ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 11(48), DOI: 10.17485/ijst/2018/v11i48/139802, December 2018* ISSN (Online) : 0974-5645

Network Intrusion Detection System Using Machine Learning

Riyazahmed A. Jamadar*

Department of Information Technology, AISSMS Institute of Information Technology / Savitribai Phule Pune University (SPPU), Sangamvadi, Pune – 411001, Maharashtra, India; riyaz.jamadar@gmail.com

Abstract

Objective: This study proposes a model for building the network intrusion detection system using a machine learning algorithm called decision tree. This system detects primarily an anomaly based intrusion. **Methods**: In this model, the categorical features from the dataset Change Control IDentifiers (CCIDS) 2017 are encoded using label encoder. Using Recursive-Feature-Elimination (RFE) some best features is selected. This data is then divided into training and testing data. Training data is then used to form a Decision-Tree-Model wherein each leaf signifies the possible outcome. **Findings**: Classification models are developed making use of the training data to classify the test data as malicious or benign. Measuring the accuracy of the classifier on future data rather than the past data is of a paramount aspect. The observed accuracy of the classifier on test data is 99%. The precision of the proposed system indicates that the True-Positive-Rate (TPR) is 99.9% and the False-Positive-Rate (FPR) is 0.1%. The proposed model uses the latest data set for training data and test data compared to the traditional systems which have been modeled using KDD-CUP-99 data set. Moreover, unlike other systems, it does not use any data-mining tool like Weka. This work provides as basis for any new algorithm using dataset CCIDS 2017. **Improvements:** The work can be extended to exploit the big data available for attacks and intrusions using big data analytics.

Keywords: Accuracy, Detection, Decision Tree, Intrusion, Machine Learning

1. Introduction

Network security and Information security is a paramount concern of the growing economy. For the sake of network security, at personal level, only the installation of antivirus and firewall on the system is performed. However, for an organization, the task of handling the network security is not that simple. It not only requires the updated type of attacks but also capable to deal the data in enormous amount. The Intrusion Detection System (IDS) would detect any intrusion and alert the administrator.

Two types of attacks are possible:

1. Signature based attack, and
2. Profile based attack (Anomaly).

Signature based attack detects all the predefined attacks. The signature based files are mapped with the attacks and if matched it will return the attack type. However, it is notable to check anomaly based intrusion. Due to the available signature based file the false positive rate is low.

Profile based attacks also termed as anomaly based attacks are those attacks which doesn't use any already predefined path. The IDS which is employed to detect such kind of attack should be flexible enough to handle such anonymous scenario. It has high false-positive-rate[1].

The datasets available for training are KDD-CUP-99 and DARPA 98/99. But these datasets have become outdated. This will become a hurdle for the researchers aiming to build anomaly based intrusion detection system. Many of the recent attacks such as SSH attacks are not covered in these datasets. Thus a dataset is required which is independent of redundancy and has real-world data. The openly available dataset DARPA98/99 and KDD-CUP-99 have some limitations over updating of

*Author for correspondence

new attacks. CICIDS 2017 is one of the latest real-world dataset which overcomes all the limitations till date. It primarily includes the labeled flows, based upon the time stamp, protocols, ports, source IPs and destination IPs and the outcomes of the network traffic analysis by CIC Flow Meter. Moreover, it provides 84 feature with 4 categorical columns.

2. Literature Survey

The work proposed in[2] analyzes the various supervised machine learning classifiers based on the data sets containing the labeled instances/objects of network traffic features/parameters obtained from genuine and malicious application. Their main focus was to build NIDS which is termed as mobile based Network Intrusion Detection system. It employs ISCX Android Botnet Dataset which contains 1929 samples from botnet families of four years. Firstly the data from the ISCX Android Botnet Dataset is passed through the genuine and malicious application after which filtration and selection of feature is done resulting in the formation of labeled data. These labeled data is divided into test and training data. The training data is used to develop a model using machine learning based algorithm classifier which in turn, is used to evaluate the test data. The system has high false-positive-rate. Random Forest classifier is used to classify the data and due to its high true-positive-rate as compared to other machine learning classifier; however its false-positive-rate is slightly high. They also used weka as data mining tool which adds more overheads.

The work proposed in[3] is two tiers architecture to detect network level intrusions. They used weka data mining tool using NSL-KDD dataset. Firstly, they processed the data by building autonomous model on training set using hierarchical agglomerative clustering, further data gets classified using KNN classification and finally misuse-detection and anomaly-detection are done using multilevel perceptron and reinforcement algorithm. The use of unsupervised learning (Hierarchical clustering) to develop the data warehouse iteratively which makes their system self-autonomous. The use of weka as a data mining tool increases the overhead. The false-positive-rate is high.

The work proposed in[4] implements a classification method which is basically combines a machine learning based decision tree algorithm and multilayer perceptron. They use Artificial Neural Networks (ANNs) primarily to counteract the dataset limitations such as nonlinear, limited and incomplete. In this paper they used KDD-CUP99 as their source dataset. Firstly the dataset from KDD-CUP99 is feed to both the decision tree based approach and multilayer perceptron which classifies them and label the data into attack or benign. This label along with the data becomes the new data set which is again feed to the well trained multilayer perceptron to evaluate the test data. The major short-comings of this approach are that it doesn't account for handling big data.

In[5] develops a learning-model for fast learning network based on Particle-Swarm-Optimization and named as PSO-FLN KDD-CUP99 dataset was used. Here they considered the data as a particle, firstly the provide weights to the particle and then for each of the particle they build a fast learning network. For each of the particle they calculate the accuracy and if the fitness-value is better than the best local fitness-value it will going to set the current fitness value as best local fitness value. Update the particle position accordingly. The major short-comings of this approach is that it does not account for handling the big data and also it gives high false-positive-rate. By increasing the number of node it is possible to get improved accuracy.

The work proposed in[6] uses the KDD-CUP99 as their data set. In their work firstly they feed the data to Principal-Component-Analysis (PCA) which reduces the higher dimension dataset to lower dimension dataset. Then this new dataset is fed into various machine-learning based algorithms such as support vector machine, k-nearest neighbor, decision tree algorithm, random forest tree classification algorithm, adaboost algorithm, naïve bayes probabilistic classifier. Experiment results are analyzed and compared among the algorithms with regard to detection-rate and detection-time in which tree algorithm achieved superior results. They use the weka interface as their machine learning tool which in turn increases the overhead.

The work proposed in[7] makes use of NSL-KDD dataset. In this study, the dataset is normalized and discretized by the k-means method and the selection of feature using Information gain algorithm which is passed to the naïve bayes machine learning algorithm. They found that k-means clustering method provides better result as relative to the discretization technique of mean and standard deviation. The data after getting labeled from the k-means method is fed to the information-gain that

uses scoring methods for nominal or weighting of continuous attributes that are discredited by using the maximum entropy. The major short-comings of this approach are that the k-means method can't handle nonlinear and incomplete data. The accuracy and false-positive-rate of the system can be further improved.

3. Dataset CICIDS 2017

Network intrusion detection system requires updated data so as to train the model to work effectively in the anonymous intrusion. The openly accessible dataset KDD-CUP99, DARPA98/99 has limitations over the updating of new attacks. CICIDS 2017 dataset has genuine as well as most common attacks resembling the true real-world data (CSV's). Moreover it has the labeled-flows based upon the time stamp, protocols, ports, source IPs and destination IPs and the outcome of the network traffic analysis by CIC Flow Meter. It also covers the complete network traffic, complete capture of the data, attack diversity such as web based, brute force, DoS, DDos, infiltration, Bot and Scan and the heterogeneity of the captured data which is not covered by earlier datasets. Moreover, it provides 84 feature with 4 categorical columns. The 11 criteria important for developing a reliable dataset are: Complete Network Configuration, Complete Traffic, Labeled Dataset, and Complete Interaction, Complete Capture, and Available protocols, Attack Diversity, Heterogeneity, Feature Set and Metadata.

4. Proposed Architecture

In the proposed architecture, the data is collected from the openly accessible dataset i.e. CCIDS 2017. Then the categorical features of the data are encoded using label encoder. The label encoder is used in order to convert string data into numerical format as any machine-learning based algorithm is incapable to accept any string data. All the features/parameters of the data are not needed for developing the model therefore; some best features are selected using RFE. This data is then divided into training and testing data. Training data is then used to form a Decision-Tree-Model. Decision tree uses a tree-like structure where each leaf signifies the possible outcome.

Some of the best features are selected using RFE as all the features of the data are not necessary for building the model. A Total of 13 features were selected from 83 features. Features such as source ip, destination ip, flow bytes etc. were selected. This data is then divided into training and testing data. Training data is then used to build a decision tree model. Decision tree is a supervised learning method which needs training data for building a model and based on this training model testing data is tested.

Decision tree uses a tree-like structure where each leaf signifies the possible outcome; in this case each leaf represents the type of attack or normal behavior (benign). Test data is passed through the training model to determine whether it is benign or attack and if it resembles any attack then it will return the type of attack.

Decision trees like algorithms work through recursive partitioning of the training set in order to obtain subsets that are as pure as possible to a given target class. Each node of the tree is associated to a particular set of records T that is split by a specific test on a feature. For instance, a split on a continuous attribute A can be induced by the test $A \leq x$. The set of records T is then divided in two subsets that lead to the left branch of the tree and the right one[8-10].

$T_l = \{t \in T : t(A) \leq x\}$

and

$T_r = \{t \in T : t(A) > x\}$

Similarly, a categorical feature B can be used to induce splits according to its values. For example:

If $B = \{b_1, \ldots, b_k\}$ each branch ' i' can be induced by the test $B = b_i$.

The divide step of the recursive algorithm to induce decision tree takes into account all possible splits for each feature and tries to find the best one according to a selected quality measure.

Figure 1 signifies that each leaf represents the type of attack or normal behavior (benign).Test data is passed through the training model to determine whether it is benign or attack and if it resembles any attack then it will return the type of attack.

5. Experimental Results

For the classification problems the TPR (Success rate of detecting malicious activity) and FPR are two important factors. Classification models are developed making use of the training data to classify the test data as malicious or benign. Therefore, it is important to measure the accuracy of the classifier on future data rather than in the past data. The observed accuracy of the classifier on test data is 99%.

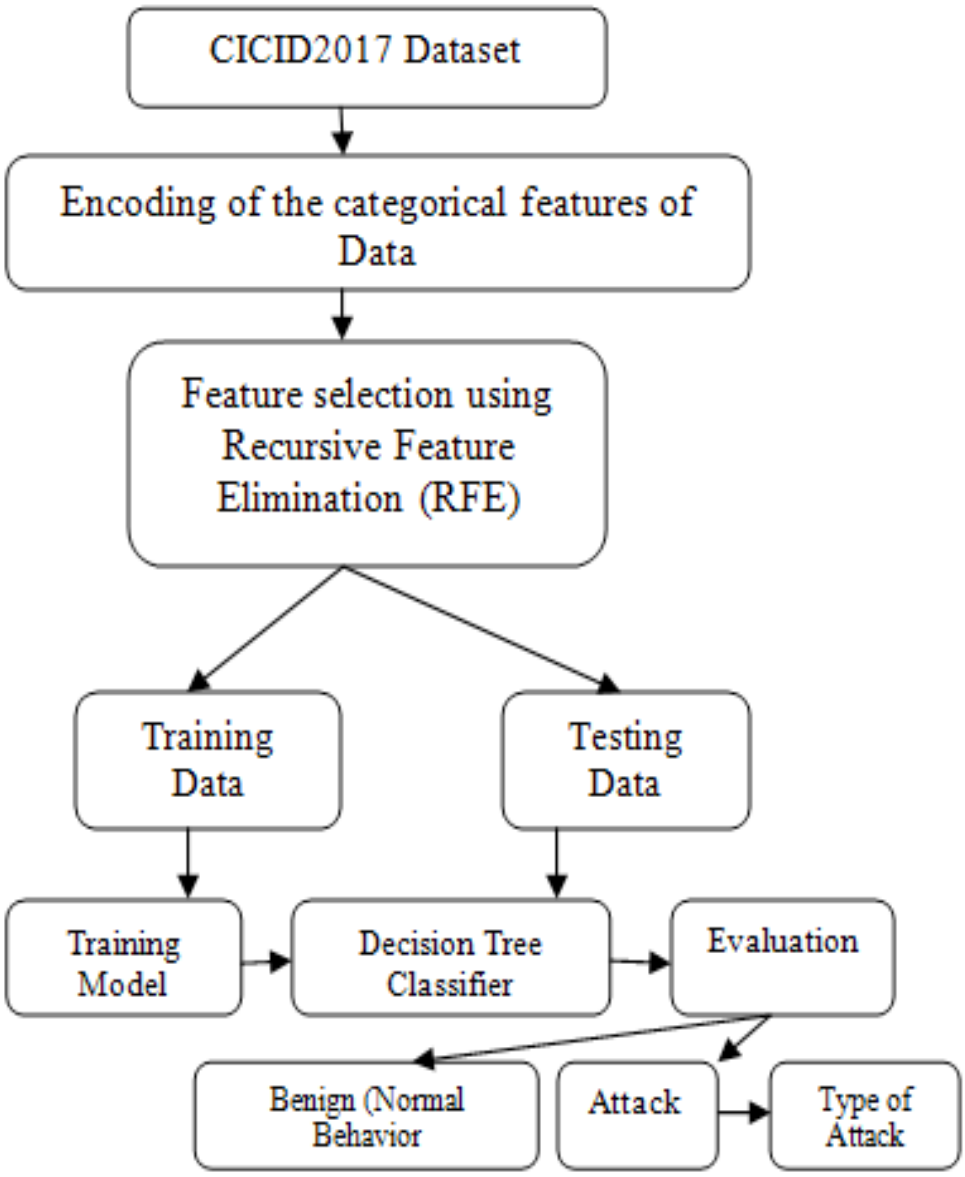

Figure 1. Working of the proposed Network Intrusion Detection system (NIDS).

In the available CICIDS 2017 dataset provide us with 84 features from which 4 are categorical feature.

Table 1 shows the f1-score provides the harmonic mean of precision and recall. The scores corresponding to every class represent the accuracy of the classifier in classifying the data points in that particular class compared to all other classes. The support indicates the number of samples of the true response that lie in that class. The training and test models have been developed using Python libraries on system of Core i3 7th Gen processor, using database SQLite3.

Table 2 shows the confusion matrix developed by analyzing the test data. This matrix helps us to calculate the accuracy of the proposed system. The precision of the proposed system indicates that the TPR is 99.9% and the FPR is 0.1%.

6. Conclusion and Future Work

Previously KDD-Cup99 Dataset was considered as the benchmark dataset for intrusion-detection but Nowadays, the network and the attack methods have changed drasti-

Table 1. Classification report

Attribute	Precision	Recall	f1-score	Support
0.0	1.00	1.00	1.00	265572
1.0	0.95	0.95	0.95	41
2.0	1.00	1.00	1.00	2479
3.0	1.00	0.98	0.99	195
4.0	1.00	1.00	1.00	5938
5.0	1.00	1.00	1.00	4
6.0	0.69	0.61	0.65	18
7.0	1.00	1.00	1.00	7625
8.0	1.00	1.00	1.00	260
9.0	1.00	1.00	1.00	740
10.0	1.00	0.86	0.92	14
11.0	1.00	1.00	1.00	323
Average / total	1.00	1.00	1.00	283209

cally so use CICIDS 2017 dataset is used. Thus, it can be used to detect attacks based on current network scenario.

The approach based on Decision Tree is presented and discussed to develop an efficient intrusion detection model. The experimental results demonstrate that the proposed approach can be used to develop an Intrusion-Detection-Model having high detection rate, high accuracy (99.9%) and low False-Positive-Rate.

The future work would be collecting real time packets from the network and testing them against the already classified training dataset. Based on results achieved this work can be extended to host based IDS or analysis on an application level.

7. Acknowledgement

The Author would like to thank UNB Canadian Institute of Cyber security for providing CICIDS 2017 dataset for this work.

8. References

1. Jamadar RA, Himani Gupta, Ankit Baghel & Rituraj. Study and analysis of hadoop based Network Intrusion Detection System, International Journal of Engineering and Science Invention. Dec 2017; 6(12):1–4. http://www.ijesi.org/papers/Vol(6)12/Version-4/A0612040104.pdf.
2. Kumar S. Viinikainen A. & Hamalainen T. Machine learning classification model for network based intrusion detection system, 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST), 5-7 Dec. 2016. IEEE, Barcelona, Spain; Dec 2016. p. 242–49. https://ieeexplore.ieee.org/document/7856705/citations#citations.
3. Divyatmika & Manasa Sreekesh. A Two-tier Network based Intrusion Detection System Architecture using Machine Learning Approach. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 3-5 March 2016. IEEE, Chennai, India; Mar. 2016. p. 42–47. https://ieeexplore.ieee.org/document/7755404.
4. Jamal Esmaily, Reza Moradinezhad & Jamal Ghasemi. Intrusion Detection System Based on Multi-Layer Perceptron Neural Networks and Decision Tree. 2015 7th Conference on Information and Knowledge Technology (IKT), 05 October 2015. IEEE, Urmia: Iran; 2015. https://doi.org/10.1109/IKT.2015.7288736.

Table 2. Confusion matrix

265563	2	0	3	0	0	7	0	0	0	0	0
2	39	0	0	0	0	0	0	0	0	0	0
0	0	2479	0	0	0	0	0	0	0	0	0
0	0	0	192	0	0	0	0	0	0	0	0
2	0	0	0	5938	0	0	0	0	0	0	0
0	0	0	0	0	4	0	0	0	0	0	0
5	0	0	0	0	0	11	0	0	0	0	0
0	0	0	0	0	0	0	7624	0	0	2	0
0	0	0	0	0	0	0	0	260	0	0	0
0	0	0	0	0	0	0	0	0	740	0	0
0	0	0	0	0	0	0	0	0	0	12	0
0	0	0	0	0	0	0	0	0	0	0	323

5. Ali MH, Bahaa Abbas Dawood Al Mohammed, Alyani Ismail & Mohamad Fadli Zolkipli. A new intrusion detection system based on Fast learning network and swarm optimization. IEEE, 2018; 6:20255–61. https://doi.org/10.1109/ACCESS.2018.2820092.
6. Chabathula KJ, Jaidhar CD & Ajay Kumara MA. Comparative study of principal component analysis based intrusion detection approach using machine learning Algorithms. 3rd International Conference on Signal Processing, Communication and Networking (ICSCN); 2015. p. 1–6. https://doi.org/10.1109/ICSCN.2015.7219853.
7. Effendy DA, Kusrini Kusrini & Sudarmawan Sudarmawan. Classification of intrusion Detection System (IDS) based on computer network. 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). IEEE, Yogyakarta: Indonesia; 2017. p. 90–94. https://ieeexplore.ieee.org/document/8285566.
8. Mathematics Behind Classification and Regression. Date accessed: 26.11.2012. https://stats.stackexchange.com/questions/44382/mathematics-behind-classification-and-regression-trees.
9. Sharafaldin I, Arash Habibi Lashkari & Ali A. Ghorbani Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. 4th International Conference on Information Systems Security and Privacy (ICISSP); Jan 2018. p. 108–16. http://www.scitepress.org/Papers/2018/66398/66398.pdf.
10. Gharib A, Iman Sharafaldin, Arash Habibi Lashkari & Ali A. Ghor. An evaluation framework for intrusion detection data set. International Conference on Information Science and Security (ICISS), 19-22 Dec. 2016. IEEE, Pattaya: Thailand; 2016. p. 1–6 https://doi.org/10.1109/ICISSEC.2016.7885840. PMid: 27047644, PMCid: PMC4818783.