Design an Intrusion Detection System based on Feature Selection Using ML Algorithms

Srinath Venkatesan,
New York University, sv778@nyu.edu

Article Info
Page Number: 702 - 710
Publication Issue:
Vol 72 No. 1 (2023)

Abstract

With the development of the Internet and Technology, cyber-attacks are growing rapidly thus cyber security is an important aspect which needs to be given attention. There are several types of attacks occurring across the internet like: Denial of Service (DoS) , R2L (Root to Local attacks), U2R (User to Root attack), Probe (Probing attacks) , DNS Spoofing attack etc for which ways to identify the vulnerabilities have to be found. Machine learning, an emerging field in the current time has many techniques that can be adopted in various domains where it shows its dominance over many traditional algorithms. These techniques can be exploited in the field of cyber security with the aim of detecting intrusions which can support or even replace the first level of security analysts. This paper aims in building a model for intrusion detection by applying various machine learning algorithms on the selected features obtained through the modelling process. NSL-KDD dataset is used in order to build IDS. The paper provides a brief description of various machine learning algorithms like SVM, Random Forests, Decision Tree in order to study the of attacks on the prescribed dataset using the selected critical features which aims in improving the accuracy. The paper also provides a comparative study of the algorithms used to determine the algorithm that works the best.

Article History
Article Received: *15 October 2022*
Revised: *24 November 2022*
Accepted: *18 December 2022*

Keywords— Intrusion Detection System, Machine learning, Decision Tree, Random Forest. SVM, feature selection, DoS, R2L, U2R, Probe.

I Introduction

With new technologies like cloud computing, big data analytics and social media our society has huge quantities of data present in all forms. When we transmit these data's over a network or internet, we face so many issues regarding security. Although many new inventions regarding intrusion prevention has been developed to reduce the impacts of these threats, attacks continue to happen and only increases in number but never decreases and for this reason we need a strong mechanism which helps in detecting any unwanted traffic causing damage to a specific network.

This problem discussed above gives rise to the need of an Intrusion Detection System (IDS) which is commonly known as group of hardware or software device which is very much capable of collecting, analyzing and detecting malicious traffic either on a host particularly or a network [1]. Therefore, to achieve all the results it is expected to receive we do a lot of statistical and mathematical methods to interpret and read the data it has collected and report if anything is malicious to the network administrator [2].

There is another important issue regarding the ID detection which is the humans participation in the interaction. The intrusion dataset maybe of big size with the increase in the number of observed attributes and increasing amount of false positive results are generated as many duplicate records may be present [5].

Machine learning, an emerging field in the present can reduce the need of human interaction. This helps in optimizing the performance using example data or past experience using a programming model. To perform this task, it mainly uses 2 techniques namely classification and clustering. Classification is nothing but predicting most probable category, class or label. In clustering classes are not predefined during the learning process. In our paper classification techniques are used for differentiating between the normal and intrusion traffic and for identifying the type clustering techniques are used [4].

In this paper, three machine learning algorithms namely Decision tree classifier, Random Forest and SVM are employed on the most well-known dataset used in the field of cyber security, NSL-KDD dataset. The paper also provides a mechanism for feature selection in order to build an accurate model. Comparative study is done on all the three prescribed algorithms to determine the algorithm which gives the best accurate results.

II LITERATURE SURVEY

In this section we discuss about various literary works obtained through various means.

In [1], the paper provides a distributed Intrusion Detection System supporting multiple freelance entities operating conjointly by exploiting the machine learning algorithm. This approach solves a number of the issues related to cyber security. The model is built using Decision tree algorithm by applying concepts of feature selection.

In [2], the paper focuses on the (DoS) attacks with the aid of pattern recognition techniques in data mining and analysis. The paper provides a brief summary on the severity of DoS that can jeopardize the IT resources that are valuable for a corporation by overloading with messages and requests from any un-authorized user.

In [3], we see that with the developments of latest technologies and usage of internet cyber-attacks are increasing thereby leading to the necessity of cyber security. Here we get to know about the usage of ML and DL algorithms in the analysis of network. And we get a brief information on the different types of datasets that can be used.

In [4], novel ML system concepts are exploited to develop an efficient system that classifies the traffic caused due to network for classification of malicious and benign attacks. In this paper they realize that the artificial neural network related machine learning algorithm worked better than the SVM algorithm.

In [6] the paper covered some classifier dependant techniques that does not include choices. Similarly [7][8] focuses on the types of ID’s schemes. In this paper it stated that vulnerabilities can be figured out without the need for any implementation.

In [9] they developed a interruption identification scheme using the DT classifier. They focus on C4.5 calculations for achieving a better accuracy thereby minimizing the error rate.

In [10] utilized the machine learning concepts like DT algorithm to retrieve important aspect set from the dataset prepared for ID. The methodology included the calculations of both ID3 and C4.5.They incorporated a new element called weight where the nodes closer to the weight value within the tree is given a non- zero value while the other are approximated to zero .Similarly [11] suggested ML methodologies like SVM and DT classifiers stating that DT produces accurate results when compared to SVM.

In [12] they developed a statistical technique for splitting up the KDD99 data set in 2011.The splitting was done by identifying important features by checking the dependency between them.

In [13] developed a new strategy called Feature-Vitality Based Reduction Method (FVBRM) that mainly involves the concept of Naive Bayes classifier. Using an in-query approach every element is eliminated one at a time until the accuracy of the classifier reaches a limit. The work does not propose any strategy to detect U2R attacks

In 2013 the paper [14] developed a ID system using the KDD dataset. The main drawback of this system lied in the preparation time which was overcome using RBF. In 2014, the paper [15], exploited the concepts of DT algorithm and RF algorithm with voting mechanism that included forward detection and backward elimination in the field of cyber security.

In [16] the authors have inclined to develop 2 DT classifier techniques for attack classification. They have used C4.5 methodology with pruning and without pruning. The results showed the former showed better results as it paid heed to the removal of elements thereby concentrating on the interested area. This study still takes a longer execution time.

III BACKGROUND STUDY

A. Machine Learning:

In this section a brief description is provided on the machine learning techniques employed in our paper.

1. *Decision Tree:*

This algorithm comes under the broad classification of the supervised learning. It is a widely used algorithm that can solve both regressions based and classification problems that occur. DT's make use of the tree representation. In order to find an optimal solution to the problem the tree is constructed in such a way that the leaf node maps to the class label and attributes present in the problem are located in the internal node of the constructed tree. In DT any Boolean function can be illustrated on discrete labels. At each and every test one feature among all is used to split the node taken into consideration according to the values that are tagged onto the features.[1] After every split a check is done to see if the instances selected belong to classes of similar type and then the split is considered pure or complete. Decision tree (DT) can be built by using ID3, Gini Index or CART based on the type of information present.

Measurement of split using ID3 is done using entropy and information gain calculation

Entropy is the information found in dataset because of the appearance of more than one possible classification. Entropy need to be calculated for the target label as well as the local variables present w.r.t the target variable.

Formula for entropy using frequency table for target attribute:

$$\text{E(Y)}=\sum_{i=1}^{l} -q_j log_2 q_j$$

Where S is the target variable and p_i are the probability function

Formula for entropy using frequency table of local attributes w.r.t target attribute.

Where X represents all the local attributes present and T is the target attribute.

The information gain (IG) is mainly dependant on the entropy. The decrease in entropy affects the IG when the dataset in split for any attribute. The DT construction mainly depends on finding attributed with highest IG.

Formula for Gain calculation: Gain is calculated by taking the difference between the entropy measures of the target and the local attributes w.r.t the target attribute.

$$\text{Gain }(T, X) = \text{Entropy}(R) - \text{Entropy }(R, Y)$$

The decision tree algorithm avoids over-fitting of data and works well even in the presence of missing data [5]

2. *Support Vector Machine (SVM) :*

A Support Vector Machine (SVM) is an efficient and accurate ML algorithm [1]. The SVM has 2 phases that include support vector Classification (SVC) and support vector regression (SVR). The SVC has a decision boundary which separates a set of instances or fields of interest into two different groups because they are differing in class values. This technique classifies the data by using a separating hyper plane. For ex. in 2D space the hyper plane divides into 2 parts having each class in one side. In case data points are found not be separable the SVM uses some kernel functions like RBF, Gaussian Kernel etc to transform into some higher dimensional space. Therefore the main objective of SVM is to find an optimal hyper plane in N D space which classifies the data points precisely [7]. The hyper plane with maximum (i.e dist between the data point of the 2 classes) margin is to be found out

The RBF is used as kernel functions to map them to higher dimensional places.

3. *Random Forest :*

It is a supervised learning algorithm. The objective of this algorithm is to build a forest and randomize it. The forest built using this technique is collection of DTs. The training is done with the ensemble mechanism called the bagging- method. Bagging method takes a combination of learning-models to elevate the overall accuracy and provide better results. In this case RF builds multiple DTs and merges them to get accurate predictions. RFs can be used for classification as well as regression.[1]

It basically draws bootstrap samples from the given data and for each of these samples an un-pruned classification [3] or regression tree grows. Here rather than choosing best split among all predictors we pick random available samples and find the best split. Next step here is that we predict a new data by doing the aggregations of the predictions of the trees after which we get an approximate prediction of error. Things we have to remember while doing is we should not predict the data from the bootstrap samples. After which all the error rate is calculated.

B. Types of Cyber attacks:

This section provides a description of the types of cyber attacks which substantiates the need of building an Intrusion Detection System (IDS).

1. *Denial of Service (DoS) :*

This is a type of attack that is prominently found in the field of cyber crimes that mainly concentrates in obstructing the services that are provided by the system or the network by masking it with the aid of junk requests thereby restricting services to the authorized users present. [5]

2. *Probe* :

In this attack the attacker or the hacker tries to completely scrutinize the data and the vulnerabilities present in the system or in the networks [11] that will be stored elsewhere to perform any attacks to the system.

3. *Remote to Local (R2L):*

In this type the attacker or the hacker obtains a firm access[1] of any computer system over a network that is considered to be unauthorized. This is done by sending packets to the system at regular intervals that will later be the prime source of attack.

4. *User to Root (U2R):*

In this type of attack that is prevalent in the field of cyber crime the attacker or the hacker obtains a quick access of any user having normal privilege restrictions and then exploits the administrative or root privileges of the system to perform some malicious attacks[3]

IV METHODOLOGY

A. Dataset

The KDD99 data set is the most commonly used data set for data mining in the field of cyber security [1]. The data set has network and operating system data collected over a period of 9 weeks. This dataset has a lot of duplicate records so a new enhanced data set was developed call NSL-KDD dataset which we have used in our study.[5]

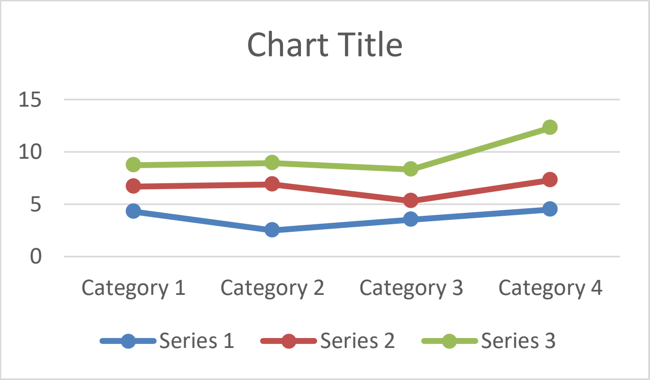
The attributes in the dataset consists of 123 features, a brief description of which is shown in the *Table 4.1*

Feature	**Description**
Basic	Duration, Service
Time oriented	Count Rerror
Content	Hot, Num_root
Host oriented	Dst_host_count

Table 4.1 Description of the attribute fields.

B. INTRUSION DETECTION SYSTEM

The proposed model of Intrusion Detection System consists of 4 steps as described in the *Figure 4.1.*



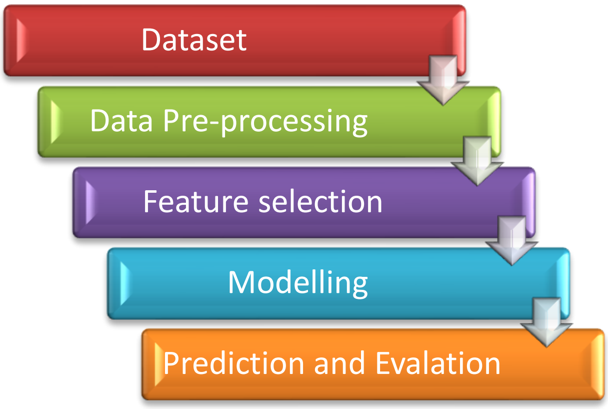


Figure 4.1 Intrusion Detection Systems

1. Data Pre-processing

Data cleaning and pre-processing is the first and the foremost step that is carried out after the data is collected in order to build an efficient and accurate model. This step involves cleaning the data set to remove any duplicate or sparse records. Since we are building a model on the NSL – KDD [1] dataset the cleaning process can be skipped because the data obtained is already cleaned. After cleaning the data we need to pre-process the dataset in order to remove non-numerical labels like categorical variables as all the classifier algorithms available in sick it works best with only numerical fields. We use one-hot encoding method to convert the necessary labels into numeric fields in order to build our model.

2. Feature Selection:

Feature selection is the most significant step in building any ML model as it improves the efficiency of the model. Before we perform any feature selection process, we need to scale the feature to avoid them from having large values which may cause an impact on the final result. Feature scaling results in features having an average of zero and a S.D of one. [7]

Feature selection eliminates irrelevant and useless features from the dataset and retains only the data set required for the model construction process. This decreases the computation time and also minimizes over-fitting of the any model. In our experiment we employ invariant feature selection algorithm using ANOVA F-Test to extract the needed features. This method examines each feature individually and closely in order to determine the strength of a relationship of the feature having labels. Once we obtain the needed feature we use the recursive feature elimination (RFE)[1] to eliminate all the unwanted features and keeping only the requires features .The features retained are ranked in order of their relevance and importance..

3. Modelling:

Once we obtain our data set with the selected features we need to build a model by splitting the dataset into test and train sample (already split in our case 80-20 split) for each category of the attacks specified. [4] In this paper, model is built on various machine learning algorithms like Decision Tree classifier [1], Random Forest Classifier [7] and Support Vector Machines [3] on each category of the attack. The model is built with selected features and also with the complete set of features in order to provide a comparative study.

4. Prediction and Evaluation:

After the model is built, the test data set is used to make prediction of the model. In our case the prediction of the category of the attack is found out i.e [Normal, DoS, Probe, R2L and U2R].Later evaluation is done using various measures like accuracy score, precision, recall and confusion matrix. The accuracies for each model under each category of attack is completely studied and a comparative study is done.[1]

V. RESULTS AND DISCUSSION

This section consists of the various results obtained from the experimentation process. The experimentation was carried using the Python Jupiter Notebook under the learning toolkit [16]. As discussed in the system above the feature selection done using ANOVA F-test and ranked using RFE gives us the following results show in *figure. All* the categories show 13 important features that are selected.

The modelling stage described in the Intrusion detection system focuses on comparing the accuracy between the model built with all the features and the selected 13 features. The following figure 5.1[1] shows the accuracy comparison for all the features vs 13 features using Decision Tree classifier. [1]

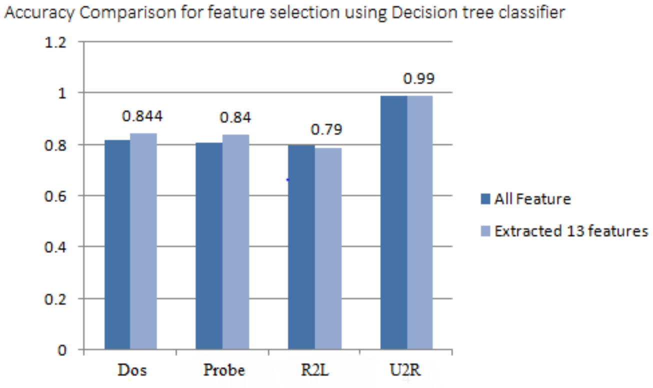


Figure 5.1 Accuracy Comparison for feature selection using Decision tree classifier.

The figure 5.1 obtained clearly shows that the decision tree classifier model built with the extracted features shows better accuracy in all the categories of the attacks thereby proving the need to employ feature selection in the modelling process to obtain better results.

A. COMPARATIVE STUDY

Here we provide a detailed comparative study regarding the accuracies for all the machine learning algorithms employed in our paper.

The figure 5.2 shows a comparison between the accuracy of the prescribed models namely Decision tree[1], Random forest[7] and Support vector machine[3].

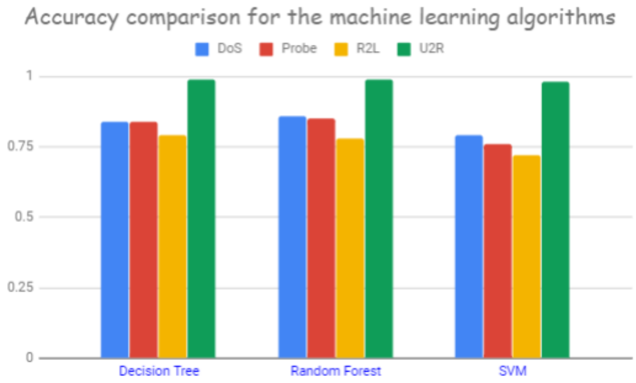


Figure 5.2 Accuracy Comparison for machine learning algorithms

The figure 5.2 clearly shows that the Random forest algorithm works best on all the categories of the dataset having an accuracy of 0.87, 0.86 ,0.76 and 0.98 for DoS [5],Probe [11], R2L [1] and U2R[3] categories respectively.

VI. CONCLUSION

The increase in the threats and attacks in the field of cyber crimes gives us the need to build an efficient Intrusion Detection System (IDS) by exploiting science and technology. Machine Learning, an emerging field in the field of data analytics can be employed to build efficient IDS.

In this paper we employ various machine learning techniques like Decision Tree [1], Random Forest and SVM in order to build efficient IDS. We were successful in building a model to predict the type of attack categorized prevalent in the field of cyber crimes by using the NSL-KDD data set [5]. The model uses feature selection technique like ANOVA F-Test and RFE to extract the required features and rank them by eliminating the rest. This method was successful in providing better accuracy when compared to building a model by considering all the features. The paper also provided a comparative study of all the algorithms employed by using evaluation metric like accuracy in order to evaluate the best algorithm that can be used for IDS. From our experimentation we were able to conclude that the Random Forest algorithm works best with the selected features for IDS.

In consideration with the future work many other machine learning algorithms and feature selection techniques can be employed in order to provide a detailed comparison with the aim of building more efficient IDS. Other fields of science and technology can also be exploited to detect new attacks as the IDS present is capable of predicting only the known attacks.

REFERENCES

[1] Nkiama, H., Said, S.Z.M. and Saidu, M., 2016. A Subset Feature Elimination Mechanism for Intrusion Detection System. *International Journal of Advanced Computer Science and Applications*, *7*(4), pp.148-157.

[2] Khan, M.A., Pradhan, S.K. and Fatima, H., 2017, March. Applying data mining techniques in cyber crimes. In *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)* (pp. 213-216). IEEE.

[3] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H. and Wang, C., 2018. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, *6*, pp.35365-35381.

[4] Taher, K.A., Jisan, B.M.Y. and Rahman, M.M., 2019, January. Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 643-646). IEEE.

[5] Thomas, R. and Pavithran, D., 2018, November. A Survey of Intrusion Detection Models based on NSL-KDD Data Set. In *2018 Fifth HCT Information Technology Trends (ITT)* (pp. 286-291). IEEE.

[6] C. F. Tsai, et al., "Intrusion detection by machine learning: A review," Expert Systems with Applications, vol. 36, pp. 11994-12000, 2009.

[7] V. Bolón-Canedo, et al., "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," Expert Systems with Applications, vol. 38, pp. 5947-5957, 2011.

[8] F. Amiri, et al., "Improved feature selection for intrusion detection system," Journal of Network and Computer Applications, 2011.

[9] Juan Wang, Qiren Yang, Dasen Ren, “An intrusion detection algorithm based on decision tree technology,” In the Proc. of IEEE Asia-Pacific Conference on Information Processing, 2009.

[10] Dewan Md. Farid, Nouria Harbi, and Mohammad Zahidur Rahman, "Combining Nave Bayes and Decision Tree for Adaptive Intrusion Detection," International Journal of Network Security & Its Applications, Vol. 2, No. 2, April 2010, pp. 12-25.

[11] Ektefa M, Memar S, Sidi F, Affendey L., "Intrusion detection using data mining techniques," 2010 International Conference on Information Retrieval & Knowledge Management(CAMP).2010.doi:10.1109/infrkm.2010.5466919.

[12] Geetha Ramani R, S.SivaSathya, SivaselviK, "Discriminant Analysisbased Feature Selection in KDD Intrusion Dataset," , International Journal of Computer Application VoI.31,No.ll, 2011

[13] S. Mukherjee and N. Sharma, “Intrusion Detection using Naive Bayes Classifier with Feature Reduction,” Procedia Technol., vol. 4, pp. 119– 128, 2012.

[14] Bhavsar Y. B, Waghmare K. C. "Intrusion Detection System Using Data Mining Technique: Support Vector Machine," International Journal of Emerging Technology and Advanced Engineering, Vol.3, Issue 3, pp.581-586(2013).

[15] O. Y. Al-Jarrah, a. Siddiqui, M. Elsalamouny, P. D. Yoo, S. Muhaidat, and K. Kim, “Machine-Learning-Based Feature Selection Techniques for Large-Scale Network Intrusion Detection,” 2014 IEEE 34th Int. Conf. Distrib. Comput. Syst. Work., pp. 177–181, 2014.

[16] N. G. Relan and D. R. Patil, “Implementation of Network Intrusion Detection System using Variant of Decision Tree Algorithm,” 2015 Int. Conf. Nascent Technol. Eng. F., pp. 3–7, 2015.