RECURSIVE FEATURE ELIMINATION METHOD HYBRID MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM

[1]P.Sampanna Laxmi, [2]K.Venkatesh, [3]K.Deepthi, [4]Guntaka Usha Rani
[1,2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of Computer Science Engineering, Brilliant Grammar School Educational Society Group of Institutions Integrated Campus, Hyderabad, India

ABSTRACT
Network analysts find it difficult to manually monitor traffic flows and spot breaches in large networks due to the astronomical volume of internet traffic and rising network complexity. Information security is significantly enhanced by intrusion detection systems (IDS). Recently, it has become more common to use Machine Learning (ML) approaches for intrusion-based detection. The network traffic generates a lot of data with irrelevant information, slowing down the detection process and deteriorating system performance. Elimination or selection is carried out prior to categorization to get around that flaw. This research presents a hybrid machine learning-based intrusion detection system that employs recursive feature removal. In order to adequately and effectively analyze network traffic for intrusions, Support Vector Machine (SVM) and Decision Tree (DT) machine learning models are combined into hybrid machine learning model. Furthermore, a multi-class classifier was built to classify not only malicious or benign traffic but also to extend labels upon the malicious data. In order toimprove the performance, this system identifiesthe features which are irrelevant and eliminated it. The feature selection is achieved by using Recursive Feature elimination method. Experiments show improved performance in network intrusion detection.
KEY WORDS: Intrusion detection system (IDS), Machine Learning (ML), SVM, DT, Recursive Feature elimination.

INTRODUCTION

The sensitive information that is readily accessible online continuously draws enemies, making it vulnerable to severe network intrusion. When an adversary uses a compromised network to send malicious packets to the host system in an attempt to steal or alter sensitive data, this is referred to as intrusion. A server or system vulnerability like a mis-configuration or a programming flaw may cause it to happen. Whenever any such activity takes place on internal networks, intrusion detection systems (IDS) are deployed to sound the alert. Since they are so good at differentiating intrusions, machine learning and data mining techniques are frequently used to construct productive IDS. A model is trained utilizing datasets during the training phase of these techniques. Datasets provide examples of the attack and normal classes that are labeled. After training these models with mathematical algorithms, trained model is tested on separate samples of data to check accuracy of prediction [1].

Machine Learning is one of the technique used in the IDS to detect attacks [2].Machine learning is concerned with the design and development of algorithms and methods that allow computer systems to autonomously acquire and integrate knowledge to continuously improve them to finish their tasks efficiently and effectively. In recent years, Machine Learning Intrusion Detection system has been giving high accuracy and good detection of novel attacks. Intrusion Detection System (IDS) is a security technique attempting to detect various attacks [3]. They are the set of techniques that are used to detect suspicious activity both on host and network level. Majority of Machine learning and datamining approaches couldn't work well with intrusion detection because of gigantic complexity and size of datasets. These techniques take huge computational time to classify attacks which makes implementation more difficult in real time environments. This is because of hugenumber of features are contained in network data which is to be processed by Intrusion Detection System [4]. For better classification, hybrid machine learning by combining advantages of two machine learning models and quantity and quality of features matter and it helps us understand their importance and their correlation [5]. If features selected are very less, then classification quality will reduce and if they are more than required, it will make loss of generalization. Therefore a hybrid machine learning classification model along with the

Recursive Feature elimination method for reduce the number of feature in data are used in this paper.

LITERATURE SURVEY

In literature, numbers of anomaly detection systems are developed based on many different machine learning techniques. Forexample, some studies apply single learning techniques, such as neural networks, genetic algorithms, support vector machines, etc. Yunpeng Wang et. al [6] proposed an advanced Naïve Bayesian classification- based machine learning model for the efficient intrusion detection system. This advanced Naïve Bayesian Classification (NBC-A) is a combination of traditional NBC and RELIEFF algorithm; both are usedto train and test the network behavior using KDD-99 dataset. According to results obtained, NBC-A was suitable for large scale and complex dataset with a higher rate of a true positive.

Bhupendra Ingre, Anamika Yadav and Atul Kumar Soni et.al [7] recommended decision tree-based Classification and Regression Tree (CART) algorithm for effective attack categorization using NSL-KDD dataset. The Correlation-based Feature Selection (CFS) subset evaluation algorithm was used to select optimal features; then the CART decision tree-based algorithm evaluates the performance of the dataset. The optimal feature selection method was used to enhance the accuracy; although significant enhancement has been found in NSL-KDD dataset after applying the decision tree algorithm and hence detection rate of all the attacks has improved.

R. A. R. Ashfaq et. al [8] proposed a Fuzzy based Semi-Supervised Learning Approach for Intrusion Detection. They proposed a Fast-Learning Mechanism for Single HiddenLayer Feed Forward Neural Network with Random Weights and Fuzziness for Intrusion Detection. Their approach is based on the principle of Divide and Conquer Algorithm Design technique. The fuzziness of each Sample is evaluated and classified into 3 categories as High FuzzinessSamples, Low Fuzziness Samples, and Mid Fuzziness Samples. Samples with High and Low Fuzziness are used to retrain theSystem. Neural Network with Random Weights has shown an Excellent Learning Performance and is Computationally Efficient. The proposed system has shown a High Accuracy Rate for Samples with High and Low Fuzziness but it has shown Low Accuracy Rate with Samples having Mid Fuzziness.

A. Midzic, Z. Avdagic and S. Omanovic et.al [9] recommended a hybrid model for intrusion detection using neural network andfuzzy logic. Self Organizing Map (SOM) block was responsible for the reduction of training data through the process of clustering data in smaller subsets or clusters. These clusters are used in the Adaptive Network-Based Inference System (ANFIS) for training the system. Fuzzy logic is used in the ANFIS system. They used the KDD- 99 dataset for the training and testing of the system. Javaid et al. [10] proposed an approach to trained Intrusion detection system using deep learning mechanisms. Self-Taught Learning (STL) and deep learning-based technique used to improve the performance of the network intrusion detection system. Self-Taught Learning (STL) is a kind of deep learning methodconsists of two classification stages. In the first step deal with the good classification of features from large unlabeled dataset (Unsupervised Feature Learning) and secondstep applied to leveled dataset for classification.

HYBRID MACHINE LEARNINGBASED IDS

The purpose was to classify malicious orbenign traffic and extend labels upon themalicious data, meaning, to classify each malicious packet. The machine learningmodels and training methods used in thisresearch are supervised learning models. The first step was to collect data. The dataset collected here is KDD cuP99 dataset, and both contained ground truth files. This is imperative because supervised learning algorithms depend on labelled data. After the datasets were obtained, data pre- processing, the second step of the process, was done on the dataset to make it usable forthe machine learning algorithm. Analyzing large amount of data is difficult and time consuming. In order to overcome that only relevant features are taken for analysis. In this IDS system feature selection is done by using recursive feature elimination method. In this technique the features are ranked. The least rank(s) are then removed. Then themodel is re-built and the least ranked is removed again. This is repeated until theoptimal subsets of features are given for classification. The Fig. 1 explains the architecture of Hybrid machine learning based IDS system with recursive feature elimination.

Data Pre-processing

The classifier takes only numerical data to give accurate result. So, in the pre- processing module non-numerical features should be converted into numerical features. After that, normalization of the numerical values is done. The data

preprocessingincludes three steps:

Data Cleaning: The cleaning process in dataset is done initially to remove duplicate records and to add missing values.

Numericalization

The process of converting non numerical dataset into numerical dataset is known as numericalization. The non-numericalfeatures can be of Protocol type, service and flag. The One-hot encoding technique is used for converting non-numerical features into numerical features. Finally, all thefeatures of dataset are converted into numerical form.

Normalization: The process of normalizing or scaling the value of the entire feature in the range of [0, 1] is known as normalization. This process is done in order to perform better classification. In thisprocess average or mean "μ" is calculatedfor each feature. Then subtract the mean value "μ" from the feature value "x'. Then divide that subtracted mean and feature "x- μ" value by its standard deviation "σ" as in equation (1). Finally each feature will have μ= 0 and σ =1. Xi = (x- μ) / σ (1) In the above equation, "μ" is mean or average, "x" is the feature value and "σ" is the standard deviation.

Recursive Feature Elimination

In Feature selection or elimination is the process of removing redundant and irrelevant data from the dataset which arenot suitable for the specific task. The KDD Cup99 dataset has 41 features, but all these features are not required for the network intrusion detection. Using feature elimination method mainly important features are chosen to acquire maximum efficiency and also to reduce the computation time. In this system selected a subset of Features from the existing 41 features. This selected subset of feature produces maximum classification accuracy. For feature selection 10 fold cross validationis done. Using recursive feature elimination method the features are passed as parameter to identify the selected features. This process is done repeatedly placed aside untilall features in the dataset are done. Finally 20 out of 41 features are selected using this method.

Hybrid ML Classification

With the help of classification model attacksand its types are detected. For the selected features classification models such as SVM- and Decision Tree (DT) into a hybrid ML classifier is applied.

Support Vector Machines (SVMs): SVM work by learning similarities between features. It is maximizes the margin between different classes of training data. There are two types of SVMs, simple and kernel. A Simple SVM works precisely like a general SVM but is used exclusively for linear datasets and data of lower dimensions (1D, 2D). A Kernel SVM functions almost the same as a Simple SVM, but the main difference is that Kernel SVMs are used for higher dimension datasets and deal with non-linearity. The purpose of the Kernel is to transform the input data into the required form.

Decision Tree (DT): A Decision Tree is a supervised learning model that classifies data in a tree structure. It creates decision trees on data samples using bagging and feature randomness and gives the prediction. It continuously splits a given dataset intosmaller subsets based on similarities and forms a tree based on an arbitrary parameter. It was chosen because it is a unique andstraightforward technique to implement. Itworks by taking test data and creating a branch for each possible outcome, after training of course. Each branch was continuously subdivided until all instances of a branch had the same class. The dataset is labeled as normal or attack class. The attack type falls into 4 categories such as DoS, R2L, U2R and probe.

RESULTS

The goal of this study was to design amachine learning based tool to be used forintrusion detection. Confusion matrices areused to represent the data associated topredicted and actual traffic classificationdone by classifiers. Following terms areused while representing a confusion matrix. True-Positive (TP): Correctly classify ananomalous sample as attack.

True-Negative (TN): Correctly classify anon attack sample as ordinary instance.

False-Positive (FP): Incorrectly classify anordinary sample as anomalous instance.

False-Negative (FN): Incorrectly classify anattack sample as ordinary instance.

		Predicted: Attack	Predicted: Normal
Actual	Attack	TP	FN
	Normal	FP	TN

Fig. 2: Confusion Matrix

Diagonal elements in matrix signifies correct predictions while remaining elements indicates wrong estimation Reduction of False negatives and False positives is a major research problem as these have very negative effects on overall security of networks. Using above mentioned terms and matrix from Fig. 1,Accuracy, Precision, Sensitivity and specificity are evaluated as,

The accuracy metric is the proportion of data that was correctly predicted. It ismathematically defined as, Classification has been performed for thedataset. This was needed because the datasetonly contains two types of data: regular traffic and DDOS traffic. For this classification, Decision Tree and Support Vector Machine algorithms were combineused as a hybrid machine learning to classifythe intrusions in the network. The performance evaluation metrics for this intrusion classification are calculated by using above metrics from confusion matrix. The following Table 1 shows the Performance evaluation on intrusion detection. Here, the SVM and DT as individual models are compared with the Hybrid ML model in terms of Accuracy, Precision, sensitivity and Specificity. From this it is apparent that the Hybrid Ml gives better results than other models.

Table 1: Performance Evaluation on Intrusion Detection

Performance Metric (%)	SVM	DT	Hybrid ML
Accuracy	91	88	96
Precision	88	84	94
Sensitivity	87	80	92
Specificity	40	20	15

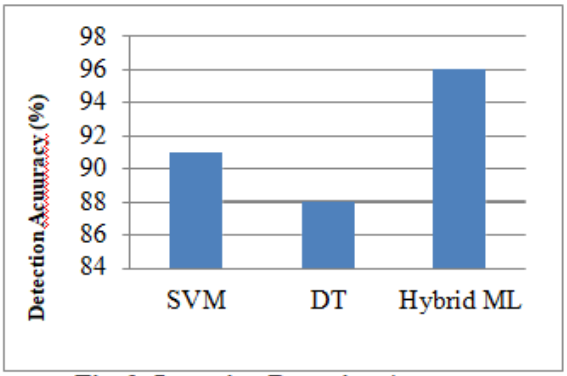

Fig. 3: Intrusion Detection Accuracy

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \text{ ------ (1)}$$

The precision metric (or Positive Predictive Value) was defined as the proportion of true positives relative to actual positives. It ismathematically defined as

$$Precision = \frac{TP}{TP+FP} \text{ ------ (2)}$$

Sensitivity is also called as recall or Probability of detection or True Positive Rate (TPR). The TPR is the rate of identifying actual positive result as positive as in equation (3)

$$Sensitivity = \frac{TP}{TP+FN} \text{ ------ (3)}$$

Specificity is also called as True Negative Rate (TNR). The TNR is the rate of identifying actual negative result as negativeas in equation (4).

$$Specificity = \frac{TN}{TN+FP} \text{ ------ (4)}$$

The hybrid machine learning model and the individual SVM and DT model were trained and tested using the dataset for classification. The results are given in Fig. 3.

The conclusion drawn is that the twoclassifiers SVM and DT did not perform well needed to be trained with more datacompared to our Hybrid ML model.

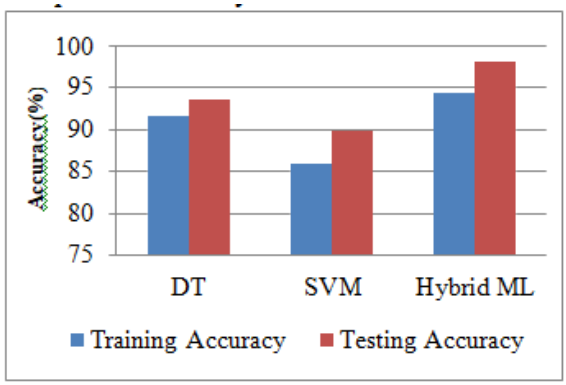

Fig. 4: Training and Testing accuracy

CONCLUSION

A hybrid Machine Learning based approach to intrusion detection system with recursive feature elimination was presented in this. Here, Support Vector Machine and Decision Tree combined into hybrid ML model were applied to perform classification on the packets from large computer networks. It was shown that with proper training, the machine learning models could identify malicious packets accurately. In addition, data pre-processing has been performed to mitigate the problem of unbalanced datasets. Because of the usage of feature selection the computation cost is decreased and detection rate of this system is increased. This IDS system is capable of detecting the traffic types with better correctness when measure up to the other existing system. The Hybrid ML approach gives 96% of Accuracy, 94% precision and 92% of Sensitivity. Thissystem shows that using Hybrid ML the detection rate is increased when compared toother model.

REFERENCES

1. S. Wang and Z.G. Jin, "IDS classification algorithm based on fuzzy SVM model", Application Research of Computers, vol. 02, pp. 501-504, 2020
2. C.Y. Hou, G.W. Wang and C.J. Wang, "Improved genetic algorithm optimizing SVM intrusion detection method", Journalof Longyan University, vol. 06, pp. 109-114,2019.
3. J. Liu and Z.X. Yang, "Improved ACO- SVM for network intrusion detection", Computer Engineering & Software, vol. 10, pp. 57-59, 2018.
4. Setareh Roshan, Yoan Miche, Anton Akusok and Amaury Lendasse, "Adaptive and Online Network Intrusion Detection System using Clustering and Extreme Learning Machines", ELSEVIER Journal of the Franklin Institute, vol. 355, no. 4, pp. 1752-1779, March 2018.
5. H. Wang, J. Gu and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation", Knowl.-Based Syst., vol. 136, pp. 130-139, Nov. 2017.
6. Wang, Yunpeng, Yuzhou Li, Daxin Tian, Congyu Wang, Wenyang Wang, Rong Hui, Peng Guo, and Haijun Zhang, "A novel intrusion detection system based on advanced naive Bayesian classification", In International Conference on 5G for Future Wireless Networks, pp. 581-588. Springer, Cham, (2017)
7. Ingre, Bhupendra, Anamika Yadav, and Atul Kumar Soni. "Decision tree based intrusion detection system for NSL-KDD dataset." In International Conference on Information and Communication Technology for Intelligent Systems, pp. 207-218 Springer, Cham, (2017)

8. Ashfaq, Rana Aamir Raza, Xi-Zhao Wang, Joshua Zhexue Huang, Haider Abbas, and Yu-Lin He, " Fuzziness based semi-supervised learning approach for intrusion detection system", Information Sciences, 378, pp. 484-497, Feb. 2017.
9. Midzic, A., Z. Avdagic, and S. Omanovic. "Intrusion detection system modeling based on neural networks and fuzzy logic", In 2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES), pp. 189-194(2016)
10. Javaid, Ahmad, Quamar Niyaz, Weiqing Sun, and Mansoor Alam, A deep learning approach for network intrusion detection system. Proc. 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), pp.21-26, May 2016