their functioning, the principal component analyses, and machine learning. The main obstacle is the high dimensionality of the collected data, which is why the use of different techniques of dimensionality reduction will help us have the most efficient Intrusion Detection System. In order to do so, we can use a hybrid classifier and that is by combining several machine learning techniques which will provide high classification accuracy. So, we will use the QR based PCA because it will be faster and more efficient, and there will not be any loss of important information, and we will use another classifier. This technique will help us obtain a FPR as low as possible and the highest possible DR.

REFERENCES

Winursito, A., Hidayat, R., Bejo, A., & Utomo, M. N. Y. (2018). *Feature Data Reduction of MFCC Using PCA and SVD in Speech Recognition System. 2018 International Conference on Smart Computing and Electronic Enterprise.*

Hadri, A., Chougdali, K., & Touahni, R. (2016*). Intrusion detection system using PCA and Fuzzy PCA techniques. 2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS).*

Li, J., Liu, Y., & Gu, L. (2010). *DDoS attack detection based on neural network. 2010 2nd International Symposium on Aware Computing*

Sharma, A., Paliwal, K. K., Imoto, S., & Miyano, S. (2012). *Principal component analysis using QR decomposition. International Journal of Machine Learning and Cybernetics, 4(6), 679–683.*

Guo, C., Zhou, Y., Ping, Y., Zhang, Z., Liu, G., & Yang, Y. (2013). *A distance sum-based hybrid method for intrusion detection. Applied Intelligence, 40(1), 178–188.*

Doctorat de l’Université de Toulouse : *Evaluation des systèmes de détection d’intrusion. Soutenue par Mohammed El-Sayed GADELRAB(2008).*

Nathalie Dagorn. *Rapport de recherche Détection et prévention d’intrusion : présentation et limites*

Liran LERMAN, *Université libre de Bruxelles _Les systèmes de détection d'intrusion basés sur du machine learning.*

J.Allen ,A. Christie ,W. Fithen, J. McHugh, J. Pickel &Ed Stoner(January 2000) *State of the Practice of Intrusion Detection Technologies.*

Labib, K., & Vemuri, V. R. (2006). *An application of principal component analysis to the detection and visualization of computer network attacks. Annals of Telecommunications - Annales Des Télécommunications, 61(1-2), 218–234.*

Math.univ-toulouse.Analyse en composante Principal.

Andy Liaw and Matthew Wiener (December 2002). *Classification and Regression by randomForest.*