Network Intrusion Detection System Based on Machine Learning

Farhana[1], Bharath Raju[2], Adarsh SN[3]
[1]Assistant Professor, Dept of Computer Science & Engineering
[2, 3]Dept of Computer Science & Engineering
[1, 2, 3] Atria Institute of Technology, Bangalore

Abstract- *Recently, it has become important to use advanced intrusion detection techniques to protect network from the developing network attacks, which are becoming more complex and difficult to detect. For this reason, machine learning techniques have been employed in the Intrusion Detection Systems (IDS), so that, more complex features can be detected in the characteristics of the packets incoming to the network. As these techniques require training data, many datasets are collected for this purpose. Some of these datasets have known issues that limit the ability to apply intrusion detection systems built, based on these datasets, in real-life applications. The main function of Intrusion Detection System is to protect the resources from threats. It analyses and predicts the behaviours of users, and then these behaviours will be considered an attack or a normal behaviour. We use k-Nearest Neighbour(k-NN), Artificial Neural Networks(ANN) and Support Vector Machine (SVM) to detect network intrusions. The results of the discussed studies show the great potential of using machine learning techniques to implement IDS, where the Artificial Neural Networks (ANN) have shown the highest average performance, among other machine learning techniques.*

Keywords- IDS, k-NN, ANN, SVM, KDD

I. INTRODUCTION

Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modelling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies. With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate. IDS detects attacks from a variety of systems and network sources by collecting information and then analyses the information for possible security breaches. The network based IDS analyses the data packets that travel over a network and this analysis are carried out in two ways. Till today anomaly based detection is far behind than the detection that works based on signature and hence anomaly based detection still remains a major area for research. The challenges with anomaly based intrusion detection are that it needs to deal with novel attack for which there is no prior knowledge to identify the anomaly.

Machine learning is the field of study that aims to provide computers with the ability of gaining knowledge from the external world, without any human interaction. The knowledge extracted by a certain machine learning technique may be different from one set of inputs, from the external world, to another. Moreover, knowledge extracted from a single set of inputs may also be different from one machine learning technique to another, according to the different approaches used to extract such knowledge. One of the main machine learning fields is data mining, where the inputs from the external world are datasets, collected from that knowledge extraction is required.

II. NETWORK TRAFFIC DATASETS

Different datasets are collected for the characteristics of packets in network traffic that includes normal and attack packets to build and evaluate the performance of data mining techniques in the field of net-work protection. As the classifiers have different approaches to extract knowledge from datasets, it is important to evaluate the performance of the classifier, as a measure of the quality of the extracted know-ledge. However, as the classifiers are used to provide predictions, labelled data are used for that evaluation, by comparing the predictions provided by the classifier to the actual classes, or labels, that these objects belong to. Thus, each dataset is split into two parts, one is used by the classifier to extract the knowledge, which is known as the training dataset, while the other is used to evaluate the performance of the classifier, by comparing the predicted classes to the actual ones, which is known as the testing dataset. One of the earlier dataset collected for network traffic to train and evaluate data mining techniques in IDS is the KDD Cup’99 dataset. This dataset includes information about 4,898,431 network packet, where each packet is characterized using 41 different features. Each packet is labelled with one of five labels, one for normal packets, and four for different network attacks that are included in the dataset, which are:

1.Probing Attack: is an attack that aims to gather any possible information about the network and thecomputers that belong to that network in order to use that information to compromise the security of the network.

2. User to Root (U2R): is an attack that exploits information of users who have legitimate access to the system using any vulnerabilities in that system.

3.**Denial of Service Attack (DoS):** is an attack that attempts to exhaust the resources available on acomputer, such as memory or processing power, in order to deny providing services to legitimate users.

4. Remote to Local Attack (R2L): is an attack where the attacker has access to the network but does nothave the necessary information to authenticate to the services provided on that network.

The first issue stated in this dataset is the tools used to collect the packets' information, such as the TCP dump tool which is expected to drop some of the network traffic during heavy traffic. Such drops are not examined during the collection of the dataset. Another issue is the definition of attacks included in the dataset, where probing attacks, for example, are not considered actual attacks, unless a certain threshold is exceeded, where such conditions are not considered in the data collection. Moreover, the number of redundant objects in the dataset is extremely high, which affects the difficulty of the analysis, as objects similar to those in the testing dataset are highly expected to be in the training dataset, which reduces the difficulty of predicting classes for these objects.

III. MACHINE LEARNING TECHNIQUES

Different machine learning techniques are used to implement intrusion detection systems using the datasets illustrated earlier. As the firewalls are the network components that are responsible of analysing the packet information in order to make a decision of allowing the packet access to the network or denying it, these techniques are used with these firewalls to protect the networks. The information of each packet is retrieved by the firewall and sent to the machine learning technique in order to predict whether it is of a normal or attack traffic. These predictions are used to make and execute the decisions in the firewall. In this section, machine learning techniques employed in intrusion detection systems are illustrated.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linearcombination of oneor more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from ***logistic unit***, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

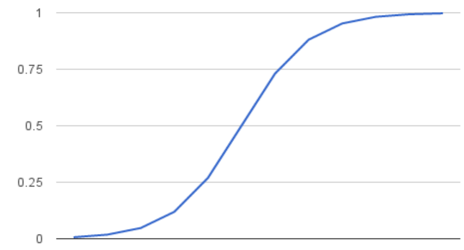

Fig.1 Logistic Regression Graph

Decision tree classifiers are utilized as a well known classification technique in different pattern recognition issues, for example, image classification and character recognition (Safavian & Landgrebe, 1991). Decision tree classifiers perform more successfully, specifically for complex classification problems, due to their high adaptability and computationally effective features. Besides, decision tree classifiers exceed expectations over numerous typical supervised classification methods (Friedl & Brodley, 1997).In particular, no distribution assumption is needed by decision tree classifiers regarding the input data. This particular feature gives to the Decision Tree Classifiers a higher adaptability to deal with different datasets, whether numeric or categorical, even with missing data. Also, decision tree classifiers are basically nonparametric. Also, decision trees are ideal for dealing with nonlinear relations among features and classes. At long last, the classification procedure through a tree-like structure is constantly natural and interpretable.Generally, a decision tree comprises of three basic segments including a root node, a few hidden nodes, and a lot of terminal nodes (known as leaves). An illustrative case of a decision tree structure is depicted in Fig. 2.21. As demonstrated, for each hidden and terminal node (known as child node), there should exist a parent node demonstrating the data source. In the interim, with respect to the root node and each hidden node (known as parent hub), at least two child nodes will be created from these parent nodes dependent on different decision rules.

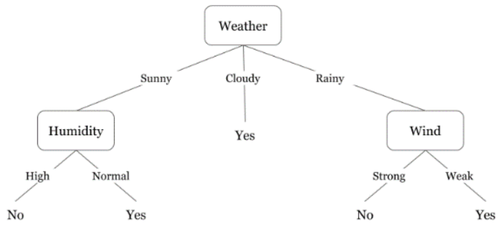

Fig. 2 Example of Decision Tree Classifier

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The first algorithm for random decision forests was created by Tin Kam Housing the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark (as of 2019, owned by Minitab, Inc.). The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

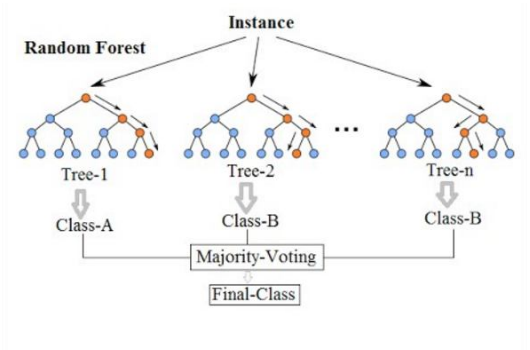

Fig. 3 Random Forest Example

IV. INTRUSION DETECTION SYSTEMS

Many intrusion detection systems are proposed based on machine learning techniques. These systems have shown different performances depending on the dataset, used for the training and evaluation, and the machine learning technique used in the system. The system proposed by Wei-Chao Lin, et al. uses the *k*-NN classifier to predict the state of each network packet, whether to be from a normal or at-tack traffic. This system is trained and evaluated using the KDD CUP’99 dataset, where the evaluation measures show a good prediction accuracy of 99.89% accurate predictions. However, as the *k*-NN classifier is a lazy classifier, the knowledge is extracted each time a prediction is required, i.e., the training data-set is scanned every time a new packet enters the network, which is a very resource-consuming process that requires either expensive servers withhigh resources, or longerexecution time that may degrade the quality of the services provided on that network.

Neha G Relan and Dharmaraj R Patil, which perform an intrusion detection system using the decision tree classifier. The performance of the proposed system has scored a highest of 95.09%, using the KDDCUP’99 dataset for both training and testing stage. The decision tree classifier generate sets of IF/ THEN rules that can be applied to the attributes’ values of each tuple, in order to predict a class for that tuple. These sets are created depend on the attributes values of the tuples in the training dataset, and the label that each record belongs to, where the sets are distributed in levels, and the condition to be investigated in the next level is selected depending on the outcome of the condition being applied in the current level.

Malek Al-Zewairi, et al. suggest an intrusion detection system depend on deep learning that include of five hidden layers with ten neurons in each layer. The deeper the neural network, the more complex attribute can be discover based on the input data, while rising the number of neurons in a layer rising the number of attributes that the layer can detect. The accuracy of the deep learning model is compared to other classifiers, such as logistic regression, decision tree, Naïve Bayes and neural network, where the experimental results show that the deep learning model has scored the highest with 98.99% accuracy when tested with the UNSW-NB15 dataset.

V. PROPOSED SYSTEM

we developed a supervised machine learning model that can classify unseen network traffic based on what is learnt from the seen traffic. We used both SVM and ANN learning algorithm to find the best classifier with higher accuracy and success rate.

The system proposed is composed of feature selection and learning algorithm. Feature selection component are responsible to extract most relevant features or attributes to identify the instance to a particular group or class. The learning algorithm component builds the necessary intelligence or knowledge using the result found from the feature selection component. Using the training dataset, the model gets trained and builds its intelligence. So when an unknown packet is coming SVM can classify easily whether it is a normal or attack packet. Using this method we could detect 95% of attack packets correctly and warning the administrator about it.

According to the administrators decision log file is created and storedfor future reference.

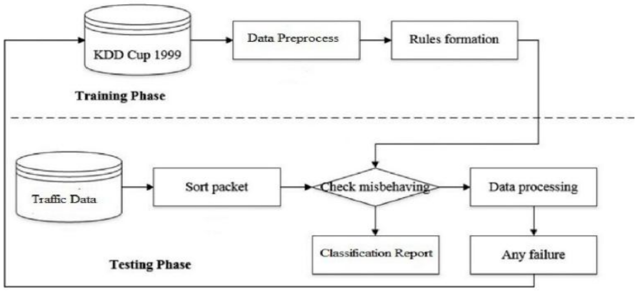

Fig. 4 Proposed Architechture

VI. CONCLUSION

Network attack methods are developing rapidly in order to execute intrusions using network traffic similar to normal traffic, so that, detecting these attacks becomes more difficult using traditional techniques. For this reason, intrusion detection systems are developed to use machine learning techniques to gain the ability of making more complex decision and protect the network from any intrusion attempts. Machine learning techniques have the ability to extract knowledge from a set of inputs collected from the external world. This knowledge is then used to assist making more appropriate decision based on the characteristics of the new data objects fed to the machine learning technique. Classification is one of the widely used supervised data mining techniques, where data mining is the field of machine learning that is concerned with processing datasets. A classifier extracts the characteristics of objects in each class to predict a class for new data objects, depending on their characteristics.

Many datasets are collected for packets in network traffic that contains both normal and attach packets, so that, these datasets can be used to train classifiers on how to detect attack packets incoming to the network, based on their characteristics. These predictions are used to come up with the appropriate decision, whether to allow the packet through to the network, or block it. Different studies are con-ducted that employ many machine learning techniques in intrusion detection systems. As these techniques have different approaches to extract knowledge from the datasets, they have shown different performance measures depending on the techniques and the training dataset. However, techniques that use Artificial Neural Networks have shown a better overall performance, compared to other techniques used for this purpose.

VII. ACKNOWLEDGMENT

We would like to thank our Research Guide Farhana, Associate Professor in Computer Science, Atria Institute Of Technology, Bangalore for their continuous support and valuable suggestions throughout this work carried out by us. Authors are also grateful to the reviewer for perilously going through the manuscript and giving valuable suggestions for the renovation of manuscript. We would also like to thank the Department of Computer Science Engineering, Atria Institute Of Technology, Bangalore for providing us with the facility for carrying out the simulations.

REFERENCES

[1] D. Acemoglu, A. Malekian, and A. Ozdaglar, “Network security and contagion,”*Journal of EconomicTheory*, vol. 166, pp. 536-585, 2016.

[2] D. Yu, Y. Jin, Y. Zhang, and X. Zheng, “A survey on security issues in services communication of Micro-services-enabled fog applications,” *Concurrency and Computation: Practice and Experience*, p. e4436.

[3] V. C. Storey and I.-Y. Song, “Big data technologies and Management: What conceptual modeling cando,” *Data & Knowledge Engineering*, vol. 108, pp. 50-67, 2017.

[4] H. Witten, E. Frank, M. A. Hall, and C. J. Pal,*Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.

[5] M. Ahmed, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,”*Journal ofNetwork and Computer Applications*, vol. 60, pp. 19-31, 2016.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,”*arXiv preprint arXiv:1409.1556,* 2014.