ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 12, December 2019

Detection of Network Intrusion by using Supervised Machine Learning Technique with Feature Selection

[1] Ch Muni Koteswara Rao, [2] T N Siva Kumar, [3] J Avinash
[1][2][3] Assistant Professor in Department of CSE, N.B.K.R Institute of Science & Technology

Abstract: To find out network traffic and classify whether it is malicious or benign A novel supervised machine learning system is used .To obtain best model considering detection success rate, both combination of supervised learning algorithm and feature selection method have been used. To classifying network traffic it is found that Artificial Neural Network (ANN) with wrapper feature selection support vector machine (SVM) technique is used. To evaluate the performance, NSL-KDD dataset is used to classify network traffic using SVM and ANN supervised machine learning techniques. With respect to intrusion detection success rate Comparative study shows that the proposed model is efficient than other existing models.

Index Terms – intrusion, machine learning, deep learning, neural network, support vector machine, feature selection.

1. INTRODUCTION

Cybercrime is happening at an increasing rate as with wide usage of Internet for accessing online contents at increasing rates, Intrusion detection is the first step to prevent security attack. IDS detect 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 over a network 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. Novel attack is the challenge with anomaly based intrusion detection for which there is no prior knowledge to identify the anomaly. Hence the system needs to have some intelligence to segregate the traffic which is safe 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. IDS cannot compensate weak authentication mechanisms or weakness in the network protocols. While network IDS that works based on signature have seen commercial success throughout the globe, anomaly based network IDS have not gained success in the same scale. Due to that IDS is currently anomaly based detection is a major focus area of research and development .And before going to deployment of anomaly based intrusion detection system in wide scale key issues remain to be solved. But the literature today is limited to apply supervised machine learning techniques to protect target systems and networks against malicious activities anomaly-based network IDS is a valuable technology.

Several anomaly based techniques have been proposed including Linear Regression, Support Vector Machines (SVM), Genetic Algorithm, Gaussian mixture model, k-nearest neighbor algorithm, Naive Bayes classifier, Decision Tree. Among them the most widely used learning algorithm is SVM as it has already established itself on different types of problem. One major issue on anomaly based detection is it can detect novel attacks but they all suffer a high false alarm rate in general. The major challenges in evaluating performance of network IDS is the unavailability of a comprehensive network based data set. In this paper we used SVM and ANN –two machine learning techniques, on NSL- KDD which is a popular benchmark dataset for network intrusion. Hence we came out that the challenge of identifying new attacks or zero day attacks facing by the technology enabled organizations today can be overcome using machine learning techniques. Here 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.

II. SYSTEM MODEL

In this proposed system is composed of feature selection and learning algorithm show in Fig.1. Feature selection component are responsible to extract most relevant features or attributes to identify to a particular group or class. To build intelligence using the result found from the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 12, December 2019

feature selection component the learning algorithm component is used.The necessary intelligence using the training dataset, the model gets trained and builds its intelligence. Then use the learned intelligences are applied to the testing dataset to measure the accuracy of how much the model correctly classified on unseen data.

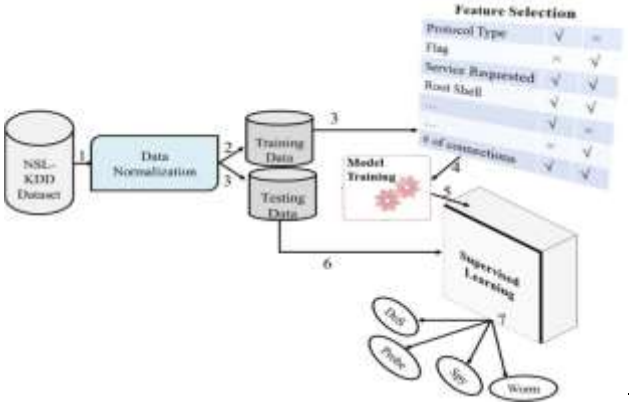


FIG 1: PROPOSED SUPERVISED MACHINE LEARNING CLASSIFIER SYSTEM

A. Feature Selection

For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model. Firstly in wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince classifiers from the features in each subset. The classifier considers the subset of feature with which the classification algorithm performs the best. To identify the subset the filter method uses an a ranker to rank all the features in the dataset. Here one feature is omitted at a time that has lower ranks and then sees the predictive accuracy of the classification algorithm. Wrapper method is useful for machine learning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features.

Fig 2: Filter Method

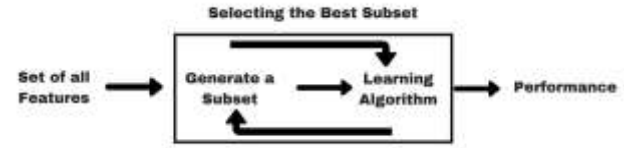


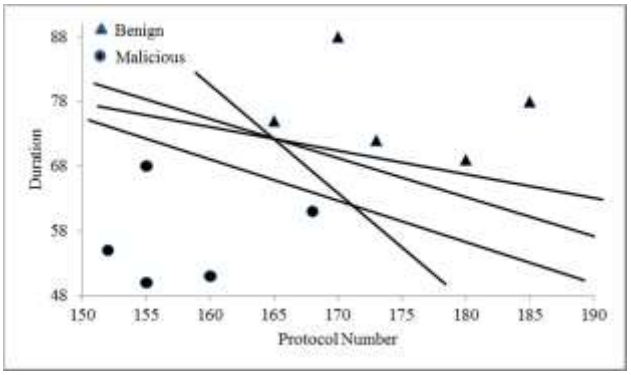
Fig 3: Wrapper Method

B. Building Machine Intelligence

Best features found in the feature selection process, Based on it learning models are developed. To develop the learning model, machine learning algorithm is used. With the selected features Training dataset is used to train the algorithm. In supervised machine learning, each instance in the training dataset has the class it belongs to.

C. Support Vector Machine

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well SVM a separating hyper plane defines the classifier depending on the type of problem and available datasets. In case where dataset is one dimensional, the hyper plane is a point, for two dimensional data it is a separating line as shown in Fig 2, for three dimensional dataset, it is a plane and if the data dimension is higher it is a hyper plane. For a linearly separable dataset, the classifier or the decision function will have the form –



ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 12, December 2019

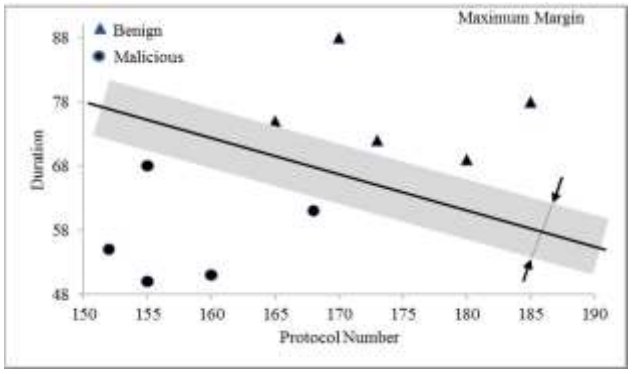


Fig 4: SVM classifier in two dimensional problem spaces

D. Artificial Neural Network (ANN)

ANN is a system inspired by human brain system and replicates the learning system of human brain. Artificial Neural Network is another tool used in machine learning. It consists of input and output layers with one or more hidden layers in most cases. It uses a technique called back propagation to adjust the outcome with the expected result or class.

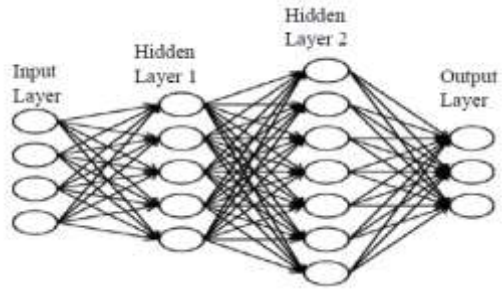


Fig 5: Artificial neural network showing the input, output and hidden layers

III. EXPERIMENTAL ANALYSIS OF THE SYSTEM

A. Feature Selection

The experiment carried out using Weka open source software for data mining and machine learning and consists of two parts. feature selection (FS) methods are used for extract most relevant features. In the wrapper method we used SVM classification algorithm with cross-validation to avoid over fitting and under fitting problem. In the filter method a ranker algorithm is used to find the best result suitable for our proposed classifier. The training data we used from NSL-KDD dataset contains 25,191 labeled instances. Results of the feature selection experiment are shown in Table 1.

FS Technique	FS Type	Input Features	Output Features
Correlation Based	Wrapper	50	19
Chi-Square Based	Filter	50	38

TABLE 1: RESULT OF FEATURE SELECTION

Correlation based feature selection found total 17 features most relevant from 41 features present in the training dataset whereas Chi-Square algorithm retained 35 features to be more relevant to the resultant class. These 17 and 35 retained features were used to train the model using training or seen dataset as well as to test the model using unseen or testing dataset.

B. Classification

SVM and ANN learning algorithm are used to training the model for each type of feature selection method. Hence we build four learning models, two model using SVM and another 2 using ANN. Among the 2 model built for each learning algorithm, one is built using 17 features and another one is built using 35 features found in the feature selection part.these four trained models were evaluated using 22,542 instances of testing data picked from the NSL-KDD testing dataset.

TABLE 2: RESULT OF CLASSIFICATION

Learning Type	Number of Features	Detection Accuracy
SVM	19	82.18%
SVM	38	83.44%
ANN	19	95.12%
ANN	38	84.78%

Table 3: PERFORMANCE COMPARISON WITH EXISTING MODELS

Learning Type	Our Model Accuracy	Existing Model	Existing Model
SVM	83.44%	92.84% [16]	69.52% [17]

ANN	95.12%	81.2% [18]	77.23% [19]

IV. DISCUSSION ON SYSTEM IMPLEMENTATION

To implement and evaluation the system we have used widely used open source machine learning software suite called Weka. Along with machine learning algorithm implemented, Weka also has several algorithm and search technique implemented to perform feature selection. In the ANN model, we experimented with different number of hidden layer and found that the detection success rate varies with the number of hidden layer. After several trial and error methods, we found best detection rate with 3 hidden layers and 0.1 learning rate. In the wrapper feature selection method, we also used SVM algorithm as classifier. The model implemented in Weka has been run on a computing platform having 64 bit 2.6 GHz Intel core i5 CPU with 8 GB RAM on Windows 7 environment with limited network traffic instances. Implementing the solution on large scale network will require additional infrastructure with some higher capacity server platform.

V. CONCLUSION

In this paper, we have presented different machine learning models using different machine learning algorithms and different feature selection methods to find a best model. The analysis of the result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 95.12%.The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero day attack still remains a research topic due to the high false positive rate of the existing system.

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 Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178–184.
[3] 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.
[4] M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection," Procedia Computer Science, vol. 89, pp. 117–123, 2016.
[5] T. Janarthanan and S. Zargari, "Feature selection in UNSW-NB15 and KDDCUP'99 datasets," in Industrial Electronics (ISIE), 2017 IEEE 26th International Symposium on, 2017, pp. 1881–1886.
[6] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446–452, 2015.
[7] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), 2016, pp. 21–26.
[8] M. Zamani and M. Movahedi, "Machine learning techniques for intrusion detection," arXiv preprint arXiv:1312.2177, 2013.