International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 09 | Sept - 2021 ISSN: 2582-3930

Supervised Machine Learning Technique for Network Intrusion Detection with Feature Selection

Abhilash L Bhat[1], Veena V[2]
[1]*Assistant Professor, Department of ISE, R.R. Institute of Technology, Bengaluru*
[2]*Assistant Professor, Department of CSE, R.R. Institute of Technology, Bengaluru*

Abstract - A novel supervised machine learning system is developed to classify network traffic whether it is malicious or benign. To find the best model considering detection success rate, combination of supervised learning algorithm and feature selection method have been used.
Through this study, it is found that Random forest Algorithm based machine learning with wrapper feature selection outperform support vector machine (SVM) technique while classifying network traffic. To evaluate the performance, NSL-KDD dataset is used to classify network traffic using SVM and Random forest techniques. Comparative study shows that the proposed model is efficient than other existing models with respect to intrusion detection success rate.
Key words: Intrusion Detection, Supervised Machine Learning, Feature Selection

1.INTRODUCTION

With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate. Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modeling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies.
IDS detects attacks from a variety of systems and network sources by collecting information and then analyzes the information for possible security breaches. The network based IDS analyzes 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. Hence the system somehow needs to have the intelligence to segregate which traffic is harmless and which one is malicious or anomalous and for that machine learning techniques are being explored by the researchers over the last few years. IDS however is not an answer to all security related problems. For example, IDS cannot compensate weak identification and authentication mechanisms or if there is a weakness in the network protocols.
The main objective of this project is to solve the problems face by existing NIDS techniques. In response to this we have proposed our novel NDAE method for unsupervised feature learning. We have then built upon this by proposing a novel classification model constructed the RF classification algorithm.

2.LITERATURE SURVEY

2.1 Network Intrusion Detection System (NIDS) using Machine Learning Perspective

Many intrusion detection systems are rule based which cannot detect novel attacks. Moreover, rule based technique is time consuming due to the encoded rule manually and it highly depend on the prior knowledge of the known attacks. Therefore, we proposed network based intrusion detection system (NIDS) using machine learning technique. NIDS is meant to be a device or a system application that monitor a network traffic and event occurring in a computer system. In network security intrusion detection system play a major role to detect different kinds of attacks. The machine learning technique can be used to increase the attack detection performance. In this paper Network intrusion detection system is proposed with the method of principle component analysis (PCA) and support vector machine (SVM). This proposed method was tested on KDD Cup dataset and attack detection accuracy is compared to decision tree and naive bayes algorithms.

2.2 Real Time Intrusion Detection System using Machine Learning

Today, world has come closer due to rapid increase internet. As technology has been developed many threats are emerged for the data security which is not at all good for sensitive data transactions, but as we know that the network security also posses equal importance in the computer infrastructure. Because of the intruders the security of the network has become serious problem. Thus to overcome this we are proposing this paper which is based on machine learning

algorithm for intrusion detection using Naïve Bayesian Classifier, which is based on probabilistic model.
This algorithm performs balance detections and keeps false positive rate at acceptable level for different types of real time networking attacks. In this, the system is trained by arranging the data attributes in a characterized format which eliminates the redundancy resulting in the reduction.

2.3 Network Intrusion Detection Using Machine Learning
In the network communications, network intrusion is the most important concern nowadays. The increasing occurrence of network attacks is a devastating problem for network services. Various research works are already conducted to find an effective and efficient solution to prevent intrusion in the network in order to ensure network security and privacy. Machine learning is an effective analysis tool to detect any anomalous events occurred in the network traffic flow. In this paper, a combination of two machine learning algorithms is proposed to classify any anomalous behavior in the network traffic. The overall efficiency of the proposed method is dignified by evaluating the detection accuracy, false positive rate, false negative rate and time taken to detect the intrusion. The proposed method demonstrates the effectiveness of the algorithm in detecting the intrusion with higher detection accuracy of 98.76% and lower false positive rate of 0.09% and false negative rate of 1.15%, whereas the normal SVM based scheme achieved a detection accuracy of 88.03% and false positive rate of 4.2% and false negative rate of 7.77%.

2.4 An Intrusion Detection System Using Machine Learning Algorithm
Security of data in a network based computer system has become a major challenge in the world today. With the high increase of network traffic, hackers and malicious users are devising new ways of network intrusion. In other to address this problem, an intrusion detection system (IDS) is developed which will detect attacks in a computer network. In this research, the KDDCup99 Test datasets is analyzed using certain machine learning algorithms (Bayes Net, J48, Random Forest, and Random Tree) to determine the accuracy of these algorithms by classifying these attacks into their various classes. A constructive research methodology is adopted throughout this research. The experimental results show that the Random Forest and Random Tree algorithms appear to be the most efficient in performing the classification technique on the Test dataset. The experimental tool used is WEKA which is used to perform a correlation based feature selection on the dataset with a Best First search method, and the parameters used for the computation are Precision, Recall and F-measure.

2.5 Evaluation of Machine Learning Algorithms for Intrusion Detection system
Intrusion detection system (IDS) is one of the implemented solutions against harmful attacks. Furthermore, attackers always keep changing their tools and techniques. However, implementing an accepted IDS system is also a challenging task. In this paper, Several experiments.

2.6 Evaluation of Machine Learning Algorithms for Intrusion Detection system
Intrusion detection system (IDS) is one of the implemented solutions against harmful attacks. Furthermore, attackers always keep changing their tools and techniques. However, implementing an accepted IDS system is also a challenging task. In this paper, several experiments have been performed and evaluated to assess various machine learning classifiers based on KDD intrusion dataset. It succeeded to compute several performance metrics in order to evaluate the selected classifiers. The focus was on false negative and false positive performance metrics in order to enhance the detection rate of the intrusion detection system. The implemented experiments demonstrated that the decision table classifier achieved the lowest value of false negative while the random forest classifier has achieved the highest average accuracy rate.

3.System Architecture

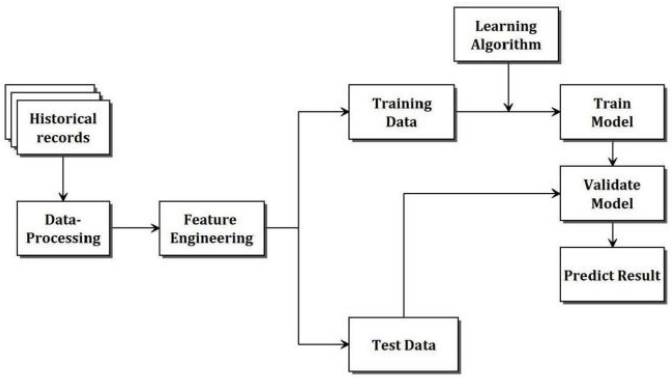

Fig - 1: System Architecture

In this system, we propose a novel deep learning model to enable NIDS operation within modern networks. The model we propose is a combination of deep and shallow learning, capable of correctly analysis awide-range of network traffic. More specifically, we combine the power of stacking our proposed Non-symmetric Deep Auto-Encoder (NDAE) (deep learning) and the accuracy and speed of Random Forest (RF)

(shallow learning). We have practically evaluated our model using GPU- enabled Tensor Flow and obtained promising results from analysing the KDD Cup '99 and NSL-KDD datasets. We are aware of the limitations of these datasets but they remain widely-used benchmarks amongst similar works, enabling us to draw direct comparisons.

This paper offers the following novel contributions:
1) A new NDAE technique for unsupervised feature learning, which unlike typical auto-encoder approaches provides non-symmetric data dimensionality reduction. Hence, our technique is able to facilitate improved classification results when compared with leading methods such as Deep Belief Networks (DBNs).
2) A novel classifier model that utilises stacked NDAEs and the RF classification algorithm. By combining both deep and shallow learning techniques to exploit their respective strengths and reduce analytical overheads. We are able to better or at least match results from similar research, whilst significantly reducing the training time.

Dataflow Diagram

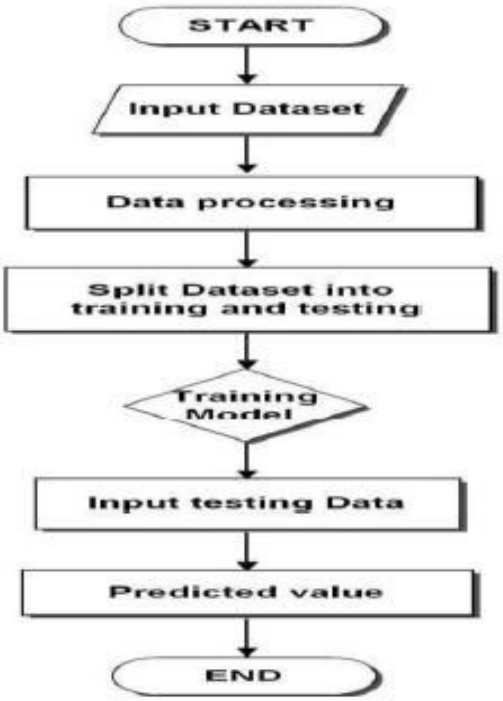

Fig - 2 : Data flow diagram

Data collection and preprocessing
In this module data are uploading into pandas data frame and it will enter into pre-processing to correct the missing data.

Training data and testing data split
Once data are preprocessed this system has to split data into division training data and testing data. Usually training should be large for accurate result.

Training process methodology
In this method the training data set with label has give to any one of the machine learning technique like random forest, this module will extract the feature from the label data keep it ready prediction process.

Prediction methodology
In this method the test data without label has to give prediction module which generate using training method this prediction module accept the test data and process. Finally it will produce the accuracy module.

Data visualization
This method uses mat plot lib python tool for producing graph from training data as well as testing data set.

4.Results

Test ID	Test Input	Excepted Result	Actual Result	Remarks
T_01	upload dataset	Uploaded data has to be stored in data frame.	Uploaded data will store in data frame.	Pass
T_02	Cleaning Process	During this process the data in data frame has to be verified and remove the all null values	The data in data frame will be verified and remove the all null values	Pass
T_03	Labelling process	During this process the uploaded twits has to labelled whether it is spam or not	The uploaded twits will be labelled whether it is spam or not	Pass
T_04	Data splitting	During this process the data set has to split into training and test data set	The data set will split into training and test data set	Pass
T_05	Training Process	This process has to read all the training dataset and create valid data model	This process will read all the training dataset and create valid data model	Pass
T_06	Testing Process	This process has to read test data and pass it to validation model and display whether the twit is spam or not	This process will read test data and pass it to validation model and display whether the twit is spam or not	Pass

Table - 1: Results

5.Conclusion and Future Enhancement

We have presented different machine learning models using different machine learning algorithms and different feature selection methods to find best model. The analysis of the result shows that the model built using SVM and Random Forest and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 94.02%.

The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero day attack still remains a research topic for future scope.

References

1. H. Song, M. J. Lynch, and J. K. Cochran, “A macro-social exploratory analysis of the rate of interstate cyber-victimization,” *American Journal of Criminal Justice*, vol. 41, no. 3, pp. 583–601,2016.

2. P. Alaei and F. Noorbehbahani, “Incremental anomaly-based intrusion detection system using limited labeled data,” in *WebResearch (ICWR), 2017 3th International Conference on*, 2017, pp. 178–184.

3. M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, "Modeling and implementation approach to evaluate the intrusion detection system," in *International Conference on Networked Systems*, 2015, pp. 513–517.

4. M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part C(Applications and Reviews)*, vol. 40, no. 5, pp. 516–524, 2010.

5. S. Ashoor and S. Gore, "Importance of intrusion detection system (IDS)," *International Journal of Scientific and EngineeringResearch*, vol. 2, no. 1, pp. 1–4, 2011.

6. M. Zamani and M. Movahedi, "Machine learning techniques for intrusion detection," *arXiv preprint arXiv:1312.2177*, 2013. [7] N. Chakraborty, "Intrusion detection system and intrusion prevention system: A comparative study," *International Journal ofComputing and Business Research (IJCBR) ISSN (Online)*, pp. 2229–6166, 2013.

7. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *computers & security*, vol. 28, no. 1–2, pp. 18–28, 2009.